

Computational Predictions of G Protein-Coupled Receptor Structures and Binding Sites

Thesis by
Andrea Kirkpatrick

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



CALIFORNIA INSTITUTE OF TECHNOLOGY

Pasadena, California

2014

Defense Date: December 10, 2014

© 2015
Andrea Kirkpatrick
All Rights Reserved

Acknowledgements

First and foremost, I would like to thank my advisor, Professor William A. Goddard III, for his guidance over the past seven years. I would not be where I am today scientifically or personally without his help. He is an unquenchable source of novel ideas and his passion for science is astounding. His knowledge over the breadth of computational chemistry is truly an inspiration.

Next, I would like to thank Professor Ravinder Abrol for the immense amount of assistance he has provided during the course of my graduate degree. His patience knows no bounds that I have found. I could not have asked for a better mentor than Ravi.

I would also like to express my gratitude to the rest of the Biogroup. Adam Griffith has always been extremely helpful in resolving any issues with our programs and the Biogroup as a whole owes him a tremendous amount. Dr. SooKyung Kim has also been a fantastic source of guidance and encouragement. Dr. Caitlin Scott experienced much of my graduate school progress alongside me, and I am thankful for having had a peer going through the same experiences as myself. Additional people who have helped me in my graduate research include Professor Jiyoung Heo, Professor Bartosz Trzaskowski, and Dr. Claude Rogers. I am additionally thankful for the remaining members of the Biogroup's support and encouragement: Fan Liu, Sija Dong, Vaclav Cvicek, Matthew Gethers, Dr. William Ford, Dr. Jenelle Bray, Dr. Ismet Caglar Tanrikulu, and Dr. Heather Wiencko.

I would like to thank the members of my thesis committee: Professor Thomas Miller, Professor Dennis Dougherty, and Professor Stephen Mayo. Their critiques and suggestions have made me a better scientist.

Lastly, I would like to thank my family, friends, and pets for their support throughout my graduate school career. In particular I want to express my gratitude towards my mother, father, brother, and grandmother (Dr. Helena Kirkpatrick, Professor Theodore R. Kirkpatrick, Paul Kirkpatrick, and Zdena Plavec respectively) and the amazingly supportive Dr. Artur R. Menzeleev. It is no understatement to say that I would not have completed my degree without them.

My graduate work at the California Institute of Technology has been funded by Sanofi, Cargill, the Chemistry Graduate Program Fellowship, the Bing Scholarship, several teaching fellowships, the Materials and Process Simulation Center, the Center for Catalytic Hydrocarbon Functionalization, and NIH grants (R01NS071112, R01NS073115).

Abstract

G protein-coupled receptors (GPCRs) are the largest family of proteins within the human genome. They consist of seven transmembrane (TM) helices, with a N-terminal region of varying length and structure on the extracellular side, and a C-terminus on the intracellular side. GPCRs are involved in transmitting extracellular signals to cells, and as such are crucial drug targets. Designing pharmaceuticals to target GPCRs is greatly aided by full-atom structural information of the proteins. In particular, the TM region of GPCRs is where small molecule ligands (much more bioavailable than peptide ligands) typically bind to the receptors. In recent years nearly thirty distinct GPCR TM regions have been crystallized. However, there are more than 1,000 GPCRs, leaving the vast majority of GPCRs with limited structural information. Additionally, GPCRs are known to exist in a myriad of conformational states in the body, rendering the static x-ray crystal structures an incomplete reflection of GPCR structures. In order to obtain an ensemble of GPCR structures, we have developed the GEnSeMBLE procedure to rapidly sample a large number of variations of GPCR helix rotations and tilts. The lowest energy GEnSeMBLE structures are then docked to small molecule ligands and optimized. The GPCR family consists of five subfamilies with little to no sequence homology between them: class A, B1, B2, C, and Frizzled/Taste2. Almost all of the GPCR crystal structures have been of class A GPCRs, and much is known about their conserved interactions and binding sites. In this work we particularly focus on class B1 GPCRs, and aim to understand that family's interactions and binding sites both to small molecules and their native peptide ligands. Specifically, we predict the full atom structure and peptide binding site of the glucagon-like peptide receptor and the TM region and small molecule binding sites for eight other class B1 GPCRs: CALRL, CRFR1, GIPR, GLR, PACR, PTH1R, VIPR1, and VIPR2. Our class B1 work reveals multiple conserved interactions across the B1 subfamily as well as a consistent small molecule binding site centrally located in the TM bundle. Both the interactions and the binding sites are distinct from those seen in the more well-characterized class A GPCRs, and as such our work provides a strong starting point for drug design targeting class B1 proteins. We also predict the full structure of CXCR4 bound

to a small molecule, a class A GPCR that was not closely related to any of the class A GPCRs at the time of the work.

Table of Contents

Acknowledgements	iii
Abstract	v
Table of Contents	vii
List of Figures and Tables	ix
Chapter I: Introduction	1
Background	2
Statement of the Problem	6
Purpose of the Study	7
Outline of the Thesis	7
References	9
Chapter II: Structure and Binding Site Prediction Methodology	10
GEnSeMBLE: Generating an Ensemble of GPCR Structures	11
Small Molecule Docking and Full Structure Optimization	16
References	20
Chapter III: The Predicted Structure of Peptide-Bound Glucagon-Like Peptide-1 Receptor, a Class B1 G Protein-Coupled Receptor	23
Abstract	24
Introduction	24
Methods	26
Results and Discussion	37
Conclusion	53
References	55
Chapter IV: Transmembrane Region and Small Molecule Binding Site Predictions for Nine Class B1 G Protein-Coupled Receptors	60
Abstract	61
Introduction	61
Methods	64
Results and Discussion	66
Conclusion	82
Appendix	83
References	91

Chapter V: The Full Structure and Small Molecule Binding Site	
Prediction of CXCR4, a Class A G Protein-Coupled Receptor	94
Abstract	95
Introduction.....	95
Methods	98
Results and Discussion.....	99
Conclusion	108
References.....	110

List of Figures and Tables

Chapter I

Figure 1. Diversity of G-protein-coupled receptor cell signaling pathways.....	3
Figure 2. Typical (A) class A and (B) B1 GPCR endogenous ligand binding sites.	4
Figure 3. GPCR x-ray crystal structures shown on the GPCR phylogenetic tree	5
Figure 4. Class A conserved interactions.....	6

Chapter II

Figure 1. GEnSeMBLE overview.....	12
Figure 2. Bihelix methodology.	15
Figure 3. Scoring complexes in Superbihelix.....	16
Figure 4. GPCR small molecule binding site generation and full structure optimization procedure overview.....	17

Chapter III

Figure 1. Creating the GLP1R/Exe4 bundle.....	28
Table 1. Top ten Bihelix/Combihelix structures for GLP1R.	36
Table 2. Top two Superbihelix/Supercombihelix structures.....	37
Table 3. Interhelical hydrogen bonds for the GLP1R/Exe4 structure.	38
Figure 2. The TM2-TM3-TM6 and TM1-TM2-TM7 conserved hydrogen bond networks.....	41
Figure 3. Overview of the ligand binding site's hydrophobic and hydrophilic interactions.....	42
Table 4. Polar interactions between GLP1R and Exendin-4.....	44
Figure 4. Exe4 hydrogen bonds with the N-terminus and transmembrane region of GLP1R.....	44
Table 5. Top ten hydrophobic interactions between GLP1R and Exe4.....	46
Figure 5. Hydrophobic interactions between GLP1R and Exendin-4 in the N-terminus and extracellular loop 1	46
Table 6. Full unified cavity analysis of the GLP1R/Exe4 Binding Site	47
Table 7. Residues in the binding site with their mutation data.	50
Table 8. Suggested mutations for GLP1R or Exendin-4.....	52

Chapter IV

Table 1. The nine class B1 GPCRs and their small molecule ligands.....	62
Figure 1. Phylogenetic tree for the class B1 receptors..	64
Figure 2. WebLogo representation of the sequence identities between the transmembrane regions of the nine class B1 predicted structures	66
Table 2. Summary of structures input to Superbihelix from Bihelix/Combihelix	67
Table 3. Superbihelix/Supercombihelix results for the nine predicted structures.....	68
Table 4. Class B1 conserved interactions	69
Figure 3. Class B1 conserved interactions, as found in the GLR predicted structure	70
Table 5. Class B1 predicted structure interhelical interactions.....	70

Figure 4. Comparison of the class A and class B1 GPCR binding sites.....	74
Figure 5. CALRL bound to MK-0974.....	75
Figure 6. CRFR1 bound to CP-376395	76
Figure 7. GLR bound to NNC0640.	77
Figure 8. GIPR bound to NNC0640.	78
Figure 9. GLP1R bound to T0632.	79
Figure 10. PACR bound to Molecule 1.....	79
Figure 11. PTH1R bound to SW106.....	80
Figure 12. VIPR1 bound to Molecule 4.....	81
Figure 13. VIPR2 bound to Molecule 6.....	81
Chapter IV Appendix	
Table A1. Top 10 structures from Bihelix/Combihelix for CALRL	83
Table A2. Top 10 structures from Bihelix/Combihelix for CRFR1	83
Table A3. Top 10 structures from Bihelix/Combihelix for GLR	84
Table A4. Top 10 structures from Bihelix/Combihelix for GIPR	84
Table A5. Top 10 structures from Bihelix/Combihelix for GLP1R.....	84
Table A6. Top 10 structures from Bihelix/Combihelix for PACR.....	85
Table A7. Top 10 structures from Bihelix/Combihelix for PTH1R.....	85
Table A8. Top 10 structures from Bihelix/Combihelix for VIPR1	85
Table A9. Top 10 structures from Bihelix/Combihelix for VIPR2	86
Table A10. Superbihelix/Supercombihelix results for CALRL.....	86
Table A11. Superbihelix/Supercombihelix results for CRFR1	87
Table A12. Superbihelix/Supercombihelix results for GLR.....	87
Table A13. Superbihelix/Supercombihelix results for GIPR.....	88
Table A14. Superbihelix/Supercombihelix results for GLP1R	88
Table A15. Superbihelix/Supercombihelix results for PACR	89
Table A16. Superbihelix/Supercombihelix results for PTH1R	89
Table A17. Superbihelix/Supercombihelix results for VIPR1.....	90
Table A18. Superbihelix/Supercombihelix results for VIPR21.....	90
Chapter V	
Figure 3. Inhibition of HIV infection by chemokines	97
Figure 4. CXCR4 inhibitor 1t.....	97
Figure 5. Two possible alignments of TMs 2 and 4 of β 1 and CXCR4	100
Table 1. Superbihelix/Combihelix results for the lowest ranked templates	101
Table 2. Top Supercombihelix/Supercombihelix structures.....	101
Table 3. RMSD matrix of the top 10 structures from Supercombihelix.....	102
Figure 6. Binding pose and unified cavity analysis for the rank 1 protein-ligand complex.....	103
Figure 8. Interhelical interactions in the rank 1 CXCR4 structure... ..	104
Figure 9. Comparisons of the predicted and x-ray crystal CXCR4 structures.	105
Table 4. Top Bihelix/Combihelix structures for the CXCR4 crystal helices	106
Table 5. Top Superbihelix/SuperCombihelix structures for the CXCR4 crystal helices.	106
Figure 10. Binding site of the GenDock docking of the crystal 1t to the crystal protein.	107

Figure 11. Comparison of four class A GPCR crystal TM bundles 108

Chapter I:
Introduction

Background

G Protein Coupled Receptors

G protein coupled receptors (GPCRs) are the largest protein superfamily in mammalian genomes.¹ They are integral membrane proteins that enable cells to respond to external stimuli. Because of their vital role in cellular signaling networks, GPCRs are involved in many diseases, and are the target of approximately 40% of all prescription pharmaceuticals on the market.²

GPCRs are activated by intercellular signaling molecules such as hormones, neurotransmitters, ions, and chemokines, or they sense light, or odorants, or taste substances (**Figure 1**). Upon interaction with their signaling molecules, GPCRs undergo conformational changes, which catalyze GDP-GTP (guanosine triphosphate-guanosine diphosphate) exchange on the heterotrimeric G proteins.³ This starts a signaling cascade through which the G proteins affect cell metabolism in two major ways: regulation (activation/inhibition) of cyclic AMP (adenosine monophosphate) concentration or stimulation of inositol phospholipid hydrolysis.⁴ Each of these responses cause further cellular reactions, the specifics of which depend on the cell signaling pathways involved.

GPCRs share a common architecture of seven transmembrane (TM) domains. These α -helical domains are roughly 20-30 amino acids long and are arranged in a tightly packed bundle.⁵ The N terminus of the protein is located in the extracellular space while the C terminus is in the intracellular area. The seven transmembrane helices are connected by six alternating intracellular (IC1-3) and extracellular (EC1-3) loops.

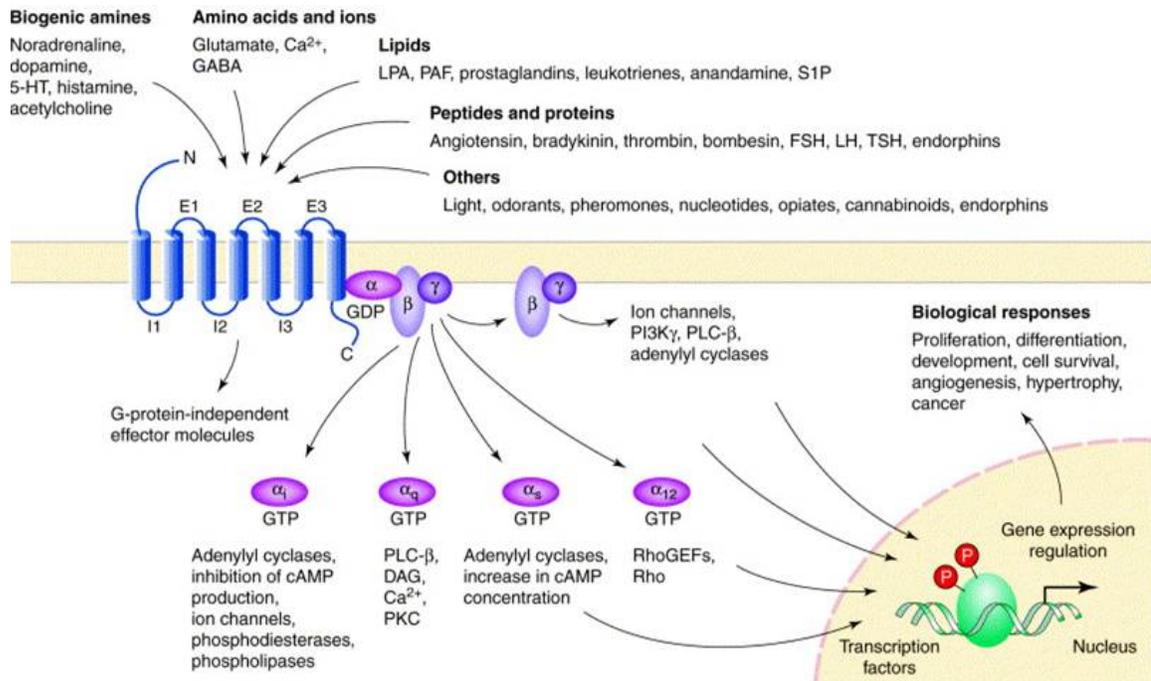


Figure 1. Diversity of G-protein-coupled receptor cell signaling pathways.

Abbreviations: DAG: diacylglycerol, FSH: follicle-stimulating hormone, GEF: guanine nucleotide exchange factor, LH: luteinizing hormone, LPA: lysophosphatidic acid, PAF: platelet-activating factor, PI3K: phosphoinositide 3-kinase, PKC: protein kinase C, PLC: phospholipase C, S1P: sphingosine-1-phosphate, TSH: thyroid-stimulating hormone.⁶

Classes of GPCRs

A phylogenetic analysis of GPCRs splits the family into five classes: class A (rhodopsin-like), class B1 (secretin-like), class B2 (adhesion-like), class C (glutamate-like), and Frizzled/Taste2.⁷ The classes share little to no sequence homology, but share similar transmembrane domain architectures and interact with the same G proteins. The N-terminal architecture of the five classes varies wildly, as do the native ligand binding sites. Class A GPCRs have short unstructured N-termini (on the order of 30 residues) and their native ligands tend to be small molecules that interact within the TM region of the proteins (**Figure 2A**). Class B1 receptors have longer, structured, N-termini (on the order of 150 residues), and their endogenous ligands are peptides whose helical portion interacts with the N-terminus while a flexible portion of the ligand interacts with the TM region (**Figure 2B**). Class B2 receptors have very large extracellular domains (on the order of 950 residues) which contain known protein motifs, and are coupled to the TM region by a GPCR-autoproteolysis inducing domain. Their native ligands primarily

interact with their large N-termini. Class C GPCRs have large N-termini (on the order of 600 residues) which consist of a cysteine-rich domain (CRD) above the TM region as well as a “Venus fly trap” structured region above the CRD which interacts with their small molecule ligands. Frizzled receptors have a large N-terminus (around 300 residues) which have a CRD. Their protein ligands are proposed to interact with both the CRD and TM regions. Taste2 receptors architectures are architecturally similar to class A GPCRs: they have short N-termini (around 10 residues or less) and their native ligands bind within the TM region.

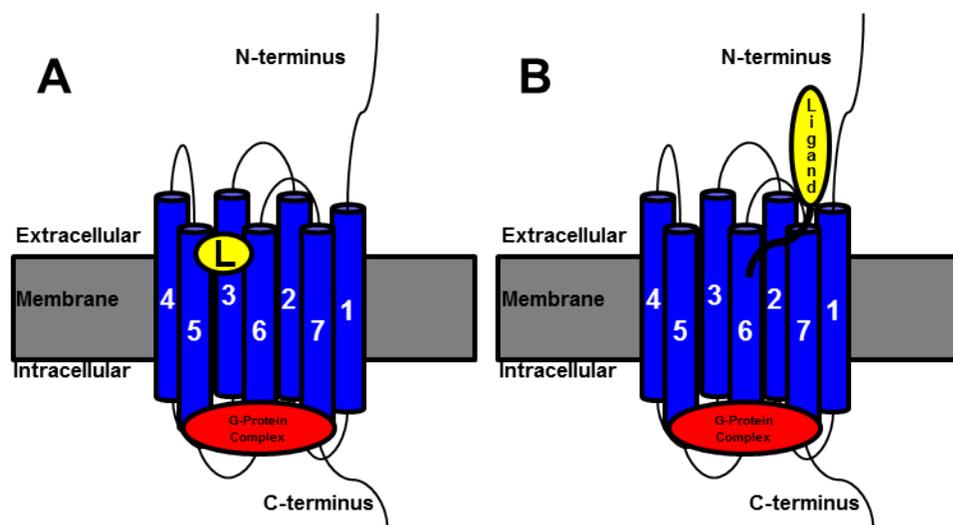


Figure 2. Typical (A) class A and (B) B1 GPCR endogenous ligand binding sites. The N-termini of the class B1 receptors have several α -helices and/or β -sheets, while the N-termini of the class A GPCRs are unstructured and approximately 100 residues shorter.

GPCR Crystal Structures

To date, there are approximately thirty GPCR TM region structures (**Figure 3**).⁸ Most have a single structure, although a few class A GPCRs have several, including both their active and inactive forms. The majority (22) of the x-ray crystal structures are from the class A GPCR family, with two structures from the class B1 family, two from the class C family, and one from the Frizzled/TAS2 family.⁸ In order to obtain these structures, modifications are often made to the wild type protein including residue mutations or insertion of large structured proteins (T4 Lysozyme or b₅₆₂RIL) to the intracellular loops or N-terminus.⁹

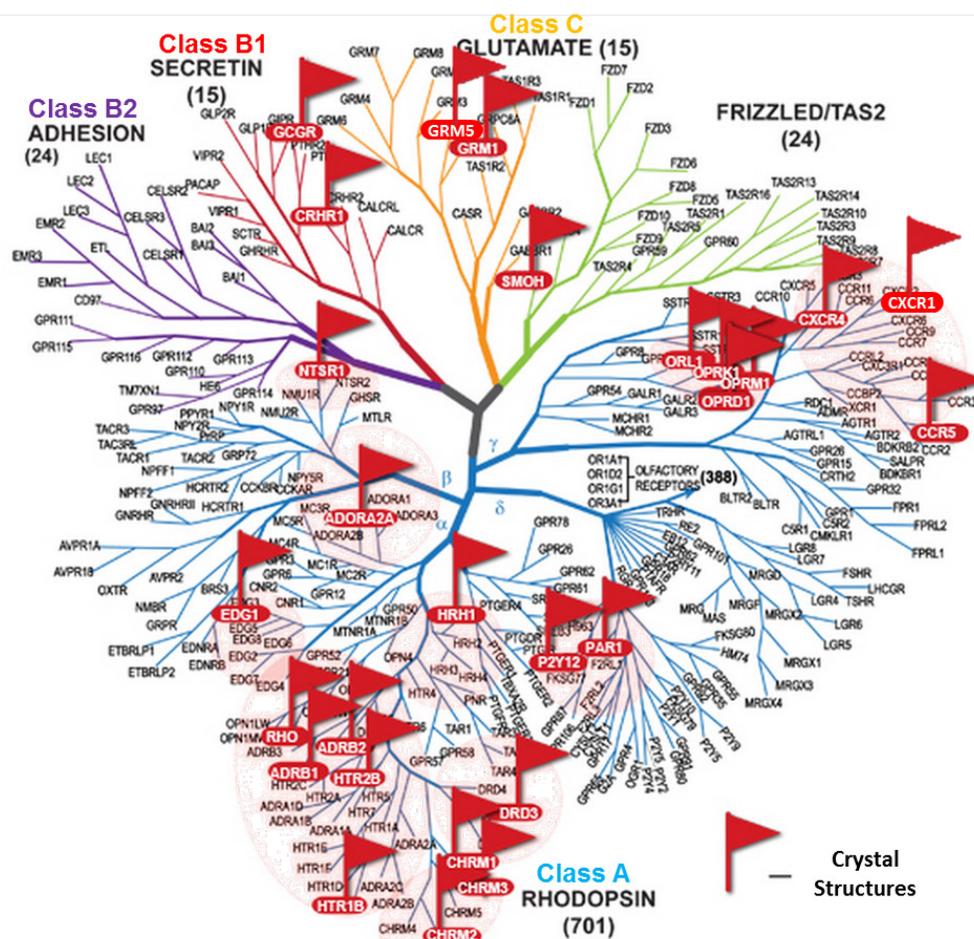


Figure 3. GPCR x-ray crystal structures shown on the GPCR phylogenetic tree. Note that the majority (22) of the structures are from the class A GPCR family (blue tree), with two structures from the class B1 family (red tree), two from the class C family (orange tree), and one from the frizzled/TAS2 family (green tree).¹⁰

Due to the larger number of x-ray crystal structures available for class A GPCRs, several conserved hydrogen bond networks have been discovered. These motifs include hydrogen bond networks between residues on TMs 1, 2, and 7 and TMs 2, 3, and 4 which are highly conserved among Class A GPCRs (**Figure 4A**). Additionally, inactive crystal structures show an interaction between TMs 3 and 6 (**Figure 4B**).¹¹ To denote residue locations with respect to the TM region's most conserved residue, Ballesteros numbering is utilized.¹² The most conserved residue for each helix is designated by its helix number followed by .50 (**Figure 4C**). Other residues are specified by counting up or down from 50. The TM 1-2-7 interaction is thus shown: N1.50-D2.50-N7.49, while the

TM 2-3-4 interaction is S/N/T3.42-S/N/T2.45-W4.50. The inactive TMs-TM6 interaction is usually R3.50-E6.30.

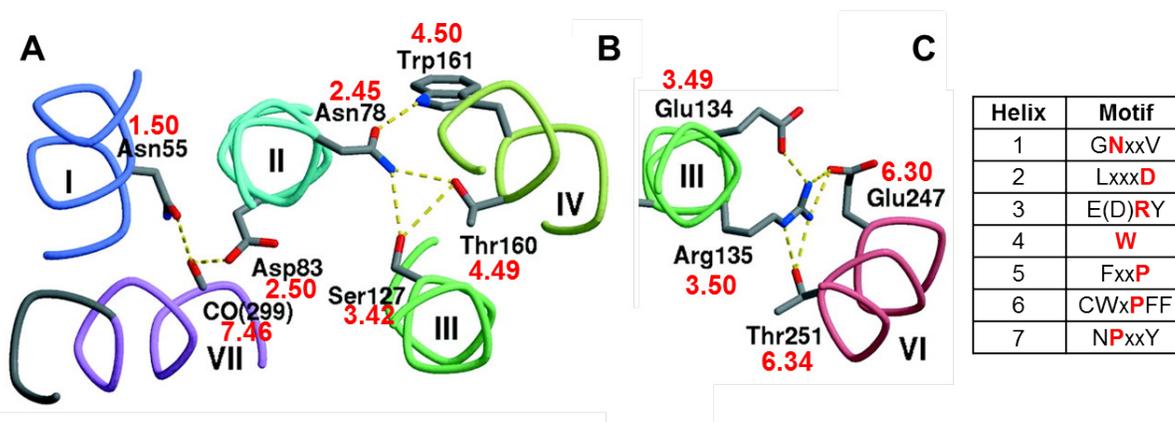


Figure 4. Class A conserved interactions. (A) The TM region 1-2-7 (N1.50-D2.50-N7.49) and TM region 2-3-4 (S/N/T3.42-S/N/T2.45-W4.50) conserved hydrogen bond networks of Class A GPCRs.¹³ (B) The conserved TM region 3-6 (R3.50-E6.30) interaction which is indicative of an inactive structure. The pictured residues are for bovine rhodopsin, and the numbers in red are the Ballesteros numbers. (C) The conserved residue motifs of Class A GPCRs. The residues marked in red are those designated as .50 in Ballesteros numbering.^{12, 14-15}

Statement of the Problem

Full-atom structures of GPCRs provide great aid to designing drugs that target them. However, such information can only be obtained in two ways: x-ray crystallography or computational predictions. Unfortunately, GPCRs, like other membrane proteins, are notoriously difficult to crystallize. While there has been a marked increase in the number of GPCRs crystallized in the last few years, there are still fewer than thirty distinct receptors with transmembrane region crystal structures out of the over 1,000 human GPCRs.⁸ In order to crystallize these GPCRs modifications are made to the wild type protein including residue mutations or insertion of large structured proteins (T4 Lysozyme or b₅₆₂RIL) to the intracellular loops or N-terminus.⁹ These insertions may potentially impact the ability of the crystal structure to accurately represent nature. Additionally, the majority (22) of the crystallized GPCR structures are of class A GPCRs, leaving the other classes, which share no significant sequence homology with class A, significantly less well understood. Finally, GPCRs exist in a myriad of activation states in

the body, depending on which ligands the protein is interacting with and its activation state.¹⁶ X-ray crystal structures only provide a snapshot of GPCR structures in specific crystallizable environments. Therefore, in order to obtain full-atom structural information about an ensemble of GPCR structures, one must turn to computational methods.

Purpose of the Study

In this work, we use computational techniques to predict the structures and binding sites of GPCRs, with a focus on class B1 GPCRs and transmembrane region predictions. The GEnSeMBLE (GPCR Ensemble of Structures in Membrane BiLayer Environment) procedure is used to optimize the transmembrane helix bundle. Specifically, the possible helix rotations and tilts are explored in great detail. GenDock is then utilized to generate small molecule binding site information for the GPCR targets of interest. The GPCR structures and binding sites are analyzed with an emphasis on conserved interactions in the class B1 family. The peptide binding site to the full TM region and N-terminus of the class B1 GPCR GLP1R is generated and compared to class A GPCRs and experimental data. Transmembrane region structure and small ligand binding sites across the class B1 family are predicted and analyzed for these receptors: CALRL, CRFR1, GLR, GIPR, GLP1R, PACR, PTH1R, VIPR1, and VIPR2. Finally, the structure and small molecule binding site of CXCR4, a class A GPCR, is also predicted.

Outline of the Thesis

Chapter II: Methodology overview focusing on the GEnSeMBLE procedure of predicting an ensemble of GPCR structures and the GenDock hierarchical docking program.

Chapter III: Full structural prediction of GLP1R, a class B1 GPCR, bound to Exendin-4, a peptide ligand. Interactions between conserved residues are analyzed, as is the peptide binding site to both the N-terminus and TM region. Experimental data is used to validate the prediction.

Chapter IV: Transmembrane region and small-molecule binding site predictions of nine class B1 GPCRs: CALRL, CRFR1, GLR, GIPR, GLP1R, PACR, PTH1R, VIPR1, and VIPR2. Conserved interactions and binding sites are compared across the B1 subfamily of GPCRs. Comparisons to the two x-ray crystal class B1 structures recently released are made, and our structure and binding site prediction methods are used on these two receptors as well.

Chapter V: Full structural prediction of CXCR4, a class A GPCR, bound to the small molecule ligand 1t. Comparisons are made to the later published CXCR4 crystal structure.

References

1. Katritch, V.; Cherezov, V.; Stevens, R. C., Structure-Function of the G Protein-Coupled Receptor Superfamily. *Annu Rev Pharmacol Toxicol* **2013**, *53*, 531-56.
2. Filmore, D., It's a GPCR World. *Modern Drug Discovery* **2004**, *7* (11), 24-28.
3. Schoneberg, T.; Schultz, G.; Gudermann, T., Structural Basis of G Protein-Coupled Receptor Function. *Mol Cell Endocrinol* **1999**, *151* (1-2), 181-93.
4. Vauzequin, G. a. v. M., B, *G Protein-Coupled Receptors*. Wiley and Sons: England, 2007; p 147.
5. Vaidehi, N.; Floriano, W. B.; Trabanino, R.; Hall, S. E.; Freddolino, P.; Choi, E. J.; Zamanakos, G.; Goddard, W. A., 3rd, Prediction of Structure and Function of G Protein-Coupled Receptors. *Proc Natl Acad Sci U S A* **2002**, *99* (20), 12622-7.
6. Marinissen, M. J.; Gutkind, J. S., G-Protein-Coupled Receptors and Signaling Networks: Emerging Paradigms. *Trends in Pharmacological Sciences* **2001**, *22* (7), 368-376.
7. Lagerstrom, M. C.; Schioth, H. B., Structural Diversity of G Protein-Coupled Receptors and Significance for Drug Discovery. *Nat Rev Drug Discov* **2008**, *7* (4), 339-57.
8. Yang, J., Zhang, Y. Gpcrs-Exp: A Database for Experimentally Solved GPCR Structures. <http://zhanglab.ccmb.med.umich.edu/GPCR-EXP> (accessed November 11, 2014).
9. Chun, E.; Thompson, A. A.; Liu, W.; Roth, C. B.; Griffith, M. T.; Katritch, V.; Kunken, J.; Xu, F.; Cherezov, V.; Hanson, M. A.; Stevens, R. C., Fusion Partner Toolchest for the Stabilization and Crystallization of G Protein-Coupled Receptors. *Structure* **2012**, *20* (6), 967-976.
10. GPCR Network. <http://gpcr.scripps.edu/index.html> (accessed November 19, 2014).
11. Kobilka, B. K., G Protein Coupled Receptor Structure and Activation. *Biochimica et Biophysica Acta (BBA) - Biomembranes* **2007**, *1768* (4), 794-807.
12. Ballesteros, J. A.; Weinstein, H., Integrated Methods for Modeling G-Protein Coupled Receptors. In *Methods of Neuroscience*, 1995; Vol. 270, pp 366-428.
13. Palczewski, K.; Kumasaka, T.; Hori, T.; Behnke, C. A.; Motoshima, H.; Fox, B. A.; Trong, I. L.; Teller, D. C.; Okada, T.; Stenkamp, R. E.; Yamamoto, M.; Miyano, M., Crystal Structure of Rhodopsin: A G Protein-Coupled Receptor. *Science* **2000**, *289* (5480), 739-745.
14. Lagerstrom, M. C.; Schioth, H. B., Structural Diversity of G Protein-Coupled Receptors and Significance for Drug Discovery. *Nat Rev Drug Discov* **2008**, *7* (4), 339-357.
15. Palczewski, K.; Kumasaka, T.; Hori, T.; Behnke, C. A.; Motoshima, H.; Fox, B. A.; Le Trong, I.; Teller, D. C.; Okada, T.; Stenkamp, R. E.; Yamamoto, M.; Miyano, M., Crystal Structure of Rhodopsin: A G Protein-Coupled Receptor. *Science* **2000**, *289* (5480), 739-45.
16. Kenakin, T.; Miller, L. J., Seven Transmembrane Receptors as Shapeshifting Proteins: The Impact of Allosteric Modulation and Functional Selectivity on New Drug Discovery. *Pharmacol Rev* **2010**, *62* (2), 265-304.

Chapter II:

Structure and Binding Site Prediction Methodology

The following chapter describes the technical details of the methodology used in the GEnSeMBLE and GenDOCK procedures. Exceptions and relevant details for each specific structural prediction will be noted in its chapter.

GEnSeMBLE: Generating an Ensemble of GPCR Structures

The GEnSeMBLE (GPCR Ensemble of Structures in Membrane BiLayer Environment) procedure is an *ab initio* methodology for predicting the multiple conformations of the different conformation states of a GPCR.¹⁻⁵ We refer to this method as a first principles one due to its minimal reliance on structural information from experimental data. The vast majority of the procedure is based solely from the amino acid sequence of the protein. This allows the methodology to be used on proteins of low sequence homology to the thirty experimentally determined transmembrane region GPCR structures out of the over 1,000 total GPCRs.⁶ This methodology is particularly useful when predicting the structures of non-class A GPCRs, since the vast majority of the crystallized structures are those of class A GPCRs, but there is little to no sequence homology between the classes.

One of the unique strengths of the GEnSeMBLE procedure is its ability to predict an ensemble of low-energy structures of GPCRs. GPCRs exist in a myriad of activation states in the body, depending on which ligands the protein is interacting with and its activation state.^{4, 7-8} X-ray crystal structures only provide a snapshot of GPCR structures in specific crystallizable environments. Additionally, they often have mutations or added cofactors to aid in the crystallization process, such as the addition of the T4 lysozyme to one of the intracellular loops or the BRIL protein to an intracellular loop or in place of the N-terminus of the protein.⁹ These cofactors, mutations, and the crystal conditions themselves may bias GPCR crystal structures away from their real structures in nature. The GEnSeMBLE methodology has none of these effects. This fact, plus its ability to efficiently predict a myriad of structures of different activation levels, makes the GEnSeMBLE procedure ideal for GPCR structure prediction.

The GEnSeMBLE methodology focuses on the prediction of the seven transmembrane (TM) helix bundle of GPCRs. This region is the primary area of interaction between small molecule ligands and GPCRs, making a full-atom

understanding of the TM structure essential for rational drug design. Additionally, many more crystal structures exist of the water-soluble portions of GPCRs, allowing homology modeling to those areas to be more effective.

The GEnSeMBLE procedure consists of two main steps (**Figure 1**): First, the TM region locations and lengths are determined via homology modeling and then the TM bundle is generated. Next, an ensemble of structures is generated via helix angle optimizations in BiHelix/Combihelix and Superbihelix/SuperCombihelix. The ensemble of structures generated are then used for docking in the GenDock methodology.

The GEnSeMBLE methodology has been validated by predicting several crystallized GPCR structures, and we find that our predicted structures are accurate.¹⁻³ For example, Abrol *et al.* predicted the inactive human adenosine A_{2A} receptor (iA_{A2}AR) conformation starting from the inactive human β 2 adrenergic receptor (i β ₂AR) template using the GEnSeMBLE method. After sampling, the iA_{A2}AR predicted structure had a small 1.4 Å backbone root-mean-squared deviation (RMSD) with respect to the actual iA_{A2}AR crystal.³

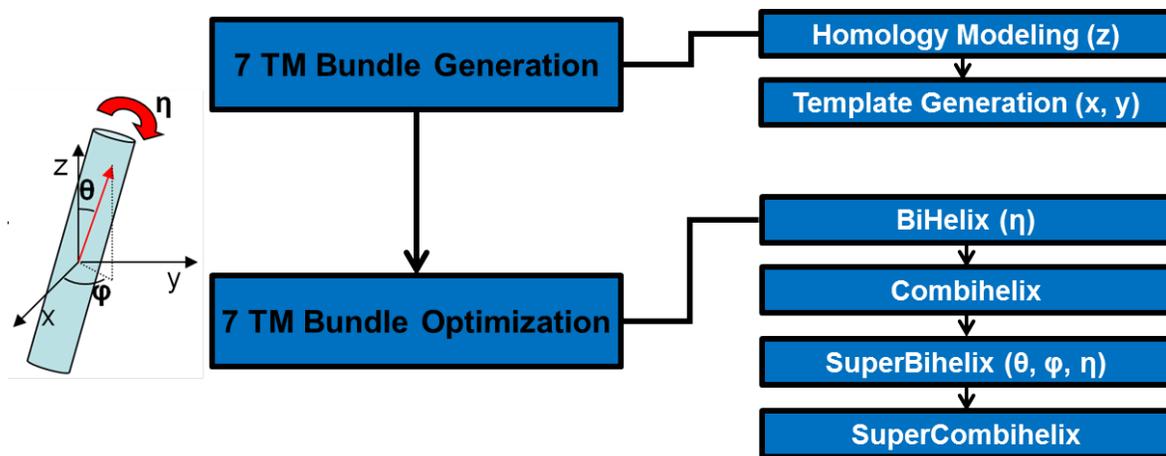


Figure 1. GEnSeMBLE overview. The GPCR Ensemble of Structures in Membrane BiLayer Environment GPCR structure prediction method samples a large variety of helix rotations and tilts, producing an ensemble of GPCR structures.

7 TM Helix Bundle Generation

The first step in GPCR structure prediction is the determination of the lengths and locations of the seven TM helices. This is achieved by homology modeling to the x-ray crystal structures with the most closely related sequence to the protein of interest, as determined by a ClustalW¹⁰ sequence alignment between the target GPCR and all known GPCR crystal sequences. Once the target and template sequences are aligned, the helical regions are directly determined from those of the crystallized GPCR.

Before any further steps are taken in the homology modeling process, the selected x-ray crystal structures must be prepared for use. This entails removing the GPCR from its crystal environment, including any fusion proteins. If any residues are missing or unresolved in the crystal structure, they are added in via Schrödinger's protein building tools¹¹⁻¹². The entire bundle is then minimized to 0.5RMS force in MPSim¹³ using the Dreiding¹⁴ forcefield and conjugate-gradient minimization (all future minimizations mentioned will use this method unless stated otherwise).

Once the full crystal protein crystal template structure is relaxed, its helical shapes and locations (x, y in **Figure 1**) are used to generate a starting conformation for the target protein's TM bundle. To do this, the helical portions are directly taken from the crystal structure. The crystal residues are then mutated to that of the target protein according to the previous ClustalW sequence alignment using the side chain optimization program SCREAM¹⁵. The individual helix backbones are then minimized. After the helices have been built they are oriented with respect to each other via the crystal x and y locations. The z location is taken from the crystal structure as determined by the Orientations of Proteins in Membranes (OPM) database.¹⁶

At the end of this procedure we have a target protein whose helical shapes and relative locations are taken directly from that of the GPCR crystal structure (or more likely, several of these, one for each crystal template used). The side chains have been optimized by SCREAM, and the helix backbones are relaxed slightly from their initial individual minimization.

7 TM Bundle Optimization

After TM bundles are generated, we optimize their angles. First, Bihelix/Combihelix¹ is used to coarsely optimize the helix rotations (η in **Figure 1**), then Superbihelix/SuperCombihelix⁵ more finely optimize the helix rotations as well as their tilts (θ , φ in **Figure 1**). All samplings of the rotations and tilts of the helices are performed with respect to the least moment of inertia vector for the helix (red arrow in **Figure 1**), obtained by diagonalizing the moment of inertia matrix for the helix using only heavy backbone atoms.¹

Bihelix/Combihelix

Bihelix involves sampling η angles in 30° increments to score 12^7 (~35 million) combinations of angles. This procedure independently considers the 12 pairs of helices that are close enough to directly interact (1-2, 1-3, 1-7, 2-3, 2-4, 2-7, 3-4, 3-5, 3-6, 3-7, 4-5, 5-6, and 6-7, pictured in **Figure 2**). For each of the 12^2 combinations of each pair, the side chains are optimized using SCREAM and are minimized for 10 steps. Summing the total intra- and interhelical values from the 12^3 pairwise interaction energies leads to an energy estimate for all ~35 million bundles (a more thorough discussion of the energy scoring can be found in our 2012 Bihelix paper, along with validations of the methodology¹). The Combihelix procedure constructs the 2,000 lowest total energy bundles and optimizes them. The energy is evaluated after SCREAM and 10 steps of minimization. The bundles with the lowest averaged charged interhelical, charged total, neutral interhelical, and neutral total energy ranks continue on to Superbihelix. If bundles with different helix lengths are being considered, the average charged interhelical and neutral interhelical energy ranks are used instead. The neutral energies mentioned here are those obtained by transferring the hydrogen of each charged residue interaction from the acceptor to the donor and a modified Dreiding force field. This technique decreases the effect of long-range Coulombic interactions between charged groups.⁵

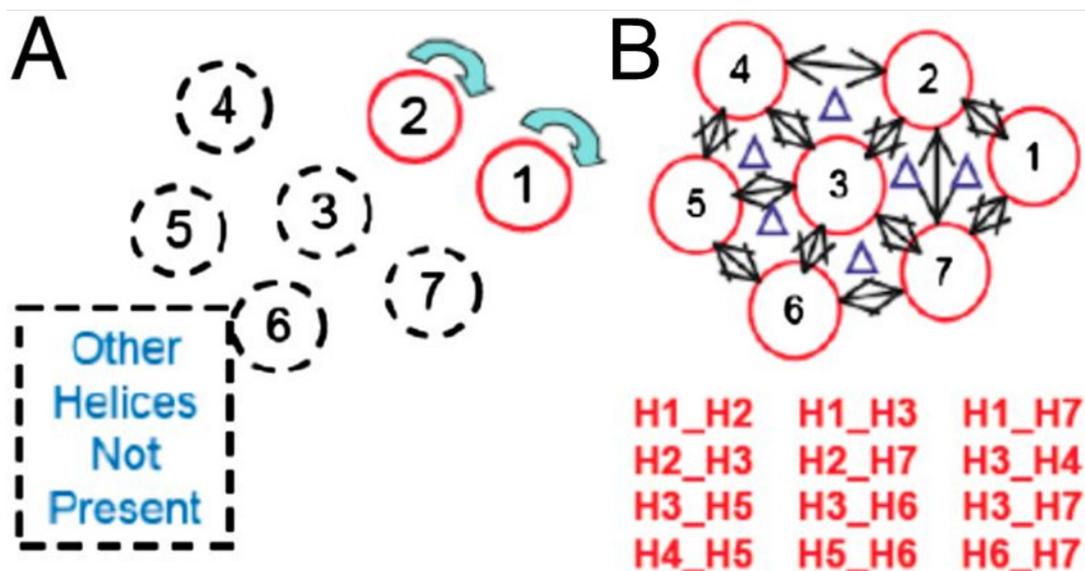


Figure 2. Bihelix methodology. (A) In Bihelix, helix rotations are sampled two at a time for the (B) 12 GPCR nearest neighbor pairs.⁵

SuperBihelix/SuperCombihelix

The SuperBihelix methodology samples the helix angles $\theta = -10/0/10$, $\phi = -30/-15/0/15/30$, and $\eta = -30/-15/0/15/30$ of the structure generated in the previous Bihelix/Combihelix step. This leads to the sampling of $(3 \times 5 \times 5)^7$ or ~ 13 trillion structures. To reduce computational time, the seven-helix bundle is further partitioned into four QuadHelix bundles (**Figure 3**). A more detailed description of the energy scoring and approximations used can be found in our 2014 paper on SuperBihelix, along with validations of the methodology.⁵ Once again the top 2,000 of those structures are built and optimized via SuperCombihelix, a procedure that follows the same steps as Combihelix. The best structures from SuperCombihelix are then used for ligand docking.

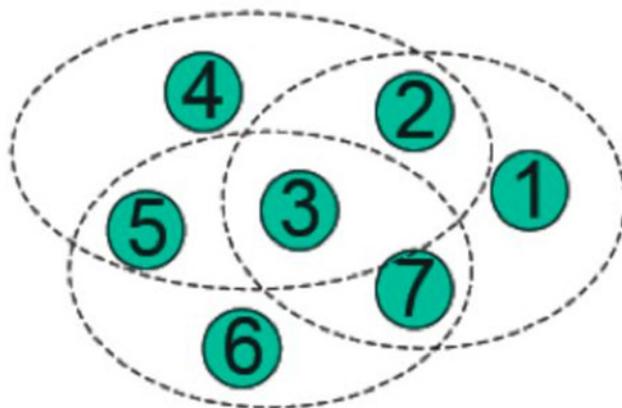


Figure 3. Scoring complexes in Superbihelix. To efficiently determine a subset of conformations for each helix most likely to lead to the lowest energy bundles, we partition the seven-helix bundle into three QuadHelix bundles: TM1-TM2-TM3-TM7, TM2-TM3-TM4-TM5, and TM3-TM5-TM6-TM7.⁵

Small Molecule Docking and Full Structure Optimization

After an ensemble of low-lying structures is generated for the GPCR of interest via GeNSeMBLE, a small molecule binding site is generated and the entire structure is optimized. This consists of several steps, as outlined in **Figure 4**. Spheres are generated for the GeNSeMBLE structures, ligands are generated, the ligands are docked to the GeNSeMBLE structures in GenDock, and the final structures undergo full-atom molecular dynamics. This procedure has been successfully utilized and validated against experimental ligand binding information in many studies.¹⁷⁻²¹ It is important to note that no experimental data is utilized in this procedure, and as such it is well-suited for the predictions of poorly understood GPCR binding sites.

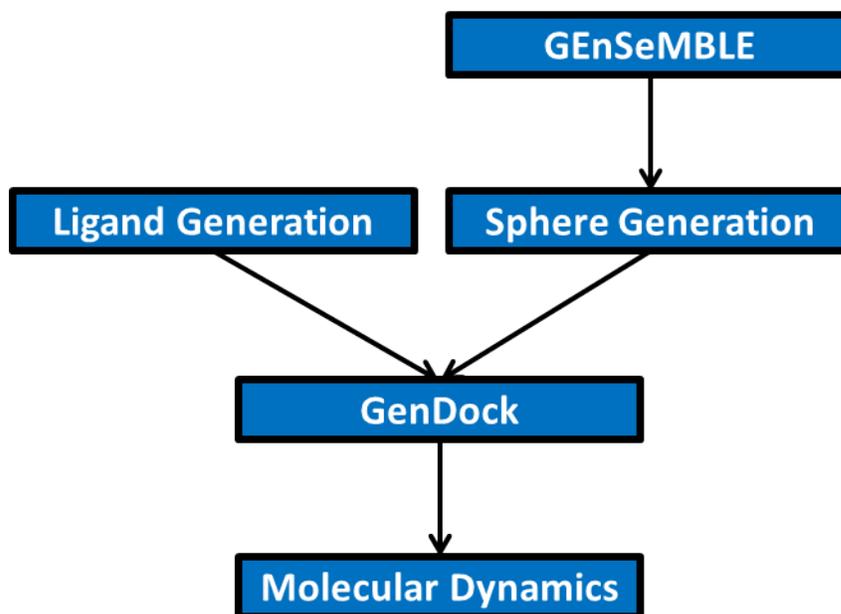


Figure 4. GPCR small molecule binding site generation and full structure optimization procedure overview.

Ligand Preparation

To obtain a small molecule ligand structure, we generate the ligand using Schrödinger's ligand building suite^{12, 22-23}. The ligand is assigned Mulliken charges from quantum mechanics: DFT with the B3LYP functional and the 6-31G** basis set using Schrödinger's Jaguar²³. Because GenDock is a rigid docking program, we generate a series of ligand conformations using Schrödinger's MacroModel¹²: torsional sampling of the rotatable bonds in 30° increments. Ligand conformations are saved which are within 10kJ/mol of the lowest energy conformation. The conformational search is performed with the OPLS force field²⁴ in a dielectric of 80 to match that of water. Clustering of ligands generated is performed using Schrödinger's LigPrep²² in two rounds: first at 2Å diversity, followed by clustering at a 1Å diversity. Each ligand conformation is then minimized.

Sphere Generation

The active site of the GPCR is defined via Dock6's²⁵ sphere generation tools, with sphere regions with a width of 10, overlap of 2, and cutoff of 75. Spheres located outside of the protein are eliminated.

GenDock

a. DarwinDock

Before DarwinDock, all bulky, nonpolar, residues (isoleucine, phenylalanine, leucine, tyrosine, valine, and tryptophan) within 4Å of the spheres are mutated to alanine using SCREAM. DarwinDock generates a large number of poses (5,000) using Dock6 and clusters them into families using Voronoi clustering at 2Å diversity. Poses are discarded if they clash into receptor residues more than a specified number of times depending on the size of the ligand. New ligand poses are added until completeness is achieved. Completeness is defined as when the number of new families created is less than 2% of the number of existing families, indicating that the binding site has been thoroughly sampled. A representative pose for each family is determined (the family head). The top 10% of the family heads and members are scored, and the top 150 structures are passed onto the next step. The DarwinDock procedure allows for a diverse number of poses to be sampled, at minimal computational cost.

b. SCREAM

After DarwinDock, the poses are refined. First, the bulky nonpolar residues are restored to their wild type residues and optimized via SCREAM. The entire complex is then minimized for 10 steps.

c. Neutralize

To reduce the effect of long-range Columbic forces, charged residues are neutralized as described above in the Combihelix section. The same neutralization scheme is applied to the ligand as well, if it is charged at physiological pH.

d. ComplexMinimize

All atoms within 4Å of the ligand are then minimized for 50 steps or to an RMS force of 0.5kcal/mol/Å.

e. Scoring

Snap binding energies and unified cavity energies are calculated. Snap binding energy is the energy difference between the complex and the sum of the receptor and ligand energies. Unified cavity energy is the energy between the ligand and the residues in the unified binding site, i.e. the energies of all of the residues in all of the poses' binding sites are scored for each pose. The top poses are selected based on their energies for undergoing molecular dynamics.

Molecular Dynamics

The full protein and ligand are then relaxed via molecular dynamics. Before undergoing molecular dynamics the intracellular and extracellular loops, N-terminus, and C-terminus of the GPCR must be constructed. This is done via homology modeling to the most closely related GPCR crystal structure.

The complex is inserted into a periodic membrane and water box sufficiently sized to leave a minimum 15Å gap between the protein and edge of the box using NAMD²⁶, with overlapping species removed. Counterions are added using NAMD. The water, lipids, ions, and any newly added residues are minimized for 0.5ns followed by equilibration at 310K for 0.5ns, and then the entire system is minimized for 0.5ns and equilibrated for at least 10ns, with CHARMM22 charges for the protein and CHARMM27 charges for the lipids²⁷⁻²⁸. Equilibration is performed in the isothermal-isobaric (NPT) ensemble using the Nosé-Hoover Langevin piston method^{26, 29}. The waters are modeled using the TIP3P potential function³⁰. Snapshots of the equilibrated dimers (neglecting the first 5ns of the full system equilibration) are taken every half nanosecond, and minimized in the Dreiding forcefield. The lowest energy snapshots are analyzed.

References

1. Abrol, R.; Bray, J. K.; Goddard, W. A., 3rd, Bihelix: Towards De Novo Structure Prediction of an Ensemble of G-Protein Coupled Receptor Conformations. *Proteins* **2012**, *80* (2), 505-18.
2. Abrol, R.; Griffith, A. R.; Bray, J. K.; Goddard, W. A., 3rd, Structure Prediction of G Protein-Coupled Receptors and Their Ensemble of Functionally Important Conformations. *Methods Mol Biol* **2012**, *914*, 237-54.
3. Abrol, R.; Kim, S. K.; Bray, J. K.; Griffith, A. R.; Goddard, W. A., 3rd, Characterizing and Predicting the Functional and Conformational Diversity of Seven-Transmembrane Proteins. *Methods* **2011**, *55* (4), 405-14.
4. Abrol, R.; Kim, S. K.; Bray, J. K.; Trzaskowski, B.; Goddard, W. A., 3rd, Conformational Ensemble View of G Protein-Coupled Receptors and the Effect of Mutations and Ligand Binding. *Methods Enzymol* **2013**, *520*, 31-48.
5. Bray, J. K.; Abrol, R.; Goddard, W. A., 3rd; Trzaskowski, B.; Scott, C. E., Superbihelix Method for Predicting the Pleiotropic Ensemble of G-Protein-Coupled Receptor Conformations. *Proc Natl Acad Sci U S A* **2014**, *111* (1), E72-8.
6. Yang, J., Zhang, Y. Gpcrs-Exp: A Database for Experimentally Solved Gpcr Structures. <http://zhanglab.ccmb.med.umich.edu/GPCR-EXP> (accessed November 11, 2014).
7. Kenakin, T.; Miller, L. J., Seven Transmembrane Receptors as Shapeshifting Proteins: The Impact of Allosteric Modulation and Functional Selectivity on New Drug Discovery. *Pharmacol Rev* **2010**, *62* (2), 265-304.
8. Kobilka, B. K.; Deupi, X., Conformational Complexity of G-Protein-Coupled Receptors. *Trends Pharmacol Sci* **2007**, *28* (8), 397-406.
9. Chun, E.; Thompson, A. A.; Liu, W.; Roth, C. B.; Griffith, M. T.; Katritch, V.; Kunkun, J.; Xu, F.; Cherezov, V.; Hanson, M. A.; Stevens, R. C., Fusion Partner Toolchest for the Stabilization and Crystallization of G Protein-Coupled Receptors. *Structure(London, England:1993)* **2012**, *20* (6), 967-976.
10. Larkin, M. A.; Blackshields, G.; Brown, N. P.; Chenna, R.; McGettigan, P. A.; McWilliam, H.; Valentin, F.; Wallace, I. M.; Wilm, A.; Lopez, R.; Thompson, J. D.; Gibson, T. J.; Higgins, D. G., Clustal W and Clustal X Version 2.0. *Bioinformatics* **2007**, *23* (21), 2947-8.
11. Schrödinger *Maestro*, 9.9; 2014.
12. Schrödinger *Macromodel*, 10.6; 2014.
13. Lim, K.-T.; Brunett, S.; Iotov, M.; McClurg, R. B.; Vaidehi, N.; Dasgupta, S.; Taylor, S.; Goddard, W. A., Molecular Dynamics for Very Large Systems on Massively Parallel Computers: The Mpsim Program. *Journal of Computational Chemistry* **1997**, *18* (4), 501-521.
14. Mayo, S. L.; Olafson, B. D.; Goddard, W. A., Dreiding: A Generic Force Field for Molecular Simulations. *The Journal of Physical Chemistry* **1990**, *94* (26), 8897-8909.
15. Tak Kam, V. W.; Goddard, W. A., Flat-Bottom Strategy for Improved Accuracy in Protein Side-Chain Placements. *Journal of Chemical Theory and Computation* **2008**, *4* (12), 2160-2169.

16. Lomize, M. A.; Lomize, A. L.; Pogozheva, I. D.; Mosberg, H. I., Opm: Orientations of Proteins in Membranes Database. *Bioinformatics* **2006**, *22* (5), 623-5.
17. Kim, S. K.; Li, Y.; Abrol, R.; Heo, J.; Goddard, W. A., 3rd, Predicted Structures and Dynamics for Agonists and Antagonists Bound to Serotonin 5-Ht2b and 5-Ht2c Receptors. *J Chem Inf Model* **2011**, *51* (2), 420-33.
18. Kim, S. K.; Riley, L.; Abrol, R.; Jacobson, K. A.; Goddard, W. A., 3rd, Predicted Structures of Agonist and Antagonist Bound Complexes of Adenosine A3 Receptor. *Proteins* **2011**, *79* (6), 1878-97.
19. Goddard, W. A., 3rd; Kim, S. K.; Li, Y.; Trzaskowski, B.; Griffith, A. R.; Abrol, R., Predicted 3d Structures for Adenosine Receptors Bound to Ligands: Comparison to the Crystal Structure. *J Struct Biol* **2010**, *170* (1), 10-20.
20. Floriano, W. B.; Vaidehi, N.; Zamanakos, G.; Goddard, W. A., 3rd, Hierarchical Docking Protocol for Virtual Ligand Screening of Large-Molecule Databases. *J Med Chem* **2004**, *47* (1), 56-71.
21. Nair, N.; Kudo, W.; Smith, M. A.; Abrol, R.; Goddard, W. A., 3rd; Reddy, V. P., Novel Purine-Based Fluoroaryl-1,2,3-Triazoles as Neuroprotecting Agents: Synthesis, Neuronal Cell Culture Investigations, and Cdk5 Docking Studies. *Bioorg Med Chem Lett* **2011**, *21* (13), 3957-61.
22. Schrödinger *Ligprep*, 3.2; 2014.
23. Bochevarov, A. D.; Harder, E.; Hughes, T. F.; Greenwood, J. R.; Braden, D. A.; Philipp, D. M.; Rinaldo, D.; Halls, M. D.; Zhang, J.; Friesner, R. A., Jaguar: A High-Performance Quantum Chemistry Software Program with Strengths in Life and Materials Sciences. *International Journal of Quantum Chemistry* **2013**, *113* (18), 2110-2142.
24. Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J., Development and Testing of the Opls All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids. *Journal of the American Chemical Society* **1996**, *118* (45), 11225-11236.
25. Lang, P. T.; Brozell, S. R.; Mukherjee, S.; Pettersen, E. F.; Meng, E. C.; Thomas, V.; Rizzo, R. C.; Case, D. A.; James, T. L.; Kuntz, I. D., Dock 6: Combining Techniques to Model Rna-Small Molecule Complexes. *RNA* **2009**, *15* (6), 1219-30.
26. Phillips, J. C.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R. D.; Kale, L.; Schulten, K., Scalable Molecular Dynamics with Namd. *J Comput Chem* **2005**, *26* (16), 1781-802.
27. Brooks, B. R.; Brooks, C. L., 3rd; Mackerell, A. D., Jr.; Nilsson, L.; Petrella, R. J.; Roux, B.; Won, Y.; Archontis, G.; Bartels, C.; Boresch, S.; Caflisch, A.; Caves, L.; Cui, Q.; Dinner, A. R.; Feig, M.; Fischer, S.; Gao, J.; Hodoscek, M.; Im, W.; Kuczera, K.; Lazaridis, T.; Ma, J.; Ovchinnikov, V.; Paci, E.; Pastor, R. W.; Post, C. B.; Pu, J. Z.; Schaefer, M.; Tidor, B.; Venable, R. M.; Woodcock, H. L.; Wu, X.; Yang, W.; York, D. M.; Karplus, M., Charmm: The Biomolecular Simulation Program. *J Comput Chem* **2009**, *30* (10), 1545-614.
28. MacKerell, A. D.; Bashford, D.; Bellott, M.; Dunbrack, R. L.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; Joseph-McCarthy, D.; Kuchnir, L.; Kuczera, K.; Lau, F. T.; Mattos, C.; Michnick, S.; Ngo, T.; Nguyen, D. T.; Prodhom, B.; Reiher, W. E.; Roux, B.; Schlenkrich, M.; Smith, J. C.; Stote, R.; Straub, J.; Watanabe, M.; Wiorkiewicz-Kuczera, J.; Yin, D.; Karplus, M., All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins. *J Phys Chem B* **1998**, *102* (18), 3586-616.

29. Feller, S. E.; Zhang, Y.; Pastor, R. W.; Brooks, B. R., Constant Pressure Molecular Dynamics Simulation: The Langevin Piston Method. *The Journal of Chemical Physics* **1995**, *103* (11), 4613-4621.
30. Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L., Comparison of Simple Potential Functions for Simulating Liquid Water. *The Journal of Chemical Physics* **1983**, *79* (2), 926-935.

Chapter III:
**The Predicted Structure of Peptide-Bound
Glucagon-Like Peptide 1 Receptor, a Class B1
G Protein-Coupled Receptor**

Abstract

The glucagon-like peptide 1 receptor (GLP1R) is a G protein-coupled receptor involved in insulin synthesis and regulation, and therefore is an important drug target for the treatment of diabetes. However, GLP1R is a member of the class B1 family of GPCRs, for which there are no complete crystal structures (i.e. transmembrane region, large structured N-terminus, and peptide ligand). To provide a structural basis for drug design and to probe class B1 GPCR activation, we predicted the transmembrane bundle structure of GLP1R bound to the peptide Exendin-4 (Exe4) (a GLP1R agonist on the market for treating diabetes) using the MembStruk method for scanning TM bundle conformations. We then used protein-protein docking methods to combine the TM bundle with the x-ray crystal structure of the 143 amino acid N-terminus coupled to the Exe4 peptide. This complex was subjected to 28ns of full-solvent full-lipid molecular dynamics. We find a total of fourteen strong polar interactions of Exe4 with GLP1R, of which eight are in the TM bundle (two confirmed by mutation studies) and six involve the N-terminus (three found in the crystal structure). We also find ten important hydrophobic interactions, of which four are in the TM bundle (two confirmed by mutation studies), and six are in the N-terminus (all present in the crystal structure). Thus, our predicted structure is in excellent agreement with available mutagenesis studies. We suggest a number of new mutation experiments for further validation of our predicted structure. The structure should be useful for guiding drug design and can provide a structural basis for understanding ligand binding and receptor activation of GLP1R and other class B1 GPCRs.

Introduction

The glucagon-like peptide 1 receptor (GLP1R) is a GPCR which belongs to the class B1 (secretin-like) family of GPCRs. Class B1 GPCRs are activated by peptide hormones. A feature distinguishing them from class A GPCRs is their large, highly structured, N-terminal ectodomain, which binds their ligands. Mechanisms for class B1 GPCR agonist binding and activation initiation have been speculated upon¹⁻³, but in the absence of

atomic level structures of the full receptor-peptide ligand complexes it is difficult to understand, probe, and expand upon these activation hypotheses.

The large ectodomain of GLP1R interacts strongly with the C-terminal half of its endogenous polypeptide agonist (in this case GLP1). The N-terminus of the ligand binds to the TM bundle and extracellular loops. Activation of GLP1R by GLP1 stimulates the adenylyl cyclase pathway which increases insulin synthesis and release of insulin in a glucose-dependent fashion ⁴. In addition, GLP1 reduces body weight by increasing satiety in the brain ⁵. Consequently GLP1 would seem to be attractive for treating both type 2 diabetes and obesity. However, GLP1 is rapidly degraded by the serine protease dipeptidyl peptidase-IV, resulting in a half-life of only 1-2 minutes ⁶⁻⁷. Exendin-4 (Exe4), a 39-amino acid peptide isolated from the venom of the Gila monster, is a more stable analog of GLP1 with a half-life of 2.5 hours in its marketed form ⁸⁻⁹. It has a 50% sequence homology with GLP1, and is a full agonist with a stronger affinity and potency for GLP1R ¹⁰. Indeed, Exe4 is currently on the market for treatment of diabetes. Despite the success of Exe4 and its derivatives, there is still a need to develop small-molecule (non-peptide) orally active agonists of GLP1R. This need is furthered by recent reports of oncogenic side effects of Exe4 ¹¹. This process of novel drug design could be aided significantly if there was knowledge of the full-atom structure of the full GLP1R bound to Exe4, which is the motivation of the research reported here. Our structure also provides testable hypotheses of GLP1R activation upon ligand binding.

In the following sections, we present the predicted structure of the full membrane-bound GLP1R/Exe4 complex in the presence of water. The TM bundle was predicted using the MembStruk methodology. This bundle, which was attached to the crystal GLP1R N-terminus and partial Exe4, was inserted into a periodic membrane in a box of water and relaxed via molecular dynamics (MD) ¹²⁻¹³. We find that this predicted structure is consistent with all available mutation data, but we suggest additional experimental tests to validate key aspects of our structure. We believe that this structural information presented here should help in the development of selective active small-molecule agonists for GLP1R and also aid in probing the activation of class B1 GPCRs.

Methods

Overview

We have been developing methods for predicting the structures of the transmembrane region of GPCRs since the late 1990s. The earlier methodology, denoted as MembStruk, focused on sequential optimization of the 7 TM helices starting from a homology template¹². More recently, we developed a new method, denoted as GEnSeMBLE, that aims at a combinatorially complete set of helix rotations and tilts¹⁴. The structure that we report here was built entirely using MembStruk several years ago, but had not been published. We applied our new GEnSeMBLE methodology to the older MembStruk structure, but we obtained essentially the same packing of the 7 TM helices. Therefore, we decided to continue with our previous structure for the 7 TM helix bundle and its connection to the N-terminus, but we replaced the previously homology built nGLP1R/Exe4 part (where nGLP1R is the N-terminus of GLP1R) with the crystal structure which appeared more recently.

A summary of the full procedure used to generate the GLP1R/Exe4 can be given in 10 steps (the procedure through step 6 is depicted in **Figure 1**):

1. The initial step in MembStruk is to use comparative hydrophobicity analysis of GLP1R and related GPCRs to identify the seven likely TM domains and then to position (x,y,z) the centers of these helices on a common plane with preselected tilts (θ , ϕ) and axis rotations (η) based on some template (in this case we used the predicted DP structure)^{12, 15}. We found that the TM bundle's first helix started with residue 144, and then seventh helix ended with residue 408.
2. This is followed by sequential optimization of each TM domain by varying η over 360° in 30° increments and side chain optimization (SCREAM with modest minimization) in a sequence that considers all TM domains multiple times¹⁶.
3. Because of the large 143AA N-terminus of GLP1R, we originally built the structure for nGLP-1R from homology to the NMR structure of the mouse CRF receptor (PDB ID:2JND). This 131AA region then underwent optimization of the side chains¹⁷.
4. We used the ZDOCK procedure to dock the NMR structure of the full Exe4 ligand (PDB ID: 1JRJ) to the nGLP1R structure from step 3¹⁸⁻¹⁹.

5. We manually docked the nGLP1R/Exe4 complex from step 4 to the TM bundle in such a way that the N-terminus of the ligand could interact with the TM region from step 1.
 6. We connected the N-terminus to the TM region (residues 131 to 145) and built the three extracellular and three intracellular loops using MODELLER, followed by SCREAM to position the side chains and then energy minimization²⁰.
 7. Then we inserted the full GLP1R protein into a periodic membrane and water box (75Å×75Å×117Å), eliminating overlapping species to obtain a system with 61,119 atoms. This was equilibrated at 300K (first the water and membrane and then all atoms for 20ns) using NAMD 2.6, with CHARMM22 charges for the protein and CHARMM27 charges for the lipids²¹⁻²³. The waters were modeled using TIP3P²⁴.
 8. We then extracted the 7 TM helices from this structure and applied the SuperBiHelix procedure of GEnSeMBLE to obtain a new 7 TM bundle. This led to essentially the exact MembStruk structure for GLP1R, (in addition we found a second high-scoring structure differing most prominently in the η for TM1 and TM7). Consequently, we decided to continue the MD on the already equilibrated original structure.
 9. However, in the meantime, an x-ray crystal structure had appeared for the ectodomain of GLP1R (nGLP1R) bound to part of Exe4 (9-39) (PDB ID: 3C5T)¹³. We matched this to our predicted structure and reoptimized (SCREAM, minimization, and MD for 18ns).
 10. After the 18ns of full-atom and full-solvent molecular dynamics we performed simulated annealing of the TM portion of the ligand binding site, and then we carried 10 more ns of full-atom and full-solvent molecular dynamics at 300K. A representative snapshot of the final 10ns of MD was chosen for the discussion below.
- Experimental information was not used during any of the above steps, except that information known about GLP1 loop structures was used to select loops from MODELLER in step 6.

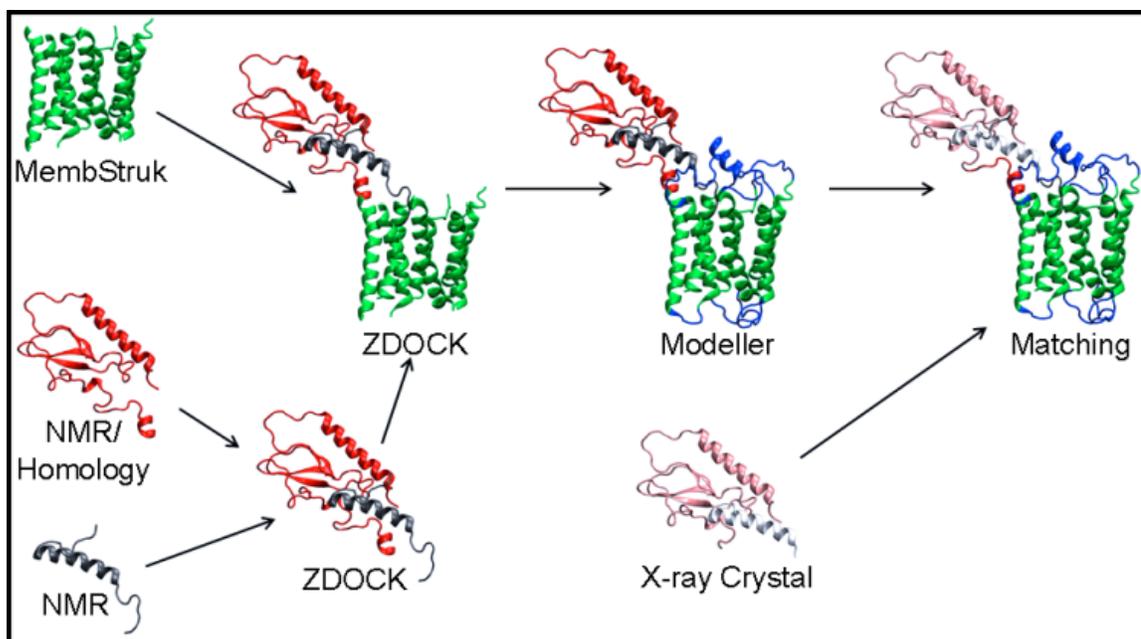


Figure 1. Creating the GLP1R/Exe4 bundle. The steps used to create our GLP1R/Exe4 structure are depicted, along with the methods used at each point. After the steps shown here, the entire complex was optimized through 28ns of molecular dynamics.

Details

Unless specified otherwise all minimizations referenced were conjugate-gradient minimizations performed in MPSim under the Dreiding forcefield (FF)²⁵⁻²⁶. Likewise, all of the molecular dynamics were performed in NAMD 2.6, using CHARMM22 charges for the protein and CHARMM27 charges for the lipids²¹⁻²³. The waters were modeled using TIP3P²⁴.

Structure Prediction of the Transmembrane Domain

The 3D structure of the GLP1R TM bundle was predicted using MembStruk (version 4.3). The details of MembStruk are described elsewhere¹². Here we outline the procedure, highlighting aspects relevant to GLP1R or that were additionally improved.

a. Prediction of Transmembrane Regions

The transmembrane (TM) region was predicted using the TM2nsS method¹². We searched for the sequences of the family B GPCRs from the UniProtKB/Swiss-Prot

database²⁷. The 166 sequences included 76 members of the LN-TM7 and 18 members of the Methuselah subfamilies. Only seven sequences (belonging to the receptors for glucagon/glucagon-like peptide and glucose-dependent insulinotropic polypeptide) showed a sequence identity higher than 40%. Among the 166 sequences we took 65 sequences with pairwise sequence identity to GLP1R of >20% to calculate the hydrophathy analysis curve. A multiple sequence alignment was performed with ClustalW²⁸, which was used as input to TM2ndS for hydrophathy analysis. In this alignment, the long amino terminus of the sequence was excluded. The hydrophobic center for each TM was determined by the position bisecting the area of each peak in the curve. In order to define the clear boundary of the TM regions, we carried out a second round of seven TM predictions where the sequence of each TM core (20 amino acids around a hydrophobic center) was used as a query in a BLAST search. Under a high gap penalty in the BLAST search, the sequences with >50% identity were identified from the set of the family B GPCR sequences. The final refined TM region and its hydrophobic center for each of the 7 TM domains were determined from this second round of prediction.

b. Assembly of Transmembrane Helical Bundle

We built a canonical α -helix for each TM region and assembled these seven helices into a template generated from the predicted structure of human prostaglandin DP receptor fully equilibrated in explicit lipids and water solvent¹⁵. As expected, all receptors available from our family A GPCR structural database were distant in sequence from the GLP1R. The human prostaglandin DP receptor was chosen since it was the only hormone receptor in our database. Here we assumed that the arrangement of TM helices of the GLP1R would be similar to that of the family A GPCR. The (x, y) coordinate, the tilt angle with respect to z-axis and the azimuthal angle needed for definition of arrangement of TM α -helices as shown for frog rhodopsin²⁹ were calculated as follows: the xy mid-plane and z-axis were determined by diagonalizing the matrix of the moment of inertia for the heavy atoms comprising the lipid bi-layers where the human prostaglandin DP receptor was embedded. The +z direction pointed towards the extracellular region. The center of the helix bundle C α atoms was set to origin and the x-

axis was defined as the axis from the center of the bundle to the center of helix 2. Under this coordinate system, the (x, y) coordinates of each helix center and the tilt/azimuthal angles of each helix (the helix axis was determined from the moment of inertia of C α atoms) were computed for the DP receptor template. The seven hydrophobic centers for GLP1R were all in the x-y plane with these (x, y) coordinates and the seven helices were inclined accordingly. Each helix was rotated about its axis so that its hydrophobic moment pointed towards the membrane.

We then carried out 200ps of molecular dynamics at 300K without solvent or lipid in order to allow the conformation of each individual helix to bend or kink as appropriate. The molecular dynamics were run in MPSim under canonical ensemble (NVT) conditions with the Dreiding forcefield. We selected the lowest potential energy snapshot after 100ps. Using this conformation, the net hydrophobic moment vector was calculated from the middle 15 residue around the hydrophobic center. Each helix was rotated so that this hydrophobic moment vector pointed away from the center of the helix bundle.

Next, the orientation of the helix was further examined with energy-based optimization. The rotational orientation of each helix was scanned over 360° in 30° increments and at each orientation the side-chains were placed using the SCREAM program. Here we used the coarse energy scoring function that was combined with the penalty score derived from the hydrophobicity scale and the FF-based SCREAM energy function³⁰. From this scanning step, we chose two orientations for TM 1, 2, 3, 4, 5, 7 and three for TM6, leading to the 192 combinatorial orientations. These 192 combinations were ordered by the number of inter-helical hydrogen bonds (IHHBs), showing a consistent orientation for TM2, 3, and 7 on the top. These structures consistently showed hydrogen bonds between R190 (TM2) and N240(TM3) and between E247(TM3) and Q394(TM7). After that, the orientation of TM1 was scanned and the best orientation was chosen based on the number of IHHBs. This was motivated by previous experimental studies reporting that inter-helical hydrogen bonding drives strong interactions in membrane proteins³¹⁻³². The orientation of TM4 was selected where the aromatic residues were inside the bundle and well packed with the adjacent helices. TM5

and 6 were then scanned combinatorially and the best rotations were chosen based on the sidechain-sidechain hydrogen bond energy.

The final helix bundle was subjected to conjugate gradient minimization to a RMS force threshold of 0.5kcal/mol/Å. Two layers of explicit lipid molecules were then added to the bundle and this 7-helix-lipid complex was optimized in order to achieve the proper packing using rigid body molecular dynamics in MPSim with the Dreiding forcefield for 50ps. The final equilibrated structure was then minimized to an RMS force threshold of 0.3kcal/mol/Å.

Structure Prediction of nGLP1R/Ligand Complex

The structures of the N-terminus of GLP1R (nGLP1R) and ligand were modeled separately and then combined. The details are described elsewhere³³ and we outline the procedure briefly.

a. Determination of the 3-D Structure for nGLP1R

The structure of nGLP1R was determined by homology modeling with the NMR structure of mouse CRF receptor (PDB ID: 2JND), which consists of 19 conformations¹⁷. We scored these 19 structures using the potential energy in the Dreiding forcefield, and selected the lowest three for equilibration in a periodic box with explicit water solvent. The final template structure was determined by considering the energy, and possible conserved structural motifs among class B1 GPCRs (i.e. the salt bridge between Asp65 and Arg101) and the C α RMS with the NMR structures¹.

We used MODELLER9v1³⁴ to construct homology models for residues 45-130 of nGLP1R. The alignment of sequences between the template (the mouse CRF receptor) and the query (nGLP1R) was determined from the multiple sequence alignment including seven other secretin-like family B GPCRs whose N-terminal sequences show >30% identity. The side chains were replaced using the SCREAM program¹⁶ and optimized with 100-steps of conjugate gradient minimization. This final optimized structure was then equilibrated in the explicit water box. The water box was chosen to extend by ~7 Å beyond the solute in each direction. The whole system was then neutralized by adding

Na⁺ ions. The solvent for each system was minimized with conjugate gradients for 5000 steps and then equilibrated for 10 ps of molecular dynamics at 310 K using the Langevin thermostat with a damping coefficient of 5 ps⁻¹ with all of the coordinates of the solute fixed. This system was then minimized without any constraint and equilibrated for 500 ps at 310 K and a pressure of 1 atm (using the Langevin piston method). All simulations used time steps of 1 fs with electrostatic interactions computed using the Particle Mesh Ewald (PME) method.

b. Docking of Ligands into nGLP1R

The NMR structure of Exe4 (PDB ID: 1JRJ) was optimized with conjugate gradient minimization and equilibrated for 500 ps in explicit water solvent at 310 K and a pressure of 1 atm using the Langevin piston method¹⁹. We docked the optimized Exe4 ligand into nGLP1R (residue 45-130) using ZDOCK rigid docking program¹⁸. By assuming that the Exe4 would bind to a region similar to that of mouse CRF receptor/astressin, we filtered the initial 2000 configurations down to 131 configurations¹⁷. After side-chain optimization and conjugate gradient minimization, we selected the top five configurations based on the FF energy, each of which were then fully minimized to a RMS force of 0.3 kcal/mol/Å. The lowest energy configuration was equilibrated in explicit water solvent with a harmonic constraint of 5 kcal/(mol Å²) for the backbone atoms as described previously.

Combination of nGLP1R/Exe4 Complex with the Transmembrane Bundle

In order to combine our nGLP1R/ligand complex with the TM bundle, we needed to grow protein residues 131 to 145. We ran secondary structure predictions on these residues in the programs nnPredict, GORIV, HNN, SOPMA, PSIPRED, Porter, and Jpred³⁵⁻⁴¹. Based on secondary structure predictions (six consensus predictions among seven) we found that residues 137 to 145 were helical. The conformation of this segment was predicted by using MODELLER9v1 under helix restraint. The overall combined structure was built by matching the common helical parts of residue 137 to 145 in the nGLP1R/Exe4 complex and the TM1 region. We first generated 1,000 conformations of

the residues 131 to 145 and selected ones with the correct helix chirality (right-handed) and the helical conformation preserved (not unraveled). Then we found the conformations where the ligand Exe4 was located properly inside the TM pocket without any bumping into helices. We examined the interactions between receptor and ligand and chose the conformations where the ligand key residues were involved⁴². We carried out minimization and annealing MD for the residues of the ligand contacting TM regions and having random coil conformation (residues 1 to 7). The energy-minimization and annealing molecular dynamics were run in MPSim for three cycles (50K→600K→50K (50K step; 1ps run for each T)).

Prediction of Loop Structures

The first and second extracellular loops (EC1 and EC2) were first modeled by using MODELLER9v1. EC1 and EC2 were 29 and 14 amino acids long, respectively. According to the secondary structure predictions from APSSP2 and PSIPRED servers, EC1 was found to be helical from residues 215 to 224^{39,43}. We generated 1000 conformations under helix restraint in MODELLER9v1 for EC1 and EC2. The restraints included both a distance restraint between Y205 on EC1 and F6 of the Exe4 ligand^a, as well as a disulphide bond between EC1 (residue 226) and EC2 (residue 296)⁴⁴⁻⁴⁶. We chose the conformations where the predicted parts were helical and right-handed (723 among 1000 conformations). The final three candidates were selected, for which the conformation was compact and not touching the presumed lipid regions, and the key residues of the EC1 and the ligand showed the favorable contacts (or at least close). The contacts emphasized were with M204, Y205, D215, and R227^{1 47-48}. The loops were optimized by side-chain replacement with SCREAM, energy-minimization of the EC1 and EC2 only, and then three-cycle annealing MD (50K→600K→50K (50K step; 1ps run for each T)) in MPSim. Based on the potential energy and maximizing the total number of contacts, the best structure of EC1 and EC2 was chosen. Next, the 10 amino acid EC3 loop was modeled. Among the 1000 conformations generated from MODELLER9v1 we chose two candidates where the loop was not touching the presumed

^a Since we were building both GLP1 and Exe4 in this process, and no corresponding data existed for Exe4, we used information from studies which were performed on GLP1.

lipid regions and was likely to contact T7 of the ligand that was known as one of key residues⁴⁵.

The remaining three intracellular (IC) loops were modeled all together. The lengths of the IC1, IC2, and IC3 loops were 5, 9, and 16 amino acids, respectively. 200 conformations were generated and the loops were chosen which sat in a closed conformation on top of the TM Region. The EC3 and IC loops were optimized together through SCREAM, energy minimization, and three-cycle annealing molecular dynamics (50K→600K→50K (50K step; 1ps run for each T)) in MPSim. Based on the potential energy and maximizing contacts, the final structure was chosen.

Relaxation in Explicit Membrane and Water Solvent

After predicting the full structure of the GLP1R/Exe4 complex, we embedded this structure in a periodic infinite membrane and solvated the system with explicit water and equilibrated with MD at 310K. The system size was 75Å×75Å×117Å with 6493 solute atoms, 5092 lipid atoms, 40302 water atoms, and 5 Na⁺ ions. We used palmitoyl-oleoyl-phosphatidylcholine (POPC) to form the lipid bilayers. We made the receptor with the acetylated N-terminus and the N-Methylamide C-terminus. Prior to full equilibration, the solvent molecules were equilibrated first at 310K for 100ps, then the whole system was equilibrated by gradually increasing the temperature to 100K, 200K, and 310K with a 500ps MD run performed at each temperature. Finally, the full MD simulation was carried out for 20ns while a constant pressure of 1 atm was maintained by using the Langevin piston method.

Incorporating the Crystal Structure

After the 2008 crystal structure of the human nGLP1R in complex with the antagonist Exe4(9-39) was revealed (PDB ID: 3C5T), we incorporated it into our GLP1R model¹³. This crystal structure was aligned to our relaxed GLP1R bound to Exe4 structure using VMD's RMSD Trajectory Tool⁴⁹. This alignment was done by the backbone atoms of those ligand residues which were well resolved in the crystal structure -- 9-33. The resolved crystal residues (9-33 for Exe4 and 28-131 for GLP1R) were then substituted for those in our structure. Five residues on either side of the newly connected

residues were minimized to 0.5RMS force with conjugate gradient minimization in MPSim with Dreiding forcefield. Then, the entire complex was minimized to 0.5RMS force. This procedure allowed us to retain our TM bundle conformation and binding site, while adding in the crystal information.

Optimizing the New Structure

a. Molecular Dynamics

To optimize our models further, we ran full-lipid and full-solvent molecular dynamics. Each complex was inserted into a fully equilibrated hydrated POPC lipid blayer having a cell size of 75Å x 75Å. All lipids within 1Å and all waters within 5Å were removed. The complex was then solvated using the solvate package of VMD⁴⁹. In order to have a net charge of zero, three sodium atoms were added to the Exe4 structure using the autoionize feature of VMD⁴⁹. Both structures then underwent 250ps, with a 1fs timestep, of conjugate gradient minimization. This was done with the protein and ligand kept fixed. These remained fixed as the systems were heated to 310K and equilibrated for 500ps under NPT conditions. Next, both entire systems were minimized for 250ps. Finally, the structures underwent NPT dynamics for 18ns.

b. Simulated Annealing

Upon inspecting the dynamics runs, it was decided to optimize the transmembrane region binding sites further by simulated annealing in MPSim with the Dreiding forcefield. The proteins and ligands were taken from their MD runs, and everything within 5Å of the ligand was allowed to be flexible, along with the loops. These structures underwent minimization to 0.5RMS, with a maximum of 1000 steps, and were then heated from 50 to 600K in 50 degree increments, with 0.1ps (100 steps) spent at each temperature five times. The final structure was again minimized to 0.5RMS.

c. More Molecular Dynamics

After annealing, the two structures were reinserted into membranes, and the above procedure for starting NPT molecular dynamics was repeated for 10ns. The trajectories

were analyzed at 1ns intervals, with each protein complex being minimized to 0.5 RMS force using MPSim with the Dreiding forcefield. A representative snapshot was chosen for each run for discussion here.

Utilizing Updated Methods on the MembStruk TM Bundle

With the goal of optimizing our MembStruk TM bundle, which had been through around twenty nanoseconds of MD, we ran Bihelix/Combihelix on the TM region⁵⁰. This sampled the eta angles from 0 to 360 in 30 degree increments. The top ten helix orientations are shown in **Table 1**. You can see that the all zero structure (our starting MembStruk structure) shows up as number nine. We then ran Superbihelix/Supercombihelix on all 10 of these structures, with the top two results shown in **Table 2**⁵¹. We sampled both the eta and phi angles of each of the 10 structures from -30 to 30 in 15 degree increments. As you can see, the second structure is almost identical to the starting structure, with only two -15 degree changes to phi for helices two and four. Because this was so close to our starting structure, we decided to continue with the MD-relaxed structure for further work. The number one case, which deviates more from the all-zero, is not discussed in this work.

Table 1. Top ten Bihelix/Combihelix structures for GLP1R

Eta						
H1	H2	H3	H4	H5	H6	H7
90	0	0	0	0	0	270
0	0	0	0	0	0	270
30	0	0	0	0	0	270
0	0	0	0	0	30	0
30	0	0	0	0	0	0
0	0	0	0	270	0	0
90	0	0	0	0	0	0
270	0	0	0	0	0	270
0	0	0	0	0	0	0
240	0	0	0	0	0	270

Table 2. Top two Superbihelix/Supercombihelix structures

Phi							Eta						
H1	H2	H3	H4	H5	H6	H7	H1	H2	H3	H4	H5	H6	H7
-15	-30	0	0	0	0	15	75	15	0	0	0	15	240
0	-15	0	-15	0	0	0	0	0	0	0	0	0	0

Results and Discussion

Intra-Protein Interactions

For class A GPCRs, several conserved interhelical interactions, such as the TMs 1-2-7 or TMs 2-3-4 hydrogen bond networks, are present in most crystal structures. The amino acids involved in the TMs 1-2-7 and TMs 2-3-4 interactions are not conserved in class the B GPCRs. However, we do find many interhelical interactions, some of which occur between residues conserved in class B1 which as such could be important in all of the class B1 GPCRs. The GLP1R/Exe4 complex has 16 hydrogen bonds within its TM bundle or connecting loops, as shown in **Table 3** and discussed in the following sections. The TM-bundle interactions are pictured in **Figure 2**.

Table 3. Interhelical hydrogen bonds for the GLP1R/Exe4 structure.

Donor	Acceptor	Energy
TM3-TM6 Ionic Lock		
R348(IC3)	E247(TM3)	-45.4
Coupling of TM2-TM3-TM6		
N182(TM2)	E247(TM3)	-10.7
R190(TM2)	N240(TM3)	-15.1
TM1-TM2-TM7 Interaction Network		
R176(TM2)	E408(TM7)	-52.9
Y152(TM1)	Q394(TM7)	-4.7
T149(TM1)	E387(TM7)	-8.1
H180(TM2)	S163(TM1)	-3.5
Extracellular Loop Couplings		
R299(EC2)	D222(EC1)	-46.2
W297(EC2)	D222(EC1)	-6.1
C296(EC2)	D222(EC1)	-5.5
H374(EC3)	M303(EC2)	-6.4

We did not include here interactions between residues within the same TM or between a TM and its adjoining loop. Energies are in kcal/mol. Red lettering indicates that the residue is fully conserved among class B1 GPCRs, green indicates that it is partially conserved, and blue means that it is conserved among some class B1 GPCRs.

TM3-TM6 Ionic Lock

We consider that the conserved E247(TM3)-R348(TM6) ionic lock is analogous to the R(3.50)-D/E(6.30) ionic lock of class A GPCRs known to stabilize the GPCR in an inactive form, even though this donor and acceptor residues are swapped in the class B1 variation of the TM3-TM6 ionic lock⁵². We find this ionic lock to be maintained throughout the dynamics, except for a brief 1ns break. We suggest that this interaction plays the role of maintaining the inactive form of class B1 GPCRs. Indeed, E247 is fully conserved among class B1 GPCRs, while R348 is either an R or K in all class B1 GPCRs. This hypothesis could be tested by mutations that break the ionic lock, which should lead to increased constitutive activity (GLP1R does not exhibit constitutive activity)⁵³. Although we assert that such mutations should be formed, we recognize that for the GLP1 system, Heller et al. studied R348G and found no activation for all concentrations of GLP1, while Takhar et al. found R348A to have no effect on binding or

activation of GLP1⁵⁴⁻⁵⁵. Perhaps these large changes to alanine or glycine led to a modified TM bundle that changed the binding site. A more subtle R348Q mutation would test the intricacies of this ionic lock in more detail. It is also possible that this ionic lock is specific to the GLP1R/Exe4 complex, and is involved in its activation, but not those of the entire class B1 family.

Since the TM3-TM6 ionic lock is present in our structure, it is possible that our complex is not yet fully activated. However our structure is bound to an agonist, and TM6 makes almost no interactions other than the ionic lock. Thus, it may be at least partially activated. This lack of TM6 interactions would allow TM6 to immediately move to interact with the G α as it does in the other active GPCR structures.

Coupling of TM2-TM3-TM6

The 3-6 ionic lock is additionally coupled to TM2 via the conserved N182(TM2)-E247(TM3) hydrogen bond, resulting in a TM2-TM3-TM6 (2-3-6) hydrogen bond network [N182(TM2)-E247(TM3)-R348(TM6)]. This N182(TM2)-E247(TM3) has only minor fluctuations through the course of the dynamics. Since N182 is either an N, H, or Q in class B1 GPCRs, this 2-3-6 hydrogen bond network may occur in any of the class B1 GPCRs. The 2-3-6 conserved network may be further stabilized by the similarly conserved R190(TM2)-N240(TM3) hydrogen bond and which is shown in conjunction with the 2-3-6 network in **Figure 2**. We suggest that this interaction might be analogous to the TM2.45(S/N/T)-3.42(S/N/T)-4.50(W) conserved interaction of class A GPCRs, which are also related to their 3-6 ionic lock. Indeed, N240(TM3) is just seven amino acids away from E247(TM3), compared to the 3.42 just eight amino acids away from 3.50. We find the R190(TM2)-N240(TM3) to be very constantly maintained throughout the dynamics. N240 is fully conserved among class B1 GPCRs while R190 is R, K, or N in all class B1s, so this interaction is possible in any class B1 GPCR, and therefore may be a feature of the class at large.

TM1-TM2-TM7 Interaction Network

The remaining two strong and conserved interactions between TM1 and TM7 [Y152(TM1)-Q394(TM7)] and TM2 and TM7 [R176(TM2)-E408(TM7)] might play an

analogous role to the TM1.50(N)-TM2.50(D)-TM7.49(N) interactions conserved among class A GPCRs. Perhaps it is important to keep the TM1-TM2-TM7 (1-2-7) region rigid to control activation. Both interactions are stable in the MD. These residues are also conserved in class B1 GPCRs: Y152 may be a Y or H, while Q394 may be a Q or H. R176 and E408 are fully conserved in all class B1 GPCRs. Note that the R176(TM2)-E408(TM7) interactions are located in the intracellular end of the TMs, and as such could also play a role in G protein coupling.

We find two additional interactions in the 1-2-7 region which involve non-conserved amino acids – T149(TM1)-E387(TM7) and H180(TM2)-S163(TM1) – which further stabilize the 1-2-7 coupling. We find that the T149(TM1)-E387(TM7) interaction is only transient during dynamics, being often mediated by the H1 of the ligand. Perhaps the agonist will eventually break this interaction as part of activation. The H180(TM2)-S163(TM1) interaction forms/breaks/reforms several times during the course of the dynamics indicating that it is less stable than the hydrogen bonds discussed previously. Despite their transience, these hydrogen bonds do help stabilize the coupling of the 1-2-7 helices, and in conjunction with the more stable conserved interactions discussed earlier, form a solid grouping of TMs 1, 2, and 7. The two conserved interactions, along with these two nonconserved interactions, are pictured in **Figure 2**.

Extracellular Loop Couplings and N-terminal Interactions

The remaining four interhelical interactions are between the extracellular loops. The first three are with D222(EC1) and adjoining EC2 residues R299, W297, and C296. In addition, there is a helical region present in EC1, from residue 215 to 225. It is the base of this helix which interacts with EC2. The final inter-loop hydrogen bond is between H374(EC3) and M303(EC2). The extracellular loops are clearly closely coupled, and provide order to the flexible loop regions, as have been seen in other GPCR crystal structures⁵⁶. These stabilizing interactions would play a role in peptide binding, since they need to accommodate a peptide reaching from the N-terminus, past EC1, and into the TM bundle interior. Finally, GLPR1 also has the TM3-EC2 disulfide coupling (C296-C226) conserved among class A GPCRs and among class B1. There are no other Cys in the EC loops.

The overall architecture of the N-terminus from the crystal structure remains stable during the course of MD. There are still the three conserved disulphide bonds, two regions of antiparallel β -sheets, and an alpha helix adjacent to the ligand such as occurs in all of the class B1 GPCR N-terminal crystal structures⁵⁷. No significant deviations from the crystal are observed.

Overall, our study of GLP1R shows that there are several conserved hydrogen bond networks which mimic those of class A GPCRs. The 3-6 ionic lock is similar to those of class A GPCRs, as well as 1-2 and 2-7 networks which mimic the 1-2-7 class A motif. Finally, the loop structures, which may have direct bearing on ligand binding, are stabilized by several inter-loop interactions.

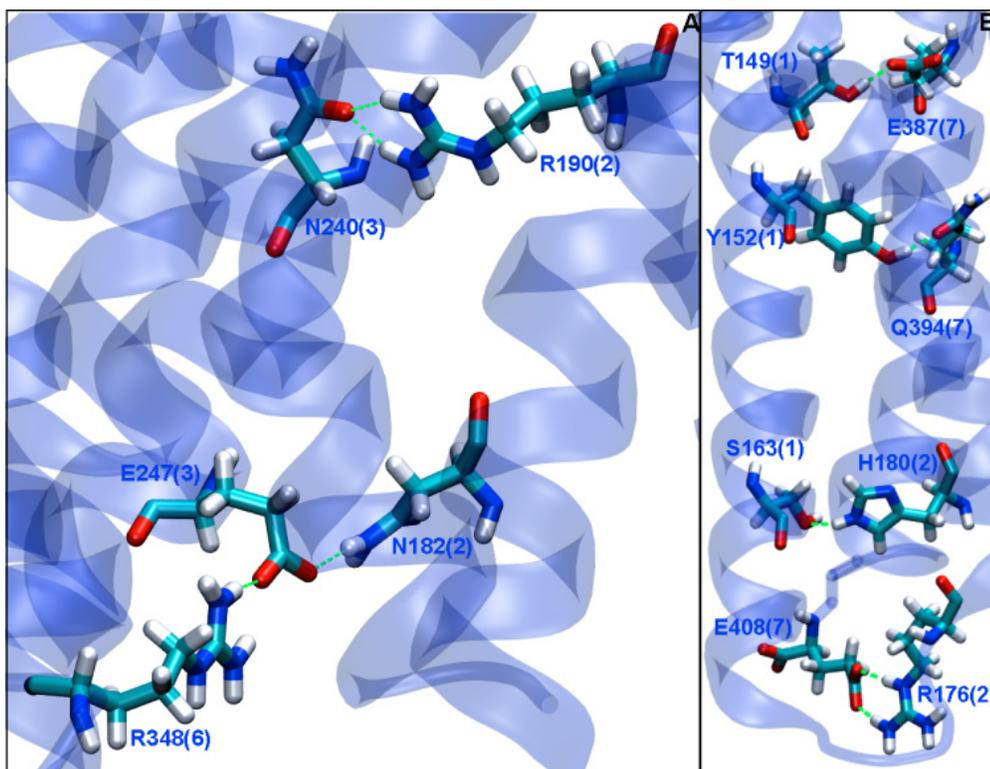


Figure 2. The (A) TM2-TM3-TM6 and (B) TM1-TM2-TM7 conserved hydrogen bond networks. (A) We believe that the E247(TM3)-R348(TM6) ionic lock may be associated with the unactivated GPCR structure (analogous to the R3.50-D/E6.30 interaction in class A GPCRs). TM3 is additionally coupled to TM2 via the conserved N182(TM2), to form a 2-3-6 interaction. This interaction is further stabilized by the R190(TM2)-N240(TM3) hydrogen bond, which may be analogous to the TM2.45(S/N/T)-3.42(S/N/T)-4.50(W) conserved interaction of class A GPCRs. (B) We also see several more transient couplings between TMs 1, 2, and 7 as shown here. These four interactions help solidify the local structure of TMs 1, 2, and 7 and their disruption may be involved in activation.

Protein-Ligand Interactions

The GLP1R/Exe4 binding site involves interactions throughout the N-terminus, TM regions, and extracellular loops with the primary interactions occurring with the N-terminus, TM1, TM2, TM7, and EC1 (see **Figure 3**). We find Exendin-4 to be helical from residues nine through twenty-nine. The strongest interactions with the N-terminus include six polar interactions (three which were present in the crystal structure) and six hydrophobic interactions (all present in the crystal structure). The TM bundle features eight polar interactions (two which were confirmed by mutation studies) and four hydrophobic interactions (two which were confirmed by mutation studies). The full cavity analysis for these interactions is shown in **Table 6**. We will discuss the binding site in two parts – hydrogen bonds and nonpolar interactions – and then the comparison to mutation data will be made in more detail.

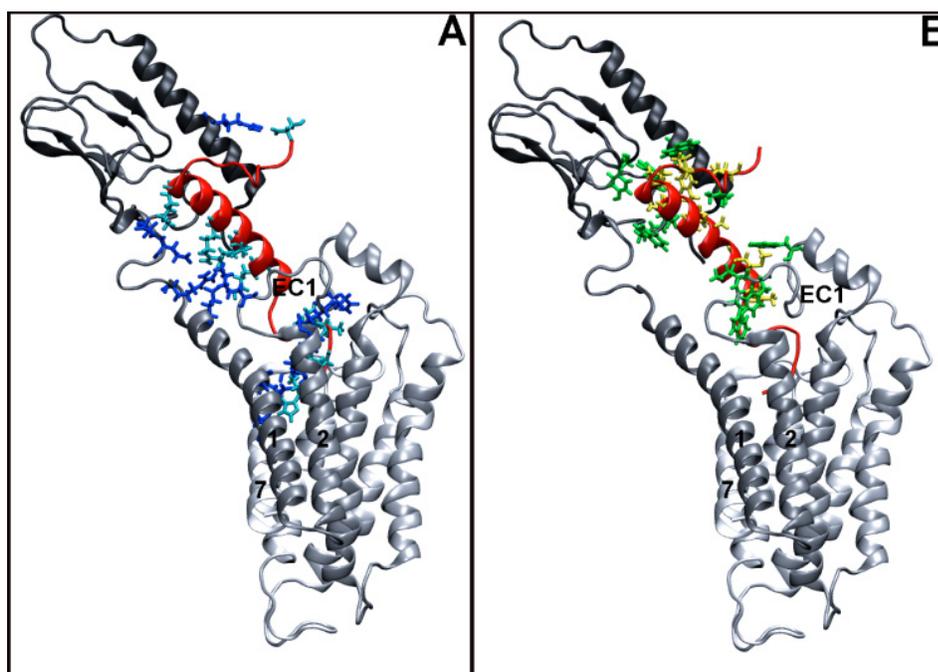


Figure 3. Overview of the ligand binding site's (A) hydrophobic and (B) hydrophilic interactions. GLP1R is shown with a color transition from black to white as the protein goes from the N-terminus to the C-terminus. Exendin-4 is shown in red. Hydrophilic interactions are shown in panel (A), with protein residues in blue and ligand residues in turquoise. Hydrophobic interactions are shown in panel (B), with protein residues in green, and ligand residues in yellow.

Polar Interactions

We find fourteen polar interactions (hydrogen bonds or salt bridges) between GLP1R and Exendin-4, as listed in **Table 4** and pictured in **Figure 4**. For the full unified cavity analysis of GLP1R/Exe4, please see **S3**. There eight polar interactions within the TM region, which reflect the primary areas of interaction between the ligand and TM bundle: TMs 1, 2, and 7, as well as EC1. The TM-region polar interactions are particularly focused at the first few residues of the ligand – specifically H1 and E3, but also with T5 and F6. Our TM-bundle interactions are further validated by site-directed mutagenesis studies for residues T149, E387, T391, and K197, which we find to interact with H1 and E3 of the ligand⁵⁸⁻⁵⁹ as shown in **Table 4**.

Our five N-terminal interactions include the two crystal salt bridges: E128(N)-R20, E127(N)-K27¹³. These two residues' importance has also been shown through mutation studies on E127 and E128⁶⁰. In addition, we find three novel very strong interactions of the N-terminus, of which two are in the flexible region between the structured N-terminus and TM1, with the other at the C-terminus of the ligand. The remaining crystal hydrogen bond between R121(N) and the backbone of K27 alternates during the MD between a water-mediated interaction and a very weak hydrogen bond. This weak interaction is confirmed by the Underwood study, who found that mutating R121 to alanine (R121A) decreased ligand binding by only 1.6-fold⁶¹.

Table 4. Polar interactions between GLP1R and Exendin-4

GLP1R Residue	Location	Exe4 Residue	Energy	Mutation	Binding Effect
Transmembrane Region Interactions					
T149	TM1	H1	-1.0	T149M ⁵⁸	5.0 - fold reduction
E387	TM7	H1	-6.2	E387A ⁶⁰	3.9 - fold reduction
T391	TM7	H1	-2.8	T391A ⁶⁰	2.8 - fold reduction
K197	TM2	E3	-39.6	K197A ⁶⁰	3.0 - fold reduction
K383	TM7	E3	-37.9		
Q211	EC1	T5 (backbone)	-1.0		
Q210	EC1	F6 (backbone)	-3.1		
K202	EC1	E17	-48.8		
N-terminal Interactions					
R134	N	E16	-47.6		
E128	N	R20	-44.8	E128A ¹³	2.4 - fold reduction *
E139	N	K12	-38.4		
E127	N	K27	-33.4	E127A ¹³	6.8 - fold reduction *
R40	N	S39	-23.3		

Energies are given in kcal/mol and are provided to show relative strength of the interactions.

* - Present in the crystal structure of nGLP1R/Exe4

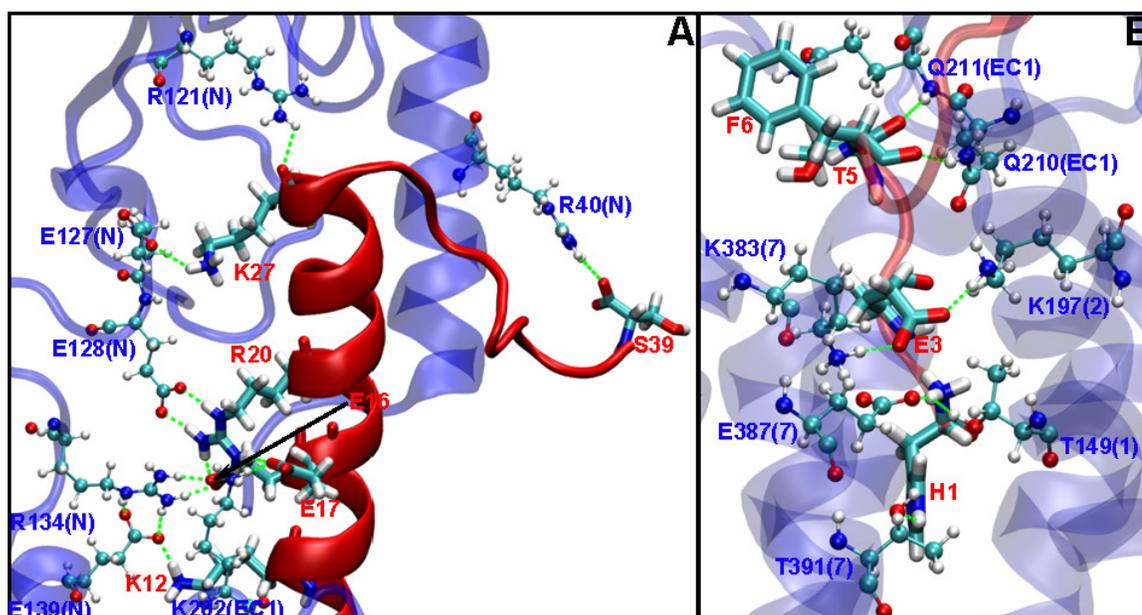


Figure 4. Exe4 hydrogen bonds with the (A) N-terminus and (B) transmembrane region of GLP1R. All of the receptor-ligand hydrogen bonds are depicted here and quantified in S3. Protein residues are shown in blue and CPK drawing method while ligand residues are shown in red and licorice drawing method. We find a total of fourteen important polar interactions of Exe4 with GLP1R of which six involve the (A) N-terminus (including three found in the x-ray crystal structure) and eight are in the (B) TM bundle or loops (two of which have been confirmed by mutation studies).

Hydrophobic Interactions

We predict 21 strong hydrophobic interactions between GLP1R and Exendin4. The ten strongest (cutoff of -3kcal/mol for the VDW energy) are shown below in **Table 5** and **Figure 5**. We find two main clusters of hydrophobic interactions.

The first cluster of hydrophobic interactions occurs in EC1. GLP1R residues W203, M204, Y205, A209, W214, and L217 interact with Exe4 residues L10 and M14. Indeed, experiments on residues M204A/Y205A found a 2.7-fold decrease in binding of Exe4⁴⁷. We find that M204 has a -7.0kcal/mol interaction energy with Exe4, while the interaction energy with Y205 is -2.6kcal/mol.

The second cluster of hydrophobic interactions occurs on one face of the helical portion of Exe4, interacting with the hydrophobic face of a helix of GLP1R in the N-terminus. These include the interactions between GLP1R residues L32, T35, V36, W39, Y69, Y88, L89, P90, W91, and L123 with Exe4 residues V18, V19, F22, L23, L26, P31, and P36. These interactions include some of the strongest hydrophobic interactions in our entire structure: L32 at -8.4kcal/mol, W39 at -6.6kcal/mol, P90 at -4.2kcal/mol, and W91 at -4.0kcal/mol. These interactions were all present in the crystal structure of nGLP1R/Exe4, and have been confirmed by mutation studies^{13,61}. Specifically, the L32A mutation had a 7.1-fold effect on Exe4 binding and 9.5-fold on activation, while P90A had a 2.1-fold effect on binding and 5.5-fold effect on activation.

We also find a final small cluster of nonpolar residues in the middle of TM3 (L232, M233, and V229, not pictured) which form a hydrophobic pocket around G2 of Exendin4. These hydrophobic interactions, plus the K202-E17 salt bridge and the weaker polar interactions with EC1 in **Table 5** (to Q210 and Q211), make it clear that EC1 is extremely important for Exe4 binding.

Table 5. Top ten hydrophobic interactions between GLP1R and Exe4.

GLP1R Residue	Location	Strongest Interacting Exe4 Residue	VDW	Coulomb	Hbond	Total	Experimental Study?
L32	N	V18, V19, F22, P36	-7.12	-1.3	0	-8.42	mutation and crystal
W203	EC1	L10	-7.38	0.01	0	-7.37	
W214	EC1	M14	-6.33	-0.8	0	-7.13	
M204	EC1	M14	-6.64	-0.35	0	-6.99	mutation
W39	N	F22, L26, P31	-5.33	-1.3	0	-6.64	crystal
P90	N	I23	-3.25	-0.9	0	-4.15	mutation and crystal
W91	N	I23	-3.85	-0.09	0	-3.95	crystal
Y69	N	L26	-3.63	0.68	0	-2.95	mutation and crystal
Y205	EC1	L10, M14	-3.5	0.91	0	-2.59	mutation
Y88	N	F22	-3.13	1.74	0	-1.39	mutation and crystal

All energies are given in kcal/mol, and are summed over the GLP1R residue interactions. The residues are ordered by their total interaction energy.

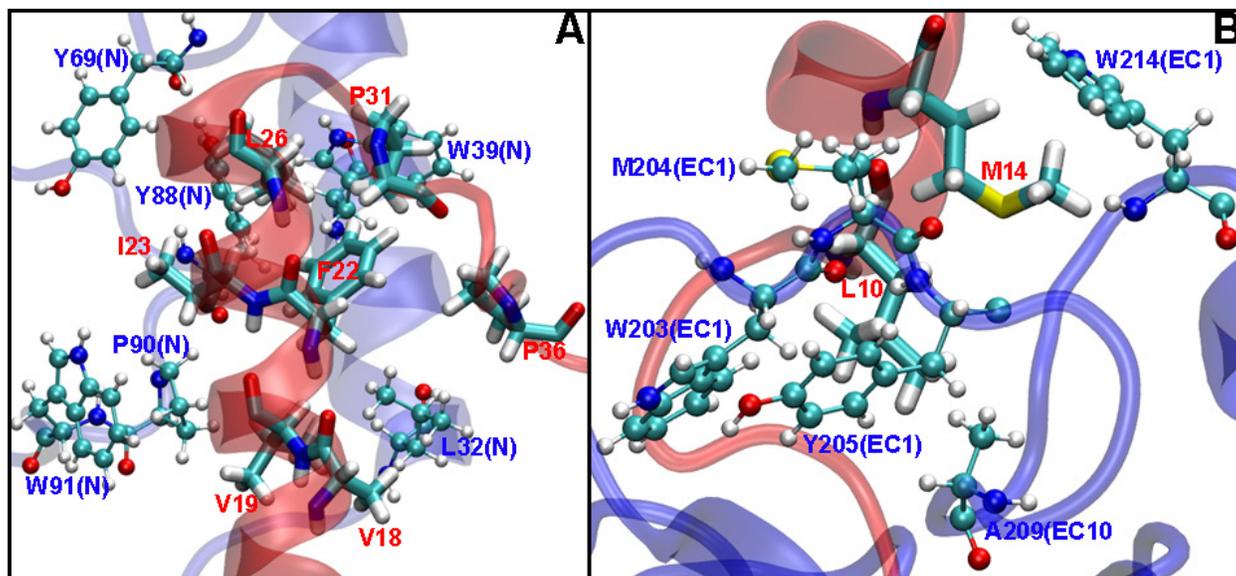


Figure 5. Hydrophobic interactions between GLP1R and Exendin-4 in the (A) N-terminus and (B) extracellular loop 1 (EC1). Protein residues are shown in blue and CPK drawing method while ligand residues are shown in red and licorice drawing method. We find ten important hydrophobic interactions of which six are in the (A) N-terminus (all confirmed by x-ray crystal structure) and four are in the (B) TM bundle (two of which have been confirmed by mutation studies).

Full Unified Cavity Analysis

Table 6. Full unified cavity analysis of the GLP1R/Exe4 Binding Site.

GLP1R Residue	Location	VDW	Coulomb	Hbond	Total
K202	EC1	4.69	-61.5	-7.98	-64.79
K197	2	6.15	-53.75	-6.01	-53.6
K383	7	4.41	-49.6	-5.58	-50.77
R134	N	10.78	-49.19	-12.24	-50.65
R40	N	3.77	-44.13	-6.39	-46.74
R121	N	-1.33	-34.39	-0.03	-35.74
K38	N	-0.25	-31.8	0	-32.05
R190	2	2.7	-33.79	0	-31.09
R43	N	-0.44	-26.51	0	-26.95
R227	EC1	-0.15	-26.75	0	-26.91
K113	N	-0.14	-24.39	0	-24.53
R44	N	-0.04	-24.46	0	-24.5
K130	N	-0.07	-22.17	0	-22.24
R102	N	-0.04	-20.61	0	-20.65
R299	EC2	-0.04	-20.33	0	-20.37
K288	4	-0.03	-20.18	0	-20.21
R376	EC3	-0.15	-19.56	0	-19.71
R131	N	-0.02	-18.75	0	-18.78
R380	EC3	-0.17	-17.68	0	-17.85
R48	N	-0.01	-16.68	0	-16.69
R64	N	-0.01	-14.67	0	-14.68
R310	5	0	-12.79	0	-12.79
R348	IC3	0	-12.56	0	-12.56
Q210	EC1	-3.71	-4.88	-3.61	-12.19
K351	6	0	-11.96	0	-11.96
R170	1	0	-11.68	0	-11.68
R176	2	0	-10.81	0	-10.81
R326	5	0	-10.05	0	-10.05
R267	4	0	-9.59	0	-9.59
K346	IC3	0	-9.26	0	-9.26
K342	IC3	0	-9.18	0	-9.18
K334	IC3	0	-0.11	0	-9.11
K336	IC3	0	-8.52	0	-8.52
L32	N	-7.12	-1.3	0	-8.42
W203	EC1	-7.38	0.01	0	-7.37
W214	EC1	-6.33	-0.8	0	-7.13
S31	N	-2.18	-4.88	0	-7.06

M204	EC1	-6.64	-0.35	0	-6.99
W39	N	-5.33	-1.3	0	-6.64
H212	EC1	-4.2	-2.11	0	-6.31
T35	N	-3.06	-1.87	0	-4.92
P90	N	-3.25	-0.9	0	-4.15
W91	N	-3.85	-0.09	0	-3.95
V36	N	-2.49	-1.2	0	-3.69
C236	3	-1.02	-2.6	0	-3.62
Q213	EC1	-0.72	-2.89	0	-3.61
L89	N	-2.35	-1.2	0	-3.55
Q211	EC1	-3.57	0.56	-0.49	-3.5
T29	N	-3.03	-0.41	0	-3.43
L232	3	-2.87	-0.51	0	-3.38
S193	2	-1.08	-2.15	0	-3.23
T391	7	-0.61	-1.33	-1.36	-3.2
Y69	N	-3.63	0.68	0	-2.95
Y205	EC1	-3.5	0.91	0	-2.59
T149	1	-1.79	-0.65	-0.05	-2.49
V229	3	-2.12	-0.2	0	-2.32
Q37	N	-0.38	-1.79	0	-2.17
L228	3	-1.43	-0.51	0	-1.93
L217	EC1	-1.07	-0.8	0	-1.87
F390	7	-1.03	-0.74	0	-1.77
A153	1	-0.43	-1.32	0	-1.76
L384	7	-1.46	-0.29	0	-1.75
F143	N	-1.23	-0.49	0	-1.72
S206	EC1	-1.55	-0.06	0	-1.62
W33	N	-0.83	-0.77	0	-1.6
A209	EC1	-0.74	-0.85	0	-1.59
L123	N	-1.61	0.04	0	-1.57
Y152	1	-0.65	-0.83	0	-1.47
Y88	N	-3.13	1.74	0	-1.39
M233	3	-2.38	1.07	0	-1.31
A239	3	-0.08	-1.13	0	-1.21
Y42	N	-0.16	-1	0	-1.17
L118	N	-0.15	-0.99	0	-1.15
V194	2	-1.12	-0.02	0	-1.14
Y235	3	-0.34	-0.77	0	-1.12
Q221	EC1	-0.18	-0.91	0	-1.09
E127	N	-1.29	14.62	-0.65	12.69
E139	N	5.11	16.23	-5.71	15.63

E128	N	11.2	16.4	-10.48	17.12
E387	7	2.87	36.66	-5.47	17.12

All interactions that are over -1kcal/mol, or include a hydrogen bond, are shown.

Energies are in kcal/mol.

Comparison to Mutation Data

Several mutation studies have been carried out on Exendin-4 with the intent of determining which residues are important for ligand binding^{47, 58, 60-62}. These studies are summarized in **Table 7**. Of the 26 mutations leading to a decrease in binding or activity, 24 are consistent with our predicted equilibrium structure, while the remaining two appear transiently in the MD. Of these 26 residues, 13 involve the N-terminus (and were in the x-ray structure) while 13 involve the TM helices and EC1. 11 of the remaining 13 residues interact with six of the seven TMs (all but TM6), plus EC1. The remaining two residues on TM6 (H363 and E364) are transiently involved in a hydrogen bond network that spans the middle of the TM bundle and extends to H1 of the ligand. Our prediction agrees with the conclusion of the mutation studies that D198 is not crucial for ligand binding⁶³. It is important to emphasize that the GLP1R structure (except for the N-terminus) was derived strictly from our MembStruk method, without any use of mutation data (except in the loop growing, which used a distance constraint between Y205(EC1)-F6, and required M204, Y205, D215, and R227 to be close to some part of Exe4).

Finally, we note that our structure preserved all interactions found in the crystal structure – both hydrophilic and hydrophobic – over the course of MD. This further validates our structure, since these interactions would be expected to be stable. In addition, all residues indicated in the literature to be potentially important for binding or activation are found in our structure to point into the TM bundle or are otherwise accessible to the ligand. As such they can either interact with the ligand itself or have structural effects. Overall, our structure strongly agrees with the available experimental data, making it valuable for further studies on GLP1R.

Table 7. Residues in the binding site with their mutation data.

GLP1R Residue	Location	Experiment	Ligand Interaction	Type of Interaction
L32	N	L32A ⁶¹ , crystal ¹³	V18, V19, F22, P36	Hydrophobic
T35	N	T35A ⁶¹ , crystal ¹³	V19, F22	Hydrophobic
V36	N	V36A ⁶¹ , crystal ¹³	F22	Hydrophobic
W39	N	crystal ¹³	F22, L26, P31	Hydrophobic
Y69	N	Y69A ⁶¹ , crystal ¹³	L26	Hydrophobic
Y88	N	Y88A ⁶¹ , crystal ¹³	F22	Hydrophobic
L89	N	L89A ⁶¹ , crystal ¹³	I23	Hydrophobic
P90	N	P90A ⁶¹ , crystal ¹³	I23	Hydrophobic
W91	N	crystal ¹³	I23	Hydrophobic
R121	N	R121A ⁶¹ , crystal ¹³	K27	Hbond
L123	N	L123A ⁶¹ , crystal ¹³	I23	Hydrophobic
E127	N	E127A/E127Q ⁶¹ , crystal ¹³	K27	Hbond
E128	N	E128A/E128Q ⁶¹ , crystal ¹³	R20	Hbond
T149	1	T149M ⁵⁸	H1	Hbond
Y152	1	Y152A ⁶⁰	E3	Polar
R190	2	R190A ⁶⁰	E3	Polar
K197	2	K197A ⁶⁰	E3	Hbond
M204	EC1	M204A/Y205A ⁴⁷	M14	Hydrophobic
Y205	EC1	M204A/Y205A ⁴⁷	L10, M14	Hydrophobic
Y235	3	Y235A ⁶⁰	E3	Polar
K288	4	K288A ⁶²	E3	Polar
R310	5	R310A ⁶⁰	E3	Polar
E387	7	E387A ⁶⁰	H1	Hbond
T391	7	T391A ⁶⁰	H1	Hbond

Testing the Predicted Binding Site

Our predicted structure of GLP1R and its binding site with Exendin-4 suggests many mutation studies for testing its validity. Indeed, we constructed new structures for 14 such mutations Q213K, S11W, R134A/Q, K202A/Q, K383A/Q, W203N/Y, and W214N/Y (see **Table 8** for their effects on energies). The procedure was to SCREAM in the desired mutation, and then minimize the protein to 0.5RMS force¹⁶. These calculations assumed that the overall backbone structure remains intact. Of the 14

mutations, two are predicted to improve binding, while 12 are predicted to decrease binding.

The first set of mutations was chosen with the goal of validating our predictions of the strongest interactions between GLP1R and Exendin-4. Ten cases were aimed at disrupting the binding site by modifying protein residues: R134A/Q, K202A/Q, K383A/Q, W203N/Y, and W214N/Y. In each case our predicted change in binding agrees with expectation. The R134A/Q mutations both break the salt bridge with E16. The K202 mutations break the E17 salt bridge, although a weaker hydrogen bond does remain in the K202Q case. The K383 mutations both break the E3 hydrogen bond. The remaining tryptophan mutations were made with the goal of disrupting the EC1-ligand hydrophobic interactions. This was achieved in all of the (W203N/Y and W214N/Y) cases. We also suggest two ligand mutations that would decrease binding: K12A and M14Q. The K12A mutation would lose the E139 salt bridge, while the M14Q would disrupt the hydrogen bond network M14 has in the EC1 area. These predictions were confirmed.

Finally, we suggest two mutations we predict to improve ligand binding. The first mutation was S11W of Exendin-4; which we predict allows a new interaction to be formed with W214. The second change is Q213K of GLP1R, which we suggest forms a new hydrogen bond with D9. Both of these changes improve our predicted interaction energies.

Table 8. Suggested mutations for GLP1R or Exendin-4.

Mutation	Location	Motivation	Change in Interaction Energy
Q213K	Protein	New Hydrogen Bond with D9	-49.4
S11W	Ligand	New Hydrophobic Interaction with W214	-1.8
WT	WT	WT	0
W214Y	Protein	Lose M14 Hydrophobic Interaction	1.8
W203Y	Protein	Lose L10 Hydrophobic Interaction	3
M14Q	Ligand	Loses Hydrophobic Interactions	4.6
W203N	Protein	Lose L10 Hydrophobic Interaction	4.8
W214N	Protein	Lose M14 Hydrophobic Interaction	7.5
K383Q	Protein	Lose E3 Hydrogen Bond	45.2
K383A	Protein	Lose E3 Hydrogen Bond	57.3
K202Q	Protein	Lose E17 Salt Bridge	60.2
R134A	Protein	Lose E16 Salt Bridge	60.7
R134Q	Protein	Lose E16 Salt Bridge	61.9
K202A	Protein	Lose E17 Salt Bridge or Hydrogen Bond	68.2
K12A	Ligand	Lose E139 Salt bridge	80.4

All energies are given in kcal/mol. WT=wild type.

Discussion of Ligand Binding and Protein Activation

Our GLP1R/Exendin-4 structure suggests several general features of binding to GLP1R and potentially class B1 proteins as a whole. The first is that we find that the TM-region polar interactions are particularly focused at the first few residues of the ligand – specifically H1 and E3, but also at T5 and F6. This is in accordance with the known importance of the N-terminus (specifically residues 1 through 7) of class B1 agonists for protein activation⁵⁷. In the specific case of Exendin-4, if the ligand is truncated by eight residues it becomes a competitive antagonist for GLP1R since it can still bind the receptor, but can no longer cause activation¹⁰.

Next, we find that Exendin-4's binding pocket shows strong polar and hydrophobic interactions with EC1. Experimental studies showed that mutations of EC2 residues to alanine dramatically decreased binding of GLP1, but had no effect on the binding of Exe4⁶⁴, consistent with our structure. We suggest that the reason for the importance of the loops in the peptide binding is to align the ligand in the correct conformation for TM-bundle entry. In the two-domain model of class B1 GPCR protein binding, the N-terminal ectodomain plays the role of recognizing the ligand and supports the initial binding⁶⁵.

We believe that the next step is for the flexible N-terminus/ligand complex to align itself to the TM bundle via loop interactions, followed by final insertion of the head region of the ligand into the TM bundle itself.

Thirdly, we note that the flexible N-terminus of the ligand is nestled in the TM1-TM2-TM7 binding pocket. This leaves the TM3-TM4-TM6 region largely open, making this area available for binding of small-molecules serving as ago-allosteric modulators⁶⁶. In addition, the residues of the ligand inserted themselves between hydrogen bonds of the apo GLP1R; for example, T149(TM1)-H1-E387(TM7). This suggests that part of the effect of Exendin-4 binding may be to break some of the TM1-TM2-TM7 strong interactions, giving the structure the flexibility to achieve its active conformation.

Any discussion of GPCR activation would be incomplete without mention of the TM3-TM6 ionic lock, which we find a variation of in our structure. Instead of the conserved R(3.50)-D/E(6.30) ionic lock of class A GPCRs we find an analogous conserved E247(TM3)-R348(TM6) ionic lock. Breaking this interaction may be crucial for GLP1R activation. To test this hypothesis, one could mutate one of the charged residues to a polar residue such as a glutamine so that the overall hydrophilicity of the region could be preserved, but the interaction broken.

Finally, it has been suggested that an N-terminal helix capping motif of a peptide agonist is a key element underlying class B1 GPCR activation². This structural feature is theorized to consist of a hydrophobic interaction between residues 6 and 10 and is stabilized by a salt bridge between residues 7 and 10 of the ligand. The result of these interactions is that the ligand forms an “L” at its N-terminus. While we do not see the exact 7-10 and 6-10 interactions – instead we find that residues Phe6 and Ser8 form a backbone hydrogen bond – this alternate interaction causes the ligand to adopt a slightly more loose “L” conformation. This structural constraint may be important for the rational drug design of peptide agonists targeting class B1 receptors.

Conclusion

We present here the predicted TM bundle for GLP1R (residues 146-408) which we combined with the crystal structure for the N-terminus (residues 28-145). The resulting structure was then equilibrated in full periodic membrane and water for 28ns. This is the

first full-structure prediction of GLP1R bound to Exendin-4. We find strong agreement with available experimental data, most of which played no role in the predictions. We suggest 14 new mutations to provide strong tests of our predicted binding site. This structure can now form the basis for rational design of other peptide ligands and greatly needed small molecule ligands.

The model we present here can be used to further explore the method of class B1 GPCR binding and activation⁶⁰. In addition, we expect that this structure will provide a basis for design and optimization of new small-molecule ligands that bind selectively and specifically to GLP1R.

Finally, one of the grand challenges in understanding GPCRs is to elucidate the mechanism of activation. This study does not address this issue directly; however, we do identify a TM3-TM6 ionic lock which is conserved across class B1 GPCRs that we believe may play a very similar role in activation to the TM3-TM6 ionic lock conserved across a subset of class A GPCRs. Mutation studies on this ionic lock will help test such speculations and provide structural signatures of active and inactive receptor conformations.

References

1. Parthier, C.; Reedtz-Runge, S.; Rudolph, R.; Stubbs, M. T., Passing the Baton in Class B Gpcrs: Peptide Hormone Activation Via Helix Induction? *Trends Biochem Sci* **2009**, *34* (6), 303-10.
2. Neumann, J. M.; Couvineau, A.; Murail, S.; Lacapere, J. J.; Jamin, N.; Laburthe, M., Class-B Gpcr Activation: Is Ligand Helix-Capping the Key? *Trends Biochem Sci* **2008**, *33* (7), 314-9.
3. Pal, K.; Melcher, K.; Xu, H. E., Structure and Mechanism for Recognition of Peptide Hormones by Class B G-Protein-Coupled Receptors. *Acta Pharmacol Sin* **2012**, *33* (3), 300-11.
4. Schmidt, W. E.; Siegel, E. G.; Creutzfeldt, W., Glucagon-Like Peptide-1 but Not Glucagon-Like Peptide-2 Stimulates Insulin Release from Isolated Rat Pancreatic Islets. *Diabetologia* **1985**, *28* (9), 704-7.
5. Kieffer, T. J.; Habener, J. F., The Glucagon-Like Peptides. *Endocr Rev* **1999**, *20* (6), 876-913.
6. Kieffer, T. J.; McIntosh, C. H.; Pederson, R. A., Degradation of Glucose-Dependent Insulinotropic Polypeptide and Truncated Glucagon-Like Peptide 1 in Vitro and in Vivo by Dipeptidyl Peptidase Iv. *Endocrinology* **1995**, *136* (8), 3585-96.
7. Mentlein, R.; Gallwitz, B.; Schmidt, W. E., Dipeptidyl-Peptidase Iv Hydrolyses Gastric Inhibitory Polypeptide, Glucagon-Like Peptide-1(7-36)Amide, Peptide Histidine Methionine and Is Responsible for Their Degradation in Human Serum. *Eur J Biochem* **1993**, *214* (3), 829-35.
8. Eng, J.; Kleinman, W. A.; Singh, L.; Singh, G.; Raufman, J. P., Isolation and Characterization of Exendin-4, an Exendin-3 Analogue, from Heloderma Suspectum Venom. Further Evidence for an Exendin Receptor on Dispersed Acini from Guinea Pig Pancreas. *J Biol Chem* **1992**, *267* (11), 7402-5.
9. Parkes, D. G.; Pittner, R.; Jodka, C.; Smith, P.; Young, A., Insulinotropic Actions of Exendin-4 and Glucagon-Like Peptide-1 in Vivo and in Vitro. *Metabolism* **2001**, *50* (5), 583-9.
10. Goke, R.; Fehmann, H. C.; Linn, T.; Schmidt, H.; Krause, M.; Eng, J.; Goke, B., Exendin-4 Is a High Potency Agonist and Truncated Exendin-(9-39)-Amide an Antagonist at the Glucagon-Like Peptide 1-(7-36)-Amide Receptor of Insulin-Secreting Beta-Cells. *J Biol Chem* **1993**, *268* (26), 19650-5.
11. Elashoff, M.; Matveyenko, A. V.; Gier, B.; Elashoff, R.; Butler, P. C., Pancreatitis, Pancreatic, and Thyroid Cancer with Glucagon-Like Peptide-1-Based Therapies. *Gastroenterology* **2011**, *141* (1), 150-156.
12. Trabanino, R. J.; Hall, S. E.; Vaidehi, N.; Floriano, W. B.; Kam, V. W.; Goddard, W. A., 3rd, First Principles Predictions of the Structure and Function of G-Protein-Coupled Receptors: Validation for Bovine Rhodopsin. *Biophys J* **2004**, *86* (4), 1904-21.
13. Runge, S.; Thogersen, H.; Madsen, K.; Lau, J.; Rudolph, R., Crystal Structure of the Ligand-Bound Glucagon-Like Peptide-1 Receptor Extracellular Domain. *J Biol Chem* **2008**, *283* (17), 11340-7.
14. Abrol, R.; Griffith, A.; Bray, J.; Goddard, W. A., 3rd, Structure Prediction of G Protein-Coupled Receptors and Their Ensemble of Functionally Important Conformations. In *Methods in Molecular Biology Special Volume on Membrane Protein Structure Determination and Prediction Methods*, Vaidehi, N.; J, K.-S., Eds. 2011.

15. Li, Y.; Zhu, F.; Vaidehi, N.; Goddard, W. A., 3rd; Sheinerman, F.; Reiling, S.; Morize, I.; Mu, L.; Harris, K.; Ardati, A.; Laoui, A., Prediction of the 3d Structure and Dynamics of Human Dp G-Protein Coupled Receptor Bound to an Agonist and an Antagonist. *J Am Chem Soc* **2007**, *129* (35), 10720-31.
16. Kam, V. W. T.; Goddard, W. A., 3rd, Flat-Bottom Strategy for Improved Accuracy in Protein Side-Chain Placements. *Journal of Chemical Theory and Computation* **2008**, *4* (12), 2160-2169.
17. Grace, C. R.; Perrin, M. H.; Gulyas, J.; Digruccio, M. R.; Cattle, J. P.; Rivier, J. E.; Vale, W. W.; Riek, R., Structure of the N-Terminal Domain of a Type B1 G Protein-Coupled Receptor in Complex with a Peptide Ligand. *Proc Natl Acad Sci U S A* **2007**, *104* (12), 4858-63.
18. Chen, R.; Li, L.; Weng, Z. P., Zdock: An Initial-Stage Protein-Docking Algorithm. *Proteins-Structure Function and Genetics* **2003**, *52* (1), 80-87.
19. Neidigh, J. W.; Fesinmeyer, R. M.; Prickett, K. S.; Andersen, N. H., Exendin-4 and Glucagon-Like-Peptide-1: Nmr Structural Comparisons in the Solution and Micelle-Associated States. *Biochemistry* **2001**, *40* (44), 13188-200.
20. Eswar, N.; Webb, B.; Marti-Renom, M. A.; Madhusudhan, M. S.; Eramian, D.; Shen, M. Y.; Pieper, U.; Sali, A., Comparative Protein Structure Modeling Using Modeller. *Curr Protoc Bioinformatics* **2006**, *Chapter 5*, Unit 5 6.
21. Phillips, J. C.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R. D.; Kale, L.; Schulten, K., Scalable Molecular Dynamics with Namd. *Journal of Computational Chemistry* **2005**, *26* (16), 1781-1802.
22. MacKerell, A. D.; Bashford, D.; Bellott, M.; Dunbrack, R. L.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; Joseph-McCarthy, D.; Kuchnir, L.; Kuczera, K.; Lau, F. T. K.; Mattos, C.; Michnick, S.; Ngo, T.; Nguyen, D. T.; Prodhom, B.; Reiher, W. E.; Roux, B.; Schlenkrich, M.; Smith, J. C.; Stote, R.; Straub, J.; Watanabe, M.; Wiorkiewicz-Kuczera, J.; Yin, D.; Karplus, M., All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins. *Journal of Physical Chemistry B* **1998**, *102* (18), 3586-3616.
23. Brooks, B. R.; Brooks, C. L., 3rd; Mackerell, A. D., Jr.; Nilsson, L.; Petrella, R. J.; Roux, B.; Won, Y.; Archontis, G.; Bartels, C.; Boresch, S.; Caflisch, A.; Caves, L.; Cui, Q.; Dinner, A. R.; Feig, M.; Fischer, S.; Gao, J.; Hodoscek, M.; Im, W.; Kuczera, K.; Lazaridis, T.; Ma, J.; Ovchinnikov, V.; Paci, E.; Pastor, R. W.; Post, C. B.; Pu, J. Z.; Schaefer, M.; Tidor, B.; Venable, R. M.; Woodcock, H. L.; Wu, X.; Yang, W.; York, D. M.; Karplus, M., Charmm: The Biomolecular Simulation Program. *Journal of Computational Chemistry* **2009**, *30* (10), 1545-614.
24. Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L., Comparison of Simple Potential Functions for Simulating Liquid Water. *Journal of Chemical Physics* **1983**, *79* (2), 926-935.
25. Lim, K. T.; Brunett, S.; Iotov, M.; McClurg, R. B.; Vaidehi, N.; Dasgupta, S.; Taylor, S.; Goddard, W. A., 3rd, Molecular Dynamics for Very Large Systems on Massively Parallel Computers: The Mpsim Program. *Journal of Computational Chemistry* **1997**, *18* (4), 501-521.
26. Mayo, S. L.; Olafson, B. D.; Goddard, W. A., 3rd, Dreiding - a Generic Force-Field for Molecular Simulations. *Journal of Physical Chemistry* **1990**, *94* (26), 8897-8909.

27. Gasteiger, E.; Gattiker, A.; Hoogland, C.; Ivanyi, I.; Appel, R. D.; Bairoch, A., ExPASy: The Proteomics Server for in-Depth Protein Knowledge and Analysis. *Nucleic Acids Res* **2003**, *31* (13), 3784-8.
28. Thompson, J. D.; Higgins, D. G.; Gibson, T. J., Clustal W: Improving the Sensitivity of Progressive Multiple Sequence Alignment through Sequence Weighting, Position-Specific Gap Penalties and Weight Matrix Choice. *Nucleic Acids Res* **1994**, *22* (22), 4673-80.
29. Unger, V. M.; Hargrave, P. A.; Baldwin, J. M.; Schertler, G. F. X., Arrangement of Rhodopsin Transmembrane Alpha-Helices. *Nature* **1997**, *389* (6647), 203-206.
30. Bray, J. K.; Goddard, W. A., 3rd, The Structure of Human Serotonin 2c G-Protein-Coupled Receptor Bound to Agonists and Antagonists. *J Mol Graph Model* **2008**, *27* (1), 66-81.
31. Popot, J. L.; Engelman, D. M., Membrane-Protein Folding and Oligomerization - the 2-Stage Model. *Biochemistry* **1990**, *29* (17), 4031-4037.
32. Zhou, F. X.; Cocco, M. J.; Russ, W. P.; Brunger, A. T.; Engelman, D. M., Interhelical Hydrogen Bonding Drives Strong Interactions in Membrane Proteins. *Nature Structural Biology* **2000**, *7* (2), 154-160.
33. Heo, J.; Goddard, W. A., Predicted Structure of the Agonist Peptide Binding Site in the N-Terminal Ectodomain of Glucagon-Like Peptide 1 Receptor. *to be submitted*.
34. Sali, A.; Blundell, T. L., Comparative Protein Modelling by Satisfaction of Spatial Restraints. *J Mol Biol* **1993**, *234* (3), 779-815.
35. Kneller, D.
Nnpredict. <http://www.bioinf.manchester.ac.uk/dbbrowser/bioactivity/nnpredictfrm.html>.
36. Garnier, J.; Gibrat, J. F.; Robson, B., Gor Method for Predicting Protein Secondary Structure from Amino Acid Sequence. *Methods Enzymol* **1996**, *266*, 540-53.
37. Lin, K.; Simossis, V. A.; Taylor, W. R.; Heringa, J., A Simple and Fast Secondary Structure Prediction Method Using Hidden Neural Networks. *Bioinformatics* **2005**, *21* (2), 152-9.
38. Levin, J. M.; Robson, B.; Garnier, J., An Algorithm for Secondary Structure Determination in Proteins Based on Sequence Similarity. *FEBS Lett* **1986**, *205* (2), 303-8.
39. Buchan, D. W.; Ward, S. M.; Lobley, A. E.; Nugent, T. C.; Bryson, K.; Jones, D. T., Protein Annotation and Modelling Servers at University College London. *Nucleic Acids Res* **2010**, *38* (Web Server issue), W563-8.
40. Pollastri, G.; McLysaght, A., Porter: A New, Accurate Server for Protein Secondary Structure Prediction. *Bioinformatics* **2005**, *21* (8), 1719-20.
41. Cole, C.; Barber, J. D.; Barton, G. J., The Jpred 3 Secondary Structure Prediction Server. *Nucleic Acids Res* **2008**, *36* (Web Server issue), W197-201.
42. Adelhorst, K.; Hedegaard, B. B.; Knudsen, L. B.; Kirk, O., Structure-Activity Studies of Glucagon-Like Peptide-1. *J Biol Chem* **1994**, *269* (9), 6275-8.
43. Raghava, G. P. S., Apssp2 : A Combination Method for Protein Secondary Structure Prediction Based on Neural Network and Example Based Learning. *CASP5* **2002**, *A-132*.
44. Lopez de Maturana, R.; Treece-Birch, J.; Abidi, F.; Findlay, J. B. C.; Donnelly, D., Met-204 and Tyr-205 Are Together Important for Binding Glp-1 Receptor Agonists but Not Their N-Terminally Truncated Analogues *Protein & Peptide Letters* **2004**, *11* (1), 15-22.

45. Adelhorst, K.; Hedegaard, B. B.; Knudsen, L. B.; Kirk, O., Structure-Activity Studies of Glucagon-Like Peptide-1. *Journal of Biological Chemistry* **1994**, *269* (9), 6275-6278.
46. Mann, R. J.; Al-Sabah, S.; Lopez de Maturana, R.; Sinfield, J. K.; Donnelly, D., Functional Coupling of Cys-226 and Cys-296 in the Glucagon-Like Peptide-1 (Glp-1) Receptor Indicates a Disulfide Bond That Is Close to the Activation Pocket. *Peptides* **2010**, *31* (12), 2289-93.
47. Lopez de Maturana, R.; Treece-Birch, J.; Abidi, F.; Findlay, J. B.; Donnelly, D., Met-204 and Tyr-205 Are Together Important for Binding Glp-1 Receptor Agonists but Not Their N-Terminally Truncated Analogues. *Protein Pept Lett* **2004**, *11* (1), 15-22.
48. Xiao, Q.; Jeng, W.; Wheeler, M. B., Characterization of Glucagon-Like Peptide-1 Receptor-Binding Determinants. *Journal of Molecular Endocrinology* **2000**, *25* (3), 321-335.
49. Humphrey, W.; Dalke, A.; Schulten, K., Vmd: Visual Molecular Dynamics. *Journal of Molecular Graphics* **1996**, *14* (1), 33-&.
50. Abrol, R.; Bray, J. K.; Goddard, W. A., 3rd, Bihelix: Towards De Novo Structure Prediction of an Ensemble of G-Protein Coupled Receptor Conformations. *Proteins* **2012**, *80* (2), 505-18.
51. Bray, J. K.; Abrol, R.; Goddard, W. A., 3rd; Trzaskowski, B.; Scott, C. E., Superbihelix Method for Predicting the Pleiotropic Ensemble of G-Protein-Coupled Receptor Conformations. *Proc Natl Acad Sci U S A* **2014**, *111* (1), E72-8.
52. Vogel, R.; Mahalingam, M.; Ludeke, S.; Huber, T.; Siebert, F.; Sakmar, T. P., Functional Role of the "Ionic Lock"--an Interhelical Hydrogen-Bond Network in Family a Heptahelical Receptors. *J Mol Biol* **2008**, *380* (4), 648-55.
53. Fortin, J. P.; Schroeder, J. C.; Zhu, Y.; Beinborn, M.; Kopin, A. S., Pharmacological Characterization of Human Incretin Receptor Missense Variants. *Journal of Pharmacology and Experimental Therapeutics* **2010**, *332* (1), 274-80.
54. Heller, R. S.; Kieffer, T. J.; Habener, J. F., Point Mutations in the First and Third Intracellular Loops of the Glucagon-Like Peptide-1 Receptor Alter Intracellular Signaling. *Biochem Biophys Res Commun* **1996**, *223* (3), 624-32.
55. Takhar, S.; Gyomory, S.; Su, R. C.; Mathi, S. K.; Li, X.; Wheeler, M. B., The Third Cytoplasmic Domain of the Glp-1[7-36 Amide] Receptor Is Required for Coupling to the Adenylyl Cyclase System. *Endocrinology* **1996**, *137* (5), 2175-8.
56. Katriitch, V.; Cherezov, V.; Stevens, R. C., Diversity and Modularity of G Protein-Coupled Receptor Structures. *Trends in Pharmacological Sciences* **2012**, *33* (1), 17-27.
57. Couvineau, A.; Laburthe, M., The Family B1 Gpcr: Structural Aspects and Interaction with Accessory Proteins. *Curr Drug Targets* **2012**, *13* (1), 103-15.
58. Beinborn, M.; Worrall, C. I.; McBride, E. W.; Kopin, A. S., A Human Glucagon-Like Peptide-1 Receptor Polymorphism Results in Reduced Agonist Responsiveness. *Regul Pept* **2005**, *130* (1-2), 1-6.
59. 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.; Granier, S., Crystal Structure of the Mu-Opioid Receptor Bound to a Morphinan Antagonist. *Nature* **2012**, *485* (7398), 321-U170.
60. Coopman, K.; Wallis, R.; Robb, G.; Brown, A. J.; Wilkinson, G. F.; Timms, D.; Willars, G. B., Residues within the Transmembrane Domain of the Glucagon-Like Peptide-1 Receptor Involved in Ligand Binding and Receptor Activation: Modelling the Ligand-Bound Receptor. *Mol Endocrinol* **2011**, *25* (10), 1804-18.

61. Underwood, C. R.; Garibay, P.; Knudsen, L. B.; Hastrup, S.; Peters, G. H.; Rudolph, R.; Reedtz-Runge, S., Crystal Structure of Glucagon-Like Peptide-1 in Complex with the Extracellular Domain of the Glucagon-Like Peptide-1 Receptor. *Journal of Biological Chemistry* **2010**, *285* (1), 723-730.
62. Al-Sabah, S.; Donnelly, D., A Model for Receptor-Peptide Binding at the Glucagon-Like Peptide-1 (Glp-1) Receptor through the Analysis of Truncated Ligands and Receptors. *Br J Pharmacol* **2003**, *140* (2), 339-46.
63. Lopez de Maturana, R.; Donnelly, D., The Glucagon-Like Peptide-1 Receptor Binding Site for the N-Terminus of Glp-1 Requires Polarity at Asp198 Rather Than Negative Charge. *FEBS Lett* **2002**, *530* (1-3), 244-8.
64. Koole, C.; Wootten, D.; Simms, J.; Miller, L. J.; Christopoulos, A.; Sexton, P. M., Second Extracellular Loop of Human Glucagon-Like Peptide-1 Receptor (Glp-1r) Has a Critical Role in Glp-1 Peptide Binding and Receptor Activation. *J Biol Chem* **2012**, *287* (6), 3642-58.
65. Hoare, S. R., Mechanisms of Peptide and Nonpeptide Ligand Binding to Class B G-Protein-Coupled Receptors. *Drug Discov Today* **2005**, *10* (6), 417-27.
66. Knudsen, L. B.; Kiel, D.; Teng, M.; Behrens, C.; Bhumralkar, D.; Kodra, J. T.; Holst, J. J.; Jeppesen, C. B.; Johnson, M. D.; de Jong, J. C.; Jorgensen, A. S.; Kercher, T.; Kostrowicki, J.; Madsen, P.; Olesen, P. H.; Petersen, J. S.; Poulsen, F.; Sidemann, U. G.; Sturis, J.; Truesdale, L.; May, J.; Lau, J., Small-Molecule Agonists for the Glucagon-Like Peptide 1 Receptor. *Proc Natl Acad Sci U S A* **2007**, *104* (3), 937-42.

Chapter IV:
Transmembrane Region and Small Molecule
Binding Site Predictions for Nine Class B1
G Protein-Coupled Receptors

Abstract

Class B1 G protein-coupled receptors (GPCRs) are a subfamily of GPCR which share little to no sequence homology with the more well-characterized class A GPCRs. In the last year, two transmembrane region crystal structures of class B1 GPCRs have been obtained, that of the corticotropin-releasing factor 1 receptor (CRFR1)¹ and the glucagon receptor (GLR)². We use these two structures as starting points for the well-validated GEnSeMBLE method of structure prediction on seven novel class B1 GPCRs (CALRL, GIPR, GLP1R, PACR, PTH1R, VIPR1, and VIPR2), as well as the two crystallized proteins. An analysis of the predicted structures shows two fully conserved interactions across the family: R2.46-E3.50 and R/K6.37-Y/F7.57, as well as several more interactions present in at least half of the structures. The nine predicted structures are docked to small molecule ligands, revealing a consistent binding site in the center of the protein, between transmembrane helices 3 and 7. This binding site differs from that of class A GPCRs, and may be characteristic of the class B1 receptors. We propose protein mutation experiments to test the validity of our binding sites. Additionally, we suggest the characteristics of an optimal ligand for the binding site. Our work provides a strong starting point for drug design targeting these receptors, on top of characterizing a protein family with little full-atom information available.

Introduction

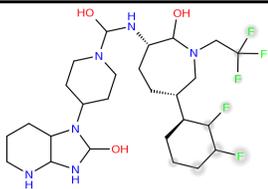
In the last year, two class B1 GPCRs have been crystallized – the corticotropin-releasing factor receptor 1 (CRFR1, 3.0Å resolution)¹ and the glucagon receptor (GLR, 3.4Å resolution)². CRFR1 was crystallized with a small molecule antagonist CP-376395. Likewise, GLR was crystallized with the small molecule antagonist NNC0640, but the ligand was not resolved in the crystal structure. In both cases, modifications were made to the protein to allow for crystallization to occur. For CRFR1, twelve amino acids were mutated and the T4 lysozyme was inserted into the intracellular loop 2.¹ For GLR, the thermally stable *E. coli* apocytochrome b₅₆₂RIL replaced the N-terminus of the protein.²

The recent availability of class B1 GPCR transmembrane region crystal structures provides high quality starting structures for the predictions of other class B1 GPCR

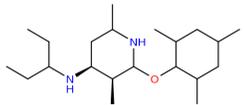
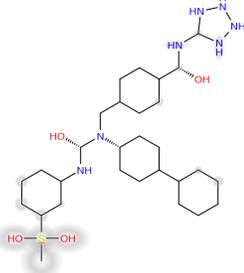
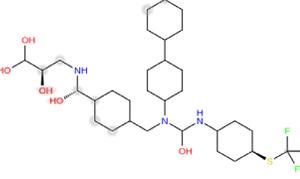
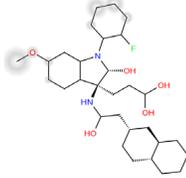
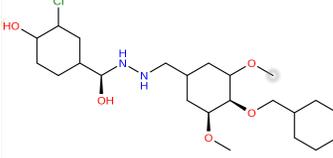
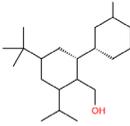
transmembrane region structures and binding sites. We present here the GENSeMBLE predicted ensemble of 3D structures of nine class B1 GPCRs, along with their binding sites to small molecule ligands (**Table 1**). Very little is known about how these ligands interact with their proteins so any knowledge we can obtain about binding sites is novel and useful in optimizing the ligands for possible drug use.

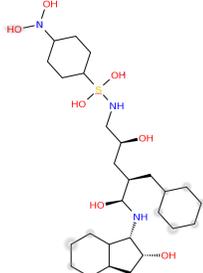
The nine class B1 GPCRs we will be predicting the structures of are the calcitonin gene-related peptide type 1 receptor (CALRL), corticotropin-releasing factor receptor 1 (CRFR1), glucagon receptor (GLR), gastric inhibitory polypeptide receptor (GIPR), glucagon-like peptide-1 receptor (GLP1R), pituitary adenylate cyclase-activating polypeptide type I receptor (PACR), parathyroid hormone/parathyroid hormone-related peptide receptor (PTH1R), vasoactive intestinal polypeptide receptor 1 (VIPR1), and vasoactive intestinal polypeptide receptor 2 (VIPR2).^{a,3} These nine receptors were chosen based on their evolutionary diversity (**Figure 1**), and the fact that small molecule ligands were known for each receptor. We also predict the structures of the wild type CRFR1 and GLR which already have x-ray crystal structures for two reasons: to validate our methodology as well as to determine if the crystallization procedures affected the conformation of the GPCRs. The goal of this work is to obtain high quality transmembrane region and binding site predictions for nine class B1 GPCRs, to examine the similarities and differences in the structures and binding sites of the receptors, and to suggest optimizations of the nine ligands which may improve their efficacy as drug candidates.

Table 1. The nine class B1 GPCRs and their small molecule ligands.^{1-2, 4-10}

Receptor	Principal Biological Actions	Small Molecule Ligand
Calcitonin gene-related peptide type 1 receptor (CALRL)	Vasodilation, transmission of pain	 MK-0974

^a The abbreviation scheme used here is that of the Universal Protein Resource (UniProt). In particular, note that the glucagon receptor is sometimes abbreviated as GCGR and the corticotropin-releasing factor receptor 1 as CRF1R.

<p>Corticotropin-releasing factor receptor 1(CRFR1)</p>	<p>ACTH release, central stress response</p>	 <p>CP-376395</p>
<p>Glucagon receptor (GLR)</p>	<p>Regulation of blood glucose</p>	 <p>NNC0640</p>
<p>Gastric inhibitory polypeptide receptor (GIPR)</p>	<p>Insulin secretion</p>	 <p>Mol 29</p>
<p>Glucagon-like peptide-1 receptor (GLP1R)</p>	<p>Insulin and glucagon secretion</p>	 <p>T0632</p>
<p>Pituitary adenylate cyclase-activating polypeptide type I receptor (PACR)</p>	<p>Neurotransmission, neuroendocrine functions</p>	 <p>Mol 1</p>
<p>Parathyroid hormone/parathyroid hormone-related peptide receptor (PTH1R)</p>	<p>Ca²⁺ homeostasis, developmental regulator</p>	 <p>SW106</p>
<p>Vasoactive intestinal polypeptide receptor 1 (VIPR1)</p>	<p>Vasodilation, neuroendocrine functions, neurotransmission</p>	 <p>Mol 4</p>

<p>Vasoactive intestinal polypeptide receptor 2 (VIPR2)</p>	<p>Vasodilation, neuroendocrine functions, neurotransmission</p>	 <p>Mol 6</p>
---	--	--

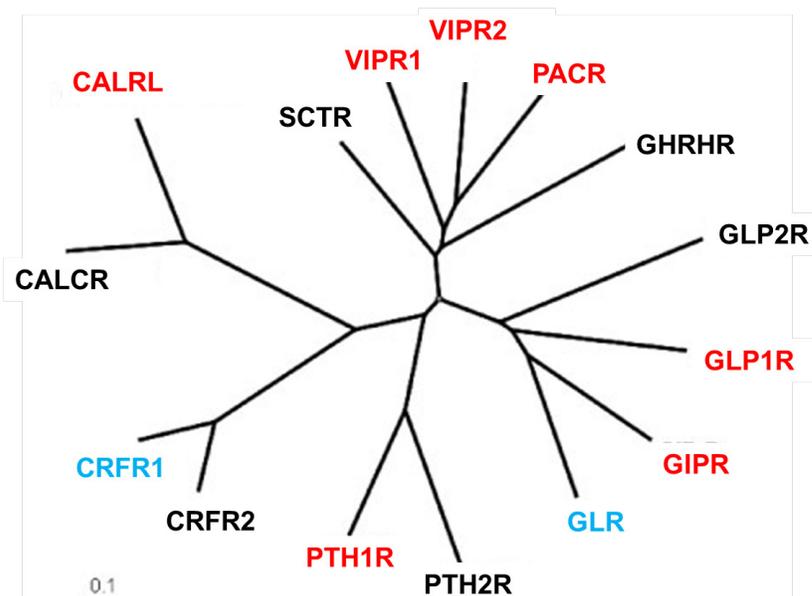


Figure 1. Phylogenetic tree for the class B1 receptors.¹¹ Proteins that we predicted are colored, red blue indicating those with x-ray crystal structures of the transmembrane region and red indicating the other seven GPCRs.

Methods

Structure Prediction

A more detailed explanation of each of these steps is provided in Chapter II.

1. Homology helix prediction – We predict helix lengths, locations, and orientations for each of the nine class B1 GPCRs using the two class B1 crystal structures as templates, resulting in eighteen starting structures.

2. Bihelix - Helices are rotated in 30° increments from 0-360° (η) and the energies are estimated for all 12^7 (35 million) structures in terms of energies for each of the twelve pairs.¹²
3. Combihelix- The top 2,000 seven-helical bundles from Bihelix are assembled and their energies calculated after optimizing side chains. The best template and helix rotations (η) are determined from the average charged and neutral interhelical energies, and are submitted to Superbihelix. Total energies were not used for the comparison due to the different TM lengths of the proteins.
4. Superbihelix – We estimate the energies for all $(3*5*5)^7$ (13 trillion) combinations of angles: θ (-10, 0, 10), ϕ (-30, -15, 0, 15, 30), and η (-30, -15, 0, 15, 30) of the best Combihelix structures, by combining Bihelix energies into sets of four.¹³
5. Supercombihelix – We assemble, optimize side chains, and score the top 2,000 Superbihelix bundles. The top structure is selected for docking.

Binding Site Prediction

6. DarwinDock/GenDock - The ligands from **Table 1** are docked to the best structures from Supercombihelix. Approximately ten diverse ligand conformations are generated using Maestro's MacroModel¹⁴ conformation search. Around 50,000 poses are generated for each using UCSF's Dock6¹⁵. The poses are clustered by CRMS, and the family head energies evaluated; then all of the children energies are evaluated for the top 10%. The best 120 poses are selected by using polar, total, and hydrophobic energies. These steps are performed with hydrophobic residues mutated to alanine.
7. Scoring - The residues are dealanized, SCREAMed¹⁶, neutralized¹³, and minimized in the Dreiding¹⁷ forcefield to obtain 120 poses with optimized side chains. The final best poses are selected by snap binding energy.

Results and Discussion

Structure Prediction

We predicted the structures for nine class B1 GPCRs. The sequence conservation for the transmembrane helices is shown in **Figure 2**, along with the most conserved residue of each transmembrane region, marked as X.50. This figure illustrates the relative conservation of the TMs: TM2 and TM4 are the most conserved, with the prolines throughout the receptor being conserved as well.

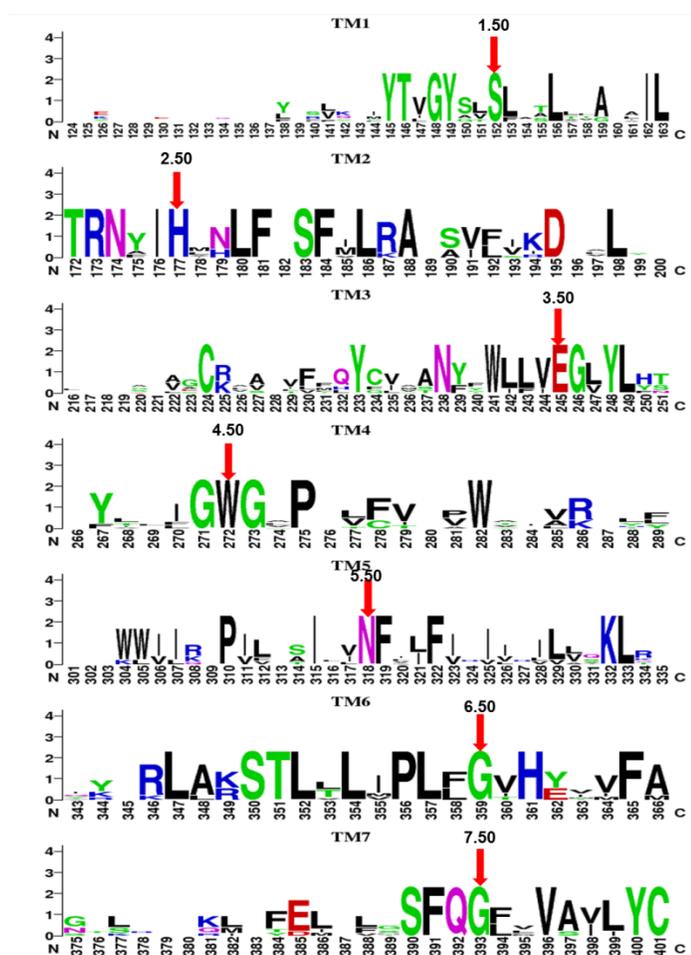


Figure 2. WebLogo representation of the sequence identities between the transmembrane regions of the nine class B1 predicted structures.¹⁸ Larger residue symbols indicate higher sequence identity. Polar residues are green, basic are blue, acidic are red, and hydrophobic are black. The red arrows indicate the X.50 residues as described in Siu FY, et al. 2013.²

Bihelix/Combihelix

Bihelix and Combihelix were run for the nine protein structures sampling the eta angles from 0-360° in 30° increments. The top structures from each case are shown in **Tables A1-A9 of the Appendix**. Seven of the nine structures prefer the glucagon receptor template. The remaining two, CALRL and CRFR1, preferred the CRFR1 template. This is not surprising when looking at the relationship between the class B1 GPCRs as shown in **Figure 1** – CALRL is much more closely related to CRFR1, while the other receptors are more closely related to GLR.

The structures in **Table 2** were chosen as starting points for the Superbihelix fine sampling of θ , ϕ , and η (θ : -10, 0, 10 / ϕ : -15, 0, 15 / η : -30, 0, 30). The structures were selected based on these criteria: the number 1 structure was always included, as was the all-zero rotation. The all-zero rotation was always included based on previous experiences showing it to often be a strong starting point for Superbihelix.¹⁹⁻²⁷

Table 2. Summary of structures input to Superbihelix from Bihelix/Combihelix

Protein	Template	Rank	Eta (η)						
			TM1	TM2	TM3	TM4	TM5	TM6	TM7
CALRL	CRFR1	1	0	0	0	0	0	0	0
CRFR1	CRFR1	1	0	0	0	0	0	0	0
GLR	GLR	1	0	0	0	0	0	-30	0
		7	0	0	0	0	0	0	0
GIPR	GLR	1	0	0	0	0	0	0	0
GLP1R	GLR	1	0	0	0	0	0	-30	0
		9	0	0	0	0	0	0	0
PACR	GLR	1	0	0	0	0	0	30	0
		2	0	0	0	0	0	0	0
PTH1R	GLR	1	0	0	0	150	30	-120	0
		3	0	0	0	0	0	0	0
VIPR1	GLR	1	0	0	0	0	0	0	0
VIPR2	GLR	1	0	0	0	0	0	0	0

Superbihelix/Supercombihelix

The Superbihelix/Supercombihelix top structures are shown in **Tables A10-A18 of the Appendix**. The rank 1 structure in each case was used for docking. **Table 3** summarizes these results. The helices with the largest variations throughout our nine receptors are: TM4 (φ), TM5 (η), and TM7 (η). The Superbihelix/Supercombihelix structures for CRFR1 and GLR are marked by red text. The CRFR1 structure has 0.9Å RMSD to the CRFR1 x-ray crystal TM region and the GLR structure has 2.1Å RMSD to the x-ray crystal TM region. It is interesting to note that throughout our structure prediction procedure, the GLR structure has consistently varied more from that of the crystal structure than CRFR1 did. It is important to remember that GPCRs exist in several low-lying states. It is entirely possible that the conditions necessary to obtain crystal structures of GLR and CRFR1 favored one of these states, while we find another.

Table 3. Superbihelix/Supercombihelix results for the nine predicted structures.

Protein	Theta (θ)							Phi (φ)							Eta (η)						
	H1	H2	H3	H4	H5	H6	H7	H1	H2	H3	H4	H5	H6	H7	H1	H2	H3	H4	H5	H6	H7
CALCRL	-10	0	0	0	0	0	0	15	0	0	-30	15	15	15	15	0	0	0	15	-30	-15
CRFR1	0	0	0	0	0	0	-10	0	0	0	-15	0	15	0	0	0	0	0	0	0	15
GCGR	0	-10	-10	0	-10	10	10	-30	0	0	-30	15	15	0	0	0	0	15	-30	-15	0
GIPR	0	-10	-10	0	-10	0	10	-15	0	0	-30	0	15	-30	0	0	0	15	-15	0	15
GLP1R	0	-10	-10	0	-10	0	10	0	-15	0	-15	-15	30	-15	0	0	-15	0	-15	0	15
PACAPR	0	0	0	0	0	0	10	-15	0	0	-30	0	0	-15	0	0	0	0	15	0	15
PTH1R	0	-10	-10	0	-10	0	10	0	-15	0	-30	-15	15	-30	15	0	-15	0	-15	-30	30
VIPR1	-10	0	0	0	0	10	10	30	0	0	-15	0	0	-30	-15	0	0	0	15	-15	15
VIPR2	0	-10	-10	-10	-10	0	10	0	-15	0	30	15	30	-15	15	0	0	-15	15	-15	15

Angles are with respect to the starting template – CRFR1 for CALCRL and CRFR1, and GLR for the rest.

Interhelical Interactions

The interhelical interactions of each of the nine Superbihelix/Supercombihelix structures are shown below in **Table 5**. A summary of the conserved interactions is shown in **Table 4** and depicted for GLR in **Figure 3**. All nine structures show an

interaction between TM2-TM3 (R2.46-E3.50) as well as an interaction between TM6 and TM7: R/K6.37-Y/F7.57. Six of the structures show additional interactions between TMs 2 and 3: R2.46-Y3.53, and/or H2.50-E3.50, and five have an additional TM2-TM3 interaction, which is weaker: N3.43-A/S2.56. Additionally, TMs 2 and 7 also interact in five of the structures via K/R2.60-Q7.49 and four of the structures via K/R2.60-G/S/E/N7.46. It is important to note that in every case where one of the conserved interactions is not present, the residues are instead forming other interactions, thereby stabilizing that conformation. From these conserved interactions, one can see that in class B1 GPCRs, TMs 2 and 3 are very closely coupled, with TMs 6-7 and TMs 2-7 also interacting tightly. We do not see the TMs 1-2-7 and TMs 2-3-4 interactions which are consistently present in the class A GPCRs because of the low sequence homology between the GPCR families. The residues involved in those class A GPCRs are not conserved in class B1.

Table 4. Class B1 conserved interactions

Protein	TM2-TM3:				TM6-7:	TM2-7:	
	R2.46-E3.50	R2.46-Y3.53	H2.50-E3.50	N3.43-A/S2.56	R/K6.37-Y/F7.57	K/R2.60-Q7.49	K/R2.60-G/S/E/N7.46
CALRL	Y	N	Y	N	Y	N	N
CRFR1	Y	N	Y	Y	Y	Y	Y
GLR	Y	Y	Y	Y	Y	Y	Y
GIPR	Y	Y	Y	Y	Y	N	Y
GLP1R	Y	Y	N	A (R2.60-N3.43)	Y	N	N
PACR	Y	Y	Y	N	A (R6.37-L7.56)	N	N
PTH1R	Y	Y	N	N	Y	Y	Y
VIPR1	Y	Y	Y	Y	Y	Y	N
VIPR2	Y	N	N	N	Y	Y	N
Total	9	6	6	5	9	5	4

A=alternate

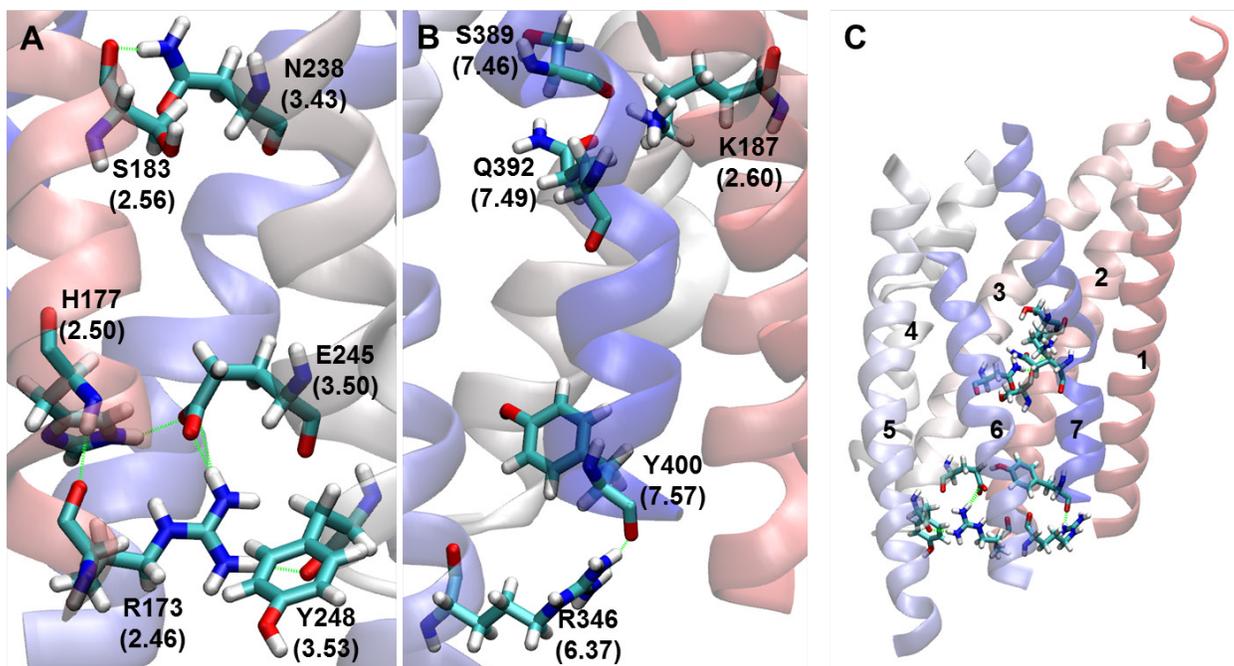


Figure 3. Class B1 conserved interactions, as found in the GLR predicted structure.

(A) TM2-TM3 interactions: R2.46-E3.50, R2.46-Y3.53, H2.50-E3.50, and N3.43-A/S2.56. (B) TM6-TM7 and TM2-TM7 interactions: R/K6.37-Y/F7.57, K/R2.60-Q7.49, and K/R2.60-G/S/E/NN7.46 (C) Overview of the structure showing the locations of the interactions.

Table 5. Class B1 predicted structure interhelical interactions

KEY

Identical in all class B1 GPCRs

Conserved mutations occur in class B1 GPCRs

Semi-conserved mutations occur in class B1 GPCRs

Not conserved across the class B1 GPCRs

Conserved interaction

Crystal interaction

CALRL					
1--2--7		2--3--4		Other	
N140_1	Y367_7	R173_2	E233_3	Y227_3	E348_6
T145_1	H374_7	H177_2	E233_3	K319_5	E327_6
S159_1	S381_7	N179_2	Y256_4	K333_6	F386_7
		K213_3	A271_4	Q376_7	E348_6
		K213_3	S275_4		
		Y221_3	W260_4		
		R274_4	Q216_3		

CRFR1						CRFR1 Crystal					
1--2--7		2--3--4		Other		1--2--7		2--3--4		Other	
W156_2	F138_1	R151_2	E209_3	N196_3	Y272_5	S130_1	F357_7	H155_2	E209_3	W246_4	D269_5
R165_2	E352_7	H155_2	E209_3	H199_3	M276_5	S130_1	S353_7	N157_2	W236_4	K250_4	D269_5
Q355_7	R165_2	R189_3	A247_4	W246_4	D269_5	R151_2	N367_7	Y197_3	W236_4	Q355_7	Y327_6
		Y197_3	W236_4	K250_4	D269_5	N152_2	N367_7				
		N202_3	A161_2	Q308_6	L366_7						
		K228_4	N157_2	Y309_6	A216_3						
				K311_6	F365_7						
				K311_6	L366_7						
				K311_6	N367_7						
GLR						GLR Crystal					
1--2--7		2--3--4		Other		1--2--7		2--3--4		Other	
Q142_1	D195_2	R173_2	Y248_3	Y239_3	E362_6	S152_1	G393_7	R225_3	C287_4	L242_2	N318_5
Y145_1	F383_7	R173_2	E245_3	R308_5	E362_6	T146_1	D195_2			Y239_3	L358_6
Y149_1	S389_7	H177_2	E245_3	R308_5	A366_6	K187_2	S389_7			H361_6	Q392_7
K187_2	Y149_1	R225_3	C287_4	N318_5	L242_3					Y400_7	T351_6
K187_2	S389_7	N238_3	S183_2	K344_6	L329_5						
K187_2	Q392_7			K344_6	K332_5						
R199_2	Q142_1			R346_6	Y400_7						

GIPR					
1--2--7		2--3--4		Other	
Y145_1	S381_7	R169_2	Y240_3	R183_2	E354_6
N170_2	L159_1	R169_2	E237_3	Y231_3	E354_6
R183_2	S381_7	H173_2	E237_3	R300_5	E377_7
R192_2	T142_1	S179_2	N230_3	R336_6	L241_3
R192_2	Q138_1	R190_2	Q220_3	R338_6	Y392_7
		R217_3	Y279_4	T343_6	Y392_7
		Y279_4	Q224_3	K373_7	F357_6
				K373_7	A358_6
				S381_7	E354_6
GLP1R					
1--2--7		2--3--4		Other	
Y152_1	T391_7	R176_2	E247_3	Y241_3	I357_6
N177_2	L166_1	R176_2	Y250_3	W306_5	R380_7
		R190_2	N240_3	R310_5	E364_6
		R190_2	V237_3	K346_6	L251_3
		R227_3	Y289_4	R348_6	Y402_7

		Y235_3	V282_4	R348_6	C403_7
		Y269_4	T253_3	T353_6	Y402_7
				R380_7	Y305_5
				R380_7	M303_5
				K383_7	F367_6
				K383_7	A368_6
PACR					
1--2--7		2--3--4		Other	
K154_1	D207_2	R185_2	E247_3	Y241_3	F362_6
T158_1	D207_2	R185_2	Y250_3	N320_5	L244_3
H189_2	Y400_7	H189_2	E247_3	Y348_6	L251_3
R199_2	E385_7	K227_3	L289_4	R350_6	L399_7
R379_7	Y150_1			T355_6	E247_3
				Y366_6	E385_7
				R381_7	V368_6
				Y400_7	E247_3
PTH1R					
1--2--7		2--3--4		Other	
Y191_1	H442_7	R219_2	E302_3	Y297_3	P366_5
Y195_1	D241_2	R219_2	Y305_3	Y297_3	S370_5
N220_2	L209_1	R282_3	A344_4	K359_5	Q440_7
R233_2	N448_7			K359_5	E444_7
R233_2	Q451_7			K360_5	F424_6
Y245_2	D185_1			K360_5	Q440_7
Y459_7	H223_2			Q364_5	F424_6
				K405_6	Y459_7
				K408_6	I458_7
				Y421_6	N448_7
				Q440_7	F424_6
VIPR1					
1--2--7		2--3--4		Other	
K143_1	D196_2	R174_2	E236_3	N308_5	L233_3
Y146_1	D196_2	R174_2	Y239_3	R317_5	Y241_3
T147_1	D196_2	H178_2	E236_3	Y336_6	L240_3
Y150_1	G377_7	S184_2	N229_3	R338_6	Y388_7
R188_2	Y150_1	F185_2	N229_3	Q380_7	Y354_6
R188_2	Q380_7	R188_2	C225_3		
		R188_2	V226_3		

		K216_3	I278_4		
		Y224_3	M271_4		
		S267_4	Y224_3		
VIPR2					
1--2--7		2--3--4		Other	
K127_1	D180_2	R158_2	E223_3	R172_2	Y341_6
K127_1	D181_2	W249_4	N164_2	N295_5	F217_3
Y134_1	D180_2	T253_4	Y211_3	K324_6	K309_5
N159_2	L148_1			R325_6	Y375_7
R172_2	Q367_7			Y341_6	G364_7
Y375_7	H162_2			Q356_7	F344_6

The CRFR1 predicted structures can be compared to the x-ray crystal structure of CRFR1. These interactions occur in both structures: H2.50-E3.50, W4.50-Y3.38 (present in CALRL as well), W4.60-D5.36, and K4.64-D5.36. The W4.60-D5.36 interaction may be present in proteins closely related to CRFR1. There are, however, six interactions of the crystal structure which we do not see between helices 1-7, 2-7, 2-4, and 6-7. We can also compare the predicted GLR structure to that of the GLR crystal structure. Both structures have the K/R2.60-G/S/E/N7.46 interaction, as well as R3.30-C4.65. There are six “missing” hydrogen bonds from the crystal structure. In both cases, the reasons for the absent crystal structure interactions are multifold: 1. Both crystal structures were modified to allow for crystallization, which may possibly affect their TM bundle conformations and interactions. 2. Our structures do not have loops or waters present, which would primarily affect the more terminal residue interactions. 3. GPCRs have several low energy conformations²⁸, and it is possible that we are seeing a low-lying structure which differs from the one crystallization conditions favored.

Binding Site Prediction

We predict the small molecule ligand (from **Table 1**) binding sites to the nine class B1 GPCRs. The binding sites are consistently placed between TMs 3 and 7. Often TMs 2, 4, and 6 are involved as well, and much more rarely 1 and 5. This localization of

the small molecule ligands is similar to that which is consistently seen in the more well-characterized class A GPCRs, but is shifted more towards the center of the TM bundle (Figure 4). In class A GPCRs, TMs 5 and 6 play a much larger role in the small molecule interactions than that which we see here.²⁹⁻³⁰

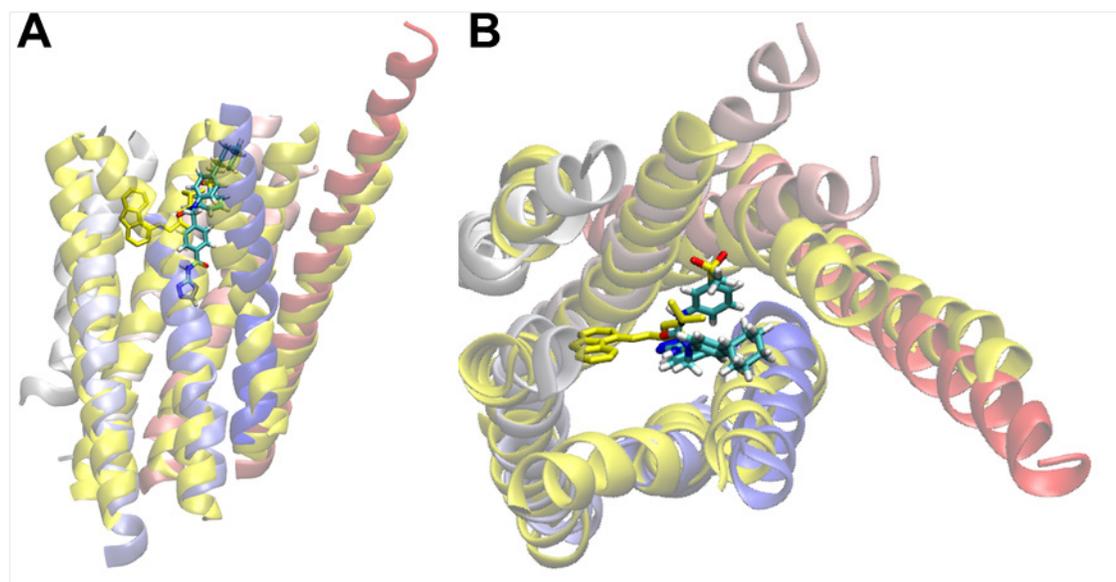


Figure 4. Comparison of the class A and class B1 GPCR binding sites. (A) Side view and (B) top view of the β 2-adrenergic receptor³⁰ (yellow) and our predicted GLR structure. Note that the ligand is more centrally located in the class B1 GPCRs, while it is closer to TM5 in the class A GPCRs.

The predicted binding sites for the nine proteins are pictured in **Figures 5-13**. For each structure we present an overview of its location, in which the TM regions are colored red->white->blue as they go from 1->7. We also present the ligand's pharmacophore, in which residues are colored: red - acidic, green - hydrophobic, purple - basic, blue - polar, and white - glycine. Finally, we present a table of the strongest interacting residues (note that these energies are from the neutralized complexes, and are therefore lower for charged residues than they would be otherwise). Because very little is known about these ligands' binding sites, we are unable to fruitfully compare to experimental data. Instead, we use our binding site information to suggest protein mutation studies which could test our binding site. Additionally, we suggest the characteristics of ideal ligands for the binding location.

For CALRL bound to MK-0974 (**Figure 5**), we find that the ligand interacts with TMs 2, 3, 5, 6, and 7. To test if our binding site is accurate, we suggest performing experiments to probe the importance of Glu348; for example, testing the effect of a Glu348Ala mutation. We note that a positively charged ligand would be ideal for interacting with Glu348, possibly with aromatic components to interact with His295 and Phe349.

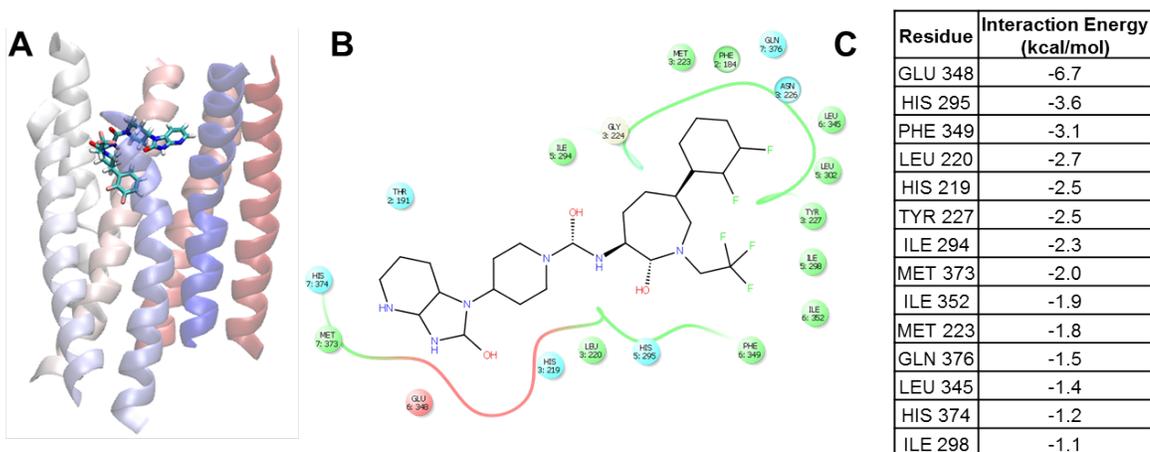


Figure 5. CALRL bound to MK-0974. (A) Predicted structure binding site overview. (B) Predicted structure pharmacophore. (C) Strongest interactions between the protein and ligand (kcal/mol). MK-0974 adopts a folded conformation which interacts with TMs 2, 3, 5, 6, and 7.

For CRFR1 we can compare our structure to that of the crystallized CRFR1- CP-376395. Our binding site is similar, but slightly shifted towards TM7 (**Figure 6**). Both binding sites are placed low in the protein – an unusual binding site which has not been seen in any of the other crystallized GPCRs, and may be unique to this receptor. To ensure that our binding site is not an artifact of our procedure, we also docked the crystal ligand to the crystal protein (**C and D of Figure 6**). This results in a binding site nearly identical to that of the crystal structure. Therefore, our CRFR1 predicted structure's binding site is a reflection of its structural variations and sequence differences from that of the crystal. It is also important to note that our structure's binding site was also slightly more favorable energetically than the crystal binding pose: the crystal snap binding energy was -50.4 kcal/mol while our structure was -51.2 kcal/mol. To test our binding

site, and also its validity in comparison to the x-ray crystal structure's, we suggest two mutations: His155Ala and Asn283Ala. If the first mutation has a large effect on CP-376395 binding, it may reflect the accuracy of our prediction, if the second does then it would indicate that the crystal ligand binding pose is more accurate.

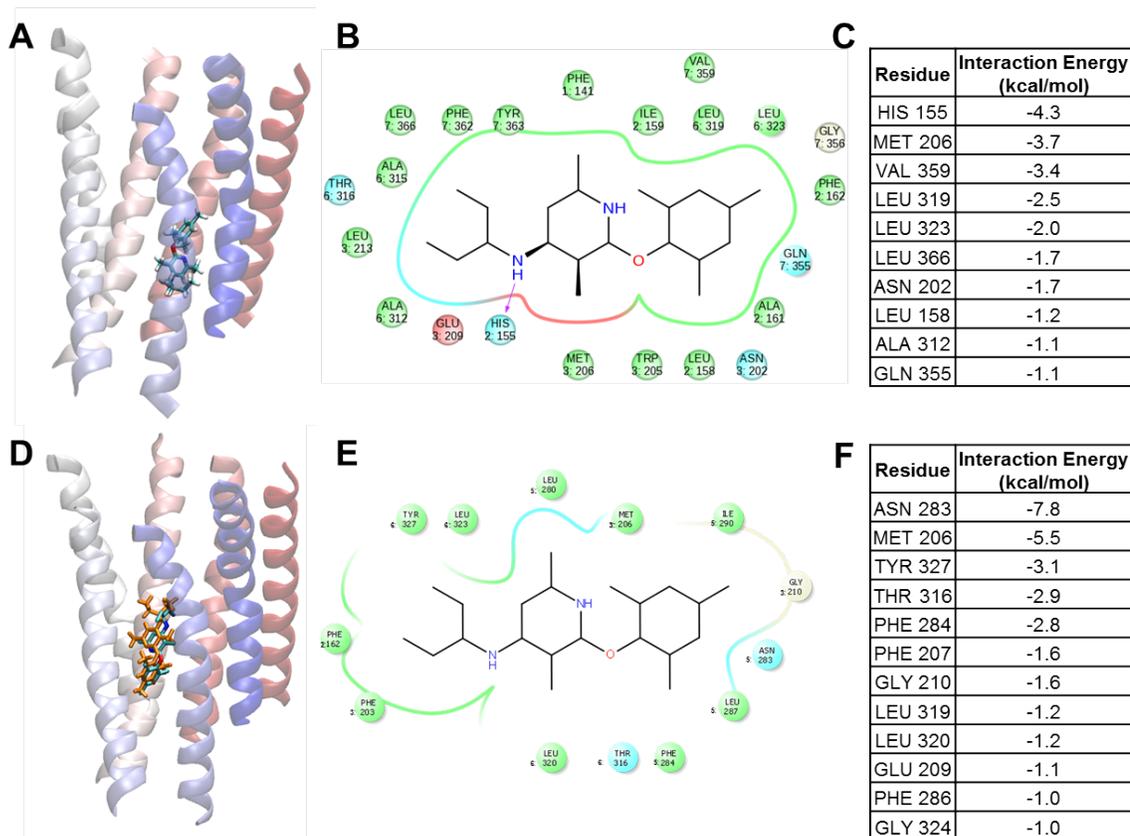


Figure 6. CRFR1 bound to CP-376395. (A) Predicted structure binding site overview. (B) Predicted structure pharmacophore. (C) Strongest interactions between the predicted CRFR1 and CP-376395 (kcal/mol). (D) Crystal structure binding site, with our docking results of the crystal ligand and protein shown in orange. (E) Crystal structure pharmacophore. The predicted CRFR1:CP-376395 binding site interacts with TMs 2, 3, 6, and 7, while the crystal binding site interacts with TMs 3, 5, and 6.

The GLR crystal structure did not have a resolved ligand, therefore no direct comparison can be made between the two structures. However, we docked NNC0640 to both our predicted GLR structure, and the crystal structure (**Figure 7**). The poses are nearly identical, despite the structural variations between the GLR predicted and crystal

structures. However, the predicted structure's binding pose was more energetically favorable than the predicted crystal structure binding pose: a snap binding energy of -89.5kcal/mol versus -80.4kcal/mol for the crystal predicted pose. Our predicted pose may be the more physical structure for binding to NNC0640, especially since the ligand was not stably in complex with the crystal protein. To evaluate our binding sites we would suggest two mutation studies: Arg225Ala and also Phe365Ala. Depending on the results, one or the other binding site may be found to be more physical. To optimally interact with this binding site, a ligand with two negative charges could be designed which targets both Lys187 and Arg378 (or Arg 308).

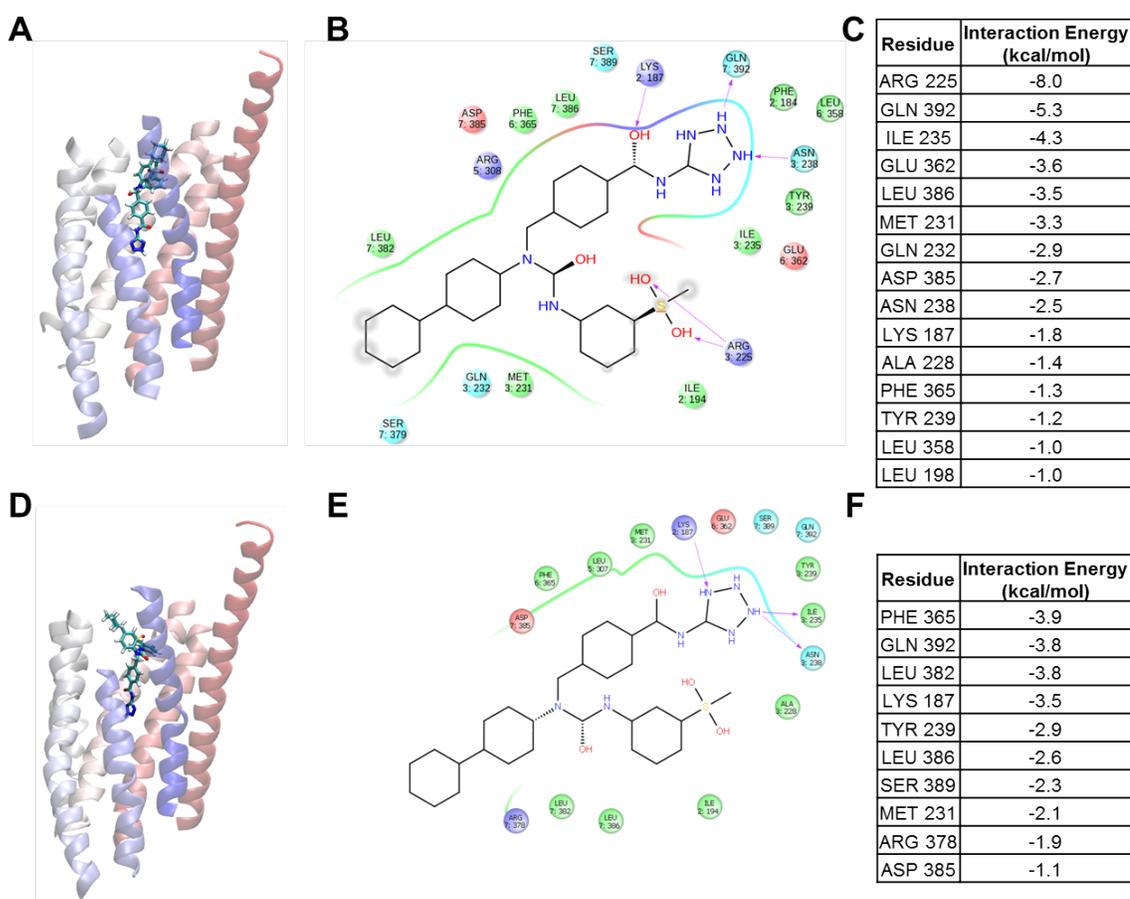


Figure 7. GLR bound to NNC0640. (A) Predicted structure binding site overview. (B) Predicted structure pharmacophore: the ligand interacts with TMs 2, 3, 5, 6, and 7. (C) Strongest interactions between the predicted protein and ligand (kcal/mol). (D) Predicted crystal structure binding site (the ligand was not resolved in the crystal structure). (E) Predicted crystal structure pharmacophore: the ligand interacts with TMs 2, 3, 5, 6, and 7. (F) Strongest interactions between the predicted crystal protein and ligand (kcal/mol).

The GIPR: Molecule 29 binding pose is presented in **Figure 8**. Molecule 29 interacts with TMs 1, 2, 3, 5, and 7. To test the validity of our binding site we suggest three mutations: Arg190Ala, Arg370Ala, and Arg278Ala. Additionally, the large number of positively charged residues in the binding site indicates that a drug designed with multiple negatively charged groups (or at least strongly polar if many negative charges proves to be energetically unfavorable) that is also large enough to hit all four arginines would be ideal. Alternatively, a ligand with one or two negative charges could hit a subset of the four arginines and be less prone to poor columbic energies.

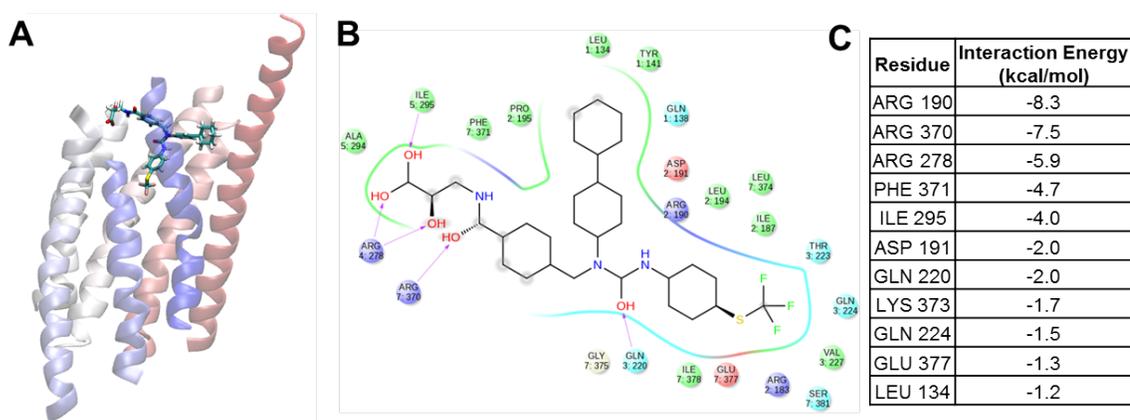


Figure 8. GIPR bound to Molecule 29. (A) Predicted structure binding site overview. (B) Predicted structure pharmacophore. (C) Strongest interactions between the protein and ligand (kcal/mol). The rather large Molecule 29 bridges most of the protein and interacts with TMs 1, 2, 3, 5 and 7.

GLP1R bound to T0632 is shown in **Figure 9**. The ligand interacts with TMs 2, 3, 5, and 7. To probe this binding site, we suggest three protein mutations: Tyr305Ala, Lys197Ala, and Glu387Ala. Additionally, the combination of Lys197 and Glu387 in the binding pocket suggests that a zwitterionic ligand may be ideal for interacting with GLP1R.

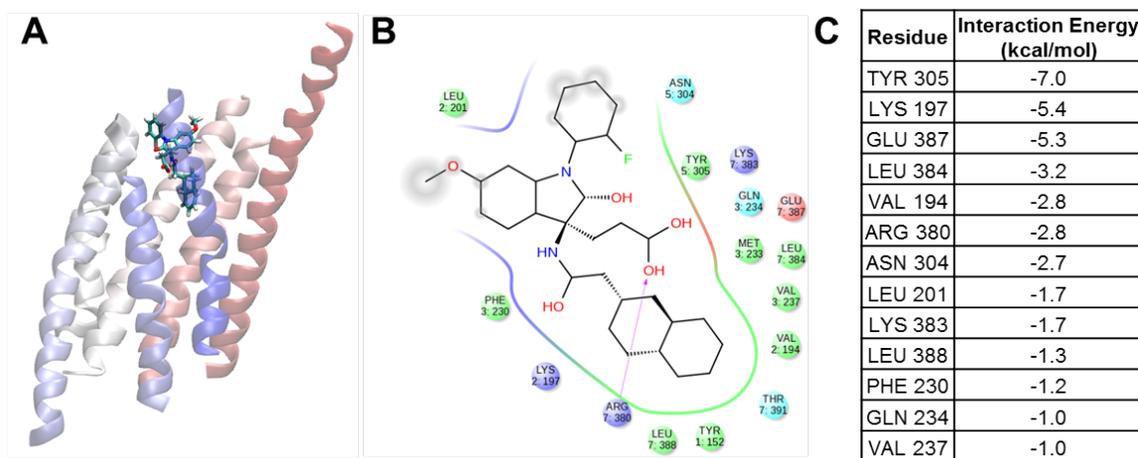


Figure 9. GLP1R bound to T0632. (A) Predicted structure binding site overview. (B) Predicted structure pharmacophore. (C) Strongest interactions between the protein and ligand (kcal/mol). T0632 orients itself vertically between TMs 2, 3, 5 and 7.

PACR bound to Molecule 1 is presented in **Figure 10**. Molecule 1 interacts with TMs 1, 2, 3, 6, and 7. We suggest several mutation studies to test our binding site: His234Ala, Tyr241Ala, Glu385Ala, and Tyr366Ala. Additionally, we note that it may be ideal to generate a ligand with a positive charge to interact with Glu383, and/or multiple aromatic and polar groups to interact with the binding pocket's tyrosines.

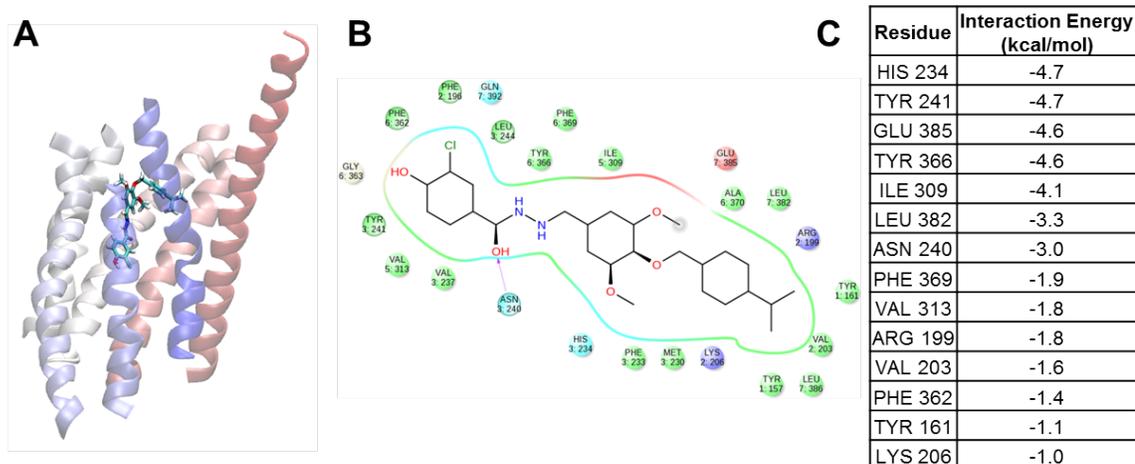


Figure 10. PACR bound to Molecule 1. (A) Predicted structure binding site overview. (B) Predicted structure pharmacophore. (C) Strongest interactions between the protein and ligand (kcal/mol). Molecule 1 is quite large, but it takes a folded conformation and interacts with TMs 1, 2, 3, 6, and 7.

SW106's binding site with PTH1R is depicted in **Figure 11**. The ligand interacts with residues on TMs 2, 3, 5, 6, and 7. To test the validity of our binding site we suggest two mutation experiments: Arg233Ala and Gln451Ala. Additionally we note that a negatively charged ligand would be ideal for interacting with this binding pocket, possibly also with polar atoms to hit Gln451 and an aromatic group to interact with Phe447.

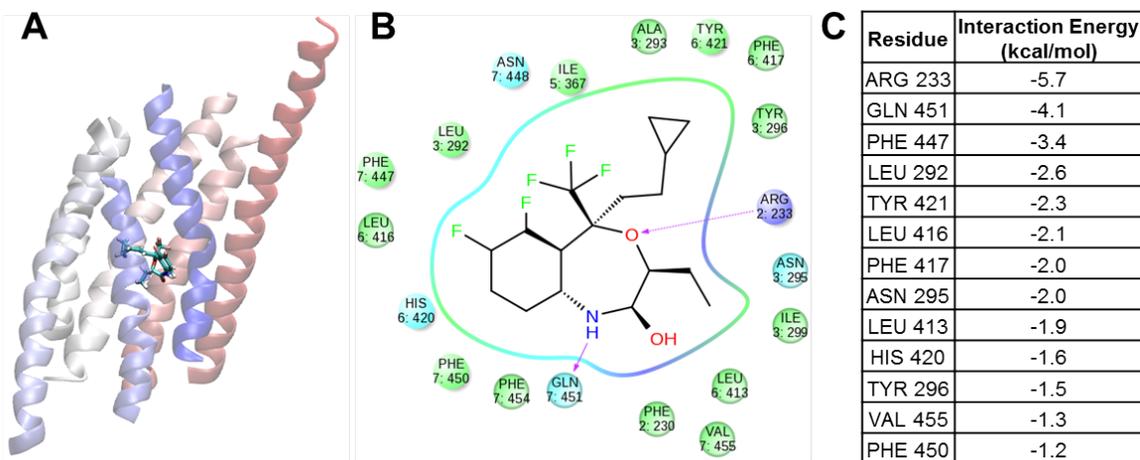


Figure 11. PTH1R bound to SW106. (A) Predicted structure binding site overview. (B) Predicted structure pharmacophore. (C) Strongest interactions between the protein and ligand (kcal/mol). SW106 is small overall, and takes a binding pose relatively low in the protein. It interacts with TMs 2, 3, 5, 6, and 7.

VIPR1 bound to Molecule 4 is shown in **Figure 12**. The ligand interfaces with TMs 2, 3, 4, 5, and 7. To probe our binding site we suggest two experiments: Gln223Ala and Lys195Ala. To optimally interact with this binding pocket, a zwitterionic ligand could be explored with the intent of binding to both Lys195 and Glu373.

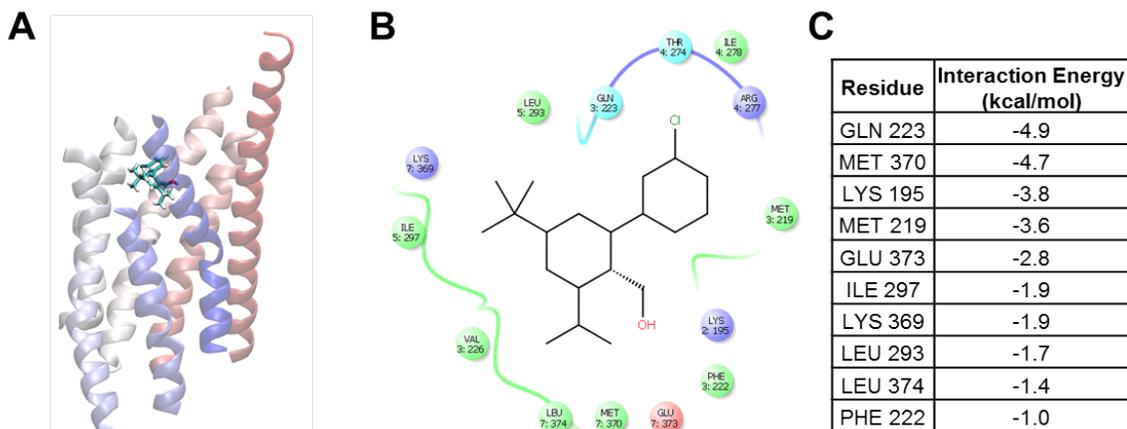


Figure 12. VIPR1 bound to Molecule 4. (A) Predicted structure binding site overview. (B) Predicted structure pharmacophore. (C) Strongest interactions between the protein and ligand (kcal/mol). The small molecule 4 interacts with TMs 2, 3, 4, 5 and 7.

The binding site of VIPR2 and Molecule 6 is pictured in **Figure 13**. Residues on TMs 2, 3, 5, 6, and 7 interact with Molecule 6. To test the validity of our binding site we suggest two mutations: Gln210Ala and Glu360Ala. Additionally we note that an ideal ligand for this binding site would be able to form charged interactions with both Arg172 and Glu360.

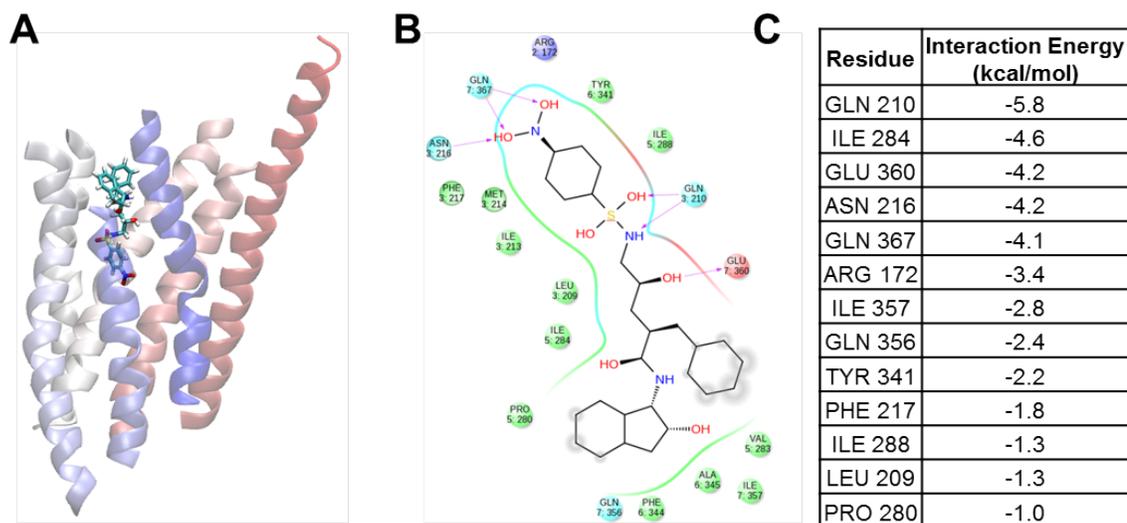


Figure 13. VIPR2 bound to Molecule 6. (A) Predicted structure binding site overview. (B) Predicted structure pharmacophore. (C) Strongest interactions between the protein and ligand (kcal/mol). The quite large molecule 6 inserts itself vertically between TMs 2, 3, 5, 6, and 7.

Conclusion

In this work, we predict the structures and small molecule binding sites of nine class B1 GPCRs: calcitonin gene-related peptide type 1 receptor (CALRL), corticotropin-releasing factor receptor 1 (CRFR1), glucagon receptor (GLR), gastric inhibitory polypeptide receptor (GIPR), glucagon-like peptide-1 receptor (GLP1R), pituitary adenylate cyclase-activating polypeptide type I receptor (PACR), parathyroid hormone/parathyroid hormone-related peptide receptor (PTH1R), vasoactive intestinal polypeptide receptor 1 (VIPR1), and vasoactive intestinal polypeptide receptor 2 (VIPR2). The nine receptors all share two conserved hydrogen bond networks: R2.46-E3.50 and R/K6.37-Y/F7.57. Additionally, most of the receptors have these interactions: R2.46-Y3.53 (six), H2.50-E3.50 (six), N3.43-A/S2.56 (five), K/R2.60-Q7.49, and K/R2.60-G/S/E/N7.46. Our small molecule binding sites are consistent across the class B1 family, with primary interactions occurring with TMs 3 and 7. We suggest mutation studies to test the validity of our binding poses, as well as suggest characteristics of optimal ligands for targeting these binding sites. The information obtained from this study is well-suited for utilization in the design of drugs targeting class B1 GPCRs, or for further optimization of the ligands docked here.

Appendix

Bihelix/Combihelix Results

The structure highlighted in yellow was used for Superbihelix.

Table A1. Top 10 structures from Bihelix/Combihelix for CALRL (CRFR1 Template)

Eta (η)							Charged InterHelical (kcal/mol)	Neutral InterHelical (kcal/mol)
TM1	TM2	TM3	TM4	TM5	TM6	TM7		
0	0	0	0	0	0	0	-499.8	-361.0
0	0	0	0	120	-60	0	-495.0	-356.3
0	-90	0	-30	90	-60	0	-491.0	-353.3
0	0	0	0	180	0	0	-473.0	-357.7
0	0	0	0	90	-60	0	-505.9	-349.8
0	0	0	0	0	-90	-30	-485.2	-350.6
0	0	0	0	0	-90	0	-463.0	-362.5
0	-90	0	120	120	-30	0	-493.8	-346.4
0	-90	0	-30	90	-30	0	-469.6	-347.3
0	0	0	0	0	-30	-30	-459.2	-350.1

Table A2. Top 10 structures from Bihelix/Combihelix for CRFR1 (CRFR1 Template)

Eta (η)							Charged InterHelical (kcal/mol)	Neutral InterHelical (kcal/mol)
TM1	TM2	TM3	TM4	TM5	TM6	TM7		
0	0	0	0	0	0	0	-481.1	-440.1
0	-90	0	0	0	30	0	-459.1	-420.8
0	-90	0	0	0	180	0	-432.3	-416.9
0	0	0	0	-60	0	0	-416.6	-401.5
0	0	0	0	0	30	0	-411.4	-419.0
30	0	0	0	0	0	0	-451.0	-389.6
0	-90	0	0	0	0	0	-417.3	-384.8
0	0	0	0	-30	0	0	-413.5	-389.7
90	0	0	0	0	0	0	-436.8	-391.3
-30	0	0	0	0	0	0	-440.0	-402.6

Table A3. Top 10 structures from Bihelix/Combihelix for GLR (GLR Template)

Eta (η)							Charged InterHelical (kcal/mol)	Neutral InterHelical (kcal/mol)
TM1	TM2	TM3	TM4	TM5	TM6	TM7		
0	0	0	0	0	-30	0	-503.2	-400.0
0	0	0	90	0	-30	0	-494.4	-392.3
0	0	0	0	0	-120	0	-478.7	-392.4
0	0	0	90	0	-120	0	-471.1	-383.1
0	0	0	0	0	30	0	-458.5	-388.3
0	0	0	120	30	-30	0	-475.6	-379.6
0	0	0	0	0	0	0	-463.0	-379.1
0	0	0	90	0	30	0	-458.2	-379.5
0	0	0	180	0	-30	0	-462.0	-364.3
0	0	0	90	0	-90	0	-449.8	-382.4

Table A4. Top 10 structures from Bihelix/Combihelix for GIPR (GLR Template)

Eta (η)							Charged InterHelical (kcal/mol)	Neutral InterHelical (kcal/mol)
TM1	TM2	TM3	TM4	TM5	TM6	TM7		
0	0	0	0	0	0	0	-315.6	-400.5
0	0	0	0	0	-60	0	-311.7	-413.1
0	0	0	0	0	-30	0	-311.2	-391.2
0	0	0	0	0	-150	0	-296.3	-398.8
0	0	0	120	30	0	0	-307.8	-388.2
0	0	0	0	30	0	0	-294.1	-395.1
0	0	0	0	0	30	0	-304.5	-382.9
0	0	0	0	0	-90	0	-291.0	-373.1
0	0	0	0	30	-120	0	-277.4	-383.3
0	0	0	0	0	60	0	-280.4	-371.7

Table A5. Top 10 structures from Bihelix/Combihelix for GLP1R (GLR Template)

Eta (η)							Charged InterHelical (kcal/mol)	Neutral InterHelical (kcal/mol)
TM1	TM2	TM3	TM4	TM5	TM6	TM7		
0	0	0	0	0	-30	0	-463.5	-417.0
0	0	0	0	0	-120	0	-428.4	-395.8
0	0	0	30	30	-90	0	-428.5	-385.0
0	0	0	30	0	-30	0	-441.1	-376.6
0	0	0	30	30	-30	0	-439.7	-367.8
0	0	0	0	30	0	0	-424.3	-374.0
0	0	0	180	0	-30	0	-421.4	-375.2
30	0	0	0	0	-30	0	-420.1	-366.3
0	0	0	0	0	0	0	-406.0	-376.6
0	0	0	0	0	0	-60	-410.3	-367.1

Table A6. Top 10 structures from Bihelix/Combihelix for PACR (GLR Template)

Eta (η)							Charged InterHelical (kcal/mol)	Neutral InterHelical (kcal/mol)
TM1	TM2	TM3	TM4	TM5	TM6	TM7		
0	0	0	0	0	30	0	-580.5	-417.6
0	0	0	0	0	0	0	-548.4	-415.3
0	0	0	0	90	0	0	-548.3	-404.5
-30	0	0	0	0	0	0	-540.4	-395.4
0	0	0	0	30	0	0	-535.6	-416.6
0	0	0	0	30	-90	0	-528.9	-396.5
0	0	0	180	0	-30	0	-539.8	-376.8
0	0	0	0	0	-60	0	-520.5	-388.6
0	0	0	0	0	-90	0	-517.9	-390.6
0	0	0	180	0	0	0	-517.8	-376.6

Table A7. Top 10 structures from Bihelix/Combihelix for PTH1R (GLR Template)

Eta (η)							Charged InterHelical (kcal/mol)	Neutral InterHelical (kcal/mol)
TM1	TM2	TM3	TM4	TM5	TM6	TM7		
0	0	0	150	30	-120	0	-399.9	-372.6
-90	0	0	0	0	0	0	-394.7	-400.9
0	0	0	0	0	0	0	-394.1	-396.4
0	0	0	150	30	0	0	-393.3	-395.8
0	0	0	150	0	0	0	-388.7	-378.6
0	0	0	0	0	-90	0	-387.8	-378.6
0	0	0	0	90	0	0	-390.8	-350.3
0	0	0	0	0	-30	0	-384.4	-388.6
0	0	0	150	30	60	0	-384.0	-375.0
-30	0	0	180	0	0	30	-374.5	-351.4

Table A8. Top 10 structures from Bihelix/Combihelix for VIPR1 (GLR Template)

Eta (η)							Charged InterHelical (kcal/mol)	Neutral InterHelical (kcal/mol)
TM1	TM2	TM3	TM4	TM5	TM6	TM7		
0	0	0	0	0	0	0	-478.9	-408.4
0	0	0	0	90	0	0	-475.7	-398.0
0	0	0	180	0	-90	0	-448.4	-361.2
0	0	0	0	0	-90	0	-437.9	-369.8
0	0	0	120	0	-90	0	-437.7	-369.9
0	0	0	0	0	-30	0	-430.9	-378.0
0	0	0	0	30	-90	0	-423.0	-376.1
0	0	0	0	30	-30	0	-430.0	-353.6
0	0	0	180	0	30	0	-418.3	-376.3
0	0	0	0	30	120	0	-422.8	-354.4

Table A9. Top 10 structures from Bihelix/Combihelix for VIPR2 (GLR Template)

Eta (η)							Charged InterHelical (kcal/mol)	Neutral InterHelical (kcal/mol)
TM1	TM2	TM3	TM4	TM5	TM6	TM7		
0	0	0	0	0	0	0	-434.1	-390.8
0	0	0	180	30	0	0	-434.2	-371.7
0	0	0	0	30	-30	0	-423.5	-361.1
0	0	0	0	0	-30	0	-436.9	-349.1
30	90	0	0	30	-90	0	-441.3	-343.1
0	0	0	120	30	0	0	-422.6	-352.7
0	0	0	0	90	0	0	-416.7	-353.7
0	0	0	180	0	0	0	-416.9	-350.6
0	0	0	0	0	-120	0	-433.0	-341.7
0	0	0	30	0	0	0	-411.6	-358.5

Superbihelix/Supercombihelix Results

The structure highlighted in yellow was used for analysis and docking.

Table A10. Superbihelix/Supercombihelix results for CALRL

Theta (θ)							Phi (ϕ)							Eta (η)							Charged Inter-Helical (kcal/mol)	Charged Total (kcal/mol)	Neutral Inter-Helical (kcal/mol)	Neutral Total (kcal/mol)
TM1	TM2	TM3	TM4	TM5	TM6	TM7	TM1	TM2	TM3	TM4	TM5	TM6	TM7	TM1	TM2	TM3	TM4	TM5	TM6	TM7				
-10	0	0	0	0	0	0	15	0	0	-30	15	15	15	15	0	0	0	15	-30	-15	-592.6	278.2	-445.0	258.0
-10	-10	0	0	0	0	-10	0	0	0	-15	15	30	-15	15	0	0	0	15	-30	0	-556.8	263.8	-423.8	257.5
-10	-10	0	0	0	0	-10	0	0	0	-30	15	30	-15	15	0	0	0	15	-30	0	-548.2	252.0	-413.4	238.6
-10	-10	0	0	0	0	-10	0	0	0	-30	0	-15	-15	15	0	0	0	15	-15	0	-530.3	290.3	-424.5	249.3
-10	0	0	0	0	0	0	15	0	0	-30	0	15	15	15	0	0	0	15	-30	-15	-560.3	293.2	-412.2	280.3
-10	-10	0	0	0	0	-10	0	0	0	-15	0	30	-15	15	0	0	0	15	-30	0	-526.6	263.3	-412.0	273.0
-10	0	0	0	0	0	0	15	0	0	-15	0	15	15	15	0	0	0	15	-30	-15	-568.8	287.3	-416.5	288.8
-10	-10	0	0	0	0	-10	0	0	0	-30	0	-15	-15	15	0	0	0	0	-15	0	-527.1	294.0	-423.1	248.5
-10	-10	0	0	0	0	-10	0	0	0	-15	0	-15	-15	15	0	0	0	0	-15	0	-526.3	298.4	-427.5	254.1
-10	0	0	0	0	0	-10	0	15	0	-30	15	30	-15	15	-15	0	0	15	-30	0	-523.1	273.0	-400.3	253.4

Angles are with respect to the starting structure built from the CRFR1 template.

Table A11. Superbihelix/Supercombihelix results for CRFR1

Theta (θ)							Phi (ϕ)							Eta (η)							Charged Inter-Helical (kcal/mol)	Charged Total (kcal/mol)	Neutral Inter-Helical (kcal/mol)	Neutral Total (kcal/mol)
TM1	TM2	TM3	TM4	TM5	TM6	TM7	TM1	TM2	TM3	TM4	TM5	TM6	TM7	TM1	TM2	TM3	TM4	TM5	TM6	TM7				
0	0	0	0	0	0	-10	0	0	0	-15	0	15	0	0	0	0	0	0	0	15	-533.2	28.5	-483.7	-98.8
-10	0	0	0	0	0	-10	0	15	0	0	0	0	-15	0	0	0	0	0	0	0	-524.9	32.2	-485.7	-98.8
-10	0	0	0	0	0	-10	0	0	0	15	0	-15	0	0	0	0	0	0	0	-15	-534.5	55.5	-492.4	-82.2
0	0	0	0	0	0	-10	0	0	0	15	0	15	0	0	0	0	0	0	0	15	-512.7	33.0	-458.8	-84.8
-10	0	0	0	0	0	-10	0	0	0	-15	0	15	0	-15	0	0	0	0	0	15	-494.3	28.6	-459.5	-82.0
-10	0	0	0	0	0	-10	15	15	0	15	0	0	0	0	0	0	0	0	0	0	-503.0	52.9	-477.7	-62.4
0	0	0	0	0	0	-10	0	0	0	0	0	15	0	0	0	0	0	0	0	15	-500.6	59.8	-453.8	-84.0
0	0	0	0	0	0	-10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-527.8	88.4	-482.1	-76.9
-10	0	0	0	0	0	-10	0	0	0	0	0	15	0	-15	0	0	0	0	0	15	-491.3	29.6	-448.2	-88.6
-10	0	0	0	0	0	-10	0	15	0	15	0	0	0	0	0	0	0	0	0	0	-496.1	38.6	-452.2	-58.5

Angles are with respect to the starting structure built from the CRFR1 template.

Table A12. Superbihelix/Supercombihelix results for GLR

Theta (θ)							Phi (ϕ)							Eta (η)							Charged Inter-Helical (kcal/mol)	Charged Total (kcal/mol)	Neutral Inter-Helical (kcal/mol)	Neutral Total (kcal/mol)
TM1	TM2	TM3	TM4	TM5	TM6	TM7	TM1	TM2	TM3	TM4	TM5	TM6	TM7	TM1	TM2	TM3	TM4	TM5	TM6	TM7				
0	-10	-10	0	-10	10	10	-30	0	0	-30	15	15	0	0	0	0	15	-30	-15	0	-524.1	114.9	-418.8	32.2
0	-10	-10	0	-10	-10	0	0	-15	0	-30	0	15	-15	15	0	-15	0	-15	-60	0	-520.9	116.1	-432.9	63.5
0	-10	-10	0	-10	0	0	-15	0	0	-30	-15	15	30	0	0	0	15	0	-15	0	-512.6	113.6	-418.3	72.7
0	-10	-10	0	-10	-10	0	0	-15	0	-30	0	15	0	15	0	-15	0	0	-45	0	-528.3	120.7	-415.0	82.1
0	-10	-10	0	-10	0	0	0	-15	0	-15	-30	15	0	15	0	-15	0	-15	0	0	-520.8	88.6	-400.7	71.5
0	-10	-10	0	-10	10	10	-15	0	0	-15	15	15	0	0	0	0	15	-30	-15	0	-506.9	96.6	-409.5	38.3
0	-10	-10	0	-10	0	0	0	0	0	-30	-15	15	0	0	0	0	-15	-15	0	15	-506.2	75.9	-409.4	72.9
0	-10	-10	-10	-10	10	10	-15	0	0	-30	15	15	0	0	0	0	0	-30	-15	0	-508.3	102.5	-403.5	45.7
0	-10	-10	0	-10	10	10	-15	0	0	-30	15	15	0	0	0	0	15	-30	-15	0	-506.6	98.5	-401.4	26.8
0	-10	-10	-10	-10	0	0	-15	0	0	-30	-15	15	30	0	0	0	0	0	-15	0	-507.7	105.6	-405.0	72.0

Angles are with respect to the starting structure built from the GLR template.

Table A13. Superbihelix/Supercombihelix results for GIPR

Theta (θ)							Phi (ϕ)							Eta (η)							Charged Inter-Helical (kcal/mol)	Charged Total (kcal/mol)	Neutral Inter-Helical (kcal/mol)	Neutral Total (kcal/mol)
TM1	TM2	TM3	TM4	TM5	TM6	TM7	TM1	TM2	TM3	TM4	TM5	TM6	TM7	TM1	TM2	TM3	TM4	TM5	TM6	TM7				
0	-10	-10	0	-10	0	10	-15	0	0	-30	0	15	-30	0	0	0	15	-15	0	15	-349.9	-399.4	-452.6	-803.5
0	-10	-10	0	-10	0	0	0	-15	0	-30	-15	0	0	15	0	0	15	-15	-15	0	-354.7	-402.5	-432.9	-760.1
0	-10	-10	0	-10	0	10	-15	0	0	-15	-15	30	-15	0	0	0	0	-15	-15	15	-354.0	-400.8	-419.1	-768.8
0	-10	-10	0	-10	0	0	15	-15	0	-15	-15	0	0	15	0	0	0	-15	-15	0	-347.4	-412.0	-420.2	-774.9
0	-10	-10	0	-10	0	0	0	-15	0	-15	-15	0	0	0	0	0	0	-15	-15	0	-349.3	-417.2	-421.8	-763.7
10	-10	-10	0	-10	0	0	0	-15	0	-15	-15	0	0	0	0	0	0	-15	-15	0	-350.6	-420.6	-414.4	-773.7
0	-10	-10	0	-10	0	0	0	0	0	0	-15	0	0	0	0	0	0	-15	-15	0	-363.8	-388.6	-428.5	-759.3
0	-10	-10	0	-10	0	0	-15	-15	0	-15	-15	0	0	15	0	0	0	-15	-15	0	-354.4	-418.0	-409.3	-765.5
-10	-10	-10	0	-10	0	10	-15	0	0	-30	0	15	-30	0	0	0	15	-15	0	15	-336.0	-391.3	-434.5	-781.9
0	-10	-10	0	-10	0	10	-15	0	0	-30	-15	30	-15	0	0	0	0	-15	-15	15	-359.8	-380.0	-437.7	-749.4

Angles are with respect to the starting structure built from the GLR template.

Table A14. Superbihelix/Supercombihelix results for GLP1R

Theta (θ)							Phi (ϕ)							Eta (η)							Charged Inter-Helical (kcal/mol)	Charged Total (kcal/mol)	Neutral Inter-Helical (kcal/mol)	Neutral Total (kcal/mol)
TM1	TM2	TM3	TM4	TM5	TM6	TM7	TM1	TM2	TM3	TM4	TM5	TM6	TM7	TM1	TM2	TM3	TM4	TM5	TM6	TM7				
0	-10	-10	0	-10	0	10	0	-15	0	-15	-15	30	-15	0	0	-15	0	-15	0	15	-492.8	103.3	-415.5	-14.0
0	-10	-10	-10	-10	-10	0	15	-15	0	-15	-15	15	-15	0	0	0	0	15	0	15	-487.9	80.0	-414.5	8.4
0	-10	-10	0	-10	0	10	0	-15	0	-15	-15	15	-15	0	0	-15	0	-15	0	15	-489.8	119.8	-411.6	6.6
0	-10	-10	0	-10	0	10	0	-15	0	-15	-15	30	-15	0	0	-15	0	0	0	15	-477.6	105.4	-406.3	0.3
0	-10	-10	-10	-10	0	10	0	-15	0	15	-15	30	-15	0	0	-15	-15	15	0	15	-469.7	100.9	-414.1	-9.1
0	-10	-10	0	-10	0	10	0	-15	0	-15	-15	0	-30	15	0	-15	0	-15	0	15	-487.7	130.1	-415.5	14.5
-10	-10	-10	0	-10	0	10	0	-15	0	-15	-15	30	-15	-15	0	-15	0	15	0	15	-491.9	105.3	-409.9	29.1
-10	-10	-10	0	-10	0	10	15	-15	0	-15	-15	0	-30	-15	0	-15	0	-15	0	15	-500.8	124.7	-413.3	38.2
0	-10	-10	-10	-10	-10	0	0	-30	0	-15	-15	15	-15	0	0	0	-15	15	0	15	-489.7	101.4	-401.6	38.7
0	-10	-10	0	-10	0	10	15	-15	0	-15	-15	15	-15	0	0	-15	0	-15	0	15	-478.0	114.6	-399.1	24.9

Angles are with respect to the starting structure built from the GLR template.

Table A15. Superbihelix/Supercombihelix results for PACR

Theta (θ)							Phi (ϕ)							Eta (η)							Charged Inter-Helical (kcal/mol)	Charged Total (kcal/mol)	Neutral Inter-Helical (kcal/mol)	Neutral Total (kcal/mol)
TM1	TM2	TM3	TM4	TM5	TM6	TM7	TM1	TM2	TM3	TM4	TM5	TM6	TM7	TM1	TM2	TM3	TM4	TM5	TM6	TM7				
0	0	0	0	0	0	10	-15	0	0	-30	0	0	-15	0	0	0	0	15	0	15	-596.5	329.1	-432.6	238.9
0	0	0	0	0	0	10	15	0	0	-30	0	0	-15	0	0	0	0	15	0	15	-611.1	338.9	-450.1	244.1
0	-10	-10	-10	-10	0	0	-15	-15	0	-15	15	30	30	0	0	0	15	15	0	0	-577.9	344.5	-423.3	226.6
0	0	0	0	0	0	0	-15	0	0	-15	0	15	15	0	0	0	0	15	0	0	-567.3	349.1	-423.8	229.6
0	0	0	0	0	0	0	-15	0	0	-15	0	15	15	0	0	0	15	15	0	0	-571.0	354.9	-432.3	249.3
0	0	0	10	0	0	0	-15	0	0	-30	0	0	0	0	0	0	-15	15	0	0	-570.2	351.6	-426.2	260.2
0	0	0	0	0	0	0	-15	0	0	-15	0	15	15	0	0	0	30	15	0	0	-566.4	355.2	-422.2	261.8
0	0	0	0	0	0	0	-30	0	0	-30	0	0	15	0	0	0	0	15	0	0	-569.4	341.2	-403.9	252.2
0	0	0	0	0	0	0	-15	0	0	0	0	15	15	0	0	0	0	15	0	0	-568.9	348.8	-407.3	258.7
0	0	0	10	0	0	0	0	0	0	-30	0	0	15	0	0	0	-15	15	0	0	-560.1	364.0	-430.6	246.6

Angles are with respect to the starting structure built from the GLR template.

Table A16. Superbihelix/Supercombihelix results for PTH1R

Theta (θ)							Phi (ϕ)							Eta (η)							Charged Inter-Helical (kcal/mol)	Charged Total (kcal/mol)	Neutral Inter-Helical (kcal/mol)	Neutral Total (kcal/mol)
TM1	TM2	TM3	TM4	TM5	TM6	TM7	TM1	TM2	TM3	TM4	TM5	TM6	TM7	TM1	TM2	TM3	TM4	TM5	TM6	TM7				
0	-10	-10	0	-10	0	10	0	-15	0	-30	-15	15	-30	15	0	-15	0	-15	-30	30	-519.6	-93.1	-429.2	-203.2
0	-10	-10	0	-10	0	10	0	-15	0	-30	0	0	-30	15	0	-15	0	30	-15	15	-474.5	-90.8	-422.0	-198.8
0	-10	-10	0	-10	0	10	0	-15	0	-30	-15	15	-30	0	0	-15	0	-15	0	30	-468.9	-108.2	-419.6	-208.0
0	-10	-10	0	-10	0	10	0	-15	0	-30	0	15	-30	15	0	-15	0	15	0	30	-476.3	-72.5	-425.6	-178.0
0	-10	-10	-10	-10	0	10	0	-15	0	0	0	15	-30	15	0	-15	15	15	0	30	-489.3	-71.1	-421.4	-168.7
-10	0	0	0	0	0	10	30	0	0	-15	0	0	-30	0	0	0	0	15	-15	15	-476.6	-57.9	-433.7	-174.1
0	-10	-10	0	-10	10	10	0	-15	0	-30	15	0	-30	15	0	-15	0	0	-15	30	-479.4	-56.9	-430.1	-170.8
0	-10	-10	0	-10	0	10	0	-15	0	-30	-15	15	-30	15	0	-15	0	0	0	15	-470.5	-75.0	-418.9	-172.7
0	-10	-10	0	-10	0	10	-15	-15	0	-15	0	0	-30	15	0	-15	0	30	-15	15	-463.2	-57.6	-424.2	-194.6
0	-10	-10	-10	-10	0	10	0	-15	0	-30	0	15	-30	15	0	-15	0	15	0	30	-465.6	-63.8	-423.8	-174.7

Angles are with respect to the starting structure built from the GLR template.

Table A17. Superbihelix/Supercombihelix results for VIPR1

Theta (θ)							Phi (ϕ)							Eta (η)							Charged Inter-Helical (kcal/mol)	Charged Total (kcal/mol)	Neutral Inter-Helical (kcal/mol)	Neutral Total (kcal/mol)
TM1	TM2	TM3	TM4	TM5	TM6	TM7	TM1	TM2	TM3	TM4	TM5	TM6	TM7	TM1	TM2	TM3	TM4	TM5	TM6	TM7				
-10	0	0	0	0	10	10	30	0	0	-15	0	0	-30	-15	0	0	0	15	-15	15	-493.5	291.3	-400.6	274.2
-10	0	0	10	0	0	10	30	0	0	-30	0	15	-30	-15	0	0	0	15	-15	15	-488.1	309.1	-404.0	256.4
-10	0	0	10	0	0	10	15	0	0	-30	0	15	-30	-15	0	0	0	15	-15	15	-484.6	298.8	-407.1	247.9
0	-10	-10	-10	-10	0	10	-15	0	0	0	15	30	-15	-15	0	0	0	30	-30	15	-506.6	328.9	-390.5	272.2
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-478.9	274.9	-408.4	228.5
-10	0	0	0	0	0	10	15	0	0	-15	0	15	-30	-15	0	0	0	15	-15	15	-481.1	308.3	-396.5	264.7
-10	0	0	0	0	0	10	30	0	0	-15	0	15	-30	-15	0	0	0	15	-30	30	-481.5	311.2	-409.8	276.9
-10	-10	-10	0	-10	10	0	-30	15	15	15	15	0	0	0	0	-15	-15	-15	-30	0	-499.4	338.4	-404.8	290.4
-10	-10	-10	-10	-10	0	10	30	0	0	-30	0	0	-30	0	0	0	0	0	-30	15	-483.2	301.8	-380.6	266.0
0	-10	-10	0	-10	-10	0	-15	-15	0	-30	-15	0	0	0	0	-15	0	0	-30	0	-480.3	324.3	-392.0	287.3

Angles are with respect to the starting structure built from the GLR template.

Table A18. Superbihelix/Supercombihelix results for VIPR2

Theta (θ)							Phi (ϕ)							Eta (η)							Charged Inter-Helical (kcal/mol)	Charged Total (kcal/mol)	Neutral Inter-Helical (kcal/mol)	Neutral Total (kcal/mol)
TM1	TM2	TM3	TM4	TM5	TM6	TM7	TM1	TM2	TM3	TM4	TM5	TM6	TM7	TM1	TM2	TM3	TM4	TM5	TM6	TM7				
0	-10	-10	-10	-10	0	10	0	-15	0	30	15	30	-15	15	0	0	-15	15	-15	15	-537.6	398.4	-428.3	356.3
0	-10	-10	-10	-10	0	10	-15	-15	0	-30	15	30	-15	0	0	0	-15	15	-15	15	-523.6	428.4	-414.8	377.3
0	-10	-10	-10	-10	0	10	-15	-15	0	0	15	15	-15	0	0	0	0	30	0	15	-519.9	424.2	-409.7	374.0
0	-10	-10	0	-10	0	10	-15	-15	0	-30	15	15	-15	0	0	-15	-15	15	0	15	-518.0	467.2	-423.6	388.6
0	-10	-10	0	-10	0	10	-15	-15	0	-30	15	30	-15	0	0	-15	-15	15	-15	15	-511.7	473.0	-426.3	396.3
0	-10	-10	-10	-10	0	10	-15	-15	0	30	15	30	-15	0	0	0	15	15	-15	15	-508.4	433.5	-405.3	389.7
0	0	0	0	0	0	10	15	0	0	-30	0	0	-30	0	0	0	0	15	0	15	-535.1	476.2	-402.8	399.1
0	-10	-10	-10	-10	0	10	-15	-15	0	-15	15	30	-15	0	0	0	15	15	-15	15	-507.8	436.3	-403.0	381.9
0	-10	-10	-10	-10	0	10	-15	-15	0	-30	0	30	-15	0	0	-15	0	30	-15	15	-503.1	448.3	-412.9	391.2
0	-10	-10	-10	-10	0	10	-15	-15	0	-15	15	30	-15	0	0	0	-15	15	-15	15	-506.7	475.7	-418.2	376.0

Angles are with respect to the starting structure built from the GLR template.

References

1. Hollenstein, K.; Kean, J.; Bortolato, A.; Cheng, R. K.; Dore, A. S.; Jazayeri, A.; Cooke, R. M.; Weir, M.; Marshall, F. H., Structure of Class B GPCR Corticotropin-Releasing Factor Receptor 1. *Nature* **2013**, *499* (7459), 438-43.
2. Siu, F. Y.; He, M.; de Graaf, C.; Han, G. W.; Yang, D.; Zhang, Z.; Zhou, C.; Xu, Q.; Wacker, D.; Joseph, J. S.; Liu, W.; Lau, J.; Cherezov, V.; Katritch, V.; Wang, M. W.; Stevens, R. C., Structure of the Human Glucagon Class B G-Protein-Coupled Receptor. *Nature* **2013**, *499* (7459), 444-9.
3. UniProt, C., Activities at the Universal Protein Resource (UniProt). *Nucleic Acids Res* **2014**, *42* (Database issue), D191-8.
4. Moore, E. L.; Burgey, C. S.; Paone, D. V.; Shaw, A. W.; Tang, Y. S.; Kane, S. A.; Salvatore, C. A., Examining the Binding Properties of Mk-0974: A Cgrp Receptor Antagonist for the Acute Treatment of Migraine. *Eur J Pharmacol* **2009**, *602* (2-3), 250-4.
5. Kodra, J. T.; Jorgensen, A. S.; Andersen, B.; Behrens, C.; Brand, C. L.; Christensen, I. T.; Guldbrandt, M.; Jeppesen, C. B.; Knudsen, L. B.; Madsen, P.; Nishimura, E.; Sams, C.; Sidelmann, U. G.; Pedersen, R. A.; Lynn, F. C.; Lau, J., Novel Glucagon Receptor Antagonists with Improved Selectivity over the Glucose-Dependent Insulinotropic Polypeptide Receptor. *J Med Chem* **2008**, *51* (17), 5387-96.
6. Beebe, X.; Darczak, D.; Davis-Taber, R. A.; Uchic, M. E.; Scott, V. E.; Jarvis, M. F.; Stewart, A. O., Discovery and SAR of Hydrazide Antagonists of the Pituitary Adenylate Cyclase-Activating Polypeptide (Pacap) Receptor Type 1 (Pac1-R). *Bioorg Med Chem Lett* **2008**, *18* (6), 2162-6.
7. Carter, P. H.; Liu, R. Q.; Foster, W. R.; Tamasi, J. A.; Tebben, A. J.; Favata, M.; Staal, A.; Cvijic, M. E.; French, M. H.; Dell, V.; Apanovitch, D.; Lei, M.; Zhao, Q.; Cunningham, M.; Decicco, C. P.; Trzaskos, J. M.; Feyen, J. H., Discovery of a Small Molecule Antagonist of the Parathyroid Hormone Receptor by Using an N-Terminal Parathyroid Hormone Peptide Probe. *Proc Natl Acad Sci U S A* **2007**, *104* (16), 6846-51.
8. Harikrishnan, L. S.; Srivastava, N.; Kayser, L. E.; Nirschl, D. S.; Kumaragurubaran, K.; Roy, A.; Gupta, A.; Karmakar, S.; Karatt, T.; Mathur, A.; Burford, N. T.; Chen, J.; Kong, Y.; Cvijic, M.; Cooper, C. B.; Poss, M. A.; Trainor, G. L.; Wong, T. W., Identification and Optimization of Small Molecule Antagonists of Vasoactive Intestinal Peptide Receptor-1 (Vipr1). *Bioorg Med Chem Lett* **2012**, *22* (6), 2287-90.
9. Chu, A.; Caldwell, J. S.; Chen, Y. A., Identification and Characterization of a Small Molecule Antagonist of Human Vpac(2) Receptor. *Mol Pharmacol* **2010**, *77* (1), 95-101.
10. Tibaduiza, E. C.; Chen, C.; Beinborn, M., A Small Molecule Ligand of the Glucagon-Like Peptide 1 Receptor Targets Its Amino-Terminal Hormone Binding Domain. *J Biol Chem* **2001**, *276* (41), 37787-93.
11. Gaylinn, B., Current Research on the Structure and Function of the Growth Hormone Releasing Hormone Receptor. *J Korean Soc Endocrinol* **2006**, *21* (3), 173-183.
12. Abrol, R.; Bray, J. K.; Goddard, W. A., 3rd, Bihelix: Towards De Novo Structure Prediction of an Ensemble of G-Protein Coupled Receptor Conformations. *Proteins* **2012**, *80* (2), 505-18.

13. Bray, J. K.; Abrol, R.; Goddard, W. A., 3rd; Trzaskowski, B.; Scott, C. E., Superbihelix Method for Predicting the Pleiotropic Ensemble of G-Protein-Coupled Receptor Conformations. *Proc Natl Acad Sci U S A* **2014**, *111* (1), E72-8.
14. Schrödinger *Macromodel*, 10.6; 2014.
15. Lang, P. T.; Brozell, S. R.; Mukherjee, S.; Pettersen, E. F.; Meng, E. C.; Thomas, V.; Rizzo, R. C.; Case, D. A.; James, T. L.; Kuntz, I. D., Dock 6: Combining Techniques to Model Rna-Small Molecule Complexes. *RNA* **2009**, *15* (6), 1219-30.
16. Tak Kam, V. W.; Goddard, W. A., 3rd, Flat-Bottom Strategy for Improved Accuracy in Protein Side-Chain Placements. *Journal of Chemical Theory and Computation* **2008**, *4* (12), 2160-2169.
17. Mayo, S. L.; Olafson, B. D.; Goddard, W. A., 3rd, Dreiding: A Generic Force Field for Molecular Simulations. *The Journal of Physical Chemistry* **1990**, *94* (26), 8897-8909.
18. Crooks, G. E.; Hon, G.; Chandonia, J. M.; Brenner, S. E., Weblogo: A Sequence Logo Generator. *Genome Res* **2004**, *14* (6), 1188-90.
19. Kim, S. K.; Goddard, W. A., 3rd; Yi, K. Y.; Lee, B. H.; Lim, C. J.; Trzaskowski, B., Predicted Ligands for the Human Urotensin-Ii G Protein-Coupled Receptor with Some Experimental Validation. *ChemMedChem* **2014**, *9* (8), 1732-43.
20. Kim, S. K.; Goddard, W. A., 3rd, Predicted 3d Structures of Olfactory Receptors with Details of Odorant Binding to Or1g1. *J Comput Aided Mol Des* **2014**, *28* (12), 1175-90.
21. Abrol, R.; Kim, S. K.; Bray, J. K.; Trzaskowski, B.; Goddard, W. A., 3rd, Conformational Ensemble View of G Protein-Coupled Receptors and the Effect of Mutations and Ligand Binding. *Methods Enzymol* **2013**, *520*, 31-48.
22. Kim, S. K.; Riley, L.; Abrol, R.; Jacobson, K. A.; Goddard, W. A., 3rd, Predicted Structures of Agonist and Antagonist Bound Complexes of Adenosine A3 Receptor. *Proteins* **2011**, *79* (6), 1878-97.
23. Kim, S. K.; Li, Y.; Abrol, R.; Heo, J.; Goddard, W. A., 3rd, Predicted Structures and Dynamics for Agonists and Antagonists Bound to Serotonin 5-Ht2b and 5-Ht2c Receptors. *J Chem Inf Model* **2011**, *51* (2), 420-33.
24. Kim, S. K.; Fristrup, P.; Abrol, R.; Goddard, W. A., 3rd, Structure-Based Prediction of Subtype Selectivity of Histamine H3 Receptor Selective Antagonists in Clinical Trials. *J Chem Inf Model* **2011**, *51* (12), 3262-74.
25. Abrol, R.; Kim, S. K.; Bray, J. K.; Griffith, A. R.; Goddard, W. A., 3rd, Characterizing and Predicting the Functional and Conformational Diversity of Seven-Transmembrane Proteins. *Methods* **2011**, *55* (4), 405-14.
26. Kim, S. K.; Li, Y.; Park, C.; Abrol, R.; Goddard, W. A., 3rd, Prediction of the Three-Dimensional Structure for the Rat Urotensin Ii Receptor, and Comparison of the Antagonist Binding Sites and Binding Selectivity between Human and Rat Receptors from Atomistic Simulations. *ChemMedChem* **2010**, *5* (9), 1594-608.
27. Goddard, W. A., 3rd; Kim, S. K.; Li, Y.; Trzaskowski, B.; Griffith, A. R.; Abrol, R., Predicted 3d Structures for Adenosine Receptors Bound to Ligands: Comparison to the Crystal Structure. *J Struct Biol* **2010**, *170* (1), 10-20.
28. Kobilka, B. K.; Deupi, X., Conformational Complexity of G-Protein-Coupled Receptors. *Trends Pharmacol Sci* **2007**, *28* (8), 397-406.

29. Granier, S.; Kobilka, B., A New Era of GPCR Structural and Chemical Biology. *Nat Chem Biol* **2012**, *8* (8), 670-673.
30. Cherezov, V.; Rosenbaum, D. M.; Hanson, M. A.; Rasmussen, S. G.; Thian, F. S.; Kobilka, T. S.; Choi, H. J.; Kuhn, P.; Weis, W. I.; Kobilka, B. K.; Stevens, R. C., High-Resolution Crystal Structure of an Engineered Human Beta2-Adrenergic G Protein-Coupled Receptor. *Science* **2007**, *318* (5854), 1258-65.

Chapter V:
The Full Structure and Small Molecule
Binding Site Prediction of CXCR4,
a Class A G Protein-Coupled Receptor

Abstract

This work presents the full predicted structure and binding site of the chemokine receptor CXCR4. The structure was obtained through homology modeling of the transmembrane helices based on the β 1 adrenergic receptor, along with the GEnSEMBLE methodology. The resulting structure is analyzed, with five interhelical hydrogen bond networks described: TM1-TM2-TM7+, TM2-TM3-TM7, TM5-TM6, TM3-TM4, and TM6-TM7. The anti-HIV drug 1t is docked to the structure with the GenDock procedure, revealing a binding site between Asp171 (TM4) and Glu288 (TM7). The N-terminus, C-terminus, and loops are added, and the structure is optimized in full solvent molecular dynamics. The final structure is compared to the later published CXCR4 crystal structure, and validations of our structure and binding site predictions are performed. It is found that the deviations between our computational structure and the crystal structure are largely caused by the helical abnormalities of the CXCR4 crystal structure.

Introduction

Chemokine receptors are class A GPCRs whose native ligands are chemokines, a family of small pro-inflammatory cytokines (signaling proteins secreted by cells). Activation of these receptors initiates a signaling cascade that involves G protein binding, protein kinase activation, Ca^{2+} mobilization from intracellular stores, and cytoskeletal rearrangement.¹ The native ligand of CXCR4 is the chemokine CXCL12, also known as stromal cell-derived factor-1 (SDF-1). This protein is expressed on hematopoietic stem cells, leukocytes, endothelial cells, platelets, and tumor cells.² CXCR4 and SDF-1 are involved in cell migration in the immune and nervous systems as well as in cancer metastasis.² The importance of CXCR4 and its ligand is highlighted by the fact that deletion of either of their genes will result in embryo lethality.

CXCR4 is also closely involved in HIV entry into cells. Viral entry begins with binding of the gp120 protein on the virus surface to CD4^a on the host cell surface. This interaction causes a conformational change, thereby allowing the complex to bind to a chemokine receptor and trigger fusion between viral and host membranes. CXCR4 is implicated in the entry of T-tropic (X4) strains of HIV into host cells, which are involved in the later stage of HIV and cause rapid CD4⁺ T cell depletion and progression towards AIDS.³ The mechanics of HIV viral attachment and entry suggest an important role for CXCR4 as a potential drug target to combat the AIDS epidemic. Two main methods to prevent the spread of HIV via targeting CXCR4 have been designed.⁴ The first involves causing the removal of the receptor from the cell surface, which the native chemokines have been known to do (d in **Figure 3**). Several peptide derivative ligands based on the structure of SDF-1 have been created to prevent the spread of HIV via this method. However, as peptides, these compounds tend to be less orally bioavailable than small molecule options. As a consequence the second method of fusion inhibition is now the target of drug design (c in **Figure 3**). In this method, the attachment of small molecules to CXCR4 influence its conformation such that the gp120 can no longer bind and trigger viral fusion, thereby preventing the spread of HIV infection. A series of orally bioavailable, highly potent, selective CXCR4 inhibitors was designed by Novartis.² The most potent ligand was 1t, which is shown below (see **Figure 4**).

In this work we predict the structure of CXCR4 utilizing the GEnSEMBLE⁵⁻⁶ method of generating an ensemble of GPCR conformations. A selection of the lowest energy conformations are docked to 1t. The complexes with the best energies are relaxed via full-atom molecular dynamics. The lowest energy structure from the molecular dynamics is analyzed and compared to the later-published CXCR4 crystal structure.⁷

^a The normal function of CD4 is to activate helper T cells in response to infectious particles.

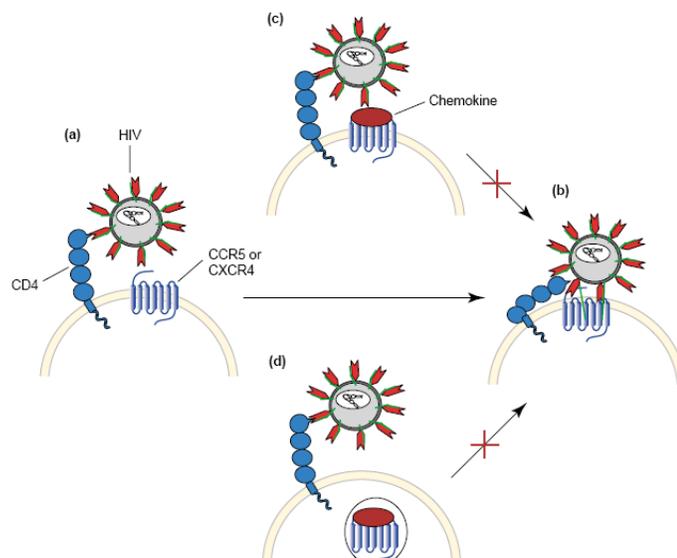


Figure 3. Inhibition of HIV infection by chemokines. The virus first undergoes a high-affinity interaction with CD4 (blue) on the cell surface (a), followed by a conformational change that enables it to interact with an essential co-receptor, usually either CXCR4 or CCR5. This in turn allows the virus to fuse with the cell membrane (d). Two methods of fusion inhibition are (c) a conformational change preventing gp120 binding, and (d) causing the removal of the receptor from the cell surface.⁴

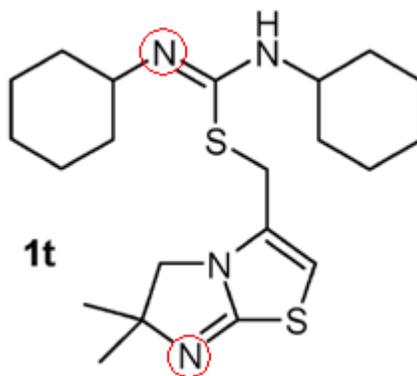


Figure 4. CXCR4 inhibitor 1t. The protonatable nitrogens are circled in red.²

Methods

A brief overview of the methodology used for structure and binding site prediction is outlined below. More details for each step are given in Chapter II:

1. Homology helix prediction - Helix lengths and locations were determined by secondary structure prediction servers. The helix translations (x/y, z) and orientations (θ , ϕ , η) were taken from the known crystal structures.
2. Bihelix - Rotation of the helices in thirty degree increments (η) and calculation of the energies of each interacting pair of helices.
3. Combihelix - Assembly of the seven-helical bundles and energy calculation to determine the best helix rotations.
4. SuperBihelix - Sampling the θ , ϕ , and η angles of the top Combihelix structures.
5. Supercombihelix- Assembly and scoring of the Superbihelix bundles.
6. Ligand conformational search
7. GenDock - Docking of 1t to the best structures from Supercombihelix.
8. N-terminus, C-terminus, and loops connecting the transmembrane helices added.
9. Molecular Dynamics - Full solvent and lipid molecular dynamics.

The first step in the CXCR4 structure prediction was the determination of the helix lengths and locations. Homology helix lengths and locations were obtained using the secondary structure prediction servers Porter, APSSP2, and SSPRO.⁸⁻¹⁰ A consensus helix length and location was determined between the three servers – if two out of three of the servers agreed that a residue was helical in character, that residue was included in the CXCR4 helix.

Homology modeling was then used to create the helices. This process took an x-ray crystal structure ($\beta 1$, $\beta 2$, and A2A)¹¹⁻¹³, removed the loops, and split it into separate helices. CXCR4 was aligned to the homology GPCR based on the known Class A GPCR motifs. Each helix was then mutated to the correct CXCR4 residues, minimized using a conjugate gradient method for 100 steps, and brought back together to form a bundle. This was done for the $\beta 1$, $\beta 2$, and A2A templates. Bovine Rhodopsin was not included due to its low sequence identity with CXCR4.

After the TM regions were predicted and the homology helices constructed, the η angles of the helices were sampled via Bihelix/Combihelix in thirty degree increments (from 0-360 for each helix) two at a time, ignoring all other helices in the procedure known as BiHelix.⁵ The conformation and template with the lowest energy was selected from the top 1,000 energies output in CombiHelix. This structure was then further optimized using Superbihelix/Supercombihelix which locally sampled the θ , ϕ , and η angles. These variations of the starting structure were explored: θ (-10, 0, 10), ϕ (-30, -15, 0, 15, 30), and η (-30, -15, 0, 15, 30).⁶ Based on the average ranking of the neutral interhelical, neutral total, charged interhelical, and charged total energies, along with a desire to fully sample the conformational space via RMSD, six protein structures were selected to continue to the binding site determination.

Next, the 1t ligand was prepared to be used in docking. A systematic torsional scan using Maestro's MacroModel conformation search was run with four starting points of the left and right cyclohexane rings: up/up, down/down, down/up, and up/down. These structures were minimized and clustered according to LigCluster. Following structure selection and ligand preparation, the GenDock docking procedure was used on 1t with the six protein structures selected from Supercombihelix. The docking results were then analyzed based on the snap binding energies. Poses with low snap binding energy that also showed a diversity of binding modes were selected. Extracellular and intracellular loops, C-termini, and N-termini were constructed based on homology modelling to the template selected in Bihelix/Superbihelix. These structures then underwent 10ns of full-atom molecular dynamics.

Results and Discussion

Structure Prediction

Helix lengths were determined for CXCR4 via the secondary structure prediction servers. For each template (β 1, β 2, and A2A), the helices were then aligned via the conserved Class A GPCR motifs. The only exceptions were for TMs 2 and 4. These helices had two different alignments- the conserved motif or the conserved prolines (see **Figure 5**). Both alignments were used in the succeeding steps.

TM2 - Proline Alignment

β 1 74 TLTNLFITSLACADLVVGLLVV**P**FGATLVVVG 105
CXCR4 70 RSMTDIYLLNLAISDLFFLLTV**P**FWAVDAVAN 101

TM2 - Conserved Residue Alignment

β 1 74 TLTNLFITSL**L**ACAD**D**LVVGLLVV**P**FGATLVVVG 105
CXCR4 71 SMTDIYLLN**L**AIS**D**LFFLLTV**P**FWAVDAVANW 102

TM4 - Proline Alignment

β 1 154 TRARAKVIICTVW AISALVSFL**P**IMM 179
CXCR4 148 RLLAEKVVYVGVWIPALLLT**P**DFI 173

TM4 - Conserved Residue Alignment

β 1 154 TRARAKVIICTV**W**AISALVSFLPIMM 179
CXCR4 149 KLLAEKVVYVGV**W**IPALLLTIPDFI 174

Figure 5. Two possible alignments of TMs 2 and 4 of β 1 and CXCR4. The helix may be aligned by conserved Class A GPCR motif (LxxxD), or via the conserved Proline, both of which are depicted in red. Both alignments were pursued for the initial template generation and BiHelix/ComBiHelix structure predictions.

Following alignment, each TM bundle was built according to its respective crystal angles, tilts, and x/y/z coordinates. Bihelix was then run, with the result that the β 1 template with proline alignments for TMs 2 and 4 was lower energy than any of the other combinations. The lowest energy bundle had all zero eta rotations with respect to the β 1 crystal structure, and is depicted in yellow in **Table 1**. To further refine this TM bundle, Superbihelix was run. The number one structure had these angles: eta (0_15_0_15_-30_0_-15), phi (15_-15_-15_-15_-15_-15_-15), and theta (-10_0_-10_10_10_-10_-10) (shown in yellow in **Table 2**). Since proteins are known to exist in a variety of low energy conformations, five structures other than the number one were selected from the top ten Superbihelix ensembles. They were chosen based on their RMSD, so that the sample space was adequately explored (see **Table 3**).

Table 1. Superbihelix/Combihelix results for the lowest ranked templates

Template	Eta							Charged Interhelical	Charged Total	Neutral Interhelical	Neutral Total
	H1	H2	H3	H4	H5	H6	H7				
β1 Proline	0	0	0	0	0	0	0	-438	522	-370	347
β 2 Proline	0	0	0	90	270	0	90	-525	653	-374	510
β 2 Conserved	0	0	0	120	270	60	60	-463	636	-357	530
β 1 Conserved	0	0	0	120	150	60	30	-428	738	-330	648

The template selected in yellow was used for further analysis. All energies are in kcal/mol. All angles are with respect to their template.

Table 2. Top Supercombihelix/Supercombihelix structures.

Eta							Phi							Theta							Average
H1	H2	H3	H4	H5	H6	H7	H1	H2	H3	H4	H5	H6	H7	H1	H2	H3	H4	H5	H6	H7	Rank
0	15	0	15	-30	0	-15	15	-15	-15	-15	-15	-15	-15	-10	0	-10	10	10	-10	-10	32
0	15	0	15	0	30	-15	-15	-15	-15	-30	0	-15	0	0	-10	10	0	-10	-10	-10	54
0	15	0	15	-15	30	-15	-15	-15	-15	-15	-30	0	-15	0	0	-10	10	0	-10	-10	58
30	15	0	15	15	0	0	0	-15	30	-30	-30	-15	0	-10	-10	-10	0	0	-10	-10	63
30	15	0	15	-15	0	0	-15	-30	0	-15	-30	0	0	-10	-10	-10	0	10	-10	0	64
0	15	0	15	15	0	15	0	0	0	-30	-30	-15	0	-10	-10	0	0	0	-10	-10	64
0	15	0	15	0	0	15	-15	0	0	0	-30	-30	0	-10	-10	0	-10	10	-10	-10	66
30	15	0	15	-30	0	0	-15	-30	0	-15	-30	0	0	-10	-10	-10	0	10	-10	-10	67
0	15	0	15	0	0	15	0	0	0	-30	-30	-30	0	-10	-10	0	-10	10	-10	-10	67
0	15	0	15	15	0	0	15	-15	-15	-15	-30	-15	-15	-10	0	-10	10	0	-10	-10	69

The structures flagged in yellow were used for docking. The pose in red was the final best structure after docking and molecular dynamics.

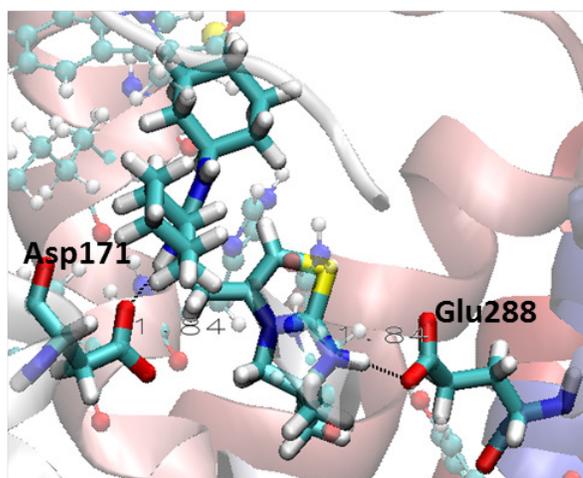
Table 3. RMSD matrix of the top 10 structures from Supercombihelix.

	1	2	3	4	5	6	7	8	9	10
1	0.0	1.9	1.9	1.7	1.4	1.2	0.6	1.5	1.5	2.0
2	1.8	0.0	1.5	1.7	2.1	1.9	2.4	2.0	2.1	0.9
3	1.7	1.5	0.0	1.7	1.9	1.8	1.9	1.8	1.9	1.6
4	2.2	0.4	0.4	0.0	1.9	1.8	2.0	1.7	1.9	1.7
5	2.0	1.7	1.8	1.8	0.0	1.4	1.9	1.6	1.7	1.6
6	2.3	1.9	1.9	1.6	1.6	0.0	1.4	0.7	1.4	2.1
7	2.4	1.8	1.8	1.4	1.9	1.9	0.0	1.9	1.2	1.7
8	1.8	1.9	2.0	1.9	1.4	1.3	1.3	0.0	0.6	2.3
9	2.1	1.8	1.7	1.6	0.7	1.9	1.5	1.5	0.0	2.1
10	1.3	1.6	1.7	1.6	2.1	1.7	2.3	2.1	2.0	0.0

The structures flagged in yellow were docked to using GenMSCDock.

Binding Site Prediction

Conformations of 1t were generated using Maestro's systematic torsional scan with four starting points. This resulted in 46 conformations of 1t. All of these structures were docked to the six protein structures from Supercombihelix. The resulting protein-ligand complexes were scored by snap binding energy. A combination of very low snap binding energies (top ten for that protein conformation), along with a desire for diverse poses was used to select six complexes for further study using molecular dynamics. After 10ns of MD, the best pose was chosen based on its snap binding energy after being run through the post-DarwinDock modules of GenDock. The pre-MD angles of this pose are shown in red in **Table 2** and ligand binding site and cavity analysis are show in **Figure 6**. As the figure shows, two hydrogen bonds are formed between the protein and the ligand's charged nitrogens.



Residue	VdW	Coulomb	H-Bond	NonBond
Glu 288	3.8	-5.1	-4.1	-5.4
His 113	-4.4	0.3	0.0	-4.2
Trp 195	-4.0	-0.1	0.0	-4.1
Lys 110	-4.0	0.0	0.0	-4.1
Gln 200	-3.0	-0.7	0.0	-3.7
Asp 171	-2.0	-0.7	-0.2	-2.9
Val 114	-2.2	-0.2	0.0	-2.4
Cys 109	-2.1	-0.1	0.0	-2.3
Tyr 256	-1.0	-1.0	-0.1	-2.1

Figure 6. Binding pose and unified cavity analysis for the rank 1 protein-ligand complex. Note the two hydrogen bonds that the charged nitrogens on the ligand make to the protein residues Glu288 and Asp171. All energies are in kcal/mol.

Structure Analysis

The best CXCR4-1t structure after molecular dynamics had five sets of interhelical hydrogen bonds: TM1-TM2-TM7+, TM2-TM3-TM7, TM5-TM6, TM3-TM4, and TM6-TM7. The 1.50-2.50 conserved Class A GPCR hydrogen bond motif was present in our structure: N56(TM1)-D84(TM2). However, CXCR4 is unable to make the Class A 3.42-2.45-4.50 hydrogen bond network because the 3.42 and 2.45 residues are not the typically conserved residues.

The largest region of interlocked hydrogen bonds occurred in the intracellular portion of primarily TMs 1, 2, and 7 (with a few residues in TMs 3, 4, 6). These are shown in **Figure 7A**. Notably, there is a strong salt bridge between R77(TM2) and E153(TM4). The second largest hydrogen bond network occurs at the extracellular ends of TMs 2, 3, and 7, as is shown in **Figure 7B**. Finally, three smaller sets of interactions occur between TMs 5-6, 3-4, and 6-7 as shown in **Figure 7C, D, and E** respectively.

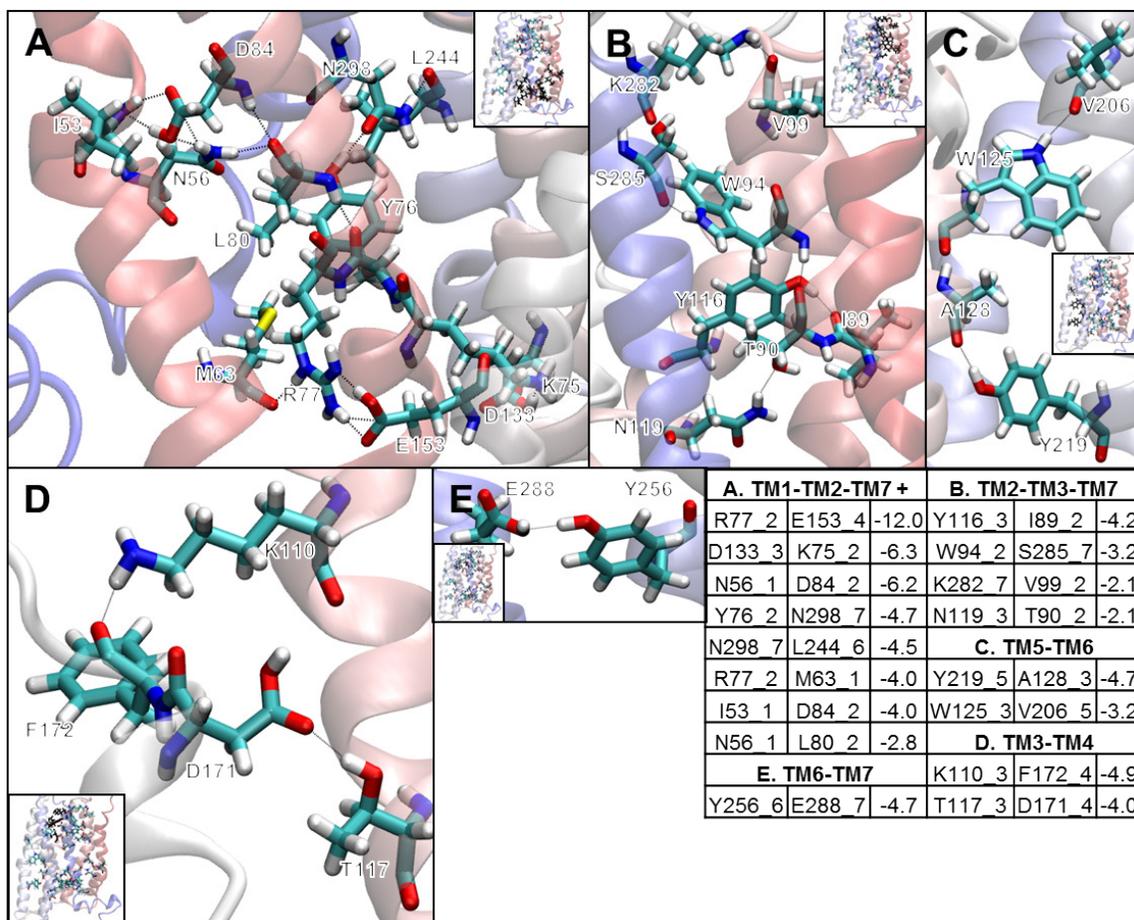


Figure 8. Interhelical interactions in the rank 1 CXCR4 structure. In the table $_X$ indicates the transmembrane helix that the residue is a member of. All energies are in kcal/mol. (A) TM1-TM2-TM7+ hydrogen bond network, (B) TM2-TM3-TM7, (C) TM5-TM6, (D) TM3-TM4, and (E) TM6-TM7.

Comparison to the CXCR4 Crystal Structure

After the previous work was completed, a crystal structure of CXCR4 bound to 1t was published (PDB ID: 3ODU).⁷ The TM bundle and ligand RMSDs can be found in **Figure 9**. The resolution of the structure was 2.5Å, and the T4 lysozyme fusion protein was inserted at the intracellular junction of TMs 5 and 6. Here we compare our structure to that of the CXCR4 crystal. **Figure 9** presents an overview of the CXCR4 crystal: CXCR4 predicted structure comparison. The average TM RMSD between the two structures was 4.0 Å. Also, the predicted ligand pose itself was not in same location – our ligand targeted Asp171 and Glu288, while the crystallized ligand interacted with Asp97

and Glu288. The possible reasons for the discrepancy between the predicted and the crystal structure are discussed below.

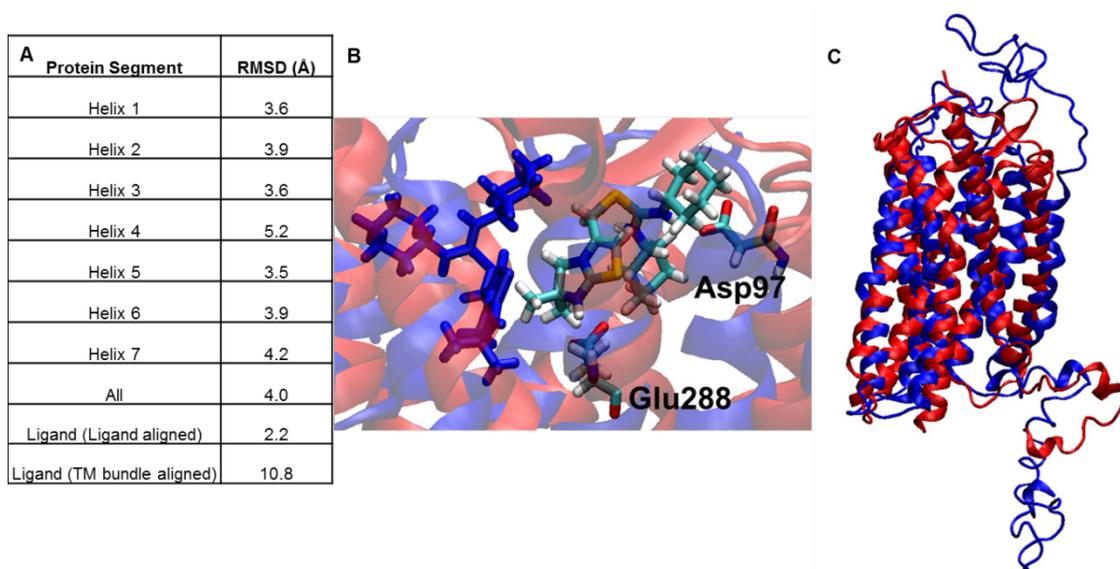


Figure 9. Comparisons of the predicted and x-ray crystal CXCR4 structures. (A) RMSD comparison of each helix of our structure to the x-ray crystal structure. All RMSD are backbone RMSD, with the TM bundle aligned, unless specified otherwise. (B) Our structure (blue), x-ray crystal structure (red and atom colored). (C) Our structure (blue), x-ray crystal structure (red).

To dissect the potential flaws in our structure prediction procedure, we validated each of them sequentially:

First, we explored Bihelix/Combihelix. When these programs were run on the crystal helices and template, we obtained an all-zero conformation as our number 1 structure by the average neutral interhelical, charged interhelical, neutral total, and charged total energy ranks, our usual scoring methodology (**Table 4**). We then ran our Superbihelix/Supercombihelix procedure, which showed the crystal all-zero rotations in the top five (specifically fourth), a rank sufficiently high to validate those procedures as well. The number one structure from Superbihelix/Supercombihelix only showed a single variation from the crystal structure: the ϕ angle of TM4 was 30° (**Table 5**).

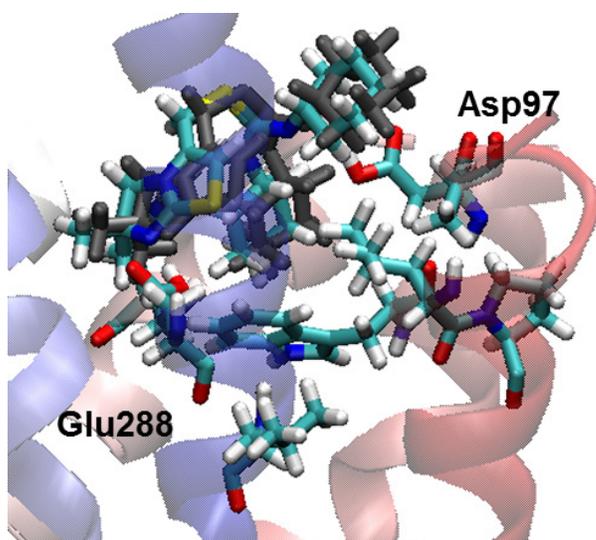
Table 4. Top Bihelix/Combihelix structures for the CXCR4 crystal helices

Eta						
H1	H2	H3	H4	H5	H6	H7
0	0	0	0	0	0	0
0	0	0	0	0	60	0
0	0	0	0	0	0	0
0	0	0	30	0	0	0
0	0	0	0	330	0	0
0	0	0	0	0	60	330
0	0	0	0	270	0	0
0	0	0	330	0	0	0
0	0	0	330	270	0	0
0	0	0	0	330	60	0

Table 5. Top Superbihelix/SuperCombihelix structures for the CXCR4 crystal helices

Theta							Phi							Eta							
H1	H2	H3	H4	H5	H6	H7	H1	H2	H3	H4	H5	H6	H7	H1	H2	H3	H4	H5	H6	H7	
0	0	0	0	0	0	0	0	0	0	30	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	30	0	-15	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	15	0	0	0	0	0	0	-15	0	0	0	0
0	0	0	0	0	0	0	0	0	0	15	0	-15	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	-30	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	30	15	-15	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	30	15	-30	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	-15	-15	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	-30	0	0	0	0	0	0	-15	0	0	0	0

Next, we considered the docking procedure. First, we looked at the ligand conformation prediction. Out of our 46 ligand conformations, two ligands were generated with RMSDs of less than 1Å between the computational and crystal ligands. Next, we considered the docking procedure itself. We docked the crystal ligand conformation to the crystal protein structure. We obtained several poses with an RMSD of less than 1Å to that of the crystal structure, including the number one structure by snap binding energy (**Figure 10**). This structure also interacted with the correct residues on CXCR4 – Asp97 and Glu288. Therefore, neither our ligand conformation nor our docking procedure was the reason for our relatively poor results.



Residue	VdW	Coulomb	H-Bond	NonBond
ASP 97	-3.3	-1.7	-0.6	-5.6
TRP 94	-4.3	-0.1	0.0	-4.4
GLU 288	3.7	-2.7	-4.7	-3.7
LEU 41	-2.2	-0.8	0.0	-3.0
HIS 113	-3.2	0.6	0.0	-2.6
LEU 290	-1.9	0.1	0.0	-1.9
ALA 98	-1.4	0.0	0.0	-1.3
PRO 42	-1.1	0.1	0.0	-1.0

Figure 10. Binding site of the GenDock docking of the crystal 1t to the crystal protein. The crystal pose is shown in gray. All energies are in kcal/mol.

This left us with one final comparison to make – that of the actual helix structures, particularly any atypical helices that overwound or unwound compared to a canonical alpha helix, since our procedure does not account for this. As discussed in the Steven paper, the CXCR4 structure is remarkably different from the other structures that have been crystallized thus far (**Figure 11**). The largest deviations, as mentioned in the Stevens paper, are:¹⁴

1. The EC end of helix 1 is shifted toward the center of the bundle by 9Å compared to β 2 and by 3Å compared to A2A.
2. Helix 2 is overwound near Pro92, which allowed both Asp97 and the conserved Asp84 to face inwards to interact with the ligand and form the 1-2-7 conserved hydrogen bond network, respectively.
3. Both ends of helix 4 show large variations (~5 and 3Å) from those of the other GPCRS.
4. The EC end of helix 5 is one turn longer in CXCR4.
5. Helix 6's EC end is shifted by ~3Å from β 2 and A2A.
6. The EC end of helix7 is two turns longer.

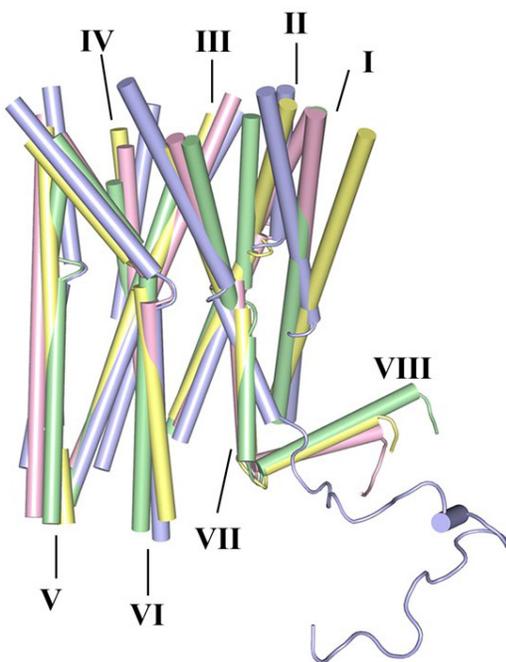


Figure 11. Comparison of four class A GPCR crystal TM bundles: CXCR4 (purple), β 2AR (yellow), A2AR (green) and rhodopsin (pink).

These deviations indicate that using any sort of homology helices or homology bundle template would result in large errors in the TM bundle predictions. We experienced this issue, as did all of the GPCR Dock 2010 participants.¹⁴ No one did particularly well on the bundle, and most of the ones who were close on the ligand binding site used mutagenesis information to pick their best poses, and as such were not the sort of *ab initio* methods that we strive for.

While our structure was not identical to that of the CXCR4 crystal structure, it is important to keep several things in mind. First of all, GPCRs exist in several low lying conformational states.¹⁵ The helix changes necessary to optimally bind to the ligand 1t, while preserving the class A GPCR conserved TMs 1-2-7 interactions and also crystallize (with T4L between TMs 5 and 6), is one conformation of CXCR4. Our methods may have obtained another viable conformation of CXCR4.

Conclusion

We present here the full structure of CXCR4 bound to 1t, a small ligand antagonist that is a potential anti-HIV drug. The structures are predicted through the GEnSeMBLE and GenDock methodologies. Our structure showed strong hydrogen

bonds between 1t and Asp171(TM4) and Glu288 (TM 7), as well as large interhelical hydrogen bond networks: TM1-TM2-TM7+, TM2-TM3-TM7, TM5-TM6, TM3-TM4, and TM6-TM7. While our final structure showed significant deviations from the later published CXCR4 crystal structure, we are able to validate all steps of our methodology, including: Bihelix/Combihelix, Superbihelix/Supercombihelix, and GenDock. However, since our methodology did not allow for helix shapes to vary, and the crystal structure helices were deformed greatly from canonical alpha-helices, our structure differed from that of the crystal structure. We suggest that our structure is another valid low-lying structure of CXCR4, just not the one which was favored by the crystallization conditions.

References

1. Ueda, S.; Oishi, S.; Wang, Z. X.; Araki, T.; Tamamura, H.; Cluzeau, J.; Ohno, H.; Kusano, S.; Nakashima, H.; Trent, J. O.; Peiper, S. C.; Fujii, N., Structure-Activity Relationships of Cyclic Peptide-Based Chemokine Receptor Cxcr4 Antagonists: Disclosing the Importance of Side-Chain and Backbone Functionalities. *J Med Chem* **2007**, *50* (2), 192-8.
2. Thoma, G.; Streiff, M. B.; Kovarik, J.; Glickman, F.; Wagner, T.; Beerli, C.; Zerwes, H. G., Orally Bioavailable Isothioureas Block Function of the Chemokine Receptor Cxcr4 in Vitro and in Vivo. *J Med Chem* **2008**, *51* (24), 7915-20.
3. Hatse, S.; Princen, K.; Bridger, G.; De Clercq, E.; Schols, D., Chemokine Receptor Inhibition by Amd3100 Is Strictly Confined to Cxcr4. *FEBS Lett* **2002**, *527* (1-3), 255-62.
4. Wells, T. N.; Power, C. A.; Shaw, J. P.; Proudfoot, A. E., Chemokine Blockers--Therapeutics in the Making? *Trends Pharmacol Sci* **2006**, *27* (1), 41-7.
5. Abrol, R.; Bray, J. K.; Goddard, W. A., 3rd, Bihelix: Towards De Novo Structure Prediction of an Ensemble of G-Protein Coupled Receptor Conformations. *Proteins* **2012**, *80* (2), 505-18.
6. Bray, J. K.; Abrol, R.; Goddard, W. A., 3rd; Trzaskowski, B.; Scott, C. E., Superbihelix Method for Predicting the Pleiotropic Ensemble of G-Protein-Coupled Receptor Conformations. *Proc Natl Acad Sci U S A* **2014**, *111* (1), E72-8.
7. Wu, B.; Chien, E. Y.; Mol, C. D.; Fenalti, G.; Liu, W.; Katritch, V.; Abagyan, R.; Brooun, A.; Wells, P.; Bi, F. C.; Hamel, D. J.; Kuhn, P.; Handel, T. M.; Cherezov, V.; Stevens, R. C., Structures of the Cxcr4 Chemokine GPCR with Small-Molecule and Cyclic Peptide Antagonists. *Science* **2010**, *330* (6007), 1066-71.
8. Pollastri, G.; McLysaght, A. Porter: A New, Accurate Server for Protein Secondary Structure Prediction. <http://distill.ucd.ie/porter/> (accessed 1/14).
9. Raghava, G. P. S. Apssp2 : A Combination Method for Protein Secondary Structure Prediction Based on Neural Network and Example Based Learning. <http://www.imtech.res.in/raghava/apssp2/> (accessed 1/14).
10. Bryson, K.; McGuffin, L.; Marsden, R.; Ward, J.; Sodhi, J.; Jones, D. Protein Structure Prediction Servers at University College London. <http://bioinf.cs.ucl.ac.uk/psipred/> (accessed 1/14).
11. Cherezov, V.; Rosenbaum, D. M.; Hanson, M. A.; Rasmussen, S. G.; Thian, F. S.; Kobilka, T. S.; Choi, H. J.; Kuhn, P.; Weis, W. I.; Kobilka, B. K.; Stevens, R. C., High-Resolution Crystal Structure of an Engineered Human Beta2-Adrenergic G Protein-Coupled Receptor. *Science* **2007**, *318* (5854), 1258-65.
12. Warne, T.; Serrano-Vega, M. J.; Baker, J. G.; Moukhametzianov, R.; Edwards, P. C.; Henderson, R.; Leslie, A. G.; Tate, C. G.; Schertler, G. F., Structure of a Beta1-Adrenergic G-Protein-Coupled Receptor. *Nature* **2008**, *454* (7203), 486-91.
13. Jaakola, V. P.; Griffith, M. T.; Hanson, M. A.; Cherezov, V.; Chien, E. Y.; Lane, J. R.; Ijzerman, A. P.; Stevens, R. C., The 2.6 Angstrom Crystal Structure of a Human A2a Adenosine Receptor Bound to an Antagonist. *Science* **2008**, *322* (5905), 1211-7.
14. Kufareva, I.; Rueda, M.; Katritch, V.; Stevens, R. C.; Abagyan, R., Status of GPCR Modeling and Docking as Reflected by Community-Wide GPCR Dock 2010 Assessment. *Structure* **2011**, *19* (8), 1108-26.

15. Kenakin, T.; Miller, L. J., Seven Transmembrane Receptors as Shapeshifting Proteins: The Impact of Allosteric Modulation and Functional Selectivity on New Drug Discovery. *Pharmacological Reviews* **2010**, *62* (2), 265-304.