IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received February 3, 2020, accepted February 16, 2020, date of publication February 27, 2020, date of current version March 6, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2976797

TM-ZC: A Deep Learning-Based Predictor for the Z-Coordinate of Residues in α-Helical Transmembrane Proteins

CHANG LU[1,2], YINGLI GONG[1,2], ZHE LIU[1,2], YUANZHAO GUO[1,2], ZHIQIANG MA[1,2,3], AND HAN WANG[1,2,3]

[1]School of Information Science and Technology, Northeast Normal University, Changchun 130117, China
[2]Institute of Computational Biology, Northeast Normal University, Changchun 130117, China
[3]Department of Computer Science, College of Humanities and Sciences, Northeast Normal University, Changchun 130117, China

Corresponding authors: Zhiqiang Ma (mazq@nenu.edu.cn) and Han Wang (wangh101@nenu.edu.cn)

This work was supported in part by the National Natural Science Funds of China under Grant 81671328 and Grant 61802057, in part by the Jilin Scientific and Technological Development Program under Grant 20180414006GH, Grant 20180520028JH, and Grant 20170520058JH, in part by the Science and Technology Research Project of the Education Department of Jilin Province under Grant JJKH20190290KJ and Grant JJKH20191309KJ, and in part by the Fundamental Research Funds for the Central Universities under Grant 2412019FZ052 and Grant 2412019FZ048.

ABSTRACT Z-coordinate is an important structural feature of α-helical transmembrane proteins (α-TMPs), which is defined as the distance from a residue to the center of the biological membrane. Since the α-TMP structures from both experimental solved and computational predicted approaches still cannot cover the requirements in relevant research fields, z-coordinate prediction provides an opportunity to partly descript α-TMP structures based on their sequences, further contributes to function annotation and drug target discovery. For the purpose of improving the prediction accuracy and providing a convenient tool, we proposed a deep learning-based predictor (TM-ZC) for the z-coordinate of residues in α-TMPs. TM-ZC used the one-hot code and the PSSM as input features for a convolutional neural network (CNN) regression model. The experimental results demonstrated that TM-ZC was a powerful predictor, which is simple and fast, and achieved a considerable performance: the average error was 1.958, the percent of prediction error within 3Å was 77.461%, and the correlation coefficient (CC) was 0.922. We further discussed the usefulness of TM-ZC predicted z-coordinate and found its high consistency with topology structure and the enhancement of the surface accessibility prediction.

INDEX TERMS α-helical transmembrane protein, convolutional neural network (CNN), regression, Z-coordinate of residues.

I. INTRODUCTION

α-helical transmembrane proteins (α-TMPs) are the major category of transmembrane proteins (TMPs). According to the statistics of the Universal Protein Resource (UniProt) [1], α-TMPs account for more than 98% of the TMPs. α-TMPs play numerous roles in basic physiology and pathophysiology, including signal transduction [2], nutrients or drugs reception [3], immune response [4], and enzyme activation [5]. Malfunction of α-TMP may cause many diseases, such as autism [6], epilepsy [7], and cancer [8]–[11]. Consequently, α-TMPs are the major targets for more than half of known drugs, the detailed structure of them would be paramount to the success of drug discovery [12], [13]. Unfortunately, despite their important biological functions, determination of high-resolution structures of α-TMPs persist technical difficulties, only approximately 5% of them are determined.

For this reason, the TMP relevant researches are promoted currently by means of many structural descriptors abstracted from primary sequences. Beyond high-resolution structural information, some low-resolution structural descriptors, such as topology structure, surface accessibility, and z-coordinate, can also provide valuable information about α-TMPs. In recent years, a lot of illuminating methods have been proposed and accessed great achievements. Such as the

The associate editor coordinating the review of this manuscript and approving it for publication was Quan Zou.

topology structure prediction methods of α-TMPs [14], [15], especially, S. H. Feng et al. firstly developed a multiscale deep learning protocol (MemBrain 3.0) that includes two submodules: transmembrane helix prediction and orientation prediction [16]. Likewise, several methods have been developed to predict the surface accessibility of α-TMPs and achieved considerable performance [17], [18]. For example, our previous work [19] presented a deep learning-based predictor (TMP-SSurface), which combined the Inception and the CapsuleNet by using one-hot code and PSSM as input features.

Z-coordinate of a residue in α-TMP is defined as the distance from the residue to the center of the membrane [20]. Similar to the topology structure, z-coordinate also reflect the relationship between the residue and the membrane, but by continuous numerical measurement. The z-coordinate is highly correlated with the ligand-binding and the protein-protein binding regions because these binding regions are always specifically located on transmembrane, water-soluble, or junction regions. The predicted z-coordinate is helpful for the topology prediction [21], structural classification [22], burial status prediction [23], and many other research fields [24], [25]. Accurate predicting the z-coordinate of residues in α-TMPs by computational methods is not only an intermediate step towards structure determination, but also a potential property that may assist the function annotation, drug target discovery, and other associated problems [21], [26], [27].

However, the z-coordinate study has not received as much attention as the study of topology structure and surface accessibility. ZPRED [20], only one z-coordinate predictor published more than a decade, where Artificial Neural Network (ANN) and Hidden Markov Model (HMM) were combined, and sequential features were used as inputs. ZPRED is the pioneering work on the z-coordinate prediction, but its webserver is no longer available. To further support α-TMP research, a z-coordinate predictor is surely needed, which should be more reliable and with high accuracy and performance. In the past 15 years, the number of α-TMPs' structures has increased more than 10 times. More data can be used to improve this research, and deep learning method provides a novel opportunity to construct a more powerful predictor that is simpler and faster, and achieve higher performance at the same time.

In this work, we proposed a deep learning-based predictor (TM-ZC) for the z-coordinate of residues in α-TMPs. TM-ZC used the one-hot code and the PSSM as input features for a convolutional neural network (CNN) regression model. The experimental result proved a considerable performance of TM-ZC. The average error was 1.865Å, the present of prediction error within 3Å was 76.703%, and the correlation coefficient (CC) was 0.917. Besides, We also tested the usefulness of TM-ZC predicted z-coordinate on the problems of surface accessibility prediction and topology structure prediction. We tried to distinguish the transmembrane residues from non-transmembrane residues by limiting the threshold of the z-coordinate predicted by TM-ZC. The experimental result demonstrated that there exists a strong correlation between the topology structure and TM-ZC predicted z-coordinate. For surface accessibility prediction, we added the TM-ZC predicted z-coordinate as an additional feature of our previous work to predict the surface accessibility of α-TMPs. The experimental result proved that TM-ZC predicted z-coordinate could enhance the prediction performance. A stable webserver is accessible freely in http://icdtools.nenu.edu.cn/TM-ZC.

II. MATERIALS AND METHODS

A. BENCHMARK DATASETS

A dataset used by ZPRED was constructed in 2005 that consisted of 101 non-homologous chains from 46 complexes. We believe that a more comprehensive benchmark dataset is urged. The Protein Data Bank of Transmembrane Proteins (PDBTM) [28] is the most widely used comprehensive data bank for transmembrane proteins. It was created by scanning all PDB entries with the TMDET algorithm [29]. According to the statistics of PDBTM, the number of α-TMPs has increased more than 10 times in the past 15 years. We downloaded 3820 complexes with 13,209 α-TMP sequences from PDBTM (version: 2019-05-10). Removing the sequences that contain residues other than 20 standard amino acids. Removing the short sequences with residues less than 30 because they are always considered as peptides. To reduce the negative effect of homology bias [30], we clustered the rest of the proteins by running CD-HIT with a 0.3 sequence identity cut-off, and the longest sequences in each cluster were collected. After pre-processing, 851 α-TMPs with 223,310 residues were left. Among them, 50 sequences were randomly selected as the independent testing dataset (ZC-test50) for the independent test to verified the robustness of TM-ZC. The remaining 801 sequences were used for training and tuning the prediction model, among them, 50 sequences were randomly selected as the validation dataset (ZC-valid50), and the remaining 751 sequences were built as the training dataset (ZC-train751). The process of selecting the validation dataset were performed for ten times for ten-fold cross validation. The performance reported in this work when training the models was the average performance of sub-models in ten-fold cross validation. All the datasets that used in this work can be found in Supplementary Materials.

B. CALCULATION OF Z-COORDINATE

The original coordinates of residues recorded in the PDB files need to be rotated and moved according to the relative positions of the protein and the membrane. The observed z-coordinate value can be calculated by using Formula 1:

$$\left[x_i^{'}, y_i^{'}, z_i^{'}\right] = [x_i, y_i, z_i]A^{-1} + \left[b_x, b_y, b_z\right], \tag{1}$$

where $[x_i, y_i, z_i]$ represents the original coordinate of the alpha-carbon atom of the ith residue recorded in the PDB files obtained from PDBTM, A is the matrix that rotated the protein

perpendicular to the membrane. $[b_x, b_y, b_z]$ is a vector that moved the protein to the right place related to the membrane. Both A and $[b_x, b_y, b_z]$ were obtained from TMDET [29]. The z-coordinate of the membrane center is 0.

Then, as the same process of ZPRED, two steps of threshold cutting was performed:

Step 1: We took the absolute value of $z_i^{'}$ that limiting the threshold from $(-\infty, +\infty)$ to $[0, +\infty)$, considering only the distance between the residues and the membrane center without the orientation. The absolute value can be calculated as:

$$z_{i[0,+\infty)} = \left| z_i^{'} \right|, \tag{2}$$

Step 2: Limiting $\left|z_i^{'}\right|$ in range [5, 25] so that all the residues with $\left|z_i^{'}\right|$ between 0 and 5 were defined to be in a central hydrophobic region and the z-coordinate values were set to 5, all the values above 25 were set to 25 because they were considered as non-transmembrane residues. The z-coordinate values within the range [5, 25] were considered as the observed z-coordinate labels of TM-ZC. It can be calculated as:

$$z_{i[5,25]} = \begin{cases} 5, & \left|z_i^{'}\right| \leq 5 \\ \left|z_i^{'}\right|, & 5 < \left|z_i^{'}\right| < 25 \\ 25, & 25 \leq \left|z_i^{'}\right|, \end{cases} \tag{3}$$

C. ENCODING OF PROTEIN FRAGMENTS

1) EVOLUTIONARY CONSERVATION (PSSM)

In the process of evolution, certain genetic characteristics of proteins have become increasingly prominent among homologous proteins. It has been proved that fragments with high evolutionary conservation always related to the structural or functional needs of the proteins [31], [32]. The position-specific score matrix (PSSM) is an effective descriptor extracted from the result of multiple sequence alignment [33]. The PSSM of a given protein was calculated by running PSI-BLAST [34] against the UniRef50 database (released on October 16, 2019) with e-value threshold 0.001 and 3 iterations. The PSSM of a protein can be defined as a $20 \times L$ matrix:

$$PSSM = \begin{bmatrix} P_{1,AA_1} & P_{1,AA_2} & \cdots & P_{1,AA_{20}} \\ P_{2,AA_1} & P_{2,AA_2} & \cdots & P_{2,AA_{20}} \\ \vdots & \vdots & \vdots & \vdots \\ P_{L,AA_1} & P_{L,AA_2} & \cdots & P_{L,AA_{20}} \end{bmatrix}, \tag{4}$$

where P_{i,AA_j} represent the element's value of PSSM, which represents the occurrence frequency of AA_j at the i-th position of the given protein in the result of multiple sequence alignment. L represents the length of the protein. Then, we used the logistic function to normalized each element of PSSM into [0, 1]:

$$P_{i,AA_j}^{'} = \frac{1}{1 + e^{-P_{i,AA_j}}}, \tag{5}$$

2) ONE-HOT CODE

One-hot coding is a sparse coding method used to represents the type of each residue in a protein sequence. It is the most direct way to describe the protein sequence, reflecting the most primitive arrangement information of 20 standard amino acids. It proved to be a valid feature for deep learning-based protein function predictors [35]–[39]. The one-hot code of a protein can be defined as a $20 \times L$ matrix:

$$one-hot = \begin{bmatrix} O_{R_1,AA_1} & O_{R_1,AA_2} & \cdots & O_{R_1,AA_{20}} \\ O_{R_2,AA_1} & O_{R_2,AA_2} & \cdots & O_{R_2,AA_{20}} \\ \vdots & \vdots & \vdots & \vdots \\ O_{R_L,AA_1} & O_{R_L,AA_2} & \cdots & O_{R_L,AA_{20}} \end{bmatrix}, \tag{6}$$

where O_{R_i,AA_j} represents the element's value of one-hot code. R_i is the type of residue on position i. AA_j is the type of 20 standard amino acids. $O_{R_i,AA_j} = 1$ if $R_i = AA_j$; $O_{R_i,AA_j} = 0$ if $R_i \neq AA_j$. L represents the length of the protein.

3) SLIDING WINDOW

Residues are not isolated in the protein sequence but are arranged in a certain order [40], [41]. The position and function of a target residue are greatly affected by its adjacent residues. Hence, we employed a sliding window scheme that presents the target residue by a protein fragment. Here, we set the window size to 25: target residue with 12 residues from upstream and 12 residues from downstream. The target residue was the center of the protein fragment. At the terminal of the sequence, the features' values of the overflow part of the sliding window were filled with 0. At last, we got a 40×25 matrix as the feature of each residue.

D. MODEL DESIGN

The convolutional neural network (CNN) is a kind of feed-forward neural network, in which the neurons can reflect the surrounding information within the coverage of the convolution kernel. The training dataset was used to training the network in each iteration, and the validation dataset was used to validate the performance of this iteration and feedback to the next iteration. It solves the problem of traditional machine learning's dependence on manual features and can learn useful features directly from primitive data. CNN performs great in the field of image and video recognition [42], natural language processing [43], and medical diagnosis [44]. Due to its effectiveness, CNN has been widely used in the field of bioinformatics, such as super-enhancer prediction [45] and drug-disease association prediction [46]. In this work, our goal was to make the prediction model as simple as possible. So we constructed a small network architecture. The design of the prediction model is shown in Fig. 1. All the convolution layers contained 256 kernels with the size of 3×3 and the stride of 1, and the activation function was ReLU. All the pooling layers were max pooling with 256 kernels with the size of 2×2 and the stride of 2. In this model, one convolution layer was first performed to extract features

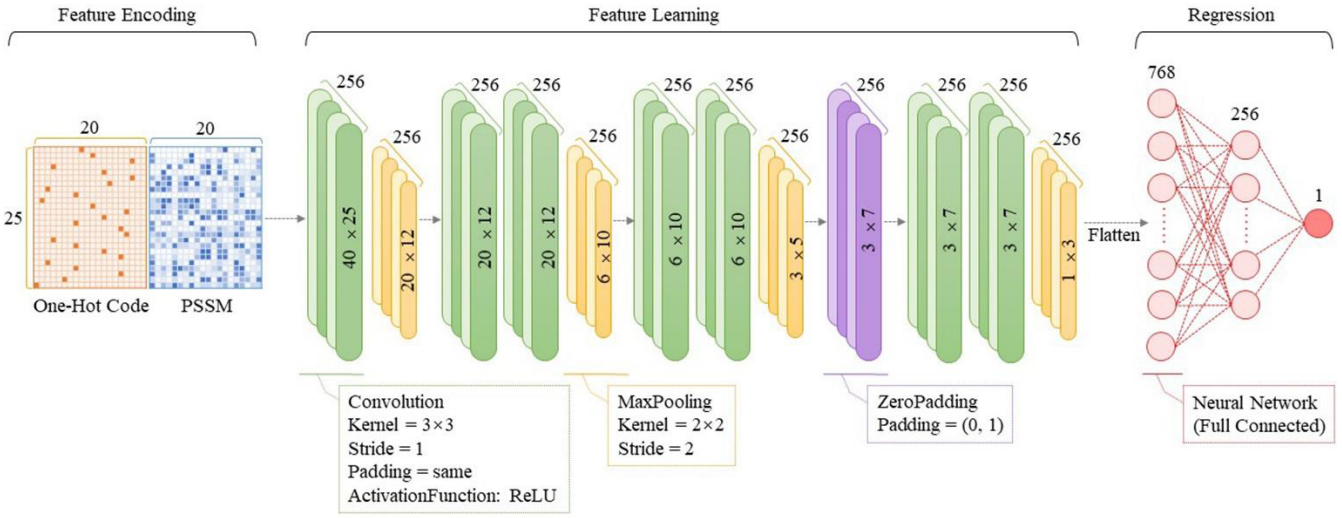

FIGURE 1. The structure of TM-ZC.

from the original encoded feature matrix. Then, one max pooling layer was performed. After that, two convolution layers, one max pooling layer, two convolution layers, and one max pooling layer were performed in order. At this time, the size of the feature matrix was 3×5, it was too small to perform the following layers. Thus, we performed a zero-padding layer to padding the feature matrix to the size of 3×7. Then, two convolution layers and one max pooling layer were performed. At this time, we extract features as 256 matrixes with the size of 1×3. We flatten them in a line, and a full-connected neural network was performed. The input layer contains 768 neurons, the hidden layer contains 256 neurons, and the output layer contains just one neuron. The opportunity of a target residue that belongs to the output neuron was considered as the prediction result.

E. PERFORMANCE EVALUATION

In order to measure the prediction performance of TM-ZC, three indicators were used to reflect it: the mean absolute error (MAE), the Pearson correlation coefficient (CC), and the percent of results that error less than 3Å ($P_{3Å}$). MAE reflects the average deviation between the predicted and observed z-coordinate of all residues. It ranged in [0, 1], the smaller the MAE value, the better the performance. CC reflects the linear correlation between predicted and observed z-coordinate. It ranged in [−1, 1], the more the CC close to 1, the better the performance. $P_{3Å}$ reflects the ratio of the ideal prediction results, the threshold inherited from ZPRED. It ranged in [0%, 100%], the higher the ratio, the better the performance.

$$\mathrm{MAE} = \frac{1}{L}\sum_{i=1}^{L} |y_i - x_i|, \tag{7}$$

$$\mathrm{CC} = \frac{\sum_{i=1}^{L} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\left[\sum_{i=1}^{L} (x_i - \bar{x})^2\right]\left[\sum_{i=1}^{L} (y_i - \bar{y})^2\right]}}, \tag{8}$$

$$\mathrm{P}_{3Å} = \frac{N_{|y_i - x_i| < 3Å}}{L}, \tag{9}$$

where L represents the number of residues, x_i and y_i represent the observed and predicted z-coordinate of the ith residue, and $\bar{x}$ and $\bar{y}$ represent the corresponding mean value, $N_{|y_i - x_i| < 3Å}$ represents the number of residues that prediction error less than 3Å.

III. RESULTS AND DISCUSSION

A. FEATURE ANALYSIS

In order to investigate the effectiveness of different kinds of features and their contribution to the prediction model, we performed the ablation study on features. Three models were built by using one-hot code, PSSM, and both of them. As shown in TABLE 1, PSSM outperforms one-hot code. This phenomenon reflects the strong correlation between evolutionary conservation and protein structure. Although the model using one-hot code alone performed poorly, it complemented the PSSM feature that the model achieved the best performance while using two features together.

B. EFFECT OF WINDOW SIZE

A sliding window scheme was used in this work, and the value of the window size greatly affected the prediction performance of TM-ZC. We test the possible value of window size from 15 to 31 with the step size of 2. The prediction performance on the validation dataset while training TM-ZC

TABLE 1. Performance comparison of the different models on the feature ablation study.

window size	MAE	CC	$P_{3Å}$ (%)
One-hot code	8.801	0.675	32.540
PSSM	4.598	0.873	52.292
One-hot code+PSSM	**4.373**	**0.917**	**55.343**

The bolded parts represent the highest value of the corresponding evaluation indicator.

TABLE 2. The prediction performance by using different window sizes.

window size	MAE	CC	$P_{3Å}$ (%)
15	5.267	0.883	47.833
17	4.918	0.894	50.073
19	1.827	0.924	78.66400
21	1.726	0.927	79.12000
23	1.755	0.925	78.80800
25	**1.673**	**0.939**	**79.10700**
27	1.741	0.93	79.69300
29	1.838	0.926	79.00700
31	1.882	0.922	80.05700

by using different window sizes is illustrated in TABLE 2. As the window size gets larger, the performance of TM-ZC gradually improved and reached the top when the value of window size reached to 25. Thus, we set the window size to 25 in all the experiments.

C. EFFECT OF CUTTING THRESHOLD

As described in the *"Section II-BCalculation of Z-coordinate"*, the original coordinates of residues recorded in the PDB files were rotated and moved according to the relative positions of the protein and the membrane. Then, two steps of threshold cutting were performed. The first step took the absolute value that cutting the threshold from $(-\infty, +\infty)$ to $[0, +\infty)$. The second step limited the value in the range $[5, 25]$. Three training labels with different cutting thresholds were obtained, and three models were built by using different labels. The performance of these models on the validation dataset (ZC-valid50) is illustrated in TABLE 3.

It is obvious that the performance of the model using the original labels with the threshold of $(-\infty, +\infty)$ was poor, and the other two models significantly outperformed it. There proposed a possible explanation for this phenomenon: Residues that are symmetrical about the membrane center always have similar features while their labels are opposite to each other, which makes it difficult for the prediction model to find the relationship between the features and the corresponding labels. It can also be seen that the models using the labels with the threshold of [5, 25] achieved the best performance.

TABLE 3. The Prediction performance of models, which are trained by using training labels with different cutting thresholds.

Cutting Threshold	MAE	CC	$P_{3Å}$ (%)
Performance on ZC-valid50 dataset			
$(-\infty, +\infty)$	30.067	0.010	17.080
$[0, +\infty)$	6.077	0.903	45.266
[5,25]	**1.65**	**0.931**	**80.109**

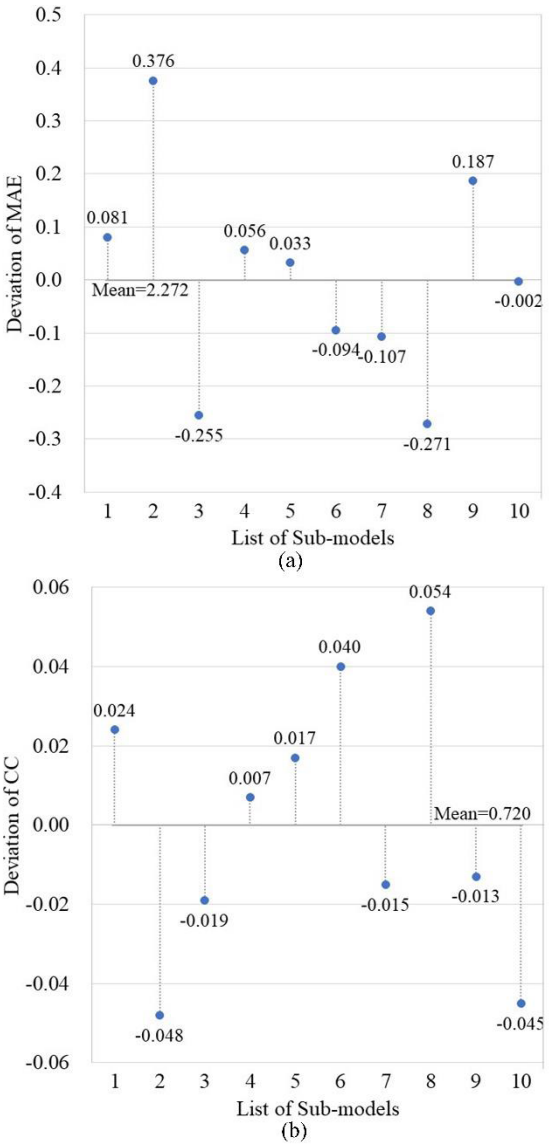

FIGURE 2. The stability of the training process. (a) The performance of MAE on the cross validation. (b) The performance of CC on the cross validation.

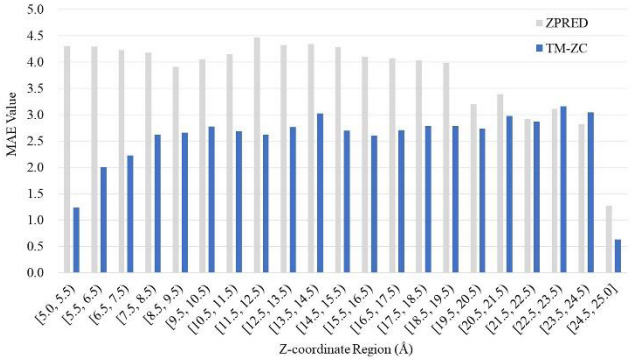

FIGURE 3. Comparison of MAE in different z-coordinate regions.

D. PERFORMANCE OF TM-ZC

In order to investigate the performance of TM-ZC and verified its stability, we performed ten-fold cross validation.

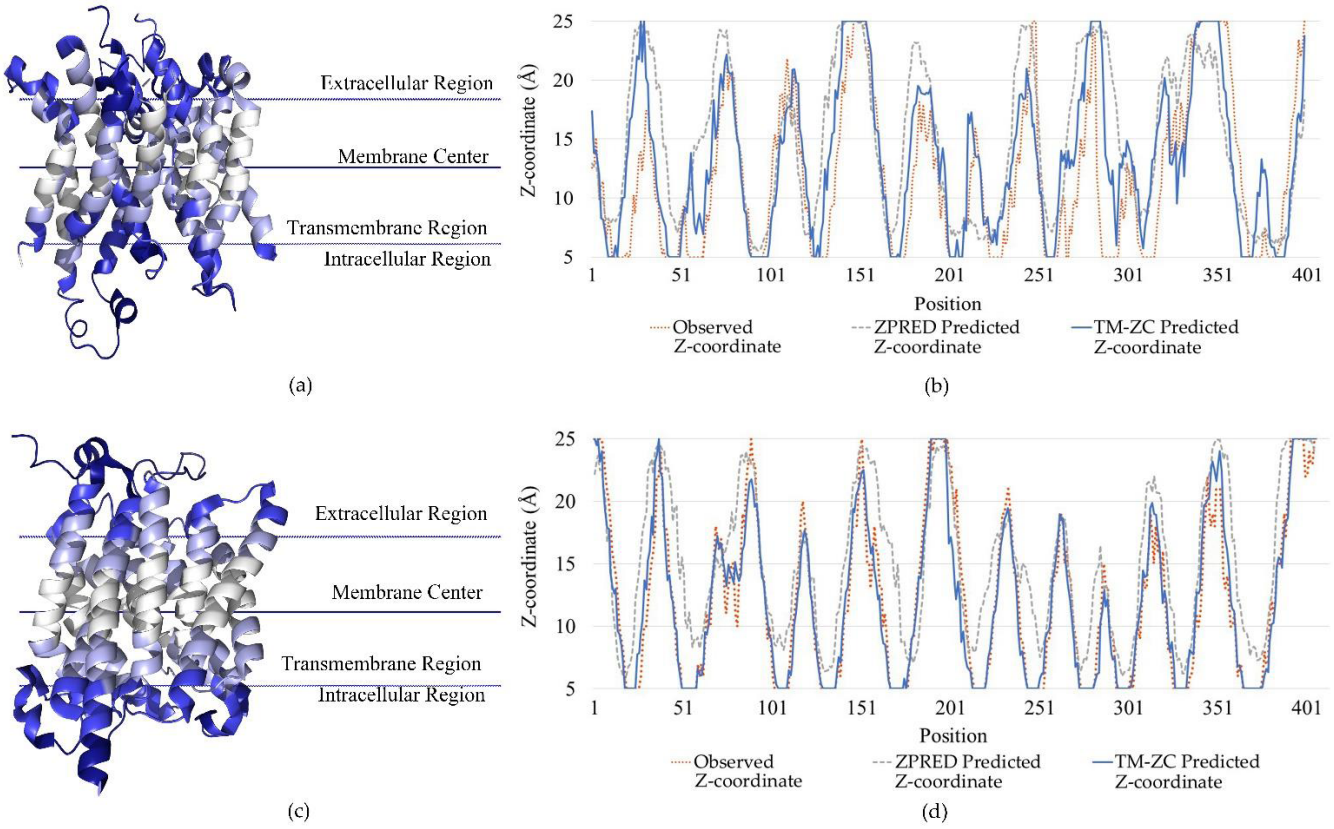

FIGURE 4. Case studies of TM-ZC: took 6EU6_A and 5L25_A as examples. (a) and (c) are 3D visualization of the prediction result of 6EU6_A and 5L25_A, respectively. As the value of predicted z-coordinate increases, the color changes from white to dark blue. (b) illustrate the curves of the z-coordinate value of 6EU6_A: the observed value (orange dotted curve), the ZPRED predicted value (gray dotted curve), and the TM-ZC predicted value (blue curve). (d) illustrate the curves of the z-coordinate value of 5L25_A, which is as same as (b).

Fig. 2 shows the performance of the TM-ZC in terms of (a) MAE value and (b) CC value. As shown in (a), the mean value of MAE of ten models is 2.215, the deviation between the MAE of each sub-model and the mean value is shown as the label value. As shown in (b), the mean value of CC of ten models is 0.915, the deviation between the CC of each sub-model and the mean value is shown as the label value. Fig. 2 exhibits that TM-ZC performed stably among cross-validation.

E. COMPARE WITH ZPRED

ZPRED is the only predictor in existence. We compared the TM-ZC with ZPRED to investigated its effectiveness. Fig. 3 illustrates the mean absolute error (MAE) of ZPRED and TM-ZC in different z-coordinate regions. It could be seen that TM-ZC outperforms ZPRED in most of the z-coordinate regions, especially in regions with low z-coordinate value. A low z-coordinate indicates that the residue locates in the hydrophobic transmembrane region. It proved that TM-ZC performed great for residues in the transmembrane region. In the region [19.5, 24.5), the advantages of TM-ZC became less obvious and even worse than ZPRED. It shows that TM-ZC's prediction performance for residues in the junction area on the membrane surface needs to be further improved.

TABLE 4. Comparison of the performance between ZPRED and TM-ZC.

Predictor	MAE	CC	$P_{3Å}$ (%)
ZPRED	3.293	0.813	59.731
TM-ZC	**1.958**	**0.922**	**77.461**

The overall prediction performance comparison between ZPRED and TM-ZC is illustrated in TABLE 4. It is obvious that TM-ZC outperformed ZPRED. The MAE reduced by approximately 43%, the percent of results that error less than 3Å ($\mathbf{P_{3Å}}$) increased more than 28% and the increment of the Pearson correlation coefficient (CC) up to 0.1.

F. CASE STUDIES

We performed case studies to demonstrate the effectiveness of TM-ZC further. 6EU6_A and 5L25_A from ZC-test50 were chosen as examples. 6EU6_A is an eleven-transmembrane protein from *Escherichia Coli*. It is the target of ATP, Dodecyl-Alpha-D-Maltoside, and other ligands. 5L25_A is a ten-transmembrane protein from *Saccharomyces Cerevisiae*. It plays an important role in the process of anion exchange and borate transport. The prediction results of both proteins are illustrated in Fig. 4.

In Fig. 4: (a) visualized the TM-ZC predicted z-coordinate of 6EU6_A, the darker the color, the larger the precited z-coordinate value. (b) is the curve of the z-coordinate of 6EU6_A, including the observed value, the ZPRED predicted value, and the TM-ZC predicted value. (c) and (d) illustrated the same information of 5L25_A as (a) and (b). It could be seen from (a) and (c) that the z-coordinate predicted by TM-ZC is basically in line with the actual situation. It is further confirmed in (b) and (d) that the curve of the TM-ZC predicted value fits the curve of the observed value to a high degree, and it is better than ZPRED.

G. Z-COORDINATE CORRELATED WITH TOPOLOGY STRUCTURE

The z-coordinate directly relates to the topology structure of residues in TMPs. We tried to distinguish the transmembrane residues from non-transmembrane residues by limiting the threshold of the z-coordinate predicted by TM-ZC. This experiment was performed on the ZC-test50 dataset and achieved the highest accuracy (79.268%) at the threshold of 14.5Å. The residues with z-coordinate within 14.5Å were considered as transmembrane residues. That means the mean thickness of the membrane was $14.5 \times 2 = 29$Å, and the biological experiment proved that the mean thickness of the membrane is about 30Å [47]. Therefore, the threshold of TM-ZC predicted z-coordinate is in line with facts, and there exists a strong correlation between topology structure and z-coordinate.

H. TM-ZC ENHANCE THE PREDICTION OF SURFACE ACCESSIBILITY

In addition to having a strong correlation with the topology structure, the TM-ZC predicted z-coordinate could also enhance the prediction of the surface accessibility of α-TMPs residues. Our team once proposed a predictor (TMP-SSurface) for the relative accessible surface area (rASA) prediction of residues in α-TMPs [19]. TMP-SSurface used one-hot code, terminal flag, and PSSM as the input features of a deep learning-based regression method. We added TM-ZC predicted z-coordinate as an additional feature to verified the usefulness of TM-ZC and we were glad to find that the value of the Pearson correlation coefficient (CC) increased from 0.581 to 0.604. This experiment proved that there is also a correlation between z-coordinates and rASA.

IV. CONCLUSION

The z-coordinate of a residue in α-TMPs is defined as the distance between the residue and the membrane center. It is an important structure descriptor that highly correlated with the function regions of α-TMPs. Up to now, ZPRED is the only one predictor for this problem and needed to be further improved. In this work, we proposed a deep learning-based predictor (TM-ZC) to predict the z-coordinate of residues in α-TMPs. TM-ZC is a simple CNN-based predictor that used one-hot code and PSSM as input features. TM-ZC achieved great performance that the MAE was 1.958, the CC was 0.922, and the percent of prediction error within 3Å was 77.461%. Experiments showed the contribution of two kinds of feature, and find that PSSM features were more powerful. We also verified the correlations between TM-ZC predicted z-coordinate and tested the usefulness of TM-ZC predicted z-coordinate on the problems of surface accessibility prediction. Experiments proved that TM-ZC could provide effective support for related problems. We are confident that TM-ZC can be further used in more researches on transmembrane proteins.

REFERENCES

[1] T. UniProt Consortium, "UniProt: A worldwide hub of protein knowledge," *Nucleic Acids Res.*, vol. 47, no. D1, pp. D506–D515, Jan. 2019, doi: 10.1093/nar/gky1049.

[2] L. He, E. B. Cohen, A. P. B. Edwards, J. Xavier-Ferrucio, K. Bugge, R. S. Federman, D. Absher, R. M. Myers, B. B. Kragelund, D. S. Krause, and D. DiMaio, "Transmembrane protein aptamer induces cooperative signaling by the EPO receptor and the cytokine receptor β-common subunit," *iScience*, vol. 17, pp. 167–181, Jul. 2019, doi: 10.1016/j.isci.2019.06.027.

[3] M. L. Colgrave, K. Byrne, S. V. Pillai, B. Dong, A. Leonforte, J. Caine, L. Kowalczyk, J. A. Scoble, J. R. Petrie, S. Singh, and X.-R. Zhou, "Quantitation of seven transmembrane proteins from the DHA biosynthesis pathway in genetically engineered canola by targeted mass spectrometry," *Food Chem. Toxicol.*, vol. 126, pp. 313–321, Apr. 2019, doi: 10.1016/j.fct.2019.02.035.

[4] L. Chen, Y. Zhang, S. Zhang, Y. Chen, X. Shu, J. Lai, H. Cao, Y. Lian, Z. Stamataki, and Y. Huang, "A novel T-cell epitope in the transmembrane region of the hepatitis B virus envelope protein responds upon dendritic cell expansion," *Arch. Virology*, vol. 164, no. 2, pp. 483–495, Feb. 2019, doi: 10.1007/s00705-018-4095-0.

[5] X. Duan, X. Liao, S. Li, Y. Li, M. Xu, Y. Wang, H. Ye, H. Zhao, C. Yang, X. Zhu, and L. Chen, "Transmembrane protein 2 inhibits zika virus replication through activation of the janus kinase/signal transducers and activators of transcription signaling pathway," *Future Virol.*, vol. 14, no. 1, pp. 9–19, Jan. 2019, doi: 10.2217/fvl-2018-0115.

[6] S. K. Rafi, A. Fernández-Jaén, S. Álvarez, O. W. Nadeau, and M. G. Butler, "High functioning autism with missense mutations in synaptotagmin-like protein 4 (SYTL4) and transmembrane protein 187 (TMEM187) genes: SYTL4-protein modeling, protein-protein interaction, expression profiling and MicroRNA studies," *Int. J. Mol. Sci.*, vol. 20, no. 13, p. 3358, Jul. 2019, doi: 10.3390/ijms20133358.

[7] Y. Tanabe, T. Taira, A. Shimotake, T. Inoue, T. Awaya, T. Kato, A. Kuzuya, A. Ikeda, and R. Takahashi, "An adult female with proline-rich transmembrane protein 2 related paroxysmal disorders manifesting paroxysmal kinesigenic choreoathetosis and epileptic seizures," *Rinsho Shinkeigaku*, vol. 59, no. 3, pp. 144–148, Mar. 2019, doi: 10.5692/clinicalneurol.cn-001228.

[8] Y. Moon, W. Lim, and B. Jeong, "Transmembrane protein 64 modulates prostate tumor progression by regulating Wnt3a secretion," *Oncol. Lett.*, vol. 18, no. 1, pp. 283–290, Jul. 2019, doi: 10.3892/ol.2019.10324.

[9] D. Tao, J. Liang, Y. Pan, Y. Zhou, Y. Feng, L. Zhang, J. Xu, H. Wang, P. He, J. Yao, Y. Zhao, Q. Ning, W. Wang, W. Jiang, J. Zheng, and X. Wu, "*In vitro* and *in vivo* study on the effect of lysosome-associated protein transmembrane 4 beta on the progression of breast cancer," *J. Breast Cancer*, vol. 22, no. 3, pp. 375–386, Sep. 2019, doi: 10.4048/jbc.2019.22.e43.

[10] J. Yan, Y. Jiang, J. Lu, J. Wu, and M. Zhang, "Inhibiting of proliferation, migration, and invasion in lung cancer induced by silencing interferon-induced transmembrane protein 1 (IFITM1)," *BioMed Res. Int.*, vol. 2019, pp. 1–9, May 2019, doi: 10.1155/2019/9085435.

[11] K. Qu, F. Gao, F. Guo, and Q. Zou, "Taxonomy dimension reduction for colorectal cancer prediction," *Comput. Biol. Chem.*, vol. 83, Dec. 2019, Art. no. 107160, doi: 10.1016/j.compbiolchem.2019.107160.

[12] T. Langó, G. Róna, É. Hunyadi-Gulyás, L. Turiák, J. Varga, L. Dobson, G. Várady, L. Drahos, B. G. Vértessy, K. F. Medzihradszky, G. Szakács, and G. E. Tusnády, "Identification of extracellular segments by mass spectrometry improves topology prediction of transmembrane proteins," *Sci. Rep.*, vol. 7, no. 1, Feb. 2017, Art. no. 42610, doi: 10.1038/srep42610.

[13] L. Yu, X. Sun, S. W. Tian, X. Y. Shi, and Y. L. Yan, "Drug and non-drug classification based on deep learning with various feature selection strategies," *Current Bioinf.*, vol. 13, no. 3, pp. 253–259, 2018, doi: 10.2174/1574893612666170125124538.

[14] A. Krogh, B. Larsson, G. von Heijne, and E. L. L. Sonnhammer, "Predicting transmembrane protein topology with a hidden Markov model: Application to complete genomes," *J. Mol. Biol.*, vol. 305, no. 3, pp. 567–580, Jan. 2001, doi: 10.1006/jmbi.2000.4315.

[15] H. Wu, K. Wang, L. Lu, Y. Xue, Q. Lyu, and M. Jiang, "Deep conditional random field approach to transmembrane topology prediction and application to GPCR three-dimensional structure modeling," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 14, no. 5, pp. 1106–1114, Sep. 2017, doi: 10.1109/TCBB.2016.2602872.

[16] S. H. Feng, W. X. Zhang, J. Yang, Y. Yang, and H. B. Shen, "Topology prediction improvement of alpha-helical transmembrane proteins through helix-tail modeling and multiscale deep learning fusion," *J. Mol. Biol.*, vol. 432, no. 4, pp. 1279–1296, Dec. 2019, doi: 10.1016/j.jmb.2019.12.007.

[17] K. Illergard, S. Callegari, and A. Elofsson, "MPRAP: An accessibility predictor for a-helical transmem-brane proteins that performs well inside and outside the membrane," *BMC Bioinf.*, vol. 11, no. 1, p. 333, Jun. 2010, doi: 10.1186/1471-2105-11-333.

[18] X. Yin, J. Yang, F. Xiao, Y. Yang, and H.-B. Shen, "MemBrain: An easy-to-use online webserver for transmembrane protein structure prediction," *Nano-Micro Lett.*, vol. 10, no. 1, 2018, Art. no. 2, doi: 10.1007/s40820-017-0156-2.

[19] C. Lu, Z. Liu, B. Kan, Y. Gong, Z. Ma, and H. Wang, "TMP-SSurface: A deep learning-based predictor for surface accessibility of transmembrane protein residues," *Crystals*, vol. 9, no. 12, p. 640, Dec. 2019, doi: 10.3390/cryst9120640.

[20] E. Granseth, H. Viklund, and A. Elofsson, "ZPRED: Predicting the distance to the membrane center for residues in -helical membrane proteins," *Bioinformatics*, vol. 22, no. 14, pp. e191–e196, Jul. 2006, doi: 10.1093/bioinformatics/btl206.

[21] A. Bernsel, H. Viklund, A. Hennerdal, and A. Elofsson, "TOPCONS: Consensus prediction of membrane protein topology," *Nucleic Acids Res.*, vol. 37, pp. W465–W468, May 2009, doi: 10.1093/nar/gkp363.

[22] H. Viklund, E. Granseth, and A. Elofsson, "Structural classification and prediction of reentrant regions in α-helical transmembrane proteins: Application to complete genomes," *J. Mol. Biol.*, vol. 361, no. 3, pp. 591–603, Aug. 2006, doi: 10.1016/j.jmb.2006.06.037.

[23] Y. Park, S. Hayat, and V. Helms, "Prediction of the burial status of transmembrane residues of helical membrane proteins," *BMC Bioinf.*, vol. 8, no. 1, p. 302, Aug. 2007, doi: 10.1186/1471-2105-8-302.

[24] C. Papaloukas, E. Granseth, H. Viklund, and A. Elofsson, "Estimating the length of transmembrane helices using Z-coordinate predictions," *Protein Sci.*, vol. 17, no. 2, pp. 271–278, Feb. 2008, doi: 10.1110/ps.073036108.

[25] B. Wallner, "ProQM-resample: Improved model quality assessment for membrane proteins by limited conformational sampling," *Bioinformatics*, vol. 30, no. 15, pp. 2221–2223, Apr. 2014, doi: 10.1093/bioinformatics/btu187.

[26] T. Nugent and D. T. Jones, "Accurate de novo structure prediction of large transmembrane protein domains using fragment-assembly and correlated mutation analysis," *Proc. Nat. Acad. Sci. USA*, vol. 109, no. 24, pp. E1540–E1547, May 2012, doi: 10.1073/pnas.1120036109.

[27] A. Rose, S. Lorenzen, A. Goede, B. Gruening, and P. W. Hildebrand, "RHYTHM–a server to predict the orientation of transmembrane helices in channels and membrane-coils," *Nucleic Acids Res.*, vol. 37, pp. W575–W580, May 2009, doi: 10.1093/nar/gkp418.

[28] D. Kozma, I. Simon, and G. E. Tusnády, "PDBTM: Protein data bank of transmembrane proteins after 8 years," *Nucleic Acids Res.*, vol. 41, no. D1, pp. D524–D529, Jan. 2013, doi: 10.1093/nar/gks1169.

[29] G. E. Tusnady, Z. Dosztanyi, and I. Simon, "TMDET: Web server for detecting transmembrane regions of proteins by using their 3D coordinates," *Bioinformatics*, vol. 21, no. 7, pp. 1276–1277, Apr. 2005, doi: 10.1093/bioinformatics/bti121.

[30] Q. Zou, G. Lin, X. Jiang, X. Liu, and X. Zeng, "Sequence clustering in bioinformatics: An empirical study," *Briefings Bioinf.*, vol. 21, no. 1, pp. 1–10, Sep. 2018, doi: 10.1093/bib/bby090.

[31] J. Cheol Jeong, X. Lin, and X.-W. Chen, "On position-specific scoring matrix for protein function prediction," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 8, no. 2, pp. 308–315, Mar. 2011, doi: 10.1109/TCBB.2010.93.

[32] B. Zeng, P. Hönigschmid, and D. Frishman, "Residue co-evolution helps predict interaction sites in α-helical membrane proteins," *J. Struct. Biol.*, vol. 206, no. 2, pp. 156–169, May 2019, doi: 10.1016/j.jsb.2019.02.009.

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

[34] S. Altschul, "Gapped BLAST and PSI-BLAST: A new generation of protein database search programs," *Nucleic Acids Res.*, vol. 25, no. 17, pp. 3389–3402, Sep. 1997, doi: 10.1093/nar/25.17.3389.

[35] F. He, R. Wang, J. Li, L. Bao, D. Xu, and X. Zhao, "Large-scale prediction of protein ubiquitination sites using a multimodal deep architecture," *BMC Syst. Biol.*, vol. 12, no. S6, p. 109, Nov. 2018, doi: 10.1186/s12918-018-0628-0.

[36] H. Ding and D. Li, "Identification of mitochondrial proteins of malaria parasite using analysis of variance," *Amino Acids*, vol. 47, no. 2, pp. 329–333, Feb. 2015, doi: 10.1007/s00726-014-1862-4.

[37] Z. Lv, C. Ao, and Q. Zou, "Protein function prediction: From traditional classifier to deep learning," *Proteomics*, vol. 19, no. 14, Jul. 2019, Art. no. 1900119, doi: 10.1002/pmic.201900119.

[38] L. Peng, M. Peng, B. Liao, G. Huang, W. Li, and D. Xie, "The advances and challenges of deep learning application in biological big data processing," *Current Bioinf.*, vol. 13, no. 4, pp. 352–359, Jul. 2018, doi: 10.2174/1574893612666170707095707.

[39] L. Wei, R. Su, B. Wang, X. Li, Q. Zou, and X. Gao, "Integration of deep feature representations and handcrafted features to improve the prediction of N6-methyladenosine sites," *Neurocomputing*, vol. 324, pp. 3–9, Jan. 2019.

[40] J. X. Tan, S.-H. Li, Z.-M. Zhang, C.-X. Chen, W. Chen, H. Tang, and H. Lin, "Identification of hormone binding proteins based on machine learning methods," *Math. Biosciences Eng.*, vol. 16, no. 4, pp. 2466–2480, 2019, doi: 10.3934/mbe.2019123.

[41] H. Ding, W. Yang, H. Tang, P.-M. Feng, J. Huang, W. Chen, and H. Lin, "PHYPred: A tool for identifying bacteriophage enzymes and hydrolases," *Virologica Sinica*, vol. 31, no. 4, pp. 350–352, May 2016, doi: 10.1007/s12250-016-3740-6.

[42] W. Raveane, P. L. Galdámez, and M. A. González Arrieta, "Ear detection and localization with convolutional neural networks in natural images and videos," *Processes*, vol. 7, no. 7, p. 457, Jul. 2019, doi: 10.3390/pr7070457.

[43] P. Li and K. Mao, "Knowledge-oriented convolutional neural network for causal relation extraction from natural language texts," *Expert Syst. Appl.*, vol. 115, pp. 512–523, Jan. 2019, doi: 10.1016/j.eswa.2018.08.009.

[44] Z. Zhao, Y. Deng, Y. Zhang, Y. Zhang, X. Zhang, and L. Shao, "DeepFHR: Intelligent prediction of fetal acidemia using fetal heart rate signals based on convolutional neural network," *BMC Med. Inform. Decis. Making*, vol. 19, no. 1, p. 286, Dec. 2019, doi: 10.1186/s12911-019-1007-5.

[45] H. Bu, J. Hao, Y. Gan, S. Zhou, and J. Guan, "DEEPSEN: A convolutional neural network based method for super-enhancer prediction," *BMC Bioinf.*, vol. 20, no. S15, p. 598, Dec. 2019, doi: 10.1186/s12859-019-3180-z.

[46] P. Xuan, H. Cui, T. Shen, N. Sheng, and T. Zhang, "HeteroDualNet: A dual convolutional neural network with heterogeneous layers for drug-disease association prediction via Chou's five-step rule," *Frontiers Pharmacol.*, vol. 10, p. 1301, Nov. 2019, doi: 10.3389/fphar.2019.01301.

[47] O. S. Andersen and R. E. Koeppe, "Bilayer thickness and membrane protein function: An energetic perspective," *Annu. Rev. Biophys. Biomol. Struct.*, vol. 36, no. 1, pp. 107–130, Jun. 2007, doi: 10.1146/annurev.biophys.36.040306.132643.

CHANG LU was born in Changchun, Jilin, China, in 1988. She received the B.S. degree from the School of Information Science and Technology, Northeast Normal University, in 2011, where she is currently pursuing the Ph.D. degree in bioinformatics. Her research interests include ligand-protein interaction, membrane protein function annotation, and drug target prediction.

YINGLI GONG was born in Shandong, China, in 1998. She is currently pursuing the B.S. degree in computer science and technology with Northeast Normal University, Changchun, China. Her current research interests include data mining, machine learning, and bioinformatics.

ZHE LIU was born in Zhejiang, China, in 1997. She is currently pursuing the bachelor's degree with the School of Information Science and Technology, Northeast Normal University, China. She joined the Institution of Computational Biology, Northeast Normal University, in 2019, where she participated in some research on transmembrane protein structure prediction using deep learning methods.

YUANZHAO GUO was born in Jilin, China, in 1998. She is currently pursuing the B.S. degree with the School of Information Science and Technology, Northeast Normal University, China. Her research interests include bioinformatics and data mining.

ZHIQIANG MA received the Ph.D. degree from the School of Computer Science, Jilin University, in 2009. He is currently a Professor with the School of Information Science and Technology, Northeast Normal University. He is also the Vice President of the Research Association of Computer Education, Normal Universities of China, and the Executive Director of the Jilin Computer Federation. His interests include bioinformatics, software engineering, molecular biology, and data mining

HAN WANG is currently the Director of the Institution of Computational Biology, Northeast Normal University, China. Major in transmembrane protein structure and function researching using artificial intelligence methods, which extend to the researches, including big biological data, protein-protein interaction, new drug target discovering, and drug affection. He has published many peer-reviewed articles in these fields, funded by the National Natural Science Foundation of China, Jilin Scientific, and Technological Development Program of China.