Integrated Structural Biology for Understanding Membrane Proteins: A TSPO Story

By

Yan Xia

Dissertation

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

in

Chemical and Physical Biology

10 Aug 2018

Nashville, Tennessee

Approved:
Charles R. Sanders, Ph.D.
Jarrod Smith, Ph.D.
Michael Stone, Ph.D.
Charles H. Manning, Ph.D.

ACKNOWLEDGEMENT

This work would not have been possible without the support from the following people. First, I would like to thank my mentor Dr. Jens Meiler. The journey to the west is a path paved in struggles and hardship, as described in the famous Chinese novel written in 16th century. In the novel, the monk had to face 81 tribulations before attaining Buddhahood. In a way, the journey towards a ph.D is similar to what was described in that ancient novel. I'm lucky to have been mentored by Jens, inspired by him, encouraged by him when I was low and celebrated with him when things go right. Jens, you have been both a mentor on my scientific advancement, and a mentor outside my professional career, I specially thank you for your encouragement and tolerance.

I would like to thank all my colleagues in and outside the Meiler lab. I start by thanking the people who initially introduced me into the realm of computational and experimental structural biology, Soumya Ganguly and Brian Weiner. Huge thanks goes to Sandeep Kothiwali, Steven Combs, Sam Daluca, Stephanie Daluca in my earlier years in graduate school, and later on Axel Fischer, Rocco Moretti, Bian Li, Xuan Zhang, and Marion Sauer for all the scientific discussions, trouble shootings and food ventures. A friend of mine, who entered the graduate school the same year with me, Amanda Duran, thank you for always helping me figuring out Rosetta symmetry mode and many other questions. A special thank goes to Heather Darling, oh how can we survive one day without your keeping the lab running. I would like to thank many others who I've interacted with professionally or just chatted in the mega lab. I would like to thank all my collaborators, especially

Dewei Tang, Jason Buck and Jun Li from the Manning lab who provided me invaluable materials for my experiments. The interface of computational biology and experimental biology demands interdisciplinary interactions. My work would not have been possible without their open-mindedness in sharing their scientific insights.

I would like to thank my committee and CPB program: Dr. Chuck Sanders, Dr. Al Beth, Dr. Charles Manning, Dr. Mike Stone and Dr. Jarrod Smith for their scientific guidance along the way. I would also thank Dr. Bruce Damon for his leadership in the program and being a truly wonderful DGS for the student.

Another group of people I would have to thank are Dr. Kathy Gould, Dr. Roger Colbran, Dr. Karoly Mirnics, and Dr. Ann Kenworthy. Thank you for bringing me into this journey through the Vanderbilt international scholar program.

Last but not least, I would like to thank my friends and family. My parents who I've been away from for 10 years. I would not be the man I am if not for their support. We've probably only been physically living together for 11 months during the past 10 years, but I know I'm never alone thanks to my parents. I would like to thank my best friends I made during graduate school: **Raaj** Gowrishankar, Mary Lynn Deer and Aparnar Shekar for all the good time spent in Vanderbilt. Lastly I would like to thank Tianyu Zhang and Yaoyue Yao.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

This document summarizes my work on integrated structural biology to study membrane proteins. My focus is utilizing experimental and computational biology tools to study membrane proteins with primary focus on the translocator protein (TSPO). It is divided up into orthogonal publications, with each publication having incorporated some aspects of experimental and computational technologies.

The introduction in chapter I outlines background knowledge in the order of: big picture, methodology, and specific biological problems. I first detail the current state of the structure biology field, the difficulties in structure characterization of membrane proteins, and the rationale to use an integrated structural biology method. I also describe the molecular modeling suites BCL and ROSETTA, which are the primary computational toolkits used in all of my work. Lastly I introduce my model system of TSPO, its role in biological functions, and its potential as therapeutic target.

The description of my research starts in chapter 2. This part of my research aims to develop integrated structure biology tools for membrane protein structure determination. I outline the sequence space and fold space for membrane protein structure determination as a major obstacle for both computational and experimental structural biology. I developed strategies in BCL and ROSETTA to fold membrane proteins using various orthogonal experimental data such as NMR, EPR and EM. The computational structure prediction pipeline developed in this work can circumvent some of the current limitations of structure biology tools and

determine large protein structures to a high accuracy from sparse or low resolution structural data. The folding algorithm developed can be used for determining membrane protein structures for a broader structural biology audience.

Chapter 3 of my thesis details the structural biology investigation on mice TSPO (mTSPO). I set out to determine the structure of TSPO at the early phase of my graduate school when there was no high-resolution structural data available to study TSPO. During this time, structures of TSPO from different sources were experimentally determined by other groups in the field. While controversies exist between the published experimentally determined structures, these structural models along with our own biochemical characterization of TSPO has brought unique opportunity for computational modeling this system. I also address the niche in structural characterization of TSPO, where model systems appear not to provide a true representation of biological activity. By constructing a unified structural model of TSPO based on currently available data, I show that integrating computational modeling with NMR data could further our understanding of crucial ligand-TSPO interactions.

In chapter 4, I describe two cases of application projects. First, I detail the computational ligand docking and protein modeling work to provide additional structural insight for a Class-C GPCR crystal structure. A case where computational ligand docking can accurately predict ligand efficacy and residue contributions to ligand interaction. I also describe the working hypothesis of Class-C GPCR activation. In the case of TSPO ligand study, I briefly describe how NMR spectroscopy could identify differences of ligand binding site and detailed

interactions. I report findings that human rs6971 polymorphism site can be structurally characterized to account for ligand sensitivity.

Finally, chapter 5 details conclusions from current studies and future directions I foresee for these projects. I suggest additional improvement of the computational folding algorithm in the of BCL and Rosetta to utilize additional information such as evolutionary data. I reiterate the current understanding of TSPO structural studies and proposed future directions for TSPO protein characterization. I also detailed experimental conditions failed to produce meaningful data on mTSPO structural characterization.

.

CHAPTER 1 INTRODUCTION

1.1 Current state of structural biology of membrane proteins

Roughly 20-35% of all proteins are integral MPs. While thousands of protein structures that have been determined to high resolution, only about 110 *distinct* MP folds with more than one transmembrane span (TMS) are represented in the PDB (Figure 1) [1, 2]. This is important as knowledge of all distinct MP folds would allow MP structure prediction via comparative modeling. However, analysis of genomes suggests up to 1200 distinct MP folds[3, 4]. At the current rate of MP fold determination, another 120 years would be required to determine the more than 1000 remaining MP folds**.** Hence, experimental determination of the complete fold space of integral MPs remains a formidable challenge for structural biology in the 21st century. As experimental data for MPs with unknown fold are often limited, computational methods gain importance to supplement missing experimental information and to model the MPs fold at atomic-detail resolution. At the same time, challenges to computational protein structure prediction exist because many MPs have no templates for template-based comparative modeling. *De novo* protein structure prediction from protein sequence alone is limited to only small MPs, simulations driven by limited experimental data are required for accurate modeling of large multi-span MPS.

98% of all protein structures currently deposited in the protein data base (PDB) [5] have been determined by X-ray crystallography and nuclear magnetic resonance (NMR) spectroscopy. However, MPs remain difficult targets for both techniques [1, 6]. MPs are often crystallized only after an exhaustive search of

crystallization conditions that far exceeds what is typically necessary for soluble proteins [6] or even evade crystallization completely [7]. The particular challenge for MP crystallization arises as crystals are inherently three-dimensional while MPs naturally assemble into two-dimensional membranes. It remains difficult to provide realistic membrane mimics in the context of a stable three-dimensional crystal although – also at this frontier – progress is made using lipoic cubic phases [8]. However, even if MP structure determination through crystallization is feasible, biological relevance of the resulting models needs to be verified with orthogonal experimental techniques to avoid over-interpretation. Solid- [9] and solution- [10] state NMR have made exceptional advances in investigations of MP structure and dynamics [11-15], but the large size of the MP in its membrane mimic (most often a micelle) continues to be the limiting factor in the applicability of NMR spectroscopy [6, 10, 16].

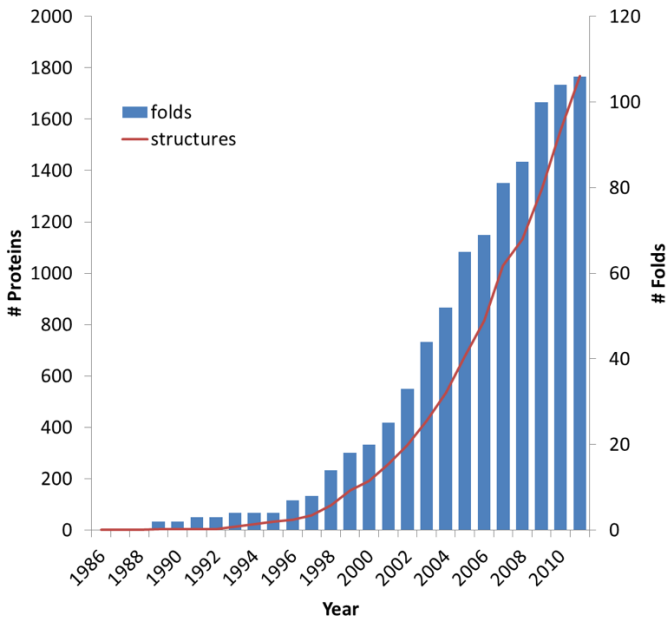


Figure 1 Cumulative number of integral MPs with experimentally determined structure (left y-axis) and total number of unique integral MP folds (right y-axis).

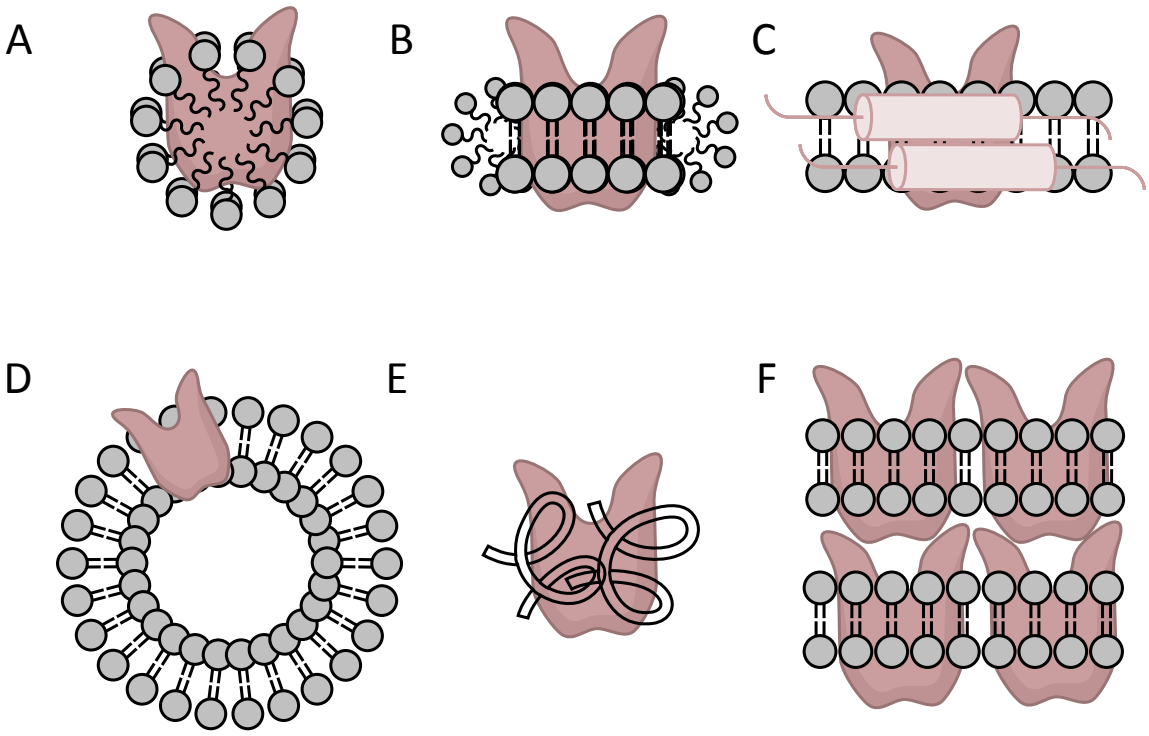


Figure 2 Membrane mimetics for studying membrane proteins. (A) detergent micelle systems; (B) lipid bicelle systems; (C) nanodisc systems; (D) liposome systems; (E) Amphipol systems; (F) crystaline systems.

Experimental MP structural model determination requires rigorous optimization in sample preparation

Structural and biochemical characterization of MPs requires expression and purification of proteins of high quality in sufficient quantities. The experimental technique most used is recombinant protein expression in model systems such as *E.coli cells* [17], yeasts cells [18, 19], insects cells [20] and mammalian cells [21]. While the bacterial recombinant expression system has been most widely used, the MPs of interest are sometimes expressed into the insoluble inclusion bodies or expressed in bacterial membrane with low yield due to cytotoxicity. The MPs in inclusion body require further extraction treatments and re-folding, which is often tedious and resource consuming [22, 23]. Additional tags and fusion proteins need

be introduced into the MPs to improve membrane-targeted expression [24, Tilegenova, 2016 #707, 25]. Additionally, recombinant expression of multi-pass mammalian membrane proteins often requires additional thermos-stabilizing mutations, thus making the method less ideal for investigating native structural functional states [26, 27]. Eukaryotic expression systems have been successful in producing functional membrane embedded MPs, but requires more resources and infrastructure building. More recently, cell-free expression methods have been successfully applied to MPs production [28, 29]. Such platforms offer other opportunities such as non-canonical amino acids labeling and robust lipid-mimetic optimization alone with protein expression [30].

To study membrane protein structure, one would have to utilize one of the many lipid mimetic system to first isolate the protein of interest (Figure 2). While detergent micelle systems can efficiently solubilize membrane proteins, it often fails to maintain the structural integrity due to its high exchange with solvent (Figure 2A). Lipid bicelles or nanodiscs are alternatives to detergent (Figure 2B, C) that have the advantage of maintaining a bilayer lipidic environment. However, these systems are often too large for NMR spectroscopy and requires high expertise in sample preparation. Smaller nanodiscs have been developed specifically for solution state NMR studies of MPs [31]. Liposome particles are often used for spectroscopic characterization of membrane proteins and solid-state NMR (Figure 2D). Amphipol polymers are emerging systems for membrane protein studies [32]. Lastly, crystallization of membrane proteins can be done in many well established systems, but requires a large amount of protein in screening for a

suitable condition. It seems that even with all the model systems available to scientists, optimizing the condition of MPs for a functional stable state is often the limiting step.

Apart from sample preparation, to determine the structure of membrane protein, several other obstacles persist in the pursuit of high quality diffraction patterns from X-ray crystallography to produce high resolution density maps. The same is that for the limited number of structural restraints from complementary experimental techniques such as nuclear magnetic resonance (NMR), site-directed spin labeling (SDSL)-electron paramagnetic resonance (EPR) spectroscopy, or cross-linking coupled with mass spectrometry (XLMS) and imaging contrast for cryo-electron microscopy (cryoEM). As a result, datasets are limited (i.e. they fail to unambiguously define all aspects of the IMP structure at atomic detail).

Computational MP structural model prediction is challenged by the fold space and energy evaluation

Computational MP modeling can successfully complement experimental data by adding atomic-detailed simulations in the MP models where experimental data are limited [33-35]. However, *de novo* prediction of IMP fold from its primary sequence remains a challenging problem [36]. The vast size of the fold space that needs to be sampled from a primary structure is a prohibitive factor in accurate prediction of large IMPs with unknown structure. Computational structural prediction algorithms often suffer from limited coverage of the fold space or approximations of the free energy, hindering accurate sampling of native-like structural models as well as discrimination of inaccurate models. The limited

sampling density results in an intrinsic, minimal deviation of the conformations sampled and the lowest energy conformation that exists in each region of the conformational space further adding to the uncertainty. In addition, the environment of the protein – the cytoplasm or the membrane – is represented in an implicit and static way, adding another layer of inaccuracy in free energy evaluation.

MP structural models must consider biologically relevant structural ensembles

While prediction of the fold of MPs is challenging, modeling the ensemble of the functional states of MPs might represent the next holy grail in computational structure modeling. Within a structurally characterized IMP fold, biologically relevant structural ensembles are observed to deviate from an experimentally determined structural model. Many IMPs function through shifting ensembles of numerous conformations. Biophysical methods that observe ensembles of molecules can be used to complement knowledge from X-ray crystallographic data.

The best examples for dynamic ensemble elucidations are perhaps the GPCRs. GPCRs exist in a dynamic conformational equilibrium of basal, activated, inactivated, G-protein coupling, and arrestin binding states as demonstrated by many recent biophysical studies [34, 37-46]. Solution NMR experiments using ^{19}F probes conjugated to specific amino acids [39] and ^{13}C methionine side chain labels [43] capture β2 adrenergic receptor in distinct population of structural states aside from observed crystal structural models. The shift of the conformational equilibrium of GPCRs during signal transduction is accompanied by major

structural changes, as revealed by SDSL-EPR in β2 adrenergic receptor [37] and rhodopsin [34, 44]. The architecture of the molecular assembly of GPCR and heterotrimeric G-protein and arrestin in a dynamic state could also be visualized using cross-linking mass spectrometry (XLMS) and EM [41]. Using fluorescence resonance transfer experiment, conformational rearrangement in oligomerized state have also been observed in Class-C GPCRs, where the activated receptor dimers undergo conformational change in individual subunit as well as the dimer interface [45, 46]. Molecular dynamics simulations and Monte-Carlo simulations utilizing these experimental data were successful in generating structural models that match the observed conformational states [34, 41, 43].

1.2 BioChemical Library (BCL) and Rosetta molecular modeling suite

BCL is developed for efficient sampling of conformational fold space to complement Rosetta.

The complexity of the conformational space grows exponentially with number of residues in the protein, rendering exhaustive sampling of the conformational space impossible. Protein structure prediction groups have come up with different approaches to address this problem. For example, Rosetta assembles the tertiary structure of proteins by using assembling short fragments collected from the Protein Data Bank (PDB) [47]. This approach substantially reduces the complexity of the sampling space because the dihedral angles are not exhaustively sampled. Rotational conformer libraries provide a similar simplification for the side chain conformations [48]. However, even with the

mentioned simplifications, the size of the conformational space remains too large for many proteins with more than 100 residues. Additionally, previous studies found that de novo prediction with Rosetta has bias towards structures with low contact order [22].

The *de novo* protein structure prediction algorithm BCL::Fold [6] was developed to overcome the aforementioned problems and efficiently predict the topology of larger proteins with up to 400 residues. It was specifically designed to complement Rosetta by predicting secondary structure elements (SSE)-only models with likely topologies of the protein and feeding them into Rosetta for loop and side chain construction as well as high-resolution refinement. The necessary reduction of complex sampling space is achieved by assembling predicted SSEs using a Monte Carlo algorithm and omitting more flexible loop regions. The energy evaluation of the sampled models is conducted using knowledge-based scoring functions [7], which provide a rapid way to approximate the free energy of the sampled conformation. In a previous study, it was demonstrated that BCL::Fold is able to efficiently sample the topologies of larger proteins [8]. Problems in model discrimination, which can arise from necessary simplifications made to sampling, scoring, and system representation, can be compensated for through incorporation of limited experimental data from electron microscopy [9–11], nuclear magnetic resonance spectroscopy [12], electron paramagnetic resonance spectroscopy [13,14], cross-linking experiments [15], small angle X-ray and neutron scattering [16], and predicted residue-residue contacts.

Rosetta is successful in computationally modeling atomic-detailed interactions

Rosetta molecular modeling suite [47] contains a package of community maintained software suite for biomolecule modeling. Being the inspiration for BCL, Rosetta utilizes a knowledge-based potential to evaluate the full atomic energy of biomolecules as well as simplified centroid models [47, 49-52] by a combination of evaluation terms such as van der Waals attractive and repulsive forces, electrostatic interactions, hydrogen bonding, solvation, likelihood of particular side-chain & backbone conformations and implicit membrane environment [47, 49-52]. Similar to BCL::MP-Fold, incorporation of individual experimental restraints into Rosetta been successful for EM [53], NMR [54, 55] and EPR [56] data.

The two particular applications I have extensively used but will not explain in detail in this document are: RosettaCM [50, 57] and RosettaLigand [57-59]. RosettaCM is a homology modeling application that takes multiple template models as input. Conventional homology modeling using single template suffers from limited sampling of conformational space when sequence divergence is large. RosettaCM circumvent such sampling limitation by allowing fragment replacement from multiple models to the target sequence using a Monte Carlo sampling algorithm. Although using the same core energy evaluation scheme in Rosetta, RosettaCM could produce more native-like models. RosettaLigand is the ligand docking algorithm developed for structure based drug discovery. It accounts for both ligand flexibility and protein flexibility during the docking process. In principle, such conformational sampling techniques can improve the chance of finding global energy minimum for ligand conformations.

1.3 The evolving understanding of TSPO proteins in biological processes

Translocator Protein 18 kDa (TSPO) was originally discovered as a secondary binding site for the widely prescribed anti-anxiety drugs, benzodiazepines, in the mitochondria of rat kidney [60]. For this reason, it became known as the peripheral-type benzodiazepine receptor (PBR). In 2006, it was renamed as Translocator Protein 18 kDa (TSPO) [61] to better represent the emerging understanding of a family of highly conserved integral membrane proteins found in Archaea through to plants and human, with a broad spectrum of functions including steroid and porphyrin transport and regulation of apoptosis, inflammation, metastatic cancer and cardiovascular disease [62].

TSPO's biological function is controversially discussed

Primarily located in the outer mitochondrial membrane and concentrated at the outer/inner membrane contact sites in mammalian cells [63], knock-out of TSPO is developmentally lethal in mouse [64]. It is expressed in all organs examined so far and at particularly high levels in steroid hormone producing tissues, such as adrenal glands and kidney. The major function of TSPO has been proposed to be transporting small molecules, including cholesterol and heme biosynthesis intermediates, into or out of mitochondria to be further metabolized [65-67]. In addition, TSPO has been found to be highly expressed in areas of brain injury and inflammation [68], and aggressive cancers [69, 70], as well as brains of Alzheimer and Huntington disease patients [71]. In these situations, TSPO has been proposed to be involved in the regulation of the mitochondrial permeability

transition pore (MPTP) [72, 73], thus exerting its effects on cellular homeostasis ([74, 75]). TSPO ligands, including benzodiazepines and more specific compounds represented by the diagnostic ligand PK11195, have been shown to attenuate cancer cell proliferation [70], to have neuro-protective effects [76], and to inhibit the MPTP [77]. There are discrepancies and controversies in the literature regarding the binding of TSPO ligands and the physiological significance of their interactions [78-80]. For instance, TSPO ligands have been reported to have pro-apoptotic effects in cancer cells [81] but to be anti-apoptotic in cardiac tissue after ischemia and reperfusion injury [77, 82, 83]. A major difficulty comes from the interpretation of binding data measured in complex systems such as total membranes or whole cells from different tissues.

TSPO is a promising target for therapeutic ligand development

The interaction of TSPO with ligands continues to be a emerging research focus for development of drugs targeting TSPO for imaging and treatment of diverse disease states (Table 1). Table 1 (adopted from [84]) summarizes several major developments of imaging ligands. From the early generation of ligand PK11195 that binds TSPO at nano-molar affinity, new generations of TSPO ligands have achieved sub nano-molar affinity to the protein and greatly increased specificity. However, one of the major obstacles remaining in the development of imaging ligands as highlighted in the table is ligand's sensitivity towards a human rs6971 polymorphism (A147T). The sensitive ligands summarized in the table display low efficiency in imaging patients with this polymorphism. Developing ligands that are tolerant for such variation in mTSPO may assist by structural

based ligand characterization and computational ligand screening. In order to achieve that, one would need to establish experimental conditions to probe TSPO structure reliably and with high resolution. The computational structure-based ligand docking also requires high quality models of TSPO with appropriate binding pocket features to distinguish minor structure perturbations arise from A147T polymorphism.

Chemical Class	Radioligand	Advantage	Limitation	Generation
Isoquinoline carboxamide	^{11}C-PK11195	rs6971-insensitive	Low signal-to-noise ratio	First
Indole acetamides	^{11}C-SSR180575			
Vinca alkaloids	^{11}C-vinpocetine			
Phenoxyarylacetamides	^{11}C-PBR28		rs6971-sensitive Reproducibility in humans Short half-time ^{11}C	Second (^{11}C)
Imidazopyridine acetamides	^{11}C-DPA713			
Imidazopyridines	^{11}C-CLINME			
Dihydro-9*H*-purinacetamides	^{11}C-DAC			
Phenoxyarylacetamides	^{18}F-PBR06			
Imidazopyridine acetamides	^{18}F-PBR111	Longer half-time than ^{11}C Improved signal-to-noise ratio	rs6971-sensitive Reproducibility in humans	Second (^{18}F)
Pyrazolopyrimidines	^{18}F-DPA-714			
Isoquinoline carboxamide	^{11}C-ER176			Third
Tetrahydrocarbazole	^{18}F-GE180			

Table 1 Summary of different generation of TSPO radiotracers (table adopted from [84])

Structural understanding of TSPO has been improved by the recent high resolution experimental structural models

TSPO structural functional studies was initiated from recombinant expression and purification of mice TSPO (mTSPO) in to proteoliposomes [85].

Although the proteins were first solubilized in SDS detergent, subsequent re-folding into lipid bilayers restored its high-affinity cholesterol and PK11195 binding. Electron microscopy image of the freeze-fracture of proteoliposome particles showed particles consistent with the size of a monomeric protein. In a subsequent study, the same group performed NMR characterization of its cholesterol recognition/interaction amino acid consensus (CRAC) motif (L/V–X (1–5)–Y–X (1–5)–R/K) [86]. The C-terminal peptide containing CRAC motif of mTSPO was purified in organic solvent, where NMR structural studies showed the motif adopted helical conformation. Additional deletion and mutation data within the CRAC motif confirmed CRAC was necessary for the nano-molar affinity of cholesterol.

NMR studies from Murail et al. produced the first 2D ^{1}H-^{15}N NMR spectrum of mTSPO in DPC detergent micelles [87]. In the absence of PK11195 ligand, the NMR spectra of mTSPO showed typical features of a dynamic molten-globular state. PK11195 drastically improved spectrum quality and NMR peak dispersions [87].

A homologue of the mammalian TSPO, Rhodobacter sphaeroides TSPO (rsTSPO), was crystalized into helical crystals in lipid membranes [88]. The 10Å resolution cryo-EM density map elucidated dimeric packing of TSPO. The low resolution EM density showed helices orientations of TSPO but not the connectivity of the respective helices. We attempted computational structure prediction based on the cryo-EM density map [89] and proposed that the dimeric interface would support potential pathway for cholesterol transport in mammalian

TSPOs.

In 2014, the first atomic detailed mTSPO structural model was determined using solution-state NMR spectroscopy [90, 91]. The structure was determined in complex with PK11195 in DPC detergent micelles. The mTSPO protein adopts a 5-helical bundle structure where PK11195 is bound in the core of the protein. Interestingly, the authors also observed the dramatic stabilization upon ligand binding in NMR spectrum [92]. The structure of A147T polymorphism mutant of mTSPO was also determined [91]. The A147T mutation does not affect the binding mode of PK11195 to mTSPO, thus explaining the molecular mechanism of PK11195's insensitivity for this polymorphism.

The subsequent year, two TSPO X-ray crystal structural models from were published [93, 94]. While the details of these bacterial homologue structures differ much from that of the mTSPO NMR structure, the overall fold is conserved through evolution. In both crystal structures, the helix 3 adopted helical conformation, where in the mTSPO NMR structure the helix 3 was partially unfolded. The authors of the crystal structures argued that in the NMR conditions, high temperature and detergent might results in deviations of the fold from its native state.

More recently, reconstituted mTSPO in liposome was shown to exists in a dimeric equilibrium modulated by cholesterol binding using solid-state NMR [95]. While cholesterol binding affected chemical shift surrounding the CRAC motif, allosteric modulation has been observed in residues away from the CRAC motif. Other attempts to produce functional homologues of mammalian TSPO were carried out in mammalian cell cultures [96]. HEK293T cells were able to produce

functional mammalian TSPOs into membrane environment, thus circumvent harsh detergent extraction processes adopted by many prior studies.

The structural characterization of mammalian TSPO has advanced greatly in the recent years. While controversies exist in discussion of the precise structural features of TSPO proteins, the three high quality structural models present valuable starting point to hypothesis cholesterol binding mechanisms, drug interaction mode as well as functional implications. It seems that the biochemical characterization of mammalian TSPO in membrane mimetics remains a challenging task, that further optimization is required for studying the native functional states of TSPO.

CHAPTER 2 INTEGRATED STRUCTURE BIOLOGY FOR ALPHA-HELICAL MEMBRANE PROTEIN STRUCTURE DETERMINATION

This chapter is a reproduction of Yan Xia, Axel W. Fischer, Pedro Teixeira, Brian Weiner, Jens Meiler, 'Integrated structural biology for alpha-helical membrane protein structure determination', manuscript published in *Structure [97].* The chapter focus on methodologies to combine orthogonal experimental data in membrane protein modeling that I developed during my Ph.D. I am the first author of the manuscript, I contributed in producing data, writing the manuscript and generating all the figures.

De novo structure prediction for large membrane proteins remains a challenging task due to the vast conformational space of such proteins. As membrane protein structure determination is entering another stage of probing the ensemble of dynamic structures, methods such as NMR, EPR and EM were used to provide orthogonal information of the structure. We hypothesized that simultaneous integration of experimental data from multiple sources can allow for accurate atomic detail prediction of IMP structure. Our current structure determination pipeline is built on the foundation of BCL::MP-Fold and many other advances in BCL and Rosetta software, integrating various frameworks of experimental data in both programs.

We have constructed a pipeline to utilize multiple experimental restraints to fold membrane protein in BCL and Rosetta. The pipeline described in the

manuscript could determine structure to an accuracy of 1.2Å RMSD to the experimentally determined structural model. My work demonstrated the potential of this methodology to be applied to membrane protein structural determination when limited experimental data were available. The structural determination pipeline described in the manuscript could be attractive to the boarder structural biology community for structure determination. This work ties in my thesis for membrane protein structure modeling as an important computational method development, and thus can be considered as core component of my thesis.

2.1 Introduction

2.1.1 Integral membrane proteins remain a formidable challenge for structure determination methods

Alpha-helical integral membrane proteins (IMPs) are important players in many cellular functions; specifically, they orchestrate the communication between the cell and external stimuli by transferring signals and chemicals across the plasma membrane. Roughly 20-35% of a genome's proteins are IMPs and yet only around 2-3% of the experimentally determined structures in the protein data bank (PDB) are IMPs [98]. Around 80% of all protein structures currently deposited in the PDB [5] have been determined by X-ray crystallography. The particular challenge for IMP crystallization arises as crystals are inherently three-dimensional while IMPs naturally assemble into membranes that extend in two dimensions. It remains difficult to provide realistic membrane mimics in the context of a stable three-dimensional crystal, although impressive progress has been made using lipidic cubic phases [8, 16, 99-101]. Even if MP structure determination through

crystallization is feasible, biological relevance of the resulting models needs to be verified using orthogonal experimental techniques to exclude artifacts introduced by crystallization aides such as thermo-stabilizing mutations, helper proteins integrated into MP loop regions, or the non-native membrane mimic.

2.1.2 Experimentally determined structures of IMPs cover a small variety of folds

Interestingly, IMPs in the PDB cluster into only about 100 *distinct* IMP folds with more than one transmembrane helix (TMH). This number is small compared to the fold space of soluble proteins and small with respect to the number of IMP sequence families (read below). Multiple factors could contribute to this finding: the fold space of IMPs might be smaller than the fold space for soluble proteins as structure might be better conserved than sequence in IMPs with IMPs of very low sequence identity adopting the same fold [102]. This seems to be the case for G-Protein Coupled Receptors (GPCRs) or in the LeuT transporter family. In addition, it is possible that once the experimental procedures have been refined to crystallize one particular class of IMPs, many members of the same fold family are experimentally studied rather than discovering new folds [103], resulting in a non-representative fold representation in the PDB.

2.1.3 A vast sequence space for IMPs remains to be represented with experimentally determined structures

Oberai and colleagues have estimated that between 700 and 1700 families of IMPs would cover 90% of the IMP sequence space [104]. By the end of 2011, structural genomic efforts have increased structural coverage of IMPs to ~28%. It

was estimated that without significant structural genomic investment it would take up to 25 years to achieve a structural coverage of 50% of IMPs [105]. Accelerating fold determination for particularly important IMPs would therefore be of great relevance to biology. It is estimated that 3305 human membrane proteins exist in the Homo sapiens proteome (Uniprot), with 90% of the sequences mapping to one Pfam family. In particular, of all the Pfam-mapped human IMPs, only 10% (around 50) have an experimentally determined structural representative that is a human protein or a sequence related non-human protein [106]. It was calculated that with the best theoretical prediction, an additional 100 protein families need to be structurally characterized to cover the human α-helical IMP proteome to 58% of all sequences [107].

2.1.4 Current experimental and computational methods to determine IMP structures suffer from limitations

The large theoretical fold space is contrasted by experimental datasets that – if available at all – are often limited by crystals that fail to defract to high resolution (X-ray crystallography), medium resolution of electron microscopy (EM) density maps, or the limited number of structural restraints from complementary experimental techniques such as nuclear magnetic resonance (NMR), site-directed spin labeling-electron paramagnetic resonance (SDSL-EPR) spectroscopy, or cross-linking coupled with mass spectrometry (XLMS). Those experimental datasets are limited in a sense that they provide insufficient information to determine a structure at atomic detail. Ideally, computational methods could be used to fill those information gaps [33-35]. However, *de novo* prediction of an

IMP's fold from its primary sequence remains a challenging problem [36]. The vast size of the theoretical fold space makes exhaustive sampling of an IMP's potential conformations prohibitive. In addition, necessary simplification when approximating a conformation's free energy frequently results in problems distinguishing accurate from inaccurate models.

2.1.5 BCL::MP-Fold and Rosetta predict membrane protein structural ensembles using experimental data

BCL::MP-Fold [108] was developed for *de novo* protein structure prediction. We demonstrated in previous studies that BCL::Fold is able to efficiently sample the fold of large IMP [53]. To achieve sufficient coverage of the fold space, the algorithm simplifies the sampling by assembling predicted secondary structure elements (SSEs) in a virtual membrane using a Monte Carlo Metropolis algorithm, while the loops connecting the SSEs are modeled implicitly [109]. After each Monte Carlo step, the free energy of the intermediate model is approximated using knowledge-based scoring functions that were specifically developed for IMPs [108]. Incorporation of limited experimental data can compensate for the simplified representation of IMPs during sampling and energy evaluation. Incorporation of individual experimental data has been established and benchmarked for EM [33, 110], NMR [111], EPR [112], and XLMS [113]. The BCL::MP-Fold algorithm outputs a simplified fold consisting of SSEs that exhibit only limited deviations from idealized dihedral angles. These models are then input for further optimization to atomic detail using the Rosetta modeling suite [52, 114, 115]. Similar to BCL::MP-Fold, incorporation of individual experimental restraints into Rosetta been

successful for EM [53], NMR [54, 55] and EPR [56] data. The membrane environment was simulated implicitly during the structural refinement.

Considering the theoretical fold space for IMPs, we estimate that billions of folds are possible, a number increasing sharply with an increasing number of trans-membrane spans. The number of IMP sequence is large as well, though most IMP families have relatively few trans-membrane spans. We hypothesize that simultaneous integration of experimental data from multiple sources can allow for accurate prediction of IMP structures at atomic detail. Here we test this hypothesis by using BCL and Rosetta to incorporate a combination of EM, EPR, and NMR data to predict the fold of Rhodopsin. The proposed computation structure prediction pipeline is not limited to the prediction of new folds. BCL::MP-Fold and Rosetta were developed to allow simulations from a given starting structure and leverage sparse experimental data to derive a model for alternative states.

2.2 Materials and Methods

Enumerating the theoretical α-helical integral membrane protein fold space

The theoretical calculation was done using Mathematica (Wolfram). We simplified the TMH fold space by defining the position of α-helices on a two-dimensional grid, where the arrangements for a given number of TMHs was plotted. For each unique TMH arrangement, we inserted the TMHs into the arrangement to generate all possible folds. A particular fold was accepted when each TMH had two direct contacts with other TMHs and every TMH was connected in a single fold. When internal symmetry was detected for an arrangement of TMHs, the number of folds was divided by the symmetry operator.

Assuming each TMH in the neighboring sequence adopts anti-parallel insertion in membrane, to account for topology of N-terminal facing extracellular or intracellular environment, we multiplied the number of folds by two. For each number of TMHs, the number of possible arrangements of TMHs and the number of possible folds were computed.

Enumerating the sequence space of α-helical integral membrane protein

The database search for all α-helical IMPs was performed using UniProt and pfam (Figure 2A). We used the UniProt server to retrieve the sequence information of α-helical IMPs by searching for the annotated keyword 'Transmembrane helix'. The TMH annotation and the pfam family id associated with each Uniprot entry was downloaded as a tab delimited table. Around 75 thousand entries were pulled and entries containing less than three TMHs were filtered out. The unique pfam families and their number of TMHs were then compiled by clustering all UniProt entries based on their pfam id. We also directly used the pfam server to download all 2100 families that were annotated as IMPs. The sequences in each pfam family were then subjected to transmembrane span prediction using SPOCTOPUS[116] until multiple sequences were predicted to have more than two TMHs. The IMP family list mined from the two methods were then combined and cleaned for duplicates and manually inspected. The XML representation of the MPtopo database [117] was downloaded to search for IMP families with a known structural fold.

Integral membrane protein structure prediction using combined experimental restraints and BCL::MP-Fold

The test dataset of bovine rhodopsin (PDB entry 1GZM) [118] was downloaded from the PDB and considered the 'native' structure. An experimentally determined electron density map for rhodopsin [119] at 5.5 Å resolution was used for the EM data. At this resolution, TMH density could be distinguished but the connectivities between density rods could not be observed directly. The NMR data was simulated in the form of backbone chemical shift (CS) using SPARTA+ [120], and ten sets of randomly selected sparse side chain Nuclear Overhauser effect (NOE)-derived distances at a 1 restraint per residue level (a total of 326 distances) using BCL with simulated uncertainties dereived from the NMR knowledge-based potential [111]. Ten sets of distance data from EPR double electron-electron resonance (DEER) spectroscopy was simulated with a distance uncertainty model [56]. Each set consisted of at least 3 restraints per TMH (a total of 27 distances). A sample restraint file containing simulated NOE and DEER distances was included in Table.S1 & S2.

The protocol was based on the protein structure prediction protocols of BCL::MP-Fold [108] and BCL::EM-Fold [33]. The SSEs were first predicted from the primary structure of Rhodopsin using the consensus of two secondary structure prediction methods, JUFO9D [121] and SPOCTOPUS [116]. When limited NMR data was included, the backbone CS was used to generate SSEs definitions from the primary structure. The SSEs were then assembled in a multi-stage approach and intermediate and final models were evaluated using a

membrane-specific knowledge-based potential and the Metropolis criterion. During the assembly process, the protein model was randomly perturbed by one of over 100 MC moves belonging to one of six categories: (1) adding SSEs, (2) removing SSEs, (3) swapping SSEs, (4) single SSE moves, (5) SSE-pair moves, and (6) moving domains. The energy function contained terms for evaluating amino acid pairwise distances, amino acid environment, loop closure, radius of gyration, contact order, secondary structure prediction agreement, environment prediction agreement, TM topology, and steric interferences. A static membrane object was utilized in conjunction with the environment-specific potentials. If experimental data were used, the scoring function was extended by the appropriate scoring terms to account for density rod agreement with experimental density map in the case of EM [33], and a knowledge-based distance agreement evaluation in the case of EPR [56] and NMR [111]. The scores were linearly combined to a sum score with weighted score components.

1000 models were sampled in each prediction experiment with hybrid data. The models were evaluated through the RMSD100 [122] metric: $RMSD100 = \frac{RMSD}{1+\ln\sqrt{N/100}}$, where the root-mean-square deviation (RMSD) of the Cα-coordinates between the predicted model and the crystal structure model is normalized by the number of amino acid (N) of the protein. When the prediction was performed from single/multi sets of experimental data and native-like models could be identifies by their score rank, the top 1% scoring models (10) were selected for the subsequent Rosetta refinement – regardless of the selected set included non-native-like models. When the prediction was performed from EM data only, the top 10 folds

were selected.

Integral membrane protein structure refinement using combined experimental restraints and Rosetta

Rosetta was used to add loop regions and side chains to the model, and refine with a high-resolution scoring function. The protocol for the Rosetta refinement was modified to incorporate multiple experimental data.

As Rosetta uses fragments from a structural database to model local sequence bias, the fragment search excluded fragments from structures that are homologous to Rhodopsin. In the case of NMR data incorporation, fragment search was performed using backbone CS information to select for fragment with preferable backbone torsion angles [123]. For each of the models from the previous BCL stage, 500 models were sampled using Rosetta's cyclic coordinate descent algorithm [124] to build loops and remodel TMHs. Further atomic detail refinement was carried out using Rosetta's "relax" application [115] once for each of the 500 model. A TMH definition file is used for the placement of a virtual membrane encompassing each input model for membrane environment evaluations. When EM experimental data was used, additional electron density scoring terms were turned on [53]. In the case of distance restraints from NMR, a bounded penalty potential $f(x)$ was used to discourage sampling of conformations that are inconsistent with the experimental data. For each distance restraint, an upper (ub) and a lower boundary (lb) are provided by the user (Table S 1). If the measured distance of xÅ fulfills the criterion lb≤x≤ub, ; if x is outside the boundary, a score penalty would be given: if x ≤ lb, $f(x) = (\frac{x-lb}{sd})^2$; if $\text{ub} \leq \text{x} \leq \text{ub} + \text{rswitch} *$

$\mathrm{sd}, f(x) = \left(\frac{x-ub}{sd}\right)^2$; if $\mathrm{ub} + \mathrm{rswitch} * \mathrm{sd} \leq \mathrm{x}, f(x) = \frac{1}{sd}\left(x - (ub + rswitch * sd)\right) + \left(\frac{rswitch*sd}{sd}\right)^2$, where the rswitch term is set to default of 0.5. The distance restraints from EPR DEER measurements (Table S 2) are treated with a knowledge-based potential as detailed in [125]. The score terms were linearly combined with respective weighting to compute the total Rosetta energy score. The RMSD100 relative to the 'native' 1GZM structure was used to quantify the prediction accuracy. The RMSD100 specific to the TMH region was computed by limiting the comparison of Cα-coordinates to residues in the predicted TMHs. Inspection of the models and their depiction was performed using Pymol [126].

2.3 Results

The Result section is divided into subsections discussing the fold space and sequence space for IMPs, followed by the results of the Rhodopsin fold prediction experiment. Results from BCL::MP-Fold and Rosetta refinement are divided into subsections describing in detail the sampling accuracy, fold discrimination and the effects of combining hybrid experimental data on protein fold prediction.

2.3.1 Estimation of the fold space for α-helical IMPs from theoretical arrangements of TMHs

To estimate a lower boundary for the theoretical size of the IMP fold space we first simplify the problem by considering only helices that actually span the membrane. We further assume that these helices are perfectly parallel and arranged on a hexagonal grid to maximize packing density. Under these assumptions we can compute the number of TMH arrangements, which is the

general relative placement of trans-membrane spans. From the number of arrangements, we can infer the number of folds, which are the distinct, compact units of protein structure that differ in specific topological order of trans-membrane spans including the order of the TMH insertion and extra-/intra-cellular location of the N-terminus. Thus, the number of unique folds for a theoretical protein with X number of TMHs will be the number of its possible arrangements with all possible helices insertion in a particular arrangement times two from whether its N-terminal being inside/outside divided by the symmetry operators under a particular arrangement (Figure 3A). An example of a Five-TMH protein's fold defined by our criteria is demonstrated here (Figure 3B). For proteins with less than five TMHs, there is only one unique arrangement, but up to 120 unique folds. As one might expect, the number of TMH arrangements and folds increases exponentially with the number of trans-membrane spans (Figure 3 B,D). For example, an IMP with nine TMHs can adopt about one million distinct folds. Consequently, exhaustive sampling of all possible folds is prohibitive for larger IMPs even if one considers that some folds might be forbidden because loops might be too short to connect trans-membrane spans distant in the fold. Next we will look at the sequence space for the number of IMP families that exist.

2.3.2 Sequence space for α-helical MPs

An analysis of protein sequence families documented by Pfam (pfam27.0) and crosschecked by Uniprot annotation (Figure 4 A) The non-redundant arrangement and fold is computed for each revealed 895 IMP families consisting of α-helical TMHs. For 108 of these IMP families, at least one structure has been determined

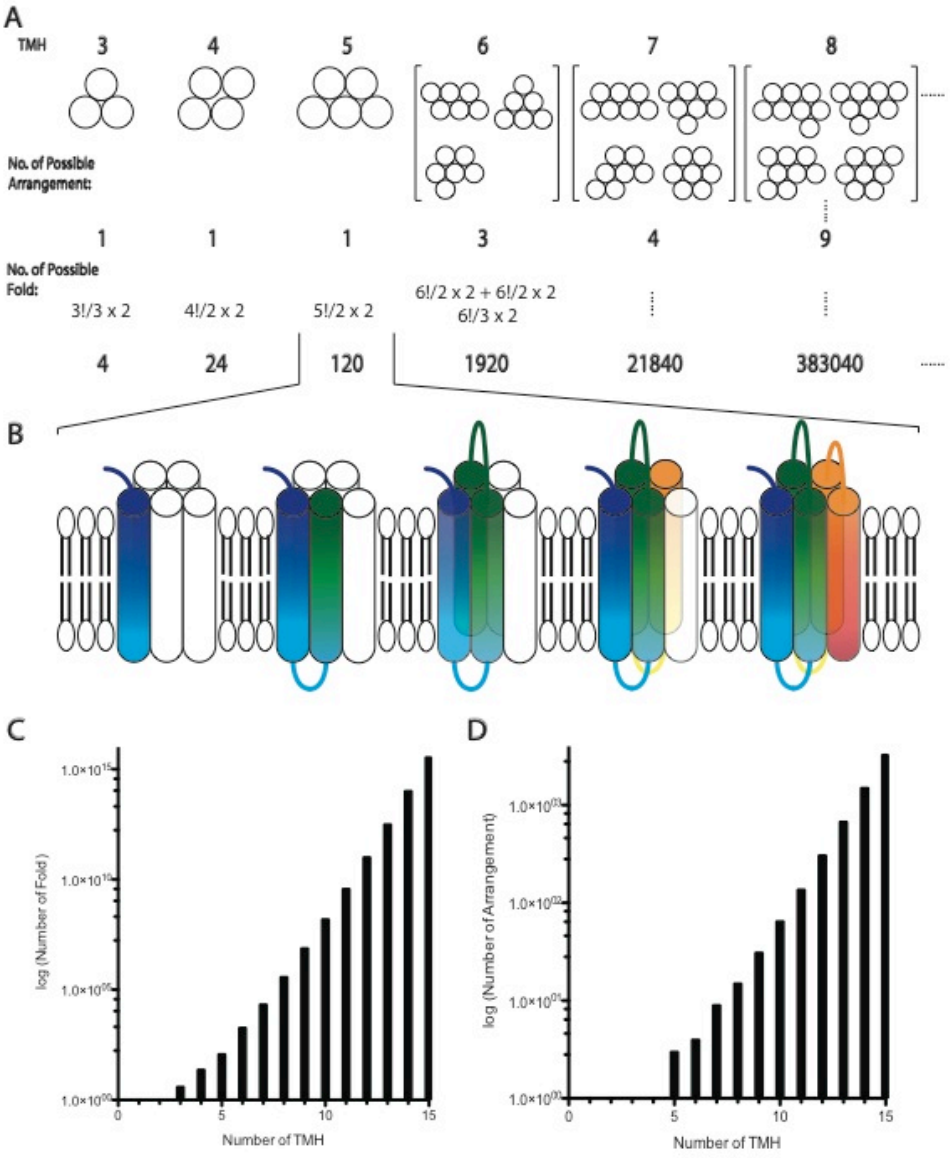


Figure 3. *The theoretical fold space for IMP.*The theoretical fold space of α-helical MPs is computed through the arrangement of helices (A) and the topological insertion of TMHs (B). The theoretical fold (C) and arrangement (D) numbers scale with the number of TMHs. The computed hexagonal grid and the TMHs are represented as circles from a top view (A): systems with more than three TMHs start to have more than one arrangement. number of TMH. An example fold of a five TMH IMP is illustrated in the rainbow diagram (B). The number of folds (C) and TMH arrangements (D) computed for different numbers of TMHs is plotted on a logarithmic scale.

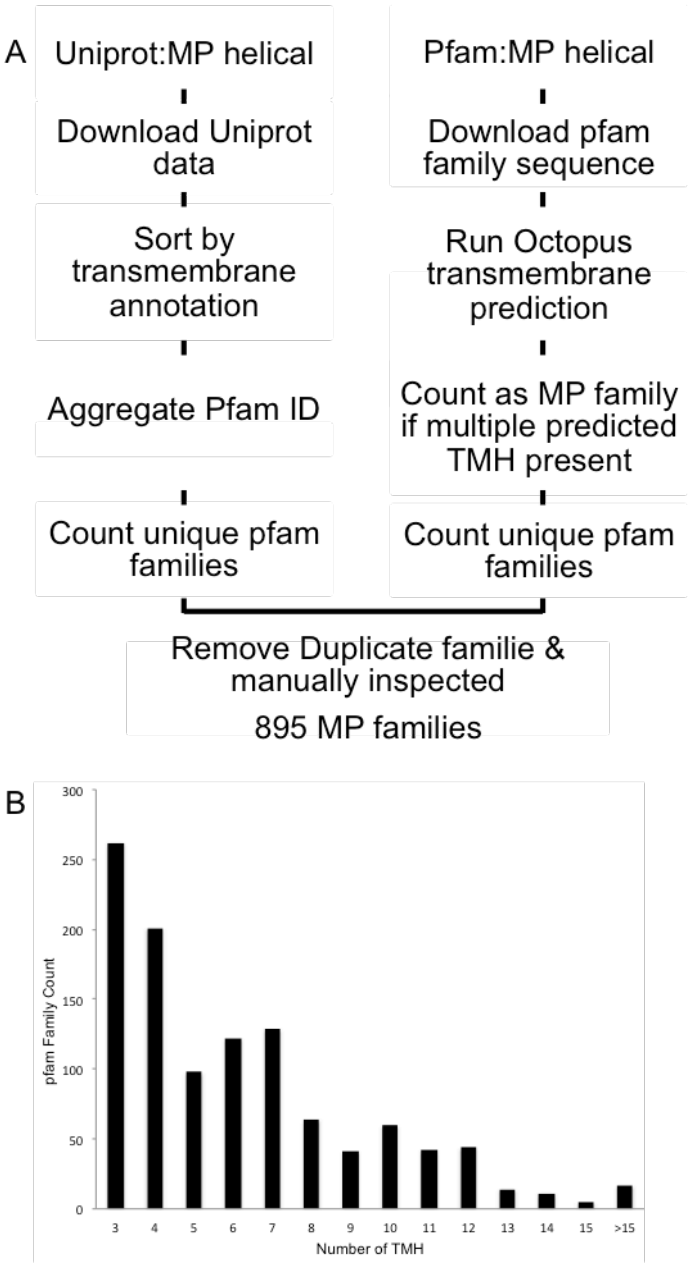


Figure 4. *The sequence space for IMPs.*A survey of the protein sequence database was performed to map the unique membrane protein families separated by sequence homology (A). Taking statistics about the average TMH numbers of each IMP family and counting the occurrences of families with average TMH numbers, a histogram (B) of families with different number of TMHs was illustrated.

experimentally based on a cross check with the MPtopo database. Since the complexity of IMP folds grows exponentially with an increasing number of TMHs, we collected TMH statistics over the 895 families from sequence space. The family counts were plotted against the TMH number: more than 70% of the IMP families have less than 7 TMHs, few have more than 12 TMHs (Figure 4B). The IMP sequence space is large, though most of the IMP families fall in the lower complexity regions of the fold space, i.e. the simplified fold space is limited to 10^5. Computational structure prediction methods such as BCL::MP-Fold are suitable for sampling such fold spaces, though distinguishing accurate from inaccurate folds would remain an obstacle for *de novo* methods without assistance from experimental data [109].

2.3.3 BCL::MP-Fold assembly of TMHs using hybrid experimental data

BCL:MP-Fold was used to simultaneously incorporate experimental data from multiple sources like EM density maps, EPR distance restraints from DEER experiments, NOE distance restraints from NMR experiments, and chemical shift information during structure prediction. In the following subsections, we present our results for α-helical IMP fold prediction from hybrid experimental data for Rhodopsin.

In addition to pure *de novo* structure prediction, predictions were performed from limited experimental data using a single dataset (NMR, EPR, or EM), integrating two data sets (NMR_EPR, EM_NMR, or EM_EPR – double-hybrid) or integrating all three experimental data sets (EM_EPR_NMR – triple-hybrid). Predicted structural models were evaluated by computing the RMSD100 [122] of

the sampled models relative to the experimentally determined structure and their respective energy scores in the BCL. Note that the RMSD100 calculation was performed over aligned regions of the structure: since BCL::MP-Fold assembles only the TMHs in the three-dimensional space, the RMSD100 values presented here only relate to the TM helices of the IMP.

Restraints BCL			RMSD100		Score	
			Best	Top 5%	Best	Top 5%
No Restraints			4.5	5.4	9.4	9.2
NMR			3.1	3.8	5.1	6.1
	EPR		3.5	4.5	8.5	7.5
		EM	2.9	4.8	8.3	8.1
NMR	EPR		2.7	3.7	3.6	4.0
NMR		EM	3.0	4.0	4.0	5.5
	EPR	EM	1.9	4.1	8.4	7.4
NMR	EPR	EM	1.4	2.9	2.5	3.9

Restraints Rosetta			RMSD100		RMSD100_TM		Score		Score (RMSD100_TM)	
			Best	Top 5%	Best	Top 5%	Best	Top 5%	Best	Top 5%
NMR			4.0	4.4	1.7	3.0	5.3	6.4	2.1	1.8
		EM	4.0	4.2	3.0	3.0	5.4	4.9	4.6	3.7
NMR	EPR		3.9	4.1	1.5	1.8	8.2	5.8	1.5	1.6
	EPR	EM	3.3	3.7	1.9	2.1	4.4	4.2	2.0	2.1
NMR		EM	3.3	3.5	1.6	1.3	3.8	4.1	1.1	1.1
NMR	EPR	EM	2.9	3.0	1.3	1.2	3.8	3.6	1.2	1.1

Table 2. *Structure prediction results from hybrid experimental data.*The RMSD100 metric is used to quantify model quality. RMSD100: RMSD100 of models ranked by RMSD100 to native model; Score: RMSD100 of models ranked by either BCL score or Rosetta score; Top5%: Averaged RMSD100 over top 5% models ranked by the respective metric. RMSD100_TM: RMSD100 value calculated from TMH regions of the models that ranked by total RMSD100. Score (TM_RMSD): RMSD100 value calculated from TMH regions of the models that ranked by Rosetta score.

In each case, BCL::MP-Fold was able to sample the native-like fold (Table 2). However, although the *de novo* sampled structures were structurally similar to the experimentally determined structure (RMSD100 = 4.5 Å) (PDB ID 1GZM

[118]), the scoring function lacked the discriminative power to identify the most accurate models. The models with the most favorable score often exhibited large structural deviations from the experimentally determined structure. Incorporation of experimental restraints improved the sampling accuracy and discriminative power of the scoring function. Prediction using a single set of experimental data (**Figure 5** A) improved sampling and scoring slightly. Notably, NMR data improved model discrimination and sampling density around the experimentally determined structure (**Figure 5** B), where the best scoring models achieved an RMSD100 of 5.1 Å. EM data helped in positioning the TMHs in the EM density rods, resulting in more accurate sampling of the native fold (RMSD100 = 2.9 Å). However, the medium resolution density map, combined with our simplified energy evaluation, was unable to unambiguously identify the most accurate models by score. EPR data suffered from the limited number of restraints leading to a moderate 1 Å improvement in sampling accuracy (RMSD100 = 3.5 Å). Again, the scoring function was unable to identify these improved models. Both, NMR- and EM-assisted models included low RMSD100 folds in the top scoring models and were further refined in next stage under the assumption that an all atom representation of the protein structure combined with higher resolution energy evaluation would allow to identify and refine the most accurate models.

Before we proceeded to the refinement step, we combined two or three experimental data sets to test if integrating data from multiple sources improved sampling accuracy and/or discriminative power. Substantial improvements were observed over *de novo* prediction or the single-experimental data prediction. The

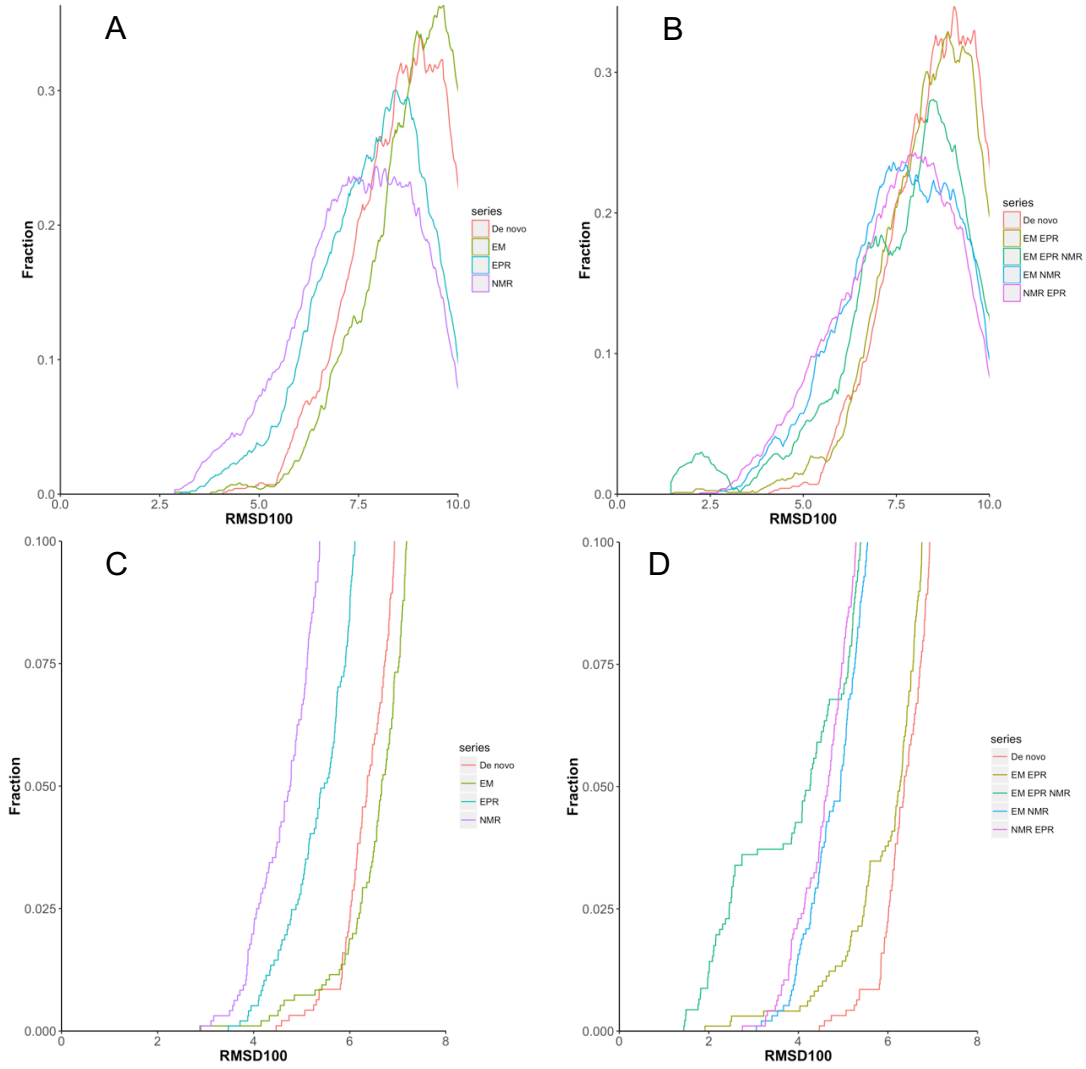


Figure 5. *Prediction accuracy of low-resolution SSE assembly using hybrid experimental data.* Density distributions of structure prediction accuracies using single experimental dataset (A,C) and hybrid experimental dataset (B,D). The fraction of models versus the predicted models' RMSD100 to the native crystal structure is shown. (A) A comparison is drawn between the sampling density for native-like models using *de novo* structure prediction (orange), EM (light green), EPR (blue) and NMR (purple). (C) Cumulative fraction of models that falls within 8Å RMSD, the y-axis is cut off at 0.1. (B) A comparison is drawn between the sampling density for native-like models using de novo structure prediction (orange), NMR_EPR (light green), EM_EPR (green), EM_NMR (blue) and EM_EPR_NMR (Purple). (D) Cumulative fraction of models that falls within 8Å RMSD, the y-axis is cut off at 0.1.

NMR_EPR, EM_NMR sets improved sampling density near the native-like folds to below 3 Å RMSD100(**Figure 5** B, D). The best models ranked by score are among the 5% most accurate models. EM_EPR prediction also improved the RMSD100 of

the most accurate folds sampled to 1.9 Å but had an accurate model ranked only second by score (RMSD100 = 2.2 Å). The improvement in sampling density was less compared to tests that included NMR restraints (**Figure 5**B), possibly due to the limited distance information in experimental data and alternative folds fulfilling the 27 EPR distance restraints.

Unsurprisingly, the EM_EPR_NMR set performed best. The most accurate model had an RMSD100 of 1.4 Å and an average RMSD100 of 2.9 Å among the top 5% sampled folds. The sampling density was also significantly larger as seen in the peak centered near 2 Å in **Figure 5** B, D. Models ranked by score in EM_EPR_NMR also had the best agreement with the crystal structure, which is demonstrated by an RMSD100 of 2.5 Å and an average RMSD100 of 3.9 Å among the top 5% scoring models. The top scoring models using two or three datasets were further processed for all-atom refinement.

2.3.4 Loop modeling and structural refinement using hybrid experimental data

The BCL::MP-Fold models consist of simplified helices with only limited deviations from idealized dihedral angles. The models don't contain loops and side-chain components. The top scoring models sampled in the BCL::MP-Fold stage using EM, NMR, EM_NMR, EM_EPR, NMR_EPR and EM_NMR_EPR datasets were refined in Rosetta using their corresponding restraint sets. In each case, except EM and EM_EPR, the top scoring BCL models used as inputs in this stage were within reasonable accuracy to the experimentally determined 1GZM structure. The single restraint set (EM and NMR) was successful in refinement and

finding native-like models among the best scoring models. However, using two or three experimental data sets, added additional layers of accuracy in refinement (Figure 6 A, B and Table 2).

The energy landscape of the rhodopsin fold, visualized by plotting the RMSD100 of each model with its respective Rosetta energy score, shows that native-like models are strongly favored by the inclusion of experimental data during the refinement stage (Figure 6 A, B). The use of hybrid experimental data improved the prediction accuracy in the core of the protein. The RMSD100_TM, which quantifies the RMSD100 of the trans-membrane region, for the top 5% sampled models using hybrid restraint data improved by at least 1 Å improvement over that of the top 5% sampled models using single restraint (Table 2, RMSD100_TM).

With EM data alone, which is not able to generate an unambiguous answer to the correct fold using the low-resolution scoring functions in BCL::MP-Fold, the Rosetta refinement successfully sampled the core of the protein with a high accuracy of 3 Å RMSD100_TM. Although input models contained models with incorrect folds that extend to an RMSD100_TM above 8 Å, the native-like folds were strongly favored by the EM density scoring function. EM_EPR and EM_NMR both showed improvements when EM data was incorporated.

Overall, whereas the inclusion of the EM density map resulted in an improvement of the total RMSD100 of the best scoring model to below 6 Å, distance type restraints (NMR, NMR_EPR) were unable to refine the best scoring models to below 6 Å (Figure 6 A, B). The EM data restrains loop conformations to

the boundary of the density map. NMR information could improve modeling the correct contacts between the TMHs, as seen in the improvement in TM_RMSD100 to below 2 Å in the best scoring models. The models that were refined using NMR data also had less favorable scores to our surprise. To break down the cause of the worse scores, rescoring using only the built-in Rosetta scoring function was performed to investigate the behavior of the membrane scoring function (see discussion).

Refinement using EM_NMR_EPR resulted in the most accurate models (RMSD100 = 2.9 Å). The total RMSD100 approached a limit of 4 Å while the RMSD100 of the core helical TM region was close to 1 Å. In our prediction pipeline, the retinal molecule is not modeled. Accurate modeling of the extra-cellular loop 2 (ECL2) that interacts with the retinal molecule and the 30 amino acids long N-terminus remains challenging using the Rosetta loop modeling algorithm.

The refinement protocol using hybrid experimental data could recapitulate backbone conformation features observed in the crystal structures such as kinks in TMHs. For example, the best scoring model from the EM_NMR_EPR dataset reproduced the backbone contacts and loop conformation similar to the native structure (Figure 6 C). Even when input models were idealized helices, the kinks in helices six and seven could be observed after the refinement. The side chain conformers from the model could be modeled to high accuracy (Figure 6 D).

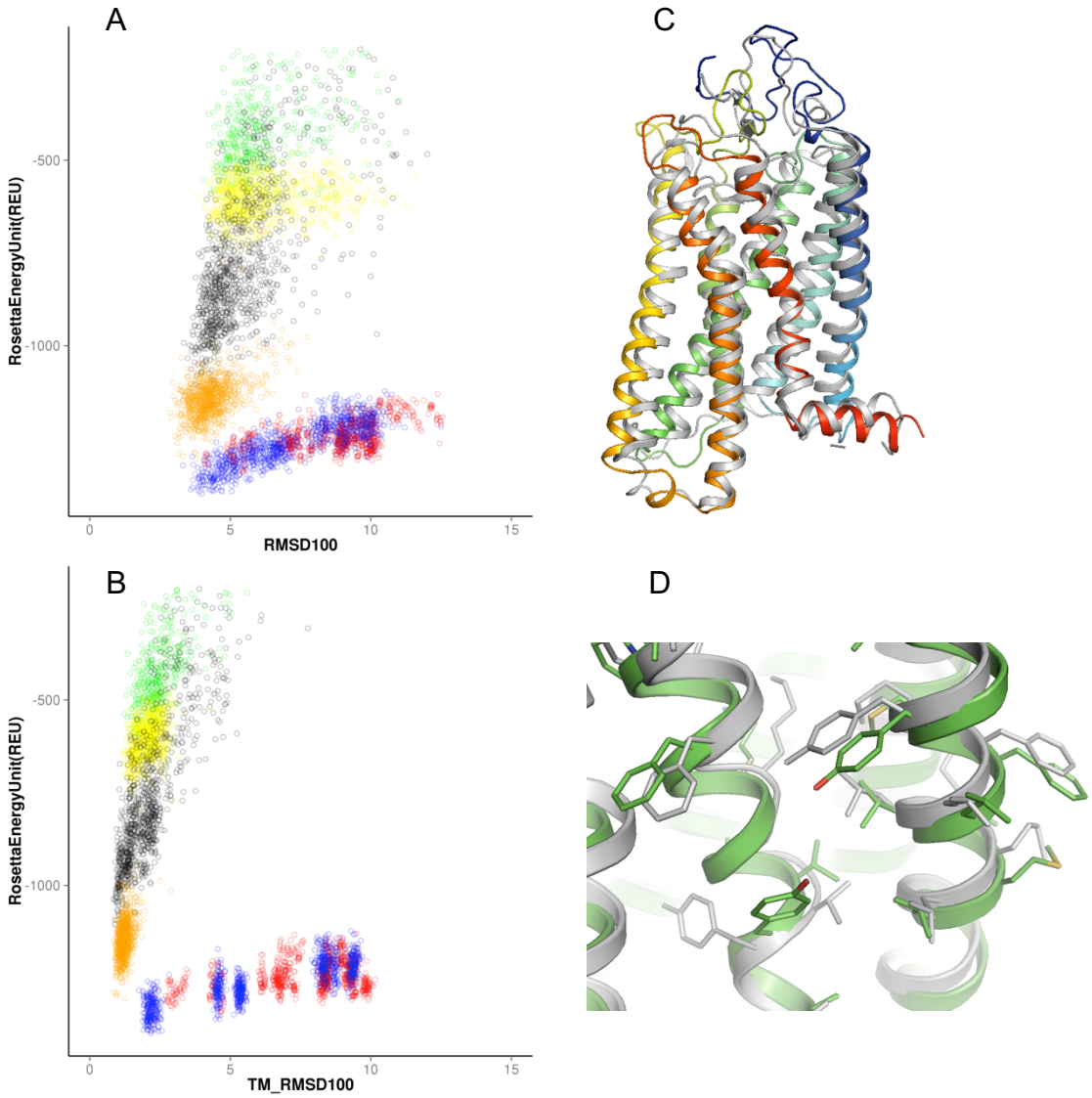


Figure 6. *Prediction accuracy of Rosetta refinement using hybrid experimental data.* The scatter plot for model quality in terms of sampling and scoring of rhodopsin models to atomic detail using Rosetta modeling suite and representative predicted models of low energy. The resulting models are plotted with their respective Rosetta score against their RMSD relative sto the experimentally determined 1GZM structure (A), or TMH RMSD to the TMH region of 1GZM (B). The scatter plot is color coded with the respective dataset used for the refinement: NMR (Green), EM (Red), NMR_EPR (Yellow), EM_EPR (Blue), EM_NMR (Black) and EM_EPR_NMR (Orange). The best scoring model from EM_EPR_NMR experiment is depicted in rainbow diagram, with the experimental structure show in grey (C). The side chain rotational conformer of the predicted structure matched the experimental structure with high accuracy (D).

2.4 Discussions

2.4.1 The fold and sequence space of α-helical MPs

Structural representatives remain to be determined for about seven hundred more IMP families to allow for comprehensive comparative modeling of all IMPs. At an average rate of six new folds per year (average over the last 5 years), experimentally determining the remaining IMP folds would take approximately 110 years. Our survey also showed that a large number of IMP families have a relatively low theoretical fold complexity. The fold space of these IMPs can be comprehensively searched with the current computational algorithms. Therefore, integrating limited experimental data into these computational algorithms should accelerate fold determination for such IMP families. For larger proteins, such as 7-TM Rhodopsin, which has a theoretical fold space of around one hundred thousand representations, additional restraints will be required to achieve high prediction accuracy.

In this survey, we also found that within a single IMP family, there are members containing different numbers of predicted TMHs, further increasing complexity for accurate homology modeling, which requires a template structure that contains most of the structural elements. Modeling such IMPs would require a combination of template-based modeling methods and de novo structure prediction in order to sample the additional fold space.

2.4.2 Available experimental data for Rhodopsin are suboptimal for IMP structure determination

Although the use of actual experimental data would be preferred in demonstrating

our algorithm's capability in application, this proved difficult in the present case: The DEER restraints published by Hubbell et al[44]. are centered on the intra-cellular side of rhodopsin to monitor the conformational changes upon receptor activation. The amino acid pairs for labeling were selected for that purpose alone and are not suited to determine the fold. In our previous studies using a single restraint type, incorporating these 16 restraints resulted in an RMSD100 improvement from 5Å to 4.5Å [112]. In our study using simulated DEER restraints, the RMSD100 of the most accurate model arrived at 3.5 Å. The additional improvement by 1Å is a direct result of selecting restraints that restrict the overall fold of the protein. There is a complete NMR dataset available for sensory Rhodopsin from bacteria [127], however the NMR restraints for bovine rhodopsin are sparse and affiliated with a higher uncertainty due to experimental limitations. In the experimental paper where the bovine rhodopsin NMR structure was described, only secondary structure restraints stem from NMR. In addition, 17 long-range constraints from EPR experiments and 58 inter-helical constraints from low resolution electron diffraction experiments are used (pdb ID: 1JFP) [128]. The experimental procedure was also based on resolving overlapping peptide constructs of bovine rhodopsin solubilized in DMSO, which is not expected to stabilize a biologically relevant conformation of a membrane protein. As a result, the limited availability of actual NMR-derived distance data prompted us to use simulated NMR restraints. For our simulated NMR distances, the lower bound of distances is 0 Å, the upper bound of the distances is below 6Å.

2.4.3 Experimental data overcomes the limitations of simplified representation of IMPs during de novo structure prediction

Due to the simplified representation of IMPs in the BCL and in Rosetta, the depth of the native energy minimum is reduced. As a consequence, energy differences between the native-like folds and non-native-like folds are small and

unambiguous identification of the correct native fold becomes often difficult or impossible. Further, by virtue of their respective sampling algorithms, not all IMP conformations are easily accessible or accessible at all. Therefore, the optimal native conformation might be missed as the algorithm fails to sample it.

In the Rosetta stage, we observed that models predicted from EM and EM_EPR datasets had more favorable scores compared to models predicted from NMR data. To confirm that those discrepancies were not caused by the experimental data, the models were re-evaluated using the original Rosetta membrane scores. Notably, the Rosetta score of the crystal structures was substantially lower than the Rosetta score of the computational models (**Figure S 2**) suggesting that the Rosetta energy function can correctly distinguish non-native states from native states. Since the overall RMSDs of the computational models are above 3 Å, the energy gap could be explained by inaccurate loop conformations that arise from insufficient sampling of the long N-terminal loop and ECL2. An energy gap is observed between models predicted from EM, EM_EPR and other datasets where NMR data were included. Analysis of the individual score components showed that EM and EM_EPR models, although exhibiting a larger deviation in the core of the protein, have less unfavorable energy contributions from fa_rep (Lennard-Jones repulsive energy between atoms in different residues[49, 129]), and fa_dun (Internal energy of sidechain rotamers as derived from Dunbrack's statistics [48]). At the same time, models predicted from triple-hybrid experimental data exhibited lower Rosetta scores compared to that of EM_NMR, NMR_EPR, and NMR. The experimental data leveraged in the

prediction constrained the model to sample conformations in a fold space close to the experimentally determined structural model and was able to correct the deviations in the scoring function used by Rosetta. The reason for such observations is perhaps the richness in the side-chain contact information provided by NMR distance data, that forces side-chain contact that are otherwise hard to be sampled due to the simplified representation of the IMPs.

The existing abundance of sequence data of IMPs are expected to challenge the limit of experimental structural determination pipelines in the near future. Hybrid approaches combining experimental and computational techniques could accelerate determination of protein structures and refine existing knowledge of protein structural functions. We demonstrated that using a combination of computational structure prediction methods and sparse experimental data enables accurate fold determination for large IMPs to atomic detail. Although combining orthogonal experimental data improved the prediction accuracy, future applications should always consider the source of the experimental data on whether they can be used complementarily since not all combinations yield results of comparable quality. Future development of the proposed structural prediction pipeline will be focused on the prediction of structural ensembles.

2.5 Supplemental Materials

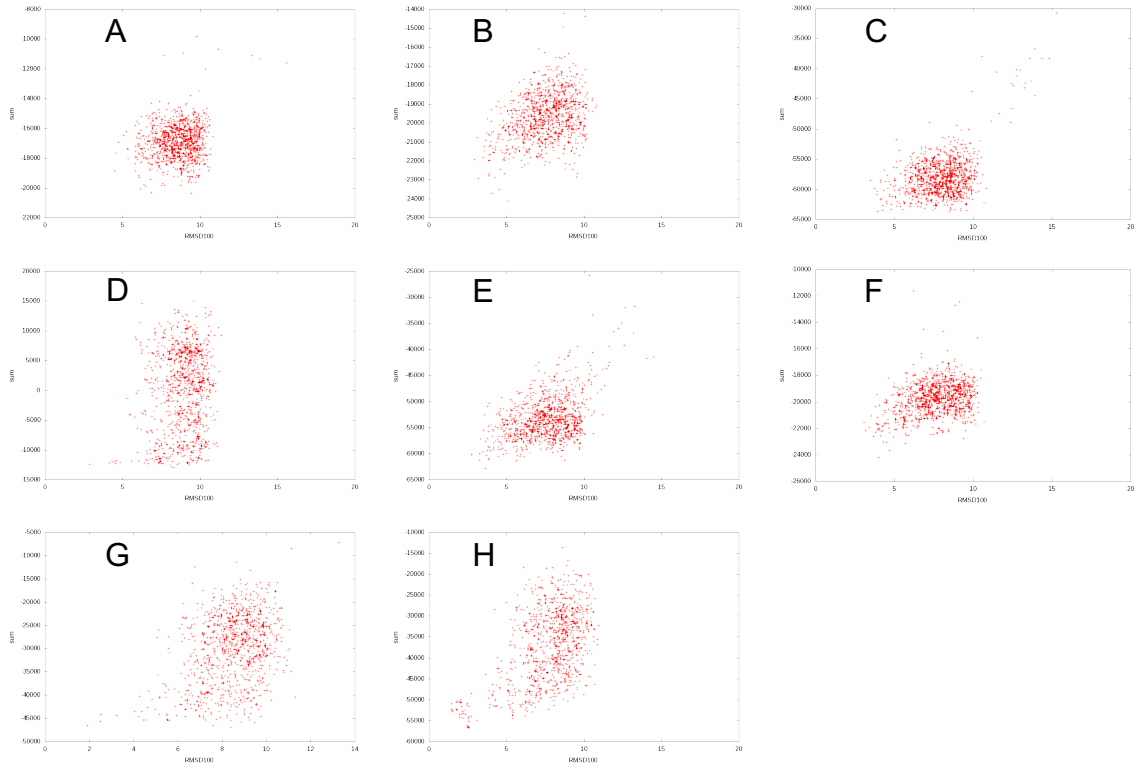


Figure S 1 Score versus RMSD100 plot for BCL::MP-Fold assembly of rhodopsin fold. (A) de novo; (B) NMR; (C) EPR; (D) EM; (E) EM_NMR; (F) NMR_EPR; (G) EM_EPR; (H) EM_NMR_EPR.

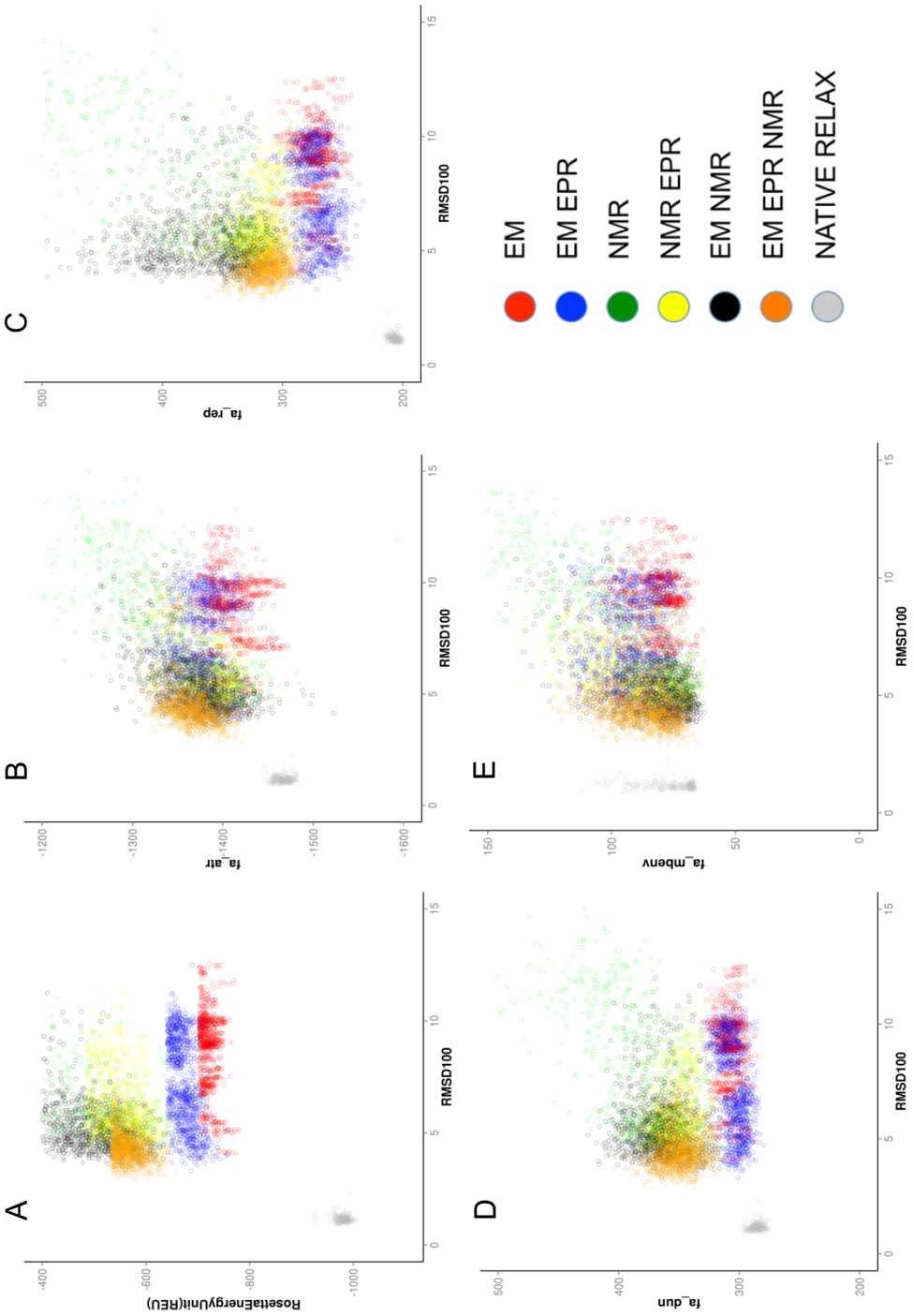


Figure S 2. Score versus RMSD100 in TMH for models rescored with original Rosetta membrane scores and individual score terms. fa_atr is the Lennard-Jones attractive energy between atoms in different residues; fa_rep is the Lennard-Jones repulsive energy between atoms in different residues; fa_dun is the internal energy of rotational tautomer of the sidechain, derived from Dunbrack;s statistics of protein sidechains observed in pdb; fa_mbenv is the solvation energy for residues in membrane environment based on amino acids' hydrophobicity.

Restraint type	Atom 1	Residue 1	Atom 2	Residue 2	Evaluation Function	lb	ub	sd	Tag /rswitch
AtomPair	H	75	1HG1	130	BOUNDED	0	5.94786	1	NOE
AtomPair	3HD1	59	3HD2	77	BOUNDED	0	5.33137	1	NOE
AtomPair	2HD1	131	1HG1	254	BOUNDED	0	4.62996	1	NOE
AtomPair	3HD1	76	1HD1	131	BOUNDED	0	4.93543	1	NOE
AtomPair	3HG2	63	3HD2	77	BOUNDED	0	5.92617	1	NOE
AtomPair	2HD1	133	1HG1	218	BOUNDED	0	4.05534	1	NOE
AtomPair	1HD2	125	3HD2	262	BOUNDED	0	5.8083	1	NOE
AtomPair	3HG1	139	3HG2	230	BOUNDED	0	4.66986	1	NOE
AtomPair	2HG2	129	H	219	BOUNDED	0	3.95145	1	NOE
AtomPair	1HG1	130	H	156	BOUNDED	0	5.00824	1	NOE
AtomPair	3HG1	209	3HD1	214	BOUNDED	0	5.63787	1	NOE
AtomPair	H	260	1HD1	305	BOUNDED	0	4.45914	1	NOE
AtomPair	1HG2	129	H	219	BOUNDED	0	5.48106	1	NOE
AtomPair	1HD1	128	3HD1	219	BOUNDED	0	4.94971	1	NOE
AtomPair	1HD2	72	3HG2	250	BOUNDED	0	5.42105	1	NOE
AtomPair	2HD2	131	3HG2	254	BOUNDED	0	4.32513	1	NOE
AtomPair	3HD1	131	H	254	BOUNDED	0	5.35469	1	NOE
AtomPair	2HG1	139	H	230	BOUNDED	0	4.91498	1	NOE
AtomPair	2HD1	75	2HG2	130	BOUNDED	0	5.69671	1	NOE
AtomPair	1HD2	72	1HG1	250	BOUNDED	0	4.5216	1	NOE
AtomPair	H	51	1HG2	87	BOUNDED	0	4.4161	1	NOE
AtomPair	2HD2	50	1HG2	304	BOUNDED	0	4.84477	1	NOE
AtomPair	H	47	1HG2	300	BOUNDED	0	5.80309	1	NOE
AtomPair	3HD1	219	2HG2	258	BOUNDED	0	3.0854	1	NOE

Table S 1. Sample NMR restraints used in Rhodopsin fold determination. NMR restraints were specified in a line-based file. AtomPair specifies the type of restraint to be used, followed by atom name and residue number of the first and second atom. BOUNDED term is used to set the restraint evaluation for calculating a bounded penalty. The lb an ub define the lower and upper bound of the NOE distances, while sd defines the standard deviation of distances. The tag is optional input to specify a rsswitch term, when tag is not numeric, the rswitch is set to default of 0.5.

Restraint type	Atom 1	Residue 1	Atom 2	Residue 2	Evaluation Function	EPR descript	Distance value	Weight	Bin size
AtomPair	CB	5	CB	215	SPLINE	EPR_DISTANCE	31.6286	1	0.5
AtomPair	CB	13	CB	210	SPLINE	EPR_DISTANCE	32.5877	1	0.5
AtomPair	CB	35	CB	241	SPLINE	EPR_DISTANCE	56.7064	1	0.5
AtomPair	CB	39	CB	293	SPLINE	EPR_DISTANCE	9.27139	1	0.5
AtomPair	CB	41	CB	159	SPLINE	EPR_DISTANCE	35.8508	1	0.5
AtomPair	CB	42	CB	136	SPLINE	EPR_DISTANCE	41.8808	1	0.5
AtomPair	CB	57	CB	295	SPLINE	EPR_DISTANCE	22.5476	1	0.5
AtomPair	CB	58	CB	223	SPLINE	EPR_DISTANCE	25.2647	1	0.5
AtomPair	CB	70	CB	221	SPLINE	EPR_DISTANCE	22.281	1	0.5
AtomPair	CB	72	CB	128	SPLINE	EPR_DISTANCE	11.9059	1	0.5
AtomPair	CB	78	CB	299	SPLINE	EPR_DISTANCE	13.5557	1	0.5
AtomPair	CB	81	CB	263	SPLINE	EPR_DISTANCE	22.6959	1	0.5
AtomPair	CB	91	CB	165	SPLINE	EPR_DISTANCE	19.359	1	0.5
AtomPair	CB	97	CB	285	SPLINE	EPR_DISTANCE	13.7844	1	0.5
AtomPair	CB	115	CB	291	SPLINE	EPR_DISTANCE	18.7978	1	0.5
AtomPair	CB	116	CB	205	SPLINE	EPR_DISTANCE	23.0765	1	0.5
AtomPair	CB	131	CB	261	SPLINE	EPR_DISTANCE	11.0256	1	0.5
AtomPair	CB	131	CB	291	SPLINE	EPR_DISTANCE	25.1033	1	0.5
AtomPair	CB	135	CB	167	SPLINE	EPR_DISTANCE	23.5482	1	0.5
AtomPair	CB	140	CB	226	SPLINE	EPR_DISTANCE	7.02843	1	0.5
AtomPair	CB	158	CB	212	SPLINE	EPR_DISTANCE	21.0541	1	0.5
AtomPair	CB	162	CB	249	SPLINE	EPR_DISTANCE	29.4457	1	0.5
AtomPair	CB	200	CB	301	SPLINE	EPR_DISTANCE	30.2289	1	0.5
AtomPair	CB	213	CB	267	SPLINE	EPR_DISTANCE	16.8667	1	0.5
AtomPair	CB	230	CB	290	SPLINE	EPR_DISTANCE	37.2075	1	0.5
AtomPair	CB	247	CB	306	SPLINE	EPR_DISTANCE	16.5986	1	0.5
AtomPair	CB	274	CB	300	SPLINE	EPR_DISTANCE	26.0892	1	0.5

Table S 2. Sample EPR restraints used in Rhodopsin fold determination. EPR DEER distance restraints were specified in a line-based file, AtomPair specifies the type of restraint to be used, followed by atom name and residue number of the first and second atom. SPLINE and EPR_DISTANCE would specify the program read in a histogram file and create cubic spline for the RosettaEPR knowledge-based potential. Experimental distances are set in the next column. Weight sets the numerical multiplier for the score term when linearly adding it to the total energy evaluation. Bin size set the histogram bins, in this case, distances are evaluated by a 0.5Å bin.

CHAPTER 3 UNIFIED STRUCTURE MODEL OF THE MAMMALIAN TRANSLOCATOR PROTEIN (TSPO)

Chapter 3 is a reprint of “Yan Xia, Jun Li, Amanda Duran, Charles R. Sanders, Charles H. Manning, Jens Meiler. A Unified Structural View of the Mammalian Translocator Protein.” Manuscript in preparation. I designed and performed the experiments, analyzed the data, generated the figures and write the manuscript or text describing the specific scientific findings.

The translocator protein (TSPO), previously named peripheral-benzodiazepin receptor (PBR), is a membrane protein located on the outer-mitochondial membrane. Structures of mouse TSPO (mTSPO) and its homologue from bacterial species determined using NMR spectroscopy and X-ray crystallography, respectively, display distinct differences that prompted discussion and additional experiments. Herein we leverage computational and experimental studies to conclude that that the lipid-mimetic system used to solubilize mTSPO thermodynamically destabilizes the protein, introduces structural perturbation, and affects its ligand binding characteristics. Further, to contribute to resolving the controversies in the structural features of mTSPO, we used Rosetta to construct a unified model that reconciles deviating features of mammalian and bacterial TSPO that are likely a result of the detergent system used for structure determination of mTSPO. The unified mTSPO model agrees with available experimental NMR, is physically realistic – i.e. thermodynamically not frustrated – as judged by the Rosetta energy function, and simultaneously shares the structural features

observed in sequence-conserved regions of the bacterial proteins. Finally, we determined the putative binding site for a new, clinically tested imaging ligand VUIIS8310 using NMR spectroscopy and propose a computational model of VUIIS8310 in complex with mTSPO.

3.1 Introduction

The 18kDa TSPO, previously known as the peripheral benzodiazepine receptor (PBR), is an evolutionarily conserved 5 transmembrane (TM) alpha-helical protein in the mitochondria [1, 2]. Research on TSPO has been conducted for decades from basic biochemical analysis to high-affinity ligand development. It has been found that TSPO is universally expressed in all human tissues, especially elevated in steroidogenic tissues and pathological tissues such as cancer and brain inflammation tissues [3, 4]. *In vivo* imaging suggests that TSPO is also a promising target of cancer and central nerve system (CNS) disorder biomarker or therapeutics [5, 6]. There is evidence that the majority of TSPO ligands interact with the loop between helix 1 and 2, while the cholesterol binding is mediated by its C-terminal cholesterol recognition/interaction amino acid consensus (CRAC) motif [7, 8].

3.1.1 TSPO is involved in important functions in mitochondrial metabolism

TSPO, together with the voltage dependent anion channel (VDAC) and adenine nucleotide translocase (ANT), is a component of the mitochondrial permeability transition pore (MPTP) during apoptosis [9, 10]. TSPO-mediated cholesterol translocation is discussed as the rate-limiting step in progesterone synthesis in mitochondria, where cholesterol is further metabolized to steroid

hormones [11]. Several studies have been conducted to probe pharmacological or biochemical properties of TSPO involvement in hormone metabolism and apoptosis [12-14]. Recently, a number of genetic studies of *in vivo* or *in vitro* knocked-out TSPO challenged the paradigm of TSPO's function role in biological processes [15-18]. Additional compensatory factors such as F0F1 ATP synthase was found in MPTP complex [19], that MPT could occur in TSPO knockout mice [18]. Some TSPO ligands were found to also inhibit F0F1 ATP synthase mediated MPTP function [20]. Conditional knockout or global knockout of TSPO produced no apparent changes in viability and steroid hormone synthesis in mice [15, 16], conflicting with previous report that TSPO knockout was embryonically lethal, albeit the methodology of generating TSPO null models were different. An *in vitro* knockout of TSPO by siRNA also failed to display inhibition of steroidogenic pathway [16]. One report also suggested that the effect of early generation TSPO ligands on steroidogenesis might be the result of alteration of membrane properties by ligands or other off-target effects [21]. TSPO is highly expressed in activated microglia, and has been proposed as a biomarker of neuroinflammation [22]. With the TSPO global knockout mice, it was subsequently demonstrated that activation of microglia after neuronal injury does not require TSPO, though microglia isolated from TSPO knockouts have altered oxygen consumption and ATP synthesis rates [23]. At this point, it seems likely that compensatory mechanisms exist for absence or malfunction of TSPO as mitochondrial function is a critical component of cellular function. Thus, the presence of redundancy in mechanisms that mediate crucial mitochondrial functions would be expected. It is

possible TSPO's exact function in biological processes remains to be elucidated. Nevertheless, TSPO remains an ancient, well-conserved protein that is highly expressed during physiological response and disease state and thus an important target of study.

3.1.2 TSPO is an emerging target for imaging and therapeutic ligands development

Despite the controversies about TSPO function, TSPO has become a target for cancer imaging via positron emission tomography (PET) as TSPO is overexpressed in several cancer types, including glioma [24, 25], breast [26, 27], and colorectal [28-30] cancers. TSPO overexpression in cancer cells has also been linked with disease progression and prognosis in patients with colorectal [28-30], breast [26, 27], and brain [24, 31] cancers. Additionally, elevated TSPO levels appear to be associated with metastatic potential in breast and colorectal cancer cells [26, 32, 33]. Thus, TSPO is an important prognostic biomarker in oncology and target for tumor-selective TSPO PET ligands for cancer imaging. The second generation of PET imaging ligands derived from pyrazolopyrimidine such as [^{18}F]DPA-714 [34] and [^{18}F]VUIIS1008 [35, 36] is used to visualize TSPO expression in tumors in preclinical settings. A recent development of PET ligand based on similar scaffold 7-chloro-N,N,5-trimethyl-4-oxo-3(6-[F]fluoropyridin-2-yl)-3,5-dihydro-4H-pyridazino[4,5-b]indole-1-acetamide (VUIIS8310) [37] is also shown to be a viable candidate developed as PET tracer for cancer imaging.

3.1.3 The structural fold of TSPO is highly conserved

The current understanding of the TSPO structure at molecular level

revolves around the solution NMR structure of mTSPO [38, 39] and two crystal structures of bacterial homologues from *Rhodobacter sphaeroides (rs*TSPO$^{X\text{-}Ray}$*)* and *Bacillus cereus* (*bc*TSPO$^{X\text{-}Ray}$) [40, 41]. The NMR structure of mTSPO (mTSPO^{NMR}) is determined in Fos-Cholin-12 (DPC) detergent system with the imagining ligand PK11195 complexed with the protein (Figure 8 A). The crystal structures of two bacterial TSPO homologs are determined by liquid cubic phase crystallization (Figure 8 A schematics). All three models share the same fold, despite bacterial TSPOs having less than 30% sequence identity to human TSPO (Figure 8 A schematics). In all structures, the CRAC motif is facing lipid environment and away from dimerization interface, complicating a hypothesis on a cholesterol transport mechanism.

3.1.4 Ligand conformation observed in experimentally determined structures differ

co-crystalized with bcTSPO displays a different binding mode when compared to PK11195 as positioned in the mTSPO structure determined by NMR spectroscopy. The ligand binding site in *bc*TSPO$^{X\text{-}Ray}$ is 5.5 Å away from the center of the protein, while the ligand binding site in mTSPO^{NMR} is 6.7 Å from the center. Use the In the crystal structure model, PK111195 establishes a π-π stacking between the indol ring of the ligand and, a polar interaction between the ligand's hydroxyl group with the W138 tryptophan side chain. In contrast, the NMR structure from both the native mTSPO and the A147T mutant show the ligand PK11195 to have predominantly hydrophobic interactions with amino acid side chains in the binding pocket. The ligand hydroxyl group points away from W143

which is in mTSPO the structural equivalent amino acid to W138 in bcTSPO, and fails to form a strong polar interaction. At the same time, the ligand flips 180º in orientation when compared between the two structural models. Structural features in the conserved regions of helix III in the NMR structure also show an 180º rotation compared to that of the bcTSPO crystal structure. However, as sequence identity between the mTSPO and bcTSPO is only ~30%. Thus, it is unclear if these structural changes and the different ligand binding mode are caused by this substantial deviation in amino acid sequence. After all, the affinity of PK11195 to mTSPO is reported to be a thousand fold higher than the affinity to bcTSPO.

3.1.5 TSPO exists in an equilibrium of oligomeric states in biological membrane

While the $mTSPO^{NMR}$ structure is monomeric in detergent systems, there are reports indicating that a fraction of mTSPO exists as oligomers in lipid bilayers [42-44]. Bacterial TSPOs are homodimers in the crystal structure with different dimerization interfaces. In the case of $bcTSPO^{X\text{-}Ray}$, helix I&II contribute to the main dimerization interface. $RsTSPO^{X\text{-}Ray}$'s main dimer interface is formed by helix I&III. mTSPO in lipid bilayer is shown to be in an equilibrium state of dimer and monomer mediated by the $^{83}GxxxG^{87}$ motif on helix III, as demonstrated by the solid state NMR studies [44]. Addition of cholesterol shifts the equilibrium towards monomeric protein, and additionally induces structural changes of mTSPO distal to the cholesterol binding site.

A manuscript from the original authors of the solution NMR structure describes the conformational state of apo-mTSPO [45]. In their findings, the apo-

mTSPO NMR spectra is partially assigned and showed pico to nano-second backbone motion of mTSPO in the absence of PK11195. Detergents of difference acyl-chain length are used to demonstrate the conformational exchange is independent of detergent system. It is also shown that apo-mTSPO lacks tertiary contact that could be observed on the NMR time scale. While the NMR structure was determined with a ligand-induced fold of mTSPO, the high temperature and the lack of a NMR titration result raises concerns of the structure's biological relevance. Understanding the conformational plasticity of mTSPO in lipid-mimetic systems is of crucial importance for investigate the binding of PET imaging ligands to mTSPO in similar methods used in the NMR structural determination.

3.1.6 TSPO structural models need additional experimental and computational validations for the use to assist drug development

Taken together, the physiological role of TSPO is still unclear. Several structural snap-shot have been revealed, that contradict each other. The biophysical and biochemical mechanism of ligand binding is not readily accessible in light of the recent structures alone. The three experimental structures of TSPO provide a starting point for understanding possible biological roles of the protein from a molecular level. However, they suffer from respective limits of technology applied in the study or their potential biological relevance to the mammalian model systems. In this study we report experimental validation of ligand binding in the reported NMR conditions. A unified structural model of mTSPO was constructed based on all three experimentally determined structures featuring its energy optimization. Such computational validation of the experimental models will be

important to support framework for future investigation. Lastly, we report the molecular interaction between imaging VUIIS8310 and mTSPO based on computational docking and NMR experiments.

3.2 Materials & Methods

Protein expression and purification

^{15}N-labeled TSPO was expressed in BL21 (DE3) *E. coli* in M9 minimal media, solubilized, and purified with DPC detergent, as adapted from previous publications [46, 47]. In short, ^{15}N labeled *m*TSPO was expressed in BL21(DE3) cells in M9 media following 1mM Isopropyl β-D-1-thiogalactopyranoside (IPTG) addition for 12 hours at 25°C. The membrane targeted *m*TSPO fraction was collected using ultracentrifugation after lysing cells by sonication. 1% (w/v) DPC was used to solubilize the membrane faction for Ni-NTA affinity chromatography. A constant concentration of 0.1% (w/v) of DPC was maintained throughout purification and *m*TSPO was eluted using 300mM imidazole. The purity of the protein was estimated on SDS-PAGE stained with coomassie blue. In case when N-Dodecyl β-D-maltoside (DDM) was used, 1% DDM (w/v) was used to solubilize the membrane fraction of mTSPO expressing E. coli and 0.1% DDM (w/v) was used during purification. The resulting sample was subjected to buffer exchange into buffer containing 25 mM MES buffer, pH 6.2, and 100 mM NaCl. Both NMR and fluorescence experiments were performed under such buffer conditions.

Tryptophan fluorescence quenching experiments of ligand binding.

A tryptophan fluorescence quenching assay was used to study the binding properties of TSPO by utilizing the intrinsic tryptophan fluorescence [48]. For the

*m*TSPO in DPC and *DDM*, 0.5 μM purified protein was titrated with increasing amounts of ligand in the desired concentration range at room temperature or 42 C°. Each titration point was monitored by a spectral scan from 300 nm to 400 nm (excitation at 290 nm) on a Horiba Jobin Yvon Fluoromax-3™ spectrofluorimeter. Control experiments were performed to evaluate the absorption/emission of ligands as well as buffers and solvents. Since mTSPO contains multiple buried and solvent exposed tryptophan residues, the area under each emission spectrum was integrated to account for all tryptophan and a percentage quenching ratio for each concentration was calculated. The binding curve was obtained by plotting the percentage quenching *%Change* against total ligand concentration [X] in mole fraction unit (Mol fraction) in Prism 6©. The actual solvent condition for membrane proteins is its lipid mimetics, in this case, the detergent. Since fluorescence and NMR experiments operates at very different protein and detergent concentrations, we used mole fraction here for the ease of comparing titration data of membrane protein cross different conditions and spectroscopic techniques. Mole fraction was calculated by dividing the amount mole of ligand by the amount of detergent in the system. Since we observed a pseudo two phase transition from raw data in fluorescence quenching experiment, Equation (1) was derived based on a two site per monomer assumption, taking into consideration ligand depletion. [*X*] is the total ligand concentration plotted on the binding curves.

(1)

$$Site1 = B_{max}^{High} \times \frac{K_{app}^{High} + E_t + X - \sqrt{K_{app}^{High^2} + {E_t}^2 + X^2 + 2K_{app}^{High} \times E_t + 2K_{app}^{High} \times X - 2X \times E_t}}{2E_t}$$

$$Site1 = B_{max}^{Low} \times \frac{K_{App}^{Low} + E_t + X - \sqrt{K_{App}^{Low^2} + {E_t}^2 + X^2 + 2K_{App}^{Low} \times E_t + 2K_{App}^{Low} \times X - 2X \times E_t}}{2E_t}$$

$$\%Change = Site1 + Site2 + Offs$$

(2) $F_{\max} = \frac{F_i - F_f}{F_i}$

%Change represents the calculated percentage quenching at each concentration, F_{max} is the maximum amplitude of fluorescence quenching, which equals the difference between the initial and final fluorescence signals divided by the initial fluorescence, [*X*] is the total ligand concentration, [Et] is the total concentration of purified *m*TSPO protein (fixed at 0.5 μM in the experiment and during the fitting procedure) and K_{app}^{High}& $K_{app}^{Low}$is the apparent dissociation constant for the two sites. *Offs* (offset) allows a fit unconstrained to go through zero. All ligands were dissolved in DMSO as a stock solution and diluted with DMSO or buffer for measurement. One representative binding curve was shown in each figure while the K_{app}^{High} & K_{app}^{Low}were reported as the average of 3 replicate measurements fitted with Equation (1).

NMR experiments of ligand binding

2D ^{1}H-^{15}N TROSY spectra of mTSPO with/without ligands were recorded on an 800 MHz NMR spectrometer at 42 °C with 0.1mM TSPO in 2% DPC (w/v), 25 mM MES buffer, pH 6.2, and 100 mM NaCl. When NMR spectra were mentioned in the manuscript, it refers to a 2D ^{1}H-^{15}N TROSY, unless specified differently. In ligand binding experiments, PK11195 was added to the sample for a

final concentration of 0.1mM TSPO with 0.1mM, 0.2mM, 0.4mM, 0.8mM, 1.6mM and 3.2mM. The ligand concentration [X] was converted to Mol fraction unit for data fitting. The Spectra were processed with NMRpipe and analyzed using Sparky. Complete backbone assignment of mice TSPO was obtained from Biological Magnetic Resonance Data Bank (http://www.bmrb.wisc.edu/ BMRB entry: 19608). The backbone ^{1}H and ^{15}N assignments were transferred to our mTSPO-PK11195 spectrum with most of the peaks aligning well with the published data, except for central region of the spectrum where overlapping peaks made unambiguous assignment difficult (Figure S 3 D, E).

The partially assigned mTSPO-PK11195 spectrum were then used for identifying resonance peaks in each concentration. The spectral changes upon ligand binding are indicative of a typical slow-exchange process. Thus, the titration was monitored by integrating peak intensity from assigned resonances belonging to mTSPO-PK11195 bound species. The binding curve was obtained by plotting the percentage intensity change *%Change* (calculated by dividing the intial peak intensity at [X] by max peak intensity at highest ligand concentration) against total ligand concentration [X] in mole fraction unit (Mol fraction). Since we observe a one phase transition from raw data in NMR titration experiment, Equation (3) was derived based on a one site per monomer assumption, taking into consideration ligand depletion. [*X*] is the total ligand concentration plotted on the binding curves. The binding curve is plotted by pooling intensities for ten residues surrounding the ligand binding site (W5, Y34, W42, W47, G72, F100, W107, S116, V118, W155).

(3)

$$\%\boldsymbol{Change} = \boldsymbol{B_{max}} \times \frac{\boldsymbol{K_{app}} + \boldsymbol{E_t} + \boldsymbol{X} - \sqrt{\boldsymbol{K_{app}}^2 + \boldsymbol{E_t}^2 + \boldsymbol{X^2} + 2\boldsymbol{K_{app}} \times \boldsymbol{E_t} + 2\boldsymbol{K_{app}} \times \boldsymbol{X} - 2\boldsymbol{X} \times \boldsymbol{E_t}}}{2\boldsymbol{E_t}}$$

$+\mathbf{Offs}$

In the case of VUIIS8310 ligand binding experiment, 2D HSQC spectra of mTSPO with/without 5mM VUIIS8310 were recorded on an 800 MHz NMR spectrometer at 42 °C with 0.1mM TSPO in 2% DPC (w/v), 25 mM MES buffer, pH 6.2, and 100 mM NaCl. The partially assigned mTSPO-PK11195 spectrum from previous NMR experiments was leveraged for identifying resonance peaks in the mTSPO-VUIIS8310 spectrum (Figure 11A). In short, residues with resonances unshifted in mTSPO-VUIIS8310 compared to mTSPO-PK11195 were marked as no shift; residues with resonances shifted in mTSPO-VUIIS8310 but traceable from near assigned mTSPO-PK11195 resonances were marked either significant or insignificant chemical shift change (ΔCS) based on a cut-off of 0.05 ppm in the proton dimension; residues with resonances completely disappeared in mTSPO-VUIIS8310 were marked significant ΔCS. The Resulting chemical shift difference mapping was used for identify potential VUIIS8310 binding site on TSPO.

RosettaCM modeling of mTSPO

The $mTSPO^{NMR}$ structural ensemble (PDB: 2GMY), bcTSPO crystal structure (PDB:4RYI), $rsTSPO^{X\text{-}Ray}$ crystal structure (PDB:4UC2) was energy optimized using Rosetta fastrelax protocol [49] to produce $mTSPO^{NMR\text{-}opt}$, $bcTSPO^{X\text{-}Ray\text{-}opt}$ and $rsTSPO^{X\text{-}Ray\text{-}opt}$ models. RosettaCM [50] was then used to construct 2000 homology models based on either 5 structures of best energy

optimized bcTSPO$^{X\text{-}Ray}$ and rsTSPO$^{X\text{-}Ray}$ or a combination of 5 bcTSPO$^{X\text{-}Ray\text{-}opt}$, 5 rsTSPO$^{X\text{-}Ray\text{-}opt}$ and top 10 mTSPO^{NMR} ensemble models. The top 1% models with favorable energy from each condition were subjected to an additional round of energy optimization. After this stage, all models were evaluated, whereby only the top 1% models were selected for a final energy optimization. The best energy mTSPO^{RosettaCM} model is used to plot Rosetta energy versus root mean square distance (RMSD) plot using R. All structural comparisons of the mTSPO^{RosettaCM} model to experimentally determined structures was done using Dali pairwise structural alignment [51]. Chemical shift calculations from constructed models were carried out using the Sparta+ server [52]. NOE distance deviations were calculated on a per residue basis. The atomic distances between the respective long range NOE restraint from the models were calculated to compare with NOE restraint values downloaded from BMRB. Proline dihedral angles were calculated for a database of membrane proteins downloaded from PDB to generate Ramachandran plot for the Pro-Pro motif. The Pro-Pro dihedrals of the top 20 models from mTSPO$^{NMR\text{-}opt}$ and top 20 models from mTSPO^{RosettaCM} were plotted on the graph.

Computational ligand docking in mTSPO binding pocket

PK11195 was computationally docked into the bcTSPO$^{X\text{-}Ray\text{-}opt}$, the mTSPO$^{NMR\text{-}opt}$ and the mTSPO^{RosettaCM} structure using RosettaLigand [53]. VUIIS8310 [37] was computationally docked into the mTSPO$^{NMR\text{-}opt}$ structure and the mTSPO^{RosettaCM} model. The Ligand conformers were generated using MOE [54] (Molecular Operating Environment; Chemical Computing Group Inc.,

Montreal, QC, Canada). In the case of docking PK11195, the ligand was placed into the binding pocket, superimposing with the experimentally determined PK11195 coordinates. For each setup, 1000 docking models were constructed with 5Å translational and 360° rotational sampling [53, 55]. In the case of docking VUIIS8310, the ligand was placed into the protein center based on proximity to residues determined from NMR experiments. Here, 2000 docking models for $mTSPO^{RosettaCM}$ and 2000 for $mTSPO^{NMR\text{-}opt}$ were constructed with 5Å translational and 360° rotational sampling. Resulting models were ranked based on their predicted binding energy using the Rosetta energy function that contains terms for van der Waals attractive and repulsive forces, electrostatic interactions, hydrogen bonding, solvation, and likelihood of particular side-chain conformations. The top scoring docked model is used for visualizing ligand-protein interface in PyMol [56]. The final model is reported after manual inspection for agreement with NMR chemical shift data.

3.3 Results

3.3.1 TSPO in DPC micelles shows highly dynamic behavior

We recombinantly expressed mTSPO, purified it, and reconstituted it into DPC micelles, the same system that was used to determine the $mTSPO^{NMR}$ structure. The purity of mTSPO was checked using SDS-PAGE (Figure S 3A). Equilibrium ligand binding experiment were performed to demonstrate mTSPO ligand binding function in micelle. The amide peaks in the NMR 2D ^{15}N-^{1}H correlation spectrum are sensitive to the local chemical environment, thus could be viewed as fingerprints of actual protein structure. The NMR spectra of the apo

mTSPO in DPC micelles displayed very narrow peak dispersion in line with literature [22] (Figure S 3 C), typical of a highly dynamic protein or protein in the molten globule state.

3.3.2 NMR Titration of PK11195 into mTSPO preparation in micelles shows significantly lower binding affinity than membrane preparation

Increasing concentration of PK11195 were added into mTSPO in NMR conditions as specified in methods. The chemical shift changes showed a typical slow-exchange, whereby the peaks belonging to the apo-protein diminished with the increase concentration of the ligand, and peaks belonging to the PK111955 bound species appeared in the 2D TROSY spectrum (Figure S 3 D, E). The changes in peak intensity with increasing concentration of PK11195 showed a single transition of ligand binding response (Figure 7 Purple line). The lower limit of the apparent binding constant (Kapp) calculated from ligand titration was 0.0057 (Mol fraction) (Table S 3), while PK11195 was reported to bind mTSPO with nanomolar affinity (which would be equivalent to ~10^{-7} Mol fraction K_d, assuming a ligand to lipid/detergent ratio of 1 to 100). A decrease of intensity for some peaks was also observed at high concentration of PK11195, possibly due to aggregation or changes in the dynamic of the protein due to excess amount of ligand. The very high apparent K_d could be a result of changes in the energetics of the protein at high temperature accompanied by the effect of detergent system

An intrinsic tryptophan fluorescence quenching assay was then used to complement the NMR binding assay. The ligand-induced quenching of intrinsic protein tryptophan fluorescence was used to monitor ligand binding (Figure S 3B).

Different concentrations of DPC were tested for mTSPO-DPC system at room temperature (25 °C) or the temperature used for NMR studies (42 °C). mTSPO purified in DDM detergent (mTSPO-DDM) was also included in the set of experiments to monitor possible detergent effects apart from concentration and temperature. In both DDM and DPC detergent systems, a two-phase transition curve was observed when titrating PK11195 (Figure 7 A-C). Without prior assumption of possible non-specific binding, we fitted the data with a two-site per monomer model for binding. For all samples of mTSPO-DPC at 42 °C, the high affinity apparent binding constant component (K_{app}^{high}) for DPC was around ~10^{-6} Mol fraction and independent from the DPC concentration; while the low affinity apparent binding constant component ($\boldsymbol{K_{app}^{Low}}$) in the second transition was at 0.0068 Mol fraction for 0.2%DPC, 0.0054 Mol fraction for 0.4%DPC and 0.012 Mol fraction for 0.1% DPC. This second transition in the fluorescence titration experiments correlated with the transition observed with NMR spectroscopy (Figure 7 A). A set of experiment at room temperature was performed for mTSPO-DPC to access temperature dependencies of the observed transitions (Figure 7 B). The $\boldsymbol{K_{app}^{Low}}$ transition appeared to shift slightly with half order of magnitude (Table S 3). The $\boldsymbol{K_{app}^{High}}$ transition for 0.4%DPC 25 °C displayed a significant shift, however the exact reasons could not be attributed. The mTSPO-DDM sample was prepared for 42°C experiments to account for detergent dependencies of the observed transitions (Figure 7 C). In this case, The $\boldsymbol{K_{app}^{Low}}$ component of titration data from mTSPO-DDM showed two order of magnitude increase (Table S 3). The $\boldsymbol{K_{app}^{Low}}$ component changed only slightly.

The fluorescence binding results demonstrate that mTSPO-DPC binds PK11195 in a similar manner as seen in NMR studies, where the quenching event would suggest a binding and reorganizing of the structure of mTSPO during the low affinity apparent binding constant (K_{app}^{Low}) transition. In NMR experiments, such a ligand binding induced reorganization of the mTSPO structure was seen as a slow-exchange transition of resonance peaks. High temperature affected the ligand binding behavior significantly in DDM micelles, suggesting the DDM micelles might not be suitable to maintain structural integrity at high temperature. The high affinity apparent binding constant (K_{app}^{High}) transitions seems to be temperature and detergent sensitive. We hypothesize that it represents an initial binding of PK11195 to a preformed partial ligand binding pocket at low ligand concentrations. As the concentration of PK11195 is increased and this preformed partial ligand binding pocket constantly occupied, the structure of mTSPO is rearranged, the protein folds, and the complete binding pocket is formed.

3.3.3 The experimentally TSPO structures might be determined in perturbed system

The available structural models of TSPO obtained from experimental data have several limitations (Figure 8 A): The $mTSPO^{NMR}$ structure was determined using micelle systems at high temperature where the energy landscape is substantially perturbed. Additionally, some features of the NMR structure conflict with conformations observed in sequence-conserved regions of bacterial homologs, specifically in helix III. At the same time, these bacterial homologs ($bcTSPO^{X\text{-}Ray}$ and $rsTSPO^{X\text{-}Ray}$), although determined in a more native-like

environment, have less than 40% of sequence identity to mTSPO. The sequence divergence could prevent them being directly used to model structural interactions of small molecule probes, imaging agents or therapeutics.

3.3.4 An Integrated Comparative Model of mTSPO is biophysically realistic and reconciles conflicting structural information from multiple templates

To better sample the energy landscape of TSPO while including structural features of experimental structures, we decided to construct homology models of mTSPO using a single bacterial structure as template ($bcTSPO^{X\text{-}Ray}$ or $rsTSPO^{X\text{-}Ray}$, resepctively) and multiple templates ($bcTSPO^{X\text{-}Ray}$, $rsTSPO^{X\text{-}Ray}$, and $mTSPO^{NMR}$) using RosettaCM. The sequence alignment from Figure S 4 was used to generate the initial threading template. We also performed energy optimization on $mTSPO^{NMR}$ structural ensemble to computationally evaluate its energy landscape. The $mTSPO^{RosettaCM}$ model produced from multiple template homology modeling displayed the lowest energy amongst other strategies after 2 rounds of selection based on energy of generated models (Figure 8 B). The $mTSPO^{NMR\text{-}opt}$ structural model has a higher energy compared to that of the homology models from single and multiple templates. When plotting the average energy of the top 20 models obtained from each sampling strategy, $mTSPO^{RosettaCM}$ models showed clearly favorable energies compared to that of the $mTSPO^{NMR\text{-}opt}$ or models obtained from a single bacterial template (Figure 8 C). This suggests that the NMR structure when removed from the micelle and embedded in a biological model membrane is in a high energy, i.e. biophysically unrealistic state with many local unfavorable interactions. Similarly, simple homology models of mTSPO based on

single bacterial template structures are energetically frustrated suggesting that major rearrangements are needed to accurately describe the conformation of mTSPO in a biological membrane. Only if structural features form all three models are combined using RosettaCM, the energy of mTSPO gets in the expected range of high-resolution experimental structures of helical membrane proteins (Figure S, to be added). The final mTSPO RosettaCM model was compared to original template structures using structural alignment tools (Figure 8 D). The model shared the same global fold with all templates, although differ from $mTSPO^{NMR}$ model by 4.1Å, $bcTSPO^{X\text{-}Ray}$ by 3.3Å and $rsTSPO^{X\text{-}Ray}$ by 3.5Å.

3.3.5 $mTSPO^{RosettaCM}$ model shares structural features with all three experimentally determined structural templates

In contrast to the $mTSPO^{NMR}$ structure the $mTSPO^{RosettaCM}$ has a Rosetta energy that is biophysically realistic in a membrane model. Globally, the model resembled the conformations observed in the bacterial proteins more closely. Hence, we wondered if this model is still generally consistent with the restraints obtained by NMR spectroscopy, keeping in mind that the NMR restraints are valid in the thermodynamically destabilized micelle conformation of the protein. RosettaCM modeling was conducted in an implicit membrane model. We expected that the overall fold as well as secondary structure are well conserved between mTSPO models in a micelle and in a membrane. We expected slightly larger deviations in local features of the tertiary structure such as bending or rotation of transmembrane segments.

3.3.6 mTSPO^{RosettaCM} model agrees with experimental NMR chemical shift data

Backbone chemical shifts from CA and CB atoms are sensitive to secondary structure and local structural perturbations. To access the agreement of mTSPO^{RosettaCM} model with the experimental NMR data, the back-calculated backbone chemical shifts (CS) were computed from the mTSPO^{RosettaCM} and mTSPO$^{NMR\text{-}opt}$ using Sparta+ [52]. The chemical shift of CA and CB were then compared with the experimentally determined backbone chemical shift. The mTSPO^{RosettaCM} model agrees with the experimental CS data as good as the mTSPO$^{NMR\text{-}opt}$ structure (Figure 9). Note, that CS were not input into the RosettaCM modeling protocol as restraints but only indirectly provided through the mTSPO^{NMR} template. This result indicates that secondary structure and many local structural features are consistent between the mTSPO^{NMR} and mTSPO^{RosettaCM} structures.

Since the mTSPO^{NMR} structure was calculated from experimentally determined NOE distance restraints, we also calculated the per residue NOE deviation for both the mTSPO^{RosettaCM} model and mTSPO$^{NMR\text{-}opt}$ model (Figure S 5 A). Figure 10 displays the per residue distance deviations. As expected, mTSPO$^{NMR\text{-}opt}$ model showed very small deviation from the experimental data (Figure 10 A). In the case of RosettaCM model, residues in helix III, IV, and V adopted alternative backbone and side conformations and deviate from the experimental data more significantly (Figure 10 B). The majority of the deviations

between mTSPO^{RosettaCM} and the NOE distance restraints occurs in the top part of the protein close to PK11195 binding site, while other parts of mTSPO^{RosettaCM} fulfill the majority of long-range NOE restraints. We interpret this finding as consistent with our experimental observations: mTSPO is destabilized in micelles and the PK11195 binding pocket only fully forms at ligand concentrations sufficiently high to induce protein folding. The induced tertiary structure in this region is the most stable conformation in micelles but deviates from the low energy conformation in a membrane.

3.3.7 Ligand binding pocket of mTSPO^{RosettaCM} shares similarity to that of the bacterial homologues

The ligand binding pocket in the core of mTSPO^{RosettaCM} structure shows a larger similarity to bcTSPO$^{X\text{-}ray}$ (Figure 10 C) compared to mTSPO^{NMR} (Figure 10 D). mTSPO^{RosettaCM} has transmembrane helix III in a conformation close to that of bcTSPO$^{X\text{-}ray}$. Helices III, IV, and V of mTSPO^{RosettaCM} are in a different backbone conformation compared to mTSPO^{NMR}. Specifically, the conserved residues W53 and W143 packing the core of the protein adopt similar conformations that point to the core of the protein as observed in bcTSPO$^{X\text{-}ray}$. Of W93 and W95, which are evolutionary not conserved, only W95 points towards the core of the protein in mTSPO^{RosettaCM} (Figure 10 D).

There are three Pro-Pro motif in mTSPO: the ^{44}Pro-^{45}Pro is sequence conserved, while the ^{96}Pro-^{97}Pro, ^{131}Pro-^{132}Pro is not conserved through evolution. ^{96}Pro-^{97}Pro is located at the partially unfolded helix III in the mTSPO NMR structures. In the mTSPO^{RosettaCM} model, the ^{96}Pro-^{97}Pro motif adopts a slightly

perturbed but still helical conformation. To test, if this is a likely conformation for Pro-Pro motifs in membrane proteins, we plotted the ψ,φ angles in a Ramachandran plot for all Pro-Pro motifs in the transmembrane regions of membrane proteins with known structure (Figure S 5B). The ^{96}Pro-^{97}Pro motif for both mTSPO^{RosettaCM} model and NMR structure agrees with the dihedral angles of other experimentally determined Pro-Pro motifs.

At last, docking PK11195 with RosettaLigand [53] into bcTSPO$^{X\text{-}ray}$ yielded a low energy binding model similar to the ligand binding pose observed experimentally (Figure S 7A). Docking of PK11195 into the binding pocket of mTSPO^{NMR_optimized} failed to converge to the NMR observed ligand pose (Figure S 7B). Docking of PK11195 into mTSPO^{RosettaCM} (Figure S 7C) converged to a low energy ligand conformation differs from bcTSPO$^{X\text{-}ray}$ and mTSPO^{NMR}. It is possible that after rearrangement of binding pocket in these RosettaCM model after homology modeling, the ligand binding pocket for PK11195 is in a conformation that does not favor either observed ligand conformation from experimentally determined structure.

3.3.8 VUIIS8310 interaction with mTSPO from NMR and computational ligand docking studies

Chemical shift information from 2D N-H correlation NMR spectrum reflects local chemical environment of backbone amide atoms. When comparing two separate NMR spectra of two distinct ligands binding to the protein, one would observe ΔCS arises from presence and absence of ligand and the chemical shift for residues in purple overlap might expect either a change or no change based on

the particular chemical environment. While PK11195 binding to mTSPO produces a chemical environment unique to its binding site and its set of NMR resonances, other ligands that binds to a slightly different site would results in a different chemical environment and producing a slightly different set of chemical shifts. When comparing the two spectra, the differences in chemical shift could be results of two phenomena: the difference in chemical moiety of ligand and difference in protein conformation. Ligand binding site mapping based on resonance shift (ΔCS) in NMR is conceptualized as following (Figure 11 A): Red circle and its dashed circle represents a molecule's influence to its surrounding chemical environment thus producing distinct resonance peaks; Blue circle and its dashed circle represents a molecule's influence to its surrounding if binding at a different site and producing distinct resonance peaks; If the NMR spectrum for the Red circle sample is assigned, and the NMR spectrum of Blue circle is not assigned, one can still deduce the approximate ligand binding site of Blue circle by comparing ΔCS in NMR spectra of Blue circle to Red circle.

Based on this principle mTSPO-VUIIS8310 interaction was studied using NMR spectroscopy. Comparing the spectra of mTSPO-PK11195 to the spectra of mTSPO-VUIIS8310 revealed many differences in the chemical shift of residues in the binding pocket (Figure S 6). Mapping the residues with and without ΔCS onto the $mTSPO^{RosettaCM}$ model revealed initially a deeper binding site for VUIIS8310 (Figure 11 B right) compared to that of PK11195 (Figure 11 B left). The VUIIS8310 might make specific interaction with W53, W95, W143, and A147. Computational ligand docking experiments were performed on VUIIS8310 to mTSPO CM Model

and energy optimized NMR structure (Figure S 7 D, E). While both structures showed convergence to low energy conformers, the energy optimized NMR structure docked with VUIIS8310 (Figure S 7 F) could not fulfil all ΔCS of the residues. VUIIS8310 docked to mTSPO^{RosettaCM} model (Figure 11 B right enlarged panel) showed pi-pi stacking of W53 to the pyrazolopyrimidine ring and potential for the flexible chemical groups on VUIIS8310 to interact with A147. The docking data can be used for optimizing imaging ligand development.

3.4 Discussions

3.4.1 Current micelle preparations of mTSPO display reduced ligand affinity, likely through a thermodynamic destabilization of the protein

mTSPO solubilized in detergent micelles displayed lowered binding affinity to PK11195 compared to previously reported data of nano-molar affinity. Based on titration results from NMR and fluorescence experiments, the ligand induced NMR spectra changes and the second transition of tryptophan quenching curve is a combined effect of ligand binding to mTSPO and induced structural changes that affects global fold of protein. The high affinity transition observed in fluorescence experiment is sensitive to temperature and detergent conditions, thus is likely to be binding of ligand to either pre-formed ligand binding pocket in the apo mTSPO or detergent micelles. Taking both fluorescence and NMR data into account, we suggest that mTSPO in micelles at 42 °C is not fully folded in its biologically relevant three-dimensional structure. It has only a partially pre-formed ligand binding pocket that allows for an initial, fast association of PK11195 with the protein. As the concentration of PK11195 is increased, its high binding energy is

sufficient to induce mTSPO folding. The resulting tertiary structure adopts the biologically relevant fold of the protein but is still locally perturbed because of the thermodynamically destabilized and altered energy landscape in the micelle environment.

We concluded that the binding of PK11195 induces stabilization effect of mTSPO is a combined effect of ligand binding and induced folding event. In detergent conditions explored by previous NMR experiments at high temperature and at our hands, mTSPO might exist in a ligand binding capable, altered high-energy state. This high-energy state, indicated by the highly dynamic characteristics in apo mTSPO NMR spectrum, appears to deviate from that of mTSPO in mitochondrial membrane extract where protein is in lipid and other protein co-factors. It is important to note that our experimental results appear not to invalidate the NMR results, as the conformational stability of apo protein does not interfere with its final ligand-bound states. The binding of ligands such as PK11195 could induce the folding of mTSPO to a conformation that reflect physiologically realistic ligand binding despite of the altered energy landscape.

In the recently published solid state NMR studies on mTSPO in DMPC liposomes, apo mTSPO also displayed dynamic characteristics [44]. Resonance assignment and further experiments in solid state NMR could only be achieved by complexing mTSPO with a ligand in a similar fashion as in the case of detergent micelles. This would reflect that at least in vitro, mTSPO alone might exist in a dynamic, ligand binding capable state. mTSPO is found to be in complex with VDAC, ANT possibly other proteins in the MPTP [9, 10].

In the current study of mTSPO, all determinant of protein function is based on the capability of binding to the ligand while ignoring the energy state of the protein might be altered or other proteins in the MPTP complex being absent from the system. Mitochondrial membrane is also asymmetric in its lipid composition and membrane potential. It is unclear whether these conditions would contribute to the normal function of TSPO, i.e. the energy state of mTSPO in its apo state. Future experiments must be able to establish the validity of mTSPO-ligand complex being physiologically realistic or the binding phenomenon is trackable to literature values.

3.4.2 The remodeled mTSPO structure is a viable model for structural studies

We presented the mTSPO^{RosettaCM} model as an alternative conformation that might exist for studying ligand binding to TSPO proteins. While all experimentally determined TSPO structures provide valuable insight to this important protein, certain features in the experimental structures should be carefully examined. Computational modeling to energetically optimize TSPO structures is able to sample a larger conformational equilibrium and achieve lower energy ensembles for mTSPO^{RosettaCM} model. The resulting structure, while resembling a fold closer to that of the bacterial TSPOs, still agrees with NMR chemical shift data experimental collected on mTSPO.

We also employed computational docking experiment to access the models' potential for future structure based ligand screening. The mTSPO^{RosettaCM} model was able to produce meaningful ligand conformation for VUIIS8310, that agrees

with NMR binding data. Imaging ligand targeting TSPO seems to be binding to the core of the protein [46, 57] and often interact with all five TM helices. These ligands might also alter the conformation of mTSPO or cause functional changes of the protein. Thus computational models that provide a larger sampling area for ligands is important for future structure based ligand development.

The It is worth noting that in all cases of ligand docking, the energy optimized NMR structure fails to produce data that can be interpreted along with our experimental supplement. Docking of PK11195 into bcTSPO$^{X\text{-}ray\text{-}opt}$ clearly showed the model can accurately generate ligand ensembles close to the observed crystal structure. It is possible that mTSPO^{NMR} structure is over-optimized in a condition that fulfills all structural restraints. The resulting conformation is in a local energy minimum in the respective computational energy evaluation that is not viable for other computational modeling software to carry forward.

3.4.3 Mechanism of ligand binding and influences on function of TSPO

The functional role of TSPO in biological processes still needs to be further investigated in light of the recent shift in paradigm of TSPO's involvement in cholesterol biosynthesis pathways. While this ancient protein has been proposed to be involved in many alternative processes in terms of oxidative stress sensing, fatty acid metabolism, the molecular basis of TSPO's function has never been tackled from a bottom up approach. Despite the controversies surrounding its biological function, TSPO still remains to be a highly anticipated target for imaging molecules. It is our intention to refine TSPO structures for a better basis to provide

computational models for future development of potent TSPO ligands. While the specific ligand pocket configuration might change regarding different ligands, the ligand binding of different molecules seems not to induce global conformational shift of the protein and maintains a 5 helices bundle for TSPO.

We have demonstrated that second generation TSPO ligand binds at a deeper pocket in mTSPO, where it would be interacting with A147. This piece of information is important for determining the molecular mechanism of rs6971 (A147T) polymorphism interfering the potency of these TSPO ligands [58, 59]. In human, the rs6971 polymorphism results in three genotypes and its associated binding phenopypes for the A147 in TSPO: C/C (Ala/Ala) and high binding affinity, C/T (Ala/Thr) and mixed binding affinity, and T/T (Thr/Thr) and low binding affinity. Clinical assessment for T/T binding is difficult for separating specific binding to non-specific bindings in human subjects. Using our computational model and NMR studies, we could access ligand binding information at A147 position. Future direction characterizing this important polymorphism site would include TSPO ligands of different scaffold and sensitivity to this site for systematic assessment of all ligands molecular mechanisms of binding.

It is unclear how binding of ligand would mechanistically interfere with the function of TSPO. It seems that ligand-bound mTSPO retains the cholesterol binding and oligomerization potential [44] in lipid. It is difficult to incorporate cholesterol molecules in DPC micelles used in solution NMR studies. Future experimental implementation should consider screening various conditions where incorporation of cholesterol can be achieved along with TSPO ligands.

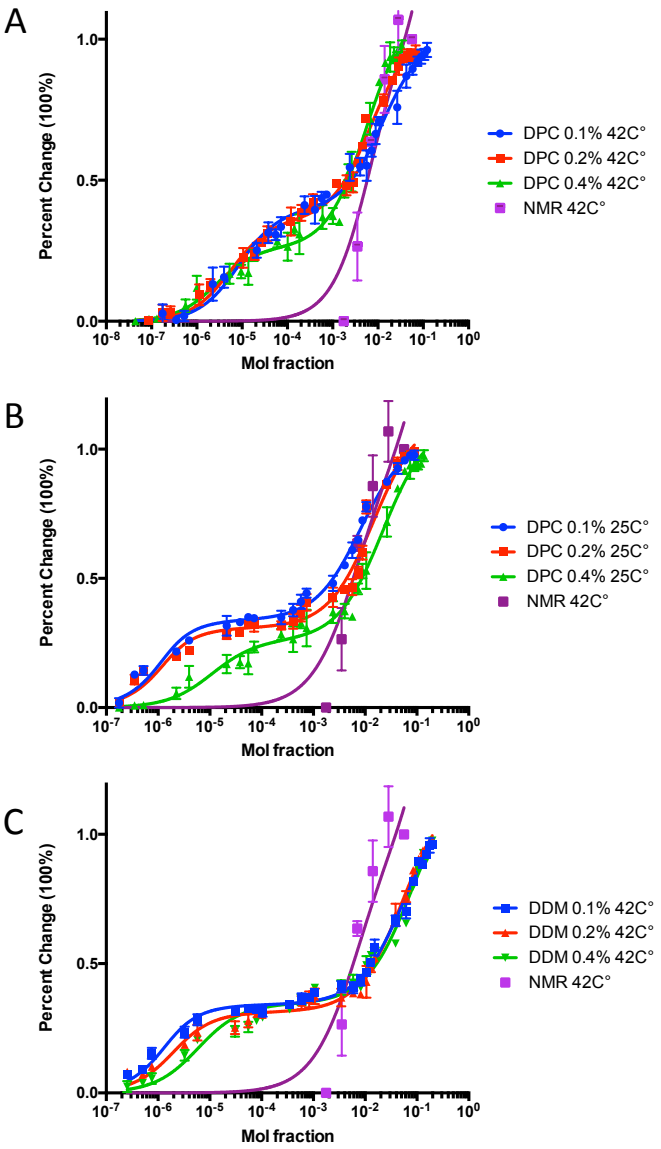


Figure 7. Fluorescence quenching and NMR ligand binding experiment of mTSPO in different conditions. X-axis is plotted as Mole fraction unit, and Y-axis is ploted as percent change to account for difference sources of data to be comparable at the same level. The titration curve obtained from NMR experiments is plotted in all panels as reference comparison. (A) Fluorescence quenching and NMR ligand titration for PK11195 and mTSPO in 0.1% (w/v), 0.2% (w/v), 0.4% (w/v), NMR condition (~2% w/v) DPC at 42C°. (B) Fluorescence quenching and NMR ligand titration for PK11195 and mTSPO in 0.1% (w/v), 0.2% (w/v), 0.4% (w/v), NMR condition (~2% w/v) DPC at 25C° (C) Fluorescence quenching and NMR ligand titration for PK11195 and mTSPO in 0.1% (w/v), 0.2% (w/v), 0.4% (w/v) DDM and NMR condition (~2% w/v) DPC at 42C°. Error bars are based on triplicate experiment of the same condition. Data fitting is performed as described in methods.

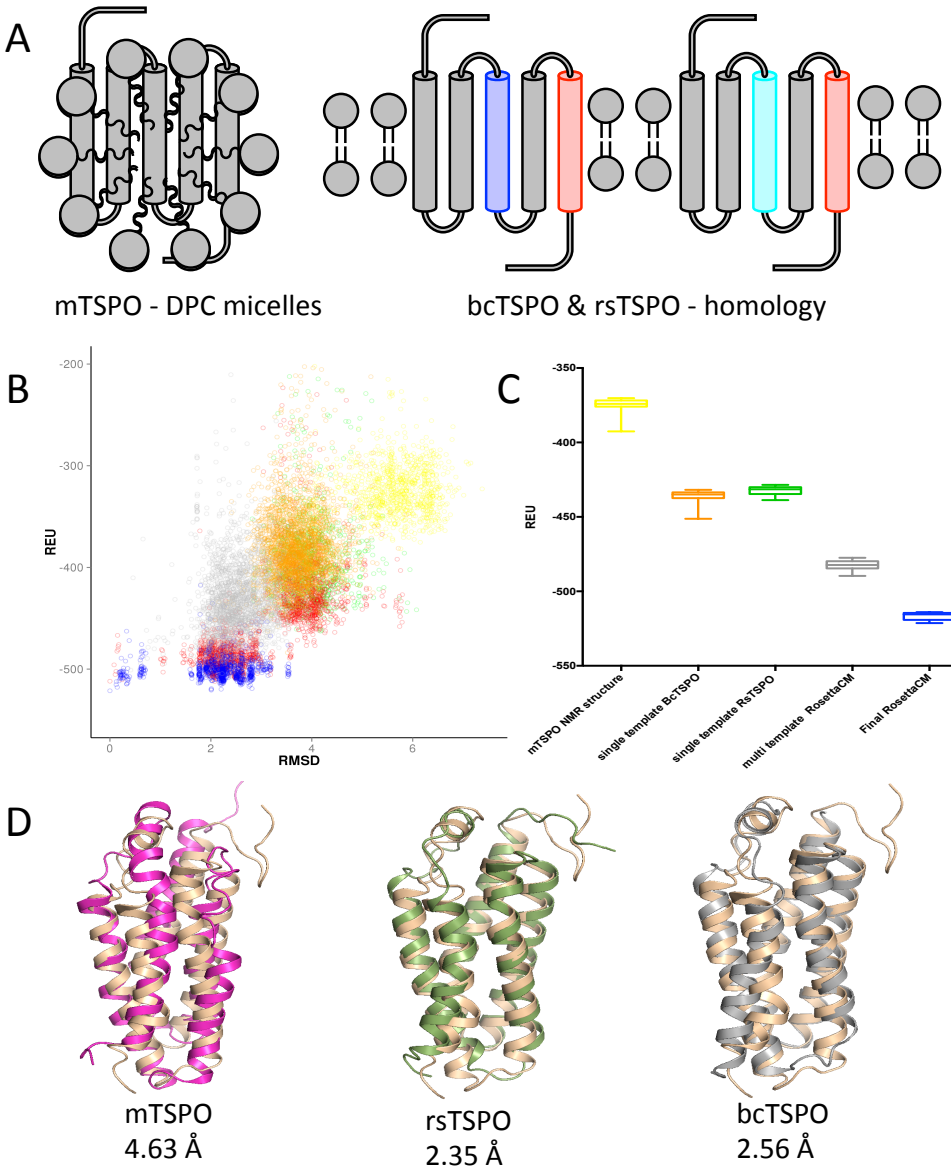


Figure 8. Remodeled mTSPO CMmodel structures have favorable energy. (A) Schematics of the current available high resolution structural data: mTSPO NMR structure utilizes micelles, while bacterial TSPO has low sequence homology to mammalian protein. (B) Energy (in REU) versus RMSD plot between the final model of mTSPO and all the models produced in each steps of the pipeline. Each point in the plot represents one model's RMSD to the final CMmodel and its respective energy. Yellow: energy optimized mTSPO NMR structure; Orange: mTSPO homology model based on bcTSPO; Green: mTSPO homology model based on rsTSPO; Gray: mTSPO homology model based on bcTSPO, rsTSPO and mTSPO using RosettaCM; Red: models after 2 rounds of energy optimization of models from above pipelines; Blue: final models from RosettaCM after another round of energy optimization. (B) Bar plot of the energy of top 20 models produced in each pipeline shows the mTSPO CMmodel has a superior energy (color scheme follows that of panel A). (C) Structural comparison of CMmodel to the experimentally determined structural models based on sequence independent alignment. mTSPO CMmodel is depicted in brown ribbon, the RMSD to bcTSPO (Gray), rsTSPO (Green), mTSPO (Magenta) are reported.

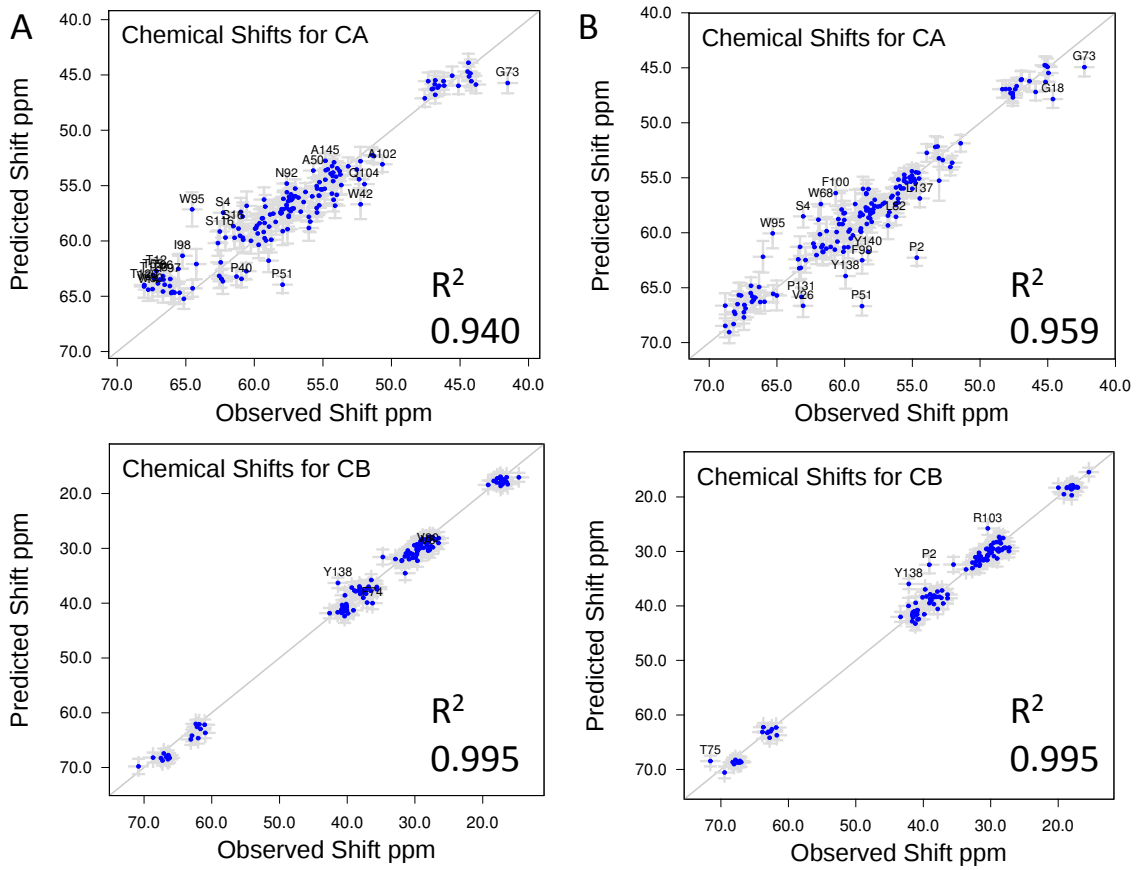


Figure 9. Back calculated chemical shift of the mTSPO CMmodel correlates with observed chemical shift from NMR data. (A) Back calculated chemical shift for CA, CB from experimental mTSPO NMR structures correlates with the observed NMR data. (B) Back calculated chemical shift for CA, CB from mTSPO CMmodel correlates with the observed NMR data. R^2 value is calculated and noted in the graph.

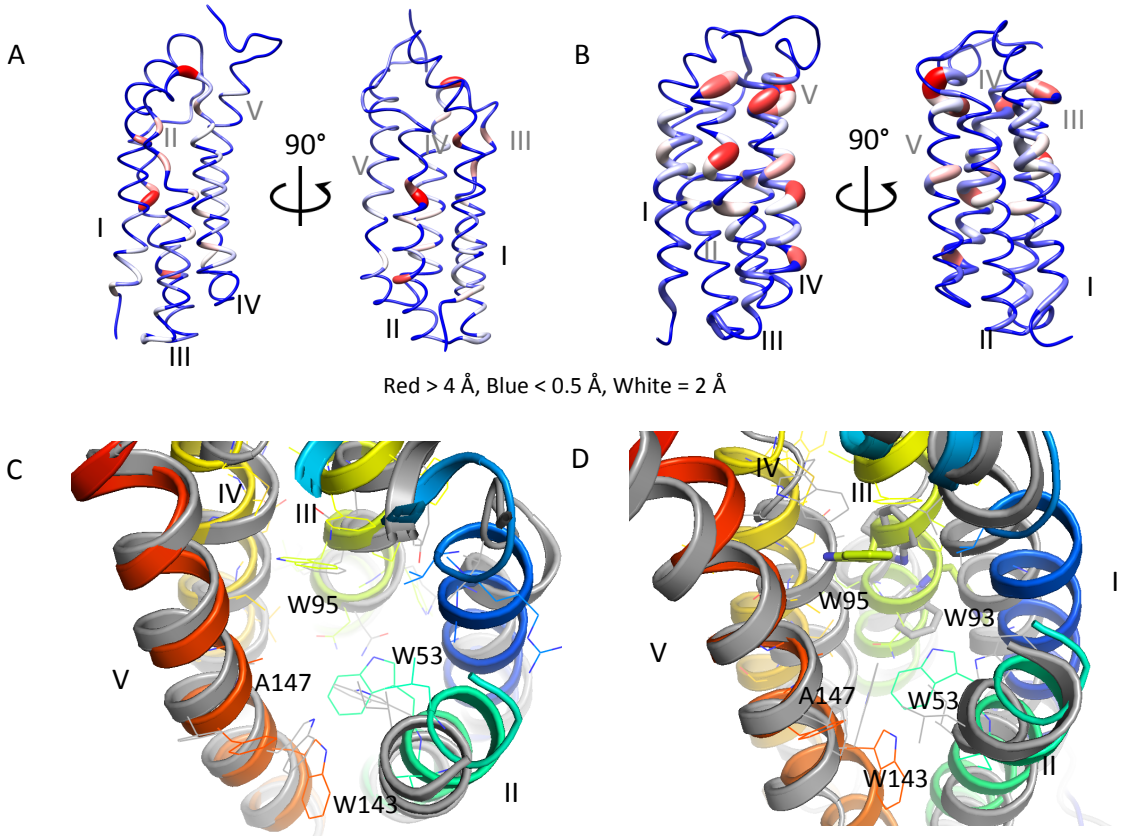


Figure 10. Energy optimized experimental mTSPO NMR structure and mTSPO CMmodel mostly agree with experimentally measured long-range NOEs, but differ in ligand pocket configuration. (A) The per residue agreement of experimental NOE derived distances and model of energy optimized NMR structure are color coded to the backbone trace of the model with distance deviation higher than 4Å coded red, lower than 0.5Å coded blue and middle ground of 2Å coded white. The diameter of backbone trace is also scaled to the deviation. (B) The per residue agreement of experimental NOE derived distances and model of mTSPO CMmodel are depicted with the same scheme as panel A. (C) The ligand binding pocket of mTSPO CMmodel resembles that of the bcTSPO, where conserved residues are also structurally conserved. (D) Comparing the ligand binding site of mTSPO CMmodel to the NMR structure shows different backbone trace for helix three and four, where the partially unfolded helix three adopts alpha helical conformation.

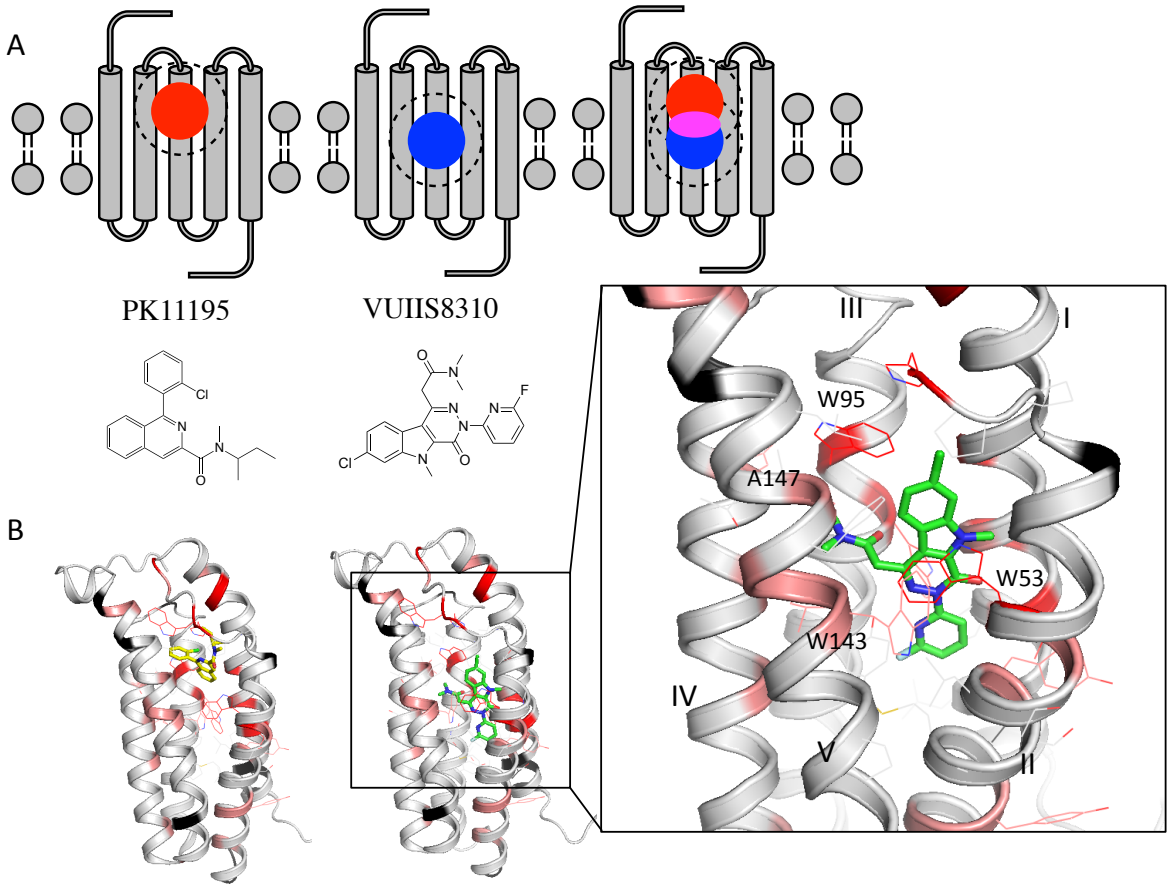


Figure 11. The Ligand binding interactions are mapped for VUIIS8310 on mTSPO. (A) Concept of ligand binding mapping based on differences of chemical shift (ΔCS) in NMR: Red circle and its dashed circle represents a molecule's influence to its surrounding chemical environment; Blue circle and its dashed circle represents a molecule's influence to its surrounding; When comparing ligand induced chemical shift of two distinct ligand binding event the net chemical shift changes is a combination of the two. The molecular representations of PK11195 and VUIIS8310 are also shown. (B) mTSPO CMmodel with mapped chemical shift differences arise from VUIIS8310 (Black: no ΔCS; Gray: insignificant ΔCS or unassigned; Pink: moderate ΔCS; Red: large ΔCS). Left: PK11195 docked to mTSPO CMmodel. Right: VUIIS8310 docked to mTSPO CMmodel at a deeper site compared to PK11195. The enlarged picture of VUIIS8310 binding site is depicted at the side with residues interacting with the ligand marked

3.6 Supplemental Materials

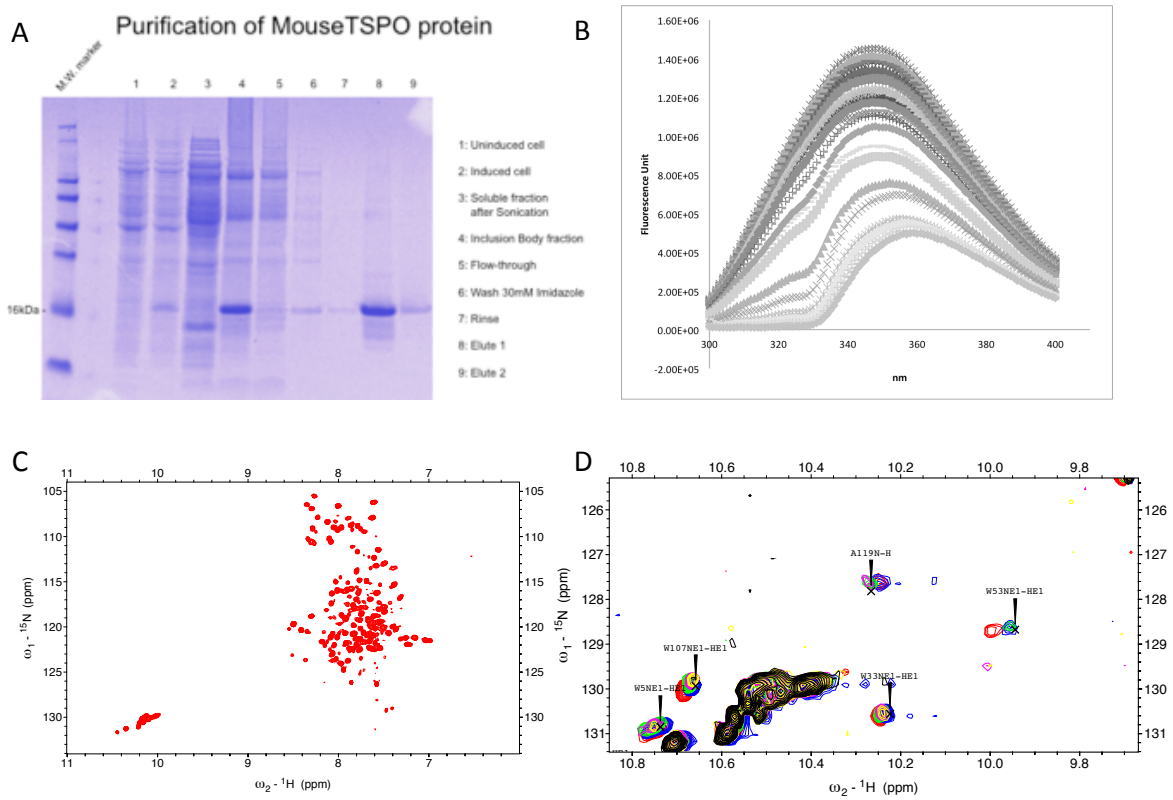
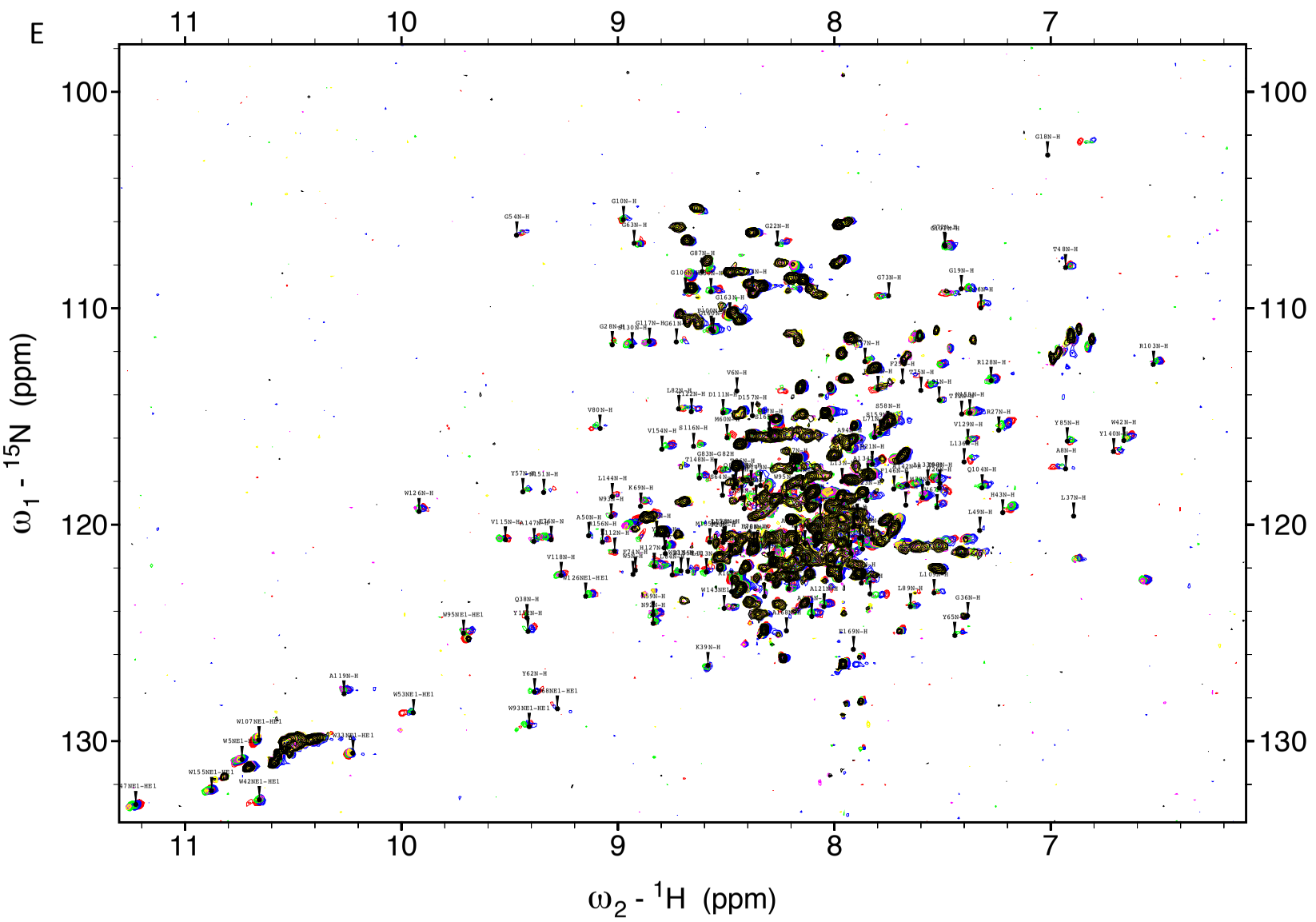


Figure S 3 mTSPO protein purification and spectroscopic data of ligand titration. (A) SDS-PAGE gel of mTSPO purification. The eluted sample were above 95% purity. (B) Sample raw data of fluorescence titration of PK11195 into mTSPO. In all PK11195 experiments, ligand does not completely abolish tryptophan spectra even at very high concentration. (C) 2D 1H-15N HSQC spectrum of apo mTSPO. (D) 2D 1H-15N HSQC NMR titration for PK11195 and mTSPO overlaid (zoom-in) (Black: 0.1mM Yellow: 0.2mM Magenta: 0.4mM Green: 0.8mM Blue: 1.6mM Red: 3.2mM). (E) 2D 1H-15N HSQC NMR titration for PK11195 and mTSPO overlaid (full plot) in the same color scheme as (D).


E
11
10
9
8
7
100
110
120
130
ω_1 - ^{15}N (ppm)
ω_2 - ^{1}H (ppm)

Sample	K_{app}^{High}	K_{app}^{Low}	K_{app}
DPC 0.1% 25C	5.79E-07	8.30E-03	
DPC 0.2% 25C	6.33E-07	1.46E-02	
DPC 0.4% 25C	1.05E-05	2.19E-02	
DPC 0.1% 42C	8.39E-06	1.23E-02	
DPC 0.2% 42C	5.60E-06	6.87E-03	
DPC 0.4% 42C	2.66E-06	5.47E-03	
DDM 0.1% 42C	7.62E-07	7.85E-01	
DDM 0.2% 42C	1.45E-06	8.66E-01	
DDM 0.4% 42C	5.42E-06	8.60E-01	
N M R 4 2 C			>5.74E-03

Table S 3 Apparent binding constant of PK11195 to mTSPO in various conditions. The values reported in the table are in Mol Fraction unit. The fluorescence binding data were fitted with equation (1) to report K_{app}^{High} and K_{app} with prior assumption of a high affinity binding event. The NMR binding experiment data was fitted with equation (2) to report a lower limit of K_{app} for apparent binding constant.

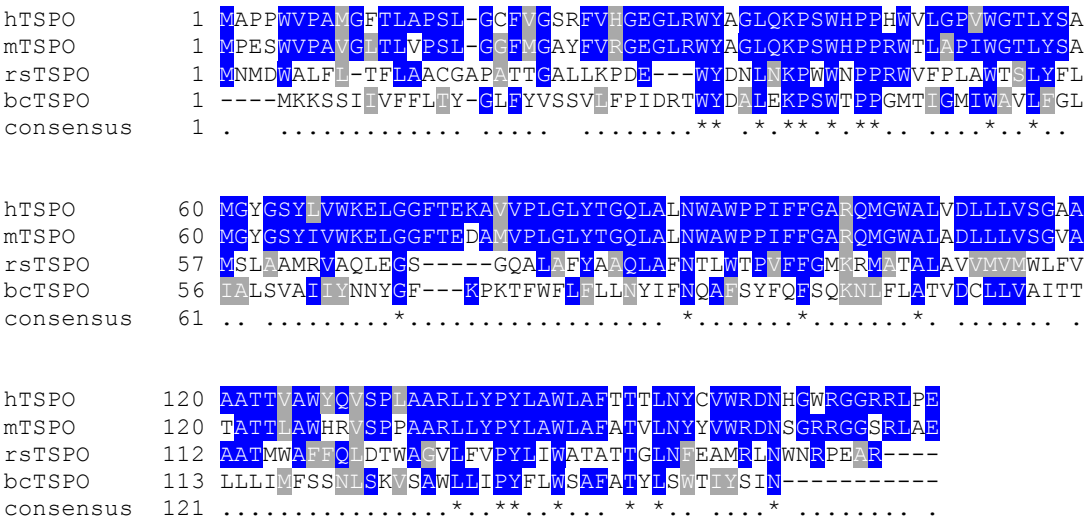


Figure S 4. Multiple sequence Alignment of mammalian TSPO and bacterial homologues. The identical amino acids are colored blue, while similar amino acids in the aligned are colored grey. This sequence alignment was converted to pairwise sequence alignment in Grishin format for RosettaCM.

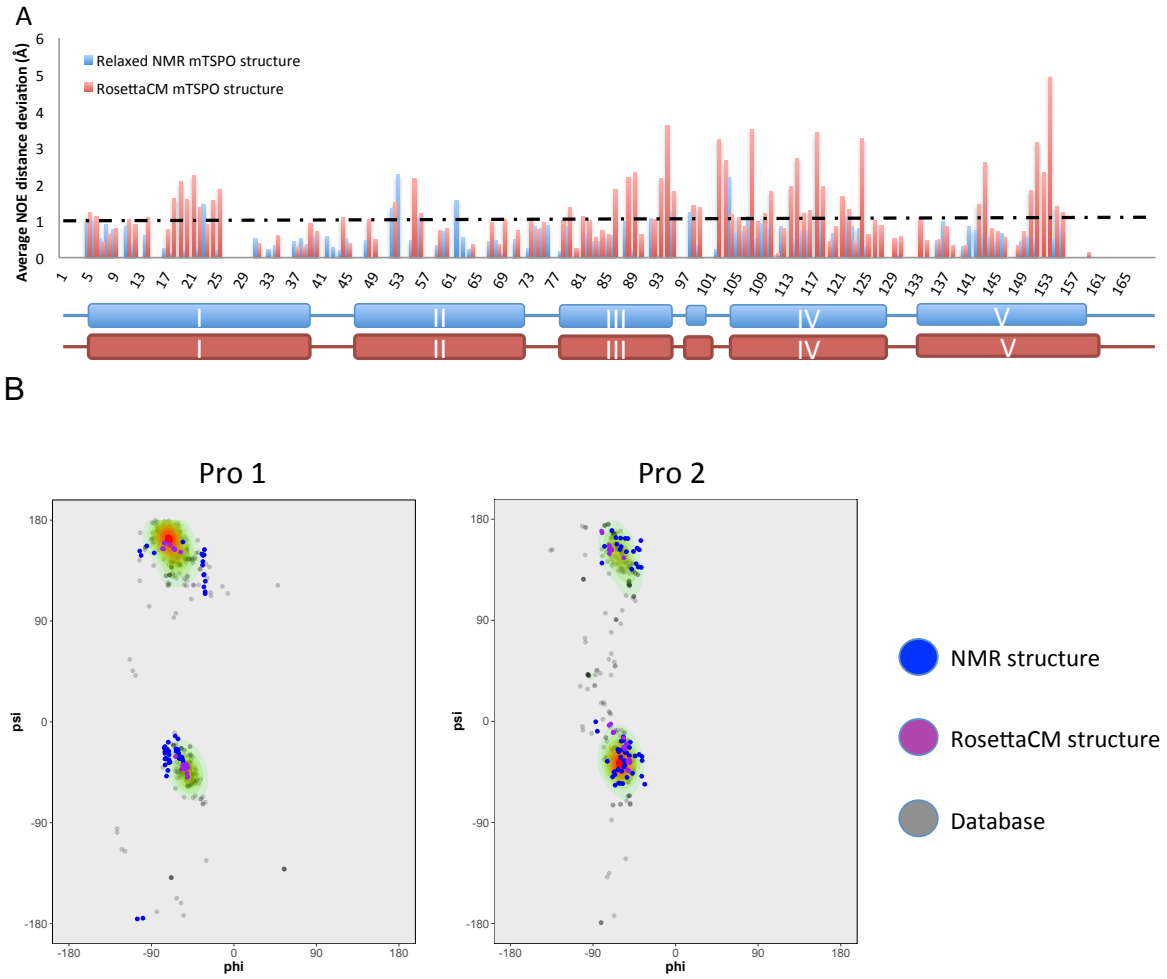


Figure S 5. NOE distance deviations and proline conformation deviations of the mTSPO$^{NMR\text{-}opt}$ structure and mTSPO^{RosettaCM} model. (A) The average per residue NOE distance deviation (Å) was plotted against the amino acid sequence, with transmembrane helix regions noted below the x-axis. The per residue NOE distance deviation was calculated by averaging the total distance deviations for a residue's backbone and sidechain NOE by the number of NOE distance restraints. The dashed line was determined at 1Å to indicate manual cut-off. (B) The pro-pro dihedral angles of were plotted in the Ramachandran plot. The leading proline in sequence was termed Pro1 and the trailing proline was termed Pro2. All pro-pro motifs in the transmembrane region of available PDB structures were taken to construct the database plot in gray and contour graph. The pro-pro dihedrals from top 20 mTSPO$^{NMR\text{-}opt}$ structures were colored blue, while the pro-pro dihedrals from top 20 mTSPO^{RosettaCM} model were colored purple.

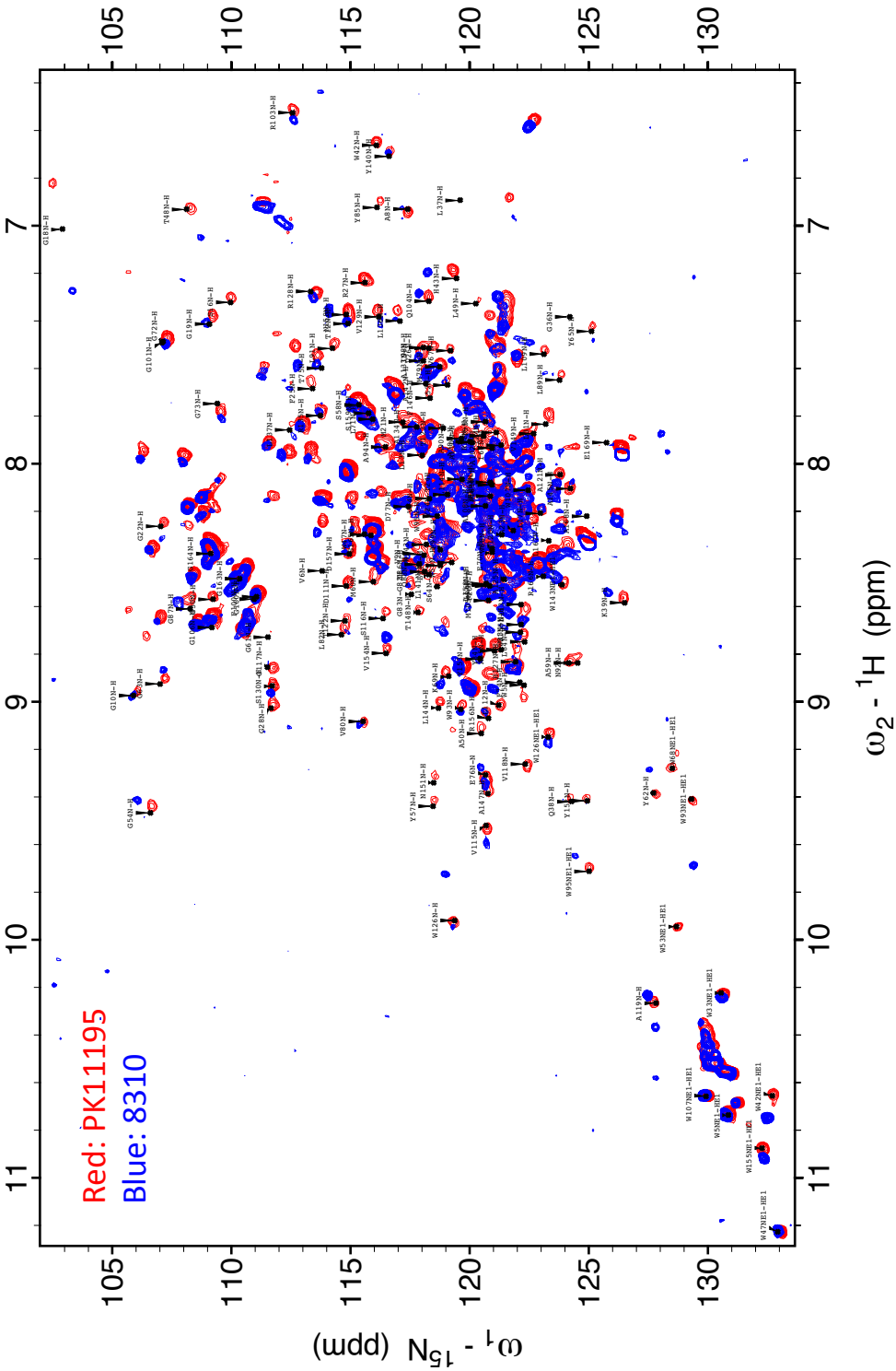


Figure S 6 2D ^{1}H-^{15}N HSQC spectrum overlay of mTSPO in presence of 5.0 mM VUIIS8310 (Blue) and mTSPO in prescence of 5.0 mM PK11195 (Red).

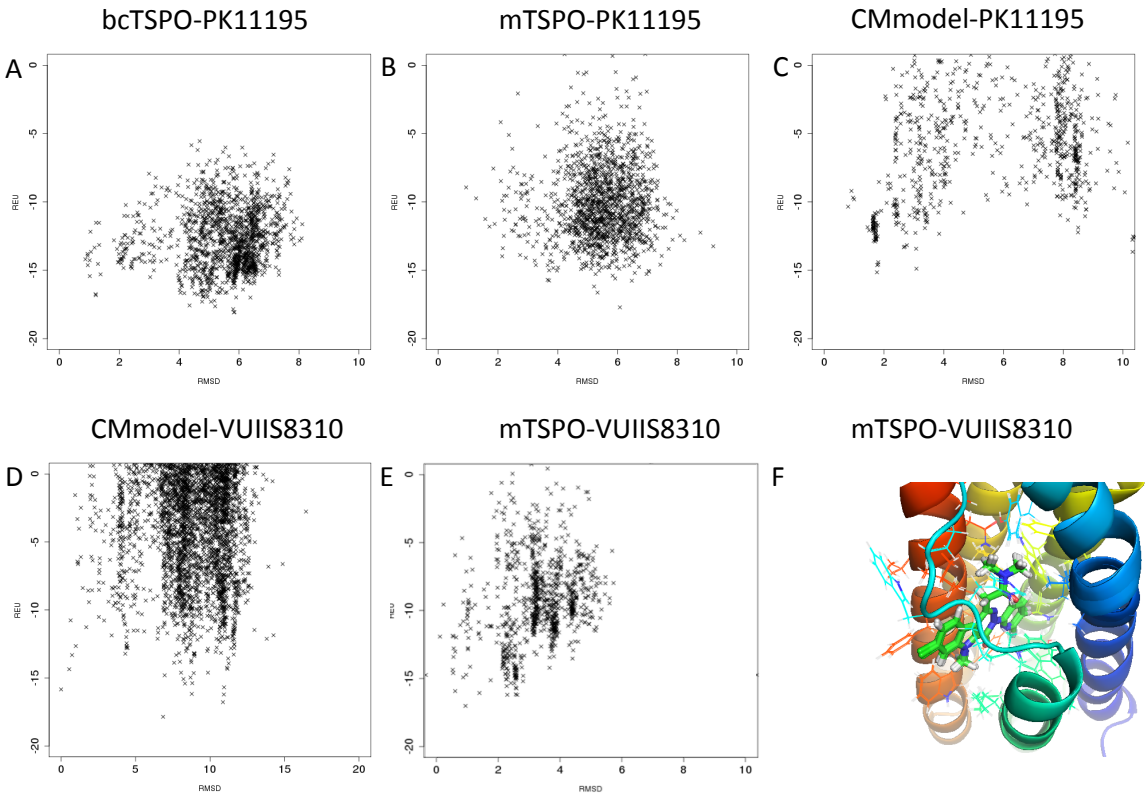


Figure S 7 Computational ligand docking of ligands to TSPO proteins. The Energy (in REU) versus RMSD plots were used to indicate energy funnels of ligand docking or deviation of ligand conformation from the experimentally determined ligand pose when available. In the case where experimentally determined ligand-protein complex is available, the ligand RMSD values in the x-axis were calculated based on deviations to the experimentally determined structural models. (A) Energy vs RMSD plot for PK11195 docking into energy optimized bcTSPO crystal structure showed a cluster of ligand conformations similar to experimentally determined ligand pose. (B) Energy vs RMSD plot for PK11195 docking into energy optimized mTSPO showed no obvious convergence to a ligand pose close to the observed NMR structure complex. (C) Energy vs RMSD plot for PK11195 docking into mTSPO CMmodel (RMSD calculated based on lowest energy ligand, since differences in protein conformation between CMmodel and NMR model is large, especially in the ligand binding pocket) showed convergence to a cluster of low energy ligand conformations. (D) Energy vs RMSD plot for VUIIS8310 to CMmodel (RMSD calculated based on lowest energy ligand) showed two major clusters. NMR data could manually filter the other ligand conformation. (E) Energy vs RMSD plot for VUIIS8310 to the energy optimized NMR model (RMSD calculated based on lowest energy ligand) showed some convergence to two low energy ligand conformers. (F) The lowest energy conformers were visualized, the ligand placement was at membrane-protein interface with ligand pointing towards membrane, the interactions between ligand and CMmodel appeared not to agree with NMR data.

CHAPTER 4 ELUCIDATING PROTEIN-LIGAND INTERFACE

This chapter presents and discusses several projects on interface modeling of membrane proteins based on experimental data. Computational modeling in protein-protein or protein-ligand interface can be useful in rationalizing molecular mechanisms of biological processes, understanding crucial structural feature of binding interactions, analyzing energy landscape of the interface of interest, etc. Application of computational modeling technologies to advance the understanding of biological problems is an important component of my thesis. The application projects in the chapter includes the TSPO project where I personally collected experimental data, and solely computational modeling work on GPCR.

The chapter begins with the computational modeling work on class C GPCR $mGlu_5$ based on an intermediate resolution crystal structural model. It is the first project I completed on modeling membrane protein-ligand interface and homology modeling of large membrane protein complexes using Rosetta modeling suite. The project serves as a proof of principle case where computational modeling could accurately predict ligand binding energy and ligand efficacies based on crystal structure of $mGlu_5$. Modeling of the full-length $mGlu_5$ is the first case of a full-length class C GPCR model constructed to elucidate molecular mechanisms of receptor activation.

The second part of the chapter is based on imagining ligand development work in collaboration of the Manning lab. TSPO is proposed to be a potential target for imagine molecules for many types of cancer and neuro-inflammation. While the

ligand described in the manuscript VUIIS1009A & B showed differential binding characteristics in *in vivo* PET imagining results from rats and mice, our NMR characterization was able to demonstrate both ligands binds at the same site on TSPO with similar binding behavior. We also mapped the ligand binding site of DPA714, VUIIS1009A & B on mTSPO based on NMR chemical shift data. This work can be used for future structure based ligand optimization and development.

4.1 An introductory project into modeling: $mGlu_5$, a Class-C GPCR

Chapter4.1 contains partial reproduction of "Wu H, Wang C, Gregory KJ, Han GW, Cho HP, **Xia Y**, Niswender CM, Katritch V, Meiler J, Cherezov V, Conn PJ, Stevens RC. Structure of a class C GPCR metabotropic glutamate receptor 1 bound to an allosteric modulator. Science. 2014 Apr 4; 344(6179):58-64". Reprinted with permission from AAAS. I contributed to the section of text describing determinant of subtype selectivity within common allosteric site and communication between the extracellular domain and transmembrane domain. I edited other section of the paper, and produced Figure 12 Figure 13Figure 14Figure 15. The manuscript describes the first crystal structure of a Class-C GPCR and discusses potential implications of the allosteric modulator binding and receptor activation through interpretation of crystal structure models in combination of in vivo assays and computational modeling. The computational modeling work in the project advances the understanding of crucial ligand interaction to downstream signaling, and showed the potential of integrated structural biology to understanding of biological problems.

4.1.1 Introduction

The human G protein-coupled receptor (GPCR) superfamily is composed of over 800 seven transmembrane (7TM) receptors that can be divided into four classes based on their sequence homology: class A, B, C, and Frizzled (F) [167]. The fifteen Class C GPCRs play important roles in many physiological processes such as synaptic transmission, taste sensation and calcium homeostasis, and include metabotropic glutamate receptors (mGlu), γ-aminobutyric acid B receptors ($GABA_B$), calcium sensing receptor (CaS), taste 1 receptors (TAS1), as well as a few orphan receptors. A distinguishing feature of class C GPCRs is constitutive homo- or hetero-dimerization mediated by a large N-terminal extracellular domain (ECD) (Figure 12A). The ECDs within homodimeric receptors (mGlu and CaS) are cross-linked via an intermolecular disulfide bond. The heterodimeric receptors ($GABA_B$ and TAS1) are not covalently linked, but their heterodimerization is required for trafficking to the cell surface and signaling [168]. The ECD of class C GPCRs consists of a Venus flytrap (VFT) domain, which contains the orthosteric binding site for native ligands (Figure 12A), and a cysteine-rich domain (CRD), except for $GABA_B$ receptors. The CRD, which mediates the communication between ECD and 7TM domains, is stabilized by five disulfide bridges, one of which connects the CRD and VFT [169].

The mGlu family was the first group of class C GPCRs to be cloned [170, 171]. Comprised of eight members, the mGlu family can be separated into three subgroups [172], termed groups I ($mGlu_1$ and $mGlu_5$), II ($mGlu_2$ and $mGlu_3$) and III ($mGlu_{4,6,7,8}$), based on their sequence homology, G protein coupling profile, and

pharmacology [173]. Group I mGlus are predominantly coupled to $G_{q/11}$ and activate phospholipase C_β, which hydrolyses phosphoinositides into inositol 1,4,5-trisphosphate (IP_3) and diacylglycerol, inducing intracellular calcium mobilization and activating protein kinase C (PKC).

The group I mGlus, $mGlu_1$ and $mGlu_5$, are considered promising therapeutic targets for the areas of cancer, pain, schizophrenia, Alzheimer's disease, anxiety, autism and others [173, 174]. However, the development of subtype-selective small molecule ligands that might serve as drug candidates for these receptors has been hampered by the conservation of the orthosteric (glutamate) binding site (Figure 12A). This problem can be overcome by the utilization of allosteric modulators that act at alternative binding sites; these compounds bind predominantly within the 7TM domain for the class C receptors. Allosteric modulators can alter the affinity or efficacy of native ligands in positive, negative, and neutral ways, demonstrating a spectrum of activity that cannot be achieved by orthosteric ligands alone.

In this study, we report the crystal structure of the human $mGlu_1$ 7TM domain bound to a negative allosteric modulator (NAM), 4-fluoro-*N*-(4-(6-(isopropylamino)pyrimidin-4-yl)thiazol-2-yl)-*N*-methylbenzamide (FITM) [175] at 2.8 Å resolution [176]. This structure provides a 3D framework for understanding the molecular recognition and facilitating the discovery of allosteric modulators for the mGlu family and other Class C GPCRs. It also complements crystallographic studies of the transmembrane domain structures of class A [177, 178], B [179, 180] and F [181] GPCRs and extends the knowledge base upon which to study the

diversity and evolution of the entire GPCR superfamily.

4.1.2 Materials and Methods

Computational ligand docking in FITM binding pocket

FITM and five derivatives reported by Satoh et al [175] were computationally docked into the $mGlu_1$ using RosettaLigand [59]. The Ligand conformers were generated using MOE [182] (Molecular Operating Environment; Chemical Computing Group Inc., Montreal, QC, Canada). The ligands are placed into the binding pocket, superimposing the large common substructure with the experimentally determined FITM coordinates. 1000 docking models per compound were constructed with 2Å translational and 10° rotational sampling [59, 166]. We limited sampling to retain the experimentally determined FITM binding mode for all derivatives. Resulting models were ranked based on their predicted binding energy using an energy function that contains terms for van der Waals attractive and repulsive forces, electrostatic interactions, hydrogen bonding, solvation, and likelihood of particular side-chain conformations. The ligand interface energy of the top 1% scoring models was used to compute the predicted ensemble average ligand binding energy and to correlate with experimental IC_{50}. The predicted per residue binding energy was extracted from top 1% models where negative values indicate residues that contribute favorable interactions, positive values indicate residues that contribute unfavorable interactions, zero values indicate residues not involve in ligand binding. The top scoring docked model is used for visualizing ligand-protein interface in PyMol.

Full length model building

The ECD of $mGlu_1$ was constructed by homology modeling of the glutamate-bound dimeric VFT domain structure from $mGlu_1$ (PDB ID: 1EWK) and CRD from $mGlu_3$ (PDB ID: 2E4U) using Rosetta ([47, 58]). The model was constructed imposing a perfect 2-fold rotational symmetry ([183]). For this purpose, the template 1EWK which is in a putative Aco (active closed-open) conformation was converted into a symmetric Acc (active closed-closed) conformation before the CRDs were added. To bring the CRDs into intimate contact [184] the loop region between Q513 and V523 and the C254-C543 disulfide bond were remodeled to allow the relative orientation of CRD and VFT to be sampled. The ECD was placed with its C-termini P580 within 4Å distance to I581, the N-terminal residue of the 7TM domain, imposing C2 symmetry. The loop region E579 to P582 was reconstructed to link CRD and 7TM domain.

4.1.3 Overall structure of the $mGlu_1$ 7TM domain

The human $mGlu_1$ 7TM domain (residues 581-860), complexed with FITM, was crystallized by the lipidic cubic phase method using the thermostabilized apocytochrome b_{562}RIL (BRIL) N-terminal fusion strategy (*10*). A series of *in vitro* pharmacological studies were performed to verify that this truncated construct binds FITM and was functional in G protein coupling. The structure was solved using a 4.0 Å single wavelength anomalous dispersion (SAD) dataset collected from a single crystal soaked with tantalum bromide cluster, followed by extending the resolution to 2.8 Å using native data collected from multiple crystals.

The $mGlu_1$ 7TM domain forms a parallel dimer in each asymmetric unit with

a dimer interface mediated mainly through helix I (Figure 12B). Interestingly, we observed six well-resolved cholesterol molecules packed against hydrophobic residues on the extracellular side of helices I and II, mediating the dimer formation. The extracellular loop (ECL) 2 adopts a β-hairpin conformation, pointing to the extracellular space, which was also observed in many peptide class A GPCRs [185, 186]. This β-hairpin is connected to the top of helix III through a disulfide bond (C657-C746) that is conserved through all classes of GPCRs. The $mGlu_1$ NAM, FITM, binds within a pocket formed by the 7TM bundle close to the extracellular side (Figure 12 B and C), in the region where ligand binding is observed for some class A GPCRs. The intracellular loop (ICL) 1 forms an ordered helical turn, while a large part of ICL2 (residues 688-695 in molecule A and residues 689-693 in molecule B) is missing in the structure due to the long and presumably flexible nature of this loop. ICL3 is well resolved and forms a short link connecting the intracellular ends of helices V and VI. In addition, we did not observe helix VIII, reported in most class A GPCR structures, as well as in class B and F. Instead, C terminal residues (844-860 in molecule A and 847-860 in molecule B) are missing in the $mGlu_1$ structure, indicating that this region can be disordered.

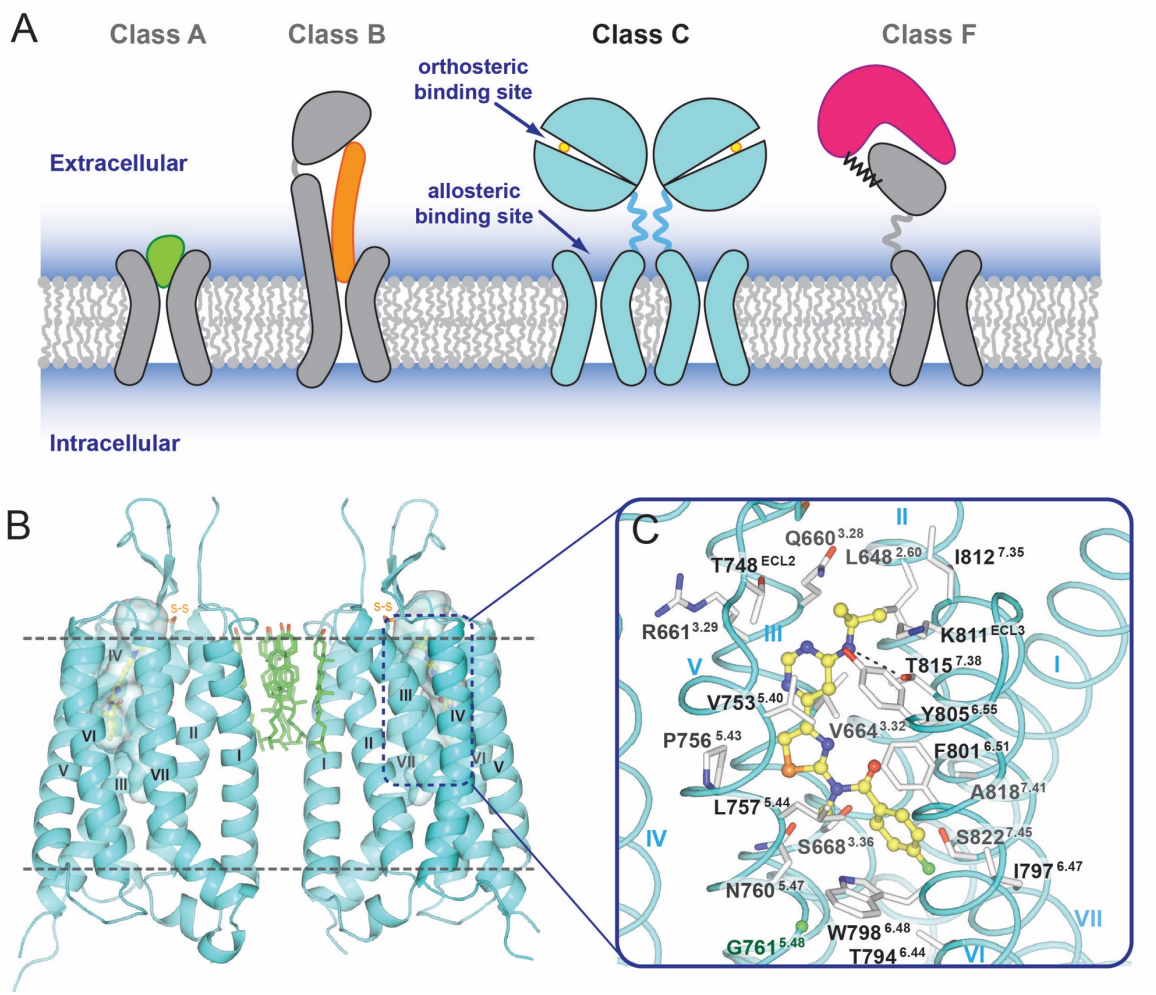


Figure 12**.** Overall structure of the $mGlu_1$ TM domain. (**A**) Cartoon models for structure and endogenous ligand recognition in different GPCRs classes. For class A, in most cases, the endogenous ligand (shown in green) is recognized by an orthosteric site in the 7TM domain. For class B, the endogenous peptide ligand (shown in orange) binds to both ECD and 7TM domains. For class C, the endogenous small molecule ligands (shown in yellow circle) are recognized by orthosteric sites in the VFT domains. For class F, lipoprotein WNT (shown in magenta) binds the CRD domain of Frizzled receptors. (**B**) The $mGlu_1$ 7TM domain crystallized as a parallel dimer shown in cyan cartoon. Cholesterols mediating the dimer interface are shown as green carbons. The surface representation of the cavity embedded in the 7TM domain is shown in transparent palecyan. (**C**) Side chains of the FITM binding pocket residues are shown as white carbons. Hydrogen bond interaction between the NAM and $T815^{7.38}$ is shown as a dashed line. The C_α carbon of $G761^{5.48}$ is shown as a green ball. In (B) and (C), the ligand FITM is shown as yellow carbons.

4.1.4 Determinants of subtype selectivity within the common allosteric site

Previous mutagenesis studies have proposed at least one common allosteric site for the mGlu family within the 7TM domain. The $mGlu_1$ binding pocket for FITM (Figure 12C) largely corresponds to this proposed common allosteric site in mGlus and probably other class C GPCRs. Despite the evidence that binding of various chemotypes of class C GPCR allosteric modulators involve similar residue positions, many mGlu modulators display a high degree of subtype selectivity, including FITM which shows high affinity (K_i = 2.5 nM) and selectivity for $mGlu_1$ over $mGlu_5$ [175]. Examination of the contact residues in the binding pocket reveals only four residues of $mGlu_1$ that differ from $mGlu_5$: $V664^{3.32}$, $S668^{3.36}$, $T815^{7.38}$, and $A818^{7.41}$, all of which have previously been implicated in subtype selectivity by mutagenesis-based studies [187-189]. Therefore, we mutated three out of four residues to their corresponding amino acid in $mGlu_5$ in comparison to the wide-type (WT) full-length human $mGlu_1$ (Figure 13 A). Methionine substitution of $T815^{7.38}$ (Figure 13 B) had the most profound effect, reducing FITM affinity ~6 fold and decreasing negative cooperativity with glutamate. This indicates that $T815^{7.38}$ is a key selectivity determinant for FITM, a result that is not surprising given the observed polar interaction between $T815^{7.38}$ and the ligand in the structure. Mutation of $S668^{3.36}$ and $A818^{7.41}$ had little effect, although there was a trend for reduced FITM affinity at $S668^{3.36}P$.

In addition, we assessed mutations known to influence the allosteric modulation of other mGlu subtypes that had not previously been explored in $mGlu_1$. $T794^{6.44}A$ and $S822^{7.45}A$ had no effect on FITM; while $P756^{5.43}S$ significantly

reduced FITM affinity (~ 3 fold) as well as negative cooperativity (Fig. 3C). Based on its location in the binding pocket, we predict that a P756$^{5.43}$S mutation may induce conformational changes in the backbone altering the geography of the binding pocket in relation to the thiazole core of FITM. Interestingly, multiple $mGlu_5$ modulator scaffolds are known to be sensitive to mutations of two non-conserved residues, S$^{6.39}$ and A$^{7.46}$ [166, 188, 190-194]; neither is observed here as contributing to the FITM binding pocket. However, both of these residues contribute to a small pocket separated from that occupied by FITM by the side chain of Y672$^{3.40}$. Given that the S$^{3.36}$ in $mGlu_1$ is replaced by P$^{3.36}$ in $mGlu_5$, it is conceivable that the unique proline induced kink in helix III of $mGlu_5$ can significantly change the shape of the pocket, making S$^{6.39}$ and A$^{7.46}$ of $mGlu_5$ accessible to ligands.

To further improve our understanding of the critical ligand-receptor interactions for FITM binding within the pocket, we docked a selection of FITM analogs (Figure 13 D-G and Figure 14) [175] into the crystal structure. Re-docking FITM (Figure 14 A-C) and analyzing the binding energy contribution per residue (Figure 14 D) revealed that T815$^{7.38}$ is forming an energetically favorable hydrogen bond with FITM (Figure 14 C and D). Compound 17 lacks not only the hydrogen bond with T815$^{7.38}$, but also misses a non-polar interaction with L648$^{2.60}$ (Figure 13 D and E); however, a potential hydrogen bond between Q660$^{3.28}$, which is not observed in FITM binding, may compensate for this loss and account for the retained activity at 10 nM. The 3-pyridyl analog (compound 14, Figure 13 D and E) lacks this potential interaction with Q660$^{3.28}$, accounting for its further decreased

potency (230 nM). Furthermore, compound 28 exhibits approximately 10-fold lower potency and differs from FITM by the introduction of a methyl group to the amine on the pyrimidine ring. Docking compound 28 reveals a major energy penalty that arises from the loss of a polar interaction and the introduction of steric clash with T815$^{7.38}$ (Figure 13 D and F). The binding pose of compound 28 also lacks a polar interaction with Q660$^{3.28}$ and requires movement of T815$^{7.38}$ and Y805$^{6.55}$ to accommodate the methyl group (Figure 13 F). Compound 22 (Figure 13 D and G), which contains an iso-propyl group on the amide linker, requires movement of two residues in helix V (P756$^{5.43}$ and L757$^{5.44}$) to fit in the pocket, and also lacks hydrogen bonding capacity with either T815$^{7.38}$ or Q660$^{3.28}$, accounting for its reduced (micromolar) potency. Collectively, by comparing the binding poses of FITM with that of other less active or inactive compounds, we attribute the superior potency observed for FITM to the polar interaction between T815$^{7.38}$ and the amine derivative on the 5' position of the pyrimidine ring, as well as the perfect fit of the ligand shape within the narrow binding pocket.

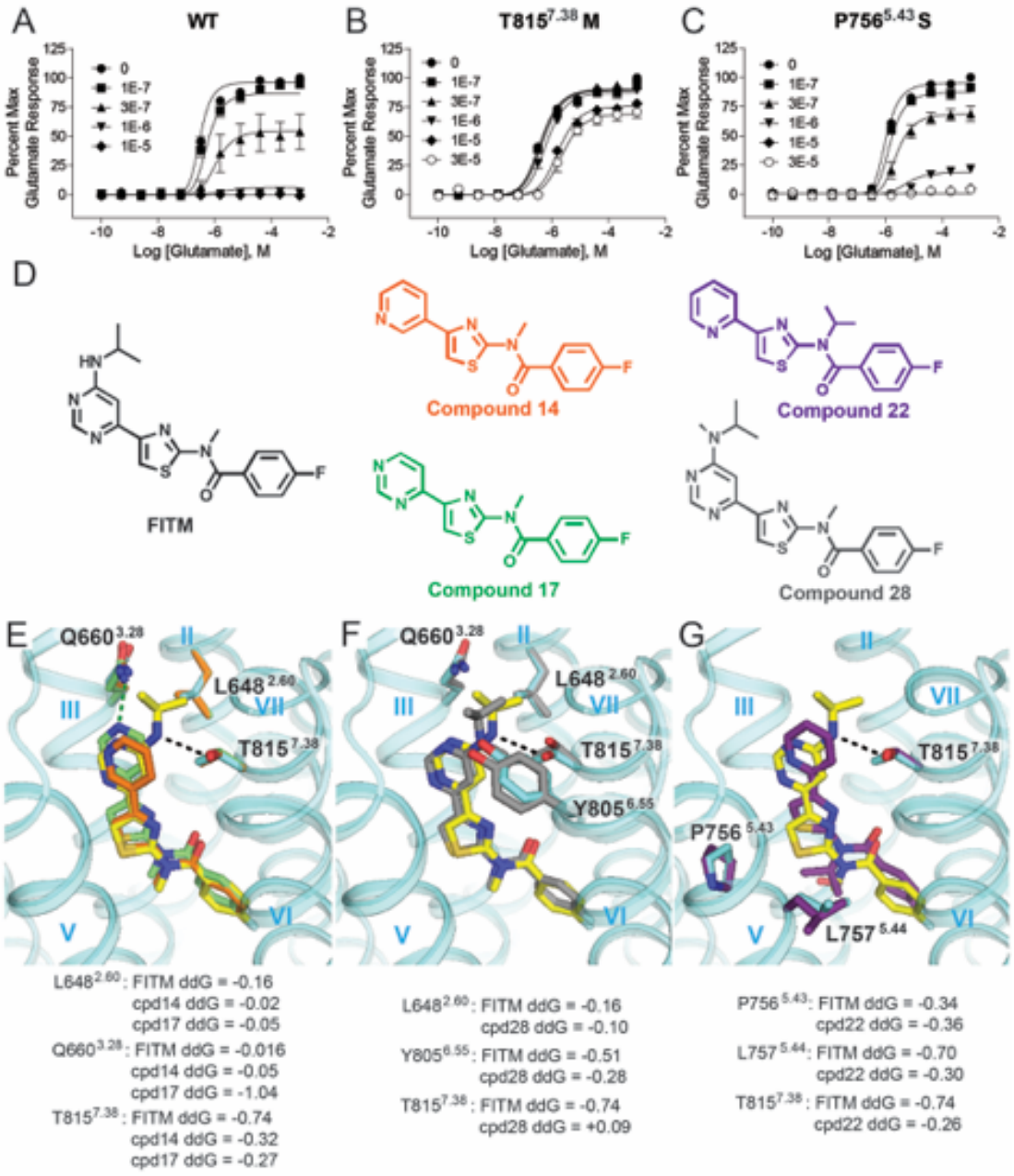


Figure 13. Critical FITM-receptor interactions are revealed by mutations and SAR. (A) FITM is a full NAM of the wild-type (WT) full-length human $mGlu_1$ receptor, while the affinity of FITM and the degree of negative cooperativity with glutamate are reduced in (B) $T815^{7.38}M$ and (C) $P756^{5.43}S$ mutants. (D) Structures of FITM and FITM-related NAMs used for study. Binding pose of FITM (yellow carbons; IC_{50}: 5 nM) in comparison with lower potency analogs, (E) compound 17 (green carbons; IC_{50}: 10 nM) and compound 14 (orange carbons; IC_{50}: 230 nM), (F) compound 28 (grey carbons; IC_{50}: 77 nM) and (G) compound 22 (purple carbons; IC_{50}: 2 μM). Per-residue binding energy ddG is predicted by Rosetta in Rosetta Energy Units (REU). In (E), (F) and (G), side chain rotamers from the top 1% of key amino acids are depicted in sticks and colored corresponding to their respective docked ligand, with the exception of those from the crystal structure shown in cyan.

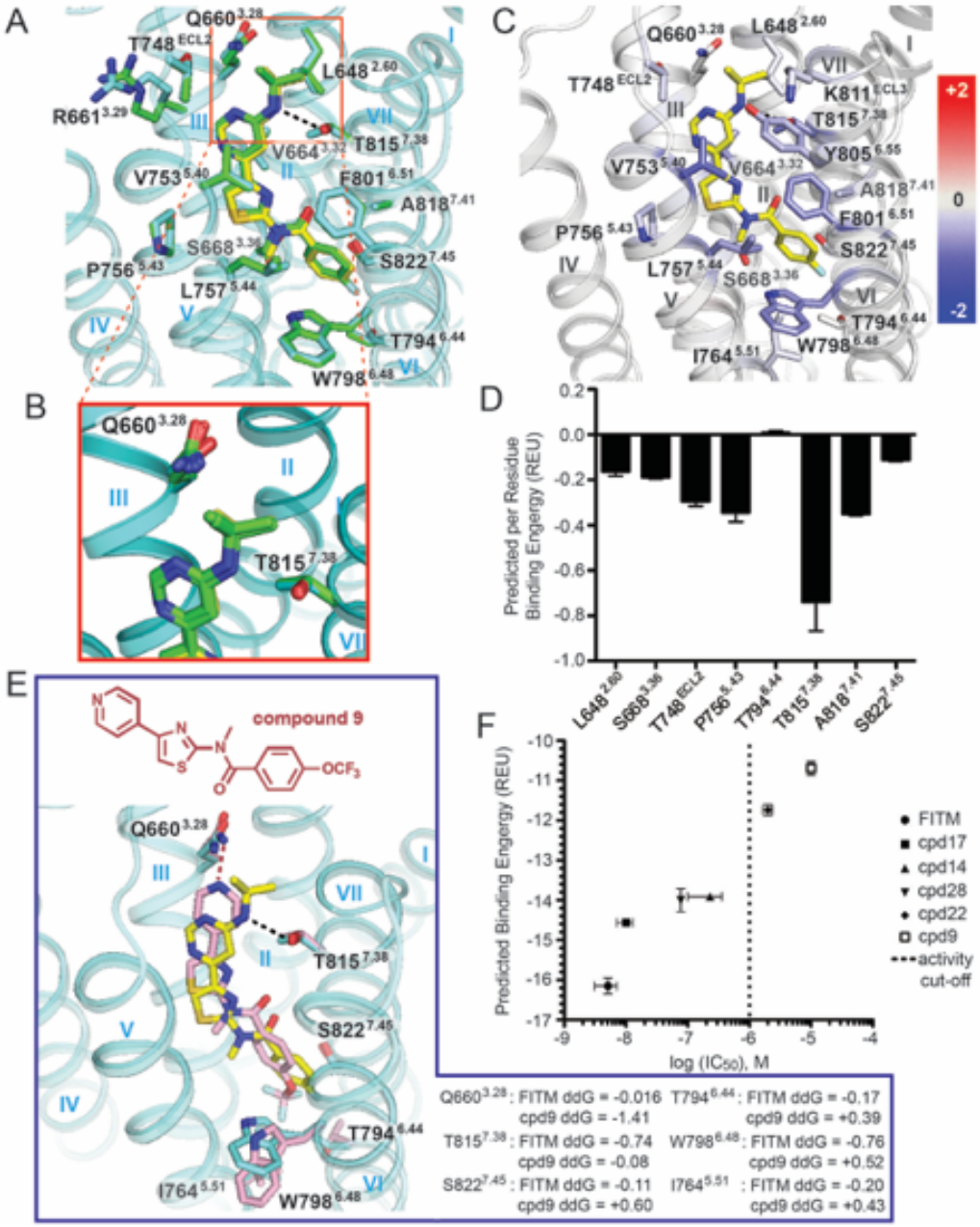


Figure 14. Analysis of compound docking to human mGlu1 7TM crystal structure. (A) Re-docking FITM with RosettaLigand recapitulates the docking pose observed in the co-crystal structure with a RMSD of 0.3Å. (B) The Q6603.28 side chain samples different rotamers. The T8157.38 side chain conformation is strictly conserved, suggesting an essential role in interacting with ligand. (C) and (D) The important role of T8157.38 in ligand binding is confirmed when comparing per-residue binding energies (ddG) as predicted by Rosetta in Rosetta Energy Units (REU). (E) Compound 9 (salmon; IC50: 10 μM) docked into the binding pocket lacks the capacity to interact with T8157.38, and forms unfavorable interactions with I7645.51, T7946.44, W7986.48 and S8227.45, as indicated by ddG. (F) Correlation plot between log IC50 and Rosetta predicted binding energy for all the tested compounds and FITM. The Pearson correlation coefficient is 0.9704 (R square: 0.9418). The dashed line separates compounds considered active and inactive.

4.1.5 The communication between ECD and 7TM domain

In the mGlu family, as well as in other class C GPCRs, a signal is initiated by the native ligand binding to the ECD, which induces large conformational changes in the ECD. In our structure, the linker region (I581-E592) between the ECD and 7TM domain is resolved. The linker forms strong interactions with the ECL2 β-sheet via main chain and side chain hydrogen bonds (Figure 15 A). ECL2 is connected via a covalent disulfide bond to the top of helix III, known to be important in triggering activation in class A GPCRs [195]. Thus, the interaction between the linker region and ECL2 may establish a communication path through which the ligand induced conformational changes of the ECD are transduced to the 7TM region, driving intracellular signaling transduction. In addition, part of the linker residues (*e.g.* W588, a residue conserved in all mGlus) insert into the lipid bilayer, where they form extensive contacts with cholesterol molecules that mediate the observed dimerization of the 7TM domain. This interaction highlights a potential role of dimerization and/or lipid components in the coupling between the ECD and 7TM domain during the activation process.

ECDs of class C GPCRs are known to mediate receptor homo- and hetero-dimerization. Several dimeric structures of mGlu receptor VFT domains have been previously solved in different conformations: putative active (A) or resting (R) state defined by the relative orientation between the VFT protomers as well as closed (c) or open (o) states defined by the conformation of each VFT (Figure 16) [196]. Comparing different conformations, the distance between the C terminal ends of the ECDs within a dimer changes dramatically [169]. In our crystal structure of the

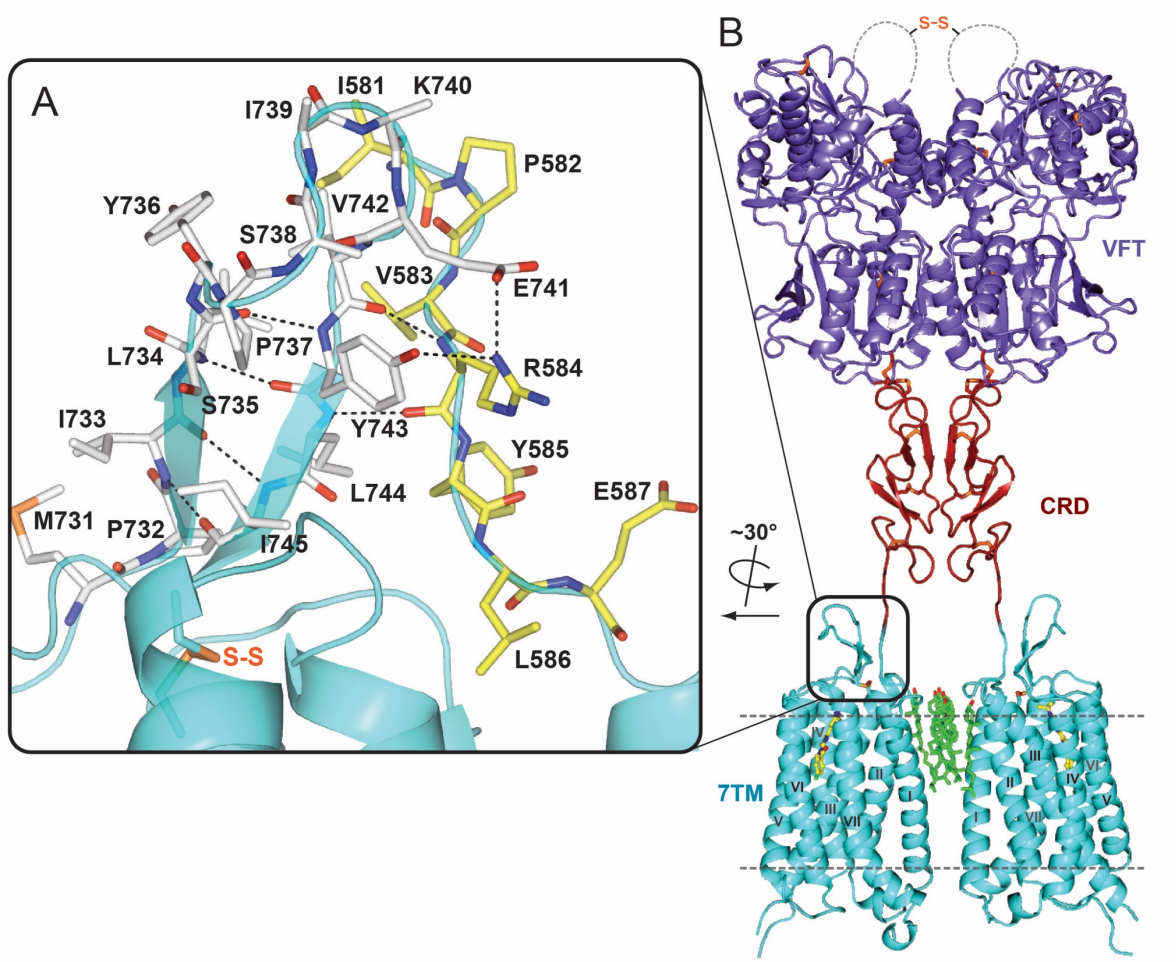


Figure 15. A full-length $mGlu_1$ dimer model with highlighted details of interactions between ECL2 and the 7TM-to-CRD linker. (A) Shown in cyan is the extracellular part of the $mGlu_1$ 7TM. ECL2 residues (M731-I745) are shown as white carbons, while the linker region residues (I581-E587) are shown as yellow carbons. Hydrogen bond interactions between ECL2 and the linker region are shown as dashed lines. (B) Full-length model of $mGlu_1$ with the VFT in the Acc (active closed-closed) state. VFT, CRD and 7TM domains are colored in slate, firebrick and cyan, respectively.

7TM domain, we observed a parallel dimer mediated by interactions of helices I and cholesterols. In this dimer conformation, the distance between the N terminal linkers that are attached to the C termini of ECDs is ~20 Å. If this is a conformation that can be adopted by the full-length receptor dimer, the CRDs of each protomer should also be in a close proximity. Disulfide bond crosslinking experiments have

shown that such a conformation is feasible in an activated receptor dimer [184]. Although our structure is solved in complex with a NAM and the 7TM domain appears to be in an inactive state, there is evidence supporting the existence of a glutamate-bound, but signaling incapable state, in the full-length mGlu dimmers [197]. Moreover, there is evidence that cholesterol can positively modulate glutamate responses by recruiting mGlus to lipid rafts [166, 198], consistent with the observation that the close proximity of the N terminus of the 7TM domain results from a dimer conformation mediated by multiple cholesterol molecules. To test the possibility of fitting the existing ECD structures into our observed 7TM dimer conformation, we created a full-length dimer model in which the VFT adopts an Acc (active closed-closed) conformation as this conformation has the closest distance (~ 50 Å) between the C termini of the ECDs [169] (Figure 15 B). A 20° rotation was applied to the CRD coupled with a conformational change in the Q513-V523 loop region that reduces this distance to 20 Å, fulfilling the CRD interface proposed in the cysteine mutant study [184], as well as matching the distance of the 7TM domain N termini observed in the crystallographic dimer. This model might represent a glutamate bound, but signaling incapable, conformation of $mGlu_1$. We acknowledge that this model, while consistent with the currently available experimental data, is only one of the several explanations for the biological role of the 7TM domain dimer we observed, and needs to be confirmed by future studies. We further acknowledge that the 7TM domain dimer conformation might vary in different states of the receptor and could be modulated by several factors in real biological systems, such as membrane lipid content or

other protein-protein interactions.

4.1.6 The conformation states of mGlu5 during receptor activation

The ECD/TMD together forms a symmetric dimer with three regions of contacts via upper ligand binding domain, CRD, and TMD (Figure 16 A). The structures of mGlu ECD have been determined in Activated-open-close conformation (Aoc), Resting-open-open (Roo) and Resting-close-close conformation (Rcc). It is hypothesized that only the Roo to Acc transition is biologically relevant. However, only the Rcc conformation resolves location and structure of the CRD. The relaxed ECD structure showed the CRDs have the potential to be as close as 20Å at resting states. Thus the Roo conformation of ECD hints that TMD may not have to move a large distance through membrane to dimerize/activate. The conformational changes from Roo to Acc might only require a small distance change between the C-termini, and reorganization of VFT domain and CRD domain. The time scale for mGlu activation is estimated to be smaller than 10ms.

The conformation of the CRD in our model significantly differs from that of crystalized mGluR3, which suggest Q513-V523 loop region would require an extent of conformational flexibility for such transition to occur. The CRD has been described to be mediating activation of mGluR2, and cysteine mutations along the 3rd beta-sheet of CRD lead to constitutive activation or dimerization of the receptor [184]. We identified the equivalent residues to the proposed CRD dimer interface to be in very close proximity: G539, E540, C544, I546 T548 E554 and G568. These residues may mediate the dimerization interface during receptor activation,

as receptor transitions from open to close conformation, and further brings TMD to the dimer conformation. The linker in mGluR1 between CRD and TMD is shorter than other mGluRs, and the bottom of the CRD domain does not have the propensity to form beta-sheet structure as seen in mGluR3. In our model, we conservatively model the linker region between CRD and TMD. There exists the possibility that CRD dimer conformational changes would alter the PIP linker region even further to affect helix 1 or the ECL2.

Taken together, Ligand binding to the VFT ligand-binding pocket leads to the transition and domain reorientation from Open to Closed state. The coupled conformational changes in the VFT loop region thus lead to domain reorientation and restricts the conformational flexibility of CRD for dimerization to occur. The CRD dimerization and further conformational changes would be relayed by linker and ECL2 to the TMD. It was later demonstrated in works by Li et al [46] that TMD dimer interface reorientation is required for receptor activation. By cross-linking experiment, it was demonstrated that TM4 & TM5 formed the dimer interface at resting state, while receptor activation required TM6 to come in contact with each other. Cross-linking TM4 & TM5 interface locked receptor in resting state that cannot be rescued by agonist. Cross-linking TM6 interface resulted in constitutively activated receptor. It is thus interesting to observe that cross-linking both CRD and TMD would lead to constitutively active receptor. It is possible that receptor activation is largely mediated by mechanical signaling upon ligand binding.

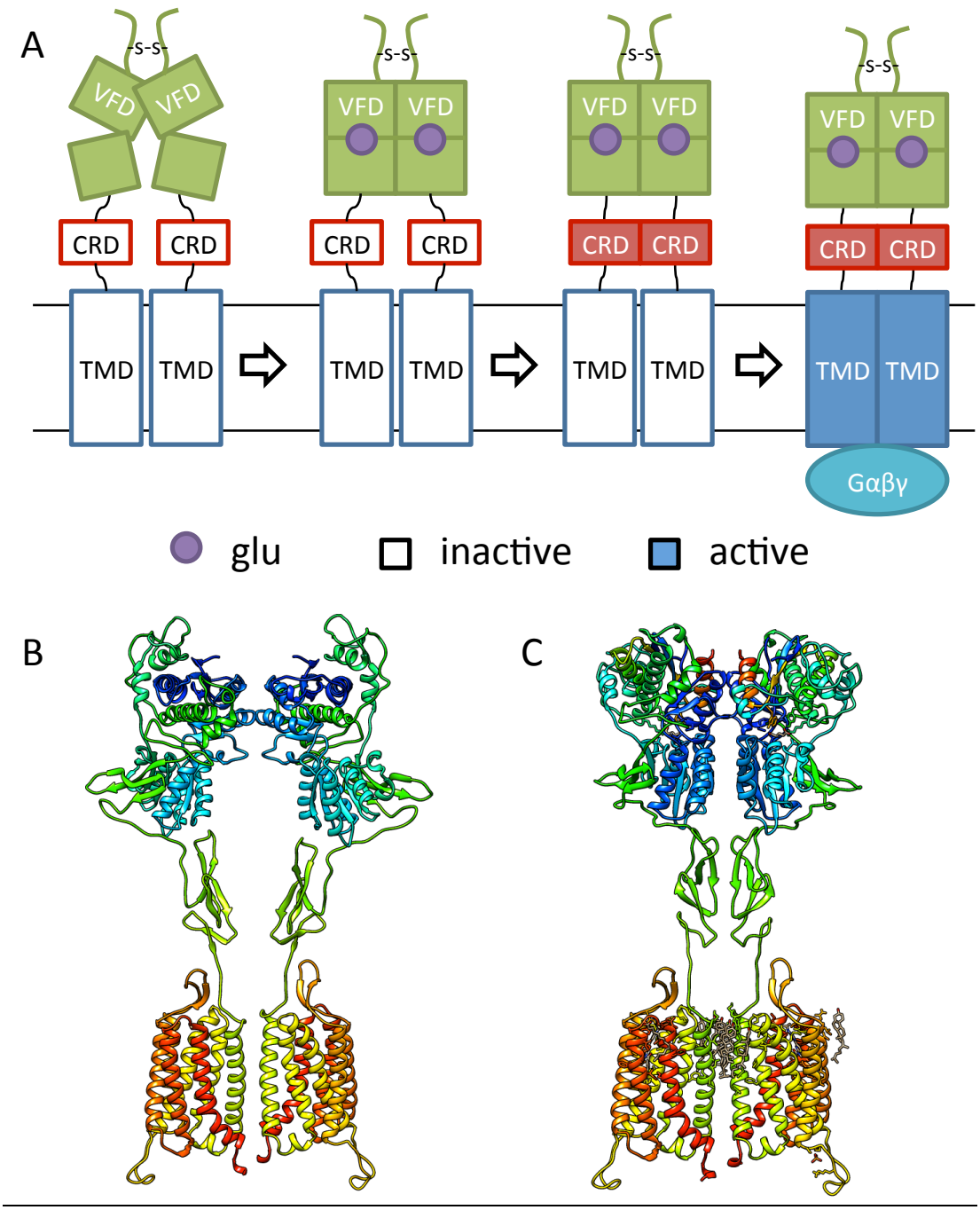


Figure 16. Schematics of $mGlu_5$ activation and conformational changes. (A) The basal state of mGlu where ECL adopt resting open state of VFT domain would change its conformation upon ligand binding to activated closed state. The conformational changes would then propagate to CRD domain, where further signal transduction was relayed to TMD dimer. The activated TMD would be in its activated state for binding G-proteins and downstream signaling. (B) The full length mGlu1 model of the Roo state. (C) The full length mGlu1 model of the Acc state.

4.1.7 Conclusions

The $mGlu_1$ 7TM structure presented here uncovers atomic details of the class C GPCRs transmembrane domain, providing the heretofore missing link in our structural understanding of the entire GPCR superfamily. As noted for the recently solved class B and F GPCR structures, and now for class C, despite a lack of sequence and motif conservation, the architecture of the 7TM bundle is generally preserved. Furthermore, while class C GPCRs are known to form obligate dimers via ECDs, the observed 7TM dimer suggests additional points of communication between protomers, mediated by multiple cholesterol molecules and direct protein-protein interactions. Moreover, as a robust structural template, the $mGlu_1$ 7TM domain structure will likely provide crucial insights into pharmacology of small molecule allosteric modulators for class C GPCRs.

4.2 Resolving Binding mode differences of imaging ligand VUIIS1009A/B for mTSPO

Chapter4.2 is largely based on manuscript "DW, Li J, Buck JR, Tantawy MN, **Xia Y**, Harp JM, Nickels ML, Meiler J, Manning HC. Evaluation of TSPO PET Ligands [18F]VUIIS1009A and [18F]VUIIS1009B: Tracers for Cancer Imaging. Mol Imaging Biol. 2016 Nov 16."

We started characterizing ligands reported in the manuscript after our collaborators reports interesting differences in *in vivo* PET imaging results for the pair of regioisomers. I performed the NMR experiments and mapping of DPA714, VUIIS1009A/B binding site on TSPO. The mapping of ligand binding sites were produced using the NMR structure of mTSPO. Our work demonstrated pyrazolopyrimidine ligands bind to the center cavity of TSPO, slightly different from the binding site for benzodiazepine ligand. After confirming the regioisomers binding to TSPO in a similar fashion, our collaborators carried out further analysis to characterize the ligands' bioavailability differences were caused by plasma protein binding.

4.2.1 Introduction

Neuroimaging enables the noninvasive evaluation of the structure and function of the nervous system and has found particular use towards glioma detection and guidance of therapy. Neuroimaging techniques routinely used in clinical neuro-oncology include magnetic resonance imaging (MRI), x-ray computed tomography (CT), single-photon emission computed tomography (SPECT), and positron emission tomography (PET). Clinically, T_1-weighted MRI,

with or without contrast, remains the gold standard for glioma imaging and treatment response evaluation [199]. However, this technique is an indirect measurement, only reflecting biological activity of the tumor due to degradation of the blood–brain barrier (BBB) [200]. As such, MRI is limited towards providing molecular information attributable to disease pathology and tends to poorly discriminate disease margins, particularly where tumors can be highly infiltrative.

With its sensitivity and quantitative nature, coupled with the ability to readily produce biologically active compounds bearing positron-emitting isotopes, such as fluorine-18 (^{18}F) and carbon-11 (^{11}C), PET imaging has the potential to overcome many of the limitations associated with conventional brain tumor imaging methods. Through the use of different PET tracers, various gliomas can be evaluated based upon: glucose metabolism ([^{18}F]FDG) [201, 202]; amino acid metabolism ([^{11}C]MET) [203]; nucleoside metabolism/proliferation ([^{18}F]FLT) [204]; hypoxia ([^{18}F]FMISO) [205]. However, all these tracers possess limitations [202, 206]. Thus, there is still an urgent need to develop and validate new biomarkers for glioma PET imaging.

Our recent efforts in the PET imaging of glioma have focused on development of novel PET probes that target translocator protein (TSPO), an 18-kD outer mitochondria membrane protein that is overexpressed in several cancer types, including glioma [147, 148] and oral [207], liver [208, 209], breast [70, 149], and colorectal cancers [150-152]. Elevated TSPO expression has also been linked with disease progression and diminished survival in patients with oral [207, 210], colorectal [150-152], breast [70, 149], and brain [147, 153] cancers. Additionally,

elevated TSPO levels appear to be associated with aggressive, metastatic behavior in breast and colorectal cancer [70, 154, 155], suggesting TSPO expression as a potentially important prognostic biomarker in oncology and suggest the utility of tumor-selective TSPO PET ligands for cancer imaging.

Our previous research has shown that high-affinity, pyrazolopyrimidine TSPO PET probes such as [^{18}F]DPA-714 (**1**) [156] and [^{18}F]VUIIS1008 (**2**) [157, 158] can be used to visualize TSPO expression in tumors in preclinical settings (**Fig. 1**). The novel ligand [^{18}F]VUIIS1008 was specifically developed within our lab through focused library synthesis and structure-activity relationship (SAR) development of the 5,6,7-substituted pyrazolopyrimidine scaffold of **1**. [^{18}F]VUIIS1008 proved to be a highly potent TSPO PET ligand, exhibiting a 36-fold enhancement in TSPO affinity, compared to **1**, and high radiochemical yield and specific activity [157, 158].

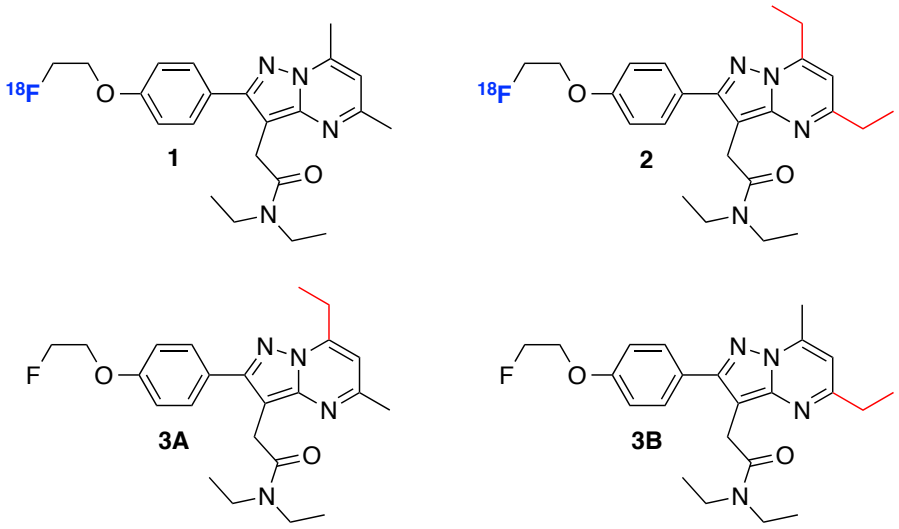


Figure 17. Pyrazolopyrimidine TSPO ligands: (**1**) ^{18}F-DPA-714; (**2**) ^{18}F-VUIIS1008; (**3A**) VUIIS1009A; (**3B**) VUIIS1009B.

With our ongoing efforts to continually develop TSPO ligands for cancer

imaging, we have discovered two novel TSPO probes, *N*,*N*-diethyl-2-(7-ethyl-2-(4-(2-fluoroethoxy)phenyl)-5-methylpyrazolo[1,5-*a*]pyrimidin-3-yl)acetamide (VUIIS1009A, **3A**) and *N*,*N*-diethyl-2-(5-ethyl-2-(4-(2-fluoroethoxy)phenyl)-7-methylpyrazolo[1,5-*a*]pyrimidin-3-yl)acetamide (VUIIS1009B, **3B**) (Figure 17). Both probes share the same pyrazolopyrimidinal scaffold with **1** and **2**, but with distinct modifications at the 5- and 7-positions. These minor structural variations yielded comparable picomolar binding affinities (*in vitro*) against [^{3}H]PK11195 that were 20-fold greater than **2** and 500-fold that of **1**. NMR titration analysis showed both ligands to bind approximately the same sites within the TSPO binding pocket and elicit similar conformational changes within the protein. Though regioisomers, **3B** demonstrated a lower plasma protein binding affinity than **3A**. Radiolabeling of corresponding precursors with fluorine-18 (^{18}F) gave [^{18}**F**]**3A** and [^{18}**F**]**3B**, which allowed exploration of quantitative, *in vivo* imaging characteristics in healthy mice and a rat model of glioma (C6). Of the two, [^{18}**F**]**3B** demonstrated overall higher uptake in tumor tissue, which facilitated better imaging contrast. Taken together, these studies highlight [^{18}**F**]**3B** as a promising and novel PET ligand for evaluating TSPO expression in cancer and potentially other disease.

4.2.2 Material and Methods

NMR experiments of ligand binding

^{15}N-labeled TSPO was expressed in BL21 (DE3) *E. coli* in M9 minimal media, solubilized, and purified with DPC detergent, as adapted from previous publications [90, 163]. 2D ^{1}H-^{15}N heteronuclear single quantum coherence (HSQC) spectra of mouse TSPO with/without ligands (5.0 mM) were recorded on

an 800 MHz NMR spectrometer at 42 °C with 0.2 mM TSPO in ~2% DPC (w/v), 25 mM MES buffer, pH 6.0, and 100 mM NaCl. Ligands tested were: **1**, **3A, 3B**, and PK11195. Spectra were processed with NMRpipe and analyzed using Sparky.

Mapping of ligand site by chemical shift differences

Complete backbone assignment of mice TSPO was obtained from Biological Magnetic Resonance Data Bank (http://www.bmrb.wisc.edu/ BMRB entry: 19608). The backbone H-N assignments were transferred to our mTSPO-PK11195 spectrum with most of the peaks aligning well with published assignment value, except for center regions where overlapping peaks made accurate assignment impossible. The assigned mTSPO-PK11195 spectrum were then overlaid with mTSPO-DPA714, mTSPO-**3A** and mTSPO-**3B** to further analyze the differences in chemical shift (ΔCS). Partial assignment of mTSPO-DPA714, mTSPO-**3A** and mTSPO-**3B** were carried out when the chemical shift of a backbone residue can be confidently assigned from overlapping peaks compared to mTSPO-PK11195 spectrum, or a shifting peak that can be unambiguously identified from the original spectrum assignment. **3A** and **3B** ligand binding differences were marked with ΔΔCS, since it is derived from assignment of the ΔCS. The putative ligand binding site for the three pyrazolopyrimidine were mapped onto the sequence and mTSPO NMR structure [90] (PDB:2GMY) using Pymol.

4.2.3 Protein Binding Pockets for **3A** and **3B** in Mammalian TSPO

Jaremko *et al.* recently reported that the structure of TSPO could be stabilized in the presence of PK11195, yielding a tight bundle of five

transmembrane α helices that form a hydrophobic binding pocket to accept PK11195. 2D ^{1}H-^{15}N heteronuclear single quantum coherence (HSQC) NMR analysis of this structure allowed for a high-quality spectrum of the ligand-bound state of the protein [90]. To investigate the differences in the structure of TSPO when bound to **3A** and **3B**, 2D HSQC spectroscopy was used to observe structural features of each ligand-bound state of the protein and compared to DPA-714 and PK11195. In a 2D ^{1}H-^{15}N HSQC spectrum, peaks arise from amides in the protein backbone and side chains that are highly dependent on the chemical environment surrounding each amide group, allowing the HSQC spectra to serve as a fingerprinting method of a protein structure in solution. The ligand-free TSPO reconstituted in dodecylphosphocholine (DPC) micelles displayed the same narrow dispersed spectrum as previously reported (Figure 18 A). Addition of PK11195 (Figure 18 B), as well as DPA-714 (Figure 18 B&C), resulted in highly dispersed spectra, with a new set of peaks representing signals from the ligand-bound state of the protein. The DPA-714 spectrum showed significant differences in chemical shift perturbation when compared with the PK11195 spectrum (Fig. 3B), possibly due to different chemical groups from the two compounds interacting with the protein, the different ligand-induced protein conformational changes, and the ligand-binding pocket formed for DPA-714 being different from that of PK11195. Titrating **3A** or **3B** into the TSPO also gave improved spectra dispersion, though in our hands, of the four compounds tested, only 3B completely saturated the ligand-bound state. No obvious changes in chemical shift perturbation were observed between DPA-714, **3A**, or **3B** (Fig. 3C), indicating that

all three ligands bind the same site within the TSPO binding pocket and elicit similar ligand-induced conformational changes within the protein.

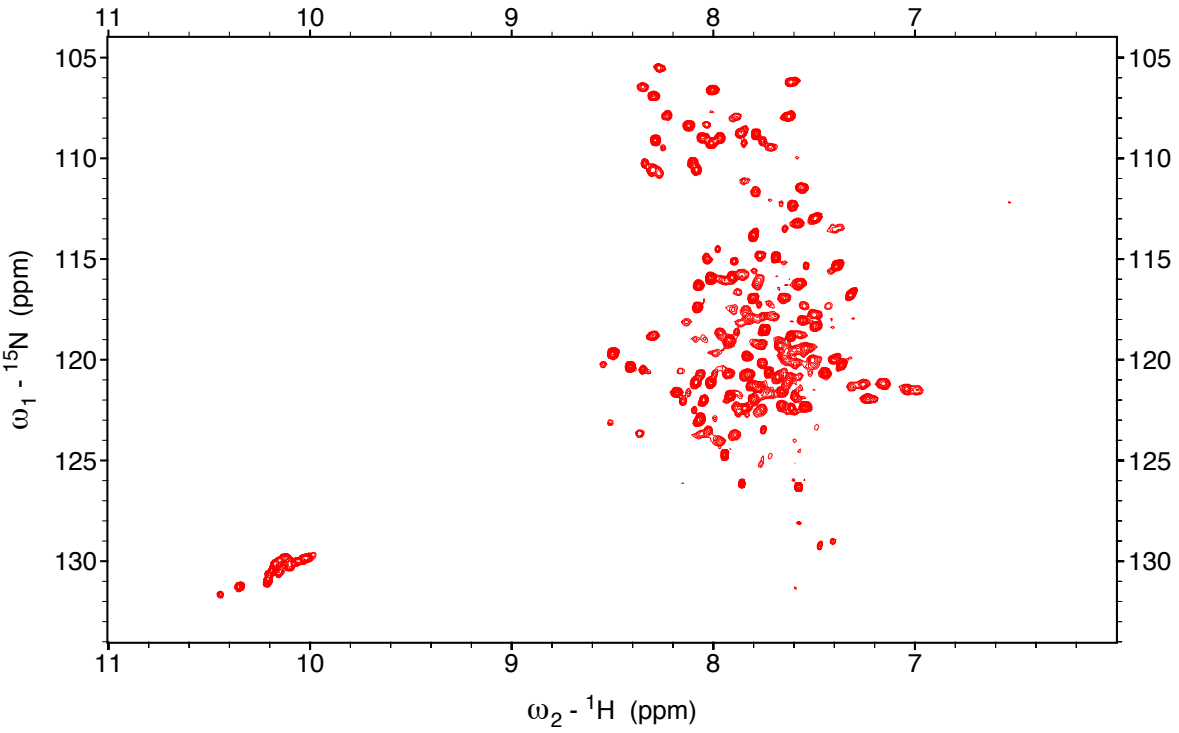
4.2.4 Ligand Binding Site mapping for DPA714, **3A** and **3B** in Mammalian TSPO

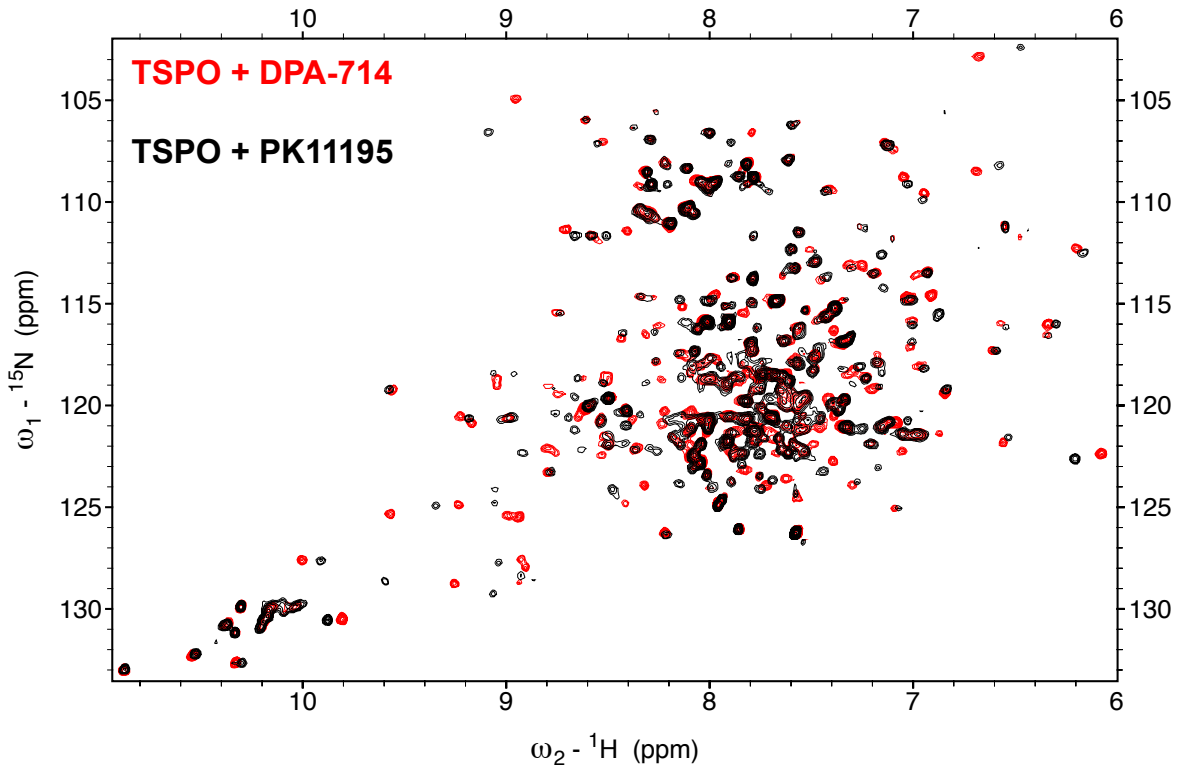
Based on chemical shift of information of mice TSPO in the high quality 2D HSQC of mTSPO-PK11195, resonance assignment transfer was performed. The assigned mTSPO-PK11195 spectrum were then overlaid with mTSPO-DPA714, mTSPO-**3A** and mTSPO-**3B** respectively to further assign resonances from these samples. In the mTSPO-DPA714, mTSPO-**3A** and mTSPO-**3B** spectra, shifted peaks from mTSPO-PK11195 were marked by the differences in chemical shift (ΔCS) (Figure 19) and the peaks that did not shift were also marked. In all samples, we observed well dispersed NMR spectra after ligand addition, indicative of the ligand stabilizing mTSPO in similar global fold. Comparing mTSPO-PK11195 and mTSPO-DPA714 spectrum showed peaks away from ligand binding site having no significant ΔCS, with few exceptions. This finding suggest the protein conformation and chemical environment surrounding these residues are the same. The residues that shows ΔCS in mTSPO-DPA714 were found to overlap partially with those comprised of the PK11175 binding pocket. In all three cases for mTSPO-DPA714, mTSPO-**3A** and mTSPO-**3B,** residue G22, R27, W53, G54, W93, W95, V118, A119, A147 and W143 showed significant ΔCS (Figure 19, Figure 20A). Among these residues that showed ΔCS, W53, G54 and V118, A119 are located further away from PK11195 binding site on TSPO, suggesting that pyrazolopyrimidine molecules might bind at a deeper site within the transmembrane domain (Figure

20A).

Additionally, for residues that can be successfully assigned in mTSPO-DPA714, **3A** and **3B,** we also observed differences in chemical shift. We termed this differences between pyrazolopyrimidine molecules to be ΔΔCS (Figure 19), which would be indicative of minor differences in binding pose of the respective ligand in mTSPO ligand binding pocket. After comparing the ΔCS between pyrazolopyrimidine molecules and PK11195 and ΔΔCS between the DPA714, **3A** and **3B**. We mapped the ligand binding site on mTSPO (Figure 20). The regioisomer also induced differential binding induced chemical shift to residue G19, G22, W93, W95 and A147 (Figure 20B). These residues would be important for ligand binding based on the chemical moiety differences in the ligand. However, specific interactions between ligand chemical moieties and amino acids could not be determined by our NMR experiment since 2D HSQC only monitors the protein resonances. Amongst the residues that showed significant ΔΔCS, A147 is of particular interes. Our finding suggests that modifications to pyrazolopyrimidine molecules would affect the ligand interaction with A147.

The ΔCS and ΔΔCS mapping also showed residues away from ligand binding site has chemical shift changes based on ligand binding. This suggest certain allosteric interactions might play a row in protein function upon ligand binding.


11
10
9
8
7
105
110
115
120
125
130
ω1 - 15N (ppm)
ω2 - 1H (ppm)


TSPO + DPA-714
TSPO + PK11195
10
9
8
7
6
105
110
115
120
125
130
ω1 - 15N (ppm)
ω2 - 1H (ppm)

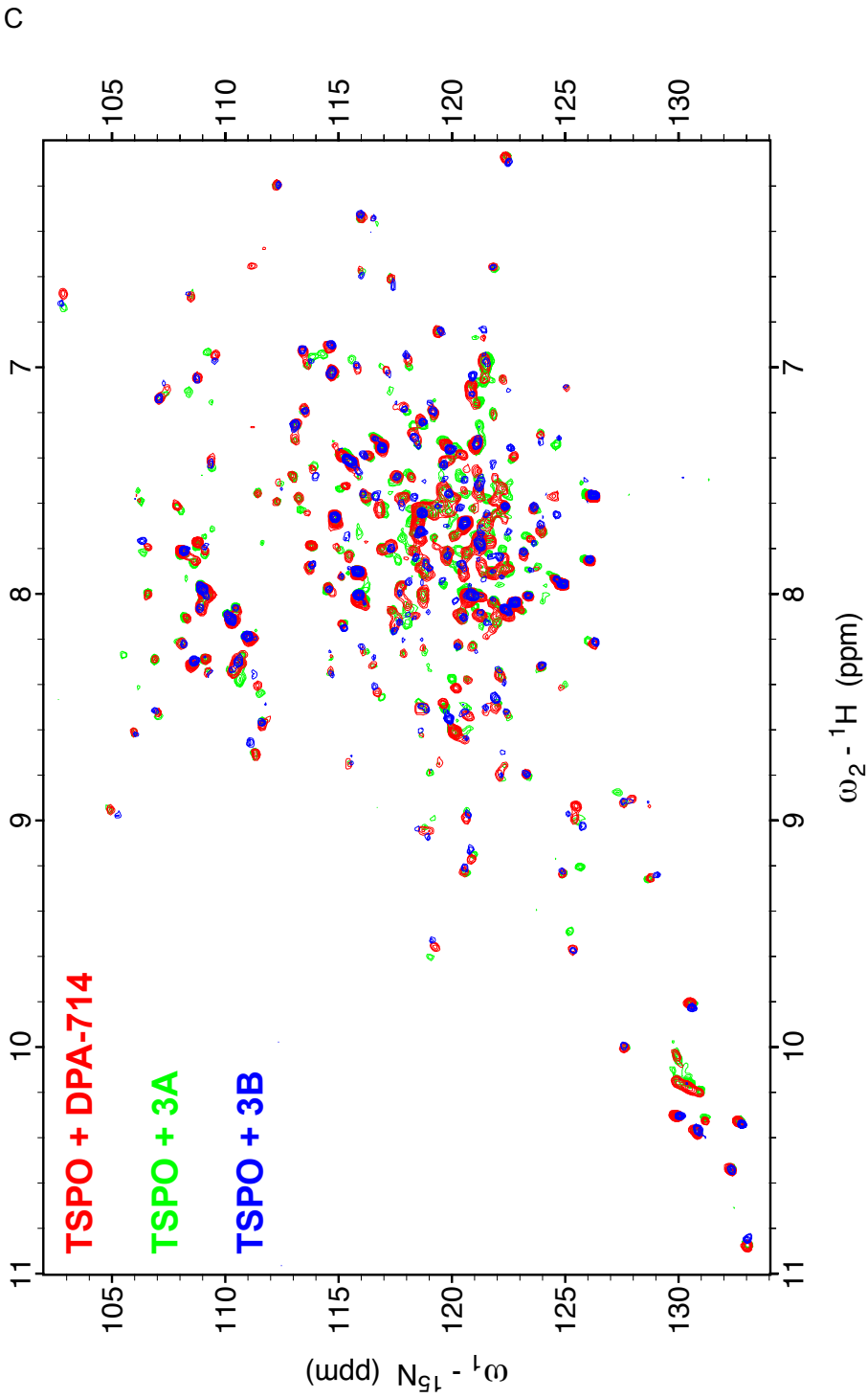


Figure 18 2D ^{1}H-^{15}N HSQC spectrum of TSPO in presence of different ligands. (**A**) The 2D HSQC spectrum of TSPO in the absence of ligand. (**B**) Comparison of TSPO spectra in presence of 5.0 mM PK11195 (Black), and 5.0 mM DPA-714 (Red). (**C**) Comparison of TSPO spectra in presence of 5.0 mM DPA-714 (Red), 5.0 mM **3A** (green), and 5.0 mM **3B** (Blue).

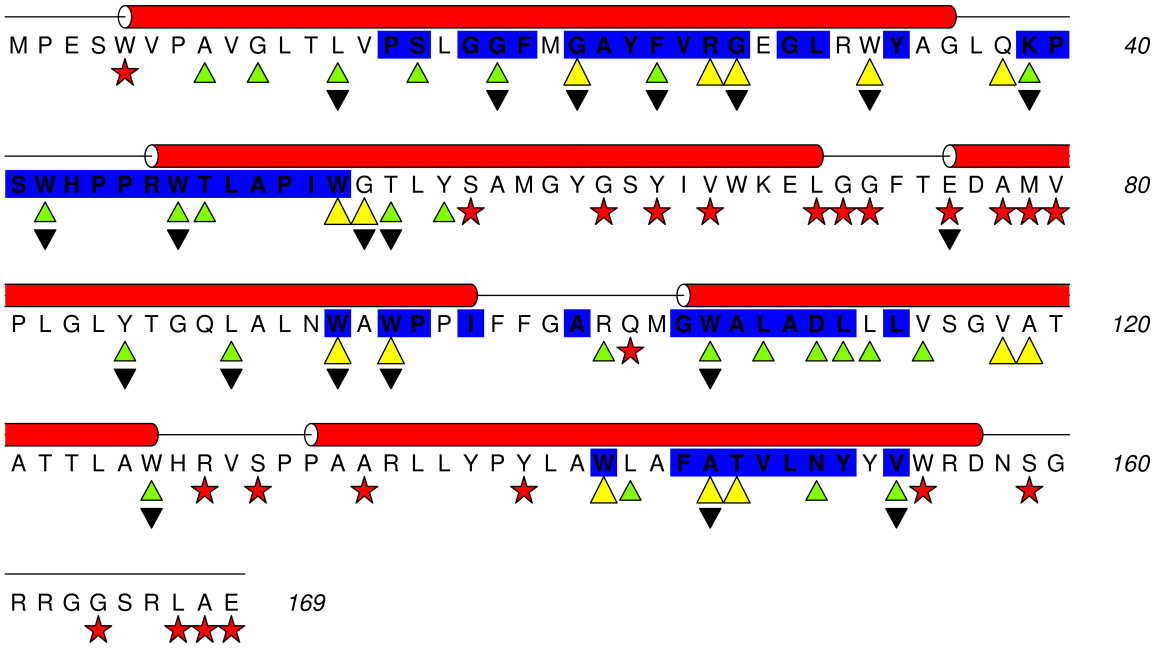
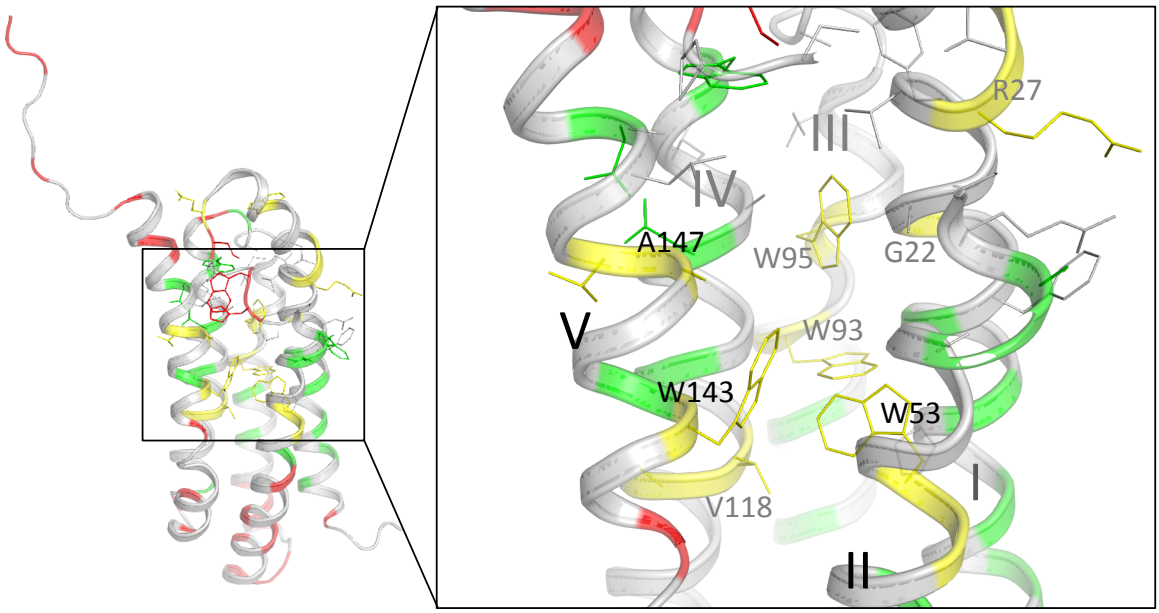


Figure 19 The ΔCS mapping of the binding pocket for DPA-714, 3A and 3B onto TSPO sequence. Transmembrane helices were highlighted as cylinders on top of the sequence based on NMR model, while amino acids that are within 6 Å to binding site for PK11195 were highlighted blue in the sequence. When comparing ΔCS of the Amino acids, ones that showed no significant CS differences were marked by red stars; Amino acids that showed minor shift were marked by green triangles, while amino acids that showed major shift changes were marked by yellow triangles. When comparing 3A and 3B chemical shift, the residues that showed significant ΔΔCS changes were marked by black inverse triangles.

A.



B.

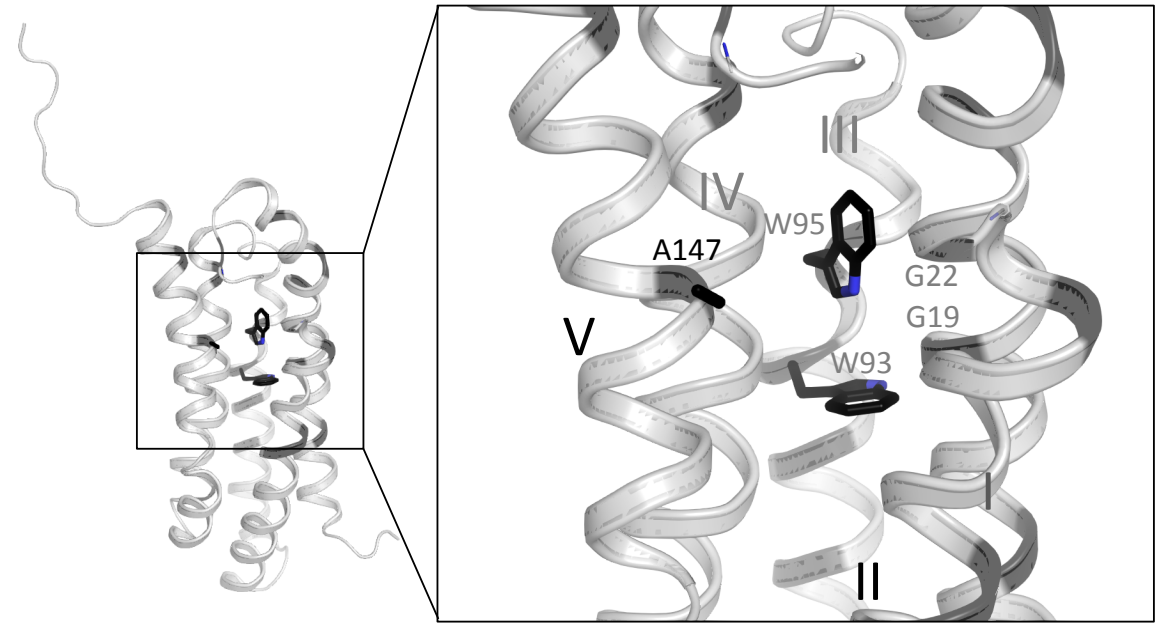


Figure 20 The ΔCS mapping of the binding pocket for DPA-714, 3A and 3B. (A) ΔCS mapping of the binding pocket for DPA-714, 3A and 3B. (B) residues with significant ΔΔCS changes in respond to 3A and 3B titration in the binding pockets are mapped according to coloring schemes as previous figure.

4.2.5 Discussions

At the molecular level, both **[^{18}F]3A** and **[^{18}F]3B** exhibited similar binding affinity and binding pockets, while at the tissue-system level, they demonstrated significant different *in vivo* performance. This difference could be explained by the pharmacokinetic analysis with kinetic modeling and kinetic parameter analysis. Based on our results of the 2-compartment, 4-paramter kinetic modeling, we found **[^{18}F]3B** to have a similar tumor BP_{ND} and lower brain BP_{ND} relative **[^{18}F]3A**. When compared with **[^{18}F]3A**, we also noticed a significantly higher K_1/k_2 ratio for **[^{18}F]3B** in both tumor and brain, indicating a higher plasma-to-tumor extraction efficacy for **[^{18}F]3B**. We propose that this higher plasma-to-tumor extraction efficacy could be the reason for the higher tumor and brain uptake for **[^{18}F]3B**.

The plasma-to-tissue extraction efficacy is important for PET imaging, especially for glioma imaging, which requires radioligands to readily cross the blood-brain barrier (BBB). This capability relies mainly on the molecule's lipophilicity and plasma binding affinity. As regioisomers, **[^{18}F]3A** and **[^{18}F]3B** have nearly the same structure and very similar lipophilicities, with their ability to cross the BBB due mainly to their different plasma binding affinities. Measurement of the plasma binding affinity of **3A** and **3B** showed that modification at the 5- and 7-position of the pyrazolopyrimidine scaffold bore significant effects on plasma protein binding affinity. Of the four pyrazolopyrimidine TSPO ligands tested (DPA-714, VUIIS1008, **3A**, **3B**), VUIIS1008 had the highest plasma protein binding affinity, and DPA-714 the lowest. Although they have very similar structures and lipophilicities, **3A** possesses a higher plasma protein binding affinity than **3B**. In

light of the comparatively higher *in vivo* uptake of **[^{18}F]3B** in glioma, and considering the similar binding in TSPO as measured by NMR, we propose the lower plasma binding affinity to be a crucial factor for the higher uptake and higher extraction efficiency for **[^{18}F]3B**, when compared with its regioisomer **[^{18}F]3A**.

In vivo radiometabolite analysis indicated both probes to possess a higher level of stability compared to **1**, which has a faster *in vivo* plasma clearance, with its parent compound ratio dropping from 95 to 27%, in a timespan of 2 to 60 minutes. Direct comparison between **[^{18}F]3A** and **[^{18}F]3B** indicated a slower clearance for **[^{18}F]3A** and thus, a higher overall *in vivo* stability. Interestingly, both probes demonstrated the same biodistribution pattern, as well as specificity and reversibility in healthy tissues in mice, indicating similar overall performance for both probes in healthy tissue.

Residue A147 is of particular interest since the human rs6971 (A147T) polymorphism. Our finding suggests that modifications to pyrazolopyrimidine molecules would affect the ligand interaction with A147. This information is important for imaging ligand development as pyrazolopyrimidine is sensitive in to human rs6971 polymorphism in pre-clinical PET imaging settings. We have demonstrated that NMR spectroscopy could determine whether A147 is involved in ligand interactions. Future ligand development could use NMR to screen for ligand that does not affect interactions with A147 site based on our protocol described here.

CHAPTER 5 CONCLUSION AND FUTURE DIRECTIONS

The work presented in this document summarizes methodology development for membrane protein structure determination and application of NMR spectroscopy and computational modeling to solve biological problems such as in the case of TSPO. I presented an integrated structural biology tool to predict membrane protein structures with limited experimental data. I have shown that structural properties of mTSPO is dependent on the detergent and ligands. I have also constructed a unified model of mTSPO that can be used to explain ligand interactions based on computational ligand docking and experimental NMR data. The crossroad of computational and experimental structure biology could also yield structural insight on large GPCR allosteric modulators or elucidate ligand-protein interactions between imaging ligands and TSPO.

However, several challenges are still present that call for further development of our computational tool kit and experimental characterization of TSPO. I outline the future direction for developing the integrated MP structure prediction pipeline in section 5.1. Section 5.2 summarizes the published production and purification methods for TSPO structural characterization, and discusses other strategies to characterize TSPO-ligand interactions. Most of the structural data of TSPO are obtained in a ligand-bound stabilized condition. The dynamics of TSPO are studied in the context of ligand-stabilization or oligomerization. In reality, TSPO might function as a component in the mitochondria permeability transition pore (MPTP) complex with VDAC and other proteins. Future research on TSPO should also consider probing the larger molecular complexes. At last, I describe the

current understanding of TSPO dynamics and its role in biological function and potential experimental strategies to study the TSPO dynamics in lipid bilayers.

5.1 Structure determination from limited experimental data will be important to determine conformational states of integral MPs

5.1.1 The experimental data needed for the current BCL and Rosetta prediction pipeline can be optimized

The integrated structure prediction pipeline has been shown to successfully determine membrane protein structures to atomic accuracy [97]. We observe consistent improvement when combining experimental data from orthogonal experimental techniques in our simulation. However, it is uncertain what is the minimum requirement of experimental data for accurately predicting large membrane proteins in the current implementation. In the test case used in the publication, a combination of NMR, EPR distance restraints and intermediate resolution EM density maps were used. Specifically, the benchmark case used simulated sets of 348 NMR NOE distance restraints (1 restraint per residue), 27 EPR DEER distance restraints and 5.5Å resolution experimental EM density map. The distance restraints were simulated due to the lack of experimental NOE data and uniformly distributed EPR spin labeling dataset.

We have previously determined the lower limit of experimental NOE and EPR data required for accurate protein structure prediction based on single type of data [112, 211]. I hypothesis that to achieve atomic accuracy modeling of MPs, a lower number of experimental restraints is required when these restraints are

simultaneously incorporated in the modeling process. A benchmark can be designed with simulated sets of different number of restraints from a set of MPs of different sizes. The simulated experimental data can be designed to 0.25, 0.5, 1 restraint per residue for long-range NOE distances, and 2,3,4 restraints per SSE for DEER distances. Combination of these sets of restraint can be used for finding the required numbers of experimental data for accurate structure modeling at atomic detail. The same principle could be applied to cryo-EM density maps, as the experimental density can be adjusted for the voxel sizes that account for lower resolution maps. The benchmark would be useful for finding the limit of our algorithm and also guide experimental data collection. After establishing such minimum requirement for accurate modeling, experimentalists can use subsets of unambiguous data to drive computational modeling or attempt structure determination when no single type of experimental data can yield unambiguous models.

While NOE measurement is often difficult in MPs due to the large size of the system, other experimental data from PRE, PCS and RDC may be implemented and incorporated into the prediction pipeline. With the resolution-revolution of cryo-EM [212], our current pipeline can also be used in structure validation, refinement, and constructing missing dynamic loops that are not accessible in the electron density.

5.1.2 Integrating evolutionary information in structure prediction may further improve prediction accuracy

Recent developments in co-evolution contact prediction offer alternative

information for IMP structure determination [213-216]. With large enough sequence databases, residue-residue contacts can be inferred from residues that co-evolve. Such restraints have been used to improve structure prediction accuracy for IMPs by restricting the sampling space in Rosetta [217]. Combining sparse experimental restraints from NMR with evolutionary constraints allows for accurate prediction of soluble protein structure in many cases [213]. Evolutionary Coupling-NMR developed by Tang et al [213] incorporates evolutionary contacts during and after the NMR data interpretation and NOE assignment. In a more recent publication by Ovchinnikov et al, 206 unknown IMP folds were predicted using integrated metagenome data and co-evolutionary analysis [218]. The sizes of the proteins involved in this study are below 300 amino acids.

In the current implementation of the membrane protein folding pipeline, only experimental data were implemented. Future works on the integrated structural prediction pipeline may focus on expanding the functionality to incorporate evolutionarily derived contact information. The residue-residue contact can be implemented a similar way to the NOE distance potential in BCL and Rosetta (See Chapter 2.2), where a contact may be defined by the lower and upper bound distances between the center of mass of two residues.

With the expansion of genomic databases coupled with computational prediction algorithms, evolutionary constraints provide viable orthogonal structural information to guide protein structure prediction. The structural conservations from sequences could also inform crucial interactions in the protein that might affect its functionality.

5.1.3 Future development of the pipeline should focus on conformational ensemble states modeling guided by experimental data

Structural models of IMPs could be obtained through X-ray crystallography: in many cases, disruptive experimental technologies are used to stabilize and alter the energy landscape of the native IMP via thermo-stabilizing mutations, chimeric protein engineering, or the non-native-like membrane mimics. Biophysical methods observe ensembles of molecules in their native-like dynamic equilibrium to complement or correct observations in crystal structures.

For example, transporters exist in multiple conformations with respect to their transport cycle. The small multidrug resistance transporter EmrE ultilizes the proton gradient to export cytotoxic molecules against their chemical gradient out of the cell and protect the bacteria. Whereas the crystal structure is limited to its substrate bound state [219], the transport cycle needs to adopt the intermediate proton-bound state to complete a cycle. A systematic SDSL-EPR study on EmrE has revealed rotation and tilting of TM helices 1 – 3 in response to the change of the protonation state [220], which in turn results in the change of substrate entry and binding site. The distance distribution from the SDSL-EPR study, combined with computational modeling was successful in producing an intermediate structural ensemble that fills the knowledge gap of the crystal structures [220].

NMR, EPR, cryo-EM data, albeit with samples at high temperature or flash-frozen, are expected to reflect native-like conditions when different conformations of the protein exist in the equilibrium. SDSL-EPR spectroscopy can observe such

ensemble states as a probability distribution of a distance within an ensemble is observed. NMR spectroscopy can observe ensemble averages of conformations or distinct conformational states depending on the time-scale of motion in the exchange process. With improvements in direct electron detection and image averaging, EM can also distinguish diverse conformational states in the sample [221]. Computational modeling algorithms

BCL::MP-Fold and Rosetta molecular modeling suite can optimize a given starting structure. Thus from any given starting structural model one can derive a model for an alternate state of an MP by fitting it to sparse experimental data observed for this state **[220].** By combining experimental data from different sources, excluding the likely outlier populations due to artifacts in the disruptive experimental methods, one can model the entire conformational ensemble of the IMPs. Such ensemble models along with the statistic inferences could provide further molecular mechanistic insight on protein function, and possibly guide small molecule modulator development against intermediate states.

5.2 TSPO biochemistry and molecular interactions require further examination

5.2.1 Experimental techniques perturb TSPO function in multiple aspects

The introduction briefly describes progresses towards TSPO structural characterization. The sample preparation protocol that has been used in the aforementioned studies can be described by two types: 1. TSPO extraction from recombinant expression host using harsh detergent and re-folding into lipid mimetics [86, 87, 90-92, 95]; 2. TSPO or TSPO homologue production and

extraction from biological membrane fractions using milder detergent [88, 93, 94, 96].

The mTSPO is produced in *E.coli* and mostly expressed into inclusion bodies [85]. Thus a harsh detergent such as SDS were used to extract mTSPO from insoluble aggregate after centrifugation. The proteins are then refolded into proteoliposomes [85, 95] or milder detergent DPC [87, 90-92]. While proteoliposome preparations have been shown to restore the re-folded mTSPO's nano-molar affinity to cholesterol and imaging ligand PK1195, the re-folding efficiency has not been reported in the literature. Ligand binding behavior in DPC micelles was not well documented in literature.

In our experimental optimization steps, we found a fraction of mTSPO are expressed into the membrane during *E.coli* recombinant expression. DPC and DDM detergent was used to solubilize mTSPO for subsequent purification and ligand binding assays. The binding affinity of mTSPO to PK11195 in detergent systems are determined to be at high micro-molar to milli-molar range. The mTSPO in DPC micelles might It in a destabilized state.

In the second approach, TSPO homologues from bacteria (rsTSPO, bcTSPO) can be expressed into the biological membranes. Milder detergent such as DM and DDM are used to extract the TSPOs. Ligand titration experiment in detergent solubilized rsTSPO and bcTSPO found these homologues binds PK11195 at micro-molar affinity. The discrepancies in ligand affinity of these bacterial homologues could be the result of sequence divergence in the ligand binding pocket. In the case of rsTSPO, DM and DDM micelles are found to

stabilize homo-dimeric states in solution.

More recently, mammalian cell expression system was used to express a number of mammalian TSPOs [95]. The HK293T cells allowed for membrane targeted expression of mammalian TSPO and extraction using mild detergent. A mixed micelle system containing DDM and CHS were used to solubilize TSPOs. Fluorescence quenching experiments showed the mammalian TSPO in DDM/CHS mix micelles had micro-molar affinity to PK11195, lower than the literature values. The results suggest that regardless of the extraction methods, detergent systems must be further investigated for conditions that stabilize TSPO.

In our optimization effort for mTSPO extraction and solubilizing, a number of conditions were tested (Table 3). In contrast to the protocol used in NMR studies, where SDS was used for solubilization, milder detergent such as DPC and DDM are used for protein extraction. However, many of the conditions tested result in poor solution-state NMR spectrum. Attempts to incorporate purified mTSPO in detergent into isotopically tumbling lipid-bicelles and nanodiscs failed due to aggregation in the sample after mixing. In the case of nanodiscs, mTSPO start to aggregate upon interaction to the mixture of detergent, lipid and scaffolding proteins. It is possible that during the mixing, TSPO is kinetically favored for aggregation than association with the lipid molecules present in the mixture. Additional lipid-mimetic systems such as amphipols can be screened,

Taken together, the current available technologies to produce and purify TSPO require further optimization. Cell-free expression might be a viable path for TSPO production, as it circumvents the step of exchange process with detergent.

Direct expression of MPs into nanodiscs haven been successfully demonstrated and could be applied for TSPO production [31, 222, 223].

Detergent	**NMR spectral quality**	**SOlubilization**
DPC[90]	+++	+
LPPG	+++	+
LPPG/DPC 1:1	+++	+
LPPG/DPC 1:2	+++	+
LPPG/Cholesterol	++	Aggregation at NMR temperature
LDAO	-	+
DDM	-	Aggregation at high concentration
DDM/CHS	-	Aggregation at high concentration
D7PC	-	Aggregation during purification
TDPC [92]	+	+
DMPC/DMPG/D7PC bicelle	-	Aggregation*
DMPC/D7PC bicelle	-	Aggregation*
LMPG	+	+
cNW9/DMPC/DMPG nanodiscs	-	Aggregation*
DMPC/DMPE liposome	-	+
DMPC liposome	-	+

Table 3. list of detergent systems used to optimize mTSPO ligand. The number of + in the table is indicative to spectra quality and peak numbers. *: In these conditions, protein aggregation is observed several minutes after mixing the components of the membrane mimetic system with mTSPO.

5.2.2 Cholesterol interaction to TSPO needs further elucidation

We have shown that TSPO ligand binding can be monitored by NMR spectroscopy in our studies of mTSPO. However, the molecular mechanisms of many of the proposed TSPO functions are still unknown. For a proposed cholesterol transporter protein, mTSPO binds cholesterol at nano-molar affinity at the CRAC site [86, 137]. Identification of potential secondary cholesterol binding site could explain TSPO's cholesterol transporting mechanism. Solid state NMR results suggest mTSPO favors the monomeric state upon cholesterol binding in liposome systems. Cholesterol not only affected the chemical shifts of the CRAC site, but also affected the chemical shifts of the GXXXG motif that is proposed to be the putative protein dimerization interface. The authors suggested that a potential allosteric effect upon cholesterol binding might influence the oligomerization states of mTSPO.

Chemical shift data from solid-state NMR data alone might not be sufficient to ascertain such claims. The allosteric effect of cholesterol molecules in mTSPO requires further examination to exclude the effects of possible cholesterol-mediated membrane re-organization. GXXXG motif has been proposed to mediate membrane protein interactions [224]. Cholesterol also seems to negatively affect GXXXG mediated protein-protein interactions [225, 226] and, in a specific case, bind to GXXXG [227]. The influence of cholesterol on the transmembrane GXXXG motif of the ErbB2 receptor has been studied using CG MD simulations of the dimer in bilayers with varying amounts of cholesterol [226]. In the study, it was observed that cholesterol generally packs more closely around the helix and that

the position of cholesterol along the membrane normal overlaps with the position of each of the two GxxxG sequences. The increased cholesterol content also causes a thickening of the membrane, thus affecting tilt angle of the helix with respect to the membrane normal and changes interaction mode for GXXXG motifs.

Additional experimental validations are need to examine cholesterol interactions with TSPO. Spin-labeled cholesterol molecules could provide information of the cholesterol interaction motifs in TSPO [228]. The spin-labeled cholesterol, when bound to TSPO, would introduce PRE effect to the back-bone amide protons in the protein. Such PRE effect can be monitored through H-^{15}N cross peak intensities changes in 2D NMR and used for quantifying the distance of the protein residues to the spin-lable on Cholesterol. Computational ligand docking guided by such distance restraint could provide atomic detailed interaction mechanism for cholesterol.

5.2.3 Characterization of TSPO's dynamic ensemble is important to understand its ligand binding behavior

The possible that conformational exchange of TSPO in its imaging ligand bound state might influence its functional role. Although in the observed crystal structures, TSPO does seem to have significant change in conformation upon ligand binding, suggesting a change in the signaling via structure is unlikely. The endogenous ligands such as cholesterol has no structural information of its binding mode experimentally determined. It is uncertain whether the currently reported structures represent a conformational species that belongs to the conformational assembly that performs its native biological functions, due to the disruptive nature

of crystallography and NMR structure determination. To date, the known natural ligands include cholesterol and porphyrin for mammalian TSPO, and porphyrin for bacterial TSPO. However, other endogenous ligand might exist as discoveries regarding the TSPO's function being explored in the new paradigm.

While cholesterol showed binding on mTSPO and influencing its structure [95], it is not clear if signal would propagate to downstream co-factors or facilitate cholesterol transport. Current structural biology effort focuses on mTSPO alone, while in mitochondria metabolism, mTSPO interacts with VDAC and ANT to for the mitochondria permeability transition pore (MPTP) complex [136, 229]. Future structural biology effort should also consider characterizing the molecular assembly of TSPO in the context of MPTP.

The apo-state of TSPO reported in crystal structures does not show large structural changes upon ligand binding. The dynamic nature of mammalian TSPO in its apo state in detergent is difficult to track by NMR alone, as suggested by the previous studies that lacks a defined tertiary structure [92]. While detergent systems destabilize mTSPO to various extents, proteoliposome preparations seems to maintain the structural integrity and function. EPR spectroscopy experiments on mTSPO in proteoliposomes could provide an alternative approach to understanding the dynamics of the mTSPO and to accessing its structural ensemble in both apo and ligand-bound states. For example, continuous wave (CW) EPR experiments on single site-directed spin labeled (SDSL) protein could provide data in the dynamic changes, while DEER EPR experiments on double SDSL protein could provide distance restraints for ensemble of protein in structural

transitions upon ligand binding. Additionally, cholesterol incorporation into proteoliposomes is robust. Investigating TSPO dynamic changes in cholesterol containing membrane could provide additional information for the structural transitions proposed by solid-state NMR studies. One could measure sets of DEER EPR distance data of mTSPO in its apo-state in liposomes, apo-state in cholesterol containing liposomes, PK11195-bound-state in liposomes, and PK11195-bound-state in cholesterol containing liposomes. The resulting dataset could explain potential structural functional modulation of imaging ligand to mTSPO's binding capacity to cholesterol.

The low resolution and sparse nature of the EPR technique would need to be complemented by other experimental techniques and computational modeling for building reliable models. The approach has been successfully applied to various receptors and transporter structural refinement [34, 220]. To study other TSPO ligand-protein interaction, more experimental structure and computational refinement of ligand binding pocket is needed.

In conclusion, TSPO remains a difficult target for structural characterization. For solution-state NMR, optimization effort must be taken to ensure a stabilized functional mTSPO in NMR compatible systems. While such further screening is needed, alternative experimental investigations, using established liposome systems could generate valuable structural information of mTSPO in response to ligand binding.

BIBLIOGRAPY

1. Tusnady, G.E., Z. Dosztanyi, and I. Simon, *Transmembrane proteins in the Protein Data Bank: identification and classification.* Bioinformatics, 2004. **20**(17): p. 2964-72.
2. White, S.H., *Membrane proteins of known 3-D structure.* 2006.
3. Weiner, B.E., et al., *BCL::MP-Fold - Folding Membrane Proteins through Assembly of Transmembrane Helices.* submitted.
4. Hopf, T.A., et al., *Three-Dimensional Structures of Membrane Proteins from Genomic Sequencing.* Cell, 2012.
5. Berman, H.M., et al., *The Protein Data Bank.* Acta Crystallogr D Biol Crystallogr, 2002. **58**(Pt 6 No 1): p. 899-907.
6. Arora, A. and L.K. Tamm, *Biophysical approaches to membrane protein structure determination.* Curr Opin Struct Biol, 2001. **11**(5): p. 540-7.
7. Bordignon, E. and H.-J. Steinhoff, *Membrane Protein Structure and Dynamics Studied by Site-Directed Spin Labeling ESR*, in *ESR Spectroscopy in Membrane Biophysics*, M.A. Hemminga and L.J. Berliner, Editors. 2007, Springer Science and Business Media: New York. p. 129 - 164.
8. Landau, E.M. and J.P. Rosenbusch, *Lipidic cubic phases: a novel concept for the crystallization of membrane proteins.* Proc Natl Acad Sci U S A, 1996. **93**(25): p. 14532-5.
9. Hong, M., *Structure, topology, and dynamics of membrane peptides and proteins from solid-state NMR spectroscopy.* J Phys Chem B, 2007. **111**(35): p. 10340-51.
10. Oxenoid, K., et al., *NMR assignments for a helical 40 kDa membrane protein.* J Am Chem Soc, 2004. **126**(16): p. 5048-9.
11. MacKenzie, K.R., J.H. Prestegard, and D.M. Engelman, *A transmembrane helix dimer: structure and implications.* Science, 1997. **276**(5309): p. 131-3.
12. Arora, A., et al., *Structure of outer membrane protein A transmembrane domain by NMR spectroscopy.* Nat Struct Biol, 2001. **8**(4): p. 334-8.
13. Liang, B., J.H. Bushweller, and L.K. Tamm, *Site-directed parallel spin-labeling and paramagnetic relaxation enhancement in structure determination of membrane proteins by solution NMR spectroscopy.* J Am Chem Soc, 2006. **128**(13): p. 4389-97.
14. Oxenoid, K. and J.J. Chou, *The structure of phospholamban pentamer reveals a channel-like architecture in membranes.* Proc Natl Acad Sci U S A, 2005. **102**(31): p. 10870-5.
15. Van Horn, W.D., et al., *Solution nuclear magnetic resonance structure of membrane-integral diacylglycerol kinase.* Science, 2009. **324**(5935): p. 1726-9.
16. Sanders, C.R. and F. Sonnichsen, *Solution NMR of membrane proteins: practice and challenges.* Magn Reson Chem, 2006. **44 Spec No**: p. S24-40.
17. Baneyx, F., *Recombinant protein expression in Escherichia coli.* Curr Opin Biotechnol, 1999. **10**(5): p. 411-21.
18. Andre, N., et al., *Enhancing functional production of G protein-coupled receptors in Pichia pastoris to levels required for structural studies via a single expression screen.* Protein Sci, 2006. **15**(5): p. 1115-26.

19. Fan, Y., et al., *Uniform isotope labeling of a eukaryotic seven-transmembrane helical protein in yeast enables high-resolution solid-state NMR studies in the lipid environment.* J Biomol NMR, 2011. **49**(2): p. 151-61.
20. Kost, T.A., J.P. Condreay, and D.L. Jarvis, *Baculovirus as versatile vectors for protein expression in insect and mammalian cells.* Nat Biotechnol, 2005. **23**(5): p. 567-75.
21. Lackner, A., et al., *A bicistronic baculovirus vector for transient and stable protein expression in mammalian cells.* Anal Biochem, 2008. **380**(1): p. 146-8.
22. Schmidt, P., et al., *Prokaryotic expression, in vitro folding, and molecular pharmacological characterization of the neuropeptide Y receptor type 2.* Biotechnology Progress, 2009: p. NA-NA.
23. Mannella, C.A., *Structure of the mitochondrial outer membrane channel derived from electron microscopy of 2D crystals.* J Bioenerg Biomembr, 1989. **21**(4): p. 427-37.
24. Dvir, H. and S. Choe, *Bacterial Expression of a Eukaryotic Membrane Protein in Fusion to various Mistic Orthologs.* Protein expression and purification, 2009. **68**(1): p. 28-33.
25. Tilegenova, C., et al., *An improved method for the cost-effective expression and purification of large quantities of KcsA.* Protein expression and purification, 2016. **127**: p. 53-60.
26. Serrano-Vega, M.J., et al., *Conformational thermostabilization of the β1-adrenergic receptor in a detergent-resistant form.* Proceedings of the National Academy of Sciences, 2008. **105**(3): p. 877.
27. Shibata, Y., et al., *Thermostabilization of the Neurotensin Receptor NTS1.* Journal of Molecular Biology, 2009. **390**(2): p. 262-277.
28. Focke, P.J., et al., *Combining in Vitro Folding with Cell Free Protein Synthesis for Membrane Protein Expression.* Biochemistry, 2016. **55**(30): p. 4212-9.
29. Shinoda, T., et al., *Cell-free methods to produce structurally intact mammalian membrane proteins.* Scientific Reports, 2016. **6**: p. 30442.
30. Loscha, K.V., et al., *Multiple-site labeling of proteins with unnatural amino acids.* Angew Chem Int Ed Engl, 2012. **51**(9): p. 2243-6.
31. Nasr, M.L., et al., *Covalently circularized nanodiscs for studying membrane proteins and viral entry.* Nature methods, 2017. **14**(1): p. 49-52.
32. Popot, J.-L., *Amphipols, Nanodiscs, and Fluorinated Surfactants: Three Nonconventional Approaches to Studying Membrane Proteins in Aqueous Solutions.* Annual Review of Biochemistry, 2010. **79**(1): p. 737-775.
33. Lindert, S., et al., *EM-fold: de novo atomic-detail protein structure determination from medium-resolution density maps.* Structure, 2012. **20**(3): p. 464-78.
34. 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.
35. Barth, P., B. Wallner, and D. Baker, *Prediction of membrane protein structures with complex topologies using limited constraints.* Proc Natl Acad Sci U S A, 2009. **106**(5): p. 1409-14.
36. Koehler Leman, J., M.B. Ulmschneider, and J.J. Gray, *Computational modeling of membrane proteins.* Proteins, 2015. **83**(1): p. 1-24.
37. Manglik, A., et al., *Structural Insights into the Dynamic Process of beta2-*

Adrenergic Receptor Signaling. Cell, 2015. **161**(5): p. 1101-11.
38. Kaiser, A., et al., *Unwinding of the C-Terminal Residues of Neuropeptide Y is critical for Y2 Receptor Binding and Activation.* Angewandte Chemie International Edition, 2015: p. n/a-n/a.
39. 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.
40. Dror, R.O., et al., *SIGNAL TRANSDUCTION. Structural basis for nucleotide exchange in heterotrimeric G proteins.* Science, 2015. **348**(6241): p. 1361-5.
41. Shukla, A.K., et al., *Visualization of arrestin recruitment by a G-protein-coupled receptor.* Nature, 2014. **512**(7513): p. 218-22.
42. Manglik, A. and B. Kobilka, *The role of protein dynamics in GPCR function: insights from the beta2AR and rhodopsin.* Curr Opin Cell Biol, 2014. **27**: p. 136-43.
43. Nygaard, R., et al., *The dynamic process of beta(2)-adrenergic receptor activation.* Cell, 2013. **152**(3): p. 532-42.
44. Altenbach, C., et al., *High-resolution distance mapping in rhodopsin reveals the pattern of helix movement due to activation.* Proc Natl Acad Sci U S A, 2008. **105**(21): p. 7439-44.
45. El Moustaine, D., et al., *Distinct roles of metabotropic glutamate receptor dimerization in agonist activation and G-protein coupling.* Proc Natl Acad Sci U S A, 2012. **109**(40): p. 16342-7.
46. Xue, L., et al., *Major ligand-induced rearrangement of the heptahelical domain interface in a GPCR dimer.* Nat Chem Biol, 2015. **11**(2): p. 134-40.
47. Das, R. and D. Baker, *Macromolecular modeling with rosetta.* Annual Review of Biochemistry, 2008. **77**: p. 363-382.
48. 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.
49. Leaver-Fay, A., et al., *Scientific benchmarks for guiding macromolecular energy function improvement.* Methods Enzymol, 2013. **523**: p. 109-43.
50. Song, Y., et al., *High-resolution comparative modeling with RosettaCM.* Structure, 2013. **21**(10): p. 1735-42.
51. Song, Y., et al., *Structure-guided forcefield optimization.* Proteins, 2011. **79**(6): p. 1898-909.
52. Yarov-Yarovoy, V., J. Schonbrun, and D. Baker, *Multipass membrane protein structure prediction using Rosetta.* Proteins, 2006. **62**(4): p. 1010-25.
53. DiMaio, F., et al., *Refinement of protein structures into low-resolution density maps using rosetta.* Journal of Molecular Biology, 2009. **392**(1): p. 181-190.
54. Bowers, P.M., C.E. Strauss, and D. Baker, *De novo protein structure determination using sparse NMR data.* J Biomol NMR, 2000. **18**(4): p. 311-8.
55. Schmitz, C., et al., *Protein structure determination from pseudocontact shifts using ROSETTA.* J Mol Biol, 2012. **416**(5): p. 668-77.
56. Alexander, N., et al., *De novo high-resolution protein structure determination from sparse spin-labeling EPR data.* Structure, 2008. **16**(2): p. 181-95.
57. Lemmon, G. and J. Meiler, *Rosetta Ligand docking with flexible XML protocols.* Methods Mol Biol, 2012. **819**: p. 143-55.
58. Combs, S.A., et al., *Small-molecule ligand docking into comparative models with*

Rosetta. Nat Protoc, 2013. **8**(7): p. 1277-98.
59. Meiler, J. and D. Baker, *ROSETTALIGAND: protein-small molecule docking with full side-chain flexibility.* Proteins, 2006. **65**(3): p. 538-48.
60. Braestrup, C. and R.F. Squires, *Specific benzodiazepine receptors in rat brain characterized by high-affinity (3H)diazepam binding.* Proc Natl Acad Sci U S A, 1977. **74**(9): p. 3805-9.
61. Papadopoulos, V., et al., *Translocator protein (18 kDa): new nomenclature for the peripheral-type benzodiazepine receptor based on its structure and molecular function.* Trends in Pharmacological Sciences, 2006. **27**(8): p. 402-409.
62. Fan, J., et al., *Structural and functional evolution of the translocator protein (18 kDa).* Curr Mol Med, 2012. **12**(4): p. 369-86.
63. Anholt, R.R., et al., *The peripheral-type benzodiazepine receptor. Localization to the mitochondrial outer membrane.* J Biol Chem, 1986. **261**(2): p. 576-83.
64. Papadopoulos, V., et al., *Peripheral benzodiazepine receptor in cholesterol transport and steroidogenesis.* Steroids, 1997. **62**(1).
65. Li, H. and V. Papadopoulos, *Peripheral-type benzodiazepine receptor function in cholesterol transport. Identification of a putative cholesterol recognition/interaction amino acid sequence and consensus patterns.* Endocrinology, 1998. **139**(12): p. 4991-4997.
66. Taketani, S., et al., *Involvement of peripheral-type benzodiazepine receptors in the intracellular transport of heme and porphyrins.* J Biochem, 1995. **117**(4): p. 875-80.
67. Verma, A., J.S. Nye, and S.H. Snyder, *Porphyrins Are Endogenous Ligands for the Mitochondrial (Peripheral-Type) Benzodiazepine Receptor.* Proceedings of the National Academy of Sciences of the United States of America, 1987. **84**(8): p. 2256-2260.
68. Chen, M.K. and T.R. Guilarte, *Translocator protein 18 kDa (TSPO): molecular sensor of brain injury and repair.* Pharmacol Ther, 2008. **118**(1): p. 1-17.
69. Brown, R.C., et al., *Location-dependent role of the human glioma cell peripheral-type benzodiazepine receptor in proliferation and steroid biosynthesis.* Cancer Lett, 2000. **156**(2): p. 125-32.
70. Hardwick, M., et al., *Peripheral-type benzodiazepine receptor (PBR) in human breast cancer: correlation of breast cancer cell aggressive phenotype with PBR expression, nuclear localization, and PBR-mediated cell proliferation and nuclear transport of cholesterol.* Cancer Res, 1999. **59**(4): p. 831-42.
71. Ji, B., et al., *Imaging of peripheral benzodiazepine receptor expression as biomarkers of detrimental versus beneficial glial responses in mouse models of Alzheimer's and other CNS pathologies.* J Neurosci, 2008. **28**(47): p. 12255-67.
72. Crompton, M., *The mitochondrial permeability transition pore and its role in cell death.* Biochem J, 1999. **341 (Pt 2)**: p. 233-49.
73. Decaudin, D., et al., *Peripheral benzodiazepine receptor ligands reverse apoptosis resistance of cancer cells in vitro and in vivo.* Cancer Research, 2002. **62**(5): p. 1388-1393.
74. Campanella, M., *Editorial: The physiology and pharmacology of the mitochondrial 18 kDa translocator protein (TSPO): an emerging molecular target for diagnosis and therapy.* Curr Mol Med, 2012. **12**(4): p. 355.

75. Gatliff, J. and M. Campanella, *The 18 kDa translocator protein (TSPO): a new perspective in mitochondrial biology.* Curr Mol Med, 2012. **12**(4): p. 356-68.
76. Rupprecht, R., et al., *Translocator protein (18 kDa) (TSPO) as a therapeutic target for neurological and psychiatric disorders.* Nat Rev Drug Discov, 2010. **9**(12): p. 971-88.
77. Xiao, J., et al., *4'-Chlorodiazepam, a translocator protein (18 kDa) antagonist, improves cardiac functional recovery during postischemia reperfusion in rats.* Exp Biol Med (Maywood), 2010. **235**(4): p. 478-86.
78. Scarf, A.M., et al., *Evidence for complex binding profiles and species differences at the translocator protein (TSPO) (18 kDa).* Curr Mol Med, 2012. **12**(4): p. 488-93.
79. Scarf, A.M., K.M. Auman, and M. Kassiou, *Is there any correlation between binding and functional effects at the translocator protein (TSPO) (18 kDa)?* Curr Mol Med, 2012. **12**(4): p. 387-97.
80. Gonzalez-Polo, R.A., et al., *PK11195 potently sensitizes to apoptosis induction independently from the peripheral benzodiazepin receptor.* Oncogene, 2005. **24**(51): p. 7503-13.
81. Decaudin, D., et al., *Peripheral benzodiazepine receptor ligands reverse apoptosis resistance of cancer cells in vitro and in vivo.* Cancer Res, 2002. **62**(5): p. 1388-93.
82. Obame, F.N., et al., *Peripheral benzodiazepine receptor-induced myocardial protection is mediated by inhibition of mitochondrial membrane permeabilization.* J Pharmacol Exp Ther, 2007. **323**(1): p. 336-45.
83. Schaller, S., et al., *TRO40303, a new cardioprotective compound, inhibits mitochondrial permeability transition.* J Pharmacol Exp Ther, 2010. **333**(3): p. 696-706.
84. Dupont, A.-C., et al., *Translocator Protein-18 kDa (TSPO) Positron Emission Tomography (PET) Imaging and Its Clinical Impact in Neurodegenerative Diseases.* International Journal of Molecular Sciences, 2017. **18**(4): p. 785.
85. Lacapère, J.-J., et al., *Structural and Functional Study of Reconstituted Peripheral Benzodiazepine Receptor.* Biochemical and Biophysical Research Communications, 2001. **284**(2): p. 536-541.
86. Jamin, N., et al., *Characterization of the cholesterol recognition amino acid consensus sequence of the peripheral-type benzodiazepine receptor.* Mol Endocrinol, 2005. **19**(3): p. 588-94.
87. Murail, S., et al., *Secondary and tertiary structures of the transmembrane domains of the translocator protein TSPO determined by NMR. Stabilization of the TSPO tertiary fold upon ligand binding.* Biochimica et biophysica acta, 2008. **1778**(6): p. 1375-1381.
88. Korkhov, V.M., et al., *Three-dimensional structure of TspO by electron cryomicroscopy of helical crystals.* Structure, 2010. **18**(6): p. 677-87.
89. Li, F., et al., *Characterization and modeling of the oligomeric state and ligand binding behavior of purified translocator protein 18 kDa from rhodobacter sphaeroides.* Biochemistry, 2013. **52**(34): p. 5884-5899.
90. Jaremko, L., et al., *Structure of the mitochondrial translocator protein in complex with a diagnostic ligand.* Science, 2014. **343**(6177): p. 1363-6.
91. Jaremko, M., et al., *Structural Integrity of the A147T Polymorph of Mammalian TSPO.* Chembiochem, 2015. **16**(10): p. 1483-9.

92. Jaremko, L., et al., *Conformational Flexibility in the Transmembrane Protein TSPO.* Chemistry, 2015. **21**(46): p. 16555-63.
93. Guo, Y., et al., *Protein structure. Structure and activity of tryptophan-rich TSPO proteins.* Science, 2015. **347**(6221): p. 551-5.
94. Li, F., et al., *Protein structure. Crystal structures of translocator protein (TSPO) and mutant mimic of a human polymorphism.* Science, 2015. **347**(6221): p. 555-8.
95. Jaipuria, G., et al., *Cholesterol-mediated allosteric regulation of the mitochondrial translocator protein structure.* Nat Commun, 2017. **8**: p. 14893.
96. Graeber, E. and V.M. Korkhov, *Expression and purification of the mammalian translocator protein for structural studies.* PLoS One, 2018. **13**(6): p. e0198832.
97. Xia, Y., et al., *Integrated Structural Biology for α-Helical Membrane Protein Structure Determination.* Structure, 2018. **26**(4): p. 657-666.e2.
98. Bill, R.M., et al., *Overcoming barriers to membrane protein structure determination.* Nat Biotechnol, 2011. **29**(4): p. 335-40.
99. Wiener, M.C., *A pedestrian guide to membrane protein crystallization.* Methods, 2004. **34**(3): p. 364-72.
100. Loll, P.J., *Membrane protein structural biology: the high throughput challenge.* J Struct Biol, 2003. **142**(1): p. 144-53.
101. White, S.H., *The progress of membrane protein structure determination.* Protein Sci, 2004. **13**(7): p. 1948-9.
102. Grant, A., D. Lee, and C. Orengo, *Progress towards mapping the universe of protein folds.* Genome Biol, 2004. **5**(5): p. 107.
103. Stevens, R.C., et al., *The GPCR Network: a large-scale collaboration to determine human GPCR structure and function.* Nat Rev Drug Discov, 2013. **12**(1): p. 25-34.
104. Oberai, A., et al., *A limited universe of membrane protein families and folds.* Protein Sci, 2006. **15**(7): p. 1723-34.
105. Khafizov, K., et al., *Trends in structural coverage of the protein universe and the impact of the Protein Structure Initiative.* Proc Natl Acad Sci U S A, 2014. **111**(10): p. 3733-8.
106. Kloppmann, E., M. Punta, and B. Rost, *Structural genomics plucks high-hanging membrane proteins.* Curr Opin Struct Biol, 2012. **22**(3): p. 326-32.
107. Pieper, U., et al., *Coordinating the impact of structural genomics on the human alpha-helical transmembrane proteome.* Nat Struct Mol Biol, 2013. **20**(2): p. 135-8.
108. Weiner, B., et al., *BCL::MP-fold: folding membrane proteins through assembly of transmembrane helices.* Structure (London, England : 1993), 2013. **21**(7): p. 1107-1117.
109. Karakas, M., et al., *BCL::Fold--de novo prediction of complex and large protein topologies by assembly of secondary structure elements.* PLoS ONE, 2012. **7**(11): p. e49240.
110. Lindert, S., et al., *EM-fold: De novo folding of alpha-helical proteins guided by intermediate-resolution electron microscopy density maps.* Structure, 2009. **17**(7): p. 990-1003.
111. Weiner, B.E., et al., *BCL::Fold--protein topology determination from limited NMR restraints.* Proteins, 2014. **82**(4): p. 587-95.
112. Fischer, A.W., et al., *BCL::MP-fold: Membrane protein structure prediction guided by EPR restraints.* Proteins, 2015. **83**(11): p. 1947-62.

113. Hofmann, T., et al., *Protein structure prediction guided by crosslinking restraints - A systematic evaluation of the impact of the crosslinking spacer length.* Methods, 2015. **89**: p. 79-90.
114. 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.
115. 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.
116. Viklund, H., et al., *SPOCTOPUS: a combined predictor of signal peptides and membrane protein topology.* Bioinformatics, 2008. **24**(24): p. 2928-9.
117. Jayasinghe, S., K. Hristova, and S.H. White, *MPtopo: A database of membrane protein topology.* Protein Sci, 2001. **10**(2): p. 455-8.
118. Li, J., et al., *Structure of bovine rhodopsin in a trigonal crystal form.* J Mol Biol, 2004. **343**(5): p. 1409-38.
119. Ruprecht, J.J., et al., *Electron crystallography reveals the structure of metarhodopsin I.* Embo j, 2004. **23**(18): p. 3609-20.
120. Shen, Y. and A. Bax, *SPARTA+: a modest improvement in empirical NMR chemical shift prediction by means of an artificial neural network.* J Biomol NMR, 2010. **48**(1): p. 13-22.
121. Koehler, J., et al., *A unified hydrophobicity scale for multispan membrane proteins.* Proteins, 2009. **76**(1): p. 13-29.
122. Carugo, O. and S. Pongor, *A normalized root-mean-square distance for comparing protein three-dimensional structures.* Protein Sci, 2001. **10**(7): p. 1470-3.
123. Vernon, R., et al., *Improved chemical shift based fragment selection for CS-Rosetta using Rosetta3 fragment picker.* J Biomol NMR, 2013. **57**(2): p. 117-27.
124. Canutescu, A.A. and R.L. Dunbrack, Jr., *Cyclic coordinate descent: A robotics algorithm for protein loop closure.* Protein Sci, 2003. **12**(5): p. 963-72.
125. 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.
126. Schrodinger, LLC, *The PyMOL Molecular Graphics System, Version 1.8*. 2015.
127. Gautier, A. and D. Nietlispach, *Solution NMR studies of integral polytopic alpha-helical membrane proteins: the structure determination of the seven-helix transmembrane receptor sensory rhodopsin II, pSRII.* Methods Mol Biol, 2012. **914**: p. 25-45.
128. Yeagle, P.L., G. Choi, and A.D. Albert, *Studies on the structure of the G-protein-coupled receptor rhodopsin including the putative G-protein binding site in unactivated and activated forms.* Biochemistry, 2001. **40**(39): p. 11932-7.
129. Rohl, C.A., et al., *Protein structure prediction using Rosetta.* Methods Enzymol, 2004. **383**: p. 66-93.
130. Batarseh, A. and V. Papadopoulos, *Regulation of translocator protein 18 kDa (TSPO) expression in health and disease states.* Mol Cell Endocrinol, 2010. **327**(1-2): p. 1-12.
131. Ching, A.S., et al., *Current paradigm of the 18-kDa translocator protein (TSPO) as a molecular target for PET imaging in neuroinflammation and neurodegenerative diseases.* Insights Imaging, 2012. **3**(1): p. 111-9.
132. Galiegue, S., N. Tinel, and P. Casellas, *The peripheral benzodiazepine receptor: a*

promising therapeutic drug target. Curr Med Chem, 2003. **10**(16): p. 1563-72.
133. Fafalios, A., et al., *Translocator protein blockade reduces prostate tumor growth.* Clin Cancer Res, 2009. **15**(19): p. 6177-84.
134. Gavish, M., et al., *Enigma of the peripheral benzodiazepine receptor.* Pharmacol Rev, 1999. **51**(4): p. 629-50.
135. McEnery, M.W., et al., *Isolation of the mitochondrial benzodiazepine receptor: association with the voltage-dependent anion channel and the adenine nucleotide carrier.* Proc Natl Acad Sci U S A, 1992. **89**(8): p. 3170-4.
136. Veenman, L. and M. Gavish, *The role of 18 kDa mitochondrial translocator protein (TSPO) in programmed cell death, and effects of steroids on TSPO expression.* Curr Mol Med, 2012. **12**(4): p. 398-412.
137. Lacapère, J.-J. and V. Papadopoulos, *Peripheral-type benzodiazepine receptor: structure and function of a cholesterol-binding protein in steroid and bile acid biosynthesis.* Steroids, 2003. **68**(7-8): p. 569-585.
138. Morohaku, K., et al., *Translocator protein/peripheral benzodiazepine receptor is not required for steroid hormone biosynthesis.* Endocrinology, 2014. **155**(1): p. 89-97.
139. Tu, L.N., et al., *Peripheral benzodiazepine receptor/translocator protein global knock-out mice are viable with no effects on steroid hormone biosynthesis.* J Biol Chem, 2014. **289**(40): p. 27444-54.
140. Tu, L.N., et al., *PK11195 Effect on Steroidogenesis Is Not Mediated Through the Translocator Protein (TSPO).* Endocrinology, 2015. **156**(3): p. 1033-9.
141. Sileikyte, J., et al., *Regulation of the Mitochondrial Permeability Transition Pore by the Outer Membrane does not Involve the Peripheral Benzodiazepine Receptor (TSPO).* J Biol Chem, 2014.
142. Giorgio, V., et al., *Dimers of mitochondrial ATP synthase form the permeability transition pore.* Proc Natl Acad Sci U S A, 2013. **110**(15): p. 5887-92.
143. Cleary, J., et al., *Inhibition of the mitochondrial F1F0-ATPase by ligands of the peripheral benzodiazepine receptor.* Bioorg Med Chem Lett, 2007. **17**(6): p. 1667-70.
144. Hatty, C.R., et al., *Investigating the interactions of the 18kDa translocator protein and its ligand PK11195 in planar lipid bilayers.* Biochim Biophys Acta, 2014. **1838**(3): p. 1019-30.
145. Liu, G.J., et al., *The 18 kDa translocator protein, microglia and neuroinflammation.* Brain Pathol, 2014. **24**(6): p. 631-53.
146. Banati, R.B., et al., *Positron emission tomography and functional characterization of a complete PBR/TSPO knockout.* Nat Commun, 2014. **5**: p. 5452.
147. Black, K.L., K. Ikezaki, and A.W. Toga, *Imaging of brain tumors using peripheral benzodiazepine receptor ligands.* J Neurosurg, 1989. **71**(1): p. 113-8.
148. Starostarubinstein, S., et al., *Imaging of a Glioma Using Peripheral Benzodiazepine Receptor Ligands.* Proc Natl Acad Sci U S A, 1987. **84**(3): p. 891-895.
149. Carmel, I., et al., *Peripheral-type benzodiazepine receptors in the regulation of proliferation of MCF-7 human breast carcinoma cell line.* Biochem Pharmacol, 1999. **58**(2): p. 273-8.
150. Deane, N.G., et al., *Targeted imaging of colonic tumors in smad3-/- mice discriminates cancer and inflammation.* Mol Cancer Res, 2007. **5**(4): p. 341-9.

151. Maaser, K., et al., *Specific ligands of the peripheral benzodiazepine receptor induce apoptosis and cell cycle arrest in human colorectal cancer cells.* British journal of cancer., 2001. **85(11)**: p. 1771-80.
152. Maaser, K., et al., *Overexpression of the peripheral benzodiazepine receptor is a relevant prognostic factor in stage III colorectal cancer.* Clin Cancer Res, 2002. **8**(10): p. 3205-9.
153. Junck, L., et al., *Pet Imaging of Human Gliomas with Ligands for the Peripheral Benzodiazepine Binding-Site.* Annals of Neurology, 1989. **26**(6): p. 752-758.
154. Han, Z., et al., *Expression of peripheral benzodiazepine receptor (PBR) in human tumors: relationship to breast, colorectal, and prostate tumor progression.* J Recept Signal Transduct Res, 2003. **23**(2-3): p. 225-38.
155. Hardwick, M., et al., *Peripheral-type benzodiazepine receptor levels correlate with the ability of human breast cancer MDA-MB-231 cell line to grow in SCID mice.* International Journal of Cancer, 2001. **94**(3): p. 322-327.
156. Tang, D., et al., *Quantitative preclinical imaging of TSPO expression in glioma using N,N-diethyl-2-(2-(4-(2-18F-fluoroethoxy)phenyl)-5,7-dimethylpyrazolo[1,5-a]pyrimi din-3-yl)acetamide.* J Nucl Med, 2012. **53**(2): p. 287-94.
157. Tang, D., et al., *Preclinical Imaging Evaluation of Novel TSPO-PET Ligand 2-(5,7-Diethyl-2-(4-(2-[F]fluoroethoxy)phenyl)pyrazolo[1,5-a]pyrimidin-3-yl)-N,N-diethylacetamide ([F]VUIIS1008) in Glioma.* Mol Imaging Biol, 2014.
158. Tang, D., et al., *Synthesis and structure-activity relationships of 5,6,7-substituted pyrazolopyrimidines: discovery of a novel TSPO PET ligand for cancer imaging.* J Med Chem, 2013. **56**(8): p. 3429-33.
159. Cheung, Y.-Y., et al., *Preclinical evaluation of 7-chloro-N,N,5-trimethyl-4-oxo-3(6-[18F]fluoropyridin-2-yl)-3,5-dihydro-4H-pyridazino[4,5-b]indole-1-acetamide: A novel pyridazinoindole ligand for PET imaging of TSPO in cancer.* Journal of Nuclear Medicine, 2014. **55**(supplement 1): p. 1053.
160. Jaremko, Ł., et al., *Structure of the Mitochondrial Translocator Protein in Complex with a Diagnostic Ligand.* Science, 2014. **343**(6177): p. 1363-1366.
161. Teboul, D., et al., *Mouse TSPO in a lipid environment interacting with a functionalized monolayer.* Biochimica et biophysica acta, 2012. **1818**(11): p. 2791-2800.
162. Papadopoulos, V., et al., *Topography of the Leydig cell mitochondrial peripheral-type benzodiazepine receptor.* Mol Cell Endocrinol, 1994. **104**(1): p. R5-9.
163. Murail, S., et al., *Secondary and tertiary structures of the transmembrane domains of the translocator protein TSPO determined by NMR. Stabilization of the TSPO tertiary fold upon ligand binding.* Biochim Biophys Acta, 2008. **1778**(6): p. 1375-81.
164. Holm, L. and P.i. Rosenstrï¿½m, *Dali server: conservation mapping in 3D.* Nucleic Acids Research, 2010. **38**(suppl_2): p. W545-W549.
165. Molecular Operating Environment, 2013.08; Chemical Computing Group Inc., 1010 Sherbooke St. West, Suite #910, Montreal, QC, Canada, H3A 2R7, 2013.
166. 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.

167. Lagerstrom, M.C. and H.B. Schioth, *Structural diversity of G protein-coupled receptors and significance for drug discovery.* Nat Rev Drug Discov, 2008. **7**(4): p. 339-57.
168. Kniazeff, J., et al., *Dimers and beyond: The functional puzzles of class C GPCRs.* Pharmacol Ther, 2011. **130**(1): p. 9-25.
169. Muto, T., et al., *Structures of the extracellular regions of the group II/III metabotropic glutamate receptors.* Proc Natl Acad Sci U S A, 2007. **104**(10): p. 3759-64.
170. Houamed, K.M., et al., *Cloning, expression, and gene structure of a G protein-coupled glutamate receptor from rat brain.* Science, 1991. **252**(5010): p. 1318-21.
171. Masu, M., et al., *Sequence and expression of a metabotropic glutamate receptor.* Nature, 1991. **349**(6312): p. 760-5.
172. Nakanishi, S., *Molecular diversity of glutamate receptors and implications for brain function.* Science, 1992. **258**(5082): p. 597-603.
173. Niswender, C.M. and P.J. Conn, *Metabotropic glutamate receptors: physiology, pharmacology, and disease.* Annu Rev Pharmacol Toxicol, 2010. **50**: p. 295-322.
174. Dolen, G., et al., *Mechanism-based approaches to treating fragile X.* Pharmacol Ther, 2010. **127**(1): p. 78-93.
175. Satoh, A., et al., *Discovery and in vitro and in vivo profiles of 4-fluoro-N-[4-[6-(isopropylamino)pyrimidin-4-yl]-1,3-thiazol-2-yl]-N-methylbenzam ide as novel class of an orally active metabotropic glutamate receptor 1 (mGluR1) antagonist.* Bioorg Med Chem Lett, 2009. **19**(18): p. 5464-8.
176. Materials and methods are available as supplementary materials on Science Online.
177. Katritch, V., V. Cherezov, and R.C. Stevens, *Structure-function of the G protein-coupled receptor superfamily.* Annu Rev Pharmacol Toxicol, 2013. **53**: p. 531-56.
178. Palczewski, K., et al., *Crystal structure of rhodopsin: A G protein-coupled receptor.* Science, 2000. **289**(5480): p. 739-45.
179. Siu, F.Y., et al., *Structure of the human glucagon class B G-protein-coupled receptor.* Nature, 2013. **499**(7459): p. 444-9.
180. Hollenstein, K., et al., *Structure of class B GPCR corticotropin-releasing factor receptor 1.* Nature, 2013. **499**(7459): p. 438-43.
181. Wang, C., et al., *Structure of the human smoothened receptor bound to an antitumour agent.* Nature, 2013. **497**(7449): p. 338-43.
182. Molecular Operating Environment, 2013.08; Chemical Computing Group Inc., 1010 Sherbooke St. West, Suite #910, Montreal, QC, Canada, H3A 2R7, 2013.
183. DiMaio, F., et al., *Modeling symmetric macromolecular structures in Rosetta3.* PLoS ONE, 2011. **6**(6).
184. Huang, S., et al., *Interdomain movements in metabotropic glutamate receptor activation.* Proc Natl Acad Sci U S A, 2011. **108**(37): p. 15480-5.
185. Wu, H., et al., *Structure of the human kappa-opioid receptor in complex with JDTic.* Nature, 2012. **485**(7398): p. 327-32.
186. Wu, B., et al., *Structures of the CXCR4 chemokine GPCR with small-molecule and cyclic peptide antagonists.* Science, 2010. **330**(6007): p. 1066-71.
187. Litschig, S., et al., *CPCCOEt, a noncompetitive metabotropic glutamate receptor 1 antagonist, inhibits receptor signaling without affecting glutamate binding.* Mol Pharmacol, 1999. **55**(3): p. 453-61.

188. Pagano, A., et al., *The non-competitive antagonists 2-methyl-6-(phenylethynyl)pyridine and 7-hydroxyiminocyclopropan[b]chromen-1a-carboxylic acid ethyl ester interact with overlapping binding pockets in the transmembrane region of group I metabotropic glutamate receptors.* J Biol Chem, 2000. **275**(43): p. 33750-8.
189. Surin, A., et al., *Cyclothiazide selectively inhibits mGluR1 receptors interacting with a common allosteric site for non-competitive antagonists.* Neuropharmacology, 2007. **52**(3): p. 744-54.
190. Malherbe, P., et al., *Comparison of the binding pockets of two chemically unrelated allosteric antagonists of the mGlu5 receptor and identification of crucial residues involved in the inverse agonism of MPEP.* Journal of Neurochemistry, 2006. **98**(2): p. 601-615.
191. Chen, Y., et al., *N-{4-Chloro-2-[(1,3-dioxo-1,3-dihydro-2H-isoindol-2-yl)methyl]phenyl}-2-hy droxybenzamide (CPPHA) acts through a novel site as a positive allosteric modulator of group 1 metabotropic glutamate receptors.* Mol Pharmacol, 2008. **73**(3): p. 909-18.
192. Molck, C., et al., *Pharmacological Characterization and Modeling of the Binding Sites of Novel 1,3-bis(pyridinylethynyl)benzenes as Metabotropic Glutamate Receptor 5-selective Negative Allosteric Modulators.* Molecular pharmacology, 2012.
193. Gregory, K.J., et al., *Investigating metabotropic glutamate receptor 5 allosteric modulator cooperativity, affinity, and agonism: enriching structure-function studies and structure-activity relationships.* Molecular pharmacology, 2012. **82**(5): p. 860-75.
194. Malherbe, P., et al., *Mutational analysis and molecular modeling of the binding pocket of the metabotropic glutamate 5 receptor negative modulator 2-methyl-6-(phenylethynyl)-pyridine.* Mol Pharmacol, 2003. **64**(4): p. 823-32.
195. Xu, F., et al., *Structure of an agonist-bound human A2A adenosine receptor.* Science, 2011. **332**(6027): p. 322-7.
196. Kunishima, N., et al., *Structural basis of glutamate recognition by a dimeric metabotropic glutamate receptor.* Nature, 2000. **407**(6807): p. 971-7.
197. Doumazane, E., et al., *Illuminating the activation mechanisms and allosteric properties of metabotropic glutamate receptors.* Proc Natl Acad Sci U S A, 2013. **110**(15): p. E1416-25.
198. Eroglu, C., et al., *Glutamate-binding affinity of Drosophila metabotropic glutamate receptor is modulated by association with lipid rafts.* Proc Natl Acad Sci U S A, 2003. **100**(18): p. 10219-24.
199. Dhermain, F.G., et al., *Advanced MRI and PET imaging for assessment of treatment response in patients with gliomas.* Lancet Neurol, 2010. **9**(9): p. 906-20.
200. van den Bent, M.J., et al., *End point assessment in gliomas: novel treatments limit usefulness of classical Macdonald's Criteria.* J Clin Oncol, 2009. **27**(18): p. 2905-8.
201. Dimitrakopoulou-Strauss, A., et al., *Pharmacokinetic studies of Ga-labeled Bombesin (Ga-BZH) and F-18 FDG PET in patients with recurrent gliomas and comparison to grading: preliminary results.* Clin Nucl Med, 2011. **36**(2): p. 101-8.
202. la Fougere, C., et al., *Molecular imaging of gliomas with PET: opportunities and limitations.* Neuro Oncol, 2011. **13**(8): p. 806-19.

203. Santoni, M., et al., *[(11) C]-methionine positron emission tomography in the postoperative imaging and followup of patients with primary and recurrent gliomas.* ISRN Oncol, 2014. **2014**: p. 463152.
204. Nowosielski, M., et al., *An intra-individual comparison of MRI, [18F]-FET and [18F]-FLT PET in patients with high-grade gliomas.* PLoS One, 2014. **9**(4): p. e95830.
205. Kawai, N., et al., *Correlation between (1)(8)F-fluoromisonidazole PET and expression of HIF-1alpha and VEGF in newly diagnosed and recurrent malignant gliomas.* Eur J Nucl Med Mol Imaging, 2014. **41**(10): p. 1870-8.
206. Goldman, S. and B.J. Pirotte, *Brain tumors.* Methods Mol Biol, 2011. **727**: p. 291-315.
207. Nagler, R., et al., *Oral cancer, cigarette smoke and mitochondrial 18 kDa translocator protein (TSPO) - In vitro, in vivo, salivary analysis.* Biochim Biophys Acta, 2010. **1802**(5): p. 454-61.
208. Venturini, I., et al., *Diazepam binding inhibitor and total cholesterol plasma levels in cirrhosis and hepatocellular carcinoma.* Regul Pept, 1998. **74**(1): p. 31-4.
209. Venturini, I., et al., *Up-regulation of peripheral benzodiazepine receptor system in hepatocellular carcinoma.* Life Sci, 1998. **63**(14): p. 1269-80.
210. Nagler, R., et al., *Cigarette smoke decreases salivary 18 kDa translocator protein binding affinity--in association with oxidative stress.* Curr Med Chem, 2010. **17**(23): p. 2539-46.
211. Weiner, B., et al., *BCL::Fold-Protein topology determination from limited NMR restraints.* Proteins, 2013.
212. Kühlbrandt, W., *The Resolution Revolution.* Science, 2014. **343**(6178): p. 1443.
213. Tang, Y., et al., *Protein structure determination by combining sparse NMR data with evolutionary couplings.* Nat Methods, 2015. **12**(8): p. 751-4.
214. Marks, D.S., et al., *Protein 3D structure computed from evolutionary sequence variation.* PLoS One, 2011. **6**(12): p. e28766.
215. Kamisetty, H., S. Ovchinnikov, and D. Baker, *Assessing the utility of coevolution-based residue-residue contact predictions in a sequence- and structure-rich era.* Proc Natl Acad Sci U S A, 2013. **110**(39): p. 15674-9.
216. Morcos, F., et al., *Direct-coupling analysis of residue coevolution captures native contacts across many protein families.* Proc Natl Acad Sci U S A, 2011. **108**(49): p. E1293-301.
217. Ovchinnikov, S., et al., *Large-scale determination of previously unsolved protein structures using evolutionary information.* Elife, 2015. **4**: p. e09248.
218. Ovchinnikov, S., et al., *Protein structure determination using metagenome sequence data.* Science, 2017. **355**(6322): p. 294-298.
219. Chen, Y.J., et al., *X-ray structure of EmrE supports dual topology model.* Proc Natl Acad Sci U S A, 2007. **104**(48): p. 18999-9004.
220. Dastvan, R., et al., *Protonation-dependent conformational dynamics of the multidrug transporter EmrE.* Proc Natl Acad Sci U S A, 2016. **113**(5): p. 1220-5.
221. Zhou, A., et al., *Structure and conformational states of the bovine mitochondrial ATP synthase by cryo-EM.* Elife, 2015. **4**: p. e10180.
222. Malhotra, K. and N.N. Alder, *Reconstitution of Mitochondrial Membrane Proteins into Nanodiscs by Cell-Free Expression.* Methods in molecular biology (Clifton,

N.J.), 2017. **1567**: p. 155-178.
223. Roos, C., et al., *High-level cell-free production of membrane proteins with nanodiscs.* Methods Mol Biol, 2014. **1118**: p. 109-30.
224. Teese, M.G. and D. Langosch, *Role of GxxxG Motifs in Transmembrane Domain Interactions.* Biochemistry, 2015. **54**(33): p. 5125-5135.
225. Yano, Y., et al., *GXXXG-Mediated Parallel and Antiparallel Dimerization of Transmembrane Helices and Its Inhibition by Cholesterol: Single-Pair FRET and 2D IR Studies.* Angew Chem Int Ed Engl, 2017. **56**(7): p. 1756-1759.
226. Prakash, A., L. Janosi, and M. Doxastakis, *GxxxG Motifs, Phenylalanine, and Cholesterol Guide the Self-Association of Transmembrane Domains of ErbB2 Receptors.* Biophysical Journal, 2011. **101**(8): p. 1949-1958.
227. Barrett, P.J., et al., *The amyloid precursor protein has a flexible transmembrane domain and binds cholesterol.* Science, 2012. **336**(6085): p. 1168-71.
228. Scheidt, H.A., et al., *The potential of fluorescent and spin-labeled steroid analogs to mimic natural cholesterol.* J Biol Chem, 2003. **278**(46): p. 45563-9.
229. Veenman, L., Y. Shandalov, and M. Gavish, *VDAC activation by the 18 kDa translocator protein (TSPO), implications for apoptosis.* J Bioenerg Biomembr, 2008. **40**(3): p. 199-205.

APPENDIX

Protocol Capture for Integrated Structural Biology for Alpha-Helical Membrane Protein Structure Determination

We documented the protocol for structure determination using hybrid experimental restraints using BCL::MP-Fold and Rosetta frame work. command lines used for generating the test restraint sets for structure determination and model production using the two stage BCL::MP-Fold and Rosetta structure prediction suite. The following steps were taken to prepare the simulated restraints files used in each stage of the pipeline and running structural prediction with experimental data in BCL::MP-Fold and Rosetta. The protocol capture presented here could be used to reproduce the published data presented in chapter two.

Step	Text	Commands	Comments
1. Simulate EPR DEER distance restraints	EPR distance restraints were simulated using BCL by first predicting the optimum a.a. pairs, then simulating spin label distances with uncertainty added.	1.1 bcl.exe restraint:OptimizeDataSetPairwise -fasta 1GZM.fasta -pool secondary_structure.pool -exclude_residue_types GLYCINE -restaint_distance_structures native.ls -read_mutates_start mutate_start.table -read_mutates_optimization mutate_opt.table -read_scores_optimization score_opt_bipolar.table -read_mutates_end mutate_end.table -nmodels 100 -mc_number_iterations 10000 1.2 bcl.exe restraint:SimulateDistances -pdb 1GZM.pdb -skip_undefined_aas -simulate_distance_restraints -add_distance_uncertainty sl-cb_distances.histograms -output_file 1GZM.epr_cst_bcl -restraint_list restraint.ls	1.1 The OptimizeDataSetPairwise outputs the set of amino acid pairing in the protein sequence and a score for the set. Input: 1GZM.fasta #fasta file secondary_structure.pool #secondary structure pool native.ls #directory to native pdb mutate_start.table #specify number of restraint set Output: bcl.data #table of optimized set of EPR spin labeling a.a. pairs 1.2 SimulateDistances outputs the EPR distance restraints of the given sets of a.a. pairs. Input: 1GZM.pdb #native pdb sl-cb_distances.histograms #spin label to C_{beta} distance histogram used to simulate uncertainty restraint.ls #table of a.a. pairs from step 1.1. Output: 1GZM.epr_cst_bcl #EPR distance file
2. Simulate NMR restraints	Sparse NMR restraints were simulated as NOEs as 1 restraint per a.a. residue using BCL. When NMR restraints were used in folding, secondary structure information predicted from backbone chemical shift simulated by TALOS+ was also incorporated.	2.1 bcl.exe restraint:SimulateDistances -pdb 1GZM.pdb -simulate_nmr_distance_restraints -num_restraint_fraction 1 -output_file 1GZM.noe_star -aaclass AAComplete -add_distance_uncertainty noe_knowledge_based.histogram 2.2.1 chemical shift simulation were done using SPARTA+ server 2.2.2 bcl.exe protein:CreatSSEPool -prefix 1GZM -pool_min_sse_length 9 5 999 -ssmethod TALOS -factory SSPredThreshold 1.1 -join_separate	2.1 SimulateDistances outputs the NOE distance restraints by randomly picking 1 restraint per residue. Input: 1GZM.pdb #native pdb (protonated) noe_knowledge_based.histogram # histogram used to add NOE distance uncertainty. Output: 1GZM.noe_star #restraint in NMR-STAR 3.1 file formate 2.2.1 Input:1GZM.pdb 1GZM.fasta Output: 1GZMSS.tab 2.2.2 Input:1GZM.fasta 1GZMSS.tab from step 2.2.1 Output: 1GZM.TALOS.pool #SSE pool for BCL folding
3. Prepare input files for BCL fold	Other input files such as TM and secondary structure prediction pool were prepared using octopus and JUFO9D. Stage file were used to guide	3.1.1 Perform SSE prediction using Octopus and JUFO9D 3.1.2 bcl.exe protein:CreatSSEPool -prefix 1GZM -pool_min_sse_length 9 5 999 -ssmethod SSE_prediction_method -factory SSPredThreshold 3.2 Generate Stage file and corresponding scoring	3.1.1 Input: 1GZM.fasta Output: SSE prediction file such as *.jufo9d And *.octo_topo 2.1.2. Input: SSE prediction files with the same prefix '1GZM' Output: *.pool #SSE pool 3.2 See sample file format below.

	prediction with different experimental data in each BCL phases of assembly and refinement. Score weights files were adjusted to account for different restraint's presence.	function weight set file for each stage specified by stage file	
4. Running BCL::MP-fold with hybrid experimental restraints	Running BCL:MP-Fold using the membrane environment and experimental restraints. The command line utilizes the files prepared in previous steps. Sections in '-restraint_types' and '-body_restraint' points to the BCL modules that could be mixed and matched to use different types of restraints.	bcl.exe protein:Fold -stages_read stages.txt -mc_temperature_fraction 0.25 0.05 -native 1GZM.pdb -quality RMSD GDT_TS -superimpose RMSD -pool_separate 1 -pool_min_sse_lengths 5 999 -sspred JUFO9D OCTOPUS TALOS -sequence_data /directory_to_SSE_prediction_data/ 1GZM -pool_prefix 1GZM -pool 1GZM.SSPredHighest_TALOS.pool -membrane -tm_helices 1GZM.SSPredHighest_OCTOPUS.pool -restraint_types NOE DistanceEPR -restraint_prefix 1GZM -body_restraint 1GZM_body.pdb 2.5 2.5 5.0 5.0 -1.0 -print_body_assignment -score_density_connectivity 1GZM.mrc -nmodels * -prefix prefix_for_output_pdbs -protein_storage /directory_to_save_models/ -random_seed	Input: stage.txt #Stage file 1GZM.pdb #native pdb 1GZM.SSPredHighest_OCTOPUS.pool #TM helices to guide membrane positioning 1GZM_body.pdb #body restraint pdb to guide EM fold movers 1GZM.mrc #EM density file Output: prefix_for_output_pdbs.pdb
5. Running Rosetta loop modeling with hybrid experimental restraints	Rosetta takes the coarse-grained models produced in previous BCL stage and models their loops and atomic details. Several preparations steps needs to be taken to generate fragment files and restraints	5.1.1 fragment_picker.default.linuxgccrelease -database /rosetta/main/database/ -in::file::vall /rosetta/tools/fragment_tools/vall.jul19.2011.gz -frags::n_frags 200 -frags::frag_sizes 3 9 -frags::sigmoid_cs_A 2 -frags::sigmoid_cs_B 4 -out::file::frag_prefix.score -frags::describe_fragments 1GZM.fsc.score -frags::scoring::config scores.score.cfg -in:file:fasta 1GZM.fasta -in:file:talos_cs 1GZM.talos -frags::ss_pred predSS.tab talos -in::file::talos_phi_psi pred.tab 5.1.2 Prepare loop definition files 5.1.3 Prepare TM definition files /rosetta/tools/membrane_tools/octopus2span.pl 1GZM. octo_topo 5.1.4 Prepare Rosetta Format restraint files bcl.exe restraint:NmrFileConvert -pdb_file 1GZM.pdb -input_file 1GZM.noe_star star noe -output_file 1GZM_nmr.cst ROSETTA	5.1.1 Input: vall database installed in rosetta 1GZM.fasta 1GZM.talos #chemical shift predicted from Sparta+ Output: aa1GZM09_05.200_v1_3 #3mer fragments aa1GZM09_05.200_v1_9 #9mer fragments 5.1.2 example loop file format: LOOP 1 22 0 0.0 0 #1st and 2nd column indicates start and end of loop residues to be modeled 5.1.3 Input: 1GZM.octo_topo #from step 3.1.1 Output: 1GZM.span 5.1.4 Input: 1GZM.noe_star #step 2.1 Output: 1GZM_nmr.cst # restraints in Rosetta format 5.2 Input: 1GZM_bcl.pdb #bcl models that will be further refined by rosetta

		5.2 loopmodel.linuxgccrelease -database /rosetta/main/database/ -in:file:s 1GZM_bcl.pdb -loops:loop_file 1GZM.loops -in:file:native 1GZM.pdb -evaluation:rmsd NATIVE _FULL FULL -evaluation:gdtmm -loops::frag_sizes 9 3 1 -loops::frag_files aa1GZM09_05.200_v1_3 aa1GZM03_05.200_v1_3 none -loops::remodel quick_ccd -loops::intermedrelax no -loops::refine refine_ccd -loops::relax fastrelax -ex1 -ex2 -constraints:cst_file 1GZM_centroid_restraint.cst -constraints:cst_weight 5.0 -constraints:cst_fa_file 1GZM_full_atom.cst -constraints:cst_fa_weight 5 -constraints:epr_distance -constraints:viol -constraints:viol_level 101 -edensity::mapfile 1GZM.mrc -edensity::sliding_window 9 -edensity::mapreso 5.5 -edensity::grid_spacing 4.0 -whole_structure_allatom_wt 0.1 -score:weights membrane_highres_Menv_smooth.wts -membrane:no_interpolate_Mpair -membrane:Menv_penalties -in:file:spanfile 1GZM.span -out:pdb -out:output -out:file:scorefile score_file.sc -out:nstruct * -out:prefix /directory_to_save_models/prefix	1GZM.loops #step 5.1.2 1GZM.pdb #native pdb aa1GZM09_05.200_v1_3 #step 5.1.1 aa1GZM09_05.200_v1_9 #step 5.1.1 1GZM_centroid_restraint.cst #step 5.1.4 1GZM_full_atom.cst #step 5.1.4 1GZM.mrc #EM density file 1GZM.span #step 5.1.3 Output: score_file.sc #score file /directory_to_save_models/prefix.pdb #models produced by loopmodeling
6. Running Rosetta refinement with hybrid experimental restraints	A final refinement step was used to relax the predicted model to their energy minimum using the Rosetta relax application. Top scoring models from loop modeling were taken	6 relax.linuxgccrelease -database /rosetta/main/database/ -in:file:s 1GZM_loopmodeled.pdb -in:file:fullatom -in:file:native 1GZM.pdb -evaluation:rmsd NATIVE _FULL FULL -evaluation:gdtmm -relax:fast -relax:membrane -ex1 -ex2 -constraints:cst_fa_file 1GZM_full_atom.cst -constraints:cst_fa_weight 5 -constraints:epr_distance -constraints:viol -constraints:viol_level 101 -edensity::mapfile 1GZM.mrc -edensity::sliding_window 9 -edensity::mapreso 5.5 -edensity::grid_spacing 4.0 -whole_structure_allatom_wt 0.1	6 Input: 1GZM_loopmodeled.pdb #best scoring models after loop modeling 1GZM.pdb #native pdb aa1GZM09_05.200_v1_3 #step 5.1.1 aa1GZM09_05.200_v1_9 #step 5.1.1 1GZM_full_atom.cst #step 5.1.4 1GZM.mrc #EM density file 1GZM.span #step 5.1.3 Output: score_file.sc #score file /directory_to_save_models/prefix.pdb #models produced by relax

		-score:weights membrane_highres_Menv_smooth.wts -membrane:no_interpolate_Mpair -membrane:Menv_penalties -in:file:spanfile 1GZM.span -out:pdb -out:output -out:file:scorefile score_file.sc -out:nstruct * -out:prefix /directory_to_save_models/prefix	

Example BCL Stage file format:

The BCL stage files sets the parameters for BCL::MP-Fold for the specific restraints used in structure prediction. In the example file, SCORE_PROTOCOLS specifies the membrane environment and the used of distance restraint and EM density; MUTATE_PROTOCOLS specifies the particular mutate moves in the program that are tailored to efficiently sample protein conformations for the given environment.

```
NUMBER_CYCLES 1
STAGE Stage_assembly_1
        SCORE_PROTOCOLS Default Membrane Restraint EM
        SCORE_WEIGHTSET_FILE assembly_1.scoreweights
        MUTATE_PROTOCOLS Default Assembly Membrane Restraint EM
        NUMBER_ITERATIONS 2000 400
STAGE_END
STAGE Stage_assembly_2
        SCORE_PROTOCOLS Default Membrane Restraint EM
        SCORE_WEIGHTSET_FILE assembly_2.scoreweights
        MUTATE_PROTOCOLS Default Assembly Membrane Restraint EM
        NUMBER_ITERATIONS 2000 400
STAGE_END
STAGE Stage_assembly_3
        SCORE_PROTOCOLS Default Membrane Restraint EM
        SCORE_WEIGHTSET_FILE assembly_3.scoreweights
        MUTATE_PROTOCOLS Default Assembly Membrane Restraint EM
        NUMBER_ITERATIONS 2000 400
STAGE_END
STAGE Stage_assembly_4
        SCORE_PROTOCOLS Default Membrane Restraint EM
        SCORE_WEIGHTSET_FILE assembly_4.scoreweights
        MUTATE_PROTOCOLS Default Assembly Membrane Restraint EM
        NUMBER_ITERATIONS 2000 400
STAGE_END
STAGE Stage_refinement_1
        SCORE_PROTOCOLS Default Membrane Restraint EM
        SCORE_WEIGHTSET_FILE refinement_1.scoreweights
        MUTATE_PROTOCOLS Default Refinement Membrane Restraint EM
        NUMBER_ITERATIONS 2000 400
        PRINT_END_MODEL true
STAGE_END
```

Example BCL weightset file format:

A sample weightset file used in BCL stage is presented below. The first line corresponds to the various scoring terms used in the BCL scoring during each Monte-Carlo Metropolis minimization step. The second line corresponds to the weight for in each scoring term. Specifically, the body_agreement and connectivity score terms are EM specific; the noe_restraint, noe_penalty, ss_TALOS, ss_TALOS_ent score terms are NMR specific; the epr_distance, epr_lower_penalty, epr_upper_penalty are EPR specific. The scoreweight used in the current manuscript makes an equal contribution to the scores. It is possible for users to tweak the weightset based on the confidence of experimental data.

```
bcl::storage::Table<double> aaclash aadist aaneigh aaneigh_ent loop loop_closure_gradient rgyr
sseclash ssepack_fr strand_fr co_score ss_JUFO9D ss_JUFO9D_ent ss_OCTOPUS ss_OCTOPUS_ent
ssealign mp_helix_topology ss_OCTOPUS_env ss_JUFO9D_env body_agreement connectivity
noe_restraint noe_penalty ss_TALOS ss_TALOS_ent epr_distance epr_lower_penalty
epr_upper_penalty
weights                 375    0.35  50    50          10   50000              5   375    8
20      0.5     1          1            1           1              8       500
20              5              300            450       5           5          10
10           20.0        24             24
```

Protocol Capture of TSPO expression, purification and on-column detergent exchange

We documented the protocol for working with TSPO protein families (including mTSPO, hTSPO and rsTSPO). The protocol described here includes all recipes used for basic purification and a timeline for the workflow. The protocol can be used for on-column detergent exchange to allow efficient detergent or lipid-mimetics screening for structural studies. It is advised that since DPC is a detergent of relatively low critical micelle concentration (c.m.c. = 1.1mM), detergent exchange following DPC should have a lower c.m.c. to allow for higher efficiency. In the case were the desired detergent has higher c.m.c. compared to DPC, the detergent used for solubilizing cells should be changed to OG or Empigen. The 50kDa cut-off concentrator is used to limit detergent build-up since we find excess detergent in samples interferes with downstream assays.

Material:

- Plasmid: pBG100 (N-terminal 6XHis tag with 3C/preScission protease site)
  Kan Resistance [1000 fold dilution of 30mg/ml Kan]
- LB culture, pH adjusted to 7.0
- M9 medium, basic (2g KH2PO4, 8g Na2HPO4, 0.5g NaCl, 0.5g MgCl2*6H2O in 1L, pH 7.2)+ 10ml Vitamin + ZnCl2
- Buffer A: 50mM HEPES-Na, 150mM NaCl, pH 7.8
- Lysis Buffer: Lysis Buffer+ LDR mix (for 50ml of lysis buffer, add 100ul LDR, 500ul PMSF)
  75mM Tris-HCl. pH 7.7 (start with Tris-base and pH titration with HCl)
  300mM NaCl
  0.2mM EDTA
  (for 1Liter:9.086g Tris-base; 17.532g NaCl; 1mL 0.2M EDTA, adjust pH to 7.7.0 and fill with ddH2O to 1L)

LDR stock (Lysozyme/DNase/RNase Mix)
100mg/mL Lysozyme
10mg/mL DNase
10mg/mL RNase
10% glycerol
(for 10mL stock: 1g Lysozyme/0.1g DNase/0.1g RNase, dissolved in 10% glycerol to 10mL, filter sterilized, stored at -80C)

- DPC Buffer A: 1% DPC + Buffer A
- Wash Buffer DPC: Buffer A + 0.1% DPC + 20mM Imidazole
- Rinse Buffer: Buffer A + Detergent of choice + 10mM Imidazole
- Elution Buffer: Buffer A pH 7.0 + 0.05% DPC/detergent if choice + 250mM Imidazole
- NMR Buffer: 10mM Sodium phosphate buffer (pH6), 100mM NaCl + detergent of choice at its c.m.c concentration.

Protein production and purification:

Day 1, 4pm

Transfect the pBG100-mTSPO / pBG100-hTSPO / pBG100-rsTSPO into BL21 (DE3) competent cells, plating onto Kan agar plates

Day 2, 8am [pre-culture]

Pick colony from plate, inoculate transformed BL21 cells into 5ml sterile LB broth containing Kan.

Day 3, 8am [scale-up]

Transfer 5 ml of pre-culture into 1L LB/M9 with Kan. Incubation at 37 degrees until OD reaches 0.7. Add 1ml 1M IPTG to induce protein production.

Harvest cells after 5 hour at 37 degree by centrifugation at 7000rpm using Bechman centrifuge. Weigh and store the cell pellet in a 50ml falcon tube.

Day 4, Purification (start in afternoon)

Re-suspend the cell pellet with 20X (cell weight) Lysis Buffer, mix and tumble for 30 min

Sonicate the cell suspension in ice bath, 45%amplitude, 10 min, 5s/5s

Alternative method for inclusion body preparation: Spin the cell down using JA25.5 at 20000rpm at 4 degree (20min). Save the supernatant, (for ultracentrifugation, collecting membrane portion). Pallet washing step: Resuspend inclusion body with Lysis buffer, sonicate 45% amplitude, 5 min, 5s/5s, Spin the cell down using JA25.5 at 20000rpm at 4 degree (20min). Homogenize inclusion body with 20X (pellet weight) DPC Buffer A, tumbling overnight at 4 degrees.

For Membrane preparation: Using ultracentrifugation, 2hour, 4C, 30000rmp. Collect the insoluble, discard solution. Homogenize the entire fraction with 20X DPC buffer A, tumbling overnight at 4 degrees.

Day 5, Purification (start in morning)

Centrifuge the homogenized cell lysates from previous night using JA25.5 at 20000 rpm for (30min); collect supernatant.

Load 1X (cell weight) ml bead volume Ni-NTA to column and equilibrate using DPC Buffer A (10x bead volume).

Add equilibrated beads into supernatant, tumbling 1 hour at 4 degrees.

Add the Bead-supernatant mix to column.

Rinse with DPC Buffer A until the peak disappear.

Wash with 10X (bead volume) Wash Buffer DPC/until the peak disappear.

Optional rinse step for detergent exchange: 20X bead volume Rinse Buffer (If using DPC buffer, no need for rinse) with the detergent of choice, restrict the flow of the column to allow for a slower drop rate.

Elute using Elution buffer, collect fractions with 1.5ml Eppendorf tubes.

Check Purity:

SDS PAGE.

Buffer Exchange:

Combine the pure elution fractions.

Using Centricon Unit (50kDa cut-off), spin until elution fraction concentrate to 500ul.

Add 5ml NMR buffer, spin, repeat for 4 times.

Check final concentration.

Prepare for NMR/ CD/ other biophysical and biochemical assays.