Knowledge-Based Modeling of G-Protein Coupled Receptors and their Binding Partners

By

Brian Joseph Bender

Dissertation

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

in

Pharmacology

January 31, 2019

Nashville, Tennessee

Approved:
Professor Vsevolod Gurevich, Chair
Professor Jens Meiler
Professor Heidi Hamm
Professor Tina Iverson
Professor Hassane Mchaourab
Professor Daniel Huster
Professor Christine Konradi

To April

ACKNOWLEDGEMENTS

This thesis is the work of so many people and it would be remiss to not acknowledge them here. First and foremost, I have to give thanks to my mentor Jens Meiler for really taking a chance on me. In my first meeting with him I openly told him I had no experience in a computational lab and he barely batted an eye at the concern. He understood what I realize now, computation is merely a tool, like protein expression or X-ray crystallography that I had to learn at one point, to ask the questions I wanted to know the answers to. His persistence that the mark of a scientist is their desire to learn and not what they already know truly defines my thesis research. I learned quickly that to become an expert I must first seek them out to guide and teach me. This of course led me to create a formidable committee who pushed me over the years to improve my research skills and, more importantly, my research questions. Professors Seva Gurevich, Heidi Hamm, Tina Iverson, and Hassaane Mchaourab all welcomed me into their labs for various experiments and provided me advice outside of the formal committee setting to which I am most grateful.

A special mention must go to Professor Daniel Huster of Leipzig University, also on my committee, who accepted me into his lab for almost a quarter of my thesis career. I came in as an untested experimentalist trying to become a novice computer scientist with the goal of learning protein expression of one of the more difficult and interesting families of proteins to work with. If it was not clear, this was also in a lab in a country that did not speak English. From day one of making buffers I had to realize that “Kalium” was not the same as “Calcium” if I wanted my cells to grow. Daniel, however, was never discouraging if I made a mistake and integrated me into his working group without a question.

The labs of Daniel Huster and also Professor Annette Beck-Sickinger of Leipzig University became my second home in graduate school. Drs. Peter Schmidt and Anette Kaiser were always there with a coffee (oder bier) to help get me situated in my foreign home. They responded to countless emails and explained to me various experiments over the years such that this thesis truly would not have been accomplished without them.

The lab of Jens Meiler here in Nashville, Tennessee could not have been a better group to work with over the last five years. Over countless hikes, karaoke performances, and patio gatherings, we developed a community unlike any other. I am deeply grateful to Drs. Amanda Duran and Yan Xia for leading the way in combining experiment and computation in their thesis work. The three of us all, likely unknowingly, bit off a real challenge. Combined with the leadership of Dr. Soumya Ganguly, we formed a tight-knit group to push the work of the Meiler lab into the difficult world of membrane protein biochemistry. On the computational side, thanks must go to Drs. Rocco Moretti and Steven Combs who put up with my endless questions as I was getting started learning Rosetta. Of course, there have been countless others who deserve mentioning, like Dr. Amandeep Sangha who taught me my first Bash script or Dr. Clara Schoeder who is the only person who shared my love of GPCRs, but the list would be nearly as long as the thesis itself.

Beyond the bench, myself and countless others would not have made it through graduate school without the tireless work of support staff. Heather Darling deserves a special mention for her support first in the wetlab and later as Jens' administrative assistant. From day one, she has helped me navigate this graduate school adventure whether it's teaching me where the pipette tips are stored or how to get a travel reimbursement. People like Karen Gieg and Karen Davis are some of the more incredible

people you will meet at Vanderbilt and yet likely get overlooked in their roles in getting us to where we are today. Trust me, if you could see Karen Davis on the dance floor you would never underestimate the energy and sheer awesomeness that she brings to the workplace on a daily basis.

Another often overlooked but incredibly vital component of my research here at Vanderbilt is the research community. Every university talks about being an open and collaborative working environment, but it is no better exemplified than here at Vanderbilt University. A special recognition must go to Dr. Walter Chazin who heads the Center for Structural Biology. While not formally a department, the CSB served as a home for all the people interested in structural biology but belonging to diverse research groups. I've met people in Biological Sciences, Pathology, Microbiology and Immunology, Chemical and Physical Biology, and Biochemistry beyond just my home department of Pharmacology that have been willing to talk shop and help me figure out experimental problems in structural biology which has been incredibly helpful over the years. This willingness to help extends beyond just graduate students and postdoctoral fellows but also to the professors—an incredibly unique and stimulating environment to call on when needed.

One more acknowledgement must go to my undergraduate research advisor Dr. Roger Rowlett. He offered me a spot in his crowded lab after many persistent emails and conversations. While not my first research experience, it was my first time in a lab in which I learned the excitement of waking up every day with a burning desire to ask questions in structural biology. He was a true believer in developing young scientists and not just working them for cheap labor. He taught me X-ray crystallography and inspired my research career. I still remember sitting in a lab meeting discussing these "new GPCR

crystal structures" and he said "People that work in GPCR structural biology are just crazy." Well, here we are. Importantly, he taught me that a scientist is a human with many varied interests and not just an encyclopedia. I learned that discussing science with a true colleague is just as easy as our conversations on flying Cessna planes or watching curling. Learning at an early stage that life outside the lab is just as important as life in the lab allowed me to prevent burnout and push myself through the long years of the PhD to keep testing new hypotheses.

Away from the university setting I am grateful to my friends here in Nashville, especially Alicia and Ryan who have each made room for me in their homes over the years. Danielle, a graduate student at Belmont, was always there to commiserate about this life we chose. Whitney and Nick are Nashville's super couple, and I could always count on a Selleh cake to cheer me up. Rae and Snow were there for me from day one and made life in Nashville an unforgettable experience.

Lastly, my family has been incredible through this whole experience. As the first in the family to go through a science PhD I dealt with questions about classes and summer breaks every year at family get-togethers. Despite a slight misunderstanding of what I was doing, they were always supportive and willing to listen to me talk about my research with attention even when the story didn't change from year to year. Over the course of the PhD we've experienced many joys and sorrows as a family and despite the distance have grown closer. Of special importance is my niece April who is exactly the same age as this thesis. Amazingly, this little five-year-old remembers her uncle at every visit home, and she has truly provided inspiration when the outlook was bleak. I couldn't have done this PhD without my family's continuing support.

TABLE OF CONTENTS

Page

TITLE PAGE i

COPYRIGHT ii

DEDICATION iii

ACKNOWLEDGEMENTS iv

TABLE OF CONTENTS viii

TABLE OF TABLES xx

TABLE OF FIGURES xxii

SUMMARY 1

CHAPTER 1: Peptide Binding at G-Protein Coupled Receptors 5

1.1 Summary 5

1.2 Introduction 5

1.2.1 G-Protein Coupled Receptors are a Significant Target of Therapeutic Intervention 5

1.2.2 Peptide Hormone Receptors Represent a Large Population of the GPCR Family 6

1.2.3 Diversity of Peptide Hormones 6

1.2.4 Diversity of Peptide Hormone Receptors 7

1.2.5 Complexity of Peptide Hormone and Receptor Interactions 8

1.3 Crystal Studies of Peptide Binding GPCRs 8

1.3.1 The First Peptide-Activated GPCR Structure 9

1.3.2 Crystallization of the First Peptide-Activated GPCR in Complex with a Peptide Ligand 9

1.3.3 Crystal Structures of Additional Peptide Ligands Reveal Diverse Binding Modes ... 10
1.3.4 Reduced Flexibility of Endogenous Peptide Ligands is Crucial for Crystallographic Success ... 11
1.3.5 Peptide Ligands Affect the Conformation of the Extracellular Surface ... 11

1.4 Structural Changes in Peptides Drive Binding at their Receptors ... 12

1.4.1 Neurotensin ... 13
1.4.2 Apelin ... 14
1.4.3 Endothelin ... 15
1.4.4 C5a ... 17
1.4.5 Ghrelin ... 18
1.4.6 Gonadotropin Releasing Hormone ... 20
1.4.7 Neuropeptide Y ... 21
1.4.8 Opioid Peptides ... 22
1.4.9 Secretin-Type Peptides ... 23

1.5 Implications for Future Studies ... 24

1.5.1 Peptides Need to be Characterized in their Bound State ... 24
1.5.2 Mimetics of the Bound-State Conformations can Aid in Structure Determination and Drug Discovery ... 26

1.6 Conclusions ... 27

CHAPTER 2: Protocols for Molecular Modeling with Rosetta3 and RosettaScripts ... 28

2.1 Summary ... 28

2.2 Introduction ... 28

2.1.1 Making Rosetta Accessible ... 30
2.1.2 Sampling and Scoring in Rosettta ... 34

2.2 De Novo Structure Prediction ... 40

2.2.1 Generating Peptide Fragments ... 41
2.2.2 TopologyBroker ... 41

2.2.3 Benchmarking 42
2.2.4 Limitations of *De Novo* 42

2.3 Comparative Modeling 43

2.3.1 Starting Templates. 44
2.3.2 Three Stages of Multi-Template Comparative Modeling. 44
2.3.4 Modeling Loops. 46

2.4 Protein-Protein Docking 47

2.4.1 Global vs. Local Docking. 47
2.4.2 Low Resolution vs. High Resolution Docking 48
2.4.3 Improvements to RosettaDock 48
2.4.4 FlexPepDock 49

2.5 Protein-Small Molecule Docking 50

2.5.1 Improvements to RosettaLigand 50
2.5.2 Customizable Small Molecule Docking Protocols 51
2.5.3 Small Molecule Docking in Membrane Proteins 52

2.6 Incorporating Experimental Data 53

2.7 Protein Design 55

2.7.1 “Inverse Folding Problem” 55
2.7.2 Design for Stability 56
2.7.3 Design for Functionality 57

2.8 Additional Rosetta Methods 59

2.8.1 Membrane 59
2.8.2 Non-canonical Amino Acids and Non-canonical Backbones 59

2.9 Conclusions 61

CHAPTER 3: Rosetta's Predictive Ability for *in silico* Fragment-Based Drug Discovery 63

3.1 Summary 63

3.2 Introduction 63

3.2.1 Computer-Assisted Drug Discovery Is an Important Component in Drug Discovery 63
3.2.2 SB-CADD can Rapidly Advance Initial Hit Discovery 64
3.2.3 RosettaLigand has Two Score Functions for Docking and Scoring 65
3.2.4 Fragment-Based Drug Discovery Identifies Hits Using a Focused Library 65
3.2.5 NMR can Determine Affinity and Binding Pose Simultaneously in FB-DD 66
3.2.6 RosettaLigand has not been Tested in *in silico* FB-DD 66

3.3 Methods 67

3.3.1 HisF Mutagenesis, Expression, and Purification 67
3.3.2 Crystallization of HisF-C9S 68
3.3.3 NMR Spectra and Assignment Transfer 68
3.3.4 Small Molecule Screening 69
3.3.5 NMR Ligand Titration to Determine Binding Affinity 70
3.3.6 RosettaLigand Blind Docking of Fragments 70
3.3.7 Assessment of RosettaLigand's Discrimination of Actives from Inactives 701
3.3.8 ROC Analysis of Rosetta Binding Site Predictions 71
3.3.9 Surflex-Sim Alignments to Binding Mode Hypotheses 72

3.4 Results 73

3.4.1 Selection of HisF-C9S for NMR and Computational Experiments 73
3.4.2 NMR Assignment of HisF-C9S Spectra 75
3.4.3 Identification of Potential Hits to HisF-C9S 75
3.4.4 Binding Analysis of Individual Hits 76
3.4.5 Binding Mode of the Small Molecules 77
3.4.6 RosettaLigand Can Discriminate Active from Inactive Drug Fragments 79
3.4.7 High Density Docking of Active Compounds with RosettaLigand 80
3.4.8 Structure Based Fragment Docking Captures Experimentally Determined Binding Pocket 80

3.4.9 Superimposition of HisF-C9S Naïve Binders with HisF/Substrate Complex 82

3.5 Discussion 83

3.5.1 Identification of Set of Low Affinity Ligands for HisF-C9S 83
3.5.2 RosettaLigand Accurately Predicts Binding of Low Affinity Fragments 84
3.5.3 Structure-Based Ligand Docking Results in Similar Ligand Orientation as Ligand-Based Results 84

3.6 Conclusion 85

3.7 Additional Scripts 86

CHAPTER 4: Redesign of HisF to Increase Binding to a Novel Ligand using RosettaLigand 90

4.1 Summary 90

4.2 Introduction 90

4.2.1 Gene Duplication Allows A Protein to Evolve to Bind a Novel Ligand 90
4.2.2 Protein-Ligand Interface Design Often Fails to Match Evolutionary Success 91
4.2.3 Protein-Interface Design Can Mimic Evolutionary Pathways 92

4.3 Methods 93

4.3.1. Ligand Docking and Interface Design 93
4.3.2 Expresion of HisF mutants 94
4.3.3 NMR Titration of VU0068924 against HisF variants 94

4.4 Results 95

4.4.1 Pocket Optimization of Ligand VU0068924 95
4.4.2 Rosetta Designs Surface Mutations in addition to Pocket Mutations 97
4.4.3 Incorporation of Rosetta-predicted Mutations did not Alter the Tertiary Structure of HisF 98
4.4.4 Design Mutation Greatly Enhances Binding Affinity of Ligand 99
4.4.5 Identification of Critical Mutations that Enhance Binding Affinity 99
4.4.6 Pairwise Mutations Identify D130M and S144W as most potent for binding

enhancement 100
4.4.7 Rosetta Score Function Does Not Distinguish Between Critical and Non-Critical Mutations 102
4.5 Discussion 103
4.5.1 RosettaLigand can Blindly Design Protein Ligand Interfaces 103
4.5.2 The Effect of Each Mutation is Hard to Predict 104
4.5.3 Avenues for Further Improvement to Affinity 105
4.6 Conclusions 103
CHAPTER 5: RosettaGPCR: An Improved Method for Multiple Template Homology Modeling of GPCRs with Rosetta 107
5.1 Summary 107
5.2 Introduction 107
5.2.1 G-protein coupled receptors represent important therapeutic targets 107
5.2.2 GPCRs are Defined by Their Fold 108
5.2.3 Atomic Structures of GPCRs are Limited but Growing 108
5.2.4 Computational Modeling can Extend our Current Understanding of GPCR Structure 109
5.2.5 Several GPCR-specific Homology Modeling Protocols Exist 110
5.2.6 Rosetta Hybridizes Multiple Templates 111
5.2.7 Development of a GPCR-Specific Multiple-Template Homology Modeling Protocol in Rosetta 112
5.3 Methods 113
5.3.1 Description of Benchmark Data Set 113
5.3.2 Generation of Alignments for Modeling 114
5.3.3 Template Selection 117
5.3.4 Generation of additional input files 117
5.4.4 Model Production 117
5.4 Results 119

5.4.1 Blended Sequence- and Structure-Based Alignment Is Critical for Modeling Success 119
5.4.2 Peptide Fragment Hybridization Improves Target Model Quality 121
5.4.3 Multiple Templates Improves Performance with Low Sequence Identity Templates 122
5.4.4 Identification of the Optimal Number of Templates 124
5.4.5 RosettaGPCR Outperforms other GPCR Modeling Servers 125
5.4.6 Accuracy of Models with Increasingly Worse Templates 126
5.4.7 Development of Database for All Human Non-Odorant GPCRs 127

5.5 Discussion 128

5.5.1 Blending of Sequence- and Structure-Based Alignment are Critical for Low Identity Template Based Modeling 128
5.5.2 Template and Peptide Hybridization are Key Drivers for Accurate Modeling 131
5.5.3 Development of the First Rosetta-based GPCR Database 133

5.6 Conclusions 134

CHAPTER 6: Structural Model of Ghrelin Bound to its G Protein Coupled Receptor 135

6.1 Summary 135

6.2 Introduction 135

6.3 Methods 138

6.3.1 Experimental Models and Subjects Details 138
6.3.2 Method Details 139
6.3.3 Quantification and Statistical Detail 146
6.3.4 Data and Software Availability 146

6.4 Results 147

6.4.1 Expression and Functional Refolding of GHSR 147
6.4.2 Investigation of Ghrelin Partitioning between Membrane- and Receptor-Bound States 148
6.4.3 Determination of Ghrelin Segments in Close Contact with GHSR by STD NMR 150

6.4.4 Structural Features of Receptor-Bound Ghrelin 152
6.4.5 Modeling of the Ghrelin/GHSR Complex 154
6.4.6 Characterization of the Ghrelin/GHSR Complex 157
6.4.7 Validation of Binding Interface via Mutagenesis 160

6.5 Discussion 163

6.5.1 Ghrelin Binds to GHSR via an Extended Binding Surface 163
6.5.2 The Binding Model of Ghrelin Adopts a Fairly Small Structural Ensemble 165
6.5.3 A Hierarchical Approach to Modeling Highly Flexible Protein-Ligand Complexes in Rosetta 166

6.6 Conclusions 168

CHAPTER 7: Structural Basis of Ligand Binding Modes at the Neuropeptide Y Y_1 receptor 169

7.1 Summary 169

7.2 Introduction 169

7.3 Methods 170

7.3.1 Cloning and Protein Expression 170
7.3.2 Purification of Y_1R–UR-MK299 and Y_1R–BMS-193885 Complexes 171
7.3.3 Lipidic Cubic Phase Crystallization of Antagonist-Bound Y_1Rs 173
7.3.4 Data Collection and Structure Determination 174
7.3.5 Immunoblotting 176
7.3.6 Radio Ligand Binding Assay 176
7.3.7 IP Accumulation Assay 179
7.3.8 Live Cell Fluorescence Microscopy 180
7.3.9 Quantification of Receptor Surface Expression in COS-7 cells 181
7.3.10 Peptide Synthesis 181
7.3.11 NMR Measurements of Y_1R-bound NPY 182
7.3.12 Molecular Docking of NPY into Y_1R 184
7.3.13 Photo-Crosslinking Experiment between Y_1R and NPY N-terminus 186
7.3.14 Data Availability 187

7.4 Results and Discussion 188

7.4.1 Characterization of the Y1R Crystal Structures 188
7.4.2 The UR-MK299 Binding Site 189
7.4.3 The BMS-193885 Binding Site 193
7.4.4 Modeling the NPY Binding Mode at Y1R 195
7.4.5 Differences between Agonist and Antagonist Binding Modes 197
7.4.6 Identification of the Role of the NPY N-terminus in Y1R Binding 200

CHAPTER 8: Conslusions 202

8.1 Summary 202

8.2 Key Findings 204

8.3 Future Outlook 205

8.3.1 Challenges of Docking into Homology Models 205
8.3.2 Extension of GPCR Model Database 207
8.3.3 Generalization of Peptide-Binding Studies 209
8.3.4 Modeling of GPCR Conformational Dynamics 209

APPENDIX A: Rosetta and the Design of Ligand Binding Sites 215

A.1 Summary 215

A.2 Introduction 215

A.3 Materials 218

A.4 Methods 219

A.5 Notes 226

APPENDIX B: α_{2A}- and α_{2C}-Adrenoceptors as Potential Targets for Dopamine and Dopamine

Receptor Ligands 233

B.1 Summary 233

B.2 Introduction 234

B.3 Methods 236

B.3.1 DNA Constructs and Transfection 236
B.3.2 BRET 237
B.3.3 DMR 237
B.3.4 Adenylyl Cyclase Activity 238
B.3.5 ERK1/2 Phosphorylation 239
B.3.6 Radioligand Binding 240
B.3.7 Binding Data Analysis 241
B.3.8 Statistical Analysis 243
B.3.9 Drugs 243
B.3.10 Homology Modeling of α_{2A}- and α_{2C}-Adrenoceptors 243
B.3.11 Protein-Ligand Docking 244

B.4 Results 245

B.4.1 Binding of DA and DA Receptor Ligands to α_2-Adrenoceptors in Cortical and Striatal Tissue 245
B.4.2 Binding of DA and NE to D_1-Like and D_2-Like Receptors in Striatal Tissue 247
B.4.3 α_{2A}- and α_{2C}-Adrenoceptor-Mediated G Protein Activation by DA and Synthetic DA Receptor Ligands 250
B.4.4 α_{2A}- and α_{2C}-Adrenoceptor-Mediated Effects of NE and DA on Adenylyl Cyclase Activity 257
B.4.5 α_{2A}- and α_{2C}-Adrenoceptor-Mediated Effects of NE, DA, and Synthetic DA Receptor Ligands on ERK1/2 Phosphorylation 259
B.4.6 Structural Basis for DA at α_{2A}- and α_{2C}-Adrenoceptors 260

B.5 Discussion 262

B.6 Conclusions 266

APPENDIX C: Modeling the Complete Chemokine-Receptor Interaction 267

C.1 Summary 267

C.2 Introduction 267

C.2.1 Overview 267
C.2.2 Structure-based modeling of chemokine-GPCR complexes 268
C.2.3 Modeling Strategy 272

C.3 Materials 273

C.3.1 Getting started 276
C.3.2 Target Sequence 277
C.3.3 Generate Fragments 277
C.3.4 Templates 278
C.3.5 Alignment 279
C.3.6 Define Topology 280

C.4 Methods 281

C.4.1 Threading 281
C.4.2 Hybridize 282
C.5 Data Analysis and Interpretation 284

C.5.1 Calculating RMSD 284
C.5.2 Decoy Clustering 285
C.5.3 Interface analyzer 287
C.5.4 Coordinate-based Metrics 289

C.6 Selection of a Final Model 291

C.7 Concluding Remarks 293

C.7.1 Scope of Modeling 293
C.7.2 Online Resources 293
C.7.3 Next Steps 294

APPENDIX D: Improved *In Vitro* Folding of the Y2 G Protein-Coupled Receptor into Bicelles 295

D.1 Summary 295

D.2 Introduction 295

D.3 Methods 299

D.3.1 Y2R Sample Preparation 299
D.3.2 Negative Stain Electron Microscopy 300
D.3.3 Assessment of Disulfide Bridge Formation 300
D.3.4 Fluorescence Polarization Ligand Binding Assay 301
D.3.5 G-Protein Activation in vitro 301
D.3.6 Peptide Synthesis 303
D.3.7 NMR Measurements 303

D.4 Results 304

D.4.1 In vitro Folding of the Y2R into Bicelles 304
D.4.2 Disulfide Bridge Formation of Refolded Y2R 308
D.4.3 Fluorescence-Based Ligand Binding of Refolded Y2R 309
D.4.4 G-protein Activation by Refolded Y2R 311
D.4.5 NMR Experiments to Assess Receptor Function in Micromolar Concentration 312
D.4.6 Carbon-Carbon Correlation MAS NMR Measurement 313

D.5 Discussion 315

D.6 Conclusions 315

BIBLIOGRAPHY 321

TABLE OF TABLES

Page

Table 1.1 Examples of Peptide Hormone Modifications 7

Table 2.1 Publicly Accessible Web Servers Running Rosetta 33

Table 2.2 Standard Rosetta Score Function Terms 36

Table 3.1 Crystallographic Data Collection and Refinement Statistics for HisF-C9S (PDB ID: 5TQL) 74

Table 4.1 Measured KD's for compound VU0068924 for Each Combination of Tested Mutations 101

Table 5.1 List of Receptors in Benchmark Set 114

Table 5.2 List of Templates for Each Target Ranked by Sequence Identity 118

Table 6.1 Peptides used in this study. Residues in red and underlined were $^{13}C/^{15}N$-labeled. 141

Table 6.2 Relative STD effects for the assigned protons of ghrelin in the presence of GHSR 152

Table 6.3 ^{13}C NMR chemical shifts determined for GHSR bound ghrelin 154

Table 6.4 Energetic analysis of the ghrelin/GHSR complex. 159

Table 6.5 Mutational analysis of GHSR binding site 162

Table 7.1 Data collection and refinement statistics 175

Table 7.2 IP accumulation assays of wild-type (WT) and mutant Y1Rs for NPY/NPY analogues. 197

Table B.1 Competitive inhibition experiments of [^{3}H]RX821002 versus NE, DA, clonidine, and D_2-like receptor ligands in the sheep brain cortex and striatum 248

Table B.2 Competitive inhibition experiments of [^{3}H]SCH 23390, [^{3}H]YM-09151-2, or

[^{3}H]RX821002 versus DA and NE in the sheep brain striatum 249

Table B.3 Potency of NE, DA, clonidine, 7-OH-PIPAT, and quinpirole obtained from G protein activation experiments mediated by α_{2A}- and α_{2C}-adrenoceptors coupled to the different Gαi/o subtypes 252

Table B.4 Efficacy of NE, DA, clonidine, 7-OH-PIPAT, and quinpirole obtained from G protein activation experiments mediated by α_{2A}- and α_{2C}-adrenoceptors coupled to the different Gαi/o subtypes 2523

TABLE OF FIGURES

Page

Figure 2.1: Multi-template comparative modeling with Rosetta. 45

Figure 2.2 Protein-peptide interface prediction using FlexPepDock ab-initio..... 49

Figure 2.3 Application of RosettaLigand docking of negative allosteric modulator MPEP into comparative model of mGlu5 transmembrane domain. 53

Figure 2.4 Design of protein-ligand interactions for high affinity and selectivity. 58

Figure 3.1 The HisF binding pocket..... 73

Figure 3.2 Identification of Low Affinity Fragments for HisF-C9S. 77

Figure 3.3 Full dataset for the 31 low affinity ligands identified for HisF-C9S. 78

Figure 3.4 ROC Analysis of Rosetta Predictions. 79

Figure 3.5 Iterative Docking to Identify Binding Pockets 81

Figure 3.6 Binding hypotheses by computational modeling. 82

Figure 4.1 Design Strategy for Compound VU0068924. 96

Figure 4.2 The final set of mutations for HisF design. 97

Figure 4.3 Incorporation of Rosetta Designs Does not Affect the Tertiary Structure of HisF .. 98

Figure 4.4 NMR Titration of VU0068924 against HisF 99

Figure 4.5 Identification of Residues Critical for Enhanced Binding to VU0068924..... 100

Figure 4.6 Comparison of Measured versus Predicted Binding Affinities for Critical and Non-Critical Mutations..... 102

Figure 5.1 Alignment of Receptors in the Benchmark..... 116

Figure 5.2 Comparison of Average RMSD Change using Various Alignment Methods. 119
Figure 5.3 Comparison of Single Template Modeling Methods with Peptide Insertion. 121
Figure 5.4 Comparison of Average RMSDs for Single versus Multiple Template Homology Modeling. 123
Figure 5.5 Comparison of Model Accuracy using Various Numbers of Starting Templates .. 124
Figure 5.6 Results of Novel Structure Prediction from Various GPCR Modeling Servers 125
Figure 5.7 Model Accuracy with Templates over Multiple Sequence Identity Ranges 126
Figure 5.8 Percent Identify of Best Available Template for Every Non-Odorant Human GPCR 128
Figure 6.1 In vitro binding assay of GHSR 147
Figure 6.2 ^{2}H NMR analysis of ghrelin binding to GHSR and membrane fragments 149
Figure 6.3 ^{1}H NMR saturation transfer difference spectroscopy of ghrelin interaction at GHSR 151
Figure 6.5 Iterative Ligand Docking and Loop Building of Ghrelin/GHSR-complex 156
Figure 6.6 Rosetta modeling of ghrelin at GHSR 157
Figure 6.7 Variability of Final Ensemble of Ghrelin at GHSR 158
Table 6.4 Energetic analysis of the ghrelin/GHSR complex 159
Figure 6.8 Impact of the residues with the highest Rosetta energy score to G-protein signal transduction 160
Figure 6.9 Fluorescence microscopy of transiently transfected HEK293 with GHSR eYFP fusion proteins 161
Figure 7.1 Overall structures of Y1R-UR-MK299 and Y1R-BMS-193885 complexes 188
Figure 7.2 Crystal packing and structural features of Y_1R and chemical structures of Y_1R

ligands 190

Figure 7.3 Ligand-binding pocket of Y_1R for UR-MK299 and BMS-193885 192

Figure 7.4 IP accumulation assays 194

Figure 7.5 Pharmacological characterization of refolded Y_1R and NMR studies of Y_1R-bound NPY 196

Figure 7.6 Docking poses of NPY 198

Figure 7.7 Photo-crosslinking experiments between NPY and Y_1R 201

Figure A.1 Flowchart of RosettaLigand design protocol 217

Figure A.2 Protein/ligand interface design with RosettaLigand 225

Figure B.1 Radioligand binding to D2-like and D1-like receptors in the brain striatum 245

Figure B.2 Radioligand binding of dopaminergic and adrenergic ligands to α_2-adrenoceptors in brain tissue 246

Figure B.3 G protein activation of α_{2A} by dopaminergic and adrenergic ligands 250

Figure B.4 G protein activation of α_{2C} by dopaminergic and adrenergic ligands 250

Figure B.5 Specificity of the effect of dopaminergic ligands on α_{2A}- and α_{2C}-adrenoceptors. 255

Figure B.6 DMR induced by NE, DA, and synthetic DA receptor ligands using label-free assay 256

Figure B.7 Effect of NE and DA on the modulation of adenylyl cyclase activity by activation of α_{2A}- and α_{2C}-adrenoceptors 258

Figure B.8 NE, DA, and synthetic DA receptor ligands signaling via ERK1/2 phosphorylation 260

Figure B.9 Binding mode of DA at each receptor type 261

Figure C.1 Chemokine binding and receptor activation is as a two-step, two-site process 270

Figure C.2 Timeline of chemokine receptor crystal structures 271

Figure C.3 Summary of homology modeling protocol 274

Figure C.4 The input structures for this tutorial are summarized 275

Figure C.5 Divergence of low homologous structures..... 282

Figure C.6 Rosetta and coordinate metrics used to select final models..... 286

Figure C.7 System of 3D metrics for quick analysis..... 290

Figure C.8 Overview of inputs and outputs 292

Figure D.1 Scheme of the three-step folding protocol for the preparation of the Y_2R in either isotropic or non-isotropic bicelles..... 305

Figure D.2 Negative staining electron microscopy images of (A) small isotropic bicelles (q = 0.25), (B) intermediate sized bicelles, and (C) non-isotropic bicelles (q > 10) 308

Figure D.3 Results of the CPM assay for testing disulfide bridge formation 309

Figure D.4 Pharmacological characterization of the Y_2R preparation at nanomolar concentration using a fluorescence polarization assay with [Dpr^{22}-atto520]-NPY 310

Figure D.5 *In vitro* folded Y_2R variants functionally activate purified G_i protein 311

Figure D.6 NPY binding tests of the Y_2R preparation in concentration of 50 μM using solution NMR spectroscopy..... 312

Figure D.7 Solid-state MAS NMR spectra of uniformly labeled Y_2R in non-isotropic bicelles showing $^{13}C/^{13}C$ correlation using DARR 311

SUMMARY

G-protein coupled receptors (GPCRs) represent the largest family of membrane proteins and one of the most heavily targeted for therapeutic intervention. Unfortunately, the current understanding of GPCR structures is limited to the ~14% of available crystallized receptor structures representing a large knowledge gap for understanding of current drug therapies and future drug development platforms. Herein, I present a series of protocols and studies for structural characterization of GPCRs using a combination of molecular modeling and experimental data.

Chapter 1 introduces a specific family of GPCRs that bind peptide ligands. This family represents about one quarter of the druggable GPCR superfamily. However, few studies have characterized the structural basis of peptide recognition. Using a series of receptor/peptide case studies, I lay out a hypothesis for conformational selection as a driver of peptide recognition. These studies demonstrate the need for hybrid method approaches for studying this complex set of binding patterns. This combination of different methods for GPCR structural characterization is the general theme that my thesis work presents. This chapter is being developed for a review paper where I will be first author.

Chapter 2 presents an overview of the Rosetta modeling suite and specific protocols that are used repeatedly throughout my thesis. This is from the review paper “Protocols for Molecular Modelling with Rosetta3 and RosettaScripts” for which I am the first author. This review was developed from our semi-annual Rosetta Workshop that I have taught at for the last 6 iterations. The protocols presented in this chapter are those which I have either directly contributed to or heavily used in my thesis work.

Chapter 3 demonstrates Rosetta's capability for *in silico* fragment-based drug discovery in a model protein. This work combines computational ligand docking with NMR-based screening and X-ray crystallography. The iterative docking approach that I developed in this paper is later extended to my work on GPCRs. This chapter comes from the research article "Rosetta's Predictive Ability for *in silico* Fragment-Based Drug Discovery" for which I am co-first author.

Chapter 4 extends on the ligand docking protocol presented in Chapter 3 to incorporate protein design. Again, iterative docking is described, and techniques covered include computational methods, NMR spectroscopy, and X-ray crystallography. While this work is in a model system and not a GPCR, the protocols developed here are useful in redesigning any protein binding pocket to recognize novel ligands. This technology is actively being pursued for GPCRs, known as DREADD technology. This work is a research article for which I am first author.

Chapter 5 describes the development of a GPCR-specific homology modeling protocol within Rosetta. This work extends and significantly improves on previous methods in the Rosetta community. I developed new rules for modeling structurally conserved, sequence diverse proteins that push the minimum accepted homology into a region that describes most GPCRs. With the new method, we can confidently model 88% of the GPCR superfamily open up avenues for novel drug development platforms. This chapter is a research article for which I am sole first author.

Chapter 6 utilizes the improved modeling of GPCRs and iterative docking combined with experimental data to characterize the ghrelin peptide binding at its GPCR. My new approach to modeling highly flexible binding partners is useful to the field of peptide-GPCR recognition events because it allows for dense sampling of conformationally

realistic binding solutions. This work demonstrates the effects of conformational selection of a peptide binding to its receptor. This chapter comes from the article "Structural Modeling of Ghrelin Binding its G-Protein Coupled Receptor" for which I am the first author.

Chapter 7 is similar in technique to Chapter 6 but characterizes the neuropeptide Y binding to the Y1 receptor. In addition to iterative docking of a peptide and combining multiple sets of experimental data, I helped to refine the crystal structure of the Y1 receptor bound to a small molecule antagonist. This work used the protocols developed in every chapter of this thesis except for Chapter 4. This chapter comes from the article "Structural Basis for Ligand Binding Modes at the Neuropeptide Y Y1 Receptor" for which I am a co-first author.

Appendix A is related to Chapter 4 and is a methods paper describing the design of protein-ligand interfaces using the protocol that I developed. This generalizes the protocol for any protein of interest, including GPCRs. This appendix is from the article "Rosetta and the Design of Protein-Ligand Interfaces" for which I am second author.

Appendix B is related to Chapters 3 and 5. This combines GPCR structure prediction and small molecule ligand docking to develop a working hypothesis for cross-talk of ligands between dopaminergic and alpha-adrenergic receptors. I contributed the modeling simulations to this mostly experimental body of work. This appendix is from the article "α_{2A}- and α_{2C}-Adrenoceptors as Potential Targets for Dopamine and Dopamine Receptor Ligands" for which I am a contributing author.

Appendix C is related to Chapter 5 and extends the general protocol of GPCR homology modeling to chemokine receptor complexes with chemokine ligands. This is a protocols paper that is developed entirely based on the method described in Chapter 5. I

worked with the first author to modify this specifically for chemokine/chemokine receptor modeling but it serves as a general protocol for modeling protein-protein interactions of GPCRs. This appendix is from the article “Modeling the Complete Chemokine/Chemokine Receptor Interaction” for which I am a contributing author.

Appendix D is related to Chapters 6 and 7 and describes our *in vitro* expression and refolding protocol for GPCRs for use in structural characterization. This appendix focuses on refolding of the neuropeptide Y type 2 receptor and validation of the function of these refolded receptors but the same principles apply to the ghrelin and neuropeptide Y type 1 receptors. This appendix comes from the article “Improved *in vitro* Folding of the Y2 G-Protein Coupled Receptor into Bicelles” for which I am the second author.

CHAPTER 1

Peptide Binding at G-Protein Coupled Receptors

1.1 Summary

This chapter introduces the structural knowledge and known issues in characterization of peptide ligands at GPCRs. Over a series of receptor/peptide case studies, I lay out a hypothesis for conformational selection as a driver for peptide recognition. This review demonstrates the need for hybrid approaches to GPCR structural biology. The chapter is under development for future publication as a review article, where I anticipate being first author.

1.2 Introduction

1.2.1 G-Protein Coupled Receptors are a Significant Target of Therapeutic Intervention

With more than 800 members, G-protein coupled receptors (GPCRs) are the largest family of human transmembrane proteins [11]. They are key player in a number of physiological functions, regulate the majority of cellular processes, and are involved in numerous disease pathologies [12]. Their substantial involvement in cellular signaling has established them as highly relevant pharmacological drug targets. About 34% of all drugs approved by the US Food and Drug Administration (FDA) achieve their therapeutic effects through GPCRs [13]. In the top 20 drug targets in terms of commercial impact are drugs targeting GPCRs for heart disease (olmesartan: AT1), diabetes (lixisenatide; dulaglutide:

GLP1R), and schizophrenia (aripiprazole: 5HT2A/5HT1A/DRD2) [12-14]. Additional targets of GPCRs include treatments against HIV [15], Alzheimer's disease [16], and a variety of cancer onco-targets [17, 18]. This impact is represented by the fact that the sales of GPCR targeting drugs have an estimated share of >27% of the global market [13].

1.2.2 Peptide Hormone Receptors Are a Large Percentage of the GPCR Family

This variety of drugs targeting GPCRs reflects the diversity of chemical signals that can interact with GPCRs including small molecules, lipids, ions, and proteins [19, 20]. In particular, the peptide (and protein) activated receptors are found to be about 25% of the pharmacologically relevant receptors (this excludes the odorant receptors). Peptide-activated receptors are found across all rhodopsin-like subfamilies (α, β, γ, and δ) and the entire secretin family [21]. Given this coverage, it is unsurprising that many of the blockbuster drugs mentioned above (e.g. olmesartan and dulaglutide) are targeted at members of this group of receptors. With such importance among current and future therapeutic development, a full understanding of these receptors is necessary. In this review, we plan to cover what is known about these receptors structurally using a variety of biophysical techniques and provide suggestions for future routes of discovery.

1.2.3 Diversity of Peptide Hormones

Peptide ligands come in all shapes and sizes though share the common theme that they are produced by ribosomal translation. Often these, peptide hormones are produced as pre-hormones that are subsequently processed to their active component. As a result, peptide hormones range in size from 3 amino acids (TRH) to >100 amino acids (chemokines). In addition to size differences, many peptide hormones undergo post-

translational modifications. Some of these modifications are necessary to increase peptide half-life by inhibition of exopeptidases such as N-terminal pyroglutamation and C-terminal amidation. However, in some cases these modifications serve dual purposes by acting as sites of molecular recognition in their cognate receptor [2]. Further types of post-translational modifications include lipidation, bromination, and disulfide bridge formation. A summary of modifications is found in Table 1. These modifications further increase the diversity of chemical space available to peptide hormones beyond the canonical 20 amino acids. The size, sequence, and chemical diversity allows for a vast degree of specificity between peptide hormones and their receptors. Further, it is common for a given peptide hormone to exist in multiple isoforms, such as the family of neuropeptide Y which consists of NPY, PYY, and PP and the endothelin peptides ET-1, ET-2, and ET-3.

Table 1.1 Examples of Peptide Hormone Modifications

Peptide Modification	Example
C-terminal amidation	NPY, PYY
N-terminal pyroglutamic acid	TSH
Bromination	NPBW
Lipidation	Ghrelin
Disulfide bridge formation	Endothelin?
Differential proteolysis	Bradykinin, kallikrein, NPY vs NPY3-36, Ape-13 vs Ape-17 vs Ape 36

1.2.4 Diversity of Peptide Hormone Receptors

Peptide hormone receptors are believed to have undergone gene duplication events over the course of human evolution. It is rare for a given peptide hormone to target a single receptor as is the case for the ghrelin and motilin receptors. Many hormone receptors have at least two subtypes (i.e. orexin and endothelin), and in the case of chemokine receptors, a large number of receptor subtypes exist. These subtypes increase the modes and sites of

action of peptide hormones through G-protein and tissue selectivity, respectively.

1.2.5 Complexity of Peptide Hormone and Receptor Interactions

This diversity of peptide hormone receptors results in a common theme of peptide hormone biology: that of multi-hormone/multi-receptor interactions. This increases the challenges of studying a given peptide hormone with its receptor because the binding mode of a hormone may differ between receptor subtypes as evidenced by NPY binding at its various receptors [22]. The same can be true at a given receptor that the binding mode of one peptide hormone may be different than the binding mode of a highly related peptide hormone as was recently demonstrated with the bradykinin receptors [23]. This prevents overarching binding and activation mechanisms that can explain this family of receptors as a whole, unlike what is known about aminergic receptors. As such, it is critical for full understanding of receptor/hormone biology to study each combination in detail before attempting to formulate generalizations that can be used for future drug development. This is a monumental task though many efforts are ongoing to attempt this feat.

1.3 Crystal Studies of Peptide Binding GPCRs

Since the first crystal structure of a GPCR that responds to a diffusible ligand in 2007 [24], we now have structures of about 50 unique receptors. Of these 50, about half represent peptide activated receptors. In this section we discuss the structures of known peptide-activated receptors and attempt to draw conclusions about this family as a whole.

1.3.1 The First Peptide-Activated GPCR Structure

The first crystal structure of a peptide-activated receptor was the CXCR4 receptor in 2010 [25]. The receptor was determined in the inactive state bound to both a small molecule antagonist and a peptidomimetic. The structure of this receptor was similar to what had previously been seen for aminergic [24, 26] and nucleotide receptors. However, an interesting difference was the presence of a β-hairpin in extracellular loop 2 (ECL2), a motif that has been present in all peptide-activated receptors reported since that time [27].

1.3.2 Crystallization of the First Peptide-Activated GPCR in Complex with a Peptide Ligand

Two more years passed before another peptide-activated receptor structure was determined. 2012 was a watershed year for this family with the structure determination of all four opioid receptor members (δOR [28], κOR [29], μOR [30], and NOP [31]), the PAR1 receptor [32], and the NTS1 receptor [33]. All of these receptors contained the β-hairpin motif in ECL2 presenting an open binding pocket for interaction with ligand. Importantly, the NTS1 receptor structure was the first determined structure of a peptide-activated receptor in complex with its endogenous peptide ligand. Unfortunately, this receptor contained extensive mutations that prevented signaling through peptide activation so questions were raised as to whether the binding mode of NT(8-13) was the native orientation. However, in the following years, a series of structures of NTS1R with varying mutations back to the wild-type that restored signaling found the endogenous peptide binding in the exact same manner suggesting that the binding mode was unaffected these mutations. Interestingly, the binding depth of NT(8-13) was not a deep as seen for the aminergic and nucleotide ligands suggesting that peptide ligands bind more superficially

and predominantly interact with the extracellular loops. As the extracellular loops of GPCRs are the most highly divergent region of this family, this greatly complicated that ability to generate hypotheses for the binding mode of peptide ligands as a whole.

1.3.3 Crystal Structures of Additional Peptide Receptors Reveal Diverse Binding Modes

Over the years, additional peptide-activated receptor structures were determined. These included additional chemokine receptors (CCR2 [34], CCR5 [35], CCR9 [36], and the viral US28 chemokine receptor [37]), both subtypes of the orexin [38, 39] and angiotensin [40, 41] receptors, the PAR2 receptor [42], the endothelin-B receptor [43], the neuropeptide Y type 1 receptor [44], and the C5a receptor [45]. The binding pockets of peptide-activated GPCRs while uniformly wide, correspondingly display a variety of hydrophobic and electrostatic conditions [27]. Of note, on a very small subset of these have been determined bound to their respective peptide ligands. These include the chemokine receptors US28, CCR5 and CXCR4 [37, 46, 47], the endothelin-B receptor [43], the apelin receptor [48], the μ opioid receptor [49], and the C5a receptor [50]. Interestingly, the binding modes of these ligands are more diverse than was previously expected. Peptide ligands can bind in extended manners to penetrate deep in the helical bundle via either their N- or C-terminus, they can bind both termini folded into the binding pocket, or in a horseshoe manner presenting a curved surface to the receptor. The ligands can bind deeply or only at the surface. The main theme connecting the binding mode of a peptide ligands is that they often contain multiple binding sites and bind over an extended surface of the receptor. This greatly complicates the ability to form general hypotheses and significantly drives the need for additional structure determination.

1.3.4 Reduced Flexibility of Endogenous Peptide Ligands is Crucial for Crystallographic Success

A significant challenge for interpretation of structures determined via crystallization of peptide-activated receptors in complex with their cognate peptide ligand is the inherent flexibility of the peptides. Small molecule antagonists and agonists will adopt a single conformation and often are fully encased in the receptor binding pocket. Peptide ligands may adopt a single conformation in the binding pocket but the remainder of the ligand can remain flexible. This is likely why NTS1R was crystallized with only residues 8-13 of the peptide as residues 1-7 are expected to extend above the receptor pocket and remain unconstrained [33, 51]. The peptide ligand of the apelin receptor, while full length, was modified to incorporate a lactam ring which greatly constrained the peptide flexibility [48]. Full length chemokine crystallization is possible as the portion of the chemokine that extends out of the binding pocket folds into a well-defined structural domain though the N-terminus of the receptor which is known to bind the chemokines has yet to be resolved in a crystal structure [37, 46, 47].

1.3.5 Peptide Ligands Affect the Conformation of the Extracellular Surface

An important consequence of the extended binding surface area of peptide ligands is that their presence affects not only the deep binding pocket but the extracellular loops as well. This was recently demonstrated by structures of the endothelin receptor [43]. This receptor was crystallized in the apo state and in complex with a peptide ligand. Interestingly, the ECL2 adopted a different conformation in the presence of the peptide ligand. This is expected to be the case for many peptide-activated receptor structures. In particular, the recent structure of the Y1R receptor in complex with a small ligand found

the N-terminus of the receptor lying over the binding pocket fully encasing the ligand [44]. Mutagenesis studies confirmed that this portion of the receptor had no effect on the binding properties of either small molecule or endogenous peptide but it is expected that the N-terminus has to be folded back to allow binding of the much larger NPY ligand. This was modeled and presented with the crystal structure with extensive use of orthogonal biophysical techniques including NMR, crosslinking mass spectrometry, and mutagenesis.

1.4 Structural Changes in Peptides Drive Binding at their Receptors

This theme of conformational change in peptides in their bound state is not unique to peptide-GPCR recognition. Studies of ubiquitin by X-ray crystallography bound to various substrates identified several unique conformations. However, NMR analysis revealed that all of these conformations existed at the same time in solution demonstrating that conformational selection drove the binding recognition event [52]. Peptide binding sites have been characterized to require unique conformations of peptide ligands in GPCRs [53], proteases [54], and other systems including antibodies and major histocompatibility complex [55]. To the best of our knowledge, there has not been a review of the conformational changes that peptide ligand must undergo from their unbound to bound states at GPCRs. These changes have relevance in future structure determination and peptidomimetic drug discovery. In the following section, we will highlight many examples of peptide structural studies with a particular focus on the conformational changes observed during the binding event.

1.4.1 Neurotensin

Neurotensin (NT) was initially characterized in the 1970s as a tridecapeptide [56]. Alanine-scanning, truncation studies, and comparison with neuromedin N, a homologous hexapeptide, identified the C-terminal six residues as the portion responsible for receptor activation known as NT(8-13) [57]. The structure of NT has been investigated by NMR for decades. Original NMR studies of the full-length NT in aqueous solution, methanol, and SDS (a membrane-mimic) found that under all conditions the peptide was unstructured [58]. Given the overall disorder, attention was turned to the C-terminal NT(8-13). Upon binding to the receptor, significant chemical shift perturbations were observed indicating a conformational change when bound [59]. This structural rearrangement was subsequently confirmed with determination of the structures of free, membrane-bound, and receptor-bound NT(8-13) with solid state NMR [60]. In this study, it was found that both the solution-state and membrane-bound states contained no defined structure while the receptor-bound peptide possessed an extended β-sheet conformation. Using a combination of solid-state NMR and molecular dynamics, the phi/psi angles were mapped over the unbound, membrane-bound, and receptor-bound states. These results confirmed that the conformation present in the receptor-bound state is distinct from the conformations found in the other states [61]. Studies of full-length NT in various membrane-mimicking environments found that while the N- and C-termini could adopt structured regions, the central region remained flexible [62, 63]. When full length NT was combined with the third extracellular loop of NTSR1, the extended conformation of the C-terminus was again found [51]. When the crystal structure of NTSR1 was determined with NT(8-13), the extended structure of this peptide fragment was very similar to what had been seen in the NMR studies [33]. Unfortunately, this initial crystal structure contained a number of

mutations that prevented activation of G-proteins. Although the mutations were outside of the ligand binding pocket, subsequent structures with reduced mutations that were capable of coupling to G-proteins found only minimal differences in the binding pose of NT(8-13) [64, 65]. Knowledge of this extended binding pose allowed for the design of constrained peptides that reinforced the need for this conformation in the bound state. The use of spirolactams between Pro^{10} and Tyr^{11} generates either linear or kinked peptides. While the linear peptide maintained WT activity at the receptor, kinked peptides lost binding affinity by at least 10,000-fold[66]. Additionally, cyclization is a strategy for increasing half-life of peptide in the body. While the end-to-end cyclization of NT(8-13) could penetrate the blood-brain barrier better than linear peptide, the activity of this peptide was significantly reduced at the receptor owing to the need for an extended conformation [67]. Later attempts at cyclization that maintained the overall linearity instead linked the side chains of Arg^{8} to Tyr^{11} while maintaining activity [68]. Further manipulations of peptidomimetics will need to ensure the proper geometry of this peptide as defined by the receptor binding pocket.

1.4.2 Apelin

The apelin peptides are a family of peptides all formed from the same prohormone but with successive N-terminal proteolytic processing. Apelin plays an important role in cardiovascular activity [69, 70] and obesity [71] and its target receptor, APJR, is involved in HIV viral fusion [70, 72]. Structure-activity relationships (SAR) analysis on the peptide identified a primary binding motif of the last five C-terminal residues with a second binding motif located 4 residues N-terminal of this motif [73-75]. This second motif which starts at the N-terminus of apelin-13 is required and further N-terminal truncations are not

active at the APJR. With the two binding motifs, all apelin peptides are equally potent, though evidence suggests the long N-terminus of apelin-36 contributes to longer activation at the receptor [69, 72]. CD studies of the peptide revealed that in solution the apelin peptides possessed no structured regions [73, 76]. Secondary structure could be induced either by lowering the temperature of the solution [76] or the addition of membrane mimetics [76]. The regions that became ordered under these conditions were the same regions that were previously identified in SAR studies as the binding motifs. This suggested that the binding motifs must become ordered in order to bind to the receptor. When the structure of APJR bound to an apelin mimetic was determined, this hypothesis was largely confirmed [48]. It was found that the apelin mimetic, a conformationally constrained peptide containing a lactam, adopted a conformation that allowed for an ordered presentation of these two binding motifs at distinct regions of the receptor. Interestingly, the lactam ring that was used to reduce the conformation of the peptide was introduced in the region between these two binding motifs thereby allowing for presentation of both motifs of the constrained ligand to the receptor. Mutagenesis and MD simulations of apelin-13 in the crystal structure revealed that native apelin peptide bind in a similar orientation as the crystalized ligand. While unlikely that a small molecule will be able to activate both binding sites of the APJR, further development of peptidomimetics with this constrained geometry will further the development of therapeutic drugs at this receptor.

1.4.3 Endothelin

The endothelin peptides are a family of three 21-amino acid long peptide containing two internal disulfide bonds. Endothelin is the most potent vasoconstrictor in human [77]

and is involved in blood pressure regulation and salt homeostasis [78]. The endothelin peptides bind two receptors, ET_A and ET_B, which share about 60% sequence similarity [79]. All three endothelin peptides bind with equal affinity at ET_B while ET-13 is less affine than the others at ET_A [80]. The two disulfide bonds stabilize a defined horseshoe structure from residues 1 through 15 of an N-terminal loop followed by an α-helix from residue 8 through the last disulfide as evidenced by multiple studies in NMR and X-ray crystallography [81-86]. The alpha helix is formed in solution independent of the cysteines but the orientation of the N-terminus is lost [87]. The C-terminus beyond residue 15 highly dynamic adopting helical structures [82], extended structures [84, 85], and in some cases so poorly resolved that a structure could not be assigned [81, 83]. The C-terminus, however, is critical for peptide activity and is expected to bind in an ordered pocket within the receptor [88]. When the cocrystal structure of ET-1 with ET_B was determined, the overall conformation of ET-1 remained largely unchanged from that in solution owing to the structural restraints of the two disulfide bonds [43]. However, the C-terminus of the ligand was bound within the receptor core in an extended manner and in close proximity to the ligand N-terminus. This orientation of the C-terminus with respect to the N-terminus is found in two of the ten ensemble structures of a snake venom toxin with high sequence similarity and identical disulfide linkage as ET-1 suggesting the peptide can sample this conformation albeit at a low population in solution [84]. Interestingly, the receptor in the bound state folds its ECL2 and N-terminus over the ligand explaining the extremely slow off-rates exhibited by these peptides *in vivo* [89, 90]. This structure clearly demonstrates conformational changes in both binding partners needed for full binding activity.

1.4.4 C5a

The complement system is a peptide-receptor system comprising two ligands (C3a and C5a) and three receptors (C3aR, C5aR1, and C5aR2, previously known as GPR77). The peptide ligands are 77 and 74 amino acids in length, respectively, and share 36% sequence identity. Both ligands contain 3 conserved disulfide bonds that play a role in defining the overall helical bundle fold that has been observed repeatedly by crystallography and NMR [91-101]. The consensus from all of these structures is that the peptides form three- or four-helix bundles depending on C-terminal truncation. While the full peptide is necessary for activation of the receptors via a two-step binding mechanism, the C-terminal segment is the activation segment that binds at the receptor core [102]. However, this C-terminal segment adopts a variety of conformations depending on the crystallized condition and lacks any secondary structure. One NMR study measures chemical shifts in the C-terminal backbone residues to find an α-helix that folded back onto the helix-bundle [101], an unlikely conformation in the active state as this peptide must be "presented" to the receptor for activation. As a result, C-terminal derivatives were studied as possible low-molecular weight activators of the receptors [103]. NMR structures of C-terminal peptide derivatives in solutions of varying hydrophobicity again found that this region was structurally sensitive to the chemical environment [104, 105]. These findings led to extensive SAR studies to define elements that are critical for receptor activation [106-108]. It was eventually found that cyclic derivatives could be potent activators that were structural insensitive to the chemical environment as measured by NMR and HD/X. These derivatives all contained a turn of central residues that were critical for receptor activation. NMR of a linear hexapeptide that capitalized on this SAR data found that this linear peptide also possessed a central turn despite cyclization suggest this was the active

conformation of the bound peptides [109]. Modeling of the C-terminus of C5a in a C5aR homology model also suggested that the endogenous linear peptide possess a dramatically different conformation in the bound state as compared to the solution state mimicking the conformations explored by the cyclic constrained analogs [110]. This proposed binding mode was in fact very similar to the conformation of the cyclic hexapeptide PMX53 that was eventually crystallized bound to C5aR [50]. The ligand formed a beta-hairpin to interact directly with ECL2 via backbone hydrogen bonding. It is now understood that the cyclization enforces the conformation of these backbone orientation such that while the endogenous ligand is linear, the predefined extended backbone can enhance this orientation necessary for binding. Additional modeling studies have found this extended conformation of C5a and a derivative peptide interaction with the extracellular loops [111, 112].

1.4.5 Ghrelin

The ghrelin receptor is involved in feeding control and its signaling is dysregulated in eating disorders [113]. Long before the receptor was identified, peptide and non-peptide mimetics were being developed to induce release of growth hormone [114]. The receptor was eventually identified in 1996 [115] while the endogenous ligand was discovered a few years later in a case of reverse pharmacology [116]. The identified ghrelin peptide was a 28 amino acid polypeptide with a lipid modification at position three. This is the only known lipid-modified peptide hormone and it has been found that this lipid modification is critical for receptor activation [116-119]. Importantly, the receptor displays high levels of constitutive activity [120, 121]. Therefore, the development of inverse agonists, and not neutral antagonists, are necessary for quieting this basal activity. Modification of the

substance P peptide has generated a variety of inverse agonists, all with a common structural motif of wFw (w representing D-Trp) that results in an L-shaped conformation with all hydrophobic side chains pointing away from the bend [120, 122-126]. However mutational analysis of the ghrelin receptor revealed that the binding regions for endogenous agonist ghrelin did not fully overlap with the binding regions of these inverse agonists [122] suggesting a need to understand ghrelin binding more fully to further develop peptide-mimetic derivatives with various activity profiles. Structure-function studies on ghrelin initially identified that the N-terminus of the peptide was critical for binding and activating the receptor via two main interactions: the positively charged amino head group and the hydrophobic octanoyl chain at Ser^3 [118, 119, 127]. As such, synthetic agonists for the receptor maintain positive charge and bulky hydrophobic moieties [123, 128, 129]. However, beyond these rules, little is known about the binding mode or conformation of ghrelin at its receptor. NMR and CD spectroscopy studies of the peptide in solution agreed that the peptide was highly disordered in the aqueous state [130]. Increasing the hydrophobicity of the solution either with organic solvents or detergents seemed to increase the helicity of the central portion of the peptide while the termini remained highly flexible [7, 131-133]. A recent study of the peptide bound to its receptor via NMR revealed that a helix is extended in the central peptide while the N-terminal binding portion converged to a well-defined extended structure (Bender et al, in press). Additionally, a much smaller number of models were present in the ensemble needed to explain the available NMR data suggesting a convergence on a single conformation in the receptor-bound state as compared to the membrane-bound state [7].This structure is likely to jump start novel drug discovery platforms by defining the conformation of the endogenous peptide in the bound state. Peptidomimetics that enforce this conformation should be able to maintain high potency at

the ghrelin receptor.

1.4.6 Gonadotropin Releasing Hormone

Gonadotropin releasing hormone (GnRH) is a decapeptide consisting of pyroGlu-His-Trp-Ser-Tyr-Gly-Leu-Arg-Pro-Gly-NH2. Evolutionary analysis reveals that the first four residues, the central Gly6 residue, and the last two residues are highly conserved [134]. This suggests a dual binding mode that requires both termini to come into close contact with the receptor. Extensive mutagenesis on both the peptide and receptor have resulted in the idea of an inverted horseshoe binding motif for receptor activation [135]. NMR studies of this peptide in solution failed to identify a single conformation [136-139], however peak sharpening increases in the presence of membranes suggested a reduction in conformational sampling [136]. Computer simulations also reveal a broad population of conformations that could exist though many low energy states contained a β-turn conformation in residues 5-7 [140]. The conformations of Gly^{6} adopt states that are inaccessible to L-amino acids but low energy for D-amino acids [141]. Given the flexibility of Gly^{6}, this is the proposed residue that induces the β-turn. Substitution of this residue with a D-amino acid enhances the likelihood of the β-hairpin turn thereby prestabilizing the conformation for receptor binding. In the native ligand, it is proposed that Arg^{8} helps to form the β-turn via an interaction with $Asp^{7.32}302$ [142]. The Gallus gallus homolgy of GnRH contains a Gln at position 8 which binds at a significantly reduced affinity to mammalian GnRHR. When studied in NMR, Gln^{8}-GnRH adopts an extended conformation as compared to wild-type GnRH which can sample β-turns [143]. Interestingly, a Gly^{6} substitution with D-Trp can overcome the loss of binding in Arg^{8} to Gln mutation [144]. This suggests that conformational constraint is more important than amino acid identity.

Medicinal chemists have taken advantage of this conformational change to create super-agonists. GnRH analogues including goserelin, napharelin, triptorelin, leuprorelin, buserelin, histrelin, and doslorelin that are used in the treatment of hormone-sensitive diseases such as breast and prostate cancer [145] often contain a D-amino acid substitution at position 6. NMR studies of napharelin find that unlike GnRH, this peptide readily adopts a β-turn conformation in aqueous solution [146]. Similar results were found in NMR analysis of leuprorelin [147]. Additionally, cyclization of the peptide has resulted in potent analogues [148, 149]. This all points to the need for GnRH to select a particular conformation from the many accessible in the unbound state in order to bind at the GnRH receptor.

1.4.7 Neuropeptide Y

The neuropeptide Y (NPY) system consists of three 36 amino acid peptides (NPY, PYY, and PP) and four receptors (Y1, Y2, Y4, Y5) with differing affinities of the various peptide/receptor combinations [22]. Initial studies to parse out the specific interactions of these peptides with the various receptors relied on alanine scanning and peptide truncation. It was revealed that the C-terminal six residues were the primary binding and activation epitope within the NPY peptides [150]. However, mutations in the central portion of the peptide could also affect receptor activation. Interestingly, the N-terminus proved to be critical for activation at the Y1 and Y4 receptors but could be dispensed with at the Y2 and Y5 receptors. Given this complexity, structural studies were pursued to understand the molecular basis of receptor recognition and subtype selectivity. An X-ray crystal structure of avian PP revealed a disordered N-terminus with an α-helix from residue 14-31 and a

disordered C-terminus [151]. Solution NMR studies revealed a dissimilar structure with the helix present through the C-terminal end of the peptide [152]. The one agreement between the two structures was that the helix was amphipathic leading to the theory that these peptides recognize their receptors by first binding to the membrane [153]. Structural characterization of NPY at membranes by NMR, CD, and EPR again found the N-terminus disordered but with an amphipathic membrane-binding helix from residues 14-36 [154, 155]. However, it was not until the peptide was structurally characterized in its Y2 receptor-bound state that it became clear that the C-terminus, though helical in its membrane-bound state, must unwind into an extended conformation for binding at the receptor [2]. This study provided a model of binding that identified a second binding site between the central helix and ECL2 in agreement with two-site binding modes observed with other peptide-GPCR interactions. The conformational change of the C-terminus was also observed in a study of NPY binding at the Y1 receptor [44]. However, in this study, photo-crosslinking revealed that the N-terminus of NPY was interacting with ECL2 instead of the central helix. This alteration of the second binding site interaction resulted in a distinct binding orientation of NPY at two if its four receptors. Further, studies will need to be pursued to contrast the binding mode of the remaining receptors to understand the basis of subtype selectivity.

1.4.8 Opioid Peptides

The opioid receptor family, comprising of δOR, μOR, κOR, and NOP, respond to a variety of endogenous peptides include endorphins, dynorphins, and enkephalins. These peptides contain a common N-terminal motif of YGGF followed by diverging residues. It

is suggested that the N-terminal motif is the activation sequence while the remaining residues confer receptor selectivity, the so called “message-address” paradigm [156]. Given this complex interaction pattern and the pharmacological relevance of the opioid system, several X-ray and NMR studies have looked at the structural features of these peptides over the last three decades [157-164]. What has been evidenced repeatedly is that there is a conformational heterogeneity within the population of these peptides in both the aqueous- and membrane- bound states. Only in 2015 was the peptide dynorphin B studied in the presence of the κOR [165]. Here it was found that the central portion of the peptide formed a well-defined α-helical turn while the N- and C-terminal residues were extended with no secondary structure. It was interesting that multiple conformations were found for the N-terminal motif in the bound state, however it was proposed that these conformations were representative of various inactive- and active-state populations. This is in contrast to molecular dynamic simulations ran on the DAMGO peptide bound in the μOR-G_i cryo-EM structure with bound synthetic peptide [49]. Here, the researchers found that the peptide was quite stable in its conformation over time within the binding pocket. At the present, it is unclear if this conformational stability is due to the alterations of the peptide backbone in this synthetic peptide derivative, stabilization due to activation state, or a difference between the binding pockets of μOR and κOR. The strategy of using conformationally stable peptides for crystallization is similar to the approach used in the apelin receptor crystal structure [166].

1.4.9 Secretin-Type Peptides

The structural basis of peptide binding to Class B secretin-like receptors has been reviewed before [167]. These peptides also exhibit two-site binding recognition with the C-

terminus first binding the N-terminal extracellular domain of the receptor followed by N-terminal insertion into the transmembrane helical bundle. Owing to the folded N-terminal extracellular domain, several crystal structures have been determined of these peptides bound to this soluble component. Repeatedly, these structures reveal a long α-helix of the bound peptides. This is in contrast to several studies of the peptides in solution which display a variety of conformations [168-173]. These NMR and CD studies often revealed that helicity of the peptides increased with increasing hydrophobicity of the solution, usually by increasing the concentration of TFE or methanol. It was expected that this alteration in hydrophobicity mimics the change in moving from solution to a bound state. The crystal structures and NMR studies of the bound peptides to the extracellular domain exhibit this helical nature of the peptide, however they often show that the N-termini are disordered. While this may have been interpreted as disorder due to lack of binding interaction, recent structures of full-length Class B receptors with bound agonist peptides confirm that the N-terminus remains in an extended conformation within the transmembrane binding pocket [174, 175]. This is in agreement with SAR studies of PACAP that demonstrate that the receptor is highly sensitive to alterations that affect secondary structure propensity of the N-terminal residues [176]. Here, as in the class A receptor peptides, there is a strong degree of conformational change between the solution state and bound state.

1.5 Implications for Future Studies

1.5.1 Peptides Need to be Characterized in their Bound State

Several peptide hormones have been examined to understand their structure via

NMR or CD in solution. These include motilin [177], prolactin-releasing peptide [178, 179], vasopressin [180], relaxin [181, 182], and somatostatin analogues [183-186]. In contrast, relatively few examples exist of peptide studied in both their solution and bound states. These include the peptides NPY, ghrelin, and bradykinin [2, 7, 23, 44, 154, 187]. A common theme that has arisen in all these studies and the ones mentioned above is that the conformations of the peptides in their unbound states are distinct from their bound state. This is perhaps unsurprising as the individual degrees of freedom in each amino acid are high in a peptide. In contrast, the receptor binding pocket imposes a stringent constraint on the conformation of these peptides. This theme of conformational sampling is analogous to the change in extracellular loop conformations in C5aR when bound to either a small molecule or peptide ligand [45, 50]. Given these differences, it is necessary to study these peptides in the presence of their cognate receptors to develop a full understanding of the molecular basis of peptide recognition.

Of note, the studies described in the above section rely on a variety of biophysical techniques for structural characterization. While X-ray crystallography and, in some cases, cryo-EM can reveal the conformations of peptides binding to GPCRs, this is currently rare. This is likely due to the inherent flexibility of peptide ligands, as described, which can hinder the crystallization process or identification of class averages. Complementary to these techniques, several studies have utilized NMR and CD to characterize the structure of the peptide ligands. CD provides readily accessible information to the overall secondary structure changes in varying environments. NMR can provide detailed information on a residue and atomic level about the structural properties of these peptides. Further, NMR reveals information on population dynamics that may provide information on the binding recognition process such as with STD-NMR. Additional techniques used in the above-

mentioned studies include EPR, H/DX-MS, crosslinking, and molecular modeling. Lastly, a significant technique used for decades in peptide ligand studies is the use of mutational analysis. Alanine scanning and backbone modification of peptides is analogous to tradition SAR studies of small molecule ligands. Future studies will likely need to combine multiple of these techniques to arrive at reliable understandings of these peptide-receptor complexes.

1.5.2 Mimetics of the Bound-State Conformations can Aid in Structure Determination and Drug Discovery

As evidenced by the apelin, μOR, and neurotensin crystal structures, conformational stabilization or truncation of flexible components within the peptide ligands can assist in the crystallization of these complexes. While NTS(8-13) and AMG3054 are minimal perturbation of the peptide structures of neurotensin and apelin, respectively, DAMGO represents are more dramatic change from the endogenous peptide ligand structure. Interpretation of these structures will need to be verified for the endogenous peptide ligands. SAR studies on peptide with no known crystal structures will be invaluable for understanding of the conformational constraints required for these peptides in the bound states. Future crystallization trials with these conformationally constrained peptide derivatives will increase the likelihood of a stable crystal with interpretable density at the ligand binding site. Simultaneously, as the conformational constraints of the ligand binding sites become better understood, development of more potent drug therapies may become possible. This has been evidenced clearly with the development of super-agonists for the gonadotropin-releasing hormone receptor. Addition of a D-amino acid enhanced the beta turn in the peptide that is needed for the bound state. It is suggested the pre-orientation of

the ligand conformation will reduce the entropic cost of ligand binding thereby increasing the affinity at the receptor. This has not yet proven true with stabilization attempts of neurotensin derivatives in which the best derivatives are still only on par with the endogenous peptide. This theme will need to be studied in future drug developments to see if this consistently holds.

1.6 Conclusions

Peptide-binding GPCRs represent nearly a quarter of the druggable human GPCR superfamily. Structural studies on either the peptide ligands or receptors are often pursued independently of one another. However, receptors influence the conformation of their peptide ligands, and peptide ligands can alter the conformation of the extracellular loops of the receptors. This suggests that future studies of complexed structures will reveal as-of-yet unknown structural motifs for a give receptor/ligand pairing. As additional structural studies reveal the specifics of peptide-receptor recognition, a general mechanism may be possible for this family of receptors. Currently, our understanding suggests that conformational selection is a prime driver of receptor recognition, and as a result each peptide/receptor pairing should be studied individually.

CHAPTER 2

Protocols for Molecular Modeling with Rosetta3 and RosettaScripts

2.1 Summary

This chapter presents an overview of the Rosetta modeling suite and specific protocols that are used repeatedly throughout my thesis. The introduction provides an overview of the Rosetta framework and subsequent sections outline methods that are used throughout the thesis. I have contributed to the development of several of these protocols and have also taught several of them at our semiannual Rosetta Workshop. This chapter comes from parts of the review paper "Protocols for Molecular Modelling with Rosetta3 and RosettaScripts" for which I am the first co-author [188].

2.2 Introduction

Obtaining atomic-detail accurate models for all proteins, natural and engineered, in all relevant functional states, alone and in complex with all relevant interaction partners by crystallography or NMR is impaired by the vast number of possible protein sequences and interactions. In some cases, it is complicated by experimental obstacles, and is often time- and cost-intensive. Additional difficulties arise when the dynamic properties of proteins and their interactions with other molecules are to be studied from crystallographic snapshots. Here, computational modeling of the structure and dynamics of proteins and interactions can complement experimental techniques. Such computational models:

- add atomic detail not present in low-resolution, limited experimental data,
- model states not tractable for experimental structure determination,
- simulate conformational flexibility and plasticity of states, and
- prioritize states for crystallization or study with other experimental techniques.

At the same time, prediction and design of protein structure *in silico* is a formidable task: the need to model thousands of atoms increase the sampling challenge of testing a large number of possible arrangements or conformations. The need to complete these calculations in a finite time creates the scoring challenge of developing an energy function that is rapid but still accurately distinguishes between biologically relevant, low free energy states and other alternative conformations.

The Rosetta software suite is a compilation of computational tools aimed at obtaining physically-relevant structural models of proteins and their interactions with other proteins, small molecules, RNA, and DNA. Rosetta has contributed to the advance of structural biology by tackling challenges in de novo protein design [189-191], comparative modeling [5, 192], protein design [3, 193-197], protein-protein docking [9, 198-200], and protein-small molecule docking [201-203]. Additionally, Rosetta can be applied to RNA/DNA structure prediction [204, 205], the incorporation of non-canonical amino acids [206, 207], and other difficult structural challenges such as membrane protein structure prediction [208] and modeling of symmetric proteins [209, 210].

Rosetta developers follow the hypothesis that a single, unified energy function should be able to accomplish all of these complex tasks; furthermore, the continuous optimization of this energy function to improve one structural problem will ultimately

improve performance for other modeling tasks. Important components of the energy function are statistically-derived, i.e. using protein models derived from high resolution crystallographic data in the Protein Data Bank (PDB) as a knowledge base [189, 193, 202, 208, 211-220]. For speed, the energy function is pair-wise decomposable and employs a distance cut-off. For many sampling tasks Rosetta employs a Monte-Carlo search steered by the Metropolis criterion (MCM) [212]. Rosetta is continually developed and rigorously tested by a consortium of international academic laboratories known as the RosettaCommons (www.rosettacommons.org). Herein we present a global review of generalized Rosetta protocols and applications, as well as descriptions of novel functionalities recently introduced [221-223].

Detailed tutorials and examples are included as Supporting Information. The tutorials herein supersede our previous tutorials put forward in “Practically useful: what the Rosetta protein modeling suite can do for you” [224].

2.1.1 Making Rosetta Accessible

Rosetta is extremely powerful for many applications in structural biology, but for many years it was limited by the fact that users needed an extensive background in C++ and the Unix environment to be able to construct new protocols. An on-going effort by many groups has been taken to eliminate these boundaries, allowing greater flexibility and ease-of-use for the novice and intermediate user. These updates include customizable protocols using XML or Python. The updates using XML (RosettaScripts) [225] or Python (PyRosetta) [226] allow users to customize protocols without learning C++, by combining pre-written Rosetta objects and defining their behavior without having to write and

recompile new C++ code. In addition, the Rosetta community now offers multiple web interfaces for application specific tasks.

Other tools have been added, not to run Rosetta, but to improve users' experience, such as graphical user interfaces (GUIs) to visualize Rosetta operations and generate input files [227], and PyMOL integration for real-time molecular visualization [228]. These tools offer users intuitive control over structural modeling without sacrificing flexibility and power.

RosettaScripts. RosettaScripts is an XML-like language for specifying modeling protocols through the Rosetta framework [225]. It allows users to define a set of Rosetta objects and execute them in a defined order to develop full protocols. Rosetta objects in RosettaScripts fall under four main categories: Movers, which are objects that modify a structure in some way; Filters, which evaluate properties of a structure; TaskOperations, which control the degrees of freedom of Rosetta's side chain placement routines; and ScoreFunctions, which evaluate the energy of a structure. By combining these four elements users are able to leverage many different sampling and scoring algorithms, with fine control over sampling degrees of freedom and protocol flow. All objects defined under these categories are customizable, which is a distinct advantage of RosettaScripts over conventional command line applications. For example, a user can define multiple score functions in the ScoreFunctions block to be used in different sections of a protocol and then combine several protocols into a single XML protocol (i.e. protein-protein docking and design). This flexibility has made a number of scientific advances possible, such as *de novo* design of an influenza binder [197], protein-protein docking based on hybrid structural methods [229], and HIV vaccine design [230].

PyRosetta. Due to the popularity of Python as a programming language in the computational biology community, a Python-based implementation of Rosetta was developed, termed PyRosetta [226]. PyRosetta consists of Python bindings for the major functions and objects of Rosetta, allowing all of these objects to be run from a Python environment. One advantage is the ability to combine Rosetta protocols with other popular structural biology software, such as PyMOL [231] and BioPython [232]. PyRosetta includes access to the same set of Rosetta objects for sampling and scoring that are described above for RosettaScripts, as well as many others. Unlike RosettaScripts, PyRosetta can be run either in script mode or interactive mode. Interactive mode allows the user to inspect their objects in real-time while prototyping a new protocol [226]. Notably, PyRosetta is available for Windows in addition to Linux and Mac OSX, expanding the availability of Rosetta to researchers who use a Windows environment.

Web Interfaces. We are aware of nine webservers that have been created to allow non-experts to make use of Rosetta's functionality (

Table 2.1). These webservers allow Rosetta to be used with almost no learning curve, making the boundary to entry even lower than the above-mentioned scripting protocols. In particular, ROSIE (the Rosetta Online Server that Includes Everyone) [233] has been set up to easily provide a web interface to new Rosetta protocols.

Table 2.1 Publicly Accessible Web Servers Running Rosetta

Server Address	Protocols Offered
rosie.rosettacommons.org	Many, including small molecule docking, protein design, RNA design, etc.[233]
robetta.bakerlab.org	Structure prediction[234]
graylab.jhu.edu/docking/rosetta	Protein-protein docking[235]
rosettadesign.med.unc.edu	Protein design[236]
flexpepdock.furmanlab.cs.huji.ac.il	Flexible peptide docking[237]
kortemmelab.ucsf.edu/backrub	Backbone remodeling & design[238]
funhunt.furmanlab.cs.huji.ac.il	Classification protein-protein complex interactions[239]
csrosetta.bmrb.wisc.edu	Structure prediction based on chemical shift data
rosettadiagrams.org	Set up protocols through visual diagrams

Other Tools. Since the publication of RosettaScripts and PyRosetta, new tools have been developed to make running a Rosetta protocol even more intuitive. An interface to PyMOL was developed by Baugh et al, which allows users to visualize their molecules being manipulated by Rosetta as a protocol is being run [228]. While the viewer was originally developed for use with PyRosetta it has since been extended for RosettaScripts. This visualization tool is especially useful for new users with experience in structural biology but new to computation.

In addition, several GUIs for Rosetta have been developed to eliminate the need to run Rosetta exclusively through the Unix command line [227]. The PyRosetta Toolkit was

developed to serve as a GUI for running PyRosetta, with options menus to guide the user through the relevant Rosetta options that are needed for a protocol [240]. InteractiveRosetta is a GUI for running Rosetta protocols, with an integrated molecular visualization window and user-friendly controls for implementing common Rosetta protocols [227]. Through these GUIs users can generate input files for Rosetta protocols using a point and click interface while also running protocols seamlessly in the same window.

2.1.2 Sampling and Scoring in Rosetta

Rosetta Sampling. While the approaches used by different protocols vary, in general Rosetta utilizes a Monte Carlo Metropolis sampling algorithm to quickly and efficiently determine the quality of structural trajectories. Rosetta further differentiates between sampling backbone and side chain conformations within two separate refinement tasks. In addition, backbone sampling can be performed on a global or local scale. Large scale backbone sampling utilizes three-mer and nine-mer fragments derived from the Protein Data Bank (PDB), while local refinements of the backbone optimize φ and ψ angles without disturbing the global fold. Side chain sampling also utilizes information derived from the PDB to create a "rotamer" library of observed conformations in order to reduce the conformational search space. For a more detailed discussion of Rosetta sampling please refer to Rohl *et al.* [212].

Rosetta Scoring. The Rosetta score, or energy function, is a linear, weighted sum of terms combining knowledge- and physics-based potentials gathered from protein structural features within the PDB. The score function is used during Rosetta modeling to evaluate Monte Carlo sampling and for scoring the final output pose. With the

implementation of Rosetta3, the score function is treated as a separate entity such that it can be repeatedly called and rapidly processed independent of the protocol at hand [223]. Additionally, score terms are grouped into a hierarchy based on potentials related to one entity (i.e. χ angle probability), two interacting entities (i.e. hydrogen bonding potential), and terms that require the analysis of the entire model (i.e. radius of gyration).

Low vs. High Resolution Scoring. In low resolution scoring mode, the side chain of each residue is represented as a super atom, or 'centroid', at the Cβ position. This greatly reduces the degrees of freedom that must be sampled during low resolution backbone movement while preserving chemical and structural features of a given residue. Low resolution sampling involves replacement of the backbone conformation with peptide fragments of three and nine amino acids in length that are derived from the PDB. Peptide fragments are generated based on the primary sequence of the protein and influence secondary structure in these regions. Centroid mode scoring and sampling is used during initial stages of protein modeling where exhaustive searches of conformational space are performed such as *de novo* protein folding, loop building, and rigid-body protein-protein docking [189, 198, 212, 220]. Common score terms used in centroid mode are listed in .

High-resolution scoring, or 'full atom' mode, allows for full representation of all atoms of each side chain. In full atom mode, conformational sampling relies on replacing side chains with rotamers found in the PDB during a Metropolis Monte Carlo simulated annealing simulation to find the global minimum [213]. Full atom scoring was originally developed for protein design but has seen several improvements throughout Rosetta's history to the current *talaris2014* score function [193, 214-216, 221].

Table 2.2 Standard Rosetta Score Function Terms

Score Term	Definition
Low resolution scoring terms	
env	Hydrophobicity term for each amino acid
vdw	Steric repulsion between two residues
pair	Probability of two residues interacting
rg	Radius of gyration
cbeta	Solvation term based on number of surrounding residues
hs_pair, ss_pair, sheet	Secondary structure terms
High resolution scoring terms (talaris2014)	
fa_atr, fa_rep, fa_intra_rep	Decomposed 6-12 Lennard-Jones potential
fa_sol	EEF1 solvation term
pro_close	Proline ring closure energy
omega	Omega backbone dihedral potential
dslf_fa13	Updated disulfide geometry potential
rama	Potential of φ/ψangles for each amino acid
p_aa_pp	Probability of an amino acid given a set of phi/psi angles
fa_dun	Rotamer likelihood
hbond_sr_bb, hbond_lr_bb, hbond_bb_sc, hbond_sc	Combined covalent-electrostatic hydrogen bond potentials for alpha helices, beta sheets, side chain-backbone, and side chain-side chain interactions, respectively
yhh_planarity	Tyrosine hydroxyl out-of-plane penalty
fa_elec	Coulombic electrostatic potential between two residues with a distance-dependent dielectric (deprecates fa_pair)

Score Function Optimization. The score function is a linear weighted sum of energy terms, therefore the weights can be parameterized to generate meaningful scores for predicted models. These are often fitted against benchmark sets of modeling challenges to guide prediction of native structures. An algorithm 'optE' was developed to streamline this weighting term optimization [216]. This algorithm excels at setting reference weights for amino acids which is important for design challenges. The *talaris2014* score function was analyzed by optE and the reference weights were set to enhance sequence recovery (~40%).

Like previous iterations of the full atom score function, *talaris2014* sums separate physics- and knowledge-based potentials. However, it was found that combining physics- and knowledge-based information in a given score term led to improved Lennard-Jones and hydrogen bonding score terms [215]. The combined covalent-electrostatic hydrogen bonding terms were further updated with improved geometry and a parameterized sp^2 hybridized acceptor [221]. Scoring potentials of knowledge-based score terms were smoothened with the use of bicubic-spline interpolation [216]. An updated rotamer library was included with an adaptive kernel formulation, which allows for smoother potentials of Ramachandran-based score terms [217]. Ideal atomic coordinates for amino acids, the geometry of disulfide bonds, and the hydroxyl sampling of serine and threonine residues were improved as well as expanded. The free energy of solvation (LK_DGFREE) were updated to improve the EEF1 solvation energy potential of buried residues. Lastly, a new term was introduced that describes the Coulombic electrostatic potential between two residues with a distance-dependent dielectric (fa_elec) and replaces a previous statistics-based potential (fa_pair) [216]. Further refinements were made to reduce the influence of

the hydrogen bonding terms. This resulted in improved sequence recovery, rotamer recovery, and model discrimination [221]. As of writing, these updates culminated in the *talaris2014* score function, which is the default for current versions of Rosetta. All *talaris2014* score terms are listed in .

Continual optimization of the Rosetta score function means that the default score function varies with Rosetta version: *score12* for versions prior to Rosetta 3.5, *talaris2013* for weekly releases until 2016.10, and *talaris2014* for Rosetta 3.6 and weekly releases since 2016.11. Further score function refinement is ongoing, and it is likely Rosetta releases in the not too distant future will have as different default score function. Additionally, while Rosetta strives to have a single all-atom score function to encompass all modeling tasks, several application specific scoring potentials have been developed to include new score terms and optimized score term weights. These include, but are not limited to, modified score functions for small molecule docking [202], protein-protein docking [198, 218], and membrane protein modeling [208, 219], as well as specialized score functions for low-resolution sampling stages.

Clustering. Many protocols have adopted a clustering step as a means of narrowing the selection of candidate models. If several low-energy solutions are found, models are grouped based on the pairwise root-mean-square deviation (RMSD) between models. Clusters are then ranked by size with the assumption that native-like structures are found in the widest, deepest free-energy well [241]. This is particularly useful when applied in the context of *de novo* modeling, where the structure of the target sequence is unknown and the native conformation is thought to have many slightly less stable, structurally similar conformations.

Limitations and Caveats. Ongoing improvements made by the Rosetta community have led to increasingly accurate modeling protocols; however, there are still several hurdles that must be overcome for Rosetta to accurately obtain native-like models. First, Rosetta sampling is stochastic in nature. Therefore, not every modeling trajectory will sample a regional minimum on the score function. Second, the score function is heuristic and abbreviated for speed. It fails to fully recapitulate the fundamental forces. Therefore, minima of the energy function are not guaranteed to describe biologically relevant states. Third, even with its rapid score function, Rosetta is unable to exhaustively sample all possible structural space due to computational time restraints. Fourth, many Rosetta protocols are optimized for local resampling and require a starting model, which may not exist for some systems.

Evaluating Interfaces. Some biological applications of Rosetta focus on improving, creating, or otherwise altering a well-defined protein-protein, protein-small molecule, or protein-DNA interfaces. These protocols typically inhabit a much smaller search space, and in some cases rely solely on rigid-body optimization to generate a desired interaction [242]. In these instances, a series of specific interactions is evaluated, and the widely used "score vs. RMSD" plot is repurposed to look at small changes at the interface; here, plotting the "interface score" against the "interface RMSD" prevents small, meaningful changes from being lost in the larger fluctuations when scoring the entire model or computing the RMSD over all atoms. Additionally, analytical tools like the Interface Analyzer provide a series of useful calculations that include binding energy [243], shape complementarity, the number of buried, unsatisfied hydrogen bonds, and the solvent accessible area buried at the interface, can be used in conjunction with RosettaHoles [244]

to generate a packing statistic score for the interface.

2.2 De Novo Structure Prediction

De novo protein structure prediction is one of the greatest remaining challenges in computational structural biology. This process models the tertiary structure of a protein from its primary amino acid sequence. Importantly, *de novo* modeling differs from template-based or comparative protein modeling in that structural predictions are not based upon a known homologous structure. To address the challenge of predicting a protein's structure *de novo*, Rosetta uses short peptide "fragments" to assemble a complete protein structure.

The Rosetta *de novo* protein-folding algorithm continues to follow the steps described in our previous review [224]. Briefly, short peptide fragments of known proteins are obtained from structures deposited in the PDB and are inserted into an extended-chain protein following a Monte Carlo strategy [189]. In that sense, Rosetta *de novo* protein-folding is not truly *de novo*; it combines a very large number of small templates. The hypothesis is that while not every protein fold is yet represented in the PDB, the conformation of small peptides is exhaustively sampled. These peptide fragments are used to alter the backbone conformation of the extended-chain protein, folding it toward a low energy tertiary structure. The process is repeated to create an ensemble of models. Finally, these low-resolution models can be filtered based on pass/fail criteria provided by the user, clustered, and an energy minimization step can be applied to refine an all-atom model with the high-resolution energy function.

2.2.1 Generating Peptide Fragments.

De novo protein folding relies on the assembly of short peptide fragments, usually generated as a pre-processing step. First, the primary protein sequence is used to generate secondary structure predictions. Next, the sequence, secondary structure predictions and NMR data (if available) are used to pick candidate three- and nine-amino acid fragments from the PDB. Finally, these candidate fragments are scored and the best N fragments are written to a fragment library file. The ROBETTA webserver is available for non-commercial use, and allows users to generate fragment libraries using a simple interface [234]. Additionally, Gront *et al.* have developed the FragmentPicker that provides users with total control over the fragment picking protocol [245].

2.2.2 TopologyBroker.

The TopologyBroker [246], a tool that allows for more complex simulations, is an improvement added to Rosetta since our last review. The conformational space searched during a Rosetta *de novo* modeling simulation is vast, and successful searches often integrate prior knowledge with sampling. In *de novo* protein folding, this prior knowledge may be the form of β-strand pairing constraints or the formation of a rigid chunk of the target fold based on a structurally homologous domain. Previously, protocol developers were restricted to a sequential sampling approach in which Rosetta could readily violate one set of these constraints while sampling to satisfy the other. The TopologyBroker was developed to create a consensus sampling approach that satisfies all of the requested constraints without requiring additional code development for each unique system; instead, the Broker provides an Application Program Interface (API) that allows for plug-and-play

applications to generate complex sampling strategies.

2.2.3 Benchmarking

The *de novo* modeling capabilities of the object-oriented Rosetta software suite ("Rosetta3") were assessed in the CASP8 (Critical Assessment of protein Structure Prediction) experiment [190]. For 13 targets in the assessment, no homologous templates were identified and Rosetta's *de novo* modeling protocol was used to predict the structure of these targets. Following the observation that Rosetta *de novo* structural predictions are sometimes improved by using non-standard fragment sizes, a range of fragment lengths were used when modeling the CASP8 targets. Longer fragment lengths were found to improve modeling of α-helical proteins, while shorter fragment lengths improved modeling of mainly β-strand proteins.

2.2.4 Limitations of *De Novo*

Since *de novo* structure prediction is such a powerful tool and yet such a computational challenge, it is critically important to understand the limitations of the algorithm. Rosetta performs well at folding small, globular, soluble proteins as well as small, simple membrane proteins containing 80-100 residues. However, large and complex proteins present additional difficulties that are not easily overcome by *de novo* techniques alone. Instead, users must incorporate other biochemical information in order to obtain native-like models. Ongoing work shows that the incorporation of residue-residue co-evolution information can significantly improve the prediction accuracy during *de novo* modeling trails [191]. Other techniques such as homology modeling and using

experimental constraints are discussed below.

Furthermore, since *de novo* structural prediction will sample many potential protein folds, it is necessary to generate large numbers of models (>10,000) in order to adequately sample the conformational space. Extensive computational resources are needed to generate this number of models, and the use of distributed computational methods (such as computational clusters) is recommended. An example tutorial for the *de novo* prediction of a protein structure with Rosetta is included in the Supporting Information.

2.3 Comparative Modeling

Comparative modeling differs from *de novo* methods in that it utilizes a known protein structure as the starting scaffold or template for structural prediction. If the template structure is a homologous protein, one speaks often of 'homology modeling'. Comparative modeling is a useful strategy for predicting protein structure and function when experimental methods fail or would be too resource-intensive to employ. It increases the probability of obtaining realistic conformational predictions, especially when the target, or desired protein, is greater than 150 amino acids in length and/or adopts a complex tertiary fold. However, it requires that a related, often homologous, structure has been determined experimentally. This is termed the template. Ideally the sequence identity between the target and the template is above 30%, although proteins with lower sequence identity may still be used for comparative modeling when their tertiary fold is conserved. The latter case will be examined within the tutorial provided with the Supporting Information.

Over the past several years, comparative modeling in Rosetta has incorporated

many improvements, specifically the use of multiple templates and a specific low-resolution scoring functions [5]. Previously published protocols of comparative modeling with Rosetta suggested using multiple templates to obtain diversity and flexibility [192]. However, models were built on individual templates. The new RosettaCM protocol allows for integration of multiple templates with *de novo* fragments into a single structural model of the protein [5]. Hence, this multi-template, multi-staged protocol samples a broader structural landscape and can select well-scoring sub-templates for different regions of the protein to be modeled.

A highly detailed description of RosettaCM design, sampling and scoring has previously been published [5]. Users are encouraged to refer to this manuscript for a comprehensive assessment of RosettaCM applications, considerations, and caveats. Herein we will briefly describe features of RosettaCM as they apply to the protocol presented.

2.3.1 Starting Templates.

Before utilizing RosettaCM, starting templates must be identified through remote homolog detection methods such as PSIBLAST [247]. When homologs are not found using sequence-based methods, 3-D fold recognition software may be used to obtain suitable templates. As with other modeling software, RosettaCM performance improves with higher sequence similarity and identity.

2.3.2 Three Stages of Multi-Template Comparative Modeling.

Multi-template RosettaCM is a three-staged process in which the best scoring

model from each stage is utilized as the input for the following (Figure 2.1). The output of stage one is a full-length, assembled model that is generally correct in topology. However, segment boundaries where templates are mended can be sub-optimal in geometry and energetically frustrated. To resolve these energetic frustrations and to explore the conformational space around this starting model, stage two of RosettaCM iteratively improves local environments through a series of fragment insertions, side chain rotamer sampling, and gradient-based energy minimization of the entire structure using a RosettaCM-specific low-resolution energy function. The best model from this cycle is then moved to stage three for a final round of all-atom refinement that improves side chain geometries, backbone conformations, and packing density before converging on a final output model.

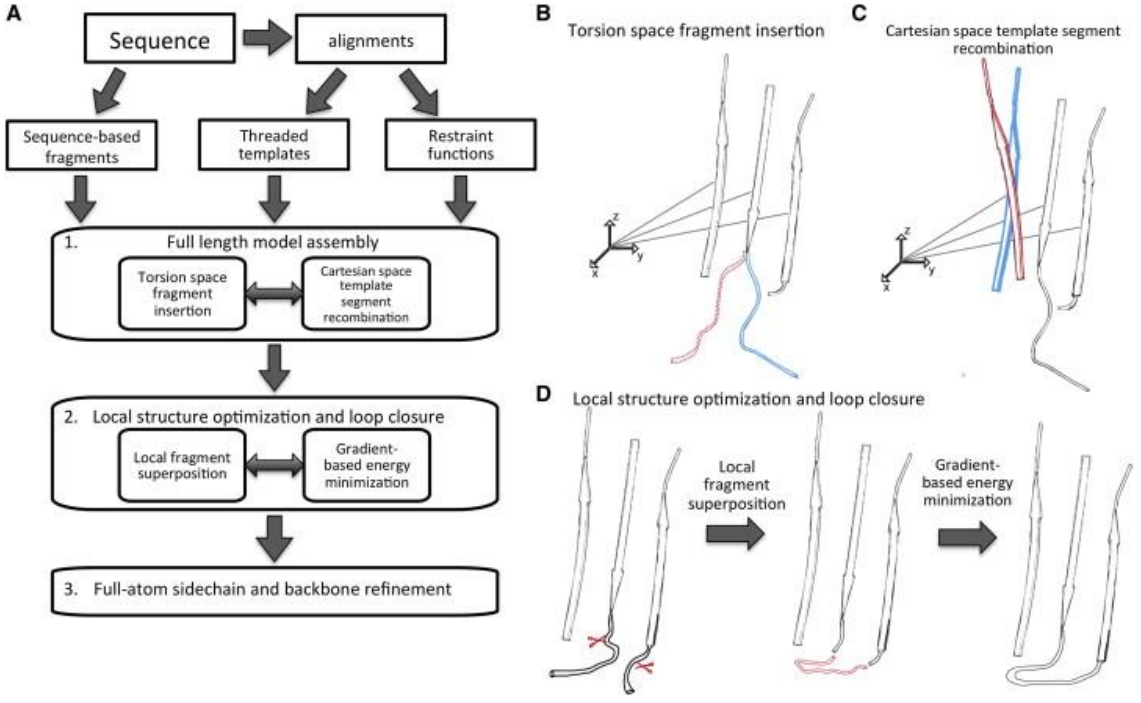


Figure 2.1: Multi-template comparative modeling with Rosetta. (A) General workflow of the RosettaCM protocol. (B) Fragment insertion (blue - before insertion, red - after insertion). (C) Recombination of template segments. (D) Fragment insertion and minimization for loop closure. Reprinted with permission from Song *et al*.[5].

2.3.4 Modeling Loops.

In previous Rosetta comparative modeling protocols, a user-defined, “loop” closure step was required to remove chain breaks, reconcile long unstructured coils, or rebuild regions of low sequence similarity (all of which are defined as “loops” within the Rosetta framework). Two different algorithms are available: Cyclic Coordinate Descent (CCD) and Kinematic Loop Closure (KIC). Briefly, CCD quickly closes roughly 99% of loops utilizing a robotics-inspired iterative approach to manipulate dihedral angles of three residue backbone atoms between user-specified C-terminal and N-terminal anchor points. The second loop-building algorithm, KIC, explicitly determines all possible combinations of torsion angles within the defined segment using polynomial resultants [248]. While being slower than CCD, KIC determines more accurate loop structures, provided the anchor points are optimally set. Both algorithms within Rosetta can be used in conjunction with fragments derived from the PDB to build regions of missing electron density, poor homology, or backbone gaps.

Unlike the single template loop building application, comparative modeling with multiple templates closes chain breaks and rebuilds loops internally during stage two. *De novo* fragment insertions are encouraged in regions of weak backbone geometry while template-based fragment insertions anneal chain-breaks and regions of low electron density. Additional smoothing occurs with the RosettaCM-specific scoring function. This internal step removes the need for additional loop closures by the user. However, it is encouraged for the user to critically examine all output models to validate structural accuracy.

2.4 Protein-Protein Docking

Determining the optimal binding orientation and interface of two or more protein binding partners has many biological and pharmaceutical applications; yet determining the structure of protein-protein complexes by biochemical techniques is slow and laborious. RosettaDock is a useful tool for computationally predicting protein-protein interactions by employing an algorithm that simulates a biophysical encounter of two or more binding partners and optimizes the conformation of the bound state. The RosettaDock algorithm includes a multi-scale, Monte Carlo-based docking algorithm that begins with a centroid-mode stage to identify docking poses, followed by an all-atom refinement stage to optimize rigid-body position and side-chain conformations [198]. An example protein-protein docking tutorial with Rosetta is provided in the Supporting Information.

2.4.1 Global vs. Local Docking.

The initial pose for docking is determined by either global docking or by local perturbation. Global docking randomly orients one of two binding partners in relation to the other binding partner to determine an initial binding interface. This is useful when there is no biological or structural evidence to suggest a starting pose. Local perturbation allows the user to define a general starting pose for the binding partners when prior experimental knowledge exists; this initial placement greatly decreases the conformational search space and improves the sampling density close to the starting pose.

2.4.2 Low Resolution vs. High Resolution Docking

The full RosettaDock algorithm begins with a low-resolution docking mode. The first step involves rigid body movements of binding partners which rotate and translate in relation to one another [199]. The score function is used to achieve a threshold acceptance rate of rigid-body moves [198]. A high-resolution docking mode follows in which the lowest energy structures and/or largest clusters assessed from the centroid-mode stage is selected for high-resolution refinement. Centroid pseudo-atoms are replaced with all-atom side chains in their initial unbound conformations followed by additional fine-grained rigid body docking.

2.4.3 Improvements to RosettaDock

The addition of RosettaScripts and PyRosetta to Rosetta now gives users the flexibility to modularize the centroid-mode and all-atom mode of RosettaDock to suit case-specific applications. This was done by splitting RosettaDock into three major classes: DockingProtocol, DockingLowRes and DockingHighRes [199]. The increase in flexibility only showed a marginal increase in successful predictions, however, it is particularly adept at predicting antibody-antigen complexes [199]. However, the modularization of RosettaDock has enabled developers to also incorporate additional features within their docking protocols, including additional parameters for non-protein moieties and protonation states [243, 249, 250], flexible peptide chain docking using FlexPepDock [200], and *de-novo* peptide docking [9].

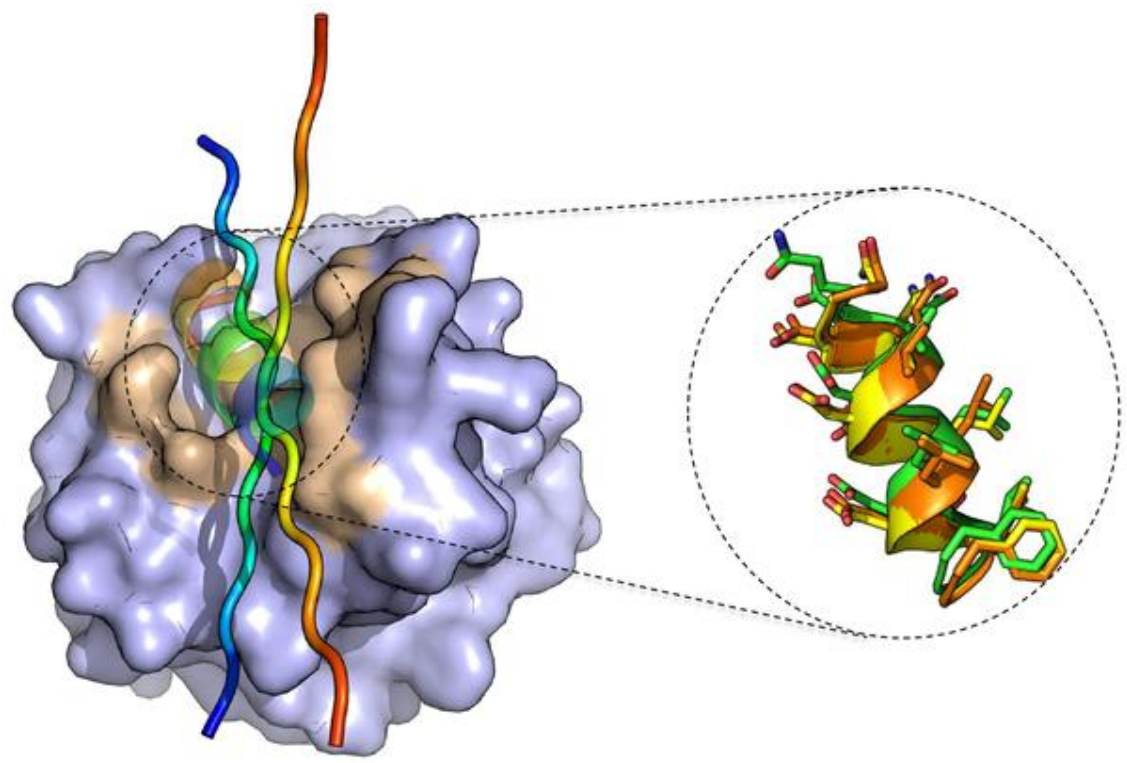


Figure 2.2 Protein-peptide interface prediction using FlexPepDock ab-initio. Structure prediction of the Che-Z derived peptide bound to CheY (PDB-ID 2FMF) from two opposite starting orientations converges onto the same final conformation resembling the structure of the native peptide. (Left) General view of the CheY receptor (in gray; interface residues colored in light brown), the two initial, extended peptide conformations (rainbow cartoons), and the final helical peptide conformation (rainbow, transparent cartoon). (Right) Detailed atomic view of the top FlexPepDock ab-initio predictions from two simulations (in yellow and orange) and the native peptide conformation (in green). Reprinted from Raveh *et al.*[9].

2.4.4 FlexPepDock

The FlexPepDock *de novo* docking algorithm is similar to the RosettaDock algorithm in that it begins with sampling rigid body moves from the initial protein-peptide complex. However, this step also includes iterative peptide fragment insertions and random moves of the peptide backbone using decreasing simulated temperature weights. Next, the low-resolution model is improved using an all-atom refinement stage by peptide side-chain placement optimization using a Monte Carlo search of ‘small’ and ‘shear’ moves described by Rohl *et al*. [189]. Each round of refinement also includes a decreasing repulsive van der

Waals weight term and an increasing attractive van der Waals term to allow greater degree of perturbations within the binding pocket without causing the peptide and protein to separate during energy minimization. The FlexPepDock *de novo* benchmark demonstrated that the protocol produces near-native models with 86% accuracy (Figure 2.2) [9].

2.5 Protein-Small Molecule Docking

Protein-small molecule docking aims to capture the binding interactions between a protein and a small molecule. This requires a combination of recapitulating the binding pose and quantifying the interaction strength. The RosettaLigand protein-small molecule docking protocol is designed to consider both protein and small molecule flexibility [201, 202]. It uses a two-phase docking approach similar to Rosetta protein-protein docking: a low-resolution phase of rapid sampling based on shape complementarity is followed by a high-resolution phase of Monte Carlo minimization of side-chain rotamers and small molecule conformers. The models undergo a final gradient minimization of the protein and molecule torsion degrees of freedom before they are output along with an interface score as a proxy for binding free energy. The "Small Molecule Docking" tutorial included in the Supporting Information demonstrates this optimized protocol.

2.5.1 Improvements to RosettaLigand

In contrast to the previously published RosettaLigand protocol [224, 251], this tutorial replaces independent Translation/Rotation low resolution sampling steps with the new Transform algorithm [203]. The Transform algorithm couples translational, rotational, and conformational sampling into a single Monte Carlo process. In a benchmark case the

Transform algorithm demonstrated a 10-15% improvement in docking success rate and an effective 30-fold speed increase over the classical methods [203]. The improved search time enables the use of RosettaLigand for screening medium-sized small molecule libraries, protocols for which are found in the supplementary information. For screening work with much larger libraries, Rosetta's Docking Approach using Ray-Casting (DARC) is a GPU-accelerated method demonstrated to be successful for protein-protein interface small molecules [252].

2.5.2 Customizable Small Molecule Docking Protocols

The RosettaLigand protocol can now be customized through the RosettaScripts XML interface, allowing for greater flexibility of use [251]. Additional features now include docking with explicit interface water molecules, which demonstrated 56% recovery for failed docking cases across a CSAR (Community Structure-Activity Resource) benchmark of 341 diverse structures [253]. Design of interfaces can now be incorporated into a single step docking and design of a protein-small molecule binding pockets [254]. These RosettaScripts based protocols have also been used to predict absolute binding energies for HIV-1 protease/inhibitor complexes with R=0.71 [255].

Research questions often focus on small molecules binding to a target protein without an experimentally determined structure. Such cases require first building models of the receptor using *de novo* Rosetta, RosettaCM or similar protein modeling protocols. When docking small molecules into protein models, Kaufmann and Meiler observed a native-like binding pose among the top ten scoring comparative models for 21 out of 30 test cases [256]. Furthermore, docking results were significantly better in cases utilizing

protein templates containing a small molecule of similar chemotype compared to templates with dissimilar small molecules or in the apo state. A full Rosetta protocol linking comparative modeling and small molecule docking is available in Combs *et al*. [192].

2.5.3 Small Molecule Docking in Membrane Proteins

Due to their biological importance and the challenges of experimentally determining their structures, membrane proteins are particularly attractive targets for the comparative model docking strategy. While the comparative modeling portion may be handled in a membrane environment, to date, Rosetta handles small molecule docking in a soluble environment. Nguyen *et al.* demonstrated the applicability of the soluble simplification for G Protein Coupled Receptors (GPCRs) [257]. RosettaLigand sampled near-native poses when docking small molecules into comparative models of GPCRs, but it remains challenging to select correct small molecule poses by Rosetta score alone (Figure 2.3). The use of high sequence identity templates, knowledge-based binding pocket filters, and experimental contacts are recommended methods to improve accuracy. Additional algorithm development and benchmarking is being pursued to fully integrate RosettaLigand with the RosettaMembrane framework [208].

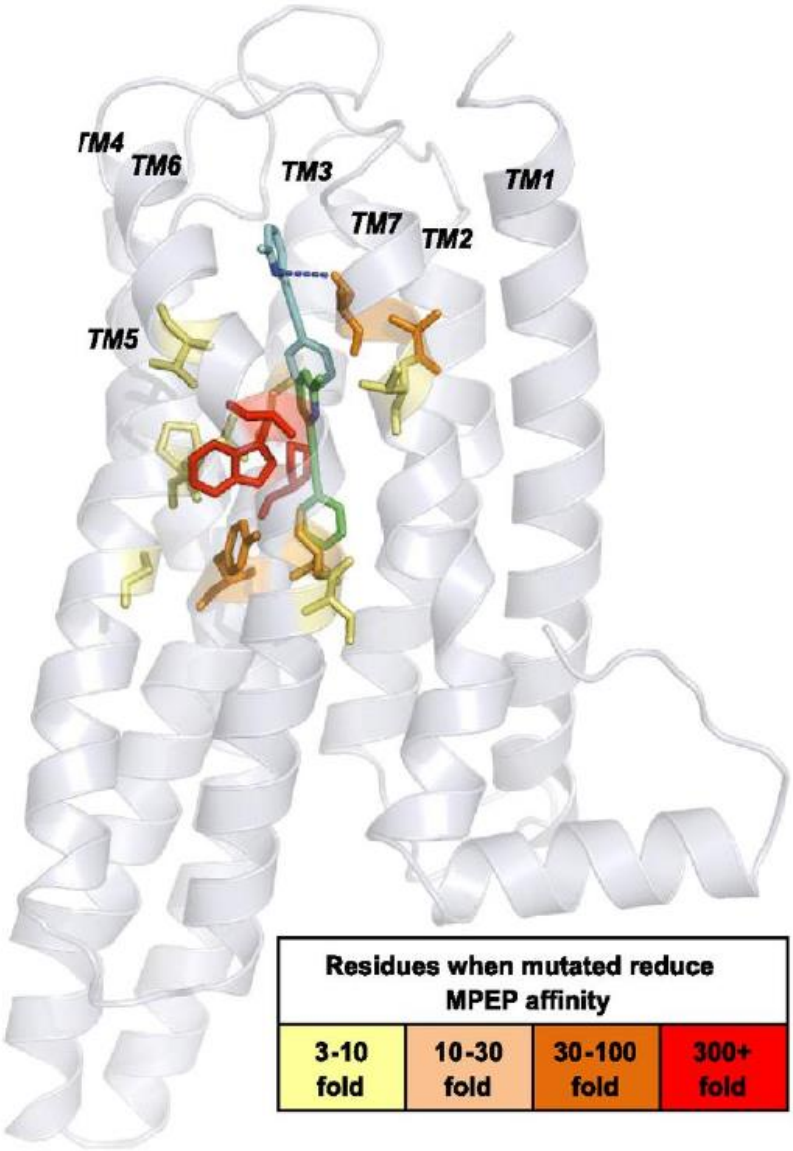


Figure 2.3 Application of RosettaLigand docking of negative allosteric modulator MPEP into comparative model of mGlu5 transmembrane domain. The predicted lowest energy MPEP docking position (cyan) is close to residues demonstrating change in MPEP modulations when mutated (yellow to red). Reprinted with permission from Gregory *et al.*[1].

2.6 Incorporating Experimental Data

While Rosetta can sample near-native structures in a variety of situations, knowledge of limited experimental information can guide sampling and discriminate conformations inconsistent with experimental data, allowing more accurate determination

of structures with less sampling. The incorporation of experimental data most commonly takes the form of modifications to the energy function. Addition of experiment-based scoring terms can make the energy landscape less rugged, allowing Rosetta sampling to more rapidly converge on relevant conformations.

For the incorporation of such information, Rosetta has a flexible restraint system (termed “constraints” in Rosetta parlance). Rosetta constraints are composed of a two-part organization: specification of structural measurements such as distances or angles, and a function which converts the measurement into an energetic penalty. A wide variety of measurements and functional transformations are currently available within Rosetta, and these can be freely mixed and matched according to the particular use case. There are also built-in tools to combine experimental data, allowing users to select only the best of a set of potentially inaccurate restraints. The flexibility of these restraints allow them to be applied in a diversity of situations, from incorporation of Nuclear Overhauser Enhancement (NOE) distances from NMR spectroscopy [258], to use of mass spectrometry cross linking information [259], to the use of custom potentials derived from probability distributions matching EPR/DEER measurements [260, 261].

Although the constraint system provides flexibility when incorporating experimental data for most Rosetta protocols, other experimental data types may reflect more complex structural parameters and require specialized scoring terms. Residual dipolar couplings [262], pseudocontact shifts [263], and small angle X-ray scattering [264] have all been incorporated into Rosetta using specialized score terms, as have several techniques for working with EM and X-ray based electron density [221, 265, 266]. An example tutorial for using X-ray crystallography data and electron density maps with Rosetta is provided in

the Supporting Information.

Improvements in image data analysis and electron detectors have led to advances in electron microscopy, producing electron density maps at resolutions as high as 3 Å for complex molecular machines. However, model building into these near-atomic resolution electron-density maps is still difficult and error prone. DiMaio *et al*. have developed methods in Rosetta that incorporate medium to high-resolution (3Å to 5Å) cryo-EM maps for density guided structure determination and structure refinement [265-267].

2.7 Protein Design

2.7.1 "Inverse Folding Problem"

Protein design is a unique protocol in that instead of finding the optimal conformation of a particular sequence, it aims to determine an optimal sequence for a given conformation. For this reason, it is often referred to as the "inverse protein folding problem" [224]. Generally, there are two main design strategies: design for stability and design for function. The stability protocol considers the entire protein for design, and the score terms of interest are generally focused on improved packing. The design for function protocol is usually a localized design, centered on a specific region, domain, pocket, etc., of a protein with a focused energy function that governs precise interactions, such as electrostatics or hydrogen bonding.

Protein design involves iterative optimization of sequence and structure. During fixed backbone and side chain optimization, sequence space is sampled simultaneously

with side chain conformational space using Monte Carlo simulated annealing by swapping out all possible amino acids at user-specified designable positions while evaluating the predicted energy [193]. This is followed by flexible backbone minimization to optimize the model. The first successful use of *de novo* RosettaDesign produced a sequence for a fold not seen in the PDB [193]. The experimentally determined structure had an RMSD of 1.1 Å to the computationally design model. An example tutorial for protein design with Rosetta is provided in the Supporting Information.

2.7.2 Design for Stability

Protein stability can be affected by a single-point mutation. Kellogg *et al.* evaluated several protocols with varying levels of flexibility and sampling and determined one method in particular to be useful for single-point mutations [194]. This method was made into the application ddg_monomer. When ddg_monomer was tested on a set of 1210 single-point mutants from the ProTherm database, the correlation of predicted ddGs to experimental ddGs was 0.69 while the stability-classification accuracy was 0.72.

While ddg_monomer is a tool to predict how a single-point mutation affects the stability of a protein, RosettaVIP (void in packing) is a design strategy that has been developed to identify single-point mutations that could improve the stability of a protein [196]. When Borgo *et al*. fully-designed proteins, they found that the hydrophobic core of the designs was poorly packed when compared to their respective native proteins. RosettaVIP was able to identify packing deficiencies and sample a much smaller sequence space to fill the void in packing resulting in a more stable design.

2.7.3 Design for Functionality

In addition to stabilizing monomeric proteins, RosettaDesign can be used to design interfaces between proteins. Fleishman *et al.* established a dock design protocol that optimizes the sequence of a protein to bind a surface patch of a target protein during design. Docking was used to optimize the positioning of the interacting proteins at the interface. Experimentally determined structures had an interface very similar to the designed models[197].

Other types of interfaces of interest for design applications are protein-small molecule interfaces. Tinberg *et al*. 2013 is a great example of using RosettaDesign to design for affinity as well as stability (Figure 2.4) [3]. First, RosettaMatch [268] was used to find a stable scaffold for design for binding a particular small molecule. Next, RosettaDesign was used to maximize the binding affinity between the protein and small molecule. Finally, a second round of design was used to minimize destabilization due to mutagenesis in the first round. To ensure these mutations were meaningful, design was guided by a multiple sequence alignment. The resulting most energetically favorable model was the highest affinity binder in experimental studies, and had a co-crystal structure which agreed with the computational model.

Most design algorithms in Rosetta are performed while considering a single fixed-backbone structure. Recently, efforts have been made to consider several structures during the design process in order to tackle more difficult design problems. A generalized multistate design protocol was introduced in 2011[195] to help cases where design should occur to satisfy multiple conformations or to design specificity towards one state and

negative design against other states. Willis *et al*. [269] showed that RosettaMultistateDesign was capable of predicting residues that were important for polyspecificity when designing the heavy chain variable region of an antibody. Sevy *et al*.

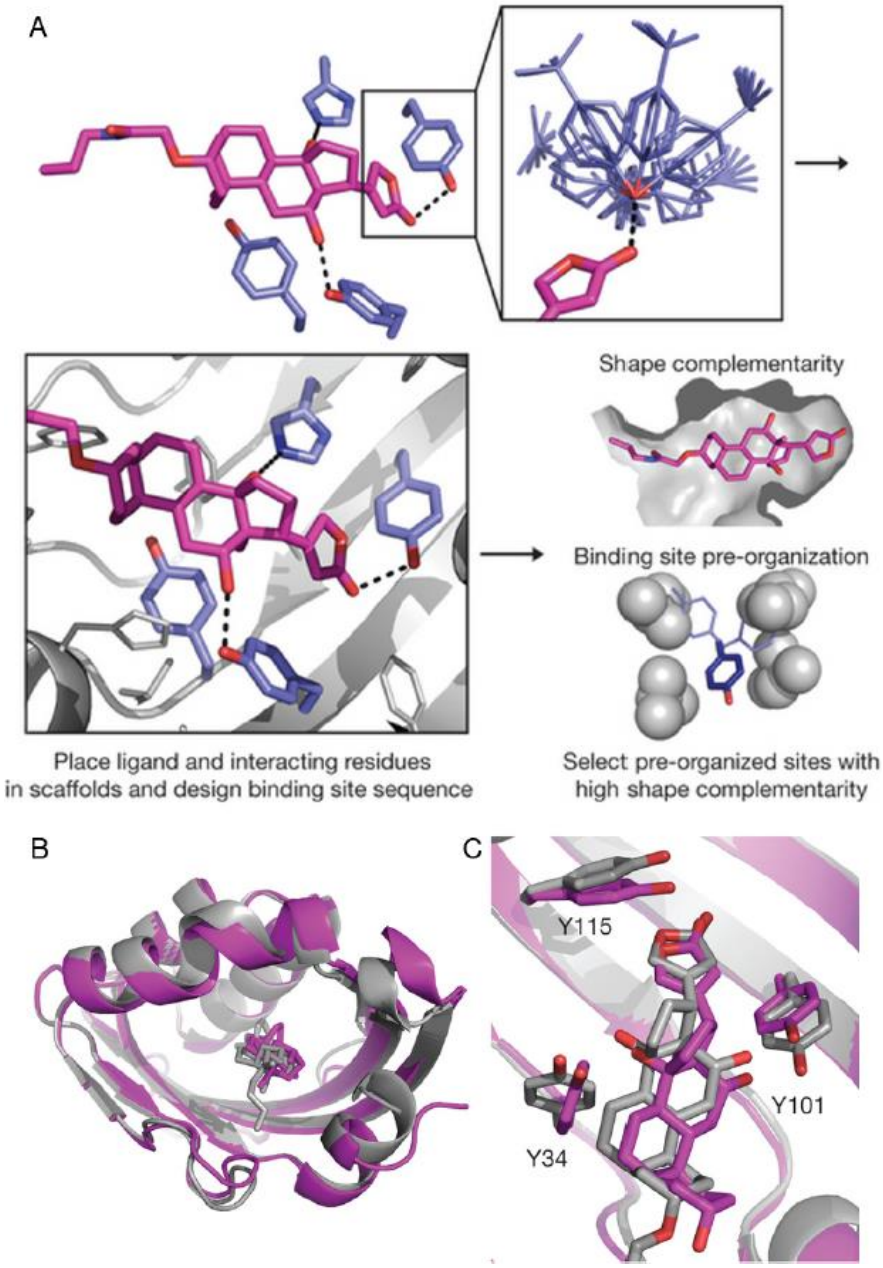


Figure 2.4 Design of protein-ligand interactions for high affinity and selectivity. (A) The design approach involved specifying binding interactions between the protein and ligand followed by design of the binding site. Finally, only designs where shape complementarity was better than what is seen in native complexes were selected for experimental characterization. (B) The designed crystal structure (purple) and computational model (grey) of the protein ligand complex resulting from design for high affinity and selectivity. The RMSD was 0.54 Å while the bound (C) had an RMSD of 0.99 Å. Reprinted by permission from Macmillan Publishers Ltd: Nature, Tinberg *et al.[3]*, copyright 2013.

introduced a new approach to multistate design that accelerates the process of multistate design by reducing the sequence search space [270], allowing more complex backbone movements to be incorporated into a design protocol.

2.8 Additional Rosetta Methods

2.8.1 Membrane

RosettaMembrane has been the method used to model helical transmembrane proteins for several years. RosettaMembrane consists of both a low-resolution [219] and a high-resolution[271] scoring function that were developed to describe how the membrane protein interacts with the membrane environment. Recently, RosettaMP, a new framework for modeling membrane proteins in Rosetta, was developed to interface the membrane environment with both scoring and conformational sampling [208]. Work is ongoing to adapt existing protocols to be compatible with RosettaMP.

2.8.2 Non-canonical Amino Acids and Non-canonical Backbones

Rosetta was initially developed to predict the three-dimensional structure of proteins using the twenty canonical amino acids. However, the expansion to include non-canonical amino-acids (NCAAs) and non-canonical backbones (NCBs) is important, as they allow for the flexibility to create more precise interactions between proteins [272], metal ions [273], or antigens [274]. While the expansion to include more diverse structures is critical, the addition is non-trivial.

The addition of NCAAs requires the modification of both the scoring function and

how the space is explored. These hurdles, however, are not easy to clear, as Rosetta is built on a foundation of knowledge-based components to its scoring function. Most of these knowledge-based score-terms come from published protein structures, and few NCAAs have a statistically relevant representation in the PDB. Therefore developers need to rework key components of the Rosetta scoring function [206]. All score terms were then re-weighted to account for the changes in the score term. Along with the new score terms, the authors created rotamer libraries for 114 NCAAs, as well as a tool, MakeRotLib, to create rotamers for user supplied NCAAs.

An effort was also undertaken to add non-canonical backbones to Rosetta, and in the initial attempt, five new backbones were added [207]. The first hurdle in the addition of an NCB is defining what a "residue" is. In Rosetta3 [223] the "residue" became the central object, therefore with NCBs, a repeating subunit must be defined. Additionally, new backbone sampling movers must be created, or the backbone must be fixed, as the NCB will have differing flexibilities as compared to a linear chain of three singly bonded atoms. The final key point in the addition of NCBs is the creation of new rotamer libraries for the side-chain. Even if the side-chain atoms are identical to a canonical side-chain, the chemical change in the backbone will cause different flexibilities, either due to sterics or electrostatics. A peptoid (a backbone structure identical to the canonical backbone, with the only change being the side-chain branches from the nitrogen instead of the α-carbon) rotamer generator has been created [275], for users to create rotamers for their own side-chains. However, care must be taken when creating rotamers for a blended backbone system [276].

The main considerations for a user attempting to use NCAAs/NCBs in Rosetta is

understanding the chemical properties of their side-chain and/or backbone and properly representing this knowledge in Rosetta. The correct score-terms need to be applied, as the standard knowledge-based score-terms will not apply. An appropriate rotamer library and/or mover must be added to allow for proper sampling of the protein landscape. And finally, the user must understand that because work on NCAAs/NCBs is still limited, novel score-terms or sampling methods may be required.

2.9 Conclusions

The Rosetta software suite represents a compilation of computational tools aimed at obtaining physically-relevant structural models of proteins, RNA, and small molecule interactions. Herein we presented a general outline of updated Rosetta applications, protocols, frameworks and functionalities with the aim of improving user success. All protocols are generalizable and can be applied to an extended list of biological queries that other structure-determining methods cannot address.

Improvements to the variety of Rosetta interfaces (RosettaScript, PyRosetta and many web interfaces) allow the user a high degree of flexibility and personalization for each specific structural problem, as well as providing a previously unavailable entry point for novice users.

The current, default Rosetta score function (*talaris2014*) has been optimized and improved with new score terms as well as reweighted knowledge- and physics-based potentials. Rosetta also incorporates a new release of the Dunbrack rotamer library [217].

De novo structure prediction has greatly improved with the implementation of the TopologyBroker which was developed to create a consensus sampling which satisfies all user-requested constraints without requiring additional code development for each unique system. Recent progress in comparative modeling applications have broadened the conformational search space possible by incorporating multiple starting templates. Protocols for protein-protein docking now include flexibility to modularize the coarse-grained and high-resolution modes of RosettaDock, giving the user more freedom to incorporate additional features in the docking process while narrowing the computational search space. Improvements in protein-small molecule docking utilizes an improved *Transform* algorithm which increases both the speed and quality of this tool in obtaining more native-like conformations. Likewise, the flexibility in incorporating experimentally-derived constraints for most protocols has also greatly improved. In order to tackle the challenge of the inverse folding problem, new implementations of multi-state design permit users to optimize sequences while considering several structures simultaneously.

Continuous developments in Rosetta have increased its functionality by adding functionality to model proteins embedded in the membrane, expansion into non-traditional protein modeling by adding non-canonical amino acids, and nucleic acids, as well as adding the ability to model ever-larger proteins by the addition of symmetry.

CHAPTER 3

Rosetta's Predictive Ability for *in silico* Fragment-Based Drug Discovery

3.1 Summary

This chapter describes a method and benchmark for virtual screening of drug fragments within the binding pocket of a receptor. The model protein used was HisF, which is easily manipulated in an experimental and computational setting. In order to generate well-converged binding poses for the drug fragments I developed a multi-round docking approach with decreased flexibility of docking in each round. Additionally, I structurally characterized the HisF variant used for screening via X-ray crystallography. This chapter comes from the research article "Rosetta's Predictive Ability for *in silico* Fragment-Based Drug Discovery" for which I am the co-first author.

3.2 Introduction

3.2.1 Computer-Assisted Drug Discovery Is an Important Component in Drug Discovery

The ability of computational software to predict the binding of drug compounds to target structures is a sought-after ability to accelerate drug discovery and lead development processes. Various computer algorithms have been developed over the decades dedicated to this problem [277]. The two main approaches to these predictions are ligand-based and structure-based computer-assisted drug discovery (LB- and SB-CADD). In LB-CADD,

cheminformatics is applied to a set of ligands of known activity to attempt to describe features that distinguish actives from non-actives such as substructure scaffold, hydrogen bond donors, and molecular weight [278-280]. In these studies, no knowledge of the structure of the protein target is needed. Alternatively, SB-CADD uses the structure of the target protein to guide docking scenarios that can estimate degree of fit for various ligands and attempt to predict binding energy [203, 281]. Both strategies have proven useful in the field of drug discovery and new developments are continuing to improve their predictive abilities through rigorous testing.

3.2.2 SB-CADD can Rapidly Advance Initial Hit Discovery

In SB-CADD, ligands are docked into a known structure of the target protein in a process called virtual high throughput screening. The goal of this strategy is to identify hits that are competent to bind in the available pocket. Since the goal is to rapidly screen a large library of compounds, the energy function is designed for speed and often uses rough estimations of physical properties to improve the throughput in finite time [282, 283]. These energy functions allow for dense sampling of conformational space and can often result in near-native predicted binding poses. However, because binding energies that discriminate actives from inactives rely on highly specific interactions like hydrogen-bond donors and acceptors and Van der Waals complementarity, these programs often fail at ranking the binding affinity of predicted hits [284, 285]. Therefore, a commonly used strategy for SB-CADD predictions is to use one program for the docking experiment and a second program for scoring compounds [286-288].

3.2.3 RosettaLigand has Two Score Functions for Docking and Scoring

Rosetta, a protein modeling software suite for protein structure prediction and design [223, 289], has two energy functions that are optimized separately for sampling and scoring. The low-resolution function rapidly samples conformational space and pose orientation while the high-resolution scoring function provides realistic energy calculations. RosettaLigand utilizes both functions in its docking strategy with an initial low-resolution grid-based docking and a subsequent atomic-based scoring [201-203]. Further, RosettaLigand is capable of fully-flexible docking in which both the ligand and protein are allowed to sample various conformations to allow for induced fit binding. This strategy has shown to be quite useful in the docking predictions of ligands for crystal structures [255], comparative models [192, 256, 257, 290, 291], and even in the case of enzyme design [292-295] which relies on extremely accurate detailing of ligand-protein contacts.

3.2.4 Fragment-Based Drug Discovery Identifies Hits Using a Focused Library

A common theme in the above studies is the use of drug compounds with high potency and several binding interactions between the drug and protein. Identification of drug compounds via standard high throughput screening requires hundreds of thousands of compounds in a library to cover the broad chemical space of compounds of this size (~500 Da). An alternative approach to screening drug-like compounds is to use fragment-based drug discovery (FB-DD) which uses smaller compounds (average molecular weight 200 Da) that provide starting hits for subsequent lead development [296]. Multiple hit fragments can subsequently be combined to increase the affinity of these compounds [297]. By screening substructures, these libraries only need to be on the order of several thousand

compounds [298].

3.2.5 NMR can Determine Affinity and Binding Pose Simultaneously in FB-DD

In FB-DD, the small size of the fragments results in few specific binding interactions between the fragment and the target protein. As such, the binding affinities are often quite weak and millimolar affinities are considered reasonable starting points [299]. Biophysical techniques such as nuclear magnetic resonance (NMR), isothermal calorimetry, and surface plasmon resonance that are sensitive to such weak binding interactions are needed to screen a fragment library. NMR has proven extremely useful in this regard as this technique provides a measure of binding affinity while also identifying the binding pose within the protein target [300]. Further, the use of SOFAST-HMQC allows for rapid screening by NMR enabling this method for high-throughput screening [301].

3.2.6 RosettaLigand has not been Tested in *in silico* FB-DD

Here, we benchmarked the predictive ability of Rosetta [188, 223] to accurately dock fragments to a protein target. This was a rigorous challenge for the docking algorithm as these fragments possessed low affinity and needed only a few interactions to bind the protein. For this test, we identified a set of fragments that bound to a TIM barrel protein HisF with affinities in the high micromolar range via NMR chemical shift analysis. These fragments were blindly docked into the binding pocket of the protein structure we determined via X-ray crystallography. This docking strategy was able to distinguish actives from inactives indicative of a discriminatory scoring function. Further, we found that the

docking poses of the actives, guided only by Rosetta's sampling algorithm, correlated strongly with the residues identified in the NMR experiments. Lastly, we performed a ligand-based prediction using SurflexSim[302] to align these compounds with the previously crystallized native ligand for HisF which matched closely to Rosetta's predicted binding pose. Taken together, these results show that Rosetta can confidently dock and score fragments despite their low affinity and number of specific interactions. This suggests that Rosetta is a valid starting tool for FB-DD and subsequent drug design.

3.3 Methods

3.3.1 HisF Mutagenesis, Expression, and Purification

The sequence of wild type HisF was obtained from PDB entry 1THF. This sequence was synthesized and subcloned into the pBG100 vector, which encodes an N-terminal hexahistidine purification tag and a TEV protease cleavage site. The primers HisF-C9S-Fwd 5'-CTGGCGAAGCGTATTATCGCGAGCCTGGACGTTAAAGACGGTCGC-3' and HisF-C9S-Rev 5'-GCGACCGTCTTTAACGTCCAGGCTCGCGATAATACGCTTCGCCAG-3' were used to generate HisF-C9S using QuikChange PCR. The plasmids encoding HisF-WT or HisF-C9S were transformed into BL21(*DE3*) cells. Overnight cultures were grown in LB media. The following morning, the cells were centrifuged and the media removed. Cells were then transferred into M9 media containing 0.5 g/L $(^{15}NH_4)_2SO_4$ (Cambridge Isotopes) and grown at 37 ^{o}C to an OD_{600} of 0.5. Flasks were then moved to room temperature and shaking continued until an OD_{600} of 0.6-0.8 before inducing protein production with ITPG for overnight growth. The cells were harvested by centrifugation,

5000 x g for 20 minutes at 4 ^{o}C, and stored at -80°C until needed. Cell pellets were lysed by sonication in 20 mM Tris-HCl, 5 mM imidazole, 200 µg/mL lysozyme, and Roche EDTA-free protease tabs. Lysate was clarified by centrifugation at 22,000 x g for 20 min at 4 ^{o}C. The clarified solute was purified over packed TALON-Cobalt resin and immobilized protein was eluted with 250 mM imidazole. The protein was dialyzed against 10 mM MES, pH 6.8, 50 mM KCl, and 1 mM EDTA and stored at -30 ^{o}C until needed.

3.3.2 Crystallization of HisF-C9S

HisF-C9S was dialyzed against 25mM potassium phosphate containing 150mM NaCl at pH 7.5. Crystals of HisF-C9S were grown using the hanging drop vapor diffusion method. In short, HisF-C9S was concentrated to 10 mg/mL and set in 24-well plates using a ratio of 1:1 with well solution. Final crystals were grown in a solution of 0.1 M Tris, pH 7.5, and 25% PEG 3350. Crystals were cryo-cooled after a brief soak in artificial reservoir solution containing 25% ethylene glycol. X-ray diffraction data were collected at 77 K on a Bruker Microstar resulting in 1.9 Å resolution diffraction data. Data was processed in HKL2000. The phases were determined by molecular replacement using Phaser on an ensemble of templates (1THF, 2AON, 1VH7) generated with Ensembler. Additional refinement was carried out with Phenix.Refine.

3.3.3 NMR Spectra and Assignment Transfer

NMR spectra were collected on uniformly labeled ^{15}N labeled HisF-WT and HisF-C9S at a concentration of 100 µM. All samples were spiked with D_2O to a concentration of 5% to lock the signal. Experiments were performed on a 600 MHz Bruker spectrometer fitted with a QCI-P cryoprobe and a sample jet. Backbone assignments of HisF-WT [303]

were transferred from a previously published ^{1}H-^{15}N TROSY-HSQC spectrum at 30 °C to our ^{1}H-^{15}N HSQC spectra collected at 25 °C. The assignments of HisF-WT at 25 °C was then transferred to ^{1}H-^{15}N SOFAST [304] HMQC spectra of HisF-C9S at 25 °C which was used for successive rounds of small molecule screenings and titration. All NMR spectra were processed using NMR pipe [305] and analyzed with Sparky [306].

3.3.4 Small Molecule Screening

The Vanderbilt fragment library was screened using a 96 well plate NMR setup to identify weak binders of HisF-C9S. Three plates consisting of 12 compounds per well were screened to identify candidate molecules. The individual concentration of the 12 compounds in each well was 600 µM. The mixture of 12 compounds in each well was dissolved with 100µM HisF-C9S protein in NMR buffer with 4% DMSO and 10% D_2O. Small molecules were screened by observing changes in ^{1}H-^{15}N backbone chemical shift of HisF-C9S residues at the binding pocket during SOFAST HMQC experiment. Wells displaying chemical shifts were partially de-convoluted by screening ligands in groups of four. Ligands in wells that displayed chemical shifts were then screened individually to identify the compound(s) that exhibited these effects. Hits identified in this screen were compared against the remaining compounds in the Vanderbilt fragment library using Chemcart. The search was filtered by atomic charge, bond type, and match primary fragment and was limited to small molecules with 0 – 3 rotatable bonds as this was the range of the search set. A total of 86 related fragments were identified and screened against HisF-C9S.

3.3.5 NMR Ligand Titration to Determine Binding Affinity

Hits identified during small molecule screening were further investigated to determine their binding affinity. An excess of up to 12 molar equivalent of small molecule was titrated to 100 µM HisF-C9S and the ^{1}H-^{15}N chemical shift perturbation was monitored through SOFAST[304]-HMQC experiments. Combined ^{1}H and ^{15}N chemical shift of the perturbed peaks were calculated using Equation 1 below [307, 308].

(Equation 1) $$\sqrt{(\Delta CS\_^{1}H)^2 + (\Delta CS\_^{15}H/6.5)^2}$$

Here, $\Delta CS\_^{1}H$ is the chemical shift difference between small molecule bound and unbound protein for amide proton and $\Delta CS\_^{15}H$ for amide nitrogen. Concentration response curves (CRC) were prepared in GraphPad Prism (GraphPad Prism version 6.04 for Windows, GraphPad Software, La Jolla California USA, www.graphpad.com) [309].

3.3.6 RosettaLigand Blind Docking of Fragments

The fragments were docked into the structure of HisF-C9S using RosettaLigand as described previously [192, 310]. Mol files of the fragments were used to generate conformers using confab. The output conformer mol files were then converted into Rosetta-readable params and pdb file. Since the binding site was unknown, fragments were placed in the center of the large convex face of HisF-C9S using the StartFrom mover in RosettaScripts [225]. The entire width of this pocket (10 Å) was used as the potential binding region and defined the box size (radius) as 6 Å such that the scoring grid would cover a space just larger than the potential binding pocket. In the first round of docking, the fragments were allowed freedoms of 1 Å step sizes and full 360º rotation in the Transform mover [203]. 5000 models were generated in the first round of docking. The models were

sorted by overall Rosetta energy and then the top 10% were sorted by interface energy. The top 50 models by interface energy were used as input in the second round of docking. In the second round of docking, fragment movement was reduced to 0.2 Å step sizes and 45^{o} rotation in the Transform mover. Again 5000 models were generated (100 from each parent model), and the same sorting scheme was used to identify the top 50 models. To ensure model diversity, no more than 4 models from a single parent model (or less than 10% of the 50 models) were allowed in the final set of 50 models. These 50 models were then used in a last round of docking in which the fragment freedom was reduced further to 0.04 Å step sizes and 5^{o} rotation in the Transform mover. 5000 models were generated. After selecting for the top 10% by Rosetta energy score and sorting by interface score, the top 10 models were selected for analysis. In every stage of docking, the InterfaceScoreCalculator [290] was used to determine the interface score of the fragment-protein complex. This is calculated by scoring the model of the complex and then moving the fragment 1000 Å away from the protein and rescoring.

3.3.7 Assessment of RosettaLigand's Discrimination of Actives from Inactives

The full set of experimentally tested drug fragments were docked in the binding pocket of HisF-C9S as described above. Average binding energies of the top 5% of docking poses were used to rank the fragment binding. These rankings were compared to the list of fragments that displayed binding interactions. A receiver-operator curve was generated using R analysis.

3.3.8 ROC Analysis of Rosetta Binding Site Predictions

Rosetta binding site predictions were analyzed using the ddg mover in

RosettaScripts [225]. Energies of the bound protein-fragment complex were compared to those of the protein *apo* state and differences in the energy of the two states were computed on a per residue level. The energies of the top 10 binding poses were averaged. Averages that displayed a non-zero energy between the two states were assigned as Rosetta-predicted binding sites and the magnitude of the energy was used to rank the strength of the predicted residues. Binding pockets residues were identified as those that displayed a displacement of the ^{1}H-^{15}N chemical shift upon titration of the fragment. The accuracy of the predicted binding site residues was determined using the ROCR library in R [311]. A modification to this analysis was that for residues that displayed a displacement in the HSQC spectrum, Rosetta energies were queried for that given residue (i) and the residues adjacent (i–1 and i+1) for non-zero values as the displacement of a backbone-amide chemical shift may result from ligand interaction with any of these three residues. Measures were calculated as a true positive rate compared to a false positive rate and the area under the curve was determined. Only those residues which have been assigned in the HSQC spectrum (63 of 253 residues) were used for this analysis to not over-calculate the true negative rate.

3.3.9 Surflex-Sim Alignments to Binding Mode Hypotheses

Surflex-Sim as implemented in the SybylX 2.1.1 suite of computational chemistry programs was used to generate alignments of HisF binders and non-binders to either computed binding mode hypotheses or the PRFAR crystal structure [278]. The PRFAR structure was trimmed to exclude the flexible sidechain past the amidine moiety to constrain alignments to a physically realistic region of space. For these alignments ring flexibility was considered, pre- and post-run energy minimizations were performed, spin density was set to 5, number of spins per molecule was set to 20, and up to 50

conformations of each molecular fragment were investigated. This choice of parameters was made to ensure that a thorough sampling of both conformational and 3-dimensional search space was performed when aligning compounds. Once completed, the most similar conformation of each compound to the hypothesis (according to score) was extracted and used for analysis.

3.4 Results

3.4.1 Selection of HisF-C9S for NMR and Computational Experiments

We selected the (βα)8-barrel (TIM-barrel) protein HisF from *Thermatoga maritima* for these studies as it has been characterized by NMR and X-ray crystallography. We obtained the protein sequence of HisF from PDB entry 1THF [312], the highest resolution structure of HisF. Of note, the protein sequence from 1THF contains a point mutation of Thr21Ser. As this mutation is conservative and outside of the ligand binding pocket we

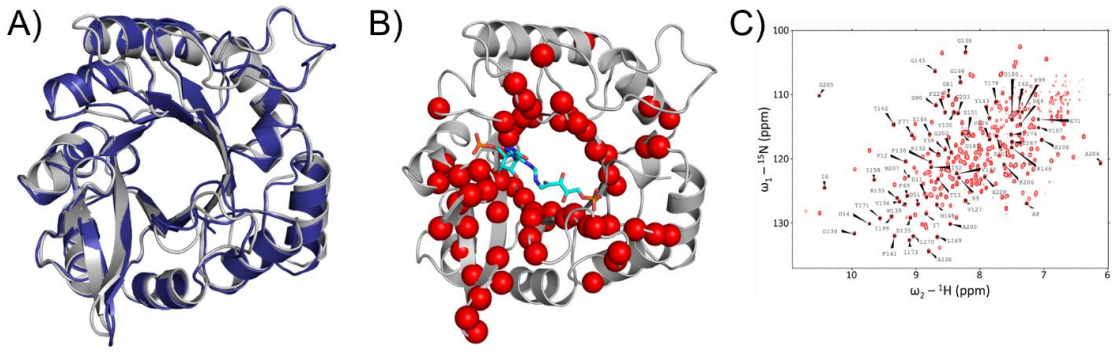


Figure 3.1 The HisF binding pocket. (A) Comparison of the Thermatoga maritima HisF wild-type structure in blue (PDB ID 1THF) with HisF-C9S in grey (PDB ID 5TQL). Overlay of the two structures yields an RMDS of 0.853 Å. (B) After transferring the NMR assignments from HisF-WT to HisF-C9S, a subset of backbone amides was selected based on their proximity to the binding pocket as inferred from the yeast HisF-PRFAR complex (PDB ID 1OX5). PRFAR from the crystal complex is shown in cyan. The selected backbone amide groups are shown as red spheres. (C) ^{1}H-^{15}N HSQC-TROSY of HisF-C9S. Tracked amide residues from panel B are marked.

consider this as our wild-type HisF sequence. Upon analysis of the ligand binding pocket prior to screening we identified a Cys at position 9 which we mutated to Ser to avoid the potential of covalent interactions between fragments and the protein. This mutant, HisF-C9S, has an identical structure to HisF-WT as determined in a crystal structure with 1.9 Å resolution (PDB 5TQL, Table 3.1). The full atom RMSD between HisF-C9S and 1THF is 0.85 Å (Figure 3.1A). Further, the Ser9 sidechain adopts the same rotameric state in HisF-C9S as Cys9 in 1THF. Taken together, we conclude that the HisF-C9S structure is nearly identical to the HisF-WT structure and provides a suitable protein for NMR screening and docking studies.

Table 3.1 Crystallographic Data Collection and Refinement Statistics for HisF-C9S (PDB ID: 5TQL)

Data Collection Statistics	
Space Group	$P6_1$
Cell Dimensions	
a, b, c (Å)	96.505, 96.505, 96.505
α, β, γ (°)	90, 90, 120
Unique Reflections	64115 (6359)
Resolution (Å)	1.90 (1.97-1.90)
Completeness (%)	99.9 (99.7)
Redundancy	7.53 (3.74)
I/σ(I)	14.08 (2.46)
Refinement Statistics	
Resolution Range	22.92-1.90 (1.97-1.90)
R(work)	0.189 (0.254)
R(free)	0.220 (0.261)
Mean B Value ($Å^2$)	21.63
RMSD	
Bond Length (Å)	0.007
Bond Angles (°)	0.81

3.4.2 NMR Assignment of HisF-C9S Spectra

Amide ^{1}H-^{15}N resonances from a previously assigned HisF spectrum [303] collected at 30 °C were transferred to a ^{1}H-^{15}N HMQC spectrum we obtained of HisF-WT at 25 °C. We subsequently generated a ^{1}H-^{15}N SOFAST-HMQC of HisF-C9S at 25 °C and transferred assignments for 150 residues from our HisF-WT spectrum. Since the goal of this study is to identify small molecules that bind within the native ligand site, attention was focused on the known binding pocket of HisF for its natural substrate 5-[(5-phospho-1-deoxyribulos-1-ylamino)methylideneamino]-1-(5-phosphoribosyl)imidazole-4-carboxamide (PRFAR). Of the 150 confirmed residues, an inclusive list of 63 residues was compiled to be within or proximal to the binding pocket, based on residues that were confidently assigned within two shells of the yeast HisF-PRFAR binding pocket and were tracked over the remaining experiments (Figure 3.1B,C).

3.4.3 Identification of Potential Hits to HisF-C9S

The Vanderbilt fragment library of around 14,000 fragments was built from drug-like fragments that contain substructures that often bind proteins such as carboxylic acids and heterocycles [313] and generally adhere to the ‘rule of three’ meaning they have a molecular weight less than 300 Da, a ClogP less than 3, and contain less than 3 hydrogen bond acceptors [314]. 3456 fragments from the library were screened against HisF-C9S to identify fragments with intrinsic binding affinity. Binding of fragments was tested using the ^{15}N-SOFAST heteronuclear multiple quantum correlation (SOFAST-HMQC) NMR technique, which allows for rapid dections of conformational changes at the binding interface induced by small molecule binding [301]. The 3456 fragments were batched on 96-well plates with 12 fragments per well [300, 301]. Fragments in wells that displayed

chemical shift perturbation were tested individually against HisF-C9S for binding activity. A set of 25 hits were identified in this screen. The remaining ~10,000 compounds in the Vanderbilt fragment library were searched for similar fragments using substructure similarity as implemented in the Chemcart software. A total of 86 related fragments were screened against HisF-C9S and 15 were identified as additional hits. Combined with the initial 25 confirmed hits, a total of 40 fragments were discovered that possessed intrinsic binding affinity for HisF-C9S.

3.4.4 Binding Analysis of Individual Hits

The 40 fragments identified as hits were confirmed via an NMR titration experiment to determine their binding affinity. Chemical shift perturbations were monitored in HisF-C9S in the presence of each fragment at concentrations up to 1200 μM. A combined ^{1}H and ^{15}N chemical shift difference was calculated against the apo NMR spectrum according to Equation 1 [307, 308]. This value was plotted against fragment concentration to derive binding curves. Figure 3.2 highlights the titration spectra and binding curve of one of these ligands. Only 31 of the 40 titration experiments could be fit with a one-site binding model (Figure 3.3). The 9 remaining hits all generated chemical shift perturbations during titration experiments but could not be fit to a simple one-site binding model and were removed from the dataset. The K_D values of these naïve binders are, as expected, weak with the best affinities approaching 400 μM. In many instances, K_D values are affiliated with a substantial error as the binding is not saturated at the highest measured fragment concentration.

3.4.5 Binding mode of the small molecules

A HisF-C9S binding pocket hotspot appeared after analyzing the compiled data of the shifting residues from the 31 naïve binders (Figure 3.2C). Among the 63 tracked residues, 35 never displayed any peak shifts and are therefore categorized as not interacting with the fragments. Hot spots for interaction include the catalytic residue Asp130 and many surrounding residues including the nearby loop (residues 142 – 144). Interestingly, catalytic residue Asp11 displayed no peak shifts. According to the crystallized yeast HisF-PRFAR complex, Asp11 interacts with the glycerol phosphate while Asp130 binds the

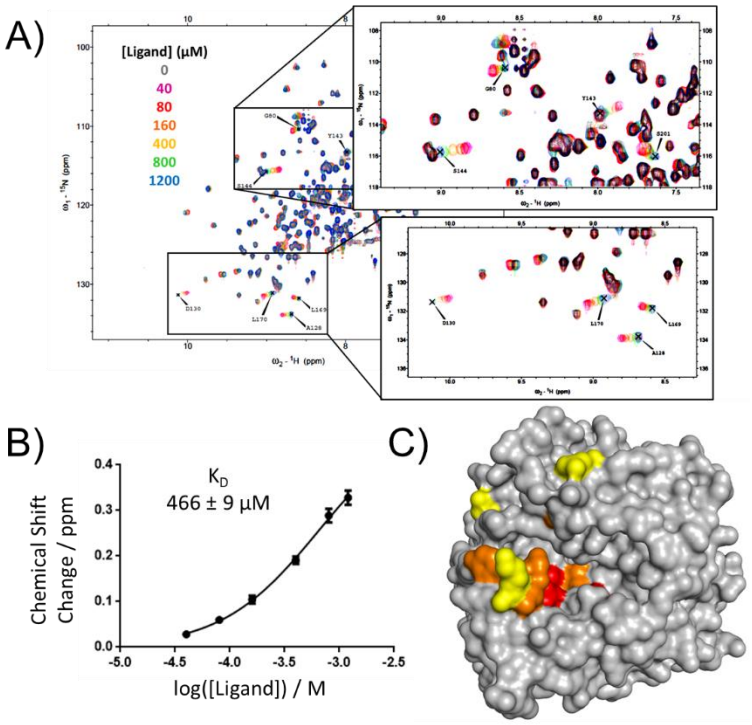


Figure 3.2 Identification of Low Affinity Fragments for HisF-C9S. (A)An example titration of compound VU0139210. HMQC-NMR spectra of HisF-C9S in the presence of 0, 40, 80, 160, 400, 800, and 1200 µM ligand reveal peaks that shift with increasing ligand concentration. An example of significantly shifting peaks have been identified on zoomed in inserts. (B) The plotted binding curve of chemical shift change (ppm) vs log ligand concentration (M) gives a K_D values of 466 µM. (C) Residues that were identified as responding to ligand binding across the full set of 31 low affinity ligands are mapped onto the protein surface. The residues are colored according to frequency of chemical shift perturbation throughout the ligand series with yellow-orange-red representing lowest to highest frequency.

imidazole-ribose region. The binding hotspots for these identified binding fragments cluster around the region that interacts with PRFAR's aromatic ring moieties which may provide insight into the why many of the novel ligands contain aromatic rings.

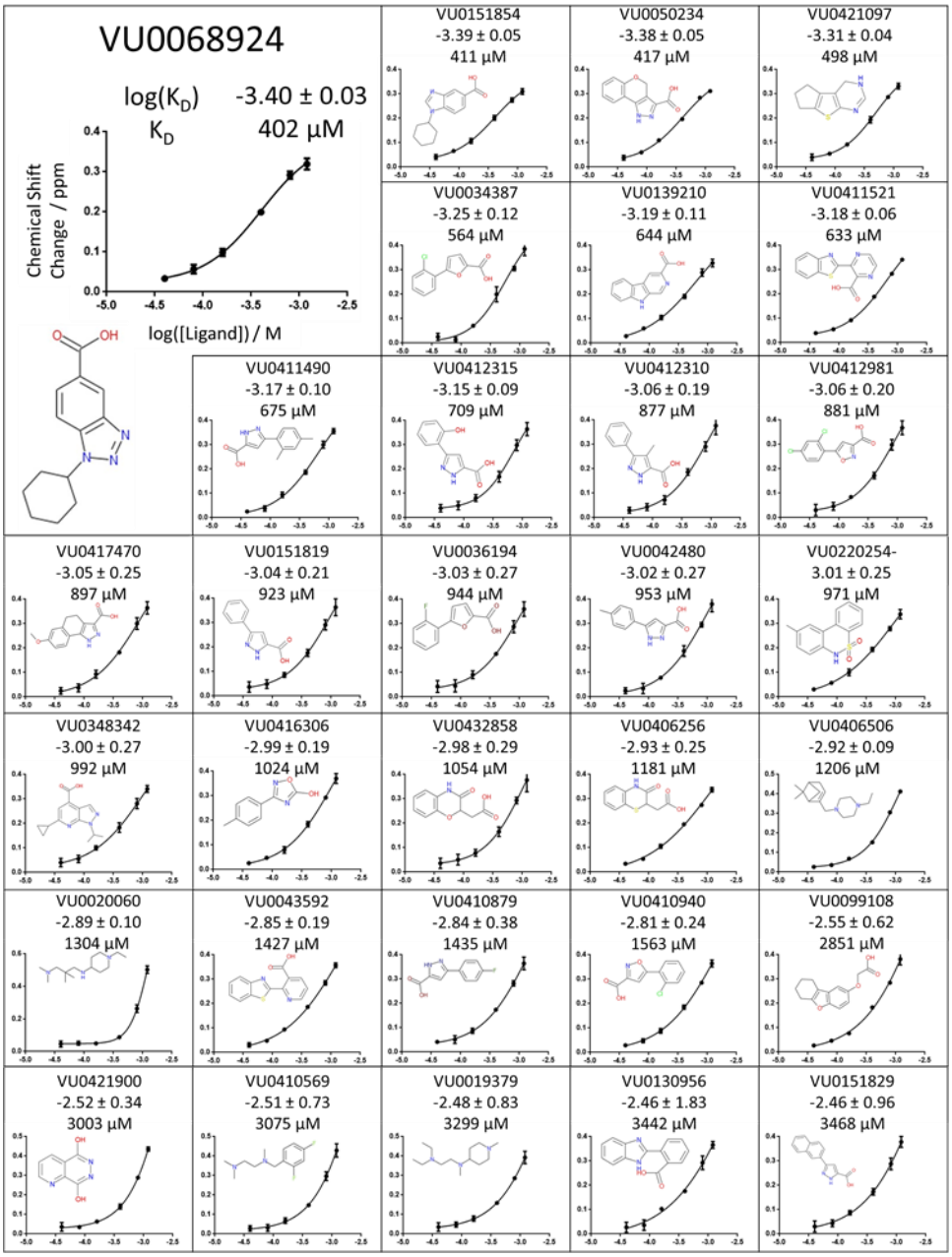


Figure 3.3 Full dataset for the 31 low affinity ligands identified for HisF-C9S. For each compound, the 2D structure and binding curve are shown. Also displayed are the log K_D and K_D.

3.4.6 RosettaLigand Can Discriminate Active from Inactive Drug Fragments

The full set of 3542 fragments that were screened experimentally were also screened *in silico* against the HisF-C9S binding pocket using RosettaLigand. The binding cavity of HisF-C9S is quite large with dimensions of 10 Å x 10 Å x 15 Å. The fragments however are mostly planar with a maximum dimension of around 5-8 Å. To test Rosetta's ability to predict the region in space that the fragments bind to we placed the fragments in the geometric center of the binding pocket prior to sampling of the space [192, 310]. A set of 1000 docked poses were generated for each fragment. The fragments were ranked using the predicted binding energy of the top 5% of these poses. The rank was compared against the set of actives and a receiver-operator characteristic (ROC) curve was generated to evaluate true positives from false positives. From this analysis, it was found that the area

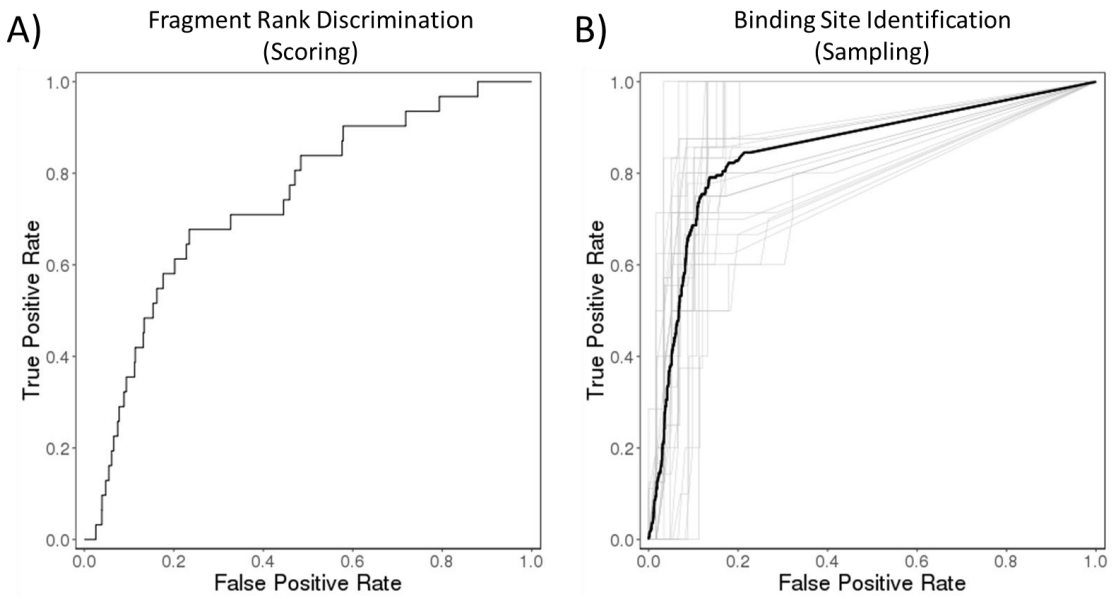


Figure 3.4 ROC Analysis of Rosetta Predictions. (A) Discrimination of actives from inactives using the rank order from average binding energies of all 3546 drug fragments screened by NMR. The AUC was measured to be 0.74 suggesting a discriminatory score function. (B) Shown is the receiver-operator curve for Rosetta predictions of residues involved in ligand binding as verified by perturbations to the chemical shift in fragment titration HSQC spectra. The individual fragment analyses are shown in light grey and the aggregate analysis is shown in dark black. We determined an aggregate ROC curve for all 31 ligands with an area under the curve of 0.85.

under the curve was 0.74 suggesting that RosettaLigand, a structure-based docking strategy, can accurately discriminate active from inactive fragments (Figure 3.4A).

3.4.7 High Density Docking of Active Compounds with RosettaLigand

The 31 confirmed binding fragments were docked into the binding cavity of HisF-C9S over multiple rounds to ensure high density sampling. A round of 5000 docking trajectories with large fragment movements were carried out allowing sampling over the entire pocket (Figure 3.5A). Top models were ranked by binding energy and overall energy. The binding poses were used as starting positions for a subsequent round of docking with tighter restrictions to ligand movement (Figure 3.5B). A last round of docking was carried out on top models again until a consensus binding site was identified (Figure 3.5C). Through each round of docking, the binding mode converged and the energies for the overall complex and binding site improved (Figure 3.5D).

3.4.8 Structure Based Fragment Docking Captures Experimentally Determined Binding Pocket

Following the last round of docking, a set of ten models were selected based on binding energy to best represent the likely binding pose. These binding poses localize to the identified binding pocket in HisF-C9S as evidenced by chemical shift perturbations. To quantify the degree of agreement between Rosetta predictions and experimental results we

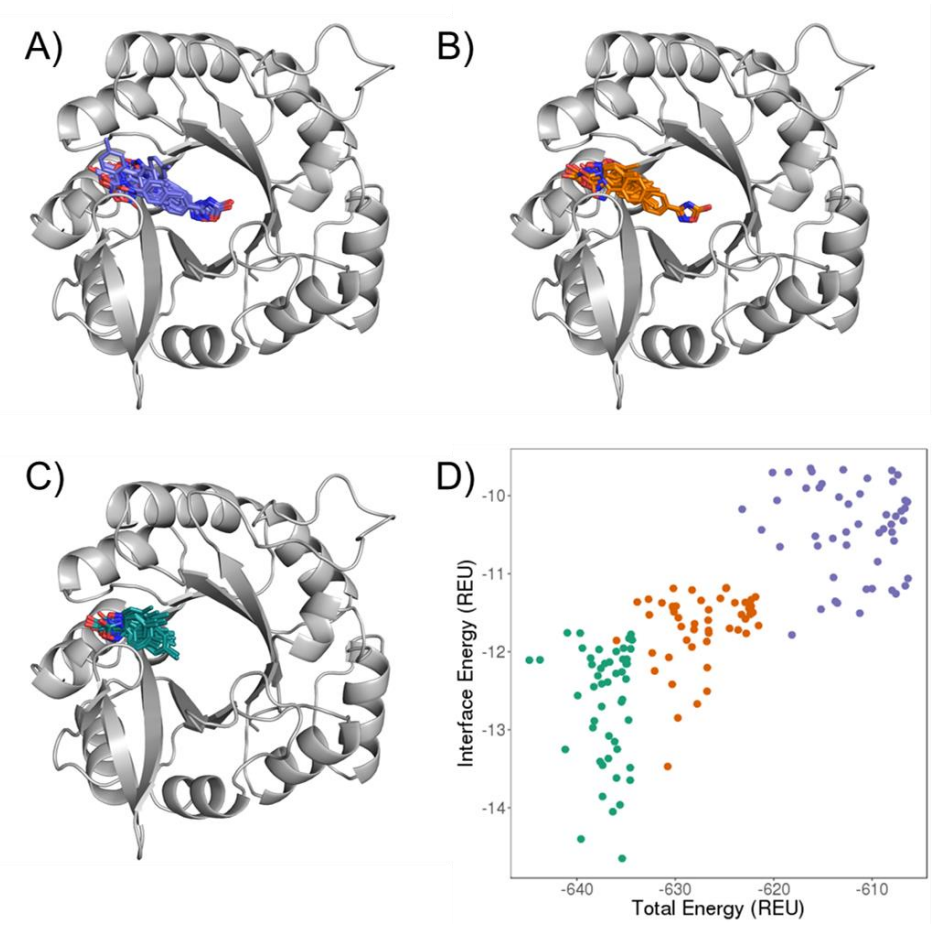


Figure 3.5 Iterative Docking to Identify Binding Pockets. An example docking strategy in which the results of the first round (A) seed docking of a subsequent round (B). A final round of docking (C) results in a highly converged binding hypothesis. (D) Top models from each round were selected based on overall energy and interface binding energy. The values of each docking poses from panels A-C are plotted to show how the energies in both metrics improved with each round of docking.

used ROC analysis to measure a true positive rate of identification. Per-residue ddGs were calculated over the Rosetta predicted protein-ligand complexes and residues involved in binding the ligand were assigned non-zero energies. These residues were compared against the list of chemical shifts that were displaced upon ligand titration. Importantly, if a residue was seen to undergo a movement in the HSQC titration, the Rosetta-determined energies were queried on the i, i+1, and i-1 residues for predicted involvement in ligand binding. This is due to the fact that a backbone-amide chemical shift reports on changes in the

chemical environment which can result from local and non-local changes[315]. Using this approach, we determined an aggregate ROC curve for all 31 ligands with an area under curve of 0.85 (Figure 3.4B). For each of the individual ligands, ROC analysis resulted in area under curve values of 0.72-0.98 suggesting that RosettaLigand strongly predicts the binding location of the fragments within the HisF-C9S binding pocket.

3.4.9 Superposition of HisF-C9S naïve binders with HisF/substrate complex

To gain more insight into trends between chemical structure and binding energy, the crystal structure of PRFAR from the complex with yeast HisF was used to flexibly align all 31 binding fragments using the Surflex-Sim program[278]. Since the NMR and Rosetta docking experiments show that these ligands all bind in the same region of the ribose subunit of PRFAR, the flexible portion of PRFAR was removed during these alignments to ensure that physically realistic alignments were made. Eight of the ten strongest binding fragments adopt a conformation orienting polar or negatively charged groups with the phosphate group of PRFAR and the remainder of the ligand aligning with a

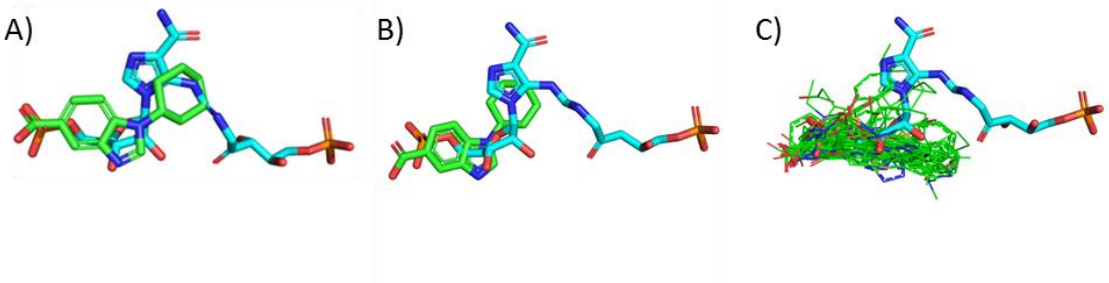


Figure 3.6 Binding hypotheses by computational modeling. (A) The major alignment solution from Surflex-Sim between the HisF naïve binders (green) and the PRFAR crystal structure (cyan). Shown here is VU0151854. (B) The binding hypothesis of VU0151854 from RosettaLigand docking again with the ligand in green and PRFAR in cyan. (C) The docking solutions of all 31 ligands from RosettaLigand (green) with PRFAR (green) highlighting the overall similarity of docked solutions.

portion of the ribose or imidazole subunits depending on the molecular size and character. As seen in Figure 3.6, the overall binding hypothesis is quite similar between the two methods. The use of independent methods that result in similar solutions lends credence to both methods and overall strengthens our binding mode hypotheses.

3.5 Discussion

3.5.1 Identification of Set of Low Affinity Ligands for HisF-C9S

Whereas most tests of ligand docking predictions are performed on high affinity ligands, we sought to test RosettaLigand's predictive ability on moderate to low affinity ligands. HisF was a good choice for these studies owing to its structured characterization by NMR spectroscopy and X-ray crystallography. The high resolution of the HisF-C9S structure is reasonable for docking studies. Additionally, the identification of low affinity binding fragments and their binding location is facilitated by titration analysis NMR. The affinity of these fragments might be too low for co-crystallization while other methods of screening such as isothermal calorimetry or surface plasmon resonance would not provide information about residues involved in ligand binding. Using NMR titration analysis we were able to generate a dataset of low affinity binding fragments to HisF-C9S. After fragment identification, our NMR titration experiments showed that about a quarter of these ligands could not be fit to a one-site binding model. We removed these ligands as the potential for multiple binding sites, cooperative binding, or general non-specificity of binding would not provide a relevant target for binding predictions.

3.5.2 RosettaLigand Predicts Binding of Low Affinity Fragments

The identification of this dataset provided an ideal opportunity to test the ability of RosettaLigand to predict the binding of such low affine fragments. Small molecule docking is dependent on highly specific small molecule and side chain positioning. Accurate prediction of binding for a small molecule fragment is a stringent test of the computational algorithm, as only few interactions are needed to confer the low affinity. This test was additionally made difficult as the ligands were placed in the geometric center of the binding surface and allowed to dock blindly. As demonstrated in the NMR experiments, the binding site of these ligands was away from center in a small region. The rank order of ligands is a task generally used in ligand-based drug screening. However, it was seen that even with low sampling RosettaLigand could accurately rank actives from inactives. On the other hand, binding pocket analysis measures the contribution of individual amino acids to the binding of the small molecule as opposed to total binding energy. Comparison of these predictions with the chemical shifts of residues in the entire binding surface of HisF acquired in the HMQC-NMR titrations showed that despite these difficulties there was a strong predictive ability of RosettaLigand to place these molecules in their respective binding pockets.

3.5.3 Structure-Based Ligand Docking Results in Similar Ligand Orientation as Ligand-Based Results

One possible disadvantage to structure-based docking as attempted here is the disregard of relevant knowledge of ligand orientation and chemical nature that could enhance the likelihood of accurate predictions. The yeast HisF protein has previously been crystallized with the native ligand PRFAR. Despite large differences between the identified

ligand and PRFAR chemical structure, rough similarities such as a negatively charged group linked to an aromatic moiety exist. Alignment of these groups between the novel ligands and PRFAR would provide an ideal starting point for docking studies. As seen in the Surflex-Sim results, such an alignment provides highly similar binding poses despite no input regarding the protein structure. The ability to dock the ligands as a group or with alignment to a known ligand would likely have decreased the number of decoys needed to identify accurate binding poses [316]. The similarity between the Surflex-Sim aligned binding poses and the correlation with experimental results further strengthen the confidence in RosettaLigand's predictive ability for even low affinity ligands.

3.6 Conclusion

Since the mid-1990's, successful application of X-ray crystallography and NMR spectroscopy has guided computational programs to predict ligand binding [317]. Benchmarking algorithms for accurate prediction of protein-ligand binding is highly relevant for improving *in silico* screening of drugs and for structure-guided ligand design. A regular pitfall of these algorithms is the range of errors in predictions for moderate to low affinity ligands. To that end, our test of RosettaLigand's ability to dock ligands with affinities in the high micromolar range provides a rigorous test for ligand docking across a range of affinities. The success of Rosetta in predicting the rank order and binding site of these compounds was significant given the few interactions needed for such weak binding. The results here could be used to enhance the algorithm using feedback from NMR on specific binding interactions. Additionally, identification of accurate binding of such low affinity drug fragments can be used for the designs of higher affinity ligands guided by the context of the binding pocket. Further, as shown here, Rosetta may prove useful in initial

selection of fragments for screening purposes in the development of new drugs for novel protein targets.

3.7 Additional Scripts

Options file, used to specify the input, output, parser file, and packing options. For input files, one must specify the path to the Rosetta database, path to the small molecule PDB file(s), and the path to the small molecule params file. For output, one must specify the type of file to be output (pdb) and number of structures to generate. For the parser, one must specify the path to the XML parser file. For packing, these are standard options to include for side chain repacking.

```
-in
        -path
                -database /path_to_database/database/
        -file
                -s /path_to_pdb/filename.pdb
                -extra_res_fa /path_to_params/ligand.params
-out
        -level 300
        -pdb_gz
        -path
                -pdb    /path_to_output_files/
                -score  /path_to_output_files
        -nstruct 100
        -mute all
        -unmute protocols.jd2.JobDistributor

-parser
        -protocol /path_to_RosettaScripts/RosettaScripts.xml

-packing
        -ex1
        -ex2
        -linmem_ig 10
```

XML Script, for the experiments discussed in this study, used to assign values for

the cut-off points to detect the protein/small molecule interface, a value for the favor native residue bonus, and values for small molecule translation and rotation. In TASKOPERATIONS:DetectProteinLigandInterface, one must specify values to determine which residues surrounding the small molecule are allowed to be designed and/or repacked (details in my first manuscript), and specify the path to the resfile. In Transform, must specify how much small molecule movement is allowed.

```
<ROSETTASCRIPTS>
     <SCOREFXNS>
          <ligand_soft_rep weights=ligand_soft_rep>
               <Reweight scoretype=fa_elec weight=0.42/>
               <Reweight scoretype=hbond_bb_sc weight=1.3/>
               <Reweight scoretype=hbond_sc weight=1.3/>
               <Reweight scoretype=rama weight=0.2/>
          </ligand_soft_rep>
          <hard_rep weights=ligandprime>
               <Reweight scoretype=fa_intra_rep
weight=0.004/>
               <Reweight scoretype=fa_elec weight=0.42/>
               <Reweight scoretype=hbond_bb_sc weight=1.3/>
               <Reweight scoretype=hbond_sc weight=1.3/>
               <Reweight scoretype=rama weight=0.2/>
          </hard_rep>
     </SCOREFXNS>
     <SCORINGGRIDS ligand_chain="X" width="16">
          <vdw grid_type="ClassicGrid" weight="1.0"/>
     </SCORINGGRIDS>
     <TASKOPERATIONS>
          <DetectProteinLigandInterface name=design_interface
cut1=6.0 cut2=8.0 cut3=10.0 cut4=12.0 design=1
resfile="/path_to_resfile/Resfile_dock"/>
     </TASKOPERATIONS>
     <LIGAND_AREAS>
          <docking_sidechain chain=X cutoff=6.0
add_nbr_radius=true all_atom_mode=true minimize_ligand=10/>
          <final_sidechain chain=X cutoff=6.0
add_nbr_radius=true all_atom_mode=true/>
          <final_backbone chain=X cutoff=7.0
add_nbr_radius=false all_atom_mode=true
Calpha_restraints=0.3/>
     </LIGAND_AREAS>
     <INTERFACE_BUILDERS>
```

```
          <side_chain_for_docking
ligand_areas=docking_sidechain/>
          <side_chain_for_final
ligand_areas=final_sidechain/>
          <backbone ligand_areas=final_backbone
extension_window=3/>
     </INTERFACE_BUILDERS>
     <MOVEMAP_BUILDERS>
          <docking sc_interface=side_chain_for_docking
minimize_water=true/>
          <final sc_interface=side_chain_for_final
bb_interface=backbone minimize_water=true/>
     </MOVEMAP_BUILDERS>
     <MOVERS>
     single movers
          <StartFrom name=start_from_X chain=X>
               <Coordinates x=25.325 y=35.021 z=22.716/>
          </StartFrom>
          <FavorNativeResidue name=favor_native bonus=1.0/>
          <ddG name=calculateDDG jump=1 per_residue_ddg=1
repack=0 scorefxn=hard_rep/>
          <Transform name="transform" chain="X"
box_size="5.0" move_distance="1.0" angle="360" cycles="500"
temperature="5" initial_perturb="5.0"/>
          <HighResDocker name=high_res_docker cycles=1
repack_every_Nth=1 scorefxn=ligand_soft_rep
movemap_builder=docking/>
          <PackRotamersMover name=designinterface
scorefxn=hard_rep task_operations=design_interface/>
          <FinalMinimizer name=final scorefxn=hard_rep
movemap_builder=final/>
          <InterfaceScoreCalculator name=add_scores chains=X
scorefxn=hard_rep/>
     </MOVERS>
     <PROTOCOLS>
          <Add mover_name=start_from_X/>
          <Add mover_name=transform/>
          <Add mover_name=favor_native/>
          <Add mover_name=high_res_docker/>
          <Add mover_name=final/>
          <Add mover_name=calculateDDG/>
          <Add mover_name=add_scores/>
     </PROTOCOLS>
</ROSETTASCRIPTS>
```

Resfile, used to indicate that residues considered for design and repack are limited

to the cut-off points specified above.

```
#These commands will be applied to all residue positions that
lack a specified behavior in the body:
NATAA       # allow only native residues (for docking only;
no design allowed)
AUTO
start
```

Rosetta binding site predictions were analyzed using the ddg mover

```
<ROSETTASCRIPTS>
     <SCOREFXNS>
          <hard_rep weights=ligandprime>
               <Reweight scoretype=fa_intra_rep
          weight=0.004/>
               <Reweight scoretype=fa_elec weight=0.42/>
               <Reweight scoretype=hbond_bb_sc weight=1.3/>
               <Reweight scoretype=hbond_sc weight=1.3/>
               <Reweight scoretype=rama weight=0.2/>
          </hard_rep>
     </SCOREFXNS>
     <MOVERS>
          <ddG name=calculateDDG jump=1 per_residue_ddg=1
     repack_bound=0 repack_unbound=1 scorefxn=hard_rep/>
          <InterfaceScoreCalculator name=add_scores chains=X
     scorefxn=hard_rep/>
     </MOVERS>
     <PROTOCOLS>
          <Add mover_name=calculateDDG/>
          <Add mover_name=add_scores/>
     </PROTOCOLS>
</ROSETTASCRIPTS>
```

CHAPTER 4

Redesign of HisF to Increase Binding to a Novel Ligand using RosettaLigand

4.1 Summary

This chapter extends on Chapter 3 to incorporate protein-ligand interface design. Again, the model protein HisF was used for easy manipulation but the protocol is applicable to any protein-ligand binding pocket. For this work I conducted the computational docking, NMR titrations, and X-ray crystallography. We show the feasibility of pocket redesign using a method that does not rely on pre-defined ligand-pocket geometries making redesign accessible to non-experts. This chapter is from an unpublished research article for which I will be first author.

4.2 Introduction

4.2.1 Gene Duplication Allows A Protein to Evolve to Bind a Novel Ligand

Given the vast diversity of protein-ligand interactions in nature, it has been postulated that generation of novel interactions proceeds via gene duplication with modification at the interface [318]. Gene duplication allows for a novel protein to evolve binding to new small molecules and increase the diversity of biologically relevant complexes. The family of $(\beta\alpha)_8$ barrels has previously been used to study the evolution of protein-small molecule interactions [319, 320]. $(\beta\alpha)_8$-barrels, also known as TIM-barrels,

consist of eight repeating (βα) units and are frequently soluble enzymes in metabolic pathways [321]. It is estimated that 10% of enzymes have adopted this fold, despite vast sequence diversity and varied catalytic reactions [322]. $(\beta\alpha)_8$-barrel proteins can be separated into two domains: a stabilization domain and a catalytic domain. Mutations in the catalytic domain can result in new protein-ligand interactions while maintaining a stable protein fold. Several studies have focused on the evolution of this family, in particular amino acid synthesis in bacteria by the protein HisF. Of note, a single amino acid change was needed to convert HisF and a related protein HisA into enzymes that recognized the native substrate of TrpF [323]. Additionally, HisA could be further designed such that a symmetric half-barrel dimer needed only a small number of mutations to establish catalytic activity similar to TrpF [324]. These studies suggested that relatively dissimilar proteins could evolve through a gene duplication and few mutations to accommodate unique ligands.

4.2.2 Protein-Ligand Interface Design Often Fails to Match Evolutionary Success

A better understanding of evolutionary pathways to create small molecule binding pockets in proteins can provide insights for engineering new protein/small molecule interactions. The ability to predict what mutations are necessary to drive a novel activity in a protein is the basis of protein design. The design of proteins is an unmet challenge with implications in the fields of signaling, therapeutics, and polymer production. The redesign of existing proteins has been used to create proteins that respond to non-native ligands such as Designer Receptors Exclusively Activated by a Designer Drug (DREADDs) which can be used for noninvasive studies of neurological phenomenon [325]. Many of these novel

proteins have been generated via random mutagenesis with positive selection. The alternative is rational, computationally predicted design. Computational enzyme design has seen some successes (Kemp elimination reaction [294], Diels-Alder reaction [326], retro-aldol enzymes [292]), but turnover rates are minimal compared to naturally evolved enzymes [327]. Morin *et al* attempted to design a protein to bind a peptide, and although the protein crystal structure superimposed with the computationally predicted structure, the peptide did not bind [328]. Tinberg *et al* described two successfully designed proteins which bind a rigid steroid hormone, but also detailed 15 additional designed proteins which were computationally favorable but showed no experimental indication of binding [3].

4.2.3 Protein-Interface Design Can Mimic Evolutionary Pathways

It is generally accepted that the low initial affinity of a protein for a small molecule provides a benefit for the organism [329]. This initial intrinsic affinity is needed to create a signal and an evolutionary pressure to optimize this interaction. We hypothesize that this intrinsic affinity can be quite low so that a limited interaction with a smaller part of a larger molecule is sufficient. Leveraging evolutionary principles aids the process for computational design of small molecule binding pockets. We argue that the computational re-engineering of proteins to recognize small molecules can be accelerated by mimicking nature's approach of repurposing existing small molecule binding pockets with an intrinsic ability to recognize a small molecule of interest or a fragment thereof. This approach has several advantages: (1) Optimization of an existing binding pocket with an intrinsic affinity is an easier task than *de novo* design of a new binding pocket, and (2) an intrinsic affinity for the small molecule of interest or a fragment thereof indicates the general ability of the protein to bind molecules of this structure. This is distinct from the previously mentioned

Rosetta designs that either used large ligands with several functional moieties or grafted binding pockets onto a variety of protein scaffolds until a match was found that could support that pocket geometry.

We sought to redesign the binding pocket of HisF to bind novel ligands using RosettaLigand, an algorithm originally developed to predict the structure of protein-ligand complexes. We used a drug-like fragment (< 250 Da) that had a weak intrinsic affinity for the protein as measured by HSQC NMR titration. The compound was computationally docked and the sequence was optimized to bind the novel substrate. A case example was expressed and tested for enhanced binding to its substrate. Mutations were tested individually to find a minimal number needed to convey the enhanced binding affinity. These experiments demonstrate the feasibility of RosettaLigand for blind protein-ligand interface redesign of a pocket to enhance binding affinity to a compound with weak intrinsic affinity analogous to evolutionary reprogramming.

4.3 Methods

4.3.1. Ligand Docking and Interface Design

The compound VU0068924 was previously docked into the structure of HisF (Chapter 3). Briefly, conformers of the ligand were generated using the BCL [330]. The ligand was placed in the geometric center of the binding pocket and 5000 decoys were generated with maximal 1 Å step sizes and full 360^{o} rotation using RosettaLigand with XML scripts [201-203, 251]. Models were sorted by total and binding energies and sorted. The top 50 models were redocked from their individual starting positions with maximal 0.2 Å step sizes and 45^{o} rotation. Each starting position was used to generate 100 models for a

final set of 5000. These models were again sorted by total and binding energies and the top 50 models were selected. These models were redocked with similar ligand constraints as the previous step but with sequence optimization turned on. Following this round of modeling the models were sorted by total and binding energies and number of mutations. A final design with 8 mutations was selected.

4.3.2 Expression of HisF mutants

The plasmids encoding HisF variants were transformed into BL21(DE3) cells. Overnight cultures were grown in LB media. Cells were centrifuged at 5,000 x g and 4 ºC for 10 min and the media removed. Cells were then transferred into M9 media containing 0.5 g/L $(^{15}NH_4)_2SO_4$ (Cambridge Isotopes) and grown at 37 ºC to an OD_{600} of 0.5. Flasks were then moved to room temperature for continued shaking until an OD_{600} of 0.6-0.8 before inducing protein production with ITPG. Protein was induced for 18 hours. The cells were harvested by centrifugation, 6500 rpm for 20 minutes at 4 ºC and stored at -80°C. Cell pellets were lysed in 20 mM Tris buffer with 5 mM imidazole containing 200 µg/mL lysozyme and Roche EDTA-free protease tabs followed by sonication. The clarified lysate was purified over packed TALON-Cobalt resin and immobilized protein was eluted with 250 mM imidazole. The protein was dialyzed against 10 mM MES, pH 6.8, 50 mM KCl, and 1 mM EDTA and stored at -30 ºC.

4.3.3 NMR Titration of VU0068924 against HisF variants

HisF variants were concentrated to 50 µM in buffer containing 10% DMSO. VU0068924 was added from concentrations of 0 to 1200 µM to ensure saturation of the binding pocket. ^{1}H-^{15}N chemical shift perturbation was monitored through SOFAST [304]-

HMQC experiments. Combined ^{1}H and ^{15}N chemical shift of the perturbed peaks were calculated using Equation 1 below [307, 308].

(Equation 1) $$\sqrt{(\Delta CS\_^{1}H)^2 + (\Delta CS\_^{15}H/6.5)^2}$$

Here, $\Delta CS\_^{1}H$ is the chemical shift difference between small molecule bound and unbound protein for amide proton and $\Delta CS\_^{15}H$ for amide nitrogen. Concentration response curves (CRC) were prepared in GraphPad Prism (GraphPad Prism version 6.04 for Windows, GraphPad Software, La Jolla California USA, www.graphpad.com)[309].

4.4 Results

4.4.1 Pocket Optimization of Ligand VU0068924

The drug fragment VU0068924 (Figure 4.1B) was identified via an NMR high throughput screen as a ligand with intrinsic affinity for HisF-C9S (Bender, Chapter 3, unpublished). The K_D of this ligand for HisF-C9S was measured to be 441 µM. We chose a ligand with weak affinity as the effects of designs should be obvious: those that improve binding would increase the affinity, and importantly, designs that weaken binding would display a loss of affinity. The placement of the ligand in the binding pocket was optimized over multiple rounds of RosettaLigand docking. In the first round (Figure 4.1C), the ligand was allowed to reorient a full 360° and move up to 5 Å in three-dimensional space. Top models by total and binding energy were docked a second time (Figure 4.1D) with only 45° of reorientation and 1 Å translation. A final round of docking with sequence optimization (Figure 4.1E) was carried out on top models from this second round and minimal ligand rearrangement. RosettaLigand detected amino acids within a 6 Å radius of the ligand and

allowed for Monte Carlo sequence optimization to find side chains that interact more favorably with the docked ligand. This often resulted in removal of charge from the binding pocket and alteration of the hydrophobic packing to enhance shape complementarity with the ligand.

Examination of the top designs showed a great improvement in interface score over the last round of docking by up to 5 REUs (Figure 4.1F). The number of mutations suggested by Rosetta ranged from 4 to 22. The top designs were sorted first by total energy, then the top 20% were arranged by interface score. Lastly, the top 50 models by this search criterion were sorted by number of mutations to select for models with the lowest deviation from the wild-type sequence. A final model containing 8 mutations was selected to explore further.

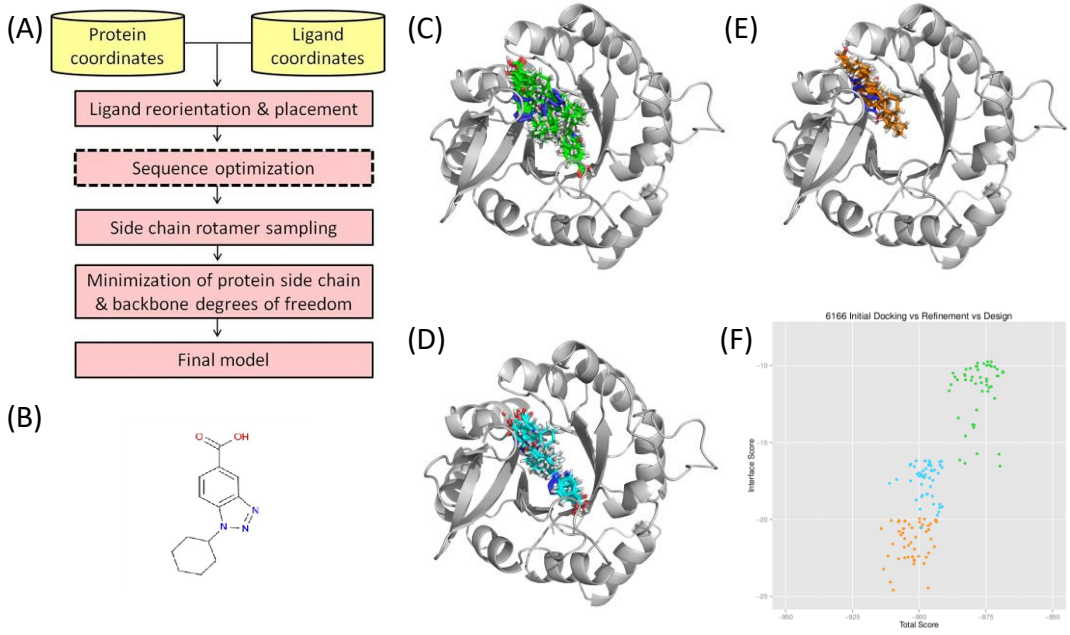


Figure 4.1 Design Strategy for Compound VU0068924. (A) RosettaLigand protocol for simultaneous docking and design. (B) Two dimensional structure of compound VU0068924. (C) Initial docking of VU0068924 with large ligand reorientation. (D) Refinement of the initial docking pose with tighter ligand reorientation restraints. (E) Last round of ligand docking with sequence optimization turned on and minimal ligand rearrangement. (F) Plot of total score versus predicted binding energy over the three rounds of RosettaLigand docking (D-F) color coded according to the corresponding panel.

4.4.2 Rosetta Designs Surface Mutations in addition to Pocket Mutations

The initial design contained the mutations S9Q, L50V, G80A, A128T, D130M, S144W, T171A, and S201H. Upon visual examination, mutations S9Q and S201H were located distal from the binding pocket and were solvent exposed. It was suspected that these mutations were allowed due to the 6 Å radius used for automatic pocket detection and that these mutations wouldn't contribute to ligand binding but likely were identified by Rosetta to improve surface electrostatics. To control for this, we reverted these two residues back to the wild-type sequence *in silico*, minimized the structures, and rescored the complex. As expected, these surface mutations contributed no energy to the interface score and were excluded from further design work. The final design for expression included only L50V, G80A, A128T, D130M, S144W, and T171A (Figure 4.2).

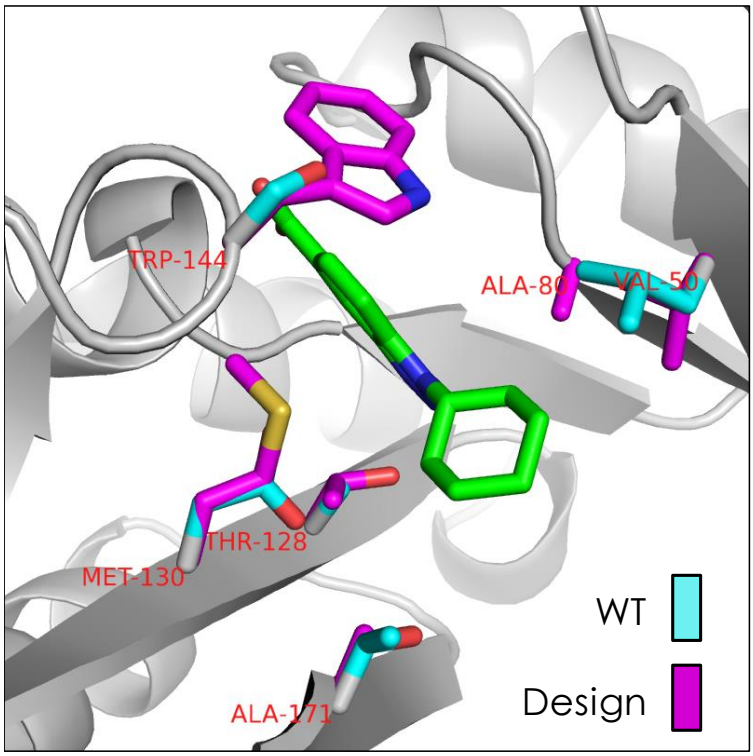


Figure 4.2 The final set of mutations for HisF design. The ligand VU0068924 is shown in green. Wild-type side chains are shown in cyan, and designed mutations are shown in magenta.

4.4.3 Incorporation of Rosetta-predicted Mutations did not Alter the Tertiary Structure of HisF

An important consideration in any design experiment is that inclusion of a predicted mutation may disturb the fold of the protein. To account for this, we expressed HisF in *E. coli* with the six predicted mutations and subjected the protein to biophysical characterization. A HMQC NMR spectrum was collected on uniformly ^{15}N-labeled designed HisF as shown in Figure 4.3A. Overlay of this spectrum with that of WT-HisF showed minimal changes between the two proteins suggesting the fold is similar and that the designed protein can be used for additional NMR experiments. Additionally, crystals of designed HisF were grown and a structure was determined at 2.5 Å using molecular replacement methods in Phenix using the WT-HisF as a search model. The structure of the designed HisF was identical to the WT-HisF structure (PDB ID 5TQL) with an RMSD of 0.4 Å (Figure 4.3B). Further, the rotameric state of the mutated side chains matched with the predicted rotamers in the HisF-VU0068924 model complex (Figure 4.3C).

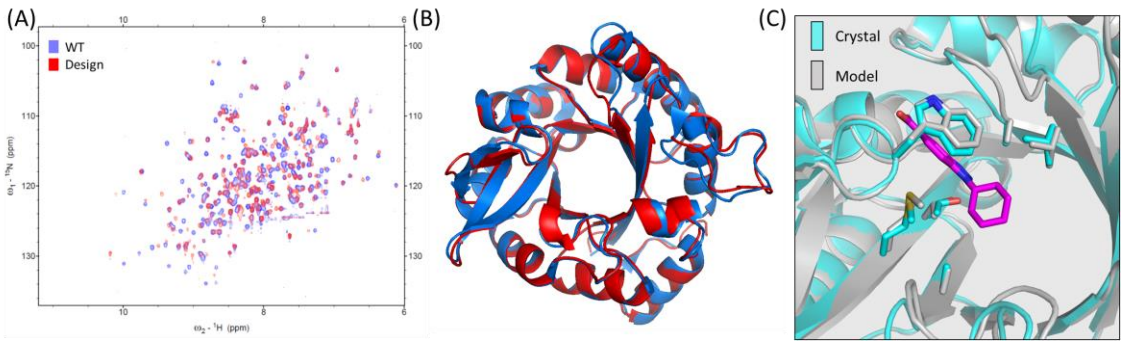


Figure 4.3 Incorporation of Rosetta Designs Does not Affect the Tertiary Structure of HisF. (A) ^{15}N-^{1}N HMQC spectra overlay between WT (blue) and Design (red). (B) Overlay of crystal structures of HisF-WT and Design, colors matching in (A). (C) Overlay of predicted side chain orientations for Rosetta model (grey) compared to the cyrstallized rotamers (cyan). The ligand VU0068924 is shown in magenta from the Rosetta prediction for reference.

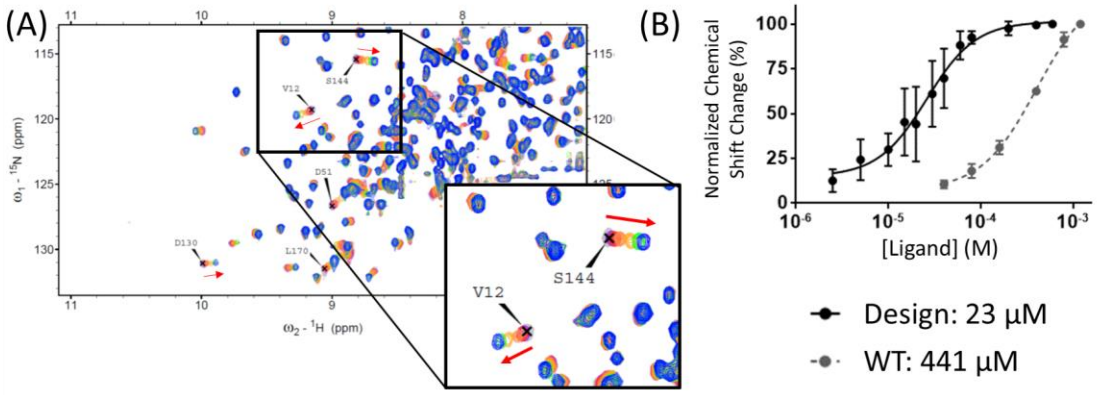


Figure 4.4 NMR Titration of VU0068924 against HisF. (A) Overlay of HSQC spectrum for the WT HisF. Each color represents the chemical shifts at increasing concentration of ligand. Inset, zoom in of select positions as tracked over the titration. (B) Titration curves for the WT (dashed grey) and designed (solid black) HisF with the measured binding affinities listed below.

4.4.4 Design Mutation Greatly Enhances Binding Affinity of Ligand

A titration of VU0068924 against 50 µM of designed HisF was measured via HMQC NMR spectroscopy. Ligand binding is measured by perturbations in the amide backbone chemical shifts (Figure 4.4A). By plotting the magnitude of the chemical shift perturbation as it relates to ligand concentration, a binding affinity can be calculated. As noted previously, titration of VU00068924 against the wild-type HisF sequence suggested a K_D of 442 µM. Incorporation of the six Rosetta mutations enhanced the binding affinity 20-fold to 23 µM (Figure 4.4B). Importantly, the designed HisF exhibits ligand saturation which is lacking in the WT HisF. This allows for accurate calculation of the K_D which is likely unreliable in the WT measurement.

4.4.5 Identification of Critical Mutations that Enhance Binding Affinity

In order to parse out the individual contribution of the designed residues towards the enhanced binding affinity, each position was reverted to the wild-type sequence in the context of the full design and again tested for ligand binding. As seen in Figure 4.5A,

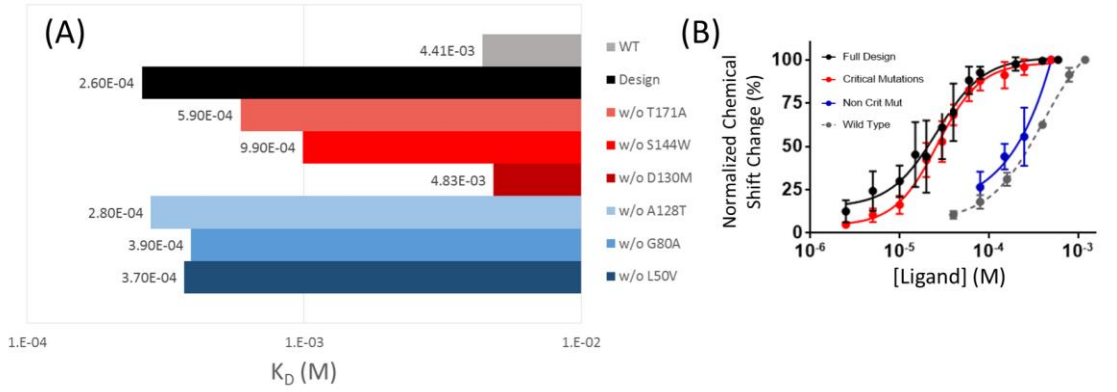


Figure 4.5 Identification of Residues Critical for Enhanced Binding to VU0068924. (A) Measured K_D's for each revertant mutant. The WT is shown in grey and the full design is shown in black. Individual reversions that strongly affected binding affinity are shown in shades of red and mutations that had minimal effect on binding affinity are shown in shades of blue. (B) Binding curves for triple mutants combining either the three mutations that were deemed critical for enhanced binding (red) or non-critical for enhanced binding (blue).

reversions of M130D, W144S, and A171T resulted in a greater than two-fold decrease in binding affinity compared to the Rosetta design. Reversions of V50L, A80G, and T128A affected the binding affinity by less than two-fold. This separated the mutations into two categories, those critical for ligand binding and those not critical. Generation of triple mutants containing only the critical or non-critical mutations (D130M/S144W/T171A and L50V/G80A/A128T, respectively) further confirmed this discrimination (Figure 4.5B). The triple mutant containing the three critical mutations nearly recapitulated the full design K_D with a binding affinity of 30 µM. Conversely, the mutant containing only the non-critical mutations possessed nearly WT affinity towards the ligand. A K_D of 601 µM was calculated for this mutant though the exact value is unreliable due to the lack of ligand saturation.

4.4.6 Pairwise Mutations Identify D130M and S144W as Most Potent for Binding Enhancement

To further deconvolve the contribution of the three critical mutations, they were

each built into the WT sequence individually and in pairs building up to the triple mutant (Table 4.1). Inclusion of either D130M or S144W individually enhanced the binding affinity by a similar but a significantly lesser degree than the full design. Pairwise addition of the mutations resulted in dissimilar effects with the pair S144W/T171A losing binding affinity while the other two combinations D130M/S144W and D130M/T171A further enhancing the binding affinity beyond any individual mutation. Additionally, D130M/S144W possessed nearly full binding affinity compared to the design with 35 μM compared to 23 μM. The results here suggest a rank order of D130M > S144W > T171A in contributing to the enhanced binding affinity of VU0068924.

Table 4.1 Measured KD's for compound VU0068924 for Each Combination of Tested Mutations

Mutation Set	K_D (μM)
WT	441
Full Design (L50V/G80A/A128T/D130M/S144W/T171A)	23
Full Design minus L50V	37
Full Design minus G80A	39
Full Design minus A128T	28
Full Design minus D130M	60
Full Design minus S144W	99
Full Design minus T171A	59
WT + L50V/G80A/A128T	601
WT + D130M/S144W/T171A	30
WT + D130M	109
WT + S144W	103
WT + T171A	378
WT + D130M/S144W	35
WT + D130M/T171A	51
WT + S144W/T171A	430

4.5.7 Rosetta Score Function Does Not Distinguish Between Critical and Non-Critical Mutations

After identifying two mutations that contributed the most to the overall enhanced binding affinity we wanted to compare how effectively Rosetta distinguished these mutations from the non-critical mutations. We altered the sequence of the HisF structure *in silico* to reflect the WT sequence with either set of the three critical (D130M/S144W/T171A) or three non-critical mutations (L50V/G80A/A128T) and resampled the binding pocket energies. While the binding energies for the WT and Full Design were dissimilar, both sets of triple mutants scored equally well (Fig 4.6This is in sharp contrast with the experimental results which showed that the three critical mutations possess nearly full design affinity and the three non-critical mutations behave as WT HisF. Further modeling of each individual mutation singly and in pairwise combination show nearly equal contribution to Rosetta's binding energy independent of the effect on the measured binding affinity (data not shown).

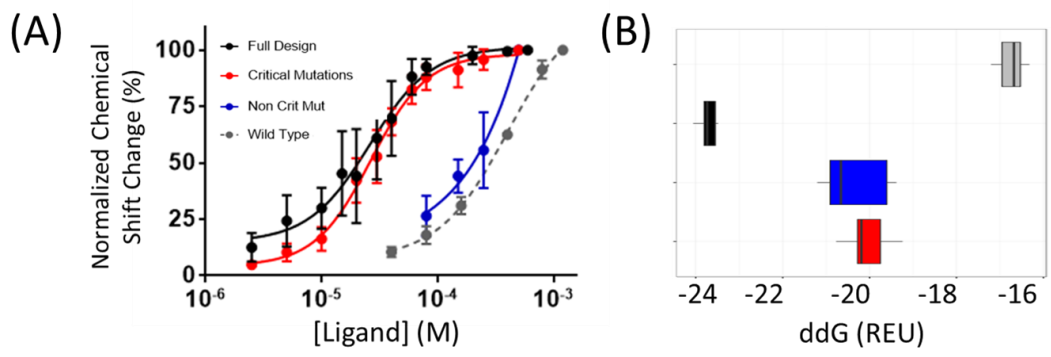


Figure 4.6 Comparison of Measured versus Predicted Binding Affinities for Critical and Non-Critical Mutations. (A) NMR titration curves of WT (grey), Full Design (black), critical mutations (red), and three non-critical mutations (blue) as in Figure 4.5B. (B) Predicted binding affinities (ddG) from Rosetta colored according to A.

4.5 Discussion

4.5.1 RosettaLigand can Blindly Design Protein Ligand Interfaces

Previous ligand interface design in Rosetta stems from enzyme design. In these experiments, ligands need to have a specific orientation with respect to catalytic side chains and the geometries define activity. The protocols EnzDes and RosettaMatch have seen success in designing new interfaces to bind ligands. However, a key consideration in both of these applications is the need to predefine the types of interactions between ligands and protein side chains and the geometries of these various orientations. Further, RosettaMatch takes the predefined ligand pocket and grafts it onto numerous protein scaffolds to find a scaffold that can maintain the correct pocket geometry. However, for the design of protein sequestering agents, drug delivery proteins, or biosensors, the exact orientation of the ligand within the pocket may not be critical. Further, as in the case of biosensors, the functionality of the protein is predefined and swapping of protein scaffolds is not possible. Here, we developed and tested the role of blind interface design with RosettaLigand. We allowed RosettaLigand to both place the ligand into a binding pocket and simultaneously optimize the sequence of the protein to enhance the binding affinity. This is a unique approach to protein-ligand interface design as all it requires is a target protein and target ligand, no other input is needed from the user. This will extend the general applicability of design protocols to non-expert users with unique biological questions. We showed that a protein with minimal affinity to a small molecular weight compound could be redesigned to effectively bind the compound with few mutations and no effect on protein fold. The need for only two mutations to see an order of magnitude increase in binding affinity speaks to the strength of the Rosetta design strategy.

4.5.2 The Effect of Each Mutation is Hard to Predict

The initial design from Rosetta contained 8 mutations. Upon visual examination of the protein binding pocket it was obvious that two of the mutations contributed nothing to the binding interface. In fact, when these mutations were reverted to the WT sequence, no change on predicted binding affinity was observed. These types of mutations will need to be controlled for in future development of the RosettaLigand algorithm. Here we used an automatic pocket detection that identified the pocket as residues within a certain radius from ligand atoms. Any residue that falls within this radius can be mutated in the design algorithm. While this allows for rapid detection of interface residues without needing to specify manually which residues should fall in the binding pocket (a strength for the general applicability of the protocol), it is suggested that an orientation dependent vector be added to pocket detection such that the residues should be pointing towards the ligand. While it has been reported that second shell mutations, or those residues just outside the pocket that stabilize the pocket geometry, will enhance the stability of the pocket resulting in enhanced binding to small molecules, these mutations are hard to predict *a priori*. These mutations are often found in high throughput mutational optimization *in vitro* following an initial *in silico* design hit. While we did not test the binding affinity experimentally of the full eight mutation prediction we feel confident that they would not have further enhanced the binding affinity beyond the six mutations that were detected. This is because the two mutations that contributed nearly the full enhancement of binding affinity to the ligand were making direct contact with the ligand. D130M removed a negative charge and S144W increased the hydrophobicity of the pocket allowing for the small, hydrophobic, uncharged ligand to bind. These two mutations alone generate a protein with binding affinity of 35

μM compared to the full design affinity of 23 μM. This suggests that the remaining four mutations play roles in packing or pocket stabilization and not ligand recognition. The crystal structure of the designed protein revealed that the rotameric state of each designed side chain matched the crystallized state. This suggests that the pocket geometry is predefined and the entropic cost of binding is minimized. Unfortunately, the effect of each mutation *in vitro* is not captured by the Rosetta scoring function. Weighting of the degree of contact or type of interaction with the ligand may help discriminate critical from non-critical mutations in the future.

4.5.3 Avenues for Further Improvement to Affinity

We report here the enhancement of a binding affinity by an order of magnitude for our ligand. However, it should be noted that the maximal binding affinity of the designed protein for the ligand is still μM, which is well below the necessary affinity for an interaction that could be used as a tool compound for *in vitro* experiments and even worse for a potential *in vivo* tool. A likely reasoning for this is the tested compound. We used a small compound with molecular weight below 250 Da possessing minimal binding affinity to the starting protein. This compound belongs to a drug fragment library which is designed to identify compounds with weak binding affinity that can be further optimized to build drug-like compounds with higher affinity. We anticipate that the use of a larger compounds with additional functional moieties would increase the likelihood of engineering stronger interactions. If we can enhance binding affinities by two or three orders of magnitude to a starting compound then we will start to move into tool compound ranges. Further, if we start with a compound that possess single digit μM affinity as opposed to nearly mM affinity, a one or two order magnitude increase in binding affinity

would result in nM affinity. Lastly, it has been shown that *in silico* predictions alone often do not result in the final result for either drug screening or protein-interface design. These predictions serve as starting tools with increased functionality over background that can be further optimized through yeast or phage display. As mentioned, these later rounds of *in vitro* design often identify second shell mutations that are difficult to predict. We suggest that the protocol described above will serve as a tool in a much larger protein design pipeline to develop either novel binding affinities for new compounds or for the enhancement of binding to intrinsically binding compounds.

4.6 Conclusions

These results suggest that Rosetta can readily design novel proteins with unique ligand binding properties. Importantly, these were naively guided predictions, i.e. we didn't specify certain interactions and geometries the ligand must have with the protein, which could allow for high-throughput ligand docking and design. These methods have implications for drug sequestration and delivery, for biomarkers, and for pharmacological intervention of designed receptor proteins (i.e. DREADD technology).

CHAPTER 5

RosettaGPCR: An Improved Method for Multiple Template Homology Modeling of GPCRs with Rosetta

5.1 Summary

This chapter describes my new protocol for GPCR-specific multiple-template homology modeling within Rosetta. Previous single- and multi-template homology modeling protocols in Rosetta often generated non-native-like conformations of transmembrane helices and/or extracellular loops. It was identified that these issues stemmed from sub-par alignments of GPCRs owing to their highly diverged sequences. Generation of a blended sequence- and structure-based alignment method that for the first time accounted for loop structure conservation allowed for the generation of highly accurate GPCR models in a single round of modeling. The concepts developed in this chapter were solely my own. This new method allows for accurate modeling of receptors down to 20% sequence identity which accounts for nearly the entire druggable GPCR family. This chapter is a research article for which I am first author.

5.2 Introduction

5.2.1 G-protein Coupled Receptors Represent Important Therapeutic Targets

G-protein coupled receptors (GPCRs) are the largest family of membrane proteins in the human body. They are important proteins for cellular response to extracellular signals and play roles in immune response, cardiopathies, and neural development. They

are ubiquitous family of proteins evolved over time to respond to a variety of stimuli including ions, small molecules, and larger peptides. Given their role in cellular functioning, they are a highly targeted class of proteins for therapeutic intervention. Current estimates suggest between 30-40% of drugs available worldwide act at a GPCR.

5.2.2 GPCRs are Defined by Their Fold

These receptors have been studies for over half a century with initial studies on whole cell tissue—guinea pig ileum being an early favorite. This work led to the development of antihistamine drugs and beta blockers. Over time, the definition became more formalized. When the β2 adrenergic receptor was first cloned in the 1980s it became apparent that this and all other receptors shared a common seven helix transmembrane spanning domain. It was this moment that first confirmed that these receptors were similar to rhodopsin, a GPCR involved in light recognition, and comprised a larger family of proteins defined by this shared topology. As the human genome was fully sequenced throughout the 1990s it was found that this family consisted of about 800 members in human. Taxonomy of the GPCR superfamily identified 5 distinct families defined by their N-terminal extracellular domain: glutamate, rhodopsin, adhesion, frizzled, and secretin [21]. The receptors could be further divided by subfamily or ligand recognition, but the unifying fold was the seven transmembrane spanning helices indicating the role of shared structure.

5.2.3 Atomic Structures of GPCRs are Limited but Growing

The earliest structural studies were electron crystallography of bacteriorhodopsin which confirmed the bundle of seven transmembrane helices [331]. Nearly twenty years

later, a low resolution structure of rhodopsin was characterized by electron crystallography [332, 333] followed five years later by a high-resolution crystal structure [334]. However, it took a further seven years for a crystal structure of a receptor bound to a diffusible ligand to be determined [24, 335]. Two structures of β2 adrenergic receptor became available within a few months of each other, one with a stabilizing fusion protein inserted in intracellular loop 3 [335] and another that had been thermostabilized through a series of mutations [24]. This delay in time indicated the difficult nature of membrane proteins and in particular GPCRs for high resolution structure determination. Nearly every structure we have today has been altered in some way either by mutation, truncation, chimeric fusion proteins, and/or addition of soluble stabilizing proteins such as nanobodies or mini-G proteins [336]. Despite this we have about 50 structures of unique receptors determined in only the first decade since the crystallization of β2AR. While this is a tremendous achievement, nearly 750 human receptors remain to have structure determined. Even if we remove the odorant receptors as they are unlikely potential drug targets, the remaining 350 structures need to be studied for either an understanding of how current therapies bind their targets or for the development of novel structure-based therapies. At least 100 of these remaining receptors have no known ligand and are classified as orphan receptors. Knowledge of the structural details of the ligand binding pocket could prove helpful in identifying the endogenous ligand.

5.2.4 Computational Modeling can Extend our Current Understanding of GPCR Structure

Given this huge knowledge gap and the shared fold of the diverse receptors, homology modeling is an important tool for generating models of as-of-yet unsolved receptor structures. Homology modeling uses a protein template with a shared topology to

map the target sequence onto its backbone coordinates in a process called threading. Early homology modeling relied entirely on fixed backbone threading but these often failed as sequence identity worsened as sequence differences alter loop orientations and local geometries. Next generation homology modeling incorporated energy minimizations to account for backbone flexibility. Despite this, there was an inherent bias towards the template structure and target models were influenced strongly by the degree of fit between the available template and the target structure. More recently, the use of multiple templates has seen success in modeling targets in which the sequence identity is below 50% to any given template. The idea is that the target protein may have higher local identity to one template in one region of the structure while a second template (or third, or fourth, etc.) has higher local identity in a region where the first template suffers. By combining multiple templates, the ability to generate high quality models down to 40% sequence became routine.

5.2.5 Several GPCR-specific Homology Modeling Protocols Exist

Looking at the success of computational modeling, several programs have been developed specifically for modeling of GPCRs. While single-template homology modeling servers exist (GoMoDo [337]), most servers have moved towards multiple templates because of the low sequence identity shared between various receptors. Use of multiple templates is now routine for GPCR-ModSim [338], GPCR-I-Tasser [339], GPCRM [340], GPCR-SSFE [341], and GPCRdb [342]. While all of the above-mentioned servers utilize Modeller as their base homology modeling software with the exception of GPCR-I-Tasser, the handling of the multiple templates is distinct between them. For example, GPCR-

ModSim and GPCR-SSFE preselect templates for specific regions of the target receptor, GPCRM averages the set of template structures. Additionally, GPCR-ModSim subjects all models to a short molecular dynamics relaxation simulation to allow sampling away from template structures. Further, closure of loops is handled differently according to each server. Again, while most utilize Modeller loop closure, GPCRM utilizes both Modeller and Rosetta loop closing algorithms to identify best case examples. While Rosetta does not contain a specific GPCR modeling platform, it's performance on single-template modeling of GPCRs has been analyzed in the past with some success [257]. In the GPCR Dock 2013 experiment, Rosetta performed best in the structure prediction of the Smoothened receptor ligand binding pose.

5.2.6 Rosetta Hybridizes Multiple Templates

While other methods predefine template segments for various parts of the target model or averages template structures, Rosetta handles multiple templates simultaneously during its modeling process [5]. Rosetta holds all templates in a defined global geometry and randomly swaps parts of each template using Monte Carlo sampling to identify regions from the various templates that best satisfy the local sequence requirements. This template swapping occurs in parallel with peptide fragment swapping as predefined from the PDB based on the target sequence and predicted secondary structure, a hallmark of Rosetta's folding algorithm [189]. This simultaneous sampling of template segments and peptide fragments allows the energy function to best select which parts to keep from the various components based on how well each part improves the overall score of the model. This hybridization of templates has been shown to be successful in CASP experiments

particularly for low template identity targets down to 40%. Below 40% identity, Rosetta is capable of producing accurate models, though it is not known *a priori* if the output models will be reliable [5].

5.2.7 Development of a GPCR-specific Multiple-Template Homology Modeling Protocol in Rosetta

Given the past success of Rosetta in single template homology modeling of GPCRs [257] and the novel strategy of multiple template modeling in the Rosetta framework [5], we sought to develop a protocol specific to GPCRs that utilized the new code. The change from the previous single-template homology modeling to multiple-template modeling was multifaceted and we needed to test each component individually. First, the use of multiple templates as compared to one begs the question of what is the optimal number of templates to use. Previous work in multiple template homology modeling suggested that there is a goldilocks effect in which multiple templates are better than one but too many templates could actually hurt the modeling process [343]. Additionally, as the other servers don't use a peptide fragment library, we wanted to evaluate its influence on moving models away from input templates and towards target structures. Additionally, loop closure is handled simultaneously in Rosetta's multiple-template homology modeling through the use of these peptide fragments. As these loops are defined by their input template, we decided to pay particular attention to the alignment we give to Rosetta in these regions. Sequence alignments of GPCRs are traditionally weak due to the low identity between targets. Groups have tried to overcome this weakness by aligning based on sequence profiles or structure-based alignments. Here we utilized a blended sequence- and structure-based alignment with particular emphasis given to the highly divergent loops. We benchmarked

our methods on 34 available GPCRs covering the four families that have structures available. Additionally, we chose to model all targets with templates below 40% sequence identity, unless otherwise noted, to mimic most likely scenarios when predicting novel target structures. We find that our GPCR-specific Rosetta-based multiple template homology modeling method (RosettaGPCR) is highly accurate due to the curated sequence alignments and peptide fragment utilization. We find, that our new method accurately models rhodopsin family receptors down to a template identity of 20%. Further, this method outperformed other GPCR servers in the prediction for four new GPCR structures. Based on this success, we established a database of all human non-odorant receptor available for use. Altogether, RosettaGPCR is currently the best available method for modeling this pharmacologically important family of proteins.

5.3 Methods

5.3.1 Description of Benchmark Data Set

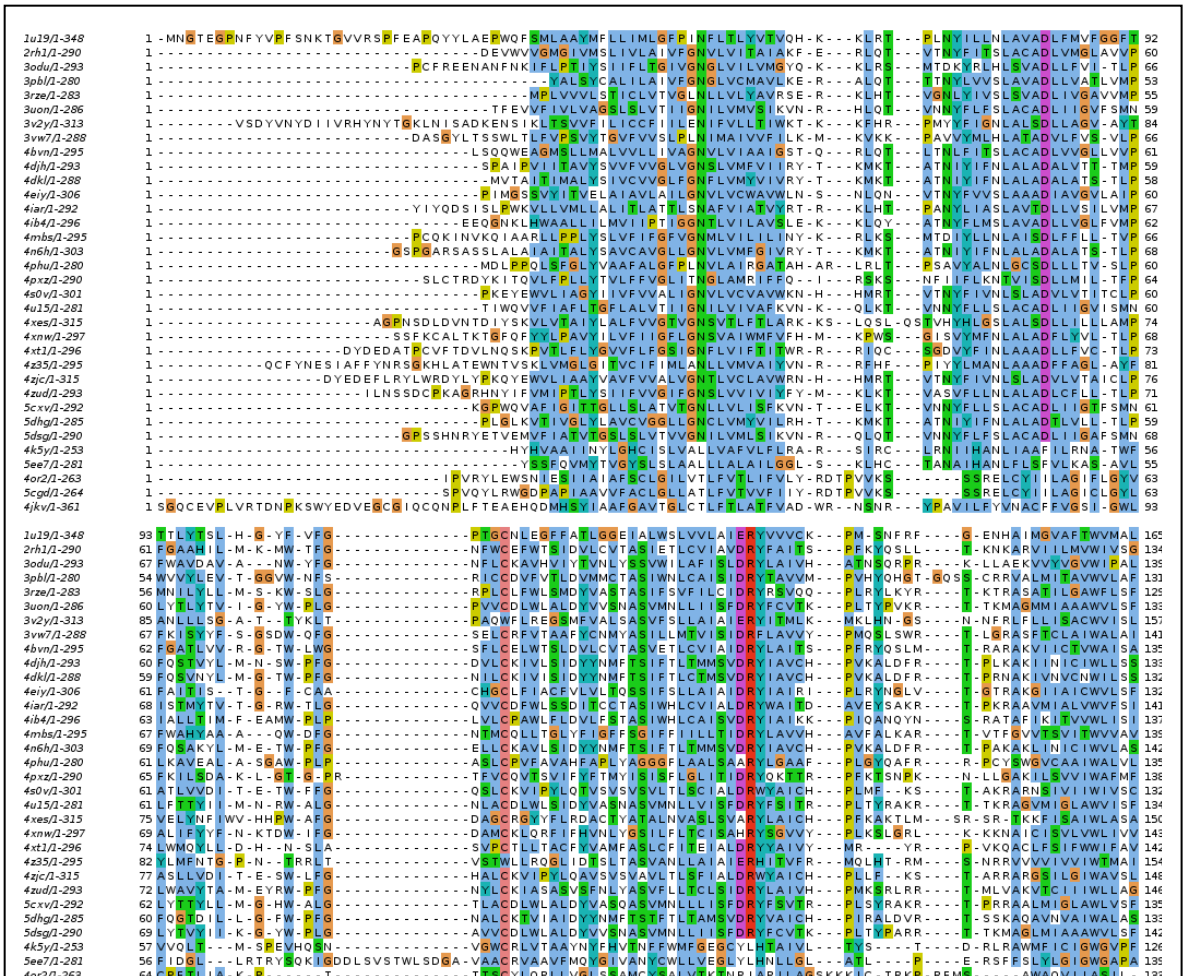
For this study we chose to model 34 crystal structures of GPCRs covering the four structurally characterized families: rhodopsin, secretin, glutamate, and Frizzled. In total there were 29 rhodopsin family members, two glutamate receptors, two secretin receptors, and one Frizzled receptor (Table 5.1). Importantly, we chose to model the receptors with templates below 40% sequence identity as this most closely resembles the majority of cases when modeling GPCRs.

Table 5.1 List of Receptors in Benchmark Set

Receptor	PDB
Rhodopsin	1u19
β1AR	4BVN
β2AR	2RH1
A2AR	4IEY
D3R	3PBL
H1R	3RZE
M1R	5CXV
M2R	3UON
M3R	4UI15
M4R	5DSG
5HT1B	4IAR
5HT2B	4IB4
LPA1	4Z35
S1P1R	3V2Y
NTSR1	4XES
OX1R	4XJC
OX2R	4XJC
δOR	4N6H
κOR	4DJH
μOR	4DKL
NOP	5DHG
CCR5	4MBS
CXCR4	3ODU
US28	4XT1
AT1R	4ZUD
PAR1	3VW7
FFAR1	4PHU
P2Y1	4XNW
P2Y12	4PXZ
CRF1R	4K5Y
GCGR	5EE7
mGluR1	4OR2
mGluR5	5CGD
SMO	4JKV

5.3.2 Generation of Alignments for Modeling

Initial alignments for the benchmark set were obtained from the GPCRdb. This largely ensured that the transmembrane helices were well aligned. To improve on these



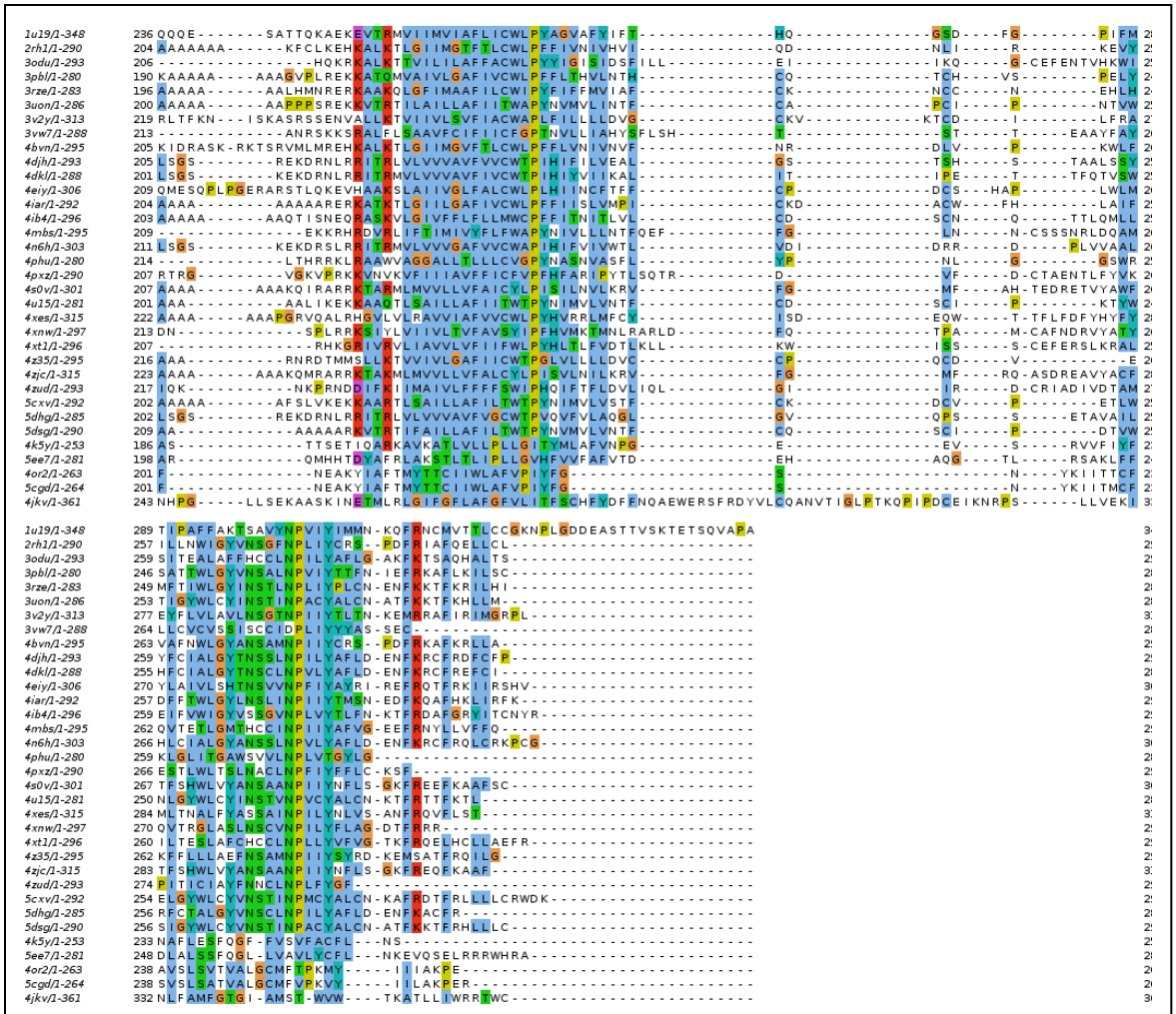


Figure 5.1 Alignment of Receptors in the Benchmark

alignments, the structures were visualized in PyMol and the structural alignments were compared to the sequence alignments. Certain helices needed to be extended or shortened by a turn depending on the resulting structure. Further, loop alignments were generated based on the alignment of vectors of Cα to Cβ atoms. If structures were present in loops such as disulfides, α-helices, and β-sheets, these were preserved in the alignment. Remaining residues that could not be aligned by any of the above metrics were moved to be adjacent to a region of defined secondary structure to ensure proper fitting of peptide fragments between ordered and unordered regions. The alignment of the 35 receptors is shown in Figure 5.1. Additional alignments were generated using the default options of

ClustalOmege, Muscle, T-Coffee TM-PSI and Espresso, and Mustang and used without further modification.

5.3.3 Template Selection

For all receptors, a pairwise identity matrix was generated using ClustalOmega. The reported identities were used to rank the templates for each receptor model. Shown in Table 5.2 is the ranked list of templates for each target receptor. While most templates have sequence identities below 40%, those highlighted in yellow were removed because they were over the 40% threshold. Bolded templates were used in single-template high identity modeling to compare to previous benchmark [257].

5.3.4 Generation of Additional Input Files

Membrane spanning topology files were generated by submitting the sequence of the target proteins to Octopus [344]. The output files were converted into Rosetta readable span files with Rosetta's built in `octopus2span.pl` script.

Disulfide files were prepared for each target protein. Disulfides were mapped for the conserved bond between TM3 and ECL2 except for LPA1 and S1P1. Additional disulfides within ECL3 were mapped as needed.

5.4.4 Model Production

With all input files in hand, target sequences were threaded onto the pre-aligned

Table 5.2 List of Templates for Each Target Ranked by Sequence Identity

Yellow highlighted templates were not used for general modeling because they have sequence identities greater than 40%. Bolded templates were used for single-template high identity modeling to compare to previous benchmark.

Rhodopsin	B1AR	B2AR	A2AR	D3R	H1R	M1R	M2R	M3R	M4R	5HT1B	5HT2B	LPA1	S1P1	NTSR1	OX1R	OX2R
3pbl	**2rh1**	**4bvn**	3rze	4iar	5cxv	4u15	**5dsg**	**5cxv**	3uon	4bvn	3pbl	3v2y	4z35	4djh	4s0v	4zjc
3uon	4iar	4iar	4bvn	4bvn	4u15	5dsg	4u15	5dsg	4u15	3pbl	2rh1	3pbl	4iar	5dhg	5dhg	4dkl
4iar	3pbl	3pbl	2rh1	2rh1	5dsg	3uon	5cxv	3uon	5cxv	3rze	4bvn	4iar	3rze	3pbl	4djh	5dhg
4dkl	3rze	4u15	4iar	5dsg	4iar	3rze	3rze	3rze	3rze	4u15	4iar	5dsg	4eiy	4zud	4eiy	4djh
5cxv	4ib4	3rze	3pbl	5cxv	3uon	4iar	4iar	4iar	3pbl	2rh1	4u15	4u15	3pbl	4zjc	4dkl	4n6h
5dhg	4u15	4ib4	4u15	4ib4	4bvn	3pbl	3pbl	2rh1	4iar	5cxv	3rze	5cxv	3uon	4xnw	4bvn	2rh1
4xnw	5cxv	5cxv	4zjc	3rze	2rh1	2rh1	4bvn	4bvn	2rh1	3uon	5cxv	3uon	4bvn	4n6h	3rze	3pbl
4mbs	4eiy	4eiy	4ib4	3uon	4eiy	4bvn	2rh1	3pbl	4djh	5dsg	5dsg	4eiy	4ib4	4iar	4n6h	4bvn
4u15	3uon	5dsg	4s0v	4u15	3pbl	4ib4	4ib4	4ib4	4bvn	4ib4	3uon	4bvn	4u15	3odu	2rh1	4eiy
2rh1	5dsg	3uon	3v2y	4eiy	4djh	4eiy	4djh	4eiy	4ib4	3v2y	4eiy	3rze	5cxv	4ib4	3pbl	4xnw
3odu	4s0v	4s0v	5cxv	4s0v	4ib4	4zud	4zud	4zud	5dhg	4eiy	5dhg	5dhg	5dsg	3uon	4iar	4iar
5dsg	4zjc	4zjc	5dsg	4z35	4zud	4xt1	5dhg	4xt1	4zud	4djh	4djh	4s0v	2rh1	4bvn	4xes	4u15
4ib4	3v2y	4n6h	3uon	4dkl	3v2y	4djh	4dkl	4s0v	4dkl	4n6h	4dkl	4zjc	4s0v	4u15	4ib4	3odu
4n6h	4n6h	4dkl	4z35	5dhg	4dkl	3v2y	3v2y	3v2y	4n6h	4s0v	4zud	2rh1	4zud	4dkl	4xnw	4zud
4xt1	5dhg	5dhg	4zud	3v2y	4n6h	4z35	4n6h	4djh	4eiy	4dkl	3v2y	4n6h	4zjc	4s0v	3w7	4ib4
3rze	4djh	4djh	4djh	4n6h	5dhg	5dhg	4eiy	4z35	4z35	4z35	4n6h	4ib4	4xnw	5dsg	4u15	3uon
4eiy	4z35	4zud	4dkl	4zjc	4zjc	4s0v	4z35	4dkl	3v2y	4zjc	4s0v	3w7	4xes	4xt1	5cxv	4xt1
4bvn	4dkl	4mbs	4n6h	4xes	4z35	4mbs	4s0v	4n6h	4xt1	5dhg	4zjc	4zud	4mbs	5cxv	3odu	3rze
4zud	4zud	3v2y	5dhg	1u19	4s0v	4n6h	4xt1	5dhg	4s0v	4xt1	3odu	4dkl	3odu	4mbs	3uon	5cxv
3v2y	4mbs	3odu	4xes	3odu	3odu	4dkl	1u19	3odu	3odu	3odu	4xt1	4djh	4xt1	4eiy	4zud	4z35
4xes	4xt1	4z35	4xnw	4djh	4mbs	4zjc	4zjc	4mbs	4mbs	4xes	4xes	4xnw	1u19	3v2y	4z35	3v2y
4djh	4xes	4xt1	3w7	4zud	3w7	4xes	3odu	4zjc	4zjc	4zud	4xnw	4xt1	4n6h	2rh1	5dsg	5dsg
4s0v	3odu	1u19	4mbs	4xt1	4xt1	3odu	4xes	4xes	4xes	4xnw	4phu	4pxz	4djh	3w7	3v2y	4xes
4pxz	4pxz	4xnw	1u19	4xnw	1u19	4xnw	3w7	4xnw	4xnw	1u19	3w7	1u19	4dkl	3rze	4xt1	3w7
3w7	1u19	4xes	3odu	4mbs	4xnw	1u19	4pxz	4pxz	4pxz	4mbs	4pxz	3odu	5dhg	1u19	4mbs	4mbs
4z35	4xnw	3w7	4xt1	4pxz	4xes	4pxz	4mbs	1u19	1u19	4pxz	4mbs	4mbs	4pxz	4pxz	4phu	4pxz
4phu	4phu	4pxz	4pxz	3w7	4pxz	3w7	4xnw	3w7	3w7	3w7	1u19	4xes	3w7	4phu	4k5y	4phu
4zjc	3w7	4phu	4phu	4phu	4phu	4phu	4phu	4jkv	4phu	5ee7	4z35	4or2	4or2	4z35	4pxz	1u19
5ee7	4k5y	4or2	5cgd	5ee7	4k5y	5ee7	4jkv	4phu	5ee7	4phu	4k5y	5ee7	4phu	4or2	5cgd	4k5y
4or2	5ee7	5cgd	4or2	4jkv	4jkv	4jkv	4k5y	4k5y	4jkv	5cgd	4or2	4phu	5ee7	5ee7	1u19	5cgd
5cgd	4or2	5ee7	4k5y	4or2	5ee7	4k5y	5ee7	5ee7	4k5y	4or2	5cgd	5cgd	5cgd	4k5y	5ee7	4or2
4k5y	4jkv	4k5y	5ee7	4k5y	5cgd	5cgd	5cgd	5cgd	5cgd	4jkv	5ee7	4k5y	4jkv	5cgd	4or2	5ee7
4jkv	5cgd	4jkv	4jkv	5cgd	4or2	4or2	4or2	4or2	4or2	4k5y	4jkv	4jkv	4k5y	4jkv	4jkv	4jkv

dOR	kOR	mOR	NOP	CCR5	CXCR4	US28	PAR1	FFAR1	P2Y1	P2Y12	AT1R	CRF1R	GCGR	mGluR1	mGluR5	SMO
4dkl	**4n6h**	**4n6h**	**4djh**	4zud	4zud	3odu	4xnw	3w7	3w7	4zud	3odu	5ee7	4k5y	**5cgd**	**4or2**	4k5y
4djh	4dkl	4djh	4n6h	3odu	4mbs	4mbs	5dhg	4zud	4djh	4mbs	4djh	4jkv	4jkv	4zud	4zud	5ee7
5dhg	5dhg	5dhg	4dkl	4xt1	4xt1	4zud	4dkl	3odu	4dkl	4djh	4dkl	4zjc	4xt1	4mbs	4zjc	4u15
4zud	4zud	4zud	4zud	4djh	5dhg	5dhg	4zud	4ib4	4n6h	4dkl	4mbs	4s0v	4n6h	4eiy	4eiy	5cxv
4s0v	3rze	4s0v	4s0v	5dhg	4dkl	4n6h	4n6h	4mbs	5dhg	3odu	4n6h	4xt1	5dhg	3odu	4dkl	3w7
4xnw	4s0v	3odu	3odu	4pxz	4djh	4dkl	4djh	4n6h	4zud	5dhg	5dhg	4n6h	5cxv	3v2y	4mbs	3uon
4iar	4xnw	4xnw	4zjc	4dkl	4n6h	4u15	4phu	5dhg	4s0v	4n6h	4xt1	5dhg	4zjc	4xes	4s0v	3pbl
5dsg	5dsg	4zjc	4ib4	4n6h	4s0v	4djh	4xt1	4xt1	4xt1	4xt1	4pxz	4djh	3w7	4s0v	5dhg	4bvn
3rze	4zjc	3rze	5dsg	4xnw	4pxz	4xnw	3odu	4pxz	3odu	4xnw	3rze	4dkl	4iar	4ib4	3odu	3rze
3w7	4mbs	5dsg	4mbs	5cxv	4xnw	5cxv	4zjc	4s0v	4mbs	4bvn	4u15	3w7	4dkl	4n6h	4xt1	5dsg
4zjc	3uon	3w7	4xnw	2rh1	3pbl	3uon	4mbs	4zjc	4xes	3uon	5dsg	4zud	4phu	4bvn	4djh	4iar
3odu	4iar	4ib4	3w7	4u15	4ib4	3w7	3uon	4dkl	4zjc	3pbl	4xnw	4ib4	4xes	4xt1	4n6h	4zud
3pbl	3odu	4mbs	3rze	3rze	4u15	4s0v	4s0v	4xnw	4pxz	4ib4	3uon	4bvn	4zud	4zjc	4phu	3odu
2rh1	4xes	3pbl	3uon	4bvn	3rze	5dsg	3pbl	4bvn	4ib4	5dsg	3w7	4eiy	1u19	1u19	4xnw	4xt1
4xt1	4ib4	4iar	3pbl	3w7	3w7	4pxz	3rze	3pbl	3pbl	4u15	4ib4	3rze	4bvn	4z35	4ib4	1u19
4bvn	4pxz	3uon	4xt1	5dsg	4iar	4iar	4ib4	5dsg	4iar	3w7	4s0v	4u15	4s0v	4xnw	4xes	4eiy
3uon	3w7	4xt1	4xes	3pbl	4xes	4ib4	4pxz	4eiy	4u15	4s0v	5cxv	4xes	4u15	3w7	3w7	4xnw
4mbs	4u15	2rh1	2rh1	3uon	2rh1	4bvn	5dsg	4djh	5dsg	5cxv	2rh1	5cgd	3v2y	4phu	4k5y	4pxz
4ib4	4xt1	4pxz	4iar	4phu	4zjc	3pbl	4eiy	3uon	3v2y	4iar	4xes	4mbs	3pbl	2rh1	1u19	2rh1
4u15	2rh1	4u15	4bvn	4iar	5dsg	4xes	4u15	1u19	5cxv	4phu	4eiy	5cxv	5dsg	4dkl	4iar	4mbs
4pxz	5cxv	4bvn	4z35	4xes	3uon	4phu	4xes	5cxv	1u19	2rh1	4bvn	4xnw	4z35	5dhg	3v2y	4ib4
5cxv	4bvn	4eiy	5cxv	4s0v	4bvn	2rh1	2rh1	2rh1	3uon	3rze	3pbl	4pxz	3odu	4iar	2rh1	4phu
4xes	3pbl	5cxv	4u15	4ib4	4phu	3rze	4z35	4u15	4eiy	4z35	4iar	1u19	4eiy	3pbl	4bvn	4n6h
4eiy	4eiy	4xes	4pxz	1u19	5cxv	1u19	5cxv	4xes	2rh1	4zjc	4phu	4iar	4pxz	4djh	4z35	4djh
4phu	1u19	1u19	4eiy	4eiy	1u19	3v2y	4bvn	3rze	3rze	3v2y	4zjc	5dsg	4djh	4k5y	5cxv	4xes
4z35	3v2y	4z35	1u19	3v2y	3v2y	4zjc	1u19	5ee7	4bvn	4eiy	3v2y	2rh1	2rh1	5cxv	3rze	4z35
1u19	4z35	4phu	4phu	4zjc	4eiy	5ee7	4iar	3v2y	4phu	4xes	4z35	3odu	4xnw	3rze	4u15	5dhg
3v2y	4phu	3v2y	3v2y	4z35	4z35	4z35	3v2y	4iar	4z35	1u19	1u19	3pbl	3rze	4pxz	5dsg	4zjc
4k5y	4k5y	4k5y	4k5y	4or2	4or2	4eiy	5ee7	5cgd	5cgd	5ee7	5cgd	3uon	4ib4	5dsg	3pbl	4or2
5ee7	5cgd	5cgd	5ee7	5cgd	5cgd	4k5y	4k5y	4z35	4or2	4k5y	4or2	4z35	3uon	4u15	3uon	4dkl

templates using Rosetta's `partial_thread` application. Threaded pdbs were passed to the `hybridization` application via use of Rosetta XML scripts [5, 225]. Either 100 models or 1000 models were generated per run as noted in the text.

5.4 Results

5.4.1 Blended Sequence- and Structure-Based Alignment Is Critical for Modeling Success

Inherent to any homology modeling protocol is an alignment between the sequence of the target protein and the template structure. As different families of GPCRs share such low identity with one another, sequence alignment is not trivial for this class of proteins. The best-known alignment method for GPCRs is Ballesteros-Weinstein numbering which identifies the most conserved residue in each helix to set as a starting position to count along the helix in reference to this residue. While highly useful, this alignment falls short in two areas. As more receptor structures became available, it was found that not all receptor families adhere strictly to the *i* to *i+4* periodicity in every helix. Insertions and deletions have resulted in alterations of this helicity, and subsequently the BW numbering,

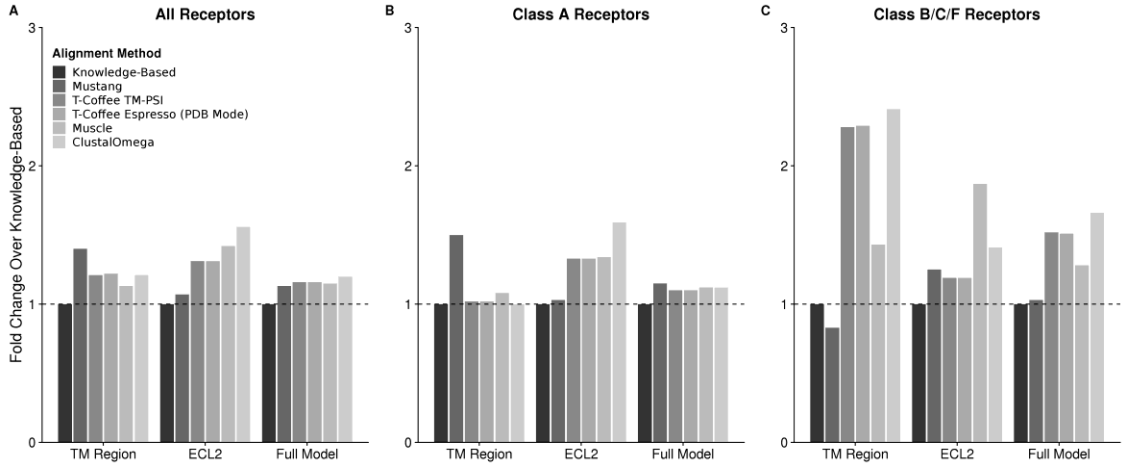


Figure 5.2 Comparison of Average RMSD Change using Various Alignment Methods.

of certain subfamilies of proteins. Secondly, BW numbering falls apart in the loop regions as different receptors have varying length helices and loops and dramatically different loop structures with ECL2 adopting disordered loops, alpha helices, and beta sheets depending on the ligand type. Further, as these loops are critical for ligand recognition, they suffer extremely low sequence identity proving even more difficult for alignment generation. Several groups have implemented various methods for improving alignments for this family of proteins due to this problem. As the purpose of a threading algorithm is to map the target sequence onto the structure of a template protein, it seems prudent to incorporate structure into the sequence alignment for this method to yield accurate results. While this has been reported for the TM region, little attention has been paid to the loop regions. Therefore, a critical component to this method has been the blending of sequence and structure into an optimized alignment for GPCRs. We sought to compare this new alignment to other well-known sequence- or structure-based alignment methods. For each receptor in the benchmark, 100 models were generated for each of the six alignment methods tested. The average RMSD for a target protein was divided by the average RMSD for the same target using the new alignment resulting in a fold change and the average across the full benchmark is reported (Figure 5.2). As seen, despite using sequence-only (ClustalOmega and Muscle), structure-only (Mustang), or blended (T-Coffee TM-PSI and Espresso (PDB Mode)), the curated alignment performs the best in all regions tested. Additionally, for Class A receptors it is found that the TM region is modeled nearly equivalently across all methods with the most improvement coming from improvements in ECL2 modeling. For Classes B/C/F there is a large improvement in modeling of all metrics, except for Mustang alignments of the TM region, the only metric and Class in which Mustang outperforms our alignment. Importantly, for all Classes, the accuracy of

ECL2 is strongly improved suggesting this curated alignment critical for modeling of this region.

5.4.2 Peptide Fragment Hybridization Improves Target Model Quality

Our previous benchmark of GPCR modeling relied on single-template threading. We wanted to recapitulate this initial study using the Hybridize code to allow for peptide fragment insertion but not template swapping to see what effect peptide sampling alone had on output quality. This benchmark dataset was limited to the 8 receptors with high identity templates available in 2013 (β1AR, β2AR, M2R, M3R, δOR, κOR, μOR, and NOPFQ). In this experiment, each target was modeled on either the best available template with either greater than 40% or less than 40% identity and allowed to hybridize with the peptide fragments (Table 2). As seen in Figure 5.3, using the exact same template as was used in the previous threading-alone method, hybridization can substantially improve output model accuracy in all measured regions. The transmembrane (TM) region improves

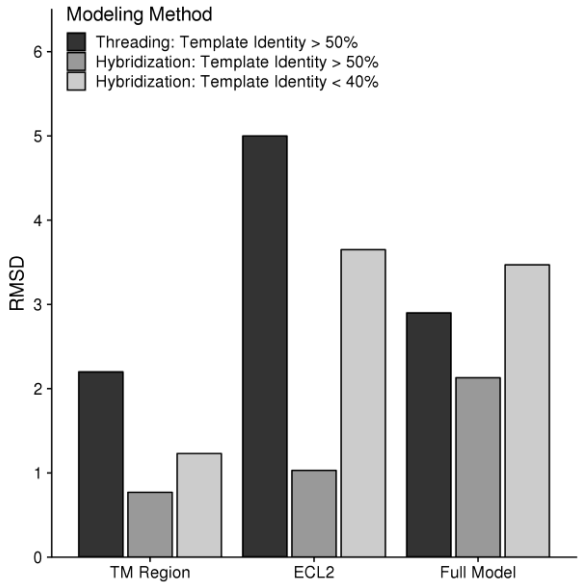


Figure 5.3 Comparison of Single Template Modeling Methods with Peptide Insertion.

on average by over an Angstrom to 0.77 Å RMSD to the crystal structure showing highly accurate modeling of this region. The extracellular loop 2 (ECL2) which has constantly proven difficult to model due to its long length, also showed a dramatic improvement with an average RMSD to the crystal structure of 1.03 Å compared to the previous method with a reported average RMSD of 5 Å. The full model RMSD, which accounts for all remaining loops and flexible termini, also showed modest improvement from 2.9 to 2.13 Å. These results were similar when using a single template with sequence identity less than 40%. Both the TM region and ECL2 improved by at least 1 Å while the Full Model RMSD actually worsened by 0.5 Å. Taken together, peptide insertion accounts for a substantial improvement over threading alone even when templates of poor sequence identity are used.

5.4.3 Multiple Templates Improves Performance with Low Sequence Identity Templates

While peptide insertion helped improved accuracy in the TM and ECL2 regions, overall model accuracy weakened when using a single template with sequence identity less than 40% to the target model. Therefore, we expected that multiple templates could overcome the shortcomings of any single template when using such a poor-quality starting template. We generated 1000 models for every receptor using either the single best template less than 40% identity or the ten best available templates under 40% identity and compared the average RMSD of the resulting models (Figure 5.4). As expected, the average RMSDs improved for almost all receptors in the ECL2 and Full Model criteria. The TM region was rather insensitive to the increase in template availability showing on average only 0.05 Å improvement for the whole set. This is likely due to the high degree of structural similarity of the fold of inactive GPCRs. A few exceptions to the overall trends deserve mention. In the TM region, both Class C receptors perform extremely poorly in

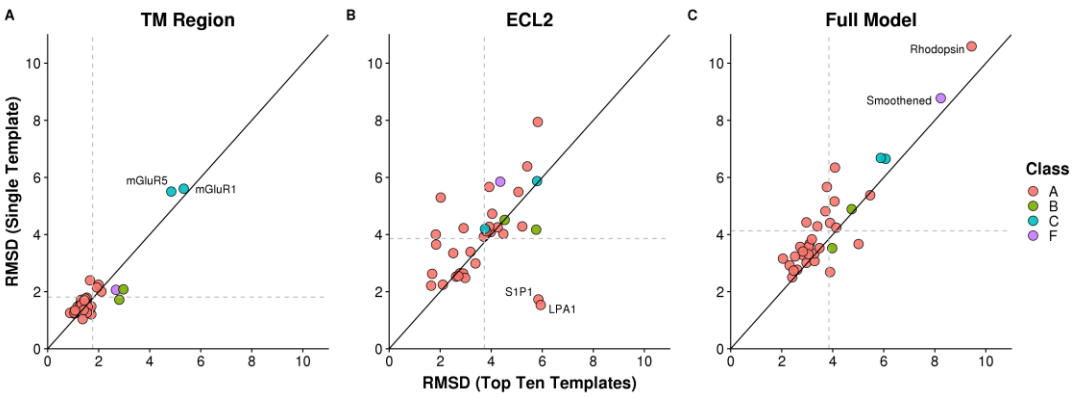


Figure 5.4 Comparison of Average RMSDs for Single versus Multiple Template Homology Modeling.

either set of templates. This is due to the fact that the 40% threshold for selecting templates removed the other Class C receptor from the template pool such that they were modeled with non-Class C templates. As the structure of the TM region is so distinct for these proteins compared to the other classes, the error was expected to be high. For the Class B receptors, the two structures have a sequence identity of 35 with respect to one another allowing these templates to be included with modeling the other. Therefore, the single template TM RMSD outperforms the ten template TM RMSD nearly half an Å. In ECL2, there are two class A receptors (S1P1 and LPA1) that perform extremely well when using a single template as compared to ten templates. These are the only two receptors in the benchmark that do not contain the conserved disulfide between ECL2 and TM3. Their loop structures are quite distinct from all other receptors and as a result, loop modeling only performs well when using the other as a template. In the Full Model RMSD, both Rhodopsin and the Smoothened receptor perform extremely poorly regardless of the modeling method used. This is because both have extremely long and unusual loops and termini. Of note, the TM regions of these two receptors are extremely accurate with 1.54 Å and 2.68 Å to the crystal structure of 1U19 and 4JKV, respectively. Additionally, only two Class A receptors perform worse in the Full Model RMSD calculation when using multiple

template. These again are S1P1 and LPA1 which performed poorly in the ECL2 modeling. It appears that the poor quality of ECL2 is reflected in the Full Model RMSD as the difference in the TM region for these two structures is only 0.1-0.2 Å.

5.4.4 Identification of the Optimal Number of Templates

As reported previously for multiple template homology modeling, the use of multiple template, while improving over a single template, will weaken model accuracy if too many templates are used. We wanted to see if we could identify an optimal number of templates for GPCR modeling using our method. Therefore, we generated an additional 1000 models for each receptor using either five or all available templates and compared the data with the previous data on one and ten templates (Figure 5.5). For both the TM region and ECL2, using all available templates was worse than any other set of templates while the average RMSD were quite similar for one, five, and ten. However, for the Full Model accuracy, using a single template was worse than all over template sets including using all

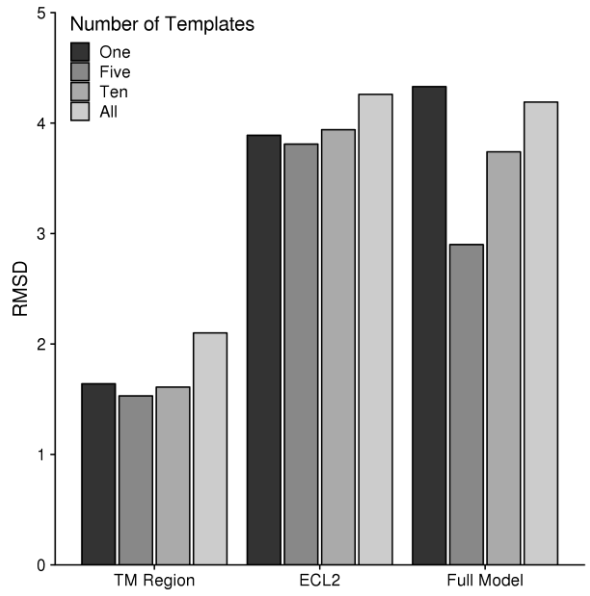


Figure 5.5 Comparison of Model Accuracy using Various Numbers of Starting Templates

available templates. For this criterion, using five templates showed a significant increase in accuracy over all other template sets. Therefore, while using five templates shows a modest improvement over using one or ten templates in the TM Region and ECL2, the overall model accuracy increase with five templates suggests this is the best number of templates for modeling GPCRs in our method.

5.4.5 RosettaGPCR Outperforms other GPCR Modeling Servers

We next sought to compare our method against other GPCR modeling servers. Three servers with publicly available datasets of GPCR models: GPCRdb, GPCR-I-Tasser, and GPCR-SSFE. Additionally, we utilized the GPCR-ModSim server which runs Modeller under the hood as it is very user-friendly and fast. We identified four human GPCR structures that have been released since the beginning of 2018 as we expected there might be a lag in updating databases with the most recent template structures. The four structures were C5aR1, Y1R, PTAFR, and D2R (PDB IDs 6C1R, 5ZBQ, 5ZKQ, and 6CM4, respectively). As it turned out, GPCR-ModSim, which is the only on-demand server we tested, had already been updated to include the structures of PTAFR and D2R, and GPCR-SSFE already had the structure of D2R. Therefore, we could not compare those

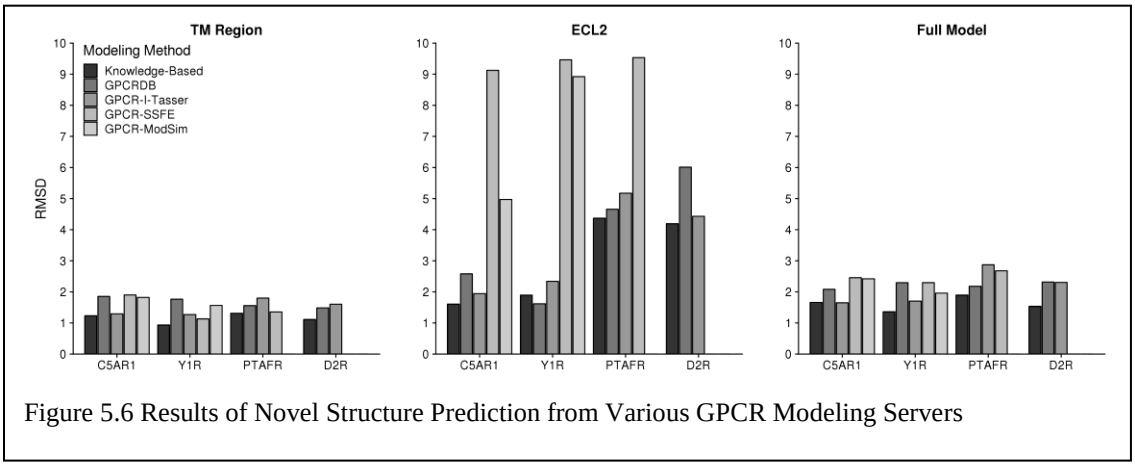


Figure 5.6 Results of Novel Structure Prediction from Various GPCR Modeling Servers

servers for those receptors. We generated 100 models of each receptor target using five template structures and selected the best model by total energy. In comparing our model with the models generated by other servers, we find that RosettaGPCR consistently outperforms the other models (Figure 5.6). Only two servers performed better than RosettaGPCR on two separate measurements. GPCRdb had a better ECL2 of Y1R compared to ours with RMSDs of 1.62 Å versus 1.90 Å, respectively, and GPCR-I-Tasser had a slightly better TM Region model of C5aR1 compared to ours with an RMSD of 1.65 Å versus 1.66 Å, respectively.

5.4.6 Accuracy of Models with Increasingly Worse Templates

In our previous work on GPCR modeling using single template threading, it was found that templates needed to be greater than 50% identity for decent models. Subsequently, the use of multiple templates modeling in Rosetta was suggested to be accurate down to 40% identity. However, with this current benchmark in which we do not use templates above 40% identity, we find that we still produce highly accurate receptor models. Therefore, we wanted to identify what the new threshold for template accuracy was for accurate output models. We devised an experiment where we binned available

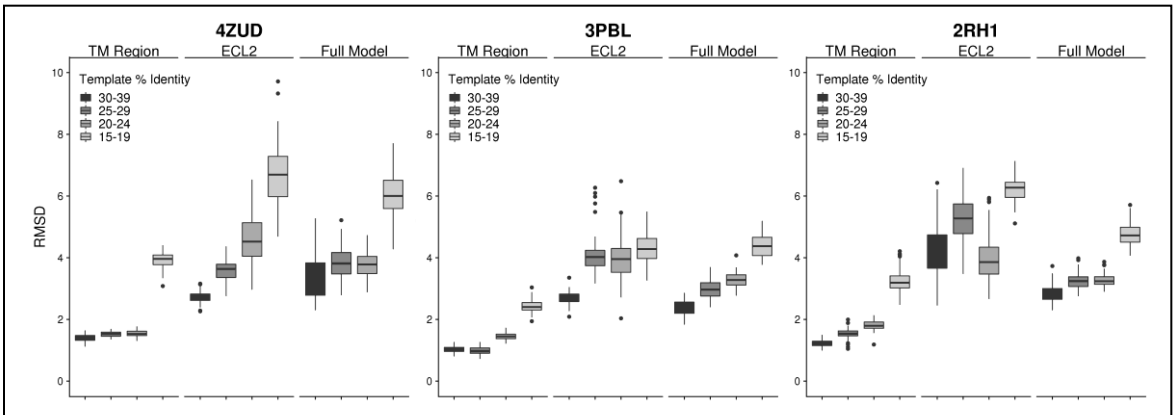


Figure 5.7 Model Accuracy with Templates over Multiple Sequence Identity Ranges

templates into identity groups of 15-19%, 20-24%, 25-29%, and 30-39%. Binning of 30-39% resulted in only a single receptor with multiple templates in both ranges. We then identified three receptors with at least five templates in each identity group and performed multiple-template homology modeling with each set of five templates. The results, shown in Figure 5.6, find that overall the TM Region accuracy is unaffected by the use of templates down to 20%. The same trend held true for the Full Model RMSDs. However, ECL2 was the most strongly affected region with accuracy drop-offs from the first lowering of template identity. This is perhaps to be expected as the structure of ECLs is often influenced by ligand identity (i.e. small molecule vs peptide) and lowering the template sequence identity often changes the ligand selectivity of the templates. Taken together, we suggest that templates down to 20% overall yield accurate models.

5.4.7 Development of Database for All Human Non-Odorant GPCRs

By effectively pushing the lower threshold to 20% sequence identity, we wanted to see what that meant for the remaining druggable GPCRs. To this effect, we identified the best templates by identity for the entire set of non-odorant human GPCRs (Figure 5.8). Out of 397 receptors, 54 have at least one structure determined (14% of the receptor family). This provides 81 receptors with a template with sequence identity above 40%, the previous threshold for accurate modeling. However, the number of receptors with a template between 20 and 40% is 214. Only 48 receptors (or 12% of the receptor family) remain in which the template is less than 20% identity and likely to have significant errors in the modeling process. Due to the likelihood of modeling accurate structures over the majority of the GPCR family we decided generate a model database of the remaining GPCRs using this new method. This is, to the best of our knowledge, the only Rosetta-based GPCR

Figure 5.8 Percent Identify of Best Available Template for Every Non-Odorant Human GPCR

server available which distinguishes it from the many Modeller-based servers. Currently all structures are in the inactive state though work is on-going to provide additional intermediate and active state models.

5.5 Discussion

5.5.1 Blending of Sequence- and Structure-Based Alignment are Critical for Low Identity Template Based Modeling

Inherent to any homology modeling protocol is an alignment between the target sequence and the template sequence. This alignment maps the target sequence onto the template structure in a process called threading. Sequence alignments are necessary for so many biological processes and a wide variety of search methods have been generated. Each sequence alignment method uses a different algorithm to weight the importance of

sequence conservation globally or locally with emphasis or not on gap penalties. Some methods rely on secondary structure propensity and others have been developed specifically for multi-pass membrane proteins. As we learn more about the structures of diverse proteins, it becomes apparent that structure is often more well conserved than sequence. As such, additional algorithms have been generated based on structural alignments and domain fold recognition. This latter case is more inherent to the family of GPCRs in which the common sequence identity between receptors is around 30% while all receptors share a similar structural domain. Therefore, any method for aligning GPCRs for the purpose of homology modeling must upweight structural alignments over sequence. That is not to say GPCRs lack critical sequence motifs. The NPxxY and DRY motifs as well as numerous proline residues and disulfides are critical for receptor function and should be maintained in any sequence alignment. Of all the alignment methods available, a blended sequence- and structure-based alignment is the best for modeling of structures that have such a low pairwise identity with any given template.

The best-known GPCR sequence alignment is the Ballesteros-Weinstein system which attempts to account for this mix of sequence and structure by identifying only a single residue in each transmembrane helix that is conserved and allowing the remaining numbering to be based on the position in the helix with respect to this one residue. While this numbering system is still relevant, it has been updated in recent years as GPCR structures became available. It was found that the periodicity in certain parts of the helices within different subfamilies could be altered and these gaps and insertions needed to be accounted for in order to ensure equivalent residue positions along the face of a given helix. This updated blending of sequence- and structure-based alignment by the GPCRdb is

still one of the best alignments for GPCRs available.

One region in alignment methods that is often overlooked are the loops. These loops are highly diverse in sequence as they have evolved over time to recognize many different ligands. Additionally, there is a variety of structures these loops display. As more GPCR structures become available, patterns are emerging for these loop regions. ECL1 is considered to be highly conserved in sequence and structure and modifications to either will affect receptor trafficking or signaling. ICL2 appears to adopt signaling-state dependent secondary structure with a transition from disorder to order upon receptor activation. ECL3 is home to a variety of disulfide bonds either within itself or connect to extended N-termini which restricts its conformational sampling. Lastly, ECL2 has been studied multiple times with debating theories. It appears to be accepted that ECL2 adopts a beta hairpin structure in all peptide and protein binding receptors. This feature is conserved in some lipid receptors while others have a non-structured loop that serves as a cap to fully encase the hydrophobic ligands. Aminergic ECL2 can adopt both disordered loops or form an alpha helix. A recent paper suggested that the structure of ECL2 is evolutionarily defined with beta hairpin being the original structure that eventually evolved to more varied structures as the ligand diversification increased.

As there is a degree of structural conservation, we focused on aligning the loops structurally. This also affected how we modeled the loops. By providing a structure-based alignment of the loops in the overall alignment, we could allow Rosetta to build the loops simultaneously during the receptor modeling step. This prevented the need for a secondary modeling step. Additionally, because these alignments were based on structure, we were confident that the resulting models would be structurally realistic. In traditional loop

closure methods, the peptide chain grows to join the two cut points through 3D space. For short loops, there can be dense sampling to ensure the native conformation is sampled. However, as the loop length increase to 10 or more residues, the search space is expansive and identifying native conformations is difficult. Further, these loop algorithms don't handle internal secondary structural changes efficiently and identification of ECL2 with accurate secondary structure can be challenging. Therefore, by performing loop building via peptide fragment insertion concomitantly with structure-based modeling of the full-length receptor, we focus the sampling of loop conformations around the native-like fold. This structural alignment and simultaneous loop building are a large reason why our alignment method outperformed all other alignment methods tested. As shown, almost all methods could accurately align the TM region, but only our method performed well also in the ECL2 which brought the Full Model RMSD down. This blend of sequence- and structure-based alignments may prove critical to successful modeling of other protein families in which available templates have poor sequence identity but likely similar folds.

5.5.2 Template and Peptide Hybridization are Key Drivers for Accurate Modeling

The Rosetta code for multiple-template homology modeling made two primary changes to the method of homology modeling. The first was the ability to leverage the peptide fragment library derived from structures in the PDB. The peptide fragment library is a set of 3mers and 9mers that are mined from the PDB based on the target sequence. These fragments help bias the local geometries of the predicted structures towards structures that are known to exist and are low energy for these short regions. The ability to swap in peptide fragments is helpful for homology modeling, even of single templates,

because it allows sampling away from the starting template structure in regions that do not serve well for the target structure. We showed here that threading alone, for even very high identity templates, can be further improved solely by incorporation of peptide fragment hybridization.

The other key change was the use of multiple templates. The original authors of Rosetta's multiple template homology modeling showed that the use of multiple structures is better than using a single structure. The reasoning is the same as stated above that for a given target, the best available template is still likely to have one or more regions that don't accurately represent geometries accessible to the target sequence. However, as multiple related templates are used, the likelihood of finding the global fold from parts of the different templates increases. For GPCRs, this has been noted before and many GPCR modeling protocols now use multiple templates. However, their treatment of the multiple templates is distinct than the process used in Rosetta. Often these methods try to pre-identify which regions of the target protein are best captured by each target, and the segments of each template serve as the starting model for energetic minimization and loop rebuilding. A key problem of this method is that segments are selected based on local sequence identity, which as shown can be a poor indicator of structural similarity in this family of proteins. Other methods have tried to move towards combining templates via averaging or summing of segments. In Rosetta, all templates are passed to the hybridization protocol which randomly swaps segments of templates throughout the entire protein. After each modification, the structure is rescored and if the score improves, the segment is maintained. As a result, segments that perhaps contained initial lower sequence identity and would have been discarded before the modeling began in other methods, may

prove to have better geometries and energies for the target sequence. This discarding by energy of the conformation can lead to sampling of more native like conformations for the target sequence. Combined with the peptide fragment insertion, Rosetta can quickly identify native-like conformations distinct from any given template and perform dense sampling around this novel conformation. This is not to suggest that Rosetta will sample conformational changes on the scale of receptor activation (~10 Å in TM 6), but around 1-2 Å of the input templates. While none of this is new to Rosetta, the application and focus on GPCRs is novel. We present here the method specific for GPCRs and identify the optimal number of templates. Further by optimizing the alignment method and simultaneous loop modeling, we were able to push the previously reported template threshold of a 40% sequence identity minimum down to 20%.

5.5.3 Development of the First Rosetta-based GPCR Database

While several GPCR model databases or on-the-fly modeling servers exist, there is as-of-yet one developed within Rosetta. The Robetta server is available for general homology modeling, but is not specified for membrane proteins. Further, it uses automatic alignment methods which we showed are not ideal for this class of proteins. Therefore, we decided to generate a database of all non-odorant human GPCRs. We have confidence in models with templates above 20% sequence identity which accounts for 88% of the receptors. It is important to note that this method was benchmarked on inactive state structures and therefore the models in the database are inactive models. These can serve as structures for understanding biochemical data and genetic variations. Further, they can serve as structures for *in silico* docking strategies. It has been shown before that docking into inactive structures can work for both identification of agonists and antagonists.

However, work is ongoing to develop the database further to include active state structures.

5.6 Conclusions

Accurate modeling of GPCRs is a critical technology for understanding the structural basis of ligand recognition and signal transduction for the ~350 druggable GPCRs that have not had structures determined. Many of these proteins already have FDA approved drugs targeting them but a deep understanding of the molecular basis of drug intervention is lacking. Further, about a third of these receptors are classified as orphan receptors because the endogenous ligand has not been identified. A structural perspective of the ligand binding pocket may help shed light on this group of receptors. Lastly, it should be noted that this protocol is dependent on novel information. As new structures become available, the template dataset will increase as will the accuracy of the alignment and resulting models. Despite this, in the current format, RosettaGPCR stands as the best modeling protocol for GPCRs.

CHAPTER 6

Structural Model of Ghrelin Bound to its G Protein Coupled Receptor

6.1 Summary

This chapter uses the improved modeling of GPCRs from Chapter 5 and extends the iterative docking of Chapter 3 to model the highly flexible ghrelin peptide bound to its receptor. The novel approach to modeling highly flexible binding partners takes into account conformational selection of both the peptide ligand and the receptor binding pocket and loops by dense local sampling of backbone structures in a hierarchical building from the deepest part of the binding pocket outwards. This method is extremely useful to the field of peptide-GPCR binding structural biology. My contribution to this chapter is the computational modeling; experimental NMR and mutational data were collected by collaborators. This chapter comes from the article "Structural Modeling of Ghrelin Binding its G-Protein Coupled Receptor" for which I am the first author.

6.2 Introduction

Ghrelin is a 28 amino acid bioactive peptide that binds to the G protein-coupled receptor (GPCR) growth hormone secretagogue receptor 1a (GHSR) to induce a release of growth hormone [116]. Of all of its biological activities, it is best characterized as an orexigenic peptide as its levels increase just before feeding times to stimulate hunger via activation of GHSR [345]. GHSR, a rhodopsin-like subfamily β GPCR, has a rich history

of pharmacological development. Its first exogenous ligands were discovered prior to the discovery of ghrelin and subsequently aided in the isolation of the peptide in a rare case of reverse pharmacology [346]. The receptor is known to have high constitutive activity on the order of 50% [347] and a high degree of dynamics in the apo state [348]. Much mutational work has attempted to determine the structural basis for this constitutive activity [349-351] and to identify ligand binding sites [122-125, 129, 347, 352, 353]. The majority of this work has focused on small molecule and peptidomimetic ligands while little work has attempted to develop a model of endogenous ghrelin binding at the GHSR.

When ghrelin was initially isolated from rat stomach it was found to contain an *n*-octanoylation at the Ser^3 residue [116]. Ghrelin is currently the only known lipidated bioactive hormone. While des-acyl ghrelin is found in high concentrations in the bloodstream, it is only the lipidated form that is active at the GHSR [117]. It is thought that the lipidation of ghrelin peptide increases its interaction at the plasma membrane which would increase its local concentration and enhance receptor binding [354]. Despite the weak partitioning of an octanoyl chain in the membrane, we previously showed that ghrelin preferentially interacts with the membrane via its lipid modification and subsequent amino acid Phe^4 [7]. The membrane-bound structure was found to be highly dynamic possessing a central α-helical core and disordered termini. This contrasted with the aqueous state of the peptide which displayed almost no helical character. This work demonstrated that the environment was crucial to the structure of the peptide and suggested that the receptor-bound state may yet be different from the membrane-bound or aqueous states.

Peptide-binding GPCRs comprise a large subset of the GPCR family and several structures of these receptors have been solved to date [27]. However, only a few of these

structures were determined with their respective peptide ligand. These include rhodopsin-like subtype β receptors like GHSR—neurotensin [33] and endothelin-B [43] receptors—and subtype γ receptors such as the γ-opioid receptor [355], C5a receptor [50], apelin receptor [166], and chemokine receptors CXCR4 [356] and US28 [357]. Nuclear magnetic resonance (NMR) has allowed structural and dynamic studies of peptides binding their receptors including bradykinin [23, 358], NPY [2, 44], and dynorphin [165]. An emerging theme in these studies is the large binding surface areas of peptide ligands compared to small molecule ligands. Many of these peptide-receptor interactions involve two sites of binding. One site correlates with the well characterized orthosteric site of rhodopsin-like GPCRs and an additional site is located in the N-terminus such as in chemokine receptors [356, 357] or extracellular loop 2 (ECL2) in NPY type 2 receptor [2].

Initial studies of ghrelin focused on identifying the binding site within the 28-residue peptide. It was determined that the N-terminus including the lipid-modified Ser^3 residue was critical for receptor binding and the first five residues were suggested to represent the minimal binding motif [117]. However, the binding affinity of this short peptide in transfected cell assays was two orders of magnitude lower than that of full-length ghrelin. Later, it was found that this short peptide could not activate GHSR in isolated membranes and had no efficacy in rat models [359]. These findings suggested a critical role for residues outside of the first five N-terminal amino acids. We report here an extension of this binding model to at least residue 9 at a site in GHSR distinct from the orthosteric binding pocket as determined by NMR spectroscopy. We expressed and purified GHSR and reconstituted it in lipid bilayers and identified secondary structure characteristics and residues in close proximity to the receptor using isotopically labeled

ghrelin. In order to generate a structural model of these highly dynamic binding partners in complex, we developed a novel modeling method that hierarchically built the ligand-receptor complex. Iterative extension of receptor loops and peptide residues allowed for a higher density sampling of the conformational space of these two binding partners. Resulting models were filtered against the experimental NMR data and final models were compared against existing and novel mutational analysis. The final ensemble of models that best fits the data identifies an extended binding pocket from the central transmembrane bundle out to the receptor's extracellular loops.

6.3 Methods

6.3.1 Experimental Model and Subject Details

Parts of this study were conducted in human and animal cell lines. HEK293 (human embryonic kidney) cells were cultured in Dulbecco's modified Eagle's medium with 4.5 g/L glucose and L-glutamine (DMEM) and HAM's F-12 (1:1, *v/v*) Lonza, Basel, Switzerland) supplied with 15% (*v/v*) heat-inactivated FBS (Biochrom, Berlin, Germany). COS-7 (African green monkey, kidney) cells were maintained in DMEM with 10% (*v/v*) heat-inactivated FBS. All cell lines were cultured under humidified atmosphere at 37°C and 5% CO_2.

HEK293 were authenticated using a PCR based multiplex assay based on the use of short tandem repeats (STR) (Authentication of Human Cell Lines: Standardization of STR Profiling, ANSI/ATCC ASN-0002-2011). Identity of COS-7 was confirmed by DNA Barcoding by PCR amplification of 5´coding region of cytochrome c oxidase I. Cycle

sequencing of respective PCR products revealed taxon assignment upon submission to BOLD [360]. All cell lines were routinely tested (negative) for mycoplasma contamination.

6.3.2 Method Details

Expression and purification of GHSR. The GHSR was prepared as described before [348]. Briefly, the receptor was expressed with a C-terminal His8-tag (pET41b(+)-GHSR_M1-T366_LE_8xHis) in *E. coli* NiCo21(DE3) (NEB, Frankfurt am Main, Germany) in an insoluble form in inclusion bodies. Subsequently, the receptor was solubilized with 15 mM SDS and 50 mM DTT. After removal of DTT, it was purified using affinity chromatography. For reconstitution of the receptor into a lipid membrane the correct disulfide was formed at room temperature by dialysis in 50 mM sodium phosphate (pH 8), 2 mM SDS, 1 mM EDTA, 2 mM reduced and 1 mM oxidized glutathione at a receptor concentration of 0.5 mg/ml. After concentrating the receptor to 1 mg/ml it was added to preformed bicelles comprised of DMPC (or DMPC-d_{54}) and DHPC with a ratio of receptor:DMPC:DHPC of 1:600:2400 (q=0.25) for solution NMR experiments, or 1 : 200 : 800 for solid-state NMR experiments. Three cycles of temperature jumps between 0°C and 42°C were performed. The solution NMR sample was then dialyzed against 50 mM sodium phosphate pH 7, 1.5 mM DHPC and it was concentrated with 20-30% (w/v) PEG 20,000. For the solid-state NMR samples, BioBeads SM-2 (BioRad, Germany) were added to a concentration of 50 mg/ml to reduce the amount of detergent to yield samples with a q value higher than 10. The resulting turbid solution was incubated with a twofold molar excess of peptide for 1 h at 37°C. The sample was finally centrifuged and the pellet loaded into the MAS rotor.

Peptide synthesis. Ghrelin was synthesized by Fmoc/*tert*-butyl strategy as described before [7, 361]. Briefly, 15 µmol of [Ser(Trt)^3]-Ghrelin were synthesized automatically on R-Wang residue on a robot system (SyroI, MultiSynTech, Bochum, Germany). After selective deprotection of Ser^3 by treatment of the resin with 10 times 1 ml of TFA/TIS/DCM (1:5:94, v/v/v), octanoic acid was coupled. First, free reactive hydroxyl groups at the resin were preactivated with 5 equivalents of DIC in DMF for 10 min. Subsequently, 5 equivalents of HOBt, 5 equivalents of octanoic acid, 5 equivalents of methylimidazole and 0.1 equivalents of DMAP in DMF were added and shaken overnight. $^{13}C/^{15}N$-labeled peptides (Table 6.1) were synthesized on R-Wang resin as described for ghrelin (Vortmeier et al., 2015). $^{13}C/^{15}N$-amino acids were coupled manually with 5 equivalents of Fmoc-protected amino acid, 5 equic HOBt and 5 equivalents of DIC in DMF overnight. Fluorescent tracer [Dpr^3-Oct; Dpr^{16}-atto520]-Ghrelin (Dpr, diaminopropionic acid; Oct, octanoic acid) for in vitro functionality assay was prepared as described [348]. Briefly, 15 µmol of Boc-[Dpr(Mtt)^3, Dpr(ivDde)^{16}]-ghrelin was synthesized automatically on R-Wang resin. After selective deprotection of methyltrityl (Mtt) with 1% TFA/DCM (10x 1 min), octanoic acid was coupled using 1-hydroxybenzotriazole/diisopropylcarbodiimide (HOBT/DIC) over night (5 eq. each). Next, ivDde (1-(4,4-dimethyl-2,6-dioxocyclohex-1-ylidene)-3methylbutyl) protection group was selectively cleaved off using 2% hydrazine/DMF (10x 10min), and 1 eq. of the atto520 dye (free acid) was coupled using HOBT/DIC in DMF overnight. Peptides were cleaved off the resin using 90/5/5 (v/v/v) trifluoroacetic acid/H_2O/triisopropylsilane for 3h and precipitated by ice cold diethyl ether. Peptide identity was confirmed by MALDI-ToF mass spectrometry (Ultraflex III MALDI ToF/ToF, Bruker), and purified to > 95% by reversed-phase HPLC applying linear gradients of H_2O + 0.1% TFA and acetonitrile (ACN) +

0.08% TFA.

Table 6.1 Peptides used in this study. Residues in red and underlined were $^{13}C/^{15}N$-labeled.

Peptide sequence	Molecular weight [Da]
GHR1: H2N- GSS(n-octanoyl-d15)FL SPEHQ RVQQR KESKK PPAKL QPR -OH	3398
GHR2: H2N- GSS(n-octanoyl)FL SPEHQ RVQQR KESKK PPAKL QPR -OH	3384
GHR3: H2N- GSS(n-octanoyl)FL SPEHQ RVQQR KESKK PPAKL QPR -OH	3385
GHR4: H2N- GSS(n-octanoyl)FL SPEHQ RVQQR KESKK PPAKL QPR -OH	3388
GHR5: H2N- GSS(n-octanoyl)FL SPEHQ RVQQR KESKK PPAKL QPR -OH	3379
∑: H2N- GSS(n-octanoyl-d15)FL SPEHQ RVQQR KESKK PPAKL QPR -OH	

In vitro binding assay. Functionality of GHSR reconstituted in DMPC/DHPC bicelles (with q=0.25) was tested by a fluorescence binding assay using [Dpr^3-Oct; Dpr^{16}-atto520]-ghrelin (ghrelin-atto520). Various concentrations of GHSR-containing bicelles in 50 mM sodium phosphate buffer pH 7 were incubated with 50 nM ghrelin-atto520 overnight. Samples were transferred into a 10 mm quartz cuvette and measurements were carried out on a FluoroMax-2 (JOBIN YVON) at 20°C with linear polarized light, an excitation wavelength of 500 nm, an emission wavelength of 540 nm, and 90° detection angle. Two independent measurements were carried out in duplicate and given data points are the mean values of the four experiments (with their standard deviation). Saturation binding data were fit with a sigmoidal dose-response curve using the Origin software. To assess the binding of ghrelin-atto520 to the membrane alone, the measurements were repeated with empty bicelles (a single measurement in duplicate).

^{2}H NMR of Deuterated Ghrelin-d_{15}. 20 mg of GHSR embedded in large bicelles composed of DMPS/DMPC (1/5 mol/mol) was prepared as stated above. The pellet was resuspended three times in 5 ml of de-deuterated water and centrifuged 15 min at 5,000 rpm. Ghrelin (5.25 mg) with a deuterated octanoyl chain (ghrelin-d_{15}) was dissolved in 1 ml de-deuterated water and the pH was adjusted to pH 7 with NaOH. The peptide solution was added to the receptor sample and the complex incubated 1 h at 37°C. Unbound ghrelin was removed by centrifugation (15 min at 5,000 rpm).

Static ^{2}H NMR spectra were acquired on a Bruker Avance I 750 MHz spectrometer with a ^{2}H resonance frequency of 115 MHz at a temperature of 30°C. A quadrupolar echo pulse sequence was used. The 90° pulses had a length of 3.1 µs, the echo time was 60 µs and the relaxation delay 1 s. ^{2}H NMR spectra were processed and simulated with a program written in Mathcad to estimate the amount of tightly bound, membrane associated and free ghrelin [362].

STD NMR of Ghrelin. To investigate ligand-receptor contacts by saturation transfer difference (STD) NMR, GHSR was reconstituted in DMPC/DHPC bicelles (q=1) to achieve a receptor concentration of 25 µM. By dialysis with 25% PEG 20,000 the volume was adjusted to 1 ml. Ghrelin peptide was added in a 100-fold molar excess. A pseudo-2D version of the STD NMR sequence was used for the interleaved acquisition of on- and off-resonance spectra [363]. For selective saturation of the protein, a cascade of 50 E-Burp-shaped pulses with a length of 50 ms each and an interpulse delay of 1 ms was used. The on-resonance frequency was set to -1 ppm and the off-resonance pulse was applied at 33.3 ppm. For suppression of the residual water signal, a Watergate sequence was used. Typically, STD NMR spectra were acquired with a 100-fold excess of ligand, a

saturation time of 3 s, 5120 scans, and 8 K complex points. Prior to Fourier transformation FIDs were multiplied by an exponential line broadening function of 1 Hz.

^{13}C/^{13}C Single Quantum Double Quantum NMR. The experiments were carried out on a Bruker Avance III 600 NMR spectrometer with a 4 mm spinning module at a temperature of -30°C and an MAS frequency of 7 kHz. ^{1}H as well as ^{13}C 90° pulses had a length of 4 µs. For heteronuclear decoupling, the SPINAL64 was used with a ^{1}H radio-frequency field strength of about 65 kHz. ^{13}C magnetization was achieved with a cross polarization step with a contact time of 700 µs. The SQ/DQ correlation spectra were acquired using the SPC5 recoupling sequence for double quantum excitation and reconversion (set to 0.571 ms each). The relaxation delay was 2 s.

Modeling of the GHSR-Ghrelin Complex. We expanded on the structures generated in the previous study which were only used to define the membrane in Rosetta simulations but did not interact with the peptide. An ensemble of starting structures were generated previously used with varying backbone conformations [7]. The extracellular loops of the receptor were removed and the first 7 amino acids of ghrelin with an acetyl group on Ser^{3} were docked using FlexPepDock [9]. A constraint between the N-terminal amino head group and the oxygen atoms of Glu 124$^{3.33}$ was used to filter the resulting models. Top models were passed through to a round of loop modeling guided by known crystal structures using RosettaCM [5] in which most of ECL1 was built and the initial residues of ECL2 and ECL3 were built. A 12mer peptide was docked into these new extended models but were trimmed to a 10mer because of significant clashes with ECL3. A subsequent round of loop modeling completed ECL1 and ECL3 and built ECL2 through the β-sheet secondary structure region seen in all peptide binding crystal structures to date.

The N-terminal 17 residues of ghrelin comprising the N-terminal flexible region and α-helical middle region were docked into this nearly complete structure. Lastly ECL2 was closed in a final loop modeling step. These last structures were modified such that the acetyl group on Ser^3 was replaced by the full octanoyl chain followed by a last round of energetic minimization. These models were filtered by overall energy then analyzed by Proshift to compare calculated chemical shifts with the experimentally determined data. This resulted in a small ensemble size of 5 peptide conformations that had the lowest RMSD to the experimental data.

Mutagenesis of GHSR plasmid. QuikChange™ site-directed mutagenesis (Agilent, Santa Clara, USA) was used to introduce mutations into the eukaryotic expression plasmid GHSR1a_EYFP_pVitro2 [125]. The correct sequences were confirmed by Sanger DNA sequencing.

Ca^{2+} Mobilization Assay. HEK293 cells were cultured in 25 cm^2 flasks until 70-80 % confluency and subsequently transfected with 4 µg plasmid DNA and 15 µl Metafectene®Pro (Biontex, Munich, Germany) as a transfection reagent. One day after transfection, the cells were seeded into poly-D-lysine coated 96-well plates (black) and incubated overnight. Prior to stimulation, the medium was replaced by assay buffer (HBSS, 20 mM HEPES, 2.5 mM Probenecid (Sigma-Aldrich, St. Louis, USA)) containing 4 µM fluorescent dye Fluo-8 AM (Abcam, Cambridge, UK) and 0.1% Pluronic F127(Sigma-Aldrich, St. Louis, USA). After incubating the cells for 60 min at 37°C and 5% CO_2, the dye solution was replaced by 150 µl assay buffer. Subsequently, fluorescence measurement (excitation 485 nm, emission 525 nm, cutoff 515 nm) was carried out with a Flexstation® 3 microplate reader (Molecular Devices, Sunnyvale, USA). Following a baseline recording

for 20 s, 50 µl of serial dilutions of ghrelin (6-fold concentrated, 10^{-11} M to 10^{-5} M) were added automatically through a multi-channel pipetter and the fluorescence was detected for another 50 s. The experiments were repeated at least two times independently in technical triplicate.

Inositol Phosphate Accumulation Assay. COS7 cells were transiently transfected with 4 µg receptor plasmid in 25 cm^2 culture flasks using 15 µl of MetafectenePro (Biontex, Munich, Germany) according to the manufacturer´s instructions. 24 h post-transfection, the cells were seeded into 48-well plates and grown to 90% confluency. The cells were labeled in DMEM supplied with 10% (*v/v*) FBS containing 2 µCi/ml myo-[2-^{3}H]-inositol (Perkin Elmer, Waltham, USA) at 37°C in the presence of 5% CO_2 for 16-18 h. After aspirating the labeling solution, the cells were washed once with DMEM containing 10 mM LiCl and stimulated for 2 h with increasing ghrelin concentrations, ranging from 10^{-11} M to 10^{-5} M in DMEM with 10 mM LiCl. Stimulation was stopped by aspiration of the medium followed by basic cell lysis with 0.1 M NaOH and subsequent neutralization with 0.13 M formic acid. After removal of the cell debris, radioactive inositol phosphate species were diluted and purified on an AG 1-X8 anion exchange resin (Bio-Rad, Hercules, USA) as described [364]. Radioactivity of the eluates was measured on a liquid scintillation counter. The experiments were repeated at least two times independently in technical triplicate.

Live cell microscopy. Membrane localization of the mutant GHSR variants was verified by fluorescence imaging. HEK293 cells were seeded onto microslide 8-wells (ibidiTreat, Martinsried, Germany) to 70-80% confluency and transiently transfected with 1 µg plasmid DNA and 1 µl Lipofectamine®2000 (Invitrogen, Carlsbad, USA) for 1 h.

Following cultivation in DMEM and Ham's F-12 (1:1, *v/v*) supplemented with 15% FBS overnight, the medium was replaced by Opti-MEM® (Life Technologies, Darmstadt, Germany) supplemented with 10 µM Hoechst33342 (Sigma-Aldrich, St. Louis, USA) and incubated at 37 °C for 10 min. Membrane localization was documented in pure Opti-MEM® using an AxioObserver.Z1 microscope with an ApoTome Imaging System (Carl Zeiss AG, Oberkochen, Germany). Representative images of at least two independent experiments are shown.

6.3.3 Quantification and Statistical Detail

Statistics from receptor activation assays were calculated in GraphPad Prism 5.03 (GraphPad Software, San Diego, USA) using standard non-linear regression (log(agonist) vs. response (three parameters)) and normalization to wild type response. All experiments were performed at least in duplicates. EC_{50} values are given as mean ± standard error of the mean (SEM). Error bars are displayed in Figure 4 and described in the caption.

6.3.4 Data and Software Availability

The ^{13}C chemical shift data has been deposited in the BMRB under ID code 27600. The ensemble of ghrelin/GHSR complexes has been deposited in the PDB-Dev under ID code PDBDEV_00000024.

6.4 Results

6.4.1 Expression and Functional Refolding of GHSR

GHSR was expressed as inclusion bodies in E. coli using batch-fed fermentation as previously described [348]. It was isolated in a denatured form and reconstituted in vitro into lipid bicelles formed from a mixture of 1,2-dimyristoyl-sn-glycero-3-phosphocholine (DMPC) and 1,2-diheptanoyl-sn-glycero-3-phosphocholine (DHPC). Either small bicelles

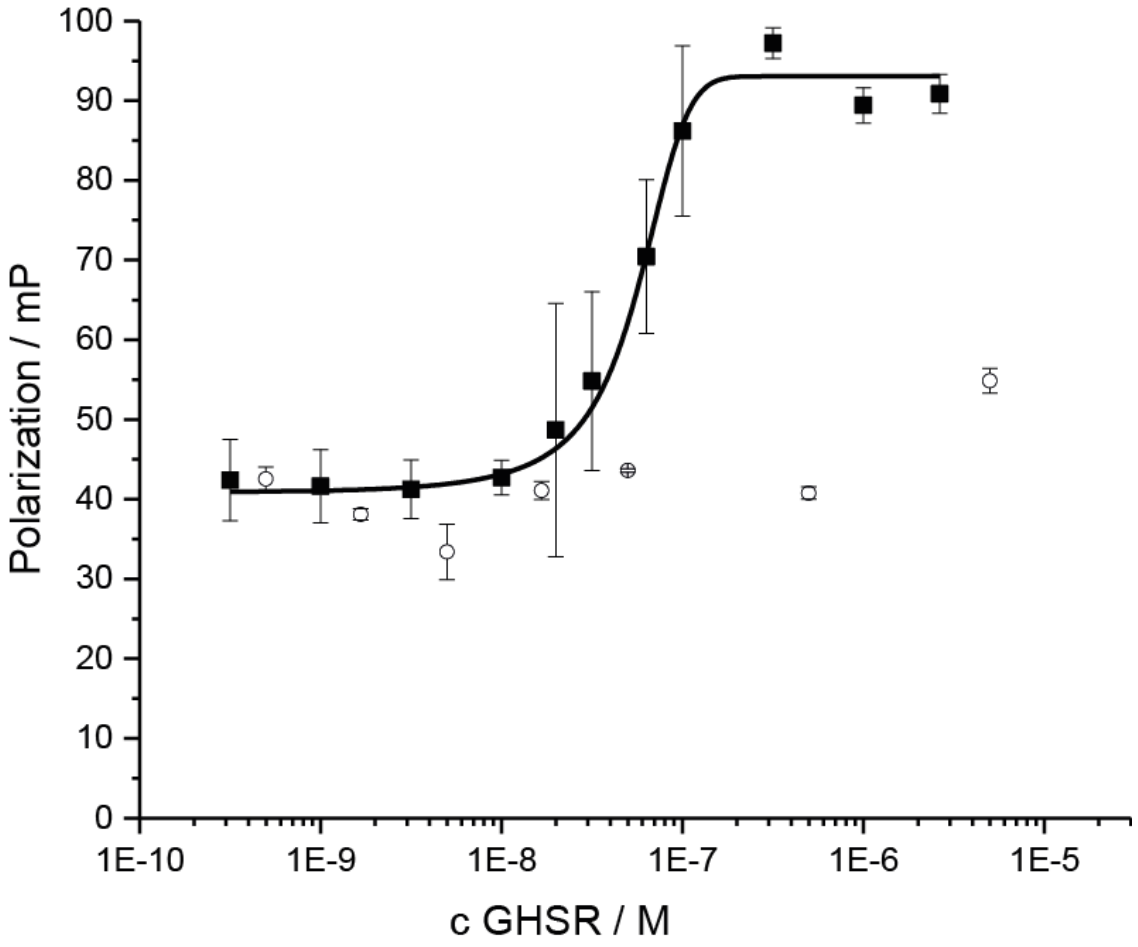


Figure 6.1 In vitro binding assay of GHSR. Titration of GHSR to ghrelin-atto520 (c=50 nM) enhanced fluorescence intensity of the peptide tracer (filled circles) with half maximal effect of 54 nM (95% CI: 18 – 38 nM). This is approximately at the assay limit of 0.5x [ligand] and presents significant enhancement upon nonspecific binding to empty bicelles (open circles), demonstrating high functionality of GHSR preparations.

(q=0.25, q being the molar ratio of DMPC to DHPC) or large bicelles (q>10) were used in solution and solid-state NMR experiments, respectively. We used bicelles, because in our hands, they provide the best lipid systems for the reconstitution of high amounts of functional receptor. Furthermore, by tuning the ratio between long and short chain phospholipids, the size of the bicelles can be tuned to be suitable for solution (q = 0.25) and solid-state NMR (q > 10). With the high q value of q > 10, the structures that result are best characterized as large lamellar membrane discs with a few holes inside, as characterized in our previous work [365]. Receptor functionality was confirmed by saturation ligand binding using atto520-labeled ghrelin peptide [348]. Representative measurements on the functionality and the pharmacological characterization of the GHSR are shown in Figure 6.1.

6.4.2 Investigation of Ghrelin Partitioning between Membrane- and Receptor-Bound States

The octanoylation of ghrelin in combination with hydrophobic residues Phe^4 and Leu^5 render the molecule lipophilic and provide the basis for membrane partitioning of ghrelin [7]. Although the affinity of ghrelin to GHSR is much higher than to the membrane, the membrane area much exceeds that of the receptor, which may lead to sizable unspecific membrane binding. To quantitatively determine this equilibrium, we attached a deuterated octanoyl chain to ghrelin (ghrelin-d_{15}) and measured the ^2H NMR spectra of ghrelin-d_{15} bound to empty lipid bicelles and bicelles containing GHSR. In order to produce comparable data to the study of ghrelin bound to membranes in the absence of GHSR [7], negatively charged DMPC/DMPS (5/1, mol/mol) were used, which leads to increased membrane binding. The ^2H NMR spectrum of ghrelin-d_{15} measured at a

temperature of 30 °C is shown in Figure 6.2. The ^{2}H NMR line shape is complicated consisting of several contributions of varying origin. Most obvious, two dominating contributions can be extracted from the spectral line shape at first sight, comprising (i) a large isotropic NMR signal indicative of a highly mobile octanoic acid, and (ii) a Pake spectrum with a dominating quadrupolar splitting of ~110 kHz for the CH_2 groups, indicative of a largely immobilized octanoyl chain. These contributions can be assigned to free and receptor-bound ghrelin, respectively. The 110 kHz quadrupolar splitting corresponds to an order parameter of 0.88 indicating that the octanoyl chain of ghrelin is highly immobilized when the molecule is bound to GHSR. As ^{2}H NMR spectra were excited by a direct pulse and acquired with sufficient relaxation delay, the individual contributions to the spectral intensity can be readily obtained from the experimental ^{2}H NMR spectrum using a quantitative line shape analysis on the basis of numerical

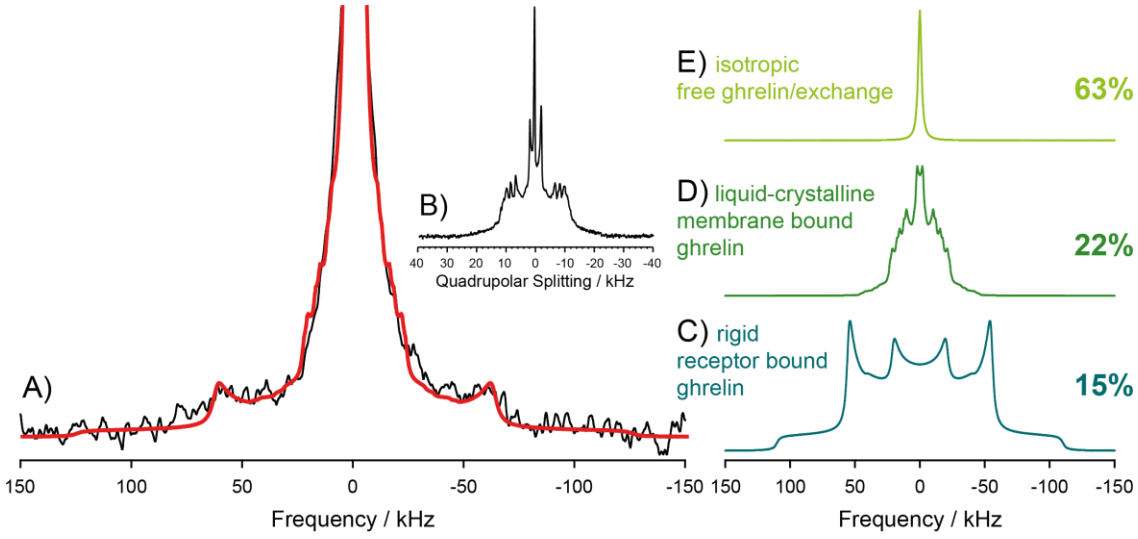


Figure 6.2 ^{2}H NMR analysis of ghrelin binding to GHSR and membrane fragments. A) Experimental ^{2}H NMR spectrum of ghrelin-d_{15} in the presence of GHSR-containing DMPS/DMPC (10/90, mol/mol) bicelles/membranes at a temperature of 30°C. For comparison, panel B) reproduces the ^{2}H NMR spectrum of ghrelin-d_{15} in DMPS/DMPC membranes as published before [7] C)-E) Simulated ^{2}H NMR spectra of a rigid (B), a liquid-crystalline (based on the order parameters determined in pure membrane environment, (C)), and an isotopically moving octanoyl-d_{15} chain (D), respectively. The red curve in panel A) represents the superposition of the simulated lineshapes at a 15:22:63 ratio.

simulations [362]. However, the superposition of the aforementioned isotropic signal and immobilized Pake spectrum do not provide a satisfying description of the experimental ^{2}H NMR line shape. Therefore, a third contribution was used in the numerical line shape analysis, which is the ^{2}H NMR spectrum of ghrelin bound to empty bicelles and is characterized by motionally averaged quadrupolar splitting as investigated in detail before (spectrum reproduced in Figure 6.2B [7]). With these three spectral contributions, a reasonable description of the ^{2}H NMR signal could be achieved as plotted in Figure 6.2. Quantification of the peak fitting yielded (i) 63% isotropic contribution, indicative of free or fast exchanging ghrelin, (ii) 15% rigid contribution, indicative of receptor bound ghrelin, and (iii) 22% liquid-crystalline contribution, indicative of membrane-bound ghrelin.

6.4.3 Determination of Ghrelin Segments in Close Contact with GHSR by STD NMR

A technique to measure specific interactions between a ligand and its receptor is saturation transfer difference (STD) NMR spectroscopy [363]. An important limitation of this technique is that the ligand must be of sufficiently low affinity to allow for rapid exchange at the receptor [366]. As ghrelin's affinity for GHSR is low nanomolar, it was not per se expected to exhibit rapid exchange. However, given the high exchange between receptor-bound, membrane-bound, and free ghrelin deduced from the ^{2}H NMR results, STD NMR spectroscopy was suggested to determine the ghrelin residues that are in close contact with GHSR in the bound state. Figure 6.3 displays a ^{1}H NMR spectrum of ghrelin (A), a ^{1}H STD NMR spectrum of ghrelin in the presence of empty zwitterionic bicelles (B), and a ^{1}H STD NMR spectrum of ghrelin in the presence of zwitterionic bicelles containing

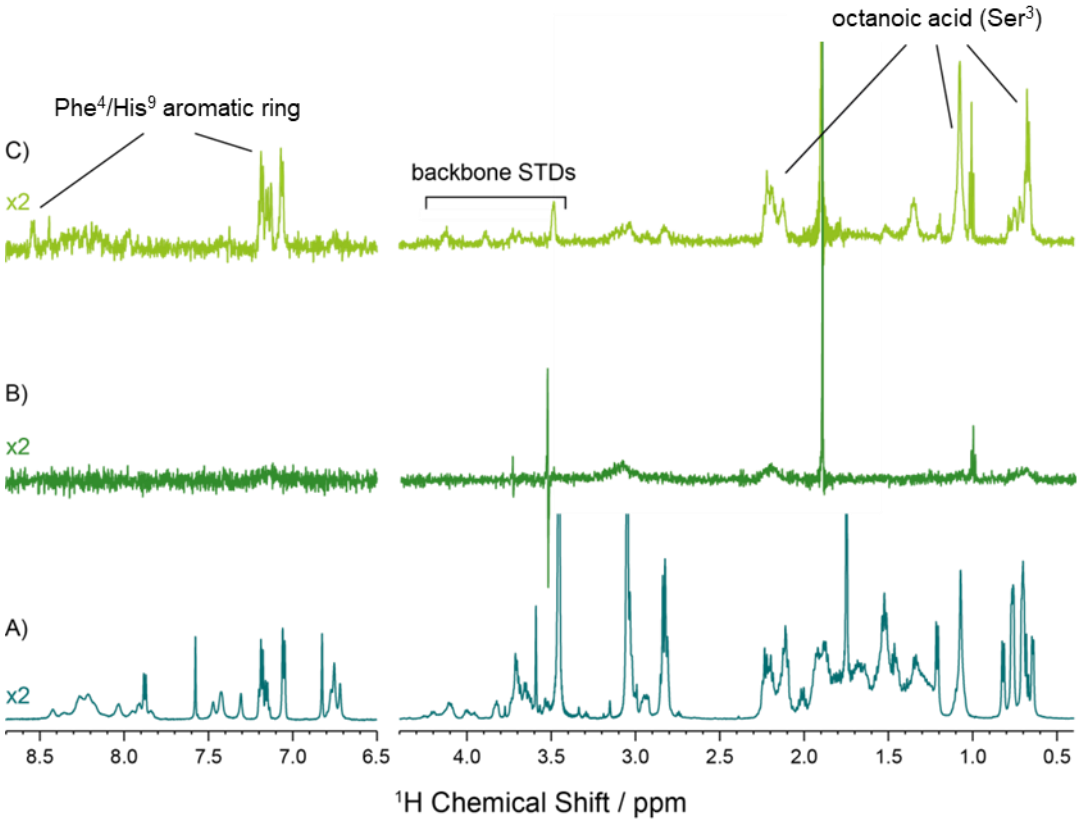


Figure 6.3 ^{1}H NMR saturation transfer difference spectroscopy of ghrelin interaction at GHSR. A) ^{1}H NMR spectrum of ghrelin; B) ^{1}H STD NMR spectrum of ghrelin in the presence of 5 mM DMPC/DHPC bicelles; C) ^{1}H STD NMR spectrum of ghrelin in the presence of 5 mM DMPC/DHPC bicelles loaded with 25 µM GHSR. All NMR spectra were acquired in 50 mM NaP buffer (pH 7) at a temperature of 20°C.

GHSR (C). Generally, only very weak STD effects are measured for receptor-bound ghrelin, which is plausible from the high affinity of the peptide to its receptor. The only ghrelin signals that are observed in the STD NMR spectrum in the presence of GHSR can safely be assigned to the octanoyl chain at Ser^{3} and hydrophobic residues Phe^{4} and His^{9}. We used empty zwitterionic lipid bicelles as a control and did not observe any of these STD effects, only weak phospholipid signals pass the STD filter. As we used zwitterionic DMPC/DHPC bicelles in these experiments, the affinity of ghrelin to the membrane surface is drastically reduced [7], allowing to exclusively determine STD effects of ghrelin in interaction with GHSR. Thus, zwitterionic bicelles provide the best model system to

study the exchange between receptor-bound and free ghrelin by STD NMR. The amplitude of signal in the STD spectrum is related to the proximity and rigidness of the binding residues with a greater signal indicating more tightly bound component. The STD effect is greatest for the octanoyl chain of Ser^3 (60-100% of saturation maximum), the Phe^4 sidechain (40-50%), and His^9 (10-30%) suggesting that these molecular groups are in very close proximity to GHSR (Table 6.2).

Table 6.2 Relative STD effects for the assigned protons of ghrelin in the presence of GHSR

Residue	Phe^4			Octanoic acid (Ser^3)			His^9		
Atom	H2,6	H4	H3,5	H13-15	H5-12	H3,4	H1,2	H3	H4
STD effect	45%	46%	40%	100%	56%	91%	61%	10%	31%

6.4.4 Structural Features of Receptor-Bound Ghrelin

We next sought to characterize structural features of ghrelin in its receptor-bound form. We produced a small library of selectively ^{13}C-labeled ghrelin peptides with 3-4 labeled residues at a time (Table 6.1). ^{13}C magic-angle spinning (MAS) NMR experiments of the ghrelin/GHSR complex in large DMPC/DHPC bicelles (q>10) were carried out at a temperature of -30°C. Using $^{13}C/^{13}C$ single quantum(SQ)/double quantum (DQ) correlation NMR spectroscopy, we were able to assign chemical shifts for each of these labeled residues. A typical ^{13}C SQ/DQ correlation NMR spectrum is shown in Figure 6.4A. The cross peaks could be assigned according to their correlation schemes within the amino acid side chain. For Pro^7, Glu^8, Gln^{14}, and Ala^{23}, we detected two sets of cross peaks giving rise

to two different sets of chemical shifts, which are reported in Table 6.3. The exact origin of the second contribution is unclear at the moment and points to slightly different backbone conformation and/or chemical environment. Pro^7 cis/trans isomerization is unlikely to have caused these two conformations as the measured chemical shifts do not support this. However, this indicates two slightly different chemical structures are possible of receptor-bound ghrelin. The chemical shift index (CSI) calculated by the difference in the Cα-Cβ chemical shift values relates information regarding the backbone geometry of these

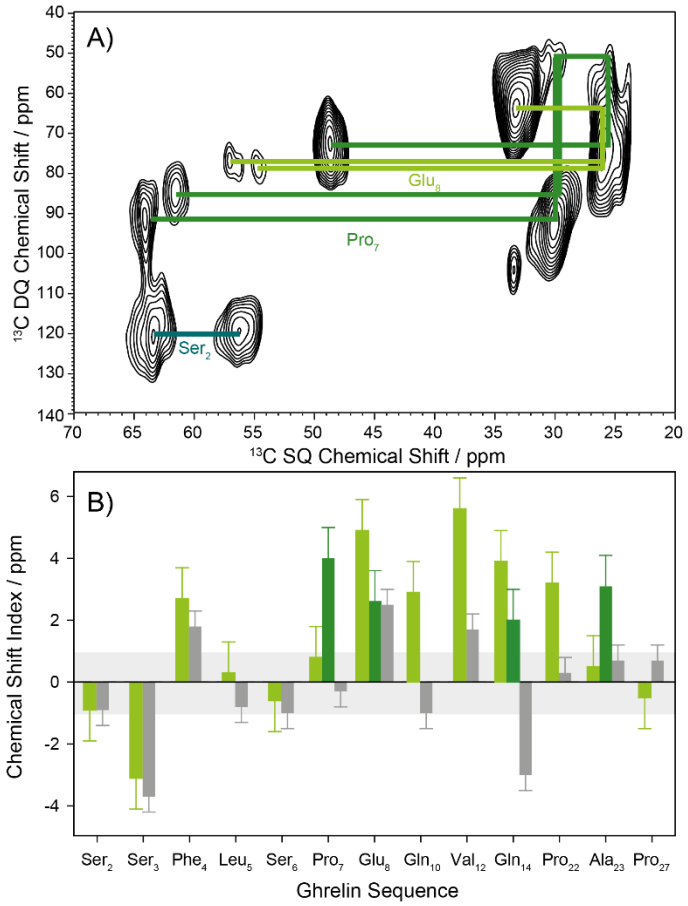


Figure 6.4 Structural data on GHSR-bound ghrelin. A) $^{13}C/^{13}C$ SQ/ DQ correlation NMR spectrum of a representative ghrelin peptide labeled in Ser_2, Pro_7, and Glu_8, with the correlation pattern indicated by lines. B) Chemical shift index (CSI) for the investigated labeled amino acids of GHS receptor bound ghrelin (light green bars). Positive values greater than 1 ppm indicate a tendency for α-helical structure, whereas values less than −1 ppm suggest β-sheet character. For comparison, the chemical shift index for membrane-bound ghrelin is given as gray bars (data adopted from literature [7]). For Pro_7, Glu_8, Gln_{14}, and Ala_{23}, two sets of chemical shifts were found, the second gives rise to the chemical shift index shown in dark green bars.

residues. Using the CSI across the collected residues we were able to generate a rough idea of the secondary structure of the bound peptide. We compared these results with our previously collected data on the membrane-bound structure [7] and found that while the peptide maintains a lack of secondary structure at the N-terminus, the degree of α-helical nature in the central region becomes more defined (Figure 6.4B). This may be due to selection by the receptor for ghrelin molecules in this particular conformation.

Table 6.3 ^{13}C NMR chemical shifts determined for GHSR bound ghrelin

Residue	Ser^{2}	Ser^{3}	Phe^{4}	Leu^{5}	Ser^{6}	Pro^{7}	Glu^{8}	Gln^{10}	Val^{12}	Gln^{14}	Pro^{22}	Ala^{23}	Pro^{27}
^{13}Cα / ppm	56.7	55.8	57.5	52.8	54.6	61.8 64.3[a]	57.0 54.7[a]	59.0	64.5	56.0 54.1[a]	61.0	50.5 53.1[a]	61.1
^{13}Cβ / ppm	63.6	64.9	37.8	40.5	61.2	30.3 29.6[a]	26.1	30.0	30.3	26.0	27.1	17.2	30.9

[a]two cross peaks were found in the 2D ^{13}C-^{13}C SQ/DQ correlation spectra

Taken together, three types of structural information were obtained from the NMR analysis, comprising (i) information on the binding equilibrium of ghrelin to the GHSR and the membrane, respectively, (ii) the important ghrelin residues/segments for the interaction with the GHSR, and (iii) the backbone conformation of ghrelin bound to the GHSR.

6.4.5 Modeling of the Ghrelin/GHSR Complex

To gain further insight into the structure of receptor-bound ghrelin, we used the Rosetta software to generate a model of the receptor/peptide complex. An initial model of the receptor was generated using RosettaCM from several backbone templates of diversely

related receptors [7]. Of note, these initial models were generated in the absence of a bound ligand and energetic minimization often collapsed the binding pocket. Additionally, as the peptide was expected to extend out of the central binding pocket and make extensive contacts with the loop regions, it was anticipated that there would be large inaccuracies in the initial models of these regions. In order to generate models of the peptide-bound receptor, we chose to remove all extracellular loops and rebuild them in the presence of the peptide. We devised an approach to achieve this in an iterative fashion by alternating between docking ghrelin and rebuilding loops with greater number of residues in each step (Figure 6.5). To begin, we docked the first seven residues of ghrelin with an acetylated Ser at position 3 to mimic the octanoyl chain using the fully flexible docking protocol FlexPepDock (Figure 6.5A). We selected this short peptide as the first five residues have been shown to comprise a minimal signaling motif at the receptor [117]. In the initial docking step, we filtered models that oriented the N-terminal amino group towards Glu124$^{3.33}$ (superscripts on receptor residues indicate Ballesteros-Weinstein nomenclature [367]) as this residue has been previously suggested to balance the charge of the positive amino head group [129]. Next, we built ECL1 using RosettaCM homology modeling based on other peptide-binding GPCRs and extended the peptide at the C-terminal end via the flexible docking protocol to ten residues (Figure 6.5B,C). ECL3 and part of ECL2 were next built in a similar fashion before extending the peptide to 17 residues and closing the loop of ECL2 (Figure 6.5D,E,F). We filtered models to ensure a β-sheet was formed in ECL2 as this has been observed in all available peptide-binding GPCR crystal structures [368]. In the computational work, we chose to model only the first 17 amino acids of ghrelin as the C-terminus has been shown to have no affinity for the receptor, does not enhance ligand binding over truncated variants [117], and is expected to be largely

disordered. Lastly, we extended the acetylation at Ser^3 to a full length octanoyl chain before a final energetic minimization of the complex. Following this step, chemical shifts were calculated for the top 20% of models by energy. These values were compared against the experimentally determined chemical shifts. This filtering step identified an ensemble of five models that collectively best represent the data (Figure 6.6A,B and Figure 6.7). Smaller ensemble sizes suffer from deviations in individual models that get canceled as a larger pool is averaged. Conversely, larger ensemble sizes suffer from additional noise from too many models. As such, the remaining analysis is performed on this ensemble of

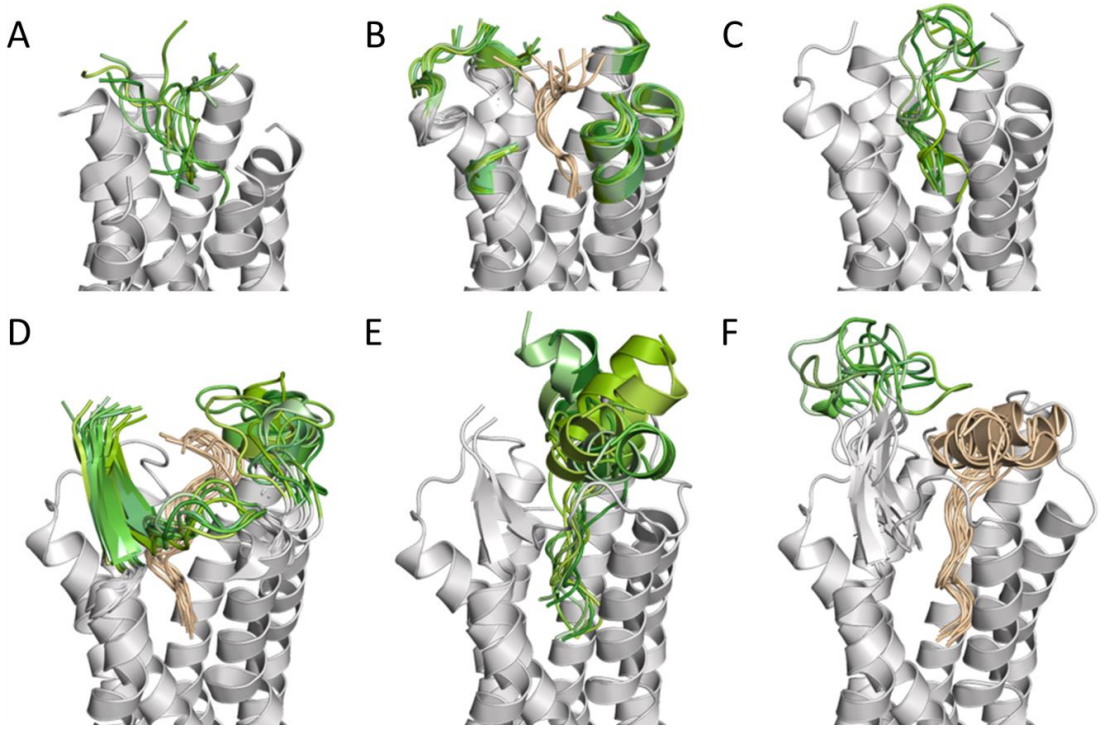


Figure 6.5 Iterative Ligand Docking and Loop Building of Ghrelin/GHSR-complex. (A) Using a previously generated homology model of GSHR (Vortmeier et al., 2015), we removed the extracellular loops and part of the external helices to provide a minimal ligand binding pocket. The first seven residues of ghrelin show in green were docked using FlexPepDock. (B) Using inputs from the previous round of modeling in which the ghrelin peptide is present (tan), the helices of GHSR were extended and most of ECL1 was built using RosettaCM to allow for extensive sampling of backbone conformations in these local regions shown in green. (C) Using the backbones generated in the previous round, an extended form of ghrelin through the first 12 residues are docked with FlexPepDock. (D) In the presence of ghrelin1-12 (tan), all of ECL1 and ECL3 are built with RosettaCM. ECL2 is built through the proposed β-sheet. New residues are shown in green. (E) A final round of FlexPepDock extended the peptide through residue 17 (green) in which it can now make contacts with loop residues. (F) In a final round of RosettaCM, the loop region connecting the β-sheet in ECL2 (green) is fully rebuilt in the presence of ghrelin 1-17.

five ghrelin-GHSR complexes. Of note, the filtering step only addressed the conformation of the ghrelin peptide backbone and ambiguities remain in the models, particularly in the GHSR loop regions and octanoyl chain (Figure 6.7).

6.4.6 Characterization of the Ghrelin/GHSR Complex

The final ensemble of models shows a tight fit in which the N-terminus of the peptide extends down into the central binding pocket of GHSR (Figure 6.6C and Figure 6.7A). This depth is similar to what is observed for small molecule-GPCR binding interactions. The N-terminus has no defined secondary structure propensity as also seen in

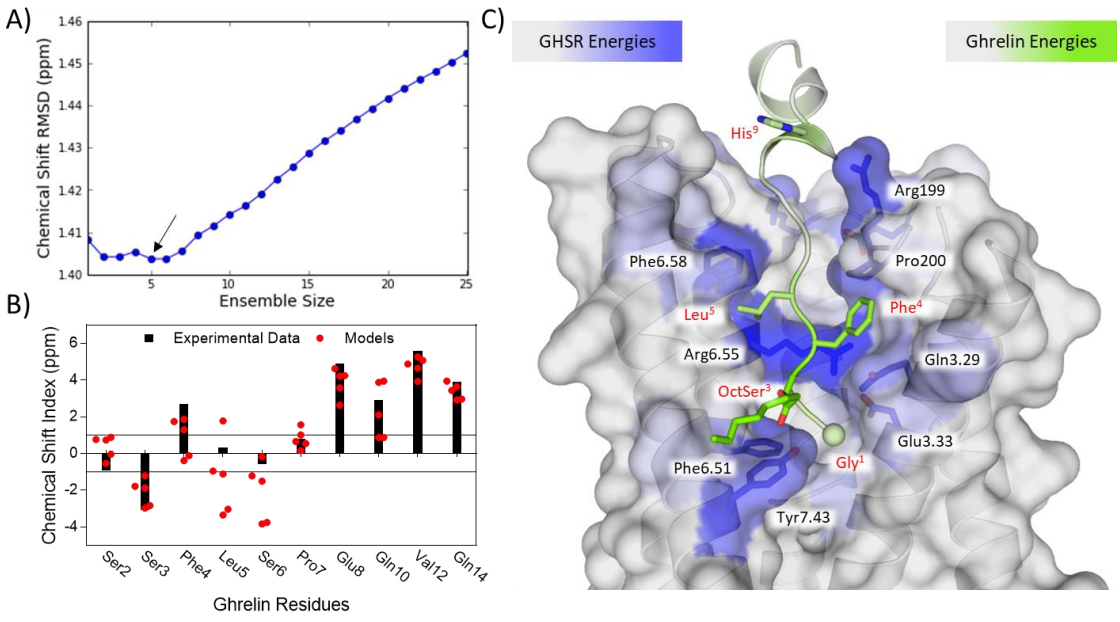


Figure 6.6 Rosetta modeling of ghrelin at GHSR. A) Filtering of models to generate a small ensemble to best represent the experimental data as measured by chemical shift RMSD in ppm. The minimum RMSD and smallest ensemble required 5 models which is highlighted with an arrow. B) Comparison of the chemical shift index between the experimental data points and the five-model ensemble. Black bars correspond to the experimentally determined CSI and red circles show the CSI for every model in the ensemble. C) A representative model of ghrelin binding in the GHSR pocket (TM 1 and 2 are removed for clarity). The residues are colored according to the Rosetta-predicted binding energies with green corresponding to strong energies in ghrelin and blue corresponding to strong energies in GHSR. The N-terminal nitrogen is shown as a sphere. The side chains shown represent those that contribute the most to binding across the interface.

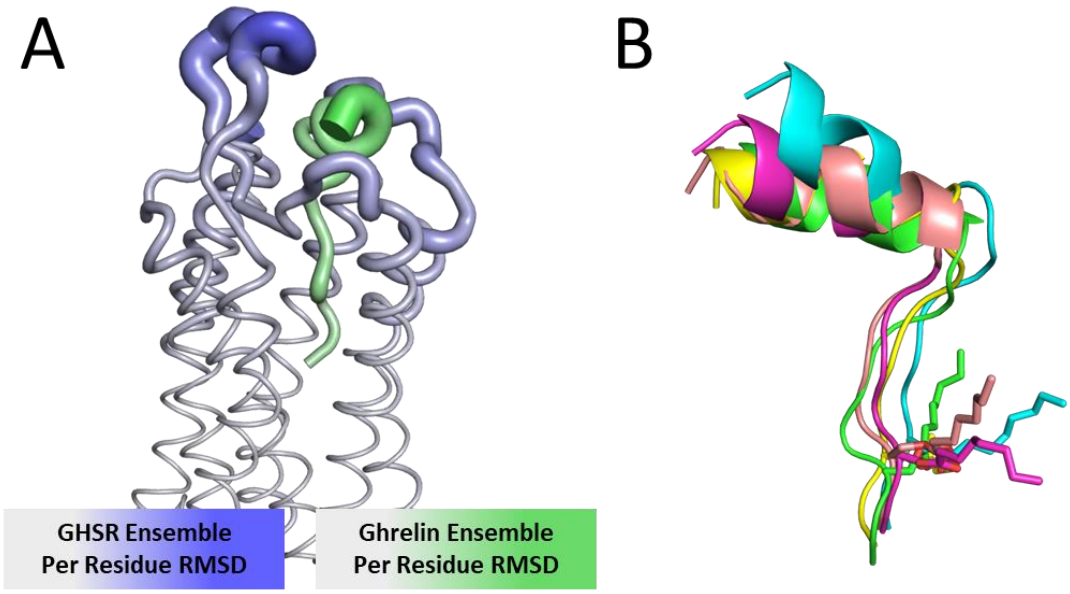


Figure 6.7 Variability of Final Ensemble of Ghrelin at GHSR. (A) A per-residue RMSD was calculated over the ensemble of five models. The thickness and color are representative of this metric with darker shades and greater thicknesses representing higher variability. (B) Close-up of the ensemble of models for the ghrelin peptide with the octanoyl chain represented in stick. The overall conformation is very similar with an extended N-terminus and an α-helix from residue 7 or 8, depending on the model, through residue 17. The orientation of the octanoyl chain is highly variable as there are no experimental data points to restrict its conformation.

the membrane-bound and solutions states. Distinct from these previously studied states, a well-defined α-helix extends from residues 8 or 9, depending on the model, through residue 17. This region of the peptide extends out of the central pocket and lies along ECL3. This is also the region where two sets of ^{13}C chemical shifts are found, suggesting some plasticity in the conformation of the bound ligand. To better understand the molecular recognition between receptor and peptide, we analyzed the energetics of binding and contributions of various residues to the binding interaction over the ensemble using a ΔΔG calculation between the energetically minimized complex and apo-states (Figure 6.6C; Table 6.4). As expected, the N-terminal binding motif of ghrelin was found to contribute significant energy to the binding interface with residues 3-5 ranking as the top three. In addition to the N-terminal motif, residues His^{9} and Arg^{11} were ranked highly in binding

energy. This is in accordance to the STD experiment in which His^{9} was shown to be tightly and closely interacting with the receptor. On the receptor, the strongest interaction partner for ghrelin was found to be Arg283$^{6.55}$ (Table 6.4). This residue has repeatedly been shown via mutagenesis studies to be critical for ghrelin signaling at GHSR [123, 347, 349, 369]. In addition, residue Glu124$^{3.33}$, which is proposed to balance the positively charged amino head group, was also found within the top ten strongest interacting residues on the receptor [123, 129, 370]. Further residues contributing significant binding energy were identified in the binding pocket along TMs 5, 6, and 7, which are common activation regions in GPCRs.

Table 6.4 Energetic analysis of the ghrelin/GHSR complex. Pairwise ΔΔG's were calculated for all interactions across the binding interface over all models in the ensemble. All pairwise energies that were determined are shown in shades of grey with darker shades referring to stronger interaction. Summation of pairwise energies for each residue in ghrelin or GHSR provides the values on the x- and y-axis, respectively. Ghrelin residue energies are shown in green and GHSR residue energies are shown in blue with darker shades corresponding to stronger energies.

		Gly1	Ser2	OctSer3	Phe4	Leu5	Ser6	Pro7	Glu8	His9	Gln10	Arg11	Val12	Gln13	Gln15	Arg15	Glu17	Receptor Energies
Phe 54	1.43			-0.29														-0.29
Cys 95	2.56			-0.21														-0.21
Met 96	2.57			-0.72														-0.72
Leu 98	2.59			-0.46														-0.46
Leu 100	2.61			-0.65														-0.65
Arg 102	2.63			-0.32		-0.92												-1.24
Leu 103	2.64			-0.84		-0.46												-1.30
Trp 109	ECL1			-0.61														-0.61
Phe 119	3.28			-0.70														-0.70
Gln 120	3.29			-0.43	-0.96													-1.39
Ser 123	3.32	-0.53																-0.53
Glu 124	3.33	-0.79		-0.79														-1.58
Thr 127	3.36	-0.62																-0.62
Ile 178	4.60				-1.29													-1.29
Leu 181	4.63				-0.29													-0.29
Val 182	ECL2				-0.76													-0.76
Glu 185	ECL2										-0.39	-0.99						-1.38
Asn 188	ECL2													-0.44				-0.44
Trp 193	ECL2									-0.31			-0.24					-0.55
Asp 194	ECL2									-0.45								-0.45
Thr 195	ECL2									-1.45								-1.45
Glu 197	ECL2							-0.41		-0.21								-0.62
Cys 198	ECL2				-0.20													-0.20
Arg 199	ECL2				-0.18		-0.73				-0.18						-1.24	-2.33
Pro 200	ECL2				-1.15						-0.32	-0.62						-2.08
Glu 202	ECL2											-1.23					-0.55	-1.78
Phe 203	ECL2															-0.19		-0.19
Ala 204	ECL2														-0.52	-0.21		-0.72
Val 205	ECL2														-0.29			-0.29
Arg 206	5.32														-1.02		-0.91	-1.93
Met 213	5.39				-0.56													-0.56
Val 216	5.42				-0.54													-0.54
Phe 220	6.44	-0.25																-0.25
Phe 279	6.51		-0.97	-1.37														-2.34
Arg 283	6.55		-1.84		-0.50	-0.95												-3.28
Phe 286	ECL3					-1.81	-0.23					-0.63						-2.68
Phe 290	ECL3								-0.40			-0.49	-0.43					-1.32
Pro 292	ECL3															-1.18		-1.18
Gly 293	ECL3								-0.17									-0.17
Ala 298	7.28								-0.26	-0.25								-0.50
Gln 299	7.29								-0.52									-0.52
Gln 302	7.32									-0.62								-0.62
Phe 309	7.39					-0.20												-0.20
Val 310	7.40			-0.38														-0.38
Tyr 313	7.43			-1.76														-1.76
Peptide Energies		-2.19	-2.80	-9.53	-6.41	-4.34	-0.96	-0.41	-1.35	-3.29	-0.88	-3.96	-0.67	-0.44	-1.82	-1.57	-2.70	

6.4.7 Validation of Binding Interface via Mutagenesis

Many of the top 15 interacting residues in GHSR have been previously investigated confirming our localization of the peptide within a central cavity (Table 6.5). In particular Glu124$^{3.33}$ and Gln120$^{3.29}$ in TM3, that in our models show interactions with Gly^{1} and Phe^{4}, displayed a dramatic loss of signal transduction after mutagenesis [122, 123, 129, 347]. To shed additional light on the structure of the binding complex, receptor mutants of the top 15 residues were tested in Ca^{2+} signaling and IP_3 accumulation assays. Alanine mutation of the top three interacting residues Arg283$^{6.55}$, Phe286$^{6.58}$, and Phe279$^{6.51}$ exhibited a strong impact on the receptor activation profiles (Figure 6.8). In accordance with previous studies, Arg283$^{6.55}$Ala and Phe279$^{6.51}$Ala revealed a complete

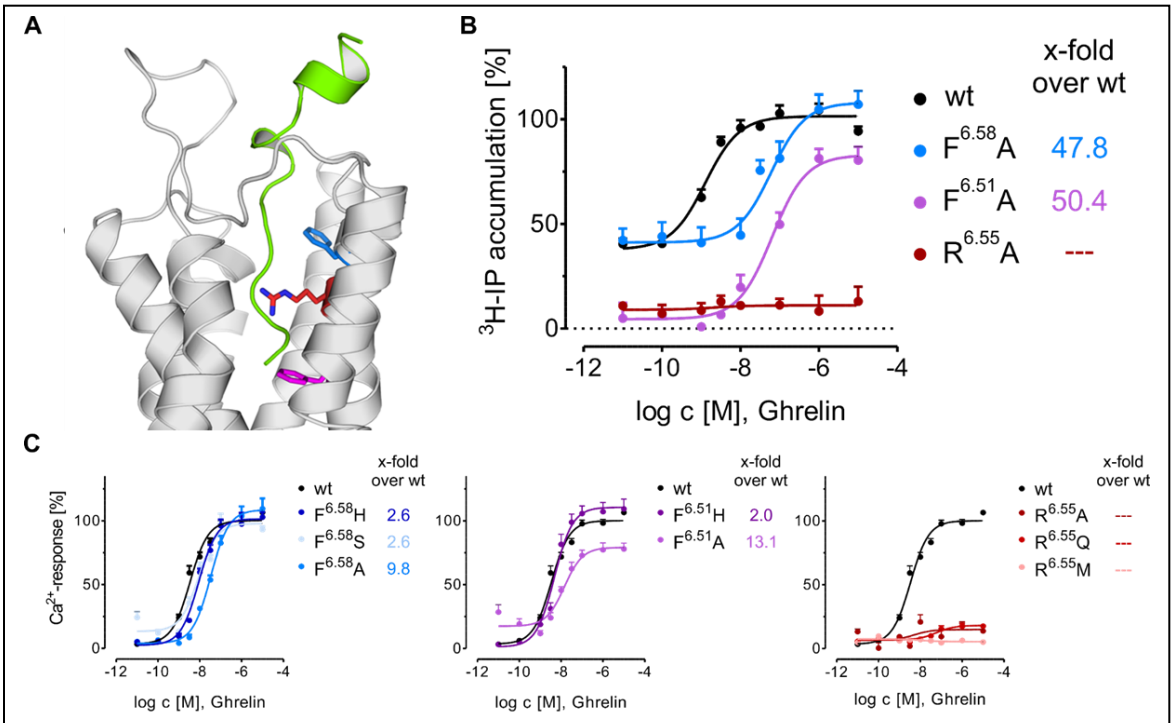


Figure 6.8 Impact of the residues with the highest Rosetta energy score to G-protein signal transduction. A) Position of the residues with the highest Rosetta energy score in the ghrelin/GHSR complex. GHSR show in grey, ghrelin in green, and side chains tested colored according to panels B and C. B) Concentration-response curves of GHSR mutants determined by IP accumulation assay in transiently transfected COS7 cells. Data are means ± SEM of ≥ 2 independent experiments performed in duplicates. C) Ca^{2+} mobilization assay of selected GHSR1a mutants in transiently transfected HEK293 cells.

loss of constitutive activity [347], whereas Phe286$^{6.58}$Ala retained wildtype-like behavior. Additionally, all of these alanine variants of GHSR led to a substantially decreased activation after stimulation with ghrelin. Compared to the wildtype GHSR, Phe286$^{6.58}$Ala and Phe279$^{6.51}$Ala showed a 50-fold reduction in activity and Arg283$^{6.55}$Ala led to a complete loss of ghrelin-induced receptor activation. However, Phe286$^{6.58}$ and Phe279$^{6.51}$ could be partly rescued by a less disruptive histidine mutation suggesting that a bulky hydrophobic sidechain is critical in these positions. According to the models, these two residues pack alongside hydrophobic segments of ghrelin including octanoyl-Ser^{3} and Leu^{5}. Another residue suggested to interact with the octanoyl chain of Ser^{3} is Tyr313$^{7.43}$.

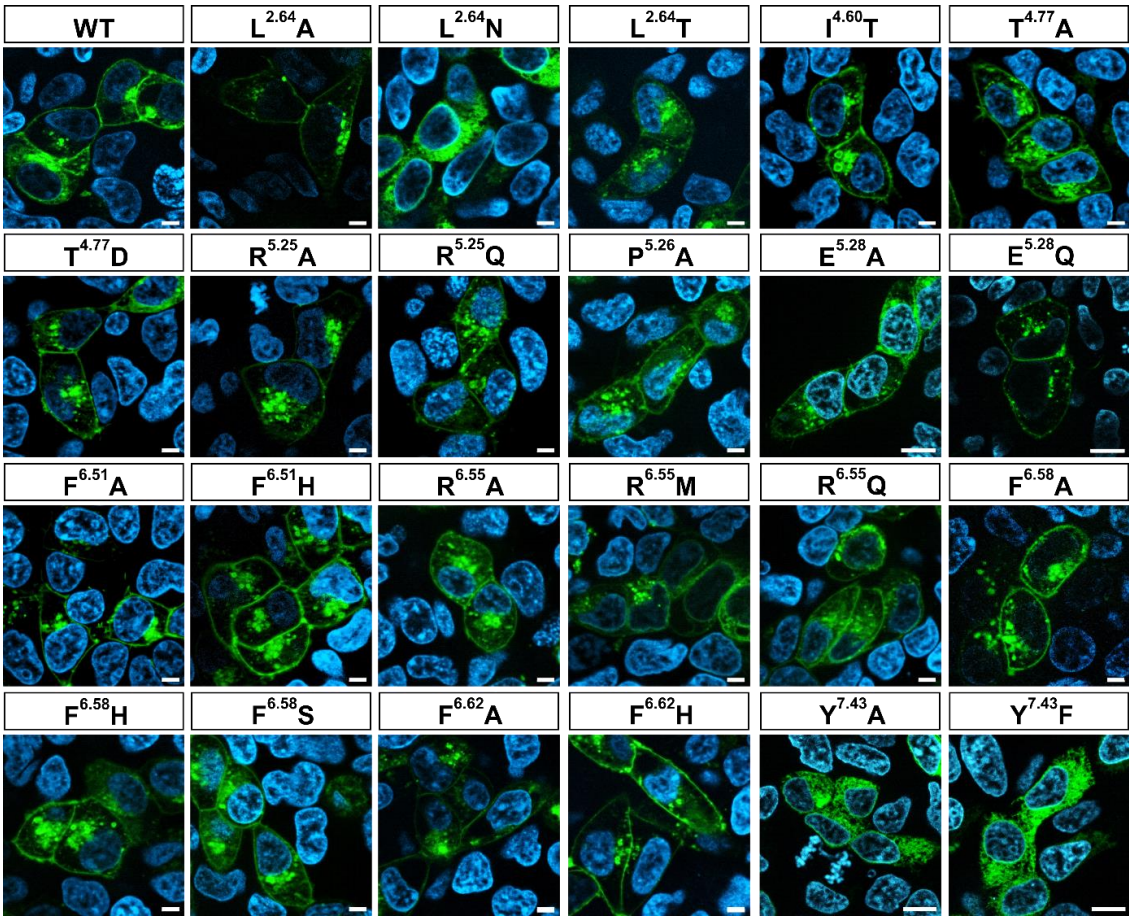


Figure 6.9 Fluorescence microscopy of transiently transfected HEK293 with GHSR eYFP fusion proteins. Green color represents eYFP-tagged receptor fluorescence, nuclei were stained with Hoechst33342 and are shown in cyan. Scale bar = 10 µm. All constructs but L2.64N, Y7.43A and Y7.43F were localized in the cell membrane.

However, mutation to both alanine and the very conservative phenylalanine resulted in loss of trafficking of the receptor to the plasma membrane suggesting this residue to be critical to the structural integrity of the receptor in general (Figure 6.9). All mutations in ECL2 including Thr195, Arg199, Pro200, and Glu202 had no effect on ghrelin signaling correlating with previous knowledge [352]. Ile178$^{4.60}$ which may interact with Phe^{4} in ghrelin displayed only a 2-fold loss of potency for alanine [347] but a 15-fold loss of potency for threonine (Table 6.5), suggesting a hydrophobic interaction between receptor and ligand. Lastly, mutation of Leu103$^{2.64}$ in TM2 near the top of the helical bundle is accepted for alanine, however an increase in polarity to a threonine resulted in a 100-fold loss of signaling while an asparagine yielded no localization of the receptor to the plasma membrane. Taken together, these mutations highlight the likelihood that our structural model represents a relevant binding pose of ghrelin at its cognate receptor.

Table 6.5 Mutational analysis of GHSR binding site.

reference: 1) [347]; 2) [353]; 3) [129]; 4) [352]; 5) new data

† method: a) IP3 accumulation/COS7; b) SRE-Luc/HEK293; c) Ca^{2+}/HEK293

Residue		Rosetta Energy	Mutation	x-fold over wt #,†
R283	6.55	-3.28	A	loss [1a. 5a, 5c]
			Q	loss [5c] / 55.9 [3a]
			M	loss [5c]
F286	6.58/ECL3	-2.68	A	47.8 [5a] / 9.8 [5c] / 25.0 [3a]
			H	2.6 [5c]
			S	2.6 [5c]
F279	6.51	-2.34	A	50.4 [5a] / 13.1 [5c] / 35.3 [3a]
			H	2.0 [5c]
			L	0.5 [2b]
R199	5.25/ECL2	-2.33	A	1.4 [5c] / 0.7 [4a]
			Q	1.7 [5c]
			E	2.1 [4a]
			L	1.4 [4a]
P200	5.26/ECL2	-2.08	A	1.7 [5c] / 0.8 [4a]

R206	5.32	-1.93	A	0.9 [4a]
			E	0.9 [4a]
			K	3.5 [4a]
			Q	4.1 [4a]
E202	5.28/ECL2	-1.78	A	2.9 [5c] / 0.9 [4a]
			Q	2.3 [5c] / 1.5 [4a]
			R	1.9 [4a]
Y313	7.43	-1.76	A	no surface expression [5]
			F	no surface expression [5]
E124	3.33	-1.58	Q	153.0 [3a]
T195	4.77/ECL2	-1.45	A	1.6 [5c]
			D	1.4 [5c]
Q120	3.29	-1.39	A	6.8 [3a]
			L	loss [1a]
E185	ECL2	-1.38	A	2.1 [4a]
			D	2.2 [4a]
			Q	1.0 [4a]
F290	6.62/ECL3	-1.32	A	2.9 [5a]
			H	1.0 [5a]
L103	2.64	-1.3	A	1.4 [5a]
			N	no surface expression [5]
			T	n.d. [5a]
I178	4.60	-1.29	A	2.4 [3a]
			T	15.2 [5a]

6.5 Discussion

6.5.1 Ghrelin Binds to GHSR via an Extended Binding Surface

Our models of ghrelin in complex with GHSR depict binding from the lower reaches of the orthosteric binding pocket outward along TMs 6 and 7 to ECL3. Our STD NMR studies found several residues to be clearly interacting with the receptor. As STD is limited to bound epitopes within ~5 Å of the receptor, these signals suggest a specific binding interaction [366]. Of note, residues Ser^3 and Phe^4 had a strong STD signal indicating tight interaction with the receptor. These residues were also identified as critical

for ghrelin association with the membrane [7]. Given that the signal for the octanoyl chain on Ser^3 was maximally polarized in the saturation transfer, it provides evidence for the location of the lipid chain to lay within the interior of the protein where it would be tightly bound as opposed to lying peripherally to the receptor in the membrane or exposed to solvent. Additionally, His^9 was identified via these STD studies to be interacting with the receptor. This residue is unique in that it is outside the first five residues that comprise the minimal ghrelin signaling peptide [117]. This identifies a secondary region of the peptide that is needed for full signaling potential. Two-site binding modes have been evidenced in the few available peptide-bound GPCR crystal structures [50, 166, 356, 357]. This secondary site may explain why the minimal 5mer peptide could not compete full length ghrelin in vivo [359]. Our modeling experiments also found His^9 to be strongly interacting with the receptor as measured in our energy calculations. Interestingly, the adjacent residue Glu^8 was found to lay in a fairly hydrophobic pocket at the top of TM7. The agrees with previous studies that showed mutation of this residue from the negatively charged Glu to hydrophobic Ala or Tyr increases the affinity for the peptide to the receptor [127]. Overall, our model shows that while the N-terminal core signaling motif binds within the central cavity in GHSR, the central α-helix including His^9 lays alongside and interacts with the extracellular loops. There is increased flexibility in the ensemble as to the exact interaction of this α-helix (Figure 6.7A) which may account for why we could not determine any high-specificity interactions between this α-helical region of ghrelin and the extracellular loops via mutagenesis. However, the sum total of interactions provides a strong signaling potential for the receptor as noted by enhanced activity for peptides from 4 up to 10 residues in length [118]. This is very similar to what was observed in the endothelin receptor in which known crystal contacts between ET-B receptor and endothelin-1 were

mutated while only yielding a maximal IC_{50} change of about 10-fold [43]. At present, the available data converge on a model in which this α-helix extends between ECLs 2 and 3 and away from the receptor N-terminus, unlike what has been seen in larger peptide ligands such as the chemokines [25]. Additionally, our use of an *E. coli* expression system suggests that no additional post-translational modifications are necessary beyond the ghrelin lipidation to achieve receptor-ligand binding. While additional data may become present in the future, we remain confident in the model given the current known experimental results.

6.5.2 The Binding Model of Ghrelin Adopts a Fairly Small Structural Ensemble

Using the chemical shift differences collected over selected residues within the ghrelin peptide, we filtered the structural models to generate an ensemble that best represents the experimental data. Five peptide models were identified that collectively satisfied the NMR restraints. This number was particularly low compared to the ensemble of 22 models needed to best represent the membrane-bound structure as determined previously [7]. The five peptide models were fairly similar in conformation (Figure 6.7B) again contrasting with the membrane-bound ensemble with highly diverse backbone structures. For residues with multiple chemical shifts, we examined the secondary structure propensities individually. Glu^{8} and Gln^{14} display α-helical conformations in both populations and therefore we ensured all models had α-helical character at these positions. Pro^{7} shows one population with distinct α-helical character and a second population of random coil. Our resultant ensemble shows that this proline is important for initiating the α-helix at the following position. Of note, we chose not to model the flexible C-terminus of ghrelin from residues 18-27 as these residues are unstructured and do not contribute to

receptor binding [117]. Addition of these residues would likely increase the ensemble size and conformational space though provide little additional information on the binding interaction between the peptide and receptor. Compared to the membrane-bound state, the flexible N-terminus is highly converged to a single conformation in the receptor-bound state. This is likely due to the presence of the receptor which acts as a constraint to the space sampled by the peptide. The specific interaction between Gly^{1} and $Glu124^{3.33}$ lock the N-terminus in place with only a small space left for the bulky hydrophobic groups of octanoyl-Ser^{3} and Phe^{4} to position. This results in a well-defined pocket that can accommodate only small variations in the backbone structure. Evidence for this defined N-terminal conformation has been confirmed in complementary NMR studies [371]. This phenomenon is also seen in the structure of leukotriene B4 which is highly flexible in solution but adopts a well-defined structure when bound to its cognate receptor [372]. Of note, the chemical shift information was only used to filter the backbone conformation and does not provide information on the side chain positions. As a result, the geometry of the octanoyl chain cannot be structured in the model with high confidence (Figure 6.7B).

6.5.3 A Hierarchical Approach to Modeling Highly Flexible Protein-Ligand Complexes in Rosetta

We introduce here a new approach for modeling protein-ligand complexes when both binding partners have a high degree of flexibility. A unique challenge of peptide ligands as compared with small molecule ligands is the large degree of freedom associated with each amino acid. This is evidenced by the low number of crystal structures of peptide-binding GPCRs in which the native peptide and not a small molecule antagonist is present [43, 368]. Additionally, GPCRs possess a vast degree of internal flexibility as seen in

spectroscopic studies that measure population states of the receptor in apo- and ligand-bound forms [348, 373-378]. One of the more difficult regions to model in GPCRs is the extracellular loops due to this intrinsic flexibility [257]. Given the high number of interactions between peptide ligands and the extracellular loops, it is expected that the conformation of these loops would differ between the apo- and ligand-bound states. This was recently shown in crystal structures of the C5a receptor in which ECL2 adopted a different conformation in a small molecule-bound state as compare to a peptide-bound state [45, 50]. Additionally, it is anticipated that the presence of these loops would only be compatible with distinct conformations of the peptide, a sort of cross-conformational selection. To account for these properties, we developed an iterative approach in which we built and docked the peptide into the receptor. While no new code was developed for this method, the iterative use of homology modeling and flexible peptide docking of progressively longer segments of the target peptide is a novel technique not previously seen in Rosetta modeling studies of peptide-binding GPCRs. This approach, though based in Monte Carlo random sampling, was significantly guided by our understanding of GPCR biology and the ghrelin system. Each step in the protocol allowed for dense sampling of flexible components at the receptor binding pocket. Further, the use of RosettaCM homology modeling of loop conformations ensures that we only build loops with relevant structure to known GPCRs. We felt this provided a more focused sampling to simultaneously identify conformations of both the peptide and receptor that would be compatible with ligand-binding as opposed to the vast conformational space that would be allowed in modeling of either component individually. We also opted out of rigid body docking of the peptide, as is done for small molecule ligands or large proteins, for similar reasons as we wanted the peptide to fold into the binding pocket. We believe the approach

presented here will be useful for ligand-GPCR modeling in general with particular emphasis in complexes that require extensive remodeling of the extracellular loops for ligand binding to occur as is the case for peptides and protein ligands.

6.6 Conclusions

Currently, there is a lack of knowledge how endogenous peptide ligands bind to their cognate receptors. While the rate of deposited structures is increasing for peptide-binding GPCRs, structures containing a bound peptide remain sparse. The methodologies presented here combine orthogonal structural methods to derive a model that best represents all available data. NMR allows the study of the receptor in its native state absent of fusion proteins and with full length termini. Computational modeling rapidly eliminates unlikely solutions and allows researchers to focus on native-like structures. Use of these two methods resulted in a novel understanding of ghrelin binding to GHSR. Further, this model adds to the growing structural evidence of multi-site binding modes within peptide-binding GPCRs and has potential to increase our knowledge of this important family of GPCRs.

CHAPTER 7

Structural Basis of Ligand Binding Modes at the Neuropeptide Y Y_1 Receptor

7.1 Summary

Chapter 7 describes the structural characterization of the neuropeptide Y Y_1 receptor (Y_1R) binding to small molecule antagonists and full-length endogenous peptide agonist. Crystal structures of Y_1R bound antagonists were collected by collaborators, and I assisted in the refinement. The binding of the NPY agonist peptide was modeled by my previously described iterative docking method (Chapter 6) based on these crystal structures and incorporation of experimental data including NMR, photo-crosslinking, and mutation. Interestingly, we demonstrate the even though the agonist and antagonists have highly similar structures, they bind in different manners at the receptor. This work utilized the protocols developed in every chapter of this thesis except for Chapter 4. This chapter comes from the article "Structural Basis for Ligand Binding Modes at the Neuropeptide Y Y_1 Receptor" for which I am a co-first author [44].

7.2 Introduction

Neuropeptide Y (NPY) receptors belong to the G protein-coupled receptor (GPCR) superfamily and have important roles in food intake, anxiety and cancer biology [379, 380]. The NPY-Y receptor system has emerged as one of the most complex networks with

three peptide ligands (NPY, peptide YY and pancreatic polypeptide) binding to four receptors in most mammals, namely Y_1, Y_2, Y_4 and Y_5 receptors, with different affinity and selectivity [381]. NPY is the most powerful stimulant of food intake and this effect is primarily mediated by the Y_1 receptor (Y_1R)[382]. A number of peptides and small-molecule compounds have been characterized as Y_1R antagonists and have shown clinical potential in the treatment of obesity [382], tumor [379] and bone loss [383]. However, their clinical usage has been hampered by low potency and selectivity, poor brain penetration ability or lack of oral bioavailability [384]. Here we report crystal structures of the human Y_1R bound to two selective antagonists UR-MK299 and BMS-193885 at 2.7 and 3.0 Å resolution, respectively. The structures combined with mutagenesis studies reveal binding modes of Y_1R to several structurally diverse antagonists and the determinants of ligand selectivity. The Y_1R structure and molecular docking of the endogenous agonist NPY, together with nuclear magnetic resonance (NMR), photo-crosslinking and functional studies, provide insights into the binding behavior of the agonist and for the first time, to our knowledge, determine the interaction of its N terminus with the receptor. These insights into Y_1R can enable structure-based drug discovery that targets NPY receptors.

7.3 Methods

7.3.1 Cloning and protein expression

DNA sequence of wild-type human Y_1R was optimized and synthesized by Genewiz and then cloned into a modified **pFastBac1** vector (Invitrogen), which contains an expression cassette with a haemagglutinin (HA) signal sequence followed by a Flag tag prior to the receptor at the N terminus and a PreScission protease site followed by a

10×His-tag at the C terminus. An engineered construct was generated by inserting a modified T4 Lysozyme (T4L)[385] at the third intracellular loop (ICL3) between residues R241 and D250 and introducing a mutation $F129^{3.41}W$[386]. Twenty-five amino acids (V359-I384) were truncated at the C terminus to further improve protein yield and stability. Bac-to-Bac Baculovirus Expression System (Invitrogen) was used to generate high-titer (>10^8 viral particles per ml) recombinant baculovirus. *Spodoptera frugiperda* (*Sf*9) cells (Invitrogen) at density of 2×10^6 cells per ml were infected by viral stock at MOI (multiplicity of infection) of 5. In company with the virus, a ligand (UR-MK299 or BMS-193885) was added to the cell culture to a final concentration of 1 μM. Transfected cells were cultured at 27 ºC for 48 h and then collected by centrifugation and stored at −80 ºC until use.

7.3.2 Purification of Y_1R–UR-MK299 and Y_1R–BMS-193885 Complexes

Frozen insect cells expressing the Y_1R–UR-MK299 complex were disrupted with thawing and repeated dounce homogenization in a hypotonic buffer containing 10 mM HEPES, pH 7.5, 10 mM $MgCl_2$, 20 mM KCl and protease inhibitor cocktail (Roche). After centrifugation at 160,000 *g* for 30 min, cell debris was re-suspended in a high osmotic buffer (10 mM HEPES, pH 7.5, 1 M NaCl, 10 mM $MgCl_2$, 20 mM KCl) followed by extensive homogenization. Soluble and membrane associated proteins were removed from the suspension by centrifugation. This procedure was repeated for 2-3 more times and then the hypotonic buffer was used to remove the high concentration of NaCl. Purified membranes were re-suspended in the hypotonic buffer with additional 30% (v/v) glycerol and stored at −80 ºC until use.

Purified membranes were thawed on ice in the presence of 100 μM UR-MK299, 2

mg ml^{-1} iodoacetamide (Sigma) and EDTA-free protease inhibitor cocktail (Roche) and incubated at 4 ºC for 1 h. Equal volume of solubulization buffer containing 100 mM HEPES, pH 7.5, 1 M NaCl, 1% (w/v) n-dodecyl-b-D-maltopyranoside (DDM, Anatrace), 0.2% (w/v) cholesterol hemisuccinate (CHS) (Sigma) was added and incubation was continued for additional 3 h. The supernatant was isolated by centrifugation at 160,000 *g* for 30 min and incubated with TALON resin (Clontech) supplemented with 10 mM imidazole, pH 7.5 at 4 ºC overnight. The resin was then washed with ten column volumes of 25 mM HEPES, pH 7.5, 500 mM NaCl, 0.05% (w/v) DDM, 0.01% (w/v) CHS, 10% (v/v) glycerol, 30 mM imidazole and 50 μM UR-MK299, followed by ten column volumes of 50 mM HEPES, pH 7.5, 500 mM NaCl, 0.05% (w/v) DDM, 0.01% (w/v) CHS, 10% (v/v) glycerol, 10 mM $MgCl_2$, 5 mM ATP and 50 μM UR-MK299 and five column volumes of 25 mM HEPES, pH 7.5, 500 mM NaCl, 0.05% (w/v) DDM, 0.01% (w/v) CHS, 10% (v/v) glycerol and 50 μM UR-MK299. The protein was eluted by five column volumes of 25 mM HEPES, pH 7.5, 500 mM NaCl, 0.05% (w/v) DDM, 0.01% (w/v) CHS, 10% (v/v) glycerol, 300 mM imidazole and 100 μM UR-MK299. A PD MiniTrap G-25 column (GE healthcare) was used to remove imidazole. The C-terminal His-tag and glycosylation was then treated by overnight digestion with His-tagged PreScission protease (custom-made) and His-tagged PNGase F (custom-made). Ni-NTA super flow resin (Qiagen) reverse binding was performed to remove the PreScission protease, PNGase F and the cleaved His-tag. The purified Y_1R–UR-MK299 complex was collected and concentrated to 20-30 mg ml^{-1} with a 100 kDa molecular weight cutoff concentrator (Sartorius Stedim Biotech). Receptor purity and monodispersity were estimated by SDS-PAGE and analytical size-exclusion chromatography (aSEC).

The Y_1R–BMS-193885 complex protein was purified following the same procedure

as above. The membranes of Y_1R construct were incubated with 50 μM BMS-193885, 2 mg ml^{-1} iodoacetamide (Sigma), and EDTA-free protease inhibitor cocktail (Roche) at 4 ºC for 1 h, and then solubilized in final concentration of 50 mM HEPES, pH 7.5, 500 mM NaCl, 0.5% (w/v) DDM, 0.1% (w/v) CHS, 10% glycerol and 25 μM BMS-193885 at 4 ºC for 3 h. The solubilized Y_1R–BMS-193885 complex bound to the TALON resin was washed with ten column volumes of 25 mM HEPES, pH 7.5, 500 mM NaCl, 0.05% (w/v) DDM, 0.01% (w/v) CHS, 10% (v/v) glycerol, 30 mM imidazole and 25 μM BMS-193885, followed by ten column volumes of 50 mM HEPES, pH 7.5, 500 mM NaCl, 0.05% (w/v) DDM, 0.01% (w/v) CHS, 10% (v/v) glycerol, 10 mM $MgCl_2$, 5 mM ATP and 25 μM BMS-193885 and five column volumes of 25 mM HEPES, pH 7.5, 500 mM NaCl, 0.05% (w/v) DDM, 0.01% (w/v) CHS, 10% (v/v) glycerol and 25 μM BMS-193885. The protein was eluted by five column volumes of 25 mM HEPES, pH 7.5, 500 mM NaCl, 0.05% (w/v) DDM, 0.01% (w/v) CHS, 10% (v/v) glycerol, 300 mM imidazole and 50 μM BMS-193885. The eluted sample was concentrated and desalted using the PD MiniTrap G-25 column (GE healthcare). Overnight digestion by Precission protease and PNGase F and Ni-NTA reverse binding were then performed to further purify the protein. The complex protein was concentrated to 10-20 mg ml^{-1} and analyzed by SDS-PAGE and aSEC.

7.3.3 Lipidic Cubic Phase Crystallization of Antagonist-Bound Y_1Rs

The Y_1R sample in complex with UR-MK299 or BMS-193885 was mixed with molten lipid (monoolein/cholesterol 10:1 by mass) at a weight ratio of 1:1.5 (protein:lipid) using two syringes to create lipidic cubic phase (LCP). The mixture was dispensed onto glass sandwich plates (Shanghai FAstal BioTech) in 40 nl drop and overlaid with 800 nl precipitant solution using a Gryphon robot (Art-Robbins). Protein reconstitution in LCP

and crystallization trials were performed at room temperature (19-22 ºC). Plates were placed in an incubator (Rock Imager, Formulatrix) and imaged at 20 ºC automatically following a schedule. Crystals of Y_1R–UR-MK299 complex showed up after 4 days and grew to the full size (150 × 50 × 5 μm^3) within two weeks in 0.1 M Tris, pH 7.4-8.0, 30-40% (v/v) PEG 400, 50-150 mM sodium tartrate and 100 μM UR-MK299. The Y_1R–BMS-193885 complex was crystallized in 0.1 M HEPES, pH 7.2-7.6, 20% PEG 400 and 50 μM BMS-193885 with the maximum size of 30 × 10 × 5 μm^3. The crystals of Y_1R–UR-MK299 and Y_1R–BMS-193885 complexes were harvested directly from LCP using 150 μm and 50 μm micro mounts (M2-L19-50/100, MiTeGen), respectively, and flash frozen in liquid nitrogen.

7.3.4 Data Collection and Structure Determination

X-ray diffraction data were collected at the SPring-8 beam line 41XU, Hyogo, Japan, on a Pilatus3 6M detector (X-ray wavelength 1.0000 Å). Crystals were exposed with a 10 μm × 8 μm mini-beam for 0.2 s and 0.2º oscillation per frame. Data from 47 best-diffracting crystals of the Y_1R–UR-MK299 complex and 33 crystals of the Y_1R–BMS-193885 complex were processed by XDS [387], respectively. Structure of the Y_1R–UR-MK299 complex was solved by molecular replacement (MR) implemented in Phaser [388] using the receptor portion of NTSR1 (PDB ID: 4GRV), converted to polyalanines, and T4L structure (PDB ID: 1C6P) as search models. The correct MR solution contained one Y_1R-T4L molecule in the asymmetric unit. Initial refinement was performed with REFMAC5 [389] and BUSTER [390], and then manual examination and rebuilding of the refined coordinates were carried out in COOT [391] using both |2Fo| - |Fc| and |Fo| - |Fc| maps. The structure has been carefully refined and the ramachandran plot analysis

indicates that 100% of the residues are in favorable (95.5%) or allowed (4.5%) regions (no outliers). The final model includes 312 residues (F18-R241 and D250-F337) of the 384 residues of Y_1R and residues N2-Y161 of T4L. The Y_1R–BMS-193885 complex structure was solved using Y_1R in the Y_1R–UR-MK299 complex and T4L as search models and refined under the same procedure. The ramachandran plot analysis indicates that 100% of the residues are in favorable (95.4%) or allowed (4.6%) regions (no outliers). The final model of the Y_1R–BMS-193885 complex contains 301 residues (D31-R241 and D250-D339) of Y_1R and the 160 residues of T4L. Helix VIII in the Y_1R–UR-MK299 structure rotates towards helix VI by about 90° compared to that in the BMS-193885-bound

Table 7.1 Data collection and refinement statistics.

* Diffraction data from 47 Y_1R–UR-MK299 crystals and 33 Y_1R–BMS-193885 crystals were used to solve the structures.

† Numbers in parentheses refer to the highest-resolution shell.

	Y_1R–UR-MK299	Y_1R–BMS-193885
Data Collection*		
Space group	$P2_1$	$C222_1$
Cell dimensions		
a, *b*, *c* (Å)	37.8, 100.7, 83.2	76.9, 126.8,170.3
α, β, γ (°)	90.0, 98.8, 90.0	90.0, 90.0, 90.0
Resolution (Å)	50.0-2.7 (2.8-2.7)†	50.0-3.0 (3.1-3.0)
R_{merge} (%)	23.7 (83.9)	16.6 (93.7)
$I / \sigma I$	5.01 (0.71)	5.30 (1.00)
Completeness (%)	93.0 (87.0)	92.4 (79.2)
Redundancy	7.4 (2.3)	3.5 (2.6)
Refinement		
Resolution (Å)	50.0-2.7	50.0-3.0
No. reflections	15,818 (758)	15,600 (797)
R_{work} / R_{free} (%)	22.4 / 26.6	22.4 / 24.9
Number of atoms		
Protein	3,791	3,654
Ligand	45	43
Overall *B* values (Å^2)		
Protein	88.3	108.0
Ligand	66.4	81.0
R.m.s. deviations		
Bond lengths (Å)	0.009	0.009
Bond angles (°)	1.01	1.00

structure most likely due to crystal packing (Table 7.1).

7.3.5 Immunoblotting

Total solubilized protein of *Sf*9 membrane preparations (see above) used in radio ligand binding assay was determined by the Bradford method according to the manufacturers' protocol (BioRad Protein Assay; BioRad, Munich, Germany). Aliquots of homogenized membrane preparations, corresponding to 100 µg of protein, were centrifuged at 50,000 *g* at 4 °C for 15 min, and the pellets were re-suspended in 50 mM Tris, pH 7.4, supplemented with 1 mM EDTA and protease inhibitors (SIGMAFAST Protease Inhibitor cocktail tablets, Sigma) at a protein concentration of 1,600 µg ml^{-1}. Membrane homogenates (15 µl) were processed and subjected to immunoblotting as described previously [392] with the following modifications: blotting onto the nitrocellulose membrane was performed at 60 mA for 60 min. Primary antibody ANTI-FLAG M1 from mouse (Sigma) was diluted 1:500. The secondary antibody, an anti-mouse IgG HRP-conjugated antibody from goat (ThermoFisher Scientific) was diluted 1:10,000. The washing steps after incubation with the primary and the secondary antibody were 3 × 10 min each. Control experiments in the absence of the primary antibody were not performed.

7.3.6 Radio Ligand Binding Assay

All binding experiments with [^{3}H]-UR-MK299 (synthesis described elsewhere [393]) were performed at *Sf9* membrane preparations in PP 96-well microplates (Greiner bio-one) at 23±1 °C using a sodium-containing, iso-osmotic HEPES buffer (10 mM HEPES, pH 7.4, 150 mM NaCl, 5 mM KCl, 2.5 mM $CaCl_2$, 1.2 mM KH_2PO_4, 1.2 mM

Mg_2SO_4 and 25 mM $NaHCO_3$ supplemented with 1% BSA) for competition binding studies with antagonists, and a sodium-free, hypo-osmotic HEPES buffer (25 mM HEPES, pH 7.4, 2.5 mM $CaCl_2$ and 1 mM $MgCl_2$ supplemented with 1% BSA) for competition binding studies with the agonist NPY (in the following, both buffers are referred to as 'binding buffer'). Prior to competition binding experiments, K_d values of [^{3}H]-UR-MK299 were determined by saturation binding using the respective binding buffer. In case of saturation binding experiments, [^{3}H]-UR-MK299 was 1:1 diluted with 'cold' UR-MK299 (in the following, the mixture is referred to as 'radioligand'). On the day of the experiment, *Sf*9 membranes were thawed and re-suspended using a 1-ml syringe equipped with a needle (20G) followed by centrifugation at 16,000 *g* at 4 °C for 10 min. The supernatant was discarded and the pellets were re-suspended in binding buffer using a 1-ml syringe equipped with a needle (27G3/4). The membrane homogenates were stored on ice until use. The total amount of protein per well was between 0.25 and 8 μg, depending on the receptor expression level.

Saturation binding experiments: for the determination of total binding, wells were pre-filled with binding buffer (160 μl), followed by the addition of binding buffer (20 μl), containing the radioligand at a 10-fold higher concentration compared to the final concentration. For the determination of unspecific binding (in the presence of UR-MK299 at a 100-fold excess), wells were pre-loaded with binding buffer (140 μl), binding buffer (20 μl) containing UR-MK299 (10-fold concentrated) and binding buffer (20 μl) containing the radioligand (10-fold concentrated). To all wells, 20 μl of the membrane suspension were added, and the plates were shaken at 23 °C for 90 min. The membranes were collected on GF/C filter mats (0.26 mm; Whatman, Maidstone, UK) (pre-treated with 0.3% polyethylenimine for 30 min) and washed with cold Tris buffer (91 g l^{-1} Tris base, 25.5 g

l^{-1} $MgCl_2 \cdot 6H_2O$ and 3.76 g l^{-1} EDTA) using a Brandel Harvester (Brandel, Gaithersburg, MD). Filter pieces were punched out and transferred into 1450-401 96-well plates (PerkinElmer, Rodgau, Germany). Rotiscint eco plus (Carl Roth, Karlsruhe, Germany) (200 µl) was added, and the plates were sealed with transparent tape (permanent seal for microplates, PerkinElmer), vigorously shaken for at least 3 h, and kept in the dark for at least 1 h prior to the measurement of radioactivity (dpm) with a MicroBeta2 plate counter (PerkinElmer). Specific binding data (dpm) were plotted against the free radioligand concentration in nM (obtained by subtracting the amount of bound radioligand (nM) (calculated from the specifically bound radioligand in dpm, the specific activity, and the volume per well) from the total radioligand concentration (nM)) and analyzed by a two-parameter equation describing hyperbolic binding (SigmaPlot 11.0, Systat Software Inc., Chicago, IL) to obtain K_d and B_{max} values. In case of K_d values < 1 nM, the B_{max} was kept below 1,200 dpm by choosing an appropriate protein concentration. In case of K_d values > 1 nM, the B_{max} was kept below 3,300 dpm.

Competition binding experiments were performed according to the procedure for saturation binding with the following modifications: [^{3}H]-UR-MK299 was used undiluted and in case of Y_1R mutants, at which [^{3}H]-UR-MK299 exhibited a K_d value > 3 nM, the total volume per well was 100 µl, i.e., in the case of total binding, wells were pre-filled with binding buffer (80 µl), followed by the addition of binding buffer (10 µl) containing [^{3}H]-UR-MK299 (10-fold concentrated), and the membrane homogenate (10 µl). The following concentrations of [^{3}H]-UR-MK299 were used for competition binding with antagonists: 0.2 nM (wild-type Y_1R, T280$^{6.52}$A, T212$^{5.39}$A), 0.3 nM (F173$^{4.60}$W), 1.1 nM (L279$^{6.51}$A), 5 nM (Q219$^{5.46}$A), 7 nM (L215$^{5.42}$G), 10 nM (I124$^{3.36}$A, F173$^{4.60}$A). 1 nM [^{3}H]-UR-MK299 was used for competition binding with NPY. The incubation time was 90

min throughout. Unspecific binding was determined in the presence of UR-MK299 (100-fold excess to [^{3}H]-UR-MK299). Total binding was between 700 and 3,500 dpm. Maximum unspecific binding amounted to ≤ 30% of total binding. Specific binding data (dpm) were plotted against log (concentration competitor) and analyzed by a four-parameter logistic equation (log(inhibitor) vs response - variable slope, GraphPad Prism Software 5.0, GraphPad Software, San Diego, CA) to obtain pIC_{50} values, which were converted to IC_{50} values. In case of incomplete displacement of [^{3}H]-UR-MK299 (specifically bound radioligand at the highest competitor concentration between 20% and 50%), pIC_{50} values were determined by plotting $\log(B/(B_0-B))$ (Hill plot; B = specifically bound radioligand in the presence of competitor (values between 10 and 90%), B_0 = specifically bound radioligand in the absence of competitor (B_0 = 100%)) against log(competitor concentration) (at least three data points) and pIC_{50} values ($\log(B/(B_0-B)) = 0$) were determined by linear regression. K_i values were calculated from the IC_{50} value as well as the respective K_d value and the concentration of [^{3}H]-UR-MK299 according to the Cheng-Prusoff equation[394].

7.3.7 IP Accumulation Assay

The signal transduction assay was performed as previously described[364, 395]. Briefly, COS-7 cells (obtained from American Type Culture Collection) were seeded into 48-well plates and were transiently co-transfected with wild-type receptor or receptor mutant and a chimeric G protein ($G_{\alpha\Delta 6qi4myr}$) plasmid DNA[396]. Cells were routinely tested for mycoplasma contamination. Cells were radioactively labelled with myo-[2-^{3}H]-Inositol (Perkin Elmer, Boston, USA) overnight, and then stimulated with an increased concentration of NPY (NPY-curve). For antagonist curves, cells were stimulated with the

antagonist (BIBP3226: 10^{-5} M, BIBO3304: 10^{-6} M, MK-HU404: 10^{-7} M, UR-MK289: 10^{-5} M, UR-MK299: 10^{-7} M) parallel to an increased concentration of NPY for 1h (standard conditions). After cell lysis, an anion exchange chromatography was performed and isolated, radioactive accumulated IP derivatives were analyzed by liquid scintillation counting (scintillation cocktail Optiphase HiSafe, Perkin Elmer).

Using GraphPad Prism 5.0 (GraphPad Software, San Diego, CA, USA) the determined concentration response-curves were analyzed. The curves were normalized to the top (100%) and bottom (0%) values of the associated NPY curve. All curves of independent experiments were summarized to one single concentration response-curve by the row means total function. Using nonlinear regression (curve fit) the EC_{50} and pEC_{50} ± SEM were examined. The shift between NPY and NPY/antagonist curve is defined as EC_{50} ratio and calculated by dividing $EC_{50(NPY/antagonist)}/EC_{50(NPY)}$, set Hill slope to 1. All experiments were performed in duplicates of at least two independent experiments.

7.3.8 Live Cell Fluorescence Microscopy

The membrane localization of Y_1R and receptor mutants was verified by fluorescence microscopy. COS-7 cells were seeded in 8 well µ-slides (IBIDItreat, Martiensried, Germany) and transiently transfected with Lipofectamine2000 transfection reagent (Invivogen, Toulouse, France). 24h post transfection, nuclei were stained with Hoechst33342 (Sigma-Aldrich, Taufkirchen, Germany) and images were recorded using an ApoTome Imaging System, AxioVert Observer Z1 (YFP: Filter Set 46, DAPI: Filter Set 49, ApoTome, 63x/1.40 oil objective, ZEISS, Jena, Germany) in a quasi-confocal setting. The data demonstrate that all the mutants are expressed at similar, and wild-type-like, level in COS-7 cells.

7.3.9 Quantification of Receptor Surface Expression in COS-7 Cells

COS-7 cells were seeded into black 96 well plates (Greiner, Kremsmünster, Austria), and transiently transfected with plasmid encoding receptor-eYFP fusion protein using MetafectenePro. 24h post transfection, cells were washed once with HBSS, and fluorescence was quantified using a plate reader (Tecan Infinite M200, Tecan, Männedorf, Switzerland) at excitation 488/5 nm and emission 530/5 nm. Data were normalized to mock transfected (0%) and wild type Y_1R-eYFP (100%), respectively. Results represent means ± s.e.m. from three independent experiments performed in quadruplicate.

7.3.10 Peptide Synthesis

Porcine NPY (YPSKPDNPGEDAPAEDLARYYSALRHYINLITRQRY-NH_2) and NPY analogues were synthesized by automated solid-phase peptide synthesis (SPPS) on an automated multiple peptide synthesis robot system (Syro, MultiSynTech, Bochum, Germany), using a 9-fluorenylmethoxycarbonyl-*tert*-butyl (Fmoc/*t*Bu) strategy in 15 µM scale as previously described [397]. NPY-tyramide was synthesized as previously described [398]. Isotopically labelled NPY variants were prepared as described [399], and $^{13}C/^{15}N$ labelled amino acids were coupled manually with 2 eq. HOBT/DIC in DMF over night. The porcine variant of NPY, which contains a single mutation at position 17 from a methionine to leucine, was used, as this variant has identical binding affinity and signalling properties as human NPY with an increased solubility to assist in handling and will therefore be referred to as wild-type NPY [400].

Modified NPY analogues [Bpa^1,K^4[$(Ahx)_2$-biotin]]NPY and [K^4[$(Ahx)_2$-biotin]]NPY were synthesized by automated solid phase peptide synthesis and Bpa/Ahx/biotin were coupled manually using orthogonal Dde protection groups, cleaved

by freshly prepared 3% (v/v) hydrazine in DMF for 10 × 5 min. Manual coupling reactions were performed by incubation of the resin with 5 eq. of the respective amino acid, 5 eq. HOBt and 5 eq. DIC in DMF for 2 h.

For biotin labelling, 3 eq. biotin was dissolved in DMF for 10 min at 60 °C. Next, 3 eq. HOBt and 3 eq. DIC were added to the mixture. Coupling was performed overnight at room temperature under constant shaking. Bpa containing peptides were cleaved from the resin and completely deprotected with a mixture of trifluoroacetic acid (TFA)/thioanisole (TA)/water (90:5:5 v/v/v).

All peptides were purified by preparative reversed-phase high-performance liquid chromatography (RP-HPLC) on a Jupiter 4u Proteo RP-C18 column (90 Å, 4 µm, Phenomenex), Kinetex 5u XB-C18 column (100 Å, 5 µm, Phenomenex), Kinetex 5u Biphenyl (100 Å, 5 µm, Phenomenex), Aeris 3.6 µm WIDEPORE XB-C18 (200 Å, 3.6 µm, Phenomenex) or Varitide RPC (200 Å, 6 µm, Agilent Technologies). All peptides were characterized by MALDI-ToF (Ultraflex III MALDI-ToF/ToF, Bruker daltonics) and ESI-HCT (High-capacity ion trap ESI-MS, Bruker Daltonics). Peptide purities were determined by two different analytical RP-HPLC systems and using 0.1% (v/v) TFA in H_2O (eluant A) and 0.08% (v/v) TFA in acetonitrile (ACN) (eluant B). Purity of all peptides was ≥ 95%.

7.3.11 NMR Measurements of Y_1R-bound NPY

Fourteen differently isotopically labelled (U-^{13}C/^{15}N) porcine NPY peptides were prepared by standard Fmoc solid phase synthesis as described previously[399]. On the basis of the structure of micelle-bound NPY, the positions of the NMR labels were chosen to avoid signal overlap in ^{13}C-^{13}C single quantum/double quantum correlation experiments

and to allow straightforward signal assignment.

Expression of the human Y_1R in *E. coli* as inclusion bodies, inclusion body preparation, solubilization of the receptor in SDS and receptor purification was performed according to Schmidt *et al.* [401] yielding ~40-50 mg Y_1R per liter expression medium. To assemble the Y_1R into a functional state, a three-step folding protocol was applied. In step 1, the purified Y_1 receptor was dialyzed against a degassed buffer containing 2 mM SDS, 50 mM NaP, pH 8.5, 1 mM EDTA, 1 mM reduced glutathione, and 0.5 mM oxidized glutathione at room temperature for 48 h. Subsequently, 25% (w/w) poly(ethylene glycol) (M.W. 20 kDa) was added to the buffer to concentrate the receptor before reconstitution. In step 2, bicelles consisting of 1,2-dimyristoyl-*sn*-glycero-3-phosphocholine (DMPC) / 1,2-diheptanoyl-*sn*-glycero-3-phosphocholine (DHPC-c7) (AvantiPolarLipids, Alabaster, USA) with a q value of 0.25 in 50 mM NaP, pH 8.0 were incubated with Y_1R, followed by three cycles of fast temperature cycles from 42°C to 0°C. Aggregated protein was removed instantly by centrifugation. In step 3, the Y_1R samples were concentrated in large bicelles ($q > 20$), which were used instead of liposomes because the high achievable receptor packing[401]. In large bicelles, all receptor binding sites are fully accessible. Subsequently, 50 mg ml^{-1} BioBeadsSM2 was added at least twice to the solution. After removal of the beads with a sieve, samples were washed four times through cycles of pelleting by centrifugation and resolubilization in 50 mM NaP, pH 7.0. Concentration determination of the membrane embedded receptor was performed by solubilization of the bicelles in 10 times the volume of 15 mM SDS and 50 mM NaP, pH 7.0 and subsequent measurement of the Y_1R intrinsic absorption at 280 nm using UV-Vis. Labelled NPY variants in a slight molar excess were added to the Y_1R after detergent removal but before concentrating.

Assessment of the binding affinity of the Y_1R was carried out using homogenous

fluorescence assays as described in the literature [402]. The reconstituted receptor was incubated in increasing concentrations with 50 nM fluorescently labelled NPY (NPY-atto520) overnight at room temperature in 50 mM NaP, pH 7.0 and 0.1% BSA. The fluorescence spectra were recorded on a FluoroMax-2 (JOBIN YVON) in a 10 mm quartz cuvette at 20 °C. The maximum signals of each spectrum were determined, normalized and plotted against the receptor concentration. The inflection point for Y_1R binding was determined (OriginPro 8G / DoseResp) at EC_{50} = 52 nM, demonstrating high functionality of the system. As a control, we used empty bicelles in concentrations, according to the bicelles concentration of the receptor samples, resulting in a lower binding ability to the ligand in comparison to the Y_1R.

NMR spectra were recorded on a Bruker Avance III 600 NMR spectrometer. The ^{13}C cross-polarization magic angle spinning (CP MAS) NMR experiments (0.7 ms contact time) were carried out using a double-resonance MAS probe with a 4 mm spinning module. Typical 90° pulse lengths were 4 µs for ^{1}H and ^{13}C and heteronuclear decoupling (SPINAL64) at a field strength of ~65 kHz. The ^{13}C chemical shifts were referenced relative to tetramethylsilane. The experiments were conducted at −30 °C and an MAS frequency of 7 kHz. The ^{13}C double quantum ^{13}C single quantum correlation spectra were acquired using the SPC-5 recoupling sequence [403] for double quantum excitation and reconversion (set to 0.571 ms each). The relaxation delay was 2.4 s.

7.3.12 Molecular Docking of NPY into Y_1R

Peptide docking of full-length porcine NPY was completed using Rosetta's FlexPepDock application [9]. Briefly, low energy backbone conformations were generated from the starting conformation of UR-MK299-bound Y_1R. Initially, the trimer of C-

terminal NPY was docked into these conformations using full flexible docking guided by mutagenesis data. For each round of docking 5,000 models were generated. The models were sorted by total energy and binding energy. Top models from a given docking round were used to seed the next round of docking in which the peptide was extended. Fragment picking was performed using the fragment_picker application within Rosetta [188]. Secondary structure during fragment picking was guided by the NMR chemical shift data. Additionally, experimentally derived restraints were used to guide docking (R35-D287$^{6.59}$, R35-N283$^{6.55}$, Y36-Y100$^{2.64}$, R33-N299$^{7.32}$, L30-I293^{ECL3}). After docking peptides of length 6, 12, 18, and 36, the binding pocket was resampled to allow the ligand binding pocket to adopt to the shape of the peptide. This was accomplished with RosettaCM [5]. The Y_1R crystal structure was used as a template along with the docked model to ensure the models did not drift too far from the starting structure though the N terminus was removed until the last docking step to provide steric bulk. Following full-length NPY docking, the N terminus of NPY was localized using loose distance constraints with the peptides identified in crosslinking experiments. Model selection from RosettaCM was accomplished using clustering to ensure backbone diversity. Following the identification of docked poses that satisfied the majority of experimental data, the chemical shifts of docked NPY peptides were calculated and filtered against the experimental NMR data to generate a final ensemble of docked poses with a 1.4 ppm root mean square distance to the experimental data. To analyze the binding interactions, per residue energetic analysis was calculated using the residue_energy_breakdown algorithm. The model with the lowest energy was selected as the representative binding pose (Fig. 4a). The ensemble is rather tight and therefore the individual binding poses are overall similar in structure (Fig. 4b).

7.3.13 Photo-crosslinking Experiment between Y_1R and NPY N-terminus

Cell-free expressed Y_1R was produced by coupled in vitro transcription/translation based on the protocol from Schwarz *et al.* [404] using a bacterial cell lysate (S30 extract) from *E.coli* BL21(*DE3*). Soluble membrane protein expression was achieved by addition of 0.1% (w/v) Brij-58, 1 mM GSSG and 5 mM GSH. Expression buffer was then exchanged to a binding buffer (0.1 M Tris/HCl, pH 7.4, 5% glycerol and 0.1% (w/v) Brij-58) and samples were purified by ligand affinity chromatography using [K^4[$(Ahx)_2$-biotin]]-NPY immobilized on Pierce Avidin Agarose beads (obtained from Thermo Fisher Scientific, Braunschweig, Germany) as previously described [405]. Elution was performed using 60 mM $CaCl_2$.

For photo-crosslinking Y_1R in binding buffer was incubated with [Bpa^1,K^4[$(Ahx)_2$-biotin]]-NPY in a molar ratio of 4:1 (5 nmol:1.25 nmol) for 30 min at room temperature. In addition, the same reaction was performed with an 8 fold-excess of NPY (Y_1R:[Bpa^1,K^4[$(Ahx)_2$-biotin]]-NPY:NPY, 4:1:8). Subsequently, the opened reaction vessels were placed on ice and irradiated with UV light (UV lamp: Atkas Fluorest forte, λ = 366 nm, 180 W) for 90 min. 50 µl of photo-crosslinked Y_1R sample (~ 200 µg) was digested with Glu-C and rLys-C (obtained from Promega, Mannheim, Germany) according to the manufacturer's protocol. Following, crosslinked fragments were isolated by affinity purification using Monomeric Avidin Agarose beads (obtained from Thermo Fisher Scientific, Braunschweig, Germany) according to the manufacturer's protocol. Possible fragments of digested Y_1R were calculated by the online tool PeptideMass [406]. To mind incomplete digestion the tool was allowed for a maximum of five missed cleavages. For the analysis the combined option "Glu C (phosphate) + Lys-C" was chosen. Same was done for the calculation of possible NPY fragments. Theoretical masses of fragments after

enzymatic cleavage of photo-crosslinked Y_1R – [Bpa^1,K^4[$(Ahx)_2$-biotin]]NPY were calculated by adding possible Y_1R fragment masses to NPY fragment masses containing the N terminus. The masses of Bpa, two times Ahx and biotin reduced by water for formation of a peptide bond were added manually. Peptide fragments of photo-crosslinked Y_1R were analyzed by matrix-assisted laser desorption ionization time of flight mass spectrometry (MALDI-TOF MS) using an Ultraflex III MALDI-TOF/TOF mass spectrometer (Bruker Daltonics, Billerica, MA, USA).

Functionality of cell-free expressed Y_1R samples was verified by a homogenous binding assay based on fluorescence polarization. We used [Dpr22-atto520]NPY (hereafter: NPY-atto520) as fluorescence tracer (IP accumulation in transiently transfected COS-7: EC_{50} = 24 nM, $pEC_{50} = -7.16 \pm 0.20$). 50 nM of NPY-atto520 was incubated with increasing concentrations of Y_1R in Brij-58 micelles in buffer (0.1 M Tris/HCl, pH 7.4, 2.5% glycerol, 0.1% (w/v) Brij-58 and 0.1% bovine serum album) for 90 min under gentle agitation in opaque 96 well plates. Fluorescence was then measured in a Tecan Spark plate reader (Tecan, Männedorf, Switzerland) using linear polarized light, excitation 510/5 nm, emission 550/10 nm, 90° detection angle. Experiments were conducted at least twice independently in duplicate.

7.3.14 Data Availability

Atomic coordinates and structure factor files for the Y_1R–UR-MK299 and Y_1R–BMS-193885 complex structures have been deposited in the Protein Data Bank with identification codes 5ZBQ and 5ZBH, respectively.

7.4 Results and Discussion

7.4.1 Characterization of the Y1R Crystal Structures

Within the β-branch of class A GPCRs, to which NPY receptors belong, the structures of four receptors, namely the neurotensin receptor NTS1[33], the OX_1 and OX_2 orexin receptors[38, 39] and the endothelin ET_B receptor[43], have been determined so. These structures reveal differences of ligand-binding modes between different receptors, suggesting that more structural information is needed to develop any consensus about the ligand recognition mechanisms for this GPCR subfamily. The Y_1R structure shares a canonical seven transmembrane helical bundle (helices I-VII) with the other known GPCR structures (Figure 7.1 and Figure 7.2a, b). The Y_1R–UR-MK299 and Y_1R–BMS-193885 complexes are structurally similar with C_α root-mean-square deviation (RMSD) of 0.75 Å within the helical bundle. Both structures exhibit inactive conformations with helix VI adopting an inward conformation that is similar to other inactive GPCR structures. UR-MK299 binds to Y_1R in a cavity within the helical bundle bordered by helices III, IV, V,

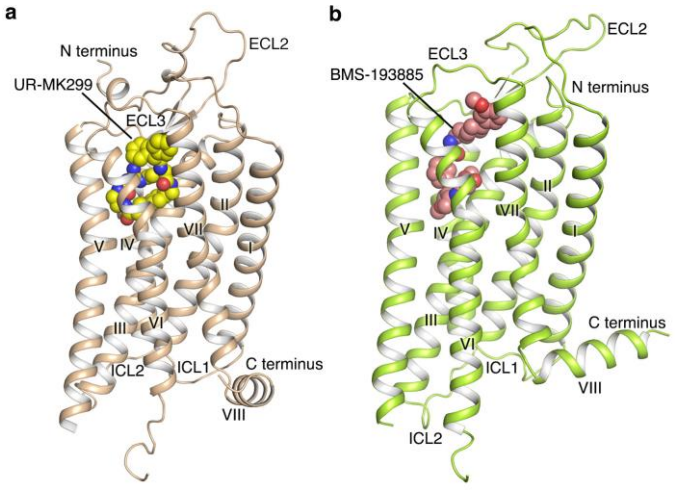


Figure 7.1 Overall structures of Y1R-UR-MK299 and Y1R-BMS-193885 complexes. **a**, Structure of Y_1R–UR-MK299 complex. The receptor is shown in brown cartoon representation. The ligand UR-MK299 is shown in sphere representation with yellow carbons. **b**, Structure of Y_1R–BMS-193885 complex. The receptor is shown in green cartoon representation. The ligand BMS-193885 is shown in sphere representation with pink carbons.

VI and VII (Figure 7.3a, b). The diphenylmethyl moiety of the antagonist interacts with a hydrophobic cluster formed by F282$^{6.54}$, F286$^{6.58}$ and F302$^{7.35}$ (superscript on residues throughout the text indicate Ballesteros-Weinstein nomenclature [407]) on helices VI and VII of Y_1R. The critical role of this hydrophobic patch in recognizing the argininamide-type Y_1R antagonist was confirmed by the NPY-induced inositol phosphate (IP) accumulation of Y_1R when inhibited by UR-MK299 and several related Y_1R antagonists—BIBP3226, BIBO3304, UR-HU404 and UR-MK289 (Figure 7.2e-i). The F302$^{7.35}$A mutation abolishes the antagonistic activity for all these antagonists, and a two- to fivefold decreased antagonistic effect of all tested antagonists was observed for the F286$^{6.58}$A mutation (Figure 7.4a-c).

7.4.2 The UR-MK299 Binding Site

The hydroxyphenyl group of UR-MK299 sits in a groove formed by helices III and VI of the receptor, enabling hydrophobic contacts with residues Q120$^{3.32}$, C121$^{3.33}$, I124$^{3.36}$, W276$^{6.48}$ and L279$^{6.51}$. In Y_1R and Y_2R, Q120$^{3.32}$ is suggested to be the interaction partner for the C terminus of NPY and crucial for receptor activation [399]. In the Y_1R–UR-MK299 structure, this residue forms a hydrophobic contact with the phenyl ring of the hydroxyphenyl group, potentially blocking the binding of Y_1R to NPY. Mutagenesis data show that the mutation Q120$^{3.32}$N does not influence the inhibitory effect of Y_1R antagonists on NPY signaling, but a mutation to histidine dramatically increases antagonistic activity of these ligands (Figure 7.4d, e), suggesting that an additional π-stacking interaction with the antagonist is beneficial at this position. The highly conserved residue W$^{6.48}$ represents the “toggle switch” and was suggested to trigger receptor activation through a conformational change in various GPCRs [408]. In the Y_1R–UR-

MK299 structure, the residue W276$^{6.48}$ is in a conformation similar to those observed in other inactive class A GPCR structures and distinct from their active-state conformations [409, 410]. The hydroxyphenyl group of UR-MK299 forms a hydrophobic contact with

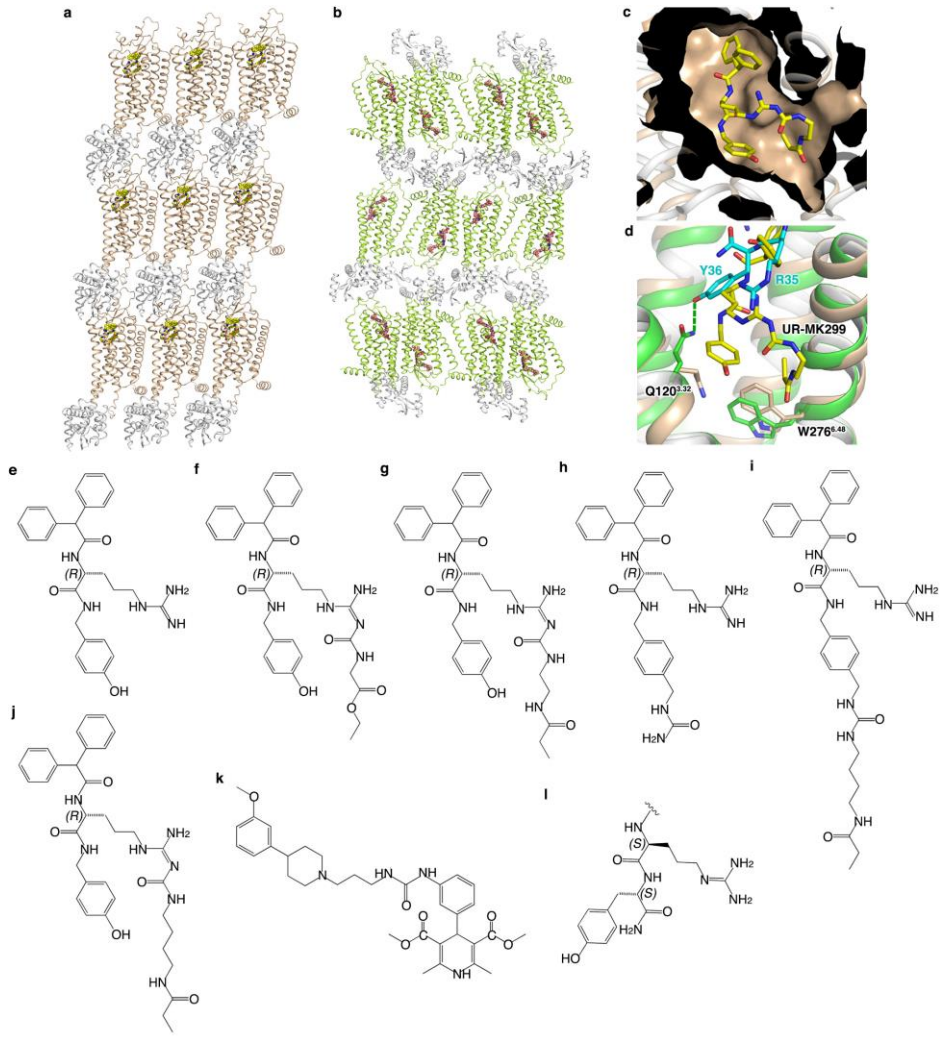


Figure 7.2 **Crystal packing and structural features of Y_1R and chemical structures of Y_1R ligands.** **a, b**, Crystal packing of Y_1R–UR-MK299 (**a**) and Y_1R–BMS-193885 (**b**) complexes. Y_1R is shown in cartoon representation and colored in brown and green in the Y_1R–UR-MK299 and Y_1R–BMS-193885 complexes, respectively. The T4L fusion is shown as grey cartoons. UR-MK299 and BMS-193885 are displayed as yellow and pink spheres, respectively. **c**, Cutaway view of UR-MK299 binding pocket in Y_1R. The receptor is shown in brown cartoon and surface representations. The ligand is shown as yellow sticks. **d**, Comparison of Y_1R between the Y_1R–UR-MK299 crystal structure (brown) and Y_1R-NPY model (green). Side chains of Q120$^{3.32}$ and W276$^{6.48}$ are shown as sticks. R35-Y36 of NPY is displayed as cyan sticks. The hydrogen bond between Q120$^{3.32}$ and Y36 of NPY is shown as a green dashed line. **e-j**, Chemical structures of argininamide Y_1R antagonists, BIBP3226 (**e**), UR-HU404 (**f**), UR-MK299 (**g**), BIBO3304 (**h**), UR-MK289 (**i**) and UR-MK136 (**j**). **k**, Chemical structure of BMS-193885. **l**, Scaffold of NPY C-terminal residues R35 and Y36. Key differences between R35-Y36 of NPY and UR-MK299 are chirality of the arginine derivative and alteration of bond connectivity leading to the hydroxyphenyl group.

W276$^{6.48}$, potentially preventing its activation-related motion and stabilizing the receptor in an inactive conformation. Compared to the wild-type receptor, the Y_1R mutant W276$^{6.48}$A displayed an over 2,000-fold decrease in binding affinity to [^{3}H]-UR-MK299 and reduced the antagonistic activity of the argininamide-type Y_1R antagonists by four- to sevenfold (Figure 7.4f), supporting its important role in antagonist recognition.

Residues N283$^{6.55}$ and D287$^{6.59}$ were suggested as the most important amino acids for Y_1R ligand recognition [411]. In the Y_1R–UR-MK299 structure, N283$^{6.55}$ is engaged in two hydrogen bonds with the α-nitrogen and the carboxylic oxygen next to the hydroxybenzylamine UR-MK299. D287$^{6.59}$ builds a salt bridge with the protonated guanidinyl moiety and a hydrogen bond with the carbamoyl group, in agreement with a decrease in antagonist affinity when the carbamoyl group was replaced by an alkoxycarbonyl, acyl or alkyl group [412]. The mutants N283$^{6.55}$A and D287$^{6.59}$N displayed a notable loss of NPY-induced receptor function, a complete abolishment of antagonistic activity for the small-molecule antagonists (Figure 7.4g, h), and an over 2,000-fold decrease in the binding affinity of Y_1R to [^{3}H]-UR-MK299. These data indicate the critical roles of these two residues in agonist and antagonist binding. Additionally, the carbamoyl substituent at the guanidine group binds deep in a sub-pocket shaped by helices V and VI, characterized by hydrophobic contacts with L216$^{5.43}$, T280$^{6.52}$ and N283$^{6.55}$, and a hydrogen bond between the oxygen of the propionyl group and Q219$^{5.46}$. The latter was reflected by a 30-fold decrease in binding affinity of [^{3}H]-UR-MK299 to the Y_1R mutants Q219$^{5.46}$A and Q219$^{5.46}$V. Extra empty space at the bottom of the sub-pocket is observed in the Y_1R–UR-MK299 structure, suggesting that a larger substituent may be allowed (Figure 7.2c). This is supported by studies showing that some other carbamoylated argininamide-type Y_1R antagonists containing longer carbamoyl chains, such as UR-MK136 (Figure

7.2j), bind to the receptor with a relatively high affinity [393].

UR-MK299 was reported to exhibit high Y_1R subtype selectivity (Y_2R, inhibition constant (K_i) > 3,000 nM, Y_4R and Y_5R, K_i > 10,000 nM) and specificity compared to two related neuropeptide FF (NPFF) receptors ($NPFF_1R$, K_i = 1,000 nM; $NPFF_2R$, K_i > 3,000

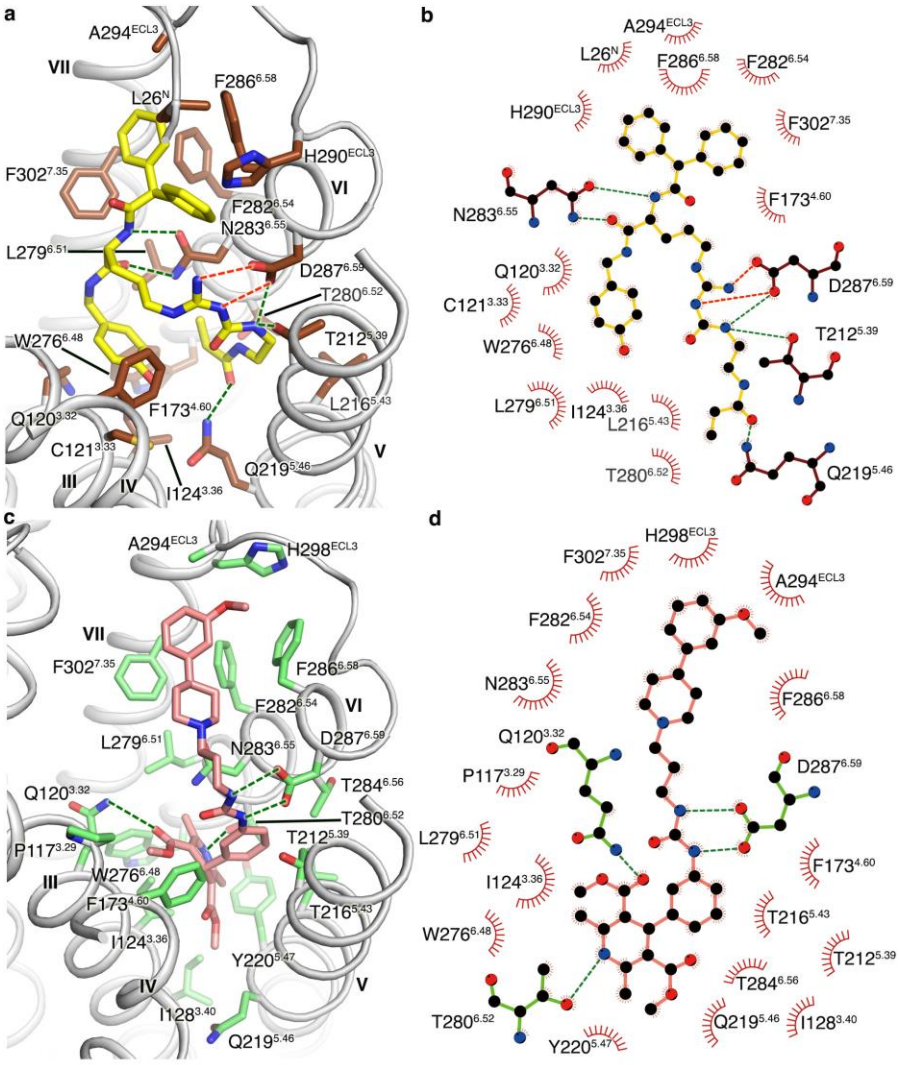


Figure 7.3 **Ligand-binding pocket of Y_1R for UR-MK299 and BMS-193885. a**, Binding pocket for UR-MK299. The receptor is shown in grey cartoon representation. UR-MK299 (yellow carbons) and receptor residues (dark brown carbons) involved in ligand binding are shown as sticks. Salt bridge and hydrogen bonds are shown as red and green dashed lines, respectively. **b**, Schematic representation of interactions between Y_1R and UR-MK299 analysed by LigPlot$^+$ (ref. [8]). The stick drawings of Y_1R residues and UR-MK299 are colored in dark brown and yellow, respectively. **c,** Binding pocket for BMS-193885. BMS-193885 (pink carbons) and receptor residues (green carbons) involved in ligand binding are shown as sticks. **d**, Schematic representation of interactions between Y_1R and BMS-193885 analysed by LigPlot$^+$ (ref. [8]). The stick drawings of Y_1R residues and BMS-193885 are colored in green and pink, respectively.

nM) [393]. Sequence alignment reveals that most of the key residues involved in UR-MK299 binding are conserved between Y_1R, the other NPY receptors and the NPFFRs, except for $F^{4.60}$, $Q^{5.46}$, $N^{6.55}$ and $F^{6.58}$, indicating the importance of these four residues in terms of selectivity and specificity of UR-MK299. Y_2R is the only NPY receptor with $L^{5.46}$ instead of $Q^{5.46}$, preventing key polar contacts. In Y_4R, $E^{6.58}$ disturbs the $F^{6.54}/F^{6.58}/F^{7.35}$ hydrophobic patch and probably mediates selectivity, supported by the $F^{6.58}E$ mutation in Y_1R reducing binding affinity for BIBP3226 [413], which contains the same diphenylmethyl group as UR-MK299. Similarly, hydrophilic residues at key positions impede high-affinity binding at Y_5R ($T^{6.58}$) and $NPFF_2R$ ($S^{6.58}$), while the hydrophobic pocket is preserved in $NPFF_1R$ ($L^{6.54}/I^{6.58}/F^{7.35}$), although with less bulk, leading to moderate affinity of BIBP3226 (K_i=18 nM) [393].

7.4.3 The BMS-193885 Binding Site

The ligands BMS-193885 and UR-MK299 occupy a similar binding pocket within the helical bundle of Y_1R (Figure 7.3c, d and Figure 7.2k). The dihydropyridine group of BMS-193885 fits in a sub-pocket formed by helices III, V and VI, which aligns with previous structure-activity relationship (SAR) studies showing that larger substituents at the position three of the dihydropyridine ring reduced Y_1R binding affinity [414]. Residue $T280^{6.52}$ forms a hydrogen bond with the nitrogen of the dihydropyridine ring as confirmed by our mutagenesis studies, which showed that the $T280^{6.52}A$ mutation decreased the binding affinity of BMS-193885 by about 330 fold, in agreement with a reported lower affinity N-methylated derivative [415]. Additionally, the dihydropyridine ring makes a hydrophobic contact with residue $I124^{3.36}$, which is consistent with a 400-fold decrease in affinity at the mutant $I124^{3.36}A$. It was also reported that methylation of either nitrogen of

the urea group of BMS-193885 significantly decreased binding ability of the methylated derivatives to Y_1R [415], suggesting that these hydrogen-bond donors are critical for Y_1R recognition. Indeed, in the BMS-193885-bound Y_1R structure, the urea group forms hydrogen bond interactions with $D287^{6.59}$. Similar to the diphenylmethyl group of UR-MK299, the piperidine and methoxyphenyl rings of BMS-193885 form extensive

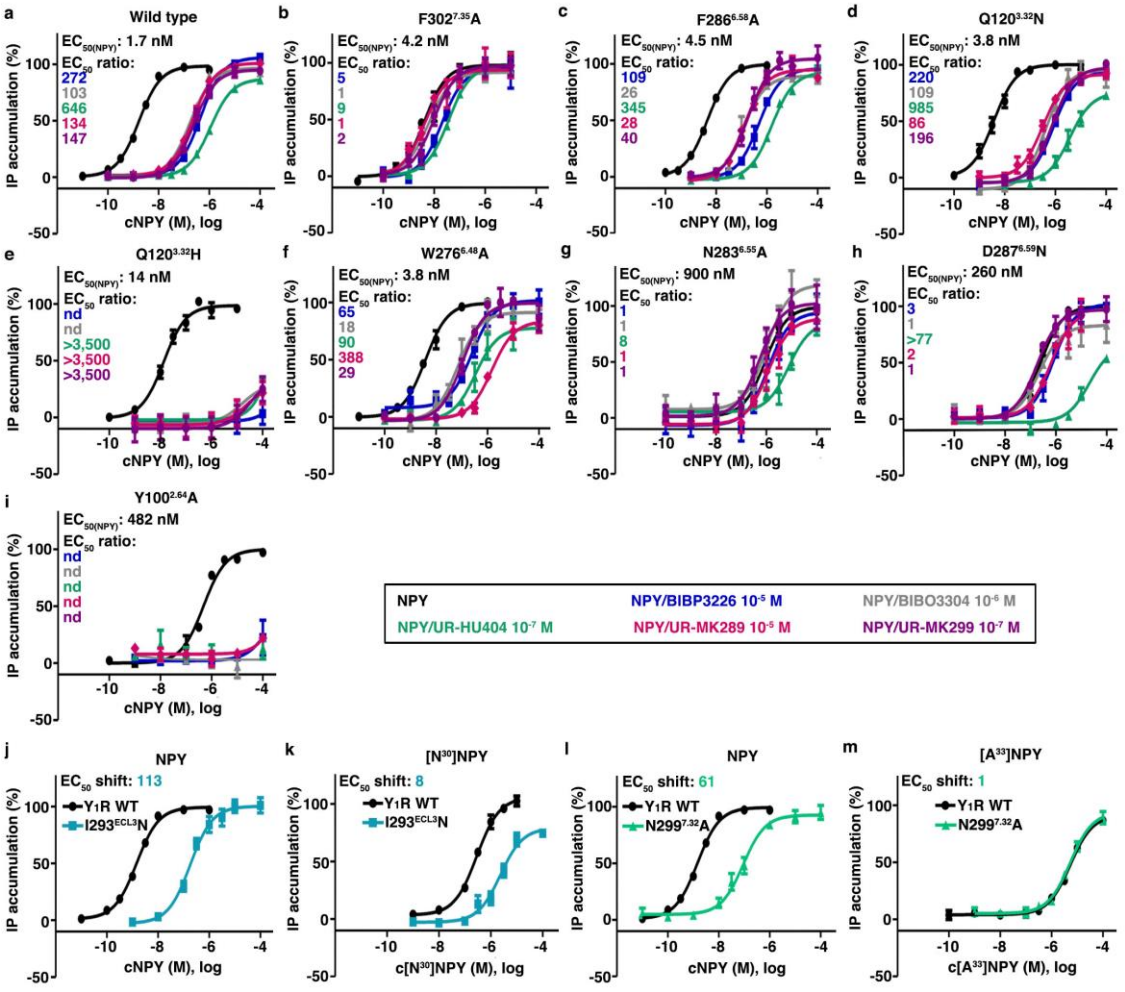
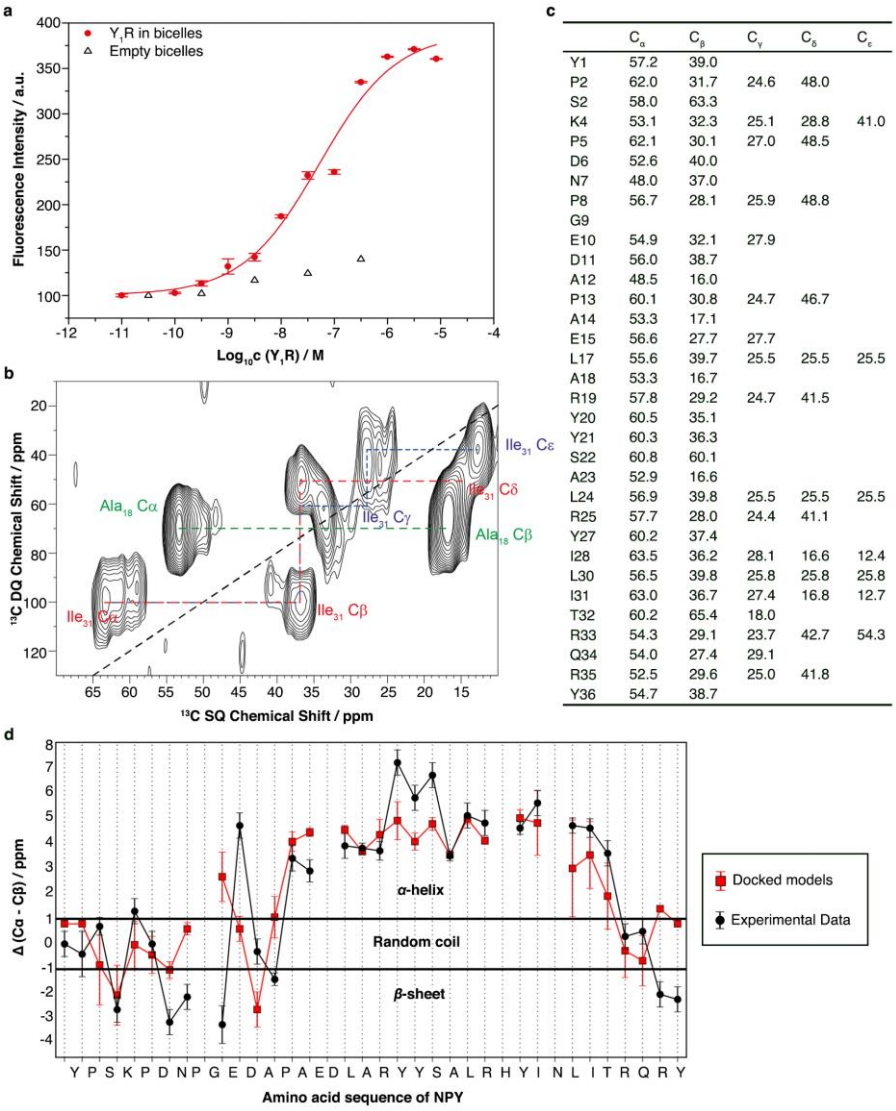


Figure 7.4 **IP accumulation assays. a-i**, NPY-induced IP accumulation of wild-type and mutant Y_1Rs in absence of antagonist or in presence of BIBP3226 (10^{-5} M), BIBO3304 (10^{-6} M), UR-HU404 (10^{-7} M), UR-MK289 (10^{-5} M) or UR-MK299 (10^{-7} M). (**a**) Wild-type Y_1R, (**b**) $F302^{7.35}A$, (**c**) $F286^{6.58}A$, (**d**) $Q120^{3.32}N$, (**e**) $Q120^{3.32}H$, (**f**) $W276^{6.48}A$, (**g**) $N283^{6.55}A$, (**h**) $D287^{6.59}A$ and (**i**) $Y100^{2.64}A$. EC_{50} values of NPY (black) and EC_{50} ratios ($EC_{50(NPY+antagonist)}/EC_{50(NPY)}$) for antagonists (colored) are given in the upper left corner for each plot. A reduced EC_{50} ratio of mutant compared to the wild-type receptor was interpreted as important for the respective antagonist. nd, not determined. **j-m**, Complementary mutagenesis assays of $[N^{30}]$NPY with $I293^{ECL3}N$ (**j**, **k**) and $[A^{33}]$NPY with $N299^{7.32}A$ (**l**, **m**). EC_{50} shifts ($EC_{50(mutant)}/EC_{50(WT)}$) are given in the upper left corner for each plot. A reduced EC_{50} shift of NPY analogue/Y_1R mutant compared to NPY/Y_1R mutant was interpreted as no further loss of function and a direct interaction between both positions. All data are shown as means ± s.e.m. from at least two independent experiments performed in duplicates.

hydrophobic contacts with the residues F282$^{6.54}$, F286$^{6.58}$ and F302$^{7.35}$. Replacement of the methoxyphenyl substituent by piperidine resulted in lower binding affinity to Y_1R[414], indicating the importance of the methoxyphenyl group in Y_1R binding and reflecting lipophilic demands at this position.

7.4.4 Modeling the NPY Binding Mode at Y1R

Understanding the binding mode of the endogenous agonist NPY at a molecular level will facilitate the rational development of Y_1R selective ligands. The C-terminal pentapeptide of NPY was identified to be essential for binding to the NPY receptors [22]. Since the hydroxyphenyl and the argininamide group of UR-MK299 mimic R35 and Y36 in the C terminus of NPY (Figure 7.2l), the crystal structure of Y_1R–UR-MK299 serves as a good template for molecular docking of the agonist. To aid docking, complementary mutagenesis studies were performed to determine corresponding interaction partners between Y_1R and NPY (Table 7.2a). Furthermore, solid-state NMR chemical shift measurements of several specifically $^{13}C/^{15}N$-labelled NPY variants revealed residue-specific alterations of the secondary structure of NPY upon binding to Y_1R (Figure 7.5). A number of key Y_1R/NPY contacts identified by the mutagenesis studies were used to guide NPY docking in Rosetta [188] with the final models being filtered against the NMR data to generate a final ensemble that best represents the combined data. The NPY-bound model reveals a relatively flat binding pose of NPY to Y_1R with the C-terminal tetrapeptide R33-Y36, identified as either a random coil or a β-strand structure in NMR, penetrating into the binding pocket (Figure 7.6a). The unstructured N terminus, Y1-P13, is in close proximity to the second extracellular loop (ECL2), while the α-helix in the middle region of NPY, A14-T32, lays along ECL1 and ECL3 and points away from ECL2.



c

	C_α	C_β	C_γ	C_δ	C_ε
Y1	57.2	39.0			
P2	62.0	31.7	24.6	48.0	
S2	58.0	63.3			
K4	53.1	32.3	25.1	28.8	41.0
P5	62.1	30.1	27.0	48.5	
D6	52.6	40.0			
N7	48.0	37.0			
P8	56.7	28.1	25.9	48.8	
G9					
E10	54.9	32.1	27.9		
D11	56.0	38.7			
A12	48.5	16.0			
P13	60.1	30.8	24.7	46.7	
A14	53.3	17.1			
E15	56.6	27.7	27.7		
L17	55.6	39.7	25.5	25.5	25.5
A18	53.3	16.7			
R19	57.8	29.2	24.7	41.5	
Y20	60.5	35.1			
Y21	60.3	36.3			
S22	60.8	60.1			
A23	52.9	16.6			
L24	56.9	39.8	25.5	25.5	25.5
R25	57.7	28.0	24.4	41.1	
Y27	60.2	37.4			
I28	63.5	36.2	28.1	16.6	12.4
L30	56.5	39.8	25.8	25.8	25.8
I31	63.0	36.7	27.4	16.8	12.7
T32	60.2	65.4	18.0		
R33	54.3	29.1	23.7	42.7	54.3
Q34	54.0	27.4	29.1		
R35	52.5	29.6	25.0	41.8	
Y36	54.7	38.7			

Figure 7.5 **Pharmacological characterization of refolded Y_1R and NMR studies of Y_1R-bound NPY.** **a**, Saturation binding of atto520-labelled NPY (50 nM) to increasing amounts of Y_1R-containing bicelles or empty bicelles. Data reflect fluorescence enhancement upon binding. An inflection point at EC_{50} = 52 nM was determined. Data shown are means ± s.e.m. from at least two independent experiments performed in duplicate. **b**, Typical ^{13}C MAS single quantum (SQ) / double quantum (DQ) correlation spectrum of NPY in presence of Y_1R reconstituted into large bicelles at −30 °C. **c**, Table showing ^{13}C NMR chemical shifts of assigned amino acids of NPY bound to Y_1R (referenced to tetramethyl-silane) as acquired in solid-state NMR experiments. **d**, ^{13}C chemical-shift index of NPY bound to Y_1R in large DMPC/DHPC-c7 bicelles (q > 20) compared with docked models. Plotted is the measured chemical shift difference (Cα-Cβ) for each individual residue of NPY subtracted by chemical shift difference of the same amino acid type in random coil conformation in black. Data shown are means ± s.e.m. from at least two independent experiments performed in duplicate. Chemical shifts were back-calculated for the top docking solutions and filtered against the experimental data to generate a final ensemble of docked poses. Their chemical-shift index and associated standard deviation are shown in red.

Table 7.2 IP accumulation assays of wild-type (WT) and mutant Y1Rs for NPY/NPY analogues.

†EC_{50} were determined using GraphPad Prism 5.0. All curves were normalized to the top and bottom values of the Y_1R/NPY curve. Nonlinear regression (curve fit) was performed for normalized response in all assays. All data are shown as means from at least two independent experiments performed in duplicate.

‡The EC_{50} shifts were determined by $EC_{50(mutant)}/EC_{50(WT)}$, set Hill slope to 1. For the wild-type receptor x-fold is set to 1. Lower EC_{50} shift of NPY analogue/mutant compared to NPY/mutant was interpreted as no further loss of function and a direct interaction between both positions.

§Data are from reference [4].

nd, not determined up to 10^{-4} M agonist concentration; /, not tested.

a. IP accumulation assays of complementary mutagenesis between NPY/NPY analogues and WT and mutant Y_1Rs

Peptides*	WT		Q120$^{3.32}$H		I293^{ECL3}N		N299$^{7.32}$A		N283$^{6.55}$A		D287$^{6.59}$A§	
	EC_{50} (nM)† (pEC_{50} ± SEM)	X-fold over WT‡	EC_{50} (nM) (pEC_{50} ± SEM)	X-fold over WT	EC_{50} (nM) (pEC_{50} ± SEM)	X-fold over WT	EC_{50} (nM) (pEC_{50} ± SEM)	X-fold over WT	EC_{50} (nM) (pEC_{50} ± SEM)	X-fold over WT	EC_{50} (nM) (pEC_{50} ± SEM)	X-fold over WT
NPY	1.5 (8.83 ± 0.02)	1	39 (7.41 ± 0.11)	26	169 (6.77 ± 0.08)	113	91 (7.04 ± 0.09)	61	1,053 (5.98 ± 0.13)	702	1,384 (5.86 ± 0.09)	461
[N^{30}]NPY	289 (6.54 ± 0.06)	1	/	/	2,334 (5.63 ± 0.09)	8	/	/	/	/	/	/
[A^{33}]NPY	5,395 (5.27 ± 0.07)	1	/	/	/	/	5,306 (5.28 ± 0.07)	1	/	/	/	/
[A^{35}]NPY	>10,000	1	/	/	/	/	/	/	/	/	>4,900	>5
[A^{36}]NPY	1,378 (5.86 ± 0.06)	1	/	/	/	/	/	/	nd	nd	/	/
NPY-tyramide	68 (7.17 ± 0.11)	1	nd	nd	/	/	/	/	/	/	/	/

b. IP accumulation assays of wide-type Y_1R for NPY/NPY analogues

	NPY	Ac-NPY	[A^1]NPY	[A^2]NPY	[A^1,A^2]NPY	NPY(3-36)	NPY(13-36)
EC_{50} (nM) (pEC_{50} ± SEM)	1.5 (8.83 ± 0.02)	1.7 (8.77 ± 0.19)	4.2 (8.38 ± 0.10)	7.3 (8.14 ± 0.11)	8.0 (8.10 ± 0.06)	82.7 (8.27 ± 0.07)	616 (6.21 ± 0.16)
X-fold over NPY	1	1	3	5	5	55	411

c. NPY-induced IP accumulation assays of WT and mutant Y_1Rs

	WT	F184^{ECL2}A	F184^{ECL2}N	V187^{ECL2}N	L189^{ECL2}N	Y192^{ECL2}S	V197^{ECL2}A	V197^{ECL2}N	F199^{ECL2}N	F202^{ECL2}N
EC_{50} (nM) (pEC_{50} ± SEM)	1.5 (8.83 ± 0.02)	19 (7.73 ± 0.08)	23 (7.64 ± 0.10)	1.9 (8.72 ± 0.08)	1.9 (8.73 ± 0.16)	3.8 (8.42 ± 0.11)	1.9 (8.71 ± 0.13)	188 (6.73 ± 0.11)	3.4 (8.47 ± 0.11)	1.2 (8.93 ± 0.15)
X-fold over WT	1	13	15	1	1	2	1	125	2	1

7.4.5 Differences between Agonist and Antagonist Binding Modes

Inspection of the NPY binding pocket of NPY reveals that the binding pose of residue R35 of NPY is similar to that of the argininamide of UR-MK299 (Figure 7.6b). R35 forms a salt bridge with the residue D287$^{6.59}$ of Y_1R and approaches to N283$^{6.55}$ (Figure 7.6c). The NPY mutant R35A displays a decrease in activity of over 6,000-fold, which represents the highest influence on agonist potency of all tested NPY analogues

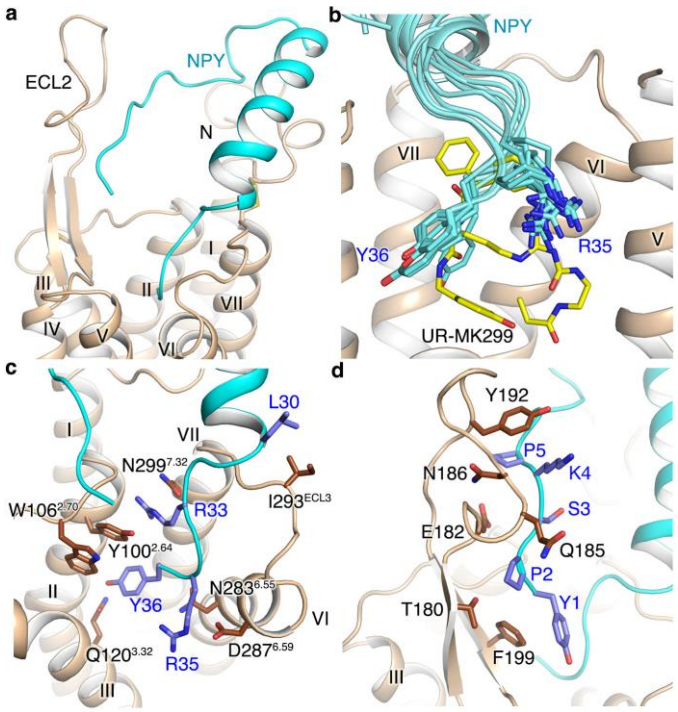


Figure 7.6 Docking poses of NPY. a, Overview of the NPY binding pose. The receptor and the lowest energy NPY conformation are shown as cartoons, and colored in brown and cyan, respectively. **b**, Comparison of UR-MK299 (yellow sticks) binding mode and predicted ensemble binding mode of NPY residues R35 and Y36 (cyan sticks). **c**. Predicted binding mode between Y_1R and NPY C terminus. Key Residues involved in Y_1R-NPY interaction are shown as sticks and colored in dark brown (Y_1R) and blue (NPY). **d**, Predicted binding mode between Y_1R's ECL2 and NPY N terminus. Residues in Y_1R's ECL2 and NPY N terminus that may form contacts are shown as dark brown and blue sticks, respectively.

(Table 7.2a), supporting the importance of the positively charged residue in NPY recognition. Aspartate or glutamate residues are not found at position 6.59 in any peptide GPCRs except for the receptors that bind to Arg-Phe-amide peptides, including NPFF, prolactin releasing peptide and pyroglutamylated Arg-Phe-amide peptide, which share a common C-terminal Arg-Phe-NH_2 motif, supporting the hypothesis that the arginine residue may function in a manner similar to that of the R35 of NPY by interacting with the conserved D/E$^{6.59}$ of the respective receptors[4]. In contrast to the similarity between the binding modes of the R35 of NPY and the guanidine group of UR-MK299, the C-terminal tyrosinamide of NPY and the hydroxyphenyl group of UR-MK299 show different orientations. The hydroxyphenyl ring is oriented towards helix V (Q219$^{5.46}$) in the UR-

MK299-bound Y_1R structure, while Y36 of NPY points toward the residue $Q120^{3.32}$ on helix III in the NPY-docked model (Figure 7.6c). This may arise from the opposite configuration of the stereo center in R35 of NPY and UR-MK299, as well as by only partial mimicking of the Y36-NH_2 of NPY by a 4-hydroxybenzyl group in UR-MK299 (Figure 7.2g, l). In the Y_1R–UR-MK299 crystal structure, residue $Q120^{3.32}$ forms a hydrophobic contact with the phenyl ring of the hydroxyphenyl group of the antagonist. By contrast, the NPY-bound model shows that side chain of $Q120^{3.32}$ points almost in the opposite direction and engages in a hydrogen bond with the hydroxy group of Y36-NH_2 (Figure 7.6c and Figure 7.2d), in a similar manner to the previously suggested interaction between the Y_2R residue $Q^{3.32}$ and NPY[416]. In Y_2R, it was also reported that $Q^{3.32}$ may interact with the C-terminal amide of NPY[399]. IP accumulation studies show that the Y_1R mutation $Q120^{3.32}H$ leads to a 26-fold decrease in potency of NPY, and NPY-tyramide lacking the C-terminal amide displays a 45-fold loss of activity. Complementary mutagenesis revealed an additional reduction of NPY-tyramide potency at the $Q120^{3.32}H$ mutant, and thus rules out a direct contact between the C-terminal amide of NPY and $Q120^{3.32}$ in Y_1R (Table 7.2a). Additionally, Y36 of NPY forms hydrophobic contacts with $Y100^{2.64}$ and $W106^{2.70}$ in Y_1R (Figure 7.6c). Although $Y100^{2.64}$ is not involved in antagonist binding, mutagenesis data suggest a critical role in agonist recognition as the Y_1R mutant $Y100^{2.64}A$ displays a 284-fold decrease in potency for NPY (Fig. 3i). Furthermore, the model reveals close contacts between L30 of NPY and I293 in ECL3 of Y_1R and between R33 of NPY and the Y_1R residue $N299^{7.32}$ (Figure 7.6c), which align with complementary mutagenesis data showing no further loss of function for combining mutant I293N with $[N^{30}]$NPY and $N299^{7.32}A$ with $[A^{33}]$NPY (Figure 7.4j-m and Table 7.2a).

7.4.6 Identification of the Role of the NPY N-terminus in Y1R Binding

Previous studies have shown that different NPY receptors behave differently when binding to the N terminus of NPY. While Y_2R and Y_5R can bind to N-terminally truncated NPY, Y_1R and Y_4R require the complete N terminus of NPY for full agonist potency[4, 22]. However, these data did not allow conclusions about the interaction of the N terminus of NPY with the receptor. To further explore the involvement of the NPY N terminus in recognition between the receptor and ligand, we performed mutagenesis studies, showing that truncation of the first two residues of NPY (NPY(3-36)) reduces peptide potency by more than 50-fold (Table 7.2b). This decrease in potency, however, is not seen when these residues are mutated to alanine ($[A^1,A^2]$NPY: 5-fold EC_{50} shift), suggesting important contributions of peptide backbone in binding to the receptor. Our NPY-bound model suggests that the N-terminal region of NPY makes close contacts with the fragment T180-F199 in ECL2 of Y_1R and is also in proximity to the receptor N terminus (Figure 7.6a, d). To experimentally verify interacting sites in the receptor, we performed photo-crosslinking studies between NPY analogues carrying the highly reactive *p*-benzoyl-phenylalanine $[Bpa^1, K^4[(Ahx)_2$-biotin]]NPY and Y_1R. Crosslinked fragments were assigned to two regions in Y_1R, the N terminus (K21-D32) and ECL2 (A191-D205) (Figure 7.7). Previous studies demonstrated that deletion of the Y_1R N terminus does not interfere with receptor signalling, but reduces NPY binding by about 95% compared to the full-length receptor[417]. This creates the possibility that the N terminus of Y_1R plays a role in recognizing and positioning the peptide ligand, which aligns with the photo-crosslinking data. Consistent with the crosslinking hits in receptor ECL2, our mutagenesis data show that mutations F184A/N and V197N in this region greatly reduce NPY potency (Table 7.2c). Together, these data suggest that the N terminus and ECL2 of Y_1R play important

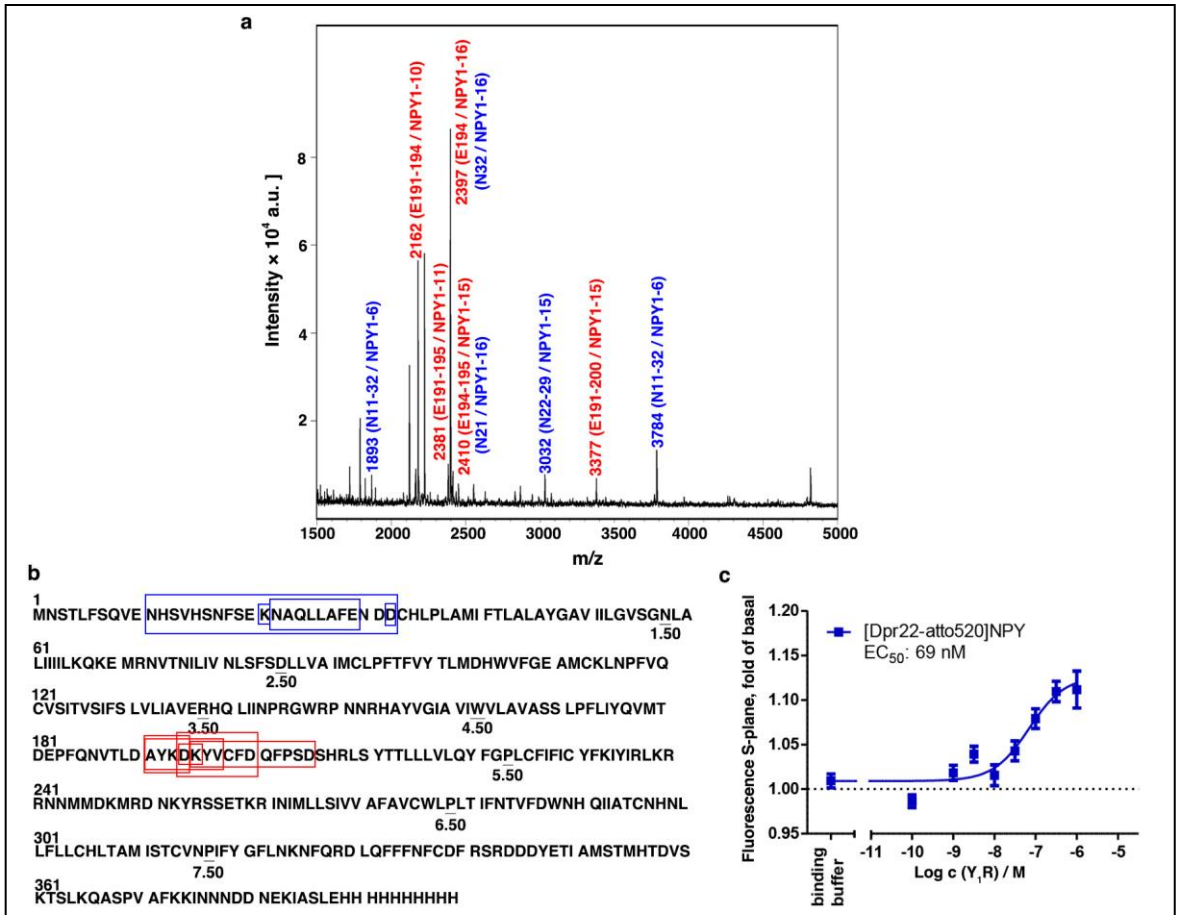


Figure 7.7 **Photo-crosslinking experiments between NPY and Y_1R.** **a**, Mass spectra of photo-crosslinked Y_1R with [Bpa[1],K[4][(Ahx)$_2$-biotin]]NPY. Exemplary MALDI-ToF mass spectra of photo-crosslinked samples enzymatically digested by rLys-C and Glu-C. Potential Y_1R fragments are labelled. N, N terminus of Y_1R (blue); E, ECL2 (red) **b**, Respective regions of NPY N terminus at Y_1R. Amino acid sequence of Y_1R with a C-terminal His-tag. The two detected regions within Y_1R (N terminus (blue), ECL2 (red)) after crosslinking with [Bpa[1],K[4][(Ahx)$_2$-biotin]]NPY are emphasized in boxes. The different sizes of the boxes represent different detected fragments (Extended Data Table 5), respectively. **c**, Saturation binding of atto520-labelled NPY (50 nM) to increasing amounts of cell-free produced Y_1R in Brij-58. Data reflect fluorescence enhancement upon binding. EC_{50} value of 69 nM was determined. Data shown are means ± s.e.m. from at least three independent experiments performed in triplicate.

roles in recognition of the N terminus of NPY and receptor activation. This contrasts with NPY binding at Y_2R, in which ECL2 may interact with the central α-helix of NPY and the peptide N terminus is flexible and not anchored by the receptor[399]. Although this study provides insights into the interactions between Y_1R and NPY, further structural details, such as structures of Y receptors bound to NPY, are required to fully understand the endogenous agonist-binding modes of the NPY receptor family.

CHAPTER 8

Conclusions

8.1 Summary

Chapter 1 introduces the problem of structural characterization of GPCRs with a specific focus on peptide-binding. Multiple studies demonstrate that conformational changes occur in both the receptor and ligands during the binding event. Characterization of either component separately will need to account for these changes in models of the binding event. A discussion of the use of hybrid methods to approach these challenging studies is presented.

Chapter 2 provides an overview of the Rosetta modeling suite with a focus on the protocols that are used repeatedly throughout the thesis. The primary topics covered are protein structure prediction, design, and ligand- and protein-protein docking. Additionally, background on the scoring function and scripting interfaces is provided as these play central roles in every protocol. This chapter provides examples of success cases for each topic and demonstrations of effective use of the various protocols. This chapter is paired with tutorials found at www.meilerlab.org.

Chapter 3 describes a method for virtual screening of drug fragments using RosettaLigand. This is a particularly challenging test of the RosettaLigand algorithm as drug fragments contain few functional moieties and possess high micromolar affinities for the receptor. The binding of these compound is likely mediated by one or two interactions within the protein binding pocket. Despite this challenge, Rosetta performed well in

sampling and correct binding pose and ,importantly, also demonstrated success at ranking compounds, a task generally best accomplished with cheminformatics protocols.

Chapter 4 extends the work in Chapter 3 to allow for redesign of the protein-ligand interface using the RosettaLigand docking algorithm. This is a proof of principle experiment as pocket design in Rosetta has traditionally required pre-defined ligand and side chain geometries that are grafted onto any available protein pocket. This protocol specifically asks Rosetta to use a single protein and a single ligand but does not require pre-defined geometries. This chapter is paired with Appendix A which describes the method in a general workflow.

Chapter 5 presents a new method and database for GPCR modeling within the Rosetta framework. This method requires a new alignment method to account for the low sequence identity but high structural conservation of this family of proteins. The new results suggest that accurate models of GPCRs can be predicted from templates with sequence identities as low as 20%, down from the previously reported threshold of 40%. Given the success of the method, a database of all druggable human GPCRs was generated. This chapter is paired with Appendices B and C which extend the protocol for combination with protein-ligand and protein-protein binding predictions of GPCRs, respectively.

Chapter 6 combines the method presented in Chapter 5 with experimental data to characterize the bound state of the ghrelin peptide at its receptor. Ghrelin exists in a multitude of conformations in its unbound state and an iterative approach to peptide docking and receptor backbone sampling was created. This new method demonstrates the ability for Rosetta to model binding events of highly flexible binding partners and account for conformational selection during the binding process.

Chapter 7 is similar to Chapter 6 but characterizes the neuropeptide Y Y_1 receptor.

Here we additionally generated crystal structures of the Y_1 receptor bound to small molecule antagonists and describe differences in binding between the peptide agonist and antagonists. These results provide meaningful interpretations for dug development platforms for specific functional activity of future compounds. Additionally, the modeling process integrated NMR, photo-crosslinking, and mutational analysis demonstrating the power of hybrid methods for characterizing these systems.

8.2 Key Findings

The development of accurate methods for structure prediction, small molecule docking, peptide docking, and interface design demonstrate the power and flexibility of the Rosetta modeling suite. Many of the base algorithms required little additional modification. Much of the forward progress presented in this work came from innovations to the inputs and iteration between protocols to better mimic biological processes. This is perhaps the one negative to Rosetta's implementation, the reliance on structures that demonstrate the final stage: either in comparing the predicted loop structures of a static crystal structure that represents a single state or in the prediction of a binding event that occurred over several stages. Rosetta tries to predict the end state and not the process which is acceptable for rigid body recognition but insufficient for induced fit or conformational selection. However, a strength of Rosetta is the ability to combine various protocols as I have done to better capture these recognition events. Movement towards structural ensembles is in progress for experimental characterization of proteins [374, 378]. As such, it is rare for a computational prediction to generate a single model but an ensemble. Plus, the incorporation of molecular dynamics simulations paired with Rosetta modeling has seen past success [418]. Despite this limitation, the virtual screening and homology modeling

benchmarks presented demonstrate the accuracy of Rosetta. The ability to accurately predict homology models with template identities as low as 20% is incredible. Using the previous method which relied on templates >40% identity, only a quarter of the GPCR family would be accurately modeled. Now we can confidently model nearly 90% of the GPCR family showing that we are rapidly closing the knowledge gap for this class of proteins.

In chapters in which experimental information is available, the search algorithm rapidly identifies a converged pose in strong agreement with the data. This speaks to the power of experimental constraints to narrow the search space. Iteration between experiment and computation resulted in the understanding of the ghrelin and NPY peptide binding poses. Either technique in isolation is unlikely to have come to the same conclusion, at least not in a meaningful timeframe. The majority of the work presented in this thesis is computational, but it would be insignificant were it not for the experimental feedback loops. For benchmarking and development of new protocols, tight feedback loops are even more critical. Therefore, while the accomplishments of the blind predictions in this thesis are notable, it is without a doubt that they will continue to improve with increasing availability of new experimental data.

8.3 Future Outlook

8.3.1 Challenges of Docking into Homology Models

The first part of this thesis focused on ligand docking into crystal structures. The results found that binding pose could be accurately identified in 90% of the cases while ranking of binding order dropped to 70%. For *in silico* virtual screening, 70% is not ideal

as it is likely that these screens will be upwards of one million compounds. This would allow for a very high false positive rate that would severely limit subsequent *in vitro* screening. While it is demonstrated before and in this thesis that pre-screening of libraries *in silico* does improve the hit rate *in vitro*, there is significant room for improvement. Unfortunately, moving towards homology models is even more fraught with difficulties [192, 257]. In these situations, there is a mix of inaccuracies in both the protein modeling component and in the ligand docking component. Appendix B represents a successful docking campaign into comparative models, but it should be noted that the docking was guided by a highly similar ligand binding to a distantly related protein. What is not presented in this thesis, are the attempts to dock ligands into homology models blindly. I find that there is a significant loss in both the scoring and sampling of ligands in these scenarios. Future work must be applied to this open question if we want to fully leverage the power of homology modeling. One possible strategy for ligand docking into these proteins is the recently developed RosettaLigandEnsemble [419]. This algorithm requires a series of related ligands to all bind in a similar fashion within a protein binding pocket which increases the likelihood for identifying native binding poses. There has been additional work towards combining ligand ensembles with protein ensembles to account for the flexibility inherent to homology modeling with early success. A second approach to ligand docking into homology modeling uses a fitness test for homology model accuracy [420]. In this scenario, a series of known actives and inactives are docked into a homology model. If the actives are ranked above the inactives it is likely that the homology model represents an accurate pose for more rigorous docking analysis. However, this requires docking several ligands into a series of models that is time and computationally expensive. A third approach would simultaneously dock ligand during the homology modeling step.

This is actually the most likely scenario, in my opinion, to yield positive results. During homology modeling of apo state proteins, energetic minimization often results in pocket contraction as the score function tries for fill voids in the protein. The presence of a ligand during this minimization step would prevent such a collapse. This has been demonstrated to be highly successful in comparative modeling of kinase inhibitors into kinases owing to the high structural similarity of the binding pocket and functional moieties within kinase inhibitors [John Karanicolas, RosettaCon2018, unpublished]. GPCRs would present a unique test case as the binding pockets and ligands are much more diverse. Pre-alignment of target ligands and template ligands will likely be of poor quality though may be sufficient. If they prove sufficient, it may be possible to present receptors with "hydrophobic spheres" representing a generic ligand to prevent collapse. Otherwise, either ligand docking into the threaded template or simultaneous ligand docking and comparative modeling may prove beneficial, though code in Rosetta does not currently support this owing to differential handling of chain identities in docking and comparative modeling algorithms. Future development of the code may alleviate this issue.

8.3.2 Extension of GPCR Model Database

The current database that has been generated in Chapter 5 represents the transmembrane regions of the druggable human GPCRs in their inactive states. While this may be sufficient for understanding available biochemical, genetic variation data, and docking strategies to the orthosteric binding pocket, there are numerous studies in which these models would be insufficient. Binding of ligands to extracellular domains as is the case for Classes B, C, and F are not currently possible. Additionally, docking of extrahelical allosteric modulators that recognize the active state conformation of a receptor

would likely yield good results. Understanding of the recognition of G-proteins and arrestins to the intracellular domain of GPCRs would also be impossible. Comparisons of inactive versus active state communication networks within a receptor would only be possible with active state structures. Understanding the role of C-terminal phosphorylation patterns as it affects arrestin recognition is also lacking. Further, the increasing role of receptor dimerization should be approached from a structural perspective. Lastly, extension to receptor sequences of mouse, fly, worm, and other organisms would expand the relativity of the database to researchers working in these model systems.

Therefore, it is prudent to extend the GPCR modeling protocol to routinely model active state receptors (with and without G-protein and arrestin present), full-length receptors (extended termini and folded domains), potential dimeric states, and genetic variability. Benchmarking the results would be impossible as many of these states have not been observed experimentally. However, it is possible to develop the protocols for benchmarking when data becomes available. My work has over time touched on all of these aspects and I will speak to this experience here. For any of these modeling strategies starting with the transmembrane regions of the inactive human receptors is prudent owing to the amount of structural data available for receptors in these conditions. For example, modeling of an active state receptor using only available active state templates would reduce the template dataset from ~50 templates to 7 reducing the likelihood of having an accurate model of the ligand binding pocket and extracellular loops. However, if one starts with modeling of the inactive structure, they can have high confidence in these regions. Then, using these output models as templates paired with the templates of the active states, they can generate more accurate models of the active receptor across its extracellular, ligand binding, and intracellular domains. These models could then easily be paired with

different G-proteins and arrestins for meta-analysis of receptor selectivity motifs and communication pathways between binding partners. The same scenario is expected for N- and C-terminal extension. Modeling of full-length receptors in the first step would generate significant noise for model selection as scoring and clustering would be biased towards termini conformations and not towards accuracy of the TM domain and extracellular loops. Additionally, the presence of these loops may provide steric hindrances to loops and generate errors in these loops. If the TM region and loops are modeled first and the termini are grafted onto the models, an overall higher accuracy should ensue.

8.3.3 Generalization of Peptide-binding Studies

Lastly, structural studies of peptide-binding GPCRs is a labor-intensive, computationally expensive study that requires integrated methods to arrive at well-defined structural models. Recombinant generation of receptors in bulk either through *E. coli*, insect cell, or cell-free expression is becoming more routine which should allow for more accessible studies of these systems. Additionally, solid-phase peptide synthesis is relatively straightforward and production of labeled peptides for NMR, EPR, fluorescence, or crosslinking is generating new avenues for hybrid studies. However, it seems there is a potentially overlooked dataset that could prove useful in understanding molecular recognition, though not necessarily conformation, of peptide-GPCR binding. Protein co-evolution has recently been used to find long range contacts for protein tertiary fold prediction [191]. In the last year, co-evolution was used to identify binding interactions across the interface of homo-oligomeric proteins [421]. Further, individual GPCRs and peptides have been analyzed over their evolutionary history [422-425]. Therefore, it seems likely that there is sufficient data to analyze peptide-binding GPCR interaction by co-

evolutionary analysis to identify novel interaction restraints. Initial studies will likely need to focus on well characterized binding interactions such as the ones listed in Chapter 1. However, if this technique proves useful, then it can be applied to the approximately 100 peptide-binding receptors that represent a quarter of the druggable GPCR family. This one technique could advance the characterization of the entire family and coupled with automated homology modeling could provide novel insights into the binding of these proteins. Of course, as is always the case, these experiments will rely on experimental feedback but the development of high-throughput GPCR screening such as the PRESTO-Tango [426] or PathHunter [427] platforms coupled with automated peptide synthesis of variants should prove useful in parsing these large dataset.

8.3.4 Modeling of GPCR Conformational Dynamics

GPCRs are highly dynamic proteins requiring distinct conformational states to achieve various signaling states. In basic receptor theory, there existed an inactive and an active state often characterized by the relative degree of openness of the cytosolic domain. However, as we learn more about the structural features of GPCRs intermediate states were required to explain various signaling phenotypes. Using UV- and infrared spectroscopy, the transition from 11-*cis*-retinal to all-*trans*-retinal in rhodopsin activation was found to proceed through at least 3 meta states [428]. Then in the "active" state, rhodopsin activation of G-protein requires either the recruitment of the N-terminal [429] or C-terminal helix [11] of Gα to initiate a series of conformational changes to occur before release of GDP and exchange of nucleotide [430, 431]. The use of crystallography and ^{19}F-NMR identified at least two inactive states of GPCRs characterized by the lock/unlock of

the DRY motif [378]. Further, co-crystallization with various ligands has been found to slightly alter the ligand binding pocket or extracellular loop conformations [45, 50]. Additionally, as the field of biased signaling grows, there are hypotheses suggesting different conformations of a receptor's active state can distinguish binding of various G-proteins or arrestin. Taken together, there is significant evidence to suggest GPCRs exist in a multitude of conformations or sets of conformations that can determine the overall signaling state of the receptor pool [432].

To be clear, there exists multiple interpretations of GPCR structural heterogeneity. The first is the existence of multiple GPCR conformations corresponding to the various signaling states of a receptor. The second is the existence of multiple conformations allowed for a single signaling state. The majority of my PhD these has focused on the latter. While crystal structures provide a snapshot of a single receptor conformation in a single signaling state it is true that an inherent heterogeneity and dynamics can be extrapolated from B factors [433] . High B factors, low occupancy, and multiple rotamers can be found throughout GPCR crystal structures. Often these indications of dynamism is localized in the loops and termini where there are fewer contacts and higher solute accessibility. Interpretation of the physiological role of this dynamism is lacking, however this speaks to the conformational dynamics at play in a single signaling state. Evidence of loop conformation rearrangement was seen in the serotonin receptors and C5a receptor with a suggestion that alterations in loop conformations play a role in biased signaling or ligand kinetics [45, 50, 434]. Additionally, for crystal structures, the resolution is an indication of the degree of structural heterogeneity present in a single crystal. As the highest resolution crystal structure for a GPCR is around 2 Å, this demonstrates a degree of uncertainty with respect to the published coordinates.

Protein modeling is inherently error prone. Oversimplification of the energy function is needed for rapid structural sampling. As such, it is unlikely, that a protein model will identify the true global minimum of a target protein but approximate it with a built-in degree of error. In my work, we used this to our advantage by selecting from the predicted minimum of the energy landscape to build ensembles of models that best represent the available experimental data. In all chapter, we do not present a single best model. Rather, we select from the top models an ensemble that taken together has the best agreement with the data. This approach acknowledges that we don't expect to identify the global minimum but a range of structures and related conformations that are near the global minimum. Additionally, it accounts for any errors inherent to a single model by averaging over the complete ensemble. We believe that this approach better represents the dynamism of a receptor in a given signaling state. In Chapter 6, I map the per-residue RMSD of the ensemble and it is found that there are large differences in the loops and termini and the extended ghrelin peptide reminiscent of the B factor distribution of crystal structures. However, unlike a crystal structure, we now have multiple related conformations that can be tested in discrete experiments to better understand the degree and role of dynamism in these regions. For certain experimental restraints, such as photo-crosslinking that have a wide range of potential linkages, it is best to represent the ensemble over a single model as there is a large uncertainty in a specific linkage, as we did in Chapter 7.

Despite our efforts to understand the dynamism in a single state for a specific ligand, there is much to be discovered about the multiple conformations that exist globally throughout a receptor population. Two studies on different receptors using 19F-NMR and DEER spectroscopy conclude opposite results, that all conformations exist and are selected for by various ligands [378] or that conformations are induced upon binding of various

ligands [374]. Additionally, studies will be needed to determine between induced fit and conformational selection for other receptors but it is likely to be a blend of the two. Despite differences in conclusions, it is agreed that there are discrete conformations for various signaling states of a GPCR. In my current implementation of GPCR modeling and ligand docking, I do not explore discrete conformations. It is common practice to dock both antagonists and agonists to the same "inactive" receptor conformation. While there is evidence that remodeling of the ligand binding pocket can induce a large conformational change in the intracellular domain of GPCRs, it is outside the sampling algorithm of Rosetta to identify these changes. RosettaLigand will only resample sidechain and backbone positions within a defined sphere around the ligand. Even if one were to increase the radius of the sphere to encompass the entire protein, it is unlike to evidence large conformational changes using a simplistic energy minimization. RosettaLigand's sampling algorithm is only meant to remove clashes between the protein and ligand and not to alter protein conformations. Additionally, RosettaCM, the workhorse of my receptor modeling method, requires a defined starting conformation and will only sample around that state. Running RosettaCM on a crystal structure will sample only about 2 Å away from the starting coordinates with the majority of change around 1 Å. However, GPCR activation through transmembrane helix 6 has been shown to translate up to 10-12 Å. In my attempts to generate "active" state conformations, I would use the active state crystal structures as templates to ensure sampling of this conformation, i.e. you cannot use an inactive conformation and expect to get a pocket to open for G-protein or arrestin binding. Further, modeling of the "active" state or even "inactive" state will result in pocket collapse upon energy minimization in the absence of a bound ligand or signaling partner. This is Rosetta's attempt to fill voids and may need changes to the energy function or addition of

coordinate constraints to avoid. However, at present, I ensure the addition of a binding partner for whichever pocket I am currently studying.

Unfortunately, even with all these modifications, we will still end up with only two discrete conformations ("active" and "inactive") albeit with ensembles around these states. In order to truly sample the full range of conformations available to a GPCR I feel that long time span molecular dynamics (MD) simulation. Though in computational time, an MD simulation that could cover the full conformational change of inactive to active state is quite expensive. A better approach would be steered dynamics which force the sampling from one starting point towards a defined ending point. Additionally, work is underway to modify Rosetta's relax algorithm to act similar to a steered dynamics run. This would have the advantage of using Monte Carlo sampling which is significantly faster than even a steered MD simulation. However, once a sampling algorithm is made available for this task, data interpretation will need to be approached cautiously. What constitutes a real intermediate from a false intermediate? The best evidence we will have will be spectroscopic methods like NMR, DEER, infrared, and CD. We will likely need to fit populations to the averages and hopefully the distributions should our sampling algorithm be sufficient. An important question to ask before attempting any such sampling will be for what purpose does this simulation serve? If the hypothesis that difference "active" states play a role in biased signaling, then these simulations could potentially "capture" the conformation of a biased pathway for subsequent *in silico* ligand screening for development of biased ligands. Further, it would be interesting to see if pockets open up along the receptor as it transitions from inactive to active as these could potentially serve as allosteric pockets for development of allosteric modulators. The options are endless, but an experimentally ground approach to intermediate sampling and interpretation is a must.

APPENDIX A

Rosetta and the Design of Ligand Binding Sites

A.1 Summary

Appendix A is related to Chapter 4 and is a methods paper describing the redesign of an existing protein pocket to bind a novel ligand using the protocol I developed in that chapter. This generalizes the protocol for any protein of interest, including GPCRs. This appendix is from the article "Rosetta and the Design of Protein-Ligand Interfaces" for which I am second author [435].

A.2 Introduction

Proteins that bind to small molecules (i.e. ligands) are involved in many biological processes such as enzyme catalysis, receptor signaling, and metabolite transport. Designing these interactions can produce reagents and therapeutics which can serve as biosensors, *in vivo* diagnostics, signal modulators, molecular delivery devices, and sequestering agents [436-440]. Additionally, the computational design of proteins that bind small molecules serves as a critical test of our understanding of the principles that drive protein/ligand interactions.

While *in vitro* techniques for the optimization of protein/small molecule interactions have shown success [441], these are limited in the number of sequence variants that can be screened and often require a modest starting affinity on which to further optimize[442]. Computational techniques allow searching larger regions of sequence space and permit design in protein scaffolds with limited to no detectable intrinsic affinity for the target ligand. Computational and *in vitro* techniques are often complimentary and starting affinity achieved via computational design can often be improved via *in vitro* techniques [3, 443]. Although challenges remain, computational design of small molecule interactions have yielded success on a number of occasions [440, 444] and further attempts will refine our predictive ability to generate novel ligand binders.

The Rosetta macromolecular modeling software suite [223, 289] has proven to be a robust platform for protein design, having produced novel protein folds [445, 446], protein/DNA interactions [447], protein/peptide interactions [448], protein/protein interactions[197], and novel enzymes [294, 326, 449]. Technologies for designing protein/ligand interactions have also been developed [3, 254, 439]. Design of ligand binding proteins using Rosetta can be approached in one of two way. One method derives from enzyme design, where predefined key interactions to the ligand are emplaced onto a protein scaffold and the surrounding context is subsequently optimized around them [3]. The other method derives from ligand docking, in which the interactions with a movable ligand is optimized comprehensively [254, 439]. Both approaches have proven successful in protein redesign and features from both can be combined using the RosettaScripts system [225], tailoring the design protocol to particular design needs.

Here we present a protocol derived from RosettaLigand ligand docking [201-203,

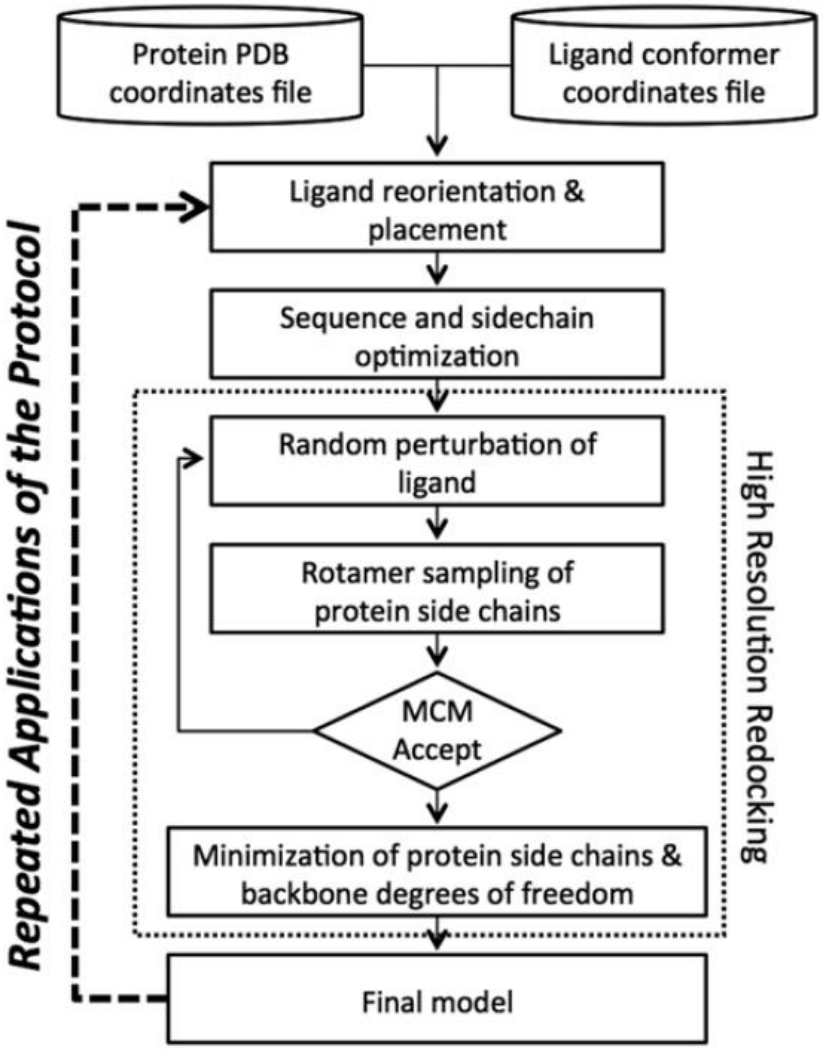


Figure A.1 Flowchart of RosettaLigand design protocol. From the combined input coordinates of the protein and ligand, the position of the ligand is optimized. Next residues in the protein/ligand interface are optimized for both the identity and position. After several cycles of small molecule perturbation, sidechain rotamer sampling, Monte Carlo minimization with Metropolis (MCM) criterion, and a final gradient-based minimization of the protein to resolve any clashes ("high resolution redocking"), the final model is the output. Further optimization can occur by using the final models of one round of design as the input models of the next round. Most variables in this protocol are user-defined, and will be varied to best fit the protein-ligand complex under study.

251], which designs a protein binding site around a given small molecule ligand (Figure A.1). After preparing the protein and ligand structures, the placement of the ligand in the binding pocket is optimized, followed by optimization of sidechain identity and conformation. This process is repeated iteratively, and the proposed designs are sorted and filtered by a number of relevant structural metrics such as predicted affinity and hydrogen

bonding. This design process should be considered as part of the integrated program of computational and experimental work, where proteins designed computationally are tested experimentally and the experimental results are used to inform subsequent rounds of computational design.

A.3 Materials

1. A computer running a Unix-like operating system such as Linux or MacOS. Use of a multi-processor computational cluster is recommended for productions runs, although test runs and small production runs can be performed on conventional laptop and desktop systems.
2. Rosetta: The Rosetta modeling package can be obtained from the RosettaCommons website (https://www.rosettacommons.org/software/license-and-download). Rosetta licenses are available free to academic users. Rosetta is provided as source code and must be compiled before use. See the Rosetta Documentation (https://www.rosettacommons.org/docs/latest/) for instructions on how to compile Rosetta. The protocol in this paper has been tested with Rosetta weekly release version 2015.12.57698.
3. A program to manipulate small molecules: OpenBabel [450] is a free software package which allows manipulation of many small molecule file formats. See http://openbabel.org/ for download an installation information. The protocol in this paper has been tested with OpenBabel version 2.3.1. Other small molecule manipulation programs can also be used.
4. A small molecule conformer generation program: We recommend the BCL [330] which is freely available from http://meilerlab.org/index.php/bclcommons for academic

use but does require an additional license to the Cambridge Structural Database [451] for conformer generation. The protocol in this paper has been tested with BCL version 3.2. Other conformer generation programs such as Omega [452], MOE [453], or RDKit [454] can also be used.

5. The structure of the target small molecule in a standard format such as SDF or SMILES (*see* **Note 1**).
6. The structure of the protein to be redesigned, in PDB format (*identification of suitable protein templates discussed in* **Note 2&3**).

A.4 Methods

Throughout the protocol $ROSETTA represents the directory in which Rosetta has been installed. File contents and commands to be run in the terminal are in *italics*. The use of a bash shell is assumed – users of other shells may need to modify the syntax of command lines.

3.1 Pre-relax the Protein Structure into the Rosetta Scoring Function [455]

Structures from non-Rosetta sources can have minor structural variations resulting in energetic penalties which adversely affect the design process (see **Notes 4&5**).

```
$ROSETTA/main/source/bin/relax.linuxgccrelease -
ignore_unrecognized_res -ignore_zero_occupancy_false -
use_input_sc -flip_HNQ -no_optH false -
relax:constrain_relax_to_start_coords -
relax:coord_constrain_sidechains -relax:ramp_constraints
false -s PDB.pdb
```

(For convenience rename the output structure:)

```
mv PDB_0001.pdb PDB_relaxed.pdb
```

3.2 Prepare the ligand

1. Convert the small molecule to SDF format including adding hydrogens as needed (*see* **Note 6**).

```
obabel LIG.smi --gen3D -O LIG_3D.sdf

obabel LIG_3D.sdf -p 7.4 -O LIG.sdf
```

2. Generate a library of ligand conformers (*see* **Note 7&8**).

```
bcl.exe molecule:ConformerGenerator -top_models 100 -
   ensemble_filenames LIG.sdf -conformers_single_file
   LIG_conf.sdf
```

3 Convert the conformer library into a Rosetta-formatted "params file" (*see* **Notes 9&10**).

```
$ROSETTA/main/source/src/python/apps/public/molfile_to_par
   ams.py -n LIG -p LIG --conformers-in-one-file
   LIG_conf.sdf
```

This will produce three files: LIG.params, a Rosetta-readable description of the ligand; LIG.pdb, a selected ligand conformer; LIG_conformers.pdb, the set of all conformers (*see* **Note 11**).

3.3 Place the ligand into the protein (*see* **Note 12&13**).

1. Identify the location of desired interaction pockets. Visual inspection using programs like PyMol or Chimera [456] is normally the easiest method (*see* **Note 14**). Use the structure editing mode of PyMol to move the LIG.pdb file from step 3.2.3 into the binding pocket. Save the repositioned molecule with its new coordinates as a new file (LIG_positioned.pdb) (*see* **Note 15**).

2. If necessary, use a text editor to make the ligand to be residue 1 on chain X (*see* **Note 16**).
3. Using a structure viewing program, inspect and validate the placement of the ligand (LIG_positioned.pdb) in the binding pocket of the protein (PDB_relaxed.pdb) (*see* **Note 17**).

3.4 Run Rosetta design

1. Prepare a residue specification file. A Rosetta resfile allows specification of which residues should be designed and which should not. A good default is a resfile which permits design at all residues at the auto-detected interface (*see* **Note 18**).

```
ALLAA
AUTO
Start
1 X NATAA
```

2. Prepare a docking and design script. ("design.xml") This suggested protocol is based off of RosettaLigand docking using the RosettaScripts framework [201-203, 251]. It will optimize the location of ligand in the binding pocket (low_res_dock), redesign the surrounding sidechains (design_interface), and refine the interactions in the designed context (high_res_dock). To avoid spurious mutations, a slight energetic bonus is given to the input residue at each position (favor_native).
3. Prepare an options file ("design.options"). Rosetta options can be specified either on the command line or in a file. It is convenient to put options which do not change run-to-run (such as those controlling packing and scoring) into an options file rather than the command line.

```
-ex1
-ex2
-linmem_ig 10
-restore_pre_talaris_2013_behavior #
```
(see **Note 25**)

4. Run the design application (*see* **Note 26&27**). This will produce a number of output PDB files (named according to the input file names, see **Note 28**) and a summary score file ("design_results.sc").

```
$ROSETTA/main/source/bin/rosetta_scripts.linuxgccrelease
   @design.options -parser:protocol design.xml -
   extra_res_fa LIG.params -s "PDB_relaxed.pdb
   LIG_positioned.pdb" -nstruct <number of output models> -
   out:file:scorefile design_results.sc
```

3.5 Filter Designs

1. Most Rosetta protocols are stochastic in nature. The output structures produced will contain a mixture of good and bad structures. The large number of structures produced need to be filtered to a smaller number of structures taken on to the next step.

A rule of thumb is that filtering should be done to remove unlikely solutions, rather than selecting the single "best" result. Successful designs are typically good across a range of relevant metrics, rather than being the best structure on a single metric (*see* **Note 29**).

The metrics to use can vary based on the desired properties of the final design. Good standard metrics include the predicted interaction energy of the ligand, the stability score of the complex as a whole, the presence of any clashes [244], shape complementarity of the protein/ligand interface [457], the interface area, the energy density of the interface

(binding energy per unit of interface area), and the number of unsatisfied hydrogen bonds formed on binding.

2. Prepare a file ("metric_thresholds.txt") specifying thresholds to use in filtering the outputs of the design runs. IMPORTANT: The exact values of the thresholds need to be tuned for your particular system (*see* **Note 30**).

```
req total_score value < -1010                # measure
of protein stability
req if_X_fa_rep value < 1.0                    # measure
of ligand clashes
req ligand_is_touching_X value > 0.5  # 1.0 if ligand is
in pocket
output sortmin interface_delta_X        # binding energy
```

3. Filter on initial metrics from the docking run. This will produce a file ("filtered_pdbs.txt") containing a list of output PDBs which pass the metric cutoffs.

```
perl
$ROSETTA/main/source/src/apps/public/enzdes/DesignSelect
.pl -d <(grep SCORE design_results.sc) -c
metric_thresholds.txt -tag_column last >
filtered_designs.sc

awk filtered_designs.sc  > filtered_pdbs.txt
```

4. Calculate additional metrics (*see* **Note 31**). Rosetta's InterfaceAnalyzer [458] calculates a number of additional metrics. These can take time to evaluate, though, so are best run on only a pre-filtered set of structures. After the metrics are generated, the structures can be filtered as in steps 3.5.1 and 3.5.2. This will produce a score file ("design_interfaces.sc") containing the calculated metric values for the selected PDBs.

```
$ROSETTA/main/source/bin/InterfaceAnalyzer.linuxgccrelease
   -interface A_X -compute_packstat -pack_separated -
```

```
score:weights ligandprime -no_nstruct_label -
out:file:score_only design_interfaces.sc -l
filtered_pdbs.txt -extra_res_fa LIG.params
```

5. Filter on additional metrics. The commands are similar to those used in step 3.5.2, but again the design_interfaces.sc score file, and with a new threshold file.

```
perl
   $ROSETTA/main/source/src/apps/public/enzdes/DesignSelect
   .pl -d <(grep SCORE design_interfaces.sc) -c
   metric_thresholds2.txt -tag_column last >
   filtered_designs2.sc


awk filtered_designs.sc  > filtered_pdbs.txt
```

Example contents of metric_thresholds2.txt:

```
req packstat value > 0.55                    # packing
metric; 0-1 higher better
req sc_value value > 0.45                    # shape
complementarity; 0-1 higher better
req delta_unsatHbonds value < 1.5    # unsatisfied
hydrogen bonds on binding
req dG_separated/dSASAx100 value < -0.5 # binding energy
per contact area
output sortmin dG_separated              # binding
energy
```

3.6 Manually Inspect Selected Sequences.

While automated procedures are continually improving and can substitute to a limited extent [459], there is still no substitute for expert human knowledge in evaluating designs. Visual inspection of interfaces by a domain expert can capture system-specific requirements that are difficult to encode into an automated filter (see **Note 32**).

3.7 Reapply the Design Protocol, Starting at Step 3.4

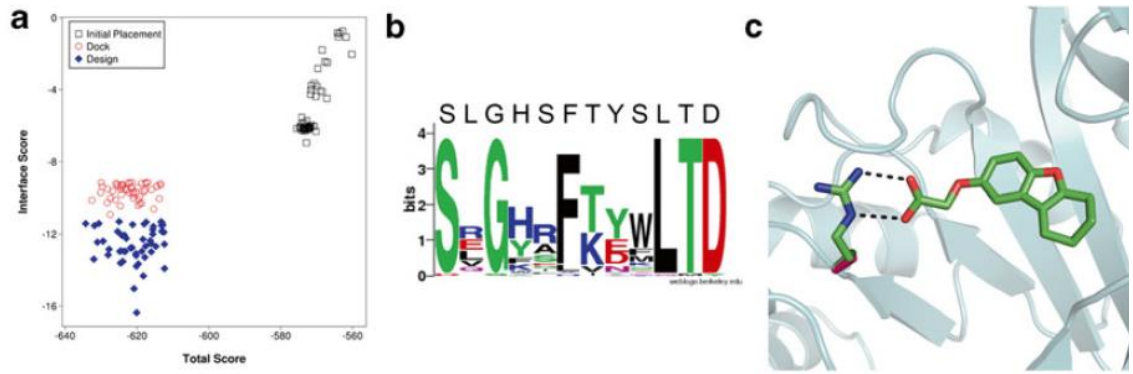


Figure A.2 Protein/ligand interface design with RosettaLigand. (a) Comparison in improvements in Interface Score and Total Score for top models from an initial placement, docking without sequence design, and docking with design. (b) Sequence logo of mutation sites among the top models from a round of interface design [6]. For most positions, the consensus sequence resembles the native sequence. Amino acids with sidechains that directly interact with the ligand show a high prevalence to mutation as seen in the positions with decreased consensus. (c) Example of a typical mutation introduced by RosettaLigand. The protein structure is represented in cartoon (cyan). The native alanine (pink) is mutated to an arginine residue (green) to match ionic interactions with the negatively charged ligand (green). Image generated in PyMol [10].

Improved results can be obtained by repeating the design protocol on the output structures from previous rounds of design. The number of design rounds depend on your system and how quickly it converges, but 3-5 rounds of design, each starting from the filtered structures of the previous one, is typical (see **Note 33**).

3.8 Extract protein sequences from the final selected designs into FASTA format.

```
$ROSETTA/main/source/src/python/apps/public/pdb2fasta.py
   $(cat final_filtered_pdbs.txt) >
   selected_sequences.fasta
```

3.9 Iteration of design.

Only rarely will the initial design from a computational protocol give exactly the desired results. Often it is necessary to perform iterative cycles of design and experiment, using information learned from experiment to alter the design process (Figure A.2).

A.5 Notes

1. While Rosetta can ignore chain breaks and missing loops far from the binding site, the structure of the protein should be complete in the region of ligand binding. If the binding pocket is missing residues, remodel them with a comparative modeling protocol, using the starting structure as a template.
2. Acceptable formats depend on the capabilities of your small molecule handling program. OpenBabel can be used to convert most small molecule representations, including SMILES and InChI, into the sdf format needed by Rosetta.
3. High resolution experimental structures determined in complex with a closely related ligand are most desirable, but not required. Experimental structures of an unliganded protein and even homology models can be used [5, 192].
4. The option "-relax:coord_constrain_sidechains" should be omitted if the starting conformation of the sidechains are from modeling rather than experimental results.
5. Rosetta applications encode the compilation conditions in their filename. Applications may have names which end with *.linuxgccrelease, *.macosclangrelease, *.linuxiccrelease, etc. Use whichever ending is produced on for your system. Applications ending in "debug" have additional error checking which slows down production runs.
6. It is important to add hydrogens for the physiological conditions under which you wish to design. At neutral pH, for example, amines should be protonated and carboxylates deprotonated. The "-p" option of OpenBabel uses heuristic rules to reprotonate molecules for a given pH value. Apolar hydrogens should also be present.

7. Visually examine the produced conformers and manually remove any which are folded back on themselves or are otherwise unsuitable for being the target design conformation.
8. It is unnecessary to sample hydrogen position during rotamer generation, although any ring flip or relevant heavy atom isomeric changes should be sampled.
9. molfile_to_params.py can take a number of options – run with the "-h" option for details. The most important ones are: "-n", which allows you to specify a three letter code to use with the PDB file reading and writing, permitting you to mix multiple ligands; "-p", which specifies output file naming; "--recharge", which is used to specify the net charge on the ligand if not correctly autodetected; "--nbr_atom", which allows you to specify a neighbor atom (*see* **Note 10**)
10. Specifying the neighbor atom is important for ligands with offset "cores". The neighbor atom is the atom which is superimposed when conformers are exchanged. By default, the neighbor atom is the "most central" atom. If you have a ligand with a core that should be stable when changing conformers, you should specify an atom in that core as the neighbor atom.
11. LIG.params expects LIG_conformers.pdb to be in the same directory, so keep them together when moving files to a new directory. If you change the name of the files, you will need to adjust the value of the PDB_ROTAMERS line in the LIG.params file.
12. Rosetta expects the atom names to match those generated in the molfile_to_params.py step. Even if you have a starting structure with the ligand correctly placed, you should align the molfile_to_params.py generated structure into the pocket so that atom naming is correct.

13. Other methods of placing the ligand in the pocket are also possible. Notably, Tinberg et al. [3] used RosettaMatch [268] both to place the ligand in an appropriate scaffold and to place key interactions in the scaffold.
14. Other pocket detection algorithms can also be used (see [460] and [461] for a review).
15. If you have a particularly large pocket, or multiple potential pockets, save separate ligand structures at different positions and perform multiple design runs. For a large number of locations, the StartFrom mover in RosettaScripts can be used to randomly place the ligand at multiple specified locations in a single run.
16. Being chain X residue 1 should be the default for molfile_to_params.py produced structures. Chain identity is important as the protocol can be used to design for ligand binding in the presence of cofactors or multiple ligands. For fixed-location cofactors, simply change the PDB chain of the cofactor to something other than X, add the cofactor to the input protein structure, and add the cofactors' params file to the -extra_res_fa commandline option. For designing to multiple movable ligands, including explicit waters, see Lemmon, et al. [253].
17. To refine the initial starting position of the ligand in the protein, you can do a few "design" runs as in step 3.4, but with design turned off. Change the value of the design option in the DetectProteinLigandInterface tag to zero. A good starting structure will likely have good total scores and good interface energy from these runs, but will unlikely to result in ideal interactions. Pay more attention to the position and orientation of the ligand than to the energetics of this initial placement docking run (*see* **Note 18**).
18. The exact resfile to use will depend on system-specific knowledge of the protein structure and desired interactions. Relevant commands are ALLAA (allow design to all

amino acids) PIKAA (allow design to only specified amino acids) NATAA (disallow design but permit sidechain movement) and NATRO (disallow sidechain movement). The AUTO specification allows the DetectProteinLigandInterface task operation to remove design and sidechain movement from residues which are “too far” from the ligand.

19. Change the name of the resfile in the XML script to match the filename of the resfile you’re using. The cut values decide which residues with the AUTO specification to design. All residues with a C-beta atom within cut1 Angstroms of the ligand will be designed, as will all residues within cut2 which are pointing toward the ligand. The logic in selecting sidechains is similar for cut3 and cut4, respectively, but this selects sidechains to repack rather than design. Anything outside of the cut shells will be ignored during the design phase, but may be moved during other phases.
20. The grid width must be large enough to accommodate the ligand. For longer ligands, increase the value to at least the maximum extended length of the ligand plus twice the value of box_size in the Transform mover.
21. Allison et al. found that a value of 1.0 for the FavorNativeSequence bonus worked best over their benchmark set [254]. Depending on your particular requirements, though, you may wish to adjust this value. Do a few test runs with different values of the bonus and examine the number of mutations which result. If there are more mutations than desired, increase the bonus. If fewer than expected, decrease the bonus.
22. More complicated native favoring schemes can be devised by using FavorSequenceProfile instead of FavorNativeSequence. For example, you can add weights according to BLOSUM62 relatedness scores, or even use a BLAST-formatted

position specific scoring matrix (PSSM) to weight the bonus based on the distribution of sequences seen in homologous proteins.

23. The value of box_size sets the maximum rigid body displacement of the ligand from the starting position. The value of rmsd sets the maximum rmsd from the starting position. Set these to smaller values if you wish to keep the designed ligand closer to the starting conformation, and to large values if you want to permit more movement. These are limits for the active sampling stage of the protocol only. Additional movement may occur during other stages of the protocol.
24. The provided protocol only does one round of design and minimization. Additional rounds may be desired for further refinement. Simply replicate the low_res_dock, design_interface, and high_res_dock lines in the PROTOCOLS section to add additional rounds of design and optimization. Alternatively, the EnzRepackMinimize mover may be used for finer control of cycles of design and minimization (although it does not incorporate any rigid body sampling).
25. Refinement of the Rosetta scorefunction for design of protein/small molecule interfaces is an area of current active research. The provided protocol uses the standard ligand docking scorefunction which was optimized prior to the scorefunction changes which occurred in 2013 and thus requires an option to revert the behavior. Decent design performance has also been seen with the “enzdes” scorfunction (which also requires the -restore_pre_talaris_2013 option) and the standard “talaris2013” scorefunction.
26. Use of a computational cluster is recommended for large production runs. Talk to your local cluster administrator for instructions on how to launch jobs on your particular cluster system. The design runs are “trivially parallel” and can either be manually split or run with an MPI-compiled version. If splitting manually, change the value of the -

nstruct option to reduce the number of structures produced by each job, and use the options -out:file:prefix or -out:file:suffix to uniquely label each run. The MPI version of rosetta_scripts can automatically handle distributing structures to multiple CPUs, but requires Rosetta to be compiled and launched in cluster-specific ways. See the Rosetta documentation for details.

27. The Rosetta option "-s" takes a list of PDBs to use as input for the run. The residues from multiple PDBs can be combined into a single structure by enclosing the filenames in quotes on the command line. Multiple filenames not enclosed in quotes will be treated as independent starting structures.
28. The number of output models needed (the value passed to -nstruct) will depend on the size of the protein pocket and the extent of remodeling needed. Normally, 1000-5000 models is a good standard run. At a certain point, you will reach "convergence" and the additional models will not show appreciable metric improvement or sequence differences. If you have additional computational resources, it's often better to run multiple smaller runs (100-1000 models) with slightly varying protocols (different starting location, number of rounds, extent of optimization, native bonus, etc.), rather than have a larger number of structures from the identical protocol.
29. Relevant metrics can be determined by using "positive controls". That is, run the design protocol on known protein-ligand interactions which resemble your desired interactions. By examining how the known ligand-protein complexes behave under the Rosetta protocol, you can identify features which are useful for distinguishing native-like interactions from non-native interactions. Likewise, "negative controls", where the design protocol is run without design (*see* **Note 17**) can be useful for establishing baseline metric values.

30. The thresholds to use are system-specific. A good rule of thumb is to discard at least a tenth to a quarter by each relevant metric. More important metrics can receive stricter thresholds. You may wish to plot the distribution of scores to see if there is a natural threshold to set the cut at. You will likely need to do several test runs to adjust the thresholds to levels which give the reasonable numbers of output sequences.
31. Other system-specific metric values are available through the RosettaScripts interface as “Filters”. Adding “confidence=0” in the filter definition tag will turn off the filtering behavior and will instead just report the calculated metric for the final structure in the final score file. Many custom metrics, such as specific atom-atom distances, can be constructed in this fashion.
32. Certain automated protocols can ease this post-analysis. For example, Rosetta can sometimes produce mutations which have only a minor influence on binding energy. While the native bonus (see **Notes 21&22**) mitigates this somewhat, explicitly considering mutation-by-mutation reversions can further reduce the number of such “spurious” mutations seen. Nivon et al. [459]presents such a protocol.
33. In subsequent rounds, you will likely want to decrease the aggressiveness of the low resolution sampling stage (the box_size and rmsd values of the Transform mover in step 3.4.2) as the ligand settles into a preferred binding orientation. As the output structure contains both the protein and ligand, the quotes on the values passed to the “-s” option (see step 3.4.4 and **Note 27**) are no longer needed. Instead, you may wish to use the “-l” option, which takes the name of a text file containing one input PDB per line. Each input PDB will each produce “-nstruct” models. Reduce this value such that the total number of unfiltered output structures in each round is approximately the same.

APPENDIX B

α_{2A}- and α_{2C}-Adrenoceptors as Potential Targets for Dopamine and Dopamine Receptor Ligands

B.1 Summary

Appendix B is related to Chapters 3 and 5 and combines GPCR structure prediction with small molecule ligand docking to develop a working hypothesis for cross-talk of ligands between dopaminergic and α-adrenergic receptors. It has long been recognized that dopamine (DA) innervation in the striatum is present while coupled α_{2A}- and α_{2C}-adrenoceptors are present despite a lack of norepinephrine (NE) innervation. Therefore it was suggested that DA could activate these adrenoceptors. BRET studies largely confirmed this cross-talk. Modeling of the various receptors found strong structural conservation of the ligand binding pocket and allowed for docking of dopamine with good scoring. I contributed the modeling simulations to this mostly experimental body of work. This appendix is from the article "α_{2A}- and α_{2C}-Adrenoceptors as Potential Targets for Dopamine and Dopamine Receptor Ligands" for which I am a middle author [462].

B.2 Introduction

The neurotransmitter NE binds and activates three subfamilies of adrenoceptors: α_1-adrenoceptors, subdivided into α_{1A}, α_{1B}, and α_{1D}; α_2-adrenoceptors, subdivided into α_{2A}, α_{2B}, and α_{2C}; and β-adrenoceptors, subdivided into β_1, β_2, and β_3 [463]. Classically, α_1-, α_2- and β-adrenoceptors couple to Gq/11, Gi/o, and Gs, respectively [463, 464]. In mammalian species, α_{2A} is the main subtype in most brain regions whereas α_{2B} subtype has a limited distribution and is mostly expressed in the thalamus [465, 466]. The α_{2C} subtype is found with particularly high density in the striatum [467, 468] with a moderately lower density than α_{2A}[469, 470]. The high density of dorsal striatal α_{2A}- and α_{2C}-adrenoceptors prompted a fundamental question in view of the well-known paucity of striatal noradrenergic terminals [471, 472] and the concomitant low extracellular levels of striatal NE [473]. Yet, a series of studies indicate that both types of receptors are fully functional in the striatum, where they seem to be localized mostly postsynaptically, preferentially in GABAergic striatal efferent neurons [474, 475]. There is also evidence for α_{2A}-adrenoceptors playing a role as autoreceptors localized in the sparse striatal noradrenergic terminals [476]. It was postulated by Zhang et al. [477] that dopamine (DA) could provide the endogenous neurotransmitter for striatal α_2-adrenoceptors. In transfected mammalian cells, using radioligand binding experiments, they found only a small preferential affinity of NE versus DA at both α_{2A}- and α_{2C}-adrenoceptors. Similar results were more recently obtained from radioligand binding studies using transfected mammalian and insect cell lines [478] and with radioligand binding and autoradiographic experiments in the bird and rat brain with a non-selective α_2-adrenoceptor ligand [479]. However, Zhang et al. [477] reported a much lower potency of DA (in the micromolar range) than NE (in the nanomolar range) at the level of α_{2C}-adrenoceptor-mediated signaling (modulation of forskolin-induced adenylyl

cyclase activation).

Due to the mismatch between dopaminergic and noradrenergic innervation and the density of their receptors in several brain areas, the controversy about the differential binding affinity of DA versus NE on adrenoceptors, and the potential functional efficacy of this binding, we wanted to study in detail the ability of DA and several synthetic DA receptor ligands to bind to the orthosteric site of α_2-adrenoceptors in transfected cells and in the sheep brain. Moreover, we further analyzed the ability of these ligands to generate functional responses: activation of G proteins, inhibition of cAMP accumulation, and ERK1/2 phosphorylation. In the present study, we first analyzed the ability of DA and several DA receptor ligands to bind to α_2-adrenoceptors in cortical tissue, which predominantly expresses α_{2A}-adrenoceptors, and in striatal tissue, which expresses both α_{2A}- and α_{2C}-adrenoceptors. We also studied the potential dopaminergic function of α_{2A}- and α_{2C}-adrenoceptors using the same methodology that recently allowed us to demonstrate the potent activation of all Gi/o-coupled DA D_2-like receptors by NE [480]. This methodology consists on sensitive bioluminescence resonance energy transfer (BRET)-based techniques that allow detection of ligand-dependent interactions between specific receptors and specific G proteins (G protein activation) or receptor-induced activation of effectors (adenylyl cyclase activity) in living cells [480]. Moreover, we compared the ability of NE, DA, and DA receptor ligands to modulate dynamic mass redistribution (DMR) and to activate MAPK signaling. Lastly, we modeled the binding of DA at α_{2A}- and α_{2C}-adrenoceptors, as compared to its binding to the crystallized D_3 receptor. Our results provide conclusive evidence for α_{2A}- and α_{2C}-adrenoceptors being not only NE but also DA receptors and common targets for other D_2-like receptor ligands.

B.3 Methods

B.3.1 DNA Constructs and Transfection

For BRET experiments, human receptor constructs were used for α_{2A}- and α_{2C}-adrenoceptors (cDNA Resource Center). The following human G protein constructs were used: Gαi1-, Gαi2-, Gαi3-, Gαo1-, or Gαo2-Renilla luciferase 8 (RLuc8) with RLuc8 inserted at position 91, untagged Gβ1, and Gγ2 fused to full-length mVenus at its N terminus. The Gα-RLuc8 constructs were kindly provided by Céline Galés (INSERM, Toulouse, France). The cAMP sensor using YFP-Epac-Rluc (CAMYEL) biosensor was obtained from the American Type Culture Collection (no. MBA-277; ATCC, Manassas, VA, USA) [481]. All the constructs were confirmed by sequencing analysis. A constant amount of plasmid cDNA (0.5 μg Gα-RLuc8, 4.5 μg Gβ1, 5 μg Gγ2-mVenus, and 5 μg receptor) was transfected into HEK-293T cells using polyethylenimine (Sigma-Aldrich) in a 1:2 ratio in 10-cm dishes. Cells were maintained in culture with Dulbecco's modified Eagle's medium supplemented with 10% fetal bovine serum and kept in an incubator at 37 °C and 5% CO_2. The transfected amount and ratio among the receptor and heterotrimeric G proteins were tested for optimized dynamic range in drug-induced BRET. HEK-293T cells were also used in BRET experiments for determination of adenylyl cyclase inhibition (see below). For DMR and ERK1/2 phosphorylation assays, Chinese hamster ovary (CHO) cells were grown in minimum essential medium (MEMα; Gibco) supplemented with 2 mM L-glutamine, 100 μg/ml sodium pyruvate, MEM non-essential amino acid solution (1:100), 100 U/ml penicillin/streptomycin, and 5% (*v*/*v*) of heat-inactivated fetal bovine serum. These cells were transfected with human α_{2A}-RLuc8 receptor using polyethylenimine in a 1:2 ratio in 25-cm^2 cell culture flasks. All experiments

were performed approximately 48 h after transfection.

B.3.2 BRET

BRET assays were performed to detect receptor ligand-induced events for Gi/o protein activation. Gi/o protein activation assay used RLuc-fused Gαi/o protein subunit and mVenus-fused Gγ2 protein for BRET pair. Receptor and untagged Gβ1 constructs were co-transfected. As reported previously [480, 481], cells were harvested, washed, and resuspended in phosphate-buffered saline. Approximately 200,000 cells/well were distributed in 96-well plates, and 5 μM coelenterazine H (substrate for luciferase) was added to each well. One minute after the addition of coelenterazine H, ligands (DA, NE, clonidine, quinpirole, 7-OH-PIPAT, and RO-105824) were added to each well. Antagonists were added 10 min before coelenterazine. The fluorescence of the acceptor was quantified (excitation at 500 nm and emission at 540 nm for 1-s recordings) in Mithras LB940 (Berthold Technologies, Bad Wildbad, Germany) to confirm the constant expression levels across experiments. In parallel, the BRET signal from the same batch of cells was determined as the ratio of the light emitted by mVenus (510–540 nm) over that emitted by RLuc (485 nm). G protein activation was calculated as the BRET change (BRET ratio for the corresponding drug minus the BRET ratio in the absence of the drug) observed 10 min after the addition of the ligands. E_{max} values were expressed as the percentage of the effect of each ligand over the effect of NE. BRET curves were analyzed by non-linear regression using the commercial Prism 4 (GraphPad Software).

B.3.3 DMR

A global cell signaling profile or DMR was measured using an EnSpire Multimode

Plate Reader (PerkinElmer, Waltham, MA, USA). This label-free approach uses refractive waveguide grating optical biosensors, integrated into 384-well microplates. Changes in local optical density are measured in a detection zone up to 150 nm above the surface of the sensor. Cellular mass movements induced upon receptor activation are detected by illuminating the underside of the biosensor with polychromatic light and measured as changes in the wavelength of the reflected monochromatic light. These changes are a function of the refraction index. The magnitude of this wavelength shift (in picometers) is directly proportional to the amount of DMR. Briefly, after 24 h of CHO cell transfection with α_{2A}-RLuc8 receptor, cells were resuspended and seeded at a density of 7000 cells per well in 384-well sensor microplates in 30 µl growing media and cultured for 24 h at 37 °C and 5% CO_2, to obtain monolayers at 70–80% confluency. Before starting the assay, cells were washed twice with assay buffer (MEMα supplemented with 20 mM HEPES, pH 7.15, 0.1% DMSO, and 0.1% BSA) and incubated 2 h in 40 µl per well in the reader at 24 °C. Hereafter, the sensor plate was scanned, and a baseline optical signature was recorded for 10 min before adding 10 µl of the agonist dissolved in assay buffer and recorded for 90 min. Kinetic results were analyzed using EnSpire Workstation Software v 4.10.

B.3.4 Adenylyl Cyclase Activity

BRET assays were performed to detect receptor ligand-induced adenylyl cyclase activity. This assay used the CAMYEL biosensor construct which contains RLuc and YFP. The biosensor detects the conformational changes in Epac that are induced upon its binding to cAMP. The conformational change triggered by an increase in cAMP induced by forskolin results in a decrease in BRET due to the relative orientation change between donor and acceptor. A decrease in forskolin-induced cAMP levels is therefore observed as

an increase in BRET [482]. To study Gαi-dependent inhibition activity, cells were treated as described above but pre-stimulated for 10 min with 1 μM forskolin (Sigma-Aldrich), in the presence of 10 μM propranolol 10 min before sample reading to control for activation of endogenous β-adrenergic receptors (see "Results").

B.3.5 ERK1/2 Phosphorylation

CHO cells were transfected with α_{2A}-RLuc8 receptor, obtaining a transfection of about 0.3 pmol/mg protein. The day of the experiment, cells were starved by treating them with serum-free media for 4 h at 37 °C. After that, cells were incubated with the indicated agonist for 5 min at 37 °C. Then, cells were rinsed with ice-cold phosphate-buffered saline and lysed by adding 200 ml ice-cold lysis buffer (50 mM Tris-HCl (pH 7.4), 50 mM NaF, 150 mM NaCl, 45 mM β-glycerophosphate, 1% Triton X-100, 20 mM phenylarsine oxide, 0.4 mM $NaVO_4$, and protease inhibitor cocktail). The cellular debris was removed by centrifugation at 13,000*g* for 5 min at 4 °C, and the protein was quantified. To determine the level of ERK1/2 phosphorylation, equivalent amounts of protein were separated by electrophoresis on a denaturing 10% SDS polyacrylamide gel and transferred onto polyvinylidene fluoride membranes. Odyssey blocking buffer (LI-COR Biosciences, Lincoln, NE, USA) was then added, and the membrane was rocked for 90 min. The membranes were then probed with a mixture of a mouse anti-phospho-ERK1/2 antibody (1:2500; Sigma-Aldrich) and rabbit anti-ERK1/2 antibody that recognizes both phosphorylated and non-phosphorylated ERK1/2 (1:40,000; Sigma-Aldrich) overnight at 4 °C. The 42- and 44-kDa bands corresponding to ERK1 and ERK2 were visualized by the addition of a mixture of IRDye 800 (anti-mouse) antibody (1:10,000; Sigma-Aldrich) and IRDye 680 (anti-rabbit) antibody (1:10,000; Sigma-Aldrich) for 2 h and scanned by the

Odyssey infrared scanner (LICOR Biosciences). Band densities were quantified using the scanner software and exported to Excel (Microsoft, Redmond, WA, USA). The level of phosphorylated ERK1/2 isoforms was normalized for differences in loading using the total ERK1/2 protein band intensities.

B.3.6 Radioligand Binding

Brains of male and female sheep of 4–6 months old were freshly obtained from the local slaughterhouse. Brain tissues (cortex and dorsal striatum) and HEK-293T cell suspensions were disrupted with a Polytron homogenizer (PTA 20 TS rotor, setting 3; Kinematica, Basel, Switzerland) for two 5-s periods in 10 volumes of 50 mM Tris-HCl buffer, pH 7.4, containing a proteinase inhibitor cocktail (Sigma, St. Louis, MO, USA). Membranes were obtained by centrifugation twice at 105,000*g* for 45 min at 4 °C. The pellet was stored at − 80 °C, washed once more as described above, and resuspended in 50 mM Tris-HCl buffer for immediate use. Membrane protein was quantified by the bicinchoninic acid method (Pierce Chemical Co., Rockford, IL, USA) using bovine serum albumin dilutions as standard. Binding experiments were performed with membrane suspensions at room temperature in 50 mM Tris-HCl buffer, pH 7.4, containing 10 mM $MgCl_2$. For competition-binding assays, membrane suspensions (0.2 mg of protein/ml) were incubated for 2 h with a constant-free concentration of 0.9 nM of the α_2R-antagonist [^{3}H]RX821002 or 1.3 nM of the D_1-like receptor antagonist [^{3}H]SCH 23390 or 0.8 nM of the D_2-like receptor antagonist [^{3}H]YM-09151-2 and increasing concentrations of each tested ligand: NE, DA, clonidine, 7-OH-PIPAT, quinpirole, and RO-105824. For α_2R saturation-binding assays, membrane suspensions (0.2 mg of protein/ml) were incubated for 3 h at room temperature in 50 mM Tris-HCl buffer, pH 7.4, containing 10 mM

$MgCl_2$ with increasing concentrations of the α_2R-antagonist [^{3}H]RX821002. Non-specific binding was determined in the presence of 10 μM of the non-radiolabeled antagonist RX821002 (for α_2R) or 30 μM of DA (for D_1R and D_2R). In all cases, free and membrane-bound ligands were separated by rapid filtration of 500-μl aliquots in a cell harvester (Brandel, Gaithersburg, MD, USA) through Whatman GF/C filters embedded in 0.3% polyethylenimine that were subsequently washed for 5 s with 5 ml of ice-cold 50 mM Tris-HCl buffer. The filters were incubated with 10 ml of Ecoscint H scintillation cocktail (National Diagnostics, Atlanta, GA, USA) overnight at room temperature, and radioactivity counts were determined using a Tri-Carb 2800 TR scintillation counter (PerkinElmer) with an efficiency of 62%.

B.3.7 Binding Data Analysis

Data were analyzed according to the "two-state dimer model" of Casadó et al. [483]. The model assumes GPCR dimers as a main functional unit and provides a more robust analysis of parameters obtained from saturation and competition experiments with orthosteric ligands, as compared with the commonly used "two-independent site model" [483, 484]. In saturation experiments with the radioligand, the model analyzes the total number of radioligand binding sites (B_{max}; more specifically, it calculates R_T, the total number of dimers, where $B_{max} = 2R_T$), the affinity of the radioligand for the first protomer in the unoccupied dimer (K_{DA1}), the affinity of the radioligand for the second protomer when the first protomer is already occupied by the radioligand (K_{DA2}), and an index of cooperativity of the radioligand (D_{CA}). A positive or negative value of D_{CA} implies either an increase or a decrease in affinity of K_{DA2} versus K_{DA1}, and its absolute value provides a measure of the degree of increase or decrease in affinity. In competition experiments, the

model analyzes the interactions of the radioligand with a competing ligand and it provides the affinity of the competing ligand for the first protomer in the unoccupied dimer (K_{DB1}), the affinity of the competing ligand for the second protomer when the first protomer is already occupied by the competing ligand (K_{DB2}) or the radioligand (K_{DAB}), and an index of cooperativity of the competing ligand (D_{CB}). A positive or negative value of D_{CB} implies either an increase or a decrease in affinity of K_{DB2} versus K_{DB1}, and its absolute value provides a measure of the degree of increase or decrease in affinity.

Radioligand competition and saturation curves were analyzed by non-linear regression using the commercial GraFit curve-fitting software (Erithacus Software, Surrey, UK), by fitting the binding data to the mechanistic two-state dimer receptor model, as described in detail elsewhere [485]. The equation describing the saturation experiment with the radioligand A in non-cooperative conditions ($K_{DA2} / K_{DA1} = 4$) is as follows: $A_{bound} = 2AR_T / (2K_{DA1} + A)$, where A represents the radioligand concentration. To calculate the macroscopic equilibrium dissociation constants from competition experiments, the following general equation must be applied: $A_{bound} = (K_{DA2}A + 2A^2 + K_{DA2}AB / K_{DAB})R_T / (K_{DA1}K_{DA2} + K_{DA2}A + A^2 + K_{DA2}AB / K_{DAB} + K_{DA1}K_{DA2}B / K_{DB1} + K_{DA1}K_{DA2}B^2 / (K_{DB1}K_{DB2}))$, where B represents the assayed competing compound concentration (F). For A, the non-cooperative and non-allosteric modulation between A and B, the equation is simplified due to the fact that $K_{DA2} = 4K_{DA1}$ and $K_{DAB} = 2K_{DB1}$; $A_{bound} = (4K_{DA1}A + 2A^2 + 2K_{DA1}AB / K_{DB1})R_T / (4K_{DA1}^2 + 4K_{DA1}A + A^2 + 2K_{DA1}AB / K_{DB1} + 4K_{DA1}^2B / K_{DB1} + 4K_{DA1}^2B^2 / (K_{DB1}K_{DB2}))$. For A and B, the non-cooperative and non-allosteric modulation between A and B, the equation can be simplified due to the fact that $K_{DA2} = 4K_{DA1}$, $K_{DB2} = 4K_{DB1}$, and $K_{DAB} = 2K_{DB1}$; $A_{bound} = (4K_{DA1}A + 2A^2 + 2K_{DA1}AB / K_{DB1})R_T / (4K_{DA1}^2 + 4K_{DA1}A + A^2 + 2K_{DA1}$

$AB / K_{DB1} + 4K_{DA1}^2B / K_{DB1} + K_{DA1}^2B^2 / K_{DB1}^2$).

B.3.8 Statistical Analysis

In binding assays, goodness of fit was tested according to reduced chi-square value given by the regression program. The test of significance for two different model population variances was based upon the *F*-distribution. Using this *F*-test, a probability greater than 95% ($p < 0.05$) was considered to be the criterion to select a more complex model (cooperativity) over the simplest one (non-cooperativity). In all cases, a probability of less than 70% ($p > 0.30$) resulted when one model was not significantly better than the other. In all cases, results are given as parameter values ± SEM and statistical differences were analyzed with Prism 4.

B.3.9 Drugs

Dopamine hydrochloride and L-(−)-norepinephrine (+)-bitartrate salt monohydrate were purchased from Sigma. (−)-Quinpirole hydrochloride, clonidine hydrochloride, 7-OH-PIPAT maleate, RO-105824 dihydrochloride, RX821002, and yohimbine hydrochloride were purchased from Tocris. [^{3}H]RX821002 (63.9 Ci/mmol), [^{3}H]SCH 23390 (81.9 Ci/mmol), and [^{3}H]YM-09151-2 (84.4 Ci/mmol) were from PerkinElmer. Pertussis toxin was purchased from Sigma.

B.3.10 Homology Modeling of α_{2A}- and α_{2C}-Adrenoceptors

Homology models of α_{2A}- and α_{2C}-adrenoceptors were constructed from multiple templates using RosettaCM [5] with a protocol previously described [188]. Sequences of each adrenoceptor were aligned with sequences of the following receptors: D_3 (PDB ID:

3PBL [26]), β_1 (PDB ID: 4BVN [486]), β_2 (PDB ID: 2RH1 [24]), $5HT_{1B}$ (PDB ID: 4IAR [487]), and $5HT_{2B}$(PDB ID: 4IB4 [488]) using BLAST and were modified to ensure alignment of secondary structure elements and conserved residues. The N-terminus was truncated through residues 28 and 46 and the C-terminus was deleted after residues 442 and 456, respectively. Additionally, the long intracellular loop 3 was deleted at residues 229–372 in α_{2A} and 243–381 in α_{2C} and replaced with an eight-residue poly-Gly linker. These sequences were threaded onto each template and hybridized to generate full-length, energy-minimized structures. Models were clustered using automatic radius detection in Rosetta, and the low-energy cluster centers from the top five clusters were selected for additional modeling.

B.3.11 Protein-Ligand Docking

The tridimensional structure of DA was obtained from PubChem (ID 3713609). Conformers of DA were generated using the BCL [330]. To identify the initial starting coordinates for ligand docking, homology models were aligned with the crystal structure of β_2-adrenoceptor (PDB ID: 4LDO (33)) and DA was aligned with the crystallized ligand. Ligand docking was performed in RosettaLigand using the small perturbation of ligand position protocol and swapping of ligand conformers [202, 251]. One thousand models for each protein-ligand complex were generated. Models were sorted initially by total energy and then culled to the top 5% of models by interface energies for analysis. Per-residue $\Delta\Delta G$ analysis was performed to identify residues involved in ligand binding.

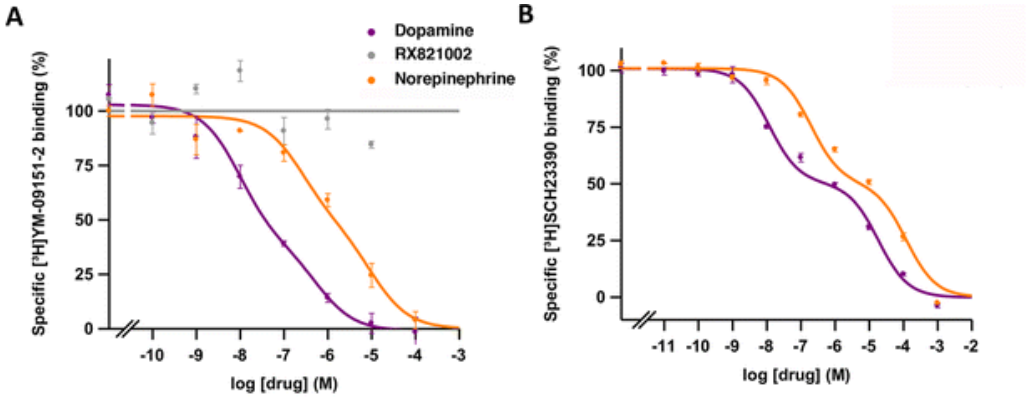


Figure B.1 Radioligand binding to D2-like and D1-like receptors in the brain striatum. Representative competition curves of D2-like receptor antagonist [3H]-YM-09151-2 vs. increasing concentrations of free competitors NE, DA, and RX821002 (a) and of D1-like receptor antagonist [3H]SCH 23390 vs. increasing concentrations of DA and NE (b) in the sheep brain striatum. Values are expressed as a percentage of the specific binding (100% is 0.13 ± 0.01 pmol/mg protein in a and 0.43 ± 0.04 pmol/mg protein in b). Experimental data were fitted to the two-state dimer receptor model equations, as described in the "Materials and Methods" section. Values are mean ± SEM from a representative experiment (n = 3–5) performed in triplicate

B.4 Results

B.4.1 Binding of DA and DA Receptor Ligands to α_2-Adrenoceptors in Cortical and Striatal Tissue

First, we analyzed the ability of NE, DA, the non-selective α-adrenoceptor agonist clonidine, the non-selective D_2-D_3-D_4 receptor agonist quinpirole, the selective D_3 receptor agonist 7-OH-PIPAT, and the selective D_4 receptor agonist RO-105824 to displace the binding of the non-selective α_2-adrenoceptor antagonist radioligand [^{3}H]RX821002 in membrane preparations from the sheep cortex and striatum with competitive inhibition experiments. See "Materials and Methods" and refs. [483-485] for description of the variables. Saturation experiments with [^{3}H]RX821002 for cortical and striatal tissue provided B_{max} values for α_2-adrenoceptors of 0.33 ± 0.02 and 0.13 ± 0.02 pmol/mg protein and affinity values (K_{DA1}) of 0.06 ± 0.01 and 0.07 ± 0.01 nM ($n = 4–8$), respectively. This

implies that the density of α_2-adrenoceptors in the cortex, which is mostly represented by α_{2A}-adrenoceptors [465, 466], is three times higher than that in the striatum, which expresses similar densities for both α_{2A}- and α_{2C}-adrenoceptors [470]. To test the binding selectivity of [^{3}H]RX821002 for α_2-adrenoceptors and not for D_2-like receptors, we developed competition experiments of the D_2-like receptor antagonist [^{3}H]YM-09151-2 with increasing concentrations of unlabeled RX821002 in sheep striatal membranes. RX821002 did not displace the radioligand binding at any concentration up to 10 μM (Figure B.1a). The same experiments were also performed in membranes from HEK-293T cells stably transfected with human D_2, D_3, or D_4 receptors, with identical results (data not shown). These results demonstrate that the radioligand [^{3}H]RX821002 does not bind to D_2-like receptors.

Competition experiments of [^{3}H]RX821002 with NE, DA, clonidine, and the D_2-like receptor ligands in cortical and striatal sheep membranes are shown in Figure B.2a, b, respectively, and the K_{DB1}, K_{DB2}, and D_{CB} values obtained are presented in Table B.1. In both tissues, NE, DA, and clonidine showed high affinity for [^{3}H]RX821002 binding sites

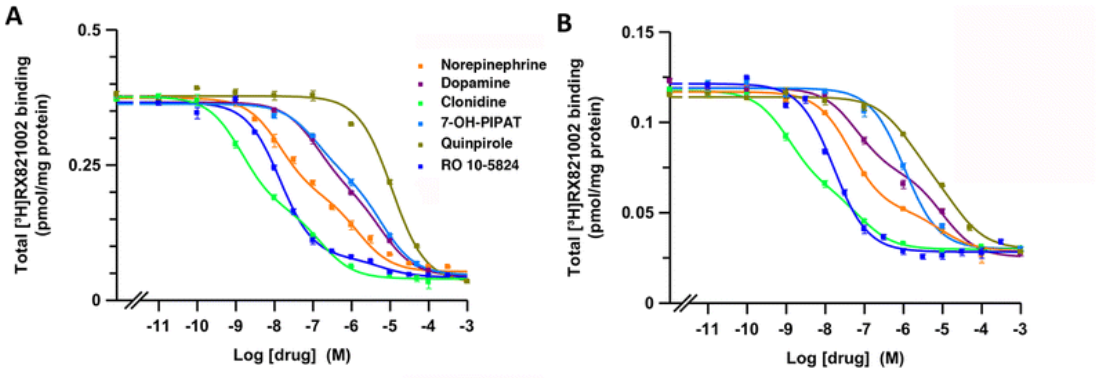


Figure B.2 Radioligand binding of dopaminergic and adrenergic ligands to α_2-adrenoceptors in brain tissue. Representative competition curves of α_2-adrenoceptor antagonist [^{3}H]RX821002 vs. increasing concentrations of free competitors (NE, DA, clonidine, quinpirole, 7-OH-PIPAT, and RO-105824) in sheep brain cortical (**a**) and striatal (**b**) membranes. Experimental data were fitted to the two-state dimer receptor model equations, as described in the "Materials and Methods" section. Values are mean ± SEM from a representative experiment (n = 3) performed in triplicate

with an order of potency of clonidine > NE > DA (Figure B.2). The three ligands showed negative cooperativity (negative D_{CB}values). The affinity of NE was higher in the cortex than in the striatum, with higher striatal K_{DB1}, K_{DB2}, and D_{CB} values (stronger negative cooperativity) (Table B.1). The affinity of DA was very similar in both tissues, with similar K_{DB1} values and a moderately but significantly higher K_{DB2} value in the striatum, resulting in similar D_{CB} values (Table B.1). The affinity of clonidine was also higher in the cortex, with a significantly higher striatal K_{DB1} value and similar D_{CB}values (Table B.1). 7-OH-PIPAT and quinpirole also displaced [^{3}H]RX821002 binding with nanomolar and submicromolar affinities, respectively (Table B.1, Figure B.2). Interestingly, 7-OH-PIPAT showed negative cooperativity in the cortex, but not in the striatum. The only measurable affinity parameter of 7-OH-PIPAT in the striatum, K_{DB1}, was significantly higher than that in the cortex, and it was almost ten times lower than the cortical K_{DB2} value (Table B.1). Quinpirole also showed differences in the binding parameters between the cortex and striatum, such as a lower K_{DB1} value but negative cooperativity in the striatum. Finally, RO-105824 also displaced [^{3}H]RX821002 binding from the cortex and striatum with high affinity (subnanomolar). No cooperativity (D_{CB} = 0) was obtained, except for RO-105824 in the cortex (D_{CB} = − 4.3) (Table B.1).

B.4.2 Binding of DA and NE to D_1-Like and D_2-Like Receptors in Striatal Tissue

Next, we compared the affinity of endogenous DA and NE binding to DA D_1-like and D_2-like receptors with the affinity, determined above, for α_2-adrenoceptors. In addition to competition experiments with the D_2-like radioligand antagonist [^{3}H]YM-09151-2 (Figure B.1A), we performed competition experiments with the D_1-like radioligand

antagonist [^{3}H]SCH 23390 in sheep striatal preparations (Figure B.1b). The equilibrium binding parameters are shown in Table B.2. Both NE and DA showed negative

Table B.1 Competitive inhibition experiments of [^{3}H]RX821002 versus NE, DA, clonidine, and D_2-like receptor ligands in the sheep brain cortex and striatum

*$p < 0.05$; **$p < 0.01$; ***$p < 0.001$

Ligand	**Binding parameters**	
	Cortex	**Striatum**
NE	K_{DB1} = 0.3 ± 0.2*	K_{DB1} = 0.8 ± 0.1
	K_{DB2} = 250 ± 100	K_{DB2} = 5000 ± 3000
	D_{CB} = − 2.3	D_{CB} = − 3.2
DA	K_{DB1} = 6.9 ± 0.2	K_{DB1} = 6 ± 1
	K_{DB2} = 350 ± 10*	K_{DB2} = 1000 ± 200
	D_{CB} = − 1.1	D_{CB} = − 1.6
Clonidine	K_{DB1} = 0.014 ± 0.003*	K_{DB1} = 0.036 ± 0.005
	K_{DB2} = 40 ± 20	K_{DB2} = 20 ± 10
	D_{CB} = − 2.8	D_{CB} = − 2.1
7-OH-PIPAT	K_{DB1} = 9 ± 2**	K_{DB1} = 51 ± 6
	K_{DB2} = 430 ± 80	
	D_{CB} = − 1.1	D_{CB} = 0
Quinpirole	K_{DB1} = 530 ± 50**	K_{DB1} = 110 ± 10
		K_{DB2} = 2700 ± 400
	D_{CB} = 0	D_{CB} = − 0.8
RO-105824	K_{DB1} = 0.055 ± 0.003***	K_{DB1} = 0.42 ± 0.03
	K_{DB2} = 4000 ± 2000	
	D_{CB} = − 4.3	D_{CB} = 0

Table B.2 Competitive inhibition experiments of [^{3}H]SCH 23390, [^{3}H]YM-09151-2, or [^{3}H]RX821002 versus DA and NE in the sheep brain striatum

For DA, $^{***}p < 0.001$ vs. D_2-like receptors; for NE, $^{\#}p < 0.05$ and $^{\#\#}p < 0.01$ vs. α_2 receptors

Receptor	Binding parameters	
	DA	NE
D_1-like	$K_{DB1} = 8 \pm 3$	$K_{DB1} = 53 \pm 90^{\#\#}$
	$K_{DB2} = 8000 \pm 1000^{***}$	$K_{DB2} = 50000 \pm 10{,}000^{\#}$
	$D_c = -2.4$	$D_c = -2.4$
D_2-like	$K_{DB1} = 3.5 \pm 0.6$	$K_{DB1} = 60 \pm 40^{\#\#}$
	$K_{DB2} = 700 \pm 200$	$K_{DB2} = 3400 \pm 100$
	$D_c = -1.7$	$D_c = -1.2$
α_2	$K_{DB1} = 6 \pm 1$	$K_{DB1} = 0.8 \pm 0.1$
	$K_{DB2} = 1000 \pm 200$	$K_{DB2} = 5000 \pm 3000$
	$D_c = -1.6$	$D_c = -3.2$

cooperativity (Figure B.1) with negative D_c values (Table B.2). The affinity of DA for the first protomer of the D_2-like receptor dimer (K_{DB1}), mostly represented by the D_2R subtype in the dorsal striatum [489], was similar to the affinity for α_2-adrenoceptors and D_1-like receptors (Figure B.1, Table B.2). The affinity of DA for the second protomer (K_{DB2}) of the D_1-like receptor was even about ten times lower than that for the α_2-adrenoceptors and D_2-like receptors (stronger negative cooperativity; see Table B.2). On the other hand, the affinity of NE for α_2-adrenoceptors in the striatum, represented by α_{2A}- and α_{2C}-adrenoceptors, was significantly higher than that for dopamine receptors (Figure B.1). Specifically, NE had an affinity for DA receptors about 60-fold lower than that for α_2-

adrenoceptors (Table B.2, Figure B.1). We can therefore assume that extracellular physiological levels of DA that are able to activate DA receptors are also able to bind α_2-adrenoceptors.

B.4.3 α_{2A}- and α_{2C}-Adrenoceptor-Mediated G Protein Activation by DA and Synthetic DA Receptor Ligands

The G protein activation BRET assay (see "Materials and Methods") was used to determine the potency and efficacy of the above-studied ligands to activate α-adrenoceptors in HEK-293T cells transfected with α2A- or α2C-adrenoceptor and one of the five different RLuc-fused Gαi/o subunits (Gαi1, Gαi2, Gαi3, Gαo1, and Gαo2) with

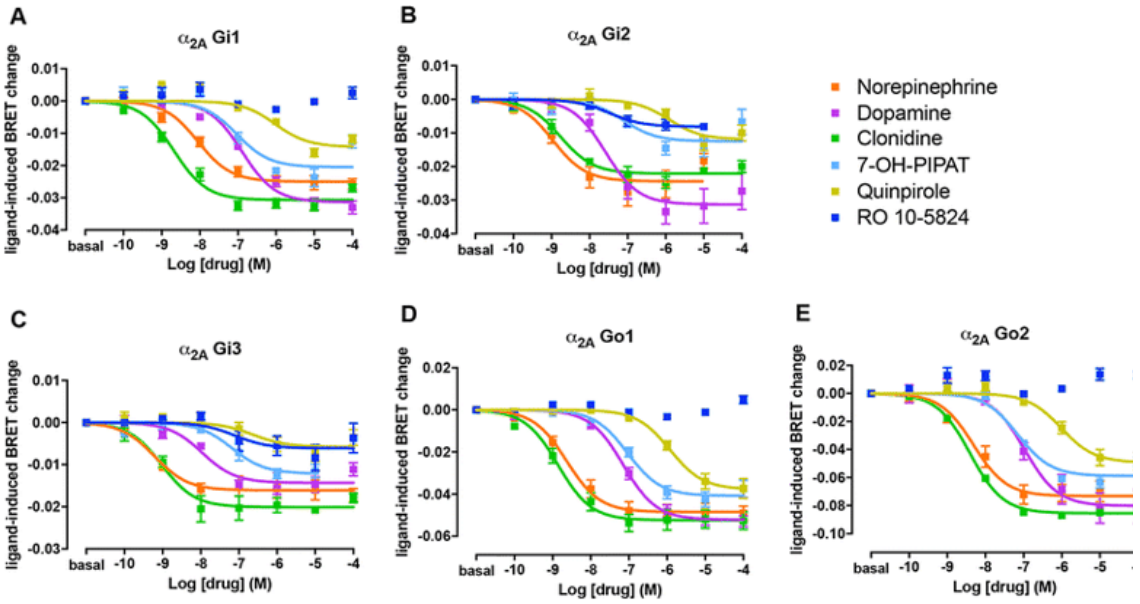


Figure B.3 G protein activation of α_{2A} by dopaminergic and adrenergic ligands. Concentration-response experiments of G protein activation by NE, DA, clonidine, and D_2-like receptor ligands mediated by α_{2A} in HEK-293T cells transiently transfected with α_{2A}-adrenoceptor; the G protein subunits Gαi1-RLuc (**a**), Gαi2-RLuc (**b**), Gαi3-RLuc (**c**), Gαo1-Rluc (**d**), or Gαo2-RLuc (**e**); γ2-mVenus, and non-fused β1. Cells were treated with coelenterazine H followed by increasing concentrations of one of the ligands. Ligand-induced changes in BRET values were measured as described in the "Materials and Methods" section. BRET values in the absence of ligands were subtracted from the BRET values for each agonist concentration. Data were adjusted to a sigmoidal concentration-response function by non-linear regression analysis and represent means ± SEM of 3 to 11 experiments performed in triplicate (see Table B.1 and Table B.2 for EC_{50} and E_{max} values and statistical analysis)

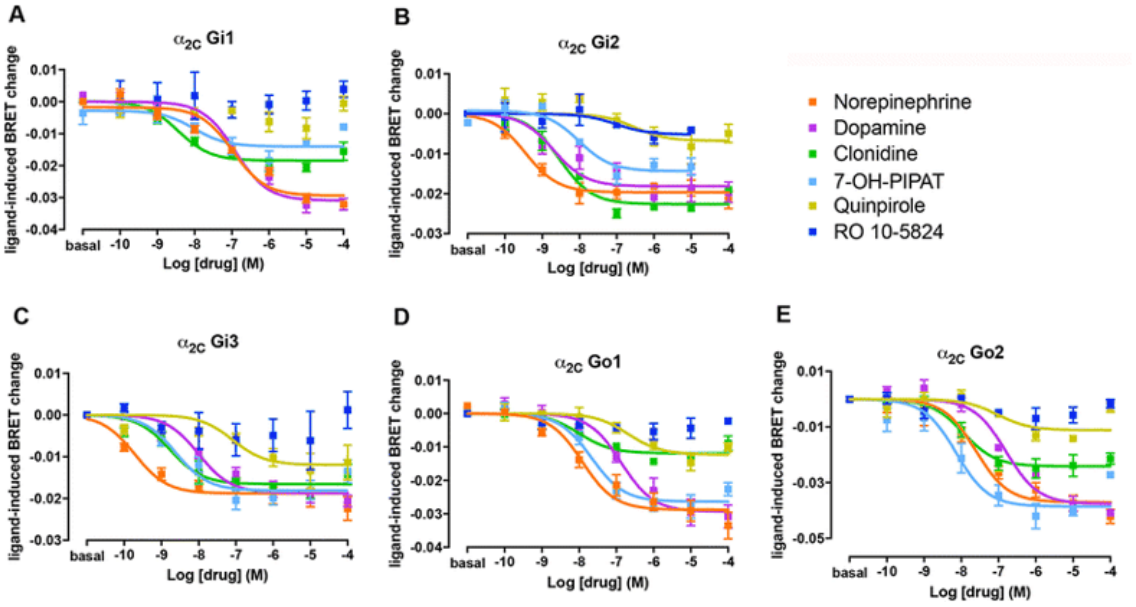


Figure B.4 G protein activation of α_{2C} by dopaminergic and adrenergic ligands. Concentration-response experiments of G protein activation by NE, DA, clonidine, and D_2-like receptor ligands mediated by α_{2C} in HEK-293T cells transiently transfected with α_{2C} receptor; the G protein subunits Gαi1-RLuc (**a**), Gαi2-RLuc (**b**), Gαi3-RLuc (**c**), Gαo1-RLuc (**d**), or Gαo2-RLuc (**e**), γ2-mVenus; and non-fused β1. Cells were treated with coelenterazine H followed by increasing concentrations of one of the ligands. Ligand-induced changes in BRET values were measured as described in the "Materials and Methods" section. BRET values in the absence of ligands were subtracted from the BRET values for each agonist concentration. Data were adjusted to a sigmoidal concentration-response function by non-linear regression analysis and represent means ± SEM of three to nine experiments performed in triplicate (see Table B.1 and Table B.2 for EC_{50} and E_{max} values and statistical analysis)

Venus-fused Gγ2 protein as BRET acceptor pair. The amount of Gαi/o subunits transfected produced values between 0.5 and 1 million luminescence units (arbitrary units). Previously, we found that luminescence values between 200,000 and 1.5 million did not produce a significant alteration of the Emax of drug-induced BRET. Moreover, the levels of α2-adrenoceptor were around 2 pmol/mg protein. We also previously reported that, using the same cell line and assay conditions, neither DA or NE produces a significant BRET change when transfected with the same fused G protein subunits but without receptor co-transfection [480]. These transfected receptor levels were only slightly higher than those obtained in binding experiments in the sheep brain cortex (see above). A concentration response of the ligand-induced change in BRET values allows to determine the potency as well as the relative efficacy (to NE) at α2A- and α2C-adrenoceptor-mediated G protein

Table B.3 Potency of NE, DA, clonidine, 7-OH-PIPAT, and quinpirole obtained from G protein activation experiments mediated by α_{2A}- and α_{2C}-adrenoceptors coupled to the different Gαi/o subtypes

*$p < 0.05$; **$p < 0.01$; ***$p < 0.001$

[a]The ratio of EC_{50} values of DA and NE for each receptor and Gαi/o protein subtype

Gα subunit	Receptor	NE	DA	Clonidine	7-OH-PIPAT	Quinpirole	DA/NE[a]
Gαi1	α_{2A}	11 ± 2**	170 ± 40	3 ± 1	80 ± 20*	700 ± 250	15
	α_{2C}	90 ± 30	150 ± 40	6 ± 2	(11 ± 4)	ND	1.7
Gαi2	α_{2A}	1.3 ± 0.3*	30 ± 3**	2.0 ± 0.8	120 ± 6*	(1000 ± 300)	23
	α_{2C}	0.4 ± 0.2	5 ± 3	3 ± 1	10 ± 2	(200 ± 100)	12.5
Gαi3	α_{2A}	0.6 ± 0.2	15 ± 5	1.0 ± 0.2	60 ± 20*	(700 ± 400)	25
	α_{2C}	0.4 ± 0.2	30 ± 20	4 ± 1	5 ± 3	400 ± 100	75
Gαo1	α_{2A}	3.0 ± 0.5*	80 ± 10**	2.0 ± 0.4*	100 ± 20**	1300 ± 200**	27
	α_{2C}	19 ± 6	126 ± 8	(12 ± 5)	20 ± 3	(230 ± 50)	7
Gαo2	α_{2A}	6 ± 1*	100 ± 20	4.0 ± 0.2**	66 ± 7*	820 ± 80***	17
	α_{2C}	50 ± 10	140 ± 20	11 ± 2	14 ± 7	(100 ± 10)	2.8

activation. Results were largely in agreement with the values obtained with binding experiments, considering that cortical values should represent ligand binding parameters of α2A-adrenoceptors, while striatal values represent combined ligand binding parameters for both α2A- and α2C-adrenoceptors. NE was more potent at α2A- than at α2C-adrenoceptor, except for Gαi2 and Gαi3. On the other hand, DA had similar potencies at both adrenoceptors, except for Gαi2 and Gαo1. At both α2A- and α2C-adrenoceptors, DA showed high potency and efficacy as compared with NE (Figure B.3 and Figure B.4, Table B.3 and Table B.4), although DA was always less potent than NE. The relative DA/NE potency depended on the α-adrenoceptor and on the Gαi/o subtype (see Table B.3). Therefore, the potencies of DA at activating α2A-adrenoceptor varied from about 15-fold lower, for Gαi1, to about 30-fold lower, for Gαo1. On the other hand, the potencies of DA as compared to NE at activating α2C-adrenoceptor were very close to those of NE and they varied from less than twofold lower, for Gαi1, to about 70-fold lower, for Gαi3 (Table B.3).

Table B.4 Efficacy of NE, DA, clonidine, 7-OH-PIPAT, and quinpirole obtained from G protein activation experiments mediated by α_{2A}- and α_{2C}-adrenoceptors coupled to the different Gαi/o subtypes

ND not detectable

$^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$

Gα subunit	Receptor	NE	DA	Clonidine	7-OH-PIPAT	Quinpirole
Gαi1	α_{2A}	100 ± 5	120 ± 6*	120 ± 5	94 ± 6	70 ± 2*
	α_{2C}	100 ± 2	105 ± 5	64 ± 4**	*41 ± 1***	ND
Gαi2	α_{2A}	100 ± 10	120 ± 20	89 ± 6	55 ± 2*	*47 ± 8**
	α_{2C}	100 ± 7	80 ± 10	107 ± 4	90 ± 20	*46 ± 9***
Gαi3	α_{2A}	100 ± 10	90 ± 10	112 ± 8	71 ± 6	*40 ± 9***
	α_{2C}	100 ± 9	95 ± 6	85 ± 5	90 ± 10	80 ± 15
Gαo1	α_{2A}	100 ± 7	108 ± 6	110 ± 10	81 ± 9	78 ± 9
	α_{2C}	100 ± 10	100 ± 10	*41 ± 3***	98 ± 7	*42 ± 3***
Gαo2	α_{2A}	100 ± 9	110 ± 10	120 ± 2	85 ± 4	72 ± 8
	α_{2C}	100 ± 8	114 ± 6	70 ± 8*	106 ± 3	*34 ± 6***

The prototypical non-selective α-adrenoceptor agonist clonidine only showed a significantly higher potency at α_{2A}- than at α_{2C}-adrenoceptors for Gαo1 and Gαo2 (Table B.3). An additional difference as compared to NE was that clonidine behaved as a full agonist at α_{2A}- and as a partial agonist at α_{2C}-adrenoceptor, except for Gαi2 and Gαi3 (Figure B.3 and Figure B.4, Table B.4). Intriguingly, the level of efficacy of clonidine for α_{2C}-adrenoceptor varied significantly with the associated Gαi/o protein subtypes, from no decrease for Gαi2 to a very significant loss of efficacy for Gαo1 (Table B.4). Previous studies have already reported a partial agonism of clonidine at α_2-adrenoceptors, but with disparate results [490, 491], which, at least for α_{2C}, could be related to the Gαi/o protein subtypes involved. In summary, the differences in the respective potency values of NE, clonidine, and DA for α_{2A}- and α_{2C}-adrenoceptors in the G protein activation BRET experiments correlate with the higher affinities of NE and clonidine in the cortex and similar affinities of DA in the cortex and striatum.

The non-selective D_2-D_3-D_4 receptor agonist quinpirole and the selective D_3 receptor

agonist 7-OH-PIPAT also activated α_{2A}- and α_{2C}-adrenoceptors, but with very different profiles (see Figure B.3 and Figure B.4, Table B.3 and Table B.4). 7-OH-PIPAT behaved as a low efficacious agonist at α_{2A}-adrenoceptors for the Gαi2 subtype. On the other hand, 7-OH-PIPAT behaved as a partial agonist at α_{2C}-Gαi1 complexes but as a full agonist with the other α_{2C}-Gαi/o complexes. This D_3 receptor agonist, at α_{2A}-adrenoceptor, was, in general, as potent as DA and, for both Gαo subtypes, was as potent α_{2C}-adrenoceptor agonist as NE. In contrast, quinpirole showed a weak potency (submicromolar range) but also behaved as a partial or full agonist depending on the Gαi/o subtype. At α_{2A}, quinpirole behaved as a partial agonist for Gi1, Gi2, and Gi3 and a full agonist for Go1 and Go2, whereas at α_{2C}, it behaved as a partial agonist for all G protein subtypes except for Gi3 (full agonist) and showed no activity when coupled with Gαi1. As shown in Figure B.5, yohimbine, a non-selective α_2-adrenoceptor antagonist, completely blocked the effect of 7-OH-PIPAT and quinpirole at both α_{2A}- and α_{2C}-adrenoceptors (for Gαo1), demonstrating the specificity of the α_2-adrenoceptor signal produced by both agonists. The potency values of 7-OH-PIPAT and quinpirole in G protein activation BRET experiments correlate with the nanomolar and submicromolar affinities, respectively, as seen in binding assays with brain membranes. Moreover, the higher potencies of 7-OH-PIPAT for α_{2C}- versus α_{2A}-adrenoceptors also correlate with our binding results due to the fact that 7-OH-PIPAT showed negative cooperativity in the cortex but not in the striatum. For quinpirole, however, it would be difficult to establish correlations with results from binding assays due to its low efficacy in BRET assays, which might lead to inaccurate values (Table B.3, in parentheses, and Table B.4Table B.4, in italics). Finally, and unexpectedly, the selective D_4 receptor agonist RO-105824 (with subnanomolar affinity for α-adrenoceptors) did not produce a significant activation of α_{2A}- or α_{2C}-adrenoceptors coupled to any of the Gαi/o

subtypes, except for a small efficacy at α_{2A} for Gαi2 and Gαi3 (Figure B.3 and Figure B.4). To confirm the binding of this putative selective D_4receptor ligand to α_{2A}- and α_{2C}-adrenoceptors, they were tested for their ability to modify the effect of clonidine. RO-105824 did counteract the respective full and partial agonistic effect of clonidine (1 μM) at the α_{2A}- and α_{2C}-adrenoceptors coupled to Gαo1 (Figure B.5c, d). Therefore, the results of BRET and radioligand binding experiments disclosed a previously unknown additional role

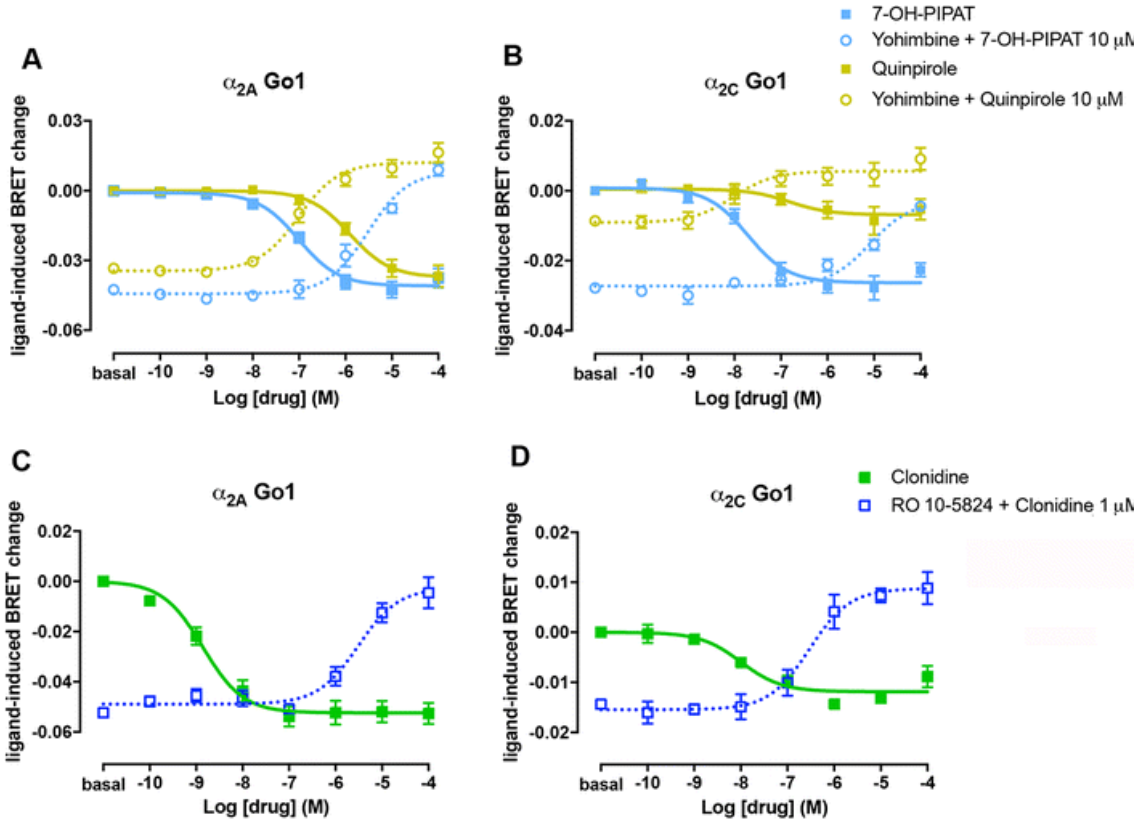


Figure B.5 Specificity of the effect of dopaminergic ligands on α_{2A}- and α_{2C}-adrenoceptors. **a**, **b** Dose-dependent inhibition by increasing concentrations of the non-selective α_2 receptor antagonist yohimbine of Gαo1 protein activation induced by 10 μM of the D_2-like receptor agonists 7-OH-PIPAT (dotted blue) or 10 μM of quinpirole (dotted yellow) in HEK-293T cells transfected with α2A(a) or α2C (b) adrenoceptors, Gαo1-RLuc, γ2-mVenus, and non-fused β1. As controls, concentration-response curves of Gαo1 protein activation by 7-OH-PIPAT (full blue) or quinpirole (full yellow) from Figs. 3d and 4d are showed. Cells were treated with coelenterazine H followed by the addition of 7-OH-PIPAT or quinpirole. Ligand-induced changes in BRET values were measured as described in the "Materials and Methods" section. c, d Dose-dependent inhibition of the agonistic effect of clonidine at 1 μM by increasing concentrations of RO-105824 (dark blue) in cells transfected with α2A (c) or α2C (d) adrenoceptors, Gαo1-RLuc, γ2-mVenus, and non-fused β1. As controls, concentration-response curves of Gαo1 protein activation by clonidine (full green) from Figs. 3d and 4d are showed. Data were adjusted to a sigmoidal concentration-response function by non-linear regression analysis and represent means ± SEM of three to six experiments performed in triplicate

of the D_4 receptor agonist RO-105824, as a very potent and low-efficacious ligand for α_2-adrenoceptors at Gαi/o activation.

We also determined the effect of DA and synthetic DA receptor ligands on a DMR label-free assay in CHO-transfected cells. This approach detects changes in local optical density due to cellular mass movements induced upon receptor activation (see the "Materials and Methods" section), and DMR responses primarily reflect G protein-dependent signaling in living cells, since it can be abrogated by toxins or inhibitors of the G proteins involved [492]. DA was as capable as NE at activating cellular signaling in CHO cells transfected with α_{2A}-RLuc8 receptor (Figure B.6a). The amount of α_{2A}-RLuc8 receptor expressed was about 0.3 pmol/mg protein. DA and NE activation decreased by adding the α_{2A}R-antagonist BRL 44408, indicating the specificity of the cell activation through α_{2A}-receptor. The synthetic DA receptor ligands 7-OH-PIPAT, quinpirole, and RO-105824 at 300 nM were also able to produce a significant response (Figure B.6b),

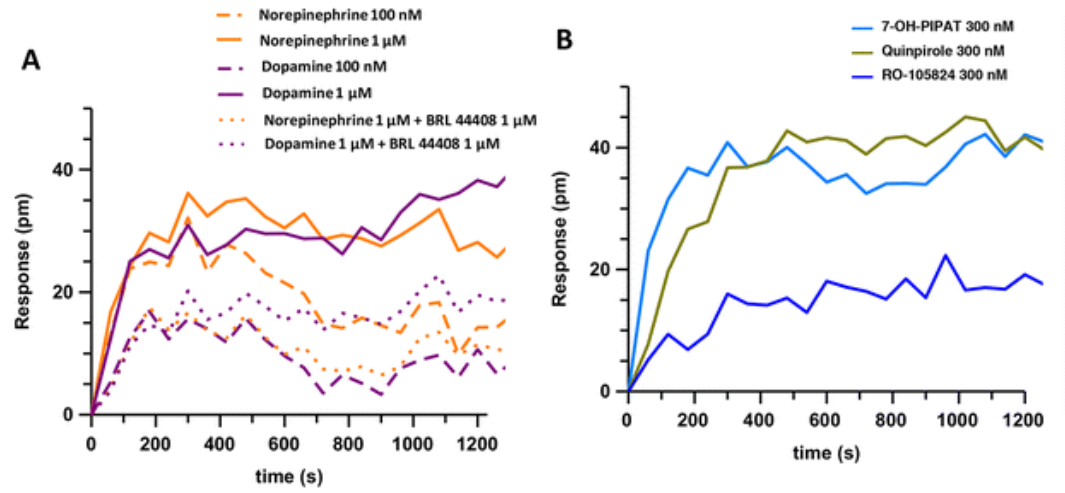


Figure B.6 DMR induced by NE, DA, and synthetic DA receptor ligands using label-free assay. DMR assay was performed in CHO cells transfected with α_{2A}-adrenoceptor. **a** Cells were pretreated (dotted lines) or not (full lines) with the α_{2A}-adrenoceptor antagonist BRL 44408 at 1 µM for 30 min before adding the endogenous agonists DA or NE at 300 nM or 1 µM. **b** Cells were treated with 100 nM of several synthetic DA receptor ligands. The resulting shifts of reflected light wavelength (pm) were monitored over time. Each panel is a representative experiment of n = 3 different experiments. Each curve is the mean of a representative optical trace experiment carried out in triplicates

substantially lower for RO-105824, which correlates with the G protein activation BRET assays (Figure B.3). These results indicate that DA and synthetic DA receptor ligands are also α_2-adrenoceptor ligands able to activate Gαi/o proteins, which correlate with their efficacy with DMR.

B.4.4 α_{2A}- and α_{2C}-Adrenoceptor-Mediated Effects of NE and DA on Adenylyl Cyclase Activity

NE- and DA-induced changes in adenylyl cyclase activity were also analyzed by measuring cAMP levels in intact cells transiently transfected with α2A- or α2C-adrenoceptor, using the CAMYEL BRET biosensor (see the "Materials and Methods" section and ref. [482]). HEK-293T cells have been reported to endogenously express β-adrenoceptors [493]. Accordingly, we recently reported that NE, in non-transfected HEK-293T cells, stimulated a Gs-mediated cAMP increase, which could be completely inhibited by the selective β-adrenergic blocker propranolol (10 μM) (ref. [480]; the same website address as above). Therefore, the β-adrenoceptor antagonist propranolol was added throughout the cAMP detection experiments. As shown in Figure B.7, NE and DA produced an increase in BRET, corresponding to a decrease in forskolin-induced cAMP accumulation for both α_{2A}- and α_{2C}-adrenoceptor-transfected cells. The decrease in adenylyl cyclase activity by NE and DA provided apparent half maximal effective concentration (EC_{50}) values that were qualitatively and quantitatively close to those observed with the Gi/o activation BRET assays, as NE was more potent than DA at α_{2A} and α_{2C} (1.4 ± 0.2 and 7 ± 4 nM for NE and 140 ± 40 and 90 ± 20 nM for DA, respectively). The putative Gi/o-dependent effects mediated by NE and DA were blocked by the non-selective α_2-adrenoceptor antagonist yohimbine, confirming the receptor specificity of the signal

(Figure B.7). In addition, cells were treated with pertussis toxin which catalyzes the ADP ribosylation of the αi subunits of the heterotrimeric G protein, preventing its interaction with the receptor. As expected, pertussis toxin treatment selectively eliminated the initial, therefore Gi-dependent, component of the NE- and DA-mediated effects (Figure B.7). Surprisingly, NE and DA showed inverted U-shaped concentration-response curves with a putative Gs-dependent profile at high ligand concentrations for both α_{2A}- and α_{2C}-adrenoceptor-transfected cells (Figure B.7). These results could be explained by considering previous studies showing that α_2-adrenoceptors functionally couple not only to

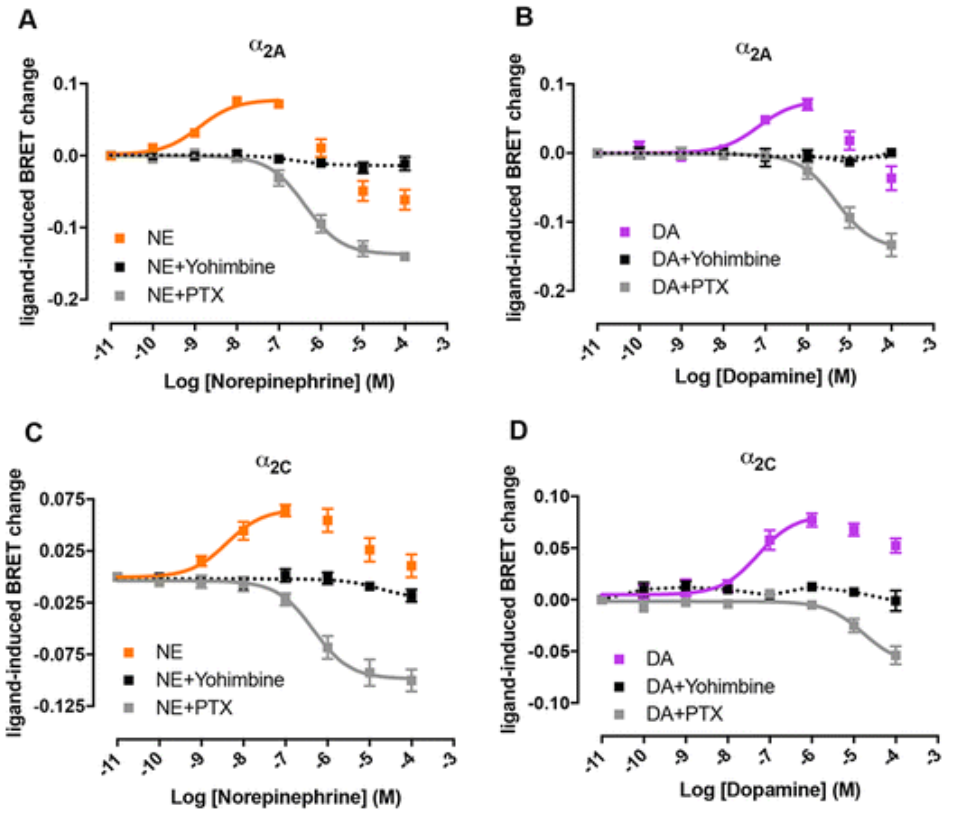


Figure B.7 Effect of NE and DA on the modulation of adenylyl cyclase activity by activation of α_{2A}- and α_{2C}-adrenoceptors. Concentration-response experiments of inhibition of forskolin-induced adenylyl cyclase activity by NE (orange) or DA (purple) mediated by α_{2A} (**a**, **b**) or α_{2C}-adrenoceptors (**c**, **d**) in HEK-293T cells transiently transfected with the CAMYEL sensor and one of the receptors. Cells were treated with forskolin (1 μM) for 10 min with or without the selective α_2 antagonist yohimbine (10 μM) followed by the addition of coelenterazine H and increasing concentrations of NE or DA. After 10 min, BRET was measured as described in the "Materials and Methods" section. In gray, cells were treated with 100 ng/ml pertussis toxin (PTX) for 16–18 h previous to the experiment. Values obtained with forskolin alone were subtracted from BRET values for each agonist concentration. Data represent the mean ± SEM of three to seven experiments performed in triplicate

Gi/o proteins but also to Gs [494-498]. Typically, the agonist concentrations necessary to elicit detectable stimulation of adenylyl cyclase are significantly higher than those for inhibition. Equivocal results were published by Zhang et al. [477] when comparing the effect of NE and DA on forskolin-induced adenylyl cyclase activation. In their cell systems, NE seemed to predominantly activate Gs with α_{2A} and Gi with α_{2C}, while DA would predominantly activate Gi with both receptors, but at high micromolar concentrations. In contrast, our experiments show that NE and DA follow the same differential concentration-dependent effects on Gi/o and Gs activation and, at least, at 10 and 100 μM, DA and NE promoted Gs activation mediated by both α_{2A}- and α_{2C}-adrenoceptors (Figure B.7).

B.4.5 α_{2A}- and α_{2C}-Adrenoceptor-Mediated Effects of NE, DA, and Synthetic DA Receptor Ligands on ERK1/2 Phosphorylation

Finally, we studied the ability of DA and synthetic DA receptor ligands to produce MAPK activation. First, we analyzed the increase on ERK1/2 phosphorylation produced by 300 nM of NE in CHO cells transfected with α_{2A}-RLuc8 receptor. This NE concentration increased ERK1/2 phosphorylation levels by threefold over basal, and this effect was similar to that produced by 1 μM of DA (Figure B.8). Next, we demonstrated that the synthetic DA receptor ligands 7-OH-PIPAT, quinpirole, and RO-105824, at 1 μM, were also able to produce MAPK activation (Figure B.8). At this concentration, the efficacy of RO-105824 was similar to that of the other ligands. Together with its very low efficacy disclosed on G protein activation and DMR assays, these results indicate that RO-105824 is a biased agonist of α_{2A}-adrenoceptors with functional selectivity for G protein-independent signaling. In summary, adenylyl cyclase activity and ERK1/2 phosphorylation

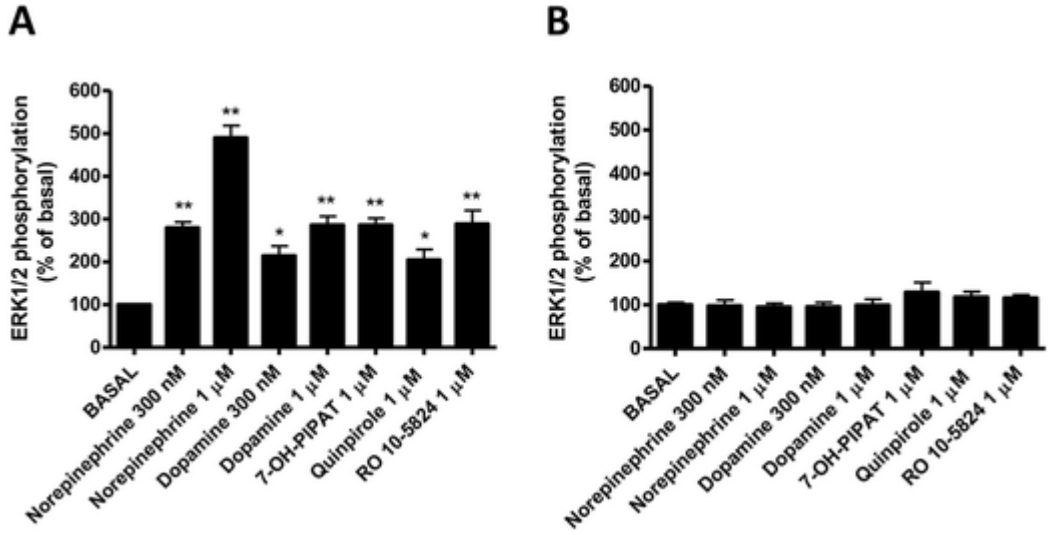


Figure B.8 NE, DA, and synthetic DA receptor ligands signaling via ERK1/2 phosphorylation. ERK1/2 phosphorylation was determined in CHO cells transfected with the α_{2A}-adrenoceptor (**a**) and non-transfected control cells (**b**), treated with 300 nM or 1 μM of the tested ligands for 5 min at 37 °C. Values are expressed as mean ± SEM (n = 6) of percentage of phosphorylation relative to basal levels in non-treated cells. Statistical differences vs. basal conditions were calculated by one-way ANOVA followed by Dunnett's post hoc test; *$p < 0.05$ and **$p < 0.01$

experiments confirm the results from radioligand binding and G protein activation assays, indicating that DA and synthetic DA receptor ligands are efficacious α_2-adrenoceptor agonists.

B.4.6 Structural Basis for DA at α_{2A}- and α_{2C}-Adrenoceptors

An examination of the binding mode of DA to the adrenoceptors was undertaken to model the activity seen in the biological assays. To generate models of α_{2A}- and α_{2C}-adrenoceptors, we used the RosettaCM [5] application within the Rosetta suite of macromolecular modeling tools [188, 223]. This method relies on the optimal alignment of a target sequence with multiple template structures obtained from the PDB which are hybridized together to generate novel models. We submitted the sequence of α_{2A} and α_{2C} to BLAST-P and found the top five crystal templates by sequence identity to be the DA D_3 receptor [26], β_1-adrenoceptor [486], β_2-adrenoceptor [24], serotonin 5-HT_{1B} receptor

[487], and serotonin 5-HT_{2B} receptor [488]. Interestingly, the α_2-adrenoceptors have more sequence identity in the transmembrane helical bundle with the D_3 receptor than with the related β-adrenoceptors. RosettaCM yielded an ensemble of low-energy models of the receptors, which were clustered by structural similarity. The top five cluster centers were included in the docking studies to account for structural diversity and uncertainty in homology modeling. To understand DA activation, we first examined the binding of DA to the D_3 receptor, the only crystal structure of a DA receptor to date. As there is not a co-crystal structure of DA/D_3 receptor, we first docked DA to the D_3receptor using RosettaLigand [202]. The starting coordinates of epinephrine bound to the β_2-adrenoceptor [499] were used to place DA for docking. Binding pocket analysis identified residue D3.32 interacting with the primary amine in DA and the catechol hydroxyls interacting with S5.42 and S5.46. Important hydrophobic packing against the central portion of DA was achieved by V3.33, H6.55, and F6.51. These are the same interactions that were previously identified in a molecular dynamics simulation of DA binding at D_3 receptor [500]. Comparing the residues at these positions to those in α_{2A}- and α_{2C}-adrenoceptors revealed

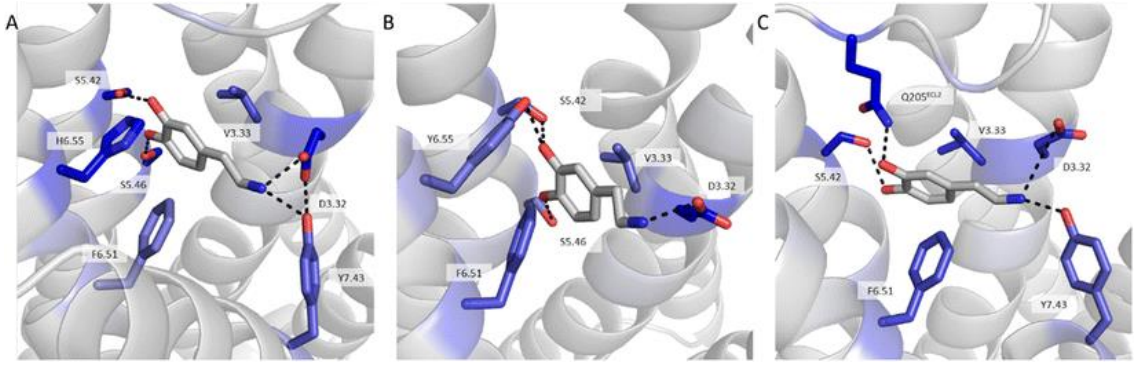


Figure B.9 Binding mode of DA at each receptor type. Shown is the docking orientation of DA at the D_3receptor (**a**), α_{2A}-adrenoceptor (**b**), and α_{2C}-adrenoceptor (**c**). Residues were scored for binding energy ΔΔG, and those most strongly contributing to the binding of DA are shown in stick representation. The strength of binding interaction is colored by the depth of blue with dark blue being the most strongly contributing. Important hydrogen bonds to the amine group and catechol hydroxyls are formed in all binding poses

identity at all residues except position 6.55, in which the His has been replaced with a Tyr in both α_{2A}- and α_{2C}-adrenoceptors. Docking results of DA at either α_{2A} or α_{2C} also identified many of these same residues as critical for binding. Particularly, D3.32, V3.33, S5.42, and F6.51 were present in all receptor models contributing more than − 0.4 Rosetta energy units each to the binding of DA (Figure B.9). Residues S5.46 and Y6.55 were also present in all receptor binding modes though contributions varied depending on which receptor type. These results coupled with those from the biological assays provide a strong structural reasoning behind the activity of DA at the α_{2A}- or α_{2C}-adrenoceptors.

B.5 Discussion

Previous studies reported DA as a potential α_2-adrenoceptor ligand on the basis of radioligand binding experiments in transfected mammalian and insect cell lines [477, 478] and in bird and rat brains [479] and also from autoradiographic experiments in tissues [479]. Furthermore, DA has been reported to decrease cAMP intracellular levels in transfected mammalian cell lines but only throughout α_{2C}-adrenoceptors, not α_{2A}-adrenoceptors, and at concentrations much higher than NE (EC_{50} in the micromolar range) [479].

In our study, we show that α_{2A}- and α_{2C}-adrenoceptors can bind DA at concentrations in the same order than NE, suggesting that they could be activated by DA at in vivo concentrations. First, our results demonstrate that endogenous DA, and also common synthetic DA receptor ligands, binds to α_2-adrenoceptors with moderate to high affinity in the mammalian brain. Second, the affinity of DA for α_2-adrenoceptors is in the same range as for D_1-like and D_2-like receptors, suggesting that endogenous levels of DA

can activate both α_2-adrenoceptors and dopamine receptors. Third, DA and synthetic DA receptor ligands can activate G protein and induce cell DMR through α_2-adrenoceptors. Finally, DA and NE show the same cell signaling pattern, being both capable to modulate adenylyl cyclase activity and ERK1/2 phosphorylation at nanomolar concentrations.

The most conclusive demonstration that DA is an α_2-adrenoceptor ligand comes from the results obtained with binding and G protein activation BRET assays, where the affinities or potencies of DA for α_2-adrenoceptors were found to be very similar or even higher than for D_1-like and some subtypes of D_2-like receptors [480]. Particularly, the EC_{50} values of DA for α_{2A}- and α_{2C}-adrenoceptors (5–170 nM) were consistently lower across all Gi/o protein subtypes as compared with the EC_{50} values (130–400 nM) for the predominant striatal D_2-like receptor D_{2L} [480]. Taking into account that the levels of tonic extracellular DA are 20–30 nM (with peaks of 500 nM) [501], DA could reach sufficient extracellular concentration to activate α_{2A}- and α_{2C}-adrenoceptors in the striatum, irrespective of the maximal concentration of extracellular NE. In fact, striatal DA release sites are designed for transmitter spillover [502] and most striatal DA receptors are primarily extrasynaptic [503, 504], as well as striatal adrenoceptors, based on the mismatched low NE innervation [469-472]. Although the specific functional role of the DA-sensitive α_2-adrenoceptors in neuronal striatal function remains to be established, a previous study suggests that they might mediate an inhibitory modulatory role of the Gs/olf-coupled striatal adenosine A_{2A} and DA D_1 receptors [475].

The possibility of DA-mediated activation of α_{2A}- and α_{2C}-adrenoceptors in extrastriatal areas should not, however, be underestimated. Cortical α_{2A}-adrenoceptors are most probably able to be activated by DA, particularly in the prefrontal cortex, which

receives a rather dense DA innervation [505]. In fact, there is recent evidence for the localization of α_{2A}-adrenoceptors in the cortical terminals from mesencephalic DA neurons [506], which could play a role as "DA autoreceptors." But, there is also evidence for the localization of both α_{2A}- and α_{2C}-adrenoceptors in the soma and dendrites of the mesencephalic DA cells of both substantia nigra and ventral tegmental areas [506, 507]. Apart from the NE input, these α_{2A}- and α_{2C}-adrenoceptors should be able to act as DA autoreceptors that control the non-synaptic somatodendritic DA release [502]. Adding the present results to our recent study that also indicates a significant role of NE as a Gi/o-coupled D_2-like receptor agonist [480], we could state that Gi/o-coupled adrenoceptors and DA receptors should probably be considered as members of one "functional" family of catecholamine receptors. A general consideration from the DA and D_2-like receptor ligand sensitivity of cortical α_{2A}-adrenoceptors is that it should also be involved in the cognitive-enhancing effects associated with their activation, with possible implications for attention deficit hyperactivity disorder [508].

Molecular modeling of DA binding to the various receptors provides a likely binding hypothesis for the results obtained in the biological assays. Of note is the striking similarity between the ligand binding pocket of the D_3 receptor and that of α_{2A}- and α_{2C}-adrenoceptors. Many of the residues that line the binding pocket are identical or chemically well conserved. Given this similarity, it was perhaps unsurprising that the docking of DA at α_{2A} and α_{2C} was nearly identical to DA binding to the D_3 receptor. The lower potency of DA at α_{2A}- and α_{2C}-adrenoceptors compared to NE seems to depend on a lower number of strong interactions as compared to those between DA and D_3 receptors. The pocket may have evolved to bind the slightly bulkier NE and, therefore, is not of an ideal size for DA.

However, the differences may also be due to the lower resolution of binding predictions for a comparative model as opposed to a crystal structure. Despite this, the structural model strongly mimics the results of the binding and activation experiments and therefore provides further evidence of DA acting as a ligand at these receptors.

Another major finding of the present study is that α_{2A}- and α_{2C}-adrenoceptors are also common targets for compounds previously characterized as D_2-like receptor ligands. Particularly striking was the ability of prototypical D_3 and D_4 receptor agonists 7-OH-PIPAT and RO-105824 to bind with high affinity to α_{2A}- and α_{2C}-adrenoceptors, which might call for revisiting results of previous studies using these compounds. Furthermore, these two compounds and the other DA-synthetic ligands assayed, as well as NE, were able to activate ERK1/2 phosphorylation by binding to α_2-adrenoceptors. The final pharmacological profile of RO-105824 was that of a potent biased agonist for α_{2A}-adrenoceptor with functional selectivity for a G protein-independent signaling. On the other hand, based on BRET experiments, both potency and efficacy dependence on the receptor and the Gαi/o protein subtype were the norm for all ligands, including the endogenous neurotransmitters. We already described that NE and DA show different receptor- and Gαi/o subtype-dependent potencies of D_2-like receptor-mediated G protein activation [480]. The present results extend these findings to other receptors and to non-endogenous ligands, as well as to differences in efficacy. Even though G proteins of the Gαs-Gαolf family do show contrasting brain expression pattern [509], to our knowledge, no clear region-specific pattern of mRNA expression for Gαi/o protein subtypes has been reported. Detailed characterization of the expression patterns for Gαi/o protein subtypes would then be central to determine their role in α_{2A}- and α_{2C}-adrenoceptor activation and

thus their possible specific targeting with Gαi/o subtype functionally selective compounds.

In conclusion, DA is a potent and efficacious ligand at α_2-adrenoceptors, which modulates forskolin-induced adenylyl cyclase activity and ERK1/2 phosphorylation. The concentration required for these effects is in the range of that for activating D_2-like and D_1-like receptors, indicating that these receptors are members of one functional family of catecholamine receptors. Our results provide a clear answer to the mismatch between the low striatal NE innervation and the high density of striatal α_2-adrenoceptors, which behave as functional DA receptors.

B.6 Conclusions

The results here show that traditional classification of adrenoceptors versus dopamine receptors is blurred. Rather, we argue that these families should be largely classified as aminergic receptors. This is explained in similarity of ligands, ability for disparate ligands to cross-talk between receptor subtypes, and, importantly, high structural conservation of the ligand binding pocket. As more receptors have their ligand binding pockets structurally characterized it may be possible to identify similar cross-talk between ligands at different receptors. This has significant implications in the ability to generate ligands with high selectivity at a specific target receptor with low off-target efficacy.

APPENDIX C

Modeling the Complete Chemokine-Receptor Interaction

C.1 Summary

Appendix C is related to Chapter 5 and extends the general protocol of GPCR homology modeling to chemokine receptor complexes with chemokine ligands. This is a protocols paper that is based on the method described in Chapter 5. I worked with the first author to modify my original protocol specifically for chemokine/chemokine receptor modeling but it serves as a general protocol for modeling of protein-protein interactions of GPCRs. This appendix is from the article "Modeling the Complete Chemokine/Chemokine Receptor Interaction" for which I am a contributing author [510]. Reprinted by permission from Springer Nature Customer Service Center GmbH: Springer Nature, Modeling the Complete Chemokine-Receptor Interaction by Michael J. Wedemeyer, Benjamin K. Mueller, Brian J. Bender et al, 2019 (2019).

C.2 Introduction

C.2.1 Overview

Recent structural breakthroughs have allowed for the crystallization of dozens of G protein-coupled receptors (GPCRs), but experimental structures of GPCR-protein complexes remain relatively sparse. Homology modeling has become a key tool for understanding GPCR complexes [44, 511, 512]. While significant strides have been made to create models of GPCRs bound to small molecules, generalizable protocols for larger

protein ligands, such as chemokines, remains a challenge. Previous publications have made excellent models of chemokine-receptor structures, of both complete complexes and partial complexes [513-515]. These models have provided insight into mechanisms of chemokine receptor activation. In most cases, these models are highly specialized and require structure-function data or structural knowledge as restraints [516]. In this manuscript, we present a general protocol for the production of full-length chemokine-GPCR complexes in the absence of experimental data.

C.2.2 Structure-based Modeling of Chemokine-GPCR Complexes

Cell migration is indispensable for nearly all essential physiologic processes, from development to immune surveillance and beyond [517, 518]. *Chemotaxis* entails the regulation of cell migration via cellular responses to secreted chemical mediators. In the human body, the most studied and biologically relevant mediator of chemotaxis is a family of small, soluble proteins called chemokines (i.e., chemotactic cytokines). Chemokines are secreted into the extracellular space, forming concentration gradients along which cells migrate. Chemokines elicit this concentration-dependent migration by binding cell-surface GPCRs, which initiates signaling pathways that provoke chemotaxis [519]. Through this process, chemokine-GPCR signaling aids in pathogen clearance, tissue repair, adaptive immune responses, and other processes through receptor specific expression on leukocytes, neutrophils, macrophages, and dendritic cells [520].

Chemokines have a highly conserved tertiary structure despite their relatively low sequence homology [521]. This is mainly attributed to the presence of four conserved cysteines that form two intramolecular disulfide bonds. The presence and location of the first two cysteines separate chemokines into one of four families: CC, CXC, CX_3C, or XC.

This familial differentiation has structural impacts, most clearly seen in the positioning of the N-terminus in CC vs CXC chemokines [37]. A large majority of chemokine-CKR pairs are not interfamily, implying that these small changes in structure have a large biological influence. Moreover, the dimerization patterns are unique among families and likely dictate how chemokine dimers can interact with their binding partners [522].

CKRs are a subset of the GPCR superfamily and as such share many common features. CKRs reside in the plasma membrane where they can bind extracellular ligands and transduce this chemical information into downstream signaling pathways. The unique intracellular pathways activated by CKRs can affect adhesion, differentiation, and proliferation, but are most well-known for stimulating a migratory phenotype. This is done mainly through interactions with the heterotrimeric G protein and β-arrestin. Unlike most class A GPCRs that bind small molecules, CKRs bind much larger protein ligands. Decades of mutational structure-function studies have developed a two-step, two-site model of chemokine-CKR interaction [523, 524]. In this paradigm, the N-terminus of the CKR first binds the globular core of the chemokine. This results in a bound chemokine-CKR complex but does not result in G-protein activation or β-arrestin recruitment. In the second step, the N-terminus of the chemokine is inserted into the orthosteric pocket of the CKR. This triggers a change in conformation of the CKR that is transferred through a series of structural toggle switches in the transmembrane domain that eventually results in activation of intracellular signaling carriers (Figure C.1).

Obtaining structural information on GPCRs has been challenging, in part due to the size and dynamics of the system, which interfere with standard NMR or crystallography methods. Only within the past decade was the first high-resolution, CKR structure solved [25]. Currently there are approximately 50 unique GPCR structures. While these structures

make up a small percentage of the estimated 826 members of the GPCR superfamily, they opportunely contain 6 unique CKRs: CCR2, CCR5, CCR9, CXCR1, CXCR4, and the viral CKR US28 [25, 34-36, 525]. These data give insight into the molecular toggles imbedded in the CKR necessary for activation. Furthermore, three of these proteins were solved with chemokines bound in the orthosteric pocket: CCR5 in complex with 5P7-CCL5, CXCR4 in complex with vMIP-II, and US28 in complex with CX_3CL1 (Figure C.2) [37, 46, 47]. Undoubtedly, these structures have propelled the field forward [526, 527], but it is important to note that the full, native chemokine-CKR interaction has not been captured.

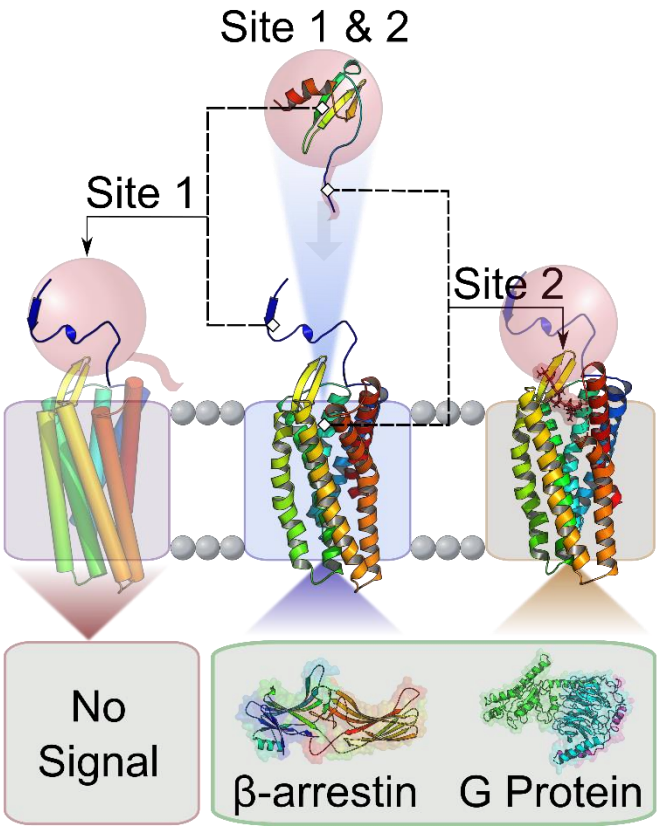


Figure C.1 Chemokine binding and receptor activation is commonly described as a two-step, two-site process. The first step in this model is the binding of the CKR N-terminus to the core of the chemokine (site 1). This interaction contributes a large amount of the binding energy but is not sufficient for signaling. The N-terminus of the chemokine then docks into a pocket within the transmembrane core of the CKR (the orthosteric site, or site 2). This leads to conformational changes in the CKR and activates internal signaling pathways such as G protein and β-arrestin. While both site 1 and 2 are necessary for potent, selective chemokine function *in vivo*, binding of a peptide ligand at site 2 can be sufficient to induce full CKR activation.

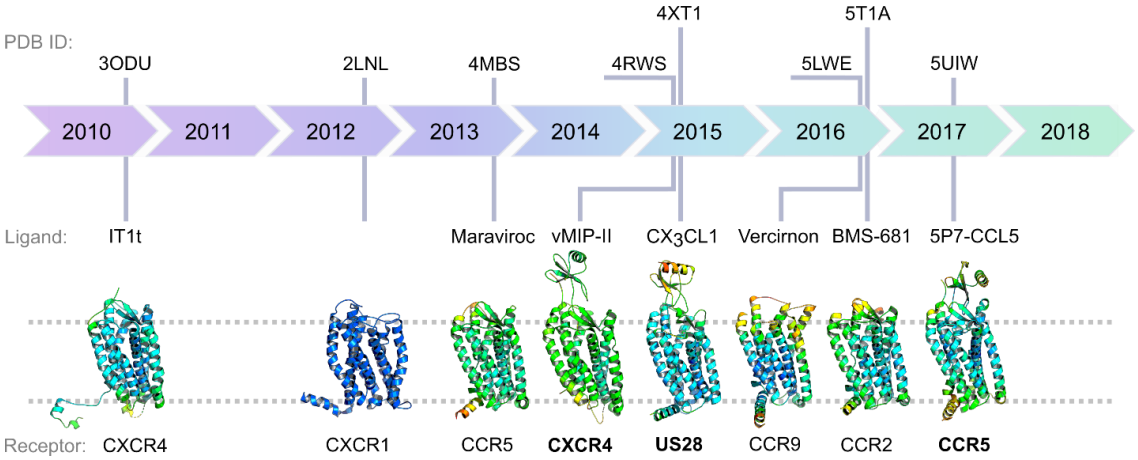


Figure C.2 The first chemokine receptor, CXCR4, was crystallized in 2010. Currently there are a total of 13 CKR structures published, with the eight non-redundant structures shown above. Each structure is colored according to B-factor with the exception of 2LNL, an NMR structure. Of note, there are three co-crystal complexes of chemokines bound to CKRs shown in bold.

The N-terminus of the CKR, the main site 1 interface, lacks electron density in all three co-crystal structures and is not resolved past a few residues. Additionally, mutations in the chemokine or CKR have been made to stabilize the complex. Analyzing these structures has shed light on the chemokine-CKR interaction and developed our knowledge beyond the two-site model [528]. These structures show that the chemokine-CKR interface is not restricted to two small peptide interactions; rather the interface is a large, continuous protein-protein interaction surface approximately 3300 A^2 [515]. This adds nuance to the two-site model by identifying both the distal and proximal extremes of the CKR N-terminus to significantly contribute to the binding interface, referred to as site 0.5 and site 1.5 respectively [529]. Analyzing these new interaction sites explains in part how different chemokine dimer geometries may hamper or prevent binding [530].

To better understand this new multi-site model of chemokine-CKR interaction, solving additional chemokine-CKR complexes is vital. This is a monumental task as including a chemokine ligand greatly complicates the crystallization process. Additionally, because of the promiscuity of the chemokine-CKR interaction network, it is estimated that

the three unique complex structures known represent less than 5% of the entire chemokine-CKR network [531]. The chemokine-GPCR system demonstrates the need for generalizable methods that can predict the structure of bound complexes in the absence of experimental data beyond homologous structures. To extend our structural understanding of this complex system to the remaining 95% of relevant co-complexes, we can use homology modeling to create 3D representations of unsolved chemokine-CKR structures to generate hypotheses and guide rational design.

C.2.3 Modeling Strategy

This paper outlines a specialized protocol to develop a structural model of a chemokine bound to its receptor using the Rosetta software suite for macromolecular modeling [223]. The protocol was designed to use all available structural information and is heavily inspired by Rosetta's comparative modeling protocol [5], which consistently performs well in the critical assessment of protein structure prediction [532]. Ultimately Rosetta was chosen for its powerful ability to model flexibility via sampling of peptide fragments within the protein data bank (PDB).

The following protocol is designed to model any desired chemokine-CKR pair from a series of homologous templates. The limitations inherent to this style of modeling are mainly derived from the available template structures and the quality of the sequence alignment. As such, as more structures become available, the accuracy of the models will improve. In an attempt to minimize error, multiple templates are used, and alignments are based on solved structures.

In overview, this protocol will sample short structural segments stochastically chosen between templates of chemokine, CKR, and complexed crystal structures [188].

NMR structures of chemokines bound to the N-terminal peptide of a CKR are also included to fill in missing information of the site 1 interaction. These templates will be stripped of their side chain identities, and new side chains will be modeled onto the peptide backbone corresponding to the target sequence in a process called threading. This library of threaded models will compose a large part of the chemical sampling space, while a second library of short peptide fragments mined from the PDB will allow sampling away from the template structures. Models of the target protein complex, called decoys, will be constructed by selecting components from each of these two libraries based on weighted probabilities. The system will evaluate each change and either accept or reject it based on the calculated energy of that change. This iterative process will guide decoys down the energy landscape while allowing them to escape shallow local minima. At the end of this process the decoy will be relaxed into its nearest free energy minimum. The entire process will be repeated to create multiple decoys that can then be analyzed and interpreted. One or more decoys will be chosen as the final model to be used as a guide for further investigations (Figure C.3). While Rosetta can be used as a black box, this protocol strives to explain each step. To begin, Rosetta must be installed onto your system. Rosetta is designed to run on clusters in which multiple cores can be used to parallelize model generation, and at minimum a lab-scale computer with 10-30 cores should be used(see https://www.rosettacommons.org/docs/latest/getting_started/Rosetta-on-different-scales for more information).

C.3 Materials

Homology modeling is the construction of an atomic model of a *target* protein based solely on the target's amino acid sequence and the experimentally determined

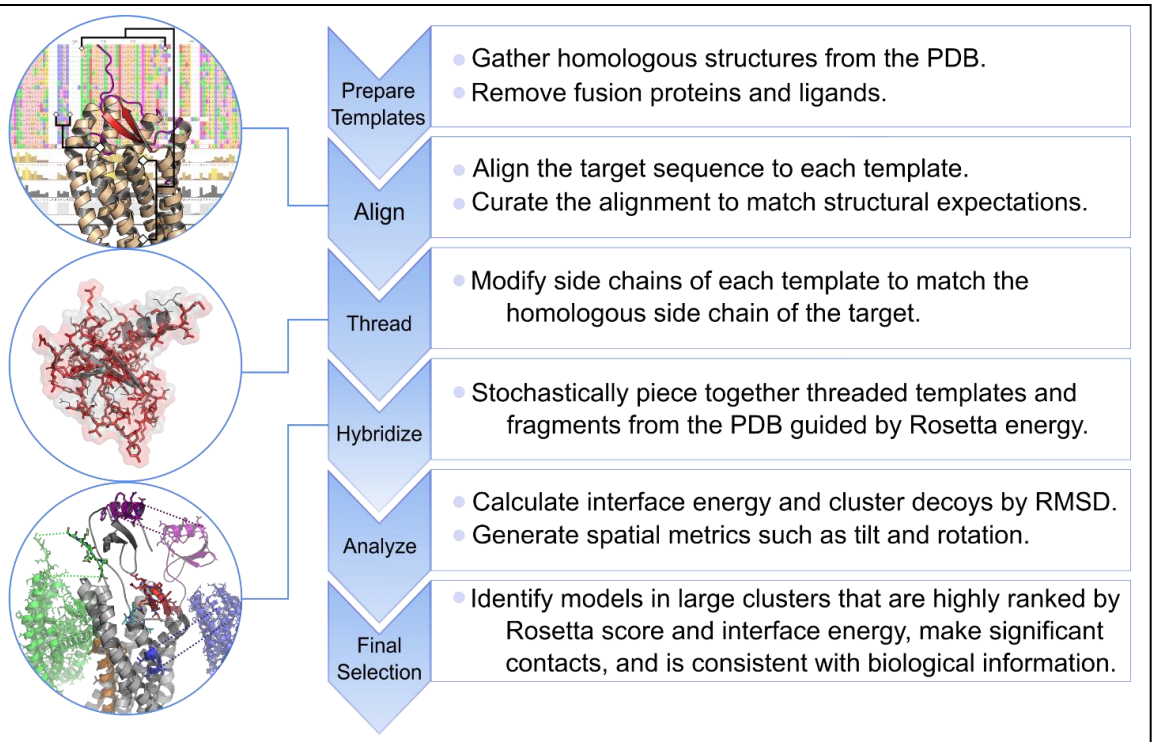


Figure C.3 This figure summarizes the steps detailed in this protocol to generate homology models of chemokine-CKR complexes. The central steps of align, thread, and hybridize are depicted to the left. In short, the method uses a curated alignment to graft a target's homologous side chains onto published structures in a process called threading. These threaded models are then used as a library to construct decoys in the hybridize step. Analysis of the decoys will guide final model selection.

structures of homologous proteins, referred to as *templates*. Traditionally, homology models are built for one target using one or more templates. This protocol modifies this approach by modeling two protein targets simultaneously, a chemokine and its CKR. This is accomplished using a variety of templates, particularly chemokine-CKR complexes.

When choosing templates, it is important to consider the value of the information you pass on to Rosetta. Selecting a variety of templates to increase sampling diversity is recommended, but templates must be structurally similar to one another enough to warrant inclusion. The signaling state of the GPCR is also an important factor for modeling active or inactive structures.

In this tutorial, we will be modeling the recent structure of 5P7-CCL5 bound to CCR5 (PDB ID: 5UIW) while only using the structure for comparative purposes(Zheng, 2017). The templates used for this tutorial are 4 unique CKR crystals (PDB ID: 3ODU, 4MBS, 5LWE, and 5T1A), the native chemokine structure (PDB ID: 5COY), 4 NMR structures showing site 1 interactions (PDB ID: 2N55, 2K05, 2MPM, and 1ILP), and 2 chemokine-CKR complexes (PDB ID: 4RWS and 4XT1) (Figure C.4). Depending on an investigator's biological question, it may be beneficial to reduce or expand these inputs or

PDB: ID	CHEMOKINE	RECEPTOR	METHOD	SEQUENCE IDENTITY TO CCL5/CCR5	RESOLUTION /CALCULATED CONFORMERS	ENGINEERED MUTATIONS
3ODU	-	CXCR4	X-ray	29.12	2.5 Å	3
4MBS	-	CCR5	X-ray	100	2.71 Å	4
5LWE	-	CCR9	X-ray	32.62	2.8 Å	0
5T1A	-	CCR2	X-ray	63.06	2.8 Å	18
5COY	CCL5	-	X-ray	100	1.44 Å	0
4XT1	CX3CL1	US28	X-ray	20.78 / 25.82	2.886 Å	1
4RWS	vMIP-II	CXCR4	X-ray	33.80 / 29.12	3.1 Å	6
2MPM	CCL11	CCR3 Peptide	Solution NMR	33.33	64	0
1ILP	CXCL8	CXCR1 Peptide	Solution NMR	18.18	40	0
2K05	CXCL12	Sulfated CXCR4 Peptide	Solution NMR	28.57	100	3
2N55	CXCL12	CXCR4 Peptide	Solution NMR	28.57	100	3

Figure C.4 The input structures for this tutorial are summarized. In total there are 7 crystal structures, including two co-crystals, and 4 NMR structures solved in the solution state. The quality of each structure is shown as resolution in angstroms or number of conformers generated for X-ray and NMR structures, respectively. The number of engineered mutations is reported, but it should be noted that many of these structures contain additional non-native components (*i.e*. fusion proteins or nanobodies).

include structures of other homologous GPCRs. A GitHub repository has been created for easy access to all of the materials required for this tutorial (available at: https://github.com/MichaelWedemeyer/Chemokine-GPCR-Complexes).

Bold font represents code to be typed, unless noted this is all on one line

"Quotation marks signify exact text or file names"

Underlined font are web resources

Software used:

- Rosetta 3.8 installed on a lab grade computer (10+ cores)
  - https://www.rosettacommons.org/demos/latest/tutorials/install_build/install_build
- Calibur
  - https://sourceforge.net/projects/calibur/
- Chimera 1.12 (optional)
  - https://www.cgl.ucsf.edu/chimera/download.html
- PyMOL
  - https://pymol.org/2/
- A text editor used for code such as Notepad ++ or Sublime
  - https://notepad-plus-plus.org/
- Python
  - https://www.python.org/
- Perl
  - https://www.perl.org/

C.3.1 Getting started

1. Obtain a license to install Rosetta from https://www.rosettacommons.org/software/license-and-download. A non-commercial license is free to academic and government laboratories. This protocol was developed for Rosetta 3.8.
2. Follow the instructions at https://www.rosettacommons.org/demos/latest/tutorials/install_build/install_build to install Rosetta.

C.3.2 Target Sequence

1. A chemokine-CKR pair, referred to as the target, must be chosen for modeling. The sequences will need to be gathered and edited as desired. Go to https://www.uniprot.org/ and search for your target CKR.
2. Scroll to the bottom third of the web page and download the fasta file of the fully processed form. Depending on your biological question, you may want to truncate the N or C terminal regions of the CKR to reduce the computational burden. Long chemokine tails may also be excluded if they are not predicted to interact with the CKR. This protocol was tested with the full CKR N-terminus with a truncated the C-terminus to match crystallographic structures. The C-terminal residues AKRFCKCCSIFQQEAPERASSVYTRSTGEQEISVGL were removed from CCR5. The sequence of 5P7-CCL5 was used for the chemokine.
3. Open both fasta files in a text editor such as Notepad ++ or Sublime. The fasta file format begins with a ">" followed by the name and information of the sequence. The following line has the sequence for the protein using the single letter amino acid code.
4. Copy the chemokine sequence and paste it after the CKR sequence with a forward slash "/" separating the two sequences. This will inform Rosetta that these are two separate chains.
5. Replace the first line with ">CCR5_CCL5" to reflect this change.
6. Save this as a new file with the name "CCR5_CCL5.fasta".

C.3.3 Generate Fragments

Rosetta uses a fragment library to guide the search of conformational space. A *fragment* is a stretch of amino acids from an experimentally determined structure that corresponds to amino acids in your protein's target sequence that Rosetta uses to guide the search of conformational space. This is important for *de novo* modeling of large flexible loops or segments with no template information. To generate this library, the robetta server will use PSI-BLAST first predict the secondary structure of your target and then find similar 3 and 9 amino acid fragments. The coordinates of these small pieces of experimental structures are saved to be used during stochastic sampling.

1. For academic users: Go to http://robetta.bakerlab.org/
2. Sign up for an account.
3. Submit a fragment request by pasting in the fasta file you generated in step 3.2.7 into the text box, this process may take a while

4. Once completed, save the fragment files “att000_03_05.200_v1_3” and “att000_03_05.200_v1_3” as “CCR5_CCL5_3.frags” or “CCR5_CCL5_9.frags”, respectively.
5. For non-academic users: Follow the instructions at https://www.rosettacommons.org/docs/latest/rosetta_basics/file_types/fragment-file

C.3.4 Templates

Templates are used by Rosetta to model homologous segments of the target protein. Starting with homologous structural components greatly increases the likelihood of generating near-native decoys and choosing the proper templates to include is vital to the success of the method. First, look up what structural information is available. Websites such as http://gpcrdb.org/ [342] or https://zhanglab.ccmb.med.umich.edu/GPCR-EXP/ are useful for curating GPCR structural information and determining the percent similarity to your target of interest. Many chemokine structures are deposited in the PDB along with several NMR structures of CKR N-terminal peptides complexed to chemokines. The PDB can quickly be searched for homologous templates using BLASTp at https://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE=Proteins.

1. Go to https://www.rcsb.org/ and download the PDB format of your desired templates. Structures used for this tutorial are: 4RWS, 5COY, 2N55, 2K05, 1ILP, 2MPM, 5T1A, 5LWE, 4MBS, 3ODU, and 4XT1.
2. Before using these structures as templates crystallization artifacts must be removed. Many GPCRs are crystalized as a fusion protein or with nanobodies bound. These must be removed manually. In addition, flexible portions of NMR structures, such as the non-interacting portion of CKR peptides may be removed to avoid bias. Open the template PDB files in PyMOL.
3. Select the non-native components you wish to remove. Type **set seq_view, 1** in the PyMOL text box or click the button labeled “S” in the lower right-hand corner of the viewer window. The sequences are now displayed above the structures. Highlight the residues you wish to remove. Note: A drastic change in the PDB numbers usually denotes the location of the inserted protein. In the case of 4RWS, the residues to remove are 1002-1164.
4. With the residues selected type **remove sele** in the command line or right-click the residues and select “remove” to delete them.

5. Align all templates to one of the chemokine–CKR structures using the "Action (A) ➔ Align ➔ all to this" command in PyMOL. Adjust the alignment as needed so the chemokines are aligned together and in approximate binding positions with the CKRs. For example, it may be necessary to align receptor templates to just chain A of 4RWS.
6. Save the edited structure with the PDB ID and a suffix of "_Isolated.pdb".
7. Clean the PDB using the clean_pdb python script in Rosetta seen below. In addition, this program also creates a fasta file of the cleaned sequence. The first input argument is the name of the PDB file to be cleaned, followed by the chain(s) to be extracted. Copy the code below and replace the "(/PATH/TO/ROSETTA/DIRECTORY)" with the correct path for your system.

python
(/PATH/TO/ROSETTA/DIRECTORY)/tools/protein_tools/scripts/clean_pdb.py 4RWS_Isolated.pdb AC

8. Repeat this process for each input template.

C.3.5 Alignment

Rosetta uses a sequence alignment to guide the threading of the target sequence onto homologous regions of the template. The target amino acids are placed onto the coordinates of the template according to the alignment of the template sequence to the target sequence. Any residues that are not aligned between the two will be discarded in this step. Due to the impact the alignment has, it is often necessary to adjust the alignment manually. This protocol focuses on chemokines and CKRs that have conserved topology, and the accuracy of the modeling can be improved by using this structural consistency. This translates into removing gaps in highly conserved secondary structures, such as the chemokine β-strands or the α-helices of the CKR, that are created by sequence-based alignment algorithms. Gaps in the alignment force Rosetta to either insert target residues or skip template residues, which may result in disruption of the secondary structures or a register shift. This will worsen the overall input and make future steps more difficult. Therefore, it is important to keep similar structures aligned.

1. Go to https://www.ebi.ac.uk/Tools/msa/clustalo/ or your preferred sequence alignment program.
2. Open the fasta files for all templates and the target. Copy and paste them into the Clustal window.
3. Keep default settings and press submit.
4. Download the alignment file once the tool has completed.
5. Open the file in a text editor such as notepad++ or a multi sequence editor such as Jalview.
6. Edit the sequence alignment to match with the structure. To get an example of structural alignments align your CKR structures using http://gpcrdb.org/. The chemokine should always come after the receptor in each of these alignments. Gaps should be included between CKR loops and transmembrane regions when required. Highly conserved residues, such as cysteines should be aligned when possible.
7. Manually convert the alignment into the Grishin format as seen in the provided examples. Each template requires its own alignment file containing only the target and template sequence. This will create as many Grishin files as you have templates. The Grishin format is as follows:

```
## Target Template.pdb
#
Scores from program: 0
0 Target_sequence
0 Template_sequence
```

Where the Target is the name in the target fasta file and the Template.pdb is the template file name. The sequences must be on continuous lines with no breaks other than "-".

C.3.6 Define Topology

Rosetta must be given information on membrane protein *topology* (*i.e.* how the complex sits in the membrane). This information can be predicted based on the sequence of amino acids or on experimental information. Rosetta does not use an explicit membrane as other programs do, instead an implicit model is used. A membrane prediction file informs Rosetta on which residues are within the membrane environment.

1. Go to http://octopus.cbr.su.se/ or your preferred topology prediction program.
2. Copy and paste the fasta generated in step 3.2.7 into the text box. In the text box remove the "/" signifying the chain break.
3. Press "Submit Octopus"
4. Once completed, right click the octopus topology file to save as a "CCR5_CCL5.octopus". Note: Inspect this file to ensure that there are seven transmembrane domains as expected. If this is not the case, manually calculate membrane spanning residues using an alignment.
5. Convert to span file using Perl script.

perl (/PATH/TO/ROSETTA/DIRECTORY)tools/membrane_tools/octopus2span.pl CCR5_CCL5.octopus > CCR5_CCL5.span

C.4 Methods

Now that our files have been generated and gathered together, we can begin to use Rosetta to make decoy structures. First, we will combine the homology information with the PDB templates and convert this into a form Rosetta can efficiently use. The workhorse of this method, the hybridize step, will then sample throughout the conformational space we have generated and build conglomerated models from the assortment of inputs. These decoys will then be analyzed and evaluated.

C.4.1 Threading

Each residue in the templates will be converted to its corresponding residue in the target fasta based on the structural alignment. The atomic coordinates of the main chain will be conserved, but the side chains will be converted to the target residue.

1. IMPORTANT: If you did not align the template structures in 3-dimensional space in step 5 of Section 2.4, do this now. If you don't place the templates in one geometric context, the next step will generate threaded templates that are also not aligned. This will result in faulty behavior in the subsequent Hybridize step (Section 3.2).
2. Make sure all of the Grishin alignments, template PDBs, and the target fasta are in one folder.
3. Run the partial_thread Rosetta application.

(/PATH/TO/ROSETTA/DIRECTORY)/main/source/bin/partial_thread.default.linux gccrelease
-in:file:fasta CCR5_CCL5.fasta -in:file:alignment CCR5_CCL5_4RWS.grishin
-in:file:template_pdb 4RWS_Isolated_AC.pdb

4. This command should yield a new PDB named 4RWS_Isolated_AC.pdb.pdb. Rename this file as appropriate, such as "CCR5_CCL5_on_4RWS.pdb".
5. Repeat step 3 and 4 for each template.

C.4.2 Hybridize

This step will partition the threaded structures generated in Section 3.1 into discrete segments. These building blocks will then be combined in a stochastic manner over the first input template. This allows Rosetta to sample multiple conformational states and select for the most energetically favorable. To model segments with no templates or long loops, the fragments generated in Section 2.3 will be randomly inserted similar for *de novo* building of these regions. The resulting decoy will be relaxed by Rosetta into the nearest low-energy conformation and a single PDB will be generated. This will be repeated 5000 times to generate 5000 decoys. After final analysis in step 5, additional rounds of hybridization may be required to sample more conformations or to improve on promising

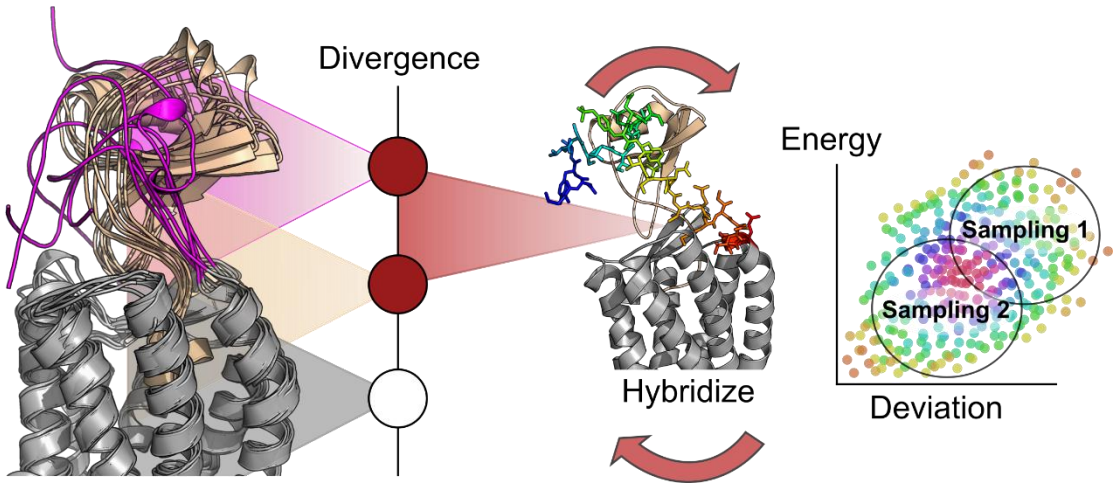


Figure C.5 Areas with low homologous structural information, such as the CKR N-terminus, likely are very divergent among decoys. Sections of promising decoys can be selected for refinement by using the PDB as an input in a new hybridize step. Iterative sampling of promising decoys is likely lead to higher quality models.

decoys. For the latter, these decoys should be used as input by modifying the XML script to include them (Figure C.5).

1. Gather the following input files into one folder:
    - 3mer and 9mer fragment files generated in Section 2.3
    - Threaded models generated in Section 3.1
    - Target fasta file generated in Section 2.2
    - Span file generated in Section 2.6
    - The xml file, options file, and weights files found in the supplied folder
2. Define any post-translational modifications to be used by editing the fasta file. In this protocol, the Tyrosines in the N-terminus of the CKR will all be sulfated. Tyrosine sulfation plays a critical role for some chemokine-CKR interactions [533]. In the fasta file, add "[TYR:sulfated]" Immediately after the reside to be altered.
3. Save this file as "CCR5_CCL5_sulfated. fasta"
4. Define expected disulfide bonds in a disulfide file. Open a blank text file using a text editor.
5. Type the two residue numbers that create a disulfide with a space separating. The numbering of residues is according to the supplied fasta sequence, i.e. the residues do not start back 1 for each new chain but continue in sequence.
6. Using a new line each time, repeat this for each disulfide predicted.
7. Save this file as "CCR5_CCL5.disulfide".
8. Modify the XML script to match your specific inputs if they differ from this example.
9. Open the options file in a text editor and modify lines 6-18 under the "Options to be changed" section
10. Run the Rosetta scripts xml file to generate and relax decoys.

(/PATH/TO/ROSETTA/DIRECTORY)/main/source/bin/rosetta_scripts.default.linuxgccrelease @CCR5_CCL5.options > CCR5_CCL5.log

After this step, the output directory should be populated with a score file named "score_CCR5_CCL5.out" and a silent file named "CCR5_CCL5.silent" containing all of the decoys generated. A silent file is a binary version of a PDB file which saves storage space and can concatenate all of the decoys into a single file. The silent file can only be read by Rosetta and decoys will need to be converted to PDB files prior to manipulation or visualization by other programs. Experimentally validated contacts can be used to guide this step if available. For more information, visit https://www.rosettacommons.org/docs/latests/rosetta_basics/file_types/constraint-file.

C.5 Data Analysis and Interpretation

The ultimate goal with protein modeling is to predict the native structure. In the presence of an experimental structure, a model can be validated by simple comparisons such as RMSD. More likely, modeling will be used in the absence of an experimentally determined structure. In this case, identifying high quality decoys with confidence that these decoys are structurally relevant requires additional analysis.

Provided that Rosetta has sampled enough of the chemical search space, near-native models exist in the set of decoys. The analysis of large production runs focuses on the top ~10% of decoys by total energy, as these have the highest chance of being structurally relevant. The following steps will cluster decoys by RMSD, analyze the protein-protein contacts, and use Rosetta scoring to guide the selection of a final model. Ultimately, a good structure will be in a large cluster, highly ranked by total score and interface energy, make significant contacts, and be consistent with biological information.

C.5.1 Calculating RMSD

This step will compare each decoy to the lowest scoring decoy via root-mean-square deviation (RMSD) of Cα positions. This gives information on the different population states sampled and the credibility of the score function to identify correct models.

1. Sort the score file generated in Section 3.2 by total Rosetta score.

cat score_CCR5_CCL5.out| awk 'NR<3(print $0;next)(print | "sort -nk2")' > scores_sorted.out

2. Identify the decoy with the lowest total score to be used as the ‘native’ state for RMSD comparison.

3. Extract the ‘native’ decoy from the silent file by replacing the “S_XXXX” with the correct decoy tag.

(/PATH/TO/ROSETTA/DIRECTORY)/main/source/bin/extract_pdbs.default.linuxgccrelease

-in:file:silent CCR5_CCL5.silent -in:file:tags S_XXXX

4. Calculate pairwise RMSD using the following command replacing the "S_XXXX.pdb" in the last line with the lowest scoring deccoy. The RMSD for the full sequence will appear in the score file under the column labeled "rms_rmsdto_000".

(/PATH/TO/ROSETTA/DIRECTORY)/main/source/bin/score_jd2.default.linuxgccrelease

-in:file:silent CCR5_CCL5.silent -in:file:silent_struct_type binary -in:file:fullatom

-in:file:spanfile CCR5_CCL5.span -score:weights membrane_highres_Menv_smooth

-score:set_weights cart_bonded 0.5 -score:set_weights pro_close 0

-out:file:scorefile RMSD_rescore.out -evaluation:rmsd S_0001.pdb _rmsdto_0000 FULL

5. Plot the top 10% of decoys by total score vs the RMSD to the lowest scoring decoy (Figure C.6).

To gauge convergence, construct a scatter plot with a Y-axis of total score in Rosetta Energy Units and an X-axis of the RMSD to the native structure. The lowest scoring decoy is used as a surrogate to the native structure if it is not available. This is a typical plot used in Rosetta, particularly *ab initio* folding, to visualize a folding funnel. Ideally, large RMSDs correlate with poorer scores and the scores quickly improve as the RMSD approaches the "native" state. Many times, this ideal situation is not observed and only a loose correlation can be drawn about the likelihood of converging on the native state. If this plot looks strikingly different, then likely Rosetta has not reached a consensus and more decoys should be generated by repeating Section 3.2.

C.5.2 Decoy Clustering

The more often a structural conformation is found by Rosetta, the more likely that the conformation actually represents a true native state. This is the logic behind clustering decoys based on their Cα RMSD. Calibur is a common tool used to cluster a large number

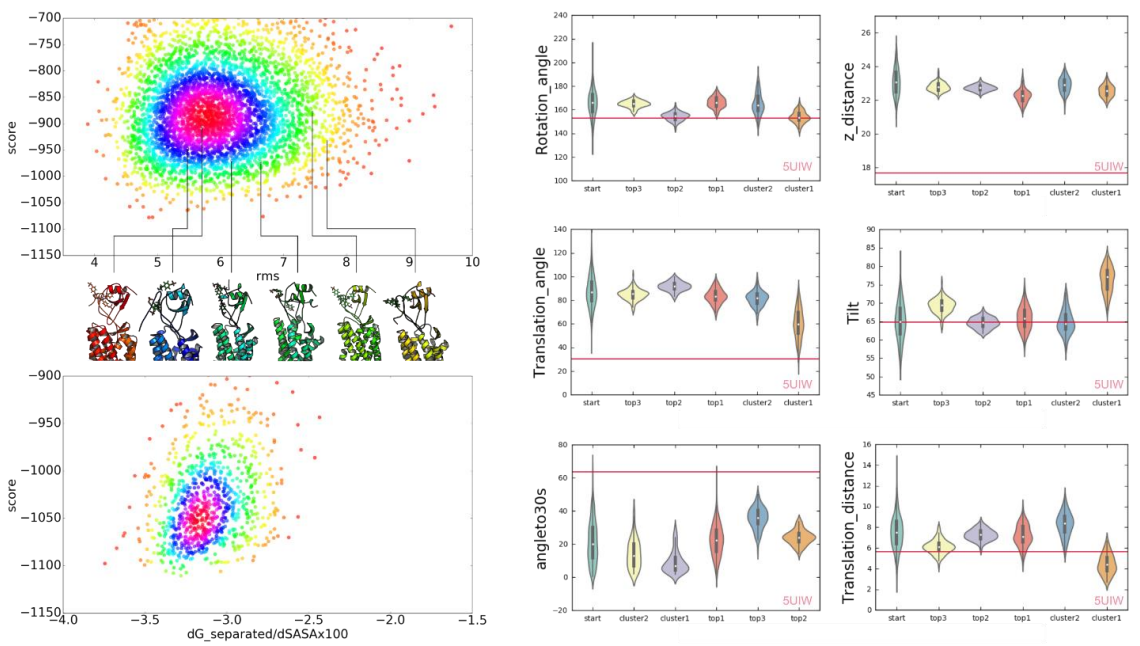


Figure C.6 Rosetta metrics and coordinate based metrics are used to select final models. Rosetta total score should be plotted against the RMSD to the native structure, or the lowest scoring decoy if the native structure is unknown. The total score should also be plotted against the calculated ΔG of the interface divided by the surface area of interaction (dG_separated/dSASAx100). Each dot in these plots represents a decoy as depicted, and there should be a positive correlation between score and RMSD as well as score and interface energy. The coordinate-based metrics show angles and distances that represent the chemokine-CKR interaction. The initial round of this protocol is shown under the label "Start" and a second round was completed using one of the top two clusters or top three scoring models as input. The red line corresponds to the distances and angles of the co-crystal structure 5UIW. Beyond this tutorial, this should also be compared to the other co-crystal complexes.

of decoy structures. This will be used to find the best scoring models in the three largest clusters as well as a representative model for each cluster. We want Calibur to select representative structures using Rosetta energy, so we will first sort the input according to this criterion. In terms of the energy landscape, Rosetta total score represents the depth of the energy well. A low score translates to an energetic minimum and therefore a candidate native conformation. The large clusters represent the wide energy wells that correspond to higher entropy.

1. Generate a list of PDBs to cluster that is ordered by the total score using the command below.

cat scores_sorted.out | awk 'NR>2(print $NF)' > pdb_list.txt

2. Extract these PDB files from the silent file. Note: this will generate 5000 PDB files and may reduce computer performance. Moving these to a new folder after this step is recommended.

(/PATH/TO/ROSETTA/DIRECTORY)/main/source/bin/extract_pdbs.default.linuxgccrelease
-in:file:silent CCR5_CCL5.silent -in:file:tagfile pdb_list.txt > extract.log

3. Add the file extension ".pdb" after each file name in "pdb_list.txt".
4. Calibur will output the results of the analysis to the "Cluster_calibur.out" file. At the bottom of the file will be a list of the largest three clusters. The first decoy is the center of the cluster and can be used as a representative of the general folding pattern this cluster sampled. After this, the number of decoys in the cluster is shown followed by every member of the cluster. This will be organized based on the input file. In this case it will be organized by Rosetta energy because we pulled the input list from the sorted score file. In terms of the energy landscape, Rosetta total score represents the depth of the energy well. A low score translates to an energetic minimum and therefore a candidate native conformation. The large clusters represent the wide energy wells that correspond to higher entropy. Cluster the PDB files using Calibur by typing the code below.

calibur -c AB pdb_list.txt > Cluster_calibur.out

5. Identify the largest three clusters and the top scoring decoys within them.

Clustering serves as a "hidden" score term to evaluate the realism of a decoy. Large clusters are more likely to contain the global minimum than small clusters. This is due to the expectation that the global minimum is surrounded by other low-energy structures. A very small cluster with good energy scores is more likely to be a false positive. In this way, clustering helps to account for errors in Rosetta scoring. Additionally, when selecting decoys for iterative refinement it is inefficient to include multiple very similar structures. It is recommended to use the center decoy in the top clusters as representatives to pass on for second stage sampling (*i.e.* more hybridize runs with these PDBs as inputs).

C.5.3 Interface Analyzer

Up to this point, Rosetta has scored the entire complex together. This step will focus on scoring the interface between the chemokine and the CKR. Many metrics, such as

the ΔG of interaction, will be calculated and used as a tool for analysis in step 5.

1. Run the interface analyzer command.

(/PATH/TO/ROSETTA/DIRECTORY)/main/source/bin/InterfaceAnalyzer.default.linuxgccrelease @ Interface_analyzer.options

After this step, a file will be generated named "interface_score_CCR5_CCL5.out".

2. Open the interface_score_CCR5_CCL5.out file in a text editor and note the values of each score term.

The meaning of each term can be found at https://www.rosettacommons.org/docs/latest/Home. Typically, the pertinent score term is the total energy. A broad rule of thumb is that the total energy score should be between negative 1-3 Rosetta energy units per residue in the system. In the benchmark test case, the top scores were approximately -1080. Another significant score term is the interface energy shown as dG_separated. This term is the difference in Rosetta energy in the bound complex compared to when the chemokine is separated from the CKR. Rosetta may favor large interfaces even if each additional contact is relatively weak. To control for this the interface energy term is divided by the solvent exposed area buried by the interface, as seen in dG_separated/dSASAx100. The top scoring decoys in the benchmark case had approximately -3 dG_separated/dSASAx100 values. Both the interface energy and total score must be used to sort successful decoys as poor scoring decoys with good interfaces presumably have no relevance.

3. Plot the top 10% of decoys by total score vs dG_separated/dSASAx100.

This plot will show the relationship between the Rosetta energy score and the calculated interface score. Similar to the RMSD plot generated in Section 4.1, this plot should have a trend where lower energy scores will have a lower ΔG of interaction (Figure C.6). If this trend is not observed it is possible that modeling is too restrictive, and additional inputs or reduced constraints may be required.

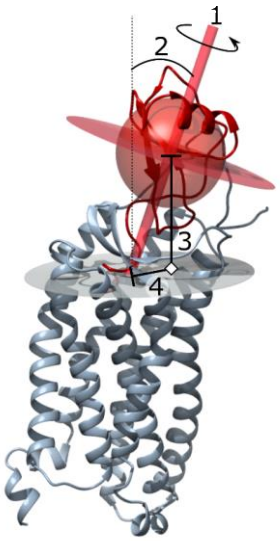
C.5.4 Coordinate-based Metrics

Each decoy generated thus far has been guided by the total Rosetta energy. A large portion of this energy comes from intrachain interactions that may not be directly involved in the chemokine-CKR interface. To give credence to the protein-protein interaction and to validate these decoys in a manner independent from Rosetta a series of spatial measurements can be calculated. A set of polar coordinate systems were designed to spatially represent any possible chemokine-chemokine receptor pair. These data can be generated by modifying the script provided or by independently calculating the metrics described in the following steps 1-8 (Figure C.6).

1. Define the 7 residues in the CKR that align with the outer lipid bilayer using predicted topology or the Orientations of Protein in Membranes database (http://opm.phar.umich.edu/).
2. Create a plane of best fit through the Cα of each of these residues with a mass weighted center point.
3. Calculate the center of mass of the chemokine and a mass weighted axis the runs the length of the chemokine.
4. Create a plane centered on the chemokines center of mass that is orthogonal to the chemokine axis.
5. Choose a CKR residue to serve as a reference point. The phenylamine at position 4.63x64 (Ballesteros-Weinstein numbering [367]) was chosen due to its conservation and its proximity to the membrane.
6. Project this residue's Cα onto the plane of the CKR. Do the same for the center of mass of the chemokine.
7. Similarly project the reference residue and the Cα of the cysteine in the chemokines 3rd β-strand onto the plane of the chemokine.
8. Measure the following angles and distances (Figure C.7).

This set of data should be generated for each decoy to describe the protein-protein interaction as well as for each chemokine-CKR complex available. Metrics may be added or modified based on your system such as the angle of the 30's loop of the chemokine, a known spot of interaction. Metrics such as these can validate and select for biologically probable decoys by comparing them to known values from the PDB.

9. Plot the coordinate-based metrics along with known chemokine-CKR complexes.



MEASUREMENT	DESCRIPTION
1. ROTATION	Angle along the chemokine plane between the F4.63, chemokine center of mass, and the β3 Cysteine.
2. TILT	Angle from the chemokine's axis to a line perpendicular to the plane of the receptor.
3. Z-DISTANCE	Distance from the chemokine's center of mass to the plane of the receptor.
4. TRANSLATION	Vector along the plane of the receptor from the center of the pocket to the chemokine's center of mass

Figure C.7 To quickly analyze the large number of decoys generated and to independently validate Rosetta modeling, a system of 3D metrics can be measured. The above scheme treats the chemokine and CKR independently and investigates the geometry of the interaction. Important metrics to consider are shown in the table to the right. These metrics can identify similarities and differences among the co-crystal complexes and the decoys generated.

By measuring specific distances and angles for the top PDBs output, we have distilled a large amount of the structural information down to a manageable size for swift analysis. These data can be compared to current chemokine-CKR complexes by constructing a box plot or violin plot of each metric. It is expected that some metrics, such as the chemokine rotation, will have large variability reflecting the degree of sampling. Other metrics, such as the Z-distance, may only vary by one or two Å but this difference may be very important to the success of modeling. Ultimately, this analysis gives insight into the amount of sampling performed and the chemical space it was performed in, as well as support the biological relevance of each individual decoy. Preferably the sampling is centered around the structure of the closest homolog. This can be used to identify aspects of modeling that were successful as well as those that are uncertain. These data can be used to optimize the modeling process by preferentially weighting the probability that a threaded input structure is sampled. These weights can be found in the CCR5_CCL5.xml file used in

Section 3.2.

C.6 Selection of a Final Model

A final model or ensemble of models should be chosen by building off of each previous section. First, evidence for convergence should be seen in the score vs. RMSD plots to validate the assumption that a probable model exists in the set of decoys generated. The score vs dG_separated/dSASAx100 plot should also show a positive correlation. Next, the top three clusters should be identified and sorted by total score. Select the top 10% by score for each cluster and sort by interface energy. Typically, Rosetta's rankings by score will identify realistic structures, but this is untrue if a decoy conflicts with biological information. For this reason, each candidate model must be visually inspected in a molecular visualization software such as PyMOL (Figure C.8). A background in structural biology coupled with experience is the best guide to parse out relevant models, but there are some main qualities to check for. The most obvious disqualifications will be unexpected chain breaks and unformed disulfide bonds. If either of these are commonly seen in the top decoys, it is likely that the input structures are not properly aligned before the hybridize step. Important contacts that are conserved in chemokines and CKRs should be maintained, and there should be significant contacts between these two proteins. The expected binding site should have a large contact area with close packing containing at most small voids between the chains. Most importantly, the structure should be consistent with experimental expectations. Pay close attention to any biological information available while you look through the top decoys**. In summary, a good structure will be in a large cluster, highly ranked by total score and interface energy, make significant contacts, and be consistent with biological information.** If several decoys meet the criteria for a

good model they can be used as an ensemble of models until new information becomes available. If no decoy meets these expectations, iterative rounds of hybridization (Section 3.2) can be done. The center decoy of the top three clusters and any promising decoys should be threaded and added as inputs.

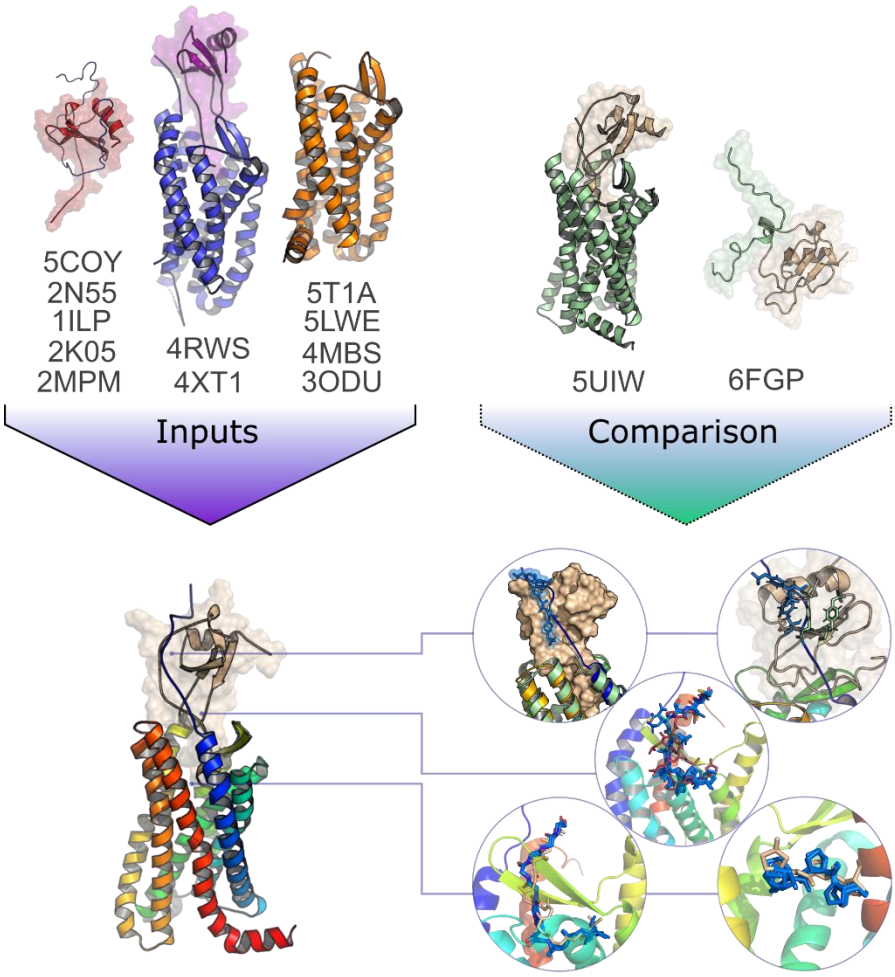


Figure C.8 In this protocol, eleven structures were used to generate decoys of 5P7-CCL5 bound to CCR5. This example was chosen to recapitulate the chemokine-CKR interface without using structures of the interaction. Promising decoys must be manually inspected and compared to the co-crystal complex (PDB ID: 5UIW) and the chemokine-peptide NMR complex (PDB ID: 6FGP). Images of site 1 are shown to demonstrate the similarity in sulfotyrosine positions of the model (color: blue) compared to the published structures (color: light green). Images of the chemokine N-terminus are compared to the co-crystal complex (color: wheat).

C.7 Concluding Remarks

C.7.1 Scope of Modeling

The goal of this protocol is to generate physically realistic models of chemokines bound to their CKR. These models were generated through homology modeling with little to no external data to look at the entire chemokine-CKR interface. If the answer to your biological question can be solved by focusing solely on partial interactions, alternative methods may be more efficient. It is also important to note that final models likely do not mimic the native state perfectly, and imperfections are likely clustered in areas of low information such as the CKR N-terminus. The underlying principles of this method can be applied to a broader range of GPCR-protein interactions such as β-arrestin or G protein complexes, although this has not been tested.

C.7.2 Online Resources

This protocol gives the user fine control over every stage of the modeling process. For entry level users or those restricted by time, there are several public servers available to quickly generate decoys using default settings. The FlexPepDock server found at http://flexpepdock.furmanlab.cs.huji.ac.il/ is useful for docking peptides to protein and can be used in place of this method to investigate site 2 interactions. It can also be used as a validation tool to dock peptides into CKRs generated from this protocol. Another validation tool can be found at https://swift.cmbi.umcn.nl/gv/whatcheck/index.html to inspect physical aspects of the model and generate a report describing structural warnings. The robetta server found at http://robetta.bakerlab.org/ can be used to quickly make homology models or perform a virtual alanine scan on the generated chemokine-CKR

models. This can be used in conjunction with the Rosetta design server found at http://rosettadesign.med.unc.edu/ to identify key residues that contribute to the chemokine-CKR interaction. Finally, the Rosetta Online Server that Includes Everyone found at http://rosie.rosettacommons.org/ has a plethora of tools such as small molecule docking that can be useful separately or in combination with this modeling.

C.7.3 Next Steps

Structural models have very little intrinsic utility unless they can be used to successfully predict biologically relevant outcomes. To support your model, *in vitro* assays should corroborate predictions *in silico*. In the chemokine-CKR system this can be accomplished using several methods: the functional effects of point mutations may be predicted by calculating the change in binding energy, small molecules that bind chemokine-CKR complexes can be identified through molecular docking, or residue-residue distances can be confirmed using biophysical techniques. Obtaining a structural model is required for many *in silico* techniques such as docking and molecular dynamics. Ultimately, these toolkits can give valuable insight into the underpinnings of chemokine-CKR interactions, generate hypotheses about binding specificity and affinity, and guide rational design projects.

APPENDIX D

Improved *In Vitro* Folding of the Y_2 G Protein-Coupled Receptor into Bicelles

D.1 Summary

Appendix D is related to Chapters 6 and 7 and describes our *in vitro* expression and refolding protocol for GPCRs for use in structural characterization. This appendix focuses on refolding of the neuropeptide Y type 2 receptor and validation of the function of these refolded receptors, but the same principles apply to the ghrelin and neuropeptide Y type 1 receptors. I helped in the refinement of the refolding protocol. Additionally, I developed the CPM assay for detection of disulfide bridge formation, and I imaged the bicelles using negative stain electron microscopy. This appendix comes from the article "Improved *in vitro* Folding of the Y_2 G-Protein Coupled Receptor into Bicelles" for which I am the second author [534].

D.2 Introduction

G protein-coupled receptors (GPCRs) play a central role in cell-cell communication and represent the largest group of membrane proteins with over 800 members in the human genome. These molecules transduce signals across the cell membrane via complex formation with extracellular ligands and intracellular interaction partners, namely G-proteins, kinases, and arrestins [27]. Interaction with intracellular effectors is mediated through structural rearrangements within the seven-transmembrane α-helix bundle (7TM)

and the loops connecting these α-helices. The dynamic nature of these binding processes has recently been shown in structural detail for the ß2-adrenergic receptor [535] and the A2A adenosine receptor [378]. Influencing these signal transduction pathways holds great potential for pharmaceutical research. Active components in several of the highest selling FDA approved pharmaceutical products in 2016 directly act on GPCRs, for instance in the treatment of depression, asthma, or pain (www.fda.gov). Structure based design of highly specific agonists and antagonists targeting GPCRs with reduced side effects requires comprehensive knowledge about the structure and dynamics of these membrane embedded molecules at different stages in their signaling process.

To date, over 150 crystal structures from 35 individual GPCRs in different activation states have been deposited in the protein database. This provides a large body of available data regarding structural features GPCRs as recently reviewed [27]. In spite of the significant breakthroughs these crystal structures provide for the GPCR field, they represent static conformations, typically achieved in a non-native solvation systme. Furthermore, for crystallography, GPCRs are usually engineered to reduce conformational heterogeneity and/or aid crystallization [536, 537] via the introduction of additional disulfide bonds [410], or replacement of loops with stabilizing proteins [335]. These snapshots of GPCRs can only partially reveal the rich dynamics. Therefore, non-crystallographic biophysical tools are required to fully characterize the dynamic nature of these conformationally complex membrane proteins [538, 539].

Complementary to crystallography and cryo-electron microscopy [174, 540], NMR spectroscopy represents a versatile method for obtaining structural information on both non-engineered GPCRs in a membrane (mimicking) environment [541] and also of their ligands bound to the receptors [2, 358, 372]. Both solution and solid-state NMR

spectroscopy provide complimentary NMR constraints for GPCR research [377, 378]. For example, chemical shift perturbation (CSP) measurements using solution NMR provided information about binding events in the receptor/water interface on G-protein in complex with the neurotensin-1 receptor [542]. In solid-state MAS NMR, the strengths of dipolar couplings were measured to obtain through space distance information for determining the CXCR1 structure [525] or characterizing the Y_2R dynamics [543]. Furthermore, the complementary use of restraints from solution and solid-state MAS NMR was demonstrated for the structure modeling of neuropeptide Y (NPY) in complex with its Y_2R [2].

In all the NMR studies mentioned above, the GPCRs were obtained from prokaryotic expression in *E. coli.* Either the receptors were expressed functionally [544], stabilized by directed evolution [545], or non-functionally in inclusion bodies [546, 547]. The latter method provides a feasible and economical method to express the required milligram amounts of non-engineered, isotopically labeled GPCRs for NMR studies. However, the molecules aggregated in inclusion bodies must be subsequently solubilized and folded *in vitro* into their functional state [548]. A number of studies using refolding of GPCRs into lipid environment have been published and demonstrated that valuable information on structure [525] or dynamics [543] can be obtained for the receptors alone or in complex with intracellular [549] or extracellular [2] binding partners. Nevertheless, developing efficient and successful folding protocols remains challenging and time consuming as each individual step in the refolding protocol introduces obstacles that must be must be overcome by optimization.

Of course, the benchmark indicating the success of the refolding protocol of GPCR samples for NMR studies are functionality assays of the folded molecules at various

concentrations. In most studies, functionality of the GPCRs, irrespective if refolded or functionally expressed, is measured in radioligand binding assays. These assays are performed at nanomolar receptor concentrations in order to determine the low nanomolar ligand affinities and to avoid the extensive use of expensive radioactive labeled material. However, in NMR measurements, GPCR concentrations in the high micro- to low millimolar range have to be used. During the necessary procedures to increase receptor concentration, the stabilizing environmental properties likely change with respect to protein/lipid or protein/detergent ratios, total receptor concentration, or solvent viscosity and might denature the protein and/or lead to protein aggregation. Hence, functionality should be confirmed at the protein concentrations required for structural measurements using the respective method.

Here, we present in detail an optimized three-step folding protocol of the human neuropeptide Y type 2 receptor (Y_2R) into phospholipid bicelles for both solution and solid-state magic-angle spinning (MAS) NMR measurements. The Y_2R is involved in the regulation of a number of physiological processes including food intake, neuroprotection, and circadian rhythm. As a consequence, the Y_2R is a putative target for therapeutics to treat obesity, epilepsy, schizophrenia, or anti-social behavior like aggression, depression, and drug addiction [550]. Furthermore, we show binding of the ligand neuropeptide Y (NPY) to the Y_2R as well as competence of the activated Y_2R to catalyze nucleotide exchange in G_i-protein using concentrations from the high nano- to the micromolar range using fluorescence and NMR spectroscopy.

D.3 Methods

D.3.1 Y_2R Sample Preparation

Expression of a cysteine deficient variant of the human Y_2R [551] in *E. coli* as inclusion bodies, receptor solubilization, and IMAC purification in 15 mM sodium dodecyl sulfate (SDS), 50 mM sodium phosphate (NaP), yielding ~20 mg Y2R per liter of expression medium, were performed as described before [401].

To refold the Y_2R into a functional state, a three-step folding protocol was developed (Figure D.1), which is explained in detail in the results section. The following buffers were used: in step 1, the purified Y_2R is dialyzed against a carefully degassed buffer containing 1 mM SDS, 50 mM NaP at pH 8.5, 1 mM EDTA, 1 mM reduced glutathione (GSH), and 0.5 mM oxidized glutathione (GSSG) at room temperature for 48 h using dialysis tubing with an 8-10 kDa molecular weight cut-off. Subsequently, 25 wt% poly(ethylene glycol) of a molecular weight of 20 kDa (PEG 20,000) is added to the same buffer to concentrate the receptor before reconstitution. In step 2, preformed bicelles consisting of 1,2-dimyristoyl-*sn*-glycero-3-phosphocholine (DMPC) and 1,2-diheptanoyl-*sn*-glycero-3-phosphocholine (DHPC-c7) (obtained from AvantiPolarLipids, Alabaster, USA) at a DMPC/DHPC-c7 molar ratio of 1:4 (q-value of 0.25) and dissolved in 50 mM NaP at pH 8.0 were incubated with the Y_2R, followed by three cycles of fast temperature changes from 42 °C to 0 °C with an incubation time of 25 min each. Visibly aggregated protein at any stage of the refolding protocol was removed by centrifugation at 5000 x g. In step 3, the Y_2R samples were either concentrated in small ($q = 0.25$) or large bicelles ($q > 10$). For small bicelle preparations, the samples were dialyzed at least three times against 20-30 wt% PEG 20,000, 1.5 mM DHPC-c7, 50 mM NaP at pH 7. For large bicelle

preparations, 50 mg/ml BioBeadsSM2 were added at least twice to the solution until the sample became slightly opaque. After removal of the beads with a sieve, the samples were washed four times through cycles of pelleting by centrifugation and resolubilization in 50 mM NaP at pH 7. Concentration determination of the membrane embedded receptors was performed by solubilization of the bicelles in 10 times the volume of 15 mM SDS, 50 mM NaP at pH7 and subsequent measurement of the Y_2R intrinsic absorption at 280 nm using UV-Vis.

D.3.2 Negative Stain Electron Microscopy

Y_2R prepared in bicelles of varying q values were diluted to 0.3-0.5 µM in 50 mM NaP, pH 7, 1 mM EDTA and 1.5 mM DHPC (only for low q values). A 3 µL sample was adsorbed onto a glow-discharged copper grid coated with a carbon film. The samples were washed with two drops of water and stained in two drops of uranyl formate (0.75%). Samples were visualized on a FEI Morgani electron microscope equipped with a 1 k × 1 k ATM CCD camera. The electron dose was set to 100 kV and magnification to 28,000×, unless otherwise noted.

D.3.3 Assessment of Disulfide Bridge Formation

To monitor the disulfide bridge formation, free cysteines were labeled with thiol-specific fluorochrome N-[4-(7-diethylamino-4-methyl-3-coumarinyl)phenyl]maleimide (CPM) (Alexandrov et al., 2008). A stock solution of CPM was dissolved in DMSO (4 mg/mL). The solution used in the experiments was further diluted using a 40 fold excess of buffer. A total of 10 µg of the Y_2R collected before and after each Step (D.3.1) were diluted in buffer containing 15 mM SDS to a final volume of 720 µL. A volume of 60 µL

of the working stock solution of CPM was added to the Y_2R and incubated at room temperature in the dark for 15 min. Data was collected on FluoroMax-2 (JOBIN YVON) in a 10 mm quartz cuvette at 20°C with an excitation wavelength of 387 nm, scanning emission wavelength from 450-500 nm, and integration time 0.5 s. All samples were scanned three times.

D.3.4 Fluorescence Polarization Ligand Binding Assay

Functionality of the Y_2R in nanomolar concentration was verified in a fluorescence polarization binding assays [348, 402] using [Dpr^{11}-atto520]-NPY. The reconstituted Y_2 receptor was incubated in increasing concentrations with the fluorescently labelled NPY at a concentration of 50 nM overnight at room temperature in 50 mM NaP at pH 7 in duplicate. The fluorescence spectra were recorded on the FluoroMax-2 using a 10 mm quartz cuvette at 20°C. The polarization units for each point were calculated from the maximal intensities of the four spectra measured in different planes and plotted against the receptor concentration as described in the literature [552]. As control, NPY binding to empty bicelles in the same concentrations as the receptor-containing bicelles were measured. In competition assays, constant concentrations of 50 nM Y_2R, 50 nM attoNPY and increasing concentrations of unlabeled NPY were used.

D.3.5 G-Protein Activation *in vitro*

Wildtype Gα protein was produced in *E. coli* and purified as described in the literature [553, 554]. Protein was stored at a concentration of 50 mM in Tris-Cl buffer, pH 8.0, 50 mM NaCl, 2 mM $MgCl_2$, 1 mM dithiothreitol, 10 µM guanosine diphosphate (GDP), and 10% glycerol at -80°C. $G\beta_1\gamma_1$ protein was isolated from bovine rod outer

segments as described earlier [555] and stored at a concentration of 10 mM in Tris-Cl buffer, pH 7.5, 100 mM NaCl, 5 mM 2-mercaptoethanol, and 10% glycerol at -80°C.

Nucleotide exchange in the basal state (Gα only) or catalyzed by activated receptor (R*-Gαβγ) was monitored as increase of intrinsic tryptophan fluorescence of W^{211}within switch II of $G\alpha_i$ [556] following binding to non-hydrolyzable GTPγS. Measurements were carried out at 16°C in semi-micro cuvettes (109.004F, Hellma, Müllheim, Germany) under constant magnetic stirring in a LS 50B fluorescence spectrometer (Perkin Elmer, Waltham, MA, USA), kinetic mode, constant photomultiplier voltage of 750 V, using excitation and emission filters of λ_{ex} 290/5 nm and λ_{em}345/5 nm, signal integration time of 800 ms, and signal interval of 2 s. For measurement of basal nucleotide exchange, fluorescence increase of 200 nM Gα (0.24 nmol in 1200 µl total volume) in 50 mM NaP/DHPC degassed buffer was monitored after addition of 82 µM GTPγS (10 µl of 10 mM stock in H_2O). For measurement of receptor-catalyzed nucleotide exchange, $G\alpha\beta_1\gamma_1$ (0.24 nmol; 10% molar excess of $\beta_1\gamma_1$) was pre-assembled in 10 µl Tris-Cl pH 7.5, 50 mM NaCl, 1 mM $MgCl_2$ for 10 min on ice. Y_2R (0.24 nmol) was activated with 10 fold excess of NPY in 20 µl NaP/DHPC (NaP for non-isotropic Y_2R preparations) for 30 min at room temperature. Y2R-NPY was allowed to bind pre-formed $G\alpha\beta_1\gamma_1$ for 10 min at 15°C, and the complex was added to the cuvette preloaded with degassed NaP/DHPC (NaP for non-isotropic Y_2R preparations). Samples were equilibrated in the cuvette for 5 min to ensure a stable baseline, and 82 µM GTPγS was added. GTPγS binding kinetics was fitted applying the built-in one-phase association function of GraphPad Prism 5.03 (GraphPad Software, San Diego, CA, USA) to obtain the apparent rate constant k.

D.3.6 Peptide Synthesis

Porcine NPY and isotopically labeled NPY variants were synthesized by combined manual/automated fluorenylmethyloxy-carbonyl/*tert*-butyl (Fmoc/*t*Bu) solid phase peptide synthesis in 15 µM scale on Rink amide resin as described before [557]. Fluorescently labeled NPY [Dpr22-atto520]NPY was synthesized as described [348]. Peptides were purified on a preparative reversed-phase high-performance liquid chromatography (RP-HPLC) system with C18 column (Jupiter 10U Proteo, Phenomenex, Aschaffenburg, Germany), applying linear gradients of 0.1% TFA in H_2O (eluent A) and 0.08% TFA in ACN (eluent B).

D.3.7 NMR Measurements

All NMR spectra were acquired on a on a Bruker 600 Avance III NMR spectrometer (Bruker BioSpin GmbH, Rheinstetten, Germany) at a resonance frequency of 600.1 MHz for ^{1}H, 150.9 MHz for ^{13}C, and 60.8 MHz for ^{15}N. Solution state experiments were conducted using a standard TXI probe. For solid-state NMR measurements, either a 4 mm MAS double or a 3.2 mm MAS triple resonance probe was used. Typical 90° pulse lengths for both probes were 4 µs for ^{1}H and ^{13}C and 5 µs for ^{15}N. ^{1}H dipolar decoupling during acquisition with and radio frequency amplitude of 65 kHz was applied using Spinal64. Chemical shifts were referenced externally (for ^{13}C relative to TMS). For the ^{13}C-^{13}C proton-driven spin diffusion spectra, a CP contact time of 700 µs and a mixing time of 500 ms was used. In the indirect dimension, 180 increments were accumulated. The relaxation delay was 2 s.

D.4 Results

D.4.1 In vitro Folding of the Y_2R into Bicelles

After *E. coli* expression and solubilization in SDS for IMAC purification, the Y_2R is ready for refolding. Figure D.1 shows a scheme of the three-step folding process for the Y2R to remove the denaturating SDS, form the native disulfide bridge, reconstitute into stable lipid environment, and finally to obtain concentrated samples of functional Y_2R in either isotropic bicelles (q = 0.25) for solution NMR or non-isotropic bicelles (q > 10) for solid-state MAS NMR studies. The folding process comprises three main steps; step 1 - the folding dialyses, step 2 –the reconstitution into bicelles, and step 3 – concentrating the sample for NMR in either isotropic or non-isotropic bicelles.

In **step 1**, the SDS concentration is reduced to just below its critical micelle concentration (cmc) by dialysis. This reduction of the SDS/receptor molar ratio enables the formation of intramolecular contacts between the receptor α-helices and allows a pre-formation of the α-helix bundle. Best results, characterized by less than 10% protein aggregation, were achieved when using a concentration of 10 µM for the Y_2R and 1 mM for the SDS in the dialysis resulting in a SDS/Y2R molar ratio of 100. Using lower SDS/Y_2R ratios resulted in lower reconstitution yields in subsequent folding step 2, while higher ratios lowered the proportion of active protein. Additionally, the glutathione based redox-shuffling system was added at this step to ensure the formation of the native disulfide bridge in this cysteine reduced variant of the Y_2R [551].

Prior to reconstitution in **step 2**, the Y_2R is concentrated to 20-30 µM by adding PEG 20,000 to the dialysis buffer. Through this concentration step, the sample volume as well as the SDS/Y_2R ratio is reduced, which slightly improved reconstitution yields. More

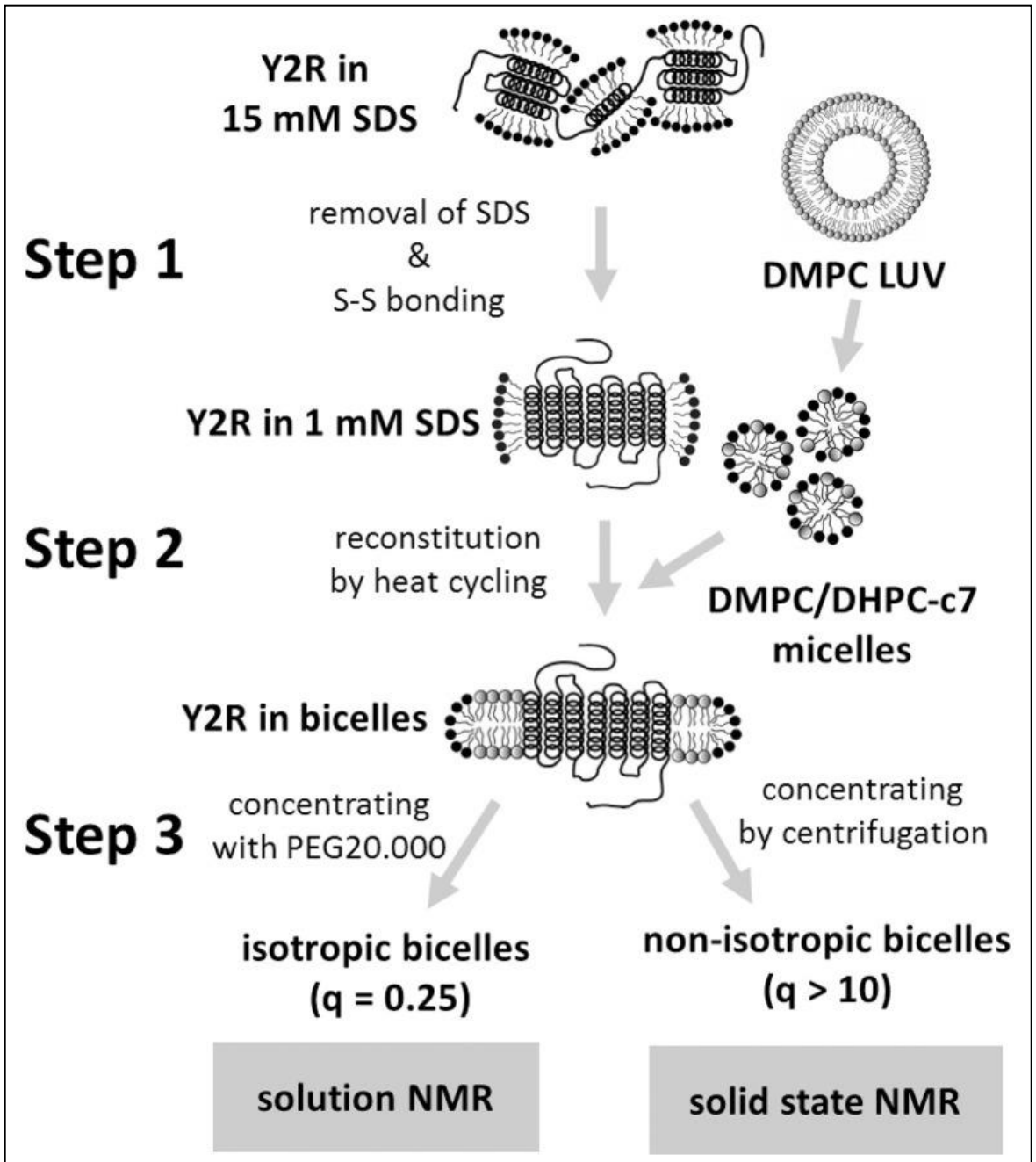


Figure D.1 Scheme of the three-step folding protocol for the preparation of the Y_2R in either isotropic or non-isotropic bicelles. The main steps are the folding dialysis (step 1), the reconstitution into bicelles (step 2), and concentrating the sample for NMR measurements (step 3).

importantly, the reduction of the sample volume simplifies and accelerates the concentration in step 3 when preparing samples for solution NMR measurements. At low SDS concentration, the concentrated receptor molecules are more prone to oligomerize and the step 2 reconstitution has to be performed directly afterwards. Reconstitution is achieved by addition of freshly prepared DHPC-c7/DMPC mixed micelles, solubilized to a q-value of 0.25 from preformed DMPC vesicles of 100 nm diameter and a concentration of 10 mg/ml DMPC [401]. The DMPC/Y_2R molar ratio depends on the final desired preparation. For the preparation of non-isotropic bicelles, a ratio of 180/1 is used, while for isotropic bicelles, a ratio of 400/1 shows the best results. Ratios below these values drastically reduce either the reconstitution yield or, in case of the isotropic bicelles, the stability in the final sample after concentrating the sample in step 3.

Step 2, reconstitution of the Y_2R from the low SDS concentration environment into the DMPC bilayer, is achieved through a heat cycling process [558], where the solution containing the receptors in SDS micelles and the DMPC/DHPC-c7 bicelles is alternately heated and cooled well above and below the phase transition of the lipid/detergent mixture, respectively. This procedure alters the lateral forces acting on the receptor between mixed micelles and isotropic bicelles or rather between bilayer and non-bilayer formation, and hence facilitates the replacement of the high cmc SDS detergents by the very low cmc DMPC lipids on the hydrophobic core of the receptors, which should form a stable bilayer around the α-helical receptor bundle. The unfavorable SDS is replaced by a zwitterionic phospholipid bilayer applying a well-defined lateral pressure profile onto the receptor [559] and finally allows for the native orientation of the transmembrane helices in the membrane mediated by side chain contacts with lipids. At the end of step 2, the Y2R is stabilized in the isotropic bicelle solution at a concentration of 15-20 µM.

In **step 3**, the sample has to be concentrated to the micromolar range required for NMR measurements, and impurities such as residual SDS molecules, glutathione from redox-shuffling system, and EDTA, which disturb the NMR measurements, have to be removed. The procedure for sample concentration depends on the measurements the sample is prepared for. For solution NMR measurements, it is important to maintain the isotropic bicelles at a q-value of 0.25. Therefore, the Y_2R sample is concentrated by dialysis against a buffer containing PEG 20,000 for the removal of the water and DHPC-c7 slightly above the CMC for maintaining the receptor/lipid/detergent ratio. Multiple buffer exchanges assure the removal of SDS, glutathione, and EDTA. A sample stability of at least 48 h, which is required for the solution NMR spectra shown below, was achieved with samples concentrated up to 160 µM Y_2R. At longer time periods or higher concentration, the isotropic bicelles started to fuse of aggregate to larger lipid complexes and were not useful for solution NMR measurements indicated by substantial line broadening.

In contrast to solution NMR, large complexes are well suited for solid-state MAS NMR measurements. Here, one limitation is the total amount of receptor in the sample, and hence the signal intensity, which is constrained by the available volume of the MAS rotor. The concentration of active protein can consequently only be increased by reducing the amount of all other components in the sample such as lipid and water. As stated above, the lowest DMPC/Y_2R ratio, required for functional reconstitution in step 2, was 180/1. The DHPC-c7 on the other hand, which is required for reconstitution in step 2 and for maintaining isotropic bicelles conditions for solution NMR measurements in step 3, is not necessary for the functionality of the Y_2R once embedded in the bilayer. Therefore, the detergent is removed in multiple incubation steps using BioBeadsSM2 [560], which

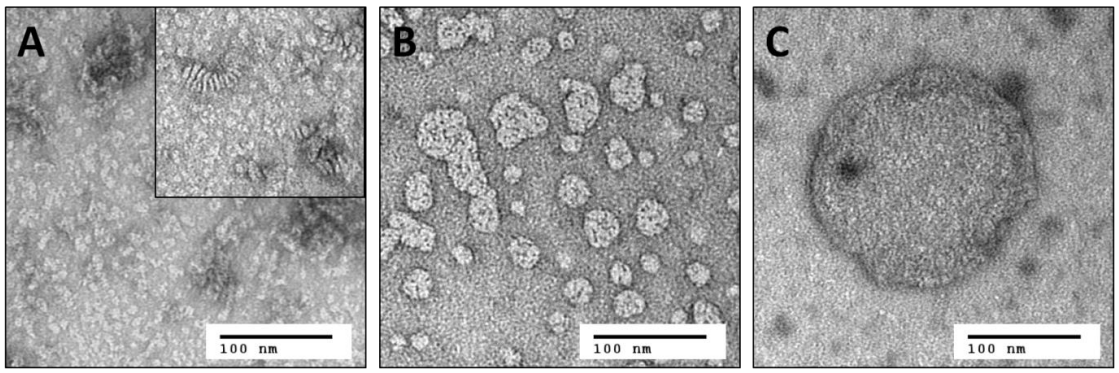


Figure D.2 Negative staining electron microscopy images of (A) small isotropic bicelles (q = 0.25), (B) intermediate sized bicelles, and (C) non-isotropic bicelles (q > 10). The inset in (A) shows the same sample after one week of storage at room temperature. Stacking of the bicelles becomes visible, which leads to reduced binding yields and substantial line broadening in solution NMR spectra. Samples from (A) and (C) are used for solution and solid-state MAS NMR, respectively. Image (B) illustrates the fusion of the small bicelles to larger patches during removal of the DHPC-c7 and hence to an increased q-value.

changes the q-value from 0.25 to above 10 and facilitates the fusion of the bicelles from small structures to large, non-isotropic bicelle-like patches with a diameter of 300 to 500 nm, as visualized in Figure D.2. The DMPC/DHPC-c7 ratio was determined from one-dimensional ^{1}H solid-state MAS NMR spectra. The large non-isotropic bicelles can easily be pelleted by centrifugation, which is used for the removal of SDS, glutathione, and EDTA using several washing steps, and for concentrating the sample. The final Y_2R sample for solid-state NMR measurements contained ~6 mg Y_2R in a 50 µl NMR rotor with a water content of ~50%, determined by weighing before and after lyophilization. In contrast to the isotropic bicelle samples, the Y_2R embedded in non-isotropic bicelles was stable for at least one month at -20°C, displaying no changes in the NMR spectra.

D.4.2 Disulfide Bridge Formation of Refolded Y_2R

To assess disulfide bridge formation during the three-step folding process the free cysteines were labeled with CPM and detected in fluorescence measurements [561], shown

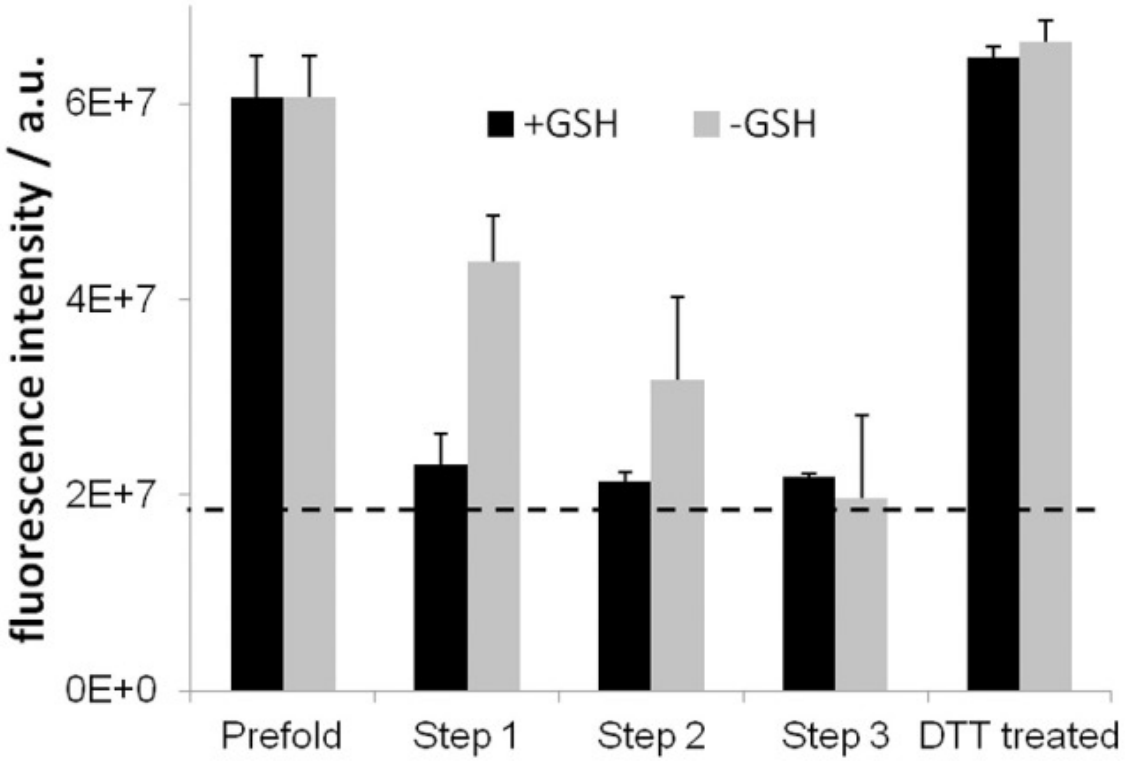


Figure D.3 Results of the CPM assay for testing disulfide bridge formation. High fluorescence intensities designate free cysteine residues. The dotted line indicates the background fluorescence intensity. Disulfide bridges are formed during the folding process to completeness after step 3. Glutathione (GSH) accelerates the formation, which is fully reversible shown by reducing the cysteines using DDT.

in Figure D.3. As expected, in the presence of the glutathione based redox-shuffling system, the two remaining cysteines in the Y_2R sequence are bridged in the step 1 folding dialysis almost completely and remains stable over all steps. Surprisingly, also in the absence of glutathione, the cysteines become connected, although to a lesser extent and after a longer time period.

D.4.3 Fluorescence-Based Ligand Binding of Refolded Y_2R

Functionality of the Y_2R samples at nanomolar concentrations were tested in fluorescence based assays to test the folding yields and the accessibility of both sides of the Y2R embedded in bicelles, the ligand binding site as well as the G-protein binding site.

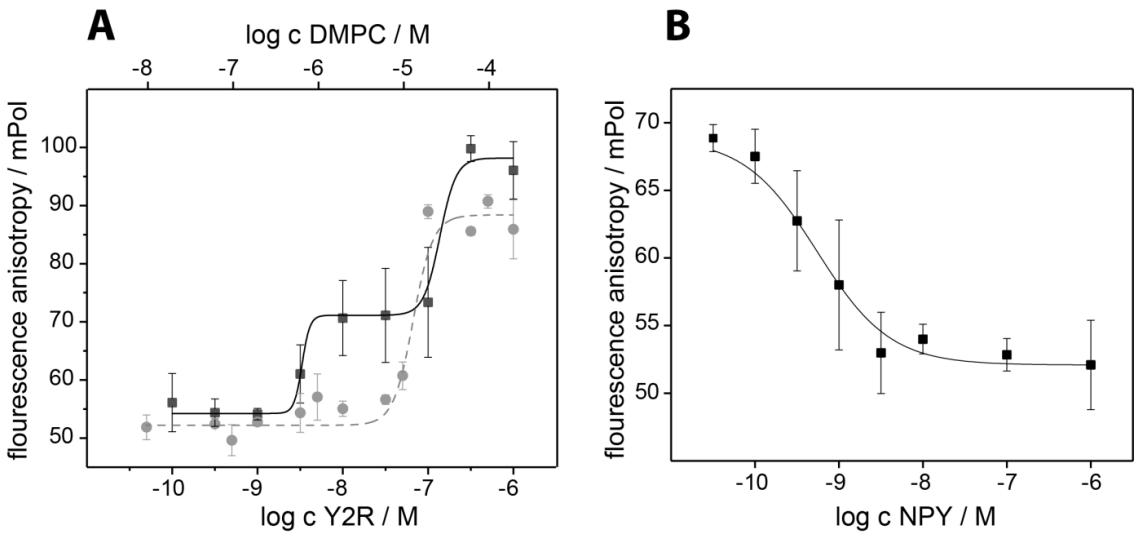


Figure D.4 Pharmacological characterization of the Y_2R preparation at nanomolar concentration using a fluorescence polarization assay with [Dpr^{22}-atto520]-NPY. The saturation assay in **(A)** recorded at increasing concentration of bicelle-reconstituted Y_2R (black) displays two inflection points at 4 and 126 nM. The latter represent the binding of NPY to the membrane, as revealed by the reference measurement with empty bicelles (gray). The specificity of the high affinity Y_2R binding of 4 nM is confirmed in the competition assay in **(B)** using increasing concentration of unlabeled NPY. The error bars were determined from three independent preparations.

In Figure D.4A, saturation curves of NPY binding to Y_2R and empty bicelles are shown. In the presence of the Y_2R, two inflection points at values of (4 ± 3) nM and (126 ± 52) nM could be detected from the fit of a two-side binding model to the data points. The higher value displays the binding of NPY to the membrane as shown by the fit to the data points in the absence of the Y_2R, while the lower value displays the low nanomolar affinity of the Y_2R. Binding assays in the presence of BSA showed weaker membrane affinities for NPY, but also shifted the affinity to the Y_2R to higher values (data not shown), implying that pre-binding of NPY to the membrane and though increasing the effective concentration supports receptor binding [153]. Assuming that about ten DMPC molecules are needed to bind one NPY, the affinity of NPY to DMPC membranes is calculated from the inflection point to 2.5 µM. The displacement assay in Figure D.4B verified the specificity of the Y_2R binding, showing an EC50 value similar to the K_D value in the saturation assay. It is of notice that this assay can only be carried out with a Y2R

concentration between the two inflection point values determined from the saturation assay, because at lower concentrations no polarization beyond background can be detected and at higher concentrations the membrane binding dominates the measurement, due to higher polarization.

D.4.4 G-protein Activation by Refolded Y_2R

To assess functionality of Y_2R preparations with respect to G-protein activation, intrinsic tryptophan fluorescence readout was used, exploiting an activity-dependent increase of W^{211} fluorescence within switch II of $G\alpha_{i1}$ [556]. While W^{211} fluorescence is low in the GDP-bound states, this residue inserts into a hydrophobic pocket upon binding to GTP or GTP analogs, which strongly increases its intrinsic fluorescence. In unbound Gα

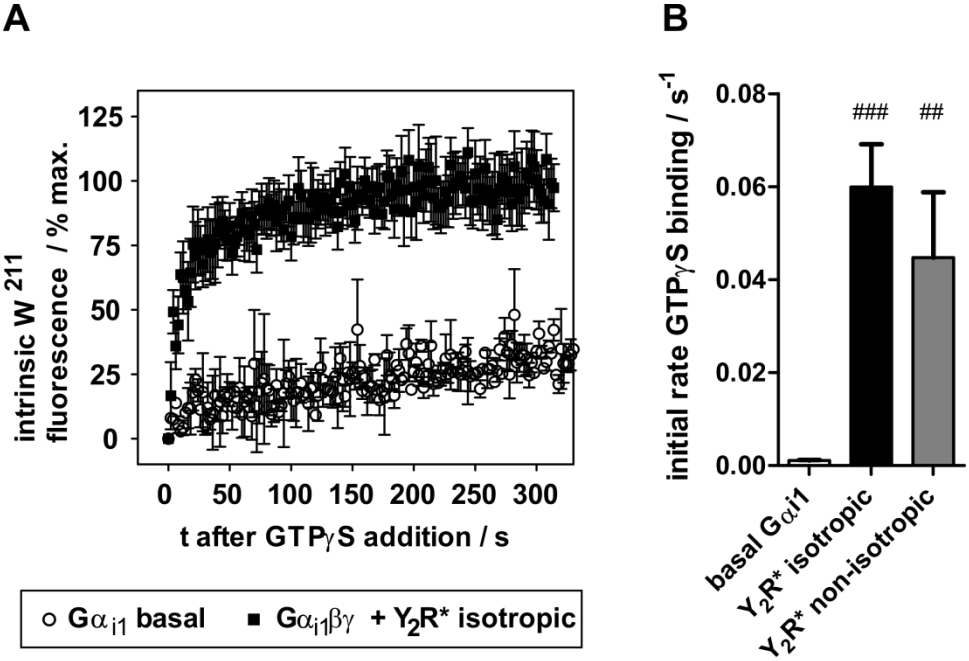


Figure D.5 *In vitro* folded Y_2R variants functionally activate purified G_i protein. **(A)** Y_2R folded into isotropic bicelles and activated with NPY drastically accelerates nucleotide exchange of wild type $G\alpha_{i1}$. The fluorescence trace is given as mean of seven independent experiments. **(B)** Resulting apparent rates of GTPγS binding of $G\alpha_{i1}$ (basal) and Y_2R-catalyzed nucleotide exchange. Statistical significance was determined using one-way ANOVA/Dunnett's post hoc test against basal $G\alpha_{i1}$ in Graph Pad Prism 5.03. $^{###}p < 0.001$; $^{##}p < 0.01$.

subunits, nucleotide exchange is very slow, and even essentially absent in G-protein heterotrimers (Gαβγ). Activated GPCRs (R*) act as nucleotide exchange factors, when binding Gαβγ-GDP, leading to the high affinity R*- Gαβγ 'empty' complex, followed by GTP binding and Gα activation (Gα-GTP). In isolated *in vitro* systems, GTPγS can be added to trap Gα in the activated state and to follow GTPγS binding kinetics by tryptophan fluorescence. An exemplary fluorescence trace is shown in Figure D.5A. While basal GTPγS binding of $G\alpha_{i1}$ is very slow ($k < 0.002\ s^{-1}$), $G\alpha_{i1}\beta\gamma$ allowed to interact with the NPY-activated receptor preparations displayed a greatly accelerated nucleotide exchange with apparent GTPγS binding rate of $0.06 \pm 0.01\ s^{-1}$. Comparable GTPγS binding rates were also observed in non-isotropic samples ($k = 0.045 \pm 0.014\ s^{-1}$) (Figure D.5B).

D.4.5 NMR Experiments to Assess Receptor Function in Micromolar Concentration

Next we acquired fingerprint NMR spectra of the Y_2R samples in both preparations, small and large bicelles, at concentrations sufficient for NMR measurements. Ligand binding of the Y_2R in small bicelles was assessed by recording chemical shift

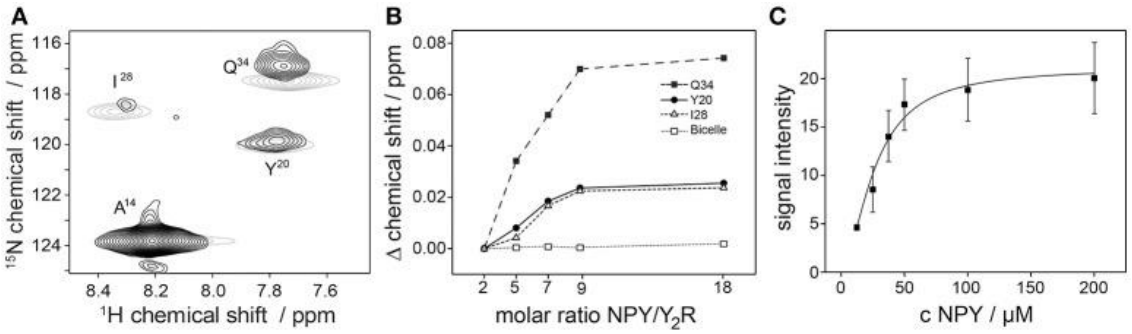


Figure D.6 NPY binding tests of the Y_2R preparation in concentration of 50 µM using solution NMR spectroscopy. In **(A)** a ^{1}H-^{15}N HSQC spectra of specifically labeled NPY in the presence of Y_2R with a NPY/Y_2R ratio of 2 (black) and 18 (gray) are shown. Chemical shift perturbations (CSP) were measured for the labeled NPY positions Y20, I28, Q34 which are involved in Y_2R binding [2], but not for A14 which is not interacting with the Y_2R. As control the same amounts of NPY were titrated to empty bicelles to exclude self-aggregation effects of NPY at high concentration. Concentration dependent binding effects were verified in **(B)** by observing the CSP at increasing concentrations of NPY binding to Y_2R in isotropic bicelles, and in **(C)** by measuring the signal intensities of bound NPY to Y_2R in non-isotropic bicelles. All spectra were recorded at 293 K.

perturbations of isotopically labeled [^{15}N-A^{14},Y^{20},I^{28},Q^{34}]NPY in interaction with the receptor (Figure D.6). To this end, specifically labeled NPY was titrated to 60 µM of Y_2R in molar ratios from 2 to 18 and ^{1}H-^{15}N HSQC spectra were acquired for each sample. The NMR spectrum at the lowest molar ratio compared to the spectrum at the highest ratio is shown in Figure D.6A. Chemical shift perturbations were observed for all labeled positions except for A^{14}, as it was reported before [2]. Furthermore, changes in the chemical shift at different ratios could be measured and are plotted as difference to the chemical shift determined at the lowest ligand to receptor ratio (Figure D.6B). Ligand binding on Y_2R in large bicelles at NMR concentration was tested using a pull-down assay. Varying concentrations of the isotopically labeled NPY were incubated with 40 µM Y_2R each for 2 h, subsequently pelleted and the unbound NPY in the supernatant was removed. The Y_2R/NPY complex containing pellets were solubilized in SDS to denature the receptor and hence release the NPY. The signal integrals of the NPY, corresponding to the amount, were recorded in ^{15}N filtered ^{1}H spectra, corrected for intensities measured using empty bicelles, and plotted over the NPY concentration used for incubation (Figure D.6C).

Although the determination of binding affinities at micromolar receptor concentration is hardly possible because too many assumptions have to be made, the assays performed here show clear concentration dependent ligand binding effects. Thereby, the presented measurements represent an option to test receptor samples in concentrations required for structural studies.

D.4.6 Carbon-Carbon Correlation MAS NMR Measurement

To demonstrate the high efficiency of the folding protocol and the feasibility of the samples for the application to NMR measurements, we recorded a solid-state MAS ^{13}C/^{13}C

DARR correlation spectra of uniformly ^{13}C-labeled Y_2R (Figure D.7) at two mixing times. Even at a mixing time of 20 ms, where polarization can be transferred only between neighboring carbons, a high number of partly resolved signals are visible. Increasing the mixing time to 500 ms allows detecting long range correlations which may indicated tertiary contacts providing valuable constraints for structural studies. Indeed, the number of crosspeaks drastically increases under these conditions.

We predicted one bond correlations (Cα/Cβ, Cβ/Cγ) from a Y_2R homology model [2] in the DARR spectrum at short mixing time using ShiftX2 [562] and superimposed them with the experimental NMR spectrum (Figure D.7). Overall we found a rather good agreement between experimental and model-based chemical shifts. An interesting exception is that we find Ala and Leu peaks that indicate β-strand like structure. According

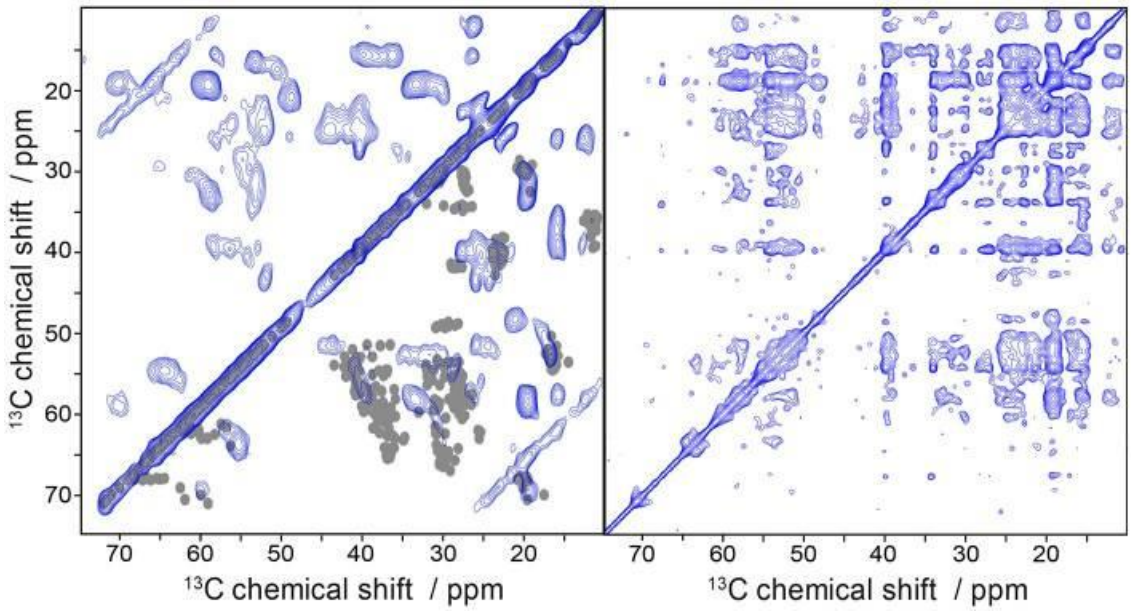


Figure D.7 Solid-state MAS NMR spectra of uniformly labeled Y_2R in non-isotropic bicelles showing $^{13}C/^{13}C$ correlation using DARR. The mixing time was varied from 20 ms **(left)** to 500 ms **(right)**. In the bottom right half of the 20 ms DARR spectrum are superimposed one bond correlations cross-signals, simulated from an Y2R homology model. The measurements were performed at a MAS frequency of 7 kHz and a temperature of 5°C.

to the model, the Y2R features one β-strand in ECL2 comprising residues 183 to 207. This stretch of amino acids contains two Ala (A184, A202) and two Leu residues (L183, L191), which could produce the β-strand like NMR shifts, although ShiftX2 does not predict such chemical shifts. Peak intensity for Ser and Thr residues agrees relatively well with the predicted chemical shifts from the model.

D.5 Discussion

NMR spectroscopy can be a valuable method in structure-based GPCR research, especially when acquiring data on non-engineered receptors in a membrane environment [378, 563]. Expressing GPCRs in *E. coli* as inclusion bodies and subsequent refolding into membrane environment provides a feasible and successful strategy to purify the required amounts of isotopically labeled molecules. The first solid-state NMR structure of a GPCR, the CXCR1 receptor, was determined by applying this strategy [525]. Further, we recently reported a structural model of the peptide NPY bound to the Y_2R based on NMR restraints from solution and solid-state MAS NMR [2] and revealed the comprehensive dynamical features of the Y_2R reconstituted into bicelle environment [543, 564]. Here, we discuss an improved folding and preparation protocol, which is a prerequisite for obtaining structural and dynamic data on the Y_2R and GPCRs in general.

After inclusion body expression and purification, the GPCRs are generally solubilized in SDS micelles [548]. Although the SDS-solubilized receptors are completely non-functional, they already contain most of the native secondary structure including all α-helices, as it has been shown for the BLT1 receptor [565], the μ-opioid receptor [566], and the Y_2R [547]. While the presence of 15 mM anionic detergent SDS suppresses crucial

intramolecular contacts for native tertiary structure resulting in non-functional receptor, it prevents aggregation and oligomerization thereby rendering it a good starting point for *in vitro* folding.

During *in vitro* folding of the Y_2R for NMR measurements, three major steps have to be applied: (*i*) removal of the denaturing SDS without losing receptor molecules by aggregation, (*ii*) formation of the native disulfide bridge between two cysteines in TM3 and ECL2, and (*iii*) high yield reconstitution of the receptor into a stable environment, in which micro- to millimolar protein concentrations can be achieved.

To remove the SDS, its concentration was decreased to 1 mM in step 1 of the folding process. This concentration is below the cmc of SDS, which was determined to 1.9 mM under these conditions [551]. Further, the Y_2R concentration was adjusted to a SDS/Y2R ratio of 100. Interestingly, this is on the order of the SDS aggregation number, specified with 62-101 molecules per micelle under similar conditions [567]. This suggests that the Y2R is not kept in a detergent micelle at step 1 of the folding process. What is conceivable instead is that the hydrophobic regions of the molecule are covered by a few SDS molecules, which apparently have a high affinity to the receptor. The hydrophobicity of the receptor seems to have a strong effect on the equilibrium of SDS between the monomeric and the micellar state. Similar effects have been reported on helical domains [568, 569]. The fact that the Y_2R is not covered in a large micelle might enable the high yield reconstitution into the phospholipid bilayer. In step 3 of the folding process, the residual SDS was removed below detection limit of ^1H NMR in the final sample.

A straightforward strategy for an effective formation of the native disulfide bridge between the two cysteines in TM3 and ECL2 and prevention of non-functional bridging between free cysteines is to reduce their number in the sequence to the possible minimum

[563, 570]. Following this strategy for the Y_2R, all cysteines were exchanged to serine or alanine except for the two cysteines involved in the required disulfide bridge. Fortunately, these mutations did not interfere with the functionality of the receptor [551]. Interestingly, even the mutation of the putative palmitoylation site at the C-terminus did not alter cell surface expression and signaling properties of the Y_2R [571]. We achieved the complete disulfide bridge formation in step 1 of the folding process accelerated through the use of glutathione. This is important because at this step the monomeric Y_2R is still protected by ionic SDS and therefore the formation of intermolecular non-native disulfide bridges is avoided.

We reconstituted the Y_2R into DMPC/DHPC-c7 bicelle-like structures, which is known to represent a much more stable environment for GPCRs than detergent micelles [405, 551]. Similar to SMA- or MSP-nanodiscs [402, 572, 573], both intra- and extracellular sides of the receptors are accessible in bicelles for interacting molecules, whereas in liposomes there is only one. This enables detection of ligand as well as G-protein interaction. Additionally, the size of the bicelles can be adjusted by varying the q-value (molar DMPC/DHPC-c7 ratio) from isotropically tumbling bicelles ($q < 0.25$) to large non-isotropic membrane structures with little residual detergent [574]. Therefore, very similar preparations of receptors in bicelles can be used in complementary methods of solution and solid-state NMR [2]. Especially in our preparation of non-isotropic bicelles, the Y_2R is densely packed within the membrane. Considering the given lipid/Y_2R ratio and assuming cross-sectional areas of 60 $Å^2$ for DMPC, 2000 $Å^2$ or the Y2R and 107 $Å^2$ per bicelle, we can estimate that each receptor is surrounded by a lipid annulus comprising 2 to 3 molecular layers.

Using fluorescence polarization assays, strong affinities of NPY to the Y_2R with a

K_D value of (4 ± 3) nM, but also of NPY to the DMPC membrane of 2.5 µM, assuming ten DMPC molecules bind to one NPY molecule, were measured. Unfortunately, the determination of the high affinity inflection point is limited by the concentration of the labeled ligand and thus does not provide a true K_D. Due to the limitations in fluorescence detection, the binding assay is conducted far above the expected sub-nanomolar equilibrium binding constant with labeled ligand and receptor present in the same concentration range. Under these conditions, the apparent K_D is no longer independent of the number of binding sites to be saturated and deviates from true K_D, as the apparent K_D = true K_D + ½ [Ligand] [575]. Thus, ½ [Ligand] represents the assay limit assuming a very high affinity K_D. Within experimental error including concentration determination of labeled ligand and receptor, this limit is met for the Y_2R preparations. Thus, high affinity binding of the receptor can be concluded, with a K_D that is at least in the low nanomolar range.

The measured fluorescence anisotropy of the atto520 labelled NPY corresponds to the degree of freedom of the fluorescence label. Therefore in the bound state the label has a lower degree of freedom and hence higher anisotropies are expected for NPY. Surprisingly, higher anisotropies were detected than for the membrane bound NPY than the receptor bound NPY. The NPY was labeled on position 22 directly before the C-terminal α-helix. When bound to the membrane, large parts of the NPY helix are in contact with the phospholipids [153] and the NPY molecule lays flat on the membrane, restricting the NPY diffusion to a two-dimensional surface. In contrast, the structural model of NPY bound to the Y_2R shows a rather steep pose of NPY with respect to the membrane normal, where only the C-terminal part is interacting with the receptor and N-terminal half, including residue 11, sticks out into the solution [2].

Before collecting data on the structure and dynamics of GPCRs using NMR experiments, it is necessary to prove functionality of the receptor samples in high concentrations as used in NMR spectroscopy. However, a comprehensive pharmacological characterization in terms of affinities or K_D values is difficult under these concentrations. Here, we present two approaches to confirm a concentration dependent binding response using NMR measuring either CSPs [576] or intensities of specifically labeled NPY bound to Y2R. Because of the size and the dynamic features of wild type GPCRs [539], the NMR spectra are dominated by short T_2 relaxation times, resulting in relatively large line widths and thus low spectral resolution in solution NMR [563] as well as in solid-state NMR [543]. Using peptides with only a few labeled amino acids reduces the number of signals in each NMR spectrum, thereby providing a simple way to avoid signal overlap and enabling straightforward signal assignment. This approach was used in the binding assays. However, in first solution NMR experiments, the signals of amino acids strongly involved in Y_2R binding were strongly broadened and signal intensities dropped below detection limit. To still be able to assign these important amino acids in the bound state, we have used a minimum of twofold molar excess of NPY in solution NMR. Therefore, the signals represent an average of the intensities from bound and free NPY. In consequence, CSPs presented in Figure D.6 also represent an average and would even be higher for the bound NPY assuming fast exchange on the NMR time scale.

Finally, we recorded first $^{13}C/^{13}C$ DARR correlation spectra to prove that receptor loading within the bicelles and hence signal intensities are sufficient for solid-state MAS NMR measurements. Both short- and long range proximities can be probed by this approach, provided (nearly) full signal assignment can be achieved. Given the high degree of control over synthesis pathways during *E. coli* protein expression using appropriately

labeled precursors, specifically isotopically labeled receptor variants can be produced which should simplify the NMR assignment [577-579].

D.6 Conclusions

In conclusion, we present a robust and efficient protocol for functional reconstitution of the Y_2R into either isotropic or non-isotropic phospholipid bicelles. The preparations provide the receptor concentrations required for spectroscopic methods, like solution and solid-state NMR or EPR. The protocols can be adapted to other GPCRs as already shown for the GHSR [348]. Further, samples can not only be prepared from GPCRs expressed in *E. coli* as inclusion bodies, but also from all other precipitated GPCRs, such as from cell-free expression produced in the PCF-mode [580].

BIBLIOGRAPHY

1. Gregory, K.J., et al., *Probing the metabotropic glutamate receptor 5 (mGlu(5)) positive allosteric modulator (PAM) binding pocket: discovery of point mutations that engender a "molecular switch" in PAM pharmacology.* Mol Pharmacol, 2013. **83**(5): p. 991-1006.
2. Kaiser, A., et al., *Unwinding of the C-Terminal Residues of Neuropeptide Y is critical for Y(2) Receptor Binding and Activation.* Angew Chem Int Ed Engl, 2015. **54**(25): p. 7446-9.
3. Tinberg, C.E., et al., *Computational design of ligand-binding proteins with high affinity and selectivity.* Nature, 2013. **501**(7466): p. 212-6.
4. Merten, N., et al., *Receptor subtype-specific docking of Asp$^{6.59}$ with C-terminal arginine residues in Y receptor ligands.* J. Biol. Chem., 2007. **282**(10): p. 7543-51.
5. Song, Y., et al., *High-resolution comparative modeling with RosettaCM.* Structure, 2013. **21**(10): p. 1735-42.
6. Crooks, G.E., et al., *WebLogo: a sequence logo generator.* Genome Res, 2004. **14**(6): p. 1188-90.
7. Vortmeier, G., et al., *Integrating solid-state NMR and computational modeling to investigate the structure and dynamics of membrane-associated ghrelin.* PLoS One, 2015. **10**(3): p. e0122444.
8. Laskowski, R.A. and M.B. Swindells, *LigPlot^{+}: multiple ligand-protein interaction diagrams for drug discovery.* J. Chem. Inf. Model., 2011. **51**(10): p. 2778-86.
9. Raveh, B., et al., *Rosetta FlexPepDock ab-initio: simultaneous folding, docking and refinement of peptides onto their receptors.* PLoS One, 2011. **6**(4): p. e18934.
10. DeLano, W.L., *The PyMOL Molecular Graphics System* 2007, DeLano Scientific LLC: Palo Alto, CA, USA.
11. Flock, T., et al., *Selectivity determinants of GPCR-G-protein binding.* Nature, 2017. **545**(7654): p. 317-322.
12. Hauser, A.S., et al., *Trends in GPCR drug discovery: new agents, targets and indications.* Nat Rev Drug Discov, 2017. **16**(12): p. 829-842.
13. Santos, R., et al., *A comprehensive map of molecular drug targets.* Nat Rev Drug Discov, 2017. **16**(1): p. 19-34.
14. Riddy, D.M., et al., *G Protein-Coupled Receptors Targeting Insulin Resistance, Obesity, and Type 2 Diabetes Mellitus.* Pharmacol Rev, 2018. **70**(1): p. 39-67.
15. Brelot, A. and L.A. Chakrabarti, *CCR5 Revisited: How Mechanisms of HIV Entry Govern AIDS Pathogenesis.* J Mol Biol, 2018. **430**(17): p. 2557-2589.
16. Zhao, J., et al., *G Protein-Coupled Receptors (GPCRs) in Alzheimer's Disease: A Focus on BACE1 Related GPCRs.* Front Aging Neurosci, 2016. **8**: p. 58.
17. Nieto Gutierrez, A. and P.H. McDonald, *GPCRs: Emerging anti-cancer drug targets.* Cell Signal, 2018. **41**: p. 65-74.
18. Moody, T.W., I. Ramos-Alvarez, and R.T. Jensen, *Neuropeptide G Protein-Coupled Receptors as Oncotargets.* Front Endocrinol (Lausanne), 2018. **9**: p. 345.

19. Wacker, D., R.C. Stevens, and B.L. Roth, *How Ligands Illuminate GPCR Molecular Pharmacology.* Cell, 2017. **170**(3): p. 414-427.
20. Bockaert, J. and J.P. Pin, *Molecular tinkering of G protein-coupled receptors: an evolutionary success.* EMBO J, 1999. **18**(7): p. 1723-9.
21. Fredriksson, R., et al., *The G-protein-coupled receptors in the human genome form five main families. Phylogenetic analysis, paralogon groups, and fingerprints.* Mol Pharmacol, 2003. **63**(6): p. 1256-72.
22. Pedragosa-Badia, X., J. Stichel, and A.G. Beck-Sickinger, *Neuropeptide Y receptors: how to get subtype selectivity.* Front Endocrinol (Lausanne), 2013. **4**: p. 5.
23. Joedicke, L., et al., *The molecular basis of subtype selectivity of human kinin G-protein-coupled receptors.* Nat Chem Biol, 2018. **14**(3): p. 284-290.
24. Cherezov, V., et al., *High-resolution crystal structure of an engineered human beta2-adrenergic G protein-coupled receptor.* Science, 2007. **318**(5854): p. 1258-65.
25. Wu, B., et al., *Structures of the CXCR4 chemokine GPCR with small-molecule and cyclic peptide antagonists.* Science, 2010. **330**(6007): p. 1066-71.
26. Chien, E.Y., et al., *Structure of the human dopamine D3 receptor in complex with a D2/D3 selective antagonist.* Science, 2010. **330**(6007): p. 1091-5.
27. Wu, F., et al., *Structure and Function of Peptide-Binding G Protein-Coupled Receptors.* J Mol Biol, 2017. **429**(17): p. 2726-2745.
28. Granier, S., et al., *Structure of the delta-opioid receptor bound to naltrindole.* Nature, 2012. **485**(7398): p. 400-4.
29. Wu, H., et al., *Structure of the human κ-opioid receptor in complex with JDTic.* Nature, 2012. **485**(7398): p. 327-32.
30. Manglik, A., et al., *Crystal structure of the micro-opioid receptor bound to a morphinan antagonist.* Nature, 2012. **485**(7398): p. 321-6.
31. Thompson, A.A., et al., *Structure of the nociceptin/orphanin FQ receptor in complex with a peptide mimetic.* Nature, 2012. **485**(7398): p. 395-9.
32. Zhang, C., et al., *High-resolution crystal structure of human protease-activated receptor 1.* Nature, 2012. **492**(7429): p. 387-92.
33. White, J.F., et al., *Structure of the agonist-bound neurotensin receptor.* Nature, 2012. **490**(7421): p. 508-13.
34. Zheng, Y., et al., *Structure of CC chemokine receptor 2 with orthosteric and allosteric antagonists.* Nature, 2016. **540**(7633): p. 458-461.
35. Tan, Q., et al., *Structure of the CCR5 chemokine receptor-HIV entry inhibitor maraviroc complex.* Science, 2013. **341**(6152): p. 1387-90.
36. Oswald, C., et al., *Intracellular allosteric antagonism of the CCR9 receptor.* Nature, 2016. **540**(7633): p. 462-465.
37. Burg, J.S., et al., *Structural biology. Structural basis for chemokine recognition and activation of a viral G protein-coupled receptor.* Science, 2015. **347**(6226): p. 1113-7.
38. Yin, J., et al., *Structure and ligand-binding mechanism of the human OX1 and OX2 orexin receptors.* Nat Struct Mol Biol, 2016. **23**(4): p. 293-9.
39. Yin, J., et al., *Crystal structure of the human OX2 orexin receptor bound to the insomnia drug suvorexant.* Nature, 2015. **519**(7542): p. 247-50.

40. Zhang, H., et al., *Structural Basis for Ligand Recognition and Functional Selectivity at Angiotensin Receptor.* J Biol Chem, 2015. **290**(49): p. 29127-39.
41. Asada, H., et al., *Crystal structure of the human angiotensin II type 2 receptor bound to an angiotensin II analog.* Nat Struct Mol Biol, 2018. **25**(7): p. 570-576.
42. Cheng, R.K.Y., et al., *Structural insight into allosteric modulation of protease-activated receptor 2.* Nature, 2017. **545**(7652): p. 112-115.
43. Shihoya, W., et al., *Activation mechanism of endothelin ETB receptor by endothelin-1.* Nature, 2016. **537**(7620): p. 363-368.
44. Yang, Z., et al., *Structural basis of ligand binding modes at the neuropeptide Y Y1 receptor.* Nature, 2018. **556**(7702): p. 520-524.
45. Robertson, N., et al., *Structure of the complement C5a receptor bound to the extra-helical antagonist NDT9513727.* Nature, 2018. **553**(7686): p. 111-114.
46. Qin, L., et al., *Structural biology. Crystal structure of the chemokine receptor CXCR4 in complex with a viral chemokine.* Science, 2015. **347**(6226): p. 1117-22.
47. Zheng, Y., et al., *Structure of CC Chemokine Receptor 5 with a Potent Chemokine Antagonist Reveals Mechanisms of Chemokine Recognition and Molecular Mimicry by HIV.* Immunity, 2017. **46**(6): p. 1005-1017 e5.
48. Ma, Y., et al., *Structural Basis for Apelin Control of the Human Apelin Receptor.* Structure, 2017. **25**(6): p. 858-866 e4.
49. Koehl, A., et al., *Structure of the micro-opioid receptor-Gi protein complex.* Nature, 2018. **558**(7711): p. 547-552.
50. Liu, H., et al., *Orthosteric and allosteric action of the C5a receptor antagonists.* Nat Struct Mol Biol, 2018. **25**(6): p. 472-481.
51. Da Costa, G., et al., *Intermolecular interactions between the neurotensin and the third extracellular loop of human neurotensin 1 receptor.* J Biomol Struct Dyn, 2013. **31**(12): p. 1381-92.
52. Lange, O.F., et al., *Recognition dynamics up to microseconds revealed from an RDC-derived ubiquitin ensemble in solution.* Science, 2008. **320**(5882): p. 1471-5.
53. Tyndall, J.D., et al., *Over one hundred peptide-activated G protein-coupled receptors recognize ligands with turn structure.* Chem Rev, 2005. **105**(3): p. 793-826.
54. Tyndall, J.D., T. Nall, and D.P. Fairlie, *Proteases universally recognize beta strands in their active sites.* Chem Rev, 2005. **105**(3): p. 973-99.
55. Siligardi, G. and A.F. Drake, *The importance of extended conformations and, in particular, the PII conformation for the molecular recognition of peptides.* Biopolymers, 1995. **37**(4): p. 281-92.
56. Carraway, R. and S.E. Leeman, *The amino acid sequence of a hypothalamic peptide, neurotensin.* J Biol Chem, 1975. **250**(5): p. 1907-11.
57. Henry, J.A., et al., *A structure-affinity study of the amino acid side-chains in neurotensin : N and C terminal deletions and Ala-scan.* Bioorganic & Medicinal Chemistry Letters, 1993. **3**(5): p. 949-952.
58. Xu, G.Y. and C.M. Deber, *Conformations of neurotensin in solution and in membrane environments studied by 2-D NMR spectroscopy.* Int J Pept Protein Res, 1991. **37**(6): p. 528-35.
59. Williamson, P.T., et al., *Probing the environment of neurotensin whilst bound to the neurotensin receptor by solid state NMR.* FEBS Lett, 2002. **518**(1-3): p. 111-5.

60. Luca, S., et al., *The conformation of neurotensin bound to its G protein-coupled receptor.* Proc Natl Acad Sci U S A, 2003. **100**(19): p. 10706-11.
61. Heise, H., et al., *Probing conformational disorder in neurotensin by two-dimensional solid-state NMR and comparison to molecular dynamics simulations.* Biophys J, 2005. **89**(3): p. 2113-20.
62. Coutant, J., et al., *NMR solution structure of neurotensin in membrane-mimetic environments: molecular basis for neurotensin receptor recognition.* Biochemistry, 2007. **46**(19): p. 5656-63.
63. Khatun, U.L., S.K. Goswami, and C. Mukhopadhyay, *Modulation of the neurotensin solution structure in the presence of ganglioside GM1 bicelle.* Biophys Chem, 2012. **168-169**: p. 48-59.
64. Krumm, B.E., et al., *Structure and dynamics of a constitutively active neurotensin receptor.* Sci Rep, 2016. **6**: p. 38564.
65. Krumm, B.E., et al., *Structural prerequisites for G-protein activation by the neurotensin receptor.* Nat Commun, 2015. **6**: p. 7895.
66. Bittermann, H., et al., *Evaluation of lactam-bridged neurotensin analogues adjusting psi(Pro10) close to the experimentally derived bioactive conformation of NT(8-13).* J Med Chem, 2004. **47**(22): p. 5587-90.
67. Van Kemmel, F.M., et al., *A C-terminal cyclic 8-13 neurotensin fragment analog appears less exposed to neprilysin when it crosses the blood-brain barrier than the cerebrospinal fluid-brain barrier in mice.* Neurosci Lett, 1996. **217**(1): p. 58-60.
68. Sousbie, M., et al., *In Search of the Optimal Macrocyclization Site for Neurotensin.* ACS Med Chem Lett, 2018. **9**(3): p. 227-232.
69. Chandrasekaran, B., O. Dar, and T. McDonagh, *The role of apelin in cardiovascular function and heart failure.* Eur J Heart Fail, 2008. **10**(8): p. 725-32.
70. Goidescu, C.M. and L.A. Vida-Simiti, *The Apelin-APJ System in the Evolution of Heart Failure.* Clujul Med, 2015. **88**(1): p. 3-8.
71. O'Carroll, A.M., et al., *The apelin receptor APJ: journey from an orphan to a multifaceted regulator of homeostasis.* J Endocrinol, 2013. **219**(1): p. R13-35.
72. Zou, M.X., et al., *Apelin peptides block the entry of human immunodeficiency virus (HIV).* FEBS Lett, 2000. **473**(1): p. 15-8.
73. Fan, X., et al., *Structural and functional study of the apelin-13 peptide, an endogenous ligand of the HIV-1 coreceptor, APJ.* Biochemistry, 2003. **42**(34): p. 10163-8.
74. Medhurst, A.D., et al., *Pharmacological and immunohistochemical characterization of the APJ receptor and its endogenous ligand apelin.* J Neurochem, 2003. **84**(5): p. 1162-72.
75. Murza, A., et al., *Elucidation of the structure-activity relationships of apelin: influence of unnatural amino acids on binding, signaling, and plasma stability.* ChemMedChem, 2012. **7**(2): p. 318-25.
76. Langelaan, D.N., et al., *Structural insight into G-protein coupled receptor binding by apelin.* Biochemistry, 2009. **48**(3): p. 537-48.
77. Yanagisawa, M., et al., *A novel potent vasoconstrictor peptide produced by vascular endothelial cells.* Nature, 1988. **332**(6163): p. 411-5.
78. Kohan, D.E., et al., *Regulation of blood pressure and salt homeostasis by endothelin.* Physiol Rev, 2011. **91**(1): p. 1-77.

79. Davenport, A.P., *International Union of Pharmacology. XXIX. Update on endothelin receptor nomenclature.* Pharmacol Rev, 2002. **54**(2): p. 219-26.
80. Kirkby, N.S., et al., *The endothelin system as a therapeutic target in cardiovascular disease: great expectations or bleak house?* Br J Pharmacol, 2008. **153**(6): p. 1105-19.
81. Andersen, N.H., et al., *Conformational isomerism of endothelin in acidic aqueous media: a quantitative NOESY analysis.* Biochemistry, 1992. **31**(5): p. 1280-95.
82. Janes, R.W., D.H. Peapus, and B.A. Wallace, *The crystal structure of human endothelin.* Nat Struct Biol, 1994. **1**(5): p. 311-9.
83. Aumelas, A., et al., *[Lys(-2)-Arg(-1)]endothelin-1 solution structure by two-dimensional 1H-NMR: possible involvement of electrostatic interactions in native disulfide bridge formation and in biological activity decrease.* Biochemistry, 1995. **34**(14): p. 4546-61.
84. Atkins, A.R., R.C. Martin, and R. Smith, *1H NMR studies of sarafotoxin SRTb, a nonselective endothelin receptor agonist, and IRL 1620, an ETB receptor-specific agonist.* Biochemistry, 1995. **34**(6): p. 2026-33.
85. Takashima, H., et al., *Hydrophobic core around tyrosine for human endothelin-1 investigated by photochemically induced dynamic nuclear polarization nuclear magnetic resonance and matrix-assisted laser desorption ionization time-of-flight mass spectrometry.* Biochemistry, 2004. **43**(44): p. 13932-6.
86. Hoh, F., et al., *High-resolution X-ray structure of the unexpectedly stable dimer of the [Lys(-2)-Arg(-1)-des(17-21)]endothelin-1 peptide.* Biochemistry, 2004. **43**(48): p. 15154-68.
87. Hewage, C.M., et al., *Solution structure of a novel ETB receptor selective agonist ET1-21 [Cys(Acm)1,15, Aib3,11, Leu7] by nuclear magnetic resonance spectroscopy and molecular modelling.* J Pept Res, 1999. **53**(3): p. 223-33.
88. Lattig, J., et al., *Structural determinants for selective recognition of peptide ligands for endothelin receptor subtypes ETA and ETB.* J Pept Sci, 2009. **15**(7): p. 479-91.
89. Hilal-Dandan, R., et al., *The quasi-irreversible nature of endothelin binding and G protein-linked signaling in cardiac myocytes.* J Pharmacol Exp Ther, 1997. **281**(1): p. 267-73.
90. Takasuka, T., et al., *Human endothelin receptor ETB. Amino acid sequence requirements for super stable complex formation with its ligand.* J Biol Chem, 1994. **269**(10): p. 7509-13.
91. Huber, R., et al., *Crystal structure analysis and molecular model of human C3a anaphylatoxin.* Hoppe Seylers Z Physiol Chem, 1980. **361**(9): p. 1389-99.
92. Nettesheim, D.G., et al., *Secondary structure of complement component C3a anaphylatoxin in solution as determined by NMR spectroscopy: differences between crystal and solution conformations.* Proc Natl Acad Sci U S A, 1988. **85**(14): p. 5036-40.
93. Zuiderweg, E.R., et al., *Tertiary structure of human complement component C5a in solution from nuclear magnetic resonance data.* Biochemistry, 1989. **28**(1): p. 172-85.
94. Williamson, M.P. and V.S. Madison, *Three-dimensional structure of porcine C5adesArg from 1H nuclear magnetic resonance data.* Biochemistry, 1990. **29**(12): p. 2895-905.

95. Fredslund, F., et al., *Structure of and influence of a tick complement inhibitor on human complement component 5.* Nat Immunol, 2008. **9**(7): p. 753-60.
96. Laursen, N.S., et al., *Substrate recognition by complement convertases revealed in the C5-cobra venom factor complex.* EMBO J, 2011. **30**(3): p. 606-16.
97. Bajic, G., et al., *Human C3a and C3a desArg anaphylatoxins have conserved structures, in contrast to C5a and C5a desArg.* Protein Sci, 2013. **22**(2): p. 204-12.
98. Laursen, N.S., et al., *Structural basis for inhibition of complement C5 by the SSL7 protein from Staphylococcus aureus.* Proc Natl Acad Sci U S A, 2010. **107**(8): p. 3681-6.
99. Cook, W.J., et al., *Structure of human desArg-C5a.* Acta Crystallogr D Biol Crystallogr, 2010. **66**(Pt 2): p. 190-7.
100. Schatz-Jakobsen, J.A., et al., *Structural and functional characterization of human and murine C5a anaphylatoxins.* Acta Crystallogr D Biol Crystallogr, 2014. **70**(Pt 6): p. 1704-17.
101. Zhang, X., et al., *Structural definition of the C5a C terminus by two-dimensional nuclear magnetic resonance spectroscopy.* Proteins, 1997. **28**(2): p. 261-7.
102. Siciliano, S.J., et al., *Two-site binding of C5a by its receptor: an alternative binding paradigm for G protein-coupled receptors.* Proc Natl Acad Sci U S A, 1994. **91**(4): p. 1214-8.
103. Finch, A.M., et al., *Biologically active conformer of the effector region of human C5a and modulatory effects of N-terminal receptor binding determinants on activity.* J Med Chem, 1997. **40**(6): p. 877-84.
104. Vogen, S.M., et al., *NMR analysis of a potent decapeptide agonist of human C5a anaphylatoxin.* J Pept Res, 1998. **51**(3): p. 226-34.
105. Vogen, S.M., et al., *Determination of structural elements related to the biological activities of a potent decapeptide agonist of human C5a anaphylatoxin.* J Pept Res, 1999. **54**(1): p. 74-84.
106. Wong, A.K., et al., *Small molecular probes for G-protein-coupled C5a receptors: conformationally constrained antagonists derived from the C terminus of the human plasma protein C5a.* J Med Chem, 1998. **41**(18): p. 3417-25.
107. Finch, A.M., et al., *Low-molecular-weight peptidic and cyclic antagonists of the receptor for the complement factor C5a.* J Med Chem, 1999. **42**(11): p. 1965-74.
108. March, D.R., et al., *Potent cyclic antagonists of the complement C5a receptor on human polymorphonuclear leukocytes. Relationships between structures and activity.* Mol Pharmacol, 2004. **65**(4): p. 868-79.
109. Gerber, B.O., et al., *An activation switch in the ligand binding pocket of the C5a receptor.* J Biol Chem, 2001. **276**(5): p. 3394-400.
110. Nikiforovich, G.V., G.R. Marshall, and T.J. Baranski, *Modeling molecular mechanisms of binding of the anaphylatoxin C5a to the C5a receptor.* Biochemistry, 2008. **47**(10): p. 3117-30.
111. Rana, S. and A.R. Sahoo, *Model structures of inactive and peptide agonist bound C5aR: Insights into agonist binding, selectivity and activation.* Biochem Biophys Rep, 2015. **1**: p. 85-96.
112. Sahoo, A.R., R. Mishra, and S. Rana, *The Model Structures of the Complement Component 5a Receptor (C5aR) Bound to the Native and Engineered (h)C5a.* Sci Rep, 2018. **8**(1): p. 2955.
113. Muller, T.D., et al., *Ghrelin.* Mol Metab, 2015. **4**(6): p. 437-60.

114. Bowers, C.Y., et al., *Structure-activity relationships of a synthetic pentapeptide that specifically releases growth hormone in vitro.* Endocrinology, 1980. **106**(3): p. 663-7.
115. Howard, A.D., et al., *A receptor in pituitary and hypothalamus that functions in growth hormone release.* Science, 1996. **273**(5277): p. 974-7.
116. Kojima, M., et al., *Ghrelin is a growth-hormone-releasing acylated peptide from stomach.* Nature, 1999. **402**(6762): p. 656-60.
117. Bednarek, M.A., et al., *Structure-function studies on the new growth hormone-releasing peptide, ghrelin: minimal sequence of ghrelin necessary for activation of growth hormone secretagogue receptor 1a.* J Med Chem, 2000. **43**(23): p. 4370-6.
118. Matsumoto, M., et al., *Structure-activity relationship of ghrelin: pharmacological study of ghrelin peptides.* Biochem Biophys Res Commun, 2001. **287**(1): p. 142-6.
119. Matsumoto, M., et al., *Structural similarity of ghrelin derivatives to peptidyl growth hormone secretagogues.* Biochem Biophys Res Commun, 2001. **284**(3): p. 655-9.
120. Holst, B., et al., *High constitutive signaling of the ghrelin receptor--identification of a potent inverse agonist.* Mol Endocrinol, 2003. **17**(11): p. 2201-10.
121. Holst, B., et al., *Nonpeptide and peptide growth hormone secretagogues act both as ghrelin receptor agonist and as positive or negative allosteric modulators of ghrelin signaling.* Mol Endocrinol, 2005. **19**(9): p. 2400-11.
122. Holst, B., et al., *Ghrelin receptor inverse agonists: identification of an active peptide core and its interaction epitopes on the receptor.* Mol Pharmacol, 2006. **70**(3): p. 936-46.
123. Holst, B., et al., *Identification of an efficacy switch region in the ghrelin receptor responsible for interchange between agonism and inverse agonism.* J Biol Chem, 2007. **282**(21): p. 15799-811.
124. Sivertsen, B., et al., *Unique interaction pattern for a functionally biased ghrelin receptor agonist.* J Biol Chem, 2011. **286**(23): p. 20845-60.
125. Els, S., et al., *An aromatic region to induce a switch between agonism and inverse agonism at the ghrelin receptor.* J Med Chem, 2012. **55**(17): p. 7437-49.
126. Maingot, M., et al., *New ligands of the ghrelin receptor based on the 1,2,4-triazole scaffold by introduction of a second chiral center.* Bioorg Med Chem Lett, 2016. **26**(10): p. 2408-2412.
127. Van Craenenbroeck, M., et al., *Ala-scan of ghrelin (1-14): interaction with the recombinant human ghrelin receptor.* Peptides, 2004. **25**(6): p. 959-65.
128. Feighner, S.D., et al., *Structural requirements for the activation of the human growth hormone secretagogue receptor by peptide and nonpeptide secretagogues.* Mol Endocrinol, 1998. **12**(1): p. 137-45.
129. Holst, B., et al., *Overlapping binding site for the endogenous agonist, small-molecule agonists, and ago-allosteric modulators on the ghrelin receptor.* Mol Pharmacol, 2009. **75**(1): p. 44-59.
130. Silva Elipe, M.V., M.A. Bednarek, and Y.D. Gao, *1H NMR structural analysis of human ghrelin and its six truncated analogs.* Biopolymers, 2001. **59**(7): p. 489-501.
131. Martin-Pastor, M., et al., *Interaction between ghrelin and the ghrelin receptor (GHS-R1a), a NMR study using living cells.* Bioorg Med Chem, 2010. **18**(4): p. 1583-90.

132. Staes, E., et al., *Acylated and unacylated ghrelin binding to membranes and to ghrelin receptor: towards a better understanding of the underlying mechanisms.* Biochim Biophys Acta, 2010. **1798**(11): p. 2102-13.
133. De Ricco, R., et al., *Conformation propensities of des-acyl-ghrelin as probed by CD and NMR.* Peptides, 2013. **43**: p. 62-7.
134. Millar, R.P., *GnRHs and GnRH receptors.* Anim Reprod Sci, 2005. **88**(1-2): p. 5-28.
135. Sealfon, S.C., H. Weinstein, and R.P. Millar, *Molecular mechanisms of ligand interaction with the gonadotropin-releasing hormone receptor.* Endocr Rev, 1997. **18**(2): p. 180-205.
136. Chary, K.V., et al., *Molecular conformation of gonadoliberin using two-dimensional NMR spectroscopy.* Eur J Biochem, 1986. **158**(2): p. 323-32.
137. Grant, G. and W. Vale, *Speculations on structural relationships between the hypothalamic releasing factors of pituitary hormones.* Nat New Biol, 1972. **237**(75): p. 182-3.
138. Chang, J.K., et al., *Luteinizing releasing hormone, synthesis and arg 8 -analogs, and conformation-sequence-activity relationships.* Biochem Biophys Res Commun, 1972. **47**(4): p. 727-32.
139. Monahan, M.W., et al., *Synthetic analogs of the hypothalamic luteinizing hormone releasing factor with increased agonist or antagonist properties.* Biochemistry, 1973. **12**(23): p. 4616-20.
140. Pincus, M.R., et al., *The low-energy conformations of gonadotropin-releasing hormone in aqueous solution.* Protein J, 2014. **33**(6): p. 565-74.
141. Pincus, M.R., et al., *Low energy conformations for gonadotropin-releasing hormone with D- and L-amino acid substitutions for Gly 6: possible receptor-bound conformations.* Protein J, 2014. **33**(6): p. 575-87.
142. Fromme, B.J., et al., *Role of aspartate7.32(302) of the human gonadotropin-releasing hormone receptor in stabilizing a high-affinity ligand conformation.* Mol Pharmacol, 2001. **60**(6): p. 1280-7.
143. Maliekal, J.C., et al., *Solution conformations of gonadotropin releasing hormone (GnRH) and [Gln8]GnRH*. Vol. 50. 1997. 217-219.
144. Millar, R.P., et al., *Chimeric analogues of vertebrate gonadotropin-releasing hormones comprising substitutions of the variant amino acids in positions 5, 7, and 8. Characterization of requirements for receptor binding and gonadotropin release in mammalian and avian pituitary gonadotropes.* J Biol Chem, 1989. **264**(35): p. 21007-13.
145. Kumar, P. and A. Sharma, *Gonadotropin-releasing hormone analogs: Understanding advantages and limitations.* J Hum Reprod Sci, 2014. **7**(3): p. 170-4.
146. Andersen, N.H. and P.K. Hammen, *A conformation-preference/potency correlation for GnRH analogs: NMR evidence*. Bioorganic & Medicinal Chemistry Letters, 1991. **1**(5): p. 263-266.
147. Laimou, D.K., et al., *Structural elucidation of Leuprolide and its analogues in solution: insight into their bioactive conformation.* Amino Acids, 2010. **39**(5): p. 1147-60.
148. Rivier, J., et al., *Design of potent cyclic gonadotropin releasing hormone antagonists.* J Med Chem, 1988. **31**(3): p. 677-82.

149. Pappa, E.V., et al., *Structure-activity studies of lGnRH-III through rational amino acid substitution and NMR conformational studies.* Biopolymers, 2012. **98**(6): p. 525-34.
150. Beck-Sickinger, A.G. and G. Jung, *Structure-activity relationships of neuropeptide Y analogues with respect to Y1 and Y2 receptors.* Biopolymers, 1995. **37**(2): p. 123-42.
151. Blundell, T.L., et al., *X-ray analysis (1. 4-A resolution) of avian pancreatic polypeptide: Small globular protein hormone.* Proc Natl Acad Sci U S A, 1981. **78**(7): p. 4175-9.
152. Saudek, V. and J.T. Pelton, *Sequence-specific 1H NMR assignment and secondary structure of neuropeptide Y in aqueous solution.* Biochemistry, 1990. **29**(19): p. 4509-15.
153. Bader, R. and O. Zerbe, *Are hormones from the neuropeptide Y family recognized by their receptors from the membrane-bound state?* Chembiochem, 2005. **6**(9): p. 1520-34.
154. Thomas, L., et al., *Membrane interaction of neuropeptide Y detected by EPR and NMR spectroscopy.* Biochim Biophys Acta, 2005. **1714**(2): p. 103-13.
155. Bader, R., et al., *Structure and dynamics of micelle-bound neuropeptide Y: comparison with unligated NPY and implications for receptor selection.* J Mol Biol, 2001. **305**(2): p. 307-29.
156. Portoghese, P.S., *Bivalent ligands and the message-address concept in the design of selective opioid receptor antagonists.* Trends Pharmacol Sci, 1989. **10**(6): p. 230-5.
157. Deber, C.M. and B.A. Behnam, *Role of membrane lipids in peptide hormone function: binding of enkephalins to micelles.* Proc Natl Acad Sci U S A, 1984. **81**(1): p. 61-5.
158. Jarrell, H.C., et al., *Interaction of opioid peptides with model membranes. A carbon-13 nuclear magnetic study of enkephalin binding to phosphatidylserine.* Biochemistry, 1980. **19**(2): p. 385-90.
159. Milon, A., T. Miyazawa, and T. Higashijima, *Transferred nuclear Overhauser effect analyses of membrane-bound enkephalin analogues by 1H nuclear magnetic resonance: correlation between activities and membrane-bound conformations.* Biochemistry, 1990. **29**(1): p. 65-75.
160. Marcotte, I., et al., *Interaction of the neuropeptide met-enkephalin with zwitterionic and negatively charged bicelles as viewed by 31P and 2H solid-state NMR.* Biophys J, 2003. **85**(1): p. 328-39.
161. D'Alagni, M., et al., *Conformational study of [Met5]enkephalin-Arg-Phe in the presence of phosphatidylserine vesicles.* Eur J Biochem, 1996. **240**(3): p. 540-9.
162. Graham, W.H., E.S. Carter, 2nd, and R.P. Hicks, *Conformational analysis of Met-enkephalin in both aqueous solution and in the presence of sodium dodecyl sulfate micelles using multidimensional NMR and molecular modeling.* Biopolymers, 1992. **32**(12): p. 1755-64.
163. Higashijima, T., et al., *Nuclear-magnetic-resonance study on Met-enkephalin and Met-enkephalinamide. Molecular association and conformation.* Eur J Biochem, 1979. **97**(1): p. 43-57.
164. Marcotte, I., et al., *A multidimensional 1H NMR investigation of the conformation of methionine-enkephalin in fast-tumbling bicelles.* Biophys J, 2004. **86**(3): p. 1587-600.

165. O'Connor, C., et al., *NMR structure and dynamics of the agonist dynorphin peptide bound to the human kappa opioid receptor.* Proc Natl Acad Sci U S A, 2015. **112**(38): p. 11852-7.
166. Ma, Y., et al., *Structural Basis for Apelin Control of the Human Apelin Receptor.* Structure, 2017. **25**(6): p. 858-866.e4.
167. Pal, K., K. Melcher, and H.E. Xu, *Structure and mechanism for recognition of peptide hormones by Class B G-protein-coupled receptors.* Acta Pharmacol Sin, 2012. **33**(3): p. 300-11.
168. Inooka, H., et al., *Pituitary adenylate cyclase activating polypeptide (PACAP) with 27 residues. Conformation determined by 1H NMR and CD spectroscopies and distance geometry in 25% methanol solution.* Int J Pept Protein Res, 1992. **40**(5): p. 456-64.
169. Inooka, H., et al., *Conformation of a peptide ligand bound to its G-protein coupled receptor.* Nat Struct Biol, 2001. **8**(2): p. 161-5.
170. Braun, W., et al., *Conformation of glucagon in a lipid-water interphase by 1H nuclear magnetic resonance.* J Mol Biol, 1983. **169**(4): p. 921-48.
171. Gronenborn, A.M., G. Bovermann, and G.M. Clore, *A 1H-NMR study of the solution conformation of secretin. Resonance assignment and secondary structure.* FEBS Lett, 1987. **215**(1): p. 88-94.
172. Fry, D.C., et al., *Solution structure of an analogue of vasoactive intestinal peptide as determined by two-dimensional NMR and circular dichroism spectroscopies and constrained molecular dynamics.* Biochemistry, 1989. **28**(6): p. 2399-409.
173. Haghjoo, K., et al., *Solution structure of vasoactive intestinal polypeptide (11-28)-NH2, a fragment with analgesic properties.* Pept Res, 1996. **9**(6): p. 327-31.
174. Zhang, Y., et al., *Cryo-EM structure of the activated GLP-1 receptor in complex with a G protein.* Nature, 2017. **546**(7657): p. 248-253.
175. Liang, Y.L., et al., *Cryo-EM structure of the active, Gs-protein complexed, human CGRP receptor.* Nature, 2018. **561**(7724): p. 492-497.
176. Bourgault, S., et al., *Molecular and conformational determinants of pituitary adenylate cyclase-activating polypeptide (PACAP) for activation of the PAC1 receptor.* J Med Chem, 2009. **52**(10): p. 3308-16.
177. Andersson, A. and L. Maler, *NMR solution structure and dynamics of motilin in isotropic phospholipid bicellar solution.* J Biomol NMR, 2002. **24**(2): p. 103-12.
178. Rathmann, D., et al., *Ligand-mimicking receptor variant discloses binding and activation mode of prolactin-releasing peptide.* J Biol Chem, 2012. **287**(38): p. 32181-94.
179. Deluca, S.H., et al., *The activity of prolactin releasing peptide correlates with its helicity.* Biopolymers, 2013. **99**(5): p. 314-25.
180. Lubecka, E.A., et al., *Arginine-, D-arginine-vasopressin, and their inverso analogues in micellar and liposomic models of cell membrane: CD, NMR, and molecular dynamics studies.* Eur Biophys J, 2015. **44**(8): p. 727-43.
181. Haugaard-Kedstrom, L.M., et al., *Solution structure, aggregation behavior, and flexibility of human relaxin-2.* ACS Chem Biol, 2015. **10**(3): p. 891-900.
182. Rosengren, K.J., et al., *Solution structure and novel insights into the determinants of the receptor specificity of human relaxin-3.* J Biol Chem, 2006. **281**(9): p. 5845-51.

183. Grace, C.R., et al., *Three-dimensional consensus structure of sst2-selective somatostatin (SRIF) antagonists by NMR.* Biopolymers, 2008. **89**(12): p. 1077-87.
184. Grace, C.R., et al., *Novel sst2-selective somatostatin agonists. Three-dimensional consensus structure by NMR.* J Med Chem, 2006. **49**(15): p. 4487-96.
185. Grace, C.R., et al., *Somatostatin receptor 1 selective analogues: 4. Three-dimensional consensus structure by NMR.* J Med Chem, 2005. **48**(2): p. 523-33.
186. Grace, C.R., et al., *Novel sst(4)-selective somatostatin (SRIF) agonists. 4. Three-dimensional consensus structure by NMR.* J Med Chem, 2003. **46**(26): p. 5606-18.
187. Denys, L., et al., *Conformational diversity of bradykinin in aqueous solution.* Biochemistry, 1982. **21**(25): p. 6531-6.
188. Bender, B.J., et al., *Protocols for Molecular Modeling with Rosetta3 and RosettaScripts.* Biochemistry, 2016. **55**(34): p. 4748-63.
189. Rohl, C.A., et al., *Protein structure prediction using Rosetta.* Methods Enzymol, 2004. **383**: p. 66-93.
190. Raman, S., et al., *Structure prediction for CASP8 with all-atom refinement using Rosetta.* Proteins, 2009. **77 Suppl 9**: p. 89-99.
191. Ovchinnikov, S., et al., *Large-scale determination of previously unsolved protein structures using evolutionary information.* Elife, 2015. **4**.
192. Combs, S.A., et al., *Small-molecule ligand docking into comparative models with Rosetta.* Nat Protoc, 2013. **8**(7): p. 1277-98.
193. Kuhlman, B., et al., *Design of a novel globular protein fold with atomic-level accuracy.* Science, 2003. **302**(5649): p. 1364-8.
194. Kellogg, E.H., A. Leaver-Fay, and D. Baker, *Role of conformational sampling in computing mutation-induced changes in protein structure and stability.* Proteins, 2011. **79**(3): p. 830-8.
195. Leaver-Fay, A., et al., *A generic program for multistate protein design.* PLoS One, 2011. **6**(7): p. e20937.
196. Borgo, B. and J.J. Havranek, *Automated selection of stabilizing mutations in designed and natural proteins.* Proc Natl Acad Sci U S A, 2012. **109**(5): p. 1494-9.
197. Fleishman, S.J., et al., *Computational design of proteins targeting the conserved stem region of influenza hemagglutinin.* Science, 2011. **332**(6031): p. 816-21.
198. Gray, J.J., et al., *Protein-protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations.* J Mol Biol, 2003. **331**(1): p. 281-99.
199. Chaudhury, S., et al., *Benchmarking and analysis of protein docking performance in Rosetta v3.2.* PLoS One, 2011. **6**(8): p. e22477.
200. Raveh, B., N. London, and O. Schueler-Furman, *Sub-angstrom modeling of complexes between flexible peptides and globular proteins.* Proteins, 2010. **78**(9): p. 2029-40.
201. Davis, I.W. and D. Baker, *RosettaLigand docking with full ligand and receptor flexibility.* J Mol Biol, 2009. **385**(2): p. 381-92.
202. Meiler, J. and D. Baker, *ROSETTALIGAND: protein-small molecule docking with full side-chain flexibility.* Proteins, 2006. **65**(3): p. 538-48.
203. DeLuca, S., K. Khar, and J. Meiler, *Fully Flexible Docking of Medium Sized Ligand Libraries with RosettaLigand.* PLoS One, 2015. **10**(7): p. e0132508.
204. Das, R. and D. Baker, *Automated de novo prediction of native-like RNA tertiary structures.* Proc Natl Acad Sci U S A, 2007. **104**(37): p. 14664-9.

205. Das, R., *Atomic-accuracy prediction of protein loop structures through an RNA-inspired Ansatz.* PLoS One, 2013. **8**(10): p. e74830.
206. Renfrew, P.D., et al., *Incorporation of noncanonical amino acids into Rosetta and use in computational protein-peptide interface design.* PLoS One, 2012. **7**(3): p. e32637.
207. Drew, K., et al., *Adding diverse noncanonical backbones to rosetta: enabling peptidomimetic design.* PLoS One, 2013. **8**(7): p. e67051.
208. Alford, R.F., et al., *An Integrated Framework Advancing Membrane Protein Modeling and Design.* PLoS Comput Biol, 2015. **11**(9): p. e1004398.
209. Andre, I., et al., *Prediction of the structure of symmetrical protein assemblies.* Proc Natl Acad Sci U S A, 2007. **104**(45): p. 17656-61.
210. DiMaio, F., et al., *Modeling symmetric macromolecular structures in Rosetta3.* PLoS One, 2011. **6**(6): p. e20450.
211. Leaver-Fay, A., et al., *ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules.* Methods Enzymol, 2011. **487**: p. 545-74.
212. Rohl, C.A., et al., *Modeling structurally variable regions in homologous proteins with rosetta.* Proteins, 2004. **55**(3): p. 656-77.
213. Dunbrack, R.L. and M. Karplus, *Backbone-dependent rotamer library for proteins. Application to side-chain prediction.* J Mol Biol, 1993. **230**(2): p. 543-74.
214. Kuhlman, B. and D. Baker, *Native protein sequences are close to optimal for their structures.* Proc Natl Acad Sci U S A, 2000. **97**(19): p. 10383-8.
215. Song, Y., et al., *Structure-guided forcefield optimization.* Proteins, 2011. **79**(6): p. 1898-909.
216. Leaver-Fay, A., et al., *Scientific benchmarks for guiding macromolecular energy function improvement.* Methods Enzymol, 2013. **523**: p. 109-43.
217. Shapovalov, M.V. and R.L. Dunbrack, Jr., *A smoothed backbone-dependent rotamer library for proteins derived from adaptive kernel density estimates and regressions.* Structure, 2011. **19**(6): p. 844-58.
218. Bazzoli, A., S.P. Kelow, and J. Karanicolas, *Enhancements to the Rosetta Energy Function Enable Improved Identification of Small Molecules that Inhibit Protein-Protein Interactions.* PLoS One, 2015. **10**(10): p. e0140359.
219. Yarov-Yarovoy, V., J. Schonbrun, and D. Baker, *Multipass membrane protein structure prediction using Rosetta.* Proteins, 2006. **62**(4): p. 1010-25.
220. Simons, K.T., et al., *Assembly of Protein Tertiary Structures from Fragments with Similar Local Sequences using Simulated Annealing and Bayesian Scoring Functions.* J. Mol. Biol., 1997. **268**: p. 209-225.
221. O'Meara, M.J., et al., *A Combined Covalent-Electrostatic Model of Hydrogen Bonding Improves Structure Prediction with Rosetta.* J Chem Theory Comput, 2015. **11**(2): p. 609-622.
222. Fleishman, S.J., et al., *Hotspot-centric de novo design of protein binders.* J Mol Biol, 2011. **413**(5): p. 1047-62.
223. Leaver-Fay, A., et al., *ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules.* Methods Enzymol, 2011. **487**: p. 545-74.
224. Kaufmann, K.W., et al., *Practically useful: what the Rosetta protein modeling suite can do for you.* Biochemistry, 2010. **49**(14): p. 2987-98.
225. Fleishman, S.J., et al., *RosettaScripts: a scripting language interface to the Rosetta macromolecular modeling suite.* PLoS One, 2011. **6**(6): p. e20161.

226. Chaudhury, S., S. Lyskov, and J.J. Gray, *PyRosetta: a script-based interface for implementing molecular modeling algorithms using Rosetta.* Bioinformatics, 2010. **26**(5): p. 689-91.
227. Schenkelberg, C.D. and C. Bystroff, *InteractiveROSETTA: a graphical user interface for the PyRosetta protein modeling suite.* Bioinformatics, 2015. **31**(24): p. 4023-5.
228. Baugh, E.H., et al., *Real-time PyMOL visualization for Rosetta and PyRosetta.* PLoS One, 2011. **6**(8): p. e21931.
229. Thornburg, N.J., et al., *Human antibodies that neutralize respiratory droplet transmissible H5N1 influenza viruses.* J Clin Invest, 2013. **123**(10): p. 4405-9.
230. Jardine, J., et al., *Rational HIV immunogen design to target specific germline B cell receptors.* Science, 2013. **340**(6133): p. 711-6.
231. Schrodinger, LLC, *The PyMOL Molecular Graphics System*. 2015.
232. Cock, P.J., et al., *Biopython: freely available Python tools for computational molecular biology and bioinformatics.* Bioinformatics, 2009. **25**(11): p. 1422-3.
233. Lyskov, S., et al., *Serverification of molecular modeling applications: the Rosetta Online Server that Includes Everyone (ROSIE).* PLoS One, 2013. **8**(5): p. e63906.
234. Kim, D.E., D. Chivian, and D. Baker, *Protein structure prediction and analysis using the Robetta server.* Nucleic Acids Res, 2004. **32**(Web Server issue): p. W526-31.
235. Lyskov, S. and J.J. Gray, *The RosettaDock server for local protein-protein docking.* Nucleic Acids Res, 2008. **36**(Web Server issue): p. W233-8.
236. Liu, Y. and B. Kuhlman, *RosettaDesign server for protein design.* Nucleic Acids Res, 2006. **34**(Web Server issue): p. W235-8.
237. London, N., et al., *Rosetta FlexPepDock web server--high resolution modeling of peptide-protein interactions.* Nucleic Acids Res, 2011. **39**(Web Server issue): p. W249-53.
238. Lauck, F., et al., *RosettaBackrub--a web server for flexible backbone protein structure modeling and design.* Nucleic Acids Res, 2010. **38 Suppl**: p. W569-75.
239. London, N. and O. Schueler-Furman, *FunHunt: model selection based on energy landscape characteristics.* Biochem Soc Trans, 2008. **36**(Pt 6): p. 1418-21.
240. Adolf-Bryfogle, J. and R.L. Dunbrack, Jr., *The PyRosetta Toolkit: a graphical user interface for the Rosetta software suite.* PLoS One, 2013. **8**(7): p. e66856.
241. Shortle, D., K.T. Simons, and D. Baker, *Clustering of low-energy conformations near the native structures of small proteins.* Proc Natl Acad Sci U S A, 1998. **95**(19): p. 11158-62.
242. Lewis, S.M., et al., *Generation of bispecific IgG antibodies by structure-based design of an orthogonal Fab interface.* Nat Biotechnol, 2014. **32**(2): p. 191-8.
243. Kilambi, K.P., et al., *Extending RosettaDock with water, sugar, and pH for prediction of complex structures and affinities for CAPRI rounds 20-27.* Proteins, 2013. **81**(12): p. 2201-9.
244. Sheffler, W. and D. Baker, *RosettaHoles: rapid assessment of protein core packing for structure prediction, refinement, design, and validation.* Protein Sci, 2009. **18**(1): p. 229-39.
245. Gront, D., et al., *Generalized fragment picking in Rosetta: design, protocols and applications.* PLoS One, 2011. **6**(8): p. e23294.

246. Porter, J.R., B.D. Weitzner, and O.F. Lange, *A Framework to Simplify Combined Sampling Strategies in Rosetta.* PLoS One, 2015. **10**(9): p. e0138220.
247. Altschul, S.F., et al., *Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.* Nucleic Acids Res., 1997. **25**: p. 3389-3402.
248. Mandell, D.J., E.A. Coutsias, and T. Kortemme, *Sub-angstrom accuracy in protein loop reconstruction by robotics-inspired conformational sampling.* Nat Methods, 2009. **6**(8): p. 551-2.
249. Guilhot-Gaudeffroy, A., et al., *Protein-RNA complexes and efficient automatic docking: expanding RosettaDock possibilities.* PLoS One, 2014. **9**(9): p. e108928.
250. Kilambi, K.P., K. Reddy, and J.J. Gray, *Protein-protein docking with dynamic residue protonation states.* PLoS Comput Biol, 2014. **10**(12): p. e1004018.
251. Lemmon, G. and J. Meiler, *Rosetta Ligand docking with flexible XML protocols.* Methods Mol Biol, 2012. **819**: p. 143-55.
252. Gowthaman, R., et al., *DARC: Mapping Surface Topography by Ray-Casting for Effective Virtual Screening at Protein Interaction Sites.* J Med Chem, 2015.
253. Lemmon, G. and J. Meiler, *Towards ligand docking including explicit interface water molecules.* PLoS One, 2013. **8**(6): p. e67536.
254. Allison, B., et al., *Computational design of protein-small molecule interfaces.* J Struct Biol, 2014. **185**(2): p. 193-202.
255. Lemmon, G., K. Kaufmann, and J. Meiler, *Prediction of HIV-1 protease/inhibitor affinity using RosettaLigand.* Chem Biol Drug Des, 2012. **79**(6): p. 888-96.
256. Kaufmann, K.W. and J. Meiler, *Using RosettaLigand for Small Molecule Docking into Comparative Models.* PLoS One, 2012. **7**(12): p. e50769.
257. Nguyen, E.D., et al., *Assessment and challenges of ligand docking into comparative models of G-protein coupled receptors.* PLoS One, 2013. **8**(7): p. e67302.
258. Zhang, Z., et al., *Robust and highly accurate automatic NOESY assignment and structure determination with Rosetta.* J Biomol NMR, 2014. **59**(3): p. 135-45.
259. Kahraman, A., et al., *Cross-link guided molecular modeling with ROSETTA.* PLoS One, 2013. **8**(9): p. e73411.
260. Alexander, N.S., et al., *RosettaEPR: Rotamer Library for Spin Label Structure and Dynamics.* PLoS One, 2013. **8**(9): p. e72851.
261. Hirst, S.J., et al., *RosettaEPR: an integrated tool for protein structure determination from sparse EPR data.* J Struct Biol, 2011. **173**(3): p. 506-14.
262. Sgourakis, N.G., et al., *Determination of the structures of symmetric protein oligomers from NMR chemical shifts and residual dipolar couplings.* J Am Chem Soc, 2011. **133**(16): p. 6288-98.
263. Schmitz, C., et al., *Protein structure determination from pseudocontact shifts using ROSETTA.* J Mol Biol, 2012. **416**(5): p. 668-77.
264. Rossi, P., et al., *A hybrid NMR/SAXS-based approach for discriminating oligomeric protein interfaces using Rosetta.* Proteins, 2015. **83**(2): p. 309-17.
265. Wang, R.Y., et al., *De novo protein structure determination from near-atomic-resolution cryo-EM maps.* Nat Methods, 2015. **12**(4): p. 335-8.
266. DiMaio, F., et al., *Improved low-resolution crystallographic refinement with Phenix and Rosetta.* Nat Methods, 2013. **10**(11): p. 1102-4.
267. DiMaio, F., et al., *Atomic-accuracy models from 4.5-A cryo-electron microscopy data with density-guided iterative local refinement.* Nat Methods, 2015. **12**(4): p. 361-5.

268. Zanghellini, A., et al., *New algorithms and an in silico benchmark for computational enzyme design.* Protein Sci, 2006. **15**(12): p. 2785-94.
269. Willis, J.R., et al., *Human germline antibody gene segments encode polyspecific antibodies.* PLoS Comput Biol, 2013. **9**(4): p. e1003045.
270. Sevy, A.M., et al., *Design of Protein Multi-specificity Using an Independent Sequence Search Reduces the Barrier to Low Energy Sequences.* PLoS Comput Biol, 2015. **11**(7): p. e1004300.
271. Barth, P., J. Schonbrun, and D. Baker, *Toward high-resolution prediction and design of transmembrane helical protein structures.* Proc Natl Acad Sci U S A, 2007. **104**(40): p. 15682-7.
272. Sievers, S.A., et al., *Structure-based design of non-natural amino-acid inhibitors of amyloid fibril formation.* Nature, 2011. **475**(7354): p. 96-100.
273. Mills, J.H., et al., *Computational design of an unnatural amino acid dependent metalloprotein with atomic level accuracy.* J Am Chem Soc, 2013. **135**(36): p. 13393-9.
274. Xu, J., et al., *Structure-based non-canonical amino acid design to covalently crosslink an antibody-antigen complex.* J Struct Biol, 2014. **185**(2): p. 215-22.
275. Renfrew, P.D., et al., *A rotamer library to enable modeling and design of peptoid foldamers.* J Am Chem Soc, 2014. **136**(24): p. 8772-82.
276. Butterfoss, G.L., et al., *Conformational preferences of peptide-peptoid hybrid oligomers.* Biopolymers, 2014. **102**(5): p. 369-78.
277. Sliwoski, G., et al., *Computational methods in drug discovery.* Pharmacol Rev, 2014. **66**(1): p. 334-95.
278. Jain, A.N., *Ligand-based structural hypotheses for virtual screening.* Journal of Medicinal Chemistry, 2004. **47**(4): p. 947-61.
279. Willett, P., *Similarity-based approaches to virtual screening.* Biochemical Society Transactions, 2003. **31**: p. 603-606.
280. Butkiewicz, M., et al., *Benchmarking ligand-based virtual High-Throughput Screening with the PubChem database.* Molecules, 2013. **18**(1): p. 735-56.
281. Sousa, S.F., P.A. Fernandes, and M.J. Ramos, *Protein-ligand docking: current status and future challenges.* Proteins, 2006. **65**: p. 15-26.
282. Rastelli, G., et al., *Fast and accurate predictions of binding free energies using MM-PBSA and MM-GBSA.* J Comput Chem, 2010. **31**(4): p. 797-810.
283. Masetti, M., et al., *Exploring complex protein-ligand recognition mechanisms with coarse metadynamics.* J Phys Chem B, 2009. **113**(14): p. 4807-16.
284. Li, Y., et al., *Comparative assessment of scoring functions on an updated benchmark: 2. Evaluation methods and general results.* J Chem Inf Model, 2014. **54**(6): p. 1717-36.
285. Damm-Ganamet, K.L., et al., *CSAR benchmark exercise 2011-2012: evaluation of results from docking and relative ranking of blinded congeneric series.* J Chem Inf Model, 2013. **53**(8): p. 1853-70.
286. Cozza, G., et al., *Identification of ellagic acid as potent inhibitor of protein kinase CK2: a successful example of a virtual screening application.* J Med Chem, 2006. **49**(8): p. 2363-6.
287. Becker, O.M., et al., *An integrated in silico 3D model-driven discovery of a novel, potent, and selective amidosulfonamide 5-HT1A agonist (PRX-00023) for the treatment of anxiety and depression.* J Med Chem, 2006. **49**(11): p. 3116-35.

288. Durrant, J.D., et al., *Novel naphthalene-based inhibitors of Trypanosoma brucei RNA editing ligase 1.* PLoS Negl Trop Dis, 2010. **4**(8): p. e803.
289. Schueler-Furman, O., et al., *Progress in modeling of protein structures and interactions.* Science, 2005. **310**(5748): p. 638-42.
290. Kaufmann, K.W., et al., *Structural determinants of species-selective substrate recognition in human and Drosophila serotonin transporters revealed through computational docking studies.* Proteins, 2009. **74**(3): p. 630-42.
291. Kufareva, I., et al., *Advances in GPCR Modeling Evaluated by the GPCR Dock 2013 Assessment: Meeting New Challenges.* Structure, 2014. **22**(8): p. 1120-39.
292. Jiang, L., et al., *De novo computational design of retro-aldol enzymes.* Science, 2008. **319**(5868): p. 1387-91.
293. Procko, E., et al., *Computational design of a protein-based enzyme inhibitor.* J Mol Biol, 2013. **425**(18): p. 3563-75.
294. Rothlisberger, D., et al., *Kemp elimination catalysts by computational enzyme design.* Nature, 2008. **453**(7192): p. 190-5.
295. Wang, L., et al., *Structural analyses of covalent enzyme-substrate analog complexes reveal strengths and limitations of de novo enzyme design.* J Mol Biol, 2012. **415**(3): p. 615-25.
296. Price, A.J., S. Howard, and B.D. Cons, *Fragment-based drug discovery and its application to challenging drug targets.* Essays Biochem, 2017. **61**(5): p. 475-484.
297. de Kloe, G.E., et al., *Transforming fragments into candidates: small becomes big in medicinal chemistry.* Drug Discov Today, 2009. **14**(13-14): p. 630-46.
298. Siegal, G., E. Ab, and J. Schultz, *Integration of fragment screening and library design.* Drug Discov Today, 2007. **12**(23-24): p. 1032-9.
299. Jhoti, H., et al., *The 'rule of three' for fragment-based drug discovery: where are we now?* Nat Rev Drug Discov, 2013. **12**(8): p. 644-5.
300. Shuker, S.B., et al., *Discovering high-affinity ligands for proteins: SAR by NMR.* Science, 1996. **274**(5292): p. 1531-4.
301. Hajduk, P.J., et al., *High-throughput nuclear magnetic resonance-based screening.* J Med Chem, 1999. **42**(13): p. 2315-7.
302. Jain, A.N., *Surflex: fully automatic flexible molecular docking using a molecular similarity-based search engine.* Journal of Medicinal Chemistry, 2003. **46**(4): p. 499-511.
303. Lipchock, J.M. and J.P. Loria, *1H, 15N and 13C resonance assignment of imidazole glycerol phosphate (IGP) synthase protein HisF from Thermotoga maritima.* Biomolecular NMR Assignments, 2008. **2**(2): p. 219-21.
304. Schanda, P. and B. Brutscher, *Very fast two-dimensional NMR spectroscopy for real-time investigation of dynamic events in proteins on the time scale of seconds.* Journal of the American Chemical Society, 2005. **127**(22): p. 8014-5.
305. Delaglio, F., et al., *NMRPipe: a multidimensional spectral processing system based on UNIX pipes.* J Biomol NMR, 1995. **6**(3): p. 277-93.
306. Lee, W., M. Tonelli, and J.L. Markley, *NMRFAM-SPARKY: enhanced software for biomolecular NMR spectroscopy.* Bioinformatics, 2015. **31**(8): p. 1325-7.
307. Wu, B., et al., *HTS by NMR of combinatorial libraries: a fragment-based approach to ligand discovery.* Chem Biol, 2013. **20**(1): p. 19-33.
308. Farmer, B.T., 2nd, et al., *Localizing the NADP+ binding site on the MurB enzyme by NMR.* Nat Struct Biol, 1996. **3**(12): p. 995-7.

309. Fielding, L., *NMR methods for the determination of protein-ligand dissociation constants.* Current Topics in Medicinal Chemistry, 2003. **3**(1): p. 39-53.
310. Lemmon, G. and J. Meiler, *Rosetta Ligand docking with flexible XML protocols.* Methods in Molecular Biology, 2012. **819**: p. 143-55.
311. Sing, T., et al., *ROCR: visualizing classifier performance in R.* Bioinformatics, 2005. **21**(20): p. 3940-1.
312. Lang, D., et al., *Structural evidence for evolution of the beta/alpha barrel scaffold by gene duplication and fusion.* Science, 2000. **289**(5484): p. 1546-50.
313. Harner, M.J., A.O. Frank, and S.W. Fesik, *Fragment-based drug discovery using NMR spectroscopy.* J Biomol NMR, 2013. **56**(2): p. 65-75.
314. Congreve, M., et al., *A 'rule of three' for fragment-based lead discovery?* Drug Discov Today, 2003. **8**(19): p. 876-7.
315. Williamson, M.P., *Using chemical shift perturbation to characterise ligand binding.* Progress in Nuclear Magnetic Resonance Spectroscopy, 2013. **73**: p. 1-16.
316. Fu, D.Y. and J. Meiler, *Predictive Power of Different Types of Experimental Restraints in Small Molecule Docking: A Review.* J Chem Inf Model, 2018. **58**(2): p. 225-233.
317. Waszkowycz, B., D.E. Clark, and E. Gancia *Outstanding challenges in protein-ligand docking and structure-based virtual screening*. Wiley Interdisciplinary Reviews: Computational Molecular Science, 2011. **1**, 229-259.
318. Chothia, C., et al., *Evolution of the protein repertoire.* Science, 2003. **300**(5626): p. 1701-3.
319. Höcker, B., *Design of proteins from smaller fragments-learning from evolution.* Current Opinion in Structural Biology, 2014. **27**: p. 56-62.
320. Fortenberry, C., et al., *Exploring symmetry as an avenue to the computational design of large protein domains.* Journal of the American Chemical Society, 2011. **133**(45): p. 18026-9.
321. Douangamath, A., et al., *Structural evidence for ammonia tunneling across the (beta alpha)(8) barrel of the imidazole glycerol phosphate synthase bienzyme complex.* Structure, 2002. **10**(2): p. 185-93.
322. Copley, R.R. and P. Bork, *Homology among (betaalpha)(8) barrels: implications for the evolution of metabolic pathways.* Journal of Molecular Biology, 2000. **303**(4): p. 627-41.
323. Jurgens, C., et al., *Directed evolution of a (beta alpha)8-barrel enzyme to catalyze related reactions in two different metabolic pathways.* Proc Natl Acad Sci U S A, 2000. **97**(18): p. 9925-30.
324. Claren, J., et al., *Establishing wild-type levels of catalytic activity on natural and artificial (beta alpha)8-barrel protein scaffolds.* Proc Natl Acad Sci U S A, 2009. **106**(10): p. 3704-9.
325. Nawaratne, V., et al., *New insights into the function of M4 muscarinic acetylcholine receptors gained using a novel allosteric modulator and a DREADD (designer receptor exclusively activated by a designer drug).* Mol Pharmacol, 2008. **74**(4): p. 1119-31.
326. Siegel, J.B., et al., *Computational design of an enzyme catalyst for a stereoselective bimolecular Diels-Alder reaction.* Science, 2010. **329**(5989): p. 309-13.
327. Feldmeier, K. and B. Höcker, *Computational protein design of ligand binding and catalysis. Current Opinion* in *Chemical Biology*, 2013. **17**(6): p. 929-33.

328. Morin, A., et al., *Computational design of an endo-1,4-beta-xylanase ligand binding site.* Protein Engineering, Design and Selection, 2011. **24**(6): p. 503-16.
329. Nannemann, D.P., et al., *Assessing directed evolution methods for the generation of biosynthetic enzymes with potential in drug biosynthesis.* Future Medicinal Chemistry, 2011. **3**(7): p. 809-19.
330. Kothiwale, S., J.L. Mendenhall, and J. Meiler, *BCL::Conf: small molecule conformational sampling using a knowledge based rotamer library.* J Cheminform, 2015. **7**: p. 47.
331. Henderson, R. and P.N. Unwin, *Three-dimensional model of purple membrane obtained by electron microscopy.* Nature, 1975. **257**(5521): p. 28-32.
332. Schertler, G.F. and P.A. Hargrave, *Projection structure of frog rhodopsin in two crystal forms.* Proc Natl Acad Sci U S A, 1995. **92**(25): p. 11578-82.
333. Unger, V.M. and G.F. Schertler, *Low resolution structure of bovine rhodopsin determined by electron cryo-microscopy.* Biophys J, 1995. **68**(5): p. 1776-86.
334. Palczewski, K., et al., *Crystal structure of rhodopsin: A G protein-coupled receptor.* Science, 2000. **289**(5480): p. 739-45.
335. Rosenbaum, D.M., et al., *GPCR engineering yields high-resolution structural insights into beta2-adrenergic receptor function.* Science, 2007. **318**(5854): p. 1266-73.
336. Bertheleme, N., et al., *Unlocking the secrets of the gatekeeper: methods for stabilizing and crystallizing GPCRs.* Biochim Biophys Acta, 2013. **1828**(11): p. 2583-91.
337. Sandal, M., et al., *GOMoDo: A GPCRs online modeling and docking webserver.* PLoS One, 2013. **8**(9): p. e74092.
338. Esguerra, M., et al., *GPCR-ModSim: A comprehensive web based solution for modeling G-protein coupled receptors.* Nucleic Acids Res, 2016. **44**(W1): p. W455-62.
339. Zhang, J., et al., *GPCR-I-TASSER: A Hybrid Approach to G Protein-Coupled Receptor Structure Modeling and the Application to the Human Genome.* Structure, 2015. **23**(8): p. 1538-1549.
340. Miszta, P., et al., *GPCRM: a homology modeling web service with triple membrane-fitted quality assessment of GPCR models.* Nucleic Acids Res, 2018. **46**(W1): p. W387-W395.
341. Worth, C.L., et al., *GPCR-SSFE 2.0-a fragment-based molecular modeling web tool for Class A G-protein coupled receptors.* Nucleic Acids Res, 2017. **45**(W1): p. W408-W415.
342. Pandy-Szekeres, G., et al., *GPCRdb in 2018: adding GPCR structure models and ligands.* Nucleic Acids Res, 2018. **46**(D1): p. D440-D446.
343. Larsson, P., et al., *Using multiple templates to improve quality of homology models in automated homology modeling.* Protein Sci, 2008. **17**(6): p. 990-1002.
344. Viklund, H. and A. Elofsson, *OCTOPUS: improving topology prediction by two-track ANN-based preference scores and an extended topological grammar.* Bioinformatics, 2008. **24**(15): p. 1662-8.
345. Cummings, D.E., et al., *A preprandial rise in plasma ghrelin levels suggests a role in meal initiation in humans.* Diabetes, 2001. **50**(8): p. 1714-9.

346. Kojima, M., et al., *Ghrelin: discovery of the natural endogenous ligand for the growth hormone secretagogue receptor.* Trends Endocrinol Metab, 2001. **12**(3): p. 118-22.
347. Holst, B., et al., *Common structural basis for constitutive activity of the ghrelin receptor family.* J Biol Chem, 2004. **279**(51): p. 53806-17.
348. Schrottke, S., et al., *Expression, Functional Characterization, and Solid-State NMR Investigation of the G Protein-Coupled GHS Receptor in Bilayer Membranes.* Sci Rep, 2017. **7**: p. 46128.
349. Goze, C., et al., *Involvement of tryptophan W276 and of two surrounding amino acid residues in the high constitutive activity of the ghrelin receptor GHS-R1a.* Eur J Pharmacol, 2010. **643**(2-3): p. 153-61.
350. Holst, B., et al., *A conserved aromatic lock for the tryptophan rotameric switch in TM-VI of seven-transmembrane receptors.* J Biol Chem, 2010. **285**(6): p. 3973-85.
351. Valentin-Hansen, L., et al., *PheVI:09 (Phe6.44) as a sliding microswitch in seven-transmembrane (7TM) G protein-coupled receptor activation.* J Biol Chem, 2012. **287**(52): p. 43516-26.
352. Mokrosinski, J., et al., *Modulation of constitutive activity and signaling bias of the ghrelin receptor by conformational constraint in the second extracellular loop.* J Biol Chem, 2012. **287**(40): p. 33488-502.
353. Liu, G., et al., *Four missense mutations in the ghrelin receptor result in distinct pharmacological abnormalities.* J Pharmacol Exp Ther, 2007. **322**(3): p. 1036-43.
354. Sargent, D.F. and R. Schwyzer, *Membrane lipid phase as catalyst for peptide-receptor interactions.* Proc Natl Acad Sci U S A, 1986. **83**(16): p. 5774-8.
355. Fenalti, G., et al., *Structural basis for bifunctional peptide recognition at human delta-opioid receptor.* Nat Struct Mol Biol, 2015. **22**(3): p. 265-8.
356. Qin, L., et al., *Crystal structure of the chemokine receptor CXCR4 in complex with a viral chemokine.* Science, 2015. **347**(6226): p. 1117-22.
357. Burg, J.S., et al., *Structural basis for chemokine recognition and activation of a viral G protein-coupled receptor.* Science, 2015. **347**(6226): p. 1113-7.
358. Lopez, J.J., et al., *The structure of the neuropeptide bradykinin bound to the human G-protein coupled receptor bradykinin B2 as determined by solid-state NMR spectroscopy.* Angew Chem Int Ed Engl, 2008. **47**(9): p. 1668-71.
359. Torsello, A., et al., *Short ghrelin peptides neither displace ghrelin binding in vitro nor stimulate GH release in vivo.* Endocrinology, 2002. **143**(5): p. 1968-71.
360. Ratnasingham, S. and P.D. Hebert, *bold: The Barcode of Life Data System (http://www.barcodinglife.org).* Mol Ecol Notes, 2007. **7**(3): p. 355-364.
361. Park, M., et al., *Bioorthogonal Labeling of Ghrelin Receptor to Facilitate Studies of Ligand-Dependent Conformational Dynamics.* Chem Biol, 2015. **22**(11): p. 1431-6.
362. Stahlberg, S., et al., *Probing the role of the ceramide acyl chain length and sphingosine unsaturation in model skin barrier lipid mixtures by (2)H solid-state NMR spectroscopy.* Langmuir, 2015. **31**(17): p. 4906-15.
363. Meyer, B. and T. Peters, *NMR spectroscopy techniques for screening and identifying ligand binding to protein receptors.* Angew Chem Int Ed Engl, 2003. **42**(8): p. 864-90.
364. Els, S., A.G. Beck-Sickinger, and C. Chollet, *Ghrelin receptor: high constitutive activity and methods for developing inverse agonists.* Methods Enzymol, 2010. **485**: p. 103-21.

365. Schmidt, P., et al., *Improved in Vitro Folding of the Y2 G Protein-Coupled Receptor into Bicelles.* Front Mol Biosci, 2018. **4**: p. 100.
366. Viegas, A., et al., *Saturation-Transfer Difference (STD) NMR: A Simple and Fast Method for Ligand Screening and Characterization of Protein Binding.* Journal of Chemical Education, 2011. **88**(7): p. 990-994.
367. Ballesteros, J.A. and H. Weinstein, *Integrated methods for the construction of three-dimensional models and computational probing of structure-function relations in Gprotein-coupled receptors*, in *Receptor Molecular Biology*, S.C. Sealfon, Editor. 1995, Academic Press. p. 366–428.
368. Krumm, B.E. and R. Grisshammer, *Peptide ligand recognition by G protein-coupled receptors.* Front Pharmacol, 2015. **6**: p. 48.
369. Holst, B. and T.W. Schwartz, *Ghrelin receptor mutations--too little height and too much hunger.* J Clin Invest, 2006. **116**(3): p. 637-41.
370. Pantel, J., et al., *Loss of constitutive activity of the growth hormone secretagogue receptor in familial short stature.* J Clin Invest, 2006. **116**(3): p. 760-8.
371. Ferré, G., et al., *Structure and Dynamics of the Receptor-Bound Ghrelin Lipopeptide.* Biophysical Journal, 2018. **114**(3): p. 235a.
372. Catoire, L.J., et al., *Structure of a GPCR ligand in its receptor-bound state: leukotriene B4 adopts a highly constrained conformation when associated to human BLT2.* J Am Chem Soc, 2010. **132**(26): p. 9049-57.
373. Kofuku, Y., et al., *Efficacy of the beta(2)-adrenergic receptor is determined by conformational equilibrium in the transmembrane region.* Nat Commun, 2012. **3**: p. 1045.
374. Manglik, A., et al., *Structural Insights into the Dynamic Process of beta2-Adrenergic Receptor Signaling.* Cell, 2015. **161**(5): p. 1101-11.
375. Schafer, C.T. and D.L. Farrens, *Conformational selection and equilibrium governs the ability of retinals to bind opsin.* J Biol Chem, 2015. **290**(7): p. 4304-18.
376. Sounier, R., et al., *Propagation of conformational changes during mu-opioid receptor activation.* Nature, 2015. **524**(7565): p. 375-8.
377. Isogai, S., et al., *Backbone NMR reveals allosteric signal transduction networks in the beta1-adrenergic receptor.* Nature, 2016. **530**(7589): p. 237-41.
378. Ye, L., et al., *Activation of the A2A adenosine G-protein-coupled receptor by conformational selection.* Nature, 2016. **533**(7602): p. 265-8.
379. Zhang, L., M.S. Bijker, and H. Herzog, *The neuropeptide Y system: pathophysiological and therapeutic implications in obesity and cancer.* Pharmacol Ther, 2011. **131**(1): p. 91-113.
380. Morales-Medina, J.C., Y. Dumont, and R. Quirion, *A possible role of neuropeptide Y in depression and stress.* Brain Res, 2010. **1314**: p. 194-205.
381. Michel, M.C., et al., *XVI. International Union of Pharmacology recommendations for the nomenclature of neuropeptide Y, peptide YY, and pancreatic polypeptide receptors.* Pharmacol Rev, 1998. **50**(1): p. 143-50.
382. Yulyaningsih, E., et al., *NPY receptors as potential targets for anti-obesity drug development.* Br. J. Pharmacol., 2011. **163**(6): p. 1170-202.
383. Sousa, D.M., H. Herzog, and M. Lamghari, *NPY signalling pathway in bone homeostasis: Y1 receptor as a potential drug target.* Curr. Drug Targets, 2009. **10**(1): p. 9-19.

384. Antal-Zimanyi, I., et al., *Pharmacological characterization and appetite suppressive properties of BMS-193885, a novel and selective neuropeptide Y_1 receptor antagonist.* Eur. J. Pharmacol., 2008. **590**(1-3): p. 224-32.
385. Rosenbaum, D.M., et al., *GPCR engineering yields high-resolution structural insights into β_2-adrenergic receptor function.* Science, 2007. **318**(5854): p. 1266-73.
386. Roth, C.B., M.A. Hanson, and R.C. Stevens, *Stabilization of the human β_2-adrenergic receptor TM4-TM3-TM5 helix interface by mutagenesis of Glu122$^{3.41}$, a critical residue in GPCR structure.* J. Mol. Biol., 2008. **376**(5): p. 1305-19.
387. Kabsch, W., *Xds.* Acta Crystallogr. D 2010. **66**(Pt 2): p. 125-32.
388. McCoy, A.J., et al., *Phaser crystallographic software.* J. Appl. Crystallogr., 2007. **40**(Pt 4): p. 658-674.
389. Murshudov, G.N., et al., *REFMAC5 for the refinement of macromolecular crystal structures.* Acta Crystallogr. D 2011. **67**(Pt 4): p. 355-67.
390. Smart, O.S., et al., *Exploiting structure similarity in refinement: automated NCS and target-structure restraints in BUSTER.* Acta Crystallogr. D 2012. **68**(Pt 4): p. 368-80.
391. Emsley, P., et al., *Features and development of Coot.* Acta Crystallogr. D, 2010. **66**(Pt 4): p. 486-501.
392. Keller, M., et al., *Mimicking of Arginine by Functionalized N^{ω}-Carbamoylated Arginine As a New Broadly Applicable Approach to Labeled Bioactive Peptides: High Affinity Angiotensin, Neuropeptide Y, Neuropeptide FF, and Neurotensin Receptor Ligands As Examples.* J. Med. Chem., 2016. **59**(5): p. 1925-45.
393. Keller, M., et al., *N^{ω}-Carbamoylation of the Argininamide Moiety: An Avenue to Insurmountable NPY Y_1 Receptor Antagonists and a Radiolabeled Selective High-Affinity Molecular Tool ([^{3}H]UR-MK299) with Extended Residence Time.* J. Med. Chem., 2015. **58**(22): p. 8834-49.
394. Cheng, Y. and W.H. Prusoff, *Relationship between the inhibition constant (K_I) and the concentration of inhibitor which causes 50 per cent inhibition (I_{50}) of an enzymatic reaction.* Biochem. Pharmacol., 1973. **22**(23): p. 3099-108.
395. Burkert, K., et al., *A Deep Hydrophobic Binding Cavity is the Main Interaction for Different Y_2R Antagonists.* ChemMedChem, 2017. **12**(1): p. 75-85.
396. Kostenis, E., *Is $G\alpha_{16}$ the optimal tool for fishing ligands of orphan G-protein-coupled receptors?* Trends. Pharmacol. Sci., 2001. **22**(11): p. 560-4.
397. Pedragosa-Badia, X., et al., *Pancreatic polypeptide is recognized by two hydrophobic domains of the human Y_4 receptor binding pocket.* J. Biol. Chem., 2014. **289**(9): p. 5846-59.
398. Hoffmann, S., et al., *Structure-affinity studies of C-terminally modified analogs of neuropeptide Y led to a novel class of peptidic Y_1 receptor antagonist.* Regul. Pept., 1996. **65**(1): p. 61-70.
399. Kaiser, A., et al., *Unwinding of the C-terminal Residues of Neuropeptide Y is critical for Y_2 Receptor Binding and Activation.* Angew. Chem. Int. Ed. Engl., 2015. **54**(25): p. 7446-9.
400. Gerald, C., et al., *A receptor subtype involved in neuropeptide-Y-induced food intake.* Nature, 1996. **382**(6587): p. 168-71.

401. Schmidt, P., et al., *A reconstitution protocol for the in vitro folded human G protein-coupled Y2 receptor into lipid environment.* Biophys Chem, 2010. **150**(1-3): p. 29-36.
402. Casiraghi, M., et al., *Functional Modulation of a G Protein-Coupled Receptor Conformational Landscape in a Lipid Bilayer.* J Am Chem Soc, 2016. **138**(35): p. 11170-5.
403. Hohwy, M., et al., *Fivefold symmetric homonuclear dipolar recoupling in rotating solids: Application to double quantum spectroscopy.* J. Chem. Phys., 1999. **110**: p. 7983-7992.
404. Schwarz, D., et al., *Preparative scale expression of membrane proteins in Escherichia coli-based continuous exchange cell-free systems.* Nat. Protoc., 2007. **2**(11): p. 2945-57.
405. Bosse, M., et al., *Assessment of a fully active class A G protein-coupled receptor isolated from in vitro folding.* Biochemistry, 2011. **50**(45): p. 9817-25.
406. Wilkins, M.R., et al., *Detailed peptide characterization using PEPTIDEMASS--a World-Wide-Web-accessible tool.* Electrophoresis, 1997. **18**(3-4): p. 403-8.
407. Ballesteros, J.A. and H. Weinstein, *[19] 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,* S. Sealfon, Editor. 1995, Elsevier. p. 366-428.
408. Venkatakrishnan, A.J., et al., *Molecular signatures of G-protein-coupled receptors.* Nature, 2013. **494**(7436): p. 185-94.
409. Rasmussen, S.G., et al., *Crystal structure of the β2 adrenergic receptor-Gs protein complex.* Nature, 2011. **477**(7366): p. 549-55.
410. Standfuss, J., et al., *The structural basis of agonist-induced activation in constitutively active rhodopsin.* Nature, 2011. **471**(7340): p. 656-60.
411. Sautel, M., et al., *Neuropeptide Y and the nonpeptide antagonist BIBP 3226 share an overlapping binding site at the human Y_1 receptor.* Mol. Pharmacol., 1996. **50**(2): p. 285-92.
412. Keller, M., et al., *Guanidine-acylguanidine bioisosteric approach in the design of radioligands: synthesis of a tritium-labeled N^G-propionylargininamide ([^{3}H]-UR-MK114) as a highly potent and selective neuropeptide Y Y_1 receptor antagonist.* J. Med. Chem., 2008. **51**(24): p. 8168-72.
413. Sjodin, P., et al., *Re-evaluation of receptor-ligand interactions of the human neuropeptide Y receptor Y1: a site-directed mutagenesis study.* Biochem. J., 2006. **393**(Pt 1): p. 161-9.
414. Poindexter, G.S., et al., *Dihydropyridine neuropeptide Y Y_1 receptor antagonists.* Bioorg. Med. Chem. Lett., 2002. **12**(3): p. 379-82.
415. Poindexter, G.S., et al., *Dihydropyridine neuropeptide Y Y_1 receptor antagonists 2: bioisosteric urea replacements.* Bioorg. Med. Chem., 2004. **12**(2): p. 507-21.
416. Xu, B., et al., *Mutagenesis and computational modeling of human G-protein-coupled receptor Y2 for neuropeptide Y and peptide YY.* Biochemistry, 2013. **52**(45): p. 7987-98.
417. Lindner, D., et al., *Functional role of the extracellular N-terminal domain of neuropeptide Y subfamily receptors in membrane integration and agonist-stimulated internalization.* Cell Signal., 2009. **21**(1): p. 61-8.

418. Lindert, S., J. Meiler, and J.A. McCammon, *Iterative Molecular Dynamics-Rosetta Protein Structure Refinement Protocol to Improve Model Quality.* J Chem Theory Comput, 2013. **9**(8): p. 3843-3847.
419. Fu, D.Y. and J. Meiler, *RosettaLigandEnsemble: A Small-Molecule Ensemble-Driven Docking Approach.* ACS Omega, 2018. **3**(4): p. 3655-3664.
420. Carlsson, J., et al., *Ligand discovery from a dopamine D3 receptor homology model and crystal structure.* Nat Chem Biol, 2011. **7**(11): p. 769-78.
421. Uguzzoni, G., et al., *Large-scale identification of coevolution signals across homo-oligomeric protein interfaces by direct coupling analysis.* Proc Natl Acad Sci U S A, 2017. **114**(13): p. E2662-E2671.
422. Elphick, M.R., O. Mirabeau, and D. Larhammar, *Evolution of neuropeptide signalling systems.* J Exp Biol, 2018. **221**(Pt 3).
423. Ocampo Daza, D. and D. Larhammar, *Evolution of the growth hormone, prolactin, prolactin 2 and somatolactin family.* Gen Comp Endocrinol, 2018. **264**: p. 94-112.
424. Tostivint, H., et al., *Molecular evolution of GPCRs: Somatostatin/urotensin II receptors.* J Mol Endocrinol, 2014. **52**(3): p. T61-86.
425. Cerda-Reverter, J.M., et al., *Molecular evolution of the neuropeptide Y (NPY) family of peptides: cloning of three NPY-related peptides from the sea bass (Dicentrarchus labrax).* Regul Pept, 2000. **95**(1-3): p. 25-34.
426. Kroeze, W.K., et al., *PRESTO-Tango as an open-source resource for interrogation of the druggable human GPCRome.* Nat Struct Mol Biol, 2015. **22**(5): p. 362-9.
427. van der Lee, M.M., et al., *Pharmacological characterization of receptor redistribution and beta-arrestin recruitment assays for the cannabinoid receptor 1.* J Biomol Screen, 2009. **14**(7): p. 811-23.
428. Heck, M., et al., *Signaling states of rhodopsin. Formation of the storage form, metarhodopsin III, from active metarhodopsin II.* J Biol Chem, 2003. **278**(5): p. 3162-9.
429. Dror, R.O., et al., *SIGNAL TRANSDUCTION. Structural basis for nucleotide exchange in heterotrimeric G proteins.* Science, 2015. **348**(6241): p. 1361-5.
430. Kaya, A.I., et al., *A conserved phenylalanine as a relay between the alpha5 helix and the GDP binding region of heterotrimeric Gi protein alpha subunit.* J Biol Chem, 2014. **289**(35): p. 24475-87.
431. N.S. Alexander, A.M.P., A.I. Kaya, R.A. Stein, H.E. Hamm, J. Meiler, *Energetic analysis of the rhodopsin-G-protein complex links the a5 helix to GDP release.* 2014. **21**(1).
432. Kim, T.H., et al., *The role of ligands on the equilibria between functional states of a G protein-coupled receptor.* J Am Chem Soc, 2013. **135**(25): p. 9465-74.
433. Schneider, B., et al., *Local dynamics of proteins and DNA evaluated from crystallographic B factors.* Acta Crystallogr D Biol Crystallogr, 2014. **70**(Pt 9): p. 2413-9.
434. Wacker, D., et al., *Crystal Structure of an LSD-Bound Human Serotonin Receptor.* Cell, 2017. **168**(3): p. 377-389 e12.
435. Moretti, R., et al., *Rosetta and the Design of Ligand Binding Sites.* Methods Mol Biol, 2016. **1414**: p. 47-62.
436. Leader, B., Q.J. Baca, and D.E. Golan, *Protein therapeutics: a summary and pharmacological classification.* Nat Rev Drug Discov, 2008. **7**(1): p. 21-39.

437. Knudsen, K.E. and H.I. Scher, *Starving the addiction: new opportunities for durable suppression of AR signaling in prostate cancer.* Clin Cancer Res, 2009. **15**(15): p. 4792-8.
438. Baeumner, A.J., *Biosensors for environmental pollutants and food contaminants.* Anal Bioanal Chem, 2003. **377**(3): p. 434-45.
439. Morin, A., et al., *Computational design of an endo-1,4-beta-xylanase ligand binding site.* Protein Eng Des Sel, 2011. **24**(6): p. 503-16.
440. Morin, A., J. Meiler, and L.S. Mizoue, *Computational design of protein-ligand interfaces: potential in therapeutic development.* Trends Biotechnol, 2011. **29**(4): p. 159-66.
441. Jackel, C., P. Kast, and D. Hilvert, *Protein design by directed evolution.* Annu Rev Biophys, 2008. **37**: p. 153-73.
442. Nannemann, D.P., et al., *Assessing directed evolution methods for the generation of biosynthetic enzymes with potential in drug biosynthesis.* Future Med Chem, 2011. **3**(7): p. 809-19.
443. Tinberg, C.E. and S.D. Khare, *Improving Binding Affinity and Selectivity of Computationally Designed Ligand-Binding Proteins Using Experiments.* Methods Mol Biol, 2016. **1414**: p. 155-71.
444. Feldmeier, K. and B. Hocker, *Computational protein design of ligand binding and catalysis.* Curr Opin Chem Biol, 2013. **17**(6): p. 929-33.
445. Kuhlman, B., et al., *Design of a Novel Globular Protein Fold with Atomic Level Accuracy.* Science, 2003. **302**(5649): p. 1364-1368.
446. Koga, N., et al., *Principles for designing ideal protein structures.* Nature, 2012. **491**(7423): p. 222-7.
447. Ashworth, J., et al., *Computational reprogramming of homing endonuclease specificity at multiple adjacent base pairs.* Nucleic Acids Res, 2010. **38**(16): p. 5601-5608.
448. Sammond, D.W., et al., *Computational design of the sequence and structure of a protein-binding peptide.* J Am Chem Soc, 2011. **133**(12): p. 4190-2.
449. Jiang, L., et al., *De novo computational design of retro-aldol enzymes.* Science, 2008. **319**(5868): p. 1387-91.
450. O'Boyle, N.M., et al., *Open Babel: An open chemical toolbox.* J Cheminform, 2011. **3**: p. 33.
451. Allen, F.H., *The Cambridge Structural Database: a quarter of a million crystal structures and rising.* Acta Crystallogr B, 2002. **58**(Pt 3 Pt 1): p. 380-8.
452. Hawkins, P.C., et al., *Conformer generation with OMEGA: algorithm and validation using high quality structures from the Protein Databank and Cambridge Structural Database.* J Chem Inf Model, 2010. **50**(4): p. 572-84.
453. Labute, P., *LowModeMD--implicit low-mode velocity filtering applied to conformational search of macrocycles and protein loops.* Journal of Chemical Information and Modeling, 2010. **50**(5): p. 792-800.
454. Ebejer, J.P., G.M. Morris, and C.M. Deane, *Freely available conformer generation methods: how good are they?* J Chem Inf Model, 2012. **52**(5): p. 1146-58.
455. Nivon, L.G., R. Moretti, and D. Baker, *A Pareto-optimal refinement method for protein design scaffolds.* PLoS One, 2013. **8**(4): p. e59004.
456. Pettersen, E.F., et al., *UCSF Chimera--a visualization system for exploratory research and analysis.* J Comput Chem, 2004. **25**(13): p. 1605-12.

457. Lawrence, M.C. and P.M. Colman, *Shape complementarity at protein/protein interfaces.* J Mol Biol, 1993. **234**(4): p. 946-50.
458. Stranges, P.B. and B. Kuhlman, *A comparison of successful and failed protein interface designs highlights the challenges of designing buried hydrogen bonds.* Protein Sci, 2013. **22**(1): p. 74-82.
459. Nivon, L.G., et al., *Automating human intuition for protein design.* Proteins, 2014. **82**(5): p. 858-66.
460. Roche, D.B. and L.J. McGuffin, *In silico Identification and Characterization of Protein-Ligand Binding Sites.* Methods Mol Biol, 2016. **1414**: p. 1-21.
461. Henrich, S., et al., *Computational approaches to identifying and characterizing protein binding sites for ligand design.* J Mol Recognit, 2010. **23**(2): p. 209-19.
462. Sanchez-Soto, M., et al., *alpha2A- and alpha2C-Adrenoceptors as Potential Targets for Dopamine and Dopamine Receptor Ligands.* Mol Neurobiol, 2018. **55**(11): p. 8438-8454.
463. Alexander, S.P., A. Mathie, and J.A. Peters, *Guide to Receptors and Channels (GRAC), 5th edition.* Br J Pharmacol, 2011. **164 Suppl 1**: p. S1-324.
464. MacDonald, E., B.K. Kobilka, and M. Scheinin, *Gene targeting--homing in on alpha 2-adrenoceptor-subtype function.* Trends Pharmacol Sci, 1997. **18**(6): p. 211-9.
465. Nicholas, A.P., V. Pieribone, and T. Hokfelt, *Distributions of mRNAs for alpha-2 adrenergic receptor subtypes in rat brain: an in situ hybridization study.* J Comp Neurol, 1993. **328**(4): p. 575-94.
466. Scheinin, M., et al., *Distribution of alpha 2-adrenergic receptor subtype gene expression in rat brain.* Brain Res Mol Brain Res, 1994. **21**(1-2): p. 133-49.
467. Fagerholm, V., et al., *Autoradiographic characterization of alpha(2C)-adrenoceptors in the human striatum.* Synapse, 2008. **62**(7): p. 508-15.
468. Lehto, J., et al., *Test-retest reliability of (11)C-ORM-13070 in PET imaging of alpha2C-adrenoceptors in vivo in the human brain.* Eur J Nucl Med Mol Imaging, 2015. **42**(1): p. 120-7.
469. Ordway, G.A., S.M. Jaconetta, and A.E. Halaris, *Characterization of subtypes of alpha-2 adrenoceptors in the human brain.* J Pharmacol Exp Ther, 1993. **264**(2): p. 967-76.
470. Uhlen, S., et al., *Quantification of alpha2A and alpha2C adrenoceptors in the rat striatum and in different regions of the spinal cord.* Acta Physiol Scand, 1997. **160**(4): p. 407-12.
471. Lindvall, O. and A. Bjorklund, *The organization of the ascending catecholamine neuron systems in the rat brain as revealed by the glyoxylic acid fluorescence method.* Acta Physiol Scand Suppl, 1974. **412**: p. 1-48.
472. Swanson, L.W. and B.K. Hartman, *The central adrenergic system. An immunofluorescence study of the location of cell bodies and their efferent connections in the rat utilizing dopamine-beta-hydroxylase as a marker.* J Comp Neurol, 1975. **163**(4): p. 467-505.
473. Gobert, A., et al., *Quantification and pharmacological characterization of dialysate levels of noradrenaline in the striatum of freely-moving rats: release from adrenergic terminals and modulation by alpha2-autoreceptors.* J Neurosci Methods, 2004. **140**(1-2): p. 141-52.

474. Holmberg, M., et al., *Adrenergic alpha2C-receptors reside in rat striatal GABAergic projection neurons: comparison of radioligand binding and immunohistochemistry.* Neuroscience, 1999. **93**(4): p. 1323-33.
475. Hara, M., et al., *Role of adrenoceptors in the regulation of dopamine/DARPP-32 signaling in neostriatal neurons.* J Neurochem, 2010. **113**(4): p. 1046-59.
476. Ihalainen, J.A. and H. Tanila, *In vivo regulation of dopamine and noradrenaline release by alpha2A-adrenoceptors in the mouse nucleus accumbens.* J Neurochem, 2004. **91**(1): p. 49-56.
477. Zhang, W., et al., *alpha2C adrenoceptors inhibit adenylyl cyclase in mouse striatum: potential activation by dopamine.* J Pharmacol Exp Ther, 1999. **289**(3): p. 1286-92.
478. Alachkar, A., J.M. Brotchie, and O.T. Jones, *Binding of dopamine and 3-methoxytyramine as l-DOPA metabolites to human alpha(2)-adrenergic and dopaminergic receptors.* Neurosci Res, 2010. **67**(3): p. 245-9.
479. Cornil, C.A. and G.F. Ball, *Interplay among catecholamine systems: dopamine binds to alpha2-adrenergic receptors in birds and mammals.* J Comp Neurol, 2008. **511**(5): p. 610-27.
480. Sanchez-Soto, M., et al., *Evidence for Noncanonical Neurotransmitter Activation: Norepinephrine as a Dopamine D2-Like Receptor Agonist.* Mol Pharmacol, 2016. **89**(4): p. 457-66.
481. Urizar, E., et al., *CODA-RET reveals functional selectivity as a result of GPCR heteromerization.* Nat Chem Biol, 2011. **7**(9): p. 624-30.
482. Jiang, L.I., et al., *Use of a cAMP BRET sensor to characterize a novel regulation of cAMP by the sphingosine 1-phosphate/G13 pathway.* J Biol Chem, 2007. **282**(14): p. 10576-84.
483. Casado, V., et al., *Old and new ways to calculate the affinity of agonists and antagonists interacting with G-protein-coupled monomeric and dimeric receptors: the receptor-dimer cooperativity index.* Pharmacol Ther, 2007. **116**(3): p. 343-54.
484. Ferre, S., et al., *G protein-coupled receptor oligomerization revisited: functional and pharmacological perspectives.* Pharmacol Rev, 2014. **66**(2): p. 413-34.
485. Casado, V., et al., *Useful pharmacological parameters for G-protein-coupled receptor homodimers obtained from competition experiments. Agonist-antagonist binding modulation.* Biochem Pharmacol, 2009. **78**(12): p. 1456-63.
486. Miller-Gallacher, J.L., et al., *The 2.1 A resolution structure of cyanopindolol-bound beta1-adrenoceptor identifies an intramembrane Na+ ion that stabilises the ligand-free receptor.* PLoS One, 2014. **9**(3): p. e92727.
487. Wang, C., et al., *Structural basis for molecular recognition at serotonin receptors.* Science, 2013. **340**(6132): p. 610-4.
488. Wacker, D., et al., *Structural features for functional selectivity at serotonin receptors.* Science, 2013. **340**(6132): p. 615-9.
489. Meador-Woodruff, J.H., et al., *Dopamine receptor mRNA expression in human striatum and neocortex.* Neuropsychopharmacology, 1996. **15**(1): p. 17-29.
490. Pohjanoksa, K., et al., *Alpha2-adrenoceptor regulation of adenylyl cyclase in CHO cells: dependence on receptor density, receptor subtype and current activity of adenylyl cyclase.* Eur J Pharmacol, 1997. **335**(1): p. 53-63.

491. Kukkonen, J.P., et al., *Ligand- and subtype-selective coupling of human alpha-2 adrenoceptors to Ca++ elevation in Chinese hamster ovary cells.* J Pharmacol Exp Ther, 1998. **287**(2): p. 667-71.
492. Schroder, R., et al., *Applying label-free dynamic mass redistribution technology to frame signaling of G protein-coupled receptors noninvasively in living cells.* Nat Protoc, 2011. **6**(11): p. 1748-60.
493. Atwood, B.K., et al., *Expression of G protein-coupled receptors and related proteins in HEK293, AtT20, BV2, and N18 cell lines as revealed by microarray analysis.* BMC Genomics, 2011. **12**: p. 14.
494. Fraser, C.M., et al., *Cloning, sequence analysis, and permanent expression of a human alpha 2-adrenergic receptor in Chinese hamster ovary cells. Evidence for independent pathways of receptor coupling to adenylate cyclase attenuation and activation.* J Biol Chem, 1989. **264**(20): p. 11754-61.
495. Jones, S.B., S.P. Halenda, and D.B. Bylund, *Alpha 2-adrenergic receptor stimulation of phospholipase A2 and of adenylate cyclase in transfected Chinese hamster ovary cells is mediated by different mechanisms.* Mol Pharmacol, 1991. **39**(2): p. 239-45.
496. Eason, M.G., M.T. Jacinto, and S.B. Liggett, *Contribution of ligand structure to activation of alpha 2-adrenergic receptor subtype coupling to Gs.* Mol Pharmacol, 1994. **45**(4): p. 696-702.
497. Eason, M.G., et al., *Simultaneous coupling of alpha 2-adrenergic receptors to two G-proteins with opposing effects. Subtype-selective coupling of alpha 2C10, alpha 2C4, and alpha 2C2 adrenergic receptors to Gi and Gs.* J Biol Chem, 1992. **267**(22): p. 15795-801.
498. Eason, M.G. and S.B. Liggett, *Identification of a Gs coupling domain in the amino terminus of the third intracellular loop of the alpha 2A-adrenergic receptor. Evidence for distinct structural determinants that confer Gs versus Gi coupling.* J Biol Chem, 1995. **270**(42): p. 24753-60.
499. Ring, A.M., et al., *Adrenaline-activated structure of beta2-adrenoceptor stabilized by an engineered nanobody.* Nature, 2013. **502**(7472): p. 575-579.
500. Feng, Z., T. Hou, and Y. Li, *Selectivity and activation of dopamine D3R from molecular dynamics.* J Mol Model, 2012. **18**(12): p. 5051-63.
501. Owesson-White, C.A., et al., *Sources contributing to the average extracellular concentration of dopamine in the nucleus accumbens.* J Neurochem, 2012. **121**(2): p. 252-62.
502. Rice, M.E., J.C. Patel, and S.J. Cragg, *Dopamine release in the basal ganglia.* Neuroscience, 2011. **198**: p. 112-37.
503. Hersch, S.M., et al., *Electron microscopic analysis of D1 and D2 dopamine receptor proteins in the dorsal striatum and their synaptic relationships with motor corticostriatal afferents.* J Neurosci, 1995. **15**(7 Pt 2): p. 5222-37.
504. Yung, K.K., et al., *Immunocytochemical localization of D1 and D2 dopamine receptors in the basal ganglia of the rat: light and electron microscopy.* Neuroscience, 1995. **65**(3): p. 709-30.
505. Goldman-Rakic, P.S., et al., *The anatomy of dopamine in monkey and human prefrontal cortex.* J Neural Transm Suppl, 1992. **36**: p. 163-77.
506. Castelli, M.P., et al., *alpha2A adrenergic receptors highly expressed in mesoprefrontal dopamine neurons.* Neuroscience, 2016. **332**: p. 130-9.

507. Lee, A., et al., *Localization of alpha2C-adrenergic receptor immunoreactivity in catecholaminergic neurons in the rat central nervous system.* Neuroscience, 1998. **84**(4): p. 1085-96.
508. Brennan, A.R. and A.F. Arnsten, *Neuronal mechanisms underlying attention deficit hyperactivity disorder: the influence of arousal on prefrontal cortical function.* Ann N Y Acad Sci, 2008. **1129**: p. 236-45.
509. Herve, D., *Identification of a specific assembly of the g protein golf as a critical and regulated module of dopamine and adenosine-activated cAMP pathways in the striatum.* Front Neuroanat, 2011. **5**: p. 48.
510. Wedemeyer, M.J., et al., *Modeling the complete chemokine-receptor interaction.* Methods Cell Biol, 2019. **149**: p. 289-314.
511. Kang, Y., et al., *Crystal structure of rhodopsin bound to arrestin by femtosecond X-ray laser.* Nature, 2015. **523**(7562): p. 561-7.
512. Komolov, K.E., et al., *Structural and Functional Analysis of a beta2-Adrenergic Receptor Complex with GRK5.* Cell, 2017. **169**(3): p. 407-421 e16.
513. Abayev, M., et al., *The solution structure of monomeric CCL5 in complex with a doubly sulfated N-terminal segment of CCR5.* FEBS J, 2018. **285**(11): p. 1988-2003.
514. Gustavsson, M., et al., *Structural basis of ligand interaction with atypical chemokine receptor 3.* Nat Commun, 2017. **8**: p. 14135.
515. Ziarek, J.J., et al., *Structural basis for chemokine recognition by a G protein-coupled receptor and implications for receptor activation.* Sci Signal, 2017. **10**(471).
516. Kufareva, I., T.M. Handel, and R. Abagyan, *Experiment-Guided Molecular Modeling of Protein-Protein Complexes Involving GPCRs.* Methods Mol Biol, 2015. **1335**: p. 295-311.
517. Gerard, C. and B.J. Rollins, *Chemokines and disease.* Nat Immunol, 2001. **2**(2): p. 108-15.
518. Sallusto, F. and M. Baggiolini, *Chemokines and leukocyte traffic.* Nat Immunol, 2008. **9**(9): p. 949-52.
519. Rossi, D. and A. Zlotnik, *The biology of chemokines and their receptors.* Annu Rev Immunol, 2000. **18**: p. 217-42.
520. Griffith, J.W., C.L. Sokol, and A.D. Luster, *Chemokines and chemokine receptors: positioning cells for host defense and immunity.* Annu Rev Immunol, 2014. **32**: p. 659-702.
521. Miller, M.C. and K.H. Mayo, *Chemokines from a Structural Perspective.* Int J Mol Sci, 2017. **18**(10).
522. Seo, Y., et al., *Differentiation of CC vs CXC chemokine dimers with GAG octasaccharide binding partners: an ion mobility mass spectrometry approach.* J Am Chem Soc, 2013. **135**(11): p. 4325-32.
523. Monteclaro, F.S. and I.F. Charo, *The amino-terminal extracellular domain of the MCP-1 receptor, but not the RANTES/MIP-1alpha receptor, confers chemokine selectivity. Evidence for a two-step mechanism for MCP-1 receptor activation.* J Biol Chem, 1996. **271**(32): p. 19084-92.
524. Pakianathan, D.R., et al., *Distinct but overlapping epitopes for the interaction of a CC-chemokine with CCR1, CCR3 and CCR5.* Biochemistry, 1997. **36**(32): p. 9642-8.

525. Park, S.H., et al., *Structure of the chemokine receptor CXCR1 in phospholipid bilayers.* Nature, 2012. **491**(7426): p. 779-83.
526. Arimont, M., et al., *Structural Analysis of Chemokine Receptor-Ligand Interactions.* J Med Chem, 2017. **60**(12): p. 4735-4779.
527. Kufareva, I., *Chemokines and their receptors: insights from molecular modeling and crystallography.* Curr Opin Pharmacol, 2016. **30**: p. 27-37.
528. Kleist, A.B., et al., *New paradigms in chemokine receptor signal transduction: Moving beyond the two-site model.* Biochem Pharmacol, 2016. **114**: p. 53-68.
529. Kufareva, I., et al., *What Do Structures Tell Us About Chemokine Receptor Function and Antagonism?* Annu Rev Biophys, 2017. **46**: p. 175-198.
530. Kufareva, I., et al., *Stoichiometry and geometry of the CXC chemokine receptor 4 complex with CXC ligand 12: molecular modeling and experimental validation.* Proc Natl Acad Sci U S A, 2014. **111**(50): p. E5363-72.
531. Stone, M.J., et al., *Mechanisms of Regulation of the Chemokine-Receptor Network.* Int J Mol Sci, 2017. **18**(2).
532. Kryshtafovych, A., et al., *Evaluation of the template-based modeling in CASP12.* Proteins, 2018. **86 Suppl 1**: p. 321-334.
533. Veldkamp, C.T., et al., *Structural basis of CXCR4 sulfotyrosine recognition by the chemokine SDF-1/CXCL12.* Sci Signal, 2008. **1**(37): p. ra4.
534. Schmidt, P., et al., *Improved in Vitro Folding of the Y2 G Protein-Coupled Receptor into Bicelles.* Front Mol Biosci, 2017. **4**: p. 100.
535. Manglik, A., et al., *Structural Insights into the Dynamic Process of beta2-Adrenergic Receptor Signaling.* Cell, 2015. **161**(5): p. 1101-1111.
536. Warne, T., et al., *Structure of a beta1-adrenergic G-protein-coupled receptor.* Nature, 2008. **454**(7203): p. 486-91.
537. Egloff, P., et al., *Structure of signaling-competent neurotensin receptor 1 obtained by directed evolution in Escherichia coli.* Proc Natl Acad Sci U S A, 2014. **111**(6): p. E655-62.
538. Kobilka, B. and G.F. Schertler, *New G-protein-coupled receptor crystal structures: insights and limitations.* Trends Pharmacol Sci, 2008. **29**(2): p. 79-83.
539. Latorraca, N.R., A.J. Venkatakrishnan, and R.O. Dror, *GPCR Dynamics: Structures in Motion.* Chem Rev, 2017. **117**(1): p. 139-155.
540. Liang, Y.L., et al., *Phase-plate cryo-EM structure of a class B GPCR-G-protein complex.* Nature, 2017. **546**(7656): p. 118-123.
541. Warschawski, D.E., et al., *Choosing membrane mimetics for NMR structural studies of transmembrane proteins.* Biochim Biophys Acta, 2011. **1808**(8): p. 1957-74.
542. Goricanec, D., et al., *Conformational dynamics of a G-protein alpha subunit is tightly regulated by nucleotide binding.* Proc Natl Acad Sci U S A, 2016. **113**(26): p. E3629-38.
543. Schmidt, P., et al., *The G-protein-coupled neuropeptide Y receptor type 2 is highly dynamic in lipid membranes as revealed by solid-state NMR spectroscopy.* Chemistry, 2014. **20**(17): p. 4986-92.
544. Vukoti, K., et al., *Stabilization of functional recombinant cannabinoid receptor CB(2) in detergent micelles and lipid bilayers.* PLoS One, 2012. **7**(10): p. e46290.

545. Schlinkmann, K.M. and A. Pluckthun, *Directed evolution of G-protein-coupled receptors for high functional expression and detergent stability.* Methods Enzymol, 2013. **520**: p. 67-97.
546. Park, S.H., et al., *Optimization of purification and refolding of the human chemokine receptor CXCR1 improves the stability of proteoliposomes for structure determination.* Biochim Biophys Acta, 2012. **1818**(3): p. 584-91.
547. Schmidt, P., et al., *Prokaryotic expression, in vitro folding, and molecular pharmacological characterization of the neuropeptide Y receptor type 2.* Biotechnol Prog, 2009. **25**(6): p. 1732-9.
548. Baneres, J.L., J.L. Popot, and B. Mouillac, *New advances in production and functional folding of G-protein-coupled receptors.* Trends Biotechnol, 2011. **29**(7): p. 314-22.
549. Damian, M., et al., *Ghrelin receptor conformational dynamics regulate the transition from a preassembled to an active receptor:Gq complex.* Proc Natl Acad Sci U S A, 2015. **112**(5): p. 1601-6.
550. Parker, S.L. and A. Balasubramaniam, *Neuropeptide Y Y2 receptor in health and disease.* Br J Pharmacol, 2008. **153**(3): p. 420-31.
551. Witte, K., et al., *Oxidative in vitro folding of a cysteine deficient variant of the G protein-coupled neuropeptide Y receptor type 2 improves stability at high concentration.* Biological Chemistry, 2013. **394**(8): p. 1045-1056.
552. Lea, W.A. and A. Simeonov, *Fluorescence polarization assays in small molecule screening.* Expert Opin Drug Discov, 2011. **6**(1): p. 17-32.
553. Alexander, N.S., et al., *Energetic analysis of the rhodopsin-G-protein complex links the alpha5 helix to GDP release.* Nat Struct Mol Biol, 2014. **21**(1): p. 56-63.
554. Medkova, M., et al., *Conformational changes in the amino-terminal helix of the G protein alpha(i1) following dissociation from Gbetagamma subunit and activation.* Biochemistry, 2002. **41**(31): p. 9962-72.
555. Mazzoni, M.R., J.A. Malinski, and H.E. Hamm, *Structural analysis of rod GTP-binding protein, Gt. Limited proteolytic digestion pattern of Gt with four proteases defines monoclonal antibody epitope.* J Biol Chem, 1991. **266**(21): p. 14072-81.
556. Hamm, H.E., et al., *Trp fluorescence reveals an activation-dependent cation-pi interaction in the Switch II region of Galphai proteins.* Protein Sci, 2009. **18**(11): p. 2326-35.
557. Beck-Sickinger, A.G., et al., *Complete L-alanine scan of neuropeptide Y reveals ligands binding to Y1 and Y2 receptors with distinguished conformations.* Eur J Biochem, 1994. **225**(3): p. 947-58.
558. De Angelis, A.A. and S.J. Opella, *Bicelle samples for solid-state NMR of membrane proteins.* Nat Protoc, 2007. **2**(10): p. 2332-8.
559. Marsh, D., *Lateral pressure in membranes.* Biochim Biophys Acta, 1996. **1286**(3): p. 183-223.
560. Rigaud, J.L., et al., *Bio-Beads: an efficient strategy for two-dimensional crystallization of membrane proteins.* J Struct Biol, 1997. **118**(3): p. 226-35.
561. Alexandrov, A.I., et al., *Microscale fluorescent thermal stability assay for membrane proteins.* Structure, 2008. **16**(3): p. 351-9.
562. Han, B., et al., *SHIFTX2: significantly improved protein chemical shift prediction.* J Biomol NMR, 2011. **50**(1): p. 43-57.

563. Wiktor, M., et al., *Biophysical and structural investigation of bacterially expressed and engineered CCR5, a G protein-coupled receptor.* J Biomol NMR, 2013. **55**(1): p. 79-95.
564. Thomas, L., et al., *The dynamics of the G protein-coupled neuropeptide Y2 receptor in monounsaturated membranes investigated by solid-state NMR spectroscopy.* J Biomol NMR, 2015. **61**(3-4): p. 347-59.
565. Baneres, J.L., et al., *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**(4): p. 801-14.
566. Muller, I., et al., *The full-length mu-opioid receptor: a conformational study by circular dichroism in trifluoroethanol and membrane-mimetic environments.* J Membr Biol, 2008. **223**(1): p. 49-57.
567. le Maire, M., P. Champeil, and J.V. Moller, *Interaction of membrane proteins and lipids with solubilizing detergents.* Biochim Biophys Acta, 2000. **1508**(1-2): p. 86-111.
568. Tulumello, D.V. and C.M. Deber, *SDS micelles as a membrane-mimetic environment for transmembrane segments.* Biochemistry, 2009. **48**(51): p. 12096-103.
569. Alvares, R.D., et al., *Effects of a polar amino acid substitution on helix formation and aggregate size along the detergent-induced peptide folding pathway.* Biochim Biophys Acta, 2013. **1828**(2): p. 373-81.
570. Li, J.H., et al., *Ligand-specific changes in M3 muscarinic acetylcholine receptor structure detected by a disulfide scanning strategy.* Biochemistry, 2008. **47**(9): p. 2776-88.
571. Walther, C., et al., *The anterograde transport of the human neuropeptide Y2 receptor is regulated by a subtype specific mechanism mediated by the C-terminus.* Neuropeptides, 2012. **46**(6): p. 335-43.
572. Logez, C., et al., *Detergent-free Isolation of Functional G Protein-Coupled Receptors into Nanometric Lipid Particles.* Biochemistry, 2016. **55**(1): p. 38-48.
573. Ravula, T., et al., *Bioinspired, Size-Tunable Self-Assembly of Polymer-Lipid Bilayer Nanodiscs.* Angew Chem Int Ed Engl, 2017. **56**(38): p. 11466-11470.
574. Son, W.S., et al., *'q-Titration' of long-chain and short-chain lipids differentiates between structured and mobile residues of membrane proteins studied in bicelles by solution NMR spectroscopy.* J Magn Reson, 2012. **214**(1): p. 111-8.
575. Hulme, E.C. and M.A. Trevethick, *Ligand binding assays at equilibrium: validation and interpretation.* Br J Pharmacol, 2010. **161**(6): p. 1219-37.
576. Kunze, G., et al., *Identification of the Glycosaminoglycan Binding Site of Interleukin-10 by NMR Spectroscopy.* J Biol Chem, 2016. **291**(6): p. 3100-13.
577. Castellani, F., et al., *Structure of a protein determined by solid-state magic-angle-spinning NMR spectroscopy.* Nature, 2002. **420**(6911): p. 98-102.
578. Loquet, A., et al., *13C spin dilution for simplified and complete solid-state NMR resonance assignment of insoluble biological assemblies.* J Am Chem Soc, 2011. **133**(13): p. 4722-5.
579. Loquet, A., B. Habenstein, and A. Lange, *Structural investigations of molecular machines by solid-state NMR.* Acc Chem Res, 2013. **46**(9): p. 2070-9.

580. Rues, R.B., et al., *Cell-Free Production of Membrane Proteins in Escherichia coli Lysates for Functional and Structural Studies.* Methods Mol Biol, 2016. **1432**: p. 1-21.