SUPPORTING INFORMATION
High resolution structure prediction of β-barrel membrane proteins

Wei Tian[1], Meishan Lin[1], Ke Tang[1], Jie Liang[*1], and Hammad Naveed[†2]

[1]University of Illinois at Chicago, Department of Bioengineering, 835 South Wolcott Avenue, Chicago, IL, 60607, USA
[2]Toyota Technological Institute at Chicago, 6045 South Kenwood Avenue, Chicago, IL, 60637, USA.

1 Dataset

We use 59 non-homologous βMPs (resolution 1.45Å– 3.2Å) with less than 30% pairwise sequence identity for this study. All 59 βMPs are used to construct the empirical potential function, but predictions are only made for 51 proteins, after excluding multichain βMPs to avoid over estimation of repeated interaction types. Based on the number of strands (or equivalently, number of residues) and the stability of the proteins [1], we divide the dataset into five subsets (Table S1).

2 Secondary structure determination

Existing computational approaches can successfully identify the location of β-strands [2, 3]. However, to assess our 3D modeling approach without any short coming from the secondary structure prediction we use the β-strands from the dssp program that uses PDB structure to calculate the location of the β-strands [4]. Only l_{cut} number of resides from the periplasmic side of each dssp strand are used for register prediction. We choose $l_{cut} = 12$ since the length of strands in the dataset has a mean of 12.7, mode of 12, and median of 12. For 3D structure construction, complete dssp strands are used.

3 Strand Register Prediction

To predict the register of a stand pair, we have developed a model incorporating both the empirical potential scores of physical interactions between strands from our previous study [5] and the se-

[*]Email: jliang@uic.edu
[†]Email: hammad.naveed@ttic.edu
To whom correspondence may be addressed.
[2]Current affiliation: Department of Computer Science, National University of Computer and Emerging Sciences, Islamabad Pakistan.

Group	Description	PDB id
1	Small βMPs ($N < 16$) without inplugs or outclamps	1bxw, 1qj8, 1p4t, 2f1t, 1thq, 2erv, 2lhf, 2mlh, 3dzm, 1qd6, 2f1c, 1k24, 1i78, 2wjr, 4pr7
2	Small βMPs ($N < 16$) with inplugs or outclamps	1t16, 1uyn, 1tly, 3aeh, 3bs0, 3dwo, 3fid, 3kvn, 4e1s
3	Medium oligomeric βMPs ($16 \leq N < 20$)	2mpr, 1a0s, 2omf, 2por, 1prn, 1e54, 2o4v, 3vzt, 4n75
4	Medium monomeric βMPs ($16 \leq N < 20$)	2qdz, 2ynk, 3emn, 3rbh, 3syb, 3szv, 4c00, 4gey
5	Large βMPs ($N \geq 20$)	1fep, 2fcp, 1kmo, 1nqe, 1xkw, 2vqi, 3csl, 3rfz, 3v8x, 4q35
6	Multichain βMPs	1ek9, 1yc9, 2gr8, 2lme, 3pik, 3b07, 3o44, 7ahl

Table S1: The groups of βMPs in this study. All the six groups are used in the construction of the empirical energy function. Structure predictions are made for only the first five groups.

quence covariation information that can identify medium-to-large range residue contacts based on the concept that spatially close residues might coevolve. In a strand pair, one strand is fixed and the other strand slides up and down, thus changing the register and also the interstrand residue contacts. Our model gives a score for each register with Equation (1)

$$E(r) = E_{emp}(r) + E_{sc}(r), \tag{1}$$

where r is a given register of the strand pair, $E_{emp}(r)$ is the empirical potential score of physical interstrand interactions, and E_{sc} is the score from sequence covariation analysis. The register with the lowest score is selected as the prediction.

3.1 Model for interstrand interactions

The model for physical interactions between a strand pair from Ref [5] is used in this study. Briefly, the model assumes that neighboring strands interact through canonical strong hydrogen bonds, weak hydrogen bonds, and non-hydrogen bonds (sidechain interactions), which is based on the observation of the periodic dyad bonding repeat pattern of antiparallel β-sheets [6] (Figure S1). The entropy for unbonded regions and left/right handedness of the strand pair are considered as well. See Ref [5] for more detailed description of the model. The total empirical score of certain register r of a given strand pair is calculated with the empirical scoring function

$$\begin{aligned} E_{emp}(r) =& \alpha \sum_{k_i} \sum_{k_{i+1}} E_{SH}(k_i, k_{i+1}; r) + \beta \sum_{k_i} \sum_{k_{i+1}} E_{WH}(k_i, k_{i+1}; r) \\ & \gamma \sum_{k_i} \sum_{k_{i+1}} E_{NH}(k_i, k_{i+1}; r) + \delta \ln(\frac{n_{ref} + \Delta L(r)}{n_{ref}}) + \varepsilon[LH(r)], \end{aligned} \tag{2}$$

where $E_{SH}(k_i, k_{i+1}; r)$, $E_{WH}(k_i, k_{i+1}; r)$, and $E_{NH}(k_i, k_{i+1}; r)$ are the empirical energies of strong, weak, and non-hydrogen bonds between the residue k_i on strand i and the residue k_{i+1} on strand $i + 1$, respectively. $n_{ref} = 8.5$ is the average length of loops. $\Delta L(r)$ is related to the number of

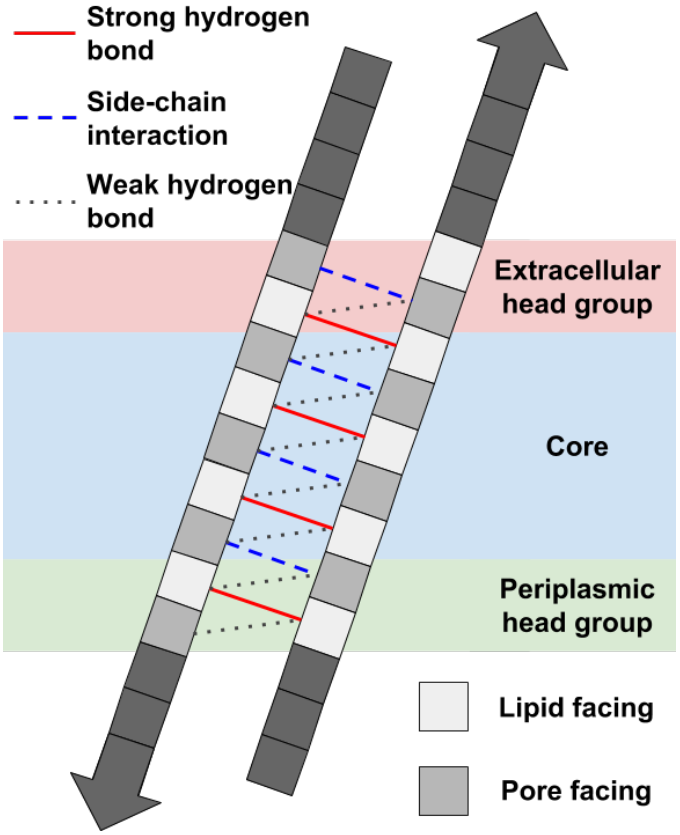

Figure S1: Model for interstrand interactions between adjacent strands.

residues that do not share a hydrogen bond with the adjacent strand in the register r, minus the difference in strand lengths. $LH(r)$ is the penalty for left handed twist ($r < 0$) since all β-sheets are right handed.

$$LH(r) = \begin{cases} r & r < 0 \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

3.2 Model for sequence covariation

We use PSICOV [7] to calculate the sequence covariation scores of each residue pairs in TM regions. The score of certain register of a strand pair is calculated as the weighted summation of sequence covariation scores of residue pairs:

$$\begin{aligned} E_{sc} =& w_0 \sum_{k_i} \sum_{k_{i+1}} \delta(d_{k_i,k_{i+1}}, 0) Q(k_i, k_{i+1}) + w_1 \sum_{k_i} \sum_{k_{i+1}} \delta(d_{k_i,k_{i+1}}, 1) Q(k_i, k_{i+1}) \\ &+ w_2 \sum_{k_i} \sum_{k_{i+1}} \delta(d_{k_i,k_{i+1}}, 2) Q(k_i, k_{i+1}) \end{aligned} \tag{4}$$

where $Q(i, j)$ is the sequence covariation score of the residues k_i and k_{i+1}, $d_{k_i,k_{i+1}}$ is the distance between the two residues in the discretized conformational state space (Figure S2), w_c ($c = 0,\ 1,$ or 2) is the weight of residue pair whose distance is c, and $\delta(d_{k_i,k_{i+1}}, c)$ is the Kronecker delta function which identifies if the distance of the residues k_i and k_{i+1} is c. All residue pairs with distance larger than 2 are ignored in the calculation, for they are unlikely to have any physical interaction.

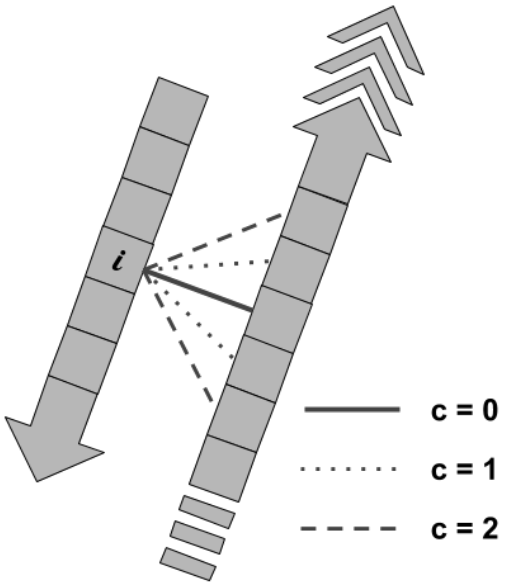

Figure S2: Model for calculating sequence covariation between adjacent strands.

Group	α	β	γ	δ	ε	w_0	w_1	w_2
1	0.026	0.038	0.036	0.245	0.050			
2	0.055	0.100	0.075	0.450	0.120			
3	0.000	0.082	0.006	0.052	0.074	-0.500	-0.136	-0.364
4	0.045	0.020	0.024	0.290	0.100			
5	0.045	0.024	0.014	0.110	0.135			

Table S2: Values for α, β, γ, δ, ε, w_0, w_1, and w_2.

3.3 Parameter determination and cross-validation

We first fix the weights (w_0, w_1, and w_2) in the sequence covariation model. Since the sequence covariation analysis comes purely from sequences and needs no prior knowledge of the dataset, we neither use the leave-one-out scheme for the searching of these three weights, nor discriminate the groups of the dataset. The weights (w_0, w_1, and w_2) are determined by searching for the values such that the score E_{sc} alone can give a best prediction of the registers of the neighboring strand pairs in the dataset.

Then the leave-one-out cross-validation (LOOCV) is used for searching the other undetermined weights (α, β, γ, δ, and ε) in Equation (2) so that the total scores calculated via Equation (1) give the best prediction. In LOOCV, we left one protein out of the data set while using the other proteins to construct the empirical potential function. The registers of the leave-out protein were predicted. This process was repeated for each protein to find the optimized values of the group-specific parameters (α, β, γ, δ, and ε), which gave the best register prediction accuracy for that group. The parameters (α, β, γ, δ, and ε) were optimized using an adaptive grid search. The final values used in the model is listed in Table S2.

3.4 Sidechain direction prediction

We only predict the sidechain direction of the first residue on the periplasmic side of each strand. As we ignore β-bulge and assume sidechain directions of a strand always follow an alternative lipid-facing-pore-facing pattern, sidechain direction of all the other residues can be obtained accordingly.

In the original reduced state space (RSS) model, there are 2, 5, and 2 residues in the extracellular headgroup, the core, and the periplasmic headgroup regions, respectively (Figure S1). However, it is known that membrane could become either thinner or thicker around transmembrane proteins adaptively. So, we use a variant of RSS where the number of the resides in each of these three regions can vary by 1 from the original RSS while the total thickness of these three regions is restricted to 7-11 residues.

We enumerated the combination of the sidechain directions and the legit conformations in the RSS variant aforementioned, and used the single body potential [1] derived from our βMP dataset to calculate the energy of each combination. The sidechain directions that give the lowest energy within the enumeration of each strand are selected as predictions, which gives a 98% accuracy.

3.5 Global register optimization

Strand register prediction considers hydrogen bonds contact two adjacent strands at a time. However, global hydrogen bond pattern is better represented by the shear number of the protein. The shear number is the displacement of the relative positions in the TM strands if one starts to follow the backbone hydrogen bond between strands, beginning from strand 1 and returning after a full circle to the same strand (see Figure S5, more examples can be found in Ref [8]). The shear number of a βMP also equals to the sum of the strand registers. When these registers are not known, the shear number can be estimated reliably from the number of TM strands [5].

We optimize strand register prediction so that the predicted shear number S is as close as possible to the most common shear number S_{com} of the βMBs of the same size. Shear number of all proteins in the dataset are given in Table S3. The predicted shear number S is calculated from the predicted registers:

$$S = \sum_{i=1}^{N} r_i, , \tag{5}$$

where r_i is the predicted strand register of the i-th strand. N is the total number of strands.

For each strand, two register candidates with lowest scores in the register prediction step are kept. The summation of the first register candidate of each strand gives the predicted shear number before optimization. This selection also gives the total score for the predicted protein conformation by summing up the score of each predicted register. The global shear optimization attempts to replace the first candidate with the second one of each strand in the final selection so that the predicted shear number is as close to the target shear as possible while keeping the total score for the protein as close to the minimum score as possible.

The register candidates are first filtered according to the predicted sidechain directions of the first periplasmic residues of that strand and of its sequential neighbor: If the first residues of the i-th strand and of the $(i+1)$-th strand have the same sidechain direction, only the candidate(s) of the i-th strand with even register number is kept; otherwise, the odd one(s). This criteria is based on the fact that hydrogen-bonded residues on adjacent strands always have the same sidechain direction. When neither of the candidates satisfy this criteria, both are kept.

Subsequently, the strands are sorted in ascending order according to the difference between the scores of the two candidates of each strand. The difference is 0 if only one candidate was kept in the

PDB id	N	S	PDB id	N	S	PDB id	N	S	PDB id	N	S
1qj8	8	8	3aeh	12	14	2o4v	16	20	3szv	18	22
1bxw	8	10	3fid	12	14	2omf	16	20	3emn	19	20
1p4t	8	10	3kvn	12	14	2por	16	20	1fep	22	24
1thq	8	10	4e1s	12	14	2qdz	16	20	1kmo	22	24
2erv	8	10	4pr7	12	14	3vzt	16	20	1nqe	22	24
2f1t	8	10	1qd6	12	16	4c00	16	20	1xkw	22	24
2lhf	8	10	1tly	12	16	4gey	16	20	2fcp	22	24
2mlh	8	10	1t16	14	14	4n75	16	22	3csl	22	24
3dzm	8	10	3bs0	14	14	2ynk	18	20	3v8x	22	24
1i78	10	12	3dwo	14	14	1a0s	18	20	2vqi	24	26
1k24	10	12	2f1c	14	16	2mpr	18	22	3rfz	24	26
1uyn	12	14	1e54	16	20	3rbh	18	22	4q35	26	30
2wjr	12	14	1prn	16	20	3syb	18	22			

Table S3: Shear number of βMPs.

previous step. We scan the strands in this order and make the final selection for each strand. For the top two strands, the second candidate will be selected if it can bring the predicted shear number S closer to the target S_{com}. For the remaining strands, the second candidate will be selected only when it can keep the predicted shear number S in same parity with the target shear number S_{com}. and can also reduce the shear number difference $|S - S_{com}|$ between prediction and target.

4 Three dimensional structure prediction

4.1 C_α trace construction

An intertwined coil model was used in our previous study [5], in which the C_α trace of a βMP was generated, followed by backbone generation, sidechain generation, and MD minimization. If we look closely at the C_α trace of a βMP structure, however, we find that the intertwined coil model is not able to capture the following geometric properties: 1) the C_α trace of a strand is not as smooth as a coil, but is zigzag-like (Figure S3a). The C_α atoms of lipid-facing residues are farther away from the vertical axis of the barrel compared to those of pore-facing residues; and 2) the C_α atoms on two adjacent strands are not equidistant as depicted by the intertwined coil model. The distance between of C_α atoms of residues sharing strong hydrogen bonds are larger than those sharing non-hydrogen bonds (Figure S3b and S3c).

To capture these geometric properties, we developed a parametric structural model of intertwined zigzag coils, in which the C_α trace of each strand is depicted by a zigzag coil that wraps around a hypothetical cylinder. To calculate the C_α position of a strand, we first build a coil basis for the strand (Figure S4a). The tilt angle θ of coil basis with respect to the vertical cylinder axis and the radius r of the cylinder are calculated using Equation (6) following McLachlan [9]:

$$\begin{aligned} \theta &= \arctan(\frac{SA}{NB}), \\ r &= \frac{B}{2\sin(\frac{\pi}{N})\cos\theta}, \end{aligned} \tag{6}$$

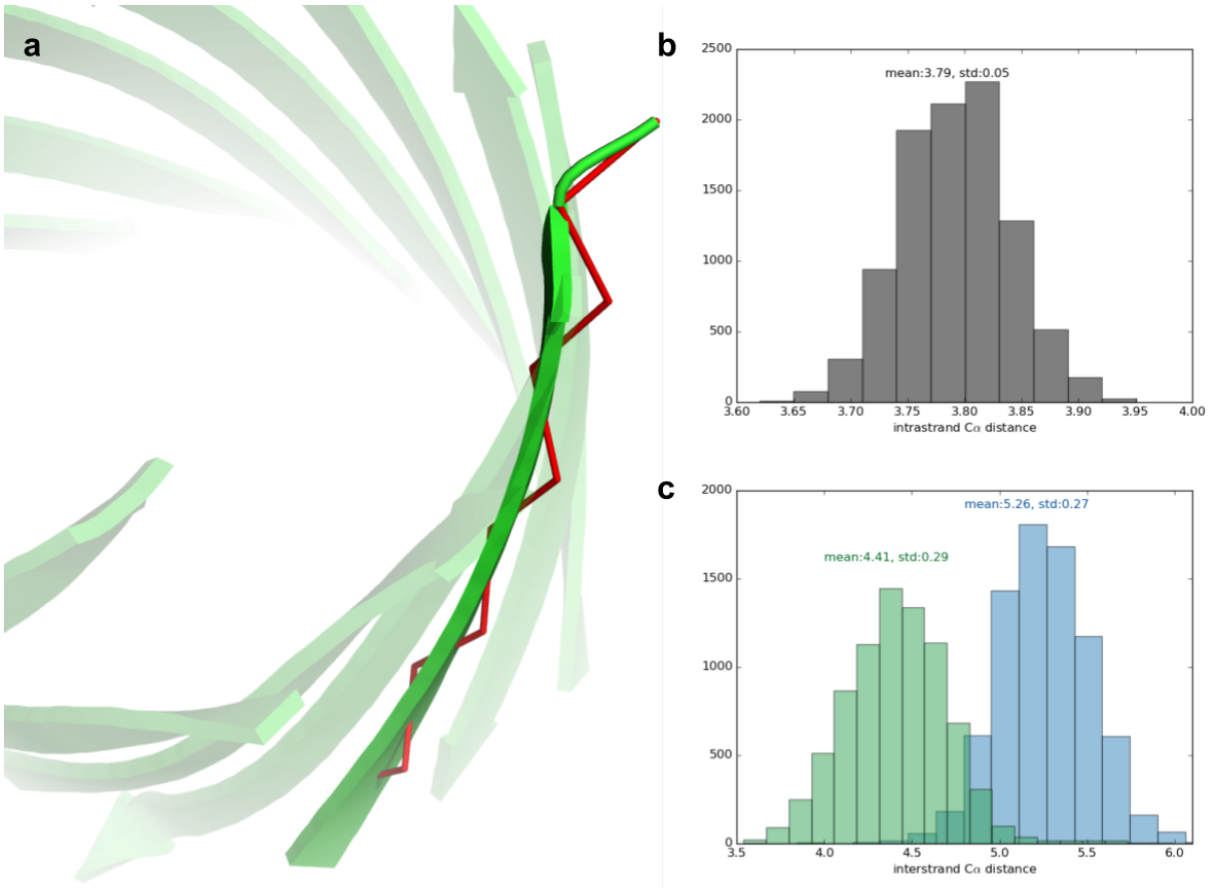

Figure S3: Geometric properties of βMPs. (a) The C_α trace of a β-strand shows a zigzag pattern (red). The structure used here is TodX (PDB id:3bs0). (b) Distribution of distance between consecutive C_α atoms on the same strand. (c) Distribution of distance between C_α atoms of residues sharing a non-hydrogen bond (green) and of those sharing a strong hydrogen bond (blue) on adjacent strands.

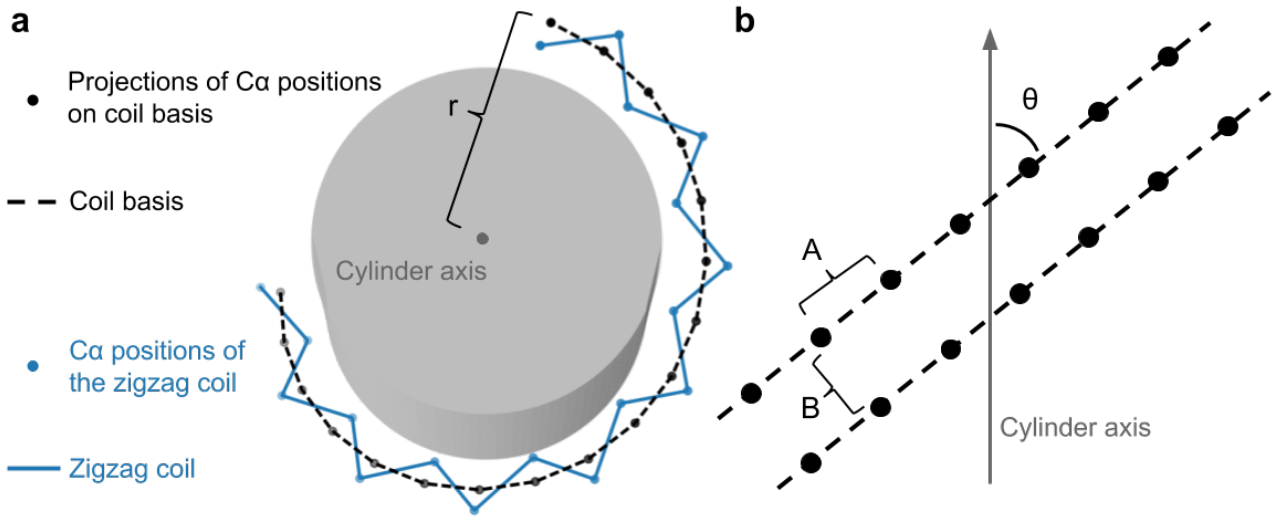

Figure S4: The parametric structural model of intertwined zigzag coils. (a) One zigzag coil (blue) and the corresponding coil basis (black) wrap around the hypothetical cylinder (grey). (b) The relative position and the corresponding parameters of coil bases are shown after unwrapping the coil bases onto a plane.

where A is the distance between projections of consecutive C_α atoms on the same coil basis, and B is the distance between projections of C_α atoms sharing a strong or non-hydrogen bond on adjacent strands. Note that A and B here are not the intra- and the inter-strand C_α distances. N is the number of β-strands, and S is the shear number for the βMP.

Using time curves from differential geometry [10], each position j of C_α projection on coil basis i is represented by a parametric curve represented by Equation (7).

$$\begin{aligned} c_i(t_{ij}) &= \left(r\cos(t_{ij} - \frac{2\pi i}{N}),\ r\sin(t_{ij} - \frac{2\pi i}{N}),\ bt_{ij} \right), \\ b &= \frac{r}{\tan\theta}, \end{aligned} \tag{7}$$

where $c_i(\cdot)$ is the parametric curve of the i-th coil basis. Let $V_r(t_{ij})$ be the vector from position j of coil basis i to position j of coil basis $i+1$, and $T_r(t_{ij})$ the tangent vector at position j of coil basis i. Given that the vector between two C_α atoms sharing a strong or non-hydrogen bond on adjacent strands is roughly perpendicular to the strands, the inner product of the two vectors should be 0:

$$\begin{aligned} V_r(t_{ij}) \cdot T_r(t_{ij}) &= 0, \\ V_r(t_{ij}) &= c_{i+1}(t_{i+1,j}) - c_i(t_{ij}), \\ T_r(t_{ij}) &= \left(-\frac{r}{c}\sin(t_{ij}),\ \frac{r}{c}\cos(t_{ij}),\ \frac{b}{c} \right), \\ c &= \sqrt{r^2 + b^2}. \end{aligned} \tag{8}$$

By solving Equation (8), t_{ij} can be written as

$$\begin{aligned} t_{ij} &= \frac{s_{ij}}{c}, \\ s_{ij} &= (j - \sum_{k=1}^{i-1} R_k)A + i\frac{2\pi r^2}{cN}, \end{aligned} \tag{9}$$

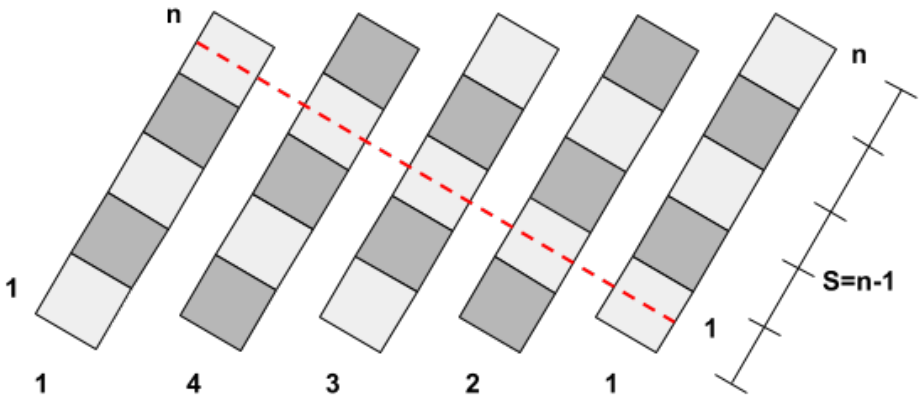

Figure S5: Shear number is the displacement of the relative positions in the TM strands if one starts to follow the backbone hydrogen bond between strands, starting from strand 1 and returning after a full circle to the same strand. For example, the shear number for the 4 strand β-barrel shown above is $n-1$.

where R_k is the register of the k-th strand.

Using different radii for the lipid-facing and the pore-facing residues, $\tilde{c}_i(t_{ij})$, the zigzag pattern of the C_α trace (Figure S4a) can be taken into account by Equation (10).

$$\begin{aligned} \tilde{c}_i(t_{ij}) &= \left(r' \cos(t_{ij} - \frac{2\pi i}{N}),\ r' \sin(t_{ij} - \frac{2\pi i}{N}),\ b t_{ij} \right), \\ b &= \frac{r}{\tan\theta}, \\ r' &= \begin{cases} r + \Delta r, & \text{if the position is lipid-facing} \\ r - \Delta r, & \text{if the position is pore-facing} \end{cases}. \end{aligned} \tag{10}$$

Considering that the distance between C_α atoms of residues sharing a strong hydrogen bond is different from the distance between those sharing a non-hydrogen bond (Figure S3c), the 3D coordinates $\mathbb{C}_i(t_{ij})$ of C_α atoms in the intertwined coiled zigzag model can be written as Equation (11)

$$\mathbb{C}_i(t_{ij}) = \begin{cases} \tilde{c}_i(t_{ij}) + \Delta w \frac{V_{r'}(t_{ij})}{\|V_{r'}(t_{ij})\|}, & \text{if } i \text{ is odd and the position is lipid-facing,} \\ & \text{or } i \text{ is even and the position is pore-facing .} \\ \tilde{c}_i(t_{ij}) - \Delta w \frac{V_{r'}(t_{i-1,j})}{\|V_{r'}(t_{i-1,j})\|}, & \text{otherwise} \end{cases} \tag{11}$$

4.2 Parameter estimation

The intrastrand C_α distance has very little variance from its mean value of 3.79 Å, while the interstrand C_α distance of residues sharing a strong or a non-hydrogen bond have different means (5.26 Å and 4.41 Å, respectively) and relatively larger variances (Figure S3c). We used $B = 4.83$Å, which is the mean value of interstrand C_α distance of residues sharing a strong or a non-hydrogen bond, and did a grid search for the values of A, Δd, and Δw that satisfy the following criteria:

1. Any value that makes the intrastrand C_α distance out of the range $[3.79 \pm 0.02]$ is rejected,
2. The average interstrand C_α distances of residues sharing a strong hydrogen bond and of residues sharing a non-hydrogen bond are as close to 5.26Å and to 4.41Å as possible,

The parameters we found are $A = 3.345$Å, $\Delta r = 0.83$Å, and $\Delta w = 0.22$Å, which give a intrastrand C_α distance of 3.77 ± 0.06, interstrand C_α distances of residues sharing a strong hydrogen bond of 5.28 ± 0.19 and interstrand C_α distances of residues not sharing a strong hydrogen bond of 4.43 ± 0.18.

As for a βMP with N strands, an approximation for the shear number S is

$$\begin{cases} S &= N, \quad N = 14, \\ S &= N + 4, \quad N = 16 \text{ or } 18, \\ S &= N + 2, \quad \text{otherwise.} \end{cases} \tag{12}$$

which is correct for all βMP structures with the exception of OmpX, OmpLA, Tsx, OmpG, Wzi, LptD and VDAC in our data set.

4.3 3D structure construction

Given the predicted C_α trace, Gront *et al.*'s algorithm (BBQ) [11] was used to construct the backbone of the barrel. As loop region is ignored in our model at this stage and the strands are disconnected with each other, BBQ tends to make mistake at the ends of the strands. So, for the C_α trace input for BBQ, we grew two additional pseudo C_α atoms on both extracellular side and periplasmic side for each strand. Then the backbone obtained from BBQ was fed to the Scwrl4 program [12] for sidechain generation. The pseudo C_α atoms were chopped after sidechain generation

5 Results of three-dimensional structure prediction

The set of 51 βMP structures are listed in Table S4, along with the PDB ids, the organism for the protein, the number of TM strands, and the RMSD values between the TM region and the TM+extended barrel regions of real and modeled structures for main chain and all atom models. It also lists the number of strands for which the strand register is correctly predicted before and after global shear optimization. The TM-regions of the predicted structures superimposed on experimentally determined structures are shown in Figure S7. A plot showing the RMSD against the size of the proteins can be seen in Figure S6.

6 Comparison with a previous study.

In a recent study, structures for 17 proteins (compared to the 51 proteins in this study) were predicted with an RMSD of 6.66 Å [13], as the number of sequences available for the remaining proteins is insufficient for computing sequence covariation. Our results show that this limitation can be removed by combining patterns of hydrogen bond and sidechain interactions derived from experimentally resolved 3D structures with the sequence covariation information. Figure S8 shows that even when the available sequences are insufficient for sequence covariation analysis alone (accuracy $\sim$ 30%), our model model can make accurate strand register prediction ($\sim$ 70%). Our improved modeling methodology can predict the 3D structures of 51 βMPs with an average RMSD of 3.48 Å. Moreover, in Ref [13], TM-align was employed to assess accuracy of predicted structures, which does not give the appropriate assessment of prediction accuracy. TM-align is used when the correspondence or residue-residue mappings between two structures are not known, as it will decide which portions of the sub-structures are sufficiently similar for RMSD/TM-score calculation.

In the case of computing the RMSD of a predicted structure and a known PDB structure, direct mappings of all TM residues between the two structures are already known and a straightforward direct RMSD calculation is required. We have carried out a direct measurement of RMSD using predicted structures of Hayat *et al.* The RMSD calculated using this approach is 6.66 Å, as compared to the reported 4.45 Å, which is the average RMSD of a subset of TM residues selected by TM-align. In addition, the authors of Ref [13] inflated their accuracy in strand register prediction by considering the predictions that were off by ± 1 register as correct. As there is a direct relationship between the sidechain orientation and the functions of the proteins, this relaxed definition of "correct" registration implies erroneous sidechain orientation and thus incorrect functional regions of the proteins.

Table S4: Data set and prediction results. Strand register prediction and RMSD between the TM region and the TM+extended barrel regions of real and modeled structures of 51 non-homologous βMPs.

Protein/PDB	Organism	Strands #	Correct register # before/after optimization	Shear # before/after optimization	TM domian C_α-RMSD		Barrel domain C_α-RMSD	
					Main chain	All atom	Main chain	All atom
OmpA/1bxw	*E. coli*	8	6/8	6/10/10	1.39	2.83	1.46	2.86
NspA/1p4t	*N. meningitidis*	8	7/8	8/10/10	1.45	2.50	1.83	2.95
OmpX/1qj8	*E. coli*	8	7/7	6/10/8	2.63	3.47	3.01	3.94
PagP/1thq	*E. coli*	8	5/6	5/10/10	3.35	4.25	3.35	4.25
PagL/2erv	*P. aeruginosa*	8	6/4	4/10/10	3.94	4.47	3.94	4.47
OmpW/2f1t	*E. coli*	8	7/6	12/10/10	3.12	4.00	3.20	4.22
OprH/2lhf	*P. aeruginosa*	8	7/8	12/10/10	1.49	2.43	1.49	2.42
Opa60/2mlh	*N. gonorrhoeae*	8	8/8	10/10/10	1.49	2.69	1.49	2.69
HB27/3dzm	*T. thermophilus*	8	6/6	4/10/10	2.85	3.41	3.00	3.65
OmpT/1i78	*E. coli*	10	9/8	14/12/12	3.69	4.53	4.84	5.86
OpcA/1k24	*N. meningitidis*	10	6/3	18/12/12	4.79	5.31	4.84	5.49
OmpLA/1qd6	*E. coli*	12	9/8	16/14/16	5.55	6.68	5.71	6.86
Txs/1tly	*E. coli*	12	7/7	18/14/16	4.71	5.57	4.72	5.59
NalP/1uyn	*N. meningitidis*	12	11/12	12/14/14	1.45	2.88	1.56	3.06
NanC/2wjr	*E. coli*	12	11/10	12/14/14	2.94	3.54	2.94	3.54
Hbp/3aeh	*E. coli*	12	12/12	14/14/14	1.78	3.02	1.80	2.97
LpxR/3fid	*S. enterica*	12	8/8	14/14/14	6.56	6.93	6.58	7.09
EstA/3kvn	*P. aeruginosa*	12	12/12	14/14/14	1.61	2.90	2.03	3.24
intimin/4e1s	*E. coli*	12	9/10	16/14/14	3.10	3.90	3.10	3.89
KdgM/4pr7	*D. dadantii*	12	9/9	14/14/14	3.91	4.52	3.99	4.55
FadL/1t16	*E. coli*	14	13/12	12/14/14	2.23	3.10	2.84	3.73
OmpG/2f1c	*E. coli*	14	10/9	12/14/16	3.09	3.96	3.15	4.10
TodX/3bs0	*P. putida*	14	14/14	14/14/14	1.30	2.13	2.01	2.97
FadL/3dwo	*P. aeruginosa*	14	12/10	10/14/14	3.15	3.76	3.82	4.42
Omp32/1e54	*C. acidovorans*	16	14/14	16/20/20	3.03	3.63	3.10	3.67
Porin/1prn	*R. balistica*	16	14/14	20/20/20	2.44	3.28	2.44	3.27
Porin P/2o4v	*P. aeruginosa*	16	15/14	18/20/20	3.28	4.26	3.58	4.44
OmpF/2omf	*E. coli*	16	15/16	18/20/20	2.60	3.90	2.79	4.01
Porin/2por	*R. capsulatus*	16	13/12	18/20/20	2.77	3.36	2.81	3.39
FhaC/2qdz	*B. pertussis*	16	12/11	23/21/20	6.18	6.68	5.97	6.49
PorB/3vzt	*N. meningitidis*	16	14/14	16/20/20	3.48	4.08	3.75	4.52
TamA/4c00	*E. coli*	16	12/13	18/20/20	4.68	5.12	4.78	5.21
OprB/4gey	*P. putida*	16	12/11	10/20/20	4.64	5.22	4.69	5.31
BamA/4n75	*E. coli*	16	14/15	18/20/22	3.44	4.07	3.53	4.23
ScrY/1a0s	*S. typhimurium*	18	15/15	25/25/20	5.14	5.60	5.17	5.64
LamB/2mpr	*S. typhimurium*	18	13/10	32/22/22	6.08	6.86	7.65	8.27
Wzi/2ynk	*E. coli*	18	14/13	20/22/20	3.61	4.42	3.92	4.70
AlgE/3rbh	*P. aeruginosa*	18	13/16	16/22/22	4.36	5.30	4.30	5.20
OpdP/3syb	*P. aeruginosa*	18	16/15	18/19/22	3.73	4.41	3.72	4.46
OpdO/3szv	*P. aeruginosa*	18	17/16	20/22/22	3.20	3.89	3.17	3.87
VADC1/3emn	*M. musculus*	19	10/10	19/19/20	3.53	4.34	3.53	4.34
FepA/1fep	*E. coli*	22	21/19	29/25/24	4.51	4.96	5.14	5.67
FecA/1kmo	*E. coli*	22	22/22	24/24/24	2.71	3.47	3.19	3.94
BtuB/1nqe	*E. coli*	22	21/20	22/24/24	2.84	3.60	3.53	4.26
FptA/1xkw	*P. aeruginosa*	22	22/22	24/24/24	3.29	3.84	3.88	4.34
FhuA/2fcp	*E. coli*	22	22/22	24/24/24	2.83	3.41	5.20	5.57
HasR/3csl	*S. marcescens*	22	22/22	24/24/24	2.71	3.35	3.16	3.89
TbpA/3v8x	*N. meningitidis*	22	21/20	26/24/24	2.58	3.68	4.96	5.74
PapC/2vqi	*E. coli*	24	22/22	23/26/26	6.06	6.62	6.42	7.02
FimD/3rfz	*E. coli*	24	20/21	30/31/26	4.74	5.47	4.79	5.60
LptD/4q35	*S. flexneri*	26	18/16	37/32/30	7.25	7.67	7.53	7.96

Table S5: Protein size and average mainchain RMSD using different prediction methods. Proteins with the number of strands ≤ 14 are grouped into the small dataset, those with > 14 and ≤ 20 strands are grouped into the medium dataset, and proteins with > 20 strands are grouped into the large dataset. In contrast to the other prediction methods, where there is considerable deterioration in the quality of predicted structures, the quality of prediction of our methods, 3D-BMPP, does not deteriorate for large-sized proteins.

Method	Small	Medium	Large
TMBpro-server, 2008	6.0	6.3	11.8
3D-SPoT, 2012	3.9	4.5	4.0
EVfold_bb, 2015	4.9	7.7	9.3
3D-BMPP, 2017	3.0	3.9	4.0

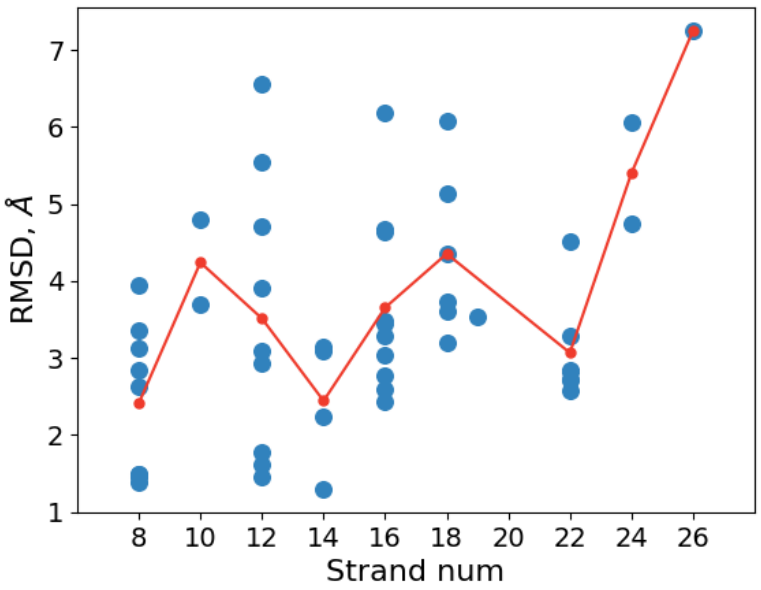

Figure S6: RMSDs of our prediction against the size of the proteins. Each blue dot represents one of the 51 predicted structures, while each red dot shows the average RMSD of predicted structures with the same corresponding strand number.

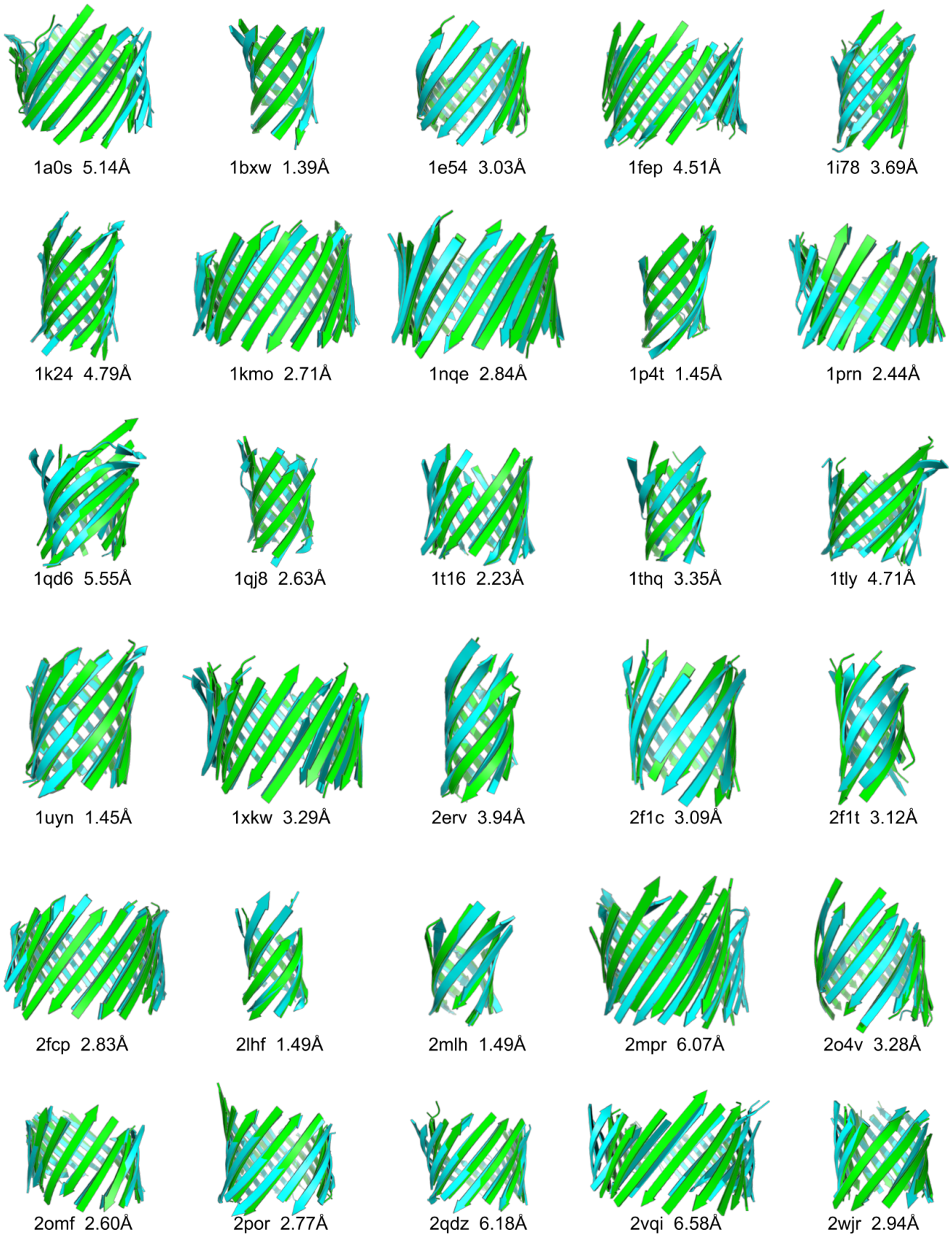

Figure S7: Structure prediction of TM regions. Predicted structures of the TM regions (green) are superimposed on experimentally determined structures (cyan).

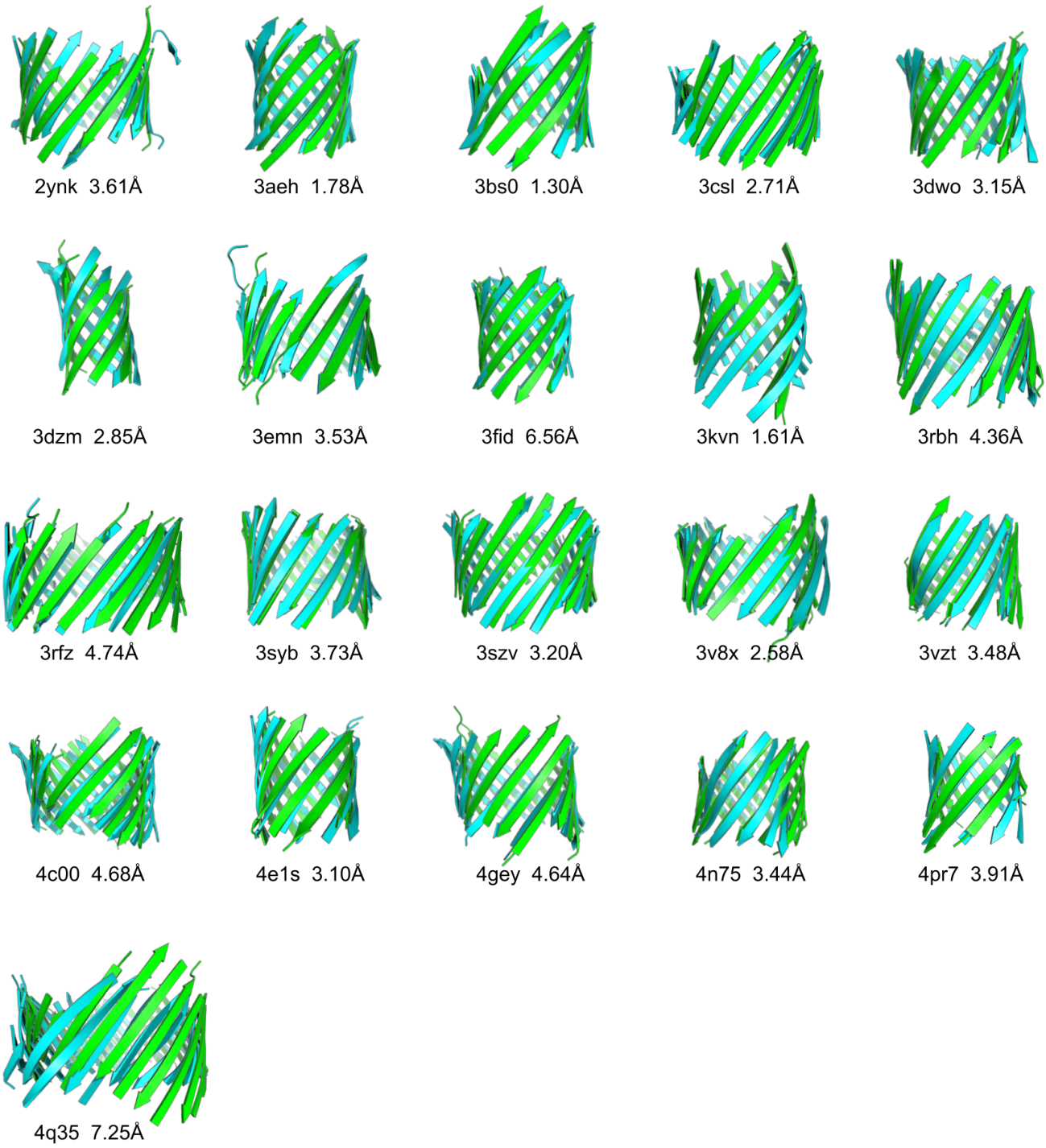

Figure S7: (Cont'd) Structure prediction of TM regions. Predicted structures of the TM regions (green) are superimposed on experimentally determined structures (cyan).

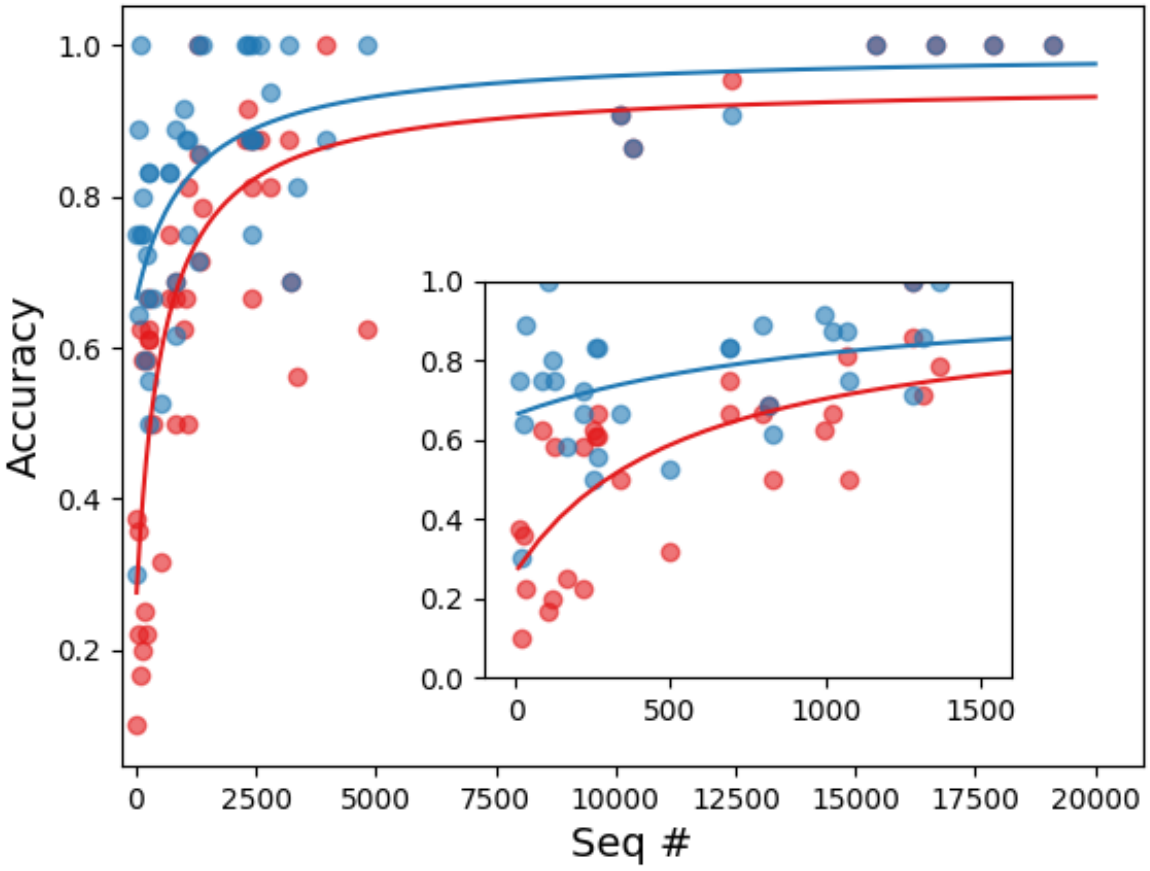

Figure S8: Our method on register prediction does not suffer from the limitation of requiring a large number ($\sim$ 1000) of available sequences for sequence covariation analysis. This figure shows how the number of available sequences affect register prediction accuracy. The numbers of sequences are found by HHblits [14]. Each blue dot represents the register prediction for one protein using our model, while each red dot represents the prediction made by the sequence covariation analysis results alone (using Equation 4 and Figure S2). The blue and red curves are fitted from the corresponding dots, respectively. The inset shows the details of proteins when the available number of sequences is limited. In these cases, our model can still make accurate prediction (accuracy at $\sim 0.7 = 70\%$) while the prediction made using covariation analysis along is not reliable ($\sim 0.3 = 30\%$)

References

[1] Naveed, H., Jackups, Jr., R. & Liang, J. Predicting weakly stable regions, oligomerization state, and protein-protein interfaces in transmembrane domains of outer membrane proteins. *Proc Natl Acad Sci U S A*, **106(31)**:12735–12740 (2009).

[2] Ou, Y., Chen, S. & Gromiha, M. Prediction of membrane spanning segments and topology in beta-barrel membrane proteins at better accuracy. *J Comput Chem*, **31(1)**:217–223 (2010).

[3] Hayat, S., Peters, C., Shu, N., Tsirigos, K. & Elofsson, A. Inclusion of dyad-repeat pattern improves topology prediction of transmembrane beta-barrel proteins. *Bioinformatics* (2016).

[4] Kabsch, W. & Sander, C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22(12)**:2577–2637 (1983).

[5] Naveed, H., Xu, Y., Jackups, Jr., R. & Liang, J. Predicting three-dimensional structures of transmembrane domains of beta-barrel membrane proteins. *J Am Chem Soc*, **134(3)**:1775–1781 (2012).

[6] Jackups Jr, R. & Liang, J. Interstrand pairing patterns in beta-barrel membrane proteins: the positive-outside rule, aromatic rescue, and strand registration prediction. *J Mol Biol*, **354(4)**:979–93 (2005).

[7] Jones, D., Buchan, D., Cozzetto, D. & Pontil, M. PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. *Bioinformatics*, **28(2)**:184–190 (2012).

[8] Liu, W. Shear numbers of protein beta-barrels: definition refinements and statistics. *J Mol Biol*, **275(4)**:541–545 (1998).

[9] McLachlan, A. Gene duplications in the structural evolution of chymotrypsin. *J Mol Biol*, **128(1)**:49–79 (1979).

[10] O'Neill, B. *Elementary Differential Geometry.*, Academic Press Inc., London UK1966.

[11] Gront, D., Kmiecik, S. & Kolinski, A. Backbone building from quadrilaterals: a fast and accurate algorithm for protein backbone reconstruction from alpha carbon coordinates. *J Comput Chem*, **28(9)**:1593–1597 (2007).

[12] Krivov, G., Shapovalov, M. & Dunbrack, Jr., R. Improved prediction of protein side-chain conformations with SCWRL4. *Proteins*, **77(4)**:778–795 (2009).

[13] Hayat, S., Sander, C., Marks, D. & Elofsson, A. All-atom 3d structure prediction of transmembrane beta-barrel proteins from sequences. *Proc Natl Acad Sci U S A*, **112(17)**:5413–5418 (2015).

[14] Remmert, M., Biegert, A., Hauser, A. & S??ding, J. HHblits: Lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nature Methods*, **9(2)**:173–175 (2012).