frontiers in Bioengineering and Biotechnology

ORIGINAL RESEARCH
published: 25 January 2021
doi: 10.3389/fbioe.2020.629937

TMPSS: A Deep Learning-Based Predictor for Secondary Structure and Topology Structure Prediction of Alpha-Helical Transmembrane Proteins

Zhe Liu[1,2], *Yingli Gong*[3], *Yihang Bao*[4], *Yuanzhao Guo*[4], *Han Wang*[4]* and *Guan Ning Lin*[1,2]*

[1] Shanghai Mental Health Center, Shanghai Jiao Tong University School of Medicine, School of Biomedical Engineering, Shanghai Jiao Tong University, Shanghai, China, [2] Shanghai Key Laboratory of Psychotic Disorders, Shanghai, China, [3] College of Intelligence and Computing, Tianjin University, Tianjin, China, [4] School of Information Science and Technology, Institute of Computational Biology, Northeast Normal University, Changchun, China

OPEN ACCESS

Edited by:
Zhibin Lv,
University of Electronic Science and Technology of China, China

Reviewed by:
Pu-Feng Du,
Tianjin University, China
Shiwei Sun,
Chinese Academy of Sciences (CAS), China
You Zhou,
Jilin University, China

***Correspondence:**
Han Wang
wangh101@nenu.edu.cn
Guan Ning Lin
nickgnlin@sjtu.edu.cn

Specialty section:
This article was submitted to Synthetic Biology, a section of the journal Frontiers in Bioengineering and Biotechnology

Received: 16 November 2020
Accepted: 10 December 2020
Published: 25 January 2021

Citation:
Liu Z, Gong Y, Bao Y, Guo Y, Wang H and Lin GN (2021) TMPSS: A Deep Learning-Based Predictor for Secondary Structure and Topology Structure Prediction of Alpha-Helical Transmembrane Proteins. Front. Bioeng. Biotechnol. 8:629937. doi: 10.3389/fbioe.2020.629937

Alpha transmembrane proteins (αTMPs) profoundly affect many critical biological processes and are major drug targets due to their pivotal protein functions. At present, even though the non-transmembrane secondary structures are highly relevant to the biological functions of αTMPs along with their transmembrane structures, they have not been unified to be studied yet. In this study, we present a novel computational method, TMPSS, to predict the secondary structures in non-transmembrane parts and the topology structures in transmembrane parts of αTMPs. TMPSS applied a Convolutional Neural Network (CNN), combined with an attention-enhanced Bidirectional Long Short-Term Memory (BiLSTM) network, to extract the local contexts and long-distance interdependencies from primary sequences. In addition, a multi-task learning strategy was used to predict the secondary structures and the transmembrane helixes. TMPSS was thoroughly trained and tested against a non-redundant independent dataset, where the Q3 secondary structure prediction accuracy achieved 78% in the non-transmembrane region, and the accuracy of the transmembrane region prediction achieved 90%. In sum, our method showcased a unified model for predicting the secondary structure and topology structure of αTMPs by only utilizing features generated from primary sequences and provided a steady and fast prediction, which promisingly improves the structural studies on αTMPs.

Keywords: protein secondary structure, protein topology structure, deep learning, alpha-helical transmembrane proteins, long short-term memory networks

INTRODUCTION

Membrane proteins (MPs) are pivotal players in several physiological events, such as signal transduction, neurotransmitter adhesion, ion transport, etc. (Goddard et al., 2015; Roy, 2015). While transmembrane proteins (TMPs), as an essential type of MPs, span the entire biological membrane with segments exposed to both the inside and the outside of the lipid bilayers (Stillwell, 2016). As the major class of TMPs, alpha-helical TMPs are given great pharmacological importance,

accounting for about 60% of known drug targets in the current benchmark (Wang et al., 2019). Nevertheless, the difficulties of acquiring their crystal structures always stand in our way due to their low solubilities in the buffers typically used in 2D-PAGE (Butterfield and Boyd-Kimball, 2004; Nugent et al., 2011). All of this is calling for accurate computational predictors.

Predicting alpha-helical TMPs' tertiary structure directly from amino acid sequences has been a challengeable task in computational biology for many years (Yaseen and Li, 2014), but some indirect measures may be worth considering. Since Pauling et al. (1951) performed the first protein secondary structure prediction in 1951, many indicators on the secondary structure level of proteins, such as topology structure (Wang et al., 2019), surface accessibility (Lu et al., 2019), have been demonstrated to be strongly associated with the 3D information of TMPs. Specifically, the secondary structure helps to identify function domains and guides the design of site-specific mutation experiments (Drozdetskiy et al., 2015), whereas the topology structure can help reveal the relative position relationship between TMPs and membranes (Tusnady and Simon, 2001). Generally, the performance of protein secondary structure prediction can be measured by Q3 accuracy in a 3-class classification, i.e., helix (H), strand (E), and coil (C), or Q8 accuracy in an 8-class classification under a more sophisticated evaluation system. Q3 is preferred according to its low cost and close ability in depicting the secondary structure compared with Q8.

Progress in the structure prediction for MPs is slower than that for soluble proteins (Xiao and Shen, 2015). At present, state-of-the-art methods aiming at predicting the secondary structure based on primary sequences, such as SSpro/ACCpro 5 (Magnan and Baldi, 2014), JPred4 (Drozdetskiy et al., 2015), PSIPRED 4 (Buchan and Jones, 2019), and MUFOLD-SSW (Fang et al., 2020), are all trained on soluble protein-specific datasets. However, none of those mentioned methods can simultaneously predict the secondary structure and topology structure of alpha-helical TMPs. More specifically, existing tools could not distinguish transmembrane helices of TMPs from non-transmembrane ones and, in-term, would weaken the TMPs' structure prediction specificity. Another common challenge among the available methods is that features fed into these models are often too miscellaneous, making the model prediction low efficient and even difficult for users to understand. Thus, a more suitable and practical tool for assisting the structure prediction of TMPs is greatly needed.

Deep learning has been employed in several protein sequence classification problems (Lv et al., 2019; Wei et al., 2019; Zeng et al., 2020). Here, we proposed a deep learning-based predictor named TMPSS to predict the secondary structure and topology structure of alpha-helical TMPs simultaneously using amino acid sequences. Equipped with a robust network and carefully screened input features, TMPSS ignored input length restriction and achieved the highest output efficiency compared with other state-of-the-art methods with an acceptable Q3 performance of secondary structure prediction in the full chain (see **Figure 1**). In addition, our TMPSS achieved the Q3 of a whopping 0.97 in the transmembrane region, suggesting that almost all the transmembrane helices were identified. Moreover, TMPSS also significantly outperformed other existing topology structure predictors with the prediction accuracy of 0.90 and the Matthew Correlation Coefficient (MCC) of 0.76 using an independently generated dataset. TMPSS implemented a deep neural network by grouped multiscale Convolutional Neural Networks (CNNs) and stacked attention-enhanced Bidirectional Long Short-Term Memory (BiLSTM) layers for capturing local contexts and global dependencies, respectively. We also utilized the multi-task learning technique to improve prediction performance by considering the mutual effects between different protein properties. We have released TMPSS as a publicly available prediction tool for the community. The pre-trained model and support materials are both available at https://github.com/NENUBioCompute/TMP-SS.

MATERIALS AND METHODS

Benchmark Datasets

As illustrated above, none of the existing secondary structure predictors available today are specific to TMPs. Thus, it is necessary to create unique datasets that contain only alpha-helical TMPs for targeted research. The Protein Data Bank of transmembrane proteins (PDBTM) (Kozma et al., 2012), the first up-to-date and comprehensive TMP selection of the Protein Data Bank (PDB) (Burley et al., 2017), was chosen to construct our datasets. We downloaded 4,336 alpha-helical TMPs from PDBTM (version: 2020-2-7) and removed the chains that contained unknown residues (such as "X") and whose length was <30 residues.

To reduce the redundancy of data and avoid the influence of homology bias (Zou et al., 2020), we utilized CD-HIT (Fu et al., 2012) with a 30% sequence identity cut-off and obtained 911 protein chains. These protein chains were then randomly divided into a training set of 811 chains, a validation set of 50 chains, and a test set (named "TEST50") of 50 chains. Secondary structure labels were obtained by the DSSP program (Kabsch and Sander, 1983) through PDB files, and topology structures were collected from PDBTM. All the experiments were conducted on five-fold cross-validation to gauge its generalization performances (Walsh et al., 2016). The results were used to evaluate our model and compare against other predictors. The overview of AA composition of the training set, validation set, and TEST50 is shown in **Table 1**.

Features and Input Encoding

Features are the key issue for the machine learning tasks (Patil and Chouhan, 2019; Zhang and Liu, 2019). Prediction of alpha-helical TMPs' secondary structure and topology structure at the residue level is formulated as follows: for a given primary protein sequence of an alpha-helical TMP, a sliding window whose length is L residues is used to predict the secondary structure and topology structure of the central residue. For example, if L is 19, each protein will be sliced into fragments of 19 amino acids. Providing valuable input features to deep learning networks is of great importance to make predictions more accurate. Here, we carefully selected two encoding features to represent the protein

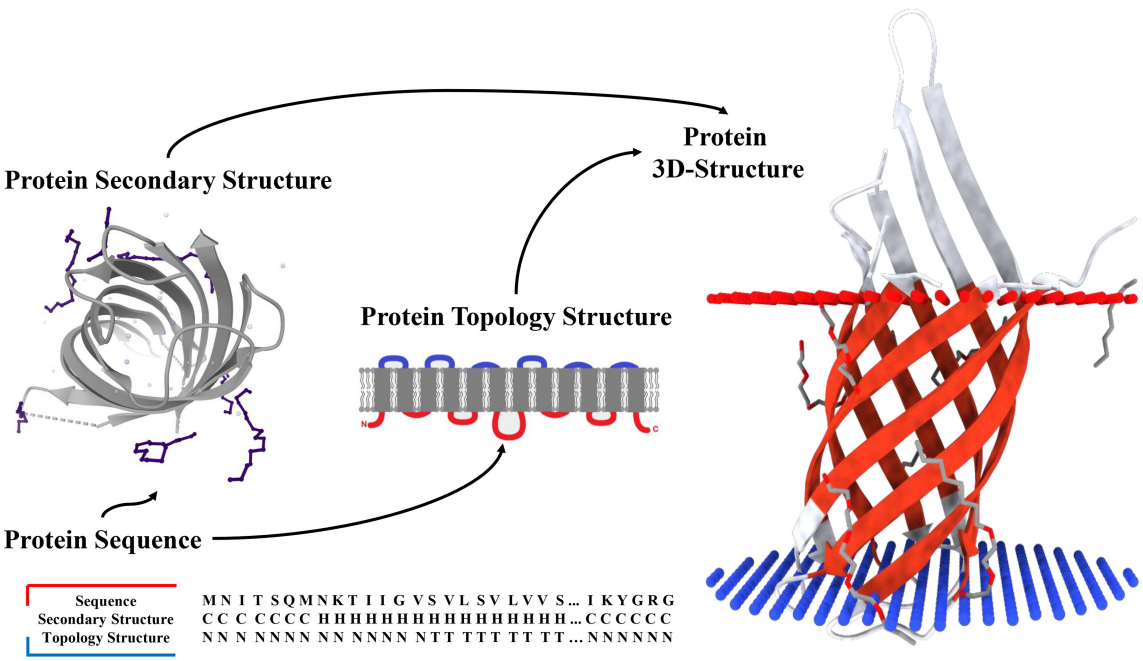

FIGURE 1 | Secondary structure and topology structure prediction of alpha-helical transmembrane proteins.

TABLE 1 | Overview of AA composition of the training set, validation set, and TEST50.

3-State	8-State	Training set		Validation set		TEST50	
Helices	G	60.1%	5,090	59.0%	438	55.1%	317
	H		119,987		7,931		7,897
	I		3,101		254		192
Strands	E	6.3%	12,226	6.5%	853	8.9%	1,240
	B		1,295		103		110
Coils	C	33.5%	34,372	34.5%	2,298	36.0%	2,607
	S		17,861		1,332		1,397
	T		19,195		1,409		1,488

fragment: one-hot code and HHblits profile (Remmert et al., 2012).

The first set came from the protein profiles generated by HHblits, which is faster, almost twice as sensitive, and provides more accurate evolutionary information for protein sequence than PSI-BLAST (Steinegger et al., 2019). We found the best results against the database named uniprot20_2016_02 with three iterations, an E-value threshold of 0.001, and other default settings. The obtained H_{hhm} matrix consisted of 31 dimensions, 30 of which were HMM profile values and one reflected *NoSeq* label (representing a gap) (Fang et al., 2018) at the last column. Each of H_{ij} in the matrix was scaled by a variation of sigmoid function [see Equation (1)], making the distribution of features more uniform and reasonable.

$$f(\mathbf{t}) = \frac{\mathbf{10}}{\mathbf{1}+e^{-\frac{t}{\mathbf{2000}}}} \quad (1)$$

We then adopted a 21-dimensional matrix O_{onehot} as our second set containing a simple one-hot encoding of 20 positions with one *NoSeq* label. The past research suggested that one-hot encoding was straightforward to generate and has been successfully used in protein structure prediction-associated tasks (Ding and Li, 2015). Therefore, we used 19 dimensional "0" vector with a "1" to represent AA at the index of a particular protein sequence. We mapped each protein fragment sliced by the sliding window with this encoding strategy into an undisturbed coding within local position information.

Model Design

Network Architecture

As a deep learning-based predictor, TMPSS can predict the secondary structure and topology structure of alpha-helical TMPs simultaneously. As we can see in **Figure 2**, the four parts of our model are feature-integration layers for input feature preprocessing, grouped multiscale CNN layers, attention-enhanced BiLSTM layer, and fully-connected layers by two softmax outputs in the end.

Our network's input carried two types of features generated from primary sequences, amino acid features, and profile features. These preprocessed features were fed into a grouped multiscale CNN layer to capture local position information and prevent their mutual interferences at the same time. Then, the merged CNN output flew into two stacked BiLSTM layers, which turned out to be skilled in extracting long-term dependencies and global information (Zhou et al., 2016). We also proposed the attention mechanism as a simple dense layer to help LSTM know which unit's output should be paid more attention. At

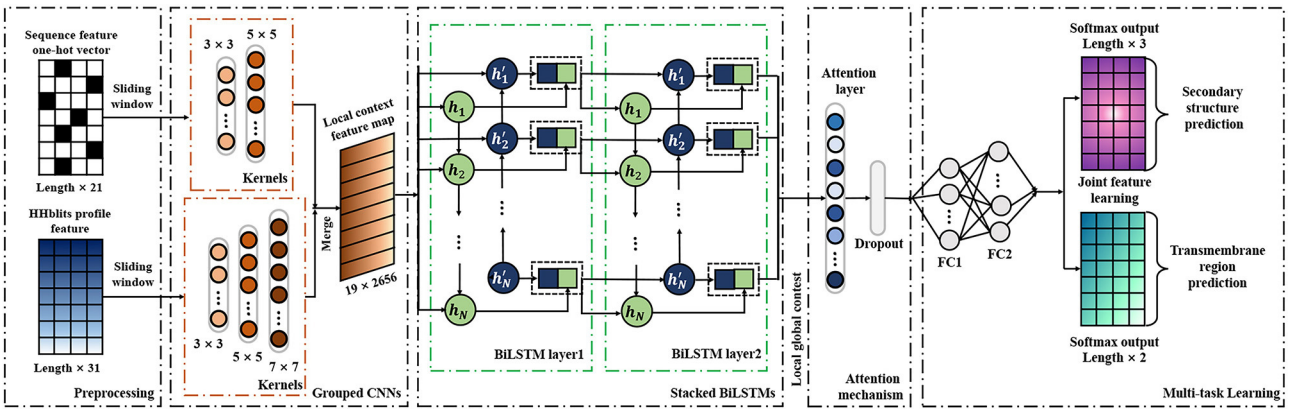

FIGURE 2 | Diagram of TMPSS.

the end of the components mentioned above, there were two fully-connected hidden layers with a softmax-activated output layer, which performed a 3-category secondary structure and 2-category topology structure classification. More details of grouped multiscale CNNs and attention-enhanced BiLSTM are discussed in the **Supplementary Material**.

Implementation Details

Our model was implemented, trained, and tested using the open-source software library Keras (Gulli and Pal, 2017) and Tensorflow (Abadi et al., 2016) on an Nvidia 1080Ti GPU. Main hyperparameters, such as sliding window length, training dropout rate, and number of LSTM units, were explored, and an early stopping strategy and a save-best strategy were adopted (Fang et al., 2018). When the validation loss did not reduce in 10 epochs during training time, the training process would be stopped, and the best model parameters would be saved. In all cases, the weights were initialized by default setting in Keras; the parameters were trained using an Adam optimizer (Bello et al., 2017) to change the learning rate during model training dynamically. Furthermore, batch normalization layers (Ioffe and Szegedy, 2015) and a Dropout layer (Gal et al., 2017) (rate = 0.30) were utilized since they were both skilled in avoiding the network from overfitting and improving the speed of the training process effectively. We set the sliding window's length as 19 residues and put 700 units in each LSTM layer according to the hyperparameter tuning results in this study.

Performance Evaluation

A commonly used evaluation metric for both secondary structure and topology structure prediction based on the residue level is accuracy (ACC), and in particular, Q3 was widely used as a performance metric for 3-category secondary structure prediction (Fang et al., 2017). To quantitatively evaluate the performance of TMPSS and other predictors at the residue level, they were assessed by six measures, including accuracy, recall, precision, specificity, MCC, and F-measure (Tan et al., 2019; Yang et al., 2019; Zhu et al., 2019). The calculation formulas of these evaluation parameters were illustrated as follows:

$$Accuracy = \frac{TN+TP}{TP+FN+FP+TN} \tag{2}$$

$$Recall = \frac{TP}{TP+FN} \tag{3}$$

$$Precision = \frac{TP}{TP+FP} \tag{4}$$

$$Specificity = \frac{TN}{FP+TN} \tag{5}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \tag{6}$$

$$F-measure = 2 \times \frac{Recall \times Precision}{Recall+Precision} \tag{7}$$

where TN, TP, FN, and FP, respectively denoted true negative, true positive, false negative, and false positive samples.

RESULTS

Prediction Performance Analysis at the Residue Level

To evaluate the prediction performance of each category in both two classification tasks at the residue level, we used the confusion matrices (see **Figure 3**), Receiver Operating Characteristic (ROC) curves, and Precision–Recall (PR) curves (see **Figure 4**) to visualize the predict results of TMPSS on TEST50. As illustrated in **Table 1**, TEST50 contains a total of 15,248 residues labeled by "H" (helix), "E" (strand), or "C" (coil) in secondary structure prediction and "T" (transmembrane helix) or "N" (non-transmembrane residue) in topology structure prediction.

Figures 3A,B shows the confusion matrices of secondary structure prediction in the full chain and non-transmembrane region, respectively. As we can see, class "H" was predicted

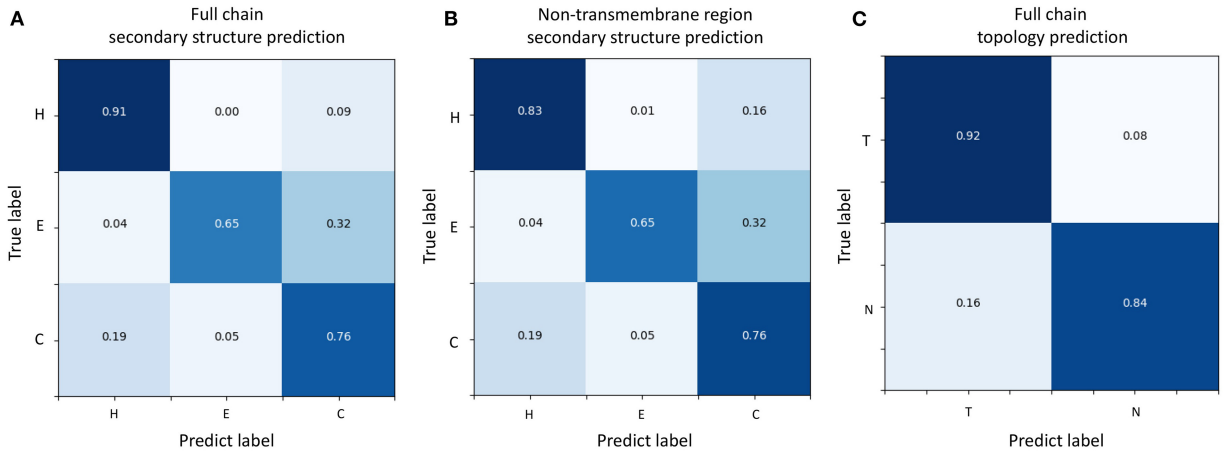

FIGURE 3 | Confusion matrices of TMPSS's prediction performance. **(A)** Confusion matrix of secondary structure prediction in the full chain. **(B)** Confusion matrix of secondary structure prediction in the non-transmembrane region. **(C)** Confusion matrix of topology structure prediction in the full chain.

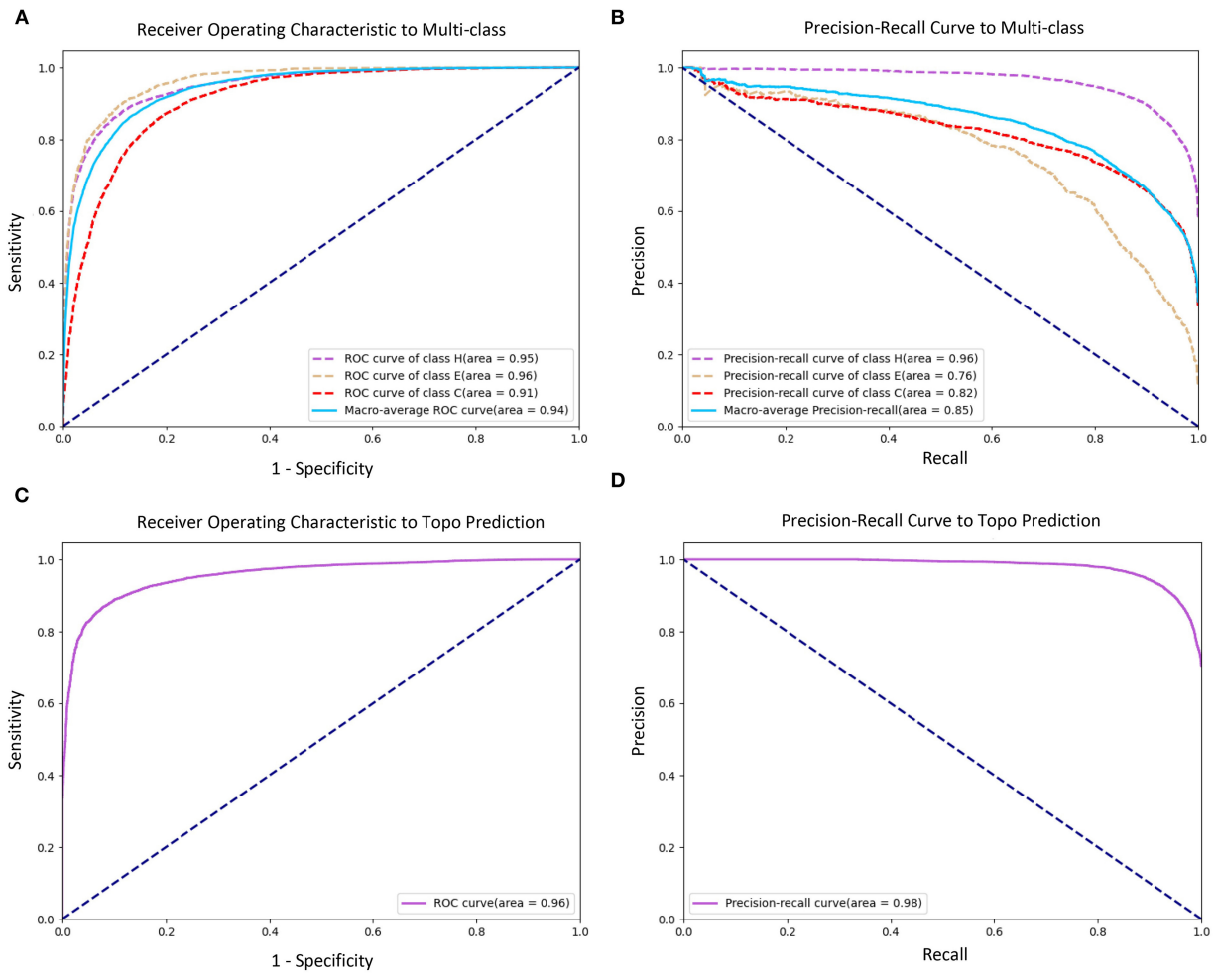

FIGURE 4 | Receiver Operating Characteristic (ROC) and Precision–Recall (PR) curves of prediction performance. **(A)** ROC curve of secondary structure prediction. **(B)** PR curve of secondary structure prediction. **(C)** ROC curve of topology structure prediction. **(D)** PR curve of topology structure prediction.

TABLE 2 | Comparison of TMPSS with previous secondary structure predictors on TEST50 in the full chain.

Method	Class	R	P	S	MCC	F	Full chain SS Q3	Limitation of input length (residues)	Time cost (min)
SSpro5 (with templates)	H	0.908	**0.942**	**0.923**	**0.826**	**0.925**	**0.90**	Limited to 1,500	980
	E	**0.908**	**0.778**	0.975	**0.824**	0.838			
	C	**0.870**	**0.854**	**0.926**	**0.792**	**0.862**			
PSIPRED 4	H	0.907	0.880	0.829	0.741	0.893	0.83	Limited to 1,500	490
	E	0.726	0.735	0.975	0.705	0.731			
	C	0.731	0.770	0.891	0.631	0.750			
RaptorX-Property	H	0.897	0.910	0.877	0.772	0.903	0.85	–	114
	E	0.771	0.761	0.977	0.743	0.766			
	C	0.786	0.770	0.883	0.666	0.778			
Porter 5	H	0.919	0.893	0.849	0.773	0.906	0.85	–	1,035
	E	0.757	0.763	0.977	0.737	0.760			
	C	0.758	0.796	0.903	0.670	0.777			
DeepCNF	H	0.867	0.908	0.879	0.741	0.887	0.83	–	3,000
	E	0.741	0.703	0.970	0.694	0.722			
	C	0.791	0.743	0.864	0.645	0.766			
Spider3	H	0.927	0.883	0.831	0.766	0.904	0.85	–	720
	E	0.751	0.765	0.978	0.734	0.758			
	C	0.737	0.803	0.910	0.662	0.769			
SPOT-1D	H	**0.931**	0.884	0.832	0.772	**0.907**	0.85	Limited to 750	2,030
	E	0.821	0.767	0.976	0.773	0.793			
	C	0.731	0.822	0.921	0.673	0.774			
MUFOLD-SSW	H	0.920	0.884	0.833	0.760	0.902	0.85	Limited to 700	150
	E	0.820	0.743	0.973	0.758	0.779			
	C	0.724	0.815	0.918	0.663	0.767			
JPred4	H	0.830	0.908	0.884	0.706	0.867	0.80	Limited to 800	110
	E	0.664	0.602	0.958	0.595	0.632			
	C	0.772	0.689	0.826	0.583	0.728			
TMPSS	H	0.907	0.888	0.842	0.752	0.897	0.84	–	**96**
	E	0.646	0.764	**0.981**	0.677	0.700			
	C	0.763	0.759	0.880	0.641	0.761			

H, helix (DSSP classes H, G, and I); E, strand (DSSP classes E and B); C, coil (DSSP classes S, T, and blank).
R, Recall; P, Precision; S, Specificity; F, F-measure. Bold fonts represent the best experimental results.

with great precision in different regions of TMPs, but the results of class "E" and class "C" were less satisfactory. A similar experimental phenomenon existed in **Figures 4A,B** simultaneously. Helices account for the largest proportion and make the prediction more significant by considering our dataset's characteristics. The matrices demonstrate that TMPSS did well in both full chain and non-transmembrane region prediction of secondary structure on TEST50, confirming it to be a suitable secondary structure predictor for TMPs.

As for topology structure prediction, TMPSS is also an effective method. The confusion matrix of topology structure prediction in the full chain (see **Figure 3C**) proves that the output results performed well, whether for class "T" or class "N." The ROC and PR curves (see **Figures 4C,D**) also support the above conclusion. After doing a thorough analysis of TMPSS's prediction performance at the residue level on TEST50, it can be seen that TMPSS is a reliable and convenient tool for predicting the secondary structure and topology structure of alpha-helical TMPs synchronously.

Assessment of Multiple Predictors on TEST50

We tested TMPSS against SSpro5 (Magnan and Baldi, 2014) (with templates), PSIPRED 4 (Buchan and Jones, 2019), RaptorX-Property (Wang et al., 2016a), Porter 5 (Torrisi et al., 2019), DeepCNF (Wang et al., 2016b), Spider3 (Heffernan et al., 2017), SPOT-1D (Hanson et al., 2019), MUFOLD-SSW (Fang et al., 2020), and JPred4 (Drozdetskiy et al., 2015) on the TEST50 we created (see **Table 2**). Experimental results illustrated that SSpro5 (with templates) was the most accurate 3-state predictor in our tests on TEST50 in the full chain with a Q3 of 0.90. It might be probably because of the contribution of templates.

TABLE 3 | Comparison of TMPSS with previous secondary structure predictors on TEST50 in the different transmembrane regions.

Method	Trans SS Q3	Non-trans SS Q3
SSpro5 (with templates)	0.90	**0.89**
PSIPRED 4	0.94	0.79
RaptorX-Property	0.95	0.80
Porter 5	0.95	0.81
DeepCNF	0.91	0.80
Spider3	0.95	0.80
SPOT-1D	0.95	0.81
MUFOLD-SSW	0.94	0.81
JPred4	0.90	0.75
TMPSS	**0.97**	0.78

Trans, transmembrane region; Non-trans, non-transmembrane region. Bold fonts represent the best experimental results.

TABLE 4 | Comparison of TMPSS with state-of-the-art topology predictors on TEST50 in the full chain.

Method	ACC	MCC
HMMTOP 2	0.84	0.64
OCTOPUS	0.87	0.71
TOPCONS	0.88	0.72
Philius	0.87	0.71
PolyPhobius	0.88	0.72
SCAMPI	0.87	0.70
SPOCTOPUS	0.87	0.71
TMPSS	**0.90**	**0.76**

Bold fonts represent the best experimental results.

TABLE 5 | Effect of loss weight during multi-task learning.

Loss weight (λ_1:λ_2)	SS Q3	Topo ACC
1:0.1	0.832	0.887
1:0.3	0.833	0.892
1:0.5	**0.835**	**0.896**
1:0.7	0.825	0.892
1:1	0.830	0.894
1:5	0.811	0.889
1:10	0.794	0.892

Bold fonts represent the best experimental results.

However, apart from SSpro5 (with templates), the remaining servers performed similarly with the maximum Q3 deviation of 0.02, and some servers, such as JPred4, even performed worse. Many methods refused to accept sequences of more than a certain length. By comparison, TMPSS was user-friendly with no length limitation of input and had the highest output efficiency among the existing methods with an acceptable Q3 of 0.84 in the full chain.

It is worth emphasizing that this comparison shown in **Table 2** is "unfair" for our experimental tool. Firstly, the existing secondary structure predictors cannot distinguish the transmembrane "H's" from non-transmembrane "H's", whereas ours can. Secondly, some tools, such as SSpro5, uses templates, which cannot be found when making predictions about unknown structural sequences and not recommended to use under normal circumstances.

However, the tools suitable for water-soluble proteins may not be suitable for handling the residues in the transmembrane region of TMPs since they cannot distinguish transmembrane helices from non-transmembrane helices. To assess different servers' secondary structure prediction ability in the different transmembrane regions, we calculated the precision of both transmembrane and non-transmembrane residues and listed the results in **Table 3**. As expected, TMPSS achieved the best Q3 performance among all exemplified servers in the transmembrane region, which signified that almost all the transmembrane helices were identified by our method.

As for topology prediction, we compared TMPSS to state-of-the-art topology predictors, including HMMTOP 2 (Tusnady and Simon, 2001), OCTOPUS (Viklund and Elofsson, 2008), TOPCONS (Tsirigos et al., 2015), Philius (Reynolds et al., 2008), PolyPhobius (Jones, 2007), SCAMPI (Bernsel et al., 2008), and SPOCTOPUS (Viklund et al., 2008). As illustrated in **Table 4**, TMPSS obtains the best ACC (= 0.90) and MCC (= 0.76) performance on TEST50 in the full chain among the listed methods. The most probable cause is that the joint feature learning helped two prediction tasks promote each other. According to this, the deep convolutional BiLSTM extracted the most effective information though there are only two features exploited.

Multi-Task Learning

Secondary structure prediction and topology structure prediction of alpha-helical TMPs are highly related tasks since the residues labeled "T" (transmembrane helix) in topology structure prediction also have the label of "H" (helix) in secondary structure prediction (Chen et al., 2002). Therefore, we put these two tasks together to support multi-task learning (Zhang and Yeung, 2012) and generated a 3-class secondary structure and a 2-class topology structure simultaneously. With the help of multi-task learning, our model's computational complexity was significantly reduced compared with other methods based on cascaded deep learning networks. The joint loss function could be formulated as follows:

$$L\left(\{s_i, t_i\}\right) = \frac{\lambda_1}{N}\sum L_s\left(s_i, s_i^*\right) + \frac{\lambda_2}{N}\sum L_t\left(t_i, t_i^*\right) \quad (8)$$

where $L_s\left(s_i, s_i^*\right) = -s_i^* log(s_i)$ and $L_t\left(t_i, t_i^*\right) = -[t_i^* log(t_i) + (1 - t_i^*) log(1 - t_i)]$ are respective loss functions for secondary structure and topology structure prediction, s_i and t_i are predicted probabilities (softmax output) of secondary structure labels and topology structure labels, respectively, s_i^* and t_i^* are ground-truth labels of secondary structure and topology structure, respectively, λ_1 and λ_2 are loss weight of combined loss function, and N is the total number of residues. **Table 5** shows the effect of different loss weights ($\lambda_1 : \lambda_2$) during multi-task learning on the validation dataset, and we set $\lambda_1 = 1$, $\lambda_2 = 0.5$

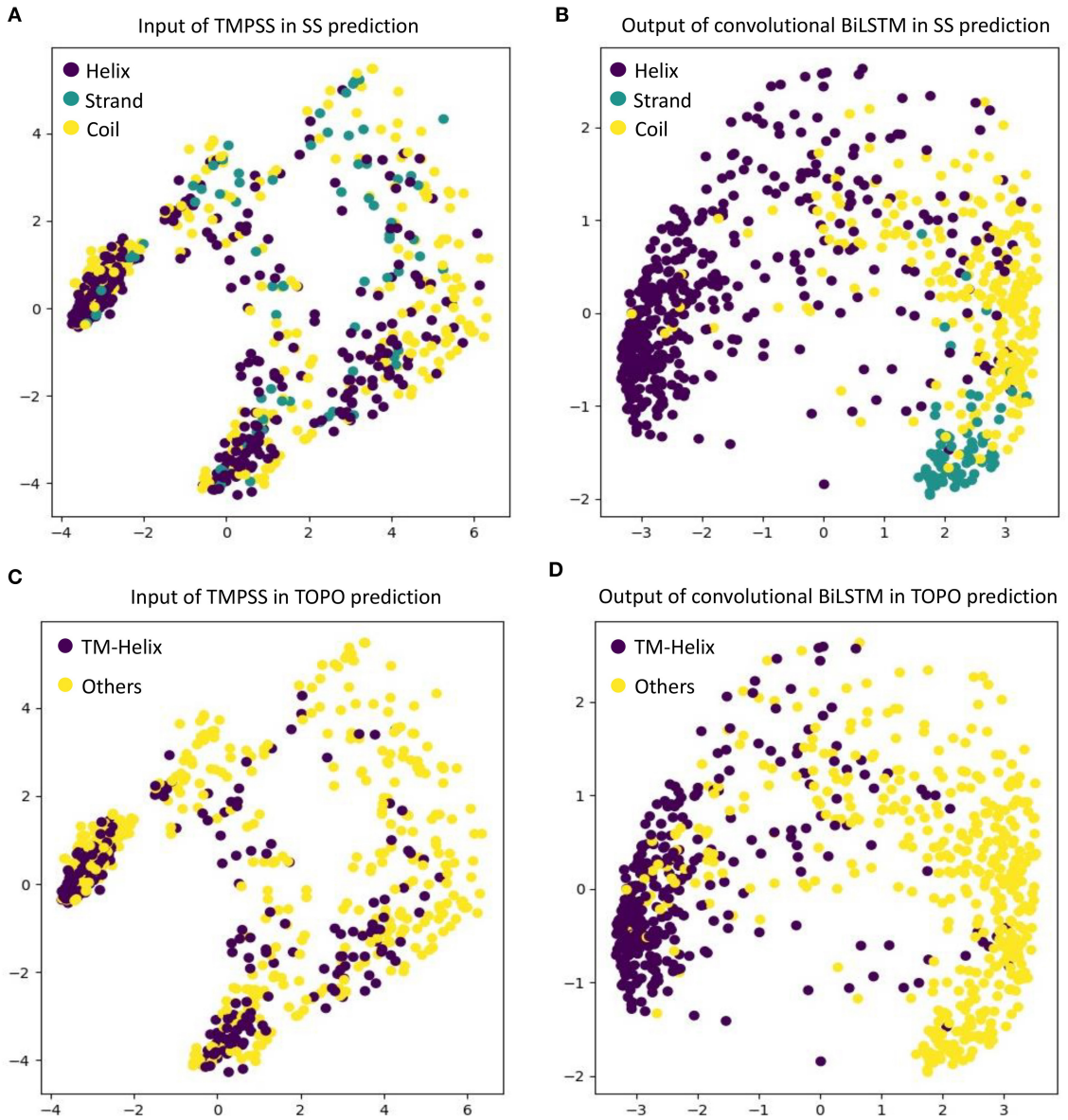

FIGURE 5 | Visualize the input features and the features learned by convolutional BiLSTM, respectively, using PCA. **(A)** Input of TMPSS in SS prediction. **(B)** Output of convolutional BiLSTM in SS prediction. **(C)** Input of TMPSS in TOPO prediction. **(D)** Output of convolutional BiLSTM in TOPO prediction.

for balancing two joint feature learning tasks and regularization terms in the end.

Visualization of the Features Learnt by Convolutional BiLSTM

As an automatic feature extraction process, deep learning can learn high-level abstract features from original inputs (Farias et al., 2016). To further explore the effectiveness of convolutional BiLSTM, Principal Component Analysis (PCA) (Shlens, 2014) was utilized to visualize the input features and each LSTM unit's output in the last bidirectional layer with TEST50. **Figure 5** shows the PCA scatter diagrams before and after TEST50 was fed into our network, respectively.

TABLE 6 | Effect of different combination ways of the attention mechanism on TEST50.

Model	SS Q3	Topo ACC
Attention with multiscale CNNs	0.826	0.893
Attention with BiLSTM	**0.835**	**0.896**
Attention with dropout	0.742	0.866

Bold fonts represent the best experimental results.

As described earlier, the input data had 52 features (i.e., 52 dimensions). PCA reduced the input features' dimensionality to two principal dimensions and visualized it. As we can

see in **Figures 5A,C**, no clear cluster can be found. However, after feeding the data into the convolutional BiLSTM that contains 1,400 dimensions (twice of the unit number in a simple LSTM) at the top layer, the data points showed apparent clustering tendency (see **Figures 5B,D**). This visualization experiment strongly proved the feature extraction efficiency of the convolutional BiLSTM.

It is worth mentioning that since multi-task joint feature learning was performed in our network, the label-based visualization results also revealed the internal relation between secondary structure prediction and topology structure prediction. We found that the points representing "helices" of secondary structure and the ones representing "transmembrane helices" of topology structure have almost completely overlapping distributions under different label-orientated predictions. This experimental phenomenon also directly confirmed the strong correlation between the two prediction tasks and the necessity and effectiveness of multi-task learning.

More results, such as the prediction performance analysis at the residue level, feature analysis, implementation details of multi-task learning, implementation details of attention mechanism, and an ablation study, can be found in the **Supplementary Material**.

TABLE 7 | An ablation study on TEST50.

Model	SS Q3	Topo ACC
Without multiscale CNNs	0.832	0.895
Without BiLSTM layers	0.759	0.743
Without multi-task learning	0.825	0.891
Without attention mechanism	0.828	0.892
TMPSS	**0.835**	**0.896**

Bold fonts represent the best experimental results.

Attention Mechanism

The attention mechanism can stimulate the model extracting features more effectively, speeding up reaching or even improving the best performance of prediction (Choi et al., 2016). To verify the effect of various binding ways of attention mechanism, which acted as a simple full-connect layer in our model, we combined it with different network layers, and the results are shown in **Table 6**. It can be seen that when we attached an attention layer to BiLSTM layers, the prediction results (SS Q3 = 0.835 and Topo ACC = 0.896) were better than doing the same thing to multiscale CNNs or the Dropout layer as expected. One reason could be that the attention mechanism enhanced the process of feature extraction. Another reason could be that BiLSTM layers just learned the most abundant contextual features, making it achieve the best effect when combining attention layer with BiLSTM layers.

Ablation Study

To discover whether a certain component of our proposed method was vital or necessary, we carried out an ablation study by removing some network elements in this section. The experiments performed in our ablation study shared the same

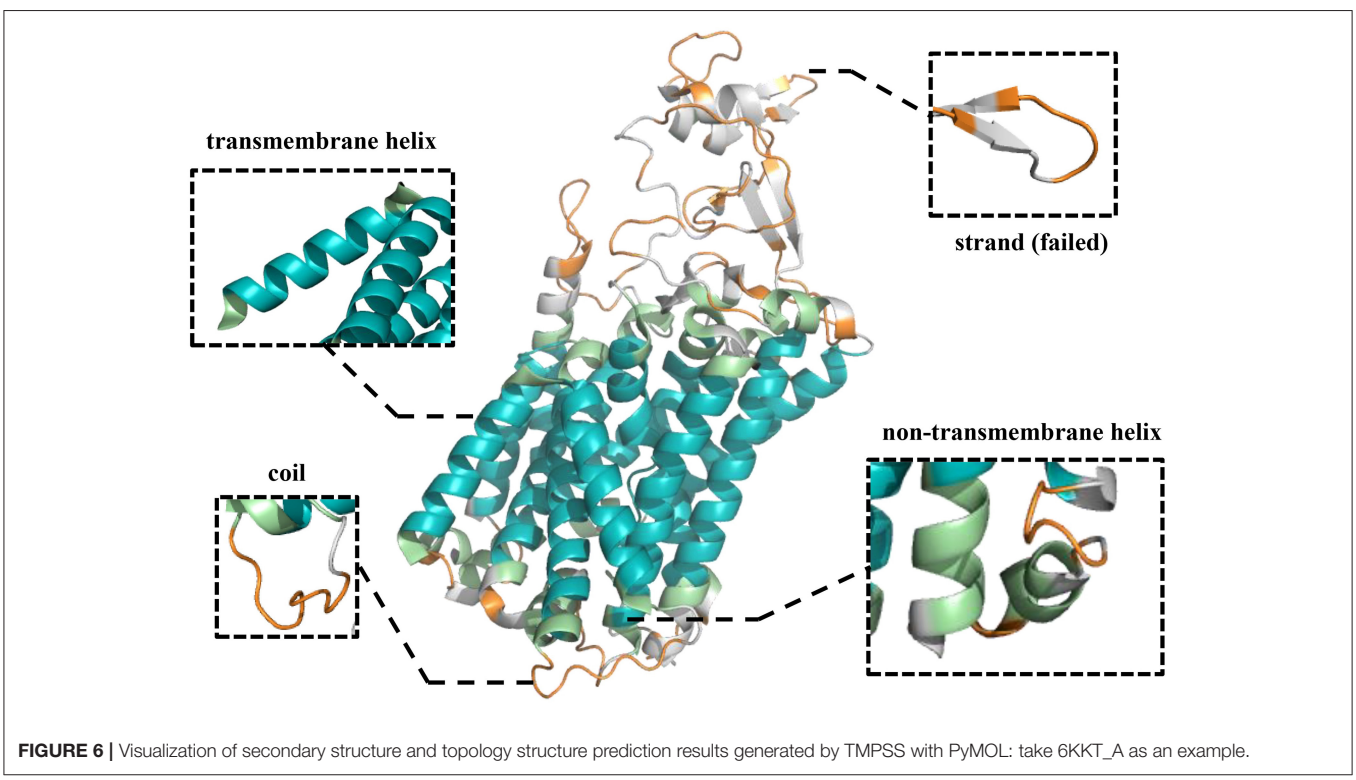

FIGURE 6 | Visualization of secondary structure and topology structure prediction results generated by TMPSS with PyMOL: take 6KKT_A as an example.

features and hyperparameters. From the results on TEST50 presented in **Table 7**, we found that those BiLSTM layers were the most contributing and effective component in our model since the Q3 accuracy of secondary structure prediction dropped to 75.9% when we roughly removed this part from the network. Multiscale CNNs were also essential for good performance as they were particularly good at dealing with local information of protein sequences. Furthermore, multi-task learning and attention mechanism were necessary at the same time because their application made contributions to the robustness of our method with the proof of study results.

Case Study

To further demonstrate the effectiveness of TMPSS on predicting the secondary structure and topology structure of alpha-helical TMPs, we randomly took 6KKT_A as an example of our case study. 6KKT is a kind of transport protein of *Homo sapiens* released on 2019-10-23 that plays vital roles in cell volume regulation, ion transport, and salt reabsorption in the kidney (Liu et al., 2019). The prediction result of TMPSS is visualized in **Figure 6** using PyMOL (DeLano, 2002).

As can be seen, our model correctly identified the helices in the transmembrane region (colored blue) and the non-transmembrane region (colored green). Additionally, most of the coils in the non-transmembrane region (colored orange) were also successfully distinguished.

CONCLUSION

In this study, we proposed a deep learning-based predictor, TMPSS, to predict the secondary structure and topology structure of alpha-helical TMPs from primary sequences. TMPSS's Q3 accuracy of secondary structure prediction in the full chain performed on par with the state-of-the-art methods statistically, and our model had the highest output efficiency with no length restriction of input at the same time. Moreover, our method achieved the best Q3 performance in the transmembrane region and significantly outperformed other topology structure predictors on the independent dataset TEST50.

TMPSS applied a deep learning network with grouped multiscale CNNs and stacked attention-enhanced BiLSTM layers for capturing local and global contexts. Multi-task learning was exploited to improve prediction performance and reduce our method's computational expense by considering the interactions between different protein properties. A series of visualization experiments and comparative tests was taken to verify the validity of the model components mentioned above.

Furthermore, we implemented TMPSS as a publicly available predictor for the research community. The pre-trained model and the datasets we used in this paper could be downloaded at https://github.com/NENUBioCompute/TMP-SS. Finally, we sincerely hope that the predictor and the support materials we released in this study will help the researchers who need them.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Materials**, further inquiries can be directed to the corresponding authors.

AUTHOR CONTRIBUTIONS

ZL, YGo, and YB conceived the idea of this research, collected the data, implemented the predictor, and wrote the manuscript. ZL and YGu tuned the model and tested the predictor. HW and GL supervised the research and reviewed the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by grants from the National Natural Science Foundation of China (nos. 81671328, 81971292), the Program for Professor of Special Appointment (Eastern Scholar) at Shanghai Institutions of Higher Learning (no. 1610000043), the Innovation Research Plan supported by Shanghai Municipal Education Commission (ZXWF082101), the Jilin Scientific and Technological Development Program (no. 20180414006GH), and the Fundamental Research Funds for the Central Universities (nos. 2412019FZ052, 2412019FZ048).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fbioe.2020.629937/full#supplementary-material

REFERENCES

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., et al. (2016). "Tensorflow: a system for large-scale machine learning," in *12th {USENIX} Symposium on Operating Systems Design and Implementation (OSDI 16)* (Savannah, GA), 265–283.

Bello, I., Zoph, B., Vasudevan, V., and Le, Q. V. (2017). "Neural optimizer search with reinforcement learning," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70: JMLR. org* (Sydney, NSW), 459–468.

Bernsel, A., Viklund, H., Falk, J., Lindahl, E., von Heijne, G., and Elofsson, A. (2008). Prediction of membrane-protein topology from first principles. *Proc. Natl. Acad. Sci. U.S.A.* 105, 7177–7181. doi: 10.1073/pnas.0711151105

Buchan, D. W., and Jones, D. T. (2019). The PSIPRED protein analysis workbench: 20 years on. *Nucleic Acids Res.* 47, W402–W407. doi: 10.1093/nar/gkz297

Burley, S. K., Berman, H. M., Kleywegt, G. J., Markley, J. L., Nakamura, H., and Velankar, S. (2017). Protein Data Bank (PDB): the single global macromolecular structure archive. *Methods Mol. Biol.* 1607, 627–641. doi: 10.1007/978-1-4939-7000-1_26

Butterfield, D. A., and Boyd-Kimball, D. (2004). Proteomics analysis in Alzheimer's disease: new insights into mechanisms of neurodegeneration. *Int. Rev. Neurobiol.* 61, 159–188. doi: 10.1016/S0074-7742(04)61007-5

Chen, C. P., Kernytsky, A., and Rost, B. (2002). Transmembrane helix predictions revisited. *Protein Sci.* 11, 2774–2791. doi: 10.1110/ps.0214502

Choi, E., Bahadori, M. T., Sun, J., Kulas, J., Schuetz, A., and Stewart, W. (2016). Retain: an interpretable predictive model for healthcare using reverse time attention mechanism. *Adv. Neural Inf. Process. Syst.* 3504–3512.

DeLano, W. L. (2002). Pymol: an open-source molecular graphics tool. *CCP4 Newslett. Protein Crystallogr.* 40, 82–92.

Ding, H., and Li, D. (2015). Identification of mitochondrial proteins of malaria parasite using analysis of variance. *Amino Acids* 47, 329–333. doi: 10.1007/s00726-014-1862-4

Drozdetskiy, A., Cole, C., Procter, J., and Barton, G. J. (2015). JPred4: a protein secondary structure prediction server. *Nucleic Acids Res.* 43, W389–W394. doi: 10.1093/nar/gkv332

Fang, C., Li, Z., Xu, D., and Shang, Y. (2020). MUFold-SSW: a new web server for predicting protein secondary structures, torsion angles, and turns. *Bioinformatics* 36, 1293–1295. doi: 10.1093/bioinformatics/btz712

Fang, C., Shang, Y., and Xu, D. (2017). MUFold-SS: Protein Secondary Structure Prediction Using Deep Inception-Inside-Inception Networks. *arXiv preprint arXiv:1709.06165*.

Fang, C., Shang, Y., and Xu, D. (2018). Improving protein gamma-turn prediction using inception capsule networks. *Sci. Rep.* 8, 1–12. doi: 10.1038/s41598-018-34114-2

Farias, G., Dormido-Canto, S., Vega, J., Ratt,á, G., Vargas, H., Hermosilla, G., et al. (2016). Automatic feature extraction in large fusion databases by using deep learning approach. *Fusion Eng. Des.* 112, 979–983. doi: 10.1016/j.fusengdes.2016.06.016

Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. (2012). CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* 28, 3150–3152. doi: 10.1093/bioinformatics/bts565

Gal, Y., Hron, J., and Kendall, A. (2017). Concrete dropout. *Adv. Neural Inf. Process. Syst.* 3581–3590.

Goddard, A. D., Dijkman, P. M., Adamson, R. J., dos Reis, R. I., and Watts, A. (2015). Reconstitution of membrane proteins: a GPCR as an example. *Methods Enzymol.* 556, 405–424. doi: 10.1016/bs.mie.2015.01.004

Gulli, A., and Pal, S. (2017). *Deep Learning With Keras.* Birmingham: Packt Publishing Ltd.

Hanson, J., Paliwal, K., Litfin, T., Yang, Y., and Zhou, Y. (2019). Improving prediction of protein secondary structure, backbone angles, solvent accessibility, and contact numbers by using predicted contact maps and an ensemble of recurrent and residual convolutional neural networks. *Bioinformatics* 35, 2403–2410. doi: 10.1093/bioinformatics/bty1006

Heffernan, R., Yang, Y., Paliwal, K., and Zhou, Y. (2017). Capturing non-local interactions by long short-term memory bidirectional recurrent neural networks for improving prediction of protein secondary structure, backbone angles, contact numbers, and solvent accessibility. *Bioinformatics* 33, 2842–2849. doi: 10.1093/bioinformatics/btx218

Ioffe, S., and Szegedy, C. (2015). Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *arXiv preprint arXiv:1502.03167.*

Jones, D. T. (2007). Improving the accuracy of transmembrane protein topology prediction using evolutionary information. *Bioinformatics* 23, 538–544. doi: 10.1093/bioinformatics/btl677

Kabsch, W., and Sander, C. (1983). DSSP: definition of secondary structure of proteins given a set of 3D coordinates. *Biopolymers* 22, 2577–2637. doi: 10.1002/bip.360221211

Kozma, D., Simon, I., and Tusnady, G. E. (2012). PDBTM: protein data bank of transmembrane proteins after 8 years. *Nucleic Acids Res.* 41, D524–D529. doi: 10.1093/nar/gks1169

Liu, S., Chang, S., Han, B., Xu, L., Zhang, M., Zhao, C., et al. (2019). Cryo-EM structures of the human cation-chloride cotransporter KCC1. *Science* 366, 505–508. doi: 10.1126/science.aay3129

Lu, C., Liu, Z., Kan, B., Gong, Y., Ma, Z., and Wang, H. (2019). TMP-SSurface: a deep learning-based predictor for surface accessibility of transmembrane protein residues. *Crystals* 9:640. doi: 10.3390/cryst9120640

Lv, Z. B., Ao, C. Y., and Zou, Q. (2019). Protein function prediction: from traditional classifier to deep learning. *Proteomics* 19:e1900119. doi: 10.1002/pmic.201900119

Magnan, C. N., and Baldi, P. (2014). SSpro/ACCpro 5: almost perfect prediction of protein secondary structure and relative solvent accessibility using profiles, machine learning, and structural similarity. *Bioinformatics* 30, 2592–2597. doi: 10.1093/bioinformatics/btu352

Nugent, T., Ward, S., and Jones, D. T. (2011). The MEMPACK alpha-helical transmembrane protein structure prediction server. *Bioinformatics* 27, 1438–1439. doi: 10.1093/bioinformatics/btr096

Patil, K., and Chouhan, U. (2019). Relevance of machine learning techniques and various protein features in protein fold classification: a review. *Curr. Bioinform.* 14, 688–697. doi: 10.2174/1574893614666190204154038

Pauling, L., Corey, R. B., and Branson, H. R. (1951). The structure of proteins: two hydrogen-bonded helical configurations of the polypeptide chain. *Proc. Natl. Acad. Sci. U.S.A.* 37, 205–211. doi: 10.1073/pnas.37.4.205

Remmert, M., Biegert, A., Hauser, A., and Söding, J. (2012). HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat. Methods* 9, 173–175. doi: 10.1038/nmeth.1818

Reynolds, S. M., Käll, L., Riffle, M. E., Bilmes, J. A., and Noble, W. S. (2008). Transmembrane topology and signal peptide prediction using dynamic bayesian networks. *PLoS Comput. Biol.* 4:e1000213. doi: 10.1371/journal.pcbi.1000213

Roy, A. (2015). Membrane preparation and solubilization. *Methods Enzymol.* 557, 45–56. doi: 10.1016/bs.mie.2014.11.044

Shlens, J. (2014). A tutorial on principal component analysis. *arXiv preprint arXiv:1404.1100*.

Steinegger, M., Meier, M., Mirdita, M., Vöhringer, H., Haunsberger, S. J., and Söding, J. (2019). HH-suite3 for fast remote homology detection and deep protein annotation. *BMC Bioinform.* 20:473. doi: 10.1186/s12859-019-3019-7

Stillwell, W. (2016). *An Introduction to Biological Membranes: Composition, Structure, and Function*. Amsterdam: Elsevier.

Tan, J.-X., Li, S.-H., Zhang, Z.-M., Chen, C.-X., Chen, W., Tang, H., et al. (2019). Identification of hormone binding proteins based on machine learning methods. *Math. Biosci. Eng.* 16, 2466–2480. doi: 10.3934/mbe.2019123

Torrisi, M., Kaleel, M., and Pollastri, G. (2019). Deeper profiles and cascaded recurrent and convolutional neural networks for state-of-the-art protein secondary structure prediction. *Sci. Rep.* 9, 1–12. doi: 10.1038/s41598-019-48786-x

Tsirigos, K. D., Peters, C., Shu, N., Käll, L., and Elofsson, A. (2015). The TOPCONS web server for consensus prediction of membrane protein topology and signal peptides. *Nucleic Acids Res.* 43, W401–W407. doi: 10.1093/nar/gkv485

Tusnady, G. E., and Simon, I. (2001). The HMMTOP transmembrane topology prediction server. *Bioinformatics* 17, 849–850. doi: 10.1093/bioinformatics/17.9.849

Viklund, H., Bernsel, A., Skwark, M., and Elofsson, A. (2008). SPOCTOPUS: a combined predictor of signal peptides and membrane protein topology. *Bioinformatics* 24, 2928–2929. doi: 10.1093/bioinformatics/btn550

Viklund, H., and Elofsson, A. (2008). OCTOPUS: improving topology prediction by two-track ANN-based preference scores and an extended topological grammar. *Bioinformatics* 24, 1662–1668. doi: 10.1093/bioinformatics/btn221

Walsh, I., Pollastri, G., and Tosatto, S. C. (2016). Correct machine learning on protein sequences: a peer-reviewing perspective. *Brief. Bioinform.* 17, 831–840. doi: 10.1093/bib/bbv082

Wang, H., Yang, Y., Yu, J., Wang, X., Zhao, D., Xu, D., et al. (2019). "DMCTOP: topology prediction of alpha-helical transmembrane protein based on deep multi-scale convolutional neural network," in *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* (San Diego, CA: IEEE), 36–43.

Wang, S., Li, W., Liu, S., and Xu, J. (2016a). RaptorX-Property: a web server for protein structure property prediction. *Nucleic Acids Res.* 44, W430–W435. doi: 10.1093/nar/gkw306

Wang, S., Peng, J., Ma, J., and Xu, J. (2016b). Protein secondary structure prediction using deep convolutional neural fields. *Sci. Rep.* 6, 1–11. doi: 10.1038/srep18962

Wei, L., Su, R., Wang, B., Li, X., Zou, Q., and Gao, X. (2019). Integration of deep feature representations and handcrafted features to improve the prediction of N 6-methyladenosine sites. *Neurocomputing* 324, 3–9. doi: 10.1016/j.neucom.2018.04.082

Xiao, F., and Shen, H.-B. (2015). Prediction enhancement of residue real-value relative accessible surface area in transmembrane helical proteins by solving the output preference problem of machine learning-based predictors. *J. Chem. Inf. Model.* 55, 2464–2474. doi: 10.1021/acs.jcim.5b00246

Yang, W., Zhu, X.-J., Huang, J., Ding, H., and Lin, H. (2019). A brief survey of machine learning methods in protein sub-Golgi localization. *Curr. Bioinform.* 14, 234–240. doi: 10.2174/1574893613666181113131415

Yaseen, A., and Li, Y. (2014). Context-based features enhance protein secondary structure prediction accuracy. *J. Chem. Inf. Model.* 54, 992–1002. doi: 10.1021/ci400647u

Zeng, X., Zhong, Y., Lin, W., and Zou, Q. (2020). Predicting disease-associated circular RNAs using deep forests combined with positive-unlabeled learning methods. *Brief. Bioinform.* 21, 1425–1436. doi: 10.1093/bib/bbz080

Zhang, J., and Liu, B. (2019). A review on the recent developments of sequence-based protein feature extraction methods. *Curr. Bioinform.* 14, 190–199. doi: 10.2174/1574893614666181212102749

Zhang, Y., and Yeung, D.-Y. (2012). A convex formulation for learning task relationships in multi-task learning. *arXiv preprint arXiv:1203.3536*.

Zhou, P., Shi, W., Tian, J., Qi, Z., Li, B., Hao, H., et al. (2016). "Attention-based bidirectional long short-term memory networks for relation classification," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (volume 2: Short papers)* (Berlin), 207–212.

Zhu, X.-J., Feng, C.-Q., Lai, H.-Y., Chen, W., and Hao, L. (2019). Predicting protein structural classes for low-similarity sequences by evaluating different features. *Knowl. Based Syst.* 163, 787–793. doi: 10.1016/j.knosys.2018.10.007

Zou, Q., Lin, G., Jiang, X., Liu, X., and Zeng, X. (2020). Sequence clustering in bioinformatics: an empirical study. *Brief. Bioinform.* 21, 1–10.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The reviewer P-FD declared a shared affiliation, with no collaboration, with the author YGo to the handling editor at the time of the review.