Resource

Structure

Integrated Structural Biology for α-Helical Membrane Protein Structure Determination

Highlights

- A computational prediction pipeline was developed to utilize EM, NMR, EPR data
- At least 800 integral membrane protein families remain to be structurally elucidated
- Hybrid experimental data improve membrane protein structure prediction accuracy

Authors

Yan Xia, Axel W. Fischer, Pedro Teixeira, Brian Weiner, Jens Meiler

Correspondence

jens.meiler@vanderbilt.edu

In Brief

Xia et al. developed a computational structure prediction pipeline to utilize multiple experimental restraints to fold membrane proteins in BCL and Rosetta. The pipeline described herein could determine structures to an accuracy of 1.2 Å RMSD relative to the experimentally determined structural model.

Xia et al., 2018, Structure *26*, 657–666
April 3, 2018
https://doi.org/10.1016/j.str.2018.02.006

CellPress

Structure
Resource

Integrated Structural Biology for α-Helical Membrane Protein Structure Determination

Yan Xia,[1,2] Axel W. Fischer,[1,2] Pedro Teixeira,[2] Brian Weiner,[2] and Jens Meiler[1,2,3,*]
[1]Department of Chemistry, Vanderbilt University, Stevenson Center, Station B 351822, Room 7330, Nashville, TN 37232, USA
[2]Center for Structural Biology, Vanderbilt University, Nashville, TN 37232, USA
[3]Lead Contact
*Correspondence: jens.meiler@vanderbilt.edu
https://doi.org/10.1016/j.str.2018.02.006

SUMMARY

While great progress has been made, only 10% of the nearly 1,000 integral, α-helical, multi-span membrane protein families are represented by at least one experimentally determined structure in the PDB. Previously, we developed the algorithm BCL::MP-Fold, which samples the large conformational space of membrane proteins *de novo* by assembling predicted secondary structure elements guided by knowledge-based potentials. Here, we present a case study of rhodopsin fold determination by integrating sparse and/or low-resolution restraints from multiple experimental techniques including electron microscopy, electron paramagnetic resonance spectroscopy, and nuclear magnetic resonance spectroscopy. Simultaneous incorporation of orthogonal experimental restraints not only significantly improved the sampling accuracy but also allowed identification of the correct fold, which is demonstrated by a protein size-normalized transmembrane root-mean-square deviation as low as 1.2 Å. The protocol developed in this case study can be used for the determination of unknown membrane protein folds when limited experimental restraints are available.

INTRODUCTION

Integral Membrane Proteins Remain a Formidable Challenge for Structure Determination Methods

α-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 PDB are IMPs (Bill et al., 2011). Around 80% of all protein structures currently deposited in the PDB (Berman et al., 2002) 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 (Sanders and Sonnichsen, 2006; Wiener, 2004; Loll, 2003; White, 2004; Landau and Rosenbusch, 1996). Even if membrane protein (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 aids such as thermostabilizing mutations, helper proteins integrated into MP loop regions, or the non-native membrane mimic.

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 with 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 (Grant et al., 2004). 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 (Stevens et al., 2013), resulting in a non-representative fold representation in the PDB.

A Vast Sequence Space for IMPs Remains to be Represented with Experimentally Determined Structures

Oberai et al. (2006) have estimated that between 700 and 1,700 families of IMPs would cover 90% of the IMP sequence space. 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 (Khafizov et al., 2014). Accelerating fold determination for particularly important IMPs would therefore be of great relevance to biology. It is estimated that 3,305 human MPs exist in the *Homo sapiens* proteome (UniProt), with 90% of the sequences mapping to Pfam families. 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 (Kloppmann et al., 2012). 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 (Pieper et al., 2013).

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 diffract 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 crosslinking coupled with mass spectrometry (XLMS). These 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 these information gaps (Lindert et al., 2012; Alexander et al., 2014; Barth et al., 2009). However, *de novo* prediction of an IMP's fold from its primary sequence remains a challenging problem (Koehler Leman et al., 2015). 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.

BCL::MP-Fold and Rosetta Predict Membrane Protein Structural Ensembles Using Experimental Data

BCL::MP-Fold (Weiner et al., 2013) was developed for *de novo* protein structure prediction. We demonstrated in previous studies that BCL::MP-Fold is able to efficiently sample the fold of large IMPs (Dimaio et al., 2009). 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 (Karakas et al., 2012). After each Monte Carlo step, the free energy of the intermediate model is approximated using knowledge-based scoring functions specifically developed for IMPs (Weiner et al., 2013). 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 (Lindert et al., 2009, 2012), NMR (Weiner et al., 2014), EPR (Fischer et al., 2015), and XLMS (Hofmann et al., 2015). 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 (Yarov-Yarovoy et al., 2006; Mandell et al., 2009; Leaver-Fay et al., 2011). Similar to BCL::MP-Fold, incorporation of individual experimental restraints into Rosetta been successful for EM (Dimaio et al., 2009), NMR (Bowers et al., 2000; Schmitz et al., 2012) and EPR (Alexander et al., 2008) 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 transmembrane spans. The number of IMP sequences is large as well, although most IMP families have relatively few transmembrane 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.

RESULTS

This 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 effects of combining hybrid experimental data on protein fold prediction.

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 transmembrane 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 transmembrane spans including the order of the TMH insertion and extra-/intracellular 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 insertions in a particular arrangement times 2 from whether its N-terminal being inside/outside divided by the symmetry operators under a particular arrangement (Figure 1A). An example of a five-TMH protein's fold defined by our criteria is demonstrated here (Figure 1B). 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 transmembrane spans (Figures 1C and 1D). For example, an IMP with nine TMHs can adopt about 1 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 transmembrane spans distant in the fold. Next we look at the sequence space for the number of IMP families that exist.

Sequence Space for α-Helical MPs

An analysis of protein sequence families documented by Pfam (pfam27.0) and crosschecked by UniProt annotation (Figure 2A)

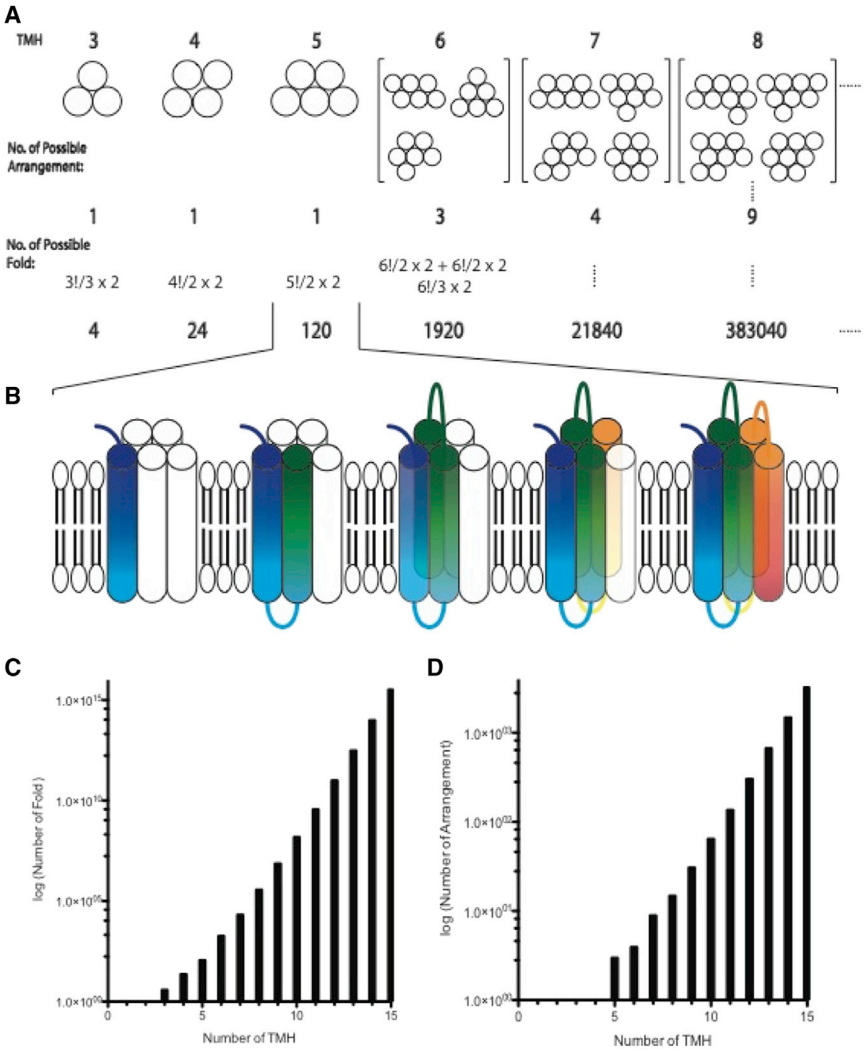

Figure 1. The Theoretical Fold Space for IMPs

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. The non-redundant arrangement and fold is computed for each 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.

revealed 895 IMP families consisting of α-helical TMHs. For 108 of these IMP families, at least one structure has been determined 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 fewer than 7 TMHs, while few have more than 12 TMHs.

The IMP sequence space is large, although 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, although distinguishing accurate from inaccurate folds would remain an obstacle for *de novo* methods without assistance from experimental data (Karakas et al., 2012).

BCL::MP-Fold Assembly of TMHs Using Hybrid Experimental Data

BCL::MP-Fold was used to simultaneously incorporate experimental data from multiple sources such as EM density maps, EPR distance restraints from double electron-electron resonance (DEER) experiments, nuclear Overhauser effect (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 datasets (NMR_EPR, EM_NMR, or EM_EPR—double-hybrid) or integrating all three experimental datasets (EM_EPR_NMR—triple-hybrid). Predicted structural models were evaluated by computing the root-mean-square deviation value $RMSD_{100}$ (Carugo and Pongor, 2001) of the sampled models relative to the experimentally determined structure and their respective energy scores in the BCL. Note that the $RMSD_{100}$ calculation was performed over aligned regions of the structure: since BCL::MP-Fold assembles only the TMHs in the three-dimensional space, the $RMSD_{100}$ values presented here only relate to the TMHs of the IMP.

In each case, BCL::MP-Fold was able to sample the native-like fold (Table 1). However, although the *de novo* sampled structures were structurally similar to the experimentally determined structure ($RMSD_{100}$ = 4.5 Å) (PDB: 1GZM; Li et al., 2004), 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 (Table 1; Figures 3A and 3C) improved sampling and scoring slightly. Notably, NMR data improved model discrimination and sampling density around the experimentally determined structure (Figure S1B), where the best scoring models achieved an $RMSD_{100}$ of 5.1 Å. EM data helped in

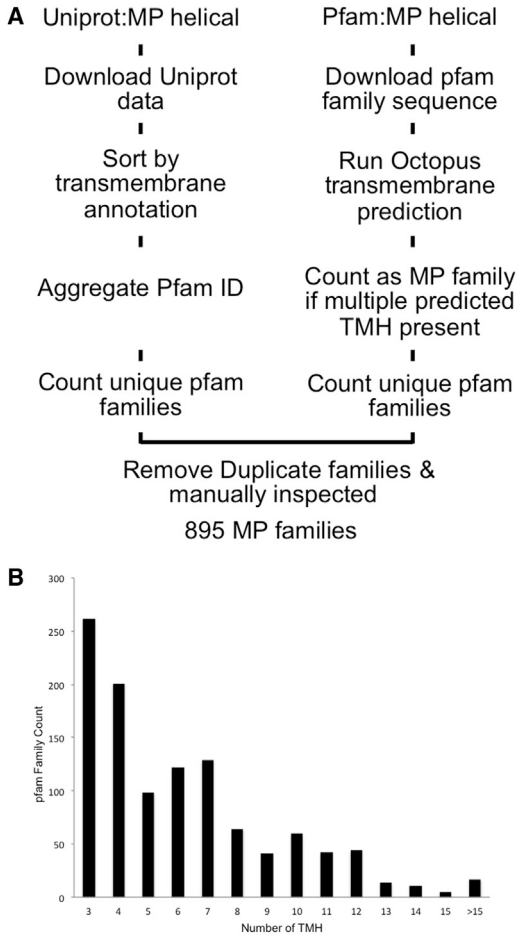

Figure 2. The Sequence Space for IMPs

A survey of the protein sequence database was performed to map the unique MP 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 constructed.

positioning the TMHs in the EM density rods, resulting in more accurate sampling of the native fold ($RMSD_{100}$ = 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 ($RMSD_{100}$ = 3.5 Å). Again, the scoring function was unable to identify these improved models. Both, NMR- and EM-assisted models included low $RMSD_{100}$ 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 datasets to test whether 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 NMR_EPR, EM_NMR sets improved sampling density near the native-like folds to below 3 Å $RMSD_{100}$ (Table 1 and Figure 3B). The best models ranked by score are among the 5% most accurate models. EM_EPR prediction also improved the $RMSD_{100}$ of the most accurate folds sampled to 1.9 Å but had an accurate model ranked only second by score ($RMSD_{100}$ = 2.2 Å). The improvement in sampling density was less compared with tests that included NMR restraints (Figure 3B), 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 $RMSD_{100}$ of 1.4 Å and an average $RMSD_{100}$ 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 Figures 3B and 3D. Models ranked by score in EM_EPR_NMR also had the best agreement with the crystal structure, which is demonstrated by an $RMSD_{100}$ of 2.5 Å and an average $RMSD_{100}$ 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.

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 do not 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 datasets added additional layers of accuracy in refinement (Figures 4A and 4B; Table 1).

The energy landscape of the rhodopsin fold, visualized by plotting the $RMSD_{100}$ 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 (Figures 4A and 4B). The use of hybrid experimental data improved the prediction accuracy in the core of the protein. The $RMSD_{100}$_TM, which quantifies the $RMSD_{100}$ of the transmembrane region, for the top 5% sampled models using hybrid restraint data improved by at least 1 Å over that of the top 5% sampled models using single restraint (Table 1, $RMSD_{100}$_TM).

With EM data alone, which are 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 $RMSD_{100}$_TM. Although input models contained models with incorrect folds that extend to an $RMSD_{100}$_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 were incorporated.

Overall, whereas the inclusion of the EM density map resulted in an improvement of the total $RMSD_{100}$ of the best scoring

Table 1. Structure Prediction Results from Hybrid Experimental Data

Restraints BCL			RMSD$_{100}$ Best	RMSD$_{100}$ Top 5%	Score Best	Score Top 5%
No restraints			4.5	5.4	9.4	9.2
NNMR			3.1	3.8	5.1	6.1
	EEPR		3.5	4.5	8.5	7.5
		EEM	2.9	4.8	8.3	8.1
NNMR	EEPR		2.7	3.7	3.6	4.0
NNMR		EEM	3.0	4.0	4.0	5.5
	EEPR	EEM	1.9	4.1	8.4	7.4
NNMR	EEPR	EEM	1.4	2.9	2.5	3.9

Restraints Rosetta			RMSD$_{100}$ Best	RMSD$_{100}$ Top 5%	RMSD$_{100}$_TM Best	RMSD$_{100}$_TM Top 5%	Score Best	Score Top 5%	Score (RMSD$_{100}$_TM) Best	Score (RMSD$_{100}$_TM) 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

The RMSD$_{100}$ metric is used to quantify model quality. RMSD$_{100}$: RMSD$_{100}$ of models ranked by RMSD$_{100}$ to native model; Score: RMSD$_{100}$ of models ranked by either BCL score or Rosetta score; Top 5%: averaged RMSD$_{100}$ over top 5% models ranked by the respective metric. RMSD$_{100}$_TM: RMSD$_{100}$ value calculated from TMH regions of the models ranked by total RMSD$_{100}$. Score (TM_RMSD): RMSD$_{100}$ value calculated from TMH regions of the models ranked by Rosetta score.

model to below 6 Å, distance type restraints (NMR, NMR_EPR) were unable to refine the best scoring models to below 6 Å (Figure 4A and Table 1). The EM data restrain 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 RMSD$_{100}$_TM to below 2 Å in the best scoring models. To our surprise, the models that were refined using NMR data also had less favorable scores. To break down the cause of the worse scores, we performed re-scoring using only the built-in Rosetta scoring function to investigate the behavior of the membrane scoring function (see Discussion).

Refinement using EM_NMR_EPR resulted in the most accurate models (RMSD$_{100}$ = 2.9 Å). The total RMSD$_{100}$ approached a limit of 4 Å while the RMSD$_{100}$ of the core helical transmembrane region was close to 1 Å. In our prediction pipeline, the retinal molecule is not modeled. Accurate modeling of the extracellular 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 4C). Even when input models were idealized helices, the kinks in helices 6 and 7 could be observed after the refinement. The side-chain conformers from the model could be modeled to high accuracy (Figure 4D).

DISCUSSION

The Fold and Sequence Space of α-Helical MPs Is Vast, but Computational Modeling Provides Promising Results

Structural representatives remain to be determined for about 700 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 100,000 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.

Recent developments in co-evolution contact prediction offer alternative information for IMP structure determination (Tang et al., 2015; Marks et al., 2011; Kamisetty et al., 2013; Morcos

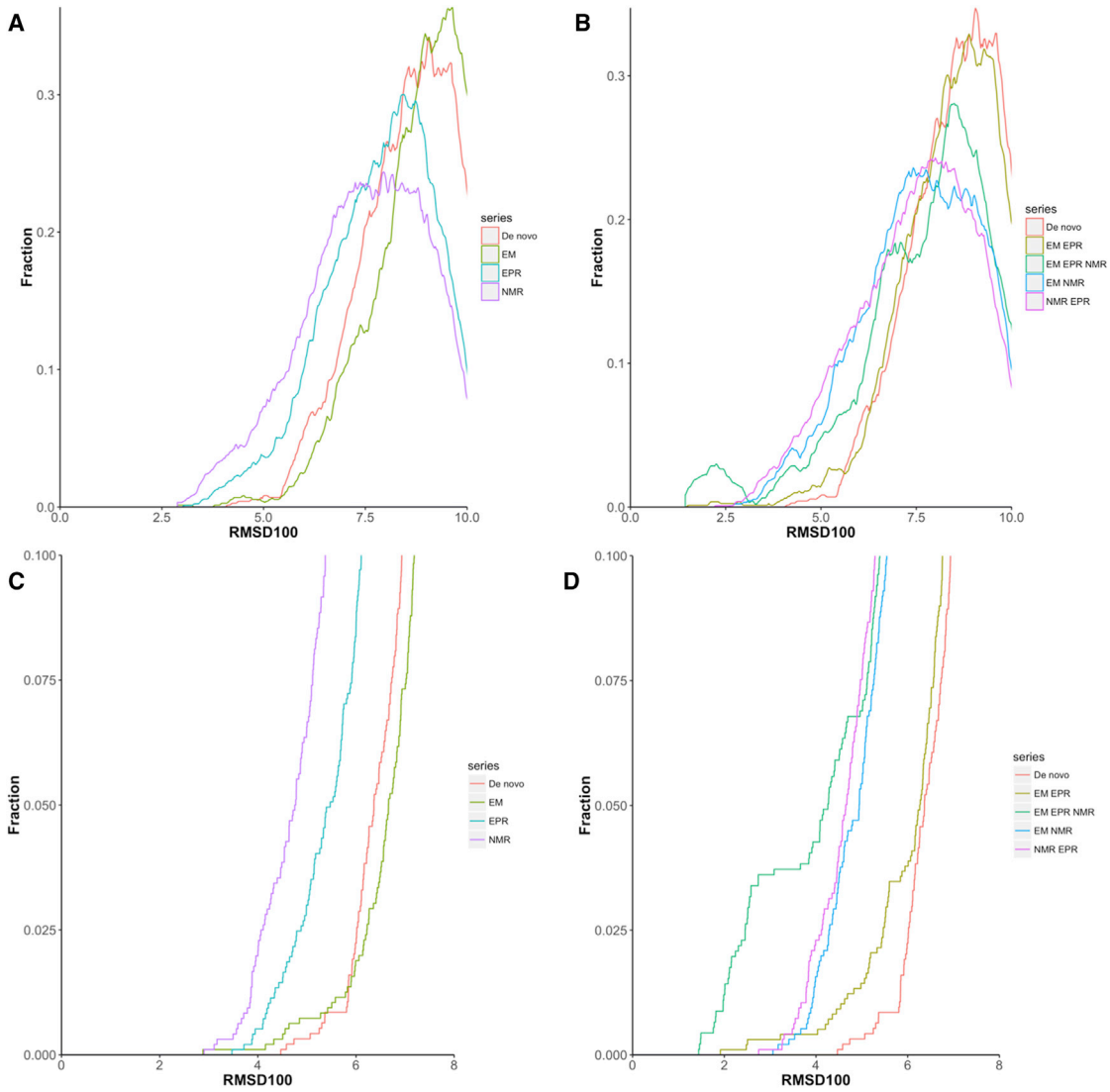

Figure 3. Prediction Accuracy of Low-Resolution SSE Assembly Using Hybrid Experimental Data

Density distributions of structure prediction accuracies using single experimental dataset (A and C) and hybrid experimental dataset (B and D). The fraction of models versus the predicted models' $RMSD_{100}$ 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 8 Å 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 8 Å RMSD; the y axis is cut off at 0.1.

et al., 2011). 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 (Ovchinnikov et al., 2015). Combining sparse experimental restraints from NMR with evolutionary constraints allows for accurate prediction of soluble protein structure in many cases (Tang et al., 2015). Evolutionary coupling-NMR, developed by Tang et al., incorporates evolutionary contacts during and after the NMR data interpretation and NOE assignment. In a more recent publication by Ovchinnikov et al. (2017), 206 unknown IMP folds were predicted using integrated metagenome data and co-evolutionary analysis. The sizes of the proteins involved in this study are below 300 amino acids. With the expansion of genomic databases coupled with computational prediction algorithms, evolutionary constraints provide viable orthogonal structural information to guide protein structure prediction.

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. (Altenbach et al., 2008). are centered

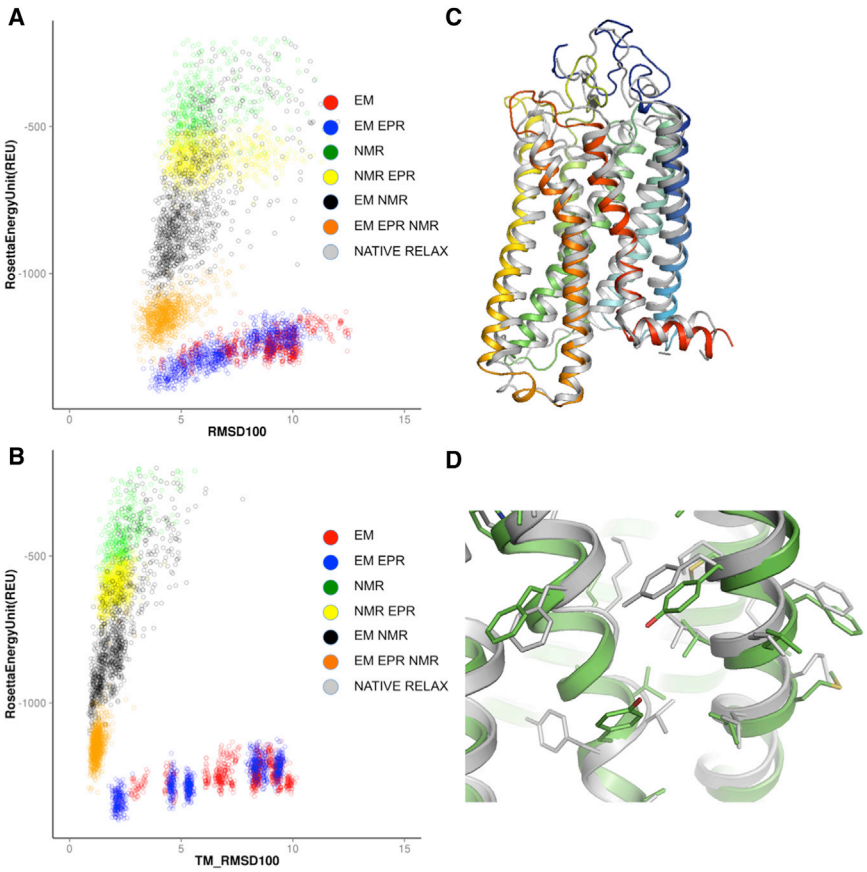

Figure 4. Prediction Accuracy of Rosetta Refinement Using Hybrid Experimental Data

The scatterplot 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 to the experimentally determined 1GZM structure (A), or TMH RMSD to the TMH region of 1GZM (B). The scatterplot 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 gray (C). The side-chain rotational conformer of the predicted structure matched the experimental structure with high accuracy (D).

on the intracellular 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 $RMSD_{100}$ improvement from 5 Å to 4.5 Å (Fischer et al., 2015). In our study using simulated DEER restraints, the $RMSD_{100}$ of the most accurate model arrived at 3.5 Å. The additional improvement by 1 Å 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 (Gautier and Nietlispach, 2012); 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 interhelical constraints from low-resolution electron diffraction experiments are used (PDB: 1JFP) (Yeagle et al., 2001). 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 an MP. 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 and the upper bound of the distances is below 6 Å.

Experimental Data Overcome 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. Furthermore, 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 with models predicted from NMR data. To confirm that such discrepancies were not caused by the experimental data, we re-evaluated the models 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 S2), 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 in which 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 fewer unfavorable energy contributions from fa_rep (Lennard-Jones repulsive energy between atoms in different residues [Rohl et al., 2004; Leaver-Fay et al., 2013]) and fa_dun (Internal energy of side-chain rotamers as derived from Dunbrack's statistics [Shapovalov and Dunbrack, 2011]). At the

same time, models predicted from triple-hybrid experimental data exhibited lower Rosetta scores compared with 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, which forces side-chain contacts that are otherwise hard to sample due to the simplified representation of the IMPs.

Structure Determination from Limited Experimental Data Will Be Important to Determine Alternative Conformational States of Integral Membrane Proteins

Many IMPs function through shifting ensembles of numerous conformations. 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 thermostabilizing 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.

The best examples are perhaps 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 (Manglik et al., 2015; Kaiser et al., 2015; Kim et al., 2013; Dror et al., 2015; Shukla et al., 2014; Manglik and Kobilka, 2014; Alexander et al., 2014; Nygaard et al., 2013; Altenbach et al., 2008; El Moustaine et al., 2012; Xue et al., 2015). Solution NMR experiments using ^{19}F probes conjugated to specific amino acids (Kim et al., 2013) and ^{13}C methionine side-chain labels (Nygaard et al., 2013) capture β_2-adrenergic receptor in distinct population of structural states aside from observed crystal structures. The shift of the conformational equilibrium of GPCRs during signal transduction is accompanied by major structural changes, as revealed by SDSL-EPR in β_2-adrenergic receptor (Manglik et al., 2015) and rhodopsin (Altenbach et al., 2008; Alexander et al., 2014). The architecture of the molecular assembly of GPCR and heterotrimeric G protein and arrestin in a dynamic state could also be visualized using XLMS and EM (Shukla et al., 2014). While the homodimeric metabotropic glutamate receptor 1 (mGluR1) is crystallized in an inactive state dimer (Wu et al., 2014), a later study using fluorescence resonance transfer and crosslinking displayed a deviating dimer interface in the biological sample. Conformational rearrangement in an oligomerized state has also been observed, whereby the activated receptor dimers undergo conformational change in individual subunits as well as the dimer interface (El Moustaine et al., 2012; Xue et al., 2015). Molecular dynamics simulations and Monte Carlo simulations utilizing these experimental data were also successful in generating structural models that match the observed conformational states (Shukla et al., 2014; Alexander et al., 2014; Nygaard et al., 2013).

Transporters also exist in multiple conformations with respect to their transport cycle. The small multidrug resistance transporter EmrE utilizes 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 (Chen et al., 2007), 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 TMHs 1–3 in response to the change of the protonation state (Dastvan et al., 2016), 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.

NMR, EPR, and cryoelectron microscopy data, albeit samples are at high temperature or flash-frozen, are expected to reflect native-like conditions when different conformations of the protein exist in 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 timescale 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 (Zhou et al., 2015).

BCL::MP-Fold and Rosetta algorithms can also be used to optimize a given starting structure. Thus, from any given starting structural model or crystal structure we can derive a model for an alternative state of an MP by fitting it to sparse experimental data observed for this state (Dastvan et al., 2016). By combining experimental data from different sources and excluding the likely outlier populations due to artifacts in the disruptive experimental methods, one can model the entire conformational ensemble of the IMP. Such ensemble models along with the statistical inference could be used to provide further molecular mechanistic insight into protein function, and possibly guide small-molecule modulator development against intermediate states.

Conclusion

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.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- CONTACT FOR REAGENTS AND RESOURCE SHARING

- METHOD DETAILS
  - Enumerating the Theoretical α-helical Integral Membrane Protein Fold Space
  - Enumerating the Sequence Space of α-helical Integral Membrane Protein
  - Integral Membrane Protein Structure Prediction Using Combined Experimental Restraints and BCL::MP-fold
  - Integral Membrane Protein Structure Refinement Using Combined Experimental Restraints and Rosetta

SUPPLEMENTAL INFORMATION

Supplemental Information includes two figures, two tables, and Supplemental text and can be found with this article online at https://doi.org/10.1016/j.str.2018.02.006.

ACKNOWLEDGMENT

The authors want to thank Rocco Moretti and Amanda Duran for computational advice regarding the Rosetta software suite. Work on this project in the Meiler laboratory is supported through NIH (R01 GM080403, R01 GM099842).

AUTHOR CONTRIBUTIONS

Conceptualization, Y.X. and J.M.; Methodology, Y.X. and J.M.; Software, Y.X., A.W.F., B.W., and P.T.; Investigation, Y.X., P.T., and A.W.F.; Formal Analysis, Y.X. and A.W.F.; Writing – Original Draft, Y.X., A.W.F., and J.M.; Writing – Review & Editing, Y.X., A.W.F., B.W., and J.M.; Funding Acquisition, J.M.; Resources, A.W.F. and B.W.; Supervision, J.M.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: January 25, 2017
Revised: June 14, 2017
Accepted: February 5, 2018
Published: March 8, 2018

REFERENCES

Alexander, N., Bortolus, M., Al-Mestarihi, A., Mchaourab, H., and Meiler, J. (2008). De novo high-resolution protein structure determination from sparse spin-labeling EPR data. Structure *16*, 181–195.

Alexander, N.S., Preininger, A.M., Kaya, A.I., Stein, R.A., Hamm, H.E., and Meiler, J. (2014). Energetic analysis of the rhodopsin-G-protein complex links the alpha5 helix to GDP release. Nat. Struct. Mol. Biol. *21*, 56–63.

Altenbach, C., Kusnetzow, A.K., Ernst, O.P., Hofmann, K.P., and Hubbell, W.L. (2008). High-resolution distance mapping in rhodopsin reveals the pattern of helix movement due to activation. Proc. Natl. Acad. Sci. USA *105*, 7439–7444.

Barth, P., Wallner, B., and Baker, D. (2009). Prediction of membrane protein structures with complex topologies using limited constraints. Proc. Natl. Acad. Sci. USA *106*, 1409–1414.

Berman, H.M., Battistuz, T., Bhat, T.N., Bluhm, W.F., Bourne, P.E., Burkhardt, K., Feng, Z., Gilliland, G.L., Iype, L., Jain, S., et al. (2002). The protein data bank. Acta Crystallogr. D Biol. Crystallogr. *58*, 899–907.

Bill, R.M., Henderson, P.J., Iwata, S., Kunji, E.R., Michel, H., Neutze, R., Newstead, S., Poolman, B., Tate, C.G., and Vogel, H. (2011). Overcoming barriers to membrane protein structure determination. Nat. Biotechnol. *29*, 335–340.

Bowers, P.M., Strauss, C.E., and Baker, D. (2000). De novo protein structure determination using sparse NMR data. J. Biomol. NMR *18*, 311–318.

Canutescu, A.A., and Dunbrack, R.L., Jr. (2003). Cyclic coordinate descent: a robotics algorithm for protein loop closure. Protein Sci. *12*, 963–972.

Carugo, O., and Pongor, S. (2001). A normalized root-mean-square distance for comparing protein three-dimensional structures. Protein Sci. *10*, 1470–1473.

Chen, Y.J., Pornillos, O., Lieu, S., Ma, C., Chen, A.P., and Chang, G. (2007). X-ray structure of EmrE supports dual topology model. Proc. Natl. Acad. Sci. USA *104*, 18999–19004.

Dastvan, R., Fischer, A.W., Mishra, S., Meiler, J., and Mchaourab, H.S. (2016). Protonation-dependent conformational dynamics of the multidrug transporter EmrE. Proc. Natl. Acad. Sci. USA *113*, 1220–1225.

Dimaio, F., Tyka, M., Baker, M., Chiu, W., and Baker, D. (2009). Refinement of protein structures into low-resolution density maps using Rosetta. J. Mol. Biol. *392*, 181–190.

Dror, R.O., Mildorf, T.J., Hilger, D., Manglik, A., Borhani, D.W., Arlow, D.H., Philippsen, A., Villanueva, N., Yang, Z., Lerch, M.T., et al. (2015). SIGNAL TRANSDUCTION. Structural basis for nucleotide exchange in heterotrimeric G proteins. Science *348*, 1361–1365.

El Moustaine, D., Granier, S., Doumazane, E., Scholler, P., Rahmeh, R., Bron, P., Mouillac, B., Baneres, J.L., Rondard, P., and Pin, J.P. (2012). Distinct roles of metabotropic glutamate receptor dimerization in agonist activation and G-protein coupling. Proc. Natl. Acad. Sci. USA *109*, 16342–16347.

Fischer, A.W., Alexander, N.S., Woetzel, N., Karakas, M., Weiner, B.E., and Meiler, J. (2015). BCL::MP-fold: membrane protein structure prediction guided by EPR restraints. Proteins *83*, 1947–1962.

Gautier, A., and Nietlispach, D. (2012). 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. *914*, 25–45.

Grant, A., Lee, D., and Orengo, C. (2004). Progress towards mapping the universe of protein folds. Genome Biol. *5*, 107.

Hirst, S.J., Alexander, N., Mchaourab, H.S., and Meiler, J. (2011). RosettaEPR: an integrated tool for protein structure determination from sparse EPR data. J. Struct. Biol. *173*, 506–514.

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

Jayasinghe, S., Hristova, K., and White, S.H. (2001). MPtopo: a database of membrane protein topology. Protein Sci. *10*, 455–458.

Kaiser, A., Müller, P., Zellmann, T., Scheidt, H.A., Thomas, L., Bosse, M., Meier, R., Meiler, J., Huster, D., Beck-Sickinger, A.G., and Schmidt, P. (2015). Unwinding of the C-Terminal residues of neuropeptide Y is critical for Y2 receptor binding and activation. Angew. Chem. Int. Ed. *54*, 7446–7449.

Kamisetty, H., Ovchinnikov, S., and Baker, D. (2013). Assessing the utility of coevolution-based residue-residue contact predictions in a sequence- and structure-rich era. Proc. Natl. Acad. Sci. USA *110*, 15674–15679.

Karakas, M., Woetzel, N., Staritzbichler, R., Alexander, N., Weiner, B.E., and Meiler, J. (2012). BCL::Fold—de novo prediction of complex and large protein topologies by assembly of secondary structure elements. PLoS One *7*, e49240.

Khafizov, K., Madrid-Aliste, C., Almo, S.C., and Fiser, A. (2014). Trends in structural coverage of the protein universe and the impact of the Protein Structure Initiative. Proc. Natl. Acad. Sci. USA *111*, 3733–3738.

Kim, T.H., Chung, K.Y., Manglik, A., Hansen, A.L., Dror, R.O., Mildorf, T.J., Shaw, D.E., Kobilka, B.K., and Prosser, R.S. (2013). The role of ligands on the equilibria between functional states of a G protein-coupled receptor. J. Am. Chem. Soc. *135*, 9465–9474.

Kloppmann, E., Punta, M., and Rost, B. (2012). Structural genomics plucks high-hanging membrane proteins. Curr. Opin. Struct. Biol. *22*, 326–332.

Koehler, J., Woetzel, N., Staritzbichler, R., Sanders, C.R., and Meiler, J. (2009). A unified hydrophobicity scale for multispan membrane proteins. Proteins *76*, 13–29.

Koehler Leman, J., Ulmschneider, M.B., and Gray, J.J. (2015). Computational modeling of membrane proteins. Proteins *83*, 1–24.

Landau, E.M., and Rosenbusch, J.P. (1996). Lipidic cubic phases: a novel concept for the crystallization of membrane proteins. Proc. Natl. Acad. Sci. USA *93*, 14532–14535.

Leaver-Fay, A., O'Meara, M.J., Tyka, M., Jacak, R., Song, Y., Kellogg, E.H., Thompson, J., Davis, I.W., Pache, R.A., Lyskov, S., et al. (2013). Scientific benchmarks for guiding macromolecular energy function improvement. Methods Enzymol. *523*, 109–143.

Leaver-Fay, A., Tyka, M., Lewis, S.M., Lange, O.F., Thompson, J., Jacak, R., Kaufman, K., Renfrew, P.D., Smith, C.A., Sheffler, W., et al. (2011). ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. Methods Enzymol. *487*, 545–574.

Li, J., Edwards, P.C., Burghammer, M., Villa, C., and Schertler, G.F. (2004). Structure of bovine rhodopsin in a trigonal crystal form. J. Mol. Biol. *343*, 1409–1438.

Lindert, S., Alexander, N., Wotzel, N., Karakas, M., Stewart, P.L., and Meiler, J. (2012). EM-fold: de novo atomic-detail protein structure determination from medium-resolution density maps. Structure *20*, 464–478.

Lindert, S., Staritzbichler, R., Wotzel, N., Karakas, M., Stewart, P.L., and Meiler, J. (2009). EM-fold: de novo folding of alpha-helical proteins guided by intermediate-resolution electron microscopy density maps. Structure *17*, 990–1003.

Loll, P.J. (2003). Membrane protein structural biology: the high throughput challenge. J. Struct. Biol. *142*, 144–153.

Mandell, D.J., Coutsias, E.A., and Kortemme, T. (2009). Sub-angstrom accuracy in protein loop reconstruction by robotics-inspired conformational sampling. Nat. Methods *6*, 551–552.

Manglik, A., Kim, T.H., Masureel, M., Altenbach, C., Yang, Z., Hilger, D., Lerch, M.T., Kobilka, T.S., Thian, F.S., Hubbell, W.L., et al. (2015). Structural insights into the dynamic process of beta$_2$-adrenergic receptor signaling. Cell *161*, 1101–1111.

Manglik, A., and Kobilka, B. (2014). The role of protein dynamics in GPCR function: insights from the beta2AR and rhodopsin. Curr. Opin. Cell Biol. *27*, 136–143.

Marks, D.S., Colwell, L.J., Sheridan, R., Hopf, T.A., Pagnani, A., Zecchina, R., and Sander, C. (2011). Protein 3D structure computed from evolutionary sequence variation. PLoS One *6*, e28766.

Morcos, F., Pagnani, A., Lunt, B., Bertolino, A., Marks, D.S., Sander, C., Zecchina, R., Onuchic, J.N., Hwa, T., and Weigt, M. (2011). Direct-coupling analysis of residue coevolution captures native contacts across many protein families. Proc. Natl. Acad. Sci. USA *108*, E1293–E1301.

Nygaard, R., Zou, Y., Dror, R.O., Mildorf, T.J., Arlow, D.H., Manglik, A., Pan, A.C., Liu, C.W., Fung, J.J., Bokoch, M.P., et al. (2013). The dynamic process of beta(2)-adrenergic receptor activation. Cell *152*, 532–542.

Oberai, A., Ihm, Y., Kim, S., and Bowie, J.U. (2006). A limited universe of membrane protein families and folds. Protein Sci. *15*, 1723–1734.

Ovchinnikov, S., Kinch, L., Park, H., Liao, Y., Pei, J., Kim, D.E., Kamisetty, H., Grishin, N.V., and Baker, D. (2015). Large-scale determination of previously unsolved protein structures using evolutionary information. Elife *4*, e09248.

Ovchinnikov, S., Park, H., Varghese, N., Huang, P.S., Pavlopoulos, G.A., Kim, D.E., Kamisetty, H., Kyrpides, N.C., and Baker, D. (2017). Protein structure determination using metagenome sequence data. Science *355*, 294–298.

Pieper, U., Schlessinger, A., Kloppmann, E., Chang, G.A., Chou, J.J., Dumont, M.E., Fox, B.G., Fromme, P., Hendrickson, W.A., Malkowski, M.G., et al. (2013). Coordinating the impact of structural genomics on the human alpha-helical transmembrane proteome. Nat. Struct. Mol. Biol. *20*, 135–138.

Rohl, C.A., Strauss, C.E., Misura, K.M., and Baker, D. (2004). Protein structure prediction using Rosetta. Methods Enzymol. *383*, 66–93.

Ruprecht, J.J., Mielke, T., Vogel, R., Villa, C., and Schertler, G.F. (2004). Electron crystallography reveals the structure of metarhodopsin I. EMBO J. *23*, 3609–3620.

Sanders, C.R., and Sonnichsen, F. (2006). Solution NMR of membrane proteins: practice and challenges. Magn. Reson. Chem. *44* (Spec No), S24–S40.

Schmitz, C., Vernon, R., Otting, G., Baker, D., and Huber, T. (2012). Protein structure determination from pseudocontact shifts using ROSETTA. J. Mol. Biol. *416*, 668–677.

Schrödinger, L.L.C. (2015). The PyMOL Molecular Graphics System, Version 1.8 (Schrödinger).

Shapovalov, M.V., and Dunbrack, R.L., Jr. (2011). A smoothed backbone-dependent rotamer library for proteins derived from adaptive kernel density estimates and regressions. Structure *19*, 844–858.

Shen, Y., and Bax, A. (2010). SPARTA+: a modest improvement in empirical NMR chemical shift prediction by means of an artificial neural network. J. Biomol. NMR *48*, 13–22.

Shukla, A.K., Westfield, G.H., Xiao, K., Reis, R.I., Huang, L.Y., Tripathi-Shukla, P., Qian, J., Li, S., Blanc, A., Oleskie, A.N., et al. (2014). Visualization of arrestin recruitment by a G-protein-coupled receptor. Nature *512*, 218–222.

Stevens, R.C., Cherezov, V., Katritch, V., Abagyan, R., Kuhn, P., Rosen, H., and Wuthrich, K. (2013). The GPCR Network: a large-scale collaboration to determine human GPCR structure and function. Nat. Rev. Drug Discov. *12*, 25–34.

Tang, Y., Huang, Y.J., Hopf, T.A., Sander, C., Marks, D.S., and Montelione, G.T. (2015). Protein structure determination by combining sparse NMR data with evolutionary couplings. Nat. Methods *12*, 751–754.

Vernon, R., Shen, Y., Baker, D., and Lange, O.F. (2013). Improved chemical shift based fragment selection for CS-Rosetta using Rosetta3 fragment picker. J. Biomol. NMR *57*, 117–127.

Viklund, H., Bernsel, A., Skwark, M., and Elofsson, A. (2008). SPOCTOPUS: a combined predictor of signal peptides and membrane protein topology. Bioinformatics *24*, 2928–2929.

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

Weiner, B.E., Alexander, N., Akin, L.R., Woetzel, N., Karakas, M., and Meiler, J. (2014). BCL::Fold—protein topology determination from limited NMR restraints. Proteins *82*, 587–595.

White, S.H. (2004). The progress of membrane protein structure determination. Protein Sci. *13*, 1948–1949.

Wiener, M.C. (2004). A pedestrian guide to membrane protein crystallization. Methods *34*, 364–372.

Wu, H., Wang, C., Gregory, K.J., Han, G.W., Cho, H.P., Xia, Y., Niswender, C.M., Katritch, V., Meiler, J., Cherezov, V., et al. (2014). Structure of a class C GPCR metabotropic glutamate receptor 1 bound to an allosteric modulator. Science *344*, 58–64.

Xue, L., Rovira, X., Scholler, P., Zhao, H., Liu, J., Pin, J.P., and Rondard, P. (2015). Major ligand-induced rearrangement of the heptahelical domain interface in a GPCR dimer. Nat. Chem. Biol. *11*, 134–140.

Yarov-Yarovoy, V., Schonbrun, J., and Baker, D. (2006). Multipass membrane protein structure prediction using Rosetta. Proteins *62*, 1010–1025.

Yeagle, P.L., Choi, G., and Albert, A.D. (2001). Studies on the structure of the G-protein-coupled receptor rhodopsin including the putative G-protein binding site in unactivated and activated forms. Biochemistry *40*, 11932–11937.

Zhou, A., Rohou, A., Schep, D.G., Bason, J.V., Montgomery, M.G., Walker, J.E., Grigorieff, N., and Rubinstein, J.L. (2015). Structure and conformational states of the bovine mitochondrial ATP synthase by cryo-EM. Elife *4*, e10180.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Software and Algorithms		
BioChemical Library	This Paper	http://meilerlab.org/index.php/bclcommons/show/b_apps_id/1
Rosetta Modeling Suite	Leaver-Fay et al., 2011	https://www.rosettacommons.org/
SPOCTOPUS	Viklund et al., 2008	http://octopus.cbr.su.se/index.php
SPARTA+	Shen and Bax, 2010	https://spin.niddk.nih.gov/bax/software/SPARTA+/
Pymol	Schrödinger, 2015	https://pymol.org/

CONTACT FOR REAGENTS AND RESOURCE SHARING

Further information and request for computational resources may be directed to, and will be fulfilled by the Lead Author, Dr. Jens Meiler (jens.meiler@vanderbilt.edu).

METHOD DETAILS

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(Viklund et al., 2008) 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 (Jayasinghe et al., 2001) 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

All detailed computational protocol and command documentations should be referred to Supplemental Text. The test dataset of bovine rhodopsin (PDB entry 1GZM) (Li et al., 2004) was downloaded from the PDB and considered the 'native' structure. An experimentally determined electron density map for rhodopsin (Ruprecht et al., 2004) 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+ (Shen and Bax, 2010), 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 (Weiner et al., 2014). Ten sets of distance data from EPR double electron-electron resonance (DEER) spectroscopy was simulated with a distance uncertainty model (Alexander et al., 2008). 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 Tables S1 and S2.

The protocol was based on the protein structure prediction protocols of BCL::MP-Fold (Weiner et al., 2013) and BCL::EM-Fold (Lindert et al., 2012). The SSEs were first predicted from the primary structure of Rhodopsin using the consensus of two secondary

structure prediction methods, JUFO9D (Koehler et al., 2009) and SPOCTOPUS (Viklund et al., 2008). 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 (Lindert et al., 2012), and a knowledge-based distance agreement evaluation in the case of EPR (Alexander et al., 2008) and NMR (Weiner et al., 2014). 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 (Carugo and Pongor, 2001) 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. All computational protocol and command documentation mentioned in this section are referred to in Supplemental Text.

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 (Vernon et al., 2013). For each of the models from the previous BCL stage, 500 models were sampled using Rosetta's cyclic coordinate descent algorithm (Canutescu and Dunbrack, 2003) to build loops and remodel TMHs. Further atomic detail refinement was carried out using Rosetta's "relax" application (Leaver-Fay et al., 2011) 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 (Dimaio et al., 2009). 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 S1). If the measured distance of xÅ fulfills the criterion $lb \leq x \leq ub$, ;if x is outside the boundary, a score penalty would be given: if $x \leq lb$, $f(x) = \left(\frac{x-lb}{sd}\right)^2$; if $ub \leq x \leq ub + rswitch*sd$, $f(x) = \left(\frac{x-ub}{sd}\right)^2$; if $ub + rswitch*sd \leq x$, $f(x) = \frac{1}{sd}(x-(ub+rswitch*sd)) + \left(\frac{rswitch*sd}{sd}\right)^2$, where the rswitch term is set to default of 0.5. The distance restraints from EPR DEER measurements (Table S2) are treated with a knowledge-based potential as detailed in (Hirst et al., 2011). The score terms were linearly combined with respective weighting to compute the total Rosetta energy score. The RMSD100 relative to the 'native' 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 (Schrödinger, 2015).