bioRxiv preprint doi: https://doi.org/10.1101/2022.04.08.487609; this version posted April 10, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

DeepTMHMM predicts alpha and beta transmembrane proteins using deep neural networks

Jeppe Hallgren[1,10], Konstantinos D. Tsirigos[2,10], Mads Damgaard Pedersen[1], José Juan Almagro Armenteros[3], Paolo Marcatili[4], Henrik Nielsen[4], Anders Krogh[5,6] and Ole Winther[7,8,9,*]

[1]BioLib Technologies, Copenhagen, Denmark
[2]Department of Energy Conversion and Storage, Technical University of Denmark, Kgs Lyngby, Denmark
[3]Department of Genetics, Stanford University School of Medicine, Stanford, CA, USA
[4]Department of Health Technology, Technical University of Denmark, Kgs Lyngby, Denmark
[5]Department of Computer Science, University of Copenhagen, Copenhagen, Denmark
[6]Center for Health Data Science, University of Copenhagen, Copenhagen, Denmark
[7]Department of Biology, Bioinformatics Center, University of Copenhagen, Denmark
[8]Center for Genomic Medicine, Rigshospitalet (Copenhagen University Hospital), Copenhagen, Denmark
[9]Section for Cognitive Systems, Department of Applied Mathematics and Computer Science, Technical University of Denmark, Kgs Lyngby, Denmark

[10] These authors have contributed equally to the presented work.

[*]Correspondence should be addressed to Ole Winther (ole.winther@bio.ku.dk)

bioRxiv preprint doi: https://doi.org/10.1101/2022.04.08.487609; this version posted April 10, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Abstract

Transmembrane proteins span the lipid bilayer and are divided into two major structural classes, namely alpha helical and beta barrels. We introduce DeepTMHMM, a deep learning protein language model-based algorithm that can detect and predict the topology of both alpha helical and beta barrels proteins with unprecedented accuracy. DeepTMHMM (https://dtu.biolib.com/DeepTMHMM) scales to proteomes and covers all domains of life, which makes it ideal for metagenomics analyses.

Main

Transmembrane (TM) proteins typically account for ~30% in a proteome[1] and are involved in many vital cellular processes[2]. Because of their importance, TM proteins like G-protein coupled receptors (GPCRs), transporters and ion channels pose a great pharmaceutical interest as drug targets[3].

TM proteins are difficult to crystallize and obtain good quality 3D structures of, thus they are greatly underrepresented in PDB[4]. This creates the need for computational tools that can accurately identify them amongst the other types of proteins and predict their topology, i.e., the position and orientation of TM segments in the protein sequence as well as their N-terminal.

Furthermore, given that signal peptides are often falsely predicted as TM segments[5], methods that predict the topology of the protein and the presence of a signal peptide at the same time have been developed. The majority of prokaryotic beta barrels also encode an N-terminal export signal peptide in the translated precursor, which signals secretion across the inner membrane via the Sec translocon machinery[6].

bioRxiv preprint doi: https://doi.org/10.1101/2022.04.08.487609; this version posted April 10, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

TMHMM, one of the first and the most widely used methods since 2001[1], focused on alpha helical TM topology. Here, we introduce DeepTMHMM, which compared to its predecessor, has three main improvements: (i) it has improved topology prediction, (ii) it can predict the presence of a signal peptide and (iii) it can predict the topology of beta-barrels.

DeepTMHMM is based upon a deep learning encoder-decoder sequence-to-sequence model that takes a protein sequence as input and outputs the corresponding per-residue sequence of labels. The per-residue labels are signal peptide (S), inside cell/cytosol (I), alpha membrane (M), beta membrane (B), periplasm (P) and outside cell/lumen of ER/Golgi/lysosomes (O). The sequence of residue labels defines the topology of the protein.

Briefly, the encoder consists of three components: a pre-trained language model (ESM-1b)[7], a bi-directional LSTM and a dense layer with drop-out. The large-scale pre-training task of prediction of masked-out amino acids trained on 250 million protein sequences, equips the encoder's representations of the protein sequence with a lot of implicit structural and evolutionary information, thus easing the need for labeled data derived from TM 3D structures. The encoder's representations are fed into a conditional random field (CRF). The CRF is closely related to the hidden Markov model used in TMHMM. The CRF decoder assigns a probability to the entire output sequence rather than treating each position in the sequence as an independent classification task. Furthermore, by expanding the dimensionality of the state space beyond the five per-residue labels and constraining the learned "interaction" matrix, it is possible to reproduce properties of TM proteins such as the length distribution of segments and let label sequences obey a basic transmembrane

bioRxiv preprint doi: https://doi.org/10.1101/2022.04.08.487609; this version posted April 10, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

grammar. The CRF allows for decoding of the most probable sequence (Viterbi decoding) and calculation of marginal probabilities at each position. We can use either type of decoding as the final topology prediction (see Supplementary Information).

We used five types of proteins for training and testing DeepTMHMM, namely alpha helical transmembrane proteins without a signal peptide (alpha TM), alpha helical transmembrane proteins with signal peptide (SP + alpha TM), beta-barrel transmembrane proteins (Beta), globular proteins with signal peptide (SP + Globular) and globular proteins without signal peptide (Globular). Because of the very limited availability of eukaryotic beta barrel structures, DeepTMHMM is primarily set up to predict the topology of prokaryotic beta barrels. However, small scale testing on the limited eukaryotic beta barrel TM structural data available indicates that DeepTMHMM can correctly identify them and capture the main features of their topology.

In order to account for possible overfitting and over-optimistic assessment of predictive performance, we performed a within-type homology reduction using the CD-HIT[8] algorithm at 30% sequence identity cut-off. This means that, within each of the five types, there are no proteins that share more than >30% sequence identity. The total number of sequences in our training set is 3,574, of which 2,000 Globular, 1,000 SP+Globular, 106 SP+alpha TM, 387 alpha TM, and 81 prokaryotic beta barrels (see Supplementary Material).

We compared DeepTMHMM to several topology prediction methods that are currently available (Supplementary Table 1). The performance of DeepTMHMM was assessed in a five-fold cross validation setup with three folds for training, one for validation and one for testing. The test performance was assessed only once after the completion of training and

bioRxiv preprint doi: https://doi.org/10.1101/2022.04.08.487609; this version posted April 10, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

model optimization steps. For benchmarking we used our full set, which inevitably favors the compared methods *a priori* because of overlap our set with their training sets. Further, only DeepTMHMM predicts all five protein types.

For the overall type classification, DeepTMHMM outperforms all other methods as seen in Figure 2a and Supplementary Tables 2a-c. Only in the case of alpha TM proteins TMHMM is slightly superior. This is mainly because TMHMM is very conservative in predicting TM proteins and only predicts two types. Supplementary Table 2a shows the performance of DeepTMHMM against methods that can operate in a proteome-wide manner, and there, DeepTMHMM shows superior performance. The same applies to beta barrel protein detection, where DeepTMHMM is better than all other methods we tested.

Regarding topology prediction accuracy (Figure 2b), in the alpha TM set, DeepTMHMM performs slightly worse than TOPCONS2[9] and slightly better than CCTOP[10] and DMCTOP[11], which, however include more than 30% of the test set proteins in their training sets (Supplementary Table 4a). All other methods perform substantially worse. In the alpha SP+TM set, DeepTMHMM is better than all other methods, with a marginal improvement over TOPCONS2 and substantial over the remaining methods (Supplementary Table 4b). Finally, for beta barrels, DeepTMHMM ranks first, followed closely by BOCTOPUS2[12] and PRED-TMBB2[13] (whose respective training sets also substantially overlap with ours). DeepTMHMM gives a sensible prediction for the eukaryotic beta barrel proteins as well, like the human voltage-dependent anion channel VDAC (PDB: 2JK4). VDAC[14] has 19 beta strands and DeepTMHMM, which is constrained by the state space model to predict an even number of strands, predicts 18. The

bioRxiv preprint doi: https://doi.org/10.1101/2022.04.08.487609; this version posted April 10, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

prediction overlaps with the ones from the solved structure except the first beta strand that is predicted as signal peptide in accordance with the DeepTMHMM's state space model.

Another substantial improvement that DeepTMHMM brings about is the accurate detection of the cleavage site (CS) of signal peptides (when present). Existing topology prediction methods are good at detecting signal peptides, but not as good in pinpointing their CS. As shown on Figure 2c and Supplementary Tables 5a-d, DeepTMHMM is far better than all other methods in CS detection and even slightly better than SignalP 6.0[15]. DeepTMHMM is also conservative in not over-predicting signal peptides.

As an overall conclusion, we can confidently state that DeepTMHMM is currently the most complete software to facilitate proteome-wide topology predictions of both alpha helical and beta barrel transmembrane proteins.

DeepTMHMM is available at https://dtu.biolib.com/DeepTMHMM/ and allows for up to 10,000 protein sequences per submission. For large-scale analyses, we recommend that users run DeepTMHMM locally on their own machine. DeepTMHMM is free for all academic users and provided for a fee to commercial users.

bioRxiv preprint doi: https://doi.org/10.1101/2022.04.08.487609; this version posted April 10, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

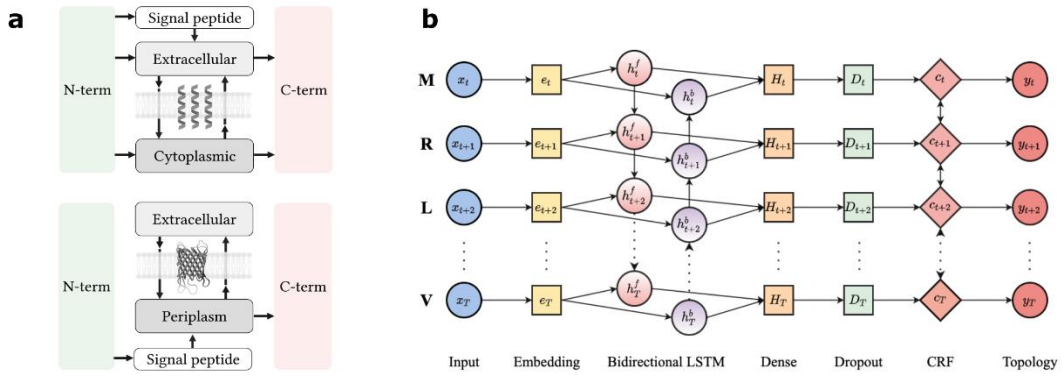


Figure 1: a. State space model of the protein topology for the considered protein types. 1a top generates Alpha TM, SP+TM, Glob and SP+Glob topologies and the bottom plot prokaryotic Beta barrels and SP+Glob. The protein sequence begins in the N-terminus (for clarity split in the plot), ends in the C-terminus and the arrows indicate transitions to other "compartments". The model can stay in each compartment for a number of residues in a pre-specified range. **b**. The DeepTMHMM Neural Network architecture consists of the ESM1-b model (shown as "Embedding"), a bi-directional LSTM, a dense layer with dropout and, finally, a CRF decoder layer.

bioRxiv preprint doi: https://doi.org/10.1101/2022.04.08.487609; this version posted April 10, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

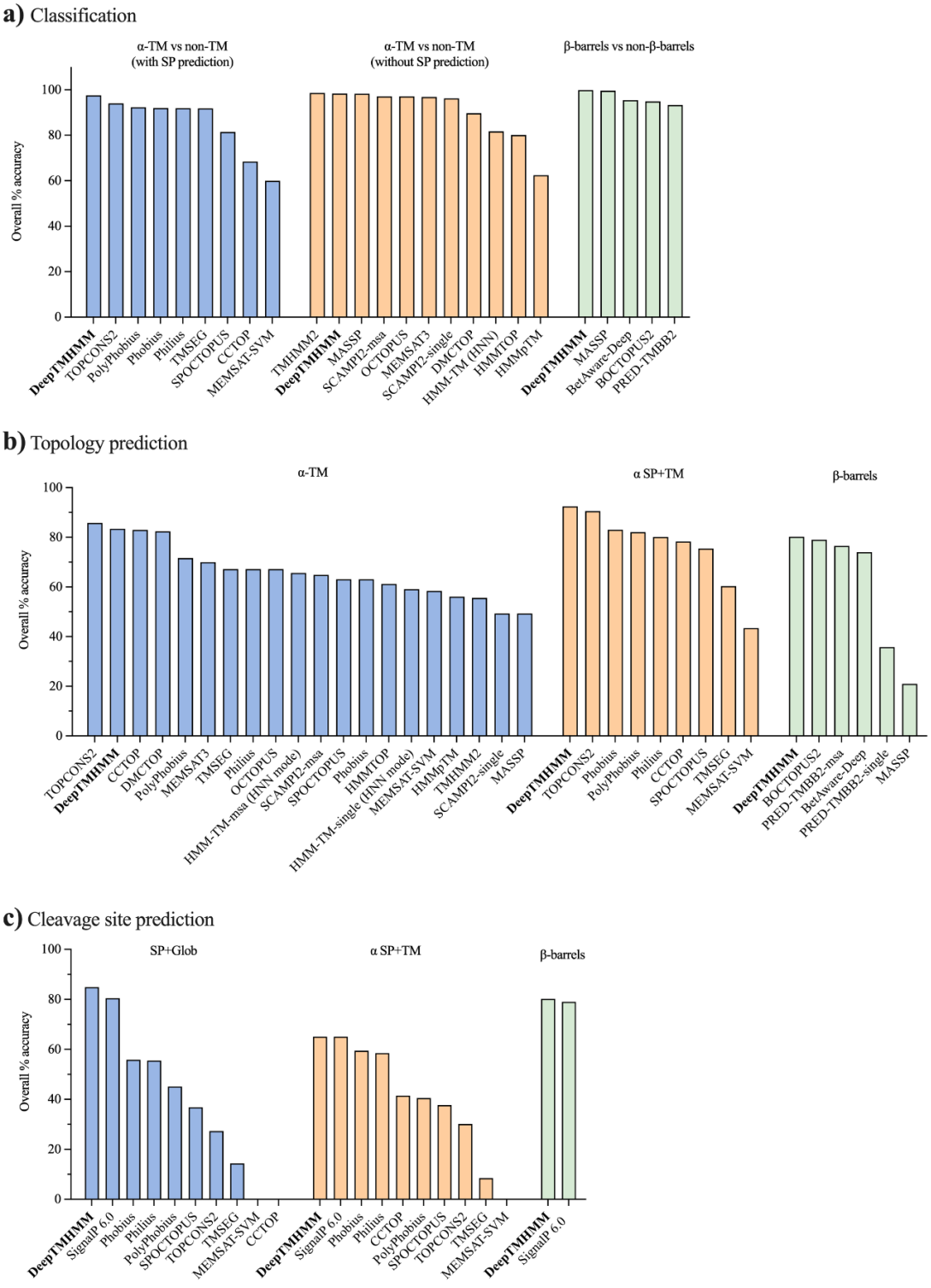


Figure 2: DeepTMHMM accuracy benchmarked against other methods on **(a)** type classification, **(b)** topology prediction and **(c)** signal peptide cleavage site prediction.

bioRxiv preprint doi: https://doi.org/10.1101/2022.04.08.487609; this version posted April 10, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Acknowledgements
The authors would like to thank Stephanie Heusser for assisting with image creation and Felix Teufel and Gunnar von Heijne for useful discussions. A.K and O.W. are supported by the Novo Nordisk Foundation (NNF20OC0062606).

Author contributions
J.H. designed the model architecture and trained the DeepTMHMM method, with input from M.D.P, J.J.A.A and O.W. K.D.T collected the training and test data, performed the benchmarks and analyzed the results. P.M., A.K. and H.N. contributed suggestions during the design of DeepTMHMM. K.D.T and O.W. wrote the paper, with input from J.H. and M.D.P. O.W. supervised and guided the project. All authors edited and approved the manuscript.

Competing financial interests
A version of DeepTMHMM has been commercialized by the Technical University of Denmark - DTU (it is provided for a fee to commercial users). The revenue from these commercial sales is divided between the program developers and DTU.

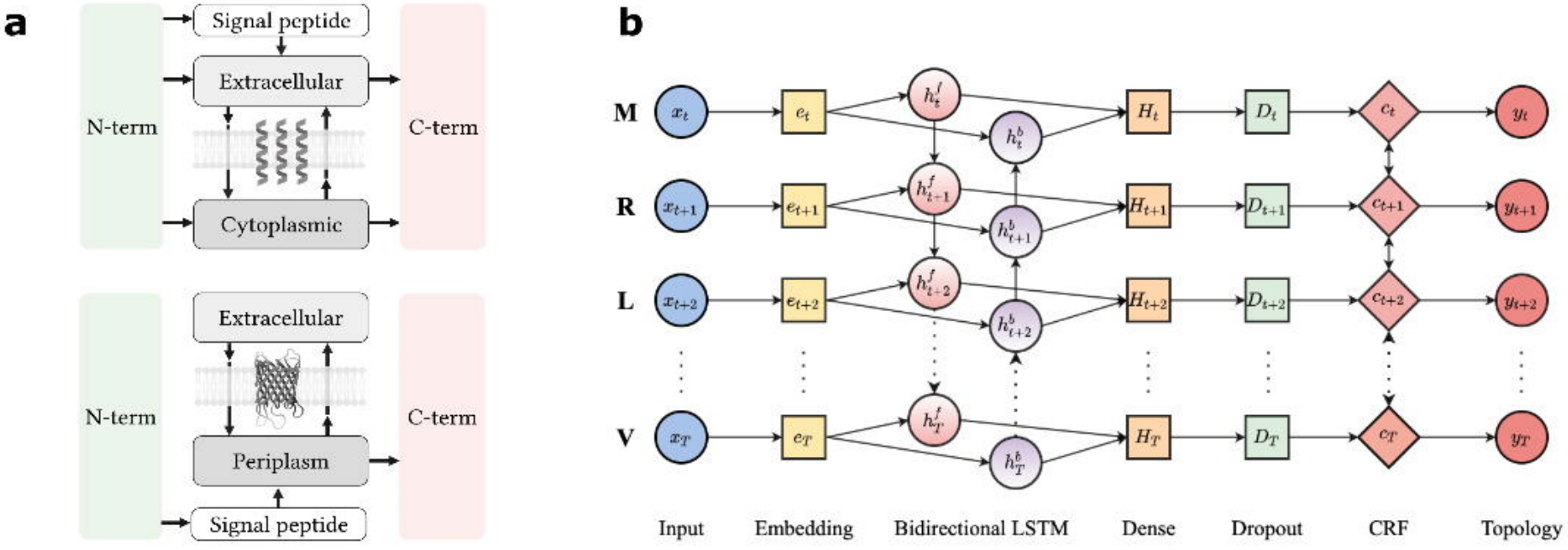
Code and availability
DeepTMHMM is available at https://dtu.biolib.com/DeepTMHMM. DeepTMHMM is free for academic users and is licensed for a fee to commercial users. The dataset used for training DeepTMHMM can be downloaded from https://dtu.biolib.com/DeepTMHMM.

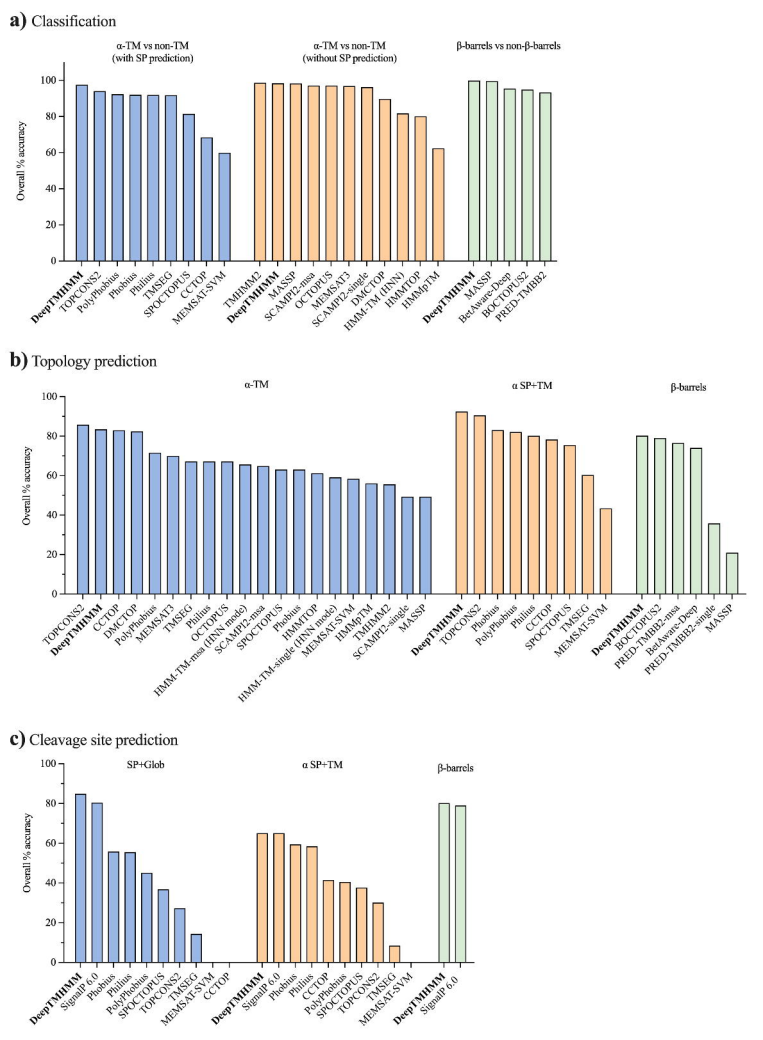
References

1. Krogh, A., Larsson, B., von Heijne, G. & Sonnhammer, E.L. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J Mol Biol* **305**, 567-580 (2001).
2. Tsirigos, K.D. et al. Topology of membrane proteins-predictions, limitations and variations. *Curr Opin Struct Biol* **50**, 9-17 (2018).
3. Gong, J. et al. Understanding Membrane Protein Drug Targets in Computational Perspective. *Curr Drug Targets* **20**, 551-564 (2019).
4. Burley, S.K. et al. RCSB Protein Data Bank: biological macromolecular structures enabling research and education in fundamental biology, biomedicine, biotechnology and energy. *Nucleic Acids Res* **47**, D464-D474 (2019).
5. Lao, D.M., Arai, M., Ikeda, M. & Shimizu, T. The presence of signal peptide significantly affects transmembrane topology prediction. *Bioinformatics* **18**, 1562-1566 (2002).
6. White, S.H. & von Heijne, G. The machinery of membrane protein assembly. *Curr Opin Struct Biol* **14**, 397-404 (2004).
7. Rives, A. et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc Natl Acad Sci U S A* **118** (2021).

bioRxiv preprint doi: https://doi.org/10.1101/2022.04.08.487609; this version posted April 10, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

8. Li, W. & Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658-1659 (2006).
9. Tsirigos, K.D., 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* **43**, W401-407 (2015).
10. Dobson, L., Remenyi, I. & Tusnady, G.E. CCTOP: a Consensus Constrained TOPology prediction web server. *Nucleic Acids Res* **43**, W408-412 (2015).
11. Yang, Y. et al. An Improved Topology Prediction of Alpha-Helical Transmembrane Protein Based on Deep Multi-Scale Convolutional Neural Network. *IEEE/ACM Trans Comput Biol Bioinform* **19**, 295-304 (2022).
12. Hayat, S., Peters, C., Shu, N., Tsirigos, K.D. & Elofsson, A. Inclusion of dyad-repeat pattern improves topology prediction of transmembrane beta-barrel proteins. *Bioinformatics* **32**, 1571-1573 (2016).
13. Tsirigos, K.D., Elofsson, A. & Bagos, P.G. PRED-TMBB2: improved topology prediction and detection of beta-barrel outer membrane proteins. *Bioinformatics* **32**, i665-i671 (2016).
14. Hiller, S. et al. Solution structure of the integral human membrane protein VDAC-1 in detergent micelles. *Science* **321**, 1206-1210 (2008).
15. Teufel, F. et al. SignalP 6.0 predicts all five types of signal peptides using protein language models. *Nat Biotechnol* (2022).


a
Signal peptide
Extracellular
N-term
C-term
Cytoplasmic
Extracellular
N-term
C-term
Periplasm
Signal peptide
b
M
R
L
V
Input
Embedding
Bidirectional LSTM
Dense
Dropout
CRF
Topology


a) Classification
α-TM vs non-TM (with SP prediction)
α-TM vs non-TM (without SP prediction)
β-barrels vs non-β-barrels
Overall % accuracy
DeepTMHMM
TOPCONS2
PolyPhobius
Phobius
Philius
TMSEG
SPOCTOPUS
CCTOP
MEMSAT-SVM
TMHMM2
DeepTMHMM
MASSP
SCAMPI2-msa
OCTOPUS
MEMSAT3
SCAMPI2-single
DMCTOP
HMM-TM (HNN)
HMMTOP
HMMpTM
DeepTMHMM
MASSP
Betaware-Deep
BOCTOPUS2
PRED-TMBB2
b) Topology prediction
α-TM
α SP+TM
β-barrels
Overall % accuracy
TOPCONS2
DeepTMHMM
CCTOP
DMCTOP
PolyPhobius
MEMSAT3
TMSEG
Philius
OCTOPUS
HMM-TM-msa (HNN mode)
SCAMPI2-msa
SPOCTOPUS
Phobius
HMMTOP
HMM-TM-single (HNN mode)
MEMSAT-SVM
HMMpTM
TMHMM2
SCAMPI2-single
MASSP
DeepTMHMM
TOPCONS2
Phobius
PolyPhobius
Philius
CCTOP
SPOCTOPUS
TMSEG
MEMSAT-SVM
DeepTMHMM
BOCTOPUS2
PRED-TMBB2-msa
Betaware-Deep
PRED-TMBB2-single
MASSP
c) Cleavage site prediction
SP+Glob
α SP+TM
β-barrels
Overall % accuracy
DeepTMHMM
SignalP 6.0
Phobius
Philius
PolyPhobius
SPOCTOPUS
TOPCONS2
TMSEG
MEMSAT-SVM
CCTOP
DeepTMHMM
SignalP 6.0
Phobius
Philius
CCTOP
PolyPhobius
SPOCTOPUS
TOPCONS2
TMSEG
MEMSAT-SVM
DeepTMHMM
SignalP 6.0