



Published in final edited form as:

*J Proteome Res.* 2021 August 06; 20(8): 4089–4100. doi:10.1021/acs.jproteome.1c00410.

## A multi-task deep-learning method for predicting membrane associations and secondary structures of proteins

Bian Li<sup>1,2</sup>, Jeffrey Mendenhall<sup>2,3</sup>, John A. Capra<sup>4</sup>, Jens Meiler<sup>2,3,5,\*</sup>

<sup>1</sup>Department of Biological Sciences, Vanderbilt University, Nashville, TN 37203, USA

<sup>2</sup>Center for Structural Biology, Vanderbilt University, Nashville, TN 37203, USA

<sup>3</sup>Department of Chemistry, Vanderbilt University, Nashville, TN 37203, USA

<sup>4</sup>Bakar Computational Health Sciences Institute and Department of Epidemiology and Biostatistics, University of California, San Francisco, CA 94143, USA

<sup>5</sup>Institute for Drug discovery, University Leipzig Medical School, Leipzig, Germany

### Abstract

Prediction of residue-level structural attributes and protein-level structural classes helps model protein tertiary structure and understand protein function. Existing methods are either specialized to only one class of proteins or predict one specific type of residue-level attribute. In this work, we develop a new deep-learning method, named Membrane Association and Secondary Structure Predictor (MASSP), for accurately predicting both residue-level structural attributes (secondary structure, location, orientation, and topology) and protein-level structural classes (bitopic,  $\alpha$ -helical,  $\beta$ -barrel, and soluble). MASSP integrates a multi-layer 2D convolutional neural network (2D-CNN) with a long short-term memory (LSTM) neural network into a multi-tasking framework. Our comparison shows that MASSP performs equally well or better than state-of-the-art methods for predicting residue-level secondary structures, boundaries of transmembrane segments, and topology. Furthermore, it achieves outstanding accuracy in predicting protein-level structural classes. MASSP automatically distinguishes the structural classes of input sequences and identifies transmembrane segments and topologies if present, making it broadly applicable

\*To whom correspondence should be addressed: jens@meilerlab.org.

Author contribution

B.L., J. Mendenhall, and J. Meiler conceived the study. B.L. and J. Mendenhall developed the method and performed data analysis. J.A.C. and J. Meiler supervised the project. All authors read, commented, and edited the manuscript.

Supporting Information

The following supporting information is available free of charge at ACS website <http://pubs.acs.org>.

Figure S1. Tuning architectural hyperparameters of the CNN component of MASSP.

Table S1. Five-letter PDB IDs of training set proteins grouped according to protein class.

Table S2. Five-letter PDB IDs of validation set proteins grouped according to protein class.

Table S3. Five-letter PDB IDs of test set proteins grouped according to protein class.

Availability

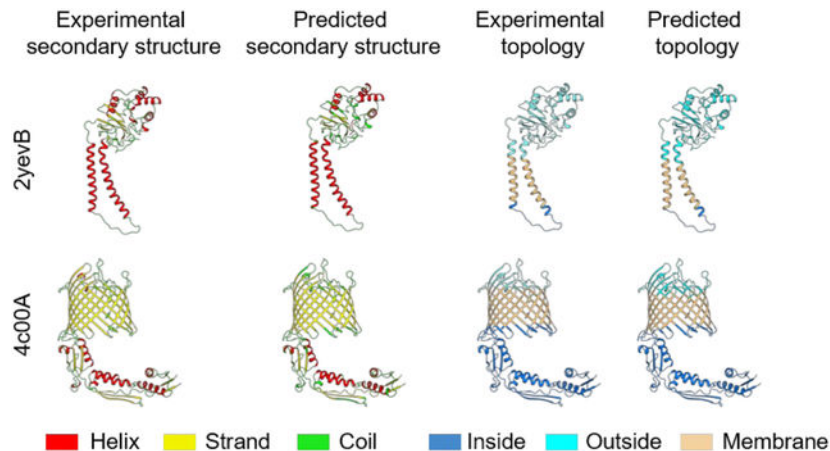
The source code of MASSP and test set predictions by all methods are made publicly available at: <https://github.com/computbiolgeek/massp>. MASSP is also available as a web server at [http://www.meilerlab.org/index.php/servers/show?s\\_id=26](http://www.meilerlab.org/index.php/servers/show?s_id=26). The C++ utility for computing position-specific-scoring matrices using multiple sequence alignments was implemented in the sequence analysis module of the BioChemical Library software, source and executables for which are available free of charge for academic use from [www.meilerlab.org/bcl\\_academic\\_license](http://www.meilerlab.org/bcl_academic_license).

Declaration of interests

None.

to different classes of proteins. In summary, MASSP's good performance and broad applicability make it well suited for annotating residue-level attributes and protein-level structural classes at the proteome scale.

## Graphical Abstract



## Keywords

multi-task deep learning; convolutional neural networks; long short-term memory networks; secondary structure prediction; transmembrane topology prediction

## Introduction

Protein structures are essential to understanding protein function, unraveling disease mechanisms, and the design of novel therapeutic molecules. While innovations in experimental techniques continue to help determine new structures at a rapid pace, determining the structures of all proteins by experimental techniques remains impractical. Thus, computational prediction of protein tertiary structures from amino acid sequences continues to be an area of active research<sup>1</sup>. While exciting progress was made in the past few years, especially in the 14<sup>th</sup> Critical Assessment of Techniques for Protein Structure Prediction (CASP14)<sup>2</sup>, protein tertiary structure prediction is still a challenging problem for many proteins<sup>2-4</sup>.

Accurate prediction of secondary structures provides useful information for predicting protein tertiary structures. Over the past decades, a plethora of methods were developed for secondary structure prediction<sup>5,6</sup>. Some of the first methods relied on statistical propensities of amino acids to form particular structures<sup>7</sup>, simple nearest-neighbor algorithms that involve finding short sequences of known structure that closely match stretches of the query sequence<sup>8</sup>, or explicit modeling of the contribution that neighboring residues make to the probability of a given structure state<sup>9</sup>. The accuracy of these methods hovered under 70% for the so-called Q3, a three-state prediction that classifies amino acids in likely helix (H), strand (extended, E), or coil (C) states. More accurate methods, better than 70% in Q3, were developed that use machine-learning-based models<sup>10</sup>. The machine-learning

models powering these methods are often artificial neural networks<sup>11–17</sup> that take sequence profiles generated from multiple sequence alignments as input<sup>11,12</sup> or couple sequence profiles with nonlocal interactions from predicted tertiary structures to boost the accuracy of secondary structure prediction<sup>17</sup>. Recently, motivated by unprecedented performance of deep neural networks in computer vision problems<sup>18</sup>, deeper neural network architectures, such as convolutional neural networks (CNNs) and recurrent neural networks were applied to protein secondary structure prediction with accuracy improving to above 80% in Q3<sup>6,19–21</sup>.

For transmembrane proteins (TMPs), an additional essential component of predicting the tertiary structure is to first locate all the TM segments and the orientation of the protein with respect to the membrane. These tasks are collectively known as “topology prediction”. Several algorithms have been developed over the past decades for predicting the topology of  $\alpha$ -helical TMPs (TM-alpha) or  $\beta$ -barrel TMPs (TM-beta)<sup>22,23</sup>. The earliest method of Kyte and Doolittle<sup>24</sup> for TM helix (TMH) prediction computes simple “hydropathy plots” to identify probable TMHs. While this method is conceptually simple and easy to implement, it can neither predict the “inside-outside phasing” of the helices relative to the cytoplasm, i.e. topology, nor the location of TM strands in TM-beta proteins. By combining hydropathy analysis and the “positive-inside rule”, the observation that positively charged residues are more abundant in cytoplasmic as compared to periplasmic regions of bacterial inner membrane proteins<sup>25</sup>, a method dubbed TOP-PRED was developed for predicting TMHs and their topologies<sup>26</sup>. While TOP-PRED identified all 135 TMHs in the 24 proteins tested with only one overprediction and correctly predicted the topology of 22 proteins, no results were reported as to how accurate TOP-PRED can locate the boundaries (N- and C- terminal ends) of TMHs. Aimed at predicting both the boundaries and topology of TMHs, the PHDhtm method, whose driving predictor is a two-level neural network system trained on 69 transmembrane proteins, was developed<sup>27,28</sup>. TMHMM and HMMTOP<sup>29</sup> are the first methods to model membrane topology of IMPs with hidden Markov models. For transmembrane segment detection, MEMSAT3<sup>30</sup>, OCTOPUS<sup>31</sup> (for TM-alpha proteins), BOCTOPUS<sup>32</sup> and PRED-TMBB<sup>33</sup> (for TM-beta proteins) are classic methods. Recent innovations in deep learning also enabled the development of more sophisticated methods, such as TMP-SS<sup>34</sup> and DMCTOP<sup>35</sup>, for membrane segment detection and topology prediction.

While these existing methods have been widely used, they are either specialized on one class of proteins or developed to predict one specific type of residue-level attributes. In the current work, we leverage recent advances in deep neural network (DNN) architectures to develop a novel method, Membrane Association and Secondary Structure Prediction (MASSP), that simultaneously predicts residue-level structural attributes (secondary structure, location, orientation, and topology) and protein-level structural classes (bitopic, TM-alpha, TM-beta, and soluble). This is achieved by integrating a multi-layer 2D-CNN with a LSTM neural network into a multi-tasking framework. We extensively compare MASSP with several classic and state-of-the-art methods and show that MASSP performs equally well or better for predicting residue-level attributes. Furthermore, MASSP achieves high accuracy in predicting protein-level structural classes and is broadly applicable to different classes of proteins. The preliminary version of this method was successfully adopted in two previous studies for *de novo* tertiary structure prediction of TM-alpha proteins<sup>36,37</sup>. In

this manuscript, we describe the method in detail, demonstrate its broad applicability, and discuss aspects of its improved performance compared to some of the most popular methods in the field.

## Methods

### Datasets

The sets of bitopic, TM-alpha, and TM-beta proteins were all obtained from the OPM database<sup>38</sup>. The datasets were pruned to 25% sequence identity using the PISCES server<sup>39</sup>, and structures for which the resolution is worse than 3.0 Å or were determined using techniques other than X-ray crystallography or have less than 40 residues were excluded. The final datasets consist of 240 TM-alpha protein chains, 54 bitopic, and 77 TM-beta protein chains. This dataset was augmented by adding 372 (accounting for 50% of the dataset) soluble protein chains. The set of soluble protein chains was randomly selected from the set of all X-ray only soluble protein chains with resolution better than 3.0 Å and culled at 25% sequence identity using the PISCES server<sup>39</sup>. The dataset was split according to the ratio 8:1:1 as training:validation:test while maintaining an approximately constant fractions of each class of proteins in each subset. The fractions of TM-alpha, bitopic, TM-beta, and soluble proteins are 31.5%, 6.8%, 11.0%, and 50.7%, respectively, in each subset. This resulted in three subsets containing 165080, 18293, 18571 training examples (residues) for which all four target attributes are available. For TM-alpha and bitopic proteins combined, there are 1371, 143, and 154 transmembrane helices that are at least 10 residues long in the training, validation, and test sets respectively, and for TM-beta proteins, there are 855, 89, and 104 membrane spanning beta strands that are at least 5 residues long, respectively. The reference secondary structure elements for each chain were derived from the consensus identification of DSSP<sup>40</sup>, Stride<sup>41</sup>, and PALSSE<sup>42</sup>. The reference residue location and topology annotations for each chain were derived from the coordinates and membrane boundaries provided by OPM<sup>43</sup>.

### Determining beta-strand orientation

The topology of each  $\beta$ -strand composing a barrel was identified by computing the difference in Z-coordinate between the first and last residue in the strand, with strands that ascend in the Z-axis labeled as up (U), while those that descend in Z-axis coordinate labeled as down (D). For orientation, residues were labeled based on whether the  $C_{\alpha}-C_{\beta}$  vector points to the pore (P), lipid bilayer (L), or is in solution (s).

Conventionally, three-state Hidden Markov models use interior/exterior/bilayer states and infer strand direction afterwards. We found this state separation to be unappealing for use in training neural networks because in several known structures there are trans-pore segments that cross from intracellular to extra-cellular space, rendering assignment of any residue connected to a pore-crossing segment as intracellular or extracellular ambiguous. Creating an intra-pore state is similarly unappealing due to known structures with dynamic helical plugs that occupy several states. Conversely, while there is debate in the literature about whether topology can be inverted in TM-alpha proteins<sup>44,45</sup>, we are aware of no TM-beta

proteins in which the barrel flips under physiologically relevant conditions, leaving the assignment of U/D/s states unambiguous.

### PSSM calculation

The position-specific-scoring matrix (PSSM) of an amino acid sequence contains the log-odds of each of the 20 amino acids observed at each of the sequence positions over evolutionary timescale<sup>46</sup>. To compute the PSSM for a target amino acid sequence, we first queried the UniRef20 protein sequence database for sequences homologous to the target sequence using the HHblits method<sup>47</sup>. The parameters used in the searching were three iterations with a e-value of 0.001, a maximum sequence identity of 0.90, a minimum coverage of 0.50 of the target sequence residues. The multiple sequence alignment generated by running HHblits was used as input to an in-house C++ utility software implementing the algorithm described in Altschul et al.<sup>46</sup> to compute the floating point valued PSSM for the target sequence.

### Network architecture

The performance of a neural network depends on many hyperparameters specifying the architecture of the neural network. In designing MASSP, we considered the number of convolutional and densely connected layers, the number of units in each layer, and the size of the input feature matrix. We tuned hyperparameters by training models using the training set and comparing model performances on the validation set. We note that the test set was never touched before the final model was selected.

We followed a workflow of developing deep-learning models recommended in<sup>48</sup>. Specifically, we started out with a very small architecture that has a total of 1339 trainable parameters. This model takes a  $7 \times 20$  input feature matrix, one convolutional layer with weight ReLU units, and one densely connected layer with eight ReLU units. Our goal at this stage was to have a model that is better than a baseline model, i.e. to make sure that patterns in the input PSSM can be learned by the chosen type of neural network. In fact, after 20 epochs of training, we observed that the validation loss was still decreasing (Fig. S1A), indicating that a model with a larger architecture may achieve higher performance.

We next scaled up the architecture of the model by adding two additional convolutional layers, one additional layer of densely connected units, and increasing the number of ReLU units in each layer to 16, 32, 64, 32, and 64, respectively. We also expanded the size of the input feature matrix to cover 15 residues on each side of the central residue. The goal was to have a model that is sufficiently powerful. This was confirmed by observing the model's performance on the validation set began to degrade only after the first four epochs of training (Fig. S1B).

Starting with this sufficiently powered model, we continued several rounds of hyperparameter tuning by scaling down the size of the architecture and adding regularization (see Fig. S1C and S1D for two examples). In the end, we selected a model with the flattened layer regularized by a 0.25 dropout rate and that appeared to have the optimal overall prediction accuracy on the validation set. The architecture of the final model is shown in Fig. 1B.

## Network construction and training

MASSP is a two-tier neural network system designed for simultaneous prediction of four 1D structural attributes and the protein class for the input amino acid sequence. The first tier is a multi-output deep 2D-CNN and was constructed using the functional API of the Keras (version 2.3.1) model-level deep-learning framework for Python (version 3.7.10), with TensorFlow (version 2.0.0) serving as the backend engine. The output layer of this multi-output 2D-CNN consists of four heads each representing a separate one of the four target structural attributes. The final model was trained using the Adam optimizer<sup>49</sup> with the following hyperparameters: maximum learning rate = 0.001,  $\beta_1 = 0.9$ ,  $\beta_2 = 0.999$ . Kernel weights of the model were initialized using the Glorot uniform initializer and updated after processing each batch of 64 training examples. The activation of the last layer was the softmax function and the loss function was the categorical cross entropy function for all output heads. By default, Keras sums all four losses into a global loss that is back propagated to update trainable weights of the 2D-CNN. We did not reweigh individual losses because no one target attribute can be said more important than another. We set the maximum number of training epochs to 100. However, the training was stopped when the loss on the validation set stopped decreasing for 5 consecutive epochs. The LSTM model was similarly trained except that the maximum number of training epochs was set to 20.

## Generating predictions using other methods

PSIPRED and MEMSAT3 predictions were generated by running PSIPRED and MEMSAT3 on test set protein sequences locally with the recommended UniRef90 sequence database (PSIPRED version 4.02, MEMSAT version 3.0). SPINE-X predictions were similarly generated by running SPINE-X 2.0 locally with the UniRef90 sequence database. RaptorX-Property predictions were generated by running RaptorX-Property version 1.02 on test protein sequences with the recommended UniRef20 sequence database. Source code of these locally run methods were downloaded and compiled locally, and the programs were set up following recommendations. TMHMM2 predictions were generated by submitting test sequences to the TMHMM web server at <http://www.cbs.dtu.dk/services/TMHMM/>. Similarly, OCTOPUS predictions were obtained from its web server at <http://boctopus.bioinfo.se/octopus/>. BOCTOPUS2 predictions were generated by submitted all test protein sequences to the BOCTOPUS2 web server at <http://boctopus.bioinfo.se/pred/> with default settings. PRED-TMBB predictions were generated by submitting all test protein sequences to the PRED-TMBB web server at <http://www.compgen.org/tools/PRED-TMBB2> in single-sequence mode with default settings. NetSurfP-2.0 predictions were generated by submitting test sequences to the NetSurfP-2.0 server at <https://services.healthtech.dtu.dk/service.php?NetSurfP-2.0>. Similarly, TOPCONS2 predictions were generated by submitting test sequences to the TOPCONS server at <https://topcons.cbr.su.se/pred/>.

## Performance measures

We employed two of the most used performance measures to evaluate MASSP and to compare it with other methods. The first measure, called the Q3 accuracy, is defined as the fraction of residues for which the three-state target attributes are correctly predicted. For secondary structure prediction in this work,

$$Q_3 = \frac{P_{helix} + P_{strand} + P_{coil}}{N},$$

where  $N$  is the total number of predicted residues and  $P_x$  is the number of correctly predicted secondary structures of type  $x$ . The accuracy measure for evaluating the prediction of location, orientation, and membrane topology were similarly defined. For example, the Q3 accuracy for residue-level topology prediction is defined as

$$Q_3 = \frac{P_{up} + P_{down} + P_s}{N}.$$

The other measure is known as the fractional overlap of segments (SOV) originally proposed by Rost et al<sup>50</sup>. SOV measures the percentage of correctly predicted secondary structure segments rather than individual residue positions, and it pays less attention to small errors in the ends of structural elements. Consistent with previous studies, we used the definition of SOV introduced by Zemla et al.<sup>51</sup> and implemented in Perl by Liu et al<sup>52</sup>. For secondary structure prediction,

$$SOV = \frac{100}{N_{SOV}} \sum_{S_o} \left[ \frac{\minov(s_{obs}, s_{pred}) + \delta(s_{obs}, s_{pred})}{\maxov(s_{obs}, s_{pred})} \right] \text{len}(s_{obs})$$

where  $s_{obs}$  and  $s_{pred}$  represent all observed and predicted segments of helices, strands, and coils.  $S_o$  is the set of all overlapping pairs of  $s_{obs}$  and  $s_{pred}$  for which the segments are in the same state.  $\text{len}(s_{obs})$  is the length in residues of any segment  $s_{obs}$ ;  $\minov(s_{obs}, s_{pred})$  is the length of the actual overlap between any segment pair  $(s_{obs}, s_{pred})$  in  $S_o$  and  $\maxov(s_{obs}, s_{pred})$  is the total extent to which at least one residue is that state.  $N_{SOV}$  is the total number of residues in  $s_{obs}$  in all pairs in plus the number of residues in any  $s_{obs}$  that are not overlapped by a predicted segment of the same state. The summation represents the fraction of the segment pair that the observed and predicted states agree.  $\delta(s_{obs}, s_{pred})$  is added to allow for some variation in segment boundaries and is defined as

$$\delta(s_{obs}, s_{pred}) = \min \left\{ \left[ \maxov(s_{obs}, s_{pred}) - \minov(s_{obs}, s_{pred}) \right], \minov(s_{obs}, s_{pred}), \text{int} \left[ \frac{\text{len}(s_{obs})}{2} \right], \text{int} \left[ \frac{\text{len}(s_{pred})}{2} \right] \right\}.$$

While SOV is typically used to evaluate performance on secondary structure prediction, it is a general measure that is well suited for, and thus was also used in, evaluating prediction of the other three structural attributes in this study. To this end, we adopted the SOV\_refine metric introduced by Liu et al.<sup>52</sup> for computing the SOVs for the other three structural attributes.

## Results

### Overview of MASSP

We designed MASSP as a two-tier prediction system to work with both soluble and membrane proteins (Fig. 1). The first tier is a multi-task multi-layer 2D-CNN that predicts residue-level structural attributes. The second tier is an LSTM neural network that treats the predictions of the first tier as natural language input and predicts the structural class of the input sequence.

Given an amino acid sequence, MASSP calls HHblits (version 3.0)<sup>47</sup> to search against the UniRef20 sequence database for homologous proteins and to create a multiple sequence alignment (MSA) of the hits to the input sequence. The MSA contains essential evolutionary information and is then used as input to an in-house C++ utility to compute the corresponding position-specific scoring matrix (PSSM, see Methods). To predict the target attributes of a given residue, MASSP takes as input a  $21 \times 20$  matrix consisting of a sliding window of 21 positions around the residue of interest (Fig. 1A). The core algorithm of MASSP is a multi-output deep 2D-CNN that simultaneously predicts the secondary structure (helix, strand, or coil, indicated by H, E, or C in Fig. 1B), location (membrane or solution, indicated by M or s in Fig. 1B), orientation (lipid-, pore-facing, or soluble indicated by L, P, or s in Fig. 1B), and topology of the residue of interest (up, down, or soluble indicated by U, D, or s in Fig. 1B). While a separate model could be trained for each of the target attributes, we chose to learn to jointly predict all target attributes simultaneously, because these attributes are not necessarily independent. Once all the four 1D residue-level structural attributes of the input protein sequence are predicted by the first tier 2D-CNN, the predicted symbols for each residue are combined to form four-letter “tokens”, which are then concatenated to form a “sentence”. Thus, a protein sequence with  $I$  residues will be represented by a sentence of  $I$  tokens. The “semantic” meaning of this sentence (i.e. the structural class of the input protein sequence) is classified using an LSTM network (Fig. 1C).

We tuned the architecture of MASSP by following a workflow of developing deep-learning models recommended in<sup>48</sup> (see Methods). For the 2D-CNN, we started out with a model that has a total of 1339 parameters. This “small” model confirmed that patterns in the input matrix are learnable (Fig. S1A). We then scaled up the model to 198187 parameters, which is sufficiently powerful to the point that overfitting is obvious (Fig. S1B). We continued several rounds of hyperparameter tuning by scaling down the size of the architecture and adding regularization (see Fig. S1C and S1D for two examples). The architecture of LSTM was similarly tuned. In the end, we selected a model with 136,651 parameters for training, shown in Fig. 1B, that has the optimal overall prediction accuracy on the validation set.

We developed MASSP using the experimental structures of 240 TM-alpha, 54 bitopic, and 77 TM-beta proteins. This data set was further augmented by adding 372 (accounting for 50% of the dataset) soluble protein subunits homology-reduced at 25% pairwise sequence identity level. The dataset was split according to the ratio 8:1:1 as training:validation:test (Tables S1, S2, S3). We used the validation set to tune hyperparameters of MASSP, such as number of layers and number of neurons in each layer, of the neural networks, and size of



the input matrix (see Methods). The final selected models of MASSP, as described above, was evaluated on the test set that consists of 5 bitopic, 23 TM-alpha, 8 TM-beta, and 37 soluble proteins, contributing a total of 19416 residues for which all four target attributes are available.

### Dataset statistics

We split the data set according to the ratio 8:1:1 as training:validation:test while maintaining the constant fractions of each class of proteins in each subset. As shown in Fig. 2A, each of the three secondary structure types is about evenly represented in the three subsets for TM-alpha, TM-beta, and soluble proteins. Some biases exist for bitopic proteins because a single bitopic protein with a large soluble domain could easily skew their representation among the three subsets. Likewise, membrane topologies are also similarly represented in the three subsets for all integral membrane protein classes (Fig. 2B). Soluble proteins are all considered to have a topology of outside the membrane. These three subsets each have 165979, 18293, 19416 instances (residues) for which all four target attributes are available. For TM-alpha and bitopic proteins combined, there are 1371, 143, and 154 transmembrane helices that are at least 10 residues long in the training, validation, and test sets respectively, and for TM-beta proteins, there are 855, 89, and 104 membrane-spanning beta strands, respectively.

### MASSP accurately predicts residue-level attributes

We evaluated the performance of MASSP using the Q3 (or Q2 in the case of residue location prediction) and the fractional segment overlap (SOV) accuracy measures. Q3 is the percentage of correctly predicted residues on all three types of secondary structure, transmembrane topology, or in the case of TM-beta proteins, residue orientation. The SOV measure counts the fractional extent to which predicted and experimental segments of secondary structure or transmembrane region overlap, with some allowance for non-matching residues at the ends<sup>50,51</sup> (see Methods).

Overall, MASSP achieves accurate simultaneous prediction of residue-level secondary structure, orientation, and transmembrane topology, with Q3s of 0.844, 0.944, and 0.928, respectively, over 19416 test set residues (Fig. 3A). The Q2 accuracy for residue location prediction over the same set of residues is 0.946. When the overall accuracy was decomposed according to protein classes, MASSP achieved the highest secondary structure prediction Q3 for TM-alpha proteins (0.884) and the lowest for soluble proteins (0.809) (Fig. 3B). This is likely because the lipid bilayer imposes several constraints on structures of integral membrane proteins, simplifying their secondary structure prediction. Remarkably, MASSP rarely predicts residues of soluble proteins to be membrane associated, as manifested by the near perfect accuracy in predicting location, orientation, and topology of residues of soluble proteins (Fig. 3B). Among the three classes of membrane associated proteins, the three membrane related attributes are much easier to predict for both bitopic and TM-beta proteins than for TM-alpha proteins. This is likely because the folds of TM-alpha proteins are generally much more complex than bitopic or TM-beta proteins.

We also computed the per-protein distributions of the Q3 and the SOV measures (Fig. 3C and 3D). For secondary structure prediction, the median Q3 and SOV are 0.845 and 0.792, respectively. For the other three membrane protein-related attributes, the median accuracies for location, orientation, and topology are 0.881, 0.890, and 0.867, and the median SOVs are 0.755, 0.803, and 0.852, respectively, across 36 membrane proteins in the test set. These results indicate that MASSP achieved excellent performance in predicting residue-level attributes.

### Accurate prediction of protein classes via a LSTM neural network

Knowing the structural class of a protein (soluble vs. transmembrane, TM-alpha vs. TM-beta, and bitopic vs. multi-spanning) is essential toward understanding its potential cellular function. MASSP builds on the residue-level predictions of secondary structures, membrane associations, residue orientations, and transmembrane topologies and leverages the power of the LSTM recurrent neural network architecture<sup>53</sup> to accurately predict protein classes. In this LSTM model, each sequence is treated as a single data point and is processed by iterating through the sequence elements and maintaining a state containing information relative to what it has seen so far. When applied to protein class prediction, the network maintains information about all residues in the sequence. This is akin to sentiment analysis in natural language processing<sup>48</sup>. To achieve this, we tokenize protein sequences at residue level by representing each residue with a “word” composed of their residue-level predictions. For instance, a residue predicted to be helix-forming (H), membrane associated (M), lipid facing (L), and going toward the cytoplasmic side (D) would be represented by the word “HMLD”, and a strand-forming (E) residue in a soluble protein would be represented by the word “Esss”. As shown in Table 1, the LSTM model achieved a notable four-class classification accuracy of 98.6%. Specifically, it correctly classified all bitopic, TM-alpha, and soluble proteins. The only mis-classified protein in the held-out test set is the pore-forming TM-beta subunit 7ahlA, which has a predominantly large soluble domain (95.9% of all residues are outside the membrane). Our results suggest a remarkable effectiveness of LSTM neural networks in predicting protein classes by learning from predicted residue-level attributes that are combined in a residue-wise manner.

### MASSP achieves state-of-the-art performance

We evaluated the performance of MASSP relative to other methods in the field using the same test set. We first compared it to Jufo9D, which was previously developed in our group<sup>54</sup>, and five other popular secondary structure prediction methods (PSIPRED<sup>11</sup>, RaptorX-Property<sup>55</sup>, SPINE-X<sup>56,57</sup>, NetSurfP-2.0<sup>58</sup>, and TMP-SS<sup>34</sup>). MASSP performs comparably or better than all secondary structure prediction methods evaluated in this work (Figure 4). Specifically, MASSP achieved a median Q3 of 0.845, which is comparable to the best method NetSurfP-2.0 (0.850), but substantially better than PSIPRED (0.829), SPINE-X (0.819), RaptorX-Property (0.750), TMP-SS (0.821). Our previous method Jufo9D achieved a moderate median Q3 of 0.766. We note that due to the unavailability of NetSurfP-2.0 training set, we were not able to check if any proteins used to train NetSurfP-2.0 was in our test set (No downloadable training data at <https://services.healthtech.dtu.dk/service.php?NetSurfP-2.0>, last accessed April. 13<sup>th</sup> 2021). We also compared these methods in terms of the SOV metric. With a median SOV of 0.792, MASSP is ranked the best among all

evaluated methods and is substantially better than TMP-SS (0.768) and NetSurfP-2.0 (0.729) whose SOVs are ranked next. PSIPRED (0.690) and SPINE-X (0.705) performed similarly in both evaluations, while RaptorX-Property (0.651) was significantly worse. Our evaluation also indicates that Jufo9D has the lowest SOVs.

We then compared the performance of MASSP with six other commonly used methods (JUFO9D<sup>54</sup>, MEMSAT3<sup>30</sup>, OCTOPUS<sup>31</sup>, TMHMM2<sup>59</sup>, TMP-SS<sup>34</sup>, and TOPCONS2<sup>60</sup>) in predicting transmembrane segments and topology for TM-alpha and bitopic proteins (Tables 2 and 3). Our comparison shows that MASSP has the best performance in predicting the number and location of TMHs. MASSP correctly identified 151 out of all 154 TMHs that are at least 10 residues long, ~9% higher than the MEMSAT3 and OCTOPUS, which are the second best. For topology prediction, OCTOPUS performs the best both at the protein level and residue level for both TM-alpha and bitopic proteins. OCTOPUS only mis-predicted the topology of 1 bitopic protein, whereas MASSP mis-predicted the topology of 1 TM-alpha and 2 bitopic proteins. At the residue level, MASSP attained a Q3 of 0.871 and SOV of 0.850 for TM-alpha proteins, which are comparable to OCTOPUS (Q3: 0.874, SOV: 0.861). However, MASSP's residue-level performance is substantially lower than OCTOPUS for bitopic proteins, due to its worse prediction of protein-level topology.

We next compared the performance of MASSP in predicting TM strands of TM-beta proteins with Jufo9D<sup>54</sup>, BOCTOPUS2<sup>32</sup>, and PRED-TMBB2<sup>33</sup>. The latter two methods were recently developed methods and were shown to have improved performance over previous methods. As shown in Table 4, MASSP achieved the highest accuracy (Q3: 0.775, SOV: 0.775) in residue-level topology prediction and in predicting the number and location of TM strands (98 out of 104 TM strands that have at least 5 residues). BOCTOPUS2 correctly predicted the highest number of protein-level topology (6 out of the 8 TM-beta proteins in the held-out test set) and achieved the same level of accuracy as MASSP in predicting the number and location of TM strands. However, the residue-level accuracy of BOCTOPUS2 (Q3: 0.696, SOV: 0.745) is substantially worse than MASSP.

In summary, our extensive comparison on secondary structure and membrane association/topology prediction tasks demonstrates that MASSP performance is comparable to or better than state-of-the-art methods, while being more versatile and applicable to any protein sequences.

### Examples of accurate MASSP predictions

We demonstrate the predictions made by MASSP on two proteins by mapping secondary structures and topologies predicted by MASSP onto tertiary structures (Fig. 5). We selected the subunit B from the *caa3*-type cytochrome oxidases from *Thermus thermophilus* (PDB ID: 2yevB) and the TamA protein from the *E. coli* (PDB ID: 4c00A) as two representative examples, because they both consist of large transmembrane and soluble domains and all three types of secondary structures are also well represented. The subunit B from the *caa3*-type cytochrome oxidases consists of an alpha-helical TM domain and a large extracellular soluble domain, resolved at a resolution of 2.4 Å. TamA is an *E. coli* Omp85 protein involved in autotransporter biogenesis. It comprises a 16-stranded transmembrane  $\beta$ -barrel and three cytoplasmic POTRA domains, resolved at a resolution of 2.3 Å. For 2yevB,

MASSP achieved a Q3 of 0.936 and a SOV of 0.893 for secondary structure prediction, and Q3 of 0.984 and a SOV of 0.977 for topology prediction. For 4c00A, MASSP achieved a Q3 of 0.815 and a SOV of 0.780 for secondary structure prediction, and Q3 of 0.938 and a SOV of 0.911 for topology prediction. These two examples illustrate that even for proteins composed of mixed large transmembrane and soluble domains, MASSP can accurately predict both secondary structures and transmembrane topologies.

## Discussion

Prediction of secondary structures and transmembrane regions are fundamental problems in bioinformatics with broad downstream applications. Decades of research into these problems have produced dozens of methods that can be broadly categorized into knowledge-based analyses, generative probabilistic modeling, and discriminative machine learning. However, methods with better performance are still needed giving the increasingly large gap between the set of known sequences and known structures<sup>5,22</sup>. Among machine-learning methods, those mainly based on neural networks, have been shown to generally perform best, with Q3 accuracies for secondary structure prediction around 80%<sup>5</sup>; however, their accuracies in predicting transmembrane segments and topologies have been low<sup>22</sup>. A newly sequenced protein can be of any one of the structural classes of soluble, bitopic, TM-alpha, or TM-beta. Thus, methods are needed for simultaneous prediction of secondary structures, transmembrane regions, topologies, and protein class.

In this work, we introduced MASSP, a multi-task deep-learning framework designed for simultaneous prediction of secondary structures, transmembrane regions, topologies, and protein classes. MASSP simultaneously predicts residue-level structural attributes for all four classes of proteins. The core algorithm behind MASSP is a two-tier deep neural network. Our framework is conceptually similar to several recent works that reported substantially improved performances in predicting signal peptides<sup>61</sup>, protein subcellular localization<sup>62</sup>, the topology of TM-alpha proteins<sup>35</sup>. The first tier is a multi-layer multi-task 2D-CNN that predicts residue-level 1D structural attributes. The second tier is a LSTM neural network that treats the predictions of the first tier from the perspective of natural language and predicts the protein class of the input amino acid sequence. We demonstrated that MASSP accurately predicts residue-level secondary structures, locations, orientations, and residue topologies, and protein-level structural classes. We also show that the performance of MASSP is comparable or better than several widely used methods for secondary structure and membrane topology prediction.

In addition to performance that is better than or comparable to existing methods, the MASSP framework has several other important strengths. First, the multi-tasking nature of MASSP makes it a versatile tool that simultaneously predicts residue-level 1D structural features. Specifically, the multi-output architecture of the 2D-CNN would allow us to easily extend MASSP to predict other structural properties, such as solvent accessibility and contact order, from amino acid sequence without the requirement of training a different network for each target property. Second, the fact that MASSP does not make *a priori* assumptions about whether the given amino acid sequence represents a TM-alpha protein, TM-beta protein, or a soluble protein gives the user a one-stop shop where the secondary structures and membrane

associations of a given amino acid sequence of any protein class can be predicted. Thus, MASSP could be applied to wholly sequenced proteomes of an organism to predict the protein class composition of the proteome as well as residue-level secondary structures and membrane associations.

Despite half-a-century of research into the problems of secondary structure prediction and transmembrane segment and topology prediction, these problems are still open and the field seems to have reached a bottleneck where improvement in prediction accuracy is small and progress is slow<sup>5</sup>. It strikes us as remarkable that the approach that we took in developing MASSP, i.e. learning patterns in PSSMs with deep 2D-CNNs, achieved comparable or better performance than previously developed, more sophisticated methods. This can be partially attributed to applying the “right” type of neural network to the “right” representation of input features, i.e. we leveraged the image-processing power inherent in 2D-CNNs and the fact that PSSMs can be thought of images with real-valued “pixels”. Recent advances in protein tertiary structure prediction also demonstrated that a boost in performance can be achieved by developing novel neural network architectures that leverage the problem’s fundamental mathematical essence<sup>4,63–65</sup>. For example, AlphaFold, the algorithm that outperformed all entrants in CASP13, relies on the fundamental principle of the interconvertibility of probability and energy, and predicts the probability distributions of residue pair distances and converts distance distributions to energy landscapes<sup>4,64</sup>. In the recent CASP14 competition, AlphaFold2, which is based on an attention-based neural network system, trained end-to-end, that attempts to interpret protein structures as “spatial graphs”, outperformed all its competitors with an unprecedented margin<sup>2,66</sup>. These advances in tertiary structure prediction together with our work here suggest that completely solving these problems will likely require the development novel neural network architectures and representation of input features.

## Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

## Acknowledgement

This work was supported by NIH awards R01 GM080403, R01 HL122010, R01 DA046138, and by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) through SFB1423, project number 421152132, subproject Z04, A07. B.L. was also supported by an American Heart Association Postdoctoral Fellowship (20POST3522000). This work was also conducted in part using the computational resources of the Advanced Computing Center for Research and Education at Vanderbilt University.

## References

- (1). Li B; Fooksa M; Heinze S; Meiler J Finding the needle in the haystack: towards solving the protein-folding problem computationally. *Crit Rev Biochem Mol Biol* 2018, 53 (1), 1–28. [PubMed: 28976219]
- (2). Jumper J; Evans R; Pritzel A; Green T; Figurnov M; Tunyasuvunakool K; Ronneberger O; Bates R; Žídek A; Bridgland A et al. Fourteenth Critical Assessment of Techniques for Protein Structure Prediction, 2020.
- (3). Shrestha R; Fajardo E; Gil N; Fidelis K; Kryshchuk A; Monastyrskyy B; Fiser A Assessing the accuracy of contact predictions in CASP13. *Proteins* 2019, 87 (12), 1058–1068. [PubMed: 31587357]

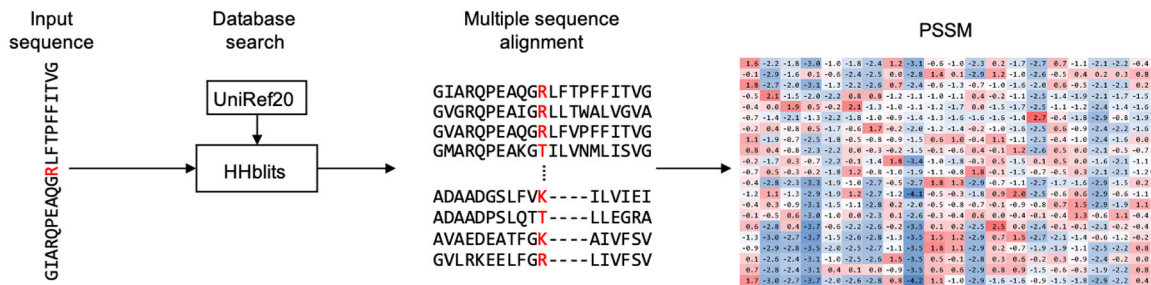
- (4). Senior AW; Evans R; Jumper J; Kirkpatrick J; Sifre L; Green T; Qin C; Zidek A; Nelson AWR; Bridgland A et al. Improved protein structure prediction using potentials from deep learning. *Nature* 2020, 577 (7792), 706–710. [PubMed: 31942072]
- (5). Yang Y; Gao J; Wang J; Heffernan R; Hanson J; Paliwal K; Zhou Y Sixty-five years of the long march in protein secondary structure prediction: the final stretch? *Brief Bioinform* 2018, 19 (3), 482–494. [PubMed: 28040746]
- (6). Torrisi M; Pollastri G; Le Q Deep learning methods in protein structure prediction. *Comput Struct Biotechnol J* 2020, 18, 1301–1310. [PubMed: 32612753]
- (7). Chou PY; Fasman GD Prediction of protein conformation. *Biochemistry* 1974, 13 (2), 222–245. [PubMed: 4358940]
- (8). Frishman D; Argos P Seventy-five percent accuracy in protein secondary structure prediction. *Proteins* 1997, 27 (3), 329–335. [PubMed: 9094735]
- (9). Garnier J; Gibrat JF; Robson B GOR method for predicting protein secondary structure from amino acid sequence. *Methods Enzymol* 1996, 266, 540–553. [PubMed: 8743705]
- (10). Rost B; Sander C Prediction of protein secondary structure at better than 70% accuracy. *J Mol Biol* 1993, 232 (2), 584–599. [PubMed: 8345525]
- (11). Jones DT Protein secondary structure prediction based on position-specific scoring matrices. *J Mol Biol* 1999, 292 (2), 195–202. [PubMed: 10493868]
- (12). Rost B PHD: predicting one-dimensional protein structure by profile-based neural networks. *Methods Enzymol* 1996, 266, 525–539. [PubMed: 8743704]
- (13). Qian N; Sejnowski TJ Predicting the secondary structure of globular proteins using neural network models. *J Mol Biol* 1988, 202 (4), 865–884. [PubMed: 3172241]
- (14). Rost B; Sander C Prediction of Protein Secondary Structure at Better Than 70-Percent Accuracy. *Journal of Molecular Biology* 1993, 232 (2), 584–599. [PubMed: 8345525]
- (15). Rost B; Sander C Improved prediction of protein secondary structure by use of sequence profiles and neural networks. *Proc Natl Acad Sci U S A* 1993, 90 (16), 7558–7562. [PubMed: 8356056]
- (16). Meiler J; Zeidler A; Schmaschke F; Muller M Generation and evaluation of dimension-reduced amino acid parameter representations by artificial neural networks. *Journal of Molecular Modeling* 2001, 7 (9), 360–369.
- (17). Meiler J; Baker D Coupled prediction of protein secondary and tertiary structure. *Proc Natl Acad Sci U S A* 2003, 100 (21), 12105–12110. [PubMed: 14528006]
- (18). LeCun Y; Bengio Y; Hinton G Deep learning. *Nature* 2015, 521 (7553), 436–444. [PubMed: 26017442]
- (19). Wang S; Peng J; Ma J; Xu J Protein Secondary Structure Prediction Using Deep Convolutional Neural Fields. *Sci Rep* 2016, 6, 18962. [PubMed: 26752681]
- (20). Zhavoronkov A; Ivanenkov YA; Aliper A; Veselov MS; Aladinskiy VA; Aladinskaya AV; Terentiev VA; Polykovskiy DA; Kuznetsov MD; Asadulaev A et al. Deep learning enables rapid identification of potent DDR1 kinase inhibitors. *Nat Biotechnol* 2019, 37 (9), 1038–1040. [PubMed: 31477924]
- (21). Liu S; Xu LY; Guan FH; Liu YT; Cui YX; Zhang Q; Zheng X; Bi GQ; Zhou ZH; Zhang XK et al. Cryo-EM structure of the human alpha 5 beta 3 GABA(A) receptor. *Cell Res* 2018, 28 (9), 958–961. [PubMed: 30140029]
- (22). Koehler Leman J; Ulmschneider MB; Gray JJ Computational modeling of membrane proteins. *Proteins* 2015, 83 (1), 1–24. [PubMed: 25355688]
- (23). Fleishman SJ; Ben-Tal N Progress in structure prediction of alpha-helical membrane proteins. *Curr Opin Struct Biol* 2006, 16 (4), 496–504. [PubMed: 16822664]
- (24). Kyte J; Doolittle RF A simple method for displaying the hydropathic character of a protein. *J Mol Biol* 1982, 157 (1), 105–132. [PubMed: 7108955]
- (25). von Heijne G The Distribution of Positively Charged Residues in Bacterial Inner Membrane-Proteins Correlates with the Trans-Membrane Topology. *Embo Journal* 1986, 5 (11), 3021–3027.
- (26). von Heijne G Membrane-Protein Structure Prediction - Hydrophobicity Analysis and the Positive-inside Rule. *Journal of Molecular Biology* 1992, 225 (2), 487–494. [PubMed: 1593632]

- (27). Rost B; Casadio R; Fariselli P; Sander C Transmembrane helices predicted at 95% accuracy. *Protein Sci* 1995, 4 (3), 521–533. [PubMed: 7795533]
- (28). Rost B; Fariselli P; Casadio R Topology prediction for helical transmembrane proteins at 86% accuracy. *Protein Sci* 1996, 5 (8), 1704–1718. [PubMed: 8844859]
- (29). Tusnady GE; Simon I Principles governing amino acid composition of integral membrane proteins: application to topology prediction. *J Mol Biol* 1998, 283 (2), 489–506. [PubMed: 9769220]
- (30). Jones DT Improving the accuracy of transmembrane protein topology prediction using evolutionary information. *Bioinformatics* 2007, 23 (5), 538–544. [PubMed: 17237066]
- (31). Viklund H; Elofsson A OCTOPUS: improving topology prediction by two-track ANN-based preference scores and an extended topological grammar. *Bioinformatics* 2008, 24 (15), 1662–1668. [PubMed: 18474507]
- (32). 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), 1571–1573. [PubMed: 26794316]
- (33). Tsirigos KD; Elofsson A; Bagos PG PRED-TMBB2: improved topology prediction and detection of beta-barrel outer membrane proteins. *Bioinformatics* 2016, 32 (17), i665–i671. [PubMed: 27587687]
- (34). Liu Z; Gong Y; Bao Y; Guo Y; Wang H; Lin GN TMPSS: A Deep Learning-Based Predictor for Secondary Structure and Topology Structure Prediction of Alpha-Helical Transmembrane Proteins. *Front Bioeng Biotechnol* 2020, 8, 629937. [PubMed: 33569377]
- (35). Yang Y; Yu J; Liu Z; Wang X; Wang H; Ma Z; Xu D An Improved Topology Prediction of Alpha-Helical Transmembrane Protein Based on Deep Multi-Scale Convolutional Neural Network. *IEEE/ACM Trans Comput Biol Bioinform* 2020, *PP*.
- (36). Li B; Mendenhall J; Nguyen ED; Weiner BE; Fischer AW; Meiler J Improving prediction of helix-helix packing in membrane proteins using predicted contact numbers as restraints. *Proteins* 2017, 85 (7), 1212–1221. [PubMed: 28263405]
- (37). Fischer AW; Heinze S; Putnam DK; Li B; Pino JC; Xia Y; Lopez CF; Meiler J CASP11--An Evaluation of a Modular BCL::Fold-Based Protein Structure Prediction Pipeline. *PLoS One* 2016, 11 (4), e0152517. [PubMed: 27046050]
- (38). Lomize MA; Lomize AL; Pogozheva ID; Mosberg HI OPM: orientations of proteins in membranes database. *Bioinformatics* 2006, 22 (5), 623–625. [PubMed: 16397007]
- (39). Wang G; Dunbrack RL Jr. PISCES: a protein sequence culling server. *Bioinformatics* 2003, 19 (12), 1589–1591. [PubMed: 12912846]
- (40). Kabsch W; Sander C Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* 1983, 22 (12), 2577–2637. [PubMed: 6667333]
- (41). Heinig M; Frishman D STRIDE: a web server for secondary structure assignment from known atomic coordinates of proteins. *Nucleic Acids Res* 2004, 32 (Web Server issue), W500–502. [PubMed: 15215436]
- (42). Majumdar I; Krishna SS; Grishin NV PALSSE: a program to delineate linear secondary structural elements from protein structures. *BMC Bioinformatics* 2005, 6, 202. [PubMed: 16095538]
- (43). 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 Research* 2012, 40 (D1), D370–D376. [PubMed: 21890895]
- (44). Woodall NB; Hadley S; Yin Y; Bowie JU Complete topology inversion can be part of normal membrane protein biogenesis. *Protein Sci* 2017, 26 (4), 824–833. [PubMed: 28168866]
- (45). Fluman N; Tobiasson V; von Heijne G Stable membrane orientations of small dual-topology membrane proteins. *Proc Natl Acad Sci U S A* 2017, 114 (30), 7987–7992. [PubMed: 28698365]
- (46). Altschul SF; Madden TL; Schaffer AA; Zhang J; Zhang Z; Miller W; Lipman DJ Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 1997, 25 (17), 3389–3402. [PubMed: 9254694]
- (47). Remmert M; Biegert A; Hauser A; Söding J HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat Methods* 2012, 9 (2), 173–175.
- (48). Chollet F Deep Learning with Python; Manning Publications: Shelter Island, NY, 2018.

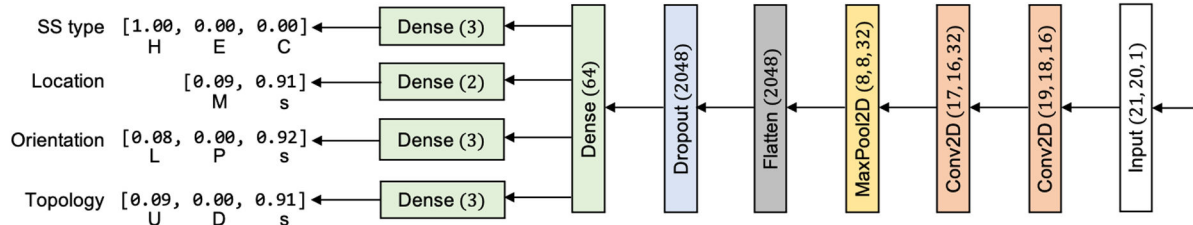
- (49). Kingma DP; Ba J, 2014.
- (50). Rost B; Sander C; Schneider R Redefining the Goals of Protein Secondary Structure Prediction. *Journal of Molecular Biology* 1994, 235 (1), 13–26. [PubMed: 8289237]
- (51). Zemla A; Venclovas C; Fidelis K; Rost B A modified definition of Sov, a segment-based measure for protein secondary structure prediction assessment. *Proteins* 1999, 34 (2), 220–223. [PubMed: 10022357]
- (52). Liu T; Wang Z SOV\_refine: A further refined definition of segment overlap score and its significance for protein structure similarity. *Source Code Biol Med* 2018, 13, 1. [PubMed: 29713370]
- (53). Hochreiter S; Schmidhuber J Long short-term memory. *Neural Comput* 1997, 9 (8), 1735–1780. [PubMed: 9377276]
- (54). Leman JK; Mueller R; Karakas M; Woetzel N; Meiler J Simultaneous prediction of protein secondary structure and transmembrane spans. *Proteins* 2013, 81 (7), 1127–1140. [PubMed: 23349002]
- (55). Buchan DWA; Jones DT The PSIPRED Protein Analysis Workbench: 20 years on. *Nucleic Acids Res* 2019, 47 (W1), W402–W407. [PubMed: 31251384]
- (56). Wang S; Li W; Liu S; Xu J RaptorX-Property: a web server for protein structure property prediction. *Nucleic Acids Res* 2016, 44 (W1), W430–435. [PubMed: 27112573]
- (57). Faraggi E; Zhang T; Yang Y; Kurgan L; Zhou Y SPINE X: improving protein secondary structure prediction by multistep learning coupled with prediction of solvent accessible surface area and backbone torsion angles. *J Comput Chem* 2012, 33 (3), 259–267. [PubMed: 22045506]
- (58). Klausen MS; Jespersen MC; Nielsen H; Jensen KK; Jurtz VI; Sonderby CK; Sommer MOA; Winther O; Nielsen M; Petersen B et al. NetSurfP-2.0: Improved prediction of protein structural features by integrated deep learning. *Proteins* 2019, 87 (6), 520–527. [PubMed: 30785653]
- (59). Kall L; Krogh A; Sonnhammer ELL A combined transmembrane topology and signal peptide prediction method. *Journal of Molecular Biology* 2004, 338 (5), 1027–1036. [PubMed: 15111065]
- (60). 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), W401–407. [PubMed: 25969446]
- (61). Almagro Armenteros JJ; Tsirigos KD; Sonderby CK; Petersen TN; Winther O; Brunak S; von Heijne G; Nielsen H SignalP 5.0 improves signal peptide predictions using deep neural networks. *Nat Biotechnol* 2019, 37 (4), 420–423. [PubMed: 30778233]
- (62). Almagro Armenteros JJ; Sonderby CK; Sonderby SK; Nielsen H; Winther O DeepLoc: prediction of protein subcellular localization using deep learning. *Bioinformatics* 2017, 33 (21), 3387–3395. [PubMed: 29036616]
- (63). Xu J Distance-based protein folding powered by deep learning. *Proc Natl Acad Sci U S A* 2019, 116 (34), 16856–16865. [PubMed: 31399549]
- (64). AlQuraishi M A watershed moment for protein structure prediction. *Nature* 2020, 577 (7792), 627–628. [PubMed: 31988401]
- (65). AlQuraishi M End-to-End Differentiable Learning of Protein Structure. *Cell Syst* 2019, 8 (4), 292–301 e293. [PubMed: 31005579]
- (66). Dauparas J; Fuchs F; <https://fabianfuchsmi.github.io/alphafold2/>, 2021; Vol. 2021.



(A) Generation of position-specific-scoring matrices (PSSM)



(B) Prediction of residue-level structural attributes with a 2D-CNN



(C) Prediction of protein class with a LSTM

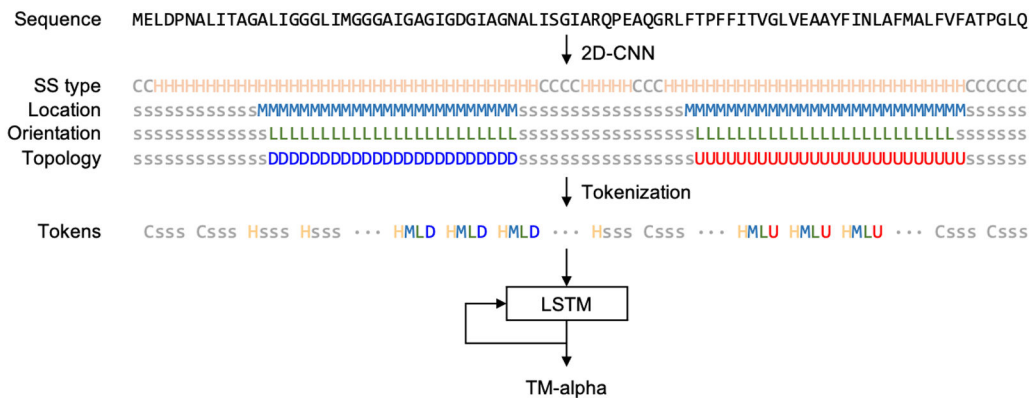
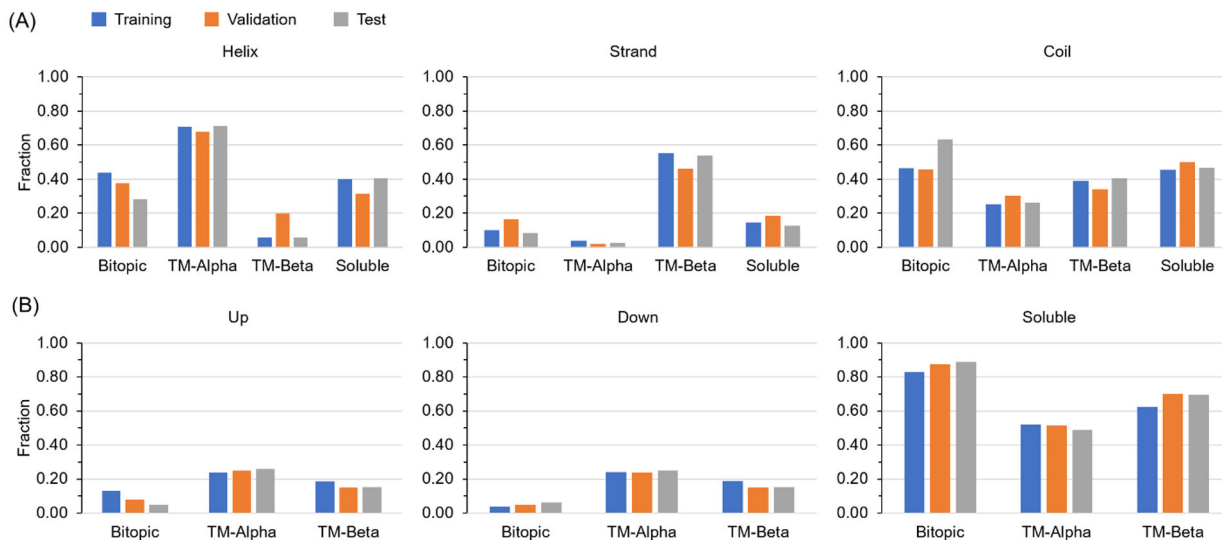


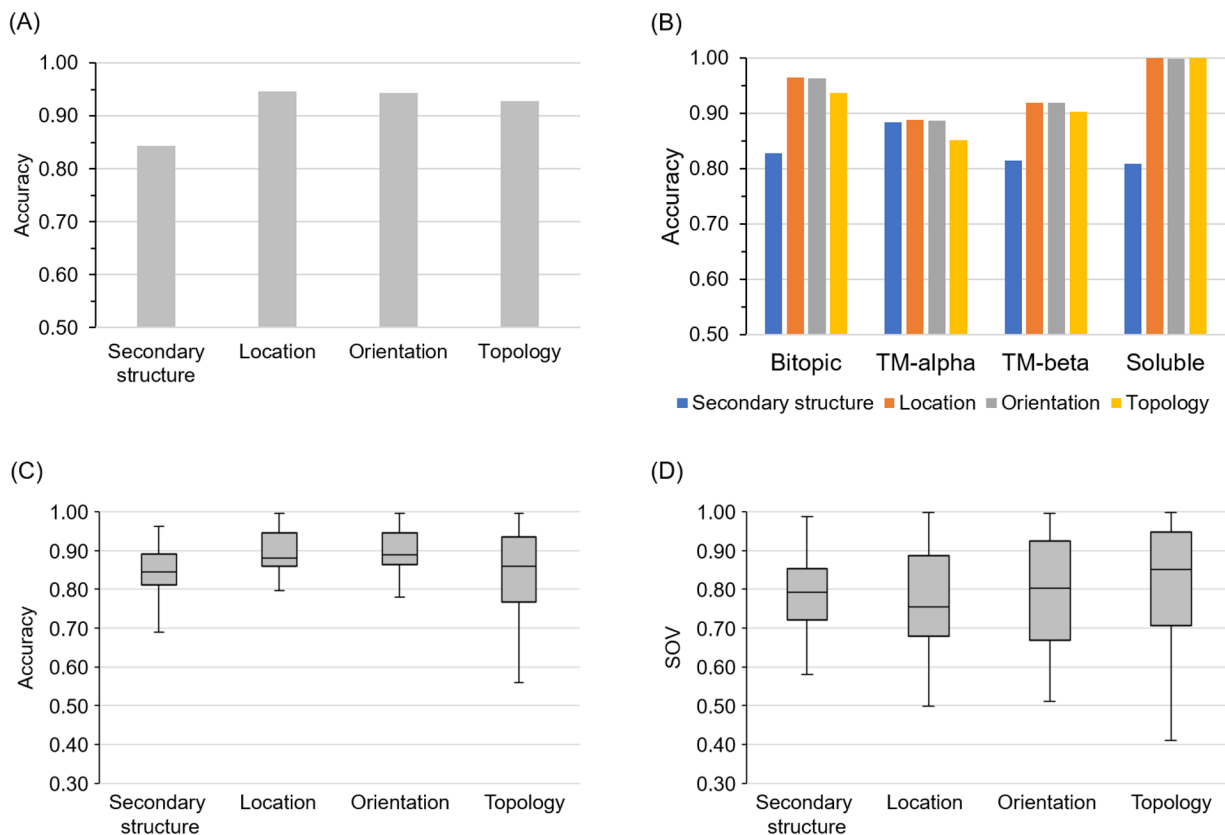
Figure 1. Design of the MASSP framework.

MASSP is a hierarchical prediction system with two predictors. The first predictor is a multi-layer 2D convolutional neural network-based (2D-CNN) residue-level classifier trained to predict structural attributes of each residue in the input protein sequence (A and B). MASSP predicts four categories of residue-level attributes, namely secondary structure types (helix, strand, or coil, indicated by H, E, or C), location (membrane or solution, indicated by M or s), orientation (lipid-, pore-facing, or soluble indicated by L, P, or s), and transmembrane topology (up, down, or soluble indicated by U, D, or s). The second predictor is a long short-term memory (LSTM) recurrent neural network-based sequence-level classifier trained to predict the protein class (bitopic, TM-alpha, TM-beta, soluble) of the input amino acid sequence (C). As input to this LSTM, we create a four-letter token by collapsing residue-level predictions made by the 2D-CNN for each residue. Thus, for a protein sequence that has *l* residues, the input to the LSTM will be *l* four-letter tokens.



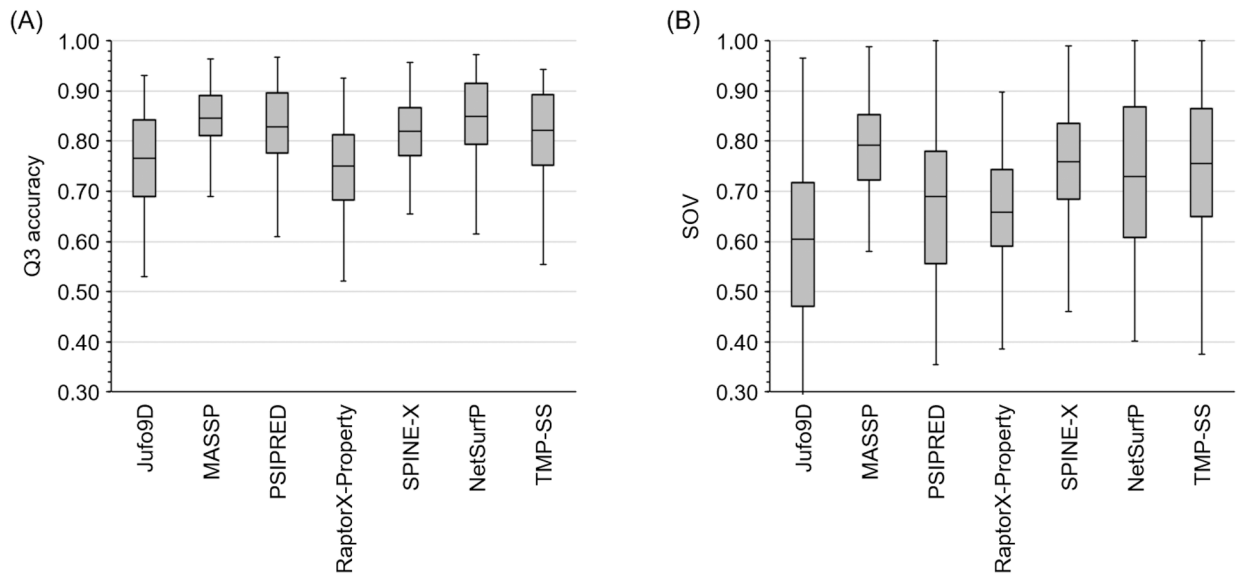
**Figure 2. Residue secondary structures and topologies are evenly partitioned across protein types for training and evaluation.**

(A) The distribution of secondary structure types among training, validation, and test sets for all four protein classes. (B) The distribution of residue topology types among training, validation, and test sets for three protein classes. The distribution for soluble proteins is omitted because the topology types of all soluble protein residues are labeled as “s”.



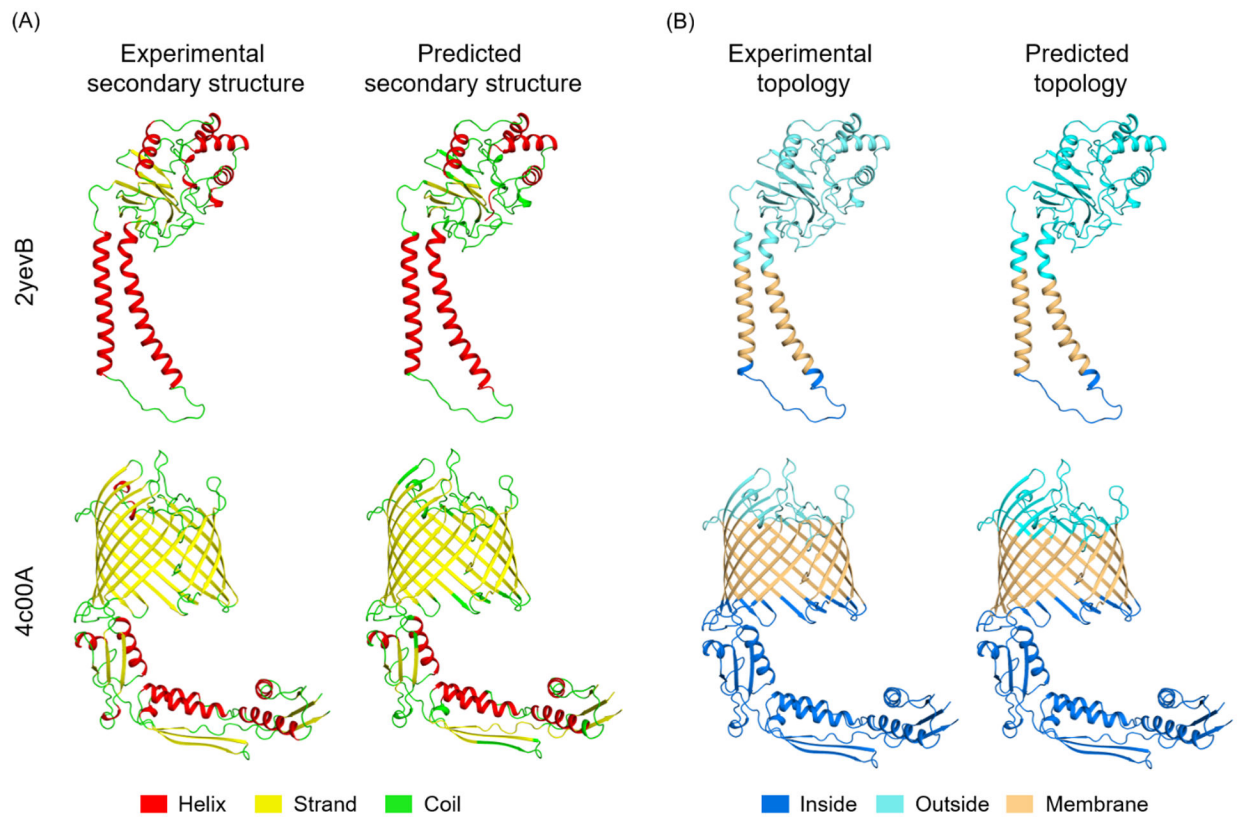
**Figure 3. Performance of MASSP on predicting structural features as tested using a held-out test set.**

(A) Overall accuracy achieved by MASSP for each target attribute. (B) Overall accuracy for each target attribute decomposed according to protein class. (C) Distribution of prediction accuracy for each target attribute on test set proteins. (D) Distribution of SOV metric for each target attribute on test set proteins.



**Figure 4. MASSP predicts secondary structure accurately.**

(A) Distribution of Q3 secondary structure accuracies of MASSP on the test set compared to six other representative methods. (B) Distribution of performance at secondary structure prediction in terms of the SOV metric for MASSP and six other representative methods.



**Figure 5. Examples of accurate secondary structure and transmembrane topology prediction by MASSP.**

(A) Comparisons of experimental and predicted secondary structures mapped to tertiary structures for cytochrome C oxidase subunit 2 (PDB: 2yevB) and translocation and assembly module subunit TamA (PDB: 4c00A) illustrating high accuracy. (B) Comparisons of experimental and predicted residue topologies mapped to tertiary structure for the same proteins.

**Table 1.**

Summary of the performance of the LSTM-based protein class predictor.

	<b>Bitopic</b>	<b>TM-alpha</b>	<b>TM-beta</b>	<b>Soluble</b>
<b>Bitopic</b>	5	0	0	0
<b>TM-alpha</b>	0	23	0	0
<b>TM-beta</b>	0	0	7	1
<b>Soluble</b>	0	0	0	37

Row headers indicate true protein classes; column headers indicate predicted protein classes; numbers in the matrices are protein counts.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

**Table 2.**

Comparison on the performance of TMH and topology prediction for TM-alpha proteins (23 proteins in total).

Method	Correct #TMH	Correct topology	Q3	SOV
Jufo9D	131 / 149 (87.9%)	NA	NA	NA
MASSP	<b>146 / 149 (98.0%)</b>	22 / 23 (95.6%)	0.871	0.850
MEMSAT3	134 / 149 (89.9%)	20 / 23 (87.0%)	0.851	0.813
OCTOPUS	134 / 149 (89.9%)	<b>23 / 23 (100%)</b>	<b>0.874</b>	<b>0.861</b>
TMHMM2	129 / 149 (86.6%)	16 / 23 (69.7%)	0.790	0.772
TMP-SS	130 / 138 (94.2%)	NA	NA	NA
TOPCONS2	130 / 149 (87.2%)	20 / 23 (87.0%)	0.830	0.801

NA: not available, Jufo9D and TMP-SS do not predict residue-level three-state topology. A predicted TMH is considered correct if it overlaps with the true TMH for at least 10 residues. Our local run of TMP-SS was not successful for 3dh4A, hence the total number of TMHs is 138.

**Table 3.**

Comparison on the performance of TMH and topology prediction for bitopic transmembrane proteins (5 proteins in total).

Method	Correct #TMH	Correct topology	Q3	SOV
Jufo9D	4 / 5	NA	NA	NA
MASSP	<b>5 / 5</b>	3 / 5	0.551	0.531
MEMSAT3	4 / 5	<b>4 / 5</b>	0.831	0.819
OCTOPUS	4 / 5	<b>4 / 5</b>	<b>0.865</b>	<b>0.869</b>
TMHMM2	4 / 5	3 / 5	0.629	0.630
TMP-SS	5 / 5	NA	NA	NA
TOPCONS2	3 / 5	4 / 5	0.589	0.584

NA: not available, Jufo9D and TMP-SS do not predict residue-level three-state topology. A predicted TMH is considered correct if it overlaps with the true TMH for at least 10 residues.



**Table 4.**

Comparison on the performance of TMS and topology prediction for TM-beta proteins (8 proteins int total).

Method	Correct #TMS	Correct topology	Q3	SOV
Jufo9D	82 / 104 (78.8%)	NA	NA	NA
MASSP	<b>98 / 104 (94.1%)</b>	5 (64.3%)	<b>0.775</b>	<b>0.775</b>
BOCTOPUS2	<b>98 / 104 (94.1%)</b>	<b>6 (75.0%)</b>	0.696	0.745
PRED-TMBB2	85 / 104 (81.7%)	5 (64.3%)	0.714	0.716

NA: not available, Jufo9D does not predict residue-level three-state topology. A predicted TMS is considered correct if it overlaps with the true TMS for at least 5 residues.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript