Outer membrane β-barrel structure prediction through the lens of AlphaFold2

Annika Topitsch[1,*], Torsten Schwede[1,2], Joana Pereira[1,2,†]

[1] *Biozentrum, University of Basel, Basel, Switzerland*

[2] *SIB Swiss Institute of Bioinformatics, Basel, Switzerland*

*Current address: Institute of Surgical Pathology, University Medical Center Freiburg, Germany

†Correspondence: joana.pereira@unibas.ch

Abstract

Most proteins found in the outer membrane of Gram-negative bacteria share a common domain: the transmembrane β-barrel. These outer membrane β-barrels (OMBBs) occur in multiple sizes, and different families with a wide range of functions evolved independently by amplification from a pool of homologous ancestral ββ-hairpins. This is part of the reason why predicting their three-dimensional (3D) structure, especially by homology modeling, is a major challenge. Recently, DeepMind's AlphaFold v2 (AF2) became the first structure prediction method to reach close-to-experimental atomic accuracy in CASP even for difficult targets. However, membrane proteins, especially OMBBs, were not abundant during its training, raising the question of how accurate the predictions are for these families. In this study, we assessed the performance of AF2 in the prediction of OMBBs of various topologies using an in-house-developed tool for the analysis of OMBB 3D structures, *barrOs*. In agreement with previous studies on other membrane protein classes, our results indicate that AF2 predicts OMBB structures at high accuracy independently of the use of templates, even for novel topologies absent from the training set. These results provide confidence on the models generated by AF2 and open the door to the structural elucidation of novel OMBB topologies identified in high-throughput OMBB annotation studies.

Introduction

Protein structure prediction is an important tool to gain insights into the function and biological role of macromolecular machines from three-dimensional (3D) models. While the number of known natural protein sequences has been increasing exponentially [1,2] since the first sequencing of a protein in the 1950s, the experimental determination of macromolecular 3D structures is a laborious task. For this reason, and even with the recent considerable improvements in experimental biophysical methods, the rate by which protein structures are deposited in the Protein Data Bank (PDB) [3] is much lower than that by which protein-coding sequences are made available through GenBank or the UniProt Knowledgebase. One way of tightening this gap is to use computational approaches such as homology modeling, threading, or *ab initio* methods for protein structure prediction [4–7].

The Critical Assessment of Protein Structure Prediction (CASP) [8] experiment provides a platform for the benchmarking of such methods and, since its onset in the early 1990s, has fostered the development of multiple approaches exploring a wide range of data sources and computational techniques. Until recently, homology modeling was the method of choice to model 3D structures of proteins with homologs of known structure in the PDB, while *ab initio* methods were preferred for all others. However, *ab initio* modeling was rarely able to reach for such difficult targets the same level of accuracy as that typically reached by homology modeling. That changed in 2020, with DeepMind's second version of the AlphaFold algorithm (AF2) providing, on average, close to experimental accuracy levels for most targets in the 14th round of CASP (CASP14) [9,10], and later significantly expanding the structural coverage of the cataloged protein sequence space [11,12].

AF2 is a deep neural network with two attention-based transformation modules, where evolutionary, physical, and geometric information is used to perform end-to-end protein structure prediction [10]. The first module, the Evoformer, uses information from multiple sequence alignments (MSAs) and templates to generate a pair representation, a contact map of sorts, for the input. The second module, the structure module, uses this representation and the input sequence to fold the target protein. The network has been trained with all protein structures in the PDB as of April 30, 2018, and it is not tailored to any specific class of proteins.

However, the PDB is biased towards those proteins that are 'easier' to experimentally characterize, with only 10 % of its content corresponding to membrane proteins [13]. For this reason, it is not expected that AF2 is able to predict the 3D structure of transmembrane proteins as accurately as soluble ones, especially when multiple domains are present. In a recent study, Hegedűs *et al.* assessed AF2 structure prediction of α-helical transmembrane proteins. They observed that the models predicted by AF2 exhibited a known fold of α-helical transmembrane proteins for all 1,137 test cases, suggesting that the prediction of transmembrane proteins by AF2 is as accurate as for soluble proteins [14].

In this short report, we focus on the second-largest class of transmembrane proteins: the outer membrane β-barrels (OMBBs). OMBBs are abundant in Gram-negative bacteria, but are also found in chloroplasts, mitochondria and mitochondria-associated organelles [15–17]. They have both medical and biotechnological importance [18–20] as they are composed of an antiparallel β-sheet that connects back to itself to form a pore that crosses the outer membrane, where they perform a large variety of biological activities essential for survival [21]. They are found in a large spectrum of protein families either as single domains, together with other domains, or in multiple copies in the same chain [22]. Different OMBB families are composed of different numbers of β-strands [22,48]. The diameter of the barrel depends on the numbers of β-strands, but also on the shear number, which is, simply put, a measure of the parallel displacement of the strands relative to each other [23–25].

OMBB structure prediction is a challenging task as they can be traced back to a pool of homologous ancestral ββ-hairpins and novel families emerge by the reuse and amplification of smaller pieces from other OMBBs [26–29]. In the special case of homology modeling, when dealing with a novel family for which no full-length template is known or for which the full-length template corresponds to part of a larger β-barrel, the resulting model will either correspond to (1) a mix of local matches with mismatching shears that prevent the proper closing of the barrel, or (2) an incomplete, open barrel incompatible with the membrane environment. Current OMBB modeling approaches circumvent these problems by using external information specific to these proteins. These include the generation of perfect barrel structures directly from a theoretical description of a barrel [15,23,25,30], the prediction of transmembrane segments from sequence features [31–36] and their fitting into a putative membrane [37], and the prediction of contacts between those segments from free energy potentials based on statistical models [38–41] or evolutionary couplings [42,43].

In this short report, we sought to evaluate how the family-agnostic AF2 network performs for OMBBs. As of the time of AF2 training, about 100 single-chained OMBBs at a maximum of 70 % sequence identity (table S1) and with 8 up to 26 β-strands were deposited in the PDB. In the meantime, the structure of a 36-stranded OMBB, the translocon of the Fibrobacteres-Chlorobi-Bacteroidetes type 9 secretion system, was solved by cryo-EM [44], and more than 30 previously unknown OMBB families were predicted at the sequence level, including the largest ever reported OMBB, with at least 38 predicted strands [22]. In the case of long-known OMBB topologies, structural information has been fed into the network during the training phase. Thus, even without using homologous structures for modeling, OMBB models of high accuracy are expected. But since the newest topology of a 36-stranded OMBB was discovered after the date limit for inclusion in the training set, it is unclear how AF2 performs in such cases and what the impact of using templates is.

Methods

Collection of test case structures

Ten OMBBs of known structure, covering topologies of 8- to 24-stranded barrels (table S1) were first used as input for searches against an HHM database of the PDB70 (as of February 2021) with HHpred [45] through the MPI Bioinformatics toolkit [46]. Default parameters were used and all PDB chains matched at a p-value better than 10 collected. These were then analyzed with *barrOs* in order to (1) identify the matched PDB IDs that carry a barrel fold, (2) extract geometric features of the barrel region, and (3) extract the sequence of the barrel domains. With this, 129 unique OMBBs of known structure were collected and the sequences of the detected barrel regions, which include the barrel-forming strands and the connecting loops, were used as input to AF2.

Identification of OMBB folds and extraction of barrel geometric features with *barrOs*

barrOs (for *barrel circle searcher*) is an in-house-developed tool that, given a PDB structure, uses a graph-based approach to identify the strands that form a barrel fold and then uses them to compute geometric features. This includes the number of strands, the barrel diameter, and the shear number of the barrel region. The method is family-agnostic and can take as input (1) a PDB structure, (2) a list of PDB IDs, or (3) HHsearch output files. It can be targeted specifically to transmembrane proteins by using the Orientations of Proteins in Membranes (OPM) database [47] as a source of 3D structures, and to OMBBs specifically by combining it with the results from HHsearch.

For each structure to be analyzed, *barrOs* starts by extracting all Cα atoms and searches for all β-strands. This is done by (1) running DSSP [48,49] and, in parallel, (2) detecting what we denote as 'regular regions'. Regular regions are continuous backbone segments where the angle between the $C\alpha_i$-$C\alpha_{i+2}$ and $C\alpha_{i+1}$-$C\alpha_{i+3}$ vectors is lower than 25°. Regular and stranded region annotations derived from the DSSP ('E') output are then fused, and the resulting continuous intervals referred to as 'strands'.

Two strands where the minimum interstrand distance of their Cα atoms is less than 5 Å are considered adjacent, allowing the construction of a strand-connectivity matrix. This matrix is then used to build an undirected, labeled graph, and the *cycle_basis* function implemented in *NetworkX* [50] is used to identify the nodes, i.e., the strands, that form a closed cycle. Given that OMBBs typically have an even number of strands (except for the 19-stranded mitochondria-specific OMBBs), if the resulting number of barrel-forming strands (the estimated topology) is uneven, this process is repeated using the regular regions or the DSSP-extracted strands until an even topology is obtained. If the result remains uneven, that topology is considered. Structures without detected structured barrels are excluded, and only those with a detected barrel fold are used for further analysis. This includes the estimation of the barrel height, average diameter and shear number, as defined in Murzin *et al.* [23].

Running AF2

AF2 models were generated for the 129 OMBBs in three independent experiments with AlphaFold v2.0.1 on a local cluster instance with three different parameter settings: The first was performed with the default pipeline, which includes the use of all templates found in the PDB (labeled 'M'). In the second, AF2 was run without considering any templates by setting the `--max_template_date` option to 1900-01-01 (labeled 'Mnotemp'). And in the third, template usage was partially allowed by setting the `--max_template_date` option to one day prior to the respective release date in order to exclude the deposited structure from being used as a template for modeling (labeled 'Mreldate').

Model comparison and visualization

All AF2 models were used as input files for *barrOs* to identify barrel topologies and extract geometrical information of the barrel domains. Calculations of the template modeling score (TM-score) and the root-mean-square deviation (RMSD) were carried out with TM-align [51]. OpenStructure was used to calculate the per-residue local distance difference test score (Cα-lDDT) for each model [52].

Results

As a first step, we evaluated how well AF2 captures the core geometric features of OMBBs in the presence and absence of templates, especially the topology of the domain, the average diameter of the channel and the shear, which measures the extent by which the strands are staggered (fig. 1). For that, the 129 experimental structures and the corresponding AF2 models were used as input for *barrOs.* The first observation is that AF2 predicted models with the correct topology for most cases; out of the 129, there were only two test cases where the number of strands in the model deviated by ±1. In these cases, visual inspection highlighted that the difference is not a result of an incorrectly modeled topology, but due to minor differences in the experimental structure and the AF2 model that misled *barrOs* during the identification of regular regions (figs. S1A-B).

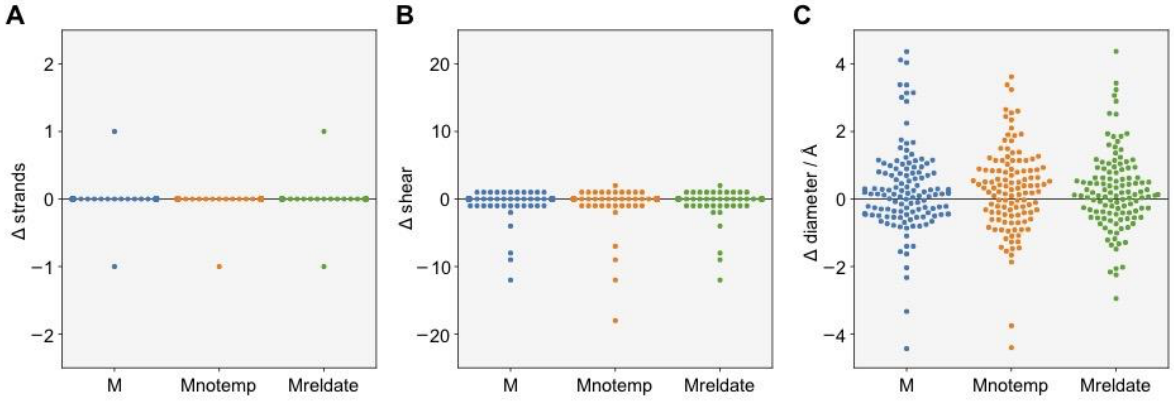

Figure 1. Predicted OMBB topologies and geometries. Differences (Δ) of the number of strands (A), shear numbers (B) and barrel diameters (C) in target structures and AF2 models. Data of AF2 models generated with using templates ('M'), without using templates ('Mnotemp') and with using templates up to the release date of the target structure ('Mreldate'), is shown in blue, orange and green, respectively.

Regarding the shear and barrel diameter, there are also only marginal differences between the target structures and the models predicted by AF2, with some noteworthy exceptions. One striking case is that of *Vibirio cholerae* OmpT (PDB ID 6EHD), where the shear of the model generated with no templates ('Mnotemp') was 18 residues smaller (fig. S1C). The reason here lies on an extracellular loop that in both AF2 models predicted using templates ('M' and 'Mreldate') and the experimental structure is modeled inwards, facing the channel of the barrel, while in the 'Mnotemp' model it faces the exterior, extending the strands that build the barrel region and leading to an incorrect value of the shear.

The agreement between the geometric features of AF2 models and their target experimental structure is also corroborated by superposition-based and superposition-free quality metrics. In the case of superposition-based metrics, high median TM-scores, and correspondingly low RMSD values, were observed for all three experiments (fig. 2A-B). The highest median TM-scores (0.98 ± 0.02) were obtained with the AF2 default pipeline, in which template information is used for the prediction of models ('M'), but also when templates up to the release date ('Mreldate') were allowed. Excluding templates completely ('Mnotemp') only lead to marginally, and not statistically significant, lower TM-scores (0.97 ± 0.02).

This testifies to an overall high accuracy of the AF2 models independently on the use of templates, yet there are a few outliers below and above the lower and upper quartile of the TM-score and RMSD distributions, respectively. The lowest TM-scores of < 0.5 (and highest RMSD values of > 4 Å) were observed for AF2 models of the 8-stranded Opa OMBB (PDB ID 2MLH), which is crucial for the recognition and engulfment of bacterial pathogens *Neisseria gonorrhoeae* or *Neisseria meningitidis* by human cells during pathogenesis [53]. The target structure used is one of the 20 calculated conformers with the lowest energy determined by solution NMR. While the barrel region was predicted accurately, only the extracellular loops did not overlap with the target structure (fig. 2C). This is further highlighted by the superposition-free per-residue Cα-lDDT (fig. S2B), where the scores are higher (> 75) for stranded regions than for the loops (< 50). In this particular case, the flexibility of those loops is in fact essential for the function of the protein, and thus it is not surprising that the AF2 prediction does not match the selected solution NMR structure.

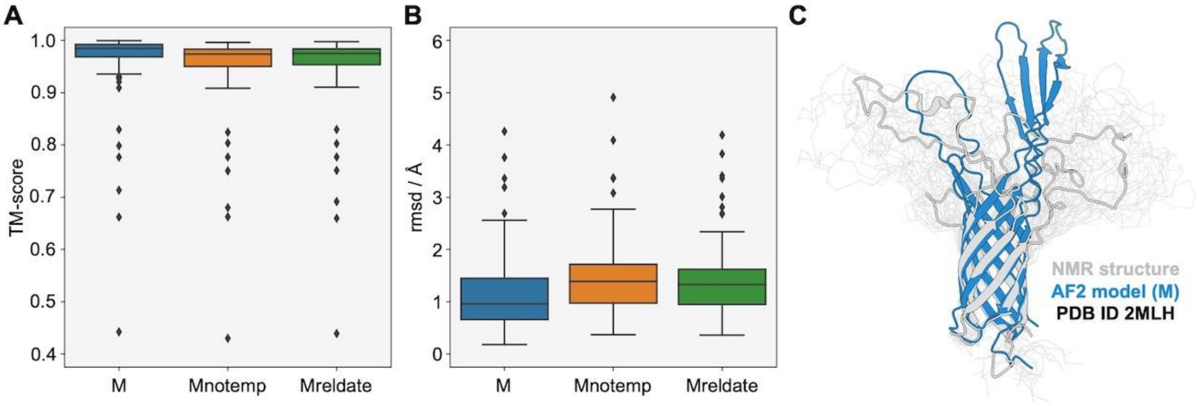

Figure 2. Full-length assessment of target structures and AF2 models. (A) The median TM-scores of the 'M', 'Mnotemp' and 'Mreldate' experiments are 0.98 ± 0.02, 0.97 ± 0.02 and 0.98 ± 0.02, respectively. (B) The median RMSD values, as computed by TMalign, of the 'M', 'Mnotemp' and 'Mreldate' experiments are 1.0 ± 0.5, 1.4 ± 0.5 and 1.3 ± 0.5 Å, respectively. (C) Solution NMR structure and AF2 model of PDB ID 2MLH shown in gray and blue, respectively. The backbone traces of the other 19 calculated conformers are shown in light gray.

This trend is also observed for other cases where the TM-score is above 0.9, an uncertainty also captured by the predicted Cα-lDDT (pLDDT) computed by AF2. Examples of an 8-stranded and a 12-stranded OMBB are shown in figure 3, highlighting the strikingly good prediction accuracy of pLDDT. Both the Cα-lDDTs and pLDDTs reach values between 95 and 100 for β-strand regions, while the loops (specially those facing the extracellular side of the outer membrane) result in lower lDDT values, with no striking differences between the three AF2 'M', 'Mnotemp' and 'Mreldate' predictions. Limiting the analysis to the β-strands forming the barrels (which we refer to as 'barrel cores' for the remaining text) (fig. 4A), resulted in extremely high median lDDT values of 98.2 ± 1.6, 97.0 ± 1.9 and 97.1 ± 1.9 for the models of the 'M', 'Mnotemp' and 'Mreldate' experiments, respectively; corroborating the marginal deviations observed in the different barrel geometric features.

Interestingly, while the lDDT and pLDDT correlate well, their distribution for the barrel core regions is different independently of the use of templates (fig. 4A), with AF2 underestimating, on average, their accuracy. In four cases, however, the confidence of the AF2 models for the barrel regions was above 85 while the lDDT was lower, but still within a reasonable range of 75-80 (fig. 4B). The three first cases are PDB IDs 6QWR, 2MLH and 2K0L, all of which are 8-stranded OMBBs whose structures were determined by NMR spectroscopy with 100 or more calculated conformers. The fourth case corresponds to PDB ID 5O8O, the first experimental structure of a 19-stranded mitochondrial import receptor subunit Tom40. Its experimental structure was determined through rigid body docking into a 6.8-Å resolution cryo-EM map of a homology model generated based on the X-ray structure of a homologous mitochondrial voltage-dependent anion channel (VDAC) [54]. While the overall topology of the AF2 model matches this experimental structure and the same residues build up the barrel core, a few strands exhibit a distinct frame-shift along its axis in all three predicted models (fig. S3A), resulting in average lDDT values below 80. A more recent structure of a homologous Tom40 determined by cryo-EM at higher resolution (PDB ID 6UCU) [55], and which was also a target in this study, agrees with the AF2 model (fig. S3B).

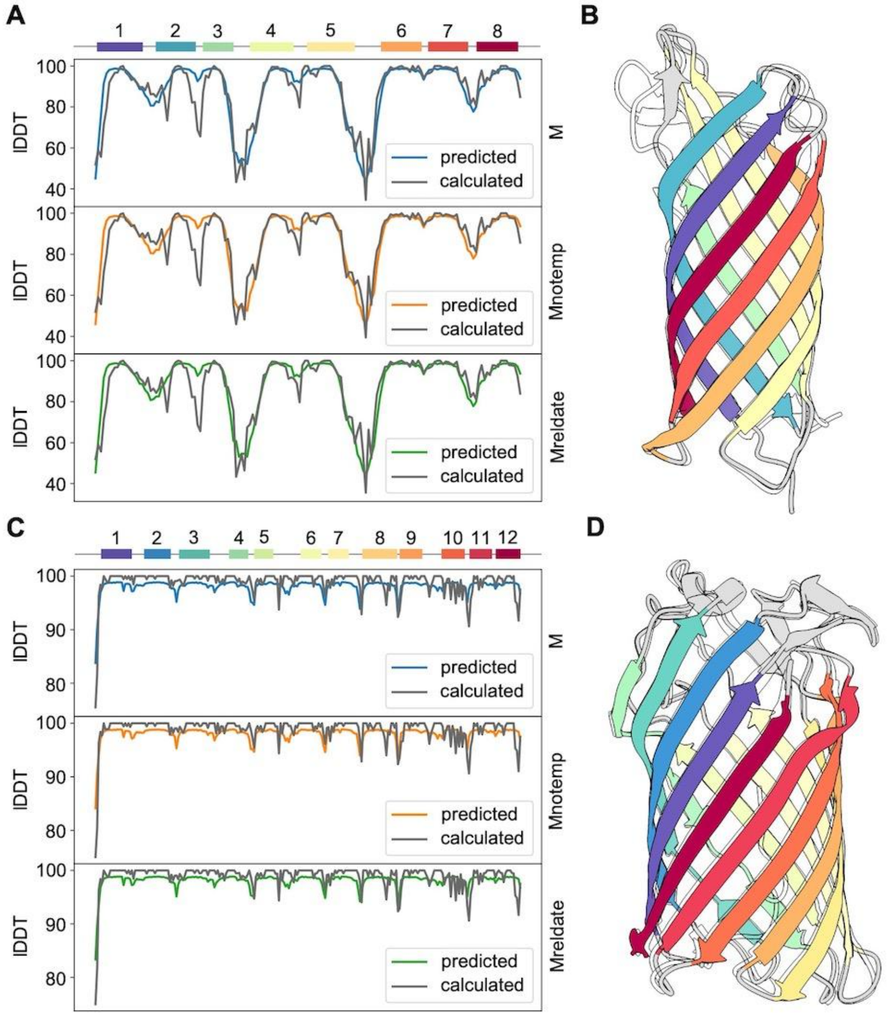

Figure 3. Predicted and calculated Cα-lDDT values per residue for two examples. (A) PDB ID 1P4T, an OMBB with 8 β-strands. (B) and of target PDB ID 4RL8, an OMBB with twelve β-strands. Boxes and numbers indicate the β-strands. The average correlation coefficients between predicted and calculated lDDTs over the three models shown in (A) and (B) are 0.895 ± 0.005 and 0.756 ± 0.009, respectively.

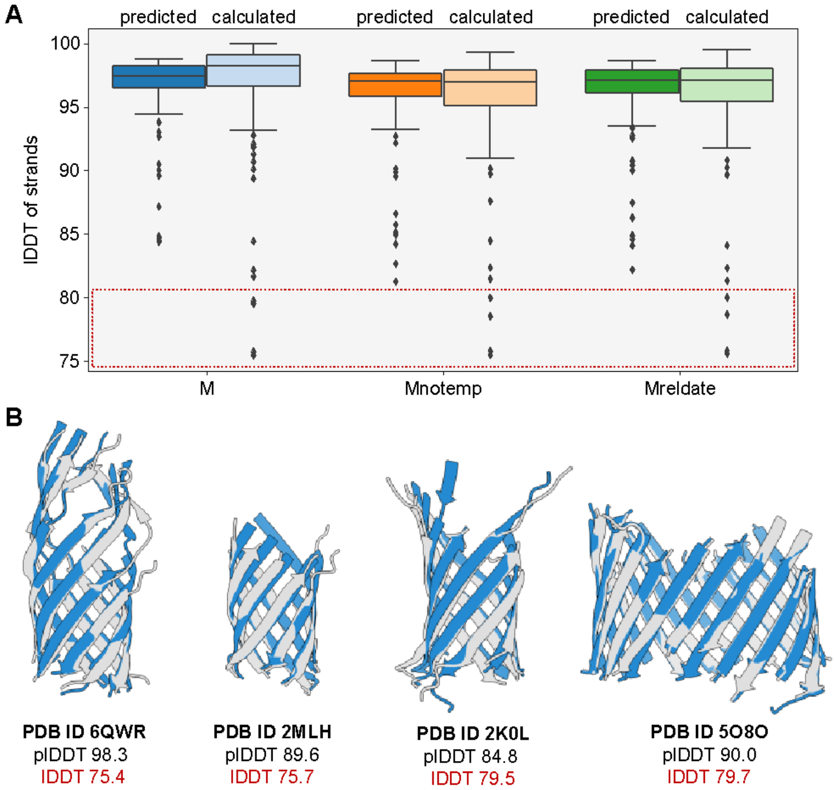

Figure 4. Superposition-free assessment of target structures and AF2 models. (A) Predicted and calculated Cα-lDDT scores of barrel core β-strands are shown in dark and light colors, respectively. (B) The four targets with lDDTs of strands below 80. Shown are only the residues considered as regular regions. Experimental structures and AF2 models are shown in gray and blue, respectively.

Most of these cases, however, were either part of or had full-length homologs in the AF2 training set, thus such high accuracy is expected *a priori.* Of higher interest is the performance of AF2 for cases of novel topology, unknown to AF2. Unfortunately, only one such case is available in the PDB and corresponds to the only known 36-stranded OMBB (PDB ID 6H3I) [44]. It forms the translocon of the Fibrobacteres-Chlorobi-Bacteroidetes type 9 secretion system and its structure was deposited in the PDB after April 30, 2018 (table S1). Although AF2 had never “seen” a 36-stranded barrel, the barrel core and its geometric features were predicted accurately, regardless of the use of templates (fig. 5). In all cases, however, local backbone conformations of the barrel region in the AF2 model are closer to standard geometries of β-sheets than those in the experimental structure. This is not a surprising result as the target is a 3.5 Å cryo-EM structure and lower resolutions lead to higher uncertainties of the atomic coordinates. In the model generated with templates, even the intra- and extracellular loops matched those in the target structure with high accuracy, as also seen in the comparison of predicted and calculated

Cα-lDDT values (fig. S2A). There were only minor displacements in the loop regions of the model predicted without any template information.

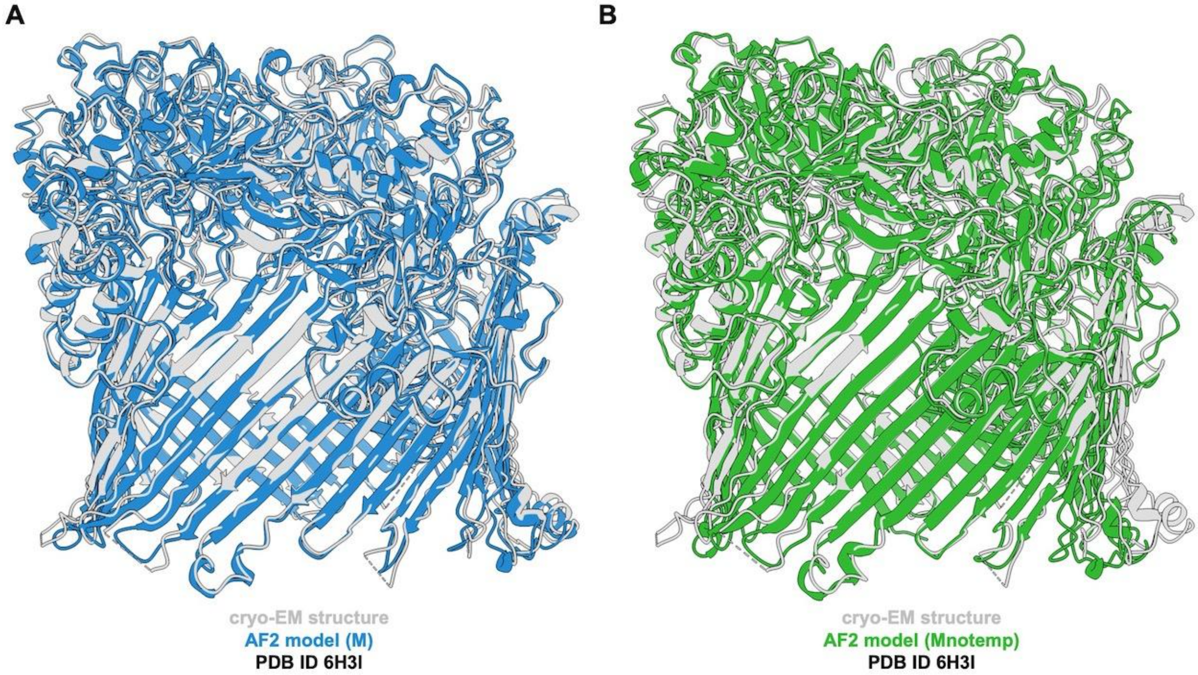

Figure 5. AF2 models of the 36-stranded translocon in the type 9 secretion system (PDB ID 6H3I). Predictions were generated with template (A) and without template (B) information. Excluding templates from the AF2 algorithm resulted in minor displacements in loop regions, while the barrel core was predicted accurately nonetheless. The experimental structure and AF2 models are shown in gray and in color, respectively.

Discussion

Given the under-representation of transmembrane proteins in the PDB, and consequently in the training set of AF2, it is imperative to evaluate how the algorithm performs for such an important class of proteins. While a study was previously carried out for α-helical transmembrane proteins [14], we focused on the second-largest category: the outer membrane β-barrels (OMBBs), especially those found at the surface of Gram-negative bacteria and their eukaryotic homologs. Gram-positive, multimeric transmembrane β-barrels, which evolved by convergence [56], as well as multimeric OMBBs, which are formed by separate polypeptide/protein chains, such as those forming the anchor domain of trimeric autotransporter adhesins [57], were not considered. We identified 129 non-redundant single-chained OMBBs in the PDB, with topologies ranging from 8 to 36 strands. In all cases, AF2 predictions were highly accurate; all experiments resulted in extremely high median TM-scores above 0.97 and low median RMSD values below 1.4 Å and, overall, no significant differences were observed. For all cases, AF2 correctly predicted the topology of the domain as well as the shear and average diameter, demonstrating that in the case of OMBBs the accuracy of the prediction is not substantially affected by the use or omission of templates.

However, targets with structures deposited in the PDB prior to April 30, 2018 were part of the training set. So even when removing the experimental structure from the template list, structural information might still be used to predict the model as it is stored in the network. The only case in our test set with a topology completely new to the AF2 network was the 36-stranded OMBB from the Fibrobacteres-Chlorobi-Bacteroidetes type 9 secretion system translocon [44]. Although the network has never seen a 36-stranded OMBB, its predictions were highly accurate, even improving on the geometry of the backbone of an experimental low-resolution structure. AF2 predicted correctly the 36-stranded topology, as well as the diameter and the shear of the barrel, but also the intricate folds of the extracellular loops at an extremely high level of detail that translates into an overall lDDT of 86. The models for this test case were of the same accuracy as for those of well-known topologies, indicating that structural information of templates or close homologs is not essential for a correct prediction.

On average, the per-residue lDDT is lower for loops, especially those facing the extracellular side of the outer membrane and independently of the use of templates. This is likely the result of the higher flexibility of extracellular loops observed in experimental structures, which is important for protein function. Such flexibility makes it difficult to predict a static snapshot of those regions at an atomic level of detail, which in turn decreases their pLDDT. Larger differences were also observed for cases where the target was either solved by solution NMR or low resolution cryo-EM. The same was observed in CASP14, where AF2 also performed worst for NMR structures [58]. More recently, Fowler *et al*. examined this by measuring the accuracy of solution NMR structures and comparing them to AF2 predictions [59]. They concluded that, in general, AF2 models are more accurate than NMR ensembles. This is especially the case of β-sheet proteins, which include OMBBs, providing a consistent explanation for the observed low lDDT values when comparing AF2 models with NMR structures. However, when evaluating these values, it must be noted that the lDDT scores are merely a measure of how similar the AF2 models and the experimental structures are, without providing information on which structure is closer to the truth. Still, and although the test set is small, these results provide confidence in the models for OMBBs generated with AF2, especially those with previously unknown topologies.

Data and code availability

The models generated, as well as the structural features extracted for them and the reference structures, are available in: https://www.modelarchive.org/doi/10.5452/ma-ombbaf2. *barrOs* can be downloaded from: https://git.scicore.unibas.ch/schwede/barrOs. In the repository, detailed instructions on how to use it for the general analysis of proteins with an expected barrel fold are provided, and the HHsearch results used in this work are provided in the `Examples` folder.

Acknowledgements

We would like to thank the SWISS-MODEL development team for insightful discussions, technical support and text revisions, and sciCORE at the University of Basel for providing computational resources and system administration support. JP would also like to thank Prof. Andrei Lupas and Jens Bassler (Max Planck Institute for Biology Tübingen) for support in the early days of *barrOs*, and Roger Bamert for contributing to the implementation of the shear number calculation into the pipeline.

References

1. NCBI Resource Coordinators. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 2018;46: D8–D13.

2. UniProt Consortium. UniProt: the universal protein knowledgebase in 2021. Nucleic Acids Res. 2021;49: D480–D489.

3. wwPDB consortium. Protein Data Bank: the single global archive for 3D macromolecular structure data. Nucleic Acids Res. 2019;47: D520–D528.

4. Biasini M, Bienert S, Waterhouse A, Arnold K, Studer G, Schmidt T, et al. SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information. Nucleic Acids Res. 2014;42: W252–8.

5. Hildebrand A, Remmert M, Biegert A, Söding J. Fast and accurate automatic structure prediction with HHpred. Proteins: Structure, Function, and Bioinformatics. 2009. pp. 128–132. doi:10.1002/prot.22499

6. Peng J, Xu J. Raptorx: Exploiting structure information for protein alignment by statistical inference. Proteins: Structure, Function, and Bioinformatics. 2011. pp. 161–171. doi:10.1002/prot.23175

7. Anishchenko I, Baek M, Park H, Hiranuma N, Kim DE, Dauparas J, et al. Protein tertiary structure prediction and refinement using deep learning and Rosetta in CASP14. Proteins. 2021;89: 1722–1733.

8. Kryshtafovych A, Schwede T, Topf M, Fidelis K, Moult J. Critical assessment of methods of protein structure prediction (CASP)-Round XIV. Proteins. 2021;89: 1607–1617.

9. Pereira J, Simpkin AJ, Hartmann MD, Rigden DJ, Keegan RM, Lupas AN. High-accuracy protein structure prediction in CASP14. Proteins. 2021. doi:10.1002/prot.26171

10. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021;596: 583–589.

11. Varadi M, Anyango S, Deshpande M, Nair S, Natassia C, Yordanova G, et al. AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. Nucleic Acids Res. 2022;50: D439–D444.

12. Callaway E. "The entire protein universe": AI predicts shape of nearly every known protein. Nature. 2022;608: 15–16.

13. Bittrich S, Rose Y, Segura J, Lowe R, Westbrook JD, Duarte JM, et al. RCSB Protein Data Bank: Improved Annotation, Search, and Visualization of Membrane Protein Structures Archived in the PDB. Bioinformatics. 2021. doi:10.1093/bioinformatics/btab813

14. Hegedűs T, Geisler M, Lukács GL, Farkas B. Ins and outs of AlphaFold2 transmembrane protein structure predictions. Cell Mol Life Sci. 2022;79: 73.

15. Schulz GE. beta-Barrel membrane proteins. Curr Opin Struct Biol. 2000;10: 443–447.

16. Duy D, Soll J, Philippar K. Solute channels of the outer membrane: from bacteria to chloroplasts. Biol Chem. 2007;388: 879–889.

17. Chaturvedi D, Mahalakshmi R. Transmembrane β-barrels: Evolution, folding and energetics. Biochim Biophys Acta Biomembr. 2017;1859: 2467–2482.

18. Choi U, Lee C-R. Antimicrobial Agents That Inhibit the Outer Membrane Assembly Machines of Gram-Negative Bacteria. J Microbiol Biotechnol. 2019;29: 1–10.

19. Heselpoth RD, Euler CW, Schuch R, Fischetti VA. Lysocins: Bioengineered Antimicrobials That Deliver Lysins across the Outer Membrane of Gram-Negative Bacteria. Antimicrob Agents Chemother. 2019;63. doi:10.1128/AAC.00342-19

20. Slusky JSG. Outer membrane protein design. Current Opinion in Structural Biology. 2017. pp. 45–52. doi:10.1016/j.sbi.2016.11.003

21. Fairman JW, Noinaj N, Buchanan SK. The structural biology of β-barrel membrane proteins: a summary of recent reports. Curr Opin Struct Biol. 2011;21: 523–531.

22. Solan R, Pereira J, Lupas AN, Kolodny R, Ben-Tal N. Gram-negative outer-membrane proteins with multiple β-barrel domains. Proc Natl Acad Sci U S A. 2021;118. doi:10.1073/pnas.2104059118

23. Murzin AG, Lesk AM, Chothia C. Principles determining the structure of beta-sheet barrels in proteins. I. A theoretical analysis. J Mol Biol. 1994;236: 1369–1381.

24. Dhar R, Feehan R, Slusky JSG. Membrane Barrels Are Taller, Fatter, Inside-Out Soluble Barrels. J Phys Chem B. 2021;125: 3622–3628.

25. Murzin AG, Lesk AM, Chothia C. Principles determining the structure of β-sheet barrels in proteins II. The observed structures. Journal of Molecular Biology. 1994. pp. 1382–1400. doi:10.1016/0022-2836(94)90065-5

26. Nanda V. Building bigger beta-barrels. Elife. 2019;8. doi:10.7554/eLife.44076

27. Franklin MW, Nepomnyachyi S, Feehan R, Ben-Tal N, Kolodny R, Slusky JS. Evolutionary pathways of repeat protein topology in bacterial outer membrane proteins. Elife. 2018;7. doi:10.7554/eLife.40308

28. Remmert M, Biegert A, Linke D, Lupas AN, Söding J. Evolution of outer membrane beta-barrels from an ancestral beta beta hairpin. Mol Biol Evol. 2010;27: 1348–1358.

29. Pereira J, Lupas AN. The Origin of Mitochondria-Specific Outer Membrane β-Barrels from an Ancestral Bacterial Fragment. Genome Biol Evol. 2018;10: 2759–2765.

30. Chou KC, Carlacci L, Maggiora GM. Conformational and geometrical properties of idealized beta-barrels in proteins. J Mol Biol. 1990;213: 315–326.

31. Bagos PG, Liakopoulos TD, Hamodrakas SJ. Evaluation of methods for predicting the topology of beta-barrel outer membrane proteins and a consensus prediction method. BMC Bioinformatics. 2005;6: 7.

32. Hayat S, Peters C, Shu N, Tsirigos KD, Elofsson A. Inclusion of dyad-repeat pattern improves topology prediction of transmembrane β-barrel proteins. Bioinformatics. 2016;32: 1571–1573.

33. Mizianty MJ, Kurgan L. Improved identification of outer membrane beta barrel proteins using primary sequence, predicted secondary structure, and evolutionary information. Proteins. 2011;79: 294–303.

34. Bigelow H, Rost B. PROFtmb: a web server for predicting bacterial transmembrane beta barrel proteins. Nucleic Acids Res. 2006;34: W186–8.

35. Tsirigos KD, Elofsson A, Bagos PG. PRED-TMBB2: improved topology prediction and detection of beta-barrel outer membrane proteins. Bioinformatics. 2016;32: i665–i671.

36. Madeo G, Savojardo C, Martelli PL, Casadio R. BetAware-Deep: An Accurate Web Server for Discrimination and Topology Prediction of Prokaryotic Transmembrane β-barrel Proteins. J Mol Biol. 2020; 166729.

37. Hayat S, Elofsson A. Ranking models of transmembrane β-barrel proteins using Z-coordinate predictions. Bioinformatics. 2012;28: i90–6.

38. Waldispühl J, Berger B, Clote P, Steyaert J-M. transFold: a web server for predicting the structure and residue contacts of transmembrane beta-barrels. Nucleic Acids Res. 2006;34: W189–93.

39. Waldispühl J, O'Donnell CW, Devadas S, Clote P, Berger B. Modeling ensembles of transmembrane β-barrel proteins. Proteins: Structure, Function, and Bioinformatics. 2008. pp. 1097–1112. doi:10.1002/prot.21788

40. Naveed H, Xu Y, Jackups R Jr, Liang J. Predicting three-dimensional structures of transmembrane domains of β-barrel membrane proteins. J Am Chem Soc. 2012;134: 1775–1781.

41. Tian W, Lin M, Tang K, Liang J, Naveed H. High-resolution structure prediction of -barrel membrane proteins. Proc Natl Acad Sci U S A. 2018;115: 1511–1516.

42. Hayat S, Sander C, Marks DS, Elofsson A. All-atom 3D structure prediction of transmembrane β-barrel proteins from sequences. Proc Natl Acad Sci U S A. 2015;112: 5413–5418.

43. Andreani J, Söding J. bbcontacts: prediction of β-strand pairing from direct coupling patterns. Bioinformatics. 2015;31: 1729–1737.

44. Lauber F, Deme JC, Lea SM, Berks BC. Type 9 secretion system structures reveal a new protein transport mechanism. Nature. 2018;564: 77–82.

45. Söding J. Protein homology detection by HMM-HMM comparison. Bioinformatics. 2005;21: 951–960.

46. Gabler F, Nam S-Z, Till S, Mirdita M, Steinegger M, Söding J, et al. Protein Sequence Analysis Using the MPI Bioinformatics Toolkit. Curr Protoc Bioinformatics. 2020;72: e108.

47. Lomize MA, Pogozheva ID, Joo H, Mosberg HI, Lomize AL. OPM database and PPM web server: resources for positioning of proteins in membranes. Nucleic Acids Res. 2012;40: D370–6.

48. Kabsch W, Sander C. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. Biopolymers. 1983. pp. 2577–2637. doi:10.1002/bip.360221211

49. Touw WG, Baakman C, Black J, te Beek TAH, Krieger E, Joosten RP, et al. A series of PDB-related databanks for everyday needs. Nucleic Acids Res. 2015;43: D364–8.

50. Hagberg AA, Schult DA, Swart PJ. Exploring Network Structure, Dynamics, and Function

using NetworkX. In: Varoquaux G, Vaught T, Millman J, editors. Proceedings of the 7th Python in Science Conference. Pasadena, CA USA; 2008. pp. 11–15.

51. Zhang Y, Skolnick J. TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Res. 2005;33: 2302–2309.

52. Biasini M, Schmidt T, Bienert S, Mariani V, Studer G, Haas J, et al. OpenStructure: an integrated software framework for computational structural biology. Acta Crystallogr D Biol Crystallogr. 2013;69: 701–709.

53. Fox DA, Larsson P, Lo RH, Kroncke BM, Kasson PM, Columbus L. Structure of the Neisserial Outer Membrane Protein Opa60: Loop Flexibility Essential to Receptor Recognition and Bacterial Engulfment. Journal of the American Chemical Society. 2014. pp. 9938–9946. doi:10.1021/ja503093y

54. Bausewein T, Mills DJ, Langer JD, Nitschke B, Nussberger S, Kühlbrandt W. Cryo-EM Structure of the TOM Core Complex from Neurospora crassa. Cell. 2017;170: 693–700.e7.

55. Tucker K, Park E. Cryo-EM structure of the mitochondrial protein-import channel TOM complex at near-atomic resolution. Nat Struct Mol Biol. 2019;26: 1158–1166.

56. Franklin MW, Nepomnyachiy S, Feehan R, Ben-Tal N, Kolodny R, Slusky JSG. Efflux Pumps Represent Possible Evolutionary Convergence onto the β-Barrel Fold. Structure. 2018;26: 1266–1274.e2.

57. Linke D, Riess T, Autenrieth IB, Lupas A, Kempf VAJ. Trimeric autotransporter adhesins: variable structure, common function. Trends Microbiol. 2006;14: 264–270.

58. Huang YJ, Zhang N, Bersch B, Fidelis K, Inouye M, Ishida Y, et al. Assessment of prediction methods for protein structures determined by NMR in CASP14: Impact of AlphaFold2. Proteins. 2021;89: 1959–1976.

59. Fowler NJ, Williamson MP. The accuracy of protein structures in solution determined by AlphaFold and NMR. bioRxiv. 2022. p. 2022.01.18.476751. doi:10.1101/2022.01.18.476751

60. Waterhouse A, Bertoni M, Bienert S, Studer G, Tauriello G, Gumienny R, et al. SWISS-MODEL: homology modelling of protein structures and complexes. Nucleic Acids Res. 2018;46: W296–W303.

Supplementary Information

Table S1 Targets used for AF2 modeling. Initial seeds for HHsearch are shown in bold print.

PDB ID and chain	Organism	Topology	Shear number	Diameter (Å)	Release date
2POR_A	*Rhodobacter capsulatus*	16	20	34.75	07/15/93
1PHO_A	*Escherichia coli*	16	20	33.17	10/31/93
1A0T_Q	*Salmonella enterica*	18	22	37.30	03/18/98
1AF6_B	***Escherichia coli***	**18**	**22**	**36.57**	**03/25/98**
1AF6_C	***Escherichia coli***	**18**	**22**	**36.39**	**03/25/98**
1A0S_Q	*Salmonella enterica*	18	22	37.87	06/10/98
3PRN_A	*Rhodobacter blasticus*	16	20	33.61	08/12/98
1FEP_A	*Escherichia coli*	22	24	41.63	01/13/99
1QJ8_A	***Escherichia coli***	**8**	**8**	**15.16**	**10/10/99**
1QD5_A	*Escherichia coli*	12	16	23.23	10/25/99
1QD6_D	*Escherichia coli*	12	16	23.18	10/25/99
1QJP_A	*Escherichia coli*	8	10	15.05	06/30/00
1I78_A	***Escherichia coli***	**10**	**12**	**21.12**	**10/03/01**
1KMO_A	*Escherichia coli*	22	24	40.40	03/06/02
1P4T_A	*Neisseria meningitidis*	8	9	15.41	07/22/03
1UYN_X	*Neisseria meningitidis*	12	14	22.62	03/18/04
1XKW_A	*Pseudomonas aeruginosa*	22	24	40.98	10/04/05
2ERV_A	*Pseudomonas aeruginosa*	8	9	14.99	04/11/06
2GUF_A	*Escherichia coli*	22	24	41.76	12/05/06
2HDI_A	*Escherichia coli*	22	24	41.31	05/08/07
2JMM_A	*Escherichia coli*	8	9	14.63	07/03/07
2VDF_A	*Neisseria meningitidis*	10	12	18.64	10/23/07
2QOM_A	*Escherichia coli*	12	13	23.21	11/13/07
2VQI_A*	***Escherichia coli***	**24**	***N/A****	***N/A****	**05/27/08**
3BRY_A	*Ralstonia pickettii*	14	16	27.70	06/10/08
2ZFG_A	*Escherichia coli*	16	19	33.42	07/29/08
3DWO_X	*Pseudomonas aeruginosa*	14	14	24.12	12/16/08
2K0L_A	*Klebsiella pneumoniae*	8	9	15.92	12/23/08
3EFM_A	***Bordetella pertussis***	**22**	**25**	**41.84**	**03/31/09**
3FHH_A	*Shigella dysenteriae*	22	24	40.35	07/14/09
2WJR_A	***Escherichia coli***	**12**	**14**	**23.10**	**10/13/09**
3AEH_A	*Escherichia coli*	12	14	22.91	07/07/10
2X55_A	*Yersinia pestis*	10	13	19.19	07/28/10
2X27_X	*Pseudomonas aeruginosa*	8	9	14.15	12/15/10
2X9K_A	*Escherichia coli*	14	16	27.22	01/26/11
3QQ2_A	*Bordetella pertussis*	12	14	24.23	04/13/11
3PGU_A	***Escherichia coli***	**14**	**14**	**23.79**	**05/25/11**
3RFZ_B	*Escherichia coli*	24	27	46.63	06/01/11
3NSG_C	*Salmonella enterica*	16	19	32.43	07/13/11
2LHF_A	*Pseudomonas aeruginosa*	8	10	15.47	08/24/11
3SLT_A	*Escherichia coli*	12	14	22.20	11/16/11
3QRA_A	*Yersinia pestis*	8	8	13.43	11/23/11
2Y0H_A	*Pseudomonas aeruginosa*	18	23	35.52	12/21/11
2Y0L_A	*Pseudomonas aeruginosa*	18	22	34.76	12/21/11
3SYB_A	*Pseudomonas aeruginosa*	18	22	36.01	02/08/12
3SZV_A	*Pseudomonas aeruginosa*	18	22	35.43	02/08/12
3SYS_A	*Pseudomonas aeruginosa*	18	22	36.11	02/08/12
3SY7_A	*Pseudomonas aeruginosa*	18	22	35.71	02/08/12
3SZD_B	*Pseudomonas aeruginosa*	18	22	35.18	02/08/12
3V8X_A	*Neisseria meningitidis*	22	24	40.35	02/29/12
4E1T_A	*Yersinia pseudotuberculosis*	12	14	23.04	06/13/12

PDB ID and chain	Organism	Topology	Shear number	Diameter (Å)	Release date
2POR_A	*Rhodobacter capsulatus*	16	20	34.75	07/15/93
4E1S_A	*Escherichia coli*	12	14	23.54	06/13/12
4EPA_A	*Yersinia pestis*	22	24	40.53	06/20/12
4GEY_A	*Pseudomonas putida*	16	21	32.95	10/24/12
4B7O_A	*Neisseria meningitidis*	22	24	42.96	01/23/13
2YNK_A	*Escherichia coli*	18	20	34.42	05/15/13
4FQE_A	*Dickeya dadantii*	12	14	23.10	06/26/13
4FUV_A	*Acinetobacter baumannii*	8	10	19.57	07/10/13
4FRT_A	*Pseudomonas aeruginosa*	18	22	35.37	07/24/13
4FSO_A	*Pseudomonas aeruginosa*	18	22	35.30	07/24/13
4FSP_A	*Pseudomonas aeruginosa*	18	22	35.42	07/24/13
4FT6_A	*Pseudomonas aeruginosa*	18	22	35.77	07/24/13
4K3B_A	*Neisseria gonorrhoeae*	16	22	36.24	09/04/13
4K3C_A	*Haemophilus ducreyi*	16	22	34.69	09/04/13
4C00_A	*Escherichia coli*	16	8	32.54	09/25/13
3WI5_A	*Neisseria meningitidis*	16	20	33.66	01/01/14
4CU4_A	*Escherichia coli*	22	24	41.19	04/09/14
4C4V_B	*Escherichia coli*	16	22	35.12	04/23/14
4MEE_A	*Escherichia coli*	12	14	23.02	06/04/14
4C69_X	***Mus musculus***	**19**	**20**	**35.03**	**06/04/14**
2MLH_A	*Neisseria gonorrhoeae*	8	9	16.96	06/25/14
4Q35_A	*Shigella flexneri*	26	30	53.57	06/25/14
4N74_A	***Escherichia coli***	**16**	**20**	**32.36**	**10/15/14**
4RLC_A	*Pseudomonas aeruginosa*	8	9	14.67	04/22/15
4QL0_A	*Bordetella pertussis*	16	20	32.73	06/17/15
4RL8_A	*Pseudomonas putida*	12	14	23.45	07/29/15
4RDR_A	*Neisseria meningitidis*	22	24	41.05	08/19/15
4RJW_A	*Pseudomonas aeruginosa*	16	20	31.27	10/21/15
5FP2_A	*Pseudomonas aeruginosa*	22	24	41.82	12/09/15
5DL6_A	*Acinetobacter baumannii*	18	22	35.73	02/03/16
5DL7_A	*Acinetobacter baumannii*	18	21	35.02	02/03/16
5DL5_A	*Acinetobacter baumannii*	18	22	34.69	02/03/16
5D0O_A	*Escherichia coli*	16	22	35.61	03/09/16
4D65_B	*Providencia stuartii*	16	20	33.75	03/09/16
4Y25_A	*Escherichia coli*	16	18	30.66	03/16/16
5FP1_A	*Acinetobacter baumannii*	22	24	40.69	05/11/16
5FOK_A	*Pseudomonas aeruginosa*	22	24	42.42	05/11/16
5FR8_A	*Acinetobacter baumannii*	22	23	38.30	05/11/16
5IV8_A	*Klebsiella pneumoniae*	26	30	50.60	05/18/16
5IXM_A	*Yersinia pestis*	26	30	51.43	05/18/16
5IVA_A	*Pseudomonas aeruginosa*	26	30	47.78	05/18/16
5FVN_B	*Enterobacter cloacae*	16	20	33.38	08/10/16
4ZGV_A	*Pectobacterium atrosepticum*	22	24	41.32	08/31/16
5LDV_A	*Campylobacter jejuni*	18	22	36.33	10/26/16
5FQ8_B	*Bacteroides thetaiotaomicron*	22	24	41.94	12/21/16
5O8O_A	***Neurospora crassa***	**19**	**20**	**35.20**	**08/16/17**
5O65_B	*Pseudomonas sp.*	12	14	24.09	08/23/17
5MDR_A	*Vibrio harveyi*	16	20	33.65	12/20/17
5ONU_A	*Vibrio cholerae*	16	20	33.39	01/03/18
5M9B_A	*Pseudomonas aeruginosa*	22	24	38.64	02/21/18
6EHB_A	*Vibrio cholerae*	16	20	33.70	04/25/18
6EHD_A	*Vibrio cholerae*	16	2	32.56	04/25/18
5O77_A	*Klebsiella pneumoniae*	16	19	32.37	06/20/18
6GIE_A	*Acinetobacter baumannii*	14	14	24.63	09/19/18
6CD2_C	*Escherichia coli*	24	25	47.13	10/03/18

PDB ID and chain	Organism	Topology	Shear number	Diameter (Å)	Release date
2POR_A	*Rhodobacter capsulatus*	16	20	34.75	07/15/93
6HCP_B	*Acinetobacter baumannii*	22	24	40.25	10/10/18
6H3I_A	*Flavobacterium johnsoniae*	36	40	71.37	11/07/18
6H3I_F	*Flavobacterium johnsoniae*	14	14	25.14	11/07/18
6EUS_B	*Acinetobacter baumannii*	16	20	33.16	11/14/18
6BPN_A	*Escherichia coli*	22	24	41.46	11/28/18
6FOK_A	*Pseudomonas aeruginosa*	22	24	41.41	03/13/19
6E4V_A	*Escherichia coli*	22	24	40.90	04/10/19
6QGW_A	*Escherichia coli*	16	22	36.21	06/26/19
6I96_A	*Pseudomonas aeruginosa*	22	24	39.50	08/28/19
6OFR_A	*Escherichia coli*	22	24	41.42	10/02/19
6TZK_A	*Escherichia coli*	16	9	29.99	10/23/19
6UCU_A	*Saccharomyces cerevisiae*	19	20	35.26	11/06/19
6QWR_A	*Pseudomonas oleovorans*	8	10	14.36	03/18/20
6V78_C	*Klebsiella pneumoniae*	16	19	33.89	04/01/20
6V78_A	*Klebsiella pneumoniae*	16	19	33.89	04/01/20
6R2Q_B	*Shewanella baltica*	26	26	43.33	04/22/20
6V81_A	*Escherichia coli*	22	24	41.62	05/06/20
6SLN_A	*Porphyromonas gingivalis*	22	24	42.11	05/20/20
6WIL_A	*Acinetobacter baumannii*	16	20	33.86	11/04/20
6Z34_A	*Pseudomonas putida*	14	14	28.35	11/04/20
6WIM_A	*Escherichia coli*	16	20	32.93	11/04/20
6ZLT_B	*Bacteroides thetaiotaomicron*	22	24	42.43	11/11/20
6Z8I_B	*Bacteroides thetaiotaomicron*	22	24	42.12	11/18/20
6Z8A_A	*Pseudomonas aeruginosa*	22	23	39.78	11/25/20
7ACG_A	*Pseudomonas aeruginosa*	18	22	35.83	12/16/20

*The initial target is not part of the PDB70 database and was not therefore used for AF2 modeling.

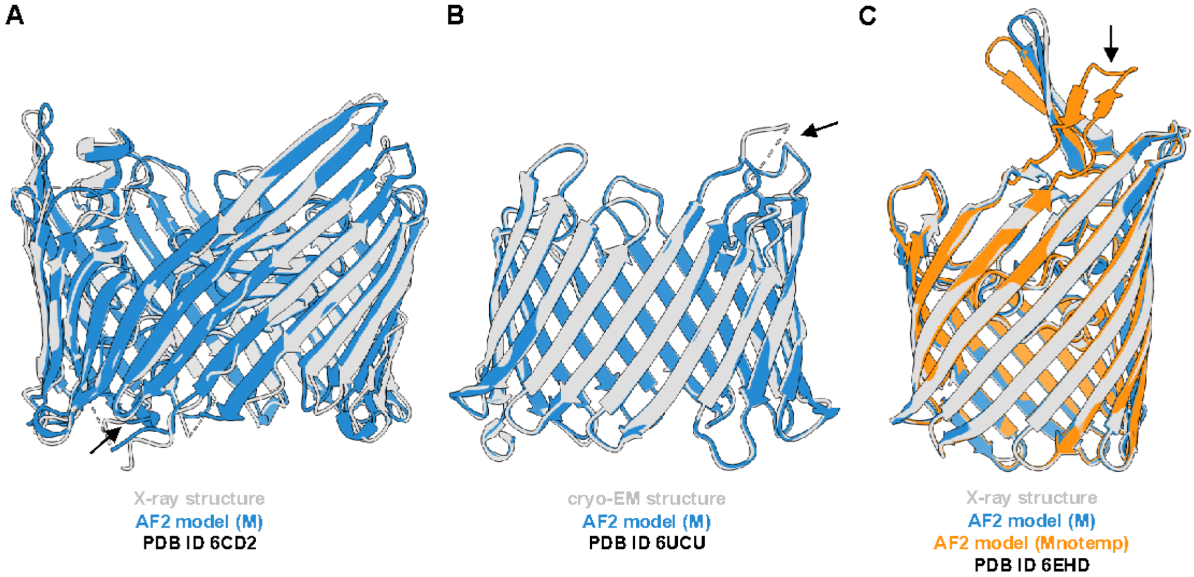

Figure S1. Outliers of the topology analysis. (A) In the AF2 model 'M' of target PDB ID 6CD2, two strands were merged into one single regular region due to a short intracellular loop, leading to a miscalculation of a 23-stranded instead of a 24-stranded topology. (B) In the cryo-EM structure of target PDB ID 6UCU, two strands were falsely identified as one due to missing coordinates of an extracellular loop, resulting in a calculated topology of 18 instead of 19 strands. (C) In the AF2 model 'Mnotemp', an extracellular loop, which in the experimental structure points inwards into the channel, was predicted facing the outside. Since parts of it were assessed as regular regions by *barrOs*, the calculated shear number differed greatly from those of the X-ray experimental structure. Experimental structures, 'M' and 'Mnotemp' AF2 models are shown in gray, blue and orange, respectively.

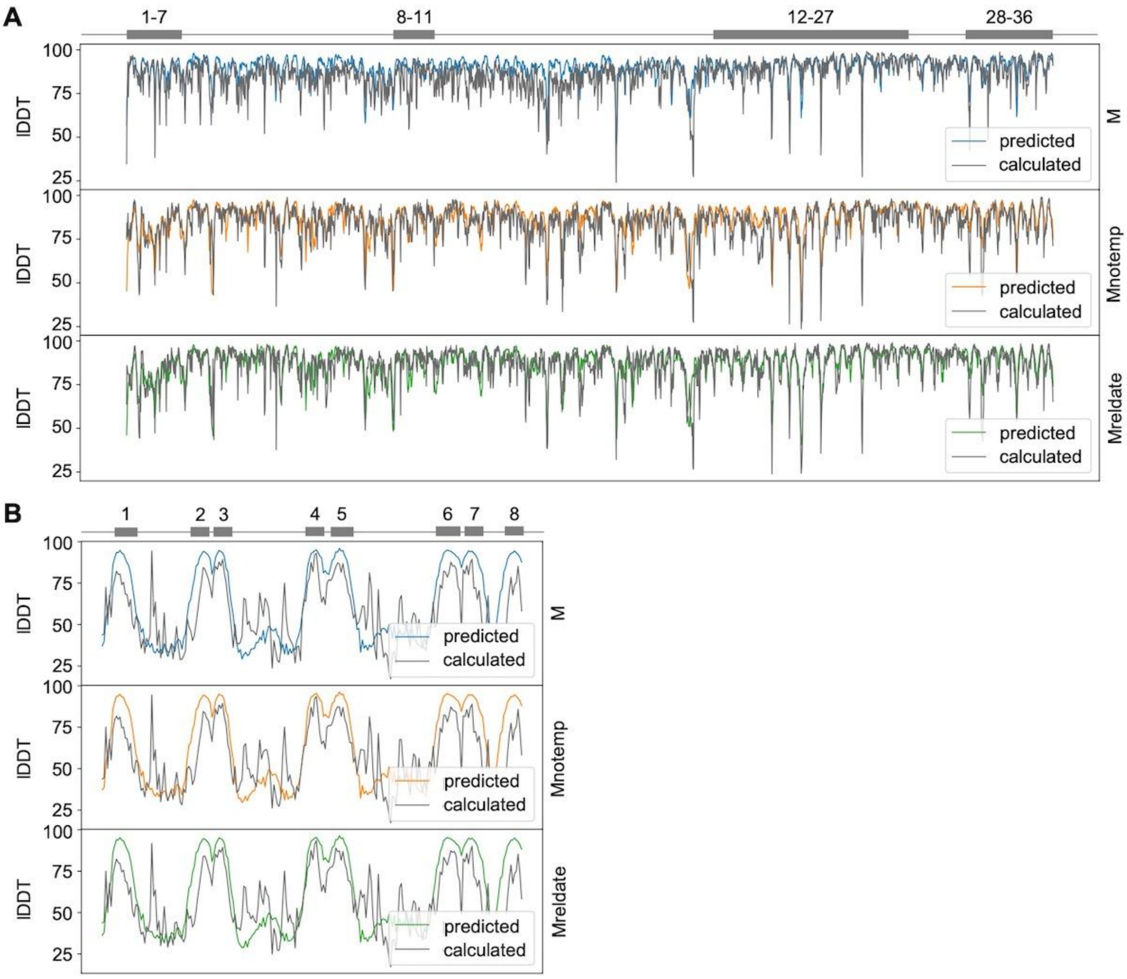

Figure S2. Predicted and calculated Cα-lDDT values for two examples. (A) Target PDB ID 6H3I, an OMBB with 36 β-strands. (B) Target PDB ID 2MLH, an OMBB with 8 β-strands. Gray boxes with numbers indicate β-strands.

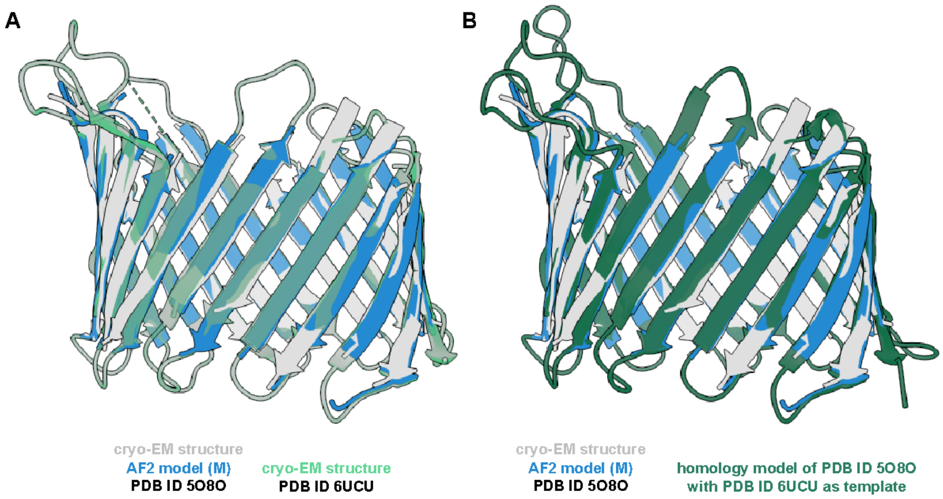

Figure S3. Experimentally determined and predicted structures of Tom40 proteins. The AF2 model (blue) of target PDB ID 5O8O displays a shift in some β-strands as compared to its deposited 6.8-Å cryo-EM structure of *N. crassa* Tom40 (gray). This shift is also observed in the 3.1-Å cryo-EM structure of *S. cerevisiae* Tom40 (PDB ID 6UCU, light green) as well as in a homology model of the target with PDB ID 6UCU as template (dark green). The homology model was generated using SWISS-MODEL [60].