MONASH University

Computational Investigations of Ligand Binding Pathways

Trayder Thomas

B.Sc.

A thesis submitted for the degree of *Doctor of Philosophy* at

Monash University in 2017

Department of Medicinal Chemistry

Copyright notice

Contents

Abstract

Knowledge of how ligands bind to their macromolecular targets is critical to rational drug design. Most of structure-based drug design and efforts to optimize the binding affinity are performed using only the final bound pose of a drug, however there is an increasing appreciation that the kinetics of drug binding can directly impact its efficacy. Experimental methods can be used to determine the bound pose of a drug (e.g. X-ray crystallography) or the on/off rates to that pose (e.g. surface plasmon resonance), but these methods are not suitable for many target systems, and it is frequently necessary to employ computational methods to augment experimental findings.

In this work, we conducted a homology modeling study and developed effective models of the not-yet-crystallized muscarinic acetyl choline receptors based on an experimentally determined structure of the β2 adrenergic receptor from another sub-family. We found that these homology models, trained with experimental knowledge, outperformed the crystal structures in virtual screening. This study also reinforced the understanding that the predictive power of both the crystal structures and models was limited to the immediate vicinity of the co-crystallized or training ligands.

It is Important to realize that drug binding is more than a two-state process. Ligands interact with the receptor long before they reach the bound pose, often in well-defined metastable states. These metastable states often exist for short timescales that are difficult to access experimentally, and as such their impact on the greater binding process is poorly understood. To investigate the role of metastable states in ligand binding, we conducted conventional molecular dynamics simulations to observe the never before simulated binding pathways of the drugs clozapine and haloperidol to the D_2 and D_3 dopamine receptors. For each ligand, we were able to identify metastable states, in addition to an entire binding pathway. Transitions between these metastable states, as well as to and from the final binding site itself, were rare events. Although we observed ligand binding, we were not able to garner an appreciation for the kinetics involved or whether other binding pathways existed.

To better characterize the binding process, we employed Markov state models (MSMs). When constructing a MSM, the molecular system being investigated, is broken down into discrete states, and the metastability and kinetics of each of these states can then be estimated. MSMs are most commonly applied to protein folding, so we developed a methodology more suitable to ligand binding. This methodology was developed and applied to two very different targets, a G protein-coupled receptor and a fatty acid-binding protein. Through the use of this methodology we were able to identify multiple binding pathways and several metastable sites in each target, furthering our understanding of the binding process.

Thesis including published works declaration

I hereby declare that this thesis contains no material which has been accepted for the award of any other degree or diploma at any university or equivalent institution and that, to the best of my knowledge and belief, this thesis contains no material previously published or written by another person, except where due reference is made in the text of the thesis.

This thesis includes 2 original papers published in peer reviewed journals and 0 submitted publications. The core theme of the thesis is ligand binding. The ideas, development and writing up of all the papers in the thesis were the principal responsibility of myself, the student, working within the Faculty of Pharmacy and Pharmaceutical Sciences under the supervision of David K. Chalmers and Elizabeth Yuriev.

(The inclusion of co-authors reflects the fact that the work came from active collaboration between researchers and acknowledges input into team-based research.)

In the case of *Chapter 2* and *3* my contribution to the work involved the following:

Thesis Chapter	Publication Title	Status	Nature and % of student contribution	Co-author name(s) Nature and % of Co-author's contribution*	Co-author(s), Monash student
2	Homology Modeling of Human Muscarinic Acetylcholine Receptors.	Published	50%. Generating and analysing data, writing manuscript.	Kimberley C. McLean, Generating and analysing data, writing manuscript. 50%	No
3	Ligand Binding Pathways of Clozapine and Haloperidol in the Dopamine D2 and D3 Receptors	Published	90%. Developing methodology, generating and analysing data, writing manuscript	Yu Fang, Developing methodology, and generating data. 10%	No

I have not renumbered sections of submitted or published papers in order to generate a consistent presentation within the thesis.

Student signature: [redacted] **Date:** 31/08/2017

The undersigned hereby certify that the above declaration correctly reflects the nature and extent of the student's and co-authors' contributions to this work. In instances where I am not the responsible author I have consulted with the responsible author to agree on the respective contributions of the authors.

Main Supervisor signature: [redacted] **Date:** 20/09/2017

Acknowledgements

Foremost, I would like to extend an especial thanks to my supervisors, Dr David K. Chalmers and Dr Elizabeth Yuriev, for their support throughout this project. I would then like to thank them again for the many constructive arguments and their support or tolerance of my quirks.

Many thanks go to Tamir, Matt, Stephen, Anitha, and Mitch for the many useful office discussions, from which I have developed a better understanding of not just my own work but of science in general.

Thanks go to my sister, Briar, who proofread parts of this thesis.

This research was supported by an Australian Government Research Training Program (RTP) Scholarship.

Lastly, I am grateful to all the great friends I have spent time with over the past few years, whether sharing a lunch break or drinks on a Friday night.

Publications

The following publications have been produced during the course of this work:

Thomas, T.; McLean, K. C.; McRobb, F. M.; Manallack, D. T.; Chalmers, D. K.; Yuriev, E. Homology Modeling of Human Muscarinic Acetylcholine Receptors. *J. Chem. Inf. Model.* 2014, 54, 243–253.

Thomas, T.; Fang, Y.; Yuriev, E.; Chalmers, D. K. Ligand Binding Pathways of Clozapine and Haloperidol in the Dopamine D2 and D3 Receptors. *J. Chem. Inf. Model*. 2015, 56, 308–321.

Thomas, T.; Chalmers, D. K.; Yuriev, E. Homology Modeling and Docking Evaluation of Human Muscarinic Acetylcholine Receptors. In Muscarinic Receptor: From Structure to Animal Models; Myslivecek, J., Jakubik, J., Eds.; *Neuromethods*; Springer New York: New York, NY, 2016; Vol. 107, pp 15–35.

Nguyen, A. T. N.; Baltos, J.-A.; **Thomas, T.**; Nguyen, T. D.; Muñoz, L. L.; Gregory, K. J.; White, P. J.; Sexton, P. M.; Christopoulos, A.; May, L. T. Extracellular Loop 2 of the Adenosine A1 Receptor Has a Key Role in Orthosteric Ligand Affinity and Agonist Efficacy. *Mol. Pharmacol.* 2016, 90, 703–714.

Nguyen, A. T. N.; Vecchio, E. A.; **Thomas, T.**; Nguyen, T. D.; Aurelio, L.; Scammells, P. J.; White, P. J.; Sexton, P. M.; Gregory, K. J.; May, L. T.; Christopoulos, A. Role of the Second Extracellular Loop of the Adenosine A1 Receptor on Allosteric Modulator Binding, Signaling, and Cooperativity. *Mol. Pharmacol.* 2016, 90, 715–725.

John, T.; **Thomas, T.**; Abel, B.; Wood, B. R.; Chalmers, D. K.; Martin, L. L. How Kanamycin A Interacts with Bacterial and Mammalian Mimetic Membranes. *Biochim. Biophys. Acta - Biomembr.* 2017, 1859, 2242–2252.

Chapter 1

Introduction

1.1 Ligand binding

Ligand binding is an extraordinarily complicated process that is the central focus for the fields of medicinal chemistry and pharmacology. The ligand-binding process is not directly observable experimentally and it is common to represent the approach with simple approximations of the time-averaged behavior. The most frequently used of these approximations is to describe ligand binding solely by binding affinity. Binding affinity is a thermodynamic property that quantifies the strength of ligand binding, and is typically reported using the dissociation constant (K_d), which is inversely proportional to the affinity. K_d is the ratio of concentrations of the reactants (ligand and receptor) and products (the receptor-ligand complex) at an assumed equilibrium or, when examining an individual complex over time, the ratio of time the system exists in each of those states. Binding affinity is well illustrated by the simplest models of ligand binding (Figure 1), e.g. the lock-and-key model,[1] in which the drug and the orthosteric site of the receptor click together like a pair of rigid jigsaw pieces. This model does not account for the flexibility and dynamics of the system, yet provides a workable approximation for the time-averaged behavior of the complex. The induced fit model[2] improves on the lock-and-key model by accounting for the flexibility of the ligand and binding site, introducing the concept that the conformation of the receptor can adapt to accommodate the ligand. Both of these models present a bound and unbound state of the complex suitable for interpretation with binding affinity.

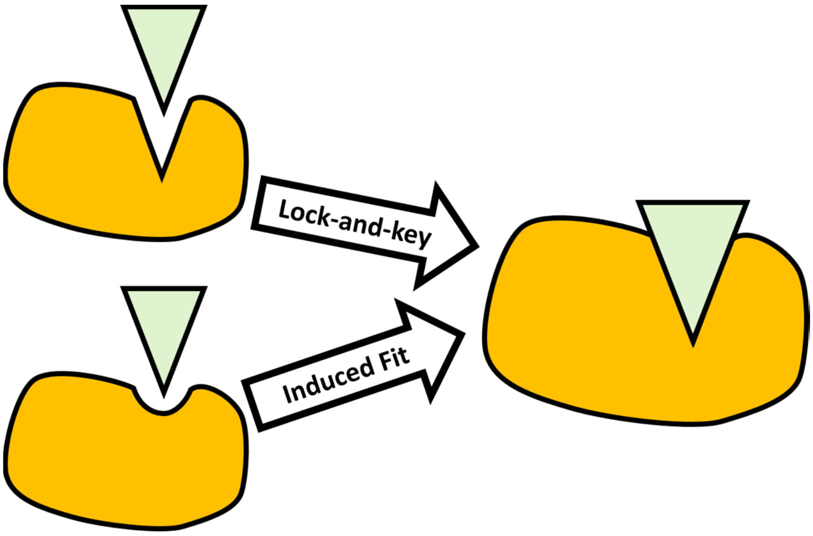


Figure 1. The lock-and-key (top) and induced fit (bottom) models of ligand binding present an easy to understand, but oversimplified, model of ligand binding.

Binding affinity can also be expressed kinetically as the ratio between the rate constants of binding and unbinding (k_{on} and k_{off}) as shown in Equation (1). It can be important to consider these rates individually, rather than only consider K_d as a ratio, as the on- and off-rates individually affect the characteristics of the drug. In the case of competitive antagonists, a slower off-rate, and therefore longer residence time, allows for a more effective blockade of the receptor, which will in-turn minimize the windows in which an agonist can bind.[3] Long residence times have been shown in many cases to correlate more strongly with the efficacy of a drug than the binding affinity does.[4] Seow et al. found in a comparison of 3 equipotent antagonists that, compared to 2 drug-like small molecules, the peptide drug with negligible oral bioavailability was still the most orally efficacious due to its significantly longer residence time. However, long residence times are not always desirable; when a blockade is too effective due to a slow off-rate drug, the receptor can be prevented from performing its biological function, leading to a more undesirable side-effect profile than a fast-on, fast-off drug, which would allow the normal biological function to proceed, albeit at a reduced rate.[3]

$$K_d = \frac{k_{off}}{k_{on}} \tag{1}$$

Whilst treating drug binding kinetically provides more insight into drug binding than using binding affinity, the approximation is still being made that the receptor only exists in two states. In reality, the ligand and receptor, both individually and as a complex, exist as an ensemble of conformations. A ligand does not “toggle” between being bound and unbound, instead it must traverse the space, both conformational and Euclidean, between these two states. A ligand can follow multiple pathways between the bound and unbound conformations, with each pathway consisting of a number of states and each state to state transition having an energy barrier that must be overcome (Figure 2).[5,6] The experimentally observed kinetics of ligand-binding arise from a combination of the rates between all of the states that make up the binding landscape. In any binding pathway, some of the intermediate states will be relatively long-lived compared to others and can be considered metastable states. The concept of metastable states has been well demonstrated by Dror et al. in their simulation of alprenolol binding to the β_2-adrenoceptor[7] where they observed the ligand passing through several well-defined metastable states on the way to the bound pose and determined that the largest barrier to drug binding was far-removed from the experimentally determined bound state[8] utilized in structure-based drug design. Departing from the two-state model of drug binding and considering metastable states opens up more avenues to drug design. An antagonist can be efficacious while occupying metastable states prior to arriving at the bound pose if in these states it blocks the binding pathways of an agonist,[9,10] likewise it can be reasoned that an agonist cannot be distinguished from an antagonist while outside of states that can induce activation of the receptor.

The receptor conformation can be as important to drug binding as the conformation of the drug and, compared to the drug, the receptor will typically explore its conformational space over a much longer timescale. Drugs may be selective for specific conformations of the receptor, and the receptor may only spend a fraction of the time in these amenable conformations.[11] Reaching an amenable conformation can be the rate-determining step to drug binding, and the conformational changes required can be largely independent to the binding of a drug.[12]

There is now great interest in compounds that bind outside the orthosteric site, functioning by allosteric modulation, or as bitopic ligands. Allosteric modulation occurs when an additional molecule binds to an allosteric site on the receptor and, either through inducing conformational change in the receptor or by blocking the unbinding pathway, alters the behavior of the orthosterically bound drug. A wide range of protein targets have been shown to be susceptible to allosteric modulation, including G protein-coupled receptors, ion channels, nuclear hormone receptors, and kinases.[13] The allosteric modulators themselves can range from small molecules to antibodies.[14] Typically, only the highest affinity states of a complex can be observed experimentally, but the existence of metastable states can often be indirectly observed in the case of allosteric modulators. The existence of metastable states is particularly evident in the case of ago-allosteric modulators,[15,16] agonists that also act as allosteric modulators, which most likely act by binding to both orthosteric and allosteric sites. Metastable sites along the binding pathway have been shown to coincide with allosteric sites; Kruse et al. performed long-timescale simulations of tiotropium binding to the M_3 muscarinic acetylcholine receptor.[17] In these simulations, they found that tiotropium paused at a distinct metastable site as it bound to, or dissociated from, the receptor, and that this site had been identified as an allosteric site through mutagenesis.[18] This correlation with allosteric sites makes metastable sites of interest as starting points for the design of allosteric or bitopic ligands.[19] Similarly to allosteric modulators, bitopic ligands also bind outside of the orthosteric site. The principle behind bitopic ligands is to combine two pharmacophores, an orthosteric pharmacophore to provide affinity, and a second pharmacophore, often allosteric, to provide selectivity.[20] While bitopic ligands may occupy an allosteric binding site, they do not necessarily act as allosteric modulators, and any metastable site that differs between two receptors could be a useful target for the second pharmacophore.

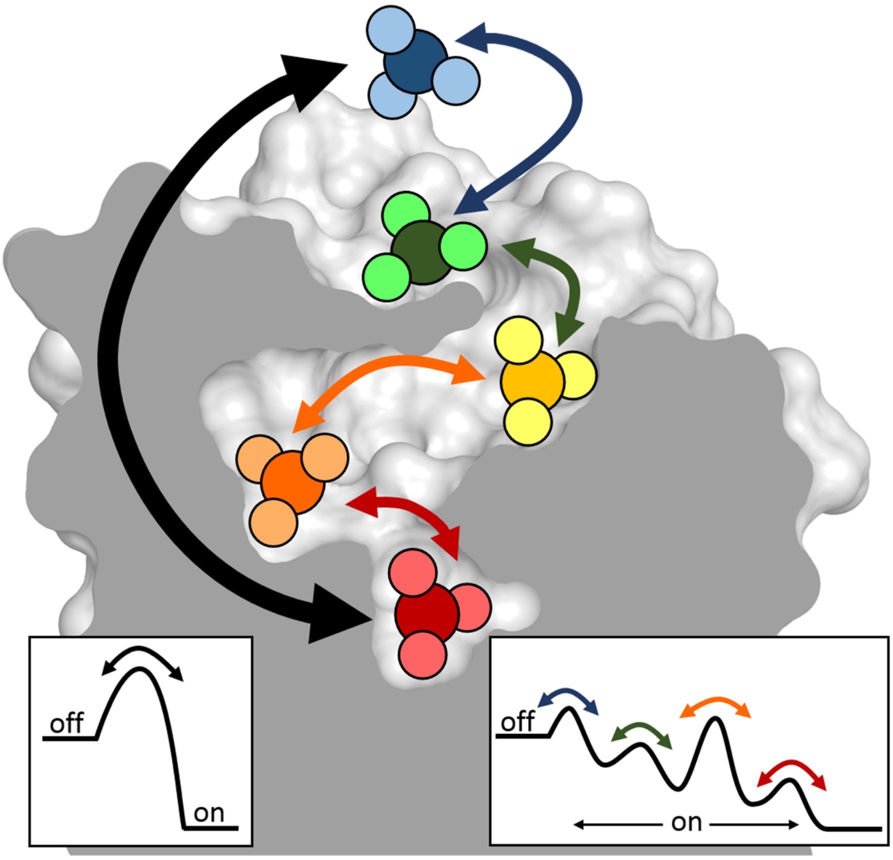


Figure 2. The two-state approximation of ligand-binding (black arrow) obscures many of the complexities of the binding pathway. The colored pathway provides a more realistic model of how a ligand binds to a receptor.

Although treatment of drug binding using simple approximations has thus far proven sufficient for drug development, subscribing to these approximations necessitates observing drug binding through a narrow lens, effectively limiting the understanding of the binding process. Now that computers are becoming powerful enough to make routine simulation of drug binding a possibility, researchers should begin to consider the behavior of ligands outside of the bound state. This would allow finer-tuning of the kinetic behavior of a drug and provide additional binding sites to target on each receptor, which may allow for greater drug selectivity or control of drug behavior through allosteric modulation. In order to investigate behavior outside of the bound state, we need to develop the methodologies that can be used to efficiently explore the intricacies of drug binding and improve our understanding of the ever-important issue of how drugs actually bind.

1.2 Target systems

1.2.1 G protein-coupled receptors

G protein-coupled receptors (GPCRs) are the largest and most diverse super-family of proteins in the human body[21] and are at the forefront of studies into the intricacies of ligand binding. GPCRs are membrane-spanning receptors that function by transducing chemical signals across the cell membrane and then effecting an intracellular response by binding to intracellular G proteins and beginning a signaling cascade.[22] All GPCRs share a similar topology[23] (Figure 3); a bundle of 7 transmembrane spanning helices that opens into an extracellular binding pocket. GPCRs are critical to the senses of sight, smell, and taste, but the crux of their pharmaceutical relevance is their critical intermediate roles in the control of the autonomic nervous system, the immune system, moods, and behavior.[24] These roles, widespread throughout the body, have led to GPCRs being the most targeted receptors in drug development, with an estimated 33% of all small-molecule drugs targeting GPCRs.[25]

There are 5 families of GPCRs according to the GRAFS classification system:[21] Glutamate, Rhodopsin, Adhesion, Frizzled/Taste2, and Secretin. Of these classes, the rhodopsin family is the largest, containing ~85% of all GPCRs, although the majority of these are predicted to relate to taste or smell. Rhodopsin family GPCRs are the most investigated class, and the most druggable sub-family are the biogenic amine receptors, including the muscarinic and dopamine receptors that are investigated in this work. The most notable disease states related to dopamine receptors are schizophrenia, addiction, and Parkinson’s disease.[26] The biological functions of muscarinic receptors is less clear due to a lack of selective drugs, but they have been found to be involved in heart disease, Parkinson’s disease, and Alzheimer’s disease.[27]

The determination of the structure of the light-activated rhodopsin through X-ray crystallography in the year 2000[28] resulted in a figurative explosion in structural studies of GPCRs. In 2007, the first crystal structure of a biogenic amine receptor, the β_2 adrenergic receptor (β_2AR), was determined with the co-crystallized antagonist carazolol.[29] In the following years, more structures were determined including structures bound to agonists[30] and those exhibiting an active conformation of the receptor.[31,32] The crystal structures we have focused on in this work are inactive conformations of the D_3 dopamine receptor (human), determined in 2010,[33] and the M_2 (human)[18] and M_3 (rat)[17] muscarinic acetylcholine receptors, determined in 2012.

It is now understood that GPCRs, and other receptors, truly exist as an ensemble of states, and through drug binding, these receptors are induced toward groups of conformations broadly classified as active or inactive states. GPCRs are often constitutively active, exploring active conformations even in the absence of an agonist, so that an inverse agonist, rather than an antagonist, is required to limit their ensemble to inactive states.[34] Important differences have been observed within the group of states that forms the active ensemble; different agonists, binding to the same general binding site, have been observed to bias the intracellular response towards different signaling pathways.[35,36] There are many states involved in the ligand binding ensemble, and the important states differ for each individual ligand. Despite the advances made in GPCR crystallography, it still remains a challenging endeavor to crystallize membrane proteins, and isolating specific states relevant to ligand binding is trickier still.

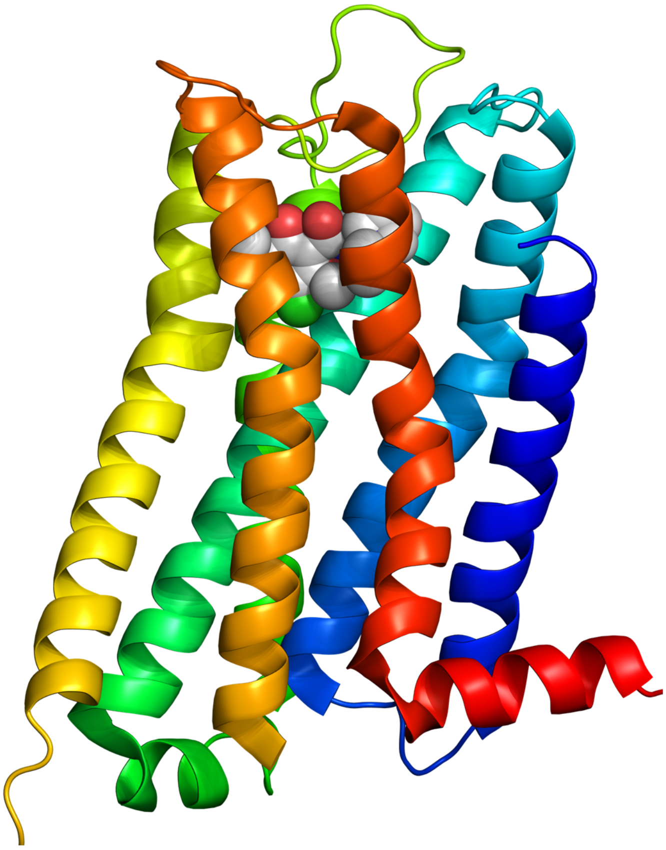


Figure 3. A cartoon representation of the D_3 dopamine receptor (PDB ID: 3PBL). The receptor is colored from the *N*-terminus (blue), through cyan, green, yellow, and orange, to the *C*-terminus (red) and the co-crystalized ligand eticlopride is shown as spheres in the orthosteric binding site.

GPCRs within a family have a very high homology between their transmembrane domains, especially in the region of the binding site. When targeting GPCRs in drug development, in many cases the largest obstacle is selectivity. Amongst the biogenic amine receptors, it is common for one drug to be active at multiple subtypes of receptor. In the case of muscarinic receptors, due to the lack of selective drugs, the individual function of each subtype is unclear. The high homology between GPCRs has led to an interest in the development of drugs that bind outside the highly conserved orthosteric site, i.e. allosteric drugs or bitopic ligands that extend outside of the orthosteric site to interact at another, often allosteric, location.

The dopamine receptor antagonists we have studied in this work are clozapine and haloperidol (Figure 4), which are both classified as antipsychotics, a therapeutic class that was first developed in the 1950s. The first generation of antipsychotics (later classified as typical antipsychotics) was found to cause problematic extrapyramidal side-effects,[37] which led to the development of a second generation of drugs, the atypical antipsychotics. Atypical antipsychotics have very similar side-effects to typical antipsychotics, and it is unclear if they should be considered an improvement over the first generation drugs.[38,39] Despite their unfavorable side-effect profiles, haloperidol (a typical antipsychotic) has been in use since the 1950s[40] and clozapine (an atypical antipsychotic) since the 1970s.[41] Both of these drugs are classified as WHO essential medicines.[42] Each of these drugs represents a basic scaffold from which many other antipsychotics have been developed,[43,44] and every antipsychotic has its own unique side-effect profile. The abundance of side-effects leaves much room for improvement in antipsychotic drugs. Current trends focus on allosteric modulation and partial agonism, but without an understanding of how these drugs bind, structural knowledge is tethered to the limited available crystal structures.[45]

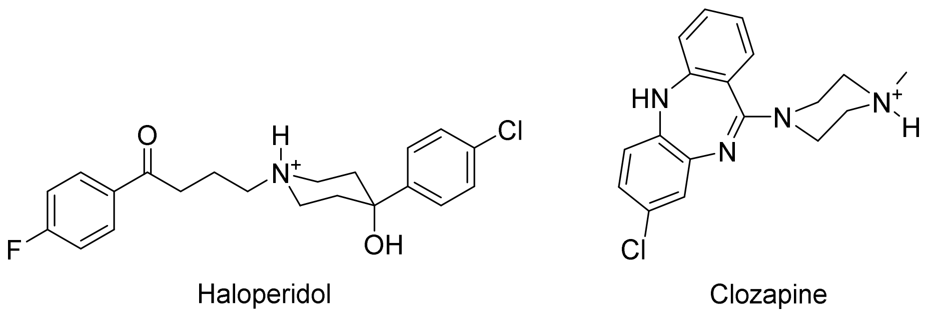


Figure 4. Haloperidol and Clozapine shown in their biologically relevant ionization states.

1.2.2 Fatty acid-binding proteins

Fatty acid-binding proteins (FABPs) are a class of small cytosolic transport proteins that are primarily responsible for the transport of fatty acids within the cell, although they are capable of binding many other lipophilic small molecules.[46] FABP expression has been shown, through mouse models, to be tied to insulin sensitivity and levels of glucose or cholesterol in the blood, making FABPs drug targets for obesity, heart disease, and diabetes.[47] The high expression levels of FABPs have also made them useful as biomarkers for these disease states.[48–50] FABPs have also been observed to cross the nuclear membrane[51] which, combined with their ability to bind lipophilic molecules, gives them a potential application as drug transporters.[52]

There are 9 types of FABPs encoded in the human genome.[53] These proteins are named FABP1-9 or by the organ they were first, or are predominantly, found in. Each type of FABP can be found in multiple organs, often alongside other types. All FABPs share the common topology of a 10-strand β-barrel, capped on one side with a helix-turn-helix cap (Figure 5).[54] Increased flexibility of the β-strands on the capped side of the receptor results in this side of the β-barrel being considered the open side through which ligands enter and exit.[55] Fatty acids bind within the hydrophobic β-barrel, forming a salt-bridge with an arginine residue. In this work we investigate L-FABP (also known as FABP1), which has a larger binding pocket than other FABPs and is capable of binding 2 fatty acids simultaneously.[56] Similar to GPCRs, the deep binding pocket within L-FABP necessitates a multiple-step ligand-binding process. In this work, we chose to use FABPs as a model system to investigate ligand binding in addition to GPCRs. Compared to GPCRs, the small size of FABPs and no need for a membrane environment make them easier to work with.

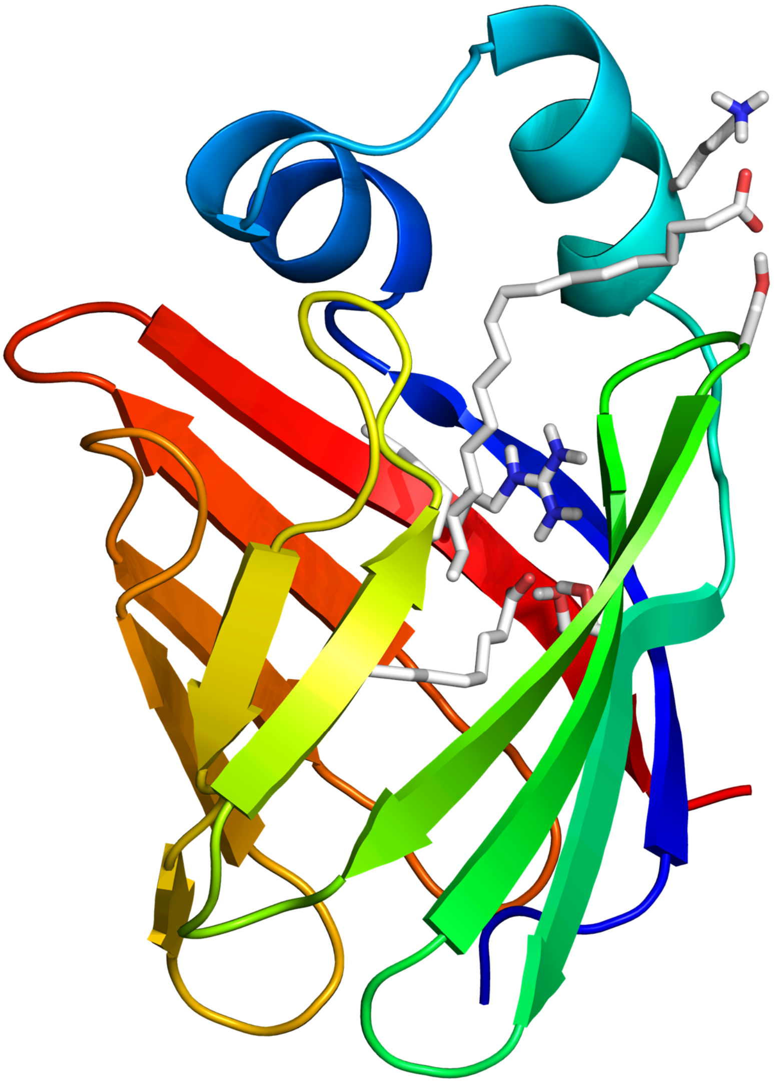


Figure 5. NMR structure of the liver-fatty acid-binding protein (PDB ID: 2LKK[56]), showing two bound oleic acids and the residues they form polar interactions with. The protein is colored from the *N*-terminus (blue), through cyan, green, yellow, and orange, to the *C*-terminus (red). The two helices form a cap to the binding pocket contained in the β-barrel.

1.3 Computational methods for studying ligand binding

1.3.1 Docking

Molecular docking is a tool used to predict the binding of existing or potentially new molecules to a target (receptor).[57] Docking involves conformational and orientational sampling and scoring to predict energetically favorable poses of the ligand in a target receptor structure. Docking is computationally efficient; high throughput docking methods such as virtual screening commonly utilize a rigid receptor model that can be likened to the lock-and-key model of ligand binding, although it is acknowledged that incorporating receptor flexibility can significantly improve the results.[58] Docking requires a structure to dock into, and these structures are generally obtained through crystallography or NMR experiments, which can limit the viable targets to the availability of specific receptor structures.

The range of viable targets for docking can be expanded through homology modeling; based on the knowledge of the homologies between a target receptor and a template crystal structure, it is possible to use homology modeling to generate a “best-guess” model of the target.[59,60] The most important pre-requisite for a homology model is to have a template that closely resembles the target receptor. The target and template sequences should have the highest level of homology possible, especially in the regions of interest such as the binding site. The target sequence is fitted to the template structure and any additional experimental information, such as secondary structure, is included. The loop regions of the protein often have a very poor homology and are modelled separately. Homology models inherit many of the weaknesses of the template crystal structures: they are likely to preserve any errors or artifacts present in the template structure, and they are still a static structure that poorly represents flexible regions. Homology modeling has proven very important in GPCR research.[61] While the number of crystal structures of GPCRs is steadily increasing,[62] there are still limited template crystal structures available, and crystallizing a single state from the receptor ensemble is a non-trivial process that must be repeated for every receptor, state, and ligand of interest. This leaves large holes in our structural knowledge that can be readily filled by homology models.

Once the initial homology model is available, it can be evaluated and trained for virtual screening. Virtual screening is the process of docking large libraries of ligands into a target receptor and is often performed for two main purposes.[63] Prospective virtual screening attempts to predict hits from a library of compounds of unknown activity, whilst retrospective virtual screening can be used to assess the predictive qualities of a model by testing its ability to rank known actives over a library of decoy compounds. Unless one is searching for ligands of a particular scaffold, it is important to optimize the homology model to allow for the binding of a diverse range of ligands. The binding site of the template structure will likely be biased towards a particular ligand or, in the absence of a ligand, the binding site may be closed and unreceptive to binding. In order to use a homology model built from a template structure to predict ligand binding, the binding site will first need to be opened up. A useful approach for this is to perform induced fit docking into the binding site with a training ligand, thereby adjusting the conformation of the receptor to accommodate a more generic or larger ligand. The bias introduced through the template structure also means that docking is typically only useful for exploring known binding locations.

1.3.2 Molecular dynamics

Molecular dynamics (MD) is a computational method in which the time-dependent behavior of a system is simulated using simple Newtonian physics, allowing an atomistic view of the system.[64] In MD simulations, the forces affecting each atom are evaluated each time step, typically a 1-5 fs interval, according to a force field. The position of the atoms is then updated and the forces are evaluated at the new positions. Repeating this evaluation over many time steps allows access to the dynamics of the system in nanosecond to microsecond timescales. Millisecond length atomic simulations are possible only with highly specialized hardware.[65] It is worth noting that, while it is also possible to employ full, or hybrid, quantum mechanical MD, we will restrict discussion to “classical” MD as only this approach was used in this work. MD simulations require a force field, which enables the evaluation of forces at each individual time step, and also the algorithms that perform the calculations, ensure the efficiency of this process, or control system properties such as temperature and pressure.[66]

Molecular mechanics force fields describe the properties of each atom in the system and how they interact with each other, allowing the calculation of the forces acting on each atom every time step of a MD simulation. Force fields are generally parametrized to match experimental values. Many different force fields have been created; in this work we have used 2 molecular dynamics force fields, the CHARMM all-atom force field[67,68] and the GROMOS united-atom force field.[69] Both force fields have both been extensively validated but were developed using different philosophies. The CHARMM force field describes hundreds of different atom types, each reflecting the different behavior of each element in subtly different environments. The GROMOS force field instead chooses to describe atoms more generally, only describing a few additional atom types for larger changes in environment and utilizing united-atom types to describe groups of bonded atoms. In general, the interactions in each force field can be divided into two major components: bonded and non-bonded interactions.[70–72] Bonded interactions include terms for the bonds, angles, and dihedrals present in a molecule (Figure 6). Bonds and angles are typically described with harmonic constraints, whilst dihedrals are described using a cosine series. Non-bonded interactions include van der Waals (vdW) forces and electrostatics, which are calculated at short range using the Leonard-Jones equation and coulombic equations respectively. It is standard practice to calculate these forces differently beyond a cut-off distance. In this work, calculation of vdW forces switches to a rapidly converging function at the long-range cut-off, whilst long-range electrostatic forces are approximated using particle mesh Ewald (PME).[73]

$$\begin{aligned} U = & \sum_{bonds} \frac{1}{2} k_r (r - r_0)^2 \\ & + \sum_{angles} \frac{1}{2} k_\theta (\theta - \theta_0)^2 \\ & + \sum_{dihedrals} k_\phi \left(1 + \cos(n\phi + \phi_0)\right) \\ & + \sum_{i} \sum_{j \neq i} 4\varepsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right] \\ & + \sum_{i} \sum_{j \neq i} \frac{q_i q_j}{4\pi\varepsilon_0 r_{ij}} \end{aligned}$$

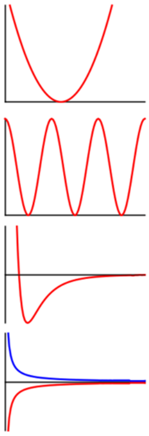


Figure 6. General equations for the GROMOS force field; the potential energy (U) of the system is described by the energetic sum of all bonds, angles, dihedrals, Lennard-Jones (van der Waals) interactions, and Coulombic (electrostatic) terms. k_r, k_θ, k_ϕ are constants, r is the bond length, r_0 is the ideal bond length, θ is the angle, θ_0 is the ideal angle, ϕ is the dihedral angle, n is the periodicity, ϕ_0 is the phase shift angle, ε_{ij} is the Lennard-Jones well depth, σ_{ij} is the Lennard-Jones radius, r_{ij} is the distance between atoms, q_i and q_j are atomic charges, ε_0 is the effective dielectric constant.

MD simulations apply several algorithms to better match experimental, or real world, conditions. Most applications do not involve closed systems, and thus algorithms are required to regulate the temperature and pressure of the system and the behavior of atoms as they reach the systems boundary. Periodic boundary conditions (PBC) are used in MD simulations to avoid the boundary effects caused by a finite system. PBCs are usually applied in 3 dimensions, infinitely tessellating the molecular system, or unit cell, to better approximate the scale of a real world system. The PBCs are usually defined such that atoms leaving the unit cell through one face, re-enter through the opposite face. While using PBCs eliminates the most severe boundary artifacts, it does not eliminate them entirely. In particular, long-range electrostatic interactions still occur between a protein and its neighboring periodic image, thus it is essential to construct a system with a sufficient divide of solvent atoms between periodic images (typically >15 Å) to dampen these interactions. Another aspect of open systems is the ability to exchange energy with the surroundings. In MD simulations, the surroundings are approximated by coupling the system to a thermostat. This is commonly called temperature coupling. The temperature is computed from the system kinetic energy; temperature coupling ensures that the size of temperature fluctuations are appropriate and that the average temperature of the system is maintained at the desired value. Calculating the forces in a system over a discrete time step invariably leads to small errors that accumulate over millions of steps causing energy drift. Temperature coupling also serves to prevent this energy drift. Commonly used thermostats are the Berendsen thermostat[74] (fastest, but does not produce the correct ensemble), the velocity re-scaling thermostat[75] (which includes a correction term on-top of the Berendsen thermostat), or the Nosé–Hoover thermostat (most accurate, but more computationally expensive).[76,77] Pressure coupling is implemented similarly to temperature coupling; the system is coupled to a barostat that allows fluctuations in the instantaneous pressure but maintains the average pressure of the system over time by scaling the volume of the unit cell. Commonly used barostats include the Berendsen barostat, and the Parrinello-Rahman barostat[78] (equivalent to the Nosé–Hoover thermostat).

Accessible simulations timescales are a major limitation of MD and simulation software packages incorporate many algorithms to increase the efficiency of each calculation.[79,80] Modern computers have multiple processing cores that can perform sets of calculations in parallel. To take advantage of this processing power, domain decomposition can be used to divide a system into smaller cells that each contain a subset of the systems' atoms. Each cell can then be assigned to an individual core that performs the calculations for the atoms in that cell. A separate process is required to control the communication between each cell, in order to calculate interactions that cross cell boundaries. Because of the additional communication required, domain decomposition results in a loss of efficiency, but the same algorithm allows calculations to be spread beyond an individual machine to a cluster of computers, increasing the available computer power. In a similar manner, algorithms can be designed so that specialized hardware like GPUs, which contain thousands of individually less-powerful cores, can be used to rapidly calculate the thousands of interactions in a system.

At the user level, a useful means to speed up a simulation is to increase the length of each time step. The length of each time step is restricted by the most rapid movements in the system, and increasing the time step commonly involves introducing approximations for these rapid movements.[81] To allow the use of a longer time step, bonds can be fixed to the average bond length to eliminate rapid bond vibrations (e.g. using the SHAKE,[82] SETTLE,[83] or LINCS[84] algorithms), or mass can be transferred to each hydrogen from its bonded heavy atom to slow its velocity while preserving its momentum.[81] Virtual sites can be used to avoid simulating hydrogen atoms entirely, instead calculating their position from the simulated heavy atoms. United-atom force fields can be used to represent groups of bonded atoms, such as methyl groups, as a single particle, thereby removing the hydrogen atoms and reducing the total number of atoms required. The united-atom philosophy can be extended into coarse-grain methods which use single particles to approximate larger groups of atoms, such as protein side chains (e.g. the MARTINI force field[85]).

Conventional MD simulations are generally poor at exploring conformational space and, somewhat realistically, continually re-sample low-energy areas rather than crossing the energy barriers that are necessary to observe a process of interest. There are many enhanced sampling methods that exist to improve the ability of MD to explore conformational space, and each provides a different balance of information, bias, and computational efficiency. Replica-exchange MD,[86] accelerated MD,[87] and metadynamics[88] are all useful for exploring binding ensembles. For cases where only the affinity is of interest, there are numerous free energy methods such as free-energy perturbation, thermodynamic integration, or potential of mean force methods, that are commonly used to determine affinity using a two-state approximation.[89] These methods have a higher computational cost than conventional MD and are best used with a simple binding model to determine the free energy of a single bound state. The increased computational cost makes free energy methods poor at exploring the conformational landscape.

1.4 Markov state models

1.4.1 Introduction

Markov state models (MSMs) are a variant of the Markov model, a stochastic model of a system that moves sequentially between states and, in which, the probability of transitioning to a future state depends only on the current state of the system. This history independence is also known as the Markov property. Markov models are named after Andrey Markov who published on the theory in the early 1900s,[90] although mathematical studies of Markov chains exist from as early as the 1600s and the theory was reinvented subsequent to Markov's work. Markov models have been applied in a diverse range of fields, including meteorology,[91] search engine page ranking,[92] and speech recognition.[93]

Markov models are typically applied to systems with clearly defined states, a property that high-dimensional molecular systems often lack. As part of the methodology of constructing a Markov model for a molecular system a state definition needs to be determined, and the resultant model is commonly referred to as a Markov state model. The key idea that enabled the advent of Markov state models was the definition of a conformation, not as a point in continuous space, but as a kinetically related area in space.[94] This idea greatly reduced the dimensionality of the problem, allowing a finite set of states that made the subsequent calculation of transition probabilities feasible.

We are interested in the application of Markov models to biomolecular simulations, specifically as a means of improving the sampling of binding events and timescales accessible by ligand binding simulations. Practically, a Markov state model describes the behavior of a system of states by gathering statistics on the transitions between those states over time.[95,96] The MSM itself is the network of transition probabilities between each pair of states. In MSMs constructed from molecular simulations, these states typically correspond to the metastable and transition states of the system. Prior to being illustrated as a network diagram, the MSM is represented as a 2D-matrix with eigenvectors that describe the slowest processes in the system and associated eigenvalues that relate to the timescale of those processes. The transition probabilities that comprise the MSM describe events observed on relatively short timescales, such as the rotation of a functional group to interact with another pocket or the breaking of an individual hydrogen bond. These probabilities can be extrapolated to longer timescales and estimate, rather than observe, the kinetics of longer timescale events such as the binding or unbinding of a ligand.

Each of the states in a MSM must be history independent (the Markov property), which means that the probability of transitioning from state A to state B depends only on state A itself and not on how the system arrived there. History independence leads to one of the major advantages of MSMs in molecular simulation, namely that there is no need for a single long simulation of the system. Instead, the observed transitions can instead be amassed from large numbers of short simulations (Figure 7), allowing many simulations to be carried out concurrently and taking full advantage of the parallel architecture available in high powered computing clusters.[97]

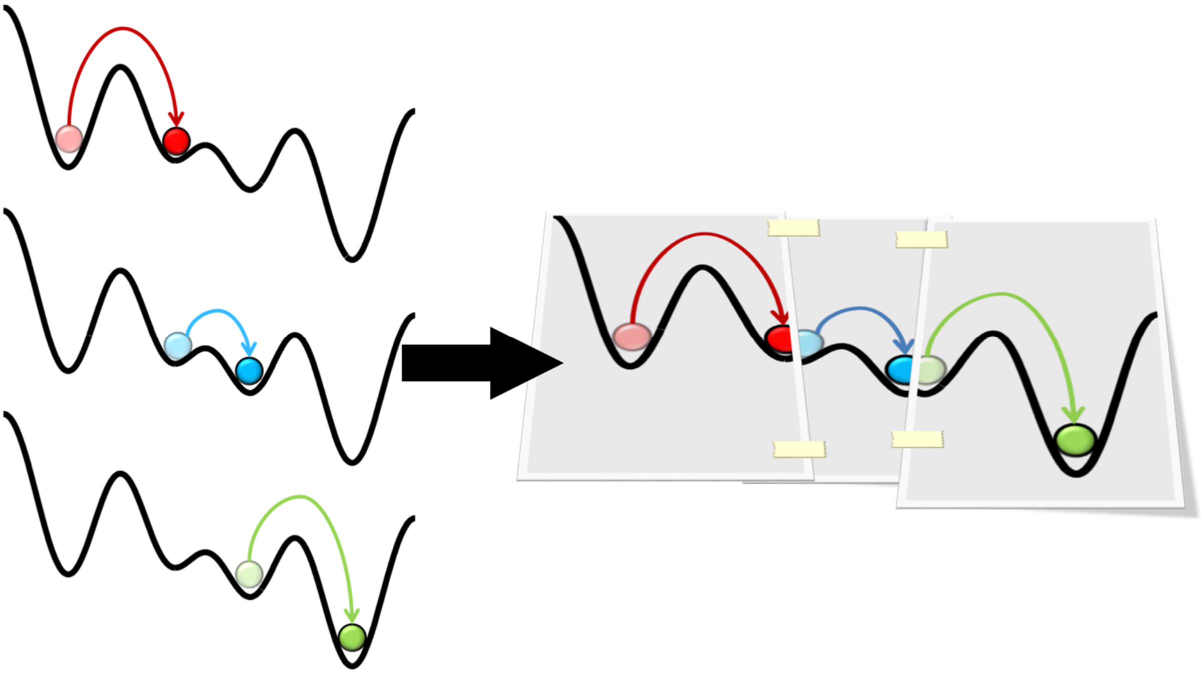


Figure 7. “Snapshots” of individual observations from multiple trajectories (left) can be stitched together into a larger free energy landscape (right).

1.4.2 Applications

The most extensive use of MSMs has been to solve protein folding problems[96] and the application of MSMs to ligand binding or conformational change in functional proteins has largely been built off this work. A key driving force for the development of MSMs has been the Folding@home project[98] and the closely associated MSMbuilder software.[99,100] Folding@home utilizes a distributed computing network consisting of a large number of volunteered personal computers that are individually able to perform only short MD simulations. The nature of this computing system results in large volumes of data in the form of short MD simulations that need to be combined to create a useful dataset, which is possible through the application of MSMs. pyEMMA[101] is another publicly available software package at the forefront of MSM development.

The same methods that can be used to study protein folding can be applied to study the conformational change in functional proteins.[102] Shukla et al. used MSMs and the Folding@Home platform to identify the conformational changes involved in the activation of SRC Kinase and predicted the existence of an allosteric pocket that might allow selective inhibition of kinase species.[103] Malmstrom et al. built on this work in a study of the activation of protein kinase A while cAMP was bound.[104] This work was accompanied by an excellent review on the methodology employed and challenges in studying functional proteins.[105] Bowman et al. constructed a MSM of the conformational ensemble of TEM-1 β-lactamase, identifying (and confirming with labelling experiments) the opening of novel pockets distal to the orthosteric site.[106] Movement of residues in these pockets were shown to correlate with movement in the orthosteric site and it was thus postulated that ligand binding to these pockets would have an allosteric effect.

Studies of GPCRs with MSMs have focused on the activation mechanism of the proteins; a phenomenon that occurs on a millisecond timescale. Kohlhoff et al. conducted a study of the activation mechanism of the β_2AR.[107,108] This study was able to map out multiple activation pathways and metastable states of the β_2AR, and additionally, by performing simulations in the presence of both an agonist and inverse agonist, was able to show that each ligand changed the preference of the activation mechanism for a different pathway. Achieving simulations of this timescale was made possible through use of the Google exacycle distributed computing network. The study also included virtual screening at a number of metastable states along the activation pathway, showing that different chemotypes preferred binding at different states of the GPCR ensemble. A second study of GPCR activation, also of the β_2AR, focused on the opening of the central water channel during deactivation, but with more limited computational resources, this study only examined a single inactivation event that occurred over 1 microsecond.[109]

Many of the MSM studies of functional proteins described above were performed with a ligand bound for the duration of the simulations, but only the effect of the ligand on the protein ensemble was considered and not the ligand itself. Studies of ligand binding using MSMs have thus far been limited and much of the method development of ligand-binding MSMs has been performed on the benzamidine trypsin system.[110–112] Individual studies have investigated a diverse range of targets: Lawrenz et al. investigated the binding of a series of small molecule drugs to the FKBP12 protein,[113] Choudhary et al. constructed a MSM of ATP binding to a voltage dependent anion channel,[114] Silva et al. constructed a MSM of lysine binding to the flexible LAO protein,[115] and Huang et al. investigated the unbinding of several weakly-binding fragments from the FK506 binding protein, and found that the dissociation occurred through multiple pathways.[116] These studies all describe the ligand position using the raw ligand coordinates. In the benzamidine trypsin system, the ligand is approximated by a single xyz coordinate.[110] This simple approximation works because benzamidine is a small, rigid body and easily rotates. In the studies by Lawrenz et al. and Choudhary et al., each determined that the proteins were rigid enough that only the ligand coordinates needed to be considered.[113,114] Huang et al. incorporated protein flexibility into their model, but found the protein to be remarkably stable.[116] A later study of the benzamidine trypsin system expanded the analysis to account for protein flexibility[111] which provides the additional complication of the vastly different timescales between changes in protein conformation and ligand binding events. While this study provides excellent insight into the ligand's effect on protein conformation, the ligand itself was only considered to be bound or unbound. Silva et al.'s study of ligand binding to the LAO protein provides an example of large protein flexibility that can be described with simple descriptors. The clam-like domain movements of the LAO protein were described using a twisting and a closing angle, although the authors had difficulty featurizing the ligand using RMSD due to the large change in the movement of the ligand after it ceased diffusing through the solvent and associated with the protein.[115] A conclusion that can be drawn from the studies described above is that the metrics, such as RMSD, that are generally used to describe protein flexibility or simple ligands are not necessarily suitable for describing the motions of more drug-like ligands, nor are these metrics suitable for dealing with the usually disparate timescales between ligand and protein movements.

1.4.3 General methodology for MSM construction

A general method for building a MSM of a molecular system is detailed in Figure 8, the components of this method are discussed in further detail in the sections below. Construction of the MSM begins with conventional MD simulation of the system. The starting structures for these simulations can be chosen in many ways; for ligand binding, a sensible approach is to start the initial simulations with the ligand in the bulk solvent, this ensures the necessary overlap of states between simulations as the binding pathways are explored. There is also the potential to draw starting structures from enhanced sampling techniques previously performed on the same system.[96] Once simulation data is obtained, the dataset needs to be featurized by reducing the dimensionality of the system down from the set of raw atomic coordinates to a set of features that describe the processes of interest in the system. The dimensionality of the system can be further reduced with time-structure independent component analysis (tICA), producing a small set of dimensions that describe the most significant kinetic events in the system. Once the dimensionality is reduced, clustering is used to assign a set of states to the system. The transitions are then counted between these states to calculate the transition network. The transition network can indicate states that are poorly sampled, and through an adaptive sampling process, additional simulations can be performed to better sample these states.

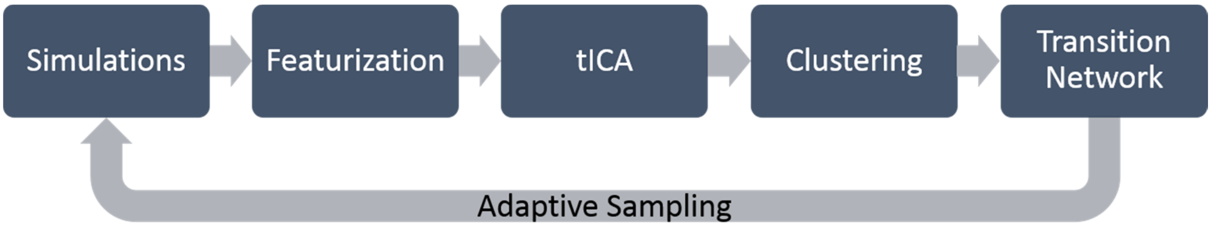


Figure 8. Generalized process of Markov state model construction

1.4.4 Featurization

While in many applications of Markov models the states are known and clearly defined, in molecular simulations the definition of the states is not known *a priori*. The raw coordinates of the MD simulations represent the system in very high-dimensional space and, before suitable states can be defined, it is necessary to choose a featurization of the system, i.e. a set of descriptors that describe kinetically related (rapidly interchanging) regions of conformational space. Once featurized, the system can be clustered to obtain the discrete states used to construct the MSM. The raw set of coordinates for each frame can be used for clustering, but the raw data contains a large proportion of redundant or highly correlated data and kinetically irrelevant noise. Practically, it is important to establish a minimal set of descriptors that describe a significant portion of the kinetic variance over the timescales of interest, for example the conformational changes in a small peptide are well represented by the rotation of a few key dihedral angles.

Without prior knowledge of the pathways followed by a particular molecular system it is not possible to devise complex featurization models, and therefore, recent MSMs[117] make use of the same simple featurization methods used in earlier models.[97,118,119] The conformations of small peptides are often described using the dihedral angles of the backbone and the presence or absence of hydrogen bonds native to the folded state.[118] Folding of larger proteins is featurized primarily based on Cα RMSD from the folded state[97] or distances between residues which make contact in the native structure.[119] With an understanding of the kinetics of a target system it is possible to devise more sophisticated featurizations, although this should be done cautiously. Sorin et al. used the presence of helix or coil secondary structure to featurize a 16-residue peptide[120] but a later study found the helix-coil transitions in this peptide to be too close kinetically to warrant separate clusters.[121]

It is frequently useful to apply a dimensionality reduction to the featurization, as even a simple featurization will have many correlated dimensions. Time-structure independent component analysis (tICA)[122] is a dimensionality reduction method that has proven useful in building MSMs.[123,124] tICA is similar to principal component analysis, which identifies sets of features that are related to the largest geometric movement in the system. tICA instead identifies the sets of correlated features that are responsible for the slowest events in the system and thus have the greatest kinetic relevance to the process of interest. A free energy surface can be projected onto the resultant tICs (Figure 9), which can indicate whether areas have been well sampled over a pair of dimensions and is a good indication of dataset quality. The downside of the tICA analysis is that the tICs are largely abstract dimensions and are difficult to map back to the original features. Nevertheless, tICA has proven highly effective for the construction of MSMs as it can reduce a featurization down to a handful of dimensions and, by eliminating large amounts of noise, increase the quality of the model.[123–125]

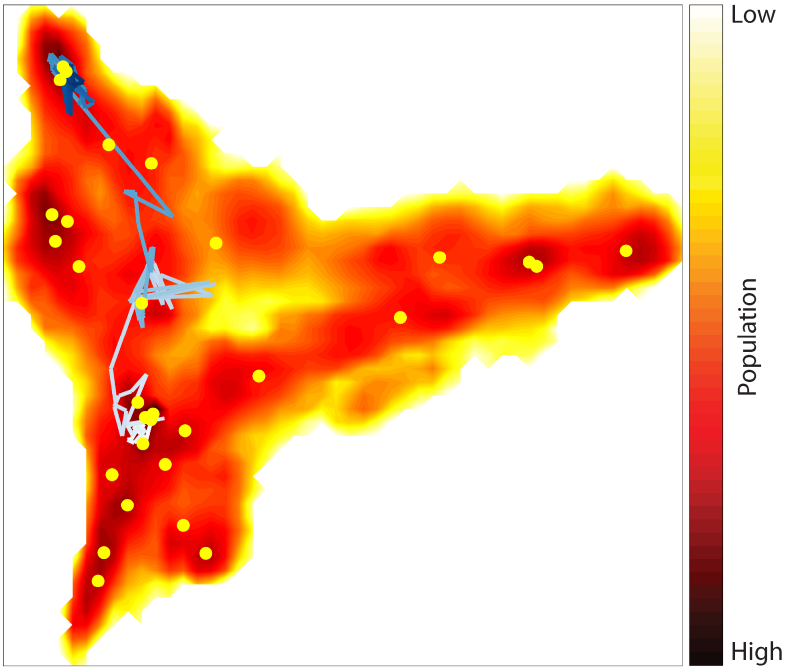


Figure 9. A free energy surface projected onto 2 tICA dimensions. Darker areas indicate a higher population and thus lower free energy. Cluster centers are shown as yellow circles. An individual trajectory is mapped onto the surface as a gradient line, changing from white to blue as the trajectory progresses.

Simple featurizations do not necessarily scale well to larger or more disordered proteins. In the case of intrinsically disordered proteins, an RMSD featurization can be misleading due to large changes regularly occurring between kinetically close states.[126] For large proteins, the increased dimensionality, due to the number of atoms and contacts to consider, leads to an increase in statistical error, requiring more data to produce the same quality model. This scaling problem is largely solved by tICA,[124] which essentially identifies kinetically relevant dimensions from the simple featurization. The use of tICA has quickly come to be considered best practice and has been shown to both reproduce the results of previous MSMs using simple featurizations[125] and to enable the construction of MSMs for intrinsically disordered proteins.[125,126]

1.4.5 Clustering

With a suitable featurization method established, the featurized data is then clustered to discretize the energy landscape into a set of states and to assign every MD trajectory frame to one of those states. There are a large number of clustering algorithms to choose from, each with advantages and disadvantages. The most frequently used clustering algorithms for MSMs are k-means clustering,[127,128] k-centers clustering,[129] and hierarchical clustering.[130] K-means clustering results in more cluster centers in regions of high density, ensuring energy minima are well described, but data in sparse areas is likely to be poorly assigned. The k-centers clustering method evenly distributes cluster centers over the free energy surface, ensuring that all space is covered, but poorly describes areas of high density. Hierarchical clustering is deterministic and maps out a dendrogram, relating every data point through similarity and allowing the number of clusters to be chosen post-calculation but at the cost of the increased computational complexity, which can make it impractical. For this work we have primarily used k-means clustering due to its efficiency when working with large datasets. Ideally, each state produced should correspond to an energy well on the real free energy surface.

Much of the earlier MSM work[118,131] was performed using smaller, manually curated cluster sets. The integration of clustering into the workflow of MSM specific software packages has allowed the automated clustering of systems using any algorithm or metric for which a software implementation is available.[99,101] The majority of studies have utilized k-means clustering, which is shown to generally outperform other commonly used algorithms, and surpassed only by the more computationally expensive hierarchical clustering with Ward's method.[132]

1.4.6 Calculating the transition network

Constructing the transition network for a Markov state model is more involved than simply counting the transitions between each state. Practically, due to featurization and discretization (clustering) error, each state is likely to have internal energy barriers which make the transitions out of the state history dependent, or non-Markovian. In order to determine the transition probability out of a given state, the system must be given time to cross the intra-state energy barriers and reach an equilibrium within that state, and by doing so, the transitions out of the state become history independent. The time that is given to the system to reach an intra-state equilibrium is called the lag time (τ).

To choose an appropriate lag time the timescale of the slowest processes in the system (and therefore the most kinetically significant) are estimated for each of a series of lag times using Equation (2):[95]

$$t_i = \frac{\tau}{\lambda_i} \tag{2}$$

where t_i is the implied timescale of process i, λ_i is the eigenvalue of the transition matrix associated with process i, and τ is the lag time. The estimated timescales of individual processes can then be plotted as a function of the lag time as shown in Figure 10. Once a lag time that is sufficient for the states in each process to reach equilibrium has been reached, the predicted timescales should remain constant for any greater lag times and the curve will be seen to flatten out. The drawback of using a lag time is that any events that occur faster than the chosen lag time will be "blurred" out in the resultant model. Therefore, a decision can be made to sacrifice accuracy for an increased resolution by choosing a lower lag time, although the resultant model may not be Markovian.

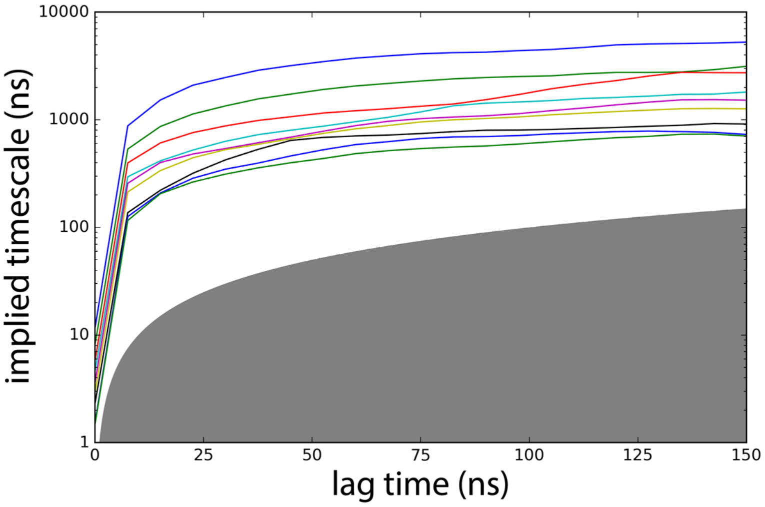


Figure 10. An implied timescale plot showing implied timescales of the 10 slowest processes in a system (y) at a range of lag times (x). Implied timescales are shown on a logarithmic scale, and the shaded area indicates the noise-dominated region where timescales are faster than the lag time.

Once the MSM states and lag time have been established, the probabilities of transitions between individual states can be calculated. Transitions between states are counted using a sliding window approach (Figure 11), giving a set of raw counts for every state-to-state transition. If both the forward and backward pathways have been simulated rigorously, these raw counts can be directly converted into transition probabilities. More commonly, unbinding or unfolding occur on much slower timescales than binding or folding, and the key transitions in these pathways are far less sampled than the forward pathways. Maximum likelihood estimation can be used to better estimate the backwards transitions by assuming an equilibrium between the forward and backward events,[133] thereby producing a set of modified counts that are used in place of the raw counts. The network of transition probabilities produced is the MSM.

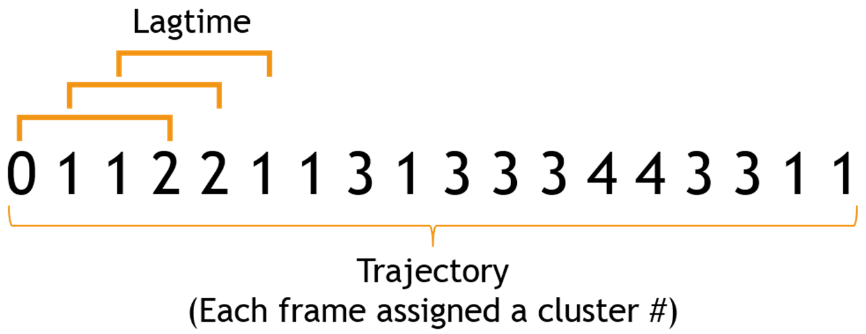


Figure 11. Transitions between states are counted using a sliding window approach; the fixed length of the window is the lag time and a transition is counted between the state at the start and end of the window as it slides along the trajectory 1 frame at a time. E.g. in this figure the first three transitions counted are between state 0-2, 1-2, and 1-1.

The MSM resulting from the above clustering and analysis process often has far more states than are human-readable, and while these are suitable for calculations using transition path theory, it is generally useful to produce a coarser macrostate model. Several methods exist (e.g. PCCA+[134] or BACE[135]) to combine kinetically close states together into a smaller set of macrostates that is much easier to interpret.

1.4.7 Adaptive sampling.

It is uncommon to construct a MSM from a set of trajectories produced in a single iteration. More often, the MSM is iteratively assembled from batches of simulations through a process called adaptive sampling.[112,136] Adaptive sampling involves deriving and analyzing an incomplete transition network after each batch of simulation to determine undersampled areas or other interesting/problematic areas of the network so they can be explored with future batches of simulations. Adaptive sampling provides a significant gain to efficiency by avoiding wasting effort resampling already sufficiently sampled areas.[137] Adaptive sampling is also critical to efforts to automate the Markov state modeling process. When running small batches of simulations manually, the manual setup time can often exceed the simulation time.

With the choice of featurization and clustering algorithms left to the user, effort has been directed towards developing automated methods of MSM construction, and the development of adaptive sampling algorithms to efficiently generate a dataset. One of the early implementations of adaptive sampling focused on improving the precision of the transition probabilities in pre-existing MSMs.[138] Applied to a model system, this adaptive sampling approach produced an order of magnitude decrease in the number of samples required to reach a given precision when compared to a naïve sampling algorithm. A later expansion of these methods, which based adaptive sampling on the variance of the distributions of the eigenvalues and eigenvectors, produced a 3 orders of magnitude gain in precision for the slowest timescale in the system.[139] In order to apply adaptive sampling to the sampling of a new model, it was found more useful to base the additional sampling on connectivity based metrics, by identifying states with poor connectivity to the rest of the network as candidates for further sampling.[137] Doerr et al. were able to automate the entire process of constructing their ligand binding MSM, from simulations through to the final model, by adaptively sampling from the states with the longest mean first passage time (MFPT) from the bulk solvent, reasoning that longer MFPTs likely indicated that states were both closer to the bound state and more in need of additional sampling.[112] They found that by using this adaptive sampling method, their predicted ΔG of binding converged to the experimental value using an order of magnitude fewer simulations.

1.4.8 Hidden Markov models

Hidden Markov models (HMMs) are a cousin of MSMs.[140,141] HMMs have seen widespread application across many fields for their usefulness in pattern recognition.[142,143] In terms of simulating biological systems, HMMs are popular for their ability to greatly decrease discretization error by interweaving the processes of clustering and transition network construction, resulting in far fewer (but less errorful) states.[144,145] As coarser states make it more difficult to isolate undersampled regions, HMMs are less suitable for adaptive sampling and are typically produced from the final dataset, in some cases using a MSM as an initial guess.[146,147]

The methods for generating HMMs are similar to those used for MSMs, up until the clustering phase. Rather than discrete states, HMMs assign states as a Gaussian distribution in the featurization dimensions. Thus every frame has a probability of belonging to multiple states.[148] A transition network is then constructed with every frame assigned to its most probable state and the entire network is assigned a likelihood score. The Gaussians are then adjusted to increase the likelihood of the network by effectively adjusting the state assignments to better match the transition probabilities and a new network is constructed. This process is continued for a specified number of iterations or until termination criteria are met and the model with the highest likelihood is kept. Construction of an HMM is illustrated in Figure 12 in which there are two states within an energy landscape (orange and blue) and each frame (circle) has a probability of belonging to either state, based on its position in the energy landscape. Frame A has a 75% probability of being orange and is observed to transition to frame B, which is blue. Two models are created, model 1 which has frame A assigned to orange, and model 2 with frame A assigned to blue. In model 1, frame A has a 75% chance of being orange and the transition to blue has a 20% chance, giving a likelihood of 0.75 × 0.20 = 0.15. In model 2, frame A has a 25% chance of being blue and the transition to blue has an 80% chance, giving a likelihood of 0.25 × 0.80 = 0.20. Therefore the state assignment in model 2 has a higher likelihood. The likelihoods for all observations are summed to generate a likelihood for the entire model. The state definitions are then revised and a new set of transition probabilities is generated.

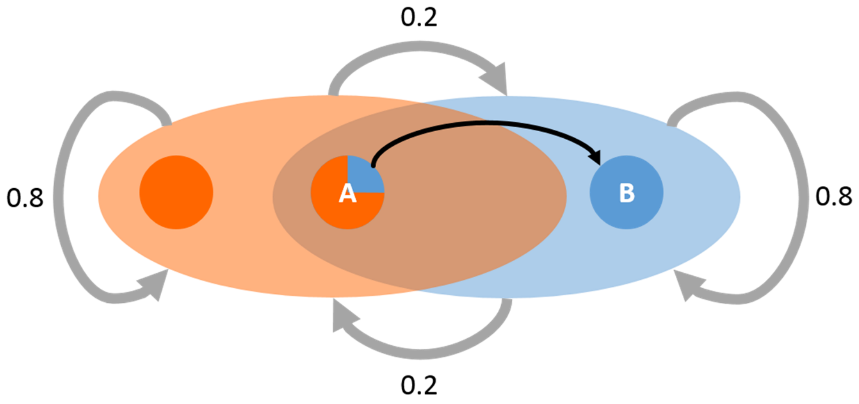


Figure 12. Two states (defined by Gaussians) on a free energy surface are represented by the orange and blue ellipses. Individual frames are represented as circles on this landscape. The possible transitions between states are shown as grey arrows with their associated probabilities, and the observed transition between two frames (A and B) is shown with a black arrow.

1.5 Aims and scope of thesis

It has been known for many years that models used to describe the ligand-binding process simplify the intricacies of ligand binding. The reason we rely on these models is that the commonly used experimental methods for structure-based drug design, namely X-ray crystallography and NMR, convey little information about the dynamics or events involved in ligand binding. In contrast, computational methods allow modeling ligand-binding systems in a time- and structure-resolution that is inaccessible to experimental methods, making computational methods particularly useful for investigating experimentally challenging systems such as G protein-coupled receptors, where the available structural information covers only a small fraction of the known targets, and a smaller still fraction of their dynamic ensembles. GPCRs are targets of immense pharmaceutical interest and are priority targets for new methodologies in both computational and experimental fields. This thesis aims to both expand the available structural and dynamic information for GPCRs, and their ligands, and improve the computational methodologies that can be used to study the behavior of ligands binding to GPCRs and other pharmaceutically important targets.

The overall aim of this work is to investigate the process of ligand binding, and its dynamics, at the atomic level using computational methods.

In Chapter 2 we present a methodology for the homology modeling of GPCRs. To address gaps in the structural knowledge of GPCRs, we develop homology models of 5 muscarinic acetylcholine receptors using the crystal structure of the β_2 adrenergic receptor as a template. In order to optimize the binding site of our homology models and improve them over naïve homology models, we include an extra step into our methodology, using induced-fit docking to include functional knowledge into the model-building process. The ability of these models to select for known actives over decoy compounds is evaluated through virtual screening and this ability is then compared to that of crystal structures of the same receptors, as well as naïve homology models constructed from these crystal structures.

In Chapter 3, we employ MD simulations to investigate the binding of haloperidol and clozapine to the D_2 and D_3 dopamine receptors. The goal of this work is to predict the bound state of these pharmaceutically important ligands and other metastable states that may be of use in drug design. This is the first time the binding pathways of these ligands have been simulated. We improve upon most studies in the literature by running simulations for a timescale long enough to examine the pathways that each ligand follows to the bound state, and compare these pathways between the 2 dopamine receptors.

A primary limitation of computational methods is the timescales of the dynamics they can access. The accessible timescale for a molecular simulation depends on the available hardware and the level of approximation used in the computation. Markov state models present a new methodology that not only makes more efficient use of the available hardware but also serves to extend the accessible timescales further, without adding new approximations or bias to the simulations. While much of the existing MSM methodology was developed to solve protein-folding problems, we devise an improved workflow and implement new system descriptors that allow us to apply this promising new method to investigate ligand-binding systems. During the process of developing the ligand-binding MSM methodology, we switch our investigations to a model system, and in Chapter 4 we employ molecular dynamics simulations to investigate the binding of oleic acid to the liver fatty acid-binding protein. We perform many simulations with an aggregate simulation time that is 2 orders of magnitude longer than previous studies of an FABP system, thereby providing insight into the process of ligand binding to FABPs. By introducing MSMs, we expand and improve on the methods employed in Chapter 3, allowing us to further improve the sampling of molecular system and gather statistics on the ligand-binding pathway.

Chapter 5 returns to our investigations of haloperidol binding to the D_3 dopamine receptor. We develop the Markov state modeling methodology presented in Chapter 4, greatly improving the ability of the method to efficiently and robustly sample ligand-binding systems. We then apply this methodology to the GPCR system and build upon our previous findings by expanding our sampling to the entire binding ensemble and identifying novel binding pathways.

Finally, we present a brief conclusion to this thesis.

1.6 References

(1) Fischer, E. Einfluss Der Configuration Auf Die Wirkung Der Enzyme. *Berichte der Dtsch. Chem. Gesellschaft* **1894**, *27*, 2985–2993.

(2) Koshland, D. E.; Jr. Application of a Theory of Enzyme Specificity to Protein Synthesis. *Proc. Natl. Acad. Sci. U. S. A.* **1958**, *44*, 98–104.

(3) Copeland, R. A; Pompliano, D. L.; Meek, T. D. Drug-Target Residence Time and Its Implications for Lead Optimization. *Nat. Rev. Drug Discov.* **2006**, *5*, 730–739.

(4) Lu, H.; Tonge, P. J. Drug-Target Residence Time: Critical Information for Lead Optimization. *Curr. Opin. Chem. Biol.* **2010**, *14*, 467–474.

(5) Bryngelson, J. D.; Onuchic, J. N.; Socci, N. D.; Wolynes, P. G. Funnels, Pathways, and the Energy Landscape of Protein Folding: A Synthesis. *Proteins Struct. Funct. Genet.* **1995**, *21*, 167–195.

(6) Tsai, C. J.; Tsai, C. J.; Kumar, S.; Kumar, S.; Ma, B.; Ma, B.; Nussinov, R.; Nussinov, R. Folding Funnels, Binding Funnels, and Protein Function. *Protein Sci* **1999**, *8*, 1181–1190.

(7) 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. U. S. A.* **2011**, *108*, 13118–13123.

(8) Wacker, D.; Fenalti, G.; Brown, M. A.; Katritch, V.; Abagyan, R.; Cherezov, V.; Stevens, R. C. Conserved Binding Mode of Human β2 Adrenergic Receptor Inverse Agonists and Antagonist Revealed by X-Ray Crystallography. *J. Am. Chem. Soc.* **2010**, *132*, 11443–11445.

(9) Elling, C. E.; Nielsen, S. M.; Schwartz, T. W. Conversion of Antagonist-Binding Site to Metal-Ion Site in the Tachykinin NK-1 Receptor. *Nature* **1995**, *374*, 74–77.

(10) Baker, J. G.; Hill, S. J. Multiple GPCR Conformations and Signalling Pathways: Implications for Antagonist Affinity Estimates. *Trends Pharmacol. Sci.* **2007**, *28*, 374–381.

(11) Monod, J.; Wyman, J.; Changeux, J. P. On the Nature of Allosteric Transitions: A Plausible Model. *J. Mol. Biol.* **1965**, *12*, 88–118.

(12) Boehr, D. D.; McElheny, D.; Dyson, H. J.; Wright, P. E. The Dynamic Energy Landscape of Dihydrofolate Reductase Catalysis. *Science* **2006**, *313*, 1638–1642.

(13) Changeux, J.-P.; Christopoulos, A. Allosteric Modulation as a Unifying Mechanism for Receptor Function and Regulation. *Cell* **2016**, *166*, 1084–1102.

(14) Gentry, P. R.; Sexton, P. M.; Christopoulos, A. Novel Allosteric Modulators of G Protein-Coupled Receptors. *J. Biol. Chem.* **2015**, *290*, 19478–19488.

(15) Lindsley, C. W.; Emmitte, K. A.; Hopkins, C. R.; Bridges, T. M.; Gregory, K. J.; Niswender, C. M.; Conn, P. J. Practical Strategies and Concepts in GPCR Allosteric Modulator Discovery: Recent Advances with Metabotropic Glutamate Receptors. *Chem. Rev.* **2016**, *116*, 6707–6741.

(16) Redka, D. S.; Pisterzi, L. F.; Wells, J. W. Binding of Orthosteric Ligands to the Allosteric Site of the M2 Muscarinic Cholinergic Receptor. *Mol. Pharmacol.* **2008**, *74*, 834–843.

(17) Kruse, A. C.; Hu, J.; Pan, A. C.; Arlow, D. H.; Rosenbaum, D. M.; Rosemond, E.; Green, H. F.; Liu, T.; Chae, P. S.; Dror, R. O.; Shaw, D. E.; Weis, W. I.; Wess, J.; Kobilka, B. K. Structure and Dynamics of the M3 Muscarinic Acetylcholine Receptor. *Nature* **2012**, *482*, 552–556.

(18) Haga, K.; Kruse, A. C.; Asada, H.; Yurugi-Kobayashi, T.; Shiroishi, M.; Zhang, C.; Weis, W. I.; Okada, T.; Kobilka, B. K.; Haga, T.; Kobayashi, T. Structure of the Human M2 Muscarinic Acetylcholine Receptor Bound to an Antagonist. *Nature* **2012**, *482*, 547–551.

(19) Fronik, P.; Gaiser, B. I.; Sejer Pedersen, D. Bitopic Ligands and Metastable Binding Sites: Opportunities for G Protein-Coupled Receptor (GPCR) Medicinal Chemistry. *J. Med. Chem.* **2017**, *60*, 4126–4134.

(20) Kamal, M.; Jockers, R. Bitopic Ligands: All-in-One Orthosteric and Allosteric. *F1000 Biol Rep* **2009**, *1*, 77.

(21) 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.* **2003**, *63*, 1256–1272.

(22) Katritch, V.; Cherezov, V.; Stevens, R. C. Structure-Function of the G Protein-Coupled Receptor Superfamily. *Annu. Rev. Pharmacol. Toxicol.* **2013**, *53*, 531–556.

(23) Lagerström, M. C.; Schiöth, H. B. Structural Diversity of G Protein-Coupled Receptors and Significance for Drug Discovery. *Nat. Rev. Drug Discov.* **2008**, *7*, 339–357.

(24) Insel, P. A.; Tang, C.-M.; Hahntow, I.; Michel, M. C. Impact of GPCRs in Clinical Medicine: Monogenic Diseases, Genetic Variants and Drug Targets. *Biochim. Biophys. Acta - Biomembr.* **2007**, *1768*, 994–1005.

(25) 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.; Overington, J. P. A Comprehensive Map of Molecular Drug

Targets. *Nat. Rev. Drug Discov. Drug Discov.* **2016**, *16*, 19–34.

(26) Beaulieu, J.-M.; Gainetdinov, R. R. The Physiology, Signaling, and Pharmacology of Dopamine Receptors. *Pharmacol. Rev.* **2011**, *63*, 182–217.

(27) Kruse, A. C.; Kobilka, B. K.; Gautam, D.; Sexton, P. M.; Christopoulos, A.; Wess, J. Muscarinic Acetylcholine Receptors: Novel Opportunities for Drug Development. *Nat. Rev. Drug Discov.* **2014**, *13*, 549–560.

(28) Palczewski, K. Crystal Structure of Rhodopsin: A G Protein-Coupled Receptor. *Science* **2000**, *289*, 739–745.

(29) Cherezov, V.; Rosenbaum, D. M.; Hanson, M. A; Rasmussen, S. G. F.; Thian, F. S.; Kobilka, T. S.; Choi, H.-J.; Kuhn, P.; Weis, W. I.; Kobilka, B. K.; Stevens, R. C. High-Resolution Crystal Structure of an Engineered Human beta2-Adrenergic G Protein-Coupled Receptor. *Science* **2007**, *318*, 1258–1265.

(30) Warne, T.; Moukhametzianov, R.; Baker, J. G.; Nehmé, R.; Edwards, P. C.; Leslie, A. G. W.; Schertler, G. F. X.; Tate, C. G. The Structural Basis for Agonist and Partial Agonist Action on a β(1)-Adrenergic Receptor. *Nature* **2011**, *469*, 241–244.

(31) Rasmussen, S. G. F.; Choi, H.-J.; Fung, J. J.; Pardon, E.; Casarosa, P.; Chae, P. S.; Devree, B. T.; Rosenbaum, D. M.; Thian, F. S.; Kobilka, T. S.; Schnapp, A.; Konetzki, I.; Sunahara, R. K.; Gellman, S. H.; Pautsch, A.; Steyaert, J.; Weis, W. I.; Kobilka, B. K. Structure of a Nanobody-Stabilized Active State of the β(2) Adrenoceptor. *Nature* **2011**, *469*, 175–180.

(32) 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.; Mathiesen, J. M.; Shah, S. T. A.; Lyons, J. A.; Caffrey, M.; Gellman, S. H.; Steyaert, J.; Skiniotis, G.; Weis, W. I.; Sunahara, R. K.; Kobilka, B. K. Crystal Structure of the β2 Adrenergic receptor–Gs Protein Complex. *Nature* **2011**, *477*, 549–555.

(33) Chien, E. Y. T.; Liu, W.; Zhao, Q.; Katritch, V.; Han, G. W.; Michael, A.; Shi, L.; Newman, A. H.; Javitch, J. A; Cherezov, V.; Stevens, R. C. Structure of the Human Dopamine D3 Receptor in Complex with a D2/D3 Selective Antagonist. *Science* **2011**, *330*, 1091–1095.

(34) de Ligt, R. A.; Kourounakis, A. P.; IJzerman, A. P. Inverse Agonism at G Protein-Coupled Receptors: (Patho)physiological Relevance and Implications for Drug Discovery. *Br. J. Pharmacol.* **2000**, *130*, 1–12.

(35) Rankovic, Z.; Brust, T. F.; Bohn, L. M. Biased Agonism: An Emerging Paradigm in GPCR Drug Discovery. *Bioorganic Med. Chem. Lett.* **2016**, *26*, 241–250.

(36) Vauquelin, G.; Van Liefde, I. G Protein-Coupled Receptors: A Count of 1001 Conformations. *Fundam. Clin. Pharmacol.* **2005**, *19*, 45–56.

(37) Tarsy, D.; Baldessarini, R. J. Epidemiology of Tardive Dyskinesia: Is Risk Declining with Modern Antipsychotics? *Mov. Disord.* **2006**, *21*, 589–598.

(38) Leucht, S.; Corves, C.; Arbter, D.; Engel, R. R.; Li, C.; Davis, J. M. Second-Generation versus First-Generation Antipsychotic Drugs for Schizophrenia: A Meta-Analysis. *Lancet* **2009**, *373*, 31–41.

(39) Iqbal, M. M.; Rahman, A.; Husain, Z.; Mahmud, S. Z.; Ryan, W.; Feldman, J. Clozapine: A Clinical Review of Adverse Effects and Management. *Ann. Clin. Psychiatry* **2003**, *15*, 33–48.

(40) López-Muñoz, F.; Alamo, C. The Consolidation of Neuroleptic Therapy: Janssen, the Discovery of Haloperidol and Its Introduction into Clinical Practice. *Brain Res. Bull.* **2009**, *79*, 130–141.

(41) Crilly, J. The History of Clozapine and Its Emergence in the US Market: A Review and Analysis. *Hist. Psychiatry* **2007**, *18*, 39–60.

(42) *WHO Model List of Essential Medicines*, 19th ed.; World Health Organization, 2015.

(43) Peprah, K.; Zhu, X. Y.; Eyunni, S. V. K.; Setola, V.; Roth, B. L.; Ablordeppey, S. Y. Multi-Receptor Drug Design: Haloperidol as a Scaffold for the Design and Synthesis of Atypical Antipsychotic Agents. *Bioorganic Med. Chem.* **2012**, *20*, 1291–1297.

(44) Li, P.; Snyder, G. L.; Vanover, K. E. Dopamine Targeting Drugs for the Treatment of Schizophrenia: Past, Present and Future. *Curr. Top. Med. Chem.* **2016**, *16*, 3385–3403.

(45) Michino, M.; Boateng, C. A.; Donthamsetti, P.; Yano, H.; Bakare, O. M.; Bonifazi, A.; Ellenberger, M. P.; Keck, T. M.; Kumar, V.; Zhu, C.; Verma, R.; Deschamps, J. R.; Javitch, J. A.; Newman, A. H.; Shi, L. Toward Understanding the Structural Basis of Partial Agonism at the Dopamine D3 Receptor. *J. Med. Chem.* **2017**, *60*, 580–593.

(46) Furuhashi, M.; Hotamisligil, G. S. Fatty Acid-Binding Proteins: Role in Metabolic Diseases and Potential as Drug Targets. *Nat. Rev. Drug Discov.* **2008**, *7*, 489–503.

(47) Wang, Y.-T.; Liu, C.-H.; Zhu, H.-L. Fatty Acid Binding Protein (FABP) Inhibitors: A Patent Review (2012-2015). *Expert Opin. Ther. Pat.* **2016**, *26*, 767–776.

(48) McMahon, B. A.; Murray, P. T. Urinary Liver Fatty Acid-Binding Protein: Another Novel Biomarker of Acute Kidney Injury. *Kidney Int.* **2010**, *77*, 657–659.

(49) Gururajan, P.; Gurumurthy, P.; Nayar, P.; Srinivasa Nageswara Rao, G.; Babu, S.; Cherian, K. M.

Heart Fatty Acid Binding Protein (H-FABP) as a Diagnostic Biomarker in Patients with Acute Coronary Syndrome. *Hear. Lung Circ.* **2010**, *19*, 660–664.

(50) Furuhashi, M.; Saitoh, S.; Shimamoto, K.; Miura, T. Fatty Acid-Binding Protein 4 (FABP4): Pathophysiological Insights and Potent Clinical Biomarker of Metabolic and Cardiovascular Diseases. *Clin. Med. Insights. Cardiol.* **2014**, *8*, 23–33.

(51) Jenkins, A. E.; Hockenberry, J. A.; Nguyen, T.; Bernlohr, D. A. Testing of the Portal Hypothesis: Analysis of a V32G, F57G, K58G Mutant of the Fatty Acid Binding Protein of the Murine Adipocyte. *Biochemistry* **2002**, *41*, 2022–2027.

(52) Kaczocha, M.; Vivieca, S.; Sun, J.; Glaser, S. T.; Deutsch, D. G. Fatty Acid-Binding Proteins Transport N-Acylethanolamines to Nuclear Receptors and Are Targets of Endocannabinoid Transport Inhibitors. *J. Biol. Chem.* **2012**, *287*, 3415–3424.

(53) Smathers, R. L.; Petersen, D. R. The Human Fatty Acid-Binding Protein Family: Evolutionary Divergences and Functions. *Hum. Genomics* **2011**, *5*, 170–191.

(54) Storch, J.; McDermott, L. Structural and Functional Analysis of Fatty Acid-Binding Proteins. *J. Lipid Res.* **2009**, *50*, S126–S131.

(55) Sacchettini, J. C.; Gordon, J. I.; Banaszak, L. J. Crystal Structure of Rat Intestinal Fatty-Acid-Binding Protein. Refinement and Analysis of the Escherichia Coli-Derived Protein with Bound Palmitate. *J. Mol. Biol.* **1989**, *208*, 327–339.

(56) Cai, J.; Lücke, C.; Chen, Z.; Qiao, Y.; Klimtchuk, E.; Hamilton, J. A. Solution Structure and Backbone Dynamics of Human Liver Fatty Acid Binding Protein: Fatty Acid Binding Revisited. *Biophys. J.* **2012**, *102*, 2585–2594.

(57) Sousa, S. F.; Ribeiro, A. J. M.; Coimbra, J. T. S.; Neves, R. P. P.; Martins, S. A.; Moorthy, N. S. H. N.; Fernandes, P. A.; Ramos, M. J. Protein-Ligand Docking in the New Millennium – A Retrospective of 10 Years in the Field. *Curr. Med. Chem.* **2013**, *20*, 2296–2314.

(58) Lionta, E.; Spyrou, G.; Vassilatis, D. K.; Cournia, Z. Structure-Based Virtual Screening for Drug Discovery: Principles, Applications and Recent Advances. *Curr. Top. Med. Chem.* **2014**, *14*, 1923–1938.

(59) McRobb, F. M.; Capuano, B.; Crosby, I. T.; Chalmers, D. K.; Yuriev, E. Homology Modeling and Docking Evaluation of Aminergic G Protein-Coupled Receptors. *J. Chem. Inf. Model.* **2010**, *50*, 626–637.

(60) N. Cavasotto, C. Homology Models in Docking and High-Throughput Docking. *Curr. Top. Med. Chem.* **2011**, *11*, 1528–1534.

(61) Beuming, T.; Lenselink, B.; Pala, D.; McRobb, F.; Repasky, M.; Sherman, W. Docking and Virtual Screening Strategies for GPCR Drug Discovery. In; Humana Press, New York, NY, 2015; pp 251–276.

(62) Lu, M.; Wu, B. Structural Studies of G Protein-Coupled Receptors. *IUBMB Life* **2016**, *68*, 894–903.

(63) Lavecchia, A.; Giovanni, C. Virtual Screening Strategies in Drug Discovery: A Critical Review. *Curr. Med. Chem.* **2013**, *20*, 2839–2860.

(64) Karplus, M.; McCammon, J. A. Molecular Dynamics Simulations of Biomolecules. *Nat. Struct. Biol.* **2002**, *9*, 646–652.

(65) Shaw, D. E.; Dror, R. O.; Salmon, J. K.; Grossman, J. P.; Mackenzie, K. M.; Bank, J. A.; Young, C.; Deneroff, M. M.; Batson, B.; Bowers, K. J.; Chow, E.; Eastwood, M. P.; Ierardi, D. J.; Klepeis, J. L.; Kuskin, J. S.; Larson, R. H.; Lindorff-Larsen, K.; Maragakis, P.; Moraes, M. A.; Piana, S.; Shan, Y.; Towles, B. *Millisecond-Scale Molecular Dynamics Simulations on Anton*. In *SC '09 Proceedings of the Conference on High Performance Computing Networking, Storage and Analysis*; New York, 2009; pp 1–11.

(66) González, M. A. Force Fields and Molecular Dynamics Simulations. *Collect. SFN* **2011**, *12*, 169–200.

(67) MacKerell, A D.; Bashford, D.; Bellott, M.; Dunbrack, R. L.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; Joseph-McCarthy, D.; Kuchnir, L.; Kuczera, K.; Lau, F. T. K.; Mattos, C.; Michnick, S.; Ngo, T.; Nguyen, D. T.; Prodhom, B.; Reiher, W. E.; Roux, B.; Schlenkrich, M.; Smith, J. C.; Stote, R.; Straub, J.; Watanabe, M.; Wiorkiewicz-Kuczera, J.; Yin, D.; Karplus, M. All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins. *J. Phys. Chem. B* **1998**, *102*, 3586–3616.

(68) Vanommeslaeghe, K.; Hatcher, E.; Acharya, C.; Kundu, S.; Zhong, S.; Shim, J.; Darian, E.; Guvench, O.; Lopes, P.; Vorobyov, I.; Mackerell, A. D. CHARMM General Force Field: A Force Field for Drug-like Molecules Compatible with the CHARMM All-Atom Additive Biological Force Fields. *J. Comput. Chem.* **2010**, *31*, 671–690.

(69) Schmid, N.; Eichenberger, A. P.; Choutko, A.; Riniker, S.; Winger, M.; Mark, A. E.; Van Gunsteren, W. F. Definition and Testing of the GROMOS Force-Field Versions 54A7 and 54B7. *Eur. Biophys.*

J. **2011**, *40*, 843–856.

(70) Leach, A. R. *Molecular Modelling : Principles and Applications*, Prentice Hall: 2001.

(71) Monticelli, L.; Tieleman, D. P. Force Fields for Classical Molecular Dynamics. In *Biomolecular Simulations: Methods and Protocols*; Monticelli, L., Salonen, E., Eds.; Humana Press, Totowa, NJ, 2013; Vol. 924, pp 197–213.

(72) Guvench, O.; MacKerell, A. D. Comparison of Protein Force Fields for Molecular Dynamics Simulations. In *Molecular Modeling of Proteins*; Humana Press, 2008; Vol. 443, pp 63–88.

(73) Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G. A Smooth Particle Mesh Ewald Method. *J. Chem. Phys.* **1995**, *103*, 8577–8593.

(74) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; DiNola, A.; Haak, J. R. Molecular Dynamics with Coupling to an External Bath. *J. Chem. Phys.* **1984**, *81*, 3684–3690.

(75) Bussi, G.; Donadio, D.; Parrinello, M. Canonical Sampling through Velocity Rescaling. *J. Chem. Phys.* **2007**, *126*, 014101.

(76) Nosé, S.; Klein, M. L. Constant Pressure Molecular Dynamics for Molecular Systems. *Mol. Phys.* **1983**, *50*, 1055–1076.

(77) Hoover, W. G. Canonical Dynamics: Equilibrium Phase-Space Distributions. *Phys. Rev. A* **1985**, *31*, 1695–1697.

(78) Parrinello, M.; Rahman, A. Polymorphic Transitions in Single Crystals: A New Molecular Dynamics Method. *J. Appl. Phys.* **1981**, *52*, 7182–7190.

(79) Abraham, M. J.; Murtola, T.; Schulz, R.; Páll, S.; Smith, J. C.; Hess, B.; Lindahl, E. GROMACS: High Performance Molecular Simulations through Multi-Level Parallelism from Laptops to Supercomputers. *SoftwareX* **2015**, *1*, 19–25.

(80) 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.

(81) Feenstra, K. A.; Hess, B.; Berendsen, H. J. C. Improving Efficiency of Large Time-Scale Molecular Dynamics Simulations of Hydrogen-Rich Systems. *J. Comput. Chem.* **1999**, *20*, 786—798.

(82) Ryckaert, J.-P.; Ciccotti, G.; Berendsen, H. J. . Numerical Integration of the Cartesian Equations of Motion of a System with Constraints: Molecular Dynamics of N-Alkanes. *J. Comput. Phys.*

1977, *23*, 327–341.

(83) Miyamoto, S.; Kollman, P. A. Settle: An Analytical Version of the SHAKE and RATTLE Algorithm for Rigid Water Models. *J. Comput. Chem.* **1992**, *13*, 952–962.

(84) Hess, B.; Bekker, H.; Berendsen, H. J. C.; Fraaije, J. G. E. M. LINCS: A Linear Constraint Solver for Molecular Simulations. *J. Comput. Chem.* **1997**, *18*, 1463–1472.

(85) Marrink, S. J.; Risselada, H. J.; Yefimov, S.; Tieleman, D. P.; De Vries, A. H. The MARTINI Force Field: Coarse Grained Model for Biomolecular Simulations. *J. Phys. Chem. B* **2007**, *111*, 7812–7824.

(86) Sugita, Y.; Okamoto, Y. Replica-Exchange Molecular Dynamics Method for Protein Folding. *Chem. Phys. Lett.* **1999**, *314*, 141–151.

(87) Hamelberg, D.; Mongan, J.; McCammon, J. A. Accelerated Molecular Dynamics: A Promising and Efficient Simulation Method for Biomolecules. *J. Chem. Phys.* **2004**, *120*, 11919–11929.

(88) Laio, A.; Parrinello, M. Escaping Free-Energy Minima. *Proc. Natl. Acad. Sci. U. S. A.* **2002**, *99*, 12562–12566.

(89) Shirts, M. R.; Mobley, D. L.; Brown, S. P. Free-Energy Calculations in Structure-Based Drug Design. In *Drug Design*; Merz, K. M., Ringe, D., Reynolds, C. H., Eds.; Cambridge University Press: Cambridge, 2010; pp 61–86.

(90) Seneta, E. Markov and the Birth of Chain Dependence Theory. *Int. Stat. Rev. / Rev. Int. Stat.* **1996**, *64*, 255–263.

(91) Gabriel, K. R.; Neumann, J. A Markov Chain Model for Daily Rainfall Occurrence at Tel Aviv. *Q. J. R. Meteorol. Soc.* **1962**, *88*, 90–95.

(92) Langville, A. N.; Meyer, C. D. Updating Markov Chains with an Eye on Google's PageRank. *SIAM J. Matrix Anal. Appl.* **2006**, *27*, 968–987.

(93) Rabiner, L. R. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *Proc. IEEE* **1989**, *77*, 257–286.

(94) Schütte, C.; Fischer, A.; Huisinga, W.; Deuflhard, P. A Direct Approach to Conformational Dynamics Based on Hybrid Monte Carlo. *J. Comput. Phys.* **1999**, *151*, 146–168.

(95) Bowman, G. R.; Pande, V. S.; Noé, F. *An Introduction to Markov State Models and Their Application to Long Timescale Molecular Simulation*; Bowman, G. R., Pande, V. S., Noé, F., Eds.;

Advances in Experimental Medicine and Biology; Springer Netherlands: Dordrecht, 2014; Vol. 797.

(96) Pande, V. S.; Beauchamp, K.; Bowman, G. R. Everything You Wanted to Know about Markov State Models but Were Afraid to Ask. *Methods* **2010**, *52*, 99–105.

(97) Jayachandran, G.; Vishal, V.; Pande, V. S. Using Massively Parallel Simulation and Markovian Models to Study Protein Folding: Examining the Dynamics of the Villin Headpiece. *J. Chem. Phys.* **2006**, *124*, 164902.

(98) Shirts, M.; Pande, V. S. COMPUTING: Screen Savers of the World Unite! *Science* **2000**, *290*, 1903–1904.

(99) Beauchamp, K. A.; Bowman, G. R.; Lane, T. J.; Maibaum, L.; Haque, I. S.; Pande, V. S. MSMBuilder2: Modeling Conformational Dynamics on the Picosecond to Millisecond Scale. *J. Chem. Theory Comput.* **2011**, *7*, 3412–3419.

(100) Harrigan, M. P.; Sultan, M. M.; Hernández, C. X.; Husic, B. E.; Eastman, P.; Schwantes, C. R.; Beauchamp, K. A.; McGibbon, R. T.; Pande, V. S. MSMBuilder: Statistical Models for Biomolecular Dynamics. *Biophys. J.* **2017**, *112*, 10–15.

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

(102) Shukla, D.; Hernández, C. X.; Weber, J. K.; Pande, V. S. Markov State Models Provide Insights into Dynamic Modulation of Protein Function. *Acc. Chem. Res.* **2015**, *48*, 414–422.

(103) Shukla, D.; Meng, Y.; Roux, B.; Pande, V. S. Activation Pathway of Src Kinase Reveals Intermediate States as Targets for Drug Design. *Nat. Commun.* **2014**, *5*, 3397.

(104) Malmstrom, R. D.; Lee, C. T.; Van Wart, A. T.; Amaro, R. E. Application of Molecular-Dynamics Based Markov State Models to Functional Proteins. *J. Chem. Theory Comput.* **2014**, *10*, 2648–2657.

(105) Malmstrom, R. D.; Kornev, A. P.; Taylor, S. S.; Amaro, R. E. Allostery through the Computational Microscope: cAMP Activation of a Canonical Signalling Domain. *Nat. Commun.* **2015**, *6*, 7588.

(106) Bowman, G. R.; Bolin, E. R.; Hart, K. M.; Maguire, B. C.; Marqusee, S. Discovery of Multiple Hidden Allosteric Sites by Combining Markov State Models and Experiments. *Proc. Natl. Acad. Sci. U. S. A.* **2015**, *112*, 2734–2739.

(107) 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.

(108) Shukla, D.; Lawrenz, M.; Pande, V. S. Elucidating Ligand-Modulated Conformational Landscape of GPCRs Using Cloud-Computing Approaches. *Methods Enzymol.* **2015**, *557*, 551–572.

(109) Bai, Q.; Pérez-Sánchez, H.; Zhang, Y.; Shao, Y.; Shi, D.; Liu, H.; Yao, X. Ligand Induced Change of β2 Adrenergic Receptor from Active to Inactive Conformation and Its Implication for the Closed/open State of the Water Channel: Insight from Molecular Dynamics Simulation, Free Energy Calculation and Markov State Model Analysis. *Phys. Chem. Chem. Phys.* **2014**, *16*, 15874–15885.

(110) Buch, I.; Giorgino, T.; De Fabritiis, G. Complete Reconstruction of an Enzyme-Inhibitor Binding Process by Molecular Dynamics Simulations. *Proc. Natl. Acad. Sci. U. S. A.* **2011**, *108*, 10184–10189.

(111) Plattner, N.; Noé, F. Protein Conformational Plasticity and Complex Ligand-Binding Kinetics Explored by Atomistic Simulations and Markov Models. *Nat. Commun.* **2015**, *6*, 7653.

(112) Doerr, S.; De Fabritiis, G. On-the-Fly Learning and Sampling of Ligand Binding by High-Throughput Molecular Simulations. *J. Chem. Theory Comput.* **2014**, *10*, 2064–2069.

(113) Lawrenz, M.; Shukla, D.; Pande, V. S. Cloud Computing Approaches for Prediction of Ligand Binding Poses and Pathways. *Sci. Rep.* **2015**, *5*, 7918.

(114) Choudhary, O. P.; Paz, A.; Adelman, J. L.; Colletier, J.-P.; Abramson, J.; Grabe, M. Structure-Guided Simulations Illuminate the Mechanism of ATP Transport through VDAC1. *Nat. Struct. Mol. Biol.* **2014**, *21*, 626–632.

(115) Silva, D.-A. A.; Bowman, G. R.; Sosa-Peinado, A.; Huang, X. A Role for Both Conformational Selection and Induced Fit in Ligand Binding by the Lao Protein. *PLoS Comput. Biol.* **2011**, *7*, e1002054.

(116) Huang, D.; Caflisch, A. The Free Energy Landscape of Small Molecule Unbinding. *PLoS Comput. Biol.* **2011**, *7*.

(117) Husic, B. E.; McGibbon, R. T.; Sultan, M. M.; Pande, V. S. Optimized Parameter Selection Reveals Trends in Markov State Models for Protein Folding. *J. Chem. Phys.* **2016**, *145*, 194103.

(118) Swope, W. C.; Pitera, J. W.; Suits, F.; Pitman, M.; Eleftheriou, M.; Fitch, B. G.; Germain, R. S.;

Rayshubski, A.; Ward, T. J. C.; Zhestkov, Y.; Zhou, R. Describing Protein Folding Kinetics by Molecular Dynamics Simulations. 2. Example Applications to Alanine Dipeptide and a β-Hairpin Peptide. *J. Phys. Chem. B* **2004**, *108*, 6582–6594.

(119) Singhal, N.; Snow, C. D.; Pande, V. S. Using Path Sampling to Build Better Markovian State Models: Predicting the Folding Rate and Mechanism of a Tryptophan Zipper Beta Hairpin. *J. Chem. Phys.* **2004**, *121*, 415–425.

(120) Sorin, E. J.; Pande, V. S. Exploring the Helix-Coil Transition via All-Atom Equilibrium Ensemble Simulations. *Biophys. J.* **2005**, *88*, 2472–2493.

(121) Chodera, J. D.; Singhal, N.; Pande, V. S.; Dill, K. A.; Swope, W. C. Automatic Discovery of Metastable States for the Construction of Markov Models of Macromolecular Conformational Dynamics. *J. Chem. Phys.* **2007**, *126*, 155101.

(122) Molgedey, L.; Schuster, H. G. Separation of a Mixture of Independent Signals Using Time Delayed Correlations. *Phys. Rev. Lett.* **1994**, *72*, 3634–3637.

(123) Naritomi, Y.; Fuchigami, S. Slow Dynamics in Protein Fluctuations Revealed by Time-Structure Based Independent Component Analysis: The Case of Domain Motions. *J. Chem. Phys.* **2011**, *134*, 065101.

(124) Schwantes, C. R.; Pande, V. S. Improvements in Markov State Model Construction Reveal Many Non-Native Interactions in the Folding of NTL9. *J. Chem. Theory Comput.* **2013**, *9*, 2000–2009.

(125) Pérez-Hernández, G.; Paul, F.; Giorgino, T.; De Fabritiis, G.; Noé, F.; P?rez-Hern?ndez, G.; Paul, F.; Giorgino, T.; De Fabritiis, G.; No?, F. Identification of Slow Molecular Order Parameters for Markov Model Construction. **2013**, *139*, 015102.

(126) Schor, M.; Mey, A. S. J. S.; MacPhee, C. E. Analytical Methods for Structural Ensembles and Dynamics of Intrinsically Disordered Proteins. *Biophys. Rev.* **2016**, *8*, 429–439.

(127) Steinhaus, H. Sur La Division Des Corps Materiels En Parties. *Bull. Polish Acad. Sci.* **1956**, *4*, 801–804.

(128) Macqueen, J. *Some Methods for Classification and Analysis of Multivariate Observations*. In *In Procedings of the Fifth Berkeley Symposium on Math, Statistics, and Probability*; 1967; Vol. 1, pp 281–297.

(129) Gonzalez, T. F. Clustering to Minimize the Maximum Intercluster Distance. *Theor. Comput. Sci.* **1985**, *38*, 293–306.

(130) Sneath, P. H. A.; Sokal, R. R. *Numerical Taxonomy : The Principles and Practice of Numerical Classification*; W.H. Freeman, 1973.

(131) Elmer, S. P.; Park, S.; Pande, V. S. Foldamer Dynamics Expressed via Markov State Models. II. State Space Decomposition. *J. Chem. Phys.* **2005**, *123*, 114903.

(132) Husic, B. E.; Pande, V. S. Ward Clustering Improves Cross-Validated Markov State Models of Protein Folding. *J. Chem. Theory Comput.* **2017**, *13*, 963–967.

(133) Prinz, J.-H.; Wu, H.; Sarich, M.; Keller, B.; Senne, M.; Held, M.; Chodera, J. D.; Sch?tte, C.; No?, F.; Schütte, C.; Noé, F. Markov Models of Molecular Kinetics: Generation and Validation. *J. Chem. Phys.* **2011**, *134*, 174105.

(134) Röblitz, S.; Weber, M. Fuzzy Spectral Clustering by PCCA+: Application to Markov State Models and Data Classification. *Adv. Data Anal. Classif.* **2013**, *7*, 147–179.

(135) Bowman, G. R. Improved Coarse-Graining of Markov State Models via Explicit Consideration of Statistical Uncertainty. *J. Chem. Phys.* **2012**, *137*, 134111.

(136) Thompson, S. K. Adaptive Sampling Designs. In; John Wiley & Sons, Inc., 2012; pp 313–318.

(137) Weber, J. K.; Pande, V. S. Characterization and Rapid Sampling of Protein Folding Markov State Model Topologies. *J. Chem. Theory Comput.* **2011**, *7*, 3405–3411.

(138) Singhal, N.; Pande, V. S. Error Analysis and Efficient Sampling in Markovian State Models for Molecular Dynamics. *J. Chem. Phys.* **2005**, *123*, 204909.

(139) Hinrichs, N. S.; Pande, V. S. Calculation of the Distribution of Eigenvalues and Eigenvectors in Markovian State Models for Molecular Dynamics (11 Pages). *J. Chem. Phys.* **2007**, *126*, 244101.

(140) Eddy, S. R. What Is a Hidden Markov Model? *Nat. Biotechnol.* **2004**, *22*, 1315–1316.

(141) Rabiner, L.; Juang, B. An Introduction to Hidden Markov Models. *IEEE ASSP Mag.* **1986**, *3*, 4–16.

(142) Gales, M.; Young, S. The Application of Hidden Markov Models in Speech Recognition. *Signal Processing* **2007**, *1*, 195–304.

(143) Eddy, S. Profile Hidden Markov Models. *Bioinformatics* **1998**, *14*, 755–763.

(144) McGibbon, R. T.; Ramsundar, B.; Sultan, M. M.; Kiss, G.; Pande, V. S. *Understanding Protein Dynamics with L1-Regularized Reversible Hidden Markov Models*. In *Proceedings of the 31st International Conference on Machine Learning*; 2014; Vol. 32, pp 1197–1205.

(145) Noé, F.; Wu, H.; Prinz, J.-H.; Plattner, N. Projected and Hidden Markov Models for Calculating Kinetics and Metastable States of Complex Molecules. *J. Chem. Phys.* **2013**, *139*, 184114.

(146) Pan, X.; Schwartz, S. D. Conformational Heterogeneity in the Michaelis Complex of Lactate Dehydrogenase: An Analysis of Vibrational Spectroscopy Using Markov and Hidden Markov Models. *J. Phys. Chem. B* **2016**, *120*, 6612–6620.

(147) Pinamonti, G.; Zhao, J.; Condon, D. E.; Paul, F.; Noè, F.; Turner, D. H.; Bussi, G. Predicting the Kinetics of RNA Oligonucleotides Using Markov State Models. *J. Chem. Theory Comput.* **2017**, *13*, 926–934.

(148) Shukla, S.; Shamsi, Z.; Moffett, A. S.; Selvam, B.; Shukla, D. Application of Hidden Markov Models in Biomolecular Simulations. In; Humana Press, New York, NY, 2017; pp 29–41.

Chapter 2

Homology Modeling of Human Muscarinic Acetylcholine Receptors

G protein-coupled receptors are targets of immense pharmaceutical interest, but often GPCR-targeting drugs exhibit undesirable side-effect profiles due to a lack of selectivity between, and within, receptor subfamilies. Although all GPCRs have a similar topology, small changes in the binding site can lead to large changes in selectivity between ligands. Obtaining structures of GPCRs is still a challenging process and experimental structures have only been determined for a limited number of receptor subtypes. Homology modeling is an effective means of developing structures for other subtypes where these structures could be used to predict the bound pose of ligands, identify hits during prospective virtual screening, or as starting structures for molecular dynamics simulations. When developing GPCR homology models, one needs to be careful to choose an appropriate template structure and validate that the binding site is receptive to the appropriate ligands.

At the time the work on this chapter began, there were no available crystal structures of muscarinic acetylcholine receptors, so we set out to create a set of homology models of the 5 human muscarinic acetylcholine receptors (M_1R – M_5R) that could be used for drug design. We refined the homology models by training the binding sites with induced-fit docking of known active ligands. As work was nearing completion, crystal structures became available for the M_2R (human) and M_3R (rat) variants, providing closer initial template structures than we had used for our models. We expanded the work to demonstrate that, not only were our models suitable for prospective virtual screening, but the training we had incorporated into our models made them superior to the newly released crystal structures or a naïve homology model constructed from the closer templates.

This chapter is the published article:

Thomas, T.; McLean, K. C.; McRobb, F. M.; Manallack, D. T.; Chalmers, D. K.; Yuriev, E. Homology Modeling of Human Muscarinic Acetylcholine Receptors. *J. Chem. Inf. Model.* **2014**, *54*, 243–253.

Journal of
CHEMICAL INFORMATION
AND MODELING

Article

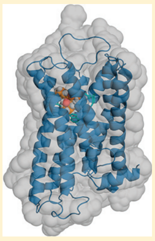
pubs.acs.org/jcim

Homology Modeling of Human Muscarinic Acetylcholine Receptors

Trayder Thomas,[†] Kimberley C. McLean,[†] Fiona M. McRobb,[§] David T. Manallack, David K. Chalmers,* and Elizabeth Yuriev*

Medicinal Chemistry, Monash Institute of Pharmaceutical Sciences, Monash University (Parkville Campus), 381 Royal Parade, Parkville, VIC 3052 Australia

S *Supporting Information*



ABSTRACT: We have developed homology models of the acetylcholine muscarinic receptors M_1R-M_5R, based on the β_2-adrenergic receptor crystal as the template. This is the first report of homology modeling of all five subtypes of acetylcholine muscarinic receptors with binding sites optimized for ligand binding. The models were evaluated for their ability to discriminate between muscarinic antagonists and decoy compounds using virtual screening using enrichment factors, area under the ROC curve (AUC), and an early enrichment measure, LogAUC. The models produce rational binding modes of docked ligands as well as good enrichment capacity when tested against property-matched decoy libraries, which demonstrates their unbiased predictive ability. To test the relative effects of homology model template selection and the binding site optimization procedure, we generated and evaluated a naïve M_2R model, using the M_3R crystal structure as a template. Our results confirm previous findings that binding site optimization using ligand(s) active at a particular receptor, i.e. including functional knowledge into the model building process, has a more pronounced effect on model quality than target–template sequence similarity. The optimized M_1R-M_5R homology models are made available as part of the Supporting Information to allow researchers to use these structures, compare them to their own results, and thus advance the development of better modeling approaches.

INTRODUCTION

The use of structure-based design methods for G protein-coupled receptors (GPCRs) is an active area of research.[1−4] It commenced in the early 2000s after the landmark report of the crystal structure of bovine rhodopsin[5] and accelerated after 2007, when the first crystal structures of ligand-infusible GPCR complexes were solved.[6−8] Technological advances have greatly improved the success of GPCR crystallization[7,9,10] and, at the time of writing, over 30 crystal structures of GPCRs have been solved.[11] However, GPCR crystallization is still an area of highly specialized expertise with most structures coming from a limited number of research groups. As a result, the number of available structures is still very small given the ~800 GPCRs present in the human genome, including 342 nonolfactory receptors.[12] Many GPCR families are still not covered by the currently available high resolution structural information, and it is accepted that, at present, solving structures for all members of the GPCR superfamily is not a realistic goal.[1,4] Consequently, in the absence of experimental structural data, researchers who wish to use structure-based methods to target GPCRs turn to homology models for docking and virtual screening (VS).[13] In several of these drug discovery campaigns, GPCR homology models have proven useful for discovering agents for a range of GPCR targets (Table 1).

While generally an established technique, generation of GPCR homology models for virtual screening can be a speculative exercise, relying on many assumptions and suppositions. Therefore, careful consideration of several related aspects is required when such an exercise is undertaken. (i) Robustness of the computational protocol. This aspect comprises quality of both homology modeling and docking algorithms and should always be evaluated against relevant targets for which experimental data is available: structural data for validating homology modeling and activity data for validating VS. The ultimate question that must be answered is whether the combination of the evaluative model and the protocol used can distinguish between known actives and drug-like decoy molecules. (ii) Quality and appropriateness of the input structural data; specifically the choice of template. Choosing a template for GPCR homology modeling has been previously evaluated;[14−17] however, with the ever-increasing number of available templates, this question cannot be resolved once and for all and requires regular re-evaluation. (iii) Predictive quality of the generated homology models. To address this final issue, homology models should be evaluated in a virtual screening scenario with a particular focus on decoy selection. Because of the importance of these issues, there is currently a considerable interest in evaluating homology modeling and VS protocols as applied to GPCRs.[18−20]

In this study, we have addressed all of the above issues by modeling the five subtypes of muscarinic acetylcholine receptors (mAChRs) and evaluating them using virtual screening. The mAChRs receptors (M_1R-M_5R) can be subdivided into two functional classes based on their G protein coupling preference.[21] The M_1R, M_3R, and M_5R selectively couple to G proteins of the G_q/G_{11} family while the M_2R and M_4R preferentially activate G_i/G_o-type G proteins. Activation of mAChRs leads to a wide range of biochemical and physiological effects, primarily depending on the mAChR location and

Received: September 2, 2013
Published: December 15, 2013

Table 1. Prospective Virtual Screening Campaigns against GPCR Homology Models

target[a]	template[a]	homology modeling program	docking/ screening program	screening library	hit rate[b]	affinity (number of compounds)[c]	ref
D_3R	β_2AR, β_1AR	MODELLER	DOCK3.6	prefiltered ZINC[28] (3000K+)	23% (20%)	K_i = 0.2 − 3.1 μM (6) optimized K_i = 81 nM (1)	29
CXCR4	rhodopsin, β_2AR, β_1AR, $A_{2A}R$	MODELLER	DOCK3.6	lead-like subset of ZINC (3300K)	4% (17%)	IC_{50} = 107 μM (1)	30
CXCR7	rhodopsin, β_2AR, β_1AR, $A_{2A}R$, CXCR4	MOE	CONSENSUS-DOCK	3 proprietary collections (187K, 402K, 196K)	3.3%	IC_{50} = 1.29−11.4 μM (21)	31
S1PR1[d]	rhodopsin	GPCRgen	Snooker	diverse subset of MSD/ Organon library (50K)	NR	pK_i = 4.3−4.7 (3)	32, 33
$A_{2A}R$	β_1AR	MODELLER, MOE	Glide	CAP, BioFocus SoftFocus (545K)	9%	pK_i = 7.5−9.0 (6) (13 to >100-fold selective vs A_1R)	34
5-HT_7R[d]	rhodopsin	MODELLER	Glide	Enamine Screening Collection (730K)	NR	K_i = 0.197 and 0.265 μM (2)	35, 36
5-HT_2R	β_2AR	MODELLER	DOCK3.5, MM-GBSA	FDA drug library, filtered by MW (1430)	NR	K_i = 1.959 mM (1)	37
MCH-1R	β_2AR	MOE	GOLD	commercial vendor catalogues (45K)	14%	IC_{50} = 131 and 213 nM (2 most potent out of 10 novel chemotypes)	38
CB_2R	β_2AR	CHARMM for “activation”	GOLD	filtered subset of ZINC (273K)	12%	K_i = 2.3 nM−71.43 μM (13)	39

[a] Receptor abbreviations: adenosine A_X receptor, A_XR; β_X-adrenergic receptor, β_XAR; cannabinoid receptor 2, CB_2R; C−X−C chemokine receptor 4, CXCR4; dopamine D_3 receptor, D_3R; melanin-concentrating hormone-1 receptor, MCH-1R; serotonin 5-HT_X receptor, 5-HT_XR; sphingosine 1-phosphate receptor, S1PR1. [b] Hit rates are estimated differently in various studies. Where available, we quote hit rates for VS against crystal structures for comparison, in parentheses. NR = not reported. [c] Reported as per original papers. [d] A combination of ligand-based and structure-based approaches were used in this campaign.

subtype. The M_1R, M_4R, and M_5R subtypes are mainly expressed in the central nervous system (CNS); whereas, the M_2R and M_3R subtypes are widely distributed both in the CNS and in peripheral tissues. Specifically, we have generated homology models of mAChRs M_1−M_5, using the β_2-adrenergic receptor (β_2AR) crystal structure (PDB ID: 2RH1)[6] as the template and have optimized their orthosteric binding sites using the induced fit docking (IFD) procedure.[22] (i) We have demonstrated the robustness of our homology modeling/VS protocol using the β_2AR crystal structure in complex with the inverse agonist carazolol and β_2AR antagonist and inverse agonist activity data. We have further verified the protocol by a validation against the β_2AR crystal structure in complex with alprenolol.[23] (ii) To assess the predictive quality of the M_1R−M_5R models, we have carried out virtual screening investigations of all five homology models. The models have been tested against property-matched decoy libraries to demonstrate their unbiased predictive capacity. (iii) Furthermore, after the M_2R (human)[24] and M_3R (rat)[25] crystal structures became available, a naïve (i.e., nonoptimized) M_2R model was generated using the M_3R crystal structure as a template. Evaluating VS performance allowed comparison between the models: naïve but based on a close-sequence template and optimized but based on a more remote-sequence template. Our results support previous findings that binding site optimization using ligand(s) active at a particular receptor, i.e. including functional knowledge into the model building process,[26] has a pronounced effect on model quality for virtual screening. It is clear from our results that carefully designed and knowledge-based homology structures, built with templates with greater than 35% overall similarity in the trans-membrane region,[1] are at least as useful in VS as crystal structures. Finally, similar to our previous work,[27] we have released the coordinates of the five optimized muscarinic receptor structures in the spirit of open science research.

■ EXPERIMENTAL SECTION

Software. Molecular modeling was performed with the Schrödinger software suite.[40,41] Homology models of the five muscarinic M_1−M_5 acetylcholine receptors were built in Prime[42] (v 3.0 and 3.1) from a multiple sequence alignment generated in ClustalW[43] using the Maestro interface (v 9.2 and 9.3). Ligand molecules were prepared using LigPrep[44] (v 2.5), and the binding site was optimized using the IFD protocol[22] following the previously developed procedure.[27] Ligands were docked into the homology models using Glide[45,46] (v 5.7 and 5.8). Default settings were used, unless otherwise stated. Physical descriptors evaluated for comparison of the decoy sets with the active compounds included molecular weight (MW), number of rotatable bonds, number of hydrogen bond donor and acceptor atoms, and calculated logP (ClogP). These physical properties, along with polar surface area (PSA) and vdW volume, were computed using the ChemAxon Marvin Calculator (cxcalc) (http://www.chemaxon.com). The 2D Tanimoto score (calculated using fragment sizes of 1−7 atoms, ignoring hydrogens) was measured to demonstrate the diversity of the structures within the ligand sets.[47] The workflow followed in this study is shown in Figure 1 and described in detail in the following sections.

Homology Modeling. The sequences of the human dopamine, serotonin, α- and β-adrenergic, adenosine, histamine, muscarinic, and bovine rhodopsin receptors were obtained from the Universal Protein Resource (http://www.uniprot.org/) and aligned using ClustalW. The multiple sequence alignment generated was manually edited to remove gaps in helices and to anchor highly conserved residues in each transmembrane (TM) helix. Naïve homology models for the five human mAChRs were built in Prime v 3.0 from the multiple sequence alignment, using the β_2-adrenergic receptor (PDB ID: 2RH1) crystal structure[6] as the template. The human muscarinic M_2 acetylcholine receptor was also built in Prime v 3.1, using the rat muscarinic M_3 acetylcholine receptor (PDB

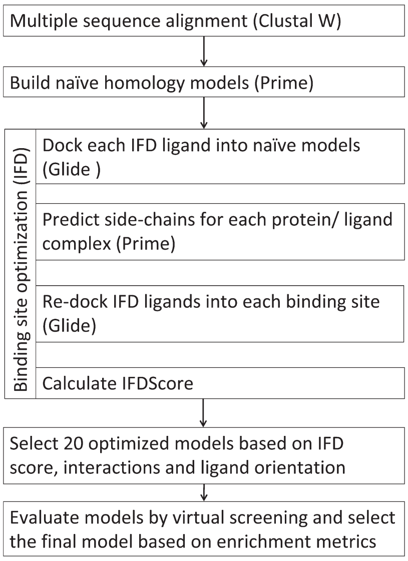

Multiple sequence alignment (Clustal W)
Build naïve homology models (Prime)
Binding site optimization (IFD)
Dock each IFD ligand into naïve models (Glide)
Predict side-chains for each protein/ ligand complex (Prime)
Re-dock IFD ligands into each binding site (Glide)
Calculate IFDScore
Select 20 optimized models based on IFD score, interactions and ligand orientation
Evaluate models by virtual screening and select the final model based on enrichment metrics

Figure 1. Flowchart of homology modeling and model evaluation.

ID: 4DAJ) crystal structure[25] as the template. Further details are described in our previous work.[27]

Binding Site Optimization. The β_2AR crystal structure (PDB ID: 2RH1) and the mAChRs homology models were treated by the Protein Preparation Wizard workflow,[44] prior to docking. Hydrogen atoms were added and minimized using the OPLS_2005 force field. The side chain conformations of the residues within the ligand binding site were refined by docking an appropriate antagonist or inverse agonist into each of the built muscarinic receptor homology models and the β_2AR crystal structure using the IFD protocol. The docking site was centered upon the residues Asp 3.32, Trp 6.48, Phe 6.52, and Tyr 7.43 (Ballesteros–Weinstein numbering[48]) and was defined by a box of dimensions 28 × Å 28 Å × 28 Å. Up to 50 poses per ligand were collected in the initial Glide docking step, with both the van der Waals (vdW) radii and the partial atomic charges scaled to 0.5 in order to collect a more extensive range of poses.

Prime was used to optimize residues within 5 Å of ligand atoms, excluding Asp 3.32 and Trp 6.48, which play a critical role in correctly orienting ligand molecules. Trp 6.48 is a key residue of the aromatic cluster of TM5 and TM6, believed to act as a "micro-switch", important for receptor activation and inactivation.[49] The IFD protocol was found to consistently cause Trp 6.48 to undergo a conformational "flip" during the Prime step, forcing the bulky indole side chain down and away from the binding pocket. For M_1R–M_5R models, when Trp 6.48 and Asp 3.32 were omitted from binding site optimization, more credible ligand poses were obtained, which led to better enrichment. Pala et al.[16] report a similar observation for VS-evaluated homology models of the MT_2 melatonin receptor, namely that residues known to form critical ligand contacts tended to adopt a conformation not favorable to forming such contacts. They have taken this observation as another reason for "calibrating" models (e.g., by VS evaluation) to determine the domain of their applicability.

Following optimization with Prime, the ligand was redocked into the optimized receptor conformations with Glide, using default vdW and charge scaling parameters. Multiple ligand–receptor poses were generated for each model. Successful poses were chosen on the basis of the position and orientation of the ligand within the binding pocket, key hydrogen bonding and vdW interactions, and the relative energy of interaction (a composite of the protein and ligand energy scores: IFDScore = GlideScore + 0.05 × PrimeEnergy). A maximum of 20 poses were collected. During the IFD optimization of the binding sites, we monitored the distance (ndist) between the ionizable or quaternary nitrogen of the ligand (for simplicity we will just refer to this atom as the "ionizable nitrogen") and the closest carboxylate oxygen of the conserved Asp 3.32 residue. This residue has been determined by site-directed mutagenesis to be crucial in the ligand-binding mode of all aminergic GPCRs.[50] The term ndist is a quantitative measure of this important ionic interaction, and receptors with ndist > 3.0 Å were excluded from further analysis.

Virtual Screening Libraries. Active compounds known to act at the β_2-adrenergic and muscarinic receptors were used to enrich the decoy compound databases (20 actives for β_2AR and 48 actives for mAChRs; see Table S1 for the lists of actives and ref 27 for chemical structures). The active compounds were downloaded from the GLIDA database[51] (http://pharminfo.pharm.kyoto-u.ac.jp/services/glida/). Protonation states and formal charges at physiological pH (pH 7.4 ± 2.0) for each active ligand and decoy compound were assigned in LigPrep. One structure per compound was selected for screening.

Three sets of decoy compounds were used in this study. Set 1, containing 1000 drug-like decoy compounds, was obtained from Schrödinger (http://www.schrodinger.com). This set had been randomly selected from a library of one million compounds having properties characteristic of drug molecules.[45,46] We have analyzed the properties of the decoy ligands and active compounds: molecular weight (g/mol), number of rotatable bonds, polar surface area (Å^2), calculated logP, number of hydrogen bond donors and acceptors, solvent accessible volume (Å^3), and 2D Tanimoto score. Generally, the properties of the active compounds were found to be similar to those of the decoy library (Table 2). The molecular weights varied from 151 to 645 g/mol, with an average of 360 g/mol. These decoys were not specifically chosen to mimic muscarinic antagonist compounds, as we first wanted to ascertain whether our models were capable of identifying active ligands from within a broad representation of drug-like compound space.

Decoy set 2 was derived from the ZINC database[28] (7 233 297 compounds, database version 7) by a process of successive

Table 2. Average Ligand Properties

ligand set	MW (g/mol)	rotatable bonds	PSA (Å^2)	ClogP	H-bond donor	H-bond acceptor	vdW volume (Å^3)	2D Tanimoto score
M_1R actives	324	5.1	31	3.03	1.4	1.6	318	0.233
				decoy sets				
1: Schrödinger	360	5.0	84	2.90	2.0	4.2	316	0.125
2: ZINC	320	4.3	38	3.43	1.4	1.7	302	0.185
3: refined Schrödinger	343	4.8	79	2.59	2.4	3.3	312	0.143

eliminations, creating a subset of molecules that closely adhered to the physical properties of the actives (Table 2). Specifically, decoys were required to fall within a similar normal distribution as the active compounds (265−434 g/mol; mean 322 g/mol; standard deviation 40 g/mol). Decoys were also required to contain an ionizable nitrogen and not to contain more than three hydrogen bond donors or four hydrogen bond acceptors. Finally, each decoy was required to have a Tanimoto score of less than 0.8 with respect to all other molecules within the set to ensure topological diversity. 1000 molecules were randomly selected from a larger subset satisfying the applied criteria, so that direct comparisons could be made between the screening results using the Schrödinger and ZINC libraries, in terms of enrichment factors and early hits. A carefully selected set of 1000 molecules seems to be sufficient to detect enrichment trends. Huang et al. found that there was little size-dependent behavior detected when screening with their entire Directory of Useful Decoys (DUD)[52] of 98 266 molecules compared to a randomly selected subset of 1000 molecules.

Decoy set 3 (refined Schrödinger) was a subset of the decoy set 1, with molecular weight limited to be consistent with that of the active compounds (260−410 g/mol). All compounds from the Schrödinger decoy library with a molecular weight which fell outside the range of the active compounds were removed. Furthermore, all decoy compounds which did not contain an ionizable nitrogen were similarly removed to create a more challenging decoy set of 261 compounds.

Enrichment Studies. Molecular docking studies were performed using Glide, which flexibly docks ligands into a rigid receptor model. The docking site was centered upon the coordinates of carazolol (the inverse agonist present in the β_2AR crystal structure) and limited to accommodate ligands up to 18 Å in length. The midpoint of each ligand was bound to an inner box of 10 Å^3. Postdocking minimization retained a single pose per ligand. Both the Standard Precision (SP) and the Extra Precision (XP) scoring functions were evaluated, and XP gave marginally better results, which are presented here. Poses were ranked using GlideScore. Following docking, models were visually inspected to ensure that the ligands were well oriented within the defined binding pocket and to ensure that important expected interactions, based on mutagenesis studies,[53] were found between ligand and receptor molecules. Enrichment factors (EF) were calculated at 2, 5, and 10% of the total number of compounds (N_{total}) screened, according to EF$^{x\%}$ = $(\text{Hits}_{\text{sampled}}/N_{\text{sampled}}) \div (\text{Hits}_{\text{total}}/N_{\text{total}})$.

■ RESULTS

Method Evaluation: β_2 Adrenergic Receptor Ligand Docking. Our modeling protocol encompasses generating multiple IFD complex structures and selecting final receptor models. To evaluate the protocol, we used the β_2AR as a test case. Thirty-three structures of the β_2AR/carazolol complex were generated using IFD. To test the ability of our modeling and VS evaluation workflow to preferentially retrieve known actives, we docked 20 known β_2AR antagonists (see Table S1 in the Supporting Information) and the library of Schrödinger decoys into all 33 receptor models. Higher enrichment factors and area under the enrichment curve (AUC) values and lower average distance between the ligand ionizable nitrogen and Asp 3.32 (ndist) correlated with greater model efficiency in selecting active molecules early in the screen. The properties of the top 5 highest ranked models are shown in Table 3, and a complete list is provided in Supporting Information Table S2. A detailed comparison between β_2AR/carazolol IFD complexes and carazalol- or alprenolol-bound crystal structures is presented in the Supporting Information. This test case shows that our protocol can retrieve correct binding modes for β_2AR/carazolol complexes (i.e., consistent with crystal structures).

Table 3. Five Top Ranked Models from Virtual Screening of the β_2AR Structures Generated by IFD Using Carazolol

	enrichment factor						
ranking	2%	5%	10%	AUC	mean ndist (Å)	Carazolol RMSD (Å)	Alprenolol RMSD (Å)
1	21.8	13.7	8.9	0.96	2.09	0.75	0.65
2	21.8	13.7	9.4	0.93	2.30	1.50	1.95
3	14.5	12.7	8.9	0.95	2.51	1.21	1.62
4	12.1	11.7	9.4	0.95	2.36	4.37	1.95
5	12.1	12.7	8.9	0.96	2.79	1.63	1.09

Homology Modeling of Muscarinic Receptors. *Binding Site Optimization by IFD.* Clozapine and atropine were chosen as the optimizing ligands for IFD since they have high affinity for the M$_1$−M$_5$ receptors; reported clozapine K_i values vary from 1.4−5.0 nM and atropine K_i values range between 0.2 and 1.5 nM.[54] Following the VS procedure, described below, we found that the atropine-optimized model for the M$_1$R gave the best enrichment, while the best M$_2$R−M$_5$R models were optimized using clozapine.

Model Quality Evaluation by VS. We evaluated the ability of the receptor models to prioritize active compounds over decoy molecules. The decoy libraries, enriched with the respective active compounds (see Table S1 in the Supporting Information), were docked into the receptor models. The IFD ligands, used for binding site optimization, were excluded from virtual screening to remove any potential structural bias. While enrichment plots and enrichment factors (EFs) are still routinely used for evaluating VS performance (e.g., ref 18), they are not ideal and do not account for several aspects of virtual screening. ROC curves are superior to enrichment plots in that they not only reflect the selection of actives, but also the nonselection of decoys.[55,56] The metric afforded by a ROC curve is the area under the receiver operating characteristic curve (ROC AUC), which gives an indication of the total number of compounds successfully docked into the model and is interpreted as the probability that a randomly chosen active has a higher score than a randomly chosen inactive. Several metrics, such as NSQ_AUC[57] and LogAUC,[58] have also been developed to focus on early, rather than overall, enrichment.

ROC curves for the M$_1$R−M$_5$R models are shown in Figure 2. Enrichment plots and semilogarithmic ROC curves are provided in the Supporting Information (Figures S2 and S3). We also report the ROC AUC and LogAUC metrics, as enrichment measures (Table 4). The LogAUC preferentially weighs early enrichment by computing the percentage of the ideal area under the semilog ROC curve.The results reveal excellent enrichment capacity for M$_2$R, M$_4$R, and M$_5$R models with the latter having particularly good early enrichment. Although, the M$_1$R and, particularly, the M$_3$R gave lower enrichments using all three decoy sets, their enrichment metrics are comparable to and sometimes better than those obtained in recent reports. For example, homology models of the MT$_2$ melatonin receptor,[16] based on the β_2AR and optimized for antagonists gave EF$_{2\%}$ = 3.1−18.7 and of antagonists against

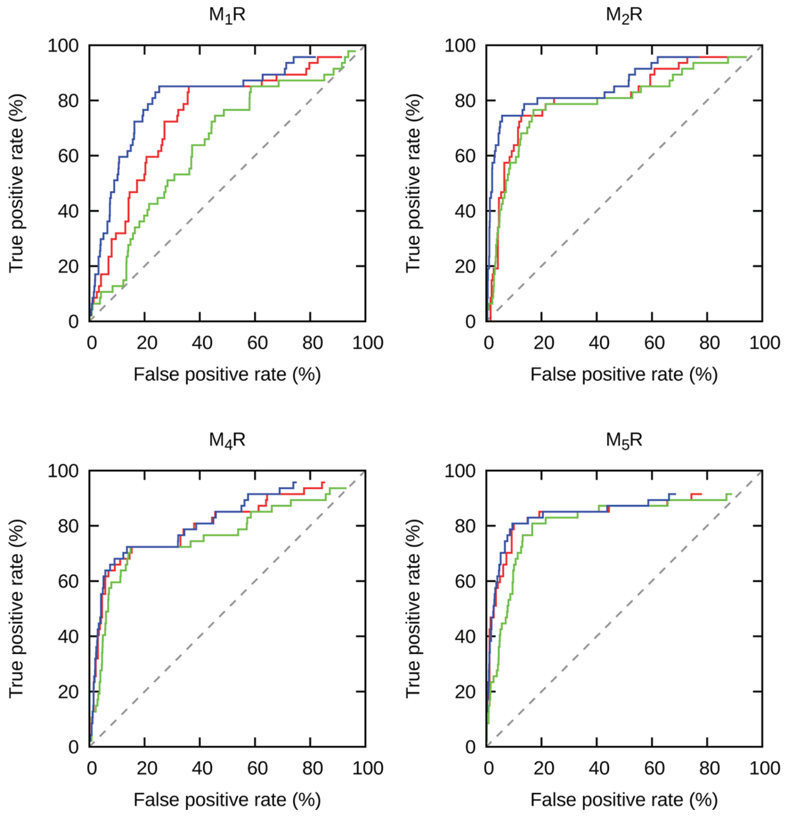


Figure 2. ROC curves for M_1R−M_5R models: (blue) set 1, Schrödinger; (green) set 2, ZINC; (red) set 3, refined Schrödinger. The dotted line indicates random choice (no enrichment).

Table 4. Virtual Screening Evaluation of Muscarinic Receptors

				EF (at *X* % of ranked database)		
receptor	ROC AUC	$\text{LogAUC}_{0.001}$	mean ndist (Å)	2	5	10
		set 1 (Schrödinger decoy set)				
M_1R	0.81	0.35	3.89	5.3	5.5	4.7
M_2R	0.86	0.50	3.72	11.7	11.4	7.4
M_3R	0.74	0.38	4.11	8.5	7.6	4.9
M_4R	0.82	0.41	5.07	7.4	8.4	6.4
M_5R	0.85	0.53	4.00	12.7	10.1	7.4
		set 2 (ZINC decoy set)				
M_1R	0.64	0.22	3.92	3.2	2.1	1.3
M_2R	0.79	0.36	3.57	3.2	5.9	5.3
M_3R	0.62	0.26	4.81	3.2	5.5	4.2
M_4R	0.76	0.35	5.10	6.4	5.5	5.7
M_5R	0.81	0.40	4.02	8.5	5.5	5.3
		set 3 (refined Schrödinger decoy set)				
M_1R	0.74	0.28	3.89	2.8	2.5	2.3
M_2R	0.81	0.36	3.72	2.8	3.7	4.0
M_3R	0.69	0.30	4.11	3.7	2.9	3.0
M_4R	0.81	0.40	5.07	4.7	4.5	4.4
M_5R	0.84	0.51	4.00	5.6	5.3	5.1

multiple β_2AR crystal structures gave $EF_{2\%}$ = 0.3−11.7 and $EF_{10\%}$ = 1.5−3.9.[59] While our results compare favorably with the cited work, such comparisons should not be overinterpreted given the studies used different actives, decoy sets, and receptor types.

The main deficiencies of the models are the failure to dock some of the actives, shown as a gap at the end of the ROC curves, and in the inability of the M_3R model to identify a substantial fraction of actives, shown by the M_3R plots dropping down to the "random" line at approximately 60% of the false positive rate when using the ZINC decoy set. The properties of actives that either did not dock or produced docked poses with a scoring energy greater than the set acceptable cutoff are reported in the Supporting Information (Table S3). This data suggests that the most likely reason for docking failure is the large size of these compounds; thus a better M_3R model might be developed by using an alternative bulkier IFD ligand.

As a simple evaluation of the binding geometries, we calculated the distance between the ionizable nitrogen of the actives and Asp 3.32 (ndist). Mean values for each set are reported in Table 4. In the majority of cases, ndist fell within the range exhibited by ligands in 22 GPCR crystal structures (2.52 Å (PDB ID: 2Y01)−4.02 Å (PDB ID: 4DAJ); mean 2.92 Å). This salt bridge and other key receptor−ligand hydrogen bonding and ionic interactions were observed among many of the top-ranked poses of active compounds. This confirmed that not only were the models capable of producing high enrichment, they were also generating the expected contacts. Figure 3 illustrates binding modes of three active ligands, demonstrating interactions with binding site residues. It could be therefore suggested that a requirement for ndist to be less than 4 Å may serve as a useful pharmacophore filter in prospective virtual screening against aminergic GPCRs. However, it should be noted that recent work by Lin et al.

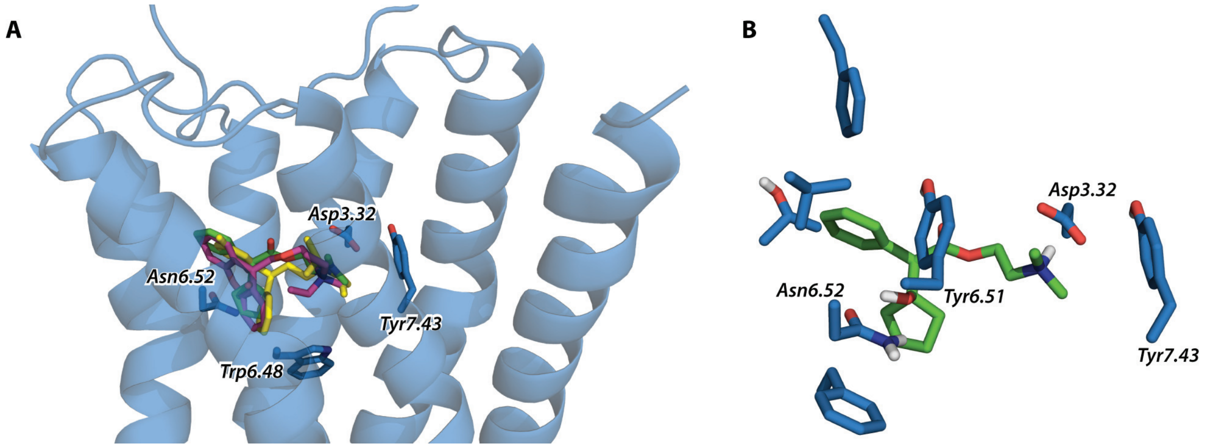


Figure 3. Cartoon representation of the M_2R model, showing the docked poses of the three highest ranked actives (A) and a close-up of cyclopentolate surrounded by its interacting residues (B). Color coding: cyclopentolate (green), tolterodine (yellow), and methantheline (pink). Binding site residues are blue.

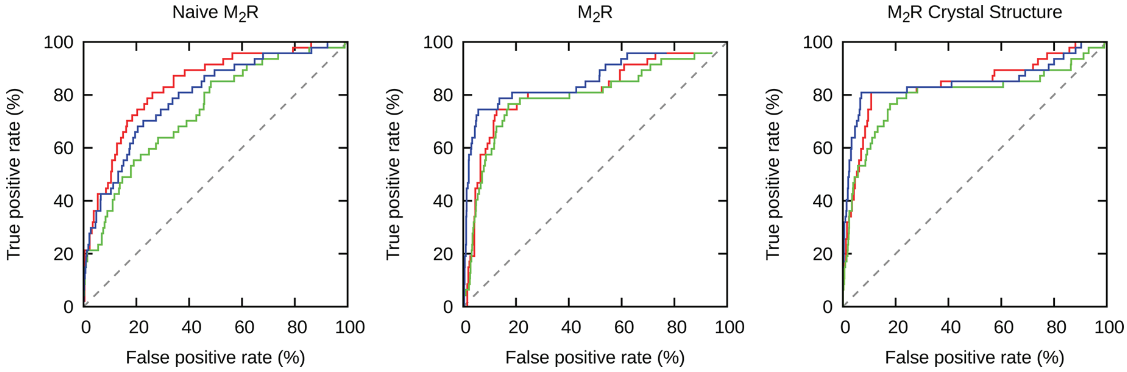


Figure 4. ROC curves for the M_2R naïve model (left), the M_2R optimized model (middle), and the M_2R crystal structure (right): (blue) set 1, Schrödinger; (green) set 2, ZINC; (red) set 3, refined Schrödinger. The dotted line indicates random choice (no enrichment).

has demonstrated that activity may be achieved without making this contact.[37]

Due to the high similarity of the five subtypes M_1R-M_5R, compounds which act at the M_1R usually also have some affinity for the other subtypes.[21] A rigorous test of model quality would be to dock compounds with a high level of specificity for individual subtypes into all subtypes so that an assessment of the selectivity of the homology models could be made. However, a significant difficulty encountered in this project has been to identify a sufficient number of compounds that are generally agreed to be selective for one receptor over the other four subtypes.

Comparison of Decoy Sets. The analysis of the VS data obtained using the Schrödinger decoy set (set 1) revealed that the results are biased toward low molecular weight compounds (across both active and decoy sets), which is reasonable given the characteristic small binding pocket of mAChRs.[24,25] Recent publications have given considerable attention to the development of receptor-appropriate decoy libraries.[15,59] Decoy sets, where the physical properties of compounds differ substantially from the corresponding active ligands, have been shown to lead to biased virtual screening results and often artificially good enrichment.[52]

Both the ZINC and refined Schrödinger decoy sets (sets 2 and 3) were matched to actives in terms of their physical properties, including the requirement to contain only compounds with an ionizable nitrogen at physiological pH. This filter was based on one of the benchmarks for model success, specifically their ability to generate the salt bridge between the ionizable nitrogen of a ligand and the Asp 3.32 residue of the receptor. Thus, these challenging sets of decoys were designed to investigate whether the docking and scoring process could select for this interaction in actives ahead of decoys that also contained an ionizable nitrogen.

The enrichment metrics (Table 4) and ROC and enrichment curves (Figures 2, S2, and S3 (Supporting Information)) demonstrate that indeed these sets of decoys are more challenging (particularly, for M_1R and M_3R). But encouragingly, the models produced enrichment and early enrichment values similar to that of nonproperty matched decoys (particularly, for M_4R and M_5R). These results indicate that our models are indeed capable of preferentially identifying active compounds among property-matched decoys.

Template Selection vs Binding Site Optimization. The choice of an appropriate template for GPCR homology modeling is an area of long-standing debate.[14−16] It has recently been demonstrated that, while important, the choice of template should be made while also considering issues such as binding site optimization and knowledge-enhancement of homology models. Specifically, Tropsha and co-workers[26] have compared the VS effectiveness of β_2AR crystal structures with a range of historical β_2AR models that were built before the crystal structures became available. They demonstrated that several models produced VS enrichment comparable to and even exceeding that of crystal structures.

Here we investigated the proposal that an optimized homology model may approach the quality of a crystal structure, even though it is based on a remote template. Using the recently solved structure of the rat M_3R[25] as a template, we built a naïve human M_2R homology model, i.e. a homology model that has not been optimized by IFD. This naïve model had an RMSD of 1.64 Å to the human M_2R crystal structure. It can be seen from the results of VS (Figures 4, S4, and S5 (Supporting Information) and Table 5) that the optimized model, based on the β_2AR template, significantly outperforms the naïve M_3R-based variant and produces results close to those for the M_2R crystal structure. These results mirror those obtained in VS against the homology models of the MT_2 melatonin receptor where the $EF_{2\%}$ increased from 0 to 5.2 for a crude model to 3.1−18.7 for an optimized model.[16] Similar to the observations for other muscarinic models (Table 4), decoy sets 2 and 3 (ZINC and refined Schrödinger) make discrimination of decoys and actives more difficult. However, even with these demanding decoys, the optimized model still outperforms the naïve model in terms of enrichment, if not early enrichment.

Table 5. Virtual Screening Evaluation of M_2 Muscarinic Receptors

				EF (at *X* % of ranked database)		
receptor	ROC AUC	$\text{LogAUC}_{0.001}$	ndist	2	5	10
		set 1 (Schrödinger decoy set)				
optimized model	0.86	0.47	3.72	11.7	11.4	7.4
naïve model	0.80	0.38	6.28	9.6	5.9	4.2
crystal structure	0.85	0.55	4.82	15.9	10.9	7.9
		set 2 (ZINC decoy set)				
optimized model	0.79	0.36	3.57	3.2	5.9	5.3
naïve model	0.74	0.33	6.02	8.5	4.2	3.4
crystal structure	0.80	0.42	5.16	8.5	8.0	5.3
		set 3 (refined Schrödinger decoy set)				
optimized model	0.81	0.36	3.72	2.8	3.7	4.0
naïve model	0.84	0.40	6.28	5.6	4.1	3.6
crystal structure	0.84	0.47	4.82	6.6	4.9	4.2

Binding site optimization takes into account the structural plasticity of a binding site and its adjustment to the structural demands of an active ligand. Our results demonstrate that the optimized M_2R model, based on the remote sequence template, was better at distinguishing actives from decoys than the naïve M_2R model, based on the close sequence template. Thus, it is clear that the choice of IFD ligand and the robustness of the IFD protocol could be as important for the production of a useful receptor model as the extent of target−template sequence similarity.

■ DISCUSSION

Several muscarinic receptor models have been generated over the past few years (summarized in Table 6), with the majority being of the M_1R.[60−69] Two models each of the M_3R[70,71] and the M_2R[72−74] and one model of the M_5R[75] have been also reported. These models were generally constructed in the course of molecular pharmacology studies to address issues of receptor activation and selectivity, allosterism, and bitopic binding, or receptor dimerization, although some groups have used predominantly modeling approaches to investigate the structural mechanisms of antagonist binding, receptor activa-

Table 6. Muscarinic Receptor Modeling Studies

receptor	template	homology modeling program	purpose	additional techniques used	ref
M_1R	rhodopsin	MODELLER	molecular pharmacology of allosteric modulation by a peptide ligand	loop modeling, MD, protein−protein docking	65
		Prime	molecular pharmacology of allosteric potentiation		62
		VEGA	modeling study to investigate receptor activation	MD in hydrated lipid bilayer	60
	β_2AR	MOE	modeling study to investigate allosteric modulation by a peptide ligand	MD in hydrated lipid bilayer, protein−protein docking	67
		QUANTA, MODELLER	molecular pharmacology of activation and selectivity	loop modeling	61, 76
			molecular pharmacology of allosterism and bitopic binding	loop modeling	63, 76
			molecular pharmacology of activation	loop modeling	64
	D_3R	MOE	molecular pharmacology of allosterism and bitopic binding	loop modeling	66
	M_3R	Prime	modeling study to investigate receptor activation	binding site refinement	77
	M_3R	MOE	molecular pharmacology of allosterism and bitopic binding		68
	M_2R	MOE	homology modeling		69
M_2R	M_3R	Prime, MODELLER, YASARA	modeling study to investigate the effect of template choice	IFD	72
	β_2AR	ICM	molecular pharmacology of allosterism and bitopic binding	flexible receptor docking of two agonists using BDMC algorithm	73, 74
M_3R	rhodopsin	MODELLER	modeling study to investigate structural mechanism of antagonist binding	MD in hydrated lipid bilayer	71
	β_1AR	Prime	molecular pharmacology of dimerization		70
M_5R	β_1AR	MODELLER	modeling study to investigate structural mechanism of antagonist binding	MD in hydrated lipid bilayer	75

tion, and allosteric modulation. A range of templates were used in these studies: rhodopsin, β_1AR and β_2AR, as well as the more recently solved D_3R, M_2R, and M_3R. Models were constructed using QUANTA/MODELLER, Prime, MOE, ICM, VEGA, and YASARA. Several approaches to additional model refinement were also implemented including MD in a hydrated lipid bilayer, loop modeling, and protein−protein docking. Significantly, the majority of the reported mAChR models were not optimized to generate knowledge-based models. In this study, we have developed such knowledge-based homology models of the muscarinic acetylcholine receptors M_1R−M_5R.

Binding site optimization has gained significant traction in the GPCR modeling field as an important way of using experimental knowledge (such as SAR and/or site-directed mutagenesis) to improve the quality and predictive power of naïve, or crude, homology models. Using 5-HT$_{2A}$R as a test-case,[27] we have previously demonstrated the importance of loop refinement and, particularly, binding site optimization for improving model quality and VS performance. Such improvements have been also achieved for GPCR models in a number of studies focused on what has been termed ligand-steered,[78] ligand-guided,[79,80] ligand-adapted,[16] or ligand-optimized[15] homology modeling (Table 7). Binding site optimization via a variety of methods—particularly those utilizing available experimental data about a target and its ligands—have been commonly used and shown to be successful in GPCR Dock assessments.[19,20]

Table 7. Binding Site Optimization Methods

method	receptor	ref
randomizing and clustering receptor complex structures	β_2AR and A_{2A}R	78
side chain conformation sampling in the presence of docked ligands	5-HT$_X$Rs	37
	A_XRs	80
backbone perturbation and binding site reshaping with elastic normal-mode analysis	CXCR4	30, 79
IFD	MT_2 melatonin receptor	16
	D_1R and D_2R	15

Ideally, an optimized model, based on a close sequence template, would be the best choice for virtual screening.[69,72] However, close sequence templates are not always available. In such cases, knowledge-based optimization, e.g. by using established actives, can improve a model (Table 7). Using the M_2R as a case study, we compared a naïve model, based on a close sequence template (M_3R), and an optimized model, based on a more remote template (β_2AR). The IFD optimized model outperformed the naïve model in virtual screening. This observation parallels that of Kolaczkowski et al.[15] who generated ligand-optimized homology models of the D_1 and D_2 dopamine receptors using IFD and tested them in VS against ZINC- and Schrödinger-based decoy libraries spiked with ligands specific for dopamine receptors. They found that binding site optimization significantly improved VS performance, while observing no advantage in using a D_3R-based D_2R model compared to a model based on the more evolutionary distant β_2AR. Our findings are also in agreement with those of Tropsha and co-workers,[26] who suggest that such knowledge-based models "may be even more useful for practical structure-based drug discovery than X-ray structures".[26] Thus, our results and those of others[15,17,26] provide evidence that binding site optimization greatly improves homology models for VS. Future work is required to evaluate homology models in a flexible receptor scenario: by on-the-fly receptor flexibility,[81,82] molecular dynamics,[83] or using receptor ensembles.[84,85]

Finally, we tested the M_1R−M_5R models against increasingly demanding decoy sets. Specifically, to avoid artificial enrichment due to active-favoring biases, we have matched physicochemical decoy properties to those of ligands active at muscarinic receptors. The results showed that indeed these sets of decoys were more challenging. However, even using our matched decoy sets, the models produced enrichment (including early enrichment), similar to that obtained using nonproperty matched decoys. Recently, Gatica and Cavasotto have published a GPCR decoy database, where 39 decoy molecules were selected for each GPCR ligand.[59] Similar to our findings, they observed a marked decrease in enrichment for matched decoys compared to bias-uncorrected decoys.

■ CONCLUSIONS

In this work, we have developed homology models of the muscarinic acetylcholine receptors M_1R−M_5R and evaluated them in VS for the identification of antagonists. The models were generated by Prime and optimized using IFD (Glide + Prime). Model refinement was guided by experimental knowledge of active compounds and critical binding site residues. The refinement resulted in ligand-induced adaptation of the receptor binding sites, which optimized them for antagonist recognition. The homology models were evaluated in retrospective VS using Glide and were capable of distinguishing known antagonists from matched decoy compounds. These results bolster confidence for prospective virtual screening using these receptor models. Even more significantly, our results support the following suppositions about homology modeling of GPCRs: (i) binding site optimization is a crucial step in model generation, (ii) knowledge-based homology models of GPCRs are appropriate for prospective VS, and (iii) property-matched decoys should be used in VS evaluation of homology models. In line with our past practice, we make the optimized M_1R−M_5R homology models freely available as part of the Supporting Information. We consider such open access as crucial in our field since it allows researchers to use these structures, compare them to their own results,[37,69] and thus advance the development of better modeling methods.

■ ASSOCIATED CONTENT

Supporting Information

List of actives used in virtual screening enrichment studies; properties of actives that either did not dock into M_1R−M_5R models or produced docked poses with a scoring energy greater than the set acceptable cutoff; enrichment plots and semilog ROC curves for M_1R−M_5R models and for the M_2R optimized model, compared to the M_2R naïve model and the M_2R crystal structure; and PDB files of homology models. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Authors

*Phone: +61 3 9903 9611. E-mail: Elizabeth.Yuriev@monash.edu.

*Phone: +61 3 9903 9110. E-mail: David.Chalmers@monash.edu.

Present Address

[§]F.M.M.: Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California, San Diego, 9500 Gilman Drive, MC 0747, La Jolla, CA, 92093-0747.

Author Contributions

[†]T.T. and K.C.M.: These authors have contributed equally to this work.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

T.T. is a recipient of an Australian Postgraduate Award (APA) scholarship. K.C.M. was a recipient of a UROP scholarship supported by BIO21, The University of Melbourne. F.M.M. was a recipient of an Australian Postgraduate Award (APA) scholarship. E.Y. was a recipient of the travel award provided by the Australian Academy of Science (COST Action CM1103). This work was supported by the Victorian Life Sciences Computation Initiative (VLSCI, grant number VR0004) and by the National Computational Infrastructure (grant number: y96), which is supported by the Australian Commonwealth Government.

■ REFERENCES

(1) Shoichet, B. K.; Kobilka, B. K. Structure-based drug screening for G-protein-coupled receptors. *Trends Pharmacol. Sci.* **2012**, *33*, 268–272.

(2) Mason, J. S.; Bortolato, A.; Congreve, M.; Marshall, F. H. New insights from structural biology into the druggability of G protein-coupled receptors. *Trends Pharmacol. Sci.* **2012**, *33*, 249–260.

(3) Granier, S.; Kobilka, B. A new era of GPCR structural and chemical biology. *Nat. Chem. Biol.* **2012**, *8*, 670–673.

(4) Stevens, R. C.; Cherezov, V.; Katritch, V.; Abagyan, R.; Kuhn, P.; Rosen, H.; Wuthrich, K. The GPCR Network: a large-scale collaboration to determine human GPCR structure and function. *Nat. Rev. Drug. Discov.* **2013**, *12*, 25–34.

(5) Palczewski, K.; Kumasaka, T.; Hori, T.; Behnke, C. A.; Motoshima, H.; Fox, B. A.; Le Trong, I.; Teller, D. C.; Okada, T.; Stenkamp, R. E.; Yamamoto, M.; Miyano, M. Crystal structure of rhodopsin: A G protein-coupled receptor. *Science* **2000**, *289*, 739–745.

(6) Cherezov, V.; Rosenbaum, D. M.; Hanson, M. A.; Rasmussen, S. G.; Thian, F. S.; Kobilka, T. S.; Choi, H. J.; Kuhn, P.; Weis, W. I.; Kobilka, B. K.; Stevens, R. C. High-resolution crystal structure of an engineered human beta2-adrenergic G protein-coupled receptor. *Science* **2007**, *318*, 1258–1265.

(7) Rasmussen, S. G.; Choi, H. J.; Rosenbaum, D. M.; Kobilka, T. S.; Thian, F. S.; Edwards, P. C.; Burghammer, M.; Ratnala, V. R.; Sanishvili, R.; Fischetti, R. F.; Schertler, G. F.; Weis, W. I.; Kobilka, B. K. Crystal structure of the human beta2 adrenergic G-protein-coupled receptor. *Nature* **2007**, *450*, 383–387.

(8) Rosenbaum, D. M.; Cherezov, V.; Hanson, M. A.; Rasmussen, S. G.; Thian, F. S.; Kobilka, T. S.; Choi, H. J.; Yao, X. J.; Weis, W. I.; Stevens, R. C.; Kobilka, B. K. GPCR engineering yields high-resolution structural insights into beta2-adrenergic receptor function. *Science* **2007**, *318*, 1266–1273.

(9) Serrano-Vega, M. J.; Magnani, F.; Shibata, Y.; Tate, C. G. Conformational thermostabilization of the β1-adrenergic receptor in a detergent-resistant form. *Proc. Natl. Acad. Sci. USA* **2008**, *105*, 877–882.

(10) Caffrey, M.; Li, D.; Dukkipati, A. Membrane protein structure determination using crystallography and lipidic mesophases: recent advances and successes. *Biochemistry* **2012**, *51*, 6266–6288.

(11) Topiol, S. X-ray structural information of GPCRs in drug design: what are the limitations and where do we go? *Expert Opin. Drug Discov.* **2013**, *8*, 607–620.

(12) Fredriksson, R.; Lagerstrom, M. C.; Lundin, L. G.; Schioth, H. B. The G-protein-coupled receptors in the human genome form five main families. Phylogenetic analysis, paralogon groups, and fingerprints. *Mol. Pharmacol.* **2003**, *63*, 1256–1272.

(13) Kooistra, A. J.; Roumen, L.; Leurs, R.; de Esch, I. J.; de Graaf, C. From heptahelical bundle to hits from the haystack: structure-based virtual screening for GPCR ligands. *Methods Enzymol.* **2013**, *522*, 279–336.

(14) Mobarec, J. C.; Sanchez, R.; Filizola, M. Modern homology modeling of G-protein coupled receptors: which structural template to use? *J. Med. Chem.* **2009**, *52*, 5207–5216.

(15) Kolaczkowski, M.; Bucki, A.; Feder, M.; Pawlowski, M. Ligand-optimized homology models of D_1 and D_2 Dopamine receptors: Application for virtual screening. *J. Chem. Inf. Model.* **2013**, *53*, 638–648.

(16) Pala, D.; Beuming, T.; Sherman, W.; Lodola, A.; Rivara, S.; Mor, M. Structure-based virtual screening of MT_2 Melatonin receptor: Influence of template choice and structural refinement. *J. Chem. Inf. Model.* **2013**, *53*, 821–835.

(17) Beuming, T.; Sherman, W. Current assessment of docking into GPCR crystal structures and homology models: successes, challenges, and guidelines. *J. Chem. Inf. Model.* **2012**, *52*, 3263–3277.

(18) Anighoro, A.; Rastelli, G. Enrichment factor analyses on G-protein coupled receptors with known crystal structure. *J. Chem. Inf. Model.* **2013**, *53*, 739–743.

(19) Kufareva, I.; Rueda, M.; Katritch, V.; Stevens, R. C.; Abagyan, R. Status of GPCR modeling and docking as reflected by community-wide GPCR Dock 2010 assessment. *Structure* **2011**, *19*, 1108–1126.

(20) Michino, M.; Abola, E.; Brooks, C. L., 3rd; Dixon, J. S.; Moult, J.; Stevens, R. C. Community-wide assessment of GPCR structure modelling and ligand docking: GPCR Dock 2008. *Nat. Rev. Drug Discov.* **2009**, *8*, 455–463.

(21) Wess, J.; Eglen, R. M.; Gautam, D. Muscarinic acetylcholine receptors: mutant mice provide new insights for drug development. *Nat. Rev. Drug. Discov.* **2007**, *6*, 721–733.

(22) Sherman, W.; Day, T.; Jacobson, M. P.; Friesner, R. A.; Farid, R. Novel procedure for modeling ligand/receptor induced fit effects. *J. Med. Chem.* **2006**, *49*, 534–553.

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

(24) Haga, K.; Kruse, A. C.; Asada, H.; Yurugi-Kobayashi, T.; Shiroishi, M.; Zhang, C.; Weis, W. I.; Okada, T.; Kobilka, B. K.; Haga, T.; Kobayashi, T. Structure of the human M2 muscarinic acetylcholine receptor bound to an antagonist. *Nature* **2012**, *482*, 547–551.

(25) Kruse, A. C.; Hu, J.; Pan, A. C.; Arlow, D. H.; Rosenbaum, D. M.; Rosemond, E.; Green, H. F.; Liu, T.; Chae, P. S.; Dror, R. O.; Shaw, D. E.; Weis, W. I.; Wess, J.; Kobilka, B. K. Structure and dynamics of the M3 muscarinic acetylcholine receptor. *Nature* **2012**, *482*, 552–556.

(26) Tang, H.; Wang, X. S.; Hsieh, J. H.; Tropsha, A. Do crystal structures obviate the need for theoretical models of GPCRs for structure-based virtual screening? *Proteins* **2012**, *80*, 1503–1521.

(27) McRobb, F. M.; Capuano, B.; Crosby, I. T.; Chalmers, D.; Yuriev, E. Homology modeling and docking evaluation of aminergic G protein-coupled receptors. *J. Chem. Inf. Model.* **2010**, *50*, 626–637.

(28) Irwin, J. J.; Sterling, T.; Mysinger, M. M.; Bolstad, E. S.; Coleman, R. G. ZINC: A free tool to discover chemistry for biology. *J. Chem. Inf. Model.* **2012**, *52*, 1757–1768.

(29) Carlsson, J.; Coleman, R. G.; Setola, V.; Irwin, J. J.; Fan, H.; Schlessinger, A.; Sali, A.; Roth, B. L.; Shoichet, B. K. Ligand discovery from a dopamine D_3 receptor homology model and crystal structure. *Nat. Chem. Biol.* **2011**, *7*, 769–778.

(30) Mysinger, M. M.; Weiss, D. R.; Ziarek, J. J.; Gravel, S.; Doak, A. K.; Karpiak, J.; Heveker, N.; Shoichet, B. K.; Volkman, B. F. Structure-based ligand discovery for the protein-protein interface of chemokine receptor CXCR4. *Proc. Natl. Acad. Sci. USA* **2012**, *109*, 5517–5522.

(31) Yoshikawa, Y.; Oishi, S.; Kubo, T.; Tanahara, N.; Fujii, N.; Furuya, T. Optimized method of G-protein-coupled receptor homology modeling: Its application to the discovery of novel CXCR7 ligands. *J. Med. Chem.* **2013**, *56*, 4236−4251.

(32) van Loenen, P. B.; de Graaf, C.; Verzijl, D.; Leurs, R.; Rognan, D.; Peters, S. L. M.; Alewijnse, A. E. Agonist-dependent effects of mutations in the sphingosine-1-phosphate type 1 receptor. *Eur. J. Pharmacol.* **2011**, *667*, 105−112.

(33) Sanders, M. P.; Roumen, L.; van der Horst, E.; Lane, J. R.; Vischer, H. F.; van Offenbeek, J.; de Vries, H.; Verhoeven, S.; Chow, K. Y.; Verkaar, F.; Beukers, M. W.; McGuire, R.; Leurs, R.; Ijzerman, A. P.; de Vlieg, J.; de Esch, I. J.; Zaman, G. J.; Klomp, J. P.; Bender, A.; de Graaf, C. A prospective cross-screening study on G-protein-coupled receptors: Lessons learned in virtual compound library design. *J. Med. Chem.* **2012**, *55*, 5311−5325.

(34) Langmead, C. J.; Andrews, S. P.; Congreve, M.; Errey, J. C.; Hurrell, E.; Marshall, F. H.; Mason, J. S.; Richardson, C. M.; Robertson, N.; Zhukov, A.; Weir, M. Identification of novel adenosine A_{2A} receptor antagonists by virtual screening. *J. Med. Chem.* **2012**, *55*, 1904−1909.

(35) Kołaczkowski, M.; Nowak, M.; Pawłowski, M.; Bojarski, A. J. Receptor-based pharmacophores for serotonin 5-HT_7R antagonists: Implications to selectivity. *J. Med. Chem.* **2006**, *49*, 6732−6741.

(36) Kurczab, R.; Nowak, M.; Chilmonczyk, Z.; Sylte, I.; Bojarski, A. J. The development and validation of a novel virtual screening cascade protocol to identify potential serotonin 5-HT(7)R antagonists. *Bioorg. Med. Chem. Lett.* **2010**, *20*, 2465−2468.

(37) Lin, X.; Huang, X. P.; Chen, G.; Whaley, R.; Peng, S.; Wang, Y.; Zhang, G.; Wang, S. X.; Wang, S.; Roth, B. L.; Huang, N. Life beyond kinases: structure-based discovery of sorafenib as nanomolar antagonist of 5-HT receptors. *J. Med. Chem.* **2012**, *55*, 5749−5759.

(38) Heifetz, A.; Barker, O.; Verquin, G.; Wimmer, N.; Meutermans, W.; Pal, S.; Law, R. J.; Whittaker, M. Fighting obesity with a sugar-based library: Discovery of novel MCH-1R antagonists by a new computational-VAST approach for exploration of GPCR binding sites. *J. Chem. Inf. Model.* **2013**, *53*, 1084−1099.

(39) Renault, N.; Laurent, X.; Farce, A.; El Bakali, J.; Mansouri, R.; Gervois, P.; Millet, R.; Desreumaux, P.; Furman, C.; Chavatte, P. Virtual screening of CB_2 receptor agonists from bayesian network and high-throughput docking: structural insights into agonist-modulated GPCR features. *Chem. Biol. Drug. Des.* **2013**, *81*, 442−454.

(40) Suite 2012: Maestro, version 9.3; LigPrep, version 2.5; Schrödinger Suite 2012 Protein Preparation Wizard; Schrödinger Suite 2012 Induced Fit Docking protocol; Glide version 5.8; Prime version 3.1; Schrödinger, LLC: New York, NY, 2012.

(41) Suite 2011: Maestro, version 9.2; LigPrep, version 2.5; Schrödinger Suite 2011 Protein Preparation Wizard; Schrödinger Suite 2011 Induced Fit Docking protocol; Glide version 5.7; Prime version 3.0; Schrödinger, LLC: New York, NY, 2011.

(42) Jacobson, M. P.; Pincus, D. L.; Rapp, C. S.; Day, T. J.; Honig, B.; Shaw, D. E.; Friesner, R. A. A hierarchical approach to all-atom protein loop prediction. *Proteins* **2004**, *55*, 351−367.

(43) Thompson, J. D.; Higgins, D. G.; Gibson, T. J. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **1994**, *22*, 4673−4680.

(44) Sastry, G. M.; Adzhigirey, M.; Day, T.; Annabhimoju, R.; Sherman, W. Protein and ligand preparation: parameters, protocols, and influence on virtual screening enrichments. *J. Comput.-Aided Mol. Des.* **2013**, *27*, 221−234.

(45) Friesner, R. A.; Banks, J. L.; Murphy, R. B.; Halgren, T. A.; Klicic, J. J.; Mainz, D. T.; Repasky, M. P.; Knoll, E. H.; Shelley, M.; Perry, J. K.; Shaw, D. E.; Francis, P.; Shenkin, P. S. Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *J. Med. Chem.* **2004**, *47*, 1739−1749.

(46) Halgren, T. A.; Murphy, R. B.; Friesner, R. A.; Beard, H. S.; Frye, L. L.; Pollard, W. T.; Banks, J. L. Glide: a new approach for rapid, accurate docking and scoring. 2. Enrichment factors in database screening. *J. Med. Chem.* **2004**, *47*, 1750−1759.

(47) Chalmers, D. K.; Roberts, B. P. *Silico—A Perl Molecular Modelling Toolkit*; Monash University: Melbourne, 2011.

(48) Ballesteros, J. A.; Weinstein, H.; Stuart, C. S. 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.

(49) Holst, B.; Nygaard, R.; Valentin-Hansen, L.; Bach, A.; Engelstoft, M. S.; Petersen, P. S.; Frimurer, T. M.; Schwartz, T. W. A conserved aromatic lock for the tryptophan rotameric switch in TM-VI of seven-transmembrane receptors. *J. Biol. Chem.* **2010**, *285*, 3973−3985.

(50) Spalding, T. A.; Birdsall, N. J.; Curtis, C. A.; Hulme, E. C. Acetylcholine mustard labels the binding site aspartate in muscarinic acetylcholine receptors. *J. Biol. Chem.* **1994**, *269*, 4092−4097.

(51) Okuno, Y.; Tamon, A.; Yabuuchi, H.; Niijima, S.; Minowa, Y.; Tonomura, K.; Kunimoto, R.; Feng, C. GLIDA: GPCR−ligand database for chemical genomics drug discovery−database and tools update. *Nucleic Acids Res.* **2008**, *36*, D907−D912.

(52) Huang, N.; Shoichet, B. K.; Irwin, J. J. Benchmarking sets for molecular docking. *J. Med. Chem.* **2006**, *49*, 6789−6801.

(53) Shi, L.; Javitch, J. A. The binding site of aminergic G protein-coupled receptors: the transmembrane segments and second extracellular loop. *Annu. Rev. Pharmacol. Toxicol.* **2002**, *42*, 437−467.

(54) Bymaster, F. P.; Felder, C. C.; Tzavara, E.; Nomikos, G. G.; Calligaro, D. O.; McKinzie, D. L. Muscarinic mechanisms of antipsychotic atypicality. *Prog. Neuro-Psychopharmacol. Biol. Psychiatry* **2003**, *27*, 1125−1143.

(55) Hawkins, P. C. D.; Warren, G. L.; Skillman, A. G.; Nicholls, A. How to do an evaluation: pitfalls and traps. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 179−190.

(56) Nicholls, A. What do we know and when do we know it? *J. Comput.-Aided Mol. Des.* **2008**, *22*, 239−255.

(57) Katritch, V.; Rueda, M.; Lam, P. C.; Yeager, M.; Abagyan, R. GPCR 3D homology models for ligand screening: lessons learned from blind predictions of adenosine A2a receptor complex. *Proteins* **2010**, *78*, 197−211.

(58) Mysinger, M. M.; Shoichet, B. K. Rapid context-dependent ligand desolvation in molecular docking. *J. Chem. Inf. Model.* **2010**, *50*, 1561−1573.

(59) Gatica, E. A.; Cavasotto, C. N. Ligand and decoy sets for docking to G protein-coupled receptors. *J. Chem. Inf. Model.* **2012**, *52*, 1−6.

(60) Espinoza-Fonseca, L. M.; Pedretti, A.; Vistoli, G. Structure and dynamics of the full-length M1 muscarinic acetylcholine receptor studied by molecular dynamics simulations. *Arch. Biochem. Biophys.* **2008**, *469*, 142−150.

(61) Lebon, G.; Langmead, C. J.; Tehan, B. G.; Hulme, E. C. Mutagenic mapping suggests a novel binding mode for selective agonists of M1 muscarinic acetylcholine receptors. *Mol. Pharmacol.* **2009**, *75*, 331−341.

(62) Ma, L.; Seager, M. A.; Wittmann, M.; Jacobson, M.; Bickel, D.; Burno, M.; Jones, K.; Graufelds, V. K.; Xu, G.; Pearson, M.; McCampbell, A.; Gaspar, R.; Shughrue, P.; Danziger, A.; Regan, C.; Flick, R.; Pascarella, D.; Garson, S.; Doran, S.; Kreatsoulas, C.; Veng, L.; Lindsley, C. W.; Shipe, W.; Kuduk, S.; Sur, C.; Kinney, G.; Seabrook, G. R.; Ray, W. J. Selective activation of the M1 muscarinic acetylcholine receptor achieved by allosteric potentiation. *Proc. Natl. Acad. Sci. USA* **2009**, *106*, 15950−15955.

(63) Avlani, V. A.; Langmead, C. J.; Guida, E.; Wood, M. D.; Tehan, B. G.; Herdon, H. J.; Watson, J. M.; Sexton, P. M.; Christopoulos, A. Orthosteric and allosteric modes of interaction of novel selective agonists of the M1 muscarinic acetylcholine receptor. *Mol. Pharmacol.* **2010**, *78*, 94−104.

(64) Kaye, R. G.; Saldanha, J. W.; Lu, Z. L.; Hulme, E. C. Helix 8 of the M1 muscarinic acetylcholine receptor: scanning mutagenesis delineates a G protein recognition site. *Mol. Pharmacol.* **2011**, *79*, 701−709.

(65) Marquer, C.; Fruchart-Gaillard, C.; Letellier, G.; Marcon, E.; Mourier, G.; Zinn-Justin, S.; Menez, A.; Servent, D.; Gilquin, B.

Structural model of ligand-G protein-coupled receptor (GPCR) complex based on experimental double mutant cycle data: MT7 snake toxin bound to dimeric hM1 muscarinic receptor. *J. Biol. Chem.* **2011**, *286*, 31661−31675.

(66) Daval, S. B.; Valant, C.; Bonnet, D.; Kellenberger, E.; Hibert, M.; Galzi, J. L.; Ilien, B. Fluorescent derivatives of AC-42 to probe bitopic orthosteric/allosteric binding mechanisms on muscarinic M1 receptors. *J. Med. Chem.* **2012**, *55*, 2125−2143.

(67) Xu, J.; Chen, H. Interpreting the structural mechanism of action for MT7 and human muscarinic acetylcholine receptor 1 complex by modeling protein-protein interaction. *J. Biomol. Struct. Dyn.* **2012**, *30*, 30−44.

(68) Daval, S. B.; Kellenberger, E.; Bonnet, D.; Utard, V.; Galzi, J. L.; Ilien, B. Exploration of the orthosteric/allosteric interface in human M1 muscarinic receptors by bitopic fluorescent ligands. *Mol. Pharmacol.* **2013**, *84*, 71−85.

(69) Jójárt, B.; Balint, A. M.; Balint, S.; Viskolcz, B. Homology modeling and validation of the human M1 muscarinic acetylcholine receptor. *Mol. Inf.* **2012**, *31*, 635−638.

(70) McMillin, S. M.; Heusel, M.; Liu, T.; Costanzi, S.; Wess, J. Structural basis of M3 muscarinic receptor dimer/oligomer formation. *J. Biol. Chem.* **2011**, *286*, 28584−28598.

(71) Martinez-Archundia, M.; Cordomi, A.; Garriga, P.; Perez, J. J. Molecular modeling of the M3 acetylcholine muscarinic receptor and its binding site. *J. Biomed. Biotechnol.* **2012**, *2012*, 789741.

(72) Jakubik, J.; Randakova, A.; Dolezal, V. On homology modeling of the M2 muscarinic acetylcholine receptor subtype. *J. Comput.-Aided Mol. Des.* **2013**, *27*, 525−538.

(73) Valant, C.; Gregory, K. J.; Hall, N. E.; Scammells, P. J.; Lew, M. J.; Sexton, P. M.; Christopoulos, A. A novel mechanism of G protein-coupled receptor functional selectivity. Muscarinic partial agonist McN-A-343 as a bitopic orthosteric/allosteric ligand. *J. Biol. Chem.* **2008**, *283*, 29312−29321.

(74) Gregory, K. J.; Hall, N. E.; Tobin, A. B.; Sexton, P. M.; Christopoulos, A. Identification of orthosteric and allosteric site mutations in M2 muscarinic acetylcholine receptors that contribute to ligand-selective signaling bias. *J. Biol. Chem.* **2010**, *285*, 7459−7474.

(75) Huang, X.; Zheng, G.; Zhan, C. G. Microscopic binding of M5 muscarinic acetylcholine receptor with antagonists by homology modeling, molecular docking, and molecular dynamics simulation. *J. Phys. Chem. B* **2012**, *116*, 532−541.

(76) Blaney, F. E.; Raveglia, L. F.; Artico, M.; Cavagnera, S.; Dartois, C.; Farina, C.; Grugni, M.; Gagliardi, S.; Luttmann, M. A.; Martinelli, M.; Nadler, G. M.; Parini, C.; Petrillo, P.; Sarau, H. M.; Scheideler, M. A.; Hay, D. W.; Giardina, G. A. Stepwise modulation of neurokinin-3 and neurokinin-2 receptor affinity and selectivity in quinoline tachykinin receptor antagonists. *J. Med. Chem.* **2001**, *44*, 1675−1689.

(77) Chin, S. P.; Buckle, M. J. C.; Chalmers, D. K.; Yuriev, E.; Doughty, S. W. Towards activated homology models of the human M_1 muscarinic acetylcholine receptor. *J. Mol. Graph. Model.* **2014**, submitted.

(78) Phatak, S. S.; Gatica, E. A.; Cavasotto, C. N. Ligand-steered modeling and docking: a benchmarking study in class a g-protein-coupled receptors. *J. Chem. Inf. Model.* **2010**, *50*, 2119−2128.

(79) Neves, M. A.; Simoes, S.; Sa e Melo, M. L. Ligand-guided optimization of CXCR4 homology models for virtual screening using a multiple chemotype approach. *J. Comput.-Aided Mol. Des.* **2010**, *24*, 1023−1033.

(80) Katritch, V.; Kufareva, I.; Abagyan, R. Structure based prediction of subtype-selectivity for adenosine receptor antagonists. *Neuropharmacology* **2011**, *60*, 108−115.

(81) Yuriev, E.; Agostino, M.; Ramsland, P. A. Challenges and advances in computational docking: 2009 in review. *J. Mol. Recognit.* **2011**, *24*, 149−164.

(82) Yuriev, E.; Ramsland, P. A. Latest developments in molecular docking: 2010−2011 in review. *J. Mol. Recognit.* **2013**, *26*, 215−239.

(83) 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.

(84) Bottegoni, G.; Rocchia, W.; Rueda, M.; Abagyan, R.; Cavalli, A. Systematic exploitation of multiple receptor conformations for virtual ligand screening. *PLoS ONE* **2011**, *6*, e18845.

(85) Rueda, M.; Bottegoni, G.; Abagyan, R. Recipes for the selection of experimental protein conformations for virtual screening. *J. Chem. Inf. Model.* **2010**, *50*, 186−193.

Chapter 3

Ligand Binding Pathways of Clozapine and Haloperidol in the Dopamine D_2 and D_3 Receptors

The dopamine receptors attract considerable interest in the drug-design community due to their involvement in several high impact disease states, such as addiction, Parkinson's disease, and schizophrenia. At the time of writing, there is only one available experimentally determined structure of a dopamine receptor (D_3R) and the co-crystallized ligand (eticlopride) bears little resemblance to most antipsychotic compounds. This leaves a pressing need for predicted binding modes of relevant antipsychotic drugs. Antipsychotics have problematic side-effect profiles due to poor selectivity between many GPCRs. Differences in the selectivity of drugs for the D_2R and D_3R are poorly explained by comparing static models of their orthosteric binding sites, leading to a need to understand the dynamics of the receptor and behavior of the ligand outside of the bound pose.

To both predict the bound poses of haloperidol and clozapine in the D_3R and to investigate the behaviors of these ligands during binding, we performed a series of unbiased simulations of clozapine or haloperidol binding to the dopamine receptors. These simulations enabled us to observe, for the first time, complete binding pathways for each of these pharmaceutically important ligands, to infer metastable binding states, and to identify common binding mechanisms between these two significantly different ligands. The bound poses of these drugs have not been experimentally determined, and our simulations predict the bound poses with a level of sophistication above the docking methods employed in the literature.

Our attempts to simulate the binding pathways of clozapine and haloperidol in this work were met with early success. The two complete binding simulations presented in this chapter were amongst the first simulations performed and we were inspired by the ease of this victory to run many more simulations to better describe the binding process. However, we found that conventional MD simulations of GPCR systems were too slow to efficiently explore the binding ensemble and much longer timescales were required to access unbinding pathways or protein dynamics. This was our first experience with the inefficiency of conventional molecular dynamics simulations and led to our use of Markov state models in future work.

This chapter is the published article:

Thomas, T.; Fang, Y.; Yuriev, E.; Chalmers, D. K. Ligand Binding Pathways of Clozapine and Haloperidol in the Dopamine D2 and D3 Receptors. *J. Chem. Inf. Model.* **2015**, *56*, 308–321.

Journal of
CHEMICAL INFORMATION
AND MODELING

Article

pubs.acs.org/jcim

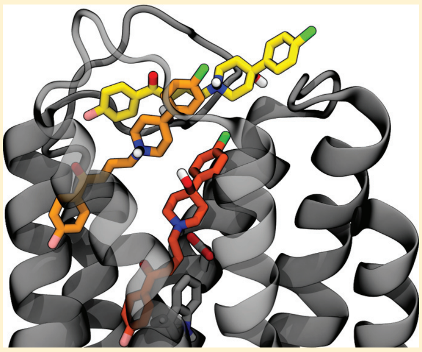
Ligand Binding Pathways of Clozapine and Haloperidol in the Dopamine D_2 and D_3 Receptors

Trayder Thomas, Yu Fang, Elizabeth Yuriev,* and David K. Chalmers*

Medicinal Chemistry, Monash Institute of Pharmaceutical Sciences, Monash University, 381 Royal Pde, Parkville, Victoria 3052, Australia

S *Supporting Information*

ABSTRACT: The binding of a small molecule ligand to its protein target is most often characterized by binding affinity and is typically viewed as an on/off switch. The more complex reality is that binding involves the ligand passing through a series of intermediate states between the solution phase and the fully bound pose. We have performed a set of 29 unbiased molecular dynamics simulations to model the binding pathways of the dopamine receptor antagonists clozapine and haloperidol binding to the D_2 and D_3 dopamine receptors. Through these simulations we have captured the binding pathways of clozapine and haloperidol from the extracellular vestibule to the orthosteric binding site and thereby, we also predict the bound pose of each ligand. These are the first long time scale simulations of haloperidol or clozapine binding to dopamine receptors. From these simulations, we have identified several important stages in the binding pathway, including the involvement of Tyr7.35 in a "handover" mechanism that transfers the ligand between the extracellular vestibule and Asp3.32. We have also performed interaction and cluster analyses to determine differences in binding pathways between the D_2 and D_3 receptors and identified metastable states that may be of use in drug design.



INTRODUCTION

The binding of a small molecule ligand to its protein target is often viewed as an "all or none" process; the ligand is either free in solution, or bound to a high-affinity binding site on the protein. This two-state model is clearly a simplification. On its way from solution to the high-affinity site, the ligand must actually traverse some pathway that likely consists of a number of intermediate lower energy states separated by higher energy transition states. Over the past few years, there has been increasing interest in characterizing ligand binding pathways for a number of reasons; first, knowledge of the binding pathway is the key to a detailed understanding of ligand binding kinetics,[1] and second, it has become apparent that intermediate states in the pathway of a ligand may themselves be additional binding sites that can be exploited in drug development.[2] Our group is particularly interested in ligand binding to G protein-coupled receptors (GPCRs),[3−7] and this study explores ligand binding pathways in the pharmaceutically important dopamine D_2 and D_3 receptors (D_2R, D_3R) using molecular dynamics (MD) simulations.

Binding pathways are difficult to investigate experimentally, but simulations can provide useful insight into the binding process for a wide range of systems. For example, an MD investigation of phosphate binding to GlpT revealed the mechanism by which the enzyme recruited the substrate, directed it to the binding site, and stabilized the bound pose.[8] In another study, an investigation of fatty acids binding to β-lactoglobulin revealed that specific residues aided the desolvation of the ligand,[9] and a study of ligand binding to the cannabinoid CB2 receptor found that the ligand passed from the lipid bilayer into the binding pocket through a specific transfer between two helices.[10] For GPCRs, many MD simulations described in the literature focus on the receptor activation mechanism.[11,12] Paired with experimental studies, these simulations have revealed that switching between the receptors' "active" and "inactive" states is not simply an on/off process but rather the selection of families of states that are capable of activating a range of cellular pathways from an extensive ensemble of states that are available to the receptor.[13] MD on short time scales has also been used as a tool to refine docked structures, accounting for receptor flexibility, and improving the prediction of the ligand bound state.[14,15]

We are most interested in the behavior of the ligand throughout the binding process, particularly for GPCR-binding ligands. One of the most interesting observations made from long time scale MD simulations of ligand binding to GPCRs[2,16] is the identification of "metastable binding sites" that are present along the binding pathway, where the ligand pauses for a significant amount of time before resuming its journey. Some of these sites have also been found to correspond to known allosteric sites. Specifically, in an MD study of tiotropium binding to the M2 muscarinic acetylcholine receptor (M_2R), the ligand paused at a previously established allosteric site.[16] As the

Received: July 22, 2015
Published: December 21, 2015

DOI: 10.1021/acs.jcim.5b00457

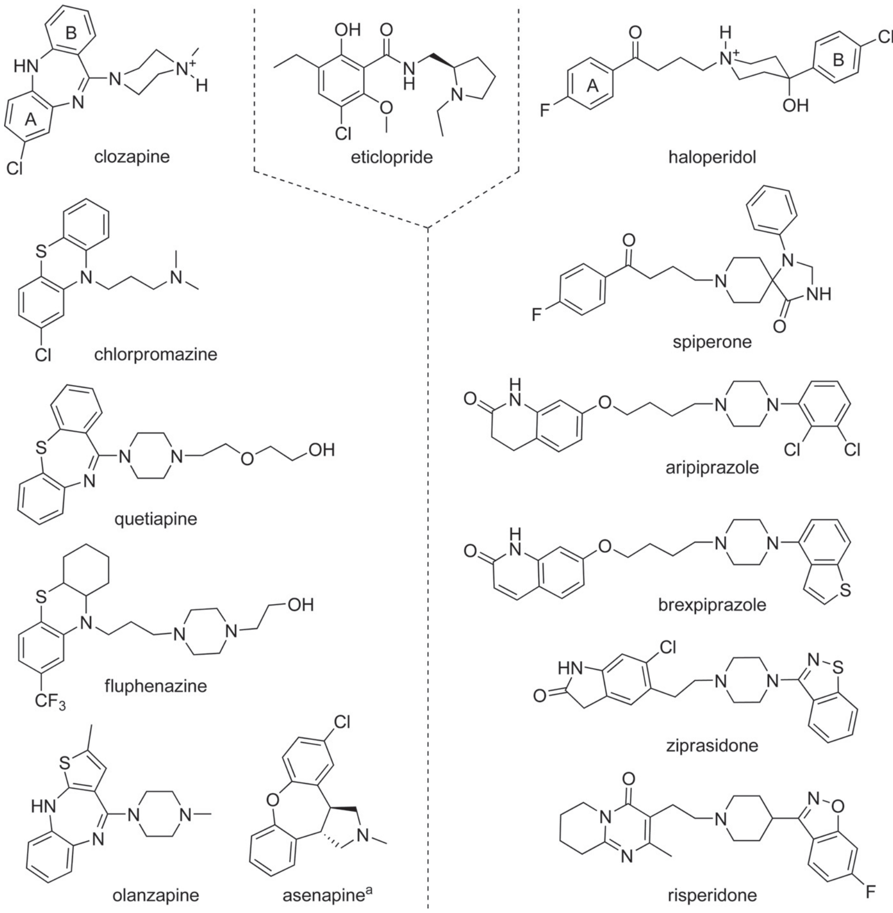


Figure 1. Structures of clozapine and haloperidol (shown in their protonated states) have scaffolds that are similar to many other antipsychotics. Eticlopride is present in the crystal structure of the D_3 receptor.[26] Aromatic rings in clozapine and haloperidol have been labeled A or B and will be referred to by these labels in the text. [a]Asenapine is a racemic mixture; one enantiomer is shown here.

residues making up the metastable binding sites are commonly less conserved than the orthosteric site itself, they represent intriguing potential targets for structure-based drug design. In an investigation of the binding of allosteric modulators to the M_2R, Dror et al. were able to use MD to predict the binding sites of a set of both positive and negative allosteric modulators, as well as the mechanism by which they functioned.[17]

It is being increasingly realized that in many cases the kinetics of binding, rather than the binding affinity, are more closely related to the efficacy of a drug[18−22] and that an understanding of the process by which a ligand binds or unbinds can assist the rational design of ligands. Detailed binding information is not provided by most experimental methods, which give static or averaged pictures of ligand binding. However, fine-grained temporal and spatial information can be gained from MD simulations. In this study, we investigate the association of the dopamine receptor antagonists clozapine and haloperidol with dopamine D_2 and D_3 receptors. Clozapine and haloperidol are representative of two common drug scaffolds from which many other antipsychotics are built (Figure 1). Both ligands bind to the dopamine D_2 and D_3 receptors with nanomolar affinity (0.12−960 nM) but have no appreciable selectivity for either receptor.[23−25] Haloperidol belongs to the broader butyrophenone-like class, containing two aromatic systems separated by a protonated amine containing linker. Clozapine is a dibenzodiazapine and is structurally similar to other tricyclic antipsychotics that contain a protonated amine located 3−4 bonds away from various tricyclic systems. These drugs have been in clinical use for over 40 years and are two of the most studied antipsychotic compounds but, despite their clinical importance, there are no reported MD studies that investigate the binding pathways of either ligand.

The only reported experimental structure of a ligand bound to a dopamine receptor is that of eticlopride (Figure 1) bound to the D_3R.[26] Eticlopride is smaller than clozapine and haloperidol and has a notably different chemical scaffold, which means that the detailed binding orientations of compounds similar to clozapine or haloperidol cannot be inferred from the eticlopride crystal structure and are currently unknown. In the absence of crystallographic information, a number of researchers have used molecular docking to predict the binding orientations of clozapine-like[27−29] and haloperidol-like[29−32] compounds. The docking studies of clozapine are in broad agreement on the overall features of the bound pose, predicting a salt-bridge between the ligand protonated amine and Asp3.32 and interactions between the tricyclic system and transmembrane helices 5 and 6 (TM5,6), but there is disagreement on the orientation of the tricyclic system, which is predicted to bind in two very similar poses that differ by a 180° rotation of the tricyclic system and correspondingly whether the A- or B-ring is buried more deeply within the receptor. The structural similarity of the two predicted clozapine poses means that the correct bound orientation cannot be determined through docking alone. In an effort to determine the correct bound orientation, Selent et al. compared the measured binding affinities of clozapine and the structurally similar olanzapine (Figure 1) in a mutagenesis study.[28] They concluded that the favored pose is where the B-ring is buried more deeply. Docking studies of haloperidol[29−31] agree on the geometric space occupied by the ligand but again differ on the gross ligand orientation. Most studies predict that the B-ring is buried more deeply in the receptor, although there is a more robust QM case that predicts the A-ring end of haloperidol is the deeper binding moiety.[32] While such docking studies attempt to predict the most stable pose in each ligand−receptor complex, they are unable to provide insight into the dynamics of the system. Similarly, experimental structures are unlikely to exist in the conformations that correspond to metastable states, and the binding pathway cannot be predicted from the experimentally determined ligand pose.

In this paper we conduct extensive unbiased simulations of clozapine and haloperidol binding to the D_2 and D_3 dopamine receptors with the aim of identifying the intermediate states on the binding pathway. These simulations have allowed us to identify metastable and bound states for ligands of these commonly employed scaffolds, which have not been identified by experimental methods. We have identified key residues involved in the recognition of these ligands and we propose a mechanism that transfers the ligand from the extracellular vestibule to the interior of the receptor, leading to the formation of the Asp3.32 salt-bridge and orthosteric binding. We expect that the knowledge gained here will assist in the rational drug design of dopamine receptor targeting ligands.

■ METHODS

The high sequence conservation in the transmembrane region of GPCRs allows residues to be commonly referred to using the Ballesteros−Weinstein numbering scheme,[33] in which the most conserved residue in each helix is numbered 50 and other residues in the helix are numbered relative to it. For example, Trp6.48 is located in TM6, two residues prior to the most conserved residue, Pro6.50. We have used the D_3R crystal structure as a reference for helical ranges and loop residues are referred to using the residue numbers, according to sequences from UniProt.[34]

System Construction. The Silico scripts package v1.01[35] was used for the initial construction of lipid bilayers. PyMol v1.5.0.4[36] was used for protein sequence alignment. Maestro v9.3[37] was used for protein preparation, positioning of ligands, removing clashes, and otherwise editing the system.

The D_2 and D_3 receptor models were constructed based on the A-chain of the D_3R crystal structure[26] (PDB ID: 3PBL). The initial D_2R homology model was that developed by our group in previous work.[38] To create the D_3R model, the T4 lysozyme, water, and ligands were deleted from the PDB structure. Side-chains that were missing in intracellular loop 2 (ICL2) and extracellular loop 3 (ECL3) of the crystal structure were added using the Maestro protein preparation wizard workflow.[39−46] The mutation L119W on the external side of TM3 was reverted to wild-type. ICL3 was not included in either structure, and the ICL3 termini were instead capped with neutral groups (N-terminus acetyl, C-terminus *N*-methyl amide) or joined together (simulation **28**). Joining the two ICL3 termini was possible without perturbation of either TM5 or TM6, and a similar approach in the β_2 adrenergic receptor showed no resulting conformation changes.[47] In the D_2R, disulfide bonds were formed between C107−C182 and C399−C401. Disulfide bonds in D_3R were included between C103−C181 and C355−C358. Asp2.50 was protonated except in simulation **28**. All histidines were protonated on the delta nitrogen.

Each simulation system was constructed by embedding the receptor in a united-atom POPC bilayer[48] consisting of 40 lipids in each layer. The system was solvated with 4000 TIP3 water molecules, packed randomly outside the bilayer plane, no salt or counterions were included. The resulting system was not neutral, but due to the implementation of PME in NAMD2 packages there is no buildup of charge between periodic images.[49] The final dimensions of the system were 64 Å × 64 Å × 78 Å and the system contained approximately 21 900 atoms.

Each protein-bilayer system was allowed to relax over a total of 100 ns, according to the following protocol. At the start of the simulations, clozapine or haloperidol were placed by hand in various orientations at the entrance of the extracellular vestibule, taking care to avoid steric clashes with the protein. Water molecules within 2 Å of the ligand were deleted. From the onset of simulations, ligands were allowed to move freely; no biasing forces were used. The resulting equilibrated bilayers were used as the starting points for all other simulations.

Ligand Parametrization. Force field parameters for each ligand were developed through a combination of charge fitting for partial charges and by analogy to the CGenFF 2b6/7 force fields[50,51] for bonded interactions. Both ligands contain an amine that is protonated under biological conditions and is essential for binding; this protonation state was used for all following steps.

To assign atomic charges, a conformational search was first performed using MacroModel v9.9[52] to produce a representative set of low-energy conformations of each ligand. Gaussian 98 vA.7[53] (Hartree−Fock, 6-31G*) was used to calculate the electronic distribution for each conformation and RED-III.51[54] was then used to assign partial charges with the RESP-A1 charge derivation model,[55] generating point charges as an average across all conformations. ParamChem[56,57] was used to generate an initial set of bonded parameters by analogy.

The haloperidol model was built with both aryl-containing substituents positioned equatorially off the piperidine ring based on the findings of a study by Sikazwe et al.[58] which compared binding data for a series of ring-locked haloperidol analogs. Clozapine required several steps to parametrize due to its complex conformational behavior. The diazepine ring is nonplanar and undergoes a butterfly-like nitrogen inversion, and the rotation of the piperazine ring is also restricted due to conjugation with the diazepine ring. No suitable parameters were available so we derived them by fitting to QM energy profiles calculated using Jaguar v7.9[59] (DFT, B3LYP, 6-31G**). The high barrier to diazepine inversion (calculated as ~6 kcal/mol) prevented clozapine molecules from inverting during individual simulations, restricting sampling of their entire conformational space on the time scale of these simulations. Both inversions of the dibenzodiazepine moiety were therefore used as starting structures. Based on QM calculations and docking, the rotation of the piperazine ring was chosen such that the hydrogen of the protonated amine was axial on the convex side of the dibenzodiazepine moiety.

MD Protocol. MD simulations were performed with NAMD 2.7−2.9[60] using the CHARMM22[61,62] and CGenFF[50,51] v. 2b6/7 force fields. Simulations used 2 fs time step with full electrostatic interactions calculated every 6 fs. The particle mesh Ewald[63] method with a grid spacing of 1 Å was used to treat long-range electrostatics. Langevin dynamics was used as a thermostat to maintain the temperature at a constant 310 K with a damping coefficient of 5 ps^{-1}. The Nosé−Hoover Langevin piston barostat was used to maintain pressure at 1 atm, period 100 fs, decay 50 fs. The unit cell was treated semi-isotropically, maintaining the cell dimensions in the plane of the bilayer at a constant ratio. The SHAKE algorithm was used to constrain all bonds to hydrogen atoms. vdW forces were calculated using a switching algorithm between 10 and 12 Å. Structures were output every 10 ps. Each simulation was initiated with a 5000-step steepest descent minimization, followed by 10 ns of dynamics with the protein constrained, and a further 10 ns with only the protein backbone constrained. The final unconstrained MD simulations were run from this point onward for durations ranging from 200 to 1000 ns. As our goal was to discover the binding pathways, simulations in which the ligand pose was stable for 50−100 ns were terminated to redistribute resources to the remaining simulations.

Analysis. VMD v1.9.1[64] was used for the visualization and analysis of the system. Analysis procedures were performed within VMD using in-house scripts. All graphs were generated using Gnuplot. Residue−ligand contacts through the course of each simulation were calculated using VMD. Contact between the ligand and protein residues was defined as when heavy atoms in each were within 3.5 Å of each other. When constructing ligand-residue contact graphs, residues were considered to be interacting with the ligand if they were in contact with the ligand for >5% of simulation time. Clusters were generated for each trajectory with MSMBuilder 2.7[65] (LPRMSD module) using the hybrid *k*-centers/*k*-medoids method. One structure per one nanosecond was aligned by the backbone of the protein, the ligand position was then clustered by RMSD with a maximum intracluster distance of 5 Å and 10 iterations of *k*-medoids.

RESULTS AND DISCUSSION

We conducted a series of 29 unconstrained MD simulations of clozapine and haloperidol binding to the dopamine D_2 and D_3 receptors (Table 1) ranging in duration from 250 to 1145 ns, with an average length of 698 ns. Each simulation commenced with the drug placed in an arbitrary starting position in the extracellular vestibule of the receptor (Figure 2). Although this procedure omits the initial ligand contact step, the ligands still demonstrated an ability to explore the extracellular vestibule and the occasional disassociation/reassociation event was observed. Most simulations produced only partial binding trajectories, however the ligands in two simulations, **28** (clozapine) and **29** (haloperidol), proceeded to a stable bound pose, deep in the orthosteric binding site. Table 1 lists all of the simulations performed and classifies them based on the pose observed in the final frame of the simulation: Group 1 proceeded only as far as the initial receptor recognition event. Group 2 bound in a secondary binding pocket, making interactions with Glu2.65. Group 3 ligands proceeded further into the receptor with their protonated amines oriented toward Asp3.32. Group 4 formed a salt-bridge to Asp3.32. Group 5 followed a full binding pathway making a salt-bridge to Asp3.32 and remained in a stable bound pose. Taken as a set, groups 1, 3, 4, and 5 reveal a linear sequence of events starting in the vestibule and ending deep within the receptor, constituting a binding pathway. In this progression, ligands occasionally stepped backward through this pathway and only stages which were stable at the end of the simulation have been tabulated.

Table 1. Simulations Categorized by the Ligand Position Reached at the End of the Simulation

No.	ligand	receptor	duration (ns)
Group 1. Ligands Making Initial Interactions with Tyr7.35			
1	clozapine	D2	294
2	clozapine	D2	500
3	clozapine	D2	500
4	clozapine	D3	500
5	haloperidol	D2	896
6	haloperidol	D2	916
7	haloperidol	D3	898
8	haloperidol	D3	855
Group 2. Ligands Interacting with Glu2.65 in the Secondary Binding Pocket			
9	clozapine	D2	246
10	clozapine	D2	970
11	clozapine	D3	500
12	clozapine	D3	400
13	haloperidol	D2	907
14	haloperidol	D2	988
Group 3. Protonated Amine Orienting Deeper into the Receptor			
15	clozapine	D2	250
16	clozapine	D3	500
17	clozapine	D3	500
18	haloperidol	D2	974
19	haloperidol	D2	1089
20	haloperidol	D2	1145
21	haloperidol	D2	1067
22	haloperidol	D3	250
23	haloperidol	D3	606
Group 4. Ligands Forming Salt-Bridge with Asp3.32			
24	clozapine	D2	994
25	clozapine	D2	400
26	clozapine	D3	500
27	haloperidol	D2	1067
Group 5. Ligands Adopting the Final Bound Pose			
28	clozapine	D3	1002
29	haloperidol	D3	516

Full Binding Simulations. In two simulations, one of clozapine (**28**) and one of haloperidol (**29**), both binding to the D_3R, the ligand proceeded from initial contact to a bound pose in the orthosteric site. The progression of these simulations is illustrated in Figures 3 and 4 and in the Supporting Information. The binding of haloperidol to the orthosteric site of the D_3R observed in simulation **29** is detailed in Figure 3. Panel A shows the distances between the protonated nitrogen of the ligand and residues Asp3.32 and Tyr7.35. Panel B shows the interactions between the ligand and the receptor residues at each nanosecond of simulation. Panels C−F show snapshots of the binding pathway at key points enroute to the orthosteric binding site. At the beginning of the simulation, the A-ring of haloperidol π-stacks with Tyr7.35 while the bulk of the ligand projects into the solvent. At 40 ns, haloperidol briefly dissociates completely from the receptor, tumbling around in the solvent before reassociating with the receptor and re-establishing the π-stacking interaction between the A-ring and Tyr7.35 (Figure 3C). From here, haloperidol draws closer to Asp3.32 in a series of steps, most easily visible in Figure 3A. First, the A-ring slides into the polar region around TM5,6 and ECL2 (Figure 3D, 65 ns) before shifting further, at 95 ns, to

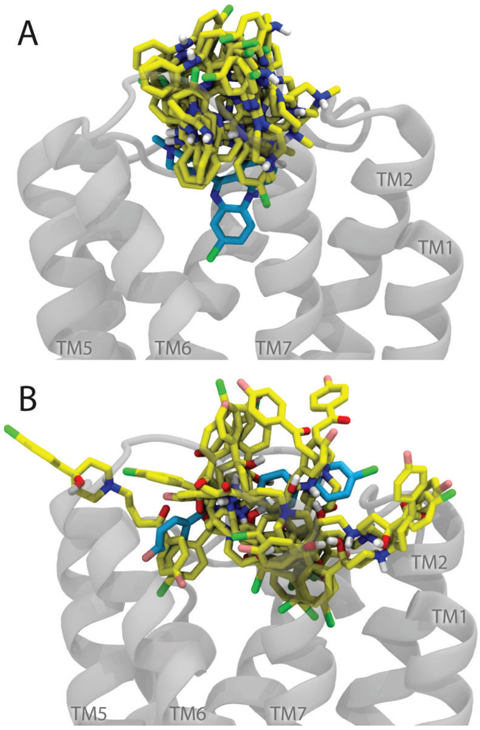


Figure 2. Overlayed starting positions of the ligand in each simulation for clozapine (A) and haloperidol (B). Starting positions for full binding simulations **28** and **29** are shown in blue.

intercalate between TM5,6 (Figure 3E), allowing the protonated amine to draw closer to Asp3.32. The salt-bridge finally snaps together at 115 ns, drawing the A-ring back into the receptor where it drops down into the orthosteric site adjacent to Trp6.48. This bound pose remains stable for the final 400 ns of simulation (Figure 3F). The final pose makes few polar interactions, aside from the salt-bridge to Asp3.32. The carbonyl and hydroxyl groups as well as the B-ring remain solvent accessible, while the buried A-ring sits alongside Trp6.48 with the fluorine located in a hydrophobic pocket and pointed toward Ile3.40. This binding process can be viewed in Supporting Information movie 1.

The final stable pose of haloperidol occupies the same space as predicted in a number of docking studies[29−31] but the molecular orientation is rotated by 180°. In our model, the butyrophenone moiety is buried most deeply in the receptor, leaving the piperazine moiety and B-ring in the extracellular vestibule, similar to the pose observed in a QM optimized study.[32] Although this result disagrees with most docking studies, the pose is well supported by experimental evidence. Barton et al.[66] performed a study in which they attached a number of bulky fluorescent tags to the compound *N*-*p*-aminophenethylspiperone (containing the same butyrophenone fragment as haloperidol. Refer to spiperone, Figure 1). Upon measuring the compound binding at the D_2R they found that the attachment of bulky fluorophores resulted in—at most—only a 10-fold reduction in affinity, suggesting that the butyrophenone moiety must be the end that binds deep into the binding pocket. Supporting the same hypothesis but from a different approach, Vangveravong et al.[67] performed a study in which they replaced the orthosteric fragment of aripiprazole with the piperazine half of haloperidol and found a > 200 fold reduction in affinity, again suggesting that the butyrophenone binds more deeply in the receptor.

The binding of clozapine to the orthosteric site of the D_3R observed in simulation **28** is detailed in Figure 4, in the same manner as Figure 3 does for haloperidol. In this simulation, clozapine first becomes wedged between Tyr7.35 and ECL2. The protonated amine then rapidly orients downward into the receptor at 40 ns (Figure 4C). From here clozapine dips down to form a tentative salt-bridge with Asp3.32 several times (best seen in Figure 4A) until, at 240 ns, the ligand plunges to the bottom of the orthosteric site (Figure 4D) before bobbing back up at 265 ns to form the salt-bridge with Asp3.32 (Figure 4E). During this event, clozapine rotates toward TM5,6 with the tricyclic system tucking under ECL2. At 315 ns, clozapine rotates further around the axis of the newly formed salt-bridge with the tricyclic system dropping deeper into the orthosteric site and then deeper again at 335 ns to reach a stable pose (Figure 4F) which is maintained for the remaining 650 ns of the simulation. Similar to haloperidol, the final pose includes few direct polar contacts aside from the salt-bridge to Asp3.32. The diazepine N−H forms a hydrogen bond with a trapped water molecule at the bottom of the orthosteric site and the B-ring interacts with Trp6.48. This binding process can be viewed in Supporting Information movie 2.

The MD bound pose of clozapine overlaid on the crystal structures of doxepin bound to the histamine H_1 receptor[68] and carazolol bound to the β_2-adrenergic receptor[69] is shown in Figure 5A. Both compounds are tricyclics with a general structural resemblance to clozapine, although doxepin has an arched shape, similar to clozapine whereas carazolol is planar. In the MD-generated pose of clozapine, the chlorine and sp^2 nitrogen remain solvent accessible in the extracellular vestibule while the tricyclic system sits deep in the orthosteric site adjacent to Trp6.48, similarly positioned to doxepin. Prior to clozapine binding in simulation **28**, Trp6.48 rotated to its "downward" rotamer and consequently was pinned in this position by the ligand (Figure 5B). While the rotation of Trp6.48 is important to the activation of many GPCRs, its rotameric state has not been found to correlate to the functional state of the receptor.[70] In contrast to reported docked structures,[28] the diazepine N−H in the MD structure does not make hydrogen bonds to Ser5.42, Ser5.43, or Ser5.46. Instead it angles deeper into the receptor, allowing the clozapine and haloperidol pharmacophores to overlap (Figure5B), and makes a hydrogen bond to a water molecule that is trapped below the ligand and is unable to exchange with the bulk solvent. To investigate the effect of the rotation of Trp6.48 on ligand binding, an additional simulation was performed beginning from a docked pose of clozapine in the D_3R crystal structure. The docked pose was similar to clozapine in Figure 4E but with the tryptophan switch still in the crystallographic rotamer. In this simulation, the clozapine dropped deeper into the receptor as soon as backbone constraints were released, matching the behavior of clozapine binding in simulation **28**. The resulting pose was angled closer to TM3 (RMSD 1.37 Å), allowing a better overlap with the haloperidol pharmacophore and there was no trapped water molecule. It should also be noted that the conformational space of the region surrounding the most deeply buried rings of clozapine and haloperidol was most likely more rigorously explored in the current MD simulations than in the reported docking studies, resulting in extension of the binding cavity.

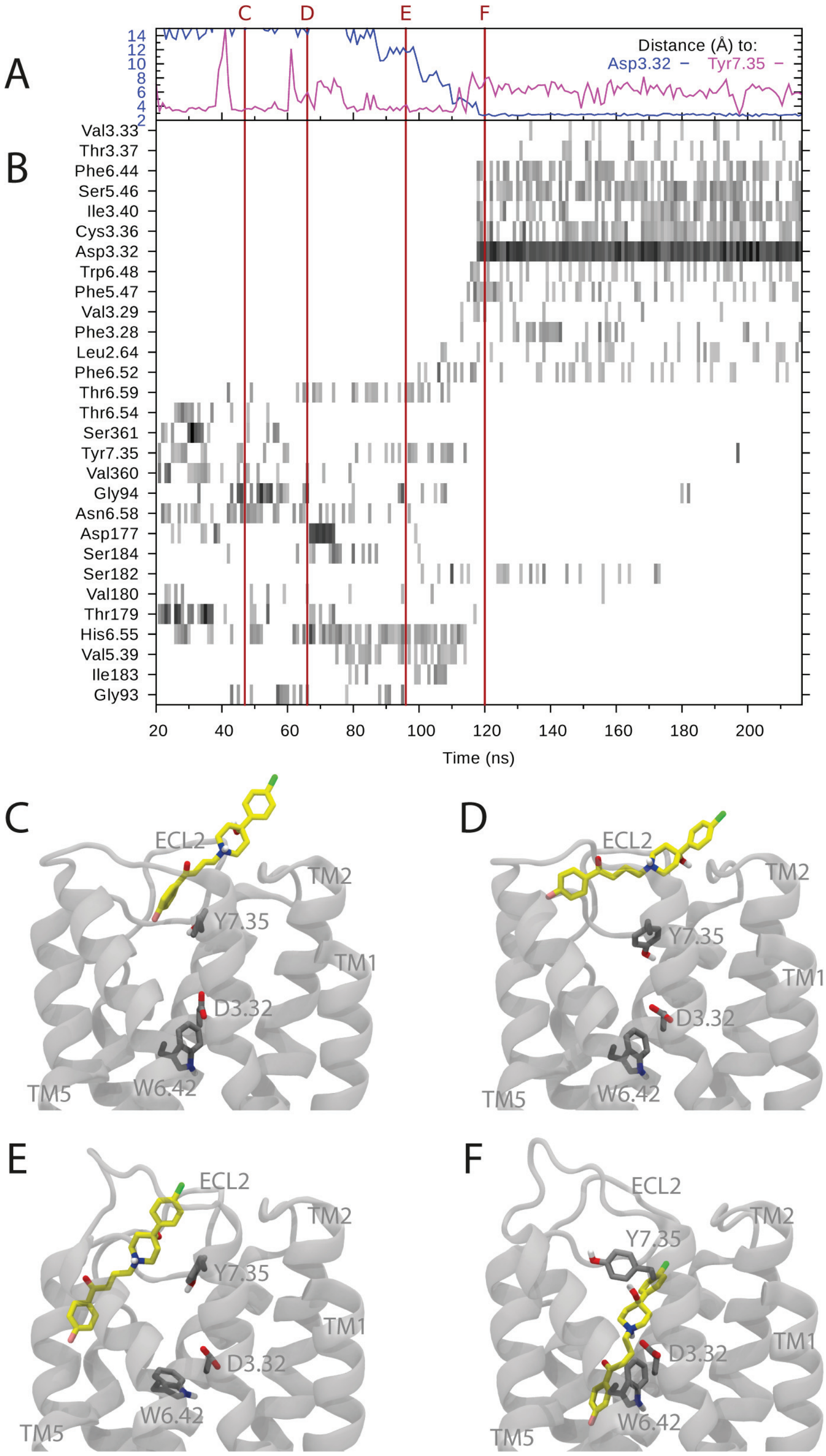


Figure 3. Graphical representations of the binding pathway of haloperidol in simulation **29**. (A) Distance between the protonated nitrogen of haloperidol and Asp3.32 or Tyr7.35. (B) Barcode graph showing which residues are interacting with the ligand. The time-points of simulation snapshots C−F are indicated by red lines. (C−F) Snapshots of the binding pathway of haloperidol showing (C) the initial recognition event involving Tyr7.35, (D) interactions with the polar region surrounding ECL2, (E) intercalation of the haloperidol A-ring into the aromatic network between TM5,6, and (F) the final bound pose.

Interactions Analysis. Although the complete binding simulations **28** and **29** provide insight into entire binding pathways, these pathways are not necessarily the dominant ones, and there are still potentially many other binding pathways that remain unexplored. By looking at ligand−receptor interactions over the entire set of simulations, the trends in behavior for

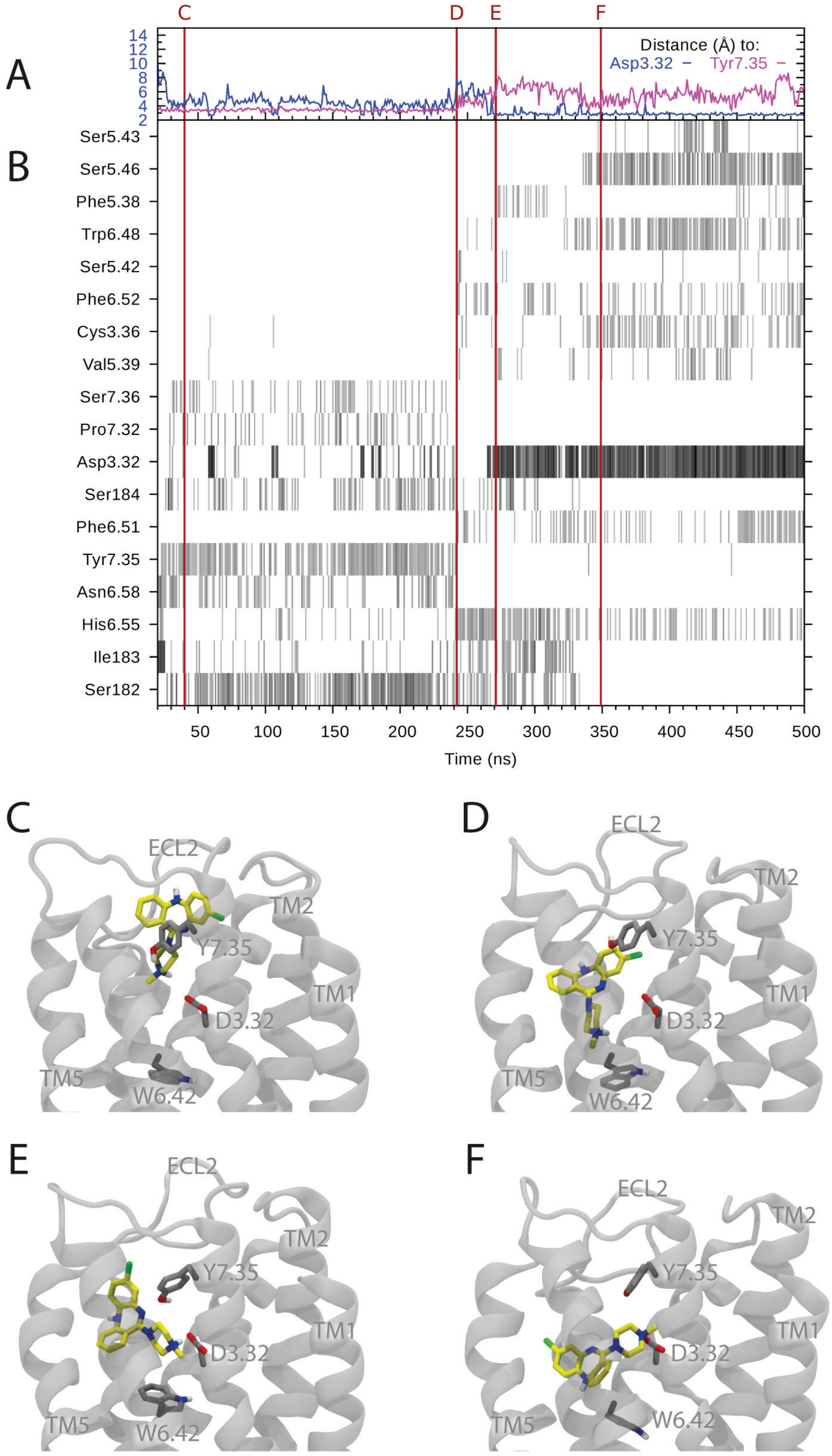


Figure 4. Graphical representations of the binding pathway of clozapine in simulation **28**. (A) Distance between the protonated nitrogen and Asp3.32 or Tyr7.35. (B) Barcode graph showing which residues are interacting with the ligand over the course of the simulation. The time-points of simulation snapshots C−F are indicated by red lines. (C−F) Snapshots of the binding pathway of clozapine showing (C) the initial interaction of the clozapine protonated amine with Asp3.32, (D) the pose following a plunge into the orthosteric site, (E) formation the salt-bridge with Asp3.32, and (F) the final pose following ligand rotation around the salt-bridge axis.

each ligand or each receptor can be characterized, and the residues most involved in metastable states or critical mechanisms can be identified.

To determine the most frequent ligand−receptor interactions present across the entire series of simulations, we conducted a residue interaction analysis using heavy-atom

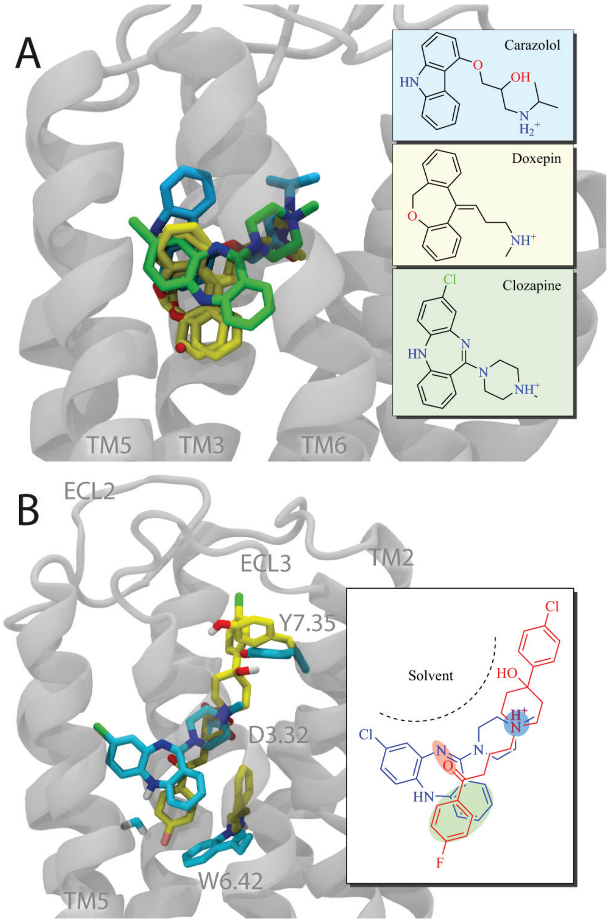


Figure 5. (A) Crystal structure bound pose of carazolol from a complex with the β_2-adrenergic receptor (blue), the two poses of doxepin resolved in the histamine H_1 receptor (yellow), and our MD bound pose of clozapine (green) superimposed onto the D_3R crystal structure. (B) Overlay of bound poses for the clozapine (blue) and haloperidol (yellow) complexes formed in MD simulations **28** and **29**, showing the well conserved pharmacophore. Tyr7.35, Trp6.48, and a trapped water molecule are also shown, colored according to the complex to which they belong. (inset) 2D representation of the overlap of pharmacophore elements, protonated amine (blue), aromatic rings (green), and electronegative elements (red).

proximity (Table 2). Contact between the ligand and protein residues was defined as when heavy atoms in each were within 3.5 Å. In addition to having large standard deviations, these interaction frequencies are weighted toward early stages in the binding pathway and, without further investigation, should be considered qualitatively.

The most prevalent protein−ligand interaction observed in both receptors was with Tyr7.35, which was observed to interact with the ligands for 30−50% of all simulation time (Table 2), often by π-stacking. After initially π-stacking with the ligands, Tyr7.35 then acts as a pivot point from which the ligands can explore the extracellular vestibule (Figure 6) with the protonated amine extending toward the nearby polar residues present in TM1,2,7 or TM5,6 and ECL2. The interaction with Tyr7.35 typically only ceased once formation of a salt-bridge between the ligand protonated amine and Asp3.32 drew the ligand out of range of Tyr7.35, a process that we denote as a "handover" mechanism. Once the salt-bridge with Asp3.32 has formed, the ligand could then rotate to explore deeper in the receptor. This mechanism was observed in the binding of both clozapine and haloperidol, and can be clearly seen in the swap-over of distances between the protonated nitrogen and Asp3.32 or Tyr7.35 shown in Figures 3A and 4A.

Table 2. Interaction Frequency for Simulations of Each Complex[a]

Most Frequent Interactions Across All **Clozapine** Simulations			
D_2	av interaction freq	D_3	av interaction freq
Tyr7.35	0.41 ± 0.19	Tyr7.35	0.47 ± 0.17
Asn6.58	0.37 ± 0.17	Asp3.32	0.30 ± 0.29
Ile403ECL2	0.24 ± 0.10	Gly94ECL1	0.24 ± 0.22
Glu2.65	0.20 ± 0.26	Ser182ECL2	0.22 ± 0.13
Ile183ECL2	0.18 ± 0.09	Thr7.39	0.17 ± 0.13
Ser7.36	0.16 ± 0.17	**His6.55**	0.16 ± 0.08
Ile184ECL2	0.15 ± 0.14	Val2.66	0.11 ± 0.10
Asp3.32	0.15 ± 0.21	Phe6.51	0.11 ± 0.14
His6.55	0.13 ± 0.15	**Glu2.65**	0.09 ± 0.12
Thr7.39	0.11 ± 0.06	Ser184ECL2	0.08 ± 0.09
Most Frequent Interactions Across All **Haloperidol** Simulations			
D_2	av interaction freq	D_3	av interaction freq
Tyr7.35	0.49 ± 0.14	**His6.55**	0.44 ± 0.30
Ile184ECL2	0.31 ± 0.22	Tyr7.35	0.33 ± 0.18
Asn6.58	0.31 ± 0.15	Ser182ECL2	0.28 ± 0.23
Phe3.28	0.25 ± 0.23	Ser184ECL2	0.20 ± 0.11
Ser7.36	0.24 ± 0.15	**Ile183ECL2**	0.18 ± 0.14
Ile183ECL2	0.22 ± 0.14	Gly94ECL1	0.16 ± 0.13
Glu2.65	0.22 ± 0.23	Asp3.32	0.15 ± 0.27
Thr7.39	0.21 ± 0.08	**Asn6.58**	0.15 ± 0.13
His6.55	0.18 ± 0.15	Ser7.36	0.15 ± 0.13
Val2.66	0.16 ± 0.07	Thr7.39	0.14 ± 0.08

[a] Values are the frequency of simulated interactions ± one standard deviation. The residues Asn6.58, His6.55, and Glu2.65, which are discussed in the text, have been highlighted in bold.

We propose that the handover mechanism is dependent on the structure of the ligand, particularly on the intramolecular distance between the A-ring in each ligand and the protonated amine (clozapine 7.76 Å, haloperidol 7.89 Å). These distances are well conserved in the pharmacophores of antipsychotic compounds.[71,72] The simulations presented here suggest that maintaining a specific distance between the A-ring and protonated-amine moieties is not only required for the ligand to bind to the orthosteric site (Figure 5) but is also important in the handover mechanism that leads to orthosteric binding. Thus, we would expect ligand interaction with Tyr7.35 to contribute to the on-rate of antipsychotic compounds matching this pharmacophore.

The region surrounding ECL2 and TM5,6 was also observed to be a hotspot for ligand interactions. Both ligands in both receptors made contacts in the polar residue-rich region surrounding His6.55, Asn6.58, and residues at the end of ECL2. In D_3R simulations, interactions were most prominently observed with the partially buried His6.55, while in D_2R simulations contact with the shallower Asn6.58 predominated (Table 2). This was likely due to the pronounced difference in polarity of the ECL2 motifs surrounding this region (D_3, Ser-Ile-Ser; D_2, Ile-Ile-Ala). Ligands in the D_2 receptor also displayed a greater tendency to make interactions with Glu2.65 at the opposite end of the extracellular vestibule (Table 2, bold). The differences in interaction frequency between the D_2 and D_3 receptors suggest that the environment outside the orthosteric site in each

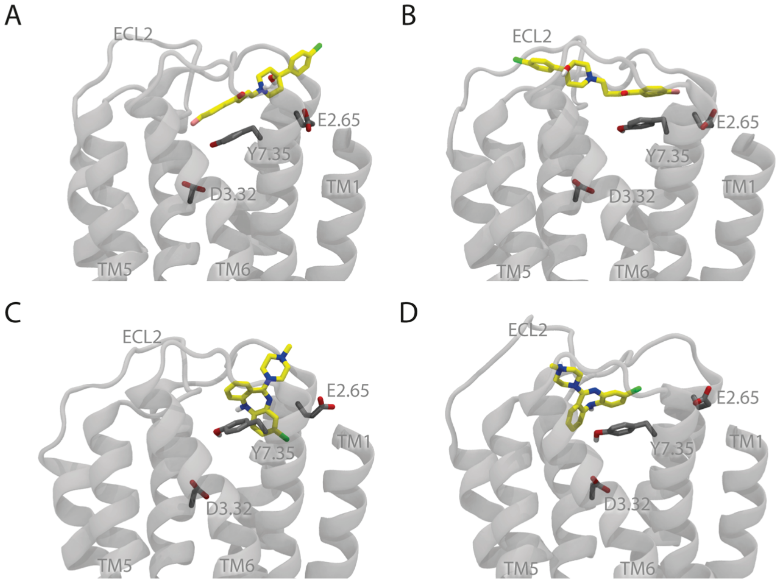


Figure 6. Haloperidol (simulation **5**) exploring the extracellular vestibule from the Tyr7.35 pivot point: (A) Oriented toward the secondary binding pocket. (B) Oriented toward ECL2. The same can be seen with clozapine (simulation **3**) oriented toward (C) the secondary binding pocket or (D) toward ECL2.

receptor can affect the binding pathway, even of nonselective ligands. In this case, the observed effect is weak but it does suggest that modifications to the ligands could direct the ligands toward either end of the extracellular vestibule.

The ligands in several simulations (Group 2, Table 1) were observed to interact with a secondary binding pocket located between TM1,2,7 where they formed polar interactions with Glu2.65 (Figure 7). All clozapine poses in this region (simulations **9**, **10** in the D_2R; **11**, **12** in the D_3R) formed a salt-bridge to Glu2.65 and were largely stable, due to the ligand nestling in the extracellular loops. Haloperidol (simulation **13**,**14** D_2R) occupied a much larger variety of poses that were generally less stable than those observed for clozapine. In the most stable haloperidol poses, the B-ring was typically buried in the secondary pocket while either the amine or hydroxyl group interacted with Glu2.65, disrupting the native hydrogen bonding to Ser7.36. Clozapine did not disrupt the hydrogen bonding between Ser7.36 and Glu2.65. This pocket between TM1,2,7 has been previously identified as the location at which bitopic ligands interact,[26] making it a focal point for the design of allosteric or bitopic ligands.[73] Although clozapine and haloperidol are neither bitopic nor allosteric, the simulations suggest that they both bind transiently to this site. Clozapine and haloperidol scaffolds differently affect the hydrogen-bonding networks of Ser7.36 and Glu2.65 within this pocket and potentially represent starting points for development of a pharmacophore for this location. Longer simulations would need to be performed to determine whether the disruption of the Glu2.65-Ser7.36 hydrogen bond propagates any effect to the intracellular side of the receptor.

Having traversed the outer portion of the binding pocket (Groups 3 and 4, Table 1), both clozapine and haloperidol initially form a salt-bridge to Asp3.32 and bind deeper in the orthosteric pocket. The observed pathway for these processes differed between the two ligands. Haloperidol was able to draw close enough to form the salt-bridge by inserting either of its own aromatic rings into the aromatic networks between TM2,3, TM4,5 or TM5,6 (Figure 8A,B). In contrast, clozapine, being a much shorter ligand, was able to form the salt-bridge by simply dropping deeper into the receptor from a central position adjacent to Tyr7.35. Once the salt-bridge was formed, clozapine either remained in place or rotated toward TM1,2,7 (Figure 8C,D). Unlike the complete clozapine binding in simulation **28**, these events did not involve clozapine plunging to the bottom of the orthosteric site prior to formation of the salt-bridge. The ability of clozapine to drop deeper into the

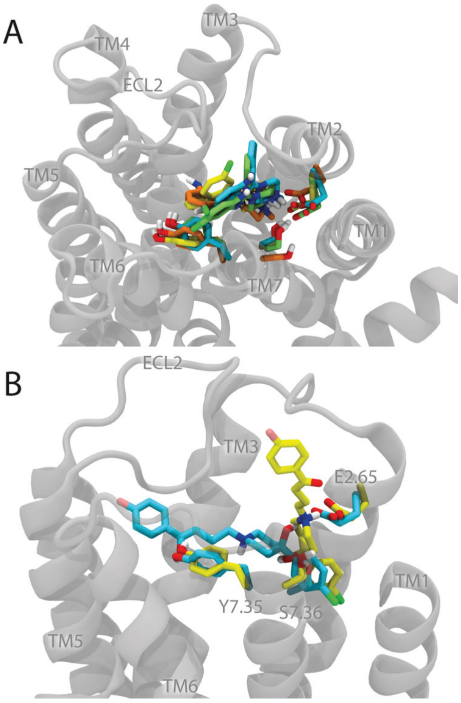


Figure 7. Poses of clozapine (A) and haloperidol (B) in the secondary binding pocket formed by TM1,2 and 7. The side chains of Glu 2.65, Ser 7.36, and Tyr 7.35 are shown.

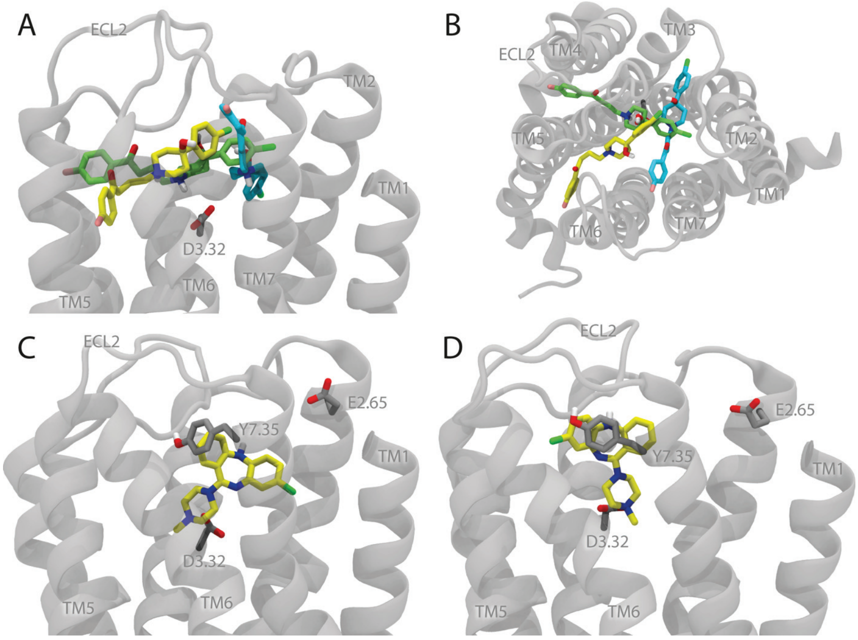


Figure 8. Positions of clozapine and haloperidol at the point of salt-bridge formation. (A) Side and (B) top view of haloperidol intercalating helices to close the distance to the salt-bridge in simulation **19** (blue), **27** (green), and **29** (yellow). (C) Clozapine, having formed the salt-bridge, leaning toward TM1,2,7 (simulation **24**) and (D) remaining around TM5,6 (simulation **25**).

receptor did not appear to be affected by the inversion of the dibenzodiazepine moiety.

Cluster Analysis of Ligand Binding. Ligand binding can be rationalized as a series of metastable states and, as such, these states are the key descriptors of the binding pathway. Metastable states have been seen to coincide with known allosteric sites,[16] and they are the most likely locations from which to influence the kinetics of drug binding. In an effort to identify potential metastable sites within the ligand binding pathway, we performed a cluster analysis of the complete set of 29 simulations. The most extensive sampling occurred in the extracellular vestibule region, and thus, the observed metastable states are weighted toward this area.

Figure 9 shows representative molecules from the five most populated clusters in each ligand−receptor combination. In clozapine simulations, we generally observed clusters where the aromatic system of the ligand is inserted between the extracellular loops. Formation of salt-bridges to Glu2.65 and Asp3.32 also led to stable poses. The D_2R-clozapine simulations (Figure 9A) produced two clusters with the ligand protonated amine located in a pocket between TM1,2,7 (orange, pink) but only one (pink) forms a salt-bridge with Glu2.65. In the other cluster (orange), the dibenzodiazepine moiety is inverted from the previous cluster and the protonated amine points away from Glu2.65, this suggests that the tight fit of the loop region makes a significant contribution to the stability of this pose. Two more clusters of clozapine in the D_2R show stable positions of the ligand in the vicinity of ECL2 and Tyr7.35 before (yellow) and after (green) alignment of the protonated amine toward Asp3.32. The final cluster (blue) shows the initial formation of the Asp3.32 salt-bridge, in this case with the clozapine ring system leaning toward the TM1,2,7 binding pocket. The clusters arising from the D_3R-clozapine simulations (Figure 9B) follow a similar pattern to the D_2R clusters, two of them (green, pink) showing similar positions of clozapine in the TM1,2,7 pocket while another (blue) shows the initial formation of the Asp3.32 salt-bridge. A fourth cluster (orange) represents a pose in which the tricyclic system of clozapine is wedged between ECL1−2 and is unable to rotate to close the distance to Glu2.65. The final (yellow) cluster shows the final bound pose.

Clustering of the haloperidol simulations revealed large differences in ligand behavior between the D_2 and D_3 receptors. In the D_2R-simulations, ligands generally favored orienting toward TM2,3 while the D_3R-simulation ligands favored orienting toward TM5,6. Four of the five most populated D_2R-haloperidol clusters (Figure 9C and E; yellow, green, pink, orange) resulted from insertion of the haloperidol aromatic rings A and B between TM2,3 with the remainder of the molecule extending toward ECL3. In the remaining cluster (blue), the B-ring is buried in the TM1,2,7 pocket while the hydroxy group interacts with Glu2.65. In the D_3R-haloperidol clustering (Figure 9D and F), one end of the ligand was always oriented toward TM5,6 and, similar to the clusters found in the D_2R, there was no preference for either aromatic ring. Four clusters (yellow, blue, green, orange) show haloperidol binding shallowly to the receptor in the polar region surrounding ECL2. The remaining cluster (pink) shows the deepest binding haloperidol, with the A-ring tucked under ECL1 allowing the initial formation of the Asp3.32 salt-bridge. These metastable states might be useful to consider when designing any ligand expected to interact in the extracellular vestibule, whether extending an orthosteric ligand into this region or developing an allosteric one. Due to the differences in the metastable binding sites for haloperidol binding to each receptor, these sites would be particularly interesting for the design of selective butyrophenone-like compounds.

■ CONCLUSION

In this work, we present a large set of extensive unbiased MD simulations that model the binding of the antipsychotic drugs

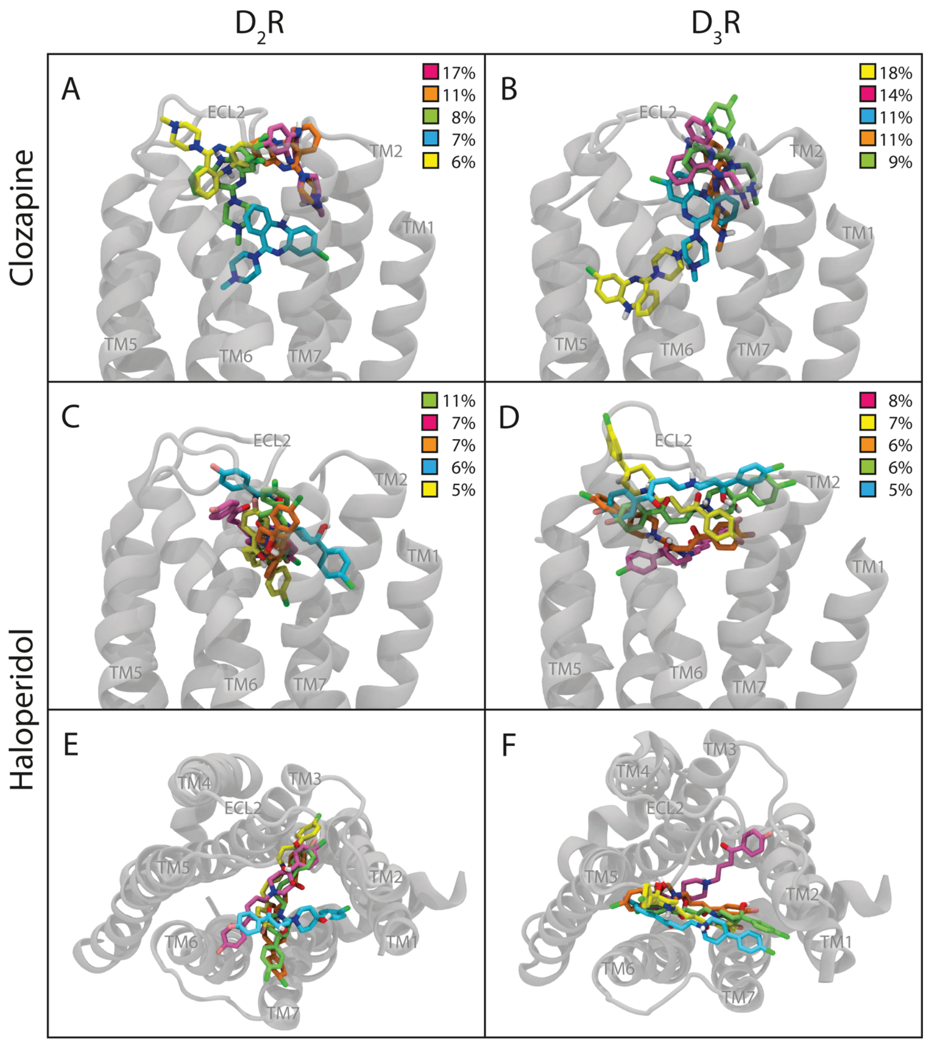


Figure 9. Cluster analysis of each receptor−ligand combination showing representative structures from each of the five most populated clusters for each ligand−receptor combination (A−D). Inset values are the percentage of the total population attributed to each cluster. Additional "top-down" views are shown for haloperidol (E and F). These clusters indicate the location of metastable states.

haloperidol and clozapine to the D_2 and D_3 dopamine receptors. Two of the simulations follow a complete binding trajectory from the extracellular vestibule to a pose buried deep within the receptor, predicting the ultimate binding pose of each ligand. These are the first simulations to show an unbiased binding pathway for clozapine or haloperidol. Although most of the simulations do not produce complete binding trajectories, they capture ligand binding events in the early stages of binding while the ligand is within the extracellular vestibule.

As a set, the simulations reveal common processes that occur in the ligand binding pathways of both clozapine and haloperidol to dopamine receptors. In all simulations, Tyr7.35 was observed to play a central role in the early phases of ligand binding. Furthermore, we observed a common handover mechanism which transferred the ligand from Tyr7.35 to form the key salt-bridge to Asp3.32. This handover mechanism takes advantage of the amine-to-aromatic distance in the well-established pharmacophore of antipsychotic compounds.

Although the D_2 and D_3 receptors are not selective for either clozapine or haloperidol, we observed differences in the binding pathways of ligands in these receptors, particularly for haloperidol. In the early stages of binding, there were only minor differences in ligand behavior between receptors with both ligands preferring to be in the vicinity of ECL2 in the D_3R and toward TM1,2,7 in the D_2R. More pronounced differences were observed once the ligands proceeded to salt-bridge formation, with haloperidol occupying two dominant metastable binding sites in the extracellular vestibule. The first site was with either aromatic ring inserted between TM2,3 (most prevalent in the D_2R). The second state involved one of the aromatic rings of haloperidol positioned near TM5,6 while the remainder of the ligand interacted with the residues around ECL2 (most prevalent in the D_3R).

A cluster analysis of the entire simulation data set revealed a number of metastable states within the binding process. Both haloperidol and clozapine occupied a secondary binding pocket between TM1,2,7, making interactions with Glu2.65, an area implicated in the binding of bitopic and allosteric ligands.[73] A metastable state in the binding of clozapine was also found on the initial formation of the salt-bridge, where the tricyclic system is still making interactions in the extracellular vestibule and prior to rotation of the ligand and deeper binding. This state was observed similarly in both D_2 and D_3 receptors.

This work is an initial study of the binding pathways of these ligands. As such, there is still much to investigate, particularly with regard to the later stages of the binding pathway that occur after salt-bridge formation. The analysis of metastable states in particular would benefit from a more quantitative approach in order to properly evaluate their usefulness as targets for structure-based drug design. We are currently pursuing further investigations in these areas.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jcim.5b00457.

Movie of the binding of haloperidol to the orthosteric site of D_3R (PDF)

Movie of the binding of clozapine to the orthosteric site of D_3R (PDF)

PDB format files containing the structures used to make Figures 3–9 (ZIP)

■ AUTHOR INFORMATION

Corresponding Authors

[redacted]

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We acknowledge computational resources and technical support provided by the Victorian Life Sciences Computational Initiative (VLSCI) through grant VR0133 and the National Computational Infrastructure (NCI), and Multimodal Australian ScienceS Imaging and Visualisation Environment (MASSIVE) through grant y96. T.T. is a recipient of an Australian Postgraduate Award (APA) scholarship. Y.F. was a recipient of a UROP scholarship supported by BIO21. We thank Ben Capuano and Rob Lane for helpful discussions. Finally, thanks very much to Ted Lloyd who got us interested in clozapine in the first place.[71] PDB format files containing the structures used to make Figures 3-9 are provide as a zip file.

■ REFERENCES

(1) Pan, A. C.; Borhani, D. W.; Dror, R. O.; Shaw, D. E. Molecular Determinants of Drug-Receptor Binding Kinetics. *Drug Discovery Today* **2013**, *18*, 667–673.

(2) 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. U. S. A.* **2011**, *108*, 13118–13123.

(3) Chin, S. P.; Buckle, M. J.; Chalmers, D. K.; Yuriev, E.; Doughty, S. W. Toward Activated Homology Models of the Human M1Muscarinic Acetylcholine Receptor. *J. Mol. Graphics Modell.* **2014**, *49*, 91–98.

(4) Manallack, D. T.; Chalmers, D. K.; Yuriev, E. Using the B 2 -Adrenoceptor for Structure-Based Drug Design. *J. Chem. Educ.* **2010**, *87*, 625–627.

(5) McRobb, F. M.; Capuano, B.; Crosby, I. T.; Chalmers, D. K.; Yuriev, E. Homology Modeling and Docking Evaluation of Aminergic G Protein-Coupled Receptors. *J. Chem. Inf. Model.* **2010**, *50*, 626–637.

(6) Thomas, T.; McLean, K. C.; McRobb, F. M.; Manallack, D. T.; Chalmers, D. K.; Yuriev, E. Homology Modeling of Human Muscarinic Acetylcholine Receptors. *J. Chem. Inf. Model.* **2014**, *54*, 243–253.

(7) Thomas, T.; Chalmers, D. K.; Yuriev, E. Homology Modeling and Docking Evaluation of Human Muscarinic Acetylcholine Receptors. In *Muscarinic Receptor: From Structure to Animal Models*; Myslivecek, J., Jakubik, J., Eds.; Neuromethods; Springer: New York, NY, 2016; Vol. *107*, pp 15–35.

(8) Enkavi, G.; Tajkhorshid, E. Simulation of Spontaneous Substrate Binding Revealing the Binding Pathway and Mechanism and Initial Conformational Response of GlpT. *Biochemistry* **2010**, *49*, 1105–1114.

(9) Bello, M.; García-Hernández, E. Ligand Entry into the Calyx of B-Lactoglobulin. *Biopolymers* **2014**, *101*, 744–757.

(10) Hurst, D. P.; Grossfield, A.; Lynch, D. L.; Feller, S.; Romo, T. D.; Gawrisch, K.; Pitman, M. C.; Reggio, P. H. A Lipid Pathway for Ligand Binding Is Necessary for a Cannabinoid G Protein-Coupled Receptor. *J. Biol. Chem.* **2010**, *285*, 17954–17964.

(11) Preininger, A. M.; Meiler, J.; Hamm, H. E. Conformational Flexibility and Structural Dynamics in GPCR-Mediated G Protein Activation: A Perspective. *J. Mol. Biol.* **2013**, *425*, 2288–2298.

(12) *G Protein-Coupled Receptors—Modeling and Simulation*; Filizola, M., Ed.; Advances in Experimental Medicine and Biology; Springer: Netherlands Dordrecht, 2014.

(13) 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.; Thian, F. S.; Kobilka, T. S.; Shaw, D. E.; Mueller, L.; Prosser, R. S.; Kobilka, B. K. The Dynamic Process of *β*2-Adrenergic Receptor Activation. *Cell* **2013**, *152*, 532–542.

(14) Alonso, H.; Bliznyuk, A. A.; Gready, J. E. Combining Docking and Molecular Dynamic Simulations in Drug Design. *Med. Res. Rev.* **2006**, *26*, 531–568.

(15) Sinko, W.; Lindert, S.; Mccammon, J. A. Accounting for Receptor Flexibility and Enhanced Sampling Methods in Computer-Aided Drug Design. *Chem. Biol. Drug Des.* **2013**, *81*, 41–49.

(16) Kruse, A. C.; Hu, J.; Pan, A. C.; Arlow, D. H.; Rosenbaum, D. M.; Rosemond, E.; Green, H. F.; Liu, T.; Chae, P. S.; Dror, R. O.; Shaw, D. E.; Weis, W. I.; Wess, J.; Kobilka, B. K. Structure and Dynamics of the M3Muscarinic Acetylcholine Receptor. *Nature* **2012**, *482*, 552–556.

(17) 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.; Baell, J. B.; Sexton, P. M.; Christopoulos, A.; Shaw, D. E. Structural Basis for Modulation of a G-Protein-Coupled Receptor by Allosteric Drugs. *Nature* **2013**, *503*, 295–299.

(18) Guo, D.; Hillger, J. M.; Ijzerman, A. P.; Heitman, L. H. Drug-Target Residence Time-A Case for G Protein-Coupled Receptors. *Med. Res. Rev.* **2014**, *34*, 856–892.

(19) Andersson, K.; Hämäläinen, M. D. Replacing Affinity with Binding Kinetics in QSAR Studies Resolves Otherwise Confounded Effects. *J. Chemom.* **2006**, *20*, 370–375.

(20) Keighley, W. The Need for High Throughput Kinetics Early in the Drug Discovery Process. *Drug Discovery World* **2011**, *12*, 39–45.

(21) Vauquelin, G.; Bostoen, S.; Vanderheyden, P.; Seeman, P. Clozapine, Atypical Antipsychotics, and the Benefits of Fast-off D2 Dopamine Receptor Antagonism. *Naunyn-Schmiedeberg's Arch. Pharmacol.* **2012**, *385*, 337–372.

(22) Ohlson, S. Designing Transient Binding Drugs: A New Concept for Drug Discovery. *Drug Discovery Today* **2008**, *13*, 433–439.

(23) Jupp, S.; Malone, J.; Bolleman, J.; Brandizi, M.; Davies, M.; Garcia, L.; Gaulton, A.; Gehant, S.; Laibe, C.; Redaschi, N.; Wimalaratne, S. M.; Martin, M.; Le Novère, N.; Parkinson, H.; Birney, E.; Jenkinson, A. M. The EBI RDF Platform: Linked Open Data for the Life Sciences. *Bioinformatics* **2014**, *30*, 1338–1339.

(24) Davies, M.; Nowotka, M.; Papadatos, G.; Atkinson, F.; van Westen, G.; Dedman, N.; Ochoa, R.; Overington, J. MyChEMBL: A Virtual Platform for Distributing Cheminformatics Tools and Open Data. *Challenges* **2014**, *5*, 334–337.

(25) Bento, A. P.; Gaulton, A.; Hersey, A.; Bellis, L. J.; Chambers, J.; Davies, M.; Krüger, F. A.; Light, Y.; Mak, L.; McGlinchey, S.;

Nowotka, M.; Papadatos, G.; Santos, R.; Overington, J. P. The ChEMBL Bioactivity Database: An Update. *Nucleic Acids Res.* **2014**, *42*, D1083−D1090.

(26) Chien, E. Y. T.; Liu, W.; Zhao, Q.; Katritch, V.; Han, G. W.; Hanson, M. A.; Shi, L.; Newman, A. H.; Javitch, J. A.; Cherezov, V.; Stevens, R. C. Structure of the Human Dopamine D3 Receptor in Complex with a D2/D3 Selective Antagonist. *Science* **2011**, *330*, 1091−1095.

(27) Selent, J.; Lopez, L.; Sanz, F.; Pastor, M. Multi-Receptor Binding Profile of Clozapine and Olanzapine: A Structural Study Based on the New β2 Adrenergic Receptor Template. *ChemMedChem* **2008**, *3*, 1194−1198.

(28) Selent, J.; Marti-Solano, M.; Rodríguez, J.; Atanes, P.; Brea, J.; Castro, M.; Sanz, F.; Loza, M. I.; Pastor, M. Novel Insights on the Structural Determinants of Clozapine and Olanzapine Multi-Target Binding Profiles. *Eur. J. Med. Chem.* **2014**, *77*, 91−95.

(29) Kalani, M. Y. S.; Vaidehi, N.; Hall, S. E.; Trabanino, R. J.; Freddolino, P. L.; Kalani, M. A.; Floriano, W. B.; Kam, V. W. T.; Goddard, W. A. The Predicted 3D Structure of the Human D2 Dopamine Receptor and the Binding Site and Binding Affinities for Agonists and Antagonists. *Proc. Natl. Acad. Sci. U. S. A.* **2004**, *101*, 3815−3820.

(30) Luedtke, R. R.; Mishra, Y.; Wang, Q.; Griffin, S. A.; Bell-Horner, C.; Taylor, M.; Vangveravong, S.; Dillon, G. H.; Huang, R. Q.; Reichert, D. E.; MacH, R. H. Comparison of the Binding and Functional Properties of Two Structurally Different D2 Dopamine Receptor Subtype Selective Compounds. *ACS Chem. Neurosci.* **2012**, *3*, 1050−1062.

(31) Wang, Q.; MacH, R. H.; Luedtke, R. R.; Reichert, D. E. Subtype Selectivity of Dopamine Receptor Ligands: Insights from Structure and Ligand-Based Methods. *J. Chem. Inf. Model.* **2010**, *50*, 1970−1985.

(32) Zanatta, G.; Nunes, G.; Bezerra, E. M.; da Costa, R. F.; Martins, A.; Caetano, E. W. S.; Freire, V. N.; Gottfried, C. Antipsychotic Haloperidol Binding to the Human Dopamine D3 Receptor: Beyond Docking Through QM/MM Refinement Toward the Design of Improved Schizophrenia Medicines. *ACS Chem. Neurosci.* **2014**, *5*, 1041−1054.

(33) 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.

(34) The UniProt Consortium. UniProt: A Hub for Protein Information. *Nucleic Acids Res.* **2015**, *43*, D204−D212.

(35) Chalmers, D. K.; Roberts, B. P. *Silico−a Perl Molecular Modelling Toolkit*, version 1.1; Monash University: Melbourne, 2011.

(36) *The PyMol Molecular Graphics System*, version 1.504; Schrodinger LLC: New York, NY, 2013.

(37) *Suite 2012: Maestro*, version 9.3; Schrodinger LLC: New York, NY, 2012.

(38) McRobb, F. M. Studies of G Protein-Coupled Receptor Structure and Function. PhD Thesis, Monash University, 2011.

(39) Madhavi Sastry, G.; Adzhigirey, M.; Day, T.; Annabhimoju, R.; Sherman, W. Protein and Ligand Preparation: Parameters, Protocols, and Influence on Virtual Screening Enrichments. *J. Comput.-Aided Mol. Des.* **2013**, *27*, 221−234.

(40) *Impact*, version 6.2; Schrödinger, LLC: New York, NY, 2014.

(41) *Prime*, version 3.5; Schrödinger, LLC: New York, NY, 2014.

(42) *Epik*, version 2.7; Schrödinger, LLC: New York, NY, 2013.

(43) Jacobson, M. P.; Friesner, R. A.; Xiang, Z.; Honig, B. On the Role of the Crystal Environment in Determining Protein Side-Chain Conformations. *J. Mol. Biol.* **2002**, *320*, 597−608.

(44) Jacobson, M. P.; Pincus, D. L.; Rapp, C. S.; Day, T. J. F.; Honig, B.; Shaw, D. E.; Friesner, R. A. A Hierarchical Approach to All-Atom Protein Loop Prediction. *Proteins: Struct., Funct., Genet.* **2004**, *55*, 351−367.

(45) Shelley, J. C.; Cholleti, A.; Frye, L. L.; Greenwood, J. R.; Timlin, M. R.; Uchimaya, M. Epik: A Software Program for pK(a) Prediction and Protonation State Generation for Drug-like Molecules. *J. Comput.-Aided Mol. Des.* **2007**, *21*, 681−691.

(46) Greenwood, J. R.; Calkins, D.; Sullivan, A. P.; Shelley, J. C. Towards the Comprehensive, Rapid, and Accurate Prediction of the Favorable Tautomeric States of Drug-like Molecules in Aqueous Solution. *J. Comput.-Aided Mol. Des.* **2010**, *24*, 591−604.

(47) 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.

(48) Hénin, J.; Shinoda, W.; Klein, M. L. United-Atom Acyl Chains for CHARMM Phospholipids. *J. Phys. Chem. B* **2008**, *112*, 7008−7015.

(49) Bader, J. S.; Chandler, D. Computer Simulation Study of the Mean Forces between Ferrous and Ferric Ions in Water. *J. Phys. Chem.* **1992**, *96*, 6423−6427.

(50) Vanommeslaeghe, K.; Hatcher, E.; Acharya, C.; Kundu, S.; Zhong, S.; Shim, J.; Darian, E.; Guvench, O.; Lopes, P.; Vorobyov, I.; Mackerell, A. D. CHARMM General Force Field: A Force Field for Drug-like Molecules Compatible with the CHARMM All-Atom Additive Biological Force Fields. *J. Comput. Chem.* **2010**, *31*, 671−690.

(51) 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.* **2012**, *33*, 2451−2468.

(52) *MacroModel*, version 9.9; Schrodinger, LLC: New York, 2014.

(53) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Zakrzewski, V. G.; Montgomery Jr, J. A.; Stratmann, R. E.; Burant, J. C.; Dapprich, S.; Millam, J. M.; Daniels, A. D.; Kudin, K. N.; Strain, M. C.; Farkas, O.; Tomasi, J.; Barone, V.; Cossi, M.; Cammi, R.; Mennucci, B.; Pomelli, C.; Adamo, C.; Clifford, S.; Ochterski, J.; Petersson, G. A.; Ayala, P. Y.; Cui, Q.; Morokuma, K.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Cioslowski, J.; Ortiz, J. V.; Baboul, A. G.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Gomperts, R.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Gonzalez, C.; Challacombe, M.; Gill, P. M. W.; Johnson, B. G.; Chen, W.; Wong, M. W.; Andres, J. L.; Head-Gordon, M.; Replogle, E. S.; Pople, J. A. *Gaussian 98*, Revision A.7; Gaussian, Inc.: Pittsburgh, PA, 1998.

(54) Dupradeau, F.-Y.; Pigache, A.; Zaffran, T.; Savineau, C.; Lelong, R.; Grivel, N.; Lelong, D.; Rosanski, W.; Cieplak, P. The R.E.D. Tools: Advances in RESP and ESP Charge Derivation and Force Field Library Building. *Phys. Chem. Chem. Phys.* **2010**, *12*, 7821−7839.

(55) Bayly, C. C. I.; Cieplak, P.; Cornell, W. D.; Kollman, P. A. A Well-Behaved Electrostatic Potential Based Method Using Charge Restraints for Deriving Atomic Charges: The RESP Model. *J. Phys. Chem.* **1993**, *97*, 10269−10280.

(56) Vanommeslaeghe, K.; MacKerell, A. D. Automation of the CHARMM General Force Field (CGenFF) I: Bond Perception and Atom Typing. *J. Chem. Inf. Model.* **2012**, *52*, 3144−3154.

(57) 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.* **2012**, *52*, 3155−3168.

(58) Sikazwe, D. M. N.; Li, S.; Mardenborough, L.; Cody, V.; Roth, B. L.; Ablordeppey, S. Y. Haloperidol: Towards Further Understanding of the Structural Contributions of Its Pharmacophoric Elements at D2-like Receptors. *Bioorg. Med. Chem. Lett.* **2004**, *14*, 5739−5742.

(59) Bochevarov, A. D.; Harder, E.; Hughes, T. F.; Greenwood, J. R.; Braden, D. A.; Philipp, D. M.; Rinaldo, D.; Halls, M. D.; Zhang, J.; Friesner, R. A. Jaguar: A High-Performance Quantum Chemistry Software Program with Strengths in Life and Materials Sciences. *Int. J. Quantum Chem.* **2013**, *113*, 2110−2142.

(60) 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.

(61) MacKerell, A. D.; Bashford, D.; Bellott, M.; Dunbrack, R. L.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; Joseph-McCarthy, D.; Kuchnir, L.; Kuczera, K.; Lau, F. T. K.; Mattos, C.; Michnick, S.; Ngo, T.; Nguyen, D. T.; Prodhom, B.; Reiher, W. E.;

Roux, B.; Schlenkrich, M.; Smith, J. C.; Stote, R.; Straub, J.; Watanabe, M.; Wiorkiewicz-Kuczera, J.; Yin, D.; Karplus, M. All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins. *J. Phys. Chem. B* **1998**, *102*, 3586−3616.

(62) Mackerell, A. D.; Feig, M.; Brooks, C. L. Extending the Treatment of Backbone Energetics in Protein Force Fields: Limitations of Gas-Phase Quantum Mechanics in Reproducing Protein Conformational Distributions in Molecular Dynamics Simulation. *J. Comput. Chem.* **2004**, *25*, 1400−1415.

(63) Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G. A Smooth Particle Mesh Ewald Method. *J. Chem. Phys.* **1995**, *103*, 8577−8593.

(64) Humphrey, W.; Dalke, A.; Schulten, K. VMD: Visual Molecular Dynamics. *J. Mol. Graphics* **1996**, *14*, 33−38.

(65) Beauchamp, K. A.; Bowman, G. R.; Lane, T. J.; Maibaum, L.; Haque, I. S.; Pande, V. S. MSMBuilder2: Modeling Conformational Dynamics on the Picosecond to Millisecond Scale. *J. Chem. Theory Comput.* **2011**, *7*, 3412−3419.

(66) Barton, A. C.; Kang, H. C.; Rinaudo, M. S.; Monsma, F. J.; Stewart-Fram, R. M.; Macinko, J. A.; Haugland, R. P.; Ariano, M. A.; Sibley, D. R. Multiple Fluorescent Ligands for Dopamine Receptors. I. Pharmacological Characterization and Receptor Selectivity. *Brain Res.* **1991**, *547*, 199−207.

(67) Vangveravong, S.; Zhang, Z.; Taylor, M.; Bearden, M.; Xu, J.; Cui, J.; Wang, W.; Luedtke, R. R.; MacH, R. H. Synthesis and Characterization of Selective Dopamine D2 Receptor Ligands Using Aripiprazole as the Lead Compound. *Bioorg. Med. Chem.* **2011**, *19*, 3502−3511.

(68) Shimamura, T.; Shiroishi, M.; Weyand, S.; Tsujimoto, H.; Winter, G.; Katritch, V.; Abagyan, R.; Cherezov, V.; Liu, W.; Han, G. W.; Kobayashi, T.; Stevens, R. C.; Iwata, S. Structure of the Human Histamine H1 Receptor Complex with Doxepin. *Nature* **2011**, *475*, 65−70.

(69) Cherezov, V.; Rosenbaum, D. M.; Hanson, M. A.; Rasmussen, S. G. F.; Thian, F. S.; Kobilka, T. S.; Choi, H.-J.; Kuhn, P.; Weis, W. I.; Kobilka, B. K.; Stevens, R. C. High-Resolution Crystal Structure of an Engineered Human beta2-Adrenergic G Protein-Coupled Receptor. *Science* **2007**, *318*, 1258−1265.

(70) Katritch, V.; Cherezov, V.; Stevens, R. C. Structure-Function of the G Protein-Coupled Receptor Superfamily. *Annu. Rev. Pharmacol. Toxicol.* **2013**, *53*, 531−556.

(71) Lloyd, E. J.; Andrews, P. R. A Common Structural Model for Central Nervous System Drugs and Their Receptors. *J. Med. Chem.* **1986**, *29*, 453−462.

(72) Boström, J.; Gundertofte, K.; Liljefors, T. A Pharmacophore Model for Dopamine D4 Receptor Antagonists. *J. Comput.-Aided Mol. Des.* **2000**, *14*, 769−786.

(73) Lane, J. R.; Chubukov, P.; Liu, W.; Canals, M.; Cherezov, V.; Abagyan, R.; Stevens, R. C.; Katritch, V. Structure-Based Ligand Discovery Targeting Orthosteric and Allosteric Pockets of Dopamine Receptors. *Mol. Pharmacol.* **2013**, *84*, 794−807.

Chapter 4

Ligand Binding of Oleic Acid to Liver Fatty Acid-Binding Protein

The work presented in Chapter 3 showed that, although conventional molecular dynamics simulations are able to illuminate ligand-binding processes, they are too slow and inefficient to be applied routinely to ligand binding in G protein-coupled receptors. We thus identified a need to employ new methods in our investigations of ligand binding. We used Markov state models to enhance the sampling of ligand-receptor complexes and to obtain statistics that could be used to compare metastable states and binding pathways.

To develop a method for applying MSMs to ligand binding, we sought a smaller ligand-binding system that could be simulated faster than GPCR systems but had binding pathways that still bore some resemblance to binding in GPCRs. The model system we chose to work with was oleic acid binding to the liver fatty acid-binding protein (FABP), for which large amounts of conventional MD data could be generated very efficiently. We performed long-timescale conventional MD for significantly longer simulation time than any previous study, providing an understanding of how ligands bind to fatty acid-binding proteins and identifying a novel metastable site. We then proceeded to construct a MSM of the system.

At the time this work was begun, the only publically available MSM programs were still in early and very active development, providing only token support for doing anything other than protein folding. The available system descriptors and clustering methods were designed and implemented with protein folding in mind and we had to select methods that could also reasonably apply to ligand-binding systems. Even once a suitable method had been chosen, persuading the software to deal with a ligand-binding problem required frequent forays into the source code.

During construction of MSMs, we quickly learned that the flexibility of both the fatty acid ligand and the protein itself made constructing an MSM a challenging task, and we had to limit the scope of the MSM to the post-association phases of the ligand-binding ensemble. Taken together, the conventional MD simulations and the dataset for the MSM provided 2 orders of magnitude more simulation time than any previous study of FAPBs. The final MSM identified novel binding pathways not observed by conventional simulations and advances our understanding of ligand binding to FABPs. In hindsight, our chosen model system was more difficult to work with than the GPCR system we had sought to avoid.

4.1 Introduction

4.1.1 Kinetics of ligand binding

Structure-based drug design is a growing field that relies on the 3D structure of the biological target to develop the effectiveness of a drug. The aim is to use this structural information as a guide to design more potent compounds. This is usually done by modifying the drug to increase binding affinity. Binding affinity is described by the equilibrium dissociation constant K_d and related to the kinetics of drug binding by Equation 1, where k_{off} and k_{on} are the rate constants for dissociation and association respectively:

$$K_d = \frac{k_{off}}{k_{on}} \tag{1}$$

This approach of optimizing K_d, although it has seen much success, is largely reliant on two approximations: that drug binding is occurring under equilibrium conditions, and that drug binding follows a simple two-state model. In reality, the concentration of a drug *in vivo* varies over time as it is metabolized or absorbed, and drug binding often involves many states each separated by their own kinetic barriers. Thus, studies have shown that k_{off} and k_{on} are often more closely related to the effectiveness of a drug than binding affinity alone.[1,2]

While, in theory, either k_{on} or k_{off} can be modified to optimize the binding of a ligand, practically, the approaches to do so are limited. Optimizing k_{on} is useful for slow-binding ligands, where binding may be gated by conformational rearrangement, but an already fast on-rate can only be increased up to the rate of diffusion. Additionally, experimental structures usually do not provide information that can be used to optimize k_{on}. When considering k_{off} in structure-based drug design, the available information is often limited to the single bound pose of the ligand provided by experimental structures. For these reasons, structure-based drug design is usually focused on decreasing the k_{off} using only the final bound state of the ligand. Focusing on this single state of the ligand is not always sufficient; targeting regions outside the binding site, by extending the ligand into nearby pockets or designing ligands that bind to allosteric sites, is a useful means for engineering selectivity or fine-tuning the behavior of a ligand. However, designing ligands to bind outside the binding site often proves far from simple because the static structures produced by experimental methods (e.g. X-ray crystallography) can be inadequate for studying these often more dynamic regions. The very nature of crystal structures denies us access to the dynamics of a ligand or flexibility of a receptor beyond the vicinity of the co-crystalized ligand, and the static structures betray no indication of how the ligand arrived. Only computational methods have sufficient temporal- and structural-resolution to investigate the trajectory of a ligand en route to its binding site in the atomic detail necessary to understand the binding pathway.

MD simulations provide a view of conformational changes and molecular interactions that is updated on the order of femtoseconds, making them particularly useful for studying the binding of ligands.[3] Modeling the dynamic behavior of a system allows simulations to overcome many of the limitations inherent in experimental structures. By using MD to predict and visualize the entire binding pathway we can, in principle, identify specific residues that facilitate the binding of ligands, ensuring that we propagate these essential interactions through the drug design process. MD simulations can also identify "metastable binding sites" where ligands pause for a significant amount of time before resuming their journey.[4] These sites have been found to correspond to known allosteric sites[5] and being far less conserved than the binding site itself are of particular interest to structure-based drug design.

4.1.2 Markov state models

Markov state models (MSMs) are an old statistical method that have only recently been applied to molecular simulations,[6] and that describes the behavior of a system of states based on the transition probabilities between those states (see Chapter 1). MSMs have seen great success in the investigation of protein folding[7] and the development of programs, such as MSMBuilder[8] and pyEMMA[9], that assist MSM construction has fostered interest in this method.[10] MSMs are able to simultaneously identify and characterize the entire binding ensemble and assume no prior knowledge of the system.

Practically, constructing MSMs rather than using more traditional MD investigation methods has a number of advantages. By breaking a molecular process into discrete states, any simulation only has to be long enough to encompass a single transition in order to provide useful data. Thus, instead of running simulations on the microsecond scale that is usually required to observe entire binding pathways, hundreds of shorter simulations can be used to take full advantage of parallel computers. These short simulations can then be stitched together into a transition network, based on the overlap of the discrete states they have in common, and the transition network can then be used to predict events beyond the timescales of the individual simulations. Nevertheless, the construction of a MSM still requires copious amounts of data to produce a statistically reliable model and care needs to be taken to ensure that all of the important state space is explored. In many cases, additional simulations may need to be performed to sample transitions where data may be scarce.

4.1.3 Fatty acid-binding proteins

Fatty acid-binding proteins (FABPs) are small cytosolic proteins that are strongly linked to metabolic and inflammatory pathways through their role as chaperones for the transport of lipophilic compounds towards sites of metabolism and signaling.[11] FABPs are known to carry drugs across the nuclear membrane[12] and FABP expression has been tied to insulin sensitivity and blood glucose and cholesterol levels.[13] The current work is centered on liver fatty acid-binding protein (L-FABP) of which both X-ray crystal and NMR structures of the human orthologue have been solved.[14,15] Both structures reveal L-FABP to have a larger binding pocket than other FABPs, allowing it to bind 2 fatty acids simultaneously in distinct high- and low-affinity binding sites (Figure 1). Similar to GPCRs, FABPs have a deep binding pocket formed by a β-barrel that consists of 10 β-strands (A-J) in 2, anti-parallel, β-sheets. The β-barrel is capped on one side by 2 α-helices (H1, H2) and it is this side of the barrel, commonly referred to as the portal region, that is thought to allow the entry and exit of ligands.[16] Once in the binding pocket of L-FABP, fatty acid head-groups form a salt-bridge with Arg122, reinforced by hydrogen bonds to Ser39 and Ser124. The deep binding pocket of L-FABP guarantees a multiple step binding pathway that is needed for method and analytical development that can later be scaled up to GPCRs. Their small size and the absence of a membrane environment make FABPs optimal for MD simulations.

Molecular dynamics simulations of ligand binding to FABPs were first performed decades ago and simulation lengths range from the picosecond[17] to microsecond[18] timescales. Access to the nanosecond timescale enabled individual simulations to begin investigating the ligand binding pathway; Friedman et al. found that in their 3 initial simulations of up to 10 ns, the hydrophobic tail of their fatty acid ligand penetrated into the anti-portal region of the FABP, the end of the barrel opposite the helical cap.[19] This binding pathway was unsupported by experimental data and is likely to be an artefact of the limited sampling. Binding to this region was observed in later simulations,[20] but it was calculated that egress through the anti-portal required much larger conformational re-arrangement of the protein than during portal binding.[21] The improved sampling provided by longer timescale MD simulations produced results that are in agreement with the portal hypothesis of ligand binding.[22,23] Most notably, Li et al. conducted simulations of 5 different ligands binding to adipocyte FABP. For each ligand they performed two 1.2 µs simulations and found several distinct binding modes within the portal region.

In this study we have investigated the binding of oleic acid to L-FABP; first by performing a series of conventional MD simulations to examine the binding pathways of oleic acid from the bulk solvent to the high-affinity binding site, then, adding an additional set of simulations, by constructing a MSM to thoroughly explore the binding ensemble of oleic acid in the portal region of the receptor. With these two approaches combined we have simulated the L-FABP-oleic acid system for a total of 215 μs.

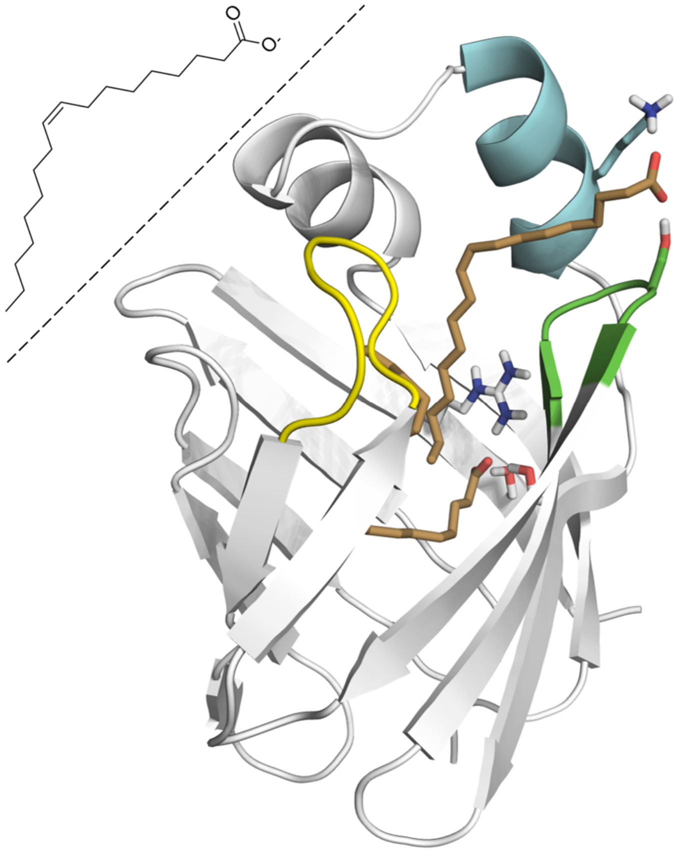


Figure 1. Cartoon representation of L-FABP (PDB ID: 2LKK) highlighting key areas and interactions: H2 (cyan), loopCD (green), and loopEF (yellow). Oleic acid is shown in the structure as brown and shown as a 2D structure in the upper-left. The ligand in the high-affinity binding site interacting with Arg122, Ser39 and Ser124 and the ligand in the low-affinity binding site interacting with Ser56 and Lys31.

4.2 Method

All Simulations were run on the MASSIVE, VLSCI and NCI clusters using Gromacs 4.6.2[24] with the GROMOS 54a7[25] force field and the TIP3P water model. Oleic acid parameters were united-atom, modified from the GROMOS 54a7 force field according to Bachar et al.[26,27] VMD[28] was used for the visualization and analysis of systems. Analysis procedures were augmented by in-house scripts written using the Perl and TCL scripting languages. All graphs were generated using gnuplot.

4.2.1 System preparation

Systems were constructed to model the spontaneous binding of oleic acid to L-FABP, from the solvent to the experimentally determined binding site. Each system consisted of a single protein, explicit solvent, and 1 or 2 oleic acid molecules. The protein model was based on the NMR structure of the human L-FABP-oleic acid complex (PDB ID: 2LKK).[15] The first conformation within the ensemble of PDB structures was used; all conformations were similar enough to be considered representative (Maximum RMSD 0.074 Å). Periodic boundary conditions were used with an orthorhombic dodecahedron unit cell, constructed such that there was a minimum of a 5 Å distance between the solute and the periodic boundary, which resulted in a minimum distance between solute atoms of neighboring images of approximately 25 Å. A two ligand model was constructed by placing 2 oleic acid molecules randomly in the periodic cell (closest distance to solute 6.5 Å). The resulting system was then solvated with the Gromacs gen_box script to a default density (4,326 water molecules), allowing solvation of the binding site. Sodium and chloride ions were added to a concentration of 150mM NaCl and ensuring that the model system was electrically neutral. A single-ligand model was created by deleting 1 Na^+ ion and 1 ligand. The entire system was represented by the GROMOS 54a7 united-atom force field.

4.2.2 Long-timescale molecular dynamics

The first set of simulations was conducted using a conventional, unbiased MD approach using simulations of 1-2 µs in length, commencing from the same starting structure but assigned different initial velocities. Each simulation sought to capture an entire binding event of oleic acid, from the solvent to the high-affinity binding site. Simulations used the Verlet cut-off scheme, in order to utilize GPU architecture. The cut-off for van der Waals and short-range electrostatics was 9 Å. The particle mesh Ewald method was used to calculate long-range electrostatics.[29] All bond lengths were constrained with the LINCS algorithm.[30] Equilibration simulations were performed with a 2 fs time step, which was increased to a 5 fs time step for production runs.

All long time-scale simulations were set up using a 5-step procedure, with each preliminary MD step being performed for 1 ns followed by the production run. The steps are as follows: (1) A steepest descent minimization was performed for up to 2000 steps. (2) Initial velocities were generated with a unique random seed, protein and ligand atoms were restrained, and the Berendsen thermostat was applied to the entire system (310 K, 0.1 ps time constant). (3) Temperature coupling was applied separately to protein and non-protein atoms, Berendsen pressure coupling was introduced (1 atm, 20 ps time constant). (4) Restraints on protein side chains were released, temperature coupling was switched to the v-rescale thermostat, and pressure coupling was switched to the Parrinello-Rahman method[31] (time constant adjusted to 2 ps). (5) The remaining restraints on the protein backbone and ligand were released. (6) In the production run, the time step was increased to 5 fs and the pressure coupling frequency was doubled to compensate. Simulation coordinates were saved every 50 ps.

Analysis was performed on a subset of production run data such that RMSDs and interactions were calculated every 1 ns. RMSDs were calculated relative to the experimental structure (PDB ID: 2LKK), the RMSD of each ligand was calculated to each of the 2 poses present in this structure.

4.2.3 Markov state models

MSMs were constructed with MSMbuilder 2.6,[8] and MSMexplorer 0.9[32] was used to visualize transition networks. Insight from the long-timescale simulations was used to identify metastable sites to use as starting structures for hundreds of shorter simulations that were used to construct MSMs. Within the metastable sites, starting structures were selected such that they had a low backbone RMSD to the experimental structure. In total, 3 structures were selected from each of the high-affinity binding site, the low-affinity binding site, and the metastable site at loopEF. Initially 50 simulations were performed from each starting structure, totalling 450 simulations. The starting structures were minimized, then the production run was performed, as described above, for 200 ns. One batch of 50 simulations from each starting point was extended to 400 ns. An initial MSM was constructed from these initial 450 simulations and it was determined that transitions into the high-affinity binding site required further sampling. To address this, 4 additional starting structures were chosen around transitions into the high-affinity binding site and an additional 50 simulations were performed from each structure to better sample this transition, resulting in a total of 650 simulations which were used to construct the final MSM.

MSMs of the binding pathway were constructed according to a modified version of the workflow described by Bowman for protein folding simulations.[33] As the first step in constructing the model, trajectories were aligned by the protein backbone residues. Data was then clustered according to the RMSD of all ligand atoms using the hybrid k-centers/k-medoids method implemented within MSMbuilder with a maximum intra-cluster distance of 2 Å. This process generated a total of 2941 clusters (microstates). A transition network was calculated for this set of microstates and the Bayesian agglomerative clustering engine (BACE)[34] algorithm was used to coarse-grain them into macrostates. The BACE algorithm works by iteratively merging the most kinetically similar (rapidly interconverting) states, and the Bayes factors produced indicate how well each step preserves the behavior of the original microstate model. From these Bayes factors it was determined that less than 16 states would result in a sharp decline in the ability of the macrostate model to preserve the behavior of the microstate mode, so a 16-state model was generated with a lag time of 5 ns. It was found that 4 states in this model were poorly connected, with an equilibrium population below 0.001% and these states were excised from the final model.

4.3 Results and discussion

4.3.1 Analysis of oleic acid binding to L-FABP using long timescale simulations

While it is known from the experimental X-ray and NMR structures that L-FABP is capable of binding two ligands, little is known about how the ligands arrive or depart. Such knowledge is important for the design, or tuning the kinetics, of ligands that bind to L-FABP. For example, if the intent is to use L-FABP to transport a drug to the nucleus, the drug needs to be able to unbind from L-FABP at an appropriate timescale to dissociate from the receptor in the nucleus. To this end we simulated the binding of oleic acid molecules to L-FABP, 1 or 2 at a time. In total, we simulated unbiased binding of oleic acid to liver fatty acid-binding protein for a total of 55 µs across 30 simulations (Table 1), with 15 simulations containing a single ligand (designated 1-1 to 1-15) and 15 containing two (designated 2-1 to 2-15). Of the 30 long-timescale simulations conducted, 5 showed the ligand spontaneously traversing from solution to the high-affinity binding site and 3 of these gave structures where the ligand was within 3 Å of the experimentally determined structure. One simulation (2-13) reproduced the experimental structure for both ligands and protein with an RMSD < 2 Å (Figure 2).

		Ligand RMSD (Å)						Interactions		
		To high-affinity		To low-affinity		Final protein RMSD (Å)		Arg122	Ser56	Met74
Single oleic acid simulations										
1-1	A	8.7	5.3	8.7	6.2	3.6	H			
1-2	A	17.8	9.8	28.8	21.0	2.7	A			
1-3	A	16.1	8.7	25.5	19.8	2.9	A			
1-4	A	12.4	8.6	9.6	2.3	2.8	P		✓ ✓	✓
1-5	A	13.8	7.9	24.6	18.5	3.6	A			
1-6	A	9.1	8.5	10.1	6.3	3.0	P		✓	✓
1-7	A	13.2	8.9	10.8	7.7	4.2	P			✓
1-8	A	11.4	9.0	9.7	5.6	3.9	P		✓	✓
1-9	A	7.6	5.6	9.5	2.3	3.0	P	✓	✓	✓
1-10	A	11.4	8.8	9.4	7.4	2.8	P		✓	✓
1-11	A	15.3	10.6	14.5	8.1	6.2	A-P			✓
1-12	A	8.4	4.9	12.1	3.1	3.4	P	✓	✓	
1-13	A	18.4	9.8	28.0	21.5	2.5	A			
1-14	A	11.4	8.6	9.2	3.5	2.7	P		✓	✓
1-15	A	13.5	7.3	12.0	7.2	3.3	P			✓
Two oleic acid simulations										
2-1	A	8.1	6.0	19.3	17.8	3.4	A			
	B	10.3	7.4	22.3	19.3		A			
2-2	A	7.3	2.2	13.1	7.5	3.2	P	✓		
	B	12.9	9.7	3.6	2.4		P		✓	✓
2-3	A	17.9	13.5	28.3	22.5	4.8	A			
	B	13.2	9.1	24.0	10.5		A			
2-4	A	13.8	8.7	13.2	2.0	4.2	P		✓	✓
	B	13.2	9.6	5.6	3.1		P		✓	✓
2-5	A	7.4	6.0	11.3	6.5	4.7	P			✓
	B	10.4	8.9	8.8	6.4		P			✓
2-6	A	13.8	9.8	13.4	7.6	3.1	P			✓
	B	10.3	8.1	9.3	1.8		P		✓	✓
2-7	A	11.0	5.9	21.5	16.9	4.5	A			
	B	11.2	8.4	7.6	1.7		P		✓	✓
2-8	A	24.9	10.5	23.1	5.5	5.0	H			
	B	4.3	3.5	12.5	2.9		P	✓		✓
2-9	A	11.1	7.2	12.0	9.2	4.3	P			✓
	B	9.3	7.9	9.5	6.3		H	✓		
2-10	A	11.3	6.4	11.2	4.1	3.9	P		✓	✓
	B	8.1	5.7	9.2	6.4		H	✓		
2-11	A	13.3	11.2	6.1	2.2	3.1	P		✓	✓
	B	16.7	8.8	16.1	8.8		P			✓
2-12	A	10.2	7.6	7.2	5.2	4.2	P			✓
	B	11.1	9.9	11.9	8.6		P			
2-13	A	13.2	9.9	2.8	1.7	1.9	P		✓	
	B	2.3	1.6	13.0	8.5		P	✓		
2-14	A	12.2	8.4	11.5	7.3	2.6	P			✓
	B	2.8	2.0	13.0	2.1		P	✓	✓	
2-15	A	11.5	8.5	6.3	2.3	3.3	P		✓	✓
	B	14.4	8.5	10.7	7.9		P			✓

Table 1. A summary of long-timescale FABP simulations. All RMSD values are given in Angstroms. Ligand regions A, H, or P indicate whether the ligand entered the anti-portal, helical cap, or portal regions respectively. Simulation 2-13 (highlighted) accurately reproduced the experimental structure. Ligand RMSDs are reported for the minimum reached and for the final frame of the simulation.

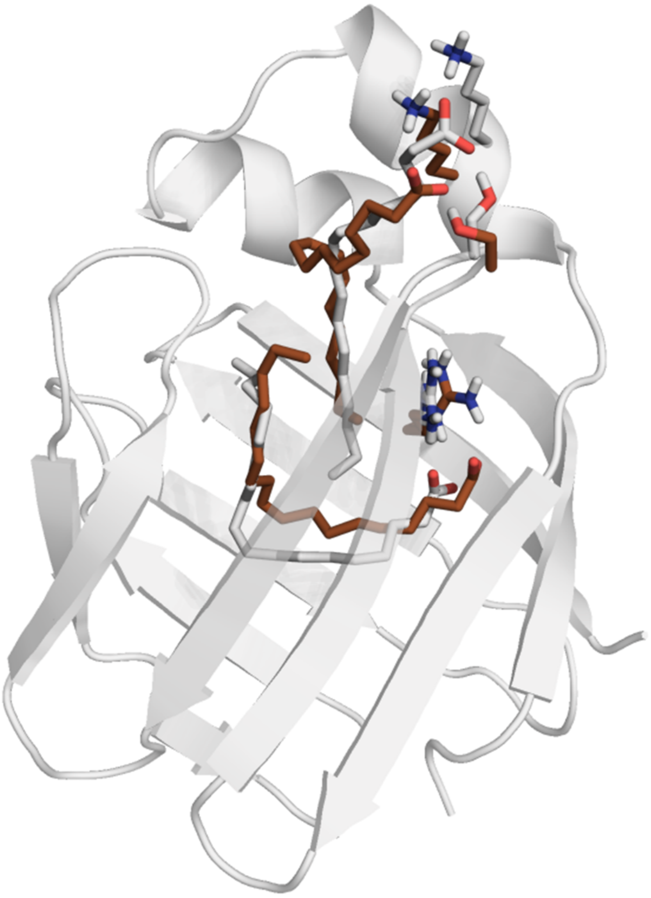


Figure 2. Simulation 2-13 (brown) reproduced the experimentally determined pose (white) within 2 Å RMSD.

Although all binding simulations were started from the same initial placement of ligands, the initial velocities of each simulation were set using a random seed unique to that simulation, causing simulations to diverge at the first simulation step. Ligands were seen to adsorb to the protein during the opening nanoseconds of simulation and roamed around the protein surface before settling in 1 of 3 regions: the portal, anti-portal, and helical cap regions (Figure 3). Of the 45 total ligands simulated, 31 arrived in the portal region, 4 inserted between the two helices in the cap, and 10 adhered to the anti-portal region (1 of which subsequently dissociated from the anti-portal late in the simulation and migrated to the portal region). These observations are in good agreement with previous MD studies of a much shorter duration that studied the locations of adsorption of fatty acids to FABPs.[20]

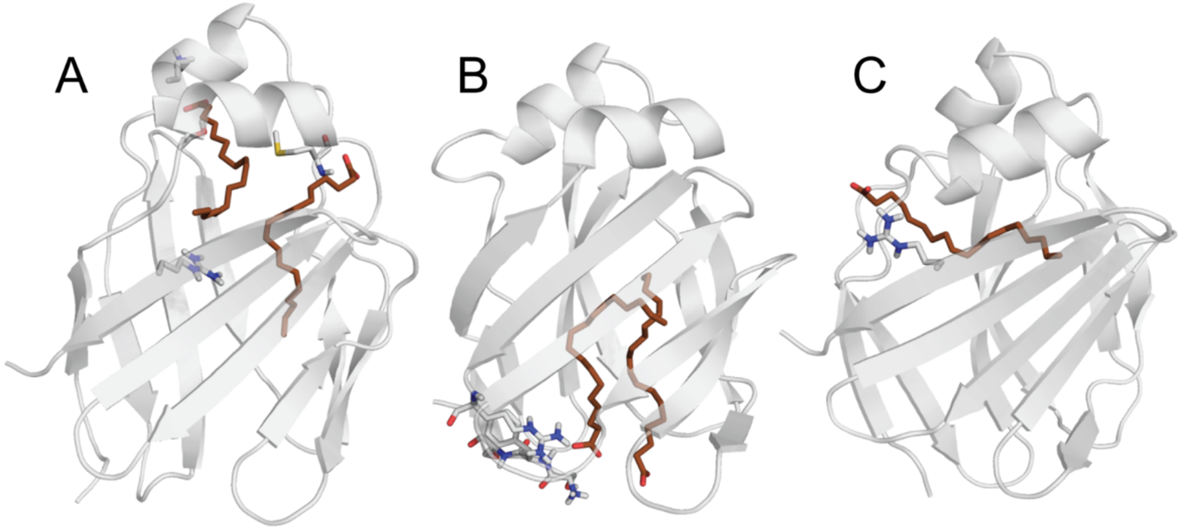


Figure 3. Three distinct ligand binding regions were observed: (A) portal, (B) anti-portal, and (C) helical cap regions.

A key finding from our long-timescale ligand-binding simulations was the observation of a metastable binding arrangement in which the ligand head-group interacts with the backbone of residues in loopEF, whilst the ligand tail-group was buried in the β-barrel (Figure 4). This binding site is not occupied in the experimental structure and presents a new location to consider when designing ligands that bind to L-FABP. A potential involvement of this loop in ligand binding was postulated by Sharma et al.[14] Based on a set of 3 crystal structures (apo, containing 1 ligand, and containing 2 ligands), they identified an interaction between the side chains of Met74 (in loopEF) and Arg122 that effectively closed the binding site in the apo structure, but which opened upon ligand binding to the high-affinity site. Combined with the observation that there are minimal differences in protein conformation between the 1-ligand and 2-ligand structures, they reasoned that the first ligand binds to the high-affinity site and changes conformation of the receptor to facilitate the binding of the second ligand. The simulations performed in the current work began with the holo conformation of the receptor and the Met74 to Arg122 interaction was only observed transiently prior to obstruction by a ligand. However, evidence from our long-timescale simulations is in general agreement, suggesting that the first ligand proceeds through the low-affinity binding site en route to the high-affinity site, paving the way for a second ligand.

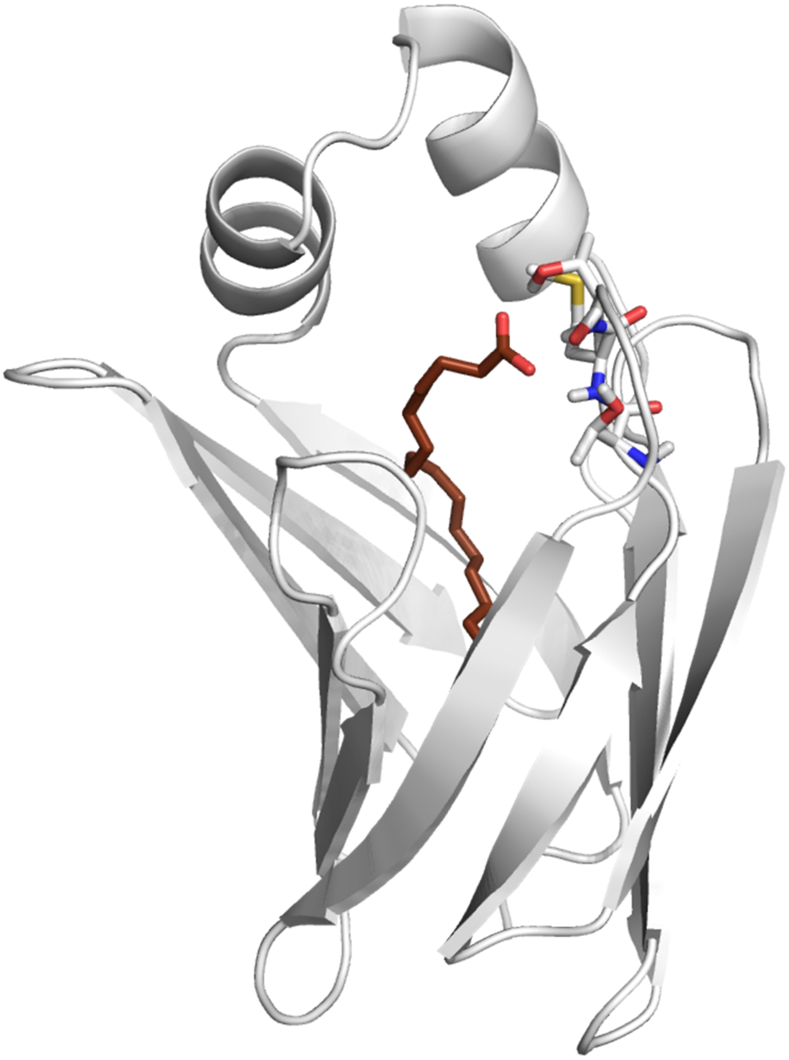


Figure 4. A representative structure of the metastable binding site at loopEF taken from the early part of simulation 2-14. Ligand A is shown interacting with Thr73, Met74, and Thr75.

Across all simulations, a total of 10 simulations showed fatty acids binding to the anti-portal region of the receptor. Binding in this region occurred through initial interaction of the ligand head-group with polar residues Ser2, Phe3, and Ser4 on the N-terminus. This interaction was followed by the hydrophobic tail inserting into the β-barrel. That such a significant number of ligands adhered to the anti-portal region is an unexpected finding. The ability of the fatty acid tails to regularly insert into the β-barrel even more so. However, the residence time of fatty acids in the anti-portal region appears to be brief, with a dissociation event observed within 1 µs (c.f. dissociation time from binding site on the timescale of seconds[35]), and the interactions with Ser2, Phe3, and Ser4, that were observed across multiple simulations (Figure 3B), were frequently broken. Although the anti-portal region seems unrelated to the binding pathways we are investigating, it is worth noting that, similar to the high affinity binding site, the anti-portal region has an Arg+2×Ser motif in a similar arrangement to the motif present in the high-affinity binding site (flexible termini allowing), which lends itself well to the binding of carboxylic groups. In the simulations performed here, once the ligand tail-group retreats into the β-barrel, the head-group loses contact with this motif, but other, perhaps longer, fatty acids may be able to maintain contact.

Across all simulations, 4 ligands were observed to bind by insertion of their tail-group between the helices of the helical cap (Figure 3C). In all cases of 'cap' binding, intercalation between the helices proved destructive to the secondary structure by causing the helices to uncoil. Of the 4 cap binding ligands, only the tail-groups were able to enter the β-barrel while the head-groups remained above the cap. The stability of this pose must also be called to question due to the substantial disruption of helical structure; one ligand was observed to dissociate shortly after the helices unwound. While this observation sheds doubt on the relevance of this helical location to ligand binding from solution, evidence suggests that the helical region interfaces with cell membranes as a key step in fatty acid transfer to membranes or vesicles.[36] Disruption of the helices would likely accelerate this process.[37]

As expected, the majority of ligands proceeded from solution to the portal region, where their head-groups interacted with polar residues in the portal-region loops while the hydrophobic tails were buried in the β-barrel (Figure 3A). The interaction of the fatty acid head-group with loopEF was observed in the majority of simulations. In total, 25/32 portal region ligands interacted with loopEF for over 5% of the simulation time. The 7 portal-region ligands that do not interact with loopEF include all 5 salt-bridge forming ligands, which skip interacting with Met74 by binding prior to 5% of simulation time or making initial contact with Ser56. The remaining 2 portal-region ligands are from simulations in which loopEF is already occupied by another ligand that is unable to contact Ser56 as the low-affinity site is obstructed by a downward shift in H2.

A downward shift in H2 (Figure 5) was an interesting phenomenon observed in many simulations. The movement in H2 typically forced loopCD outwards, where Ser56 maintained a hydrogen bond with the outward facing Lys31 in H2. This conformational change proved either conducive or detrimental to ligand binding depending on the location of the individual ligand. In the majority of cases the outward displacement of loopCD made Ser56 inaccessible to ligands occupying the site at loopEF. This conformational change can be observed in the interactions of the 2 portal binding ligands in simulation 2-15 (Figure 6) – at 1000 ns a change in the protein RMSD can be observed, indicating the return of H2 to the experimentally determined pose. This movement leads to a shift in ligand interactions as they shuffle around from loopEF and loopGH to occupy the low-affinity binding site (loopCD) and loopEF respectively. In cases where the low-affinity binding site was already occupied when the downward shift occurred, the ligand was pushed deeper into the receptor via the gap between strands B and C. In 2 simulations (1-9, 1-12, 2-14) the downward shift resulted in the low-affinity bound ligand making contact with Arg122. This is shown for simulation 2-14 in Figure 5 and Figure 7 – a change in the backbone RMSD at 200 ns indicates the downward movement of H2, pushing the ligand head-group into the β-barrel where it forms a salt-bridge with Arg122, eventually assuming the high-affinity pose. Similar motions were observed in simulation 2-13, but with the high-affinity binding site already occupied, the ligand in the low-affinity pose was able to stand its ground.

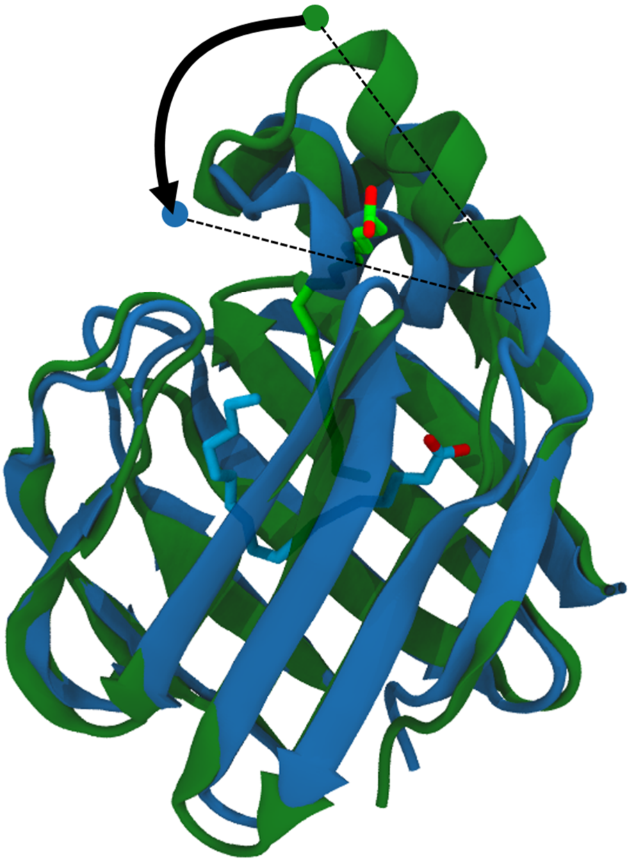


Figure 5. Two overlaid frames of simulation 2-14 showing the conformation of the protein and ligand B before (green) and after (blue) the downward movement of H2. The new position of H2 in blue can be seen to occupy the space previously occupied by the ligand in the low-affinity binding site (green).

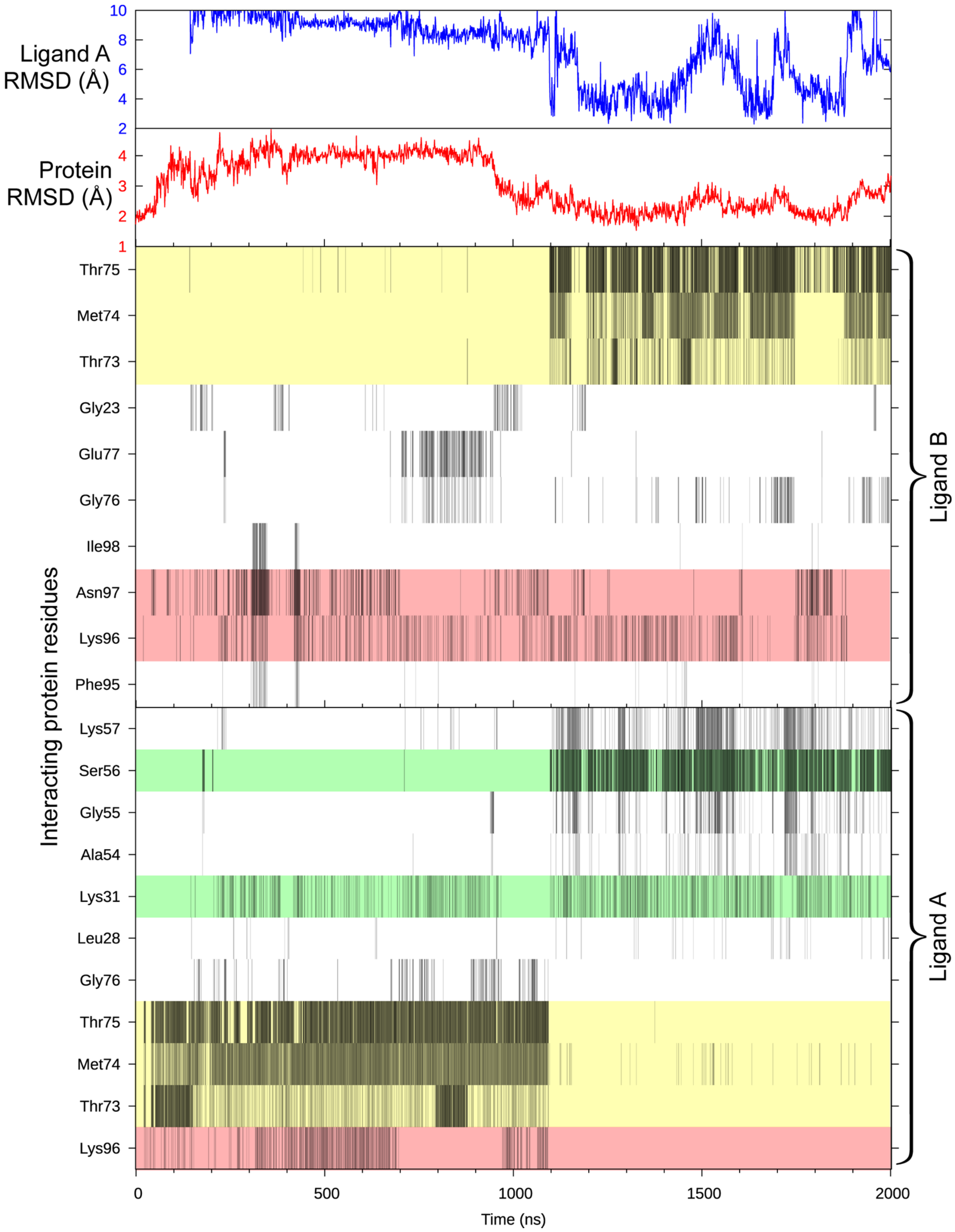


Figure 6. Interaction map representing the movements of the two ligands in simulation 2-15. The top 2 panels show the RMSD of the protein backbone from the experimental pose (red) and the RMSD of ligand A from the low-affinity pose (blue). The lower 'barcode' panes show contacts between the ligands and protein residues (ligand oxygen to protein heavy atom distance of <3.5 Å). Darker barcode lines indicate closer contact. Residues that interacted for less than 5% of the simulation time have been omitted. Residues are shaded based on their affiliation with the low-affinity binding site (green), loopEF (yellow), or loopGH (red).

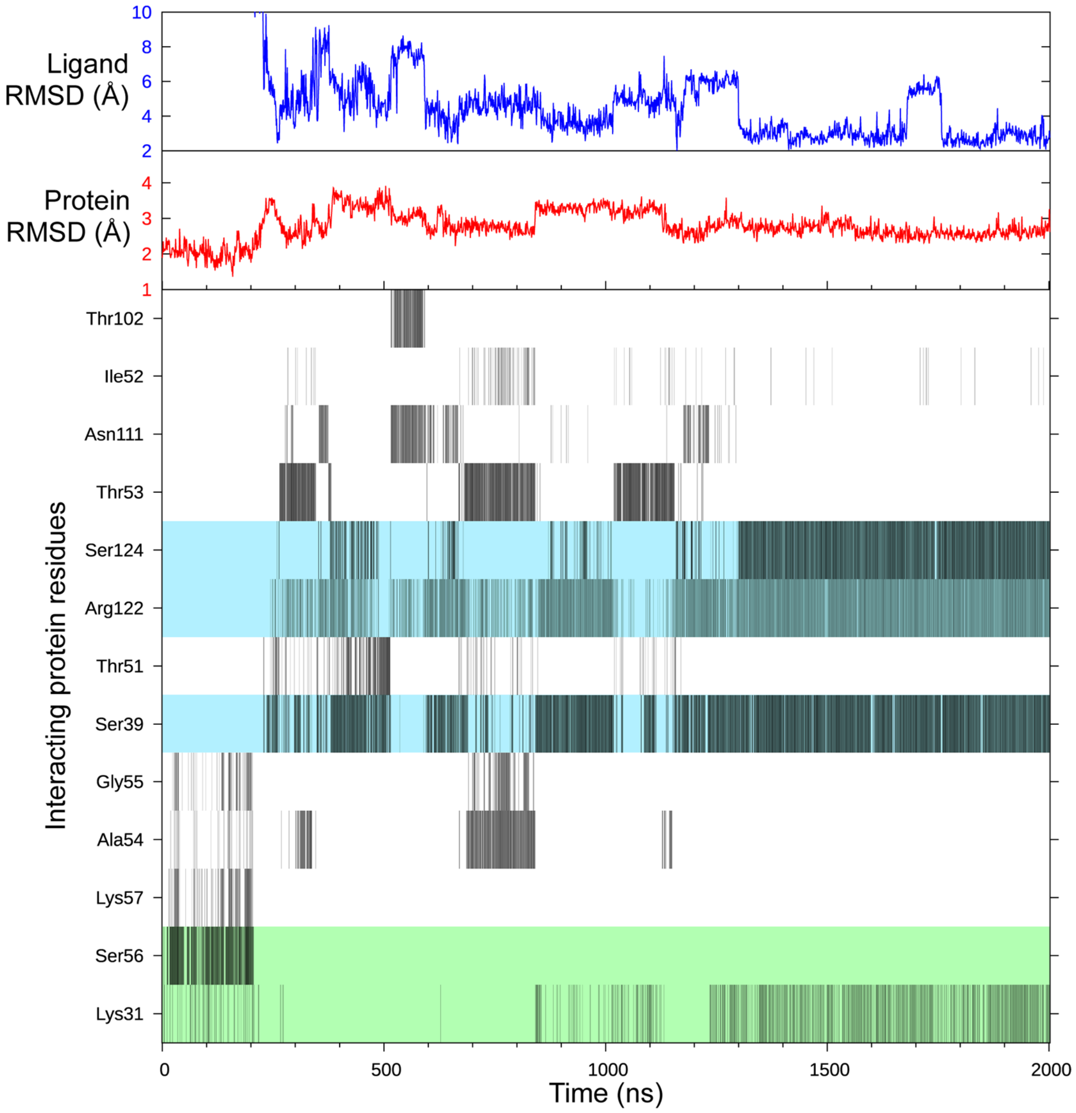


Figure 7. Interaction map of simulation 2-14, ligand B. The top two panels show RMSD of the ligand from the high-affinity pose (blue) and the RMSD of the protein backbone from the experimental pose (red). The lower 'barcode' panel shows contacts between the ligand and protein residues (ligand oxygen to protein heavy atom distance of <3.5 Å). Darker barcode lines indicate closer contact. Residues that interacted for less than 5% of the simulation time have been omitted. Residues are shaded based on their affiliation with the low-affinity binding site (green), or high-affinity binding site (blue).

Despite the frequency of ligand interactions at loopEF, there was no observed case of a ligand first interacting with loopEF and then proceeding to binding. Instead, individual simulations showed ligands proceeding from loopEF to loopCD or loopCD to the high-affinity binding site but never both. From analysis of the 2-ligand simulations, it was apparent that ligands preferred to interact at the 'earlier' loops: the low affinity binding site at loopCD and metastable site at loopEF, rather than the rarely occupied loopGH. Whilst the site at loopEF was the most frequently occupied, the site at loopCD seemed preferred when it was unobstructed by H2 (Figure 5). Similarly the position at loopGH was only occupied when the site at loopEF was unavailable. Thus preliminary observations suggest an order of site preference or a plausible binding pathway (Figure 8).

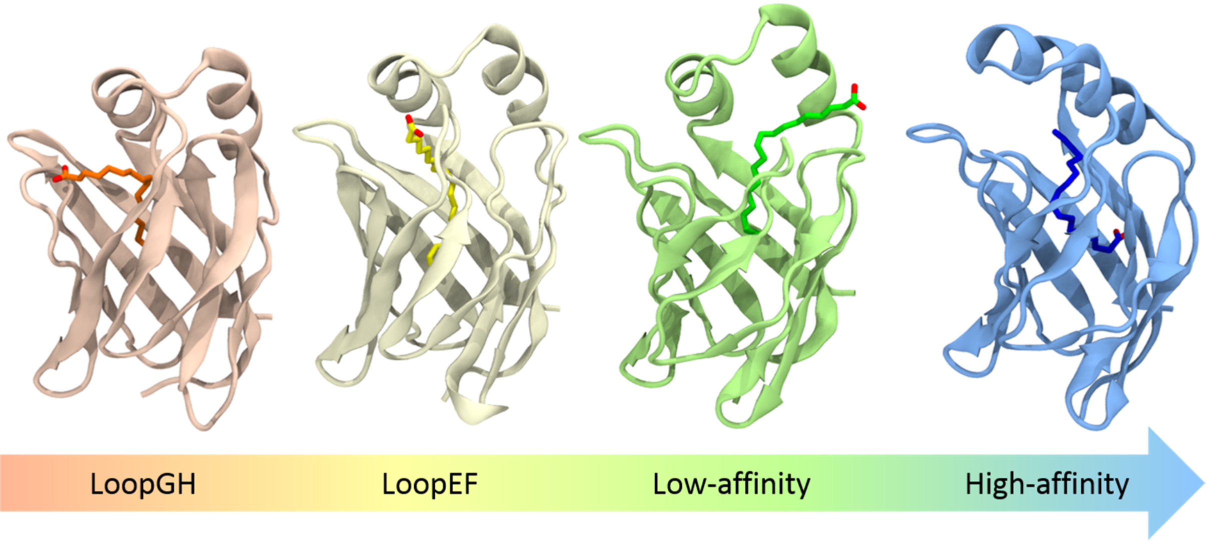


Figure 8. Proposed binding pathway for oleic acid to L-FABP based on long-timescale conventional MD simulations. The ligand begins interacting with loopGH (orange) and proceeds through loopEF (yellow), and the low-affinity binding site at loopCD (green), before forming a salt-bridge with Arg122 at the high-affinity binding site (blue).

Using a set of 30 long-timescale simulations we have been able to identify the key metastable states and transitions that occur on the binding pathway of oleic acid from solution to the high-affinity bound pose in L-FABP and we have used this information to propose a binding pathway. The simulations were able to reproduce the experimental structure of both the protein and the bound ligand to within 2 Å RMSD. We have performed a total of 55 µs of simulations, and the insight this data has provided us presents a significant advancement over previous simulations of ligand binding to L-FABP.

4.3.2 Markov state model for oleic acid binding to L-FABP

Although the long-timescale conventional MD simulations have provided us with great insight into ligand binding to L-FABP, many of the important transitions in binding are still rarely sampled and our proposed binding pathway is never observed in its entirety. To obtain a more detailed understanding of oleic acid binding to L-FABP we have developed a Markov state model that more completely characterizes the ligand-binding process. Expanding on the 2 µs of portal region-binding simulations discussed above, we performed 650 additional shorter simulations (each 200-400 ns) commencing from the high- and low-affinity binding sites and the metastable site and loopEF. These simulations total 160 µs of simulation time. An MSM was constructed from these new simulations as a transition network of 3000 states. Kinetically related states were then lumped together to produce a macrostate model, and poorly connected states were excised to produce a final 12-state MSM (Figure 9). The key metastable states shown in this MSM are the low-affinity binding site (green, states 1-3), the metastable site at loopEF (orange, states 10-12), and the high-affinity binding site (blue, state 5).

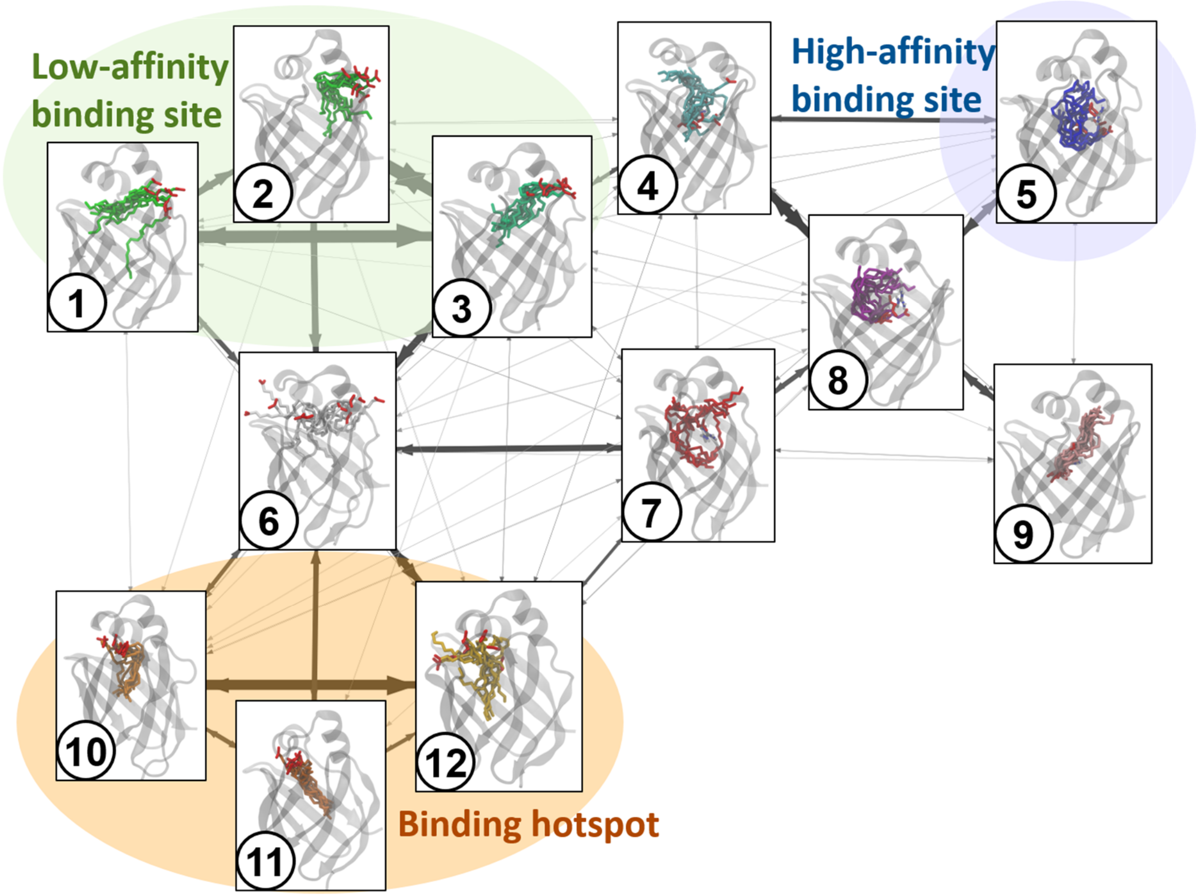


Figure 9. A 12 state MSM. Colored backgrounds indicate states affiliated with the low-affinity binding site (green), loopEF (orange), and high-affinity binding site (blue). Thicker arrows indicate a higher transition probability. States are numbered as they are referred to in the text.

The MSM identifies clearly defined states for the low-affinity binding site (green) as well as the metastable binding site at loopEF (orange). The model reproduces the previously observed transitions between the low-affinity binding site and loopEF, and shows that this transition does not proceed via any specific intermediate but rather proceeds through a loosely clustered state (state 6) where the ligand head group diffuses throughout the portal region while the tail-group remains buried within the β-barrel. In addition to the binding pathway revealed by the long-timescale simulations (states 1-3 > 4 > 5), the MSM also reveals a second major binding pathway, in which the ligand proceeds to salt-bridge formation from the metastable site at loopEF (states 10-12 > 7 > 8 > 5). The first step on both pathways is formation of the salt-bridge to Arg122 via state 4 or 7. The salt-bridge formation from the low-affinity binding site commonly only passes through state 4 on the way to the high-affinity binding site. Salt-bridge formation from loopEF requires additional steps of re-arrangement via state 7 and 8 before the ligand head-group arrives at the high-affinity pose.

Equilibrium populations can be estimated from the MSM, although the rarity of transitions between the states involved in the key step of salt-bridge formation leads to a poor estimate of the population of the high-affinity state. The MSM estimates the populations of the low-affinity and metastable site at loopEF to be equal, at 24% each, whilst the population of salt-bridge containing states is estimated to be only 42%. The low estimate of the bound population at the high affinity site is likely because the current model is the product of only 2 iterations of simulations.

4.4 Conclusion

We have conducted an investigation of ligand binding of oleic acid to the liver fatty acid-binding protein using conventional molecular dynamics simulations totaling 55 µs and extended this work to develop the first MSM of ligand binding to FABPs. Through our initial long-timescale simulations we identified several metastable binding sites in the portal and anti-portal regions of L-FABP. We also identified a plausible binding pathway in which the ligand head-group proceeds from a metastable site at loopEF, through the low-affinity binding site, enters the β-barrel, and forms a salt-bridge with Arg122, from which the ligand adopts the high-affinity pose. The metastable binding site at loopEF presents a novel bound mode that has not been observed experimentally and that may, through disruption of the interaction between Met74 in loopEF and Arg122, facilitate the transition of the receptor to the holo conformation required for ligand binding. This site provides an intriguing target to consider for structure-based drug design and could be used to optimize the kinetics of ligand transport by FABPs, thereby controlling the delivery of nuclear receptor targeting drugs.

Starting from the metastable binding sites identified in the long-timescale simulations, we performed an addition 160 µs of simulations and used this data to construct a Markov state model of oleic acid binding to L-FABP. This MSM supported the binding pathway proposed from the conventional simulations and also revealed a second major binding pathway, proceeding from the metastable site at loopEF. This MSM is an advance over conventional simulations. It clearly shows the series of transitions that form the binding pathways and enables us to estimate the relative populations of each metastable state, furthering our understanding of the ligand binding of oleic acid to the L-FABP.

The current MSM suffers from two limitations. The first of these limitations is that the model only covers the portion of the binding pathway after the ligand has associated with the portal region of the receptor. It was necessary to exclude the pre-association states of the ligand (where the ligand was freely diffusing in the solvent) from the MSM as these states cover far more geometric space than the bound state. In order to cluster freely diffusing states with a maximum intra-cluster size of 2 Å and adequately sample transitions between them, there would need to be multiple orders of magnitude more clusters and data. This could instead be handled by manually grouping these freely diffusing states into a single state, or identifying a more suitable metric for clustering. The second limitation of the model is that the rare transitions surrounding salt-bridge formation are undersampled. Additional sampling around the salt-bridge formation transitions is required to better model the kinetics of the high-affinity bound state. In this study we only performed one iteration of adaptive sampling. A third batch of simulations may solve this issue, but running several more iterations – with fewer simulations per batch – should be able to more efficiently produce a better model. Both of these limitations are specifically addressed in our subsequent work on the development of an MSM for the binding of haloperidol to the dopamine D_3 receptor.

4.5 References

(1) Copeland, R. A; Pompliano, D. L.; Meek, T. D. Drug-Target Residence Time and Its Implications for Lead Optimization. *Nat. Rev. Drug Discov.* **2006**, *5*, 730–739.

(2) 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.

(3) Durrant, J. D.; McCammon, J. A. Molecular Dynamics Simulations and Drug Discovery. *BMC Biol.* **2011**, *9*, 71.

(4) 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. U. S. A.* **2011**, *108*, 13118–13123.

(5) Kruse, A. C.; Hu, J.; Pan, A. C.; Arlow, D. H.; Rosenbaum, D. M.; Rosemond, E.; Green, H. F.; Liu, T.; Chae, P. S.; Dror, R. O.; Shaw, D. E.; Weis, W. I.; Wess, J.; Kobilka, B. K. Structure and Dynamics of the M3 Muscarinic Acetylcholine Receptor. *Nature* **2012**, *482*, 552–556.

(6) Pande, V. S.; Beauchamp, K.; Bowman, G. R. Everything You Wanted to Know about Markov State Models but Were Afraid to Ask. *Methods* **2010**, *52*, 99–105.

(7) Lane, T. J.; Shukla, D.; Beauchamp, K. A.; Pande, V. S. To Milliseconds and beyond: Challenges

in the Simulation of Protein Folding. *Curr. Opin. Struct. Biol.* **2013**, *23*, 58–65.

(8) Beauchamp, K. A.; Bowman, G. R.; Lane, T. J.; Maibaum, L.; Haque, I. S.; Pande, V. S. MSMBuilder2: Modeling Conformational Dynamics on the Picosecond to Millisecond Scale. *J. Chem. Theory Comput.* **2011**, *7*, 3412–3419.

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

(10) Bowman, G. R.; Huang, X.; Pande, V. S. Using Generalized Ensemble Simulations and Markov State Models to Identify Conformational States. *Methods* **2009**, *49*, 197–201.

(11) Furuhashi, M.; Hotamisligil, G. S. Fatty Acid-Binding Proteins: Role in Metabolic Diseases and Potential as Drug Targets. *Nat. Rev. Drug Discov.* **2008**, *7*, 489–503.

(12) Kaczocha, M.; Vivieca, S.; Sun, J.; Glaser, S. T.; Deutsch, D. G. Fatty Acid-Binding Proteins Transport N-Acylethanolamines to Nuclear Receptors and Are Targets of Endocannabinoid Transport Inhibitors. *J. Biol. Chem.* **2012**, *287*, 3415–3424.

(13) Wang, Y.-T.; Liu, C.-H.; Zhu, H.-L. Fatty Acid Binding Protein (FABP) Inhibitors: A Patent Review (2012-2015). *Expert Opin. Ther. Pat.* **2016**, *26*, 767–776.

(14) Sharma, A.; Sharma, A. Fatty Acid Induced Remodeling within the Human Liver Fatty Acid-Binding Protein. *J. Biol. Chem.* **2011**, *286*, 31924–31928.

(15) Cai, J.; Lücke, C.; Chen, Z.; Qiao, Y.; Klimtchuk, E.; Hamilton, J. A. Solution Structure and Backbone Dynamics of Human Liver Fatty Acid Binding Protein: Fatty Acid Binding Revisited. *Biophys. J.* **2012**, *102*, 2585–2594.

(16) Banaszak, L.; Winter, N.; Xu, Z.; Bernlohr, D. A.; Cowan, S.; Jones, T. A. Lipid-Binding Proteins: A Family of Fatty Acid and Retinoid Transport Proteins. *Adv. Protein Chem.* **1994**, *45*, 89–151.

(17) Rich, M. R.; Evans, J. S. Molecular Dynamics Simulations of Adipocyte Lipid-Binding Protein: Effect of Electrostatics and Acyl Chain Unsaturation. *Biochemistry* **1996**, *35*, 1506–1515.

(18) Li, Y.; Li, X.; Dong, Z. Concerted Dynamic Motions of an FABP4 Model and Its Ligands Revealed by Microsecond Molecular Dynamics Simulations. *Biochemistry* **2014**, *53*, 6409–6417.

(19) Friedman, R.; Nachliel, E.; Gutman, M.; Ran Friedman; Esther Nachliel, A.; Gutman*, M. Molecular Dynamics Simulations of the Adipocyte Lipid Binding Protein Reveal a Novel Entry Site for the Ligand. *Biochemistry* **2005**, *44*, 4275–4283.

(20) Levin, L. B.-A.; Ganoth, A.; Amram, S.; Nachliel, E.; Gutman, M.; Tsfadia, Y. Insight into the Interaction Sites between Fatty Acid Binding Proteins and Their Ligands. *J. Mol. Model.* **2010**, *16*, 929–938.

(21) Mihajlovic, M.; Lazaridis, T. Modeling Fatty Acid Delivery from Intestinal Fatty Acid Binding Protein to a Membrane. *Protein Sci.* **2007**, *16*, 2042–2055.

(22) Tsfadia, Y.; Friedman, R.; Kadmon, J.; Selzer, A.; Nachliel, E.; Gutman, M. Molecular Dynamics Simulations of Palmitate Entry into the Hydrophobic Pocket of the Fatty Acid Binding Protein. *FEBS Lett.* **2007**, *581*, 1243–1247.

(23) Long, D.; Mu, Y.; Yang, D. Molecular Dynamics Simulation of Ligand Dissociation from Liver Fatty Acid Binding Protein. *PLoS One* **2009**, *4*, e6081.

(24) Pronk, S.; Páll, S.; Schulz, R.; Larsson, P.; Bjelkmar, P.; Apostolov, R.; Shirts, M. R.; Smith, J. C.; Kasson, P. M.; van der Spoel, D.; Hess, B.; Lindahl, E. GROMACS 4.5: A High-Throughput and Highly Parallel Open Source Molecular Simulation Toolkit. *Bioinformatics* **2013**, *29*, 845–854.

(25) Schmid, N.; Eichenberger, A. P.; Choutko, A.; Riniker, S.; Winger, M.; Mark, A. E.; Van Gunsteren, W. F. Definition and Testing of the GROMOS Force-Field Versions 54A7 and 54B7. *Eur. Biophys. J.* **2011**, *40*, 843–856.

(26) Bachar, M.; Brunelle, P.; Tieleman, D. P.; Rauk, A. Molecular Dynamics Simulation of a Polyunsaturated Lipid Bilayer Susceptible to Lipid Peroxidation. *J. Phys. Chem. B* **2004**, *108*, 7170–7179.

(27) Martinez-Seara, H.; Róg, T.; Karttunen, M.; Reigada, R.; Vattulainen, I. Influence of Cis Double-Bond Parametrization on Lipid Membrane Properties: How Seemingly Insignificant Details in Force-Field Change Even Qualitative Trends. *J. Chem. Phys.* **2008**, *129*, 105103.

(28) Humphrey, W.; Dalke, A.; Schulten, K. VMD: Visual Molecular Dynamics. *J. Mol. Graph.* **1996**, *14*, 33–38.

(29) Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G. A Smooth Particle Mesh Ewald Method. *J. Chem. Phys.* **1995**, *103*, 8577–8593.

(30) Hess, B.; Bekker, H.; Berendsen, H. J. C.; Fraaije, J. G. E. M. LINCS: A Linear Constraint Solver for Molecular Simulations. *J. Comput. Chem.* **1997**, *18*, 1463–1472.

(31) Bussi, G.; Donadio, D.; Parrinello, M. Canonical Sampling through Velocity Rescaling. *J. Chem. Phys.* **2007**, *126*, 014101.

(32) Cronkite-Ratcliff, B.; Pande, V. MSMExplorer: Visualizing Markov State Models for Biomolecule Folding Simulations. *Bioinformatics* **2013**, *29*, 950–952.

(33) Bowman, G. R. A Tutorial on Building Markov State Models with MSMBuilder and Coarse-Graining Them with BACE. In *Methods in molecular biology (Clifton, N.J.)*; 2014; Vol. 1084, pp 141–158.

(34) Bowman, G. R. Improved Coarse-Graining of Markov State Models via Explicit Consideration of Statistical Uncertainty. *J. Chem. Phys.* **2012**, *137*, 134111.

(35) Ogata, R. T. Thermodynamic and Kinetic Properties of Fatty Acid Interactions with Rat Liver Fatty Acid-Binding Protein. *J. Biol. Chem.* **1996**, *271*, 31068–31074.

(36) Falomir-Lockhart, L. J.; Laborde, L.; Kahn, P. C.; Storch, J.; Córsico, B. Protein-Membrane Interaction and Fatty Acid Transfer from Intestinal Fatty Acid-Binding Protein to Membranes: Support for a Multistep Process. *J. Biol. Chem.* **2006**, *281*, 13979–13989.

(37) Steele, R. A; Emmert, D. A; Kao, J.; Hodsdon, M. E.; Frieden, C.; Cistola, D. P. The Three-Dimensional Structure of a Helix-Less Variant of Intestinal Fatty Acid-Binding Protein. *Protein Sci.* **1998**, *7*, 1332–1339.

Chapter 5

Markov State Model Analysis of Haloperidol Binding to the D_3 Dopamine Receptor

Having previously developed an understanding of the behavior of haloperidol binding to the D_3R (in Chapter 3) and a methodology for constructing ligand-binding Markov state models (in Chapter 4), we set out to construct an MSM of the D_3R-haloperidol system. Ligand binding MSMs had previously only been constructed for relatively simple systems and, although MSMs had previously been used to study the activation dynamics of GPCRs, they had never been used to study the binding pathway of a GPCR ligand. The construction of a Markov state model of ligand binding to a challenging GPCR system, and the development of the required methodology, would present a significant advancement in the field.

The beginning of this work coincided with significant improvements to the available MSM software, and we were able to far more easily implement our own code and further develop our methodology. We developed effective descriptors of the system that allowed us to account for the flexibility and hydrogen-bonding character of our ligand and developed an adaptive sampling methodology that enabled each iterative batch of simulations to reduce errors in undersampled transitions and to further the exploration of the binding ensemble. This work presents the final MSM produced, but the methodology was developed over a long series of datasets, in which we identified flaws in the methodology and solved them for the next iteration.

In Chapter 4 we encountered difficulties in describing the flexibility of the system and only implemented a minimal adaptive sampling methodology, which was not able to correct the most undersampled transitions. The methodology presented here is a significant advancement over our earlier work and is suitable for application to other complicated ligand-binding systems, providing a significant advancement to the field. Using MSMs we were able to explore the ligand binding pathways of haloperidol and support our findings with robust statistics. Where in Chapter 3 we were able to produce a binding pathway for haloperidol, here we improve upon that work by exploring the ensemble of the binding complex and predicting the relative importance of the many metastable states.

This chapter is a manuscript to be submitted for publication.

Markov State Model Analysis of Haloperidol Binding to the D_3 Dopamine Receptor

Trayder Thomas[a], Elizabeth Yuriev[a,1], David K. Chalmers[a,1]

[a] Medicinal Chemistry, Monash Institute of Pharmaceutical Sciences, Monash University, 381 Royal Pde., Parkville, Victoria 3052, Australia

[1]To whom correspondence may be addressed. Email: [redacted] or Elizabeth [redacted]

Introduction

It has long been recognized that drug-binding involves more than just the endogenous binding site and that drugs must first undergo association and recognition steps prior to reaching the high affinity bound state. The early stages of binding typically occur too fast to observe experimentally, leading to the use of the general approximation that drug binding is a one-step ‘on or off’ process. This simplification to a single parameter is, surprisingly, often adequate for drug development, but for many targets the off-rate of a drug has proven more important than the affinity as a whole (1–3). In reality there are a multitude of steps to the binding process any of which could be the rate determining step. An agonist may have no effect prior to arriving at its final bound pose, yet an antagonist can block access to a binding site prior to its own arrival, and either could have additional allosteric modes of action. There is simply not enough information in a two-step approximation to explain the complexities of the binding pathway that may result in unique ligand behaviors.

In contrast to laboratory experiments, molecular dynamics (MD) simulations can expose drug binding processes in atomic detail, revealing metastable binding sites along the drug-binding pathway where the drug pauses during the binding process. Using a series of microsecond timescale simulations, Dror et al. investigated the binding of 3 antagonists, and 1 agonist to the β_2AR, and observed that the binding pathways were characterized by a common series of metastable states (4). Additionally, they also found that the primary barrier to binding was located far outside the binding pocket and involved the dewetting of the binding site so the key salt-bridge could form. A similar investigation of tiotropium binding to the M_2/M_3 muscarinic receptors found an analogous barrier and suggested that the difference in the rate of tiotropium dissociation from the two receptors was due to differences in this energy barrier rather than the highly homologous orthosteric sites (5). Conversely, our own studies of haloperidol binding to the D_2 and D_3 dopamine receptors suggested that different binding pathways are present in the two receptors despite the fact that haloperidol is non-selective (6). Studying drug binding through unbiased simulation can be considered a brute force approach due to their lack of efficiency and often requires computational power that is still out of reach of most researchers (7).

Markov state models (MSMs) are general statistical models which estimate the transition probabilities in a system of states by observing the behavior of that system over time. When applied to large sets of conventional, unbiased MD simulations, MSMs can increase both the computational efficiency and accessible timescales. By running many shorter – and therefore efficiently parallelizable – simulations, MSMs can predict a transition probability that can be extrapolated to longer, otherwise inaccessible timescales.

MSMs are at their simplest (8, 9) a network of transition probabilities between a set of discrete states. These states can be imagined as the local minima in an *n*-dimensional energy landscape that are separated by energy barriers of various heights. The probability of transition between two adjacent states is governed by the height of the energy barrier and conversely, the height of the barrier can be determined by measuring the frequency of transitions between pairs of states. The theory of MSMs depends on the Markov property of history independence. In order to determine the height of an energy barrier to transition out of a state, the system must be allowed to reach an equilibrium within that state, thereby ‘forgetting’ where in the state it entered. For this reason, transitions between states are counted after a “lag time” that allows the system to equilibrate, and these counts are used to calculate the transition probabilities from which the kinetics are derived. For more detailed theory on MSMs the reader is recommended references (10, 11).

MSMs have been extensively applied in computational biology to elucidate protein folding pathways, an area in which they have had an impressive impact (12). The use of MSMs in studying ligand binding pathways has been far more limited, due to both the increased complexity of treating the ligand and protein separately, and the disparate timescales on which ligand-binding and protein conformational changes occur. The key to overcoming these difficulties is in the methodology used to discretize the coordinate space of the molecular system into individual states. If the system is discretized too coarsely, intra-state energy barriers prevent the equilibration of the system within a reasonable lag time. If the system is discretized too finely, a prohibitive amount of data is needed to reach a statistical certainty for each transition probability. Discretizing the coordinate space then becomes an optimization problem, a set of arbitrary dimensions must be combined to form an energy landscape that eliminates noise whilst clearly representing the important minima.

Once a suitable set of features has been chosen, dimensionality reduction algorithms can be used to reduce the large volume of data and focus the system on the kinetically relevant movements of the protein or ligand. Time-structure independent component analysis (tICA) is a dimensionality reduction recently applied to useful effect in several MSM studies (13–15). tICA identifies the combinations of correlated components that result in the longest time-scale movements of the system, essentially decoupling geometrically large yet kinetically irrelevant movements such as the diffusion of the ligand through the solvent, or the flapping of protein termini. In this way, the raw MD data, a set of coordinates for every atom, most of which are redundant, or irrelevant to this ligand binding pathway, can be efficiently denoised. It is possible, through a combination of feature selection and further dimensionality reduction, to usefully represent the states of system composed of many thousands of atoms using a handful of carefully chosen dimensions.

To date, most of the methodology developed for ligand binding MSMs has been confined to systems that are adequately described using just one or two simple descriptors. The benzamidine-trypsin complex has been used extensively as a test case. This system was initially studied by approximating the receptor as a rigid body and representing the ligand position with a single atom (16). It was subsequently used as a test case for developing adaptive sampling methodology (17) and later re-investigated on much longer timescales to observe protein motions (18). The binding of phosphate to phosphate binding protein has also been studied using similar approximations (19). More flexible systems also benefit from simple descriptors. It was found that the large-scale "clam-like" domain movements of the LAO protein during L-arginine binding were able to be simply approximated with two angles to describe the opening and twisting of the two domains (20).

While simple descriptors are ideal for simple systems, they lack the descriptive power to adequately describe more complex complexes. Representation of the ligand as a single atom or rigid body is unsuitable for flexible ligands, where the orientation and conformation of the ligand have a greater impact on the kinetics. Similarly the approximation of the protein as a rigid body is blind to energetic barriers due to protein flexibility. In order to construct a ligand-binding MSM of a relatively complicated G protein-coupled receptor (GPCR) system, we have found that more sophisticated descriptors are required.

G protein-coupled receptors are major pharmaceutical targets, and are amongst the most-studied classes of proteins. GPCRs consist of a bundle of 7 trans-membrane helices which bind to small-molecule ligands in their extra-cellular binding pockets and transduce signals through binding with intra-cellular G proteins. Whilst MSMs have been used to explore the conformational ensemble of the β_2AR receptor (21), there have been no similar studies of dopamine receptors nor any studies focusing on the behavior of the ligand during the binding process. Thus we have conducted a study of the inverse-agonist haloperidol (Figure 1) binding to the inactive state of the D_3 dopamine receptor. Haloperidol has a residence time of 58 minutes in the highly homologous D_2R (22), which leads to a challenging system to model. This study greatly extends from our previous work, where we simulated a binding pathway for haloperidol through conventional MD approaches (6).

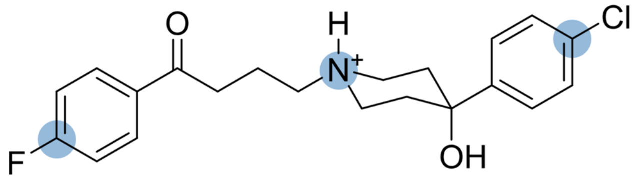


Figure 1. Haloperidol. The atoms used to define the ligand position during analysis are highlighted in blue.

In this work we demonstrate the application of MSMs to create a model for ligand binding to a GPCR that incorporates both protein and ligand flexibility, revealing the ligand binding pathways and identifying intermediate metastable sites. Our methodology utilizes adaptive sampling to increase the computational efficiency, and decrease the statistical error. This is the first MSM of ligand-binding to a GPCR that focuses on the behavior of the ligand and we anticipate that this methodology will be useful for a wide range of systems.

Methods

Molecular Dynamics Simulations.

Molecular simulations were performed with GROMACS 5.0 (23) using the GROMOS 54a7 force field (24, 25). Parameters for haloperidol were calculated using Automated Topology Builder (26–28) with additional symmetry corrections. The coordinates of the initial system were those of the pre-equilibrated D_3 dopamine receptor embedded in a POPC lipid bilayer developed in our previous work (6). The protein termini, including those created by truncating loop ICL3, were charged and heavy hydrogens were used to allow a 5 fs time step (29). The bilayer and protein were resolvated into an orthorhombic dodecahedron unit cell with the insane script (30) such that there was a distance of 30 Å between bilayers. The initial position of haloperidol was as far from the protein as possible. The resultant system contained ~5000 water molecules, 0.15 M salt, and 80 lipids. The temperature of all simulations was 310 K. Following an initial 2000 step minimization the system was equilibrated for 1 ns using the NVT ensemble with the Berendsen thermostat, for 1 ns using the NPT ensemble with Berendsen thermostat and barostat, and 20 ns with the NPT ensemble using v-rescale thermostat and Parrinello-Rahman barostat. The protein backbone and ligand were restrained during all equilibration steps. The initial simulations began from this equilibrated system. Production simulations commenced with a 2000 step minimization. Random initial velocities were applied and each simulation was performed for 200 ns with an unbiased NPT ensemble using the v-rescale thermostat, Parinello-Rahman barostat.

Development of Markov State Models Using Adaptive Sampling.

Markov state models were constructed using MSMBuilder (31). The final models were developed using an iterative, adaptive sampling process using a total of 519 production runs totaling 121800 ns. Adaptive sampling utilized an exploration phase and a refinement phase, each phase was achieved using the process outlined in Figure 2.

(A) An initial batch of 100 simulations was performed each starting with the ligand located in the solvent. Of these simulations, 53 finished with the ligand in the extracellular vestibule. The final frames of these trajectories were used as the starting structures for the first iteration of simulations. The initial set of 100 simulations was then put aside and only re-incorporated at the refinement stage, to avoid cases where the ligand explored the solvent or bilayer excessively.

(B) Following the initial simulations from the solvent, iterations were performed using sets of 20-30 concurrent simulations. The starting frames were selected according to the adaptive sampling criteria given below.

(C) The high-dimension system coordinates were reduced by treating the protein and ligand separately, creating a set of features for each of the protein and ligand. The protein was featurized based on the coordinates of each α-carbon. The ligand was featurized as a binary protein contact fingerprint with additional bits for hydrogen bonds. The ligand location was first simplified using the 3 atoms highlighted in Figure 1: the 2 carbons attached to halogens and the central nitrogen atom. One bit (true or false) of the fingerprint was assigned for each of these three atoms to each residue in the extra-cellular half of the receptor (483 bits). Where the ligand atom was within 6 Å of a protein residue heavy atom, the corresponding bit was assigned as true. An additional 54 bits were used to encode the presence of every possible hydrogen bond between the ligand and every non-loop residue in the extracellular vestibule. All –NH and –OH groups were treated as both acceptor and donor, and both backbone and sidechains were considered for the protein. A hydrogen bond was defined as being within 2.5 Å and 120° (Baker-Hubbard definition) (32).

(D) To further reduce the dimensionality of the system, tICA (lag time 500 ps) was applied separately to the protein and ligand featurizations, 3 dimensions were retained from the protein and 7 from the ligand. The reduced protein and ligand coordinates were then combined to describe the protein-ligand complex in 10-dimensional space.

(E) The tICA dimensions were clustered coarsely with k-means clustering (initially with 50 clusters), assigning each frame of data to a cluster. As more space was explored, additional clusters were added, totaling 200 clusters at the end of the exploration phase.

(F) An MSM with a lag time of 1 frame (50 ps) was constructed from the clustered data to estimate the transition network. This transition network was analyzed to determine appropriate starting structures for the next iteration. The contribution of each state to the uncertainty of each eigenvalue was calculated and the three states with the highest contributions to the first ten eigenvalues were selected as starting states for the next iteration. States contributing to the uncertainty of multiple eigenvalues were selected only once and this typically led to around 20 unique states. This procedure effectively selects under-sampled states that are related to the slowest processes. Starting structures for the subsequent iteration were selected from each of the under-sampled states by randomly picking conformations that were closer to the cluster center than the mean distance of all cluster members, ensuring that the selected conformations are representative of the cluster and preventing the selection of outliers as a starting point for the next iteration. As k-means clustering tends to place cluster centers in densely sampled regions, and the exploration of new space is a relatively rare event, additional starting conformations were manually selected following visual inspection of the trajectories. Where a trajectory transitioned to new space, the last frame of the trajectory was used in triplicate as a starting conformation in the next iteration.

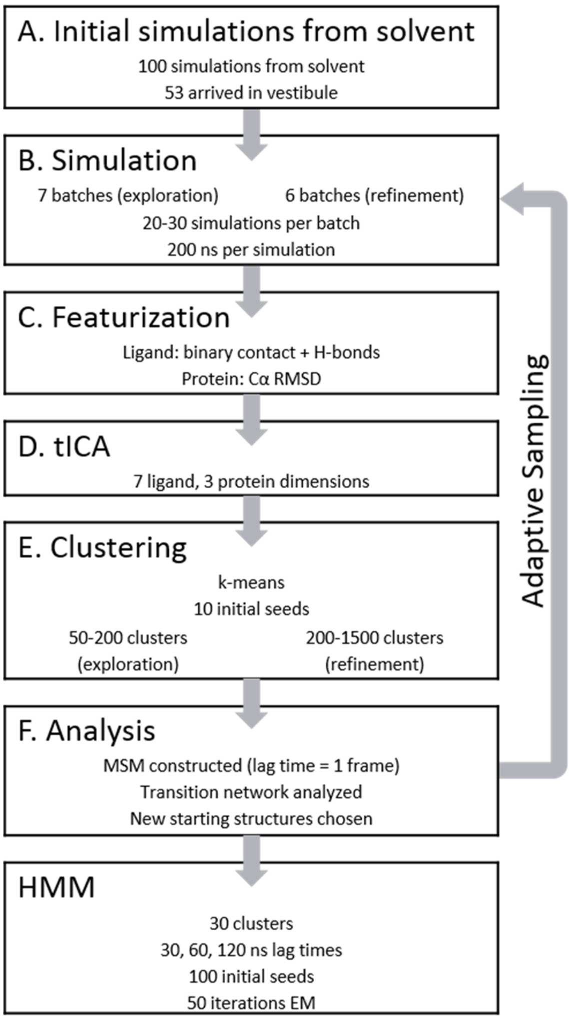

A. Initial simulations from solvent
100 simulations from solvent
53 arrived in vestibule
B. Simulation
7 batches (exploration)
6 batches (refinement)
20-30 simulations per batch
200 ns per simulation
C. Featurization
Ligand: binary contact + H-bonds
Protein: Cα RMSD
D. tICA
7 ligand, 3 protein dimensions
E. Clustering
k-means
10 initial seeds
50-200 clusters (exploration)
200-1500 clusters (refinement)
F. Analysis
MSM constructed (lag time = 1 frame)
Transition network analyzed
New starting structures chosen
HMM
30 clusters
30, 60, 120 ns lag times
100 initial seeds
50 iterations EM
Adaptive Sampling

Figure 2. Flowchart illustrating the key processes in the development of Markov state and hidden Markov models for the binding of haloperidol to the dopamine D_3 receptor.

The exploration phase of adaptive sampling was continued until no new conformational space was being discovered (7 iterations). At this point, the focus of each iteration was switched to refinement of the model. In the subsequent 4 iterations, the number of clusters used in step E was increased to 300, 500, 1000 and 1500, allowing the identification of tighter clusters with few non-self transitions. In the refinement phase, as in the exploration phase, the starting conformations for the subsequent iteration were selected based on each state's contribution to the overall model uncertainty. Additional starting conformations were selected from the states with the lowest number of non-self transitions.

Hidden Markov Model

Following the adaptive sampling procedure, we used the entire set of MD trajectories to construct a Hidden Markov Model (HMM). HMMs use many fewer clusters than MSMs and produce a coarser, macrostate model. They also reduce the discretization error during clustering by interlinking the clustering and construction of the transition network into a single iterative step. The HMM was generated using a 30 ns lag time and the best model, in terms of log probability, was chosen from 100 trial models, each generated from different initial k-means clustering, and each trial using 50 iterations of the expectation-maximization algorithm.

Analysis of our initial HMM revealed problems that had not been identifiable in the MSM. Groups of states were found that, although internally well connected, were not well connected to the rest of the network. To address these problems, additional simulations were performed to improve sampling of transitions between the poorly connected regions of the transition network. The least connected states were identified by sequentially merging the most connected pairs of states together until only one state remained. A connectivity score was assigned to each pair of states which was defined as the lower value of the forward or backward transition probability. States with the highest connectivity score were merged first. The 3 least connected groups of states were identified and sets of 10 additional simulations were performed from the poorly connected states using starting structures that were drawn from the cluster in each group that was most connected to the rest of the network. This refinement process was repeated for 2 further iterations, after which one group of states still remained poorly connected. This poorly connected group of states was estimated by a HMM to contain >95% of the total population, vastly overwhelming our expected bound state. Therefore, to better establish the true populations of each state, 5 new starting structures were drawn from the expected bound state (state A), and the poorly connected group of states (state B). The ligand in both of these states formed a salt-bridge with Asp3.32 but each state arose from an opposite alignments of the ligand in the extracellular vestibule. A set of 2 µs simulations was performed from each state and the resultant 2 µs trajectories were added to create an expanded dataset.

Our final HMM model was selected from a set of 30-state HMMs generated using 30, 60, and 120 ns lag times. State B remained the dominant state in each model, although state A still retained a significant share of the population (75% to 14% in the 120 ns lag time model). The most likely state assignment from the 120 ns lag time HMM was used to generate a series of MSMs, and the resultant implied timescales for the slowest processes in the system are shown in Figure S1. To further characterize the populations of the 2 states, 10 bootstrapped datasets were generated for each model by randomly drawing entire trajectories with replacement from the full dataset until each bootstrapped dataset matched the size of the original dataset. A new HMM was constructed for each bootstrapped dataset, 10 initial seeds of k-means clustering were used, each with 30 iterations of the expectation-maximization algorithm.

Results and Discussion

Markov state models are an effective tool for exploring complex molecular processes but, to date, their use in studies of ligand binding have been limited to simple ligand binding processes. In this work, we have constructed a series of hidden Markov models for the more complex binding of haloperidol to the deeply buried binding pocket of the dopamine D3 receptor. The models were built from a dataset of 509 individual 200 ns MD simulations of haloperidol binding to the D_3R, augmented with a further ten 2 µs simulations and assign the ligand-protein system into 30 states. As part of this process, we built models using 30, 60 and 120 ns lag times to investigate the impact of this parameter on the HMM. In general, shorter lag times provide more structural detail about the binding, but can underestimate the populations of important states, while simulations with longer lag times obscure intermediate states but provide better population estimates. Accordingly the populations and bootstrapped data presented here are for the 120 ns HMM and the network diagram presented in Figure 3 is for the 30 ns lag time HMM. The network diagram for the 120 ns model is presented in Figure S2.

Figure 3 maps the system ensemble, showing the 30 states of the HMM, color coded into kinetically related groupings. The binding process begins with the ligand in free and loosely bound ‘solvent’ states (cyan) which then progress to a set of states where the ligand occupies the extracellcular vestibule, these states rapidly intertransition and remain uncolored in the network. From the vestibule, productive binding can be grouped into two broad pathways which depend on the orientation of the ligand; when the haloperidol F-ring is directed towards TM 5 or 6, haloperidol follows pathway A (orange) to deeply buried state A; when the F-ring is oriented towards TM 1, 2 and 7, haloperidol follows pathway B (blue) to buried state B. State C is also present near the end of pathway B and, although very closely related to state B, exhibits a notably different conformation of haloperidol. States that do not lead to deep binding are also observed (kinetic traps) denoted as trap 1 (red) and trap 2 (green). A group of states was also observed where haloperidol interacts in the secondary binding pocket (pink).

The most important states in this model, from a drug design perspective, are state A and state B. These states are metastable beyond a microsecond timescale, whereas other states in the model have a lifetime of nanoseconds. To estimate the metastability of states A and B, we performed a batch of five 2 µs simulations beginning from each state. In these 20 cumulative microseconds of simulation, no transitions away from state A or state B were observed, suggesting that both of these states make a significant contribution to the residence time of haloperidol. Longer timescales would be required to determine which of these states contribute to haloperidol’s efficacy as an inverse agonist.

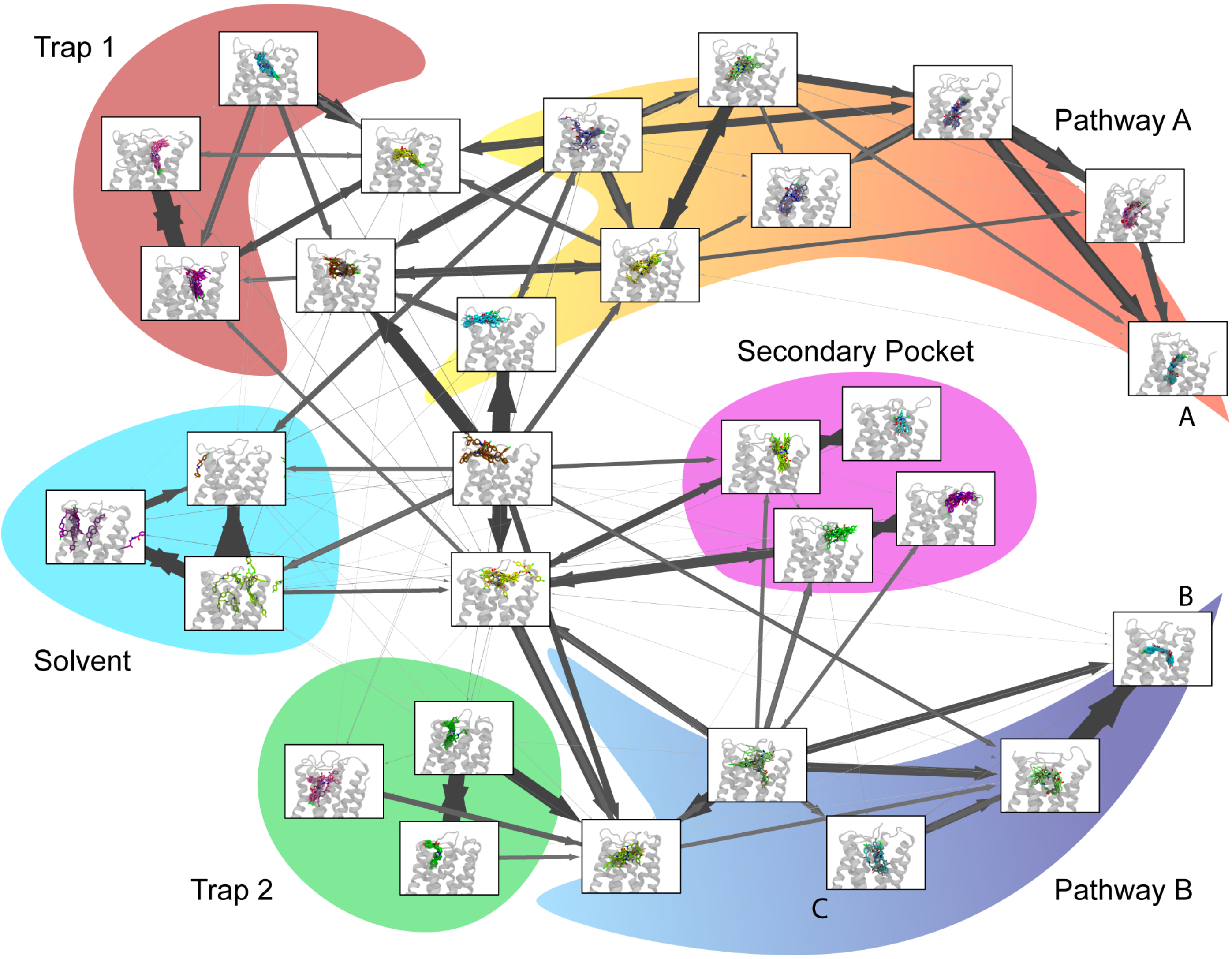


Figure 3. Transition network of the final HMM built using a 30 ns lag time, showing the progress of the ligand from the solvent state (left) to the bound states (right). Images for each state show ten, randomly selected, representative ligand conformations. Thicker arrows indicate a higher transition probabilities. States in pathway A and B are ordered based on their committor probability towards each bound state.

Figure 4 shows representative bound conformations of key states. In deeply bound state A, the haloperidol nitrogen forms a salt bridge with Asp3.32 while the F-ring sits deep in the receptor alongside Trp6.42, the Cl-ring rests between TM 2-3 while the hydroxyl group remains exposed to the solvent in the vestibule. This state was identified in our previous unbiased MD study of haloperidol binding (6). State B is a novel bound state that has the highest equilibrium population in each HMM. Here haloperidol lies horizontally across the top of Asp3.32 with the hydroxyl group forming hydrogen bonds to ECL2 whilst the F-ring dips into a pocket formed by TMs 1, 2 and 7, the Cl-ring stacks with His6.55, and the F-ring sits alongside Tyr7.43, which maintains a hydrogen bond with Asp3.32. State C is very closely kinetically related to State B but the F-ring sits deeper in the receptor, and haloperidol straightens out to adopt a conformation that more closely resembles state A.

Additional metastable states were identified in the form of kinetic traps 1 and 2, and a group of states in the secondary binding pocket. In trap 1, the hydroxyl group of haloperidol is hydrogen bonded with Asp3.32 while the Cl-ring sits in the pocket between TM 1, 2, and 7 and the F-ring points into the vestibule entrance. In kinetic trap 2 the hydroxyl group makes interactions with Asp3.32 and haloperidol sits nearly parallel with TM5 whilst the F-ring interacts with ECL2. In the secondary binding pocket haloperidol adopted 2 kinetically distinct orientations, either interacting with Glu2.65 from within the vestibule with the F-ring projecting into the bilayer between TM 1 and 7, or interacting with Glu2.65 from above with the Cl-ring directed towards TM 1.

In addition to identifying kinetically distinct states, Markov models also calculate the flux between states, allowing the characterization of binding pathways. Figure 3 reveals distinct pathways from the solvent to the deeply bound states A and B. Pathway A commonly starts with haloperidol in the vestibule, located within the polar region formed by ECL2, His6.55, and Asn6.58 and with the F-ring of haloperidol directed towards TM 5 and 6. From here, haloperidol drops deeper into the receptor forming the well-known salt-bridge with Asp3.32. The F-ring briefly interacts with the aromatic network between TM 5 and 6 before haloperidol swivels deeper into the receptor and arrives at state A. Pathway B typically begins with haloperidol aligned opposite to the starting orientation for pathway A, with the F-ring pointed towards TM 1, 2 and 7. Haloperidol then drops deeper into the receptor, taking up a seesaw-like position above Asp3.32. When the Cl-ring drops deeper into the receptor haloperidol enters kinetic trap 2, whereas when the F-ring drops deeper into the receptor haloperidol enters state B.

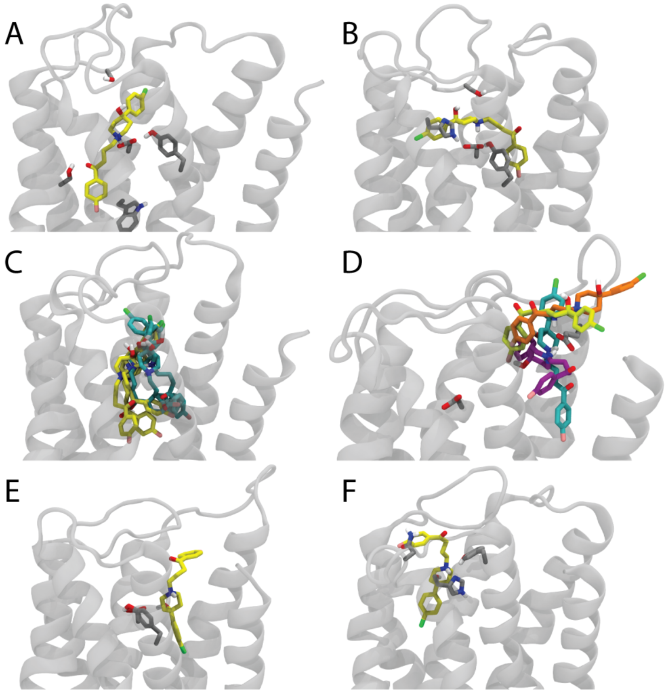


Figure 4. Representative structures drawn from the key states presented in Figure 3. (A) State A, (B) state B, (C) state C (cyan) compared to similar conformations from pathway A (yellow), (D) conformations of haloperidol from each state in the secondary binding pocket, (E) trap 1, (F) trap 2.

The kinetic data also provides information about the equilibrium populations of each state. The 120 ns lag time HMM predicts the equilibrium populations of the deeply bound states A and B to be 14% and 75% of the total population respectively. State C, being closely related to state B, is predicted to have an equilibrium population of 10%. To assess the robustness of these predictions, we conducted a bootstrap analysis, creating synthetic datasets by resampling from the full dataset which results in the omission of some trajectories. Because the HMM produced using each synthetic dataset has a unique discretization, the set of substates that form each bound state differs in every bootstrapped model. Accordingly, we made a manual comparison between bootstrapped models where we identified states A and B based on formation of the ligand salt-bridge and the positions of the ligand aromatic rings, because state C was not always separated from state B the two populations were combined. A summary of the bootstrapped data is given in Table 1, and the full set can be found in SI Table 1. The bootstrap analysis confirmed that the model is robust. All but one bootstrapped model assign most of the population to states A and B and, in the majority of models, state B is the highest population state. This analysis also reveals a sensitivity to omitted trajectories that can lead to state A being predicted as the highest population state, suggesting that, although a healthy number of transitions into each state were observed, these only occurred in a small number of trajectories. Using shorter trajectories would reduce this sensitivity, however this would also reduce the length of the lag times that could be used.

		Populations	
		State A	**State B**
All data		23.3%	74.7%
Bootstrap	1	0.0%	88.4%
	2	3.2%	61.2%
	3	28.2%	49.5%
	4	42.3%	32.0%
	5	32.2%	53.0%
	6	60.0%	36.3%
	7	40.5%	47.6%
	8	16.2%	2.9%
	9	17.0%	57.0%
	10	46.6%	24.6%
Mean		28.6%	45.2%
Std EOM		18.5%	21.9%
Std dev		6.2%	7.3%

Table 1. Populations of states A and B for 120 ns lag time HMMs generated from all data and 10 bootstrap datasets. Only bootstrapped datasets are used to calculate mean and errors.

Free energy surfaces are a useful way to visualize the state decomposition and general connectivity of the data. The free energy surfaces for the slowest tICs are shown in Figure 5 and free energy surfaces for the remaining tICs are shown in Figure S3. These surfaces show a good connectivity between the data for any given pair of tICA dimensions. The centroid of each state in the 30 ns HMM is shown on the free energy surface and colored as per Figure 3. An examination of the ligand tICs shows that the 2 slowest tICs loosely correlate to pathway A and pathway B respectively. The free energy surface also gives a good indication of correlation between the slowest modes of the system. Figure 5 shows that, as we would expect, pathway A and B are largely uncorrelated and mutually exclusive. The slowest-protein and slowest-ligand tICs are also shown to be largely uncorrelated, but the slowest protein motions appear to occur during, but independent of, ligand binding. Because haloperidol is an inverse agonist and the initial conformation of the receptor was an inactive state we would not expect to see the long timescale protein conformational changes that would be expected to accompany activation of the receptor.

In this work we have developed a protocol that improves the construction of MSMs for protein ligand binding. Specifically, we address the issues of ligand featurization and adaptive sampling. In developing the protocol, we found that binary contact featurization of the ligand to be superior to RMSD based featurizations or contact distances, which have been used previously. The use of binary contacts prevents kinetically close states from being separated due to noise caused by uncorrelated movements beyond the 6 Å cutoff. Importantly, use of binary contacts produces in a well-defined solvent state where the ligand makes no contacts with the protein. A contact distance of 6 Å provided the best balance between reducing noise and information loss. Inclusion of hydrogen-bonding information in the fingerprint enabled us to describe situations where the ligand could rotate around its principle axis whilst occupying the same geometric volume, which is particularly important during the formation of the salt-bridge. Incorporation of, at the very least, salt-bridge formation into the featurization is likely to be especially important for aminergic GPCRs due to the highly conserved Asp3.32. There is further scope for improvement of the protein and ligand featurization; models encoding protein side chain rotamers may be useful to investigate 'gating' processes during binding while solvent descriptors may be useful to understand dewetting of the orthosteric site.

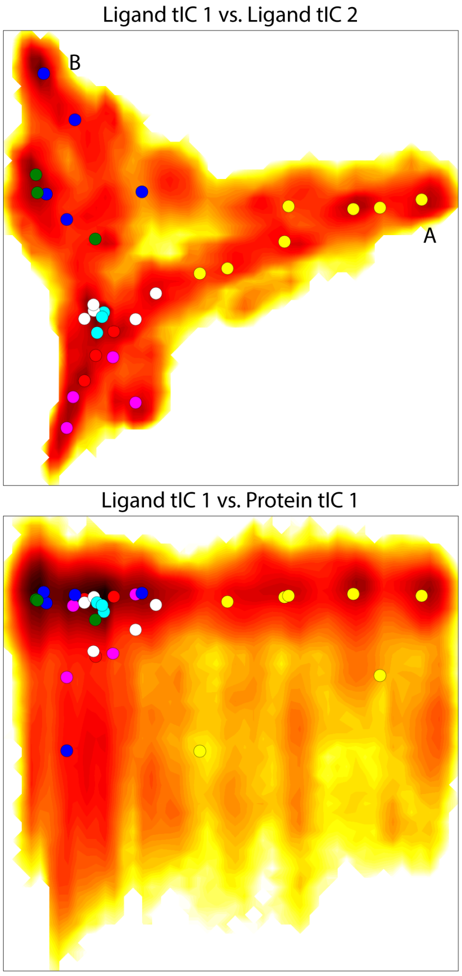


Figure 5, Free energy surface between (top) the slowest two ligand tICs and (bottom) the slowest ligand and slowest protein tICs. Means are shown as dots and colored according to the groupings in Figure 3. The positions of state A and state B are labelled.

A great advantage of applying MSMs to ligand binding is that adaptive methods can be used to direct sampling, increasing the computational efficiency. We used two protocols to direct sampling. In the early stages we chose to further sample states based on their contribution to model error and few non-self transitions. This protocol effectively determines under-sampled states and produces a well-connected MSM, but one that contains groups of highly interconnected states with poor connection to the rest of the network. We subsequently used a HMM to identify the states close to the under-sampled transitions and performed additional simulations to solve this issue. We recommend the occasional construction of a HMM throughout the adaptive sampling process to identify if discretization error is causing such a problem.

During construction of the Markov model for ligand binding, we also identified a number of issues associated with the use of k-means clustering. K-means clustering proved to be adequate but its tendency to produce more clusters in regions of high sampling density leads to poor assignment of outliers into appropriate clusters, including those that arise from the exploration of new, relevant conformational space. To solve this issue, we included manual selection of starting points as part of the adaptive sampling process, and when randomly drawing from clusters we avoided outliers by drawing from the half of the population closest to the cluster mean. We considered agglomerative hierarchical clustering as a superior clustering algorithm, however the size of MSM datasets limits the application of this method. Hierarchical clustering also suffers from the opposite problem to k-means, outlier clusters would lead to exploration of uninteresting areas such as bilayer binding, requiring a different selection process during adaptive sampling. The ability of k-means clustering to quickly and efficiently handle large volumes of data is greatly in its favor.

This study makes many advances over our previous study of haloperidol binding to the D_3R (6) in which we performed 14 conventional long-timescale MD simulations of haloperidol with the D_2R or D_3R but observed only a single complete binding event. In the previous work we identified the bound orthosteric pose to be state A. The RMSD of ligand position between state A and the earlier study is 2.9 Å. This pose itself was well supported by the overlap of the haloperidol pharmacophore with other D_3R antagonists, as well as being in agreement with other reported modeling studies. Binding pathway A of our current model is consistent with that previous work, although there are some minor differences. In this current work we observe shallower insertion of the aromatic ring between TM5-6 and a greater involvement of the alcohol group interacting with Asp3.32 prior to salt-bridge formation. Although these differences could be due to the poorly-sampled nature of the previous work, it should also be noted that these studies have used different force fields. The GROMOS force field is used here, while the previous study used the CHARMM force field.

In addition to the bound states of the ligand, our model also identified kinetic traps, metastable binding sites that are unconducive to the ligand reaching the orthosteric binding site. Traps 1 and 2 are of primary interest, as they represent metastable bound states where haloperidol is in close proximity with Asp3.32 and able to obstruct binding to the orthosteric site. Whilst our model suggests that binding to these sites is too short-lived to compete effectively as an antagonist, the pose of haloperidol in these sites presents a potential starting point for structure-based drug design.

Whilst there is no experimental evidence that suggests haloperidol acts an allosteric inhibitor, we do observe metastable binding in a secondary binding pocket located between TMs 1, 2, and 7 (Figure 4D). Glu2.65 in this secondary binding pocket is complimentary to the conserved amino group present in dopamine receptor targeting drugs and has been a chief focus in the exploration of allosterically or bitopically acting compounds in dopamine receptors (33). A structure-activity study of the “SB” series of compounds found that hydrogen bonding to Glu2.65 (in the D_2R) was able to induce an allosteric effect (34). Our model contains 2 kinetically distinct pairs of states of haloperidol bound in this pocket; in both states the ligand interacts with Glu2.65 but each state has an antiparallel orientation compared to the other. Despite the two orientations of haloperidol in this pocket, both pairs of states only feed into pathway B. The haloperidol conformations in these kinetic groupings are quite variable, likely due to solvent exposure. While the ligand maintains either a salt-bridge or hydrogen-bond with Glu2.65, the aromatic rings flit about the vestibule, interacting with Tyr7.35 or aromatic rings in ECL1. We proposed in our previous work that this secondary binding pocket appeared to favor the same pharmacophore as the orthosteric site (similar distances between aromatic and charged acidic residues) and we repeat that observation in this study, although the HMM also suggests that these poses are not stable over longer timescales. Nevertheless, drugs based on a haloperidol scaffold may have a potential allosteric effect through salt-bridge formation, or hydrogen bonding of the alcohol to Glu2.65.

Conclusion

We have developed an improved protocol for the construction of MSMs and HMMs for the binding of complex, flexible ligands to deeply buried binding sites, such as those present in GPCRs. This protocol utilizes a minimal set of descriptors, combining: binary contact fingerprints, hydrogen bonding, and Cα RMSD. The protocol provides an adaptive sampling approach that effectively explores conformational space simultaneously with error reduction. This methodology is applicable to a wide range of systems.

In this particular application we have built a MSM of haloperidol binding to the D_3 dopamine receptor which reveals multiple binding pathways, and several distinct metastable states. These metastable states include 2 bound states and 2 kinetic traps that could be considered in structure based drug design. Additional metastable sites in the secondary binding pocket might also be interesting from an allosteric-ligand design perspective.

We also raise questions on how these metastable states would affect the ligand's ability to act in its primary mechanism of action (in this case antagonism versus inverse agonism). A question that to answer properly, at least in terms of GPCRs, is left to a future in which we can explore the receptor ensemble.

Perhaps the most unexpected revelation of our model was the co-existence of state B, predicted to be the highest affinity bound state, with state A, the bound pose predicted in our previous work. Both of these states appear to exhibit kinetics on timescales longer than we can practically simulate, and while we therefore cannot be certain of their role in haloperidol's pharmacology, they both appear to be significant kinetically.

Acknowledgements

We acknowledge computational resources and technical support provided by Melbourne Bioinformatics through grant VR0133, the National Computational Infrastructure (NCI), and Multimodal Australian ScienceS Imaging and Visualization Environment (MASSIVE) through grant y96. T.T. is a recipient of an Australian Postgraduate Award (APA) scholarship.

References

1. Pan AC, Borhani DW, Dror RO, Shaw DE (2013) Molecular determinants of drug-receptor binding kinetics. *Drug Discov Today* 18(13–14):667–673.

2. Lu H, Tonge PJ (2010) Drug-target residence time: critical information for lead optimization. *Curr Opin Chem Biol* 14(4):467–474.

3. Hoffmann C, et al. (2015) Ligand Residence Time at G-protein-Coupled Receptors--Why We Should Take Our Time To Study It. *Mol Pharmacol* 88(3):552–560.

4. Dror RO, 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.

5. Kruse AC, et al. (2012) Structure and dynamics of the M3 muscarinic acetylcholine receptor. *Nature* 482(7386):552–556.

6. Thomas T, Fang Y, Yuriev E, Chalmers DK (2015) Ligand Binding Pathways of Clozapine and Haloperidol in the Dopamine D2 and D3 Receptors. *J Chem Inf Model* 56(2):308–321.

7. Shaw DE, et al. (2009) Millisecond-Scale Molecular Dynamics Simulations on Anton. *SC '09 Proceedings of the Conference on High Performance Computing Networking, Storage and Analysis* (New York), pp 1–11.

8. Pande VS, Beauchamp K, Bowman GR (2010) Everything you wanted to know about Markov State Models but were afraid to ask. *Methods* 52(1):99–105.

9. Bowman GR, Pande VS, Noé F eds. (2014) *An Introduction to Markov State Models and Their Application to Long Timescale Molecular Simulation* (Springer Netherlands, Dordrecht).

10. Prinz J-H, Keller B, Noé F (2011) Probing molecular kinetics with Markov models: metastable states, transition pathways and spectroscopic observables. *Phys Chem Chem Phys* 13(38):16912–16927.

11. Prinz J-H, et al. (2011) Markov models of molecular kinetics: Generation and validation. *J Chem Phys* 134(17):174105.

12. Shukla D, Hernández CX, Weber JK, Pande VS (2015) Markov State Models Provide Insights into Dynamic Modulation of Protein Function. *Acc Chem Res* 48(2):414–422.

13. Naritomi Y, Fuchigami S (2011) Slow dynamics in protein fluctuations revealed by time-structure based independent component analysis: The case of domain motions. *J Chem Phys* 134(6):65101.

14. Pérez-Hernández G, Paul F, Giorgino T, De Fabritiis G, Noé F (2013) Identification of slow molecular order parameters for Markov model construction. *J Chem Phys* 139(1):15102.

15. Schwantes CR, Pande VS (2013) Improvements in Markov State Model Construction Reveal Many Non-Native Interactions in the Folding of NTL9. *J Chem Theory Comput* 9(4):2000–2009.

16. Buch I, Giorgino T, De Fabritiis G (2011) Complete reconstruction of an enzyme-inhibitor binding process by molecular dynamics simulations. *Proc Natl Acad Sci U S A* 108(25):10184–10189.

17. Doerr S, De Fabritiis G (2014) On-the-fly learning and sampling of ligand binding by high-throughput molecular simulations. *J Chem Theory Comput* 10(5):2064–2069.

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

19. Held M, Metzner P, Prinz J-HH, Noé F (2011) Mechanisms of protein-ligand association and its modulation by protein mutations. *Biophys J* 100(3):701–710.

20. Silva D-AA, Bowman GR, Sosa-Peinado A, Huang X (2011) A Role for Both Conformational Selection and Induced Fit in Ligand Binding by the LAO Protein. *PLoS Comput Biol* 7(5):e1002054.

21. Kohlhoff KJ, et al. (2014) Cloud-based simulations on Google Exacycle reveal ligand modulation of GPCR activation pathways. *Nat Chem* 6(1):15–21.

22. Kapur S, Seeman P (2000) Antipsychotic agents differ in how fast they come off the dopamine D2 receptors. Implications for atypical antipsychotic action. *J Psychiatry Neurosci* 25(2):161–166.

23. Abraham MJ, et al. (2015) GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* 1:19–25.

24. Schmid N, et al. (2011) Definition and testing of the GROMOS force-field versions 54A7 and 54B7. *Eur Biophys J* 40(7):843–856.

25. Poger D, Van Gunsteren WF, Mark AE (2010) A new force field for simulating phosphatidylcholine bilayers. *J Comput Chem* 31(6):1117–1125.

26. Koziara KB, Stroet M, Malde AK, Mark AE (2014) Testing and validation of the Automated Topology Builder (ATB) version 2.0: prediction of hydration free enthalpies. *J Comput Aided Mol Des* 28(3):221–233.

27. Canzar S, et al. (2013) Charge Group Partitioning in Biomolecular Simulation. *J Comput Biol* 20(3):188–198.

28. Malde AK, et al. (2011) An Automated Force Field Topology Builder (ATB) and Repository: Version 1.0. *J Chem Theory Comput* 7(12):4026–4037.

29. Hopkins CW, Le Grand S, Walker RC, Roitberg AE (2015) Long-Time-Step Molecular Dynamics through Hydrogen Mass Repartitioning. *J Chem Theory Comput* 11(4):1864–1874.

30. Wassenaar TA, Ingólfsson HI, Böckmann RA, Tieleman DP, Marrink SJ (2015) Computational lipidomics with insane: A versatile tool for generating custom membranes for molecular simulations. *J Chem Theory Comput* 11(5):2144–2155.

31. Beauchamp K a., et al. (2011) MSMBuilder2: Modeling conformational dynamics on the picosecond to millisecond scale. *J Chem Theory Comput* 7(10):3412–3419.

32. Baker EN, Hubbard RE (1984) Hydrogen bonding in globular proteins. *Prog Biophys Mol Biol* 44(2):97–179.

33. Chien EYT, et al. (2011) Structure of the human dopamine D3 receptor in complex with a D2/D3 selective antagonist. *Science (80-)* 330(6007):1091–1095.

34. Shonberg J, et al. (2015) Structure–Activity Study of N-((trans)-4-(2-(7-Cyano-3,4-dihydroisoquinolin-2(1H)-yl)ethyl)cyclohexyl)-1H-indole-2-carboxamide (SB269652), a Bitopic Ligand That Acts as a Negative Allosteric Modulator of the Dopamine D $_2$ Receptor. *J Med Chem* 58(13):5287–5307.

Conclusion

Ligand binding is commonly modeled as a two-state process but, in reality, ligands must traverse the space between the solvent and the binding site. There are many metastable states along the binding and unbinding pathways and each can be important for drug design, either through its influence on the rates of binding/unbinding or as a separate target for allosteric or bitopic ligands. Our goal in this work was to develop methods that would allow us to better understand the behavior of ligands during binding and apply those methods to pharmaceutically important systems. To that end, we have presented various ways of investigating the ligand-binding process: we have constructed homology models to allow prediction of bound poses, conducted molecular dynamics simulations to investigate binding pathways, and constructed Markov state models to map out the entire binding ensemble.

GPCRs are receptors of great pharmaceutical interest but, due to the challenging nature of structure determination for these receptors, there are many important receptor subtypes without available crystal structures. To address the lack of crystal structures for the M_1R-M_5R muscarinic receptors, we constructed a series of homology models of these subtypes based on a β_2AR template (Chapter 2). Functional knowledge was incorporated into the model building process to produce optimized homology models that would be more useful in drug design. Crystal structures of the M_2R (human) and M_3R (rat) later became available and we developed a naïve homology model of the M_3R based on the closer rat M_3R template. We then used virtual screening to evaluate our optimized homology models alongside this M_2R crystal structure and M_3R naïve homology model, testing the ability of these structures to distinguish between known antagonists and decoy compounds. We found that our optimized homology models were a significant improvement over the untrained models, and therefore a better choice for future drug design projects. Homology modeling continues to be a useful tool to fill gaps in structural knowledge amongst families of receptors, and the development of novel homology models is often the first step in opening a target up to more advanced computational investigation. As we gain more knowledge of ligand binding to GPCRs, it is becoming increasingly apparent that the receptors need to be treated as an ensemble, and users of homology models will need to adapt to this new paradigm.

While homology models are useful for virtual screening and structure-based drug design, they are based on static experimental structures that do not convey the dynamics of the receptor, especially in the often more flexible regions outside of the binding site, and nor are they suitable for locating metastable states. To investigate the behavior of drugs outside of the orthosteric binding site, we employed molecular dynamics simulations. Molecular dynamics models can, like docking methods, predict the bound pose of a ligand, but they are also able to predict the binding pathway of a drug, and capture the dynamics of the receptor in atomic resolution. We chose to perform such a simulation study on the D_2 and D_3 dopamine receptors, which are important targets in the treatment of schizophrenia, Parkinson's disease, and addiction (Chapter 3). The D_2R and D_3R have extremely high homology in the orthosteric binding site, so understanding the behavior of ligands outside of this site is critical for the understanding of the factors that contribute to the selectivity of dopaminergic ligands, and for reducing the serious side-effects that afflict this important class of drugs. We performed the first MD simulations of the binding pathway of the clinically important antipsychotic ligands clozapine and haloperidol binding to the D_2 and D_3 dopamine receptors. This set of simulation data accessed timescales that are much longer than most other studies in the literature, allowing the simulations to not only predict the bound pose of the 2 ligands, but also revealing the pathway of each ligand takes from the extracellular vestibule of the receptor to the orthosteric site. A cluster analysis was performed on this simulation data for each receptor, and we identified differences between the metastable binding states in the D_2R and D_3R, despite the non-selective nature of our ligands. These metastable sites are therefore points of interest to consider in simulation or design of selective dopaminergic antagonists.

Of the long-timescale MD simulations performed on the dopamine receptors, most only revealed partial binding pathways, and the complete binding pathways were not replicated. This presented 2 problems, we needed to access longer timescales so that we could observe more binding events, and we needed to support our findings with statistics so that we could have confidence in our observations. To solve both of these problems, we turned to the construction of Markov state models. While there is specialized hardware that makes running MD simulations without enhanced sampling techniques feasible for many targets of interest, there is always a larger, more complicated target that is of interest. MSMs extend the timescales reachable by current conventional MD simulations and provide a useful framework for analyzing the resulting data. MSMs have been largely developed for the analysis of protein folding simulations and have seen little use in ligand binding. A more widespread application of MSMs would greatly benefit the MSM methodology as well as our knowledge of the systems being studied.

We began our work on MSMs using oleic acid binding to the liver fatty acid-binding protein as a model system that would still present a more challenging ligand-binding scenario than the systems used in the ligand binding MSM literature. In the initial studies of the L-FABP-oleic acid complex, we performed an extensive set of conventional molecular dynamics simulations on this system (Chapter 4). This set of simulations suffered from the same limitations as our initial simulations of the dopamine receptor but, because of the smaller system, we were able to access longer timescales and observe many more binding events. From these simulations, we observed complete trajectories of oleic acid to the L-FABP and reproduced the experimentally determined complex. In addition, we identified a novel metastable site at the entrance to the binding pocket. L-FABP binds 2 ligands simultaneously, and this metastable site appeared to be more occupied than the experimentally-determined low-affinity binding site, suggesting that it was important in the bound ensemble. We then began a larger set of shorter simulations from which to construct a MSM. The size of the resultant dataset was 2 orders of magnitude larger than any previous study on FABPs. To construct a MSM from this data we borrowed methodology from protein folding MSMs and modified the algorithms to apply them to our ligand-binding system. We used the coordinates of the ligand relative to the protein to describe the system and implemented basic adaptive sampling by performing the simulations in two batches. The first batch identified areas that needed additional sampling, and the second batch further sampled those areas. The resultant MSM predicted a new binding pathway involving our previously identified metastable state, which was not observed in the conventional MD simulations. The model also predicted that this metastable state and the low-affinity binding site were approximately equal in population. There were 2 main flaws in the MSM methodology applied to the FABP system: firstly, the RMSD descriptor poorly handled describing both the flexibility of the system and the behavior of the ligand in the solvent and, secondly, the minimal adaptive sampling used was not enough to correct undersampling of the slowest transitions.

Continuing the development of our MSM methodology, we returned to the D_3R-haloperidol system, where updates to the available MSM software allowed us to implement our own code and make significant improvements to our approach. We developed an MSM methodology using the D_3R-haloperidol system, primarily by incorporating an improved adaptive sampling process and developing a set of improved descriptors for the system (Chapter 5). In the adaptive sampling regime, we performed many small batches of simulations and after each constructed a transition network. We then analyzed the transition network to identify the states that had the highest contribution to the error and the states that were exploring new space, and used these states to launch the next batch of simulations. This adaptive sampling methodology was a significant improvement over our earlier FABP work. However, where other adaptive sampling methods attempt to fully automate the sampling process, our methodology still requires a manual selection of states. Manual selection was necessary to avoid sampling binding to the membrane environment of the GPCR system, but in other proteins the method could be automated alongside an appropriate clustering algorithm. We also devised a set of improved descriptors for the location of haloperidol in the system, describing the position of the ligand by the residues it contacted and the hydrogen bonds that it formed. These descriptors are applicable to other ligand-binding systems, and their ability to efficiently describe the binding behavior of a drug-like ligand makes them a useful addition to the field. This combined methodology presents a means to investigate the binding of drug-like ligands with MSMs, where previously only the binding of relatively simple ligands could be studied.

We used our new methodology to construct a MSM of haloperidol binding to the D_3R. The resultant MSM is a significant advancement from our previous work on the D_3R-haloperidol system, and provides a more sophisticated model of the behavior of a ligand binding to a GPCR than methods employed in the literature. In addition to replicating our previously predicted binding pathways, we were able to further explore the binding ensemble and identify additional binding pathways, as well as support our observations with statistics. We were also able to identify a variety of metastable states and predict their importance in drug binding.

MSMs, combined with the methodology developed in this work, are capable of mapping the ligand-binding ensembles of a diverse range of targets. However it is still unclear to what extent knowledge of the ligand-binding ensemble allows drug binding to be manipulated experimentally. Metastable states already present targets for the design of allosteric or bitopic ligands, but there are not yet any studies in which a metastable state is used to modify the kinetics of a drug. Such a study would greatly increase the tools available for drug design.

Improvements to computer hardware continually push computationally accessible timescales and increase the complexity of the systems that can be simulated. Now that simulating on the timescales of ligand binding is becoming a routine possibility, there is a need for methodologies that can both further improve the efficiency of these simulations, and scale to more complex systems. As computer power increases even further, the limits to the size and detail of systems that we can simulate will increase too. In this work we have simulated GPCRs on timescales that can capture the binding of a ligand, and simulations in the literature have begun to capture the conformational ensemble of these receptors. It will eventually be possible to simulate GPCRs in much larger systems that can model their association with G proteins, or model their interactions with other proteins in the cell membrane. Rather than examine larger systems, it will also be possible to study current systems in increased detail. Where in this work we have studied the binding of ligands in atomic detail, eventually it will be possible to study ligand-binding pathways on a quantum level. Much as the work we have presented here has improved the existing methods and advanced the ligand binding field, future work will need to improve upon these methods to tackle ever-larger volumes of data, and describe the behavior of increasingly complicated systems.