STRUCTURAL STUDIES OF THE INTERACTION BETWEEN MGLU5 AND ALLOSTERIC MODULATORS

By

Elizabeth Dong Nguyen

Dissertation

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

in

Chemical and Physical Biology

August, 2013

Nashville, Tennessee

Approved:

Professor Walter J. Chazin

Professor P. Jeffrey Conn

Professor Terry P. Lybrand

Professor David L. Tabb

ACKNOWLEDGEMENTS

The work presented here was made possible by Public Health Service award T32 GM07347 from the National Institute of General Medical Studies for the Vanderbilt Medical-Scientist Training Program (MSTP), the Paul Calabresi Medical Student Research Fellowship from the Pharmaceutical Research and Manufacturers of America (PhRMA) Foundation and a Deutscher Akademischer Austausch Dienst (DAAD) research grant from the German Academic Exchange Service. Computational resources were provided by the Advanced Computing Center for Research and Education and the Center for Structural Biology at Vanderbilt University.

I would like to thank my advisor, Dr. Jens Meiler, who fostered in me the desire to pursue research as a summer undergraduate student and continued to support, guide and mentor me through my graduate studies. His excitement for science is infectious and has inspired my own ambitions to spread the passion of science to others.

I would like to thank other faculty members who aided in my scientific growth while at Vanderbilt University, particularly my committee members: Dr. Walter Chazin, Dr. Jeff Conn, Dr. Terry Lybrand and Dr. David Tabb. They have continually provided guidance and encouragement for my research progress and career development.

The Vanderbilt MSTP program has been a steady and consistent source of support and the strongest advocates for my success. I'm thankful to Dr. Terry Dermody for initially recruiting me into the program and for continual inspiration, as well as Dr. Michelle Grundy, Dr. Jim Bills, Dr. Larry Swift and Melissa Krasnove for their continual support. Thanks also to my MSTP mentor Dr. Dan Roden, who supported me from the day I interviewed at Vanderbilt. The Vanderbilt Chemical and Physical Biology program has been instrumental

to my success in graduate school and I'd like to thank the following members for their help: Dr. Hassane Mchaourab, Dr. Al Beth, Dr. Bruce Damon and Lindsay Meyers.

I want to say a special thank you to Dr. Karen Gregory from the Conn lab for her mentorship. She taught me everything I know about the wet lab and was instrumental in helping me put together my thesis work. I am ever grateful for her guidance and all I learned from her about being a great scientist. I'd also like to thank Sabuj, Roy, Jarrod and Jonathan from the Center of Structural Biology for their computational expertise.

Current and former members of the Meiler lab have been invaluable as friends and colleagues, both in the lab and outside of it. Dr. Julia Leman has been a wonderful friend and her encouragement has shaped my growth as a scientist and a person. My thanks also go out Stephanie DeLuca and Dr. Nathan Alexander of Room 5154G for support on a daily basis. For everyone else in the lab who patiently answered my questions and with whom I've been to the mountains and beach together and back, I'm thankful as well.

I am certain I have the best MSTP class, whom I consider my family here in Nashville, particularly Frances, Eric, Tanner, John, Brian, & Caroline. Additionally, my friends from around the world have been there for me in the best and worst of times and for their support and prayers I am eternally grateful: Linda, Lisa, Elaine, Suzanne, Mari, Lauren, Anna, Veronica, Thuy, Meryl and many more.

I have been blessed with the most wonderful family. My husband, Kevin Nguyen, has been by my side as a source of love and support from the moment I embarked on this journey. I couldn't have done any of this without him. My dad (Hoa), mom (Lori), sister (Emily) and brother (Jon) have always been an inspiration to me and my strongest supporters. I strive to be more like them each day. Above all, I am thankful to my God for filling me with strength, love, hope and purpose (Jeremiah 29.11).

TABLE OF CONTENTS

Page

Appendix

LIST OF TABLES

Table Page

LIST OF FIGURES

SUMMARY

The overall focus of this dissertation was to design and validate methods for the computational modeling of ligands in complex with G protein-coupled receptors (GPCRs) and apply the methods to the modeling of allosteric modulators in metabotropic glutamate receptor subtype 5 ($mGlu_5$).

Chapter I introduces the importance of GPCRs as a target for drug development, with $mGlu_5$ being an especially important target for disorders of cognitive function. The introduction discusses the structural information known about $mGlu_5$ and other GPCRs up until this point and highlights computational modeling in combination with experimental data as a strategy towards elucidating new structural information on GPCR drug targets that do not have an experimentally determined structure. Portions of Chapter I came from a review entitled “Allosteric modulation of metabotropic glutamate receptors: structural insights and therapeutic potential” written by Karen Gregory, Elizabeth Dong, Jens Meiler and Jeff Conn. The author of this dissertation contributed all portions of review regarding structural information on metabotropic glutamate receptors as well as the computational modeling results.

Chapter II outlines the currently established protocol for docking small-molecules into comparative models of proteins using the Rosetta protein prediction software suite. Chapter II is based on the publication entitled “Small-molecule ligand docking into comparative models with Rosetta” by Steven Combs, Sam DeLuca, Stephanie DeLuca, Gordon Lemmon, David Nannemann, Elizabeth Nguyen, Jordan Willis, Jonathan Sheehan and Jens Meiler. While compilation of this work was a collaborative effort between eight co-first authors, the protocol on which this work was based was developed and presented by the author of this dissertation. The author of this dissertation also generated the modeling and docking data and analysis presented in this work.

Chapter III presents specific tailoring of the Rosetta comparative modeling and docking protocol to GPCRs. It also presents a thorough benchmark that evaluates the accuracy of modeling the ligand-receptor complex with a set of 14 Class A GPCRs. Chapter III is based on the publication entitled “Assessment and challenges of ligand docking into comparative models of G protein-coupled receptors” by Elizabeth Nguyen, Christoffer Norn, Thomas Frimurer and Jens Meiler. The author of this dissertation developed the initial strategy for this work, designed the experiments and generated the comparative models. Collaborative efforts in data analysis and writing of the text resulted in co-first authorship between the first two authors of this paper.

Chapter IV presents the first application of the comparative modeling and docking protocols to $mGlu_5$. The computational modeling is used to understand how MPEP as well as picolinamide and nicotinamide allosteric modulators based on the MPEP scaffold interact with the receptor. In particular, we examined why particular receptor mutations cause a switch in modulator function. Chapter IV is based on the publication entitled “Probing the metabotropic glutamate receptor 5 ($mGlu_5$) positive allosteric modulator (PAM) binding pocket: discovery of point mutations that engender a “molecular switch” in PAM pharmacology” by Karen Gregory, Elizabeth Nguyen, Sean Reiff, Emma Squire, Shaun Stauffer, Craig Lindsley, Jens Meiler and Jeff Conn. The author of this dissertation contributed all aspects of computational modeling to this work, interpreted in light of the mutagenesis data performed by Dr. Karen Gregory.

Chapter V continues the investigation of how allosteric modulators bind to $mGlu_5$, but expands across four diverse chemical scaffolds. This chapter was written for the dissertation. Computational studies were interpreted in light of mutagenesis data performed by Dr. Karen Gregory and SAR data collected by Chrysa Malosh. Contributions from this work are being incorporated into a manuscript in preparation entitled “Investigating the binding modes of $mGlu_5$ allosteric modulators from diverse scaffolds” by

Elizabeth Nguyen, Karen Gregory, Chrysa Malosh, Jeffrey Mendenhall, Brittney Bates, Meredith Noetzel, Emma Squire, Eric Turner, Kyle Emmitte, Shaun Stauffer, Jason Manka, Ya Zhou, Mark Turlington, Andrew Felts, Craig Lindsley, Jeff Conn and Jens Meiler.

Chapter VI discusses the preparation of an $mGlu_5$ construct with a cysteine-less background for the collection of structural restraints that would aid in the interpretation of the computational modeling of $mGlu_5$. This chapter was written for the dissertation. The majority of the experiments presented in this chapter were performed by the author of this dissertation, with experimental work supplemented by Dr. Karen Gregory and Emma Squire as indicated in the text. Compound synthesis was performed by Shaun Stauffer, Aspen Chun, Ya Zhao and Craig Lindsley. This work will be incorporated into a future manuscript that applies the cys-less receptor to the collection of experimental structural restraints.

To overcome the limitations of comparative modeling, Chapter VII presents improvements upon the BioChemical Library membrane protein structure prediction method (BCL::MP-Fold) for *de novo* of membrane proteins by incorporating sequence-based residue exposure prediction. While this work has yet to be widely applied to GPCRs, it is a major step forwards in computational modeling for targets with no available structural templates as it allows for the generation of unique membrane protein topologies. This chapter was written for this dissertation, but will be prepared as a manuscript entitled “Incorporation of sequence-based exposure prediction in *de novo* membrane protein structure prediction” by Elizabeth Nguyen, Jeffrey Mendenhall, Brian Weiner, Kelly Gilmore and Jens Meiler.

Chapter VIII provides the major conclusions for this work and how it relates to the current findings in the field. This chapter was written for the dissertation.

The Appendix includes supplemental figures, detailed command lines, scripts and experiments not described in the manuscript for the work presented in Chapter I through Chapter VII. In addition, the Appendix includes an application of the GPCR comparative modeling and docking protocol to the Class A GPCR, $P2Y_{12}$. This section is based on a publication entitled “Identification of determinants required for agonistic and inverse agonistic ligand properties at the ADP receptor $P2Y_{12}$” by Philip Schmidt, Lars Ritscher, Elizabeth Dong, Thomas Hermsdof, Maxi Coester, Doreen Wittkopf, Jens Meiler and Torsten Schoeneberg. The author of this dissertation provided the computational modeling presented in light of experimental studies performed by Philip Schmidt and Lars Ritscher. The final section of the Appendix outlines the alignment method used in the BCL, which was evaluated and benchmarked by the author of the dissertation and is based on the publication entitled “BCL::Align-sequence alignment and fold recognition with a custom scoring function online” by Elizabeth Dong, Jarrod Smith, Sten Heinze, Nathan Alexander and Jens Meiler.

CHAPTER I

INTRODUCTION

Part of the work presented in this chapter was published in (Gregory, Dong, Meiler, & Conn, 2011)

G protein-coupled receptors as therapeutic targets

The diverse range of physiological function within the human body is made possible by the efficiency of communication that occurs between cells. This communication is driven by signals transmitted through molecules such as neurotransmitters, hormones or growth factors, as well as external stimuli such as light or odor. These signals are captured by G protein-coupled receptors (GCPRs), a family of proteins that sit within the cell membrane and trigger a downstream cascade of proteins necessary to carry out the specified function. At this moment, in fact, the rhodopsin receptor is allowing your eyes to process light as you read these words while GABA and glutamate receptors are allowing you to understand and remember these ideas. Angiotensin and adrenergic receptors are working to regulate your blood pressure as well as your urinary and digestive systems.

There are more than 800 GPCRs encoded by the human genome, which makes it the largest protein family in the human proteome (Stevens et al., 2012). GPCRs have been further categorized by sequence similarity by the International Union of Pharmacology (IUPHAR) into three main classes: class A rhodopsin-like, class B secretin-like and class C metabotropic glutamate-like (Foord et al., 2005). Class A is by far the largest with 700 members and contains receptors that bind opsins, olfactory molecules, opioids,

neurotransmitters and hormones (Katritch, Cherezov, & Stevens, 2013). Class B receptors primarily bind neuropeptides and other peptide hormones. Class C includes the metabotropic glutamate, $GABA_B$ and calcium-sensing receptors. The classification of the frizzled and smoothened receptors of Class F as GPCRs is still controversial. Although they have distinct signaling properties through the hedgehog pathway, members of Class F are still able to couple to G proteins (Chong Wang, Wu, et al., 2013). Within each class, receptors share over 25% sequence similarity, but there is less than 25% sequence similarity across classes (Jacoby, Bouhelal, Gerspacher, & Seuwen, 2006).

Despite their divergent sequences, however, the overall structure and function of all GPCRs is quite similar. GPCRs have seven transmembrane (TM) helices that are connected by flexible intra- and extra-cellular loop regions which often play an important role in ligand binding. When a ligand, whether it be a small molecule or peptide, binds the receptor, downstream signaling via a G protein catalyzes GDP-GTP exchange, which then further modulates downstream proteins. The conformational switch from an inactive to active state occurs when TM5 and TM6 undergo an outward “swinging” motion to expose the G protein binding site on the intracellular side of the receptor (Katritch et al., 2013).

Because of their key functional roles, GPCRs are a common drug target. In fact, about 30% of today’s drugs target GPCRs (Overington, Al-Lazikani, & Hopkins, 2006). Traditionally, drugs are developed to bind the receptor at the same site as the endogenous ligand, known as the orthosteric site. One of the key problems with such drugs, however, is that multiple GPCR subtypes that bind the same ligand are localized to different organ systems in the body. As a result, targeting a disease localized to a specific organ system is difficult to achieve with a drug that binds the orthosteric site. Carvedilol, for example, binds to a number of alpha- and beta-adrenergic receptors. While it is primarily used to treat heart failure and hypertension, common side effects are seen across multiple organ

systems including dizziness, fatigue, hyperglycemia, diarrhea, weight gain and muscular weakness (Frishman, 1998).

One way to enforce subtype selectivity and reduce side effects is to design drugs that bind in a site other than the endogenous ligand binding site, one that is unique to the receptor being targeted. This site is known as the allosteric binding site (Milligan & Smith, 2007). Ligands that bind to an allosteric site and enhance receptor activity are known as positive allosteric modulators, or PAMs. Ligands inhibiting receptor activity are known as negative allosteric modulators, or NAMs. Neutral allosteric ligands also exist, which do not modulate receptor activity. Allosteric modulators are more difficult to develop than orthosteric ligands because often, there is no known chemical scaffold to build from. However, there have been two GPCR allosteric modulators that have in fact been marketed for use: cincalcet targets the Calcium-sensing receptor to treat hyperparathyroidism (Lindberg et al., 2005) and maraviroc targets CCR5 to treat HIV infections (Dorr et al., 2005). The exciting potential for the discovery of other GPCR allosteric modulators has inspired the work presented in this dissertation.

Targeting the metabotropic glutamate receptor subtype 5 for cognitive function disorders

Arguably, one of the most well studied GPCR families with respect to allosteric modulation are the metabotropic glutamate receptors, or mGlus. The mGlus are sub-classified into class C GPCRs along with Calcium-sensing (CaSR), $GABA_B$, pheromone and taste receptors. Most class C GPCRs are distinguished by their large extracellular N-terminal domain, termed the Venus Flytrap domain (VFD), that contains the endogenous ligand binding site as shown in Figure 1 (Pin, Galvez, & Prézeau, 2003). Allosteric modulators therefore target the transmembrane helical region of the receptor. There are eight mGlu subtypes that are classified into three major groups based on sequence

homology, pharmacological properties, and coupling to different second-messenger pathways. Group I includes $mGlu_1$ and $mGlu_5$; group II, $mGlu_2$ and $mGlu_3$; and group III, $mGlu_4$, $mGlu_6$, $mGlu_7$ and $mGlu_8$. mGlus of the same group show ~70% sequence identity whereas between groups this percentage falls to ~45% (Conn & Pin, 1997). Group I mGlus preferentially couple to activation of the $G_{q/11}$ family of G proteins activating phosphoinositide hydrolysis as the major signaling mechanism. In contrast, group II and group III mGlus preferentially couple to $G_{i/o}$ and inhibition of adenylyl cyclases. Members of each group have a unique pharmacological profile and can be selectively activated by specific agonists or allosteric modulators that have no effects on members of the other groups (Gregory et al., 2011).

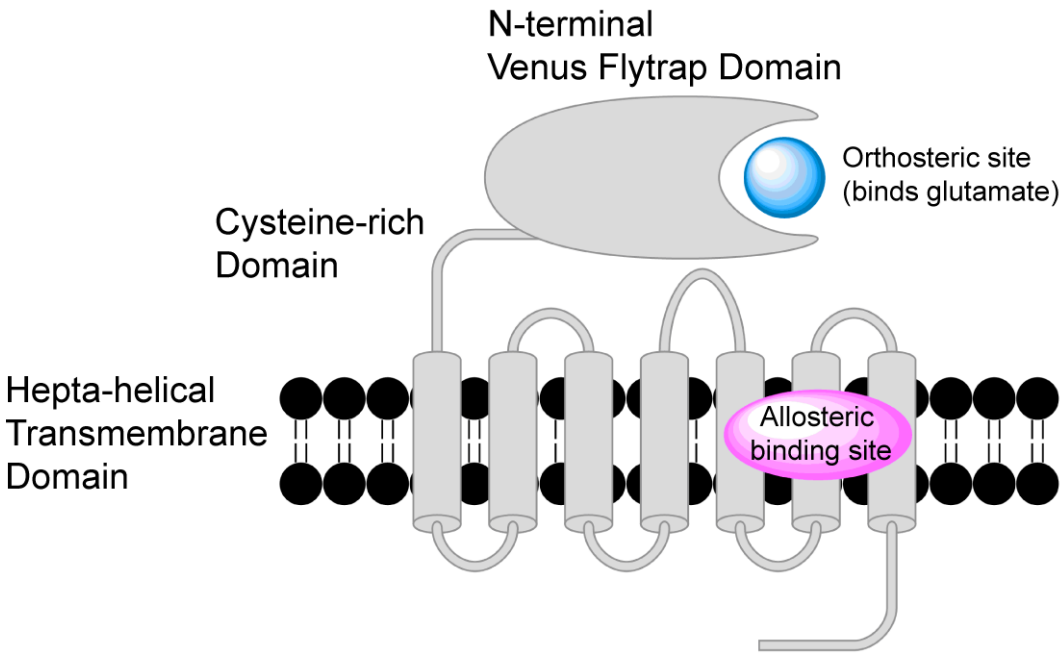


Figure 1
Schematic of the metabotropic glutamate receptor structure.
The large N-terminal extracellular domain where the endogenous ligand (glutamate) binds at the orthosteric site is also known as the Venus Flytrap Domain (VFD). The VFD is connected to the transmembrane region via a cysteine-rich domain. Allosteric ligands bind within the hepta-helical transmembrane domain.

The metabotropic glutamate receptor subtype 5 ($mGlu_5$) is involved in mammalian cognitive function through diverse signaling pathways that modulate synaptic plasticity (Hollmann & Heinemann, 1994). As a result, it was speculated that $mGlu_5$ involvement in learning and memory might make it a suitable target for cognitive function disorders such as Fragile X Syndrome and schizophrenia.

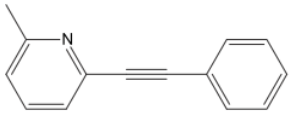
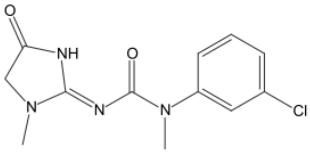
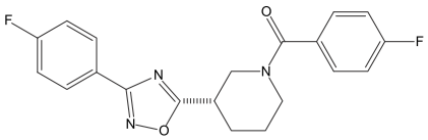
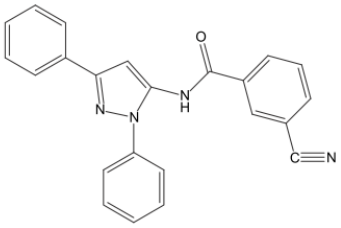
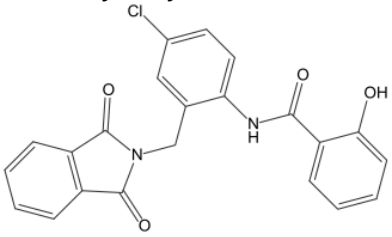
Fragile X syndrome (FXS), a disorder caused by a trinucleotide (CGG) repeat expansion in the FMR1 gene on the X chromosome, results in mental retardation, autism, seizures, anxiety and physical abnormalities (Hagerman et al., 2009). Studies with FMR1 knockout mice suggest that $mGlu_5$ signaling is increased in FXS, leading to an increase in long-term depression (LTD) of neuronal synapses in the hippocampus but no change or a decrease in long-term potentiation (LTP) (Dölen & Bear, 2008). These studies lead to the hypothesis that negative modulators of $mGlu_5$ could repair the balance between LTD and LTP in the hippocampus and reduce the cognitive defects seen in FXS. Indeed, it has been shown that administration of MPEP, an $mGlu_5$ NAM (Table 1), to FXS mice reduces anxiety and seizures (Yan, Rammal, Tranfaglia, & Bauchwitz, 2005).

Schizophrenia is a psychiatric illness characterized by positive symptoms (thought disorder, delusions, hallucinations, and paranoia), negative symptoms (social withdrawal, anhedonia, apathy, and paucity of speech) and cognitive impairments (Conn, Lindsley, & Jones, 2009). Current antipsychotics have little or no efficacy in treating the negative and cognitive symptoms associated with schizophrenia, but a strategy targeting $mGlu_5$ aims to relieve cognitive impairments as well as positive symptoms (Conn et al., 2009). Recent studies suggest that antagonists of the N-methyl-D-aspartate (NMDA) receptor, such as phencyclidine (PCP) and ketamine, produce the same cluster of symptoms seen in schizophrenic patients (Conn et al., 2009, Tsai & Coyle, 2002). Conversely, ligands that enhance NMDA receptor function have proven to be efficient in the treatment of schizophrenia (Lindsley et al., 2006). While direct agonists of NMDA receptors are

typically neurotoxic, the $mGlu_5$ receptor has been identified as a closely associated signaling partner with NMDA receptors in the forebrain, leading to the hypothesis that activators of $mGlu_5$ may lead to the amelioration of symptoms associated with schizophrenia (Conn et al., 2009). CPPHA (Table 1) is an example of a PAM of $mGlu_5$ which has no effect on glutamate affinity, but acts as a receptor agonist (O'Brien et al., 2004).

Several studies have now reported positive and negative allosteric modulators for $mGlu_5$ (see Table 1), providing an exciting and novel approach to $mGlu_5$-specfic therapeutic agents that can be used to improve cognitive function. These activators and inhibitors of $mGlu_5$ can subtly alter transmission in glutamatergic circuits in a therapeutically beneficial manner, making such compounds exciting potential treatments for a variety of CNS-related disorders. The majority of compounds that have been discovered, however, rarely enter preclinical studies due to low solubility, potency and selectivity. While high-throughput screening methods can identify more therapeutic candidates by testing hundreds of thousands of compounds per day, the cost and time of drug development would be greatly reduced if existing drugs could be modified to increase specificity and affinity. To carry out this structure-based drug development, further studies are necessary to understand exactly how these allosteric modulators bind $mGlu_5$ and which residues are critical for their interaction. This information will help develop derivatives with improved chemical properties to further support the therapeutic potential of such molecules.

Table 1
Therapeutic indications for mGlu5 in CNS disorders and examples of allosteric modulators.

Therapeutic indication	Intervention	Representative Allosteric Modulators	
		Compound	Chemical Name
Anxiety, Fragile X Syndrome, chronic pain, depression, migraine, Parkinson's disease levodopa-induced dyskinesia	NAM	MPEP (Spooren et al., 2000)	2-Methyl-6-(phenylethynyl)pyridine 
		Fenobam (Porter et al., 2005)	*N*-(3-chlorophenyl)-*N*'-(4,5-dihydro-1-methyl-4-oxo-1H-imidazole-2-yl)urea 
Schizophrenia, cognition disorders	PAM	ADX47273 (Schlumberger et al., 2009)	*S*-(4-fluoro-phenyl)-{3-[3-(4-fluoro-phenyl)-[1,2,4]oxadiazol-5-yl]-piperidin-1-yl}-methanone 
		CDPPB (Uslaner et al., 2009)	3-cyano-*N*-(1,3-diphenyl-1*H*-pyrazol-5-yl)benzamide 
		CPPHA (Zhao et al., 2007)	*N*-{4-chloro-2-[(1,3-dioxo-1,3-dihydro-2*H*-isoindol-2-yl)methyl]phenyl}-2-hydroxybenzamide 

Detecting the binding site for allosteric modulators

Unique experimental studies have been developed to determine where allosteric modulators bind $mGlu_5$. The first step is to confirm that a ligand is indeed binding to an allosteric site. One way of determining that a putative allosteric modulator is not binding the orthosteric site of an mGlu is through the use of chimeric receptors. Chimeric receptors are often constructed by exchanging the VFD of the receptor for which a modulator is selective with that of a different subtype that the ligand does not affect. If the ligand remains active at the chimeric construct, its activity must then be localized to the transmembrane region or C-terminal region, i.e. an allosteric site. Chimeric receptors were first used to determine agonist selectivity of orthosteric ligands (Takahashi, Tsuchida, Tanabe, Masu, & Nakanishi, 1993) and have now become increasingly useful for determining an allosteric mode of action. CPCCOEt was characterized as an allosteric modulator of $mGlu_1$ using a chimeric CaSR and $mGlu_1$ construct (Bräuner-Osborne, Jensen, & Krogsgaard-Larsen, 1999) as well as using chimeras with other mGlus (Gasparini et al., 2001, Litschig et al., 1999). This strategy has proved effective for characterization of numerous NAMs and PAMs, with chimeric receptor constructs often used as the first step in the validation of an allosteric mechanism (Carroll et al., 2001, Knoflach et al., 2001, Maj et al., 2003, Mitsukawa et al., 2005, Pagano et al., 2000).

Another method of confirming allosteric binding is to eliminate the extracellular VFD from the receptor altogether through the construction of a 'headless' mGlu. The headless receptor lacks the N-terminal extracellular VFD but retains an intact transmembrane region and a functional C terminus (Goudet et al., 2004). Headless mGlus behave like wild-type receptors in terms of G protein coupling and can be positively or negatively regulated by ligands, like any other class A GPCR. However, they no longer respond to orthosteric ligands. In the headless receptor, PAMs acts as agonists and NAMs become inverse agonists. Because these allosteric modulators retain activity in cells

expressing the headless receptor, these constructs are useful tools to identify allosteric ligands (Y. Chen et al., 2007). The headless construct of $mGlu_5$ has been used to localize the binding site of the $mGlu_5$ allosteric modulators MPEP, VU29 and CPPHA to the transmembrane domain (Y. Chen et al., 2007, Y. Chen, Goudet, Pin, & Conn, 2008). Chimeric and headless receptors constructs are useful for delineating the location of allosteric binding sites and investigating allosteric interactions, however, they do not provide detailed structural information.

Localization of an allosteric modulator's activity can be narrowed down further to functionally important residues and binding determinants using site-directed mutagenesis. Second messenger assays are commonly used in conjunction with mutagenesis to probe the binding site of GPCRs. The Ca^{2+} fluorescence assay is commonly used, which measures the release of Ca^{2+} stores from the endoplasmic reticulum due to activation of a $G_{\alpha q}$-coupled GPCR (Emkey & Rankl, 2009). This measurement is often carried out using a fluorometric imaging plate reader (FLIPR), which measures the intensity of Ca^{2+} sensitive dye that has been added to the cells. Compared to measurements done with the receptor alone, PAMs will demonstrate an increase in percent maximum fluorescence and shift the concentration response curve to the left while NAMs will do the opposite (Conn et al., 2009). This technology has allowed for rapid assessment of the functional effect of ligands binding $mGlu_5$.

While second messenger assays are useful for probing the functional effect of a mutation on the interaction between an orthosteric agonist and allosteric modulator, radioligand binding based studies can be used to quantify the influence of a mutation on the affinity of an allosteric modulator. Such efforts have been greatly facilitated by the development of radioligands for mGlu allosteric sites. The selective $mGlu_5$ radioligands [^{3}H]-M-MPEP (Gasparini, Kuhn, Pin, & others, 2002), [^{3}H]methoxy-PEPy (Cosford, Roppe, et al., 2003) and [^{3}H]-methoxymethyl-MTEP (Cosford, Roppe, et al., 2003), provide the

opportunity for the characterization of the MPEP binding site on $mGlu_5$. In addition, inhibition binding experiments can be used to determine if a novel allosteric modulator is competitive for known allosteric sites. A summary of mutational studies that have determined functionally critical residues and binding determinants of PAMs and NAMs of mGlu receptors is found in Table 2.

Table 2:
Functionally critical residues and binding determinants of PAMs and NAMs of group 1 mGlu receptors.

	Positive Allosteric Modulators (PAMs)						Negative Allosteric Modulators (NAMs)				
	$mGlu_1$		$mGlu_5$				$mGlu_1$			$mGlu_5$	
7TM position**	CPPHA	Ro 67-7476	CPPHA	VU-29	DFB	EM-TBPC	CFMMC	LY45606, YM298198 FTICD	CPCCOEt	MPEP	fenobam
TM1											
1.42	F599		F585								
TM3											
3.29										R648	R648
3.36		S668			P655					P655*	P655*
3.39		C671			S658					S658*	S658*
3.40						Y672				Y659*	Y659*
TM4											
4.45											
4.46											
4.55							I725				
EC2											
45.51						N747					
45.54					N734	N750					
TM5											
5.47		L757			L744	L757					L744*
5.48											
5.50							N760	N760†			
TM6											
6.43					T781					T781	T781*
6.47					W785	W798	W798	W798†		W785*	W785*
6.51					F788	F801*	F801	F801		F788*	F788*
6.55					Y792	Y805*	Y805	Y805		Y792	Y792*
TM7											
7.32					M802	T815*	T815	T815	T815	M802*	
7.35									A818	S805*	
7.40				A810	A810					A810*	A810*

Table compiled by Dr. Karen Gregory *Residues implicated in binding have been determined using selective allosteric radioligands
**The position of each residue in the mGlu 7TMD is given by the Ballesteros and Weinstein numbering system (Ballesteros & Weinstein, 1995), which allows for the comparison of equivalent positions within GPCRs. The first number represents the TM helix and the second number is its position relative to a highly conserved residue in the group A GPCRs from that TM, assigned the number 50.Highly conserved residues (assigned to position 50) are from the bovine rhodopsin sequence: N551.50, D832.50, R1353.50, W1614.50, C18745.50, P2155.50, P2676.50, P3037.50. Residues in extracellular loop 2 are labeled '45' to indicate location between helix 4 and 5.
† Effects LY456066 only

Structural analysis of G protein-coupled receptors

While mutagenesis studies allow for the identification of residues that are important to the function and binding of an allosteric modulator, it is difficult to know if these effects are due to a direct interaction between the ligand and receptor or if they are due to global changes in receptor conformation. There have been studies where a point mutation distal to a ligand binding site resulted in a dramatic change in ligand binding affinity (Baugh et al., 2010, El Omari, Liekens, Bird, Balzarini, & Stammers, 2006, Rod, Radkiewicz, & Brooks III, 2003). However, structural studies can reveal the direct interactions between a ligand bound to a receptor at an atomic level. In nuclear magnetic resonance (NMR) spectroscopy, resonance frequencies of distinct nuclei in a ligand-protein complex are measured in response to pulses from an external magnetic field. Multidimensional NMR experiments can be designed to determine ligand affinity and specificity, as well as identification of specific residues responding directly to ligand binding (Goldflam, Tarragó, Gairi, & Giralt, 2012). In electron paramagnetic resonance (EPR), the change in free electron spin of a ligand tagged with a methane-thio-sulfonate (MTSL) spin label bound to a cysteine can indicate direct interaction between the ligand and receptor (Loo & Clarke, 1997). X-ray crystallography is the most common approach used to determine a 3-dimensional structure of a ligand-receptor complex. After over-expression and solubilization of a receptor and addition of the ligand, a receptor-ligand complex can undergo crystallization under specific conditions. An X-ray diffraction pattern of the resulting crystallized complex then yields a model of the receptor and ligand in real space after Fourier transform (McPherson, 1991).

Crystal structures of the N-terminal domains of $mGlu_1$, $mGlu_3$ and $mGlu_7$ have provided extensive information about how the endogenous ligand, glutamate, binds $mGlu_5$. The N-terminal VFD of the mGlus is made up of two lobes (Kunishima et al., 2000, Muto, Tsuchiya, Morikawa, & Jingami, 2007, Tsuchiya, Kunishima, Kamiya, Jingami, &

Morikawa, 2002). This forms a clam shell-shaped structure, with the glutamate binding site residing between the two lobes, as seen in Figure 1. Evidence suggests that the mGlus dimerize via interactions between their VFDs. When glutamate binds, the globular domains close into a stable conformation with glutamate inside (Bessis et al., 2000, Kunishima et al., 2000, Tsuchiya et al., 2002). The conformational changes induced by glutamate binding at the VFD are transmitted via a cysteine-rich domain. The cysteine-rich domain, unique to class C GPCRs (with the exception of the $GABA_B$ receptor which does not have one), links the VFD to the transmembrane-spanning α-helices by a conserved disulfide bridge (see Figure 1), subsequently promoting coupling to intracellular G proteins and activation of second messenger pathways (X. Liu et al., 2004, Muto et al., 2007, Rondard et al., 2006).

To understand the interaction between allosteric ligands and $mGlu_5$, however, it is important to obtain structural information about the hepta-helical transmembrane domain in which they bind. This has proven to be very difficult. The first structural determination of the seven transmembrane-helical region of a GPCR was bovine rhodopsin in 2000 (Palczewski et al., 2000). Because they are large in size, have extremely flexible loop regions and are very difficult to solubilize as membrane proteins, more experimentally determined GPCR structures were slow to follow. With advances in protein engineering and x-ray crystallography, however, the last six years have seen an exponential growth in the number of available GPCR X-ray crystal structures. As of June 2013, there have been 17 unique GPCR X-ray crystal structures determined (Katritch et al., 2013), as seen in Table 3. All the experimentally determined structures of GPCRs have been from Class A, with the exception of the latest crystal structure of Smoothened from Class F.

Most GPCR structures have also been crystallized with either an agonist or antagonist small-molecule or peptide bound within the transmembrane helices. Agonist-bound experimental structures of rhodopsin, A2Ar and beta-adrenergic receptor have

identified a trend of a ligand-mediated conformational change involving TM3 and TM6 during activation (Katritch et al., 2013). However, the specific residues critical to ligand activation vary across each receptor. A receptor-specific model, then, is necessary for ligand-based studies of a particular complex. In the absence of an experimentally determined structure, as is the case with $mGlu_5$, computational modeling has been shown to provide substantial insight into ligand-protein interactions (Congreve, Murray, & Blundell, 2005).

Table 3:
Distinct G protein-coupled receptor experimental structures deposited in the Protein Data Bank as of June 2013.
Details of the highest resolution structure for each unique receptor is reported.

Protein name	PDBID	Resolution (Å)	Ligand
Rhodopsin (bRh) (Okada et al., 2004)	1U19	2.20	Retinal (inverse agonist)
Beta2-Adrenergic (B2Ar) (Cherezov et al., 2007)	2RH1	2.40	Carazolol (partial inverse agonist)
A2A adenosine (A2Ar) (Jaakola et al., 2008)	3EML	2.60	ZM241385 (antagonist)
CXCR4 chemokine (CXCR4) (B. Wu et al., 2010)	3ODU	2.50	IT1t (antagonist)
Dopamine D3 (D3R) (Chien et al., 2010)	3PBL	2.89	Eticlopride (antagonist)
Histamine H1 (H1R) (de Graaf et al., 2011)	3RZE	3.10	Doxepin (antagonist)
M2 muscarinic acetylcholine (M2R) (Haga et al., 2012)	3UON	3.00	3-quinuclidinyl-benzilate (antagonist)
S1P1 sphingosine 1-phospate (S1P1R) (Hanson et al., 2012)	3V2Y	2.80	ML056 (antagonist)
Beta1-Adrenergic (B1Ar) (Warne, Edwards, Leslie, & Tate, 2012)	4AMJ	2.30	Carvedilol (inverse agonist)
M3 muscarinic acetylcholine (M3R) (Kruse et al., 2012)	4DAJ	3.40	Tiotropium (inverse agonist)
Kappa-opioid (KOR) (H. Wu et al., 2012)	4DJH	2.90	JDTic (antagonist)
Mu-opioid (MOR) (Manglik et al., 2012)	4DKL	2.80	β-FNA (antagonist)
N/OFQ opioid (NOP) (A. A. Thompson et al., 2012)	4EA3	3.01	C-24 (antagonist)
Delta-opioid (DOR) (Granier et al., 2012)	4EJ4	3.40	Naltrindole (antagonist)
5-hydroxytryptamine 1B (5HT1B) (Chong Wang, Jiang, et al., 2013)	4IAR	2.70	Ergotamine (agonist)
5-hydroxytryptamine 2B (5HT2B) (Wacker et al., 2013)	4IB4	2.70	Ergotamine (agonist)
Smoothened (SMO) (Chong Wang, Wu, et al., 2013)	4JKV	2.45	LY2940680 (antagonist)

Current computational methods for G protein-coupled receptor structure prediction

In his work with ribonuclease A leading up to the 1972 Nobel Prize in chemistry, Christian Anfisen postulated that the native structure of a protein was unique, stable and a kinetically accessible minimum of the free energy. In addition, he hypothesized that a protein's native structure is determined only by its amino acid sequence (Anfinsen, 1973). Not long after Anfisen's discovery, computational methods began to develop which sought to model the structure prediction process and its structural outcomes *in silico* (Moult, 2006). It has proven to be a difficult challenge, however, as predicted by Cyrus Levinthal in 1969. In his thought experiment, Levinthal illustrated that for a 100 residue protein, where each residue has just 10 different conformations, the number of possible protein conformations is 10^{100}. If a protein samples all of these possible conformations, even at just picoseconds each, it would still take longer than the age of the universe to fold this single protein (Levinthal, 1969).

At over 300 residues and with the extra complication of being embedded in a membrane, modeling the 7TM helical region of GPCRs is a particularly difficult challenge. This work addresses these challenges through comparative modeling and *de novo* protein structure prediction with the use of membrane-specific scoring functions. Development and application of two specific protein structure prediction software suites were involved in these studies. The unified Rosetta software package performs *de novo* protein structure prediction, ligand docking, and structure prediction of biological macromolecules and macromolecular complexes (Das et al., 2009, Kuhlman et al., 2003, Meiler & Baker, 2006, Rohl, Strauss, Chivian, & Baker, 2004). The BioChemistry Library Project (BCL::Commons) was designed to perform *de novo* protein structure prediction with sparse experimental restraints and also has a large toolkit of cheminformatic and machine learning methods (Dong, Smith, Heinze, Alexander, & Meiler, 2008, Karakas et al., 2012,

Sliwoski, Lowe, Butkiewicz, & Meiler, 2012, Woetzel et al., 2012, Woetzel, Lindert, Stewart, & Meiler, 2011).

Comparative modeling of a target protein relies on the assumption that there exists a structurally similar protein that can be used as a template. Despite the low sequence identity between the different classes of GPCRs, confirmation of a common hepta-helical architecture in the mGlu transmembrane region provides support for the use of class A templates as a starting point for comparative modeling (Bhave et al., 2003). Once a structural template has been identified, its coordinates are used as a scaffold for the highly conserved secondary structure regions of the target protein. Variable loop regions are then rebuilt and the protein model can then undergo full energy minimization (D. Baker & Sali, 2001). In Rosetta, loop regions are rebuilt using a cyclic coordinate descent algorithm (Canutescu & Dunbrack, 2009, Chu Wang, Bradley, & Baker, 2007). For residues in the loop regions, ϕ-ψ angles of backbone segments of homologous sequence amino acid fragments from the Protein Data Bank (PDB) are introduced. After the fragment substitution, slight changes in the ϕ-ψ angles are considered to close breaks in the protein chain. The resulting full sequence models are subjected to side chain repacking and gradient minimization of ϕ, ψ, and χ angles until the Rosetta energy score converges to a minimum value. The high-resolution energy function includes knowledge-based terms representing solvation, electrostatic interactions, van der Waals attraction/repulsion, and hydrogen bonding terms (Qian et al., 2007). Full details regarding the comparative modeling process is the focus of Chapters II and III.

It is possible that the 7TM region of the Class C $mGlu_5$ has a modified structure compared to the Class A templates. The recent experimentally determined structure of Smoothened, a Class F GPCR with less than 10% sequence similarity to Class A GPCRs, has many of its own unique features including an extended TM6, an inward-facing tip of

TM5 and lack of helical kinks in TM 5, 6 and 7 as seen in Figure 2 (Chong Wang, Wu, et al., 2013). In such cases where a unique fold is predicted and no template is available, *de novo* protein structure prediction from the primary sequence can be employed. Rosetta builds up a protein model using three and nine residue fragments from the Protein Data Bank that are homologous to the sequence of the target protein (Bradley, Misura, & Baker, 2005, Simons et al., 1999, Simons, Kooperberg, Huang, & Baker, 1997). Rosetta folds a protein from a continuous chain, which becomes computationally limiting for proteins larger than 100 residues due to the exponential growth of non-local contacts that are necessary to sample. BCL::Fold is able to *de novo* fold large proteins and sample unique protein topologies by breaking apart the chain into distinct secondary structure elements (SSE) before sampling (Karakas et al., 2012, Weiner, Woetzel, Karakaş, Alexander, & Meiler, 2013). Further discussion of BCL::Fold is discussed in Chapter VII.

Both comparative modeling and *de novo* structure prediction have been extensively optimized for soluble proteins, but only recently have methods been modified to model membrane proteins. Recently, an implicit membrane potential was added to Rosetta that models the membrane as two parallel planes separated by 60 Å (Yarov-Yarovoy, Schonbrun, & Baker, 2005). The Rosetta Membrane energy function rewards the exposure of hydrophobic residues to the membrane region, dense packing of helices within the membrane and placement of helices perpendicular to membrane. The energy function penalizes such moves as allowing loops to dip into the membrane. In BCL::Fold, the scoring function is also adjusted for membrane proteins by first simulating an apolar membrane environment and then rewarding preferred SSE orientations and agreement of amino acid placement in the membrane based on sequence prediction (Weiner et al., 2013). These membrane-specific scoring functions have made it possible to model GPCRs using existing computational methods.

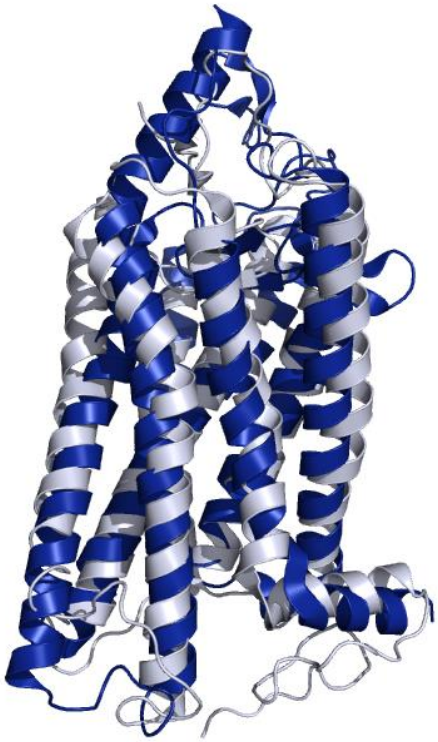


Figure 2
Comparing the structures of GPCRs from Class A and Class F.
Experimentally determined structures from X-ray crystallography for bovine rhodopsin from Class A in grey (PDB ID: 1U19) and Smoothened from Class F in blue (PDB ID: 4JKV) are overlaid for comparison. Smoothened, a Class F GPCR with less than 10% sequence similarity to Class A GPCRs, has many of its own unique features including an extended TM6, an inward-facing tip of TM5 and lack of helical kinks in TM 5, 6 and 7.

Computational methods for small-molecule docking

The value of a GPCR model is extended further when used in combination with ligand docking algorithms to predict binding modes. Such ligand docking studies can be used to identify critical residues for binding and receptor modulation which are valuable in the optimization of existing ligands as therapeutic agents. Docking into comparative models have even been shown to obtain hit rates during virtual screening that exceeded those using X-ray crystal structures (Tang, Wang, Hsieh, & Tropsha, 2012). Ligand docking is best accomplished as an iterative process, occurring concurrently with experimental studies that validate the residues in the ligand binding site. RosettaLigand is

a module of the Rosetta software suite that focuses on the interactions between protein and small molecules (Davis & Baker, 2009, Lemmon & Meiler, 2012, Meiler & Baker, 2006). Both protein and ligand flexibility is considered, as well as multiple ligand conformations. After a rigid body orientation of the ligand is performed in the initial binding site, a Monte Carlo sampling of rotamers for both the ligand and receptor side chains in the binding site are explored until an energy minimum is reached. The scoring terms in the Rosetta energy function are primarily knowledge-based and include a Lennard-Jones potential, solvation potential, inter-atomic electrostatic interactions, long and short-range hydrogen bind potentials and preference towards Dunbrack rotamer and Ramachandran bond angles (Meiler & Baker, 2006). A variety of other ligand docking algorithms exist. FlexX uses geometric hashing for ligand placement and scores binding modes using chemical descriptor models (Rarey, Kramer, Lengauer, Klebe, & others, 1996). Glide uses systematic sampling (Friesner et al., 2004) while GOLD uses genetic algorithms (Verdonk, Cole, Hartshorn, Murray, & Taylor, 2003). However, Rosetta is one of the few programs that integrate the ability to construct protein models and perform ligand docking with both ligand and receptor flexibility in a single framework with an integrated energy function.

Predicting interactions between G protein-coupled receptors and small molecules by combining experimental data with computational predictions

The advantage of computationally modeling of protein-ligand interactions is its speed, cost-efficiency and most importantly, its ability to sample accurate results (Schneider & Fechner, 2005). However, determining the most accurate model from the thousands of energy-minimized models that are created is not possible without validation from experimental data. Computational modeling is most valuable when performed concurrently with experimental studies that validate structural models and ligand binding conformations (D. Baker & Sali, 2001). Such models are then able to propose mechanisms

for protein-ligand interactions which are valuable in the optimization of existing drugs towards more effective therapeutic agents (Schneider & Fechner, 2005).

Predictions from computational models of $mGlu_5$ have already been verified by experimental results throughout literature. The location of the allosteric site in $mGlu_5$ was first predicted using molecular modeling of the receptor with rhodopsin. Residues initially predicted to be within the allosteric binding site were chosen based on where cis-retinal, an inverse agonist, binds in bovine rhodopsin (Palczewski et al., 2000). Several residues critical for cis-rhodopsin binding were verified to also be important for binding of mGlu allosteric ligands by several experimental studies, validating the computational model (Malherbe, Kratochwil, Knoflach, et al., 2003, Malherbe, Kratochwil, Zenner, et al., 2003).

Further probing of the allosteric site through site-directed mutagenesis identified several more key binding determinants for allosteric modulation by MPEP and fenobam (Table 2). These residues have been mapped onto a comparative model at $mGlu_5$ and were found to be on TMs 3, 6 and 7, as seen in Figure 3 (Malherbe, Kratochwil, Zenner, et al., 2003, Malherbe et al., 2006, Pagano et al., 2000). Similarly, functionally important residues for positive allosteric modulation by CDPPB and VU29 show a similar distribution (Y. Chen et al., 2008, Mühlemann et al., 2006). CDPPB was shown to inhibit allosteric binding of the MPEP analog [3H] methoxyPEPy in a competitive manner (Y. Chen et al., 2007). In addition, point mutations reducing the binding of MPEP also decrease the ability of CDPPB to potentiate $mGlu_5$ response to glutamate (Y. Chen et al., 2007). These results suggests that CDPPB, along with its derivative VU29, share a common or overlapping binding site with MPEP (Y. Chen et al., 2007, Kinney et al., 2005). Figure 3C demonstrates the clustering of these residues on the top half of the TMs, located on the inside face of the helices.

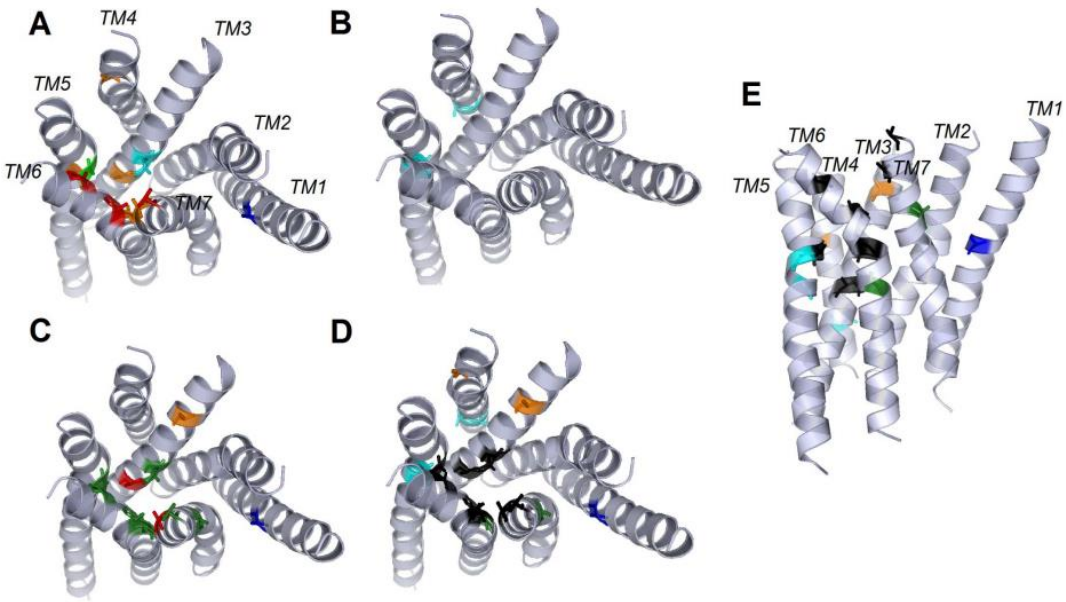


Figure 3
Functionally critical residues and binding determinants for mGlu allosteric modulators mapped onto the hepta-helical transmembrane domain of β2-adrenergic receptor X-ray crystal structure (2RH1) backbone.
The residues reported in Table 2 are shown here, mapped onto a model of mGlus based on the β2-adrenergic receptor. Shown in sticks is the backbone beta-carbon of functionally important residues & binding determinants. A) $mGlu_1$, B) $mGlu_2$, C) $mGlu_5$, D) residues from $mGlu_1$, $mGlu_2$ and $mGlu_5$ and E) side view of residues from $mGlu_1$, $mGlu_2$ and $mGlu_5$. Highlighted are: residues functionally important for CPPHA (blue) or other PAMs (cyan), residues important for NAMs functionally (orange) and through binding (red), and residues important for both PAMs and NAMs functionally (light green) or both PAM function and NAM binding (dark green). In D and E, residues important for both mGlu1 and $mGlu_5$ allosteric modulation are in black. The sequence alignment used to determine the placement of residues is found in the Appendix, Figure 42.

As these studies show, comparative models have been successfully utilized to generate hypotheses for mutagenesis-based studies and allow for three-dimensional visualization of these regions of the receptor. These models have inspired the first rounds of experimental studies, which have begun to collect detailed information about the binding site of various allosteric modulators (Table 2). This data, together with the growing number of available GPCR structural templates, can lead to information-rich comparative models for the transmembrane region of a class C GPCR developed in the absence of a crystal structure. The studies presented in the remainder of this work demonstrate the strength of the integrating computational structure prediction of GPCRs with experimental studies.

This work is will aid in the identification of novel chemotypes and optimization of allosteric modulators to improve their specificity and potency for use as therapeutic agents.

CHAPTER II

SMALL-MOLECULE LIGAND DOCKING INTO COMPARATIVE MODELS WITH ROSETTA

This work is based on publication (Combs, DeLuca, S.L., DeLuca, S.H., Lemmon, Nannemann, Nguyen, Willis, Sheehan & Meiler, 2013).

Summary

Structure-based drug design is frequently used to accelerate the development of small-molecule therapeutics. While substantial progress has been made in X-ray crystallography and nuclear magnetic resonance (NMR) spectroscopy, the availability of high-resolution structures is limited due to the frequent inability to crystallize or obtain sufficient NMR restraints for large or flexible proteins. Computational methods can be used to both predict unknown protein structures and model ligand interactions when experimental data is unavailable. This paper describes a comprehensive and detailed protocol using the Rosetta modeling suite to dock small-molecule ligands into comparative models. In the protocol presented here, the comparative modeling process, including sequence alignment, threading, and loop building, is reviewed. In addition, criteria that can improve the ligand docking results are discussed. Next, docking a small-molecule ligand into the protein comparative model is covered. Finally, and importantly, a strategy for assessing model quality is presented. The entire protocol is presented on a single example selected solely for didactic purposes. The results are therefore not representative and do not replace benchmarks published elsewhere. A tutorial is also provided in the Appendix so that the user can gain hands-on experience in using Rosetta. The protocol is

anticipated to take 5-7 hours, with additional time allocated for computer generation of models.

Introduction

The Rosetta software suite focuses on computational modeling and analysis of protein structures and is free for non-commercial users. It has enabled notable scientific advances in computational biology, including *de novo* protein design, enzyme design, ligand docking, and structure prediction of biological macromolecules and macromolecular complexes (Das & Baker, 2008, Davis & Baker, 2009, Davis, Raha, Head, & Baker, 2009, Kuhlman et al., 2003, Misura, Chivian, Rohl, Kim, & Baker, 2006, Rohl, Strauss, Chivian, et al., 2004, Rohl, Strauss, Misura, Baker, & others, 2004, Siegel et al., 2010, Yarov-Yarovoy et al., 2005). The broad spectrum of applications available through Rosetta allows for multiple computational problems to be addressed in one software framework. In this protocol, we discuss how Rosetta can be used to create a comparative model of a protein and extend this application by introducing ligand docking with comparative models. Ligand docking into comparative models (Davis & Baker, 2009, Davis et al., 2009, Misura et al., 2006) is a common technique used in structure-based drug design and provides an excellent introduction to the Rosetta software suite (Kaufmann, Lemmon, DeLuca, Sheehan, & Meiler, 2010). The protocol, which is outlined in Figure 4, is generalizable and will extend to a majority of protein-ligand systems. To aid in the understanding of Rosetta-specific language, a glossary has been provided in the Appendix.

Small-molecule docking into comparative models allows for structure-based drug design and hypothesis generation for protein/ligand systems for which there is no high-resolution structure. In such cases, there is frequently a homologous structure that has been structurally characterized at sufficient resolution for ligand docking and which can be utilized as a template for comparative modeling of the target protein. Several studies have

proven the utility in ligand docking into comparative models. Here, we provide a generalized protocol and tips for improving the success of ligand docking into comparative models. Furthermore, we use a single piece of software for comparative modeling and ligand docking. Therefore, the scoring function used remains the same throughout the process, and minimal processing is necessary between steps.

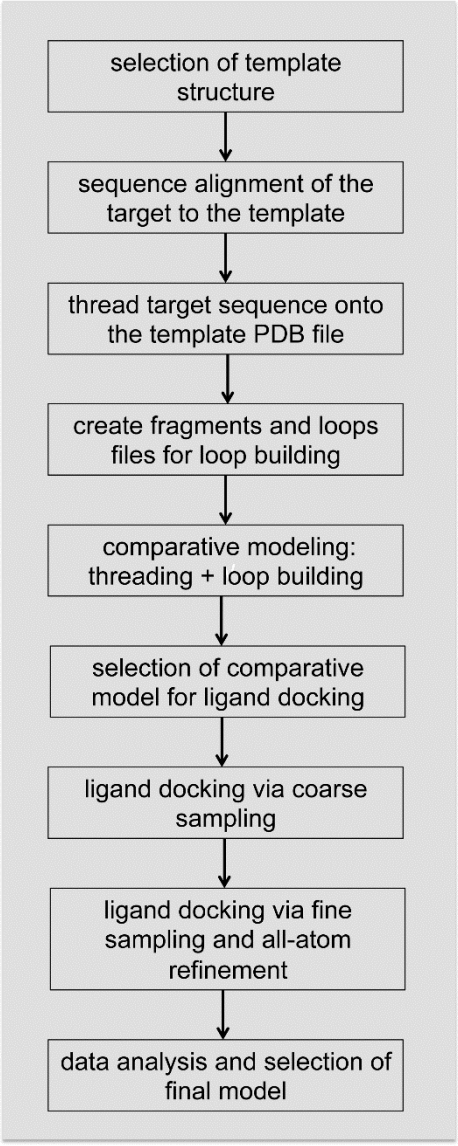


Figure 4
Outline of Rosetta modeling protocol.
This flowchart summarizes the complete protocol for docking small-molecule ligands into comparative models using Rosetta 3.4.

Comparative modeling with Rosetta

One of the most common applications of Rosetta is protein structure prediction via *de novo* structure prediction and comparative modeling (Kaufmann et al., 2010, Rohl, Strauss, Chivian, et al., 2004). When only the primary sequence of a protein is known, *de novo* structure prediction can sometimes be used to predict the protein's tertiary structure. However, to date, Rosetta has been shown to successfully fold only small, soluble proteins (fewer than 150 amino acids) and performs best if the proteins are mainly composed of secondary structural elements (alpha-helices and beta-strands) (Meiler & Baker, 2003). Helical membrane proteins between 51-145 residues were predicted within 4Å of the native structure (Yarov-Yarovoy et al., 2005). Accurate prediction of larger and/or more complex proteins can be achieved with the addition of experimental data, such as NMR chemical shifts and distance data (Lange et al., 2012, Lange & Baker, 2012, Rohl, 2005). Further, only sequences of very small proteins (up to 80 residues) have been predicted to atomic-detail accuracy (Bradley, Malmström, et al., 2005, Bradley, Misura, et al., 2005, Das et al., 2007). Therefore, whenever an experimental structure of a related protein is available, comparative modeling is the method of choice.

Comparative modeling refers to the elucidation of the tertiary fold of a protein, guided by the known structure of another, often homologous, protein. The unknown structure is commonly called the "target," while the protein of known structure, upon which the primary sequence of the target is threaded, is termed the "template." The known template structure reduces the conformational search space by providing a protein backbone scaffold; areas where the template and target sequences diverge significantly are typically remodeled and refined via the loop building application. Although the application is known as "loop building," a "loop" is defined here as any area where the backbone is to be rebuilt *de novo*, which most often occurs in flexible regions but can also

include secondary structural elements. Comparative models have played a major role in aiding experimental design and the interpretation of experimental results. They can be employed to help predict structure-function relationships (Kaufmann et al., 2009), predict binding pockets for ligands during structure-based drug design (Lees-Miller et al., 2009), and aid in the determination of target residues for site-directed mutagenesis (Fortenberry et al., 2011, Keeble et al., 2008).

In addition to Rosetta, Modeller (Eswar, Eramian, Webb, Shen, & Sali, 2008) is often used to generate comparative models. Comparative modeling with Modeller is highly automated and, as with Rosetta, works best for cases in which the sequence identity between the target sequence and the template structure is 30% or greater. It works by optimizing the comparative model's satisfaction of spatial restraints derived from one or multiple templates. Comparative modeling in Rosetta (Misura et al., 2006) is a multiple-step process that requires more input from the user; specifically, user-defined alignment and loop definitions are taken into account throughout the process. These definitions can be provided to Modeller but are not necessary for the program to generate a model.

Ligand docking with RosettaLigand and comparison to other ligand docking software

After a comparative model of the target protein has been constructed, computational ligand docking can be performed. Small-molecule ligand docking applications attempt to predict the protein/small-molecule binding free-energy, as well as critical binding interactions (Perola, Walters, & Charifson, 2004). These predictions can provide structural information of a ligand binding site (Davis et al., 2009), filter high-throughput screening libraries for likely hits (Ballester, Westwood, Laurieri, Sim, & Richards, 2010, Carlsson et al., 2011), or guide *de novo* drug design (Schneider et al., 2009, Schneider & Fechner, 2005). The protocol presented here details the process for

small-molecule ligand docking and focuses on locating critical residues for binding a specific ligand.

RosettaLigand requires input structures of a receptor (protein) and a ligand (small molecule) (Davis & Baker, 2009, Meiler & Baker, 2006). Because it does not perform binding pocket detection, the user must have prior knowledge of the location of the binding site. Other programs, such as SURFNET (Laskowski, 1995), LIGSITE (Huang & Schroeder, 2006), and PocketDepth (Kalidas, Chandra, & others, 2008), can be used to identify the ligand binding site before using RosettaLigand for small-molecule docking. Ligand and receptor side chain conformations are explored through Monte Carlo sampling of rotamers (Dunbrack Jr & Karplus, 1993). Predicted protein-ligand interactions are deemed favorable and are accepted if they improve the Rosetta energy score, described below (Davis & Baker, 2009). Backbone flexibility of the protein is modeled via a gradient-based minimization of phi and psi torsion angles (Z. Li & Scheraga, 1987). Performing ligand docking on an ensemble of ligand conformations and protein backbones can be used to increase the conformational space sampled if the protein-ligand interaction does not fit the simple lock-and-key paradigm (Siegel et al., 2010). Previously, Davis, *et al*. assessed RosettaLigand's accuracy via both retrospective and prospective benchmark studies (Davis & Baker, 2009). In 54 of 85 cases (64%), RosettaLigand's top scoring pose was within 2.0Å root mean square distance (RMSD) from the native pose. These results include backbone and side chain flexibility, as well as ligand flexibility, through conformer selection and torsion angle adjustments.

Ligand docking algorithms can be categorized based on their scoring function and search methodology. RosettaLigand uses a knowledge-based scoring function derived from statistical analysis of the Protein Data Bank (PDB) (Simons et al., 1997). The conformational search of the binding site is accomplished using a Metropolis Monte Carlo algorithm (Davis & Baker, 2009, Davis et al., 2009, Kuhlman et al., 2003, Metropolis,

Rosenbluth, Rosenbluth, Teller, & Teller, 1953, Misura et al., 2006, Rohl, Strauss, Misura, et al., 2004, Siegel et al., 2010). There are a wide range of search methodologies and scoring functions used by different ligand docking applications. Search strategies include geometric hashing (FlexX) (Rarey et al., 1996), genetic algorithms (GOLD) (Verdonk et al., 2003), and systematic sampling (Glide) (Friesner et al., 2004). Different scoring functions include physics-based force fields (Dock) (Ewing, Makino, Skillman, & Kuntz, 2001), chemical descriptor models (FlexX) (Rarey et al., 1996), and knowledge-based potentials (RosettaLigand (Davis & Baker, 2009, Meiler & Baker, 2006), DrugScore (Gohlke, Hendlich, Klebe, & others, 2000)).

A 2009 study compared the performance of the RosettaLigand docking method to nine other commonly used ligand docking programs (Dock, Dockit, FlexX, Flo, Fred, Glide, GOLD, LigandFit, MOE, and MVP) (Davis et al., 2009). Ligand docking algorithm performance was compared using a benchmark set of 136 ligands and eight target receptors provided by Glaxo-Smith-Kline. This study demonstrated that RosettaLigand performance was similar to or better than other ligand docking algorithms. This study used crystallographic protein structures as input rather than homology models. Kaufmann, *et al*. demonstrated the predictive power of Rosetta ligand docking into Rosetta-built comparative models (Kaufmann et al., 2009). In another study, RosettaLigand and AutoDock 4.0 were used to dock twenty protein-ligand complexes (Davis & Baker, 2009). In ten cases, RosettaLigand's flexible backbone docking protocol found top-scoring models under 2.0Å RMSD. In contrast, AutoDock identified only four such structures. However, the authors note that RosettaLigand consumed significantly more computational resources (40-80 CPU hours per input) than AutoDock (5-22 CPU hours per input).

Applying the comparative modeling and ligand docking protocols to a single problem

To illustrate the entire comparative modeling and ligand docking protocol on a single example, including a detailed analysis, we selected a target protein that has been co-crystallized with a small-molecule ligand and where an experimental structure of a distantly related homolog is available to serve as a template. We also selected a relatively small protein and ligand to facilitate rapid reproduction of the protocols by the reader. Specifically, T4-lysozyme in complex with 1-methylpyrrole (PDB ID: 2ou0) (Mobley et al., 2007) was chosen as the target and P22-lysozyme (PDB ID: 2anv) (Mooers & Matthews, 2006) as the template. Note that this selection was made with the above-mentioned didactic priorities in mind and not to find an optimal system to benchmark the accuracy of Rosetta. Throughout the chapter, we will refer to dedicated benchmark papers relevant for the individual steps to serve as references for expected Rosetta performance. Additionally, Kaufmann and Meiler recently performed a benchmark of ligand docking into comparative models with Rosetta, to which the reader is encouraged to refer for further information concerning RosettaLigand's performance for ligand docking into comparative models (Kaufmann & Meiler, 2012).

Usage of experimental restraints during Rosetta modeling and analysis

Incorporation of experimental data into structure prediction and analysis has been shown to improve the quality of the final model or ensemble of models (Alexander, Al-Mestarihi, Bortolus, Mchaourab, & Meiler, 2008, Hirst, Alexander, Mchaourab, & Meiler, 2011, Meiler, Baker, & others, 2005, Rohl, 2005). Numerous types of experimental data have been incorporated into such protocols, including electron density from X-ray crystallography (DiMaio et al., 2011) and electron microscopy, NMR distance and orientation data (Meiler et al., 2005, Shen et al., 2010), EPR distance data (Alexander et al., 2008, Hirst et al., 2011),crosslinking restraints (Herzog et al., 2012), small angle X-ray

scattering data (Grishaev, Guo, Irving, & Bax, 2010), and deuterium exchange mass spectrometry data (Pandit et al., 2012). While these types of data are more often applied to *de novo* protein structure elucidation, they can also be of some utility in the building of loops (Rathmann et al., 2012), reorientation of domains during comparative modeling, or identification of residues involved in ligand binding.

Caveats and challenges

As with all computational techniques, there are caveats associated with using Rosetta for comparative modeling and ligand docking. While comparative modeling can be used to model large proteins more reliably than *de novo* structure prediction methods, it is limited by the availability of high-quality structural templates in the PDB. Finite computational resources can also limit the size of conformational spaces that can be searched (A. R. Leach, Shoichet, & Peishoff, 2006). The comparative modeling and ligand docking processes discussed in this protocol allow for protein backbone movement; however, these models represent only static structures of local energy minima. For consideration of dynamics, conformational changes, and large-scale changes due to induced-fit or conformational flexibility during ligand docking, molecular dynamics simulations have been shown to be useful (A. R. Leach et al., 2006).

Despite these limitations, Rosetta has been employed to produce protein models that have proven invaluable where no experimentally determined protein structure exists (S. Combs, Kaufmann, Field, Blakely, & Meiler, 2010, Lees-Miller et al., 2009, Smith, Vanoye, George, Meiler, & Sanders, 2007). The tutorial presented in the Appendix, which uses T4-lysozyme as a simple example, provides a generalized workflow for comparative modeling and ligand docking in the Rosetta framework and demonstrates its ability to model accurate protein structures.

Availability

Rosetta is available through software licenses processed by the RosettaCommons (www.rosettacommons.org). Licenses for academia and non-profit institutions are free of charge. The Rosetta software suite can be installed on a Linux or OSX operating system (see Appendix). This setup allows other researchers to adopt the described protocol for their biological system of interest.

Experimental design

An overview of the entire protocol is summarized as a flowchart in Figure 4. The protocol matches that of the tutorial, which is provided in the Appendix.

Template selection

In Rosetta, construction of a comparative model for a desired target protein can be divided into distinct steps. First, an experimentally determined structure (template) must be identified. The quality of a comparative model is heavily dependent on the experimentally determined structure that is chosen as a template for the final model. If a low-quality, low-resolution template structure is chosen, the resulting models will also be low-quality. The following discussion provides insight into the process of identifying a proper template for comparative modeling.

A template can be located by BLAST (www.ncbi.nlm.nih.gov/BLAST/), which searches the PDB for proteins with high sequence identity. When arriving at the BLAST server online, use “protein blast,” and under “Database,” choose to search the “Protein Data Bank (PDB),” which contains all experimentally determined protein structures. A modified version of BLAST, PSI-BLAST allows for the identification of distant members of a protein family using position-specific scoring matrices (PSSMs) (Altschul et al., 1997).

Conversely, Pattern Hit Initiated BLAST (PHI-BLAST) treats two occurrences of the same pattern within the target sequence as two independent sequences and is useful for filtering out false positives when pattern occurrences are random (Z. Zhang et al., 1998). PSI-BLAST is the most commonly used method for identifying homologous proteins. While there is no strict cut-off value for what is considered homologous, proteins with at least 30% sequence identity to the target protein and a BLAST e-value, the probability of seeing the alignment by chance, of less than 10^{-5}, are suitable metrics for identifying homologous templates.

Note that, while BLAST, PSI-BLAST, and PHI-BLAST are commonly used for detecting homologs, other homology detection tools, have been shown to be more accurate. For example, HHPred (Söding, Biegert, & Lupas, 2005) and HMMER3 (Eddy, 2011, Eddy & others, 2009) employ profile Hidden Markov Models (HMMs) to perform multiple sequence alignments. Like simple sequence profiles, profile HMMs contain information concerning amino acid frequencies in each column of a multiple sequence alignment, but they also contain information about the frequency of insertions and deletions at each column. Therefore, methods that use profile HMMs potentially can be more sensitive than methods that use simple sequence alignments (e.g., BLAST, PSI-BLAST).

Sometimes, homologous, experimentally determined structures cannot be identified for use as templates, in which case homology modeling is not applicable. However, as structure is better conserved evolutionarily than sequence, proteins with low sequence identity can have similar folds. In this case, 3D-fold recognition meta-servers, such as Phyre (Kelley & Sternberg, 2009) can be used. Phyre constructs a “fold library” via three steps: 1) combining a library of proteins of known structure from the Structural Classification of Proteins (SCOP) database (Murzin, Brenner, Hubbard, & Chothia, 1995) with new entries from the PDB, 2) scanning the sequences against a non-redundant

sequence database, and 3) constructing a sequence profile from the previous step. When a query sequence is submitted to the server, Phyre produces a sequence profile of the query with potential homologs by running PSI-BLAST, generates a consensus secondary structure prediction of the query after running a plethora of secondary structure prediction methods, and performs a profile-profile alignment of the results from PSI-BLAST and secondary structure prediction by scanning these inputs against the fold library. The resulting alignments are scored and ranked (Bennett-Lovsey, Herbert, Sternberg, & Kelley, 2008). Once a suitable template has been identified, a sequence alignment should be performed between the target and template sequences.

Additional considerations should be taken into account when ligand docking into a comparative model. It has been demonstrated that ligand docking into templates of experimentally determined holo structures is more likely to be successful than docking into apo structures. Further, use of a holo structure as a template was more predictive of success than the overall template-to-target sequence identity or sequence similarity of residues in the binding site. Ligand-bound template structures in which the ligands are similar to the target ligand should be prioritized; particular emphasis should be on ligands that share functional group placement to the target ligand. Finally, in order to obtain a diversity of models that span the probable conformations of the target, multiple templates should be identified and carried through the comparative modeling process. Then, ligand docking should be performed into each template. For example, if five potential template structures are identified, then there will be ligand-docked complexes with each of these five templates. Given that sequence similarity alone is not a reliable predictor of success *a priori*, ranking comparative models using templates chosen based on sequence similarity is often futile (Kaufmann & Meiler, 2012).

Sequence alignment and threading

Once a template, or templates, has been identified, the primary sequence of the target protein is threaded onto the three-dimensional (3D) backbone of the template structure according to a sequence alignment of the two proteins. If the alignment of the two proteins results in a gap during alignment, the gap regions, which are usually indicated as dashes (“-“) or spaces (“ “) in the alignment text file, are marked as loops in the newly generated threaded PDB file (Figure 5). Further, the Cartesian coordinates for the gap region are set to 0.000 and the occupancy column is set to −1.00. For information on the PDB file format, see www.pdb.org/docs.html. Regions between secondary structure elements and areas where there is low confidence in the sequence alignment between the target and template proteins are then reconstructed via a loop building protocol (Canutescu & Dunbrack, 2009, Coutsias, Seok, Jacobson, & Dill, 2004, Mandell, Coutsias, & Kortemme, 2009). These “loop” regions, which refer to any sequence to be re-built *de novo,* can be rebuilt using a fragment insertion-based or an analytic method (see below).

Defining loop regions

The loop definitions are chosen from the alignment between the target and template sequences. Regions having at least one of the following characteristics should be rebuilt as loops: 1) long coil regions with low sequence identity found in both template and target sequences, 2) regions with discrepancies in secondary structural elements between the template and target secondary structure prediction (i.e., a beta-sheet in the template was predicted to be a loop in the target), or 3) missing density after threading the target sequence onto the template. This process is illustrated in Figure 5.

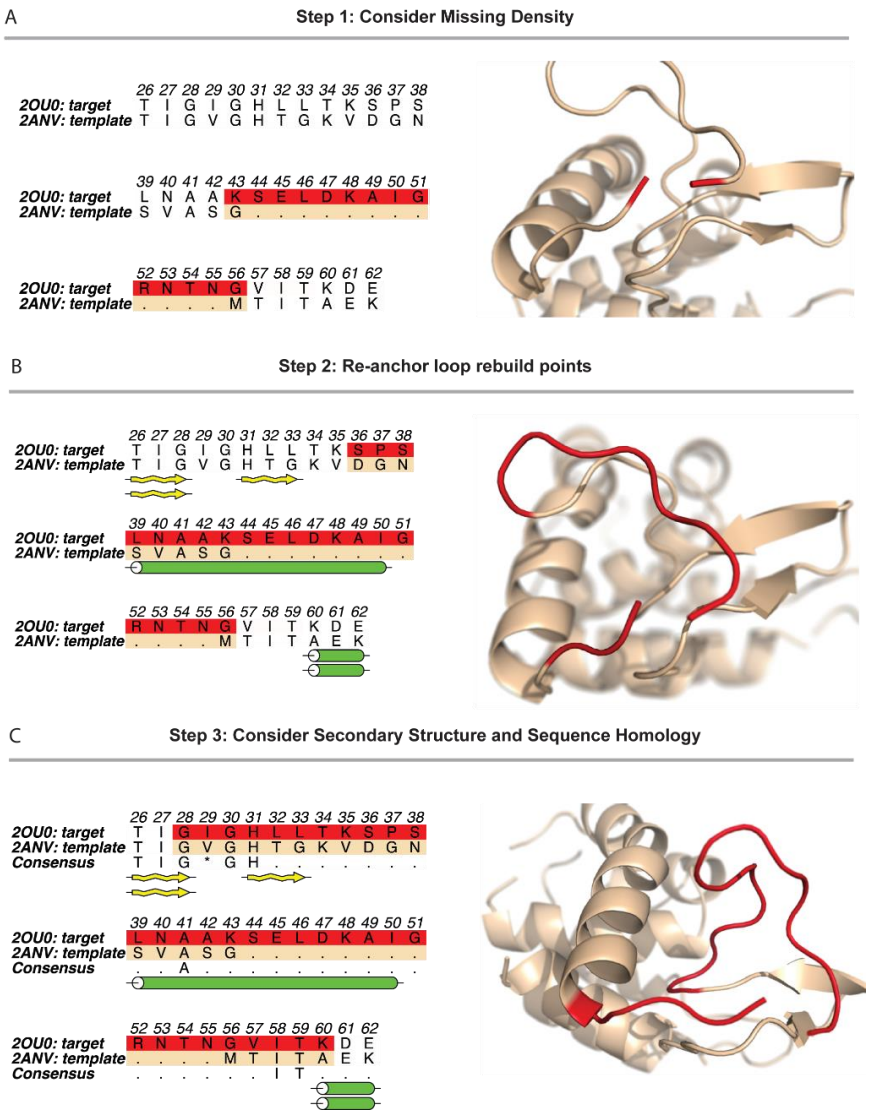


Figure 5

Criterion for selecting regions for de novo loop building.

A) The target sequence is threaded over the template backbone; the initial structure is shown in beige. There are twelve amino acids from the target sequence that do not have a corresponding amino acid from the template sequence (amino acids 44-55). The resulting alignment produces an insertion into the backbone of the template structure. To rebuild missing density, two anchor points, N- and C-terminal from the missing region, are chosen to remain fixed. The flanking amino acids of the areas of missing density (K43 and G56, highlighted in red) are chosen as the initial anchor points. Rosetta will perform *de novo* loop building in the area of missing density. B) The two anchor points are repositioned, allowing enough space to rebuild the twelve amino acids. In addition to the twelve-residue insertion, the region highlighted in red will be rebuilt with the *de novo* loop modeling protocol. C) During *de novo* loop rebuilding, secondary structure is also taken into consideration. Target residues 39-50 and 31-33 are both predicted to have secondary structural elements, but the template sequence does not contain secondary structural elements at these positions. Therefore, the loop to be built is extended to include residues 39-50 and 31-33. The final anchor points, G28 and K60, are chosen, allowing 31 amino acids to be rebuilt (shown in red).

Rosetta includes two loop-building algorithms. Cyclic coordinate descent (CCD), inspired by inverse kinematic applications in robotics, adjusts residue dihedral angles to minimize the sum of the squared distances between three backbone atoms of the moving N-terminal anchor and the three backbone atoms of the fixed C-terminal anchor (Canutescu & Dunbrack, 2009, Chu Wang et al., 2007). The advantages of CCD are its speed and its ability to close a loop over 99% of the time. Conversely, kinematic loop closure (KIC) analytically determines all mechanically accessible conformations for torsion angles of a peptide chain using polynomial resultants (Coutsias et al., 2004, Mandell et al., 2009). While KIC has been shown to recover loops from experimentally determined structures more accurately, it relies heavily on the location of the N- and C-terminal anchors and may not be an ideal choice for comparative modeling.

Rosetta loop building via CCD uses fragment libraries for generating loop coordinates for missing density in the threaded model. The fragment file is comprised of the target sequence divided into 3- and 9-amino acid overlapping sequence windows. There are 200 peptide fragments for each sequence window. After dividing the target primary sequence into 3- and 9-amino acid sequence windows, both Robetta and the fragment picker (Gront, Kulp, Vernon, Strauss, & Baker, 2011) application query a structural database of non-redundant proteins (G. Wang & Dunbrack, 2003) for each peptide sequence and store the corresponding Cartesian coordinates and secondary structure information in fragment files. For more detailed background and information on this application, see Gront, *et al*. (Gront et al., 2011) or go to www.rosettacommons.org/manuals/archive/rosetta3.4_user_guide/dc/d10/app_fragment_picker.html. Fragments can also be generated using NMR data using RosettaNMR (Rohl, 2005). For details on the procedure, please visit spin.niddk.nih.gov/bax/software/CSROSETTA/.

In this comparative modeling protocol, loop building takes place in two stages. In the first stage, a fast, low-resolution remodeling step with CCD consisting of broad sampling of backbone conformations is performed. In the second stage, the model is represented in all-atom detail and evaluated by Rosetta's all-atom scoring function. While it often makes little difference if loop building is performed in the absence or presence of the target ligand, it may be beneficial to manually place the ligand in the binding site of the threaded model *a priori*. This is especially true if the ligand is expected to interact with predicted loop regions.

All-atom refinement of the comparative model

Finally, the newly built model of the target protein undergoes refinement using the Rosetta all-atom scoring function (see below) to yield an all-atom protein model (Bradley, Malmström, et al., 2005). Both comparative modeling and ligand docking in Rosetta involve an all-atom refinement of the protein. The protocol used for structural refinement, visually described in Figure 6, is often referred to as “relax.” The goal of the relax protocol is to explore the local conformational space and to energetically minimize the protein. During this process, local interactions are improved by iterative residue side chain repacking, in which new side chain conformations, or “rotamers,” are selected from the Dunbrack library (Leaver-Fay, Kuhlman, & Snoeyink, 2005), and gradient-based minimization of the entire model, where the energy of the model is minimized as a function of the score. These small structural changes are evaluated according to the all-atom scoring function and are sampled in a Metropolis Monte Carlo (Metropolis et al., 1953) method. The relax protocol has been shown to dramatically lower the overall energy of the Rosetta model and is essential to achieving atomic detail accuracy (Bradley, Misura, et al., 2005, Rohl, Strauss, Misura, et al., 2004).

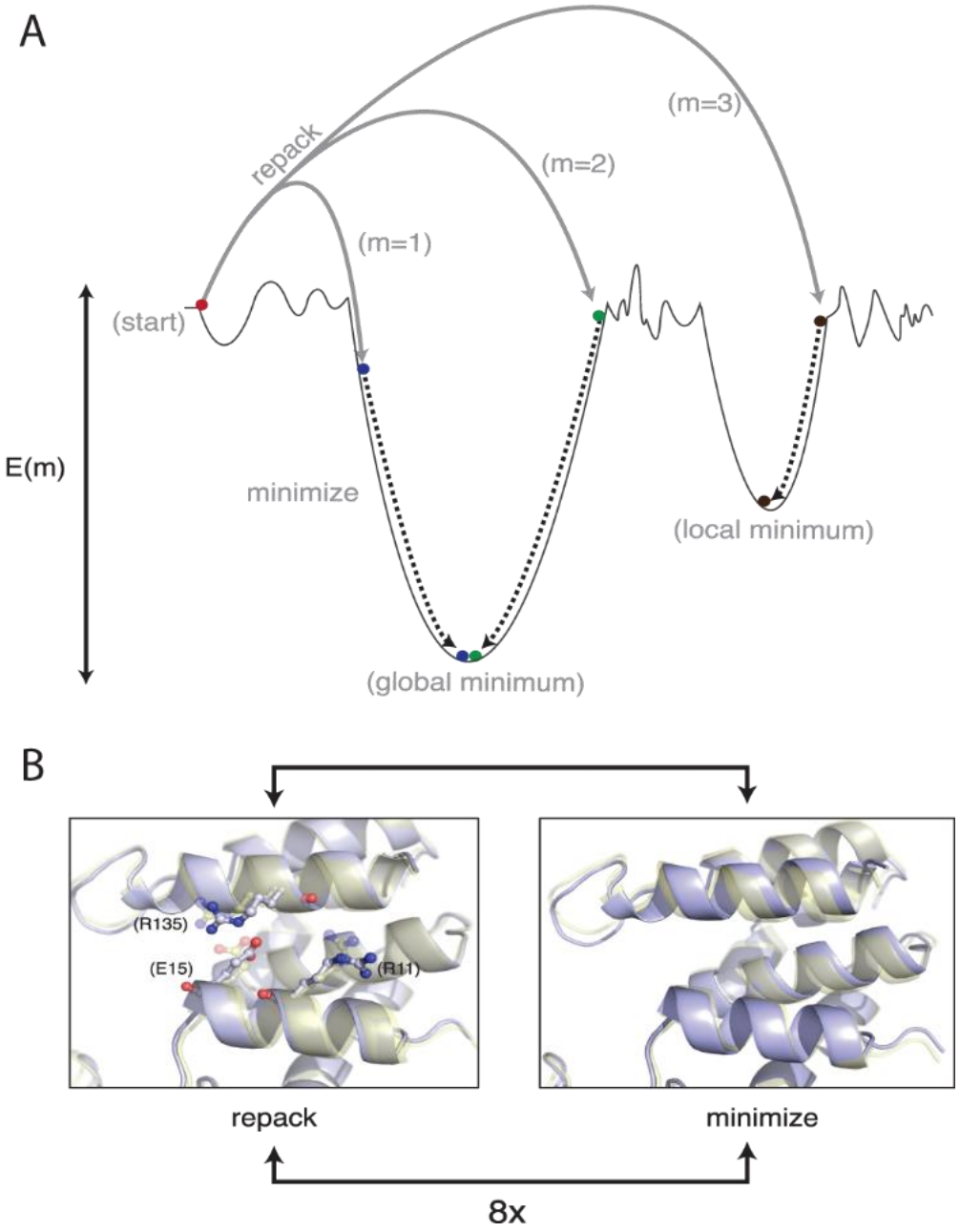


Figure 6
An overview of Rosetta energetic minimization and all-atom refinement via the relax protocol.
A) Simplified energy landscape of a protein structure. The relax protocol combines small backbone perturbations with side chain repacking. The coupling of Monte Carlo sampling with the Metropolis selection criterion (Metropolis et al., 1953) allows for sampling of diverse conformations on the energy landscape. The final step is a gradient-based minimization of all torsion angles to move the model into the closest local energy minimum. B) Comparison of structural perturbations introduced by the repack and minimization steps. During repacking, the backbone of the input model is fixed while side chain conformations from the rotamer library (Dunbrack & Cohen, 1997) are sampled. Comparison of the initial (transparent yellow) and final (light blue) models reveals conservation of the R135 rotamer but changes to the R11 and E15 rotamers. Minimization affects all angles and changes the backbone conformation.

Choosing a receptor model for ligand docking

The quality of each comparative model is evaluated by a scoring function consisting of solvation, electrostatic interactions, van der Waals attraction/repulsion, and hydrogen bonding terms (Bradley, Misura, et al., 2005, Jones, Taylort, & Thornton, 1992). As with template selection, it is difficult, if not impossible, to identify the single model that will ultimately provide the correct ligand-docked model. Therefore, as before, multiple structures resulting from comparative modeling should be used as receptors for ligand docking. These models are selected by pooling all models from each template and then selecting a small percentage of those models within each pool that fall below a certain energy cutoff. These top-scoring models are then clustered (see Appendix), and the cluster centers are carried forward into ligand docking. Clustering ensures that a maximally diverse set of models is used.

Ligand docking into comparative models

Next, the small molecule to be docked is placed into the binding site of each Rosetta model. For the best results, the target ligand is initially placed in a similar position to small molecules found in the original template structures. Ideally, biochemical information, such as results obtained from mutagenesis studies, can be used to inform the docking by restricting the conformational sampling space. If water molecules and co-factors are known to bind to the receptor, they can be added to the comparative models and docked simultaneously (Lemmon & Meiler, 2012), For simplicity, this feature is not demonstrated in our tutorial.

The Rosetta ligand docking algorithm first translates the ligand within a sphere having a user-specified radius (Davis & Baker, 2009, Meiler & Baker, 2006). These translations are repeated until the ligand's geometric center sits in a position not occupied by atoms in the receptor. These translations are followed by up to 1,000 cycles of random

rotation. A conformation resulting from rotation, in which the attractive and repulsive scores fall below a threshold value, is chosen for further refinement. Alternatively, if the position and orientation of the ligand is known, particularly if the target protein/ligand complex is highly homologous to an experimentally determined structure, then the translation/rotation movements described above may not be necessary and can be omitted.

In the high-resolution refinement step, six cycles of side chain rotamer sampling are coupled with small (0.1Å, 0.05 radians) ligand movements. Each cycle includes minimization of ligand torsion angles with harmonic constraints, where 0.05 radians of movement is equal to one standard deviation. Amino acid side chains are repacked using a backbone-dependent rotamer library (Dunbrack Jr & Karplus, 1993). During refinement, the weight of Rosetta's repulsive score term is decreased, thus preventing model rejection due to minor inter-atomic clashes. In a final energy minimization step, side chain rotamer sampling is coupled with minimization of backbone torsion angles. This is conducted with harmonic constraints on the alpha-carbon atoms (0.2Å standard deviation).

Several metrics can be employed to evaluate the results obtained from ligand docking. The most common evaluation method is analysis of the Rosetta energy, which is measured in Rosetta energy units (REU). Generally, models having lower, more negative Rosetta energies are considered to be more native-like (Raman et al., 2009). Another way in which binding modes can be confirmed is by comparing them to experimental data; further, these restraints can be employed to guide the modeling process (see Appendix). Alternatively, the models can be clustered in order to group the models by structural similarity (see Appendix). Clustering and score analysis reduce the data from thousands of models to a manageable number necessary to carry out an accurate, meaningful analysis.

Experimental data can also be used to filter out models during post-processing. Post hoc analysis allows for incorporation of data not easily represented as a restraint during model building. By performing rank-order predictions of binding energies, enzyme activities, or mutational effects and comparing these to known biochemical data, the correct model can be differentiated from those that do not agree with experimental observations (S. Combs et al., 2010, Kaufmann et al., 2009, Nannemann, Kaufmann, Meiler, & Bachmann, 2010). If restraints are not available, validation of the model is typically encouraged–and done--via experiments inspired by the computational results.

The Rosetta Energy Function

The energy, or scoring, function in Rosetta is derived empirically through analysis of observed geometries of a subset of proteins in the PDB. The measurements include, but are not limited to: radius of gyration, packing density, distance/angle between hydrogen bonds, and distance between two polar atoms. The measurements are converted into an energy function through Bayesian statistics (Dunbrack & Cohen, 1997, Simons et al., 1997).

The scoring function in Rosetta can be separated into two main categories: centroid-based scoring and all-atom scoring. The former is used for *de novo* structure prediction and initial rounds of loop building (Rohl, Strauss, Misura, et al., 2004, Simons et al., 1999, 1997). The side chains are represented as “super-atoms,” or “centroids,” which limit the degrees of freedom to be sampled while preserving some of the chemical and physical properties of the side chain. Although this centroid-based scoring function is important for *de novo* structure prediction, the structure prediction protocol is not covered within the scope of this article.

The all-atom scoring function represents side chains in atomic detail. Like the centroid-based scoring function, the all-atom scoring function is comprised of weighted

individual terms that are summed to create a total energy for a protein. Most of the scoring terms are derived from knowledge-based potentials. The scoring function contains Newtonian physics-based terms, including a 6-12 Lennard-Jones potential and a solvation potential. The 6-12 Lennard-Jones potential is split into two terms, an attractive term (fa_atr) and a repulsive term (fa_rep) for all van der Waals interactions (Kuhlman & Baker, 2000, Neria, Fischer, & Karplus, 1996). The solvation potential (fa_sol) models water implicitly and penalizes the burial of polar atoms (Lazaridis & Karplus, 1999). Inter-atomic electrostatic interactions are captured through a pair potential (fa_pair) (Simons et al., 1999), and an orientation-dependent hydrogen bond potential for long range and short range hydrogen bonding (hbond_sc, hbond_lr_bb, hbond_sr_bb, hbond_bb_sc, respectively) (Gordon, Marshall, & Mayot, 1999, Wedemeyer & Baker, 2003). In addition to the electrostatic terms, the Rosetta all-atom scoring function contains terms that dictate side chain conformations according to the Dunbrack rotamer library (fa_dun) (Dunbrack Jr & Karplus, 1993, Dunbrack & Cohen, 1997) preference for a specific amino acid given a pair of phi/psi angles (p_aa_pp), and preference for the phi/psi angles in a Ramachandran plot (rama)(Ramachandran, Ramakrishnan, & Sasisekharan, 1963, Rohl, Strauss, Chivian, et al., 2004, Wedemeyer & Baker, 2003).

The Rosetta Options File

The Rosetta options file allows users to pass specific protocol-related parameters to a specific Rosetta application. The options file is often called the “flags” file. Options can be accessed via the command line, placed within a file, or some combination of both. Below is an example of a Rosetta options file. Note that lines beginning with # are comments and are ignored when running Rosetta. Words in <> indicate where, in a specific case, the actual path to the necessary file would go (with no <>).

```
-database <database> #database location
-in
    -file
       -s <protein.pdb> #name of PDB file
-out
-prefix <desired_prefix> #desired output prefix of results files
-packing
-ex1 #use extra rotamer conformations chi 1
-ex2 #use extra rotamer conformations chi 2
-repack_only #prevents Rosetta from performing design
```

The space formatting of the options file is critical. In the example above, each new "namespace" (e.g., database, in, out, packing) starts a new line, and the "subspaces" (e.g., file) are indented by a space or a tab. However, tabs and spaces cannot be mixed within the same file. An alternate format for the options file is:

```
-database <database>
-in:file:s <protein.pdb>
-out:prefix <desired_prefix>
-packing:ex1
-packing:ex2
-packing:repack_only
```

In the above example, subspaces are designated by a colon (e.g. ex1 is a subspace option of the namespace packing; therefore, -packing:ex1.). If using RosettaScripts (Fleishman et al., 2011), which requires the input of an XML file, the options specified in this XML file override the options specified in the options file or those passed over the command line; therefore, it is important to avoid conflicting or contradicting options.

RosettaScripts XML File

RosettaScripts is an XML (Extensible Markup Language) scriptable interface to the Rosetta software with a variety of movers, scoring functions, and filters that can be tailored to a custom protocol (Fleishman et al., 2011). Movers are defined as steps in the protocol that can change the conformation of the system being modeled, or "pose." Examples of

movers include docking, loop building, and gradient-based minimization. Filters are utilized to decide if a given pose should proceed to the next step of the protocol. RosettaScripts protocols are versatile and can consist of a “mix-and-match” set of user-defined movers, filters, and scoring functions. This allows for complete customization of a protocol without manually editing the Rosetta source code. Because the RosettaScripts interface is able to access methods found throughout the Rosetta software suite, the user can instruct Rosetta to complete different tasks with a single XML file instead of running separate applications. The XML file is divided into five sections: scoring functions, filters, movers, constraints, and protocols. The format is shown below with generic names given for each section. For UserScoreFunctionName, UserFilterName, and UserMoverName the user can choose a name for the scoring function or filter. For RosettaMoverName, the name of the mover, as well as the options that accompany it, must be specified. Further information can be found at www.rosettacommons.org/manuals/archive/rosetta3.4_user_guide/Movers_(RosettaScripts).html.

```
<ROSETTASCRIPTS>
      <SCOREFXNS>
            <UserScoreFunctionName weights=“standard”/>
      </SCOREFXNS>
      <FILTERS>
            <UserFilterName name=“filter”/>
      </FILTERS>
      <MOVERS>
            <RosettaMoverName name=“UserMoverName” score=Scorefxnname/>
            <RosettaMoverName name=“userMoverName1” score=Scorefxnname/>
            <RosettaMoverName name=“UserMoverName2” score=Scorefxnname/>
      </MOVERS>
      <APPLY_TO_POSE>
      </APPLY_TO_POSE>
      <PROTOCOLS>
            <Add mover_name=“UserMoverName”/>
            <Add mover_name=“UserMoverName1” filter_name=“UserFilterName”/>
            <Add mover_name=“UserMoverName2”/>
       </PROTOCOLS>
</ROSETTASCRIPTS>
```

This generic XML file combines three separate movers that are scored by a user-defined scoring function (UserScoreFunctionName), where UserMoverName1 will be repeated until UserFilterName is satisfied. The input protein, the pose, steps through each mover iteratively until the final step is completed. The output is the final score of the pose and is given as a score file and/or the 3D coordinates of the final pose.

Anticipated results

For most applications of this protocol, biological systems will be used in which the structure of the protein or position of the docked ligand is not known, and results can only be compared to experimental data. In these cases, analysis of the results is best done using protein metrics and clustering. However, it is often beneficial to characterize the model population with respect to a single representative model in a fashion analogous to comparison to a crystal structure. In these cases, the best-scoring structure is often used.

Protein metrics are specific properties of the models. These can include van der Waal's packing, hydrogen bonds, and electrostatic interactions. Protein metrics can be tested with online servers or Visual Molecular Dynamics (VMD) (Humphrey, Dalke, & Schulten, 1996). The Rosetta energy function aims to minimize the energy of the protein with these properties in mind. In the case of ligand docking, the interface_delta score provides a measure of binding energy between the ligand and receptor. The interface_delta score is defined as the contribution to the total score for which the presence of the ligand is responsible.

Clustering refers to the process in which structurally similar models with a specified root mean square distance (RMSD) to each other are placed into groups, or clusters. After aligning the protein coordinates of all RosettaLigand models, RMSDs between all pairs of ligand-binding modes are computed. In order to assess the differences between the ligand binding modes, the RMSD of the ligand position must be calculated without ligand

superposition. In the presented tutorial, comparative models were superimposed. Since RosettaLigand docking does not alter the global position of the protein, ligand RMSDs can be calculated without additional protein superposition. The RMSD is computed as:

$$RMSD(A,B) = \sqrt{\left(\frac{1}{N}\sum_i^N {d_{(a_i-b_i)}}^2\right)},$$

where *A* refers to the first structure, *B* refers to the second structure, *N* is the number of atoms, *a* is an atom in structure *A*, *b* is an atom in structure *B*, and *d* is the Euclidean distance. Superposition of the complex must be performed prior to calculation of the ligand RMSD. The RMSD values are then used to cluster the models into structurally similar groups. The lowest-energy models in the largest clusters are considered to be the most "native-like" because 1) these binding modes were highly sampled by Rosetta and 2) are energetically favorable as determined by Rosetta's score function. Because the Rosetta score function is largely knowledge-based, Rosetta-built low-energy models are considered to recapitulate what is found to be energetically favorable in nature.

Although a Rosetta clustering application exists for protein structures (see Appendix), clustering small-molecule ligands is currently not possible within Rosetta. Alternative tools to cluster ligands include the BioChemical Library, or BCL (www.meilerlab.org/index.php/bclcommons), 3DLigandSite (Wass, Kelley, & Sternberg, 2010), Canvas by Schrödinger, the VcPpt extension for AutoDock Vina from BiochemLab Solutions (biochemlabsolutions.com) (Morris et al., 2009), the ptraj tool in the AMBER suite (http://ambermd.org) (Case et al., 2005), and RDKit (rdkit.org).

In the example presented in the tutorial, bcl::ScoreProtein was used to compute RMSD values between ligands, and bcl::Cluster was used to cluster the top ten percent of ligands into structurally similar bins with a cluster girth cutoff of 2Å. The binding mode with the lowest `interface_delta` score from the largest cluster is often chosen as a representation of Rosetta's best prediction for the ligand docking experiment (Figure 8).

Because of the imperfect nature of the Rosetta scoring function, it is possible that Rosetta ranks an incorrect binding mode better than the correct binding mode (Figure 8B). For this reason, it is suggested that after clustering, the lowest-energy models from each of the top five clusters are considered as putative binding modes. Kaufmann, et al. describe how biochemical data, such as mutagenesis studies, can be used to select from among several low-scoring, RosettaLigand-predicted binding modes (Kaufmann et al., 2009).

The appropriate RMSD cutoff for clustering will vary depending on the characteristics of the protein binding site and the ligand being docked. In this example, due to the ligand size, a conservative cluster RMSD cutoff of 2Å was used. If larger ligands are used, the cluster cutoff can be increased to 3 to 5Å (S. Combs et al., 2010). To determine the size of the cluster RMSD cutoff, multiple RMSDs can be tested. Once the clusters have been generated, the cluster sizes (i.e., the number of models in each cluster) can be measured. If any single cluster contains a large percentage of the total models used, a larger cutoff may be used.

In addition to clustering ligands, experimental data can be used to determine the correct ligand binding mode. Kaufmann, *et al*. used the relative rank of ligand energies from analogs of serotonin to determine the binding mode of serotonin into a Rosetta-built comparative model of the human serotonin transporter (hSERT) (Kaufmann et al., 2009). Experimental binding affinities were correlated to the rank of each small-molecule in the binding pocket of hSERT. With the same comparative model, Combs, *et al*. performed computational mutagenesis of the hSERT binding pocket to determine the binding mode of *S*- and *R*-citalopram (S. Combs et al., 2010).

In the example of MR3 docked into a comparative model of T4-lysozyme demonstrated in this tutorial, results from Rosetta can be compared with the crystal structure from the PDB (PDB ID: 2ou0). By computing RMSDs between models generated by Rosetta and the experimentally determined, or native, structure, the accuracy of the

models can be evaluated. The methods used to generate these RMSD values can be found in the Appendix.

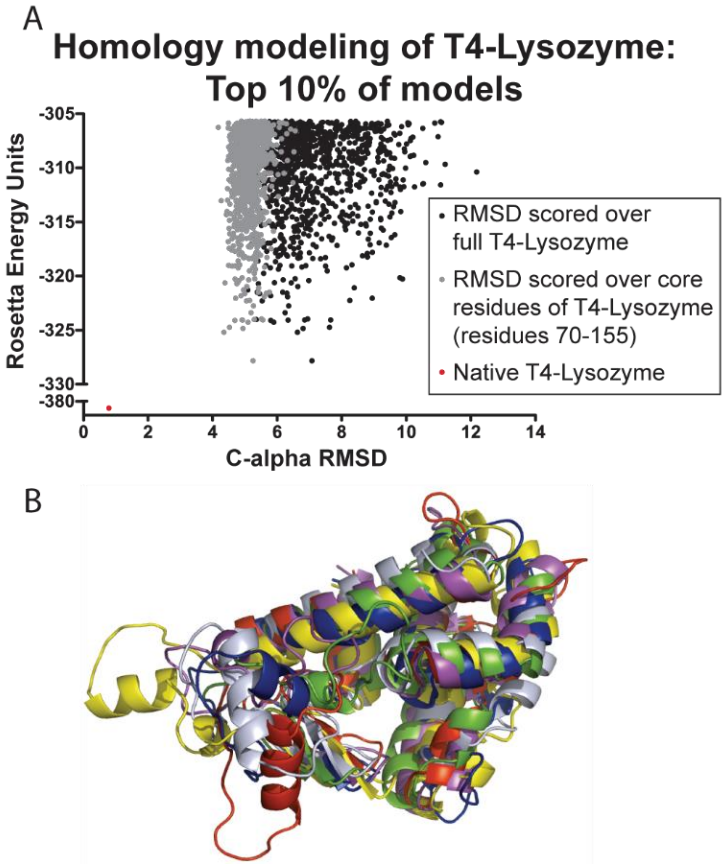


Figure 7
Building loops in comparative models of T4-lysozyme.
Loops were rebuilt in comparative models of T4-lysozyme using P22-lysozyme as a template as detailed in STEPS 1-13 of the protocol. A) The RMSD of Cα atoms between 10,000 models and the native protein (PDB ID: 2ou0) was computed over the full protein (black) and the core residues of T4-lysozyme (gray). The top 10% of models by Rosetta energy are shown here. Generally, a low Rosetta energy correlates with a low RMSD. For comparison, the Rosetta energy for the energy-minimized native crystal structure is shown in red. B) Five of the lowest energy models are seen in comparison to the native structure (shown in gray).

Plotting the interface_delta score vs. RMSD of the ligand models, as shown in Figure 8, demonstrates that a local minimum exists in a different binding site from that observed in the native crystal structure. Often, when a small ligand is docked into a large

binding pocket, several local energy minima in which the ligand can bind are detected. Rosetta is able to identify alternative binding pockets besides that which is found in the crystal structure. However, the binding mode closest to that of the crystal structure still ranks within the top three percent of the total docked binding modes. An energy funnel is often observed in score vs. RMSD plots, indicating the presence of a single energy minimum. However, the energy funnel in Figure 8 is poorly formed, which is not surprising given the nature of the complex.

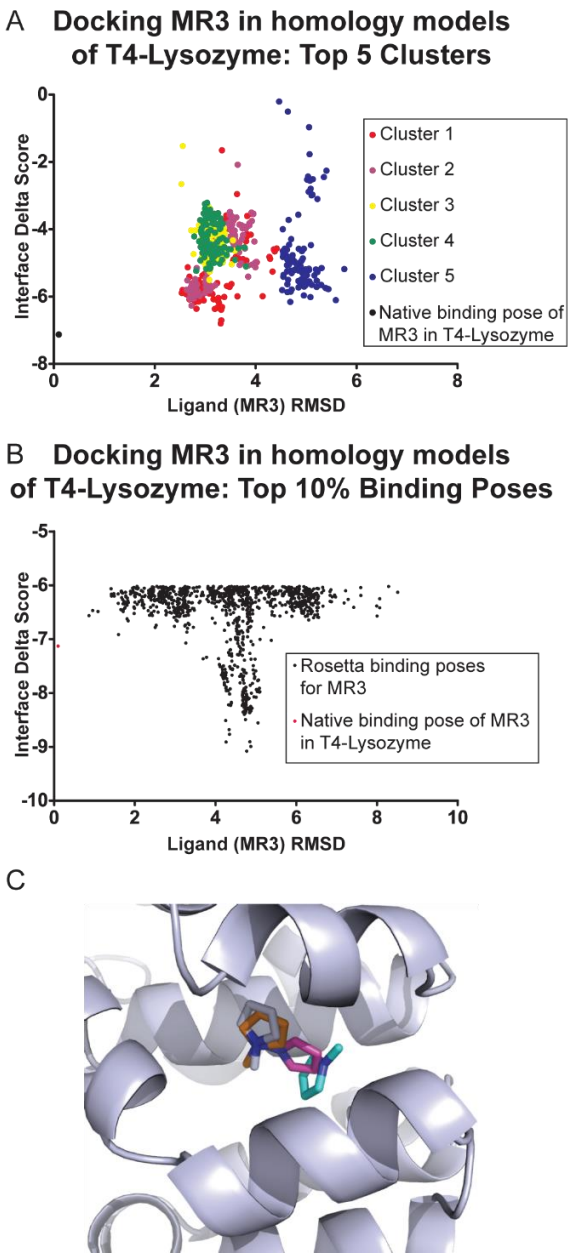


Figure 8
Docking MR3 into comparative models of T4-lysozyme.
The MR3 ligand was docked into the ten lowest-energy comparative models of T4-lysozyme as detailed in STEPS 17-22 of the protocol. A) 10,000 binding modes were clustered by RMSD using applications available in the bcl::Commons. The largest five clusters are shown, with the interface_delta score plotted against the RMSD to the native ligand binding mode (shown in black). Generally, the largest clusters are also those with the lowest RMSD to the native binding mode. B) The RMSD between 10,000 binding modes and the native binding mode (shown in red) was computed. The top ten percent of models by interface_delta score are shown here. Sub-angstrom binding modes are within the top ten percent of models, but Rosetta also identifies an alternative lower-energy binding mode within the site. C) The lowest RMSD binding mode (orange) is closer to the native binding mode (gray) than the lowest-energy binding mode of the largest cluster (magenta) and the lowest-energy binding mode overall (cyan).

The results from the modeling example presented in the tutorial point to the challenges associated with docking a small, symmetric ligand into a relatively large binding pocket. Obtaining the correct binding conformation and position of the ligand is further hampered by the low confidence of the comparative model. However, the results also show that Rosetta is capable of sampling the correct binding conformation and assigning this conformation a relatively low energy according to the knowledge-based scoring function. Further, in a situation in which the binding conformation is unknown, Rosetta may be used to predict potential interacting residues. The predicted model will then need to be tested experimentally to confirm its validity.

Author Contributions

All authors contributed equally to this work. All authors wrote substantial portions of the main text, the figures, and the supplemental information. S.A.C. proposed the composition of the work for the benefit of the scientific community, tested the presented protocol, and managed submission. S.L.D. wrote instructions on how to install the software, generated the comparative models, wrote data processing scripts, and managed references. S.H.D. wrote the supplemental glossary and was responsible for overall editing of the work. G.H.L. wrote the RosettaLigand program in its present form. D.P.N. carefully read through the manuscript for consistency and accuracy and aided in the analysis of the generated models. E.D.N. also generated comparative models, performed all of the ligand docking, and performed the data analysis. J.R.W. contributed several figures, data processing scripts, specialty movers, wrote large sections of the tutorial, and managed references. J.H.S. tested the presented protocol and wrote the Troubleshooting section. J.M. helped define the scope of the work and guided the process of establishing the protocol.

Acknowledgements

The authors would like to thank Dr. Jarrod Smith, as well as Rebecca Levinson, for testing the protocol and for their helpful comments. Work in the Meiler laboratory is supported through NIH (R01 GM080403, R01 MH090192) and NSF (Career 0742762). The authors declare that they have no competing financial interests.

CHAPTER III

ASSESSMENT AND CHALLENGES OF LIGAND DOCKING INTO COMPARATIVE MODELS OF G PROTEIN-COUPLED RECEPTORS

This work is based on publication (Nguyen, Norn, Frimurer, & Meiler, 2013).

Summary

The rapidly increasing number of high-resolution X-ray structures of G protein-coupled receptors (GPCRs) creates a unique opportunity to employ comparative modeling and docking to provide valuable insight into the function and ligand binding determinants of novel receptors, to assist in virtual screening and to design and optimize drug candidates. However, low sequence identity between receptors, conformational flexibility, and chemical diversity of ligands present an enormous challenge to molecular modeling approaches. It is our hypothesis that rapid Monte-Carlo sampling of protein backbone and side-chain conformational space with Rosetta can be leveraged to meet this challenge. This study performs unbiased comparative modeling and docking methodologies using 14 distinct high-resolution GPCRs and proposes knowledge-based filtering methods for improvement of sampling performance and identification of correct ligand-receptor interactions. On average, top ranked receptor models built on template structures over 50% sequence identity are within 2.9 Å of the experimental structure, with an average root mean square deviation (RMSD) of 2.2 Å for the transmembrane region and 5 Å for the second extracellular loop. Furthermore, these models are consistently correlated with low Rosetta energy score. To predict their binding modes, ligand conformers of the 14 ligands co-crystalized with the GPCRs were docked against the top ranked comparative models.

In contrast to the comparative models themselves, however, it remains difficult to unambiguously identify correct binding modes by score alone. On average, sampling performance was improved by 103 fold over random using knowledge-based and energy-based filters. In assessing the applicability of experimental constraints, we found that sampling performance is increased by one order of magnitude for every 10 residues known to contact the ligand. Additionally, in the case of DOR, knowledge of a single specific ligand-protein contact improved sampling efficiency 7 fold. These findings offer specific guidelines which may lead to increased success in determining receptor-ligand complexes.

Introduction

Being able to model the complex interactions between receptors and small molecule ligands offers immense opportunities for the basic biochemical understanding of signaling processes and for the development of pharmacological tool compounds and drugs that modulate receptor function. The human genome encodes for approximately 800 G protein-coupled receptors (GPCRs) that orchestrate the communication between a cell and its surroundings – an obvious place for small molecule drugs to interfere (Lagerström & Schiöth, 2008). While more than 26% of our current small molecule drugs target Class A GPCRs alone (Overington et al., 2006), structure-based drug discovery has played a limited role in developing these molecules. GPCRs have been the subject of many structural, comparative modeling and docking studies. However in many cases, models are affiliated with high uncertainty and inaccuracy. Primary reasons include a lack of adequate template structures, the existence of multi-conformational states which require intense conformational sampling of not only the protein side chain but also backbone conformational space, in combination with the large variety of ligands that interact with GPCRs, including very flexible molecules which are notoriously challenging

subjects for accurate docking (Kufareva, Rueda, Katritch, Stevens, & Abagyan, 2011, Michino, Abola, Dock, & others, 2009). Nevertheless, increasing the availability of reliable GPCR models for structure-based drug discovery would be beneficial in the development of novel, potent and subtype-selective molecules. Since the landmark publication by Rasmussen et al. in 2007 (Rasmussen et al., 2007), the number of experimentally determined GPCR structures has been increasing rapidly and now totals to 18 distinct GPCR structures that are recorded in the Protein Data Bank (PDB). While this is still just a small subset of the GPCR space, it provides a more substantial basis for comparative modeling and docking simulations.

Despite the increase in experimental structural information, it remains difficult to predict ligand-binding conformations in comparative models of GPCRs for all except the receptors most similar to those which have been determined experimentally (Beuming & Sherman, 2012, Kufareva et al., 2011, Michino et al., 2009). This difficulty originates in part from the necessity of sampling both receptor and ligand flexibility which, due to the necessarily approximate nature of the force fields and protein-less/ligand-less sampling methods, results in the sampling of biologically irrelevant conformations. This complicates discrimination between the global minimum energy conformation (GMEC) and the local minimum energy conformations (LMEC) of the binding complex, as deeper sampling reveals many different energy-equivalent binding modes. The reason for the difficulty in GMEC and LMEC discrimination is, as discussed by Fleishman and Baker (Fleishman & Baker, 2012), related to the small energy gap in ligand binding, which moreover is challenging to measure as it is often mediated by polar contacts and water molecules.

For GPCRs, the ligand docking problem is even more difficult for three reasons. Firstly, the alignment is not trivial, as the transmembrane helices occasionally contain bulges and kinks and the length of the transmembrane helix is not conserved. Secondly, GPCRs are able to assume multiple different conformations with approximately the same

energy, as demonstrated by studies on the beta-adrenergic receptors (K.-Y. M. Chen, Zhou, Fryszczyn, & Barth, 2012, Nygaard et al., 2013). Thirdly, three extracellular loops must be modeled, as they often contact the ligand and are involved in ensemble stabilization in some receptors (Klco, Wiegand, Narzinski, & Baranski, 2005, Schwartz & Rosenkilde, 1996).

At the same time, improved algorithms and high-performance computing revolutionize our ability to sample protein conformational space swiftly, enhancing the possibility to accurately dock ligands into comparative models (Kaufmann & Meiler, 2012). This, combined with the increasing number of available templates, allows us to assess the applicability of rapid Monte Carlo Metropolis (MCM) sampling as implemented in the Rosetta suite of programs for GPCR comparative modeling and docking (Leaver-Fay et al., 2011). Specifically, we address the accuracy of backbone placement in transmembrane and extracellular loops, sampling of ligand binding modes and side-chain conformations in the binding site, and strategies to select accurate models from the large conformational space sampled.

Methods

Database generation

The highest resolution experimental structure for each unique GPCR in the Protein Data Bank (PDB) at the time of writing was chosen for comparative modeling and ligand docking, as shown in Table 4. This includes the following G protein-coupled receptors: rhodopsin (Okada et al., 2004), β1-adrenergic (Warne et al., 2008), β2-adrenergic (Cherezov et al., 2007), A2A adenosine (Jaakola et al., 2008), CXCR4 chemokine (B. Wu et al., 2010), dopamine D3 (Chien et al., 2010), histamine H1 (Shimamura et al., 2011), S1P1 sphingosine 1-phosphate (Hanson et al., 2012), M2 muscarinic acetylcholine (Haga

et al., 2012), M3 muscarinic acetylcholine (Kruse et al., 2012), mu-opioid (Manglik et al., 2012), kappa-opioid (H. Wu et al., 2012), N/OFQ opioid (A. A. Thompson et al., 2012) and delta-opioid (Granier et al., 2012). Comparative models were constructed of each GPCR using the other 13 structures as templates. Ligand docking was performed with the small molecules crystallized within each receptor (Figure 43). A flowchart demonstrating the full protocol carried out in this study is shown in Figure 9. Full command lines for each step are included in the Appendix.

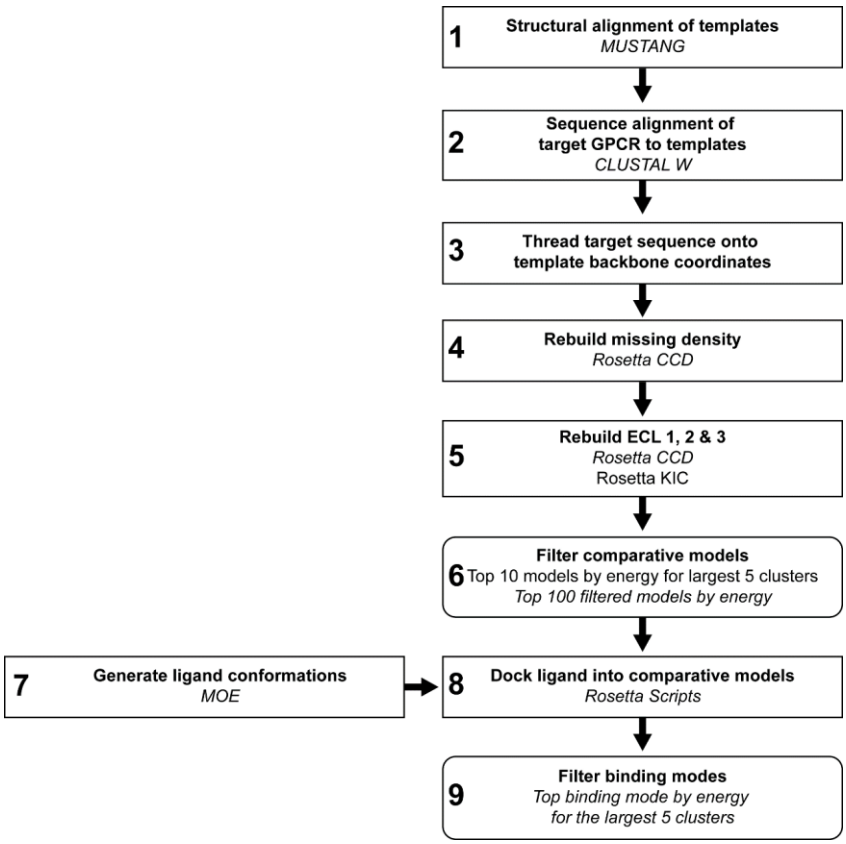


Figure 9
Flowchart of the comparative modeling and ligand docking protocol.
For each step in the protocol, the name of the application or method used to execute each step is included. Where multiple methods are mentioned, the results from the method in italics were carried on to the next step.

Table 4
G protein-coupled receptor experimental structures and their ligands used in this study.
Data collected from the Protein Data Bank.

Protein name	PDBID / Chain	Loop length ECL1/2/3[a]	Resolution (Å)	Ligand	Rotatable bonds	Fold decrease in sampling efficiency when using ligand conformers [b]	Waters within 4Å [c]	Contacts (pocket/ loops) [d]
Rhodopsin (bRh)	1U19 / A	5/27/7	2.2	Retinal	5	4.0	1	19/6
Beta1-Adrenergic (B1Ar)	2VT4 / B	6/26/6	2.7	Cyanopindolol	6	3.2	0	19/2
Beta2-Adrenergic (B2Ar)	2RH1 / A	5/26/6	2.4	Carazolol	6	4.6	0	18/2
A2A adenosine (A2Ar)	3EML / A	6/25/8	2.6	ZM241385	4	15.0	7	13/5
CXCR4 chemokine (CXCR4)	3ODU / A	5/18/7	2.5	IT1t	7	68.1	7	7/6 (+1N term)
Dopamine D3 (D3R)	3PBL / A	8/16/7	2.89	Eticlopride	7	6.6	N/A	20/2
Histamine H1 (H1R)	3RZE / A	6/17/5	3.1	Doxepin	3	3.1	N/A	18/1
S1P1 sphingosine 1-phospate (S1P1R)	3V2W / A	10/18/8	3.35	ML056	11	42.4	N/A	21/5 (+2N term)
M2 muscarinic acetylcholine (M2R)	3UON / A	7/17/6	3	3-quinuclidinyl-benzilate	5	6.4	2	18/1
M3 muscarinic acetylcholine (M3R)	4DAJ / A	7/17/7	3.4	Tiotropium	5	6.4	N/A	19/1
Mu-opioid (MOR)	4DKL / A	7/20/6	2.8	β-FNA	8	27.2	4	15/0
Kappa-opioid (KOR)	4DJH / A	5/22/4	2.9	JDTic	7	8.8	5	22/2
N/OFQ opioid (NOP)	4EA3 / A	4/19/6	3.01	C-24	8	50.6	1	17/4
Delta-opioid (DOR)	4EJ4 / A	4/19/7	3.4	Naltrindole	2	1.6	N/A	15/0

[a]Loop lengths determined by DSSP.[b] Ligand conformers generated by MOE. The fold decrease in sampling efficiency is the uniform sampling efficiency within a 2.0 Å radius (USE2.0) for the bioactive ligand conformation divided by USE2.0 for ligand conformers. [c]Number of water molecules bridging the receptor and ligand within 4 Å of the ligand crystallized in the binding pocket. [d] Number of residues in the receptor in contact with the ligand. Indicated is the number of contacts within the transmembrane region of the receptor binding pocket versus the loop or N-terminal regions.

Sequence alignment and threading

The first step in constructing the models was performing a sequence alignment of the target sequence with a structural alignment of the other 13 GPCRs to be used as templates. A structure-based sequence alignment was generated of all 14 GPCR templates using MUSTANG (Konagurthu, Whisstock, Stuckey, & Lesk, 2006) as seen in Figure 44 (Figure 9, Step 1). The sequence of the target GPCR was then aligned with the profile of structurally aligned templates using CLUSTALW (J. D. Thompson, Higgins, & Gibson, 1994) (Figure 9, Step 2). The sequence of the target GPCR was then placed onto the helical backbone coordinates of each template structure (Figure 9, Step 3). Any missing density and variable loop regions were constructed using the *ab initio* cyclic coordinate descent protocol in Rosetta (Canutescu & Dunbrack, 2009, Chu Wang et al., 2007) (Figure 9, Step 4).

Building in missing density and extracellular loop regions in the comparative models

Missing density in the threaded models due to gaps or insertions in the sequence alignment were built in Rosetta using Monte Carlo Metropolis (MCM) fragment replacement combined with cyclic coordinate descent loop closure (Canutescu & Dunbrack, 2009, Chu Wang et al., 2007) (Figure 9, Step 4). Cyclic coordinate descent (CCD) was inspired by inverse kinematic applications in robotics and closes loops by minimizing the sum of the squared distances between three backbone atoms of the moving N-terminal anchor and the three backbone atoms of the fixed C-terminal anchor through the adjustment of dihedral angles. Its speed and its ability to close a loop over 99% of the time gives CCD an advantage over other loop closure methods. In brief, loop regions defined by the user are chosen in a random order and for each loop, ϕ-ψ angles of backbone segments from homologous sequence fragments from the PDB, excluding

those from the target experimental structure, are introduced into the loop regions. After the fragment substitution, small movements in the ϕ-ψ angles are performed to close breaks in the protein chain. After each defined loop has been closed, resulting full sequence models were subjected to eight iterative cycles of side chain repacking and gradient minimization of ϕ, ψ, and χ angles using the Rosetta scoring function with an implicit membrane potential (Yarov-Yarovoy et al., 2005). A total of 200 models were constructed with each threaded model and the lowest energy model was chosen for a full remodeling of the extracellular loops (Figure 9, Step 5). Extracellular loops, as shown in Figure 45, were extensively rebuilt using both the cyclic coordinate descent loop closure method described above and the kinematic loop closure method described below. Approximately 1000 models were built for each target-template pair, resulting in a minimum of 13,000 comparative models per target structure.

A limited benchmark over the comparative modeling of six GPCRs was performed to compare the results of the kinematic loop closure (KIC) method in Rosetta (Mandell et al., 2009) with CCD. KIC analytically determines all mechanically accessible conformations for six pivot torsion angles of a peptide chain using polynomial resultants. During kinematic loop closure, all mechanically accessible conformations for ϕ and ψ dihedral angle torsions from the first, middle and last residues in a loop segment, designated as pivot torsions, are sampled. The remaining torsion angles are randomly sampled using Monte Carlo minimization from Ramachandran probabilities of each amino acid. The six pivot torsions are solved analytically to close the loop. The protocol is performed for 720 rounds of high resolution loop closure and models accepted by the Metropolis criterion are subjected to side chain repacking and gradient minimization as described above. The data from the benchmark set comparing the two methods indicated that overall, CCD produced comparative models with an average root mean square

deviation (RMSD) of 2.0 Å over extracellular loops (ECL) 1 and 3, which was significantly lower than the average RMSD over ECL1 and 3 for KIC at 2.6 Å (Table 21). The difference between CCD and KIC for the average RMSD over the full receptor was even more striking at 3.8 Å and 6.9 Å respectively. Results from CCD loop modeling were consequently used for further analysis.

Selection methods of comparative models for docking

Comparative models were filtered for ligand docking using two different methods (Figure 9, Step 6). Both methods partially build on the observation that receptor accuracy is correlated with the Rosetta energy function (Figure 46). The first method was based on clustering of the 10% best scoring structures. Clusters were determined based on pairwise RMSD of all C-alpha atoms using bcl::Cluster (Alexander, Woetzel, & Meiler, 2011) and a cluster radius of 3.0 Å. The best scoring models in each of the clusters were used for further analysis. The second method was created to avoid sampling of non-native ligand binding pocket conformations. Pocket residue positions were defined across all GPCRs as positions in the sequence alignment where C-alpha atoms of the residues had a distance of less than 4.0 Å to the ligand in at least one experimental structure. This yielded a list of 29 residue positions, which was reduced to 25 residue positions when the four residue positions at the top of transmembrane helices (TM) two and five were removed to avoid bias from structural alignment of the proteins. Pocket residues are shown in the alignment in Figure 44. Comparative models passed the filter only if C-alpha atoms of all pocket residues had an alignment equivalent pocket residue in another GPCR within a distance of a residue position specific cutoff. The cutoffs were chosen to be residue specific to represent varying flexibility in different parts of the receptor. The maximum distance between a specific pocket residue in any receptor and an equivalent pocket residue in any other GPCR, according to the alignment shown in Figure 44, was chosen

as a distance cutoff for that particular residue position. When applying the knowledge-based filter, the self-experimental structure was not considered to avoid circular bias.

Generation of ligand conformers

In preparation for docking, ligand conformers were generated by MOE (Molecular Operating Environment, Chemical Computing Group, Ontario, Canada) using the MMFF94x force field and Generalized Born implicit solvent model (Figure 9, Step 7). Conformers were generated using 10,000 iterations of the Low Mode MD method (Labute, 2010) with a redundancy cutoff of 0.25 Å. Energy cutoffs for ligand conformers were dependent on the number of rotatable bonds: 3 kcal/mol for 1-6 rotatable bonds, 5 kcal/mol for 7-9 rotatable bonds and 7 kcal/mol for 10-12 rotatable bonds (Perola & Charifson, 2004). The RMSD distribution for the generated ligand conformers compared to the bioactive ligand conformation is shown in Figure 52.

The ligand conformers were protonated as shown in Figure 43. These protonation states were determined based on the local environment in the individual experimental structures. In the case of ligand C-24, the protonation state is not what would be predicted without information from the experimental structure. We note that this adds some bias to the method. Likewise, the stereochemistry of the ligand (*E)*-IDT in CXCR4 was taken directly from the experimental structure. Of note, the experimental structure was determined with a mixture of (E) and (Z)-form, which cannot be clearly distinguished from the electron density (Raymond Stevens, personal communication).

Docking ligands into a chosen ensemble of comparative models

Ligand docking into the comparative models was performed with Rosetta Scripts (Davis & Baker, 2009, Fleishman et al., 2011, Lemmon & Meiler, 2012, Meiler & Baker,

2006) (Figure 9, Step 8). Each ligand was allowed to sample binding modes in a 5.0 Å radius from the coordinate representing the center of the ligand binding mode as given in the experimental structure. This adds some bias to the sampling, as the smallest unbiased docking sphere enclosing all ligand binding conformations has a radius greater than 5.0 Å. During the low-resolution phase of docking, rigid body orientation of the ligand centroid is performed through translation until the geometric center of the ligand is in a position not occupied by atoms in the receptor. High-resolution docking then begins with 1000 cycles of full rotational freedom until the attractive and repulsive forces fall below a threshold value. Six cycles of side-chain rotamer and ligand conformer sampling are then coupled with 0.1 Å, 0.05 radian ligand movements simultaneously in a Monte Carlo simulated annealing algorithm. All rotatable bonds within the ligand, except for planar conjugated bonds, were allowed full flexibility as indicated within the ligand parameters file. Ligand conformers are randomly chosen until the Monte Carlo criterion has been satisfied. A final minimization combines side-chain rotamer sampling with backbone torsion angle minimization with harmonic constraints on the C-alpha atoms.

The energy function used during the docking procedure contains terms for van der Walls attractive and repulsive forces, statistical energy derived from the probability of observing a particular side-chain conformation in the PDB, hydrogen bonding, electrostatic interactions between pairs of amino acids, and solvation assessing the effects of both side-chain/side-chain interactions and side-chain/ligand interactions. For each ligand, over 2,000 docked complexes were generated and evaluated in comparison to the experimental ligand binding mode using RMSD to the heavy atoms.

Assessing the size of the ligand conformational space

We propose a new measure to enable comparison of docking benchmark studies across targets and to test how the methods compare to random sampling – the uniform

sampling efficiency (USE2.0). The proposed measure is equivalent to the sampling frequency of better-than-2.0-Å-RMSD-binding-modes that would occur by random sampling in a 5.0 Å docking sphere with no occluding protein, given a set of ligand rotamers and full rotational and translational freedom. To calculate USE2.0, each i rotamer of the N rotamers in the generated ligand ensemble was aligned to the experimental structure and rotated along its principal axes (φ,θ,ψ) using M (40) uniform spacings. For sampling to be uniform, a correction factor, Cφ, is needed to account for the fact that the number of ways of choosing θ, given φ, is proportional to the circumference of the circle that θ draws on the φ,θ sphere (Bowie, 1997). The translation distance that increased the RMSD to 2.0 Å was determined for each rotamer-rotation set r(i,φ,θ,ψ). USE2.0 was then determined as the fraction between the volume of space containing binding modes below 2.0 Å and sampled volume of the 5.0 Å (R) docking sphere.

$$\mathrm{USE2.0} = \sum_{i=1}^{N} \sum_{\varphi}^{M} \sum_{\theta}^{M} \sum_{\psi}^{M} \mathrm{r}_{\mathrm{i},\varphi,\theta,\psi}^{3} C_{\varphi} / R^{3}$$

The distribution of RMSDs that arise from uniform sampling of ligand conformations is nontrivial, dependent on the ligand size and on the generated conformers. For the ligands considered in this dataset, USE2.0 varies from 10^{-5} for DOR to 10^{-7} for S1P1.

Enrichment of native-like binding modes using known contacts between the ligand and GPCR

When a mutation of a residue strongly affects ligand binding, this residue is often interpreted as having a direct contact to the ligand. To assess how this type of constraints enriches for the correct binding mode in our ligand-protein ensembles, we determined the

average enrichment through 10,000 iterations of n randomly chosen known contacts with n running from 0 to all known contacts for a particular receptor-ligand interaction.

Ligand-protein evaluation through RMSD-based clustering and binding energy

Results from the ligand docking study were evaluated using clustering on pairwise RMSD values calculated over the ligand heavy-atoms using bcl::Cluster with a 2.0 Å cutoff (Figure 9, Step 9). The lowest energy binding modes of the five largest clusters were chosen for further analysis. The coverage and accuracy of correct ligand-receptor contacts compared to the experimental structure was calculated on the top ranked models using SimiCon (Rueda, Katritch, Raush, & Abagyan, 2010). Contact coverage is calculated as the number of correct ligand contacts from the model divided by the total number of ligand contacts made in the experimental structure. Accuracy of correct contacts is calculated as the number of correct contacts divided by the total number of ligand contacts made in the model.

All plots were made with the Python 2D plotting library, matplotlib (Hunter, 2007) and Prism 5.01 (GraphPad Software, San Diego, CA). The alignment figure was created used Aline (Bond & Schuttelkopf, 2009). Structural figures were created with PyMOL (PyMOL Molecular Graphics System, Version 1.5.0.4, Schrödinger, LLC).

Results and Discussion

The results are presented in four parts. In the first part, we discuss the accuracy of comparative models generated by sequential sampling algorithms of the transmembrane and loop regions. Secondly, we discuss methods to select the most accurate models for ligand docking. Thirdly, we assess the equivalency of the local minimum interaction energy conformation (LMIEC) with the lowest energy that we sample

and the experimental ligand binding mode and ask how much receptor flexibility can be sampled before the lowest energy LMIEC deviates from the experimental ligand-protein complex. In the fourth part, we assess the sampling efficiency of native-like ligand binding modes by docking the ligand ensemble into the ensemble of comparative models. Various methods used to identify native-like ligand binding modes in the resulting ensembles are explored.

Templates of higher sequence identity produce more accurate comparative models

We generated 13,000 comparative models of each receptor through minimization and loop building in a sequential fashion as described in the Methods section. To assess the parameters which determine comparative model accuracy in the initial receptor ensemble, we considered the total energy of the models and the sequence identity of the template. Sequence identity was calculated on the aligned GPCR sequences as seen in Figure 44. As shown in Figure 10A, the average root mean square deviation (RMSD) of comparative models built with templates having greater than 50% sequence identity are consistently below 5.0 Å compared to the experimental structure. For residues in the ligand binding pocket (the pocket residues), the average RMSD of comparative models built with templates above 70% sequence identity within the pocket residues are frequently within 2.0 Å of the experimental structure (Figure 10B). In fact, only those targets with templates above 50% sequence identity were able to sample ligand binding pockets within 1.0 Å of the experimental structure (Figure 47).

As can be expected, this includes all target-template pairs within the same sub-family; for example, B1Ar and B2Ar serve as the best templates for each other at 74% sequence identity, as does M2R and M3R at 75% sequence identity. The β-adrenergic receptors also produced accurate models when used as templates for the muscarinic receptors at 48% sequence identity. The opioid receptors produced the most accurate

comparative models when used as templates for each other at 65% to 75% sequence identity. In most cases, templates with high sequence identity also generate the lowest energy models in comparison to models based on other templates (Figure 47). Where there were exceptions, the lowest energy models were produced with templates with at least 45% sequence identity to the target receptor. Without a template having sequence identity above 50%, it continues to be difficult to get accurate models of the ligand binding pocket. While it was demonstrated that building multiple models based on different templates provides a better opportunity to sample the correct conformation and this is leveraged here (Kaufmann & Meiler, 2012), the generation of a smaller but improved conformational receptor ensemble could benefit from using multiple template structures in a single model (Mobarec, Sanchez, & Filizola, 2009). Recently, Worth et al. 2011 demonstrated that similar or improved comparative models could be generated using a multi-template approach, where rotameric states as well as specific sequence and structural features could be modeled in light of the entire set of available experimental structures which otherwise might be absent when using a single template (Worth, Kreuchwig, Kleinau, & Krause, 2011).

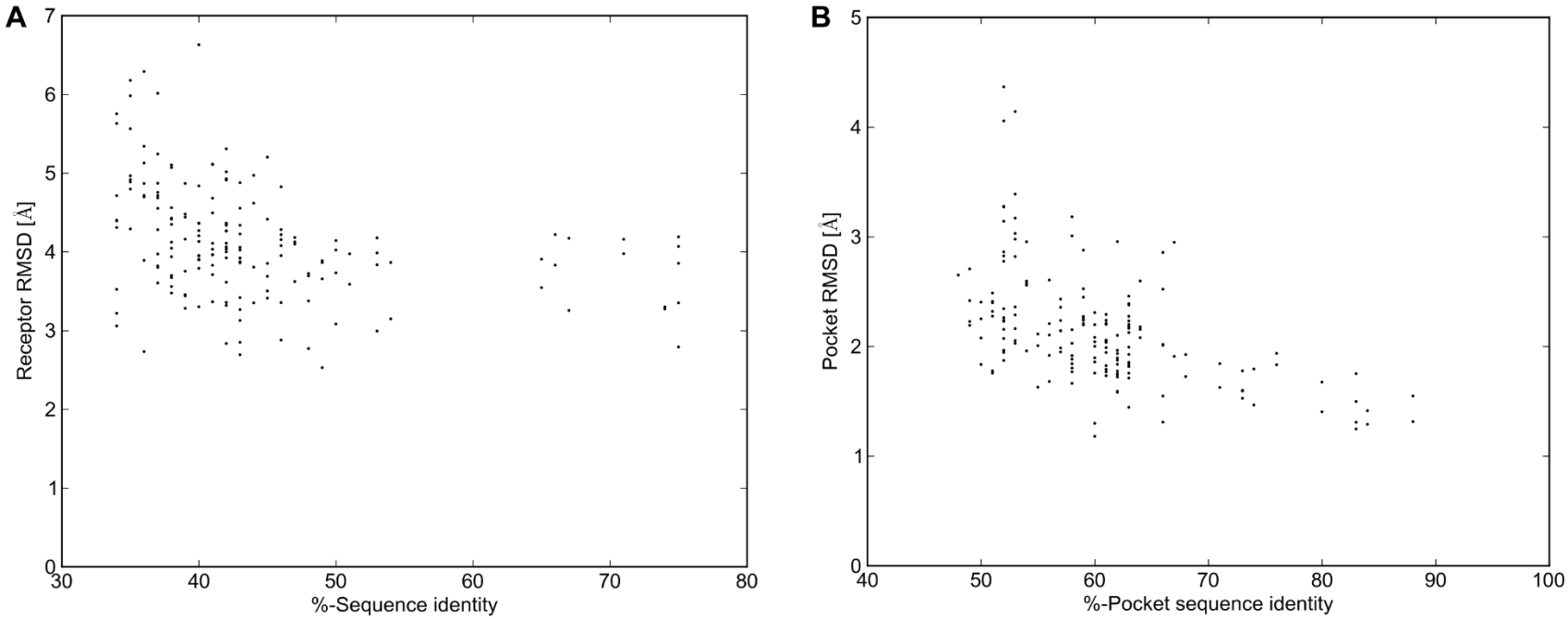


Figure 10
Template sequence identity versus comparative model RMSD.
Each point represents the average RMSD over all comparative models of a target GPCR built using a particular template. For each target-template pair, percent sequence identity was calculated on the sequence alignment shown in Figure 44. Sequence identity is shown here to correlate with low average A) receptor RMSD, calculated over the C-alpha atoms of the full receptor and B) pocket residue RMSD, calculated over the C-alpha atoms of residues within the ligand binding pocket.

Correct helical conformations are recovered in regions of aligned secondary structure

C-alpha RMSD in comparison to the experimental structure was measured for the full receptor, transmembrane region and second extracellular loop (ECL2) region in the lowest energy models and the top ranked models by clustering (Table 5). Among the top ranked models for all 14 receptors, the transmembrane region was modeled with an average RMSD of 2.5 Å compared to their corresponding experimental structures. This average drops to 2.2 Å when considering only those models with template sequence identities above 50%.

The maximum transmembrane region RMSD among top ranked models was seen for CXCR4 at 3.2 Å. In this case, helical placement of TM7 was shifted by six residues between the target sequence and the sequence of the templates, resulting in a gap in the alignment (Figure 44). Without reliable backbone coordinates to model the top of TM7, the resulting models rely on Rosetta to *de novo* fold the region using the CCD loop closure algorithm. The helical structure is recovered, but the top two helical turns of TM7 in the models are displaced from that of the experimental structure by distance of 13.3 Å (Figure 11A).

However, the conformation of the transmembrane helices is reasonably accurate throughout regions where transmembrane helices are aligned, specifically in terms of helical kinks. Deviations from ideal helical conformation are typically caused by proline or glycine residues and are important for both function and structure prediction (Yohannan, Faham, Yang, Whitelegge, & Bowie, 2004). Major helical kinks occur in regions where proline residues are highly conserved between the GPCR sequences, particularly in TM 5, 6 and 7. In the two cases where templates had a proline or glycine-induced kink that was not present in the target, Rosetta was able to remove the kink and recover the correct conformation. The template of the top ranked S1P1R model, D3R, contains a proline at P84 which causes a kink in TM2 that was resolved by Rosetta (Figure 11B). The same is

seen for TM4 in KOR, where the glycine-induced induced kink at G178 in DOR was removed during Rosetta energy minimization to recover the correct conformation (Figure 11C).

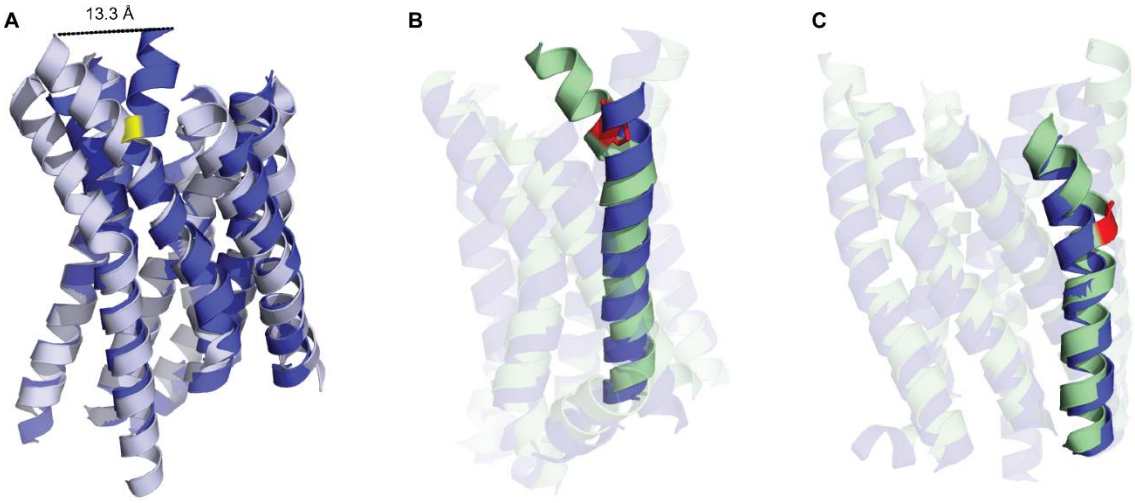


Figure 11
Structural representations of transmembrane helical regions from GPCR comparative models.
A) TM7 in the top ranked comparative model of CXCR4 (blue) deviates from experimental structure (gray), specifically at W283 (highlighted in yellow). Cases where helical kinks exist in the template but are resolved in the comparative model include B) S1P1R, where the top ranked model (blue) resolves the kink in TM2 cause by P84 (highlighted in red) in the D3R template (green) and C) KOR, where the top ranked model (blue) resolves the kink in TM4 caused by G178 (highlighted in red) from the DOR template (green). The top ranked model is the best scoring model of the largest cluster, where clustering is performed on pairwise full receptor C-alpha RMSD over the top ten percent of comparative models by energy.

Table 5

Benchmark results for comparative modeling of G protein-coupled receptors.

Full receptor, transmembrane region and extracellular loop two RMSD over C-alpha atoms compared to the experimental structure is reported for models in each category.

Protein name	Best sampled ECL2 Full receptor/TM region/ECL2 RMSD (Å)	Lowest energy model Full receptor/TM region/ECL2 RMSD (Å)	Top ranked model via clustering[a] Full receptor/TM region/ECL2 RMSD (Å)	Percent of models with ECL2b under 2.0 Å[c]
bRh	4.2 / 1.3 / 3.3	4.7 / 2.5 / 6.2	4.1 / 1.4 / 7.5	0.4
B1Ar	2.8 / 1.6 / 2.9	3.7 / 1.7 / 4.5	3.2 / 1.2 / 4.3	53.9
B2Ar	3.4 / 1.4 / 2.7	3.2 / 1.9 / 5.9	3.7 / 1.7 / 4.4	45.2
A2Ar[b]	-	3.6 / 2.5 / -	3.6 / 2.5 / -	6.8
CXCR4	3.9 / 3.0 / 2.3	4.6 / 3.3 / 3.2	4.2 / 3.2 / 6.5	20.4
D3R	3.9 / 1.9 / 1.8	3.0 / 1.8 / 3.6	3.0 / 1.8 / 3.6	85.2
H1R[b]	-	1.6 / 2.5 / -	2.4 / 2.4 / -	0.8
S1P1R	3.3 / 2.0 / 3.4	5.8 / 2.1 / 4.8	3.6 / 2.0 / 5.4	0.6
M2R	2.2 / 2.4 / 1.9	2.2 / 2.3 / 4.2	2.2 / 2.3 / 4.2	7.0
M3R	3.1 / 2.4 / 2.3	3.1 / 2.9 / 5.3	2.7 / 2.4 / 5.2	9.7
MOR	4.7 / 3.1 / 1.6	3.6 / 1.9 / 4.6	2.4 / 2.8 / 5.8	14.6
KOR	3.7 / 2.6 / 2.7	4.4 / 2.5 / 6.7	3.1 / 2.9 / 5.8	3.8
NOP	2.2 / 2.6 / 2.6	3.2 / 2.8 / 6.6	3.0 / 2.4 / 5.8	10.5
DOR	2.0 / 2.8 / 2.5	3.3 / 2.2 / 5.5	3.3 / 2.2 / 5.5	10.6

[a] Top ranked model is determined by the lowest energy model from the largest cluster.

[b] ECL2 of A2Ar and H1R could not be evaluated because of unresolved structure in this region of the experimental structure in the Protein Data Bank.

[c] ECL2b represents the C-terminal half of ECL2, after the disulfide bond, which contains the residues that contribute to ligand binding as represented in the experimental structures from the PDB

Native-like loop conformations are sampled but are difficult to identify by score

Rebuilding the three extracellular loops was a point of focus during the modeling protocol because of their role in ligand binding. The first and third extracellular loops range in length from five to ten residues, which is within the range of successful loop prediction for Rosetta when applied on experimental structures (Canutescu & Dunbrack, 2009, Chu Wang et al., 2007). Here we find that the first and third extracellular loops are built with an average RMSD of 2.0 Å to the loop conformation from the experimental structures (Table 21). In several cases, identification of the correct loop conformation was possible using the energy of the loop (Figure 48, Figure 50).

The second extracellular loop (ECL2) ranges in length from 16 to 31 residues. While the length of ECL2 is beyond the capability of prediction for loops excised from experimental structures, restriction of the sampling space was provided by requiring formation of the conserved disulfide bonds. The results demonstrate that ranking the most accurate ECL2 is difficult based on energy and clustering, since no top ranked models contained ECL2 RMSDs under 2.0 Å (Table 5 and Figure 49). However, it is possible to sample these native-like loop conformations, which is needed during docking to generate the correct ligand binding mode as observed in the experimental structure. Specifically, loop conformations were sampled below 2.0 Å for MOR, M2R and D3R (Table 5 and Figure 12). When focusing on the C-terminal region of ECL2, which is most often involved in ligand binding, we find native-like sampling for all models, with 0.4% to 85% of sampled ECL2 conformations below 2.0 Å (Table 5).

Accurate ECL2 conformations often recover secondary structure elements

On average, the ECL2 RMSD for top ranked models by clustering was 5.3 Å, with the most accurate ECL2 conformations given for D3R at 3.6 Å, M2R at 4.2 Å and B1Ar at

4.3 Å (Table 5). Compared to the length of other ECL2s, which are about 21 residues long, D3R was relatively the easiest ECL2 to model with only 16 residues (Figure 12A). Other comparative models where ECL2 regions were most accurately sampled were those with secondary structure within the loop. Applying a fragment-based approach to *de novo* loop modeling allows for the insertion of secondary structure into loop regions where it is predicted from the sequence. In four of six cases, Rosetta was able recover helical elements found in ECL2 of experimental structures and one of five cases where β-sheets were found in ECL2 (Figure 51). MOR was the one case where β-sheets were conserved in the model, and the most accurately sampled ECL2 had an RMSD of 1.6 Å (Figure 12B). B1Ar (Figure 12C), B2Ar, M2R (Figure 12D) and M3R were the models in which helical elements were correctly sampled. In cases where predicted secondary structure in the target agrees with that of the template, such as with B1Ar and B2Ar, it would be beneficial to keep the loop conformation of the template and enforce the helical element (Goldfeld, Zhu, Beuming, & Friesner, 2012). The most difficult loop conformations to model were in S1P1R (Figure 12E) and bRh (Figure 12F), where the top ranked models only came within 5.4 Å and 7.5 Å of the experimental structure respectively. Both receptors have ECL2s longer than twenty residues with little secondary structure to stabilize the conformation. Additionally, ECL2 in both receptors packs against the N-terminal region, which was removed prior to comparative modeling. Therefore, inclusion of the N-terminal region into comparative modeling might be beneficial in these cases.

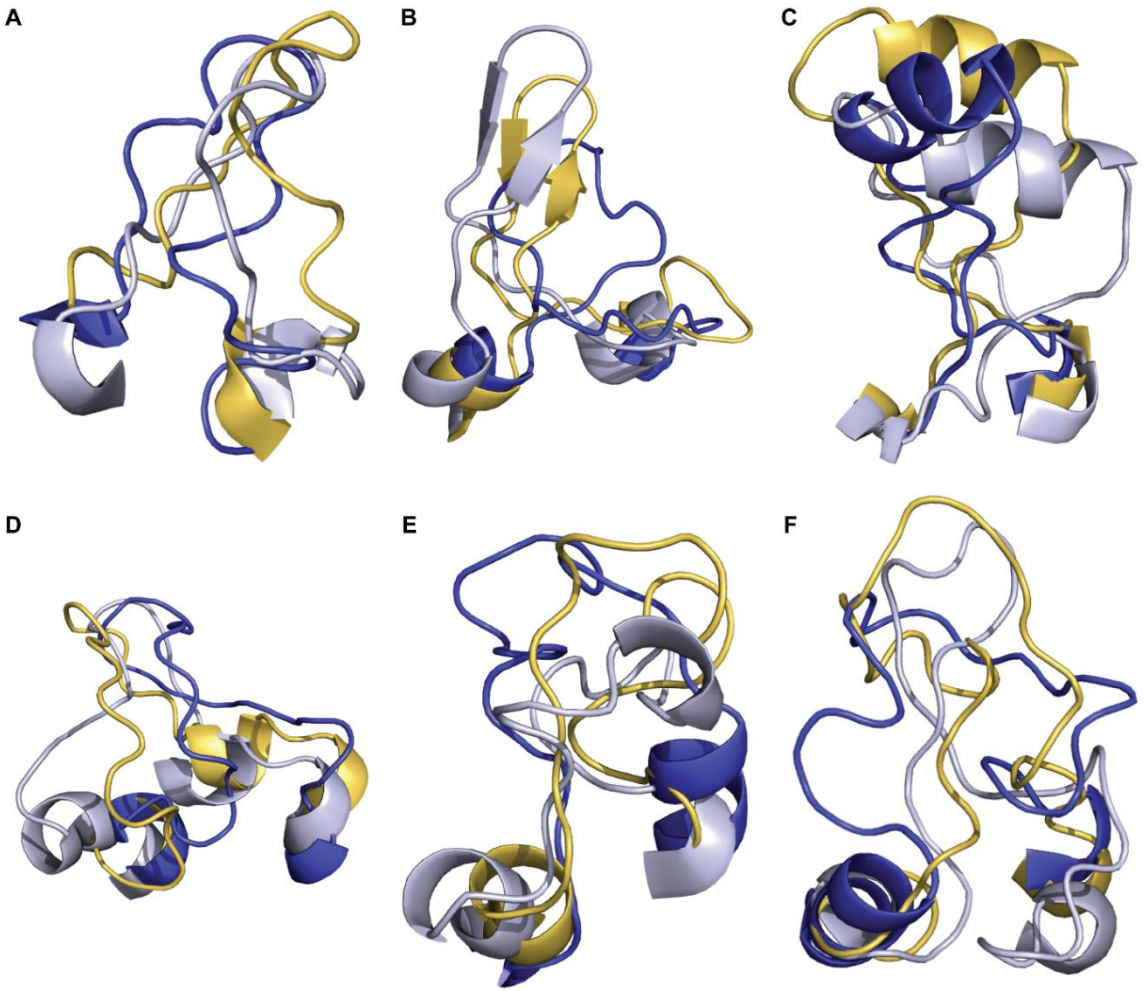


Figure 12
Structural representations of extracellular loop two from comparative models compared to experimental structures.
For A) D3R, B) MOR, C) B1Ar, D) M2R, E) S1P1R and F) bRh, the experimental structure is represented in gray, the most accurately sampled model is represented in yellow and the top ranked model is represented in blue. The top ranked model is the best scoring model of the largest cluster, where clustering is performed on pairwise full receptor C-alpha RMSD over the top ten percent of comparative models by energy.

Comparison with previous studies on GPCR loop modeling

Other studies have likewise addressed GPCR loop modeling. They include the protein local optimization program (PLOP) (Goldfeld, Zhu, Beuming, & Friesner, 2011, Goldfeld et al., 2012), which samples amino acid rotamers in loop regions and evaluates models using a physics-based energy function while explicitly modeling membrane molecules. Modeller uses the CHARMM-22 force field and knowledge-based energy terms to optimize the loop conformation (Fiser, Do, & Sali, 2008). The algorithm employed by Nikiforovich et al. (Nikiforovich, Taylor, Marshall, & Baranski, 2010) performs geometric sampling of the loops using all possible conformations of the peptide backbone. In comparison to their study, Rosetta was able to rank loop conformations in comparative models more accurately than the loop conformations built *de novo* in experimental structures by Nikiforovich et al. (Nikiforovich et al., 2010): bRh was modeled to 7.5 Å RMSD compared to 8.4 Å, B1Ar was modeled to 4.3 Å compared to 6.4 Å, and B2Ar was modeled to 4.4 Å compared to 7.4 Å. In their most recent study, Goldfeld et al. reported top ranked loop conformations built *de novo* in experimental structures as 2.7 Å for B1Ar and 2.2 Å for B2Ar (Goldfeld et al., 2012). However, the algorithm they used enforced the helical bounds within ECL2 for these structures. When considering the results from true *de novo* constructed loop conformations without the helical constraints, top ranked loop conformations from Rosetta are also more accurate than PLOP, whose top ranked ECL2 conformations were 9.1 Å for bRh, 5.6 Å for B1Ar and 13.8 Å for B2Ar.

These results indicate that even current state-of-the-art methods for loop modeling continue to have difficulty determining loop conformations, especially within comparative models. However, the experimental structures which we attempt to reproduce still only represent one of many possible loop conformations for these flexible regions and it is possible that more of the sampled conformations are in fact realistic (Cozzini et al., 2008, Groban, Narayanan, & Jacobson, 2006).

Rosetta captures native-like ligand binding pocket conformations and samples beyond the flexibility evident from experimental structures

To assess the sampling density of residues lining the ligand-binding pocket, we aligned all the models to the experimental structures and measured the collapse of the pocket as the change in distance for each residue C-beta atom (C-alpha for glycine) to the closest ligand atom as determined from the experimental structures (Figure 13). The models display increased flexibility at the top of the transmembrane (TM) helices, as would be expected due to the variability represented by the crystallographic templates. With an average collapse of -0.1 Å and a standard deviation of 3.6 Å within all the comparative models generated, Rosetta samples beyond the flexibility that is represented by the experimental structures, which have an average collapse of -0.1 Å with a standard deviation of 1.0 Å (Figure 13). As the experimental structures are still a small and biased representation of the GPCR space, it is unclear if Rosetta is introducing too much flexibility in these regions. However, for the present study, many comparative models within our ensemble will not make constructive interactions with the ligand due to non-native placement of the residues involved in ligand binding.

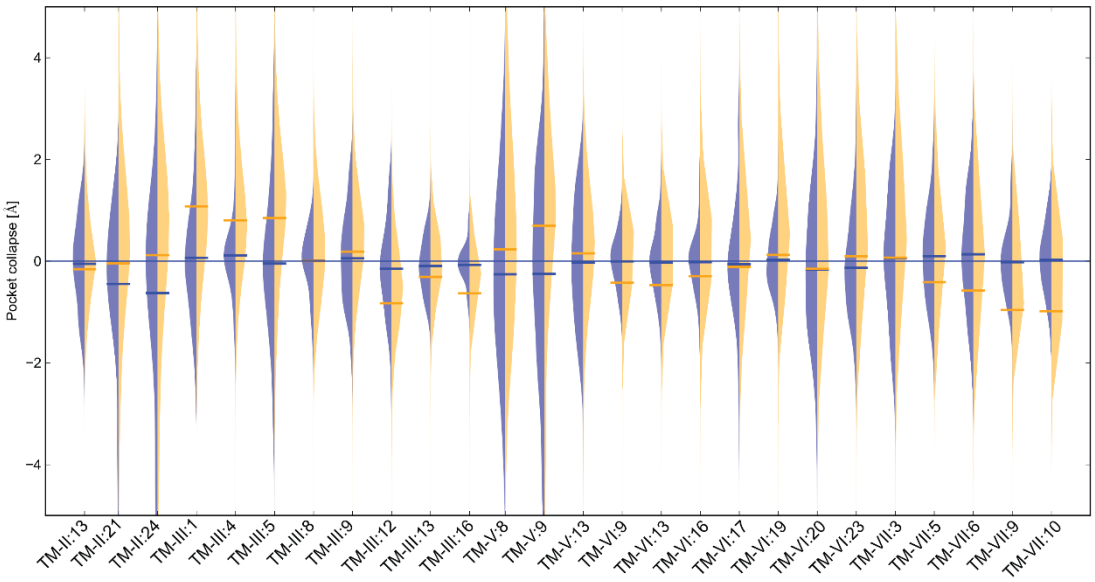


Figure 13
Ligand binding pocket flexibility within comparative models compared to experimental structures.
For each of the 26 pocket residues in the ligand binding pocket of the receptor, pocket collapse is defined as the change in distance from each pocket residue to the ligand, measured between the model and the experimental structure. A positive pocket collapse value indicates that the pocket residue moves closer to the ligand in the model compared to the experimental structure, while a negative pocket collapse value indicates movement towards the receptor. The width of the beanplot area represents the number of models having a pocket collapse of a certain value for was measured for the threaded models (blue) and for all generated comparative models after loop rebuilding and energy minimization (orange), with corresponding .blue and orange horizontal lines representing the average pocket collapse over the given set of models.

Knowledge-based filters improve the accuracy of the ligand-binding pocket

Because Rosetta samples the flexibility of the transmembrane region beyond the variability that is represented in the experimental structures, a knowledge-based filter was created which focused on the pocket residues alone to identify models that would be suitable for ligand docking. Models with structural deviation beyond the maximum flexibility observed within the binding pocket in existing experimental structures were removed, as described in the Methods section. The filter accepted between 0.2% and 10% of the models from the initial receptor ensemble and the overall RMSD of these models were

comparable to those identified by traditional clustering methods (Figure 14A). For several receptors, there was a correlation between pocket RMSD and receptor energy (Figure 47) and based on this correlation, energy was used to reduce the filtered ensemble to a maximum of 100 structures. In this filtered ensemble, the receptor collapse was 0.3 Å with a standard deviation of 0.8 Å, which is slightly greater compared to what is seen in the experimental structures. This is possibly due to favorable energy when collapsing the pocket. The average RMSD of residues constituting the common ligand binding pocket is significantly improved compared to a receptor ensemble selected by clustering of the initial receptor ensemble (Figure 14B).

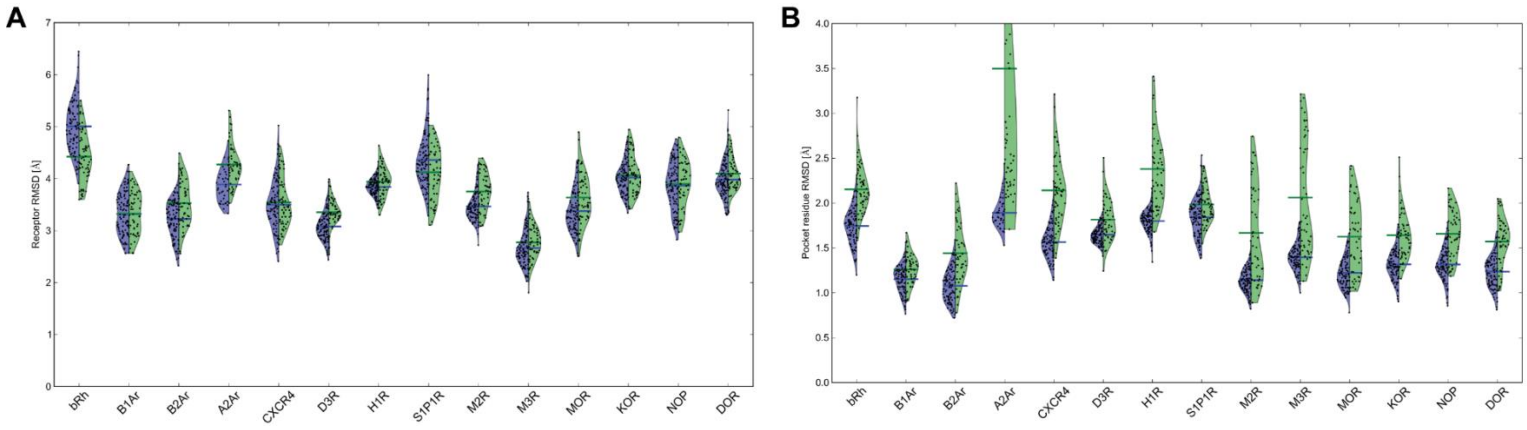


Figure 14
Comparison of two comparative models analysis methods: filtering by ligand binding pocket residues and clustering on RMSD.
For each receptor, the ten lowest energy models of the largest five clusters are represented in green and the one hundred lowest energy models from the knowledge-based filter on residues in the ligand binding pocket are represented in blue. The width of the beanplot area represents the number of models having a particular A) receptor RMSD and B) pocket residue RMSD, with the corresponding horizontal lines representing the average RMSD for all models clustered by RMSD (green) and models from the knowledge-based filter (blue).

Accuracy of the ligand conformer ensemble is highly dependent on ligand flexibility

The generation of ligand conformations is not a trivial process, as the bioactive ligand conformation need not occupy its aqueous GMEC (Boström, Norrby, & Liljefors, 1998, Nicklaus, Wang, Driscoll, & Milne, 1995, Perola & Charifson, 2004). Our approach was to create ensembles of low energy ligand conformations (LMECs) and allow Rosetta to minimize these conformations in the context of a fully flexible receptor. Nevertheless, ligand ensembles will inevitably contain irrelevant conformations which results in the search of irrelevant binding modes. Low energy ligand conformations were generated with MOE and the energy cutoff was determined by the number of rotatable bonds within the ligand (Perola & Charifson, 2004), as discussed in the Methods section.

To evaluate how the use of such ligand ensembles would affect sampling efficiency of ligand binding modes with RMSD below 2.0 Å, the Uniform Sampling Efficiency of binding modes below 2.0 2.0 Å (USE2.0) was calculated for all ligand ensembles and compared to that of the bioactive ligand conformation. While the uniform sampling efficiency dropped by only 6.0 ± 3.8 fold for the majority of the ligands by using ligand conformers instead of the bioactive ligand conformation, it dropped by 68.1, 50.6, 42.4 and 27.2 fold for IT1t, C-24, ML056 and beta-FNA respectively (Table 4 and Figure 52). These ligands are characterized by many degrees of freedom, which contribute to the difficulty of sampling the bound conformation accurately.

While a benchmark of ligand ensemble generation methods was beyond the scope of this study, we noted some reduction in the number of non-native ligand conformers by using the Generalized Born implicit solvent instead of the distance-dependent dielectric constant. Further improvement might be possible in some cases by using MD-simulations to generate a canonical ligand ensemble weighted according to the Boltzmann distribution to identify the most populated and thereby most probable ligand conformations. For large, flexible ligands, a fragment-based docking approach might be more suitable and has

already been applied in many drug design studies (A Kumar, Voet, & Zhang, 2012, Ashutosh Kumar & Zhang, 2012, Mortier, Rakers, Frederick, & Wolber, 2012).

Interaction energy is not reliable for identification of the experimental ligand binding mode when docking into flexible GPCR comparative models

The bioactive ligand conformation from the experimental structure as well as ligand conformers generated by MOE were docked into the top ranked comparative models as evaluated by clustering and the knowledge-based filter. While Rosetta considers the energy of the receptor while sampling ligand binding modes, noise generated by the multiple loop conformations makes it difficult to identify low energy binding modes using the total Rosetta energy for the receptor-ligand complex. As a result, we choose to make the assumption that all structures from the comparative model ensemble have equal energy and accuracy when docking the ligands. Thus the local minimum interaction energy conformation (LMIEC) with the lowest energy that we sample needs to equate the energy of the experimental ligand binding mode in order to be useful for its identification. To test the extent to which this occurs, the bioactive ligand conformation was re-docked into the experimental structure it came from with no relaxation of the complex after docking (Figure 53). In 13 out of 14 cases we find that the lowest energy LMIEC was within 2.0 Å of the experimental ligand binding mode and that other LMIECs were significantly separated from the lowest energy LMIEC by 4.6 ± 3.0 Rosetta Energy Units (REU). The one case where the lowest energy LMIEC deviated from the experimental ligand binding mode is IT1t in CXCR4, which has exceptionally few interactions to the protein, 7 water molecules within 4 Å, and contacts a residue in the N-terminal region, which is not represented in our models.

When repeating the protocol with the addition of flexibility within the receptor through a minimization step, the lowest energy LMIEC deviated from the experimental

ligand binding mode in 6 out of 14 cases and with an insignificant energy gap of 0.3 ± 2.1 REU, showing that lowest energy LMIEC is not suitable for identification of the experimental ligand binding mode in flexible models. However, while the ligand binding mode within a RMSD of 2.0 Å to the experimental structure could not be identified consistently by interaction energy in this highly biased analysis, it was possible to sample the correct binding mode in all 14 cases (Table 6).

Table 6

Sampling efficiency for ligand docking results.

Reported is the negative log of the sampling efficiency of ligand binding modes within 2.0 Å RMSD of the bioactive ligand conformation within the experimental structure as measured over the ligand heavy-atoms.

Protein name	Bioactive ligand docked to experimental structure, no minimization, n=1000[1]	Bioactive ligand docked to experimental structure, minimized, n=1000[1]	Bioactive ligand docked to top models from knowledge-based filter, n=2000[1]	Bioactive ligand docked to top models from clustering by RMSD, n=6000[1]	Ligand conformers docked to top models from knowledge-based filter, n=10000[2]	Ligand conformers docked to top models from clustering by RMSD, n=6000[2]
bRh	0.05 ($5 \cdot 10^4$)	0.04 ($5 \cdot 10^4$)	1.94 ($7 \cdot 10^2$)	2.40 ($2 \cdot 10^2$)	2.66 ($5 \cdot 10^2$)	3.52 (70)
B1Ar	0.03 ($5 \cdot 10^4$)	0.02 ($5 \cdot 10^4$)	2.05 ($5 \cdot 10^2$)	2.11 ($4 \cdot 10^2$)	2.18 (10^3)	2.07 (10^3)
B2Ar	0.03 ($4 \cdot 10^4$)	0.03 ($4 \cdot 10^4$)	2.19 ($3 \cdot 10^2$)	2.10 ($3 \cdot 10^2$)	2.27 (10^3)	2.40 ($8 \cdot 10^2$)
A2Ar	0.08 ($9 \cdot 10^3$)	0.81 ($2 \cdot 10^3$)	2.85 (20)	2.92 (10)	3.22 (10^2)	3.70 (30)
CXCR4	0.54 (10^4)	0.96 ($5 \cdot 10^3$)	2.68 (10^2)	2.62 (10^2)	4.00 ($3 \cdot 10^2$)	ND*[#]
D3R	0.04 ($4 \cdot 10^4$)	0.08 ($3 \cdot 10^4$)	2.22 ($3 \cdot 10^2$)	2.40 ($2 \cdot 10^2$)	2.59 ($7 \cdot 10^2$)	2.70 ($6 \cdot 10^2$)
H1R	0.01 ($3 \cdot 10^4$)	0.03 ($3 \cdot 10^4$)	2.07 ($3 \cdot 10^2$)	1.85 ($5 \cdot 10^2$)	2.17 ($7 \cdot 10^2$)	2.15 ($8 \cdot 10^2$)
S1P1R	0.05 ($7 \cdot 10^4$)	0.07 ($7 \cdot 10^4$)	3.30 (40)	2.40 ($3 \cdot 10^2$)	4.00 ($4 \cdot 10^2$)	3.30 ($2 \cdot 10^3$)
M2R	0.02 ($4 \cdot 10^4$)	0.02 ($4 \cdot 10^4$)	1.84 ($6 \cdot 10^2$)	2.38 ($2 \cdot 10^2$)	2.57 ($7 \cdot 10^2$)	2.74 ($5 \cdot 10^2$)
M3R	0.03 ($3 \cdot 10^4$)	0.03 ($3 \cdot 10^4$)	1.94 ($4 \cdot 10^2$)	2.26 ($2 \cdot 10^2$)	2.44 ($8 \cdot 10^2$)	2.74 ($4 \cdot 10^2$)
MOR	0.14 ($2 \cdot 10^4$)	0.35 (10^4)	2.02 ($3 \cdot 10^2$)	2.27 ($2 \cdot 10^2$)	2.43 ($3 \cdot 10^3$)	3.40 ($3 \cdot 10^2$)
KOR	0.04 ($2 \cdot 10^5$)	0.03 ($2 \cdot 10^5$)	2.07 ($2 \cdot 10^3$)	2.55 ($7 \cdot 10^2$)	3.10 ($2 \cdot 10^3$)	2.92 ($3 \cdot 10^3$)
NOP	0.11 ($5 \cdot 10^4$)	0.11 ($5 \cdot 10^4$)	2.24 ($4 \cdot 10^2$)	2.15 ($5 \cdot 10^2$)	3.52 (10^3)	3.70 ($7 \cdot 10^2$)
DOR	0.08 ($2 \cdot 10^4$)	0.07 ($2 \cdot 10^4$)	1.51 ($8 \cdot 10^2$)	1.77 ($5 \cdot 10^2$)	1.57 (10^3)	1.82 ($6 \cdot 10^2$)

[*] denotes where sampling efficiency of Rosetta is worse than the worst-case uniform sampling scenario.

[#] ND denotes not defined. No binding modes within 2.0 Å were sampled for this case.

[1] fold improvement over USE2.0 of bioactive ligand

[2] fold improvement over USE2.0 of ligand conformers

Sampling of native-like ligand binding modes is on average 10^3 fold increased over random sampling

Having generated receptor-ligand complexes through docking, we asked how frequently the experimental ligand binding mode was sampled within an RMSD of 2.0 Å and compared this to the sampling efficiency that can be achieved using uniform sampling in a 5.0 Å docking sphere with no protein (USE2.0). In the receptor ensemble that was selected based on the knowledge-based filter, we found that the experimental ligand binding mode was sampled in all cases with an average of 10^3 fold increase over USE2.0 (Figure 15 and Table 6). The correct binding mode for S1P1R and CXCR4 was sampled correctly least often with only one correct binding mode out of approximately 10^4 generated models. The reason for the difficulty of sampling the correct S1P1R ligand binding mode is most likely related to its flexibility and its contacts to the N-terminus, which is lacking in our models. The low number of ligand-protein contacts in the model seems to be the main reason for the poor sampling efficiency of IT1t in CXCR4, which as discussed above, was not in an interaction energy minimum when its bioactive conformation was docked into a backbone static receptor. Even so, sampling the experimental ligand binding mode within 2.0 Å RMSD was increased 300 fold over USE2.0 for both S1P1R and CXCR4, demonstrating preference of biologically relevant ligand-protein interactions during the docking procedure.

On the other end of the spectrum of sampling efficiency is DOR, which sampled the correct binding mode in 266 out of approximately 10,000 cases – 10^3 times better than USE2.0. The ligand in DOR, Naltrindole, has only 2 degrees of freedom, with all conformers below 1.0 Å of the bioactive ligand conformation (Table 4), and binds to the receptor mainly through hydrophobic contacts and one salt bridge. For all other cases, docking multiple ligand conformations into the comparative models sampled binding modes within 2.0 Å of the experimental binding mode less than 1% of the time (Figure 15).

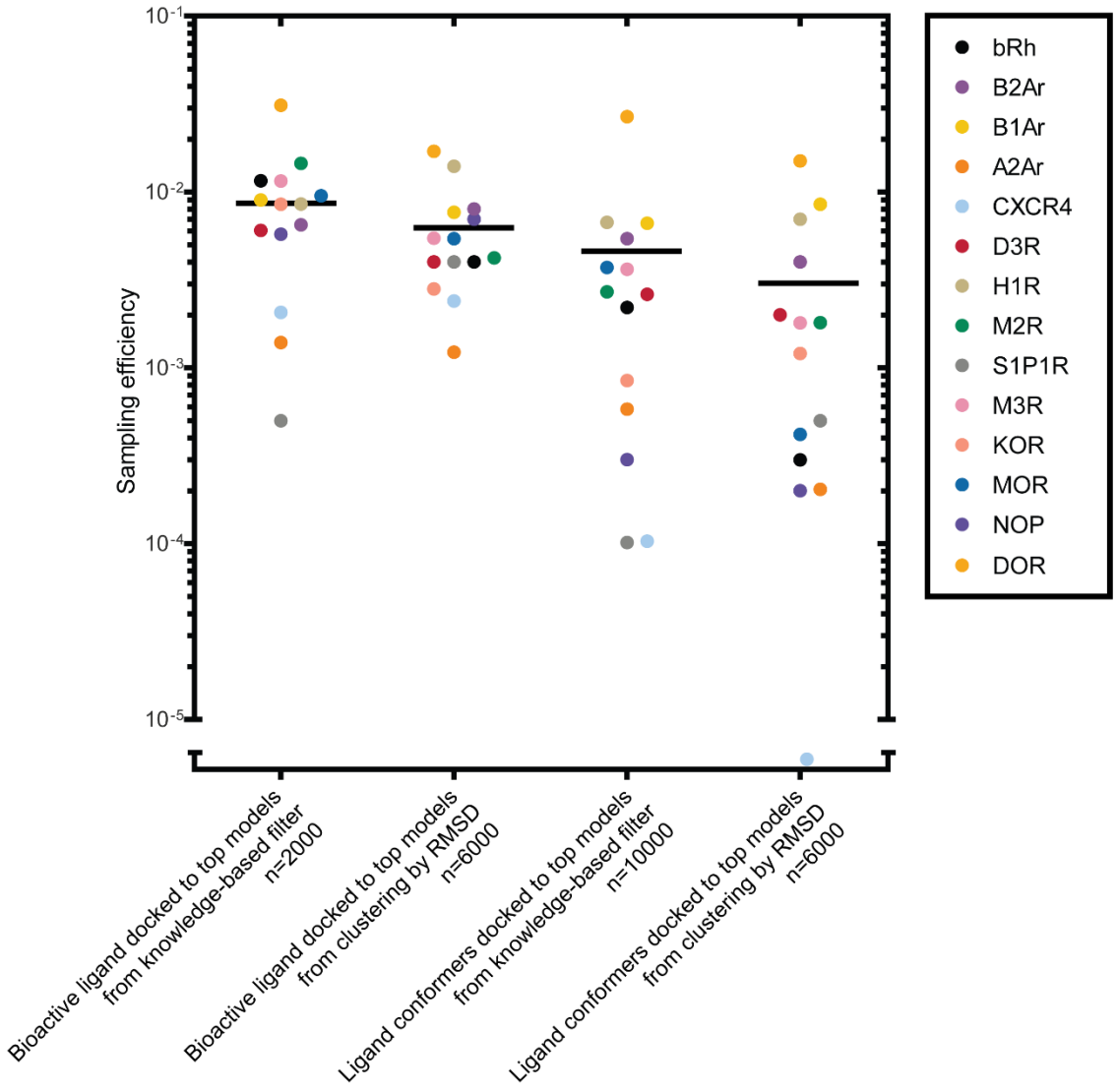


Figure 15
Sampling efficiency for docking into comparative models.
For each receptor, the fraction of binding modes sampled within 2.0 Å of the experimental binding mode is represented for docking the bioactive ligand and ligand conformers into the top models chosen by the knowledge-based filter and clustering by RMSD. The average sampling efficiency for each method is represented by the black solid line.

Sampling of native-like ligand binding modes improve within the knowledge-based filtered comparative model ensemble

To assess the effect of the knowledge-based filter we compared the sampling efficiency in models selected with the knowledge-based filter receptor ensemble with

those chosen by traditional clustering methods and found that sampling efficiency is improved in 10 out of the 14 cases (Table 5 and Figure 15). Additionally, as an attempt to identify which parameters of the receptor models yielded native-like ligand binding modes, we examined the importance of the pocket residue RMSD in Figure 16. Accuracy between the C-alpha atoms of the pocket residues does not guarantee accurate ligand placement, as side chain placement varies greatly and creates many non-native binding pockets. Also, given the flexibility of the ligand conformations, it is expected that the docking algorithm detects alternate binding modes within a particular binding pocket conformation. Despite this, we show that more accurate placement of the residues within the ligand binding pocket leads to more binding modes sampled within 2 Å of the experimental ligand binding mode using the knowledge-based filters and templates of high sequence identity (Figure 15, Figure 54). Importantly, while we show a correlation between pocket RMSD and ligand RMSD, the same effect cannot be shown when selecting receptor models using full receptor RMSD based clustering. This is likely due to the irrelevant noise that arises from non-pocket residues. Interaction energy enriches for experimental ligand binding modes.

Despite the lack of robustness in the use of interaction energy to identify the correct binding modes in relaxed experimental structures, we expected that it might be useful for enrichment of correct binding modes in our docking ensembles by removing obvious non-fit ligand-protein interactions. We found that an enrichment of approximately three fold can be achieved in most cases by taking the 10% best scored structures, as shown in Figure 17. However, when taking the top 10% of structures for bRh, CXCR4 and M3, sampling efficiency dropped. There was no correlation between optimal cutoff value and the overall sampling efficiency or the size of the largest cluster (data not shown).

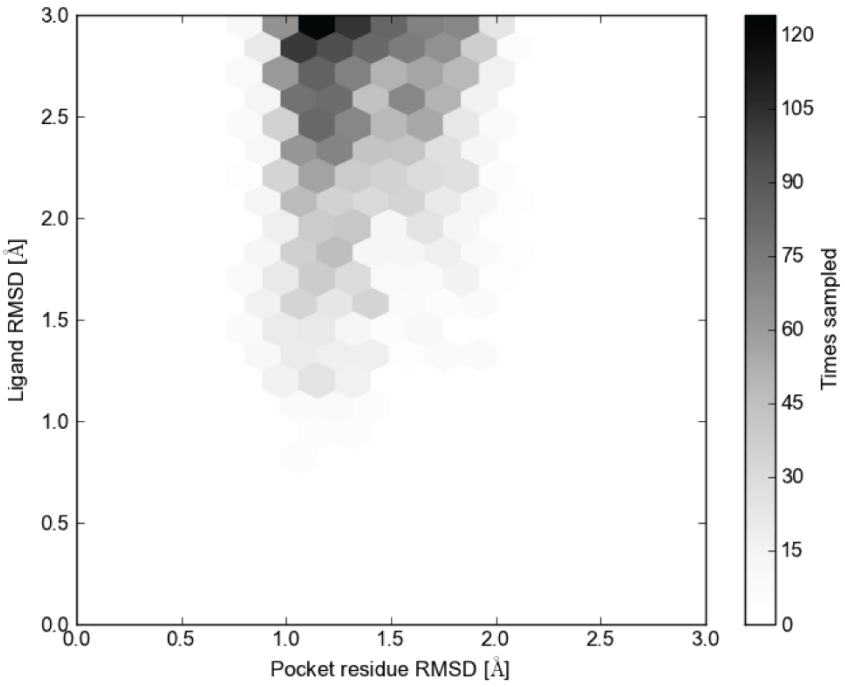


Figure 16
Sampling density of ligand binding modes versus pocket residue RMSD.
The number of binding modes within the given RMSD of the experimental ligand binding mode is given for each pocket residue RMSD.

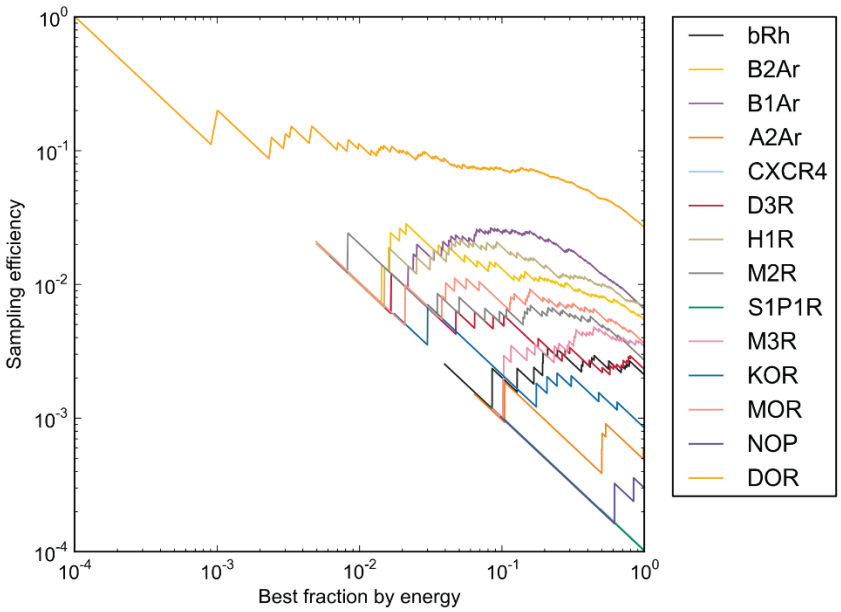


Figure 17
Enriching sampling efficiency with energy cutoffs.
The sampling efficiency for binding modes sampled within 2.0 Å of the experimental binding mode at each fraction of comparative models selected by Rosetta interaction energy is presented for each receptor.

Clustering aids in selecting native-like models

In spite of the low sampling efficiency of the experimental ligand binding modes, we hypothesized that clustering mediated through total energy optimization during docking might occur, and thus offer a method to identify native-like ligand binding modes. Notably, however, the Rosetta ligand docking algorithm does not in principle generate a Boltzmann distribution, but instead emphasizes sampling of rare binding modes, in hope of identifying a rare native-like global minimum interaction energy conformation (Lemmon & Meiler, 2012). This might blur any tendencies for clustering around an experimental ligand binding mode. Clustering was performed on the heavy-atom ligand RMSD with a cutoff of 2.0 Å and the lowest energy binding modes of the largest five clusters were examined further. Other cutoffs of 2.5, 3.0 and 3.5 Å, were also considered, but did not provide any improvement of clustering performance. The percentage of models in the largest clusters was below 1% for most receptors (Figure 18). For the receptors in which convergence occurred, however, there was some correlation between cluster size and ligand RMSD (Figure 18A).

Within the largest clusters for each receptor, there was on average 12% coverage and 10% accuracy of the correct contacts between the ligand and receptor (Table 4 and Figure 18B). For CXCR4, KOR and NOP, alternate modes are preferred over the experimental binding mode. In examining cases where the experimental binding mode is not preferred, several problems are identified which keeps the ligand from binding in the correct mode. For CXCR4 and NOP, less than 30% of the ligand conformers came within 2.0 Å of the bioactive ligand conformation, resulting in inaccurate docking results. For ligands binding high within the receptor binding sites such as A2Ar, incorrect loop placement in the models blocks the ligand from docking in the correct mode (Figure 19A). Incorrect loop placement can also induce hydrogen bonds favorable to the ligand which move it into an incorrect binding mode, as shown in KOR (Figure 19B). For many ligands,

Rosetta places the ligand in the correct position but is unable to discriminate the correct interactions and flips the ligand orientation as seen in H1R (Figure 19C), indicating possible inaccuracies within the force field and improper treatment of polar interactions. There were two cases, DOR and M2R, in which Rosetta was able to identify the correct binding mode within 2.0 Å in the top ranked clusters (Table 7), shown in Figure 19D and 19E. Docking Naltrindole in DOR and 3-quinuclidinyl-benzilate in M2R was simplified by the limited number of rotatable bonds in the ligand and high sequence identities of the templates.

Table 7
Top ranked binding modes for ligands docked into G protein-coupled receptor comparative models.

Protein name	Cluster Rank[a]	Ligand RMSD	Coverage of correct contacts[b]	Accuracy of correct contacts[c]
bRh	5	3.4	0.11	0.10
B1Ar	2	2.13	0.60	0.50
B2Ar	2	3.48	0.11	0.10
A2Ar	5	3.48	0.22	0.15
CXCR4	3	6.23	0.04	0.04
D3R	4	2.26	0.20	0.14
H1R	5	4.81	0.10	0.10
S1P1R	2	3.37	0.12	0.15
M2R	4	1.86	0.40	0.30
M3R	2	3.5	0.11	0.10
MOR	1	2.7	0.18	0.11
KOR	4	5.67	0.01	0.01
NOP	1	6.72	0.02	0.01
DOR	1	1.78	0.37	0.21

[a] The lowest energy binding mode from the largest 5 clusters, determined by heavy-atom ligand RMSD with a cutoff of 2 Å, was used for evaluation. Given here is the cluster rank for the lowest ligand RMSD of the top 5 ranked models.
[b] Coverage of correct contacts was calculated with SimiCon (Rueda et al., 2010) and is the number of correct contacts divided by the total number of ligand contacts made in the experimental structure.
[c] Accuracy of correct contacts was calculated with SimiCon (Rueda et al., 2010) and is the number of correct contacts divided by the total number of ligand contacts made in the model.

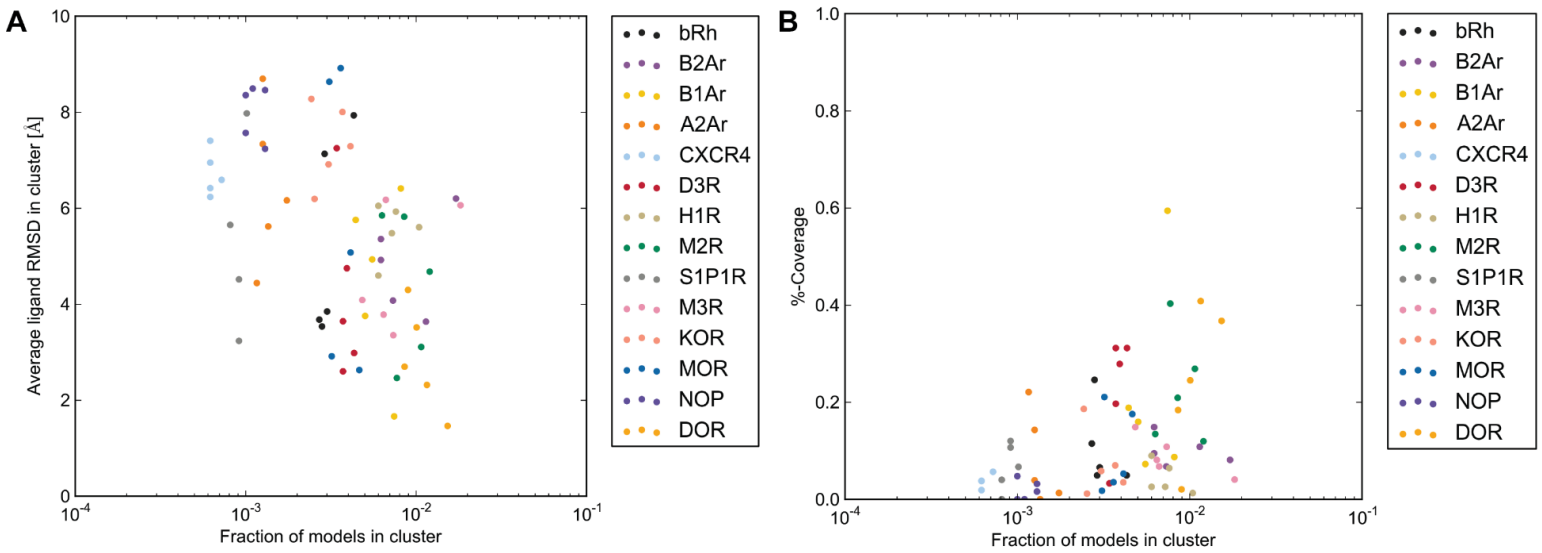


Figure 18
Clustering captures binding modes with lower RMSD and increased contact coverage.
Binding modes for each receptor were clustered by ligand heavy-atom RMSD with a cutoff of 2.0 Å. When compared to smaller cluster sizes, the large cluster sizes were more likely to capture A) lower average ligand RMSD to the experimental binding mode and B) a higher percentage of correct ligand contacts. Contact coverage was calculated using SimiCon (Rueda et al., 2010).

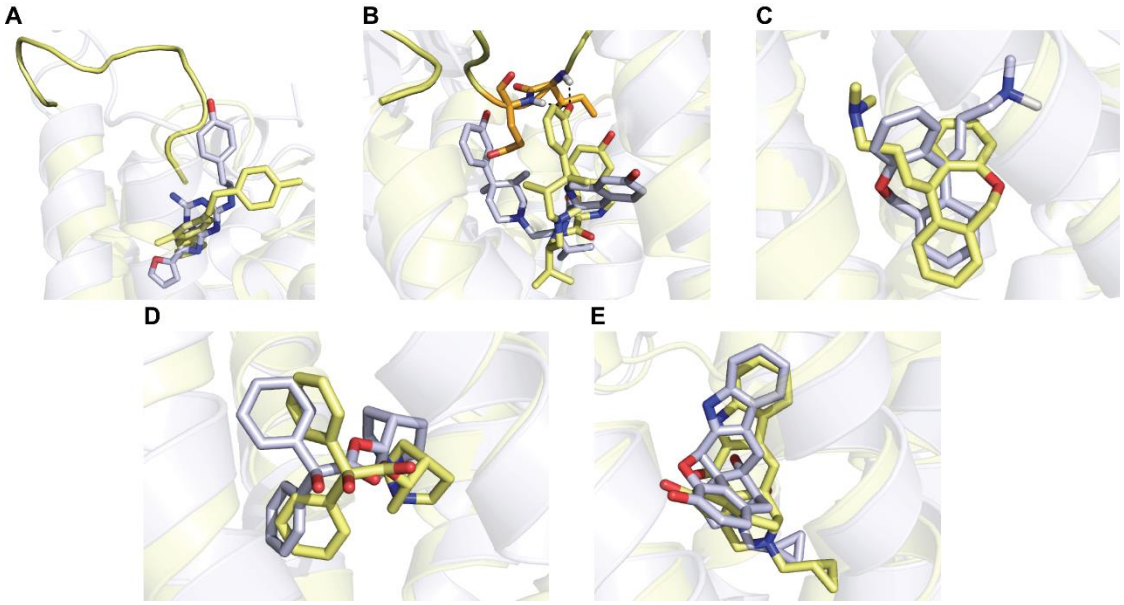


Figure 19
Structural representations of ligand binding modes compared to experimental structures. Incorrect loop placement and incorrect ligand orientation often prevent Rosetta from converging on the experimental ligand binding mode. Ligand binding modes from the experimental structures are shown in gray and the top ranked model via clustering by ligand RMSD is shown in yellow for A) A2Ar, B) KOR and C) H1R. Cases where top ranked binding modes captured the experimental binding mode within 2.0 Å were D) DOR and E) M2R.

Comparison with previous studies on ligand docking into GPCR comparative models

Using Glide (Friesner et al., 2004) and Induced Fit Docking (Sherman, Day, Jacobson, Friesner, & Farid, 2006) to dock ligands within biased comparative models of GPCRs, Beuming and Sherman (Beuming & Sherman, 2012) ranked ligand binding modes within 2.5 Å of the experimental ligand binding mode in six out of the ten receptors they modeled. In these six cases, success was likely due to the structural similarity of the templates, which always came from receptors of the same sub-family: β-adrenergic receptors were used as templates for each other and for H1R and the muscarinic receptors were used as templates for each other. Alignments were manually refined to ensure correct alignment of loop regions and the disulfide bridge within ECL2. Only regions with missing density according to the alignment were rebuilt using PLOP (Goldfeld et al., 2012).

Additionally, the ligand from the template structure remained within the model during the comparative modeling process, which may have assisted in the preservation of the ligand binding pocket. While remaining relatively unbiased in sequence alignment and *de novo* loop rebuilding, Rosetta was able to sample binding modes within 2.5 Å of the experimental ligand binding mode in all cases. However, inaccuracies in the energy function and flexibility introduced within the pocket residues made it difficult to identify native-like binding modes as top ranked. As discussed above, Rosetta had success in ranking the correct binding mode only in the cases of M2R and DOR.

Sampling efficiency is increased by one order of magnitude for every 10 known ligand-protein contacts

Docking into comparative models guided by mutational data is a widespread and largely non-validated method in the literature. Typically, side-chain alterations that heavily affect ligand binding are interpreted as having direct contacts to the ligand. To assess how such information can be used as experimental constraints in our ligand-protein ensembles, we tested to which extent these constraints would allow us to detect the correct binding mode. Enrichment of the correct binding modes was determined through 10,000 iterations of randomly chosen contacts between 0 and total number of all 4.0 Å contacts between the ligand and receptor. When docking ligand conformers into comparative models, the sampling efficiency for native-like binding modes increased on average by one log scale for every 10 known contacts assumed between the binding mode and the receptor (Figure 20). The greatest improvement was seen for receptors where sampling efficiency of the experimental binding mode was already above 0.1%, particularly for DOR, NOP and B1Ar. Little or no improvement in sampling efficiency was observed for those receptor ensembles already sampling less than 0.1% of the experimental binding mode, including A2Ar, B2Ar, S1P1R and CXCR4. Experimental data with higher information density, such

as the ionic interactions used for blinded prediction of the binding mode of Eticlopride in the dopamine D3 receptor, can be expected to provide a significantly higher improvement in sampling efficiency – in our ensemble, the sampling efficiency was improved by 7 fold by requiring a distance of less than 3.0 Å between the positively charged hydrogen atom on the tertiary amine and the oxygen atoms in the carboxyl acid group of the aspartic acid.

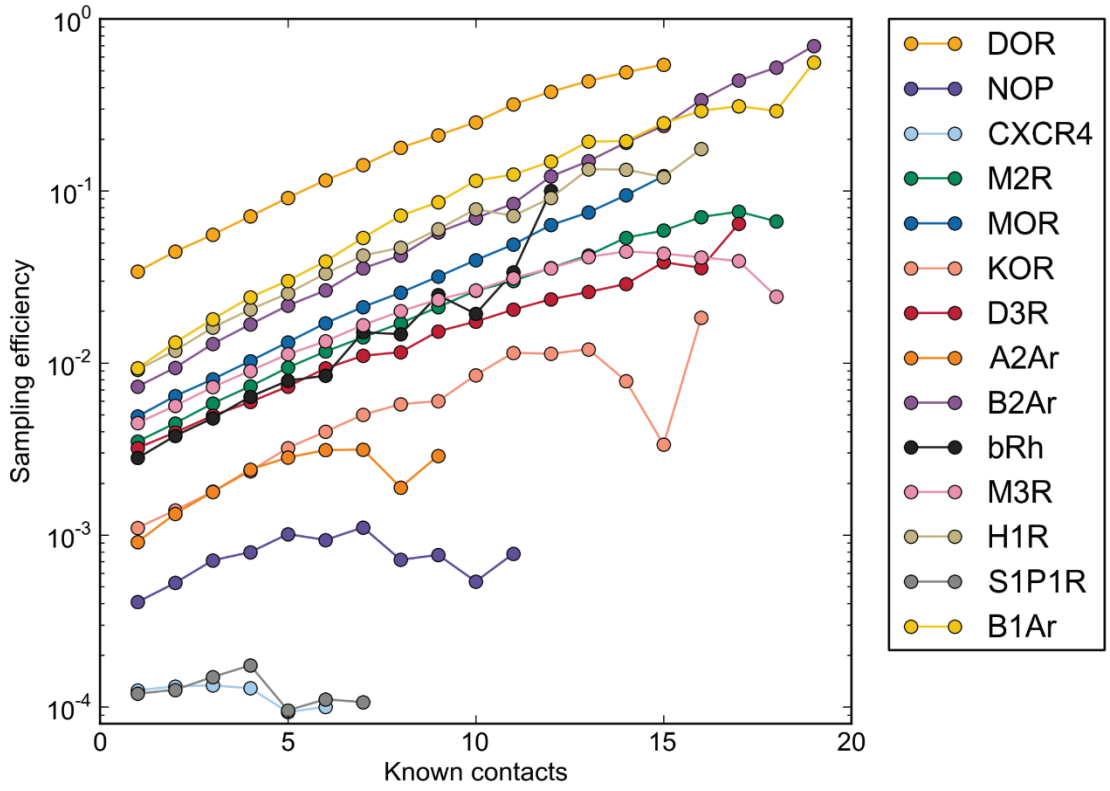


Figure 20
Enriching binding modes with known receptor-ligand contacts increase sampling efficiency of native-like binding modes.
For binding modes generated for each receptor, a random number of known contacts from 0 to the greatest possible number of contacts were chosen for 50,000 iterations and the fraction of binding modes sampled within 2.0 Å of the experimental binding mode is given.

Concluding Remarks

This study provides an analysis of the sampling performance that can be expected when docking ligands into comparative models of GPCRs. Previous studies of ligand

docking into GPCR comparative models have demonstrated that the problem is highly challenging for all but the simplest of cases that require the least sampling of receptor space (Beuming & Sherman, 2012, Kufareva et al., 2011, Michino et al., 2009). This is in agreement with recent docking studies for flexible ligand docking into multiple static structures (Barril & Morley, 2005, Bottegoni, Kufareva, Totrov, & Abagyan, 2008, Rueda, Bottegoni, & Abagyan, 2009), which consistently report that while the performance in docking and screening accuracy of a 'small' conformational ensemble is superior to that of a single conformer, that performance starts to rapidly decline when the size of the conformational ensemble begins to grow. The study presented here sought to quantify the challenges of docking ensembles of ligand conformers into comparative models through deep and relatively unbiased sampling using full receptor and ligand flexibility.

Comparative models of 14 unique GPCRs were constructed using the other 13 experimental structures as templates. Threading was based on the unbiased alignment between the target and template sequences and loops were constructed *de novo* with a fragment-based loop closure algorithm in Rosetta. When compared to corresponding experimental structures, the most accurate comparative models demonstrated a correlation to Rosetta energy. Top ranked structures with templates within 50% sequence identity were modeled with an average RMSD of 2.2 Å in the transmembrane region, with the best models coming within 1.2 Å RMSD. Extracellular loops with lengths ranging between 5 and 7 residues were modeled with an average RMSD of 2.0 Å, while ECL2 was modeled with an average RMSD of 5.3 Å. The most difficult cases to model were those in which helical regions were unable to align to suitable templates and those cases in which N-terminal residues necessary for ECL2 packing were missing. Despite these challenges, Rosetta was still able to rank more accurate loop conformations than other leading methods.

Using the ligands found in the crystallized GPCR structures, docking was performed in the top ranked comparative models. Docking ensembles of ligand conformers into comparative models sampled the correct ligand binding mode for each of the 14 receptors, but often less than 1% of the time. While the lack of energy gap makes discrimination of the correct binding modes difficult, certain techniques for filtering the models and binding modes demonstrated some success in this study. Using templates with a sequence identity above 50% provides a higher chance for correctly modeling of the ligand binding pocket as also observed in previous studies (Beuming & Sherman, 2012). In cases where such templates do not exist, using a knowledge-based filter to identify models for which the binding pocket is within the variability that is represented in the experimental structures is beneficial for docking, significantly increasing sampling efficiency in 10 of the 14 cases. Inaccuracies in the minimized structures strongly affected the accuracy in the loop regions, which in turn affected the resulting ligand binding modes. Therefore, it may be best to limit the flexibility introduced by sampling when using a highly homologous template, such as the case for B1Ar and B2Ar.

As demonstrated in this study, clustering can provide improvement over energy for identifying correct binding modes, but only if clusters contain at least 1% or more of the total binding modes. Selection of the correct binding mode from an ensemble of models might be further improved using information from structure activity relationship of active ligands, as proposed by Katritch et al., to select the best performing models from an ensemble (Katritch, Rueda, & Abagyan, 2012). However, this requires knowledge about active ligands which is typically limited for novel protein or receptor targets and the approach is based on the assumption that different ligands share a common binding mode.

In addition, many papers are published under the premise that experimental information such as mutational data can aid in finding the correct ligand binding mode

within a large ensemble of models (Gelis, Wolf, Hatt, Neuhaus, & Gerwert, 2012, Hoyer et al., 2013, Jacobson, Jayasekara, & Costanzi, 2012, Parry, Chen, Andrews, Lears, & Rogers, 2012, C.-D. Wang, Buck, & Fraser, 1991). Considering the challenges faced in this study, application of mutational data as experimental constraints seems to be an appealing strategy. Sampling efficiency for binding modes within 2.0 Å of the experimental ligand binding mode increased on average by one log scale for every 10 known contacts between the binding mode and the receptor. However, the expected benefit should be evaluated carefully on one or more experimental ligand-receptor complexes to access the true value of such constraints – in particular since indirect effects are known to occur and could blur the identification and the selection of the “correct” binding modes.

Through the use of unbiased sequence alignments and sampling algorithms using the Rosetta software suite, the most challenging scenario for GPCR comparative modeling and ligand docking was explored. As with other studies on comparative modeling and docking, however, there were still minor biases introduced in both aspects of this work which may limit the scope of this approach. Bias in the comparative modeling experiments included the addition of constraints on the disulfide connectivity of the loops based on the experimental structures, which influenced the conformations of ECL2. Bias in the ligand docking experiments included the ligand stereochemistry and protonation state of CXCR4, restricting the conformational search space by centering a sphere of 5Å radius at the center of the experimentally determined binding mode. Also, bias to the experimentally determined structures could have been eliminated with a leave-one-out cross-validation of the knowledge based filter. Despite these biases, the findings of this study identified specific avenues for improvement to approach this challenging problem. Knowledge-based and energy-based filters are able to improve sampling performance over random by 10^3 fold. Additionally, sampling performance is increased by one order of magnitude for every 10 residues known to contact the ligand. Contacts with high information density,

specifically the salt bridge between the oxygen atoms of an aspartic acid in DOR and the positively charged hydrogen atom on the tertiary amine of its ligand, improved sampling efficiency 7 fold. As the number of GPCR experimental structures being determined increases, so does the opportunity to find suitable templates for comparative modeling. With the guidelines suggested by the results from this study, relevant ligand docking studies may be able to generate structural hypotheses to guide experimental designs.

Acknowledgments

We thank Sarel J. Fleishman (Weizmann Institute of Science, Israel) for valuable discussions on the development of the uniform sampling efficiency of ligand binding modes below 2.0 Å (USE2.0), a method for establishing a null model for sampling efficiency in ligand docking.

CHAPTER IV

PROBING THE METABOTROPIC GLUTAMATE RECEPTOR 5 (MGLU5) POSITIVE ALLOSTERIC MODULATOR (PAM) BINDING POCKET: DISCOVERY OF POINT MUTATIONS THAT ENGENDER A "MOLECULAR SWITCH" IN PAM PHARMACOLOGY

This work is based on publication (Gregory et al., 2013).

Summary

Positive allosteric modulation of metabotropic glutamate receptor subtype 5 ($mGlu_5$) is a promising novel approach for the treatment of schizophrenia and cognitive disorders. Allosteric binding sites are topographically distinct from the endogenous ligand-(orthosteric) binding site, allowing for co-occupation of a single receptor with the endogenous ligand and an allosteric modulator. Negative allosteric modulators (NAMs) inhibit, while positive allosteric modulators (PAMs) enhance, the affinity and/or efficacy of the orthosteric agonist. The molecular determinants that govern $mGlu_5$ modulator affinity versus cooperativity are not well understood. Focusing on the modulators based on the acetylene scaffold, we sought to determine the molecular interactions that contribute to PAM versus NAM pharmacology. Generation of a comparative model of the transmembrane-spanning region of $mGlu_5$ served as a tool to predict and interpret the impact of mutations in this region. Application of an operational model of allosterism allowed for determination of PAM and NAM affinity estimates at receptor constructs that possessed no detectable radioligand binding as well as delineation of effects on affinity versus cooperativity. Novel mutations within the transmembrane domain regions were identified that had differential effects on acetylene PAMs versus 2-methyl-6-

(phenylethynyl)-pyridine (MPEP), a prototypical NAM. Three conserved amino acids (Y658, T780, S808) and two non-conserved residues (P654, A809) were identified as key determinants of PAM activity. Interestingly, we identified two point mutations in TM6 and 7 that, when mutated, engender a mode switch in the pharmacology of certain PAMs.

Introduction

Metabotropic glutamate receptors (mGlus) are seven-transmembrane spanning (G protein-coupled) receptors (7TMRs) that include eight subtypes, $mGlu_1$- $mGlu_8$, for the major excitatory neurotransmitter, glutamate (Niswender & Conn, 2010). Historically, it has been difficult to develop highly mGlu subtype selective ligands due to the high sequence conservation of the endogenous ligand (i.e., glutamate) orthosteric binding site. This led to the search for compounds that interact at "allosteric" sites, topographically distinct from the orthosteric site. Referred to as allosteric modulators, the presence of such compounds can affect the affinity and/or efficacy of an orthosteric ligand, a property referred to as cooperativity. Modulators that inhibit orthosteric ligand binding and/or activity are negative allosteric modulators (NAMs) while those that enhance are positive allosteric modulators (PAMs); a third category, silent (or neutral) allosteric modulators (SAMs), includes compounds that bind but do not modulate receptor activity.

Efforts to develop mGlu allosteric modulators have been especially successful for $mGlu_5$; a broad range of allosteric modulators as well as allosteric radioligands has been developed including pure PAMs, PAMs with agonist activity, weak and full NAMs and SAMs (Ametamey et al., 2007, Y. Chen et al., 2007, 2008, Cosford, Roppe, et al., 2003, Gasparini et al., 1999, Honer, Stoffel, Kessler, Schubiger, & Ametamey, 2007, Kinney et al., 2005, F. Liu et al., 2008, Noetzel et al., 2012, O'Brien et al., 2004, Rodriguez et al., 2005, 2009, 2010, Treyer et al., 2007, Varney et al., 1999). $mGlu_5$ PAMs have potential utility for treatment of cognitive disorders and schizophrenia, whereas NAMs are being

pursued for treatment of Fragile X Syndrome, depression, anxiety and L-DOPA-induced dyskinesia (Gregory et al., 2011).

In addition to improvements in receptor selectivity, allosteric modulators offer a number of theoretical advantages over their competitive counterparts (Melancon et al., 2012). Modulators that possess no intrinsic efficacy have potential for spatial and temporal modulation of receptor activity. This is an especially important consideration for potential CNS therapeutics, where 'fine-tuning' neurotransmission is likely to yield a better therapeutic outcome than the sustained blockade or activation by an orthosteric ligand. Furthermore, the cooperativity between the two sites is saturable; thus, allosteric modulators have a built-in "ceiling level" to their effect, and may therefore have a larger therapeutic index in the case of overdose.

Structure-activity relationships (SAR) for mGlu modulators, particularly with respect to targeting $mGlu_5$, are also notoriously difficult; SAR is often 'steep' or 'flat' with minimal changes to the structure resulting in a complete loss of activity (Zhao et al., 2007). Furthermore, numerous mGlu modulator chemotypes display 'molecular switches' whereby a PAM or SAM arises from a NAM scaffold or vice versa (Wood, Hopkins, Brogan, Conn, & Lindsley, 2011); originally observed during discovery of the first $mGlu_5$ PAM, difluorobenzaldazine (DFB) (O'Brien et al., 2004). This phenomenon continues to be a challenge for medicinal chemists, with PAMs being derived from NAM scaffolds (Rodriguez et al., 2010, Sharma et al., 2009, Y. Zhou et al., 2010), SAMs from either NAM or PAM chemotypes (Hammond et al., 2010, Rodriguez et al., 2005), and NAMs from PAMs (Lamb et al., 2011). Furthermore, molecular switches have also been described with respect to unanticipated alterations in receptor selectivity (Sheffler et al., 2012). Steep or flat SAR and "molecular switches" may be attributed to changes in the affinity and/or cooperativity of an allosteric modulator. Therefore, we were interested in probing the determinants of allosteric modulator affinity and cooperativity, focusing on the common

allosteric site of group I mGlu's as previously identified for $mGlu_5$ selective modulators such as 2-methyl-6-(phenylethynyl)-pyridine (MPEP) (Malherbe, Kratochwil, Zenner, et al., 2003, Malherbe et al., 2006, Mühlemann et al., 2006, Pagano et al., 2000). Two classes of acetylene PAMs, picolinamides and nicotinamides, that originally evolved from a NAM high-throughput screen (HTS) lead (Rodriguez et al., 2010) were selected for in-depth characterization in comparison with MPEP. We identified seven novel residues that, when mutated, significantly decrease MPEP affinity. Moreover, a single point mutation (W784A) reduced the cooperativity of MPEP, such that it no longer fully blocked the response to glutamate. PAMs were found to interact with the common allosteric site utilized by MPEP, although these compounds showed differential sensitivities to certain mutations. Two different point mutations were identified that conferred a "molecular switch" in the pharmacology of PAMs: T780A converted N-tert-butyl-6-[2-(3-fluorophenyl)ethynyl]pyridine-3-carboxamide (VU0415051) to a weak NAM while S808A converted ((5-((3-fluorophenyl)ethynyl)pyridin-2-yl)(3-hydroxyazetidin-1-yl)methanone (VU0405398) from a weak PAM to a full NAM and N-(tert-butyl)-5-((3-fluorophenyl)ethynyl)picolinamide (VU0405386) from a PAM to a neutral modulator. Our findings build on the existing understanding of the location of the common allosteric site. Quantification of the effect of mutations on modulator pharmacology has allowed delineation of determinants for cooperativity versus affinity.

Materials

Materials and Methods

Dulbecco's Modified Eagle's Medium (DMEM), fetal bovine serum (FBS) and antibiotics were purchased from Invitrogen (Carlsbad, CA). [^{3}H] methoxyPEPy (76.3 Ci/mmol) was custom synthesized by PerkinElmer Life and Analytical Sciences (Waltham,

MA). VU0360172 (N-cyclobutyl-6-((3-fluorophenyl)ethynyl)nicotinamide hydrochloride), VU0405398 ((5-((3-fluorophenyl)ethynyl)pyridin-2-yl)(3-hydroxyazetidin-1-yl)methanone), VU0405386 (N-(tert-butyl)-5-((3-fluorophenyl)ethynyl)picolinamide) and VU0415051 (N-tert-butyl-6-[2-(3-fluorophenyl)ethynyl]pyridine-3-carboxamide) were all synthesized in-house using previously reported methodologies (Gregory et al., 2012, Rodriguez et al., 2010). VU0360173 ((6-((3-fluorophenyl)ethynyl)pyridin-3-yl)(3-hydroxyazetidin-1-yl)methanone) and VU0403602 (N-cyclobutyl-5-((3-fluorophenyl)ethynyl)ethynyl)picolinamide hydrochloride) were synthesized in-house utilizing the methods described in Appendix. Unless otherwise stated, all other reagents were purchased from Sigma-Aldrich (St. Louis, MO) and were of an analytical grade.

Cell culture and mutagenesis

Mutations were introduced into the wild-type rat $mGlu_5$ in pCI:Neo using site-directed mutagenesis (Quikchange II, Agilent, Santa Clara, CA) and verified by sequencing. Wild-type and mutant rat $mGlu_5$ receptor constructs were transfected into HEK293A cells, using Fugene6^{TM} (Promega, Madison, WI) as the transfection reagent. Polyclonal stable cell lines were derived for rat $mGlu_5$ mutant constructs by maintaining the cells at sub-confluence for a minimum of four passages in the presence of 1 mg/ml G418 (Mediatech, Manassas, VA). Stably transfected cell lines were subsequently maintained in complete DMEM supplemented with 10% fetal bovine serum (FBS), 2 mM L-glutamine, 20 mM HEPES, 0.1 mM Non-Essential Amino Acids, 1 mM sodium pyruvate, antibiotic-antimycotic and 500 µg/ml G418 at 37°C in a humidified incubator containing 5% CO_2, 95% O_2.

Intracellular Ca^{2+} mobilization

The day prior to assay, HEK293A-rat $mGlu_5$ cells were seeded at 50,000 cells/well in poly-D-lysine coated black-walled, clear bottom 96 well plates in assay medium (DMEM supplemented with 10% dialyzed fetal bovine serum, 20 mM HEPES and 1 mM sodium pyruvate). On the day of assay, the cell permeant Ca^{2+} indicator dye Fluo-4 (Invitrogen, Carlsbad, CA) was used to assay receptor-mediated Ca^{2+} mobilization as described previously (Hammond et al., 2010) using a Flexstation II (Molecular Devices, Sunnyvale, CA). A 5-point smoothing function was applied to the raw fluorescent Ca^{2+} traces and basal fluorescence of individual wells determined during the first 20 sec. The peak increase in fluorescence over basal was determined prior to normalization to the maximal peak response elicited by glutamate.

Radioligand binding

Radioligand binding assays were performed on HEK293A cell membranes as described previously (Gregory et al., 2012). Briefly, for saturation binding experiments, membranes (20-50 µg/well) were incubated with a range of [^{3}H]-3-methoxy-5-(pyridin-2-ylethynyl)pyridine ([^{3}H]methoxyPEPy) concentrations (0.5 nM-60 nM) for 1 hr at room temperature with shaking in Binding Buffer (50 mM Tris-HCl, 0.9% NaCl, pH7.4). 10 µM MPEP was used to determine non-specific binding. For inhibition binding experiments, membranes were incubated with ~2 nM [^{3}H] methoxyPEPy and a range of concentrations of test ligand (100 pM-100 µM) for 1 hr at room temperature with shaking in Ca^{2+} assay buffer with 1% DMSO final. Assays were terminated by rapid filtration through GF/B Unifilter plates (PerkinElmer Life and Analytical Sciences, Boston, MA) using a Brandel 96-well plate Harvester (Brandel Inc., Gaithersburg, MD), and three washes with ice-cold Binding Buffer, separating bound from free radioligand. Plates were allowed to dry

overnight and radioactivity counted using a TopCount Scintillation Counter (PerkinElmer Life and Analytical Sciences, Boston, MA).

Generation of an $mGlu_5$ comparative model

A comparative model of $mGlu_5$ was constructed using the protein structure prediction software package, Rosetta version 3.4 (Leaver-Fay et al., 2011). Based on its high similarity (e-value of 3e-15 with a sequence coverage of 90%) to $mGlu_5$ according to a search using NCBI BLASTP on sequences from the Protein Data Bank (PDB), the X-ray crystal structure human β_2-adrenergic receptor (PDB ID: 2RH1) (Cherezov et al., 2007) was chosen as a template. Both β_2-adrenergic receptor and $mGlu_5$ also share a conserved disulfide bond between a cysteine at the top of transmembrane helix three and a cysteine in extracellular loop two. Members of the Family C 7TMRs, namely the human mGlus and Calcium-sensing receptor (CaSR) sequences, were first aligned with CLUSTALW. Alignment of TM regions between Family C 7TMRs and Family A crystal structure templates were directly adopted from (Malherbe et al., 2006), with the exception of TM's 2, 4 and 7, which were based on the alignment of CaSR with Family A 7TMRs from (Miedlich, Gama, Seuwen, Wolf, & Breitwieser, 2004) and is shown in Figure 55. In the construction of the comparative models, the backbone coordinates of the β_2-adrenergic receptor were retained in the comparative model of $mGlu_5$ while the loop coordinates were built in Rosetta using Monte Carlo Metropolis (MCM) fragment replacement combined with cyclic coordinate descent loop closure. Rosetta ensures that ϕ-ψ angles of backbone segments from homologous sequence fragments from the PDB are introduced into the loop regions. After the fragment substitution, small movements in the ϕ-ψ angles are performed to close breaks in the protein chain. The resulting full sequence models were subjected to eight iterative cycles of side chain repacking and

gradient minimization of ϕ, ψ and χ angles in Rosetta Membrane (Yarov-Yarovoy et al., 2005).

Docking of allosteric modulators

The negative allosteric modulator MPEP and six acetylene PAMs (VU0360173, VU0405398, VU0360172, VU0403602, VU0415051 and VU0405386) were computationally docked into the comparative model of $mGlu_5$ using Rosetta Ligand (Davis & Baker, 2009, Lemmon & Meiler, 2012, Meiler & Baker, 2006). Each modulator was allowed to sample docking poses in a 5 Å radius centered at the putative binding site for MPEP, determined by the residues known to affect modulator affinity and/or function. For MPEP, separate docking experiments were carried out, centered on two residues shown to greatly influence modulator affinity when mutated: P654 and S808. For the six acetylene PAMs, docking experiments were all centered on P654. Once a binding mode had been determined by the docking procedure, 10 low energy conformations of the ligand created by MOE (Molecular Operating Environment, Chemical Computing Group, Ontario, Canada) were tested within the site. Side-chain rotamers around the ligand were optimized simultaneously in a Monte-Carlo minimization algorithm. The energy function used during the docking procedure contains terms for van der Waals attractive and repulsive forces, hydrogen bonding, electrostatic interactions between pairs of amino acids, solvation, and a statistical term derived from the probability of observing a side-chain conformation from the PDB. For each modulator, over 2,000 docked complexes were generated and clustered for structural similarity using bcl::Cluster (Alexander et al., 2011). The lowest energy binding mode from the five largest clusters for each modulator were used for further analysis. A detailed protocol capture for protein modeling and ligand docking, including links to input and output files, is provided in the Appendix.

Data Analysis

All computerized nonlinear regression was performed using Prism 5.01 (GraphPad Software, San Diego, CA). Inhibition [^{3}H]methoxyPEPy binding data sets were fitted to a one-site inhibition binding model and estimates of inhibitor dissociation constants (K_I) were derived using the Cheng-Prusoff equation for competitive ligands (Cheng, Prusoff, & others, 1973) and the following version of the allosteric ternary complex model for ligands that did not fully displace radioligand (Lazareno & Birdsall, 1995):

$$\frac{Y}{Y_{max}} = \frac{[D]}{[D] + \frac{K_D\left(1 + \frac{[B]}{K_B}\right)}{\left(1 + \frac{\alpha[B]}{K_B}\right)}}$$

where Y/Y_{max} is the fractional specific binding, D is the radioligand concentration, B is the molar concentration of the allosteric modulator, K_D is the radioligand equilibrium dissociation constant, and K_B is the allosteric modulator equilibrium dissociation constant. α denotes the cooperativity factor, where values of α > 1 describe positive cooperativity, values of α < 1 (but greater the 0) denote negative cooperativity and α = 1 denotes neutral cooperativity.

Shifts of glutamate concentration-response curves by allosteric modulators were globally fitted to an operational model of allosterism (K. Leach, Sexton, & Christopoulos, 2007):

$$Effect = \frac{E_m(\tau_A A(K_B + \alpha\beta B) + \tau_B B K_A)^n}{(AK_B + K_A K_B + K_A B + \ \alpha AB)^n + (\tau_A A(K_B + \alpha\beta B) + \tau_B B K_A)^n}$$

where A is the molar concentration of orthosteric agonist glutamate and B is the molar concentration of the allosteric modulator. K_A is the equilibrium dissociation constant of the orthosteric agonist, glutamate, and K_B is the allosteric modulator equilibrium dissociation constant.

Affinity modulation is governed by the cooperativity factor α, and efficacy modulation is governed by β. The parameters τ_A and τ_B relate to the ability of the orthosteric and allosteric ligands, respectively, to engender receptor activation. E_m and *n* denote the maximal possible system response and the transducer function that links occupancy to response, respectively.

Allosteric modulator and agonist concentration-response curves were fitted to a four parameter logistic equation in order to determine potency estimates:

$$y = \frac{bottom + (top - bottom)}{(1 + 10^{(logEC_{50} - A)HillSlope}}$$

where *bottom* and *top* are the lower and upper plateaus, respectively, of the concentration-response curve, HillSlope is the Hill coefficient that describes the steepness of the curve, and EC_{50} is the molar concentration of modulator required to generate a response halfway between the *top* and *bottom.*

All affinity, cooperativity and potency parameters were estimated as logarithms and are expressed as the mean ± S.E.M. (Christopoulos, 1998). Statistical analyses were performed where appropriate as indicated using one-way ANOVA with Dunnett's post test when comparing to control, or Tukey's post-test when making multiple comparisons.

Results

Refining the alignment of $mGlu_5$ to Family A 7TMRs and prediction of amino acids within the common allosteric "MPEP" site

In contrast to previous models of $mGlu_5$ (Malherbe, Kratochwil, Zenner, et al., 2003, Malherbe et al., 2006, Pagano et al., 2000), the alignment of $mGlu_5$ presented here is based on a previously reported CaSR alignment (Miedlich et al., 2004) where the PKxY motif in transmembrane domain 7 (TM7) of the human mGlus is aligned with the NPxxY motif in the Family A 7TMRs (Figure 55). This adjustment in the alignment shifts TM7 by

seven residues, predicting that S806, S808 and T810 face the MPEP binding pocket. Indicated on the alignment of the human mGlus with bovine rhodopsin and human β_2-adrenergic receptor (Figure 55) are point mutations included in the current study, which includes mutations previously reported to perturb NAM or PAM activity of $mGlu_5$ modulators (Y. Chen et al., 2008, Malherbe, Kratochwil, Zenner, et al., 2003, Malherbe et al., 2006, Molck et al., 2012, Mühlemann et al., 2006, 2006, Pagano et al., 2000) and mutations novel to this study (highlighted in grey). Based on the localization of these previously known residues, we elected to mutate additional residues predicted to be on the same inward-facing helical face of TMs 3, 5, 6 (I650A, G651F, V739M, P742S, N746A, G747V, T779A, I783A, V788A, Y791F, F792A). Conserved residues were substituted for Ala, whilst non-conserved residues were mutated to the corresponding amino acid in either group II or group III mGlus. In order to validate the alignment of TM7, seven residues were mutated that were predicted to line the inward-facing helical face of TM7: S806A, S808A, S808T, T810A, T810S, A812S, L813A, C815A, M816A. As a negative control, a point mutation previously shown to affect PAM activity (by CPPHA) at the second allosteric site on $mGlu_5$, F585I, was also included (Y. Chen et al., 2008).

Identification of eight novel point mutations that perturb MPEP inhibition of glu

To assess the contribution of both novel and previously identified residues to the MPEP binding pocket, MPEP (Figure 21) was screened for its effect on the maximal response to glutamate in the Ca^{++} mobilization assay at a single concentration. For screening purposes 10 nM MPEP was selected as this concentration caused a significant decrease in glutamate E_{max} to ~45% at $mGlu_5$-wt, such that increases or decreases in the % inhibition caused by MPEP could be detected (Figure 21A). All mutations were functional and expressed at levels ranging from 0.4-3.8 pmol/mg (Table 8, Table 22). Mutations that showed a lack of [^{3}H]methoxyPEPy binding showed a similar range of

expression levels (between that of the low-expressing $mGlu_5$-wt cell line and the higher-expressing, polyclonal $mGlu_5$-wt line) as confirmed by immunoblotting (Figure 52). Twelve point mutations, corresponding to nine different amino acids, significantly reduced inhibition of glu E_{max} by 10 nM MPEP (Figure 21B). From the original 33 point mutations screened, 24 were selected for further characterization, including the negative control F585I.

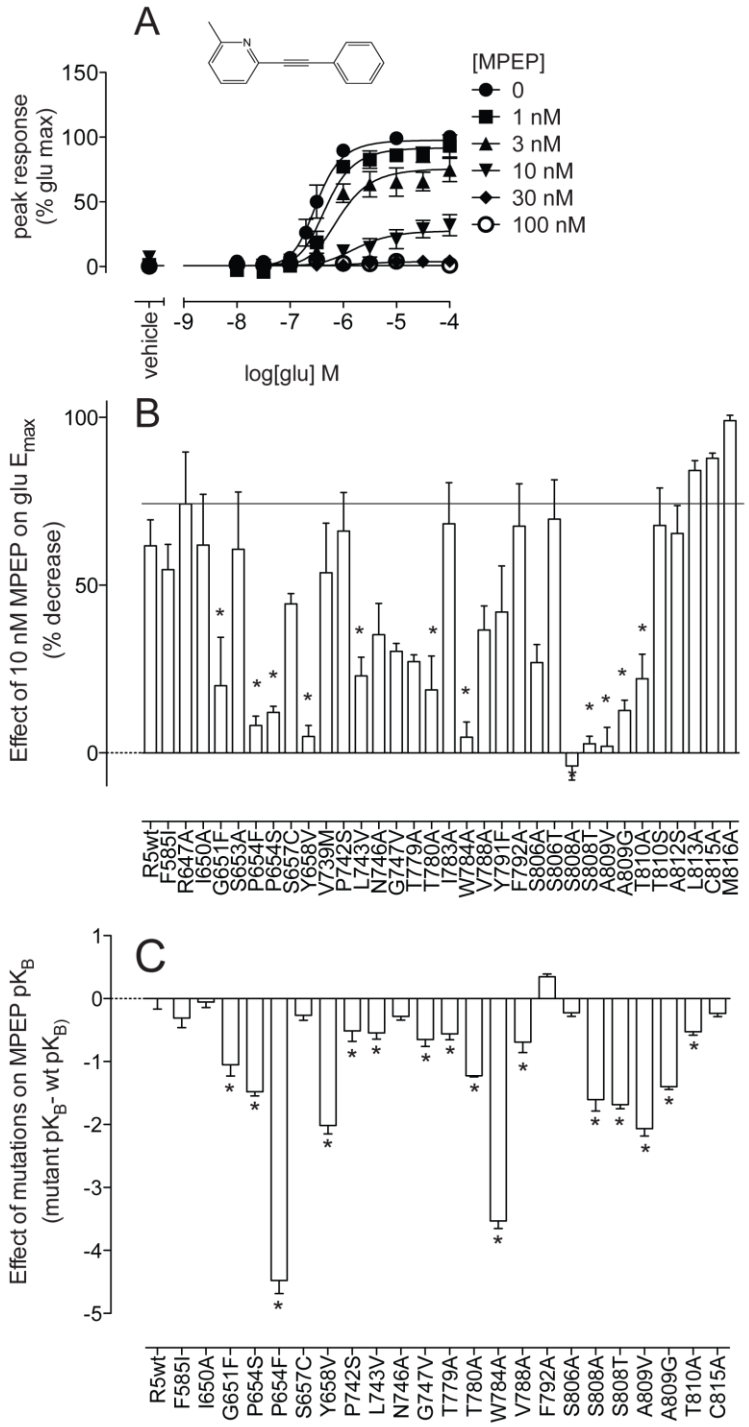


Figure 21

Probing the common allosteric binding site on mGlu5 with negative allosteric modulator, MPEP.

A) At the wild-type rat mGlu5 receptor, MPEP inhibits glutamate-mediated mobilization of intracellular Ca2+, depressing the maximal response. B) Single point mutations of mGlu5 were screened for their ability to impact inhibition of the maximal response to glutamate in the presence of 10nM MPEP. C) Comparison of MPEP affinity estimates at mutants with wild-type. Data represent the mean ± S.E.M of 3-6 experiments performed in duplicate. Error bars not shown lie within the dimensions of the symbol.

Table 8
Equilibrium binding parameters for [3H] methoxyPEPy and MPEP at mGlu5 mutations.
Data represent the mean ± s.e.m of a minimum of three independent determinations.

	[^{3}H]methoxyPEPy pK_D [a]	B_{max} [b] (pmol/mg)	MPEP pK_i [c]
R5-wt (low)	8.24 ± 0.09	0.6 ± 0.0*[#]	7.87 ± 0.04[#]
R5-wt (poly)	8.23 ± 0.10	3.8 ± 0.8	7.77 ±0.03
F585I	7.91 ± 0.19	1.8 ± 0.6	7.93 ± 0.09
R647A	8.22 ± 0.08	1.6 ± 1.3	n.d.
I650A	8.43 ± 0.11	1.0 ± 0.1*	8.34 ± 0.11*
G651F	No appreciable binding		
P654S	No appreciable binding		
P654F	No appreciable binding		
S657C	8.35 ± 0.28	0.8 ±0.3*	8.12 ±0.09
Y658V	No appreciable binding		
V739M	8.19 ± 0.02	1.6 ± 0.3	7.91 ± 0.07
P742S	7.71 ± 0.29	0.4 ± 0.2*	7.67 ± 0.14
L743V	7.92 ± 0.09	1.1 ± 0.2*	7.23 ± 0.10*
N746A	7.72 ± 0.05	1.2 ± 0.4*	7.60 ± 0.08
G747V	8.37 ± 0.07	1.9 ± 0.6	7.88 ± 0.06
T779A	8.17 ± 0.13	0.8 ± 0.2*	
T780A	No appreciable binding		
W784A	No appreciable binding		
V788A	8.13 ± 0.08	1.0 ± 0.3*	7.71 ± 0.13
F792A	8.11 ± 0.21	1.7 ± 0.6	7.90 ± 0.13
S806A	7.71 ± 0.07	1.2 ± 0.1*	7.21 ± 0.07*
S808A	No appreciable binding		
S808T	No appreciable binding		
A809V	No appreciable binding		
A809G	No appreciable binding		
T810A	7.86 ± 0.15	2.0 ± 0.4	7.28 ± 0.09*
C815A	8.00 ± 0.19	0.8 ± 0.1*	7.57 ± 0.02

* denotes significantly different to wild-type (polyclonal) value, $p<0.05$, one-way ANOVA, Dunnett's post test.
n.d. denotes not determined.
[#] data previously reported (Gregory et al., 2012).
[a] negative logarithm of the equilibrium dissociation constant of [^{3}H] methoxyPEPy.
[b] maximal number of binding sites.
[c] negative logarithm of the equilibrium dissociation constant of MPEP.

Delineation of impact of mutations on MPEP affinity versus cooperativity

Progressive fold-shift analysis by MPEP of the glutamate concentration-response curve for Ca^{++} mobilization was performed using the operational model of allosterism (K. Leach et al., 2007). This model has previously been validated for estimating affinity and cooperativity of $mGlu_5$ allosteric modulators (Gregory et al., 2012). As expected, given that mutations were introduced into the TMs, little or no change was observed in the potency

and efficacy ($logτ_A$) of glutamate across all point mutations compared to wild-type (Table 22). Furthermore, the assumption that glutamate affinity was unaffected by point mutations in the TMs had no effect on the estimates of modulator pK_B (Figure 57). The affinity (pK_B) of MPEP was found to be significantly reduced compared to wild-type at 16 point mutations (Table 8, Figure 21C). MPEP affinity estimates were reduced 3-10 fold at: P742S, L743V, G747V, T779A, V788A and T810A; 10-30 fold at G651F, P654S, T780A and A809G; 30-100 fold at Y658V, S808A, S808T and A809V. Greater than 1000-fold lower MPEP affinity was observed at P654F and W784A compared with wild-type. Binding and functional affinity estimates showed good agreement. No appreciable [^{3}H]methoxyPEPy binding was observed at G651F, P654S, P654F, Y658V, T780A, W784A, S808A, S808T, A809V and A809G (Table 8), corresponding to mutations where MPEP affinity was estimated to be decreased 10 fold or greater compared to wild-type. MPEP completely blocked the maximal response to glutamate at all constructs (data not shown) with the exception of P654F and W784A (Figure 22A & B). At W784A (Figure 22B), inhibition of glutamate by MPEP approached a limit, where logβ = -0.27±0.03, indicating that MPEP negative cooperativity is weaker at W784A compared to wild-type.

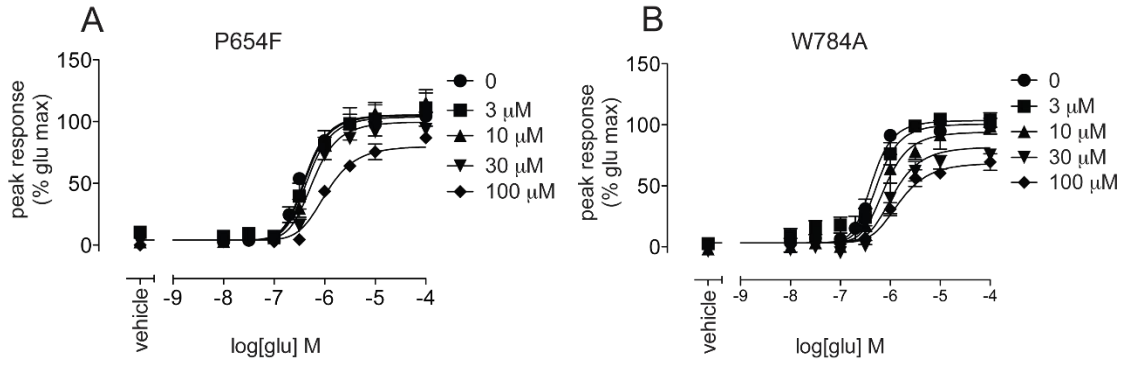


Figure 22
At two point mutations, MPEP did not fully depress the maximal response to glutamate.
MPEP inhibition of glutamate-mediated mobilization of intracellular Ca^{2+} at mGlu5-P654F (A) and W784A (B). In the presence of MPEP at concentrations up to 100 uM, glutamate retained some activity in both cell lines. Data represent the mean ± S.E.M of 3-5 experiments performed in duplicate. Error bars not shown lie within the dimensions of the symbol.

Modeling MPEP binding to $mGlu_5$ comparative model

MPEP was docked into the comparative model of $mGlu_5$ to aid interpretation of these mutational data. Docking experiments were centered on two sites, represented by two residues that were demonstrated to reduce MPEP affinity when mutated, P654 and S808. Representative binding modes from each experiment demonstrate possible binding modes for MPEP interacting with $mGlu_5$ (Figure 23A). The lowest energy MPEP binding modes from the largest clusters for docking to both sites are shown in Figure 23B and 23C. Docking to P654 identified a common binding site within the top three clusters, representing 37% of all models; five out of seven point mutations that reduce MPEP affinity are predicted to line the pocket depicted by these poses. However, given the linearity of MPEP, the orientation of the ligand proved more difficult for Rosetta to differentiate; in two of the top three clusters the pyridine ring of MPEP points towards the extracellular space. Interestingly, despite mutations causing 30-100 fold reduction in affinity, S808 and A809 were not predicted to interact with MPEP in the P654-based docking runs. The lowest energy MPEP binding modes of the largest five clusters (accounting for 28% of models) with docking centered on S808 demonstrated greater diversity (Figure 23B). In four out of these five binding modes, the nitrogen from the pyridine ring forms a hydrogen bond with S808.

A similar pose for MPEP has recently been reported (Molck et al., 2012). This interaction may also account for the impact of substitutions to A809, potentially influencing the conformation of its neighboring residue S808. Given the location of S808 at the top of TM7 closer to extracellular loop regions, the greater diversity in binding modes is not unreasonable. We hypothesize that S808 and A809 are important for the initial recognition of the receptor by MPEP, ultimately facilitating binding deeper within the pocket created by TMs 3, 5, 6 and 7. Residues that when mutated significantly reduced MPEP affinity are

highlighted in Figure 23 and color-coded based on their relative impact on MPEP affinity. Notably, residues that have the greatest effect on MPEP affinity are found in close proximity to MPEP, whereas as those with less of an effect (3-10 fold reduction) are more remote. Hydrophobic interactions with MPEP are likely occurring with W784 and Y658. Given its placement in relation to MPEP, the P654F mutation may introduce steric clash into the MPEP binding site preventing MPEP binding and consequently reducing affinity. Ser substitution of this same amino acid did not reduce affinity to the extent of the Phe mutation, with MPEP negative cooperativity retained. Thus it is also possible that replacing Pro with Phe both influences the helix conformation (by removing Pro induced kink) and introduces a larger hydrophobic amino acid, thereby dramatically changing the geography of the MPEP binding pocket, perturbing MPEP affinity and potentially cooperativity.

The effect of mutations to S806, S808 and T810 provide evidence in favor of the alignment of the PKxY motif in the mGlus with NPxxY motif in the Family A 7TMRs. As predicted by the model, the S808A mutation affects MPEP affinity, likely because its position facing the binding pocket provides an interaction with MPEP. S806 and T810 face away from the binding pocket and are not predicted to affect affinity, which is in agreement with the functional data (Figure 21C). This verification of the MPEP binding mode through point mutations encouraged the further analysis with positive allosteric modulators of $mGlu_5$.

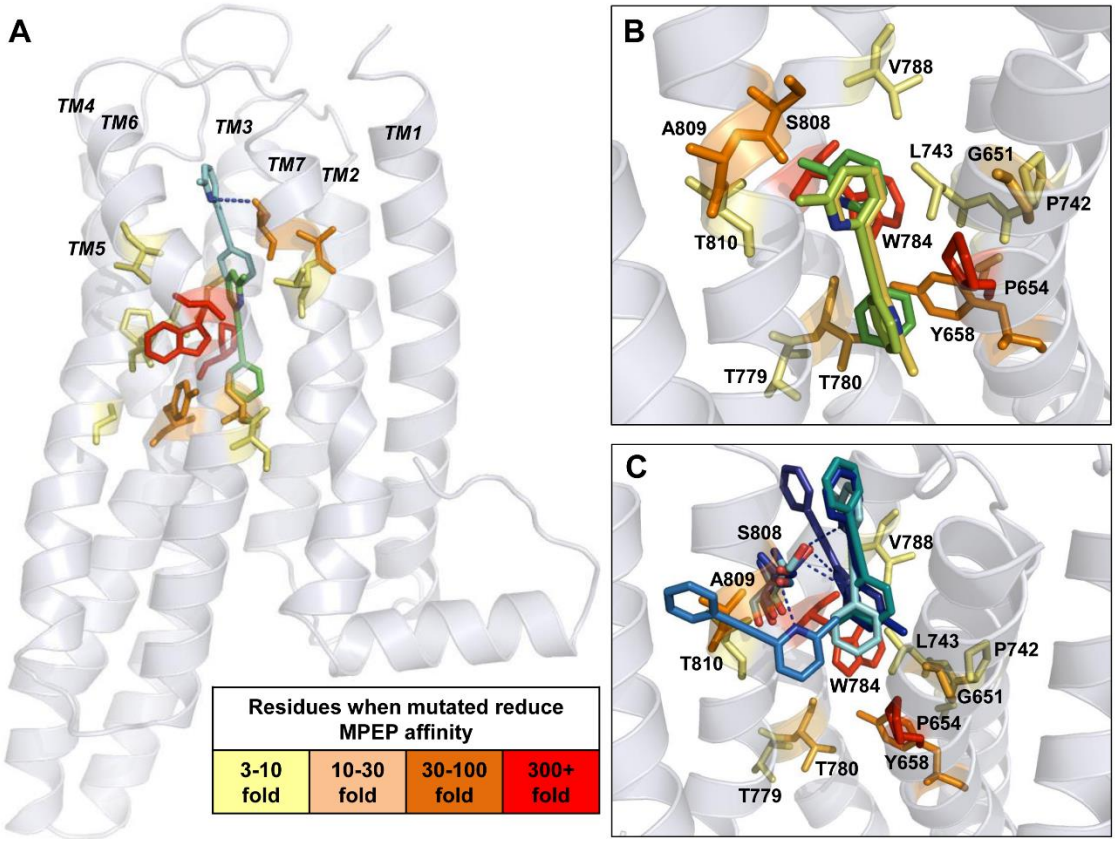


Figure 23
The negative allosteric modulator MPEP docked into mGlu5 comparative model.
MPEP was docked into the mGlu5 comparative model in two separate experiments, centered at P654 and S808. (A) The lowest energy conformation for MPEP from the largest cluster docked at P654 is shown in green and at S808 is shown in cyan. Highlighted are the residues that caused decreases in MPEP affinity when mutated, colored by graded effect compared to wild-type. (B) The lowest energy models from the largest three clusters for MPEP docked at P654. (C) The lowest energy models from the largest five clusters for MPEP docked at S808. Predicted hydrogen bonds between the nitrogen on the pyridine ring and S808 are depicted by dotted blue lines.

Identification of point mutations that affect positive allosteric modulation of glu activity

Three pairs of picolinamide and nicotinamide acetylene $mGlu_5$-selective positive allosteric modulators were selected for investigation of the binding mode of these two chemical scaffolds in the common allosteric site of $mGlu_5$ (Figure 24). These six PAMs span varying degrees of affinity at the wild-type receptor from low (VU0360173, 8 μM) to high (VU0403602, 6 nM; VU0405386, 10 nM) (Tables 2 & 3). Picolinamide PAMs show

higher affinity ~10 fold for the wild-type receptor than their nicotinamide counterparts (Tables 2 & 3). Furthermore, these PAMs show varying degrees of cooperativity, i.e. their ability to induce leftward shifts (fold-shifts) of the glutamate concentration-response curve (Figure 24; Table 11). Similar to the approach utilized for MPEP above, activity of mutant receptors was assessed at a single concentration of PAM. Based on the ability of these compounds to potentiate glutamate Ca^{++} mobilization at the wild-type receptor (Figure 24), PAM concentrations were selected that caused sub-maximal, but reproducible, fold-shifts in the concentration-response curve to glutamate. The fold-shift at a single concentration of PAM was compared with the fold-shift at wild-type (Figure 25). Point mutations that significantly increased or decreased potentiation were selected for further characterization using analysis of progressive fold-shift of the glutamate concentration-response curve to delineate their effects on PAM affinity versus cooperativity.

Table 9

Affinity estimates (pKI) for positive allosteric modulators at mGlu5-wt and mutants from inhibition binding assays.

Data represent the mean ± s.e.m of a minimum of three independent determinations.

	VU0360172 (nicotinamide)	VU0403602 (picolinamide)	VU0360173 (nicotinamide)	VU0405398 (picolinamide)	VU0415051 (nicotinamide)	VU0405386 (picolinamide)
R5-wt (low)	6.57 ± 0.02[#]	8.26 ± 0.15	5.12 ± 0.07	6.05 ± 0.12[#]	6.88 ± 0.04[#]	7.98 ± 0.05[#]
I650A	n.d.	7.89 ± 0.10	5.30 ± 0.26	6.41 ± 0.14	6.62 ± 0.11	7.69 ± 0.15
P742S	5.90 ±0.12*	7.24 ± 0.23*	5.12 ± 0.27	5.78 ± 0.14	6.35 ± 0.13*	7.19 ± 0.12*
L743V	6.32 ± 0.06	6.98 ± 0.10*	4.86 ± 0.08	5.71 ± 0.01*	6.25 ± 0.17*	7.58 ± 0.04
N746A	6.22 ± 0.07	7.49 ± 0.07*	4.63 ± 0.10	5.75 ± 0.03	6.35 ± 0.26*	7.45 ± 0.30
G747V	6.76 ± 0.05	n.d.	n.d.	n.d.	7.06 ± 0.27	8.17 ± 0.35
V788A	7.08 ± 0.23	8.70 ± 0.02	6.10 ± 0.09*	7.87 ± 0.05*	8.15 ± 0.07*	9.20 ± 0.10*
F792A	7.06 ±0.15	7.85 ± 0.04	5.68 ± 0.06	6.74 ± 0.08*	7.09 ± 0.06	7.71 ± 0.19
C815A	6.74 ± 0.24	8.01 ± 0.03	5.01 ± 0.07	6.26 ± 0.04	6.92 ± 0.11	7.96 ± 0.16

* denotes significantly different to value at wild-type receptor, $p<0.05$, one-way ANOVA, with Dunnett's post test.

n.d. indicates not determined

[#] data previously reported (Gregory et al., 2012)

Table 10
Affinity estimates (pKB) for allosteric modulators at mGlu5-wt and mutants derived from operational model analysis of interactions with glutamate.
Data represent mean ± S.E.M. from 4-8 independent experiments performed in duplicate.

	MPEP	VU0360172 (nicotinamide)	VU0403602 (picolinamide)	VU0360173 (nicotinamide)	VU0405398 (picolinamide)	VU0415051 (nicotinamide)	VU0405386 (picolinamide)
R5-wt (poly)	8.58 ± 0.17[#]	6.68 ± 0.15	8.13 ± 0.26	5.45 ± 0.27	6.94 ± 0.17	7.34 ± 0.18	8.04 ± 0.26
R647A	8.89 ± 0.19	6.29 ± 0.12	7.99 ± 0.09	4.91 ± 0.14	6.94 ± 0.10	7.28 ± 0.25	7.74 ± 0.12
I650A	8.62 ± 0.03	7.12 ± 0.19	7.97 ± 0.13	n.d.	7.29 ± 0.15	n.d.	8.10 ± 0.15
G651F	7.53 ± 0.18*	5.86 ± 0.18*	6.50 ± 0.18*	No PAM	5.99 ± 0.28*	6.94 ± 0.19	7.14 ± 0.08
P654S	7.10 ± 0.06*	6.39 ± 0.24	7.33 ± 0.16	<5	5.80 ± 0.27*	5.69 ± 0.46*	7.81 ± 0.27
P654F	4.11 ± 0.21*	5.91 ± 0.17	5.99 ± 0.26*	No PAM	5.33 ± 0.27*	5.25 ± 0.19*	6.23 ± 0.22*
S657C	8.32 ± 0.08	n.d.	7.41 ± 0.17	n.d.	6.99 ± 0.31	6.78 ± 0.12	7.68 ± 0.06
Y658V	6.57 ± 0.13[#]*	No PAM	No PAM	No PAM	4.97 ± 0.35*	No PAM	No PAM
P742S	8.07 ± 0.17*	6.57 ± 0.10	7.50 ± 0.12	4.65 ± 0.17	6.83 ± 0.20	7.09 ± 0.04	7.96 ± 0.16
L743V	8.04 ± 0.10[#]*	6.97 ± 0.15	7.93± 0.14	5.46 ± 0.12	6.77 ± 0.11	7.75 ± 0.20	8.36 ± 0.10
N746A	8.30 ± 0.06	6.66 ± 0.10	7.99 ± 0.11	n.d.	6.50 ± 0.23	n.d.	8.06 ± 0.17
T780A	7.36 ± 0.02*	No PAM	6.03 ± 0.13*	No PAM	5.59 ± 0.05*	5.82 ± 0.40*	5.32 ± 0.24*
W784A	5.50 ± 0.29*	6.76 ± 0.14	7.09 ± 0.34*	5.03 ± 0.44	6.23 ± 0.20	7.14 ± 0.15	7.53 ± 0.14
V788A	7.89 ± 0.13*	7.28 ± 0.17	8.43 ± 0.20	6.28 ± 0.18	7.84 ± 0.14	8.07 ± 0.19	8.79 ± 0.08
F792A	8.93 ± 0.04	7.25 ± 0.33	8.03 ± 0.02	4.80 ± 0.20	7.34 ± 0.20	8.06 ± 0.24	7.02 ± 0.47*
S806A	8.36 ± 0.06	n.d.	n.d.	n.d.	n.d.	n.d.	7.29 ± 0.45
S808A	6.98 ± 0.18*	6.15 ± 0.30	6.47 ± 0.10*	No PAM	6.28 ± 0.09	7.20 ± 0.44	7.48 ±0.20
S808T	6.90 ± 0.06*	6.59 ± 0.20	7.15 ± 0.33*	5.04 ± 0.33	6.40 ± 0.23	6.36 ± 0.26*	8.05 ± 0.27
A809V	6.52 ± 0.12[#]*	5.64 ± 0.19*	6.52 ± 0.21*	No PAM	No PAM	5.56 ± 0.12*	6.22 ± 0.14*
A809G	7.18 ± 0.04*	5.93 ± 0.18	6.59 ± 0.05*	5.13 ± 0.13	5.88 ± 0.05*	6.28 ± 0.11*	6.60 ± 0.24*
C815A	8.35 ± 0.05	6.58 ± 0.22	7.29 ± 0.26*	4.81 ± 0.13	6.73 ± 0.35	7.26 ± 0.11	7.63 ± 0.15

n.d. denotes not determined
"No PAM" indicates no observed positive allosteric modulation.
* denotes significantly different from wild-type value, $p<0.05$, one-way ANOVA, Dunnett's post-test.
[#] data previously reported (Gregory et al., 2012).

Table 11
Functional cooperativity factors (logβ) for allosteric modulators at mGlu5-wt and mutants derived from operational model analysis of interactions with glutamate.
Data represent mean ± S.E.M. from 4-8 independent experiments performed in duplicate.

	VU0360172 (nicotinamide)	VU0403602 (picolinamide)	VU0360173 (nicotinamide)	VU0405398 (picolinamide)	VU0415051 (nicotinamide)	VU0405386 (picolinamide)
R5-wt (poly)	0.37 ± 0.05	0.61 ± 0.08	0.17 ± 0.04	0.37 ± 0.04	0.40 ± 0.02	0.55 ± 0.07
R647A	0.44 ± 0.13	0.60 ± 0.09	0.26 ± 0.03	0.38 ± 0.05	0.41 ±0.11	0.58 ± 0.06
I650A	0.58 ± 0.05	0.55 ± 0.08	n.d.	0.51 ± 0.08	n.d.	0.56 ± 0.08
G651F	0.54 ± 0.07	0.79 ± 0.05	No PAM	0.43 ± 0.04	0.32 ± 0.05	0.70 ± 0.07
P654S	0.26 ± 0.01	0.56 ± 0.05	<5	0.34 ± 0.08	0.22 ± 0.10	0.32 ± 0.05
P654F	0.25 ± 0.04	0.58 ± 0.14	No PAM	0.17 ± 0.04	0.66 ± 0.09*	0.53 ± 0.09
S657C	n.d.	0.82 ± 0.15	n.d.	0.43 ± 0.15	0.55 ± 0.01	0.56 ± 0.12
Y658V	No PAM	No PAM	No PAM	NAM	No PAM	No PAM
P742S	0.87 ±0.10*	1.23 ± 0.06*	0.69 ± 0.07*	0.70 ± 0.15	0.92 ± 0.08*	1.12 ± 0.17*
L743V	0.36 ± 0.14	0.60 ± 0.06	0.43 ± 0.03	0.54 ± 0.08	0.36 ± 0.04	0.80 ± 0.03
N746A	0.62 ± 0.04	0.67 ± 0.04	n.d.	0.57 ± 0.05	n.d.	0.63 ± 0.02
T780A	No PAM	0.77 ± 0.08	No PAM	0.19 ± 0.01	-0.28 ± 0.02*	0.39 ± 0.04
W784A	0.90 ± 0.14*	1.05 ± 0.11*	0.41 ± 0.10	0.48 ± 0.09	0.88 ± 0.10*	0.66 ± 0.15
V788A	0.57 ± 0.10	0.68 ± 0.09	0.66 ± 0.12*	0.47 ± 0.05	0.53 ± 0.06	0.56 ± 0.11
F792A	0.66 ±0.12	0.97 ± 0.10	0.29 ± 0.11	0.73 ± 0.32	0.29 ± 0.02	0.85 ± 0.16
S806A	n.d.	n.d.	n.d.	n.d.	n.d.	0.79 ± 0.13
S808A	0.50 ± 0.06	0.57 ± 0.10*	No PAM	NAM	0.22 ± 0.02	neutral
S808T	0.46 ± 0.06	0.76 ± 0.12	0.28 ± 0.07	0.34 ± 0.04	0.41 ± 0.04	0.45 ± 0.06
A809V	0.60 ± 0.07	0.58 ± 0.09	No PAM	No PAM	0.37 ± 0.07	0.58 ± 0.11
A809G	0.55 ± 0.04	1.29 ± 0.07*	0.41 ± 0.07	0.70 ± 0.03	0.78 ± 0.12*	1.12 ± 0.17*
C815A	1.17 ± 0.12*	1.29 ± 0.10*	0.66 ± 0.06*	0.75 ± 0.24	0.86 ± 0.05*	1.38 ± 0.14*

"No PAM" indicates no observed positive allosteric modulation.
n.d. denotes not determined
* denotes significantly different from wild-type value, $p<0.05$, one-way ANOVA, Dunnett's post-test.

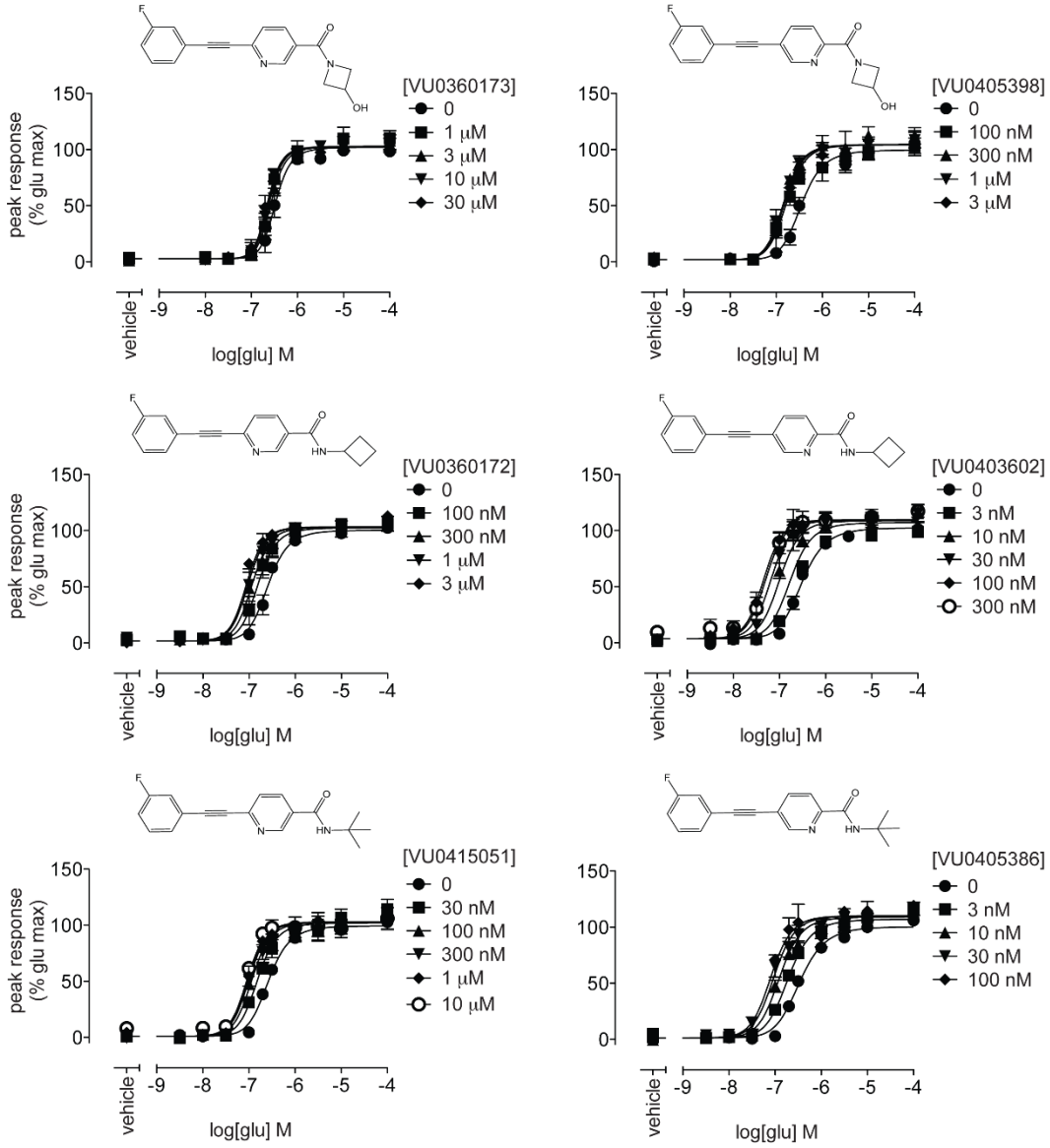


Figure 24

Potentiation of glutamate-mediated Ca2+ mobilization by nicotinamide and picolinamide acetylene PAMs at wild-type mGlu5.

The six PAMs included in this study potentiate the response to glutamate at mGlu5-wild-type in a Ca++ mobilization assay with varying degrees of cooperativity, as evidenced by increased glutamate potencies in the presence of PAMs. Nicotinamide acetylene PAMs are shown on the left, with the corresponding picolinamide acetylene PAM on the right. Data represent the mean ± S.E.M of 3-7 experiments performed in duplicate. Error bars not shown lie within the dimensions of the symbol.

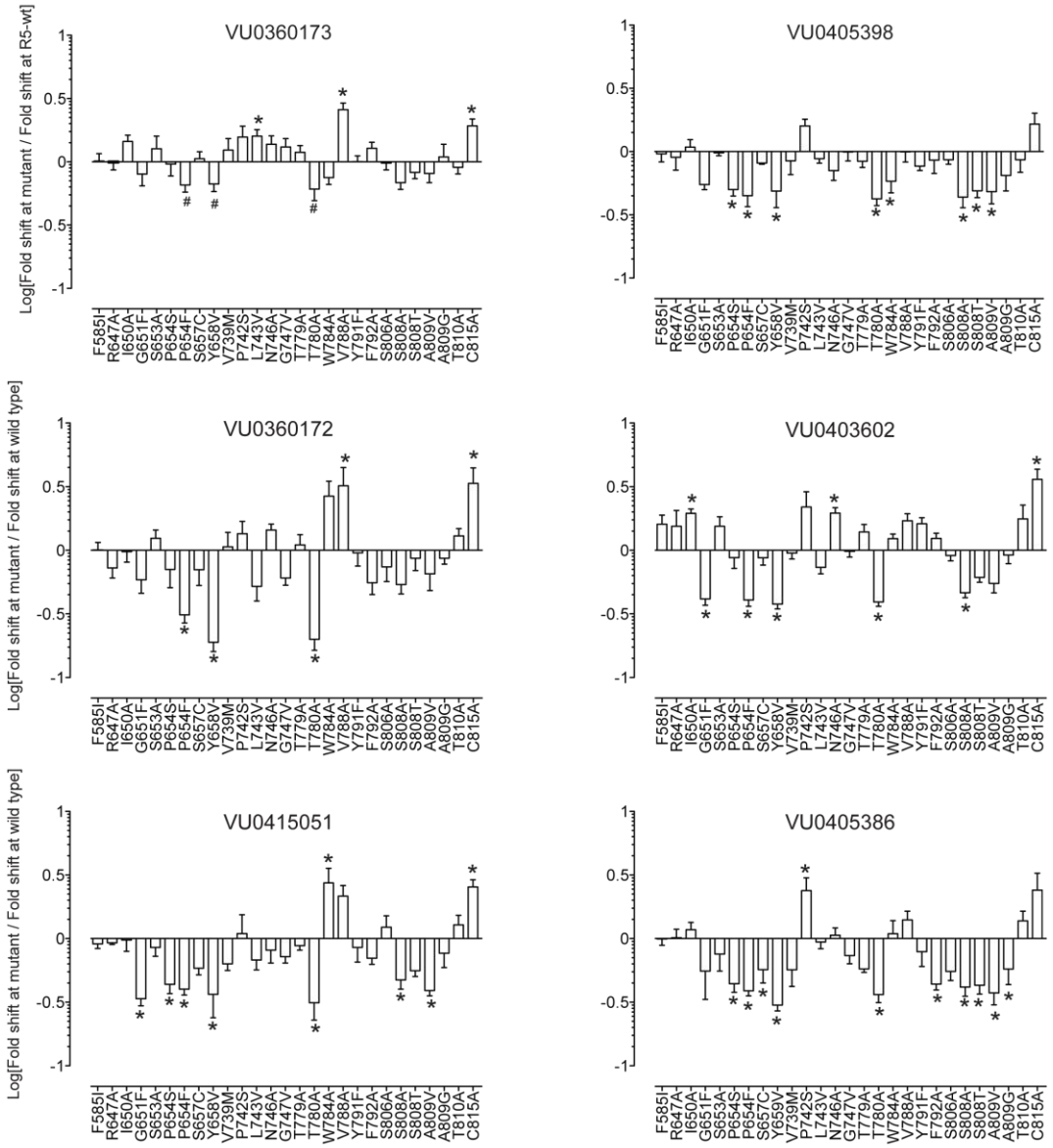


Figure 25

Effect of mutations on the fold-shift caused by a single concentration of PAM

Nicotinamide acetylene PAMs are shown on the left, with the corresponding picolinamide acetylene PAM on the left as indicated. The increase in glutamate potency in the presence of PAM, or fold-shift, at each mutant is expressed relative to that observed for the same concentration at the wild-type receptor. Specifically, PAM concentrations used were: 10 μM VU0360173; 1 μM VU0360172; 100 nM VU0415051; 100 nM VU0405398; 10 nM VU0403602; and 10 nM VU0405386. # denotes no detectable PAM activity. * denotes significantly different to wild-type, $p<0.05$, one-way ANOVA, Dunnett's post-test. Data represent the mean ± S.E.M of 3-7 experiments performed in duplicate. Error bars not shown lie within the dimensions of the symbol.

Impact of mutations on PAM affinity: interactions within the common allosteric pocket

To ensure that potential determinants were not missed in the initial screen, mutations that perturbed potentiation by at least one member of class (either picolinamide or nicotinamide) were assessed across all three members. In addition, mutations that affected one member of a picolinamide/nicotinamide PAM pair were further assessed at both. Affinity estimates (pK_B) for PAMs, derived from modulation of glutamate- mediated Ca^{++} mobilization, are shown in Table 10; for the most part, mutations influenced modulator affinity (Figure 26). Where practical, PAM affinity was also assessed using inhibition of [^{3}H] methoxyPEPy binding (Table 11). Functional (pK_B) and binding (pK_I) estimates of affinity showed strong correlation (Figure 57B). All mutations in TM3 that showed significant effects could not be assessed using radioligand binding-based approaches, relying instead on affinity estimates derived using the operational model of allosterism.

No potentiation was evident for VU0360173 (up to 30 μM) at G651F, P654S, P654F, Y658V, T780A, S808A, and A809V. The corresponding picolinamide, VU0405398, also showed no detectable potentiation at A809V and 10-30 fold decreased affinity at G651F, P654S, T780A and A809G; 30-100 fold reductions were noted at P654F and Y658V. Collectively, these results suggest that the lack of potentiation observed for VU0360173 at these mutants is a result of decreased affinity.

VU0360172 affinity was reduced approximately 10 fold at G651F and A809V. No positive allosteric modulation was observed at Y658V or T780A. Interestingly, P654S had no significant effect on VU0360172 affinity, while at the Phe substitution VU0360172 affinity was also reduced (6 fold; this did not reach significance). The picolinamide counterpart of VU0360172, VU0403602, also showed no appreciable PAM activity at Y658V, alongside marked reductions in affinity (greater than 100 fold) at P654F and T780A. VU0403602 affinity was also decreased 30-100 fold at G651F, S808A, A809V and

A809G; and 10-30 fold at S808T. At P654S and C815A, VU0403602 pK_B values were reduced 3-10 fold compared to wild-type; although this did not reach significance at P654S.

Similar to that reported for the previous four PAMs, VU0415051 did not potentiate glutamate at Y658V (up to 30 μM). Decreased VU0415051 affinity was noted at: S808T, A809G (10-30 fold); P654S, T780A, A809V (30-100); and greater than 100 fold at P654F. VU0405386 also showed no discernible PAM activity at Y658V nor S808A. The affinity of VU0405386 was decreased compared to wild-type at F792A, A809G (10-30 fold); P654F, A809V (30-100 fold); and greater than 300 fold at T780A. Interestingly, neither of the VU0415051/VU0405386 pair were unaffected by G651F.

Three point mutations in TM5, P742S, L743V and N746A, showed no significant effects in functional assays; however, in inhibition binding assays, significant reductions in affinity were observed for some PAMs (Table 9). P742S significantly decreased pK_I values 3-10 fold for VU0361072, VU0403602, VU0415051 and VU0405386. Significantly decreased pK_I values were also noted for VU0403602 (19 fold), VU0405398 (2 fold), VU0415051 (4 fold) at L743V, and for VU0403602 (6 fold), VU0415051 (3 fold) at N746A. In contrast, V788A, in TM6, showed a trend for increased pK_B and pK_I values (2-10 fold).

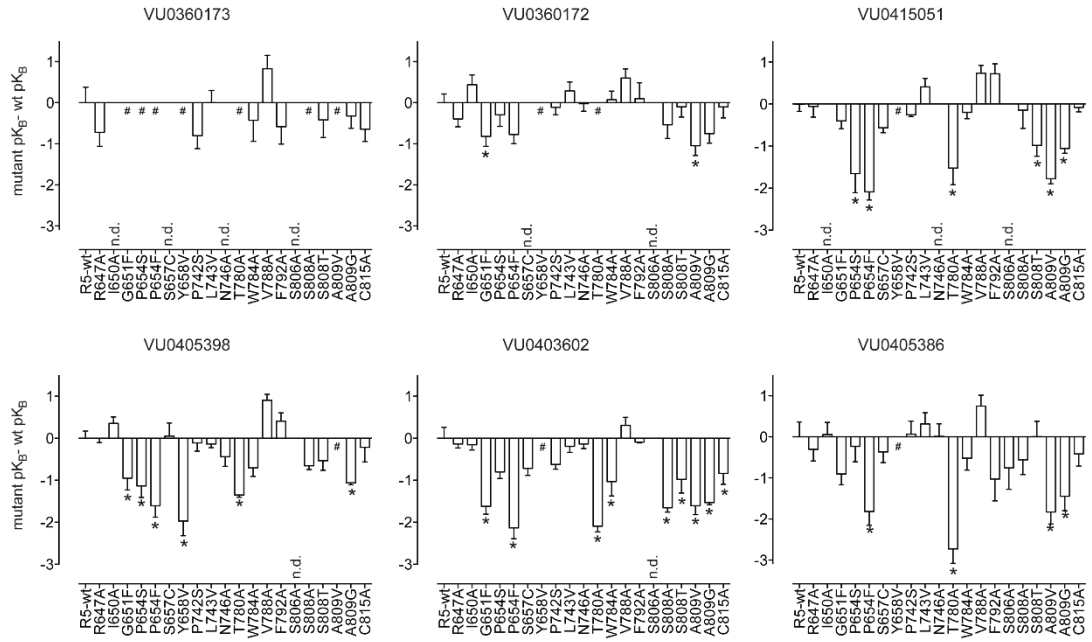


Figure 26
Effect of mutations on positive allosteric modulator affinity estimates.
Nicotinamide acetylene PAMs are shown in the top three panels, with the corresponding picolinamide acetylene PAM in the bottom three panels. Affinity estimates (pKB) were derived using an operational model of allosterism (Gregory et al., 2012, K. Leach et al., 2007) from progressive fold-shifts of the glutamate concentration-response curve for Ca2+ mobilization. The difference between the pKB for the mutant versus wild-type is plotted. # denotes no detectable PAM activity. * denotes significantly different to wild-type, $p<0.05$, one-way ANOVA, Dunnett's post-test. Data represent the mean ± S.E.M of 3-7 experiments performed in duplicate. Error bars not shown lie within the dimensions of the symbol.

Docking of PAMs to $mGlu_5$ comparative model

Docking of the positive allosteric modulators in the $mGlu_5$ comparative model provides insight into the significant residues identified to reduce affinity in the binding pocket. Each of the six positive allosteric modulators were allowed to explore a 5 angstrom radius around P654, and the lowest energy binding modes of the largest 5 clusters were shown to bind in the same pocket as MPEP (Figure 27A-C). As seen with MPEP, mutation of P654 in TM3 to a bulky residue (Phe) likely introduces a steric clash, reducing the ability of PAMs to engage the common $mGlu_5$ binding pocket, accounting for the reduced affinity for all six PAMs. As noted previously, the amino acid in this position in not conserved across the mGlu family, such that it is likely P654 also contributes to the subtype selectivity

of these PAMs. Introduction of steric bulk is also likely to underscore the impact of the G651F mutation on all modulators except for the VU0415051/VU0405386 pair. The impact of substitution of Y658 with Val can likely be attributed to disruption of key polar interactions between modulators and Y658, T780 and W784 (Highlighted by dashed lines in Figure 27A-C). Similarly, T780 likely participates in a polar interaction with the carbonyl of the PAMs, with the exception of VU0403602 where the tertiary amide is predicted to interact with this residue, such that Ala substitution of this amino acid causes the drastic loss of affinity seen for all modulators. W784 likely contributes to hydrophobic interactions within the pocket, which may account for the trend for picolinamide PAM affinity to be perturbed by W784A; it is also noteworthy that the W784 may be involved in a hydrogen bonding network linking the modulators, TM3 and TM7. V788A shows a trend for increased affinity, which may be due to secondary effects on protein conformation and the previously implicated F787, where Ala substitution reduced NAM and PAM interactions (Malherbe, Kratochwil, Zenner, et al., 2003, Malherbe et al., 2006, Mühlemann et al., 2006). S808 in TM7 may participate in hydrogen bonding with the fluorine of the modulators when in close proximity, although this was not evident in the lowest energy binding modes. Substitution of A809 with Gly or Val reduced modulator activity across the board, pointing to the importance of the alanine in maintaining the correct helix conformation for binding. With respect to C815 and VU0403602, it is likely that there is an interaction with the cyclobutane of the modulator; however, it is not immediately apparent whether this is a direct or indirect effect.

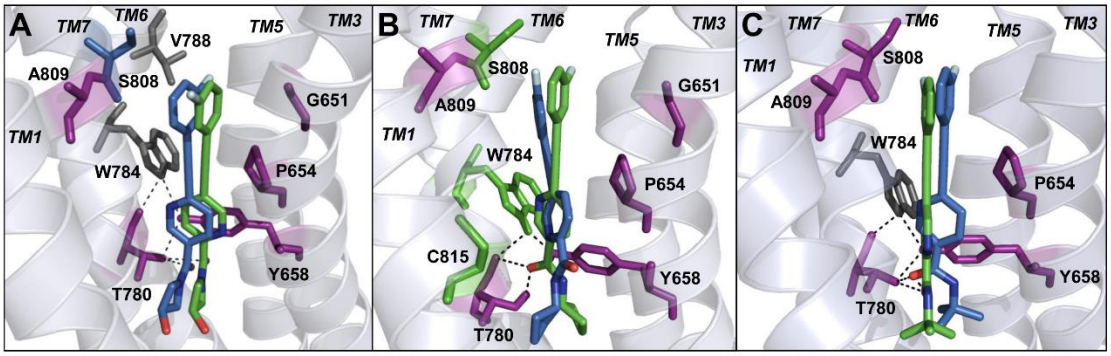


Figure 27
Computational docking of three pairs of nicotinamide and picolinamide acetylene positive allosteric modulators into mGlu5.
A) VU0360173 in blue and VU0405398 in green, B) VU0360172 in blue and VU0403602 in green and C) VU0415051 in blue and VU0405386 in green. Residues that when mutated caused a significant decrease in modulator affinity in mGlu5 are highlighted in the respective color of the modulator. Residues that affect both the nicotinamide and picolinamide in the pair are highlighted in purple. A predicted hydrogen bond network involving the modulators, Y658, T780A and W784 is represented by the dashed black lines. Highlighted in grey are residues that influence the cooperativity of certain modulators.

Quantifying effects on cooperativity: identification of mutations that engender "molecular switches"

When applying the operational model of allosterism, the interaction between glutamate and PAMs was assumed to be exclusively via efficacy modulation, an assumption previously validated for $mGlu_5$ PAMs from this scaffold (Gregory et al., 2012). Cooperativity estimates (logβ) are summarized in Table 11 and comparisons with wild-type are shown in Figure 28. Alongside marked reductions in PAM affinity, P654F also significantly increased cooperativity of VU0415051 (~2 fold). P742S in TM5 had no effect on PAM affinity, yet increased cooperativity (~3 fold). In TM7, A809G significantly increased the cooperativity of the higher affinity PAMs: VU0403602, VU0415051 and VU0405386 (2-5 fold), with a similar trend observed for VU0360172, VU0360173 and VU0405398. Also in TM7, C815A significantly increased the cooperativity of all PAMs (2-7 fold) with the exception of VU0405398. P742S, A809G, C815A all represent mutations where the glutamate potency and/or efficacy was lower than the wild-type (Table 22), such

that these increases in PAM cooperativity may be attributable to PAMs stabilizing an unstable mutant receptor. Significantly increased cooperativity (3 fold) was also observed at W784A for VU0360172, VU0403602 and VU0415051, with VU0360173 showing higher, although not significant, cooperativity. Interestingly, this was not a global PAM phenomenon as the cooperativity of VU0405386 and VU0405398 unchanged. Given the putative involvement of W784 in a key hydrogen bonding network for the PAMs (dotted lines in Figure 27A-C), perhaps this differential effect is driven by the relative importance of this interaction over that of the PAM functional head group. Also in TM6, V788A increased the cooperativity of VU0360173 (3 fold) alone; no direct interaction is predicted from the docking between VU0360173 and V788, suggesting this is an indirect effect on the geography of the binding pocket. The selective effect of V788A on VU0360173 cooperativity may be attributable to the fact that this PAM has the weakest cooperativity and lowest affinity.

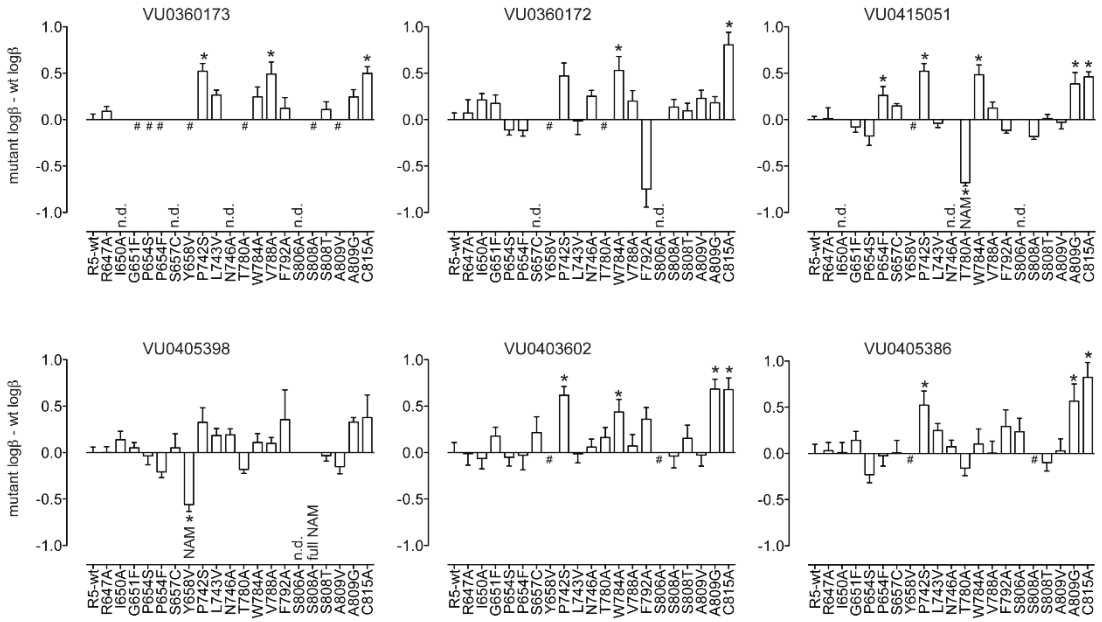


Figure 28
Effect of mutations on positive allosteric modulator cooperativity factors.
Nicotinamide acetylene PAMs are shown in the top three panels, with the corresponding picolinamide acetylene PAM in the bottom three panels. Cooperativity estimates (logβ) were derived using an operational model of allosterism (Gregory et al., 2012; Leach et al., 2007) from progressive fold-shifts of the glutamate concentration-response curve for Ca2+ mobilization. The difference between the logβ for the mutant versus wild-type is plotted. # denotes no detectable PAM activity. *denotes significantly different to wild-type, p<0.05, one-way ANOVA, Dunnett's post-test. Data represent the mean ± S.E.M of 3-7 experiments performed in duplicate. Error bars not shown lie within the dimensions of the symbol.

Three point mutations altered cooperativity drastically, such that PAMs behaved as NAMs. At Y658V, VU0405398 was a weak NAM, reducing the maximal response to glutamate by ~25% at 30 μM (Figure 29A). This same mutation resulted in a loss of potentiation by all other PAMs. At T780A, VU0415051 became a weak NAM or "partial antagonist", where logβ = -0.28, corresponding to ~40% depression in the glutamate maximal response (Figure 29B). This inhibition approached saturation, the hallmark feature of an allosteric interaction.

The most profound molecular switch engendered by a single point mutation was that of S808A, where VU0405398 behaved as a full NAM, abolishing the maximal

response to glutamate (Figure 29C). As mentioned earlier, at this same mutation VU0405386 showed no discernible PAM activity. Given that this mutation has little or no impact on the affinity of the other PAMs tested, we hypothesized that the lack of potentiation by VU0405386 was caused by a molecular switch from PAM to neutral. To test this hypothesis, concentration response curves for VU0405398 inhibition of an EC_{80} glutamate concentration were performed at S808A in the presence of varied concentrations of VU0405386. As shown in Figure 29D, VU0405386 has no effect on the response to glutamate, but causes parallel rightward shifts in the VU0405398 curve, in a manner consistent with a competitive interaction (Figure 29E; where the Schild slope was not significantly different from unity (0.92 ± 0.09) and pK_B: 7.48 ±0.20).

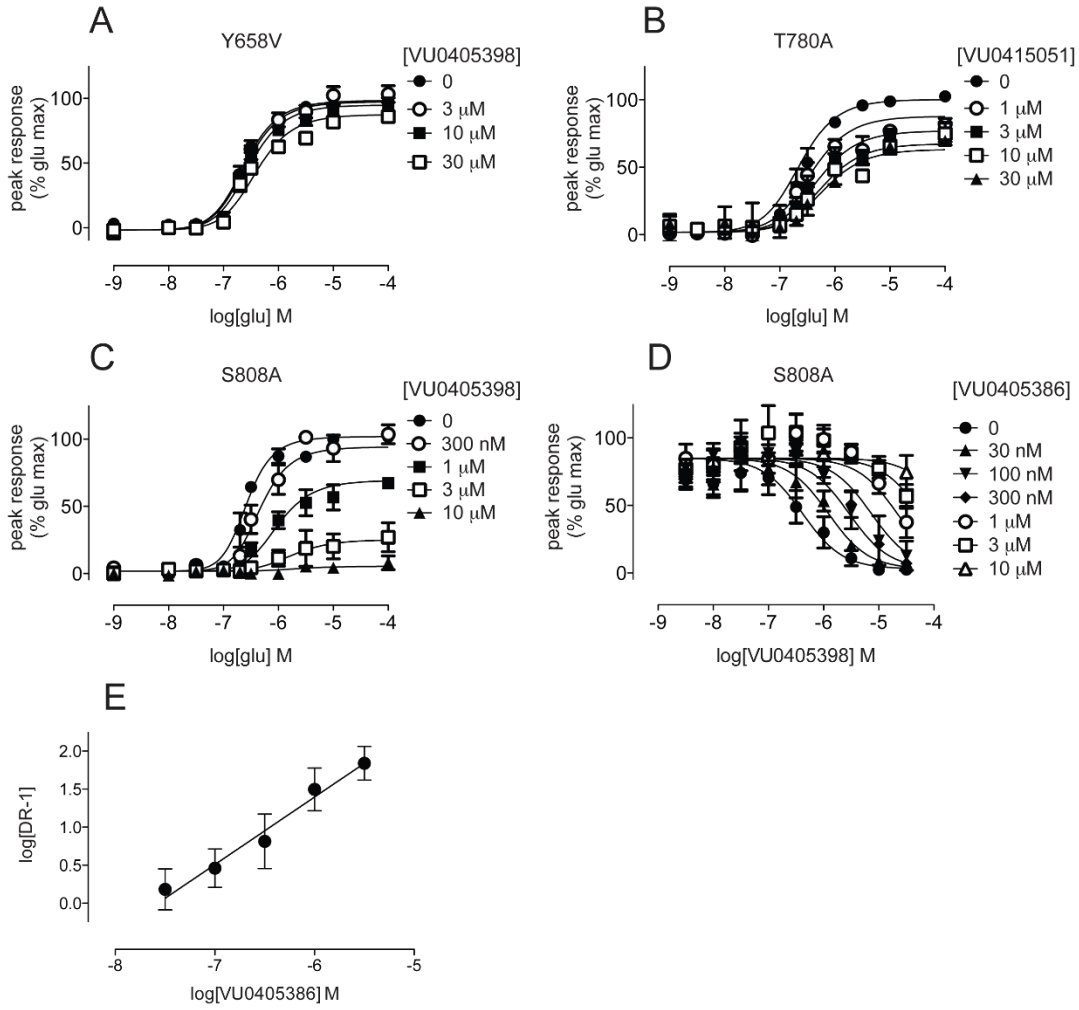


Figure 29
Characterization of mutations that engender a molecular switch in PAM pharmacology.
A) At mGlu5-Y658V, 10 μM and 30 μM VU0405398 inhibited the response to maximal glutamate. B) The interaction between glutamate and VU0415051 at the T780A mutant is negative, with inhibition approaching a limit as defined by the cooperativity. C) At S808A, VU0405398 causes a reduction in glutamate potency and depresses the maximal response to glutamate. D) Concentration-response curves for VU0405398 inhibition of an ~EC80 of glutamate in the absence and presence of the indicated concentrations of VU0405386. E) Schild regression of the interaction between VU0405386 and VU0405398 at S808A. Data represent the mean ± S.E.M of 3-6 experiments performed in duplicate. Error bars not shown lie within the dimensions of the symbol.

Modeling PAM molecular switches

To investigate the molecular cause of the PAM to NAM or neutral switches, modulators were docked into $mGlu_5$ models containing the mutation engendering the

switch. VU0405398 docked in wild-type $mGlu_5$ was compared to VU0405398 docked into the S808A mutant (Figure 30A). Interestingly, introducing this single point mutation resulted in movement of TM7, such that in the mutant receptor VU0405398 is further away from this helix. It is unclear as to why this substitution results in such a conformational change of the receptor. Although not evident in the docked results, it is possible that rotation of the side chain of S808 in the wild-type receptor would allow hydrogen bonding to occur between the fluorine of the modulator and S808, with direct interactions with TM7 being important for stabilizing active receptor conformations. At the mutant receptor, such an interaction is no longer available and VU0405398 instead stabilizes an inactive receptor conformation at the mutated receptor. From the docked poses it is also evident that the picolinamide functional group of VU0405398 adopts a strikingly different orientation within the pocket. Similarly, VU0415051 was docked into the T780A mutant receptor and binding modes compared with those in wild-type $mGlu_5$ (Figure 30B). At the T780A mutant, a polar interaction is no longer formed between the carbonyl of the modulator and the mutant A780; hydrogen bonding is no longer evident with the side chain hydroxyl of Y658 and the modulator carbonyl, nor between the tertiary nitrogen and S657. In the mutant receptor construct, the nicotinamide tertbutyl moiety of VU0415051 adopts a different orientation with decreased affinity. Docking of VU0405386 in the S808A receptor shows subtle differences in the binding mode of the ligand when compared with wild-type (Figure 30C). The movement of TM7 observed with docking VU0405398 to this same receptor mutation is not evident, which may account for the switch from PAM to neutral rather than to a robust NAM as seen for VU0405386. Although not evident in the static docked pose, we hypothesize that the side chain of S808 may rotate to form a polar interaction with the fluorine of the modulator when interacting with the wild-type receptor that is not possible in the mutant. It is also clear that this point mutation at the top of TM7 results in the ligand

adopting a different pose with respect to the picolinamide tertbutyl group, which may contribute to the observed switch to neutral cooperativity.

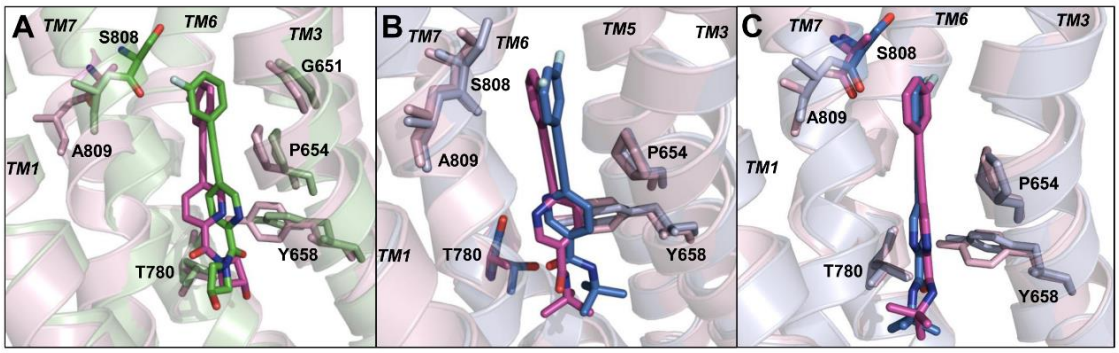


Figure 30
Mutations engendering a molecular switch for mGlu5 allosteric modulators.
A) VU0405398 docked into wild-type mGlu5 (green) and the S808A mutant (magenta). B) VU0415051 docked into wild-type mGlu5 (blue) and the T780A mutant (magenta). C) VU0405386 docked into wild-type mGlu5 (blue) and the S808A mutant (magenta). Mutated residues are colored by element. Key affinity determinants are highlighted to show conformational changes in the binding pocket.

Discussion

By utilizing an operational model of allosterism (Gregory et al., 2012, K. Leach et al., 2007), we have quantitatively assessed the interactions of positive allosteric modulators within the common allosteric site of $mGlu_5$, successfully delineating the impact of point mutations on cooperativity versus affinity. Seven novel point mutations were discovered that negatively impact the MPEP affinity, building on our understanding of the common allosteric binding pocket. Furthermore, Ala substitution of W784 reduced MPEP negative cooperativity. For the six PAMs studied, three conserved (Y658, T780, S808) and two non-conserved residues (P654, A809) were identified as critical determinants of PAM affinity. Interestingly, two point mutations engendered molecular switches in certain PAMs, changing their pharmacology to either NAMs or neutral modulators.

Close examination of the MPEP mutational data reveals a number of interesting observations. First, [^{3}H] methoxyPEPy non-binding mutants corresponded to mutations where MPEP affinity estimates, as derived from functional assays, were decreased 10 fold or greater compared to wild-type. Indeed, four previously reported mutations: Y658V, W784A, A809V, and A809G, that cause a loss of NAM binding to $mGlu_1$ and $mGlu_5$ and/or decreased MPEP potency for glutamate inhibition (Malherbe, Kratochwil, Zenner, et al., 2003, Malherbe et al., 2006, Molck et al., 2012, Mühlemann et al., 2006, Pagano et al., 2000) were confirmed herein and attributed to >30 fold reductions in MPEP affinity. Also in agreement with previous data (Malherbe, Kratochwil, Zenner, et al., 2003, Malherbe et al., 2006, Pagano et al., 2000), P654S, L743V and T780A all reduced MPEP affinity, although the effect of T780A was more pronounced in the current study than previously reported (17 versus 5 fold; (Malherbe, Kratochwil, Zenner, et al., 2003)). Second, in mutating residues predicted to line the helical face towards the binding site as shown in the comparative model, we have further validated utilization of Family A 7TMR crystal structures as templates for mGlu transmembrane spanning domain comparative modeling. Our TM7 alignment, aligns PKxY of $mGlu_5$ to the conserved NPxxY motif, differs by seven residues compared to earlier reports (Malherbe et al., 2006, Pagano et al., 2000), agreeing with the recent report by Molck and colleagues (Molck et al., 2012). In support of this new alignment for TM7, S808 was hypothesized to contribute to the common allosteric site of $mGlu_5$ as opposed to S806 and T810. Mutation of S808 to Ala perturbed MPEP affinity significantly more than S806A and T810A. S808T had a similar impact on MPEP affinity, suggesting a polar interaction may occur between S808 and MPEP that is not achieved by Thr. Alternatively, S808 may be important for maintaining the allosteric binding pocket geography, perhaps via a hydrogen bonding network.

Molecular models of receptor-ligand complexes provide important tools for hypothesis generation, predicting binding modes where experimental structures are

unavailable. In the current study, the model provides an orientation of transmembrane helices, known to be aligned well across GPCR templates, as well as the TM7 helix/loop transition, that reasonably explains experimental results demonstrated by S806, S808 and T810 mutations. The comparative model has less confidence predicting receptor loop regions, as these have less than 20% sequence homology to the template GPCR structures used. The predicted modulator binding site captures residues identified in experimental studies, lending confidence to the binding site location depth within the receptor. The long axes of the ligands were consistently aligned parallel with the helices; however, computationally it was difficult predicting the orientation of these linear ligands and distinguishing whether the functional group points towards the intracellular or extracellular space. Despite these challenges, the $mGlu_5$ computational model with allosteric modulators has provided valuable hypotheses, validated experimentally herein.

Comparison of the PAM data with that of MPEP shows a number of marked differences in mutation susceptibility. Interestingly, more mutations perturbed MPEP affinity than the PAMs; however, with one exception (VU0403602 at C815A), there were no mutations that influenced PAM affinity without affecting MPEP. It is clear that these acetylene PAMs interact with the common allosteric site on $mGlu_5$. These differences likely underscore the determinants that contribute to a NAM versus PAM interacting with the receptor. Interestingly, W784A caused a ~1000 fold reduction in MPEP affinity; however, nicotinamide PAMs were insensitive to this mutation, whilst the picolinamides showed 3-10 fold decreased affinity. W784A increased cooperativity of some PAMs; in agreement with the previous report that W784A enhanced DFB potentiation (Mühlemann et al., 2006). Ala substitution of the equivalent Trp in $mGlu_1$ (W798) has differential effects on $mGlu_1$ NAMs (Fukuda et al., 2009, Suzuki et al., 2007). W784 is analogous to the W of the CWxP motif in Family A 7TMRs that is involved in the well-known rotamer-toggle activation switch (Holst et al., 2010, Shi et al., 2002, Visiers, Ballesteros, & Weinstein, 2002). The modeling

herein predicts a network of polar interactions involving W784 and T780 in TM6, Y658 in TM3 and the PAMs. We hypothesize that the differential effect of W784A on $mGlu_5$ PAMs versus MPEP is likely underscored by PAMs interacting with an active versus inactive receptor conformation.

In each case, picolinamide modulators have higher affinity than their corresponding nicotinamide modulators. From the studies herein, it is not entirely clear which interactions within the binding pocket drive this higher affinity for picolinamides. However, a trend was observed where a greater reliance upon residues in TM7 was observed for compounds, both nicotinamides and picolinamides, with higher affinity (sub 100 nM). Picolinamides also tended to be susceptible to W784A; however, this only reached significance for VU0403602.

For all PAMs, Y658V abolished potentiation, except for VU0405398 where a weak NAM switch was observed. Trp substitution of the equivalent residue in $mGlu_8$ suppresses the activity of a constitutively active mutant receptor (Yanagawa, Yamashita, & Shichida, 2009), raising the possibility that this $mGlu_5$ mutation may have a global receptor activation effect. However, previously DFB potentiation was reportedly unaffected by Y658V (Mühlemann et al., 2006); supporting the hypothetical modulator binding modes shown, where an interaction is predicted with this residue, rather than Y658V impacting active receptor states.

A number of previous studies have identified PAMs as competitive with the common $mGlu_5$ allosteric (or “MPEP”) site on the basis of a single point mutation in TM7, A809V (Y. Chen et al., 2008, Hammond et al., 2010). Previously, A809V was reported to cause a ~30 fold decrease in the affinity of VU29, an $mGlu_5$ PAM from the CDPPB series (Gregory et al., 2012). Validating the utilization of this single point mutation as a read-out of interaction with this common allosteric site, all six PAMs exhibited decreased affinity for this mutant construct. While this interaction appears to be crucial for all PAMs that are

competitive with MPEP tested to date, it should be noted that this does not necessarily have to be the case. Notably, L743V and the equivalent mutation in $mGlu_1$ were previously shown to enhance potentiation by PAMs (Knoflach et al., 2001, Mühlemann et al., 2006); however, for acetylene PAMs, L743V had no effect on cooperativity or affinity. Further studies are underway to probe interactions within the common allosteric site by modulators from distinct chemotypes to better inform our understanding of the molecular determinants of modulator affinity and cooperativity (Manka et al., 2012).

A key finding arising from this study was the identification of residues that, when mutated, engendered a mode switch in allosteric modulator pharmacology, specifically T780A in TM6 and S808A in TM7. Considering that these PAMs originated from a NAM HTS lead (Rodriguez et al., 2010), such drastic changes in modulator cooperativity are not altogether surprising. Indeed, the acetylene series of mGlu allosteric modulators is prone to "molecular switches" (Wood et al., 2011); the SAR plagued by unanticipated changes in the mode of pharmacology and selectivity (Sheffler et al., 2012, Wood et al., 2011). Muhlemann and colleagues previously reported a similar result for the early $mGlu_5$ PAM, DFB, where at the F787A mutant, DFB behaved as a weak NAM (Mühlemann et al., 2006). Pharmacological mode switches were also noted during the discovery of DFB and related compounds (O'Brien et al., 2003). As noted above, movements in TM6 have been implicated in the transition of Family A 7TMRs from inactive to active states. TM7 contains the NPxxY motif, also well-known for its role in receptor activation (Barak, Menard, Ferguson, Colapietro, & Caron, 1995, Fritze et al., 2003, Prioleau, Visiers, Ebersole, Weinstein, & Sealfon, 2002). Furthermore, a water-hydrogen bond network involving polar residues in TMs 1, 2, 6 and 7 is postulated to play an integral role in receptor activation (Nygaard, Valentin-Hansen, Mokrosinski, Frimurer, & Schwartz, 2010). Given the importance of TMs 6 and 7 for the transitioning of receptors into active conformations, these mode switches may be attributed to either a loss of an important

direct contact that facilitates receptor activation upon modulator binding or a global (secondary) effect on protein conformations that prevents some modulators, but not others, from engendering active conformations.

Collectively, these findings highlight the subtleties of interactions within the common mGlu allosteric binding pocket that determine allosteric modulator affinity and cooperativity. The identification of point mutations that engender a molecular switch in PAM pharmacology provides the first clues from the protein side of the equation as to the underlying determinants for this phenomenon. The prevalence of “molecular switches” raises concerns regarding metabolite pharmacology. A deeper understanding of the molecular basis of allosteric modulation has the potential to aid rational drug design efforts to predict and avoid undesirable pharmacology, including mode switches.

CHAPTER V

INVESTIGATING THE BINDING MODES OF METABOTROPIC GLUTAMATE RECEPTOR 5 (MGLU5) ALLOSTERIC MODULATORS FROM DIVERSE SCAFFOLDS

This chapter is based on a manuscript in preparation of the same title by Elizabeth Nguyen, Karen Gregory, Chrysa Malosh, Jeffrey Mendenhall, Brittney Bates, Meredith Noetzel, Emma Squire, Eric Turner, Kyle Emmitte, Shaun Stauffer, Jason Manka, Ya Zhou, Mark Turlington, Andrew Felts, Craig Lindsley, Jeff Conn and Jens Meiler.

Summary

The metabotropic glutamate receptor subtype 5 ($mGlu_5$) is a potential drug target for cognitive function disorders. While both positive and negative allosteric compounds have been discovered that target $mGlu_5$, many of these are not suitable for clinical use due to problems with selectivity and solubility. As a result, recent structure-activity relationship studies have produced a number of new small molecules from different chemical scaffolds that demonstrate a gain in selectivity and potency at $mGlu_5$. In addition, site-directed mutagenesis studies have been performed to determine which residues are functionally important for representative members of four of these chemical scaffolds, including 2-methyl-6-(phenylethynyl)-pyridine (MPEP), *N*-(3-chloro-2-fluorophenyl)-3-cyano-5-fluorobenzamide (VU0366248), 2-(1,3-benzoxazol-2-ylamino)-4-(4-fluorophenyl)pyrimidine-5-carbonitrile (VU0366058) and 4-nitro-*N*-(1,3-diphenyl-1*H*-pyrazol-5-yl)benzamide (VU29). The structural relationship between compounds from each of these four scaffolds and residues that have been found to be functionally important

is not well understood. In this study, computational docking of 32 ligands across four different chemical scaffolds is performed in a comparative model of $mGlu_5$ to understand how the structural and chemical properties of these compounds relate to functionally important residues in the receptor. In order to spatially and chemically compare binding modes of ligands with different chemical composition, the PropertyRMSD measure is introduced. A common binding mode exists for the ligands docked in this study that places the long axes of the ligands parallel to the transmembrane helices at the level of P654, Y658 and W784. The results indicate that hydrophobic interactions between Y658 and W784 with the phenyl groups of ligands from the MPEP series were consistently found. Important polar interactions are predicted between the fluorine on the benzamide group of ligands in the VU0366248 series and S808 as well as the cyano group on ligands from the VU0366058 series with Y658. Results from the computational modeling are consistent with functional mutagenesis data as well as structure activity relationship across a range of compounds and are a valuable tool in the synthesis of pharmacologic and chemical data.

Introduction

Glutamate is the primary excitatory neurotransmitter that guides activity in the mammalian central nervous system (Dingledine, Borges, Bowie, & Traynelis, 1999). Glutamate-gated cation channels called ionotropic glutamate receptors are responsible for the fast excitatory synapses within the CNS. With the discovery of metabotropic glutamate receptors, glutamate was found to participate in an even wider variety of key CNS functions through G proteins coupled to second messenger systems (Hollmann & Heinemann, 1994).

Cognitive function disorders are especially in need of new therapeutic strategies. Current drugs that commonly target dopaminergic and serotonergic pathways have

debilitating metabolic and neurological side effects (Rummel-Kluge et al., 2012, Tschoner et al., 2007). As an alternative to current strategies, glutamatergic systems have been shown to play a role in specific disorders of cognitive function, including anxiety (Porter et al., 2005), Parkinson's disease (Marino et al., 2003), fragile X syndrome (Dölen & Bear, 2008) and schizophrenia (Conn et al., 2009). Specifically, studies found that targeting $mGlu_5$ in fact suppressed hyperactivity and audiogenic seizure susceptibility in FXS mice models (Yan et al., 2005) and rescued cognitive function in FXS Drosophila models (McBride et al., 2005). Also, recent studies suggest that antagonists of the N-methyl-D-aspartate (NMDA) ionotropic glutamate receptor, such as phencyclidine (PCP) and ketamine, produce the same cluster of symptoms seen in schizophrenic patients (Conn et al., 2009, Tsai & Coyle, 2002). Further studies have found that NMDA receptors physically interact with $mGlu_5$ via binding to scaffolding proteins, producing a reciprocal positive-feedback system (Ehlers, 1999). Because $mGlu_5$ has been demonstrated to potentiate NMDA activity in the forebrain, it is hypothesized that activators of $mGlu_5$ may lead to the amelioration of symptoms associated with schizophrenia (Conn et al., 2009).

However, selective targeting of $mGlu_5$ has been a challenge. The glutamate binding site is highly conserved across all subtypes of mGlus and as a result, ligands binding the orthosteric site have a variety of adverse effects (Conn & Pin, 1997). An alternative approach is targeting allosteric binding sites of specific receptor subtypes (Gregory et al., 2011). Allosteric binding sites allow for indirect modulation of the $mGlu_5$ receptor that can also be highly subtype selective. Both positive (PAM) and negative (NAM) allosteric modulators for $mGlu_5$ have been discovered, providing an exciting novel approach to $mGlu_5$-specfiic therapeutic agents that can be used to improve cognitive function (O'Brien et al., 2004).

The first well-characterized potent, selective and systemically active NAM for $mGlu_5$ was 2-methyl-6-(phenylethynyl)-pyridine (MPEP) (Gasparini et al., 1999). MPEP is

an optimization of 2-methyl-6-(2-phenylethenyl)pyridine (SIB-1893), a compound that was identified in an automated high-throughput screen for antagonists of $mGlu_5$ (Varney et al., 1999). MPEP itself, while a useful tool for studying $mGlu_5$ because of its selectivity and because it is noncompetitive for glutamate, cannot be used as a drug because it blocks NMDA receptors (Cosford, Tehrani, et al., 2003). SAR optimization of the MPEP structure has led to the development of a variety of other NAMs within the same scaffold. Since the acetylene connecting the pyridine and phenyl rings was found to be necessary for high affinity binding, focus was placed on the substitutions on the two rings (Jaeschke, Wettstein, Nordquist, & Spooren, 2008). No changes were tolerated at the pyridine ring; the nitrogen in the 2-position was found necessary for antagonist activity, as was the position of the methyl at the 6-position. Replacing the 6-methyl with bromine or methoxy decreased activity. Also, any substitution at the 5-position resulted in inactivity (Alagille et al., 2005). However, 3-((2-methyl-4-thiazolyl)ethynyl)pyridine (MTEP), a 1,3-thiazole replacement of the pyridine ring, was discovered to be five times more potent via intraperitoneal injection (IP) and more selective than MPEP (Cosford, Tehrani, et al., 2003). Addition of a 3-pyridyl ring to MTEP as well 3-methoxy, 3-benzo or 3-nitro substitutions increased oral availability (Alagille et al., 2005, Roppe et al., 2004).

Substitutions at the phenyl ring of MPEP were tolerated at the 3-position. M-MPEP (2-(3-Methoxy-phenylethynyl)-6-methylpyridine) features a 3-methoxy at the phenyl ring and was found to significantly increase NAM activity (Alagille et al., 2005).Substitutions at the 3-methoxy to nitro, fluoro and cyano analogs increased activity further (Alagille et al., 2005, Kulkarni et al., 2009). While complete replacement of the phenyl ring with a pyrimidine reduced activity, the pyrimidine replacement in addition to a 3-methoxy or 3-bromo returned NAM activity (Alagille et al., 2005).

Building from the SAR of MPEP led to the discovery of a novel chemical scaffold that is competitive for the MPEP binding site, presented here as the VU0366248 (*N*-(3-

chloro-2-fluorophenyl)-3- cyano-5-fluorobenzamide) series. Substitution of the acetylene in MPEP to amide-linker between rings resulted in potent diarylamide $mGlu_5$ antagonists with 3-cyano substituents (Kulkarni et al., 2009). The 3-cyano-5-fluorobenzamide template was expanded upon and it was found that the heteroaryl amine with addition of a 6-methyl group improved potency. An even larger improvement occurred with addition of a 4-methylthiazole (Felts et al., 2010) or a 3-chlorophenyl (Felts et al., 2010).

The VU0366058 (2-(1,3-benzoxazol-2-ylamino)-4-(4-fluorophenyl)pyrimidine-5-carbonitrile) series featuring a novel 4-aryl-5-cyanopyrimidine scaffold was discovered through a virtual high-throughput screen from an artificial neural network (ANN) trained to predict the quantitative structure-activity relationship of chemical scaffolds (Mueller et al., 2012). This ANN in particular was trained over a dataset of 345 compounds to predict $mGlu_5$ NAMs using 35 different atom descriptors. The model was then used to virtually screen the ChemDiv Discovery Chemistry dataset of 708,416 compounds. From this screen came two 2-(2-benzoxazolylamino)-4-phenylpyrimidines predicted to be potent NAMs. The compound with a 5-cyano on the pyrimidine was further optimized with a 4-substituted fluorophenyl to result in a compound that was a potent, selective, noncompetitive, antagonist of $mGlu_5$ that showed efficacy and bioavailability via IP dosing (Mueller et al., 2012).

A high-throughput screen for PAMs and fragment library synthesis generated the *N*-(1,3-diphenyl-1*H*-pyrazol-5-yl)benzamide scaffold, which was the basis for the 3-cyano-N-(1,3-diphenyl-1H-pyrazol-5-yl)benzamide (CDPPB) series (Lindsley et al., 2004). CDPPB showed significant potentiation of glutamate-mediated $mGlu_5$ activation (Lindsley et al., 2004). To improve upon its potency and because it was a weak antagonist at $mGlu_{8,}$ SAR optimization was performed on CDPPB as was done for the MPEP series (Kinney et al., 2005). In an evaluation of 50 analogues of CDPPB, it was found that generally, substitutions at the pyridine ring seemed to influence efficacy and substitutions at the

phenyl ring influenced affinity (Kinney et al., 2005). A 6-methyl substitution on pyridyl ring was found to significantly increase efficacy in particular. Electronegative aromatic substituents in the para-position of the benzamide moiety increased potency (Kinney et al., 2005). Also, improvement were seen with a halogen atom in the ortho-position of the 1-phenyl ring (de Paulis et al., 2006).

A constant theme in the SAR optimization of $mGlu_5$ compounds is that dramatic differences in potency and efficacy, and even molecular switches between positive, negative and neutral modulation are observed with only slight changes to the molecule (Gregory et al., 2012, Wood et al., 2011). This may suggest different binding modes between ligands of different scaffolds for $mGlu_5$. The purpose of this study is to explore the structural determinants within $mGlu_5$ that are required for ligand binding within and across different chemical scaffolds. The tools used in this study include the combination of computational modeling of the $mGlu_5$ in complex with known allosteric modulators from different scaffolds with structure-activity relationship information and mutagenesis data.

Materials and Methods

Generating the comparative model of $mGlu_5$

The previously published comparative model of $mGlu_5$ (Gregory et al., 2013) was constructed using Rosetta 3.4, the protein structure prediction software package (Leaver-Fay et al., 2011). In brief, the X-ray crystal structure for human β2-adrenergic receptor (PDB ID: 2RH1) (Cherezov et al., 2007) was chosen as a template based on its high sequence similarity to $mGlu_5$. A profile to profile sequence alignment of TM regions between Class C hepta-helical transmembrane regions and Class A crystal structure templates was directly adopted from (Mühlemann et al., 2006), with the exception of

TM2,TM4 and TM7, which were based on the alignment of CaSR with Class A hepta-helical regions (Miedlich et al., 2004). The sequence alignment was used to thread the amino acid sequence of the $mGlu_5$ transmembrane helical region onto the backbone coordinates of the β2-adrenergic receptor. The loop regions between the helices were rebuilt in Rosetta using Monte Carlo Metropolis (MCM) fragment replacement combined with cyclic coordinate descent loop closure (CCD) (Canutescu & Dunbrack, 2009, Chu Wang et al., 2007). The resulting full sequence models were subjected to eight iterative cycles of side chain repacking and gradient minimization of ϕ, ψ and χ angles in Rosetta Membrane (Yarov-Yarovoy et al., 2005). Over 5,000 comparative models of $mGlu_5$ were generated and clustered for structural similarity using bcl::Cluster (Alexander et al., 2011). The lowest energy model from the largest cluster was used for ligand docking studies.

Generating ligand conformations

A total of 32 ligands from four different scaffolds were chosen for computational ligand docking in this study based on their efficacy, unique structure-activity relationship (SAR) to $mGlu_5$ and their role as a pharmacologic probe (Table 12). Twenty-one of the ligands are considered active modulators because of the level to which they potentiate glutamate in intracellular Ca^{2+} mobilization assays. The other 11 ligands are considered non-functional because they exhibit little to no glutamate modulation in $mGlu_5$. The chemical structures for each ligand can be found in the Appendix (Figures 58-61). In preparation for docking, conformers for each ligand were generated with MOE (Molecular Operating Environment, Chemical Computing Group, Ontario, Canada) using the MMFF94x force field and Generalized Born implicit solvent model. Conformers were generated using 10,000 iterations of the Low Mode MD method (Labute, 2010) with a redundancy cutoff of 0.25 Å. The number of conformers generated for each ligand was dependent on its number of rotatable bonds and is reported in Table 12.

Table 12
mGlu5 allosteric modulators from four diverse scaffolds chosen for this study.
Structures for each compound are found in the Appendix in Figures 58 through 61.

Series	Compound identifier (compound name where given)	Number of conformers generated	Potency (nM)
MPEP	1A	76	0.42
	1B	13	0.5
	1C	571	8
	1D	21	20
	1E (MPEP)	735	12
	1F	7	82
	1G (M5-MPEP)	459	114
	1H^{+}	651	1980
	1I^{+}	553	2400
	1J^{+}	82	>10000
VU0366248	2A	11	4.6
	2B	10	13.7
	2C	8	45
	2D	26	59
	2E (VU0366248)	8	347
	2F (VU0366249)	8	377
	2G	10	409
	2H^{+}	4	5440
	2I^{+}	12	>10000
VU0366058	3A	8	62
	3B	8	89
	3C (VU0366058)	8	91
	3D	22	216
	3E	16	223
	3F^{+}	32	>10000
	3G^{+}	8	>10000
	3H^{+}	16	>10000
VU29	4A (VU29)	7	10.7
	4B	13	20
	4C (CDPPB)	16	20
	4D	15	39
	4E	8	43
	4F	14	54
	4G	5	203
	4H^{+}	2	>10000
	4I^{+}	15	3410
	4J^{+}	87	3530

$^{+}$ligands that do not modulate glutamate response are considered non-functional

Docking allosteric modulators into the $mGlu_5$ comparative model

The 21 allosteric modulators from Table 12 that modulated glutamate response in functional assays with $mGlu_5$ were computationally docked into the comparative model of $mGlu_5$ using Rosetta Ligand (Davis & Baker, 2009, Lemmon & Meiler, 2012, Meiler & Baker, 2006). The initial ligand docking experiments started at a manually placed position centered at the proline residue at position 654. P654 was chosen because it was demonstrated to be functionally significant for each of the modulators (Table 13). The first round of docking allowed each modulator to sample binding modes in a 5.0 Å radius. Rigid body orientation of the ligand centroid is performed through translation during the low-resolution phase of docking until the geometric center of the ligand is in a position that does not conflict with the receptor. During high-resolution docking, the ligand undergoes 1000 cycles of full rotational freedom until the attractive and repulsive forces fall below a threshold value. Small 0.1 Å, 0.05 radian ligand movements are then simultaneously coupled with six cycles of side-chain rotamer and ligand conformer sampling in a Monte Carlo simulated annealing algorithm. All rotatable bonds within the ligand, except for planar conjugated bonds, were allowed full flexibility as indicated within the ligand parameters file. A final minimization combines backbone torsion angle minimization with harmonic constraints on the C-alpha atoms and side-chain rotamer sampling. The energy function used during the docking procedure contains terms for van der Waals attractive and repulsive forces, hydrogen bonding, electrostatic interactions between pairs of amino acids, solvation, and a statistical term derived from the probability of observing a side-chain conformation from the PDB.

After generating 5,000 models of $mGlu_5$-ligand complexes during the first round of ligand docking, the top 10% of models by ligand interaction energy was carried on to a second round of ligand docking as described above. The only difference during the second round was that the starting binding mode from the model generated in the first round was

used instead of a manual binding site assignment. After generating 5,000 models during a second round of docking, the top 10% of models again were used to seed a third round of docking. After the third round of docking, the top 10% of models were clustered based on ligand root mean square deviation (RMSD) with a cutoff of 3 Å (Alexander et al., 2011) and the center of each cluster with more than 2 members was used for further analysis. This iterative method of ligand docking maximizes the sampling of ligand and receptor flexibility to allow for unique low energy binding mode to be explored.

Analysis within and across ligand scaffolds based on structure and atom properties

After five rounds of iterative ligand docking, the final binding modes from each of the 21 ligands were compared within and across their scaffold families to detect common binding patterns. Within the binding modes for each ligand, the top 10% of models were clustered based on ligand RMSD (Alexander et al., 2011) and the center of each cluster was identified for all clusters with more than 2 members. Within each of the five scaffolds, the clusters centers were compared against each other with a new measure called PropertyRMSD. This new measure is used to distinguish between common binding modes across different ligands of the same scaffold, aligning ligand atoms in space as well aligning any user-defined atom properties. The measure is based loosely off the equation for $RMSD_{100}$, which normalizes the root mean square deviation between pairs of three-dimensional structures of different sizes (Carugo & Pongor, 2001). The equation has been modified as appropriate for ligands as such:

$$Property\ RMSD_N = \frac{PRMSD}{1 + a \cdot \log\frac{N}{B}}$$

where B is the number of atoms the user has specified should be aligned for a given pair of ligands, N is the actual number of atoms aligned, and a is computed such that the denominator equals 1/B at N = 1, to prevent non-physical (negative or infinite) RMSDs.

PRMSD is equivalent to the traditional formula for RMSD, but has been extended by distance between properties, e.g.

$$PRMSD = \sqrt{\sum_{n=1}^{N}(x_{1,n} - x_{2,n})^2 + (y_{1,n} - y_{2,n})^2 + (z_{1,n} - z_{2,n})^2 + \sum_{i} b_i (p_{i,1,n} - p_{i,2,n})^2}$$

where $x_{1,n}$ refers to the x-position of the n-th atom on the first molecule, $x_{2,n}$ refers to the second molecule, etc. $p_{i,1,n}$ refers to the i-th property value of interest for atom n on the first molecule; b_i is an adjustable, property-dependent constant with units 1/A that is used to adjust for the ranges of various properties and the relative penalty for difference in distance versus property. In this application, we used the product of sigma charge and van der Walls volume as the only property and b_i was set to 5, which was the inverse standard deviation of this property over the ensemble of molecules. Once pairwise PropertyRMSD values were calculated for all ligands within a particular scaffold, the ligands were clustered with a cutoff of 1 unit and the largest clusters, representing the most common binding modes within a particular scaffold, were further evaluated.

Docking non-functional ligands into the $mGlu_5$ comparative model

The 11 ligands from Table 12 that were considered non-functional because of their lack of glutamate modulation in intracellular Ca^{2+} mobilization studies were evaluated for their likelihood to bind $mGlu_5$ in the same mode occupied by the active modulators. As

with the set of functional ligands, non-functional ligands were computationally docked into the comparative model of $mGlu_5$ using Rosetta Ligand (Davis & Baker, 2009, Lemmon & Meiler, 2012, Meiler & Baker, 2006). The ligands started at the position occupied by the largest cluster from their scaffold. Each modulator was allowed to sample binding modes in a 5.0 Å radius and full rotational freedom. After generating 5,000 models, the top 10% of models were clustered based on ligand root mean square deviation (RMSD) with a cutoff of 3 Å (Alexander et al., 2011) and the center of each cluster with more than 2 members was used for further analysis.

Filtering based on functional mutagenesis data

The effect of $mGlu_5$ single-point mutations on ligand binding affinity estimates for a representative probe ligand from each scaffold was determined by Dr. Karen Gregory. Glutamate concentration-response curves (CRC) for single-point mutations of $mGlu_5$ with the addition of the different ligands were generated from a Ca^{2+} mobilization assay. Affinity estimates were calculated from progressive fold-shifts of the glutamate CRC by applying an operational model of allosterism (Gregory et al., 2012, K. Leach et al., 2007). Statistical analysis on the effect of mutations on affinity was compared to wild-type by one-way analysis of variation (ANOVA) and Dunnett's post-test. Mutations causing a significant decrease or loss of affinity were considered functionally important residues for the binding of that ligand scaffold. The binding mode represented by the cluster center from each scaffold was compared to the results from this experimental data.

Table 13
Functionally critical residues significantly reduce PAM and NAM activity when mutated.
When mutated, residues listed here caused a significant decrease or loss of affinity of PAM or NAM during glutamate-mediated mobilization of intracellular Ca2+ compared to wild-type mGlu5. Experiments were performed by Dr. Karen Gregory.

Location in receptor	MPEP series (MPEP)[+]	VU0366248series (VU0366248)	VU0366058 series (VU0366058)	VU29 series (VU29)
TM3	G651F	G651F		
	P654S	P654S		
	P654F	P654F	P654F	P654F
	Y658V	Y658V	Y658V	
TM5	P742S			
	L743V			
		N746A		
	G747V			G747V
TM6	T779A			
	T780A			T780A
	W784A	W784A	W784A	
		F787A		
	V788A			
TM7	S808A	S808A		
	S808T			
	A809V	A809V		A809V
	A809G			A809G
	T810A			

[+]from (Gregory et al., 2013)

For details on the $mGlu_5$ comparative modeling and ligand docking protocol, including command lines and links to input and output files, refer to the protocol capture for Chapter V provided in the Appendix.

Results and Discussion

Frequently sampled binding modes are in proximity to functionally critical residues

To determine the preferred binding modes for active modulators of $mGlu_5$ within a particular chemical scaffold, ligands potentiating or inhibiting glutamate response were docked into a previously published comparative model of $mGlu_5$ (Gregory et al., 2013). After three rounds of iterative docking with full receptor and ligand flexibility that allowed

for ligand translation in a 5 Å radius, binding modes for ligands within each scaffold were clustered according to similarities in spatial and chemical properties with the PropertyRMSD quality measure. The correct binding mode has been identified within clusters containing at least 1% or more of the total binding modes in studies with Class A GPCRs (see Chapter III); as such, the largest clusters for each scaffold in this study were evaluated in detail (Figure 31). A wide range of binding modes were sampled by modulators in each scaffold, indicating that the allosteric binding pocket is able to accommodate a range of ligand conformers and sizes. This is not surprising, given the range of chemical scaffolds that are competitive for the common allosteric site in $mGlu_5$. For MPEP (Figure 31A) and CDPPB (Figure 31D) in particular, binding modes where the ligand lies beyond the transmembrane helices are observed. While these binding modes are a possible result of the computational docking studies, they are unsupported due to the body of experimental work supporting an allosteric binding site that lies within the mGlu transmembrane helical bundle (Gregory et al., 2013, 2011, Malherbe, Kratochwil, Zenner, et al., 2003, Pagano et al., 2000).

Notably, the binding mode from the largest cluster from each scaffold is within interaction distance of the residues indicated by mutagenesis to be functionally important (Figure 32). A common binding pocket is found to encompass members of the four diverse scaffolds. This binding mode generally places the long axes of the ligand parallel to the transmembrane helices at the level of P654, Y658 and W784. Similarities in ring placement are seen between ligands from different scaffolds as well. The phenyl rings in ligands from the VU0366248 series structurally align with the pyridine and phenyl rings of MPEP. The pyrimidine on VU0366058 aligns with the phenyl from MPEP. In CDPPB, there is a linear relationship between one of the symmetric diphenyl groups and the benzamide. The phenyl group in CDPPB aligns closely with the phenyl of MPEP, whereas the benzamide is in approximate position to the pyridine of MPEP.

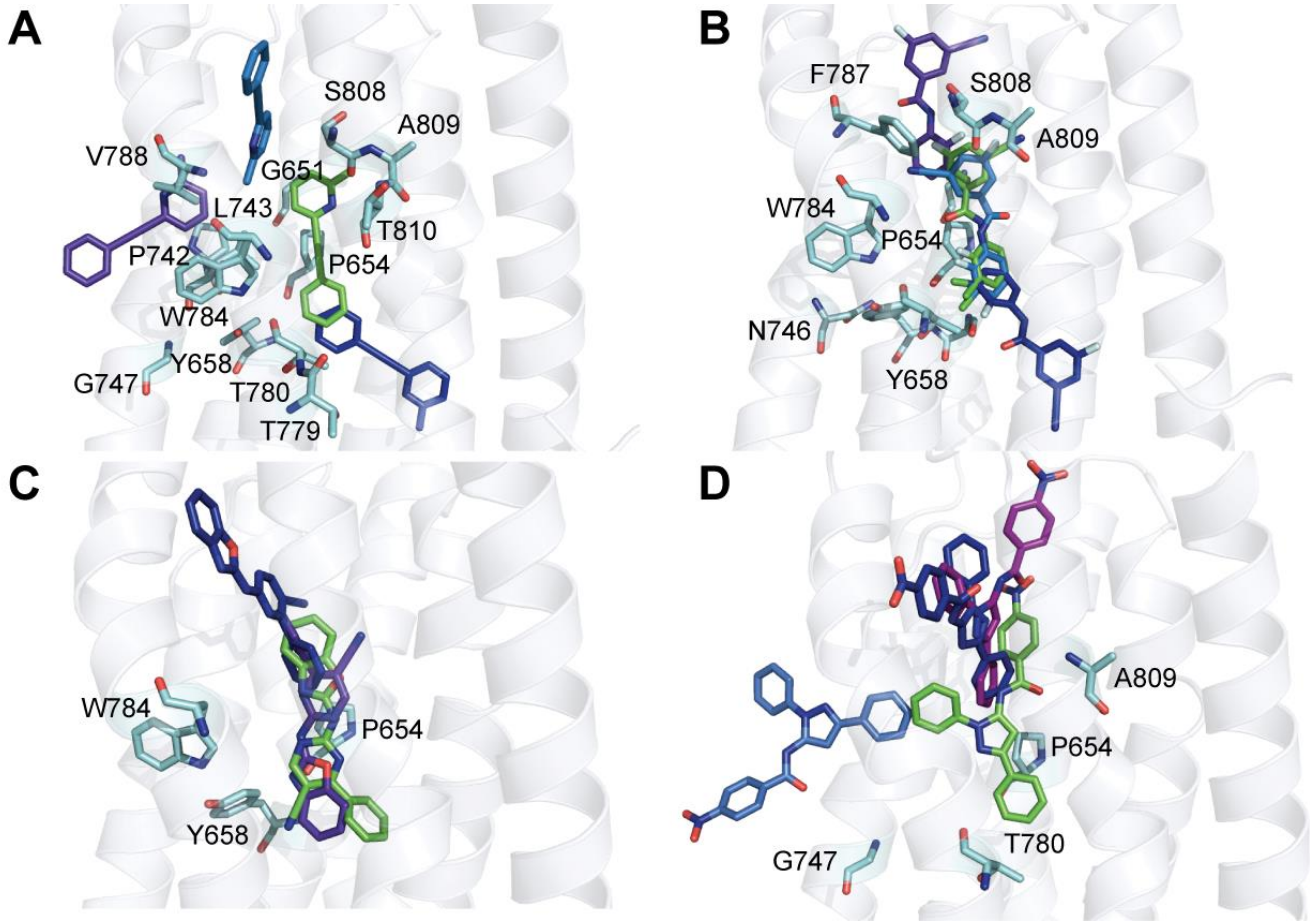


Figure 31
Binding modes from the largest cluster for each scaffold are within interaction distance of functionally important residues. Representative members from the largest clusters for each scaffold are shown docked into a comparative model of mGlu5. The binding mode from the largest cluster is shown in green for the A) MPEP series, B) VU0366248 series, C) VU0366058 series and D) VU29 series. Binding modes from other large clusters are shown in blue and purple. Side chains for functionally important residues as indicated in Table 13 are shown in cyan.

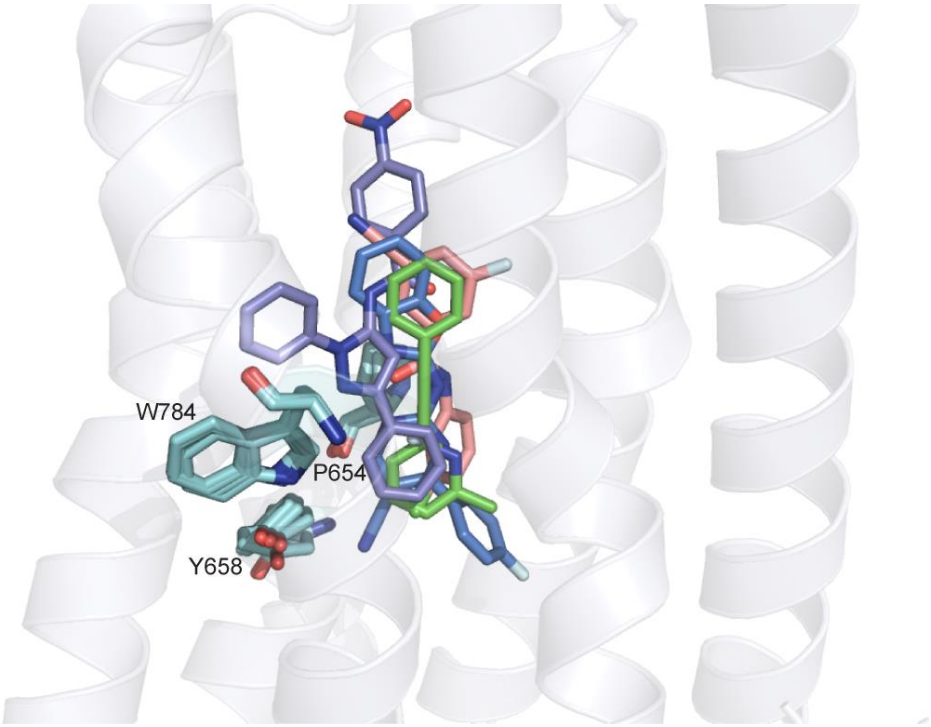


Figure 32
A representative ligand from the binding mode of the largest cluster for each scaffold align within an allosteric site supported by functional data.
MPEP is shown in green, VU0366058 is in blue, VU0366248 is in pink and VU29 is in purple. The side chain conformations for P654, Y658 and W784 from each binding mode are shown in cyan.

Different ligands within a single scaffold align in spatial and chemical properties

Within the largest cluster for each scaffold, representatives from each ligand were determined, as seen in Figure 32. Within the MPEP series, the top cluster contained representatives from all seven ligands that were docked from that scaffold with consistent alignment between the two rings separated by the acetylene (Figure 33A). The common binding mode for all the ligands from the MPEP series sits in range of all the residues found to be functionally important for MPEP, bounded by S808 and Y658 (Gregory et al., 2013). For 1B and 1D, the additional phenyl ring is in range of π-π stacking interactions with Y658. However, it is difficult to determine from the results whether the pyridine ring points towards or away from the extracellular surface, which has been seen in previous

studies (Gregory et al., 2013). The ligands from this series are small and rigid compared to the binding pocket and as such, both orientations may be possible.

Within the VU0366248 series, consistent alignment between the amide and cyano functional groups was observed (Figure 33B). The fluoro group on the benzamide is in range of forming hydrogen bonds with S808. However, the largest cluster only contained representatives from 2C, 2E and 2F. Unlike the other ligands from the VU0366248 scaffold, these three ligands in particular all have the chloro-group in the 3-position of the phenyl ring. These chemically similar ligands are found to dock in a consistent binding mode that is energetically unfavorable for other ligands within this scaffold.

The top cluster for the VU0366058 series was found to represent all five ligands in the scaffold and held a binding mode in which the cyano and amide groups were consistently aligned. In this binding mode, the cyano group points towards Y658, a functionally important residues for VU0366058 (Figure 33C). Besides the Y568 residues, however, there is little mutational data on VU0366058 to confirm or deny the binding mode provided by the docking studies.

For the VU29 series, the top binding mode represents all but the structurally dissimilar 4E, which is much larger due to the addition of a fourth phenyl group. Within the common binding mode, the position of the diphenyl groups is conserved, as is the alignment of the benzamide ring (Figure 33D). There are no clear interactions between the ligands and the functionally important residues for ligands within this series. For this bulky scaffold, steric hindrance may prohibit alternative binding modes in which the ligands are able to actively modulate glutamate signal.

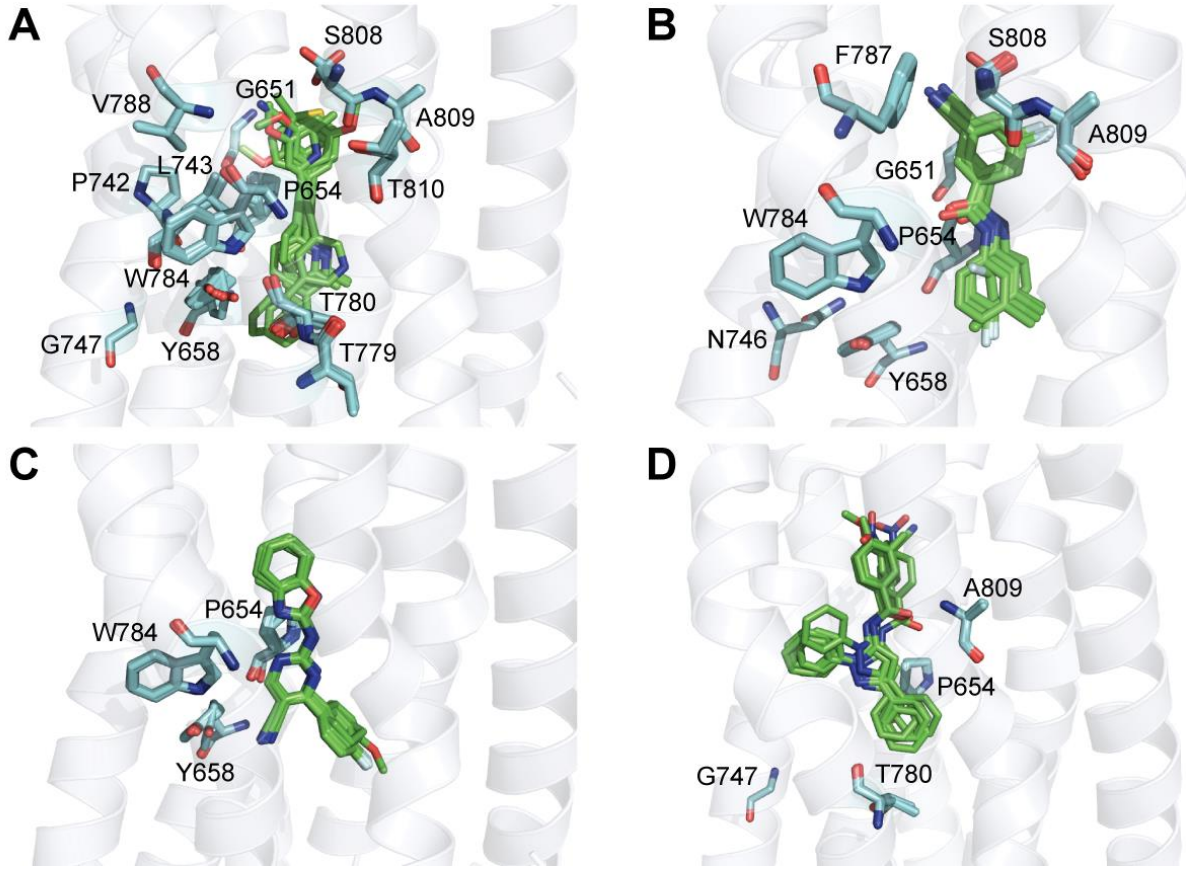


Figure 33
The largest cluster for each scaffold aligns different ligands by structural and chemical similarity.
Members from the largest cluster for each scaffold are shown in green, docked into a comparative model of mGlu5. Functionally important residues for allosteric modulation are shown in cyan sticks. A) MPEP series, B) VU0366248 series, C) VU0366058 series and D) VU29 series.

Non-functional ligands from the MPEP and VU0366248 series can accommodate binding modes of active ligands

Ligands from SAR studies that do not demonstrate the ability to modulate glutamate response were evaluated to see whether they could bind in the same mode as those within their chemical scaffold. These non-functional ligands were allowed to dock with full receptor and ligand flexibility within a radius of 5 Å, as with the active ligands. The top cluster from the docking results for each non-functional ligand was compared to the common binding mode from its scaffold.

Non-functional ligands that were tested from the MPEP scaffold (1H, 1I and 1J) contained the phenyl and pyridine rings with the acetylene linker, similar to the active ligands from the same scaffold. Docking studies were performed to explore why a change in the position of the methoxy on the pyridine group (as in 1H), the addition of methyl acetate (as in 1I) or the addition of a benzenesulfonic acid, 4-methyl-, methyl ester group (as in 1J) greatly reduces glutamate modulation as suggested by SAR studies (Alagille et al., 2005, Kulkarni et al., 2009). For ligands in the VU0366248 group, removal of chloro group (as in 2H) or substitution of the phenyl with an adamantane (as in 2I) greatly reduces ligand functionality (Felts et al., 2010, Kulkarni et al., 2009).

For ligands in both the MPEP and VU0366248 series, non-functional ligands demonstrated the preference to dock in the same binding mode as their functionally active counterparts (Figure 34). For both series, the ligands show limited conformational range due to the linear preference of the acetylene and amide linkers between the two rings. Docking results show a preference for these rigid ligands to bind with their long axis parallel to the helices to avoid clashing with side chains in the receptor. As a result, the positioning of the rings is highly conserved in both series.

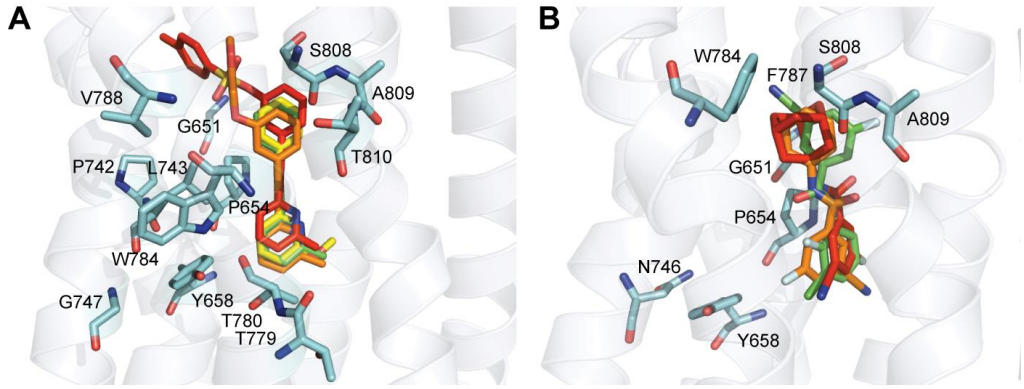


Figure 34
Non-functional ligands from the MPEP and VU0366248 series accommodate the same binding mode as functional ligands.
The common binding mode for active modulators from each scaffold is shown in green, and non-functional ligands are shown in yellow, orange and red. Residues known to be functionally important for MPEP and VU0366248 are shown in cyan. A) Ligands from the MPEP series. MPEP is shown in green, 1H is in yellow, 1I is in orange and 1J is in red. B) Ligands from the VU0366248 series. VU0366248 is in green, 2H is in orange and 2I is in red.

Steric hindrance may prevent activity in non-functional ligands from the VU0366058 series

Three non-functional ligands from the VU0366058 scaffold were evaluated for their ability to accommodate the common binding mode observed among the active ligands in the scaffold. The addition of two methoxy groups on the phenyl ring (as in 3F) and the substitution of the phenyl to a pyridine (as in 3G) or naphthalene (as in 3H) reduces functionality substantially (Mueller et al., 2012). In evaluating the non-functional ligands from the VU0366058 scaffold, added bulk due to the addition of two methoxy groups (as in 3F) or an extra phenyl ring (as in 3H) flips the ligand such that the benzoxazole faces towards the intracellular side of the receptor. As seen in Figure 35A, 3F and 3H has also been pushed further from the functionally important residues into an alternate binding mode due to steric hindrance. 3G is similar in size and chemical structure to the active ligands and as a result, it is able to accommodate the same general binding mode as the active ligands from the VU0366058 scaffold as seen in Figure 35B. However, the

positioning of the nitrogen in the benzoxazole is flipped, which may suggest interactions with the heteroatom that reduces the functionality of the ligand.

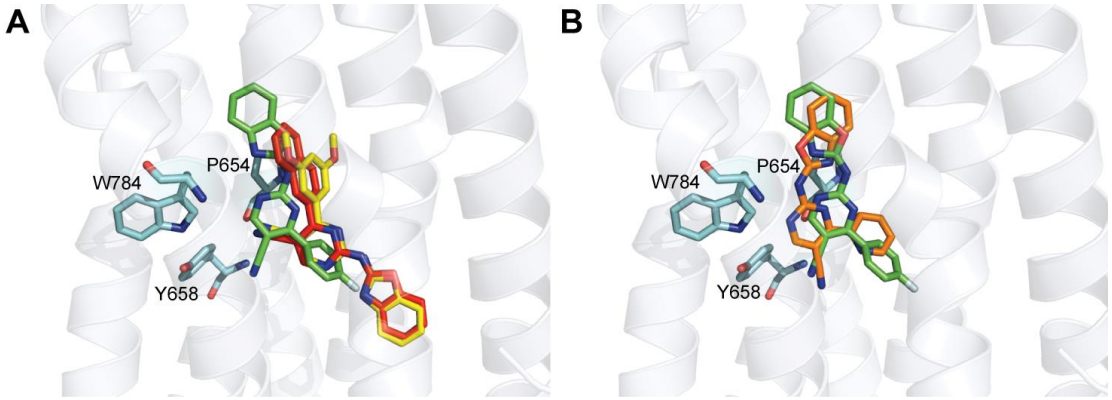


Figure 35
Non-functional ligands from the VU0366058 series docked into the comparative model of mGlu5.
The common binding mode for active modulators from the VU0366058 series is shown by VU0366058 in green. Residues known to be functionally important for VU0366058 are shown in cyan. The top ranked binding mode for A) 3F is shown in yellow, 3H is in red, and B) 3G is shown in orange.

Position of phenyl groups are highly conserved in functional and non-functional ligands from the VU29 series

Three non-functional ligands from the VU29 series were introduced into the binding site of active modulators through docking studies. SAR studies found that the substitution of one of the phenyl groups to a pyridine (as in 4H) causes a loss in potency, as well as the substitution of the phenyl ring of the benzamide to a cyclopentane (as in 4I) or the addition of two methoxy groups onto the benzamide (as in 4J) (Kinney et al., 2005). With the heteroatom substitution, the binding mode of 4H was flipped such that the phenyl ring of the benzamide acted as the counterpart to the other phenyl ring (Figure 36A). However, changes to the benzamide resulted in no differences in the preferred binding mode of 4I or 4J (Figure 36B). This result suggests that the position of the phenyl groups seem to be highly conserved. These results also demonstrate a preference of the functional group of

the ligands in the VU29 series to be pointing towards the extracellular surface in the receptor core where there is room for bulky substituents to be accommodated.

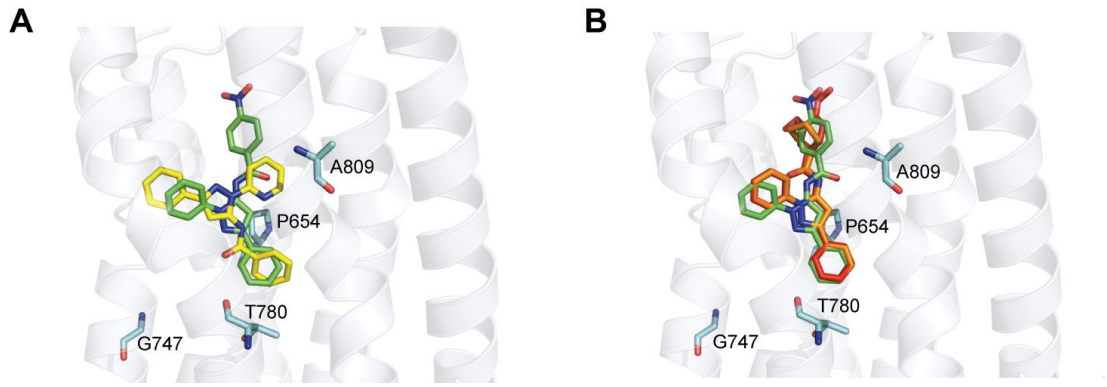


Figure 36
Non-functional ligands from the VU29 series docked into the comparative model of mGlu5.
The common binding mode for active modulators from the VU29 series is shown by VU29 in green. Residues known to be functionally important for VU29 are shown in cyan. The top ranked binding mode for A) 4H is shown in yellow, B) 4I is shown in orange and 4J is shown in red.

Conclusions

In this study, a set of ligands known to positively or negative modulate $mGlu_5$ activity from four different chemical scaffolds were computationally docked into a previously published comparative model of $mGlu_5$ (Gregory et al., 2013) to understand the structural and chemical binding determinants on both the receptor and the small molecules. The proposed binding modes from docking studies were evaluated in light of site-directed mutagenesis studies done on $mGlu_5$ with probe ligands as well as SAR studies from each scaffold. In order to compare common binding modes across different ligands within the same scaffold, a new measure was introduced called PropertyRMSD that not only compares ligand conformations by their spatial similarly, but also by conservation of chemical properties. When using PropertyRMSD as the quality measure during clustering, conserved ligand SAR becomes a key factor when determining favored

binding modes. This is different than traditional RMSD calculations, which only capture structural similarities between the common atoms of ligands.

The results found here indicate that the top cluster for each scaffold is in line with the functionally important residues from mutagenesis studies. The ligands bound in a common binding pocket surrounded by P654, Y658 and W784, even when they are dissimilar in chemical structure. The long axes of the ligands were found to be parallel to the transmembrane helices. Hydrophobic interactions between Y658 and W784 with the phenyl groups of ligands from the MPEP series were consistently found within compounds with the scaffold. Important polar interactions were predicted between the fluorine on the benzamide group of ligands in the VU0366248 series and S808 as well as the cyano group on ligands from the VU0366058 series with Y658. These results are in agreement with findings on other MPEP-based compounds, such as the role of W784, Y658 and S808 on 2-, 3- and 4-BisPEB (Molck et al., 2012).

These results validate the computational model in light of functional studies as they demonstrate that the ligand docking favors sites that are also relevant in functional mutagenesis studies. These results also supports previous studies that favor clustering analysis for computational docking studies over evaluation by energy scores alone (Kaufmann & Meiler, 2012). Particularly with low-resolution docking studies done with comparative models, the ability for the Rosetta energy function to discriminate between relevant binding modes is decreased. Evaluating binding modes that are highly sampled through clustering provides biologically relevant results across a variety of systems (Kaufmann et al., 2009, Kaufmann & Meiler, 2012, London & Schueler-Furman, 2007).

Furthermore, clustering based on the PropertyRMSD measure allowed for illumination of important chemical determinants of each scaffold. Highly conserved chemical groups within the analysis of each top cluster includes the phenyl and pyrimidine rings in the MPEP series, the amide and cyano groups in the VU0366249 and VU0366058

series and the phenyl rings in the VU29 series. These groups are also the ones that are conserved during SAR studies on these scaffolds and as such, it is not surprising that they are aligned in the top binding modes for each scaffold. For MPEP, changes at the pyridine ring were not tolerated during SAR substitutions (Alagille et al., 2005). The most potent ligands in the VU0366248 series were found to have the 3-cyano substitution on the benzamide ring (Kulkarni et al., 2009). Similarly, the 5-cyano substitution was found to be necessary for potency at the pyrimidine (Mueller et al., 2012). For PAMs in the VU29 series, substitutions on the phenyl ring was found to be highly sensitive to changes in potency (Lindsley et al., 2004).

To determine whether chemically similar ligands with little to no modulator activity would accommodate the same binding mode as functionally active ligands, two to three non-functional ligands were docked with the $mGlu_5$ comparative model at the position given by the top cluster for each scaffold. Even when given full range of translational and rotational flexibility, non-functional ligands for ligands in each scaffold were found to prefer the same binding mode as the functional ligands. Small differences included the introduction of the heteroatom in 4H of the VU29 series and 3G of the VU0366058 group which flipped the ligand, suggesting alternate hydrogen bond formations. Also, added bulk through methoxy groups in 3F or an additional phenyl group in 3H caused steric hindrances that pushed the ligand to an alternate binding mode. However, all other non-functional ligands were able to dock in the same position as that of their functional counterparts. If the binding modes for these non-functional ligands are truly similar to that of the functional ligands, this may explain why the SAR of these ligands is so steep and why small changes to chemical structure can lead to compounds with a large range of functionality, even switches between positive, negative and neutral modulation (Wood et al., 2011). Alternatively, the scoring function used in the computational docking of these ligands may not be able to capture the subtle changes in SAR for the non-functional

ligands and as a result, showed no differences in the binding modes of active versus inactive modulators. This suggests that relevant interactions between ligands and the side chains of the receptor are not being captured by the Rosetta energy function during docking and that only steric effects are being considered.

The comprehensive docking study presented here with 32 functional and non-functional ligands across a diverse set of four chemical scaffolds highlight the ability for the $mGlu_5$ allosteric site to accommodate ligands of different size and chemical content. In addition, computational docking with receptor and ligand flexibility was found to sample relevant binding modes that interact with functionally important residues as determined by mutagenesis. Clustering the binding modes with PropertyRMSD highlights important chemical properties on the ligand that are also found to be critical in SAR studies. The agreement between the docking studies with both functional mutagenesis and small-molecule SAR validate the data provided the pharmacology, chemistry and computational work that can not only apply to future drug discovery efforts for mGlus, but across different therapeutically relevant systems as well.

CHAPTER VI

CHARACTERIZING THE ROLE OF CYSTEINE RESIDUES ON METABOTROPIC GLUTAMATE RECEPTOR 5 (MGLU5) FUNCTION

Summary

The metabotropic glutamate receptor subtype 5 ($mGlu_5$) plays an important role in the regulation of cognitive function. Allosteric modulators that bind $mGlu_5$ hold promise as subtype-selective therapeutics for schizophrenia and other cognitive disorders. Because they do not bind the orthosteric site, these allosteric compounds offer the possibility of treatment without off-target effects. While structure-based drug discovery would assist in the development of such compounds, there are no experimentally determined structures of Class C GPCRs to date. Substituted cysteine accessibility method (SCAM) and disulfide cross-linking studies are established approaches for determining experimental restraints to confirm the orientation, relative positioning and membrane depth for helices involved in the allosteric binding site. Information about the ligand binding pocket can be directly determined using thiol-reactive ligands. These experiments require a functional $mGlu_5$ receptor with a cysteine-less background, the first of which is described here. All cysteine residues were systemically substituted in the N-terminal domain truncated “headless” $mGlu_5$ receptor and the effect of single point mutations as well as multiple cysteine residue mutations were evaluated through functional and radioligand binding assays. The sensitivity of $mGlu_5$ receptors to methanethiosulfonate (MTS) reagents was also evaluated for the full and headless wild-type and mutant $mGlu_5$ receptors. Two cysteine residues were found to be sensitive to substitution, C630 and C781. Single point mutations of these residues as well as mutation in a background of an otherwise cysteine-less receptor

demonstrated reduction in the potency of positive allosteric modulators. Despite this, a headless $mGlu_5$ receptor with a cysteine-less background was constructed that retained functionality with positive allosteric modulators. An $mGlu_5$ receptor construct with one cysteine, C781, potentiates glutamate response with positive allosteric modulators as well as retains concentration-specific binding of a radiolabelled ligand.

Introduction

Glutamate is a major excitatory neurotransmitter in the mammalian central nervous system (CNS) (Dingledine et al., 1999). Glutamate initiates a wide variety of key CNS functions when bound to metabotropic glutamate receptors, a member of the class C G protein-coupled receptor (GPCR) family (Hollmann & Heinemann, 1994). The heterogeneous distribution and diverse physiological roles of mGlu subtypes provide an opportunity to develop therapeutic agents that will selectively target and modulate a range of neurological and psychiatric disorders (Conn & Pin, 1997, Schoepp, Jane, & Monn, 1999). The metabotropic glutamate receptor subtype 5 ($mGlu_5$) is involved in mammalian cognitive function through diverse signaling pathways that modulate synaptic plasticity (Hollmann & Heinemann, 1994). Selective modulators of $mGlu_5$ have the exciting potential for the development of novel treatment strategies for disorders that disrupt cognitive function (Conn et al., 2009).

Because the orthosteric binding site, located in the extracellular N-terminal domain, is well conserved between all subtypes of mGlus relative to other regions on the receptor (Conn & Pin, 1997), an approach to selectively target $mGlu_5$ is to identify ligands with allosteric binding sites (Gasparini et al., 2002). Ligands that bind to an allosteric site and enhance receptor activity are known as positive allosteric modulators, or PAMs. Ligands inhibiting receptor activity are known as negative allosteric modulators, or NAMs. Negative allosteric modulators (NAMs) MPEP and fenobam alleviate cognitive and

neuropsychiatric impairments in patients with fragile X syndrome (Dölen & Bear, 2008, McBride et al., 2005, Yan et al., 2005). Positive allosteric modulators (PAMs) such as CPPHA have been shown to improve hippocampus-dependent spatial learning and have antipsychotic efficacy in behavioral model systems, holding promise as therapeutic agents for schizophrenia (Ayala et al., 2009, Kinney et al., 2005).

However, progress towards clinically applicable drugs targeting $mGlu_5$ is hampered by the particularly steep structure activity relationships (SAR) in allosteric modulators. Small changes in allosteric ligands can cause complete loss of activity or a "molecular switch" from a PAM to NAM or NAM to PAM (Wood et al., 2011). While structure-based drug design may help to shed light on the complex SAR of $mGlu_5$ allosteric modulators, a crystal structure of a Class C GPCR has yet to be determined. Computational models of $mGlu_5$ have helped increase understanding of the allosteric binding pocket of $mGlu_5$ (Gregory et al., 2013, Malherbe, Kratochwil, Zenner, et al., 2003, Mühlemann et al., 2006). However, there is limited confidence in the resolution of the computational models due to the low sequence identity between $mGlu_5$ and the Class A GPCR structural templates. Experimental restraints are necessary to provide structural information that will validate the computational models that have been generated for $mGlu_5$. In particular, structural restraints obtained through experiments such as substituted cysteine accessibility method (SCAM) (Karlin & Akabas, 1998) and disulfide cross-linking studies (Wess, Han, Kim, Jacobson, & Li, 2008) would provide a way to confirm the orientation, relative positioning and membrane depth for helices involved in allosteric binding. A thiol-reactive ligand tagged with a methanethio-sulfonate (MTSL) spin label bound to a cysteine can indicate interactions between a ligand and residue on the receptor, providing direct information about the allosteric binding site (Loo & Clarke, 1997).

The prerequisite for the abovementioned studies is an $mGlu_5$ background where the cysteine residues have been removed. Once a cys-less receptor background has been

constructed, cysteines can be mutated back into the receptor at residues that would be important to probe for structural restraints, such as residues within the proposed binding site of a ligand. A cys-less model of the entire $mGlu_5$ receptor could not remain active, since there is a crucial cysteine-rich domain that links the extracellular N-terminal domain to the transmembrane helices which transmits receptor activation via a conserved disulfide bridge (X. Liu et al., 2004, Muto et al., 2007, Rondard et al., 2006). However, the "headless" $mGlu_5$ with truncation of the extracellular N-terminal domain has been demonstrated to behave like the wild-type receptor in terms of G protein coupling (Goudet et al., 2004). In the headless receptor, PAMs act as agonists and NAMs act as inverse agonists (Y. Chen et al., 2007). Therefore, a cys-less system constructed with the headless $mGlu_5$ receptor would still be an important tool for structural studies of the allosteric modulator binding. Within the headless $mGlu_5$ receptor, there are 14 cysteine residues as opposed to the 38 native cysteine residues in the full receptor. Ten cysteine residues are in the transmembrane (TM) region (C587, C617, C625, C630, C633, C643, C701, C753, C781, C802) and four are in the loop regions (C680, C690, C732, C815). Of the 14 cysteine residues, two are stabilized by a conserved disulfide bond between TM3 and the second extracellular loop (EC2) (C643, C732). All of these cysteine residues are conserved amongst the group 1 mGlus ($mGlu_1$ and $mGlu_5$) except for C625, C633, C690 and C701. The cysteine residues conserved across all mGlus are C643, C732, C753 and C781.

The purpose of this study is to construct a functional headless $mGlu_5$ receptor with a cysteine-less background. In doing so, the role of cysteine residues in the function of $mGlu_5$ was determined. While all cysteine residues were successfully mutated to either an alanine, serine or valine, the mutation of C630 and C781 significantly decreased the efficacy of PAMs compared to the wild-type receptor. These residues are therefore postulated to be important for ligand interaction.

Materials and Methods

Materials

Dulbecco's Modified Eagle's Medium (DMEM), fetal bovine serum (FBS) and antibiotics were purchased from Invitrogen (Carlsbad, CA). [^{3}H] methoxyPEPy (76.3 Ci/mmol) was custom synthesized by PerkinElmer Life and Analytical Sciences (Waltham, MA). CDPPB (3-Cyano-N-(1,3-diphenyl-1H-pyrazol-5-yl) benzamide), CPPHA (N-{4-chloro-2-[(1,3-dioxo-1,3-dihydro-2H-isoindol-2-yl)methyl]phenyl}-2-hydroxybenzamide), VU29 (4-nitro-N-(1,3-diphenyl-1H-pyrazol-5-yl)benzamide), VU0415051 (N-tert-butyl-6-[2-(3-fluorophenyl)ethynyl]pyridine-3-carboxamide), VU0357121 (4-butoxy-N-(2,4-difluorophenyl)benzamide), VU0360172 (N-cyclobutyl-6-((3-fluorophenyl)ethynyl)nicotinamide hydrochloride), VU0360173 (6-((3-fluorophenyl)ethynyl) pyridin-3-yl)(3-hydroxyazetidin-1-yl)methanone), VU0404211 (1-(4-(2,4-difluorophenyl)piperazin-1-yl)-2-((4-fluorobenzyl)oxy)ethan-1-one), VU0364289 (2-{4-[2-(benzyloxy)acetyl]piperazin-1-yl}benzonitrile), VU0405386 (N-(tert-butyl)-5-((3-fluorophenyl)ethynyl)picolinamide), VU0405398 ((5-((3-fluorophenyl)ethynyl)pyridin-2-yl)(3-hydroxyazetidin-1-yl)methanone) and VU0403602 (N-cyclobutyl-5-((3-fluorophenyl)ethynyl)ethynyl) picolinamide hydrochloride) were all synthesized in-house using previously reported methods (Y. Chen et al., 2007, 2008, Gregory et al., 2012, 2013, Kinney et al., 2005, Rodriguez et al., 2010, Xiong et al., 2010, Y. Zhou et al., 2010). Unless otherwise stated, all other reagents were purchased from Sigma-Aldrich (St. Louis, MO) and were of an analytical grade.

Site-directed mutagenesis and cell culture

Site-directed mutagenesis was used to substitute cysteine residues of the wild-type rat $mGlu_5$ to serine, valine or alanine in pCI:Neo (Quikchange II, Agilent, Santa Clara, CA). Mutations were verified by sequencing. Wild-type and mutant rat $mGlu_5$ receptor constructs were transfected into HEK293A cells, using Fugene6™ (Promega, Madison, WI) as the transfection reagent. Polyclonal stable cell lines were derived for rat $mGlu_5$ mutant constructs by maintaining the cells at sub-confluence for a minimum of four passages in the presence of 1 mg/ml G418 (Mediatech, Manassas, VA). Stably transfected cell lines were subsequently maintained at 37°C in complete DMEM supplemented with 10% fetal bovine serum (FBS), 2 mM L-glutamine, 20 mM HEPES, 0.1 mM Non-Essential Amino Acids, 1 mM sodium pyruvate, antibiotic-antimycotic and 500 µg/ml G418 in a humidified incubator containing 5% CO_2, 95% O_2.

Intracellular Ca^{2+} mobilization assay

A day before assays were run, HEK293A-rat $mGlu_5$ cells were seeded at 50,000 cells/well in poly-D-lysine coated black-walled, clear bottom 96 well plates in assay medium (DMEM supplemented with 10% dialyzed fetal bovine serum, 20 mM HEPES and 1 mM sodium pyruvate). On the day that the assay was run, the cell permeant Ca^{2+} indicator dye Fluo-4 (Invitrogen, Carlsbad, CA) was used to assay receptor-mediated Ca^{2+} mobilization as described previously (Hammond et al., 2010) using a Flexstation II (Molecular Devices, Sunnyvale, CA). A 5-point smoothing function was applied to the raw fluorescent Ca^{2+} traces and basal fluorescence of individual wells determined during the first 20 sec.

Radioligand binding assay

Radioligand binding assays were performed on cell membranes prepared from HEK293A cells expressing wild-type and mutant rat $mGlu_5$. The cells were harvested through trypsinization and pelleted through centrifugation for 3 min and 300*xg*. Cell pellets were then resuspended in ice-fold homogenization buffer (50mM Tris-HCl, 10mM EDTA, 0.9% NaCl, pH7.4). The suspended cell pellets were then homogenized using three 10 se bursts, separated by 30 sec periods on ice, with a Tekemar TP-18/10S1 homogenizer (Teledyne Tekmar, Cincinnati, OH). Cell fractions were then separated through centrifugation for 10 minutes at 1000g. The supernatant was centrifuged for 1 hr at 30,000*xg* and the resulting pellet was then resuspended in ice-cold Ca^{2+} assay buffer. For saturation binding experiments, membranes (20-50 µg/well) were incubated with a range of [^{3}H]-3-methoxy-5-(pyridin-2-ylethynyl)pyridine ([^{3}H]methoxyPEPy) concentrations (0.5 nM-60 nM) for 1 hr at room temperature with shaking in Binding Buffer (50 mM Tris-HCl, 0.9% NaCl, pH7.4). MPEP (10 µM) was used to determine non-specific binding. For inhibition binding experiments, membranes were incubated with ~2 nM [^{3}H]methoxyPEPy and a range of concentrations of test ligand (100 pM-100 µM) in Ca^{2+} assay buffer with 1% dimethylsulfoxide (final concentration) with shaking. The radioligand binding assays were terminated by rapid filtration through GF/B Unifilter plates (PerkinElmer Life and Analytical Sciences, Boston, MA) using a Brandel 96-well plate Harvester (Brandel Inc., Gaithersburg, MD), and three washes with ice-cold Binding Buffer, separating bound from free radioligand. Plates were allowed to dry overnight and radioactivity counted using a TopCount Scintillation Counter (PerkinElmer Life and Analytical Sciences, Boston, MA).

Evaluating substituted-cysteine accessibility during Ca^{2+} mobilization assay with MTS reagents

The Ca^{2+} mobilization assay was run with wild-type $mGlu_5$ in the presence of [2-(trimethylammonium)-ethyl] methanethiosulfonate (MTSET), methanethiosulfonate-ethylammonium (MTSEA) and (2-sulfonatoethyl)-methanethiosulfonate (MTSES) to evaluate the effect of methanethiosulfonate (MTS) reagents on PAM potency. As with the previously mentioned Ca^{2+} mobilization assay, HEK293A-rat $mGlu_5$ cells were seeded at 50,000 cells/well in poly-D-lysine coated black-walled, clear bottom 96 well plates in assay medium (DMEM supplemented with 10% dialyzed fetal bovine serum, 20 mM HEPES and 1 mM sodium pyruvate) a day before running the assay. On the day of the assay, cells were incubated with MTS reagents for 2 min at room temperature. Concentrations of the specific MTS reagents were 0.5 mM MTSET, 2.5 mM MTSEA, and 5 mM MTSES. The plate was washed after 2 min to dilute the MTS reagent and slow the reaction as described in previous literature (Liapakis, Simpson, & Javitch, 2001). The cell permeant Ca^{2+} indicator dye Fluo-4 (Invitrogen, Carlsbad, CA) was then used to assay receptor-mediated Ca^{2+} mobilization as described previously (Hammond et al., 2010) using a Flexstation II (Molecular Devices, Sunnyvale, CA). A 5-point smoothing function was applied to the raw fluorescent Ca^{2+} traces and basal fluorescence of individual wells determined during the first 20 sec.

Statistical analysis

All computerized nonlinear regression was performed using Prism 5.01 (GraphPad Software, San Diego, CA). Inhibition [^{3}H]methoxyPEPy binding data sets were fitted to a one-site inhibition binding model and estimates of inhibitor dissociation constants (K_I) were derived using the Cheng-Prusoff equation for competitive ligands (Cheng et al., 1973) and

the following version of the allosteric ternary complex model for ligands that did not fully displace radioligand (Lazareno & Birdsall, 1995):

$$\frac{Y}{Y_{max}} = \frac{[D]}{[D] + \frac{K_D\left(1 + \frac{[B]}{K_B}\right)}{\left(1 + \frac{\alpha[B]}{K_B}\right)}}$$

where Y/Y_{max} is the fractional specific binding, D is the radioligand concentration, B is the molar concentration of the allosteric modulator, K_D is the radioligand equilibrium dissociation constant, and K_B is the allosteric modulator equilibrium dissociation constant. α denotes the cooperativity factor, where values of α > 1 describe positive cooperativity, values of α < 1 (but greater the 0) denote negative cooperativity and α = 1 denotes neutral cooperativity.

Allosteric modulator and agonist concentration-response curves were fitted to a four parameter logistic equation in order to determine potency estimates:

$$y = \frac{bottom + (top - bottom)}{(1 + 10^{(logEC_{50} - A)HillSlope}}$$

where *bottom* and *top* are the lower and upper plateaus, respectively, of the concentration-response curve, HillSlope is the Hill coefficient that describes the steepness of the curve, and EC_{50} is the molar concentration of modulator required to generate a response halfway between the *top* and *bottom.*

All affinity and potency parameters were estimated as logarithms and are expressed as the mean ± S.E.M. (Christopoulos, 1998). Statistical analyses were

performed where appropriate as indicated using one-way ANOVA with Dunnett's post-test when comparing to control, or Tukey's post-test when making multiple comparisons.

Results

Evaluation of mutant receptor functionality through PAM potency compared to wild-type

In our construction of a headless $mGlu_5$ receptor with a cysteine-less background, we chose to evaluate functionality as a measure of how a mutant receptor affected the potency of PAMs. PAMs act as agonists in the headless receptor, and so a significant (greater than 10-fold) decrease in agonist activity at the mutant receptor indicated a lack functionality. Mutant receptors that demonstrated PAM potency within 10-fold of the wild-type receptor were regarded as functional. PAMs with the highest potency available at the headless $mGlu_5$ receptor were used so that a change in potency would be unambiguous. These included CDPPB (3-cyano-N-(1,3-diphenyl-1H-pyrazol-5-yl) benzamide), CPPHA (N-{4-chloro-2-[(1,3-dioxo-1,3-dihydro-2H-isoindol-2-yl)methyl]phenyl}-2-hydroxybenzamide), VU29 (4-nitro-N-(1,3-diphenyl-1H-pyrazol-5-yl)benzamide), VU0415051 (N-tert-butyl-6-[2-(3-fluorophenyl)ethynyl]pyridine-3-carboxamide), VU0357121 (4-butoxy-N-(2,4-difluorophenyl)benzamide), VU0360172 (N-cyclobutyl-6-((3-fluorophenyl)ethynyl)nicotinamide hydrochloride), VU0360173 (6-((3-fluorophenyl)ethynyl) pyridin-3-yl)(3-hydroxyazetidin-1-yl)methanone), VU0404211 (1-(4-(2,4-difluorophenyl)piperazin-1-yl)-2-((4-fluorobenzyl)oxy)ethan-1-one), VU0364289 (2-{4-[2-(benzyloxy)acetyl]piperazin-1-yl}benzonitrile), VU0405386 (N-(tert-butyl)-5-((3-fluorophenyl)ethynyl)picolinamide), VU0405398 ((5-((3-fluorophenyl)ethynyl)pyridin-2-yl)(3-hydroxyazetidin-1-yl)methanone) and VU0403602 (N-cyclobutyl-5-((3-fluorophenyl)ethynyl)ethynyl) picolinamide hydrochloride). The potency of these PAMs at wild-type are included in Tables 1, 2, X and Y.

Single cysteine substitutions result in functional mutants

The effect of each individual cysteine residue was evaluated through single point mutations. Each cysteine was first mutated to alanine and functional effects were evaluated through an intracellular Ca^{2+} mobilization assay in response to known positive allosteric modulators. Experiments were performed by Dr. Karen Gregory, Emma Square and Elizabeth Nguyen. All cysteine residues except for C630, C617 and C781 tolerated substitution to alanine with no appreciable effect on PAM potency (Table 14). Serine mutations were evaluated at the three cysteine residues where alanine was not tolerated, and C630S and C617S were found to be functional and respond to PAMs (Table 14). However, it was not until the valine mutation was made that C781V was able to regain functionality in response to PAMs (Table 14, Figure 62). Full evaluation of C781V with additional PAMs VU0404211, VU0415051, VU0405386 and CPPHA is given in the Appendix (Table 23, Figure 62).

Table 14
PAM potency (pEC50) as agonists (initial peak) at N-truncated mGlu5-wt and single cysteine mutants.
Data represents the mean ± S.E.M. of 3-5 experiments performed in triplicate, unless indicated otherwise. Data collected by Dr. Karen Gregory, Emma Square and Elizabeth Nguyen.

Mutation	VU29	VU0403602	VU0360172	VU0415051	VU0405398	VU0405386
mGlu5 wt	7.02 [b]	8.01 [b]	7.02±0.09	7.73±0.12	6.65±0.11	8.05±0.29
C587A	6.66 [b]	8.25 [b]	7.11±0.15	7.61 [a]	6.62±0.13	7.86±0.04
C625A	7.66 [b]	9.03 [b]	7.73±0.11	8.03 [a]	7.01±0.05	8.24 [a]
C633A	7.15 [b]	8.51 [b]	7.44±0.15	7.77±0.05	6.78±0.14	7.86±0.10
C680A	6.86 [b]	8.27 [b]	7.16±0.15	7.55 [a]	6.80±0.06	7.74±0.06
C690A	7.01 [b]	8.49 [b]	7.44±0.15	7.74 [a]	6.80±0.20	7.91±0.09
C701A	7.05 [b]	8.69 [b]	7.36±0.25	7.76 [a]	6.87±0.09	8.06 [a]
C815A	6.58 [b]	8.47 [b]	6.87±0.08	7.23±0.03[#]	6.43±0.05	7.14±0.21[#]
C753A	6.17 [b]	7.99 [b]	6.78±0.08	7.00±0.07[#]	6.25±0.15	7.10±0.20[#]
C802A	6.44 [b]	7.64 [b]	6.82±0.13	7.24±0.10[#]	6.16±0.17	7.45±0.04
C630A	6.13 [b]	7.34 [b]	6.54±0.16	6.90±0.09[#]	6.29±0.17	7.10±0.09[#]
C630S	6.63 [b]	8.28 [b]	7.25±0.05	7.56 [a]	6.76±0.12	7.87 [a]
C617A	No PAM[# b]	No PAM[# b]	7.09	N.D.	6.78	N.D.
C617S	6.89 [b]	8.54 [b]	7.15±0.09	7.42±0.08	6.69±0.04	7.50±0.11
C781A	No PAM[# b]	No PAM[# b]	No PAM[#]	N.D.	No PAM[#]	N.D.
C781S	No PAM[# b]	No PAM[# b]	No PAM[#]	N.D.	No PAM[#]	N.D.
C781V	5.98[# b]	N.D.	7.39	N.D.	N.D.	N.D.

[#] represents mutants at which a greater than 10-fold shift decrease in potency compared to wild-type was observed
[a] mean of n=2 determinations only
[b] single determination.
N.D., not determined

Effect of multiple cysteine residue substitutions on the headless $mGlu_5$ receptor

Cysteine substitutions were made consecutively in the headless $mGlu_5$ receptor to generate constructs with multiple cysteine substitutions. The first cysteine mutation was C680A and the following mutations were generated in the order indicated by Figure 37. Functionality of the constructs with six (5+C587A), nine (8+C617S), ten (9+C630S), eleven (10+C815A) and twelve (11+C781V) mutations were evaluated by intracellular Ca^{2+} mobilization in response to PAMs. As seen in Table 15, PAM potency remained unchanged for the majority of PAMs at the $mGlu_5$ constructs 5+C587A and 8+C617S. PAM potency begins to drop below 10-fold for VU0360172 and VU0405386 at the 9+C630S construct. At the $mGlu_5$ construct with 11 mutations (10+C815A), PAM potency was maintained at all but VU0405398, VU0360172 and VU0405386. At the $mGlu_5$ construct with all 12 non-disulfide bonded cysteine residues mutated to alanine, serine or

valine, PAM potency showed less than 10-fold change from wild-type with CPPHA, VU0360172, VU0404211 and VU0415051 but a loss in potency occurred at VU29 and a 10-fold decrease in potency occurred at VU0405386. Concentration response curves for cys-less headless $mGlu_5$ characterization are found in the Appendix (Figure 66).

While a cysteine-less construct is ideal for designing experiments, a receptor with one cysteine remaining can already be used to collect experimental restraints. Because the 10+C815A construct demonstrated greater stability with less changes in PAM potency, this construct was further functionally characterized and used in radioligand binding experiments. Complete characterization of the $mGlu_5$ 10+C815A construct with 13 PAMs is reported in the Appendix (Table 24, Figure 63-65).

C680A

C690A

C701A

C633A

C625A

C587A

C802A

C753A

C617S

C630S

C815A

C781V

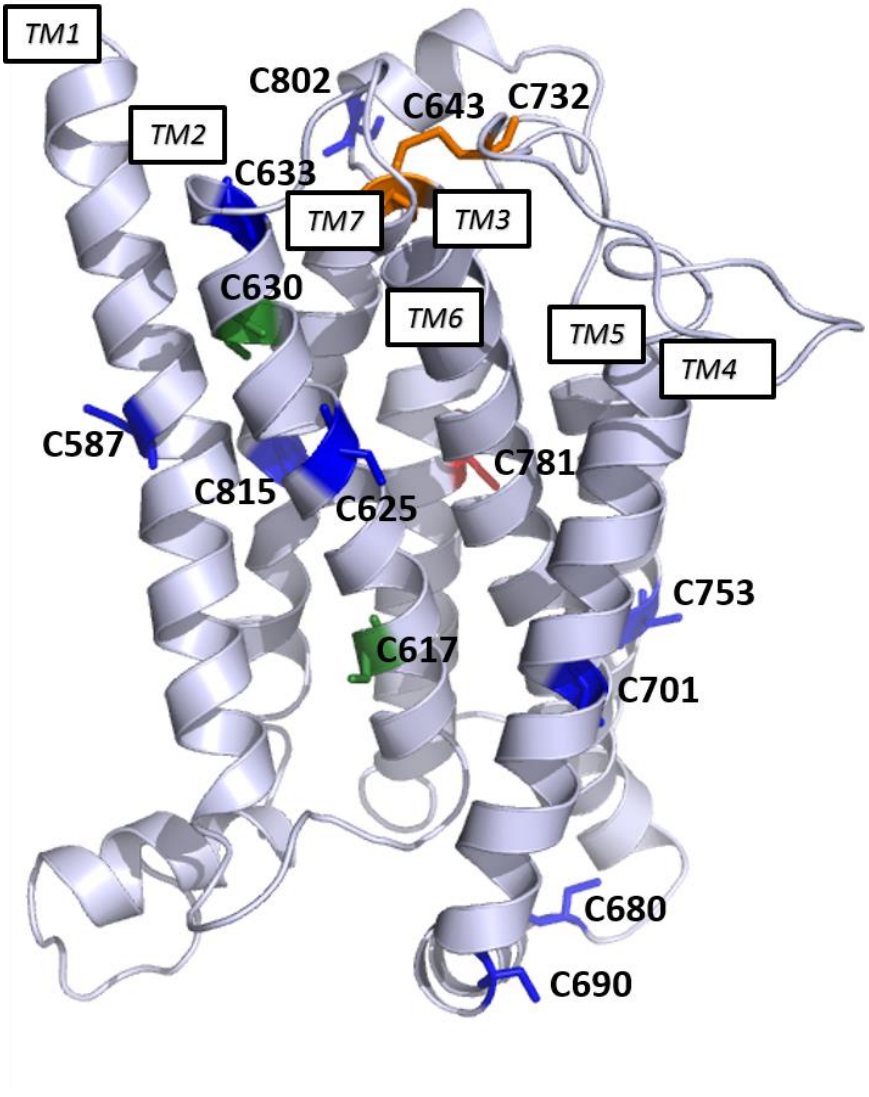


Figure 37
Order of consecutive cysteine substitutions performed on the headless mGlu5 construct. Cysteine residues from the mGlu5 sequence are shown mapped onto the structural backbone of the β2-adrenergic receptor X-ray crystal structure (2RH1). Residues in blue were mutated to alanine, residues in green were mutated to serine, residues in red were mutated to valine and residues in orange are involved in disulfide bonding.

Table 15

PAM potency (pEC50) as agonists (initial peak) at N-truncated mGlu5-wt and multiple cysteine mutant constructs.

Data represents the result of n=1 experiments performed in triplicate, unless indicated otherwise. Data collected by Dr. Karen Gregory, Emma Square and Elizabeth Nguyen.

Mutant	CPPHA	VU29	VU0415051	VU0360172	VU0404211	VU0405386	VU0405398	VU0403602
Wild-type	4.87	7.02	7.73±0.12 [b]	7.02±0.09 [b]	5.46	8.05±0.29 [b]	6.65±0.11 [b]	8.01
5+C587A	5.20	7.63	7.71 [a]	7.77 [a]	6.41	7.71 [a]	7.29 [a]	8.98
8+C617S	4.41	6.18	7.08	6.79	4.48	6.93[#]	N.D.	N.D.
9+C630S	N.D.	N.D.	6.37	5.87[#]	N.D.	6.77[#]	N.D.	N.D.
10+C815A	4.95	5.73	6.76	6.48[#]	3.76	6.91[#]	5.43[#]	7.07
11+C781V	5.49	No PAM[#]	6.83	6.58	4.60	6.93[#]	N.D.	N.D.

[#]represents mutants at which a greater than 10-fold shift decrease in function compared to wild-type was observed

[a] mean of n=2 determinations

[b] mean ± S.E.M. from 3-5 experiments

N.D. not determined

Effect of multiple cysteine residue substitutions on saturation binding experiments

The headless $mGlu_5$ 10+C815A construct was evaluated using saturation binding experiments with the radioligand ^{3}[H]methoxyPEPy. Concentration-dependent binding was exhibited by the 10+C815A construct, as shown in Figure 38. Binding affinity, however, was reduced compared to the wild-type receptor. The maximum specific binding, estimated by Bmax, was 20.88 fmol/mg for the 10+C815A construct as opposed to 90.22 fmol/mg for WT. The equilibrium binding constant Kd was calculated to be 64.11nM for the 10+C815A construct, two times more than the Kd at the wild-type receptor at 30.60nM. While these values indicate decreased binding affinity for the mutant construct, repeated experiments will be necessary to determine if binding affinity is significantly altered.

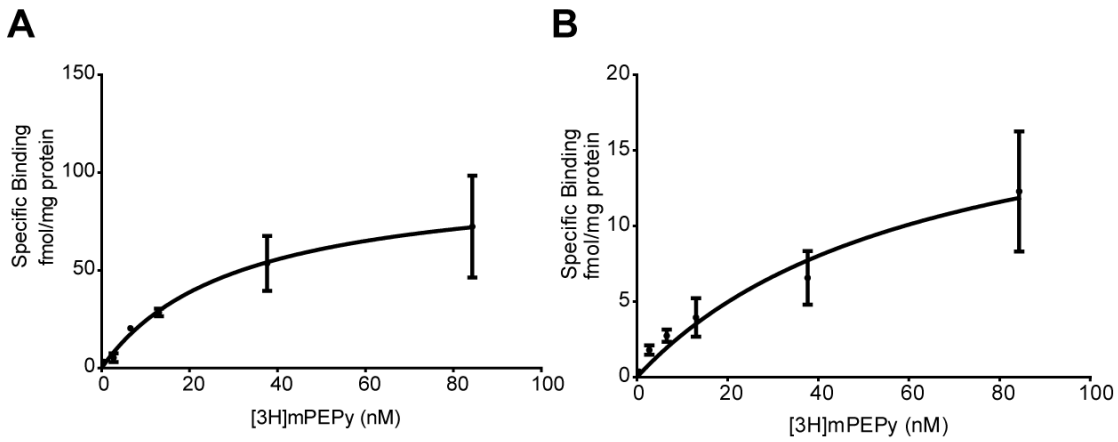


Figure 38
Concentration depending binding demonstrated by six point saturation binding curves for the 10+C815A mGlu5 construct.
Saturation binding experiments were conducted with 3[H]methoxyPEPy in HEK293 cell membranes expressing A) wild-type headless mGlu5 WT membranes at 250 ug and B) the headless 10+C815A mGlu5 construct at 500 ug. Data is the result of n=1.

Sensitivity to MTS reagents

MTS reagents are commonly used in EPR and thiol-reactive experiments as labels to tag specific cysteine residues in a receptor with an otherwise cysteine-less background. To test its applicability to the $mGlu_5$ system, the sensitivity of the full and headless $mGlu_5$

wild-type receptors to MTS reagents was evaluated. The ideal concentration of MTS reagents for use with the truncated and full $mGlu_5$ receptors was first optimized by varying the amount of MTS reagent during intracellular Ca^{2+} mobilization assay with PAMs VU0360172 and VU0415051 (see Appendix, Figure 68-69). The optimal concentrations resulted in 2.5mM for MTSEA, 5mM for MTSES and 0.5mM for MTSET. These concentrations were used in assaying for PAM potency in the presence of MTS reagents during intracellular Ca^{2+} mobilization. While the reduction in functionality of VU0360172 and VU0415051 due to the effect of the MTS reagents was not significant, it was consistent across the various MTS reagents for the headless $mGlu_5$ WT receptor (Table 6). Difficulty was encountered in gathering consistent data from Ca^{2+} mobilization with the MTS reagents during assay with the full length WT receptor and the results from the experiments often did not converge to allow for analysis. Difficulty in collected this data may be due to the MTS reagents exhibiting fluorescence themselves. Full concentration response curves are found in the Appendix (Figure 70-71).

Table 16
PAM potency (pEC50) as agonists (initial peak) at N-truncated mGlu5-wt and multiple cysteine mutant constructs in the presence of MTS reagents.
Data represents the mean ± S.E.M. of 3 experiments performed in triplicate.

$mGlu_5$ construct	VU0360172	VU0415051
HANT WT	7.38±0.30	7.25±0.26
HANT WT + MTSEA	6.55±0.48	6.67±0.26
HANT WT + MTSES	7.14±0.18	7.08±0.23
HANT WT + MTSET	7.12±0.22	6.98±0.21
Full WT	6.23±0.49	Not converged
Full WT + MTSEA	Not converged	8.42±0.55
Full WT + MTSES	Not converged	6.74±0.63
Full WT + MTSET	5.84±0.67	Not converged

Discussion and Conclusion

In this study, receptor constructs resulting from the systematic mutation of cysteine residues in the headless $mGlu_5$ receptor were characterized by how they affect PAM

potency and radioligand binding. Mutation of C781 was found to reduce PAM activity, while the 10+C815A construct potentiated glutamate response with positive allosteric modulators CPPHA, VU29, VU0415051, VU0404211 and VU0403602 as well as retained concentration-specific binding of the radiolabelled ligand ^{3}[H]methoxyPEPy.

The aim of this work was to generate a headless $mGlu_5$ receptor with a cysteine-less background to be used for further structural studies including use with thiol-reactive ligand probes for binding site detection, EPR measurements for distance restraints and substituted cysteine accessibility measurements and disulfide cross-linking studies for investigations into the receptor's helical structure. Cysteine-less receptor constructs have been generated for Class A GPCRs and have found success in collecting valuable structural data. SCAM has been used to map residues lining the binding site of the dopamine D2 receptor using MTS reagents (Javitch, Shi, & Liapakis, 2002) and has more recently been used to identified eight residues on prolactin-releasing peptide receptor (PrRPR) that were crucial for binding and function of PrRP (Rathmann, Pedragosa-Badia, & Beck-Sickinger, 2013). Disulfide cross-linking studies have revealed conformational changes that occur in rhodopsin and the M3 muscarinic receptor as a result of activation (Hubbell, Altenbach, Hubbell, & Khorana, 2003, S. D. Ward, Hamdan, Bloodworth, & Wess, 2002). Rhodopsin has also been extensively studied using EPR distance restraints (Van Eps et al., 2011, Hubbell et al., 2003). Only one study has generated a cysteine-less mutant of a class C GPCR, in which three cysteine mutations were made in $mGlu_6$ (Judith, 2012). This $mGlu_6$ cys-less system, however, has yet to be applied to any structural studies.

In performing single point mutations of cysteine residues, C781 was found to be particularly sensitive to substitution, not able to tolerate mutation to either alanine or serine. Valine substitution, however, was able to maintain PAM potency with VU0360172, VU0404211, VU0415051 and CPPHA. According to computational models of $mGlu_5$,

C781 is within range of the allosteric binding pocket and as such, it may play a key role in maintaining the structural integrity of the pocket. These results are also in agreement with cysteine substitutions done with $mGlu_6$, where the corresponding cysteine at position 793 on TM6 was found to be important for agonist response (Judith, 2012). C630 and C617 also demonstrated sensitivity to alanine substitution, which may also indicate their importance in maintaining the structure of $mGlu_5$.

Cysteine substitutions were performed consecutively to create $mGlu_5$ receptor constructs with multiple cysteine mutations. Constructs of $mGlu_5$ with six and nine substituted cysteine residues demonstrated little to no fold change in PAM potency, indicating that these cysteine residues do not influence the binding pocket of the receptor. Receptor constructs with 11 and 12 substituted cysteine residues maintained PAM potency within 10 fold for several PAMs; however, a significant drop in potency was seen with VU0405398, VU0360172 and VU0405386 in the construct with 11 substituted cysteine residues. For the cys-less headless $mGlu_5$ receptor construct, no PAM activity was seen with VU29 and a significant drop greater than 10-fold was seen with VU0405386. This is likely due to the sensitivity of the residues mentioned above during the single cysteine substitution screening. C630 and C781 are among the last residues to be substituted in the construction of the cys-less mutant and are likely changing the conformation of the receptor in such a way that impact PAM activity. To recover PAM potency, other constructs can be considered such as one with C630V. Also, leaving the cysteine at position 630 in place would still provide a system in which structural experiments can be performed without compromising receptor functionality.

The 10+C815A $mGlu_5$ receptor construct with 11 substituted cysteine residues was evaluated for sensitivity to radioligand binding by ^{3}[H]methoxyPEPy. While binding was reduced as seen by saturation binding experiments, there was concentration-dependent binding occurring in the mutant receptor constructions. This demonstrates further that the

allosteric binding pocket in these mutant receptors are still able to bind ligands and as such, they would be useful for further studies probing the binding pocket further.

The purpose of mutating out the cysteine residues is to provide a cys-less background in which cysteine residues can be strategically placed. These cysteine residues can then be tagged with MTS reagents to retrieve restraints between residues in the receptor or between a tagged ligand and a residue in the receptor. In order to perform experiments with MTS reagents, $mGlu_5$ sensitivity to these reagents must be characterized in light of the functional and binding assays performed with the receptor. The optimal concentration of MTS reagents were determined for use with $mGlu_5$ and their ability to block the binding pocket of allosteric modulators was evaluated in wild-type receptors. Although little change or even an increase was seen in PAM potency in the presence of MTS reagents, this may be due to an artificial increase in potency reported by the Ca^{2+} mobilization assay as a result of the fluorescence exhibited by the MTS reagents themselves. Alternative second messenger functional assays such as inositol phosphate (IP) accumulation should be considered instead when evaluating $mGlu_5$ with MTS reagents.

Structural information on Class C GPCRs is currently limited by the lack of experimental information on the receptors. An $mGlu_5$ receptor construct with a cysteine-less background would provide a crucial tool necessary for structural exploration of the receptor through SCAM, disulfide cross-linking, EPR and thiol-reactive ligand studies. Our results indicated that such a cysteine-less receptor construct can be both functionally active and able to bind ligands. Two cysteine residues in particular, C630 and C781, are important for allosteric modulation and may be necessary to maintain for optimal receptor functionality. Further studies will include a more thorough characterization of the cys-less headless $mGlu_5$ receptor with a C630V mutation and alternate functional assays to prevent false positive results from MTS reagents.

CHAPTER VII

INCORPORATION OF SEQUENCE-BASED EXPOSURE PREDICTION IN DE NOVO MEMBRANE PROTEIN STRUCTURE PREDICTION

This work is based on the manuscript in preparation of the same title by Elizabeth Nguyen, Jeff Mendenhall, Brian Weiner, Kelly Gilmore and Jens Meiler.

Summary

The number of short-range contacts for an amino acid affects its placement in tertiary structure during protein structure prediction and as a result, should facilitate in protein structure prediction. This is especially true for helical membrane proteins, where amino acids with many neighboring residues are oriented on the buried face of the helix while residues with few neighbors are oriented towards the membrane. An artificial neural network-based neighbor count prediction method was developed for membrane protein sequences. The model predicts the correct neighbor count to within 2.3 residues of the neighbor count calculated from the protein structure. The neighbor count prediction method was incorporated directly into a *de novo* structure prediction algorithm for membrane proteins, BCL::MP-Fold. Sampling of membrane protein folds was improved in 78% of cases based on RMSD100, although sampling of models based on contact recovery remained unchanged. However, with the inclusion of the neighbor count prediction, top ranked models had higher contact recovery in 67% of cases. The utility of neighbor count prediction, as illustrated in this study, motivate additional ideas for refining our prediction method results even further.

Introduction

Membrane proteins play an important role in signal transduction, substrate transport, ionic and proton gradient maintenance and other important biological processes. Their importance is reflected in the fact that they comprise over 25% of the human genome (Fagerberg, Jonasson, von Heijne, Uhlén, & Berglund, 2010) and are the target for over 50% of drugs on the market today (Overington et al., 2006). However, investigations of membrane protein structures using X-ray crystallography and NMR have been stalled by the difficulty in achieving a stable form of the protein, often because of their large size and lack of solubility (Caffrey, 2003). To date, only 2% of the Protein Data Bank (PDB) is comprised of membrane proteins. Nevertheless, advances in computational methods for membrane protein structure prediction have allowed for the investigation of new membrane protein topologies.

The challenging problem of computational *de novo* structure prediction of membrane proteins has seen a great deal of progress in recent years. With an implicit membrane potential incorporated into the protein structure prediction software Rosetta, twelve α-helical multi-pass membrane proteins with less than 150 residues were predicted to a root mean square deviation (RMSD) of less than 4 Å from the native structure (Yarov-Yarovoy et al., 2005). High-resolution modeling using Rosetta of three membrane protein domains with less than 150 residues achieved accuracy up to 2.5 Å RMSD (Barth, Schonbrun, & Baker, 2007). For the structure prediction of larger membrane proteins, secondary structure element assembly with BCL::MP-Fold achieved the correct topology in 29 of 34 cases, while RosettaMembrane was able to sample the correct topology in just 25 cases (Weiner et al., 2013). With current *de novo* structure prediction methods, however, contact recovery above 20% is still difficult to achieve. Helices are often placed correctly within the membrane, but their rotation relative to each other is incorrect (Weiner et al., 2013). Helical orientation is important when using the models in ligand docking

studies, especially with proteins such as G protein-coupled receptors where the binding pocket is formed specifically by the pore-facing sides of the helices (Katritch et al., 2013). Differentiating between sides of a helix with residues that prefer to form contacts with other residues versus being solvent or membrane exposed should aid *de novo* structure prediction to achieve more accurate structure predictions on proteins with high contact-ordering.

Incorporating the general preference of a particular amino acid class to be buried versus exposed has already shown to improve protein structure prediction (Simons et al., 1997, Woetzel et al., 2012). A knowledge-based environment potential per amino acid class, calculated over a dataset of known protein structures, has been applied to multiple protein structure prediction algorithms. During *de novo* structure prediction of a protein, the environment of a particular residue is compared to that of its amino acid class. Because direct calculation of the solvent accessible surface area of a residue can be computationally expensive, a commonly approximation of a residue's environmental exposure is the inverse of its neighborhood density, which is the weighted sum of a residue's neighboring atoms (Stouten, Frömmel, Nakamura, & Sander, 1993). Evaluating structure quality using a residue's environment, estimated by the number of beta-carbon atoms within 10 Å, was found to increase the quality of structure prediction simulations using Rosetta (Simons et al., 1997). Calculating neighborhood density between residue centroids, defined as pseudo-atoms at the center of mass of the side-chain, improved structure prediction quality even further (Rohl, Strauss, Misura, et al., 2004). In fold recognition methods, calculating the neighbor count over a 14 Å radius between beta-carbon atoms and comparing it to amino acid class-based environment potentials was found to achieve the best results (Karchin, Cline, & Karplus, 2004). To use the exposure prediction within structure prediction algorithms where gradient-based minimization is applied, however, the exposure potential must be continuously differentiable. To that end,

the neighbor count definition was expanded by applying a weight between 0.0 and 1.0 according to a transition function between upper and lower threshold values optimized to correlate with standard SASA calculated by the program MSMS (Durham, Dorr, Woetzel, Staritzbichler, & Meiler, 2009, Sanner, Olson, & Spehner, 1996).

If evaluating residue neighbor count based on a general amino acid class-based environment potential during protein structure prediction improves accuracy, predicting specific neighbor count values for a protein sequence should lead to even further improvements. Sequence-based predictions of environmental exposure are able to account for sequence motifs that are likely to be buried or exposed rather than preferences of amino acid identity alone. The purpose of this study is to develop and implement a sequence-based neighbor count prediction method for membrane protein structure prediction. The method is trained using artificial neural networks on a database of membrane proteins from Protein Data Bank and integrated into the *de novo* protein fold method, BCL::MP-Fold. We find that incorporating the predicted neighbor count into *de novo* protein structure prediction improves RMSD and ranks top models with higher contact recovery compared to the original prediction method in a set of 18 topologically diverse membrane proteins.

Materials and Methods

Membrane protein training set

A set of 177 membrane proteins from the Protein Data Bank (PDB) having less than 30% sequence identity and 3 Å or lower resolution was used to train the method presented in this work. This training set is the same as the set used to train the BCL::JUFO9D secondary structure and transmembrane span prediction method (Leman,

Mueller, Karakas, Woetzel, & Meiler, 2013). In brief, all protein chains with determined structure of 0 to 3 Å resolution, R-factor of 0.25 and sequence length of 40 to 10,000 residues were retrieved from the PDB (November 2011) (Berman, Henrick, Nakamura, & Markley, 2007). Non-X-ray structures and C-alpha only chains were excluded. PISCES was used to filter out structures with similar sequences with at cutoff of 30% sequence identity (G. Wang & Dunbrack, 2003, 2005). The resulting training set consists of 226 chains from 177 membrane protein structures.

The environment potential for each residue in the training set was evaluated using the BCL by calculating its number of neighbors within a defined radius. Neighbor count is inversely proportional to the exposure of an amino acid. A sequence separation of at least three residues is required before neighbor count is calculated to reduce the bias introduced by sequence proximity. This is a necessary step to prevent artificially low neighbor count values at the ends of SSE where no loop regions are considered. The function used to calculate neighbor count is as follows (Durham et al., 2009).

$$NeighborWeight(r, r_{low}, r_{high}) = \begin{cases} 1, if\ r\ \leq r_{low} \\ \frac{1}{2}\left[\cos\left(\frac{r - r_{low}}{r_{high} - r_{low}} * \pi\right) + 1\right] \\ 0, if\ r\ \leq r_{high} \end{cases}$$

$$NeighborCount(aa_i) = \sum_{|i-j|>3} NeighborWeight(r_{ij}, r_{low}, r_{high})$$

where r is the distance between amino acids i and j being considered as neighbors, r_{low} is set to 4.0 Å and r_{high} is set to 11.4 Å. These values were chosen because the resulting neighbor count has a high correlation with MSMS (Sanner et al., 1996). Residues with a distance above 11.4 Å receives a weight of 0, indicating that it is not a neighbor, and residues with a distance below 4.0 Å receive a weight of 1, indicating that it is a complete neighbor. Distances are calculated between beta-carbon atoms, where a pseudo beta-

carbon is introduced for glycine. The neighbor count is weighted between these threshold values to smoothen the potential, making it continuously differentiable to allow for gradient-based minimization. Distances in between the low and high values are given a weight calculated with a half cosine function.

Artificial neural network training with cross-validation

An artificial neural network (ANN) machine learning algorithm was trained using cross-validation on the protein database split into five subsets. Three of the five subsets were used for ANN training, the fourth was used for monitoring the training process to avoid overtraining and the fifth was used as an independent dataset to compute the accuracy of predictions. The independent and monitoring datasets were permuted through the five datasets, resulting in training of twenty separate networks. Inputs to the ANN included a set of sequence-specific descriptors generated from the sequence database and is described in detail below. The ANN was trained on one output, the residue-specific neighbor count which was calculated as described above.

The ANNs being trained were two-layer feed-forward networks with a sigmoidal activation function trained through resilient back-propagations of error (Riedmiller & Braun, 1993). The two hidden layers contained 400 and 50 neurons, respectively. This number was optimized by testing one hidden layer with 4, 8 and 128 neurons and two hidden layers with 128 and 4 neurons. Training consisted of 1000 steps with a weight update after each step with a minimum change of 0.001 and maximum change of 0.1. Prediction accuracies are reported as the difference between predicted and true neighbor count values, taken as an average over four subsets with the same independent subset. The average prediction over the twenty neural networks is reported as the neighbor count prediction for a particular amino acid.

Determining descriptors for predicting neighbor count using Garson's Connection-Weights algorithm

As input to the ANN training, sequence-specific descriptors were calculated over a window of residues across each sequence in the database. These sequence-based descriptors included 1) five amino acid chemical properties including steric parameter, volume, polarizability, isoelectric point and solvent-accessible surface area, b) six free energies for secondary structure type (helix, strand, coil), residue environment (membrane bilayer, interface and solution) and the nine combinations of both and c) evolutionary information via BLAST position-specific scoring matrices, including the BLAST profile for each amino acid type and the Shannon entropy of the BLAST profile information. A window of 4 residues was used in free energy calculations and a window of 16 residues was used in chemical property and BLAST profile calculations. Additionally, parameters over the entire protein sequence were included: a) the number of residues in the protein chain, 2) average and standard deviation of amino acid specific chemical properties, free energies and BLAST evolutionary information.

A total of 2028 descriptors were calculated. Previous work in our lab has highlighted the importance of descriptor selection in producing robust models. Our general framework is to score each descriptor based on some method, select the descriptors with the highest scores, and use them to train the next set of cross-validated models. Model-independent scoring functions such as FScore, information gain, and Pearson correlation, were initially tested, but found to offer little or no substantial improvement over using the entire descriptor set (data not shown). This is unsurprising given that the model-independent scoring methods are generally very low resolution and incapable of detecting whether certain descriptors contain essentially identical information as others or if they are just scaled differently. Additionally, descriptor interdependencies are not captured by these low-resolutions scoring terms. Following this logic, we elected to use a model-

dependent scoring term to assess which descriptors were most influential and held the most information relevant to calculating neighbor count. The scoring term we used is based off Garson's Connection-Weights algorithm (Olden & Jackson, 2002), that was previously implemented in the BCL. A total of thirty rounds of iterative ANN training took place. After each round of training, each descriptor was given a score based on its relative importance for predicting the output value. The bottom 10% of descriptors by score was removed before the next iteration of training.

Each round of training was evaluated by its prediction accuracy over the independent set, which measured the root means square deviation (RMSD) between the predicted and calculated neighbor count. While 30 rounds of training were performed, prediction accuracy only improved up to 24 rounds of training. As a result, the trained model at 24 rounds with the 259 most sensitive descriptors was taken as the most accurate model and incorporated into BCL::MP-Fold.

Incorporating the neighbor count score into BCL::MP-Fold

Progress during *de novo* structure prediction in BCL::MP-Fold is evaluated by the knowledge-based BCL::Score, which linearly combines the following potentials: amino acid pair distance, amino acid environment, secondary structure element packing, beta-strand pairing, loop length, radius of gyration, contact order and secondary structure prediction agreement (Woetzel et al., 2012). Neighbor count prediction was incorporated into BCL::MP-Fold as a score to assess model quality. The weight for the neighbor count score was optimized as a function of enrichment over contact recovery and was determined to be 947. Difference between residue neighbor count of a model and its predicted neighbor count value penalizes the overall score of that model during structure prediction. In this way, models with neighbor count scores within those predicted by the ANN are preferentially sampled by the Metropolis Monte-Carlo algorithm.

Structure prediction a benchmark set of MP with neighbor count score

A database of 22 diverse single-chain membrane proteins, from the same benchmark set used for BCL::MP-Fold, was used to evaluate the influence of the neighbor count score on *de novo* protein structure prediction (Weiner et al., 2013). This database included membrane proteins from the PDB with sequence identity less than 25%, resolution of less than 3.0 Å, R-value of less than 0.3 and a sequence length of at least 40 residues. Four proteins that were in the training set were removed from the benchmark set, leaving a total of 18 membrane protein chains. The dataset ranges from 91 to 565 residues with 2 to 15 transmembrane helices. Protein structures were folded 1000 times each with and without the neighbor count score to evaluate the influence of the score. As a proof of concept to illustrate the true effect of the neighbor count score, proteins from the benchmark dataset were also folded 1000 times with neighbor count values calculated from the protein structure. The quality of protein structure prediction was evaluated with RMSD100 (root mean square deviation calculated over the C-alpha residues, normalized to a protein length of 100 residues) and contact recovery.

A detailed protocol capture for setting up, training and testing the sequence-based neighbor count method as well as its incorporation into BCL::MP-Fold is provided in the Appendix.

Results and Discussion

Determining optimal input descriptors using Garson's Connection-Weights score

To determine which of the 2028 calculated input descriptors have the greatest contribution to the prediction of neighbor count, the connection weights between neurons

were used to rank descriptors after ANN training. 24 rounds of iterative ANN training where the bottom 10% of descriptors were removed after each round resulted in a list of 259 descriptors which trained the most accurate ANN for predictor neighbor count. The top 25 descriptors are listed in Table 17, with the full list of descriptors listed in the Appendix. The amino acid free energy for transition helix is a highly weighted input descriptor. This is likely because residues at the transition region of helices are more likely to be exposed because of their proximity to the soluble region in transmembrane proteins. Evolutionary information from the BLAST profile is consistently ranked highly as an input descriptor. The conservation of hydrophobic residues (A, L, F, Y and W) likely to interact with the membrane region were found to be correlated with neighbor count prediction. As the presence of P and G are highly unfavorable in helical secondary structure and as such, less, likely to be buried, the ANN found that their presence was correlated with neighbor count. The frequency of occurrence of L, G, F, and S were also found to be associated with its likelihood to be buried in previous studies (Park, Hayat, & Helms, 2007). Amino acid conservation, calculated from the BLAST profile for each sequence, scores highly among input descriptors, which is consistent with studies reporting that residues with the highest conservation prefer interaction with the membrane (Park et al., 2007, Russ, Engelman, & others, 2000, Senes, Gerstein, Engelman, & others, 2000, Chengqi Wang, Xi, Li, Liu, & Yao, 2012). The position of the amino acid in the sequence also plays an important role, likely because residues at the end of a protein chain are more likely to be exposed than residues in the middle of a sequence.

Table 17
Summary of the top 25 ANN input descriptors scored by connection weight.

Descriptor	Calculated range	Score
Maximum AA Free Energy: Transition Helix	Full sequence	1
AA Blast Profile, Leucine	Window of 4 residues per AA	0.83
AA Blast Profile, Lysine	Window of 4 residues per AA	0.80
AA Blast Profile, Phenylalanine	Window of 4 residues per AA	0.80
Mean AA Blast Conservation	Full sequence	0.71
AA Blast Profile, Glycine	Window of 4 residues per AA	0.71
Standard Deviation of AA Blast Profile, Tryptophan	Full sequence	0.67
Number of residues	Full sequence	0.62
AA Blast Probability, Alanine	Window of 4 residues per AA	0.61
Minimum AA Free Energy: Transition	Full sequence	0.60
Standard Deviation of AA Type, Histidine	Full sequence	0.57
Standard Deviation of AA Free Energy: Transition Helix	Full sequence	0.5
AA residue position (Sequence ID)	Full sequence	0.57
Minimum AA Free Energy: Transition Helix	Full sequence	0.56
AA Blast Profile, Aspartic Acid	Window of 4 residues per AA	0.56
Standard Deviation, AA Free Energy: Transition	Full sequence	0.55
Mean AA Blast Profile, Glycine	Full sequence	0.55
Number of residues left in sequence	Full sequence	0.55
Standard Deviation, AA Transfer Free Energy Punta Maritan 3D for Hydrophobicity	Full sequence	0.54
Mean AA Blast Profile, Phenylalanine	Full sequence	0.53
AA Blast Probability, Tryptophan	Window of 4 residues per AA	0.53
AA Blast Profile, Threonine	Window of 4 residues per AA	0.52
Standard Deviation of AA Blast Probability, Serine	Full sequence	0.52
AA Blast Profile, Tryptophan	Window of 4 residues per AA	0.51
AA Blast Profile, Proline	Window of 4 residues per AA	0.51

Correlation between predicted and true neighbor count is 65%

The accuracy of predicted neighbor count from the ANN was evaluated over the independent datasets during cross-validation and the ANN with the highest prediction accuracy was chosen for implementation into BCL::MP-Fold. The most accurate ANN predicted neighbor count with a Pearson correlation value of 0.65 with an average deviation from true neighbor count by 2.27 residues. The correlation between actual and predicted neighbor count is shown in Figure 39. Neighbor count is generally under-predicted, which may be due to the lack of representatives from the training set with

neighbor counts above 12 residues. Over a set of 18 membrane proteins in the benchmark set, the neighbor count score was shown to enrich for the selection of native-like models (under 8 Å RMSD) on average 1.2 fold.

In order to visualize the predicted neighbor count and uncover regions of lower accuracy, predictions were mapped onto the protein structures from the independent dataset. Predicted neighbor count is compared to true neighbor count for the proteins bovine rhodopsin (1U19) and the integral membrane enzyme DsbB (2K73) in Figure 40. Generally, the ANN correctly predicts residues at the ends of helices to be exposed, but is less accurate within the helix. While the prediction method was able to distinguish between the inward and outward faces of the helices to some extent, the range of predicted neighbor counts is substantially smaller towards the center of the helix. Many of the residues in the helices are predicted to have neighbor counts in the middle range of 3 to 6 residues.

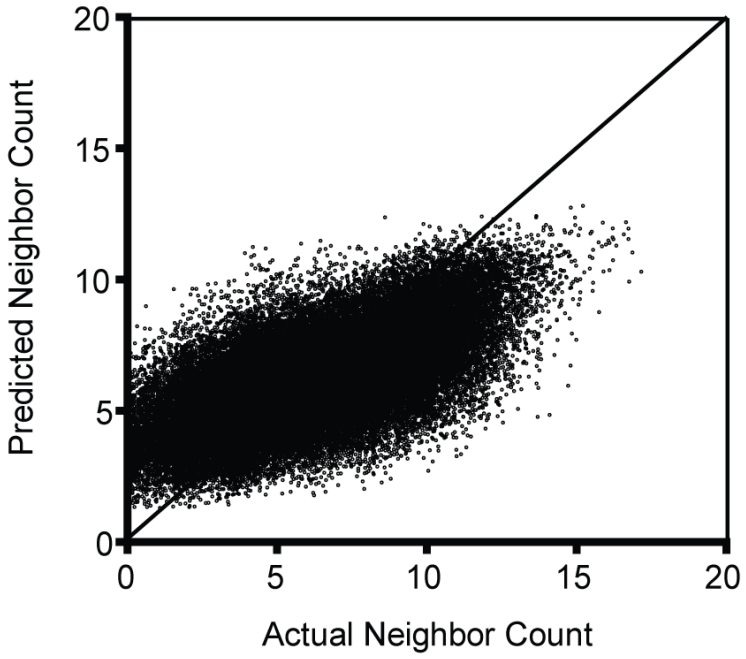


Figure 39
Correlation between true and predicted neighbor count over independent datasets.
Actual neighbor count was calculated from corresponding structures from the Protein Data Bank. Predicted neighbor count came from the trained ANN. Correlation is calculated as R^2=0.42.

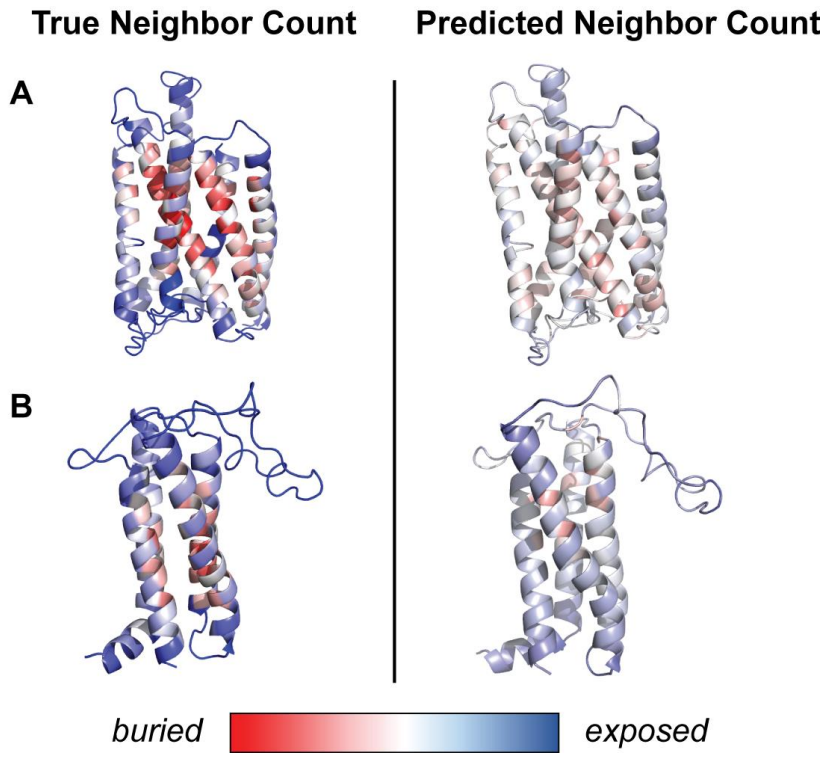


Figure 40
Mapping neighbor count prediction onto known structures of membrane proteins.
Proteins structures are colored by relative exposure and burial based on true neighbor counts were calculated for each residue in the protein structure and predicted neighbor count based on the trained ANN for A) rhodopsin (1U19) and B) DsbB (2K73)

Membrane protein structure prediction with neighbor count score samples accurate helical positions

The ANN-based neighbor count prediction method was incorporated into the BCL::MP-Fold algorithm, which has already been demonstrated to sample the correct topology for membrane proteins up to 500 residues and a range of secondary structure element content from 70% to 92% (Weiner et al., 2013). BCL::MP-Fold was used to *de novo* fold 22 membrane proteins with and without the neighbor count score. Four proteins were removed from the final analysis because they were included in the training set. Results for all 22 membrane proteins are found in the Appendix (Figures 72 & 73, Table 26).

With the addition of the neighbor count score, the BCL::MP-Fold algorithm was able to sample topologies with an average RMSD of 5.5 Å, compared to the 5.9 Å average RMSD using the original method (Table 18). The structure prediction algorithm with the added neighbor count score performed better than or equal to the original algorithm in 14 of 18 (78%) cases. With RMSD100 above 8 Å being considered native-like, incorporating the neighbor count score sampled native-like topologies for 16 of 18 (89%) cases, which is the same as the original algorithm. The structure prediction of a single subunit of aquaporin (1J4N) was a case in which an improvement in RMSD100 of almost 2 Å as well as a 6% improvement in contact recovery was seen. The neighbor count prediction algorithm was able to delineate between the inward and outward facing sides of the helices for 1J4N. This assisted in the positioning of the transmembrane helices such that residues with high neighbor count faced each other (Figure 41A).

Table 18
Model accuracy for structure prediction of a set of 18 benchmark proteins with and without inclusion of the neighbor count score.
Contact recovery and RMSD100 are reported for the best sampled and top ranked models.

		Contact Recovery		RMSD100	
		Best	Top Rank	Best	Top Rank
1IWG	No NC	26.9	1.1	4.6	8.8
	Predicted NC	22.9	6.9	4.3	8.7
	Calculated NC	37.7	12	3.6	7.6
1J4N	No NC	26.9	9.7	5.4	9.2
	Predicted NC	33.3	18.3	3.5	6.4
	Calculated NC	38.7	38.7	4.1	4.9
1KPL	No NC	14	2.3	8.9	13.1
	Predicted NC	12.4	1	9.4	13.4
	Calculated NC	14.3	3.6	7.9	12.9
1OCC	No NC	48.4	12.7	3.9	5.8
	Predicted NC	36.9	13.4	3.9	8.2
	Calculated NC	62.4	17.8	2.4	7.4
1OKC	No NC	13.2	0	7	10.3
	Predicted NC	12	6.6	7.2	11.4
	Calculated NC	11.6	5.8	6.2	9.7
1PV6	No NC	16	4.6	5.9	9.6
	Predicted NC	17.1	1.7	5.6	8.4
	Calculated NC	19.4	8	5.3	9
1PY6	No NC	24.5	4	4.4	5.7
	Predicted NC	26.1	4.4	3.4	9.8
	Calculated NC	34.9	13.7	3.3	8.2

Table 18, continued.
Contact recovery and RMSD100 are reported for the best sampled and top ranked models. RMSD100 is calculated over all C-alpha atoms in the SSEs.

		Contact Recovery		RMSD100	
		Best	Top Rank	Best	Top Rank
1PY7	No NC	31.5	12	3.8	4.9
	Predicted NC	34.3	4.6	3.3	10.4
	Calculated NC	51.9	39.8	2.6	3.3
1RHZ	No NC	22.9	8.5	6.4	9.6
	Predicted NC	22	2.5	6.1	11.2
	Calculated NC	29.7	2.5	5.1	11.2
2BG9	No NC	65	17.5	3	9.9
	Predicted NC	45	20	2.4	10.5
	Calculated NC	87.5	25	1.8	10
2BS2	No NC	31.8	7.9	5.6	6.9
	Predicted NC	25.8	13.9	4.8	11.4
	Calculated NC	41.7	15.9	4.5	9.4
2IC8	No NC	18.4	3.2	6.3	10.3
	Predicted NC	22.1	8.4	5.6	9.4
	Calculated NC	22.6	16.3	5.6	8.3
2NR9	No NC	22.7	3.4	5.8	9.9
	Predicted NC	29.5	4	5.3	10.7
	Calculated NC	24.4	3.4	6	8.6
2XQ2	No NC	5.1	4	8.1	10.9
	Predicted NC	5.9	0	7.6	12.6
	Calculated NC	5.9	0.8	8.3	13.3
2XUT	No NC	8.8	1.6	7.1	9.1
	Predicted NC	11	2.9	8.4	10.6
	Calculated NC	11	1.8	8.2	11.6
3KCU	No NC	14.6	3.1	7.3	10.4
	Predicted NC	14	9.9	7.1	9.8
	Calculated NC	15.8	10.9	7.7	8.8
3KJ6	No NC	22.3	1.6	5.7	9.3
	Predicted NC	19.1	8	5.2	9.8
	Calculated NC	21.3	5.3	5.9	6.3
3P5N	No NC	21.3	7.3	6.3	11.2
	Predicted NC	18.3	1.2	6.7	7.4
	Calculated NC	24.4	10.4	5.7	8.9
Average	No NC	24.1	5.8	5.9	9.2
	Predicted NC	22.7	7.1	5.5	10.0
	Calculated NC	30.8	12.9*	5.2	8.9

* $p < 0.1$ vs no neighbor count score, Student's T-test

Membrane protein structure prediction with neighbor count score shows little improvement in contact recovery

With the addition of the predicted neighbor count values to the structure prediction algorithm, improvements in contact recovery was expected as the exposure information was added to correctly orient the helices. Contrary to our expectations, however,

membrane structure prediction with the neighbor count score had very little impact on the sampling of models with contact recovery above 20%. Contact recovery sampling averaged 23% with the addition of the neighbor count score, compared to 24% average contact recovery for the original method. Compared to the original BCL::MP-Fold method, addition of the neighbor count score only improved the sampling of models with higher contact recovery for 8 of the 18 benchmark protein cases. With contact recovery above 20% being considered native-like, incorporating the neighbor count score resulting in sampling of native-like topologies for 10 of 18 cases (56%), as opposed to the 11 of 18 (61%) cases for the original algorithm. The greatest improvement in contact recovery was seen for GlpG (2NR9), seen in Figure 41B. The neighbor count prediction for this six-helical bundle was able to guide helical orientation for an increase in contact recovery of almost 7%. In this case, as well as 1J4N, the helices were packed close enough for direct contacts to be measured as neighbors, which directed folding to recovery contacts to about 33%.

Inclusion of the neighbor count score selects for models with higher contact recovery

While addition of the neighbor count score does not increase sampling of models with increased contact recovery, the score does allow for the selection of models with higher contact recovery. With the addition of the neighbor count score, highest ranked models by score had higher contact recovery in 12 of 18 cases (67%), with an average contact recovery of 7%. The top ranked model of 2BG9 had contact recovery of 20% and was found using the neighbor count score. With the inclusion of the neighbor count score, the range of model scores is increased by about 50% compared to the original energy function. However, the distribution of scores compared to RMSD100 is similar across the benchmark dataset of proteins between the two methods (Figure 74).

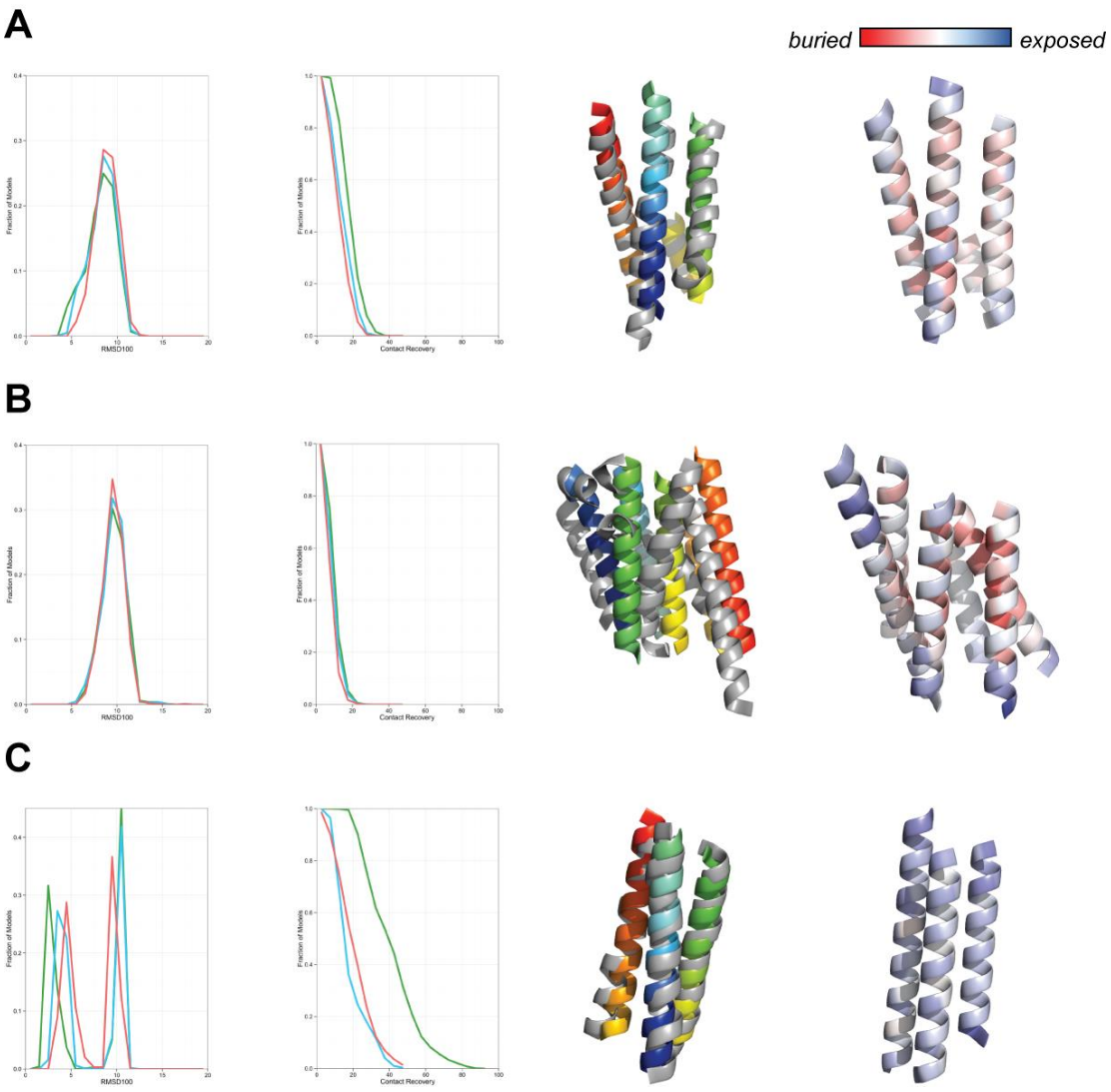


Figure 41
Gallery of select benchmark results.
The fraction of models at a range of RMSD100 values and contact recovery is shown for models predicted without neighbor count score (red), with predicted neighbor count (blue) and with calculated neighbor count (green). The top model ranked by RMSD100 is shown in rainbow compared to the native structure in grey. The top model ranked by contact recovery is colored by relative exposure and burial based on predicted neighbor count for A) 1J4N, B) 2NR9 and C) 2BG9. Complete results for all 22 proteins from the benchmark dataset can be found in the Appendix, Figure 72.

Structure prediction results for calculated neighbor count score show improvements in both RMSD100 and contact recovery

To demonstrate the maximal effect of the neighbor count score on membrane protein structure prediction, proteins from the benchmark dataset were folded with neighbor count values calculated from their known structures. An improvement in both RMSD100 and contact recovery was found across all proteins with less than 190 residues (Table 18). Inclusion of the neighbor count score also ranked models by contact recovery significantly higher than scoring without neighbor count. The range of scores for models with calculated neighbor count values are even higher than scores using predicted neighbor count and reflects in greater discrimination between native-like models in many cases, particularly 2BG9 (74). The greatest improvement in structure prediction was seen for 2BG9. Models were sampled with contact recovery up to 87% and 1.8 Å RMSD, the best sampled results seen for any protein from the benchmark dataset (Figure 41C). Little to no improvement was observed for proteins above 190 residues, which indicates an upper limit to the size of proteins that the neighbor count value can help towards sampling in BCL::MP-Fold.

Conclusion

An ANN-trained method for sequence-based neighbor count prediction has been developed and implemented in the *de novo* membrane protein structure prediction method, BCL::MP-Fold. The prediction method was trained on a database of membrane proteins where the output is the neighbor count calculated from the known structure. The prediction method has a Pearson correlation of 0.65 and an average difference of 2.27 residues from the calculated neighbor count value. Incorporating the neighbor count prediction as a score in evaluating membrane protein structure prediction retains sampling of native topologies while ranking top models with higher contact recovery compared to

the original scoring function. Sampling of models with higher accuracy based on RMSD100 was increased when using the neighbor count score, particularly for proteins with closely packed helices.

It has been shown that buried surface area drives protein folding and is proportional to the number of attractive interactions between residues (D. Baker, 2000). Additionally, a protein's three-dimensional structure is hypothesized to be determined by its amino acid sequence (Anfinsen, 1973). With this in mind, residue burial and exposure should be encoded in the protein sequence and useful for structure prediction methods (D. Baker, 2000, Durham et al., 2009, Karchin et al., 2004). While several methods have sought to predict sequence-based residue exposure through the prediction of its solvent accessibility surface area (Illergård, Callegari, & Elofsson, 2010, Joo, Lee, & Lee, 2012, Sim, Kim, & Lee, 2005, Yuan, Zhang, Davis, Bodén, & Teasdale, 2006), exposure can be more quickly and directly estimated through neighbor count (Durham et al., 2009, Karchin et al., 2004). Through the inclusion of calculated neighbor count values in structure prediction, we have provided proof of principle that per-residue neighbor count can result in sampling of protein structures with contact recovery up to 87% and 1.8 Å RMSD. Top ranked models also have significantly higher contact recovery when neighbor count is included in the scoring function.

While our current prediction method has yet to demonstrate such accurate results, there is room for improvement. Neighbor count values are currently being under-predicted and there are several ways that this could be alleviated. First, balancing datasets such that residues with neighbor count values above 12 are oversampled may improve results. Secondly, changing the program from a linear regression to a categorical problem, where residues are predicted to be either exposed or buried based on neighbor count, may improve prediction accuracy. Currently, the prediction method is trained to predict the exact neighbor count value and results close to the predicted value are being penalized.

This strict criteria may be causing the prediction values to be under-predicted. However, prediction of the neighbor count to an exact value is not necessary for the structure prediction algorithm, which only needs to know if the residue is exposed or buried. If the ANN was trained such that residues were predicted on this binary output, this may increase prediction accuracy and be a more useful score for structure prediction.

In summary, inclusion of per-residue neighbor count in structure prediction methods have been shown to improve sampling of native protein topologies. Continued improvements on sequence-based prediction methods will allow for this information to be applied to the *de novo* prediction of membrane proteins with unknown structure.

CHAPTER VIII

CONCLUSIONS

Summary of this work

In the quest for directed therapeutic approaches to effectively treat disorders of cognitive function, positive (PAM) and negative (NAM) allosteric modulators of metabotropic glutamate receptor subtype 5 ($mGlu_5$) have emerged as a promising new strategy. It has been a challenge, however, to develop compounds that are adequately soluble, potent and selective enough to be used as clinical candidates (Gregory et al., 2011). To guide drug development, structure-based drug design has demonstrated success in using the three-dimensional structure of a receptor to guide the optimization of small molecules for its binding site (Anderson, 2003, B Singh & M Tice, 2010, Becker et al., 2004, 2006, Congreve & Marshall, 2009, Congreve et al., 2005, Kandil et al., 2009, Lu et al., 2006, Shoichet & Kobilka, 2012, Tang et al., 2012). Due to the difficulties in crystallizing large membrane proteins, however, an experimentally determined structure does not yet exist for $mGlu_5$ or any other class C G protein-coupled receptors (GPCR). The basis of this dissertation was to generate and validate protocols for small-molecule docking into computational models of GPCRs and apply them to $mGlu_5$ drug discovery efforts. Key results from this work are summarized below.

Improved computational methods for membrane protein structure prediction

The primary protein structure software packages used in this dissertation are Rosetta and the BioChemical Library (BCL), both actively developed within the Meiler lab.

Comparative modeling and ligand docking with Rosetta has been shown to generate high-resolution structures for small soluble ligand-receptor complexes (Kaufmann & Meiler, 2012, Misura et al., 2006). The advantage of using Rosetta over other popular software packages such as Modeller (Eswar et al., 2008) is that Rosetta allows for both ligand and receptor flexibility while docking and allows for maximum user customizability. Because there are constant updates to the Rosetta software, however, tutorials for its use are difficult to generate and maintain. To make the process more user-friendly and applicable, an up-to-date detailed protocol for comparative modeling and ligand docking was compiled in Chapter II.

While validation of the Rosetta comparative modeling and ligand docking protocol has been completed for soluble proteins (Kaufmann & Meiler, 2012, Misura et al., 2006), a protocol for its use with membrane proteins had not been established. In Chapter III, a thorough and unbiased evaluation of the Rosetta comparative modeling and ligand docking protocol was completed on a set of 14 Class A GPCRs from the Protein Data Bank (PDB). The protocol presented in Chapter III is generalizable for use on all GPCRs and as an example, application of these methods on the Class A GPCR, $P2Y_{12}$, can be found in the Appendix. Structural models are generally considered to be comparable to the native crystal structure if they are within 3 Å root mean square deviation (RMSD). Rosetta was found to generate comparative models with an average of 2.2 Å RMSD to experimentally determined structures when templates within 50% sequence identity were used. Notably, Rosetta was able to build long extracellular loop regions (10 or more residues) with an average of 5.3 Å RMSD to experimental determined structures, which demonstrates higher accuracy than existing methods designed for such purposes (Goldfeld et al., 2012, Nikiforovich et al., 2010). While binding modes below 2 Å RMSD to experimentally determined ligand binding modes are considered to be accurate enough for drug discovery efforts and are sampled by Rosetta, they were only detected when

models were filtered by native-like binding site pocket residues, clustering and known experimental restraints. These findings are in agreement with recent studies of ligand docking into static comparative models of GPCRs (Beuming & Sherman, 2012, Katritch & Abagyan, 2011). In particular, the importance of experimental restraints in filtering for biologically relevant ligand binding modes has been demonstrated in several collaborative efforts between computational modeling, structural biology and pharmacology (Abrol, Kim, Bray, Trzaskowski, & Goddard III, 2013, Hirst et al., 2011, Katritch & Abagyan, 2011, Lange & Baker, 2012, Szymczyna et al., 2009). For this reason, application of the computation modeling methods to $mGlu_5$ system was performed in iteration with pharmacological studies of allosteric modulators in the receptor, which is detailed in Chapter IV and Chapter V.

Uncovering novel interactions between allosteric modulators and $_{mGlu5}$

In Chapter IV, the protocol developed in Chapter III was applied to the Class C GPCR, $mGlu_5$. Because the structural templates available are limited to Class A GPCRs, which have less than 20% sequence identity to the Class C GPCRs, the results from the computational modeling were analyzed in light of the functional data collected through site-directed mutagenesis studies performed by Dr. Karen Gregory. The comparative model of $mGlu_5$ for these studies was built using the backbone structure of the beta-2 adrenergic receptor (PDB ID: 2RH1) and based on a previously reported sequence alignment with the Ca^{2+} sensing receptor (CaSR) (Miedlich et al., 2004).

As one of the first selective allosteric modulators developed for $mGlu_5$, MPEP has been well characterized and several residues in the receptor have been found to influence function and binding of this compound. These residues include P654, Y658, L743, T780, W784, F787, Y791 and S809 (Malherbe, Kratochwil, Zenner, et al., 2003, Pagano et al., 2000). The computational model of $mGlu_5$ developed in Chapter IV was therefore first

evaluated by how it compared to existing knowledge about the MPEP binding site. The comparative model placed each of these residue such that the side chains were facing the binding pocket. In addition, the model predicted residue S808 to be facing towards the binding pocket and predicted S806 and T810 to be facing away. This prediction was validated by function and binding assays which demonstrated reduced MPEP potency and affinity to the S808A mutant of $mGlu_5$ (Gregory et al., 2013), which was also validated by a separate study (Molck et al., 2012).

When MPEP was docked into the comparative model of $mGlu_5$, several energy equivalent binding modes were revealed for the ligand, particularly when comparing docking studies centered at P654 and S808 (Gregory et al., 2013). Most frequently, however, Rosetta predicts the binding mode for MPEP to be in range of TM 5, 6 and 7 at the level of S808 and Y658, with the possibility of hydrogen bond interactions between both residues as well as π-π stacking interactions with Y658 or W784. While the binding modes presented here agree with the mutagenesis data presented in the literature, previously studies have demonstrated a range of energetically equivalent binding modes for MPEP as well. Two alternate binding modes for MPEP have been observed in docking studies by Molck et al., one where MPEP was located between TMs 3, 5 and 6 and another where MPEP was located between TMs 2, 3 and 6. (Molck et al., 2012). In addition, MPEP binding modes reported in earlier studies places MPEP in interaction distance with TMs 5 and 7 (Malherbe, Kratochwil, Zenner, et al., 2003) or TMs 3, 5 and 7 (Pagano et al., 2000) While these binding modes seem to be quite different, no placement of the ligand contradicts the residues determined to be consistently important, including Y658, S808, P654 and W784. As MPEP is a small and rigid ligand where few interactions are possible, it is likely that it can bind in several different conformations in the allosteric site. In evaluation of the literature and the studies done here, there seems to be a strong possibility of multiple binding modes for the ligand.

After being experimentally validated with MPEP, the molecular model of $mGlu_5$ was used to investigate interesting effects of compounds that allosterically modulated receptor function. In particular, Chapter IV discusses how the nicotinamide and picolinamide acetylene PAMs VU0405398, VU0405386 and VU0415051 were found to switch to NAMs or neutral modulators when mutations were made at S808 or T780 (Gregory et al., 2013). The molecular mechanism for the switch was investigated using ligand docking studies of these compounds in the $mGlu_5$ model. Polar interactions between the fluorine in VU0405398 and VU0405386 were hypothesized to form key interactions with S808 for receptor activation, and as such, the S808A mutant no longer allows for PAM activity. The same is thought to occur when VU0415051 is introduced to the T780A mutant, as the polar interaction between the carbonyl of the compound and T780 seen in the model is no longer formed.

Structure relationship studies on the MPEP scaffold inspired the development of several new selective allosteric modulators for $mGlu_5$. A large range compounds from diverse chemical scaffolds have now been found to produce allosteric modulation at a site competitive for MPEP (Alagille et al., 2005, Felts et al., 2010, Kinney et al., 2005, Kulkarni et al., 2009, Lindsley et al., 2004, Mueller et al., 2012, de Paulis et al., 2006). In Chapter V, 32 compounds from 4 diverse chemical scaffolds were docked into the comparative model of $mGlu_5$. In this study, we hoped to understand how structural and chemical determinants on both the ligand and receptor cause these compounds to range in their ability to potentiate or inhibit $mGlu_5$ glutamate response. Frequently sampled binding modes from the computational docking studies were found to agree with functionally important residues determined by site-directed mutagenesis studies performed with a representative probe ligand from each scaffold by Dr. Karen Gregory from the laboratory of Dr. Jeff Conn. Chemical determinants on the ligands identified to be important through comparison of binding modes agreed with SAR data collected on each scaffold by Chrysa

Malosh from the laboratory of Dr. Craig Lindsley. Together, synthesis of this data predicted a common binding mode that places the long axes of the ligands parallel to the transmembrane helices at the level of P654, Y658 and W784. Important hydrophobic interactions were predicted between Y658 and W784 with the phenyl groups of ligands from the MPEP series. Also, important polar interactions are predicted between the fluorine on the benzamide group of ligands in the VU0366248 series and S808 as well as between the cyano group on ligands from the VU0366058 series with Y658. These results are in agreement with findings on other MPEP-based compounds, particularly the role of W784, Y658 and S808 on the activity of 2-, 3- and 4-BisPEB (Molck et al., 2012).

Construction of the first $mGlu_5$ construct with a cysteine-less background

Without a structure determined by X-ray crystallography or nuclear magnetic resonances studies (NMR), experimental data on the $mGlu_5$ system is currently limited to functional data collected via intracellular Ca^{2+} mobilization studies with known modulators and radioligand binding studies with the probes [^{3}H]-M-MPEP (Gasparini et al., 2002), [^{3}H]methoxy-PEPy (Cosford, Roppe, et al., 2003) and [^{3}H]-methoxymethyl-MTEP (Cosford, Roppe, et al., 2003). Residues that modulate function or binding are inferred to direcly interact with the ligand, but this is not always the case. Studies have demonstrated instances where a point mutation distal to a ligand binding site can result in a change in ligand binding affinity (Baugh et al., 2010, El Omari et al., 2006, Rod et al., 2003). Structural studies through substituted cysteine accessibility method (SCAM) and disulfide cross-linking studies as well as information about the ligand binding pocket using thiol-reactive ligands have been shown to provide useful restraints for understanding the binding pocket of a receptor and further validating a comparative model (Field, Henry, & Blakely, 2010, J. H. Li et al., 2008, Liapakis et al., 2001, Loo & Clarke, 1997, S. D. Ward et al., 2006). To date, there has only been one published study generating a cysteine-less

mutant of a class C GPCR in which three cysteine mutations were made to $mGlu_6$. In this study, C793 on TM6 was found to be important for agonist response (Judith, 2012). This $mGlu_6$ cys-less system, however, has yet to be applied to any structural studies. In Chapter VI, a functional $mGlu_5$ construct with a cysteine-less background was developed for use in the abovementioned structural studies by mutating 12 cysteine residues to either alanine, serine or valine. C781 on TM 6 of $mGlu_5$, correlating to C793 in $mGlu_6$, was found to be play a crucial role in allosteric modulation of the receptor. Application of the $mGlu_5$ cysteine-less system in future studies will provide the first structural restraints on Class C GPCRs of its kind.

Improvements to de novo membrane protein folding through sequence-based exposure prediction

Comparative modeling has demonstrated its strength in modeling GPCRs when templates of high sequence identity are applied, as seen in Chapter III. Even when templates are low in sequence identity, comparative models of Class C GPCRs have been validated by experimental data and have generated useful hypotheses for drug discovery efforts, as discussed in Chapters IV and V. Until structural restraints are obtained with the cysteine-less $mGlu_5$ receptor as discussed in the Chapter VI, however, assumptions that the Class C GPCR topology matches that of the Class A structural templates cannot be verified. A different approach to structure prediction through *de novo* folding of proteins allows for the generation of unique membrane protein topologies. BCL::MP-Fold has been demonstrated to sample the correct fold in 34 of 40 cases where membrane proteins ranged from 91 to 565 residues. Chapter VII discusses improvements made to BCL::MP-Fold by adding residue exposure information via sequence-based neighbor count prediction. Residue exposure has been shown to drive the formation of protein tertiary structure and therefore should be encoded within the amino acid sequence itself (Eyal,

Najmanovich, Mcconkey, Edelman, & Sobolev, 2004, Karchin et al., 2004). Neighbor count can reasonably approximate the burial or exposure of a residue (Durham et al., 2009). An artificial neural network was trained to predict per-residue neighbor count from the protein sequence alone. Our studies show that adding per-residue neighbor count values as a means to evaluate progress during protein structure prediction has the ability to sample models with an average of 5.2 Å RMSD and 31 % contact recovery, as opposed to 5.9 Å RMSD and 24% contact recovery without the neighbor count score. The currently implemented prediction method has a Pearson's correlation of 0.65 to true neighbor count values. While this prediction method has yet to be fully optimized, its application with the structure prediction algorithm resulted in a sampling improvement in 78% of cases based on RMSD100 and ranked models with a higher contact recovery in 67% of cases. Once optimized, the improved *de novo* structure prediction methods using BCL::MP-Fold holds promise for determining unique topologies for protein targets with low sequence identity to the Class A GPCRs.

Implication of the results

During the course of this work, there were two overarching goals: to advance the development of computational protein structure prediction methods and to apply these methods to the development of novel treatments for schizophrenia. The first goal of method development sought to impact scientific development at a basic level. Advancing the technology used to construct models of protein structure and protein-ligand interactions has far-reaching application to all biological systems and disease processes. The second goal of applying the methods to the development of drugs for schizophrenia seeks to develop better medical treatments for a specific patient population. Novel interactions determined between $mGlu_5$ and its allosteric modulators will inform the development of drugs, where the goal is to generate a compound that can enter clinical

trials for the treatment of schizophrenia as well as other disorders of cognitive function. Implications of the results from this dissertation towards the achievement of these two goals are discussed below. Additionally, these research projects were not performed in isolation. Rather, they required the expertise and insight of scientists from various fields of research. The impact of this work on increasing the collaborative spirit between structural biology, computation, pharmacology and chemistry is also discussed below.

Impact of high-resolution GPCR modeling on all biological systems

With over 800 human GPCRs carrying out a range of essential biological functions, there is little wonder that GPCRs are such an important target for therapeutics. With only 17 GCPR structures crystallized, however, there is clearly a gap in our current knowledge about these important receptors. Computational modeling has long been used to model GPCRs where no crystal structure has been determined. As models have been compared to newly deposited crystal structures, the strength and limitations of these computational methods has been uncovered. Analysis of computational models have been documented in context of the GPCR Dock assessment, carried out the GPCR Network at Scripps in 2008 (Michino et al., 2009) and 2010 (Kufareva et al., 2011). In these assessments, the opportunity to generate models of a GPCR-ligand complex was offered to the community of scientists before a new GPCR structure was released to the Protein Data Bank. These models were then assessed for their accuracy in comparison to the crystal structure. The most common software packages used to perform the comparative modeling included Modeller (Eswar et al., 2008), MOE (Molecular Operating Environment, Chemical Computing Group, Ontario, Canada), I-TASSER (H. Zhou, Pandit, & Skolnick, 2009), Prime (Beuming & Sherman, 2012) and Rosetta (Yarov-Yarovoy et al., 2005). Ligand docking is often performed using Glide (Friesner et al., 2004), GOLD (Verdonk et al., 2003), Dock (Ewing et al., 2001) and FlexX (Rarey et al., 1996). Although the same

software is used in many instances, expertise in the parameterizing the software and knowledge about the biological system set apart the most accurate models. Many limitations to the computational models were found. Inaccuracies in the ligand binding conformation were due to incorrect placement of the transmembrane helices, variation in helical kinks, inaccuracies in modeling the loop regions and the exclusion of structurally important water molecules (Kufareva et al., 2011, Michino et al., 2009). The most accurate models took full advantage of the experimental data collected on the receptors and used this knowledge to restrain their models (Michino et al., 2009).

The GPCR comparative modeling protocol using Rosetta demonstrated in this work focused on generating high-resolution models of the receptor-ligand complex by addressing the issues of helical placement, kinks and loop reconstruction. Ongoing research in the lab outside the scope of this work focuses on the inclusion of water molecules in the docking algorithm (Lemmon & Meiler, 2013). The comparative modeling protocol was rigorously benchmarked over 14 Class A GPCRs in the Protein Data Bank. The protocol was able to correct inappropriate helical kinks and achieve high accuracy within the transmembrane bundles as well as the loop regions. Overall, the accuracy of the receptor models were comparable to the best models generated by Modeller (Kufareva et al., 2011, Michino et al., 2009). Additionally, Rosetta is widely available and free for academic users. The studies performed as part of this work further strengthens the credibility of the Rosetta method and will increase its applicability among scientists as a tool for modeling any GPCR.

There are over 780 GPCRs that still do not have an experimentally determined structure. The general method for modeling GPCRs established in this work can be applied to all GPCRs with missing structural information, which would provide a high-resolution model for any ligand-receptor complex. These comparative models have a vast range of applicability. Experimental biologists can use these models to generate

hypotheses regarding the three-dimensional location of residues for mutational studies. These models are also often used in virtual high-throughput screens of potential therapeutics. The details of these experiments will be discussed later as future directions for these studies.

Advancing progress towards new drug treatments for schizophrenia

Therapeutics for schizophrenia are currently limited to those assuming the dopamine or serotonin hypothesis, where dopaminergic or serotonergic hyperactivity is believed to drive the progression of the disease (van Bennekom, Gijsman, & Zitman, 2013). These drugs include citalopram, chlorpromazine, haloperidol and apripiprazole. Unfortunately, these drugs are not specific for target receptors causing the symptoms of schizophrenia and in addition to not fully treating the negative, positive and cognitive symptoms of the disease, these drugs have a host of side effects. These include severe muscle stiffness, seizures, slurred speech, hyperglycemia, anxiety and even permanent tardive dyskinesia (van Bennekom et al., 2013, Tschoner et al., 2007).

The glutamate hypothesis opens doors to new possibilities for treating schizophrenia. Clinical findings suggest that the increase in glutamatergic transmission in response to NMDA receptor antagonism results in symptoms closely related to those found in schizophrenia (Conn et al., 2009). Since $mGlu_5$ receptors feature a reciprocal positive feedback system with NMDA receptors, activation of $mGlu_5$ is considered a viable therapeutic option to treat such symptoms (Conn et al., 2009). As demonstrated by this work, several allosteric ligands have been shown to target $mGlu_5$ in a selective manner. However, these compounds still demonstrate off-target effects and require improvements in solubility and efficacy. A deeper knowledge of the allosteric binding pocket within $mGlu_5$ is needed to design the most effective drugs for this disease.

Before this work began, there was limited knowledge about the allosteric binding site of $mGlu_5$. Most of what was known about the $mGlu_5$ allosteric site was due to studies performed with MPEP and fenobam. The residues known from the literature to impact the functionality of these ligands are listed in Table 2. From the studies done as part of this research work, eight novel residues lining the allosteric binding pocket on TM 5, 6 and 7 were found to impact the activity of MPEP (Gregory et al., 2013). Furthermore, as new compounds have been discovered to be selective for $mGlu_5$ by the Vanderbilt Center for Neuroscience Drug Discovery (VCNDD), the allosteric binding site continues to be refined and characterized. The current structural model of $mGlu_5$ is consistent with other structural studies of mGlus, including the importance of a key hydrogen bonding network (Nygaard et al., 2010, Yanagawa et al., 2009, Yanagawa, Yamashita, & Shichida, 2013) and the impact of the serine residue at position 808 (Molck et al., 2012). Additionally, the model presented in this work is rich with mutational pharmacological data from over 40 docked ligands that sets it apart from other studies of this receptor.

By mapping the ligand-receptor interaction using pharmacological and computational methods, we find that the allosteric binding pocket is larger than we thought as well as more flexible. The pocket is able to accommodate ligands with a range of diverse structural scaffolds. It should then be possible for ligands, especially small and rigid compounds like MPEP, to bind in many different positions. This would explain why Rosetta, while being able to place ligands parallel to the helices, has difficulty differentiating between the two 180 degree orientations of the ligand. While the comparative modeling of $mGlu_5$ presented in this work brings us closer to a structural understanding of the allosteric binding pocket, it also reveals how much more we don't know. The vast range of ligand scaffolds and their ability to bind in multiple orientations within $mGlu_5$ certainly makes the problem of understanding the receptor's allosteric

modulation more complex. However, this also presents even greater options in terms of drug design that may soon lead to a candidate for clinical trials.

Synergy of computation, structure, pharmacology and chemistry drives drug discovery progress

The progression of scientific research requires the collaboration of minds from a wide range of fields of study. Each field brings a fresh perspective to a problem that would otherwise be missed. In this research, expertise in computational structure biology was integrated with expertise in pharmacology of allosteric modulation of GPCRs and chemical synthesis. The work presented here was a result of close collaboration at Vanderbilt University between the Center for Structural Biology and various branches of the VCNDD. Primarily, expertise of allosteric modulation of GPCRs and pharmacological studies came from the lab of Dr. Jeffrey Conn through the work of Dr. Karen Gregory. Synthesis of chemical compounds and insight into their structure activity relationship (SAR) came from the lab of Dr. Craig Lindsley through Dr. Shaun Stauffer and Chrysa Malosh in particular.

From the viewpoint of the structural biologist, structure informs function. The conformation held by an allosteric modulator in $mGlu_5$ and the chemical interactions formed as a result can help explain why a modulator might acts as a PAM, NAM or neither. In the context of the studies in Chapter IV, a structural understanding of molecular switches from PAM to NAM or neutral functionality was presented in light of the models produced by docking studies. However, especially in the case of allosteric modulation, a direct correlation between structure and function is sometimes not found and the true explanation of pharmacologic changes is much more complex. This complexity is unraveled through the work of molecular pharmacology. Patterns are detected through site-directed mutagenesis paired with functional and binding assays to characterize the relationship between these allosteric modulators and $mGlu_5$ that are impossible to see in

static models, including concentration-dependencies in modulation, the detection of molecular switches and the differences in compound potency and efficacy, to name a few. These relationships are brought to light by Dr. Gregory's work.

The relationship between a modulator and $mGlu_5$ is also heavily influenced by the structure of the compound itself. For deeper insight into the structure activity relationship of these compounds, expertise in chemical synthesis is integrated into the research. While docking results may predict that a particular hydrogen bond donor on a ligand stabilizes it in its active conformation, SAR may show that removing that hydrogen bond donor does not change the affinity of the compound. At other times, SAR may support hypotheses presented by the modeling results, such as instances where bulk is prohibited due to the positioning of a ligand in a tight pocket as indicated by the modeling and supported by SAR. These are just a few examples where SAR help to support or reject hypotheses formed by structural models, which influenced the results in the Chapter V.

The work presented here is unique in that it advances in the field through the integration of methods from various different fields of study. This collaborative effort has certainly impacted the work performed at Vanderbilt and hopes to inspire similar efforts elsewhere in the scientific community. While this work is specific to the $mGlu_5$ system, it can serve as an example for future work in other drug discovery efforts to increase creativity and productivity. The combination of different studies breathes new life into work through the presentation of new perspectives and ideas.

Future directions

When this work began, only two comparative models of $mGlu_5$ had been published with only four crystal structure GPCR templates available at the time (Malherbe, Kratochwil, Zenner, et al., 2003, Pagano et al., 2000). With a burst of innovation giving rise to an additional 13 GPCR crystal structures over the past four years, this has been

the most exciting time to study GPCR structure and function. While the field is starting to understand Class A GPCR structure, however, knowledge about Class C GPCRs is still very limited. The work presented here begins to map out the particular residues important for allosteric modulation of $mGlu_5$. In order to have a deeper understanding of the complex relationship between these modulators and $mGlu_5$ from a structural standpoint without an experimentally determined structure, a number of computational methods need to be put into place and experiments need to be carried out with the cysteine-less $mGlu_5$ receptor developed in Chapter VI to further refine the structural models. These methods will be discussed further below. Beyond the scope of this work, but before new compounds go into clinical trials, *in vivo* characterization of these compounds must occur to understand how they are metabolized, absorbed, distributed and cleared by the body. Characterization of any physical and behavioral side effects must also take place.

Using multiple templates and rigid body sampling in comparative modeling

In the course of validating the comparative modeling method developed for Rosetta in Chapter III, it was found that the most accurate models were built using templates with over 50% sequence identity. This is because the resulting models primarily retain the backbone conformation of the template employed. Slight changes in backbone structure account for the resolution of helical kinks and side chain placement, but large helical shifts are not performed. Also, additional improvement has been seen when using multiple template structures in the generation of a single model (Mobarec et al., 2009). Improvements to the Rosetta sampling algorithm during protein folding can be improved to incorporate multiple templates as well as rigid body sampling of secondary structure, which could shift sampling of GPCR models more towards the native conformation.

During the course of this research, these two ideas have begun to be developed in Rosetta by other scientists. Dr. Yifan Song in the lab of Dr. David Baker has begun to

implement a method that allows for a comparative model to be constructed using segments from multiple different template structures. If applied to GPCRs, comparative models of each helix could then be generated using a different template according to sequence similarity. For example, the construction of a comparative model of CXCR4 is difficult because it has only 43% sequence identity to the best available template, the mu-opioid receptor. However, TM 1 in CXCR4 contains key sequence motifs seen also in the kappa-opioid receptor and TM3 in CXCR4 contains key sequence motifs also present in the dopamine D3 receptor. Together, these three receptors could serve as templates for a single model of CXCR4 that would demonstrate more backbone accuracy within individual helices.

The next step in this new method for comparative modeling would require the rigid body sampling of these helices into an energy minimized, biologically realistic conformation. This is not yet possible with the comparative modeling protocol, which only performs small helical movements (J. Thompson & Baker, 2011). It is also not yet possible with the *de novo* folding protocol in Rosetta, which requires proteins to be folding from a single, continuous chain and is only accurate for small, soluble proteins (Bradley, Misura, et al., 2005). A graduate student, Stephanie DeLuca, in the lab of Dr. Jens Meiler has begun implementing into Rosetta an algorithm that performs rigid body sampling of helices in membrane proteins. This exciting new method will allow for the *de novo* folding of large membrane proteins and the possibility of using multiple templates for individual secondary structure elements when modeling GPCRs, as discussed here. These methods will make it possible to model GPCRs even if there does not exist a structural template within 50% sequence identity, as is the case for $mGlu_5$. After the new technologies have been put into place, this improved method can be developed and rigorously benchmarked for validity as demonstrated by this work.

Docking ensembles of ligands into ensembles of protein models

When research began on allosteric modulators of $mGlu_5$, the binding site was primarily characterized using MPEP and its derivatives of a similar scaffold because it was the allosteric modulator with the highest measurable affinity at the time (Gasparini et al., 1999, Malherbe, Kratochwil, Zenner, et al., 2003, Pagano et al., 2000). With the work of the VCNDD in the past five years, the number of allosteric modulators with high affinity to $mGlu_5$ has grown to include the six diverse scaffolds presented in this work, as well as many other compounds that are being explored by other academic labs and pharmaceutical companies. The process by which these compounds were docked into the comparative models of $mGlu_5$ and its mutants was very time-intensive and required a great deal of manual oversight. Each of the 43 compounds were docked individually into a particular model and post-processing using the Property RMSD as discussed in Chapter V was used to detect patterns in binding conformations across ligands of the same scaffold. This manually intensive work not only takes time, but is also limited to searching for patterns in binding conformations already sampled by the docking experiments.

A new idea to improve the sampling of ligand conformations docked into an ensemble of receptor models was borne out of the limitations seen in this project. The Rosetta Ligand Ensemble docking algorithm would be designed to read in an ensemble of small molecules and dock it simultaneously into models of the protein receptor. In the case of a focused library, a superimposed ensemble of small molecules as a meta-ligand with a rigid scaffold would be defined. During the docking process, transformations would be applied to the ligand ensemble simultaneously until a binding pose that optimizes the chemical properties of all active ligands within the same binding site is found. By transforming the ligand ensemble, the low-resolution docking would occur very quickly. High-resolution gradient-based all-atom minimization then occurs to optimize ligand-protein interactions further and optimize the receptor conformation specifically for each

ligand. The advantage to such a method would not only be the improved accuracy and efficiency while docking several ligands into a model of a receptor. It would also allow for the SAR of the ligands to inform the docking in real-time, and not just used as a post-processing tool. This enhances ligand sampling and allows for only the biologically relevant conformations to be accepted during the course of the docking experiment.

This method is currently being put into place by Darwin Fu, a graduate student in the Meiler lab. With over 40 allosteric modulators available for $mGlu_5$ and over 30 mutants of the receptor generated in the studies presented in this work, it serves as a perfect test case for such a method. Once Rosetta Ligand Ensemble Dock is complete, it can be applied widely to any system in which multiple ligands bind in the same site of a receptor.

Structural insight on Class C GPCRs through restraints collected with the cys-less $mGlu_5$ construct

As this work presents the first $mGlu_5$ receptor construct with a cysteine-less background, it paves the way for a vast range of experiments which will be key for collecting structural information on class C GPCRs. One of these experiments include the direct detection of the allosteric modulator binding site through the use of a thiol-reactive ligand. In these experiments, a thiol-reactive ligand tagged with a methanethio-sulfonate (MTSL) spin label bound to a cysteine can indicate interactions between a ligand and residue on the receptor, providing direct information regarding the location of the allosteric binding site (Loo & Clarke, 1997). Dr. Gregory intends on carrying out this work with a collaborator from Duquesne University in Pittsburg, PA who is generating these thiol-labelled modulators.

Additional experiments that can be performed are EPR measurements for distance restraints and substituted cysteine accessibility measurements (SCAM) and disulfide cross-linking studies for investigations into the receptor's helical structure. SCAM has

been used to map residues lining the binding site of the dopamine D2 receptor using MTS reagents (Javitch et al., 2002) and has more recently been used to identified eight residues on prolactin-releasing peptide receptor (PrRPR) that were crucial for binding and function of PrRP (Rathmann et al., 2013). Disulfide cross-linking studies have revealed conformational changes that occur in rhodopsin and the M3 muscarinic receptor as a result of activation (Hubbell et al., 2003, S. D. Ward et al., 2002). Rhodopsin has been extensively studied using EPR distance restraints (Van Eps et al., 2011, Hubbell et al., 2003). These examples provide support for the usefulness of such experiments. There are not yet current plans to carry out SCAM, disulfide or EPR experiments because of the inability to express enough of the protein for such experiments. However, Yan Xia, a graduate student in the Meiler lab, is working on protocols for large-scale membrane protein expression that might make these experiments feasible soon.

Improving hit rate through virtual high throughput screening with structural models of $mGlu_5$

Once the $mGlu_5$ model is refined using structural restraints, it should be at a resolution suitable for use in virtual high throughput screens. The purpose of a high throughput screen (HTS) is to quickly evaluate large databases of compounds in order to detect hits that can be further optimized as selective compounds for the receptor. Currently, high throughput screens are performed at Vanderbilt on various compounds received from commercial, academic and government suppliers using assays measuring mobilization of calcium as a functional output. To supplement experimental high throughput screens, virtual screens (vHTS) are also performed. These vHTS use machine learning algorithms which learn to recognize the characteristics of active compounds and uses this information to screen for new hits. An artificial neural network (ANN) trained to recognize biological potency from quantitative structure-activity relationship (QSAR)

models of $mGlu_5$ identified 749 compounds as active and suitable for experimental testing during a screen of 708,416 commercially available compounds (Mueller et al., 2012). Of these 749 compounds, 88 (12%) were determined to be truly active compounds. One compound in particular, a 2-(2- benzoxazolylamino)-4-phenylpyrimidine (VU0366058), was a novel and potent chemotype that has since been derivatized based on SAR and explored for therapeutic use.

At a 12% hit rate, there is room for improvement in the virtual high throughput screening method. Many compounds are selected that may have the characteristics of an active compound, but are not able to bind within the allosteric site. Docking the compounds predicted to be active by the ANN to a structural model of $mGlu_5$ as a secondary step in the virtual screening process may eliminated those compounds show a low binding affinity for the receptor (Shoichet & Kobilka, 2012). Studies with the human beta-2 adrenergic receptor have shown that virtual screening of large chemical libraries using comparative models can have a higher hit rate than using crystal structures (Tang et al., 2012). Virtual screening is particularly effective with models of high accuracy within the binding pocket that can capture the ligand interaction with the receptor. Comparative models have effectively been used to improve the hit rate in vHTS; recent examples include the discovery of novel MCH-1R antagonists with a 14% hit rate (Heifetz et al., 2013) and the identification of new active compounds for the thyrotropin receptor (Hoyer et al., 2013). Comparative models have yet to be integrated into the vHTS pipeline for the detection of $mGlu_5$ allosteric modulators. With future improvements to the $mGlu_5$ receptor model that would refine the placement of residues within the binding pocket, it may serve as a useful tool to increase the hit rate in vHTS of novel allosteric compounds.

Concluding remarks

Computational methods for studying the structure of GPCRs are extremely important because there are still over 700 GPCR targets without experimental structures, many of which serve as important drug targets. One of these receptors is $mGlu_5$, which shows exciting potential as a possible target for the treatment of schizophrenia and other cognitive disorders. Compounds that allosterically modulate $mGlu_5$ activity inhibit glutamate release, thereby alleviating the symptoms of schizophrenia. With a greater understanding of how these compounds bind and activate $mGlu_5$, advances towards a new therapeutic strategy to treat schizophrenia can be achieved.

This dissertation discusses the implementation of comparative modeling and ligand docking methods for GPCRs and demonstrates their application to $mGlu_5$. The timing of such work is well placed, as our understanding of GPCR structure has grown exponentially with the number of crystal structures being deposited into the Protein Data Bank. The location of the work is also ideal, as it takes advantage of the most advanced technologies and world-renowned expertise in computational structure biology and allosteric modulator of glutamate receptors. The synergy between chemical, computational, structural, and molecular biology as well as pharmacology makes this work unique in terms of its impact and applicability. A generalized protocol for building native-ligand models of GPCR-ligand complexes provides avenues for predicting structural interactions that are previously unknown. Applied to the $mGlu_5$ system, this protocol has already increased our understanding of the complex interaction between the receptor and its allosteric modulators, bring us one step closer to new drug compounds for a debilitating disease.

Future efforts towards structure-based drug design includes the development of new technologies that will improve our ability to build accurate comparative models, quickly assess the binding conformation of existing compounds and allow us to perform

virtual high throughput screens to discover novel modulators of the receptor. With these new developments and the continued interface between experimental and computational studies, new information will continue to be revealed about previously unknown systems.

APPENDIX TO CHAPTER I

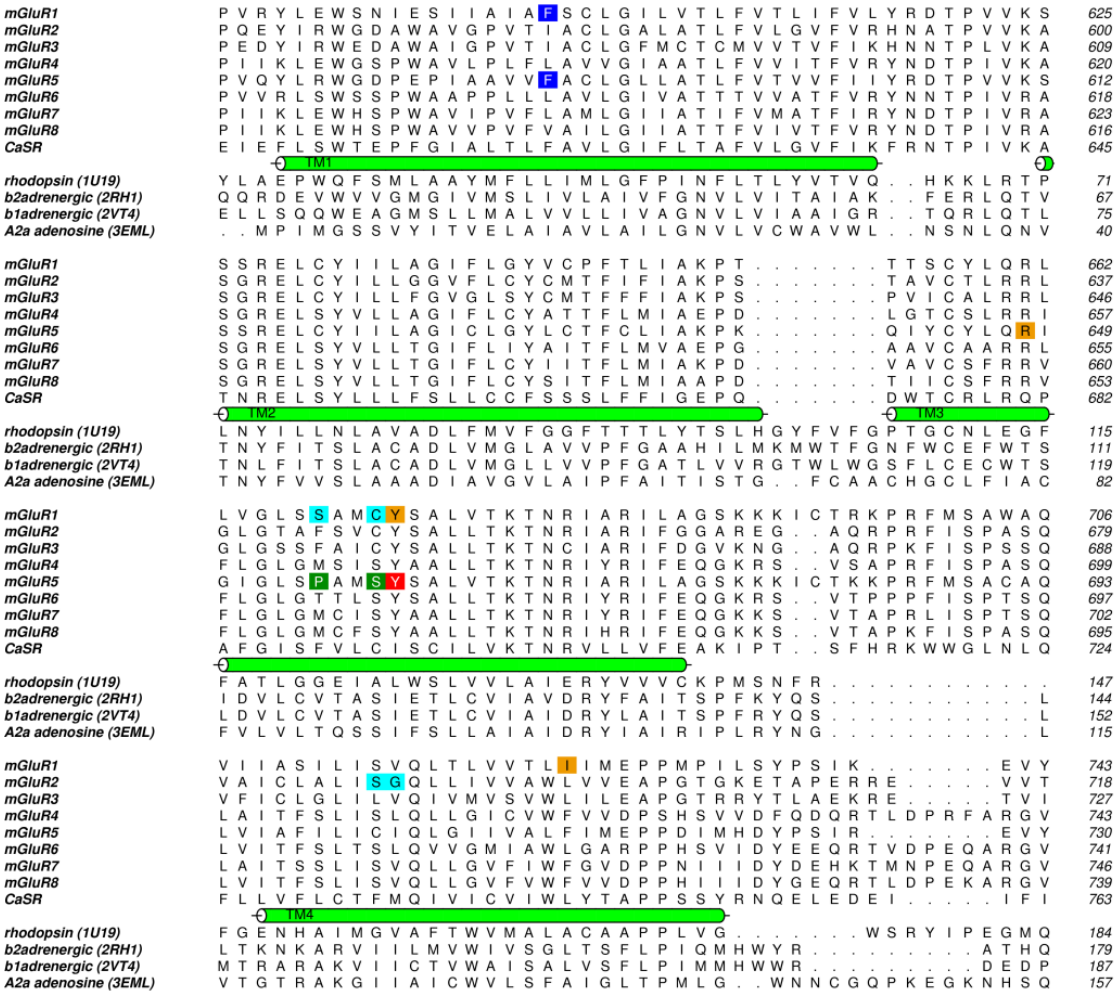
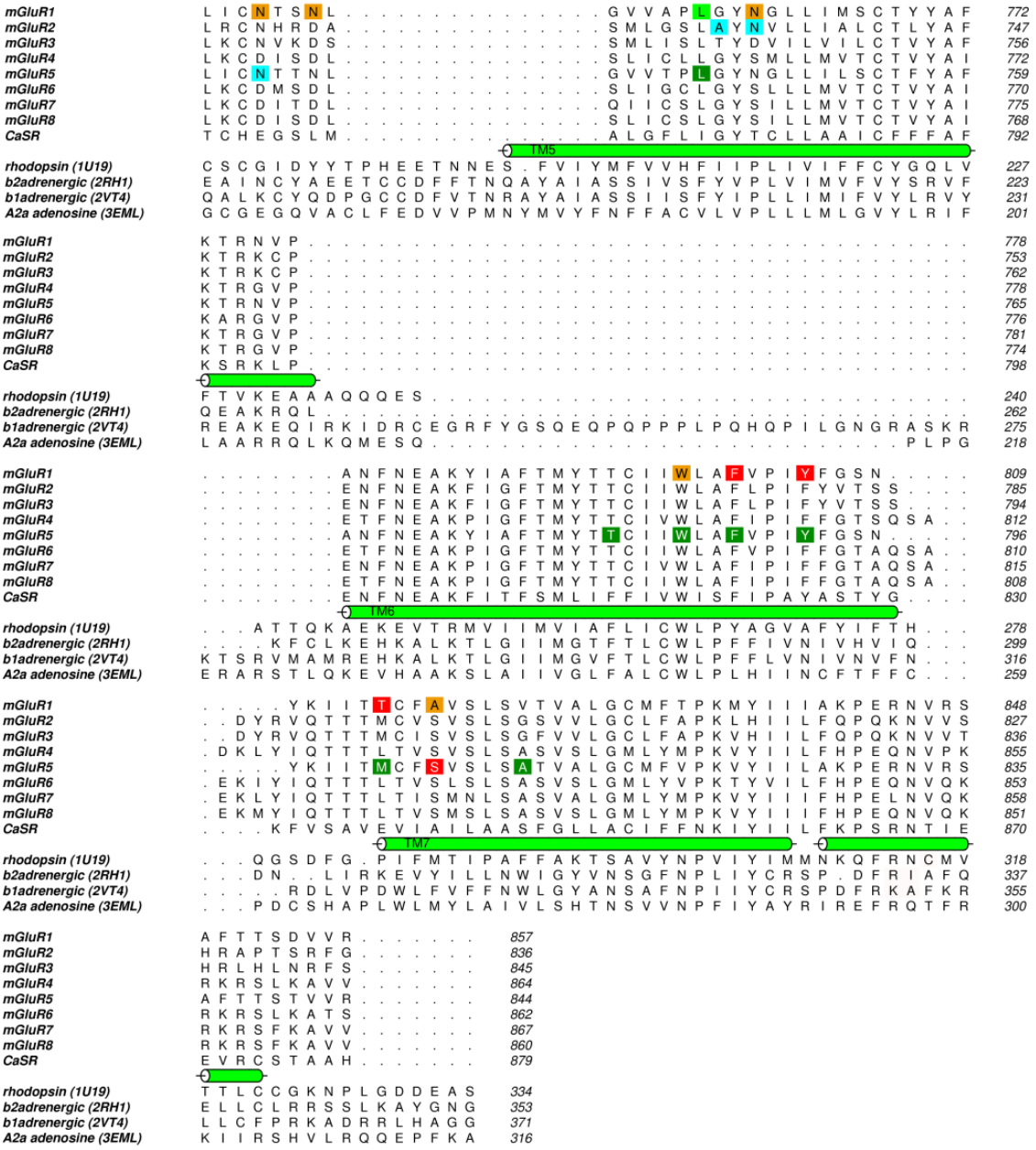


Figure 42
Alignment of the human mGlus and Calcium-sensing receptor (CaSR) sequences (aligned with CLUSTALW) relative to β2-adrenergic receptor (2RH1) and bovine rhodopsin (1U19) sequences (aligned with MUSTANG).
Alignment between class C GPCRs and bovine rhodopsin were directly adopted from Malherbe et al., 2006 (Malherbe et al., 2006), except TM2, 4 and 7, which were based on the alignment of CaSR with bovine rhodopsin from Miedlich et al., 2004 (Miedlich et al., 2004). Highlighted are: residues functionally important for CPPHA (blue) or other PAMs (cyan), residues important for NAMs functionally (orange) and through binding (red), and residues important for both PAMs and NAMs functionally (light green) or both PAM function and NAM binding (dark green). Continues on next page.



(Figure 42 continued from previous page) Alignment between class C GPCRs and bovine rhodopsin were directly adopted from Malherbe et al., 2006 (Malherbe et al., 2006), except TM2, 4 and 7, which were based on the alignment of CaSR with bovine rhodopsin from Miedlich et al., 2004 (Miedlich et al., 2004). Highlighted are: residues functionally important for CPPHA (blue) or other PAMs (cyan), residues important for NAMs functionally (orange) and through binding (red), and residues important for both PAMs and NAMs functionally (light green) or both PAM function and NAM binding (dark green).

APPENDIX TO CHAPTER II

Materials

- Primary sequence of target protein
- High-resolution protein structure of a protein homologous to the target sequence
- Linux or MacOS based workstation with internet access
- Plain text editor, such as vi, vim, or emacs.
- Academic or commercial copy of Rosetta obtained from www.rosettacommons.org/software/
- Access to the Robetta server (robetta.bakerlab.org) if a non-profit licensee or installation of the protein database from which fragments are selected, called the vall database, which can be obtained at www.bioshell.pl/rosetta-related/vall.apr24.2008.extended.gz. These will be used at Step 10.
- Desired small molecule for ligand docking
- Python, with BioPython and numpy installed (See “Installing Rosetta 3.4”)
- OPTIONAL**:** Linux or BlueGene/L based cluster

Overview of Procedure

The procedure outlined below is the complete protocol for both the construction of a comparative model of T4-lysozyme (PDB ID: 2ou0) (Mobley et al., 2007) based on the structure of P22-lysozyme (PDB ID: 2anv) (Mooers & Matthews, 2006) and of docking the ligand MR3 into the comparative model. For this example, the structure of T4-lysozyme is presumed to be unknown. The instructions in this Procedure assume that the user is working with a Linux operating system. The following procedure should be performed on a system running Linux or Mac OS X.

- Step 1: Template selection
- Steps 2-4: Prepare the PDB file of template structure
- Steps 5-7: Sequence alignment
- Steps 8-9: Threading
- Step 10: Prepare fragment files of the target sequence
- Step 11: Creating a Rosetta loops file
- Step 12: Preparation of comparative modeling options file
- Step 13: Running the comparative modeling job
- Step 14-16: Analysis of comparative modeling results and choosing receptor models for ligand docking
- Steps 17-19: Prepare the ligand file
- Step 20: Preparation of the ligand docking XML file
- Step 21: Preparation of the ligand docking options file
- Step 22: Running the ligand docking job

Procedure

Select a template

1. Select a template for comparative modeling of the target protein. Template selection and sequence alignment are the conceptual first steps of any comparative modeling procedure and is discussed in the Experimental design section of Chapter 2. It is often beneficial to explore multiple template/target sequence alignments and, in the case of lower sequence similarity, to perform manual curating of sequence alignments. This accounts for conserved motifs in the protein family that are not correctly aligned. In this procedure, the target protein to be modeled is T4-lysozyme (PDB ID: 2ou0) (Mobley et al., 2007) and the

template being applied is the structure of P22-lysozyme (PDB ID: 2anv) (Mooers & Matthews, 2006).

Prepare the PDB file of template structure

2. Download the template PDB file from the PDB at www.rcsb.org. The template PDB can be found by searching for the four-letter PDB ID, “2anv.”
3. Format, or “clean,” the protein to avoid errors when Rosetta reads in the PDB file. Cleaning the PDB file removes non-ATOM records, renumbers residues and atoms from 1, and corrects chain ID inconsistencies. The script `clean_pdb.py`, located in the `rosetta_tools/protein_tools/scripts/` directory, will be used to format the template PDB file (see “Installing Rosetta 3.4” for instructions on installing the necessary python modules).

The script follows the format:

```
python rosetta_tools/protein_tools/scripts/clean_pdb.py <raw_pdb_file> <chain>
```

Execute the script by typing:

```
python rosetta_tools/protein_tools/scripts/clean_pdb.py 2anv.pdb A
```

The script will output two files: `2anv_A.pdb` and `2anv_A.fasta`

4. Relocate the created FASTA and PDB files from the script to an `input_files` directory, which will be used in subsequent steps. The output to the screen is used for error checking and can be disregarded if no errors occurred.

Sequence alignment

5. Generate a FASTA file for the target sequence. A FASTA file is a text file that contains a header line, which consists of the name of the protein, followed by the amino acid sequence of the protein on a separate line; this is indicated below:

```
>2ou0:X|PDBID|CHAIN|SEQUENCE
MNIFEMLRIDEGLRLKIYKDTEGYYTIGIGHLLTKSPSLNAAKSELDKAIGRNTNGVITKDEAEKLFNQDVDAAVRGI
LRNAKLKPVYDSLDAVRRAAAINMVFQMGETGVAGFTNSLRMLQQKRWDEAAVNLAKSRWYNQTPNRAKRVITTFRTG
TWDAYK
```

The target FASTA file that is used comes from the T4-lysozyme sequence (PDB ID 2ou0). The FASTA file can be downloaded from the PDB (www.pdb.org) by searching for “2ou0” in the search bar at the top of the webpage. Download the FASTA file for the target by clicking on “Download” to the right of the PDB ID and selecting “FASTA.” Save the file as `2ou0_.fasta`. The header line for the `2ou0_.fasta` file must be edited. Replace the text

```
>2ou0:X|PDBID|CHAIN|SEQUENCE
```

with

```
>2ou0_
```

The FASTA file `2anv_A.fasta` that was created from step 3 can be used for the template sequence.

6. Run ClustalW on the webserver (www.ebi.ac.uk/Tools/msa/clustalw2/) by pasting the contents of the two FASTA files into the box labeled “STEP 1 - Enter your input sequences.” The order of the FASTA files is irrelevant. Both sequences should be in FASTA format, i.e., they must start with a header line such as `>target_sequence` or `>template_sequence`, followed by the sequence on a new line (see step 5). On “STEP 2,” select “slow” for the “alignment type.” This will provide the most accurate alignment for the two sequences. Do not change anything in the “STEP 3” box and hit the “Submit” button on “STEP 4.” After a short wait, a new page will be loaded where the alignment can be downloaded and saved. Click on the button labeled “Download Alignment File.” Several sequence alignment tools are publicly available; however, the web server, ClustalW (Larkin et al., 2007), is used as an example in this protocol due to its commonality and simplicity. If a different alignment tool is used, the output from the alignment must be in one of the following formats: Clustal, EMBOSS, FASTA, FASTA-M10, IG, Nexus, PHYLIP, or Stockholm.

7. Save the alignment file as `alignment.aln` in the current working directory. The default suffix provided by ClustalW is `.clustalw`.

Threading

8. Thread the target sequence over the template structure using the included script. The script has the format:

```
python rosetta_tools/protein_tools/scripts/thread_pdb_from_alignment.py --
template=<name of template in alignment file> --target=<name of target in
alignment file> --chain=<chain in pdb> --align_format=clustal <alignment file>
<template.pdb> <output.pdb>
```

The `--template` and `--target` must match the names given in the header file of the FASTA file. Check the target and template names by opening the alignment file that was created in step 6. If the naming has been consistent according to the previous steps, the command used to thread the template PDB should be:

```
python rosetta_tools/protein_tools/scripts/thread_pdb_from_alignment.py --
template=2anv_A --target=2ou0_ --chain=A –-align_format=clustal alignment.aln
2anv_A.pdb 2ou0_threaded.pdb
```

CAUTION**:** The result of step 8 (`2ou0_threaded.pdb`) is a PDB file that Rosetta will use as input. Examine this file with a text editor and also with a 3D protein structure viewer, such as PyMOL (www.PyMOL.org/) or Chimera (Pettersen et al., 2004) (www.cgl.ucsf.edu/chimera/), to ensure that 1) it contains the intended target sequence, 2) the conserved regions (especially helices and strands) between target and template are preserved, and 3) insertions (residues present in target but not template) should have zero (0.000) Cartesian coordinate values and −1.00 occupancy values.

9. Verify that the `2ou0_threaded.pdb` sequence matches the target primary sequence by generating a FASTA file out from the PDB using the included script. This script has the following syntax:

```
python rosetta_tools/protein_tools/scripts/get_fasta_from_pdb.py
<template_pdb> <chain> <output fasta file>
```

Submit the two sequences to the ClustalW server (www.ebi.ac.uk/Tools/msa/clustalw2) to verify that the sequences match.

Prepare fragment files of the target sequence

10. Generate fragment files of the target sequence. There are two commonly used methods to generate fragments for comparative modeling. Users affiliated with a non-profit institution can use the Robetta server (www.robetta.org/), which is described in Option A. Conversely, for-profit organizations should follow Option B to use the fragment picker application that comes with the Rosetta source code.

Creating fragment files with Robetta

a. Register for a username and password at the Robetta webserver (robetta.bakerlab.org/fragmentsubmit.jsp).
b. Input the sequence name 2ou0_, and load the target FASTA file `2ou0_.fasta` from step 5.
c. Submit the FASTA file. The webpage will reload and state: “Successfully added your request to the queue.” The status of the fragment file generation can be checked at http://robetta.bakerlab.org/fragmentqueue.jsp.
d. Click the link to get a list of files generated by Robetta after the status has changed to “Complete.” If following the example of 2ou0 for the target sequence, the fragment files should be called `aa2ou0_003_05.200_v1_3` for fragments of length 3 and `aa2ou0_009_05.200_v1_3` for fragments of length 9. Save all the files to the working directory by right-clicking and selecting “Save as.”

Creating fragment files with the fragment picker

a. Generate a secondary structure prediction file, such as from PSI-PRED (J. Ward, McGuffin, Buxton, & Jones, 2003) (bioinf.cs.ucl.ac.uk/psipred/) or JUFO9D (Leman et al., 2013) (www.meilerlab.org/index.php/servers/show). In either case, the primary sequence of the target protein must be in FASTA format.

CAUTION: The fragment picker expects PSI-PRED vertical format for all secondary structure prediction files. If PSI-PRED is used to generate secondary structure predictions, make sure to select the "machine learning scores" option when downloading the results. If JUFO9D is used, download the 3-state secondary structure prediction file (`.jufo9d_ss`), and make the following PSI-PRED header the first line of the JUFO9D prediction file, followed by a blank line:

```
# PSIPRED VFORMAT (PSIPRED V3.0)
```

b. Generate a sequence profile (checkpoint file). The checkpoint file, also known as a sequence profile, is created by PSI-BLAST. This file can be generated by running the Rosetta `make_fragments.pl` script as follows:

```
rosetta_tools/fragment_tools/make_fragments.pl -id 2ou0_ -nopsipred –
psipredfile <psi_pred_file> -nosam –nojufo 2ou0_.fasta
```

The `psi_pred_file` is generated from secondary structure prediction. The checkpoint file will be named `2ou0_.checkpoint`. Make sure that all paths in the `make_fragments.pl` script exist in the working environment. They will need to be altered after downloading a fresh copy of Rosetta.

c. Create a fragment picking weights file called `QuotaProtocol.wts`

```
# score name priority weight min_allowed extras
SecondarySimilarity       350    0.5    -      psipred
SecondarySimilarity       250    0.5    -      jufo
RamaScore                 150    1.0    -      psipred
RamaScore                 150    1.0    -      jufo
ProfileScoreL1            200    1.0    -
```

```
PhiPsiSquareWell          100    0.0    -
FragmentCrmsd30           0.0    -
```

d. Create a Quota definition file called `QuotaProtocol.def`

```
#pool_id     pool_name    fraction
1            psipred      0.6
2            JUFO         0.2
```

e. Create a fragment picking options file called `fragment.options` in a text editor. The file should have this format:

```
-database <path to Rosetta Database>
-in:file:vall <path to Vall Database> # available from Rosetta checkout
-in:file:fasta 2ou0_.fasta
-in:file:s 2ou0_threaded.pdb
-in:file:checkpoint 2ou0_.checkpoint
-frags:ss_pred 2ou0_.psipred.ss2 psipred 2ou0_.jufo9d_ss JUFO
-frags:scoring:config QuotaProtocol.wts
-frags:picking:quota_config_file QuotaProtocol.def
-frags:frag_sizes 9 3
-frags:n_candidates 1000
-frags:n_frags 200
-out:file:frag_prefix 2ou0_quota
-frags:describe_fragments 2ou0_quota.sc
```

f. Run the command line:

```
rosetta_source/bin/fragment_picker.default.OperatingSystemrelease
@fragment.options
```

When using the fragment picker, the fragment files will be output as `2ou0_quota.200.3mers` and `2ou0_quota.200.9mers.`

Creating a Rosetta loops file

11. Create a file which defines loop regions to be rebuilt. One line is created per loop to be built in a file with the extension, `.loops` (e.g, `2ou0_.loops`). The Rosetta loop building protocol will close gaps between residues specified in the loops file. The information in the loops file is explained further in Table 19.

```
LOOP 28 60 0 0 0
LOOP 81 93 0 0 0
LOOP 112 126 0 0 0
LOOP 135 151 0 0 0
```

Table 19
Explanation of information contained in the loops file

Column 1	LOOP	The loops file identity tag
Column 2	<integer>a	Loop start residue number. NOTE: The starting structure must have real coordinates for all residues outside the loop definition, plus the first and last residue of each loop region.
Column 3	<integer>	Loop end residue number
Column 4	<integer>	Cutpoint residue number, must be greater than the first residue of the loop and less than the end residue of the loop. Default (0) - let loop rebuild protocol choose cutpoint
Column 5	<float>	Skip rate. Default (0) - never skip
Column 6	<boolean>	Extend loop. Default (0) – false

Preparation of comparative modeling options file

12. Create an options file with the name `modeling.options` and add the lines below.

    Comments within the options file are ignored when the '#' tag is present.

```
#input file, should be the output pdb
-loops:input_pdb 2ou0_threaded.pdb
#input will be in all-atom mode
-loops:fa_input
#loop definitions file
-loops:loop_file 2ou0_.loops
#sizes of fragments
-loops:frag_sizes 9 3 1
#location of the fragment files. Fragments files will have the extension
.quota.200.3mers and .quota.200.9mers if created with the fragment picker.
-loops:frag_files aa2ou0_09_05.200_v1_3 aa2ou0_03_05.200_v1_3 none
#the centroid phase of loop modeling using CCD
-loops:remodel quick_ccd
#the all-atom phase of loop modeling
-loops:refine refine_kic
#force extended on loops (phi-psi angles set to 180 degrees in the first step)
#independent of loop input file. For rebuilding loops entirely.
-loops:extended true
#give idealized phi and psi angles after it has been closed
-loops:idealize_after_loop_close
#does a minimization of the structure in the torsion space
-loops:relax fastrelax
#decreases the monte carlo inner and outer cycles of loop rebuilding, greatly
#decreasing computation time
-loops:fast
#rotamer libraries used in the repack steps
-ex1
-ex2
```

Running the comparative modeling job

13. Generate comparative models of the target protein using Rosetta. At this point, the only Rosetta application needed is loop building. The following command can be used to run loop modeling in Rosetta. In the command line, `OperatingSystem` should be replaced with the operating system of the machine on which the job is running. For example, if the job is running on a Linux machine, the name of the executable will be `loopmodel.default.linuxgccrelease`. This command is executed on a single processor and produces 10,000 models.

```
rosetta_source/bin/loopmodel.default.OperatingSystemrelease @modeling.options
-database rosetta_database -nstruct 10000
```

It is advised to split the generation of models across multiple CPUs. A single Rosetta process can be started for each CPU in a multi-processor machine, or the work can be distributed by starting processes on independent machines. For example, one could start four different jobs on four different processors. Each job would have its own command line. Further, the `-out:prefix <prefix>` or `-out:suffix <suffix>` options can be specified to give each job its own unique name. Each job would only generate 2,500 unique structures, summing to 10,000 when all four jobs are complete. A benchmarking study of loop building in Rosetta with cyclic coordinate descent can be found in Wang *et al*., 2007 (Chu Wang et al., 2007).

Analysis of comparative modeling results and choosing receptor models for ligand docking

14. Choose the ten lowest energy comparative models for the ligand docking steps below. Rename the files to `model_01.pdb`, `model_02.pdb` … `model_10.pdb`. The process of choosing models to be used in ligand docking can vary depending on the user's specific biological problem (see "Experimental design" in Chapter 2). As seen in Figure 12, the lowest energy models are reasonably close to the native structure and, as

such, are a good starting point for ligand docking. These models are commonly chosen based on overall energy according to Rosetta's scoring function because they are usually low-energy models with fully closed loops and minimal inter-atomic clashes. Other modes of filtering, such as model satisfaction of experimental restraints (see "Using Constraints as Filters in RosettaScripts") or clustering (see "Clustering Using Rosetta"), can also be used. Furthermore, increased sampling of regions that do not converge on one or several conformations can improve the final model during *de novo* protein folding (Das et al., 2007, Qian et al., 2007, Raman et al., 2009).

15. Use a 3D protein structure viewer to check the receptor models to be used for ligand docking and ensure that they do not have chain breaks.
16. Align the ten comparative models using any 3D protein structure viewer, and save the resulting coordinates to new, individual PDB files before moving on to the ligand docking part of the protocol.

A benchmarking study of comparative modeling in Rosetta can be found in Misura, *et al*., 2006 (Misura et al., 2006).

Prepare the ligand file

17. Obtain a representation of the ligand to be docked of the type `mol`, `mol2`, or `sdf`. If a protein structure is determined in the presence of a ligand of interest, an `sdf` file can be downloaded from the PDB (www.pdb.org); however, hydrogen atoms are usually not present and must be added. To generate a `mol` file from a `pdb` (PDB) file, many different software packages can be used, including MOE (www.chemcomp.com/index.htm), PyMOL (www.pymol.org/), and ChemDraw (www.cambridgesoft.com/software/ChemDraw/). Generation of the `mol` file is not covered within this tutorial.

18. Run the following command to convert a `mol` file into a `params` file. Rosetta reads ligand files from a `params` file. The `params` file contains information about the atoms, bonds, charge, and coordinates of a ligand. The `params` file is generated from a molecule file and which can be of the type `mol`, `mol2`, or `sdf`.

```
python rosetta_source/src/python/apps/public/molfile_to_params.py <mol file>
```

In this specific example, 1-methyl-1H-pyrrole (MR3), is in complex with 2ou0 and is the ligand that will be docked into the comparative model of T4-lysozyme (2ou0). The command line used to create a `params` file for MR3 is as follows:

```
python rosetta_source/src/python/apps/public/molfile_to_params.py -n MR3
MR3.mol
```

where `-n MR3` is the three-letter name for the ligand. The resulting output will be `MR3.params` and `MR3_0001.pdb`.

19. Copy and paste the lines from `MR3_0001.pdb` to the bottom of each of the ten model PDB files from steps 14-16. These files will be used in subsequent steps.

Preparation of the ligand docking XML file

20. Create a ligand docking XML file. The scoring functions, filters, and movers (specific Rosetta functionalities for the protocol) are specified in the XML file. Below is an example of an XML file, named `ligand_dock.xml`, that will be used to dock the ligand, MR3, into the T4-lysozyme comparative models chosen in steps 14-16. Comments on specific steps are shown outside of the `<>`. It should be noted that comments are handled differently between the XML file and the options file. We recommend beginning with the provided XML file and altering key variables to suit the specific needs of the study.

```
<ROSETTASCRIPTS>
	<SCOREFXNS>
		<ligand_soft_rep weights=ligand_soft_rep>
			<Reweight scoretype=hack_elec weight=0.42/>
			<Reweight scoretype=hbond_bb_sc weight=1.3/>
			<Reweight scoretype=hbond_sc weight=1.3/>
			<Reweight scoretype=rama weight=0.2/>
		</ligand_soft_rep>
		<hard_rep weights=ligand>
			<Reweight scoretype=fa_intra_rep weight=0.004/>
			<Reweight scoretype=hack_elec weight=0.42/>
			<Reweight scoretype=hbond_bb_sc weight=1.3/>
			<Reweight scoretype=hbond_sc weight=1.3/>
			<Reweight scoretype=rama weight=0.2/>
		</hard_rep>
	</SCOREFXNS>
	<LIGAND_AREAS>
		<docking_sidechain_X chain=X cutoff=6.0 add_nbr_radius=true
all_atom_mode=true minimize_ligand=10/>
		<final_sidechain_X chain=X cutoff=6.0 add_nbr_radius=true
all_atom_mode=true/>
		<final_backbone_X chain=X cutoff=7.0 add_nbr_radius=false
all_atom_mode=true Calpha_restraints=0.3/>
</LIGAND_AREAS>
	<INTERFACE_BUILDERS>
		<side_chain_for_docking ligand_areas=docking_sidechain_X/>
		<side_chain_for_final ligand_areas=final_sidechain_X/>
		<backbone ligand_areas=final_backbone_X extension_window=3/>
	</INTERFACE_BUILDERS>
	<MOVEMAP_BUILDERS>
		<docking sc_interface=side_chain_for_docking
minimize_water=true/>
		<final sc_interface=side_chain_for_final bb_interface=backbone
minimize_water=true/>
	</MOVEMAP_BUILDERS>
	<MOVERS>
single movers
		<StartFrom name=start_from_X chain=X>
			<Coordinates x=-18.8922 y=24.5837 z=-5.7085/>
		</StartFrom>
		<CompoundTranslate name=compound_translate randomize_order=false
allow_overlap=false>
			<Translate chain=X distribution=uniform angstroms=2.0
cycles=50/>
		</CompoundTranslate>
		<Rotate name=rotate_X chain=X distribution=uniform degrees=360
cycles=500/>
		<SlideTogether name=slide_together chain=X/>
		<HighResDocker name=high_res_docker cycles=6 repack_every_Nth=3
scorefxn=ligand_soft_rep movemap_builder=docking/>
		<FinalMinimizer name=final scorefxn=hard_rep
movemap_builder=final/>
		<InterfaceScoreCalculator name=add_scores chains=X
scorefxn=hard_rep/>
compound movers
```

```
            <ParsedProtocol name=low_res_dock>
                  <Add mover_name=start_from_X/>
                  <Add mover_name=compound_translate/>
                  <Add mover_name=rotate_X/>
                  <Add mover_name=slide_together/>
            </ParsedProtocol>
            <ParsedProtocol name=high_res_dock>
                  <Add mover_name=high_res_docker/>
                  <Add mover_name=final/>
            </ParsedProtocol>
      </MOVERS>
      <PROTOCOLS>
            <Add mover_name=low_res_dock/>
            <Add mover_name=high_res_dock/>
            <Add mover_name=add_scores/>
      </PROTOCOLS>
</ROSETTASCRIPTS>
```

`LIGAND_AREAS` are used to describe the degree of protein and ligand flexibility in proximity to the protein/ligand interface. A cutoff value of 6.0Å means that any residue within 6.0Å of the ligand will be considered part of the interface. These values can be increased or decreased to enlarge or reduce the number of protein residues selected for rotamer sampling or backbone flexibility. The `minimize_ligand` value can be increased or decreased to alter the degree of ligand flexibility. This value represents the size of one standard deviation of movement in degrees. The `Calpha_restraints` value represents one standard deviation of alpha-carbon (CA) movement in angstroms (Å) and can be enlarged or reduced to alter the degree of backbone flexibility.

The coordinates given to the `StartFrom` mover should be adjusted to represent starting points for ligand docking. Typically, experimental data is used to determine the initial site of ligand docking. For this example, extensive experimental data has identified a small, buried hydrophobic binding site centered at A99[41]. An average was taken over the Cartesian coordinates for the beta-carbon atom of A99 for each of the ten models for the `StartFrom` mover in the script above.

The `Translate` mover's "`angstroms`" field should be adjusted to represent the size of the binding pocket that needs to be sampled. Because the ligand in this case is small,

the ligand is allowed to translate within a 2.0Å radius of the starting coordinates. As familiarity with the provided ligand docking XML protocol is accrued, experiment with developing a custom protocol. Typically, if no experimental data on binding of the ligand is present, a 5.0Å radius is used.

Preparation of the ligand docking options file

21. Create an options file called `ligand_dock.options`. In addition to the input PDB (`-in:file:s`) and the database location (`-database`), ligand `params` files must be provided (`-in:file:extra_res_fa`). These files were generated in step 18. The name of the XML file must be provided (`-parser:protocol`). PDB files are output by default. Below is the options file used for ligand docking in this example:

```
-in:file:s model_01.pdb #this option will need to be changed for each of the
ten models used in the docking protocol, e.g. model_10.pdb
-in:file:extra_res_fa MR3.params
-packing:ex1
-packing:ex2
-parser:protocol ligand_dock.xml
```

Accurate predictions of interfaces often rely on fine-grained placement of side chain atoms. Thus, it is recommended that the number of side chain rotamers is increased to include the standard rotamer plus or minus one standard deviation. This is accomplished as shown under the `packing` option group (`-packing:ex1, -packing:ex2`). See the Rosetta documentation for additional rotamer selection options.

Running the ligand docking job

22. Run the ligand docking job by specifying the executable and the options file on the command line:

```
rosetta_source/bin/rosetta_scripts.default.OperatingSystemrelease
@ligand_dock.options -database rosetta_database -nstruct 1000
```

The number of models (`-nstruct`) necessary to produce high-quality predictions will depend on the size of the binding pocket and the flexibility of the protein and small-molecule ligand. The number of models needed is directly proportional to the number of degrees of freedom in the system under study. For this example, the MR3 ligand is docked 1,000 times within each of the ten comparative models, for a total of 10,000 models docked with MR3.

A benchmarking study of docking ligands with Rosetta can be found in Lemmon, Kaufmann, and Meiler, 2012 (Lemmon, Kaufmann, & Meiler, 2012).

Timing

The indicated timing of each step is a rough estimate. The actual running time of steps that rely on external servers will depend on the number of jobs those servers are processing at the time, and these steps may therefore take much longer than the time estimates specified. Additionally, the run times of the Rosetta simulation steps will be longer than specified if a large protein and/or ligand are used. If the alignment or modeling steps are performed iteratively, the total run time for the iterative process will be longer than the listed time.

- Step 1: Template selection, 5 minutes

  Steps 2-4: Prepare the PDB file of template structure, 5 minutes
- Steps 5-7: Sequence alignment, 15 minutes
- Steps 8-9: Threading, 15 minutes
- Step 10: Prepare fragment files of the target sequence, 15-60 minutes
- Step 11: Creating a Rosetta loops file, 5 minutes
- Step 12: Preparation of comparative modeling options file, 5 minutes

- Step 13: Running the comparative modeling job, 5,000 CPU hours for 10,000 models (30 minutes per model). These models can be run on independent CPUs, decreasing total run time.
- Step 14-16: Analysis of comparative modeling results and choosing receptor models for ligand docking, 60 minutes
- Steps 17-19: Prepare the ligand file, 15 minutes
  Step 20: Preparation of the ligand docking XML file, 5 minutes
- Step 21: Preparation of the ligand docking options file, 5 minutes
- Step 22: Running the ligand docking job, 16 CPU hours for 10,000 models (5 seconds per model). These models can be run on independent CPUs, decreasing total run time.

Installing Rosetta 3.4

The Rosetta modeling suite is free of cost to all academic users after registration. The package comes with a user's guide; a database containing pertinent files for applications in Rosetta; the Rosetta source code, and a fragments directory containing a peptide fragment database of proteins for known structures. In addition, the modeling suite comes with FoldIt[3], an interactive graphical interface that manually folds proteins using the Rosetta scoring function and structure prediction algorithm. ProteinTools (`rosetta_tools/protein_tools`), a collection of ancillary tools commonly used in conjunction with the Rosetta software suite, is also included. Lastly, Rosetta comes with a software construction tool called SCons (http://www.scons.org/), which analyzes the source code and builds specified binary files using multiple processors. SCons interfaces with the standard GNU gcc compiler to build the source code.

1. Rosetta is free to academic users. For an academic license, apply here: http://depts.washington.edu/uwc4c/express-licenses/assets/rosetta+pyrosetta/
2. After obtaining a username and password, source code for Rosetta 3.4 can be downloaded here: https://www.rosettacommons.org/software/academic or https://www.rosettacommons.org/software/commercial
3. Create a new directory called rosetta (mkdir rosetta), and copy rosetta3.4_bundles.tgz from your downloads directory to rosetta/ (cp rosetta3.4_bundles.tgz rosetta/)
4. In the `rosetta` directory, unpack the `tar` file by typing the following: `tar -zxvf rosetta3.4_bundles.tgz`. This will unpack multiple files, including the Rosetta source code (`rosetta3.4_source.tgz`) and the Rosetta database (`rosetta3.4_database.tgz`).
5. Unpack the database and source code using the following:

```
tar -zxvf rosetta3.4_database.tgz
tar -zxvf rosetta3.4_source.tgz
```

6. Change directories into the newly created `rosetta3.4_source` (`cd rosetta3.4_source`), and build the binaries using the following command: `external/scons-local/scons.py mode=release bin/`

CRITICAL: Make sure you have the GNU gcc3.4 compiler or higher (by typing `gcc -v`) and that a working copy of Python2.5 or higher installed. Access a Python executable by `/usr/bin/env python` or explicitly type the path to your python executable. Additionally, you must have zlib installed, if you see errors referencing a missing –lz library, install the zlib-dev libraries for your operating system.

Note for Mac OS X Users: the compiler “clang” is recommended for compiling Rosetta in OS X. Make sure you have the OS X Developer tools installed, and then compile using the following build command: `external/scons-local/scons.py cxx=clang mode=release bin/`

If an error similar to `KeyError: "Unknown version number 4.1 for compiler 'clang'"` occurs, open the file `rosetta_source/tools/build/options.settings` and modify the line:

```
"clang" : [ "1.7", "2.1", "2.0", "2.8", "2.9", "3.0", "3.1","*" ],
```

To instead read:

```
"clang" : [ "1.7", "2.1", "2.0", "2.8", "2.9", "3.0", "3.1","4.1","*" ],
```

Executables compiled in this way will have the suffix ".macosclangrelease"

7. Change directory into `bin/ (cd bin)`, and confirm that all 166 binaries have been built.

The Rosetta modeling suite comes with a set of python scripts, which can greatly simplify the analysis of Rosetta models. These scripts are used throughout this tutorial. The scripts rely on several python dependencies, which need to be installed. BioPython and numpy, which are freely available python packages, are required to use these scripts. The specific installation instructions for BioPython (http://biopython.org/wiki/Biopython) and numpy (http://numpy.scipy.org/) will vary based on your specific operating system details. Consult the documentation for these packages for installation instructions.

Rosettautil is a python module with a number of useful functions for handling Rosetta output. The installation package for this Python module is located in `<rosetta_dir>/rosetta_tools/protein_tools`. To install this module, and the associated scripts, first change directories to the `<rosetta_dir>/rosetta_tools/protein_tools` directory. If you have root access and want to install the module so that it is usable by all users on the system, enter the following command:

```
python setup.py --install-scripts=/path/to/scripts/directory
```

This command will install the python module, and then copy the scripts into the directory you specify. If you do not have root access, use this command:

```
python setup.py --user --install-scripts=/path/to/scripts/directory
```

Clustering Using Rosetta

Predicted structures generated by comparative modeling are often clustered to help identify structurally similar models. Clustering is performed with the assumption that the deepest energy well, and hence the global energy minimum, will also be the widest[2]. As a result, it is expected that the largest clusters will potentially contain the predicted model that is closest to the native structure. Rosetta includes a tool for clustering protein models. The cluster application avoids the memory requirements associated with computing a complete distance matrix for large numbers of models. The Rosetta clustering method starts by computing a distance matrix for the first 400 input models. Each model in the distance matrix is assigned to the cluster to which it is nearest (typically in terms of RMSD). If the model is not within a specified radius of any cluster, it is assigned to a new cluster.

Because the Rosetta clustering application outputs most of its statistical information in its log file, a script has been provided to run the clustering application and produce a clear summary of the results. Given a set of PDB or Rosetta silent files and a Rosetta options file, `clustering.py` will produce a set of clustered PDBs, a histogram file showing the distribution of pairwise RMSDs between models, and a summary file showing which models are in which clusters. The Rosetta options file can contain a number of options that control the behavior of the cluster application. The acceptable options are listed below.

`-run:shuffle` - Input structures in a random order. Use this if you have reason to believe that the output models are not in random order

`-cluster:gdtmm` - cluster using GDTMM (global distance test) instead of RMSD distances

`-cluster:radius <float>` - The maximum radius of each cluster in Å (RMSD mode) or inverse GDT_TS (GDTMM mode)

`-cluster:input_score_filter <float>` - do not cluster scores above a given energy

`-cluster:exclude_res <int> <int> ….<int>` - do not include the listed residues in distance calculation

`-cluster:limit_cluster_size <int>` - maximum number of models in each cluster

`-cluster:limit_clusters <int>` - maximum number of clusters

`-cluster:limit_total_structures <int>` - maximum number of models to cluster

`-cluster:sort_groups_by_energy` - sort clusters by total energy during output

Using Constraints as Filters in RosettaScripts

Filters can be used to guide a RosettaScripts (Fleishman et al., 2011) protocol in producing only high quality models that pass constraints specified by the user. These filters can take on a variety of forms, where the protocol can be repeated until a certain score is met (filter by score), continuing the protocol until the model converges on another structure (filter by root mean square distance, or RMSD), or continuing the protocol until atomic contacts are made which agree with experimental observations (filter by experimental constraints). For example, an atomic contact or residue contact constraint is essentially a distance constraint that can be derived from several types of experimental data, such as NMR nuclear Overhauser effects (NOEs), distances determined by electron paramagnetic resonance (EPR), distances derived from cysteine mutagenesis, and more. The protocol repeats for all or specified movers until the experimental constraints are satisfied. In the XML script used by RosettaScripts, all filters are specified in the `FILTERS` section as shown below:

```
<FILTERS>
<filter1/>
      <filter2/>
      ...
</FILTERS>
```

Then, in the PROTOCOLS section:

```
<PROTOCOLS>
       ...
<Add mover_name=mover1 filter_name=filter1>
       <Add filter_name=filter2/>
       ...
</PROTOCOLS>
```

The general format for a filter placed in the XML script is:

```
<"filter_name" name="your_filter_name" ..parameter_name=<parameter_value>,
```

where "filter_name" is one of a predefined set of filters recognized by RosettaScripts, and name is a unique identifier for this particular filter, followed by the parameters that the specific filter needs to be defined. In the example, mover1 would continue until the constraints given in filter1 are satisfied. For filter2 that is not specified with a mover, the entire protocol would repeat up to filter2, until the constraints defined by the filter are met.

An example of an AtomicContact filter is:

```
<AtomicContact name=”res32_res45_noe” residue1=32 residue2=45 sidechain=0
backbone=1 protons=0 distance=10 confidence=0.25/>
```

This filter will check if the model generated by Rosetta satisfies the specified constraint between residues 32 and 45. If any pair of backbone residues between the two residues is within 10Å of each other, the filter will return TRUE 75% of the time. When the filter returns a TRUE value, the protocol is continued from where the filter is called. In addition, the “sidechain” and “protons” options have designated values of 0 (as opposed to 1), which means they are turned off. The side chain and hydrogen atoms are not taken into account, and only the main chain backbone atoms will be evaluated in this filter example. If the confidence is 1.0 then the filter is evaluated as either true or false. When the confidence value is less than 0.999, the filter will return TRUE in (1.0 – confidence) fraction

of the times it is evaluated. This so-called “fuzzy” filter is useful for instances of ambiguous or uncertain experimental data.

A less sophisticated type of constraint filter conceptually similar to the AtomicContact constraint filter is the ResidueDistance filter. This filter queries the distance between the beta-carbons of two specified residues. An example similar to the AtomicContact filter is:

```
<ResidueDistance name=”res32_res45_noe” res1_res_num=32 res2_res_num=45
distance=10.0 confidence=0.25/>
```

Finally, a DisulfideFilter can be applied, in which Rosetta tries to select models that position the specified residues such that they can potentially form a disulfide bond:

```
<DisulfideFilter name=disulfide targets=32,45,46 confidence=1.0/>
```

Notice that “`targets`” is defined as a comma-separated list of residues, which means that all numbers in the comma-separated list are considered when searching for a disulfide bond. For more information on using constraints in RosettaScripts[1], including constraint-type filters, go to:

http://www.rosettacommons.org/manuals/archive/rosetta3.4_user_guide/RosettaScripts.html

Testing Rosetta

Rosetta is often used to recapitulate known experiments. In such studies, several Rosetta options or protocol steps are changed carefully and methodically until the computational and experimental results correlate. Before conducting experiments with Rosetta, it is advised to test, or benchmark, the proposed protocol on known experimental data.

The definition of a successful benchmark varies and depends on the protocol. For a loop modeling benchmark, if the Rosetta model has a sub-angstrom RMSD to the experimental structure and is in the top ten lowest-scoring models built, the benchmark is

considered to be successful (Mandell et al., 2009). For ligand docking, success is achieved when one of the top ten scoring models has an RMSD to the native structure below 2Å (Davis & Baker, 2009).

The data obtained from loop building for the T4-lysozyme comparative model (see Figure 12) can be used as an example of benchmarking. The RMSD from the native ligand position can be calculated across all generated models via the scripts provided with a copy of Rosetta. To calculate the RMSD, run the provided script:

```
scripts/score_vs_rmsd_singleproc.py --native=target_A.pdb --
table=score_vs_rmsd_loops_ca.txt --term=total --ca_mode=ca --chain=A
loops_final_*.pdb --res=residues.ls
```

NOTE: The `InterfaceScoreCalculator`'s "`native`" option is used for benchmarking (if the native structure is known). To demonstrate the accuracy of the protocol in this example, the native structure is included in the script as `2ou0_native.pdb`. However, when running on a system where the native structure is not available, this option should be omitted. For reference, the results from this example will be compared against the experimentally determined T4-lysozyme crystal structure (referred to below as `2ou0_native.pdb`), which must also be aligned in the same coordinate system as the homology models used for docking. To enable RMSD calculation to the native, modify the `*.xml` script to include the native option:

```
<InterfaceScoreCalculator name=add_scores chains=X scorefxn=hard_rep
native="2ou0_native.pdb"/>
```

In your resulting `score.sc` file, the `interface_delta_X` is the score, and `ligand_rms_no_super_X` is the RMSD. The RMSDs for the lowest-energy ligand docking models for this example are considered relatively large; however, this is not unexpected due to the small size and the symmetry of the ligand. The binding conformation and position of this low-energy cluster of models could be another energy minimum separate from that found in the crystal structure. Further, Rosetta is able to sample the native ligand

conformation and position in a slightly higher-energy minimum. If the method you use is not yielding satisfactory test results, the size of the binding pocket search space can be decreased. For example, for this ligand (1-methyl-1H-pyrrole (MR3), a search radius of 2Å was used. In cases where the ligand is small and rigid, it is better to limit the degrees of freedom in which the ligand moves around the binding pocket. However, for larger ligands, a radius of up to 4-5Å may be needed to accommodate the ligand.

Glossary

- all-atom - in the case of sampling, synonymous with fine movements and often including side chain information; also referred to as high-resolution
- benchmark – another word for a test of a method, scoring function, algorithm, etc. by comparing results from the method to accepted methods/models
- binary file – a file in machine-readable language that can be executed to do something *in silico*
- BioPython – a set of tools for biological computing written and compatible with Python http://biopython.org/wiki/Biopython
- build – to compile the source code so it can be used as a program
- centroid – in Rosetta centroid mode, side chains are represented as unified spheres centered at the residue's center of mass
- cluster center – the geometric center of a cluster, or group, of models
- clustering – in this case, grouping models with similar structure together
- comparative model – a protein model where the primary sequence from one protein (target) is placed, or threaded, onto the three dimensional coordinates of a protein of known structure (template)language (binary)

- cyclic coordinate descent (CCD) – based on robotics, CCD loop closure is used to build loops in Rosetta by fragment assembly and close loops by decreasing the gap between two termini in three-dimensional space (Canutescu & Dunbrack, 2009)
- *de novo* – in this case, from the sequence; also called *ab initio*
- directory – synonymous with a folder, usually contains one or more files or other folders
- distance matrix – a matrix containing the pairwise distances for every point in a set of points
- Dunbrack rotamer library – a set of likely side chain conformations for the twenty canonical amino acids based on protein structures in the Protein Data Bank (PDB) (Dunbrack & Cohen, 1997)
- executable – binary file used to execute the program
- force field / scoring function / energy function / potential – often used interchangeably; a means of assessing the energy of the generated models
- fragment – in Rosetta folding and loop building, a set of three-dimensional coordinates corresponding to a given amino acid sequence
- fragment database – also called the fragment library; contains the Cartesian coordinates for 200 amino acid fragments (obtained from the Vall database) for each sequence window of the entire primary sequence of the protein
- gap – in sequence alignment, a gap is inserted when the sequences are of low homology; usually appear as a dash (-); the gaps form a sequence alignment correspond to areas where loops are built during comparative modeling
- GDT / GDT_TS – global distance test (total score); a measure of similarity between two protein structures having the same amino acid sequence; the largest set of

residues' Cα atoms in the model structure falling within a defined distance cutoff of their position in the experimental structure

- gradient-based minimization – also known as minimization by steepest descent; in this case, a means of energy minimization in which one takes steps proportional to the negative of the gradient of the function (energy) at the current point
- high-resolution – in the case of sampling, synonymous with fine movements and often including side chain information
- homology model – a more specific type of comparative model where the protein sequence of interest (target) is a homolog of the protein of known structure (template)
- interface delta - The interface delta score is defined as the contribution to the total score for which the presence of the ligand is responsible.
- kinematic loop closure (KIC) – robotics-inspired loop closure algorithm which analytically determines all mechanically accessible conformations for torsion angles of a peptide chain using polynomial resultants (Coutsias et al., 2004)
- knowledge-based – in the case of Rosetta, based on information obtained from structures found in the PDB
- libraries – in computing, a collection of code and data (classes and functions) used by a piece of software and is often used in software development
- ligand – in this case, a small molecule that binds to a protein to serve some biological purpose; in the presented protocol, the ligand (small molecule) is docked into the receptor (protein).
- low-resolution – a somewhat subjective term, in the case of sampling, synonymous with coarse movements of the protein and/or ligand backbone and side chains; the

individual atoms of low-resolution structures or models cannot be resolved, or observed.

- metropolis criterion – often combined with the monte carlo sampling algorithm; allows for generation of an ensemble that represents a probability distribution; see Metropolis, *et al*., 1953 (Metropolis et al., 1953)
- model – in the case of this protocol, a structure generated by Rosetta; sometimes called a decoy
- monte carlo sampling – a randomized and repetitive computational sampling method
- mover - a generic class that takes as input a pose and performs some modification on that pose; for example, a mover might take in a pose and rotate every residue
- namespace – in computer science, an abstract container holding a logical grouping of unique identifiers or symbols; in Rosetta, examples of namespaces are `loops`, `relax`, etc.
- native-like – close to the experimentally determined structure; a model that is "native-like" usually has an RMSD to the experimentally determined structure of < 2Å.
- options file – often called a flags file; a file containing Rosetta options that can be passed to a Rosetta executable after the @ symbol; can be easier to use than passing Rosetta options over the command line
- pack / repack – in Rosetta, side chains are packed/repacked by switching out rotamers and scoring them using the Rosetta scoring function
- params file – ligand file; defines the Rosetta atom-typing and internal coordinates of a small molecule; in the chapter, specified by *.params
- path – in this case, the location in the file system of a file or directory

- physics-based – in the case of scoring functions, based on Newtonian physics; for example, two atoms are considered to be balls connected by a spring; often used in molecular mechanics
- pose – in this Rosetta protocol, a three-dimensional conformation of the ligand, protein, or ligand/protein complex
- Python – interpreted, object-oriented, high-level programming language http://www.python.org/
- relax – in Rosetta, an iterative protocol of side chain repacking and gradient-based minimization; often referred to as full-atom (or all-atom) refinement
- Robetta – Rosetta structure prediction server http://robetta.bakerlab.org/ freely available to not-for-profit users
- RosettaCommons – a group of more than twenty labs that develop the Rosetta software suite
- Rosetta energy units (REU) – *not* experimental energy units; these are arbitrary energy units specific to the Rosetta scoring function
- RosettaScripts – also called the Scripter or RosettaXML; an XML-like language that allows for specifying modeling tasks in Rosetta (Fleishman et al., 2011)
- rotamer – rotational conformer of an amino acid or ligand side chain
- SCons – a tool for constructing software from its source code http://www.scons.org/
- script - in computer programming, a script is a sequence of instructions that is interpreted or carried out by another program rather than by the computer processor (as a compiled program is)
- source code – human-readable files that are the implementation of the program; in Rosetta, these are written in C++.

- target – in comparative, or homology, modeling, the protein for which we are generating a model; the target sequence is the primary sequence of the protein for which we want to make a model.
- template – in comparative modeling, the protein of known structure on which the target is threaded
- threading – placing the primary sequence of one protein (target) on the three-dimensional coordinates of a protein of known structure (template) based on a sequence alignment
- Vall database – database from which amino acid fragments are picked for folding and CCD loop building in Rosetta
- XML – Extensible Markup Language; in this case, used to write protocols to pass to RosettaScripts

Troubleshooting Tips

Table 20
Troubleshooting tips for comparative modeling and ligand docking with Rosetta.

Step	Problem	Possible Reason	Solution
1	No suitable template structure is found.	It is possible that no experimental structure has been determined for a homologous protein with greater than 30% sequence identity.	Remote homolog detection using methods such as threading may be able to identify a more distantly related template structure. This will result in a model of lower confidence. In some cases, Rosetta can be used to perform *de novo* structure prediction instead of comparative modeling.
3	clean_pdb.py script gives message: “Found preoptimized or otherwise fixed pdbfile.”	There are no HETATM or non-ATOM records to remove.	This is not actually a problem. No action required. But it is always a good idea to examine the PDB file in a text editor and a structure viewer to understand the details of the template structure.
3	The `clean_pdb.py` script does not run.	The script was not made executable when it was downloaded.	The Python script needs to be given executable permissions with a command similar to this: `chmod +x ./clean_pdb.py`
6	The resulting sequence alignment between the target and template sequences contains evident errors.	Ultimately, no automated sequence alignment algorithm is as good as an experienced biologist.	Do not hesitate to hand-edit the sequence alignment to ensure that wherever possible, functionally important residues align properly, secondary structural elements are conserved, and insertions/deletions are localized to loop regions. This will greatly increase the quality of the model.
6	The sequence alignment contains unaligned N- or C- terminal extensions.	The target sequence is longer than template structure (or *vice versa*).	Before aligning, trim the target sequence so that the N- and C-termini match the termini in the template PDB file.
8	thread_pdb_from_alignment.py does not run.	BioPython is not installed.	Install Python (version 2.5 or later) with the optional BioPython package in order to run these scripts.

Step	Problem	Possible Reason	Solution
8	thread_pdb_from_alignment.py “We cannot completely thread this protein in an automatic way, manual inspection and adjustment of loops files will be required.”	There may be gaps within the original template protein. This script will give this message when handling an alignment containing gaps greater than 3 residues in the template sequence.	While a set of loop definitions will be output to the screen, these loop definitions only include regions of the threaded protein that contain gaps corresponding to unaligned regions of the protein. In cases where additional regions need remodeling, it will be necessary to correct the loop definitions by hand (Figure 10). See *Experimental design* for details on how to determine the suggested loop definitions.
8	thread_pdb_from_alignment.py “can’t find alignment in alignment file.”	The alignment is in wrong file format, or the template or target names are not what was specified on the command line.	Make sure the file is in ClustalW format. Edit the alignment manually so that the target and template names exactly match the arguments passed to the script.
8	thread_pdb_from_alignment.py gives a Traceback with an AttributeError	Missing arguments on the command line	Be sure to specify all necessary options, including --template=<x> --target=<x> --chain=<x> followed by the 3 input files: <alignment.filename> <template.pdb> <output.filename>
8	thread_pdb_from_alignment.py says “residue mismatch between alignment and PDB”	The sequence in the template PDB is not identical to the template sequence in the alignment.	Use the FASTA sequence extracted from the PDB file using `clean_pdb.py` to generate the sequence alignment.
8	`thread_pdb_from_alignment.py` gives a loops suggestion in which one loop is only one residue long.	There is a point insertion in the alignment.	Edit the loop to include 1 or 2 residues on each side of the point insertion, to give greater flexibility for closing the loop.
10	Difficulty generating fragments file.	Using the fragment picker and the `make_fragments.pl` script that comes with the Rosetta source code has many prerequisites. The installation of the prerequisite programs is an involved procedure.	The simplest way to generate a fragments file is to use the Robetta server at: www.robetta.bakerlab.org This service is available free to all non-commercial users. See STEP 10A.

Step	Problem	Possible Reason	Solution
11	Loops file not recognized by Rosetta	Spaces and tabs were used interchangeably in the file.	Use either spaces OR tabs in the loop file, but NOT both. Make sure it is a plain text file, not e.g., a formatted Word document. Make sure a current loop file format is used (although Rosetta will try to automatically translate older formats).
11	Loops file is not functional. Rosetta runs, but gives errors during loop sampling.	Loops are too long for Rosetta to adequately sample.	Guidelines for loops: 1) Individual loops should not be greater than 10 to 12 amino acids long. 2) Rosetta can have trouble with N-terminal and C-terminal tails. It is best to trim the target termini to match the template.
12	Options file is not recognized	Spaces and tabs were used interchangeably in the options file.	Use either spaces OR tabs in the options file, but NOT both. Make sure it is a plain text file, not e.g., a formatted Word document.
13	Rosetta fails to run or contains `ERRORS` in the log file referencing the input PDB file of the template structure.	An input PDB file containing non-standard residues, including certain ions, small-molecules, and post-translational modifications that are not included in the standard residue database, or with missing density that includes backbone atoms cannot be used.	1) Manually edit the PDB file to remove or rename the offending residues with standard names. 2) Ensure that the input PDB file is properly formatted, especially with respect to column spacing. The reference for the format is here: http://www.wwpdb.org/docs.html 3) In some cases, custom parameters for the non-standard residues will need to be made and those files included in the command line.
13	Rosetta fails to complete the comparative modeling run.	This can happen when the input file has missing backbone density in non-loop regions.	To supply a starting point for the missing density, re-build loops by “modeling” the template PDB file on its own complete sequence.
13	Rosetta fails to complete the comparative modeling run. The log file contains “`permanent failure.`”	This can happen when loops are inserted into regions where they do not fit.	Extend the endpoints of the loops to increase the flexibility of the loops.
13	Rosetta fails to complete the comparative modeling run.	The alignment is bad, e.g., a proline is placed in the middle of a helix.	Manually edit the alignment file to make it more biophysically reasonable.

Step	Problem	Possible Reason	Solution
14	Cannot select the best model by Rosetta Energy Units because the scores are too similar.	Inadequate sampling can reduce the ability to distinguish good models from bad by score alone.	Increase the number of models generated by a factor of 10. Or use a clustering approach to identify the most commonly generated conformations. Or incorporate experimental restraints to filter the resulting models.
18	The `molfile_to_params.py` script does not run.	The script was not made executable when it was downloaded.	The Python script needs to be given executable permissions with a command similar to this: `chmod +x ./molfile_to_params.py`
20	The XML file is not recognized by Rosetta.	Formatting was included in the XML file.	Make sure it is a plain text file, not e.g. a formatted Word document. See Box 3 for details.
22	The ligand docking job does not run.	Rosetta cannot find the input files.	Make sure the path options are correct, and point towards the actual location of the input files.

APPENDIX TO CHAPTER III

Protocol capture for GPCR comparative modeling and ligand docking computational methods

This protocol capture contains the steps necessary to obtain the results presented in Chapter III. While the actual protocol was carried on every pairwise combination of GPCRs from Table 4, this protocol capture uses the comparative modeling of bRh onto the template B2Ar as an example for simplification. The Rosetta 3.4 software suite is publically available and the license is free for non-commercial users at http://www.rosettacommons.org/.

1. Structural alignment of GPCR templates

Step	Text	Commands	Comment
1A. Prepare GPCR crystal structures from the Protein Data Bank.	The highest resolution experimental structure for each unique GPCR in the Protein Data Bank (PDB) at the time of writing was chosen for comparative modeling and ligand docking, as shown in Table 1.	Obtain PDB files: Download GPCR crystal structures from the Protein Data Bank at http://www.rcsb.org. Clean PDB files: Clean PDB files using the following script, written here for use with B2Ar: rosetta_tools/protein_tools/scripts/clean_pdb.py 2RH1 A > 2rh1A_clean.pdb Remove lines in the PDB file representing the N-terminal, C-terminal and T4-lysozyme regions.	Input: GPCR crystal structure PDB files from the Protein Data Bank at http://www.rcsb.org. Output: 1u19A_clean.pdb 2vt4A_clean.pdb 2rh1A_clean.pdb 3emlA_clean.pdb 3oduA_clean.pdb 3pblA_clean.pdb 3rzeA_clean.pdb 3v2wA_clean.pdb 3uonA_clean.pdb 4dajA_clean.pdb 4dklA_clean.pdb 4djhA_clean.pdb 4ea3A_clean.pdb 4ej4A_clean.pdb

1B. Perform a structural alignment of GPCRs using crystal structures from the Protein Data Bank.	A structure-based sequence alignment was generated of all 14 GPCR templates using MUSTANG.	mustang -p . -i 1u19A_clean.pdb 2vt4A_clean.pdb 2rh1A_clean.pdb 3emlA_clean.pdb 3oduA_clean.pdb 3pblA_clean.pdb 3rzeA_clean.pdb 3v2wA_clean.pdb 3uonA_clean.pdb 4dajA_clean.pdb 4dklA_clean.pdb 4djhA_clean.pdb 4ea3A_clean.pdb 4ej4A_clean.pdb -o all_gpcrs -F fasta -D 2.5 -s ON	Input: 1u19A_clean.pdb 2vt4A_clean.pdb 2rh1A_clean.pdb 3emlA_clean.pdb 3oduA_clean.pdb 3pblA_clean.pdb 3rzeA_clean.pdb 3v2wA_clean.pdb 3uonA_clean.pdb 4dajA_clean.pdb 4dklA_clean.pdb 4djhA_clean.pdb 4ea3A_clean.pdb 4ej4A_clean.pdb Output: all_gpcrs.fasta

2. Sequence alignment of the target GPCR to template sequences

Step	Text	Commands	Comment
2A. Obtain sequence of the target GPCR.		Save sequence output from `clean_pdb.py` into a FASTA file called `1u19A.fasta`.	Input: 1u19A_clean.pdb Output: 1u19A.fasta
2B. Sequence alignment of the target GPCR [bRh] to templates [B2Ar].	The sequence of the target GPCR was then aligned with the profile of structurally aligned templates using CLUSTALW.	Input target sequence `1u19A.fasta` and profile alignment `all_gpcrs.fasta` to http://mobyle.pasteur.fr/cgi-bin/portal.py#forms::clustalO-profile. Default settings were used.	Input: 1u19A.fasta, all_gpcrs.fasta Output: 1u19A.aln

3. Thread target sequence onto template backbone coordinates

Step	Text	Commands	Comment
3. Thread target sequence bRh onto template B2Ar backbone coordinates.	The sequence of the target GPCR was then placed onto the helical backbone coordinates of each template structure.	rosetta_tools/protein_tools/scripts/thread_pdb_from_alignment.py --template=2rh1A_clean --target=1u19A --chain=A --align_format=clustal 1u19A.aln 2rh1A_clean.pdb 1u19A_on_2rh1A.pdb	Input: 1u19A.aln 2rh1A_clean.pdb Output: 1u19A_on_2rh1A.pdb

4. Rebuild missing density

Step	Text	Commands	Comment
4A. Generate secondary structure prediction, constraint file and fragments for bRh.		Secondary structure- Jufo9D: http://meilerlab.org/index.php/servers/show?s_id=5 Secondary structure- PSIPRED: http://bioinf.cs.ucl.ac.uk/psipred/ Transmembrane span prediction based on Jufo9D: `perl scripts/jufo9d_span.pl 1u19A.jufo9d > 1u19A.span` Disulfide bond constraint file: Create file that lists residue number of cysteine residues predicted to disulfide bond according to the alignment with the template. Fragment files: http://www.robetta.org Check for exclusion of bRh from the fragment database.	Input: 1u19A.fasta Output: 1u19A.jufo_ss, 1u19A.psipred_ss2, 1u19A.span, 1u19A.disulfide, aa1u19A03_05.200_v1_3, aa1u19A09_05.200_v1_3
4B. Rebuild missing density caused by gaps in the sequence alignment.	Any missing density and variable loop regions were constructed using the ab initio cyclic coordinate descent protocol in Rosetta.	Generate loops file: In this case, the loop definitions will span regions where gaps were located in the sequence alignment. List the residue numbers in the loop file as shown in `1u19A_on_2rh1A.loops`. Generate options file: List the desired options for rebuilding loop regions in an options file as shown in `ccd_initial.options`. Run loop building: `rosetta_source/bin/loopmodel.linuxgccrelease @ccd_initial.options -database rosetta_database`	Input: ccd_initial.options, 1u19A_on_2rh1A.pdb, 1u19A.span, 1u19A.disulfide, 1u19A_on_2rh1A.loops, aa1u19A09_05.200_v1_3, aa1u19A03_05.200_v1_3 Output: 200 models of 1u19A from 2rh1A template with missing density rebuilt, for example: 1u19A_on_2rh1A_initial.pdb

5. Rebuild ECL 1,2 and 3

Step	Text	Commands	Comment
5A. Construct comparative model by rebuilding loop regions in Rosetta with CCD.	Extracellular loops were extensively rebuilt using both the cyclic coordinate descent loop closure method described above and the kinematic loop closure method described below.	Generate loops file: The loop definitions span the region between transmembrane helices. List the residue numbers for extracellular loops in the loop file as shown in `1u19A.loops`. Generate options file: List the desired options for rebuilding loop regions in an options file as shown in `ccd.options`. Run loop building: `rosetta_source/bin/loopmodel.linuxgccrelease @ccd.options -database rosetta_database`	Input: ccd.options, 1u19A_on_2rh1A_initial.pdb, 1u19A.span, 1u19A.disulfide, 1u19A.loops, aa1u19A09_05.200_v1_3, aa1u19A03_05.200_v1_3 Output: 1000 models of 1u19A from 2rh1A template with ECLs rebuilt, for example: 1u19A_rmsd01.pdb
5B. Construct comparative model by rebuilding loop regions in Rosetta with KIC.	Extracellular loops were extensively rebuilt using both the cyclic coordinate descent loop closure method described above and the kinematic loop closure method described below.	Generate loops file: The loop definitions span the region between transmembrane helices. List the residue numbers for extracellular loops in the loop file as shown in `1u19A.loops`. Generate options file: List the desired options for rebuilding loop regions in an options file as shown in `kic.options`. Run loop building: `rosetta_source/bin/loopmodel.linuxgccrelease @kic.options -database rosetta_database`	Input: kic.options, 1u19A_on_2rh1A_initial.pdb, 1u19A.span, 1u19A.disulfide, 1u19A.loops Output: 1000 models of 1u19A with ECLs rebuilt, for example: 1u19A_rmsd01.pdb

6. Evaluate comparative models by clustering by full-receptor RMSD and knowledge-based pocket residue filter

Step	Text	Commands	Comment
6A. Analyze results by clustering top ten percent of comparative models by full receptor RMSD.	The first method was based on clustering of the 10% best scoring structures. Clusters were determined based on pairwise RMSD of all C-alpha atoms using bcl::Cluster and a cluster radius of 3.0 Å. The best scoring models in each of the clusters were used for further analysis.	Filter for the top ten percent of models by energy: Rosetta energy units incorporating the implicit membrane potential for each model are found in the *.out file. Generate table of pairwise RMSD values: bcl::PDBCompare was used to generate a table of pairwise RMSD values between comparative models. Download the bcl software suite at (the license is free for non-commercial users). http://www.meilerlab.org/index.php/bclcommons/show/b_apps_id/12 bcl.exe PDBCompare -quality RMSD -atoms CA -pdb_list 1u19A_models.ls -aaclass AACaCb -prefix 1u19A_10percent_ Cluster models by RMSD: bcl.exe Cluster -distance_input_file 1u19A_10percent_RMSD.txt -input_format TableLowerTriangle -output_format Rows Centers -output_file cluster3_1u19A -linkage Average -remove_internally_similar_nodes 3	Input: PDB files for top ten percent of 1u19A comparative models by Rosetta energy and names of those PDB files in a list called `1u19A_models.ls` Output: `1u19A_10percent_RMSD.txt`, `cluster3_1u19A.Centers`, `cluster3_1u19A.Rows`

6B. Analyze results by filtering comparative models with a knowledge-based filter.	The second method was constructed to interrogate and avoid sampling of non-native ligand binding pocket conformations. Pocket residues were defined as a residue in any GPCR that had at least a 4.0 Å distance to the ligand in the experimental structure. This yielded a list of 29 residues that was reduced to 25 residues when four residues at the top of transmembrane helix (TM) two and five were removed to avoid bias from structural alignment of the proteins. Comparative models passed the filter only if C-alpha atoms of all pocket residues had an alignment equivalent pocket residue in another GPCR within a distance of a residue specific cutoff.	Calculate the minimum distance to any alignment equivalent position in any GPCR: For all pocket residues the minimum distance to any sequence alignment equivalent residue in any GPCR is determined with PyMOL.
		scripts/evaluate_score_vs_pocket_rmsd/01_make_distances.csh
		
		scripts/evaluate_score_vs_pocket_rmsd/02_filter_models.py
			Input:
			Structures to be filtered should be in:
			scripts/evaluate_score_vs_pocket_rmsd/structures/ID/ID_struc_id.pdb
			Crystal structures for distance calculations are placed in:
			`crystal_pockets/`
			The residue numbering of the models must be identical to that of the crystal structures.
			
			Output:
			A list of filtered structures is generated in
			`scripts/evaluate_score_vs_pocket_rmsd/pdb_lists_filtered/`

7. Generate ligand conformations in MOE

Step	Text	Commands	Comment
7. Create ligand conformations in MOE.	In preparation for docking, ligand conformers were generated by MOE (Molecular Operating Environment, Chemical Computing Group, Ontario, Canada) using the MMFF94x force field and Generalized Born implicit solvent model. Conformers were generated using 10,000 iterations of the Low Mode MD method with a redundancy cutoff of 0.25 Å. Energy cutoffs for ligand conformers were dependent on the number of rotatable bonds: 3 kcal/mol for 1-6 rotatable bonds, 5 kcal/mol for 7-9 rotatable bonds and 7 kcal/mol for 10-12 rotatable bonds.	Generate ligand conformations in MOE: See MOE operating guide. LowModeMD with the MMFFx94 force field and Generalized Born solvation model was used to generate conformations within the specified energy cutoff. The ligand conformations were then saved as an `.sdf` file for conversion to `.pdb` and `.params` files for Rosetta. Convert .sdf file of ligand conformations to .pdb and .params file for Rosetta input: `rosetta_source/src/python/apps/public/molfile_to_params.py -n 1u19A -p 1u19A 1u19A.sdf` Combine all individual ligand conformations in pdb format to a file called `1u19A_confs.pdb`. Add the line "`PDB_ROTAMERS 1u19A_confs.pdb`" to the bottom of the `1u19A.params` file.	Input: ligand coordinates in mol format: `1u19A.sdf` Output: `1u19A.params`, `1u19A_confs.pdb`

8. Dock ligand into comparative models

Step	Text	Commands	Comment
8A. Generate input files necessary for docking with Rosetta Scripts.	Ligand docking into the comparative models was performed with Rosetta Scripts. Each ligand was allowed to sample binding modes in a 5.0 Å radius from the crystallized binding mode. This adds some bias to the sampling, as the smallest unbiased docking sphere enclosing all ligand binding conformations has a radius greater than 5.0 Å.	Prepare input pdb files: Align the comparative model for docking to the crystal structure `1u19A_clean.pdb`. Copy one ligand conformation from the `1u19A_confs.pdb` file to the bottom of the pdb file of the starting model, `1u19A_cluster01_01.pdb`. Save as `1u19A_cluster01_01_ligand.pdb`. Prepare options file for docking: List the desired options for docking in an options file as shown in `dock.options`. Prepare XML file for docking: List the desired specifications for docking in an options file as shown in `dock.xml`.	Input: 1u19A_clean.pdb, 1u19A_confs.pdb, 1u19A_cluster01_01.pdb Output: 1u19A_cluster01_01_ligand.pdb, dock.options, dock.xml

8B. Dock ligand within bRh comparative models.	For each ligand, over 2,000 docked complexes were generated.	`rosetta_source/bin/rosettascripts.linuxgccrelease @dock.options -database rosetta_database`	Input: `1u19A_cluster01_01_ligand.pdb, 1u19A.params, dock.xml, dock.options` Output: 1000 models of retinal bound to bRh, for example: 1u19A_cluster01_01_ligand_011u19A_cluster01_01_ligand_0001.pdb

9. Analyze results by clustering binding modes by ligand RMSD

Step	Text	Commands	Comment

9. Analyze results by clustering binding modes by ligand RMSD.	Results from the ligand docking study were evaluated using clustering on pairwise RMSD values calculated over the ligand heavy-atoms using bcl::Cluster with a 2.0 Å cutoff. The lowest energy binding modes of the five largest clusters were chosen for further analysis.	Align PDBs :
		Use PyMOL to align receptor backbone coordinates.
		Extract lines for ligand coordinates into an .sdf file:
		`/scripts/rmsd.tcsh *.pdb`
		Generate table of pairwise RMSD values:
		bcl::ScoreSmallMolecule was used to generate a table of pairwise RMSD values between ligand coordinates. Download the bcl software suite at (the license is free for non-commercial users).
		http://www.meilerlab.org/index.php/bclcommons/show/b_apps_id/12
		`bcl.exe ScoreSmallMolecule all.sdf output.sdf -comparison`
		RMSD
		Cluster models by RMSD:
		`bcl.exe Cluster -distance_input_file 1u19A_ligand.cluster.mat -input_format TableLowerTriangle -output_format Rows Centers -output_file cluster3_1u19A_ligand -linkage Average -remove_internally_similar_nodes 3`
			Input:
			PDB files for 1u19A docked models.
			
			Output:
			all.sdf,
			1u19A_ligand.cluster.mat,
			cluster3_1u19A_ligand.Centers,
			cluster3_1u19A_ligand.Rows

Table 21

Rosetta loop modeling in comparative models with cyclic coordinate descent compared to kinematic loop closure.

Reported is the average RMSD and standard deviation for all comparative models of target receptors, calculated over C-alpha atoms in the loop regions compared to the corresponding experimental structure from the Protein Data Bank. Loop closure with KIC was only performed on a subset of the GPCR dataset.

	Extracellular Loop 1 RMSD (Å)		Extracellular Loop 2 RMSD (Å)		Extracellular Loop 3 RMSD (Å)		Full Receptor RMSD (Å)	
	CCD	KIC	CCD	KIC	CCD	KIC	CCD	KIC
bRh	1.7±0.7*	2.0±0.8	7.3±1.4	7.3±4.2	2.6±0.6	2.6±0.8	4.7±0.6*	10.6±8.5
B1Ar	1.8±0.6*	2.5±0.8	6.0±1.2	5.7±3.3	1.9±0.7*	2.1±0.5	3.6±0.6*	4.3±3.4
B2Ar	1.5±0.7*	3.2±1.2	6.2±1.3*	6.7±3.8	1.7±0.5*	2.1±0.6	3.7±0.5*	4.9±3.4
A2Ar	1.4±0.5*	2.2±1.2	n.d.[a]	n.d.[a]	2.5±0.6*	2.7±1.4	3.4±0.5*	8.5±5.9
CXCR4	1.6±0.4*	2.5±1.1	5.3±1.0*	6.2±3.3	2.8±0.7*	4.5±3.0	4.5±0.4*	9.3±8.7
D3R	2.1±0.6*	2.3±0.9	4.3±1.2*	5.3±3.0	2.1±0.5*	2.9±1.8	2.9±0.4*	4.0±3.1
H1R	1.5±0.6	n.d.	n.d.[a]	n.d.[a]	2.0±0.6	n.d.	2.5±0.6	n.d.
S1P1R	2.9±0.6	n.d.	5.9±0.8	n.d.	3.4±0.8	n.d.	4.2±0.9	n.d.
M2R	1.6±0.8	n.d.	4.8±0.8	n.d.	1.4±0.7	n.d.	2.6±0.4	n.d.
M3R	1.8±0.5	n.d.	5.1±0.9	n.d.	1.1±0.4	n.d.	2.8±0.4	n.d.
MOR	1.5±0.6	n.d.	5.6±1.4	n.d.	1.6±0.8	n.d.	3.2±0.9	n.d.
KOR	1.0±0.8	n.d.	5.3±1.0	n.d.	n.d.[a]	n.d.	3.5±0.4	n.d.
NOP	1.3±0.7	n.d.	5.5±1.1	n.d.	2.9±0.9	n.d.	3.1±0.6	n.d.
DOR	1.4±0.6	n.d.	5.7±1.1	n.d.	2.4±0.6	n.d.	3.2±0.5	n.d.

n.d. denotes not determined

* Indicated significant improvement using the CCD method over KIC for the given category, evaluated with the unpaired t-test, $p < 0.05$

[a] Could not be evaluated because of unresolved structure in this region of the experimental structure in the Protein Data Bank.

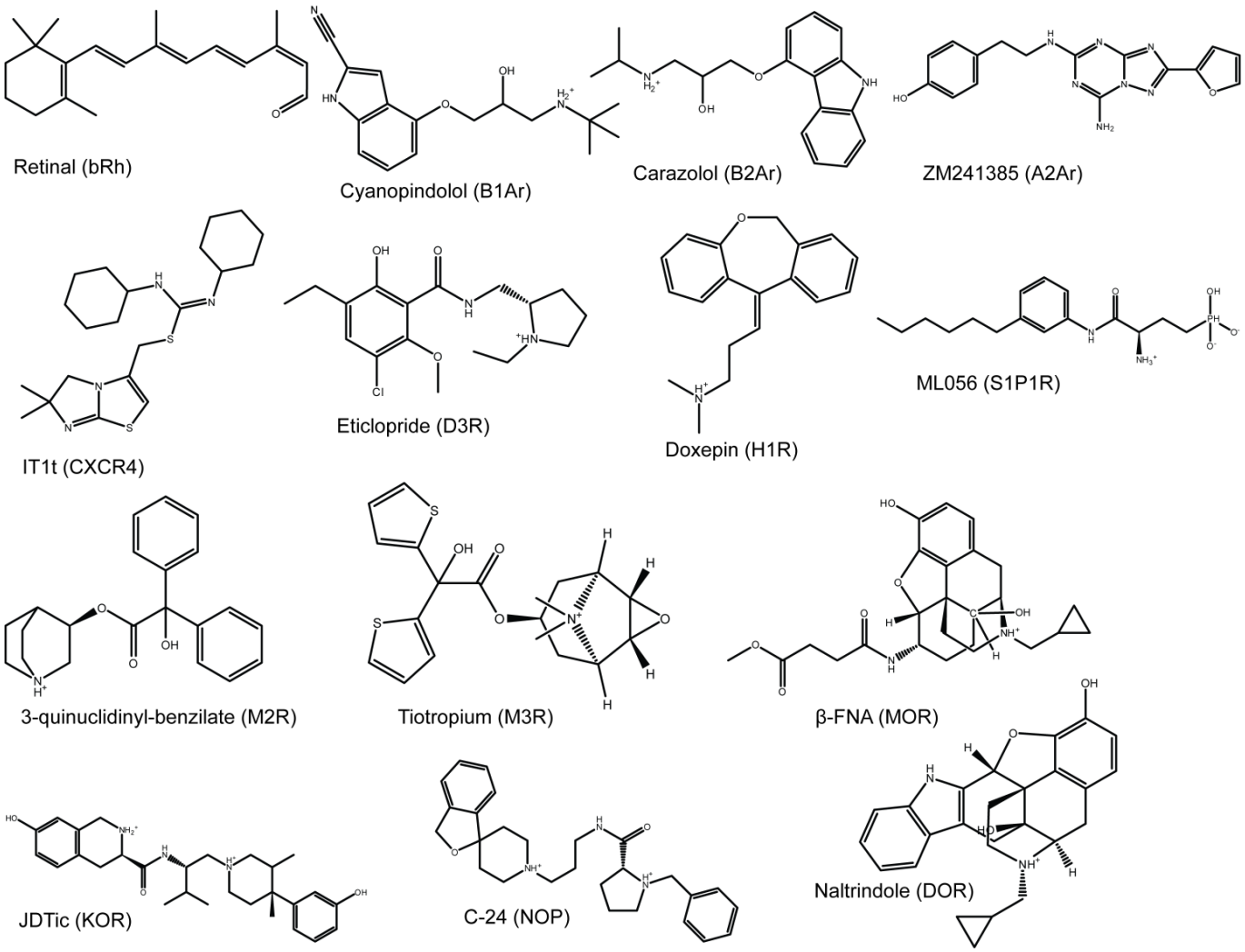


Figure 43
Ligand structures used in this study.
Ligand structures depicted here were crystallized with the G protein-coupled receptors used in this study and were obtained from the Protein Data Bank.

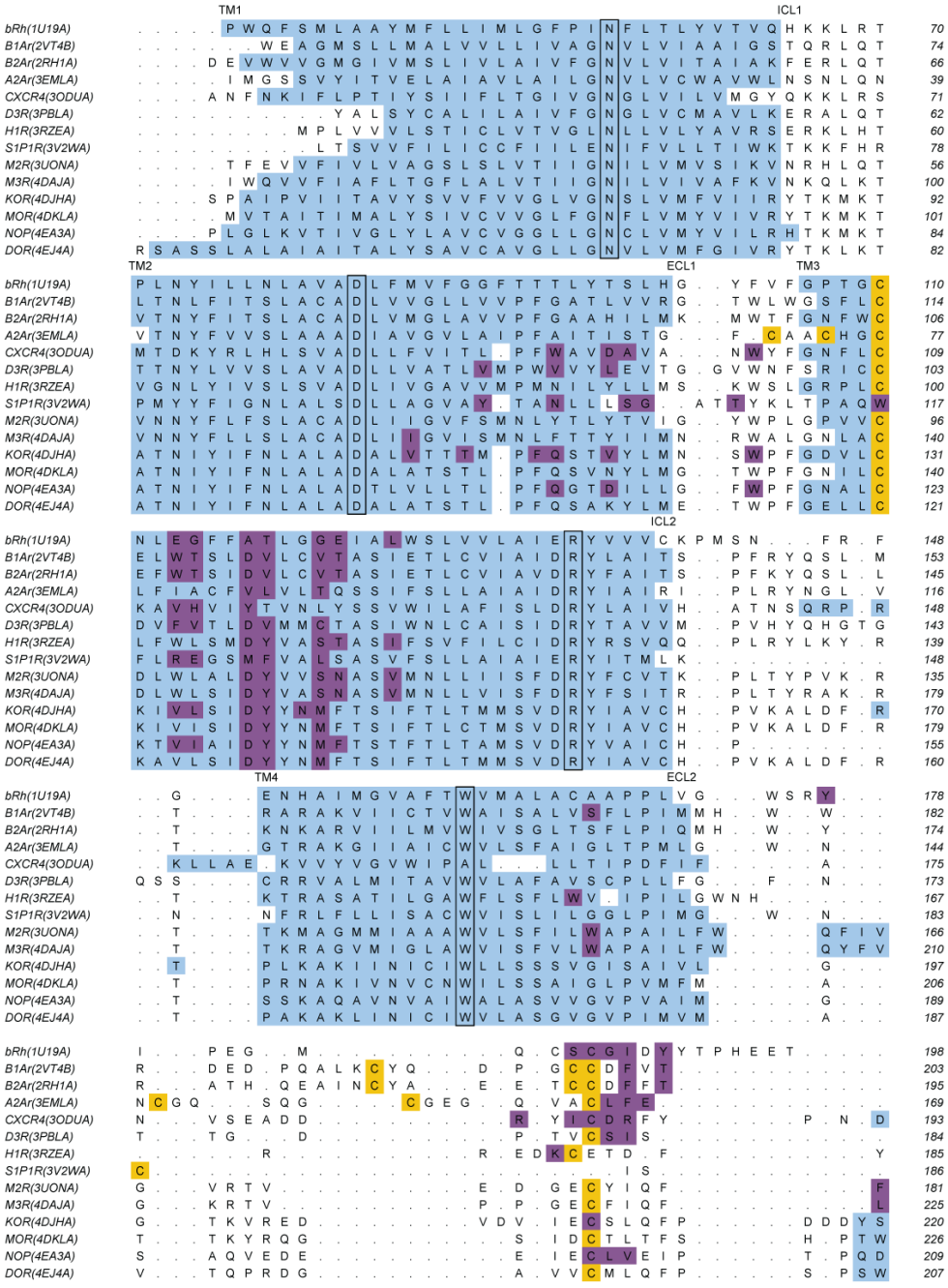
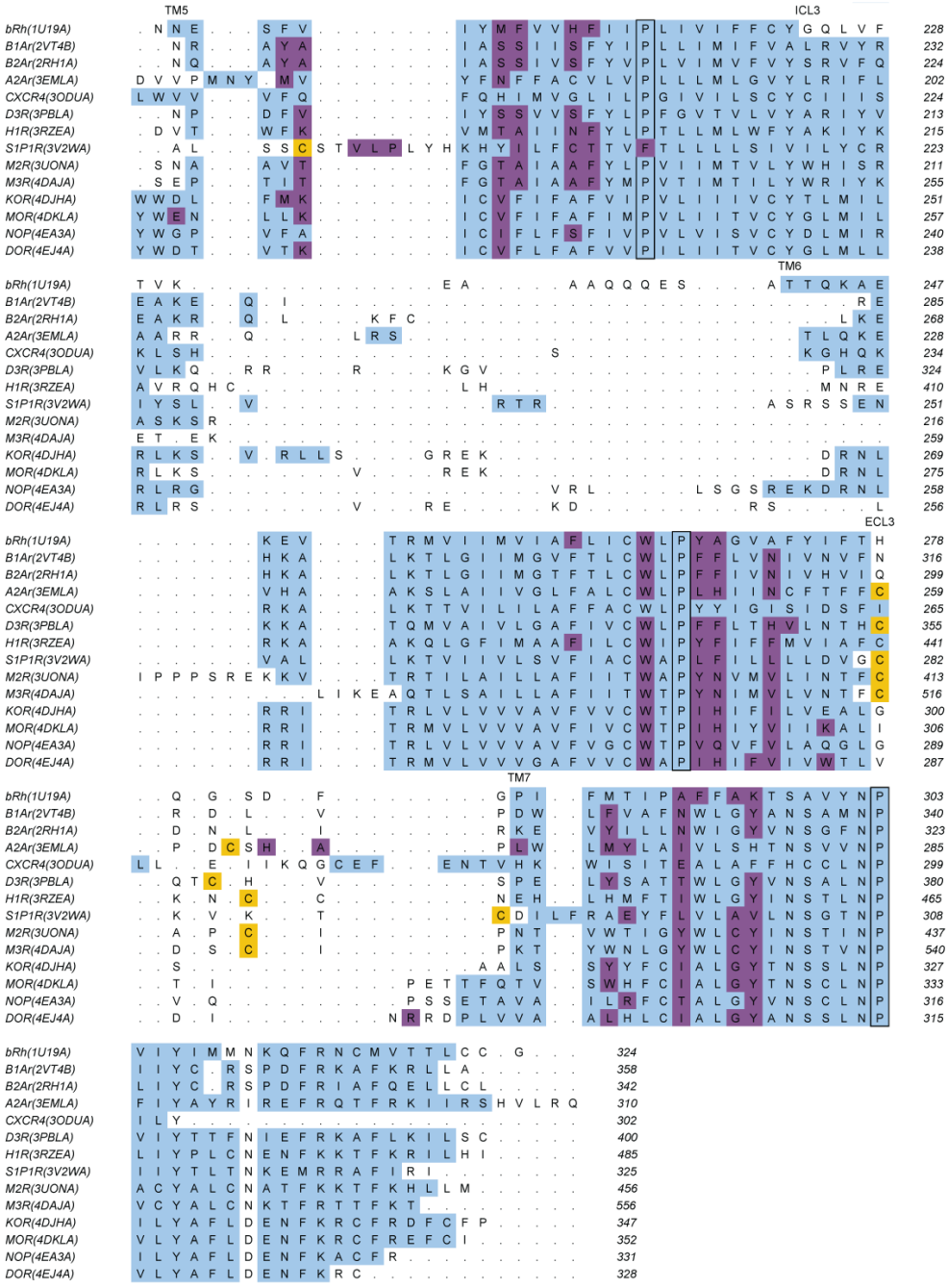


Figure 44
Structure-based sequence alignment of G protein-coupled receptors.
This sequence alignment of the fourteen GPCRs used in this study was obtained through a structural alignment of the receptors in MUSTANG (Konagurthu et al., 2006). Transmembrane regions are highlighted in blue, cysteine residues forming disulfide bonds are highlighted in yellow and residues in contact with their respective ligands are highlighted in purple. Conserved residues representing Ballesteros-Weinstein x.50 are outlined with a black box. The figure was generated using Aline (Bond & Schuttelkopf, 2009). Continued on next page.



(Continuation of Figure 44 from previous page) This sequence alignment of the fourteen GPCRs used in this study was obtained through a structural alignment of the receptors in MUSTANG (Konagurthu et al., 2006). Transmembrane regions are highlighted in blue, cysteine residues forming disulfide bonds are highlighted in yellow and residues in contact with their respective ligands are highlighted in purple. Conserved residues representing Ballesteros-Weinstein x.50 are outlined with a black box. The figure was generated using Aline (Bond & Schuttelkopf, 2009).

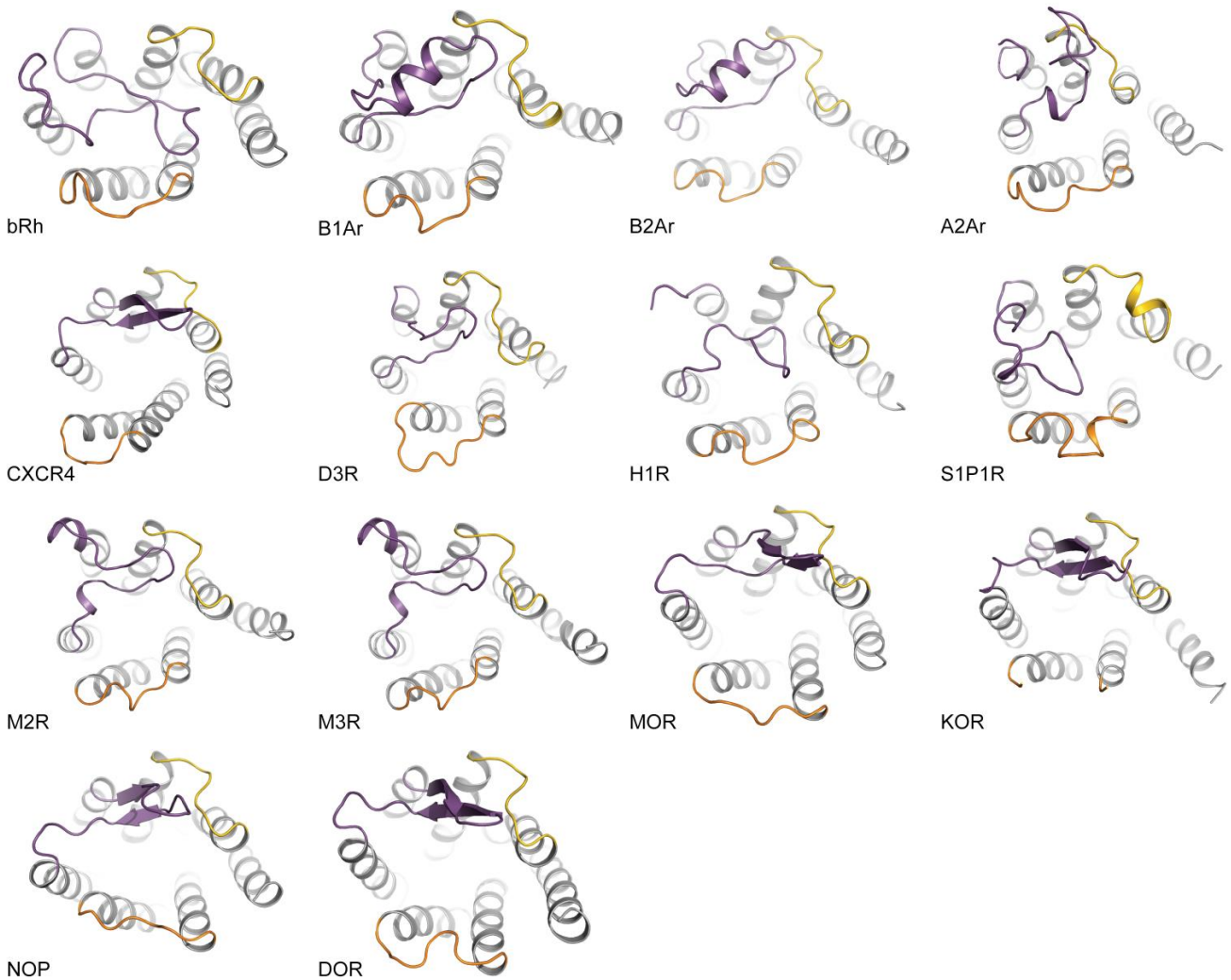


Figure 45
Structures of G protein-coupled receptors used in this study.
Experimental structures of the fourteen G protein-coupled receptors used in this study were obtained from the Protein Data Bank. Extracellular loop (ECL) 1 is shown in yellow, ECL2 in purple, and ECL3 in orange.

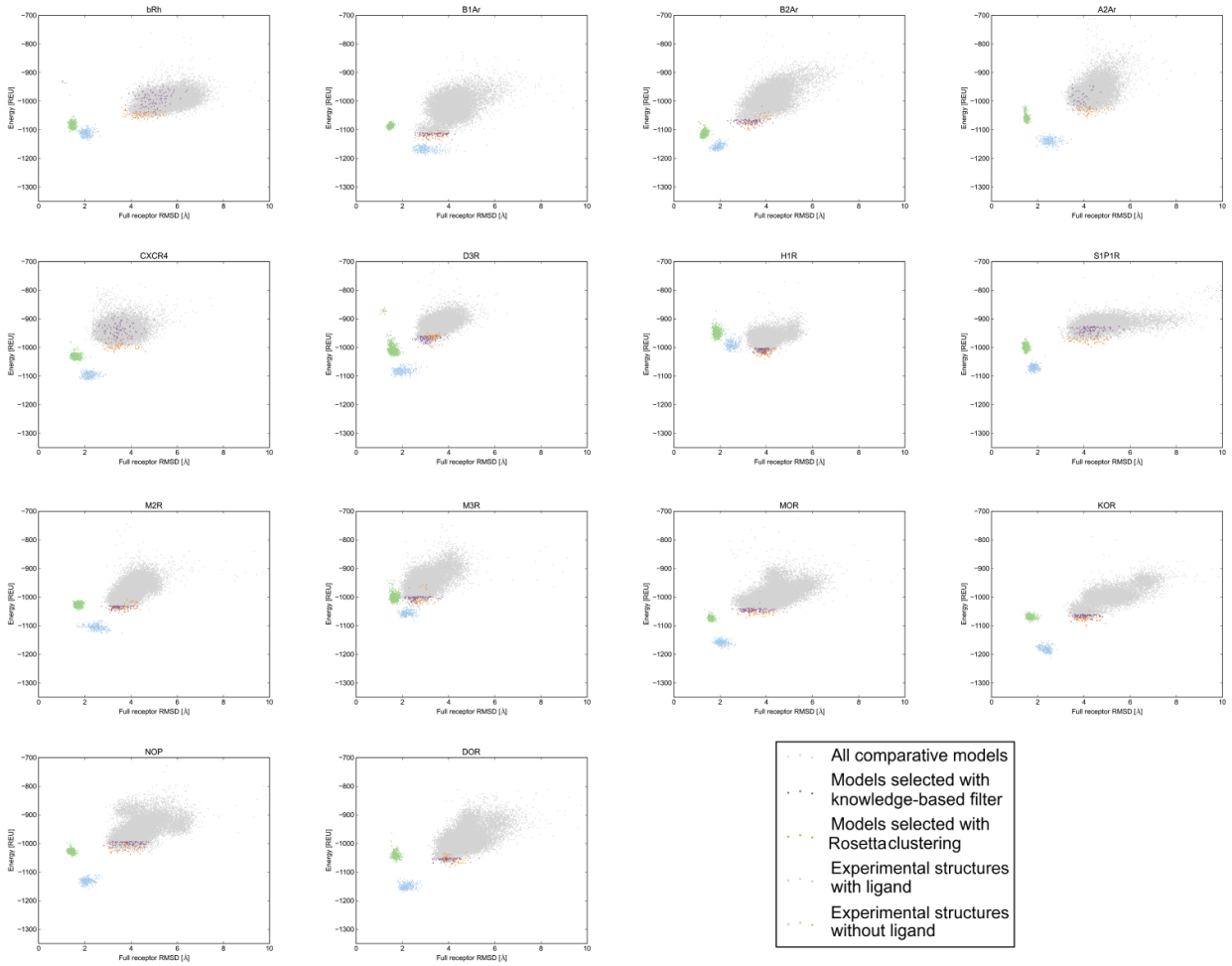


Figure 46
Energy plot of relaxed experimental structures and comparative models compared with full receptor RMSD.
For each structure, full receptor RMSD is plotted against total Rosetta energy. The experimental structure was minimized in the Rosetta force field without the ligand (in green) and with the ligand (in blue). Comparative models are in grey, with models selected through clustering in orange and models selected by the knowledge-based filter in purple.

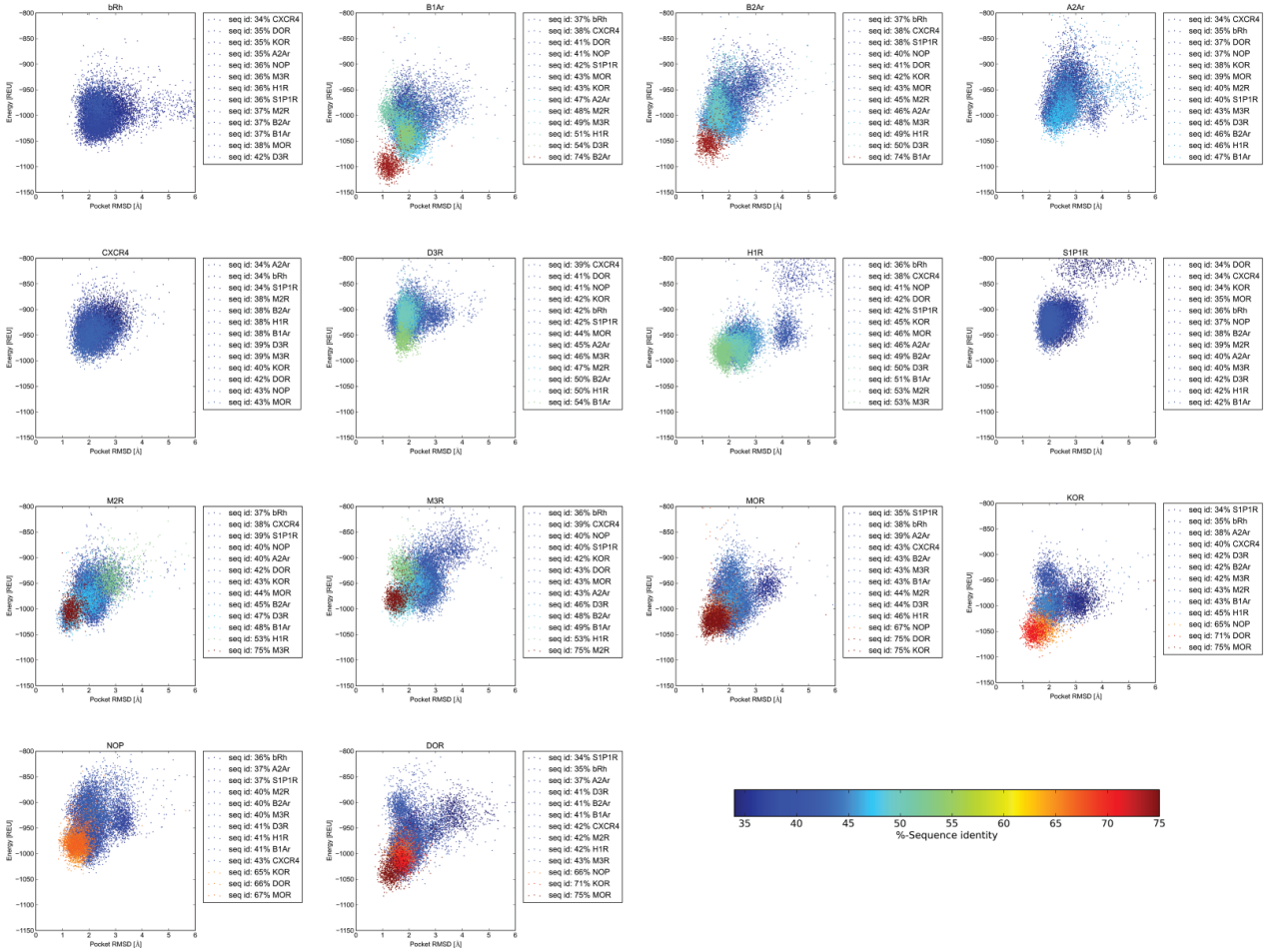


Figure 47
Energy plot of comparative models based on templates of varying sequence identity.
For each comparative model, pocket residue RMSD is plotted against total Rosetta energy. Each point is colored by the template by which the model was built, with color varying from blue to red with increasing sequence identity.

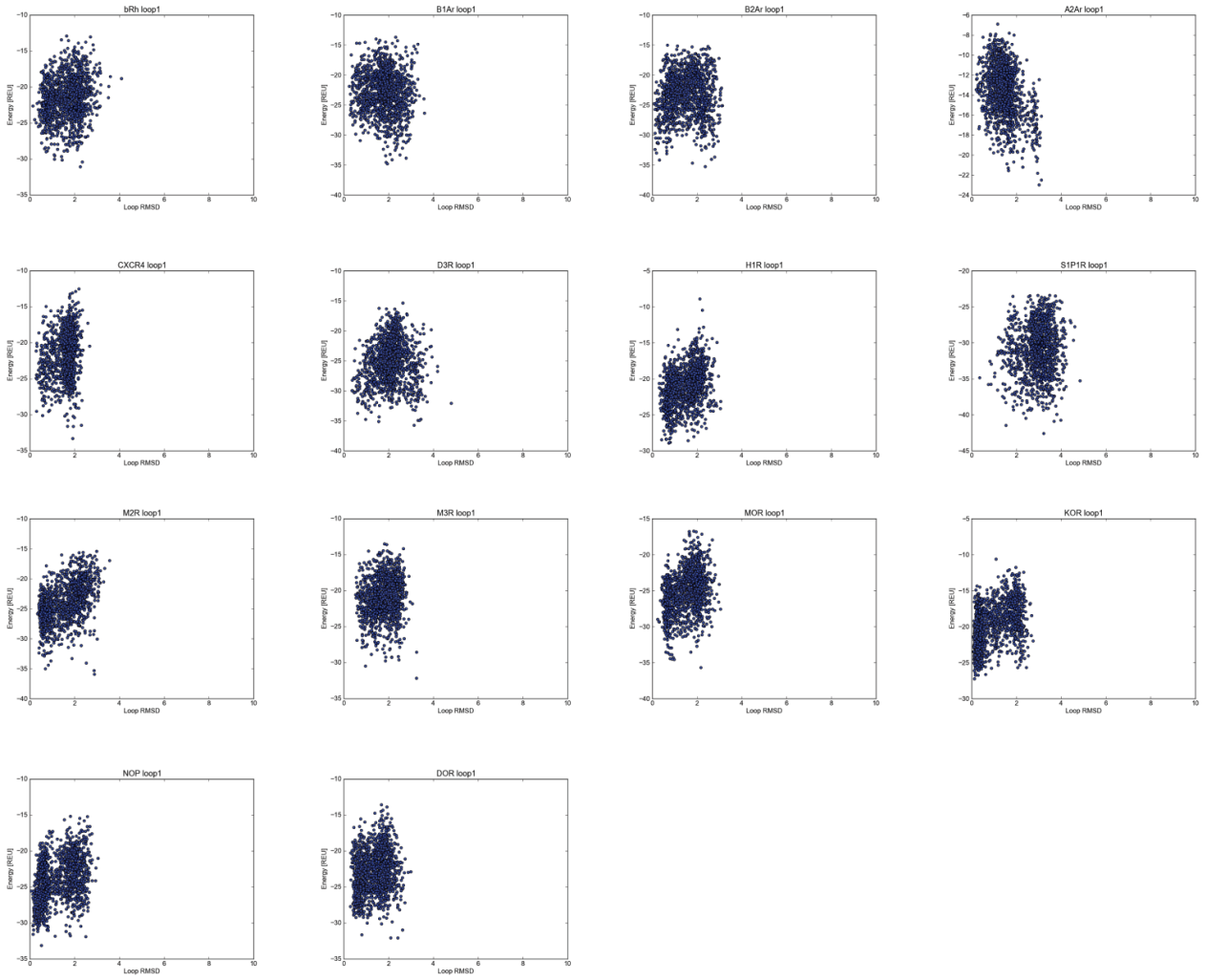


Figure 48
Energy plot of ECL1 in comparative models.
For each comparative model, ECL1 RMSD is plotted against the Rosetta energy for residues in ECL1.

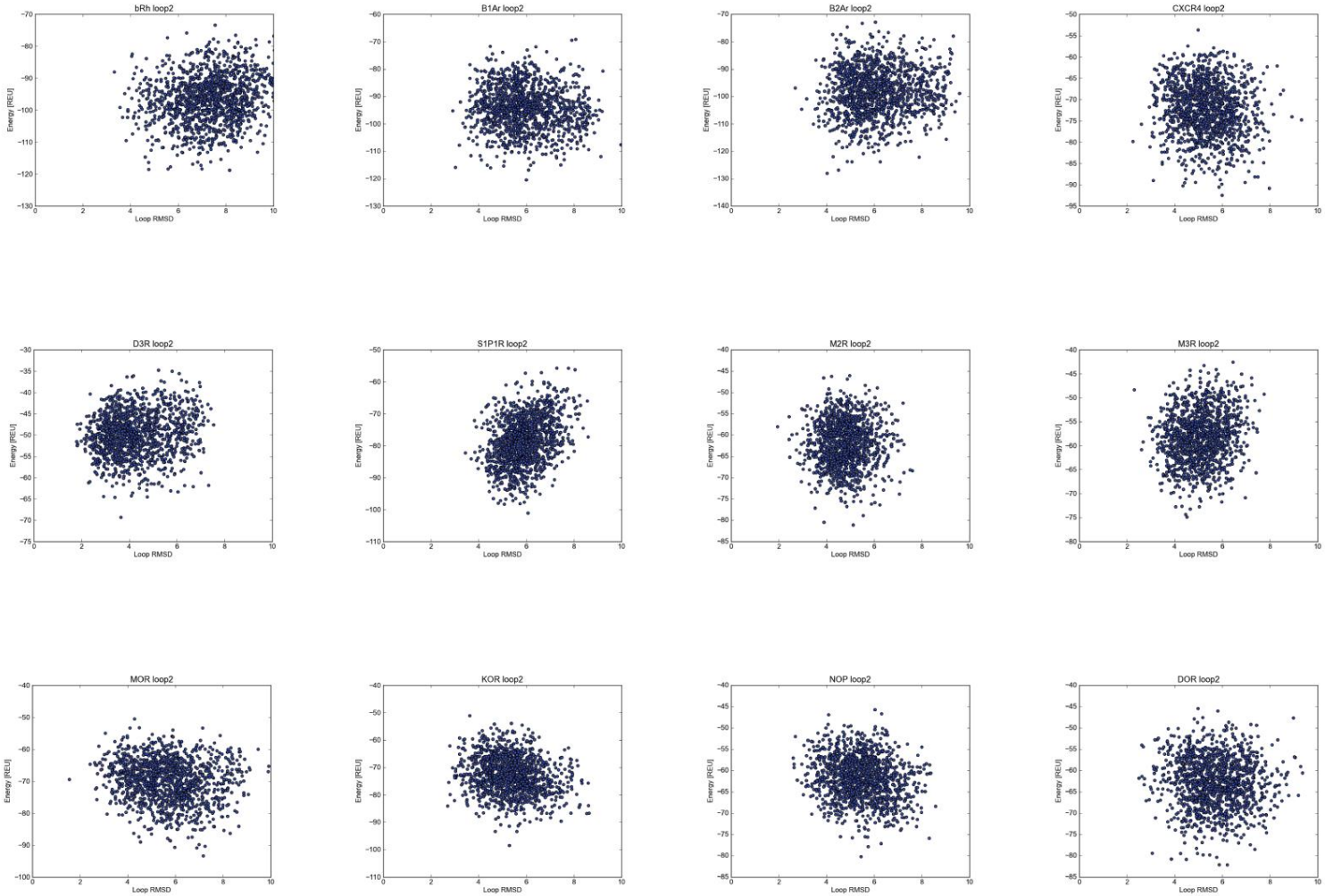


Figure 49
Energy plot of ECL2 in comparative models.
For each comparative model, ECL2 RMSD is plotted against the Rosetta energy for residues in ECL2. ECL2 for A2Ar and H1R could not be evaluated because of unresolved structure in this region of the experimental structure in the Protein Data Bank.

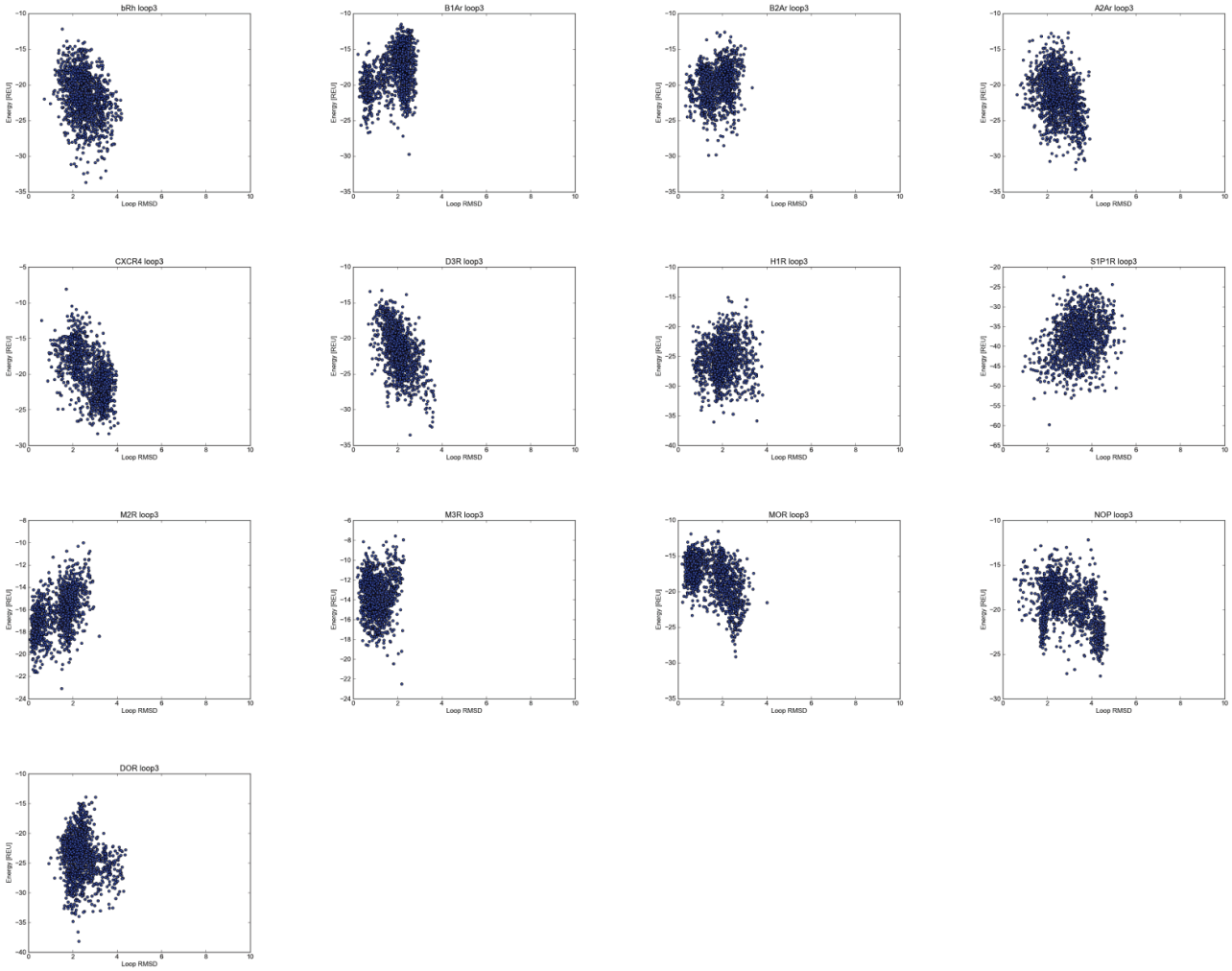


Figure 50
Energy plot of ECL3 in comparative models.
For each comparative model, ECL3 RMSD is plotted against the Rosetta energy for residues in ECL3. ECL3 for KOR could not be evaluated because of unresolved structure in this region of the experimental structure in the Protein Data Bank.

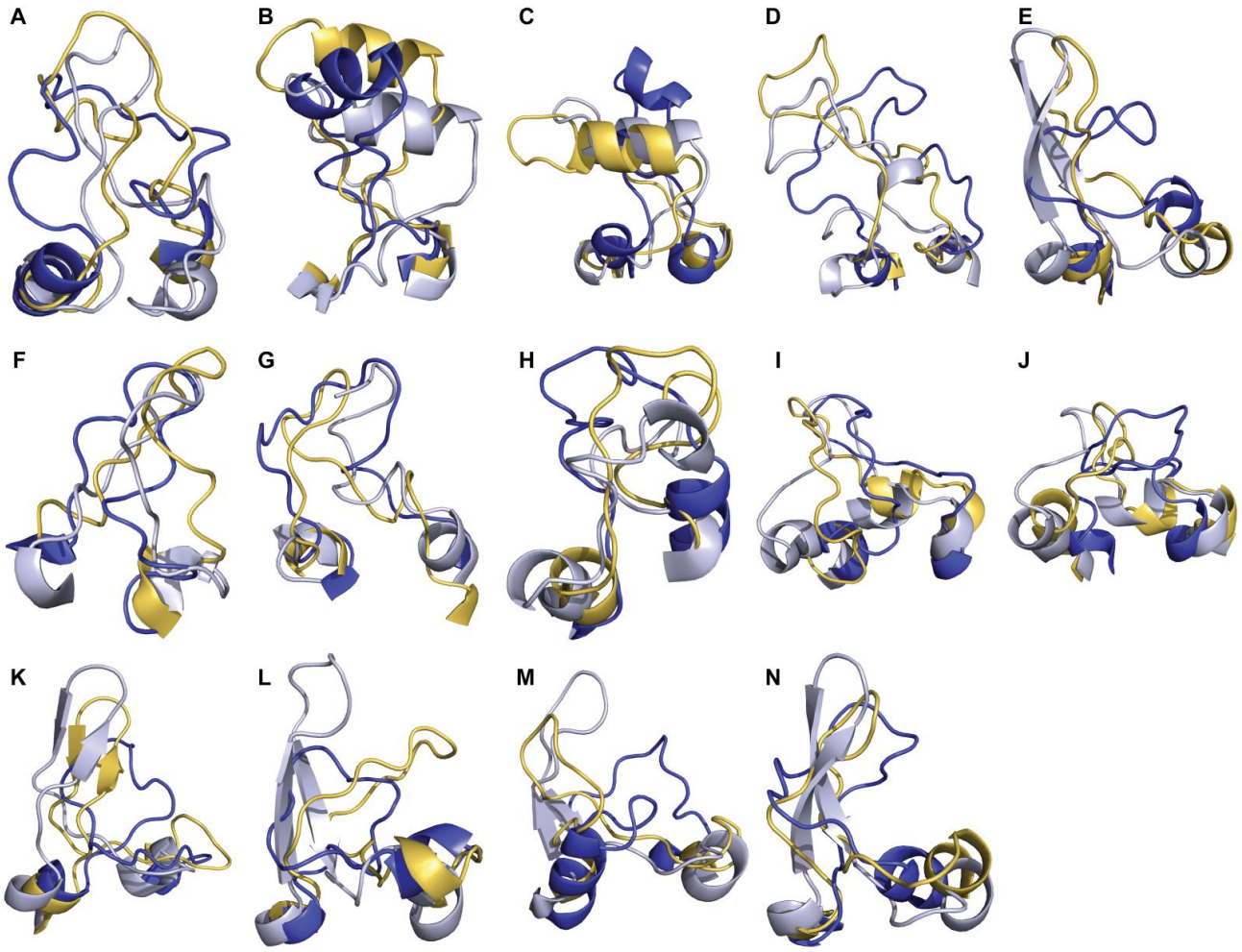


Figure 51
Structural representations of extracellular loop two from comparative models compared to experimental structures.
For A) bRh, B) B1Ar, C) B2Ar, D) A2Ar, E) CXCR4, F) D3R, G) H1R, H) S1P1R, I) M2R, J) M3R, K) MOR, L) KOR, M) NOP and N) DOR, the experimental structure is represented in gray, the most accurately sampled model is represented in yellow and the top ranked model is represented in blue. The top ranked model is the lowest energy model of the largest cluster, where clustering is performed on pairwise full receptor C-alpha RMSD over the top ten percent of comparative models by energy.

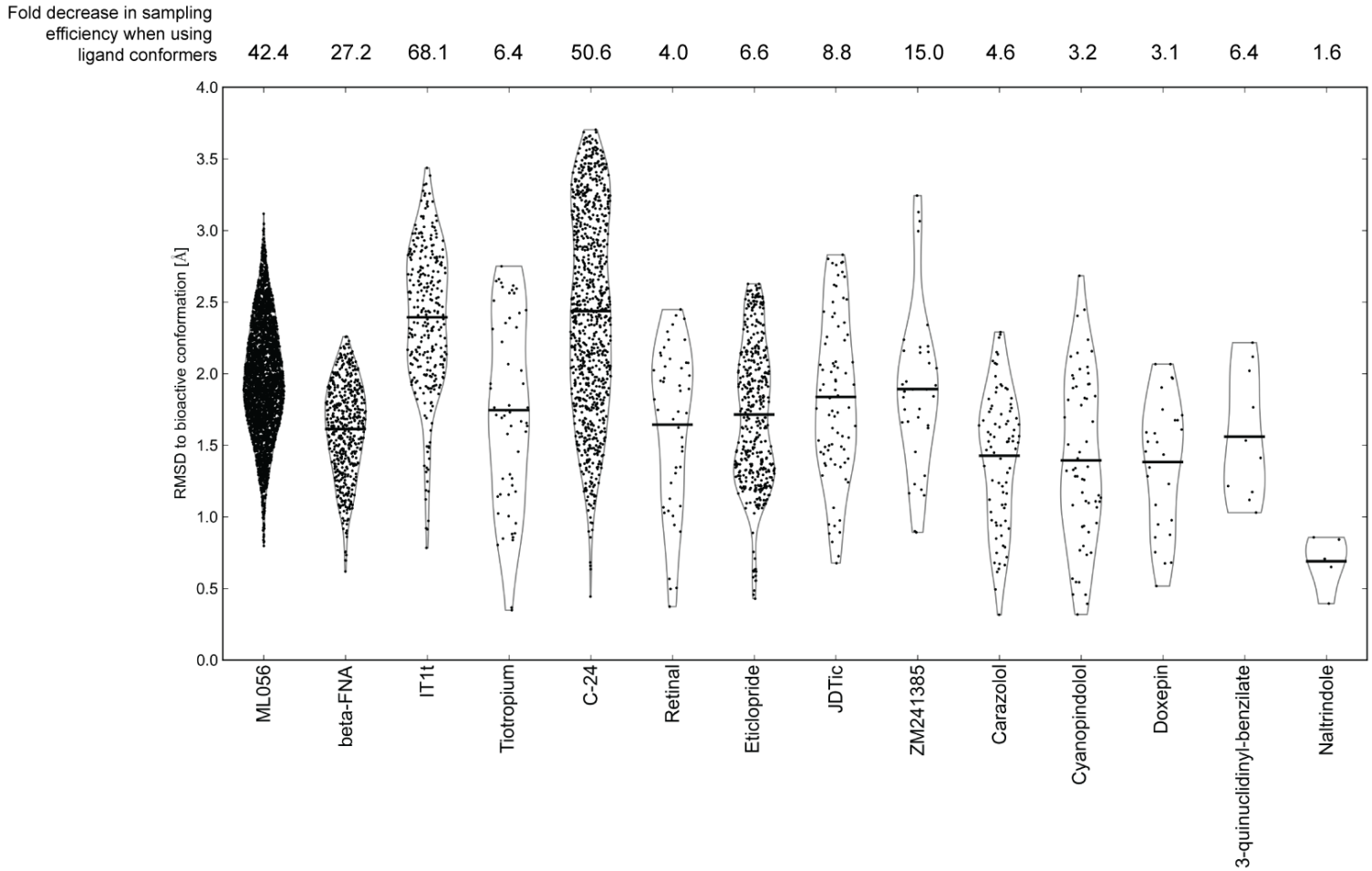


Figure 52
RMSD of ligand conformations generated by MOE.
Ligand conformations generated by MOE using the MMFF94x force field and Generalized Born solvation model were compared to the bioactive conformation found in the experimental structure by RMSD to heavy atoms in the ligand. The average RMSD is represented by a black line. The fold decrease in sampling efficiency is calculated by the uniform sampling efficiency within a 2.0 Å radius (USE2.0) for the bioactive ligand conformation divided by the uniform sampling efficiency within a 2.0 Å radius for ligand conformers.

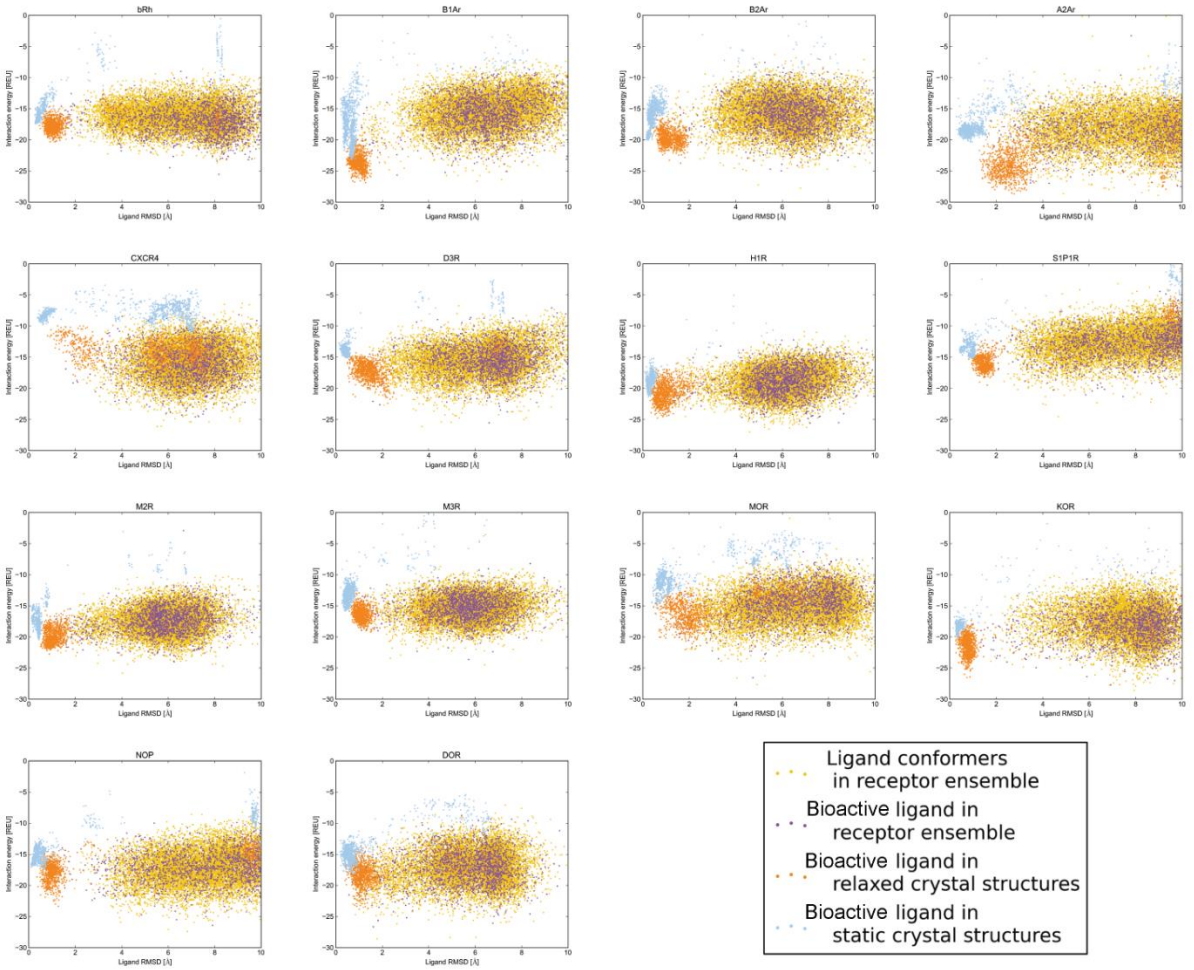


Figure 53
Interaction energy plot of binding modes from docking into experimental structures and comparative models.
For each structure, ligand heavy-atom RMSD is plotted against Rosetta interaction energy. The bioactive ligand conformation was docked into the static experimental structure (in blue), the energy minimized experimental structure (in orange) and comparative models (in purple). Ligand conformers generated by MOE were docked into comparative models, shown in yellow.

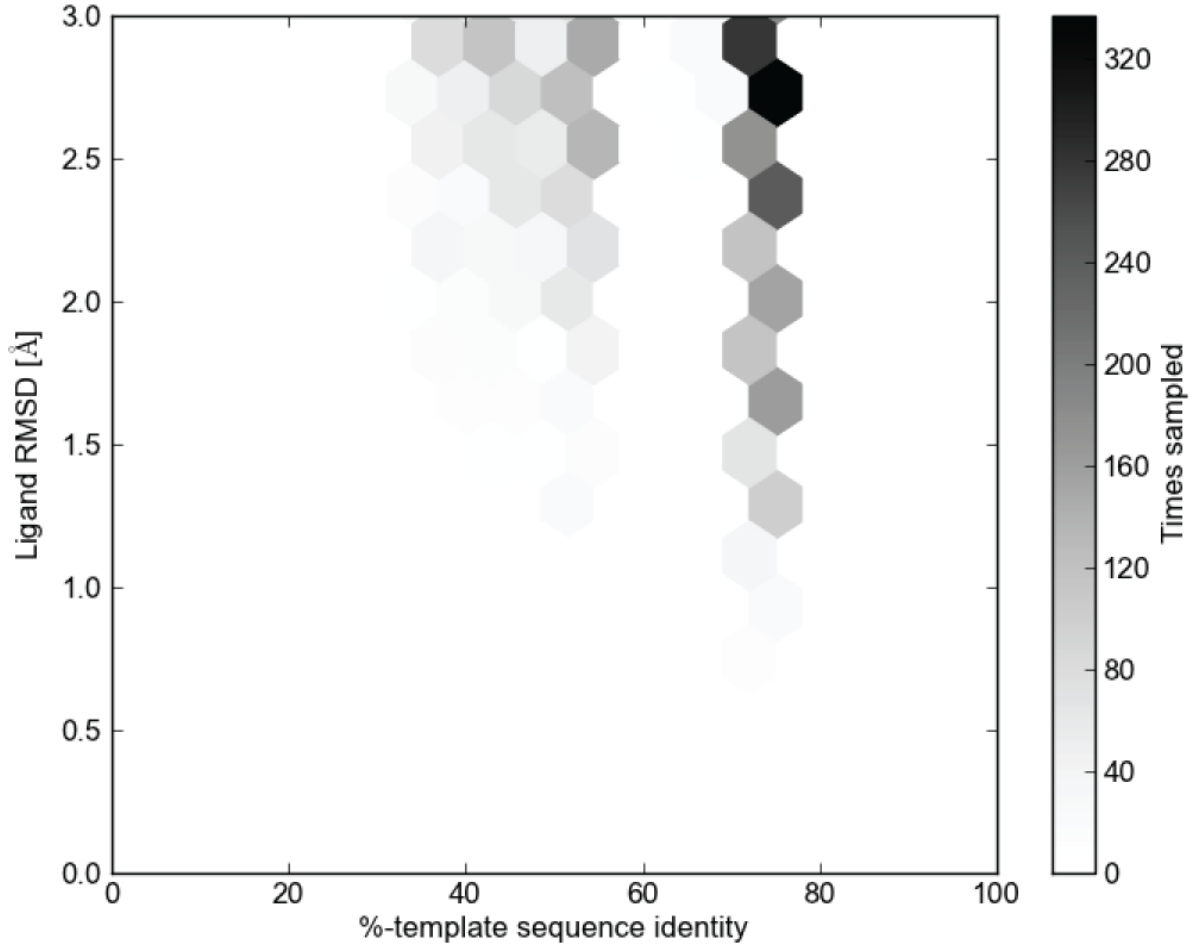
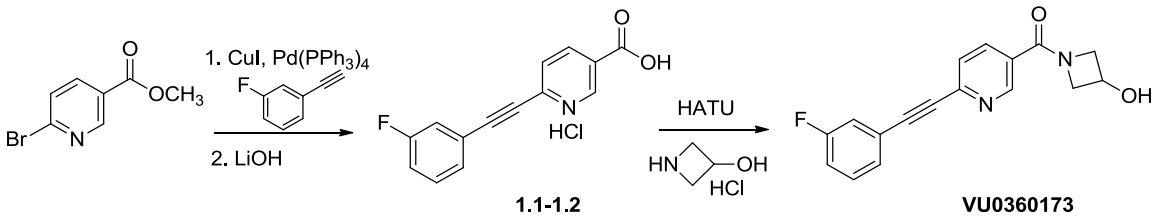


Figure 54
High sequence identity templates produce models with more accurate binding modes.
Each point represents the average ligand RMSD over all binding modes produced by docking the ligand into target GPCR comparative models built using a particular template. For each target-template pair, percent sequence identity was calculated on the sequence alignment shown in Figure 44. Sequence identity is shown here to correlate with low average ligand heavy-atom RMSD.

APPENDIX TO CHAPTER IV

VU0360173 and VU0403602 Synthesis

All NMR spectra were recorded on a Bruker 400 mHz instrument. ^{1}H chemical shifts are reported in δ values in ppm downfield from TMS as the internal standard in d_3-MeOH. Data are reported as follows: chemical shift, multiplicity (s = singlet, d = doublet, t = triplet, q = quartet, br = broad, m = multiplet), integration, coupling constant (Hz). Low resolution mass spectra were obtained on an Agilent 1200 series 6130 mass spectrometer. High resolution mass spectra were recorded on a Waters Q-TOF API-US. Analytical thin layer chromatography was performed on Analtech silica gel GF 250 micron plates. Analytical HPLC was performed on an HP1100 with UV detection at 214 and 254 nm along with ELSD detection, LC/MS (J-Sphere80-C18, 3.0 x 50 mm, 4.1 min gradient, 5%[0.05%TFA/CH_3CN]:95%[0.05%TFA/H_2O] to 100%[0.05%TFA/CH_3CN]. Preparative RP-HPLC purification was performed on a custom HP1100 automated purification system with collection triggered by mass detection or using a Gilson Inc. preparative UV-based system using a Phenomenex Luna C18 column (50 x 30 mm I.D., 5 μm) with an acetonitrile (unmodified)-water (0.1% TFA) custom gradient. Normal-phase silica gel preparative purification was performed using an automated Combi-flash companion from ISCO. Solvents for extraction, washing and chromatography were HPLC grade. All reagents were purchased from Aldrich Chemical Co. and were used without purification. All polymer-supported reagents were purchased from Argonaut Technologies and Biotage.

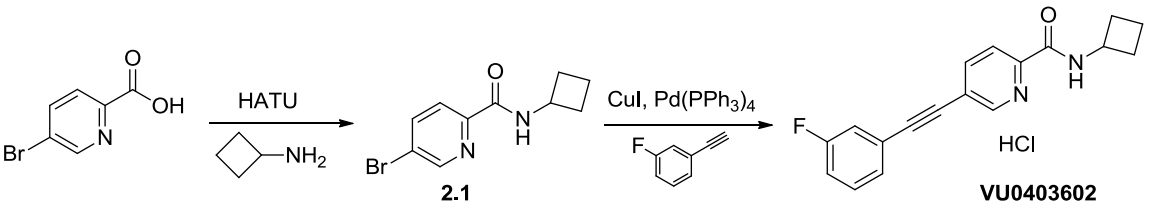


Methyl 6-((3-fluorophenyl)ethynyl)nicotinate (1.1): To a solution of methyl-6-bromonicotinate (1.5 g, 6.94 mmol) in DMF (15 mL) were added 1-ethynyl-3-fluorobenzene (0.97 mL, 8.3 mmol), tetrakis(triphenylphosphine) palladium (0) (0.49 g, 0.35 mmol), copper iodide (0.132 g, 0.69 mmol), and diethylamine (4.3 mL, 41.6 mmol). The mixture was subjected to microwave irradiation for 45 min at 90 °C. The mixture was passed through a celite pad and extracted between H_2O (100 mL) and EtOAc (100 mL x 3). The organic layers were combined, washed with brine, dried over $MgSO_4$, and concentrated under reduced pressure. The residue was purified by silica gel column chromatography (0-30% EtOAc/Hexanes) to afford methyl ester **1.1** (1.52 g, 86%) as a yellow solid.

6-((3-fluorophenyl)ethynyl)nicotinic acid hydrochloride (1.2): Methyl 6-((3-fluorophenyl)ethynyl)nicotinate (**1.1**, 1.50 g, 5.9 mmol) was dissolved in THF (15 mL), MeOH (3.5 mL) and H_2O (3.5 mL). To the suspension was added lithium hydroxide (0.99 g, 23.5 mmol). The reaction stirred at room temperature until complete conversion was observed by LC/MS. The reaction was acidified wit h HCl (2M) until pH ~ 2.0. The resulting precipitate was filtered and washed sequentially with cold water and ether to afford upon drying *in vacuo* overnight **1.2** (1.61 g, 99%) as an HCl salt: LC/MS (>98%) *m/z* = 242.1 [M+H].

(6-((3-fluorophenyl)ethynyl)pyridin-3-yl)(3-hydroxyazetidin-1-yl)methanone (VU0360173): To a solution of 6-((3-fluorophenyl)ethynyl)nicotinic acid hydrochloride (150 mg, 0.54 mmol) in DMF (2mL) were added azetidin-3-ol hydrochloride (89 mg, 0.81 mmol),

diisopropylethylamine (0.47 mL, 2.7 mmol), and HATU (246 mg, 0.65 mmol). The reaction stirred at room temperature until LC/MS indicated complete disappearance of starting material. The reaction was concentrated under reduced pressure and purified by silica gel column chromatography (50-100% EtOAc/Hexanes) to afford VU0360173 (125 mg, 78%) as a tan solid: ^{1}H NMR (400 MHz, $CDCl_3$) δ 8.87 (d, *J* = 1.6 Hz, 1H), 8.07-8.04 (dd, *J* = 6.0, 2.0 Hz, 1H), 7.64-7.62 (dd, *J* = 7.2, 0.8 Hz, 1H), 7.44-7.32 (m, 3H), 7.17-7.12 (m, 1H), 4.85-4.77 (m, 1H), 4.59-4.50 (m, 2H), 4.30-4.22 (m, 1H), 4.19-4.08 (m, 1H); LC/MS (>98%), 1.15 min, *m/z* = 297.0 [M+H]; HRMS = 297.1039 [M+H], calculated for $C_{17}H_{14}FN_2O_2$, 297.1039.



2.1 VU0403602

5-bromo-N-cyclobutylpicolinamide (2.1): To a solution of 5-bromopicolinic acid (400 mg, 2.0 mmol) in DMF (10 mL) were added cyclobutanamine (0.26 mL, 3.0 mmol), diisopropylethylamine (1.04 ml, 6.0 mmol), and HATU (0.99 g, 2.6 mmol). The reaction stirred at room temperature for 4h when LC/MS indicated full conversion. To the reaction was added H_2O (30 mL) and a white precipitate formed. The precipitate was filtered to afford 2.1 (366 mg, 72%) as a white solid.

N-cyclobutyl-5-((3-fluorophenyl)ethynyl)ethynyl)picolinamide hydrochloride (VU0403602): 5-bromo-N-cyclobutylpicolinamide (170 mg, 0.67 mmol) was combined in a microwave vial with 1-ethynyl-3-fluorobenzene (77 μL, 0.67 mmol), tetrakis(triphenylphosphine) palladium (0) (39 mg, 0.034 mmol), diethylamine (42 uL, 4.0 mmol), copper iodide (13 mg, 0.067 mmol), and DMF (15 mL). The reaction was subjected to microwave irradiation for 1h at 80 °C. The LC/MS indicated full conversion to desired product. The reaction was quenched with H_2O (45 mL) and extracted with EtOAc (30 mL

x 3). The combined organics were dried under $MgSO_4$, and concentrated under reduced pressure. The crude residue was purified by silica gel column chromatography (0-30% EtOAc/Hexanes). The pure isolated compound was dissolved in dioxane (3 mL) and 4M HCl in dioxane was added dropwise into the solution at room temperature until a pale yellow precipitate formed. The precipitate was filtered to afford the HCl salt of VU0403602 (192 mg, 87%) as an off-white powder: ^{1}H NMR (400 MHz, DMSO-d_6) δ 9.03 (d, J = 8.4, 1H), 8.83 (s, 1H), 8.19-8.16 (dd, J = 6.0, 2.0, 1H), 8.06 (d, J = 8.0, 1H), 7.56-7.48 (m, 3H), 7.38-7.33 (m, 1H), 4.51-4.43 (m, 1H), 2.22-2.16 (m, 4H), 1.70-1.64 (m, 2H); LC/MS (>98%), 3.78 min, *m/z* = 295.1 [M+H]; HRMS = 317.1066 [M+Na], calculated for $C_{18}H_{15}FN_2ONa$, 317.1066.

Protocol capture for protein modeling and ligand docking computational methods

This protocol capture contains the steps necessary to obtain the results presented in Chapter IV. The modeling and docking steps are found in separate folders and all files are labeled according to the associated steps below. The Rosetta 3.4 software suite is publically available and the license is free for non-commercial users at http://www.rosettacommons.org/.

1. Generation of an $mGlu_5$ comparative model

Step	Text	Commands	Comment
1A. Identify template for modeling	Based on its high similarity (e-value of 3e-15 with a sequence coverage of 90%) to $mGlu_5$ according to a search using NCBI BLASTP on sequences from the Protein Data Bank (PDB), the X-ray crystal structure human β_2-adrenergic receptor (Protein Data Bank ID: 2RH1) (Cherezov et al., 2007) was chosen as the template.	Input sequence from $mGlu_5$`.fasta` to http://blast.ncbi.nlm.nih.gov/Blast.cgi. Protein-protein BLAST was used with the PDB database.	Input: `A_`$mGlu_5$`.fasta` Output: Produced by Blast webserver
1B. Align sequences of family C 7TMRs	Members of the Family C 7TMRs, namely the human mGlus and Calcium-sensing receptor (CaSR) sequences, were first aligned with CLUSTALW.	Input sequences of mGlus and CaSR from `B_classC.fasta` to http://www.ebi.ac.uk/Tools/msa/clustalw2/. Slow alignment type with default settings was used.	Input: B_classC.fasta Output: B_classC.aln
1C. Align family C 7TMR sequences to template sequences	Alignment of TM regions between Family C 7TMRs and Family A crystal structure templates were directly adopted from Malherbe et al., 2006, with the exception of TM's 2, 4 and 7, which were based on the alignment of CaSR with Family A 7TMRs from Miedlich et al., 2004.	Manual alignment between Class C 7TMRs and a structural alignment of Class A 7TMRs was done by following the alignment from Malherbe et al., 2006 and Miedlich et al., 2004 papers. Visualization of the alignment through Jalview or Aline aids in the manual alignment.	Input: B_classC.aln C_classA.afasta Output: C_2rh1A.$mGlu_5$.aln

1D. Thread $mGlu_5$ sequence on β_2-adrenergic receptor backbone	In the construction of the comparative models, the backbone coordinates of the β_2-adrenergic receptor were retained in the comparative model of $mGlu_5$.	rosetta_tools/protein_tools/scripts/thread_pdb_from_alignment.py --template=2rh1A --target=$mGlu_5$ --chain=A --align_format=clustal C_2rh1A.$mGlu_5$.aln D_2rh1A.pdb D_$mGlu_5$_on_2rh1A.pdb	Input: C_2rh1A.$mGlu_5$.aln, D_2rh1A.pdb Output: D_$mGlu_5$_on_2rh1A.pdb
1E. Generate secondary structure prediction, constraint file and fragments for $mGlu_5$		Secondary structure- Jufo9D: http://meilerlab.org/index.php/servers/show?s_id=5 Secondary structure- PSIPRED: http://bioinf.cs.ucl.ac.uk/psipred/ Transmembrane span prediction based on Jufo9D: `perl bin/jufo9d_span.pl E_`$mGlu_5$`.jufo9d` Disulfide bond constraint file: Create file that lists residue number of cysteine residues predicted to disulfide bond in $mGlu_5$ predicted by alignment with Class A 7TMRs. Fragment files: http://www.robetta.org	Input: A_$mGlu_5$.fasta Output: E_$mGlu_5$.jufo9d, E_$mGlu_5$.psipred_ss2, E_$mGlu_5$_jufo9d.span, E_$mGlu_5$.disulfide, E_aamGlu509_05.200_v1_3, E_aamGlu503_05.200_v1_3
1F. Construct comparative model by rebuilding loop regions in Rosetta	In the construction of the comparative models, the backbone coordinates of the β_2-adrenergic receptor were retained in the comparative model of $mGlu_5$ while the loop coordinates were built in Rosetta using Monte Carlo Metropolis (MCM) fragment replacement combined with cyclic coordinate descent loop closure.	Generate loops file: The loop definitions span the region between transmembrane helices. List the residue numbers in the loop file as shown in `F_`$mGlu_5$`.loops`. Generate options file: List the desired options for rebuilding loop regions in an options file as shown in `F_loops.options`. Run loop building: `rosetta_source/bin/loopmodel.linuxgccrelease @ccd.options -database rosetta_database`	Input: D_$mGlu_5$_on_2rh1A.pdb, E_$mGlu_5$.jufo9d, E_$mGlu_5$.psipred_ss2, E_$mGlu_5$_jufo9d.span, E_$mGlu_5$.disulfide, E_aamGlu509_05.200_v1_3, E_aamGlu503_05.200_v1_3, F_loops.options, F_$mGlu_5$.loops Output: 5000 comparative models of $mGlu_5$ with loop regions rebuilt. F_$mGlu_5$_best_model.pdb

1G. Generate models with the S808A and T780A mutations in $mGlu_5$	To investigate the molecular cause of the PAM to NAM or neutral switches, modulators were docked into $mGlu_5$ models containing the mutation engendering the switch.	Prepare input files: Remove lines in `F_mGlu5_best_model.pdb` corresponding to the sidechain of residues 808 or 780. Change references of Ser or Thr to Ala for the remaining lines corresponding to the backbone. Generate options file for relax: List the desired options for relax (energy minimization) in an options file as shown in `G_relax.options`. Relax (energy minimization) using Rosetta: `rosetta_source/bin/relax.linuxgccrelease @relax.options -database rosetta_database`	Input: G_S808A_mGlu5.pdb, G_T780A_mGlu5.pdb, G_relax.options Output: 1000 models of mutants S808A or T780A $mGlu_5$. G_S808A_mGlu5_best_model.pdb, G_T780A_mGlu5_best_model.pdb

2. Docking of Allosteric Modulators

Note: The steps below use MPEP (shortened to "mpe") as an example. The same protocol was used on all seven ligands.

Step	Text	Commands	Comment
2A. Create ligand conformations in MOE	10 low energy conformations of the ligand created by MOE (Molecular Operating Environment, Chemical Computing Group, Ontario, Canada)	See MOE operating guide. LowModeMD with the MMFFx94 force field was used to generate conformations within 5 energy units from the minimum energy. The lowest 10 energy conformations were then saved as an .sdf file for conversion to .pdb and .params files for Rosetta. Convert .sdf file of ligand conformations to .pdb and .params file for Rosetta input: `rosetta_source/src/python/apps/public/molfile_to_params.py -n mpe -p mpe mpe.sdf` Combine all individual ligand conformations in pdb format to a file called `mpe_confs.pdb`.	Input: ligand coordinates in mol format Output: `A_mpe.params, A_mpe_confs.pdb`

2B. Generate input files necessary for docking in Rosetta	Each modulator was allowed to sample docking poses in a 5 Å radius centered at the putative binding site for MPEP, determined by the residues known to affect modulator affinity and/or function. Side-chain rotamers around the ligand were optimized simultaneously in a Monte-Carlo minimization algorithm. The energy function used during the docking procedure contains terms for van der Waals attractive and repulsive forces, hydrogen bonding, electrostatic interactions between pairs of amino acids, solvation, and a statistical term derived from the probability of observing a side-chain conformation in the PDB.	Prepare input pdb files: Copy one ligand conformation from the `mpe_confs.pdb` file to the bottom of the pdb file of the starting model, `F_`mGlu$_5$`_best_model.pdb`. Save as `B_mpe_`mGlu$_5$`.pdb`. Prepare options file for docking: List the desired options for docking in an options file as shown in `B_dock.options`. Prepare XML file for docking: List the desired specifications for docking in an options file as shown in `B_dock.xml`. Starting coordinates correspond to P654 (P82 in files renumbered starting with 1) in the starting PDB file.	Input: F_mGlu$_5$_best_model.pdb, A_mpe_confs.pdb Output: B_mpe_mGlu$_5$.pdb, B_dock.options, B_dock.xml
2C. Dock ligand within mGlu$_5$ comparative model	For each modulator, over 2,000 docked complexes were generated.	rosetta_source/bin/rosettascripts.default.linuxgccrelease @dock.options -database rosetta_database	Input: `B_mpe_`mGlu$_5$`.pdb, A_mpe.params, B_dock.xml, B_dock.options` Output: 2500 models of MPEP in low energy binding modes
2D. Analyze results by clustering binding modes	For each modulator, over 2,000 docked complexes were generated and clustered for structural similarity using bcl::Cluster (Alexander et al., 2011).	Generate table of pairwise RMSD values between binding modes, then use bcl::Cluster. Cluster using bcl::Cluster Download the bcl software suite at http://www.meilerlab.org/index.php/bclcommons/show/b_apps_id/4 (the license is free for non-commercial users). bcl.exe Cluster -distance_input_file mpe.cluster.mat -input_format TableLowerTriangle -output_format Rows Centers -output_file D_cluster_mpe -linkage Average -output_pymol 1000 5 100 10000 10 dendogram_mpe.py -remove_internally_similar_nodes 3 -pymol_label_output_string	Input: table of pairwise RMSD values between binding modes Output: `D_cluster_mpe.Centers, D_cluster_mpe.Rows`

2E. Identify most likely binding mode based on clustering and energy analysis	The lowest energy binding mode from the five largest clusters for each modulator were used for further analysis.	Sort models to identify lowest energy binding modes in the largest clusters.	Input: models from step 2C, D_cluster_mpe.Centers, D_cluster_mpe.Rows Output: E_mpe_mGlu$_5$_best_model.pdb all models from figures can be found in E_gregory_etal_2012_mGlu$_5$_pam_site.pse

Table 22
Potency and operational model parameters for glutamate stimulated Ca++ mobilization at HEK293A-mGlu5-wt and mutants in the absence and presence of MPEP.
Data represent mean ± S.E.M. from 3-6 independent determinations performed in duplicate.

	pEC_{50}[a]	$log\tau_A$[b]	n[c]	E_m[d]	basal[e]
R5-wt (poly)	6.49 ± 0.09	0.52 ± 0.04	3.7 ± 1.0	101.5 ± 5.3	0.6 ± 2.5
F585I	6.31 ± 0.09	0.37 ± 0.06	4.0 ± 0.7	109.6 ± 3.8	0.8 ± 0.7
R647A	6.11 ± 0.17	0.13 ± 0.12*	3.0 ± 0.4	167.0 ± 24.5	5.2 ± 3.1
I650A	6.44 ± 0.11	0.57 ± 0.05	3.7 ± 0.4	104.6 ± 1.3	1.5 ± 0.6
G651F	6.35 ± 0.10	0.37 ± 0.06	2.5 ± 0.3	114.4 ± 0.3	-3.0 ± 2.9
P654S	6.41 ± 0.03	0.46 ± 0.01	3.1 ± 0.3	100.1 ± 1.7	0.2 ± 1.7
P654F	6.48 ± 0.03	0.44 ± 0.02	3.2 ± 0.3	106.5 ± 7.9	3.3 ± 1.3
S657C	6.36 ± 0.07	0.41 ± 0.04	2.4 ± 0.3	111.2 ± 3.7	0.9 ± 1.2
Y658V	6.51 ± 0.08	0.52 ± 0.04	3.6 ± 0.8	103.8 ± 2.2	1.3 ± 1.5
P742S	5.90 ± 0.08*	0.19 ± 0.05*	5.0 ± 1.2	119.3 ± 13.1	0.6 ± 0.9
L743V	6.52 ± 0.11	0.52 ± 0.06	2.6 ± 0.4	95.7 ± 2.9	2.6 ± 1.3
N746A	6.55 ± 0.06	0.56 ± 0.04	3.8 ± 0.3	98.0 ± 1.8	1.5 ± 2.2
G747V	6.50 ± 0.10	0.50 ± 0.05	3.5 ± 0.8	102.3 ± 3.2	2.6 ± 0.4
T779A	6.63 ± 0.06	0.57 ± 0.03	2.3 ± 0.3	108.6 ± 1.8	1.9 ± 1.6
T780A	6.68 ± 0.02	0.58 ± 0.03	2.4 ± 0.6	104.1 ± 5.4	0.8 ± 0.5
W784A	6.15 ± 0.14	0.35 ± 0.05	3.7 ± 0.3	103.9 ± 1.3	2.9 ± 1.2
V788A	6.32 ± 0.06	0.38 ± 0.04	3.0 ± 0.6	107.4 ± 3.6	0.8 ± 0.8
F792A	6.64 ± 0.07	0.61 ± 0.09	2.5 ± 0.3	107.0 ± 6.9	-0.4 ± 0.9
S806A	6.66 ± 0.04	0.58 ± 0.05	2.0 ± 0.2	102.4 ± 4.5	3.2 ± 0.2
S808A	6.76 ± 0.08	0.71 ± 0.07	3.0 ± 0.5	101.8 ± 0.5	1.5 ± 0.8
S808T	6.72 ± 0.06	0.71 ± 0.06	2.8 ± 0.2	103.1 ± 1.6	1.0 ± 1.1
A809V	6.28 ± 0.23	0.38 ± 0.07	3.0 ± 0.6	110.1 ± 4.9	-1.4 ± 1.7
A809G	6.15 ± 0.14	0.31 ± 0.08	3.3 ± 0.4	115.2 ± 8.0	1.2 ± 1.1
T810A	6.40 ± 0.05	0.36 ± 0.04	2.1 ± 0.1	111.2 ± 4.1	1.5 ± 0.7
C815A	6.22 ± 0.06	0.36 ± 0.04	4.5 ± 0.5	100.5 ± 0.5	0.7 ± 0.6

* denotes significantly different from wild-type value, $p<0.05$, one-way ANOVA, Dunnett's post-test.

[a] negative logarithm of the concentration of agonist that elicits a half maximal response

[b] logarithm of the coupling efficiency of glutamate.

[c] the transducer function that links occupancy to response

[d] the maximal possible system response

[e] basal denotes the baseline level (ligand-independent) of the system response

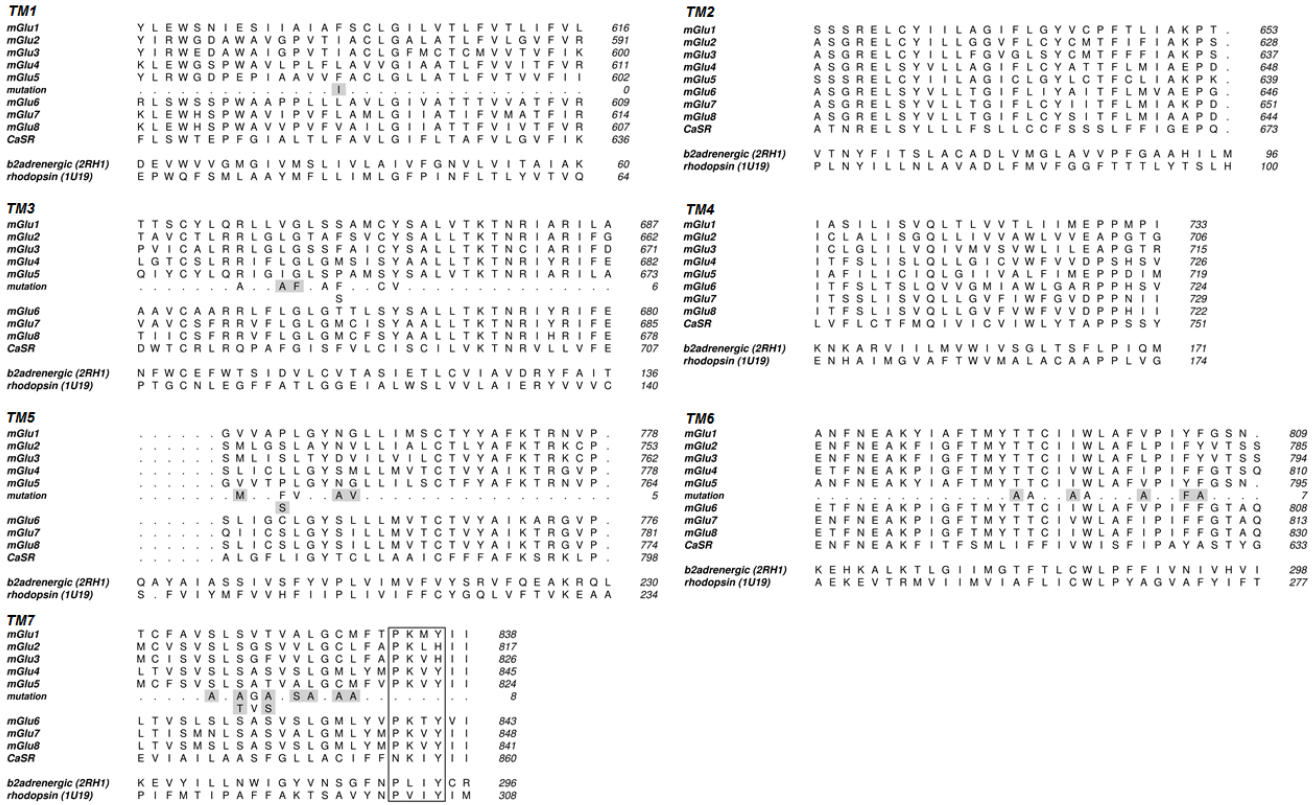


Figure 55

Alignment of the 7TMD of human mGlus with Calcium-sensing receptor, β2-adrenergic receptor and bovine rhodopsin sequences.

The 7TMD of the human mGlus and Calcium-sensing receptor (CaSR) sequences were aligned with CLUSTALW and are shown relative to β2-adrenergic receptor (2RH1) and bovine rhodopsin (1U19) sequences (aligned with MUSTANG). Alignment of TM regions between Family C 7TMRs and bovine rhodopsin were directly adopted from (26) except TM2, 4 and 7, which were based on the alignment of CaSR with bovine rhodopsin from (Miedlich et al., 2004). The PxxY alignment unique to this alignment of mGlu5 TM7 is outlined. All mutations investigated in this publication are indicated, with those novel to this work highlighted in gray.

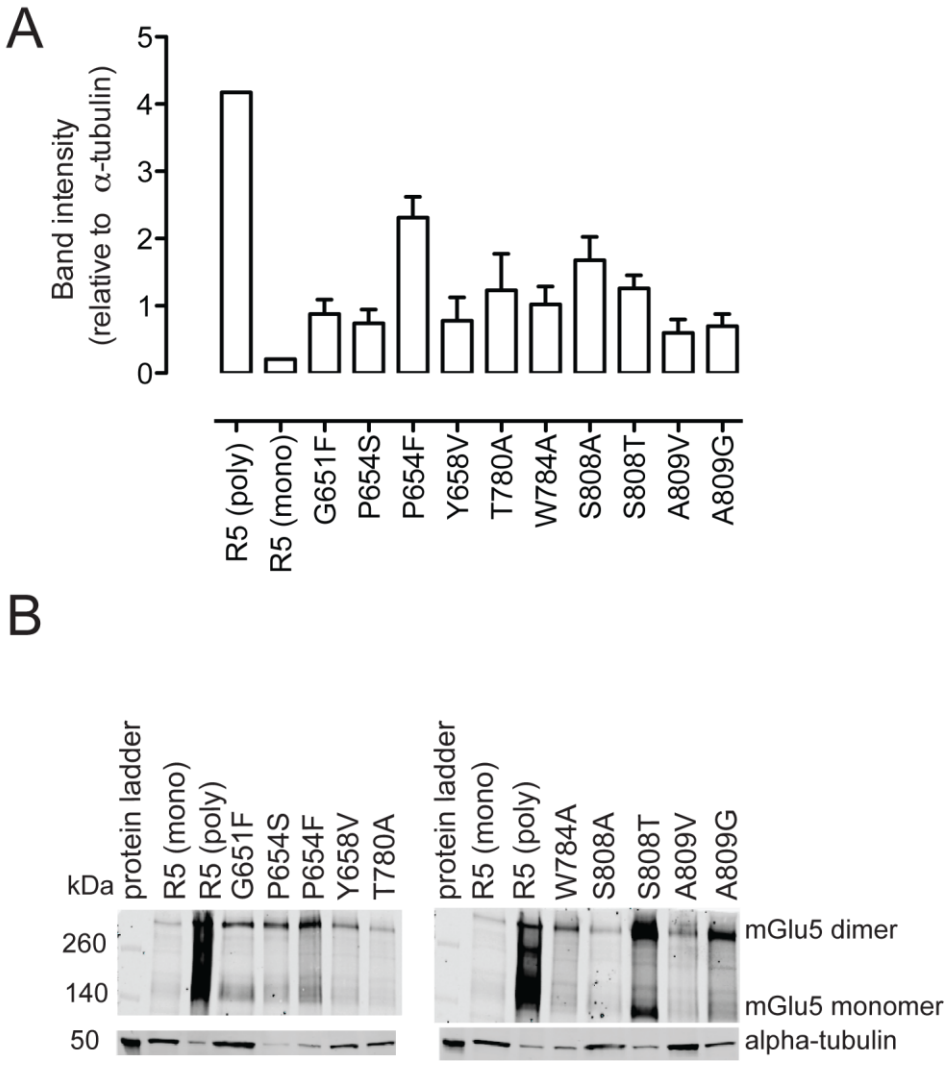


Figure 56
Immuno-blotting of non-binding point mutations of mGlu5.
A) Expression of mGlu5 in non-binding mutations was detected by immuno-blotting with rabbit polyclonal anti-mGlu5 (Chemicon) and imaged with donkey anti-rabbit-IR-680 (LI-COR) using an Odyssey imager. mGlu5 band intensity was normalized to the intensity for alpha-tubulin for the same sample to control for protein loading and is shown relative to that for the low-expressing mGlu5 wild-type (R5 mono) cell line and the higher-expressing polyclonal mGlu5 wild-type cell line (R5 poly). B) Representative blots for mGlu5 and alpha-tubulin are shown.

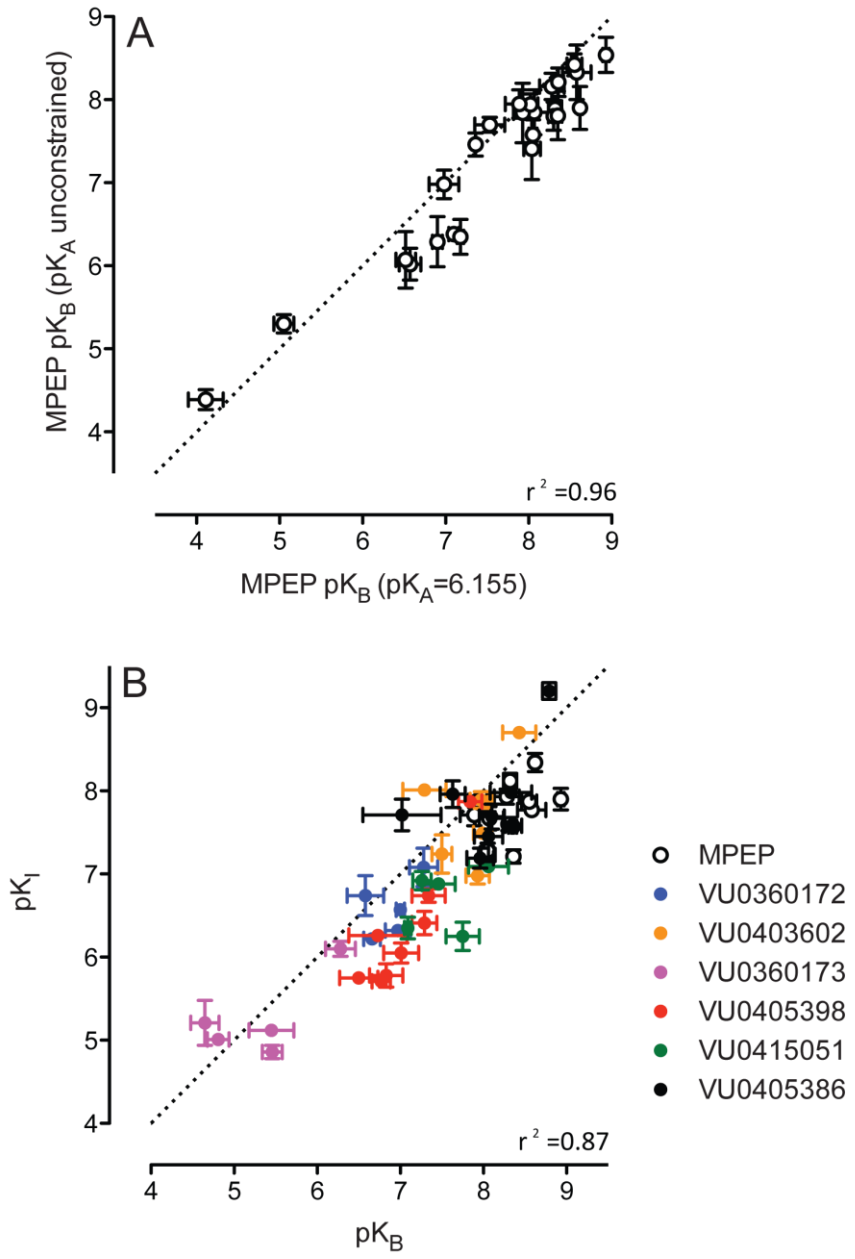


Figure 57
Affinity estimate correlations for mGlu5 allosteric modulators.
A) Constraining the affinity of glutamate to a previously determined estimate (pKA=6.155) had no effect on the affinity estimate (pKB) for MPEP derived from modulation of glutamate mediated calcium mobilization across all point mutations tested. B) Functional affinity estimates and those derived from radioligand binding inhibition assays (pKI) showed strong correlation. Data represent mean and s.e.m from a minimum of three independent determinations. The dashed line represents unity.

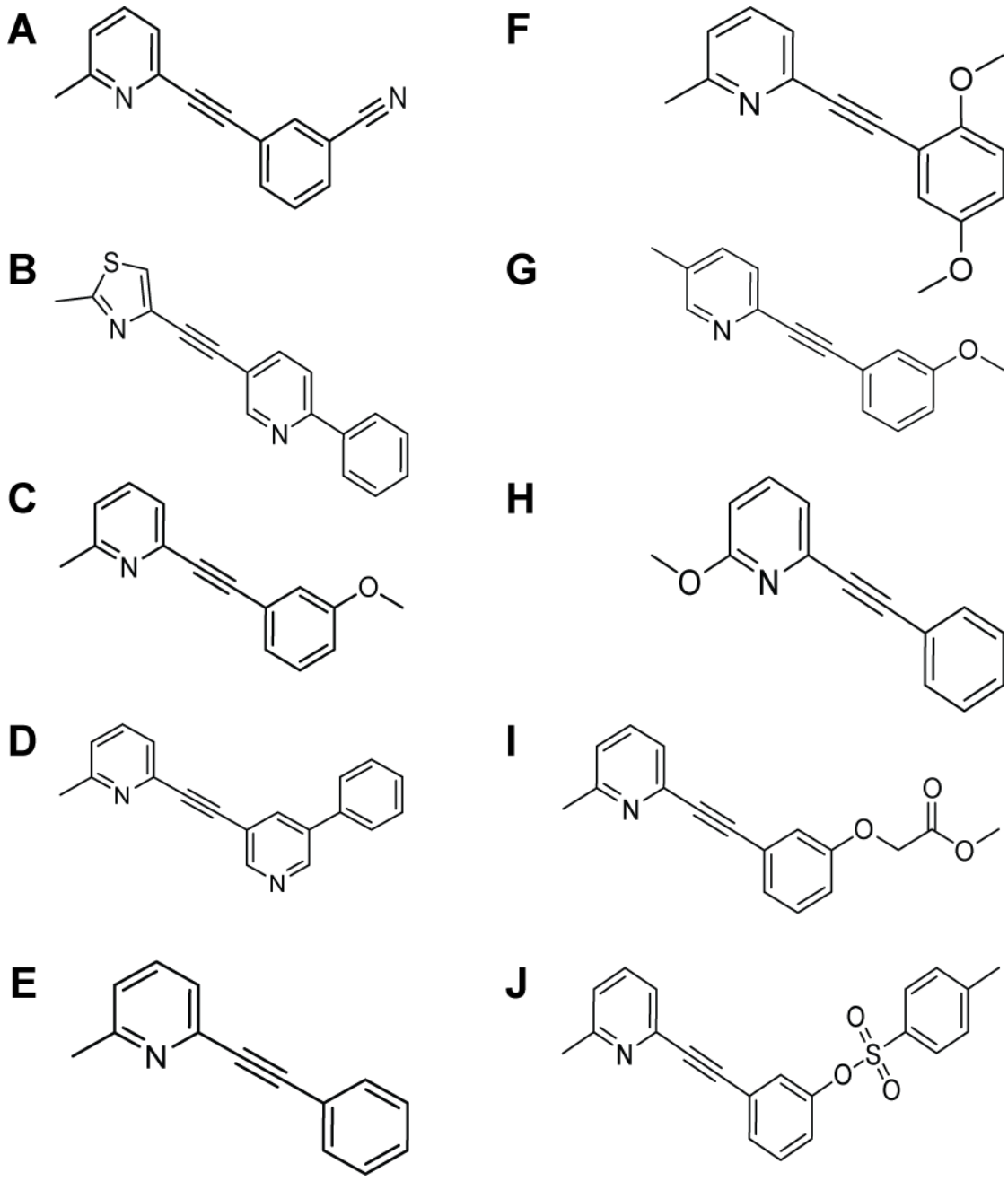


Figure 58
Ligands from the MPEP series.
Ligands in this figure correspond to ligands 1A through 1J in Table 12. Ligands H, I and J are considered to be non-functional compounds.

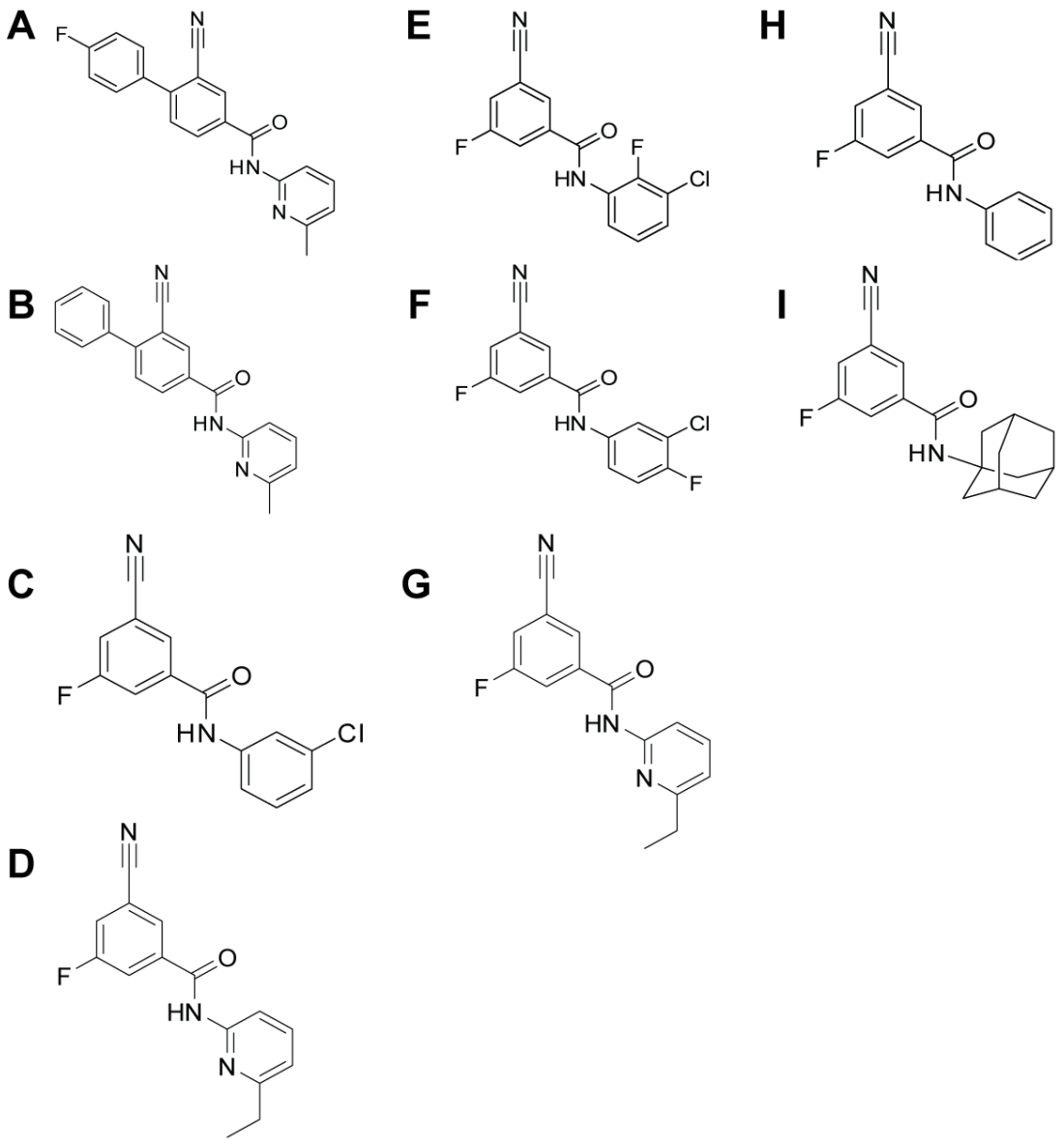


Figure 59
Ligands from the VU0366248 series.
Ligands in this figure correspond to ligands 2A through 2I in Table 12. Ligands H and I are considered to be non-functional compounds.

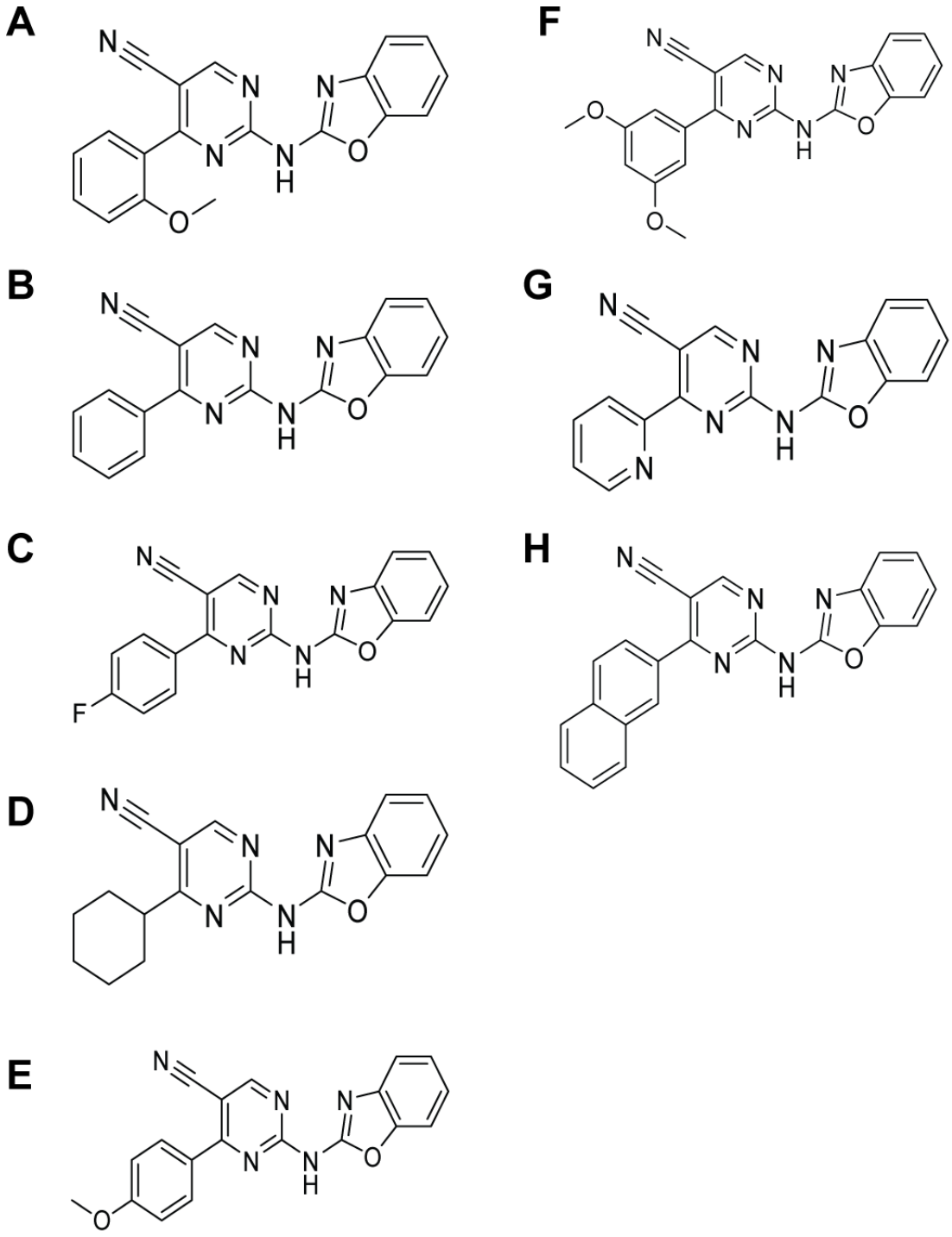


Figure 60
Ligands from the VU0366058 series.
Ligands in this figure correspond to ligands 3A through 3H in Table 12. Ligands F, G and H are considered to be non-functional compounds.

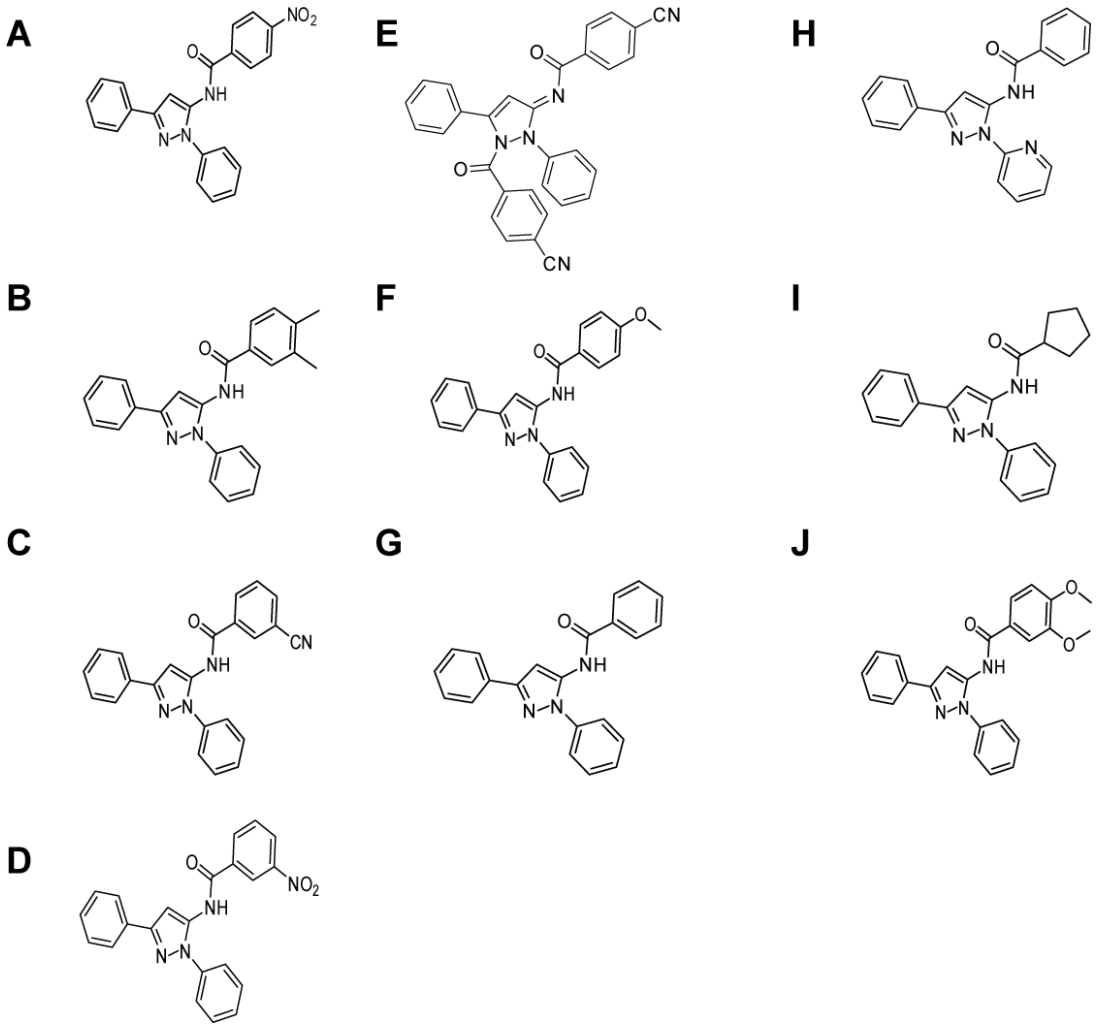


Figure 61
Ligands from the VU29 series.
Ligands in this figure correspond to ligands 4A through 4J in Table 12. Ligands H, I and J are considered to be non-functional compounds.

Protocol capture for protein modeling and ligand docking computational methods

Refer to the command lines, input and output files in the protocol capture for Chapter IV for the computational modeling and ligand docking aspects of the work in Chapter V. Small differences in methods between the two chapters are indicated in the text.

APPENDIX TO CHAPTER VI

Single cysteine mutations

Intracellular Ca^{2+} mobilization studies

Table 23
PAM potency (pEC50) for delayed (100-120 sec post add) response at headless mGlu5 C781V construct.
Data represents 1 experiment performed in triplicate.

	VU0404211	VU0415051	VU0405386	CPPHA
Wild-type	5.46	7.66	8.42	4.87
C781V	5.50	7.75	7.21	5.21

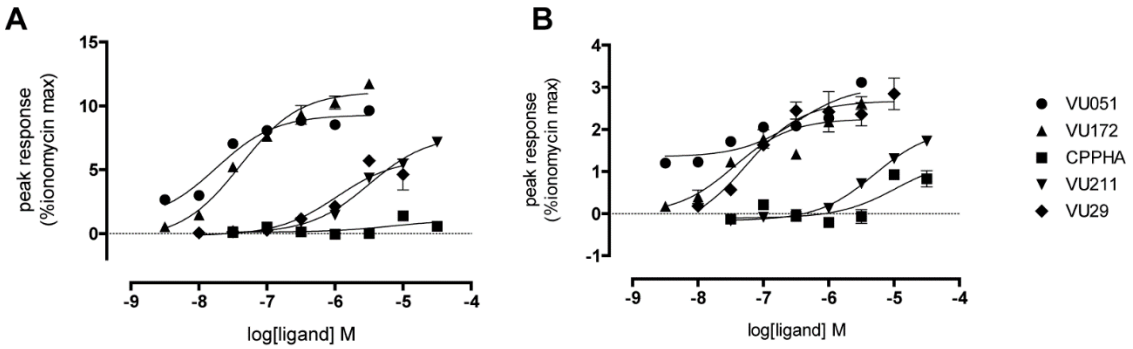


Figure 62
Effect of single point mutation C781V of mGlu5 on potentiation by PAMs of mGlu5-mediated Ca2+ mobilization.
A) Peak response, B) Plateau. Data represents the results from n=1.

Full characterization of 11-cysteine and 12-cysteine mutants

Intracellular Ca^{2+} mobilization studies

Table 24
PAM potency (pEC50) for delayed (100-120 sec post add) response at headless mGlu5 10+C815A construct.
Data represents 1 experiment performed in triplicate.

	VU0357121	VU0360173	VU120	VU0364289	CDPPB
Wild-type	5.80	5.58	5.80	5.55	6.86
10+C815A	4.19	3.98	4.59	3.19	6.46

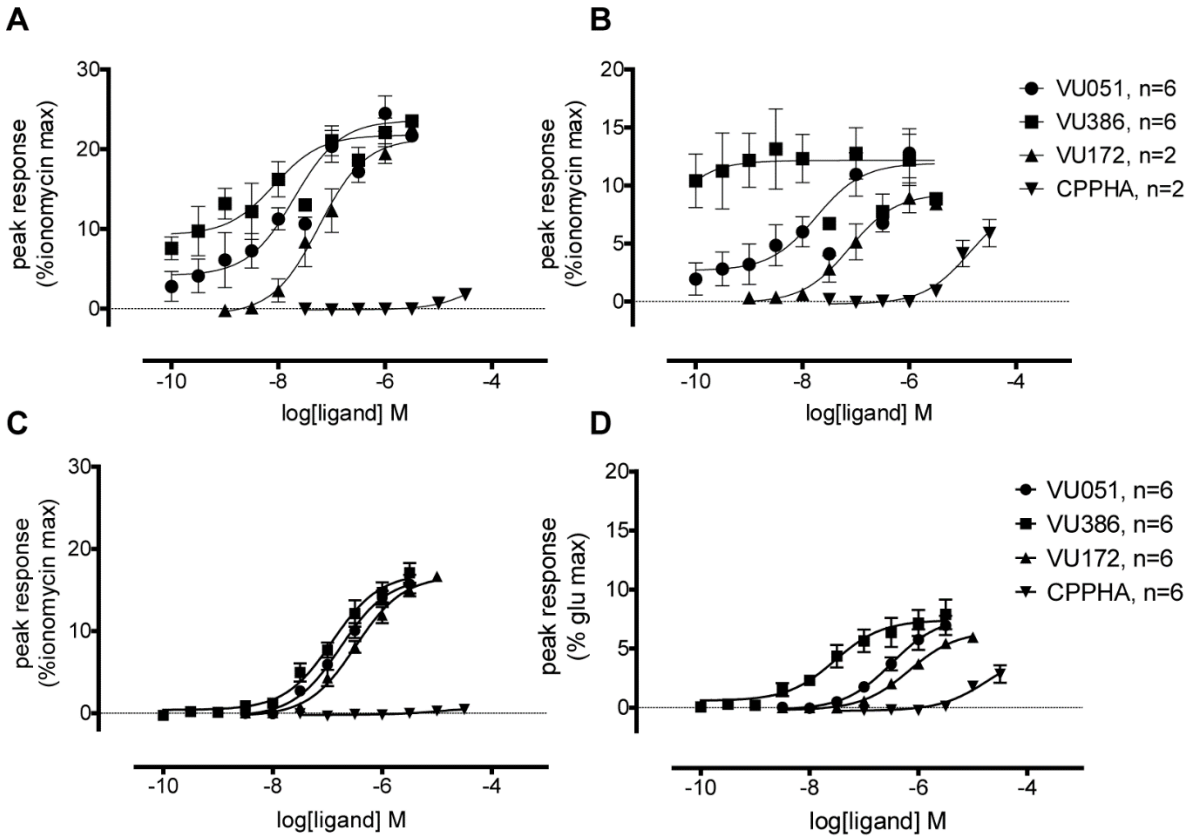


Figure 63
Effect of 11 cysteine mutations (10+C815A) of mGlu5 on potentiation by VU0415051, VU0405386, VU0360172 and CPPHA of mGlu5-mediated Ca2+ mobilization compared to headless mGlu5 wild-type.
Data represents the mean ± S.E.M. from 2 to 6 independent determinations for mGlu5 headless WT A) peak and B) plateau response and mGlu5 headless 10+C815A mutant C) peak and D) plateau response.

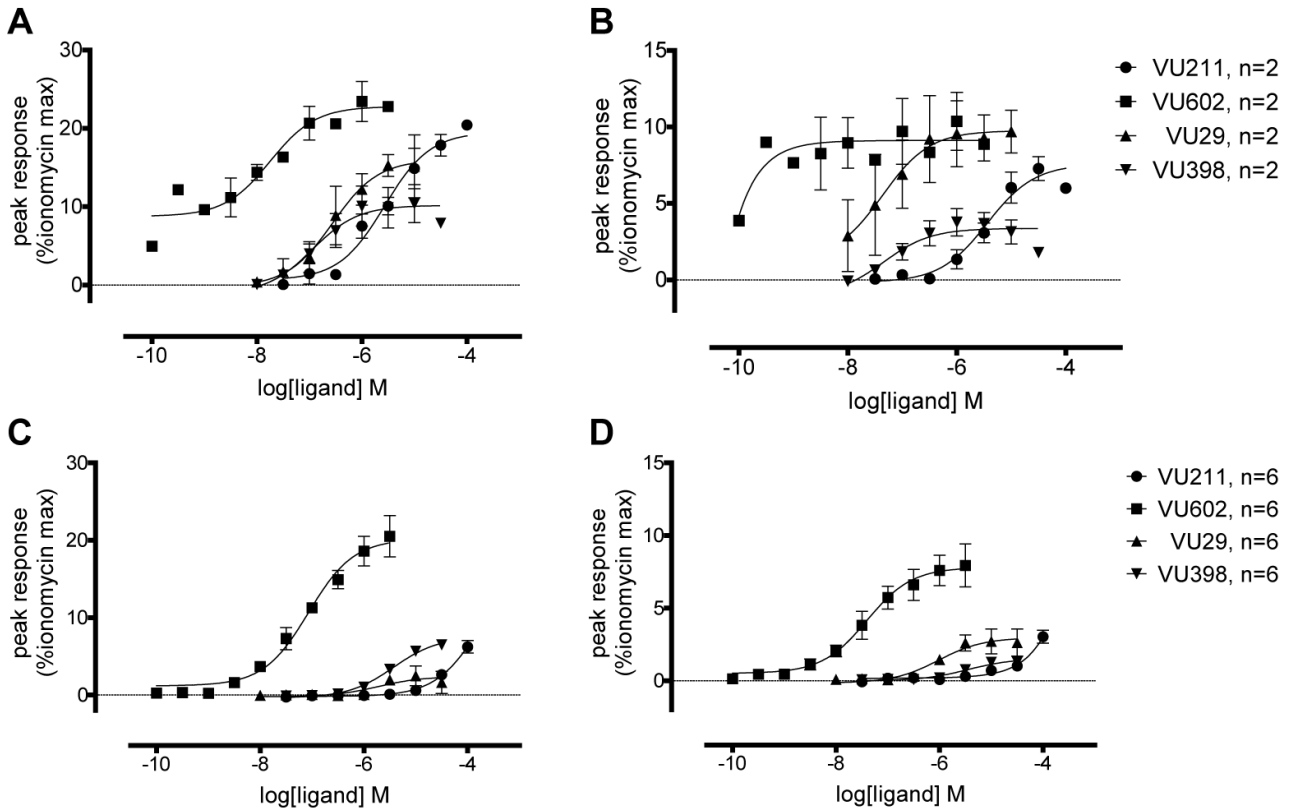


Figure 64
Effect of 11 cysteine mutations (10+C815A) of mGlu5 on potentiation by VU0404211, VU0403602, VU29 and VU0405398 of mGlu5-mediated Ca2+ mobilization compared to headless mGlu5 wild-type.
Data represents the mean ± S.E.M. from 2 to 6 independent determinations for mGlu5 headless WT A) peak and B) plateau response and mGlu5 headless 10+C815A mutant C) peak and D) plateau response.

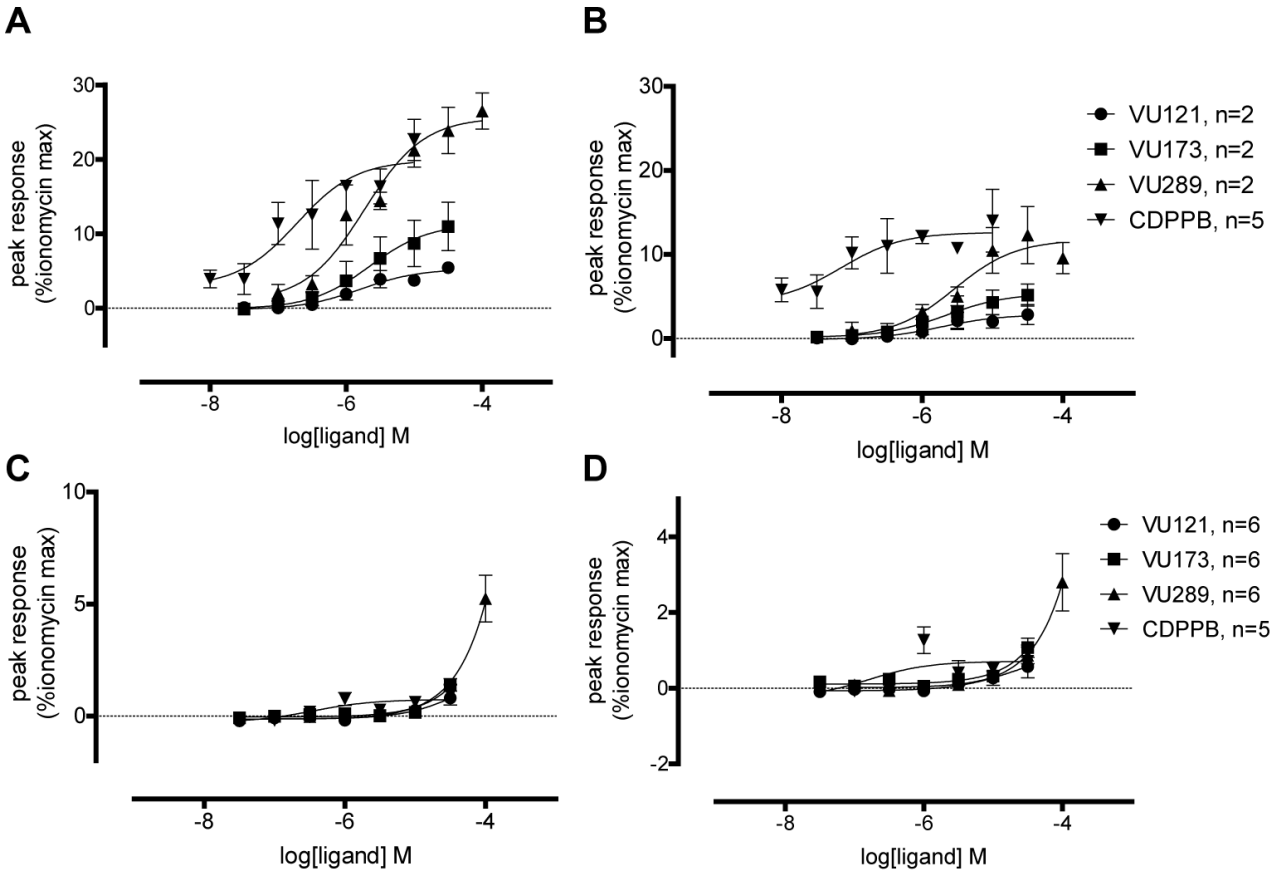


Figure 65
Effect of 11 cysteine mutations (10+C815A) of mGlu5 on potentiation by VU0357121, VU0360173, VU0364289 and CDPPB of mGlu5-mediated Ca2+ mobilization compared to headless mGlu5 wild-type.
Data represents the mean ± S.E.M. from 2 to 6 independent determinations for mGlu5 headless WT A) peak and B) plateau response and mGlu5 headless 10+C815A mutant C) peak and D) plateau response.

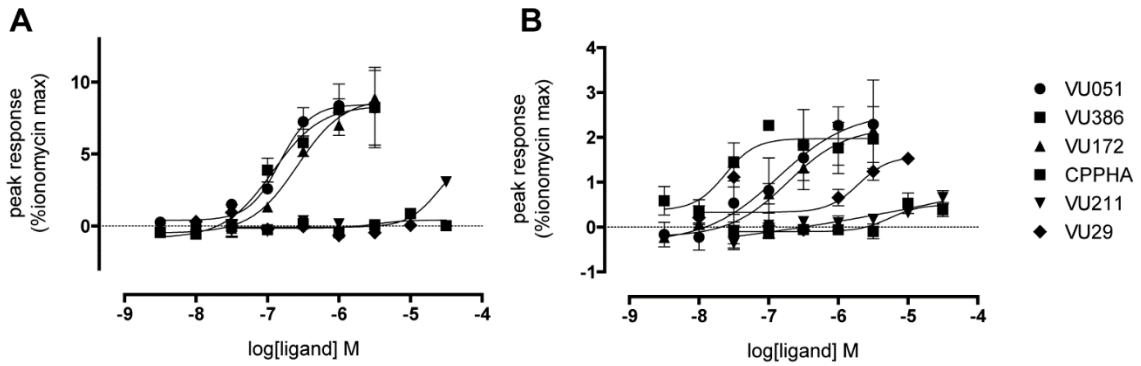


Figure 66
Effect of mutating out all cysteine residues from headless mGlu5 on potentiation by PAMs of mGlu5-mediated Ca2+ mobilization.
A) Peak response, B) Plateau. Data represents the results from n=1.

Saturation binding studies

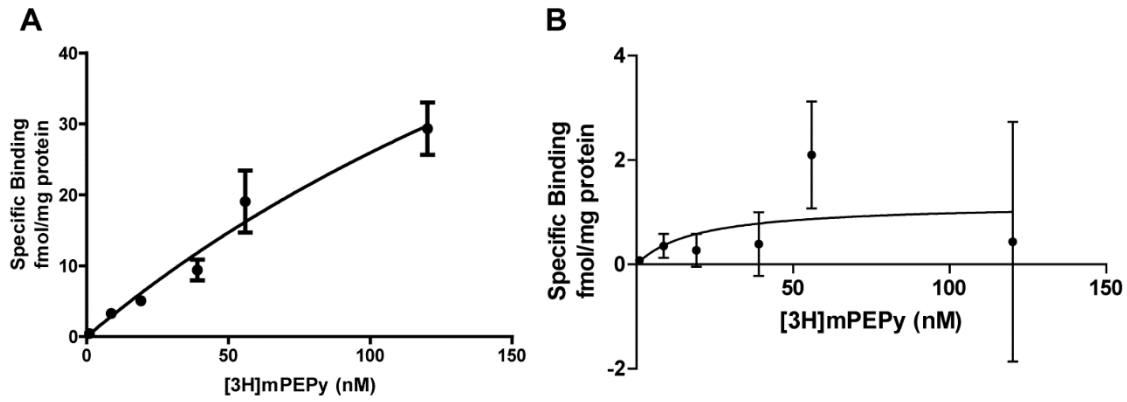


Figure 67
Concentration depending binding demonstrated by six point saturation binding curves
A) wild-type headless mGlu5 (at 500ug) and B) 10+C815A mGlu5 construct (at 1000ug) expressing HEK293A cell membranes. Data is the result of n=1.

$mGlu_5$ sensitivity to MTS reagents

Determining optimal concentrations of MTS reagents for use with $mGlu_5$

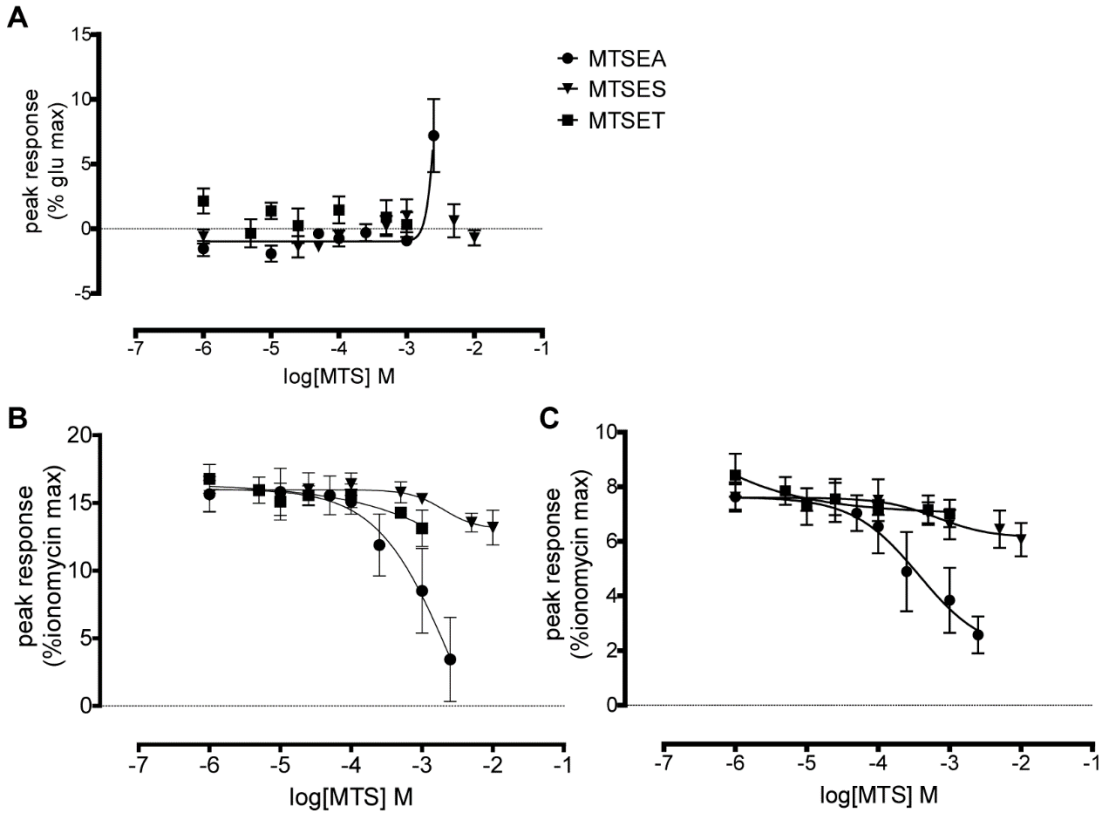


Figure 68
Effect of varying concentrations of MTS reagents on full and headless mGlu5 WT receptors on PAM potentiation of mGlu5-mediated Ca2+ mobilization by VU0360172.
A) Maximal glutamate response in full mGlu5 WT receptors, B) Peak response in headless mGlu5 WT receptors C) Plateau response in headless mGlu5 WT receptors. Data represents the mean ± S.E.M. of 3 experiments performed in triplicate.

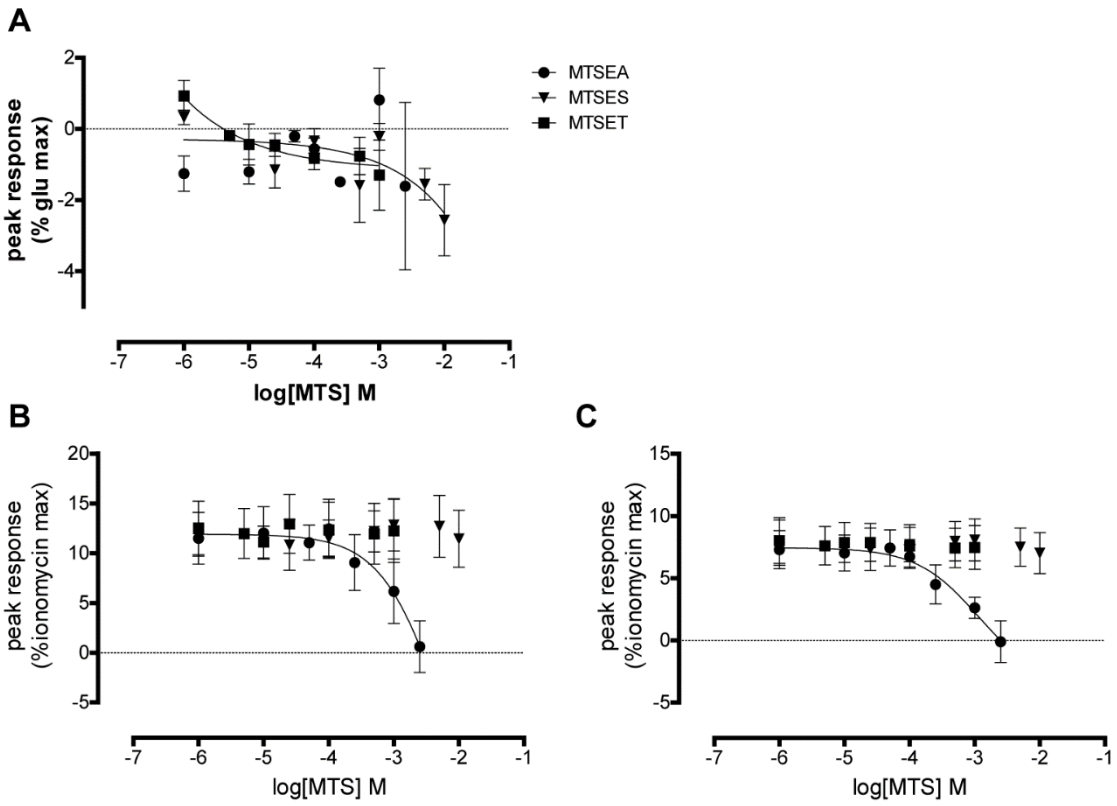


Figure 69

Effect of varying concentrations of MTS reagents on full and headless mGlu5 WT receptors on PAM potentiation of mGlu5-mediated Ca2+ mobilization by VU0415051.

A) Maximal glutamate response in full mGlu5 WT receptors, B) Peak response in headless mGlu5 WT receptors C) Plateau response in headless mGlu5 WT receptors. Data represents the mean ± S.E.M. of 3 experiments performed in triplicate.

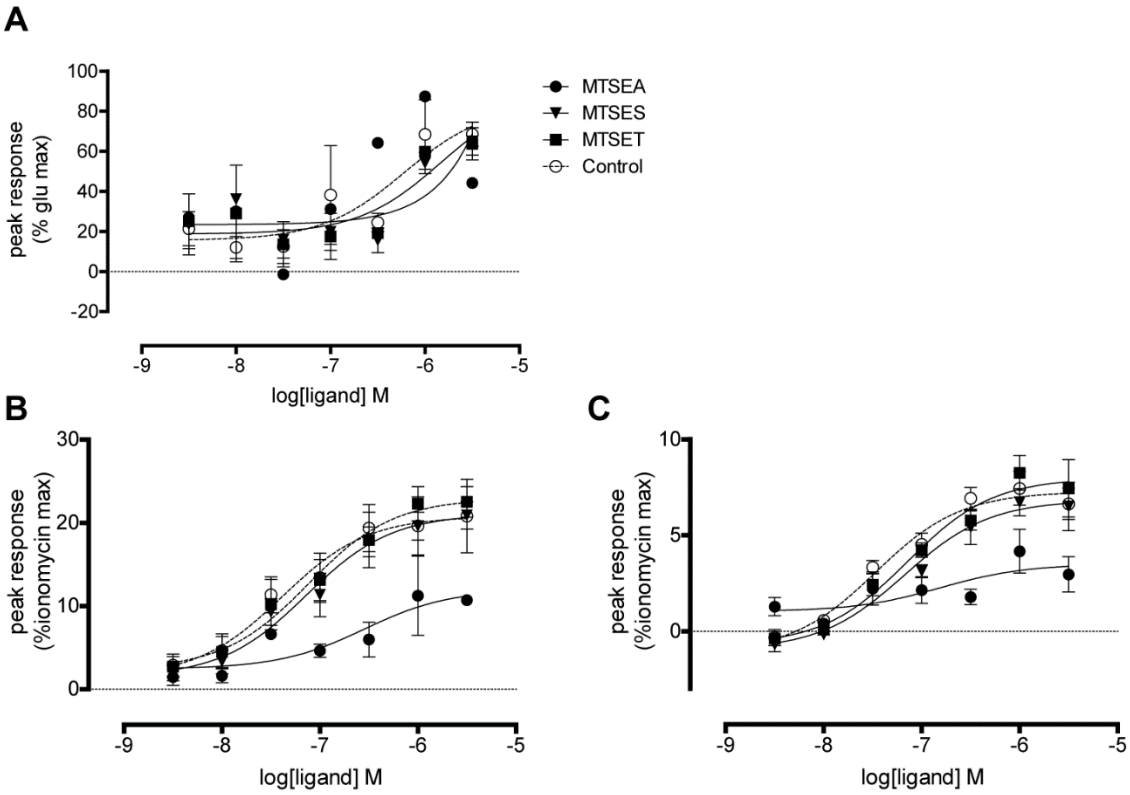


Figure 70

Effect of MTS reagents on full and headless mGlu5 WT receptors on PAM potentiation of mGlu5-mediated Ca2+ mobilization by varying concentrations of VU0360172.

A) Maximal glutamate response in full mGlu5 WT receptors, B) Peak response in headless mGlu5 WT receptors C) Plateau response in headless mGlu5 WT receptors. Data represents the mean ± S.E.M. of 3 experiments performed in triplicate.

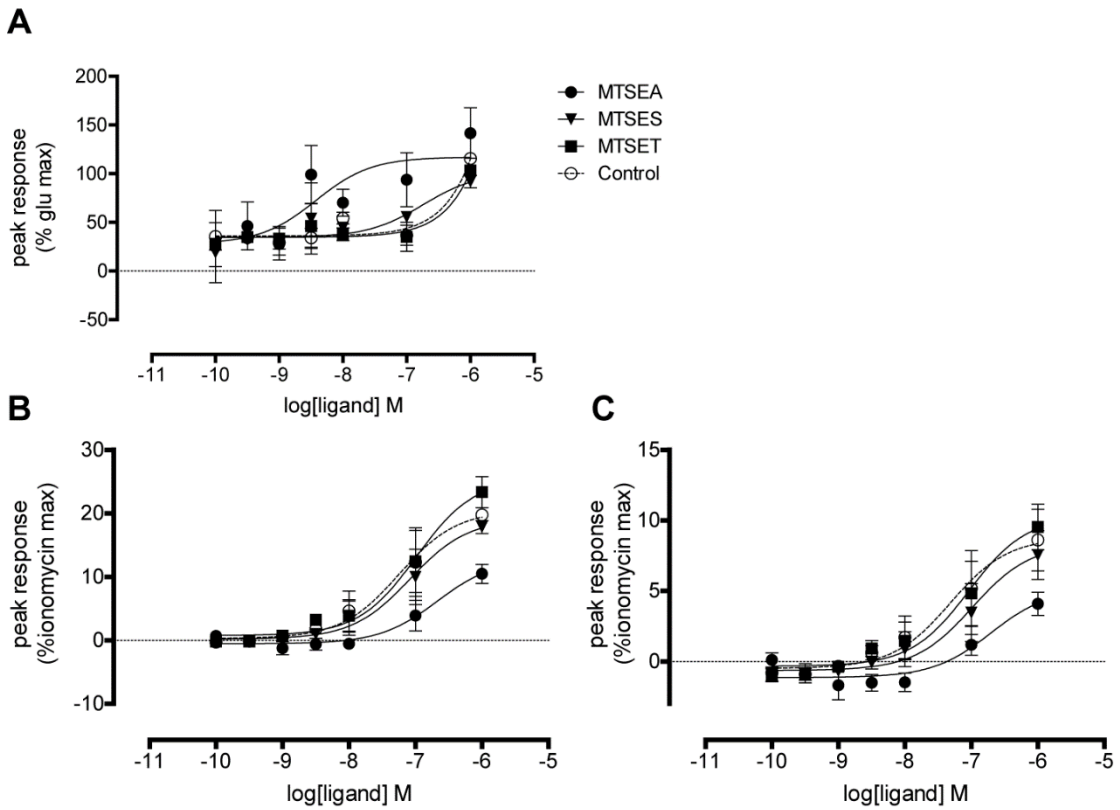


Figure 71

Effect of MTS reagents on full and headless $mGlu_5$ WT receptors on PAM potentiation of $mGlu_5$-mediated Ca^{2+} mobilization by varying concentrations of VU0415051.

A) Maximal glutamate response in full $mGlu_5$ WT receptors, B) Peak response in headless $mGlu_5$ WT receptors C) Plateau response in headless $mGlu_5$ WT receptors. Data represents the mean ± S.E.M. of 3 experiments performed in triplicate.

APPENDIX TO CHAPTER VII

Full list of ANN input descriptors

Table 25
Full list of 259 input descriptors to the ANN-trained method for neighbor count prediction. Scores are based on connection weights between input, hidden and output layers as calculated by the Garson's Connection-Weights algorithm

Score	Descriptor
1	SequenceMax(ReflectingWindowAverage(AA_FreeEnergyTransitionHelix,size=16,alignment=Center,weighting=Triangular))
0.829047	Partial(ReflectingWindow(AABlastProfile,size=4,alignment=Center),indices(10))
0.797703	Partial(ReflectingWindow(AABlastProfile,size=4,alignment=Center),indices(11))
0.797376	Partial(ReflectingWindow(AABlastProfile,size=4,alignment=Center),indices(13))
0.713307	SequenceMean(AA_BlastConservation)
0.709197	Partial(ReflectingWindow(AABlastProfile,size=4,alignment=Center),indices(7))
0.66722	Partial(SequenceStandardDeviation(AABlastProfile),indices(17))
0.623196	NElements
0.606669	Partial(ReflectingWindow(AABlastProbability,size=4,alignment=Center),indices(0))
0.599828	SequenceMin(ReflectingWindowAverage(AA_FreeEnergyTransition,size=16,alignment=Center,weighting=Triangular))
0.570723	Partial(SequenceStandardDeviation(AAType),indices(8))
0.570558	SequenceStandardDeviation(ReflectingWindowAverage(AA_FreeEnergyTransitionHelix,size=16,alignment=Center,weighting=Triangular))
0.566582	AASeqID
0.563347	SequenceMin(ReflectingWindowAverage(AA_FreeEnergyTransitionHelix,size=16,alignment=Center,weighting=Triangular))
0.557327	Partial(ReflectingWindow(AABlastProfile,size=4,alignment=Center),indices(3))
0.553776	SequenceStandardDeviation(ReflectingWindowAverage(AA_FreeEnergyTransition,size=16,alignment=Center,weighting=Triangular))
0.550682	Partial(SequenceMean(AABlastProfile),indices(7))
0.549656	Subtract(lhs=NElements,rhs=AASeqID)
0.544907	SequenceStandardDeviation(ReflectingWindowAverage(AA_TransferFreeEnergyPuntaMaritan3D,size=16,alignment=Center,weighting=Triangular))
0.531037	Partial(SequenceMean(AABlastProfile),indices(13))
0.530669	Partial(ReflectingWindow(AABlastProbability,size=4,alignment=Center),indices(17))
0.520673	Partial(ReflectingWindow(AABlastProfile,size=4,alignment=Center),indices(16))
0.518459	Partial(SequenceStandardDeviation(AABlastProbability),indices(15))
0.514299	Partial(ReflectingWindow(AABlastProfile,size=4,alignment=Center),indices(17))
0.510271	Partial(ReflectingWindow(AABlastProfile,size=4,alignment=Center),indices(14))
0.504071	SequenceMean(ReflectingWindowAverage(AA_FreeEnergyStrand,size=16,alignment=Center,weighting=Triangular))
0.501064	Partial(SequenceStandardDeviation(AABlastProbability),indices(1))
0.481452	SequenceMin(ReflectingWindowAverage(AA_FreeEnergySolutionStrand,size=16,alignment=Center,weighting=Triangular))

0.481075	Partial(CumulativeReflectingWindowAverage(AA_BlastLogProbabilityWeighted(property=AA_FreeEnergyTransitionHelix),size=16,alignment=Center,weighting=Triangular),indices(15))
0.477469	Partial(CumulativeReflectingWindowAverage(AA_BlastLogPTTest(property=AA_FreeEnergyStrand),size=16,alignment=Center,weighting=Triangular),indices(3))
0.47479	Partial(CumulativeReflectingWindowAverage(AA_BlastLogPTTest(property=AA_FreeEnergyTransitionHelix),size=16,alignment=Right,weighting=Triangular),indices(15))
0.47423	SequenceStandardDeviation(ReflectingWindowAverage(AA_FreeEnergyCoreStrand,size=16,alignment=Center,weighting=Triangular))
0.468926	Partial(ReflectingWindow(AABlastProfile,size=4,alignment=Center),indices(80))
0.464858	Partial(ReflectingWindow(AABlastProfile,size=4,alignment=Center),indices(91))
0.464424	SequenceMax(ReflectingWindowAverage(AA_Polarizability,size=16,alignment=Center,weighting=Triangular))
0.459403	Partial(SequenceStandardDeviation(AABlastProfile),indices(18))
0.458146	SequenceMax(ReflectingWindowAverage(AA_FreeEnergyCoreCoil,size=16,alignment=Center,weighting=Triangular))
0.447164	DescriptorStandardDeviation(ReflectingWindow(AA_BlastLogPTTest(property=AA_FreeEnergyStrand),size=16,alignment=Center))
0.443412	Partial(CumulativeReflectingWindowAverage(AA_BlastLogPTTest(property=AA_FreeEnergyTransition),size=16,alignment=Center,weighting=Triangular),indices(0))
0.442138	Partial(SequenceStandardDeviation(AABlastProbability),indices(9))
0.441808	Partial(ReflectingWindow(AABlastProbability,size=4,alignment=Center),indices(15))
0.437699	Partial(ReflectingWindow(AABlastProfile,size=4,alignment=Center),indices(170))
0.436836	DescriptorStandardDeviation(ReflectingWindow(AA_FreeEnergySolution,size=16,alignment=Center))
0.432894	DescriptorMax(Periodogram(AA_BlastLogPTTest(property=AA_FreeEnergyCoil),size=16,periods(2,3,3.6,4.2),alignment=Center))
0.431713	DescriptorStandardDeviation(ReflectingWindow(AA_BlastLogProbabilityWeighted(property=AA_FreeEnergyCoreHelix),size=16,alignment=Center))
0.429765	Partial(Periodogram(AA_BlastLogPTTest(property=AA_TransferFreeEnergyPuntaMaritan3D),size=16,periods(2,3,3.6,4.2),alignment=Center),indices(0))
0.427857	Partial(ReflectingWindow(AABlastProfile,size=4,alignment=Center),indices(9))
0.421331	DescriptorStandardDeviation(ReflectingWindow(AA_BlastLogPTTest(property=AA_FreeEnergyTransitionHelix),size=16,alignment=Center))
0.419777	Partial(ReflectingWindow(AABlastProfile,size=4,alignment=Center),indices(1))
0.41741	Partial(CumulativeReflectingWindowAverage(AA_BlastLogPTTest(property=AA_FreeEnergyTransitionCoil),size=16,alignment=Center,weighting=Triangular),indices(15))
0.416287	DescriptorStandardDeviation(ReflectingWindow(AA_FreeEnergyTransitionStrand,size=16,alignment=Center))
0.415021	Partial(ReflectingWindow(AABlastProbability,size=4,alignment=Center),indices(80))
0.414229	Partial(ReflectingWindow(AABlastProfile,size=4,alignment=Center),indices(162))
0.412525	SequenceMax(ReflectingWindowAverage(AA_FreeEnergyTransitionCoil,size=16,alignment=Center,weighting=Triangular))
0.403964	Partial(Periodogram(AA_BlastLogPTTest(property=AA_Hydrophobicity),size=16,periods(2,3,3.6,4.2),alignment=Center),indices(3))
0.400202	Partial(CumulativeReflectingWindowAverage(AA_FreeEnergyCoreStrand,size=16,alignment=Center,weighting=Triangular),indices(7))

0.398634	Partial(CumulativeReflectingWindowAverage(AA_BlastLogPTTest(property=AA_FreeEnergyStrand),size=16,alignment=Center,weighting=Triangular),indices(15))
0.397818	Partial(ReflectingWindow(AABlastProbability,size=4,alignment=Center),indices(163))
0.396704	Partial(ReflectingWindow(AABlastProfile,size=4,alignment=Center),indices(90))
0.395629	Partial(SequenceStandardDeviation(AABlastProfile),indices(8))
0.394072	Partial(Periodogram(AA_BlastLogPTTest(property=AA_Polarizability),size=16,periods(2,3,3.6,4.2),alignment=Center),indices(3))
0.392556	SequenceStandardDeviation(ReflectingWindowAverage(AA_FreeEnergyCore,size=16,alignment=Center,weighting=Triangular))
0.391928	DescriptorStandardDeviation(ReflectingWindow(AA_BlastLogProbabilityWeighted(property=AA_FreeEnergyStrand),size=16,alignment=Center))
0.391166	SequenceMax(ReflectingWindowAverage(AA_TransferFreeEnergyPuntaMaritan3D,size=16,alignment=Center,weighting=Triangular))
0.389927	Partial(ReflectingWindow(AABlastProfile,size=4,alignment=Center),indices(92))
0.389659	Partial(ReflectingWindow(AABlastProfile,size=4,alignment=Center),indices(8))
0.388102	Partial(ReflectingWindow(AABlastProbability,size=4,alignment=Center),indices(16))
0.387903	DescriptorMax(Periodogram(AA_BlastLogPTTest(property=AA_Hydrophobicity),size=16,periods(3,3.6,4.2),alignment=Center))
0.386894	Partial(ReflectingWindow(AABlastProbability,size=4,alignment=Center),indices(162))
0.386772	Partial(CumulativeReflectingWindowAverage(AA_BlastConservation,size=16,alignment=Right,weighting=Triangular),indices(3))
0.386451	DescriptorStandardDeviation(ReflectingWindow(AA_BlastLogPTTest(property=AA_FreeEnergySolution),size=16,alignment=Center))
0.384777	Partial(ReflectingWindow(AABlastProfile,size=4,alignment=Center),indices(144))
0.384739	DescriptorMax(Periodogram(AA_BlastLogPTTest(property=AA_FreeEnergyHelix),size=16,periods(2,3,3.6,4.2),alignment=Center))
0.384205	Partial(Periodogram(AA_BlastLogPTTest(property=AA_SASA),size=16,periods(2,3,3.6,4.2),alignment=Center),indices(3))
0.382913	Partial(ReflectingWindow(AABlastProbability,size=4,alignment=Center),indices(99))
0.381526	Partial(ReflectingWindow(AABlastProfile,size=4,alignment=Center),indices(176))
0.380893	Partial(SequenceStandardDeviation(AABlastProbability),indices(2))
0.380504	DescriptorStandardDeviation(ReflectingWindow(AA_BlastLogProbabilityWeighted(property=AA_Hydrophobicity),size=16,alignment=Center))
0.38012	Partial(ReflectingWindow(AABlastProfile,size=4,alignment=Center),indices(105))
0.378867	Partial(ReflectingWindow(AABlastProfile,size=4,alignment=Center),indices(97))
0.378321	DescriptorStandardDeviation(ReflectingWindow(AA_StrandProbability,size=16,alignment=Center))
0.376871	Partial(CumulativeReflectingWindowAverage(AA_BlastLogPTTest(property=AA_FreeEnergyCoreStrand),size=16,alignment=Right,weighting=Triangular),indices(3))
0.375472	Partial(ReflectingWindow(AABlastProfile,size=4,alignment=Center),indices(0))
0.37392	DescriptorStandardDeviation(ReflectingWindow(AA_BlastLogPTTest(property=AA_FreeEnergySolutionStrand),size=16,alignment=Center))
0.373603	Partial(CumulativeReflectingWindowAverage(AA_FreeEnergyCoreStrand,size=16,alignment=Right,weighting=Triangular),indices(15))
0.373133	Partial(ReflectingWindow(AABlastProfile,size=4,alignment=Center),indices(173))
0.372806	DescriptorStandardDeviation(ReflectingWindow(AA_SASA,size=16,alignment=Center))

0.370322	SequenceMax(ReflectingWindowAverage(AA_FreeEnergyCore,size=16,alignment=Center,weighting=Triangular))
0.369694	Partial(ReflectingWindow(AABlastProfile,size=4,alignment=Center),indices(167))
0.369556	Partial(SequenceMean(AABlastProbability),indices(1))
0.367085	Partial(CumulativeReflectingWindowAverage(AA_BlastLogProbabilityWeighted(property=AA_FreeEnergyTransitionStrand),size=16,alignment=Center,weighting=Triangular),indices(7))
0.364502	Partial(ReflectingWindow(AABlastProfile,size=4,alignment=Center),indices(84))
0.363661	Partial(SequenceStandardDeviation(AABlastProbability),indices(0))
0.363362	Partial(ReflectingWindow(AABlastProfile,size=4,alignment=Center),indices(49))
0.360095	Partial(CumulativeReflectingWindowAverage(AA_BlastLogProbabilityWeighted(property=AA_HelixProbability),size=16,alignment=Left,weighting=Triangular),indices(7))
0.359714	Partial(CumulativeReflectingWindowAverage(AA_BlastLogPTTest(property=AA_FreeEnergyTransitionHelix),size=16,alignment=Left,weighting=Triangular),indices(1))
0.35808	Partial(CumulativeReflectingWindowAverage(AA_BlastLogPTTest(property=AA_FreeEnergyCoreStrand),size=16,alignment=Left,weighting=Triangular),indices(7))
0.357764	Partial(Periodogram(AA_TransferFreeEnergyPuntaMaritan3D,size=16,periods(2,3,3.6,4.2),alignment=Center),indices(2))
0.357693	Partial(CumulativeReflectingWindowAverage(AA_BlastLogPTTest(property=AA_FreeEnergyTransition),size=16,alignment=Right,weighting=Triangular),indices(1))
0.356259	Partial(ReflectingWindow(AABlastProfile,size=4,alignment=Center),indices(19))
0.355625	Partial(ReflectingWindow(AABlastProfile,size=4,alignment=Center),indices(39))
0.354808	Partial(SequenceStandardDeviation(AAType),indices(5))
0.354031	SequenceMax(ReflectingWindowAverage(AA_Hydrophobicity,size=16,alignment=Center,weighting=Triangular))
0.35226	DescriptorStandardDeviation(ReflectingWindow(AA_FreeEnergyHelix,size=16,alignment=Center))
0.351483	Partial(SequenceMean(AABlastProbability),indices(5))
0.350242	Partial(ReflectingWindow(AABlastProbability,size=4,alignment=Center),indices(140))
0.350094	Partial(ReflectingWindow(AABlastProfile,size=4,alignment=Center),indices(178))
0.348178	Partial(ReflectingWindow(AABlastProfile,size=4,alignment=Center),indices(47))
0.347879	Partial(CumulativeReflectingWindowAverage(AA_BlastLogPTTest(property=AA_FreeEnergyTransitionCoil),size=16,alignment=Center,weighting=Triangular),indices(3))
0.346925	Partial(ReflectingWindow(AABlastProfile,size=4,alignment=Center),indices(177))
0.34679	Partial(ReflectingWindow(AABlastProfile,size=4,alignment=Center),indices(24))
0.345924	Partial(CumulativeReflectingWindowAverage(AA_BlastLogProbabilityWeighted(property=AA_FreeEnergyTransitionCoil),size=16,alignment=Right,weighting=Triangular),indices(7))
0.345904	Partial(ReflectingWindow(AABlastProfile,size=4,alignment=Center),indices(171))
0.345699	Partial(ReflectingWindow(AABlastProfile,size=4,alignment=Center),indices(77))
0.343949	Partial(ReflectingWindow(AABlastProbability,size=4,alignment=Center),indices(20))
0.343838	Partial(CumulativeReflectingWindowAverage(AA_BlastLogPTTest(property=AA_FreeEnergyTransition),size=16,alignment=Center,weighting=Triangular),indices(1))
0.343371	Partial(ReflectingWindow(AABlastProbability,size=4,alignment=Center),indices(176))

0.342961	Partial(ReflectingWindow(AABlastProfile,size=4,alignment=Center),indices(71))
0.341275	Partial(ReflectingWindow(AAType,size=4,alignment=Center),indices(127))
0.340834	Partial(CumulativeReflectingWindowAverage(AA_BlastLogProbabilityWeighted(property=AA_FreeEnergyTransitionStrand),size=16,alignment=Left,weighting=Triangular),indices(7))
0.340337	Partial(ReflectingWindow(AABlastProfile,size=4,alignment=Center),indices(129))
0.340036	Partial(Periodogram(AA_BlastLogProbabilityWeighted(property=AA_StericalParameter),size=16,periods(2,3,3.6,4.2),alignment=Center),indices(2))
0.337931	Partial(SequenceStandardDeviation(AAType),indices(1))
0.336901	Partial(CumulativeReflectingWindowAverage(AA_BlastLogPTTest(property=AA_SASA),size=16,alignment=Center,weighting=Triangular),indices(15))
0.336802	Partial(Periodogram(AA_SASA,size=16,periods(2,3,3.6,4.2),alignment=Center),indices(0))
0.335816	Partial(ReflectingWindow(AABlastProfile,size=4,alignment=Center),indices(154))
0.33443	Partial(CumulativeReflectingWindowAverage(AA_StrandProbability,size=16,alignment=Right,weighting=Triangular),indices(15))
0.334304	Partial(CumulativeReflectingWindowAverage(AA_BlastLogPTTest(property=AA_HelixProbability),size=16,alignment=Center,weighting=Triangular),indices(0))
0.331963	Partial(Periodogram(AA_Hydrophobicity,size=16,periods(2,3,3.6,4.2),alignment=Center),indices(1))
0.330889	Partial(ReflectingWindow(AABlastProfile,size=4,alignment=Center),indices(34))
0.330131	Partial(ReflectingWindow(AABlastProfile,size=4,alignment=Center),indices(64))
0.32762	Partial(CumulativeReflectingWindowAverage(AA_BlastLogProbabilityWeighted(property=AA_FreeEnergyTransitionCoil),size=16,alignment=Right,weighting=Triangular),indices(3))
0.326527	Partial(ReflectingWindow(AABlastProfile,size=4,alignment=Center),indices(131))
0.326372	Partial(Periodogram(AA_SASA,size=16,periods(2,3,3.6,4.2),alignment=Center),indices(2))
0.326288	Partial(CumulativeReflectingWindowAverage(AA_BlastLogProbabilityWeighted(property=AA_FreeEnergyCoreStrand),size=16,alignment=Right,weighting=Triangular),indices(15))
0.326014	Partial(ReflectingWindow(AABlastProbability,size=4,alignment=Center),indices(100))
0.325112	Partial(ReflectingWindow(AABlastProfile,size=4,alignment=Center),indices(145))
0.324867	Partial(ReflectingWindow(AABlastProfile,size=4,alignment=Center),indices(61))
0.324814	Partial(ReflectingWindow(AABlastProfile,size=4,alignment=Center),indices(164))
0.324664	Partial(ReflectingWindow(AABlastProfile,size=4,alignment=Center),indices(169))
0.324649	Partial(CumulativeReflectingWindowAverage(AA_BlastLogProbabilityWeighted(property=AA_Hydrophobicity),size=16,alignment=Center,weighting=Triangular),indices(7))
0.324237	Partial(CumulativeReflectingWindowAverage(AA_BlastConservation,size=16,alignment=Right,weighting=Triangular),indices(15))
0.323859	Partial(ReflectingWindow(AABlastProfile,size=4,alignment=Center),indices(58))
0.323408	Partial(ReflectingWindow(AABlastProbability,size=4,alignment=Center),indices(151))
0.322934	Partial(CumulativeReflectingWindowAverage(AA_BlastLogPTTest(property=AA_FreeEnergyTransitionCoil),size=16,alignment=Left,weighting=Triangular),indices(15))
0.321706	Partial(ReflectingWindow(AABlastProfile,size=4,alignment=Center),indices(15))
0.321636	Partial(Periodogram(AA_StericalParameter,size=16,periods(2,3,3.6,4.2),alignment=Center),indices(0))

0.321293	Partial(ReflectingWindow(AABlastProfile,size=4,alignment=Center),indices(93))
0.321214	Partial(CumulativeReflectingWindowAverage(AA_BlastLogPTTest(property=AA_FreeEnergyHelix),size=16,alignment=Center,weighting=Triangular),indices(0))
0.32074	Partial(CumulativeReflectingWindowAverage(AA_BlastConservation,size=16,alignment=Center,weighting=Triangular),indices(0))
0.320105	Partial(ReflectingWindow(AABlastProfile,size=4,alignment=Center),indices(44))
0.319412	Partial(CumulativeReflectingWindowAverage(AA_BlastLogProbabilityWeighted(property=AA_FreeEnergyTransitionStrand),size=16,alignment=Right,weighting=Triangular),indices(7))
0.31884	Partial(ReflectingWindow(AABlastProbability,size=4,alignment=Center),indices(115))
0.317147	Partial(ReflectingWindow(AABlastProfile,size=4,alignment=Center),indices(99))
0.316556	Partial(ReflectingWindow(AAType,size=4,alignment=Center),indices(178))
0.315596	Partial(ReflectingWindow(AA_BlastConservation,size=4,alignment=Center),indices(0))
0.315482	Partial(ReflectingWindow(AABlastProfile,size=4,alignment=Center),indices(76))
0.314785	Partial(CumulativeReflectingWindowAverage(AA_BlastLogPTTest(property=AA_FreeEnergyTransitionStrand),size=16,alignment=Right,weighting=Triangular),indices(1))
0.314195	Partial(CumulativeReflectingWindowAverage(AA_BlastLogPTTest(property=AA_FreeEnergySolutionHelix),size=16,alignment=Right,weighting=Triangular),indices(3))
0.313827	Partial(ReflectingWindow(AABlastProfile,size=4,alignment=Center),indices(6))
0.312878	Partial(ReflectingWindow(AABlastProfile,size=4,alignment=Center),indices(124))
0.312069	Partial(CumulativeReflectingWindowAverage(AA_FreeEnergyCoreCoil,size=16,alignment=Center,weighting=Triangular),indices(15))
0.311443	Partial(ReflectingWindow(AABlastProfile,size=4,alignment=Center),indices(120))
0.311367	Partial(ReflectingWindow(AABlastProbability,size=4,alignment=Center),indices(177))
0.311268	Partial(ReflectingWindow(AABlastProbability,size=4,alignment=Center),indices(174))
0.310677	Partial(CumulativeReflectingWindowAverage(AA_BlastLogPTTest(property=AA_FreeEnergySolutionCoil),size=16,alignment=Left,weighting=Triangular),indices(1))
0.310658	Partial(ReflectingWindow(AABlastProfile,size=4,alignment=Center),indices(31))
0.310224	Partial(CumulativeReflectingWindowAverage(AA_BlastLogProbabilityWeighted(property=AA_FreeEnergyCoreCoil),size=16,alignment=Left,weighting=Triangular),indices(7))
0.30997	Partial(ReflectingWindow(AABlastProfile,size=4,alignment=Center),indices(155))
0.309182	Partial(ReflectingWindow(AABlastProfile,size=4,alignment=Center),indices(136))
0.308369	Partial(CumulativeReflectingWindowAverage(AA_BlastLogPTTest(property=AA_SASA),size=16,alignment=Left,weighting=Triangular),indices(15))
0.307779	Partial(SequenceMean(AAType),indices(5))
0.30774	Partial(CumulativeReflectingWindowAverage(AA_BlastLogProbabilityWeighted(property=AA_FreeEnergyCoreCoil),size=16,alignment=Center,weighting=Triangular),indices(0))
0.307356	Partial(CumulativeReflectingWindowAverage(AA_FreeEnergySolutionCoil,size=16,alignment=Center,weighting=Triangular),indices(0))
0.307021	Partial(ReflectingWindow(AABlastProbability,size=4,alignment=Center),indices(27))

0.306966	Partial(CumulativeReflectingWindowAverage(AA_BlastLogPTTest(property=AA_FreeEnergyTransitionStrand),size=16,alignment=Right,weighting=Triangular),indices(7))
0.306506	Partial(ReflectingWindow(AA_StrandProbability,size=4,alignment=Center),indices(2))
0.306112	Partial(ReflectingWindow(AABlastProfile,size=4,alignment=Center),indices(35))
0.305586	Partial(ReflectingWindow(AA_BlastConservation,size=4,alignment=Center),indices(6))
0.305104	Partial(ReflectingWindow(AABlastProfile,size=4,alignment=Center),indices(174))
0.305061	Partial(ReflectingWindow(AABlastProfile,size=4,alignment=Center),indices(48))
0.30476	Partial(ReflectingWindow(AABlastProbability,size=4,alignment=Center),indices(57))
0.30425	Partial(ReflectingWindow(AA_StrandProbability,size=4,alignment=Center),indices(5))
0.303649	Partial(ReflectingWindow(AABlastProbability,size=4,alignment=Center),indices(48))
0.303477	Partial(CumulativeReflectingWindowAverage(AA_BlastLogProbabilityWeighted(property=AA_FreeEnergyCoreCoil),size=16,alignment=Center,weighting=Triangular),indices(7))
0.303011	Partial(CumulativeReflectingWindowAverage(AA_FreeEnergyCoreCoil,size=16,alignment=Left,weighting=Triangular),indices(15))
0.302552	Partial(CumulativeReflectingWindowAverage(AA_BlastLogPTTest(property=AA_SASA),size=16,alignment=Center,weighting=Triangular),indices(7))
0.302151	Partial(ReflectingWindow(AABlastProbability,size=4,alignment=Center),indices(160))
0.301756	Partial(ReflectingWindow(AABlastProfile,size=4,alignment=Center),indices(98))
0.301228	Partial(ReflectingWindow(AABlastProfile,size=4,alignment=Center),indices(56))
0.30048	Partial(CumulativeReflectingWindowAverage(AA_BlastLogProbabilityWeighted(property=AA_FreeEnergyTransitionStrand),size=16,alignment=Center,weighting=Triangular),indices(0))
0.300092	Partial(CumulativeReflectingWindowAverage(AA_BlastLogProbabilityWeighted(property=AA_FreeEnergyTransitionCoil),size=16,alignment=Center,weighting=Triangular),indices(0))
0.299952	Partial(ReflectingWindow(AA_BlastConservation,size=4,alignment=Center),indices(8))
0.299751	Partial(ReflectingWindow(AA_FreeEnergyCore,size=4,alignment=Center),indices(5))
0.299116	Partial(ReflectingWindow(AABlastProfile,size=4,alignment=Center),indices(5))
0.298629	Partial(CumulativeReflectingWindowAverage(AA_BlastLogPTTest(property=AA_FreeEnergySolutionHelix),size=16,alignment=Center,weighting=Triangular),indices(1))
0.298328	Partial(ReflectingWindow(AA_FreeEnergyCoreCoil,size=4,alignment=Center),indices(1))
0.298277	Partial(CumulativeReflectingWindowAverage(AA_BlastLogPTTest(property=AA_SASA),size=16,alignment=Center,weighting=Triangular),indices(1))
0.29821	Partial(ReflectingWindow(AABlastProfile,size=4,alignment=Center),indices(59))
0.297822	Partial(CumulativeReflectingWindowAverage(AA_BlastLogPTTest(property=AA_StrandProbability),size=16,alignment=Center,weighting=Triangular),indices(1))
0.297745	Partial(ReflectingWindow(AA_BlastConservation,size=4,alignment=Center),indices(2))
0.296822	SequenceMin(ReflectingWindowAverage(AA_FreeEnergyCoreStrand,size=16,alignment=Center,weighting=Triangular))

0.29667	Partial(CumulativeReflectingWindowAverage(AA_BlastLogProbabilityWeighted(property=AA_FreeEnergyTransition),size=16,alignment=Right,weighting=Triangular),indices(7))
0.29608	Partial(CumulativeReflectingWindowAverage(AA_BlastLogPTTest(property=AA_FreeEnergyTransitionCoil),size=16,alignment=Left,weighting=Triangular),indices(1))
0.29591	Partial(ReflectingWindow(AABlastProbability,size=4,alignment=Center),indices(168))
0.295243	DescriptorStandardDeviation(ReflectingWindow(AA_Polarizability,size=16,alignment=Center))
0.295234	Partial(ReflectingWindow(AAType,size=4,alignment=Center),indices(6))
0.294967	Partial(ReflectingWindow(AAType,size=4,alignment=Center),indices(106))
0.29487	Partial(ReflectingWindow(AABlastProfile,size=4,alignment=Center),indices(172))
0.294759	Partial(CumulativeReflectingWindowAverage(AA_BlastLogPTTest(property=AA_FreeEnergyCoreCoil),size=16,alignment=Center,weighting=Triangular),indices(15))
0.294715	Partial(ReflectingWindow(AA_FreeEnergyCoreStrand,size=4,alignment=Center),indices(8))
0.293526	Partial(ReflectingWindow(AABlastProfile,size=4,alignment=Center),indices(147))
0.293507	Partial(CumulativeReflectingWindowAverage(AA_FreeEnergyCoreCoil,size=16,alignment=Center,weighting=Triangular),indices(7))
0.293257	Partial(ReflectingWindow(AAType,size=4,alignment=Center),indices(120))
0.291696	Partial(ReflectingWindow(AABlastProbability,size=4,alignment=Center),indices(94))
0.291571	SequenceStandardDeviation(ReflectingWindowAverage(AA_FreeEnergySolutionCoil,size=16,alignment=Center,weighting=Triangular))
0.291479	Partial(CumulativeReflectingWindowAverage(AA_BlastLogPTTest(property=AA_FreeEnergyCore),size=16,alignment=Left,weighting=Triangular),indices(15))
0.290126	Partial(CumulativeReflectingWindowAverage(AA_BlastLogProbabilityWeighted(property=AA_FreeEnergyTransitionCoil),size=16,alignment=Right,weighting=Triangular),indices(1))
0.289985	DescriptorMax(Periodogram(AA_BlastLogPTTest(property=AA_StericalParameter),size=16,periods(3,3.6,4.2),alignment=Center))
0.28969	Partial(ReflectingWindow(AABlastProfile,size=4,alignment=Center),indices(175))
0.289587	Partial(ReflectingWindow(AABlastProfile,size=4,alignment=Center),indices(151))
0.288476	Partial(CumulativeReflectingWindowAverage(AA_BlastLogProbabilityWeighted(property=AA_FreeEnergyCoreCoil),size=16,alignment=Right,weighting=Triangular),indices(7))
0.28762	Partial(ReflectingWindow(AABlastProbability,size=4,alignment=Center),indices(169))
0.287605	Partial(ReflectingWindow(AABlastProfile,size=4,alignment=Center),indices(33))
0.286271	Partial(ReflectingWindow(AABlastProfile,size=4,alignment=Center),indices(156))
0.286007	Partial(ReflectingWindow(AABlastProbability,size=4,alignment=Center),indices(172))
0.285649	Partial(CumulativeReflectingWindowAverage(AA_BlastLogProbabilityWeighted(property=AA_FreeEnergyTransitionCoil),size=16,alignment=Center,weighting=Triangular),indices(1))
0.284572	Partial(ReflectingWindow(AABlastProbability,size=4,alignment=Center),indices(2))
0.284471	Partial(CumulativeReflectingWindowAverage(AA_BlastLogPTTest(property=AA_StrandProbability),size=16,alignment=Right,weighting=Triangular),indices(1))
0.284007	Partial(ReflectingWindow(AAType,size=4,alignment=Center),indices(0))

0.283072	Partial(ReflectingWindow(AAType,size=4,alignment=Center),indices(19))
0.282801	Partial(ReflectingWindow(AABlastProbability,size=4,alignment=Center),indices(161))
0.281683	Partial(ReflectingWindow(AABlastProbability,size=4,alignment=Center),indices(35))
0.281571	Partial(ReflectingWindow(AAType,size=4,alignment=Center),indices(15))
0.281328	SequenceMax(ReflectingWindowAverage(AA_FreeEnergyCoreStrand,size=16,alignment=Center,weighting=Triangular))
0.279376	Partial(CumulativeReflectingWindowAverage(AA_BlastLogPTTest(property=AA_FreeEnergyTransitionHelix),size=16,alignment=Right,weighting=Triangular),indices(1))
0.277735	Partial(ReflectingWindow(AABlastProbability,size=4,alignment=Center),indices(157))
0.274458	Partial(ReflectingWindow(AAType,size=4,alignment=Center),indices(162))
0.270877	Partial(SequenceStandardDeviation(AAType),indices(4))
0.268002	Partial(CumulativeReflectingWindowAverage(AA_BlastLogPTTest(property=AA_FreeEnergySolutionHelix),size=16,alignment=Right,weighting=Triangular),indices(1))
0.265953	Partial(ReflectingWindow(AABlastProbability,size=4,alignment=Center),indices(5))
0.265032	Partial(SequenceMean(AAType),indices(15))
0.262695	Partial(CumulativeReflectingWindowAverage(AA_BlastLogPTTest(property=AA_FreeEnergyTransition),size=16,alignment=Right,weighting=Triangular),indices(3))
0.261599	Partial(ReflectingWindow(AABlastProfile,size=4,alignment=Center),indices(134))
0.260537	Partial(ReflectingWindow(AABlastProbability,size=4,alignment=Center),indices(41))
0.259643	SequenceMax(ReflectingWindowAverage(AA_FreeEnergySolution,size=16,alignment=Center,weighting=Triangular))
0.25533	Partial(CumulativeReflectingWindowAverage(AA_HelixProbability,size=16,alignment=Left,weighting=Triangular),indices(7))
0.247675	Partial(SequenceStandardDeviation(AAType),indices(14))
0.246449	Partial(ReflectingWindow(AABlastProfile,size=4,alignment=Center),indices(142))
0.245979	Partial(CumulativeReflectingWindowAverage(AA_BlastLogPTTest(property=AA_FreeEnergySolutionHelix),size=16,alignment=Center,weighting=Triangular),indices(3))
0.225753	Partial(ReflectingWindow(AABlastProfile,size=4,alignment=Center),indices(25))
0.221431	Partial(SequenceMean(AABlastProfile),indices(12))
0.204678	Partial(SequenceStandardDeviation(AABlastProbability),indices(11))
0.196424	Partial(SequenceStandardDeviation(AABlastProfile),indices(1))
0.192793	Partial(ReflectingWindow(AAType,size=4,alignment=Center),indices(3))
0.180357	Partial(CumulativeReflectingWindowAverage(AA_BlastLogPTTest(property=AA_FreeEnergySolutionCoil),size=16,alignment=Center,weighting=Triangular),indices(1))
0.178886	Partial(ReflectingWindow(AABlastProfile,size=4,alignment=Center),indices(179))
0.128566	Partial(CumulativeReflectingWindowAverage(AA_StericalParameter,size=16,alignment=Right,weighting=Triangular),indices(7))
0.01	Partial(ReflectingWindow(AABlastProfile,size=4,alignment=Center),indices(153))

Structure prediction results on proteins from training database

Table 26
Model accuracy for structure prediction a set of 4 proteins from the training database with and without inclusion of the neighbor count score.
Models were ranked by highest contact recovery, lowest RMSD100 and lowest total score.

		Contact Recovery		RMSD 100	
		Best	Top Rank	Best	Top Rank
1U19	No NC	24.1	0.8	8	5.2
	Predicted NC	20.4	7.3	10.1	4.8
	Calculated NC	24.9	6.5	8.1	4.3
2K73	No NC	37.6	3.2	9.8	4.9
	Predicted NC	46.2	12.9	9.4	3.6
	Calculated NC	47.3	38.7	5.9	3.6
2KSY	No NC	23.8	10.7	10.2	4.3
	Predicted NC	24.2	4.4	7.1	4.1
	Calculated NC	29	11.9	10	4.6
3GIA	No NC	5.7	1.3	11.7	9.4
	Predicted NC	5.2	1	12.7	9.4
	Calculated NC	5.7	1.8	11.6	9.5
Average	No NC	22.8	4	9.925	5.95
	Predicted NC	24	6.4	9.825	5.475
	Calculated NC	26.725	14.725	8.9	5.5

Structure prediction the benchmark membrane protein dataset with and without neighbor count score

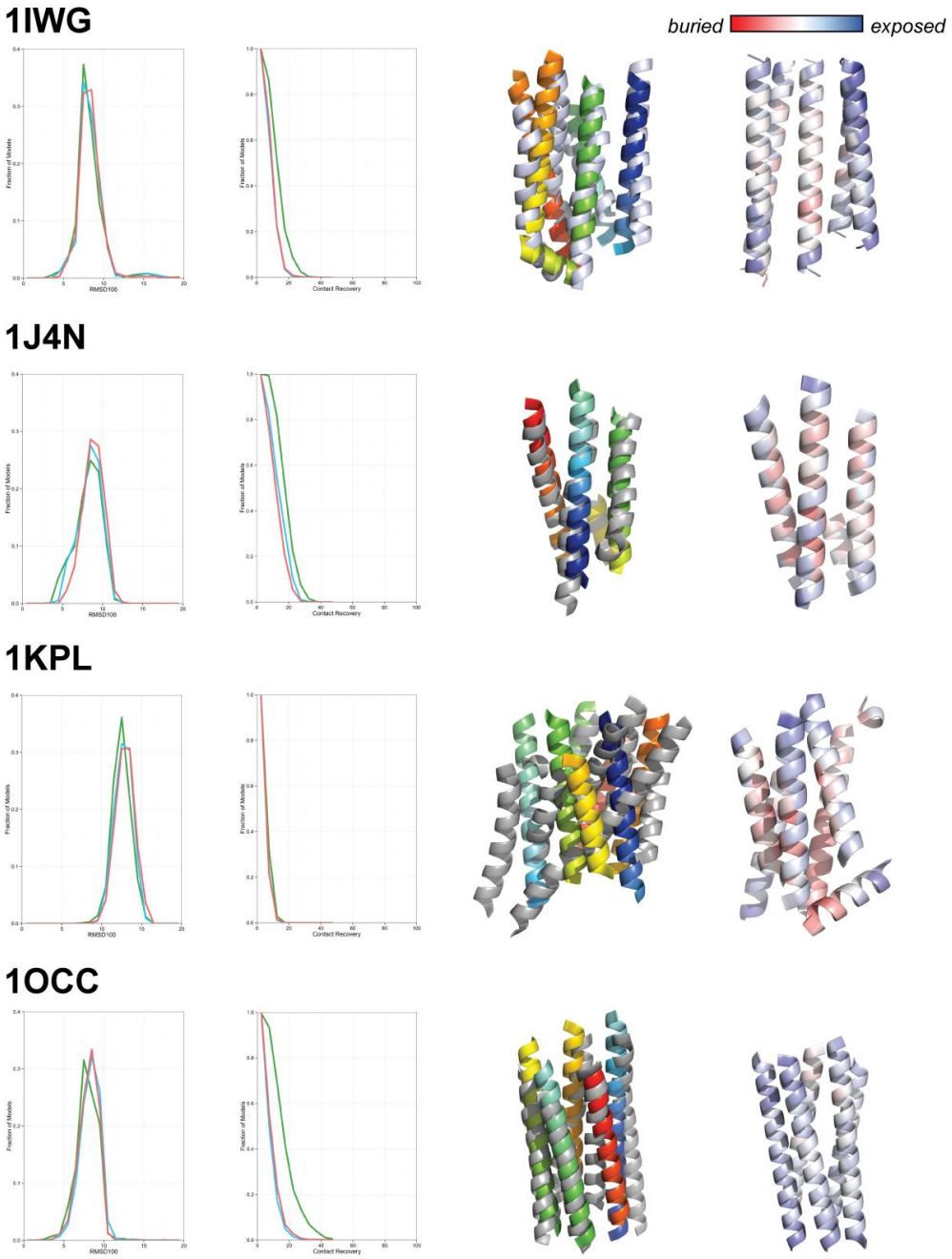
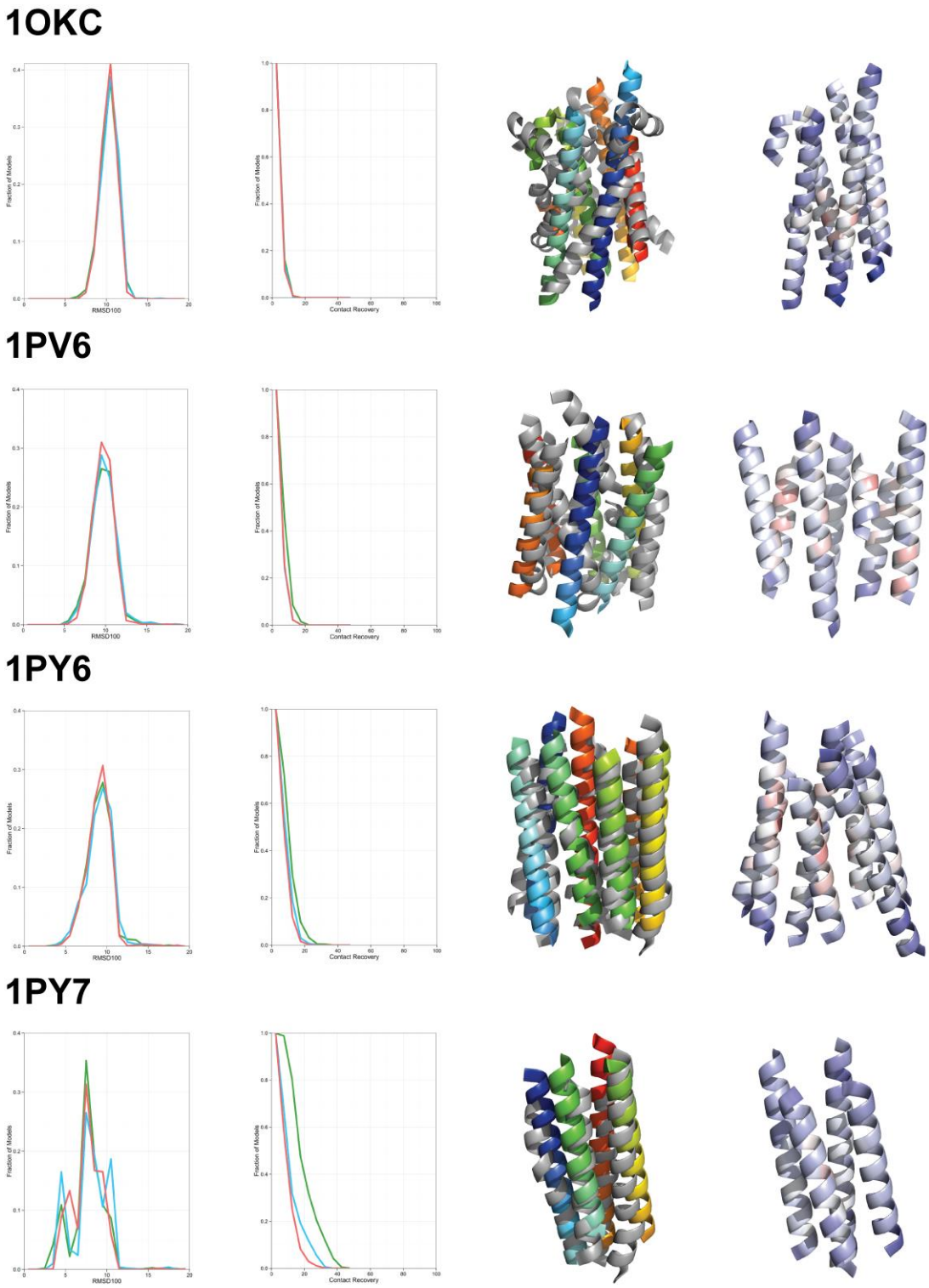
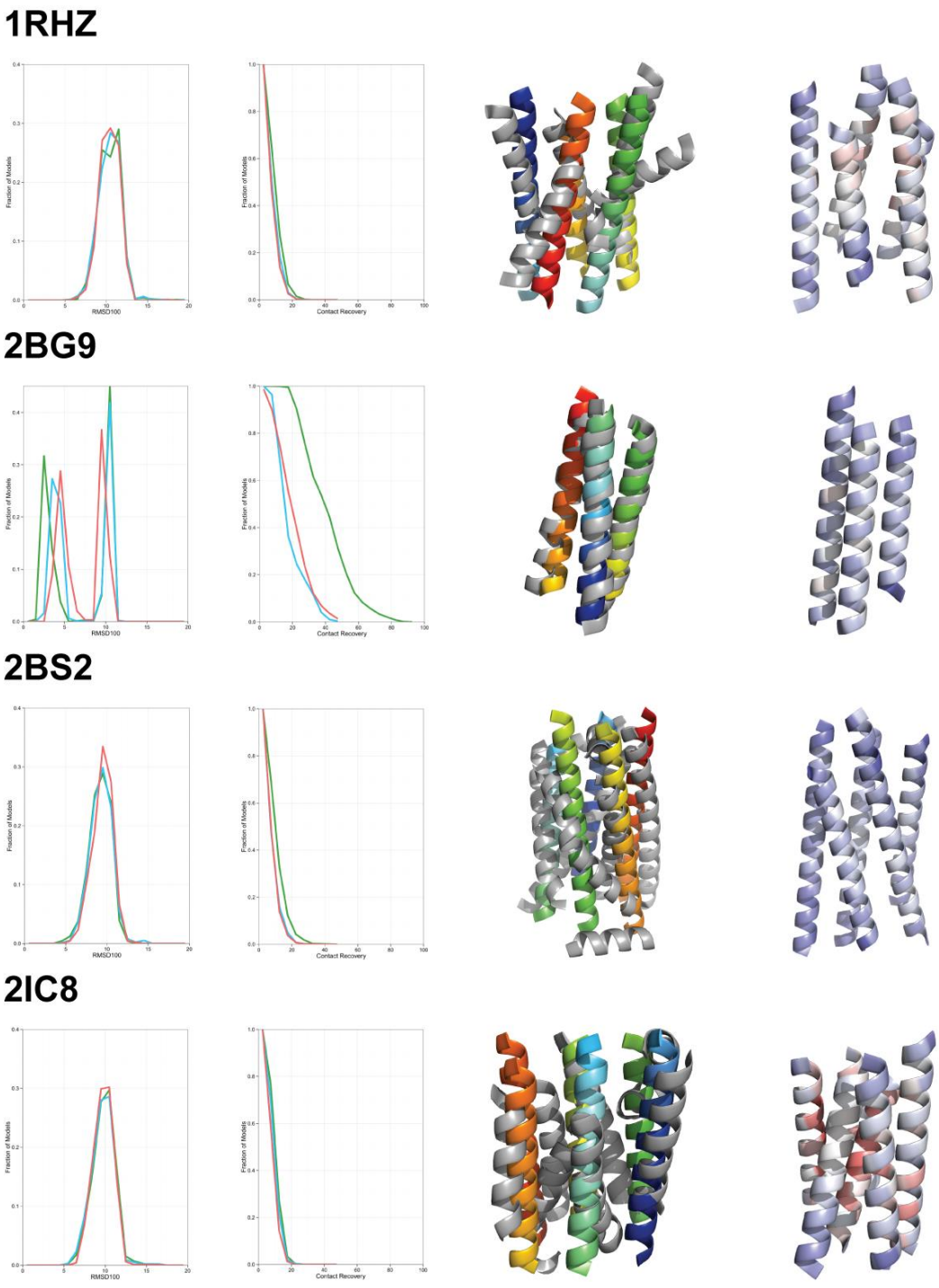


Figure 72
Gallery of results for 18 membrane proteins in the benchmark dataset.
The fraction of models at a range of RMSD100 values and contact recovery is shown for models predicted without neighbor count score (red), with predicted neighbor count (blue) and with calculated neighbor count (green). The top model ranked by RMSD100 is shown in rainbow compared to the native structure in grey. The top model ranked by contact recovery is colored by relative exposure and burial based on predicted neighbor count. Continues on next page.

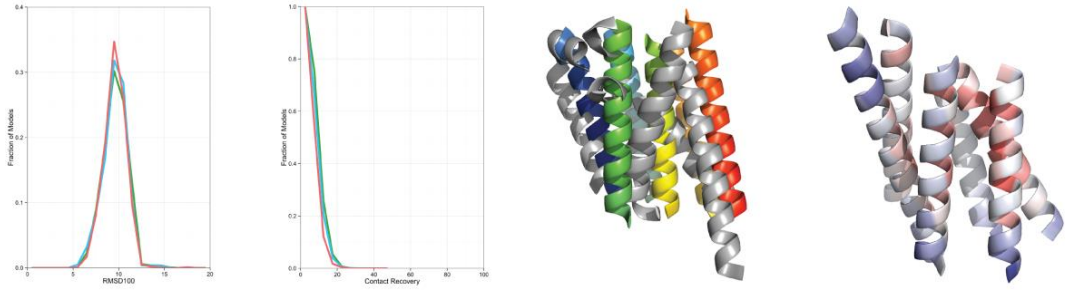


(Figure 69, continued from the previous page) The fraction of models at a range of RMSD100 values and contact recovery is shown for models predicted without neighbor count score (red), with predicted neighbor count (blue) and with calculated neighbor count (green). The top model ranked by RMSD100 is shown in rainbow compared to the native structure in grey. The top model ranked by contact recovery is colored by relative exposure and burial based on predicted neighbor count. Continues on next page.

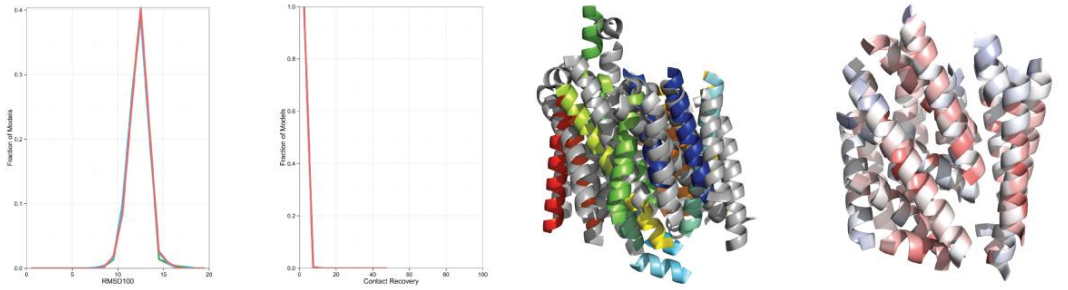


(Figure 69, continued from the previous page) The fraction of models at a range of RMSD100 values and contact recovery is shown for models predicted without neighbor count score (red), with predicted neighbor count (blue) and with calculated neighbor count (green). The top model ranked by RMSD100 is shown in rainbow compared to the native structure in grey. The top model ranked by contact recovery is colored by relative exposure and burial based on predicted neighbor count. Continues on next page.

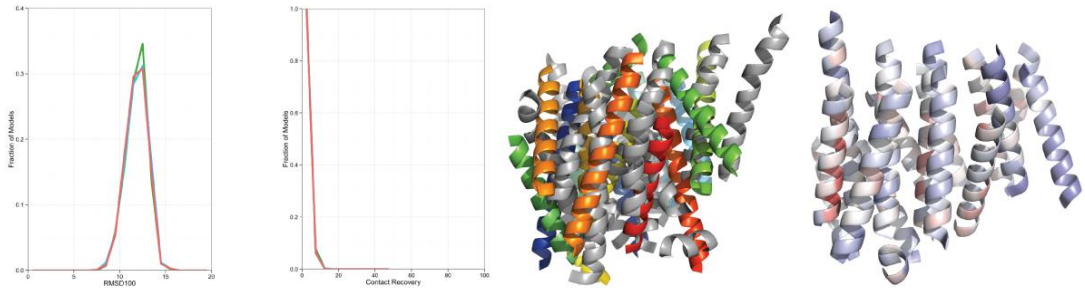
2NR9



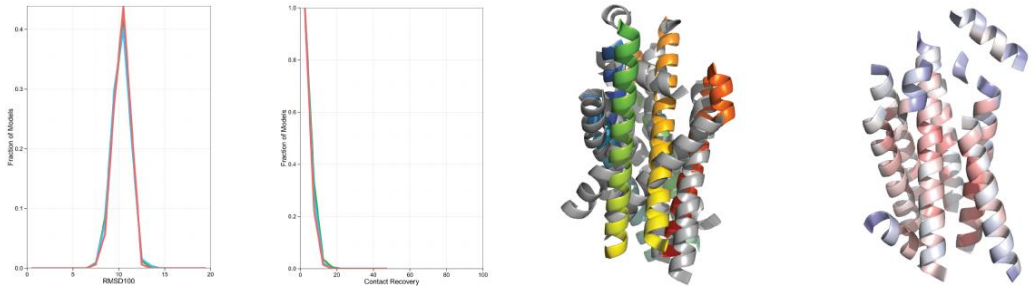
2XQ2



2XUT

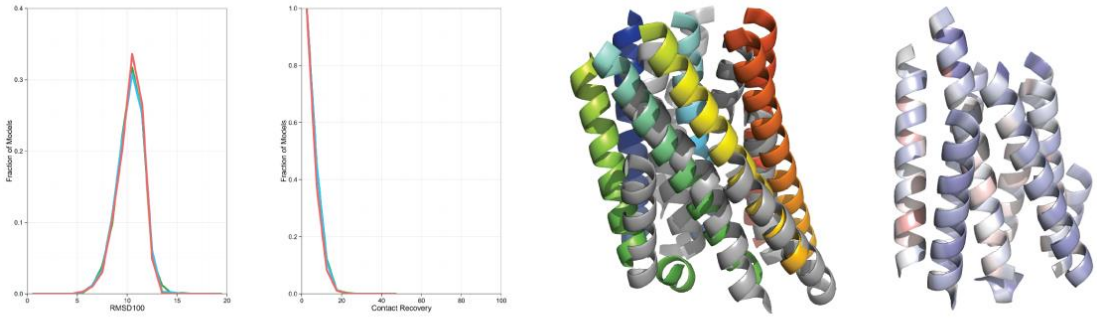


3KCU

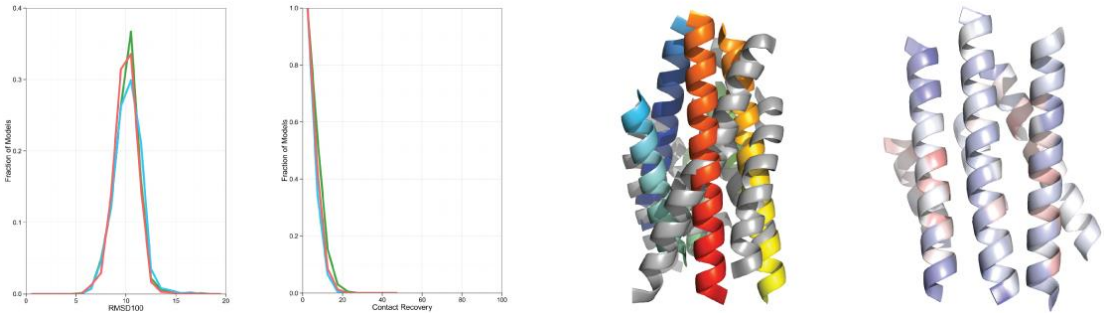


(Figure 69, continued from the previous page) The fraction of models at a range of RMSD100 values and contact recovery is shown for models predicted without neighbor count score (red), with predicted neighbor count (blue) and with calculated neighbor count (green). The top model ranked by RMSD100 is shown in rainbow compared to the native structure in grey. The top model ranked by contact recovery is colored by relative exposure and burial based on predicted neighbor count. Continues on next page.

3KJ6



3P5N



(Figure 69, continued from the previous page) The fraction of models at a range of RMSD100 values and contact recovery is shown for models predicted without neighbor count score (red), with predicted neighbor count (blue) and with calculated neighbor count (green). The top model ranked by RMSD100 is shown in rainbow compared to the native structure in grey. The top model ranked by contact recovery is colored by relative exposure and burial based on predicted neighbor count.

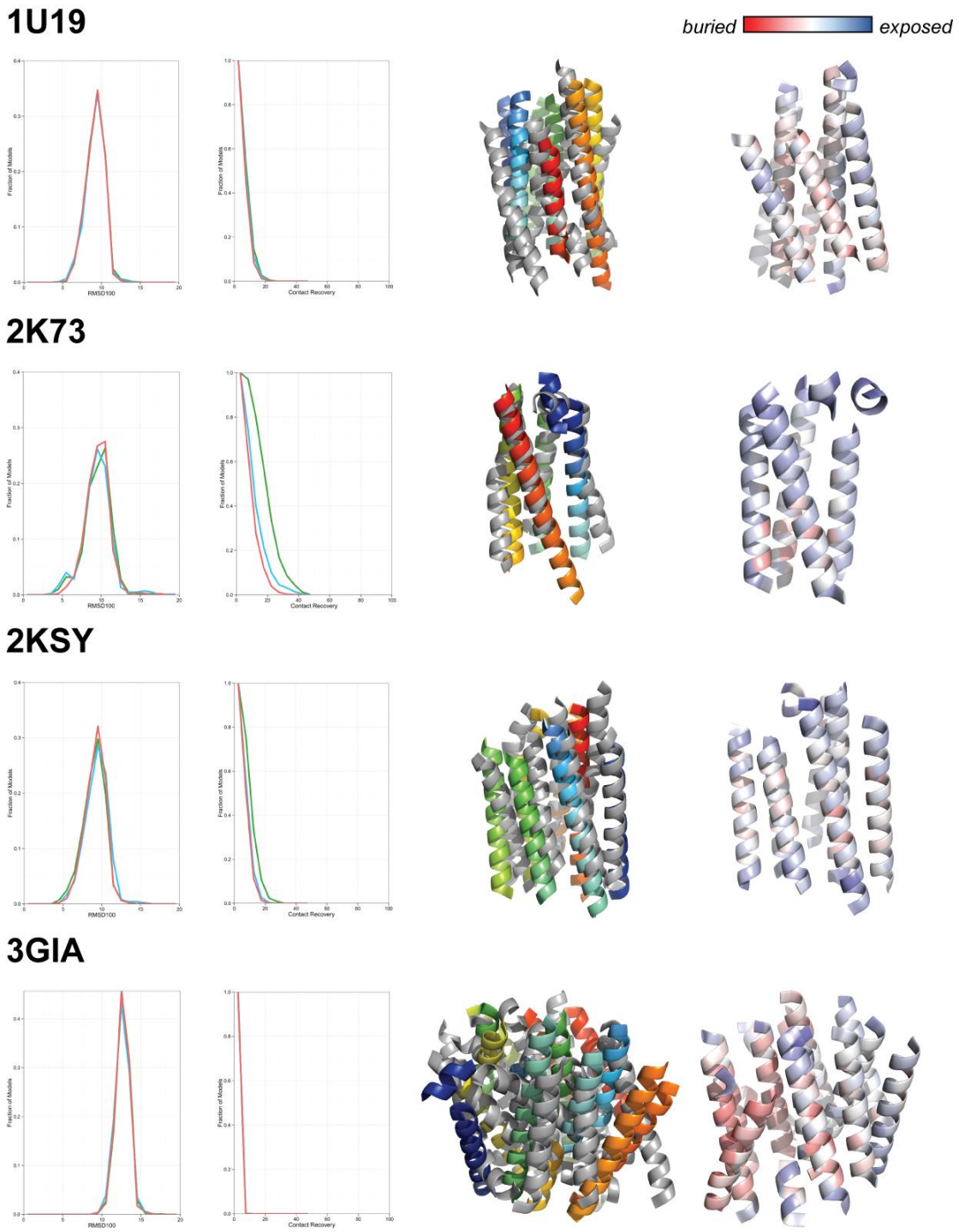


Figure 73
Gallery of results for 4 membrane proteins in the benchmark dataset that were also in the training dataset.
The fraction of models at a range of RMSD100 values and contact recovery is shown for models predicted without neighbor count score (red), with predicted neighbor count (blue) and with calculated neighbor count (green). The top model ranked by RMSD100 is shown in rainbow compared to the native structure in grey. The top model ranked by contact recovery is colored by relative exposure and burial based on predicted neighbor count.

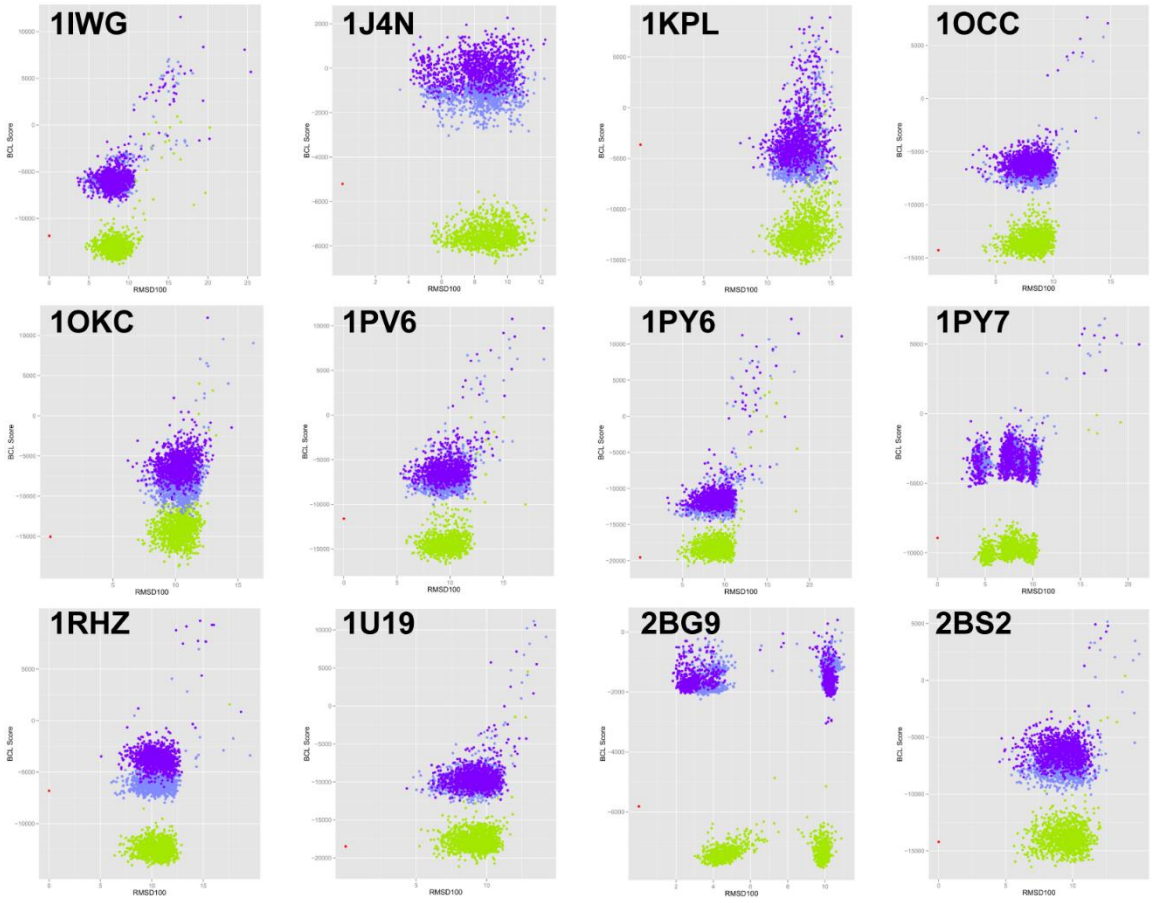
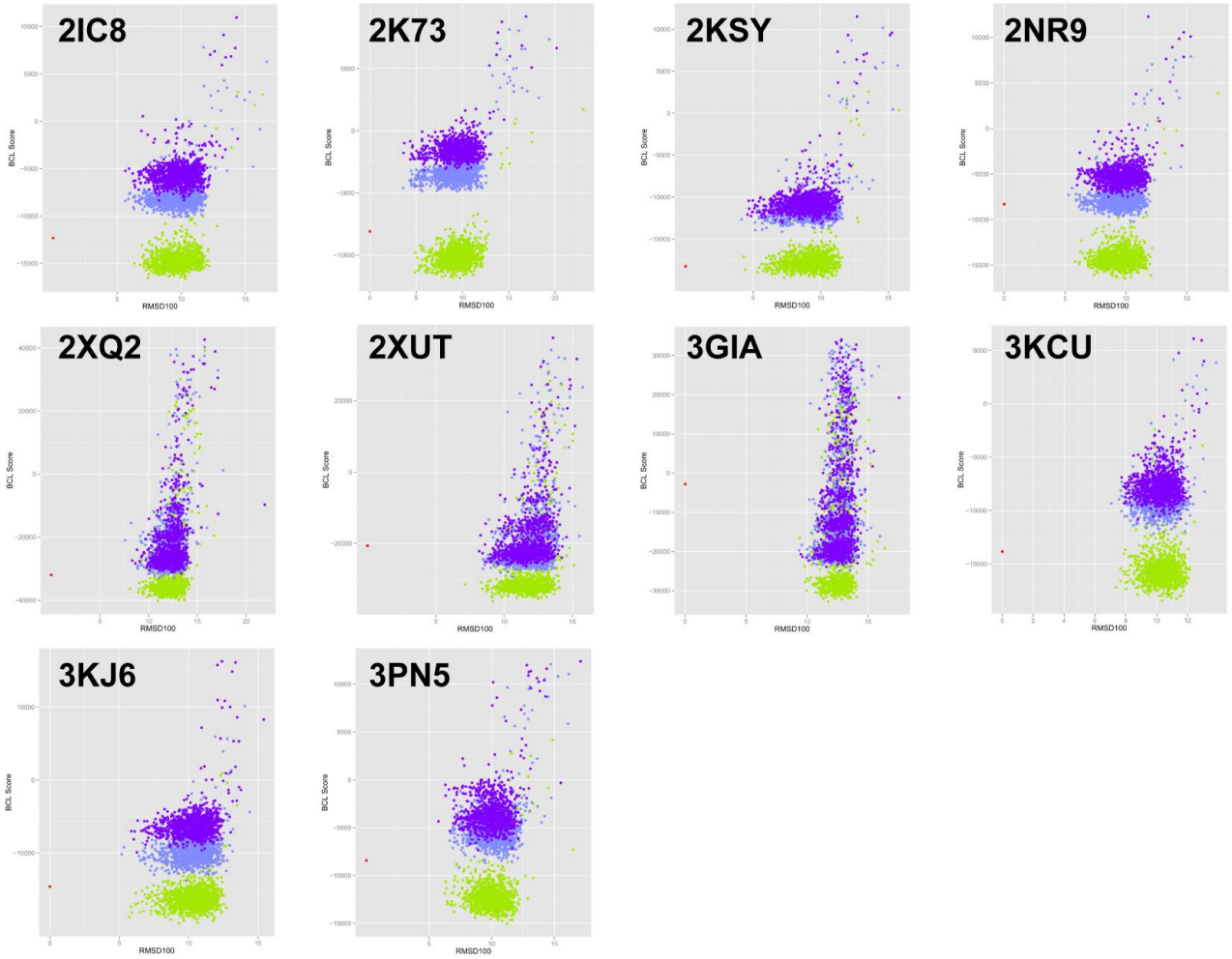


Figure 74
Score versus RMSD100 plots for the 22 membrane proteins in the benchmark dataset with and without the neighbor count score. Models folded without the neighbor count score are shown in green, models folded with predicted neighbor count values are in light purple and models folded with calculated neighbor count values are in dark purple. The score of the native protein is shown in red. Continued on the next page.



(Figure 71, continued from previous page) Models folded without the neighbor count score are shown in green, models folded with predicted neighbor count values are in light purple and models folded with calculated neighbor count values are in dark purple. The score of the native protein is shown in red.

Protocol capture

This protocol capture contains the steps necessary to obtain the results presented in Chapter VII "Incorporation of sequence-based exposure prediction in *de novo* membrane protein structure prediction". All input, output and source files can be found at 2013-05-30_bcl_exposure. The BCL software suite is publically available and the license is free for non-commercial users at http://www.meilerlab.org/bclcommons. The particular version of the BCL used for this project is found at 2013-05-30_bcl_exposure/bin/bcl-all-static.exe. Where you see "PDB", this will always indicate that the PDB ID of a particular protein should be inserted at that point.

1. Generation of ANN training datasets

Step	Text	Commands	Comment
1A. Prepare input files for training dataset of membrane proteins.	A set of 177 membrane proteins from the Protein Data Bank (PDB) having less than 30% sequence identity and 3 Å or lower resolution was used to train the method presented in this work. This training set is the same as the set used to train the JUFO9D secondary structure and transmembrane span prediction method.	See (Leman et al., 2013) for the details on how this membrane protein training dataset was set up. A list of the chains in this training dataset is included here.	Output: `fastalist_dball_MP.ls`

1B. Calculate per-residue neighbor count output value for training dataset.	The environment potential for each residue in the training set was evaluated using the BCL by calculating its number of neighbors.	1. Generate neighbor count for each PDB.pdb:
		`bcl.exe protein:Score -pdb PDB.pdb -weight_set nc.scoreweights -detailed -logger File PDB.log`
		
		2. Extracts neighbor counts from log file:
		`perl /bin/print_nc.pl  PDB > PDB_nc.ls`
		
		3. Combine output for all proteins into a single csv file:
		`ls PDB_nc.ls > nc_files.ls`
		`perl bin/print_msms_file.pl nc_files.ls > nc_all.csv`
			Input:
			pdb files for each protein in the training dataset,
			`nc.scoreweights`
			
			Output:
			`nc_all.csv`

2. Training the ANN with cross-validation and descriptor selection

Step	Text	Commands	Comment
2A. Generate input descriptors for ANN.	As input to the ANN training, sequence-specific descriptors were calculated over a window of 16 residues across each sequence in the database.	`bcl.exe descriptor:GenerateDataset -source 'SequenceDirectory([path to training dataset],suffix=.pdb,key file=fastalist_dball_MP.ls)' -feature_labels code_input.obj -result_labels code_output_nc.obj -id_labels code_id.obj -output exposure_MP_dataset_nc.bin -scheduler PThread 8 -logger File exposure_MP_dataset_nc.log`	Input: `fastalist_dball_MP.ls, nc_all.csv, code_id.obj, code_input.obj, code_output_nc.obj, path to training dataset` Output: `exposure_MP_dataset_nc.bin`
2B. Iteratively train ANN with 5-fold cross-validation and descriptor selection.	A total of thirty rounds of iterative ANN training took place. After each round of training, each descriptor was given a score based on its relative importance for predicting the output value. Each round of training was evaluated by its prediction accuracy over the independent set. Prediction accuracy increased until 24 rounds of training had taken place.	`bin/BclANNDescriptorSelectionByConnectionWeights.py -m 0 4 -i 0 4 -n 5 -d exposure_MP_dataset_nc.bin -l 'NeuralNetwork(transfer function = Sigmoid, weight update = Resilient(min change=0.001,max change=0.1), objective function = RMSD, steps per update=0, hidden architecture(400,50), shuffle = False, modulate error=True, max dropout=0.8)' --final-objective-function RMSD --show-status --max-minutes 75 --max-iterations 1000 --results code_output_nc.obj --id NNW_400_50 --max_unimproved_iterations 20 --local --ds-rounds 30 --attrition-rate 0.1`	Input: `exposure_MP_dataset_nc.bin, code_output_nc.obj` Output: The script outputs the following directories: `feature-scores, log_files, models, results` Results for NNW_400_50_rnd24 are given in the input_output directory.

3. Incorporating neighbor count into BCL::MP-Fold

Step	Text	Commands	Comment
3. Optimize weights for neighbor count score.	The weight for the neighbor count score was optimized as a function of enrichment over contact recovery and was determined to be 947.	1. Generate a decoy set of models by perturbing a set of membrane proteins. The benchmark set of 18 MP was used here. 2. Score the decoys set of proteins `awk '{print "bcl.exe protein:Score -membrane -histogram_path ~/bcl/histogram/ -quality RMSD RMSD_XYSuperimposition GDT_TS -message_level Critical -score_table_write "$1"_exposure.scores -pdblist "$1"A_pdbs.ls -native "$1".pdb -pool "$1".SSPredMC_OCTOPUS.pool -tm_helices "$1".SSPredMC_OCTOPUS.pool -sspred JUFO9D OCTOPUS -contact -score_exposure -sequence_data [path to database] "$1" -scheduler PThread 8}' ../benchmark_pdbs.ls`  3. Minimize weights for exposure scores. `bcl.exe MinimizeScoreWeightSet -list bcl_perturb_tables.ls -weight_set weights_bcl_perturb.table -weight_set_write minimized_0_cr12_50_20p_200_ -enrichment cr12 0.20 0.1 10 100 -sort_order greater -mc_tot_unimproved 10000 500 -number_repeats 1 -keep_positive -random_seed 1000 -number_weights_mutated 1`	Input: Decoy models for benchmark dataset pdbs, path to benchmark dataset, `benchmark_pdbs.ls`, `weights_bcl_perturb.table` Output: `bcl_perturb_tables.ls`, `minimized_0_cr12_50_20p_200_0.weights`

4. Structure prediction a benchmark set of MP with neighbor count score

Step	Text	Commands	Comment
4A. Prepare a benchmark dataset of 22 MP for structure prediction.	A database of 22 diverse single-chain membrane proteins, from the same benchmark set as that used for BCL::MP-Fold, was used to evaluate the influence of the neighbor count score on *de novo* protein structure prediction.	See [citation] for details on the benchmark dataset. A list of the chains in this benchmark dataset is included here.	Output: `benchmark_pdbs.ls`

4B. Calculate per-residue neighbor count output value for benchmark dataset.		1. Generate neighbor count for each PDB.pdb:
		`bcl.exe protein:Score -pdb PDB.pdb -weight_set nc.scoreweights -detailed -logger File PDB.log`
		
		2. Extracts neighbor counts from log file:
		`perl /bin/print_nc.pl PDB > PDB_nc.ls`
		
		3. Combine output for all proteins into a single csv file:
		
		`ls PDB_nc.ls > nc_files.ls`
		`perl bin/print_msms_file.pl nc_files.ls > nc_benchmark_fasta.csv`
			Input:
			pdb files for each protein in the benchmark dataset,
			`nc.scoreweights`
			
			Output:
			`nc_benchmark_fasta.csv`

4C. Calculate neighbor count prediction for benchmark dataset.	A database of 22 diverse single-chain membrane proteins, from the same benchmark set as that used for BCL::MP-Fold, was used to evaluate the influence of the neighbor count score on *de novo* protein structure prediction.	1. Change `models/model.result` to: `Combine(Mapped(file=nc_benchmark_fasta.csv,key=Combine(ProteinId,AASeqID),delimiter=","))` 2. Predict neighbor count with ANN model `bcl.exe descriptor:GenerateDataset -source 'SequenceDirectory([path to benchmark dataset],suffix=.fasta,key file=benchmark_pdbs.ls)' -feature_labels code_input_nc_predict.obj -result_labels code_output_benchmark_nc.obj -id_labels code_id.obj -output nc_benchmark_predict.csv -logger File predict_nc_benchmark_pdbs.log`  3. Extract csv file to individual exposure prediction files for each pdb. `awk '{print "grep "$1" nc_benchmark_predict.csv | awk -F\",\" \x27\{print $2\"\\t\"$5\}\x27 > "substr($1,2,2)"/"$1".exposure"}' benchmark_pdbs.ls` 4. Move the exposure prediction files to the directory of the benchmark dataset	Input: `nc_benchmark_fasta.csv, benchmark_pdbs.ls, code_id.obj, code_input_nc_predict.obj, code_output_benchmark_nc.obj,` path to benchmark dataset Output: `nc_benchmark_predict.csv, PDB.exposure` files for each PDB in the benchmark dataset

4D. Fold MP without neighbor count score.	Proteins were predicted 1000 each with and without the neighbor count score to evaluate the influence of the score.	`awk '{print "perl bin/create_fold_pbs_noex.pl 1 20 50 "substr($1,1,4)" . 02"}' ../ benchmark_pdbs.ls`
		
		`sh PDB/build_02/pbs/qsub_pbs.sh`
			Input:
			`benchmark_pdbs.ls, pred_stages_noex.txt assembly_01_noex.scoreweights, assembly_02_noex.scoreweights, assembly_03_noex.scoreweights, assembly_04_noex.scoreweights, assembly_05_noex.scoreweights, refinement_noex.scoreweights`
			
			Output:
			1000 models for each PDB

4E. Fold MP with neighbor count score.	Proteins were predicted 1000 each with and without the neighbor count score to evaluate the influence of the score.	`awk '{print "perl bin/create_fold_pbs.pl 1 20 50 "substr($1,1,4)" . 06"}' ../ benchmark_pdbs.ls`
		
		`sh PDB/build_06/pbs/qsub_pbs.sh`
			Input:
			`benchmark_pdbs.ls, pred_stages.txt assembly_01.scoreweights, assembly_02.scoreweights, assembly_03.scoreweights, assembly_04.scoreweights, assembly_05.scoreweights, refinement.scoreweights`
			
			Output:
			1000 models for each PDB

4F. Evaluate enrichment of contact recovery for benchmark dataset with ANN model.	Over a set of 18 membrane proteins in the benchmark set, the neighbor count score was shown to enrich for the selection of native-like models (under 8 Å) on average 1.2 fold.	Calculate enrichment on the set of proteins built without the neighbor count score from 4D.
		
		`awk '{print "bcl.exe protein:Score -pdblist build_02_pdbs_"$1".ls -detailed -score_table_write "$1"_score_exp1000.sc -native "$1".pdb -quality RMSD -enrichment 0.1 8.0 10 RMSD100 less enrichments_"$1"_score_exp1000.tbl -weight_set score_exp1000.weights -contact -membrane -sspred JUFO9D OCTOPUS -score_exposure -sequence_data [path to benchmark database] "$1" -pool "$1".SSPredMC_OCTOPUS.pool -tm_helices "$1".SSPredMC_OCTOPUS.pool"}' benchmark_pdbs.ls > score_exp1000.sh`
			Input:
			List of models built from step 4D for each PDB:
			`build_02_pdbs_PDB.ls score_exp1000.weights`
			
			Output:
			`enrichments_PDB_score_exp1000.tbl`

5. Analyze results

Step	Text	Commands	Comment
5. Analyze models based on score, RMSD100 and contact recovery.	The quality of protein structure prediction was evaluated with RMSD100 and contact recovery.	1. Get score files from models. Run this with “build_02” and “build_06”. `bcl.exe protein:FoldAnalysis -table_from_prefix build_06 -protein_storage ./PDB/pdbs/ -sort RMSD100 -output_table ./PDB/scores_06.ls -message_level Critical` 2. Pull out best models by overall score, contact recovery or RMSD100. `touch PDB_best_06.txt` `sort -nk21 PDB/scores_06.ls | awk '{print "best_exposure_score: "$1" "$23" "$21" "$32" "$28}' | head -10 >> PDB_best_06.txt` `sort -rnk28 PDB/scores_06.ls | awk '{print "best_contact_recovery: "$1" "$23" "$21" "$32" "$28}' | head -10 >> PDB_best_06.txt` `sort -nk32 PDB/scores_06.ls | awk '{print "best_RMSD100: "$1" "$23" "$21" "$32" "$28}' | head -10 >> PDB_best_06.txt` `sort -nk23 PDB/scores_06.ls | awk '{print "best_overall_score: "$1" "$23" "$21" "$32" "$28}' | head -10 >> 2KSF_best_06.txt` `more PDB_best_06.txt` 3. Generate distribution plots for contact recovery and RMSD100 with gnuplot. `bcl.exe bcl:DistributionPlotFromTables -gnuplot_input_table_filenames ./PDB/scores_02.ls ./PDB/scores_06.ls -gnuplot_table_columns_x RMSD100 -histogram_min 0 -histogram_binsize 1 -histogram_numbins 20 -gnuplot_output_filename ./pngs/PDB_dist_02_06_RMSD100.gnuplot -gnuplot_title PDB -gnuplot_x_label RMSD100 -gnuplot_series_names no_NC with_NC -gnuplot_set_key`  `gnuplot ./pngs/PDB_dist_02_06_RMSD100.gnuplot`  `gthumb ./pngs/PDB_dist_02_06_RMSD100.gnuplot.png &`  `bcl-all-static.exe bcl:DistributionPlotFromTables -gnuplot_input_table_filenames ./PDB/scores_02.ls ./PDB/scores_06.ls -gnuplot_table_columns_x cr12 -histogram_min 0 -histogram_binsize 5 -histogram_numbins 10 -gnuplot_output_filename ./pngs/PDB_dist_02_06_cr12.gnuplot -gnuplot_title PDB -gnuplot_x_label cr12 -gnuplot_series_names no_NC with_NC -gnuplot_set_key`  `gnuplot ./pngs/PDB_dist_02_06_cr12.gnuplot`  `gthumb ./pngs/PDB_dist_02_06_cr12.gnuplot.png &`	Input: List of models built from step 4D and 4E for each PDB Output: `/PDB/scores_02.ls,` `/PDB/scores_06.ls,` `PDB_best_02.txt,` `PDB_best_06.txt,` `/pngs/PDB_dist_02_06_RMSD100.gnuplot.png` `/pngs/PDB_dist_02_06_cr12.gnuplot.png`

IDENTIFICATION OF DETERMINANTS REQUIRE FOR AGONISTIC AND INVERSE AGONISTIC LIGAND PROPERTIES AT THE ADP RECEPTOR P2Y12

This work is based on publication (Schmidt et al., 2013).

Summary

The ADP receptor $P2Y_{12}$ belongs to the superfamily of G protein-coupled receptors (GPCRs) and its activation triggers platelet aggregation. Therefore, potent antagonists, e.g. clopidogrel, are of high clinical relevance in prophylaxis and treatment of thromboembolic events. $P2Y_{12}$ displays an elevated basal activity *in vitro* and as such, inverse agonists may be therapeutically beneficial compared to antagonists. Only a few inverse agonists of $P2Y_{12}$ have been described. To expand this limited chemical space and improve understanding of structural determinants of inverse agonist-receptor interaction, we screened an adenosine nucleotide library for lead structures using wild-type (WT) human $P2Y_{12}$ and 28 constitutively active mutants. We found that ATP and ATP derivatives are agonists at $P2Y_{12}$. The potency at $P2Y_{12}$ was 2-(methylthio)-ADP > 2-(methylthio)-ATP > ADP > ATP. Determinants required for agonistic ligand activity were identified. Molecular docking studies revealed a binding pocket for the ATP derivatives that is bordered by transmembrane helices 3, 5, 6 and 7 in the human $P2Y_{12,}$ with Y^{105}, E^{188}, R^{256}, Y^{259} and K^{280} playing a particularly important role in ligand interaction. N-methyl-anthraniloyl modification at the 3' OH of the 2'deoxiribose leads to ligands (mant-dATP, mant-dADP) with inverse agonist activity. Inverse agonist activity of mant-dATP was found at the WT human $P2Y_{12}$ and half of the constitutive active $P2Y_{12}$ mutants. Our study showed that, in addition to ADP and ATP, other ATP derivatives are not only ligands of

$P2Y_{12}$ but also agonists. Modification of the ribose within ATP can result in inverse activity of ATP-derived ligands.

Introduction

The ADP receptor $P2Y_{12}$ is a Gi protein-coupled receptor and a key player in platelet aggregation (Hollopeter et al., 2001). Inactivating mutations in $P2Y_{12}$ are responsible for bleeding disorders in humans and dogs (Boudreaux & Martin, 2011, Cattaneo et al., 2003, Cattaneo, 2005, Daly et al., 2009, Fontana, Ware, & Cattaneo, 2009, Hollopeter et al., 2001, Remijn et al., 2007, Shiraga et al., 2005). With significant relevance in pathophysiology, $P2Y_{12}$ is also the major target of the antithrombotic drugs ticlopidin and clopidogrel. The thienopyridine clopidogrel is a prodrug which requires the cytochrome P450 2C19 enzyme for its conversion to an active thiol metabolite. Several mechanisms of antagonistic action have been proposed for the active metabolite of clopidogrel, including interaction with extracellular cysteine residues of $P2Y_{12}$ (Ding, Kim, Dorsam, Jin, & Kunapuli, 2003) and receptor dimer disruption (Pierre Savi et al., 2006). Although very specific and effective, clopidogrel produces a variable platelet inhibition based on genetic polymorphisms and drug interactions (Muñoz-Esparza et al., 2011, Nawarskas & Clark, 2011). This has triggered the search for alternative $P2Y_{12}$ blockers such as prasugrel, cangrelor and ticagrelor. The latter two compounds are ATP analogues and bind reversibly at $P2Y_{12}$ (Storey, 2011).

$P2Y_{12}$ displays a high constitutive activity when expressed in vitro (Chee et al., 2008, Schulz & Schöneberg, 2003). Therefore, inverse agonists may be therapeutically beneficial compared to antagonists. Because only a few inverse agonists of $P2Y_{12}$ have been described (Ding, Kim, & Kunapuli, 2006), we, therefore, screened for compounds that reduce the basal activity of constitutively active $P2Y_{12}$ mutants.

Functional characterization of P2Y receptors and their mutants in mammalian expression systems is problematic because of the abundance of endogenous nucleotide receptors, nucleosidases and nucleotide release. In previous experiments, we and others demonstrated that the human $P2Y_{12}$ is functionally expressed in the yeast system (Pausch et al., 2004, Schulz & Schöneberg, 2003) which lacks such problems. Numerous constitutively activating mutations have been described for GPCRs in natural or recombinant systems, but only a few have been reported for P2Y receptors (Ding et al., 2006). From over 1000 single point mutations, we identified 28 constitutively active $P2Y_{12}$ mutants. Screening a purine compound library, we discovered several new agonists and inverse agonists for the WT $P2Y_{12}$ and constitutively active mutants, respectively.

Materials and Methods

Materials

If not stated otherwise, all standard substances were purchased from Sigma–Aldrich, Merck and Care Roth. Cell culture material was obtained from Sarstedt. MeS-ADP (2-(methylthio)adenosine 5´-diphosphate trisodium salt hydrate), salmon sperm DNA, lithium acetate (LiAc) dehydrate, lithium chloride (LiCl), PEG 3350 (cat. no. P-3640) and apyrase from potatoe (Grade III) were obtained from Sigma-Aldrich (Munich, Germany). MeS-ADP ($P2Y_{12}$ agonist) was dissolved in water and aliquots of stock solutions (10 mM) were stored at -20 °C. Yeast medium components were purchased from Sigma-Aldrich (Munich, Germany) and from BD Biosciences (Heidelberg, Germany). Restriction enzymes were purchased from New England Biolabs (Frankfurt a. M., Germany), primers were synthesized by Life Technologies (Darmstadt, Germany) and $P2Y_{12}$ mutant libraries were provided by Sloning BioTechnology (Puchheim, Germany).

The adenine nucleotide library was from Jena Bioscience (Jena, Germany). For compound details see http://www.jenabioscience.com/images/7c63e6fc71/LIB-101.pdf.

Generation of $P2Y_{12}$ mutants

Mutants were generated by subcloning SlonoMax-SINGLE libraries (synthesized double-stranded DNA fragments containing individual mutants, fragment sizes 100 – 150 bp) via unique endogenous or silently introduced restriction sites. $P2Y_{12}$ mutants were introduced into the yeast expression plasmid p416GPD (provided by Dr. Mark Pausch, Wyeth Research, Princeton, NJ, U.S.A.) and transformed into *E. coli* DH5α (Life Technologies, Darmstadt, Germany). Plasmids from individual clones were isolated (plasmid preperation kit; Promega, Mannheim, Germany) and mutations were identified by DNA sequencing. Because full coverage was not achieved after sequencing 96 clones, missing mutants (4 mutants/position on average) were generated by PCR-based site-directed mutagenesis using mutant-specific mutagenesis primers.

Expression and functional testing of $P2Y_{12}$ mutants in yeast and mammalian cells

The *Saccharomyces cerevisiae* yeast strain MPY578t5 (provided by Dr. Mark Pausch) was used for yeast expression and functional testing of the $P2Y_{12}$ mutants. Cells were transformed with plasmid DNA using the LiAc/salmon sperm carrier DNA/PEG method. In brief, an overnight culture grown at 30 °C in YPAD (yeast extract, peptone, dextrose medium with adenine) was diluted to an optical density of 0.2 at 600 nm ($OD_{600\ nm}$) in 50 ml YPAD. This culture was incubated at 30 °C until the $OD_{600\ nm}$ reached 0.7-0.9. Cells were then harvested (2,500 g for 5 min at room temperature) and washed once with 25 ml of water. The pellet was dissolved in 700 µl of LiAc (100 mM) and incubated for 10 min at 30 °C. A pellet of 50 µl from the yeast cell suspension was then mixed with 90 µl

PEG 3350 (50 % w/v), 13.5 µl LiAc (1 M), 18.75 µl salmon sperm carrier DNA (2.0 mg/ml), 2.75 µl of sterile water and plasmid DNA (1 µg) before being incubated for 30 min at 30 °C, then for 30 min at 45 °C.

For selection of constitutively active clones, cells were plated on agar plates not containing uracil and histidine (U^-/H^-). After incubation at 30 °C for 4 days, clones were prepared for concentration-response curves. Cells transformed with $P2Y_{12}$ mutants were pre-cultured for 2 days at 30 °C in U^-/H^- with 10 µM MeS-ADP. To remove MeS-ADP, cells were washed twice with water and grown in U^-/H^- overnight without MeS-ADP. The yeast cell suspension was then diluted to an $OD_{600\,nm}$ of 0.1. From this cell suspension, 100 µl were pipetted into each well of a 96-well plate and to this, 100 µl of a 2x agonist solution or medium was added. Background growth was suppressed by the addition of 20 mM 3-aminotriazole. Mutants were screened for growth and/or constitutive activity at 10 µM MeS-ADP. All positive mutants were further evaluated through MeS-ADP concentration-response (growth) curves.

The adenine nucleotide compound library was screened for agonists and inverse agonists at the WT $P2Y_{12}$ and constitutively active mutants. 100 µl of the respective yeast cell suspension (100 µl, $OD_{600\,nm}$ = 0.1) was pipetted into each well of a 96-well plate and, to these samples, 100 µl of a 2x ligand solution or medium was added. OD measurements were performed 24 and 48 hours later. Compounds identified to be agonists or inverse agonists were further characterized in concentration response setups. IC_{50} and EC_{50} values were calculated using GraphPadPrism 4 software (GraphPad).

To determine the stability of ATP in the 24-h yeast assay we performed a phosphoenolpyruvate (PEP)/pyruvate kinase test. Thus, 2 mM PEP and 2 µl of a pyruvatkinase solution (final 6 U/ml) were added every 5 h to the yeast growth U^-/H^- medium (200 µl total volume additional containing 100 mM imidazol, 5 mM $MgCl_2$, pH = 7.15 to assure proper pyruvatkinase function). In case ADP is formed due to degrading,

PEP is utilized by the pyruvate kinase to produce ATP. Then, the concentration of PEP in the medium was monitored over 24 h using a coupled optical enzyme-test. Thus, 50 µl yeast medium harvested after 0 h, 4 h and 24 h (or 1 mM phosphoenolpyruvate for control purposes) were incubated with 0.7 ml assay buffer (100 mM imidazol, 5 mM $MgCl_2$, pH = 7.15), 1.5 µl lactate dehydrogenase (10 U/ml), 1 µl pyruvate kinase (10 U/ml), 8 µl of a 100 mM ADP solution and 8 µl 2 mM NADH. NADH concentration was determined photometrically at 340 nm.

For expression in mammalian cells, Chinese hamster ovary (CHO)-K1 cells were grown in Dulbecco's modified Eagle's medium (DMEM supplemented with 10 % (v/v) fetal bovine serum, 100 units/ml penicillin and 100 µg/ml streptomycin) at 37 °C in a humidified 5 % CO_2 incubator. A CHO-K1 cell line stably expressing the chimeric G protein $G\alpha_{qi4}$ (Kostenis, Waelbroeck, & Milligan, 2005) was established. Transient transfection experiments of CHO-K1 cells with the respective $P2Y_{12}$ constructs and inositol phosphate (IP_1) accumulation assays were performed as described (Schulz & Schöneberg, 2003). In brief, $G\alpha_{qi4}$-stable cells were seeded into twelve-well plates (about 0.15×10^6 cells/well), transient transfected and about 48 h after this labeled with 2 µCi/ml of [*myo*-^{3}H]inositol (Perkin Elmer). After a 16 h-labeling period, cells were washed once with serum-free DMEM containing 10 mM lithium chloride (LiCl) and then incubated for 60 min at 37 °C with serum-free DMEM containing 10 mM LiCl with or without a compound. After this time, the assay medium was removed, and the reaction was stopped by adding 0.3 ml of 0.1 N NaOH, followed by a 5 min-incubation at 37 °C. The alkaline solution was then neutralized by adding 0.1 ml of 0.2 M formic acid, and the IP_1 fraction was isolated by anion exchange chromatography as described (Berridge, 1983) and counted on a liquid scintillation counter.

For cAMP assays, transfected cells were labeled with [^{3}H]adenine (2 µCi/ml; PerkinElmer) for 12 h and washed once in serum-free DMEM containing 1 mM 3-isobutyl-

1-methylxanthine (Sigma–Aldrich), followed by incubation in the presence of the indicated compounds and forskolin (10 μM) for 1 h at 37°C. Reactions were terminated by aspiration of the medium and addition of 1 ml of 5% (w/v) trichloroacetic acid. The cAMP content of cell extracts was determined by anion exchange chromatography as described (Salomon, Londos, & Rodbell, 1974).

To measure label-free receptor activation, a dynamic mass redistribution (DMR) assay (Corning Epic® Biosensor Measurements) with stably transfected Human Embryonic Kidney (HEK) cells (HEK-FlpIn, $P2Y_{12}$ in pcDNA5/FRT) was performed as described previously (Lars et al., 2012, Schröder et al., 2010). Briefly, cells were seeded into fibronectin-coated Epic 384-well microplates (60,000 cells per well) and exposed to the various compounds. In DMR measurements, polarized light is passed through the bottom of the biosensor microtiter plate, and a shift in wavelength of reflected light is indicative of intracellular mass redistribution triggered by receptor activation. DMR was recorded as a measure of cellular activity over 60 min. Agonist-induced DMR is concentration-dependent and concentration-effect curve were calculated from response peak maxima (approximately 6 min after adding the compound) of optical traces.

Generation of a $P2Y_{12}$ comparative model and ligand docking

A comparative model of $P2Y_{12}$ was constructed using the protein structure prediction software package, Rosetta version 3.2 (Leaver-Fay et al., 2011). The X-ray crystal structure of CXCR4 (Protein Data Bank ID: 3ODU) (Gupta, Pillarisetti, Thomas, & Aiyar, 2001) was chosen as a template based on its high similarity to $P2Y_{12}$ (e-value of $3e^{-15}$ with a sequence coverage of 90%) according to a search using NCBI BLASTP on sequences from the Protein Data Bank (PDB) (Figure 81). CXCR4 and $P2Y_{12}$ also share a conserved disulfide bond between the N-terminal C17 and C270 in extracellular loop 3 (Deflorian & Jacobson, 2011). The backbone coordinates of CXCR4 were retained in the

comparative model of $P2Y_{12,}$ while the loop coordinates were built in Rosetta using Monte Carlo Metropolis (MCM) fragment replacement combined with cyclic coordinate descent loop closure. In brief, ϕ-ψ angles of backbone segments from homologous sequence fragments from the PDB are introduced into the loop regions. After the fragment substitution, small movements in the ϕ-ψ angles are performed to close breaks in the protein chain. The resulting full sequence models were subjected to eight iterative cycles of side chain repacking and gradient minimization of ϕ, ψ, and χ angles Rosetta Membrane (Yarov-Yarovoy et al., 2005).

Ligand docking into the comparative model of $P2Y_{12}$ with ADP, ATP, MeS-ADP, MeS-ATP, mant-ADP, mant-ATP, mant-dADP and mant-dATP was performed with Rosetta Ligand (Davis & Baker, 2009, Meiler & Baker, 2006). Each ligand was allowed to sample docking poses in a 5 Å radius centered at the putative binding site for ADP, determined by averaging the coordinates of critical residues for ligand recognition: R^{256}, Y^{259} and K^{280} (Hoffmann, Sixel, Di Pasquale, & von Kügelgen, 2008). Once a binding mode had been determined by the docking procedure, 100 conformations of the ligand created by MOE (Molecular Operating Environment, Chemical Computing Group, Ontario, Canada) were tested within the site. Side-chain rotamers around the ligand were optimized simultaneously in a MCM simulated annealing algorithm. The energy function used during the docking procedure contains terms for van der Waals attractive and repulsive forces, statistical energy derived from the probability of observing a side-chain conformation in the PDB, hydrogen bonding, electrostatic interactions between pairs of amino acids, and solvation assessing the effects of both side-chain/side-chain interactions and side-chain/ligand interactions. For each ligand, over 3,000 docked complexes were generated and clustered for structural similarity using bcl::Cluster (Alexander et al., 2011). The lowest energy binding mode from the five largest clusters for each ligand were used

for further analysis. The change in free energy with and without ligands bound to $P2Y_{12}$ was calculated for each residue in the receptor. Residues with the greatest difference in predicted energy are suggested to be important for ligand interaction (Figure 82).

Results

Expression of the human $P2Y_{12}$ in yeast and determination of agonist specificity

In previous experiments, we and others have already demonstrated that human $P2Y_{12}$ can be functionally expressed in the yeast system. In this system, $P2Y_{12}$-expressing yeast grows in 96-well cell plates and regular OD measurements are taken. OD values measure cell growth, which is used as a strong indicator for receptor activity.

The WT $P2Y_{12}$ was functionally tested with the compound library for agonists. $P2Y_{12}$ expressed in yeast showed a similar EC_{50} value for MeS-ADP (EC_{50} value: 6 nM, Figure 75) as when expressed in mammalian cells ranging from low nanomolar to 25-80 nM concentrations (Bodor et al., 2003, Ding et al., 2006, Hoffmann et al., 2008, Simon et al., 2002, F. L. Zhang et al., 2001, Zhong, Kriz, Seehra, & Kumar, 2004). ADP was about 500 fold less potent to MeS-ADP, which is consistent with previous findings showing an about 30 fold to 1000 fold lower potency in mammalian cells (Bodor et al., 2003, Simon et al., 2002, F. L. Zhang et al., 2001). In addition to the highly potent agonist MeS-ADP, we identified further $P2Y_{12}$ agonists: ATP and MeS-ATP. ATP was a partial agonist at the human $P2Y_{12}$ when expressed in yeast (Figure 75A). EC_{50} values were ranked: MeS-ADP < MeS-ATP < ADP < ATP. We considered the possibility that the registered ATP activity might be due to the fraction of the nucleotides converted to ADP derivatives by nucleotidases or hydrolysis and quantified the possible decay of ATP during the 24-h assays. Thus, we indirectly quantified ATP degradation in the assay using the

PEP/pyruvate kinase system. The pyruvate kinase catalyzes the transfer of a phosphate group from PEP to ADP, yielding one molecule of pyruvate and one molecule of ATP. PEP concentration in the medium is therefore a measure for degraded ATP (see Materials and Methods). We found that PEP concentration in the yeast medium remained almost unchanged during 24 h yeast growth ($\Delta E_{0\,h}$ = 0.44, $\Delta E_{4\,h}$ = 0.44, $\Delta E_{24\,h}$ = 0.48). Only 3.4 % of PEP (initial concentration 2 mM) were utilized by pyruvate kinase for ATP generation. This indicated high stability of ATP (96.6 %) in the assay over 24 h. The functionality of the pyruvate kinase to convert ADP to ATP was verified in control experiments performed in parallel.

The agonistic properties of the adenine nucleotides were verified in the mammalian cell line COS-7 and CHO cells (data not shown) where the human $P2Y_{12}$ was co-expressed with the chimeric G_{qi4} protein, which redirects receptor activation to the phospholipase C/inositol phosphate (IP) pathway (Kostenis et al., 2005). Since ATP produces a cellular response via endogenous nucleotide receptors in most cell lines we performed additional measurements of $P2Y_{12}$ activation on stably transfected mammalian HEK with a dynamic mass redistribution assay (Corning Epic® Biosensor Measurements) (Schröder et al., 2010). Responses of endogenous nucleotide receptors were subtracted from the specific response of $P2Y_{12}$-transfected cells. As shown in Figure 75B, the concentration-response curves were similar to the data from the yeast expression system except we found that ATP was a full agonist in this mammalian expression system. We also performed Epic measurements in $P2Y_{12}$-stably transfected astrocytoma cells 1321N1, which should not express nucleotide receptors (Filtz, Li, Boyer, Nicholas, & Harden, 1994). However, ATP-mediated responses in 1321N1 cells were less than in HEK cells having no advantage over transfected HEK cells. In sum, the yeast expression system is free of endogenous nucleotide receptors and, therefore, the most straight forward system to use in analyzing P2Y receptors. $P2Y_{12}$ expressed in yeast displays

pharmacological properties very comparable to mammalian expression systems. Our screening revealed additional compounds with agonistic activity at $P2Y_{12}$: ADPβS, TNP-ADP, ATPγS, 2I-ATPγS, and $AppNH_2$. We did not follow the pharmacology of these ADP and ATP derivatives further, but they support the fact that derivatives of ATP, as well as of ADP, also have agonistic activities at the human $P2Y_{12}$.

It should also be noted that a number of nucleotides and nucleosides (e.g. AMP, GTP, cAMP, adenosine) which do not activate $P2Y_{12}$ in mammalian expression systems (F. L. Zhang et al., 2001) did not activate $P2Y_{12}$ expressed in yeast. The comprehensive compound library allowed identification of all determinants necessary for agonist function at the $P2Y_{12}$ (Table 28). In detail:

1. The purine ring is absolutely required. Some modifications (methylthio iodine) at the adenine are tolerated, but guanine- and inosine-based nucleotides are not agonistic.
2. Deoxidation of the ribose is not tolerated. The trinitrophenyl modification (TNP-ADP) is tolerated.
3. Adenine nucleotides with two or three phosphate residues are agonistic, while < 2 phosphate residues or cyclic phosphates are insufficient for agonistic activity. Some substitutions of phosphate moieties as in ADPβS, ATPγS, and $AppNH_2$ are tolerated.
4. Purine multimers displayed no agonistic activity.

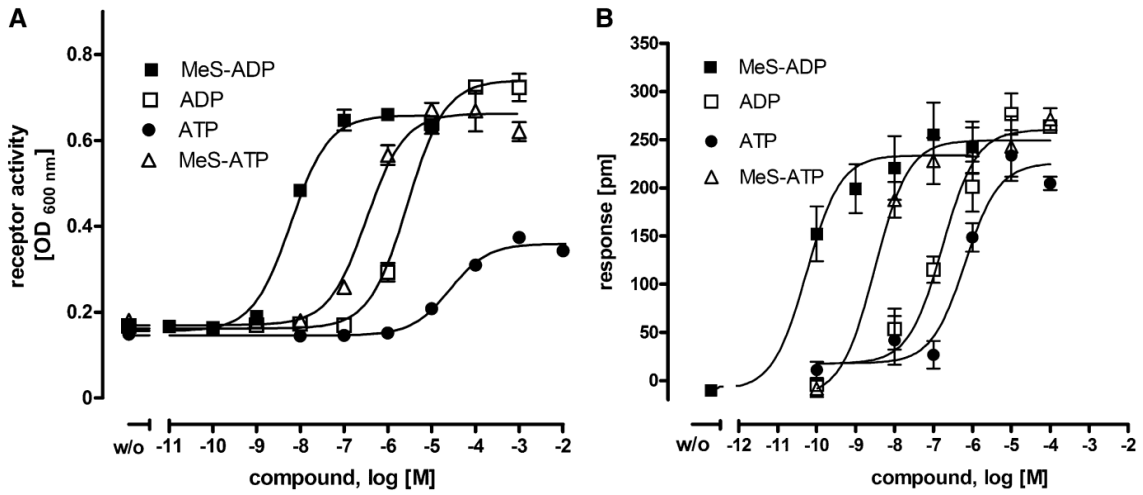


Figure 75
Nucleotide agonists at the human P2Y12.
A) The human P2Y12 was transformed into yeast cells and incubated with different concentrations of P2Y12 agonists. Receptor activation-dependent growth was measured as OD600 nm after 24 h. Data are given as mean ± SD of three independent experiments all performed in triplicate. B) For label-free measurements of receptor activation, a dynamic mass redistribution assay (Corning Epic® Biosensor Measurements) with stably transfected HEK cells was performed essentially as described previously (Schroder et al., 2010). The response is shown 6 min after compound application. The response of each compound at non-transfected HEK cells was subtracted from the respective response at P2Y12 transfected HEK cells. Data are presented as mean ± S.E.M. of two independent experiments, each carried out in triplicate.

Structural model of agonist binding

To estimate whether the different agonists may have similar binding properties, we simulated binding by docking the agonists into the comparative model of $P2Y_{12}$ (Figure 76 and Figure 80). The model suggested that ADP, ATP, MeS-ADP, MeS-ATP, mant-ADP, mant-ATP, mant-dADP and mant-dATP bind in the site bordered by transmembrane helices (TM) 3, 5, 6 and 7. Ligands were oriented such that the phosphate groups generally pointed towards TM 3 and 7, forming hydrogen bonds with Y^{105} and K^{280}. Adenosine rings frequently interacted with the hydrophobic residues on TM 5, namely L^{184}, V^{185} and F^{177} in the second extracellular loop. In agreement with previous docking studies, R^{256} and K^{280} were found to be critical residues in the ADP binding pocket (Deflorian & Jacobson, 2011, Ignatovica, Megnis, Lapins, Schiöth, & Klovins, 2012). R^{256} frequently

interacts with the hydroxyl groups and the oxygen from the furanose. K^{280} is demonstrated to interact with the negatively charged phosphate groups of the ligands. In addition to the R^{256} and K^{280}, Y^{105}, E^{188} and Y^{259} are consistently found to interact with the ligand. Y^{105} and E^{188} form hydrogen bonds with the phosphate tail, while Y^{259} seems to stabilize the adenine.

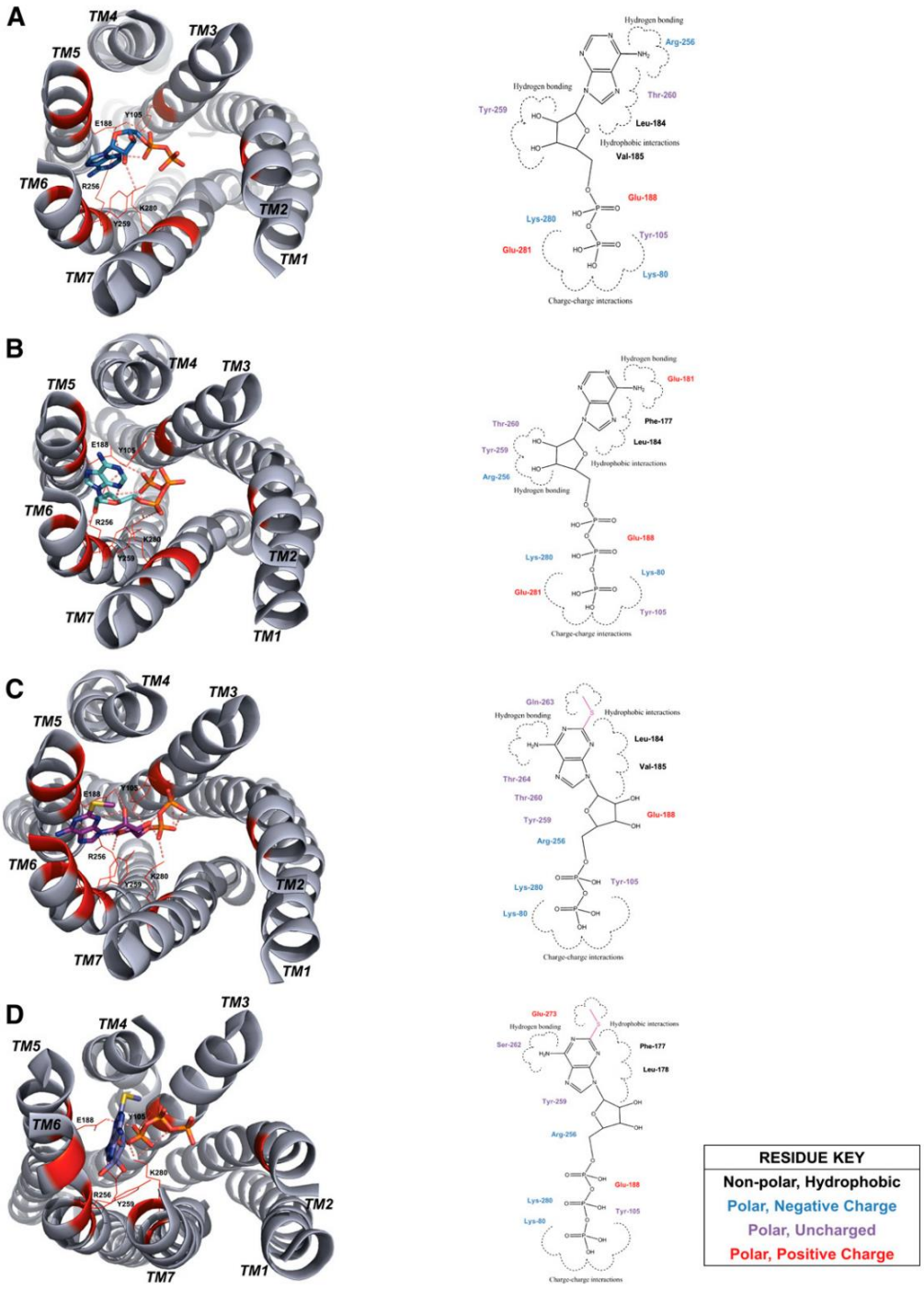


Figure 76
P2Y12 docked in complex with agonists ADP, ATP, MeS-ADP and MeS-ATP.
The docked binding modes in the comparative model of P2Y12 for agonists A) ADP, B) ATP, C) MeS-ADP and D) MeS-ATP in relation to residues Y105, E188, R256, Y259 and K280. All side chains within the binding site important for ligand interaction according to calculations of free energy change with and without ligands bound to P2Y12 are highlighted in red in the model and also shown in relation to the 2-D ligand depiction.

Identification of constitutively active mutants

It is still impossible to predict mutations leading to constitutive activity of a given GPCR. Furthermore, at positions where some mutations activate the GPCR, not all mutations will result in constitutive receptor activation (Bakker et al., 2008, Lalueza-Fox et al., 2007). Therefore, screening of mutant libraries is required. Mutations induced via error-prone derived mutant libraries cannot provide mutational saturation of every codon, and instead, most alleles will contain more than one mutated codon (B. Li et al., 2007, Thor, Schulz, Hermsdorf, & Schoneberg, 2008). Recent advances in gene synthesis technology (see Methods) have made it possible to generate comprehensive mutant libraries.

Here, we mutated every single position to all possible amino acids in a receptor region known from other GPCRs to be sensitive for mutational induced constitutive activity. In sum, 1,254 $P2Y_{12}$ mutants were generated covering 66 positions (amino acid position 236 – 301) of the receptor and yielding 28 constitutive active mutants at 10 positions (positions are given in Figure 77). Most mutations were found at positions that faced the lipid, while three positions faced the receptor pore (F^{246}, F^{249} and N^{290}) and three were near the C-terminal receptor tail (F^{299}, F^{300} and L^{301}) All data are available and organized in a $P2Y_{12}$ mutant database (http://www.ssfa-7tmr.de/$P2Y_{12}$).

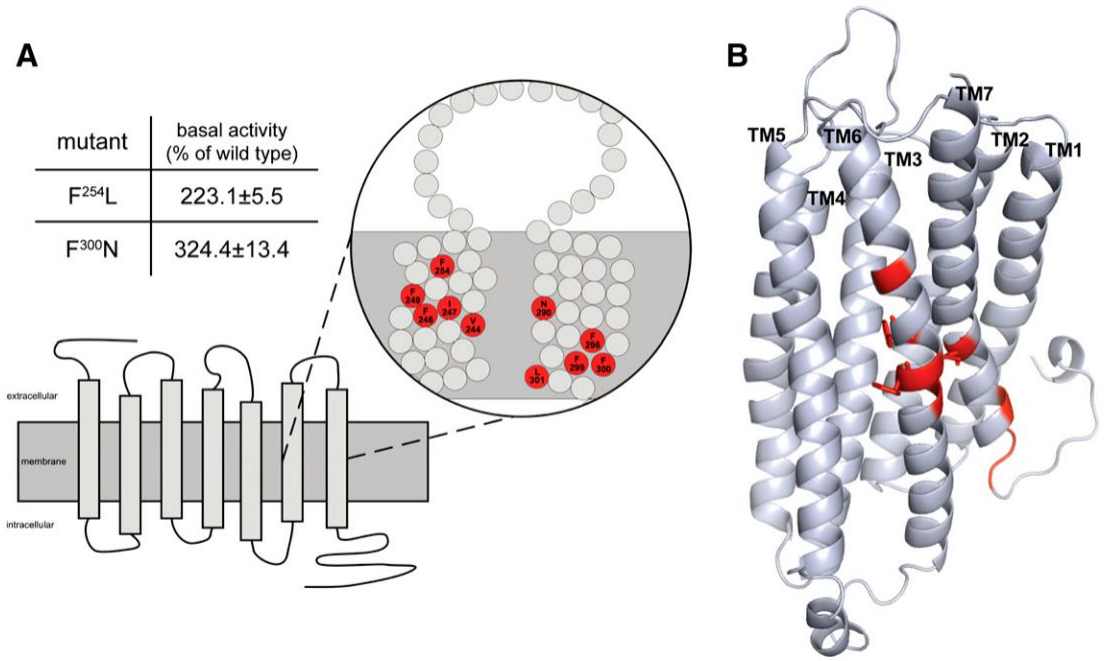


Figure 77
Position and basal activity of constitutively active P2Y12 mutants.
A) The position of constitutively active mutations in TM6 and TM7 are depicted. Basal activities of the individual mutants expressed in yeast are given in the table. Data are presented as mean ± SD of three independent experiments, each carried out in triplicate. The basal activity of the WT P2Y12 was OD600 nm: 0.074 ± 0.016. Complete functional data are available and organized in a P2Y12 mutant database (http://www.ssfa-7tmr.de/P2Y12). B) The comparative model of P2Y12 based on the CXCR4 template is depicted. Residues producing constitutively active mutants on TM6 and TM7 are highlighted in red. Residue side chains facing the pore of the receptor (F246, F249 and N290) are shown in sticks.

Identification of mant-dATP as inverse agonist at constitutively active $P2Y_{12}$ mutants

Constitutively active mutants were expressed and the purine compound library was tested for inverse agonists. N-methyl-anthraniloyl- (mant-) dATP reduced basal activity of many constitutively active $P2Y_{12}$ mutants (Table 27). For several mutants, mant-N6-methyl-ATP was also an inverse agonist (see Table 27). There is no obvious structural overlap or difference between the mutants at which the different inverse agonists act or do not act.

Inverse agonist activity was studied at $F^{254}L$ in more detail. As shown in Figure 78A, mant-dATP suppressed basal activity in a concentration-dependent manner with an IC_{50} value in a µM range. Interestingly, the potency of mant-dADP was lower compared to mant-

dATP (see Figure 78A). Both the deoxy- and the mant- modifications are required, since mant-ATP and dATP had no effect on basal activity of $P2Y_{12}$ mutants. It should be noted that the basal activities of several mutants ($V^{244}E$, $F^{246}C$, $F^{246}G$, $F^{246}P$, $F^{246}S$, $F^{246}T$, $I^{247}F$, $F^{249}Y$, $N^{290}W$, $N^{290}Y$, $F^{296}I$, $F^{296}L$, $F^{296}V$, $F^{299}I$, $F^{299}V$, $L^{301}C$, $L^{301}G$, $L^{301}T$) were not reduced by mant-dATP or any other compound tested. It is known already that WT $P2Y_{12}$ displays increased basal activity when compared to non-transfected mammalian cells (Schulz & Schöneberg, 2003).

To verify that mant-dATP mediates its inverse agonistic activity at the constitutive activity of the WT $P2Y_{12}$ expressed in mammalian cells as well, CHO-K1 cells were co-transfected with chimeric G_{qi4} and IP accumulation assays were performed. As shown in Figure 78B, the WT $P2Y_{12}$ displayed a high basal activity and MeS-ADP increased IP levels only 2-fold. Mant-dATP almost completely blocked basal IP formation at the WT $P2Y_{12}$ and $N^{300}F$ (Figure 78B). Also in cAMP inhibition assays at CHO-K1 cells, mant-dATP displayed strong inverse agonistic activity on the inhibition of basal cAMP formation at the WT $P2Y_{12}$ (Figure 78C).

It has been shown that some cell lines release receptor-function relevant amounts of nucleotides into the cell culture medium (Lazarowski, Homolya, Boucher, & Harden, 1997, Parr et al., 1994). This may account for high basal activity of $P2Y_{12}$ heterologously expressed in mammalian cell lines. Therefore, we performed similar control experiments with CHO-K1 cells stably transfected with Gαqi4. As shown in Figure 83, Gαqi4-CHO-K1 cells transiently transfected with $P2Y_{12}$ presented an increased basal IP_1 level compared to cells transfected with the control plasmid (GFP). Incubation with apyrase did not reduce this elevated IP_1 level. This clearly indicates that $P2Y_{12}$ does induce signal transduction by intrinsic active receptor conformation and not by nucleotides released from the cells into the medium. Proper apyrase function was demonstrated by loss of ADP action on $P2Y_{12}$.

Table 27

Mant-dATP is an inverse agonist at different constitutively active mutants.

Yeast cells expressing different basal active mutants were incubated with a 10 µM adenosine nucleotide library to identify inverse agonists. In a screen of over 80 adenine nucleotides and their derivatives mant-dATP and mant-N6-Methyl-ATP showed inverse activity on several constitutively active mutants. All the mutants listed showed activation (> 2-fold above increased basal activity) upon stimulation with ADP, MeS-ADP, ATP and mant-ADP. dATP and mant-ATP had no significant effects on the mutants.

Position	Mutation	Inverse activating substances (fold over basal ≤ 0.8)
F^{246}	V	mant-dATP
F^{254}	I	mant-dATP, mant-N6-Methyl-ATP
	L	mant-dATP, mant-N6-Methyl-ATP
	V	mant-dATP
F^{296}	A	mant-dATP
	C	mant-dATP, mant-N6-Methyl-ATP
	M	mant-dATP
F^{300}	N	mant-dATP
L^{301}	N	mant-dATP
	M	mant-dATP

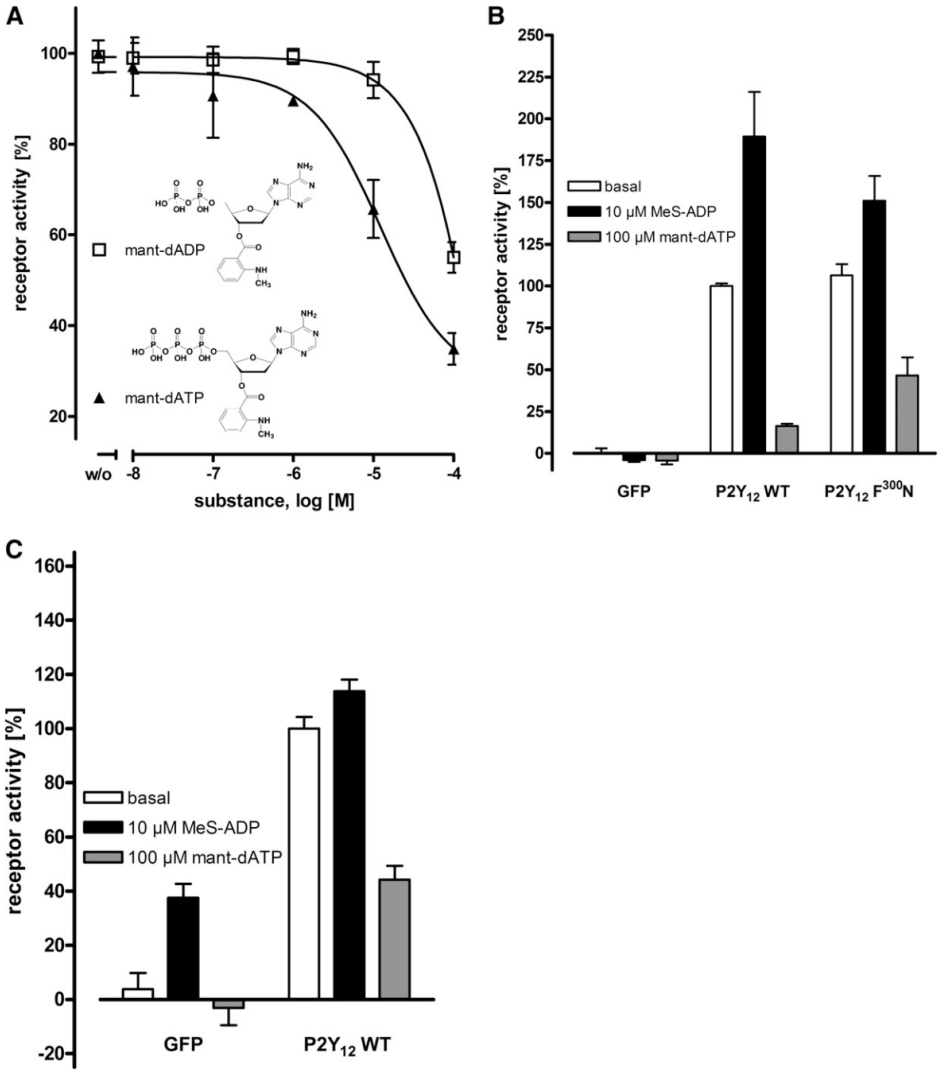


Figure 78
Mant-dADP and mant-dATP are inverse agonists at constitutively active P2Y12 F254L.
A) Yeast cells expressing F254L were incubated with increasing concentrations of the indicated compounds and yeast growth was measured after 24 h incubation. The ligand-induced decrease of basal activity of F254L is shown relative to the basal activity of the WT $P2Y_{12}$ (OD600 nm =0.074; set to 0%) and the basal activity of F254L (OD600 nm =0.165; set to 100%). Data are given as mean ± SD of three independent experiments all performed in triplicate. B) To evaluate inverse agonist specificity, CHO-K1 cells, stably expressing the chimeric G protein Gαqi4, were transfected with plasmids encoding GFP (control) or the human ADP receptor. IP formation under basal conditions (white), in the presence of 10 µm MeS-ADP (black bars) and in presence of 100 µM mant-dATP (light grey bars). The basal IP for GFP was 321 cpm/well and set 0%, the basal IP for WT $P2Y_{12}$ was 970 cpm/well and set 100%. Data are presented as mean ± SD (cpm/well) of three independent experiments, each carried out in duplicate. C) Forskolin-induced cAMP levels in CHO-K1 cells stably expressing human ADP receptor were determined under basal conditions (white bars), in the presence of 10 µm MeS-ADP (black bars) and in presence of 100 µM mant-dATP (light grey bars). The decrease of basal activity of $P2Y_{12}$ wild-type receptor is shown relative to GFP basal activity (7486 cpm/well; set 0%) and $P2Y_{12}$ wild-type basal activity (4533 cpm/well; set 100%). Data are given as mean ± SD of three independent experiments all performed in triplicates.

Mant-dATP is most likely an orthosteric ligand at $P2Y_{12}$

To evaluate whether mant-dATP mediates its inverse agonistic action through an orthosteric or an allosteric binding site, the ADP-concentration-response curves at $F^{254}L$ were determined in the presence of different concentrations of mant-dATP. As shown in Figure 79A, increasing concentrations of mant-dATP shifted the concentration-response curves to higher ADP concentrations. Similar results were obtained for mant-dADP, but with lower potency (Figure 79B). This competition is indicative of an orthosteric binding modus for the inverse agonists. Although, functional and docking data (see below) support orthosteric binding we cannot rule out the possibility of an allosteric binding of the inverse agonists given the limited concentration range of mant-dATP investigated herein.

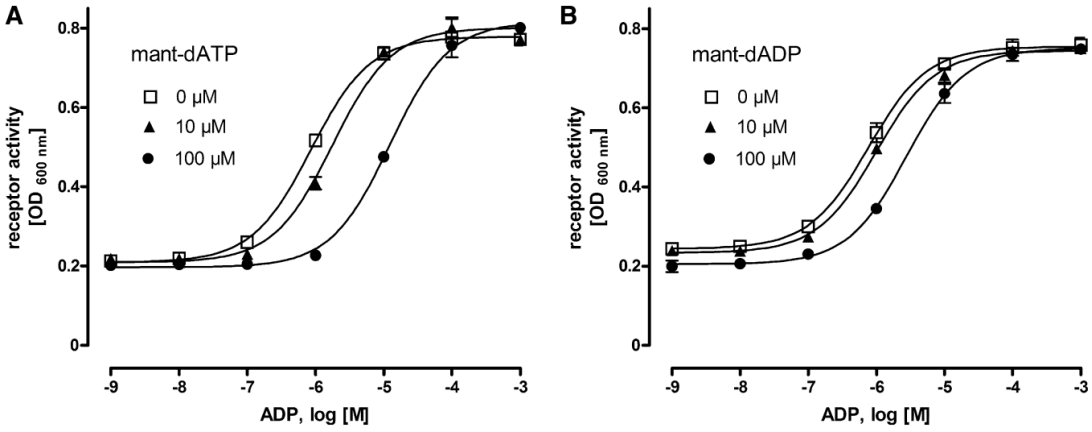


Figure 79
Mant-dADP and mant-dATP are most likely orthosteric ligands at P2Y12 F254L.
To evaluate the modus of inverse agonist binding, ADP-concentration-response curves at F254L-transformed yeast cells were determined in the presence of 0, 10 and 100 µM mant-dATP (A) and mant-dADP (B). Data are given as mean ± SD of three independent experiments all performed in duplicate.

Structural model of inverse agonist binding

Using our $P2Y_{12}$ model we investigated whether mant-dATP can dock into the agonist binding pocket of $P2Y_{12}$ and if specific interactions may explain inverse agonistic activity (Figure 80). As with the other ATP derivatives, mant-dATP sits between TMH 3, 5,

6 and 7 with Y^{105} and K^{280} forming hydrogen bonds with the phosphate tail and R^{256} stabilizing the oxygen connecting the furanose to the mant group. Unlike ATP, the extra bulk of the mant group is further stabilized by interactions with I^{257}, H^{253} and Q^{263}. However, similar interactions are seen with mant-ATP, which does not exhibit inverse agonism. We conclude that the inverse agonistic activity is likely not the result of a different binding mode. It is possibly caused by smaller scale modulations in the strengths of specific interactions between ligand and protein. Pin-pointing these changes to reveal the mechanism behind the inverse agonistic activity are beyond the accuracy of the present comparative model but will be the focus of future mutational studies.

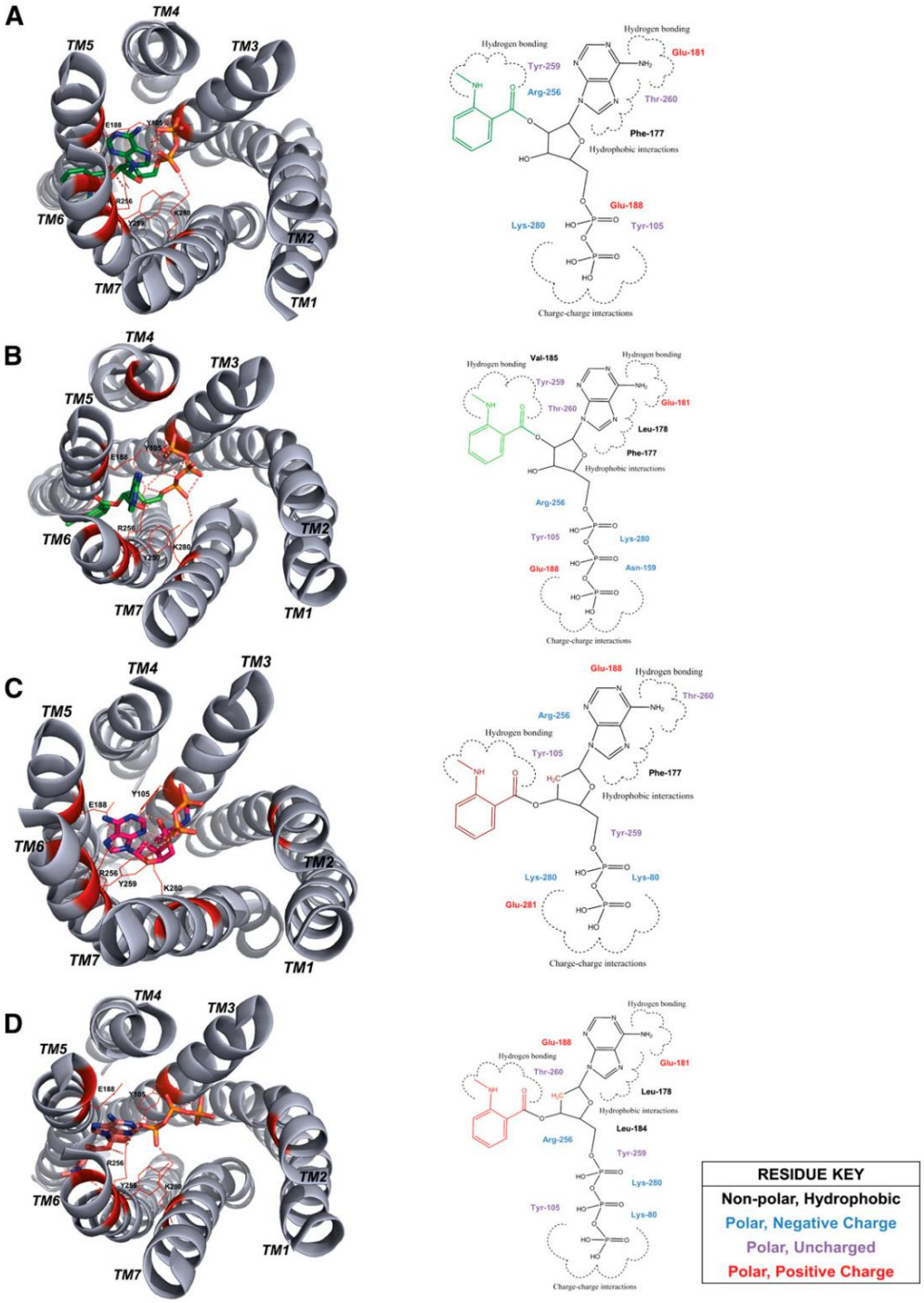


Figure 80
P2Y12 docked in complex with agonists mant-ADP and mant-ATP and inverse agonists, mant-dADP and mant-dATP.
The docked binding modes in the comparative model of P2Y12 for agonists A) mant-ADP and B) mant-ATP and the inverse agonists C) mant-dADP and D) mant-dATP in relation to residues Y105, E188, R256, Y259 and K280. All side chains within the binding site important for ligand interaction according to calculations of change in free energy with and without ligands bound to P2Y12 are highlighted in red and shown in relation to the 2-D ligand depiction.

Discussion

We used a genetically modified yeast strain (Pausch et al., 2004) to heterologously express and functionally test the human ADP receptor $P2Y_{12}$. This expression system offers some advantages over mammalian cells lines, specifically in characterizing nucleotide receptors, because it lacks endogenous nucleotide receptors. ADP and MeS-ADP are full agonists in this expression system with EC_{50} values of 2.8 µM and 6 nM, respectively (Figure 1A). Screening an adenine nucleotide-derived compound library we identified ATP and some derivatives as partial agonists at $P2Y_{12}$ in addition to ADP and its derivatives (Table 28). The agonistic activity of ATP was not only found in the heterologous yeast expression system but also in different mammalian cell lines and signaling assays.

That MeS-ATP and ATP binds to human $P2Y_{12}$ has been shown (P Savi et al., 2001) but the ligand properties of ATP at $P2Y_{12}$ are controversially discussed, ranging from antagonism (Bodor, Waldo, Blaesius, & Harden, 2004, Springthorpe et al., 2007) to agonism (Barnard & Simon, 2001, Simon et al., 2002). These contrary results are likely due to differences in mammalian expression systems and functional assays used. Introduction of a 2'-methylthio group increased ligand potency at $P2Y_{12}$ and made ATP a highly potent full agonist (Figure 75A) consistent with previous findings (Simon et al., 2002, F. L. Zhang et al., 2001). Through ligand docking into a structural comparative model of $P2Y_{12}$, ATP derivatives are found to bind in a similar binding site. While our structural $P2Y_{12}$ model is not at the resolution to reveal what fine-structural requirements are essential to turn a nucleotide into an agonist at $P2Y_{12}$, specific residues critical to ligand interaction can be predicted from the model. Notably, we find that for six of the seven residues indicated to be significant for ligand interaction that are also in the mutant database (H^{253}, I^{257}, Y^{259}, T^{260}, Q^{263}, T^{264} and K^{280}), mutation of the residues to any other amino acid results in a loss of WT function (see our $P2Y_{12}$ mutant database:

http://www.ssfa-7tmr.de/$P2Y_{12}$). Therefore, there is agreement between the residues predicted to be critical for agonist function through docking studies and experimental results. Our model and the docking studies are consistent with the fact that ATP fits into the same binding pocket as well-characterized agonists.

These findings raise a relevant question whether ATP can serve as $P2Y_{12}$ agonist also *in vivo*. The ATP-to-ADP ratio in human platelet dense granules is approximately two (Cattaneo, Lecchi, Lombardi, Gachet, & Zighetti, 2000, Weiss et al., 1979). If one assumes that ATP and ADP secretion from dense granule undergo with the same kinetics, previous data suggest that the surface concentration of ADP following thrombin stimulation will transiently reach 7–10 μM (Beigi, Kobatake, Aizawa, & Dubyak, 1999). This is sufficient for activation of the platelet $P2Y_{12}$ by ADP but also by ATP and consistent with feed-forward autocrine/paracrine activation of platelet responses.

Many WT GPCR, such as histamine receptors, thyrotropin receptor and melanocortin receptors present high basal activity (Seifert & Wenzel-Seifert, 2002). In contrast to antagonists inverse agonists suppress both agonist-dependent and -independent activity and are therefore developed in priority. For example, many β-blockers and atropine are inverse agonists at β1-adrenoceptors and muscarinic acetylcholine receptors, respectively (J. G. Baker, Hill, & Summers, 2011, Thor, Le Duc, Strotmann, & Schöneberg, 2009). Therapeutically used $P2Y_{12}$ ligands are high affinity antagonists but inverse activity was described only for the experimental $P2Y_{12}$ blocker AR-C78511 (Ding et al., 2006, Vasiljev, Uri, & Laitinen, 2003). AR-C78511 is a 2-alkylthio-substituted ATP analog but, in contrast to mant-dATP, has no modification at the 2' or 3' OH residues of the ribose. Mant-dATP binds most likely at the orthosteric ligand-binding site and inverse agonistic activity mutually depends on the desoxy-ribose, since mant-ATP lacks inverse agonistic activity. At present we cannot explain or predict inverse activity, even with a receptor model in hand, since the pharmacological properties of a ligand are

the result of a tight interplay of the ligand and the receptor molecule. It is however of interest that, as for AR-C78511 (Springthorpe et al., 2007), modification of an ATP backbone resulted again in an inverse agonist (mant-dATP). This also supports our findings that $P2Y_{12}$ naturally recognizes not only ADP but also ATP and binding of ATP and other ATP derivatives induce conformational changes within $P2Y_{12}$.

In sum, we clearly show that, in addition to ADP and ATP, some ATP derivatives are not only ligands of $P2Y_{12}$ but also agonists. Keeping with a ATP/ADP ratio >1 *in vivo* and the small differences in concentration response curves (see Figure 75B), $P2Y_{12}$ should rather be referred to as adenine nucleotide receptor without suggesting ADP specificity. Modification of the ribose within ATP can result in inverse activity of ATP-derived ligands.

Acknowledgments

This work was supported by the Deutsche Forschungsgemeinschaft (FOR 748, Scho 624/7-1; Sfb 610) and a student fellowship of the Medical Faculty, University of Leipzig, to D.W.. Work in the Meiler laboratory is supported through NIH (R01 GM080403, R01 MH090192) and NSF (Career 0742762). E. N. D. is supported through the Paul Calabresi Medical Student Research Fellowship from the PhRMA Foundation.

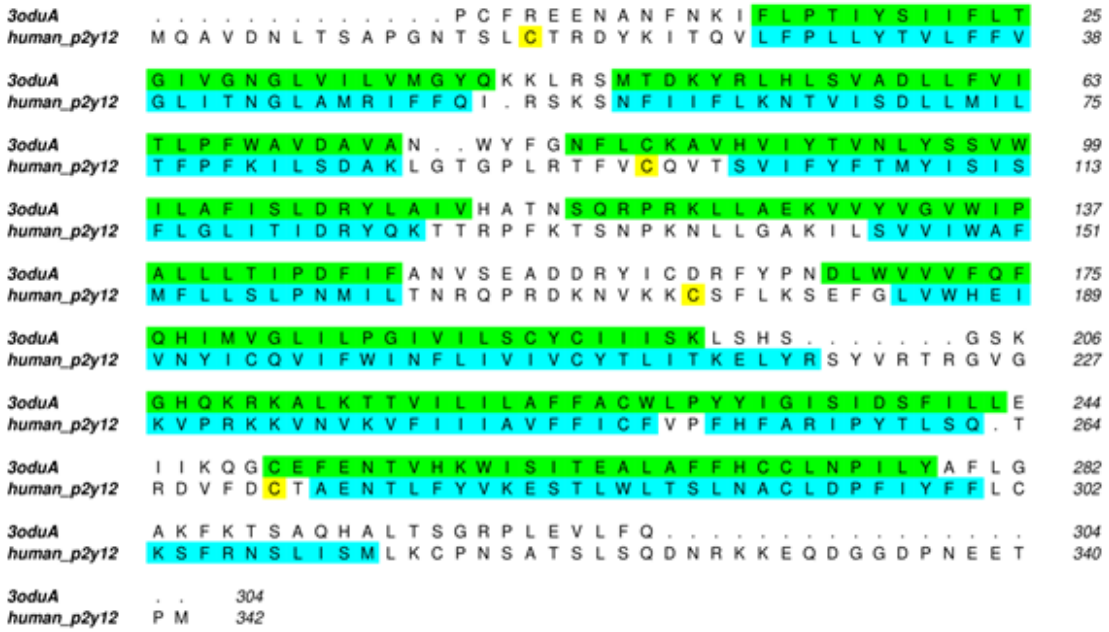


Figure 81
Sequence alignment of P2Y12 with CXCR4.
The sequence of human P2Y12 was aligned with the sequence of human CXCR4 (PDB ID: 3ODU) (Gupta et al., 2001) using CLUSTALW (Larkin et al., 2007). Transmembrane helical regions of the CXCR4 receptor are highlighted in green. Regions predicted to be transmembrane helical regions of P2Y12 according to secondary structure prediction server PSIPRED (McGuffin, Bryson, & Jones, 2000) are highlighted in cyan. Cysteine residues known to form disulphide bonds are highlighted in yellow.

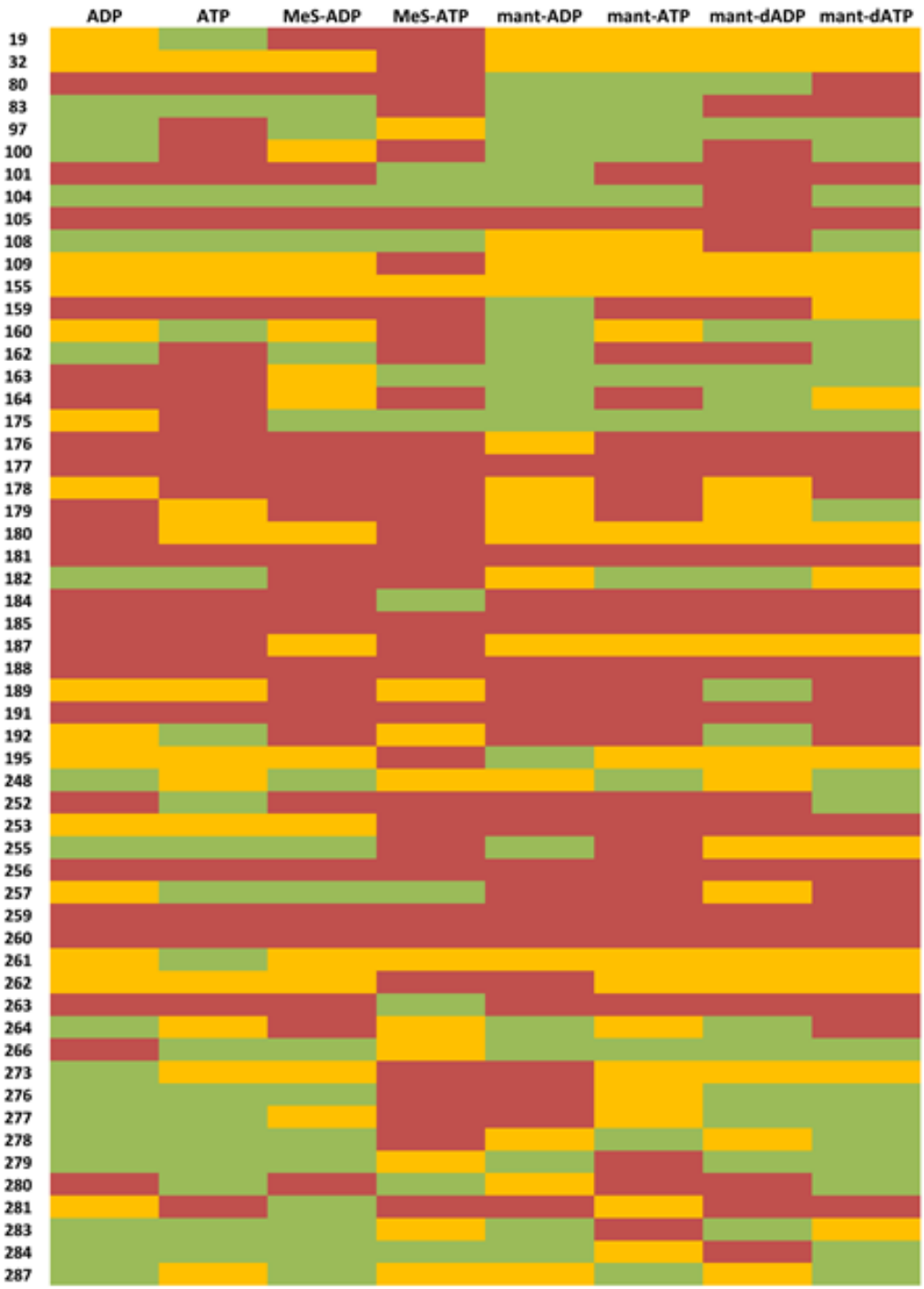


Figure 82
Per residue change in free energy with and without ATP derivatives bound.
Rosetta was used to determine the difference in change in free energy (ΔΔG) between P2Y12 with and without ATP derivatives bound. Residues that demonstrated the largest difference in free energy change in respect to bound ADP, ATP, MeS-ADP, MeS-ATP, mant-ATP, mant-ADP, mant-dADP and mant-dATP are listed. For each ligand, residues with ΔΔG above the average energy change (-0.24) are in red, while those with energy change above average but still negative are in yellow and residues with a positive energy change are in green.

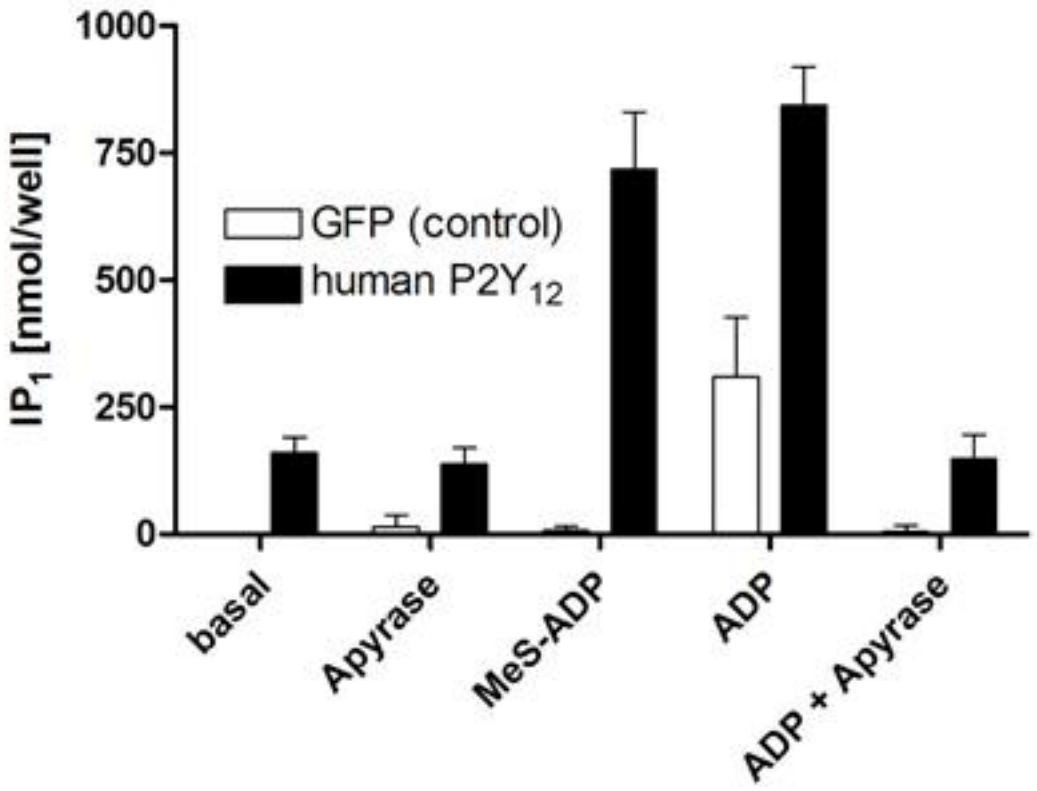


Figure 83
Analysis of the functional impact of endogenous nucleotide released from CHO cells
To analyze of whether endogenous nucleotide released from CHO cells contribute to basal activity of P2Y12 we performed control experiments with CHO-K1 cells stably transfected with Gαqi4. Intracellular inositol phosphate (IP1) levels were determined with an immunological assay (cisbio Bioassays, IP-One ELISA, part-no. 72IP1PEA). Gαqi4-CHO-K1 cells transiently transfected with P2Y12 presented an increased basal IP1 level compared to cells transfected with the control plasmid (GFP). Incubation with apyrase did not reduce this elevated IP1 level. This clearly indicates that P2Y12 does induce signal transduction by intrinsic active receptor conformation and not by nucleotides released from the cells into the medium. Proper P2Y12 transfection was control by application of ADP and MeS-ADP. Proper apyrase function was demonstrated by loss of ADP action on P2Y12.

Table 28

Adenosine nucleotide library screening at the P2Y12 WT receptor

The activity of P2Y12 WT receptor is shown relative to the basal activity (OD600 nm =0.089; set 0%) and the stimulation with MeS-ADP (OD600 nm =0.667; set 100%).

AMP -2.19	ADP 84.05	ATP 19.93	AP4 -0.29	cAMP -0.99
3'-dATP -2.02	7-Deaza-dAMP -1.33	NPE-caged-ATP -0.64	7-Deaza-dATP -1.33	DMB-caged-ATP 5.24
dATPαS -1.68	ADPβS 92	ATPγS 71.09	β-Methylene-APS -1.68	ApCp -0.81
AP3A -0.99	AP4A -1.5	AP5A -1.68	AP6A -1.68	AP4U -0.99
AP4(8I)G -2.37	AP5(8I)G -1.68	2'I-AMP -1.33	2'I-ADP -1.68	2'I-ATP -1.16
8Br-ADP 0.57	8Br-ATP -1.68	8Br-dATP -0.99	γ-[6-Aminohexyl]-ATP -1.5	2I-ATPγS 90.1
7-Deaza-7Br-dATP -2.02	γ-Aminophenyl-ATP 2.99	γ-[(6-Aminohexyl)-imido]-ATP -1.16	γ-[(8-Aminooctyl)-imido]-ATP -1.16	N6-(4-Amino)butyl-ATP -1.33
EDA-ATP -3.75	γ-[6-Aminohexyl]-N6-Benzyl-ATP -2.02	2-Hydroxy-ATP -1.5	TNP-ADP 72.81	TNP-ATP 4.03
1-Methyl-AMP -0.99	1-Methyl-ADP 1.61	1-Methyl-ATP 0.05	dATP 0.57	ddATP 0.57
2'.5'-pAp -0.47	ara-ATP 0.4	AMPαS -0.12	ATPαS 2.3	dADPαS 0.4
ApCpp -1.33	AppCp 0.4	dApCpp -0.47	AppNp 0.57	AppNH2 76.79
AP5U -1.16	AP4T -0.29	AP5T -0.47	AP4G 0.22	AP5G 0.05
2'Br-ADP -1.16	2'Br-ATP -0.64	2'-Ome-ATP -0.81	mante-ATPγS -0.4	8Br-cAMP 0.05
2'F-AMP -0.81	2'F-ATP -0.47	2'Cl-ATP -0.81	BzBzATP -0.12	7-Deaza-7I-dATP -0.12
N6-(6-Amino)hexyl-ATP -0.81	N6-(6-Amino)hexyl-dATP -0.12	8-[(4-Amino)butyl]-amino-ATP -0.12	8-[(6-Amino)hexyl]-amino-ATP 0.57	EDA-ADP 1.95
mant-ADP 6.79	mant-ATP -0.47	mant-dATP -0.64	mant-N6-Methyl-ATP -0.29	ε-ATP 0.4

BCL::ALIGN – SEQUENCE ALIGNMENT AND FOLD RECOGNITION WITH A CUSTOM SCORING FUNCTION ONLINE

This work is based on publication (Dong et al., 2008).

Summary

BCL::Align is a multiple sequence alignment tool that utilizes the dynamic programming method in combination with a customizable scoring function for sequence alignment and fold recognition. The scoring function is a weighted sum of the traditional PAM and BLOSUM scoring matrices, position specific scoring matrices output by PSI-BLAST, secondary structure predicted by a variety of methods, chemical properties, and gap penalties. By adjusting the weights, the method can be tailored for fold recognition or sequence alignment tasks at different levels of sequence identity. A Monte Carlo algorithm was used to determine optimized weight sets for sequence alignment and fold recognition that most accurately reproduced the SABmark reference alignment test set. In an evaluation of sequence alignment performance, BCL::Align ranked best in alignment accuracy (Cline score of 22.90 for sequences in the Twilight Zone) when compared with Align-m, ClustalW, T-Coffee, and MUSCLE. ROC curve analysis indicates BCL::Align's ability to correctly recognize protein folds with over 80% accuracy. The flexibility of the program allows it to be optimized for specific classes of proteins (e.g. membrane proteins) or fold families (e.g. TIM-barrel proteins). BCL::Align is free for academic use and available online at http://www.meilerlab.org/.

Introduction

Sequence alignment and fold recognition are key computational tools for predicting the evolutionary history of proteins and detecting structurally related proteins from their

amino acid sequence. The importance of these methods continues to increase with the exponential growth of sequence databases driven by various genome projects (Benson, Karsch-Mizrachi, Lipman, Ostell, & Wheeler, 2007, Mewes et al., 2002). With the help of these tools, relationships are being determined between newly discovered sequences and existing sequence databases (Bairoch & Apweiler, 1998, Benson, Karsch-Mizrachi, Lipman, Ostell, & Wheeler, 2006) along with proteins of known structure collected in the protein data bank (Berman, Henrick, & Nakamura, 2003). While sequence similarity frequently accompanies structural similarity as well as evolutionary relation to a common ancestor (Castillo-Davis, Kondrashov, Hartl, & Kulathinal, 2004, Phillips, Janies, Wheeler, & others, 2000), one major goal of these comparisons is the assignment of a function to newly discovered sequences.

Yet it is known that many structurally homologous proteins can have very low sequence identity (Rychlewski, Li, Jaroszewski, & Godzik, 2000) and in these cases sequence alignment methods alone provide little information. Threading algorithms (Jones & others, 1999, Lindahl & Elofsson, 2000) and sequence-only methods (Karplus, Barrett, & Hughey, 1998, Rychlewski et al., 2000) for fold recognition have been specifically developed to predict structural similarity. However, the accuracy of most sequence alignment methods as well as the reliability of fold recognition methods is greatly diminished when comparing sequences in the so-called “Twilight Zone” with less than 25% sequence identity (Rost & Sander, 1993, J. D. Thompson, Plewniak, & Poch, 1999).

Approaches to improve the accuracy of automatic sequence alignments start with the introduction of common substitution matrices such as PAM (Dayhoff & Schwartz, 1978) or BLOSUM (Henikoff & Henikoff, 1992). With the progressive algorithm (Feng & Doolittle, 1987, Hogeweg & Hesper, 1984), MUSCLE (R. Edgar, 2004) uses probabilities derived from the PAM 240 matrix and position specific gap penalties with iterative score refinement. ClustalW (J. D. Thompson et al., 1994) also uses a progressive alignment

method and improves its accuracy by weighting sequences, customizing substitution matrix usage and changing gap penalties depending on the surrounding residues. Align-m (Van Walle, Lasters, & Wyns, 2004) uses a non-progressive local approach to guide a global alignment. T-Coffee (Notredame, Higgins, Heringa, & others, 2000) combines information from global and local sequence alignments to determine the optimal alignment. However, BLAST (Altschul, Gish, Miller, Myers, & Lipman, 1990) and PSI-BLAST (Altschul et al., 1997) continue to dominate the field of sequence alignment tools with their rapid word-based algorithm and the iterative search using position-specific score matrices.

While there is some overlap between the tools used for sequence alignment and fold recognition, there is significant emphasis on secondary structure prediction in fold recognition methods. Sequence-based methods (Lindahl & Elofsson, 2000, Rychlewski et al., 2000) include structural information within the sequence profiles used for sequence-structure alignment. Threading-based algorithms like THREADER (Jones & others, 1999) directly use a structure template which is based on residue-residue contacts and hydrophobicity among other structural information in its double dynamic programming algorithm. ORFeus (Ginalski, Pas, et al., 2003) uses a scoring matrix based on the PSI-BLAST profile and secondary structure prediction from PSIPRED (Jones & others, 1999). K*sync (Chivian & Baker, 2006) is a comparative modeling approach that uses various weight sets to create an ensemble of sequence-sequence alignments. Based on this ensemble a library of models is created from which the optimal model is selected by tertiary structure analysis and energy prediction.

More recently, it was shown that fold recognition can be improved by incorporating the output of several primary fold recognition approaches in a secondary approach. Such meta-servers work by analyzing the consensus of several primary methods using either artificial neural networks (P-Cons) (Lundström, Rychlewski, Bujnicki, & Elofsson, 2001) or

more straight-forward structure comparison tools (3D-Jury) (Ginalski, Elofsson, Fischer, & Rychlewski, 2003).

With the growing number of sequence analysis and fold recognition tools being developed, it became clear that different scoring schemes can perform quite differently depending on the protein class, sequence identity level, or type of problem (fold recognition vs. sequence alignment). In turn, the researcher often needs to invoke multiple tools to accomplish these tasks and it is difficult to determine which method produces the most accurate result given a particular scenario.

In the present study we seek to address this shortcoming by introducing BCL::Align. The program gives the user maximum flexibility in tailoring the scoring function to fit the specific problem. The effective scoring function used by BCL::Align is a linear combination of various substitution matrices, position-specific scoring matrices, secondary structure predictions, chemical properties, and gap penalties. Here, the algorithms implemented in BCL::Align are described and optimized parameter sets for four typical tasks are presented (sequence alignment and fold recognition in the 0-25% sequence identity and 25-50% sequence identity regime). Results for the SABmark benchmark database are compared with other leading sequence alignment tools. The significance of the weights is discussed in terms of their importance for sequence alignment and fold recognition at different levels of sequence identity.

Materials and Methods

Needleman and Wunsch algorithm is employed for generation of optimal pair-wise sequence alignment

BCL::Align uses a standard dynamic programming algorithm (Needleman & Wunsch, 1970) to optimally align two sequences *A* and *B* of length *m* and *n*. In order to

execute the alignment, a scoring scheme for matches as well as gaps needs to be provided. The dynamic programming algorithm will output the optimal score $S_{m,n}$ together with the alignment.

Dynamic programming solves optimization problems by dividing the problem into independent subproblems. Since the sequence alignment problem has optimal substructure property, a subproblem can be defined as aligning prefixes of two sequences up to a point (i,j) with $0<i\leq m$ and $0<j\leq n$. To find the alignment with the highest score $S_{m,n}$, a two-dimensional matrix with the dimensions m and n is filled at each position (i,j) with the best score $S_{i,j}$ of these prefix sequences ("matrix filling"). The optimal score $S_{i,j}$ builds upon the best score computed so far. The second part of the algorithm – so-called "trace back" – starts at the lower right corner of the matrix which now contains the best possible score $S_{m,n}$. It traces back step-by-step the pathway through the matrix that lead to this optimal score, thereby generating the optimal alignment of the two sequences.

Setup of parametric scoring function as a sum of weighted Z-scores

The scoring function of BCL::Align is a weighted sum of multiple scoring schemes that have been successfully used in prior sequence alignment and fold recognition approaches. The user can choose the individual weight of each scheme and BCL::Align will recalibrate them to add up to 100%.

Raw scores obtained from each of the different scoring schemes are not directly comparable. Therefore all scores are first translated into Z-scores. For every scoring scheme, a random distribution was created by computing the score S for 10^6 arbitrarily chosen pairs of amino acids out of a representative database consisting of 1,800 protein sequences. This database was created by culling the PDB (Berman et al., 2003) for sequences with less than 25% sequence identity (G. Wang & Dunbrack, 2003). For each

of the different scores an average S_{av} and a standard deviation S_{sd} was derived (see Table 29) which are used within BCL::Align to rescale all scores into Z-scores with $Z=(S-S_{av})/S_{sd}$. Therefore, positive scores larger than 1 indicate that two positions align with a score that is least one standard deviation above the average. Since the total score is a sum of weighted Z-scores, this statement holds not only for the individual scores but also for the total score, making all scores obtained with BCL::Align directly comparable even if the composition of the scoring function was altered.

Table 29
Adjustable parameters and gap penalties.

Description	Parameters	S_{av}[a]	S_{sd}[b]
Amino Acid Identity	Identity		
Substitution Matrices	PAM 100, 120, 160, 250 (Dayhoff et al., 1978)	-0.0824	0.2498
	BLOSUM 90, 80, 62, 45 (Henikoff and Henikoff, 1992)	-0.0821	0.2273
Position specific scoring matrix	BLAST profile (Altschul et al., 1997)	-0.0072	0.0881
Secondary Structure Predictions	PSIPRED (Jones, 1999)	-0.1431	0.4728
	JUFO (Meiler and Baker, 2003)	-0.0388	0.2451
	SAM (Karplus et al., 1998; Hughey and Korgh, 1996)	-0.0056	0.2076
Chemical Properties	steric parameter	-1.1514	0.8981
	polarizability	-0.1061	0.0814
	volume	-1.9938	1.5660
	hydrophobicity	-1.0737	0.7871
	isoelectric point	-1.6180	1.8058
Gap Penalties	open gap		
	extend gap		
	open boundary gap		
	extend boundary gap		

[a] Average score for Z-score correction
[b] Standard deviation for Z-score correction

Use of the affine gap penalty is essential for alignment of distant sequence homologs

The affine gap penalty approach (Barton & Sternberg, 1987) improves sequence alignment by customizing gap penalties to the sequence, making them length- and location-dependent. BCL::Align distinguishes gap open penalties P_{open} from gap extension penalties P_{extend}. It also distinguishes boundary gaps at the beginning or end of an alignment P^B from enclosed gaps P^E. In turn, a total of four gap penalties are defined that can be chosen by the user. The total penalty for a gap is computed using $P = P_{open} + length \bullet P_{extend}$.

Scoring function components were chosen from successful sequence alignment benchmarks and can be easily extended

Table 29 lists the parameter options open to the user. While substitution matrices of various sequence identity are available, the PAM250 (Dayhoff & Schwartz, 1978) and BLOSUM45 (Henikoff & Henikoff, 1992) matrices were used for sequence alignment because these matrices are most suitable for aligning sequences with low sequence identity. The logarithm of the probability of replacing amino acid *i* with *j* is used as the score.

The BLAST profile is iteratively built from members of the homologous family by scanning a sequence database (Altschul et al., 1997). In this work, the BLAST profile was determined by 3 PSI-Blast iterations at an E-value cutoff of 0.001. The logarithm of the scalar product of the probability vectors for position *i* and *j* is used as the score. One advantage of using these parameters is that the scoring matrix obtained can be used directly for running PSIPRED and JUFO (see below).

The secondary structure predictions used in BCL::Align include PSIPRED (McGuffin et al., 2000), JUFO (Meiler & Baker, 2003) and SAM (Hughey & Krogh, 1996,

Karplus et al., 1998). The logarithm of the scalar product of the 3-state (helix, strand, coil) probability vectors for position *i* and *j* is used as the score.

The chemical properties used include sterical parameters, polarizability, volume, hydrophobicity, and the isoelectric point. For scoring, the negative absolute difference for amino acids *i* and *j* is computed. After Z-score normalization, all five properties were combined with equal weights into a single score for weight optimization.

The SABmark benchmark database

For parameter optimization, we chose to use a subset of the 1.65 version of the SABmark reference alignment database (Van Walle, Lasters, & Wyns, 2005), which is itself divided into two subsets. Sequences in the Superfamily subset have 25-50% sequence identity and are divided into test groups that represent different SCOP superfamilies. The Twilight Zone subset has sequences with 0-25% sequence identity and each test group represents a different SCOP fold.

SABmark also includes a second set of Twilight Zone and Superfamily subsets with the same sequences, plus the addition of up to the same number of false positive sequences. These false positives differ in fold from the true positives. They were selected from a BLAST search of the original sequences against a 70% identity subset of SCOP. The database covers the entire known fold space and each pairwise reference alignment is a consensus structural alignment provided by SOFI (Boutonnet, Rooman, Ochagavia, Richelle, & Wodak, 1995) and CE (Shindyalov & Bourne, 1998).

Because SABmark contained pairwise sequence alignments as well as fold information, we were able to use the benchmark to optimize the parameters for both the sequence alignment and fold recognition methods.

Optimizing the Cline score avoids over- and underprediction in sequence alignment

A total of eleven parameters and four gap penalties were optimized in our experiment (Table 29). For sequence alignment parameter and gap penalty optimization, we chose to maximize the Cline score (Cline, Hughey, & Karplus, 2002) as a measure of alignment quality, finding in agreement with previous publications that maximizing the developer's score (f_d) alone leads to overprediction while maximizing modeler's score (f_m) leads to underprediction (R. C. Edgar & Sjölander, 2004, Sauder, Arthur, & Dunbrack Jr, 2000). Scores were calculated using the qscore program (R. Edgar, 2004) found at http://www.drive4.com/qscore/.

ROC curve analysis predicts accuracy of fold recognition

For fold alignment parameter optimization, we performed a receiver operating characteristic (ROC) curve analysis on the rate of correct versus incorrect fold assignment. A ROC curve plots the false positive rate against the true positive rate. Calculating the area underneath the ROC curve provides a measure of fold alignment accuracy, where an area of 50% would represent a program with no ability to recognize folds. The area underneath the ROC curve was maximized during parameter optimization.

Parameter and gap penalty optimization using a Monte Carlo algorithm

For both the sequence alignment and fold recognition methods, we performed two different optimizations, one with Twilight Zone sequences with low (0-25%) sequence identity and one with Superfamily sequences with intermediate (25-50%) sequence identity. For sequence alignment, the parameter and gap penalty optimization was performed on 50% of the Twilight Zone subset and 36% of the Superfamily subset. For

fold recognition, 45% of the Twilight Zone subset and 22% of the Superfamily subset was used for the training set.

Using a Monte Carlo approach, we started the optimization with random values between 0 and 1 for the parameters and values between -2 and 0 for the gap penalties. For 100 Monte Carlo iterations, we adjusted the weights for the parameters and gap penalties by a random value between -0.2 and 0.2, maximizing the Cline score for sequence alignment and the area under the ROC curve for fold recognition. Fifteen rounds of this optimization procedure were carried out on each subset and weights from the top ten scoring rounds were averaged to determine the optimal weight set. The most favorable range for a particular weight is defined by average and standard deviation of the top ten scoring rounds of each trained subset.

Cross validation was used to avoid over-training

Since a subset of the SABmark database was used to determine the weight sets, we had to verify that the scores resulting from the parameter and gap penalty optimization were not affected by overtraining. To do so, the scores for the trained and untrained subset were compared with each other. They were found to be within the standard deviation (Table 33), validating that the scores taken from the weight optimization can be directly compared with other leading methods.

Performance assessment

We assessed the sequence alignment performance of BCL::Align using the entire SABmark database. The average Cline scores for pairwise alignments in a group were calculated, and those scores were averaged to determine the final Cline score for each subset of SABmark: Twilight Zone, Superfamilies, Twilight Zone with False Positives, and Superfamilies with False Positives.

Implementation

The benchmarking and testing methods were written in C (using MPI for automated load balancing across a number of processors), with additional scripts written in Perl. Parameter optimization and performance assessment were performed on the PowerPC nodes of the Vanderbilt University Advanced Computing Center for Research and Education (ACCRE).

Results and Discussion

Optimal parameters and gap penalties

Details of the training set are given in Table 30, along with the average of the top ten scoring rounds of the Monte Carlo optimization. The optimized sequence alignment training set had an average Cline score of 27 for Twilight Zone sequences and 49 for Superfamily sequences. For fold recognition, the area underneath the ROC curve for the optimized training set scored an average of 82 for both subsets.

Table 30
Training set on SABmark for parameter optimization.

Problem	Sequence Identity Level [a]	Fraction of SABmark database used [b]		Score [c]
Sequence Alignment	Twilight Zone	50%	873 of 1740 seq.	27
	Superfamilies	36%	1197 of 3280 seq.	49
Fold Recognition	Twilight Zone	45%	1552 of 3458 seq.	82
	Superfamilies	22%	1460 of 6526 seq.	82

[a] The sequence identity level is 0-25% for the Twilight Zone subset and 25-50% for Superfamily subset.
[b] The fraction of the SABmark database used for weight optimization is given as a percentage and in absolute sequences.
[c] Cline scores are reported for sequence alignment methods and the area under the ROC curve is reported for fold recognition methods. All scores are multiplied by 100. The maximum for both scores is 100.

Tables 3 and 4 give the distribution of the sequence-identity dependent optimal weight sets for BCL::Align parameters and gap penalties for sequence alignment and fold recognition. The standard deviation on most weights is five percentage points or less, demonstrating the robust nature of the Monte Carlo optimization. However, we find that there is flexibility in the use of secondary structure elements for sequence alignment, particularly PSIPRED and JUFO. PSIPRED weights can vary up to 8 percentage points for the alignment of Twilight Zone sequences and 5 for Superfamily sequences. JUFO weights can vary up to 11 percentage points for Twilight Zone sequences and 8 for Superfamily sequences. The increase in standard deviation may be due to the various methods of secondary structure prediction compensating for each other in weight value, making their individual weights vary from one round to another.

For the gap penalties, we find that the same score is given by a consistent set of weights and the only range larger than 0.5 is found in the weight for the extension boundary gap for the alignment of sequences in the Twilight Zone subset at 0.6.

The relative weight of the parameters, expressed as percentages in Table 31, suggest that the BLAST profile and PSIPRED secondary structure information carry equal weight, within the standard deviation, for each of the four tasks. For sequence alignment, the BLAST profile has the highest average weight at 36% for the Twilight Zone subset and 40% for the Superfamily subset. This reiterates the power of position specific scoring matrices created with PSI-BLAST as tools for sequence analysis. Amongst the secondary structure elements weights for alignment and fold recognition, we find that PSIPRED consistently carries the largest weight, with JUFO and SAM following behind. Only in the fold recognition of the Twilight Zone sequences do we find that JUFO and SAM carry equal weight at an average of 5%. For all other tasks, we find that JUFO outweighs SAM by over 10%. It is remarkable that the sum of the three secondary structure prediction weights is the largest contribution to the composite scoring function for all four benchmark cases.

The chemical properties of amino acids carry more weight in aligning sequences from the Twilight Zone at 11% compared to the 7% for Superfamily sequences. However, we find that the chemical properties are even more important in fold recognition, carrying 8% of the weight for the fold recognition of Twilight Zone sequences and 24% of the weight for the Superfamily subset. The relative importance of the PAM and BLOSUM substitution matrices is minimal in sequence alignment with weights below 2%, but we find that the BLOSUM matrix carries considerable weight in fold recognition at an average of 19% for Twilight Zone sequences and 13% for Superfamily sequences.

Large open gap and open boundary gap penalties were generally favored during parameter optimization of both the Twilight Zone and Superfamily subsets. The open gap penalty was -0.8 or more and the open boundary gap penalty was greater than -0.6 for all fold recognition and sequence alignment tasks. Generally, the extension gap and extension boundary gaps were penalized less, demonstrating the importance of the use of an affine gap penalty. We find that the extension boundary gap was penalized less than -0.3 for sequence alignment and fold recognition, as well as the extension gap for both sequence alignment tasks. However, for fold recognition there is a -1.3 penalty for Twilight Zone sequences and -1.4 for Superfamily sequences, indicating a particular emphasis on a penalty of this gap for fold recognition.

Table 31
Distribution of weights for parameters [a]

Problem	Sequence Identity Level [b]	PAM 250	BLOSUM 45	BLAST	PSIPRED	JUFO	SAM	Chemical Properties [c]
Sequence	Twilight Zone	0±0%	1±2%	36±5%	33±8%	16±11%	2±3%	11±3%
Alignment	Superfamilies	1±1%	2±2%	40±1%	35±5%	14±8%	1±2%	7±1%
Fold Recognition	Twilight Zone	1±0%	19±6%	33±4%	30±4%	5±5%	5±5%	8±2%
	Superfamilies	0±1%	13±5%	18±5%	20±3%	18±3%	7±6%	24±4%

[a] Weight values, varying from 0 to 1.0, were normalized to calculate percentage of weight value out of 100%. Scores may not add to 100% due to rounding. Ranges represent the standard deviation of the top ten scores.
[b] The sequence identity level is 0-25% for the Twilight Zone subset and 25-50% for Superfamily subset.
[c] Chemical properties include sterical parameters, polarizability, volume, hydrophobicity, and the isoelectric point.

Table 32
Optimized weights for gap penalties

Problem	Sequence Identity Level [a]	Open Gap	Extension Gap	Open Boundary Gap	Extension Boundary Gap
Sequence Alignment	Twilight Zone	-1.4±0.3	-0.1±0.1	-0.7±0.4	-0.3±0.6
	Superfamilies	-1.9±0.1	-0.1±0.1	-0.9±0.2	0.0±0.1
Fold Recognition	Twilight Zone	-1.2±0.2	-1.3±0.2	-0.6±0.4	-0.2±0.1
	Superfamilies	-0.8±0.4	-1.4±0.4	-1.7±0.3	-0.1±0.1

[a] Gap penalty values vary from -2 to 0. Ranges represent the standard deviation of the top ten scores.
[b] The sequence identity level is 0-25% for the Twilight Zone subset and 25-50% for Superfamily subset

Cross-validation confirms absence of over-training

The scores for the trained and untrained subsets of SABmark for each of the four tasks are given in Table 33. In the Twilight Zone subset, the untrained subset had a Cline score of 24 whereas the trained subset had a score of 23. For the Superfamily subset, the untrained subset scored 51 while the trained subset had a score of 49. The scores for the untrained subsets of SABmark for sequence alignment are higher than those of the trained subset, providing evidence that the Monte Carlo optimization did not over-train the weight set and thus the method is not biased towards this particular subset. Although the scores of the untrained subset are lower than those of the trained subset for fold recognition, the difference is still within 2 percentage points. Nevertheless, BCL::Align would still benefit from future benchmarking tests on fold recognition benchmark databases such as the Lindahl Benchmark for fold-recognition sensitivity (Lindahl & Elofsson, 2000).

Table 33
Scores on trained and untrained subsets of SABmark with optimal weight set.

Problem	Sequence Identity Level [a]	Score for trained subset[b]	Score for test subset[b]
Sequence Alignment	Twilight Zone	23	24
	Superfamilies	49	51
Fold Recognition	Twilight Zone	87	86
	Superfamilies	88	86

[a] The sequence identity level is 0-25% for the Twilight Zone subset and 25-50% for Superfamily subset.
[b] Cline scores are reported for sequence alignment methods and the area under the ROC curve is reported for fold recognition methods. All scores are multiplied by 100. The maximum for both scores is 100.

Comparison of sequence alignment methods

We compared the results of BCL::Align sequence alignment with Align-m (Van Walle et al., 2004), ClustalW (J. D. Thompson et al., 1994), T-Coffee (Notredame et al., 2000) and MUSCLE (R. Edgar, 2004) on the SABmark benchmark database using the Cline score. Scores for the methods listed above are from Blackshields et al., 2006

(Blackshields, Wallace, Larkin, & Higgins, 2006). BCL::Align results on the entire SABmark benchmark database are shown in Table 34. In each subset, BCL::Align ranks the highest in alignment accuracy, demonstrating the superiority of BCL::Align's scoring function and the power of weight flexibility when compared to other programs that also use the dynamic programming algorithm (see Figure 84). According to the data provided by Blackshields et al., ProbCons (Do, Mahabhashyam, Brudno, & Batzoglou, 2005) was the only program that consistently scored somewhat higher than BCL::Align. This is likely due to the fact that ProbCons does not employ dynamic programming but combines posterior-probabilities from pair-hidden Markov models (HMM) with a consistency-based method to determine scoring matrices.

Table 34
Performance comparison of multiple sequence alignment programs on SABmark.[a]

	Superfamilies[b]		Twilight Zone[b]	
	No FP[c]	With FP[c]	No FP[c]	With FP[c]
Align-m	44.75	41.53	15.93	13.72
ClustalW	47.60	47.82	18.57	17.98
T-Coffee	50.20	45.58	20.80	16.94
MUSCLE	44.52	40.38	15.45	12.44
BCL Align	50.74	50.80	23.02	23.66

[a] Cline scores are reported for each multiple alignment program. The highest score in each subset is displayed in bold. Scores for all methods except BCL::Align are from Blackshields et al. (2006).
[b] The sequence identity level is 0-25% for the Twilight Zone subset and 25-50% for Superfamily subset.
[c] Subsets include the addition of up to the same number of false positive sequences. False positives differ in fold from the true positives and were selected from a BLAST search of the original sequences against a 70% identity subset of SCOP.

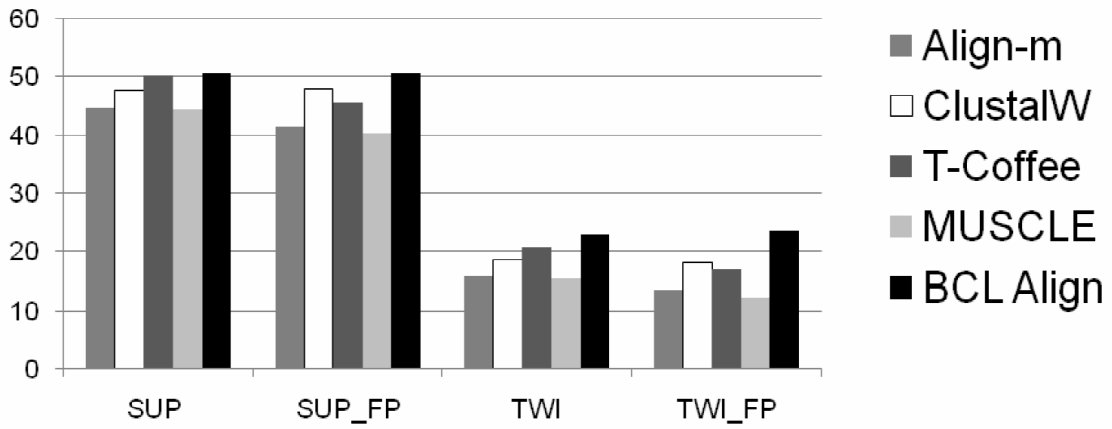


Figure 84
Performance comparison of multiple sequence alignment programs on SABmark.
Cline scores are reported for each multiple alignment program. The highest score in each subset is displayed in bold. Scores for all methods except BCL::Align are from Blackshields et al. (2006).

Performance in fold recognition

There is not a universal score for measuring fold recognition accuracy. To determine the fold recognition accuracy of BCL::Align on the SABmark benchmark database subsets that included false positives, ROC curve analysis was performed. We find that BCL::Align has a strong performance, predicting the correct structure with 86% accuracy for the Superfamily subset and 83% accuracy for the Twilight Zone subset (see Figure 85). However, the limiting factor for BCL::Align's ability to perform fold recognition is in the length of time it takes for the program to scan large databases in search of match fold, family and superfamily. Future improvements to increase the speed of BCL::Align using a word-based algorithm will allow for a more comprehensive study of the program's ability to perform fold recognition.

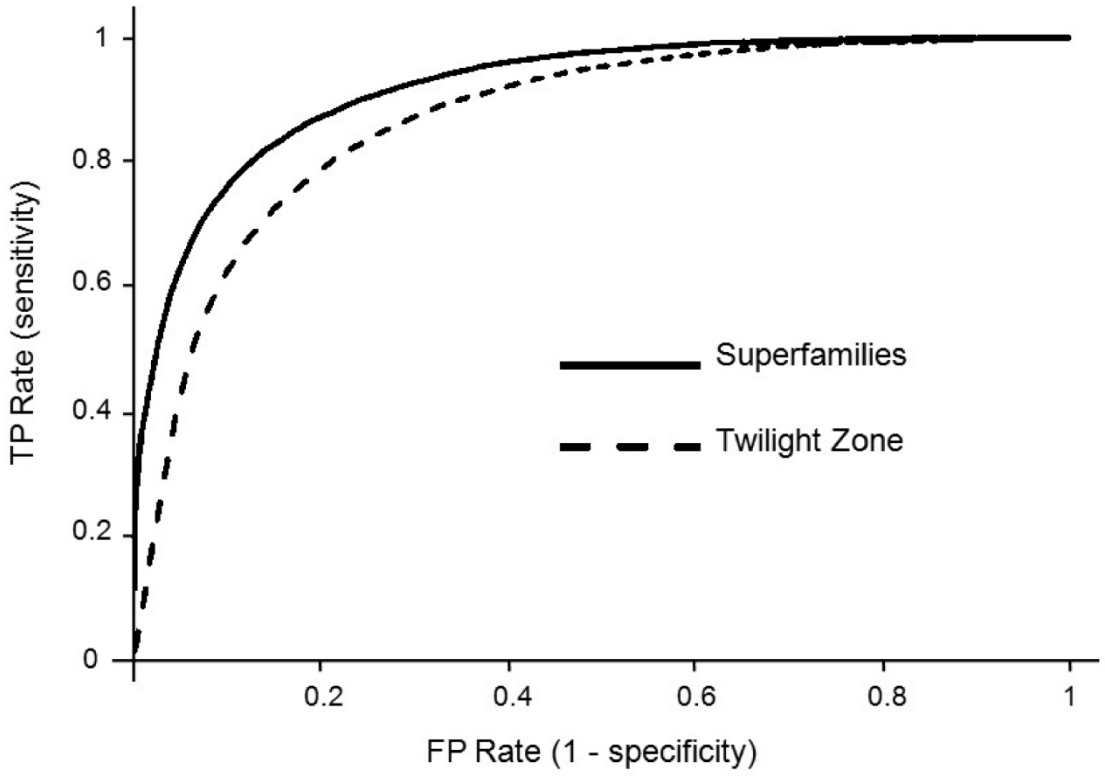


Figure 85
ROC curve analysis of fold recognition on SABmark

Conclusions

Sequence alignment and fold recognition at varying levels of sequence identity benefits from the use of customized weight sets because of the emphasis of different parameters for each situation. For the Superfamily subset, fold recognition puts an average of 12% more weight on chemical properties than sequence alignment. The BLOSUM45 substitution matrix has over 10% more weight in fold recognition than sequence alignment. Of the secondary structure predictions, PSIPRED carries the most weight with at 20-30% on average for all categories. JUFO follows behind with weights between 5 and 18%, and SAM has minimal involvement at less than 10% weight in all categories. In all cases, however, large weights for the BLAST profile and affine gap penalties provide optimal alignment and fold recognition. Using its optimal customized

weight set, BCL::Align performed better than other dynamic-programming based methods with the highest rank of sequence alignment accuracy. With the future implementation of a faster word-based algorithm and the incorporation of HMM, we expect BCL::Align to have the efficiency to quickly align multiple sequences at once and perform fold recognition over large databases of protein structures. BCL::Align is available on an online web server at http://www.meilerlab.org/.

BIBLIOGRAPHY

Abrol, R., Kim, S.K., Bray, J. K., Trzaskowski, B. & Goddard III, W. A. (2013). Conformational Ensemble View of G Protein-Coupled Receptors and the Effect of Mutations and Ligand Binding. *Methods in Enzymology*, (520), 31–48. Academic Press.

Alagille, D., Baldwin, R. M., Roth, B. L., Wroblewski, J. T., Grajkowska, E. & Tamagnan, G. D. (2005). Functionalization at position 3 of the phenyl ring of the potent mGluR5 noncompetitive antagonists MPEP. *Bioorganic & medicinal chemistry letters*, *15*(4), 945–949. Elsevier.

Alexander, N., Al-Mestarihi, A., Bortolus, M., Mchaourab, H. & Meiler, J. (2008). De novo high-resolution protein structure determination from sparse spin-labeling EPR data. *Structure*, *16*(2), 181–195. Elsevier.

Alexander, N., Woetzel, N. & Meiler, J. (2011). Bcl:: Cluster: A method for clustering biological molecules coupled with visualization in the Pymol Molecular Graphics System. *Computational Advances in Bio and Medical Sciences (ICCABS), 2011 IEEE 1st International Conference on* (pp. 13–18).

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of molecular biology*, *215*(3), 403–410. Elsevier.

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research*, *25*(17), 3389–3402. Oxford Univ Press.

Ametamey, S. M., Treyer, V., Streffer, J., Wyss, M. T., Schmidt, M., Blagoev, M., Hintermann, S., et al. (2007). Human PET studies of metabotropic glutamate receptor subtype 5 with 11C-ABP688. *Journal of Nuclear Medicine*, *48*(2), 247–252. Soc Nuclear Med.

Anderson, A. C. (2003). The process of structure-based drug design. *Chemistry & biology*, *10*(9), 787–797. Elsevier Science.

Anfinsen, C. (1973). Principles that govern the folding of protein chains. *Science*, *181*(96), 223–230.

Ayala, J. E., Chen, Y., Banko, J. L., Sheffler, D. J., Williams, R., Telk, A. N., Watson, N. L., et al. (2009). mGluR5 positive allosteric modulators facilitate both hippocampal LTP and LTD and enhance spatial learning. *Neuropsychopharmacology*, *34*(9), 2057–2071. Nature Publishing Group.

B Singh, S. & M Tice, C. (2010). Structure Based Design of 11-HSD1 Inhibitors. *Current Pharmaceutical Biotechnology*, *11*(7), 779–791. Bentham Science Publishers.

Bairoch, A. & Apweiler, R. (1998). The SWISS-PROT protein sequence data bank and its supplement TrEMBL in 1998. *Nucleic Acids Research*, *26*(1), 38–42. Oxford Univ

Press.

Baker, D. (2000). A surprising simplicity to protein folding. *Nature*, *405*(6782), 39–42. Nature Publishing Group.

Baker, D. & Sali, A. (2001). Protein structure prediction and structural genomics. *Science Signalling*, *294*(5540), 93. AAAS.

Baker, J. G., Hill, S. J. & Summers, R. J. (2011). Evolution of beta-blockers: from anti-anginal drugs to ligand-directed signalling. *Trends in pharmacological sciences*, *32*(4), 227–234. Elsevier.

Bakker, R. A., Jongejan, A., Sansuk, K., Hacksell, U., Timmerman, H., Brann, M. R., Weiner, D. M., et al. (2008). Constitutively active mutants of the histamine H1 receptor suggest a conserved hydrophobic asparagine-cage that constrains the activation of class AG protein-coupled receptors. *Molecular pharmacology*, *73*(1), 94–103. ASPET.

Ballester, P. J., Westwood, I., Laurieri, N., Sim, E. & Richards, W. G. (2010). Prospective virtual screening with Ultrafast Shape Recognition: the identification of novel inhibitors of arylamine N-acetyltransferases. *Journal of The Royal Society Interface*, *7*(43), 335–342. The Royal Society.

Ballesteros, J. A. & Weinstein, H. (1995). Integrated methods for the construction of three-dimensional models and computational probing of structure-function relations in G protein-coupled receptors. *Methods in neurosciences*, *25*, 366–428. Elsevier.

Barak, L. S., Menard, L., Ferguson, S. S., Colapietro, A.M. & Caron, M. G. (1995). The Conserved Seven-Transmembrane Sequence NP (X) 2, 3Y of the G-Protein-Coupled Receptor Superfamily Regulates Multiple Properties of the. beta. 2-Adrenergic Receptor. *Biochemistry*, *34*(47), 15407–15414. ACS Publications.

Barnard, E. A. & Simon, J. (2001). An elusive receptor is finally caught: P2Y12, an important drug target in platelets. *Trends in pharmacological sciences*, *22*(8), 388–391. Elsevier.

Barril, X. & Morley, S. D. (2005). Unveiling the full potential of flexible receptor docking using multiple crystallographic structures. *Journal of medicinal chemistry*, *48*(13), 4432–4443. ACS Publications.

Barth, P., Schonbrun, J. & Baker, D. (2007). Toward high-resolution prediction and design of transmembrane helical protein structures. *Proceedings of the National Academy of Sciences*, *104*(40), 15682–15687. National Acad Sciences.

Barton, G. J. & Sternberg, M. J. (1987). Evaluation and improvements in the automatic alignment of protein sequences. *Protein engineering*, *1*(2), 89–94. Oxford Univ Press.

Baugh, L., Le Trong, I., Cerutti, D. S., Gülich, S., Stayton, P. S., Stenkamp, R. E. & Lybrand, T. P. (2010). A Distal Point Mutation in the Streptavidin- Biotin Complex Preserves Structure but Diminishes Binding Affinity: Experimental Evidence of Electronic Polarization Effects? *Biochemistry*, *49*(22), 4568–4570. ACS Publications.

Becker, O. M., Dhanoa, D. S., Marantz, Y., Chen, D., Shacham, S., Cheruku, S., Heifetz, A., et al. (2006). An integrated in silico 3D model-driven discovery of a novel, potent, and selective amidosulfonamide 5-HT1A agonist (PRX-00023) for the treatment of anxiety and depression. *Journal of medicinal chemistry*, *49*(11), 3116–3135. ACS Publications.

Becker, O. M., Marantz, Y., Shacham, S., Inbal, B., Heifetz, A., Kalid, O., Bar-Haim, S., et al. (2004). G protein-coupled receptors: in silico drug discovery in 3D. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(31), 11304–11309. National Acad Sciences.

Beigi, R., Kobatake, E., Aizawa, M. & Dubyak, G. R. (1999). Detection of local ATP release from activated platelets using cell surface-attached firefly luciferase. *American Journal of Physiology-Cell Physiology*, *276*(1), C267–C278. Am Physiological Soc.

van Bennekom, M. W. L., Gijsman, H. J. & Zitman, F. G. (2013). Antipsychotic polypharmacy in psychotic disorders: a critical review of neurobiology, efficacy, tolerability and cost effectiveness. *Journal of Psychopharmacology*. SAGE Publications.

Bennett-Lovsey, R. M., Herbert, A. D., Sternberg, M. J. & Kelley, L. A. (2008). Exploring the extremes of sequence/structure space with ensemble fold recognition in the program Phyre. *Proteins: Structure, Function, and Bioinformatics*, *70*(3), 611–625. Wiley Online Library.

Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J. & Wheeler, D. L. (2006). GenBank. *Nucleic acids research*, *34*(suppl 1), D16–D20. Oxford Univ Press.

Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J. & Wheeler, D. L. (2007). GenBank. *D21-D25*.

Berman, H., Henrick, K. & Nakamura, H. (2003). Announcing the worldwide protein data bank. *Nature Structural & Molecular Biology*, *10*(12), 980–980. Nature Publishing Group.

Berman, H., Henrick, K., Nakamura, H. & Markley, J. L. (2007). The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. *Nucleic acids research*, *35*(suppl 1), D301–D303. Oxford Univ Press.

Berridge, M. J. (1983). Rapid accumulation of inositol trisphosphate reveals that agonists hydrolyse polyphosphoinositides instead of phosphatidylinositol. *Biochemical journal*, *212*(3), 849. Portland Press Ltd.

Bessis, A.S., Acher, F., Bertrand, H.O., Galvez, T., De Colle, C. & Pin, J.P. (2000). Three-dimensional model of the extracellular domain of the type 4a metabotropic glutamate receptor: New insights into the activation process. *Protein Science*, *9*(11), 2200–2209. Wiley Online Library.

Beuming, T. & Sherman, W. (2012). Current Assessment of Docking into GPCR Crystal Structures and Homology Models: Successes, Challenges, and Guidelines. *Journal of chemical information and modeling*, *52*(12), 3263–3277. ACS Publications.

Bhave, G., Nadin, B. M., Brasier, D., Glauner, K. S., Shah, R. D., Heinemann, S. F., Karim, F., et al. (2003). Membrane topology of a metabotropic glutamate receptor. *Journal of Biological Chemistry*, *278*(32), 30294–30301. ASBMB.

Blackshields, G., Wallace, I. M., Larkin, M. & Higgins, D. G. (2006). Analysis and comparison of benchmarks for multiple sequence alignment. *In silico biology*, *6*(4), 321–339. IOS Press.

Bodor, E. T., Waldo, G. L., Blaesius, R. & Harden, T. K. (2004). Delineation of ligand binding and receptor signaling activities of purified P2Y receptors reconstituted with heterotrimeric G proteins. *Purinergic signalling*, *1*(1), 43–49. Springer.

Bodor, E. T., Waldo, G. L., Hooks, S. B., Corbitt, J., Boyer, J. L. & Harden, T. K. (2003). Purification and functional reconstitution of the human P2Y12 receptor. *Molecular pharmacology*, *64*(5), 1210–1216. ASPET.

Bond, C. S. & Schuttelkopf, A. W. (2009). ALINE: a WYSIWYG protein-sequence alignment editor for publication-quality alignments. *Acta Crystallographica Section D: Biological Crystallography*, *65*(5), 510–512. International Union of Crystallography.

Boström, J., Norrby, P.O. & Liljefors, T. (1998). Conformational energy penalties of protein-bound ligands. *Journal of computer-aided molecular design*, *12*(4), 383–383. Springer.

Bottegoni, G., Kufareva, I., Totrov, M. & Abagyan, R. (2008). Four-dimensional docking: a fast and accurate account of discrete receptor flexibility in ligand docking. *Journal of medicinal chemistry*, *52*(2), 397–406. ACS Publications.

Boudreaux, M. K. & Martin, M. (2011). P2Y12 receptor gene mutation associated with postoperative hemorrhage in a Greater Swiss Mountain dog. *Veterinary Clinical Pathology*, *40*(2), 202–206. Wiley Online Library.

Boutonnet, N. S., Rooman, M. J., Ochagavia, M.E., Richelle, J. & Wodak, S. J. (1995). Optimal protein structure alignments by multiple linkage clustering: application to distantly related proteins. *Protein engineering*, *8*(7), 647–662. Oxford Univ Press.

Bowie, J. U. (1997). Helix packing angle preferences. *Nature structural biology*, *4*(11), 915.

Bradley, P., Malmström, L., Qian, B., Schonbrun, J., Chivian, D., Kim, D. E., Meiler, J., et al. (2005). Free modeling with Rosetta in CASP6. *Proteins: Structure, Function, and Bioinformatics*, *61*(S7), 128–134. Wiley Online Library.

Bradley, P., Misura, K. M. & Baker, D. (2005). Toward high-resolution de novo structure prediction for small proteins. *Science*, *309*(5742), 1868–1871. American Association for the Advancement of Science.

Bräuner-Osborne, H., Jensen, A. A. & Krogsgaard-Larsen, P. (1999). Interaction of CPCCOEt with a chimeric mGlu1b and calcium sensing receptor. *Neuroreport*, *10*(18), 3923–3925. LWW.

Caffrey, M. (2003). Membrane protein crystallization. *Journal of structural biology*, *142*(1), 108–132. Elsevier.

Canutescu, A. A. & Dunbrack, R. L. (2009). Cyclic coordinate descent: a robotics algorithm for protein loop closure. *Protein Science*, *12*(5), 963–972. Wiley Online Library.

Carlsson, J., Coleman, R. G., Setola, V., Irwin, J. J., Fan, H., Schlessinger, A., Sali, A., et al. (2011). Ligand discovery from a dopamine D3 receptor homology model and crystal structure. *Nature Chemical Biology*. Nature Publishing Group.

Carroll, F. Y., Stolle, A., Beart, P. M., Voerste, A., Brabet, I., Mauler, F., Joly, C., et al. (2001). BAY36-7620: a potent non-competitive mGlu1 receptor antagonist with inverse agonist activity. *Molecular pharmacology*, *59*(5), 965–973. ASPET.

Carugo, O. & Pongor, S. (2001). A normalized root-mean-square distance for comparing protein three-dimensional structures. *Protein science*, *10*(7), 1470–1473. Wiley Online Library.

Case, D. A., Cheatham, T. E., Darden, T., Gohlke, H., Luo, R., Merz, K. M., Onufriev, A., et al. (2005). The Amber biomolecular simulation programs. *Journal of computational chemistry*, *26*(16), 1668–1688. Wiley Online Library.

Castillo-Davis, C. I., Kondrashov, F. A., Hartl, D. L. & Kulathinal, R. J. (2004). The functional genomic distribution of protein divergence in two animal phyla: coevolution, genomic conflict, and constraint. *Genome research*, *14*(5), 802–811. Cold Spring Harbor Lab.

Cattaneo, M. (2005). The P2 receptors and congenital platelet function defects. *Seminars in thrombosis and hemostasis* (Vol. 31, pp. 168–173).

Cattaneo, M., Lecchi, A., Lombardi, R., Gachet, C. & Zighetti, M. L. (2000). Platelets From a Patient Heterozygous for the Defect of P2CYC Receptors for ADP Have a Secretion Defect Despite Normal Thromboxane A2 Production and Normal Granule Stores Further Evidence That Some Cases of Platelet “Primary Secretion Defect”Are Heterozygous for a Defect of P2CYC Receptors. *Arteriosclerosis, thrombosis, and vascular biology*, *20*(11), e101–e106. Am Heart Assoc.

Cattaneo, M., Zighetti, M. L., Lombardi, R., Martinez, C., Lecchi, A., Conley, P. B., Ware, J., et al. (2003). Molecular bases of defective signal transduction in the platelet P2Y12 receptor of a patient with congenital bleeding. *Proceedings of the National Academy of Sciences*, *100*(4), 1978–1983. National Acad Sciences.

Chee, M. J., Mörl, K., Lindner, D., Merten, N., Zamponi, G. W., Light, P. E., Beck-Sickinger, A. G., et al. (2008). The third intracellular loop stabilizes the inactive state

of the neuropeptide Y1 receptor. *Journal of Biological Chemistry*, *283*(48), 33337–33346. ASBMB.

Chen, K.Y. M., Zhou, F., Fryszczyn, B. G. & Barth, P. (2012). Naturally evolved G protein-coupled receptors adopt metastable conformations. *Proceedings of the National Academy of Sciences*, *109*(33), 13284–13289. National Acad Sciences.

Chen, Y., Goudet, C., Pin, J.P. & Conn, P. J. (2008). N-4-Chloro-2-[(1, 3-dioxo-1, 3-dihydro-2H-isoindol-2-yl) methyl] phenyl-2-hydroxybenzamide (CPPHA) acts through a novel site as a positive allosteric modulator of group 1 metabotropic glutamate receptors. *Molecular pharmacology*, *73*(3), 909–918. ASPET.

Chen, Y., Nong, Y., Goudet, C., Hemstapat, K., De Paulis, T., Pin, J.P. & Conn, P. J. (2007). Interaction of novel positive allosteric modulators of metabotropic glutamate receptor 5 with the negative allosteric antagonist site is required for potentiation of receptor responses. *Molecular pharmacology*, *71*(5), 1389–1398. ASPET.

Cheng, Y.C., Prusoff, W. H. & others. (1973). Relationship between the inhibition constant (K1) and the concentration of inhibitor which causes 50 per cent inhibition (I50) of an enzymatic reaction. *Biochemical pharmacology*, *22*(23), 3099.

Cherezov, V., Rosenbaum, D. M., Hanson, M. A., Rasmussen, S. G., Thian, F. S., Kobilka, T. S., Choi, H.J., et al. (2007). High-resolution crystal structure of an engineered human 2-adrenergic G protein coupled receptor. *Science Signalling*, *318*(5854), 1258. AAAS.

Chien, E. Y., Liu, W., Zhao, Q., Katritch, V., Won Han, G., Hanson, M. A., Shi, L., et al. (2010). Structure of the human dopamine D3 receptor in complex with a D2/D3 selective antagonist. *Science Signaling*, *330*(6007), 1091. AAAS.

Chivian, D. & Baker, D. (2006). Homology modeling using parametric alignment ensemble generation with consensus and energy-based model selection. *Nucleic acids research*, *34*(17), e112–e112. Oxford Univ Press.

Christopoulos, A. (1998). Assessing the distribution of parameters in models of ligand-receptor interaction: to log or not to log. *Trends in pharmacological sciences*, *19*(9), 351–357. Elsevier.

Cline, M., Hughey, R. & Karplus, K. (2002). Predicting reliable regions in protein sequence alignments. *Bioinformatics*, *18*(2), 306–314. Oxford Univ Press.

Combs, S., Kaufmann, K., Field, J. R., Blakely, R. D. & Meiler, J. (2010). Y95 and E444 interaction required for high-affinity S-Citalopram binding in the human serotonin transporter. *ACS Chemical Neuroscience*, *2*(2), 75–81. ACS Publications.

Congreve, M. & Marshall, F. (2009). The impact of GPCR structures on pharmacology and structure-based drug design. *British journal of pharmacology*, *159*(5), 986–996. Wiley Online Library.

Congreve, M., Murray, C. W. & Blundell, T. L. (2005). Keynote review: Structural biology and drug discovery. *Drug discovery today*, *10*(13), 895–907. Elsevier.

Conn, P. J., Lindsley, C. W. & Jones, C. K. (2009). Activation of metabotropic glutamate receptors as a novel approach for the treatment of schizophrenia. *Trends in pharmacological sciences*, *30*(1), 25. NIH Public Access.

Conn, P. J. & Pin, J.P. (1997). Pharmacology and functions of metabotropic glutamate receptors. *Annual review of pharmacology and toxicology*, *37*(1), 205–237. Annual Reviews 4139 El Camino Way, PO Box 10139, Palo Alto, CA 94303-0139, USA.

Cosford, N. D., Roppe, J., Tehrani, L., Schweiger, E. J., Seiders, T. J., Chaudary, A., Rao, S., et al. (2003). [3H]-Methoxymethyl-MTEP and [3H]-Methoxy-PEPy: potent and selective radioligands for the metabotropic glutamate subtype 5 (mGlu5) receptor. *Bioorganic & medicinal chemistry letters*, *13*(3), 351–354. Elsevier.

Cosford, N. D., Tehrani, L., Roppe, J., Schweiger, E., Smith, N. D., Anderson, J., Bristow, L., et al. (2003). 3-[(2-Methyl-1, 3-thiazol-4-yl) ethynyl]-pyridine: a potent and highly selective metabotropic glutamate subtype 5 receptor antagonist with anxiolytic activity. *Journal of medicinal chemistry*, *46*(2), 204–206. ACS Publications.

Coutsias, E. A., Seok, C., Jacobson, M. P. & Dill, K. A. (2004). A kinematic view of loop closure. *Journal of computational chemistry*, *25*(4), 510–528. Wiley Online Library.

Cozzini, P., Kellogg, G. E., Spyrakis, F., Abraham, D. J., Costantino, G., Emerson, A., Fanelli, F., et al. (2008). Target Flexibility: An Emerging Consideration in Drug Discovery and Design†. *Journal of medicinal chemistry*, *51*(20), 6237. NIH Public Access.

Daly, M. E., Dawood, B. B., Lester, W. A., Peake, I. R., Rodeghiero, F., Goodeve, A. C., Makris, M., et al. (2009). Identification and characterization of a novel P2Y12 variant in a patient diagnosed with type 1 von Willebrand disease in the European MCMDM-1VWD study. *Blood*, *113*(17), 4110–4113. American Society of Hematology.

Das, R., André, I., Shen, Y., Wu, Y., Lemak, A., Bansal, S., Arrowsmith, C. H., et al. (2009). Simultaneous prediction of protein folding and docking at high resolution. *Proceedings of the National Academy of Sciences*, *106*(45), 18978–18983. National Acad Sciences.

Das, R. & Baker, D. (2008). Macromolecular modeling with rosetta. *Annu. Rev. Biochem.*, *77*, 363–382. Annual Reviews.

Das, R., Qian, B., Raman, S., Vernon, R., Thompson, J., Bradley, P., Khare, S., et al. (2007). Structure prediction for CASP7 targets using extensive all-atom refinement with Rosetta@ home. *Proteins: Structure, Function, and Bioinformatics*, *69*(S8), 118–128. Wiley Online Library.

Davis, I. W. & Baker, D. (2009). RosettaLigand docking with full ligand and receptor flexibility. *Journal of molecular biology*, *385*(2), 381–392. Elsevier.

Davis, I. W., Raha, K., Head, M. S. & Baker, D. (2009). Blind docking of pharmaceutically relevant compounds using RosettaLigand. *Protein science*, *18*(9),

1998–2002. Wiley Online Library.

Dayhoff, M. O. & Schwartz, R. M. (1978). A model of evolutionary change in proteins. *In Atlas of protein sequence and structure*.

Deflorian, F. & Jacobson, K. A. (2011). Comparison of three GPCR structural templates for modeling of the P2Y 12 nucleotide receptor. *Journal of computer-aided molecular design*, *25*(4), 329–338. Springer.

DiMaio, F., Terwilliger, T. C., Read, R. J., Wlodawer, A., Oberdorfer, G., Wagner, U., Valkov, E., et al. (2011). Improved molecular replacement by density-and energy-guided protein structure optimization. *Nature*, *473*(7348), 540–543. Nature Publishing Group.

Ding, Z., Kim, S., Dorsam, R. T., Jin, J. & Kunapuli, S. P. (2003). Inactivation of the human P2Y12 receptor by thiol reagents requires interaction with both extracellular cysteine residues, Cys17 and Cys270. *Blood*, *101*(10), 3908–3914. American Society of Hematology.

Ding, Z., Kim, S. & Kunapuli, S. P. (2006). Identification of a potent inverse agonist at a constitutively active mutant of human P2Y12 receptor. *Molecular pharmacology*, *69*(1), 338–345. ASPET.

Dingledine, R., Borges, K., Bowie, D. & Traynelis, S. F. (1999). The glutamate receptor ion channels. *Pharmacological reviews*, *51*(1), 7–62. ASPET.

Do, C. B., Mahabhashyam, M. S., Brudno, M. & Batzoglou, S. (2005). ProbCons: Probabilistic consistency-based multiple sequence alignment. *Genome Research*, *15*(2), 330–340. Cold Spring Harbor Lab.

Dong, E., Smith, J., Heinze, S., Alexander, N. & Meiler, J. (2008). BCL:: Align—Sequence alignment and fold recognition with a custom scoring function online. *Gene*, *422*(1), 41–46. Elsevier.

Dorr, P., Westby, M., Dobbs, S., Griffin, P., Irvine, B., Macartney, M., Mori, J., et al. (2005). Maraviroc (UK-427,857), a potent, orally bioavailable, and selective small-molecule inhibitor of chemokine receptor CCR5 with broad-spectrum anti-human immunodeficiency virus type 1 activity. *Antimicrobial agents and chemotherapy*, *49*(11), 4721–4732. Am Soc Microbiol.

Dunbrack Jr, R. L. & Karplus, M. (1993). Backbone-dependent rotamer library for proteins. *J. Mol. Biol*, *230*(2), 543–574.

Dunbrack, R. L. & Cohen, F. E. (1997). Bayesian statistical analysis of protein side-chain rotamer preferences. *Protein Science*, *6*(8), 1661–1681. Wiley Online Library.

Durham, E., Dorr, B., Woetzel, N., Staritzbichler, R. & Meiler, J. (2009). Solvent accessible surface area approximations for rapid and accurate protein structure prediction. *Journal of molecular modeling*, *15*(9), 1093–1108. Springer.

Dölen, G. & Bear, M. F. (2008). Role for metabotropic glutamate receptor 5 (mGluR5) in the pathogenesis of fragile X syndrome. *The Journal of physiology*, *586*(6), 1503–1508. Wiley Online Library.

Eddy, S. R. (2011). Accelerated profile HMM searches. *PLoS computational biology*, *7*(10), e1002195. Public Library of Science.

Eddy, S. R. & others. (2009). A new generation of homology search tools based on probabilistic inference. *Genome Inform* (Vol. 23, pp. 205–211).

Edgar, R. (2004). MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC bioinformatics*, *5*(1), 113. BioMed Central Ltd.

Edgar, R. C. & Sjölander, K. (2004). A comparison of scoring functions for protein sequence profile alignment. *Bioinformatics*, *20*(8), 1301–1308. Oxford Univ Press.

Ehlers, M. D. (1999). Synapse structure: glutamate receptors connected by the shanks. *Current biology*, *9*(22), R848–R850. Elsevier.

El Omari, K., Liekens, S., Bird, L. E., Balzarini, J. & Stammers, D. K. (2006). Mutations distal to the substrate site can affect varicella zoster virus thymidine kinase activity: implications for drug design. *Molecular pharmacology*, *69*(6), 1891–1896. ASPET.

Emkey, R. & Rankl, N. B. (2009). Screening G protein-coupled receptors: measurement of intracellular calcium using the fluorometric imaging plate reader. *Methods Mol Biol*, *565*, 145–158. Springer.

Van Eps, N., Preininger, A. M., Alexander, N., Kaya, A. I., Meier, S., Meiler, J., Hamm, H. E., et al. (2011). Interaction of a G protein with an activated receptor opens the interdomain interface in the alpha subunit. *Proceedings of the National Academy of Sciences*, *108*(23), 9420–9424. National Acad Sciences.

Eswar, N., Eramian, D., Webb, B., Shen, M.Y. & Sali, A. (2008). Protein structure modeling with MODELLER. *Structural Proteomics* (pp. 145–159). Springer.

Ewing, T. J., Makino, S., Skillman, A. G. & Kuntz, I. D. (2001). DOCK 4.0: search strategies for automated molecular docking of flexible molecule databases. *Journal of computer-aided molecular design*, *15*(5), 411–428. Springer.

Eyal, E., Najmanovich, R., Mcconkey, B. J., Edelman, M. & Sobolev, V. (2004). Importance of solvent accessibility and contact surfaces in modeling side-chain conformations in proteins. *Journal of computational chemistry*, *25*(5), 712–724. Wiley Online Library.

Fagerberg, L., Jonasson, K., von Heijne, G., Uhlén, M. & Berglund, L. (2010). Prediction of the human membrane proteome. *Proteomics*, *10*(6), 1141–1149. Wiley Online Library.

Felts, A. S., Lindsley, S. R., Lamb, J. P., Rodriguez, A. L., Menon, U. N., Jadhav, S., Jones, C. K., et al. (2010). 3-Cyano-5-fluoro-N-arylbenzamides as negative allosteric modulators of mGlu5: Identification of easily prepared tool compounds with CNS

exposure in rats. *Bioorganic & medicinal chemistry letters*, *20*(15), 4390–4394. Elsevier.

Feng, D.F. & Doolittle, R. F. (1987). Progressive sequence alignment as a prerequisitetto correct phylogenetic trees. *Journal of molecular evolution*, *25*(4), 351–360. Springer.

Field, J. R., Henry, L. K. & Blakely, R. D. (2010). Transmembrane domain 6 of the human serotonin transporter contributes to an aqueously accessible binding pocket for serotonin and the psychostimulant 3, 4-methylene dioxymethamphetamine. *Journal of Biological Chemistry*, *285*(15), 11270–11280. ASBMB.

Filtz, T. M., Li, Q., Boyer, J. L., Nicholas, R. A. & Harden, T. K. (1994). Expression of a cloned P2Y purinergic receptor that couples to phospholipase C. *Molecular pharmacology*, *46*(1), 8–14. ASPET.

Fiser, A., Do, R. K. G. & Sali, A. (2008). Modeling of loops in protein structures. *Protein science*, *9*(9), 1753–1773. Wiley Online Library.

Fleishman, S. J. & Baker, D. (2012). Role of the biomolecular energy gap in protein design, structure, and evolution. *Cell*, *149*(2), 262–273. Elsevier.

Fleishman, S. J., Leaver-Fay, A., Corn, J. E., Strauch, E.M., Khare, S. D., Koga, N., Ashworth, J., et al. (2011). RosettaScripts: a scripting language interface to the Rosetta macromolecular modeling suite. *PloS one*, *6*(6), e20161. Public Library of Science.

Fontana, G., Ware, J. & Cattaneo, M. (2009). Haploinsufficiency of the platelet P2Y12 gene in a family with congenital bleeding diathesis. *haematologica*, *94*(4), 581–584. Ferrata Storti Foundation.

Foord, S. M., Bonner, T. I., Neubig, R. R., Rosser, E. M., Pin, J.P., Davenport, A. P., Spedding, M., et al. (2005). International Union of Pharmacology. XLVI. G protein-coupled receptor list. *Pharmacological reviews*, *57*(2), 279–288. American Society for Pharmacology and Experimental Therapeutics.

Fortenberry, C., Bowman, E. A., Proffitt, W., Dorr, B., Combs, S., Harp, J., Mizoue, L., et al. (2011). Exploring symmetry as an avenue to the computational design of large protein domains. *Journal of the American Chemical Society*, *133*(45), 18026–18029. ACS Publications.

Friesner, R. A., Banks, J. L., Murphy, R. B., Halgren, T. A., Klicic, J. J., Mainz, D. T., Repasky, M. P., et al. (2004). Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *Journal of medicinal chemistry*, *47*(7), 1739–1749. ACS Publications.

Frishman, W. H. (1998). Carvedilol. *The New England journal of medicine*, *339*(24), 1759.

Fritze, O., Filipek, S., Kuksa, V., Palczewski, K., Hofmann, K. P. & Ernst, O. P. (2003). Role of the conserved NPxxY (x) 5, 6F motif in the rhodopsin ground state and during

activation. *Proceedings of the National Academy of Sciences*, *100*(5), 2290–2295. National Acad Sciences.

Fukuda, J., Suzuki, G., Kimura, T., Nagatomi, Y., Ito, S., Kawamoto, H., Ozaki, S., et al. (2009). Identification of a novel transmembrane domain involved in the negative modulation of mGluR1 using a newly discovered allosteric mGluR1 antagonist, 3-cyclohexyl-5-fluoro-6-methyl-7-(2-morpholin-4-ylethoxy)-4H-chromen-4-one. *Neuropharmacology*, *57*(4), 438–445. Elsevier.

Gasparini, F., Floersheim, P., Flor, P. J., Heinrich, M., Inderbitzin, W., Ott, D., Pagano, A., et al. (2001). Discovery and characterization of non-competitive antagonists of group I metabotropic glutamate receptors. *Il Farmaco*, *56*(1), 95–99. Elsevier.

Gasparini, F., Kuhn, R., Pin, J.P. & others. (2002). Allosteric modulators of group I metabotropic glutamate receptors: novel subtype-selective ligands and therapeutic perspectives. *Current opinion in pharmacology*, *2*(1), 43.

Gasparini, F., Lingenhöhl, K., Stoehr, N., Flor, P. J., Heinrich, M., Vranesic, I., Biollaz, M., et al. (1999). 2-Methyl-6-(phenylethynyl)-pyridine (MPEP), a potent, selective and systemically active mGlu5 receptor antagonist. *Neuropharmacology*, *38*(10), 1493–1503. Elsevier.

Gelis, L., Wolf, S., Hatt, H., Neuhaus, E. M. & Gerwert, K. (2012). Prediction of a Ligand-Binding Niche within a Human Olfactory Receptor by Combining Site-Directed Mutagenesis with Dynamic Homology Modeling. *Angewandte Chemie International Edition*, *51*(5), 1274–1278. Wiley Online Library.

Ginalski, K., Elofsson, A., Fischer, D. & Rychlewski, L. (2003). 3D-Jury: a simple approach to improve protein structure predictions. *Bioinformatics*, *19*(8), 1015–1018. Oxford Univ Press.

Ginalski, K., Pas, J., Wyrwicz, L. S., Von Grotthuss, M., Bujnicki, J. M. & Rychlewski, L. (2003). ORFeus: detection of distant homology using sequence profiles and predicted secondary structure. *Nucleic acids research*, *31*(13), 3804–3807. Oxford Univ Press.

Gohlke, H., Hendlich, M., Klebe, G. & others. (2000). Knowledge-based scoring function to predict protein-ligand interactions. *Journal of molecular biology*, *295*(2), 337–356. Elsevier Science.

Goldfeld, D. A., Zhu, K., Beuming, T. & Friesner, R. A. (2011). Successful prediction of the intra-and extracellular loops of four G-protein-coupled receptors. *Proceedings of the National Academy of Sciences*, *108*(20), 8275–8280. National Acad Sciences.

Goldfeld, D. A., Zhu, K., Beuming, T. & Friesner, R. A. (2012). Loop prediction for a GPCR homology model: Algorithms and results. *Proteins: Structure, Function, and Bioinformatics*. Wiley Online Library.

Goldflam, M., Tarragó, T., Gairi, M. & Giralt, E. (2012). NMR Studies of Protein-Ligand Interactions. *Protein NMR Techniques* (pp. 233–259). Springer.

Gordon, D. B., Marshall, S. A. & Mayot, S. L. (1999). Energy functions for protein design. *Current opinion in structural biology*, *9*(4), 509–513. Elsevier.

Goudet, C., Gaven, F., Kniazeff, J., Vol, C., Liu, J., Cohen-Gonsaud, M., Acher, F., et al. (2004). Heptahelical domain of metabotropic glutamate receptor 5 behaves like rhodopsin-like receptors. *Proceedings of the National Academy of Sciences*, *101*(1), 378–383. National Acad Sciences.

de Graaf, C., Kooistra, A. J., Vischer, H. F., Katritch, V., Kuijer, M., Shiroishi, M., Iwata, S., et al. (2011). Crystal structure-based virtual screening for fragment-like ligands of the human histamine H1 receptor. *Journal of medicinal chemistry*, *54*(23), 8195–8206. ACS Publications.

Granier, S., Manglik, A., Kruse, A. C., Kobilka, T. S., Thian, F. S., Weis, W. I. & Kobilka, B. K. (2012). Structure of the delta-opioid receptor bound to naltrindole. *Nature*, *485*(7398), 400–404. Nature Publishing Group.

Gregory, K. J., Dong, E. N., Meiler, J. & Conn, P. J. (2011). Allosteric modulation of metabotropic glutamate receptors: structural insights and therapeutic potential. *Neuropharmacology*, *60*(1), 66–81. Elsevier.

Gregory, K. J., Nguyen, E. D., Reiff, S. D., Squire, E. F., Stauffer, S. R., Lindsley, C. W., Meiler, J., et al. (2013). Probing the Metabotropic Glutamate Receptor 5 (mGlu5) Positive Allosteric Modulator (PAM) Binding Pocket: Discovery of Point Mutations That Engender a “Molecular Switch” in PAM Pharmacology. *Molecular pharmacology*, *83*(5), 991–1006. doi:10.1124/mol.112.083949

Gregory, K. J., Noetzel, M. J., Rook, J. M., Vinson, P. N., Stauffer, S. R., Rodriguez, A. L., Emmitte, K. A., et al. (2012). Investigating mGlu5 Allosteric Modulator Cooperativity, Affinity and Agonism: Enriching Structure-function Studies and Structure-activity Relationships. *Molecular Pharmacology*. ASPET.

Grishaev, A., Guo, L., Irving, T. & Bax, A. (2010). Improved fitting of solution X-ray scattering data to macromolecular structures and structural ensembles by explicit water modeling. *Journal of the American Chemical Society*, *132*(44), 15484–15486. ACS Publications.

Groban, E. S., Narayanan, A. & Jacobson, M. P. (2006). Conformational changes in protein loops and helices induced by post-translational phosphorylation. *PLoS computational biology*, *2*(4), e32. Public Library of Science.

Gront, D., Kulp, D. W., Vernon, R. M., Strauss, C. E. & Baker, D. (2011). Generalized fragment picking in rosetta: design, protocols and applications. *PloS one*, *6*(8), e23294. Public Library of Science.

Gupta, S. K., Pillarisetti, K., Thomas, R. A. & Aiyar, N. (2001). Pharmacological evidence for complex and multiple site interaction of CXCR4 with SDF-1alpha: implications for development of selective CXCR4 antagonists. *Immunology letters*, *78*(1), 29–34. Elsevier.

Haga, K., Kruse, A. C., Asada, H., Yurugi-Kobayashi, T., Shiroishi, M., Zhang, C., Weis, W. I., et al. (2012). Structure of the human M2 muscarinic acetylcholine receptor bound to an antagonist. *Nature*. Nature Publishing Group.

Hagerman, R. J., Berry-Kravis, E., Kaufmann, W. E., Ono, M. Y., Tartaglia, N., Lachiewicz, A., Kronk, R., et al. (2009). Advances in the treatment of fragile X syndrome. *Pediatrics*, *123*(1), 378–390. Am Acad Pediatrics.

Hammond, A. S., Rodriguez, A. L., Townsend, S. D., Niswender, C. M., Gregory, K. J., Lindsley, C. W. & Conn, P. J. (2010). Discovery of a novel chemical class of mGlu5 allosteric ligands with distinct modes of pharmacology. *ACS Chemical Neuroscience*, *1*(10), 702. American Chemical Society.

Hanson, M. A., Roth, C. B., Jo, E., Griffith, M. T., Scott, F. L., Reinhart, G., Desale, H., et al. (2012). Crystal structure of a lipid G protein-coupled receptor. *Science Signaling*, *335*(6070), 851. AAAS.

Heifetz, A., Barker, O., Verquin, G. P., Wimmer, N., Meutermans, W., Pal, S., Law, R. J., et al. (2013). Fighting Obesity with a Sugar-Based Library: Discovery of Novel MCH-1R Antagonists by a New Computational-VAST Approach for Exploration of GPCR Binding Sites. *Journal of chemical information and modeling*. ACS Publications.

Henikoff, S. & Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences*, *89*(22), 10915–10919. National Acad Sciences.

Herzog, F., Kahraman, A., Boehringer, D., Mak, R., Bracher, A., Walzthoeni, T., Leitner, A., et al. (2012). Structural probing of a protein phosphatase 2A network by chemical cross-linking and mass spectrometry. *Science Signaling*, *337*(6100), 1348. AAAS.

Hirst, S. J., Alexander, N., Mchaourab, H. S. & Meiler, J. (2011). RosettaEPR: An integrated tool for protein structure determination from sparse EPR data. *Journal of structural biology*, *173*(3), 506–514. Elsevier.

Hoffmann, K., Sixel, U., Di Pasquale, F. & von Kügelgen, I. (2008). Involvement of basic amino acid residues in transmembrane regions 6 and 7 in agonist and antagonist recognition of the human platelet P2Y12-receptor. *Biochemical pharmacology*, *76*(10), 1201–1213. Elsevier.

Hogeweg, P. & Hesper, B. (1984). The alignment of sets of sequences and the construction of phyletic trees: an integrated method. *Journal of molecular evolution*, *20*(2), 175–186. Springer.

Hollmann, M. & Heinemann, S. (1994). Cloned glutamate receptors. *Annual review of neuroscience*, *17*(1), 31–108. Annual Reviews.

Hollopeter, G., Jantzen, H.M., Vincent, D., Li, G., England, L., Ramakrishnan, V., Yang, R.B., et al. (2001). Identification of the platelet ADP receptor targeted by antithrombotic drugs. *Nature*, *409*(6817), 202–207. Nature Publishing Group.

Holst, B., Nygaard, R., Valentin-Hansen, L., Bach, A., Engelstoft, M. S., Petersen, P. S., Frimurer, T. M., et al. (2010). A conserved aromatic lock for the tryptophan rotameric switch in TM-VI of seven-transmembrane receptors. *Journal of Biological Chemistry*, *285*(6), 3973–3985. ASBMB.

Honer, M., Stoffel, A., Kessler, L. J., Schubiger, P. & Ametamey, S. M. (2007). Radiolabeling and in vitro and in vivo evaluation of [18-F]-FE-DABP688 as a PET radioligand for the metabotropic glutamate receptor subtype 5. *Nuclear medicine and biology*, *34*(8), 973–980. Elsevier.

Hoyer, I., Haas, A., Kreuchwig, A., Schülein, R., Krause, G. & others. (2013). Molecular sampling of the allosteric binding pocket of the TSH receptor provides discriminative pharmacophores for antagonist and agonists. *Biochemical Society transactions*, *41*(1), 213–217.

Huang, B. & Schroeder, M. (2006). LIGSITEcsc: predicting ligand binding sites using the Connolly surface and degree of conservation. *BMC structural Biology*, *6*(1), 19. BioMed Central Ltd.

Hubbell, W. L., Altenbach, C., Hubbell, C. M. & Khorana, H. G. (2003). Rhodopsin structure, dynamics, and activation: a perspective from crystallography, site-directed spin labeling, sulfhydryl reactivity, and disulfide cross-linking. *Advances in protein chemistry*, *63*, 243–290. Elsevier.

Hughey, R. & Krogh, A. (1996). Hidden Markov models for sequence analysis: extension and analysis of the basic method. *Computer applications in the biosciences: CABIOS*, *12*(2), 95–107. Oxford Univ Press.

Humphrey, W., Dalke, A. & Schulten, K. (1996). VMD: visual molecular dynamics. *Journal of molecular graphics*, *14*(1), 33–38. Elsevier.

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, 90–95. IEEE Computer Society.

Ignatovica, V., Megnis, K., Lapins, M., Schiöth, H. B. & Klovins, J. (2012). Identification and analysis of functionally important amino acids in human purinergic 12 receptor using a Saccharomyces cerevisiae expression system. *FEBS Journal*, *279*(1), 180–191. Wiley Online Library.

Illergård, K., Callegari, S. & Elofsson, A. (2010). MPRAP: An accessibility predictor for a-helical transmem-brane proteins that performs well inside and outside the membrane. *BMC bioinformatics*, *11*(1), 333. BioMed Central Ltd.

Jaakola, V.P., Griffith, M. T., Hanson, M. A., Cherezov, V., Chien, E. Y., Lane, J. R., Ijzerman, A. P., et al. (2008). The 2.6 angstrom crystal structure of a human A2A adenosine receptor bound to an antagonist. *Science Signaling*, *322*(5905), 1211. AAAS.

Jacobson, K. A., Jayasekara, M. & Costanzi, S. (2012). Molecular structure of P2Y receptors: mutagenesis, modeling, and chemical probes. *Wiley Interdisciplinary Reviews: Membrane Transport and Signaling*, *1*(6), 815–827. Wiley Online Library.

Jacoby, E., Bouhelal, R., Gerspacher, M. & Seuwen, K. (2006). The 7 TM G-Protein-Coupled Receptor Target Family. *ChemMedChem*, *1*(8), 760–782. Wiley Online Library.

Jaeschke, G., Wettstein, J. G., Nordquist, R. E. & Spooren, W. (2008). mGlu5 receptor antagonists and their therapeutic potential. Informa.

Javitch, J. A., Shi, L. & Liapakis, G. (2002). Use of the substituted cysteine accessibility method to study the structure and function of G protein-coupled receptors. *Methods in enzymology*, *343*, 137–156. Elsevier.

Jones, D. T. & others. (1999). GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences. *Journal of molecular biology*, *287*(4), 797–815. Elsevier Science.

Jones, D. T., Taylort, W. & Thornton, J. M. (1992). A new approach to protein fold recognition. Nature Publishing Group.

Joo, K., Lee, S. J. & Lee, J. (2012). Sann: Solvent accessibility prediction of proteins by nearest neighbor method. *Proteins: Structure, Function, and Bioinformatics*, *80*(7), 1791–1797. Wiley Online Library.

Judith, K.S. (2012). Role of Conserved Transmembrane Domain Cysteines in Activation of Metabotropic Glutamate Receptor Subtype 6. *Medicinal chemistry*.

Kalidas, Y., Chandra, N. & others. (2008). PocketDepth: a new depth based algorithm for identification of ligand binding sites in proteins. *Journal of structural biology*, *161*(1), 31.

Kandil, S., Biondaro, S., Vlachakis, D., Cummins, A.C., Coluccia, A., Berry, C., Leyssen, P., et al. (2009). Discovery of a novel HCV helicase inhibitor by a de novo drug design approach. *Bioorganic & medicinal chemistry letters*, *19*(11), 2935–2937. Elsevier.

Karakas, M., Woetzel, N., Staritzbichler, R., Alexander, N., Weiner, B. E. & Meiler, J. (2012). BCL:: Fold-De Novo Prediction of Complex and Large Protein Topologies by Assembly of Secondary Structure Elements. *PloS one*, *7*(11), e49240. Public Library of Science.

Karchin, R., Cline, M. & Karplus, K. (2004). Evaluation of local structure alphabets based on residue burial. *Proteins: Structure, Function, and Bioinformatics*, *55*(3), 508–518. Wiley Online Library.

Karlin, A. & Akabas, M. H. (1998). Substituted-cysteine accessibility method. *Methods in enzymology*, *293*, 123.

Karplus, K., Barrett, C. & Hughey, R. (1998). Hidden Markov models for detecting remote protein homologies. *Bioinformatics*, *14*(10), 846–856. Oxford Univ Press.

Katritch, V. & Abagyan, R. (2011). GPCR agonist binding revealed by modeling and crystallography. *Trends in pharmacological sciences*. Elsevier.

Katritch, V., Cherezov, V. & Stevens, R. C. (2013). Structure-function of the G protein-coupled receptor superfamily. *Annual review of pharmacology and toxicology*, *53*, 531–556. Annual Reviews.

Katritch, V., Rueda, M. & Abagyan, R. (2012). Ligand-guided receptor optimization. *Homology Modeling* (pp. 189–205). Springer.

Kaufmann, K. W., Dawson, E. S., Henry, L. K., Field, J. R., Blakely, R. D. & Meiler, J. (2009). Structural determinants of species-selective substrate recognition in human and Drosophila serotonin transporters revealed through computational docking studies. *Proteins: Structure, Function, and Bioinformatics*, *74*(3), 630–642. Wiley Online Library.

Kaufmann, K. W., Lemmon, G. H., DeLuca, S. L., Sheehan, J. H. & Meiler, J. (2010). Practically useful: what the Rosetta protein modeling suite can do for you. *Biochemistry*, *49*(14), 2987–2998. ACS Publications.

Kaufmann, K. W. & Meiler, J. (2012). Using RosettaLigand for Small Molecule Docking into Comparative Models. *PloS one*, *7*(12), e50769. Public Library of Science.

Keeble, A. H., Joachimiak, L. A., Maté, M. J., Meenan, N., Kirkpatrick, N., Baker, D. & Kleanthous, C. (2008). Experimental and computational analyses of the energetic basis for dual recognition of immunity proteins by colicin endonucleases. *Journal of molecular biology*, *379*(4), 745–759. Elsevier.

Kelley, L. A. & Sternberg, M. J. (2009). Protein structure prediction on the Web: a case study using the Phyre server. *Nature protocols*, *4*(3), 363–371. Nature Publishing Group.

Kinney, G. G., O'Brien, J. A., Lemaire, W., Burno, M., Bickel, D. J., Clements, M. K., Chen, T.B., et al. (2005). A novel selective positive allosteric modulator of metabotropic glutamate receptor subtype 5 has in vivo activity and antipsychotic-like effects in rat behavioral models. *Journal of Pharmacology and Experimental Therapeutics*, *313*(1), 199–206. ASPET.

Klco, J. M., Wiegand, C. B., Narzinski, K. & Baranski, T. J. (2005). Essential role for the second extracellular loop in C5a receptor activation. *Nature structural & molecular biology*, *12*(4), 320–326. Nature Publishing Group.

Knoflach, F., Mutel, V., Jolidon, S., Kew, J. N., Malherbe, P., Vieira, E., Wichmann, J., et al. (2001). Positive allosteric modulators of metabotropic glutamate 1 receptor: characterization, mechanism of action, and binding site. *Proceedings of the National Academy of Sciences*, *98*(23), 13402–13407. National Acad Sciences.

Konagurthu, A. S., Whisstock, J. C., Stuckey, P. J. & Lesk, A. M. (2006). MUSTANG: a multiple structural alignment algorithm. *Proteins: Structure, Function, and Bioinformatics*, *64*(3), 559–574. Wiley Online Library.

Kostenis, E., Waelbroeck, M. & Milligan, G. (2005). Techniques: promiscuous Galpha proteins in basic research and drug discovery. *Trends in pharmacological sciences*, *26*(11), 595.

Kruse, A. C., Hu, J., Pan, A. C., Arlow, D. H., Rosenbaum, D. M., Rosemond, E., Green, H. F., et al. (2012). Structure and dynamics of the M3 muscarinic acetylcholine receptor. *Nature*, *482*(7386), 552–556. Nature Publishing Group.

Kufareva, I., Rueda, M., Katritch, V., Stevens, R. C. & Abagyan, R. (2011). Status of GPCR modeling and docking as reflected by community-wide GPCR Dock 2010 assessment. *Structure*, *19*(8), 1108–1126. Elsevier.

Kuhlman, B. & Baker, D. (2000). Native protein sequences are close to optimal for their structures. *Proceedings of the National Academy of Sciences*, *97*(19), 10383–10388. National Acad Sciences.

Kuhlman, B., Dantas, G., Ireton, G. C., Varani, G., Stoddard, B. L. & Baker, D. (2003). Design of a novel globular protein fold with atomic-level accuracy. *Science*, *302*(5649), 1364–1368. American Association for the Advancement of Science.

Kulkarni, S. S., Zou, M.F., Cao, J., Deschamps, J. R., Rodriguez, A. L., Conn, P. J. & Newman, A. H. (2009). Structure- Activity Relationships Comparing N-(6-Methylpyridin-yl)-Substituted Aryl Amides to 2-Methyl-6-(substituted-arylethynyl) pyridines or 2-Methyl-4-(substituted-arylethynyl) thiazoles as Novel Metabotropic Glutamate Receptor Subtype 5 Antagonists†. *Journal of medicinal chemistry*, *52*(11), 3563–3575. ACS Publications.

Kumar, A, Voet, A. & Zhang, K. (2012). Fragment Based Drug Design: From Experimental to Computational Approaches. *Current medicinal chemistry*, *19*(30), 5128–5147. Bentham Science Publishers.

Kumar, Ashutosh, & Zhang, K. Y. (2012). Computational fragment-based screening using RosettaLigand: the SAMPL3 challenge. *Journal of computer-aided molecular design*, *26*(5), 603–616. Springer.

Kunishima, N., Shimada, Y., Tsuji, Y., Sato, T., Yamamoto, M., Kumasaka, T., Nakanishi, S., et al. (2000). Structural basis of glutamate recognition by a dimeric metabotropic glutamate receptor. *Nature*, *407*(6807), 971–977. Nature Publishing Group.

Labute, P. (2010). LowModeMD- Implicit Low-Mode Velocity Filtering Applied to Conformational Search of Macrocycles and Protein Loops. *Journal of chemical information and modeling*, *50*(5), 792–800. ACS Publications.

Lagerström, M. C. & Schiöth, H. B. (2008). Structural diversity of G protein-coupled receptors and significance for drug discovery. *Nature reviews Drug discovery*, *7*(4), 339–357. Nature Publishing Group.

Lalueza-Fox, C., Römpler, H., Caramelli, D., Stäubert, C., Catalano, G., Hughes, D., Rohland, N., et al. (2007). A melanocortin 1 receptor allele suggests varying pigmentation among Neanderthals. *Science*, *318*(5855), 1453–1455. American

Association for the Advancement of Science.

Lamb, J. P., Engers, D. W., Niswender, C. M., Rodriguez, A. L., Venable, D. F., Conn, P. J. & Lindsley, C. W. (2011). Discovery of molecular switches within the ADX-47273 mGlu5 PAM scaffold that modulate modes of pharmacology to afford potent mGlu5 NAMs, PAMs and partial antagonists. *Bioorganic & medicinal chemistry letters*, *21*(9), 2711–2714. Elsevier.

Lange, O. F. & Baker, D. (2012). Resolution-adapted recombination of structural features significantly improves sampling in restraint-guided structure calculation. *Proteins: Structure, Function, and Bioinformatics*, *80*(3), 884–895. Wiley Online Library.

Lange, O. F., Rossi, P., Sgourakis, N. G., Song, Y., Lee, H.W., Aramini, J. M., Ertekin, A., et al. (2012). Determination of solution structures of proteins up to 40 kDa using CS-Rosetta with sparse NMR data from deuterated samples. *Proceedings of the National Academy of Sciences*, *109*(27), 10873–10878. National Acad Sciences.

Larkin, M., Blackshields, G., Brown, N., Chenna, R., McGettigan, P., McWilliam, H., Valentin, F., et al. (2007). Clustal W and Clustal X version 2.0. *Bioinformatics*, *23*(21), 2947–2948. Oxford Univ Press.

Lars, R., Eva, E., Claudia, S., Ines, L., Philipp, S., Thomas, H., Holger, R., et al. (2012). The ligand specificity of the G-protein-coupled receptor GPR34. *Biochemical Journal*, *443*(3), 841–850. Portland Press Ltd.

Laskowski, R. A. (1995). SURFNET: a program for visualizing molecular surfaces, cavities, and intermolecular interactions. *Journal of molecular graphics*, *13*(5), 323–330. Elsevier.

Lazareno, S. & Birdsall, N. (1995). Detection, quantitation, and verification of allosteric interactions of agents with labeled and unlabeled ligands at G protein-coupled receptors: interactions of strychnine and acetylcholine at muscarinic receptors. *Molecular pharmacology*, *48*(2), 362–378. ASPET.

Lazaridis, T. & Karplus, M. (1999). Effective energy function for proteins in solution. *Proteins: Structure, Function, and Bioinformatics*, *35*(2), 133–152. Wiley Online Library.

Lazarowski, E. R., Homolya, L., Boucher, R. C. & Harden, T. K. (1997). Direct demonstration of mechanically induced release of cellular UTP and its implication for uridine nucleotide receptor activation. *Journal of Biological Chemistry*, *272*(39), 24348–24354. ASBMB.

Leach, A. R., Shoichet, B. K. & Peishoff, C. E. (2006). Docking and scoring. *J Med Chem*, *49*, 5851–5855.

Leach, K., Sexton, P. M. & Christopoulos, A. (2007). Allosteric GPCR modulators: taking advantage of permissive receptor pharmacology. *Trends in pharmacological sciences*, *28*(8), 382–389. Elsevier.

Leaver-Fay, A., Kuhlman, B. & Snoeyink, J. (2005). Rotamer-pair energy calculations using a trie data structure. *Algorithms in bioinformatics* (pp. 389–400). Springer.

Leaver-Fay, A., Tyka, M., Lewis, S. M., Lange, O. F., Thompson, J., Jacak, R., Kaufman, K., et al. (2011). ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. *Methods Enzymol*, *487*, 545–574.

Lees-Miller, J. P., Subbotina, J. O., Guo, J., Yarov-Yarovoy, V., Noskov, S. Y. & Duff, H. J. (2009). Interactions of H562 in the S5 helix with T618 and S621 in the pore helix are important determinants of hERG1 potassium channel structure and function. *Biophysical journal*, *96*(9), 3600–3610. Elsevier.

Leman, J. K., Mueller, R., Karakas, M., Woetzel, N. & Meiler, J. (2013). Simultaneous prediction of protein secondary structure and transmembrane spans. *Proteins*. doi:10.1002/prot.24258

Lemmon, G., Kaufmann, K. & Meiler, J. (2012). Prediction of HIV-1 Protease/Inhibitor Affinity using RosettaLigand. *Chemical Biology & Drug Design*, *79*(6), 888–896. Wiley Online Library.

Lemmon, G. & Meiler, J. (2012). Rosetta Ligand docking with flexible XML protocols. *Methods in molecular biology (Clifton, NJ)*, *819*, 143. Springer.

Lemmon, G. & Meiler, J. (2013). Towards Ligand Docking Including Explicit Interface Water Molecules. *PLOS ONE*, *8*(6), e67536. Public Library of Science.

Levinthal, C. (1969). How to fold graciously. *Mossbauer spectroscopy in biological systems*, 22–24. University of Illinois Press, Urbana, IL.

Li, B., Scarselli, M., Knudsen, C. D., Kim, S.K., Jacobson, K. A., McMillin, S. M. & Wess, J. (2007). Rapid identification of functionally critical amino acids in a G protein-coupled receptor. *Nature methods*, *4*(2), 169–174. Nature Publishing Group.

Li, J. H., Hamdan, F. F., Kim, S.K., Jacobson, K. A., Zhang, X., Han, S.J. & Wess, J. (2008). Ligand-Specific Changes in M3 Muscarinic Acetylcholine Receptor Structure Detected by a Disulfide Scanning Strategy. *Biochemistry*, *47*(9), 2776–2788. ACS Publications.

Li, Z. & Scheraga, H. A. (1987). Monte Carlo-minimization approach to the multiple-minima problem in protein folding. *Proceedings of the National Academy of Sciences*, *84*(19), 6611–6615. National Acad Sciences.

Liapakis, G., Simpson, M. M. & Javitch, J. A. (2001). The Substituted-Cysteine Accessibility Method (SCAM) to Elucidate Membrane Protein Structure. *Current Protocols in Neuroscience*, 4–15. Wiley Online Library.

Lindahl, E. & Elofsson, A. (2000). Identification of related proteins on family, superfamily and fold level. *Journal of molecular biology*, *295*(3), 613–625. Elsevier.

Lindberg, J. S., Culleton, B., Wong, G., Borah, M. F., Clark, R. V., Shapiro, W. B., Roger, S. D., et al. (2005). Cinacalcet HCl, an oral calcimimetic agent for the treatment of secondary hyperparathyroidism in hemodialysis and peritoneal dialysis: a randomized, double-blind, multicenter study. *Journal of the American Society of Nephrology*, *16*(3), 800–807. Am Soc Nephrol.

Lindsley, C. W., Wisnoski, D. D., Leister, W. H., O'Brien, J. A., Lemaire, W., Williams, D. L., Burno, M., et al. (2004). Discovery of positive allosteric modulators for the metabotropic glutamate receptor subtype 5 from a series of N-(1, 3-diphenyl-1 H-pyrazol-5-yl) benzamides that potentiate receptor function in vivo. *Journal of medicinal chemistry*, *47*(24), 5825–5828. ACS Publications.

Lindsley, C. W., Zhao, Z., Leister, W. H., O'Brien, J., Lemaire, W., Williams Jr, D. L., Chen, T.B., et al. (2006). Design, Synthesis, and In Vivo Efficacy of Glycine Transporter-1 (GlyT1) Inhibitors Derived from a Series of [4-Phenyl-1-(propylsulfonyl) piperidin-4-yl] methyl Benzamides. *ChemMedChem*, *1*(8), 807–811. Wiley Online Library.

Litschig, S., Gasparini, F., Rueegg, D., Stoehr, N., Flor, P. J., Vranesic, I., Prézeau, L., et al. (1999). CPCCOEt, a noncompetitive metabotropic glutamate receptor 1 antagonist, inhibits receptor signaling without affecting glutamate binding. *Molecular pharmacology*, *55*(3), 453–461. ASPET.

Liu, F., Grauer, S., Kelley, C., Navarra, R., Graf, R., Zhang, G., Atkinson, P. J., et al. (2008). ADX47273 [S-(4-fluoro-phenyl)-3-[3-(4-fluoro-phenyl)-[1, 2, 4]-oxadiazol-5-yl]-piperidin-1-yl-methanone]: a novel metabotropic glutamate receptor 5-selective positive allosteric modulator with preclinical antipsychotic-like and procognitive activities. *Journal of Pharmacology and Experimental Therapeutics*, *327*(3), 827–839. ASPET.

Liu, X., He, Q., Studholme, D. J., Wu, Q., Liang, S. & Yu, L. (2004). NCD3G: a novel nine-cysteine domain in family 3 GPCRs. *Trends in biochemical sciences*, *29*(9), 458–461. Elsevier.

London, N. & Schueler-Furman, O. (2007). Assessing the energy landscape of CAPRI targets by FunHunt. *Proteins: Structure, Function, and Bioinformatics*, *69*(4), 809–815. Wiley Online Library.

Loo, T. W. & Clarke, D. M. (1997). Identification of residues in the drug-binding site of human P-glycoprotein using a thiol-reactive substrate. *Journal of Biological Chemistry*, *272*(51), 31945–31948. ASBMB.

Lu, I.L., Huang, C.F., Peng, Y.H., Lin, Y.T., Hsieh, H.P., Chen, C.T., Lien, T.W., et al. (2006). Structure-based drug design of a novel family of PPAR-gamma partial agonists: virtual screening, X-ray crystallography, and in vitro/in vivo biological activities. *Journal of Medicinal Chemistry*, *49*(9), 2703–2712. ACS Publications.

Lundström, J., Rychlewski, L., Bujnicki, J. & Elofsson, A. (2001). Pcons: A neural-network-based consensus predictor that improves fold recognition. *Protein Science*, *10*(11), 2354–2362. Wiley Online Library.

Maj, M., Bruno, V., Dragic, Z., Yamamoto, R., Battaglia, G., Inderbitzin, W., Stoehr, N., et al. (2003). (-)-PHCCC, a positive allosteric modulator of mGluR4: characterization, mechanism of action, and neuroprotection. *Neuropharmacology*, *45*(7), 895–906. Elsevier.

Malherbe, P., Kratochwil, N., Knoflach, F., Zenner, M.T., Kew, J. N., Kratzeisen, C., Maerki, H. P., et al. (2003). Mutational analysis and molecular modeling of the allosteric binding site of a novel, selective, noncompetitive antagonist of the metabotropic glutamate 1 receptor. *Journal of Biological Chemistry*, *278*(10), 8340–8347. ASBMB.

Malherbe, P., Kratochwil, N., Mühlemann, A., Zenner, M.T., Fischer, C., Stahl, M., Gerber, P. R., et al. (2006). Comparison of the binding pockets of two chemically unrelated allosteric antagonists of the mGlu5 receptor and identification of crucial residues involved in the inverse agonism of MPEP. *Journal of neurochemistry*, *98*(2), 601–615. Wiley Online Library.

Malherbe, P., Kratochwil, N., Zenner, M.T., Piussi, J., Diener, C., Kratzeisen, C., Fischer, C., et al. (2003). Mutational analysis and molecular modeling of the binding pocket of the metabotropic glutamate 5 receptor negative modulator 2-methyl-6-(phenylethynyl)-pyridine. *Molecular pharmacology*, *64*(4), 823–832. ASPET.

Mandell, D. J., Coutsias, E. A. & Kortemme, T. (2009). Sub-angstrom accuracy in protein loop reconstruction by robotics-inspired conformational sampling. *Nature methods*, *6*(8), 551–552. Nature Publishing Group.

Manglik, A., Kruse, A. C., Kobilka, T. S., Thian, F. S., Mathiesen, J. M., Sunahara, R. K., Pardo, L., et al. (2012). Crystal structure of the [micro]-opioid receptor bound to a morphinan antagonist. *Nature*, *485*(7398), 321–326. Nature Publishing Group.

Manka, J. T., Vinson, P. N., Gregory, K. J., Zhou, Y., Williams, R., Gogi, K., Days, E., et al. (2012). Optimization of an ether series of mGlu5 positive allosteric modulators: molecular determinants of MPEP-site interaction crossover. *Bioorganic & medicinal chemistry letters*. Elsevier.

Marino, M. J., Williams Jr, D. L., O'Brien, J. A., Valenti, O., McDonald, T. P., Clements, M. K., Wang, R., et al. (2003). Allosteric modulation of group III metabotropic glutamate receptor 4: a potential approach to Parkinson's disease treatment. *Proceedings of the National Academy of Sciences*, *100*(23), 13668–13673. National Acad Sciences.

McBride, S. M., Choi, C. H., Wang, Y., Liebelt, D., Braunstein, E., Ferreiro, D., Sehgal, A., et al. (2005). Pharmacological Rescue of Synaptic Plasticity, Courtship Behavior, and Mushroom Body Defects in a Drosophila Model of Fragile X Syndrome. *Neuron*, *45*(5), 753–764. Elsevier.

McGuffin, L. J., Bryson, K. & Jones, D. T. (2000). The PSIPRED protein structure prediction server. *Bioinformatics*, *16*(4), 404–405. Oxford Univ Press.

McPherson, A. (1991). Review Current approaches to macromolecular crystallization. *EJB Reviews 1990* (pp. 49–71). Springer.

Meiler, J. & Baker, D. (2003). Coupled prediction of protein secondary and tertiary structure. *Proceedings of the National Academy of Sciences*, *100*(21), 12105–12110. National Acad Sciences.

Meiler, J. & Baker, D. (2006). ROSETTALIGAND: Protein-small molecule docking with full side-chain flexibility. *PROTEINS: Structure, Function, and Bioinformatics*, *65*(3), 538–548. Wiley Online Library.

Meiler, J., Baker, D. & others. (2005). The fumarate sensor DcuS: progress in rapid protein fold elucidation by combining protein structure prediction methods with NMR spectroscopy. *Journal of Magnetic Resonance*, *173*(2), 310–316. San Diego: Academic Press, 1993-1996.

Melancon, B. J., Hopkins, C. R., Wood, M. R., Emmitte, K. A., Niswender, C. M., Christopoulos, A., Conn, P. J., et al. (2012). Allosteric modulation of 7 transmembrane spanning receptors: theory, practice and opportunities for CNS drug discovery. *Journal of medicinal chemistry*, *55*(4), 1445. NIH Public Access.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. & Teller, E. (1953). Equation of state calculations by fast computing machines. *The journal of chemical physics*, *21*, 1087.

Mewes, H.W., Frishman, D., Güldener, U., Mannhaupt, G., Mayer, K., Mokrejs, M., Morgenstern, B., et al. (2002). MIPS: a database for genomes and protein sequences. *Nucleic acids research*, *30*(1), 31–34. Oxford Univ Press.

Michino, M., Abola, E., Dock, G. & others. (2009). Community-wide assessment of GPCR structure modelling and ligand docking: GPCR Dock 2008. *Nature Reviews Drug Discovery*, *8*(6), 455–463. Nature Publishing Group.

Miedlich, S. U., Gama, L., Seuwen, K., Wolf, R. M. & Breitwieser, G. E. (2004). Homology modeling of the transmembrane domain of the human calcium sensing receptor and localization of an allosteric binding site. *Journal of Biological Chemistry*, *279*(8), 7254–7263. ASBMB.

Milligan, G. & Smith, N. J. (2007). Allosteric modulation of heterodimeric G-protein-coupled receptors. *Trends in pharmacological sciences*, *28*(12), 615–620. Elsevier.

Misura, K. M., Chivian, D., Rohl, C. A., Kim, D. E. & Baker, D. (2006). Physically realistic homology models built with ROSETTA can be more accurate than their templates. *Proceedings of the National Academy of Sciences*, *103*(14), 5361–5366. National Acad Sciences.

Mitsukawa, K., Yamamoto, R., Ofner, S., Nozulak, J., Pescott, O., Lukic, S., Stoehr, N., et al. (2005). A selective metabotropic glutamate receptor 7 agonist: activation of receptor signaling via an allosteric site modulates stress parameters in vivo.

Proceedings of the National Academy of Sciences of the United States of America, *102*(51), 18712–18717. National Acad Sciences.

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

Mobley, D. L., Graves, A. P., Chodera, J. D., McReynolds, A. C., Shoichet, B. K. & Dill, K. A. (2007). Predicting absolute ligand binding free energies to a simple model site. *Journal of molecular biology*, *371*(4), 1118. NIH Public Access.

Molck, C., Harpsoe, K., Gloriam, D. E., Clausen, R. P., Madsen, U., Pedersen, L. O., Jimenez, H. N., et al. (2012). Pharmacological Characterization and Modeling of the Binding Sites of Novel 1, 3-bis (pyridinylethynyl) benzenes as Metabotropic Glutamate Receptor 5-selective Negative Allosteric Modulators. *Molecular Pharmacology*. ASPET.

Mooers, B. H. & Matthews, B. W. (2006). Extension to 2268 atoms of direct methods in the ab initio determination of the unknown structure of bacteriophage P22 lysozyme. *Acta Crystallographica Section D: Biological Crystallography*, *62*(2), 165–176. International Union of Crystallography.

Morris, G. M., Huey, R., Lindstrom, W., Sanner, M. F., Belew, R. K., Goodsell, D. S. & Olson, A. J. (2009). AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. *Journal of computational chemistry*, *30*(16), 2785–2791. Wiley Online Library.

Mortier, J., Rakers, C., Frederick, R. & Wolber, G. (2012). Computational Tools for In Silico Fragment-Based Drug Design. *Current topics in medicinal chemistry*, *12*(17), 1935–1943. Bentham Science Publishers.

Moult, J. (2006). Rigorous performance evaluation in protein structure modelling and implications for computational biology. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *361*(1467), 453–458. The Royal Society.

Mueller, R., Dawson, E. S., Meiler, J., Rodriguez, A. L., Chauder, B. A., Bates, B. S., Felts, A. S., et al. (2012). Discovery of 2-(2-Benzoxazoyl amino)-4-Aryl-5-Cyanopyrimidine as Negative Allosteric Modulators (NAMs) of Metabotropic Glutamate Receptor 5 (mGlu5): From an Artificial Neural Network Virtual Screen to an In Vivo Tool Compound. *ChemMedChem*, *7*(3), 406–414. Wiley Online Library.

Mühlemann, A., Ward, N. A., Kratochwil, N., Diener, C., Fischer, C., Stucki, A., Jaeschke, G., et al. (2006). Determination of key amino acids implicated in the actions of allosteric modulation by 3, 3′-difluorobenzaldazine on rat mGlu5 receptors. *European journal of pharmacology*, *529*(1), 95–104. Elsevier.

Muñoz-Esparza, C., Jover, E., Hernández-Romero, D., Saura, D., Valdés, M., YH Lip, G. & Marin, F. (2011). Interactions between clopidogrel and proton pump inhibitors: a review of evidence. *Current medicinal chemistry*, *18*(16), 2386–2400. Bentham Science Publishers.

Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. *Journal of molecular biology*, *247*(4), 536–540. Elsevier.

Muto, T., Tsuchiya, D., Morikawa, K. & Jingami, H. (2007). Expression, purification, crystallization and preliminary X-ray analysis of the ligand-binding domain of metabotropic glutamate receptor 7. *Acta Crystallographica Section F: Structural Biology and Crystallization Communications*, *63*(7), 627–630. International Union of Crystallography.

Nannemann, D. P., Kaufmann, K. W., Meiler, J. & Bachmann, B. O. (2010). Design and directed evolution of a dideoxy purine nucleoside phosphorylase. *Protein Engineering Design and Selection*, *23*(8), 607–616. Oxford Univ Press.

Nawarskas, J. J. & Clark, S. M. (2011). Ticagrelor: a novel reversible oral antiplatelet agent. *Cardiology in review*, *19*(2), 95–100. LWW.

Needleman, S. B. & Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of molecular biology*, *48*(3), 443–453. Elsevier.

Neria, E., Fischer, S. & Karplus, M. (1996). Simulation of activation free energies in molecular systems. *The Journal of chemical physics*, *105*, 1902.

Nicklaus, M. C., Wang, S., Driscoll, J. S. & Milne, G. W. (1995). Conformational changes of small molecules binding to proteins. *Bioorganic & medicinal chemistry*, *3*(4), 411–428. Elsevier.

Nikiforovich, G. V., Taylor, C. M., Marshall, G. R. & Baranski, T. J. (2010). Modeling the possible conformations of the extracellular loops in G-protein-coupled receptors. *Proteins: Structure, Function, and Bioinformatics*, *78*(2), 271–285. Wiley Online Library.

Niswender, C. M. & Conn, P. J. (2010). Metabotropic glutamate receptors: physiology, pharmacology, and disease. *Annual review of pharmacology and toxicology*, *50*, 295. NIH Public Access.

Noetzel, M. J., Rook, J. M., Vinson, P. N., Cho, H. P., Days, E., Zhou, Y., Rodriguez, A. L., et al. (2012). Functional impact of allosteric agonist activity of selective positive allosteric modulators of metabotropic glutamate receptor subtype 5 in regulating central nervous system function. *Molecular pharmacology*, *81*(2), 120–133. ASPET.

Notredame, C., Higgins, D. G., Heringa, J. & others. (2000). T-Coffee: A novel method for fast and accurate multiple sequence alignment. *Journal of molecular biology*, *302*(1), 205–218. London, New York, Academic Press.

Nygaard, R., Valentin-Hansen, L., Mokrosinski, J., Frimurer, T. M. & Schwartz, T. W. (2010). Conserved water-mediated hydrogen bond network between TM-I,-II,-VI, and-VII in 7TM receptor activation. *Journal of Biological Chemistry*, *285*(25), 19625–

19636. ASBMB.

Nygaard, R., Zou, Y., Dror, R. O., Mildorf, T. J., Arlow, D. H., Manglik, A., Pan, A. C., et al. (2013). The Dynamic Process of beta-2 Adrenergic Receptor Activation. *Cell*, *152*(3), 532–542. Elsevier.

O'Brien, J. A., Lemaire, W., Chen, T.B., Chang, R. S., Jacobson, M. A., Ha, S. N., Lindsley, C. W., et al. (2003). A family of highly selective allosteric modulators of the metabotropic glutamate receptor subtype 5. *Molecular pharmacology*, *64*(3), 731–740. ASPET.

O'Brien, J. A., Lemaire, W., Wittmann, M., Jacobson, M. A., Ha, S. N., Wisnoski, D. D., Lindsley, C. W., et al. (2004). A novel selective allosteric modulator potentiates the activity of native metabotropic glutamate receptor subtype 5 in rat forebrain. *Journal of Pharmacology and Experimental Therapeutics*, *309*(2), 568–577. ASPET.

Okada, T., Sugihara, M., Bondar, A.N., Elstner, M., Entel, P. & Buss, V. (2004). The retinal conformation and its environment in rhodopsin in light of a new 2.2 Å crystal structure. *Journal of molecular biology*, *342*(2), 571–583. Elsevier.

Olden, J. D. & Jackson, D. A. (2002). Illuminating the "black box": a randomization approach for understanding variable contributions in artificial neural networks. *Ecological modelling*, *154*(1), 135–150. Elsevier.

Overington, J. P., Al-Lazikani, B. & Hopkins, A. L. (2006). How many drug targets are there? *Nature reviews Drug discovery*, *5*(12), 993–996. Nature Publishing Group.

Pagano, A., Rüegg, D., Litschig, S., Stoehr, N., Stierlin, C., Heinrich, M., Floersheim, P., et al. (2000). The non-competitive antagonists 2-methyl-6-(phenylethynyl) pyridine and 7-hydroxyiminocyclopropan [b] chromen-1a-carboxylic acid ethyl ester interact with overlapping binding pockets in the transmembrane region of group I metabotropic glutamate receptors. *Journal of Biological Chemistry*, *275*(43), 33750–33758. ASBMB.

Palczewski, K., Kumasaka, T., Hori, T., Behnke, C. A., Motoshima, H., Fox, B. A., Trong, I. L., et al. (2000). Crystal structure of rhodopsin: AG protein-coupled receptor. *Science Signalling*, *289*(5480), 739. AAAS.

Pandit, D., Tuske, S. J., Coales, S. J., Liu, A., Lee, J. E., Morrow, J. A., Nemeth, J. F., et al. (2012). Mapping of discontinuous conformational epitopes by amide hydrogen/deuterium exchange mass spectrometry and computational docking. *Journal of Molecular Recognition*, *25*(3), 114–124. Wiley Online Library.

Park, Y., Hayat, S. & Helms, V. (2007). Prediction of the burial status of transmembrane residues of helical membrane proteins. *BMC bioinformatics*, *8*(1), 302. BioMed Central Ltd.

Parr, C. E., Sullivan, D. M., Paradiso, A. M., Lazarowski, E. R., Burch, L. H., Olsen, J. C., Erb, L., et al. (1994). Cloning and expression of a human P2U nucleotide receptor, a target for cystic fibrosis pharmacotherapy. *Proceedings of the National Academy of Sciences*, *91*(8), 3275–3279. National Acad Sciences.

Parry, J. J., Chen, R., Andrews, R., Lears, K. A. & Rogers, B. E. (2012). Identification of Critical Residues Involved in Ligand Binding and G Protein Signaling in Human Somatostatin Receptor Subtype 2. *Endocrinology*, *153*(6), 2747–2755. Endocrine Soc.

de Paulis, T., Hemstapat, K., Chen, Y., Zhang, Y., Saleh, S., Alagille, D., Baldwin, R. M., et al. (2006). Substituent effects of N-(1, 3-diphenyl-1 H-pyrazol-5-yl) benzamides on positive allosteric modulation of the metabotropic glutamate-5 receptor in rat cortical astrocytes. *Journal of medicinal chemistry*, *49*(11), 3332–3344. ACS Publications.

Pausch, M. H., Lai, M., Tseng, E., Paulsen, J., Bates, B. & Kwak, S. (2004). Functional expression of human and mouse P2Y12 receptors in Saccharomyces cerevisiae. *Biochemical and biophysical research communications*, *324*(1), 171–177. Elsevier.

Perola, E. & Charifson, P. S. (2004). Conformational analysis of drug-like molecules bound to proteins: an extensive study of ligand reorganization upon binding. *Journal of medicinal chemistry*, *47*(10), 2499–2510. ACS Publications.

Perola, E., Walters, W. P. & Charifson, P. S. (2004). A detailed comparison of current docking and scoring methods on systems of pharmaceutical relevance. *Proteins: Structure, Function, and Bioinformatics*, *56*(2), 235–249. Wiley Online Library.

Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C. & Ferrin, T. E. (2004). UCSF Chimera—a visualization system for exploratory research and analysis. *Journal of computational chemistry*, *25*(13), 1605–1612. Wiley Online Library.

Phillips, A., Janies, D., Wheeler, W. & others. (2000). Multiple sequence alignment in phylogenetic analysis. *Molecular phylogenetics and evolution*, *16*(3), 317–330.

Pin, J.P., Galvez, T. & Prézeau, L. (2003). Evolution, structure, and activation mechanism of family 3/C G-protein-coupled receptors. *Pharmacology & therapeutics*, *98*(3), 325–354. Elsevier.

Porter, R. H., Jaeschke, G., Spooren, W., Ballard, T. M., Büttelmann, B., Kolczewski, S., Peters, J.U., et al. (2005). Fenobam: a clinically validated nonbenzodiazepine anxiolytic is a potent, selective, and noncompetitive mGlu5 receptor antagonist with inverse agonist activity. *Journal of Pharmacology and Experimental Therapeutics*, *315*(2), 711–721. ASPET.

Prioleau, C., Visiers, I., Ebersole, B. J., Weinstein, H. & Sealfon, S. C. (2002). Conserved Helix 7 Tyrosine Acts as a Multistate Conformational Switch in the 5HT2C Receptor. *Journal of Biological Chemistry*, *277*(39), 36577–36584. ASBMB.

Qian, B., Raman, S., Das, R., Bradley, P., McCoy, A. J., Read, R. J. & Baker, D. (2007). High-resolution structure prediction and the crystallographic phase problem. *Nature*, *450*(7167), 259–264. Nature Publishing Group.

Ramachandran, G., Ramakrishnan, C. & Sasisekharan, V. (1963). Stereochemistry of polypeptide chain configurations. *Journal of molecular biology*, *7*, 95.

Raman, S., Vernon, R., Thompson, J., Tyka, M., Sadreyev, R., Pei, J., Kim, D., et al. (2009). Structure prediction for CASP8 with all-atom refinement using Rosetta. *Proteins: Structure, Function, and Bioinformatics*, *77*(S9), 89–99. Wiley Online Library.

Rarey, M., Kramer, B., Lengauer, T., Klebe, G. & others. (1996). A fast flexible docking method using an incremental construction algorithm. *Journal of molecular biology*, *261*(3), 470–489. London, New York, Academic Press.

Rasmussen, S. G., Choi, H.J., Rosenbaum, D. M., Kobilka, T. S., Thian, F. S., Edwards, P. C., Burghammer, M., et al. (2007). Crystal structure of the human beta-2 adrenergic G-protein-coupled receptor. *Nature*, *450*(7168), 383–387. Nature Publishing Group.

Rathmann, D., Lindner, D., DeLuca, S. H., Kaufmann, K. W., Meiler, J. & Beck-Sickinger, A. G. (2012). Ligand-mimicking Receptor Variant Discloses Binding and Activation Mode of Prolactin-releasing Peptide. *Journal of Biological Chemistry*, *287*(38), 32181–32194. ASBMB.

Rathmann, D., Pedragosa-Badia, X. & Beck-Sickinger, A. G. (2013). In vitro modification of substituted cysteines as tool to study receptor functionality and structure-activity relationships. *Analytical biochemistry*. doi:10.1016/j.ab.2013.04.015

Remijn, J. A., IJsseldijk, M. J., Strunk, A. L., Abbes, A. P., Engel, H., Dikkeschei, B., Dompeling, E. C., et al. (2007). Novel molecular defect in the platelet ADP receptor P2Y12 of a patient with haemorrhagic diathesis. *Clinical Chemical Laboratory Medicine*, *45*(2), 187–189.

Riedmiller, M. & Braun, H. (1993). A direct adaptive method for faster backpropagation learning: The RPROP algorithm. *Neural Networks IEEE International Conference* pp. 586–591.

Rod, T. H., Radkiewicz, J. L. & Brooks III, C. L. (2003). Correlated motion and the effect of distal mutations in dihydrofolate reductase. *Proceedings of the National Academy of Sciences*, *100*(12), 6980–6985. National Acad Sciences.

Rodriguez, A. L., Grier, M. D., Jones, C. K., Herman, E. J., Kane, A. S., Smith, R. L., Williams, R., et al. (2010). Discovery of novel allosteric modulators of metabotropic glutamate receptor subtype 5 reveals chemical and functional diversity and in vivo activity in rat behavioral models of anxiolytic and antipsychotic activity. *Molecular pharmacology*, *78*(6), 1105–1123. ASPET.

Rodriguez, A. L., Nong, Y., Sekaran, N. K., Alagille, D., Tamagnan, G. D. & Conn, P. J. (2005). A close structural analog of 2-methyl-6-(phenylethynyl)-pyridine acts as a neutral allosteric site ligand on metabotropic glutamate receptor subtype 5 and blocks the effects of multiple allosteric modulators. *Molecular pharmacology*, *68*(6), 1793–1802. ASPET.

Rodriguez, A. L., Williams, R., Zhou, Y., Lindsley, S. R., Le, U., Grier, M. D., David Weaver, C., et al. (2009). Discovery and SAR of novel mGluR5 non-competitive antagonists not based on an MPEP chemotype. *Bioorganic & medicinal chemistry letters*, *19*(12), 3209–3213. Elsevier.

Rohl, C. A. (2005). Protein structure estimation from minimal restraints using Rosetta. *Methods in enzymology*, *394*, 244–260. Elsevier.

Rohl, C. A., Strauss, C. E., Chivian, D. & Baker, D. (2004). Modeling structurally variable regions in homologous proteins with rosetta. *Proteins: Structure, Function, and Bioinformatics*, *55*(3), 656–677. Wiley Online Library.

Rohl, C. A., Strauss, C. E., Misura, K. M., Baker, D. & others. (2004). Protein structure prediction using Rosetta. *Methods in enzymology*, *383*, 66.

Rondard, P., Liu, J., Huang, S., Malhaire, F., Vol, C., Pinault, A., Labesse, G., et al. (2006). Coupling of agonist binding to effector domain activation in metabotropic glutamate-like receptors. *Journal of Biological Chemistry*, *281*(34), 24653–24661. ASBMB.

Roppe, J. R., Wang, B., Huang, D., Tehrani, L., Kamenecka, T., Schweiger, E. J., Anderson, J. J., et al. (2004). 5-[(2-Methyl-1, 3-thiazol-4-yl) ethynyl]-2, 3′-bipyridine: a highly potent, orally active metabotropic glutamate subtype 5 (mGlu5) receptor antagonist with anxiolytic activity. *Bioorganic & medicinal chemistry letters*, *14*(15), 3993–3996. Elsevier.

Rost, B. & Sander, C. (1993). Improved prediction of protein secondary structure by use of sequence profiles and neural networks. *Proceedings of the National Academy of Sciences*, *90*(16), 7558–7562. National Acad Sciences.

Rueda, M., Bottegoni, G. & Abagyan, R. (2009). Recipes for the selection of experimental protein conformations for virtual screening. *Journal of chemical information and modeling*, *50*(1), 186–193. ACS Publications.

Rueda, M., Katritch, V., Raush, E. & Abagyan, R. (2010). SimiCon: a web tool for protein-ligand model comparison through calculation of equivalent atomic contacts. *Bioinformatics*, *26*(21), 2784–2785. Oxford Univ Press.

Rummel-Kluge, C., Komossa, K., Schwarz, S., Hunger, H., Schmid, F., Kissling, W., Davis, J. M., et al. (2012). Second-generation antipsychotic drugs and extrapyramidal side effects: a systematic review and meta-analysis of head-to-head comparisons. *Schizophrenia bulletin*, *38*(1), 167–177. MPRC.

Russ, W. P., Engelman, D. M. & others. (2000). The GxxxG motif: a framework for transmembrane helix-helix association. *Journal of molecular biology*, *296*(3), 911–919. Elsevier Science.

Rychlewski, L., Li, W., Jaroszewski, L. & Godzik, A. (2000). Comparison of sequence profiles. Strategies for structural predictions using sequence information. *Protein*

Science, *9*(2), 232–241. Wiley Online Library.

Salomon, Y., Londos, C. & Rodbell, M. (1974). A highly sensitive adenylate cyclase assay. *Analytical biochemistry*, *58*(2), 541–548. Elsevier.

Sanner, M. F., Olson, A. J. & Spehner, J.C. (1996). Reduced surface: an efficient way to compute molecular surfaces. *Biopolymers*, *38*(3), 305–320. Wiley Online Library.

Sauder, J. M., Arthur, J. W. & Dunbrack Jr, R. L. (2000). Large-scale comparison of protein sequence alignment algorithms with structure alignments. *Proteins: Structure, Function, and Bioinformatics*, *40*(1), 6–22. Wiley Online Library.

Savi, P, Labouret, C., Delesque, N., Guette, F., Lupker, J. & Herbert, J. (2001). P2Y12, a New Platelet ADP Receptor, Target of Clopidogrel. *Biochemical and biophysical research communications*, *283*(2), 379–383. Elsevier.

Savi, Pierre, Zachayus, J.L., Delesque-Touchard, N., Labouret, C., Hervé, C., Uzabiaga, M.F., Pereillo, J.M., et al. (2006). The active metabolite of Clopidogrel disrupts P2Y12 receptor oligomers and partitions them out of lipid rafts. *Proceedings of the National Academy of Sciences*, *103*(29), 11069–11074. National Acad Sciences.

Schlumberger, C., Pietraszek, M., Gravius, A., Klein, K.U., Greco, S., Morè, L., Danysz, W., et al. (2009). Comparison of the mGlu (5) receptor positive allosteric modulator ADX47273 and the mGlu (2/3) receptor agonist LY354740 in tests for antipsychotic-like activity. *European journal of pharmacology*, *623*(1-3), 73.

Schmidt, P., Ritscher, L., Dong, E. N., Hermsdorf, T., Cöster, M., Wittkopf, D., Meiler, J., et al. (2013). Identification of Determinants Required for Agonistic and Inverse Agonistic Ligand Properties at the ADP Receptor P2Y12. *Molecular Pharmacology*, *83*(1), 256–266. ASPET.

Schneider, G. & Fechner, U. (2005). Computer-based de novo design of drug-like molecules. *Nature Reviews Drug Discovery*, *4*(8), 649–663. Nature Publishing Group.

Schneider, G., Hartenfeller, M., Reutlinger, M., Tanrikulu, Y., Proschak, E. & Schneider, P. (2009). Voyages to the (un) known: adaptive design of bioactive compounds. *Trends in biotechnology*, *27*(1), 18–26. Elsevier.

Schoepp, D. D., Jane, D. E. & Monn, J. A. (1999). Pharmacological agents acting at subtypes of metabotropic glutamate receptors. *Neuropharmacology*, *38*(10), 1431–1476. Elsevier.

Schröder, R., Janssen, N., Schmidt, J., Kebig, A., Merten, N., Hennen, S., Müller, A., et al. (2010). Deconvolution of complex G protein-coupled receptor signaling in live cells using dynamic mass redistribution measurements. *Nature biotechnology*, *28*(9), 943–949. Nature Publishing Group.

Schulz, A. & Schöneberg, T. (2003). The structural evolution of a P2Y-like G-protein-coupled receptor. *Journal of Biological Chemistry*, *278*(37), 35531–35541. ASBMB.

Schwartz, T. W. & Rosenkilde, M. M. (1996). Is there a'lock'for all agonist'keys' in 7TM receptors? *Trends in pharmacological sciences*, *17*(6), 213.

Seifert, R. & Wenzel-Seifert, K. (2002). Constitutive activity of G-protein-coupled receptors: cause of disease and common property of wild-type receptors. *Naunyn-Schmiedeberg's archives of pharmacology*, *366*(5), 381–416. Springer.

Senes, A., Gerstein, M., Engelman, D. M. & others. (2000). Statistical Analysis of Amino Acid Patterns in Transmembrane Helices: The GxxxG Motif Occurs Frequently, and in Association with beta-branched Residues at Neighboring Positions. *Journal of molecular biology*, *296*(3), 921–936. Elsevier Science.

Sharma, S., Kedrowski, J., Rook, J. M., Smith, R. L., Jones, C. K., Rodriguez, A. L., Conn, P. J., et al. (2009). Discovery of molecular switches that modulate modes of metabotropic glutamate receptor subtype 5 (mGlu5) pharmacology in vitro and in vivo within a series of functionalized, regioisomeric 2-and 5-(phenylethynyl) pyrimidines. *Journal of medicinal chemistry*, *52*(14), 4103–4106. ACS Publications.

Sheffler, D. J., Wenthur, C. J., Bruner, J. A., Carrington, S. J., Vinson, P. N., Gogi, K. K., Blobaum, A. L., et al. (2012). Development of a novel, CNS-penetrant, metabotropic glutamate receptor 3 (mGlu3) NAM probe (ML289) derived from a closely related mGlu5 PAM. *Bioorganic & medicinal chemistry letters*. Elsevier.

Shen, Y., Bryan, P. N., He, Y., Orban, J., Baker, D. & Bax, A. (2010). De novo structure generation using chemical shifts for proteins with high-sequence identity but different folds. *Protein Science*, *19*(2), 349–356. Wiley Online Library.

Sherman, W., Day, T., Jacobson, M. P., Friesner, R. A. & Farid, R. (2006). Novel procedure for modeling ligand/receptor induced fit effects. *Journal of medicinal chemistry*, *49*(2), 534–553. ACS Publications.

Shi, L., Liapakis, G., Xu, R., Guarnieri, F., Ballesteros, J. A. & Javitch, J. A. (2002). beta-2 Adrenergic Receptor Activation. *Journal of Biological Chemistry*, *277*(43), 40989–40996. ASBMB.

Shimamura, T., Shiroishi, M., Weyand, S., Tsujimoto, H., Winter, G., Katritch, V., Abagyan, R., et al. (2011). Structure of the human histamine H1 receptor complex with doxepin. *Nature*, *475*(7354), 65–70. Nature Publishing Group.

Shindyalov, I. N. & Bourne, P. E. (1998). Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein engineering*, *11*(9), 739–747. Oxford Univ Press.

Shiraga, M., Miyata, S., Kato, H., Kashiwagi, H., Honda, S., Kurata, Y., Tomiyama, Y., et al. (2005). Impaired platelet function in a patient with P2Y12 deficiency caused by a mutation in the translation initiation codon. *Journal of Thrombosis and Haemostasis*, *3*(10), 2315–2323. Wiley Online Library.

Shoichet, B. K. & Kobilka, B. K. (2012). Structure-based drug screening for G-protein-coupled receptors. *Trends in pharmacological sciences*. Elsevier.

Siegel, J. B., Zanghellini, A., Lovick, H. M., Kiss, G., Lambert, A. R., Clair, J. L. S., Gallaher, J. L., et al. (2010). Computational design of an enzyme catalyst for a stereoselective bimolecular Diels-Alder reaction. *Science*, *329*(5989), 309–313. American Association for the Advancement of Science.

Sim, J., Kim, S.Y. & Lee, J. (2005). Prediction of protein solvent accessibility using fuzzy k-nearest neighbor method. *Bioinformatics*, *21*(12), 2844–2849. Oxford Univ Press.

Simon, J., Filippov, A. K., Göransson, S., Wong, Y. H., Frelin, C., Michel, A. D., Brown, D. A., et al. (2002). Characterization and Channel Coupling of the P2Y12 Nucleotide Receptor of Brain Capillary Endothelial Cells. *Journal of Biological Chemistry*, *277*(35), 31390–31400. ASBMB.

Simons, K. T., Kooperberg, C., Huang, E. & Baker, D. (1997). Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. *Journal of molecular biology*, *268*(1), 209–225. Elsevier Science.

Simons, K. T., Ruczinski, I., Kooperberg, C., Fox, B. A., Bystroff, C. & Baker, D. (1999). Improved recognition of native-like protein structures using a combination of sequence-dependent and sequence-independent features of proteins. *Proteins: Structure, Function, and Bioinformatics*, *34*(1), 82–95. Wiley Online Library.

Sliwoski, G., Lowe, E. W., Butkiewicz, M. & Meiler, J. (2012). BCL:: EMAS—Enantioselective Molecular Asymmetry Descriptor for 3D-QSAR. *Molecules*, *17*(8), 9971–9989. Molecular Diversity Preservation International.

Smith, J. A., Vanoye, C. G., George, A. L., Meiler, J. & Sanders, C. R. (2007). Structural models for the KCNQ1 voltage-gated potassium channel. *Biochemistry*, *46*(49), 14141–14152. ACS Publications.

Spooren, W. P., Vassout, A., Neijt, H. C., Kuhn, R., Gasparini, F., Roux, S., Porsolt, R. D., et al. (2000). Anxiolytic-like effects of the prototypical metabotropic glutamate receptor 5 antagonist 2-methyl-6-(phenylethynyl) pyridine in rodents. *Journal of Pharmacology and Experimental Therapeutics*, *295*(3), 1267–1275. ASPET.

Springthorpe, B., Bailey, A., Barton, P., Birkinshaw, T. N., Bonnert, R. V., Brown, R. C., Chapman, D., et al. (2007). From ATP to AZD6140: The discovery of an orally active reversible P2Y12 receptor antagonist for the prevention of thrombosis. *Bioorganic & medicinal chemistry letters*, *17*(21), 6013–6018. Elsevier.

Stevens, R. C., Cherezov, V., Katritch, V., Abagyan, R., Kuhn, P., Rosen, H. & Wüthrich, K. (2012). The GPCR Network: a large-scale collaboration to determine human GPCR structure and function. *Nature Reviews Drug Discovery*, *12*(1), 25–34. Nature Publishing Group.

Storey, R. F. (2011). Pharmacology and clinical trials of reversibly-binding P2Y12 inhibitors. *Thrombosis and haemostasis*, *105*(1), S75.

Stouten, P. F., Frömmel, C., Nakamura, H. & Sander, C. (1993). An effective solvation term based on atomic occupancies for use in protein simulations. *Molecular Simulation*, *10*(2-6), 97–120. Taylor & Francis.

Suzuki, G., Kimura, T., Satow, A., Kaneko, N., Fukuda, J., Hikichi, H., Sakai, N., et al. (2007). Pharmacological characterization of a new, orally active and potent allosteric metabotropic glutamate receptor 1 antagonist, 4-[1-(2-fluoropyridin-3-yl)-5-methyl-1H-1, 2, 3-triazol-4-yl]-N-isopropyl-N-methyl-3, 6-dihydropyridine-1 (2H)-carboxamide (FTIDC). *Journal of Pharmacology and Experimental Therapeutics*, *321*(3), 1144–1153. ASPET.

Szymczyna, B. R., Taurog, R. E., Young, M. J., Snyder, J. C., Johnson, J. E. & Williamson, J. R. (2009). Synergy of NMR, computation, and X-ray crystallography for structural biology. *Structure*, *17*(4), 499–507. Elsevier.

Söding, J., Biegert, A. & Lupas, A. N. (2005). The HHpred interactive server for protein homology detection and structure prediction. *Nucleic acids research*, *33*(suppl 2), W244–W248. Oxford Univ Press.

Takahashi, K., Tsuchida, K., Tanabe, Y., Masu, M. & Nakanishi, S. (1993). Role of the large extracellular domain of metabotropic glutamate receptors in agonist selectivity determination. *Journal of Biological Chemistry*, *268*(26), 19341–19345. ASBMB.

Tang, H., Wang, X. S., Hsieh, J.H. & Tropsha, A. (2012). Do crystal structures obviate the need for theoretical models of GPCRs for structure-based virtual screening? *Proteins: Structure, Function, and Bioinformatics*. Wiley Online Library.

Thompson, A. A., Liu, W., Chun, E., Katritch, V., Wu, H., Vardy, E., Huang, X.P., et al. (2012). Structure of the nociceptin/orphanin FQ receptor in complex with a peptide mimetic. *Nature*, *485*(7398), 395–399. Nature Publishing Group.

Thompson, J. & Baker, D. (2011). Incorporation of evolutionary information into Rosetta comparative modeling. *Proteins: Structure, Function, and Bioinformatics*, *79*(8), 2380–2388. Wiley Online Library.

Thompson, J. D., Higgins, D. G. & Gibson, T. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic acids research*, *22*(22), 4673–4680. Oxford Univ Press.

Thompson, J. D., Plewniak, F. & Poch, O. (1999). A comprehensive comparison of multiple sequence alignment programs. *Nucleic acids research*, *27*(13), 2682–2690. Oxford Univ Press.

Thor, D., Le Duc, D., Strotmann, R. & Schöneberg, T. (2009). Luciferase activity under direct ligand-dependent control of a muscarinic acetylcholine receptor. *BMC biotechnology*, *9*(1), 46. BioMed Central Ltd.

Thor, D., Schulz, A., Hermsdorf, T. & Schoneberg, T. (2008). Generation of an agonistic binding site for blockers of the M3 muscarinic acetylcholine receptor. *Biochem. J*, *412*, 103–112.

Treyer, V., Streffer, J., Wyss, M. T., Bettio, A., Ametamey, S. M., Fischer, U., Schmidt, M., et al. (2007). Evaluation of the metabotropic glutamate receptor subtype 5 using PET and 11C-ABP688: assessment of methods. *Journal of Nuclear Medicine*, *48*(7), 1207–1215. Soc Nuclear Med.

Tsai, G. & Coyle, J. T. (2002). Glutamatergic mechanisms in schizophrenia. *Annual Review of Pharmacology and Toxicology*, *42*(1), 165–179. Annual Reviews.

Tschoner, A., Engl, J., Laimer, M., Kaser, S., Rettenbacher, M., Fleischhacker, W., Patsch, J., et al. (2007). Metabolic side effects of antipsychotic medication. *International journal of clinical practice*, *61*(8), 1356–1370. Wiley Online Library.

Tsuchiya, D., Kunishima, N., Kamiya, N., Jingami, H. & Morikawa, K. (2002). Structural views of the ligand-binding cores of a metabotropic glutamate receptor complexed with an antagonist and both glutamate and Gd3+. *Proceedings of the National Academy of Sciences*, *99*(5), 2660–2665.

Uslaner, J. M., Parmentier-Batteur, S., Flick, R. B., Surles, N. O., Lam, J. S., McNaughton, C. H., Jacobson, M. A., et al. (2009). Dose-dependent effect of CDPPB, the mGluR5 positive allosteric modulator, on recognition memory is associated with GluR1 and CREB phosphorylation in the prefrontal cortex and hippocampus. *Neuropharmacology*, *57*(5), 531–538. Elsevier.

Varney, M. A., Cosford, N. D., Jachec, C., Rao, S. P., Sacaan, A., Lin, F.F., Bleicher, L., et al. (1999). SIB-1757 and SIB-1893: selective, noncompetitive antagonists of metabotropic glutamate receptor type 5. *Journal of Pharmacology and Experimental Therapeutics*, *290*(1), 170–181. ASPET.

Vasiljev, K. S., Uri, A. & Laitinen, J. T. (2003). 2-Alkylthio-substituted platelet P2Y12 receptor antagonists reveal pharmacological identity between the rat brain Gi-linked ADP receptors and P2Y12. *Neuropharmacology*, *45*(1), 145–154. Elsevier.

Verdonk, M. L., Cole, J. C., Hartshorn, M. J., Murray, C. W. & Taylor, R. D. (2003). Improved protein-ligand docking using GOLD. *Proteins: Structure, Function, and Bioinformatics*, *52*(4), 609–623. Wiley Online Library.

Visiers, I., Ballesteros, J. A. & Weinstein, H. (2002). Three-dimensional representations of G protein-coupled receptor structures and mechanisms. *Methods in enzymology*, *343*, 329–371. Elsevier.

Wacker, D., Wang, C., Katritch, V., Han, G. W., Huang, X.P., Vardy, E., McCorvy, J. D., et al. (2013). Structural Features for Functional Selectivity at Serotonin Receptors. *Science (New York, N.Y.)*. doi:10.1126/science.1232808

Van Walle, I., Lasters, I. & Wyns, L. (2004). Align-m—a new algorithm for multiple alignment of highly divergent sequences. *Bioinformatics*, *20*(9), 1428–1435. Oxford Univ Press.

Van Walle, I., Lasters, I. & Wyns, L. (2005). SABmark—a benchmark for sequence alignment that covers the entire known fold space. *Bioinformatics*, *21*(7), 1267–1268. Oxford Univ Press.

Wang, Chong, Jiang, Y., Ma, J., Wu, H., Wacker, D., Katritch, V., Han, G. W., et al. (2013). Structural Basis for Molecular Recognition at Serotonin Receptors. *Science (New York, N.Y.)*. doi:10.1126/science.1232807

Wang, Chong, Wu, H., Katritch, V., Han, G. W., Huang, X.P., Liu, W., Siu, F. Y., et al. (2013). Structure of the human smoothened receptor bound to an antitumour agent. *Nature*. doi:10.1038/nature12167

Wang, C.D., Buck, M. A. & Fraser, C. M. (1991). Site-directed mutagenesis of alpha 2A-adrenergic receptors: identification of amino acids involved in ligand binding and receptor activation by agonists. *Molecular pharmacology*, *40*(2), 168–179. ASPET.

Wang, Chengqi, Xi, L., Li, S., Liu, H. & Yao, X. (2012). A sequence-based computational model for the prediction of the solvent accessible surface area for α-helix and β-barrel transmembrane residues. *Journal of computational chemistry*, *33*(1), 11–7. doi:10.1002/jcc.21936

Wang, Chu, Bradley, P. & Baker, D. (2007). Protein-protein docking with backbone flexibility. *Journal of molecular biology*, *373*(2), 503–519. Elsevier.

Wang, G. & Dunbrack, R. L. (2003). PISCES: a protein sequence culling server. *Bioinformatics*, *19*(12), 1589–1591. Oxford Univ Press.

Wang, G. & Dunbrack, R. L. (2005). PISCES: recent improvements to a PDB sequence culling server. *Nucleic acids research*, *33*(suppl 2), W94–W98. Oxford Univ Press.

Ward, J., McGuffin, L. J., Buxton, B. F. & Jones, D. T. (2003). Secondary structure prediction with support vector machines. *Bioinformatics*, *19*(13), 1650–1655. Oxford Univ Press.

Ward, S. D., Hamdan, F. F., Bloodworth, L. M., Siddiqui, N. A., Li, J. H. & Wess, J. (2006). Use of an in situ disulfide cross-linking strategy to study the dynamic properties of the cytoplasmic end of transmembrane domain VI of the M3 muscarinic acetylcholine receptor. *Biochemistry*, *45*(3), 676–685. ACS Publications.

Ward, S. D., Hamdan, F. F., Bloodworth, L. M. & Wess, J. (2002). Conformational Changes That Occur during M3 Muscarinic Acetylcholine Receptor Activation Probed by the Use of an in Situ Disulfide Cross-linking Strategy. *Journal of Biological Chemistry*, *277*(3), 2247–2257. ASBMB.

Warne, T., Edwards, P. C., Leslie, A. G. & Tate, C. G. (2012). Crystal Structures of a Stabilized beta1-Adrenoceptor Bound to the Biased Agonists Bucindolol and Carvedilol. *Structure*, *20*(5), 841–849. Elsevier.

Warne, T., Serrano-Vega, M. J., Baker, J. G., Moukhametzianov, R., Edwards, P. C., Henderson, R., Leslie, A. G., et al. (2008). Structure of a beta1-adrenergic G-protein-coupled receptor. *Nature*, *454*(7203), 486–491. Nature Publishing Group.

Wass, M. N., Kelley, L. A. & Sternberg, M. J. (2010). 3DLigandSite: predicting ligand-binding sites using similar structures. *Nucleic acids research*, *38*(suppl 2), W469–W473. Oxford Univ Press.

Wedemeyer, W. J. & Baker, D. (2003). Efficient minimization of angle-dependent potentials for polypeptides in internal coordinates. *Proteins: Structure, Function, and Bioinformatics*, *53*(2), 262–272. Wiley Online Library.

Weiner, B. E., Woetzel, N., Karakaş, M., Alexander, N. & Meiler, J. (2013). BCL::MP-Fold: Folding Membrane Proteins through Assembly of Transmembrane Helices . *Structure* , . doi:10.1016/j.str.2013.04.022

Weiss, H. J., Witte, L. D., Kaplan, K. L., Lages, B. A., Chernoff, A., Nossel, H. L., Goodman, D., et al. (1979). Heterogeneity in storage pool deficiency: studies on granule-bound substances in 18 patients including variants deficient in alpha-granules, platelet factor 4, beta-thromboglobulin, and platelet-derived growth factor. *Blood*, *54*(6), 1296–1319. American Society of Hematology.

Wess, J., Han, S.J., Kim, S.K., Jacobson, K. A. & Li, J. H. (2008). Conformational changes involved in G-protein-coupled-receptor activation. *Trends in pharmacological sciences*, *29*(12), 616–625. Elsevier.

Woetzel, N., Karakas, M., Staritzbichler, R., Müller, R., Weiner, B. E. & Meiler, J. (2012). BCL:: Score—Knowledge Based Energy Potentials for Ranking Protein Models Represented by Idealized Secondary Structure Elements. *PloS one*, *7*(11), e49242. Public Library of Science.

Woetzel, N., Lindert, S., Stewart, P. L. & Meiler, J. (2011). BCL:: EM-Fit: Rigid body fitting of atomic structures into density maps using geometric hashing and real space refinement. *Journal of structural biology*, *175*(3), 264–276. Elsevier.

Wood, M. R., Hopkins, C. R., Brogan, J. T., Conn, P. J. & Lindsley, C. W. (2011). “Molecular switches” on mGluR allosteric ligands that modulate modes of pharmacology. *Biochemistry*, *50*(13), 2403–2410. ACS Publications.

Worth, C. L., Kreuchwig, A., Kleinau, G. & Krause, G. (2011). GPCR-SSFE: A comprehensive database of G-protein-coupled receptor template predictions and homology models. *BMC bioinformatics*, *12*(1), 185. BioMed Central Ltd.

Wu, B., Chien, E. Y., Mol, C. D., Fenalti, G., Liu, W., Katritch, V., Abagyan, R., et al. (2010). Structures of the CXCR4 chemokine GPCR with small-molecule and cyclic peptide antagonists. *Science Signalling*, *330*(6007), 1066. AAAS.

Wu, H., Wacker, D., Mileni, M., Katritch, V., Han, G. W., Vardy, E., Liu, W., et al. (2012). Structure of the human [kgr]-opioid receptor in complex with JDTic. *Nature*, *485*(7398), 327–332. Nature Publishing Group.

Xiong, H., Brugel, T. A., Balestra, M., Brown, D. G., Brush, K. A., Hightower, C., Hinkley, L., et al. (2010). 4-Aryl piperazine and piperidine amides as novel mGluR5 positive allosteric modulators. *Bioorganic & medicinal chemistry letters*, *20*(24), 7381–7384. Elsevier.

Yan, Q., Rammal, M., Tranfaglia, M. & Bauchwitz, R. (2005). Suppression of two major Fragile X Syndrome mouse model phenotypes by the mGluR5 antagonist MPEP. *Neuropharmacology*, *49*(7), 1053–1066. Elsevier.

Yanagawa, M., Yamashita, T. & Shichida, Y. (2009). Activation Switch in the Transmembrane Domain of Metabotropic glutamate receptor. *Molecular pharmacology*, *76*(1), 201–207. ASPET.

Yanagawa, M., Yamashita, T. & Shichida, Y. (2013). Glutamate acts as a partial inverse agonist to metabotropic glutamate receptor with a single amino acid mutation in the transmembrane domain. *Journal of Biological Chemistry*. ASBMB.

Yarov-Yarovoy, V., Schonbrun, J. & Baker, D. (2005). Multipass membrane protein structure prediction using Rosetta. *PROTEINS: Structure, Function, and Bioinformatics*, *62*(4), 1010–1025. Wiley Online Library.

Yohannan, S., Faham, S., Yang, D., Whitelegge, J. P. & Bowie, J. U. (2004). The evolution of transmembrane helix kinks and the structural diversity of G protein-coupled receptors. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(4), 959–963. National Acad Sciences.

Yuan, Z., Zhang, F., Davis, M. J., Bodén, M. & Teasdale, R. D. (2006). Predicting the solvent accessibility of transmembrane residues from protein sequence. *Journal of proteome research*, *5*(5), 1063–1070. ACS Publications.

Zhang, F. L., Luo, L., Gustafson, E., Lachowicz, J., Smith, M., Qiao, X., Liu, Y. H., et al. (2001). ADP is the cognate ligand for the orphan G protein-coupled receptor SP1999. *The Journal of biological chemistry*, *276*(11), 8608–15. doi:10.1074/jbc.M009718200

Zhang, Z., Miller, W., Schäffer, A. A., Madden, T. L., Lipman, D. J., Koonin, E. V. & Altschul, S. F. (1998). Protein sequence similarity searches using patterns as seeds. *Nucleic Acids Research*, *26*(17), 3986–3990. Oxford Univ Press.

Zhao, Z., Wisnoski, D. D., O'Brien, J. A., Lemaire, W., Williams Jr, D. L., Jacobson, M. A., Wittman, M., et al. (2007). Challenges in the development of mGluR5 positive allosteric modulators: the discovery of CPPHA. *Bioorganic & medicinal chemistry letters*, *17*(5), 1386–1391. Elsevier.

Zhong, X., Kriz, R., Seehra, J. & Kumar, R. (2004). N-linked glycosylation of platelet P2Y12 ADP receptor is essential for signal transduction but not for ligand binding or cell surface expression. *FEBS letters*, *562*(1), 111–117. Elsevier.

Zhou, H., Pandit, S. B. & Skolnick, J. (2009). Performance of the Pro-sp3-TASSER server in CASP8. *Proteins: Structure, Function, and Bioinformatics*, *77*(S9), 123–127.

Wiley Online Library.

Zhou, Y., Manka, J. T., Rodriguez, A. L., Weaver, C. D., Days, E. L., Vinson, P. N., Jadhav, S., et al. (2010). Discovery of N-aryl piperazines as selective mGluR5 potentiators with improved in vivo utility. *ACS medicinal chemistry letters*, *1*(8), 433–438. ACS Publications.