

# Supporting Online Material

## TMbed – Transmembrane proteins predicted through Language Model embeddings

Michael Bernhofer, Burkhard Rost

### Table of Contents for Supporting Online Material

TABLE OF CONTENTS FOR SUPPORTING ONLINE MATERIAL.....	1
SHORT DESCRIPTION OF SUPPORTING ONLINE MATERIAL .....	2
MATERIAL.....	3
Note S1: Performance metrics explained. ....	3
Note S2: Other prediction methods. ....	4
Note S3: Viterbi decoder. ....	5
Table S1: Hardware specifications. ....	6
Table S2: Final hyperparameters. ....	6
Table S3: Signal peptide performance. ....	7
Table S4: Performance relative to number of transmembrane segments. ....	7
Table S5: Cross-Validation protein performance. ....	8
Table S6: Cross-Validation TM segment performance. ....	8
Table S7: Cross-Validation confusion matrix. ....	9
Table S8: Ensemble confusion matrix. ....	9
Table S9: Effect of Gaussian filter and Viterbi decoder. ....	10
Table S10: Expected misclassifications. ....	11
Table S11: Out-of-distribution protein performance. ....	12
Table S12: Out-of-distribution segment performance. ....	13
Table S13: Segment performance on new membrane proteins. ....	14
Figure S1: TMbed model architecture. ....	15
Figure S2: Viterbi decoder state transitions. ....	16
Figure S3: Nested Cross-Validation. ....	17
Figure S4: Segment validation criteria. ....	18
Figure S5: More potential transmembrane proteins in the globular data set. ....	19
Figure S6: Out-of-distribution segment length statistics. ....	20
REFERENCES FOR SUPPORTING ONLINE MATERIAL.....	21

## Short description of Supporting Online Material

Here, we provide short explanations for the different performance metrics used during evaluation (Note S1, Fig. S4); how we created the results for several other prediction methods (Note S2); we briefly explain the idea behind using a simple Viterbi decoder and its limitations (Note S3, Fig. S2); and we provide illustrative sketches of our model architecture (Fig. S1) and nested cross-validation process (Fig. S3).

We list the hardware specifications of the machines used during the project (Table S1) and the optimal hyperparameters for each of the final models (Table S2). Further, we provide performance statistics for signal peptides (Table S3) and protein groups based on their number of transmembrane segments (Table S4). We also list statistics for the individual models and cross-validation splits (Tables S5 & S6), confusion matrices for the cross-validation and final models (Table S7 & S8), the effect of the Gaussian filter and Viterbi decoder on segment performance (Table S9), and estimate the expected number of mistakes made in a hypothetical proteome (Table S10). Additionally, we show a few more “false positives” that might actually be transmembrane proteins (Fig. S5). Finally, we provide performance and annotation statistics for an out-of-distribution data set gathered from DeepTMHMM (Tables S11 & S12, Fig. S6) and a CASP-like data set of novel membrane proteins (Table S13).

## Material

### Note S1: Performance metrics explained.

We evaluated our and other methods using several standard and non-standard performance metrics, listed below. Statistics referring to a specific type of segment (i.e., transmembrane beta strands or helices, and signal peptides) are calculated using only the corresponding subset of proteins. For example, the *precision* and  $Q_{ok}$  values for transmembrane helices take only the 571 alpha helical transmembrane proteins (TMPs) into account, ignoring any false positive predictions made in beta barrel TMPs or globular proteins.

*Recall*, also called Sensitivity, is the percentage of positive samples (proteins or segments) that have been correctly predicted as such. For example, TMbed correctly identified 557 of the 571 alpha helical TMPs, i.e., a *recall* of about 98%.

$$Recall = \frac{Number\ of\ correct\ positive\ predictions}{Number\ of\ positive\ samples} * 100\% \quad (Eq. S1)$$

*Precision* reflects the percentage of positive predictions that are actually correct. Since only 557 of the 584 alpha helical TMPs predicted by TMbed are correct (the other 27 are globular proteins), it has a *precision* of about 95%.

$$Precision = \frac{Number\ of\ correct\ positive\ predictions}{Number\ of\ positive\ predictions} * 100\% \quad (Eq. S2)$$

*False Positive Rate (FPR)* gives the percentage of negatives samples incorrectly predicted as positive ones. For example, 27 out of 5711 globular and beta barrel TMPs incorrectly predicted as alpha helical TMPs correspond to a *FPR* of about 0.5%.

$$FPR = \frac{Number\ of\ incorrect\ positive\ predictions}{Number\ of\ negative\ samples} * 100\% \quad (Eq. S3)$$

$Q_{ok}$  is the percentage of proteins for which all predicted segments of a given type are correct, i.e., segment *recall* and *precision* are both 100% for those proteins. A  $Q_{ok}$  of 79% for beta barrel TMPs corresponds to 45 of 57 proteins that do not have any false positive or false negative predictions of transmembrane beta strands.

$$Q_{ok} = \frac{Number\ of\ proteins\ with\ all\ segment\ of\ type\ T\ correct}{Number\ of\ proteins\ with\ segments\ of\ type\ T} * 100\% \quad (Eq. S3)$$

$Q_{num}$  shows the percentage of proteins that have the correct number of predicted segments of a given type, regardless of their exact position. TMbed predicted the correct number of transmembrane beta strands in 50 of 57 beta barrel TMPs, i.e., it has a  $Q_{num}$  of about 88%.

$$Q_{num} = \frac{Number\ of\ proteins\ with\ correct\ number\ of\ segments\ of\ type\ T}{Number\ of\ proteins\ with\ segments\ of\ type\ T} * 100\% \quad (Eq. S4)$$

$Q_{top}$  gives the percentage of correctly predicted segments of a given type that also have the correct inside/outside orientation, i.e., its endpoints are on the correct sides of the membrane. For example, TMbed correctly predicts 730 of 768 transmembrane beta strands (*recall* of about 95%). Of those 730 segments, 714 also have the correct inside/outside orientation, i.e., the  $Q_{top}$  value is about 98%. We consider only the first residue on each side of a segment to determine its orientation.

$$Q_{top} = \frac{\text{Number of correct segments of type } T \text{ with correct orientation}}{\text{Number of correct segments of type } T} * 100\% \quad (\text{Eq. S5})$$

We estimate the error margin of our performance values with the 95% confidence interval (CI), i.e., 1.96 times the standard error (SE) based on the sample standard deviation (SD):

$$CI = 1.96 * SE; \quad SE = \frac{SD}{\sqrt{N}}; \quad SD = \sqrt{\frac{1}{N-1} * \sum_{i=1}^N (x_i - \bar{x})^2}, \quad (\text{Eq. S6-S8})$$

where  $N$  is the number of measurements  $x_i$  performed and  $\bar{x}$  is the mean over those. In our case,  $N$  usually refers to the five cross-validation iterations.

## Note S2: Other prediction methods.

In order to put the performance of TMbed into context, we made predictions for the proteins in our data set using several other methods.

DeepTMHMM (1) uses ESM-1b (2) embeddings to predict alpha helical and beta barrel transmembrane proteins (TMP). We generated all predictions using a local installation as described on their web server homepage (<https://dtu.biolib.com/DeepTMHMM>).

TOPCONS2 (3), OCTOPUS (4), Philius (5), PolyPhobius (6), and SPOCTOPUS (7) all predict alpha helical TMPs. TOPCONS2, Philius, PolyPhobius, and SPOCTOPUS additionally predict signal peptides. With the exception of Philius, all other five methods use evolutionary information in the form of BLAST profiles or MSAs as additional input to the protein sequence. As TOPCONS2 is a consensus prediction method combining all of the above methods, we got all predictions from its web server (<https://topcons.net>). Unfortunately, the web server rejected one of the globular proteins, [P05790](#), due to its high GA content (incorrectly thought to be a DNA sequence).

CCTOP (8, 9) is another consensus prediction method for alpha helical TMPs. It combines a total of 10 prediction methods and topology constraint determined by a homology lookup. We used their web server (<https://cctop.ttk.hu/>) to generate predictions for our data sets. Due to sequence length restrictions (up to 5,000 residues) we are missing predictions for one alpha helical TMP and six globular proteins.

SCAMPI2 (10) is an improved version of the older SCAMPI (11) method employed as part of TOPCONS2. We downloaded the software from its [GitHub repository](#)<sup>1</sup> and used UniRef90 as the BLAST search database to generate the alignments needed for the MSA version of SCAMPI2.

HMM-TM (12) and PRED-TMBB2 (13) are methods predicting alpha helical and beta barrel TMPs, respectively. We computed predictions for our data set using their respective web servers (<http://www.compgen.org/tools>). For both methods, we used the most recent improvements employing hidden neural networks (14). As the online services only allow batch submissions for the single sequence versions, i.e., without the use of MSAs as input, we also installed local versions of their methods (15) and

<sup>1</sup> <https://github.com/ElofssonLab/scampi2>

ran them offline. However, the offline version of PRED-TMBB2 does not include the protein filtering using pHMMs that the web server employs, significantly increasing its false positive rate. Unfortunately, the local MSA versions failed for some of the proteins and we were unable to fix the issue. Thus, we are missing predictions for 155 proteins (6 beta barrel TMPs, 19 alpha helical TMPs, 130 globular proteins) by HMM-TM (MSA) and 26 proteins (2 alpha helical TMPs, 24 globular proteins) by PRED-TMBB2 (MSA).

The authors of BetAware-Deep (16) kindly provided us with predictions for our data set as the web server only allows for submission of a single sequence at a time. Their method combines sequence profiles with several machine learning architectures (LSTM, CRF) to predict beta barrel TMPs.

We installed and ran an offline version of BOCTOPUS2 (17) to predict beta barrel TMPs in our data set. We generated the sequence profiles and results according to the descriptions on their [GitHub repository](#)<sup>2</sup>.

For TMSEG (18) and PROFtmb (19) we used the predictions generated by our PredictProtein (20) pipeline. TMSEG and PROFtmb predict alpha helical and beta barrel TMPs, respectively, both using BLAST profiles as additional input.

We used the SignalP 6.0 (21) [web server](#)<sup>3</sup> to generate additional signal peptide predictions. We chose the “slow” model mode to get accurate predictions. Just like TMbed, SignalP 6.0 uses a protein language model to generate embeddings (22).

---

### Note S3: Viterbi decoder.

We use an untrained Viterbi decoder to translate the class probability distributions generated by our models into actual class labels for each residue in a sequence. The decoder scores state transitions according to the class probabilities predicted by the CNN model, trying to find the path with the highest sum of probabilities. We apply a score penalty of  $-100$  to transitions not intended by our defined grammar (Fig. S2), effectively preventing the decoder from considering those transitions.

The main purpose of the decoder is to enforce a small set of rules:

- 1) Signal peptides may only start at the N-terminus of a sequence.
- 2) Signal peptides and transmembrane segments must be at least five residues long.
- 3) The inside/outside orientation of non-membrane parts must change after every transmembrane segment.

We explicitly model the transition from IN to OUT state to allow for sequence parts that pass the membrane boundaries without actually being in contact with the membrane. For example, this includes parts of beta barrel TMPs that pass through the pore formed by their own beta barrel structure (Manuscript: Fig. 1).

---

<sup>2</sup> <https://github.com/ElofssonLab/boctopus2>

<sup>3</sup> <https://services.healthtech.dtu.dk/service.php?SignalP-6.0>

A downside to this is that the model is free to change the IN/OUT state to accommodate transmembrane segments even if the majority of the non-membrane residues on both sides of the segment are on the same side of the membrane, circumventing rule 3.

For example: `...iiiiHHHHHHHHHHHoiiii...`

However, disallowing the direct transitions between IN and OUT would prevent correctly modelling sequences where such transitions do happen and can encourage the model to split transmembrane segments to insert a single non-membrane residue, thereby accommodating the same orientation on both sides of the original segment.

For example: `...iiiiHHHHHHoHHHHHiiii...`

Given those two alternatives, we decided to allow for direct transitions.

**Table S1: Hardware specifications.**

	CPU	RAM	GPU	VRAM	Storage
<b>Desktop Machine</b>	Intel Core i5-2500K	24GB DDR3	NVIDIA GeForce RTX 3060	12GB GDDR6	SSD
<b>Server Machine</b>	Intel Xeon Gold 6248	400GB DDR4 ECC	NVIDIA Quadro RTX 8000	48GB GDDR6	SSD

\* List of most relevant hardware components in the two machines used during method development. We used the server machine to create sequence embeddings, while training and testing our new method on the Desktop machine.

**Table S2: Final hyperparameters.**

	Learning rate	Weight decay
<b>Model 0</b>	0.005	0.01
<b>Model 1</b>	0.010	0.10
<b>Model 2</b>	0.005	0.10
<b>Model 3</b>	0.010	0.10
<b>Model 4</b>	0.010	0.10

\* Optimal learning rate and weight decay values selected for each model during nested cross-validation. Numbers in the model name indicate the test set; for example, model 0 was trained and optimized on sets 1-4, and evaluated on set 0.

**Table S3: Signal peptide performance.**

	Proteins (661 vs. 4993)		Segments (661)	
	Recall (%)	FPR (%)	Recall (%)	Precision (%)
<b>TMbed</b>	98.8±0.8	<b>0.1±0.1</b>	93.3±1.4	94.5±1.3
<b>SignalP 6.0</b>	97.0±1.0	0.2±0.1	92.6±1.7	95.5±1.7
<b>DeepTMHMM</b>	<b>99.2±0.5</b>	0.2±0.2	<b>95.3±1.8</b>	<b>96.0±1.7</b>
<b>TMSEG</b>	95.0±1.5	5.3±0.4	77.4±3.7	81.5±3.7
<b>TOPCONS2<sup>1</sup></b>	93.9±1.7	2.3±0.5	81.8±3.0	87.1±2.3
<b>Philius<sup>1</sup></b>	93.3±1.6	5.9±0.6	84.5±1.8	90.6±1.7
<b>PolyPhobius<sup>1</sup></b>	92.4±1.7	1.9±0.3	84.1±2.7	91.0±1.8
<b>SPOCTOPUS<sup>1</sup></b>	93.5±1.8	3.6±0.5	85.3±2.7	91.2±1.5

\* Protein and segment performance for signal peptide (SP) prediction based on 661 globular proteins with SPs and 4993 globular proteins without SPs. Performance values were averaged over the five independent cross-validation test sets; error margins given for the 95% confidence interval (1.96\*standard error); **bold**: best values for each column; *italics*: differences statistically significant with over 95% confidence (only computed between best and 2nd best).

<sup>1</sup> Evaluation includes only 660 of the 661 globular proteins with SPs due to one sequence being rejected by the prediction web server.

**Table S4: Performance relative to number of transmembrane segments.**

	TMB		TMH		
	2, 4 (13)	8+ (44)	1 (164)	2-5 (175)	6+ (232)
<b>Q<sub>ok</sub> (%)</b>	65.0±36.7	85.6±8.6	78.1±3.1	61.0±5.5	52.7±3.2
<b>Recall (%)</b>	65.0±36.7	96.6±4.4	79.3±4.5	84.1±1.9	90.6±0.7
<b>Precision (%)</b>	75.0±38.0	99.5±0.6	83.4±2.7	83.8±2.3	90.3±0.8

\* TMbed segment performance for transmembrane beta strand (TMB) and helix (TMH) prediction based on 57 beta barrel and 571 alpha helical TMPs. TMPs are subdivided into groups based on their number of transmembrane segments: a) 2 or 4 TMBs, b) 8 or more TMBs, c) a single TMH, d) 2-5 TMHs, and e) 6 or more TMHs. The numbers in parenthesis indicate the number of proteins within that group. Performance values were averaged over the five independent cross-validation test sets; error margins given for the 95% confidence interval (1.96\*standard error).

**Table S5: Cross-Validation protein performance.**

	$\beta$ -TMP		$\alpha$ -TMP		Globular	
	Recall (%)	FPR (%)	Recall (%)	FPR (%)	Recall (%)	FPR (%)
<b>Model 0</b>	85.7	0.1	97.5	0.2	99.7	3.7
<b>Model 1</b>	100.0	0.0	97.3	0.7	99.3	2.5
<b>Model 2</b>	83.3	0.0	96.5	0.7	99.3	4.8
<b>Model 3</b>	100.0	0.2	98.2	0.5	99.3	1.6
<b>Model 4</b>	100.0	0.0	98.3	0.3	99.7	1.6
<b>Average <math>\pm 1.96SE</math></b>	<i>93.8<math>\pm</math>7.5</i>	<i>0.1<math>\pm</math>0.1</i>	<i>97.5<math>\pm</math>0.7</i>	<i>0.5<math>\pm</math>0.2</i>	<i>99.5<math>\pm</math>0.2</i>	<i>2.8<math>\pm</math>1.2</i>

\* Protein prediction performance for each TMbed model on the corresponding independent test set. Numbers in the model name indicate the test set; for example, model 0 was trained on sets 1-4 and evaluated on set 0. Last row lists the average over all five sets and 1.96 times the sample standard error, i.e., the 95% confidence interval.

**Table S6: Cross-Validation TM segment performance.**

	TMB			TMH		
	Q <sub>ok</sub> (%)	Recall (%)	Precision (%)	Q <sub>ok</sub> (%)	Recall (%)	Precision (%)
<b>Model 0</b>	64.3	95.6	98.5	56.7	88.7	88.4
<b>Model 1</b>	83.3	87.1	98.4	68.2	89.6	89.3
<b>Model 2</b>	75.0	94.9	99.2	61.1	88.4	89.3
<b>Model 3</b>	80.0	97.6	100.0	62.8	87.8	87.5
<b>Model 4</b>	100.0	100.0	100.0	63.5	88.9	89.1
<b>Average <math>\pm 1.96SE</math></b>	<i>80.5<math>\pm</math>11.4</i>	<i>95.0<math>\pm</math>3.4</i>	<i>99.2<math>\pm</math>0.7</i>	<i>62.4<math>\pm</math>3.7</i>	<i>88.7<math>\pm</math>0.6</i>	<i>88.7<math>\pm</math>0.7</i>

\* Segment prediction performance for each TMbed model on the corresponding independent test set. Numbers in the model name indicate the test set; for example, model 0 was trained on sets 1-4 and evaluated on set 0. Last row lists the average over all five sets and 1.96 times the sample standard error, i.e., the 95% confidence interval.



**Table S7: Cross-Validation confusion matrix.**

		Predicted Label			
		$\beta$ -TMP	$\alpha$ -TMP	SP	No SP
True Label	$\beta$ -TMP	53	0	3	1
	$\alpha$ -TMP	0	557	4	10
	SP	3	9	643	6
	No SP	0	18	6	4969

\* Aggregated confusion matrix for the five TMbed models on the corresponding independent test sets. Protein categories are beta barrel TMPs ( $\beta$ -TMP), alpha helical TMPs ( $\alpha$ -TMP), globular proteins with signal peptides (SP) and globular proteins without signal peptides (No SP).

**Table S8: Ensemble confusion matrix.**

		Predicted Label			
		$\beta$ -TMP	$\alpha$ -TMP	SP	No SP
True Label	$\beta$ -TMP	56	0	1	0
	$\alpha$ -TMP	0	567	0	4
	SP	3	1	656	1
	No SP	0	7	4	4982

\* Confusion matrix for the final TMbed ensemble on the complete data set, i.e., for every sequence there are four models that have seen it during training and one model that has not. Protein categories are beta barrel TMPs ( $\beta$ -TMP), alpha helical TMPs ( $\alpha$ -TMP), globular proteins with signal peptides (SP) and globular proteins without signal peptides (No SP).

**Table S9: Effect of Gaussian filter and Viterbi decoder.**

	TMB			TMH		
	Q <sub>ok</sub> (%)	Recall (%)	Precision (%)	Q <sub>ok</sub> (%)	Recall (%)	Precision (%)
<b>TMbed</b>	80.5±11.4	95.0±4.3	99.2±0.7	62.4±3.7	88.7±0.6	88.7±0.7
<b>Viterbi</b>	77.7±15.9	95.4±4.0	99.1±1.1	61.6±3.5	88.9±0.8	88.7±0.8
<b>Gaussian</b>	79.1±13.5	95.3±3.8	98.2±1.6	57.6±4.3	86.1±0.7	84.7±1.2
<b>CNN</b>	45.5±13.9	94.7±4.0	92.0±6.2	50.0±4.7	87.4±1.0	79.6±2.5

\* Comparison between the CNN model, models combining the CNN with either the Gaussian filter or the Viterbi decoder, and the final TMbed model combining all three components. Segment performances for transmembrane beta strand (TMB) and helix (TMH) prediction are based on 57 beta barrel and 571 alpha helical TMPs. Performance values were averaged over the five independent cross-validation test sets; error margins given for the 95% confidence interval (1.96\*standard error).

**Table S10: Expected misclassifications.**

	Misclassifications	
	$\beta$ -TMP	$\alpha$ -TMP
<b>TMbed</b>	22	193
<b>DeepTMHMM</b>	53 (+31)	278 (+85)
<b>TMSEG</b>		521 (+328)
<b>TOPCONS2</b>		683 (+490)
<b>OCTOPUS</b>		1666 (+1473)
<b>Philius</b>		766 (+573)
<b>PolyPhobius</b>		933 (+740)
<b>SPOCTOPUS</b>		2701 (+2508)
<b>SCAMPI2 (MSA)</b>		1135 (+942)
<b>CCTOP</b>		744 (+551)
<b>HMM-TM (MSA)</b>		3352 (+3159)
<b>BOCTOPUS2</b>	880 (+858)	
<b>BetAware-Deep</b>	984 (+962)	
<b>PRED-TMBB2</b>	1573 (+1551)	
<b>PROFtmb</b>	1049 (+1027)	

\* Number of expected misclassified proteins in a hypothetical proteome with 20,000 proteins. Misclassifications are the sum of false positive and false negative predictions for the specific tasks of predicting proteins with transmembrane alpha helices ( $\alpha$ -TMP) or beta strands ( $\beta$ -TMP). The number in parentheses is the difference to our method, TMbed. Expected prediction error rates for the methods are based on their confusion matrices on our data set of 6282 proteins. The hypothetical proteome contains 5000 (25%) alpha helical and 200 (1%) beta barrel transmembrane proteins.

**Table S11: Out-of-distribution protein performance.**

	$\beta$ -TMP		$\alpha$ -TMP		Globular	
	Recall (%)	FPR (%)	Recall (%)	FPR (%)	Recall (%)	FPR (%)
<b>TMbed</b>	<b>100.0±0.0</b>	0.6±0.6	95.5±4.5	<b>1.2±0.9</b>	98.1±1.2	3.9±3.9
<b>DeepTMHMM</b>	<b>100.0±0.0</b>	<b>0.5±0.5</b>	95.4±4.4	<b>1.2±0.9</b>	<b>98.2±1.1</b>	3.9±3.8
<b>TOPCONS2</b>	-	-	98.8±2.4	6.8±2.0	93.2±2.0	1.2±2.4
<b>CCTOP<sup>1</sup></b>	-	-	<b>98.9±2.2</b>	5.5±1.9	94.5±1.9	1.1±2.2
<b>BOCTOPUS2</b>	<b>100.0±0.0</b>	5.9±1.8	-	-	94.1±1.8	<b>0.0±0.0</b>

\* Evaluation of the ability to distinguish between 14 beta barrel TMPs ( $\beta$ -TMP), 86 alpha helical TMPs ( $\alpha$ -TMP) and 567 globular, water-soluble non-TMP proteins from the DeepTMHMM data set; all proteins are non-redundant with respect to the TMbed data sets. Recall and false positive rate (FPR) were averaged over 1000 bootstrap iterations (random sampling with replacement); error margins given for the 95% confidence interval ( $1.96 \times$  standard deviation); **bold**: best values for each column; *italics*: differences statistically significant with over 95% confidence (only computed between best and 2<sup>nd</sup> best, or all methods ranked 1 and those ranked lower).

<sup>1</sup> Evaluation missing for one of 85  $\alpha$ -TMPs.

**Table S12: Out-of-distribution segment performance.**

	TMB			TMH		
	Q <sub>ok</sub> (%)	Recall (%)	Precision (%)	Q <sub>ok</sub> (%)	Recall (%)	Precision (%)
<b>TMbed</b>	64.2±25.7	93.3±5.3	93.3±5.3	52.6±10.8	82.8±4.7	83.8±4.8
<b>DeepTMHMM</b>	<b>78.8±22.4</b>	<b>98.2±1.8</b>	<b>98.2±1.8</b>	46.8±10.4	79.8±5.5	80.0±5.6
<b>TOPCONS2</b>	-	-	-	<b>65.0±10.0</b>	<b>88.9±3.7</b>	<b>89.4±3.7</b>
<b>CCTOP<sup>1</sup></b>	-	-	-	63.5±10.0	<b>88.9±4.0</b>	89.0±3.9
<b>BOCTOPUS2</b>	43.0±26.1	91.0±5.3	92.0±4.9	-	-	-
<b>OPM<sup>2</sup></b>	61.1±14.3	92.0±6.4	91.6±6.4	52.7±7.2	83.9±2.9	83.3±3.0

\* Segment performance for transmembrane beta strand (TMB) and helix (TMH) prediction based on 14 beta barrel and 86 alpha helical TMPs from the DeepTMHMM data set; all proteins are non-redundant with respect to the TMbed data sets. Q<sub>ok</sub>, recall and precision were averaged over 1000 bootstrap iterations (random sampling with replacement); error margins given for the 95% confidence interval (1.96\*standard deviation); **bold**: best values for each column; *italics*: differences statistically significant with over 95% confidence (only computed between best and 2<sup>nd</sup> best, or all methods ranked 1 and those ranked lower; ignores the OPM baseline).

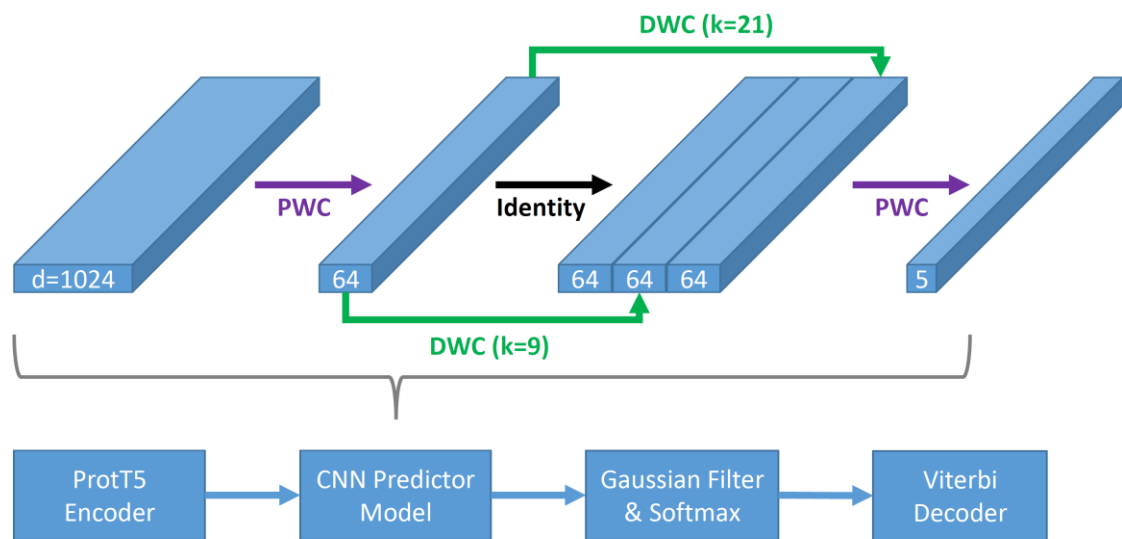
<sup>1</sup> Evaluation missing for one of 85  $\alpha$ -TMPs.

<sup>2</sup> OPM represents the baseline for how much the DeepTMHMM data set annotations agree with our annotations collected from the OPM database, i.e. we are using the OPM annotations as predictions for the DeepTMHMM data set. The performance statistics were evaluated for a set of 44 beta barrel and 184 alpha helical TMPs common to both data sets (TMbed and DeepTMHMM).

**Table S13: Segment performance on new membrane proteins.**

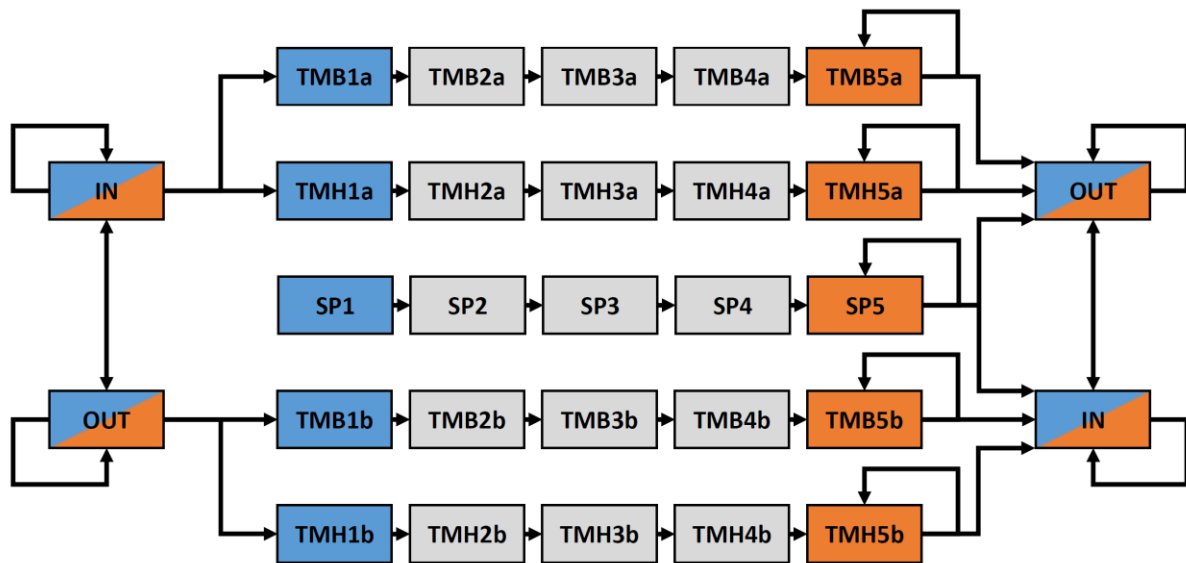
	TMB			TMH		
	Q <sub>ok</sub> (%)	Recall (%)	Precision (%)	Q <sub>ok</sub> (%)	Recall (%)	Precision (%)
<b>TMbed</b>	<b>0.0</b>	<b>93.3</b>	<b>100.0</b>	<b>25.0</b>	<b>60.0</b>	<b>62.5</b>
<b>DeepTMHMM</b>	<b>0.0</b>	86.7	92.9	<b>25.0</b>	40.0	43.5
<b>TOPCONS2</b>	-	-	-	<b>25.0</b>	52.0	48.1
<b>CCTOP</b>	-	-	-	<b>25.0</b>	48.0	50.0
<b>BOCTOPUS2</b>	<b>0.0</b>	0.0	0.0	-	-	-

\* Segment performance for transmembrane beta strand (TMB) and helix (TMH) prediction based on one beta barrel and four alpha helical TMPs; all proteins are non-redundant with respect to the TMbed data sets and the DeepTMHMM data set. Error margins omitted due to extremely small data set size; none of the performance differences are statistically significant; **bold**: best values for each column.



**Figure S1: TMbed model architecture.**

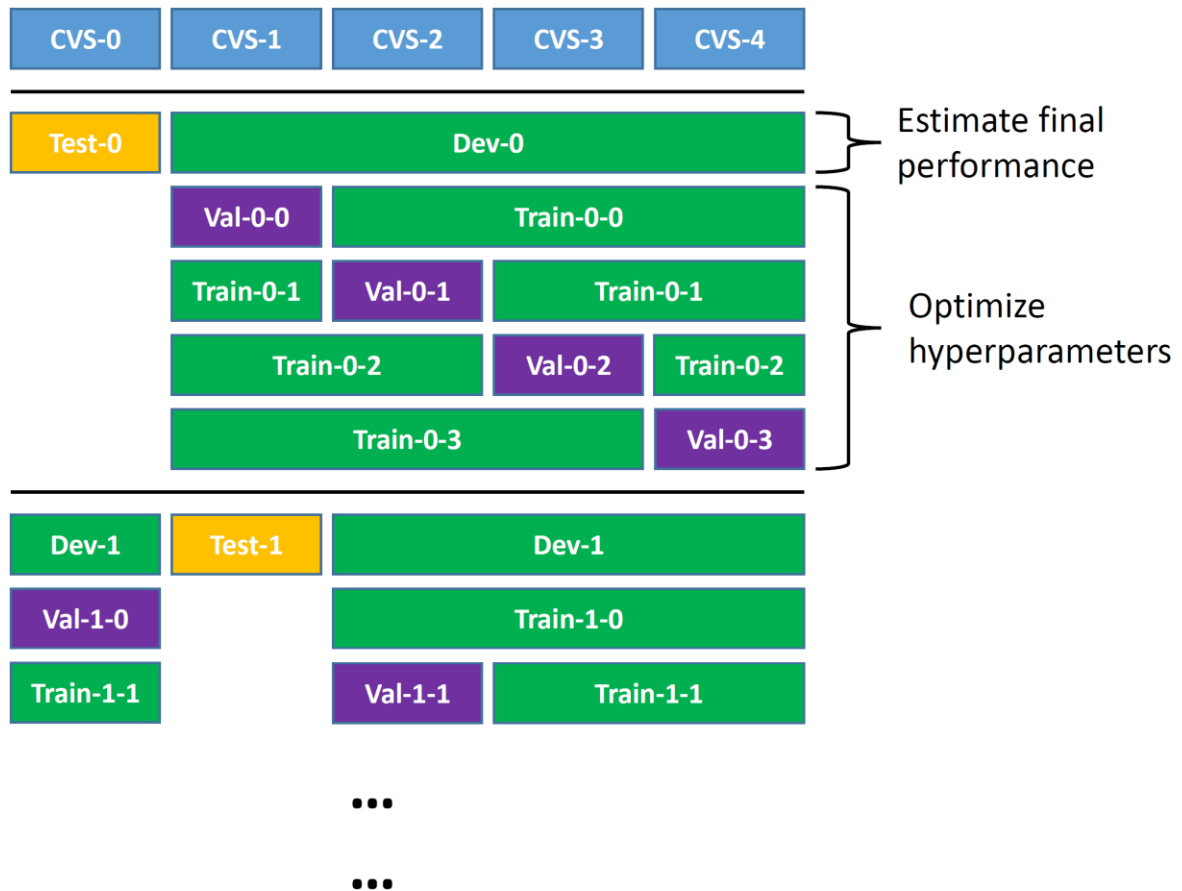
The TMbed model consists of four parts: a) The ProtT5 encoder converts the input sequence into per-residue embeddings with 1024 dimensions for each residue in the sequence; b) a convolutional neural network (CNN) predicts class scores based on those embeddings; c) a Gaussian filter smooths the class scores and converts them into class probabilities via the softmax function; d) a Viterbi decoder assigns class labels to each of the residues in the sequence. The CNN consists of four layers: two pointwise convolutions (PWC) and two depthwise convolutions (DWC; kernel sizes of 9 and 21). The output of the first PWC and both DWCs also passes through layer normalization and a ReLU activation function.



**Figure S2: Viterbi decoder state transitions.**

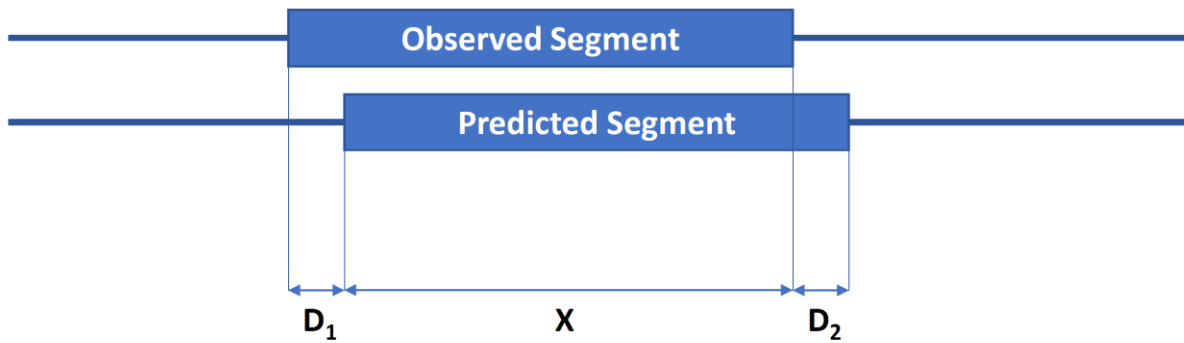
Transitions encoded in the Viterbi decoder to go from one state to another state. We split transmembrane beta strands (TMB), helices (TMH), and signal peptides (SP) into sub-states to enforce minimum segment lengths of five residues. A decoded sequence must start with one of the blue states and may only end with one of the orange states. The IN and OUT states on both sides represent the same two internal states and are only duplicated to simplify the graph.





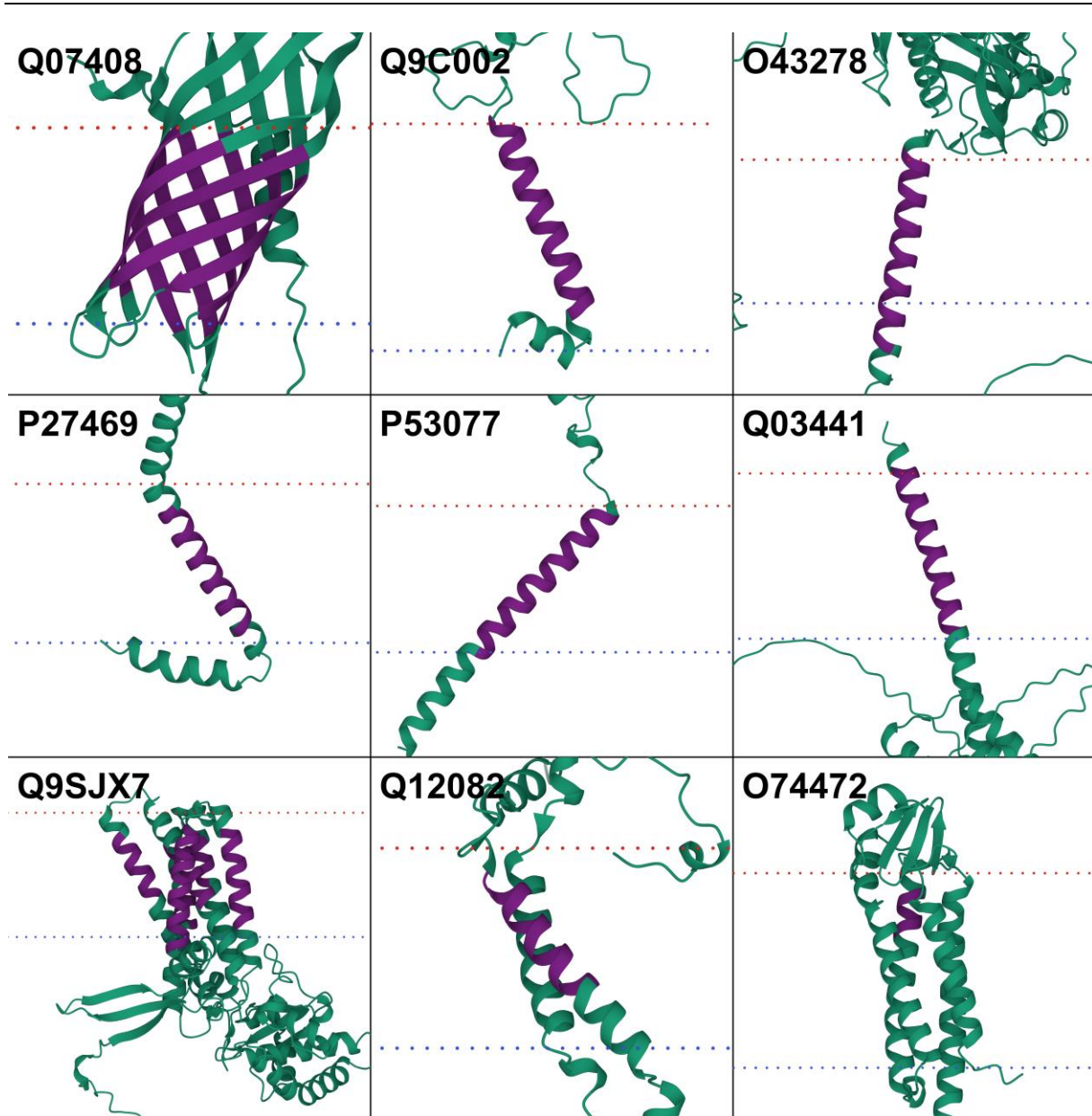
**Figure S3: Nested Cross-Validation.**

For the nested cross-validation process, we split the data set into five cross-validation splits (CVS). During each of the outer five iterations, we used one split as the test set to estimate the models final performance and the other four to develop the model. We further divided those four splits into training set and validation set. We then trained the model on the training set and evaluated the performance on the validation set, repeating the process for each hyperparameter combination. We repeated this process three more times, each time using a different split for the validation set. We chose the best set of hyperparameters based on the average performance on all four validation sets, trained the model using those parameters on the development set, and evaluated its performance on the test set. We repeated this overall process four more times, each time choosing a different CVS for the test set, until each CVS had been used for testing once. This process yielded five trained models, which we used for the final TMbed ensemble.



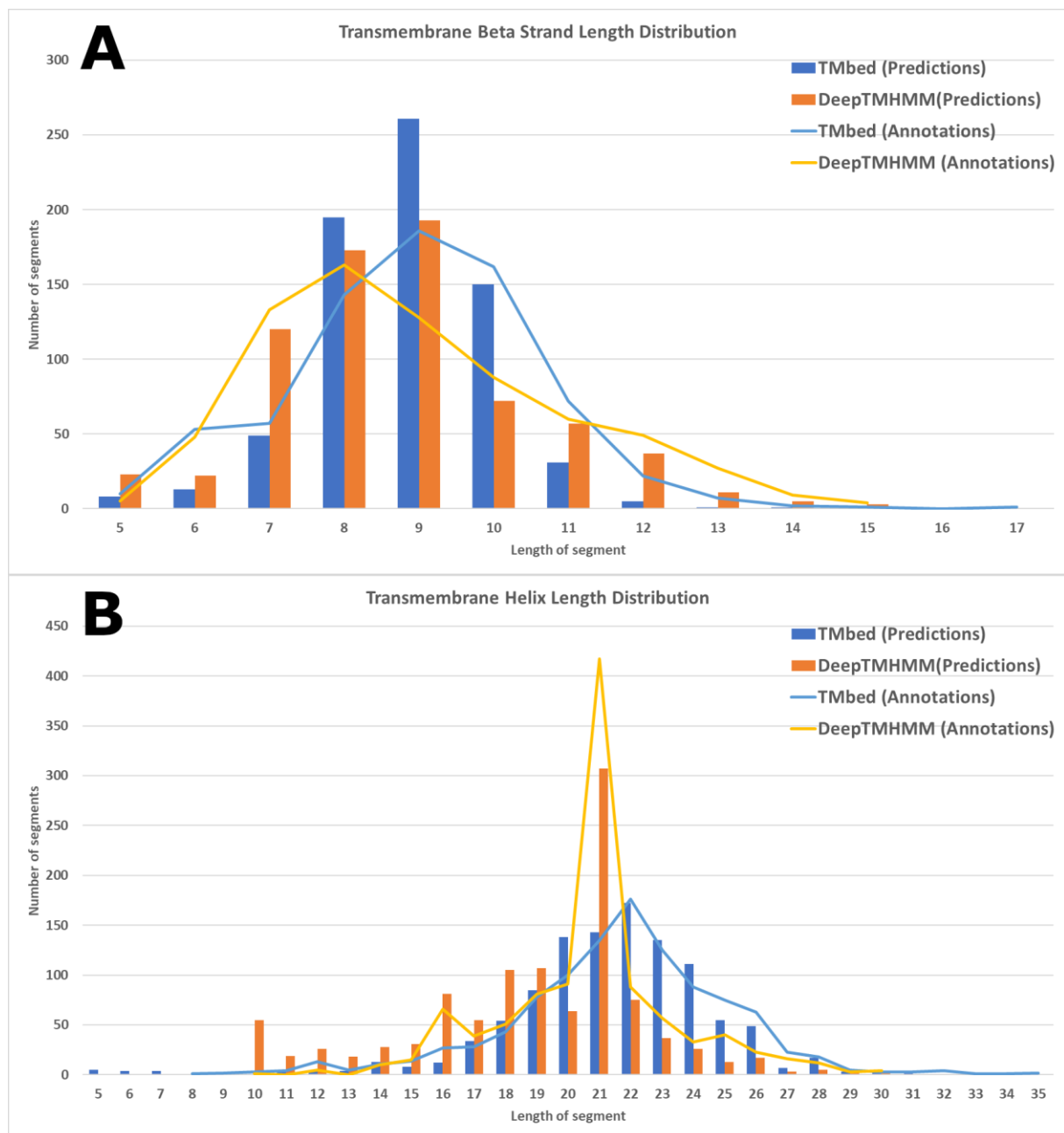
**Figure S4: Segment validation criteria.**

Illustration of the two criteria for a predicted segment to be correct: 1) start and end positions must not deviate by more than five residues, i.e.,  $\max(D_1, D_2) \leq 5$ , and 2) the intersection (overlap) between the observed and predicted segment must be at least half of their union, i.e.,  $\frac{X}{X+D_1+D_2} \geq 0.5$ .



**Figure S5: More potential transmembrane proteins in the globular data set.**

AlphaFold2 (23, 24) structures of nine proteins from the globular data set: major surface antigen 4 ([Q07408](#)), normal mucosa of esophagus-specific gene 1 protein ([Q9C002](#)), Kunitz-type protease inhibitor 1 ([O43278](#)), G0/G1 switch protein 2 ([P27469](#)), maintenance of telomere capping protein 3 ([P53077](#)), sporulation protein RMD1 ([Q03441](#)), protein root UVB sensitive 2 ([Q9SJX7](#)), uncharacterized protein YDL157C ([Q12082](#)), and meiotically up-regulated gene 33 protein ([O74472](#)). For most proteins, transmembrane segments (dark purple) predicted by TMbed correlate well with membrane boundaries (dotted lines: red=outside, blue=inside) predicted by the PPM (25) web server. Images created using Mol\* Viewer (26). Though our data set lists them as globular proteins, the predicted structures indicate transmembrane domains, which align with segments predicted by our method. Predictions were made with the final TMbed ensemble model.



**Figure S6: Out-of-distribution segment length statistics.**

Transmembrane segments length distributions for 44 beta barrel (**A**) and 184 alpha helical (**B**) transmembrane proteins common to both the TMbed and DeepTMHMM data sets. **Lines:** statistics for the annotated segments in each data set; **Bars:** statistics for the segments predicted by each method during its individual cross-validation. Panel B is cropped to the right, missing two annotated segments (L: 40, 44) and four predicted segments (L: 37, 38, 40, 43) for TMbed.

## References for Supporting Online Material

1. Hallgren J, Tsirigos KD, Pedersen MD, Almagro Armenteros JJ, Marcatili P, Nielsen H, et al. DeepTMHMM predicts alpha and beta transmembrane proteins using deep neural networks. *bioRxiv*. 2022:2022.04.08.487609.
2. Rives A, Meier J, Sercu T, Goyal S, Lin Z, Liu J, et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc Natl Acad Sci U S A*. 2021;118(15).
3. Tsirigos KD, Peters C, Shu N, Kall L, Elofsson A. The TOPCONS web server for consensus prediction of membrane protein topology and signal peptides. *Nucleic Acids Res*. 2015;43(W1):W401-7.
4. Viklund H, Elofsson A. OCTOPUS: improving topology prediction by two-track ANN-based preference scores and an extended topological grammar. *Bioinformatics*. 2008;24(15):1662-8.
5. Reynolds SM, Kall L, Riffle ME, Bilmes JA, Noble WS. Transmembrane topology and signal peptide prediction using dynamic bayesian networks. *PLoS Comput Biol*. 2008;4(11):e1000213.
6. Kall L, Krogh A, Sonnhammer EL. An HMM posterior decoder for sequence feature prediction that includes homology information. *Bioinformatics*. 2005;21 Suppl 1:i251-7.
7. Viklund H, Bernsel A, Skwark M, Elofsson A. SPOCTOPUS: a combined predictor of signal peptides and membrane protein topology. *Bioinformatics*. 2008;24(24):2928-9.
8. Dobson L, Remenyi I, Tusnady GE. CCTOP: a Consensus Constrained TOPology prediction web server. *Nucleic Acids Res*. 2015;43(W1):W408-12.
9. Dobson L, Remenyi I, Tusnady GE. The human transmembrane proteome. *Biol Direct*. 2015;10:31.
10. Peters C, Tsirigos KD, Shu N, Elofsson A. Improved topology prediction using the terminal hydrophobic helices rule. *Bioinformatics*. 2016;32(8):1158-62.
11. Bernsel A, Viklund H, Falk J, Lindahl E, von Heijne G, Elofsson A. Prediction of membrane-protein topology from first principles. *Proc Natl Acad Sci U S A*. 2008;105(20):7177-81.
12. Bagos PG, Liakopoulos TD, Hamodrakas SJ. Algorithms for incorporating prior topological information in HMMs: application to transmembrane proteins. *BMC Bioinformatics*. 2006;7:189.
13. Tsirigos KD, Elofsson A, Bagos PG. PRED-TMBB2: improved topology prediction and detection of beta-barrel outer membrane proteins. *Bioinformatics*. 2016;32(17):i665-i71.
14. Tamposis IA, Sarantopoulou D, Theodoropoulou MC, Stasi EA, Kontou PI, Tsirigos KD, et al. Hidden neural networks for transmembrane protein topology prediction. *Comput Struct Biotechnol J*. 2021;19:6090-7.
15. Tamposis IA, Tsirigos KD, Theodoropoulou MC, Kontou PI, Tsaousis GN, Sarantopoulou D, et al. JUCHMME: a Java Utility for Class Hidden Markov Models and Extensions for biological sequence analysis. *Bioinformatics*. 2019;35(24):5309-12.

16. Madeo G, Savojardo C, Martelli PL, Casadio R. BetAware-Deep: An Accurate Web Server for Discrimination and Topology Prediction of Prokaryotic Transmembrane beta-barrel Proteins. *J Mol Biol.* 2021;433(11):166729.
17. Hayat S, Peters C, Shu N, Tsirigos KD, Elofsson A. Inclusion of dyad-repeat pattern improves topology prediction of transmembrane beta-barrel proteins. *Bioinformatics.* 2016;32(10):1571-3.
18. Bernhofer M, Kloppmann E, Reeb J, Rost B. TMSEG: Novel prediction of transmembrane helices. *Proteins.* 2016;84(11):1706-16.
19. Bigelow HR, Petrey DS, Liu J, Przybylski D, Rost B. Predicting transmembrane beta-barrels in proteomes. *Nucleic Acids Res.* 2004;32(8):2566-77.
20. Bernhofer M, Dallago C, Karl T, Satagopam V, Heinzinger M, Littmann M, et al. PredictProtein - Predicting Protein Structure and Function for 29 Years. *Nucleic Acids Res.* 2021;49(W1):W535-W40.
21. Teufel F, Almagro Armenteros JJ, Johansen AR, Gislason MH, Pihl SI, Tsirigos KD, et al. SignalP 6.0 predicts all five types of signal peptides using protein language models. *Nat Biotechnol.* 2022.
22. Elnaggar A, Heinzinger M, Dallago C, Rehawi G, Wang Y, Jones L, et al. ProtTrans: Towards Cracking the Language of Lifes Code Through Self-Supervised Deep Learning and High Performance Computing. *IEEE Trans Pattern Anal Mach Intell.* 2021;PP.
23. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. *Nature.* 2021;596(7873):583-9.
24. Varadi M, Anyango S, Deshpande M, Nair S, Natassia C, Yordanova G, et al. AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res.* 2022;50(D1):D439-D44.
25. Lomize MA, Pogozheva ID, Joo H, Mosberg HI, Lomize AL. OPM database and PPM web server: resources for positioning of proteins in membranes. *Nucleic Acids Res.* 2012;40(Database issue):D370-6.
26. Sehnal D, Bittrich S, Deshpande M, Svobodova R, Berka K, Bazgier V, et al. Mol\* Viewer: modern web app for 3D visualization and analysis of large biomolecular structures. *Nucleic Acids Res.* 2021;49(W1):W431-W7.