# Is It a Barrel? General Features of Transmembrane Beta Barrels from a Large Database

**Authors:** Daniel Montezano[a], Rebecca Bernstein[a,1], Matthew M. Copeland[a], Joanna S. G. Slusky[a,b,2]

**Affiliations:** [a] Center for Computational Biology, University of Kansas, Lawrence, KS 66045; [b] Department of Molecular Biosciences, University of Kansas, Lawrence, KS 66045;

[1] Currently at University of California, Berkeley.
[2] To whom correspondence should be addressed: slusky@ku.edu

## Abstract

Large datasets for particular classes of proteins can contribute a birds-eye view that is more accurate than that gathered from studying particular exemplars. We used co-evolution data in the form of predicted contact maps to create a large, high-quality database of transmembrane β-barrels (TMBB). We classify sequences for our database using IsItABarrel, an approach combining topology prediction and rule-based TMBB classification. By applying simple feature detection on generated contact maps, our method achieves 95.88% balanced accuracy when discriminating this structural fold from other protein classes. Our database contains almost two million bacterial TMBB proteins from 20 classes, respectively 17 and 2.2 times larger than the previous sets TMBB-DB and OMPdb. Moreover, comparison with IsItABarrel revealed a high rate of false positives in previous TMBB algorithms. We assessed the TMBB-content of 2,959 proteomes of bacterial organisms across several taxonomic categories and find tremendous variance among TMBB-containing organisms with some having as many as 300 TMBBs (6.79% of proteome) and others as few as 10 (0.27% of proteome). The distribution of the lengths of the TMBBs is suggestive of previously hypothesized duplication events. In addition, we find that the C-terminal β-signal varies among different classes of bacteria though it is most commonly $GH_yGH_yG\varphi+\varphi$. The database is available through our IsItABarrel webapp and we anticipate that its quality and size will serve as a useful resource where high quality TMBB sequence data is required.

**Keywords**: database, classification, protein sequence, contact map, transmembrane β-barrel, outer membrane protein, structure prediction, bacteria, phylogeny, β-signal

## Introduction

The quality and size of biological databases controls comparative studies of organisms, genomes and protein families. These studies form the basis for answering questions about the evolutionary history of different protein families and for identifying and classifying the diversity of their structures and functions. The proteins found in the outer membrane of Gram-negative bacteria, generally have a highly similar β-barrel fold (Schulz, 2002), which has potentially evolved mostly by replication of an ancestral hairpin motif (Franklin, Nepomnyachyi, et al., 2018; Remmert et al., 2010). Despite the significant structural similarity, this fold emerges from a rich diversity of primary sequences (Franklin, Nepomnyachiy, et al., 2018). The functional roles of transmembrane β-barrels (TMBBs) are also very diverse, including nutrient uptake, membrane stabilization, catalysis, cell adhesion, cell signaling and efflux (Schulz, 2002).

The hydrophobic environment of the outer membrane imposes constraints on the sequence and topology of TMBB proteins. Since the outside of barrels is in contact with the membrane and the inside of the barrel with the water filled lumen, strand residues that point into the channel are usually

hydrophilic, while residue side chains pointing outside into the membrane are hydrophobic and aromatic. Such constraint induces an imperfect hydrophobicity alternation pattern on residues of each strand. (Dhar et al., 2021; Slusky & Dunbrack, 2013) Also, there are significant differences in amino acid abundance at different depths across the membrane, with more hydrophobic residues in the central region of the membrane. (Wimley, 2002) Attempts have been made to computationally identify TMBBs using amino acid composition (Gromiha & Suwa, 2006; Liu et al., 2003; Zhai & Saier, 2002), using specific knowledge about TMBBs and their environment (Freeman & Wimley, 2010; Hayat et al., 2016; Wimley, 2002) and by identification of sequence motifs (Berven et al., 2004; Gromiha, 2005). In addition to classification within the fold, several computational methodologies also predict more specific features of TMBB topology such as strand number and which parts of the protein localize to the periplasm, to the outside of the cell in the extracellular milieu, or are embedded in the membrane. (Hayat & Elofsson, 2012; Hayat et al., 2016; Tian et al., 2018; Tsirigos et al., 2016; Waldispuhl et al., 2006; Waldispuhl et al., 2008). Also, homology has been explored and combined with many of these previous methods to augment available data and make explicit use of evolutionary cues in order to improve TMBB classification and prediction (Bigelow & Rost, 2006; Garrow et al., 2005; Remmert et al., 2009). Despite the scarcity of large datasets of TMBB data, machine learning methods were also employed to the tasks of TMBB topology prediction and fold classification, including feed-forward neural networks and SVMs (Natt et al., 2004), recurrent neural networks (Randall et al., 2008), hidden Markov models (Bagos et al., 2004; Martelli et al., 2002), extreme learning machines (Savojardo et al., 2011) and ensemble approaches (Kazemian & Grimaldi, 2020; Randall et al., 2008). In two instances the algorithms were applied to establish large-scale redundant databases of TMBBs or to perform proteome-wide search for new TMBBs. The TMBB-DB (Freeman & Wimley, 2012) contains 1,881,712 sequences collected from 600 different bacterial proteomes with sequences ranked by their likelihood of encoding a TMBB. The OMPdb (Tsirigos et al., 2011) contains 1,198,558 (as of Jan 2021) protein sequences predicted to be TMBBs by the PRED-TMBB2 algorithm.

Many of these previous methods were dependent on the availability of experimentally solved structures, which is very scarce for β-barrels. Methods that used sequences were also trained on a comparatively small set of putative TMBB sequences from which sequence features or extracted. Overall these methods used protein information in a limited sense. Deep learning models trained on a large number of protein families can predict residue contacts from co-evolutionary data. A probabilistic map of protein contacts is a rich and readily interpretable source of structural information and in this study we show that it can provide a simple avenue for accurate discrimination of the barrel fold.

Evolutionary contact maps are a 2D representation that two residues in a protein sequence are in contact and may be used for predicting protein structure or for refining homology models. In our work we use contact information to discriminate between the barrel fold and other protein folds. Co-evolution information is obtained from analyzing covariation of residues in an alignment of protein sequences that are homologous to a protein of interest. Strong statistical co-variation of two or more residues indicates a possible co-evolutionary event, where one set of mutations is compensated by another set of mutations. A co-evolutionary event can be interpreted as evidence that the residues are in proximity, from which contact probabilities can be derived, if we assume that proximity between compensatory mutations indicates contact. While some models derive contact information from only second order approximations of co-varying residues, extracting higher order interactions predicts contacts with higher accuracy (Kamisetty et al., 2013). In this work we use contact maps predicted using a deep learning model trained on alignments of multiple protein families. The method uses higher order patterns of covariation, producing a better representation of the sequence-to-contacts mapping and better accuracy in contact prediction even for shallow alignments. (Wang et al., 2017) Although contact information can be extracted from protein structure models, generated for example, by AlphaFold (Jumper et al., 2021), information from contact maps is not only easier to extract, but contains probabilities of co-variation for all residues, and not only limited information on residue proximity. Therefore contact maps allow for a more succinct and complete representation of variability in patterns of residue co-variation.

Here we present a new methodology called IsItABarrel that uses co-evolution information and extracts discriminative features from contact maps for the identification of TMBBs from sequences. We

apply IsItABarrel to identify almost two million TMBB proteins from 21,756 bacterial organisms. In addition we used IsItABarrel to search 2,959 bacterial proteomes for TMBBs and provide updated estimates of the TMBB content in these organisms. We also show how the IsItABarrel approach is useful to classify large sets of sequences by similarity and transfer of prediction labels. We also analyze general features of our new, large TMBB dataset.

# Results

*Algorithm for TMBB Prediction*

Up-down β-barrels such as those found in the outer membrane have contact maps that are easily distinguished from other proteins (Figs. 1A and 1B). The two features that nearly definitively identify TMBBs from contact maps are strand-strand interactions (Fig. 1C) and closing contacts (Fig. 1D). Our algorithm, called IsItABarrel, was developed in two steps: 1) creation of the initial algorithm and 2) adaptation for maps that the algorithm did not recognize but could be visually identified as barrels from contact maps. We ultimately found that our best algorithm identified TMBBs by either the existence of a closing contact or the existence of at least 4 β-strand contacts of 16-25 residues each (Fig. 1E and S1). We validated IsItABarrel on a set of 1,121 proteins with solved crystal structures, 121 of which are TMBBs and 1000 are non-TMBBs, yielding 95.88% balanced accuracy with 8 instances of false positives and 9 instances of false negatives (Table 1, top row). In addition, we validated IsItABarrel on a dataset containing only membrane proteins (identified by SignalP 5.0 (Almagro Armenteros et al., 2019)), and measure a balanced accuracy of 94.91%.

| | **Dataset** | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | NONTMBB1K (121 TMBBs / 1000 non-TMBBs) | | | | | MP1929 (121 TMBBs / 1929 non-TMBBs) | | | | |
| **Algorithm** | **PPV** | **TP** | **TN** | **FP** | **FN** | **PPV** | **TP** | **TN** | **FP** | **FN** |
| IsItABarrel | 0.933 | 112 | 992 | 8 | 9 | 0.679 | 112 | 1876 | 53 | 9 |
| BOMP | 0.841 | 111 | 979 | 21 | 10 | 0.493 | 111 | 1815 | 114 | 10 |
| FW-BB Analysis | 0.617 | 121 | 925 | 75 | 0 | 0.259 | 121 | 1582 | 347 | 0 |
| PRED-TMBB2 | 0.688 | 117 | 947 | 53 | 4 | 0.398 | 117 | 1752 | 177 | 4 |

**Table 1. Comparison of predictors on validation sets.** *The column on the left labeled NONTMBB1K includes 121 non-homologous TMBBs as positive examples and 1,000 non-homologous non-TMBBs as the negative set. The column on the right, MP1929, includes the same set of 121 TMBBs and a different set with only non-TMBB membrane proteins, identified with SignalP5.0. IsItABarrel (top row) provides the lowest number of false positives and intermediate performance for true positives. TP – true positives; TN – true negatives; FP – false positives; FN – false negatives. Positive Predictive Value (PPV), or precision, calculated as the ratio of true positives to all sequences predicted positive, PPV=TP/(TP+FP).*

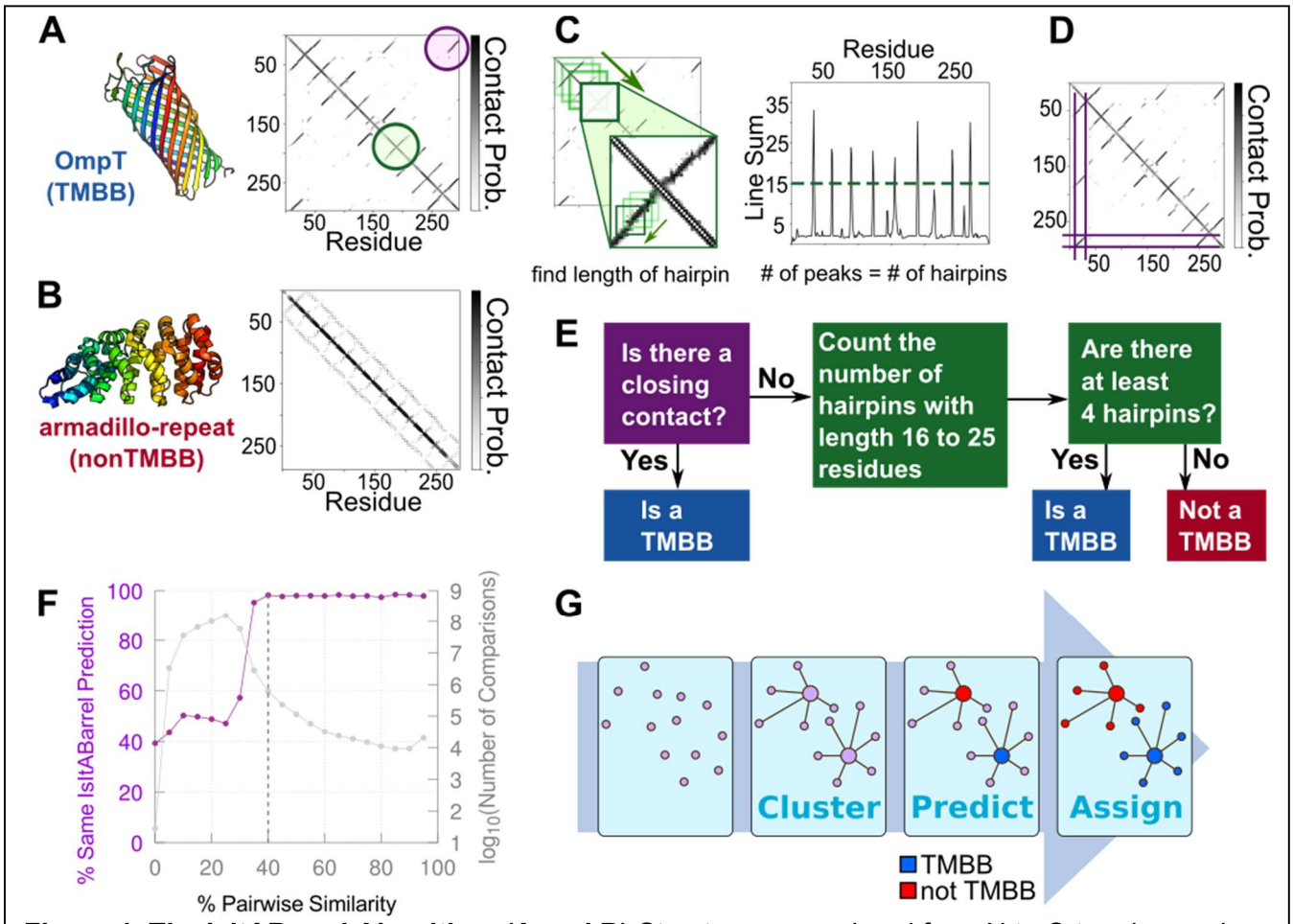



**Figure 1. The IsItABarrel Algorithm. (A and B)** Structures are colored from N to C-terminus using blue to red rainbow colors. Probability of contact is indicated by pixel grey value, with darker pixel indicating stronger contact between residues in *x* and *y* axes. **(A)** OmpT (PDB ID: 1I78) (left) and corresponding contact map predicted by RaptorX-Contact (right). The map shows typical features of β-barrel proteins, anti-parallel contacts between neighboring strands (green circle) and the closing contact between the first and last strand of the barrel (purple circle). **(B)** Structure of the non-TMBB armadillo-repeat protein (PDB ID: 4V3O) (left) and corresponding contact map predicted by RaptorX-Contact (right) without the typical features found in maps of beta-barrels.
**(C)** Strand-strand interaction (β-hairpin) detection. Map is scanned along the main diagonal in search of possible β-hairpins. Detected β-hairpins are characterized in terms of the length of their hydrogen bonding network (inset). The sum of the contact probabilities along a successfully detected β-hairpin is compared against the threshold for inclusion in the final TMBB classification step. **(D)** Closing contact detection: a closing contacts line is detected that aligns with the first and last strands and **(C)** a minimum number of hairpins is detected. **(E)** Classification by IsItABarrel is a two-step process requiring either a closing contact or four hairpins. **(F)** Sequences sharing more than 40% identity tend to have the same IsItABarrel classification. Percentage of pairs that share the same IsItABarrel prediction (y-axis) for each bin of percent similarity (x-axis) (purple), total number of comparisons ($\log_{10}$) at each similarity level (gray). Dashed line at 40% pairwise sequence similarity. **(G)** Steps of the algorithm used to assign IsItABarrel predictions by similarity. A large set of sequences is clustered by sequence similarity. IsItABarrel predictions are made for the cluster representatives. Cluster members are assigned the same prediction as their cluster representative.

*Prediction of a Large Data Set with IsItABarrel: OMPdb*

Map creation is a computationally demanding process. To increase the prediction scale, we determined the sequence similarity level that maintained consistent IsItABarrel classification. We determined that new maps do not need to be made for proteins with 40% or greater pairwise sequence similarity as homologs above that threshold have above a 95% similar IsItABarrel prediction (Fig. 1E), a success rate similar to the balanced accuracy of our algorithm overall. We determined this metric using the TMBB29183 dataset (Table S1). We evaluated 425,809,153 pairwise sequence comparisons on the set of 29,183 proteins and sequence similarity between them ranging from 4% to 99.9%.

*Expanding the dataset*

By culling previous datasets we created a database of 514,728 bacterial TMBB proteins. However, previous sets were created by finding homologs of known (structurally resolved) TMBBs. Since the IsItABarrel classifier does not rely on homology to structurally resolved barrels, we undertook a more thorough search of genetic space to uncover 43,822 previously unannotated bacterial TMBBs. We considered 2,959 NCBI prokaryote reference and representative organisms. These proteomes span 35 phyla and 78 classes (Supplemental Info S3 and S4). Many of these organisms were not diderms and did not have any TMBBs. We found consistent representation of TMBBs (i.e. TMBBs in 10 or more organisms) in 20 classes. Of those classes 6 of them were consistent among phyla (Bacteroidetes and Firmicutes), such that we could represent them as 2 phyla and 14 classes. TMBB-containing organismal categories included those that are known to include monoderms (Actinomycetia and Firmicutes). Other classes, such as Thermotogae and Deinococci, are diderms but usually contain atypical membrane compositions or thicker peptidoglycan layers which set them apart from other diderms. We searched those TMBB-containing proteomes for possible new TMBBs by extracting all membrane proteins from the largest organism (in terms of proteome size) in each class/phylum and removing all TMBB sequences that we had already found in the previous search. We predicted localization to the membrane using SignalP5.0 and kept in our set only the proteins that were predicted to have a signal peptide. We then predicted all of the remaining sequences of putative membrane proteins with IsItABarrel. In this search a total of 43,822 new protein sequences were found to be potential TMBBs, producing a 71.3% increase over what we had previously found (Fig. 2A and B). Similarly, we performed a new search procedure starting with the non-redundant protein database from NCBI, from which we selected a subset of putative bacterial membrane proteins and used IsItABarrel to identify all TMBB sequences. This new search resulted in our final set of predicted TMBBs we call IsItABarrelDB, containing almost 2 million bacterial TMBB sequences, out of the total set of known proteins (~230 million sequences).

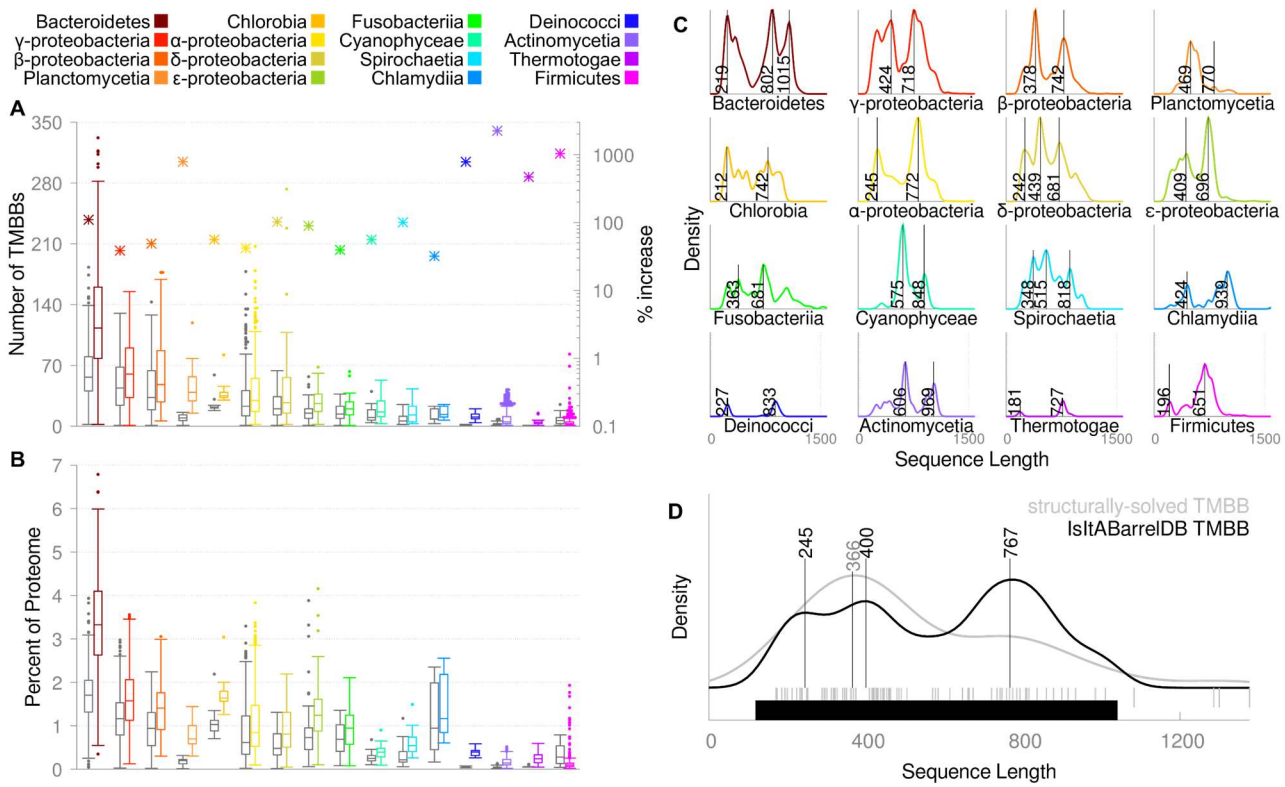



**Figure 2. TMBB Content in Bacterial Proteomes. (A and B)** Colored by class/phylum, categories are sorted by decreasing median of number of TMBBs. **(A)** Distribution of number of TMBBs in 2,959 reference and representative bacterial proteomes. TMBBs predicted with IsItABarrel that were within other datasets (gray), and updated numbers after a full search with IsItABarrel (color)**.** **(A)** Distribution of number of TMBB proteins per organism by class/phylum (boxplot), percent increase in number of TMBBs from the initial search (stars using $\log_{10}$ scale at right). **(B)** Distribution of percentage of proteome predicted to encode TMBB proteins (same organisms as in A). **(C)** Kernel density estimates for the lengths of proteins predicted to be TMBBs from the organisms in each organismal category. The numbers indicate the protein sequence lengths where the modes occur. **(D)** Kernel density estimate of the distribution of sequence lengths for structurally-solved (gray) and IsItABarrel predicted (black) TMBBs, rug plot showing instances below. The three numbers in black (245, 400, 767) indicate the protein sequence lengths where the modes occur for IsItABarrelDB, and the grey number (366) indicates the mode for the structurally-solved set.

*Distribution of TMBBs in Representative Organisms*

A highly accurate dataset of TMBBs allows for the quantification of the TMBB distribution among different organisms (Fig. 2 A and B). We find that Bacteroidetes tend to have the most TMBBs with an average of 127.5 per organism (average of 3.4% of the proteome). Thermotogae, which is known to have an atypical outer membrane (Megrian et al., 2020), has the fewest with an average of 5.28 per organism (average of 0.28% of the proteome).

*Lengths of Predicted TMBBs*

We assessed the length of the proteins by organism and computed kernel density estimates for the distribution on lengths of the TMBBs proteins (Fig. 2C). While some classes have only a small number of predicted TMBBs, it is of note the multimodal distribution of most of these classes, indicating that TMBBs usually cluster around more than one average length. Many of the modes are approximately double the previous mode, for example in β-proteobacteria modes are at sequence

lengths 378 and 742, in Firmicutes 424 and 939. Such doubling is consistent with previously described barrel domain duplications that yield larger barrels. (Franklin, Nepomnyachyi, et al., 2018; Remmert et al., 2010)

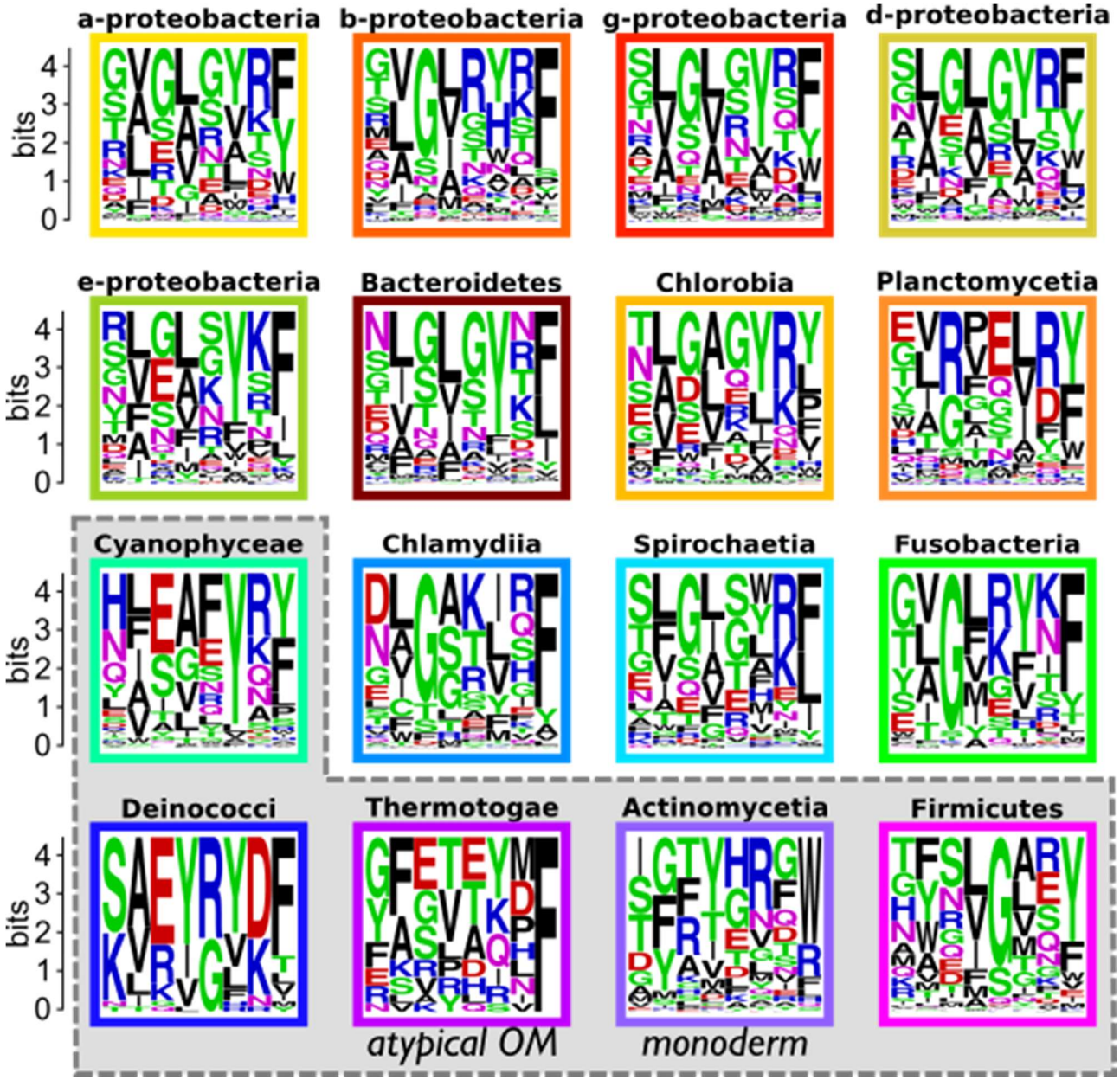
Overall we find the length distribution of TMBBs to differ from currently structurally resolved TMBBs (Fig. 2D). When comparing the IsItABarrelDB (Table S1) to 100 non-homologous structurally-resolved TMBBs we find that the most frequent size of TMBBs is 767 amino acids with another two modes for smaller barrels at 228 and 400 residues. The main peak of structurally solved proteins is at 366. Thus, while the smaller sized TMBBs are well represented in the structurally resolved set (peak of distribution at 374 amino acids) the larger barrels are not. In addition, though the structurally-resolved set has one mode, three modes can be disambiguated for IsItABarrelDB. Notably, these modes are all roughly double each other, suggestive of two possible duplications, though only the 8-stranded barrel to 16-stranded barrel duplication has been seen previously. (Franklin, Nepomnyachyi, et al., 2018)

*β-signal by organism*

The C-termini of TMBBs are known to have a sequence motif (Struyvé et al., 1991) that assists membrane insertion by interacting with the outer membrane translocation protein BAM (Doyle & Bernstein, 2019; Gessmann et al., 2014; Wang et al., 2021) called a β-signal.(Robert et al., 2006) We analyzed the predicted TMBB sequences for the presence of the β-signal in the different organismal categories using the MEME (Bailey et al., 2015) sequence-motif detector (Fig. 3). MEME searched the final quarter of each sequence so that the signal could be detected even if it is in a non-terminal strand.(Hagan et al., 2015) Though there are differences among the sequence motifs, many features are consistent. Eleven of fifteen classes/phyla most commonly have a phenyl alanine at the C-terminal position (-1) of the motif. The other four have a tyrosine. Even more likely (13 out of 15 classes/phyla) most commonly have a charged amino acid at the penultimate position (-2). The other even positions (-4, -6, and -8) tend to have small amino acids like glycine in Gracilicutes, and charged amino acids more likely in Terrabacteria. Tyrosine is the most common amino acid in the -3 position for 11 of 15 classes/phyla. The overall β-signal sequence is $H_yGH_yGY+F$ (where $H_y$ is hydrophobic amino acids L or V, and + is positively charged amino acids K or R).

However, the Actinomycetia class has none of these features and has a β-signal more consistent with the previously identified LamB-like β-signal. (Wang et al., 2021)

**Figure 3. Predicted C-terminal Motif in Bacterial Proteomes.** In relation to the most ancient recognized bacterial divide, the five classes on the bottom left are Terrabacteria (Thermotogae, Deinococci, Actinomycetia, Cyanophyceae and Firmicutes). The other 11 classes are the Gracilicutes. We find more consensus among Gracilicutes, and more variability among Terrabacteria.




*Comparison of IsItABarrel Classification with Other Predictors*

We compared IsItABarrel with other predictors used to create previous datasets: the Freeman-Wimley algorithm that was used to create the TMBB-DB database (Freeman & Wimley, 2010) (Fig. 4A) and PRED-TMBB2 (Tsirigos et al., 2016) that was used to create the OMPdb (Tsirigos et al., 2011) (Fig. 4B). We also compared IsItABarrel with the BOMP predictor (Berven et al., 2004) which

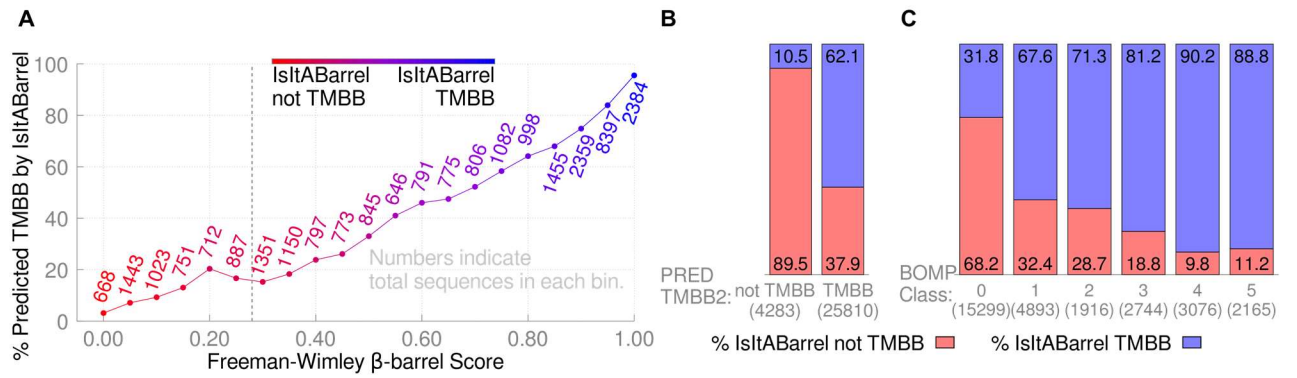



**Figure 4. Comparison of IsItABarrel with Three Other TMBB predictors. (A)** For the Freeman-Wimley algorithm a β-barrel score is computed and the sequence is classified as a TMBB for a score above 0.28. β-barrel scores for a set of 29,183 sequences were grouped in bins of width 0.05. In each bin IsItABarrel detects a percentage of barrels, which decreases with lower Freeman-Wimley score, showing a correlation between the two algorithms. **(B)** Two-class prediction by PRED-TMBB2. There is significant overlap for non-TMBB predictions, while IsItABarrel finds 37.9% of false positives. **(C)** The BOMP predictions for TMBBs also show significant overlap with IsItABarrel. For non-TMBBs (BOMP class 0), IsItABarrel identifies many false negatives (31.8% in class 0).

improved the original Freeman-Wimley algorithm by adding a β-signal motif predictor and a clustering procedure using isoleucine and asparagine abundances to reduce number of false positives. We classified two non-redundant datasets OMPDB-NR and the TMBB29183 (Table S1) using each of the algorithms. Though it is not possible to derive conclusions about the absolute prediction accuracy of each algorithm since we do not have ground truth labels for the datasets, we find that all three methods correlate with IsItABarrel. For sets with ground truth labels (Table 1), the other algorithms all substantially over-predict barrels.

The Freeman-Wimley algorithm uses a score based on the statistics of amino acid abundances and uses knowledge about lengths of β-strands and β-hairpins in outer membrane protein sequences. The predictions are combined into an overall β-barrel-score that gives the likelihood a sequence will assume the transmembrane barrel fold. According to the IsItABarrel classifier, the Freeman-Wimley proteins at the Freeman-Wimley threshold of 0.28 contain less than 20% barrels.

The PRED-TMBB2 is a topology predictor and detects periplasmic loop regions, transmembrane strands and extracellular loops using a three-part hidden Markov model. It performs TMBB classification by applying a logistic regression model using the prediction probability output from the HMM along with seven other protein features. According to IsItABarrel only 62.1% of PRED-TMBB2 barrel predictions are accurate.

We find that while the other two classifiers over-predict TMBBs, BOMP under-predicts barrels. BOMP predicts whether a protein sequence likely encodes a TMBB by using a modified version of the original Freeman-Wimley β-barrel-score and prediction of the β-signal, the highly conserved motif at the C-terminus of outer membrane β-barrels. BOMP classifies a putative TMBB sequence in one of six classes (Fig. 4C), of which classes four and five indicate a strong TMBB prediction and class zero indicates a negative prediction (not a TMBB). There is strong correlation in positive BOMP classes four and five. However, 31.8% of non-TMBB predictions (class zero) by BOMP are considered to be false negatives by IsItABarrel.

# Discussion

One of the difficulties of generalizing about the family of outer membrane β-barrels is the small number of solved crystal structures for proteins with low sequence similarity. (Jackups & Liang, 2005) With almost two million sequences, our high-accuracy IsItABarrelDB database contributes a large set

of sequences for analyzing TMBBs. The IsItABarrel algorithm will also identify more TMBBs for genomes sequenced in the future.

In addition to IsItABarrelDB being at least 2 times larger size than any previous TMBB databases the algorithm used to create it selects between two and nine times fewer false positives (Table 1). So, the database is both larger and more reliable. However, the fact that IsItABarrel is more selective also results in a higher false negative rate. IsItABarrel incorrectly predicts 8% of TMBBs as non-TMBBs. We find that the algorithms that have fewer false negatives have more false positives. For example the Freeman-Wimley algorithm has no false positives but has nine times as many false negatives as IsItABarrel. Unlike IsItABarrel, that uses contact maps generated by homology, the other three predictors use independent discriminatory features such as hydrophobicity alternation in strand residues, average length of strands, detection of β-signal motif, differences in amino acid composition across the outer membrane and topology modeling with hidden Markov models. We speculate that the contact maps we use include a combination of all these information into a richer set of features that allows IsItABarrel to improve on prediction accuracy over methods that use only certain individual features for classification. Thus though the IsItABarrel 8% false negative rate shows there is room for future improvement, because the low false negative rate and high false positive rate are correlated for other algorithms and because the features used in other algorithms have implicit representation in IsItABarrel, it is unclear that implementing features from other algorithms would help our algorithm reduce false negatives.

Even though the contact maps from RaptorX are a key element of our algorithm, the main drivers of prediction accuracy are our features that describe strand-strand interactions and closing contacts. Therefore other contact map generation methods could have been employed, such as those from GREMLIN (Kamisetty et al., 2013) or AlphaFold (Jumper et al., 2021), with some computational complexity penalty.

Two sources of error for our algorithm are the algorithm itself which has a balanced accuracy of 95.88% and the labeling assignments by sequence similarity > 40% which has accuracy of approximately 98% according to Fig. 1F.

We caution the reader that appropriate assessment of the impact of this inaccuracy in downstream tasks should be performed. In IsItABarrelDB we identify sequences that were predicted directly using IsItABarrel from those that had their predictions assigned by similarity. Future analysis of the cases that get different prediction labels when predicted by IsItABarrel and when being assigned by sequence similarity may reveal avenues for further improvement.

When identifying common lengths for TMBBs by class/phyla or for all organisms (Fig. 2C and D), we find multiple modes. Many of the larger modes are approximately double the size of the smaller modes possibly indicating domain duplication. We observe that α-proteobacteria, ε-proteobacteria and Firmicutes have a preference for larger barrel sizes in their bimodal distribution, while the Terrabacteria phyla Cyanophyceae and Actinomycetia have a preference for the shorter barrels. Some distributions in Fig. 2C are unimodal or not clearly bimodal, including Bacteroidetes, Planctomycetia, Spirochaetia, δ-proteobacteria and Chlorobia. It is possible that these classes have more diverse set of extracellular loop structures and plugs that do not directly affect the size of the barrel, but rather the length of the sequence, and probably its function. From the comparison in Fig. 2D it is clear that there is a large number of pores with length around 800 residues that are underrepresented in the PDB. Also, our analysis did not include length above 1040 residues, and we are certainly missing a large population of TMBBs of very large size. We did however include lengths down to 120 residues, and this range of barrels could also be further explored for solving new barrels structures.

We find that previous reports of TMBB content being 2 – 3% of diderm bacterial genomes (Wimley, 2003) are accurate for the most frequently studied of γ-proteobacteria genomes. However, over a larger range of organisms there is higher variance, from almost 7% to less than 1% with an overall mean of 1.54% and two modes at 0.4% and 1.4%. The category with the largest median genomic percentage of TMBBs is Bacteroidetes at a median of 3.34% per genome.

Firmicutes, the second largest phylum found in the human microbiome after Bacteroidetes, has only a small number of organisms with TMBBs (which is likely due to the fact that the organisms in this phylum are low-GC Gram-positives with only some organisms being diderms with outer membranes).

A number of organisms in the Actinomycetia class report a small number of TMBBs, of which some may be an artifact of our method since these organisms are high-GC Gram-positive. However, we find a few organisms with less than 5 predicted TMBBs (most with one predicted TMBB) in 8 other Gram-positive classes and 10 Archaea classes out of the total of 78 classes of prokaryotic organisms in our set but this is most certainly an artifact or our method of searching by sequence similarity.

In our analysis we did not find any TMBBs for several classes of organisms, including Halobacteria and Methanogenes. Even though organisms of Archaeal origin have cell envelopes that may resemble the double membrane of Gram-negative diderms (Klingl et al., 2019), it appears that either the proteins in these environments assume a structure other than the β-barrel (Burghardt et al., 2007), or that their structure is sufficiently different (longer strands for example) that they are not captured by IsItABarrel.

In our organismal search we found 2,283 proteins that included domains labeled as DUF (domain of unknown function), which represents 5.21% of the set of new TMBB proteins. A total of 52 different folds categorized as DUFs are found among those 2,283 proteins (Supplemental Info S5). Some of these domains include proteins with similarity to other barrels (DUF481, DUF3187) while others are less similar to whole barrels, and may be a globular domain attached to a barrel protein and exposed on the cellular surface such as domains for cell adhesion (DUF1566). Interestingly, DUF3108 domains are found in proteins with different characteristics: a protein that likely folds as a barrel porin, another one that is very likely not a membrane β-barrel, and a third example that appears to be a deformed β-sheet, as reported for the associated protein family PF11306 in the PFAM database (PFAM accessed Mar. 2022).

We found these “DUF proteins” in all but Chlorobia of our 16 classes/phyla of organisms (Supplemental Info S6). The DUF proteins are unevenly distributed across the organismal categories (Supplemental Info S6). There is a general correlation between organisms with more known TMBBs and more DUF proteins labeled TMBB by IsItABarrel, although Actinomycetia and Firmicutes seem to be outliers in that generalization. Some DUFs are very class- or phylum-specific while others spread more broadly across phyla. The domains DUF1080, DUF2147, DUF3458, DUF2494 and DUF3828 are found only in Bacteroidetes. Only Firmicutes contain DUF370, and DUF1553 was only found in Planctomycetia. DUF1207 was found only once, in Chlamydiia. Some other DUFs are less phylum-specific, for example DUF4139, which is found in Actinomycetia, Bacteroidetes, all Proteobacteria, Planctomycetia and Spirochaetia. While the previous discussion concerned only the proteins we found in the 2,959 organisms, we also found 76,893 predicted TMBBs with domains of unknown function in the larger IsItABarrelDB dataset, our final dataset that contains a total of almost 2 million sequences. These cover 284 DUFs. We have listed these in Supplemental Info S7.

DUF proteins may have roles in a number of useful cellular processes. These processes may be important for the discovery of new antimicrobial targets that block functions associated with interactions between the cell and the extracellular environment (nutrient utilization, efflux, sensing and control). We anticipate that the discovery of new types of outer membrane proteins will propel future biologically and medicinally-important research.

Finally, by looking at so many TMBBs we find a slightly more specific β-signal than previously reported. The vast majority of classes have the motif $H_yGH_yGY+F$. Variations in this motif likely reflect the different pathways of membrane insertion. In Terrabacteria organisms and Planctomycetia (but not Firmicutes) the glycine positions of the motif can be replaced by charged amino acids, while these glycines are more conserved in Bacteroidetes and Proteobacteria. The two Terrabacteria groups, Deinococci and Actinomycetia, are the only groups with a tyrosine replacing the glycine in position 5. Another noticeable variation is the replacement of the glycine at position 5 by smaller amino acids (alanine, valine) in the three classes Chlorobia, Cyanophyceae and Chlamydiia. Position 5 is clearly dominated by glycine in all proteobacteria, Spirochaetia, Fusobacteria and Firmicutes. Possibly, our computational approach detected a consensus motif in some categories stronger that in others due to effects of sample size and sequence redundancy, which may have impacted detection of the true motif

in some of the phyla with more variation. Nonetheless, understanding the plasticity of this motif across different organisms is important to understand its function *in vivo*, and will require more experimental evidence of how well mutations in the several positions of the β-signal motif are tolerated across organismal classes.

# Conclusion

We have developed an algorithm for TMBB prediction and created a new database of curated TMBB sequences that is 40% more accurate and at least 2 times larger than the ones created with previous algorithms. 2.5% of our dataset is proteins that are not known to be homologous to known β-barrels and are annotated as domains of unknown function in the bacterial NCBI dataset. Our algorithm uses contact maps generated from co-evolutionary information of homologous sequences to derive two features that classify protein sequences with high accuracy between TMBBs and non-TMBBs. Bacterial organisms vary tremendously with more some having more than 6% of the genome devoted to encoding TMBBs and some with less than 1%. Although there is variation among different bacterial organisms, the β-signal consensus is L/V-G-L/V-G-Y-R/K-F. We offer our algorithm and database as a useful tool for researchers working with bacterial TMBBs who need curated set of these proteins at http://isitabarrel.ku.edu and https://gitlab.ku.edu/sluskylab/isitabarrel.

# Methods

*Contact Map Generation and Knowledge-Based Features*

To classify sequences as TMBBs or non-TMBBs we combined RaptorX-Contact, a deep learning method that generates contact maps from co-evolution data (Ma et al., 2015) with the development of a rule-based algorithm for feature extraction and classification. Contact maps predicted by RaptorX are given by a symmetric matrix of numbers in the range [0,1], such that the upper triangular part above the main diagonal is identical to the lower triangle below the diagonal.

RaptorX-Contact computes contact predictions for all possible pairs of residues in a sequence. However, only a fraction of the points on the map represents strong and confident predictions. To reduce the impact of weak predictions on our algorithm and improve the robustness of classification we threshold the contact map, retaining only a predefined number of predictions proportional to the length of the sequence. For a sequence of length $L$ residues, in our tests with a small set of proteins, we observed that preserving the top strongest $2L*\log_2(L)$ contacts improved the robustness of the classification without incurring significant loss of information that would preclude correct classification.

*The IsItABarrel Algorithm*

Our algorithm was based on the following principles: Because the β-strands in TMBBs are ordered contiguously (strand 1 is next to strand 2, strand 2 is next to strand 3, etc.) TMBBs have consistent and distinctive contact map features. These features include lines emanating perpendicularly from the diagonal indicating strand-strand interactions and a line in the corner indicating the closing contacts between the first strand and the last strand (Fig. 1A). Non-TMBB folds rarely have these features (Fig. 1B). Our IsItABarrel predictor extracts features from the generated contact maps that quantify the presence of a closing contact between the first and last strand and anti-parallel β-strand contacts. These two features are designated *strand-strand contact* and *closing contact*.

Our *strand-strand contact* rule detects the number of lines perpendicular to the main diagonal as a proxy for the number of pairs of antiparallel β-strands interacting (Fig. 1C). We consider 52x52 boxes going through the diagonal of the contact map and calculate the sum of the elements lying on the box diagonal that is perpendicular to the main map diagonal. To capture perpendicular lines that are near the corner of the map, we pad the map with zeros on all four sides before scanning the diagonal. We can graph the sums of the perpendicular lines by the residue number at the center of their scanning boxes (Fig. 1C). For any position where the sum is above a threshold of 18, the area to

the bottom left of the main diagonal is searched to locate the position of the perpendicular line that is closest to the main diagonal (perpendicular lines are often disconnected from the main diagonal). Then, the bottom left of the starting position is searched and the number of residues between the starting position and the next strand–strand interaction position is summed. This sum is the length of the strand-strand contact. Additionally, to accurately compare β-hairpin lengths (two neighboring β-strands with a loop region connecting them is commonly called a "β-hairpin"), we calculate the length of the longest part of each β-hairpin by calculating the lengths of several lines along the hairpin and choosing the maximum value. Contact lengths between 16 and 25 residues are defined as strand-strand contacts. All numerical parameters and thresholds used were identified through grid search (Supplemental Fig. S1).

Our closing contact identification algorithm works in two steps. First, it detects all β-hairpins as previously described. Second, it iterates over the detected strands and then uses their location to search for the slanting line of closing contacts aligned to a pair of the first and last strands (Fig. 1D).

Once a potential closing contact is located, the position is saved and a 10x10 box is created at the right upper corner of the map. If there is a full diagonal line from the left bottom corner to the right upper corner of the 10x10 box, the algorithm detects a closing contact. If not, a larger box with height 69 residues and width 7 residues is defined instead of the original 10x10 box. The particular parameters, such as box size (10x10) and scan region (69x7), were optimized through grid search using a validation set. If a full diagonal line is found when scanning from the left bottom corner to the right upper corner in any 10x10 box inside this larger outline, it is defined as a closing contact.

The strand-strand contact prediction and closing contact prediction are then combined (Fig. 1E). If a closing contact is predicted, then the map is labeled a TMBB. In the absence of a detectable closing contact, if at least four strand-strand contacts are identified then the map is also labeled a TMBB. Otherwise, the map is defined as a non-TMBB.

*Datasets*

Seven different datasets of protein sequences were used in this work. We summarize their features in Table S1. There is only one dataset of known TMBB proteins with solved structures that is used throughout. The reason is due to the small number of non-redundant TMBB with solved structures currently available in the Protein Data Bank (PDB). The NONTMBB1K and MP1929 datasets include only non-TMBB sequences, all with solved structures in PDB. The datasets TMBB29183 and OMPDB-NR were derived from the original databases (respectively the TMBB-DB and the OMPdb) and are non-redundant datasets we used to compare different predictors. Although there is a possible overlap with the TMBB121, this does not in any way impact our analysis of predictors. The IsItABarrelDB is the largest set of predicted TMBBs resulting from our analysis.

We created three datasets for training IsItABarrel—one of TMBBs (TMBB121), one dataset of non-TMBB (NONTMBB1K) with sequences that had solved structures in the PDB, and structurally resolved non-TMBBs that had a signal peptide (MP1929). For the TMBB set, sequences were obtained from a previously published (Franklin, Nepomnyachiy, et al., 2018) dataset of outer membrane proteins that was restricted to less than 60% similarity with CD-HIT (Huang et al., 2010). This set was smaller than all other sets because of the limited number of structurally resolved non-homologous TMBBs available in the protein data bank (PDB).

For the NONTMBB1K set, one thousand non-TMBB sequences were randomly selected from PDBSELECT100 (downloaded Nov. 18, 2017) (Griep & Hobohm, 2010). We removed all TMBBs by searching descriptions for any of the words "*outer, membrane, barrel, transport, channel, porin*" and manually removing any potential TMBBs (Supplemental Info S1). A balanced dataset containing 100 TMBB sequences and a subset of 100 non-TMBB sequences from NONTMBB1K was used to develop our IsItABarrel algorithm.

We desired to test IsItABarrel on proteins containing a signal peptide that are not TMBBs since we planned to cull datasets first by removing proteins without a signal sequence. We developed the MP1929 dataset by randomly selecting 1,929 non-TMBB sequences from the non-redundant PISCES set cullpdb_pc20_res3.0_R1 (Wang & Dunbrack, 2005). Selection continued until 1929 were found with the following four required characteristics: 1) a signal peptide by SignalP 5.0 (Almagro

Armenteros et al., 2019) 2) overall sequence length between 120 and 1040, 3) match with a sequence in Uniprot, 4) no description of a possible barrel (as described above for the PDBSelect set).

The IsItABarrel algorithm depends on various parameters and after developing the general method we set out to optimize these parameters to fine tune its accuracy (Supplemental Fig. S1). To create a training set, we combined the full TMBB121 set with the full NONTMBB1K set into an unbalanced dataset containing 1,121 sequences and optimized the parameters of the algorithm by grid search. In addition, to lower the IsItABarrel false negative rate, during parameter refinement 21 more TMBB sequences were selected based on visual inspection of contact maps that had TMBB signatures but were not being classified by IsItABarrel. Those 21 sequences were selected from 200 sequences that were identified as β-barrels by PRED-TMBB2 (Tsirigos et al., 2016), BOMP (Berven et al., 2004), and the Freeman-Wimley (Freeman & Wimley, 2010) predictors, though not by initial parametrization of IsItABarrel.

To create IsItABarrelDB, cluster representatives were obtained by clustering the initial set down to 30% similarity using MMSeqs2 (Hauser et al., 2016) and predicting signal sequences with SignalP5.0. Contact maps for the set of 931,269 cluster representatives were generated with RaptorX-Contact and predicted with IsItABarrel. Predictions from the cluster representatives were assigned to all other cluster members according to sequence similarity. In this set only sequences with length in the range of 120-1040 residues from bacterial organisms were included. The IsItABarrelDB dataset is the result of running IsItABarrel on the putative bacterial membrane proteins available in the large non-redundant protein set from NCBI (accessed March 2020).

*Distribution of TMBBs in Representative Organisms and Analysis of Sequences*

To compute the distribution of TMBB populations in prokaryotic organisms, we searched bacterial genome assemblies of 2,944 representative organisms and 15 reference organisms for TMBB homologs. Genomes were retrieved from the NCBI Assembly database (April 2021) using filters "Latest", "RefSeq", "Complete Genome" and "Exclude anomalous". Homologs were found between the 95% sequence-similarity-filtered OMPDB-NR set and the RefSeq proteomes using blastp from BLAST version 2.9.0+ (Altschul et al., 1997), E-value of 1.0, query coverage of 95%, and all other parameters at their default values. We removed any redundant hits and filtered the results of each search by retaining only hits with 90% or more positive matching residues. Different TMBBs matching the same protein sequence were counted only once.

To compute the distribution of TMBBs per organism we grouped the 2,959 organisms by taxonomic class with the exception of Bacteroidetes and Firmicutes which were grouped by phylum.

*Code availability*

All sequences datasets are available for download and contact maps may be obtained upon request. The original TMBB-DB and OMPdb may be obtained from the webpages of the original authors at http://beta-barrel.tulane.edu/downloads.html and http://www.ompdb.org respectively. Code for RaptorX is freely available for local installation from the original authors and can be downloaded at https://github.com/j3xugit/RaptorX-Contact. All parameters were left at their defaults, but differences in contact map generation may arise due to the differences in the underlying databases required by the software. The set of all sequences predicted with our IsItABarrel algorithm is available for search and download at our webapp available at http://isitabarrel.ku.edu. In this database the user can select a subset of sequences from our resulting TMBB dataset set by filtering the records using various available criteria, including the original Freeman-Wimley β-barrel score, the new IsItABarrel prediction and the SignalP 5.0 prediction. The database search can be refined further by selecting between the three different types of prokaryotic signal peptide pathways that SignalP 5.0 is able to distinguish: SP (the general secretory pathway with signal peptidase I), TAT (the twin-arginine translocation pathway with signal peptidase I, and LIPO (the translocation pathway for bacterial lipoproteins with cleavage by signal peptidase II). The user can also search by organism using partial matching of boolean expression (e. g. coli AND pestis) and by sequence length. Currently, the filtered results are returned in a table format and made available for download in a compressed file format that includes two individual files, a records file containing the returned records in TAB-separated format and a FASTA

file containing the amino acid sequences of the corresponding records with the non-redundant protein accession number as the FASTA header to use as a key between the two tables.

*Programming*
All steps of dataset creation were implemented with custom BASH shell and Python scripts. Contact map estimation using RaptorX was performed on the high-performance computer cluster of the University of Kansas.

## Acknowledgments

We thank Dr. Robert Unckless from the University of Kansas for help with our analysis of distribution of TMBBs in prokaryotes and Rachel Kolodny for important discussions about the algorithm. Financial support for this work came from DP2 GM128201 to Joanna Slusky and KU Startup funding.

## Supplemental Information

**Supplemental Info S1.** Uniprot (version Jan. 2021) IDs for TMBB sequences removed from the NONTMBB11603 set.

A0A084CH10, A0A0H3AME7, A0QR29, A5VZA8, A5W3Z9, A5W4F4, A6KZB0, A6L326, C4IN73, F2VN85, G3XD89, L0RVU0, O05442, O07624, O88093, P02929, P02930, P06129, P06971, P0A901, P0A910, P0A915, P0A921, P0A940, P0ADC1, P0ADE4, P0C2W0, P10384, P13794, P16869, P17315, P17811, P18895, P24305, P26466, P26949, P30130, P31243, P32977, P35077, P37001, P43261, P48632, P54170, P69434, P69856, P72412, P76045, P76115, Q1I8U1, Q48245, Q48391, Q5LGM0, Q60932, Q659I5, Q72JD8, Q76HJ9, Q7MXJ7, Q83SQ0, Q8A5H5, Q8AB42, Q8D0Z7, Q8ECH9, Q8GKS5, Q8GNN6, Q8ZPT3, Q8ZRK4, Q934G3, Q93K99, Q9AE79, Q9HVD1, Q9I202, Q9JZN9, Q9K0U9, Q9S3Q1, Q9X4B7

**Supplemental Info S2.** List of organisms that were manually curated for Class. Data obtained from the LPSN.dsmz.de database and from the WoRMS database.

| Taxonomic ID | Organism Name | Class |
|---|---|---|
| 1140 | *Synechococcus elongatus* | Cyanophyceae |
| 13035 | *Dactylococcopsis salina* | Cyanophyceae |
| 56110 | *Oscillatoria acuminata* | Cyanophyceae |
| 63737 | *Nostoc punctiforme* | Cyanophyceae |
| 65393 | *Gloeothece citriformis* | Cyanophyceae |
| 111780 | *Stanieria cyanosphaera* | Cyanophyceae |
| 118163 | *Pleurocaps sp PCC7327* | Cyanophyceae |
| 146919 | *Salinibacter ruber* | Sphingobacteria |
| 167539 | *Prochlorococcus marinus* | Cyanophyceae |
| 179408 | *Oscillatoria nigro-viridis* | Cyanophyceae |
| 197221 | *Thermosynechococcus elongatus* | Cyanophyceae |

| | | |
|---|---|---|
| 203124 | *Trichodesmium erythraeum* | Cyanophyceae |
| 251229 | *Crhooccidiopsis thermalis* | Cyanophyceae |
| 272123 | *Arabaena cylindrica* | Cyanophyceae |
| 292564 | *Cyanobium gracile* | Cyanophyceae |
| 329726 | *Acaryochloris marina* | Cyanophyceae |
| 375899 | *Trichormus variabilis* | Cyanophyceae |
| 395962 | *Rippkaea orientalis* | Cyanophyceae |
| 411966 | *Leptolyngbya boryana* | Cyanophyceae |
| 436308 | *Nitrosopumilus maritimus* | Nitrososphaeria |
| 446679 | *Nostoc sphaeroides* | Cyanophyceae |
| 449447 | *Microcystis aeruginosa* | Cyanophyceae |
| 497965 | *Gloethece verrucosa* | Cyanophyceae |
| 518766 | *Rhodothermus marinus* | Sphingobacteria |
| 522448 | *Euhalothere natronophila* | Cyanophyceae |
| 525904 | *Thermobaculum terrenum* | Thermomicrobia |
| 551115 | *Nostoc azollae* | Cyanophyceae |
| 755178 | *Cyanobacterium aponium* | Cyanophyceae |
| 1170562 | *Calothrix parietina* | Cyanophyceae |
| 1173022 | *Crinalium epipsammum* | Cyanophyceae |
| 1470066 | *Nitrosopumilus cobalaminigenes* | Nitrososphaeria |
| 1470067 | *Nitrosopumilus* | Nitrososphaeria |
| 1472755 | *Nostoc edophicum* | Cyanophyceae |
| 1580092 | *Nitrosopumilus adriaticus* | Nitrososphaeria |
| 1582439 | *Nitrosopumilus piranensis* | Nitrososphaeria |
| 1641165 | *Halomicronema Hongdechloris* | Cyanophyceae |
| 1871025 | *Ndongobacter massiliensis* | Clostridia |
| 1917166 | *Synechococcus lividus* | Cyanophyceae |
| 1959104 | *Nitrosopumilus oxyclinae* | Nitrososphaeria |
| 2005468 | *Thermosynechococcus vulcanus* | Cyanophyceae |
| 2038116 | *Nostoc flagelliforme* | Cyanophyceae |
| 2666332 | *Cylindrospermopsis curvispora* | Cyanophyceae |

**Supplemental Info S3.** List of the 78 classes covered by the 2,959 representative and reference organisms used in the search for TMBBs.

*Acidimicrobiia, Acidithiobacillia, Acidobacteriia, Actinomycetia, Alphaproteobacteria, Anaerolineae, Aquificae, Archaeoglobi, Bacilli, Bacteroidia ,Betaproteobacteria, Blastocatellia, Caldilineae, Caldisericia, Calditrichae, Chitinophagia, Chlamydiia, Chlorobia, Chloroflexia, Chrysiogenetes,*

*Clostridia, Conexivisphaeria, Coprothermobacteria, Coriobacteriia, Cyanophyceae, Cytophagia, Deferribacteres, Dehalococcoidia, Deinococci, Deltaproteobacteria, Dictyoglomia, Elusimicrobia, Endomicrobia, Epsilonproteobacteria, Erysipelotrichia, Fimbriimonadia, Flavobacteriia, Fusobacteriia, Gammaproteobacteria, Gemmatimonadetes, Gloeobacteria, Halobacteria, Hydrogenophilalia, Ignavibacteria, Kiritimatiellae, Ktedonobacteria, Limnochordia, Methanobacteria, Methanococci, Methanomicrobia, Methanopyri, Methylacidiphilae, Mollicutes, Negativicutes, Nitriliruptoria, Nitrososphaeria, Nitrospira, Oligoflexia, Opitutae, Phycisphaerae, Planctomycetia, Rubrobacteria, Saprospiria, Sphingobacteriia, Spirochaetia, Synergistia, Tepidiformia, Thermococci, Thermodesulfobacteria, Thermoleophilia, Thermomicrobia, Thermoplasmata, Thermoprotei, Thermotogae, Tissierellia, Verrucomicrobiae, Vicinamibacteria, Zetaproteobacteria*

**Supplemental Info S4.** List of phyla covered by the 2,959 representative and reference organisms used in the search for TMBBs.

*Acidobacteria, Actinobacteria, Aquificae, Armatimonadetes, Bacteroidetes, Caldiserica, Calditrichaeota, Candidatus Thermoplasmatota, Chlamydiae, Chlorobi, Chloroflexi, Chrysiogenetes, Coprothermobacterota, Crenarchaeota, Cyanobacteria, Deferribacteres, Deinococcus-Thermus, Dictyoglomi, Elusimicrobia, Euryarchaeota, Firmicutes, Fusobacteria, Gemmatimonadetes, Ignavibacteriae, Kiritimatiellaeota, Nitrospirae, Planctomycetes, Proteobacteria, Spirochaetes, Synergistetes, Tenericutes, Thaumarchaeota, Thermodesulfobacteria, Thermotogae, Verrucomicrobia*

**Supplemental Info S5.** List of 52 domains of unknown function (DUF) found in predicted TMBBs in the 2,959 representative and reference organisms.

DUF11, DUF370, DUF481, DUF756, DUF839, DUF916, DUF945, DUF1036, DUF1045, DUF1080, DUF1104, DUF1207, DUF1259, DUF1302, DUF1396, DUF1488, DUF1513, DUF1540, DUF1549, DUF1553, DUF1566, DUF1573, DUF1579, DUF2135, DUF2147, DUF2191, DUF2219, DUF2442, DUF2808, DUF3108, DUF3187, DUF3298, DUF3458, DUF3459, DUF3494, DUF3570, DUF3734, DUF3828, DUF3943, DUF3999, DUF4105, DUF4139, DUF4163, DUF4331, DUF4381, DUF4403, DUF4440, DUF4595, DUF4625, DUF4982, DUF5050, DUF5103

**Supplemental Info S6.** Table with predicted TMBBs with domains of unknown function (DUF) by organismal category in the 2,959 representative and reference organisms.

| Organismal category | How many proteins? | Which DUFs? |
|---|---|---|
| Actinomycetia | 159 | DUF1259, DUF1396, DUF2191, DUF3298, DUF3459, DUF4139, DUF4331, DUF4982, DUF756, DUF839, DUF916 |
| Alphaproteobacteria | 246 | DUF1036, DUF1045, DUF11, DUF1259, DUF1302, DUF1579, DUF3108, DUF3459, DUF4139, DUF4163, DUF4331, DUF4403, DUF4440, DUF481, DUF4982, DUF756, DUF839 |
| Bacteroidetes | 723 | DUF1080, DUF1573, DUF2135, DUF2147, DUF2219, DUF3458, DUF3459, DUF3494, DUF3570, DUF3828, DUF3943, DUF3999, DUF4139, DUF4163, DUF4331, DUF4403, DUF4440, DUF4595, DUF4625, DUF481, DUF4982, DUF5050, DUF5103, DUF756, DUF839 |
| Betaproteobacteria | 105 | DUF1259, DUF1302, DUF1579, DUF2135, DUF3570, DUF4105, DUF4139, DUF4331, DUF4982, DUF756, DUF839, DUF945 |
| Chlamydiia | 1 | DUF1207 |
| Chlorobia | 0 | - - - - - - |
| Cyanophyceae | 50 | DUF1579, DUF2808, DUF3459, DUF4331, DUF839, |

| | | |
|---|---|---|
| Deinococci | 46 | DUF11, DUF3459, DUF4403, DUF839, |
| Deltaproteobacteria | 42 | DUF1259, DUF1302, DUF1549, DUF1573, DUF1579, DUF2219, DUF3187, DUF3570, DUF4105, DUF4139, DUF4163, DUF4331, DUF4381, DUF4403, DUF839 |
| Epsilonproteobacteria | 19 | DUF1104, DUF3570, DUF4105, DUF4139, DUF839 |
| Firmicutes | 100 | DUF11, DUF1259, DUF1540, DUF3298, DUF370, DUF3999, DUF4163, DUF4982, DUF839 |
| Fusobacteriia | 6 | DUF3298, DUF3999, DUF4163 |
| Gammaproteobacteria | 352 | DUF1259, DUF1302, DUF1488, DUF1513, DUF1566, DUF1579, DUF2135, DUF2219, DUF2442, DUF3187, DUF3459, DUF3570, DUF3734, DUF3999, DUF4105, DUF4139, DUF4163, DUF4331, DUF4403, DUF4982, DUF756, DUF839, DUF945 |
| Planctomycetia | 338 | DUF11, DUF1549, DUF1553, DUF1573, DUF1579, DUF4139, DUF4331, DUF4982, DUF839 |
| Spirochaetia | 16 | DUF1566, DUF3187, DUF3298, DUF3943, DUF4105, DUF4139, DUF4163, DUF839 |
| Thermotogae | 5 | DUF4163 |

**Supplemental Info S7.** List of 284 domains of unknown function (DUF) found in predicted TMBBs in the IsItABarrelDB dataset.

DUF11, DUF58, DUF187, DUF255, DUF285, DUF303, DUF323, DUF333, DUF348, DUF362, DUF377, DUF378, DUF400, DUF427, DUF459, DUF461, DUF481, DUF547, DUF554, DUF560, DUF615, DUF732, DUF748, DUF756, DUF799, DUF839, DUF861, DUF882, DUF916, DUF928, DUF940, DUF945, DUF1028, DUF1036, DUF1058, DUF1080, DUF1083, DUF1192, DUF1194, DUF1202, DUF1207, DUF1223, DUF1239, DUF1249, DUF1302, DUF1311, DUF1326, DUF1396, DUF1410, DUF1416, DUF1486, DUF1501, DUF1513, DUF1549, DUF1551, DUF1553, DUF1554, DUF1555, DUF1559, DUF1565, DUF1566, DUF1573, DUF1579, DUF1587, DUF1592, DUF1593, DUF1597, DUF1641, DUF1682, DUF1735, DUF1768, DUF1775, DUF1778, DUF1793, DUF1794, DUF1800, DUF1812, DUF1906, DUF1928, DUF1939, DUF1961, DUF1983, DUF2006, DUF2012, DUF2066, DUF2125, DUF2135, DUF2141, DUF2207, DUF2219, DUF2233, DUF2268, DUF2285, DUF2300, DUF2320, DUF2334, DUF2341, DUF2380, DUF2394, DUF2436, DUF2490, DUF2531, DUF2712, DUF2715, DUF2804, DUF2807, DUF2808, DUF2817, DUF2846, DUF2860, DUF2889, DUF2911, DUF2961, DUF2968, DUF3011, DUF3034, DUF3068, DUF3078, DUF3108, DUF3124, DUF3138, DUF3179, DUF3187, DUF3224, DUF3225, DUF3233, DUF3237, DUF3238, DUF3244, DUF3280, DUF3298, DUF3308, DUF3316, DUF3324, DUF3352, DUF3373, DUF3374, DUF3383, DUF3387, DUF3427, DUF3440, DUF3442, DUF3455, DUF3459, DUF3471, DUF3472, DUF3494, DUF3500, DUF3520, DUF3568, DUF3570, DUF3573, DUF3575, DUF3604, DUF3607, DUF3616, DUF3659, DUF3769, DUF3794, DUF3808, DUF3823, DUF3826, DUF3836, DUF3857, DUF3859, DUF3872, DUF3943, DUF3996, DUF3999, DUF4012, DUF4038, DUF4073, DUF4082, DUF4091, DUF4101, DUF4104, DUF4105, DUF4114, DUF4115, DUF4124, DUF4129, DUF4139, DUF4140, DUF4163, DUF4165, DUF4168, DUF4175, DUF4179, DUF4185, DUF4198, DUF4199, DUF4200, DUF4214, DUF4215, DUF4232, DUF4249, DUF4252, DUF4255, DUF4265, DUF4271, DUF4292, DUF4331, DUF4340, DUF4344, DUF4347, DUF4360, DUF4366, DUF4369, DUF4381, DUF4382, DUF4384, DUF4394, DUF4397, DUF4403, DUF4412, DUF4419, DUF4421, DUF4424, DUF4430, DUF4431, DUF4434, DUF4440, DUF4465, DUF4469, DUF4474, DUF4480, DUF4493, DUF4595, DUF4623, DUF4625, DUF4652, DUF4785, DUF4832, DUF4838, DUF4855, DUF4874, DUF4876, DUF4880, DUF4886, DUF4892, DUF4906, DUF4929, DUF4959, DUF4961, DUF4962, DUF4964, DUF4965, DUF4968, DUF4974, DUF4976, DUF4979, DUF4981, DUF4982, DUF4988, DUF4989, DUF4990, DUF5005, DUF5006, DUF5008, DUF5010, DUF5011, DUF5016, DUF5018, DUF5019,

DUF5020, DUF5024, DUF5050, DUF5060, DUF5068, DUF5074, DUF5103, DUF5104, DUF5107, DUF5110, DUF5115, DUF5116, DUF5117, DUF5121, DUF5123, DUF5125, DUF5126, DUF5320, DUF5686, DUF5753

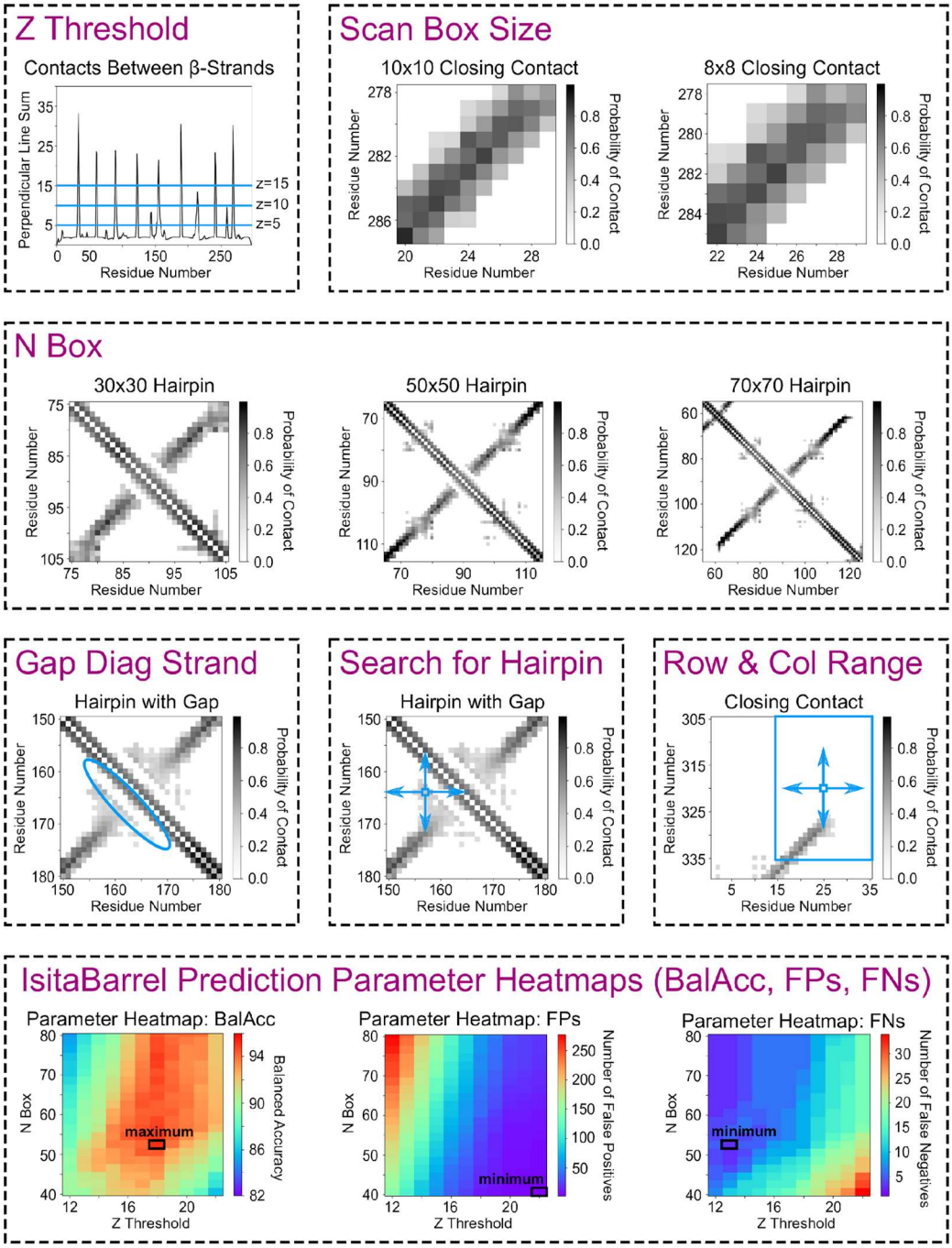



**Figure S1. The Parameters of the IsItABarrel Algorithm.** *Top 3 rows show some of the different parameters for the IsItABarrel algorithm: Z Threshold (threshold in which a peak is counted as a hairpin), Scan Box Size (size of box when scanning for closing contact), N Box (size of box when scanning for hairpins), Gap Diag Strand (maximum number of residues to go out past middle diagonal when searching for start of hairpin to account for possible gap between hairpin and middle diagonal), Search for Hairpin (maximum number of residues to go out from presumed start of hairpin in four directions to find a new start of hairpin that starts at the point where hairpin has the largest length), Row Range (width of box created around residue that was pinpointed for closing contact), and Column Range (length of box created around residue that was pinpointed for closing contact). Bottom row shows three parameter heatmaps created after running 451 simulations in which Z Threshold and N Box are changed simultaneously. For each simulation, the NONTMBB1K with TMBB121 dataset is used to test the balanced accuracy, number of false positives, and number of false negatives of the IsItABarrel algorithm for different combinations of Z Threshold and N Box. These three parameter heatmaps show the results of the simulations: the first heatmap shows the balanced accuracy of the IsItABarrel algorithm for different combinations of Z Threshold and N Box; the second heat map shows the number of false positives for different combinations of Z Threshold and N Box; the third heatmap shows the number of false positives for different combinations of Z Threshold and N Box.*

| Name | Description | Size | Usage |
|---|---|---|---|
| *TMBB121* | *Non-redundant TMBBs with solved structures.* | *121* | *Algorithm development*<br>*Validation*<br>*Comparison of predictors*<br>*Distribution of lengths* |
| *NONTMBB1K* | *Non homologous non-TMBBs with solved structures.* | *1,000* | *Parameter optimization* |
| *MP1929* | *Non-homologous membrane non-TMBBs with solved structures..* | *1,929* | *Validation* |
| *TMBB29183* | *Non-redundant set of sequences from the TMBB-DB.* | *29,183* | *Comparison of predictors* |
| *OMPDB-NR* | *Non-redundant set of sequences from the OMPdb.* | *651,874* | *Comparison of predictors* |
| *IsItABarrelDB* | *Non-redundant set of bacterial sequences from NCBI NR database predicted TMBB.* | *1,949,081* | *Distribution of Lengths*<br>*Amino acid distribution* |

**Supplementary Table S1. Datasets used to develop, test and apply IsItABarrel.** In the text, whenever the sets NONTMBB1K and MP1929 were used, it is assumed that these are the negative examples, while the TMBB121 set was used in all cases as the set of positive examples.

# References


Almagro Armenteros, J. J., Tsirigos, K. D., Sonderby, C. K., Petersen, T. N., Winther, O., Brunak, S., von Heijne, G., & Nielsen, H. (2019). SignalP 5.0 improves signal peptide predictions using deep neural networks. *Nat Biotechnol*, *37*(4), 420-423. https://doi.org/10.1038/s41587-019-0036-z

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*, *25*(17), 3389-3402. https://doi.org/10.1093/nar/25.17.3389

Bagos, P. G., Liakopoulos, T. D., Spyropoulos, I. C., & Hamodrakas, S. J. (2004). A Hidden Markov Model method, capable of predicting and discriminating beta-barrel outer membrane proteins. *BMC Bioinformatics*, *5*, 29. https://doi.org/10.1186/1471-2105-5-29

Bailey, T. L., Johnson, J., Grant, C. E., & Noble, W. S. (2015). The MEME Suite. *Nucleic Acids Res*, *43*(W1), W39-49. https://doi.org/10.1093/nar/gkv416

Berven, F. S., Flikka, K., Jensen, H. B., & Eidhammer, I. (2004). BOMP: a program to predict integral beta-barrel outer membrane proteins encoded within genomes of Gram-negative bacteria. *Nucleic Acids Res*, *32*(Web Server issue), W394-399. https://doi.org/10.1093/nar/gkh351

Bigelow, H., & Rost, B. (2006). PROFtmb: a web server for predicting bacterial transmembrane beta barrel proteins. *Nucleic Acids Res*, *34*(Web Server issue), W186-188. https://doi.org/10.1093/nar/gkl262

Burghardt, T., Nather, D. J., Junglas, B., Huber, H., & Rachel, R. (2007). The dominating outer membrane protein of the hyperthermophilic Archaeum Ignicoccus hospitalis: a novel pore-forming complex. *Mol Microbiol*, *63*(1), 166-176. https://doi.org/10.1111/j.1365-2958.2006.05509.x

Dhar, R., Feehan, R., & Slusky, J. S. G. (2021). Membrane Barrels Are Taller, Fatter, Inside-Out Soluble Barrels. *J Phys Chem B*, *125*(14), 3622-3628. https://doi.org/10.1021/acs.jpcb.1c00878

Doyle, M. T., & Bernstein, H. D. (2019). Bacterial outer membrane proteins assemble via asymmetric interactions with the BamA beta-barrel. *Nat Commun*, *10*(1), 3358. https://doi.org/10.1038/s41467-019-11230-9

Franklin, M. W., Nepomnyachiy, S., Feehan, R., Ben-Tal, N., Kolodny, R., & Slusky, J. S. G. (2018). Efflux Pumps Represent Possible Evolutionary Convergence onto the beta-Barrel Fold. *Structure*, *26*(9), 1266-1274 e1262. https://doi.org/10.1016/j.str.2018.06.007

Franklin, M. W., Nepomnyachyi, S., Feehan, R., Ben-Tal, N., Kolodny, R., & Slusky, J. S. (2018). Evolutionary pathways of repeat protein topology in bacterial outer membrane proteins. *Elife*, *7*. https://doi.org/10.7554/eLife.40308

Freeman, T. C., Jr., & Wimley, W. C. (2010). A highly accurate statistical approach for the prediction of transmembrane beta-barrels. *Bioinformatics*, *26*(16), 1965-1974. https://doi.org/10.1093/bioinformatics/btq308

Freeman, T. C., Jr., & Wimley, W. C. (2012). TMBB-DB: a transmembrane beta-barrel proteome database. *Bioinformatics*, *28*(19), 2425-2430. https://doi.org/10.1093/bioinformatics/bts478

Garrow, A. G., Agnew, A., & Westhead, D. R. (2005). TMB-Hunt: an amino acid composition based method to screen proteomes for beta-barrel transmembrane proteins. *BMC Bioinformatics*, *6*, 56. https://doi.org/10.1186/1471-2105-6-56

Gessmann, D., Chung, Y. H., Danoff, E. J., Plummer, A. M., Sandlin, C. W., Zaccai, N. R., & Fleming, K. G. (2014). Outer membrane beta-barrel protein folding is physically controlled by periplasmic lipid head groups and BamA. *Proc Natl Acad Sci U S A*, *111*(16), 5878-5883. https://doi.org/10.1073/pnas.1322473111

Griep, S., & Hobohm, U. (2010). PDBselect 1992-2009 and PDBfilter-select. *Nucleic Acids Res*, *38*(Database issue), D318-319. https://doi.org/10.1093/nar/gkp786

Gromiha, M. M. (2005). Motifs in outer membrane protein sequences: applications for discrimination. *Biophys Chem*, *117*(1), 65-71. https://doi.org/10.1016/j.bpc.2005.04.005

Gromiha, M. M., & Suwa, M. (2006). Discrimination of outer membrane proteins using machine learning algorithms. *Proteins*, *63*(4), 1031-1037. https://doi.org/10.1002/prot.20929

Hagan, C. L., Wzorek, J. S., & Kahne, D. (2015). Inhibition of the beta-barrel assembly machine by a peptide that binds BamD. *Proc Natl Acad Sci U S A*, *112*(7), 2011-2016. https://doi.org/10.1073/pnas.1415955112

Hauser, M., Steinegger, M., & Soding, J. (2016). MMseqs software suite for fast and deep clustering and searching of large protein sequence sets. *Bioinformatics*, *32*(9), 1323-1330. https://doi.org/10.1093/bioinformatics/btw006

Hayat, S., & Elofsson, A. (2012). Ranking models of transmembrane beta-barrel proteins using Z-coordinate predictions. *Bioinformatics*, *28*(12), i90-96. https://doi.org/10.1093/bioinformatics/bts233

Hayat, S., Peters, C., Shu, N., Tsirigos, K. D., & Elofsson, A. (2016). Inclusion of dyad-repeat pattern improves topology prediction of transmembrane beta-barrel proteins. *Bioinformatics*, *32*(10), 1571-1573. https://doi.org/10.1093/bioinformatics/btw025

Huang, Y., Niu, B., Gao, Y., Fu, L., & Li, W. (2010). CD-HIT Suite: a web server for clustering and comparing biological sequences. *Bioinformatics*, *26*(5), 680-682. https://doi.org/10.1093/bioinformatics/btq003

Jackups, R., Jr., & Liang, J. (2005). Interstrand pairing patterns in beta-barrel membrane proteins: the positive-outside rule, aromatic rescue, and strand registration prediction. *J Mol Biol*, *354*(4), 979-993. https://doi.org/10.1016/j.jmb.2005.09.094

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Zidek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., . . . Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, *596*(7873), 583-589. https://doi.org/10.1038/s41586-021-03819-2

Kamisetty, H., Ovchinnikov, S., & Baker, D. (2013). Assessing the utility of coevolution-based residue-residue contact predictions in a sequence- and structure-rich era. *Proc Natl Acad Sci U S A*, *110*(39), 15674-15679. https://doi.org/10.1073/pnas.1314045110

Kazemian, H. B., & Grimaldi, C. M. (2020). Cascading classifier application for topology prediction of transmembrane beta-barrel proteins. *J Bioinform Comput Biol*, *18*(6), 2050034. https://doi.org/10.1142/S0219720020500341

Klingl, A., Pickl, C., & Flechsler, J. (2019). Archaeal Cell Walls. *Subcell Biochem*, *92*, 471-493. https://doi.org/10.1007/978-3-030-18768-2_14

Liu, Q., Zhu, Y., Wang, B., & Li, Y. (2003). Identification of beta-barrel membrane proteins based on amino acid composition properties and predicted secondary structure. *Comput Biol Chem*, *27*(3), 355-361. https://doi.org/10.1016/s1476-9271(02)00085-3

Ma, J., Wang, S., Wang, Z., & Xu, J. (2015). Protein contact prediction by integrating joint evolutionary coupling analysis and supervised learning. *Bioinformatics*, *31*(21), 3506-3513. https://doi.org/10.1093/bioinformatics/btv472

Martelli, P. L., Fariselli, P., Krogh, A., & Casadio, R. (2002). A sequence-profile-based HMM for predicting and discriminating beta barrel membrane proteins. *Bioinformatics*, *18 Suppl 1*, S46-53. https://doi.org/10.1093/bioinformatics/18.suppl_1.s46

Megrian, D., Taib, N., Witwinowski, J., Beloin, C., & Gribaldo, S. (2020). One or two membranes? Diderm Firmicutes challenge the Gram-positive/Gram-negative divide. *Mol Microbiol*, *113*(3), 659-671. https://doi.org/10.1111/mmi.14469

Natt, N. K., Kaur, H., & Raghava, G. P. (2004). Prediction of transmembrane regions of beta-barrel proteins using ANN- and SVM-based methods. *Proteins*, *56*(1), 11-18. https://doi.org/10.1002/prot.20092

Randall, A., Cheng, J., Sweredoski, M., & Baldi, P. (2008). TMBpro: secondary structure, beta-contact and tertiary structure prediction of transmembrane beta-barrel proteins. *Bioinformatics*, *24*(4), 513-520. https://doi.org/10.1093/bioinformatics/btm548

Remmert, M., Biegert, A., Linke, D., Lupas, A. N., & Soding, J. (2010). Evolution of outer membrane beta-barrels from an ancestral beta beta hairpin. *Mol Biol Evol*, *27*(6), 1348-1358. https://doi.org/10.1093/molbev/msq017

Remmert, M., Linke, D., Lupas, A. N., & Soding, J. (2009). HHomp--prediction and classification of outer membrane proteins. *Nucleic Acids Res*, *37*(Web Server issue), W446-451. https://doi.org/10.1093/nar/gkp325

Robert, V., Volokhina, E. B., Senf, F., Bos, M. P., Van Gelder, P., & Tommassen, J. (2006). Assembly factor Omp85 recognizes its outer membrane protein substrates by a species-specific C-terminal motif. *PLoS Biol*, *4*(11), e377. https://doi.org/10.1371/journal.pbio.0040377

Savojardo, C., Fariselli, P., & Casadio, R. (2011). Improving the detection of transmembrane beta-barrel chains with N-to-1 extreme learning machines. *Bioinformatics*, *27*(22), 3123-3128. https://doi.org/10.1093/bioinformatics/btr549

Schulz, G. E. (2002). The structure of bacterial outer membrane proteins. *Biochimica et Biophysica Acta (BBA) - Biomembranes*, *1565*(2), 308-317. https://doi.org/10.1016/s0005-2736(02)00577-1

Slusky, J. S., & Dunbrack, R. L., Jr. (2013). Charge asymmetry in the proteins of the outer membrane. *Bioinformatics*, *29*(17), 2122-2128. https://doi.org/10.1093/bioinformatics/btt355

Struyvé, M., Moons, M., & Tommassen, J. (1991). Carboxy-terminal phenylalanine is essential for the correct assembly of a bacterial outer membrane protein. *Journal of Molecular Biology*, *218*(1), 141-148. https://doi.org/10.1016/0022-2836(91)90880-f

Tian, W., Lin, M., Tang, K., Liang, J., & Naveed, H. (2018). High-resolution structure prediction of beta-barrel membrane proteins. *Proc Natl Acad Sci U S A*, *115*(7), 1511-1516. https://doi.org/10.1073/pnas.1716817115

Tsirigos, K. D., Bagos, P. G., & Hamodrakas, S. J. (2011). OMPdb: a database of {beta}-barrel outer membrane proteins from Gram-negative bacteria. *Nucleic Acids Res*, *39*(Database issue), D324-331. https://doi.org/10.1093/nar/gkq863

Tsirigos, K. D., Elofsson, A., & Bagos, P. G. (2016). PRED-TMBB2: improved topology prediction and detection of beta-barrel outer membrane proteins. *Bioinformatics*, *32*(17), i665-i671. https://doi.org/10.1093/bioinformatics/btw444

Waldispuhl, J., Berger, B., Clote, P., & Steyaert, J. M. (2006). transFold: a web server for predicting the structure and residue contacts of transmembrane beta-barrels. *Nucleic Acids Res*, *34*(Web Server issue), W189-193. https://doi.org/10.1093/nar/gkl205

Waldispuhl, J., O'Donnell, C. W., Devadas, S., Clote, P., & Berger, B. (2008). Modeling ensembles of transmembrane beta-barrel proteins. *Proteins*, *71*(3), 1097-1112. https://doi.org/10.1002/prot.21788

Wang, G., & Dunbrack, R. L., Jr. (2005). PISCES: recent improvements to a PDB sequence culling server. *Nucleic Acids Res*, *33*(Web Server issue), W94-98. https://doi.org/10.1093/nar/gki402

Wang, S., Sun, S., Li, Z., Zhang, R., & Xu, J. (2017). Accurate De Novo Prediction of Protein Contact Map by Ultra-Deep Learning Model. *PLoS Comput Biol*, *13*(1), e1005324. https://doi.org/10.1371/journal.pcbi.1005324

Wang, X., Peterson, J. H., & Bernstein, H. D. (2021). Bacterial Outer Membrane Proteins Are Targeted to the Bam Complex by Two Parallel Mechanisms. *mBio*, *12*(3). https://doi.org/10.1128/mBio.00597-21

Wimley, W. C. (2002). Toward genomic identification of beta-barrel membrane proteins: composition and architecture of known structures. *Protein Sci*, *11*(2), 301-312. https://doi.org/10.1110/ps.29402

Wimley, W. C. (2003). The versatile β-barrel membrane protein. *Current Opinion in Structural Biology*, *13*(4), 404-411. https://doi.org/10.1016/s0959-440x(03)00099-x

Zhai, Y., & Saier, M. H., Jr. (2002). The beta-barrel finder (BBF) program, allowing identification of outer membrane beta-barrel proteins encoded within prokaryotic genomes. *Protein Sci*, *11*(9), 2196-2207. https://doi.org/10.1110/ps.0209002