2018

COMPUTATIONAL AND EXPERIMENTAL APPROACHES TO OVERCOME THE G PROTEIN-COUPLED RECEPTOR STRUCTURAL KNOWLEDGE GAP

Lee Howard Wink

COMPUTATIONAL AND EXPERIMENTAL APPROACHES TO OVERCOME THE G PROTEIN-COUPLED RECEPTOR STRUCTURAL KNOWLEDGE GAP

by

Lee Howard Wink

A Dissertation

Submitted in Partial Fulfillment of the

Requirements for the Degree of

Doctor of Philosophy

Major: Chemistry

The University of Memphis

December 2018

Abstract

Wink, Lee Howard. PhD. The University of Memphis. December 2018. Computational and Experimental Approaches to Overcome the G Protein-Coupled Receptor Structural Knowledge Gap. Co-Major Professors: Abby L. Parrill-Baker, PhD. and Daniel L. Baker, PhD.

G protein-coupled receptors (GPCR) represent one of the largest families of integral membrane proteins in eukaryotes. Approximately 34% of FDA-approved drugs interact with GPCR. Despite recent GPCR structure reports, the number of unique structures known represents only a small percentage of all therapeutically-relevant GPCR. Experimental and computational approaches have been applied to the goal of decreasing this structural knowledge gap. Experimentally, water-soluble (WS) mutation sets intended to be transferrable throughout the GPCR family have been designed. A Green Fluorescent Protein (GFP) fusion partner was used as a folding reporter with a WS-GPCR mutant to more efficiently screen culture conditions for optimal protein expression in *E. coli*. Increasing the inducer concentration and post-induction growth time at low temperatures were observed to increase the amount of protein found in the soluble fraction based on whole-cell fluorescence measurements. Chaperone (GroEL) discovery in purified WS-GPCR mutant samples prompted the search for removal strategies during protein purification. Attempts to remove GroEL by applying strategies found in the literature were unsuccessful. The MOE Patch Analyzer tool was used to calculate hydrophobic surface patches on WS-GPCR mutants generated *in silico*. Addition of mutations at positions 1.31, 6.39, 6.57, 7.37, 7.48, and 7.51 on the β2-adrenergic receptor M15 mutant decreases the calculated size of major hydrophobic surface patches and may prove useful in future efforts to eliminate the GroEL interaction with WS-GPCR mutants. A computational benchmark study comparing performance of loop modeling algorithms available in the Rosetta and Molecular Operating Environment (MOE) software suites was also performed. Extracellular loop 2 (ECL2) from various GPCR

served as the modeling targets. Results from this benchmark study suggest that out of 1000 total loop models generated, the KIC with Fragments (KICF) and Next Generation KIC (NGK) algorithms within Rosetta sampled more models with near-atomic accuracy than the Cyclic Coordinate Descent (CCD) algorithm or the *de novo* search method within MOE for ECL2 targets with 24 residues or less. For longer loop modeling targets, it is likely that increased sampling is necessary to obtain models with near-atomic accuracy.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

5-HT1B	Serotonin $5HT_{1B}$ receptor
5-HT2B	Serotonin $5HT_{2B}$ receptor
5-HT4A	Serotonin $5HT_{4A}$ receptor
6His-tag	Hexa-histidine fusion tag
7TM	Seven-transmembrane domain
AA1R	Adenosine A_{1A} receptor
AA2AR	Adenosine A_{2A} receptor
ACM1	Muscarinic acetylcholine receptor M1
ACM2	Muscarinic acetylcholine receptor M2
ACM3	Muscarinic acetylcholine receptor M3
ACM4	Muscarinic acetylcholine receptor M4
AT1R	Angiotensin II receptor type 1
AT2R	Angiotensin II receptor type 2
β1AR	β1 adrenergic receptor
β2AR	β2 adrenergic receptor
bR	Bacteriorhodopsin
BW	Ballesteros-Weinstein numbering scheme
cAAT	Cytosolic aspartate aminotransferase
CB1	Cannabinoid receptor 1
CCD	Cyclic coordinate descent
CCR2	C-C chemokine receptor 2
CCR5	C-C chemokine receptor 5

CCR9	C-C chemokine receptor 9
CHAPS	3-[(3-Cholamidopropyl)dimethylammonio]-1-propanesulfonate
CRFR1	Corticotropin-releasing factor 1
CXCR1	CXC chemokine receptor 1
CXCR4	CXC chemokine receptor 4
DLS	Dynamic light scattering
DRD3	Dopamine receptor D3
E. coli	*Escherichia coli*
ECL	Extracellular loop
EM	Cryo-electron microscopy
EPR	Electron paramagnetic resonance
ETBR	Endothelin receptor type-B
FFAR1	Free fatty acid receptor 1
GEF	Guanine nucleotide exchange factor
GFP	Green fluorescent protein
GLR	Glucagon receptor
GPCR	G protein-coupled receptor(s)
GR	GroEL removal
GRK	G protein-coupled receptor kinases
GRM1	Glutamate receptor 1
H1R	Histamine H1 receptor
HSP	Heat-shock protein

HTS	Hight-throughput screening
ICL	Intracellular loop
IPTG	Isopropyl β-D-1-thiogalactopyranoside
kD	Kilodaltons
KIC	Kinematic closure
KICF	Kinematic closure with fragments
LB	Lysogeny broth
LBDD	Ligand-based drug design
LC-MS/MS	Liquid chromatography-mass spectrometry tandem mass spectrometry
LOMETS	Local meta-threading-server
LPAR1	Lysophosphatidic acid receptor 1
LRM	Lowest RMSD model
LRM-T10	Lowest RMSD model in the top 10 scored
LRM-T25	Lowest RMSD model in the top 25 scored
mAAT	Mitochondrial aspartate aminotransferase
MAPK	Mitogen-activated protein kinases
MBP	Maltose binding protein
MC	Monte Carlo
MOE	Molecular operating environment
MP	Membrane protein
NGK	Next-generation kinematic closure
NMR	Nuclear magnetic resonance

NOP	Nociceptin receptor
NTR1	Neurotensin receptor 1
OD	Optical density
OPRD	Delta-opioid receptor
OPRK	Kappa-opioid receptor
OPRM	Mu-opioid receptor
OX1R	Orexin receptor 1
OX2R	Orexin receptor 2
P2Y12R	Purinergic P2Y receptor 12
P2Y1R	Purinergic P2Y receptor 1
PAR1	Protease-activated receptor 1
PAR2	Protease-activated receptor 2
PDB	Protein data bank
PLB	Phospholamban
RHO	Rhodopsin
RMSD	Root mean square deviation
RPM	Revolutions per minute
S1PR1	Sphingosine 1-phosphate receptor 1
SBDD	Structure-based drug design
SDS-PAGE	Sodium dodecyl sulfate polyacrylamide gel electrophoresis
SMO	Smoothened receptor
SUMO	Small ubiquitin-like modifier
TB	Terrific broth

TBE	Terrific broth enriched
TM	Transmembrane
TSM	Top scored model
US28	Viral GPCR US28
VLS	Virtual ligand screening
WB	Western blot
WS	Water soluble
WT	Wild type
XLMS	Chemical cross-linking with mass spectrometry

Chapter 1

An Overview of Modeling Methods Applicable to G Protein-Coupled Receptors

1.1 Introduction

G protein-coupled receptors (GPCR) represent one of the largest families of integral membrane proteins in eukaryotes. GPCR serve central roles in amplifying and regulating a wide range of intracellular responses from extracellular stimuli. It is estimated that ~34% of FDA-approved drugs interact with GPCR targets including classes of drugs such as beta-blockers, antihistamines, and antipsychotics.[1,2]

In response to ligand activation, GPCR undergo conformational changes that influence coupling with intracellular signaling components, including with heterotrimeric G proteins (Gα, Gβ, Gγ subunits), β-arrestins, G protein-coupled receptor kinases (GRK), and other effectors.[3-5] In addition to ligand activation, some GPCR exhibit constitutive activity which is characterized by a basal level of activation in the absence of ligand stimulation.[6] After receptor-G protein coupling, the receptor functions as a guanine nucleotide exchange factor (GEF) to catalyze nucleotide (GDP/GTP) exchange in the Gα subunit which causes the heterotrimeric G protein subunits to dissociate. The GTP-bound Gα and Gβγ subunits are then free to independently interact with effector proteins to mediate the stimulation/inhibition of second messengers such as cyclic AMP, inositol triphosphate, and Ca^{2+}.[7] Different subtypes of the Gα subunit (Gαs, Gαi, Gαq, Gα12/13) are largely responsible for the selectivity between the various downstream signaling pathways.

Members of the GRK protein family play a major role in the termination of G protein signaling by phosphorylating Ser/Thr residues on the intracellular C-terminal tail of GPCR. Subsequently, the ubiquitous β-arrestin proteins (β-Arr1, β-Arr2) are recruited to the

phosphorylated GPCR and physically hindering further G protein signaling which leads to desensitization.[8] Additionally, β-arrestins function as scaffolds for proteins that are important for receptor internalization, and recycling/degradation through endocytic pathways. Furthermore, these proteins have the ability to act as scaffolds for signaling partners involved in modulating downstream pathways, such as the mitogen-activated protein kinases (MAPK), nonreceptor tyrosine kinases, and transcription factors, independently of heterotrimeric G protein signaling (biased signaling).[9,10]

There are over 800 human GPCR that have been classified into major families and subfamilies based on sequence identity.[11] One classification scheme organizes GPCR into six distinct groups, Classes A-F as follows: Class A (rhodopsin), B (secretin), C (metabotropic glutamate), D (fungal mating pheromone receptors), E (cyclic AMP receptors), and F (frizzled/smoothened).[12,13] Another common classification scheme is the GRAFS system that strictly considers the GPCR of vertebrates (excluding Classes D and E).[14] The GRAFS classification with their Class A-F equivalents in parentheses include the glutamate (Class C), rhodopsin (Class A), adhesion (Class B2), frizzled/taste2 (Class F), and secretin (Class B1) families.

1.2 GPCR structure-function relationships

All GPCR share a common topology featuring a core bundle of seven transmembrane (TM) α-helices with the N-termini and C-termini located on the extracellular and intracellular sides of the cell membrane, respectively (Figure 1.1). The extracellular and intracellular loops (ECL and ICL) are the protein segments that connect adjacent TM domains (ECL1-3, ICL1-3). The Ballesteros-Weinstein (BW) numbering system is commonly used to denote residue numbers for class A GPCR with a format which has two components (x.y): where x represents

the TM helix 1-7 and y represents the residue position relative to the most conserved residue in each TM which is designated x.50.[15] The variable y will incrementally decrease or increase from x.50 for residues proceeding towards the N-terminus or C-terminus, respectively.

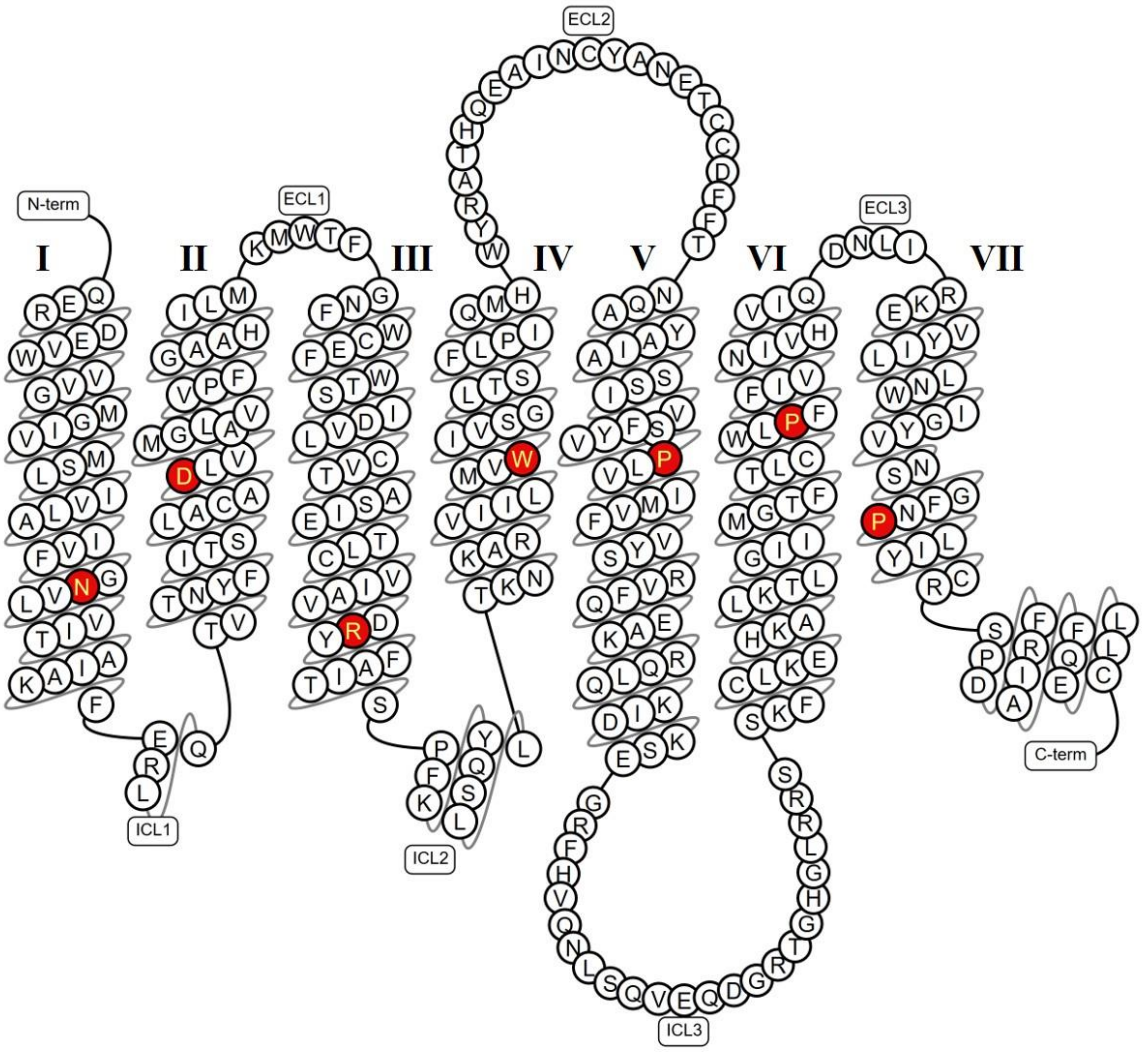

Figure 1.1: Snake plot diagram of the Class A GPCR, β_2 adrenergic receptor (β2AR). The roman numeral labels I-VII represent TM helices 1-7, ECL1-3 and ICL1-3 represent the extracellular and intracellular loops 1-3, respectively. The red highlighted amino acids represent the most conserved residues in each TM domain for Class A GPCR that is used in the Ballesteros-Weinstein numbering system (Table 1.1). The snake plot diagram was downloaded from the GPCRdb.[36]

Table 1.1: Conserved residues and sequence motifs of Class A GPCR.
The most conserved residues for each TM are in the BW row denoted by x.50. The variable x corresponds to each TM number.

BW	TM1	TM2	TM3	TM4	TM5	TM6	TM7
x.46	-	L	-	-	-	-	-
x.47	-	x	-	-	F	C	-
x.48	-	x	-	-	x	W	-
x.49	G	x	E/D	-	x	x	N
x.50	1.50 **N**	2.50 **D**	3.50 **R**	4.50 **W**	5.50 **P**	6.50 **P**	7.50 **P**
x.51	x	-	Y	-	-	F	x
x.52	x	-	-	-	-	F	x

GPCR activation mechanisms share common features that involve overall TM domain movements associated with rotamer changes in conserved side chains within the intracellular ends of the TM segments. These rotamer changes, collectively termed microswitches, stabilize the TM movements and prepare the intracellular side of the GPCR for G-protein binding.[16] The D/ERY sequence motif in TM3 is highly conserved among Class A GPCR. This motif features a salt bridge between the side chains of adjacent residues R (BW: R3.50) and D/E (BW: D/E3.49) that is found in inactive GPCR structures.[17] In addition, the conserved R3.50 residue forms a salt bridge with D6.30 that stabilizes the proximity of the intracellular ends of TM3 and TM6 for GPCR in an inactive state. This interhelical salt bridge is known as the ionic lock and was first observed in structures of bovine rhodopsin in the fully inactive state.[18] However, the stabilizing role of the ionic lock may be less important for other receptors as the acidic residue in position 6.30 is only conserved in 30% of GPCR.[16] Two examples of GPCR crystallized in the active state (rhodopsin[19,20] and β2AR*Gs complex[21]) show the R3.50 – D/E3.49 salt bridge broken and rotamer changes to R3.50 that result in interactions with the Gα subunit C-terminal helix. The NPxxY motif on the intracellular end of TM7 contains the highly conserved residue Y7.53 microswitch. Inactive state GPCR structures feature the Y7.53 side chain pointing towards TM1,

2 or Helix 8. On the other hand, all active state GPCR structures show a Y7.53 rotamer change resulting in the tyrosine side chain pointing towards the center axis of the TM bundle.[16]

In the context of rational drug design, it is important to understand how a diverse range of GPCR-ligand interactions result in global TM rearrangements that effectively expands the intracellular side of the TM bundle for heterotrimeric G protein (and other signaling partners) coupling. In comparisons of the inactive and active state adenosine A2A receptor (AA2AR) crystal structures, an agonist-induced rotamer change of the W6.48 residue (CWxPFF motif) corresponds with the outward movement (away from TM bundle) of TM6 on the intracellular side associated with GPCR activation.[22,23] Conversely, inverse agonists directly stabilize the inactive rotamer state of the W6.48 residue in observed complexes of the rhodopsin/retinal,[24] AA2AR/ZM241385,[25] and histamine H1/doxepin[26] crystal structures. However, the functional role of W6.48 in ligand-dependent activation is not a universal aspect of GPCR. For example, W6.48 rotamer shifts were associated with inactive functional states of the M2 and M3 muscarinic receptor crystal structures.[27,28] In addition, agonist- and inverse agonist-bound structures of β1AR and β2AR show no direct contacts between ligands and W6.48.[21,29,30] Analysis of inactive[30,31] and active[21,32] β2AR crystal structures suggests that the ligand-dependent interactions resulting in the outward movement of TM6 initially involve two serine residues in TM5, S5.42 and S5.46. Polar interactions between bound agonist and S5.42 and S5.46 stabilize an inward shift of the extracellular portion of TM5. The global rearrangement of TM5 leads to subsequent rotamer changes in residues I3.40 and F6.44, which corresponds to the outward movement of TM6. Further progress in GPCR structure determination will be necessary to expand our knowledge of GPCR structure-function relationships.

1.3 Challenges associated with GPCR structure determination

In general, membrane proteins have been difficult targets for traditional structure determination methods such as x-ray crystallography and solution NMR. Bovine rhodopsin, published in 2000, was the first atomic-resolution crystal structure of a GPCR (PDB: 1F88[24]). In most crystallization efforts, the target protein is expressed and purified from recombinant host systems, but many studies have highlighted the difficulties in expressing and purifying GPCR.[33] Recombinant protein expression was unnecessary in the case of rhodopsin due to the high amounts of endogenously expressed rhodopsin within bovine rod cell membranes. To date, rhodopsin remains the only GPCR to be purified and crystallized from its native source. This approach is not generally applicable due to the relatively low abundance of natively-expressed GPCR with respect to the amount of protein required for crystallization studies. Seven years after the structure of rhodopsin was solved, β2AR was the second GPCR to be successfully crystallized (PDB: 2RH1[34], 2R4R, 2R4S[35]). Currently (as of May 2018), there is a total of 50 unique GPCR structures (43 Class A, 4 Class B1, 2 Class C, 1 Class F) that have been experimentally determined (Figure 1.2).[36]

The seven-year lag and ensuing exponential increase in unique GPCR structures solved reflects many of the methodological advancements that were necessary to progress in the field.[37] GPCR structure determination faced challenges in all major stages of traditional crystallography workflows including functional protein expression/purification, crystallization (crystal formation), and diffraction/data collection. Early efforts in functional GPCR production for structural studies often suffered from low expression levels, poor receptor solubilization and stability. Improvements in protein engineering (thermostabilizing mutations and fusion partners),

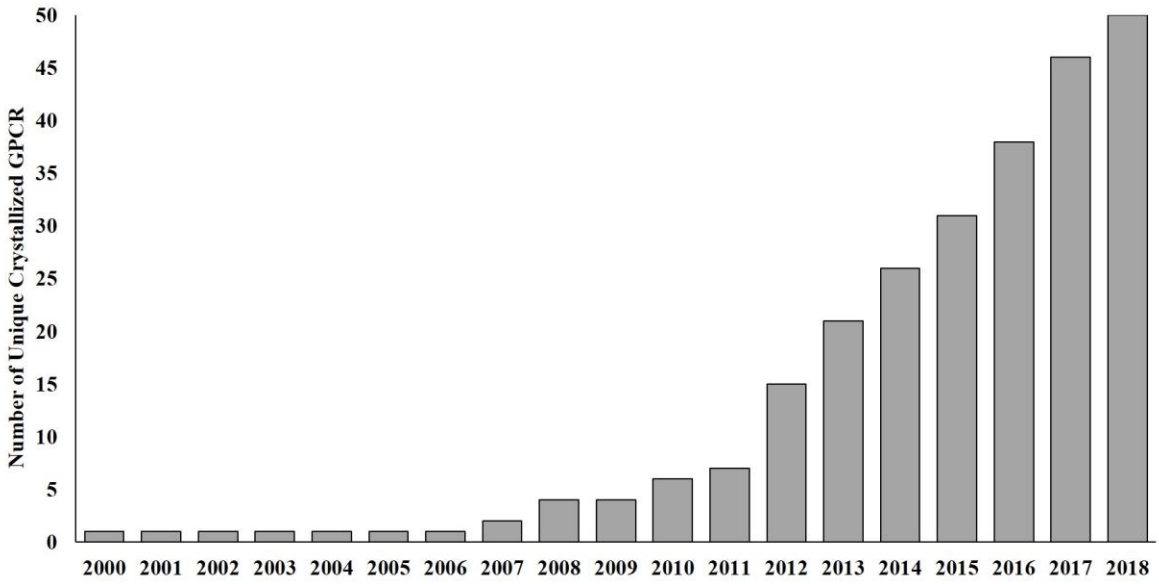

Figure 1.2: Timeline of unique GPCR crystal structure publications.
The timeline above shows the number of individual GPCR crystal structures that was published in each year since 2000. Crystal structure data was collected from GPCRdb structure statistics.[36]

detergent screening, and lipidic cubic phase-based crystallization were critical in the recent successes of GPCR structure determination.[38,39]

1.4 GPCR models can be used in structure-based drug design

Drug design efforts can be categorized into two methodological approaches: structure-based drug design (SBDD) and ligand-based drug design (LBDD).[40] In the absence of a receptor crystal structure or model, LBDD methods are more suitable for studying potential interaction sites on known ligand molecules that correlate with GPCR-mediated responses. SBDD methods can be applied to understudied GPCR and orphan receptors using GPCR crystal structures or models where very little (if anything) is known about cognate ligands (endogenous or synthetic). The cost advantage in using computational methods to guide ligand screening is significant when considering instrumentation and reagent costs associated with high-throughput screening (HTS) campaigns.

The absence of experimentally determined structures for many GPCR remains a major challenge in identifying endogenous ligands or designing novel drugs that target receptors with high selectivity and affinity.[41] In lieu of experimental structures, computational models of GPCR provide a predictive tool for studying potential receptor-ligand interactions *in silico* and developing hypotheses to prioritize experimental screening efforts. Virtual ligand screening (VLS) with GPCR structures (computational models or experimental structures) is a strategy that uses computational docking experiments to identify novel ligands from large databases for the target receptor.

Docking algorithms consist of two overall stages: sampling and scoring. For a target protein structure and a set of potential cognate ligands, docking algorithms sample the conformational space of ligands placed within the defined binding pocket of the target receptor. In the second stage, algorithms use scoring functions to evaluate the complementarity of predicted receptor-ligand docking poses based on estimates of binding energy.[42]
VLS requires some general knowledge of the spatial and energetic requirements for receptor-ligand binding *a priori*.[43]

1.5 Two major categories of protein modeling

Most protein modeling approaches can be categorized as either comparative modeling or *ab initio*/*de novo* methods. Comparative modeling methods rely on the idea that evolutionarily related proteins share a similar structure and function.[44] Therefore, the criteria for choosing a template for modeling usually include high amino acid sequence similarity and common function, which together suggest common evolutionary origin. Due to sharing a similar overall protein structure, GPCR are reasonable targets for comparative modeling with ~50 unique GPCR for which crystal structures are now available.

Traditional comparative modeling protocols start with database searches for suitable template structures that share high amino acid sequence identity to a target query. Subsequently, multiple sequence alignments between a target and appropriate candidate template(s) are performed. This is often the most critical stage in the comparative modeling procedure. GPCR sequence alignments often require manual inspection/adjustments to ensure that the most conserved residues of each TM helix are aligned and preferably excluding gaps from the highly conserved TM regions. If more than one potential template sequence/structure exists, multi-template modeling is typically preferable to single-template modeling if different templates in the overall alignment share greater local sequence identity in non-redundant regions of the target sequence.[45] In other words, the optimal components of different templates can be integrated to produce a better model than any single template.

Model building requires generating coordinates for the target sequence based on aligned template(s) structures. Three common classes of approaches are used: assembly of rigid bodies, segment matching, and satisfaction of spatial restraints.[46] The rigid body assembly method starts by identifying the conserved and variable regions of the target-template(s) sequence alignment. A framework for the superimposed template(s) is derived by taking an average of the atomic coordinates from the structurally conserved regions of the alignment. Often, this is weighted based on sequence similarity (greater sequence similarity being weighted heavier). The target model is built by fitting the core (conserved) rigid bodies onto the framework.[47] If only a single template is being used for model construction, the template coordinates of the aligned residues are copied directly to the target model. The non-conserved (often loop) regions are subsequently built by database searching or *ab initio* methods. The segment matching approach constructs models by using subsets of the template structure(s) atomic positions as guiding positions to

identify/assemble short, all-atom segments that fit the guiding positions.[48] The all-atom segments can come directly from the template, protein structure databases or conformational searches guided by a score function. The satisfaction of spatial restraints approach starts by generating restraints on the target structure to be modeled based on the alignment to the template structure.[49] The template-derived restraints are usually supplemented with confines on bond lengths, bond angles, dihedral angles, and atom contacts between non-bonded atoms which can be obtained from a molecular mechanics force field or knowledge-based statistics. In MODELLER, the distance and angle spatial relationships are expressed as conditional probability density functions to be directly used as spatial restraints for generating models.

For reliable GPCR models without any additional structure refinement steps, the lower threshold for target-template sequence identity is widely accepted to be in the 25-35% range.[50-52] Unfortunately, many orphan receptors (GPCR with unknown endogenous ligand and potentially unknown physiological role/function) and under-studied GPCR have low sequence identities (< 20%) relative to available crystallized GPCR templates.

When comparative modeling is not feasible, *ab initio* or *de novo* methods are common alternatives for protein structure prediction. The terms *ab initio* (Latin: "from the beginning") and *de novo* (Latin: "from the new") are often used interchangeably to categorize methods that attempt to predict the protein structure from a target amino acid sequence lacking any available homologous structure templates. *Ab initio* traditionally refers to structure prediction approaches that solely rely on physics-based principles, and *de novo* is a more general category of methods that exclude template-based strategies used in comparative modeling. There is a semantic debate among those in the field about whether strategies that use statistically-derived (knowledge-based) information from known protein structures can be truly considered *ab initio* methods. In

the context of this dissertation, the term *ab initio* will be used for approaches that do not use homologous template-based modeling methods.

In most cases, *ab initio* modeling methods need to accomplish at least two general tasks for protein structure prediction: 1) sampling conformational space to produce an ensemble of potential models, and 2) scoring/filtering models to discriminate putative native and near-native structures from non-native structures. There have been many excellent reviews highlighting the different approaches in conformational sampling[53-55] and scoring protein models.[56-58] *Ab initio* methods will often perform conformational sampling, filtering, and scoring tasks within distinct coarse-grained and full-atom stages of protein representation. Due to the rugged energy landscape associated with full-atom protein models, coarse-grained models are typically used for broad sampling of the conformational space, initially. Coarse-grained representations have a much smoother energy landscape which allows for more efficient conformational searches for the global energy minima.[59] Conformational searches for the global energy minima within the rugged energy landscape of full-atom models can potentially become trapped in local minima.

It is generally accepted that *ab initio* methods are more feasible for modeling smaller proteins (<150 residues) or segments of larger proteins (including loop regions, discussed in Chapter 3) that are unsuitable for template-based methods. However, GPCR (and integral membrane proteins, in general) are often much larger than 150 residues and therefore pose a greater challenge in adequately sampling the vast conformational space of the overall protein fold. One of the many approaches to reducing the conformational search space is the use of fragment-assembly.[60] In general, this approach uses a library of fragments derived from experimentally-determined protein structures and an algorithm that assigns optimal fragment combinations with segments of the target protein sequence based on sequence similarity or

predicted secondary structure.[61] Preliminary models of the overall protein fold are produced by a combinatorial fragment assembly algorithm and subsequently optimized/refined for model evaluation. The fragment-assembly approach has been implemented in the widely successful structure prediction method, I-TASSER[62,63] and many protocols within the modeling software suite Rosetta.[64,65]

1.6 Hybrid modeling approaches applied to membrane proteins and GPCR

The Rosetta software suite offers comparative modeling and *ab initio* structure prediction tools that are applicable to GPCR: Rosetta Comparative Modeling (RosettaCM) and Rosetta Membrane Protein (RosettaMP) structure prediction.[66,67] RosettaCM is a template-based method that requires alignments of the target sequence with one or multiple template sequences. This method implements Monte Carlo (MC) moves for the following: inserting fragments in unaligned segments, replacing segments derived from different template structures randomly, and Cartesian-space minimization with the Rosetta centroid (coarse-grained representation of amino acid side chains) energy function. MC moves are followed by optimization in the full-atom stage. The RosettaMP framework has developed over time to improve membrane protein modeling by implementing a modified version of the Rosetta energy function that accounts for the membrane environment and membrane protein-specific constraints (helix-helix packing in α-helical bundles) in structure prediction applications.[68-70]

GPCR-I-TASSER is a hybrid modeling method that incorporates either threading by the Local Meta-Threading-Server (LOMETS) to find template structures in the PDB or *ab initio* TM helix assembly to build the TM helix bundle framework.[71] GPCR models are built by assembling fragments derived from the LOMETS target-template threading alignments or *ab initio* TM helix models following a procedure extended from I-TASSER.[72] The major difference is that structure

assembly simulations are guided by a knowledge-based force field that is specific to GPCR. Models are built with structural restraints derived from a database that contains GPCR-specific experimental mutagenesis data (GPCR-RD).

Continued progress in the membrane protein structure prediction field is highlighted by promising examples of composite approaches that incorporate modeling restraints from sparse experimental data (i.e. nuclear magnetic resonance (NMR) spectroscopy,[73] cryo-electron microscopy (EM),[74] electron paramagnetic resonance (EPR) spectroscopy,[75] and chemical cross-linking with mass spectrometry (XLMS)[76]) with additional sequence-based methods (secondary structure predictions, TM span/topology predictions) to improve the accuracy of membrane protein structure prediction.[67,77] For example, the Meiler group has developed the BCL::MP-Fold method for *de novo* membrane protein structure prediction.[78] To efficiently sample the vast conformational space of large membrane proteins, the MC algorithm uses predicted secondary structural elements to assemble the overall protein fold within a virtual membrane environment. Knowledge-based scoring functions designed for membrane proteins are used to approximate the intermediate models' free energy after each MC step. Experimental data can be incorporated into the BCL::MP-Fold algorithm to account for model simplifications during the sampling and scoring steps.[79] The output from the BCL::MP-Fold algorithm is a simplified fold composed of secondary structural elements with only limited deviations from idealized dihedral angles. These simplified models are subsequently used as input for optimization/refinement steps to produce full-atom models using tools from the Rosetta modeling suite.

1.7 Perspectives

Results from the most recent GPCR DOCK blind prediction experiments (2010, 2013) showed that overall comparative modeling techniques perform better than *ab initio*/*de novo*

methods when suitable templates are available for the modeling target.[80,81] With increasing numbers of GPCR structures available in the PDB and other publicly available experimental data, more receptors can be confidently modeled using template-based methods. Based on the trend of unique GPCR structures solved to date, the number of available structures in the PDB will inevitably increase. Receptors that lack suitable comparative modeling templates can be modeled using *ab initio* methods due to improvements in conformational sampling techniques and membrane protein-specific score functions. Incorporating experimental data in the form of modeling constraints can also be used in structure prediction methods.

The topics of the subsequent chapters will cover additional experimental and computational approaches to studying GPCR structures. In regard to experimental approaches, Chapter 2 will go more in depth about a novel family-wide protein engineering strategy to aid in X-ray crystallography structure determination methods for GPCR. As for computational approaches, Chapter 3 will focus on loop modeling methods for one of the most challenging components to GPCR structure prediction via template-based methods.

Chapter 2

Designing and Characterizing Water-Soluble G Protein-Coupled Receptors

2.1 Introduction

Nearly half of the 800 GPCR sequences in the human genome are unique, non-olfactory GPCR that are potential therapeutic targets.[82,83] As of May 2018, crystal structures for 50 unique GPCR were available in the Protein Data Bank (PDB: www.rcsb.org).[84] While this is a marked increase from the two unique GPCR for which structures were available in 2007, structures for an overwhelming majority of GPCR have not been experimentally determined. This lack of structures applies to membrane proteins in general, when compared to the number of water-soluble protein structures that have been deposited to the PDB. According to the Mpstruc database, there are currently 1,441 publications of membrane protein structures and a total of 2,605 coordinate files in the database.[85,86] In comparison, there are a total of 131, 837 protein structures available in the PDB.[84] This dearth of structures for membrane proteins is due, in part, to their intrinsic instability in aqueous solutions that are typically required for structural studies. Structural characterization of GPCR is further hindered by the dynamic range of receptor activation states dependent on conformational changes inherent to GPCR structure and function.[87]

This chapter details our group's efforts toward developing water-soluble mutants of GPCR to overcome the challenges of determining membrane protein structures. Figure 2.1 is a schematic overview of our efforts divided into three distinct sub-topics. The overall goal and novelty of the project was to identify a minimal number of mutations on commonly outward-facing, lipid-exposed residues in most GPCR family members to generate a transferable mutation set (i.e. a single approach that could be applied across all class A GPCR) that would allow for the

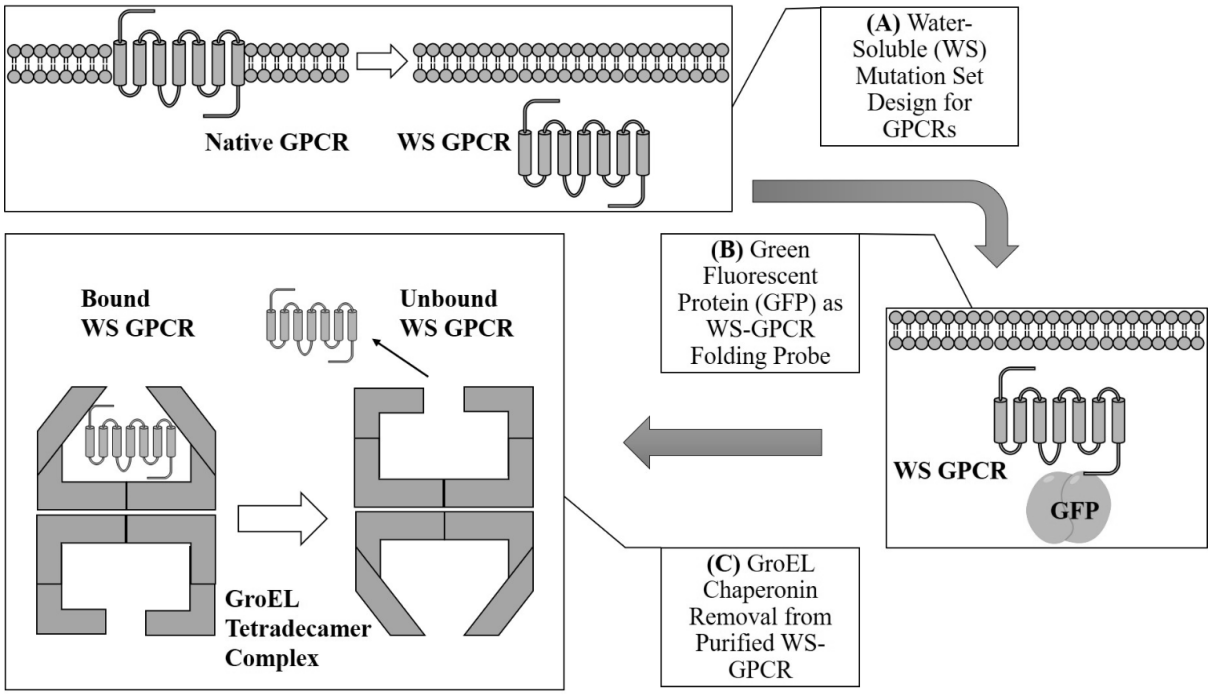

Figure 2.1: Schematic overview of the water-soluble GPCR project progression. The following sub-topics are covered in this chapter: (A) The design process for developing transferable mutation sets for WS-GPCR, (B) The role of Green Fluorescent Protein (GFP) as a fusion partner folding probe during WS-GPCR mutant expression in *E. coli*. (C) The strategies employed to remove GroEL chaperonin contaminants from purified WS-GPCR mutants.

water-soluble expression of functional GPCR for subsequent structural studies.

Historically, many questions related to the biophysical characteristics of membrane proteins went unanswered due to the difficulties in studying lipophilic biomolecules. In 1959, Kauzmann hypothesized that, in general, soluble (globular) proteins can be characterized by having polar/hydrophilic amino acid residues on the solvent-exposed surface and non-polar/hydrophobic residues buried within the interior fold of the protein as a result of satisfying components of the free energy required to drive protein folding.[88] Kauzmann's ideas were supported by the results of John Kendrew in 1960, when he published one of the first three-dimensional structures of a globular protein, myoglobin, at 2 Å resolution.[89] In the following years, hypotheses were formed about the general structure of integral membrane proteins. A

popular notion was that membrane proteins could be characterized as having an "inside-out" molecular structure relative to soluble proteins in aqueous solution.[90] This notion presumed that membrane proteins contain polar residues buried within the core protein fold and non-polar residues on the lipid-exposed surface. Initially, this idea was attractive because it could explain the basis for helix-helix associations of α-helical transmembrane (TM) domains and the lipid-exposed surface interactions.

However, pioneering research in the late 1980s and 1990s challenged this notion of membrane proteins simply being "inside-out" versions of soluble proteins by analyzing the similarities and differences between members of the two categories.[90-92] Sequence analysis confirmed that the lipid-exposed residues of membrane proteins tend to be more hydrophobic than the buried core residues, and the solvent-exposed surface residues of water soluble proteins were more hydrophilic than the interior. However, the buried residues of the interiors were comparably hydrophobic for both soluble and membrane proteins. Structural analysis of experimentally known α-helical TM domains suggested that there is no correlation between the hydrophobicity of a residue and any preferred location in the protein. Furthermore, the hydrophilic moment of an α-helix is a weak indicator of helix orientation in a lipid membrane – i.e. the hydrophilic residues are not uniformly directed towards the protein center of mass. Therefore, the term "inside-out" inaccurately describes the relative polarities of surface and interior residues for membrane proteins compared to soluble proteins. By extension, these findings suggest the potential for engineering water-soluble (WS) analogs of membrane proteins by mutating hydrophobic surface residues that are typically exposed to a lipid environment into polar residues that would interact more favorably with an aqueous environment.

2.1.1 Review of engineered water-soluble membrane protein analogs

The general idea of engineering individual membrane proteins into water-soluble analogs has been explored in previous work. One of the first published efforts toward engineering WS analogs of membrane proteins focused on bacteriorhodopsin (bR), an integral membrane protein with seven TM α-helices.[93] The Engelman group used two different strategies to identify and mutate solvent-accessible residues to generate different candidate WS constructs of bR. Since bR exists as a trimer in the native lipid environment, they engineered a bR trimer construct where 14.9% of surface residues were mutated to polar, helix-preferring residues. In addition, three different candidate WS constructs were developed based on the monomeric form of bR with varying degrees of surface mutations (13.5-24.3%). The WS trimeric bR construct expressed poorly, if at all, in E. coli. While the WS monomeric bR constructs expressed very well in E. coli, they were only soluble in solutions containing urea (2M – 8M). Further analysis revealed that the retinal pocket of bR was likely misfolded in this construct. Schiff base formation between retinal and Lys216 is an indicator of functional and properly-folded bR. Schiff base formation was not observed with the mutant bR constructs. Despite the limited success in engineering a WS membrane protein analog, the bR example is significant in that it inspired other research projects aimed at the same overall goal. In subsequent years, successful reports of engineering WS membrane proteins would emerge with different target proteins: Phospholamban (PLB),[94,95] the KcsA potassium channel,[96] the μ-opioid receptor (OPRM),[97,98] and the KchAfu104 potassium channel.[99]

The DeGrado group successfully used available crystal structures and mutagenesis data to computationally design WS analogs of PLB, a single-pass TM protein that forms a pentamer in the native lipid environment, and the KcsA potassium channel, a tetrameric complex with

monomer subunits containing two α-helical TM domains. The WS PLB construct had 11 mutations (21% of protein sequence) to lipid-exposed residues and displayed properties that were equivalent to the wild-type PLB in terms of secondary structure (α-helical), oligomerization states, and phosphorylation response.[94] The WS KcsA potassium channel went through a preliminary design of 29 mutations that yielded a water-soluble protein with consistent secondary structure and agitoxin-2 binding compared to the wild-type, but higher-order oligomer states were also observed. An optimized design using 31 mutations (25% of protein sequence) resulted in a protein that retained the structural and functional aspects of the native protein and was soluble in aqueous solution.[96]

The KchAfu104 potassium channel posed a different challenge in developing a WS variant compared to the previous examples discussed in that it lacked an experimentally-determined structure at high resolution upon project initiation.[99] While most known potassium channel homologs share the same overall tertiary and quaternary structure (tetramers), sequence analysis suggested that KchAfu104 has an irregularly polar C-terminus compared to other potassium channels which could have unknown structural impact. In this case, multiple sequence alignments with known prokaryotic potassium channels and TM sequence prediction tools were used to select lipid-exposed residues to mutate into polar/hydrophilic residues. After initial design iterations, the final WS KchAfu104 construct contained 10 mutations (~10% of protein sequence) in the transmembrane helix region and a flexible linker with an N-terminal Maltose Binding Protein (MBP, 40 kD; KchAfu104, 14 kD) fusion partner that has been shown to improve recombinant protein solubility.[100] The final MBP-KchAfu104 mutant expressed as a soluble protein in bacteria that could be purified without aggregation problems. Importantly, the

WS mutant retained the tetrameric state that is expected of native potassium channels in lipid membranes.

To our knowledge, OPRM is currently the only successful example from the literature of engineering a WS GPCR variant. The Liu group published an initial WS-OPRM design that altered 53 amino acids (~18% of expressed protein sequence) in the lipid-facing TM domain to polar, charged amino acids (mostly Glu and Lys). The first design targeted TM residues with >40% solvent-exposure for mutation based on comparative models of OPRM built using crystal structures of the β_2 adrenergic receptor (β2AR), and bovine rhodopsin as templates.[97] A modification to the WS design was described after the crystal structure for OPRM was published (4DKL[101]) that reverted 7 of the 53 mutations (46 total, ~16% of expressed protein sequence) back to the original wild-type sequence.[98] Based on the experimentally-determined structure, the 7 residues were identified to be located near the ligand binding pocket in the experimentally-determined structure. Mutating inward-facing residues near the binding pocket is potentially detrimental to receptor function and likely does not enhance protein solubility in aqueous solution. The re-engineered WS OPRM retained functionality based on ligand binding experiments with naltrexone and was soluble in aqueous solution. However, the authors note that while the purified WS receptor remains predominantly in a monomeric form, dimers and higher-order oligomers are observed to form over time (after ~2 months). The purification process of both versions of the WS OPRM required using an initial 0.1% sodium dodecyl sulfate (SDS) concentration; subsequently, a 0.01% SDS concentration was used in solutions to avoid protein aggregation in functional assays. This caveat requiring detergent to purify the WS GPCR variant will be readdressed in the discussion section.

2.1.2 The Green Fluorescent Protein (GFP) fusion partner is an efficient probe to monitor aspects of recombinant protein expression

Since its discovery, Green Fluorescent Protein (GFP) has been well-established as a versatile tool for protein characterization studies. This section briefly highlights the wide-range of applications for GFP-fusion proteins in the literature but will focus on its reported use as a protein-folding reporter to make recombinant protein expression trials more time-efficient. The utility of GFP is derived from the stable chromophore formed from residues 65-67 (Ser-Tyr-Gly) upon proper folding into the native tertiary structure, characterized by the β-barrel motif.[102] C-terminal (Ct-) GFP fusion proteins have been used to study protein trafficking and localization dynamics in various cell hosts.[103-105] It is often a reasonable concern to question the potential effects a fusion protein can have on the native protein of interest in terms of structure and function. However, several studies have shown that GPCR-GFP fusions retain functionality equivalent to the native protein. A Ct-GFP fusion partner helped visualize internalization of a putative GPCR, TGR5, upon response to bile acid, the confirmed ligand for TGR5 (hGPCR19).[106] Additionally, Ct-GFP fusion partners were used in independent studies of the β_2 adrenergic receptor and the angiotensin II type 1a receptor, and do not appear to perturb the native functionality of the respective GPCR in terms of membrane trafficking or ligand-induced G-protein signaling in mammalian cells.[107,108] If these examples of GPCR-GFP fusion proteins are able to retain the pharmacological properties of their native form, then it is logical to conclude that these receptors were able to fold properly and localize to the cell membrane undisturbed by the Ct-GFP moiety.

The wild-type GFP exhibits fluorescence excitation/emission maxima at 395/508 nm.[102] The fluorescent properties of GFP are distinguishable from tryptophan fluorescence which

typically shows major excitation/emission at ~280/350 nm.[109] Fluorescence signals attributed to the GFP chromophore will not be observed if the protein becomes kinetically trapped in a misfolded or aggregated state (thus preventing formation of the functional Ser-Tyr-Gly chromophore).[110] In the context of optimizing recombinant protein expression, a Ct-GFP fusion partner can function as a sensitive probe for proper folding of the N-terminal (Nt) protein of interest.[111] In a study of 20 proteins with and without Ct-GFP tags expressed in *E. coli*, the fluorescence signal from the GFP-tagged proteins correlated with the productive folding of the same protein lacking GFP.[112] This has been shown to apply to both soluble proteins and membrane proteins in their respective aqueous and lipid environments.[112-116]

A Ct-GFP fusion strategy was utilized herein with water-soluble GPCR targets in an attempt to save time during protein expression optimization trials. This stage of the project is depicted in the overview Figure 2.1 (B). Typically, optimization trials are conducted with cell culture samples taken at certain intervals after the induction of protein expression in the host. Sample preparation traditionally involves cell lysis to obtain the whole-cell lysate samples and centrifugation to obtain soluble/insoluble fraction samples for visualization via SDS-PAGE and western blot analysis. With traditional gel electrophoresis and western blotting protocols, an extra day or two is necessary to analyze results and plan subsequent experiments. Furthermore, to assess the three-dimensional fold and function of a target protein, it is often necessary to scale-up expression for purification which diverts more time away from progressing to the next step. GFP fusions provide a more efficient method of assessing protein expression and folding under various conditions via whole-cell fluorescence measurements in high-throughput formats (i.e. 96-well plate). However, this method does not depreciate the value of performing SDS-PAGE in tandem with whole-cell fluorescence measurements because gel analysis provides crucial

information about experimental outcomes. Additionally, properly-folded GFP fusions can be unambiguously identified via in-gel fluorescence bands that correspond to the expected molecular weight of the target protein. The stability of the GFP chromophore allows for in-gel fluorescence visualization under SDS-PAGE conditions.[117]

2.1.3 Strategies to remove GroEL chaperonin complex during WS-GPCR purifications

Concurrently with the implementation of the GFP fusion strategy to optimize WS-GPCR expression, parallel work by previous lab members, Dr. Alexandra Kikonyogo and Dr. Samantha Gacasan, had identified expression & purification conditions that were producing sufficient amounts of soluble protein and were in the process of functional and structural characterization of purified WS-GPCR. The target protein construct was a WS mutant of the β_2-adrenergic Receptor (SUMO-β2AR:M13-FLAG) that contained a relatively small Nt-fusion protein, SUMO (Small Ubiquitin-like Modifier, ~12 kD), and a Ct-FLAG sequence (DYKDDDDK, ~1 kD) for Anti-FLAG M2 antibody affinity purification. Due to the abundance of literature data for the β2AR, it was chosen as a proof-of-principle target for testing and validating the WS mutation designs. In collaboration with Dr. Darcie Miller at St. Jude Children's Research Hospital, crystallization trials were performed with purified SUMO-β2AR:M13-FLAG and produced protein crystals of sufficient size for data collection. Despite the 2.6 Å resolution X-ray diffraction data, the final stages of structure determination by molecular replacement were unsuccessful with available β2AR or SUMO crystal structures already deposited in the PDB. At that point, collaborators at St. Jude further characterized the 60 kD protein by in-gel trypsin digestion and liquid chromatography-tandem mass spectrometry (LC-MS/MS). For protein identification, database searches with St. Jude's JUMP software produced matches that suggested the 60 kD band contained a bacterial chaperonin protein, GroEL. This protein went

unnoticed in purification sample analysis because the 60 kD GroEL monomer co-migrated with the target protein, SUMO-β2AR:M13-FLAG (60 kD), during SDS-PAGE.

GroEL/ES is an ATP-dependent chaperonin system that functions as an integral component of the cellular protein folding machinery in bacteria.[118] However, chaperonins are prevalent in many forms of life outside of bacteria.[119] In general, chaperonins are structurally distinct from other ATP-dependent chaperones in that they construct large double ring complexes ranging from 800-1000 kD in size. Each ring structure contains an internal chamber where protein substrates are isolated from the crowded molecular environment of the cell to simultaneously prevent aggregation and promote proper protein folding.[120] Specifically, seven GroEL monomers (57 kD, observed ~60 kD) form the two stacked heptameric rings that constitute the tetradecameric complex. Each ring of the GroEL tetradecamer interacts with heptameric complexes of the GroES co-chaperonin (10 kD), which forms a dome-shaped structure that caps access to the GroEL-substrate chambers. Although the preferred size range for GroEL substrates is estimated to be ~20 – 60 kD, at least 30 of the 252 known interacting proteins are over 60 kD.[121] In fact, it is known to interact with an 82 kD protein, the yeast mitochondrial aconitase.[122] Each GroEL monomer contains three distinct domains: apical, equatorial, and intermediate. The apical domain is the component responsible for substrate protein (native/non-native) and GroES binding. In the tetradecameric complex, the apical domains form the two entrances to the ring cavities. The equatorial domain is responsible for the ATPase activity of the chaperonin and the intermediate domain acts as a hinge region between the apical and equatorial domains during conformational changes induced by ATP.[121] A schematic diagram of the GroEL/GroES mediated protein folding cycle is illustrated in Figure 2.2. Apo-GroEL subunits are in equilibrium between a high affinity ATP state and low affinity

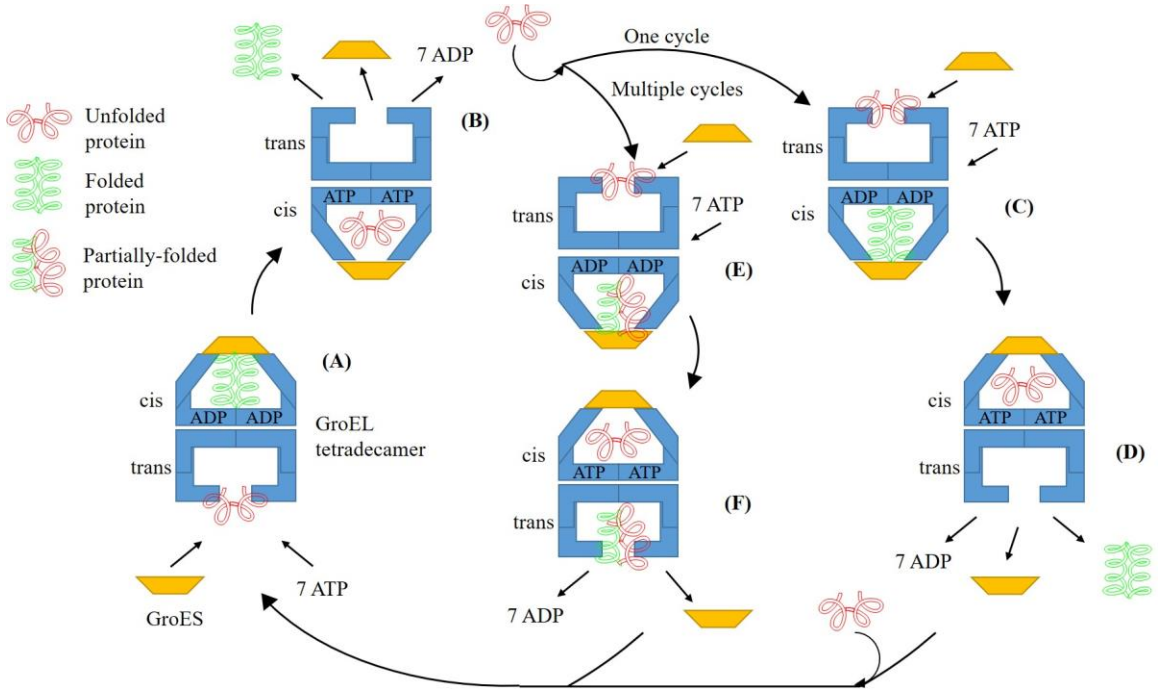

Figure 2.2: GroEL/ES/ATP protein folding cycle schematic diagram.
The predominant model of GroEL-assisted protein folding involves the cis/trans heptamer rings functioning in an alternating cycle with GroES. The ring that is associated with GroES is denoted the cis ring, and the opposite ring with an empty central cavity is denoted the trans ring. (A) The asymmetric GroEL/ES complex forms as the unfolded protein substrate interacts with the apical domains of the empty trans ring. (B) ATP-binding induces conformational changes in the lower ring that allows for GroES to bind at the apical domain and unfolded protein to enter the central cavity. Concurrently, ADP and GroES dissociate from the upper ring which permits the release of folded substrate protein bound from the previous cycle. (C – E) The unfolded protein in the new cis ring can fold in the central cavity during the time required for ATP hydrolysis. Depending on the folded state of the substrate protein, one or multiple cycles may be necessary to attain folded protein. (D) If only one cycle of folding is required, the binding of new unfolded protein, ATP, and GroES to the trans ring will cause release of the cis ring components, folded protein, ADP, and GroES. (F) If the substrate protein folding kinetics is slower than the rate of ATP hydrolysis, then multiple rounds of the folding cycle may be necessary.

ATP state. Within each heptameric GroEL ring, ATP binding exhibits positive cooperativity. However, ATP binding exhibits negative cooperativity between the two heptameric rings.[123] Cooperative ATP binding to one ring stabilizes the high ATP affinity state and induces a conformational change involving the apical domain to facilitate GroES binding. The ATP-bound GroEL/ES side of the dual ring complex is termed the cis ring and the opposite ring lacking

ATP/GroES is termed the trans ring. The negative cooperativity between the two GroEL rings after ATP/GroES binding causes the trans ring to be in a low ATP affinity state and results in the asymmetric nature of the GroEL/ES functional cycle. Conformational changes in the cis ring are induced when ATP is hydrolyzed to ADP. In the ADP-bound state, the negative cooperativity between the cis/trans rings is reduced which allows for the trans ring to bind ATP and the cis ring to release ADP. When GroES subsequently binds to the ATP-bound ring, a new cis-ring is designated, and a single cycle is complete.[120]

In the presence of non-native (unfolded) substrate protein, the apical domains of GroEL subunits in the trans ring (ATP/GroES-free) bind with the substrate protein via hydrophobic interactions. After ATP binding/hydrolysis and GroES dissociation from the cis ring, ATP can bind to the trans ring with substrate bound and induce a similar conformational change as in the apo state. However, the conformational change in the context of a bound substrate protein is followed by encapsulation of the substrate as GroES binding to the apical domains displaces the substrate into the cavity of the GroEL/ES complex. The energy penalty for encapsulating the substrate protein is likely the reason for the ATP-dependence of this functional cycle. Once the substrate protein is isolated in the GroEL/ES complex cavity, it can undergo "proper" folding unhindered by a crowded molecular environment. While it is still debated whether the relevant functional cycle of GroEL/ES with substrate protein involves the two GroEL rings acting sequentially (asymmetric GroEL/GroES complex cycle) or simultaneously (symmetric GroEL/$(GroES)_2$), it is likely that the asymmetric GroEL/ES complex cycle model is more relevant *in vivo*.[120] Although it was observed that the presence of permanently/partially unfolded substrate proteins (reduced α-Lactalbumin, and casein) promoted the formation of symmetric GroEL/$(GroES)_2$ complexes; symmetric complexes were not observed in significant populations

in the presence of the foldable substrate proteins tested.[124] Thus, the exact mechanism of action may vary depending on the nature of the substrate protein and GroEL/ES complex interactions. Table 2.1 summarizes the numerous reports of GroEL co-purification with recombinant proteins expressed in E. coli hosts and strategies to remove the GroEL contaminant. One approach uses an ATP (ATP-$MgCl_2$-KCl) wash buffer in a two-hour incubation step during protein purification to release GroEL from the protein of interest.[125]

Table 2.1: Summary of GroEL removal strategies during protein purification.
All referenced strategies use purification buffers that include 2-10 mM ATP in addition to the corresponding reagents below. A few references pertain to removing Cpn60, a cold-adapted chaperonin from *O. Antarctica* that shares very high sequence identity (74%) with GroEL.

Reagent	Reagent function	References
ATP	Stimulates GroEL protein substrate release	125,126
Casein	Competes for GroEL substrate binding	127-130
Detergent	Disrupts hydrophobic binding between GroEL and protein substrate	131,132
Urea	Destabilizes GroEL complex to disrupt binding of protein substrate	133-135
GroES	Stimulates GroEL protein substrate release	126,136

The rationale behind the use of ATP to remove GroEL is derived from the ATP-dependence of the GroEL conformational changes/functional cycle explained previously. It is not uncommon for various detergents and glycerol to also be included in purification buffers intended to remove GroEL. Co-purification of a GroEL contaminant was reported with the protein, MdtM, a ~45 kD multidrug resistance transporter protein that contains 12 transmembrane spanning segments.[132] GroEL was successfully removed from the transporter protein by using an ATP wash buffer containing n-dodecyl-β-D-maltoside (DDM), a nonionic detergent often used to stabilize and

purify membrane proteins, and glycerol, a polyol (major class of osmolytes or chemical chaperones) that can assist protein folding by increasing stability and preventing aggregation.[137] de Marco et al. reported success in using an ATP wash buffer containing glycerol and Triton X-100 (another common nonionic detergent) to remove GroEL and DnaK (70 kD chaperone protein, bacterial analog of HSP70) contaminants from recombinant protoporphyrinogen IX oxidase (PPO), a nuclear-encoded membrane-bound enzyme that is <57 kD.[131] The observation that non-ionic detergents can replace GroEL in facilitating the folding of denatured rhodanese (a mitochondrial thiosulfate sulfurtransferase) suggests that GroEL substrate binding/re-binding can be disrupted in the presence of detergent.[138,139]

Another GroEL removal strategy involves using casein, a set of proteins (αS1, αS1, β, κ) found in mammalian milk with many characteristics of intrinsically disordered proteins and partially denatured proteins due to the high proline content, lack of disulfide bonds and exposed hydrophobic regions.[140,141] In one of the early GroEL protein refolding studies, casein was shown to have a high binding affinity to GroEL in the presence of Mg-ATP *in vitro*.[142] This characteristic allows for casein to compete for the GroEL substrate binding cavities and displace the bound protein(s). While examples in which casein was used to release GroEL-bound proteins have been relatively successful, there have been instances where casein was insufficient.[127-130] In a study of GroEL interactions with a cytosolic and mitochondrial aspartate aminotransferase (cAAT and mAAT, respectively), Scherrer et al. found that cAAT could be recovered from GroEL in an active form by adding casein in 100 – 150-fold molar excess over cAAT. Instead of casein, GroES (and Mg-ATP) addition in 2-fold molar excess over GroEL was necessary for mAAT to be recovered in an active form.[129] To further illustrate the importance of GroES, a

study involving the C-C chemokine receptor 5 (CCR5) produced in a cell-free expression system showed that GroEL + GroES assisted folding was more efficient than GroEL alone.[143]

2.2 Methods

2.2.1 Designing a transferable, water-soluble mutation set

The amino acid sidechains of glutamate (E), aspartate (D), and serine (S) contribute to protein solubility in aqueous solution more than the other common amino acids.[144] Additionally, E, D, and S have relatively similar alpha helix propensities to alanine, the amino acid with the highest such propensity.[145] Therefore, we rationalized that using these amino acids to replace lipid-exposed sites on GPCR alpha helices should preserve the TM secondary structure and contribute to protein solubility in aqueous solution. To determine which residues are commonly outward-facing in the TM region, we used solvent-accessible surface areas for amino acid sidechains of crystallized GPCR relative to the same sidechain in a random coil configuration of Gly-X-Gly.[146,147] Using the modeling suite MOE 2013.08 (Molecular Operating Environment[148]), we created databases for E, D, S point mutants at the outward-facing residue for all GPCR crystal structures listed in Table 2.3. The change in hydrophobic surface area was calculated for each set of point mutants relative to the wild-type GPCR structure to determine which mutations would contribute most to protein solubility.

2.2.2 Sub-cloning genes into expression vectors

All standard molecular biology protocols that did not accompany the various kits used for sub-cloning were performed according to *Molecular Cloning: A Laboratory Manual*.[149] The β2AR:M13 DNA sequence was PCR amplified with the Q5® High-Fidelity DNA Polymerase kit (New England Biolabs, Ipswich, MA) using the verified SUMO-β2AR:M13-FLAG sequence as a template. Primers that were used for sub-cloning are listed in Appendix Table A.1. The

QIAquick Gel Extraction kit (Qiagen, Germantown, MD) was used for DNA purification after isolation by agarose (0.7% - 1.2%) gel electrophoresis and gel excision. All restriction enzymes used for DNA digestion were purchased from New England Biolabs (NEB). The pRSET-EmGFP vector (Invitrogen, Carlsbad, CA) was given to our lab by Dr. Andrew Feig at Wayne State University. The restriction enzymes BamHI (5' end) and EcoRI (3' end) were used to produce "sticky-end" DNA fragments with complementary overhangs for the pRSET-EmGFP vector and the β2AR:M13 gene insert. The β2AR:M13 gene was placed at the 5' end of EmGFP in the multiple cloning site of pRSET. To create a control expression vector, we also prepared the β2AR:M13 gene with a stop codon just before the EcoRI restriction site to prevent the transcription/translation of the C-terminal EmGFP fusion in the pRSET vector. DNA ligation between digested vector and insert was performed with T4 DNA Ligase (NEB, Ipswich, MA). To propagate pRSET- β2AR:M13-EmGFP, ligation reaction mixes were used to transform NovaBlue competent cells (Millipore Sigma, St. Louis, MO) which were plated onto LB agar plates with 50 μg/mL of ampicillin (Millipore Sigma, St. Louis, MO) for antibiotic selection of colonies that carry the successfully ligated DNA. After colony formation, single colonies were isolated and grown in starter cultures containing 5 mL of LB (lysogeny broth) media (ThermoFisher Scientific, Hampton, NH) and 50 μg/mL of ampicillin. The starter cultures were placed in incubator shakers overnight at 250 rpm at 37°C. Subsequently, the overnight starter cultures were centrifuged at 9,000 g for 10 minutes to form a cell pellet. QIAprep Spin Miniprep kits (Qiagen, Germantown, MD) were used to isolate and purify propagated DNA from the bacterial cell pellet. Miniprep samples were always verified independently by DNA sequencing at the Molecular Resource Center (University of Tennessee Health Sciences Center) and gel

electrophoresis with samples from PCR test of Miniprep template with primers matching the 5' and 3' ends of the gene insert.

Trial expressions of the M13-EmGFP mutant were unsuccessful in the pRSET vector. This could have been due to the pUC origin of pRSET, which allows for high copy replication/growth in *E. coli* hosts.[150] Generally, plasmids with high copy numbers are unsuitable for expressing recombinant proteins with potential toxicity to the host (misfolding/aggregation potential). After sub-cloning into the widely used pET-28a expression vector, very low expression levels were observed again. Ultimately, sub-cloning the β2AR:M13-EmGFP gene into the pE-SUMOpro vector (Life Sensors, Malvern, PA) led to much improved protein expression.

Prior to sub-cloning out of the pET-28a vector, mutagenesis experiments using the QuikChange Site-Directed Mutagenesis kit (Agilent, Santa Clara, CA) were performed to insert a linker sequence between β2AR:M13 and EmGFP. The linker sequence (LINDPPVAT) had been published in a previous study using WT β2AR fusions with C-terminal fluorescent proteins that were functional.[107] From this point forward, the constructs labeled β2AR:M13-EmGFP contain the linker sequence.

The restriction enzymes BsmBI (5' end) and XhoI (3' end) were used to digest the gene inserts (β2AR:M13-EmGFP and the control with the stop codon before EmGFP). Initially, the pE-SUMO vector must be digested with BsaI (Class IIS restriction enzyme) which leaves the pE-SUMO multiple cloning site with two unique overhangs. We also digested the pE-SUMO vector with XhoI to match 3' end complements of the double digested gene inserts. The 5' ends of our inserts were originally PCR amplified to contain the BsmBI restriction site and an AGGT overhang. After BsmBI digestion, the gene inserts contained the 5' AGGT overhang that

complements the corresponding end of pE-SUMO created by BsaI digestion. The rest of the processes involved in sub-cloning the gene inserts with pE-SUMO were performed as outlined previously and all resulting constructs were verified by DNA sequencing.

2.2.3 Recombinant protein expression in *E. coli* cells

Two strains of *E. coli* that are commonly used for recombinant protein expression were tested by previous group members in preliminary trial expressions of SUMO-β2AR:M13-FLAG. OverExpress C43(DE3) cells (Lucigen, Randor, PA) displayed greater expression levels of target proteins in the soluble fraction than BL21(DE3) cells (Millipore Sigma, St. Louis, MO). C43(DE3) is a mutant strain derived from the commonly used BL21(DE3) expression strain, but C43(DE3) cells have been preferential hosts for troubleshooting poor production of "difficult-to-express" proteins[151,152] – especially membrane proteins.[153] The (DE3) suffix designates that the strains have the prophage DE3 (derived from bacteriophage λ) which contains the T7 RNA polymerase (T7 RNAP) under control of the lacUV5 promoter.[154] This designation indicates that isopropyl β-D-1-thiogalactopyranoside (IPTG; ThermoFisher Scientific, Hampton, NH) is used to induce expression of the T7 RNAP. Recombinant genes will be expressed by T7 RNAP when they are sub-cloned downstream of the T7 promoter in an appropriate vector/plasmid. Previously, low levels of SUMO-β2AR:M13-FLAG expression were observed with traditional bacterial culture conditions: growth in LB media at 37°C for 6 hours. The following section details the optimized expression conditions developed by previous lab members.[155]

2.2.3.1 Previous optimization of SUMO-β2AR:M13-FLAG overexpression in the soluble fraction

OverExpress C43(DE3) cells were transformed with 100-200 ng of SUMO-β2AR:M13-FLAG DNA and plated onto LB agar plates supplemented with 50 μg/mL of ampicillin for

antibiotic selection. Starter cultures with 5 mL of LB media supplemented with 50 μg/mL ampicillin were inoculated with a single colony isolated from the transformation plates. The starter cultures were grown in an incubator shaker at 37°C (250 rpm) overnight. The next day, the starter culture was used to inoculate a flask containing 100 mL of TB (Terrific Broth) media (ThermoFisher Scientific, Hampton, NH) enriched with 1% glucose, 1 mM NaCl, and 2 mM $MgSO_4$ (TBE) and placed in the incubator shaker at 37°C (250 rpm) until it reached an optical density (at 600 nm, OD_{600}) of 0.6 – 1.2 A.U. All bacterial cell culture OD_{600} measurements were performed using an Agilent 8453 UV-Vis Spectrophotometer. Culture aliquots (1 mL) were taken to represent the 0-hour pre-induction samples for analysis. After reaching an appropriate OD_{600} value, the culture flask was cooled in an ice-water bath for 15 minutes before inducing protein expression with 0.1 mM IPTG and moving the culture flask to an 18°C (250 rpm) incubator shaker. After 12 hours post-induction, the culture flask was re-enriched with 1% glucose, 1 mM NaCl, and 2 mM $MgSO_4$ and continued growing until 22 hours post-induction with culture samples (normalized to 0-hour culture density and 1 mL sample volume) taken at various time points by centrifugation at 6,000 g for 5 minutes.

Cell samples were resuspended in lysis buffer (20 mM Na_2PO_4, 500 mM NaCl, 40 mM imidazole, pH 7.4) and sonicated using a Biologics Ultrasonic Homogenizer 150 V/T at 30% maximum power with 50 % pulse on ice. Aliquots of the whole-cell lysate samples (W) are taken to proceed to SDS-PAGE analysis (section 2.2.5). The rest of the whole-cell lysate was centrifuged at 10,000 g for 10 minutes and the soluble fraction (S) is taken from the supernatant for analysis as well.

2.2.3.2 Trial expression protocols with SUMO-β2AR:M13-EmGFP

For the initial trial expressions, OverExpress C43(DE3) cells were transformed with 100-200 ng of SUMO-β2AR:M13-EmGFP and SUMO-β2AR:M13 DNA and plated onto LB agar plates supplemented with 50 μg/mL of ampicillin for antibiotic selection. Two starter cultures with 5 mL of LB media supplemented with 50 μg/mL ampicillin were inoculated with a single colony from the SUMO-β2AR:M13-EmGFP and SUMO-β2AR:M13 transformation plates. The (2) starter cultures were grown in an incubator shaker at 37°C (250 rpm) overnight. The next day, 2.5 mL of each starter culture was used to inoculate 2 culture flasks (4 cultures total) containing 50 mL TBE media and returned to the incubator shaker for growth at 37°C (250 rpm) until cultures reached an OD_{600} of 0.6 – 0.7 A.U. One pair of cultures expressing SUMO-β2AR:M13-EmGFP and SUMO-β2AR:M13 were cooled in an ice-water bath for 15 minutes before inducing protein expression with 0.1 mM IPTG and moving the culture flask to an 18°C (250 rpm) incubator shaker. After 12 hours post-induction, the culture flask was re-enriched with 1% glucose, 1 mM NaCl, and 2 mM $MgSO_4$ and continued growing for another 10 hours (22 hours post-induction total). The other pair of cultures expressing the same two constructs were kept at 37°C (250 rpm), induced with 0.1 mM IPTG, and continued growing for another 6 hours. For whole-cell fluorescence and SDS-PAGE analysis, culture samples were taken at 0, 3, 4, 5 and 6-hour time points for trial expressions at 37°C and at 0, 12, 18, 20, and 22-hour time points (all normalized to the 0-hour culture density) for trial expressions at 18°C. Samples for analysis were centrifuged at 6,000 g for 5 minutes, decanted, and stored at -20°C until analysis.

For the trial expression with glucose variation, the above procedures for expression at 18°C were repeated with the SUMO-β2AR:M13-EmGFP construct in TBE media with either 0, 0.5, 1, or 2% glucose. For all the cultures, glucose was excluded in the re-enrichment step at 12

hours post-induction. For the trial expression with IPTG variation, the above procedures were also repeated with the SUMO-β2AR:M13-EmGFP construct in TBE (TB media enriched with 1% glucose, 1 mM NaCl, and 2 mM $MgSO_4$). The concentration of inducer used to initiate protein expression was varied with 0.075, 0.1, 0.25, or 0.5 mM IPTG.

2.2.4 Ni-NTA purifications and GroEL removal protocols

For all Ni-NTA purifications and GroEL removal experiments, overexpression of SUMO-β2AR:M13-EmGFP, SUMO-β2AR:M13-FLAG, or SUMO-LPAR1:M15-FLAG (all SUMO constructs have an implied 6His-tag at N-terminus) was achieved using the previously optimized protocols described in section 2.2.3.1, except with 250 mL culture volumes. Cells were harvested from 250 mL cultures by centrifugation (6,000 g for 30 minutes at 4°C) and the cell pellets were stored in the -80°C freezer until lysis and purification/GroEL removal. All buffers used in these experiments are listed in Table 2.2. GroEL removal experiments II-IX were performed with Ni-NTA purifications in a standard column format. The amount of Ni-NTA slurry (Qiagen, Germantown, MD) used for each of these experiments was 3.5 mL (1.75 mL column volume). GroEL removal experiment I was performed with smaller amounts of the Ni-NTA slurry (1 mL) and clarified cell lysate (5 mL) in 15 mL conical centrifuge tubes. Instead of applying buffers to a column and collecting eluate, buffers were applied and gently mixed in

Table 2.2: Buffers used in Ni-NTA purifications and GroEL removal (GR) experiments. All lysis buffers contained 0.5 mL protease inhibitor (PI) cocktail (Sigma-Aldrich, St. Louis, MO) and 3.5 µL benzonase nuclease (Sigma-Aldrich). TBS3 = 50 mM Tris, 300 mM NaCl, 40 mM imidazole. HMK3 = 20 mM HEPES, 20 mM $MgCl_2$, 300 mM KCl, 40 mM imidazole.

GR Exp.	Lysis buffer	Wash buffer	GR buffer	Elution buffer
I	50 mM KH_2PO_4, 150 mM NaCl, 5% glycerol, 5 mM imidazole, pH 7.9	50 mM KH_2PO_4, 150 mM NaCl, 10 mM imidazole, pH 7.9	20 mM HEPES, 150 mM KCl, 10 mM $MgCl_2$, (with or without) 5 mM ATP, pH 7.4	50 mM Tris, 150 mM NaCl, 5% glycerol, 250 mM imidazole, pH 7.4
II	HMK3, 10 mM ATP, 0.25 mg/mL casein, pH 7.8	HMK3, 10 mM ATP, pH 7.8	HMK3, 10 mM ATP, [(1) 0.25, (2) 1.0, (3) 2.5 mg/mL casein], pH 7.8	HMK3, 300 mM imidazole, pH 7.8
III	HMK3, 10 mM ATP, 5 mg/mL casein, pH 8	HMK3, 10 mM ATP, pH 8	HMK3, 10 mM ATP, 5 mg/mL casein, pH 8	HMK3, 300 mM imidazole, pH 8
IV	HMK3, 5 mM ATP, 1.5 mg/mL casein, 5% glycerol, pH 8	HMK3, 5 mM ATP, 5% glycerol, pH 8	HMK3, 5 mM ATP, 1.5 mg/mL casein, 5% glycerol, pH 8	HMK3, 300 mM imidazole, pH 8
V	HMK3, 5 mM ATP, 1.5 mg/mL casein, 5% glycerol, pH 8	HMK3, 5 mM ATP, 10% glycerol, pH 8	HMK3, 5 mM ATP, 1.5 mg/mL casein, 10% glycerol, pH 8	HMK3, 300 mM imidazole, pH 8
VI	TBS3, 0.01% glycerol, 0.001% SDS, pH 7.4	TBS3, 10% glycerol, pH 7.4	TBS3, 0.1% glycerol, [(1) 0.001%, (2) 0.01%, (3) 0.01% SDS], pH 7.4	TBS3, 300 mM imidazole, pH 7.4
VII	TBS3, 10% glycerol, 0.001% SDS, pH 7.5	TBS3, 10% glycerol, pH 7.5	TBS3, 10% glycerol, [(1) 0.001%, (2) 0.01%, (3) 0.1% SDS], pH 7.5	TBS3, 10% glycerol, 300 mM imidazole pH 7.5
VIII	TBS3, 10% glycerol, 2 mM CHAPS, pH 7.5	TBS3, 10% glycerol, 4 mM CHAPS, pH 7.5	TBS3, 10% glycerol, [(FT2) 4 mM, (FT3) 8 mM CHAPS], pH 7.5	TBS3, 10% glycerol, 400 mM imidazole pH 7.5
IX	TBS3, 10% glycerol, 2 mM CHAPS, pH 7.5	TBS3, 10% glycerol, 4 mM CHAPS, pH 7.5	TBS3, 10% glycerol, [(FT2) 4 mM, (FT3) 8 mM CHAPS], pH 7.5	TBS3, 10% glycerol, 400 mM imidazole pH 7.5

tubes containing the Ni-NTA slurry and bound proteins. To remove and collect samples, tubes were centrifuged at 500 g for 3 minutes and the buffers were pipetted off carefully. Due to the tendency of GroEL to release and re-bind to certain substrate proteins, the column format was preferred to achieve GroEL removal. The constant stream of mobile phase passing over bound proteins on the stationary phase allows for GroEL separation while decreasing transient interactions leading to re-binding events. For all experiments, the Ni-NTA slurry was equilibrated with lysis buffer (without protease inhibitor or nuclease) prior to each experiment.

Cells were resuspended in lysis buffer (Table 2.2), sonicated at 40% power, 70% pulse on ice, and centrifuged at 10,000 g for 30 minutes. The supernatant was clarified using a syringe with a 0.22 μm filter. Subsequently, the clarified lysate was mixed with Ni-NTA slurry during batch binding on ice (1 hour). After batch binding, the lysate/Ni-NTA slurry mix was poured into the column to collect the flow-through (FT) samples, apply wash buffers and GroEL removal (GR) buffers. Elution buffer was applied in the last step and collected in 2-3 fractions (E1, E2, E3) with volumes equivalent to the column volume (1.75 mL). Aliquots of all samples collected throughout the purification were prepared for SDS-PAGE analysis.

2.2.5 SDS-PAGE and western blotting protocols

Sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE) experiments were performed in a Mini-Protean Tetra cell (Bio-Rad, Hercules, CA) using Any kD Mini-Protean TGX Pre-cast Protein Gels (Bio-Rad, Hercules, CA) with Tris/glycine/SDS running buffer (25 mM Tris-HCl, 192 mM glycine, 0.1% SDS). SDS-PAGE samples were mixed 1:1 with 2X Laemmli sample buffer (65.8 mM Tris-HCl, pH 6.8, 2.1% SDS, 26.3% glycerol, 0.01% bromophenol blue, 355 mM β-mercaptoethanol; Bio-Rad, Hercules, CA) before pipetting up to 15 μL into sample wells. A molecular weight (MW) standard, Precision Plus Protein

Kaleidoscope Pre-Stained Protein Standard (Bio-Rad, Hercules, CA), was included on each gel. In addition to the multi-colored standards of a wide MW range (250 – 10 kD), this MW standard also contained the 25 and 75 kD fluorescent markers that were visualized with in-gel fluorescence assays. After sample loading, gels were run for 35 – 40 minutes at constant voltage (200 V; Bio-Rad PowerPac power supply) and rinsed 3 times for 5 minutes with DI H_2O after completion. To visualize proteins, gels were stained overnight in Bio-Safe Coomassie Brilliant Blue G-250 (Bio-Rad, Hercules, CA) and destained in DI H_2O for a minimum of one hour before imaging using a Fotodyne station (using FOTO/Analyst PC Image Version 5.00 software; Fotodyne Inc., Hartland, WI) equipped with a Coomassie/Methylene blue filter. To visualize proteins with in-gel fluorescence, gels were rinsed 3 times for 5 minutes with DI H_2O and directly imaged using a Fotodyne station equipped with a fluorescent green filter.

All western blotting (WB) experiments were performed using wet electroblotting techniques (tank transfer) to transfer proteins from SDS-PAGE gels to a polyvinylidene difluoride (PVDF; ThermoFisher Scientific, Hampton, NH) membrane in WB transfer buffer (25 mM Tris-HCl pH 8.3, 192 mM glycine, 20% (v/v) methanol). Transfer cassette components (filter paper and fiber pads), polyacrylamide gel, and the PVDF membrane were equilibrated in WB transfer buffer (after pre-soak of PVDF membrane in methanol for ~3 seconds) before assembling the gel/membrane sandwich. Protein transfer was performed under constant voltage (100V) with an ice pack and rotating stir bar for one hour. After rinsing the membrane with tris buffered saline and Tween 20 (TBST: 50 mM Tris-HCl pH 7.5, 150 mM NaCl, 0.05% (v/v) Tween 20), it was incubated in blocking solution (TBST with 5% (w/v) non-fat dry milk) overnight to prevent non-specific binding.

After blocking overnight, the membrane was incubated with primary antibody (1° Ab) diluted in TBST with 3% (w/v) non-fat dry milk for 1.5 hours. The following dilutions were used for each primary antibody incubation step: SUMO 1° Ab, 1: 10,000 (Abcam, Cambridge, U.K.); FLAG 1° Ab, 1: 2,000 (Sigma-Aldrich, St. Louis, MO); GroEL 1° Ab, 1: 80,000 (Sigma-Aldrich, St. Louis, MO). After three TBST wash steps, the membrane was incubated with corresponding secondary antibody (2° Ab) diluted in TBST with 3% (w/v) non-fat dry milk for one hour. The following dilutions were used for each secondary antibody incubation step: SUMO 2° Ab dilution: 1: 20,000 (rabbit anti-chicken IgY; ThermoFisher Scientific, Hampton, NH); FLAG 2° Ab, 1: 2,000 (rabbit anti-mouse IgG; Sigma-Aldrich, St. Louis, MO); GroEL 2° Ab, 1: 40,000 (goat anti-rabbit IgG; Abcam, Cambridge, U.K.). For visualization by luminescence, each secondary antibody was provided as a horse radish peroxidase (HRP) conjugated antibody. After all antibody incubations and wash steps, the membrane was incubated with HRP substrate (50:50 mix of luminol and peroxide solution) for 5 minutes before exposure (30 sec. – 1 min.) and imaging with a Fotodyne station.

2.3 Results and Discussion

2.3.1 Designing transferable, water-soluble mutation sets for GPCR

Twenty-three sets of fifteen GPCR point mutants were created based on the crystal structures listed in Table 2.3 using the MOE modeling software package. Glutamate, aspartate, and serine mutations were made at each of the outward-facing residues (listed in Tables 2.3 and 2.4) and the hydrophobic solvent-accessible surface area (SASA) was calculated for each wild-type GPCR structure and point mutant. In most cases, we observed that glutamate mutations resulted in a greater overall decrease in hydrophobic SASA relative to the wild-type GPCR. In addition to glutamate, serine was incorporated into the mutation set design when residue

positions i and i+4 in TM α-helices were identified as commonly outward-facing. This alteration was to avoid any destabilizing effects related to sidechain charge repulsions of proximal glutamates. Therefore, mutation sets containing glutamate and serine were expected to be more effective in conferring protein solubility in aqueous solution while avoiding helix destabilization from sidechain charge repulsion. The final WS design consisted of 9 glutamate and 6 serine mutations (M15) at the 15 outward-facing residues listed in Tables 4-5.

Table 2.3: Twenty-three GPCR crystal structures used for solvent-accessible surface analysis.
Overview of MOE solvent-accessible surface analysis (SASA) results for the wild-type GPCR and their respective mutants to determine optimal mutations at outward-facing residues. Under each E, D, S mutation is the number of GPCR that displayed the greatest decrease in hydrophobic SASA relative to the wild-type GPCR.

GPCR Crystal Structure	**PDB ID:**	**GPCR Crystal Structure**	**PDB ID:**	**Mutation Sites**	**GLU (E)**	**ASP (D)**	**SER (S)**
RHO	(1U19,)[156]	NOP	(4EA3)[157]	1.37	18	4	1
CXCR1	(2LNL)[158]	NTR1	(4GRV)[159]	1.41	0	1	22
β1AR	(2VT4)[160]	5HT1B	(4IAR)[161]	1.44	15	6	2
β2AR	(4LDO)[162]	5HT2B	(4IB4)[163]	1.48	2	0	21
CXCR4	(3OEO)[164]	SMO	(4JKV)[165]	1.55	15	3	5
DRD3	(3PBL)[166]	CRFR1	(4KY5)[167]	1.58	5	0	18
H1R	(3RZE)[26]	GLR	(4L6R)[168]	3.23	15	8	0
ACM2	(3UON)[27]	CCR5	(4MBS)[169]	4.44	15	6	2
S1PR1	(3V2Y)[170]			4.47	2	0	21
PAR1	(3VW7)[171]			4.51	14	5	4
ACM3	(4DAJ)[28]			4.55	2	2	19
AA2AR	(3EML)[25]			5.49	18	4	1
OPRD	(4EJ4)[172]			5.52	0	1	22
OPRK	(4DJH)[173]			5.56	16	6	1
OPRM	(4DKL)[101]			6.53	17	5	1

A previously designed WS mutation set (M13) applied to β2AR featured 13 mutations of outward-facing residues to glutamine. A comparison of the targeted residues in the M13/M15 designs shows that only 5 outward facing residues are shared between the mutation sets. Since the M15 design was based on a larger data set of known GPCR crystal structures than the M13 design, it was expected that the M15 mutation set would have wider applications among structurally uncharacterized GPCR.

Table 2.4: The M13 and M15 mutant designs.
The M15 (M9E6S) mutation set is based on common outward-facing residues among 23 different GPCR crystal structures. The M13 (M13Q) mutation set based on 7 different GPCR crystal structures was designed by previous lab members.

TM Helix: Index	Positions of Outward-facing Residues from 23 GPCR Set	M15 Water-soluble Mutations	M13 Water-soluble Mutations
1: N1.50	1.37 (18), 1.41 (18), 1.44 (21), 1.48 (20), 1.55 (14), 1.58 (14)	1.37 **E**, 1.41 **S**, 1.44 **E**, 1.48 **S**, 1.55 **E**, 1.58 **S**	1.41 **Q**, 1.43 **Q**, 1.51 **Q**, 1.58 **Q**
2: D2.50	N/A	N/A	2.63 **Q**
3: R3.50	3.23 (20)	3.23 **E**	N/A
4: W4.50	4.44 (18), 4.47 (21), 4.51 (21), 4.55 (20)	4.44 **E**, 4.47 **S**, 4.51 **E**, 4.55 **S**	4.51 **Q**, 4.55 **Q**
5: P5.50	5.49 (18), 5.52 (21), 5.56 (20)	5.49 **E**, 5.52 **S**, 5.56 **E**	5.40 **Q**
6: P6.50	6.53 (19)	6.53 **E**	6.46 **Q**, 6.53 **Q**
7: P7.50	N/A	N/A	7.34 **Q**, 7.37 **Q**, 7.51 **Q**

2.3.2 Optimizing SUMO-β2AR:M13 expression in *E. coli* using an Emerald Green Fluorescent Protein (EmGFP) fusion partner

After unsuccessful recombinant protein expression in the pRSET and pET-28a vectors (data not shown), β2AR:M13-EmGFP was sub-cloned into the pE-SUMO vector (Life Sensors,

Malvern, PA) which had been the most successful bacterial expression vector for the β2AR:M13 constructs in our lab at the time. The result is an N-terminal SUMO (Small Ubiquitin-like Modifier) fusion protein with an N-terminal 6x-histadine tag (6His-tag) for purification via immobilized metal affinity chromatography (IMAC). Figure 2.3 shows the whole-cell fluorescence data reflecting recombinant protein expression over time. The SUMO-β2AR:M13

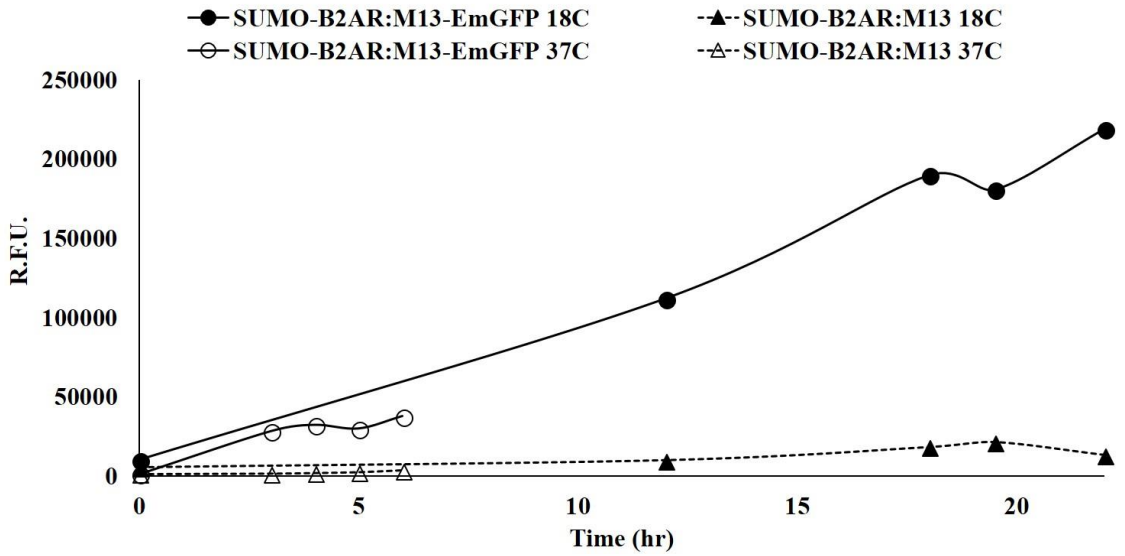

Figure 2.3: Whole-cell fluorescence during trial expressions at two temperatures. R.F.U. stands for relative fluorescence units. SUMO-β2AR:M13 with C-terminal EmGFP (solid lines, circles) and without EmGFP (dashed lines, triangles) in C43(DE3) cells at 18°C (solid symbols) and 37°C (unfilled symbols).

construct serves as a control to account for any background fluorescence signals from bacterial cells. Given that both sets of control samples displayed very low relative fluorescence, it is likely that increased expression of SUMO-β2AR:M13-EmGFP is the source of increased fluorescence over time.

To verify the presence and localization of SUMO fusion proteins, SDS-PAGE and western blot analysis with an Anti-SUMO antibody (Life Sensors, Malvern, PA) was performed on the time course samples (Figure 2.4 A, C). Additionally, in-gel fluorescence with SDS-PAGE

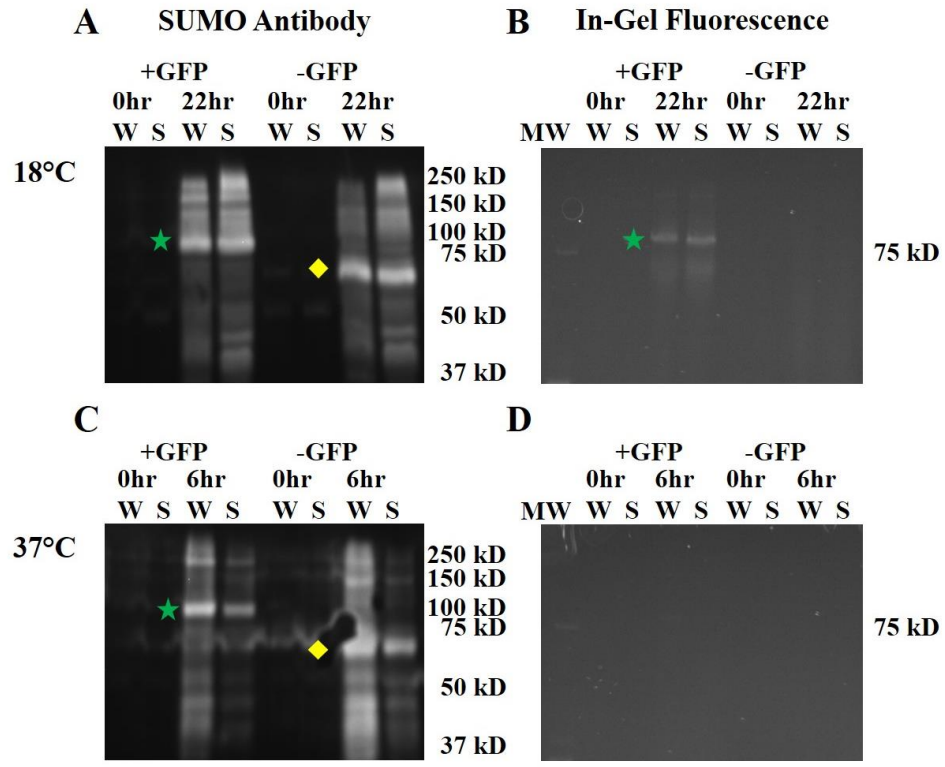

Figure 2.4: SDS-PAGE and western blot analysis of SUMO-β2AR:M13-EmGFP (+GFP) and SUMO-β2AR:M13 (-GFP) expressed in C43(DE3) cells cultured at 18°C (A, B) and 37°C (C, D).
The whole-cell lysate sample and soluble fraction after centrifugation are represented by W and S lanes, respectively. MW represents the Kaleidoscope protein standard, which has a fluorescent 75 kD band. The samples represented in the 0-hour and 6/22-hour lanes were taken prior to inducing protein expression and 6 or 22 hours after induction, respectively. (A, C) Western Blot analysis using SUMO primary antibody to detect SUMO-β2AR:M13-EmGFP (87 kD, green star) and SUMO-β2AR:M13 (~60 kD, yellow diamond). (B, D) SDS-PAGE and in-gel fluorescence analysis to detect properly-folded SUMO-β2AR:M13-EmGFP construct.

was used to identify properly-folded EmGFP fusion proteins (Figure 2.4 B, D). The relative band intensities in the soluble versus whole cell fraction for both constructs appear to be greater at lower temperatures (18°C), rather than the optimal temperature for E. coli growth (37°C). This observation correlates with the in-gel fluorescence analysis showing fluorescent bands only for the 18°C samples. There are many cases in the literature that demonstrate lower temperature growth conditions can improve solubility and stability during expression of some proteins.[174,175] It is believed that the lower temperatures can improve protein expression in part by dampening

the rate of protein synthesis which increases the amount of time for correct protein folding and decreases the potential for misfolding or aggregation events. The results shown in Figure 2.4 support the theory that C-terminal GFP can act as a fluorescent reporter for expressing properly-folded protein in bacterial cells. This provided our lab with an efficient screening tool for optimizing recombinant protein expression.

We tested some of the variables that had been optimized for trial expressions of SUMO-β2AR:M13-FLAG and summarized the results in Figures 2.5 and 2.6. The optimized protocol for SUMO-β2AR:M13-FLAG expression was developed by previous lab members (Dr. Kikonyogo and Dr. Gacasan) and is detailed in the Methods section. The protocol involved using Terrific Broth (TB) media enriched with 2mM $MgSO_4$, 1 mM NaCl, and 1% glucose (TBE) at inoculation, re-enrichment with the same concentrations of $MgSO_4$, NaCl, and glucose at 12 hours (TBE+12), and growth for 22 hours after induction with 0.1 mM IPTG. Figure 2.5 shows the fluorescence data from independent trial expressions varying glucose content in the starting TBE media (with no glucose added at re-enrichment) or IPTG inducer concentration for protein expression. These data suggest that the optimized protocol for expressing SUMO-β2AR:M13-FLAG could be improved by increasing the concentration of inducer, IPTG, from 0.1 to 0.25

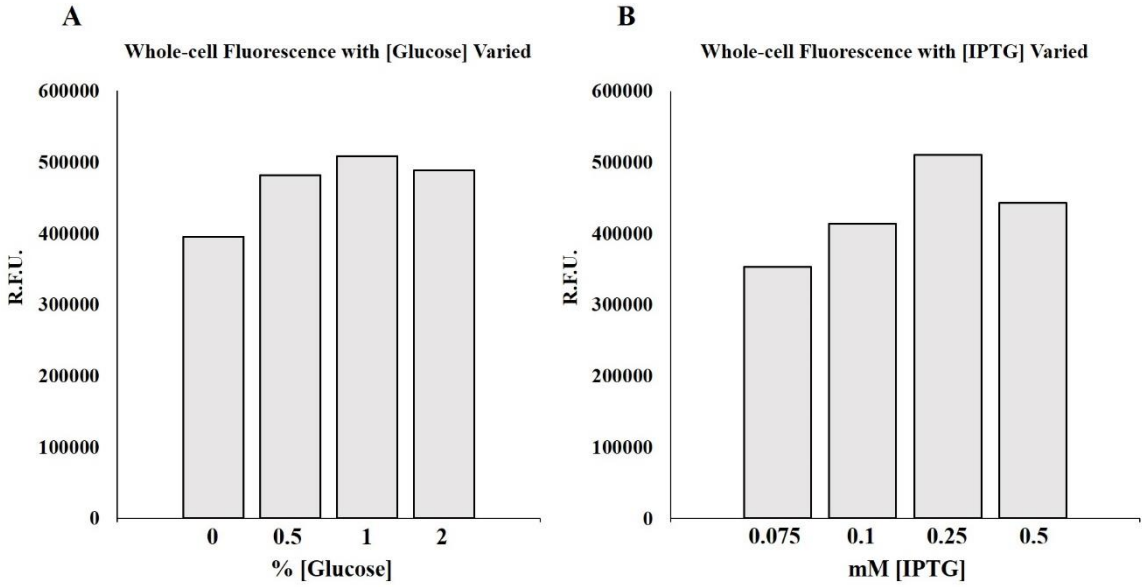

Figure 2.5: Normalized whole-cell fluorescence data from independent trial expressions of SUMO-β2AR:M13-EmGFP in C43(DE3) cells.
Cells were cultured in enriched Terrific Broth (TBE = TB enriched with 2mM $MgSO_4$, 1 mM NaCl, 1% glucose concentration – or varied in panel A) at 18°C after inducing protein expression with 0.1 mM isopropyl β-D-1-thiogalactopyranoside (IPTG), or varied IPTG concentrations in panel B. Fluorescence data above reflect cell samples at 22 hours post-induction normalized to the 0-hour cell culture density. (A) Initial glucose concentration of TBE culture media was varied from 0 – 2%, and expression. (B) IPTG concentration used to induce protein expression was varied from 0.075 – 0.5 mM.

mM (Figure 2.5 B). In terms of glucose concentration, no further improvement to the optimized protocol could be observed since 1% glucose was associated with the highest relative fluorescence (Figure 2.5 A). The culture media conditions were further investigated and summarized in Figure 2.6. The media conditions used in the optimized protocol produce relatively the same amount of fluorescence as those same conditions without the 12-hour enrichment step at 22 hours post-induction. Additionally, relative fluorescence increased substantially when culture growth was extended to 26 hours in TB, TBE, and TBG (5, 6, 7 in Figure 2.6) before harvesting cells.

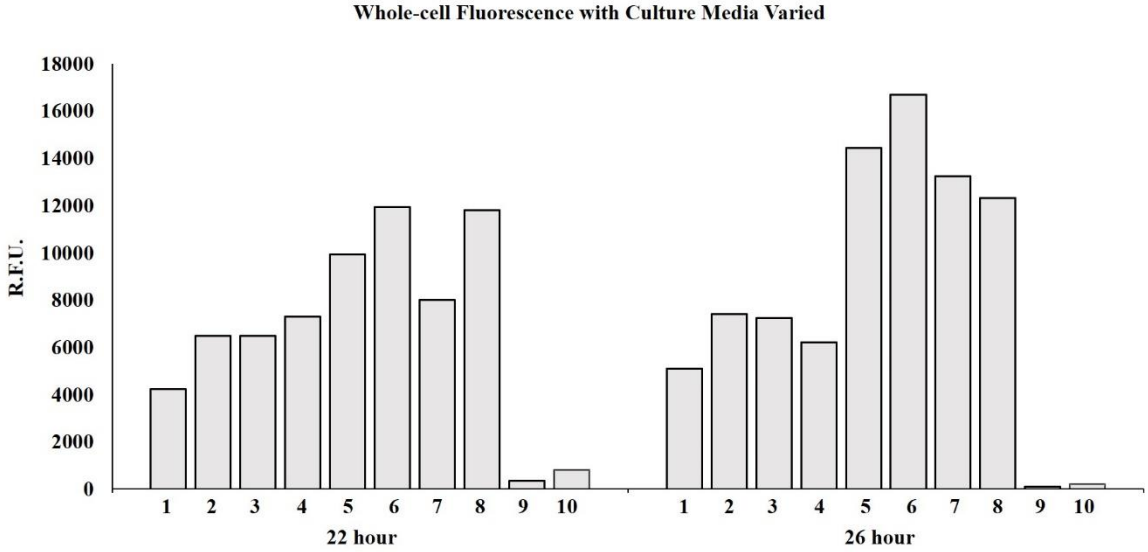

Figure 2.6: Normalized whole-cell fluorescence data from trial expressions of SUMO-β2AR:M13-EmGFP in C43(DE3) cells cultured at 18°C in LB/TB media variations. 1: LB = Lysogeny broth, 2: LBE = LB enriched (2mM $MgSO_4$, 1 mM NaCl, 1% glucose), 3: LBG = LB enriched (2 mM $MgSO_4$, 1 mM NaCl, 0.5% glycerol), 4: LBE+12 = LB enriched + 12 hr re-enrichment, 5: TB = Terrific broth, 6: TBE = TB enriched (2mM $MgSO_4$, 1 mM NaCl, 1% glucose), 7: TBG = TB enriched (2 mM $MgSO_4$, 1 mM NaCl, 0.5% glycerol), 8: TBE+12 = TB enriched + 12 hr re-enrichment, 9-10: Plasmid_LBE/TBE+12 = Empty pE-SUMO vector in both LBE/TBE+12 growth conditions to serve as negative controls. For all cultures, protein expression was induced with 0.1 mM IPTG.

Therefore, the following protocol adjustments were suggested to further optimize the expression of properly folded SUMO-β2AR:M13-FLAG: 1) Maintain the usage of TBE media from the previously optimized protocol without the 12-hour re-enrichment step. 2) Increase the amount of IPTG used to induce protein expression to 0.25 mM. 3) Increase the post-induction growth time from 22 to 26 hours before harvesting cells.

2.3.3 Strategies to remove GroEL complex during WS-GPCR purifications

Identification of GroEL (60 kD) in purified samples of SUMO-β2AR:M13-FLAG (60 kD) by collaborators at St. Jude prompted us to utilize an Anti-GroEL antibody (Sigma-Aldrich) as a specific probe for GroEL. Figure 2.7 below shows the results obtained by SDS-PAGE and western blot analysis of cell lysate samples from a previous trial expression of SUMO-

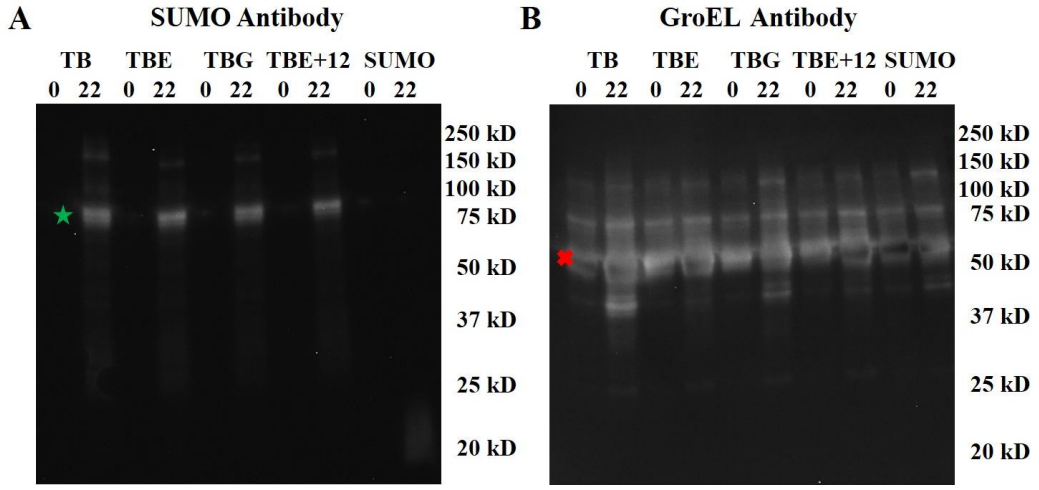

Figure 2.7: Western blot analysis on soluble fraction samples from trial expression of SUMO-β2AR:M13-EmGFP in C43(DE3) cells cultured at 18°C in TB with different supplements.
TB media abbreviations are defined in the legend of Figure 2.6. Zero represents the 0-hour culture sample taken prior to induction and 22 represents the normalized 22-hour culture sample taken post-induction. (A) SUMO primary antibody is used to detect SUMO-β2AR:M13-EmGFP (87 kD, green star) 22 hours post-induction. (B) GroEL primary antibody is used to detect GroEL (~60 kD, red x) presence during trial expression.

β2AR:M13-EmGFP (87 kD). In Figure 2.7 (B), it is evident that the antibody detects GroEL in all trial expression samples. The ubiquitous nature of GroEL should be expected as it functions as an obligate folding chaperone for many proteins that are vital for bacterial cell survival.[176] In Figure 2.7 (A), the target protein can be detected at the expected molecular weight and only in samples collected 22 hours post-induction.

The temperature conditions used during recombinant protein expression in *E. coli* does not appear to affect the relative amounts of GroEL expression (Figure 2.8 (B)). Overall, there seems to be slightly less GroEL in the 0-hour samples than in the 6-/22-hour, but no significant difference between the two temperatures. However, expression at the lower temperatures and longer growth time yields more target protein in the soluble fraction than with the higher temperatures and shorter growth time (Figure 2.8 (A)) as we have observed in previous

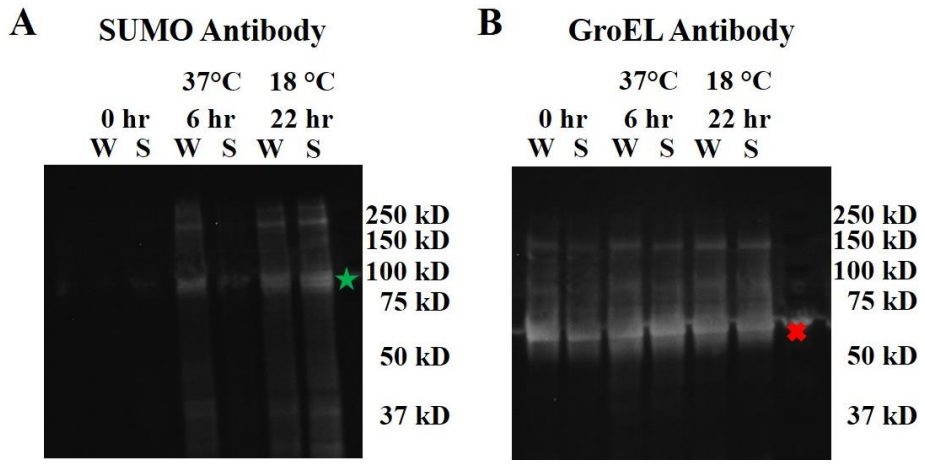

Figure 2.8: Western blot analysis of samples from trial expression of SUMO-β2AR:M13-EmGFP in C43(DE3) cells cultured at 18°C or 37°C in TBE.
(A) SUMO primary antibody is used to detect SUMO-β2AR:M13-EmGFP (87 kD, green star) in trial expression samples. (B) GroEL primary antibody is used to detect GroEL (~60 kD, red x) in trial expression samples.

experiments (Figure 2.4). Regardless, it became evident that avoiding GroEL during protein expression would necessitate changing expression systems, but even then, it is possible that orthologs of GroEL and other chaperones will be similarly problematic. Therefore, we proceeded to implement strategies for removing GroEL during protein purification as described in the introduction. The following results (Figures 2.9-2.16) are representative of the many GroEL removal trials performed during IMAC purification (Ni-NTA agarose) of either the SUMO-β2AR:M13-EmGFP or SUMO-β2AR:M13-FLAG constructs, which both have a 6His-Tag on the N-terminal end of SUMO. Data from the GroEL Removal I experiment (Figure 2.9) showed that 5 mM ATP was partially effective in promoting GroEL-substrate release. In Figure 2.9 A, the elution sample from the +ATP set (lane 8) had qualitatively less GroEL than the No-ATP set (lane 13) based on the relative band intensity.

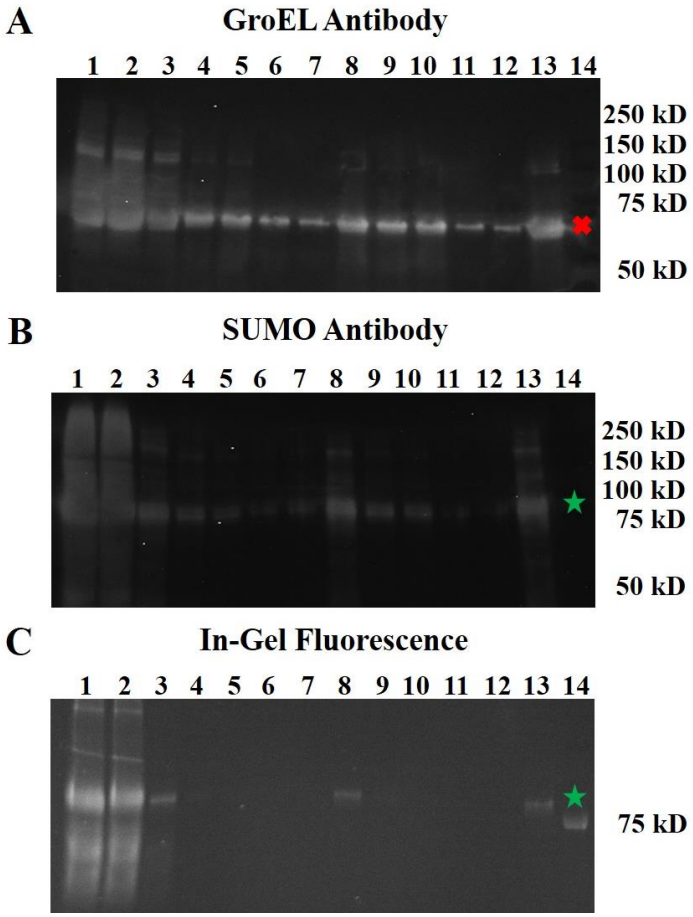

Figure 2.9: GroEL Removal I: SUMO-β2AR:M13-EmGFP Ni-NTA purification and ATP-stimulated GroEL substrate release.
Sample description: 1 = Soluble lysate, 2 = Flow-through, 3 = Wash, 4-7 = GroEL release buffer with 5mM ATP, 8 = Elution (+ATP set), 9-12 = GroEL release buffer without ATP, 13 = Elution (No-ATP set), 14 = Kaleidoscope protein standard. (A) SDS-PAGE/Western blot analysis of purification samples using GroEL antibody to detect GroEL presence (~60 kD, red x). (B) SDS-PAGE/Western blot analysis of purification samples using SUMO antibody to detect SUMO-β2AR:M13-EmGFP presence (87 kD, green star). (C) SDS-PAGE and in-gel fluorescence analysis to detect SUMO-β2AR:M13-EmGFP presence in parallel.

Since the amounts of SUMO-β2AR:M13-EmGFP in both elution samples were roughly equivalent (Figure 2.9 (B, C)), we attributed the decreased GroEL contamination in lane 8 to the presence of ATP in the GroEL release (GR) buffer. In GroEL Removal II – V experiments (Figures 2.10 – 2.13), a combination of the strategies using ATP and casein to stimulate GroEL-substrate release were implemented. Initially, we tested a protocol of increasing concentrations of casein (0.5, 1.0, 2.5 mg/mL) in the GR buffers (Figure 2.10) coupled with (10 mM) ATP

addition to all purification buffers, except elution. Given that ATP hydrolysis is a fast process, GroEL can presumably re-bind to the target protein during various steps of the purification process. Thus, our rationale behind including ATP in the lysis, wash, and GR buffers was to prevent subsequent binding events after GroEL-substrate release by constantly stimulating ATP hydrolysis. Figure 2.10 (A) shows the GroEL specific bands throughout the GR and wash samples which indicates that the chaperones are consistently being released from the target protein bound to the column. In Figure 2.10 (B), the fluorescent bands attributed to our target protein were visible in the elution fractions. Since GroEL is also present in the elution fractions, we decided to test higher concentrations of casein in the GR buffer.

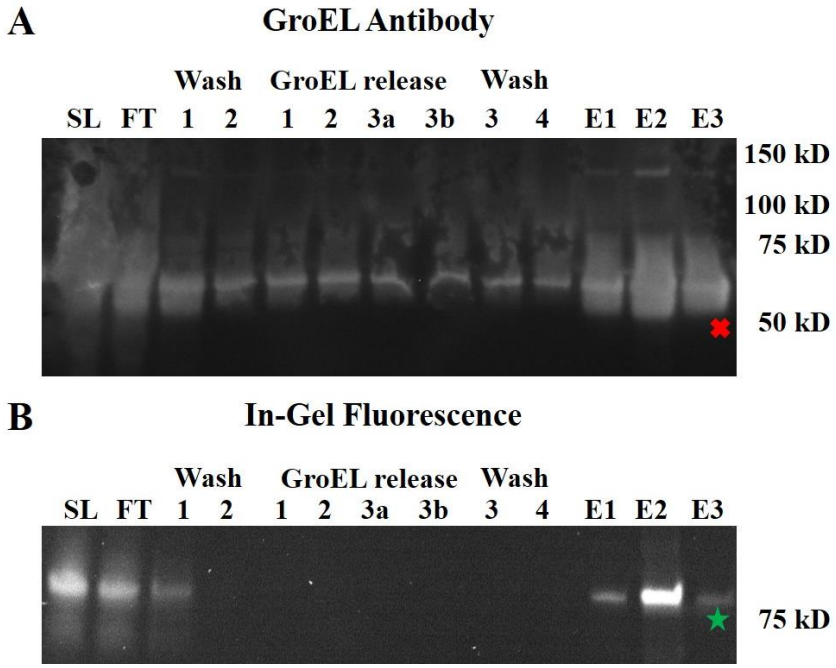

Figure 2.10: GroEL Removal II: SUMO-β2AR:M13-EmGFP Ni-NTA purification and GroEL substrate release with ATP and casein.
Sample description: SL = Soluble lysate, FT = Flow-through, Wash1-4 = Wash buffer applied twice before and after GroEL release steps contained 10 mM ATP, GroEL release1-3a/b = GroEL release buffers contained 10 mM ATP and increasing amounts of casein (30 mL each): 0.5 mg/mL (1), 1.0 mg/mL (2), 2.5 mg/mL (3a/b), E1-3 = Elution fractions. (A) SDS-PAGE/Western blot analysis of purification samples using GroEL antibody to detect GroEL presence (~60 kD, red x). (B) SDS-PAGE and in-gel fluorescence analysis to detect SUMO-β2AR:M13-EmGFP presence (87 kD, green star).

In GroEL Removal III, we increased the casein concentration to 5 mg/mL and the GR buffer volume to 150 mL total. Figure 2.11 (A) shows the Coomassie stained gel of the samples collected throughout the experiment. SUMO-β2AR:M13-EmGFP is clearly visible in the E2 lane above the 75 kD marker. Furthermore, the bands running above the 75 kD marker are visible with in-gel fluorescence in Figure 2.11 (B). However, there were still bands that likely correspond to GroEL in the elution fractions based on migration during SDS-PAGE. Since it appears that GroEL was consistently released through the GR samples 1 – 5, we drastically increased the volume of GR buffer used in GroEL Removal IV (Figure 2.12) to 1 liter. However, we also needed to decrease the casein concentration because of flow rate issues during the

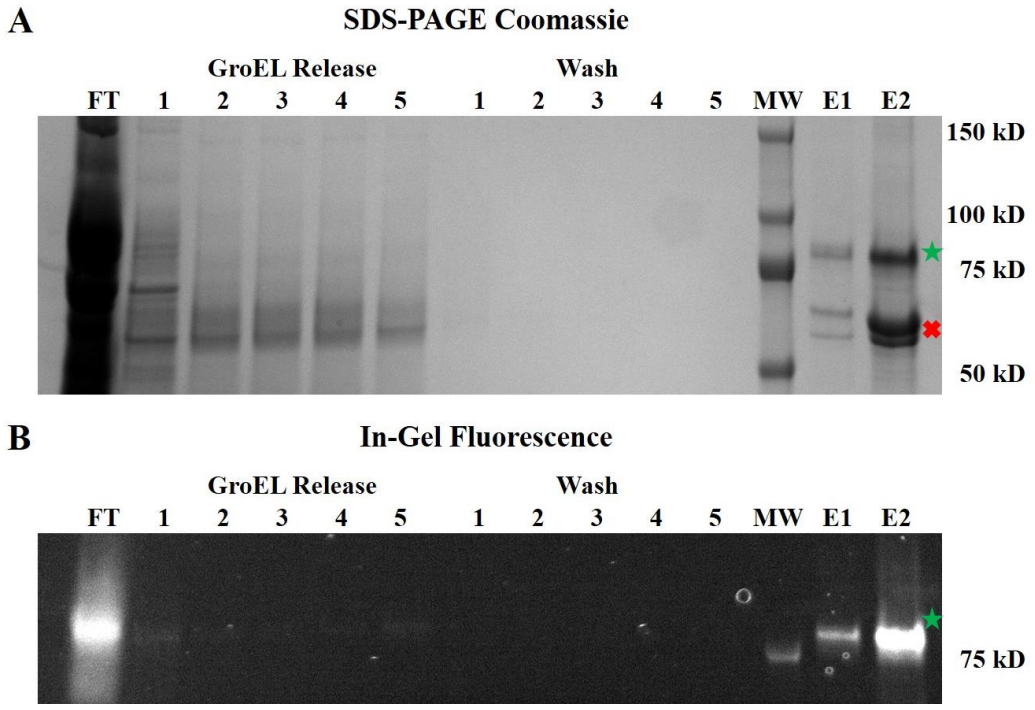

Figure 2.11: GroEL Removal III: SUMO-β2AR:M13-EmGFP Ni-NTA purification and GroEL substrate release with ATP and Casein.
Sample description: FT = Flow-through, GroEL release1-5 = GroEL release buffer contained 10 mM ATP and 5 mg/mL casein (150 mL total), Wash1-5 = Wash buffer applied after GroEL release steps, MW = Kaleidoscope protein standard contains a fluorescent 75 kD marker, E1-2 = Elution fractions. (A) SDS-PAGE analysis of purification samples using Coomassie stain to detect proteins. (B) SDS-PAGE and in-gel fluorescence analysis to detect SUMO-β2AR:M13-EmGFP presence.

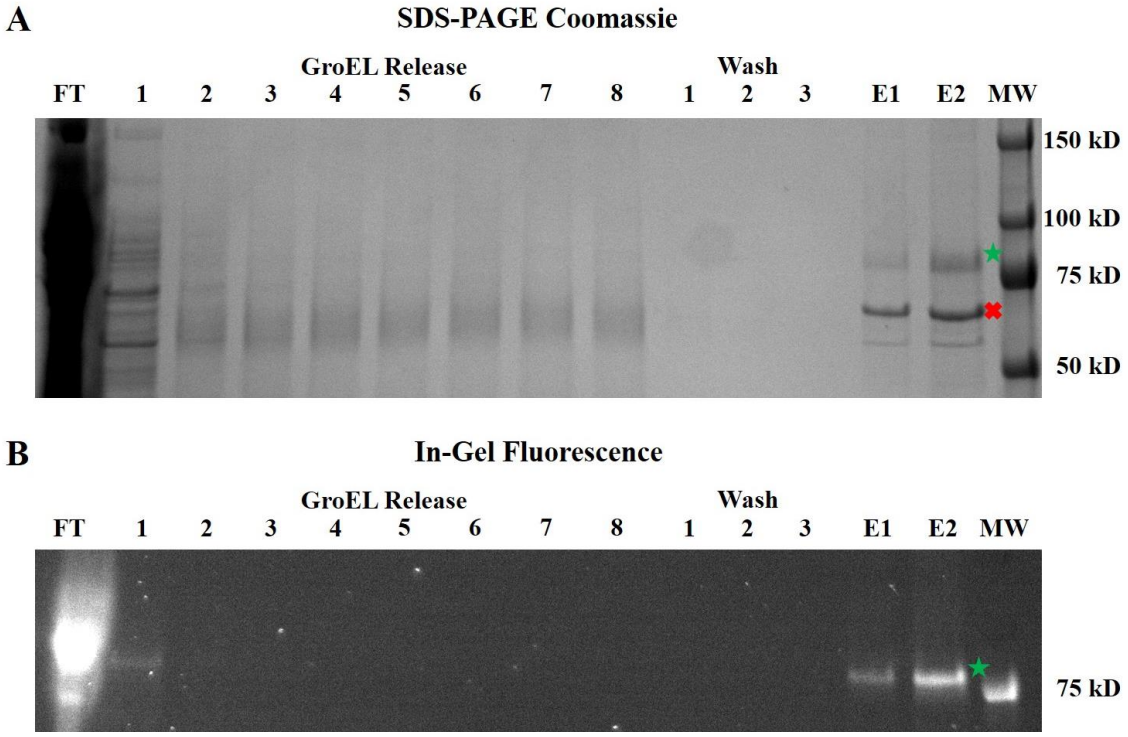

Figure 2.12: GroEL Removal IV: SUMO-β2AR:M13-EmGFP Ni-NTA purification and GroEL substrate release with ATP and Casein.
Sample description: FT = Flow-through, GroEL release1-8 = GroEL release buffer contained 5mM ATP, 5% glycerol, and 1.5 mg/mL casein (1000 mL total), Wash1-3 = Wash buffer applied after GroEL release steps, E1-2 = Elution fractions, MW = Kaleidoscope protein standard contains a fluorescent 75 kD marker. (A) SDS-PAGE analysis of purification samples using Coomassie stain to detect proteins. (B) SDS-PAGE and in-gel fluorescence analysis to detect SUMO-β2:M13-EmGFP presence.

purification with GroEL Removal III. In the elution fractions of Figure 2.12 (A), we observed the least amount of chaperone compared to previous attempts, but GroEL remained bound to our target protein despite the volume of GR buffer used. Additionally, we observed a smaller amount of our target protein (Figure 2.12 (B)) relative to previous attempts. It was suggested that we attempt the GroEL removal protocols with the construct that was originally found to have GroEL associated. The rationale here was to ensure that no aspects of the C-terminal EmGFP fusion partner were impeding the release of GroEL from our target protein. In GroEL Removal V (Figure 2.13), the same protocols and buffers used in the previous experiment were applied to the SUMO-β2AR:M13-FLAG construct. Since the target protein and GroEL have similar molecular

weights, it was necessary to use antibody detection for the two proteins. The results showed that GroEL remained bound to the target protein in the elution fractions (Figure 2.13 (A)). It is important to note that the culture volumes (250 mL) used in the purifications/GroEL removal experiments were much smaller in scale to the amounts needed for crystallization trials with St. Jude (2 liters). Therefore, even if one liter of GR buffer were sufficient to completely remove GroEL, the volume of GR buffer needed to separate chaperone from target protein would be much larger for the scale necessary for crystallization trials.

The next set of strategies we implemented involved detergent-containing buffers to remove GroEL from our target protein during Ni-NTA purification. Initially, we tried sodium

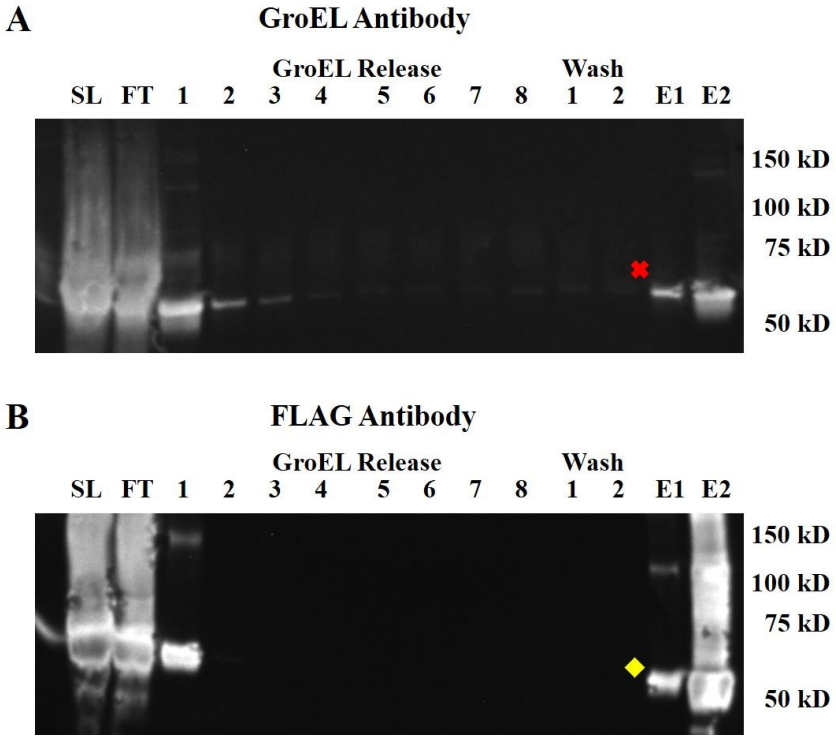

Figure 2.13: GroEL Removal V: SUMO-β2AR:M13-FLAG Ni-NTA purification and GroEL substrate release with ATP and Casein.
Sample description: SL = Soluble lysate, FT = Flow-through, GroEL release1-8 = GroEL release buffer contained 5 mM ATP, 10% glycerol, and 1.5 mg/mL casein (1000 mL total), Wash1-3 = Wash buffer applied after GroEL release steps, E1-2 = Elution fractions. (A) SDS-PAGE/Western blot analysis of purification samples using GroEL antibody. (B) SDS-PAGE/Western blot analysis of purification samples using FLAG antibody.

dodecyl sulfate (SDS) because the concentration at which it is used in sample loading buffer (1% SDS) for SDS-PAGE is effective in disrupting/denaturing the GroEL ring complexes. While not recommended, anionic detergents, such as SDS, have been used in concentrations up to 0.3% during purification according to the Ni-NTA agarose user manual (Qiagen, Germantown, MD). We initially tried very low concentrations of detergent in GroEL Removal VI (Figure 2.14), with GR buffers containing 0.001% - 0.01% SDS. The bands that were visible through in-gel fluorescence in Figure 2.14 (B) correspond with the target protein. The lower bands in the elution lanes of the Coomassie-stained gel (Figure 2.14 (A)) correspond to GroEL. The SDS concentrations used were partially effective, but not sufficient in completely removing

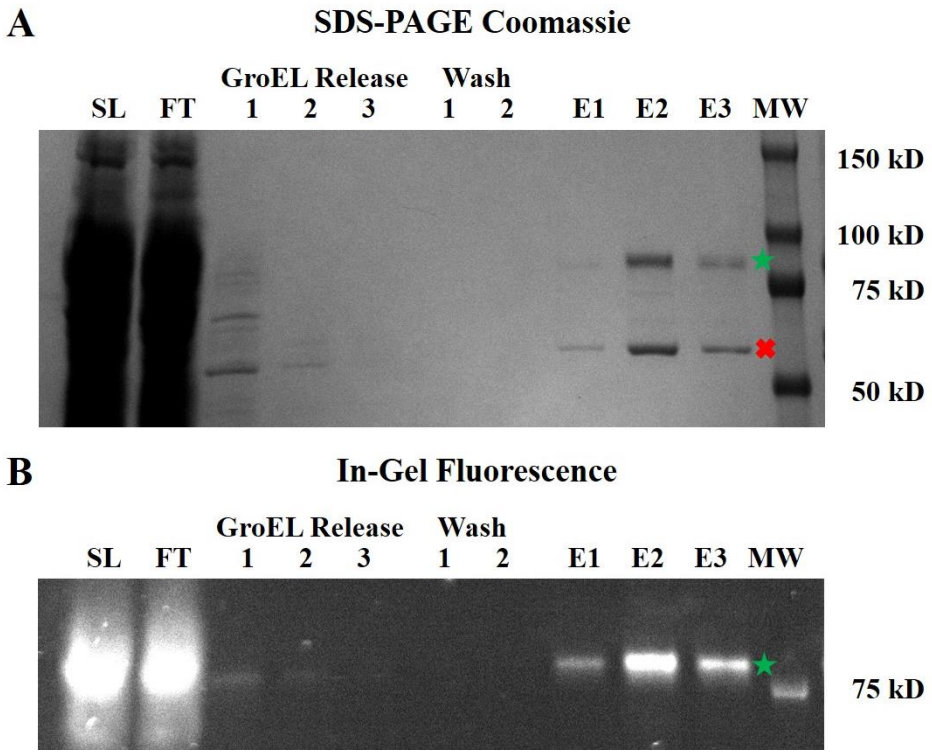

Figure 2.14: GroEL Removal VI: SUMO-β2AR:M13-EmGFP Ni-NTA purification and GroEL substrate release with SDS detergent.
Sample description: SL = Soluble lysate, FT = Flow-through, GroEL release1-3 = GroEL release buffer contained 0.001% SDS (1) and 0.01% SDS (2-3), Wash1-2 = Wash buffer applied after GroEL release steps, E1-3 = Elution fractions. (A) SDS-PAGE analysis of purification samples using Coomassie G250 for general protein visualization. (B) SDS-PAGE and in-gel fluorescence analysis to detect SUMO-β2AR:M13-EmGFP presence.

GroEL from the target protein. However, when we tried to increase the SDS concentration to 0.1% in GroEL Removal VII (Figure 2.15), GroEL and the target protein were completely removed from the column. In the Coomassie-stained gel and in-gel fluorescence assay, bands corresponding to the target protein and GroEL were absent in the expected regions of the elution lanes. These results implied that 0.1% SDS was excessively harsh and disrupted binding of target protein to the stationary phase of the column. Thus, we attempted to use a milder detergent in GroEL Removal VIII (Figure 2.16). CHAPS is a zwitterionic detergent that is often used in solubilizing membrane proteins without denaturation.[177] For this experiment, it was suggested to

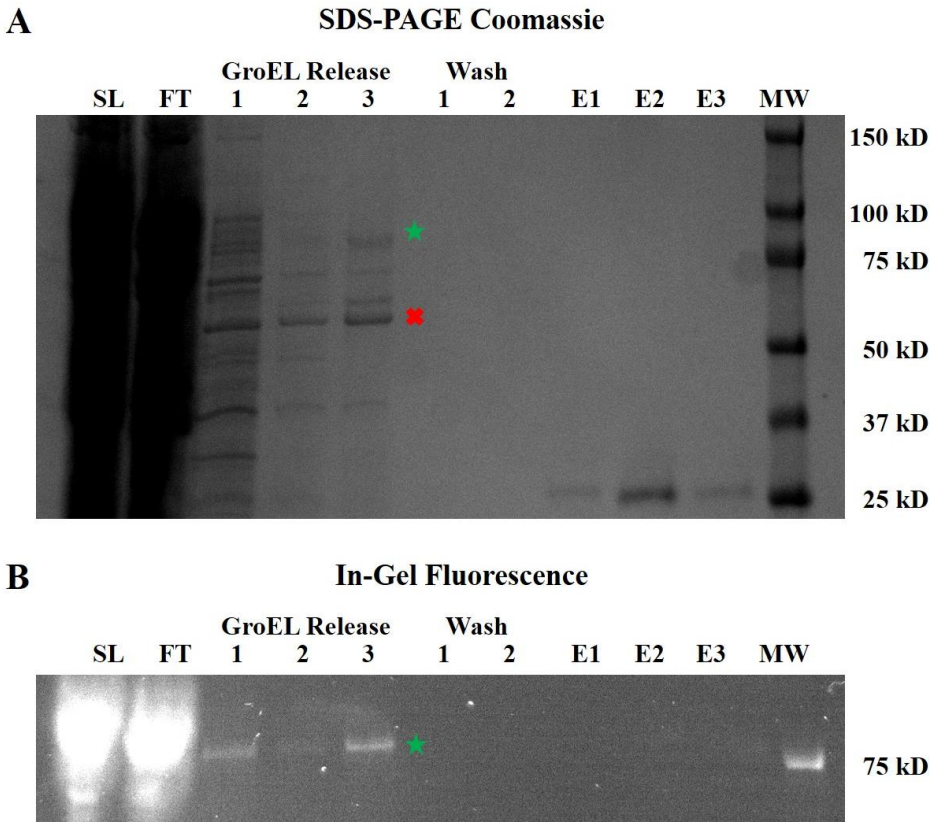

Figure 2.15: GroEL Removal VII: SUMO-β2AR:M13-EmGFP Ni-NTA purification and GroEL substrate release with SDS detergent.
Sample description: SL = Soluble lysate, FT = Flow-through, GroEL release1-3 = GroEL release buffer contained 0.001% SDS (1), 0.01% SDS (2), and 0.1% SDS (3), Wash1-2 = Wash buffer applied after GroEL release steps, E1-3 = Elution fractions. (A) SDS-PAGE analysis of purification samples using Coomassie G250 for general protein visualization. (B) SDS-PAGE and in-gel fluorescence analysis to detect SUMO-β2AR:M13-EmGFP presence.

use an increasing concentration of CHAPS in successive steps of the soluble lysate binding to the Ni-NTA agarose in the column. The prepared soluble fraction of the cell lysate initially contained 2 mM CHAPS. After the first round of binding to Ni-NTA agarose, FT1 was increased to 4 mM CHAPS and re-loaded in the column. A small volume of the resulting flow-through sample (FT2) was taken for SDS-PAGE analysis in Figure 2.16. Stock buffer was added to the remaining FT2 to give a final concentration of 8 mM CHAPS. Subsequently, FT2 was re-loaded in the column and collected as FT3 for gel samples. The rationale behind this change in GroEL

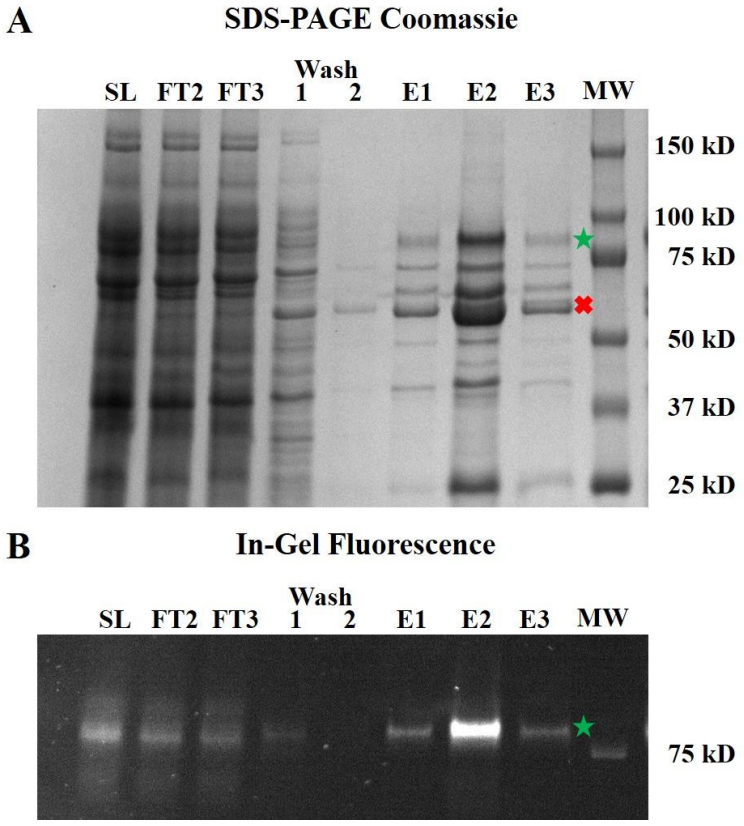

Figure 2.16: GroEL Removal VIII: SUMO-β2AR:M13-EmGFP Ni-NTA purification and GroEL substrate release with CHAPS detergent.
Sample description: SL = Soluble lysate, FT1 (not shown) = CHAPS added to flow-through 1 (2 mM) prior to reloading on column, FT2 = CHAPS added to flow-through 2 (4 mM) prior to reloading on column, FT3 = CHAPS added to flow-through 3 (8 mM) prior to reloading on column, Wash1-2 = Wash buffer applied after GroEL release steps assumed in FT1-3 steps, E1-3 = Elution fractions. (A) SDS-PAGE analysis of purification samples using Coomassie G250 for general protein visualization. (B) SDS-PAGE and in-gel fluorescence analysis to detect SUMO-β2AR:M13-EmGFP presence.

removal method was to disrupt the GroEL-target protein binding and remove as much chaperone as possible in the initial stages of binding to the Ni-NTA agarose. However, this approach was not sufficient in completely removing GroEL based on the elution fractions in Figure 2.16 (A).

Based on the nature of GroEL-substrate binding, the strong association between GroEL and the M13 mutants suggests there are still significant regions of hydrophobicity in the transmembrane (TM) region. This situation served as a useful test case for the M15 (M9E6S) mutation design. Also, it was suggested that we test the M15 design with a different GPCR to rule out the possibility of β2AR itself being the bottleneck in GroEL removal. We had sub-cloned the water-soluble LPA_1 receptor mutant (r:M15) DNA sequence into the pE-SUMO vector. The LPAR1:M15 DNA sequence was initially inserted in the pET-28a vector for trial expressions in *E. coli*, but protein expression was very poor (data not shown) in all the conditions tested. A FLAG tag (DYKDDDDK) had been inserted at the C-terminus of SUMO-LPAR1:M15 by Dr. Alexandra Kikonyogo and Binh Nguyen. The FLAG tag allows for affinity chromatography with the Anti-FLAG M2 agarose gel and produces cleaner preparations during protein purifications. At the time of this project, affinity purification with FLAG was preferred to the traditional IMAC methods. However, FLAG purification conditions must be milder relative to IMAC methods due to the sensitivity of the Anti-FLAG M2 antibody toward denaturing. Thus, for GroEL Removal IX (Figure 2.17) with SUMO-LPAR1:M15-FLAG we necessarily had to continue using the Ni-NTA purification method. The protocol that was followed in GroEL Removal VIII was repeated for this experiment with the same CHAPS detergent concentrations. As opposed to the β2AR:M13 constructs in the previous experiments, SUMO- LPAR1:M15-FLAG has a molecular weight slightly below that of GroEL. It is difficult to identify the protein bands with only the Coomassie-stained gel image in Figure 2.17 (A). Since this LPAR1:M15

mutant lacked an EmGFP fusion partner, it was necessary to perform western blot analysis to identify both proteins with similar molecular weight (Figure 2.17 (B, C)).

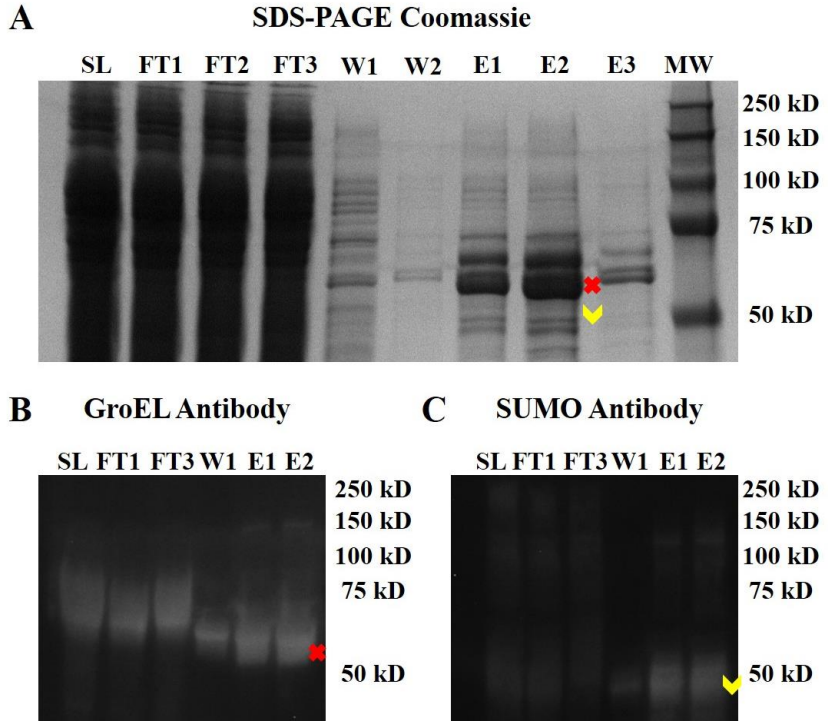

Figure 2.17: GroEL Removal IX: SUMO-LPAR1:M15-FLAG Ni-NTA purification and GroEL substrate release with CHAPS detergent.
Sample description: SL = Soluble lysate, FT1 = CHAPS added to Flow-through 1 (2 mM) prior to reloading on column, FT2 = CHAPS added to Flow-through 2 (4 mM) prior to reloading on column, FT3 = CHAPS added to Flow-through 3 (8 mM) prior to reloading on column, Wash1-2 = Wash buffer applied after GroEL release steps assumed in FT1-3 steps, E1-3 = Elution fractions. (A) SDS-PAGE analysis of purification samples using Coomassie R250 for general protein visualization. (B) SDS-PAGE/Western blot analysis of purification samples using GroEL antibody to detect GroEL presence (~60 kD, red x). (C) SDS-PAGE/Western blot analysis of purification samples using SUMO antibody to detect SUMO-LPAR1:M15-FLAG (~55 kD, yellow chevron) presence.

Again, these data showed that 2 – 8 mM CHAPS detergent was not sufficient to remove GroEL from the target protein as it was clearly present in the elution fractions. In terms of GroEL disassociation in the presence of CHAPS detergent, it appeared that both M13/M15 mutation sets affected β2AR and LPAR1 similarly. Also, there did not seem to be anything specific about the β2AR or LPAR1 that was causing strong association to GroEL. While it is difficult to interpret

comparisons between the different mutation sets when they were applied to different receptors, it was clear that a different strategy would be necessary to produce WS GPCR mutants.

In some of the gel images (Figures 2.16 and 2.17), an additional band was observed that runs slightly higher than the proposed GroEL bands (>60 kD). We speculate this higher molecular weight band could be the 70 kD chaperone protein, DnaK (HSP70 homolog), that functions both independently, and in tandem with the GroEL system to promote protein folding.[176,178] It has been proposed that the proteins interact in a successive manner in which DnaK shuttles polypeptide substrates from the ribosome to the GroEL/ES system, but with the assistance of other molecular chaperones (DnaJ and GrpE).[179] Similar to GroEL, it has been reported to be a contaminant bound to target proteins after purifcations.[180]

2.4 Conclusion

As noted in Table 2.1, GroES and urea have been used independently in reported strategies to separate target proteins from GroEL during purification. However, there were several different problems that made these strategies impracticable. For the GroES strategy, the issue was project-funding dependent because the high reagent pricing constrained us to very few hypothetical test experiments. Even if we assumed successful outcome, we still lacked funding for the greater amount of GroES necessary in a large-scale preparation of WS-GPCR intended for crystallization trials. The issues with the urea strategy were related to concerns with the potential need for subsequent GPCR refolding experiments. In the literature, there are some examples of protein refolding experiments with GPCR including the leukotriene receptors BLT1,[181,182] BLT2,[181,183] cannabinoid receptor 1 (CB1),[181,184] serotonin receptor 4A (5-HT4A),[181,185] and the glucagon-like peptide-1 receptor (GLP1R).[186] These methods involved refolding in detergents, lipid-detergent, or synthetic polymer (amphipols) mixtures, but resulted

in varying degrees of success and often recovered sub-optimal (50% or less of total protein) yields of functional GPCR.[187] Furthermore, conditions for successful refolding experiments are highly receptor-specific. Therefore, it is likely that a large amount of time would have to be invested in testing and optimizing GPCR refolding conditions.

Chapter 3

A Benchmark of Loop Modeling Methods Applied to G Protein-Coupled Receptors

3.1 Introduction

Many proteins contain loop segments that can adopt a wide range of dynamic structures, including disordered random-coil loops, short "turn" loops, and long loops with embedded secondary structures. These loop segments are often flanked by secondary structural elements or distinct protein domains. Due to their inherent conformational flexibility and structural diversity, loops serve a wide range of important functions for many proteins.[188,189] For example, loops that are exposed on the protein surface often interact with other biomolecules.[190-192]

3.1.1 Significance of extracellular loop 2 (ECL2) in GPCR: Structural characteristics and functional roles

As mentioned in Chapter 1, the extracellular and intracellular loops (ECLs and ICLs) of GPCR have lower sequence and structural conservation than the TM domains. With respect to the available known GPCR structures (mostly Class A GPCR), ECL2 is generally the longest and most diverse in terms of amino acid identity and three-dimensional structure. Despite low sequence conservation, an overwhelming majority of GPCR contain a disulfide bond between highly conserved cysteine residues in ECL2 and the top (extracellular) of TM3. Based on the analysis of 367 GPCR sequences representing members from Class A, B1, B2, C, and F that were downloaded from the GPCRdb alignment tools,[193,194] 89% (327 out of 367) of sequences contain the conserved cysteine residues in ECL2/TM3. Also, the disulfide bond between the cysteine sidechains is observable in 94% (47 out of 50) of representative crystal structures of unique GPCR (as of May 2018). Three of the lipid receptors (LPAR1, S1PR1, CB1) that have known crystal structures lack the conserved disulfide bond between ECL2/TM3. Instead, they

contain an intra-loop disulfide bond that constrains the loop conformation. Herein, the conserved cysteine in ECL2 will be denoted by C45.50 because the loop connects TM4-5. This disulfide bond provides a structural constraint that tethers ECL2 to the GPCR TM bundle. For some GPCR, interfering with the conserved ECL2-TM3 disulfide bond is detrimental to receptor function.[195,196]

For many GPCR, ECL2 plays important roles in GPCR activation, orthosteric ligand binding, and allosteric ligand interactions.[197,198] Mutagenesis experiments on ECL2 of the complement C5a receptor and thrombin receptor resulted in constitutively active GPCR (signaling activity without agonist binding).[199,200] These findings suggest that in some GPCR, ECL2 functions as a negative regulator that dampens signaling by restricting the transition to active receptor states in the absence of endogenous ligand. However, ECL2 mutagenesis does not uniformly confer constitutive GPCR activity. For example, the A204E mutation in ECL2 of the ghrelin receptor resulted in diminished constitutive activity.[201] Thus, it is generally understood that ECL2 plays a role in GPCR function, but many of the details are receptor-dependent.

Given that the binding pocket features of closely-related GPCR are relatively similar, there must be other structural aspects that give rise to observed differences in receptor-ligand specificity. Indeed, the diversity among ECL2 amino acid sequences and structural features of closely-related GPCR contributes to the receptor-specific ligand interactions.[195] Table 3.1 below shows examples of GPCR crystal structures with ECL2-ligand interactions.

Table 3.1: GPCR structures with ECL2 contacts to the crystallized ligand.
Cys 45.50 indicates the conserved ECL2 residue that forms a disulfide bond with Cys 3.25. ECL2 residue interactions are indicated by the following abbreviations: Hydrophobic (HYD), Backbone H-bond (BBH), Sidechain H-bond (SCH) Aromatic (ARO), Charge-assisted H-bond (CAH+/-). Each set of ECL2 interactions were compiled from the pre-calculated GPCR-ligand interaction tool on GPCRdb[36].

GPCR	PDB ID	Cys 45.50	ECL2 Residue Interactions with Crystallized Ligand
DRD3	3PBL[166]	C181	I183, HYD
CXCR4	3ODU[164]	C186	C186, BBH
LPAR1	4Z35[202]	N/A	M198, HYD
S1PR1	3V2Y[170]	N/A	V194, BBH
CB1	5U09[203]	N/A	F268, HYD
P2Y12R	4PXZ[204]	C175	C175, BBH; K179, CAH+
SMO	4JKV[165]	C390	V386, HYD; Y394, ARO
RHO	2Z73[205]	C186	F188, HYD
B2AR	2RH1[34]	C191	F193, ARO
PAR2	5NDD[206]	C226	C226, HYD; H227, ARO; D228, CAH-
PAR1	3VW7[171]	C254	L258, BBH; L262, HYD
APJR	5VBL[207]	C181	L173, HYD; N175, N177, BBH; T176, SCH
FFAR1	5TZR[208]	C170	L158, HYD; W174, HYD
AA1R	5UEN[209]	C169	F171, ARO
AA2AR	4EIY[210]	C166	F168, ARO; E169, CAH-

3.1.2 ECL2 is a problematic modeling target with traditional template-based methods

Template-based modeling methods rely heavily on target-template sequence alignments where gaps often lead to errors in the target model. As mentioned previously, GPCR loop segments display low sequence conservation relative to the TM domains and tend to exhibit variable lengths which inevitably introduces gaps in GPCR sequence alignments. Therefore, GPCR loops tend to be difficult modeling targets with template-based methods.

Ab initio loop modeling can be described as a "mini-protein folding problem" with success largely depending on two general components: sampling and scoring. An extensive search to sample loop conformational space is implemented with the target sequence. Also, a method to evaluate or score the loop model conformations that are most likely to be near-native (close to the reference structure) is necessary. ECL2 within GPCR tends to be a more

challenging modeling target for conformational sampling because longer loops have more degrees of freedom. ECL2 makes contacts to various ligands in experimentally determined structures (Table 3.1), the impact of which will necessarily be neglected by the scoring function when ligand is absent during loop modeling.

3.1.3 GPCR ECL2 modeling benchmark: Research questions and goals

Our primary research question at the start of this benchmark study was, which of the available modeling software methods can accurately predict loop models that are close to the reference loop structures? The prediction involves generating model loop structures and ranking them based on criteria (i.e. energy functions as in Rosetta and MOE) that desirably correlate with experimentally determined structures. Our primary goal was to establish a method for re-modeling the ECL2 of GPCR models and choosing candidate models based on the benchmark results. The reference GPCR crystal structures used in this benchmark for ECL2 modeling are listed in Table 3.2. Many of the GPCR crystal structures that were available at the start of this study (August 2017) were included in this benchmark study to represent the "native" ECL2 reference structures. Crystal structures with poorly resolved or completely missing residues in ECL2 were excluded from the benchmark. The highest resolution structures were chosen as references for individual GPCR that had multiple crystal structures available.

Table 3.2: GPCR reference structures used in loop modeling benchmark.
The reference structures were divided into four categories, Groups 1-4, of increasing expected difficulty based on loop lengths.

Group	GPCR	PDB ID / Resolution (Å)	ECL2 Residues	ECL2 Length	ECL2 Features
1	GLR	5XF1[211] / 3.2	F289 - N300	12	Loop, no 2°
	GLP1R	5NX2[212] / 3.7	Y291 - N302	12	3-10 helix
	DRD3	3PBL[166] / 2.9	F172 - N185	14	Loop, no 2°
	CRFR1	4Z9G[213] / 3.2	D254 - T268	15	3-10 helix
	ACM4	5DSG[214] / 2.6	K177 - N192	16	3-10 helix
	US28	4XT1[215] / 2.9	V166 - E181	16	β-hairpin
	CXCR4	3ODU[164] / 2.5	N176 - N192	17	β-hairpin
2	LPAR1	4Z35[202] / 2.9	W186 - S203	18	3-10 helix
	S1PR1	3V2Y[170] / 2.8	W182 - H199	18	3-10 helix
	CB1	5U09[203] / 2.6	W255 - D272	18	α-helix
	P2Y12R	4PXZ[204] / 2.5	T163 - S180	18	3-10 helix
	SMO	4JKV[165] / 2.5	A379 - N396	18	β-hairpin
	OPRD	4N6H[216] / 1.8	A187 - P205	19	β-hairpin
	CCR5	5UIW[217] / 2.2	T167 - S185	19	β-hairpin
3	P2Y1R	4XNV[218] / 2.2	S190 - E209	20	β-hairpin
	AT2R	5UNG[219] / 2.8	R182 - E202	21	β-hairpin
	RHO	2Z73[205] / 2.5	G173 - D194	22	β-hairpin
	GRM1	4OR2[220] / 2.8	E728 - S749	22	β-hairpin
	NTR1	4XES[221] / 2.6	M208 - D230	23	β-sheet, 3-10 helix
	B2AR	2RH1[34] / 2.4	W173 - N196	24	α-helix
	ETBR	5GLI[222] / 2.5	F240 - T263	24	β-hairpin
4	PAR2	5NDD[206] / 2.8	V211 - L235	25	β-hairpin
	PAR1	3VW7[171] / 2.2	L239 - E264	26	β-hairpin
	APJR	5VBL[207] / 2.6	R168 - S193	26	β-hairpin
	OX1R	4ZJ8[223] / 2.8	M183 - D209	27	β-hairpin
	FFAR1	5TZR[208] / 2.2	A146 - D175	30	Loop, β-sheet
	AA1R	5UEN[209] / 3.2	G145 - S176	32	(2) α-helices
	AA2AR	4EIY[210] / 1.8	G142 - P173	32	(2) α-helices

Although not a comprehensive list, Table 3.3 shows several GPCR with more than two crystal structures available at the beginning of this study. The dynamic nature of these loops is apparent in these sets of superposed crystal structures. The ECL2 Cα atoms in different structures of a single protein have root mean square deviation (RMSD) values ranging from 0.3 – 1.5 Å. The RMSD is a measure of the average atomic distance between superimposed protein structures, and is calculated using the following equation

$$RMSD = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\delta_i^2} \quad (1)$$

in which the distance between the i^{th} pair of corresponding atoms is represented by δ_i. Traditionally the gold standard of structure prediction is achieving top-ranking models with sub-angstrom accuracy (Cα RMSD under 1.0 Å) to the reference "native" structure. However, ECL2 experimental variability in different crystal structures of the same GPCR can exceed this value. For the ECL2 targets in this benchmark, it is more reasonable to consider methods that produce top-scoring models with near-atomic accuracy (Cα RMSD within 2.5 Å) the threshold for success.[224] Often, models with near-atomic accuracy are sufficient for applications downstream of modeling.[50]

Table 3.3: ECL2 structure variation among GPCR with more than 2 crystal structures.
The ECL2 residues of each set of GPCR PDBs were superposed and the Cα RMSD values were calculated.

GPCR	PDB IDs	ECL2 Residues	ECL2 RMSD
GLR	4L6R, 5EE7, 5XEZ, 5XF1[168,211,225]	F289 - N300	1.29 Å
GLP1R	5NX2, 5VEW, 5VEX, 6B3J[212,226,227]	Y291 - N302	1.52 Å
ACM2	3UON, 4MQS, 4MQT[27,228]	V168 - S182	0.76 Å
ACM3	4U14, 4U15, 4U16, 4DAJ[28,229]	K212 - S226	0.46 Å
CXCR4	3ODU, 3OE0, 3OE6, 3OE8, 3OE9, 4RWS[164,230]	N176 - N192	0.81 Å
LPAR1	4Z34, 4Z35, 4Z36[202]	W186 - S203	0.28 Å
CB1	5XRA, 5XR8, 5TGZ, 5U09[203,231,232]	W255 - D272	0.64 Å
P2Y12R	4NTJ, 4PXZ, 4PY0[204,233]	T163 - S180	0.93 Å
SMO	4JKV, 4O9R, 4QIN, 4QIM, 4N4W, 5L7I, 5L7D[165,234-236]	A379 - N396	0.40 Å
OPRD	4N6H, 4RWD, 4RWA[216,237]	A187 - P205	0.43 Å
NOP	4EA3, 5DHG, 5DHH[157,238]	T207 - P224	0.75 Å
AT2R	5UNG, 5UNF, 5UNH[219]	R182 - E202	0.59 Å
NTR1	3ZEV, 4GRV, 4XEE, 4XES, 4BWB, 4BV0, 4BUO, 5T04[159,221,239,240]	M208 - D230	0.75 Å
B1AR	2Y00, 2Y01, 2Y02, 2Y03, 2Y04, 2YCW, 2YCX, 2YCY, 2YCZ, 2VT4, 3ZPR, 3ZPQ, 4AMI, 4AMJ, 4BVN, 4GPO, 5F8U, 5A8E[29,160,241-246]	W181 - T203	0.43 Å
B2AR	2RH1, 2R4R, 2R4S, 3D4S, 3NY8, 3NYA, 3NY9, 3KJ6, 3P0G, 3PDS, 3SN6, 4LDE, 4LDL, 4LDO, 4GBR, 4QKX, 5JQH, 5D5B, 5D5A, 5D6L[21,30,32,34,35,162,247-254]	W173 - N196	0.40 Å
ETBR	5GLI, 5GLH, 5X93, 5XPR[222,255]	F240 - T263	1.14 Å
RHO	1U19, 1GZM, 1L9H, 1F88, 1HZX, 2G87, 2PED, 2J4Y, 2HPY, 2I35, 2I36, 2I37, 2X72, 3OAX, 3CAP, 3DQB, 3PQR, 3PXO, 3C9M, 3C9L, 4X1H, 4J4Q, 4BEZ, 4BEY, 4PXF, 4A4M, 5TE3, 5TE5, 5DYS, 5EN0[19,20,24,156,256-274]		0.65 Å
AA2AR	2YDV, 2YDO, 3REY, 3PWH, 3RFM, 3EML, 3QAK, 3UZC, 3UZA, 3VGA, 3VG9, 4EIY, 4UHR, 4UG2, 5IUB, 5IU8, 5IUA, 5IU4, 5IU7, 5G53, 5UIG, 5K2A, 5K2B, 5K2C, 5K2D[22,23,25,275-282]	G142 - P173	0.95 Å
	Average ECL2 RMSD: 0.73 Å, Std. Dev.: 0.33		

3.1.4 Description of loop modeling software: Rosetta and MOE

Rosetta loop modeling protocols can be used for loop refinement or loop reconstruction. This study only implements modeling in the context of loop reconstruction – *ab initio*/*de novo* prediction of the "native" loop conformation based on the amino acid sequence, but the initial backbone and sidechain conformations are discarded prior to modeling. Loop refinement, on the other hand, is utilized in finding lower energy conformations starting from a given loop conformation that is potentially close to the "native" structure.

The modeling process is implemented in two stages with iterations of Monte Carlo simulated annealing: An initial low-resolution/coarse-grained stage where the sidechain atoms are represented as "centroids" and a high-resolution/full-atom stage where the sidechain atoms are explicitly represented. Figure 3.1 shows a schematic overview of the general Rosetta loop

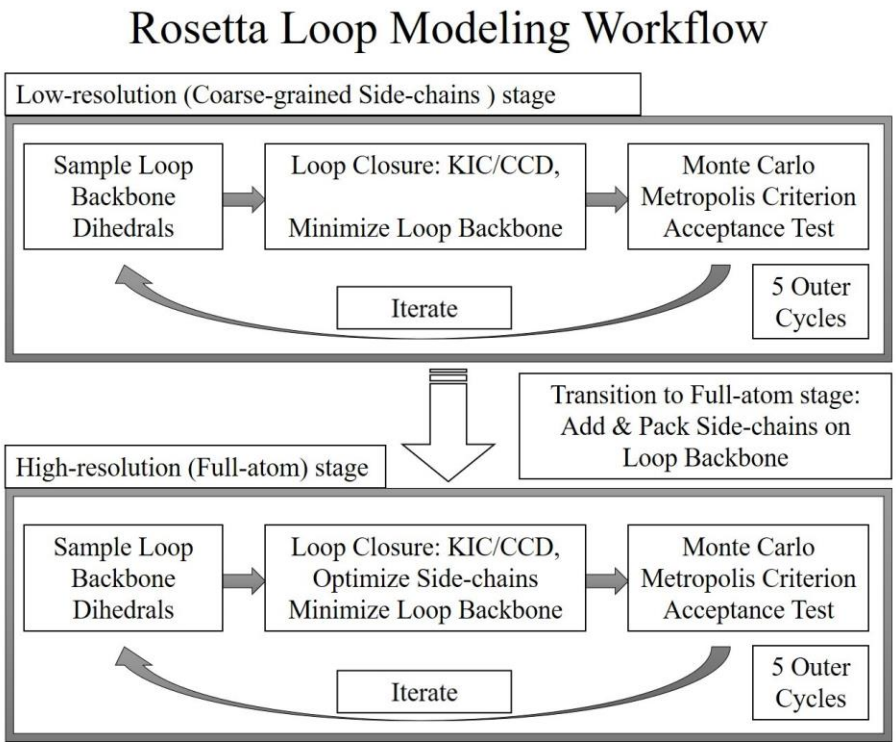

Figure 3.1: Overview schematic of Rosetta loop modeling process.

modeling process. The available loop modeling algorithms in Rosetta differ in conformational search (sampling) strategy and solutions to the loop closure problem.

The Cyclic Coordinate Descent (CCD) algorithm proceeds by optimizing the dihedral angles through consecutive loop residues from the N- to C-terminus where the goal is to minimize the distance between the free C-terminus end of the loop and the fixed anchor position.[283,284] The CCD algorithm in Rosetta uses experimentally-derived fragment libraries to guide the conformational search during loop modeling. The fragment libraries contain the coupled phi/psi dihedrals of peptide segments with 9 and 3 residues (9mers, 3mers) from the PDB.

The Kinematic Closure (KIC) method selects 3 pivot atoms (remaining loop backbone atoms are designated non-pivot) and divides the loop into two segments for conformational sampling of the non-pivot phi, psi dihedral angles. Subsequently, the pivot dihedral angles (6 phi, psi angles for 3 pivots) are analytically solved to position each rigid segment for loop closure.[285] The standard KIC protocol for loop modeling has subsequently been replaced by the Next Generation KIC (NGK) and KIC with Fragments (KICF) methods.

The Next Generation KIC (NGK) algorithm employs intensification strategies during non-pivot conformational sampling in both low- and high-resolution stages of the loop modeling process.[286] In the high-resolution stage, NGK implements additional annealing strategies that modulate the energy function to overcome large energy barriers. The intensification strategies involve 1) using neighbor-dependent Ramachandran distributions (Rama2b term) to select phi/psi dihedral combinations during sampling and 2) independently sampling ω angles based on observations in high-resolution crystal structures. Traditionally, the planar character of the peptide bond restricts the ω dihedral angle to either 180° for the common trans-configuration or

0° for the less common cis-configuration. However, analyses of high-resolution protein structures concluded that trans peptide ω values can vary by more than 25° from planarity in some cases, and that the non-planar character of peptides are more common than previously known.[287] During the NGK method, ω sampling is performed independently of the phi/psi dihedrals from a Gaussian around the observed mean of 179.1° ±6.3.[286] The annealing strategies implemented in the NGK method involve ramping the weights of 1) the repulsive component of the Lennard-Jones potential and 2) the Rama score (distinct from Rama2b term used in intensification strategy), which is the likelihood of a phi/psi combination occurring given an amino acid type. While the intensification strategies (Rama2b and ω sampling) are applied in both low- and high-resolution stages of loop modeling, the annealing strategies are only implemented in the high-resolution stage. Overall, these intensification and annealing strategies were found to greatly improve loop modeling accuracy compared to the standard KIC method.

The KIC with Fragments (KICF) method combines the fragment library sampling strategy from the CCD method with the KIC loop closure method. The main difference between this method and the NGK method is the way in which loop backbone conformations are sampled. The fragment-based sampling of phi/psi/omega dihedral angles consists of four major steps: (1) one of the given fragment libraries is selected at random and searched for alignment frames where fragments overlap with subsegments of the loop; (2) one of the alignment frames and fragments within that frame is selected at random; (3) the phi/psi/omega dihedral angles of that fragment are applied to the loop subsegment; and finally (4) kinematic closure (KIC) calculations are performed to achieve loop closure. The Rosetta all-atom energy function to evaluate/score biomolecular structures and models has evolved over many versions (Score12, Talaris2013, Talaris2014), with the Rosetta Energy Function 2015 (REF2015) becoming the

default scoring function as of July 2017.[288-291] However, the most recently available scoring function (REF2015) and loop modeling algorithm, (KICF), had not been tested on the 12-residue, or 14-17 residue loop modeling benchmark sets used by Rosetta developers (as of May 2018).

The MOE Loop Modeler application has a *de novo* search method and a PDB search method for generating potential loop backbone conformations. For this study, only the *de novo* search method was used to model the ECL2 of the benchmark GPCR. MOE loop modeling protocols also consist of distinct low-/high-resolution stages for loop modeling. The initial *de novo* search stage only deals with the loop backbone atoms, generating potential loop conformations that are ranked by an initial coarse scoring function before advancing to the full-atom stage. Figure 3.2 shows a schematic overview of the general MOE loop modeling process. MOE loop modeling uses an extension of the CCD algorithm, Full CCD (FCCD).[292] This method differs from CCD by solely operating on the Cα backbone atoms with pseudo bond angles and

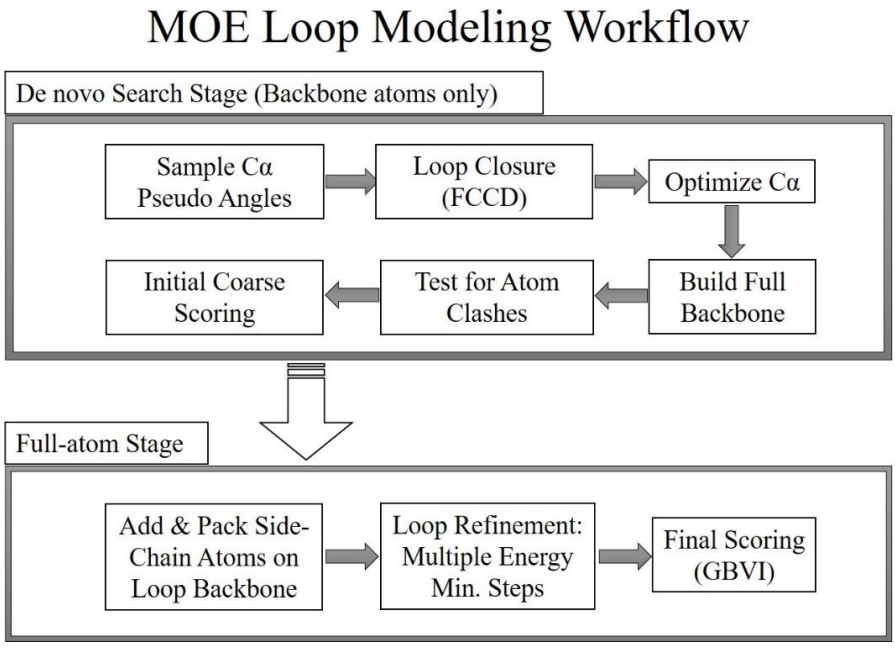

Figure 3.2: Overview schematic of MOE loop modeling process.

dihedral angles, optimizing both terms to achieve loop closure. Probability densities calculated from high-resolution PDB structures resulted in specific profiles for the Cα pseudo bond and dihedral angles. These profiles are used for random sampling of loop Cα conformational space during the *de novo* search stage, followed by FCCD loop closure.

After loop closure by FCCD, it is necessary to optimize the Cα backbone atoms. This is accomplished by using a component of the PULCHRA (Protein Chain Reconstruction Algorithm) method which employs a steepest descent gradient minimization and a simple harmonic potential to optimize the Cα positions before full backbone reconstruction.[293] To reconstruct backbone atom positions from the Cα loop traces generated, MOE uses the BBQ (Backbone Building from Quadrilaterals) method which is based on proximal distance geometries for sets of four sequential Cα atoms in the loop. Additionally, MOE backbone packer performs a minimization to relieve any strained backbone geometries and atom clashes. This step is followed by a final geometry and duplicate check to ensure that non-redundant backbone conformations with reasonable bond and dihedral angles are being evaluated by the coarse scoring function. The top-ranking loop backbone conformations are advanced to the full-atom stage.

In the full-atom stage, sidechain atoms are added to the loop and optimized with respect to the sidechain orientations. The entire loop segment is energy minimized through multiple steps before the final scoring step. The full-atom loop conformations generated are scored using GBVI (Generalized-Born Volume Integral). The potential energy of the system using GBVI has been shown to recover loop conformations close to the native from the Jacobson Loop Decoy Dataset.[294]

3.2 Methods

The Molecular Operating Environment (MOE version 2016.08) software package was used for GPCR structure preparation and visualization. The Rosetta software suite (Rosetta version 3.8) and MOE were used for modeling the ECL2 segment of the benchmark reference structures. For each of the GPCR in the benchmark, the Robetta fragment server was used to generate the "9-mer" and "3-mer" fragment library files that are necessary for Rosetta CCD and KICF loop modeling methods.[295] Fragment generation for each reference GPCR excluded any PDB data from the GPCR crystal structure itself, as verified by manual inspection of PDB ID codes in each fragment file.

3.2.1 Structure preparation protocol using MOE for GPCR targets

The benchmark GPCR structures listed in Table 3.2 were downloaded from the PDB[84] and prepared for loop modeling with MOE. Ligands and water molecules were deleted from the PDB structure files. For each GPCR structure, the "QuickPrep" process was used to streamline the structure preparation process. QuickPrep corrects any structural issues (i.e. residues with alternate locations, missing atoms, chain breaks, etc.) that often accompany structural data, adds explicit hydrogens and partial charges with "Protonate 3D," and performs a tethered-receptor energy minimization with the Amber12EHT[296-298] forcefield (RMS gradient of 0.1 kcal/mol/Å). The final energy minimization step is performed to improve any inaccurate geometries derived from the crystallographic data. During the minimization process, the receptor atoms are tethered to ensure that changes to the initial positions are modest.

3.2.2 Rosetta and MOE loop modeling methods

A total of 1000 ECL2 models were generated for the GPCR benchmark targets after structure preparation using each loop modeling method. All 28 reference structures in the

benchmark were reconstructed with the Rosetta methods discussed previously, but only 12 of the 28 reference structures were tested with MOE. This was due to time and software license limitations (limit on the number of allowed MOE batch jobs running concurrently), and most importantly the preliminary data from the shorter ECL2 targets which will be discussed later.

The following descriptions are the main Rosetta command line options used to perform NGK, KICF, and CCD loop modeling methods in low-resolution and full-atom stages. The remodel stage is the term for the initial coarse-grained modeling step. The remodel stage samples loop backbone conformations using a reduced representation of amino acid sidechains and a Rosetta low-resolution scoring function. This is initiated by the following options associated with the '-loops:remodel' command: 'perturb_kic', 'perturb_kic_with_fragments', 'perturb_ccd'. A loop definition file is separately generated for each reference GPCR to define the residues of the loop (ECL2) to be remodeled. By enabling the "extend loop" field in the loop files, the target loop segment's bond lengths, angles, and omega torsions are idealized, and all phi/psi values are replaced randomly from Ramachandran space to give an initial closed conformation at the start of remodel stage. This is to ensure that loop reconstruction is not influenced by the initial loop conformation. Subsequently, the loop phi/psi dihedrals are sampled using the fragment data described previously, followed by KIC or CCD calculations to achieve loop closure. Finally, the loop undergoes minimization using the BFGS (Broyden-Fletcher-Goldfarb-Shanno) algorithm and a Metropolis criterion acceptance test using the Rosetta low-resolution scoring function, score4L. The number of Monte Carlo steps in both stages of loop modeling is determined by the number of outer and inner cycles, (outer_cycles * inner_cycles). The default number of outer (5) and inner cycles (evaluated by: min(1000, number_of_loop_residues * 20)) were used for all Rosetta loop modeling jobs in this benchmark study. Loop poses are set to the lowest energy

conformation evaluated at the end of each outer cycle. The temperature decreases exponentially from 1.5 KT to 0.5 KT from the first step to the last. The refine stage is the term for the all-atom loop modeling step and is activated by the following options associated with the '-loops:refine' command: 'refine_kic', 'refine_kic_with_fragments', 'refine_ccd'. This stage implements a similar scheme to the perturb stage, with major differences in the all-atom treatment of the loop during conformational sampling and the scoring function used to guide and evaluate loop conformations during model production.

The following MOE loop modeling options were used during the *de novo* search stage, full-atom model generation, and final model scoring. The loop sequence for each GPCR ECL2 was selected in the sequence editor window and the following options were provided in the main MOE Loop Modeler window: only the *de novo* search method was enabled, the default RMSD limit of 0.50 was decreased to 0.25 Å, the max iterations and energy window were set to 1000 and 10, respectively. The number of *de novo* search runs, and final models built were set to 1000 total. Subsequently, SVL batch files were created to run the loop modeler jobs on a high-performance computing cluster. Due to the molecular database (.mdb files) storage limitations, 10 sets of 100 final models or 20 sets of 50 final models were generated.

3.3 Results and Discussion

Loop modeling performance throughout this benchmark study is assessed by comparing *de novo* models of ECL2 for proteins from Table 3.2 to the crystallographic reference structures indicated in the same table. Superpositions were performed for residues not modeled *de novo* before calculation of (Cα) RMSD values. The metrics reported throughout this section include: lowest RMSD model (LRM), top scored model (TSM), lowest RMSD model in the top 10 scored (LRM-T10), and lowest RMSD model in the top 25 scored (LRM-T25).

To decide which scoring function to use with Rosetta loop modeling, the NGK loop modeling algorithm was used to sample ECL2 conformations for the group 1 GPCR from Table 3.2 using both the most recently available Rosetta Energy Function 2015 (REF2015) and its predecessor, Talaris 2014. The LRM results show that models meeting the near-atomic accuracy metric of 2.5 Å were sampled in each run (Figure 3.3). However, when modeling loops of unknown structure, the TSM is more relevant than the LRM. In 3 out of 4 cases, REF2015 improved reference structure prediction accuracy (lower RMSD) for the top scored models. These data suggest that the REF2015 scoring function is more suitable for identifying models closer to the reference ECL2 structures in the benchmark. Furthermore, the most recent energy function has been parametrized to estimate energies in units of kilocalories per mole, whereas all

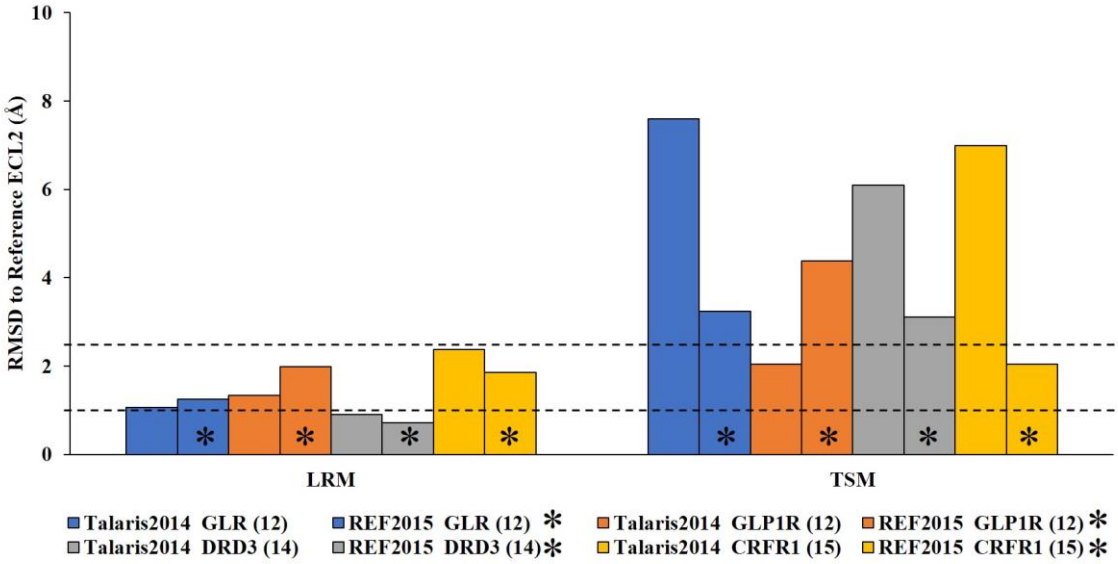

Figure 3.3: Energy function comparison with Rosetta NGK loop modeling.
Comparisons of the energy function influence on loop modeling performance is shown for the shortest GPCR ECL2 targets in the benchmark. Lowest RMSD Models (LRM) and Top Scored Models (TSM) are shown from 1000 models generated for each ECL2 target. The number in parentheses represents the ECL2 length.

previous Rosetta energy functions used arbitrary units.[291] Therefore, REF2015 was used in the Rosetta loop modeling protocols for all benchmark targets and method comparisons.

Loop modeling was performed using three algorithms in Rosetta and one algorithm in MOE for the Group 1 ECL2 targets from Table 3.2. The results from Group 1 (Figures 3.4 and 3.5) show that the NGK and KICF methods were able to sample models with better accuracy than either CCD or MOE-LM based on the RMSD of the LRM to the reference ECL2 structures. Notably, ECL2 of the muscarinic acetylcholine receptor M4 (ACM4, Figure 3.6) and dopamine receptor D3 (DRD3) targets were modeled with sub-angstrom accuracy using both NGK (LRM = 0.34, 0.71 Å, respectively) and KICF (LRM = 0.35, 0.76 Å, respectively) algorithms. For the ECL2 of CXCR4 (Figure 3.7), loop modeling using KICF also displayed sub-angstrom accuracy for the LRM (0.50 Å). In all of the figures showing superposed ECL2 of the LRM and reference structures, the green loop represents the reference crystal structure and the orange loop represents the LRM using NGK (panel A) and KICF (panel B). Six out of the seven ECL2 targets in Group 1 were modeled with near-atomic accuracy (RMSD ≤ 2.5 Å) using the NGK algorithm. Additionally, all seven of the targets were modeled with near-atomic accuracy using the KICF algorithm. Similarly, six out of the seven ECL2 targets were modeled with near-atomic accuracy using the CCD algorithm. However, none of the targets were modeled with sub-angstrom accuracy using the CCD algorithm. The average RMSD for the LRM produced by the NGK, KICF, and CCD algorithms were 1.66, 1.09, and 2.07 Å, respectively. In comparison, the average RMSD for the LRM produced by the MOE-LM *de novo* search algorithm was much higher, 3.37 Å. In addition, there were only two cases where the LRM had RMSD

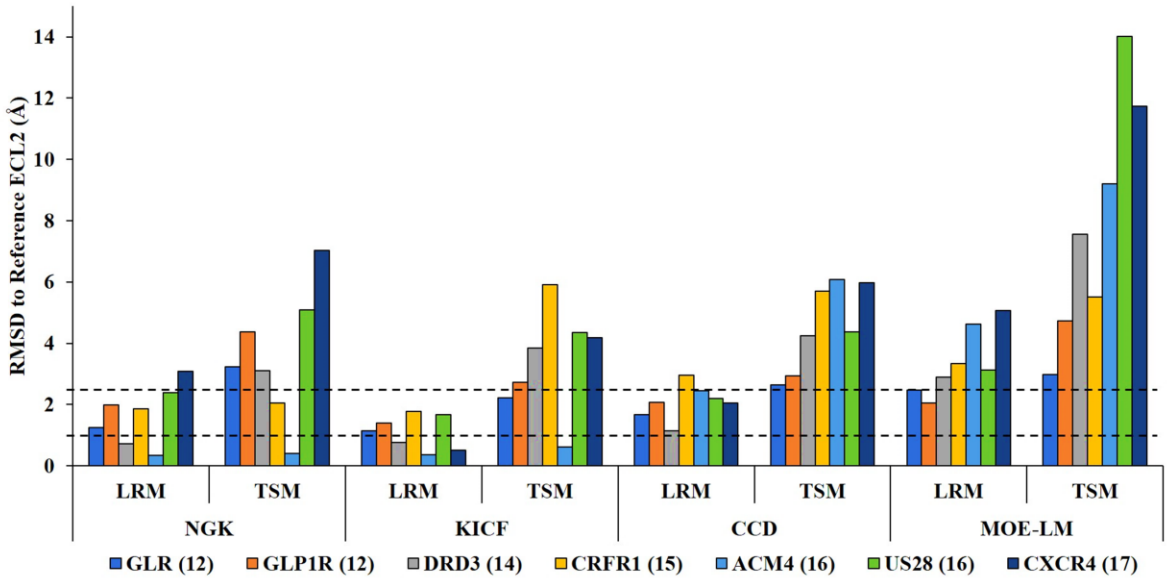

Figure 3.4: Group 1 results of the Lowest RMSD Models and Top Scored Models. The LRM and TSM are shown out of 1000 models generated for Group 1 ECL2 targets using NGK, KICF, CCD, and MOE-LM. In total, the LRM had sub-atomic accuracy in two and three cases when using the NGK and KICF algorithms, respectively. Additionally, the LRM had near-atomic accuracy in seven cases when using the KICF algorithm and six cases when using the NGK or CCD algorithms. The upper and lower dotted lines represent the near-atomic and sub-angstrom accuracy thresholds, respectively.

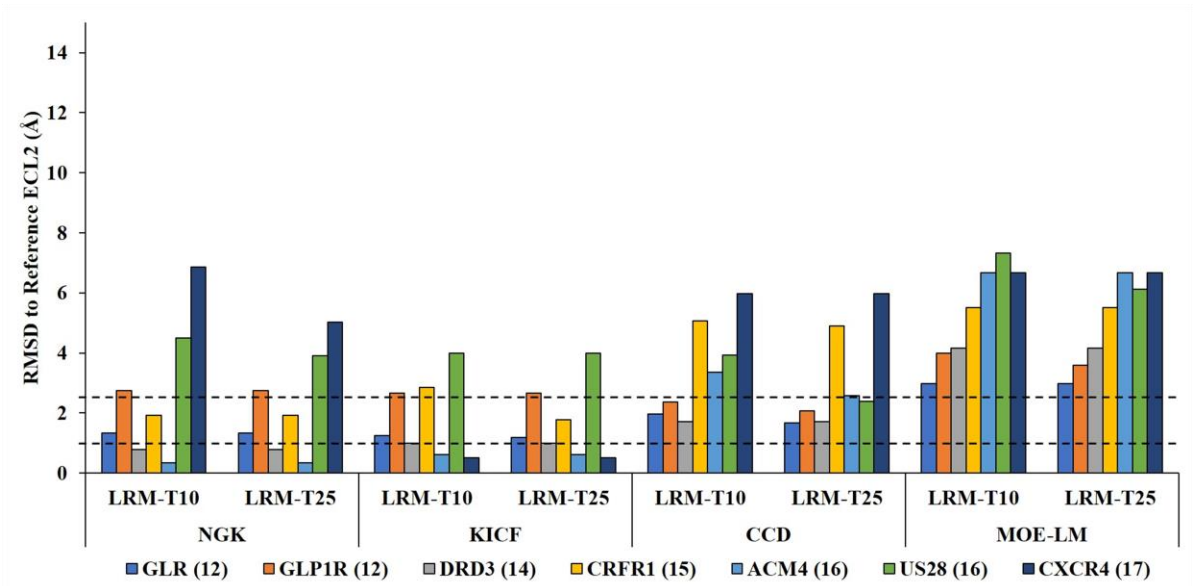

Figure 3.5: Group 1 results of the Lowest RMSD Models in the Top 10 and 25 Scored. The LRM-T10 and LRM-T25 are shown out of 1000 models generated for Group 1 ECL2 targets using NGK, KICF, CCD, and MOE-LM. The upper and lower dotted lines represent the near-atomic and sub-angstrom accuracy thresholds, respectively.

values below 2.5 Å. While generating ECL2 models with sub-angstrom or near-atomic accuracy overall is desirable, sampling loop conformations is only one aspect of structure prediction when the target structure is unknown. Loop models with low RMSD values to the target must also be scored or ranked favorably so they can be distinguished from the rest of the generated models. To evaluate the scoring component of loop modeling protocols, the RMSD of the top 1, 10, and 25 scored models (TSM, LRM-T10, LRM-T25) compared to the reference ECL2 structure were analyzed and tabulated in Figures 3.4 and 3.5. Overall, MOE-LM had a much higher average RMSD value for the TSM (7.96 Å) compared to the Rosetta loop modeling algorithms. The NGK, KICF, and CCD algorithms had average TSM values of 3.61, 3.40, and 4.56 Å, respectively. It is common practice to retain a subset of the top scored final models in structure prediction projects. For example, retaining the top 10 or 25 scored models increases the probability of selecting lower RMSD models for a target structure. For every ECL2 target in Group 1, the RMSD values for the LRM-T10/-T25 are lower than the RMSD value for the TSM. However, expanding the number of retained ECL2 models from the top 10 to the top 25 scored

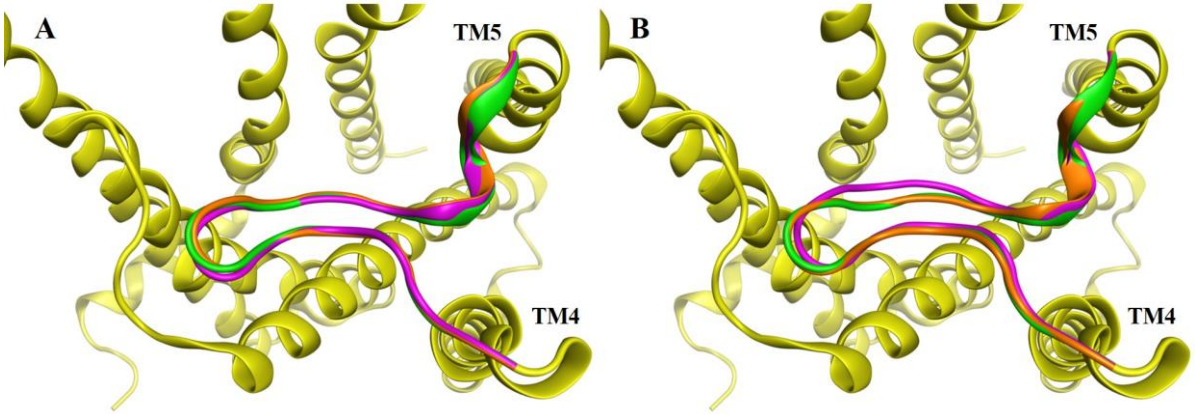

Figure 3.6: ECL2 models of ACM4 superposed with reference structure.
(A) The LRM (orange) and TSM (magenta) out of 1000 total models generated using NGK loop modeling had RMSD values of 0.34 Å and 0.40 Å to the reference structure (green). (B) The LRM and TSM out of 1000 total models generated using KICF loop modeling had RMSD values of 0.35 Å and 0.62 Å to the reference structure.

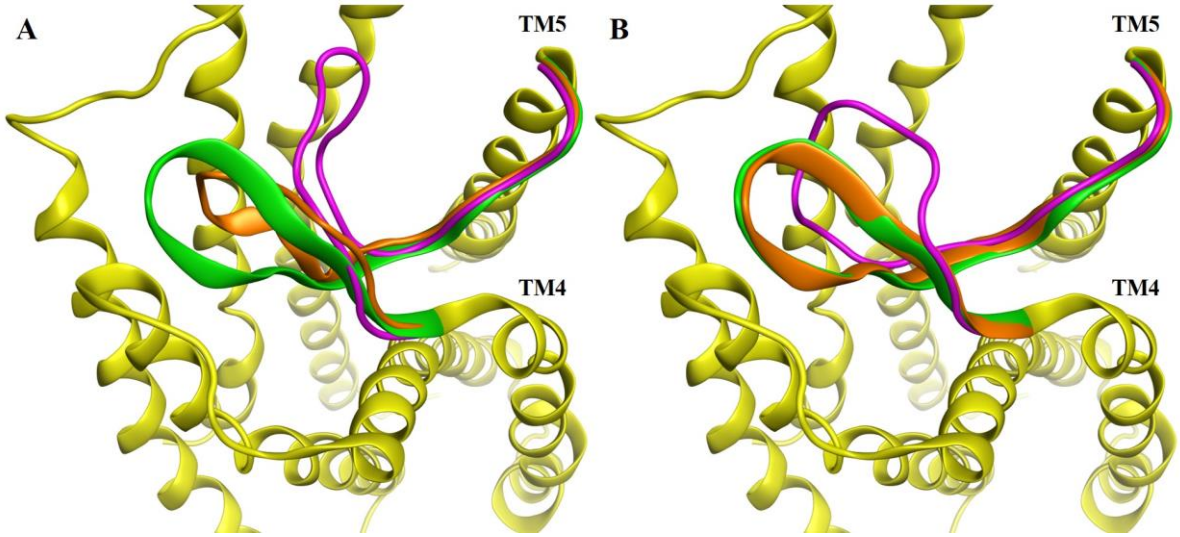

Figure 3.7: ECL2 models of CXCR4 superposed with reference structure.
(A) The LRM (orange) and TSM (magenta) out of 1000 total models generated using NGK loop modeling had RMSD values of 3.08 Å and 7.02 Å to the reference structure (green). (B) The LRM and TSM out of 1000 total models generated using KICF loop modeling had RMSD values of 0.50 Å and 4.19 Å to the reference structure.

did not result in identifying drastically lower RMSD models. In Group 1, there were only three of seven cases in which the LRM-T25 had a drastically lower RMSD value compared to the LRM-T10 (Figure 3.5). Models with near-atomic accuracy were generated and scored within the top 25 for the targets CRFR1 (using KICF), ACM4 (using CCD), and US28 (using CCD) that were not scored within the top 10. However, for most of the targets there was not a significant advantage in retaining the top 25 scored models rather than the top 10 out of 1000 models total.

All Group 2 ECL2 targets from Table 3.2 were modeled using the three loop modeling algorithms in Rosetta, and five of the seven were modeled with MOE-LM. Results for Group 2 targets (Figures 3.8 and 3.9) show that the NGK and KICF methods were able to sample loop conformations of the cannabinoid receptor type 1 (CB1) ECL2 with sub-angstrom accuracy overall (NGK/KICF LRM = 0.85/0.82 Å, Figure 3.10). Loop modeling with the KICF algorithm

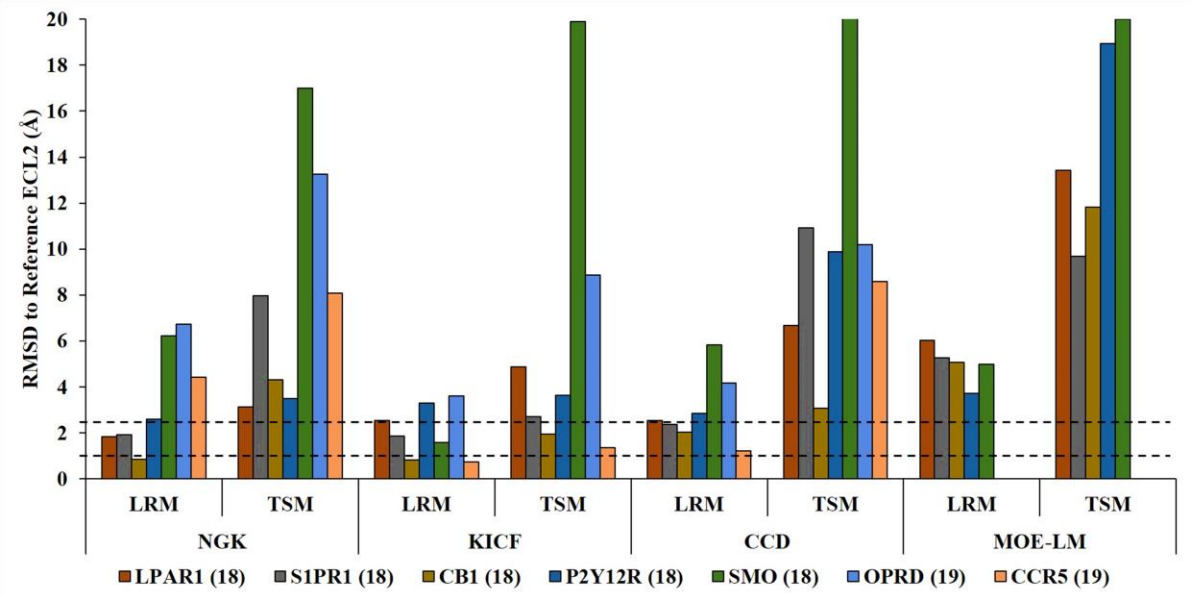

Figure 3.8: Group 2 results of the Lowest RMSD Models and Top Scored Models. The LRM and TSM are shown out of 1000 models generated for Group 2 ECL2 targets using NGK, KICF, CCD, and MOE-LM. In total, the LRM had sub-atomic accuracy in one and two cases when using the NGK and KICF algorithms, respectively. Additionally, the LRM had near-atomic accuracy in four cases when using the KICF algorithm and in three cases when using the NGK or CCD algorithms. The upper and lower dotted lines represent the near-atomic and sub-angstrom accuracy thresholds, respectively.

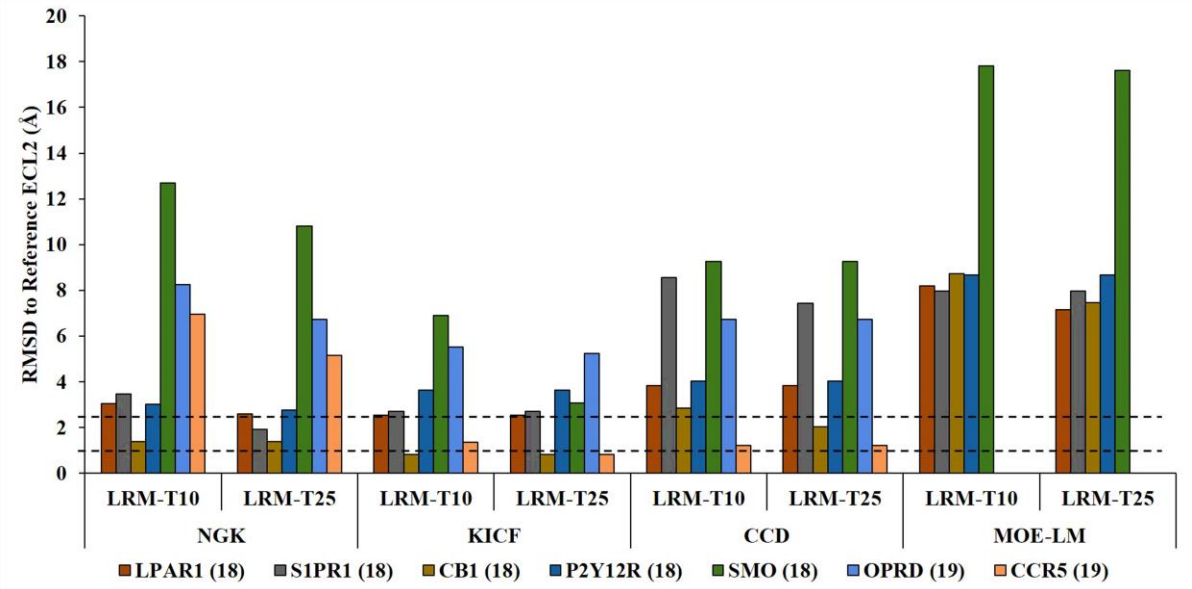

Figure 3.9: Group 2 results of the Lowest RMSD Models in the Top 10 and 25 Scored. The LRM-T10 and LRM-T25 are shown out of 1000 models generated for Group 2 ECL2 targets using NGK, KICF, CCD, and MOE-LM. The upper and lower dotted lines represent the near-atomic and sub-angstrom accuracy thresholds, respectively.

also achieved sub-angstrom accuracy with the longest loop in Group 2, CCR5 (LRM = 0.75 Å). In terms of near-atomic accuracy, there were three and four cases where the NGK and KICF algorithms sampled loop conformations with RMSD values ≤ 2.5 Å. There were three cases where the CCD algorithm sampled loop conformations with near-atomic accuracy, but no models with sub-angstrom accuracy were generated.

In Figure 3.8, the TSM for a majority of Group 2 targets had a much larger RMSD value than the LRM which is consistent with the results from Group 1. However, for the CB1 ECL2 target the TSM using KICF displayed near-atomic accuracy to the reference structure with an RMSD of 1.93 Å and the LRM was scored within the top 10 models (Figures 3.9 and 3.10 B). On the other hand, the TSM found using NGK had an RMSD of 4.32 Å to the reference structure and the LRM was not scored within the top 10 or 25 scored models. The lowest RMSD model found in the top 10 scored models using the NGK method displayed near-atomic accuracy to the

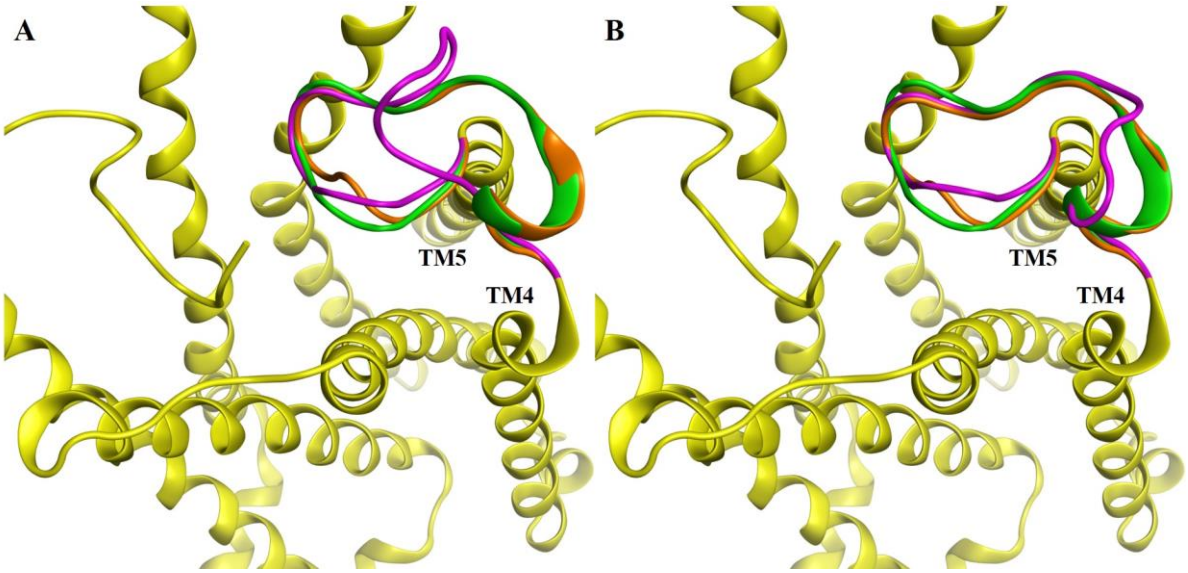

Figure 3.10: ECL2 models of CB1 superposed with reference structure.
(A) The LRM (orange) and TSM (magenta) out of 1000 models total using NGK loop modeling had RMSD values of 0.85 Å and 4.32 Å to the reference structure (green). (B) The LRM and TSM when KICF loop modeling was used had RMSD values of 0.82 Å and 1.93 Å to the reference structure.

reference structure with an RMSD of 1.39 Å. For the CCR5 ECL2 target, both methods that use fragment assembly (KICF and CCD) outperformed the NGK method in all 4 metrics. Notably, the LRM and LRM-T25 values from the KICF method with CCR5 were 0.75 and 0.81 Å.

In Group 2, the smoothened receptor (SMO) ECL2 was the most troublesome target for all three Rosetta loop modeling methods and MOE-LM. Out of all the loop modeling algorithms tested, the KICF algorithm yielded the most accurate LRM with an RMSD of 1.58 Å to the reference structure. The LRM found using the NGK, CCD, and MOE-LM algorithms had RMSD values of 6.22, 5.84, and 4.97 Å, respectively. However, the top scored model from this method had an RMSD of ~20 Å. While the ECL2 is situated just above the center of the TM bundle in the reference structure, SMO (Class F GPCR) differs from the other benchmark structures in many ways. Particularly, SMO has a longer ECL1 than other known GPCR structures (mostly Class A) and an extracellular domain (ECD) linker region that essentially form a lid over ECL2 and the TM bundle center (Figure 3.11). The long ECL1 and ECD linker regions might sterically hinder ECL2 loop conformations that are close to the reference structure. In other words, loop models that position ECL2 away from the TM bundle center, ECL1, and the ECD linker regions may be scored better. Since the SMO reference structure contained a co-crystallized ligand with contacts to the ECLs (ECL2 contacts shown in Table 3.1), it is also plausible that the conformation of the reference ECL2 is not as energetically favorable when the ligand is absent. A second set of ECL2 models (n = 1000) was generated for SMO using the NGK algorithm, but the ECD linker domain on the N-terminus was deleted prior to loop modeling. Of the Rosetta loop modeling algorithms used in this benchmark, NGK was the quickest at generating a given set of loop models. Thus, subsequent ECL2 modeling of benchmark targets were performed with the NGK algorithm due to time constraints. Out of the second set, the TSM and LRM were the

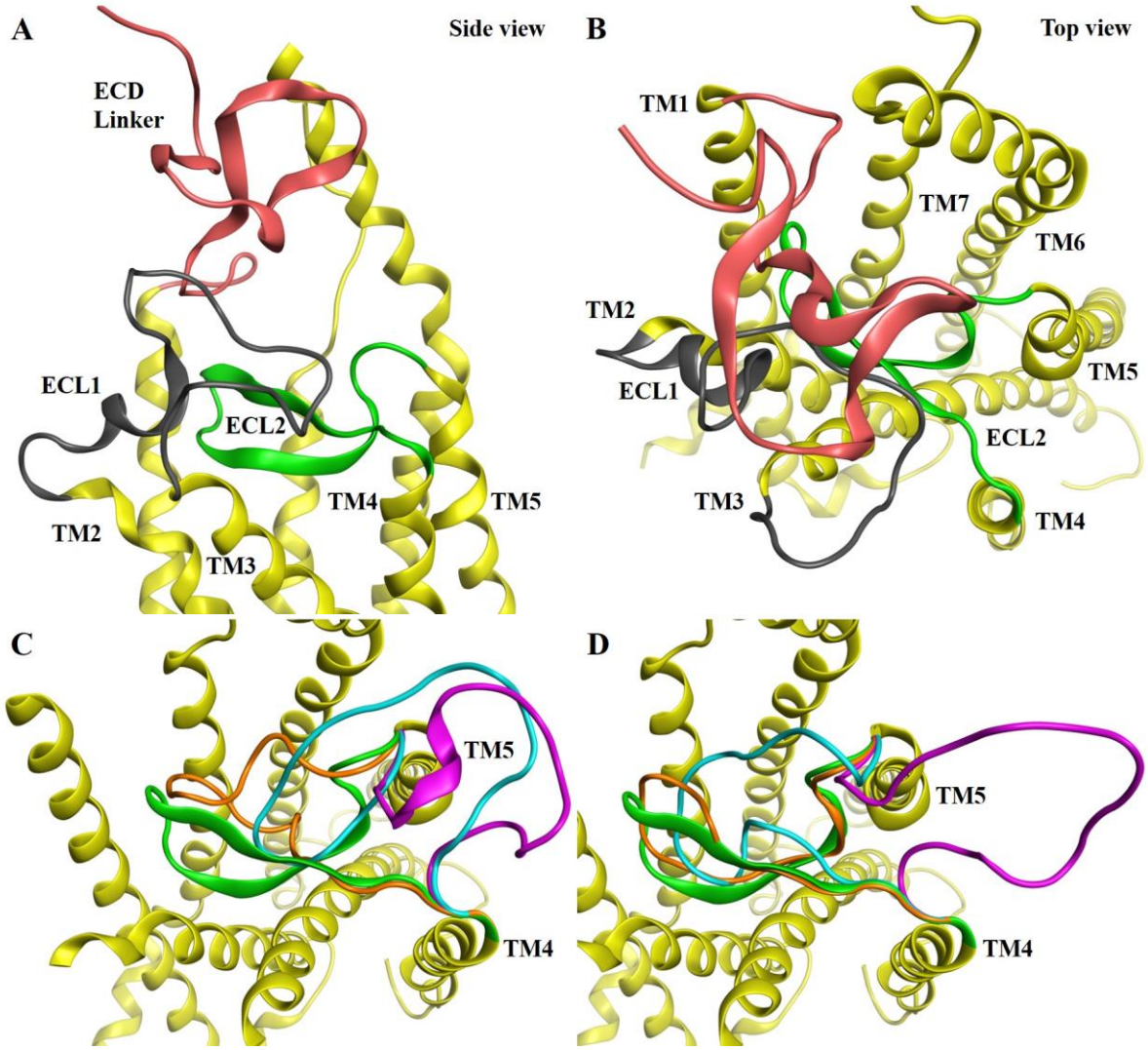

Figure 3.11: ECL2 models of SMO superposed with reference structure.
(A) Side view of SMO crystal structure (PDB:4JKV) highlighting the native ECL2 (green) buried underneath ECL1 (gray) and the ECD linker (salmon). (B) Top view of the extracellular side of the SMO crystal structure (C) The LRM (orange), TSM (magenta), and LRM-T10 (cyan) using the NGK algorithm had RMSD values of 6.22 Å, 17.0 Å, and 12.7 Å to the reference structure, respectively. (D) The LRM, TSM, and LRM-T10 using the KICF algorithm had RMSD values of 1.58 Å, 19.9 Å, and 6.89 Å to the reference structure, respectively. The ECD linker and ECL1 regions were hidden in panels C and D to visualize ECL2 and models clearly.

same and had an RMSD value of 3.30 Å to the reference ECL2. This is a significant improvement over the TSM and LRM RMSD values from the initial set of models produced by NGK which were 17.0 Å and 6.22 Å, respectively. This demonstrates that the steric hindrance

provided by the ECD linker domain is one impediment to sampling loop conformations similar to the reference SMO structure.

The MOE-LM method was used to model ECL2 for five of the seven total targets in Group 2 (Figures 3.8 and 3.9). Similar to the results from Group 1, MOE-LM was consistently less accurate than the Rosetta loop modeling algorithms based on the 4 metrics used (LRM, TSM, LRM-T10, LRM-T25). For the first five targets of Group 2, there were no models found with sub-angstrom or near-atomic accuracy to the reference structure. In addition to the technical limitations with MOE batch jobs, these initial results indicated that using the *de novo* search algorithm in MOE-LM was not feasible for the remaining Group 3-4 benchmark targets.

Group 3 targets from Table 3.2 were modeled using only the Rosetta loop modeling algorithms. The results from modeling the ECL2 of Group 3 targets (Figures 3.12 and 3.13) show that KICF was the only algorithm capable of sampling ECL2 models with sub-angstrom accuracy relative to the reference structures. The LRM generated by KICF for the P2Y1R (Figure 3.14 B) and AT2R ECL2 targets had sub-angstrom accuracy with RMSD values of 0.63 and 0.54 Å, respectively. However, the LRM of P2Y1R had near-atomic accuracy (RMSD ≤ 2.5 Å) to the reference structure for all three loop modeling algorithms.

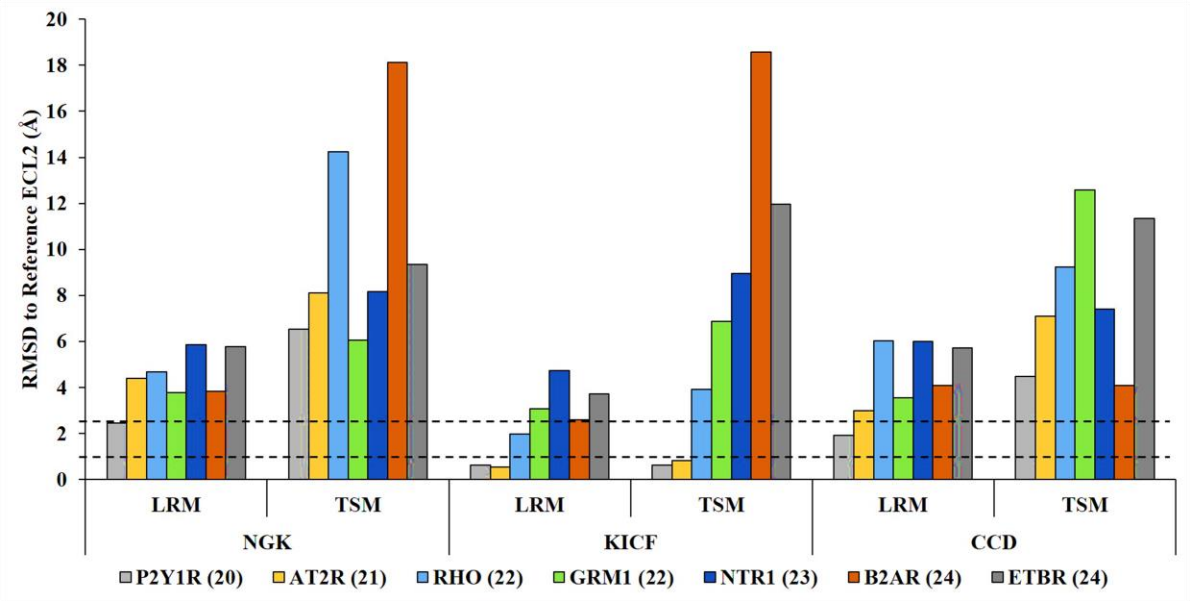

Figure 3.12: Group 3 results of the Lowest RMSD Models and Top Scored Models. The LRM and TSM are shown out of 1000 models generated for Group 3 ECL2 targets using NGK, KICF, and CCD. In total, the LRM had sub-atomic accuracy in two cases when using the KICF algorithm. Additionally, the LRM had near-atomic accuracy in three cases when using the KICF algorithm and in single cases when using the NGK or CCD algorithms. The upper and lower dotted lines represent the near-atomic and sub-angstrom accuracy thresholds, respectively.

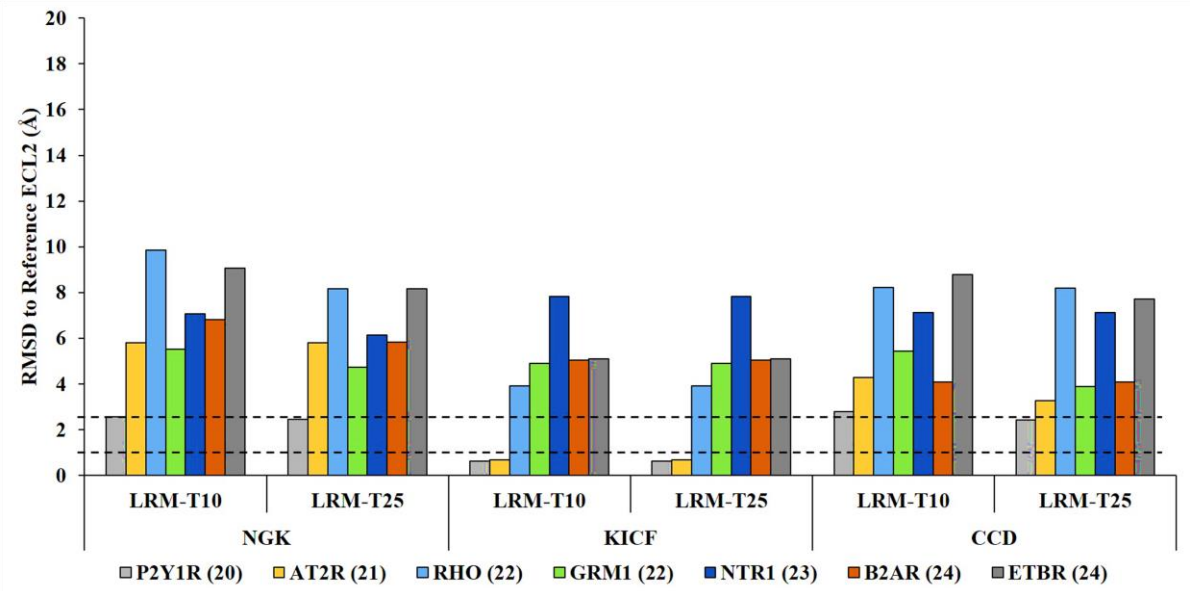

Figure 3.13: Group 3 results of the Lowest RMSD Models in the Top 10 and 25 Scored. The LRM-T10 and LRM-T25 are shown out of 1000 models generated for Group 3 ECL2 targets using NGK, KICF, and CCD. The upper and lower dotted lines represent the near-atomic and sub-angstrom accuracy thresholds, respectively.

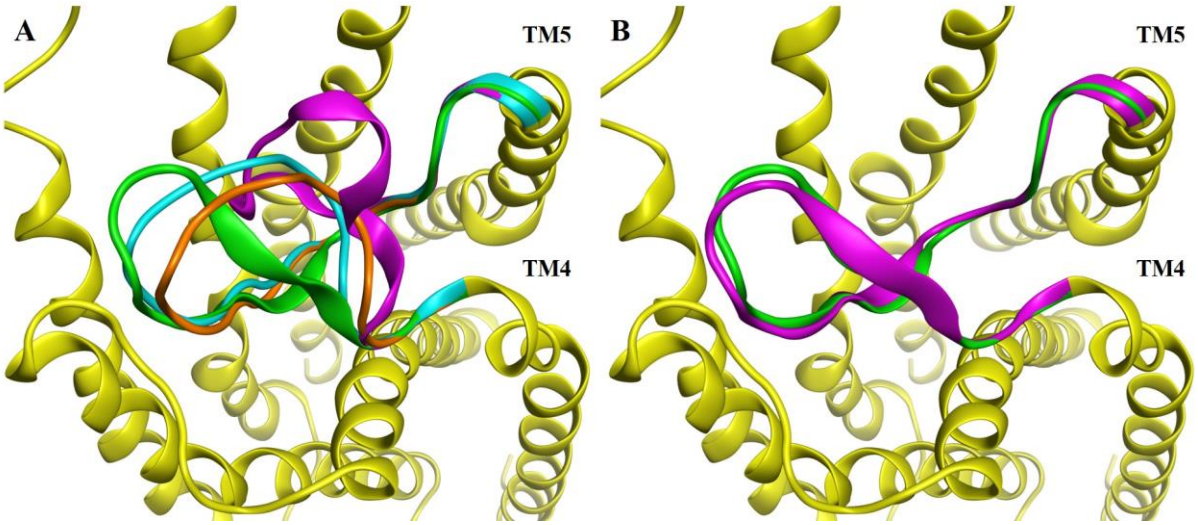

Figure 3.14: ECL2 models of P2YR1 superposed with reference structure.
(A) The LRM (orange), TSM (magenta), and LRM-T10 (cyan) out of 1000 models generated using NGK loop modeling had RMSD values of 2.44 Å, 6.53 Å, and 2.55 Å to the reference structure (green), respectively. (B) In this case, the LRM was also the TSM (magenta) when KICF loop modeling was used (LRM = TSM = 0.63 Å RMSD to reference structure).

In the case of the P2Y1R ECL2 target, the lowest RMSD model was also the top scored model. This was not the case for the AT2R ECL2 target, but the top scored model produced by the KICF method still had sub-angstrom accuracy (RMSD = 0.81 Å) when compared to the reference structure. On the other hand, the KICF and NGK algorithms produced top scored models with high RMSD values (~18 Å for both methods) compared to the ECL2 reference structure of β2AR (Figure 3.15). For the top 25 scored β2AR ECL2 models, none of the methods were able to generate loop models with RMSD values below 4 Å. Since this target had one of the longer loops in the benchmark, it is possible that increased sampling was necessary to produce models closer to the reference structure. To determine if increased sampling would drastically improve models, a second set of 4000 ECL2 models was generated for β2AR using the NGK loop modeling algorithm. From the larger set of models, the LRM had an RMSD of 2.93 Å to the reference target. While this is only a slight improvement from the LRM from the initial set of 1000 models, the TSM from the set of 4000 models had a RMSD value of 5.46 Å which is

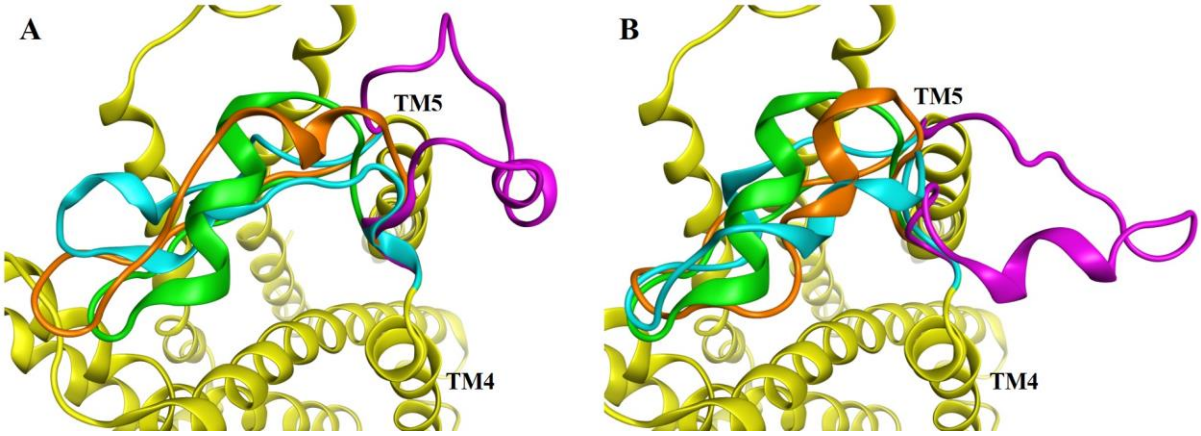

Figure 3.15: ECL2 models of B2AR superposed with reference structure.
(A) The LRM (orange), TSM (magenta), and LRM-T10 (cyan) out of 1000 models generated using NGK loop modeling had RMSD values of 3.84 Å, 18.1 Å, and 6.81 Å to the reference structure (green), respectively. (B) The LRM, TSM, and LRM-T10 generated using KICF loop modeling had RMSD values of 2.60 Å, 18.6 Å, and 5.04 Å to the reference structure, respectively.

drastically lower than the TSM from the initial set (18.1 Å). Out of the 4000-model set, the LRM-T10 and LRM-T25 both had an RMSD of 3.67 Å to the reference structure which was also lower relative to the LRM-T10 and LRM-T25 from the initial set (6.81 and 5.83 Å).

Group 4 targets from Table 3.2 were also modeled using the Rosetta loop modeling algorithms (Figures 3.16 and 3.17). The average LRM values resulting from modeling with the NGK, KICF, and CCD algorithms were 4.55, 4.98, and 5.38 Å, respectively. None of the loop modeling methods used was able to generate models with sub-angstrom or near-atomic accuracy to the reference structures. Based on these results, an increase in conformational sampling is likely necessary. Potentially due to inadequate sampling of conformational space, deficiencies in the scoring function to distinguish accurate models were evident for the longer loops in this benchmark set. The average ECL2 RMSD values for the TSM using the NGK, KICF and CCD algorithms were 13.18, 11.01, and 11.70 Å, respectively. A decrease in the average ECL2 RMSD values was observed for the LRM-T10 using the same three algorithms (7.65, 7.37, and 7.72 Å),

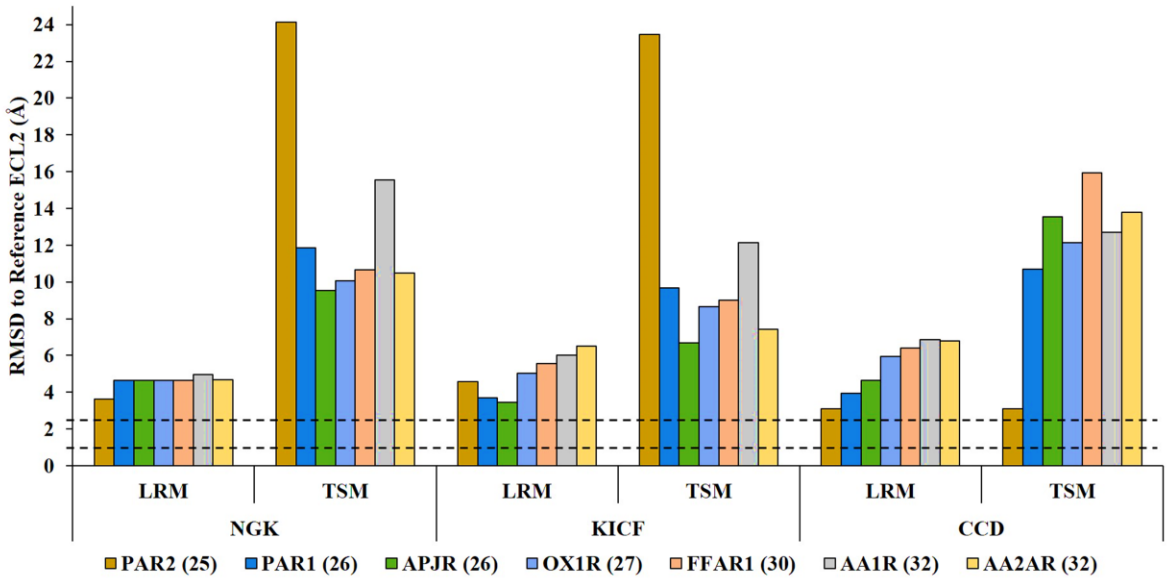

Figure 3.16: Group 4 results of the Lowest RMSD Models and Top Scored Models. The LRM and TSM are shown out of 1000 models generated for Group 4 ECL2 targets using NGK, KICF, and CCD. The upper and lower dotted lines represent the near-atomic and sub-angstrom accuracy thresholds, respectively.

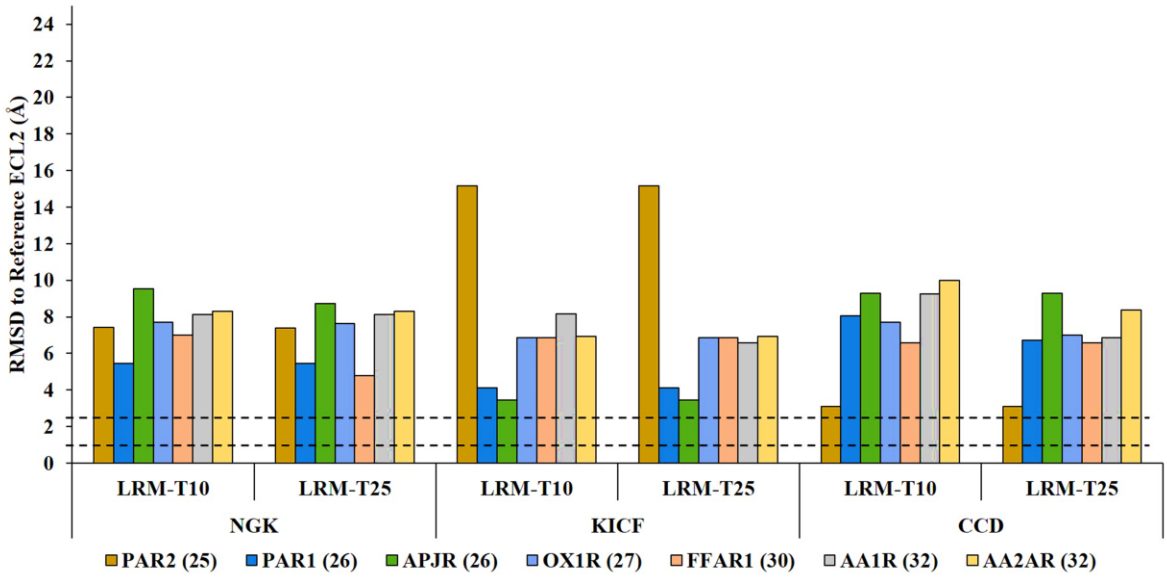

Figure 3.17: Group 4 results of the Lowest RMSD Models in the Top 10 and 25 Scored. The LRM-T10 and LRM-T25 are shown out of 1000 models generated for Group 4 ECL2 targets using NGK, KICF, and CCD. The upper and lower dotted lines represent the near-atomic and sub-angstrom accuracy thresholds, respectively.

but were still much higher than the 2.5 Å threshold.

Advances in GPCR structure prediction have been assessed in the community-wide GPCR DOCK experiments in 2008, 2010, 2013. In these experiments, researchers were tasked with modeling a target GPCR with a bound ligand prior to the publication of the target crystal structures. In GPCR DOCK 2010, two of the targets were the crystal structures of the DRD3/eticlopride (PDB: 3PBL) and CXCR4/IT1t (PDB: 3ODU) receptor-ligand complexes.[81] Overall, there were no models submitted for either target where the ECL2 had a backbone RMSD within 2.5 Å to the crystal structure. While the best DRD3 models had ECL2 RMSD values of 2.69 Å, CXCR4 was a more difficult modelling target where the two best models had ECL2 RMSD values of 4.32 and 6.61 Å. Based on all submitted models of DRD3 and CXCR4, the median RMSD values for ECL2 were 4.11 and 9.19 Å, respectively. While the RMSD values presented in this benchmark study show significant improvement in modeling the ECL2 of DRD3 and CXCR4 in terms of sampling, it is remarkable that the LRM obtained from KICF loop modeling with CXCR4 was ranked within the top 10 scored models and had an RMSD value of 0.50 Å. However, it should be noted that there is a significant advantage in modeling loops starting with a GPCR crystal structure versus a homology model. In general, a template-based GPCR model (without additional refinement) has an equivalent backbone structure to the aligned segments of the template. Model inaccuracies can originate from gaps in target-template sequence alignments which are frequently placed near the loops and anchor positions at the TM helical ends. Arora et al. showed that variations in loop anchor positions can have significant influence on modeling accuracy for GPCR loops.[299] Therefore, the results obtained from this benchmark study represent a best-case scenario for modeling GPCR loops where the starting structure contained no errors in the anchor positions.

3.4 Conclusion

Overall, the results from modeling Group 1 and 2 ECL2 targets showed that KICF sampled the most loop conformations with sub-angstrom and near-atomic accuracy to the reference structure (Figure 3.18). The NGK algorithm followed just behind KICF in terms of building loop models with sub-angstrom accuracy, but both algorithms had the same number of cases of modeling loops with near-atomic accuracy. Although the CCD algorithm was not able to build any loop models with sub-angstrom accuracy for the targets in Group 1 and 2, CCD produced models with near-atomic accuracy for nine of the fourteen targets. The results from modeling Group 3 ECL2 targets showed that only the KICF algorithm was able to sample loop models with sub-angstrom accuracy.

Regardless of loop length, the RMSD of the TSM was much higher than the LRM in most cases. This was also observed in another benchmark study targeting 13 GPCR ECL2 using the CABS modeling software.[300] Overall, these data suggest that the KIC with Fragments and Next Generation KIC methods within Rosetta perform better than Cyclic Coordinate Descent method or the *de novo* search method within MOE Loop Modeler for loops with up to 20 residues (Groups 1 & 2). For longer loops (20 – 32 residues), KIC with Fragments outperforms all the other methods. Out of all 28 GPCR loops modeled, KICF generated the most models under 2.5 Å out of 1000 produced total. However, for targets analogous to Group 4 (loop lengths 25-32 residues) it is recommended that a greater number of models be produced (i.e. > 4000).

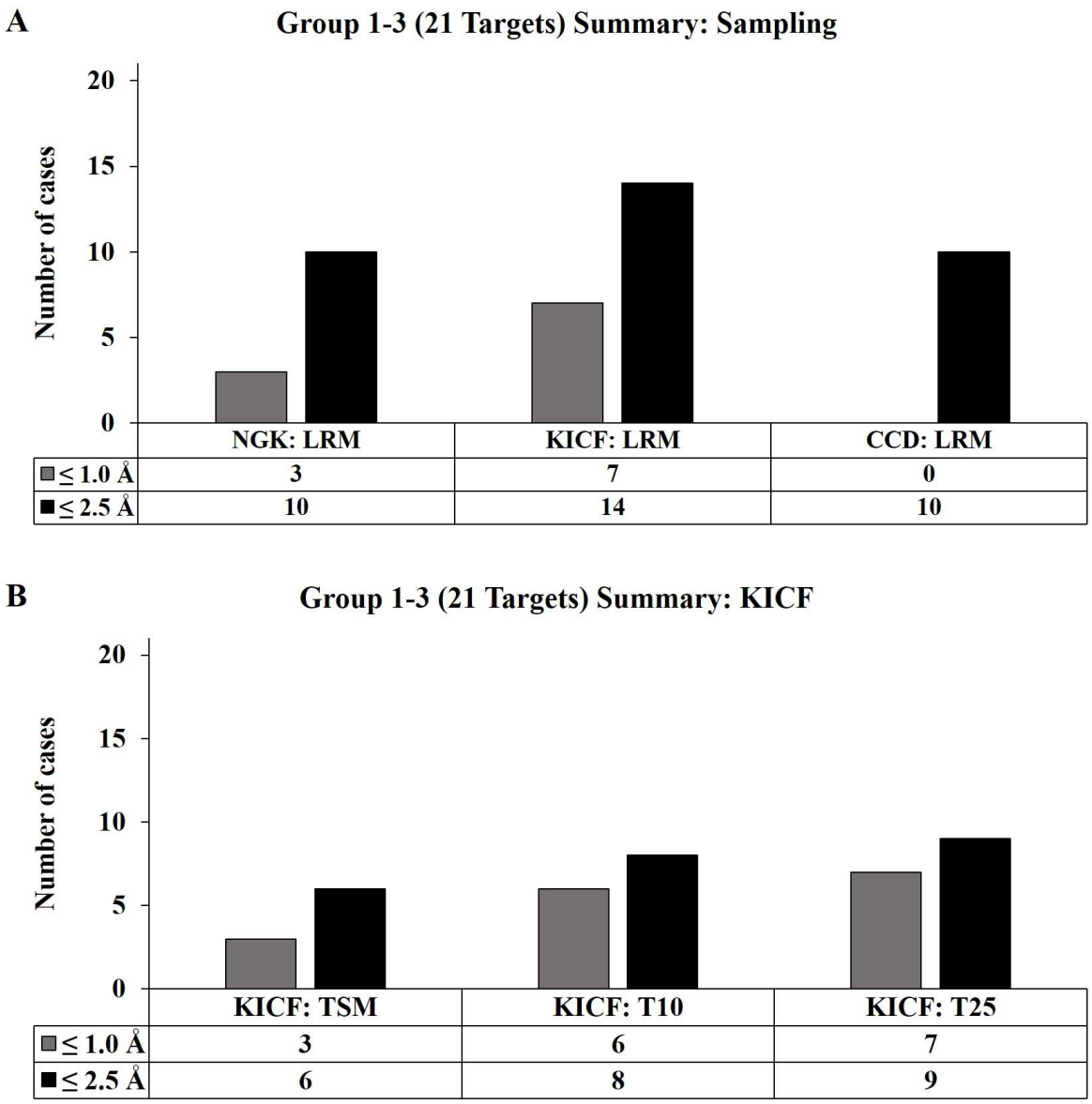

Figure 3.18: Overview of ECL2 models with sub-angstrom or near-atomic accuracy. Model accuracy is assessed by the RMSD to the reference ECL2 being under 1.0 or 2.5 Å for sub-angstrom or near-atomic accuracy, respectively. (A) For the three Rosetta loop modeling algorithms, the number of cases where the LRM had RMSD values with sub-angstrom or near-atomic accuracy to the reference structures in Groups 1-3. None of the loop models produced from the Group 4 targets had RMSD values below 2.5 Å. (B) For the Rosetta KICF loop modeling algorithm, the number of cases where the TSM, LRM-T10, and LRM-T25 had RMSD values with sub-angstrom or near-atomic accuracy to the reference structures in Groups 1-3.

Since none of the tested loop modeling algorithms were able to sample models within the 2.5 Å RMSD threshold for the Group 4 ECL2 targets, it is likely that increased sampling is necessary.

For loops that lacked models with RMSDs within 2.5 Å in the Top 1, 10, or 25 (out of 1000 total models) scored models but had models with RMSDs within 2.5 Å in the LRM category; adjustments to the scoring function, refinement methods, or loop structure environment could improve loop prediction.

Based on these data, it is recommended that researchers modeling loops of unknown structure should at least retain the top 10 scored models using the most recent Rosetta energy function for further refinement or validation. This benchmark study suggests that expanding the number of retained top scored models to the top 25 does not consistently yield a more accurate model that would compensate for the much larger computational burden of working with 25 model structures in post-modeling procedures versus 10 models.

Chapter 4

Future Considerations

4.1 Potential improvements to loop modeling methods for GPCR

Additional Rosetta loop modeling jobs are in progress for some of the group 4 benchmark targets (25-32 residue ECL2) to further test the impact of increasing the number of final models generated for very long loops. If increased sampling does not drastically improve loop modeling accuracy with respect to the native structure, then alternative approaches within the Rosetta modeling suite could be explored in future benchmarks. A potential test with KICF could evaluate loop modeling accuracy when using fragments with varied lengths). Previous analysis of the Rosetta *de novo* structure prediction method found that models were sampled closer to the native structure when using long (5-19 residues) and short (3-9 residues) sets of varied fragment sizes instead of the standard 3 and 9 residue fragments.[301]

In a separate set of tests, loop modeling with target GPCR models embedded in a virtual membrane environment guided by an energy function specific to membrane proteins (implemented with key elements available in RosettaMP.[67] Incorporating these features could disfavor the sampling and/or scoring of final loop models that are positioned away from the central helical bundle (similar to the TSM in Figure 3.15 A and B). While the field of protein structure prediction has progressed in terms of modeling accuracy over the past several years, there is still room for improvement in developing more efficient sampling and scoring strategies for GPCR structural features.

4.2 Alternative expression systems and WS mutant designs for GPCR.

By the end of Chapter 2, the evidence was clear that the strategies implemented to liberate bound WS-GPCR from GroEL were insufficient. In an alternative approach, we decided to test different protein expression systems that could be more conducive to recombinant WS-GPCR production. The GroEL/ES system optimally functions at a temperature of ~30°C, but only retains about 30% of its folding activity at 12°C.[302] Since our expression protocol involved culture growth at 18°C, it is possible that the GroEL/ES system was a major bottleneck for protein expression and proper protein folding. Thus, we tested WS-GPCR expression in ArcticExpress(DE3) bacterial cells (Agilent, Santa Clara, CA), which constitutively co-express the cold-adapted chaperonins, Cpn60 and Cpn10, that have 54% and 74% protein sequence identity to GroEL and GroES, respectively. At low temperatures in the 4 – 12°C range, Cpn60/10 showed much higher protein folding activities than the GroEL/ES counterparts.[303] Unfortunately, SUMO-β2AR:M13-EmGFP expression in ArcticExpress(DE3) cells resulted in lower levels of target protein than our previous protocols (data shown in appendix Figure A.1).

Finally, we wanted to test the WS-GPCR constructs, His-β2AR:M13/M15-FLAG, with insect cell expression systems. To avoid any similar problems with the eukaryotic homolog of GroEL, HSP60, we decided to use the pMIB vector (ThermoFisher Scientific, Hampton, NH) transient transfection system which results in the target protein with an Nt-signal peptide (Honeybee melittin secretion signal). This feature allows for secreted expression of target proteins into the cell culture media. Purifying recombinant proteins from the cell culture media can have the benefit of having reduced amounts of background proteins that will non-specifically bind to the column. For this portion of the project, Dr. Truc Chi T. Pham was essential in guiding the insect cell culture handling and maintenance work, in addition to performing the first round

of lipid-mediated transient transfections in *Spodoptera frugiperda* (Sf9) and High Five™ cell lines. After a few trial transfections, we were unable to observe expression of either the M13 or M15 WS-GPCR constructs. Additionally, Dr. Pham performed multiple expression trials with both M13 and M15 constructs in the same insect cell lines mentioned previously but instead, using a baculovirus expression system (Bac-to-Bac; ThermoFisher Scientific, Hampton, NH) which uses a recombinant virus to carry the gene of interest, infect the insect cell hosts, and drive protein expression.[304] After expression and purification by Ni-NTA, she performed in-gel trypsin digestion and LC-MS/MS to identify the proteins in the elution fractions. SDS-PAGE bands corresponding to proteins of MW ~48 and ~70 kD were identified as most likely His-β2AR:M15-FLAG and HSP70, respectively. Given that protein yields with the insect expression systems were low (or unobservable) and the potential presence of another chaperone contaminant, our group decided to suspend the WS-GPCR project until more funding was available.

Future WS-GPCR designs with additional mutations could be implemented to further disrupt hydrophobic regions on the outer surface of the TM domains. This strategy aims to decrease the binding affinity of WS-GPCR to GroEL (or re-bind to GroEL after substrate release). Thus, GroEL removal from purified WS-GPCR may be possible using some of the strategies discussed in Chapter 2 or, ideally, unnecessary due to the absence of GroEL after purification.

The Protein Patch Analysis tool in MOE was used to calculate and visualize hydrophobic surface patches to identify regions where potential protein-protein interactions occur via (i.e. GroEL substrate binding interactions). Figure 4.1 shows the hydrophobic surface patches of the

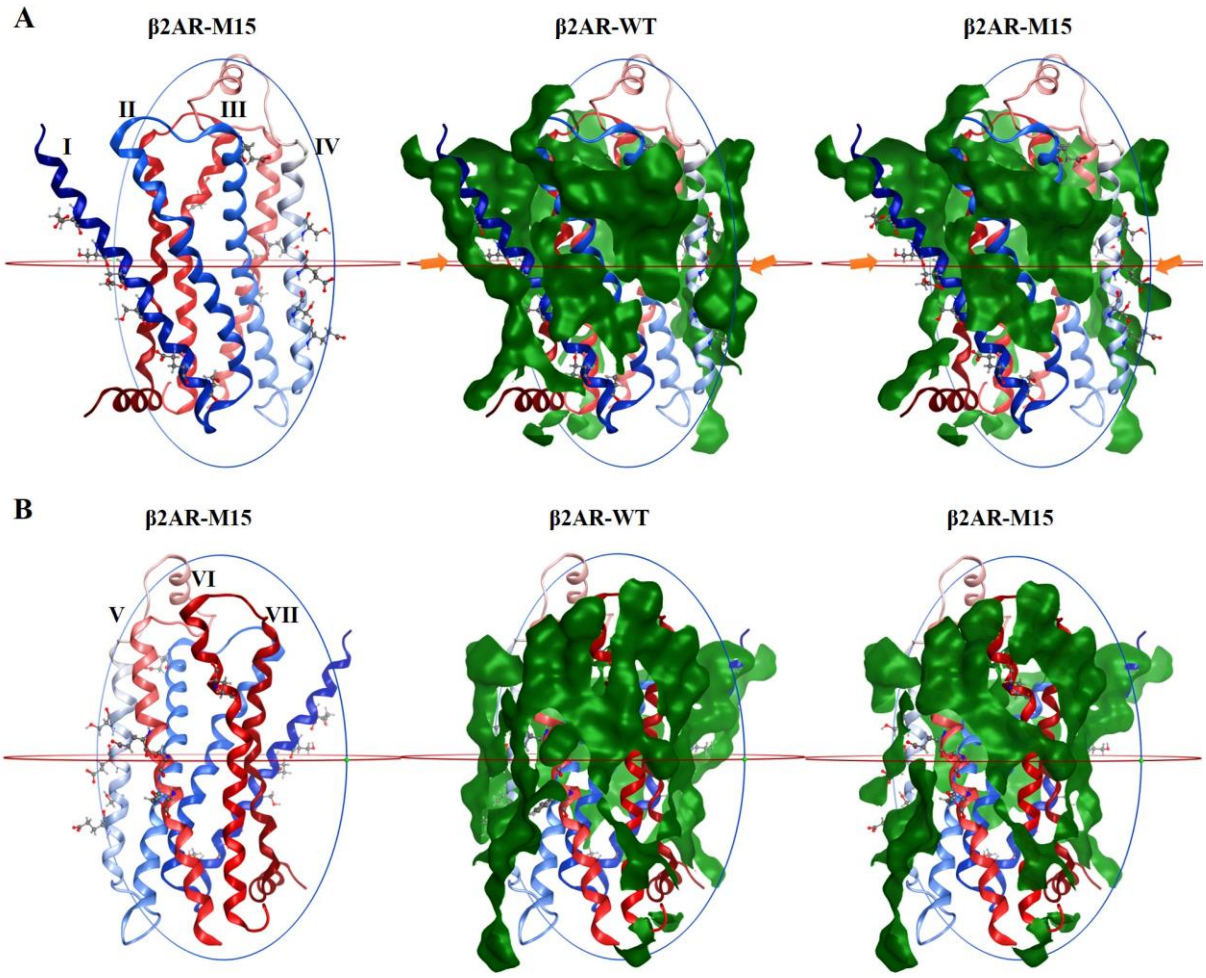

Figure 4.1: Protein Patch surface analysis of β2AR-WT and β2AR-M15 using MOE. A crystal structure of the β2AR (PDB ID: 2RH1) was used to create the M15 mutant using MOE. The atoms shown represent the mutated residues of the M15 design. The Protein Patch Analysis tool in MOE was used to calculate and visualize hydrophobic surface patches (green) to show significant regions where potential protein-protein interactions occur. For the wild-type and M15 mutant structures, (A) the TM1-4 side view and (B) TM5-7 side view are displayed and correspond to the alternate views shown in Figure 4.2.

β2AR-WT and the resulting patches after applying the M15 mutations on three-dimensional structures of the proteins. The orange arrows in Figure 4.1 A highlight significant regions where the M15 mutations disrupt hydrophobic surface patches with respect to β2AR-WT. Figure 4.2 shows the hydrophobic surface patches of β2AR-WT and β2AR-M15 represented within two-dimensional map projections generated using MOE. In all figures containing 2D maps, two viewing perspectives for each protein are provided for the TM1-4 and TM5-7 sides. For greater

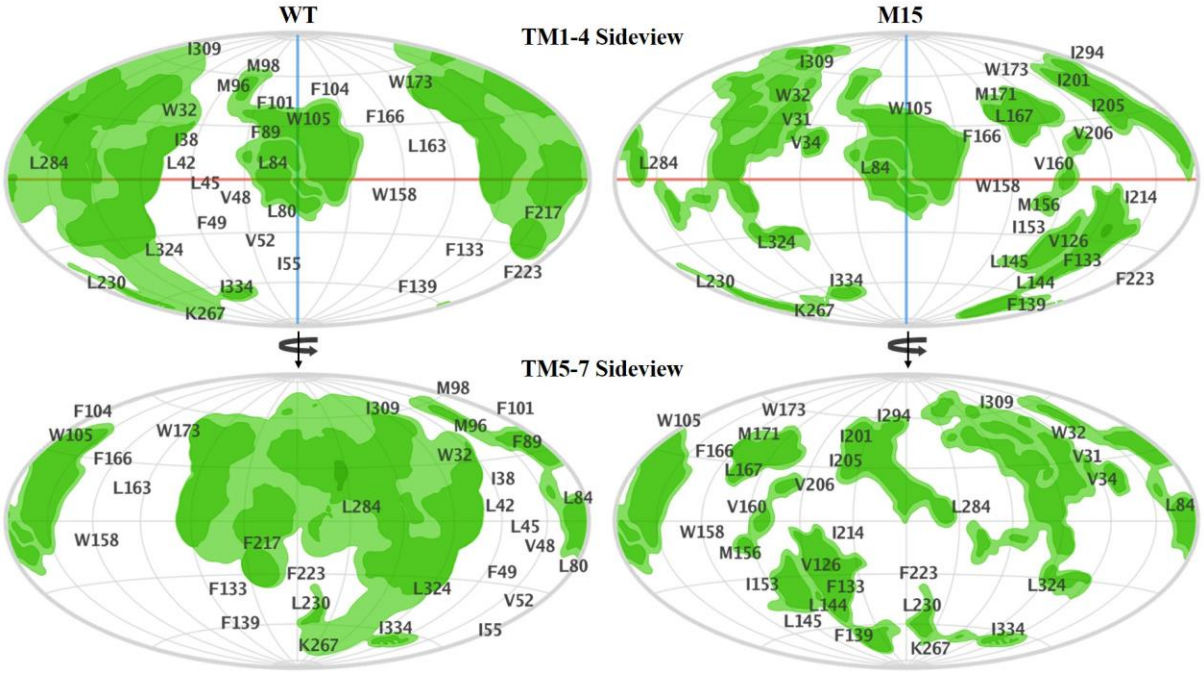

Figure 4.2: Protein Patch 2D map projection of hydrophobic surfaces on β2AR-WT and β2AR-M15.
The 2D Maps application within MOE 2016.08 creates a two-dimensional projection of the Protein Patches (Figure 4.1). The hydrophobic surface patches (green) are shown for the β2AR-WT (left) and β2AR-M15 (right) proteins when viewed from the orientations depicted in Figure 4.1 A and B. The 2D projection is dependent on the protein structure's orientation within the view window of MOE. After setting the desired view, the equator (red) and central meridian (blue) of the 2D projection are established.

detail, Table 4.1 lists the approximate area (Å^2) and average energy (kcal/mol) of significant hydrophobic surface patches as well as the individual residues that contribute to the patches. Table 4.2 shows the TM domain distribution of residues attributed to the four largest hydrophobic surface patches by area for the β2AR-M15 mutant. The data in Table 4.1 suggests that the M15 mutations disrupt the largest hydrophobic area associated with the WT protein. The largest hydrophobic patch associated with β2AR-WT has an area of 3100 Å^2 which is a little over 2.5 times greater than the largest patch associated with the β2AR-M15 mutant.

Table 4.1: Protein Patch surface analysis of β2AR-WT and β2AR-M15.
The largest hydrophobic surface patches by area with the associated patch amino acid residues are listed in descending order for the β2AR-WT and β2AR-M15 constructs.

WT vs M15 Surfaces	Hydrophobic Patch	Area (Å^2)	Avg. Energy per Å^2 (kcal/mol)
	Patch Residues		
β2AR-WT Protein Patches	**1**	**3100**	-0.15
	V31 W32 V34 M36 I38 V39 M40 L42 I43 L45 F49 G50 L53 T56 K60 I94 T110 V114 V117 C125 V126 V129 Y132 F133 P138 F139 L144 L145 K149 V152 I153 L155 M156 V157 I159 V160 L163 T164 L167 P168 M171 W173 F193 T195 Q197 Y199 I201 I205 V206 Y209 V210 P211 L212 V213 I214 M215 V216 F217 V218 Y219 R221 F223 A226 K263 F264 K273 G276 I277 M279 T281 T283 L284 C285 W286 L287 P288 F289 F290 I291 V292 I294 V295 H296 V297 I298 Q299 L302 I303 R304 K305 V307 Y308 I309 L310 L311 W313 I314 Y316 V317 F321 L324 I325 C327 R328 R333 F336 L339 L340 L342		
	2	**690**	-0.15
	V44 V48 V52 I55 A59 T73 C77 L80 V81 L84 A85 P88 F89 A92 M96 M98 T100 F101 F104 W105 F108 I112 L115 W158 F166 Q170 T189		
	3, 4	**60, 50**	-0.13, -0.13
	3 (K267 P330 I334), 4 (L230 L266)		
β2AR-M15 Protein Patches	**1**	**1170**	-0.14
	W32 M36 V39 M40 I43 L53 I94 T110 V114 V117 F193 T195 Y199 M215 Y219 F223 A226 K263 F264 K273 G276 I277 M279 T281 C285 W286 P288 F289 F290 V292 V295 H296 Q299 L302 I303 R304 K305 V307 Y308 I309 L310 L311 W313 I314 Y316 V317 F321 L324 I325 C327 R328 R333 F336 L340 L342		
	2	**610**	-0.15
	V44 V48 V52 I55 T73 C77 L80 V81 L84 A85 P88 F89 A92 M96 M98 T100 F101 W105 F108 I112 L115 W158 F166 Q170 T189		
	3	**370**	-0.15
	Q197 I201 I205 Y209 L212 V216 T283 L284 L287 I294 V297 I298		
	4	**340**	-0.15
	C125 V129 Y132 F133 P138 F139 L144 P211 I214 V218 R221		
	5, 6	**190, 100**	-0.16, -0.13
	5 (L167 P168 M171 W173 Y199), 6 (M156 V160 T164 V206)		
	7, 8	**100, 90**	-0.14, -0.13
	7 (V126 L145 K149 I153 V157), 8 (V31 V34)		
	9, 10	**60, 50**	-0.13, -0.13
	9 (K267 P330 I334), 10 (L230 L266)		

Table 4.2: Hydrophobic surface patch distribution across TM domains of β2AR-M15. The residues of the four largest hydrophobic patches are grouped by TM domain for β2AR-M15. The Ballesteros-Weinstein residue numbers are given in parentheses.

β2AR-M15	Patch # 1	Patch # 2	Patch # 3	Patch # 4
TM1	W32 (1.31), M36 (1.35), V39 (1.38), M40 (1.39), I43 (1.42), L53 (1.52)	V44 (1.43), V48 (1.47), V52 (1.51), I55 (1.54)	n/a	n/a
TM2	I94 (2.65)	T73 (2.44), C77 (2.48), L80 (2.51), V81 (2.52), L84 (2.55), A85 (2.56), P88 (2.59), F89 (2.60), A92 (2.63), M96 (2.67)	n/a	n/a
TM3	T110 (3.29), V114 (3.33), V117 (3.36)	W105 (3.24), F108 (3.27), I112 (3.31), L115 (3.34)	n/a	C125 (3.44), V129 (3.48), Y132 (3.51), F133 (3.52)
TM4	n/a	W158 (4.50), F166 (4.58), Q170 (4.62)	n/a	n/a
TM5	Y199 (5.38), M215 (5.54), Y219 (5.58), F223 (5.62), A226 (5.65)	n/a	Q197 (5.36), I201 (5.40), I205 (5.44), Y209 (5.48), L212 (5.51), V216 (5.55)	P211 (5.50), I214 (5.53), V218 (5.57), R221, (5.60)
TM6	K273 (6.35), G276 (6.38), I277 (6.39), M279 (6.41), T281 (6.43), C285 (6.47), W286 (6.48), P288 (6.50), F289 (6.51), F290 (6.52), V292 (6.54), V295 (6.57), H296 (6.58)	n/a	T283 (6.45), L284 (6.46), L287 (6.49), I294 (6.56), V297 (6.59), I298 (6.60)	n/a
TM7	K305 (7.32), V307 (7.34), Y308 (7.35), I309 (7.36), L310 (7.37), L311 (7.38), W313 (7.40), I314 (7.41), Y316 (7.43), V317 (7.44), F321 (7.48), L324 (7.51), I325 (7.52), C327 (7.54), R328 (7.55)	n/a	n/a	n/a

To improve upon the M15 design, a new series of WS mutants are proposed in Table 4.3. The development of the M21, M27, and M33 mutation sets were designed to improve the M15 design by incrementally disrupting the remaining hydrophobic patches identified in Table 4.1 and 4.2 with a minimal number of mutations. The additional mutations incorporated into the M21, M27, and M33 designs target some of the outward-facing residues that compose Patch 1 (M21), Patch 2 (M27), and Patches 3-4 (M33), respectively. It is important to note that not all of the listed patch residues in Tables 4.1-4.3 are considered outward-facing. The methods for determining outward-facing residues and choosing amino acid point mutations that were described in Chapter 2 were applied in these proposed WS mutant designs. The three proposed WS mutants were generated from the β2AR crystal structure (PDB: 2RH1[34]) by using MOE to apply the point mutations listed in Table 4.3. Protein patch analysis was performed on each WS mutant to calculate the hydrophobic surface patches (Patch details in Table 4.4) and generate 2D maps (Figure 4.3) following the same process applied previously for β2AR-WT and β2AR-M15. For each of the new WS mutants, the largest hydrophobic surface patch decreases in area relative to the corresponding patch in the preceding mutation set. Based on these data, the proposed M33 mutant would display smaller area hydrophobic surface patches compared to the initial M15 design. The largest hydrophobic patch on the M33 mutant has an area of 350 $Å^2$ which is greatly diminished compared to the surface patches associated with WT and M15 proteins (3100 and 1170 $Å^2$, respectively). With respect to the total combined area of hydrophobic patches on the protein surface (Table 4.5), the M15 mutant exhibits hydrophobic patches with smaller total area by 820 $Å^2$ relative to the WT. In comparison, the total combined area of hydrophobic patches for the M33 mutant is almost 900 $Å^2$ smaller than the M15.

Table 4.3: Future WS-GPCR mutant design sets applied to β2AR.
As extensions of the initial M15 design, the M21, M27, and M33 mutation sets incorporate the point mutations from each of their respective predecessors and an additional six mutations. The underlined residues represent the additional point mutants relative to the preceding mutation set.

β2AR-M15		β2AR-M21		β2AR-M27		β2AR-M33	
1.37 E	I38	1.31 E	W32	1.31 E	W32	1.31 E	W32
1.41 S	L42	1.37 E	I38	1.37 E	I38	1.37 E	I38
1.44 E	L45	1.41 S	L42	1.38 S	V39	1.38 S	V39
1.48 S	F49	1.44 E	L45	1.41 S	L42	1.41 S	L42
1.55 E	T56	1.48 S	F49	1.44 E	L45	1.44 E	L45
1.58 S	A59	1.55 E	T56	1.48 S	F49	1.48 S	F49
3.23 E	F104	1.58 S	A59	1.51 E	V52	1.51 E	V52
4.44 E	V152	3.23 E	F104	1.55 S	T56	1.55 S	T56
4.47 S	L155	4.44 E	V152	1.58 E	A59	1.58 E	A59
4.51 E	I159	4.47 S	L155	2.55 E	L84	2.55 E	L84
4.55 S	L163	4.51 E	I159	2.63 E	A92	2.63 E	A92
5.49 E	V210	4.55 S	L163	2.67 S	M96	2.67 S	M96
5.52 S	V213	5.49 E	V210	3.23 E	F104	3.23 E	F104
5.56 E	F217	5.52 S	V213	3.24 S	W105	3.24 S	W105
6.53 E	I291	5.56 E	F217	4.44 E	V152	3.52 E	F133
		6.39 E	I277	4.47 S	L155	4.44 E	V152
		6.53 E	I291	4.51 E	I159	4.47 S	L155
		6.57 S	V295	4.55 S	L163	4.51 E	I159
		7.37 E	L310	5.49 E	V210	4.55 S	L163
		7.48 S	F321	5.52 S	V213	5.36 E	Q197
		7.51 E	L324	5.56 E	F217	5.40 S	I201
				6.39 E	I277	5.44 E	I205
				6.53 E	I291	5.49 E	V210
				6.57 S	V295	5.52 S	V213
				7.37 E	L310	5.55 S	V216
				7.48 S	F321	5.56 E	F217
				7.51 E	L324	6.39 E	I277
						6.46 S	L284
						6.53 E	I291
						6.57 S	V295
						7.37 E	L310
						7.48 S	F321
						7.51 E	L324

Table 4.4: Protein Patch surface analysis of β2AR-21, β2AR-M27, and β2AR-M33.

WS-GPCR Surfaces	Hydrophobic Patch	Area (Å^2)	Avg. Energy per Å^2 (kcal/mol)
	Patch Residues		
β2AR-M21 Protein Patches	**1**	**610**	-0.15
	V44 V48 V52 I55 T73 C77 L80 V81 L84 A85 P88 F89 A92 M96 M98 T100 F101 W105 F108 I112 L115 W158 F166 Q170 T189		
	2	**400**	-0.14
	M36 V39 M40 I43 I94 L95 T110 V114 V117 F193 T195 Y199 W286 F289 F290 H296 K305 Y308 I309 W313 Y316 V317		
	3, 4	**370, 340**	-0.15, -0.15
	3 (Q197 I201 I205 Y209 L212 V216 T283 L284 L287 I294 V297 I298), 4 (C125 V129 Y132 F133 P138 F139 L144 P211 I214 V218 R221)		
	5, 6, 7	**210, 190, 140**	-0.14, -0.16, -0.15
	5 (C285 P288 V292 Q299 L302 I303 R304 V307 E310 L311 I314), 6 (L167 P168 M171 W173 Y199), 7 (L53 C327 R333 F336 L340 L342)		
β2AR-M27 Protein Patches	**1**	**370**	-0.15
	Q197 I201 I205 Y209 L212 V216 T283 L284 L287 I294 V297 I298		
	2	**350**	-0.14
	M36 M40 I43 I94 L95 T110 V114 V117 F193 T195 Y199 W286 F289 F290 H296 K305 Y308 I309 W313 Y316 V317		
	3, 4	**340, 210**	-0.15, -0.14
	3 (C125 V129 Y132 F133 P138 F139 L144 P211 I214 V218 R221), 4 (C285 P288 V292 Q299 L302 I303 R304 V307 E310 L311 I314)		
	5, 6, 7	**210, 190, 140**	-0.16, -0.16, -0.15
	5, (V81 A85 F108 I112 L115 W158 F166 Q170), 6 (L167 P168 M171 W173 Y199), 7 (L53 C327 R333 F336 L340 L342)		

Table 4.4 (continued): Protein Patch surface analysis of β2AR-21, β2AR-M27, and β2AR-M33.

WS-GPCR Surfaces	Hydrophobic Patch	Area (Å^2)	Avg. Energy per Å^2 (kcal/mol)
	Patch Residues		
β2AR-M33 Protein Patches	**1**	**350**	-0.14
	M36 M40 I43 I94 L95 T110 V114 V117 F193 T195 Y199 W286 F289 F290 H296 K305 Y308 I309 W313 Y316 V317		
	2, 3	**210, 210**	-0.14, -0.16
	2 (C285 P288 V292 Q299 L302 I303 R304 V307 E310 L311 I314), 3 (V81 A85 F108 I112 L115 W158 F166 Q170)		
	4, 5	**190, 180**	-0.16, -0.14
	4 (L167 P168 M171 W173 Y199), 5 (E205 Y209 L212 T283 L287 I294 I298)		
	6, 7	**140, 140**	-0.14, -0.15
	6 (C125 V129 Y132 P211 I214 V218 R221), 7 (L53 C327 R333 F336 L340 L342)		

Table 4.5: Total area of all hydrophobic patches for β2AR-WT and WS β2AR mutants.

Construct	Total hydrophobic patch area
β2AR-WT	**3900 Å^2**
β2AR-M15	**3080 Å^2**
β2AR-M21	**2730 Å^2**
β2AR-M27	**2460 Å^2**
β2AR-M33	**2190 Å^2**

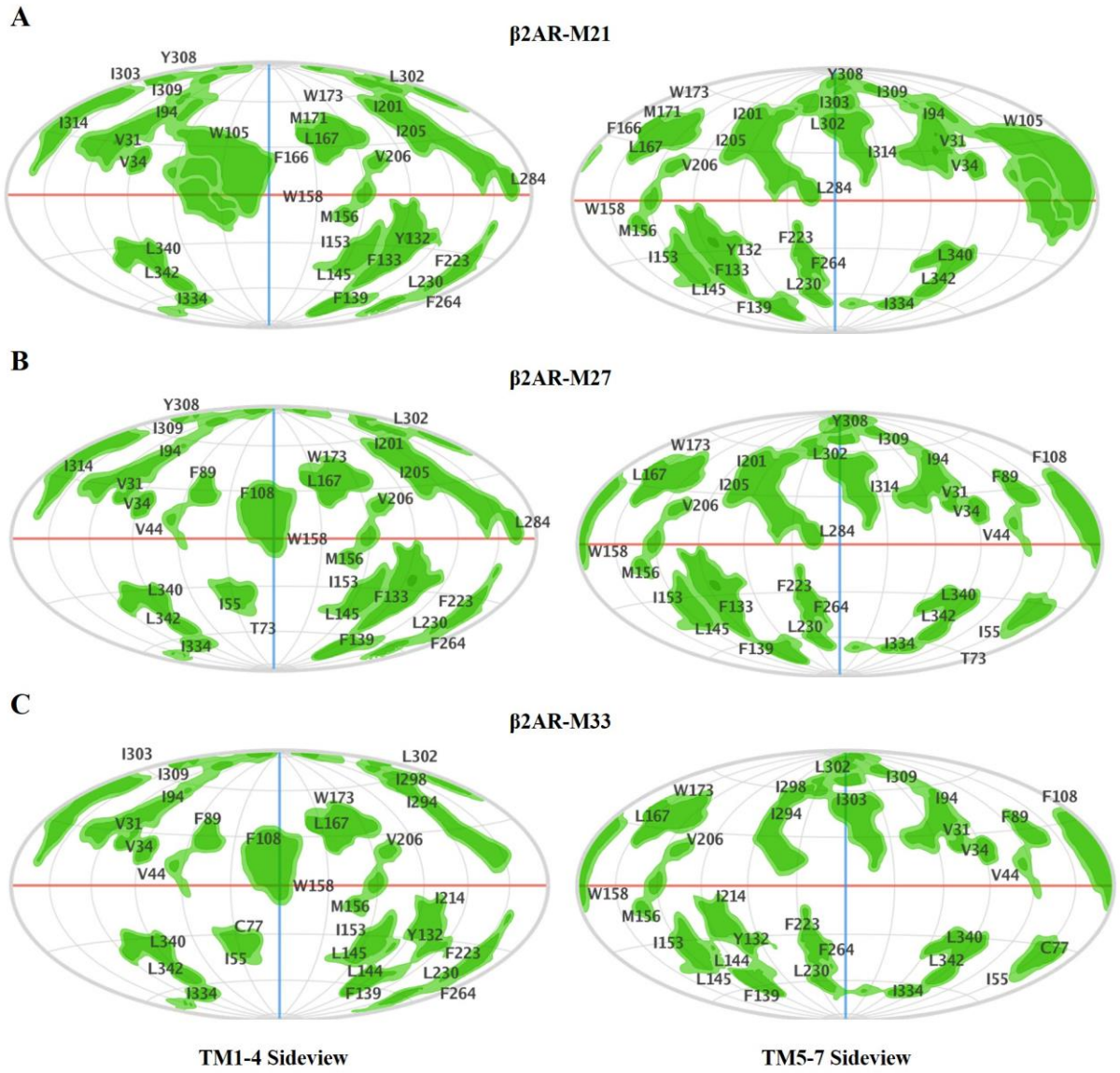

Figure 4.3: Protein Patch 2D map projections of hydrophobic surfaces on future β2AR WS mutant sets.

The hydrophobic surface patches (green) are shown for (A) β2AR-M21, (B) β2AR-M27, and (C) β2AR-M33 WS mutant proteins when viewed from the orientations depicted in Figure 4.1 A and B.

In conclusion, additional WS mutations that disrupt remaining hydrophobic surface patches could be implemented in future work. The proposed WS mutant series, M21, M27, and M33, could be tested in preliminary expression and purification trials using the same methods outlined in Chapter 2 to probe for GroEL/chaperone presence in the elution fractions. To estimate whether the maximum number of proposed mutations in the M33 design would be sufficient in disrupting hydrophobic patches, we applied the MOE Protein Patch surface analysis tools to compare the WT mu opioid receptor (OPRM) with the second generation WS OPRM mutant designed by the Liu group (described in Chapter 2) which contained 46 mutations (OPRM-M46). Although an initial exposure to 0.1% SDS was required to solubilize and purify OPRM-M46, the purified protein was soluble in aqueous solution (130 mM NaCl, 20 mM $NaHPO_4$) and retained ligand binding properties comparable to the native receptor.[97,98] The OPRM crystal structure (PDB: 4DKL[101]) was used to construct the OPRM-M46 mutant in MOE. The data from the hydrophobic patch analysis and 2D map projections are shown in Tables 14-15 and Figure 4.4. The hydrophobic surface patches remaining on the OPRM-M46 mutant relative to its WT counterpart is proportional to the patches found on the β2AR-M33 mutant. The OPRM-M46 mutations reduce the total hydrophobic surface patch area on OPRM-WT from 4,560 $Å^2$ to 2,650 $Å^2$. Similarly, the β2AR-M33 mutations reduce the total hydrophobic surface patch area on β2AR-WT from 3,900 $Å^2$ to 2,190 $Å^2$. Comparisons between different receptors with different WS mutation come with an obvious caveat that the β2AR-M33 mutant may still behave differently than the reported OPRM-M46. Thus, experimental validation will be necessary to assess the viability of the alternative WS mutation designs.

Table 4.6: Protein Patch surface analysis of OPRM-WT.

Hydrophobic Patch #	**Area (Å^2)**	**Avg. Energy per Å^2 (kcal/mol)**
1	**2700**	-0.15
T67 I71 L74 I77 V78 V80 V81 F84 L88 V89 Y91 V92 R95 F152 F156 C159 V163 Y166 I167 V169 H171 K174 F178 C190 L194 I198 P201 V202 M205 F221 P224 W226 Y227 W228 L231 L232 K233 I234 C235 F237 I238 F239 F241 I242 M243 V245 L246 I247 I248 V250 C251 Y252 L254 M255 L257 L259 K271 L275 T279 R280 L283 V284 V286 V288 I290 C292 T294 P295 I298 Y299 I301 I302 A304 L305 I306 I308 P309 F320 C321 A323 L324 T327 C330 L331 V334 L335 A337 F338 L339 F343 K344 F347 R348 F350 C351 I352		
2	**1140**	-0.15
M65 V66 A68 I69 T70 M72 C79 L83 F87 M90 I93 V94 K100 I105 F108 L112 A115 L116 T120 P122 F123 V126 L129 M130 T132 P134 F135 I138 L139 K141 I142 I146 Y149 P181 K185 W192 L200 M203 F204		
3	**230**	-0.14
Q124 W133 V143 I144 Y148 M151 C217 L219 F221 E229 L232 K233 V236 F237 W293 I296 H297 V300 I301 W318 I322 Y326		
4, 5, 6	**100, 100, 90**	-0.16, -0.14, -0.13
4 (F313 V316 S317), 5 (P172 V173 L176), 6 (V189 I193)		
7, 8, 9	**70, 70, 60**	-0.15, -0.14, -0.14
7 (T208 Y210 I215), 8 (I186 C190), 9 (I256)		

Table 4.7: Protein Patch surface analysis of OPRM-M46.

Hydrophobic Patch #	**Area (Å^2)**	**Avg. Energy per Å^2 (kcal/mol)**
1	**660**	-0.15
M65 V66 A68 I69 K70 C79 M90 I93 V94 K100 I105 F108 K112 L116 T120 P122 F123 V126 L129 M130 T132 P134 F135 L139 K142 I146 Y149 P181 K185 W192		
2	**430**	-0.14
V169 F237 F241 V245 I248 Y252 M255 L259 K271 L275 T279 L283 V286 I290 P295 K298 Y299 I301 L305 I306 I308 P309 Q314 C321		
3	**370**	-0.15
F152 F156 C159 K190 L194 I198 K205 F221 P224 W226 Y227 W228 K231 L232 K239		
4, 5, 6	**240, 230, 230**	-0.14, -0.14, -0.14
4 (L88 V89 E91 V92 R95 Y96 Y336 F343 F347 R348 F350 C351 I352), 5 (Q124 W133 V143 I144 Y148 M151 C217 L219 F221 E229 L232 K233 V236 F237 W293 I296 H297 V300 I301 W318 I322 Y326), 6 (V163 Y166 I167 H171 K174 F178 K247 C251 L254 L257)		
7, 8, 9	**100, 100, 90**	-0.14, -0.14, -0.16
7 (R280 V284 V288 C292 E331 L335 L339), 8 (P172 V173 L176), 9 (F313 V316 S317)		
10, 11, 12	**70, 70, 60**	-0.16, -0.14, -0.14
10 (T208 Y210 I215), 11 (I138 K141), 12 (I256)		

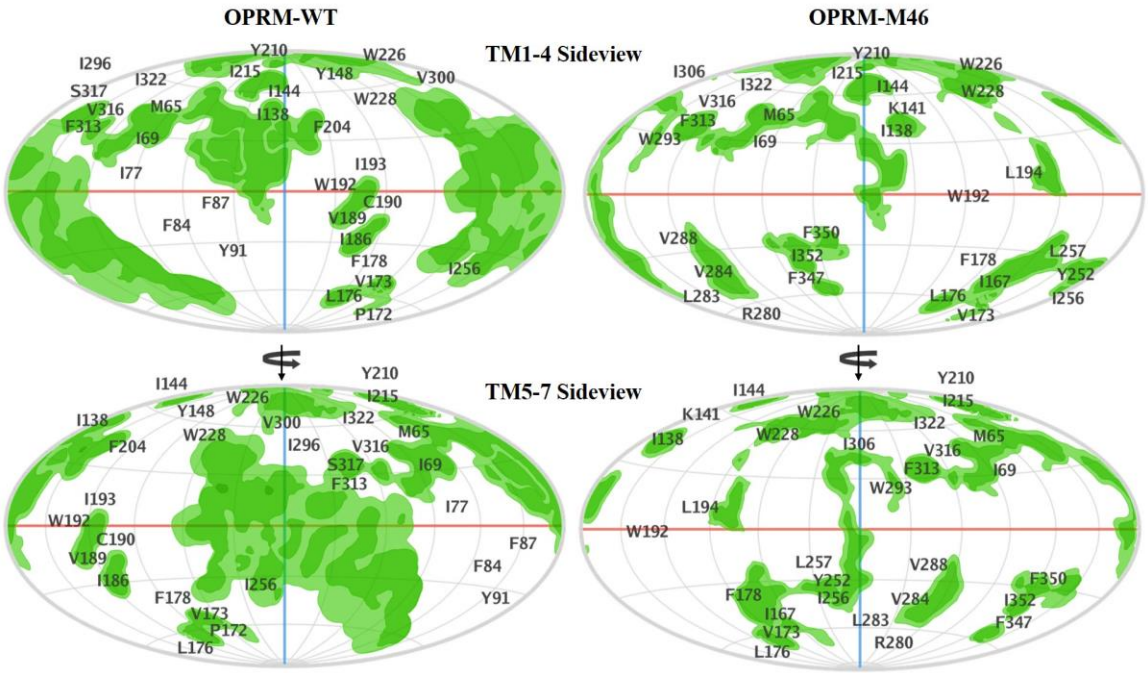

Figure 4.4: Protein Patch 2D map projection of hydrophobic surfaces on OPRM-WT and OPRM-M46.

References

1. Hauser, A. S.; Chavali, S.; Masuho, I.; Jahn, L. J.; Martemyanov, K. A.; Gloriam, D. E.; Babu, M. M. Pharmacogenomics of GPCR Drug Targets. *Cell* **2018**, *172*, 54. e19.

2. Overington, J. P.; Al-Lazikani, B.; Hopkins, A. L. How many drug targets are there? *Nat Rev Drug Discov* **2006**, *5*, 993-996.

3. Ribas, C.; Penela, P.; Murga, C.; Salcedo, A.; García-Hoz, C.; Jurado-Pueyo, M.; Aymerich, I.; Mayor, F. The G protein-coupled receptor kinase (GRK) interactome: Role of GRKs in GPCR regulation and signaling. *Biochimica et Biophysica Acta (BBA) - Biomembranes* **2007**, *1768*, 913-922.

4. Lagerström, M. C.; Schiöth, H. B. Structural diversity of G protein-coupled receptors and significance for drug discovery. *Nat Rev Drug Discov* **2008**, *7*, 339-357.

5. Rosenbaum, D. M.; Rasmussen, S. G. F.; Kobilka, B. K. The structure and function of G-protein-coupled receptors. *Nature* **2009**, *459*, 356-363.

6. Gether, U.; Ballesteros, J. A.; Seifert, R.; Sanders-Bush, E.; Weinstein, H.; Kobilka, B. K. Structural instability of a constitutively active G protein-coupled receptor. Agonist-independent activation due to conformational flexibility. *J. Biol. Chem.* **1997**, *272*, 2587-2590.

7. Oldham, W. M.; Hamm, H. E. Heterotrimeric G protein activation by G-protein-coupled receptors. *Nat. Rev. Mol. Cell Biol.* **2008**, *9*, 60-71.

8. DeWire, S. M.; Ahn, S.; Lefkowitz, R. J.; Shenoy, S. K. β-Arrestins and Cell Signaling. *Annual Review of Physiology* **2007**, *69*, 483-510.

9. Shenoy, S. K.; Lefkowitz, R. J. Seven-transmembrane receptor signaling through beta-arrestin. *Sci STKE* **2005**, *2005*, cm10.

10. Shukla, A. K.; Singh, G.; Ghosh, E. Emerging structural insights into biased GPCR signaling. *Trends in Biochemical Sciences* **2014**, *39*, 594-602.

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

12. Attwood, T. K.; Findlay, J. B. Fingerprinting G-protein-coupled receptors. *Protein Eng.* **1994**, *7*, 195-203.

13. Kolakowski, L. F. GCRDb: a G-protein-coupled receptor database. *Recept. Channels* **1994**, *2*, 1-7.

14. Schiöth, H. B.; Fredriksson, R. The GRAFS classification system of G-protein coupled receptors in comparative perspective. *Gen. Comp. Endocrinol.* **2005**, *142*, 94-101.

15. Ballesteros, J. A.; Weinstein, H. Integrated methods for the construction of three-dimensional models and computational probing of structure-function relations in G protein-coupled receptors. In *Methods in Neurosciences*; Sealfon, S. C., Ed.; Academic Press: 1995; Vol. 25, pp 366-428.

16. Katritch, V.; Cherezov, V.; Stevens, R. C. Structure-function of the G protein-coupled receptor superfamily. *Annu. Rev. Pharmacol. Toxicol.* **2013**, *53*, 531-556.

17. Sandoval, A.; Eichler, S.; Madathil, S.; Reeves, P.; Fahmy, K.; Böckmann, R. The Molecular Switching Mechanism at the Conserved D(E)RY Motif in Class-A GPCRs. *Biophysical Journal* **2016**, *111*, 79-89.

18. Vogel, R.; Mahalingam, M.; Lüdeke, S.; Huber, T.; Siebert, F.; Sakmar, T. P. Functional role of the "ionic lock"--an interhelical hydrogen-bond network in family A heptahelical receptors. *J. Mol. Biol.* **2008**, *380*, 648-655.

19. Scheerer, P.; Park, J. H.; Hildebrand, P. W.; Kim, Y. J.; Krauss, N.; Choe, H.; Hofmann, K. P.; Ernst, O. P. Crystal structure of opsin in its G-protein-interacting conformation. *Nature* **2008**, *455*, 497-502.

20. Standfuss, J.; Edwards, P. C.; D'Antona, A.; Fransen, M.; Xie, G.; Oprian, D. D.; Schertler, G. F. X. The structural basis of agonist-induced activation in constitutively active rhodopsin. *Nature* **2011**, *471*, 656-660.

21. Rasmussen, S. G. F.; DeVree, B. T.; Zou, Y.; Kruse, A. C.; Chung, K. Y.; Kobilka, T. S.; Thian, F. S.; Chae, P. S.; Pardon, E.; Calinski, D.; Mathiesen, J. M.; Shah, S. T. A.; Lyons, J. A.; Caffrey, M.; Gellman, S. H.; Steyaert, J.; Skiniotis, G.; Weis, W. I.; Sunahara, R. K.; Kobilka, B. K. Crystal structure of the β2 adrenergic receptor-Gs protein complex. *Nature* **2011**, *477*, 549-555.

22. Xu, F.; Wu, H.; Katritch, V.; Han, G. W.; Jacobson, K. A.; Gao, Z.; Cherezov, V.; Stevens, R. C. Structure of an agonist-bound human A2A adenosine receptor. *Science* **2011**, *332*, 322-327.

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

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

25. Jaakola, V.; Griffith, M. T.; Hanson, M. A.; Cherezov, V.; Chien, E. Y. T.; Lane, J. R.; IJzerman, A. P.; Stevens, R. C. The 2.6 Angstrom Crystal Structure of a Human A_{2A} Adenosine Receptor Bound to an Antagonist. *Science* **2008**, *322*, 1211-1217.

26. Shimamura, T.; Shiroishi, M.; Weyand, S.; Tsujimoto, H.; Winter, G.; Katritch, V.; Abagyan, R.; Cherezov, V.; Liu, W.; Han, G. W.; Kobayashi, T.; Stevens, R. C.; Iwata, S. Structure of the human histamine H1 receptor complex with doxepin. *Nature* **2011**, *475*, 65.

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

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

29. Warne, T.; Edwards, P. C.; Leslie, A. G. W.; Tate, C. G. Crystal structures of a stabilized β1-adrenoceptor bound to the biased agonists bucindolol and carvedilol. *Structure* **2012**, *20*, 841-849.

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

31. Vilar, S.; Karpiak, J.; Berk, B.; Costanzi, S. In silico analysis of the binding of agonists and blockers to the β2-adrenergic receptor. *J. Mol. Graph. Model.* **2011**, *29*, 809-817.

32. Rasmussen, S. G. F.; Choi, H.; Fung, J. J.; Pardon, E.; Casarosa, P.; Chae, P. S.; Devree, B. T.; Rosenbaum, D. M.; Thian, F. S.; Kobilka, T. S.; Schnapp, A.; Konetzki, I.; Sunahara, R. K.; Gellman, S. H.; Pautsch, A.; Steyaert, J.; Weis, W. I.; Kobilka, B. K. Structure of a nanobody-stabilized active state of the β(2) adrenoceptor. *Nature* **2011**, *469*, 175-180.

33. Lundstrom, K.; Wagner, R.; Reinhart, C.; Desmyter, A.; Cherouati, N.; Magnin, T.; Zeder-Lutz, G.; Courtot, M.; Prual, C.; André, N.; Hassaine, G.; Michel, H.; Cambillau, C.; Pattus, F. Structural genomics on membrane proteins: comparison of more than 100 GPCRs in 3 expression systems. *J Struct Funct Genomics* **2006**, *7*, 77-91.

34. Cherezov, V.; Rosenbaum, D. M.; Hanson, M. A.; Rasmussen, S. G. F.; Thian, F. S.; Kobilka, T. S.; Choi, H.; Kuhn, P.; Weis, W. I.; Kobilka, B. K.; Stevens, R. C. High-Resolution Crystal Structure of an Engineered Human Beta2- Adrenergic G Protein-Coupled Receptor. *Science* **2007**, *318*, 1258-1265.

35. Rasmussen, S. G. F.; Choi, H.; Rosenbaum, D. M.; Kobilka, T. S.; Thian, F. S.; Edwards, P. C.; Burghammer, M.; Ratnala, V. R. P.; Sanishvili, R.; Fischetti, R. F.; Schertler, G. F. X.;

Weis, W. I.; Kobilka, B. K. Crystal structure of the human beta2 adrenergic G-protein-coupled receptor. *Nature* **2007**, *450*, 383-387.

36. Pándy-Szekeres, G.; Munk, C.; Tsonkov, T. M.; Mordalski, S.; Harpsøe, K.; Hauser, A. S.; Bojarski, A. J.; Gloriam, D. E. GPCRdb in 2018: adding GPCR structure models and ligands. *Nucleic Acids Res* **2018**, *46*, D446.

37. Carpenter, E. P.; Beis, K.; Cameron, A. D.; Iwata, S. Overcoming the challenges of membrane protein crystallography. *Curr Opin Struct Biol* **2008**, *18*, 581-586.

38. Ghosh, E.; Kumari, P.; Jaiman, D.; Shukla, A. K. Methodological advances: the unsung heroes of the GPCR structural revolution. *Nature Reviews* **2015**, *16*, 69-81.

39. Milić, D.; Veprintsev, D. B. Large-scale production and protein engineering of G protein-coupled receptors for structural studies. *Front. Pharmacol.* **2015**, *6*.

40. Coudrat, T.; Simms, J.; Christopoulos, A.; Wootten, D.; Sexton, P. M. Improving virtual screening of G protein-coupled receptors via ligand-directed modeling. *PLOS Computational Biology* **2017**, *13*, e1005819.

41. Ranganathan, A.; Rodríguez, D.; Carlsson, J. Structure-Based Discovery of GPCR Ligands from Crystal Structures and Homology Models. In *SpringerLink*Springer, Berlin, Heidelberg: 2017; pp 1-35.

42. Kitchen, D. B.; Decornez, H.; Furr, J. R.; Bajorath, J. Docking and scoring in virtual screening for drug discovery: methods and applications. *Nat Rev Drug Discov* **2004**, *3*, 935-949.

43. Klebe, G. Virtual ligand screening: strategies, perspectives and limitations. *Drug Discov. Today* **2006**, *11*, 580-594.

44. Martí-Renom, M. A.; Stuart, A. C.; Fiser, A.; Sánchez, R.; Melo, F.; Sali, A. Comparative protein structure modeling of genes and genomes. *Annu Rev Biophys Biomol Struct* **2000**, *29*, 291-325.

45. Larsson, P.; Wallner, B.; Lindahl, E.; Elofsson, A. Using multiple templates to improve quality of homology models in automated homology modeling. *Protein Sci* **2008**, *17*, 990-1002.

46. Sali, A. Modeling mutations and homologous proteins. *Curr. Opin. Biotechnol.* **1995**, *6*, 437-451.

47. Srinivasan, N.; Guruprasad, K.; Blundell, T. L. Comparative Modelling of Proteins. In *Protein Structure Prediction: A Practical Approach*; Sternberg, M. J. E., Ed.; Oxford University Press, USA: 1996; pp 111-140.

48. Levitt, M. Accurate modeling of protein conformation by automatic segment matching. *Journal of Molecular Biology* **1992**, *226*, 507-533.

49. Sali, A.; Blundell, T. L. Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.* **1993**, *234*, 779-815.

50. Baker, D.; Sali, A. Protein structure prediction and structural genomics. *Science* **2001**, *294*, 93-96.

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

52. Isberg, V.; Mordalski, S.; Munk, C.; Harpsoe, K.; Rataj, K.; Hauser, A. S.; Vroling, B.; Bojarski, A. J.; Vriend, G.; Gloriam, D. E. GPCRdb: an information system for G protein-coupled receptors. *Nucleic Acids Res.* **2016**, *44*, D364.

53. Christen, M.; Gunsteren, W. F. v. On searching in, sampling of, and dynamically moving through conformational space of biomolecular systems: A review. *Journal of Computational Chemistry* **2008**, *29*, 157-166.

54. Liwo, A.; Czaplewski, C.; Ołdziej, S.; Scheraga, H. A. Computational techniques for efficient conformational sampling of proteins. *Current Opinion in Structural Biology* **2008**, *18*, 134-139.

55. Hatfield, M. P. D.; Lovas, S. Conformational sampling techniques. *Curr. Pharm. Des.* **2014**, *20*, 3303-3313.

56. Lazaridis, T.; Karplus, M. Effective energy functions for protein structure prediction. *Current Opinion in Structural Biology* **2000**, *10*, 139-145.

57. Gao, C.; Stern, H. A. Scoring function accuracy for membrane protein structure prediction. *Proteins: Structure, Function, and Bioinformatics* **2007**, *68*, 67-75.

58. Dorn, M.; e Silva, M. B.; Buriol, L. S.; Lamb, L. C. Three-dimensional protein structure prediction: Methods and computational strategies. *Computational Biology and Chemistry* **2014**, *53*, 251-276.

59. Kmiecik, S.; Gront, D.; Kolinski, M.; Wieteska, L.; Dawid, A. E.; Kolinski, A. Coarse-Grained Protein Models and Their Applications. *Chem. Rev.* **2016**, *116*, 7898-7936.

60. Trevizani, R.; Custódio, F. L.; Santos, K. B. d.; Dardenne, L. E. Critical Features of Fragment Libraries for Protein Structure Prediction. *PLOS ONE* **2017**, *12*, e0170131.

61. Haspel, N.; Tsai, C.; Wolfson, H.; Nussinov, R. Reducing the computational complexity of protein folding via fragment folding and assembly. *Protein Sci.* **2003**, *12*, 1177-1187.

62. Wu, S.; Skolnick, J.; Zhang, Y. Ab initio modeling of small proteins by iterative TASSER simulations. *BMC Biol.* **2007**, *5*, 17.

63. Zhang, Y. I-TASSER server for protein 3D structure prediction. *BMC Bioinformatics* **2008**, *9*, 40.

64. Rohl, C. A.; Strauss, C. E. M.; Misura, K. M. S.; Baker, D. Protein structure prediction using Rosetta. *Meth. Enzymol.* **2004**, *383*, 66-93.

65. Bender, B. J.; Cisneros, A.; Duran, A. M.; Finn, J. A.; Fu, D.; Lokits, A. D.; Mueller, B. K.; Sangha, A. K.; Sauer, M. F.; Sevy, A. M.; Sliwoski, G.; Sheehan, J. H.; DiMaio, F.; Meiler, J.; Moretti, R. Protocols for Molecular Modeling with Rosetta3 and RosettaScripts. *Biochemistry* **2016**, *55*, 4748-4763.

66. Song, Y.; DiMaio, F.; Wang, R.; Kim, D.; Miles, C.; Brunette, T. J.; Thompson, J.; Baker, D. High-Resolution Comparative Modeling with RosettaCM. *Structure* **2013**, *21*, 1735-1742.

67. Alford, R. F.; Koehler Leman, J.; Weitzner, B. D.; Duran, A. M.; Tilley, D. C.; Elazar, A.; Gray, J. J. An Integrated Framework Advancing Membrane Protein Modeling and Design. *PLoS Comput. Biol.* **2015**, *11*, e1004398.

68. Yarov-Yarovoy, V.; Schonbrun, J.; Baker, D. Multipass membrane protein structure prediction using Rosetta. *Proteins* **2006**, *62*, 1010-1025.

69. Barth, P.; Schonbrun, J.; Baker, D. Toward high-resolution prediction and design of transmembrane helical protein structures. *PNAS* **2007**, *104*, 15682-15687.

70. Barth, P.; Wallner, B.; Baker, D. Prediction of membrane protein structures with complex topologies using limited constraints. *PNAS* **2009**, *106*, 1409-1414.

71. Zhang, J.; Yang, J.; Jang, R.; Zhang, Y. GPCR-I-TASSER: A Hybrid Approach to G Protein-Coupled Receptor Structure Modeling and the Application to the Human Genome. *Structure* **2015**, *23*, 1538-1549.

72. Roy, A.; Kucukural, A.; Zhang, Y. I-TASSER: a unified platform for automated protein structure and function prediction. *Nat Protoc* **2010**, *5*, 725-738.

73. Weiner, B. E.; Alexander, N.; Akin, L. R.; Woetzel, N.; Karakas, M.; Meiler, J. BCL::Fold—Protein topology determination from limited NMR restraints. *Proteins: Structure, Function, and Bioinformatics* **2013**, *82*, 587-595.

74. Lindert, S.; Alexander, N.; Wötzel, N.; Karakaş, M.; Stewart, P.; Meiler, J. EM-Fold: De Novo Atomic-Detail Protein Structure Determination from Medium-Resolution Density Maps. *Structure* **2012**, *20*, 464-478.

75. Fischer, A. W.; Alexander, N. S.; Woetzel, N.; Karakas, M.; Weiner, B. E.; Meiler, J. BCL::MP-fold: Membrane protein structure prediction guided by EPR restraints. *Proteins: Structure, Function, and Bioinformatics* **2015**, *83*, 1947-1962.

76. Hofmann, T.; Fischer, A. W.; Meiler, J.; Kalkhof, S. Protein structure prediction guided by crosslinking restraints – A systematic evaluation of the impact of the crosslinking spacer length. *Methods* **2015**, *89*, 79-90.

77. Almeida, J. G.; Preto, A. J.; Koukos, P. I.; Bonvin, Alexandre M J J; Moreira, I. S. Membrane proteins structures: A review on computational modeling tools. *Biochim. Biophys. Acta* **2017**, *1859*, 2021-2039.

78. Weiner, B. E.; Woetzel, N.; Karakaş, M.; Alexander, N.; Meiler, J. BCL::MP-fold: folding membrane proteins through assembly of transmembrane helices. *Structure* **2013**, *21*, 1107-1117.

79. Xia, Y.; Fischer, A. W.; Teixeira, P.; Weiner, B.; Meiler, J. Integrated Structural Biology for α-Helical Membrane Protein Structure Determination. *Structure* **2018**, *26*, 666.e2.

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

81. Kufareva, I.; Katritch, V.; Stevens, R. C.; Abagyan, R. Advances in GPCR modeling evaluated by the GPCR Dock 2013 assessment: meeting new challenges. *Structure* **2014**, *22*, 1120-1139.

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

83. Roth, B. L.; Kroeze, W. K. Integrated Approaches for Genome-wide Interrogation of the Druggable Non-olfactory G Protein-coupled Receptor Superfamily. *The Journal of biological chemistry* **2015**, *290*, 19471-19477.

84. Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Research* **2000**, *28*, 235-242.

85. White, S. H. Biophysical dissection of membrane proteins. *https://www.nature.com/articles/nature08142* (accessed Jul 15, 2018).

86. White, S. Membrane proteins of known 3D structure . *http://blanco.biomol.uci.edu/mpstruc/* (accessed Jul 15, 2018).

87. Latorraca, N. R.; Venkatakrishnan, A. J.; Dror, R. O. GPCR Dynamics: Structures in Motion. *Chem. Rev.* **2017**, *117*, 139-155.

88. Baldwin, R. L. Dynamic hydration shell restores Kauzmann's 1959 explanation of how the hydrophobic factor drives protein folding. *Proceedings of the National Academy of Sciences of the United States of America* **2014**, *111*, 13052-13056.

89. Kendrew, J. C.; Dickerson, R. E.; Strandberg, B. E.; Hart, R. G.; Davies, D. R.; Phillips, D. C.; Shore, V. C. Structure of myoglobin: A three-dimensional Fourier synthesis at 2 A. resolution. *Nature* **1960**, *185*, 422-427.

90. Stevens, T. J.; Arkin, I. T. Are membrane proteins "inside-out" proteins? *Proteins* **1999**, *36*, 135-143.

91. Rees, D. C.; DeAntonio, L.; Eisenberg, D. Hydrophobic organization of membrane proteins. *Science* **1989**, *245*, 510-513.

92. Renthal, R. Transmembrane and water-soluble helix bundles display reverse patterns of surface roughness. *Biochem. Biophys. Res. Commun.* **1999**, *263*, 714-717.

93. Mitra, K.; Steitz, T. A.; Engelman, D. M. Rational design of `water-soluble' bacteriorhodopsin variants. *Protein Eng Des Sel* **2002**, *15*, 485-492.

94. Slovic, A. M.; Summa, C. M.; Lear, J. D.; DeGrado, W. F. Computational design of a water-soluble analog of phospholamban. *Protein Science* **2003**, *12*, 337-348.

95. Slovic, A. M.; Lear, J. D.; DeGrado, W. F. De novo design of a pentameric coiled-coil: decoding the motif for tetramer versus pentamer formation in water-soluble phospholamban. *J. Pept. Res.* **2005**, *65*, 312-321.

96. Slovic, A. M.; Kono, H.; Lear, J. D.; Saven, J. G.; DeGrado, W. F. Computational design of water-soluble analogues of the potassium channel KcsA. *PNAS* **2004**, *101*, 1828-1833.

97. Perez-Aguilar, J. M.; Xi, J.; Matsunaga, F.; Cui, X.; Selling, B.; Saven, J. G.; Liu, R. A Computationally Designed Water-Soluble Variant of a G-Protein Coupled Receptor: The Human Mu Opioid Receptor. *PLoS ONE* **2013**, *8*, e66009.

98. Zhao, X.; Perez-Aguilar, J.; Matsunaga, F.; Lerner, M.; Xi, J.; Selling, B.; Johnson, J., A. T; Saven, J.; Liu, R. Characterization of a Computationally Designed Water-soluble Human μ-Opioid Receptor Variant Using Available Structural Information. *Anesthesiology* **2014**, *121*, 866-875.

99. Roosild, T. P.; Choe, S. Redesigning an integral membrane K+ channel into a soluble protein. *Protein Eng. Des. Sel.* **2005**, *18*, 79-84.

100. Raran-Kurussi, S.; Waugh, D. S. The Ability to Enhance the Solubility of Its Fusion Partners Is an Intrinsic Property of Maltose-Binding Protein but Their Folding Is Either Spontaneous or Chaperone-Mediated. *PLOS ONE* **2012**, *7*, e49589.

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

102. Roger Y. Tsien The green fluorescent protein . **1998**, *67*, 509-544.

103. Epel, B. L.; Padgett, H. S.; Heinlein, M.; Beachy, R. N. Plant virus movement protein dynamics probed with a GFP-protein fusion. *Gene* **1996**, *173*, 75-79.

104. Elliott, G.; O'Hare, P. Intercellular trafficking of VP22-GFP fusion proteins. *Gene Therapy* **1999**, *6*, 149-151.

105. Shashidharan, P.; Chalmers-Redman, R. M. E.; Carlile, G. W.; Rodic, V.; Gurvich, N.; Yuen, T.; Tatton, W. G.; Sealfon, S. C. Nuclear translocation of GAPDH–GFP fusion protein during apoptosis. *NeuroReport* **1999**, *10*, 1149-1153.

106. Kawamata, Y.; Fujii, R.; Hosoya, M.; Harada, M.; Yoshida, H.; Miwa, M.; Fukusumi, S.; Habata, Y.; Itoh, T.; Shintani, Y.; Hinuma, S.; Fujisawa, Y.; Fujino, M. A G Protein-coupled Receptor Responsive to Bile Acids. *J. Biol. Chem.* **2003**, *278*, 9435-9440.

107. Bubnell, J.; Pfister, P.; Sapar, M. L.; Rogers, M. E.; Feinstein, P. β2 Adrenergic Receptor Fluorescent Protein Fusions Traffic to the Plasma Membrane and Retain Functionality. *PLOS ONE* **2013**, *8*, e74941.

108. Chen, R.; Mukhin, Y. V.; Garnovskaya, M. N.; Thielen, T. E.; Iijima, Y.; Huang, C.; Raymond, J. R.; Ullian, M. E.; Paul, R. V. A functional angiotensin II receptor-GFP fusion protein: evidence for agonist-dependent nuclear translocation. *American Journal of Physiology-Renal Physiology* **2000**, *279*, F448.

109. Ghisaidoobe, A. B. T.; Chung, S. J. Intrinsic Tryptophan Fluorescence in the Detection and Analysis of Proteins: A Focus on Förster Resonance Energy Transfer Techniques. *Int J Mol Sci* **2014**, *15*, 22518-22538.

110. Gregoire, S.; Kwon, I. A revisited folding reporter for quantitative assay of protein misfolding and aggregation in mammalian cells. *Biotechnol J* **2012**, *7*, 1297-1307.

111. Omoya, K.; Kato, Z.; Matsukuma, E.; Li, A.; Hashimoto, K.; Yamamoto, Y.; Ohnishi, H.; Kondo, N. Systematic optimization of active protein expression using GFP as a folding reporter. *Protein Expression and Purification* **2004**, *36*, 327-332.

112. Waldo, G. S.; Standish, B. M.; Berendzen, J.; Terwilliger, T. C. Rapid protein-folding assay using green fluorescent protein. *Nature Biotechnology* **1999**, *17*, 691-695.

113. Drew, D. E.; von Heijne, G.; Nordlund, P.; De Gier, J. L. Green fluorescent protein as an indicator to monitor membrane protein overexpression in Escherichia coli. *FEBS Letters* **2001**, *507*, 220-224.

114. Pédelacq, J.; Piltch, E.; Liong, E. C.; Berendzen, J.; Kim, C.; Rho, B.; Park, M. S.; Terwilliger, T. C.; Waldo, G. S. Engineering soluble proteins for structural genomics. *Nature Biotechnology* **2002**, *20*, nbt732.

115. Cabantous, S.; Rogers, Y.; Terwilliger, T. C.; Waldo, G. S. New Molecular Reporters for Rapid Protein Folding Assays. *PLOS ONE* **2008**, *3*, e2387.

116. Marino, J.; Hohl, M.; Seeger, M. A.; Zerbe, O.; Geertsma, E. R. Bicistronic mRNAs to enhance membrane protein overexpression. *J. Mol. Biol.* **2015**, *427*, 943-954.

117. Geertsma, E. R.; Groeneveld, M.; Slotboom, D.; Poolman, B. Quality control of overexpressed membrane proteins. *PNAS* **2008**, *105*, 5722-5727.

118. Yan, X.; Shi, Q.; Bracher, A.; Miličić, G.; Singh, A. K.; Hartl, F. U.; Hayer-Hartl, M. GroEL Ring Separation and Exchange in the Chaperonin Reaction. *Cell* **2018**, *172*, 617.e11.

119. Kim, Y. E.; Hipp, M. S.; Bracher, A.; Hayer-Hartl, M.; Hartl, F. U. Molecular chaperone functions in protein folding and proteostasis. *Annu. Rev. Biochem.* **2013**, *82*, 323-355.

120. Hayer-Hartl, M.; Bracher, A.; Hartl, F. U. The GroEL–GroES Chaperonin Machine: A Nano-Cage for Protein Folding. *Trends in Biochemical Sciences* **2016**, *41*, 62-76.

121. Kerner, M. J. The *Escherichia coli* GroEL interaction proteome: Identification and classification, Max-Planck Institute for Biochemistry, Martinsried, Technische Universität München, 2005.

122. Chaudhuri, T. K.; Farr, G. W.; Fenton, W. A.; Rospert, S.; Horwich, A. L. GroEL/GroES-Mediated Folding of a Protein Too Large to Be Encapsulated. *Cell* **2001**, *107*, 235-246.

123. Saibil, H. R.; Fenton, W. A.; Clare, D. K.; Horwich, A. L. Structure and allostery of the chaperonin GroEL. *J. Mol. Biol.* **2013**, *425*, 1476-1487.

124. Haldar, S.; Gupta, A. J.; Yan, X.; Miličić, G.; Hartl, F. U.; Hayer-Hartl, M. Chaperonin-assisted protein folding: Relative population of asymmetric and symmetric GroEL:GroES complexes. *J. Mol. Biol.* **2015**, *427*, 2244-2255.

125. Joseph, R. E.; Andreotti, A. H. Bacterial expression and purification of Interleukin-2 Tyrosine Kinase: Single step separation of the chaperonin impurity. *Protein Expression and Purification* **2008**, *60*, 194-197.

126. Thain, A.; Gaston, K.; Jenkins, O.; Clarke, A. R. A method for the separation of GST fusion proteins from co-purifying GroEL. *Trends in Genetics* **1996**, *12*, 209-210.

127. Brown, A. I. Novel strategies for the purification and inhibition of Protein-Tyrosine Kinases, University of Rhode Island, Kingston, Rhode Island, 2014.

128. Dessauer, C. W.; Bartlett, S. G. Identification of a chaperonin binding site in a chloroplast precursor protein. *Journal of Biological Chemistry* **1994**, *269*, 19766-19776.

129. Scherrer, S.; Iriarte, A.; Martinez-Carrion, M. Stability and release requirements of the complexes of GroEL with two homologous mammalian aminotransferases. *Journal of Protein Chemistry* **2000**, *19*, 591-602.

130. Seto, M.; Ogawa, T.; Kodama, K.; Muramoto, K.; Kanayama, Y.; Sakai, Y.; Chijiwa, T.; Ohno, M. A novel recombinant system for functional expression of myonecrotic snake phospholipase A_2 in Escherichia coli using a new fusion affinity tag. *Protein Expression and Purification* **2008**, *58*, 194-202.

131. de Marco, A.; Volrath, S.; Bruyere, T.; Law, M.; Fonne-Pfister, R. Recombinant maize PPO expressed in E. Coli forms complexes with GroEL and DnaK chaperones. *Protein Expression and Purification* **2000**, *20*, 81-86.

132. Alegre, K. O.; Law, C. J. Purification of a multidrug resistance transporter for crystallization studies. *Antibiotics* **2015**, *4*, 113-135.

133. Rohman, M.; Harrison-Lavoie, K. J. Separation of copurifying GroEL from Glutathione-S-Transferase fusion proteins. *Protein Expression and Purification* **2000**, *20*, 45-47.

134. Chen, X. S.; Casini, G.; Harrison, S. C.; Garcea, R. L. Papillomavirus capsid protein expression in Escherichia coli: Purification and assembly of HPV11 and HPV16 L1. *J. Mol. Biol.* **2001**, *307*, 173-182.

135. Belval, L.; Marquette, A.; Mestre, P.; Piron, M.; Demangeat, G.; Merdinoglu, D.; Chich, J. A fast and simple method to eliminate Cpn60 from functional recombinant proteins produced by E. coli Arctic Express. *Protein Expr. Purif.* **2015**, *109*, 29-34.

136. Weaver, J.; Jiang, M.; Roth, A.; Puchalla, J.; Zhang, J.; Rye, H. S. GroEL actively stimulates folding of the endogenous substrate protein PepQ. *Nature Communications* **2017**, *8*, 15934.

137. Ghahghaei, A.; Bathaie, S. Z.; Shahraki, A.; Asgarabad, F. R. Comparison of the Chaperoning Action of Glycerol and β-Casein on Aggregation of Proteins in the Presence of Crowding Agent. *Int J Pept Res Ther* **2011**, *17*, 101.

138. Mendoza, J. A.; Rogers, E.; Lorimer, G. H.; Horowitz, P. M. Chaperonins facilitate the in vitro folding of monomeric mitochondrial rhodanese. *Journal of Biological Chemistry* **1991**, *266*, 13044.

139. Tandon, S.; Horowitz, P. The effects of lauryl maltoside on the reactivation of several enzymes after treatment with guanidinium chloride. *Biochim. Biophys. Acta* **1988**, *955*, 19-25.

140. Ostoa-Saloma, P.; Ramirez, J.; Perez-Montfort, R. Causes of the decrease in fluorescence due to proteolysis of alpha-casein. *Biochim. Biophys. Acta* **1990**, *1041*, 146-152.

141. Morgan, P. E.; Treweek, T. M.; Lindner, R. A.; Price, W. E.; Carver, J. A. Casein proteins as molecular chaperones. *J. Agric. Food Chem.* **2005**, *53*, 2670-2683.

142. Martin, J.; Langer, T.; Boteva, R.; Schramel, A. L.; Horwich, A.; Hartl, F. -. Chaperonin-mediated protein folding at the surface of GroEL through a 'molten globule'-like intermediate. *Nature* **1991**, *352*, 36-42.

143. Chi, H. e. a. Folding of newly translated membrane protein CCR5 is assisted by the chaperonin GroEL-GroES. *Sci. Rep.* **2015**, *5 (17037)*.

144. Trevino, S.; Scholtz, J. M.; Pace, C. N. Amino Acid Contribution to Protein Solubility: Asp, Glu, and Ser Contribute more Favorably than the other Hydrophilic Amino Acids in RNase Sa. *J. Mol. Biol.* **2007**, *366*, 449-460.

145. Pace, C. N.; Scholtz, J. M. A helix propensity scale based on experimental studies of peptides and proteins. *Biophysical journal* **1998**, *75*, 422-427.

146. Fraczkiewicz, R.; Braun, W. Exact and Efficient Analytical Calculation of the Accessible Surface Areas and their Gradients for Macromolecules. *Journal of Computational Chemistry* **1998**, *19*, 319-333.

147. Negi, S.; Zhu, H.; Fraczkiewicz, R.; Braun, W. Calculation of Solvent Accessible Surface Areas, Atomic Solvation Energies and Their Gradients for Macromolecules. *http://curie.utmb.edu/getarea.html*.

148. Chemical Computing Group Inc. Molecular Operating Environment (MOE). , *2016.08*.

149. Sambrook, J.; Maniatis, T.; Fritsch, E. F.; Russell, D. W. *Molecular Cloning: A Laboratory Manual;* Cold Spring Harbor Laboratory Press: Cold Spring Harbor, N.Y., 2001; .

150. Lin-Chao, S.; Chen, W. T.; Wong, T. T. High copy number of the pUC plasmid results from a Rom/Rop-suppressible point mutation in RNA II. *Mol. Microbiol.* **1992**, *6*, 3385-3393.

151. Miroux, B.; Walker, J. E. Over-production of proteins in Escherichia coli: mutant hosts that allow synthesis of some membrane proteins and globular proteins at high levels. *J. Mol. Biol.* **1996**, *260*, 289-298.

152. Dumon-Seignovert, L.; Cariot, G.; Vuillard, L. The toxicity of recombinant proteins in Escherichia coli: a comparison of overexpression in BL21(DE3), C41(DE3), and C43(DE3). *Protein Expr. Purif.* **2004**, *37*, 203-206.

153. Wagner, S.; Klepsch, M. M.; Schlegel, S.; Appel, A.; Draheim, R.; Tarry, M.; Högbom, M.; van Wijk, K. J.; Slotboom, D. J.; Persson, J. O.; de Gier, J. Tuning Escherichia coli for membrane protein overexpression. *Proc. Natl. Acad. Sci. U. S. A.* **2008**, *105*, 14371-14376.

154. Jeong, H.; Kim, H. J.; Lee, S. J. Complete Genome Sequence of Escherichia coli Strain BL21. *Genome Announc* **2015**, *3*.

155. Gacasan, S. B. Approaches for Family-Wide Investigations of G Protein-Coupled Receptors , University of Memphis, 2017.

156. Okada, T.; Sugihara, M.; Bondar, A.; Elstner, M.; Entel, P.; Buss, V. The Retinal Conformation and its Environment in Rhodopsin in Light of a New 2.2Å Crystal Structure††This paper is dedicated to Dr Yoshimasa Kyogoku. *Journal of Molecular Biology* **2004**, *342*, 571-583.

157. Thompson, A. A.; Liu, W.; Chun, E.; Katritch, V.; Wu, H.; Vardy, E.; Huang, X.; Trapella, C.; Guerrini, R.; Calo, G.; Roth, B. L.; Cherezov, V.; Stevens, R. C. Structure of the nociceptin/orphanin FQ receptor in complex with a peptide mimetic. *Nature* **2012**, *485*, 395.

158. Park, S. H.; Das, B. B.; Casagrande, F.; Tian, Y.; Nothnagel, H. J.; Chu, M.; Kiefer, H.; Maier, K.; Angelis, A. A. D.; Marassi, F. M.; Opella, S. J. Structure of the chemokine receptor CXCR1 in phospholipid bilayers. *Nature* **2012**, *491*, 779.

159. White, J. F.; Noinaj, N.; Shibata, Y.; Love, J.; Kloss, B.; Xu, F.; Gvozdenovic-Jeremic, J.; Shah, P.; Shiloach, J.; Tate, C. G.; Grisshammer, R. Structure of the agonist-bound neurotensin receptor. *Nature* **2012**, *490*, 508.

160. Warne, T.; Serrano-Vega, M. J.; Baker, J. G.; Moukhametzianov, R.; Edwards, P. C.; Henderson, R.; Leslie, A. G. W.; Tate, C. G.; Schertler, G. F. X. Structure of a β1-adrenergic G-protein-coupled receptor. *Nature* **2008**, *454*, 486.

161. Wang, C.; Jiang, Y.; Ma, J.; Wu, H.; Wacker, D.; Katritch, V.; Han, G. W.; Liu, W.; Huang, X.; Vardy, E.; McCorvy, J. D.; Gao, X.; Zhou, X. E.; Melcher, K.; Zhang, C.; Bai, F.; Yang, H.; Yang, L.; Jiang, H.; Roth, B. L.; Cherezov, V.; Stevens, R. C.; Xu, H. E. Structural Basis for Molecular Recognition at Serotonin Receptors. *Science* **2013**, *340*, 610-614.

162. Ring, A. M.; Manglik, A.; Kruse, A. C.; Enos, M. D.; Weis, W. I.; Garcia, K. C.; Kobilka, B. K. Adrenaline-activated structure of β2-adrenoceptor stabilized by an engineered nanobody. *Nature* **2013**, *502*, 575.

163. Wacker, D.; Wang, C.; Katritch, V.; Han, G. W.; Huang, X.; Vardy, E.; McCorvy, J. D.; Jiang, Y.; Chu, M.; Siu, F. Y.; Liu, W.; Xu, H. E.; Cherezov, V.; Roth, B. L.; Stevens, R. C.

Structural Features for Functional Selectivity at Serotonin Receptors. *Science* **2013**, *340*, 615-619.

164. Wu, B.; Chien, E. Y. T.; Mol, C. D.; Fenalti, G.; Liu, W.; Katritch, V.; Abagyan, R.; Brooun, A.; Wells, P.; Bi, F. C.; Hamel, D. J.; Kuhn, P.; Handel, T. M.; Cherezov, V.; Stevens, R. C. Structures of the CXCR4 Chemokine GPCR with Small-Molecule and Cyclic Peptide Antagonists. *Science* **2010**, *330*, 1066-1071.

165. Wang, C.; Wu, H.; Katritch, V.; Han, G. W.; Huang, X.; Liu, W.; Siu, F. Y.; Roth, B. L.; Cherezov, V.; Stevens, R. C. Structure of the human smoothened receptor bound to an antitumour agent. *Nature* **2013**, *497*, 338.

166. Chien, E. Y. T.; Liu, W.; Zhao, Q.; Katritch, V.; Won Han, G.; Hanson, M. A.; Shi, L.; Newman, A. H.; Javitch, J. A.; Cherezov, V.; Stevens, R. C. Structure of the Human Dopamine D3 Receptor in Complex with a D2/D3 Selective Antagonist. *Science* **2010**, *330*, 1091-1095.

167. Hollenstein, K.; Kean, J.; Bortolato, A.; Cheng, R. K. Y.; Doré, A. S.; Jazayeri, A.; Cooke, R. M.; Weir, M.; Marshall, F. H. Structure of class B GPCR corticotropin-releasing factor receptor 1. *Nature* **2013**, *499*, 438.

168. Siu, F. Y.; He, M.; Graaf, C. d.; Han, G. W.; Yang, D.; Zhang, Z.; Zhou, C.; Xu, Q.; Wacker, D.; Joseph, J. S.; Liu, W.; Lau, J.; Cherezov, V.; Katritch, V.; Wang, M.; Stevens, R. C. Structure of the human glucagon class B G-protein-coupled receptor. *Nature* **2013**, *499*, 444.

169. Tan, Q.; Zhu, Y.; Li, J.; Chen, Z.; Han, G. W.; Kufareva, I.; Li, T.; Ma, L.; Fenalti, G.; Li, J.; Zhang, W.; Xie, X.; Yang, H.; Jiang, H.; Cherezov, V.; Liu, H.; Stevens, R. C.; Zhao, Q.; Wu, B. Structure of the CCR5 Chemokine Receptor–HIV Entry Inhibitor Maraviroc Complex. *Science* **2013**, *341*, 1387-1390.

170. Hanson, M. A.; Roth, C. B.; Jo, E.; Griffith, M. T.; Scott, F. L.; Reinhart, G.; Desale, H.; Clemons, B.; Cahalan, S. M.; Schuerer, S. C.; Sanna, M. G.; Han, G. W.; Kuhn, P.; Rosen, H.; Stevens, R. C. Crystal Structure of a Lipid G Protein-Coupled Receptor. *Science* **2012**, *335*, 851-855.

171. Zhang, C.; Srinivasan, Y.; Arlow, D. H.; Fung, J. J.; Palmer, D.; Zheng, Y.; Green, H. F.; Pandey, A.; Dror, R. O.; Shaw, D. E.; Weis, W. I.; Coughlin, S. R.; Kobilka, B. K. High-resolution crystal structure of human protease-activated receptor 1. *Nature* **2012**, *492*, 387.

172. Granier, S.; Manglik, A.; Kruse, A. C.; Kobilka, T. S.; Thian, F. S.; Weis, W. I.; Kobilka, B. K. Structure of the δ-opioid receptor bound to naltrindole. *Nature* **2012**, *485*, 400.

173. Wu, H.; Wacker, D.; Mileni, M.; Katritch, V.; Han, G. W.; Vardy, E.; Liu, W.; Thompson, A. A.; Huang, X.; Carroll, F. I.; Mascarella, S. W.; Westkaemper, R. B.; Mosier, P. D.;

Roth, B. L.; Cherezov, V.; Stevens, R. C. Structure of the human κ-opioid receptor in complex with JDTic. *Nature* **2012**, *485*, 327.

174. Rosano, G. L.; Ceccarelli, E. A. Recombinant protein expression in Escherichia coli: advances and challenges. *Front Microbiol* **2014**, *5*.

175. Vasina, J. A.; Baneyx, F. Expression of aggregation-prone recombinant proteins at low temperatures: a comparative study of the Escherichia coli cspA and tac promoter systems. *Protein Expr. Purif.* **1997**, *9*, 211-218.

176. Hartl, F. U. Molecular chaperones in cellular protein folding. *Nature* **1996**, *381*, 571.

177. Leonard M. Hjelmeland A Nondenaturing Zwitterionic Detergent for Membrane Biochemistry: Design and Synthesis. *Proceedings of the National Academy of Sciences of the United States of America* **1980**, *77*, 6368-6370.

178. Kerner, M. J.; Naylor, D. J.; Ishihama, Y.; Maier, T.; Chang, H.; Stines, A. P.; Georgopoulos, C.; Frishman, D.; Hayer-Hartl, M.; Mann, M.; Hartl, F. U. Proteome-wide Analysis of Chaperonin-Dependent Protein Folding in Escherichia coli. *Cell* **2005**, *122*, 209-220.

179. Hartl, F. U. Unfolding the chaperone story. *MBoC* **2017**, *28*, 2919-2923.

180. Rial, D. V.; Ceccarelli, E. A. Removal of DnaK contamination during fusion protein purifications. *Protein Expr. Purif.* **2002**, *25*, 503-507.

181. Dahmane, T.; Damian, M.; Mary, S.; Popot, J.; Banères, J. Amphipol-assisted in vitro folding of G protein-coupled receptors. *Biochemistry* **2009**, *48*, 6516-6521.

182. Baneres, J.; Martin, A.; Hullot, P.; Girard, J.; Rossi, J.; Parello, J. Structure-based analysis of GPCR function: conformational adaptation of both agonist and receptor upon leukotriene B4 binding to recombinant BLT1. *J. Mol. Biol.* **2003**, *329*, 801-814.

183. Arcemisbéhère, L.; Sen, T.; Boudier, L.; Balestre, M.; Gaibelet, G.; Detouillon, E.; Orcel, H.; Mendre, C.; Rahmeh, R.; Granier, S.; Vivès, C.; Fieschi, F.; Damian, M.; Durroux, T.; Banères, J.; Mouillac, B. Leukotriene BLT2 Receptor Monomers Activate the Gi2 GTP-binding Protein More Efficiently than Dimers. *J Biol Chem* **2010**, *285*, 6337-6347.

184. Michalke, K.; Huyghe, C.; Lichière, J.; Gravière, M.; Siponen, M.; Sciara, G.; Lepaul, I.; Wagner, R.; Magg, C.; Rudolph, R.; Cambillau, C.; Desmyter, A. Mammalian G protein-coupled receptor expression in Escherichia coli: II. Refolding and biophysical characterization of mouse cannabinoid receptor 1 and human parathyroid hormone receptor 1. *Anal. Biochem.* **2010**, *401*, 74-80.

185. Banères, J.; Mesnier, D.; Martin, A.; Joubert, L.; Dumuis, A.; Bockaert, J. Molecular characterization of a purified 5-HT4 receptor: a structural basis for drug efficacy. *J. Biol. Chem.* **2005**, *280*, 20253-20260.

186. Schröder-Tittmann, K.; Bosse-Doenecke, E.; Reedtz-Runge, S.; Ihling, C.; Sinz, A.; Tittmann, K.; Rudolph, R. Recombinant Expression, in Vitro Refolding, and Biophysical Characterization of the Human Glucagon-like Peptide-1 Receptor. *Biochemistry* **2010**, *49*, 7956-7965.

187. Banères, J.; Popot, J.; Mouillac, B. New advances in production and functional folding of G-protein-coupled receptors. *Trends in Biotechnology* **2011**, *29*, 314-322.

188. Leszczynski, J. F.; Rose, G. D. Loops in globular proteins: a novel category of secondary structure. *Science* **1986**, *234*, 849-855.

189. Papaleo, E.; Saladino, G.; Lambrughi, M.; Lindorff-Larsen, K.; Gervasio, F. L.; Nussinov, R. The Role of Protein Loops and Linkers in Conformational Dynamics and Allostery. *Chem. Rev.* **2016**, *116*, 6391-6423.

190. Saraste, M.; Sibbald, P. R.; Wittinghofer, A. The P-loop — a common motif in ATP- and GTP-binding proteins. *Trends in Biochemical Sciences* **1990**, *15*, 430-434.

191. Fetrow, J. S. Omega loops: nonregular secondary structures significant in protein function and stability. *The FASEB Journal* **1995**, *9*, 708-717.

192. Sehnke, P. C.; Laughner, B.; Cardasis, H.; Powell, D.; Ferl, R. J. Exposed Loop Domains of Complexed 14-3-3 Proteins Contribute to Structural Diversity and Functional Specificity. *Plant Physiol.* **2006**, *140*, 647-660.

193. van der Kant, R.; Vriend, G. Alpha-Bulges in G Protein-Coupled Receptors. *Int J Mol Sci.* **2014**, *15*, 7841–7864.

194. Isberg, V.; de Graaf, C.; Bortolato, A.; Cherezov, V.; Katritch, V.; Marshall, F. H.; Mordalski, S.; Pin, J.; Stevens, R. C.; Vriend, G.; Gloriam, D. E. Generic GPCR Residue Numbers - Aligning Topology Maps While Minding The Gaps. *Trends Pharmacol Sci* **2015**, *36(1)*, 22–31.

195. Wheatley, M.; Wootten, D.; Conner, M.; Simms, J.; Kendrick, R.; Logan, R.; Poyner, D.; Barwell, J. Lifting the lid on GPCRs: the role of extracellular loops. *British Journal of Pharmacology* **2012**, *165*, 1688-1703.

196. Woolley, M. J.; Conner, A. C. Understanding the common themes and diverse roles of the second extracellular loop (ECL2) of the GPCR super-family. *Mol. Cell. Endocrinol.* **2017**, *449*, 3-11.

197. Avlani, V. A.; Gregory, K. J.; Morton, C. J.; Parker, M. W.; Sexton, P. M.; Christopoulos, A. Critical Role for the Second Extracellular Loop in the Binding of Both Orthosteric and Allosteric G Protein-coupled Receptor Ligands. *J. Biol. Chem.* **2007**, *282*, 25677-25686.

198. Peeters, M. C.; van Westen, G J P; Li, Q.; IJzerman, A. P. Importance of the extracellular loops in G protein-coupled receptors for ligand recognition and receptor activation. *Trends in Pharmacological Sciences* **2011**, *32*, 35-42.

199. Klco, J. M.; Wiegand, C. B.; Narzinski, K.; Baranski, T. J. Essential role for the second extracellular loop in C5a receptor activation. *Nature Structural & Molecular Biology* **2005**, *12*, 320.

200. Nanevicz, T.; Wang, L.; Chen, M.; Ishii, M.; Coughlin, S. R. Thrombin Receptor Activating Mutations: Alteration of an Extracellular Agonist Recognition Domain Causes Constitutive Signaling. *J. Biol. Chem.* **1996**, *271*, 702-706.

201. Pantel, J.; Legendre, M.; Cabrol, S.; Hilal, L.; Hajaji, Y.; Morisset, S.; Nivot, S.; Vie-Luton, M.; Grouselle, D.; Kerdanet, M. d.; Kadiri, A.; Epelbaum, J.; Bouc, Y. L.; Amselem, S. Loss of constitutive activity of the growth hormone secretagogue receptor in familial short stature. *J Clin Invest* **2006**, *116*, 760-768.

202. Chrencik, J. E.; Roth, C. B.; Terakado, M.; Kurata, H.; Omi, R.; Kihara, Y.; Warshaviak, D.; Nakade, S.; Asmar-Rovira, G.; Mileni, M.; Mizuno, H.; Griffith, M. T.; Rodgers, C.; Han, G. W.; Velasquez, J.; Chun, J.; Stevens, R. C.; Hanson, M. A. Crystal Structure of Antagonist Bound Human Lysophosphatidic Acid Receptor 1. *Cell* **2015**, *161*, 1633-1643.

203. Shao, Z.; Yin, J.; Chapman, K.; Grzemska, M.; Clark, L.; Wang, J.; Rosenbaum, D. M. High-resolution crystal structure of the human CB1 cannabinoid receptor. *Nature* **2016**, *540*, 602.

204. Zhang, J.; Zhang, K.; Gao, Z.; Paoletta, S.; Zhang, D.; Han, G. W.; Li, T.; Ma, L.; Zhang, W.; Müller, C. E.; Yang, H.; Jiang, H.; Cherezov, V.; Katritch, V.; Jacobson, K. A.; Stevens, R. C.; Wu, B.; Zhao, Q. Agonist-bound structure of the human P2Y12 receptor. *Nature* **2014**, *509*, 119.

205. Murakami, M.; Kouyama, T. Crystal structure of squid rhodopsin. *Nature* **2008**, *453*, 363.

206. Cheng, R. K. Y.; Fiez-Vandal, C.; Schlenker, O.; Edman, K.; Aggeler, B.; Brown, D. G.; Brown, G. A.; Cooke, R. M.; Dumelin, C. E.; Doré, A. S.; Geschwindner, S.; Grebner, C.; Hermansson, N.; Jazayeri, A.; Johansson, P.; Leong, L.; Prihandoko, R.; Rappas, M.; Soutter, H.; Snijder, A.; Sundström, L.; Tehan, B.; Thornton, P.; Troast, D.; Wiggin, G.; Zhukov, A.; Marshall, F. H.; Dekker, N. Structural insight into allosteric modulation of protease-activated receptor 2. *Nature* **2017**, *545*, 112.

207. Ma, Y.; Yue, Y.; Ma, Y.; Zhang, Q.; Zhou, Q.; Song, Y.; Shen, Y.; Li, X.; Ma, X.; Li, C.; Hanson, M. A.; Han, G. W.; Sickmier, E. A.; Swaminath, G.; Zhao, S.; Stevens, R. C.; Hu,

L. A.; Zhong, W.; Zhang, M.; Xu, F. Structural Basis for Apelin Control of the Human Apelin Receptor. *Structure* **2017**, *25*, 866.e4.

208. Lu, J.; Byrne, N.; Wang, J.; Bricogne, G.; Brown, F. K.; Chobanian, H. R.; Colletti, S. L.; Salvo, J. D.; Thomas-Fowlkes, B.; Guo, Y.; Hall, D. L.; Hadix, J.; Hastings, N. B.; Hermes, J. D.; Ho, T.; Howard, A. D.; Josien, H.; Kornienko, M.; Lumb, K. J.; Miller, M. W.; Patel, S. B.; Pio, B.; Plummer, C. W.; Sherborne, B. S.; Sheth, P.; Souza, S.; Tummala, S.; Vonrhein, C.; Webb, M.; Allen, S. J.; Johnston, J. M.; Weinglass, A. B.; Sharma, S.; Soisson, S. M. Structural basis for the cooperative allosteric activation of the free fatty acid receptor GPR40. *Nature Structural & Molecular Biology* **2017**, *24*, 570.

209. Glukhova, A.; Thal, D. M.; Nguyen, A. T.; Vecchio, E. A.; Jörg, M.; Scammells, P. J.; May, L. T.; Sexton, P. M.; Christopoulos, A. Structure of the Adenosine A1 Receptor Reveals the Basis for Subtype Selectivity. *Cell* **2017**, *168*, 877.e13.

210. Liu, W.; Chun, E.; Thompson, A. A.; Chubukov, P.; Xu, F.; Katritch, V.; Han, G. W.; Roth, C. B.; Heitman, L. H.; IJzerman, A. P.; Cherezov, V.; Stevens, R. C. Structural Basis for Allosteric Regulation of GPCRs by Sodium Ions. *Science* **2012**, *337*, 232-236.

211. Zhang, H.; Qiao, A.; Yang, D.; Yang, L.; Dai, A.; Graaf, C. d.; Reedtz-Runge, S.; Dharmarajan, V.; Zhang, H.; Han, G. W.; Grant, T. D.; Sierra, R. G.; Weierstall, U.; Nelson, G.; Liu, W.; Wu, Y.; Ma, L.; Cai, X.; Lin, G.; Wu, X.; Geng, Z.; Dong, Y.; Song, G.; Griffin, P. R.; Lau, J.; Cherezov, V.; Yang, H.; Hanson, M. A.; Stevens, R. C.; Zhao, Q.; Jiang, H.; Wang, M.; Wu, B. Structure of the full-length glucagon class B G-protein-coupled receptor. *Nature* **2017**, *546*, 259.

212. Jazayeri, A.; Rappas, M.; Brown, A. J. H.; Kean, J.; Errey, J. C.; Robertson, N. J.; Fiez-Vandal, C.; Andrews, S. P.; Congreve, M.; Bortolato, A.; Mason, J. S.; Baig, A. H.; Teobald, I.; Doré, A. S.; Weir, M.; Cooke, R. M.; Marshall, F. H. Crystal structure of the GLP-1 receptor bound to a peptide agonist. *Nature* **2017**, *546*, 254.

213. Dore, A. S.; Bortolato, A.; Hollenstein, K.; Cheng, R. K. Y.; Marshall*, Randy J Read and Fiona H Decoding Corticotropin-Releasing Factor Receptor Type 1 Crystal Structures. *http://www.eurekaselect.com/149116/article* (accessed May 1, 2018).

214. Thal, D. M.; Sun, B.; Feng, D.; Nawaratne, V.; Leach, K.; Felder, C. C.; Bures, M. G.; Evans, D. A.; Weis, W. I.; Bachhawat, P.; Kobilka, T. S.; Sexton, P. M.; Kobilka, B. K.; Christopoulos, A. Crystal structures of the M1 and M4 muscarinic acetylcholine receptors. *Nature* **2016**, *531*, 335.

215. Burg, J. S.; Ingram, J. R.; Venkatakrishnan, A. J.; Jude, K. M.; Dukkipati, A.; Feinberg, E. N.; Angelini, A.; Waghray, D.; Dror, R. O.; Ploegh, H. L.; Garcia, K. C. Structural basis for chemokine recognition and activation of a viral G protein-coupled receptor. *Science* **2015**, *347*, 1113-1117.

216. Fenalti, G.; Giguere, P. M.; Katritch, V.; Huang, X.; Thompson, A. A.; Cherezov, V.; Roth, B. L.; Stevens, R. C. Molecular control of δ-opioid receptor signalling. *Nature* **2014**, *506*, 191.

217. Zheng, Y.; Han, G. W.; Abagyan, R.; Wu, B.; Stevens, R. C.; Cherezov, V.; Kufareva, I.; Handel, T. M. Structure of CC Chemokine Receptor 5 with a Potent Chemokine Antagonist Reveals Mechanisms of Chemokine Recognition and Molecular Mimicry by HIV. *Immunity* **2017**, *46*, 1017.e5.

218. Zhang, D.; Gao, Z.; Zhang, K.; Kiselev, E.; Crane, S.; Wang, J.; Paoletta, S.; Yi, C.; Ma, L.; Zhang, W.; Han, G. W.; Liu, H.; Cherezov, V.; Katritch, V.; Jiang, H.; Stevens, R. C.; Jacobson, K. A.; Zhao, Q.; Wu, B. Two disparate ligand-binding sites in the human P2Y1 receptor. *Nature* **2015**, *520*, 317.

219. Zhang, H.; Han, G. W.; Batyuk, A.; Ishchenko, A.; White, K. L.; Patel, N.; Sadybekov, A.; Zamlynny, B.; Rudd, M. T.; Hollenstein, K.; Tolstikova, A.; White, T. A.; Hunter, M. S.; Weierstall, U.; Liu, W.; Babaoglu, K.; Moore, E. L.; Katz, R. D.; Shipman, J. M.; Garcia-Calvo, M.; Sharma, S.; Sheth, P.; Soisson, S. M.; Stevens, R. C.; Katritch, V.; Cherezov, V. Structural basis for selectivity and diversity in angiotensin II receptors. *Nature* **2017**, *544*, 327.

220. Wu, H.; Wang, C.; Gregory, K. J.; Han, G. W.; Cho, H. P.; Xia, Y.; Niswender, C. M.; Katritch, V.; Meiler, J.; Cherezov, V.; Conn, P. J.; Stevens, R. C. Structure of a Class C GPCR Metabotropic Glutamate Receptor 1 Bound to an Allosteric Modulator. *Science* **2014**, *344*, 58-64.

221. Krumm, B. E.; White, J. F.; Shah, P.; Grisshammer, R. Structural prerequisites for G-protein activation by the neurotensin receptor. *Nature Communications* **2015**, *6*, 7895.

222. Shihoya, W.; Nishizawa, T.; Okuta, A.; Tani, K.; Dohmae, N.; Fujiyoshi, Y.; Nureki, O.; Doi, T. Activation mechanism of endothelin ETB receptor by endothelin-1. *Nature* **2016**, *537*, 363.

223. Yin, J.; Babaoglu, K.; Brautigam, C. A.; Clark, L.; Shao, Z.; Scheuermann, T. H.; Harrell, C. M.; Gotter, A. L.; Roecker, A. J.; Winrow, C. J.; Renger, J. J.; Coleman, P. J.; Rosenbaum, D. M. Structure and ligand-binding mechanism of the human OX1 and OX2 orexin receptors. *Nature Structural & Molecular Biology* **2016**, *23*, 293.

224. Chen, K. M.; Sun, J.; Salvo, J. S.; Baker, D.; Barth, P. High-resolution modeling of transmembrane helical protein structures from distant homologues. *PLoS Comput. Biol.* **2014**, *10*, e1003636.

225. Jazayeri, A.; Doré, A. S.; Lamb, D.; Krishnamurthy, H.; Southall, S. M.; Baig, A. H.; Bortolato, A.; Koglin, M.; Robertson, N. J.; Errey, J. C.; Andrews, S. P.; Teobald, I.; Brown, A. J. H.; Cooke, R. M.; Weir, M.; Marshall, F. H. Extra-helical binding site of a glucagon receptor antagonist. *Nature* **2016**, *533*, 274-277.

226. Song, G.; Yang, D.; Wang, Y.; de Graaf, C.; Zhou, Q.; Jiang, S.; Liu, K.; Cai, X.; Dai, A.; Lin, G.; Liu, D.; Wu, F.; Wu, Y.; Zhao, S.; Ye, L.; Han, G. W.; Lau, J.; Wu, B.; Hanson, M. A.; Liu, Z.; Wang, M.; Stevens, R. C. Human GLP-1 receptor transmembrane domain structure in complex with allosteric modulators. *Nature* **2017**, *546*, 312-315.

227. Liang, Y.; Khoshouei, M.; Glukhova, A.; Furness, S. G. B.; Zhao, P.; Clydesdale, L.; Koole, C.; Truong, T. T.; Thal, D. M.; Lei, S.; Radjainia, M.; Danev, R.; Baumeister, W.; Wang, M.; Miller, L. J.; Christopoulos, A.; Sexton, P. M.; Wootten, D. Phase-plate cryo-EM structure of a biased agonist-bound human GLP-1 receptor-Gs complex. *Nature* **2018**, *555*, 121-125.

228. Kruse, A. C.; Ring, A. M.; Manglik, A.; Hu, J.; Hu, K.; Eitel, K.; Hübner, H.; Pardon, E.; Valant, C.; Sexton, P. M.; Christopoulos, A.; Felder, C. C.; Gmeiner, P.; Steyaert, J.; Weis, W. I.; Garcia, K. C.; Wess, J.; Kobilka, B. K. Activation and allosteric modulation of a muscarinic acetylcholine receptor. *Nature* **2013**, *504*, 101-106.

229. Thorsen, T. S.; Matt, R.; Weis, W. I.; Kobilka, B. K. Modified T4 Lysozyme Fusion Proteins Facilitate G Protein-Coupled Receptor Crystallogenesis. *Structure* **2014**, *22*, 1657-1664.

230. Qin, L.; Kufareva, I.; Holden, L. G.; Wang, C.; Zheng, Y.; Zhao, C.; Fenalti, G.; Wu, H.; Han, G. W.; Cherezov, V.; Abagyan, R.; Stevens, R. C.; Handel, T. M. Structural biology. Crystal structure of the chemokine receptor CXCR4 in complex with a viral chemokine. *Science* **2015**, *347*, 1117-1122.

231. Hua, T.; Vemuri, K.; Nikas, S. P.; Laprairie, R. B.; Wu, Y.; Qu, L.; Pu, M.; Korde, A.; Jiang, S.; Ho, J.; Han, G. W.; Ding, K.; Li, X.; Liu, H.; Hanson, M. A.; Zhao, S.; Bohn, L. M.; Makriyannis, A.; Stevens, R. C.; Liu, Z. Crystal structures of agonist-bound human cannabinoid receptor CB1. *Nature* **2017**, *547*, 468-471.

232. Hua, T.; Vemuri, K.; Pu, M.; Qu, L.; Han, G. W.; Wu, Y.; Zhao, S.; Shui, W.; Li, S.; Korde, A.; Laprairie, R. B.; Stahl, E. L.; Ho, J.; Zvonok, N.; Zhou, H.; Kufareva, I.; Wu, B.; Zhao, Q.; Hanson, M. A.; Bohn, L. M.; Makriyannis, A.; Stevens, R. C.; Liu, Z. Crystal Structure of the Human Cannabinoid Receptor CB1. *Cell* **2016**, *167*, 762.e14.

233. Zhang, K.; Zhang, J.; Gao, Z.; Zhang, D.; Zhu, L.; Han, G. W.; Moss, S. M.; Paoletta, S.; Kiselev, E.; Lu, W.; Fenalti, G.; Zhang, W.; Müller, C. E.; Yang, H.; Jiang, H.; Cherezov, V.; Katritch, V.; Jacobson, K. A.; Stevens, R. C.; Wu, B.; Zhao, Q. Structure of the human P2Y12 receptor in complex with an antithrombotic drug. *Nature* **2014**, *509*, 115-118.

234. Byrne, E. F. X.; Sircar, R.; Miller, P. S.; Hedger, G.; Luchetti, G.; Nachtergaele, S.; Tully, M. D.; Mydock-McGrane, L.; Covey, D. F.; Rambo, R. P.; Sansom, M. S. P.; Newstead, S.; Rohatgi, R.; Siebold, C. Structural basis of Smoothened regulation by its extracellular domains. *Nature* **2016**, *535*, 517-522.

235. Weierstall, U.; James, D.; Wang, C.; White, T. A.; Wang, D.; Liu, W.; Spence, J. C. H.; Bruce Doak, R.; Nelson, G.; Fromme, P.; Fromme, R.; Grotjohann, I.; Kupitz, C.; Zatsepin, N. A.; Liu, H.; Basu, S.; Wacker, D.; Han, G. W.; Katritch, V.; Boutet, S.; Messerschmidt, M.; Williams, G. J.; Koglin, J. E.; Marvin Seibert, M.; Klinker, M.; Gati, C.; Shoeman, R. L.; Barty, A.; Chapman, H. N.; Kirian, R. A.; Beyerlein, K. R.; Stevens, R. C.; Li, D.; Shah, S. T. A.; Howe, N.; Caffrey, M.; Cherezov, V. Lipidic cubic phase injector facilitates membrane protein serial femtosecond crystallography. *Nat Commun* **2014**, *5*, 3309.

236. Wang, C.; Wu, H.; Evron, T.; Vardy, E.; Han, G. W.; Huang, X.; Hufeisen, S. J.; Mangano, T. J.; Urban, D. J.; Katritch, V.; Cherezov, V.; Caron, M. G.; Roth, B. L.; Stevens, R. C. Structural basis for Smoothened receptor modulation and chemoresistance to anticancer drugs. *Nat Commun* **2014**, *5*, 4355.

237. Fenalti, G.; Zatsepin, N. A.; Betti, C.; Giguere, P.; Han, G. W.; Ishchenko, A.; Liu, W.; Guillemyn, K.; Zhang, H.; James, D.; Wang, D.; Weierstall, U.; Spence, J. C. H.; Boutet, S.; Messerschmidt, M.; Williams, G. J.; Gati, C.; Yefanov, O. M.; White, T. A.; Oberthuer, D.; Metz, M.; Yoon, C. H.; Barty, A.; Chapman, H. N.; Basu, S.; Coe, J.; Conrad, C. E.; Fromme, R.; Fromme, P.; Tourwé, D.; Schiller, P. W.; Roth, B. L.; Ballet, S.; Katritch, V.; Stevens, R. C.; Cherezov, V. Structural basis for bifunctional peptide recognition at human δ-opioid receptor. *Nat. Struct. Mol. Biol.* **2015**, *22*, 265-268.

238. Miller, R. L.; Thompson, A. A.; Trapella, C.; Guerrini, R.; Malfacini, D.; Patel, N.; Han, G. W.; Cherezov, V.; Caló, G.; Katritch, V.; Stevens, R. C. The Importance of Ligand-Receptor Conformational Pairs in Stabilization: Spotlight on the N/OFQ G Protein-Coupled Receptor. *Structure* **2015**, *23*, 2291-2299.

239. Egloff, P.; Hillenbrand, M.; Klenk, C.; Batyuk, A.; Heine, P.; Balada, S.; Schlinkmann, K. M.; Scott, D. J.; Schütz, M.; Plückthun, A. Structure of signaling-competent neurotensin receptor 1 obtained by directed evolution in Escherichia coli. *Proc. Natl. Acad. Sci. U. S. A.* **2014**, *111*, 655.

240. Krumm, B. E.; Lee, S.; Bhattacharya, S.; Botos, I.; White, C. F.; Du, H.; Vaidehi, N.; Grisshammer, R. Structure and dynamics of a constitutively active neurotensin receptor. *Sci Rep* **2016**, *6*, 38564.

241. Warne, T.; Moukhametzianov, R.; Baker, J. G.; Nehmé, R.; Edwards, P. C.; Leslie, A. G. W.; Schertler, G. F. X.; Tate, C. G. The structural basis for agonist and partial agonist action on a β(1)-adrenergic receptor. *Nature* **2011**, *469*, 241-244.

242. Moukhametzianov, R.; Warne, T.; Edwards, P. C.; Serrano-Vega, M. J.; Leslie, A. G. W.; Tate, C. G.; Schertler, G. F. X. Two distinct conformations of helix 6 observed in antagonist-bound structures of a beta1-adrenergic receptor. *Proc. Natl. Acad. Sci. U. S. A.* **2011**, *108*, 8228-8232.

243. Christopher, J. A.; Brown, J.; Doré, A. S.; Errey, J. C.; Koglin, M.; Marshall, F. H.; Myszka, D. G.; Rich, R. L.; Tate, C. G.; Tehan, B.; Warne, T.; Congreve, M. Biophysical

fragment screening of the β1-adrenergic receptor: identification of high affinity arylpiperazine leads using structure-based drug design. *J. Med. Chem.* **2013**, *56*, 3446-3455.

244. Miller-Gallacher, J. L.; Nehmé, R.; Warne, T.; Edwards, P. C.; Schertler, G. F. X.; Leslie, A. G. W.; Tate, C. G. The 2.1 Å resolution structure of cyanopindolol-bound β1-adrenoceptor identifies an intramembrane Na+ ion that stabilises the ligand-free receptor. *PLoS ONE* **2014**, *9*, e92727.

245. Huang, J.; Chen, S.; Zhang, J. J.; Huang, X. Crystal structure of oligomeric β1-adrenergic G protein-coupled receptors in ligand-free basal state. *Nat. Struct. Mol. Biol.* **2013**, *20*, 419-425.

246. Sato, T.; Baker, J.; Warne, T.; Brown, G. A.; Leslie, A. G. W.; Congreve, M.; Tate, C. G. Pharmacological Analysis and Structure Determination of 7-Methylcyanopindolol-Bound β1-Adrenergic Receptor. *Mol. Pharmacol.* **2015**, *88*, 1024-1034.

247. Hanson, M. A.; Cherezov, V.; Griffith, M. T.; Roth, C. B.; Jaakola, V.; Chien, E. Y. T.; Velasquez, J.; Kuhn, P.; Stevens, R. C. A specific cholesterol binding site is established by the 2.8 A structure of the human beta2-adrenergic receptor. *Structure* **2008**, *16*, 897-905.

248. Bokoch, M. P.; Zou, Y.; Rasmussen, S. G. F.; Liu, C. W.; Nygaard, R.; Rosenbaum, D. M.; Fung, J. J.; Choi, H.; Thian, F. S.; Kobilka, T. S.; Puglisi, J. D.; Weis, W. I.; Pardo, L.; Prosser, R. S.; Mueller, L.; Kobilka, B. K. Ligand-specific regulation of the extracellular surface of a G-protein-coupled receptor. *Nature* **2010**, *463*, 108-112.

249. Rosenbaum, D. M.; Zhang, C.; Lyons, J. A.; Holl, R.; Aragao, D.; Arlow, D. H.; Rasmussen, S. G. F.; Choi, H.; Devree, B. T.; Sunahara, R. K.; Chae, P. S.; Gellman, S. H.; Dror, R. O.; Shaw, D. E.; Weis, W. I.; Caffrey, M.; Gmeiner, P.; Kobilka, B. K. Structure and function of an irreversible agonist-β(2) adrenoceptor complex. *Nature* **2011**, *469*, 236-240.

250. Zou, Y.; Weis, W. I.; Kobilka, B. K. N-terminal T4 lysozyme fusion facilitates crystallization of a G protein coupled receptor. *PLoS ONE* **2012**, *7*, e46039.

251. Weichert, D.; Kruse, A. C.; Manglik, A.; Hiller, C.; Zhang, C.; Hübner, H.; Kobilka, B. K.; Gmeiner, P. Covalent agonists for studying G protein-coupled receptor activation. *Proc. Natl. Acad. Sci. U. S. A.* **2014**, *111*, 10744-10748.

252. Staus, D. P.; Strachan, R. T.; Manglik, A.; Pani, B.; Kahsai, A. W.; Kim, T. H.; Wingler, L. M.; Ahn, S.; Chatterjee, A.; Masoudi, A.; Kruse, A. C.; Pardon, E.; Steyaert, J.; Weis, W. I.; Prosser, R. S.; Kobilka, B. K.; Costa, T.; Lefkowitz, R. J. Allosteric nanobodies reveal the dynamic range and diverse mechanisms of G-protein-coupled receptor activation. *Nature* **2016**, *535*, 448-452.

253. Huang, C. Y.; Olieric, V.; Ma, P.; Howe, N.; Vogeley, L.; Liu, X.; Warshamanage, R.; Weinert, T.; Panepucci, E.; Kobilka, B.; Diederichs, K.; Wang, M.; Caffrey, M. In meso in

situ serial X-ray crystallography of soluble and membrane proteins at cryogenic temperatures. *Acta Crystallogr D Struct Biol* **2016**, *72*, 93-112.

254. Ma, P.; Weichert, D.; Aleksandrov, L. A.; Jensen, T. J.; Riordan, J. R.; Liu, X.; Kobilka, B. K.; Caffrey, M. The cubicon method for concentrating membrane proteins in the cubic mesophase. *Nat Protoc* **2017**, *12*, 1745-1762.

255. Shihoya, W.; Nishizawa, T.; Yamashita, K.; Inoue, A.; Hirata, K.; Kadji, F. M. N.; Okuta, A.; Tani, K.; Aoki, J.; Fujiyoshi, Y.; Doi, T.; Nureki, O. X-ray structures of endothelin ETB receptor bound to clinical antagonist bosentan and its analog. *Nat. Struct. Mol. Biol.* **2017**, *24*, 758-764.

256. Li, J.; Edwards, P. C.; Burghammer, M.; Villa, C.; Schertler, G. F. X. Structure of bovine rhodopsin in a trigonal crystal form. *J. Mol. Biol.* **2004**, *343*, 1409-1438.

257. Okada, T.; Fujiyoshi, Y.; Silow, M.; Navarro, J.; Landau, E. M.; Shichida, Y. Functional role of internal water molecules in rhodopsin revealed by X-ray crystallography. *Proc. Natl. Acad. Sci. U. S. A.* **2002**, *99*, 5982-5987.

258. Teller, D. C.; Okada, T.; Behnke, C. A.; Palczewski, K.; Stenkamp, R. E. Advances in determination of a high-resolution three-dimensional structure of rhodopsin, a model of G-protein-coupled receptors (GPCRs). *Biochemistry* **2001**, *40*, 7761-7772.

259. Nakamichi, H.; Okada, T. Crystallographic analysis of primary visual photochemistry. *Angew. Chem. Int. Ed. Engl.* **2006**, *45*, 4270-4273.

260. Nakamichi, H.; Buss, V.; Okada, T. Photoisomerization mechanism of rhodopsin and 9-cis-rhodopsin revealed by x-ray crystallography. *Biophys. J.* **2007**, *92*, 106.

261. Standfuss, J.; Xie, G.; Edwards, P. C.; Burghammer, M.; Oprian, D. D.; Schertler, G. F. X. Crystal structure of a thermally stable rhodopsin mutant. *J. Mol. Biol.* **2007**, *372*, 1179-1188.

262. Nakamichi, H.; Okada, T. Local peptide movement in the photoreaction intermediate of rhodopsin. *Proc. Natl. Acad. Sci. U. S. A.* **2006**, *103*, 12729-12734.

263. Salom, D.; Lodowski, D. T.; Stenkamp, R. E.; Le Trong, I.; Golczak, M.; Jastrzebska, B.; Harris, T.; Ballesteros, J. A.; Palczewski, K. Crystal structure of a photoactivated deprotonated intermediate of rhodopsin. *Proc. Natl. Acad. Sci. U. S. A.* **2006**, *103*, 16123-16128.

264. Makino, C. L.; Riley, C. K.; Looney, J.; Crouch, R. K.; Okada, T. Binding of more than one retinoid to visual opsins. *Biophys. J.* **2010**, *99*, 2366-2373.

265. Park, J. H.; Scheerer, P.; Hofmann, K. P.; Choe, H.; Ernst, O. P. Crystal structure of the ligand-free G-protein-coupled receptor opsin. *Nature* **2008**, *454*, 183-187.

266. Choe, H.; Kim, Y. J.; Park, J. H.; Morizumi, T.; Pai, E. F.; Krauss, N.; Hofmann, K. P.; Scheerer, P.; Ernst, O. P. Crystal structure of metarhodopsin II. *Nature* **2011**, *471*, 651-655.

267. Stenkamp, R. E. Alternative models for two crystal structures of bovine rhodopsin. *Acta Crystallogr. D Biol. Crystallogr.* **2008**, *D64*, 902-904.

268. Blankenship, E.; Vahedi-Faridi, A.; Lodowski, D. T. The High-Resolution Structure of Activated Opsin Reveals a Conserved Solvent Network in the Transmembrane Region Essential for Activation. *Structure* **2015**, *23*, 2358-2364.

269. Park, J. H.; Morizumi, T.; Li, Y.; Hong, J. E.; Pai, E. F.; Hofmann, K. P.; Choe, H.; Ernst, O. P. Opsin, a structural model for olfactory receptors? *Angew. Chem. Int. Ed. Engl.* **2013**, *52*, 11021-11024.

270. Singhal, A.; Ostermaier, M. K.; Vishnivetskiy, S. A.; Panneels, V.; Homan, K. T.; Tesmer, J. J. G.; Veprintsev, D.; Deupi, X.; Gurevich, V. V.; Schertler, G. F. X.; Standfuss, J. Insights into congenital stationary night blindness based on the structure of G90D rhodopsin. *EMBO Rep.* **2013**, *14*, 520-526.

271. Szczepek, M.; Beyrière, F.; Hofmann, K. P.; Elgeti, M.; Kazmin, R.; Rose, A.; Bartl, F. J.; von Stetten, D.; Heck, M.; Sommer, M. E.; Hildebrand, P. W.; Scheerer, P. Crystal structure of a common GPCR-binding interface for G protein and arrestin. *Nat Commun* **2014**, *5*, 4801.

272. Deupi, X.; Edwards, P.; Singhal, A.; Nickle, B.; Oprian, D.; Schertler, G.; Standfuss, J. Stabilized G protein binding site in the structure of constitutively active metarhodopsin-II. *Proc. Natl. Acad. Sci. U. S. A.* **2012**, *109*, 119-124.

273. Gulati, S.; Jastrzebska, B.; Banerjee, S.; Placeres, Á L.; Miszta, P.; Gao, S.; Gunderson, K.; Tochtrop, G. P.; Filipek, S.; Katayama, K.; Kiser, P. D.; Mogi, M.; Stewart, P. L.; Palczewski, K. Photocyclic behavior of rhodopsin induced by an atypical isomerization mechanism. *Proc. Natl. Acad. Sci. U. S. A.* **2017**, *114*, E2615.

274. Singhal, A.; Guo, Y.; Matkovic, M.; Schertler, G.; Deupi, X.; Yan, E. C.; Standfuss, J. Structural role of the T94I rhodopsin mutation in congenital stationary night blindness. *EMBO Rep.* **2016**, *17*, 1431-1440.

275. Doré, A. S.; Robertson, N.; Errey, J. C.; Ng, I.; Hollenstein, K.; Tehan, B.; Hurrell, E.; Bennett, K.; Congreve, M.; Magnani, F.; Tate, C. G.; Weir, M.; Marshall, F. H. Structure of the adenosine A(2A) receptor in complex with ZM241385 and the xanthines XAC and caffeine. *Structure* **2011**, *19*, 1283-1293.

276. Congreve, M.; Andrews, S. P.; Doré, A. S.; Hollenstein, K.; Hurrell, E.; Langmead, C. J.; Mason, J. S.; Ng, I. W.; Tehan, B.; Zhukov, A.; Weir, M.; Marshall, F. H. Discovery of 1,2,4-triazine derivatives as adenosine A(2A) antagonists using structure based drug design. *J. Med. Chem.* **2012**, *55*, 1898-1903.

277. Hino, T.; Arakawa, T.; Iwanari, H.; Yurugi-Kobayashi, T.; Ikeda-Suno, C.; Nakada-Nakura, Y.; Kusano-Arai, O.; Weyand, S.; Shimamura, T.; Nomura, N.; Cameron, A. D.; Kobayashi, T.; Hamakubo, T.; Iwata, S.; Murata, T. G-protein-coupled receptor inactivation by an allosteric inverse-agonist antibody. *Nature* **2012**, *482*, 237-240.

278. Lebon, G.; Edwards, P. C.; Leslie, A. G. W.; Tate, C. G. Molecular Determinants of CGS21680 Binding to the Human Adenosine A2A Receptor. *Mol. Pharmacol.* **2015**, *87*, 907-915.

279. Segala, E.; Guo, D.; Cheng, R. K. Y.; Bortolato, A.; Deflorian, F.; Doré, A. S.; Errey, J. C.; Heitman, L. H.; IJzerman, A. P.; Marshall, F. H.; Cooke, R. M. Controlling the Dissociation of Ligands from the Adenosine A2A Receptor through Modulation of Salt Bridge Strength. *J. Med. Chem.* **2016**, *59*, 6470-6479.

280. Carpenter, B.; Nehmé, R.; Warne, T.; Leslie, A. G. W.; Tate, C. G. Structure of the adenosine A(2A) receptor bound to an engineered G protein. *Nature* **2016**, *536*, 104-107.

281. Sun, B.; Bachhawat, P.; Chu, M. L.; Wood, M.; Ceska, T.; Sands, Z. A.; Mercier, J.; Lebon, F.; Kobilka, T. S.; Kobilka, B. K. Crystal structure of the adenosine A2A receptor bound to an antagonist reveals a potential allosteric pocket. *Proc. Natl. Acad. Sci. U. S. A.* **2017**, *114*, 2066-2071.

282. Batyuk, A.; Galli, L.; Ishchenko, A.; Han, G. W.; Gati, C.; Popov, P. A.; Lee, M.; Stauch, B.; White, T. A.; Barty, A.; Aquila, A.; Hunter, M. S.; Liang, M.; Boutet, S.; Pu, M.; Liu, Z.; Nelson, G.; James, D.; Li, C.; Zhao, Y.; Spence, J. C. H.; Liu, W.; Fromme, P.; Katritch, V.; Weierstall, U.; Stevens, R. C.; Cherezov, V. Native phasing of x-ray free-electron laser data for a G protein-coupled receptor. *Sci Adv* **2016**, *2*, e1600292.

283. Canutescu, A. A.; Dunbrack, R. L. Cyclic Coordinate Descent: A Robotics Algorithm for Protein Loop Closure. *Protein Science* **2003**, *12*, 963-972.

284. Wang, C.; Bradley, P.; Baker, D. Protein-protein docking with backbone flexibility. *J. Mol. Biol.* **2007**, *373*, 503-519.

285. Mandell, D. J.; Coutsias, E. A.; Kortemme, T. Sub-Angstrom Accuracy in Protein Loop Reconstruction by Robotics-Inspired Conformational Sampling. *Nature Methods* **2009**, *6*, 551-552.

286. Stein, A.; Kortemme, T. Increased sampling of near-native protein conformations. *PLoS ONE* **2013**, *8(5)*.

287. Berkholz, D. S.; Driggers, C. M.; Shapovalov, M. V.; Dunbrack, R. L.; Karplus, P. A. Nonplanar peptide bonds in proteins are common and conserved but not biased toward active sites. *Proc Natl Acad Sci U S A* **2012**, *109*, 449-453.

288. Leaver-Fay, A.; O'Meara, M. J.; Tyka, M.; Jacak, R.; Song, Y.; Kellogg, E. H.; Thompson, J.; Davis, I. W.; Pache, R. A.; Lyskov, S.; Gray, J. J.; Kortemme, T.; Richardson, J. S.; Havranek, J. J.; Snoeyink, J.; Baker, D.; Kuhlman, B. Scientific Benchmarks for Guiding Macromolecular Energy Function Improvement. *Methods Enzymol* **2013**, *523*, 109-143.

289. O'Meara, M. J.; Leaver-Fay, A.; Tyka, M. D.; Stein, A.; Houlihan, K.; DiMaio, F.; Bradley, P.; Kortemme, T.; Baker, D.; Snoeyink, J.; Kuhlman, B. Combined Covalent-Electrostatic Model of Hydrogen Bonding Improves Structure Prediction with Rosetta. *J. Chem. Theory Comput.* **2015**, *11*, 609-622.

290. Park, H.; Bradley, P.; Greisen, P.; Liu, Y.; Mulligan, V. K.; Kim, D. E.; Baker, D.; DiMaio, F. Simultaneous Optimization of Biomolecular Energy Functions on Features from Small Molecules and Macromolecules. *J. Chem. Theory Comput.* **2016**, *12*, 6201-6212.

291. Alford, R. F.; Leaver-Fay, A.; Jeliazkov, J. R.; O'Meara, M. J.; DiMaio, F. P.; Park, H.; Shapovalov, M. V.; Renfrew, P. D.; Mulligan, V. K.; Kappel, K.; Labonte, J. W.; Pacella, M. S.; Bonneau, R.; Bradley, P.; Dunbrack, R. L.; Das, R.; Baker, D.; Kuhlman, B.; Kortemme, T.; Gray, J. J. The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design. *J. Chem. Theory Comput.* **2017**, *13*, 3031-3048.

292. Boomsma, W.; Hamelryck, T. Full cyclic coordinate descent: solving the protein loop closure problem in Cα space. *BMC Bioinformatics* **2005**, *6*.

293. Rotkiewicz, P.; Skolnick, J. Fast procedure for reconstruction of full-atom protein models from reduced representations. *J Comput Chem* **2008**, *29*, 1460-1465.

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

295. Kim, D. E.; Chivian, D.; Baker, D. Protein structure prediction and analysis using the Robetta server. *Nucleic Acids Res.* **2004**, *32*, 526.

296. Gerber, P. R.; Müller, K. MAB, a generally applicable molecular force field for structure modelling in medicinal chemistry. *J Computer-Aided Mol Des* **1995**, *9*, 251-268.

297. Jakalian, A.; Jack, D. B.; Bayly, C. I. Fast, efficient generation of high-quality atomic charges. AM1-BCC model: II. Parameterization and validation. *J Comput Chem* **2002**, *23*, 1623-1641.

298. Case, D.; Darden, T.; Cheatham, T.; Simmerling, C.; Wang, J.; Duke, R. E.; Luo, R.; Walker, R. C.; Zhang, W.; Merz, K.; Roberts, B.; Hayik, S.; Roitberg, A.; Seabra, G.; Swails, J.; Götz, A.; Kolossváry, I.; Wong, K. F.; Paesani, F.; Kollman, P. A. *AMBER 12, University of California, San Francisco;* 2012; Vol. 79, pp 926-935.

299. Arora, B.; Coudrat, T.; Wooten, D.; Christopoulos, A.; Noronha, S. B.; Sexton, P. M. Prediction of Loops in G Protein-Coupled Receptor Homology Models: Effect of Imprecise Surroundings and Constraints. *J. Chem. Inf. Model.* **2016**, *56*, 671-686.

300. Kmiecik, S.; Jamroz, M.; Kolinski, M. Structure Prediction of the Second Extracellular Loop in
G Protein-Coupled Receptors. *Biophysical Journal* **2014**, *106*, 2408–2416.

301. Raman, S.; Vernon, R.; Thompson, J.; Tyka, M.; Sadreyev, R.; Pei, J.; Kim, D.; Kellogg, E.; DiMaio, F.; Lange, O.; Kinch, L.; Sheffler, W.; Kim, B.; Das, R.; Grishin, N. V.; Baker, D. Structure prediction for CASP8 with all-atom refinement using Rosetta. *Proteins* **2009**, *77 Suppl 9*, 89-99.

302. Ferrer, M.; Chernikova, T. N.; Yakimov, M. M.; Golyshin, P. N.; Timmis, K. N. Chaperonins govern growth of Escherichia coli at low temperatures. *Nat. Biotechnol.* **2003**, *21*, 1266-1267.

303. Ferrer, M.; Chernikova, T. N.; Timmis, K. N.; Golyshin, P. N. Expression of a Temperature-Sensitive Esterase in a Novel Chaperone-Based Escherichia coli Strain. *Appl Environ Microbiol* **2004**, *70*, 4499-4504.

304. Massotte, D. G protein-coupled receptor overexpression with the baculovirus-insect cell system: a tool for structural and functional studies. *Biochim. Biophys. Acta* **2003**, *1610*, 77-89.

Appendix

Table A.1: Primers used in PCR for sub-cloning.
The forward primer sequences in bold correspond to the beginning (5') of the respective templates used in PCR. The reverse complement primer sequences in bold correspond to the end (3') of the respective templates. The underlined portion of the primer sequences correspond to the restriction enzyme recognition site. BsaI is unlike the other restriction enzymes in that it cuts outside of its recognition site (underlined portion without italics) to give non-palindromic overhangs for ligations. The underlined and italicized portion of the pE-SUMO forward primer is the AGGT overhang required for vector ligation.

Primer Identity	Restriction Enzyme	Primer sequence (5' – 3')
pRSET Forward	BamHI	GTAGCGCGCGGATCC**ATGGGTCATCACCATCATCATCACGGGTCCCTGC**
pRSET Reverse	EcoRI	CGCCCGCAACAGCGCCAGCGAATTC**CAGCAGGCTATCATTGGTGC**
pE-SUMO Forward	BsaI	GCAGATGTGGAGGTGGTCTCT*AGGT***ATGGGGCAACCCGGGAACGGC**
pE-SUMO Reverse	XhoI	CGCCAAGATGCCACCCTCGAG**TTACTTGTACAGCTCGTCCATGCCGAG**

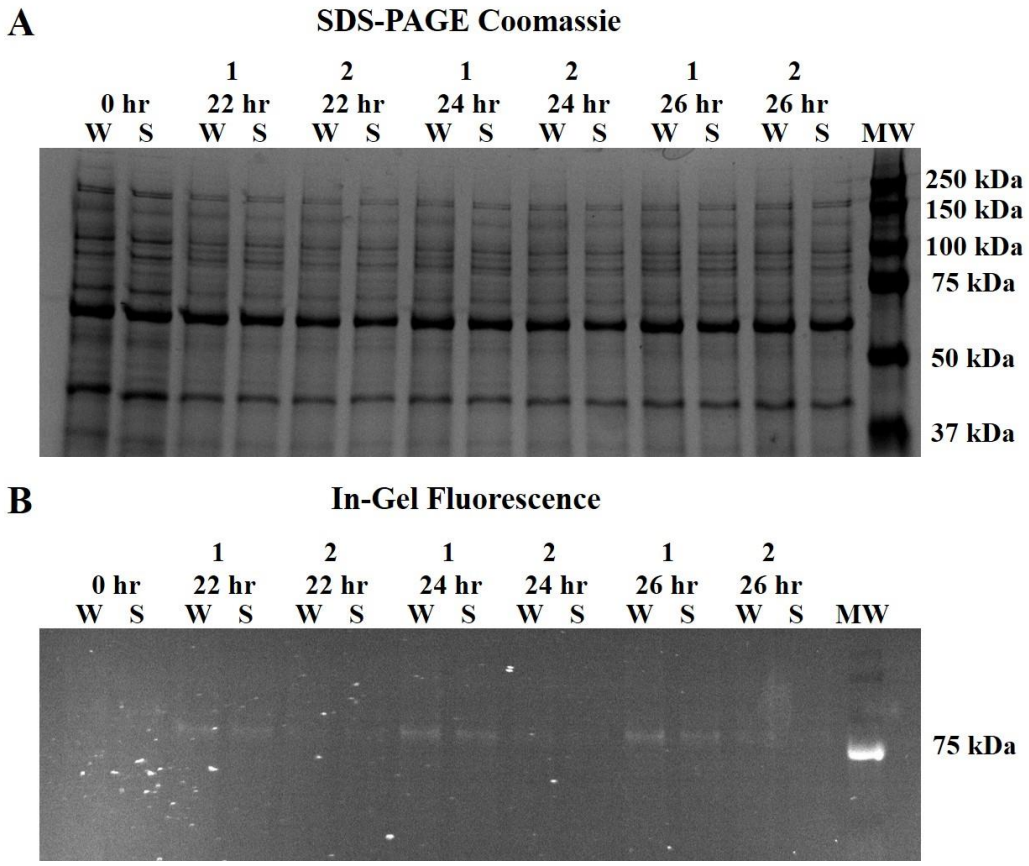

Figure A.1 Trial expression of SUMO-β_2AR:M13-EmGFP in ArcticExpress(DE3) Cells grown in TBE media at 12°C.

An O/N culture (1 mL) containing LB, 20 µg/mL gentamycin, 50 µg/mL ampicillin was inoculated with a single colony from the ArcticExpress (DE3)/SUMO-β2AR:M13-EmGFP transformation plate. The O/N culture was used to inoculate flasks 1 and 2 containing (25 mL) TBE media. The differences between the two culture conditions were, flask 1 contained no additional antibiotics (as recommended by the product manual) and flask 2 contained the same concentrations of antibiotics from the O/N culture (20 µg/mL gentamycin – selection pressure for Cpn60/10 expression plasmid, 50 µg/mL ampicillin – selection pressure for pE-SUMO plasmid). Cultures were grown at 30°C (250 rpm) until they reached an OD_{600} = 0.6 A.U. Subsequently, they were grown at 12°C (250 rpm) for 26 hours after inducing protein expression with 0.1 mM IPTG. A (1 mL) 0-hour sample was taken before IPTG addition and all samples taken afterwards were normalized to the 0-hour OD_{600} value. The whole-cell lysate and soluble lysate samples are represented by W and S, as in previous figures.