First-Principles-Based Simulations for G Protein-Coupled Receptor Activation and for Large-Scale Nonadiabatic Electron Dynamics

Thesis by

Sijia Dong

In Partial Fulfillment of the Requirements for the
degree of
Doctor of Philosophy

CALIFORNIA INSTITUTE OF TECHNOLOGY
Pasadena, California

2017
Defended December 20, 2016

Acknowledgements

I would like to thank my advisor, Prof. William A. Goddard III, for allowing me to work on several projects in very different areas that were fun and really broadened my horizons. His patience, guidance, encouragement, and being a role model as a dedicated scientist, have been indispensable in getting me where I am.

I would like to thank my thesis committee, Prof. Jim Heath, Prof. Harry Gray and Prof. Rudy Marcus, and my candidacy committee member, Prof. Vincent McKoy, for their criticism, advice, and encouragement. They pointed out some of the aspects of how to become a better scientist from a higher level, and I really appreciate their input.

I would like to thank my mentor, Prof. Ravinder Abrol, for not only discussing research with me on a regular basis over the course of the G protein-coupled receptor (GPCR) project, but also giving me some of the most important career advice I got in graduate school.

I would like to thank Hai Xiao, whom I worked with on developing the new generation of electron force field (GHA-QM), for those long hours of discussions on research, career, and life, for answering my questions tirelessly, and for bearing with me when the progress on our project was slow.

I would like to thank the members of the office 07, Samantha Johnson, Yufeng Huang, Robert "Smith" Nielsen and Jason Crowley, for being officemates, colleagues, and friends that I could hang out with both in and out of the office. Sam was really the glue that stuck us together, and made our office a warm place to be in.

I would like to thank people involved in the various projects I have worked on during graduate school: Andres Jaramillo-Botero (GHA-QM), Julius Su (GHA-QM), Tao Cheng (ReaxFF, antifreeze protein), Sergey Zybin (ReaxFF), Saber Naserifar (pQEq/ReaxFF), Fan Liu (GPCR, including bitter taste receptors), Soo-Kyung Kim (GPCR, including taste receptors), Adam Griffith (GPCR), Vaclav Cvicek (GPCR), Audrey Kishishita (antifreeze protein), Arundhati Nag (solvation of peptides), and Tod Pascal (2PT). Darius Teo was an undergraduate student I mentored on investigating metallocorroles' inhibition of a cancer-related protein (Hsp90) and enzyme catalysis. I would like to thank him for giving me the opportunity to have a particularly fruitful mentoring experience. Biogroup members in the Goddard group not in the previous lists: Caitlin Scott, Andrea Kirkpatrick, Matt Gethers, Si-Ping

Han, and Henry Levenson. To name a few other Goddard group members I have scientifically, professionally, or personally interacted with a bit more (in alphabetic order): Qi An, Daniel Brooks, Yalu Chen, Mu-Jeng Cheng, Shane Flynn, Ross Fu, Si-Ping Han, Jamil Tahir-Kheli, Yan-Choi Lam, Wei-Guang Liu, Yuanyue Liu, Yuan Ping, Jin Qian, Ravishankar Sundararaman, Ho-Cheng Tsai, and Ted Yu. During the initial couple years of graduate school, I worked with Prof. Geoff Blake on terahertz spectroscopy and Prof. Lea Goentoro on systems biology. I would like to thank them for allowing me to explore my interests and try new scientific territories I had no previous experience with.

I would like to thank Caltech Karate Club, Shotokan Karate of America, and their founder, Mr. Ohshima, for keeping an oasis of martial arts alive for so many years. Starting karate practice at Caltech was one of the best decisions I have made in life. It has ensured my well-being in graduate school both physically and mentally. Chris Finch, David Berry, and Kei Teramoto led most of the practices during the years I was in the club. They enabled the club traditions to transcend generations of club members.

I would also like to thank all my personal friends that are too many to name. I could not express my gratitude enough for you being part of my life all these years.

Last but not the least, I would like to thank my family, particularly mom, dad, and Shuo Han. I could not have achieved whatever I have without your love and support.

Abstract

This thesis focuses on simulating large molecular systems within and beyond the Born-Oppenheimer framework from first principles. Two approaches have been developed for very different but important applications.

The first one is a hybrid method based on classical force fields that predicts the high-energy ensemble of three-dimensional structures of a class of proteins critical in human physiology: the G protein-coupled receptors (GPCRs). GPCRs' functions rely on their activation marked by a series of conformational changes related to binding of certain ligands, but the short of experimental structures has hampered the study of their activation mechanism and drug discovery. Our method, combining homology modeling, hierarchical sampling, and nanosecond (ns)-scale molecular dynamics, is one of the very few computational methods that can predict their active-state conformations and is one of the most computationally inexpensive. It enables the conformational landscape and the first quantitative energy landscape of GPCR activation to be efficiently mapped out.

This method, named ActiveGEnSeMBLE, allows the inactive- and active-state conformations of GPCRs without an experimental structure to be systematically predicted. We have validated the method with one of the most well-studied GPCRs, human β_2 adrenergic receptor (hβ_2AR), and applied the method on a GPCR without an experimental structure, human somatostatin receptor 5 (hSSTR5). Insights on GPCR activation as well as structure prediction methods are discussed.

The second one is a semiclassical approach for large-scale nonadiabatic dynamics of condensed systems in extreme conditions, termed Gaussian Hartree Approximated Quantum Mechanics (GHA-QM). Many nonadiabatic processes related to important applications (e.g. renewable energy) happen in large systems, but existing excited state dynamics methods are too computationally demanding for their long timescale simulations. GHA-QM is based on the electron force field (eFF) framework where we model electrons as Gaussian wavepackets and nuclei as classical point charges, and obtain a simplified solution to the time-dependent Schrödinger equation as the equation of motion. We employ a force field philosophy approximating the total energy as a sum of electronic kinetic energies, electrostatic energies and a Pauli correction, which corrects for the lack of explicit antisymmetry in the wavefunctions. New designs of the Pauli potential and preliminary results on hydrogen systems are

discussed. With the new development, we hope to improve the accuracy and range of applications of eFF to simulate the nonadiabatic dynamics of hundreds of thousands of electrons on nanosecond timescale.

Published Content and Contributions

(*1*) Dong, S. S., Abrol, R., and Goddard, W. A., (2015). The Predicted Ensemble of Low-Energy Conformations of Human Somatostatin Receptor Subtype 5 and the Binding of Antagonists. *ChemMedChem 10*, 650–661, DOI: 10.1002/cmdc.201500023,
S.S.D. designed and performed research, and wrote the manuscript.

(*2*) Dong, S. S., Goddard, W. A., and Abrol, R., (2016). Conformational and Thermodynamic Landscape of GPCR Activation from Theory and Computation. *Biophysical Journal 110*, 2618–2629, DOI: 10.1016/j.bpj.2016.04.028,
S.S.D. designed and performed research, and wrote the manuscript.

Table of Contents

List of Illustrations

Number *Page*

List of Tables

Number *Page*

Chapter 1

Introduction

Large and complex chemical systems are ubiquitous in nature, and the understanding of these systems is critical for the scientific and technological development in some of the most important areas of the society, such as human health and renewable energy. Theoretical and computational studies of such systems are essential in advancing these areas, because they complement experimental studies and are able to provide insights and key information that experiments cannot provide.

Developments in physics-based models and computing technology have enabled computational studies of systems people could not imagine being able to study not many years ago, but the simulation of large molecular systems and long timescale processes remains one major challenge in the field of computational chemistry. The *ab initio* quantum chemistry methods are relatively accurate but computationally expensive, while the classical force fields are computationally cheaper but not as accurate, and not able to provide electronic structures.

In addition, there are systems (Case 1) that have a large number of atoms undergoing processes that span a large timescale, whose dynamics are challenging to be simulated even using classical force fields. There are also systems (Case 2) that have a size acceptable for a reasonably-long classical simulation, but the nature of the processes requires methods that take quantum effects into account, and existing quantum mechanical, mixed quantum-classical or semiclassical methods are either expensive or not accurate enough, or both.

In this thesis, two computational approaches that respectively aim at dealing with these two aforementioned cases and their applications will be discussed.

The first one greatly extends the scope of structure prediction and mechanistic study of a class of membrane proteins called the G protein-coupled receptor (GPCR), that are "large and slow". When the focus is on a specific system, one could take advantage of certain features specific to the system to make the simulation more efficient, and this is the general idea of our approach.

GPCRs exist in many organs of the human body, and play indispensable roles in human physiology, including but not limited to the senses of vision, taste and smell, mood regulation, cell growth and death, hormone secretion, and immune system regulation. Therefore, they are important drug targets, and 30%-40% of drugs on the market target on them. However, only ~30 of the ~800 human GPCRs have three-dimensional (3D) atomistic structures available. In addition, these proteins are conformationally dynamic in the membrane, but this is very difficult to be captured in x-ray crystallography. Most of the time they adopt conformations that are called the inactive states, because these conformations are relatively lower in energy. An agonist can promote the GPCR to more frequently adopt a collection of conformations that can accommodate the binding with a G protein or other cellular effectors such as a β-arrestin. These conformations are called the active states. Most of the ~30 GPCRs that have experimental structures only have one inactive-state structure available, and only 5 of them have an active-state experimental structure. Therefore, there is a great potential for computation to help elucidating the different conformations of GPCRs along their activation pathways, to study their activation mechanisms, and to enable rational design of drugs targeting on them. Because of the greater availability of inactive-state structures, predicting active-state structures from inactive-state structures would be particularly valuable. However, the long timescale for a GPCR to transform from its inactive state to active state poses challenge for computation too. This long timescale means it is hardly feasible for a brute-force molecular dynamics to take a GPCR from its inactive state to active state.

To address this challenge, based on the GPCR Ensemble of Structures in Membrane BiLayer Environment (GEnSeMBLE) method our group developed, we have developed a new hybrid computational method, ActiveGEnSeMBLE, that combines homology modeling, hierarchical discrete screening of protein conformations, and molecular dynamics (MD) simulations, and enables the prediction of active and inactive GPCR structures systematically. Chapter 2 devotes to the development of the ActiveGEnSeMBLE methodology and the validation of the method with experimental GPCR structures. The concluding remarks of Chapter 2 also discuss the outlook of the method from a higher level. Chapter 3 and 4 discuss the application of ActiveGEnSeMBLE on a GPCR without an experimental structure, human somatostatin receptor 5 (hSSTR5). Chapter 3 focuses on the analysis of the active-state conformations of hSSTR5 and the insights the analysis provides on GPCR activation as well as GPCR structure prediction strategies. Chapter 4 focuses on the structure prediction of hSSTR5 inactive-state conformations and their binding with antagonists. Chapter 2 and 3 are mainly based on the publication (*1*), and Chapter

4 is mainly based on the publication (*2*).

The second one has the goal of simulating the nonadiabatic electron dynamics of large and condensed systems. These nonadiabatic processes are the pillars of some important phenomena that hold the key to questions from fundamental science to creating clean and renewable energy. Simulation methods for excited state dynamics are not as well developed as ground state dynamics where the Born-Oppenheimer approximation applies, especially for long timescale dynamics of large systems with dense electronic states. Because of the requirements in both computational efficiency and the description of quantum mechanical properties, a semiclassical method such as the electron force field (eFF) is a reasonable starting point for large-scale nonadiabatic dynamics. Based on the eFF framework, we have developed the Gaussian Hartree Approximated Quantum Mechanics (GHA-QM) framework that has a new and more complete description on the quantum mechanical contribution to the total energy. Chapter 5 presents the motivation of the work, reviews the first generation eFF and GHA-QM methods, and discusses our recent efforts on improving GHA-QM.

GPCR and nonadiabatic dynamics relate in that there are GPCRs that can be photoactivated through nonadiabatic transitions between adiabatic potential energy surfaces (PESs). These GPCRs, called rhodopsin, have a covalently linked ligand (retinal) that undergoes photoinduced isomerization that promotes the protein to adopt active-state conformations and enables vision. The photoisomerization is a nonadiabatic process. Because GHA-QM is developed with different applications in mind and has a mean-field assumption, in its current form it may not apply to rhodopsin photoactivation. There have already been other methods that can better simulate this process, so currently we do not see the urgency of developing a new method for simulating photoisomerization in proteins.

The unifying theme of this thesis is to simulate large molecular systems and their long-timescale processes. ActiveGEnSeMBLE can be viewed as an exploration strategy of a rugged ground state PES: The system is directly taken to sparsely spaced grid points on the PES that cover a wide area of the PES, and then refinement of the grid surrounding selected grid points allows the local minima satisfying certain physical features to be efficiently discovered. A key ingredient is a selection rule that is based on suitable collective variables specific to the system. GHA-QM, on the other hand, simulates systems that nonadiabatic transitions between different PESs constantly happen (a.k.a. the Born-Oppenheimer approximation breaks down) by taking a mean-field approach and doing wave packet dynamics of electrons, with the total energy being the sum of electronic kinetic energy, electrostatic energy, and a potential accounting for the Pauli exclusion principle. To sum up, this thesis has covered the two major classes of problems in large-scale simulations, one within the Born-Oppenheimer framework and the other one beyond it.

References

(*1*) Dong, S. S., Goddard, W. A., and Abrol, R., (2016). Conformational and Thermodynamic Landscape of GPCR Activation from Theory and Computation. *Biophysical Journal 110*, 2618–2629, DOI: `10.1016/j.bpj.2016.04.028`,

(*2*) Dong, S. S., Abrol, R., and Goddard, W. A., (2015). The Predicted Ensemble of Low-Energy Conformations of Human Somatostatin Receptor Subtype 5 and the Binding of Antagonists. *ChemMedChem 10*, 650–661, DOI: `10.1002/cmdc.201500023`,

Chapter 2

Predicting the Conformational and Thermodynamic Landscape of G Protein-Coupled Receptor (GPCR) Activation: Development of ActiveGEnSeMBLE

This chapter is based on the following publication:

Dong, S. S.; Goddard, W. A.; Abrol, R. "Conformational and Thermodynamic Landscape of GPCR Activation From Theory and Computation". *Biophys. J.*, **2016**, *110* (12), 2618-2629. doi: 10.1016/j.bpj.2016.04.028

2.1 Introduction

G protein-coupled receptors (GPCRs) play critical signaling functions for numerous cellular processes, making important targets for therapeutics. Developing such therapeutics is complicated because the activation of GPCRs that is integral to their function involves multiple distinct conformations along the pathway for activation. Moreover, some GPCRs are capable of activating more than one intracellular (IC) pathway,(*1*) making it essential to identify multiple active conformations possibly involved with different functions. In order to understand GPCR activation mechanisms and to carry out structure-based drug design, it is necessary to obtain accurate three-dimensional structures for each of these important conformations. This creates a huge problem for structure determination experiments, since it requires stabilizing each of these structures sufficiently to have an ordered crystal. Indeed, despite huge efforts and remarkable experimental breakthroughs, there are experimental structures for only 4% of the ~800 human GPCRs. Of these most are for an inactive conformation and only 5 are for active-like conformations (when this paper was published in June 2016, there were only 4). In addition, among those 5 GPCRs with both the active and inactive structures crystalized, only one (rhodopsin) has more than one active experimental structure.

There is a huge potential role for theory and simulation to fill in this crystal structure gap for GPCRs. However, there are major problems using theory to predict the activation of such complex membrane bound proteins. Millisecond-long molecular dynamics (MD) simulations have not been successful in following activation from an inactive state to an active-like state along an activation pathway. In addition, the sequence homologies between different GPCRs are low to get structures for all the remaining GPCRs, most of which would be only possible for inactive structures. To begin to address this problem, we had developed the GEnSeMBLE (*2*) method to predict the ensemble of low energy (stable) 3D structures of GPCRs. This method has successfully predicted the structures for several Class A and Class B GPCRs: C-C chemokine receptor type 5 (CCR5),(*3*) adenosine A3 receptor (AA_3R),(*4*) cannabinoid receptor type 1 (CB1),(*5*) taste receptor type 2 member 38 (Tas2R38),(*6*) olfactory receptor 1G1 (OR1G1),(*7*) glucagon-like peptide 1 receptor (GLP1R),(*8*) prostaglandin D2 receptor (DP).(*9*) Most of the predictions are for inactive state structures, but we were able to predict and validate active state structures of AA_3R(*4*) and CB1.(*5*)

The GEnSeMBLE methodology starts with several template configurations specifying the initial helix packing (helix locations, tilts and rotations) for the seven helical transmembrane domains (TMDs) based either on an experimental structure of a similar GPCR or based on a previous predicted structure. Then we consider all possible simultaneous rotations by multiples of 30° about the helical axes for all seven TMDs. We estimate the energies of all $12^7 \approx 35$ million rotations by calculating the energies for all combinations of the 12 interacting pairs of helices (BiHelix), including optimized side chains. We build the best 1000 of these conformations by energy into seven-helix bundles with optimized side chains and then select ~20 of the lowest conformations by energy for further consideration. This is done for all plausible templates to identify which templates are best and which rotations best accommodate the target sequence. This is followed by an exhaustive sampling of simultaneous tilts and rotations of the seven helices (SuperBiHelix), leading to ~13 trillion helix tilts/rotation combinations. The energies for all 13 trillion conformations are again estimated by combining BiHelix energies to identify the best 2000 conformations, which are then built into seven-helix bundles for final side chain optimization. From this list, the ~20 lowest-energy conformations are selected for binding to different ligands and for further studies.(*2*)

Although exhaustive, this conformational sampling tends to be biased toward inactive conformations since a) the available templates are mainly inactive, b) inactive conformations usually have lower energy than active conformations and the procedure seeks lower energy structures, c) the agonist that might stabilize the active configurations are not present during the reduction from 13 trillion to 20 structures. Even so, GEnSeMBLE has successfully modeled some active GPCRs. In this paper, we propose a hybrid method, denoted ActiveGEnSeMBLE, that builds upon the original GEnSeMBLE method to systematically predict multiple potentially-active conformations of GPCRs. It utilizes a hierarchical sampling scheme that first samples conformations on a coarse grid followed by another conformational sampling with a finer grid. In addition, rather than using only energy-based scoring, ActiveGEnSeMBLE method uses both structural and energy information to identify the higher-energy candidates for active conformations that may reside in local energy wells stabilized by appropriate agonists or other cellular effectors (e.g. G protein or β-arrestin).

Herein we first validate the ActiveGEnSeMBLE method against experimental structures of GPCRs that have been obtained in active conformations with non-covalently bound ligands: human β_2 adrenergic receptor (hβ_2AR),(*10*, *11*) and human M2 muscarinic acetylcholine receptor (hM2),(*12*, *13*). [The x-ray structure of the agonist-bound mouse μ-opioid receptor (mOPRM or μOR) stabilized with a nanobody(*14*) was not available at the time of this study, so we did not include it as a test case.] Then we apply ActiveGEnSeMBLE to predict multiple active and inactive forms of human somatostatin receptor subtype 5 (hSSTR5), which are discussed in Chapters 3 and 4. We selected hSSTR5 because:

a. It plays an important role in anti-proliferation, hormone secretion, and human diseases such as pancreas cancer;(*15*)

b. There is no experimental structure available for use in drug development; and

c. A recent study identified hSSTR5 as the most valuable template for homology modeling of the non-orphan and non-olfactory class A GPCRs that constitute the majority of this class of GPCRs.(*16*)

2.2 Developing ActiveGEnSeMBLE for the systematic prediction of the GPCR conformations along its activation pathway

Our original GEnSeMBLE method follows the following steps, described in detail elsewhere:(*2*)

Step 1: Align the target sequence to the other GPCR sequences homologous up to an Expect (E) value of 0.1, and use the PredicTM method to determine the lengths and ranges of the helical hydrophobic core regions. Then use secondary structure prediction servers to predict helical regions that might extend beyond the hydrophobic core outside the membrane.

Step 2: Generate the structures for the helical regions of the target receptor with a variety of helical shapes using either (a) OptHelix which generates helical shape using minimization and molecular dynamics; or (b) homology modeling based on TMD from known GPCR crystal structures with high sequence identity to the target protein sequence.

Step 3: For each template-based structure (using OptHelix or homology helix shapes), sample 12^7 ≈35 million combinations of simultaneous rotations (η) of all seven helices (using BiHelix), then select the best 1000 (based on energy) to build full seven-helix bundles (CombiHelix) with optimized side chains. From these conformations, select a few diverse structures with lowest predicted energies as the starting points for the subsequent simultaneous optimization of tilt angles (θ, ϕ) and helix rotation angles (η). This SuperBiHelix sampling generally involves $(5{\times}5{\times}3)^7$ ≈13 trillion combinations. We then select the best 2000 that are then built into seven-helix bundles (SuperComBiHelix) from which the best ~20 conformations are selected (by energy) as the conformational ensemble that might play roles in GPCR function. These 20 conformations are then analyzed in terms of interhelical hydrogen bonds, particularly whether there are salt bridges between different TMD including the salt bridge interaction between domains TM3 and TM6, which is associated with inactive conformations for many Class A GPCRs.(*17*) Typically, helix shapes based on homology lead to more stable conformations, except for CB1 (*5*) and DP (*9*) where OptHelix was the best.

Step 4: Use the DarwinDock/GenDock (*18*, *19*) method to validate the predicted GPCR structures by exhaustive sampling of poses of known agonists and antagonists (including numerous torsional conformations) over possible binding regions. It involves assessing the contributions of binding from various protein residues in the ligand binding cavity (cavity analysis) and often comparisons with the binding for a range of ligands (structure-activity relationship analysis).

Step 5: Use homology or Monte Carlo procedures to add the loops and the amino (N-) and carboxyl (C-) terminal segments to the TMD of the predicted ligand-GPCR structures and optimize the loops through annealing. Build these complete predicted structures into the lipid membrane surrounded by explicit water and salt (~60 000 atoms per cell) and carry out modest (10-50 ns) of MD simulation to validate the stability of the predicted structures. The

> goal of the MD simulations is to allow water and ions to interact with the ligand-protein complex to relax the predicted structures. Here we analyze the changes in the strong interhelical interactions within the GPCR and strong ligand-protein interactions. For a valid structure, we expect to gain new interactions (sometimes due to water molecules) while not losing the original strong couplings. Such short MD would not allow an inactive structure to become activated.

Two GPCRs (rhodopsin and hβ_2AR) have reported crystal structures in which the full or partial Gα subunit C-terminus of the G protein is bound to the IC side of the GPCR. It is believed that these structures capture a stable GPCR active state. Comparison of these active structures to their inactive state counterparts shows that the active states have a different packing of the transmembrane (TM) helices and that TM6 changes shape during activation. In hβ_2AR, bovine rhodopsin (bRho), hM2 and mOPRM, the TM6 IC end moves horizontally away from its position relative to TM3 in the inactive state with a residue near the hydrophobic plane as the pivot point, resulting in the TM3-TM6 space on the IC side opening up by about 3-5 Å (Table 2.1). Both experiment and previous predictions provide strong evidence that each GPCR sequence may have multiple active states, with the active states generally higher in energy (less stable) than the inactive state.(*20*)

These insights inspired a modification of Steps 2 and 3 of GEnSeMBLE to develop the ActiveGEnSeMBLE method. In ActiveGEnSeMBLE, for Step 2, the template for the homology model is based, in addition to an inactive-state crystal structure, either on an available active-state crystal structure (for validation purposes only), or on another model based on a hybrid template in which only TM6 comes from an active-state crystal structure while the other 6 TMs (1 through 5, and 7) come from an inactive-state crystal structure. In ActiveGEnSeMBLE, Step 3 of GEnSeMBLE is replaced by a two-step conformational sampling scheme for SuperBiHelix that includes a coarse conformational sampling aimed at locating structures in the active-state potential energy wells followed by a finer conformational sampling starting from specific potential functionally-diverse conformations identified by the coarse sampling. Coarse conformational sampling casts a wide net to catch conformations that are potentially active. Specific conformations from coarse sampling are identified along an activation coordinate by using a geometric criterion prior to the energetic criterion as described below. Fine conformational sampling starting from these conformations relaxes them in their local potential energy wells. We

Protein (PDB ID)	Protein Description	Coupled Protein	R_{36} (Å)	$R_{36}^{(a)} - R_{36}^{(i)}$ (Å)
hβ_2AR (3SN6)(*11*)	active	G protein heterotrimer	13.83	4.12
hβ_2AR (2RH1)(*10*)	inactive	-	9.71	-
bRho (3PQR)(*21*)	active (metarhodopsin II)	Gα C-terminal fragment	11.44	3.69
bRho (1U19)(*22*)	inactive (rhodopsin)	-	7.75	-
hM2 (4MQS)(*13*)	active	G protein mimetic camelid antibody fragment	11.34	2.84
hM2 (3UON)(*12*)	inactive	-	8.50	-
mOPRM (5C1M)(*14*)	active	G protein mimetic camelid antibody fragment	12.21	5.08
mOPRM (4DKL)(*23*)	inactive	-	7.13	-
hAA$_{2A}$R (2YDO)(*24*)	agonist-bound	-	7.48	1.34
hAA$_{2A}$R (3EML)(*25*)	antagonist-bound	-	6.14	-

Table 2.1: Comparison of R_{36} between inactive-state and active-state GPCR crystal structures. R_{36} is the minimal approach distance between the backbone atoms of the intracellular ends of TM3 and TM6, defined in Section 2.2. R_{36} of active state structure is $R_{36}^{(\mathrm{a})}$; R_{36} of inactive state structure is $R_{36}^{(\mathrm{i})}$.

contend that this procedure is much faster and more efficient than using a standard MD simulation to identify and relax active-like conformations.

In summary, the ActiveGEnSeMBLE method (Figure 2.1) is as follows:

Step 1: Same as Step 1 of GEnSeMBLE.

Step 2: Same as Step 2 of GEnSeMBLE, except that we include a template based on an active-state crystal structure (for validation purposes only) and a hybrid

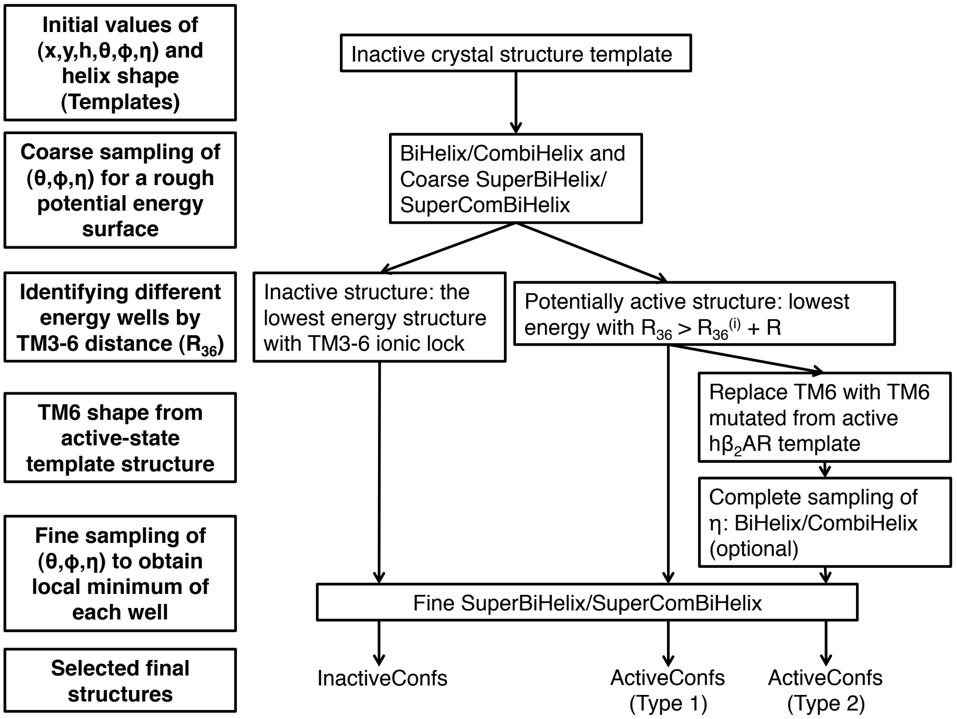

Initial values of (x,y,h,θ,ϕ,η) and helix shape (Templates)
Coarse sampling of (θ,ϕ,η) for a rough potential energy surface
Identifying different energy wells by TM3-6 distance (R36)
TM6 shape from active-state template structure
Fine sampling of (θ,ϕ,η) to obtain local minimum of each well
Selected final structures
Inactive crystal structure template
BiHelix/CombiHelix and Coarse SuperBiHelix/ SuperComBiHelix
Inactive structure: the lowest energy structure with TM3-6 ionic lock
Potentially active structure: lowest energy with R36 > R36(i) + R
Replace TM6 with TM6 mutated from active hβ2AR template
Complete sampling of η: BiHelix/CombiHelix (optional)
Fine SuperBiHelix/SuperComBiHelix
InactiveConfs
ActiveConfs (Type 1)
ActiveConfs (Type 2)

Figure 2.1: Schematic view of ActiveGEnSeMBLE. R_{36} is the minimal approach distance between the backbone atoms of the intracellular ends of TM3 and TM6, defined in Section 2.2. $R_{36}^{(i)}$ is that of the inactive structure. R is a distance, usually chosen around 4 Å. Among all cases involved in this paper, the optional step was only carried out for hSSTR5 structure prediction, and it was found to be unnecessary as the final selected structures are the same as those without this step carried out. The sampling space of BiHelix is $\Delta\eta$ from 0 to 360° in 30° increments. If not noted otherwise, the sampling space of Coarse SuperBiHelix is $\Delta\theta$: 0, ±15°; $\Delta\phi$: 0, ±45°, ±90°; $\Delta\eta$: 0, ±30°, selected angles from BiHelix/CombiHelix, starting with the best from BiHelix/CombiHelix. The sampling space of Fine SuperBiHelix is $\Delta\theta$: 0, ±15°; $\Delta\phi$: 0, ±15°, ±30°; $\Delta\eta$: 0, ±30°. The active state template for TM6 is chosen to be hβ_2AR (PDB ID: 3SN6) in this study because it is by far the only GPCR co-crystalized with a full G protein heterotrimer.

template based on the inactive-state template with an active-state TM6 for the active conformation prediction.

Step 3.1: Sample the orientations of the helices using BiHelix/CombiHelix as in GEnSeMBLE followed by a coarse SuperBiHelix/SuperComBiHelix ($\Delta\phi$ from -90° to 90° in 45° increments; $\Delta\theta$ = 0, ±15° and $\Delta\eta$ = 0, ±30° as in GEnSeMBLE). In contrast to GEnSeMBLE, in which we select the conformations corresponding to 20 lowest-energy states for further analysis, in ActiveGEnSeMBLE we measure the distance between the TM3 and TM6 IC ends (R_{36}) of 1000 lowest-energy structures generated during coarse conformational sampling. We also measure R_{36} for the inactive-state template, and denote it as $R_{36}^{(\mathrm{it})}$.

To select potential active-state structures from the coarse conformational sampling we will search for the lowest-energy structure with R_{36} - $R_{36}^{(\mathrm{it})} > 4$ Å. We denote the selected structures S2.1 (for the case with hybrid or active-state initial template) and S3.1 (for the case with inactive-state initial template). Potential inactive-state structures from the coarse conformational sampling (S4.1) are selected using the same criteria as those in GEnSeMBLE.

The definition of R_{36} and the rational behind its usage are as follows:

- Defining the distance between the TM3 and TM6 IC ends (R_{36})

  We define R_{36} to be the minimal approach distance between the IC ends of TM3 and TM6 backbone atoms. We do not define it as the distance between the two residues 3.50 and 6.30 (denoted in the Ballesteros-Weinstein numbering scheme)(*26*) that usually form a salt bridge in the inactive-state crystal structure because pure rotations of TMs 3 and 6 can increase the distance between these two (or any two) residues without opening any space between the two TMs for Gα to couple to the receptor. In order to calculate R_{36}, we use the following algorithm:

  i) The GPCR is oriented such that its hydrophobic plane is in the x-y plane (z=0), the extracellular (EC) end has positive z-coordinates, and the IC end has negative z-coordinates. For the IC ends of domains TM3 and TM6, the one with the less negative z-coordinate value is named as shortTM, and the other one as longTM.

  ii) Select a range of neighboring residues $\mathrm{r}_1, \mathrm{r}_2, \ldots, \mathrm{r}_n$ starting from the most intracellular residue of shortTM. In our example here we used $n = 4$ because there are usually 4 residues per turn on a peptide α-helix.

  iii) For each given residue rm selected in ii), determine the z-coordinate z_{m} for each of its backbone atoms. Calculate all distances between the shortTM backbone atoms in rm and the longTM backbone atoms with z-coordinate in the range (z_{m}-Δz_{IC}, z_{m}+Δz_{EC}). In general, the value of Δz_{IC} is chosen to be 5.4 Å since this is the height of one turn of the α-helix. For structure prediction steps, Δz_{EC} is chosen to be 5.4 Å. For analyzing the trajectory from the MD simulation step discussed below, the value of Δz_{EC} is chosen such that z_{m}+Δz_{EC} is about the same as the least negative z-coordinate of the Gα subunit C-terminus in the Gα-coupled case. For the structure-prediction cases discussed in this article, the latter choice of Δz_{EC} gives the same R_{36} value as choosing Δz_{EC} = 5.4 Å.

iv) The smallest distance among all distances between TM3 and TM6 calculated in iii) is R_{36}.

This definition provides a robust geometric and steric measure of the IC distance between TM3 and TM6, which correlates with the potential of G protein coupling to the active conformations.

– Using R_{36} values to facilitate selection of the active state candidates

Let R_{36} for the active state structure be $R_{36}^{(\mathrm{a})}$, and that for the inactive state structure be $R_{36}^{(\mathrm{i})}$. Define ΔR_{36} = $R_{36}^{(\mathrm{a})}$ - $R_{36}^{(\mathrm{i})}$. Class A GPCRs hβ_2AR and bRho were crystalized with the G protein or the C-terminus of the Gα subunit in complex with the GPCR, and they both have $\Delta R_{36} \approx 4$ Å (Table 2.1). For hM2 and mOPRM, their active states were crystalized with a G protein mimetic camelid antibody fragment and their ΔR_{36} are about 3 Å and 5 Å, respectively. Since the G protein couples to the receptor with the C-terminus of the Gα subunit inserted into the IC side of the GPCR in between TM3 and TM6, it is reasonable to estimate ΔR_{36} by adding the diameter of a peptide α-helix (2.3 Å) to a C-C single bond length (1.5 Å) which leads to 3.8 Å. Thus we take $\Delta R_{36} > 3$ Å as a reasonable target separation to locate active-like conformations. We show below that our final predicted active hSSTR5 structures result in $\Delta R_{36} \approx 3.5$ Å.

Step 3.2a: The structure S2.1, S3.1 or S4.1 is used directly as the starting structure of a finer SuperBiHelix/SuperComBiHelix (ϕ from -30° to 30° in 15° increments; θ and η similar to GEnSeMBLE).

Step 3.2b: The structure S3.1 has its TM6 replaced by a TM6 shape from an existing active state crystal structure according to method described in Appendix A. The resulting structure S3.1b is then used as the starting structure for fine conformational sampling ($\Delta\phi$ from -30° to 30° in 15° increments; $\Delta\theta$ and $\Delta\eta$ similar to GEnSeMBLE).

Step 3.3: To select final active-state candidates, check whether the lowest-energy structure from Step 3.2a) or Step 3.2b) at least satisfies the criterion R_{36} - $R_{36}^{(\mathrm{it})}$ > 3 Å. If it does, then select this structure as a potential active-state conformation for the target protein. If it does not, then check the second lowest-energy structure and so on. This step is carried out separately for different initial templates to have a set of candidate structures diverse in TMD shapes. The final inactive-state candidates are again selected using the same criteria as those in GEnSeMBLE.

Step 4: Same as Step 4 of GEnSeMBLE.

Step 5: Same as Step 5 of GEnSeMBLE, except that we also do MD simulation of the docked active-state candidates with the agonist bound and the G protein bound.

2.3 Validating ActiveGEnSeMBLE: Methods

Structure prediction

Starting from the TMD of crystal structures of each validation case, hβ_2AR (PDB ID: 3SN6, 2RH1) and hM2 (PDB ID: 4MQS, 3UON), we did Step 3 following the ActiveGEnSeMBLE protocol as described above. The energy E_{CNti} was used in energy ranking and is defined in Appendix A. The resulting active and inactive conformations were compared to those observed in experimental structures.

Molecular dynamics of hβ_2AR crystal structures

We did MD simulation of hβ_2AR starting from its active-state crystal structure (PDB ID: 3SN6) and inactive-state crystal structure (PDB ID: 2RH1). The MD was carried out for the following cases with explicit lipid and water environment:

- Agonist + active GPCR + Gα_s,
- Agonist + active GPCR,
- Active GPCR (apo) + Gα_s,
- Active GPCR (apo),
- Agonist + inactive GPCR,
- Inactive GPCR (apo),

- Gα_s, and
- Agonist.

We chose the agonist to be BI-167107 which is in the binding site of the active-state crystal structure for hβ_2AR. Only the Gα subunit of the G protein is included in the simulation because our main focus is on the binding interface between the GPCR and the G protein, which only involves the Gα subunit and the GPCR. The case "active GPCR (apo) + Gα_s" starts from the equilibrated "agonist + active GPCR (apo) + Gα_s" with the agonist removed.

MD simulation We used AMBER force field engine implemented in NAMD 2.9.(*27*) The conjugate gradient method was used in minimization. The Nosé-Hoover Langevin piston pressure control was used in the NPT dynamics. The 51 ns MD simulation followed the protocol detailed in Appendix A.

Energy analysis of the MD trajectories We used the self-interaction energy function of NAMD 2.9 to do a single-point energy calculation of each component in the complexes along the trajectories obtained from MD simulation above. A minimization of 5000 steps was carried out on each frame before the single-point energy was calculated. The energies were computed for the receptor, the Gα protein, the ligand, and for the interactions of Gα/ligand with the receptor, for whichever non-solvent molecules present in the complex to be studied. We then clustered the complexes along each trajectory that were saved every 100 ps such that complexes within an RMSD of 2 Å of each other were grouped into one family. The families were then classified as "inactive", "intermediate", or "active" states which are defined as $R_{36} < 10$ Å, 10 Å $< R_{36} < 13$ Å, and $R_{36} > 13$ Å, respectively. For each of these activation states of a trajectory, the mean value of the energy of each component was calculated, which is labeled E_R, E_G, E_L, E_LR, E_RG, or E_LRG with the component in consideration in the subscript. R denotes the receptor, G denotes the Gα protein, L denotes the ligand, LRG denotes the ligand+receptor+Gα complex, LR denotes the ligand+receptor complex, and RG denotes the receptor+Gα complex. We also calculated the corresponding standard deviations. We were then able to calculate the total energy of the receptor plus the stabilization from the interaction between the receptor and the ligand and/or the Gα protein as follows:

$$\begin{aligned}
E^{\mathrm{LRG}}_{\mathrm{R+interaction}} &= E^{\mathrm{LRG}}_{\mathrm{LRG}} - E^{\mathrm{LRG}}_{\mathrm{L}} - E^{\mathrm{LRG}}_{\mathrm{G}} \\
E^{\mathrm{LR}}_{\mathrm{R+interaction}} &= E^{\mathrm{LR}}_{\mathrm{LR}} - E^{\mathrm{LR}}_{\mathrm{L}} \\
E^{\mathrm{RG}}_{\mathrm{R+interaction}} &= E^{RG}_{\mathrm{RG}} - E^{\mathrm{RG}}_{\mathrm{G}} \\
E^{\mathrm{R}}_{\mathrm{R+interaction}} &= E^{\mathrm{R}}_{\mathrm{R}}.
\end{aligned}$$

The total energy of the system, with the internal energy of the ligand and the Gα protein also considered, was calculated as follows:

$$\begin{aligned}
E^{\mathrm{LRG}}_{\mathrm{Total}} &= E^{\mathrm{LRG}}_{\mathrm{LRG}} \\
E^{\mathrm{LR}}_{\mathrm{Total}} &= E^{\mathrm{LR}}_{\mathrm{LR}} + E^{\mathrm{G}}_{\mathrm{G}} \\
E^{\mathrm{RG}}_{\mathrm{Total}} &= E^{\mathrm{RG}}_{\mathrm{RG}} + E^{\mathrm{L}}_{\mathrm{L}} \\
E^{\mathrm{R}}_{\mathrm{Total}} &= E^{\mathrm{R}}_{\mathrm{R}} + E^{\mathrm{L}}_{\mathrm{L}} + E^{\mathrm{G}}_{\mathrm{G}}.
\end{aligned}$$

The superscript indicates which MD simulation case the energy is from, and the subscript indicates which components of the case were grouped to obtain the energy.

2.4 Validating ActiveGEnSeMBLE: Results and Discussion

Structure prediction

We validated the ActiveGEnSeMBLE method with hβ_2AR and hM2 receptor systems, which have both their active-state and inactive-state structures crystalized. For bRho, its ligand, retinal, is covalently bound to the GPCR. As this does not represent the majority of Class A GPCRs which do not have covalently bound ligands, we did not consider bRho as a validation case. In addition, it is nontrivial to quantify the energies of the receptor with and without the ligand for a covalently bound ligand to account for the effect of the receptor-ligand interaction energy on the thermodynamic state of the receptor system. Figure 2.2 summarizes the methods tested, with each final structure sharing the same numbering as the method that generated the structure. Starting from an active-state structure, an inactive-state structure, and a hybrid structure mixing active state (TM6) and inactive state (TM1-5 and TM7) helices, we compared the best final structures from different methods with the active-state crystal structure. The structural features and energy value of the last node in every pipeline in Figure 2.2 (i.e. the final structure of each method) are summarized in Table 2.2, with further details in Table 2.3. We assumed that the energy values and R_{36} values are the only information that will be available in selecting candidate structures for prediction of an unknown structure. For hβ_2AR and hM2, using the active-state crystal structure as a starting structure aiming at predicting active-state structures (Methods 1.x), the procedure is able to reproduce the TM orientations in the active-state crystal structure (Table 2.2 and 2.3). This is a good but necessary test of the overall methodology and the force field as they are able to identify a conformation close to the experimentally observed conformation out of $(5\times3\times3)^7 \approx 374$ billion conformations sampled. Similarly, using the inactive-state crystal structure as a starting structure aiming at predicting inactive-state structures (Methods 4.x), the procedure is able to reproduce the TM orientations in the inactive-state crystal structure (Table 2.2 and 2.3).

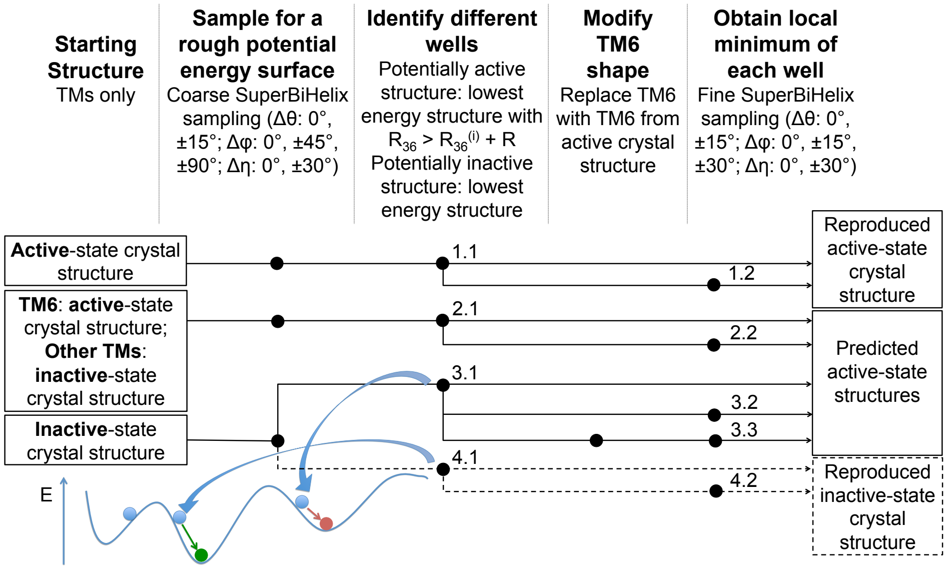

Starting Structure
TMs only
Sample for a rough potential energy surface
Coarse SuperBiHelix sampling (Δθ: 0°, ±15°; Δφ: 0°, ±45°, ±90°; Δη: 0°, ±30°)
Identify different wells
Potentially active structure: lowest energy structure with R36 > R36(i) + R
Potentially inactive structure: lowest energy structure
Modify TM6 shape
Replace TM6 with TM6 from active crystal structure
Obtain local minimum of each well
Fine SuperBiHelix sampling (Δθ: 0°, ±15°; Δφ: 0°, ±15°, ±30°; Δη: 0°, ±30°)
Active-state crystal structure
TM6: active-state crystal structure; Other TMs: inactive-state crystal structure
Inactive-state crystal structure
1.1
1.2
2.1
2.2
3.1
3.2
3.3
4.1
4.2
E
Reproduced active-state crystal structure
Predicted active-state structures
Reproduced inactive-state crystal structure

Figure 2.2: Methods for validating inactive- and active-state predictions using hβ_2AR and hM2 as test cases. The black node indicates that the corresponding step is applied to the structure from the previous black node on the same line on the left. Each number beside a black node denotes the optimal structure obtained from the method that is denoted by the same number.

Next, as a test of ActiveGEnSeMBLE on more practical cases, we find that both active-state prediction methods either starting from the inactive-state crystal structure or the hybrid structure (Methods 2.x and 3.x, see rows with the first column 2.x and 3.x in Table 2.2) can reduce the RMSD of the predicted active-state candidate to the active-state crystal structure by 1.0 Å for hβ_2AR and 0.8 Å for hM2. These numbers are significant as the RMSD between the inactive-state and active-state crystal structures are 2.48 Å and 2.30 Å for hβ_2AR and hM2, respectively. Com-

Structure / Method Identifier	Method Name	Method Description			hβ_2AR			hM2		
		Sampling Target	Starting Structure	Sampling Method	E_{CNti} (kcal mol^{-1})	RMSD to Target Crystal Structure (Å)	R_{36} (Å)	E_{CNti} (kcal mol^{-1})	RMSD to Target Crystal Structure (Å)	R_{36} (Å)
Active crystal structure	-	Active State	Active crystal structure	None	-	0.00	13.72	-	0.00	11.14
Inactive crystal structure	-	Active State	Active crystal structure	None	-	2.48	9.71	-	2.30	8.55
1.1	AfromA_Coarse	Active state	Active crystal structure	Coarse	-146.8	0.04	13.68	-168.5	0.37	11.15
1.2	AfromA_Fine	Active state	Active crystal structure	Fine after Coarse	-146.8	0.04	13.68	-168.5	0.37	11.15
2.1	AfromH_Coarse	Active state	Active crystal TM6 + inactive other TMs	Coarse	-158.0	1.41	15.12	-96.9	1.55	13.60
2.2	AfromH_Fine	Active state	Active crystal TM6 + inactive other TMs	Fine after Coarse	-196.9	1.48	14.64	-100.3	1.69	13.45
3.1	AfromI_Coarse	Active state	Inactive crystal structure	Coarse	-95.0	1.79	12.85	-77.4	1.94	14.45
3.2	AfromI_Fine	Active state	Inactive crystal structure	Fine after Coarse	-114.5	1.86	13.41	-127.2	2.68	15.20
								-111.6	1.92	15.22
3.3	AfromIH_Fine	Active state	Inactive crystal structure	Fine after replacing TM6 of 3.1 result by active crystal structure TM6	-169.5	1.47	14.03	-113.2	1.52	14.44
Inactive crystal structure	-	Inactive state	Inactive crystal structure	None	-	0.00	9.71	-	0.00	8.55
4.1	IfromI_Coarse	Inactive state	Inactive crystal structure	Coarse	-220.6	0.07	9.72	-168.8	0.04	8.54
4.2	IfromI_Fine	Inactive state	Inactive crystal structure	Fine after Coarse	-220.6	0.07	9.72	-168.8	0.04	8.54

Table 2.2: Summary of structural features and energies of hβ_2AR and hM2 optimal structures generated from different methods. The "Structure Identifier" corresponds to the numbers in Figure 2.2. Each RMSD value is between the backbone atoms of the resulting optimal structure of the particular method and the backbone atoms of the active-state crystal structure preprocessed according to Appendix A. For hM2, the second best choice of 3.2 is also listed for comparison. Unlike hβ_2AR for which the inactive state is 73.8 kcal mol^{-1} more stable than the active state, this value for hM2 is only 3.7 kcal mol^{-1}. Therefore, the number of seven-helix bundles built from SuperBiHelix results is increased to 2500 to capture more candidates in the active-like regime. (For hβ_2AR, building 1000 or 2500 bundles has the same the final results shown in this table.) While hM2's lowest-energy structure from AfromI_Fine that satisfies the active-like R_{36} criterion (upper row of 3.2 in the table) has a much larger RMSD than Structure 3.1 to the active-state crystal structure, the second-lowest-energy structure (lower row of 3.2 in the table) has an improved RMSD comparing to 3.1. This suggests selecting a small number of diverse structures from the potential energy well may help in active-state structure prediction as well.

a)

Structure / Method Identifier	Method Name	Δθ (°)	Δφ (°)							Δη (°)							E_{CNti} (kcal mol^{-1})	RMSD to crystal active (Å)	R_{36} (Å)
		H1 – H7	H1	H2	H3	H4	H5	H6	H7	H1	H2	H3	H4	H5	H6	H7			
Active crystal structure	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	N/A	0.00	13.72
Inactive crystal structure	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	N/A	2.48	9.71
1.1	AfromA_Coarse	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-146.8	0.04	13.68
1.2	AfromA_Fine	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-146.8	0.04	13.68
2.1	AfromH_Coarse	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-158.0	1.44	15.12
2.2	AfromH_Fine	0	0	0	0	-15	0	15	15	0	0	0	0	0	0	0	-196.9	1.48	14.64
3.1	AfromI_Coarse	0	0	0	0	0	0	-45	0	0	0	0	0	0	-30	0	-95.0	1.79	12.85
3.2	AfromI_Fine	0	0	0	0	-15	0	-75	30	0	0	0	0	0	-60	0	-114.5	1.86	13.41
3.3	AfromIH_Fine	0	0	0	0	15	0	15	15	0	0	0	0	0	0	0	-169.5	1.47	14.03
4.1	IfromI_Coarse	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-220.6	2.47	9.72
4.2	IfromI_Fine	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-220.6	2.47	9.72

b)

Structure / Method Identifier	Method Name	Δθ (°)							Δφ (°)							Δη (°)							E_{CNti} (kcal mol^{-1})	RMSD to crystal active (Å)	R_{36} (Å)
		H1	H2	H3	H4	H5	H6	H7	H1	H2	H3	H4	H5	H6	H7	H1	H2	H3	H4	H5	H6	H7			
Active crystal structure	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	N/A	0.00	11.14
Inactive crystal structure	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	N/A	2.30	8.55
1.1	AfromA_Coarse	0	0	0	0	0	0	0	0	0	0	-45	0	0	0	0	0	0	0	0	0	0	-168.5	0.37	11.15
1.2	AfromA_Fine	0	0	0	0	0	0	0	0	0	0	-45	0	0	0	0	0	0	0	0	0	0	-168.5	0.37	11.15
2.1	AfromH_Coarse	0	0	0	0	0	0	-15	0	0	0	0	0	0	-90	0	0	0	0	0	-30	0	-96.9	1.55	13.60
2.2	AfromH_Fine	0	0	0	0	0	0	0	0	0	0	-15	15	15	-60	0	0	0	0	0	-30	0	-100.3	1.69	13.45
3.1	AfromI_Coarse	0	0	0	0	15	15	-15	0	0	0	0	0	-90	0	0	0	0	0	0	0	0	-77.4	1.94	14.45
3.2	AfromI_Fine	0	0	0	0	30	15	0	0	0	0	-15	15	-60	0	0	0	0	0	-30	0	0	-127.2	2.68	15.20
		0	0	0	0	15	15	0	0	0	0	-15	30	-60	15	0	0	0	0	0	0	0	-111.6	1.92	15.22
3.3	AfromIH_Fine	0	0	0	0	0	0	0	0	0	0	0	15	0	15	0	0	0	0	0	-30	0	-113.2	1.52	14.44
4.1	IfromI_Coarse	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-168.8	0.04	8.54
4.2	IfromI_Fine	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-168.8	0.04	8.54

Table 2.3: Summary of structural features and energies of a) hβ_2AR and b) hM2 optimal structures generated from different methods. The "Method Identifier" corresponds to the numbers in Figure 2.2. Each RMSD value is between the backbone atoms of the resulting optimal structure of the particular method and the backbone atoms of the active-state crystal structure. For TM4, none-zero $\Delta\phi$ values do not have as big an effect as $\Delta\phi$ on the orientations of other TMs because the absolute value of θ for TM4 is close to zero.

paring different methods that start from inactive-state crystal structure or the hybrid structure (Methods 2.1 with 3.1 as well as Methods 3.3 with 3.2 in Table 2.2), it can be seen that replacing inactive-state TM6 shape with active-state TM6 shape further lowers the RMSD value between the candidate structure from sampling and the actual active-state crystal structure by 0.4 Å more. This implies that the active-like TM6 shape plays an important role and may be necessary for a high-accuracy computational prediction of active-state GPCR structures.

From the coarse sampling results, two structures are picked. One is for the inactive-state prediction (Structure 4.1 of Figure 2.2, Tables 2.2 and 2.3), and the other is for the active-state prediction (Structure 3.1 of Figure 2.2, Tables 2.2 and 2.3). From the energy profiles in Figure 2.3 (hβ_2AR) and Figure 2.4 (hM2), it can be found that the fine samplings can effectively achieve lower energies while keeping R_{36} of its lowest-energy final structures in the active-like energy well to be similar (within 1 Å) to the starting structures of the corresponding fine sampling. In other words, the fine sampling helps to locate the local minimum of a potential energy well (defined by R_{36}) even when the activation coordinate R_{36} was not sampled explicitly. These potential energy profiles, while being crude, are also consistent with the hypotheses that structures with R_{36} that deviate more than 3-4 Å from $R_{36}^{(\mathrm{i})}$ correspond to potential energy wells with higher energy local minima. In addition, they show that there are multiple higher-energy potential energy wells, which is consistent with biophysical evidence for multiple active states for a given GPCR. Furthermore, the recently published arrestin-bound receptor structure (*28*) was used to match arrestin to all the minimum energy structures of these energy wells for hβ_2AR. We found conformations for hβ_2AR that could accommodate arrestin but not the G$_\mathrm{s}$ protein (Figure 2.5). This is also consistent with the hypothesis that some of these different active states might be capable of activating different signaling pathways by coupling to different regulators. These results demonstrate that ActiveGEnSeMBLE method is able to predict functionally distinct active-like conformations of GPCRs that might be responsible for coupling to different signaling pathways. The method has the potential to map out the activation landscape of a GPCR for multiple signaling pathways efficiently.

MD simulation and analysis

So far, only qualitative energy landscapes have been generated for GPCR activation, among which the most well-studied case is hβ_2AR.(*29*) To obtain a quantitative energy landscape that may provide more insights into GPCR activation, and to have

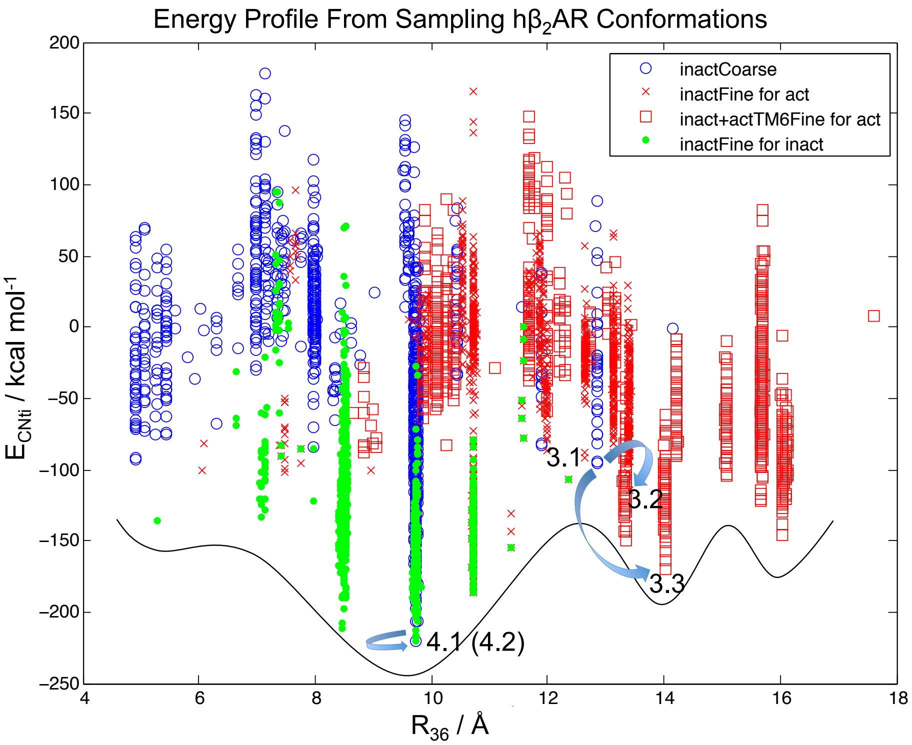


Figure 2.3: Potential energy profile from sampling hβ_2AR conformations. The black curve illustrates how our sampling results can be qualitatively translated into a potential energy curve using R_{36} as the x axis and does not quantitatively represent any real data. Results of the coarse sampling starting from the inactive-state crystal structure are in blue circles. Starting from structure 3.1, results of methods that generated structures 3.2 and 3.3 are shown in red crosses and red squares, respectively. Starting from structure 4.1, results of the fine sampling that generated structure 4.2 are in green dots. Every blue arrow points from a starting structure toward the optimal structure from the fine sampling of the corresponding method.

a strategy that can facilitate the understanding of activation of GPCRs for which only predicted structures are available, we carried out MD simulation of hβ_2AR starting from its crystal structures as described in Section 2.3. The analysis of R_{36} during the MD simulations showed that the inactive conformation with and without the agonist remains inactive (Figure 2.6), the Gα_s-bound active conformation remains stable during MD, and the active conformation not bound to the Gα_s slowly loses its activity (decrease in R_{36}) during the 51 ns MD simulation as expected (Figure 2.7). In addition, the apo-GPCR on average always has a slightly smaller R_{36} value than the agonist-bound GPCR towards the end of the 51 ns, which is consistent with the picture that the agonist shifts the equilibrium towards more activated states.

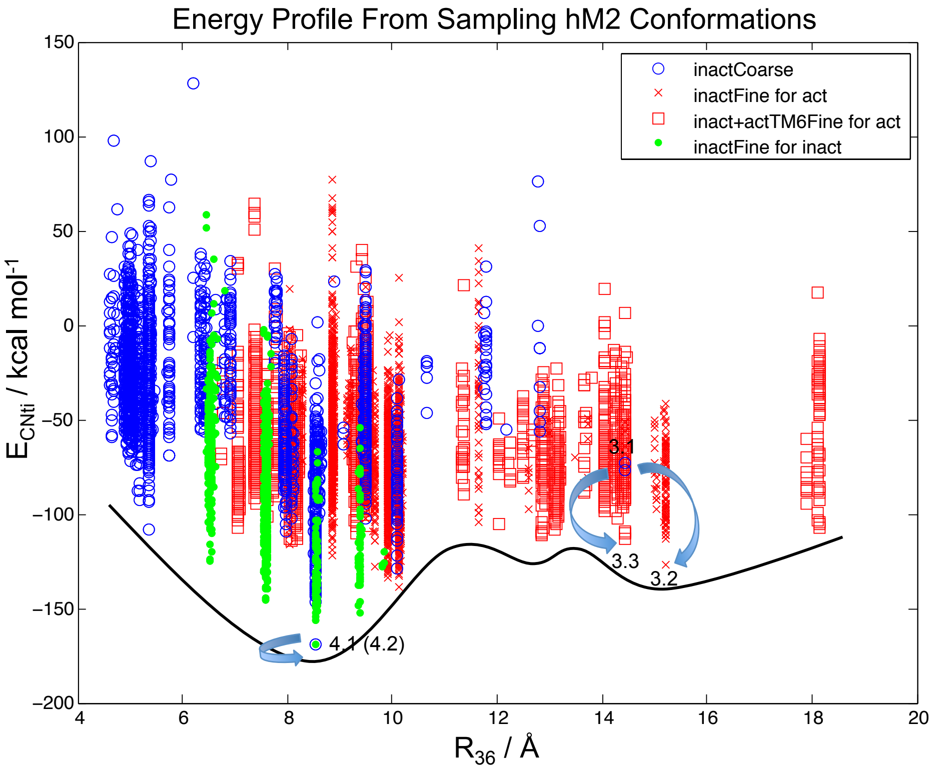


Figure 2.4: Potential energy profile from sampling hM2 conformations. The black curve is an illustration showing how our sampling results can qualitatively be translated into a potential energy curve using R_{36} as the x-axis and does not quantitatively represent any real data. Results of the coarse sampling starting from the inactive-state crystal structure are in blue circles. Starting from Structure 3.1, results of methods that generated Structures 3.2 and 3.3 are shown in red crosses and red squares respectively. Starting from the Structure 4.1, results of the fine sampling that generated Structure 4.2 are in green dots. Every blue arrow points from a starting structure towards the optimal structure from the corresponding sampling method.

We then grouped the conformations in corresponding trajectories by R_{36}, with “inactive”, “intermediate”, and “active” states defined as R_{36} < 10 Å, 10 Å < R_{36} < 13 Å, and R_{36} > 13 Å, respectively, and calculated $E_{\mathrm{R+interaction}}$ for each group. Note that each of the “inactive”, “intermediate”, and “active” state here contain multiple 3D conformational states that satisfy the respective R_{36} criterion. The resulting energy landscape (Figure 2.8) is overall consistent with the qualitative picture from experiments [Figure 2.9 adapted from ref.(*29*)].

For agonist-bound (BI-167107 being the agonist) and apo-GPCR systems, our inactive-state energy is lower than the active state but higher than the interme-

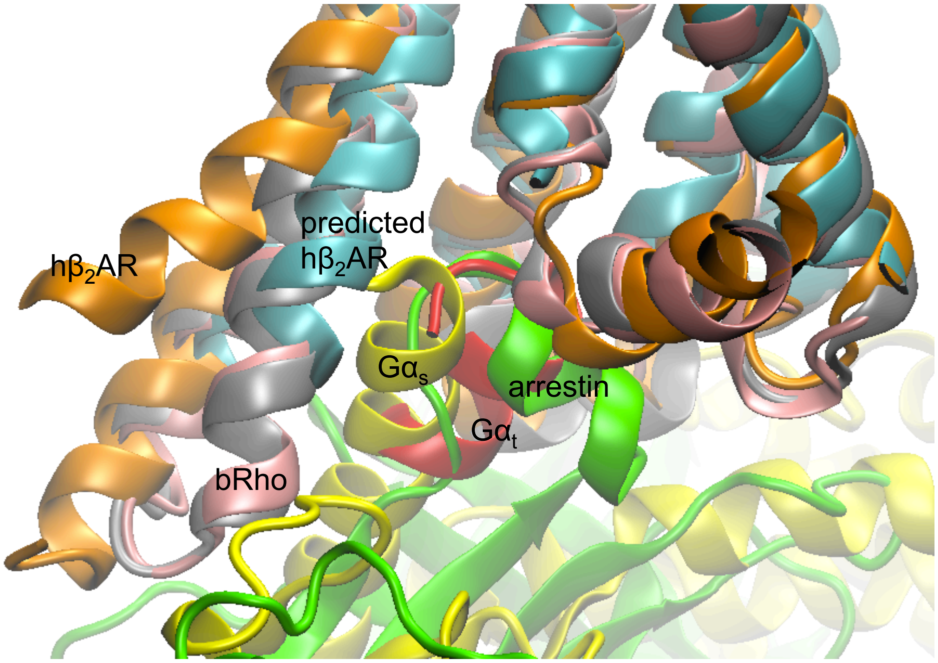


Figure 2.5: Predicted minimum-energy hβ_2AR structure (cyan) in one of the energy wells can accommodate arrestin (green) but not Gα_s (yellow). An energy well is defined by clustering the structures with R_{36} values within 0.25 Å of each other. We positioned the Gα_s protein and the arrestin by aligning hβ_2AR (orange) in its Gα_s-coupled crystal structure (PDB ID: 3SN6)(*11*), bRho (grey) in its arrestin-coupled crystal structure (PDB ID: 4ZWJ)(*28*) and bRho (pink) in its Gα_t-C-terminus-coupled crystal structure (PDB ID: 3PQR)(*21*) to the predicted hβ_2AR structure without changing the relative orientation within each crystal structure. We find the minimum energy structure of hβ_2AR in the $R_{36} \approx 12.4$ Å well clashes with Gα_s but not with arrestin. It is hard to determine by speculation whether it distinguishes arrestin with the G$_i$ protein because the C-terminus helix of Gα_t subunit (red) co-crystalized with the active-state bRho is similar in position with the part of arrestin inside the GPCR. (Therefore, similar analysis was not carried out on hM2 which couples with the G$_i$ protein instead of G$_s$ in vivo.) This structure is from the ensemble of structures generated using Method 3.2 which uses the inactive-state TM shapes for all TMs. It is characterized by $\Delta\theta = 0$ for all helices; $\Delta\phi$ = -15°, 0, 0, 15°, 0, -30°, 15°; $\Delta\eta$ = 0, 0, 0, 0, 0, -30°, 0 with the values ordered from TM1 to TM7. It has E_{CNti} = -114.1 kcal mol^{-1}.

diate state (Fig. 3). This is a signature of BI-167107-bound hβ_2AR as opposed to isoproterenol-bound hβ_2AR and has been found in these experiments.(*29*)

When the Gα protein is present, both agonist-bound and apo-GPCR energies are significantly lowered for the “active state”, which is the only group of states that can accommodate the G protein. This finding for the agonist-bound GPCR’s energy

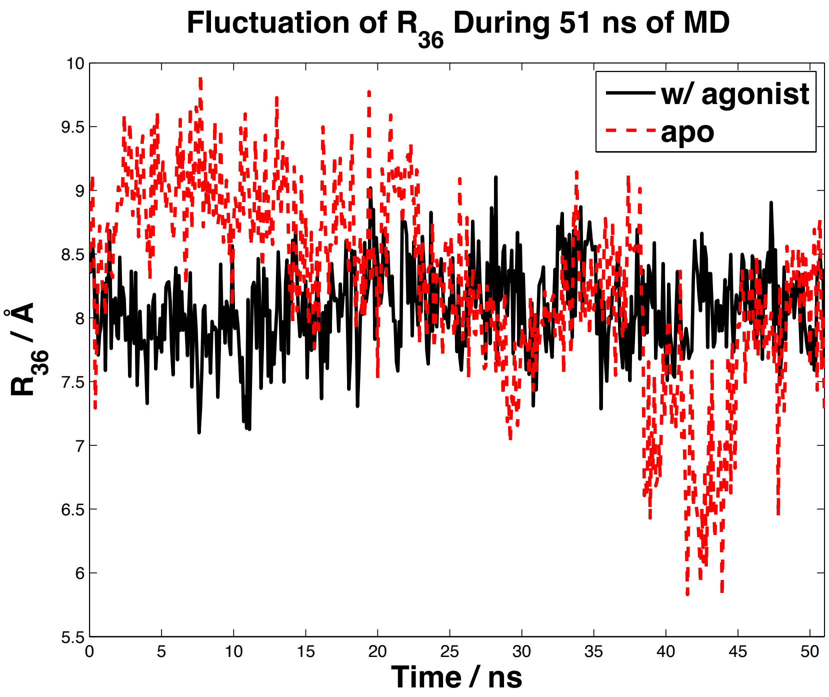


Figure 2.6: Fluctuation of R_{36} of h$\beta 2_2$AR during 51 ns molecular dynamics simulation starting from the inactive-state structure.

lowering upon coupling with Gα is consistent with experiments. Apo-GPCR coupled with the G protein is not in the picture of the experimental energy landscape, but we are able to do MD simulation and analysis of this system, and rationalize our finding on its relative energy. To be specific, our results suggest that if an apo-GPCR can indeed couple with the G protein, its energy will be lowered upon coupling but not as low as the agonist-bound GPCR. This explains the basal activity of hβ_2AR and suggests that an agonist is able to increase the activity of a GPCR by shifting the equilibrium from the apo-GPCR+Gα complex towards the more stable agonist+GPCR+Gα complex.

Furthermore, the agonist-bound GPCR has lower energy than the apo-GPCR for all other states as well, including the inactive states. This is also what has been found in the experiments.(*29*) As the agonist-bound "intermediate state" remains the lowest-energy state among the three, and the energy is lowered to a greater extent than the other states, this supports a second route for GPCR activation, which starts from an increased equilibrium population of the intermediate state upon binding with the agonist. As the agonist-bound receptor is more dynamic in conformation than apo-GPCR,(*29*) the agonist-bound receptor can then be stabilized by the G

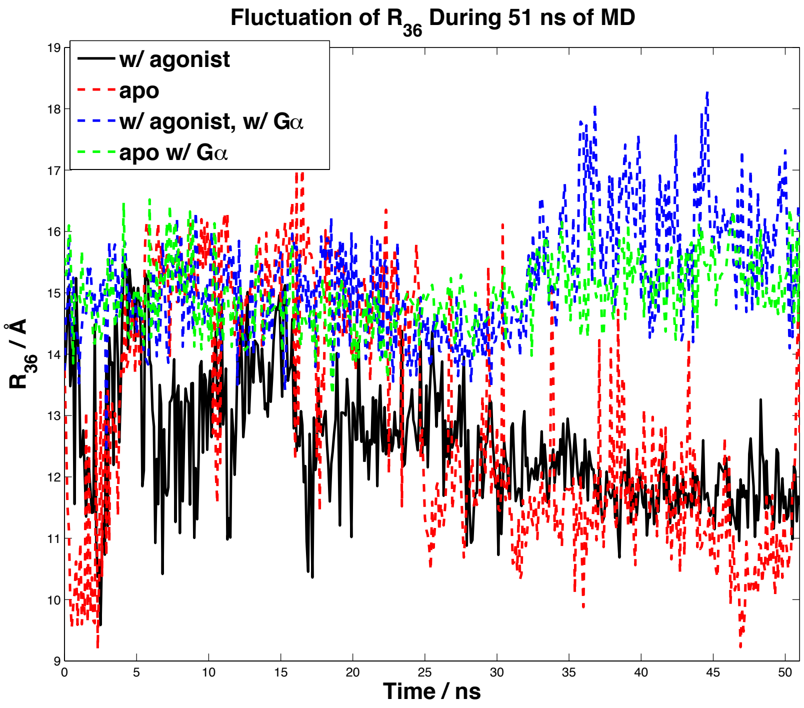


Figure 2.7: Fluctuation of R_{36} of hβ_2AR during 51 ns molecular dynamics simulation starting from the active-state structure.

protein as it transits from the intermediate state to the active state.

The above analysis is based on the quantitative energy landscape plotted with $E_{\text{R+interaction}}$, which is chosen because we would like to eliminate the possible internal energy changes in the Gα and the ligand caused by Gα not coupling with the other subunits of the G protein and by the ligand strain, which may not be captured accurately by the computational methods. Nevertheless, energy profiles of E_{R} and E_{Total} are shown in Figure 2.10 and Figure 2.11, respectively. Although E_{Total} is qualitatively the same as $E_{\text{R+interaction}}$ for hβ_2AR, later we would see that they are different for hSSTR5 (Figure 3.9). The energy profile of the receptor by itself (E_{R}) shows that the receptor is actually destabilized by the G protein if the agonist is not present. Combining with the previous analysis, we conclude that, upon the GPCR coupling to the G protein, it is the interaction between the G protein and the GPCR that stabilizes the system, instead of the G protein directly lowering the energy of the GPCR by itself. This insight was lacking from experimental profiles, but we are able to deduce it from our analysis.

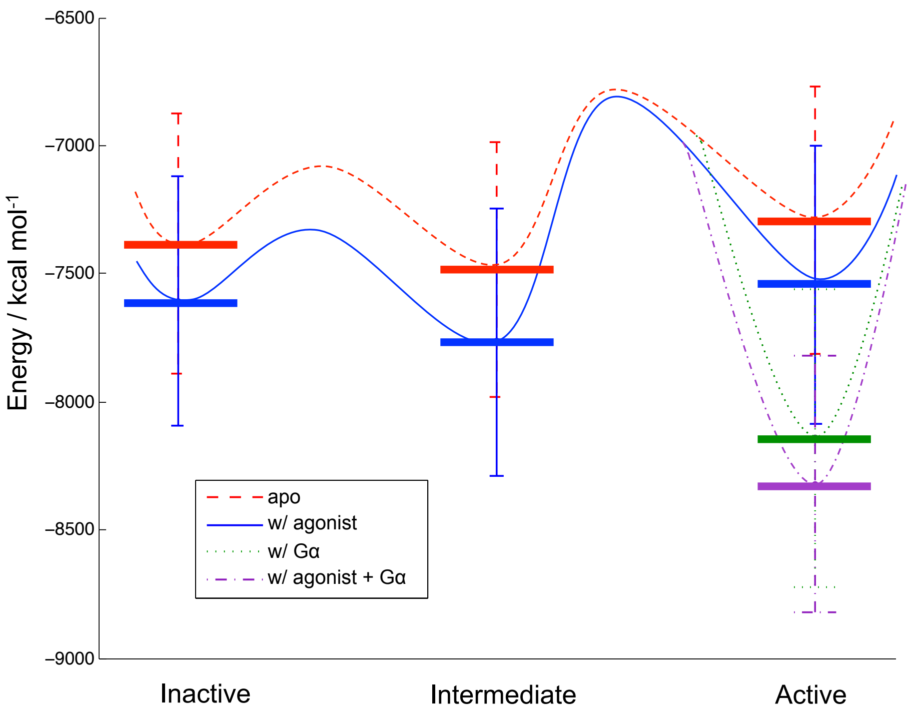


Figure 2.8: Energy landscape of hβ_2AR activation from MD simulation. The horizontal bars are $E_{\mathrm{R+interaction}}$ calculated as described in Section 2.3. The curved lines are fictitious energy surfaces, with the barriers being qualitative and the minima defined by the corresponding $E_{\mathrm{R+interaction}}$ values. Inactive, intermediate, and active states in the figure are defined as R_{36} < 10 Å, 10 Å < R_{36} < 13 Å, and R_{36} > 13 Å, respectively.

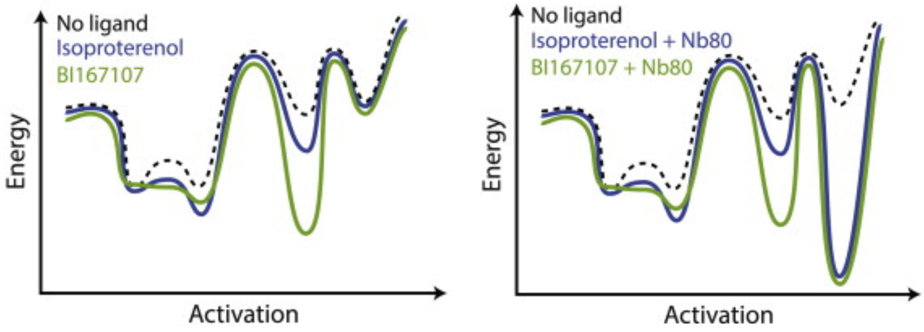


Figure 2.9: Qualitative energy landscape of hβ_2AR from experiments. This figure is adapted from Manglik et al. 2015. Cell 161:1101-1111 with permission.

2.5 Conclusions and Future Work

We have presented a new, to our knowledge, method for GPCR structure prediction, termed ActiveGEnSeMBLE, that overcomes the conformational sampling limits

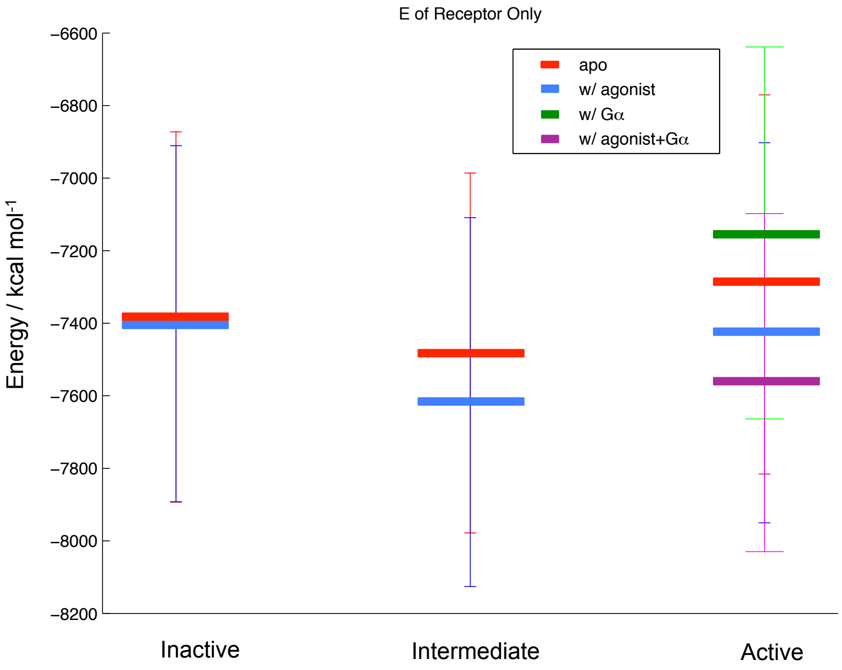


Figure 2.10: Energy profile of hβ_2AR during activation. The horizontal bars are E_R obtained according to Section 2.3. "Inactive", "intermediate", and "active" states in the figure are defined as $R_{36} < 10$ Å, 10 Å $< R_{36} < 13$ Å, and $R_{36} > 13$ Å respectively.

of MD simulations. This method can be used to identify multiple energetically accessible conformations for a GPCR that might play a role in its activation in addition to multiple lower-energy structures that might correspond to in inactive states. We validate ActiveGEnSeMBLE by predicting the active hβ_2AR and hM2 crystal structures. We found that ActiveGEnSeMBLE sampled the orientations of the TM helices and located structures in various energy wells spanning the range of TM3–TM6 distances (R_{36}) traversed in the process of activation. Subsequent analysis finds a local minimum in each of these energy wells that was close or identical to a crystal-structure conformation with a similar R_{36} value. MD simulations of the crystal structures of hβ_2AR with and without the G protein and the agonist generated energy profiles that are consistent with the qualitative energy landscape of hβ_2AR obtained from experiments, providing information about how the ligand and G protein may play are role in activation. These results indicate that the agonist alone is not enough to stabilize the active state and that the Gα C-terminal chain needs to be bound to the GPCR to promote activation, in agreement with conclusions from experiments.

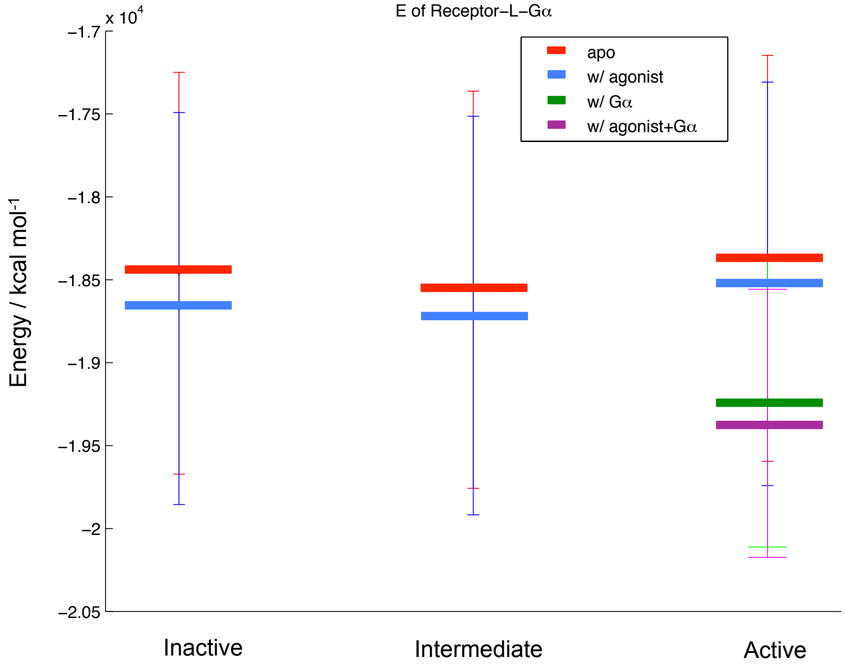


Figure 2.11: Energy profile of the system during hβ_2AR activation. The horizontal bars are E_{Total} calculated according to Section 2.3. "Inactive", "intermediate", and "active" states in the figure are defined as R_{36} < 10 Å, 10 Å < R_{36} < 13 Å, and R_{36} > 13 Å respectively.

Combined with results from applying ActiveGEnSeMBLE to hSSTR5, which are discussed in the next chapter, it is confirmed that ActiveGEnSeMBLE is an effective new method in predicting active-state conformations of at least class A GPCRs. To our knowledge, it enables the first quantitative energy profile for GPCR activation that is consistent with the qualitative profile deduced from experiments.

Future work can follow two directions: First, we could calculating the free energy instead of potential energy to generate the energy landscape of GPCR activation. This free energy profile should be quantitatively more accurate than a potential energy profile. Second, we could apply ActiveGEnSeMBLE on a large number of different GPCRs to map out their energy landscape of activation and compare the differences in their activation mechanisms. This will allow us to study GPCR activation systematically, and to further study the interactions between different GPCRs and their signaling pathways. This is highly meaningful not only for understanding the fundamentals of biology, but also for designing more effective drugging strategies, such as combining several drugs that respectively target on several selected

GPCRs to diminish drug resistance.

In addition, the ActiveGEnSeMBLE strategy is possible to be generalized to other macromolecules that the conformations and energetics from the large-scale motion of well-defined and relatively-rigid domains are of interest. As discussed in Chapter 1, apart from the hierarchical screening of grid points on the potential energy surface, a key ingredient of ActiveGEnSeMBLE is a selection rule that is based on suitable collective variable(s) specific to the target system. The identification of such variables in this thesis was based on human observation of available data to find the most prominent features, partly because of the small size of the available data set and the uncertainty in data being too large for more subtle features. However, with generalization of our strategy in mind, a more systematic identification process (e.g. applying a feature selection algorithm) can be more proper for certain systems that more high-quality data is available.

References

(*1*) Kenakin, T., and Miller, L. J., (2010). Seven transmembrane receptors as shapeshifting proteins: the impact of allosteric modulation and functional selectivity on new drug discovery. *Pharmacological reviews 62*, 265–304.

(*2*) Abrol, R., Griffith, A. R., Bray, J. K., and Goddard III, W. A., (2012). Structure prediction of G protein-coupled receptors and their ensemble of functionally important conformations. *Methods in Molecular Biology (Clifton, NJ) 914*, 237.

(*3*) Abrol, R., Trzaskowski, B., Goddard, W. A., Nesterov, A., Olave, I., and Irons, C., (2014). Ligand-and mutation-induced conformational selection in the CCR5 chemokine G protein-coupled receptor. *Proceedings of the National Academy of Sciences 111*, 13040–13045.

(*4*) Kim, S.-K., Riley, L., Abrol, R., Jacobson, K. A., and Goddard III, W. A., (2011). Predicted structures of agonist and antagonist bound complexes of adenosine A3 receptor. *Proteins: Structure, Function, and Bioinformatics 79*, 1878–1897.

(*5*) Scott, C. E., Abrol, R., Ahn, K. H., Kendall, D. A., and Goddard III, W. A., (2013). Molecular basis for dramatic changes in cannabinoid CB1 G protein-coupled receptor activation upon single and double point mutations. *Protein Science 22*, 101–113.

(*6*) Tan, J., Abrol, R., Trzaskowski, B., and Goddard III, W. A., (2012). 3D structure prediction of TAS2R38 bitter receptors bound to agonists phenylthiocarbamide (PTC) and 6-n-Propylthiouracil (PROP). *Journal of chemical information and modeling 52*, 1875–1885.

(*7*) Kim, S.-K., and Goddard III, W. A., (2014). Predicted 3D structures of olfactory receptors with details of odorant binding to OR1G1. *Journal of computer-aided molecular design*, 1–16.

(*8*) Kirkpatrick, A., Heo, J., Abrol, R., and Goddard, W. A., (2012). Predicted structure of agonist-bound glucagon-like peptide 1 receptor, a class B G protein-coupled receptor. *Proceedings of the National Academy of Sciences 109*, 19988–19993.

(*9*) Li, Y., Zhu, F., Vaidehi, N., Goddard, W. A., Sheinerman, F., Reiling, S., Morize, I., Mu, L., Harris, K., and Ardati, A., (2007). Prediction of the 3D structure and dynamics of human DP G-protein coupled receptor bound to an agonist and an antagonist. *Journal of the American Chemical Society 129*, 10720–10731.

(*10*) 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., and Kobilka, B. K., (2007). High-resolution crystal structure of an engineered human β2-adrenergic G protein-coupled receptors. *Science 318*, 1258–1265.

(*11*) Rasmussen, S. G. F., et al. (2011). Crystal structure of the β2 adrenergic receptor-Gs protein complex. *Nature 477*, 549–555, DOI: `10.1038/nature10361`.

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

(*13*) Kruse, A. C., Ring, A. M., Manglik, A., Hu, J., Hu, K., Eitel, K., Hübner, H., Pardon, E., Valant, C., and Sexton, P. M., (2013). Activation and allosteric modulation of a muscarinic acetylcholine receptor. *Nature 504*, 101–106.

(*14*) Huang, W., Manglik, A., Venkatakrishnan, A., Laeremans, T., Feinberg, E. N., Sanborn, A. L., Kato, H. E., Livingston, K. E., Thorsen, T. S., and Kling, R. C., (2015). Structural insights into μ-opioid receptor activation. *Nature 524*, 315–321.

(*15*) Li, D., Tanaka, M., Brunicardi, F. C., Fisher, W. E., Gibbs, R. A., and Gingras, M.-C., (2011). Association between somatostatin receptor 5 gene polymorphisms and pancreatic cancer risk and survival. *Cancer 117*, 2863–2872.

(*16*) Mobarec, J. C., Sanchez, R., and Filizola, M., (2009). Modern homology modeling of G-protein coupled receptors: which structural template to use? *Journal of medicinal chemistry 52*, 5207–5216.

(*17*) Trzaskowski, B., Latek, D., Yuan, S., Ghoshdastider, U., Debinski, A., and Filipek, S., (2012). Action of molecular switches in GPCRs-theoretical and experimental studies. *Current medicinal chemistry 19*, 1090.

(*18*) Floriano, W. B., Vaidehi, N., Zamanakos, G., and Goddard, W. A., (2004). HierVLS hierarchical docking protocol for virtual ligand screening of large-molecule databases. *Journal of medicinal chemistry 47*, 56–71.

(*19*) Goddard, W. A., Kim, S.-K., Li, Y., Trzaskowski, B., Griffith, A. R., and Abrol, R., (2010). Predicted 3D structures for adenosine receptors bound to ligands: Comparison to the crystal structure. *Journal of structural biology 170*, 10–20.

(*20*) Manglik, A., and Kobilka, B., (2014). The role of protein dynamics in GPCR function: Insights from the β2AR and rhodopsin. *Current opinion in cell biology 27*, 136–143.

(*21*) Choe, H.-W., Kim, Y. J., Park, J. H., Morizumi, T., Pai, E. F., Krauß, N., Hofmann, K. P., Scheerer, P., and Ernst, O. P., (2011). Crystal structure of metarhodopsin II. *Nature 471*, 651–655.

(*22*) Okada, T., Sugihara, M., Bondar, A.-N., Elstner, M., Entel, P., and Buss, V., (2004). The retinal conformation and its environment in rhodopsin in light of a new 2.2 Å crystal structures. *Journal of molecular biology 342*, 571–583.

(*23*) Manglik, A., Kruse, A. C., Kobilka, T. S., Thian, F. S., Mathiesen, J. M., Sunahara, R. K., Pardo, L., Weis, W. I., Kobilka, B. K., and Granier, S., (2012). Crystal structure of the μ-opioid receptor bound to a morphinan antagonists. *Nature 485*, 321–326.

(*24*) Lebon, G., Warne, T., Edwards, P. C., Bennett, K., Langmead, C. J., Leslie, A. G., and Tate, C. G., (2011). Agonist-bound adenosine A2A receptor structures reveal common features of GPCR activation. *Nature 474*, 521–525.

(*25*) Jaakola, V.-P., Griffith, M. T., Hanson, M. A., Cherezov, V., Chien, E. Y., Lane, J. R., Ijzerman, A. P., and Stevens, R. C., (2008). The 2.6 angstrom crystal structure of a human A2A adenosine receptor bound to an antagonist. *Science 322*, 1211–1217.

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

(*27*) Phillips, J. C., Braun, R., Wang, W., Gumbart, J., Tajkhorshid, E., Villa, E., Chipot, C., Skeel, R. D., Kale, L., and Schulten, K., (2005). Scalable molecular dynamics with NAMD. *Journal of computational chemistry 26*, 1781–1802.

(*28*) Kang, Y., Zhou, X. E., Gao, X., He, Y., Liu, W., Ishchenko, A., Barty, A., White, T. A., Yefanov, O., and Han, G. W., (2015). Crystal structure of rhodopsin bound to arrestin by femtosecond X-ray laser. *Nature 523*, 561–567.

(*29*) Manglik, A., Kim, T. H., Masureel, M., Altenbach, C., Yang, Z., Hilger, D., Lerch, M. T., Kobilka, T. S., Thian, F. S., and Hubbell, W. L., (2015). Structural Insights into the Dynamic Process of $\beta 2$-Adrenergic Receptor Signaling. *Cell 161*, 1101–1111.

Chapter 3

The Activation of Human Somatostatin Receptor 5 (hSSTR5)

This chapter is based on the following publications:

Dong, S. S.; Goddard, W. A.; Abrol, R. "Conformational and Thermodynamic Landscape of GPCR Activation From Theory and Computation". *Biophys. J.*, **2016**, *110* (12), 2618-2629. doi: 10.1016/j.bpj.2016.04.028

Dong, S. S.; Abrol, R; Goddard, W. A. "The Predicted Ensemble of Low Energy Conformations of Human Somatostatin Receptor Subtype 5 and the Binding of Antagonists". *ChemMedChem*, **2015**, *10* (4), 650–661. doi: 10.1002/cmdc.201500023

3.1 Introduction

Somatostatins (SSTs) are regulatory peptides involved in inhibition of a number of endocrine and exocrine secretion functions through somatostatin receptors, which are G protein-coupled receptors (GPCRs). (*1*, *2*) They regulate the secretion of factors such as insulin and growth hormone. All five somatostatin receptor subtypes (SSTRs) are able to down regulate cell proliferation, but they vary in a number of other functions such as the regulation of ion channels.[1a] Thus, their effects on cell proliferation and apoptosis are of interest for developing nonpeptidic agonists to enhance tumor growth suppression.(*3*) The subtype 5, SSTR5, forms heterodimer with SSTR2 and presents enhanced cell growth inhibition ability.(*4*)

Of the five human somatostatin receptor subtypes (hSSTRs), hSSTR5 is the only one that has different affinities for the two endogenous ligands SST-14 and SST-28, which are cyclic peptides with 14 and 28 residues respectively.(*5*) SSTR5 has a higher affinity for SST-28, which is shown to suppress glucagon-like peptide-1 (GLP-1) secretion more effectively than SST-14.(*6*, *7*) Therefore, hSSTR5 antagonists are potentially useful in diabetes treatment. Indeed, it has been shown that certain nonpeptidic antagonists are able to improve glucose tolerance in rodent models of type 2 diabetes.(*8*)

Two peptide-based somatostatin mimics, octreotide and vapreotide, have been commercialized to treat various diseases or conditions such as metastatic carcinoid tumors and esophageal variceal bleeding. In recent years, a number of small-molecule agonists and antagonists have also been published.(*9–12*) However, to the best of our knowledge, none have passed clinical trials. We expect that designing higher potency and higher selectivity small-molecule ligands would be useful to minimize off-target side effects.

For rational design of improved hSSTR5 ligands, it is essential to know the molecular details of the receptor binding pockets for both the active and inactive states. Since no experimental structures are available for any of the SSTRs, to predict the 3D structures of hSSTR5, we used the computational method ActiveGEnSeMBLE we developed based on our previous method GEnSeMBLE, as described in Chapter 2.

Apart from the biological functions described above, a recent study has identified hSSTR5 as one of the most valuable templates for homology modeling of non-orphan and non-olfactory class A GPCRs which represent the majority of the class A GPCRs (highest sequence identity sum, and a percentage of sequences for accurate models value of 31%).(*13*) This has increased the significance of obtaining the hSSTR5 structure.

In this chapter, we focus on discussing the active-state hSSTR5 structures we predicted using ActiveGEnSeMBLE, their binding with agonists and the G protein, and their energy profiles of activation. Results on inactive-state hSSTR5 structures and their binding with antagonists are in Chapter 4.

3.2 Methods

Structure prediction and ligand docking

We applied the ActiveGEnSeMBLE method to the hSSTR5 receptor, for which there are no experimental structures. The exact procedures of Steps 1) - 3) follow the ActiveGEnSeMBLE protocol in Chapter 2 and are described in detail in our previous publication (*14*, *15*) as well as Appendix A and B. The agonists (L-817,818 and F21) were docked to each of the five predicted hSSTR5 structures, InactiveConf1,2,3 and ActiveConf1,2, as described in Appendix A.

Molecular dynamics simulation and analysis

We carried out MD simulations of the complexes

- agonist+ActiveConf2+Gα_i,
- agonist+ActiveConf2,

- apo-ActiveConf2+Gα_{i},
- apo-ActiveConf2,
- agonist+InactiveConf2,
- apo-InactiveConf2, Gα_{i} alone, and
- the agonist alone.

Similar to the MD simulation of hβ_2AR, these systems were chosen so that we can use interaction energy analysis of the MD trajectories to extract meaningful information for studying the GPCR activation mechanism. The starting agonist-GPCR complexes were the lowest-energy L-817,818-bound inactive- and active-state structures from docking. For apo-ActiveConf2+Gα_{i}, its starting structure was from removing the agonist from the last frame of 51 ns MD simulation of agonist+ActiveConf2+Gα_{i}.

The detailed procedure of building the starting system of the MD simulation is in Appendix A. The MD procedure and energy analysis were performed as described above. The "inactive", "intermediate", or "active" states in the energy analysis are defined as $R_{36} < 8$ Å, 8 Å $< R_{36} < 11$ Å, and $R_{36} > 11$ Å, respectively.

3.3 Results and Discussion

Structure Prediction

We have applied ActiveGEnSeMBLE to a GPCR without a known experimental structure, hSSTR5. The workflow is shown in Figure 2.1, with the starting crystal structure template being mOPRM. When generating active conformations from hybrid templates, the optional BiHelix step in the flow chart in Figure 2.1 was carried out. The detailed methods used for the inactive-state structural prediction have been discussed in our previous publication (*14*) as well as Chapter 4. When selecting the potentially active structure from the coarse sampling, 10.04 Å is used as a criterion because $R_{36}^{(\mathrm{it})}$ is 6.04 Å. Similar to hβ_2AR and hM2, Table 3.1 shows the fine conformational sampling is able to lower the energies starting from the results of coarse conformational sampling. In addition, all Structures 3.x have R_{36} within 1.14 Å of each other, and all Structures 4.x have R_{36} within 1.58 Å of each other. In other words, the lowest-energy structures from fine samplings have similar R_{36} to the structures the fine samplings start with. This means we have successfully explored lower energy structures for the inactive- and active-like states in their respective energy wells.

Structure / Method Identifier	Method Name	Δθ (°)		Δφ (°)							Δη (°)							E_{CNti} (kcal mol^{-1})	R_{36} (Å)
		H1 – H5, H7	H6	H1	H2	H3	H4	H5	H6	H7	H1	H2	H3	H4	H5	H6	H7		
Homology w/ mOPRM inactive-state crystal structure	-	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	N/A	6.04
3.1	AfromI_Coarse	0	0	0	0	0	0	0	-90	0	0	0	0	0	30	270	0	-265.8	11.34
3.2 (ActiveConf1)	AfromI_Fine	0	0	0	0	0	0	0	-60	-15	0	0	0	0	30	270	0	-313.2	10.71
3.3 (ActiveConf2)	AfromIH_Fine	0	-15	0	0	-15	0	-30	-60	0	0	0	0	0	30	240	0	-306.6	10.20
4.1	IfromI_Coarse	0	0	0	0	0	90	0	0	0	0	0	0	-30	30	0	0	-303.3	6.01
4.2.1 (InactiveConf1)	IfromI_Fine	0	0	0	0	0	105	0	-15	0	0	0	0	-30	0	0	0	-350.7	7.18
4.2.2 (InactiveConf2)	IfromI_Fine	0	0	0	0	-15	60	30	15	0	0	0	0	0	60	0	0	-345.7	7.59
4.2.3 (InactiveConf3)	IfromI_Fine	0	0	0	0	-15	60	-30	15	0	0	0	0	0	30	0	0	-328.9	7.56

Table 3.1: Summary of starting structures (Methods 3.1, 4.1) and best resulting structures (Method 3.2, 3.3, 4.2.x) of sampling hSSTR5 conformations using different strategies. Method 3.x are active state samplings, and Method 4.x are inactive state samplings. The original homology template is from the mOPRM inactive-state crystal structure. The methods used are outlined in Figure 2.1 and detailed in our previous publication.(*14*)

In addition to the methods described above, we have also tested ActiveGEnSeMBLE using hβ_2AR active-state x-ray structure as the template of all the hSSTR5 TMs. The lowest-energy structure has E_{CNti} of -224.7 kcal mol^{-1}, which is less stable than any of Structures 3.x (-265.8 to -306.6 kcal mol^{-1}), which used either the mOPRM inactive-state template or the hybrid template. As a result, we did not use this structure for the later steps. This suggests that the crystal structure of the hβ_2AR active-state conformation may not be the best template for building an active-state homology model for many GPCRs. As there are more inactive-state templates than active, if a target receptor has the best sequence homology to a receptor with only an inactive-state template, one should build the starting structure using that template. Then, one can use ActiveGEnSeMBLE can be used on that structure to predict the active-state and inactive-state conformations of the target receptor.

The energy profile of hSSTR5 plotted against R_{36} is illustrated in Figure 3.1. Unlike hβ_2AR and hM2, hSSTR5 seems to have a flatter energy surface. This may be a feature of hSSTR5, but can also be affected by the homology template used. As the energies of ActiveConf1 and ActiveConf2 are comparable and we would like to have a diverse set of candidate structures, we chose both of these structures as active-state candidates. To see whether our predicted structures are reasonable, ligand-binding studies and MD simulation have been carried out.

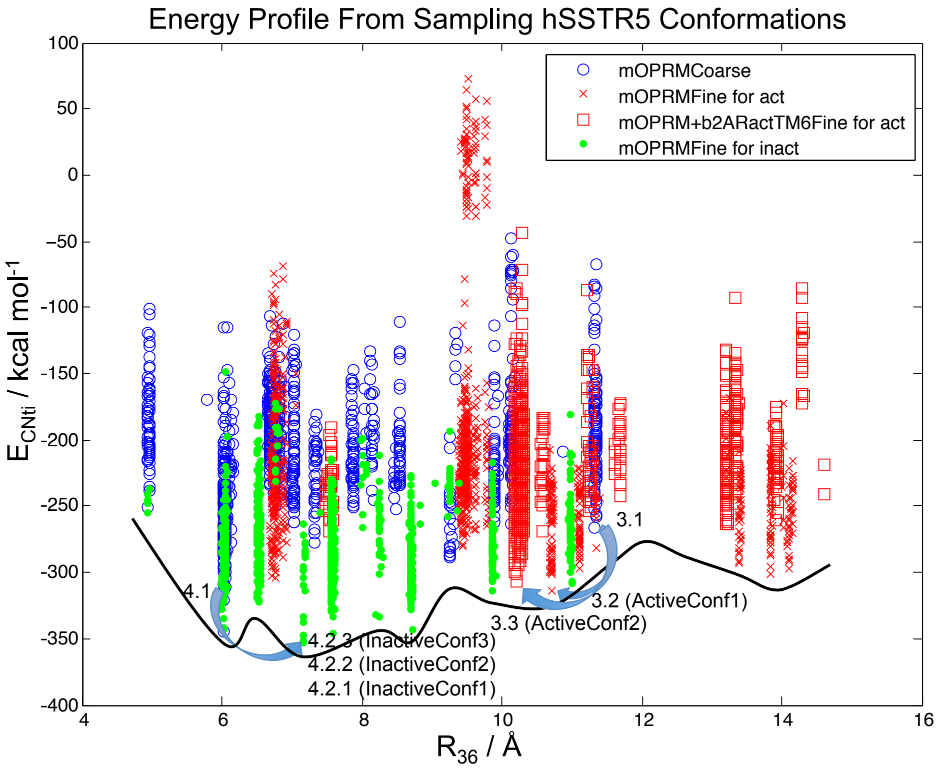


Figure 3.1: Energy profile from hSSTR5 samplings. The black curve is an illustration showing how our sampling results can qualitatively be translated into a potential energy curve using R_{36} as the x-axis and does not quantitatively represent any real data. Results of the coarse sampling are in blue circles. Starting from Structure 3.1, results of methods that generated Structures 3.2 (ActiveConf1) and 3.3 (ActiveConf2) are shown in red crosses and red squares respectively. Starting from the Structure 4.1, results of the fine sampling that generated Structures 4.2.1, 4.2.2, and 4.2.3 (InactiveConf1,2,3) are in green dots. Every blue arrow points from a starting structure towards the optimal structure from the corresponding sampling method.

Ligand Binding Studies

In order to verify the hSSTR5 structures predicted, we have docked five antagonists (*11*) (Figure 3.2) and two agonists (*10*) (Figure 3.3) to all the five candidate structures, InactiveConf1, InactiveConf2, InactiveConf3, ActiveConf1, and ActiveConf2. The detailed binding studies of the antagonists docking are in Dong et al.,(*14*) partly reproduced in Chapter 4, in which the predicted binding energy of the antagonist series is consistent with experimental binding constants.

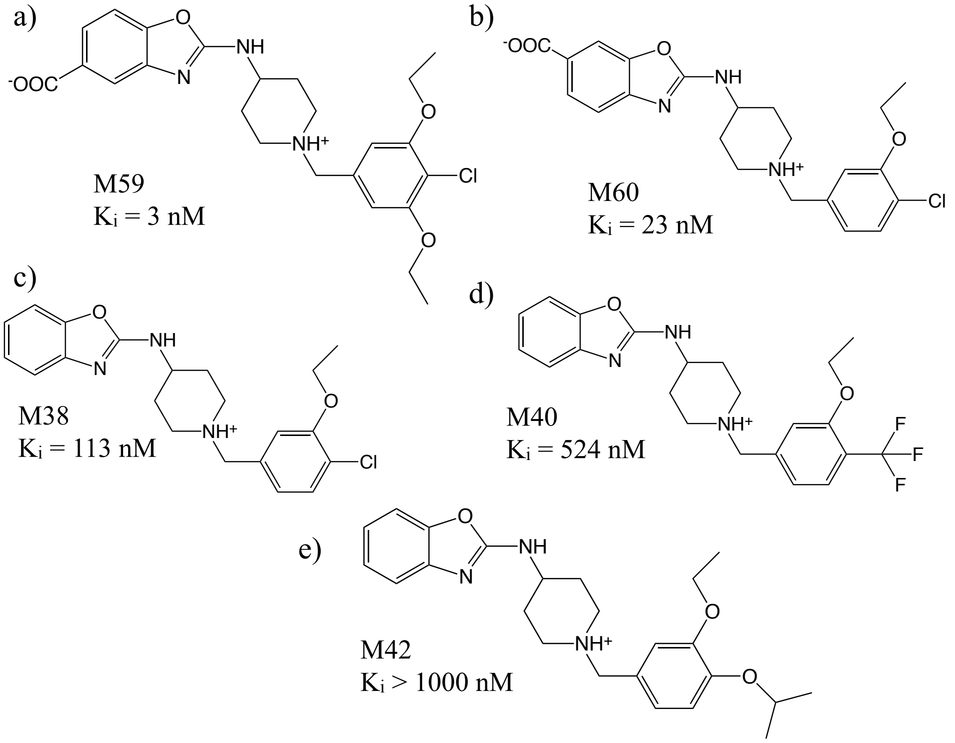


Figure 3.2: Antagonists series docked to hSSTR5 predicted structures.

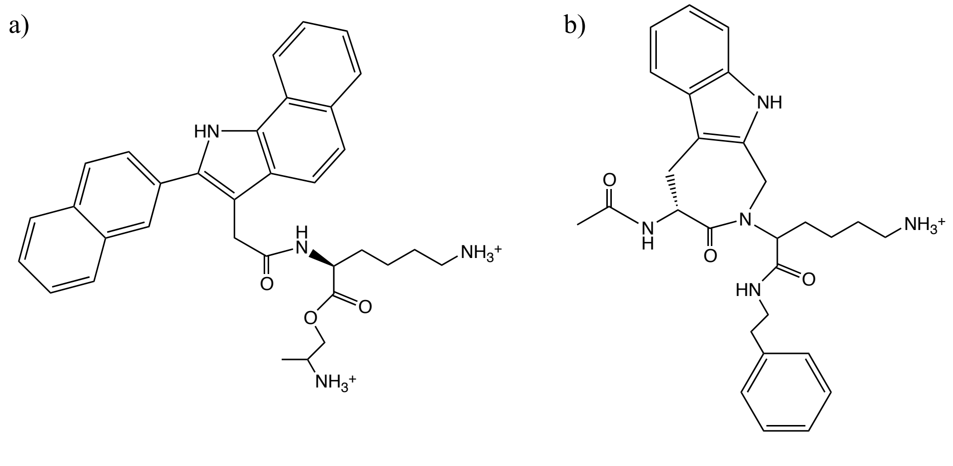


Figure 3.3: Structures of the docked agonists a) L-817,818 (K_i = 0.4 nM) and b) F21 (IC_{50} = 0.56 nM).

The agonists L-817,818 and F21 are selected because they have very high affinities (subnanomolar binding constants) with hSSTR5, and they display similar but different structural features. We did not select a series of agonists with the same scaffold for structure-activity relationship (SAR) analysis due to the lack of published experimental SAR data. The selected agonists' structures are shown in Figure 3.3. They are both peptide mimics derived from the endogenous ligand, somatostatin. They both have polycyclic aromatic groups, and have a group mimicking lysine. However, L-817,818 has one more positively charged amine group than F21.

For each agonist, docking to the five predicted structures finds that the lowest-energy poses with the inactive state and the active state are only slightly different (Table 3.2, 3.3 and 3.4). This suggests an easy pathway for the agonist-bound receptor to interconvert between the inactive state and the active states, consistent with experiments on hβ_2AR.(*16*) Binding energies from docking have also shown that, in the absence of a G protein, agonist stabilizes its inactive conformation. This is again consistent with experiments on hβ_2AR (*16*) and is further supported by interaction energy analysis of MD trajectories described in the next section.

Pose Name	Ligand	Protein Structure	SnapBE (kcal mol^{-1})
L_i2	L-817,818	InactiveConf2	-118.32
L_a2	L-817,818	ActiveConf2	-106.05
F_i2	F21	InactiveConf2	-99.41
F_a2	F21	ActiveConf2	-92.39

Table 3.2: Best pose of L-817,818 and F21 docked to active structures determined by lowest snap binding energy (SnapBE).

We also find that for both agonists, L-817,818 and F21, the best active-state pose is with ActiveConf2 (Table 3.2). Since ActiveConf2 was generated using structure prediction Method 3.3, this again suggests that TM6 shape is an important factor in GPCR activation and a more active-like TM6 shape makes computational prediction of active-state conformations more likely to succeed. Contrary to the antagonist M59, the highest binding affinity antagonist we docked, which favors InactiveConf1, the best inactive-state poses for both agonists are with InactiveConf2, which has slightly higher energy (by 5 kcal mol^{-1}) and greater R_{36} (by 0.4 Å) than InactiveConf1. This implies that agonists may stabilize a slightly "more-active" inactive state than the antagonist, and is a direct demonstration of the ability of ActiveGEnSeMBLE to predict multiple states that is crucial in elucidating GPCR activation mechanisms.

There are no published mutation studies that have probed the interaction between hSSTR5 and small molecule ligands. In addition, there is only limited mutation data of the SSTRs' binding with the endogenous S-14 and S-28, especially in the TM

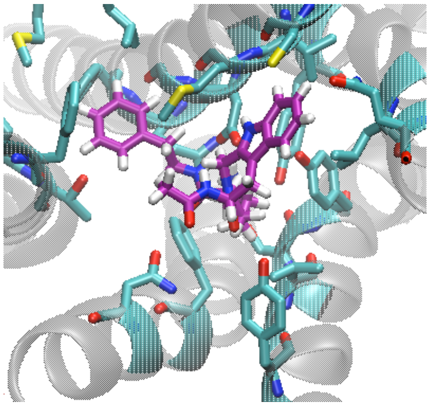
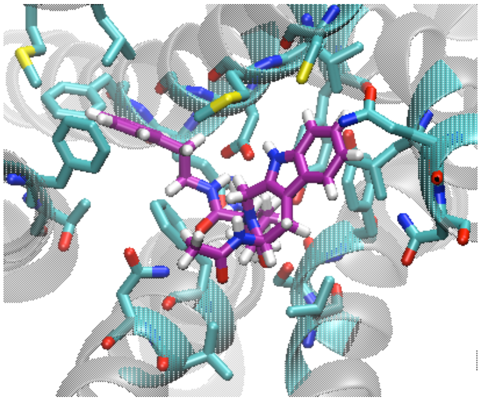
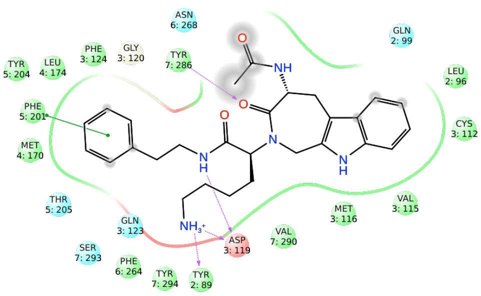
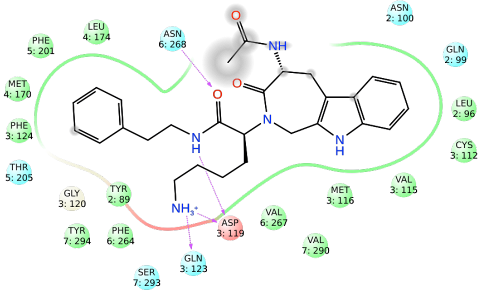
F_i2	F_a2
a) 	b) 
c)  ASN 6: 268 GLN 2: 99 TYR 5: 204 LEU 4: 174 PHE 3: 124 GLY 3: 120 TYR 7: 286 LEU 2: 96 CYS 3: 112 PHE 5: 201 MET 4: 170 THR 5: 205 GLN 3: 123 SER 7: 293 PHE 6: 264 TYR 7: 294 TYR 2: 89 ASP 3: 119 VAL 7: 290 MET 3: 116 VAL 3: 115 O NH O N N H N H NH_3^+	d)  ASN 2: 100 LEU 4: 174 ASN 6: 268 PHE 5: 201 GLN 2: 99 MET 4: 170 LEU 2: 96 PHE 3: 124 CYS 3: 112 THR 5: 205 VAL 3: 115 GLY 3: 120 TYR 2: 89 MET 3: 116 VAL 6: 267 TYR 7: 294 PHE 6: 264 ASP 3: 119 VAL 7: 290 GLN 3: 123 SER 7: 293 O NH O N N H N H NH_3^+

Table 3.3: a) 3D visualization of the docking pose F_i2, the lowest-energy complex with agonist F21. b) 3D visualization of the docking pose F_a2. The ligand is shown in purple. c) Ligand interaction diagram (LID) of the pose F_i2. d) LID of the pose F_a2. LIDs are generated by Maestro9.3.(26) The cutoff distance for the residues shown is 4.0 Å. Hydrophobic interaction: green; polar interaction: blue; hydrogen bonds (cutoff distance 2.5 Å): purple arrows; π-π stacking: straight green lines.

regions. Here we have attempted to compare our binding analysis with the available experimental mutation data, and will suggest mutations that can be experimentally tested to probe the binding of agonists L-817,818 and F21.

As shown in Table 3.3, the positively charged amine group in the ligand F21 forms a salt bridge with D119$^{3.32}$ on TM3, which is common among the closely related opioid receptors. We can also find that ActiveConf2 forms a hydrophobic pocket that is in contact with the majority non-polar groups in F21. Note that the hydrophobic pocket includes F264$^{6.51}$ on TM6. It is known that the mutation F6.52Y in rat SSTR5 (rSSTR5) can increase the binding affinity of S-14 to rSSTR5 by 20-fold,(*17*) and this means F265$^{6.52}$ is involved in the SSTR5-agonist binding. Since F265$^{6.52}$ and

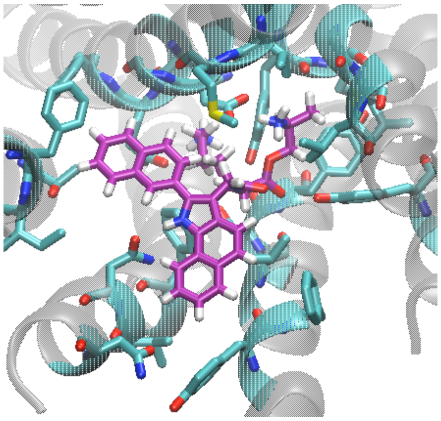
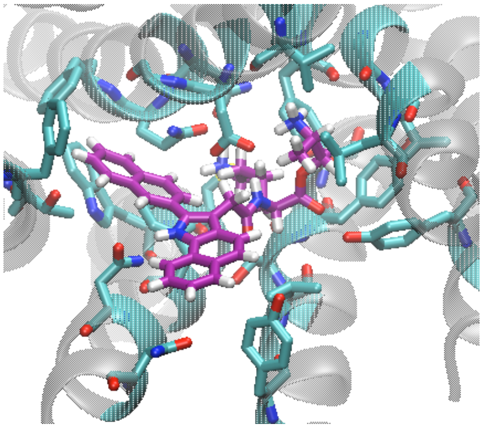
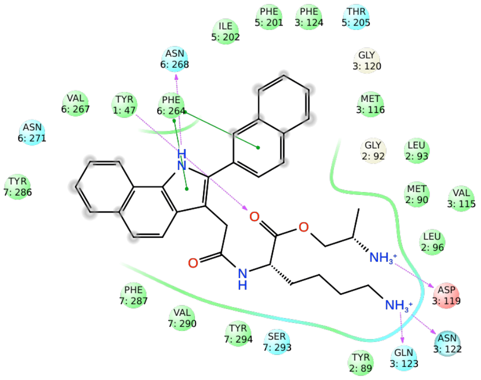
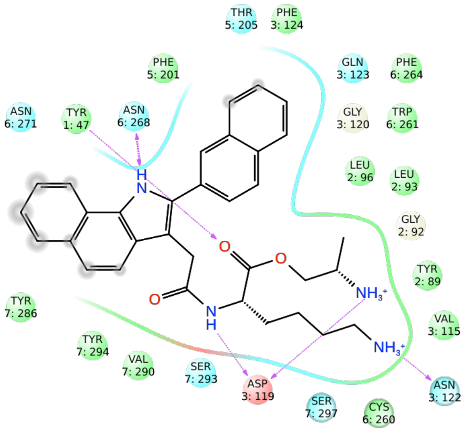
L_i2	L_a2
a) 	b) 
c)  PHE 5: 201 PHE 3: 124 THR 5: 205 ILE 5: 202 ASN 6: 268 GLY 3: 120 VAL 6: 267 TYR 1: 47 PHE 6: 264 MET 3: 116 ASN 6: 271 GLY 2: 92 LEU 2: 93 TYR 7: 286 H N MET 2: 90 VAL 3: 115 O O LEU 2: 96 NH3+ O N H ASP 3: 119 PHE 7: 287 NH3+ VAL 7: 290 TYR 7: 294 SER 7: 293 ASN 3: 122 TYR 2: 89 GLN 3: 123	d)  THR 5: 205 PHE 3: 124 PHE 5: 201 GLN 3: 123 PHE 6: 264 ASN 6: 271 TYR 1: 47 ASN 6: 268 GLY 3: 120 TRP 6: 261 H N LEU 2: 96 LEU 2: 93 GLY 2: 92 O O TYR 2: 89 NH3+ O N H TYR 7: 286 VAL 3: 115 NH3+ TYR 7: 294 VAL 7: 290 SER 7: 293 ASP 3: 119 SER 7: 297 CYS 6: 260 ASN 3: 122

Table 3.4: a) 3D visualization of the docking pose L_i2, the lowest-energy complex with agonist L-817,818. b) 3D visualization of the docking pose L_a2. The ligand is shown in purple. c) Ligand interaction diagram (LID) of the pose L_i2. d) LID of the pose L_a2. LIDs are generated by Maestro9.3.(26) The cutoff distance for the residues shown is 4.0 Å. Hydrophobic interaction: green; polar interaction: blue; hydrogen bonds (cutoff distance 2.5 Å): purple arrows; π-π stacking: straight green lines.

F264$^{6.51}$ are neighboring residues and the ligand F21 is much smaller than S-14, we can hypothesize that F264$^{6.51}$ is important in F21 binding with the activated hSSTR5. Notice that residue 6.52 is tyrosine for all other SSTRs, but F6.51 is conserved in all SSTRs. As a result, we suggest F6.51 is responsible for the affinity of SSTRs to the ligand F21 but not the selectivity.

L-817,818 has different binding modes because of its two positively charged amine groups. The 3D visualization and ligand interaction diagrams (LIDs) of the best pose with the inactive state (L_i2) and the active state (L_a2) are shown in Table 3.4.

Similar to ligand F21, the salt bridge between one amine group with D119$^{3.32}$ is present, and there is π-π interaction between the aromatic groups in L-817,818 and F264$^{6.51}$. The other lysine-like amine group in L-817,818 forms a hydrogen bond with Q123$^{3.36}$ and has the potential to have electrostatic interaction with D86$^{2.50}$. In the MD simulation discussed in more detail in the next section, we find that this amine group is indeed forming a salt bridge with D86$^{2.50}$ in the active-state simulation. Thus, we suggest D86$^{2.50}$ is involved in hSSTR5 activation by L-817,818. Also, hSSTR5 mutant F264Y is predicted to have higher affinity for L-817,818 due to the potential of an additional hydrogen bond between the ester oxygen atom in L-817,818 and Y264.

MD Simulation and analysis

To further investigate the activation mechanism of hSSTR5, we have carried out molecular dynamics simulations on the L-817,818-bound and apo predicted structures. We have considered six cases: 1) agonist+ActiveConf2+Gα_i, 2) agonist+ActiveConf2, 3) apo-ActiveConf2+Gα_i, 4) apo-ActiveConf2, 5) agonist+InactiveConf2, and 6) apo-InactiveConf2. The binding site of agonist+ActiveConf2 is from L_a2, and that of agonist+InactiveConf2 is from L_i2. As expected, similar to hβ_2AR, the inactive state remains inactive during the MD simulation (Figure 3.4), and the coupling of Gα_i is able to keep both the agonist-bound and apo-GPCR active during the MD simulation (Figure 3.6). Interestingly, contrary to hβ_2AR for which both the agonist-bound and apo-GPCR lose their activity during the MD simulation, for hSSTR5, starting from active-state conformations without Gα_i, the apo-GPCR is able to keep hSSTR5 with a large R_{36} capable of coupling to Gα_i, while the agonist-bound GPCR goes back to its "intermediate state" within 5 ns and goes back to its inactive state in 28 ns (Figure 3.5). This suggests the constitutive activity of hSSTR5 plays an important role in its activation mechanism. The G protein is more likely to couple with hSSTR5 before the agonist binds.

For each case, we have analyzed the potential energy of the agonist-bound and apo-GPCR along the MD trajectory and the resulting energy landscape (Figure 3.7) is consistent with the illustration derived from experimental findings of hβ_2AR.(*16*, *18*) To be more specific, although the active state of a GPCR has higher energy than its inactive state, binding of the G protein lowers the energy of the active state of the GPCR, and $E_{\mathrm{R+interaction}}$ of the apo-GPCR+Gα_i complex is not as low as the agonist-bound complex. This activation picture is another example besides hβ_2AR that has quantitatively confirmed the belief that the G protein facilitates GPCR

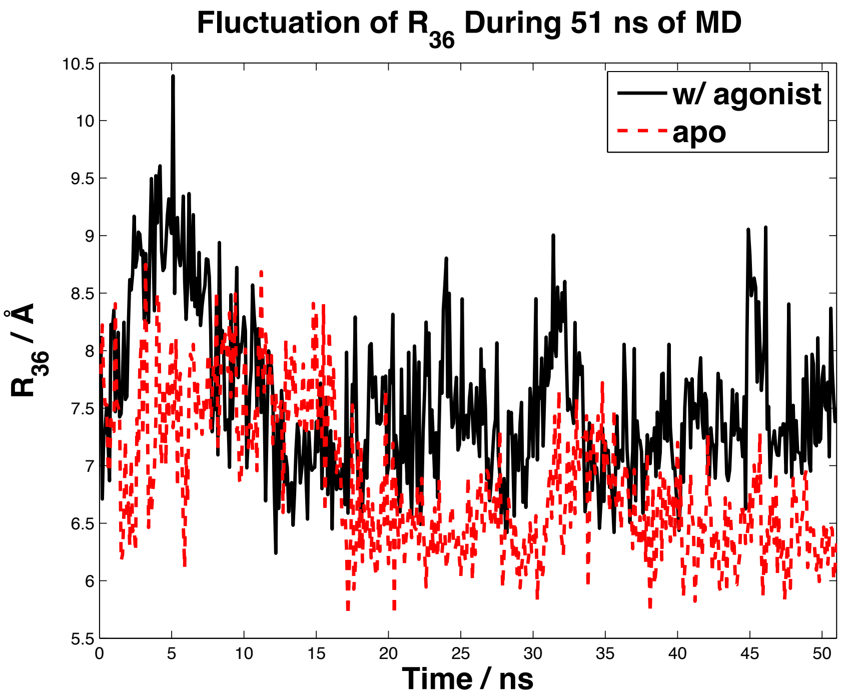


Figure 3.4: Fluctuation of R_{36} of hSSTR5 during 51 ns molecular dynamics simulation starting from the predicted inactive-state structure.

activation by stabilizing both the agonist-bound and apo-GPCR. The former leads to agonist-induced activity and the latter leads to constitutive activity.

Comparing the energy landscape of hβ_2AR in Figure 2.8 and that of hSSTR5 in Figure 3.7, we find that their intermediate states have different features. The intermediate states of the BI-167107-bound hβ_2AR have lower energy than the inactive states, while those of the L-817,818-bound hSSTR5 have higher energy than the inactive states. This can be reasonable because the energy landscape of even the same receptor can be altered by different agonists. For example, another hβ_2AR agonist isoproterenol makes hβ_2AR have higher energy intermediate states than inactive states (Figure 2.9).(*16*) In addition, the definition of "intermediate state" here include many different states that have R_{36} between the inactive state and the states that can couple to the G protein, and the average energy of these intermediate states depends on the distribution of these states in the MD trajectory.

Furthermore, there have been experimental structures of several agonist-bound GPCRs that present features from inactive to partially-active of various degrees without the presence of the G protein, suggesting a role of the agonist in the initial steps of GPCR activation that varies with different GPCRs. For example,

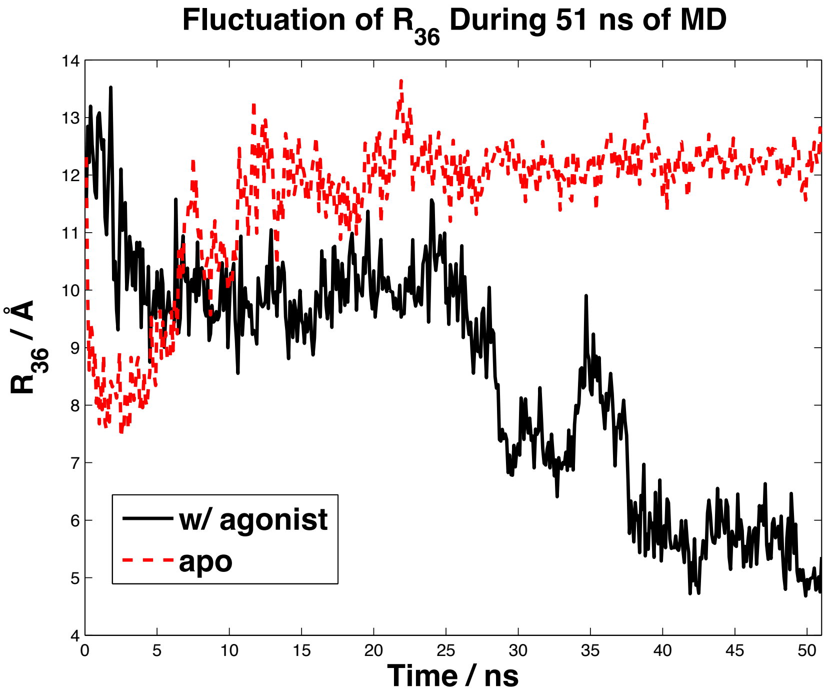


Figure 3.5: Fluctuation of R_{36} of hSSTR5 during 51 ns molecular dynamics simulation starting from the predicted active-state structure without the presence of the Gα protein.

- In turkey β_1 adrenergic receptor, the agonist-bound structure is nearly identical as the antagonist-bound inactive-state structure except for a 1 Å contraction of the binding pocket.(*19*)

- In contrast, for human adenosine A_{2A} receptor, the agonist-bound structure has all active-like features except that the IC end of TM6 is only partially opened for coupling with the G protein (Table 2.1).(*20*) The agonist-bound human serotonin 1B (5-HT_{1B}) receptor, human serotonin 2B (5-HT_{2B}) receptor and rat neurotensin 1 receptor (NTSR1) do not have an antagonist-bound inactive-state structure of the same receptor to compare to, but they show partially-active features.

- Bound with the same agonist, the 5-HT_{2B} receptor has a less-active TM6 and a more-active TM7 than the 5-HT_{1B} receptor, while they both have an outward shift of TM6 IC end comparing to inactive-state structures of other aminergic receptors.(*21*, *22*)

- NTSR1 also shows active-like features found in bRho and hβ_2AR, with a

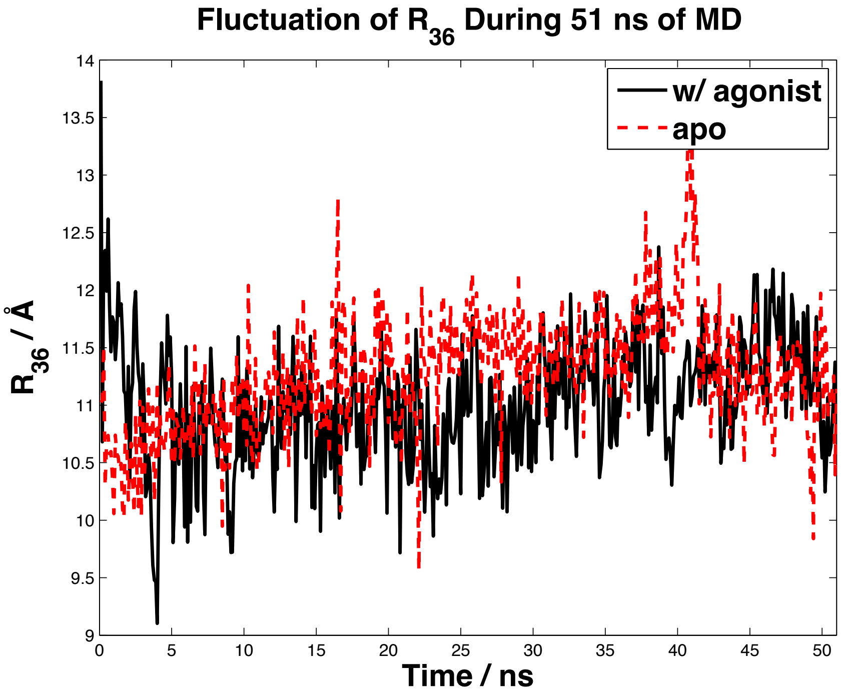


Figure 3.6: Fluctuation of R_{36} of hSSTR5 during 51 ns molecular dynamics simulation starting from the predicted active-state structure in complex with the Gα protein.

> TM6 position similar to that of the active-state bRho but not to the extent of active-state hβ_2AR.(*23*)

Therefore, we conclude that different GPCRs may adopt different activation pathways with different agonists in terms of the energy ordering of different states dictated by respective energy landscapes of activation. It is reasonable that our agonist-bound inactive states of hSSTR5 have the lowest energy among its various states displaying different degrees of activation. Thus, we consider that the computation of energy landscape of GPCRs using the method detailed in this article is valuable since it may allow for the activation mechanisms of a broader variety of GPCRs to be mapped out efficiently.

Similar to the discussion of hβ_2AR, hSSTR5 energy profile plotted with E_{R} (Figure 3.8) also shows that the G protein destabilizes the receptor, except that in the case of hSSTR5 this is regardless of whether the agonist is bound. This may be reasonable as the hSSTR5 agonist L-817,818 destabilizes hSSTR5 while the hβ_2AR agonist BI-167107 stabilizes hβ_2AR. As $E_{\mathrm{R+interaction}}$ shows a lower energy for the agonist-bound hSSTR5, we can conclude that the stabilization of hSSTR5 comes

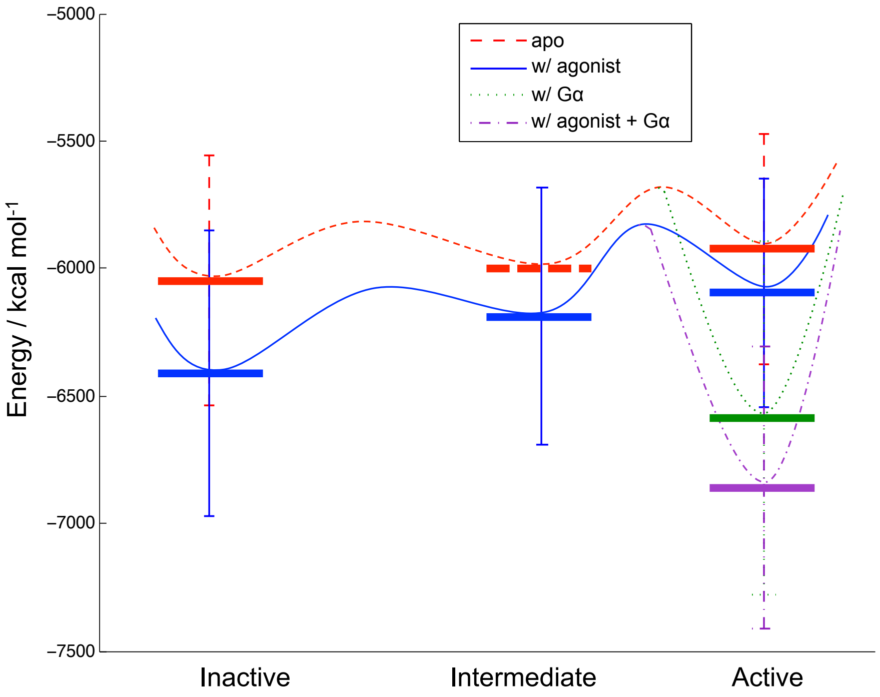


Figure 3.7: Energy landscape of hSSTR5 activation from MD simulation. The solid horizontal bars are $E_{R+interaction}$ calculated as described in Section 2.3. The dashed horizontal bar is fictitious. The curved lines are fictitious energy surface, with the barriers being qualitative and minima defined by the corresponding $E_{R+interaction}$ values. Inactive, intermediate, and active states in the figure are defined as $R_{36} < 8$ Å, 8 Å $< R_{36} <$ 11 Å, and $R_{36} >$ 11 Å, respectively. Energy landscape of hSSTR5 activation from MD simulation.

from the interaction between the agonist and the receptor, in addition to the interaction between Gα_i and the receptor. As mentioned in the previous chapter, unlike hβ_2AR, however, E_{Total} (Figure 3.9) is qualitatively different from $E_{\text{R+interaction}}$ for hSSTR5.

The stabilizing effect of the G protein on the active state of the GPCR can indeed be explained by the specific interactions between the Gα subunit and the GPCR. During the MD simulation of agonist+ActiveConf2+Gα_i, salt-bridge and hydrogen-bond networks are able to form between the C-terminal helix of Gα_i and ActiveConf2 as shown in Figure 3.10. In particular, the formation of the salt-bridge network involving Gα_i's D261$^{G.h3s5.2}$ and D350$^{G.H5.22}$, and hSSTR5's K72 on IC loop 1 (ICL1), R151 on ICL2, and the formation of the hydrogen bond between R137$^{3.50}$ and Gα_i's C351$^{G.H5.23}$, replace the inactive state's R151-D136$^{3.49}$-R137$^{3.50}$-T247$^{6.34}$

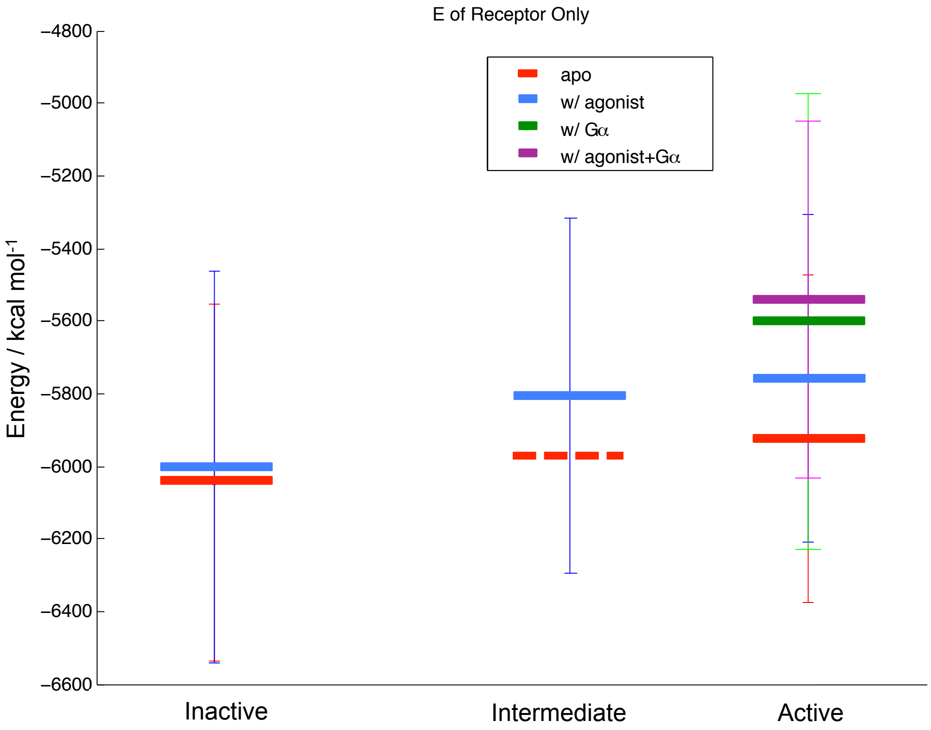


Figure 3.8: Energy profile of hSSTR5 during activation. The horizontal bars are E_R obtained according to Section 2.3. The dashed horizontal bar is fictitious. “Inactive”, “intermediate” and “active” states in the figure are defined as $R_{36} < 8$ Å, 8 Å $< R_{36} <$ 11 Å, and $R_{36} >$ 11 Å, respectively.

network. Additional salt-bridge network is formed between the carboxylate group on the Gα_{i} C-terminal residue F354$^{\mathrm{G.H5.26}}$ and hSSTR5’s K245$^{6.32}$ on TM6 and R239 on ICL3. In addition, a hydrogen bond is formed between R248$^{6.35}$ and G352$^{\mathrm{G.H5.24}}$, and a weaker hydrogen bond is formed between W150 on ICL2 and N347$^{\mathrm{G.H5.19}}$. Furthermore, the highly conserved L348$^{\mathrm{G.H5.20}}$ and L353G.H5.25 are in a hydrophobic pocket that consists of V141$^{3.54}$, I224$^{5.61}$, V246$^{6.33}$, M249$^{6.36}$, and V228 on ICL3. Since experimental mutagenesis studies have shown that the mutations L348A$^{\mathrm{G.H5.20}}$, L353A$^{\mathrm{G.H5.25}}$, and G352A$^{\mathrm{G.H5.24}}$ severely hindered the coupling between Gα_{i} and light-activated bRho (bRho*),(*24*) which has the conserved residues V139$^{3.54}$, L226$^{5.61}$, V250$^{6.33}$, M253$^{6.36}$, and R252$^{6.35}$, we may conclude that the corresponding interactions we found between hSSTR5 and Gα_{i} are consistent with experiments. The same experimental paper showed that N347A$^{\mathrm{G.H5.19}}$ did not have a significant effect on the coupling between Gα_{i} and bRho*, which is consistent with the weak interaction we found between W150 and N347$^{\mathrm{G.H5.19}}$. Interestingly, the mutation D350A$^{\mathrm{G.H5.22}}$ seems to stabilize the bRho*-Gα_{i} complex.

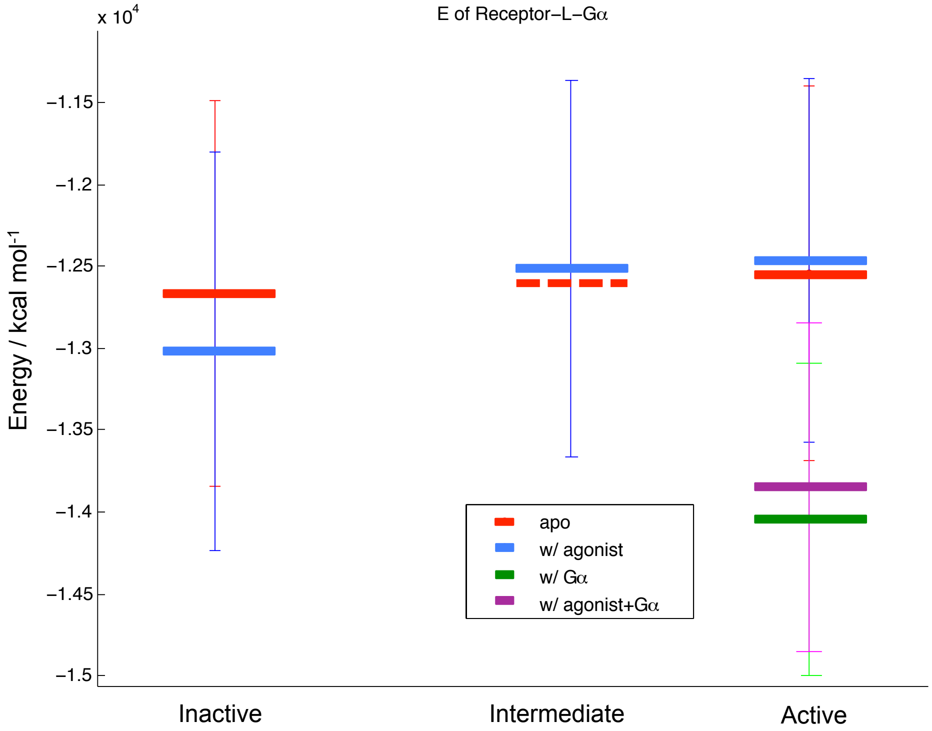


Figure 3.9: Energy profile of the system during hSSTR5 activation. The horizontal bars are E_{Total} calculated according to Section 2.3. The dashed horizontal bar is fictitious. "Inactive", "intermediate" and "active" states in the figure are defined as $R_{36} < 8$ Å, 8 Å $< R_{36} < 11$ Å, and $R_{36} > 11$ Å, respectively.

This might be a property specific to bRho*-Gα_{i} arising from the differences in helix packing and ICL sequences between bRho and hSSTR5.

After the agonist is removed, K72 on ICL1 breaks away from D350$^{\mathrm{G.H5.22}}$ and D261$^{\mathrm{G.h3s5.2}}$ and forms a hydrogen bond with the backbone oxygen atom of D350$^{\mathrm{G.H5.22}}$. In addition, the hydrogen bond between R137$^{3.50}$ and C351$^{\mathrm{G.H5.23}}$ backbone oxygen atom and the hydrogen bond between R248$^{6.35}$ and G352$^{\mathrm{G.H5.24}}$ backbone oxygen atom become water-mediated. The weakening of the interaction between the apo-GPCR and Gα is consistent with the picture that the agonist stabilizes the binding of the G protein with GPCR.

Looking into the protein-ligand interactions in more detail, we have found a characteristic interaction formed at around 25 ns of the agonist+ActiveConf2+Gα_{i} MD simulation, but absent in the entire agonist+InactiveConf2 MD trajectory: the salt bridge between the lysine-like positively charged amine of L-817,818 and D86$^{2.50}$ on hSSTR5 (Figures 3.11-3.13). The highly conservative residue 2.50 has been

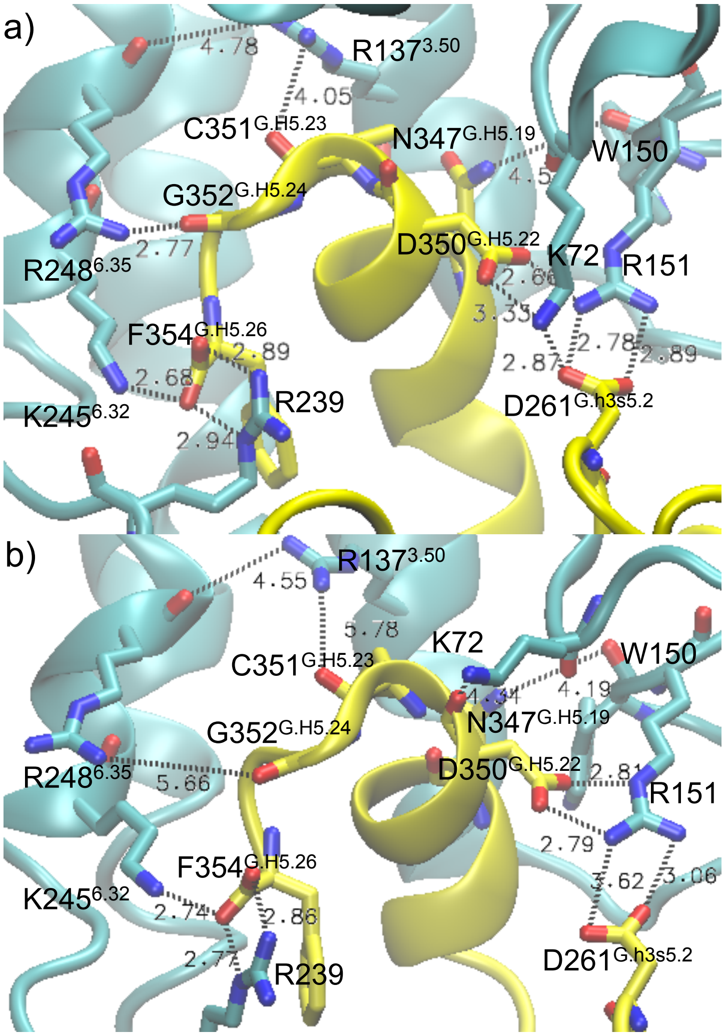
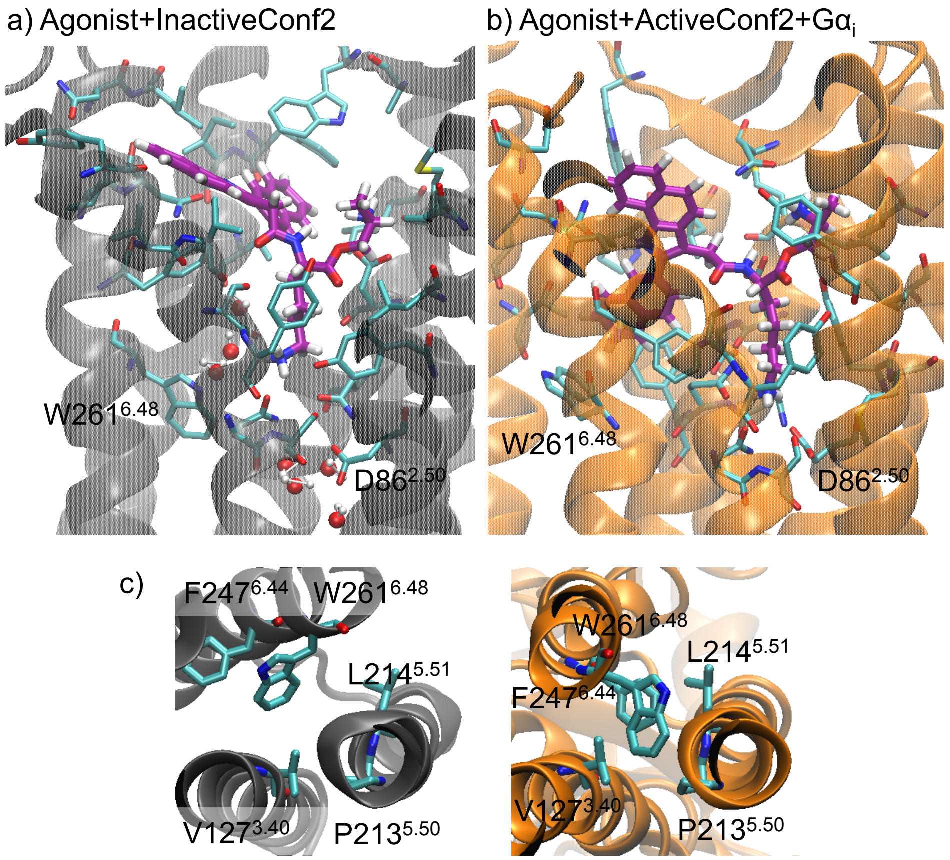


Figure 3.10: Interactions between the Gα_i C-terminus and hSSTR5 after a 51 ns MD simulation of (a) agonist+ActiveConf2+Gα_i and (b) apo-ActiveConf2+Gα_i. Gα_i is shown in yellow and hSSTR5 is shown in cyan. The superscript of Gα residue numbers follows the common Gα numbering system.(*25*)

studied in several GPCRs, but its role is not well understood and varies across different systems.(*26*, *27*) In particular, the mutation D2.50N presents different effects in SSTR1 and SSTR2. Although D2.50 is widely viewed as an allosteric site, our result suggests that the orthosteric site of L-817,818 in hSSTR5 may extend to D2.50. Therefore, our result opens up the possibility of D86$^{2.50}$ being crucial in hSSTR5 activation by engaging in the interaction with the agonist, and further experimental investigation of this residue would be worthwhile. If the significance of D86$^{2.50}$ is verified, designing agonists that are able to form salt bridges with both D119$^{3.32}$ and D86$^{2.50}$ may be a desired path towards drugs targeting hSSTR5.



a) Agonist+InactiveConf2

b) Agonist+ActiveConf2+Gα$_i$

Figure 3.11: a) The salt bridge between L-817,818 (the agonist) and $D86^{2.50}$ of hSSTR5 is not present during the MD simulation of the agonist+InactiveConf2 complex. Instead, water molecules are surrounding $D86^{2.50}$. Water molecules within 10 Å of the side chain of $D86^{2.50}$ are displayed. b) A salt bridge between L-817,818 and $D86^{2.50}$ of hSSTR5 is formed during the MD simulation of the agonist+ActiveConf2+$G\alpha_\mathrm{i}$ complex. In addition, there is π-π stacking between L-817,818 and $W261^{6.48}$ in the agonist+ActiveConf2+$G\alpha_\mathrm{i}$ complex. There is no water molecule within 10 Å of the side chain of $D86^{2.50}$. c) The presence of the transmission switch: $W261^{6.48}$ and F2576.44 are oriented towards P2135.50 in ActiveConf2 (right panel) but not in InactiveConf2 (left panel) partly due to the rotation of TM6. The secondary structure in agonist+InactiveConf2 is shown in grey, and that in agonist+ActiveConf2+$G\alpha_\mathrm{i}$ is shown in orange. Carbons in the residues on hSSTR5 are shown in cyan. The agonist carbon atoms are shown in purple.

3.4 Conclusions

We applied the validated ActiveGEnSeMBLE method to the hSSTR5 receptor, for which there is no available experimental structure. Importantly, we found that a hybrid template consisting of the TM6 from the available active-state crystal structure combined with TM1-5 and TM7 of inactive-state crystal structures from GPCRs

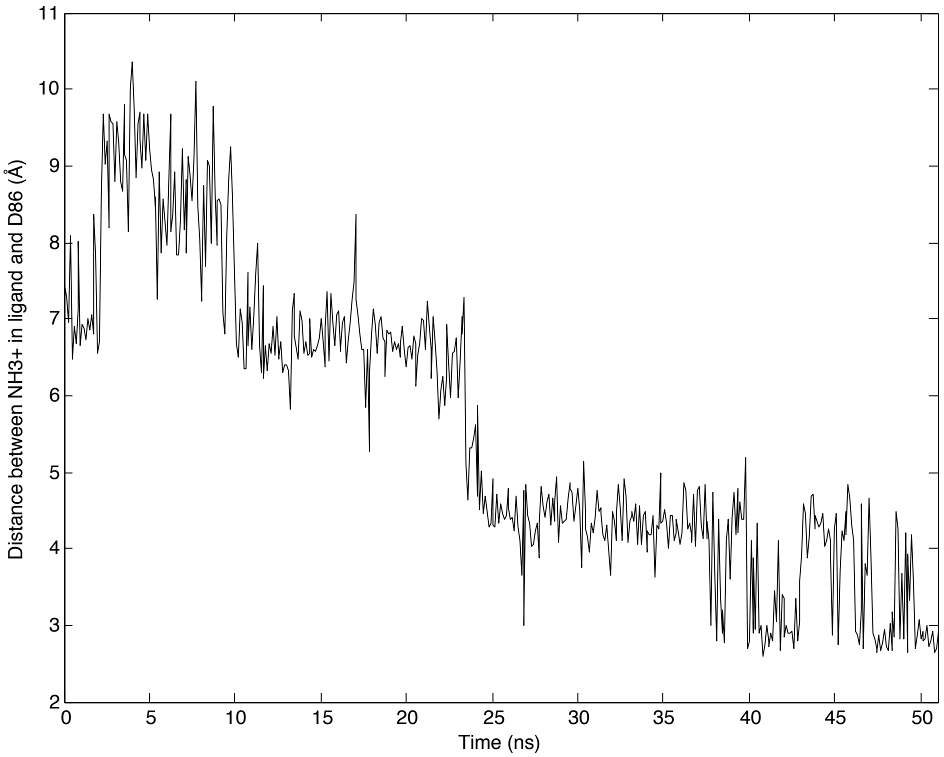


Figure 3.12: Distance between the N atom in an amine group of L-817,818 and a carboxylic acid oxygen atom in D86$^{2.50}$ of hSSTR5 along the trajectory of MD simulation of the agonist+ActiveConf2+Gα_i complex.

with high-sequence identity generated even lower energy active-like structures than a template based purely on the available active-state crystal structures. Thus, it is not necessary to have the full structure for an active GPCR to apply ActiveGEnSeMBLE. In other words, the method's application to the somatostatin receptor hSSTR5 shows that, to predict an active conformation, it is better to start from an inactive-state template based on a close homolog than to start from an active template based on a distant homolog.

Docking of agonists and the subsequent MD simulations identified important residues involved in hSSTR5 activation by the respective agonists. MD simulations of the predicted structures of hSSTR5 with and without the G protein and the agonist generated energy profiles that are consistent with the qualitative energy landscape of hβ_2AR obtained from experiments and also with the quantitative energy landscape of hβ_2AR presented in this study. The differences are compatible with previous findings from agonist-bound experimental structures for various GPCRs in that the agonist promoted the initial steps of GPCR activation to degrees that varied among

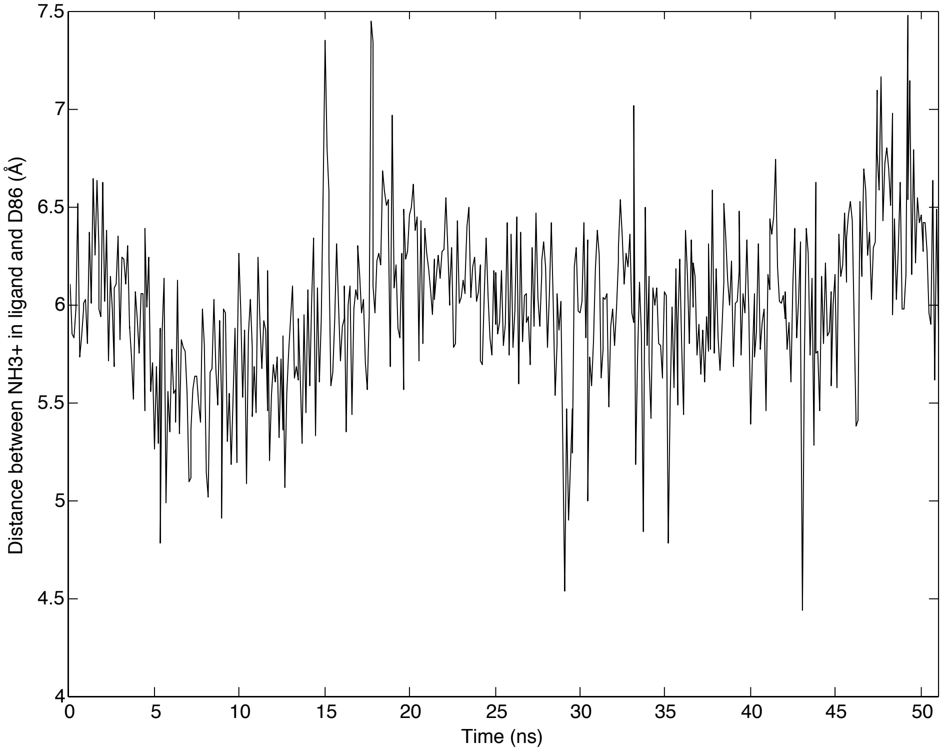


Figure 3.13: Distance between the N atom in an amine group of L-817,818 (the same amine group in Fig. S15) and a carboxylic acid oxygen atom in D86$^{2.50}$ of hSSTR5 along the trajectory of MD simulation of the agonist+InactiveConf2 complex.

different GPCRs. These energy profiles indicate that the G protein helps to stabilize the agonist-bound GPCR. These results confirm that ActiveGEnSeMBLE is effective in predicting the active-state conformations of at least class A GPCRs, and provides a powerful new tool for elucidating the activation mechanisms of GPCRs by identifying the sequence of conformations along the pathway for activation. We hope that this will accelerate the rational design of new, more potent and selective agonists.

References

(*1*) Weckbecker, G., Lewis, I., Albert, R., Schmid, H. A., Hoyer, D., and Bruns, C., (2003). Opportunities in somatostatin research: biological, chemical and therapeutic aspects. *Nature Reviews Drug Discovery 2*, 999–1017.

(*2*) Olias, G., Viollet, C., Kusserow, H., Epelbaum, J., and Meyerhof, W., (2004). Regulation and function of somatostatin receptors. *Journal of neurochemistry 89*, 1057–1091.

(*3*) Albini, A., Florio, T., Giunciuglio, D., Masiello, L., Carlone, S., Corsaro, A., Thellung, S., Cai, T., Noonan, D. M., and Schettini, G., (1999). Somatostatin controls Kaposi's sarcoma tumor growth through inhibition of angiogenesis. *The FASEB Journal 13*, 647–655.

(*4*) Grant, M., Alturaihi, H., Jaquet, P., Collier, B., and Kumar, U., (2008). Cell growth inhibition and functioning of human somatostatin receptor type 2 are modulated by receptor heterodimerization. *Molecular Endocrinology 22*, 2278–2292.

(*5*) Siehler, S., Seuwen, K., and Hoyer, D., (1999). Characterisation of human recombinant somatostatin receptors. 1. Radioligand binding studies. *Naunyn-Schmiedeberg's archives of pharmacology 360*, 488–499.

(*6*) Hansen, L., Hartmann, B., Bisgaard, T., Mineo, H., Jørgensen, P. N., and Holst, J. J., (2000). Somatostatin restrains the secretion of glucagon-like peptide-1 and-2 from isolated perfused porcine ileum. *American Journal of Physiology-Endocrinology and Metabolism 278*, E1010.

(*7*) Chisholm, C., and Greenberg, G. R., (2002). Somatostatin-28 regulates GLP-1 secretion via somatostatin receptor subtype 5 in rat intestinal cultures. *American Journal of Physiology-Endocrinology and Metabolism 283*, E311–E317.

(*8*) Sprecher, U., Mohr, P., Martin, R. E., Maerki, H. P., Sanchez, R. A., Binggeli, A., Künnecke, B., and Christ, A. D., (2010). Novel, non-peptidic somatostatin receptor subtype 5 antagonists improve glucose tolerance in rodents. *Regulatory peptides 159*, 19–27.

(*9*) D'Addona, D., Carotenuto, A., Novellino, E., Piccand, V., Reubi, J. C., Di Cianni, A., Gori, F., Papini, A. M., and Ginanneschi, M., (2008). Novel sst5-Selective Somatostatin Dicarba-Analogues: Synthesis and Conformation-Affinity Relationships. *Journal of medicinal chemistry 51*, 512–520.

(*10*) Feytens, D., De Vlaeminck, M., Cescato, R., Tourwe, D., and Reubi, J. C., (2009). Highly Potent 4-Amino-Indolo [2, 3-c] Azepin-3-One-Containing Somatostatin Mimetics With a Range of sst Receptor Selectivities. *Journal of Medicinal Chemistry 52*, 95–104.

(*11*) Martin, R. E., Mohr, P., Maerki, H. P., Guba, W., Kuratli, C., Gavelle, O., Binggeli, A., Bendels, S., Alvarez-Sánchez, R., Alker, A., et al. (2009). Benzoxazole piperidines as selective and potent somatostatin receptor subtype 5 antagonists. *Bioorganic & medicinal chemistry letters 19*, 6106–6113.

(*12*) Alker, A., Binggeli, A., Christ, A. D., Green, L., Maerki, H. P., Martin, R. E., and Mohr, P., (2010). Piperidinyl-nicotinamides as potent and selective somatostatin receptor subtype 5 antagonists. *Bioorganic & medicinal chemistry letters 20*, 4521–4525.

(*13*) Mobarec, J. C., Sanchez, R., and Filizola, M., (2009). Modern homology modeling of G-protein coupled receptors: which structural template to use? *Journal of medicinal chemistry 52*, 5207–5216.

(*14*) Dong, S. S., Abrol, R., and Goddard, W. A., (2015). The Predicted Ensemble of Low-Energy Conformations of Human Somatostatin Receptor Subtype 5 and the Binding of Antagonists. *ChemMedChem 10*, 650–661, DOI: 10.1002/cmdc.201500023,

(*15*) Dong, S. S., Goddard, W. A., and Abrol, R., (2016). Conformational and Thermodynamic Landscape of GPCR Activation from Theory and Computation. *Biophysical Journal 110*, 2618–2629, DOI: 10.1016/j.bpj.2016.04.028,

(*16*) Manglik, A., Kim, T. H., Masureel, M., Altenbach, C., Yang, Z., Hilger, D., Lerch, M. T., Kobilka, T. S., Thian, F. S., and Hubbell, W. L., (2015). Structural Insights into the Dynamic Process of β2-Adrenergic Receptor Signaling. *Cell 161*, 1101–1111.

(*17*) Ozenberger, B. A., and Hadcock, J. R., (1995). A single amino acid substitution in somatostatin receptor subtype 5 increases affinity for somatostatin-14. *Molecular pharmacology 47*, 82–87.

(*18*) Manglik, A., and Kobilka, B., (2014). The role of protein dynamics in GPCR function: Insights from the β2AR and rhodopsin. *Current opinion in cell biology 27*, 136–143.

(*19*) Warne, T., Moukhametzianov, R., Baker, J. G., Nehmé, R., Edwards, P. C., Leslie, A. G., Schertler, G. F., and Tate, C. G., (2011). The structural basis for agonist and partial agonist action on a β1-adrenergic receptors. *Nature 469*, 241–244.

(*20*) Lebon, G., Warne, T., Edwards, P. C., Bennett, K., Langmead, C. J., Leslie, A. G., and Tate, C. G., (2011). Agonist-bound adenosine A2A receptor structures reveal common features of GPCR activation. *Nature 474*, 521–525.

(*21*) Wacker, D., Wang, C., Katritch, V., Han, G. W., Huang, X.-P., Vardy, E., McCorvy, J. D., Jiang, Y., Chu, M., and Siu, F. Y., (2013). Structural features for functional selectivity at serotonin receptors. *Science 340*, 615–619.

(*22*) Wang, C., Jiang, Y., Ma, J., Wu, H., Wacker, D., Katritch, V., Han, G. W., Liu, W., Huang, X.-P., and Vardy, E., (2013). Structural basis for molecular recognition at serotonin receptors. *Science 340*, 610–614.

(*23*) White, J. F., Noinaj, N., Shibata, Y., Love, J., Kloss, B., Xu, F., Gvozdenovic-Jeremic, J., Shah, P., Shiloach, J., and Tate, C. G., (2012). Structure of the agonist-bound neurotensin receptor. *Nature 490*, 508–513.

(*24*) Sun, D., Flock, T., Deupi, X., Maeda, S., Matkovic, M., Mendieta, S., Mayer, D., Dawson, R. J., Schertler, G. F., and Babu, M. M., (2015). Probing Gαi1 protein activation at single-amino acid resolution. *Nature structural & molecular biology 22*, 686–694.

(*25*) Flock, T., Ravarani, C. N., Sun, D., Venkatakrishnan, A., Kayikci, M., Tate, C. G., Veprintsev, D. B., and Babu, M. M., (2015). Universal allosteric mechanism for Gα activation by GPCRs. *Nature 524*, 173–179.

(*26*) Kong, H., Raynor, K., Yasuda, K., Bell, G. I., and Reisine, T., (1993). Mutation of an aspartate at residue 89 in somatostatin receptor subtype 2 prevents Na+ regulation of agonist binding but does not alter receptor-G protein association. *Molecular pharmacology 44*, 380–384.

(*27*) Roche, J. P., Bounds, S., Brown, S., and Mackie, K., (1999). A mutation in the second transmembrane region of the CB1 receptor selectively disrupts G protein signaling and prevents receptor internalization. *Molecular pharmacology 56*, 611–618.

Chapter 4

The Predicted Ensemble of Low-Energy Conformations of hSSTR5 and the Binding of Antagonists

This chapter is based on the following publication:

Dong, S. S.; Abrol, R; Goddard, W. A. "The Predicted Ensemble of Low Energy Conformations of Human Somatostatin Receptor Subtype 5 and the Binding of Antagonists". *ChemMedChem*, **2015**, *10* (4), 650–661. doi: 10.1002/cmdc.201500023

In this chapter, we focus on predicting the ensemble of structures for hSSTR5 using the ActiveGEnSeMBLE method and their binding with several known antagonists. An account on the significance of the research is in Chapter 3.

4.1 Structure Prediction of hSSTR5

Our predictions of the ensemble of low energy 3D structures for hSSTR5 followed ActiveGEnSeMBLE. An overview of ActiveGEnSeMBLE applied to hSSTR5 structure prediction is provided in Figure C.1. The procedure is detailed in Appendix B.

To prepare the starting structures for the ActiveGEnSeMBLE procedure, we first carried out PredicTM and secondary structure predictions to determine which residues are in the seven transmembrane domains (TMDs). The PredicTM result is in Figure C.2, and the final assignment of each TMD is in Figure C.3. Then we carried out multiple sequence alignments between hSSTR5 and the GPCRs with x-ray structures available, which identified human nociceptin receptor (hOPRX), mouse μ-opioid receptor (mOPRM), and human κ-opioid receptor (hOPRK) as the best candidate templates to model hSSTR5 structure. To determine the shapes of the helices, we used OptHelix and homology modeling. Then the TMD bundle of hSSTR5 was assembled based on the helix positions of each template. A total of 15 starting structures with different helical shapes and positions were generated.

Among the six parameters (x,y,h,θ,ϕ,η) that uniquely define the orientation of a rigid TMD, the hydrophobic center (HPC) residue h and the Cartesian coordinates (x,y) of the HPC were taken from the template. Among the helical tilts and rotations

(θ,ϕ,η), the helical rotations η were first sampled using the BiHelix method with a sampling range of $\Delta\eta$ from 0° to 360° and a step size of 30°.

The top 10 structures from the BiHelix step are shown in Table C.2, where we see that all 10 use homology helix shapes. Since all three templates were represented in the top 10, we used the rotations for the best candidate from each template in the next step, SuperBiHelix (optimizing tilts).

SuperBiHelix optimizes (θ,ϕ,η) based on the best structures from BiHelix. For each of these, we first carried out a coarse sampling step ($\Delta\theta$ = 0, ±15°; $\Delta\phi$ = 0, ±45°, ±90°; $\Delta\eta$ = 0, ±30°, other selected angles) from angles optimized in BiHelix. This sampled at least $(3\times5\times3)^7 \approx 374$ billion configurations from which we built and optimized the lowest 2000 seven-helix bundles. This was done for all three starting templates (mOPRM, hOPRK, hOPRX), with the sampling space for each template summarized in Table C.3, and the resulting lowest-energy structure for each template shown in Table 4.1.

Template	$\Delta\varphi$ (°)							$\Delta\eta$ (°)							$\Delta\theta$ (°)	E_{CNti} (kcal mol^{-1})
	H1	H2	H3	H4	H5	H6	H7	H1	H2	H3	H4	H5	H6	H7	All	
mOPRM	0	0	0	90	0	0	0	0	0	0	-30	30	0	0	0	-303.3
hOPRX	0	0	0	-90	0	-45	0	0	0	0	0	0	60	0	0	-324.9
hOPRK	0	0	0	-30	0	0	30	0	0	0	0	0	0	0	0	-306.5

Table 4.1: Coarse SuperBihelix/SuperCombiHelix optimal structures of the three templates. The angles $\Delta\phi$, $\Delta\eta$ and $\Delta\theta$ are deviations from the respective homology templates. The $\Delta\theta$ values for all helices are zero for the cases listed.

To predict the structures for inactive states, we selected the lowest-energy structure with a TM3-TM6 ionic lock for each template. Then we carried out a finer SuperBiHelix sampling ($\Delta\theta$ = 0, ±15°; $\Delta\phi$ = 0, ±15°, ±30°; $\Delta\eta$ =0, ±30°). This examined $(3\times5\times3)^7 \approx 374$ billion configurations from which we built and optimized the lowest 2000 seven-helix bundles.

As shown in Table 4.2, the top 10 structures from this fine SuperBiHelix sampling all come from the mOPRM template except for the one ranked 5th which is from hOPRK. Thus, we focused on structures using the mOPRM template in the subsequent procedures. The structure ranked 4th by E_{CNti} has the largest number of interhelical hydrogen bonds among the top 10 among which

- the TM3-TM6 (3-6) ionic lock [R137$^{3.50}$-E243$^{6.30}$],

Rank	Template	$\Delta\theta$ (°)	$\Delta\varphi$ (°)							$\Delta\eta$ (°)							Number of HB	Number of InterHHB	Conserved Interactions	E_{CNti} (kcal mol^{-1})
		H1	H1	H2	H3	H4	H5	H6	H7	H1	H2	H3	H4	H5	H6	H7				
1	mOPRM	0	0	0	0	120	15	-15	0	0	0	0	-30	0	0	0	14	11	3-6, 1-2-7	-352.9
2	mOPRM	0	0	0	0	105	15	-15	0	0	0	0	0	0	0	0	13	9	3-6, 1-2-7	-352.8
3	mOPRM	0	0	0	0	120	15	-15	0	0	0	0	0	0	0	0	16	10	3-6, 1-2-7	-351.3
4	mOPRM	0	0	0	0	105	0	-15	0	0	0	0	-30	0	0	0	19	16	3-6, 1-2-7, 2-3-4	-350.7
5	hOPRK	-15	0	0	-15	-45	0	-15	15	0	0	0	0	-30	0	0	20	11	3-6, 2-7	-347.8
6	mOPRM	0	0	0	0	75	15	-15	0	0	0	0	-30	0	0	0	14	12	3-6, 1-2-7	-346.9
7	mOPRM	0	0	0	-15	60	30	15	0	0	0	0	0	60	0	0	14	10	3-6, 1-7, 2-4	-345.7
8	mOPRM	0	0	0	0	105	15	0	15	0	0	0	0	0	0	0	16	10	3-6, 1-2-7	-343.7
9	mOPRM	0	0	0	-15	75	15	0	0	0	0	0	0	60	0	0	17	10	1-2-7, 2-4	-342.8
10	mOPRM	0	0	0	0	120	0	0	15	0	0	0	-30	0	0	0	18	14	3-6, 1-2-7, 2-3-4	-341.4

Table 4.2: Top 10 structures from fine SuperBiHelix/SuperCombiHelix. The $\Delta\theta$ values of H2, H3, H4, H5, H6 and H7 are all zero for the cases listed. The angles $\Delta\phi$, $\Delta\eta$ and $\Delta\theta$ are deviations from the respective homology templates. HB: hydrogen bond. InterHHB: interhelical hydrogen bond. Conserved interactions: interhelical hydrogen bonds conserved in Class A GPCRs. The numbers represent which TMDs the hydrogen bonds are forming between.

- the TM1-TM2-TM7 (1-2-7) network [N58$^{1.50}$-S297$^{7.46}$-D86$^{2.50}$, N300$^{7.49}$-D86$^{2.50}$] and
- the TM2-TM3-TM4 (2-3-4) network [W164$^{4.50}$-N81$^{2.45}$-C129$^{3.42}$]

are conserved from the inactive states across class A GPCR x-ray structures (Figure 4.1). In addition, there are 7 strong interactions involving residues that are conserved in all or most hSSTRs (Figure 4.1). They are

- Y78$^{2.43}$-D136$^{3.49}$,
- N122$^{3.35}$-A85$^{2.49}$ (A85$^{2.49}$ backbone),
- T125$^{3.38}$-S167$^{4.53}$,
- K227$^{5.64}$-Y138$^{3.51}$,
- Y294$^{7.43}$-D119$^{3.32}$,
- R113$^{3.26}$-L174$^{4.60}$ (L174$^{4.60}$ backbone; valine in hSSTR1,4), and
- K245$^{6.32}$-F306$^{7.55}$ (F306$^{7.55}$ backbone; R254$^{6.32}$-F315$^{7.55}$ in hSSTR3).

Therefore, we considered this structure as the most promising inactive-state candidate, denoted InactiveConf1. We also found an important polar interaction,

- Y286$^{7.35}$-N271$^{6.58}$,

that seems likely only for hSSTR3 and hSSTR5 because the other hSSTRs do not have a tyrosine on 7.35. In addition, we found one interaction,

- T117$^{3.30}$-S171$^{4.57}$,

that we expect to be unique to hSSTR5. These interactions are shown in Figure 4.1.

In order to obtain a diverse set of low energy protein structures, we selected two other protein conformations from the lowest-energy 25 predicted hSSTR5 structures (listed in Table C.4). Here we selected the two that have the largest root-mean-square deviation (RMSD) with InactiveConf1 and with each other. These two are labeled InactiveConf2 (ranked 7th in Table 4.2, and 6th in Table C.4) and InactiveConf3 (ranked 16th in Table C.4).

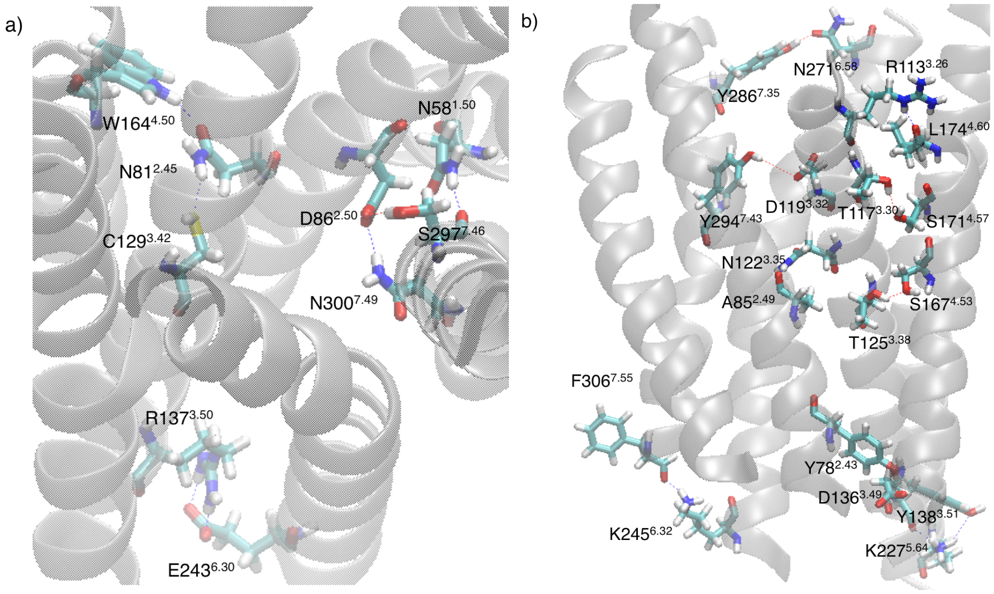


Figure 4.1: Interhelical hydrogen bonds in the predicted hSSTR5 structure that are a) conserved across inactive-state Class A GPCRs, and b) possibly conserved among hSSTRs or unique to hSSTR5.

To obtain structures that might be candidates for active-states (with TM6 well separated from TM3), we carried out a finer sampling starting from the best structures from the coarse sampling that satisfy specific structural criterion, which is described below. The resulting optimal active-state structures are named ActiveConf1 and ActiveConf2.

To distinguish potential active-state from inactive-state structures, we defined R_{36}, the measure of the distance between the intracellular (IC) ends of TM3 and TM6, to be the shortest distance between the backbone atoms of the four residues at the IC ends of TM3 and TM6. We selected the lowest-energy structure with an R_{36} value 4 Å larger than the R_{36} value of 7.18 Å from the inactive-state structure InactiveConf1. The active-inactive R_{36} difference of 4 Å was chosen because the x-ray structures for both human β_2-adrenergic receptor (hβ_2AR) and bovine rhodopsin display this feature. Then we carried out a finer sampling of (θ,ϕ,η) on this selected structure to obtain the first putatively active conformation ActiveConf1. Substituting the TM6 shape in this starting structure with the TM6 from the homology model with active hβ_2AR x-ray structure followed by finer sampling gives the second putatively active conformation ActiveConf2, discussed in Chapter 3.

A summary of all structures used in the following antagonist binding study is in Table 4.3.

Structure	$\Delta\theta$ (°)	$\Delta\varphi$ (°)							$\Delta\eta$ (°)							E_{CNti} (kcal mol^{-1})	R_{36} (Å)	RMSD (Å)	
	H6	H1	H2	H3	H4	H5	H6	H7	H1	H2	H3	H4	H5	H6	H7			Inactive-Conf1	Inactive-Conf2
Inactive-Conf1	0	0	0	0	105	0	-15	0	0	0	0	-30	0	0	0	-350.7	7.18	0.0	1.7
Inactive-Conf2	0	0	0	-15	60	30	15	0	0	0	0	0	60	0	0	-345.7	7.59	1.7	0
Inactive-Conf3	0	0	0	-15	60	-30	15	0	0	0	0	0	30	0	0	-328.9	7.56	1.4	1.8
Active-Conf1	0	0	0	0	0	0	-60	-15	0	0	0	0	30	-90	0	-313.2	10.71	-	-
Active-Conf2	-15	0	0	-15	0	-30	30	0	0	0	0	0	30	-30	0	-306.6	10.20	-	-

Table 4.3: Summary of structures from fine SuperBihelix/SuperCombiHelix sampling and docked by the antagonists. The $\Delta\theta$ values for H1, H2, H3, H4, H5, and H7 are all zero for the cases listed. The angles $\Delta\phi$, $\Delta\eta$, and $\Delta\theta$ are deviations from the respective initial homology templates.

4.2 Antagonist Binding

To validate the structures predicted for the hSSTR5 receptor, we predicted the binding site and energy for 5 known small molecule antagonists to the 5 predicted protein structures (InactiveConf1,2,3; ActiveConf1,2). The antagonists chosen for the docking were from a series of benzoxazole piperidines screened by Martin and coworkers(*1*) which exhibit a wide range of binding affinities while retaining the same structural scaffold of the ligand (Figure 4.2). The molecules are labeled "Mx", with "x" preserving the numbering scheme from Martin et al., 2009.(*1*) We selected M38, M40, M42, M59, M60 based on their experimentally determined binding affinities (K_i) for hSSTR5—M59 is the most potent derivative reported, exhibit- ing a K_i value of 3 nm, while the other compounds selected ex- hibit a diverse range of binding affinities (from 23 nm to over 1000 nm).

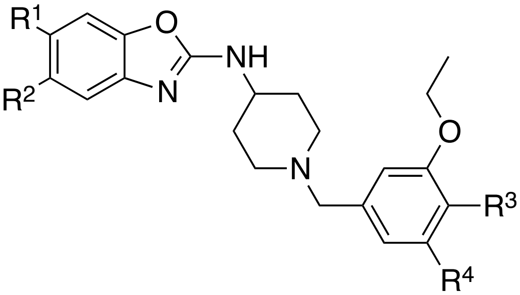


Figure 4.2: The common scaffold of the docked antagonists.

The experimental binding affinity, predicted binding site and energy for the 5 antagonists chosen are summarized in Table 4.4. To determine the best pose for each antagonist, we allowed each antagonist to select its preferred conformations out of the docking results to all 5 predicted protein structures. For each ligand, we selected the optimal pose for the ligand's preferred binding mode using the best unified-cavity (UCav) energy in comparing the ligand binding. The UCav energy ranked the five ligands as M59 (best) < M60 < M40 < M42 < M38 (worst) while the binding energy calculated from $\Delta G_1 - \Delta G_2 = RT\ln(K_{i2}/K_{i1})$ using experimental binding constants should give M59 < M60 < M38 < M40 < M42. Thus, only M38 is an outlier. The UCav energies range over a factor of 9 of the binding energies calculated from the experimental binding constants, which range over 3.57 kcal mol^{-1} (Table C.5).

To further investigate how well our predicted binding energies correlate with the experimental results, we plot the UCav energy against the negative logarithm of the experimental binding constants in Figure 4.3. This shows that the UCav energies

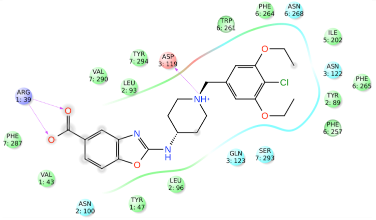
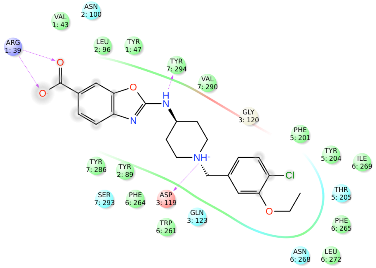
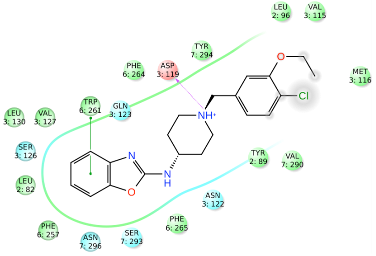
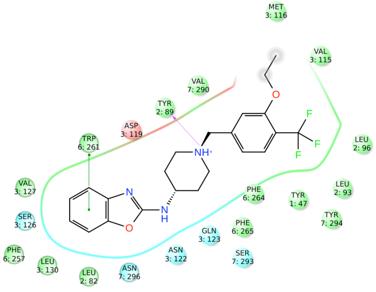
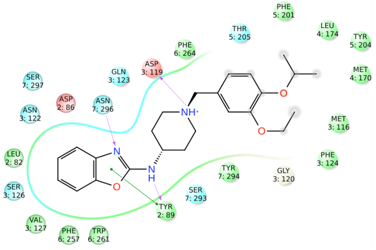
Ligand	K_i (nM)	UCav E (kcal mol^{-1})	Ligand Interaction Diagram
M59	3	-118.48	 PHE 6: 264 ASN 6: 268 TRP 6: 261 ILE 5: 202 TYR 7: 294 ASP 3: 119 VAL 7: 290 LEU 2: 93 ASN 3: 122 PHE 6: 265 ARG 1: 39 TYR 2: 89 PHE 6: 257 PHE 7: 287 GLN 3: 123 SER 7: 293 VAL 1: 43 LEU 2: 96 TYR 1: 47 ASN 2: 100
M60	23	-115.00	 ASN 2: 100 VAL 1: 43 ARG 1: 39 LEU 2: 96 TYR 1: 47 TYR 7: 294 VAL 7: 290 GLY 3: 120 PHE 5: 201 TYR 5: 204 ILE 6: 269 TYR 7: 286 TYR 2: 89 THR 5: 205 SER 7: 293 PHE 6: 264 ASP 3: 119 GLN 3: 123 TRP 6: 261 PHE 6: 265 ASN 6: 268 LEU 6: 272
M38	113	-86.35	 LEU 2: 96 VAL 3: 115 TYR 7: 294 PHE 6: 264 ASP 3: 119 MET 3: 116 TRP 6: 261 GLN 3: 123 LEU 3: 130 VAL 3: 127 SER 3: 126 TYR 2: 89 VAL 7: 290 LEU 2: 82 ASN 3: 122 PHE 6: 265 PHE 6: 257 ASN 7: 296 SER 7: 293
M40	524	-92.88	 MET 3: 116 VAL 3: 115 VAL 7: 290 TYR 2: 89 ASP 3: 119 TRP 6: 261 LEU 2: 96 LEU 2: 93 VAL 3: 127 PHE 6: 264 TYR 1: 47 SER 3: 126 TYR 7: 294 GLN 3: 123 PHE 6: 265 PHE 6: 257 LEU 3: 130 ASN 3: 122 SER 7: 293 LEU 2: 82 ASN 7: 296
M42	>1000	-87.04	 PHE 5: 201 THR 5: 205 LEU 4: 174 TYR 5: 204 PHE 6: 264 MET 4: 170 SER 7: 297 GLN 3: 123 ASP 3: 119 ASP 2: 86 ASN 7: 296 ASN 3: 122 MET 3: 116 LEU 2: 82 PHE 3: 124 TYR 7: 294 GLY 3: 120 SER 3: 126 SER 7: 293 TYR 2: 89 VAL 3: 127 PHE 6: 257 TRP 6: 261

Table 4.4: The predicted binding energy and pharmacophore for the antagonists in the binding study. Each ligand interaction diagram (LID) was generated using Maestro 9.3.(2) The cutoff distance for the residues shown is 4.0 Å. Hydrophobic interactions: green; polar interactions: blue; hydrogen bonds (cutoff distance: 2.5 Å): purple arrows; π-π stacking: straight green lines.

for the optimal poses of the antagonist series correlate with the experimental pK_i with a linear regression coefficient of 0.78. This suggests that UCav captures the essential aspects of the relative binding affinities of these antagonists.

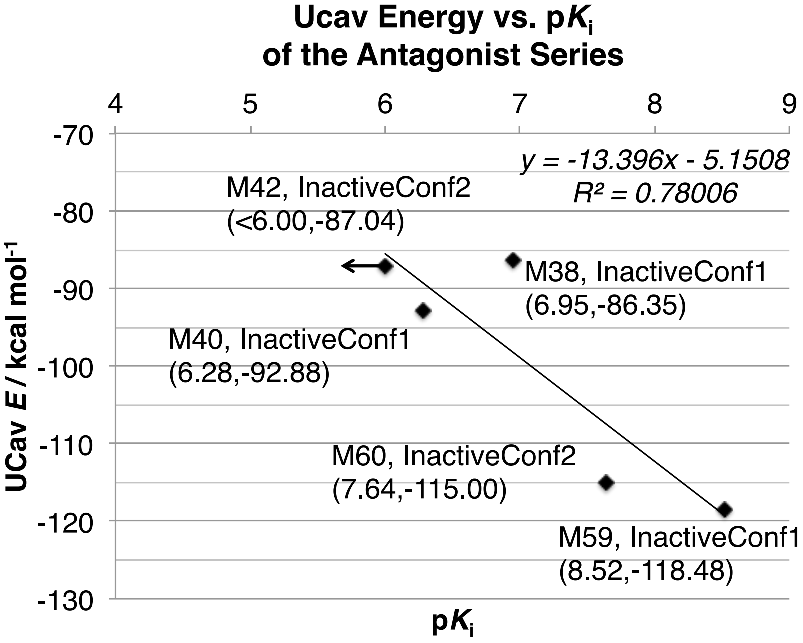


Figure 4.3: Relationship between the UCav energy from docking and the experimental pK_i of the antagonist series. R^2 is the coefficient of determination. The arrow represents the possible direction of change of the M42 data point.

We also find that M42 and M60 favor the InactiveConf2 conformation while the other three favor InactiveConf1, which means InactiveConf1 and InactiveConf2 could possibly be two inactive conformations selected by the antagonist series. This suggests that structures of hSSTR5 predicted to be more stable are more likely to be in the inactive state than the less stable ones.

Table 4.4 shows the ligand interaction diagram (LID) of the best pose for each antagonist. We find that the antagonists predominantly bind with a pocket defined by TMDs 1-2-3-6-7. In these best poses, all antagonists form a salt bridge between their positively charged piperidine amine group and D119$^{3.32}$. This aspartic acid on TM3 is conserved in all somatostatin receptors, and mutagenesis studies have shown it is essential in SST binding by forming an electrostatic interaction with a positively charged group in SST.(*3*, *4*) Therefore, our result has further confirmed that D119$^{3.32}$ is a critical residue in hSSTRs' binding with positively charged ligand groups.

The 3D visualization of these poses is shown in Figure 4.4. For M59 and M60, the ligand position is clearly dominated by the two salt bridges: one between the positively charged amine group in the ligand and D119$^{3.32}$, and the other between the carboxylic group in the ligand and R39$^{1.31}$. For M59, we did molecular dynamics simulation with lipid membrane and a water box, and both salt bridges became water-mediated after the MD simulation (discussed in the next section). Unlike D119$^{3.32}$, R39$^{1.31}$ is not conserved in any hSSTRs. This might explain why the antagonists with polar groups at R^1 or R^2 are extremely selective towards the subtype 5.(*1*) Mutating R39$^{1.31}$ to a negatively charged residue, a nonpolar residue, or Ser (as in hSSTR1, 2, 3) should be able to test this hypothesis.

Other residues playing an important role in these high affinity antagonists are polar residues N100$^{2.64}$, Q123$^{3.36}$, N268$^{6.55}$ and S293$^{7.42}$, and nonpolar residues V43$^{1.35}$, Y47$^{1.39}$, Y89$^{2.53}$, W261$^{6.48}$, F264$^{6.51}$, F265$^{6.52}$, V290$^{7.39}$ and Y294$^{7.43}$. Residues shared by M59, M60, M38 and M40 are polar residues Q123$^{3.36}$ and S293$^{7.42}$, and nonpolar residues Y89$^{2.53}$, W261$^{6.48}$, F264$^{6.51}$, F265$^{6.52}$, V290$^{7.39}$ and Y294$^{7.43}$. This is shown in the pharmacophore mapping in Table 4.4. The residues interacting with the strongly binding antagonists that are missing in the predicted pose for the nonbinding molecule M42 are F265$^{6.52}$ and V290$^{7.39}$. Since M42 binding is not experimentally detected, we can deduce that Phe265Ala and/or Val290Ala mutations may cause the other antagonists to have a decreased affinity towards hSSTR5.

In the predicted binding poses, M38 and M40 both form a π-π stacking interaction between the benzoxazole and W261$^{6.48}$, but both lack the salt bridge with R39$^{1.31}$ that is found in the predicted interactions for M59 and M60. Therefore, without the strong electrostatic interaction constraining the ligand position, the weaker π-π interaction becomes a dominating force of the ligand with the protein. Although M38 does not have a stronger binding in docking than M40 and M42 as predicted, its exposed chlorine group may lower the binding energy once solvation is taken into account.

Apart from the polar groups, the ligand size also plays a role in determining the binding affinities. Figure 4.4 shows that the ethoxy group at the R^4 position in M59 prevents the phenyl group of M59 from being parallel to the hydrophobic plane of the GPCR as in the M60 pose. The ethoxy group has directed the M59 phenyl head to go deeper into the binding pocket and reach more polar and nonpolar residues than M60. This explains why M59 has a higher affinity to hSSTR5 than M60.

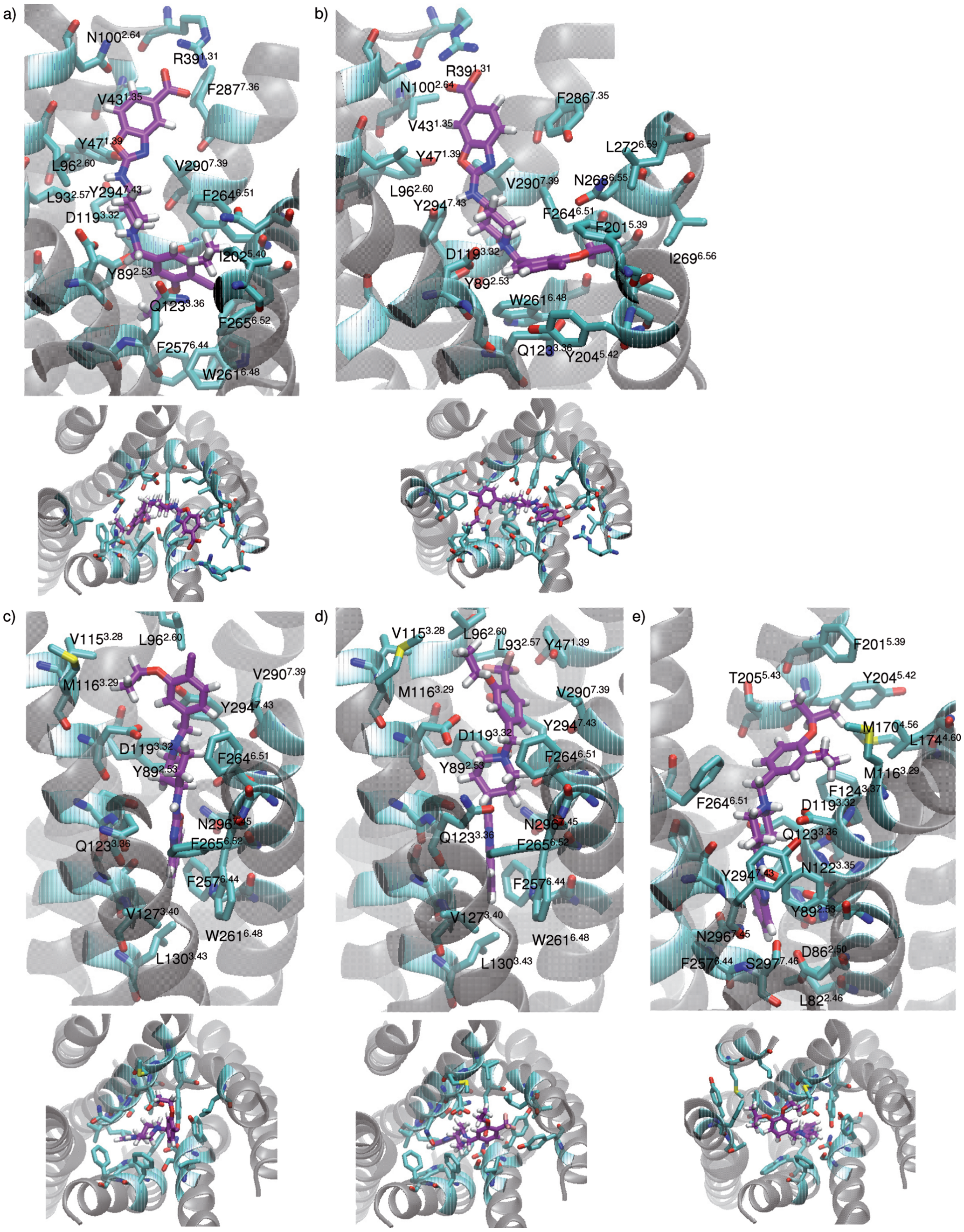


Figure 4.4: Predicted 3D structures of the best docking pose of a) M59, b) M60, c) M38, d) M40, and e) M42. Each pose is presented in both the side view and the top view. Ligand carbon: purple, protein carbon: cyan.

4.3 Molecular Dynamics Simulation

In order to anneal and validate our predicted structure, we carried out 50 ns of MD simulation of the system with the protein embedded in explicit lipid and water box starting with the predicted structure of M59-bound InactiveConf1. The RMSD analysis of the trajectory (Figures C.4 and C.5) and fluctuation of R_{36} (Figure 4.5) all suggest that the protein starts to rearrange to a different state at ~41 ns. Such fluctuations between slightly different states of the GPCR along the trajectory are typical in GPCR MD simulations during which water is diffusing into and throughout the protein, modulating various hydrogen bonds and other interactions.

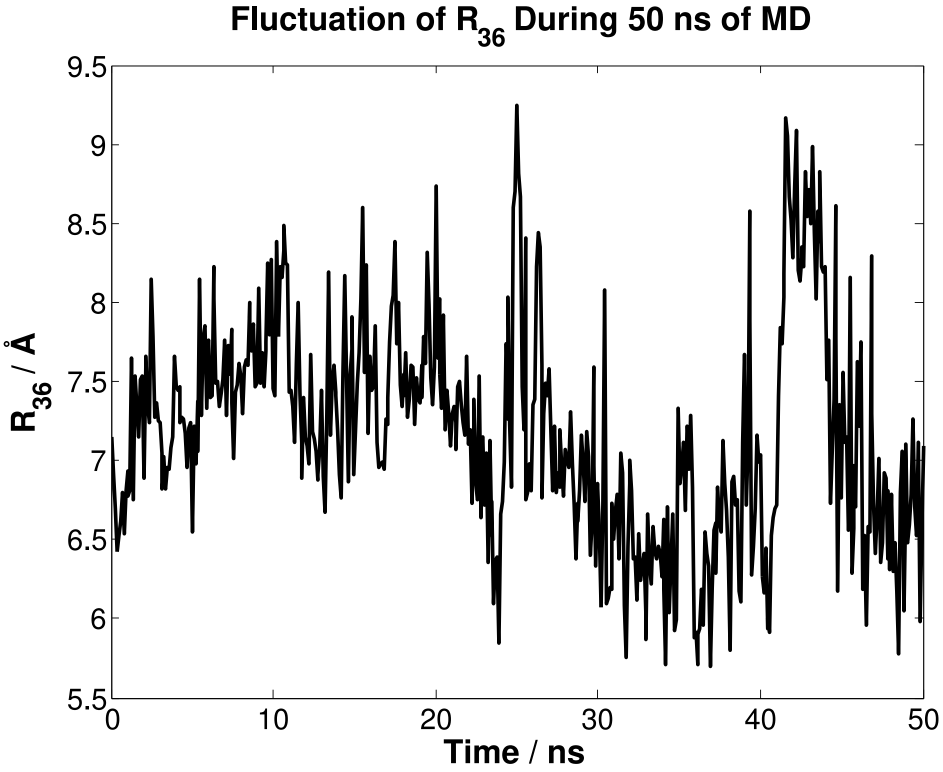


Figure 4.5: The fluctuation of R_{36} during the 50 ns MD simulation of the M59-bound predicted hSSTR5 structure. R_{36} is the shortest distance between the backbone atoms of the intracellular ends of TM3 and TM6.

The hydrogen bond distances for various interactions are shown in Figures C.6-C.14 along the trajectory. The constancy of these interactions suggest that the overall protein structural features from the region of 33 ns to 41 ns and that of later times are similar with most structural features maintained at the end of 50 ns trajectory. Thus we consider these structural features to provide a reliable representation of the structure.

We find that the ionic lock between the R137$^{3.50}$ in the DRY motif on TM3 and E243$^{6.30}$ on the IC end of TM6 breaks after 20 ns (Figure C.15). Then, R137$^{3.50}$ establishes electrostatic interactions with D136$^{3.49}$, while E243$^{6.30}$ forms a salt bridge with R241$^{6.28}$ on the IC loop 3 (ICL3). The strong electrostatic interaction with the loop explains the changes in the interatomic distance between R137$^{3.50}$ and E243$^{6.30}$ in Figure S14. D136$^{3.49}$ also makes a polar interaction with R151 on the IC loop 2 (ICL2) as shown in Figure 4.6. Similar polar interaction patterns were also observed in the x-ray crystallographic structure of mOPRM, coupling D164$^{3.49}$ and R165$^{3.50}$ in the DRY motif, and coupling D164$^{3.49}$ and R179 on ICL2 (Figure 4.6).(*5*) The x-ray structure for hβ_2AR was also found to have analogous patterns (R1313.50 interacts with D130$^{3.49}$ rather than E268$^{6.30}$, and at the same time D130$^{3.49}$ has polar interaction with S143 on ICL2).(*6*) The observation that the salt bridge between the IC sides of TM3 and TM6 is not formed in the antagonist-bound mOPRM structures supports our observation for the predicted hSSTR5. In addition, all the other interhelical interactions between side chains found in the apo hSSTR5 structure remain intact except for the one involving D119$^{3.32}$ because D119$^{3.32}$ now engages in interaction with the ligand. Among the interhelical interactions, K227$^{5.64}$-Y138$^{3.51}$ as well as T117$^{3.30}$-S171$^{4.57}$ become water-mediated.

Figure 4.5 shows how R_{36} changes during the 50 ns MD. It remains in the inactive state range 87% of the time if we set the inactive/active cut-off to be 8.0 Å, and 71% of the time if we set the cut-off 7.5 Å. We do not find much rotation of TM3 or TM6 relative to each other because the closest backbone atoms between the IC side of TM3 and TM6 remain to be between the Cα atoms of R137$^{3.50}$ and T247$^{6.34}$ 81% of the time. In addition, the polar interaction between R137$^{3.50}$ and T247$^{6.34}$ side chains converges to a water-mediated hydrogen bond with a length of ~4.5 Å although the bond length starts from 5.2 Å and quickly drops to 1.9 Å (Figure C.12). This is an intriguing result because the polar interaction R165$^{3.50}$-T279$^{6.34}$ is also found in the antagonist-bound mOPRM x-ray structure, and mutating T279$^{6.34}$ to lysine (which most likely breaks this polar interaction) can result in a constitutively active mOPRM receptor.(*5*) We can also infer that the distance between R137$^{3.50}$ and T247$^{6.34}$ is critical in determining hSSTR5 activity too, and mutating T247$^{6.34}$ to a lysine is likely to give a constitutively active hSSTR5. More importantly, since both R137$^{3.50}$ and T247$^{6.34}$ are conserved in all hSSTRs, this hypothesis may be extendable to all hSSTRs.

Analysis of the changes of several protein-ligand interactions during the dynamics finds that all the protein-ligand salt bridges become water-mediated during the MD simulation. The salt bridge with D119$^{3.32}$ starts to be water-mediated after 4.47 ns, and the one with R39$^{1.31}$ starts to be water-mediated after 2.41 ns. Although there are fluctuations during the 50 ns process, during 33-41 ns, distance between

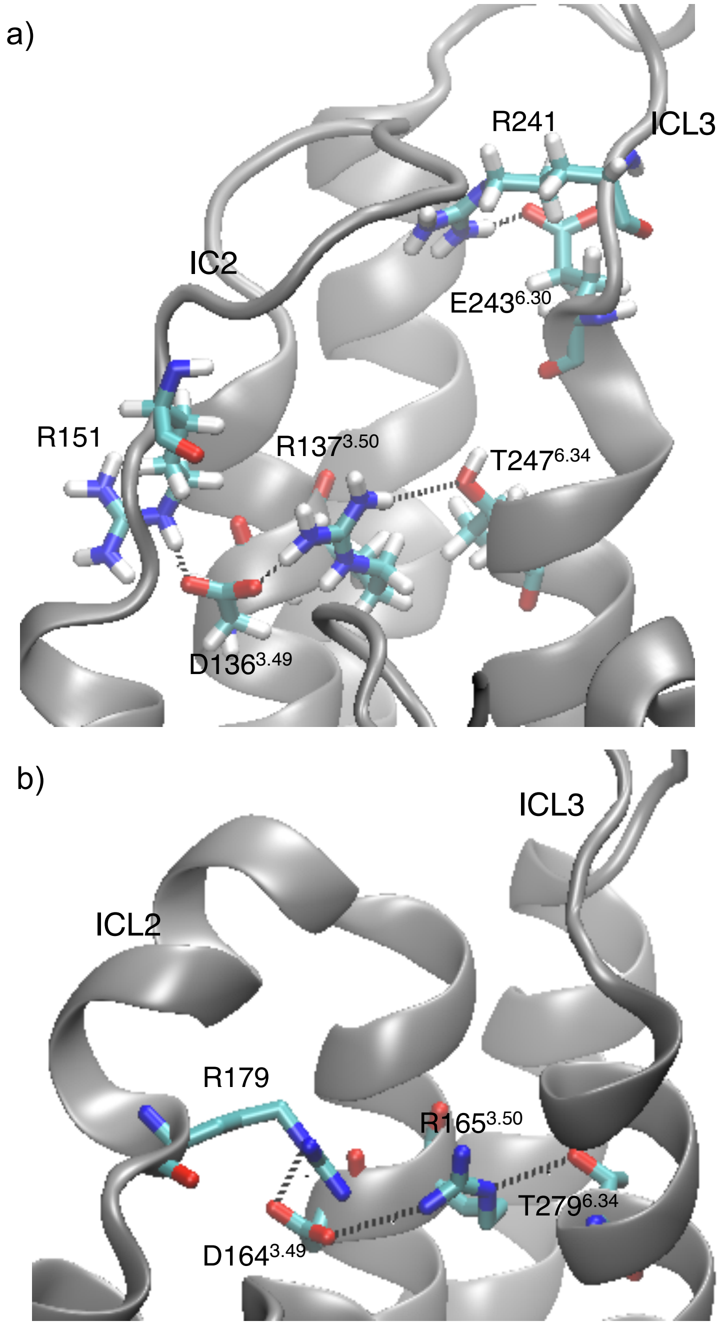


Figure 4.6: Interactions between important residues on the intracellular end of a) the predicted hSSTR5 structure after 50 ns MD with antagonist M59, and b) mOPRM X-ray structure (Protein Data Bank identifier (PDB ID): 4DKL (*5*)). The intracellular loop 2 is labeled ICL2, and the intracellular loop 3 is labeled ICL3.

M59 piperidine amine nitrogen atom and D119$^{3.32}$ carboxylic acid oxygen atom fluctuates around 4.5 Å (Figure C.13). The distance between the M59 carboxylic acid oxygen atom and R39$^{1.31}$ amine nitrogen atom fluctuates more vigorously as shown in Figure C.14, but this could be because R39$^{1.31}$ becoming part of the loop during the dynamics.

Although the protein has a structural shift right after 41 ns and the protein can be in different states a few nanoseconds before and after 41 ns, the two states have many similar antagonist-bound inactive-state characters. Therefore, we conclude that the MD simulation retains the character of an antagonist-bound inactive-state structure.

4.4 Conclusions

In this study, we predicted the ensemble of low energy structures of hSSTR5 and found plausible binding sites for a series of antagonists with a common scaffold but a diverse set of binding constants. We obtained binding energies consistent with the experimental binding constants. These structures exhibit a TM3-TM6 coupling associated with an inactive GPCR. This indicates that predicted structures InactiveConf1 and InactiveConf2 are reasonable hSSTR5 inactive-state structures, and that ActiveGEnSeMBLE is effective in predicting inactive-state GPCR structures as well as active-state structures discussed in Chapters 2 and 3 . In addition, we have identified residues that might be critical in antagonist binding to hSSTR5, and the results are able to rationalize the order of experimentally determined binding affinities for the five antagonists in the series. Furthermore, the MD simulations show that our antagonist-bound InactiveConf1 structure gains features consistent with those experimentally found in closely related GPCRs. We also introduced an approach aimed at systematically sampling structures in which TM6 is well separated from TM3 as candidates for active structures in addition to sampling small TM3-TM6 separation inactive structures.

In conclusion, this study provides structural information for understanding the antagonist binding of hSSTR5 that likely to be useful in designing new small molecule antagonists for hSSTR5. We have also provided structural features that are possible to be extended to other hSSTRs.

References

(*1*) Martin, R. E., Mohr, P., Maerki, H. P., Guba, W., Kuratli, C., Gavelle, O., Binggeli, A., Bendels, S., Alvarez-Sánchez, R., Alker, A., et al. (2009). Benzoxazole piperidines as selective and potent somatostatin receptor subtype 5 antagonists. *Bioorganic & medicinal chemistry letters 19*, 6106–6113.

(*2*) Maestro, version 9.3, Schrödinger, LLC, New York, NY, 2012.

(*3*) Nehrung, R. B., Meyerhof, W., and Richter, D., (1995). Aspartic acid residue 124 in the third transmembrane domain of the somatostatin receptor subtype 3 is essential for somatostatin-14 binding. *DNA and cell biology 14*, 939–944.

(*4*) Strnad, J., and Hadcock, J. R., (1995). Identification of a critical aspartate residue in transmembrane domain three necessary for the binding of somatostatin to the somatostatin receptor SSTR2. *Biochemical and biophysical research communications 216*, 913–921.

(*5*) Manglik, A., Kruse, A. C., Kobilka, T. S., Thian, F. S., Mathiesen, J. M., Sunahara, R. K., Pardo, L., Weis, W. I., Kobilka, B. K., and Granier, S., (2012). Crystal structure of the μ-opioid receptor bound to a morphinan antagonists. *Nature 485*, 321–326.

(*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., and Kobilka, B. K., (2007). High-resolution crystal structure of an engineered human β2-adrenergic G protein-coupled receptors. *Science 318*, 1258–1265.

Chapter 5

Gaussion Hartree Approximated Quantum Mechanics for Large-Scale Nonadiabatic Electron Dynamics

5.1 Introduction

Many chemical events, those where the nuclei move on a single potential energy surface (PES), can be simulated by quantum chemistry approaches with the assumption that electrons adjust instantaneously to the slower nuclear motion. With this assumption, which is called the Born-Oppenheimer (BO) approximation, the electronic component and the nuclear component of the wavefunction of the system can be decoupled mathematically. This forms the basis of the standard quantum chemistry methods.

However, there are also a large number of important processes in nature, such as photochemical and electrochemical reactions, that the PESs are close in energy and the nuclei transit between different PESs. During the transition, the motion of electrons and nuclei are strongly coupled, and the BO approximation breaks down. Such a transition is called a nonadiabatic transition, the theory for which was first proposed independently by Landau (*1*) and Zener (*2*) in 1932. A number of other approaches have been developed since then, including two mixed quantum-classical methods that are most widely adopted for nonadiabatic dynamics simulations, Ehrenfest dynamics (*3*–*6*) and surface hopping (*7*, *8*). Ehrenfest dynamics uses a mean-field approach in which the nuclei move on a single PES that averages over all quantum states, while the surface hopping method allows stochastic electronic transitions between states with the transition probability calculated using quantum mechanical or semiclassical methods. A number of variants of these methods have been developed, but they are still computationally demanding due to the use of advanced quantum chemistry methods to combine with these schemes, such as combining Ehrenfest dynamics with time-dependent Hartree Fock (TDHF) (*9*) or time-dependent density functional theory (TDDFT) (*10*), and the use of *ab initio* methods or TDDFT to compute the PESs for surface hopping. As many nonadiabatic phenomena happen in large systems, such as photosynthesis, etching of silicon, and insulator-to-metal transition of dense deuterium liquid, it is necessary to develop new methods to enable nonadiabatic dynamics simulations in large scale.

Aiming at solving this problem, Su et al. from our group have developed the electron force field (eFF)(*11*) framework for simulating large-scale nonadiabatic dynamics in condensed matter. In the eFF framework, the particles follow wave packet dynamics, with the nuclei represented by classical point charges propagated classically and electrons represented by wave packets. The use of wave packets relates to the perspective that, in systems that are highly excited with a high density of quasi-degenerate electronic states, there are continuous nonadiabatic transitions and large fluctuations of electronic states, which should be described by wave packets dynamics of electrons.(*12*)

Unlike the *ab initio* excited state dynamics methods which can only be realistic for hundreds of electrons on picosecond timescale, the eFF method expects to simulate hundreds of thousands of electrons on nanosecond timescale in a reasonable amount of time. It has been applied to studying problems including the shock Hugoniot curves of various materials,(*11*, *13*, *14*) Auger-induced chemistry,(*15*) Coulomb explosion in silicon and carbon,(*16*) and brittle fracture of silicon.(*17*) However, it bears certain shortcomings that inspired us to develop the Gaussion Hartree Approximated Quantum Mechanics (GHA-QM) method (*18*) based on its framework. A brief review of the eFF formulation and a discussion on the exploration and new improvements in GHA-QM are in the next sections.

5.2 The Electron Force Field (eFF) Framework

The eFF method relates to Fermion molecular dynamics (FMD),(*19*) wave packet molecular dynamics (WPMD),(*20*) and floating spherical Gaussian orbital (FSGO) methods.(*21*) In eFF, the electrons $\{j\}$ are represented by floating spherical Gaussian (FSG) wave packets with the Cartesian coordinates of the center of the FSG $\overrightarrow{R_j}$ and the FSG width (size) s_j as variables. The total wavefunction is a Hartree product of single-electron wave packets:

$$\Psi(\overrightarrow{r_i}) \propto \prod_j \exp\left[-\left(\frac{1}{s_j^2} - \frac{2p_{s_j}}{s_j}\frac{i}{\hbar}\right)\left(\overrightarrow{r_j} - \overrightarrow{R_j}\right)^2\right] \exp\left[\frac{i}{\hbar}\overrightarrow{p}_{\overrightarrow{R_j}} \cdot \overrightarrow{r_j}\right]. \qquad (5.1)$$

The normalized wavefunction of each electron is

$$\phi_j(\overrightarrow{r_j}) = \left(\frac{\sqrt{2}}{\sqrt{\pi}s_j}\right)^{\frac{3}{2}} \exp\left[-\frac{(\overrightarrow{r_j} - \overrightarrow{R_j})^2}{s_j^2}\right]. \qquad (5.2)$$

The use of Hartree product instead of antisymmetric wavefunctions reduces the $O(N^4)$ scaling of pairwise electrostatic energy evaluations to $O(N^2)$, which is desirable. To account for the energy contribution from antisymmetrization, we include a Pauli potential in the total energy expression. The total energy E is thus the sum of Hartree product electronic kinetic energy E_{ke}, Hartree product electrostatic energies $E_{\text{nuc-nuc}}$, $E_{\text{nuc-elec}}$, $E_{\text{elec-elec}}$, and Pauli energy E_{Pauli}:

$$E = E_{\text{ke}} + E_{\text{nuc-nuc}} + E_{\text{nuc-elec}} + E_{\text{elec-elec}} + E_{\text{Pauli}}, \tag{5.3}$$

where

$$\begin{aligned}
E_{\text{ke}} &= \sum_j \langle j| -\frac{1}{2}\nabla_j^2 |j\rangle = \sum_j \frac{3}{2}\frac{1}{s_j^2} \\
E_{\text{nuc-nuc}} &= \sum_{m<n} \frac{Z_m Z_n}{R_{mn}} = \sum_{m<n} \frac{Z_m Z_n}{R_{mn}} \\
E_{\text{nuc-elec}} &= \sum_{n,j} \langle j| -\frac{Z_n}{r_{nj}} |j\rangle = -\sum_{n,j} \frac{Z_n}{R_{nj}} \text{Erf}\left(\frac{\sqrt{2}R_{nj}}{s_j}\right) \\
E_{\text{elec-elec}} &= \sum_{i<j} \langle ij| \frac{1}{r_{ij}} |ij\rangle = \sum_{i<j} \frac{1}{R_{ij}} \text{Erf}\left(\frac{\sqrt{2}R_{ij}}{\sqrt{s_i^2 + s_j^2}}\right) \\
E_{\text{Pauli}} &= \sum_{\sigma_i=\sigma_j} E(\uparrow\uparrow)_{ij} + \sum_{\sigma_i\neq\sigma_j} E(\uparrow\downarrow)_{ij}
\end{aligned} \tag{5.4}$$

in which Z_n represents the charge of the nucleus n, σ_j represents the spin of electron j, and

$$\begin{aligned}
E(\uparrow\uparrow)_{ij} &= \left(\frac{\bar{S}_{ij}^2}{1-\bar{S}_{ij}^2} + (1-\rho)\frac{\bar{S}_{ij}^2}{1+\bar{S}_{ij}^2}\right)\Delta\bar{T}_{ij} \\
E(\uparrow\downarrow)_{ij} &= \frac{\rho\bar{S}_{ij}^2}{1+\bar{S}_{ij}^2}\Delta\bar{T}_{ij},
\end{aligned} \tag{5.5}$$

where $\rho = -0.2$.

The overlap integral of electrons i and j is

$$S_{ij} = \langle i|j\rangle = \left(\frac{2s_i s_j}{s_i^2 + s_j^2}\right)^{3/2} \exp\left(-\frac{R_{ij}^2}{s_i^2 + s_j^2}\right), \tag{5.6}$$

and the change of electronic kinetic energy upon antisymmetrization of the wavefunction is

$$\begin{aligned}
\Delta T_{ij} &= \langle \Psi_{\text{Slater}} | -\frac{1}{2}\nabla_i^2 - \frac{1}{2}\nabla_j^2 | \Psi_{\text{Slater}} \rangle - \langle \Psi_{\text{Hatree}} | -\frac{1}{2}\nabla_i^2 - \frac{1}{2}\nabla_j^2 | \Psi_{\text{Hartree}} \rangle \\
&= \frac{S_{ij}^2}{1 - S_{ij}^2}(t_{ii} + t_{jj} - 2t_{ij}/S_{ij}) \\
&= \frac{S_{ij}^2}{1 - S_{ij}^2}\left[\frac{3}{2}\frac{1}{s_i^2} + \frac{3}{2}\frac{1}{s_j^2} - \frac{6}{s_i^2 + s_j^2} + \frac{4R_{ij}^2}{(s_i^2 + s_j^2)^2}\right], \qquad (5.7)
\end{aligned}$$

where

$$\begin{aligned}
\Psi_{\text{Slater}} &= \frac{1}{\sqrt{2 - 2S_{ij}^2}}(\phi_i \phi_j - \phi_j \phi_i) \qquad (5.8) \\
\Psi_{\text{Hartree}} &= \phi_i \phi_j. \qquad (5.9)
\end{aligned}$$

While S_{ij} and ΔT_{ij} are functions of s_i and R_{ij}, we define $\bar{S}_{ij}$ and $\Delta \bar{T}_{ij}$ as functions of $\bar{s}_i = a_s s_i$ and $\bar{R}_{ij} = a_R R_{ij}$, where $a_s = 0.9$ and $a_R = 1.125$. The parameters ρ, a_s, a_R are universal parameters that were adjusted to reproduce the geometries for a range of structures.(*22*)

By substituting the wave packet into the time-dependent Schrödinger equation, one could derive the Hamilton equation of motion:

$$\begin{aligned}
\dot{\vec{p}}_{\vec{R_j}} &= -\nabla_{\vec{R_j}} E,\ \dot{p}_{s_j} = -\frac{\partial E}{\partial s_j} \\
\vec{p}_{\vec{R_j}} &= m_{elec} \dot{\vec{R_j}},\ p_{s_j} = \frac{3m_e}{4}\dot{s}_j, \qquad (5.10)
\end{aligned}$$

where m_e is the mass of the electron. The nuclear motion is governed by

$$\begin{aligned}
\dot{\vec{p}}_{\vec{R_n}} &= -\nabla_{\vec{R_n}} E, \qquad (5.11) \\
\vec{p}_{\vec{R_n}} &= m_n \dot{\vec{R_n}}, \qquad (5.12)
\end{aligned}$$

where m_n is the mass of the nucleus.

This equation of motion shows the average position of the wave packet follows a classical trajectory, consistent with Ehrenfest's theorem. It extends the Ehrenfest's theorem in that the size of the wave packet also follows a classical trajectory.

The above summarizes the eFF formulation. Two assumptions were made to develop the Pauli potential in Equation 5.5: 1) the Pauli energy can be approximated by the sum of pair-wise interaction energies of elections, and 2) the kinetic energy component dominates the energy change caused by antisymmetrization of the wavefunction (the Pauli energy). The functional form comes from mixing

$$E_u = \langle \Psi_{\mathrm{Slater}} | -\frac{1}{2}\nabla_i^2 - \frac{1}{2}\nabla_j^2 | \Psi_{\mathrm{Slater}} \rangle - \langle \Psi_{\mathrm{Hatree}} | -\frac{1}{2}\nabla_i^2 - \frac{1}{2}\nabla_j^2 | \Psi_{\mathrm{Hartree}} \rangle$$

and a correlation energy

$$E_g = \langle \Psi_{\mathrm{VB}} | -\frac{1}{2}\nabla_i^2 - \frac{1}{2}\nabla_j^2 | \Psi_{\mathrm{VB}} \rangle - \langle \Psi_{\mathrm{Hatree}} | -\frac{1}{2}\nabla_i^2 - \frac{1}{2}\nabla_j^2 | \Psi_{\mathrm{Hartree}} \rangle,$$

where

$$\Psi_{\mathrm{VB}} = \frac{1}{\sqrt{2 + 2S_{ij}^2}} (\phi_i \phi_j + \phi_j \phi_i). \tag{5.13}$$

Klakow (*23*) used $E(\uparrow\uparrow) = E_u$ and $E(\uparrow\downarrow) = 0$ in kinetic-energy-based Pauli potentials. To reduce the likelihood of coalescence for both same spin and opposite spin electrons, eFF chose

$$\begin{aligned} E(\uparrow\uparrow) &= E_u - (1-\rho)E_g \\ E(\uparrow\downarrow) &= -\rho E_g. \end{aligned} \tag{5.14}$$

This is discussed in detail in (*22*) by Su.

5.3 The Gaussion Hartree Approximated Quantum Mechanics Framework

Recently, Xiao (*18*) questioned the justification of the above two assumptions in eFF Pauli potential. This led to the early development of the GHA-QM formulation of the Pauli potential in 2014. He suggested that the electron-electron and electron-nucleus Coulomb components of the energy change upon antisymmetrization are not negligible. Also, the Pauli potential considering only pair-wise interactions does not scale correctly with the number of electrons, and a scaling factor is introduced.

The total energy change upon antisymmetrization in GHA-QM is written as

$$E_{\mathrm{Pauli}} = \sum_{\sigma_i = \sigma_j} F(\uparrow\uparrow)_{ij} E_{\mathrm{Pauli}}^{\mathrm{base}}(\uparrow\uparrow)_{ij} + \sum_{\sigma_i \neq \sigma_j} F(\uparrow\downarrow)_{ij} E_{\mathrm{Pauli}}^{\mathrm{base}}(\uparrow\downarrow)_{ij}, \tag{5.15}$$

where $F(\uparrow\uparrow)_{ij}$ and $F(\uparrow\downarrow)_{ij}$ are the scaling factors,

$$E_{\mathrm{Pauli}}^{\mathrm{base}}(\uparrow\uparrow)_{ij} = \Delta T_{ij} + \Delta C_{ee,i,j} + \sum_n \Delta C_{ne,i,j} \tag{5.16}$$

and

$$E_{\mathrm{Pauli}}^{\mathrm{base}}(\uparrow\downarrow)_{ij} = -\frac{p_0 S_{ij}^{p_1} + p_2 S_{ij}^{p_3} \bar{s}_{ij}^{p_4}}{1 + p_5 S_{ij}^{p_6} + p_7 S_{ij}^{p_8} \bar{s}_{ij}^{p_9}} \tag{5.17}$$

in which ΔT_{ij} is defined the same as Equation 5.7, S_{ij} is defined the same as Equation 5.6, and

$$\bar{s}_{ij} = \frac{s_i s_j}{\sqrt{s_i^2 + s_j^2}}. \tag{5.18}$$

$E^{\text{base}}_{\text{Pauli}}(\uparrow\downarrow)_{ij}$ accounts for electron correlation and the form is inspired by the Wigner correlation functional, and $p_k,\ k = 0, ...9$ are parameters fitted to singlet H_2 bonding curve calculated using B3LYP/FSG.

$\Delta C_{ee,ij}$ and $\Delta C_{ne,ij}$ come from the electron-electron and electron-nucleus Coulomb terms in the total Hamiltonian of an electron pair

$$\hat{H} = -\sum_{k=1,2} \frac{1}{2}\nabla_k^2 + \frac{1}{r_{12}} - \sum_n \sum_{k=1,2} \frac{Z_n}{r_{nk}}. \tag{5.19}$$

Similar to the derivation of ΔT_{12} in Equation 5.7,

$$\begin{aligned} \Delta C_{ee,12} &= \langle \Psi_{\text{Slater}} | \frac{1}{r_{12}} | \Psi_{\text{Slater}} \rangle - \langle \Psi_{\text{Hatree}} | \frac{1}{r_{12}} | \Psi_{\text{Hartree}} \rangle \\ &= O_{12}(J_{12} - K_{12}/S_{12}^2), \end{aligned} \tag{5.20}$$

where

$$O_{12} = \frac{S_{12}^2}{1 - S_{12}^2}, \tag{5.21}$$

$$J_{12} = \langle 12 | \frac{1}{r_{12}} | 12 \rangle = \frac{1}{R_{12}} \text{Erf}\left(\frac{\sqrt{2} R_{12}}{\sqrt{s_1^2 + s_2^2}} \right), \tag{5.22}$$

and the exchange energy

$$K_{12} = \langle 12 | \frac{1}{r_{12}} | 21 \rangle = \sqrt{\frac{2}{\pi}} \frac{\sqrt{s_1^2 + s_2^2}}{s_1 s_2} S_{12}^2. \tag{5.23}$$

For the nucleus-electron Coulomb interaction contribution,

$$\begin{aligned} \Delta C_{ne,12} &= \langle \Psi_{\text{Slater}} | -\frac{Z_n}{r_{n1}} - \frac{Z_n}{r_{n2}} | \Psi_{\text{Slater}} \rangle - \langle \Psi_{\text{Hatree}} | -\frac{Z_n}{r_{n1}} - \frac{Z_n}{r_{n2}} | \Psi_{\text{Hartree}} \rangle \\ &= O_{12}(j_{11} + j_{22} - 2j_{12}/S_{12}^2), \end{aligned} \tag{5.24}$$

where

$$j_{kk} = \langle k | -\frac{Z_n}{r_{nk}} | k \rangle = -Z_n \frac{\text{Erf}\left(\frac{\sqrt{2}}{s_k} R_{nk} \right)}{R_{nk}} \quad (k = 1, 2) \tag{5.25}$$

$$j_{12} = \langle 1| - \frac{Z_n}{r_{n1}}|2\rangle = -\frac{Z_n}{\bar{s}_{12}} \frac{\mathrm{Erf}\left(\sqrt{\frac{R_{n1}^2}{s_1^2} + \frac{R_{n2}^2}{s_2^2} - \frac{R_{12}^2}{s_1^2+s_2^2}}\right)}{\sqrt{\frac{R_{n1}^2}{s_1^2} + \frac{R_{n2}^2}{s_2^2} - \frac{R_{12}^2}{s_1^2+s_2^2}}} S_{12}. \tag{5.26}$$

The full forms of the scaling factors in Equation 5.15 are

$$F(\uparrow\uparrow)_{ij} = F(\uparrow\uparrow)_{ij,\mathrm{sym}}(\Delta T, \Delta C_{ee}, \Delta C_{ne}) F(\uparrow\uparrow)_{ij,\mathrm{asym}}(\Delta T, \Delta C_{ee}, \Delta C_{ne}) \tag{5.27}$$

and

$$F(\uparrow\downarrow)_{ij} = F(\uparrow\downarrow)_{ij,\mathrm{sym}} F(\uparrow\downarrow)_{ij,\mathrm{asym}}. \tag{5.28}$$

Define

$$\overline{\sum} S_{ij} = \frac{1}{2}\left(\sum_k S_{ik} + \sum_{k\prime} S_{ik\prime}\right) \tag{5.29}$$

$$\overline{\sum} S_{ij}^2 = \frac{1}{2}\left(\sum_k S_{ik}^2 + \sum_{k\prime} S_{ik\prime}^2\right). \tag{5.30}$$

For same spin electrons that each has exactly the same environment, the scaling factors for ΔT and ΔC_{ne} can be derived based on D_{3h} quartet H_3 or T_d quintet H_4 to give

$$F(\uparrow\uparrow)_{ij,\mathrm{sym}}(\Delta T, \Delta C_{ne}) = \frac{1 + S_{ij}}{1 + \overline{\sum} S}. \tag{5.31}$$

Therefore, we separate the scaling factor into two terms multiplying each other, F_{sym} and F_{asym}, which account for the "symmetric" environment and "asymmetric" environment, respectively.

The forms of the other components of the same spin electron scaling factors are

$$\begin{aligned} F(\uparrow\uparrow)_{ij,\mathrm{sym}}(\Delta C_{ee}) &= p_{se0}\left(\frac{p_{se1} + S_{ij}}{p_{se1} + \overline{\sum} S_{ij}} + \frac{\overline{\sum} S_{ij} - S_{ij}}{p_{se1} S_{ij} + \overline{\sum} S_{ij}}\right)^{p_{se2}} \\ &+ (1 - p_{se0})\left(\frac{p_{se3} + S_{ij}^2}{p_{se3} + \overline{\sum} S_{ij}^2} + \frac{\overline{\sum} S_{ij}^2 - S_{ij}^2}{p_{se3} S_{ij}^2 + \overline{\sum} S_{ij}^2}\right)^{p_{se4}}, \end{aligned} \tag{5.32}$$

where the parameters $p_{sek}, k = 0, ..., 4$ are fitted with the exact quantum mechanics [unrestricted Hartree Fock (UHF)/FSG] results of D_{3h} quartet H_3 and T_d quintet H_4 symmetric stretching, and

$$\begin{aligned} F(\uparrow\uparrow)_{ij,\mathrm{asym}}(\Delta T, \Delta C_{ee}, \Delta C_{ne}) &= 1 + p_{a0}\left(S_{ij}\overline{\sum} S_{ij} - \overline{\sum} S_{ij}^2\right) \\ &+ p_{a1}\left(S_{ij}\overline{\sum} S_{ij} - \overline{\sum} S_{ij}^2\right)^2 \\ &+ p_{a2}\left(S_{ij}\overline{\sum} S_{ij} - \overline{\sum} S_{ij}^2\right)^3, \end{aligned} \tag{5.33}$$

where $p_{ak}, k = 0, 1, 2$ are parameters whose values depend on whether it is for ΔT, ΔC_{ee} or ΔC_{ne}, and fitted with D_{3h} to C_{2v} transition of quartet H_3 and T_d to C_{3v} transition of quintet H_4.

For opposite spin scaling factors, the forms are

$$F(\uparrow\downarrow)_{ij,\mathrm{sym}} = p_{os0}\left(\frac{1 + p_{os1}S_{ij}}{1 + p_{os1}\overline{\sum}S_{ij}}\right)^{p_{os2}} + (1 - p_{os0})\left(\frac{1 + p_{os3}S_{ij}^2}{1 + p_{os3}\overline{\sum}S_{ij}^2}\right)^{p_{os4}}, \tag{5.34}$$

where the parameters $p_{osk}, k = 0, \ldots, 4$ are fitted with $D_{\infty h}$ doublet H_3 and D_{4h} singlet H_4 symmetric stretching, and

$$\begin{aligned} F(\uparrow\downarrow)_{ij,\mathrm{asym}} &= 1 + p_{oa0}\left(\frac{S_{ij}\overline{\sum}S_{ij} - \overline{\sum}S_{ij}^2}{S_{ij}\overline{\sum}S_{ij} + \overline{\sum}S_{ij}^2}\right) \\ &+ p_{oa1}\left(\frac{S_{ij}\overline{\sum}S_{ij} - \overline{\sum}S_{ij}^2}{S_{ij}\overline{\sum}S_{ij} + \overline{\sum}S_{ij}^2}\right)^2 \\ &+ p_{oa2}\left(\frac{S_{ij}\overline{\sum}S_{ij} - \overline{\sum}S_{ij}^2}{S_{ij}\overline{\sum}S_{ij} + \overline{\sum}S_{ij}^2}\right)^3, \end{aligned} \tag{5.35}$$

where the parameters $p_{oak}, k = 0, 1, 2$ are fitted with $D_{\infty h}$ doublet H_3 and D_{4h} singlet H_4 asymmetric stretching.

In summary, the total Pauli potential for same spin electron pairs is

$$\begin{aligned} E_{\mathrm{Pauli}}(\uparrow\uparrow)_{ij} &= F(\uparrow\uparrow)_{ij,\mathrm{sym}}(\Delta T)F(\uparrow\uparrow)_{ij,\mathrm{asym}}(\Delta T)\Delta T_{ij} \\ &+ F(\uparrow\uparrow)_{ij,\mathrm{sym}}(\Delta C_{ee})F(\uparrow\uparrow)_{ij,\mathrm{asym}}(\Delta C_{ee})\Delta C_{ee,ij} \\ &+ \sum_n F(\uparrow\uparrow)_{ij,\mathrm{sym}}(\Delta C_{ne})F(\uparrow\uparrow)_{ij,\mathrm{asym}}(\Delta C_{ne})\Delta C_{ne,ij} \end{aligned} \tag{5.36}$$

and the total Pauli potential for opposite spin pairs is

$$E_{\mathrm{Pauli}}(\uparrow\downarrow)_{ij} = F(\uparrow\downarrow)_{ij,\mathrm{sym}}F(\uparrow\downarrow)_{ij,\mathrm{asym}}E_{\mathrm{Pauli}}^{\mathrm{base}}(\uparrow\downarrow)_{ij}. \tag{5.37}$$

The above has been discussed in more detail in (*18*).

5.4 Improvements on GHA-QM

The GHA-QM framework in the previous section produces QM quality bonding curve for singlet and triplet H_2, and gives good results for symmetric and asymmetric stretching of H_3 and H_4 molecules in mild conditions. However, when we tested the system more substantially and tried to simulate hydrogen systems with more electrons in high density, we found certain problems. This section discusses several strategies that were taken to fix these problems.

New Opposite Spin Pauli Potential

We found that the size of the electron in the center of linear H_3 sometimes goes to zero. This is because in Equation 5.17 where p_4 is negative and all other parameters are positive, one electron having zero size ($\bar{s}$ going to zero) means $E^{\text{base}}_{\text{Pauli}}(\uparrow\downarrow)$ becomes negative infinite and energetically favored. To avoid this problem, we introduced a positive shift p_{10} of electron size in Equation 5.17. The new form is

$$E^{\text{base}}_{\text{Pauli}}(\uparrow\downarrow)_{ij} = -\frac{p_0 S_{ij}^{p_1} + p_2 S_{ij}^{p_3} (\bar{s}_{ij} + p_{10})^{p_4}}{1 + p_5 S_{ij}^{p_6} + p_7 S_{ij}^{p_8} \bar{s}_{ij}^{p_9}}. \tag{5.38}$$

We refitted the parameters against the exact QM results of singlet H_2 bonding curve.(*24*) The new parameters are $p_0 = 0.439387, p_1 = 2.914263, p_2 = 8.180823, p_3 = 6.100496, p_4 = -18.088005, p_5 = 0.698750, p_6 = 3.305666, p_7 = 2.031224, p_8 = 7.359878, p_9 = 5.552337, p_{10} = 8.345865$. The resulting GHA-QM H_2 bonding curve is nearly exact in the bonding regions, and performs better than B3LYP/6-311++G** at intermediate bond lengths (Figure 5.1). Because the base energy of the opposite spin Pauli potential has changed, we refitted the parameters in the opposite spin scaling factors too. We used genetic algorithm to train the parameters for both symmetric and asymmetric scaling factors at the same time. The best parameters are $p_{os0} = 0.968658, p_{os1} = 1.021811, p_{os2} = 2.720616, p_{os3} = 15.859202, p_{os4} = 20.201212$ and $p_{oa0} = 2.376509, p_{oa1} = 18.590298, p_{oa2} = 13.245066$. These parameters give much improved energy for H_3 reaction path (Figure 5.2), and overall

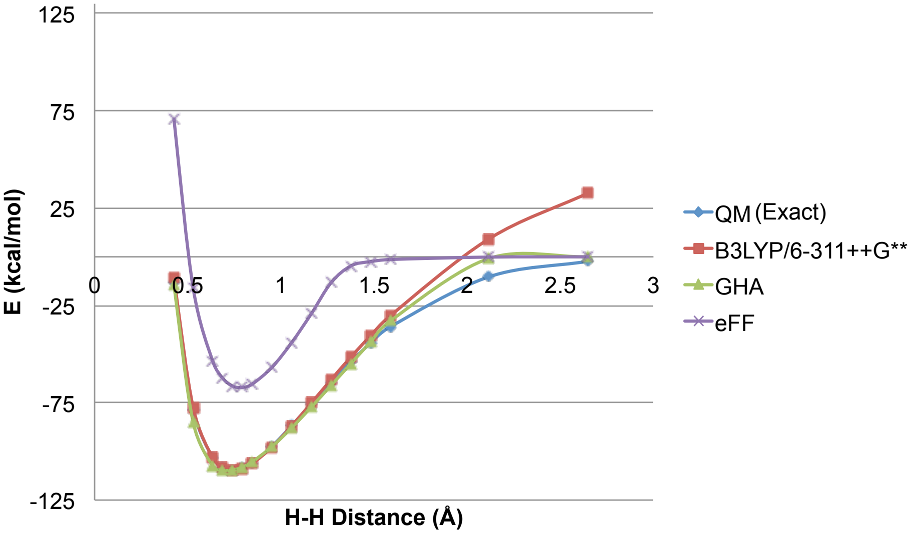


Figure 5.1: Performance of GHA-QM on single H_2.

a more accurate H_3 potential energy surface (Figure 5.4) than eFF (Figure 5.5) comparing to *ab initio* H_3 potential energy surface (Figure 5.5).(*25*)

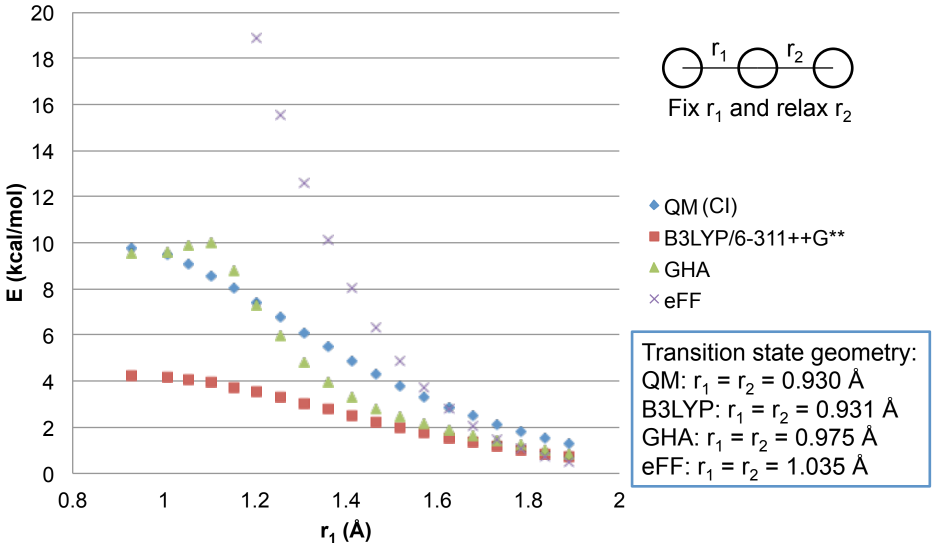


Figure 5.2: Performance of GHA-QM on H_3 reaction path. The QM configuration interaction (CI) data is from (*26*).

Preventing Same Spin Electron Coalescence

When we tried to simulate the equation of state (EOS) of a dense H_2 liquid (Wigner radius $r_s = 2.2$ bohr), we find the same spin electrons are energetically favorable to diffuse and cluster at some configurations. The following paragraphs are devoted to attempts at solving this problem.

A New Form of ΔT We'd like to shift the numerator in Equation 5.7 by a small amount dd so that ΔT goes to infinity when $S \to 1$. For convenience, the subscripts of $\Delta T, S, O$ are omitted.

The new form is

$$\begin{aligned}\Delta T &= \frac{S^2(t_{11}+t_{22}-2t_{12}/S)+dd}{1-S^2} \\ &= \frac{S^2\left[\frac{3}{2}\frac{1}{s_1^2}+\frac{3}{2}\frac{1}{s_2^2}-\frac{6}{s_1^2+s_2^2}+\frac{4R_{12}^2}{(s_1^2+s_2^2)^2}\right]+dd}{1-S^2} \\ &= O\cdot T+\frac{dd}{1-S^2}, \end{aligned} \tag{5.39}$$

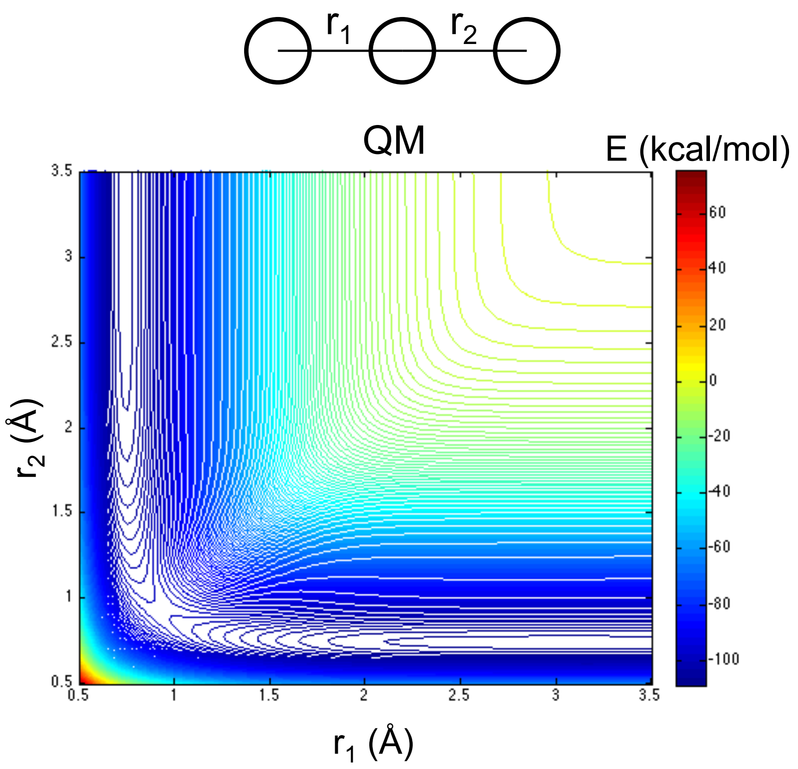


Figure 5.3: Analytical H_3 potential energy surface. The data is from (*25*).

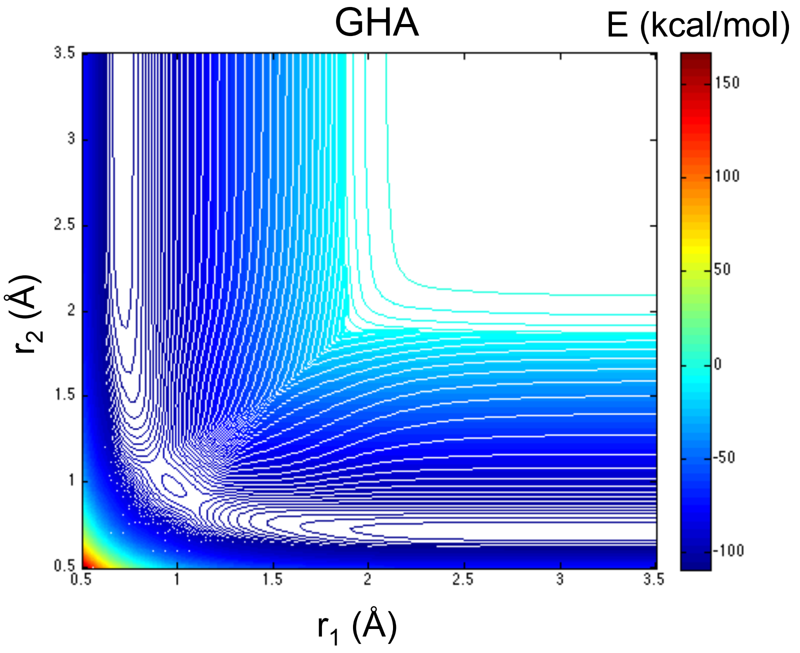


Figure 5.4: GHA-QM H_3 potential energy surface. r_1 and r_2 are defined in Figure 5.3.

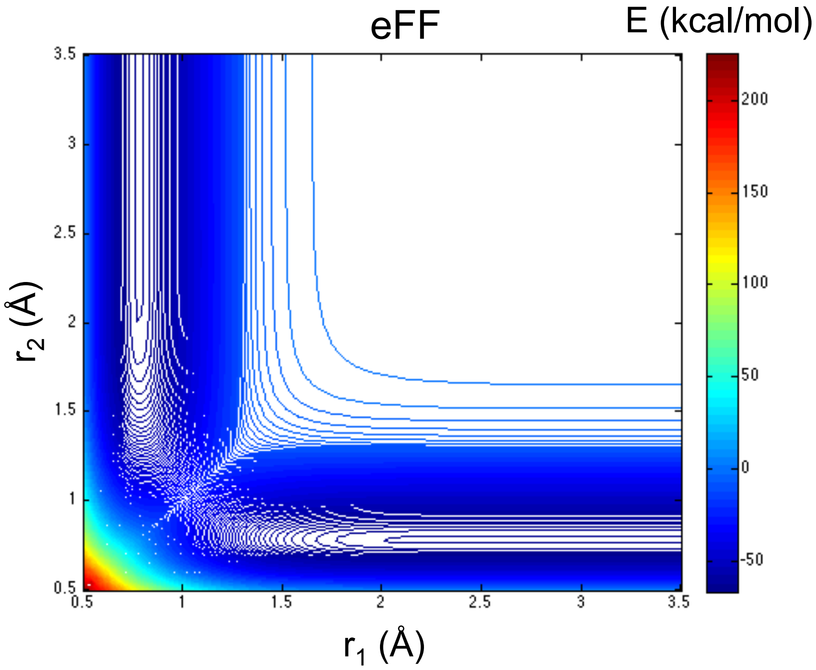


Figure 5.5: eFF H_3 potential energy surface. r_1 and r_2 are defined in Figure 5.3.

where

$$T \equiv \frac{3}{2}\frac{1}{s_1^2} + \frac{3}{2}\frac{1}{s_2^2} - \frac{6}{s_1^2 + s_2^2} + \frac{4R_{12}^2}{(s_1^2 + s_2^2)^2}.$$

We chose $dd = 1.0 \times 10^{-4}$.

To avoid numerical instability when $S \to 1$, $s_1 \to s_2$ and $R_{12} \to 0$, we did series expansion of ΔT at this condition (Appendix D). In its practical implementation, when $1 - S^2 \leq 0.001$, instead of the form in Equation 5.39, we take

$$\Delta T = \left(\frac{1}{s_1 s_2} - \frac{R_{12}^2}{2s_1^2 s_2^2} + \frac{R_{12}^4}{12s_1^3 s_2^3}\right) + dd\left(\frac{s_1 s_2}{R_{12}^2} + \frac{1}{2} + \frac{R_{12}^2}{12s_1 s_2} - \frac{R_{12}^6}{720s_1^3 s_2^3}\right). \quad (5.40)$$

Although this scheme can prevent the same spin electrons to have the same coordinates, it does not prevent same spin electrons of similar sizes to become unphysically close (overlap $S \to 1$). As can be seen in Figure 5.6, the energy penalty from the new form of ΔT is not big enough for most configurations the dynamics simulation can reach. Nevertheless, we kept this form because it does prevent the same spin electrons to have the same coordinates and enforce the Pauli exclusion principle.

New Same Spin Scaling Factors The form of Equation 5.33 may cause serious issues during simulation because it is unbounded. In addition, the sign of the

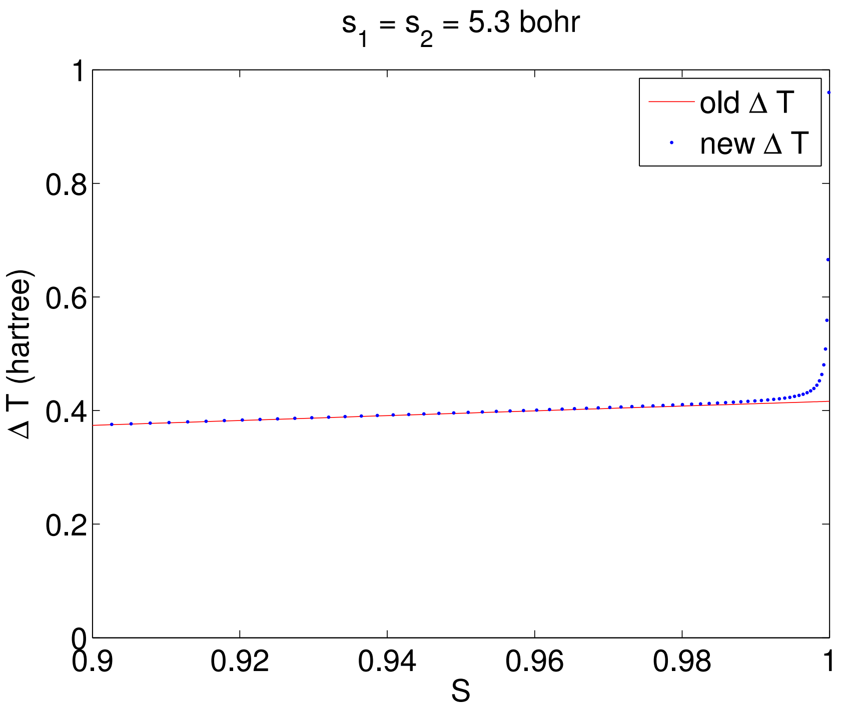


Figure 5.6: Comparison of ΔT values from the old and new forms for diffused electrons of the same size and spin.

function may change due to the odd powers and the possibility for $(S_{ij}\overline{\sum}S_{ij} - \overline{\sum}S_{ij}^2)$ to become negative in some cases. To make the function bounded, we adopt the following form instead:

$$F(\uparrow\uparrow)_{ij,\mathrm{asym}}(\Delta T, \Delta C_{ee}, \Delta C_{ne}) = 1 + p_{a0}D_{ij} + p_{a1}D_{ij}^2 + p_{a2}D_{ij}^3, \tag{5.41}$$

where $D_{ij} = \left(\frac{S_{ij}\overline{\sum}S_{ij} - \overline{\sum}S_{ij}^2}{S_{ij}\overline{\sum}S_{ij} + \overline{\sum}S_{ij}^2 + d}\right)$. The small constant $d = 1.0 \times 10^{-6}$ is to prevent the denominator from going to zero when all the overlap $S \to 0$. In fact, this shift has been added to the opposite spin asymmetric scaling factor too so that now same spin and opposite spin asymmetric scaling factors adopt the same form, with different parameters. The parameters for same spin asymmetric scaling factors are summarized in Table 5.1.

For the fitting to cover the whole range of S values from 0 to 1, we fitted against C_{2v} quartet H_3, with one H-H distance fixed at 0.1 bohr and 1.7 bohr. The size of the electrons were kept to 1.5 bohr. Figures 5.7 shows that the whole range of S has been covered, especially for large S values that were not considered in previous fitting.

	ΔT	ΔC_{ee}	ΔC_{ne}
p_{a0}	1.31546298328185	0.727335080043498	0.468486643682531
p_{a1}	3.39853157886649	1.311794439120800	13.20277753109270
p_{a2}	2.33306859558464	0.584459359077299	12.98429088741020

Table 5.1: New parameters for same spin asymmetric scaling factors $F(\uparrow\uparrow)_{ij,\text{asym}}(\Delta T, \Delta C_{ee}, \Delta C_{ne})$.

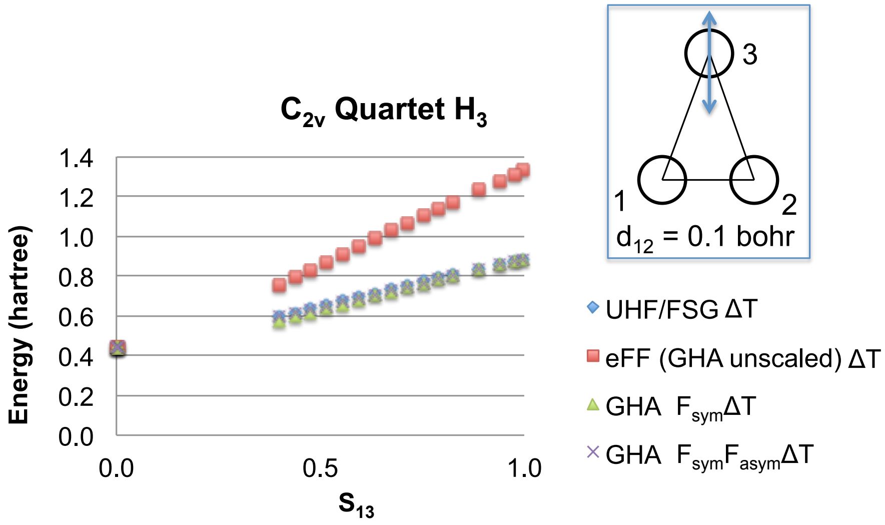


Figure 5.7: Performance of GHA-QM for C_{2v} quartet H_3 with one H-H distance fixed at 0.1 bohr.

Figure 5.8 shows the GHA-QM same spin symmetric scaling factor does not work for C_{2v} quartet H_3 when one H-H distance is fixed at 1.7 bohr and the other H-H distances are much smaller than 1.7 bohr. This suggests that the symmetric scaling factor is not likely able to deal with linear H_3. Since the same spin symmetric scaling factor for ΔT is exact for D_{3h} H_3, we hypothesized that the scaling factors are highly symmetry-dependent and the $F(\uparrow\uparrow)_{\text{sym}}$ derived from D_{3h} H_3 cannot be universally applied to all molecules with other symmetry. Interestingly, eFF seems to go to the correct limit as the system approaches being linear. To further explore this issue, we developed a scaling factor $F(\uparrow\uparrow)_{\text{linear}}$ to replace $F(\uparrow\uparrow)_{\text{sym}}$ in linear molecules:

$$F(\uparrow\uparrow)_{ij,\text{linear}} = 1 + p_{l0}v + p_{l1}v^2 + p_{l2}v^3, \tag{5.42}$$

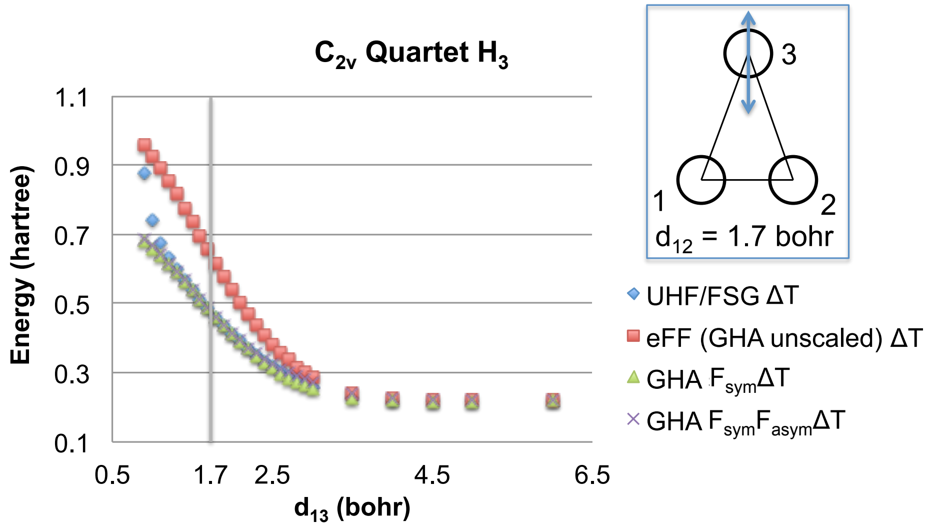


Figure 5.8: Performance of GHA-QM for C_{2v} quartet H_3 with one H-H distance fixed at 1.7 bohr.

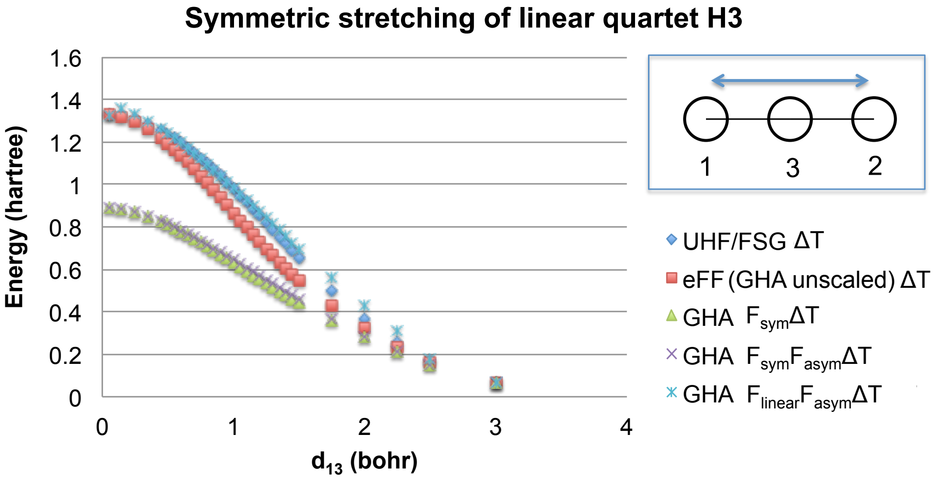


Figure 5.9: Comparison of the effect of different scaling factors in the performance of symmetric stretching of linear quartet H_3.

where

$$v = \left(\frac{(S_{ij} \sum S_{ij} - \sum S_{ij}^2)^2}{(p_{l3} S_{ij} \sum S_{ij} - \sum S_{ij}^2)^2 + d_l} \right), \tag{5.43}$$

we chose $d_l = 1.0 \times 10^{-7}$ and fitted the parameters $p_{l0} = 0.0098195829828946$, $p_{l1} = 0.0758036465880992$, $p_{l2} = 0.440549994711866$, $p_{l3} = 0.999965221776706$.

Intuitively, when the H-H distances become very small, linear H_3 and D_{3h} H_3 energy should go to the same limit. However, this is not the case as seen from Figure 5.9.

This confirms that the scaling factors should be symmetry-dependent or geometry-dependent. However, in practice it is likely impossible to implement. Since our goal is to simulate condensed matters, there will be too many electrons in the system for the program to classify the configuration to determine suitable scaling factors.

Removing the Same Spin Scaling Factors The analysis in the previous subsection suggests it might be a good idea to remove the same spin scaling factors completely (i.e. make them take the value of 1). In addition, minimization of the dense H_2 liquid (r_s = 1.76 bohr or denser) shows that same spin scaling factors for ΔT and ΔC_{ne} can scale these positive energy contributions down to $\sim \frac{1}{10}$ times the base values while maintaining the negative energy contribution from ΔC_{ee} similar to the base value. This energetically favors the same spin electrons to cluster, even when we keep the electron sizes fixed at 1.55 bohr (the electron size in H_2 optimized by GHA-QM). The different behaviors of the scaling factors for different energy components may be due to the empirical nature of some of the functional forms of the scaling factors, and may also be due to the symmetry issue discussed in the previous subsection.

To see how much removing same spin scaling factors could affect molecular interactions, we calculated the H_2-H_2 association curve with GHA-QM without same spin scaling factors. The results look reasonable as they are close to QM results (*27*) in Figure 5.10. They are also an improvement on eFF.

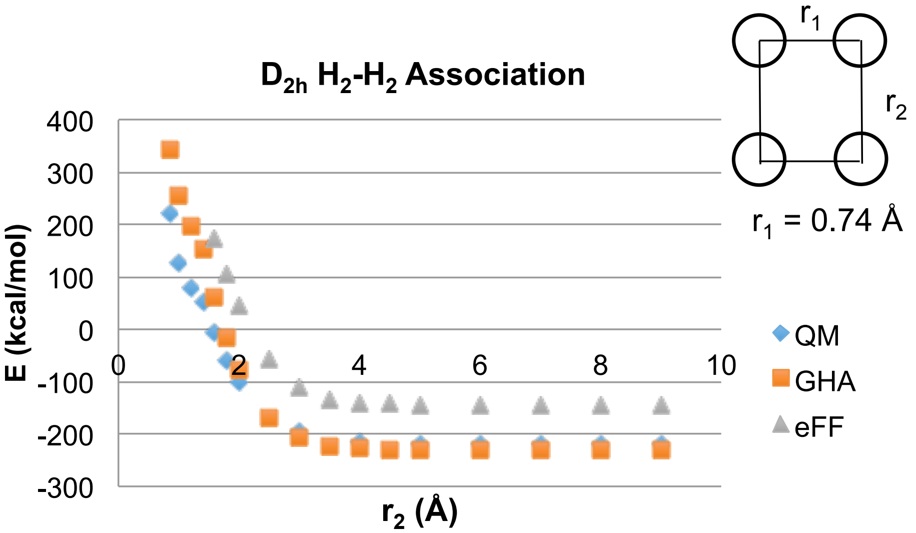


Figure 5.10: D_{2h} H_2-H_2 association without same spin scaling factors. The energy values are $E(H_4) - 4E(H)$. The QM data is from (*27*).

Therefore, with the same spin scaling factors turned off, we went on to calculate the liquid H_2 EOS and compared it to QM results. A cubic box of 108 H_2 molecules has been used with minimum image on each side of the box as the periodic boundary condition. We took $m_e = m_H$ so that we can use a relatively large time step 0.02 femtosecond. A total of 2 ps dynamics was carried out for each temperature, and the first 0.5 ps was discarded when calculating the average pressure. During the dynamics, for each time step we optimized the electron sizes based on the x, y, z coordinates of the electrons and nuclei. For simplicity, we approximated the optimization procedure by carrying out only the first step of the Newton-Raphson method and considering the second derivative of only the electronic kinetic energy. Figure 5.11 shows both GHA-QM and eFF are quite close to QM results, with both eFF and GHA-QM having a tendency to underestimate the pressure at low density. At high density, GHA-QM has a tendency to underestimate the pressure, while eFF tends to overestimate the pressure. In terms of the energy-volume relationship, GHA-QM is clearly an improvement upon eFF (Figure 5.12).

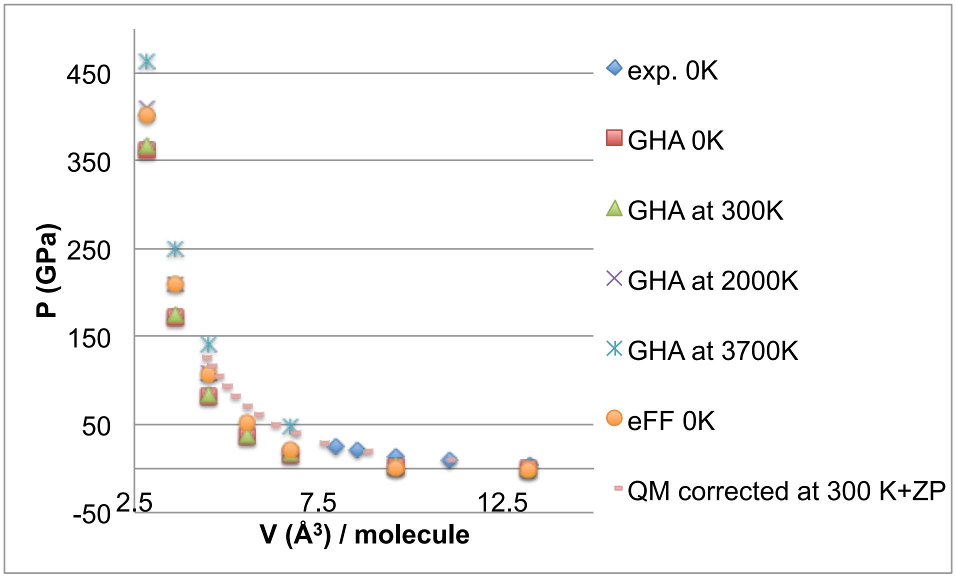


Figure 5.11: Pressure-volume diagram of H_2 liquid. The experimental values and QM extrapolation are from (*28*).

5.5 Conclusions and Future Work

We have overcome a few technical obstacles including but not limited to those discussed in the above section to enable large-scale simulations of warm dense hydrogen EOS using GHA-QM. Our next step is to simulate the shock hugoniot. We would like to see whether the new GHA-QM can correctly predict the pressure

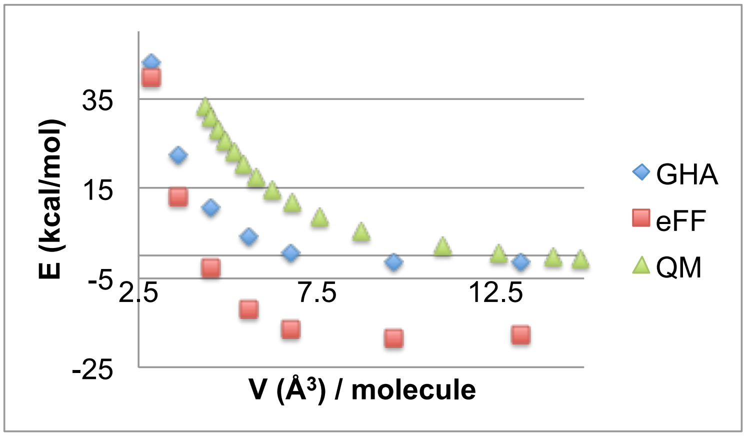


Figure 5.12: Energy-volume diagram of H_2 liquid. The QM data is generated by PBE-D3 calculation and provided by Saber Naserifar (unpublished).

and temperature where the hydrogen insulator-to-metal transition occurs. We may need to include the angular momentum projected effective core potential (AMPERE) extension,(*18*) which is not discussed here, to account for the cusp condition at the nuclei for more accurate predictions. When we move up in the periodic table, AMPERE will be necessary to obtain correct bond energy and nodal structures. In the future, we hope GHA-QM along with AMPERE will be able to simulate processes such as silicon etching that no existing method could simulate well.

References

(*1*) Landau, L., (1932). On the theory of transfer of energy at collisions II. *Phys. Z. Sowjetunion 2*, 7.

(*2*) Zener, C., In *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 1932; Vol. 137, pp 696–702.

(*3*) McLachlan, A., (1964). A variational solution of the time-dependent Schrodinger equation. *Molecular Physics 8*, 39–44.

(*4*) Micha, D. A., (1983). A self-consistent eikonal treatment of electronic transitions in molecular collisions. *The Journal of Chemical Physics 78*, 7138–7145.

(*5*) Sawada, S.-I., Nitzan, A., and Metiu, H., (1985). Mean-trajectory approximation for charge-and energy-transfer processes at surfaces. *Physical Review B 32*, 851.

(*6*) Kirson, Z., Gerber, R., Nitzan, A., and Ratner, M. A., (1984). Dynamics of metal electron excitation in atom-surface collisions: A quantum wave packet approach. *Surface Science 137*, 527–550.

(*7*) Tully, J. C., and Preston, R. K., (1971). Trajectory surface hopping approach to nonadiabatic molecular collisions: the reaction of H+ with D2. *The Journal of Chemical Physics 55*, 562–572.

(*8*) Tully, J. C., (1990). Molecular dynamics with electronic transitions. *The Journal of Chemical Physics 93*, 1061–1071.

(*9*) Li, X., Tully, J. C., Schlegel, H. B., and Frisch, M. J., (2005). Ab initio Ehrenfest dynamics. *The Journal of chemical physics 123*, 084106.

(*10*) Isborn, C. M., Li, X., and Tully, J. C., (2007). Time-dependent density functional theory Ehrenfest dynamics: collisions between atomic oxygen and graphite clusters. *The Journal of chemical physics 126*, 134307–134307.

(*11*) Su, J. T., and Goddard III, W. A., (2007). Excited electron dynamics modeling of warm dense matter. *Physical review letters 99*, 185003.

(*12*) Takatsuka, K., Yonehara, T., Hanasaki, K., and Arasaki, Y., *Chemical Theory beyond the Born-Oppenheimer Paradigm: Nonadiabatic Electronic and Nuclear Dynamics in Chemical Reactions*; World Scientific: 2014.

(*13*) Kim, H., Su, J. T., and Goddard, W. A., (2011). High-temperature high-pressure phases of lithium from electron force field (eFF) quantum electron dynamics simulations. *Proceedings of the National Academy of Sciences 108*, 15101–15105.

(*14*) Theofanis, P. L., Jaramillo-Botero, A., Goddard III, W. A., Mattsson, T. R., and Thompson, A. P., (2012). Electron dynamics of shocked polyethylene crystal. *Physical Review B 85*, 094109.

(*15*) Su, J. T., and Goddard, W. A., (2009). Mechanisms of Auger-induced chemistry derived from wave packet dynamics. *Proceedings of the National Academy of Sciences 106*, 1001–1005.

(*16*) Chenard-Lemire, C., Lewis, L., and Meunier, M., (2012). Laser-induced Coulomb explosion in C and Si nanoclusters: The determining role of pulse duration. *Applied Surface Science 258*, 9404–9407.

(*17*) Theofanis, P. L., Jaramillo-Botero, A., Goddard III, W. A., and Xiao, H., (2012). Nonadiabatic study of dynamic electronic effects during brittle fracture of silicon. *Physical review letters 108*, 045501.

(*18*) Xiao, H., First principles based multiparadigm modeling of electronic structures and dynamics., Ph.D. Thesis, California Institute of Technology, 2015.

(*19*) Feldmeier, H., and Schnack, J., (2000). Molecular dynamics for fermions. *Reviews of Modern Physics 72*, 655.

(*20*) Heller, E. J., (1975). Time-dependent approach to semiclassical dynamics. *The Journal of Chemical Physics 62*, 1544–1555.

(*21*) Frost, A. A., (1967). Floating spherical Gaussian orbital model of molecular structure. I. Computational procedure. LiH as an example. *The Journal of Chemical Physics 47*, 3707–3713.

(*22*) Su, J. T.-l., An electron force field for simulating large scale excited electron dynamics., Ph.D. Thesis, California Institute of Technology, 2007.

(*23*) Klakow, D., Toepffer, C., and Reinhard, P.-G., (1994). Semiclassical molecular dynamics for strongly coupled Coulomb systems. *The Journal of chemical physics 101*, 10766–10774.

(*24*) Kolos, W., and Wolniewicz, L., (1968). Improved Theoretical Ground-State Energy of the Hydrogen Molecule. *The Journal of Chemical Physics 49*, 404–410.

(*25*) Boothroyd, A. I., Keogh, W. J., Martin, P. G., and Peterson, M. R., (1996). A refined H3 potential energy surface. *The Journal of chemical physics 104*, 7139–7152.

(*26*) Liu, B., (1973). Ab initio potential energy surface for linear H3. *The Journal of Chemical Physics 58*, 1925–1937.

(*27*) Boothroyd, A., Martin, P., Keogh, W., Peterson, M., et al. (2002). An accurate analytic H4 potential energy surface. *Journal of Chemical Physics 116*, 666–689.

(*28*) Hemley, R., Mao, H., Finger, L., Jephcoat, A., Hazen, R., and Zha, C., (1990). Equation of state of solid hydrogen and deuterium from single-crystal x-ray diffraction to 26.5 GPa. *Physical Review B 42*, 6458.

Appendix A

Additional Computational Details for the Validation and Application of ActiveGEnSeMBLE: Preparing Input Files, Ligand Docking, and Molecular Dynamics Simulations

A.1 Pre-processing of crystal structure templates for structure prediction

For each validation case, hβ_2AR and hM2, we determined the range of TM regions by taking a consensus of the helix assignment in the Protein Data Bank (PDB) of its active state structure and inactive structure, taken from the Orientations of Proteins in Membranes (OPM) database.(*1*) We then cropped out the TM regions and added missing atoms using tleap in AmberTools1.4.(*2*) To have meaningful comparisons of the energetics of structures sampled from inactive state crystal structure and from active state crystal structure, we minimized all these structures (TM regions only) using MPSim (*3*) with a convergence criterion of root mean square force (RMS force) = 0.25 kcal mol^{-1} Å^{-1} before proceeding to the conformational sampling steps.

A.2 Constructing the hybrid templates

For Validation Method 2.1

We aligned the active state crystal structure to the inactive-state crystal structure using Visual Molecular Dynamics (VMD).(*4*) We considered only the backbone atoms in TM1-5 and TM7 in the alignment. Then we replaced the TM6 in the inactive-state crystal structure with the TM6 in the active-state crystal structure.

For Validation Method 3.3

We obtained an active state candidate, denoted S1, from coarse SuperBiHelix sampling of the inactive-state crystal structure. Then we aligned the active-state crystal structure to S1 using VMD. We considered only the backbone atoms in the alignment. Last, we replaced the TM6 in S1 with the TM6 in the active-state crystal structure.

For application to hSSTR5

We obtained a homology model of the target protein using an inactive-state crystal structure as the template, and did coarse SuperBiHelix to obtain an active-state candidate T1. We also obtained the homology model of the target protein using an active-state crystal structure as the template, and name this model T2. Then we aligned T2 to T1 using VMD. We considered only the backbone atoms in the alignment. At last, we replaced the TM6 in T1 with the TM6 in T2. The x,y coordinates of the hydrophobic center (HPC) of the hybrid template remain to be that of T1.

A.3 Energy scoring function in structure prediction

We considered 4 ways to compare the energies: a) the total energy of the charged protein (CTotal), b) the interhelical energy of the charged protein (CInterH), which neglects the intrahelical energy of each chain, c) the total energy (NTotal), and d) the interhelical energy (NInterH) of the neutralized protein. We find that the isolated net charges on Asp, Glu, Lys, and Arg, particularly on the external surfaces of the protein, can cause what we consider to be artifacts in the energetics. Thus we neutralize these charged residues by adding or subtracting a proton for surface residues and transferring a proton within each salt bridge. This leads to two sets of energy: NTotal and NInterH.

From many previous studies we found that the most reliable scoring criterion for identifying the best structures is to combine these four criteria. Thus whenever energy ranking was performed in structure prediction procedures, we used energy “E_{CNti}”, which is the average energy of CTotal, CInterH, NTotal, and NInterH. This averaging method puts more weight on interhelical energies than on intrahelical energies. We have validated that for the known x-ray structures, and this procedure correctly identifies the known x-ray rotation angles as the lowest-energy structures.(*5*)

A.4 Ligand docking

Our strategy for docking (GenDock) is to select a diverse set of low energy ligand conformations for each of which we sample a complete set of poses.

Scanning regions in protein for docking

The first step of GenDock (*6*, *7*) is DarwinDock, which modifies the protein structure to replace the six types of hydrophobic residues by alanine, and then samples the complete set of poses for regions that could potentially bind a ligand. To do this sampling, the potential binding region is filled by SphGen with "spheres" having 2 Å overlaps with each other and the spheres classified into "boxes" of 10 Å sides.

Boxes containing 75 or more spheres were kept. For docking purpose, we have discarded all spheres except for those that are in the extracellular half of the GPCR TMDs and are not potentially in contact with the membrane lipids (i.e. are in the interior of the GPCR helix bundle).

GenDock

For each ligand conformation and for the “spheres” selected in the previous step, we generated 200 000 poses without energy evaluation aiming at providing a complete set of poses. The poses were clustered into ~8000-9000 Voronoi families based on RMSD and the binding energy of the family head evaluated. Then for the top 10% of families, we evaluated the energy for all children. Then we selected the top 50 based on each three energy scores: polar energy, hydrophobic energy, and total energy. Then for these 150, we dealanized (mutating alanine back to the original hydrophobic residues) and optimized the side chains using SCREAM. Then the protein-ligand complexes (poses) were subject to minimization for 50 steps.

Docking procedure of antagonists and agonists to hSSTR5

For each ligand, we first did a conformational search using MacroModel 9.7 (*8*) in Maestro 9.1. (*9*) The conformational search was a torsional sampling of the rotatable bonds that could cause large conformational changes using a Monte Carlo Multiple Minimum (MCMM) method (*10*, *11*) with the force field OPLS2005.(11) The energy window for the generated structure to be kept was set to be 10.04 kcal mol^{-1}. Then we clustered the resulting conformations with an RMSD cutoff of 2 Å.

The initial structure of M59 was built based on the x-ray crystallographic structure of a molecule, M48, which has the same benzoxazole piperidine scaffold.(*12*) To reduce the torsional sampling space we replaced the ethoxy groups in M59 with methoxy groups for the sampling of conformations. This modified molecule is labeled “M59m”. The conformational search involved six rotatable bonds other than those in the piperidine ring. Subsequently, another clustering was performed to identify 12 distinct ligand conformations expected to represent the entire set of 705 conformations generated from the previous step. This clustering criterion is RMSD of 0.5 Å. Then we added the terminal methyl groups back to M59m, rotated the O-C bond in the ethoxy groups and generated 5 possible M59 conformations from each M59m conformation. Similarly, we modified the M59 structures to obtain structures for the other antagonists.

For L-817,818, we sampled 6 rotatable bonds. For F21, the –CH_2-CH_2-Ph group was first replaced by a methyl group. This modified molecule is labeled F21m. For F21m, 7 rotatable bonds were sampled. For L-817,818 and F21m, the clustering resulted in 18 and 8 distinct conformations, respectively. Then each F21m conformation was modified into F21 by adding back the –CH_2-Ph group with each of the 9 possible conformations.

The ligand conformations were then docked to each candidate protein structure using our standard docking strategy, GenDock, described above. The charge distribution used in docking was obtained by the Mulliken population analysis using B3LYP/6-311G** in Jaguar 7.6.(*13*)

For each antagonist conformation and for each protein structure to be docked, we generated 200,000 poses without energy evaluation aiming at providing a complete set of poses. The poses were clustered into ~7300 Voronoi families based on RMSD and the binding energy of the family head evaluated. Then for the top 10% of families we evaluated the energy for all children. Then we selected the top 50 based on each three energy scores: polar energy, hydrophobic energy, and total energy. Then for these 150 we dealanized (mutating Ala back to the original hydrophobic residues) and optimized the side chains using SCREAM. Then the protein-ligand complexes (poses) were subject to minimization. Then a simulated annealing was performed on the lowest-energy 15 complexes before another minimization was done. All final poses were scored together by unified-cavity (UCav) energy. The UCav energy of a particular pose is defined as the binding energy of the ligand of this pose and the union of the binding pocket (cavity) of all poses. A "cavity" is defined as residues within 5 Å of the ligand of a particular pose.

For each agonist and each protein structure to be docked, we collected 1000 lowest unified-cavity (UCav) energy complexes. Then for each ligand, we collected 15 lowest UCav energy complexes and matched ligands at different positions in these complexes into each of the protein conformations in these complexes. We did a simulated annealing on the resulting complexes' ligands and residues within 10 Å from the ligands. In the end, we minimized each of the complexes. The final complexes were scored by snap binding energy (SnapBE). For each ligand, the lowest-energy complex among all complexes with an active-state GPCR was selected as the final active-state pose, and the lowest-energy complex among all complexes with an inactive-state GPCR was selected as the final inactive-state pose.

A.5 Molecular dynamics simulation

Preparing the crystal structures of hβ_2AR and the G$_s$ protein for MD simulation

Starting from the active-state crystal structure coupled with the G$_s$ protein (PDB ID: 3SN6), we modeled the missing loops in the Gα domain of the G protein using the SWISS-MODEL server,(*14*) and built the missing loops in the GPCR with a Monte Carlo technique that grows geometrically allowed loop structures from the two TM ends. We then relaxed the modeled loops by simulated annealing of 10 cycles (lowest temperature: 50 K, highest temperature: 600 K) followed by minimization till RMS force reached 0.25 kcal mol^{-1} Å^{-1}. For the inactive-state case, we aligned the inactive-state crystal structure to the active-state crystal structure to match the agonist into the inactive-state structure. We then minimized the complex till RMS force reached 0.25 kcal mol^{-1} Å^{-1} and did simulated annealing of the binding site within 6 Å of the ligand, followed by minimization.

Loop building of hSSTR5

We modeled ICL1, ECL1 and ECL2 from mOPRM crystal structure (PDB ID: 4DKL)(*15*) using homology modeling. The remaining loops, ICL2, ICL3, and ECL3, were built with a Monte Carlo technique that grows geometrically allowed loop structures from the two fixed TM ends. Then we added the C-terminus of hSSTR5 up to the C-terminus of Helix8 (C320) by attaching Helix8 of the template after aligning their NPxxY motifs followed by mutating to hSSTR5. In addition, we added the N-terminus from residue 36 to 38. Minimization was then carried out on the final structure while keeping the TM domains (except the end residues) fixed until energy was converged.

Modeling Gα_i

We used SWISS-MODEL (*14*) to model Gα_i (UniProt ID: P63096, canonical) from Gα_s in the G protein heterotrimer crystalized with hβ_2AR (PDB ID: 3SN6).(*16*) Using MPSim,(*3*) we minimized Gα_i in vacuum until the RMS force is lowered to 0.1 kcal mol^{-1} Å^{-1}. Then Gα_i was placed with ActiveConf2 by aligning ActiveConf2 to hβ_2AR in 3SN6 and aligning Gα_i to Gα_s in 3SN6. Only backbone atoms were considered in the alignment.

For the ActiveConf2+Gα_i complex, we used MPSim to optimize all loops plus one more residue at each end of each TM the GPCR together with residues 352 to 354 of Gα_i in the ActiveConf2+Gα_i complex.

Building the lipid/water environment for hβ_2AR

- For the agonist+GPCR complex: We used Visual Molecular Dynamics (VMD),(*4*) to insert the prepared complex into a 75 Å × 85 Å lipid bilayer structure (for the inactive state) or into a 75 Å × 95 Å lipid bilayer structure (for the active state). This system was then placed into a water box with a total of ~11300 (for the inactive state) water molecules or ~13000 (for the active state) water molecules in the 15 Å and 35 Å thick space on the extracellular and intracellular sides of the lipid bilayer.

- For the agonist+GPCR+Gα_s complex: Using VMD, the prepared agonist+GPCR+Gα_s complex was inserted into a 120 Å × 130 Å lipid bilayer structure. This system was then placed into a water box with a total of ~39500 water molecules in the 15 Å and 60 Å thick space on the extracellular and intracellular sides of the lipid bilayer.

- For Gα_s alone: Using VMD, the prepared Gα_s protein was placed into a rectangular water box with 10 Å thick of water padded on each of x-, y-, z-direction of the protein. There were a total of ~19400 water molecules.

- For the agonist alone: Using VMD, the agonist taken from the prepared agonist+GPCR+Gα_s complex was placed into a rectangular water box with 15 Å thick of water padded on each of x-, y-, z- direction of the molecule. There were a total of ~1700 water molecules.

For all the cases above, in the end, Na^+ and Cl^- ions were placed into the system using tleap for a physiological NaCl concentration (0.9% w/v) and a neutral system.

Building the lipid/water environment for hSSTR5

- For ligand-bound GPCR: Using VMD, the final hSSTR5 structure with loops built was inserted into a 75 Å × 75 Å lipid bilayer structure. The system was then placed into a water box with a total of ~8300 water molecules in the 15 Å and 25 Å thick space on the extracellular and intracellular sides of the lipid bilayer.

- For agonist+GPCR+Gα_i complex: Using VMD, the agonist+GPCR+Gα_i complex was inserted into a 120 Å × 100 Å lipid bilayer structure. The system was then placed into a water box with a total of ~29600 water molecules in the 15 Å and 60 Å thick space on the extracellular and intracellular sides of the lipid bilayer.

- For Gα_{i} alone: Using VMD, the modeled Gα_{i} was placed into a rectangular water box with 10 Å thick of water padded on each of x-, y-, z- direction of the protein. There were a total of ~20800 water molecules.
- For the agonist alone: Using VMD, the agonist taken from the agonist+GPCR+Gα_{i} complex was placed into a rectangular water box with 15 Å thick of water padded on each of x-, y-, z- direction of the molecule. There were a total of 1855 water molecules.

For all the cases above, in the end, Na^+ and Cl^- ions were placed into the system with tleap for a physiological NaCl concentration and a neutral system.

MD simulation protocol

MD_Step1) With the ligand and TM regions of the receptor fixed, the loops and Helix8 of the receptor, lipids and water molecules in the system were minimized for 10000 steps using the conjugate gradient method. In the case of the agonist+GPCR+Gα complex, Gα is fixed in this step too. In the case of Gα alone in solvent, only Gα is fixed in this step, and the minimization was carried out for 5000 steps.

MD_Step2) With the ligand and TM regions of the receptor fixed, the loops and Helix8 of the receptor, the lipids and water molecules were equilibrated at 310 K and 1 atm for 1 ns using the NPT ensemble. This allowed the water molecules to defuse into the ligand-protein system filling any artificial voids in the simulation system. In the case of the agonist+GPCR+Gα complex, Gα is fixed in this step too. In the case of Gα alone in solvent, only Gα is fixed in this step, and the equilibration was for 1.5 ns.

MD_Step3) The whole system was then minimized for 10000 steps using the conjugate gradient method. In the case of Gα alone in solvent, the minimization was carried out for 5000 steps.

MD_Step4) The whole system was heated from 0 K to 310 K in hundreds of ps and then equilibrated using the NPT ensemble for a total of 51 ns MD simulation.

For the simulation of apo-GPCR+Gα, we took the last frame of agonist+GPCR+Gα complex from the above procedure, removed the ligand from the system, adjusted the number of Na^+ or Cl^- ions to make the system neutral again if the ligand was charged, and repeated the above MD protocol MD_Step1) to MD_Step4) with the all parts of the proteins fixed in MD_Step1).

References

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

(*2*) Case, D. A., Darden, T., Cheatham, T., Simmerling, C. L., Wang, J., Duke, R. E., Luo, R., Walker, R., Zhang, W., and Merz, K., Amber 11., Report, University of California, 2010.

(*3*) Lim, K.-T., Brunett, S., Iotov, M., McClurg, R. B., Vaidehi, N., Dasgupta, S., Taylor, S., and Goddard, W. A., (1997). Molecular dynamics for very large systems on massively parallel computers: the MPSim program. *Journal of Computational Chemistry 18*, 501–521.

(*4*) Humphrey, W., Dalke, A., and Schulten, K., (1996). VMD: visual molecular dynamics. *Journal of molecular graphics 14*, 33–38.

(*5*) Bray, J. K., Abrol, R., Goddard III, W. A., Trzaskowski, B., and Scott, C. E., (2014). SuperBiHelix method for predicting the pleiotropic ensemble of G-protein-coupled receptor conformations. *Proceedings of the National Academy of Sciences 111*, E72–E78.

(*6*) Floriano, W. B., Vaidehi, N., Zamanakos, G., and Goddard, W. A., (2004). HierVLS hierarchical docking protocol for virtual ligand screening of large-molecule databases. *Journal of medicinal chemistry 47*, 56–71.

(*7*) Goddard, W. A., Kim, S.-K., Li, Y., Trzaskowski, B., Griffith, A. R., and Abrol, R., (2010). Predicted 3D structures for adenosine receptors bound to ligands: Comparison to the crystal structure. *Journal of structural biology 170*, 10–20.

(*8*) MacroModel, version 9.7, Schrödinger, LLC, New York, NY, 2009.

(*9*) Maestro, version 9.1, Schrödinger, LLC, New York, NY, 2010.

(*10*) Chang, G., Guida, W. C., and Still, W. C., (1989). An internal-coordinate Monte Carlo method for searching conformational space. *Journal of the American Chemical Society 111*, 4379–4386.

(*11*) Saunders, M., Houk, K. N., Wu, Y. D., Still, W. C., Lipton, M., Chang, G., and Guida, W. C., (1990). Conformations of cycloheptadecane. A comparison of methods for conformational searching. *Journal of the American Chemical Society 112*, 1419–1427.

(*12*) Martin, R. E., Mohr, P., Maerki, H. P., Guba, W., Kuratli, C., Gavelle, O., Binggeli, A., Bendels, S., Alvarez-Sánchez, R., Alker, A., et al. (2009). Benzoxazole piperidines as selective and potent somatostatin receptor subtype 5 antagonists. *Bioorganic & medicinal chemistry letters 19*, 6106–6113.

(*13*) Jaguar, version 7.6, Schrödinger, LLC, New York, NY, 2009.

(*14*) Schwede, T., Kopp, J., Guex, N., and Peitsch, M. C., (2003). SWISS-MODEL: an automated protein homology-modeling server. *Nucleic acids research 31*, 3381–3385.

(*15*) Manglik, A., Kruse, A. C., Kobilka, T. S., Thian, F. S., Mathiesen, J. M., Sunahara, R. K., Pardo, L., Weis, W. I., Kobilka, B. K., and Granier, S., (2012). Crystal structure of the μ-opioid receptor bound to a morphinan antagonists. *Nature 485*, 321–326.

(*16*) Rasmussen, S. G. F., et al. (2011). Crystal structure of the β2 adrenergic receptor-Gs protein complex. *Nature 477*, 549–555, DOI: `10.1038/nature10361`.

Appendix B

Additional Computational Details for the Application of ActiveGEnSeMBLE: Predicting hSSTR5 Structures

B.1 PredicTM and secondary structure prediction: determining the 7 transmembrane domains (TMDs) and any helical extensions past the membrane

We carried out multiple sequence alignments using MAFFT (*1*) method over all GPCRs having a sequence identity higher than 8.8% (from BLAST)(*2*) with hSSTR5. Then we used the hydrophobicity values from the White and von Heijne scales (*3*, *4*) to predict the hydrophobicity along the target sequence. Then, we removed noise in the hydrophobicity profile by using the mean hydrophobicity values obtained from averaging windows ranging from 7-residues to 21-residues. Regions with a hydrophobicity value above zero in the final smooth hydrophobicity profile for hSSTR5 (Figure C.2), are defined as "raw" TMDs, leading to exactly seven continuously positive regions expected to correspond to the seven TMDs buried inside the membrane.

The x-ray structures for GPCRs often find helical extensions of the TMDs well past what would correspond to the boundary of the membrane (for example, in squid opsin, TM5 and TM6 are helical 25Å beyond the membrane).(*5*) To identify these helical extensions protruding from the membrane for each helix, we predicted helix propensity using a cross comparison of consensus results from protein secondary structure prediction servers Porter,(*6*) SSpro,(*7*, *8*) APSSP,(*9*) Jpred,(*10*) and PSIPRED,(*11*, *12*) all of which predict helical regions using trained neural networks. The raw results are in Figure C.3. The final TM helical domains extended from the raw TMDs are denoted as "cap" regions, as indicated in Figure C.3.

In PredicTM, we specified the hydrophobic center (HPC) of each helix by one of two criteria:

- "Rawmid" takes HPC to be the geometric midpoint of the raw TMD.
- "Area" integrates the hydrophobicity over the raw TMD, and takes HPC to be the centroid (half the total area on each side).

The HPCs for all chains were taken to be in the same x-y plane (the midplane of the lipid membrane bilayer).

B.2 Template selection

The sequence alignment from PredicTM identified three GPCRs for which x-ray structures were available that had the highest sequence identity over the TM region with hSSTR5:

- nociceptin receptor (OPRX_HUMAN, hOPRX, 46.79%),
- μ-opioid receptor (OPRM_MOUSE, mOPRM, 44.62%) and
- κ-opioid receptor (OPRK_HUMAN, hOPRK, 40.33%)

The next closest was human C-X-C chemokine receptor type 4 (CXCR4_HUMAN, hCXCR4, 32.13%). Thus, for exploring a diverse set of relatively high sequence identity templates, we used hOPRX (PDB ID: 4EA3(*13*)), mOPRM (PDB ID: 4DKL(*14*)) and hOPRK (PDB ID: 4DJH(*15*)) as templates for our structure predictions. There was no experimental structure available at the time of this study for δ-opioid receptor (OPRD_MOUSE). A summary of the selected sequence identity comparison is in Table C.1. All template structures used in the following steps were from the Protein Data Bank (PDB).

B.3 Predicting the helical shape: OptHelix vs homology modeling

Predicted shape of the TMD is to be used in the future step that determines the tilts and rotations of the TMD. Two methods were used to predict the shape of TMDs:

- OptHelix generates the helical shape features using energy minimization and molecular dynamics starting with an α-helix based on the peptide sequence in which residues other than Pro, Gly, Ala, Ser, and Thr are replaced with Ala. Then after dynamics the average structure is mutated to the correct sequence.
- Homology to the template shape. Here we start with the backbone from the template (usually from an x-ray structure), mutate it to the new sequence, optimize side chains using Side Chain Rotamer Energy Analysis Method (SCREAM),(*16*) and minimize the TMD.

The following is a detailed description of these two methods:

OptHelix

This method treats each of the seven helices separately. It first takes the TM lengths predicted above, and generates seven separate canonical polyalanine helices accordingly. Then, it mutates the Pro and Gly back to their respective positions on the helices using Side Chain Rotamer Energy Analysis Method (SCREAM). A first structural optimization is then done to minimize the energy of each helix. Subsequently, the Ser and Thr adjacent to Pro are mutated back, and a molecular dynamics simulation on each helix is run for 2 ns. Finally, all remaining residues are mutated back to have their original side chains, and a second energy minimization is performed. For each helix, the structures that go to the final step are selected based on "minrmsd", which takes the snapshot that has the average root mean square deviation (RMSD) closest to the average structure from the MD, and based on "mineng", which takes the snapshot that has the lowest energy from the latter 75% of the MD.

Homology modeling

The template structures were taken from the Orientations of Proteins in Membranes (OPM) database. For each template protein, the sequence was aligned with that of the target protein hSSTR5, and the corresponding residues in the template structure were mutated to be that of the target protein using SCREAM. Then each helix was truncated or extended to the previously determined start/end residues, which was followed by a geometry optimization of each individual helix for 100 steps using the DREIDING-III force field.(*17*, *18*)

B.4 Assembling the bundle

With the shape of each TMD determined, the next step is to assemble the helices into a bundle.

Assuming each TMD to be rigid, six parameters uniquely define the orientation of the TM helices: 1) the HPC residue h (which is taken to be at z = 0 so that all TMD have their HPC on the same plane); 2) & 3) the Cartesian coordinates (x,y) of the HPC; 4) the inclination angle θ of the helical axis relative to the z-axis; 5) the azimuthal angle ϕ; 6) the rotation angle η of the helix around its own helical axis. Except for TM3, we define the reference point for the rotation angle η using the most conserved residue in each TM, which is denoted as n.50 in the Ballesteros numbering scheme. For TM3, we choose 3.32 rather than 3.50 because 3.32 is closer to the center of the helix, and is also well conserved.

The hydrophobic centers used in assembling OptHelix helices were obtained using PredicTM (based either on the "area" or "rawmid" criterion as described in Experimental Section 1.1). The other parameters (x,y,θ,ϕ,η) were all based on the template protein structure from the Orientations of Proteins in Membranes (OPM) database.(*19*) We chose helical shapes from both the "minrmsd" (minimum root mean square deviation to the average structure) and "mineng" (minimum energy) criteria. Thus, for each of the three template proteins we generated a total of four structures based on OptHelix, for a total of 12. For homology helices, we selected (x,y,h,θ,ϕ,η) from the template.

Now GEnSeMBLE starts with the x, y, h, θ, ϕ, η parameters of a starting structure. It then optimizes first η using the BiHelix method,(*20*) and then optimizes θ, ϕ, η using the SuperBiHelix(*21*) method, as described in the next section.

B.5 Determining the optimum helical rotations and tilts

Selecting the optimum helical rotations (η) using the BiHelix method

Mutating from the template sequence to the target sequence, hSSTR5 in this case, is likely to make dramatic changes in some of interhelical interactions. Thus the first step of GEnSeMBLE is to sample all changes in the rotation angles, $\Delta\eta$, from 0° to 360° with a step size of 30°. This leads to $12^7 \approx 35$ million combinations. Rather than construct seven-helix bundles for all 35 million, the BiHelix method simplifies the problem by considering the 12 pairs of interacting helices independently. Thus for each pair we considered $12^2 = 144$ cases, for each of which we optimized the residue side chain conformations with SCREAM. This set of $12\times144 = 1728$ numbers was used to estimate the energies for all 35 million combinations. We then took the lowest 1000 combinations to analyze in the CombiHelix step.

In the CombiHelix step for each of the 1000 combinations from BiHelix, we built the full seven-helix bundle, reoptimized the side chains using SCREAM and minimized for 10 steps. Then the total energies from these 1000 were ordered and the lowest-energy cases were kept for consideration of the optimum tilt angles. The conformations were ranked by "E_{CNti}".

Selecting the optimum helical tilts (θ,ϕ) and rotations (η) using the SuperBiHelix method

Our previous studies showed that, starting with the x-ray structure of one GPCR, we could not match the structure of a different GPCR without allowing both the helix rotations and helix tilts to change. To make this search practical, we developed the

SuperBiHelix method.(*21*) Starting with the optimum rotation angles from BiHelix, we first carried out a coarse sampling step ($\Delta\theta$ = 0, ±15°; $\Delta\phi$ = 0, ±45°, ±90°; $\Delta\eta$ =0, ±30°, selected angles from BiHelix) which involved more than $(3\times5\times3)^7 \approx 374$ billion configurations. The selected angles of $\Delta\eta$ apart from 0 and ±30° were those appearing more than twice in top 20 structures of the BiHelix result or appearing in top 3, and are shown in Table C.2. The energies for all these configurations were estimated using the BiHelix energies but combined into 3 groups of quad helices as explained by Bray and coworkers.(*21*) For the top 2000 combinations of tilts and rotations, we built the seven-helix bundles, optimized the side chains, and selected the best case based on the energy ranking.

The subsequent finer SuperBiHelix sampling was based on the selected coarse sampling resulting structures, and the sampling range was $\Delta\theta$ = 0, ±15°; $\Delta\phi$ = 0, ±15°, ±30°; $\Delta\eta$ =0, ±30°. Again, the top 2000 combinations were built into seven-helix bundles.

References

(*1*) Katoh, K., Kuma, K., Toh, H., and Miyata, T., (2005). MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic acids research 33*, 511–518.

(*2*) Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J., (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research 25*, 3389–3402.

(*3*) Wimley, W. C., Creamer, T. P., and White, S. H., (1996). Solvation Energies of Amino Acid Side Chains and Backbone in a Family of Host-Guest Pentapeptides. *Biochemistry 35*, 5109–5124.

(*4*) Hessa, T., Meindl-Beinker, N. M., Bernsel, A., Kim, H., Sato, Y., Lerch-Bader, M., Nilsson, I. M., White, S. H., and Von Heijne, G., (2007). Molecular code for transmembrane-helix recognition by the Sec61 translocon. *Nature 450*, 1026–1030.

(*5*) Murakami, M., and Kouyama, T., (2008). Crystal structure of squid rhodopsin. *Nature 453*, 363–367.

(*6*) Pollastri, G., and McLysaght, A., (2005). Porter: a new, accurate server for protein secondary structure prediction. *Bioinformatics 21*, 1719–1720.

(*7*) Pollastri, G., Przybylski, D., Rost, B., and Baldi, P., (2002). Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles. *Proteins: Structure, Function, and Bioinformatics 47*, 228–235.

(*8*) Cheng, J., Randall, A. Z., Sweredoski, M. J., and Baldi, P., (2005). SCRATCH: a protein structure and structural feature prediction server. *Nucleic acids research 33*, W72–W76.

(*9*) Raghava, G. P. S., (2002). APSSP2: A combination method for protein secondary structure prediction based on neural network and example based learning. *Casp5*, A–132.

(*10*) Cole, C., Barber, J. D., and Barton, G. J., (2008). The Jpred 3 secondary structure prediction server. *Nucleic acids research 36*, W197–W201.

(*11*) Jones, D. T., (1999). Protein secondary structure prediction based on position-specific scoring matrices. *Journal of molecular biology 292*, 195–202.

(*12*) Buchan, D. W. A., Ward, S. M., Lobley, A. E., Nugent, T. C. O., Bryson, K., and Jones, D. T., (2010). Protein annotation and modelling servers at University College London. *Nucleic acids research 38*, W563–W568.

(*13*) Thompson, A. A., Liu, W., Chun, E., Katritch, V., Wu, H., Vardy, E., Huang, X.-P., Trapella, C., Guerrini, R., and Calo, G., (2012). Structure of the nociceptin/orphanin FQ receptor in complex with a peptide mimetic. *Nature 485*, 395–399.

(*14*) Manglik, A., Kruse, A. C., Kobilka, T. S., Thian, F. S., Mathiesen, J. M., Sunahara, R. K., Pardo, L., Weis, W. I., Kobilka, B. K., and Granier, S., (2012). Crystal structure of the μ-opioid receptor bound to a morphinan antagonists. *Nature 485*, 321–326.

(*15*) Wu, H., Wacker, D., Mileni, M., Katritch, V., Han, G. W., Vardy, E., Liu, W., Thompson, A. A., Huang, X.-P., and Carroll, F. I., (2012). Structure of the human κ-opioid receptor in complex with JDTic. *Nature 485*, 327–332.

(*16*) Tak Kam, V. W., and Goddard III, W. A., (2008). Flat-bottom strategy for improved accuracy in protein side-chain placements. *Journal of Chemical Theory and Computation 4*, 2160–2169.

(*17*) Weckbecker, G., Lewis, I., Albert, R., Schmid, H. A., Hoyer, D., and Bruns, C., (2003). Opportunities in somatostatin research: biological, chemical and therapeutic aspects. *Nature Reviews Drug Discovery 2*, 999–1017.

(*18*) Olias, G., Viollet, C., Kusserow, H., Epelbaum, J., and Meyerhof, W., (2004). Regulation and function of somatostatin receptors. *Journal of neurochemistry 89*, 1057–1091.

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

(*20*) Abrol, R., Bray, J. K., and Goddard III, W. A., (2012). Bihelix: Towards de novo structure prediction of an ensemble of G-protein coupled receptor conformations. *Proteins: Structure, Function, and Bioinformatics*.

(*21*) Bray, J. K., Abrol, R., Goddard III, W. A., Trzaskowski, B., and Scott, C. E., (2014). SuperBiHelix method for predicting the pleiotropic ensemble of G-protein-coupled receptor conformations. *Proceedings of the National Academy of Sciences 111*, E72–E78.

Appendix C

Additional Figures and Tables for hSSTR5 Structure Prediction and Antagonist Binding

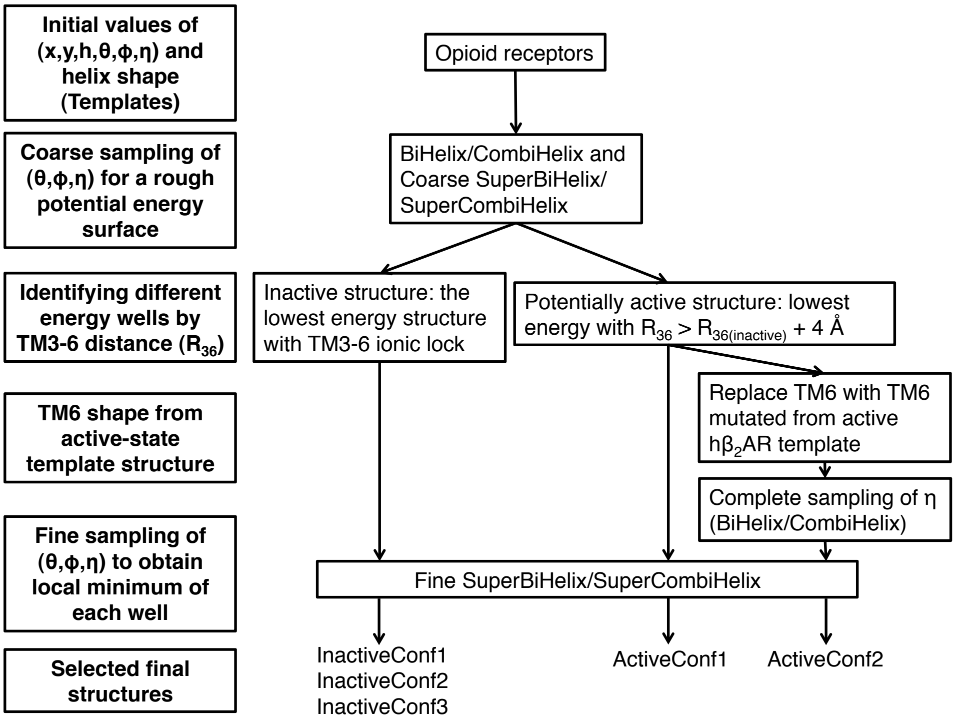

Initial values of (x,y,h,θ,φ,η) and helix shape (Templates)
Coarse sampling of (θ,φ,η) for a rough potential energy surface
Identifying different energy wells by TM3-6 distance (R36)
TM6 shape from active-state template structure
Fine sampling of (θ,φ,η) to obtain local minimum of each well
Selected final structures
Opioid receptors
BiHelix/CombiHelix and Coarse SuperBiHelix/ SuperCombiHelix
Inactive structure: the lowest energy structure with TM3-6 ionic lock
Potentially active structure: lowest energy with R36 > R36(inactive) + 4 Å
Replace TM6 with TM6 mutated from active hβ2AR template
Complete sampling of η (BiHelix/CombiHelix)
Fine SuperBiHelix/SuperCombiHelix
InactiveConf1
InactiveConf2
InactiveConf3
ActiveConf1
ActiveConf2

Figure C.1: Flow chart of ActiveGEnSeMBLE for hSSTR5 structure prediction. R_{36} is the shortest distance between the backbone atoms of the intracellular ends of transmembrane (TM) 3 and TM6. The sampling space of BiHelix is $\Delta\eta$ from 0 to 360° in 30° increments. The sampling space of Coarse SuperBiHelix is $\Delta\phi$: 0, ±45°, ±90°; $\Delta\eta$: 0, ±30°, selected angles from BiHelix/CombiHelix, starting with the best from BiHelix/CombiHelix. The sampling space of Fine SuperBiHelix is $\Delta\phi$: 0, ±15°, ±30°; $\Delta\eta$: 0, ±30°.

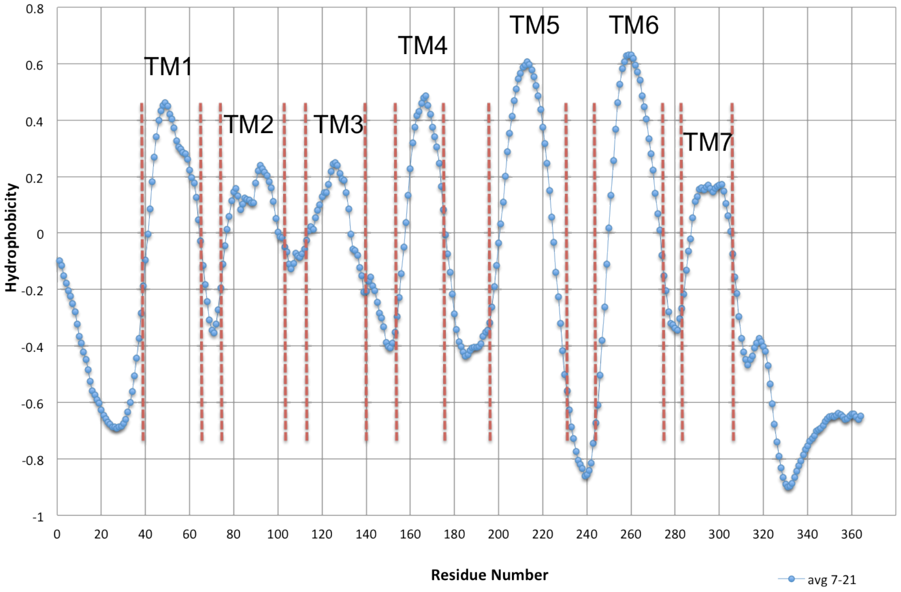
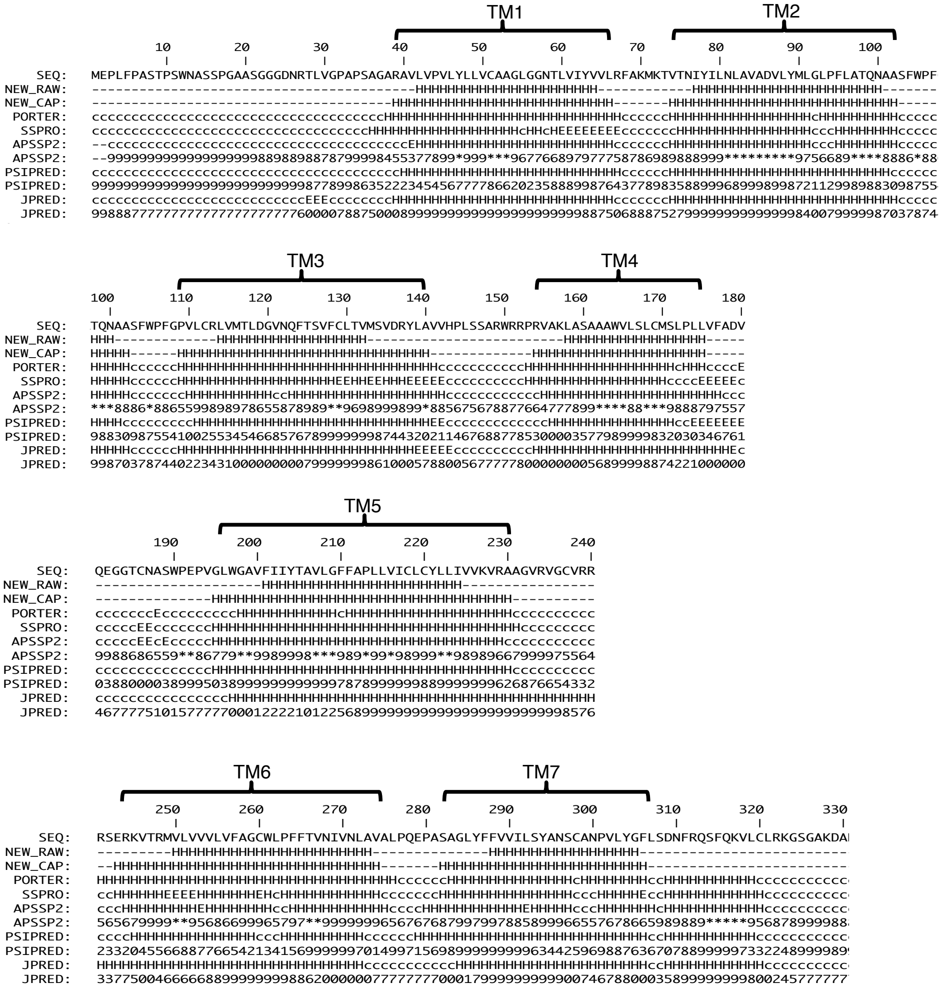


Figure C.2: Hydrophobicity profile of hSSTR5 before applying the capping rules. The residues expected to lie in the membrane are indicated by red dashed lines.

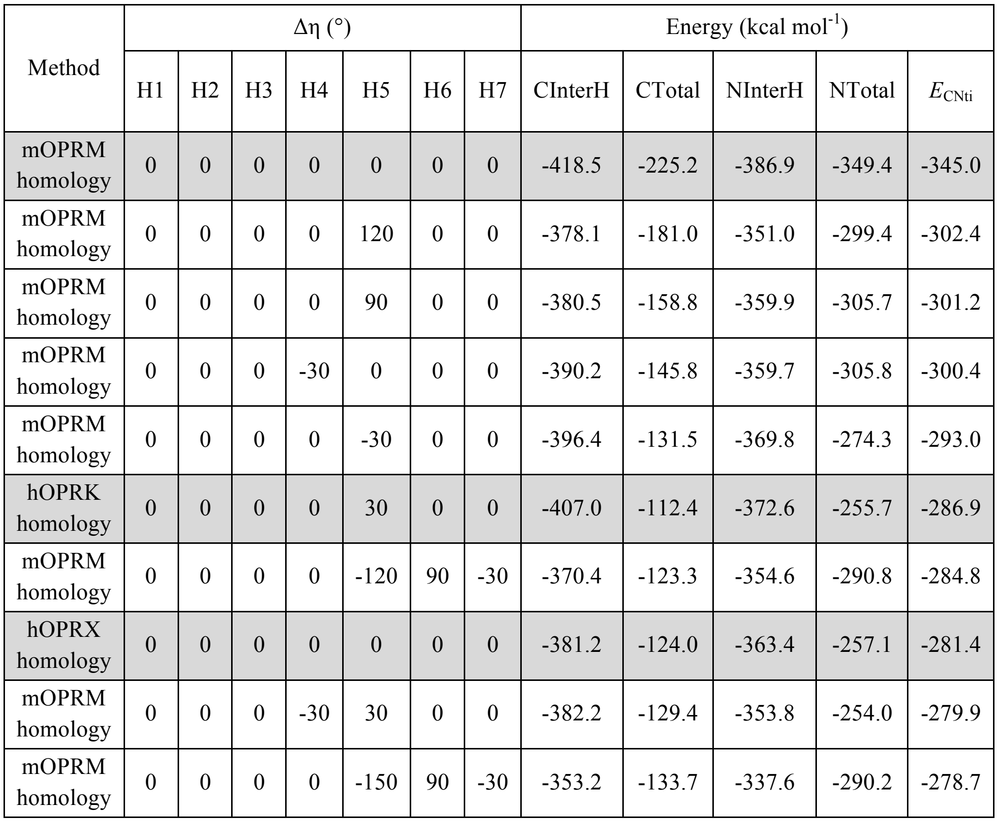


```
                   10        20        30        40        50        60        70        80        90       100
                    |         |         |         |         |         |         |         |         |         |
      SEQ:  MEPLFPASTPSWNASSPGAASGGGDNRTLVGPAPSAGARAVLVPVLYLLVCAAGLGGNTLVIYVVLRFAKMKTVTNIYILNLAVADVLYMLGLPFLATQNAASFWPF
  NEW_RAW:  -----------------------------------------HHHHHHHHHHHHHHHHHHHHHHH------------HHHHHHHHHHHHHHHHHHHHHHHH-------
  NEW_CAP:  --------------------------------------HHHHHHHHHHHHHHHHHHHHHHHHHHHH-------HHHHHHHHHHHHHHHHHHHHHHHHHHHHH-----
   PORTER:  cccccccccccccccccccccccccccccccccccccHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHccccccHHHHHHHHHHHHHHHHHHcHHHHHHHHHHccccc
    SSPRO:  cccccccccccccccccccccccccccccccccccHHHHHHHHHHHHHHHHHHHcHHcHEEEEEEEEcccccccHHHHHHHHHHHHHHHHHcccHHHHHHHHccccc
   APSSP2:  --ccccccccccccccccccccccccccccccccccccccEHHHHHHHHHHHHHHHHHHHHHHHHHcccccccHHHHHHHHHHHHHHHHHHcccHHHHHHHHccccc
   APSSP2:  --99999999999999999998898898878799998455377899*999***967766897977758786989888999*********9756689****8886*88
  PSIPRED:  cccccccccccccccccccccccccccccccccccccccHHHHHHHHHHHHHHHHHHHHHHHHHHHHccccccHHHHHHHHHHHHHHHHHHHHHHHHHHHHcccccc
  PSIPRED:  9999999999999999999999999998778998635222345456777786620235888998764377898358899689998998721129989883098755
    JPRED:  cccccccccccccccccccccccccccEEEccccccccHHHHHHHHHHHHHHHHHHHHHHHHHHHHcccccccHHHHHHHHHHHHHHHHHHHHHHHHHHHHHccccc
    JPRED:  99888777777777777777777777600007887500089999999999999999999998875068887527999999999999999840079999987037874

           100       110       120       130       140       150       160       170       180
             |         |         |         |         |         |         |         |         |
      SEQ:  TQNAASFWPFGPVLCRLVMTLDGVNQFTSVFCLTVMSVDRYLAVVHPLSSARWRRPRVAKLASAAAWVLSLCMSLPLLVFADV
  NEW_RAW:  HHH-------------HHHHHHHHHHHHHHHHHHH-------------------------HHHHHHHHHHHHHHHHHH-----
  NEW_CAP:  HHHHH------HHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHH-------------HHHHHHHHHHHHHHHHHHHHHH-----
   PORTER:  HHHHHccccccHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHcccccccccccHHHHHHHHHHHHHHHHHHHcHHHccccE
    SSPRO:  HHHHHccccccHHHHHHHHHHHHHHHHHHHHEEHHEEHHHEEEEEccccccccccHHHHHHHHHHHHHHHHHHccccEEEEEc
   APSSP2:  HHHHHcccccccHHHHHHHHHHHccHHHHHHHHHHHHHHHHHHHHccccccccccccHHHHHHHHHHHHHHHHHHHHHHHccc
   APSSP2:  ***8886*8865599898978655878989**9698999899*885675678877664777899****88***9888797557
  PSIPRED:  HHHHcccccccccHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHEEcccccccccccccHHHHHHHHHHHHHHHHccEEEEEEE
  PSIPRED:  98830987554100255345466857678999999987443202114676887785300003577989999832030346761
    JPRED:  HHHHHccccccHHHHHHHHHHHHHHHHHHHHHHHHHHHHHEEEEEcccccccccccHHHHHHHHHHHHHHHHHHHHHHHHHEc
    JPRED:  99870378744022343100000000007999999986100057880056777778000000568999998874221000000

                  190       200       210       220       230       240
                    |         |         |         |         |         |
      SEQ:  QEGGTCNASWPEPVGLWGAVFIIYTAVLGFFAPLLVICLCYLLIVVKVRAAGVRVGCVRR
  NEW_RAW:  --------------------HHHHHHHHHHHHHHHHHHHHHHHH----------------
  NEW_CAP:  --------------HHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHH----------
   PORTER:  cccccccEcccccccccHHHHHHHHHHHHcHHHHHHHHHHHHHHHHHHHHcccccccccc
    SSPRO:  cccccEEcccccccHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHccccccccc
   APSSP2:  cccccEEcEcccccHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHccccccccccc
   APSSP2:  9988686559**86779**9989998***989*99*98999**98989667999975564
  PSIPRED:  ccccccccccccccHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHcccccccccc
  PSIPRED:  038800003899950389999999999999787899999988999999962687665433
    JPRED:  ccccccccccccccccHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHH
    JPRED:  467777510157777700012222101225689999999999999999999999998576

                  250       260       270       280       290       300       310       320       330
                    |         |         |         |         |         |         |         |         |
      SEQ:  RSERKVTRMVLVVVLVFAGCWLPFFTVNIVNLAVALPQEPASAGLYFFVVILSYANSCANPVLYGFLSDNFRQSFQKVLCLRKGSGAKDA
  NEW_RAW:  ---------HHHHHHHHHHHHHHHHHHHHHHHH--------------HHHHHHHHHHHHHHHHHH-------------------------
  NEW_CAP:  --HHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHH-------HHHHHHHHHHHHHHHHHHHHHHHHH------------------------
   PORTER:  HHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHcccccHHHHHHHHHHHHHHcHHHHHHHHHHccHHHHHHHHHHHccccccccccc
    SSPRO:  ccHHHHHEEEEHHHHHHHHEHcHHHHHHHHHHHHccccccHHHHHHHHHHHHHHHHHcccHHHHHEccHHHHHHHHHHHccccccccccc
   APSSP2:  cccHHHHHHHHHEHHHHHHHHccHHHHHHHHHHHHccccHHHHHHHHHHHHHEHHHHHcccHHHHHHHcHHHHHHHHHHccccccccccc
   APSSP2:  565679999**95686699965797**9999999656767687997997885899966557678665989889*****956878999988
  PSIPRED:  ccccHHHHHHHHHHHHHHHcccHHHHHHHHHHccccccHHHHHHHHHHHHHHHHHHHHHHHHHHHHccHHHHHHHHHHHccccccccccc
  PSIPRED:  233204556688776654213415699999970149971569899999999996344259698876367078899997332248999898
    JPRED:  HHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHccccccccccHHHHHHHHHHHHHHHHHHHHHHHHccHHHHHHHHHHHccccccccccc
    JPRED:  337750046666688999999999886200000077777777000179999999999007467880003589999999800245777777
```

Figure C.3: Secondary structure prediction for hSSTR5 from various servers. NEW_RAW is PredicTM prediction; NEW_CAP is from consensus of different secondary structure prediction servers.

Protein Identifier	Rank	All	TM Avg	TM1	TM2	TM3	TM4	TM5	TM6	TM7
P35346|SSR5_HUMAN	1	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
P41146|OPRX_HUMAN	**24**	**33.79**	**46.79**	**56.52**	**58.33**	**47.37**	**22.22**	**41.67**	**45.83**	**55.56**
P42866|OPRM_MOUSE	**29**	**33.24**	**44.62**	**52.17**	**50.00**	**42.11**	**27.78**	**33.33**	**45.83**	**61.11**
P32300|OPRD_MOUSE	37	34.07	43.96	47.83	50.00	47.37	27.78	33.33	45.83	55.56
P41145|OPRK_HUMAN	**42**	**31.04**	**40.33**	**39.13**	**50.00**	**47.37**	**16.67**	**33.33**	**45.83**	**50.00**
P61073|CXCR4_HUMAN	83	23.63	32.13	39.13	45.83	31.58	16.67	33.33	25.00	33.33
P02699|OPSD_BOVIN	432	20.05	29.99	26.09	33.33	15.79	33.33	45.83	33.33	22.22
P07700|ADRB1_MELGA	464	20.33	33.73	34.78	25.00	26.32	16.67	29.17	54.17	50.00
P07550|ADRB2_HUMAN	574	18.68	27.59	30.43	25.00	21.05	16.67	20.83	45.83	33.33
P31356|OPSD_TODPA	764	17.03	26.44	30.43	29.17	15.79	11.11	45.83	25.00	27.78
P08172|ACM2_HUMAN	1027	19.23	30.37	17.39	25.00	36.84	44.44	25.00	25.00	38.89
P29274|AA2AR_HUMAN	1042	16.76	27.74	21.74	20.83	21.05	33.33	25.00	33.33	38.89
P08483|ACM3_RAT	1231	18.13	30.62	26.09	29.17	36.84	27.78	25.00	25.00	44.44

Table C.1: Sequence similarity (in percentage) between hSSTR5 and Class A GPCRs with experimentally available structures. The bold underlined cases were used as templates in this study.



Method	Δη (°)							Energy (kcal mol^{-1})				
	H1	H2	H3	H4	H5	H6	H7	CInterH	CTotal	NInterH	NTotal	E_{CNti}
mOPRM homology	0	0	0	0	0	0	0	-418.5	-225.2	-386.9	-349.4	-345.0
mOPRM homology	0	0	0	0	120	0	0	-378.1	-181.0	-351.0	-299.4	-302.4
mOPRM homology	0	0	0	0	90	0	0	-380.5	-158.8	-359.9	-305.7	-301.2
mOPRM homology	0	0	0	-30	0	0	0	-390.2	-145.8	-359.7	-305.8	-300.4
mOPRM homology	0	0	0	0	-30	0	0	-396.4	-131.5	-369.8	-274.3	-293.0
hOPRK homology	0	0	0	0	30	0	0	-407.0	-112.4	-372.6	-255.7	-286.9
mOPRM homology	0	0	0	0	-120	90	-30	-370.4	-123.3	-354.6	-290.8	-284.8
hOPRX homology	0	0	0	0	0	0	0	-381.2	-124.0	-363.4	-257.1	-281.4
mOPRM homology	0	0	0	-30	30	0	0	-382.2	-129.4	-353.8	-254.0	-279.9
mOPRM homology	0	0	0	0	-150	90	-30	-353.2	-133.7	-337.6	-290.2	-278.7

Table C.2: Top 10 structures from the BiHelix/CombiHelix predictions for all 15 starting structures. Shaded rows represent the lowest-energy case for each template. These cases were used for the SuperBiHelix step. The $\Delta\eta$ values are the deviations of helix (H) rotation angles from the respective homology templates.

Template	$\Delta\theta$ (°)	$\Delta\varphi$ (°)	$\Delta\eta$ (°)
mOPRM	0, ±15	0, ±45, ±90	0, ±30, -60 (H6), -90 (H5, H6), -120 (H5), -150 (H5)
hOPRK	0, ±15	0, ±45, ±90	0, ±30
hOPRX	0, ±15	0, ±45, ±90	0, ±30, ±120 (H5), 60 (H6), 90 (H6), -60 (H7), -90 (H7)

Table C.3: Sampling space of the coarse SuperBiHelix for each of the three templates. The deviation angles apply to every helix (H) unless otherwise labeled. The angles apply specifically to one helix are from top 20 BiHelix results of the respective template. The starting structure of the sampling is the lowest-energy case in BiHelix for each template (the shaded cases in Table C.2).

Rank	$\Delta\theta$ (°)	$\Delta\varphi$ (°)							$\Delta\eta$ (°)							E_{CNti} (kcal mol^{-1})
	All	H1	H2	H3	H4	H5	H6	H7	H1	H2	H3	H4	H5	H6	H7	
1	0	0	0	0	120	15	-15	0	0	0	0	-30	0	0	0	-352.9
2	0	0	0	0	105	15	-15	0	0	0	0	0	0	0	0	-352.8
3	0	0	0	0	120	15	-15	0	0	0	0	0	0	0	0	-351.3
4	0	0	0	0	105	0	-15	0	0	0	0	-30	0	0	0	-350.7
5	0	0	0	0	75	15	-15	0	0	0	0	-30	0	0	0	-346.9
6	0	0	0	-15	60	30	15	0	0	0	0	0	60	0	0	-345.7
7	0	0	0	0	105	15	0	15	0	0	0	0	0	0	0	-343.7
8	0	0	0	-15	75	15	0	0	0	0	0	0	60	0	0	-342.8
9	0	0	0	0	120	0	0	15	0	0	0	-30	0	0	0	-341.4
10	0	0	0	0	105	0	-15	0	0	0	0	0	0	0	0	-337.5
11	0	0	0	0	90	0	0	15	0	0	0	0	0	0	0	-334.8
12	0	0	0	-15	75	30	15	0	0	0	0	0	60	0	0	-333.8
13	0	0	0	0	90	15	-15	0	0	0	0	0	0	0	0	-333.8
14	0	0	0	0	105	15	-30	0	0	0	0	0	0	0	0	-333.3
15	0	0	0	0	120	15	-30	0	0	0	0	0	0	0	0	-332.7
16	0	0	0	-15	60	-30	15	0	0	0	0	0	30	0	0	-328.9
17	0	0	0	-15	60	-15	0	0	0	0	0	0	0	0	0	-328.1
18	0	0	0	-15	60	-15	15	0	0	0	0	0	0	0	0	-328.1
19	0	0	0	-15	60	-15	0	0	0	0	0	0	30	0	0	-328.0
20	0	0	0	-15	75	-15	0	0	0	0	0	0	0	0	0	-327.8
21	0	0	0	0	120	-15	0	15	0	0	0	-30	0	0	0	-327.6
22	0	0	0	-15	60	30	15	15	0	0	0	0	60	0	0	-327.5
23	0	0	0	-15	75	15	15	0	0	0	0	0	60	0	0	-326.8
24	0	0	0	-15	60	0	0	0	0	0	0	0	0	0	0	-325.9
25	0	0	0	-15	60	-30	0	0	0	0	0	0	0	0	0	-324.3

Table C.4: Top 25 structures from fine SuperBiHelix/SuperCombiHelix with mO-PRM as the initial template.

Antagonist	K_{i} (nM)	Relative Binding Energy (kcal mol^{-1})
M59	3	0.00
M60	23	1.25
M38	113	2.23
M40	524	3.17
M42	>1000	3.57

Table C.5: Antagonists' experimental binding constants and their corresponding calculated binding energies relative to M59. The binding energies are calculated according to equation $\Delta G_1 - \Delta G_2 = RT\ \ln(K_{i2}/K_{i1})$.

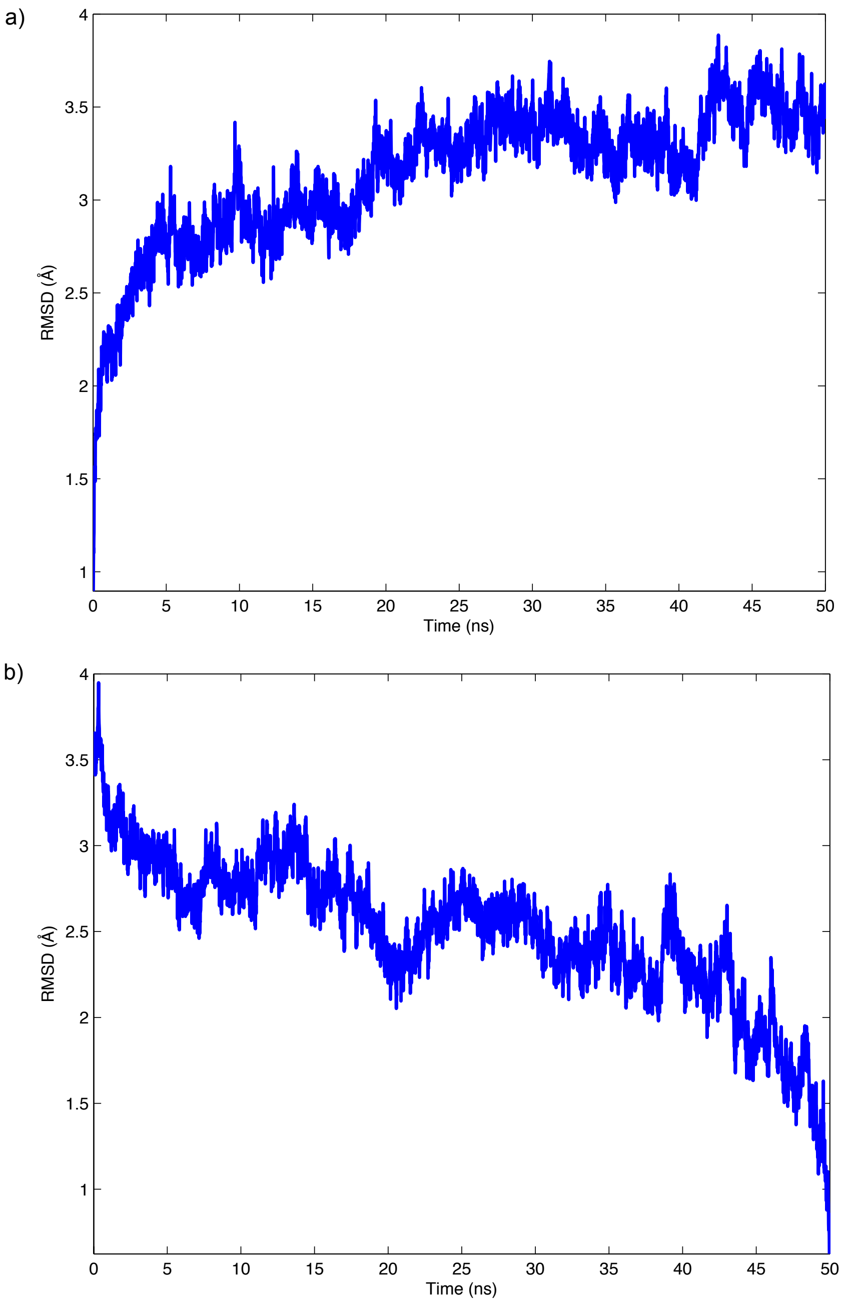


Figure C.4: RMSD changes of the protein along the MD trajectory of M59-bound predicted hSSTR5 structure. a) The snapshots were aligned against the first frame, and RMSD values were calculated with the first frame as the reference. b) The snapshots were aligned against the last frame, and RMSD values were calculated with the last frame as the reference. Only backbone atoms were considered in calculating RMSD values.

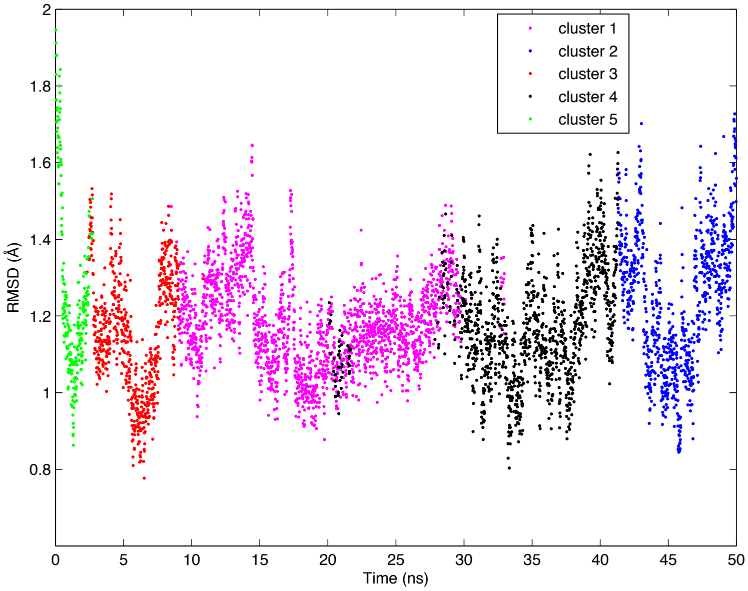


Figure C.5: Results of clustering by RMSD along the MD trajectory of M59-bound predicted hSSTR5 structure. The *k*-means algorithm was used. The *k*-means clustering radius is 2 Å.

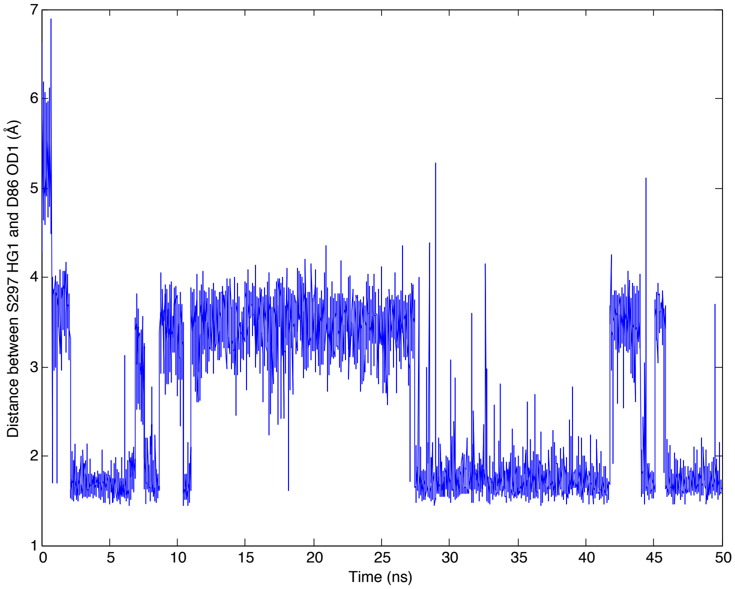


Figure C.6: The fluctuation of interatomic distance between S297$^{7.46}$ O-donated H atom and D86$^{2.50}$ O atom on their side chains during the 50 ns MD simulation of M59-bound predicted hSSTR5 structure.

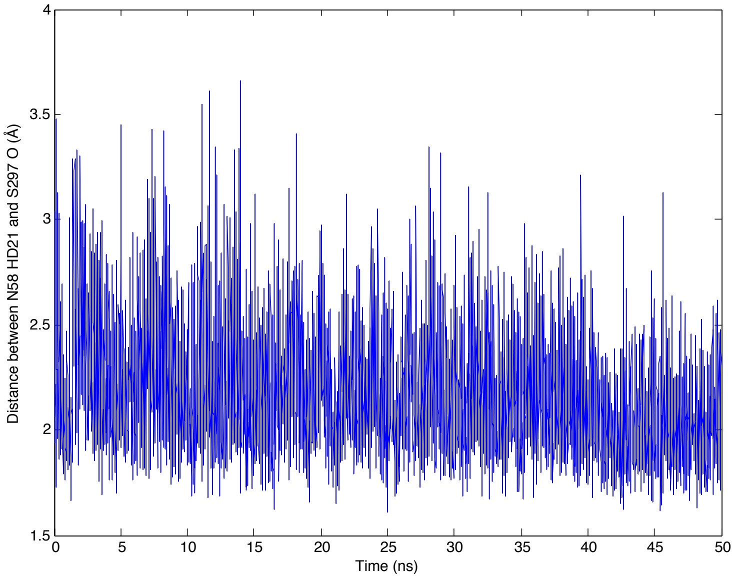


Figure C.7: The fluctuation of interatomic distance between N58$^{1.50}$ N-donated H atom on its side chain and S297$^{7.46}$ O atom on its backbone during the 50 ns MD simulation of M59-bound predicted hSSTR5 structure.

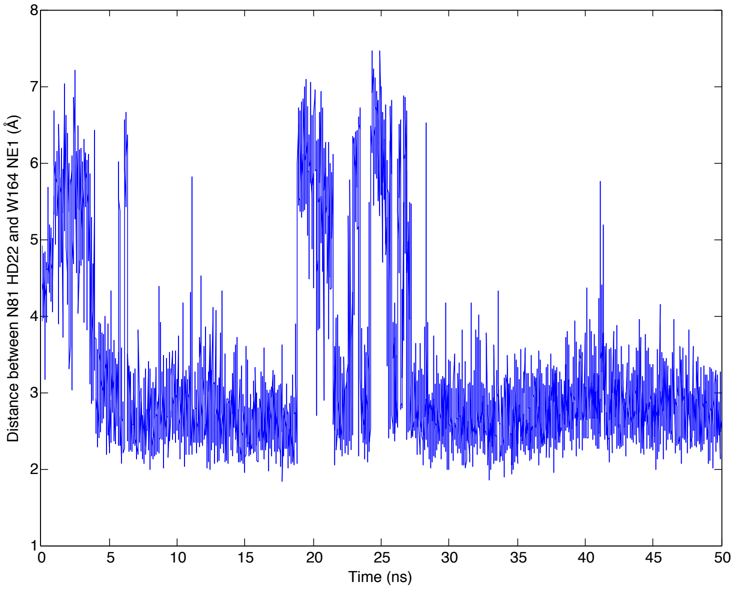


Figure C.8: The fluctuation of interatomic distance between N81$^{2.45}$ N-donated H atom and W164$^{4.50}$ N atom on their side chains during the 50 ns MD simulation of M59-bound predicted hSSTR5 structure.

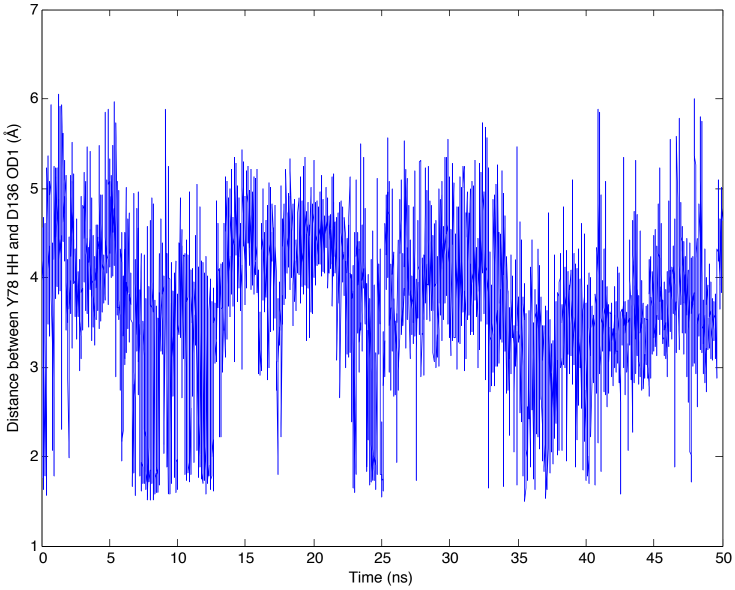


Figure C.9: The fluctuation of interatomic distance between Y78$^{2.43}$ O-donated H atom and D136$^{3.49}$ O atom on their side chains during the 50 ns MD simulation of M59-bound predicted hSSTR5 structure.

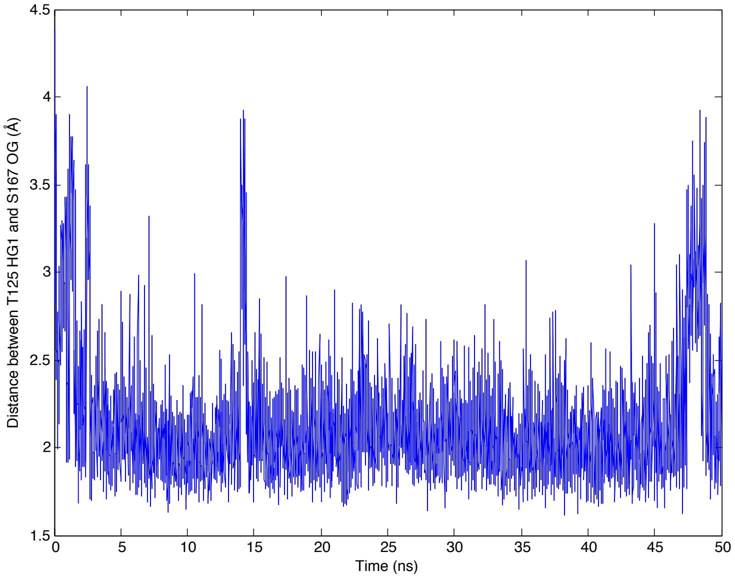


Figure C.10: The fluctuation of interatomic distance between T125$^{3.38}$ O-donated H atom and S167$^{4.53}$ O atom on their side chains during the 50 ns MD simulation of M59-bound predicted hSSTR5 structure.

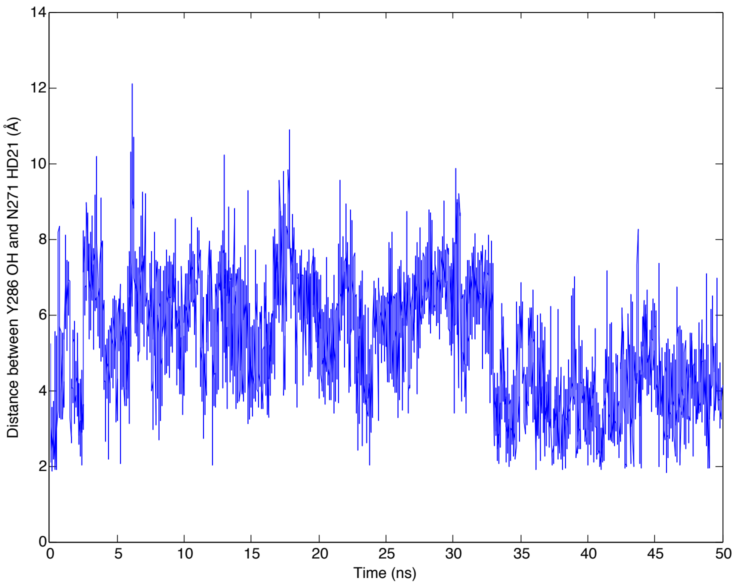


Figure C.11: The fluctuation of interatomic distance between N271$^{6.58}$ N-donated H atom and Y286$^{7.35}$ O atom on their side chains during the 50 ns MD simulation of M59-bound predicted hSSTR5 structure.

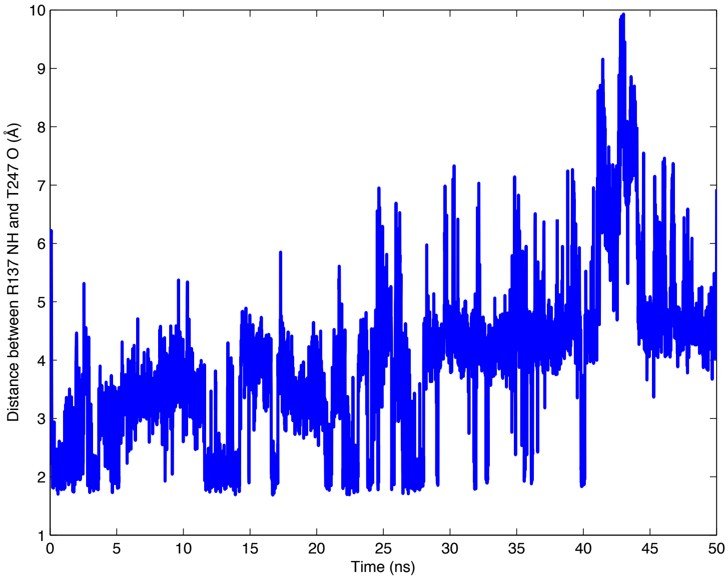


Figure C.12: The fluctuation of interatomic distance between R137$^{3.50}$ N-donated H atom and T247$^{6.34}$ O atom on their side chains during the 50 ns MD simulation of M59-bound predicted hSSTR5 structure. Visualization of this interaction is in Figure 4.6.

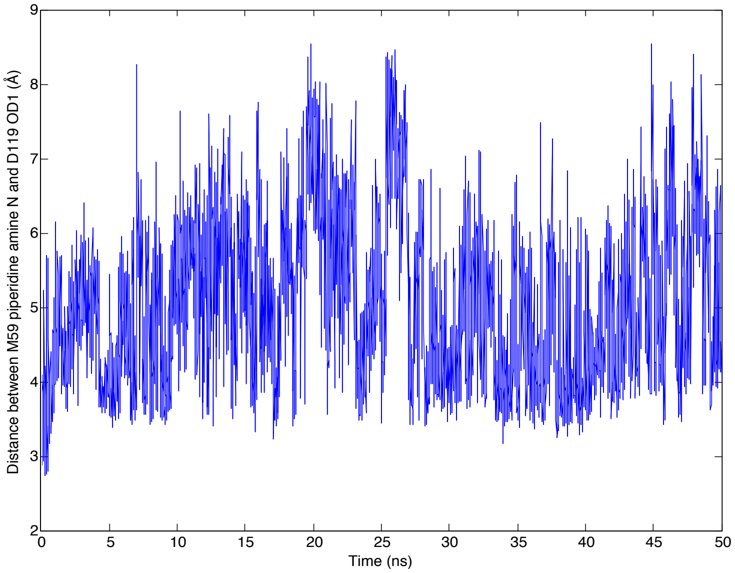


Figure C.13: The fluctuation of distance between the M59 piperidine amine N atom and D119$^{3.32}$ carboxylic acid O atom during the 50 ns MD simulation of M59-bound predicted hSSTR5 structure.

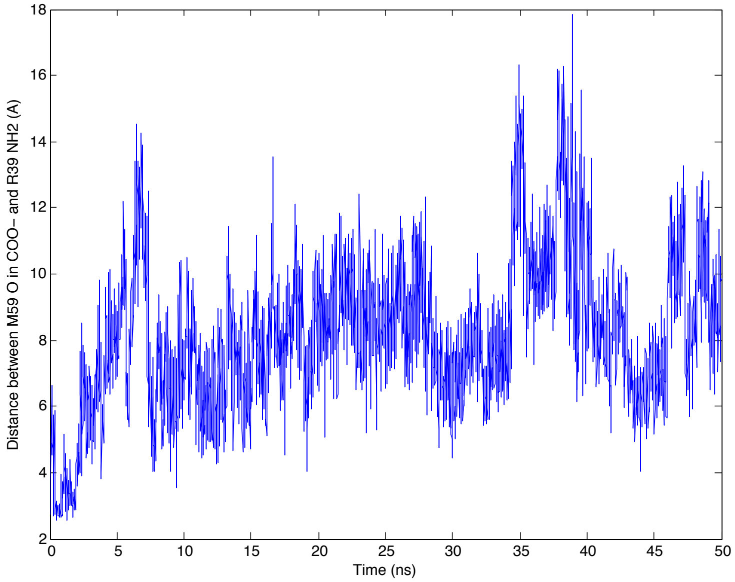


Figure C.14: The fluctuation of distance between M59 carboxylic acid O atom and R39$^{1.31}$ amine N atom during the 50 ns MD simulation of M59-bound predicted hSSTR5 structure.

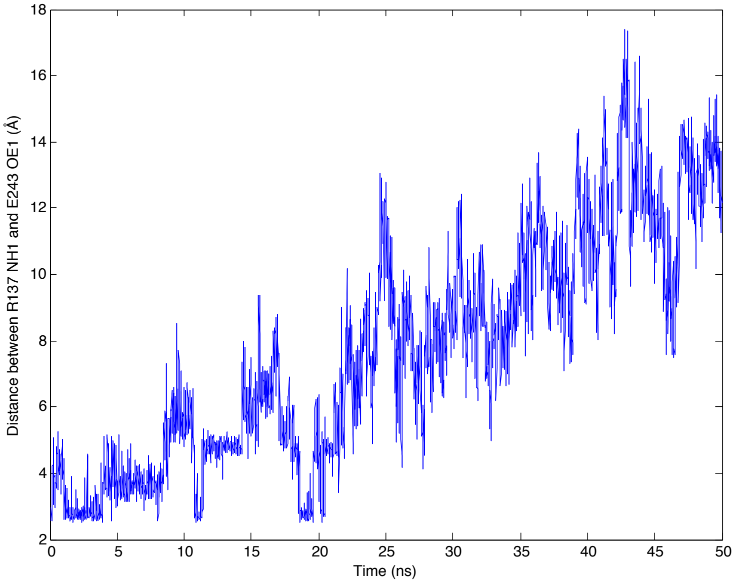


Figure C.15: The fluctuation of distance between the R137$^{3.50}$ side chain N atom and E243$^{6.30}$ side chain O atom during the 50 ns MD simulation of M59-bound predicted hSSTR5 structure. After 20 ns, the intracellular end of TM6 including E243$^{6.30}$ unwinds to accommodate the salt bridge between E243$^{6.30}$ and R241 on the loop, and E243$^{6.30}$ can be considered part of the loop.

Appendix D

Derivatives and Series Expansion for the New Form of ΔT

The new form of ΔT in GHA-QM is

$$\Delta T = O \cdot T + \frac{dd}{1 - S^2}$$

where

$$T \equiv \frac{3}{2}\frac{1}{s_1^2} + \frac{3}{2}\frac{1}{s_2^2} - \frac{6}{s_1^2 + s_2^2} + \frac{4R_{12}^2}{(s_1^2 + s_2^2)^2}.$$

The following shows the useful equations for its practical implementation.

D.1 Derivatives of the New Form

$$\frac{\partial \Delta T}{\partial R_{12}} = \frac{\partial O}{\partial R_{12}} T + O \frac{\partial T}{\partial R_{12}} + \frac{dd \cdot 2S}{(1 - S^2)^2} \frac{\partial S}{\partial R_{12}}$$

$$\frac{\partial \Delta T}{\partial s_1} = \frac{\partial O}{\partial s_1} T + O \frac{\partial T}{\partial s_1} + \frac{dd \cdot 2S}{(1 - S^2)^2} \frac{\partial S}{\partial s_1}$$

$$\frac{\partial \Delta T}{\partial s_2} = \frac{\partial O}{\partial s_2} T + O \frac{\partial T}{\partial s_2} + \frac{dd \cdot 2S}{(1 - S^2)^2} \frac{\partial S}{\partial s_2}$$

D.2 Series Expansion for the New Form

Define

$$a \equiv \frac{2s_1 s_2}{s_1^2 + s_2^2}, \; x \equiv \frac{R_{12}^2}{s_1 s_2}$$

When $S \to 1$, $s_1 \to s_2$ and $R_{12} \to 0$, we have

$$\begin{aligned}
\Delta T &\equiv \frac{a^3\exp(-ax)(\frac{3}{a}-3a+a^2x)\frac{1}{s_1s_2}+dd}{1-a^3\exp(-ax)} \\
&= \frac{\exp(-x)x\frac{1}{s_1s_2}+dd}{1-\exp(-x)} \text{ when } a \to 1 \\
&= \frac{x}{\exp(x)-1}\frac{1}{s_1s_2}+\frac{dd}{1-\exp(-x)} \\
&= \frac{x}{\sum_{k=0}^{\infty}\frac{x^k}{k!}-1}\frac{1}{s_1s_2}+dd\sum_{n=-1}^{\infty}\frac{x^nB_{1+n}(1)}{(1+n)!} \\
&= \frac{1}{s_1s_2}\left(1-\frac{x}{2}+\frac{x^2}{12}-\frac{x^4}{720}+O(x^6)\right)+dd\left(\frac{1}{x}+\frac{1}{2}+\frac{x}{12}-\frac{x^3}{720}+\frac{x^5}{30240}+O(x^6)\right) \\
&= \frac{1}{s_1s_2}\left(1-\frac{x}{2}+\frac{x^2}{12}+O(x^4)\right)+dd\left(\frac{1}{x}+\frac{1}{2}+\frac{x}{12}-\frac{x^3}{720}+O(x^4)\right) \\
&= \left(\frac{1}{s_1s_2}-\frac{R_{12}^2}{2s_1^2s_2^2}+\frac{R_{12}^4}{12s_1^3s_2^3}+O(x^4)\right)+dd\left(\frac{s_1s_2}{R_{12}^2}+\frac{1}{2}+\frac{R_{12}^2}{12s_1s_2}-\frac{R_{12}^6}{720s_1^3s_2^3}+O(x^4)\right)
\end{aligned}$$

$$\frac{\partial x}{\partial R_{12}} = \frac{2R_{12}}{s_1s_2} \equiv 2R$$

$$\frac{\partial x}{\partial s_1} = -\frac{R_{12}^2}{s_1^2s_2} \equiv -R^2s_2$$

$$\frac{\partial x}{\partial s_2} = -\frac{R_{12}^2}{s_1s_2^2} \equiv -R^2s_1$$

$$\frac{\partial \Delta T}{\partial R_{12}} = -\frac{R}{s_1s_2}+\frac{R^3}{3}+dd\left(-\frac{2R}{x^2}+\frac{R}{6}-\frac{x^2R}{120}\right)$$

$$\frac{\partial \Delta T}{\partial s_1} = -\frac{1}{s_1^2s_2}+\frac{R^2}{s_1}-\frac{R^4s_2}{4}+dd\left(\frac{s_2}{R_{12}}-\frac{x}{12s_1}+\frac{x^3}{240s_1}\right)$$