Transmembrane Topology Identification by Fusing Evolutionary and Co-evolutionary Information with Cascaded Bidirectional Transformers

Zhen Li*

lizhen@cuhk.edu.cn

The Chinese University of Hong
Kong, Shenzhen, Shenzhen
Research Institute of Big Data,
Toyota Technological Institute at
Chicago
Shenzhen, Guangdong

Chongming Ni*

nickelcming@icloud.com

Tsinghua University, Toyota
Technological Institute at Chicago
Beijing, Beijing

Jinbo Xu†

jinbo.xu@gmail.com

Toyota Technological Institute at
Chicago
Chicago, Illinois

Xin Gao†

xin.gao@kaust.edu.sa

King Abdullah University of
Science and Technology
Thuwal, Kingdom of Saudi Arabia

Shuguang Cui†

shuguangcui@cuhk.edu.cn

The Chinese University of Hong
Kong, Shenzhen, Shenzhen
Research Institute of Big Data
Shenzhen, Guangdong

Sheng Wang†

wangsheng@ttic.edu

King Abdullah University of
Science and Technology, Toyota
Technological Institute at Chicago
Thuwal, Kingdom of Saudi Arabia

ABSTRACT

The transmembrane topology is the key to understand the 3D structures of multi-pass Transmembrane Proteins (mTPs). However, accurate prediction of the 1D topology label for each residue of an mTP from evolutionary information alone is very challenging, if not infeasible. In this work, we propose a novel approach to identify the transmembrane topology under an object detection framework that takes as the input the predicted 2D distance matrix from the co-evolutionary information, followed by several bidirectional Transformer blocks that effectively fuse both 2D and 1D features for accurate label prediction. Specifically, we employ the Faster-RCNN module to simultaneously predict the rectangular bounds that cover the interacted transmembrane regions, as well as the confidence scores to discriminate them from the non-transmembrane regions. To integrate the 2D pairwise features and the 1D sequential features, we establish several bidirectional Transformer blocks consisting of self-attention units for capturing long-range dependencies in the transmembrane topology. Tested on the 330 non-redundant mTPs and the newly released 45 mTPs, in terms of the

Segment Overlap (SOV) score, our approach achieves 0.927 and 0.843, which are about **4.5%** and **6.6%** better than the cutting-edge consensus methods, respectively.

KEYWORDS

protein prediction, membrane protein prediction, multi-pass transmembrane topology identification, computer vision, co-evolutionary information analysis, evolutionary and co-evolutionary information fusion, bidirectional transformers

ACM Reference Format:

Zhen Li, Chongming Ni, Jinbo Xu, Xin Gao, Shuguang Cui, and Sheng Wang. 2019. Transmembrane Topology Identification by Fusing Evolutionary and Co-evolutionary Information with Cascaded Bidirectional Transformers. In *10th ACM Int'l Conference on Bioinformatics, Computational Biology and Health Informatics (ACM-BCB '19)*, September 7–10, 2019, Niagara Falls, NY, USA. ACM, New York, NY, USA, 8 pages. <https://doi.org/10.1145/3307339.3342140>

1 INTRODUCTION

Multi-pass Transmembrane Proteins (mTPs) have been targeted by around 50% of therapeutic drugs and are encoded by about 20% genes of the human genome [23]. They are the key players in the cell-environment interactions by serving as environment sensing receptors, transporters, and channels [24]. Nonetheless, the inherent structural and physiochemical properties of mTPs lead to challenges in solving their structures by structural biology experiments, such as X-ray crystallography, nuclear magnetic resonance, and single particle cryo-EM. Currently, only less than 400 non-redundant human mTPs with solved structures in Protein Data Bank (PDB) are available, which only constitute <15% of all human mTPs. Therefore, it is important to develop efficient computational methods to predict mTP structures from the

*Both authors contributed equally to this research.

†All authors shared the corresponding authors.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

ACM-BCB '19, September 7–10, 2019, Niagara Falls, NY, USA

© 2019 Association for Computing Machinery.

ACM ISBN 978-1-4503-6666-3/19/09...\$15.00

<https://doi.org/10.1145/3307339.3342140>

Figure 1: Illustration of the transmembrane topology of a multi-pass transmembrane protein. Left: the 1D amino acid sequence of an alpha-helical membrane protein folds into a 3D structure embedded in the lipid bilayer membrane, in which those residues embedded are denoted as the transmembrane topology. Right: the 2D distance map of the membrane protein contains all square distances between each pair of residues. The patterns of pairwise transmembrane helical-helical interactions are shown in different colors with respect to the minimum distance between the two helices, indicating different interaction strength.

sequence information, where the first and critical step is to predict the transmembrane topology.

The transmembrane topology of an mTP is a 1D 0/1 string to indicate the location of each residue that resides in (label 1) or out of (label 0) the membrane (left part in Fig. 1). The ratio between label 0 and label 1 is around 1.6 for multi-pass transmembrane protein. As almost all the transmembrane regions in eukaryotic mTPs are alpha-helical [2], in this work we will focus only on the prediction of transmembrane topology of the alpha-helical transmembrane proteins. Till now, a variety of approaches have been proposed to predict the 1D transmembrane topology from the input sequence of an mTP [20]. These approaches can be roughly categorized into three groups: (a) single-sequence-based methods that only rely on the input sequence information produce the prediction whose Q2 accuracy (i.e., the accuracy of two-class classification of membrane vs. non-membrane) is around 75%; (b) evolutionary-based methods that also consider the information embedded in the homologous Multiple Sequence Alignment (MSA) through evolutionary analysis increase the Q2 accuracy to around 78%; and (c) consensus methods that combine the outputs from different predictors can reach the cutting-edge Q2 accuracy around 80%. Yet, all the previous prediction methods only make use of the 1D sequential

features (such as the evolutionary information embedded in MSA), without taking the long-range pairwise features into account.

It is obvious that pairwise transmembrane helical-helical interactions are critical for the formation of transmembrane topology (right part in Fig. 1). The underlying pairwise features are the 2D distance map that contains all square distances between each pair of the residues. Recently, it has been shown that these pairwise features, as described by the contact map or distance map, could be predicted from the co-evolutionary analysis (e.g., Direct Coupling Analysis (DCA) [3, 12, 13]) and deep learning methods upon the MSA [26]. However, there are still two challenges remaining: (i) how to further exploit the predicted distance information to describe the pairwise transmembrane helical-helical interactions, and (ii) how to effectively merge the 2D pairwise features and the 1D sequential features into the transmembrane topology prediction.

In this work, our proposed model first formulates a 2D object detection problem that takes as the input the predicted distance matrix (which could be regarded as an “image”) [28], and then fuses the pairwise and sequential features for binary labeling of the transmembrane topology under a self-attention model [21] (shown in Fig. 2). As indicated in the right part of Fig. 1, the pairwise transmembrane helical-helical interactions become different bounded objects for detection based on their interaction strength measured by the minimum distance. Then, we apply the Faster-RCNN [16] module to simultaneously predict the bounds that cover the pairwise transmembrane regions, as well as the scores to discriminate them from the background non-transmembrane regions. To combine the 2D pairwise features (such as the 2D object detection confidence scores) and the 1D sequential features (such as the classical evolutionary information), we employ the bidirectional Transformer framework [4] to effectively merge them together to predict the binary 1D transmembrane topology label. Both the Faster-RCNN module and the Transformer framework are the current state-of-the-art in object detection and natural language processing. The critical function of the Faster-RCNN module is to generate a set of rectangular “object” efficiently and accurately, whereas the key contribution from the Transformer is the direct description of the relationships between all residues in a sequence under a self-attention mechanism.

In summary, the main contributions of this paper are as follows.

- We propose a novel deep learning framework that simultaneously fuses 1D and 2D features for sequence labeling, and apply this framework for transmembrane topology prediction. This deep network consists of (i) 1D feature encoding layers to exploit local and global contextual sequential features; (ii) 2D feature encoding layers for the explicit extraction of pairwise features; and (iii) fusion layers that integrate the 1D and 2D features for directly describing the relationships between all sequential labels.

- To the best of our knowledge, this work is the first to exploit the predicted distance matrix under an object detection framework for identifying the transmembrane topology. Experimental results on the newly released 45 mTPs dataset demonstrate that the proposed approach significantly outperforms existing consensus methods that only utilize 1D sequential features. We further show that the 2D pairwise features and the bidirectional Transformer are the main contributors to the success of the proposed method.

The rest of the paper is organized as follows. In Section 2, we introduce our proposed model in details. In Section 3, we show experimental results on two datasets. We conclude this work with remarks in Section 4.

2 MODEL ARCHITECTURE

As shown in Fig. 2, the architecture of our model for predicting the transmembrane topology label has four major components: (i) the 1D feature encoder, (ii) the 2D feature encoder with distance prediction and object detection, (iii) the 1D and 2D feature fusion, and (iv) the 1D decoder. The input to our model carries two types of information about the mTP sequence: the 1D evolutionary and the 2D co-evolutionary information, both are derived from the MSA of the input amino acid sequence. The 1D feature encoder, consisting of two stacked Bidirectional Transformer Encoder (BTE) blocks, is responsible for extracting the informative representations embedded in the sequential evolutionary information. In order to extensively exploit the pairwise co-evolutionary information, we first feed it into the distance prediction module to obtain the predicted distance distribution. Then the predicted distance map is further pushed forward to the object detection module for generating the 2D scores that describe the interaction strength between the transmembrane helices. Inspired by the success of the fusion model [9], the two parts are concatenated and fed to a 2D Transformer module consisting of cascaded horizontal and vertical BTEs for extracting the informative representations of the pairwise features. Finally, we fuse the representations from the 1D and 2D feature encoders together by two stacked BTE blocks, and send the output to the Bidirectional Transformer Decoder (BTD) for binary classification of the transmembrane topology label.

2.1 1D Feature Encoder

For better understanding, the transmembrane topology prediction can be formulated as follows. Given an amino acid sequence $X = x_1, x_2, \dots, x_L$ with length L , we predict the transmembrane topology label of each amino acid, $S = s_1, s_2, \dots, s_L$, where s_i is a 2-state transmembrane topology label. Different from the previous methods for transmembrane topology prediction that only use the evolutionary information [14], the input features of x_i in this paper consist of two parts derived from MSA, i.e., the evolutionary information $f \in \mathbb{R}^{L \times n}$ as an n -dimensional feature vector and the co-evolutionary information $g \in \mathbb{R}^{L \times L \times m}$ as a feature map with m channels. For the evolutionary information input,

Recurrent Neural Networks (RNNs) have been adopted as a typical structure for sequential representative learning [10]. However, previous sequential processing methods make RNNs difficult to take full advantage of modern parallel computing devices. In this paper, we adopt the encoder part of recently proposed ubiquitous Transformer architecture [21] with the self-attention mechanism for the 1D feature encoder. Inspired by the great success of appealing BERT [4] with the bidirectional Transformer, we adopt a multi-layer BTE, i.e., left-context-right and right-context-left. Considering the relatively small mTPs dataset, we empirically set the number of layers L as 2, multi-attention head $M = 2$, the hidden size H as 16 and the feed-forward size as $2H = 32$ accordingly.

2.2 2D Feature Encoder

As illustrated in Fig. 2, the second component of our proposed architecture is the 2D feature encoder, which is the novelty of our work from the biological aspect that extensively exploits the co-evolutionary information. Specifically, the 2D feature encoder consists of three modules: (a) the distance prediction, (b) the object detection and (c) the 2D Transformer encoder module (consisting of cascaded horizontal and vertical BTEs). Given the evolutionary information f and the co-evolutionary information g , we first predict the distance map that represents the distance distribution of any two residues [24, 26, 28]. Then the predicted distance map flows into the object detection module, i.e., Faster-RCNN [16], for detecting different interaction patterns between the transmembrane helices. Finally, we concatenate the predicted distance map and the object detection results to the 2D Transformer module for further extraction of pairwise features. The details of the three modules are described as follows.

2.2.1 Distance Prediction. We adopt a similar Deep Learning (DL) framework as RaptorX-contact [26] to predict the Euclidean distance distribution of two atoms (of different residues) in a protein. In particular, the inter-atom distance is discretized into 25 bins: $<4\text{\AA}$, $4 - 4.5\text{\AA}$, $4.5 - 5\text{\AA}$, $5 - 5.5\text{\AA}$, ..., $15 - 15.5\text{\AA}$, $15.5 - 16\text{\AA}$, and $>16\text{\AA}$. The DL model is formed by one 1D deep Residual Neural Network (ResNet) [6], one 2D deep ResNet and one logistic regression. The 1D ResNet is used to capture the contextual evolutionary information of one residue, while the 2D ResNet is used to capture the pairwise contextual co-evolutionary information of a residue pair. We also use 6 layers of 1D ResNet with the filter size 15 and around 60 layers of 2D ResNet with the filter size 3×3 accordingly. As the solved structures of mTPs in PDB are not sufficient to train a deep learning model, we transfer the knowledge learned from the non-membrane proteins through deep transfer learning according to the scheme proposed in [26].

2.2.2 Object Detection. We adopt Faster-RCNN to perform object detection from the input of the predicted distance distribution (specifically a $L * L * 25$ tensor), which could be regarded as an “image”. We define four types of “object”

Figure 2: The overview of our model for predicting the transmembrane topology label. The inputs of our model consist of 1D evolutionary and 2D co-evolutionary information derived from the multiple sequence alignment (MSA) of the input amino acid sequence. The evolutionary information is passed to the 1D feature encoder consisting of two layers of Bidirectional Transformer Encoders (BTEs, shown as yellow diamond) to extract informative representations. The evolutionary and co-evolutionary information are fed into the 2D feature encoder from the distance prediction module to the object detection module, and the concatenated distance and detection predictions then flow into cascaded horizontal and vertical BTEs to extract informative representations embedded in these pairwise features. The extracted informative representations from the 1D and 2D feature encoders are further fused together by two stacked BTEs. At last, the decoder consisting of two stacked Bidirectional Transformer Decoders (BTDs, shown as golden decagon) takes as inputs the Keys and Values matrix from the final representations generated by 1D-2D feature fusion module.

for detection according to the patterns of pairwise transmembrane helical-helical interactions (right part in Fig. 1). The first object is a self interaction whereas the other three objects are inter-helix interactions with different interaction strength. Specifically, the strong, medium, and weak interactions are defined as the minimum distances between the two transmembrane helices that is $<8\text{\AA}$, $8\text{\AA}$ to $16\text{\AA}$, and $>16\text{\AA}$, respectively.

According to the four types defined above, the objects are mostly with small scales. We use k-means clustering to analyze the scale of the defined objects. As a result, 6 clusters are obtained as (13×13) , (15×15) , (17×17) , (20×20) , (23×23) , (27×27) . As Faster-RCNN is the most accurate single model for small object detection [7], we choose Faster-RCNN with ResNet50 as the backbone. In the Faster-RCNN setting, detection happens in two stages. In the first stage, called the Region Proposal Network (RPN), images are processed by a

feature extractor (i.e., ResNet50), and features at some selected intermediate level are used to predict the class-agnostic box proposals. According to previous 6 clusters, the anchor size is set to be 15, 20, 25, 30, 35 and the anchor aspect ratio is set to be 0.8, 1, 1.2. In the second stage, these box proposals are used to crop features from the same intermediate feature map, which are subsequently fed to the remainder of the feature extractor (e.g, fc6 followed by fc7) in order to predict a class and the class-specific box regression refinement for each proposal. Considering that Feature Pyramid Networks (FPN) [11] build an architecture that combines low-resolution, semantically weak features via a top-down pathway and lateral connections, we add FPN to our Faster-RCNN structure for feature enhancement, since Faster-RCNN is most commonly performed on a single-scale feature map. Furthermore, we apply Group Normalization (GN) [27] instead of Batch Normalization (BN) to further improve the object detection performance.

2.2.3 2D Transformer Encoder. Since object detection can generate bounding boxes with class labels and confidence scores simultaneously, the output of object detection can be formed as a feature map of $L * L * 2$ through padding operations, i.e., all pixels within a detected bounding box are filled with the same class label and confidence score while other background without a detected bounding box are padded with two zeros. As illustrated in Fig. 1, this $L * L * 2$ feature map is a spare tensor due to the small number of objects, which inspires us to design a skip connection to combine the predicted distance map and the object detection feature map. On top of the concatenated feature map of $L * L * 27$, a 2D Transformer, that consists of cascaded horizontal and vertical BTEs, are exploited for the 2D feature encoder. Specifically, each row flows into a horizontal BTE to obtain a feature learning separately with parameters shared among all rows. Then only the last column of the previous BTE outputs is taken as the input to the vertical BTE. At last, the output of vertical BTE is regarded as the encoded feature of the 2D feature block, whose shape is $L * 32$. All the BTE settings is the same as those introduced in Sec. 2.1. Note that the predicted distance map is a symmetric tensor; 2D Transformer encoder with cascaded horizontal and vertical BTEs is the same as the cascaded vertical and horizontal BTEs.

2.3 Fusion of 1D and 2D Features

As pairwise transmembrane helical-helical interactions are critical for the formation of transmembrane topology and the co-evolutionary information can provide auxiliary information, a fusion module is required to adaptively integrate encoded 1D features and 2D features. In this paper, we use two stacked BTEs for the fusion of concatenated 1D features (i.e., $L * 32$) and 2D features (i.e., $L * 32$). All the BTE setting is the same as those introduced in Sec. 2.1.

2.4 Topology Prediction Decoder

Similar to the Transformer decoder in [4, 21] used for translation or text generation, two stacked Bidirectional Transformer Decoder (BTD) layers are leveraged as the decoder module. Different from the BTE according to [21], a third sub-layer performing multi-head attention over the fused 1D and 2D features is inserted in the decoder. Specifically, the fused encoder of the Keys and Values matrices from 1D-2D fusion module is exploited for encoder-decode attention, with details described in [21].

3 EXPERIMENTAL RESULTS

3.1 Datasets and Features

The datasets used to evaluate the performance of our proposed method against the other methods are 330 mTPs released before 2016, and 45 mTPs released after 2017. Any two proteins in both datasets share less than 25% sequence identity. There are no homologous sequences between the two datasets. Specifically, to remove redundancy between

the two datasets, we use a strict rule that there is no protein in the 45 test set that shares $> 25\%$ sequence identity or BLAST E-value < 0.001 with any proteins in the 330 training/validation dataset. All the ground-truth transmembrane topology labels of the two datasets are derived from the PDBTM database [8]. The 330 dataset is a relatively large set containing 330 non-redundant mTPs, which include 270 proteins for training, 30 proteins for validation and 30 proteins for testing. Note that we also use this dataset to train the Faster-RCNN framework for object detection. The 45 dataset contains 45 non-redundant mTPs, which are used for testing the performance of our proposed method against with other methods. It should be noted that all the other methods are released before 2017, therefore the comparison on the 45 dataset is fair to all methods.

The features used in our method are categorized into two parts: (i) the evolutionary information and (ii) the co-evolutionary information, which are derived from the MSA of the input amino acid sequence (using HHblits [15] with 3 iterations and E-value of 0.001 to search the UniProt20 database released in February 2016). For the evolutionary information, we use 40 channels of data per residue, among which 20 channels are position specific frequencies, and the other 20 channels are position specific scores. For the co-evolutionary information, we use 3 channels of data for each pair of residues, which include the Mutual Information (MI), pairwise contact potential and Direct Information (DI) generated by CCMpred [19]. More details about the features could be found in [26]. And our data can be obtained through ¹.

3.2 Comparing Methods and Evaluation Metric

We compare our method with the following existing methods: Philius [17], SCAMPI [1], OCTOPUS [22] and TOPCONS [20] for the 2-state transmembrane topology prediction. Philius is a single-sequence-based method that only relies on the input sequence information; SCAMPI and OCTOPUS are evolutionary-based methods relying on the MSA; TOPCONS is a consensus approach that combines the outputs from different predictors. All the methods are executed with their parameters set according to their respective papers. For TOPCONS, we submit the 330+45 sequences in the two datasets to its server to obtain the prediction results. We only choose TOPCONS as the comparison algorithm as CCTOP [5] is also a consensus algorithm which achieves similar performance as TOPCONS.

We measure the prediction results in terms of the following evaluation criteria: the protein-level accuracy, the segment-level accuracy, and the residue-level accuracy. The protein-level accuracy refers to the definition that a correct prediction of the whole protein should have the correct number of transmembrane segments at approximately correct locations (overlap of at least five residues). The segment-level accuracy is defined as the approximately correct prediction

¹<https://bit.ly/2YD30Fz>

of the transmembrane segment with overlap of at least five residues, which is denoted as the fragment recall.

The residue-level accuracy is defined at each residue, which consists of the Q2 accuracy, SOV score, recall, precision and Matthews correlation coefficient (Mcc). The Q2 accuracy is defined as the percentage of residues for which the predicted transmembrane topology labels are correct. The SOV score measures how well the ground-truth and the predicted transmembrane regions match, especially at the middle region instead of the terminal regions [25]. In order to calculate recall, precision and Mcc, we define the True Positives (TP) and True Negatives (TN) as the numbers of correctly predicted transmembrane and non-transmembrane residues, respectively, where False Positives (FP) and False Negatives (FN) are the numbers of misclassified transmembrane and non-transmembrane residues, respectively. Then recall and precision is defined as $TP/(TP + FN)$ and $TP/(TP + FP)$, respectively. Mcc is defined as:

$$\text{Mcc} = \frac{(TP \times TN - FP \times FN)}{\sqrt{(TP + FP)(TN + FP)(TP + FN)(TN + FN)}}.$$

3.3 Implementation Details

In our experiments, the bidirectional Transformers are all set to have $H = 16$ hidden states, and the output has $2H = 32$ channels. The concatenated 1D features and 2D features are regularized with dropout ($=0.5$) to avoid overfitting. The 330 dataset is divided into the train set (270 training samples), the validation set (30 validation samples) and the test set (30 test samples) by cross-validations.

Our code is implemented in Pytorch, a publicly available deep learning software. The pretrained model for the distance map prediction is fixed, and the pretrained Resnet50 feature extractor in Faster-RCNN is also fixed for object detection. Other weights in our neural networks are initialized using the default setting. We train all the layers in our deep network simultaneously using the Adam optimizer. The batch size is set to 32. The entire deep network is trained on a single NVIDIA GeForce GTX TITAN X GPU with 12GB memory. It takes about one hour for training.

3.4 Performance

We evaluate the overall performance of the proposed model by performing two sets of experiments. In the first set of experiments, we perform both training and testing on the 330 mTPs released before 2016. In the second set, we perform training on the 330 mTPs dataset and testing on 45 mTPs released after 2017, which is more difficult. In both sets of experiments, we compare our method with the cutting-edge methods representing the single-sequence-based, the evolutionary-based and the consensus methods, respectively.

3.4.1 Training and Testing on 330 mTPs. As shown in Table 1, our model trained on the 330 mTPs dataset through cross-validations achieves 0.767 topology prediction accuracy at the protein-level, which improves about **3%** over the previously

Table 1: Prediction of transmembrane topology label on the 330 dataset. The first (second) row is the assessment on the protein (segment) level, while the remaining rows are evaluated on the residue level. Boldface numbers indicate the best performance.

	Philius	SCAMPI	OCTOPUS	TOPCONS	Ours
Accuracy	0.648	0.691	0.661	0.736	0.767
Fragment recall	0.916	0.938	0.923	0.938	0.966
Q2	0.786	0.825	0.833	0.843	0.872
SOV	0.857	0.867	0.858	0.882	0.927
Recall	0.839	0.801	0.818	0.831	0.836
Precision	0.735	0.735	0.741	0.761	0.840
Mcc	0.632	0.622	0.640	0.660	0.726

best result obtained by TOPCONS. Our approach achieves 0.966 accuracy at the fragment-level, which also improves around **3%** over TOPCONS.

This trend continues to the residue-level, in which our method ranks the first in terms of Q2 accuracy, SOV score, and Mcc. Among them, we achieve 0.872 Q2 accuracy, 0.927 SOV score and 0.726 Mcc, respectively, which are **3.0%**, **4.5%** and **6.6%** better than the previously best predictor. Specifically, Q2 is an accuracy measurement that takes both transmembrane and non-transmembrane regions into account; SOV score tolerates small wrong predictions at the terminal regions of a segment while penalizing more on the erroneous predictions in the middle region of a segment; Mcc takes into account true and false positives and negatives altogether, which is regarded as a balanced measure. The significant improvement of these measures shows that our method could yield much more meaningful topology predictions.

One outlier performance is recall, where for our method 0.836 is slightly worse than but still comparable to the best performance 0.839. However, our better performance in terms of Fragment recall and the significantly better performance in Q2, SOV and Mcc indicates that our model could produce much more meaningful and reliable predictions without many false positives compared to the other methods. This is mainly due to the fact that the distance prediction module removes those false positive transmembrane regions that are not interacted with the other true positive regions; also the object detection module polishes the boundary of the transmembrane region during prediction.

Overall, we believe that the better performance is not only due to the power of the new framework of neural networks that we adopt, but also due to the effective exploitation of the 1D evolutionary and 2D co-evolutionary information, in addition to the fusion of these features together.

3.4.2 Training on 330 mTPs and Testing on 45 mTPs. To show the performance on “real-world” cases, we challenge our method using the 45 mTPs dataset, in which all data were released at least one year later than the 330 dataset and not homologous to the entries in the 330 training set. Moreover, the 45 mTPs were released at year 2017, which means that they were not included as the training set for the other comparison methods (since the latest dataset was

released at year 2015). Thus, the performance results made on the 45 dataset is fair and reliable to all methods for comparison.

Table 2: Prediction of transmembrane topology labels on the 45 dataset.

	Philius	SCAMPI	OCTOPUS	TOPCONS	Ours
Accuracy	0.511	0.511	0.511	0.533	0.600
Fragment recall	0.714	0.748	0.732	0.737	0.793
Q2	0.752	0.797	0.791	0.804	0.833
SOV	0.740	0.771	0.785	0.777	0.843
Recall	0.678	0.632	0.643	0.658	0.676
Precision	0.597	0.606	0.639	0.628	0.704
Mcc	0.486	0.492	0.523	0.514	0.578

Not surprisingly, as shown in Table 2, our method again achieves the best performance of 0.600 and 0.793 at the protein-level and fragment-level, which improves about **6.7%** and **4.5%** over the second best method TOPCONS and SCAMPI, respectively. In terms of the assessments at the residue-level, our method also achieves 0.833 Q2 accuracy, 0.843 SOV score and 0.578 Mcc, respectively, which are about **3.0%**, **5.8%** and **5.5%** better than the second best method. These results show that the improvement of our method is even higher than that over the 330 training set, especially in the protein-level and fragment-level, which means a good generalization capacity of our model, i.e., our approach can be applied over a newly released mTP with possibly novel transmembrane topology.

3.5 Ablation Study

3.5.1 Different components of our model. To discover the vital elements in the success of our proposed model, we conduct an ablation study by removing or replacing individual components in our model. Specifically, we have tested the performance of models without the 2D features, the object detection module, the 1D-2D fusion module, or replacing bidirectional Transformer with the single directional Transformer. From the ablation experiment results on the 330 dataset and 45 dataset presented in Table 3, we find that the bidirectional Transformer and 2D features are the most important component in our model as the SOV performance drops to 0.895 and 0.900 correspondingly on the 330 dataset when we remove the 2D features and run a single directional Transformer rather than the bidirectional one. The same performance degradation of SOV occurs over the 45 dataset as well. Besides, object detection is also critical as SOV drops to 0.910 once we remove the object detection module. Last but not the least, the 1D-2D fusion can be applied to further improve the accuracy of our method.

3.5.2 Different object detection models. The Faster RCNN module in 2D encoder is pre-trained on the ILSVRC CLS-LOC dataset [18]. The resulting model is pre-trained using SGD with initial learning rate 10-3, 0.9 momentum, 0.0001 weight decay, and batch size 4. As shown in Fig. 3, we add several tricks to improve the object detection results, i.e., Feature Pyramid Networks (FPN) [11], cluster anchor sizes

Table 3: An ablation study on 330 dataset and 45 dataset.

Method	330 dataset			45 dataset		
	Q2	SOV	MCC	Q2	SOV	MCC
w/o 2D features	0.856	0.895	0.678	0.813	0.792	0.536
w/o object detection	0.862	0.910	0.691	0.819	0.803	0.550
w/o 1D-2D fusion	0.868	0.905	0.713	0.825	0.830	0.563
w/o bidirectional	0.864	0.900	0.702	0.821	0.808	0.545

and Group Normalization (GN) [27]. However, it is worth mentioning that there still exists a large gap between the train and test performance, i.e., train-mAP_{0.5} can arrive at 94.09 while test mAP_{0.5} is only 54.68. One possible reason is the sparse distribution of the defined helical-helical interactions in the predicted distance map. Another reason is the serious class-imbalance in the four categories. Fortunately, the average precision of category 1 (self-self interaction) can achieve 0.8 AP, which is not too bad for the one we care most. Thus, the 2D encoder for pair-wise features based on object detection can provide auxiliary information for transmembrane topology identification.

Figure 3: Performances mAP for different object detection modules. (a) Original Faster RCNN model, (b) Faster RCNN + FPN, (c) Faster RCNN + FPN + Cluster anchors size, (d) Faster RCNN + FPN + Cluster anchors size + GN.

4 CONCLUSIONS

In this paper, we have presented a novel deep learning framework with stacked bidirectional Transformers to fuse 1D and 2D features simultaneously for sequence labeling, and applied it to the transmembrane topology prediction. This novel framework not only extracts local and global contexts embedded in the 1D sequential features, but also explicitly considers

the 2D pairwise features that bring in informative interactions between the labels that we want to predict. The object detection module exploits the predicted distance matrix to extract the interactions between transmembrane regions for further refining the performance and making the prediction more understandable. Through the fused 1D and 2D features from the evolutionary and co-evolutionary information, the previous state-of-the-art performance in transmembrane topology prediction has been significantly improved. In addition to transmembrane topology prediction as a proof of concept, our framework could be directly applied to a number of other challenging sequence labeling problems in computational biology such as protein secondary structure prediction or disorder prediction. However, our current approach is still not an end-to-end framework since the distance prediction and object detection modules are trained separately due to limited mTPs samples. In the future, it is possible to fix this issue by establishing a framework over large training sample datasets with limited mTPs training samples through transfer learning.

5 ACKNOWLEDGMENTS

This work is supported by Shenzhen Fundamental Research Fund under grants No. KQTD2015033114415450 and No. ZDSYS201707251409055, and by grant No. 2017ZT07X152. This work was supported by the King Abdullah University of Science and Technology (KAUST) Office of Sponsored Research (OSR) under Awards No. FCC/1/1976 17 01, FCC/1/19 76 18 01, FCC/1/1976 23 01, FCC/1/1976 25 01, FCC/1/1976 26 01, and URF/1/3450 01 to X.G. This work was also supported by National Institutes of Health (NIH) [R01GM089753] and National Science Foundation (NSF) [DBI 1564955] to J.X.

REFERENCES

- [1] Andreas Bernsel, Håkan Viklund, Jenny Falk, Erik Lindahl, Gunnar von Heijne, and Arne Elofsson. 2008. Prediction of membrane-protein topology from first principles. *Proceedings of the National Academy of Sciences* 105, 20 (May 2008), 7177–7181.
- [2] Deepti Chaturvedi and Radhakrishnan Mahalakshmi. 2017. Transmembrane β -barrels: evolution, folding and energetics. *Biochimica et Biophysica Acta (BBA)-Biomembranes* 1859, 12 (December 2017), 2467–2482.
- [3] David De Juan, Florencio Pazos, and Alfonso Valencia. 2013. Emerging methods in protein co-evolution. *Nature Reviews Genetics* 14, 4 (2013), 249.
- [4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).
- [5] László Dobson, István Reményi, and Gábor E Tusnady. 2015. CCTOP: a Consensus Constrained TOPology prediction web server. *Nucleic acids research* 43, W1 (2015), W408–W412.
- [6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.
- [7] Jonathan Huang, Vivek Rathod, Chen Sun, Menglong Zhu, Anoop Korattikara, Alireza Fathi, Ian Fischer, Zbigniew Wojna, Yang Song, Sergio Guadarrama, et al. 2017. Speed/accuracy trade-offs for modern convolutional object detectors. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 7310–7311.
- [8] Daniel Kozma, Istvan Simon, and Gabor E Tusnady. 2012. PDBTM: Protein Data Bank of transmembrane proteins after 8 years. *Nucleic acids research* 41, D1 (January 2012), D524–D529.
- [9] Zhen Li, Yukang Gan, Xiaodan Liang, Yizhou Yu, Hui Cheng, and Liang Lin. 2016. Lstm-cf: Unifying context modeling and fusion with lstms for rgb-d scene labeling. In *European Conference on Computer Vision*. Springer, 541–557.
- [10] Zhen Li and Yizhou Yu. 2016. Protein secondary structure prediction using cascaded convolutional and recurrent neural networks. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*. AAAI Press, 2560–2567.
- [11] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. 2017. Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2117–2125.
- [12] Jianzhu Ma, Sheng Wang, Zhiyong Wang, and Jinbo Xu. 2015. Protein contact prediction by integrating joint evolutionary coupling analysis and supervised learning. *Bioinformatics* 31, 21 (August 2015), 3506–3513.
- [13] Debora S Marks, Thomas A Hopf, and Chris Sander. 2012. Protein structure prediction from sequence variation. *Nature biotechnology* 30, 11 (2012), 1072.
- [14] Timothy Nugent and David T Jones. 2012. Detecting pore-lining regions in transmembrane protein sequences. *BMC bioinformatics* 13, 1 (July 2012), 169.
- [15] Michael Remmert, Andreas Biegert, Andreas Hauser, and Johannes Söding. 2012. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nature methods* 9, 2 (February 2012), 173.
- [16] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*. 91–99.
- [17] Sheila M Reynolds, Lukas Käll, Michael E Riffle, Jeff A Bilmes, and William Stafford Noble. 2008. Transmembrane topology and signal peptide prediction using dynamic bayesian networks. *PLoS computational biology* 4, 11 (November 2008), e1000213.
- [18] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. 2015. Imagenet large scale visual recognition challenge. *International journal of computer vision* 115, 3 (2015), 211–252.
- [19] Stefan Seemayer, Markus Gruber, and Johannes Söding. 2014. CCMpredfast and precise prediction of protein residue–residue contacts from correlated mutations. *Bioinformatics* 30, 21 (November 2014), 3128–3130.
- [20] Konstantinos D Tsirigos, Christoph Peters, Nanjiang Shu, Lukas Käll, and Arne Elofsson. 2015. The TOPCONS web server for consensus prediction of membrane protein topology and signal peptides. *Nucleic acids research* 43, W1 (July 2015), W401–W407.
- [21] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*. 5998–6008.
- [22] Håkan Viklund and Arne Elofsson. 2008. OCTOPUS: improving topology prediction by two-track ANN-based preference scores and an extended topological grammar. *Bioinformatics* 24, 15 (August 2008), 1662–1668.
- [23] Sheng Wang, Shiyang Fei, Zongan Wang, Yu Li, Jinbo Xu, Feng Zhao, and Xin Gao. 2018. PredMP: a web server for de novo prediction and visualization of membrane proteins. *Bioinformatics* 1 (August 2018), 3.
- [24] Sheng Wang, Zhen Li, Yizhou Yu, and Jinbo Xu. 2017. Folding membrane proteins by deep transfer learning. *Cell systems* 5, 3 (September 2017), 202–211.
- [25] Sheng Wang, Jian Peng, Jianzhu Ma, and Jinbo Xu. 2016. Protein secondary structure prediction using deep convolutional neural fields. *Scientific reports* 6 (January 2016), 18962.
- [26] Sheng Wang, Siqi Sun, Zhen Li, Renyu Zhang, and Jinbo Xu. 2017. Accurate de novo prediction of protein contact map by ultra-deep learning model. *PLoS computational biology* 13, 1 (January 2017), e1005324.
- [27] Yuxin Wu and Kaiming He. 2018. Group normalization. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 3–19.
- [28] Jinbo Xu. 2018. Distance-based Protein Folding Powered by Deep Learning. *arXiv preprint arXiv:1811.03481* (2018).