EPiC Series in Computing

Volume 101, 2024, Pages 1–10

Proceedings of the 16th International Conference on Bioinformatics and Computational Biology (BICOB-2024)

TMHC-MSAT: Accurate Prediction of Inter-Helical Residue Contacts in Transmembrane Proteins Using MSA Transformer

Bander Almalki[1] and Li Liao[1]

[1]Department of Computer and Information Sciences, University of Delaware, Delaware, U.S.A.
alathwny@udel.edu, liliao@udel.edu

Abstract

Residue-residue contact prediction in a protein is one of the most used and informative middle steps to ultimately predict the complete 3D structure of a protein. While most previous studies use methods relying on statistical analysis of sequential properties to infer these contacts, some recent methods based on natural language processing models have gained success in accomplishing the task. However, most of these methods and models are built for globular proteins and not intended for specific types of proteins such as Transmembrane Proteins, which actually comprise about 30% of the proteome in most organisms and play important roles in cellular processes. In this study, we propose a Transmembrane Protein Helices Contacts predictor (TMHC-MSA) that utilizes features extracted by a protein language model called MSA Transformer and incorporates neighborhood information to enhance the quality of the produced contact map. Our proposed model shows that it can successfully outperform the state-of-the-art method by an average of 7% in terms of L precision and even surpass the MSA Transformer by an average of 2.5% on the same metric. Furthermore, we demonstrate that the more accurate contact map produced by our model can be used to generate a more accurate 3D structure.

1 Introduction

Membrane proteins constitute approximately a third of all proteins in the cell[1], and most of them are Transmembrane proteins. Transmembrane (TM) proteins are an integral type that spans the entire cell membrane and play a significant role in many biological processes such as facilitating the transport of molecules, ions, and information between a cell and its external environment[2]. In addition, they are involved in critical functions such as signal transduction, cell adhesion, and the maintenance of cell structure and integrity[3]. Alpha-helical transmembrane proteins make up a significant portion of the total integral membrane protein content. It's estimated that they account for approximately 20-30% of all protein-coding genes in humans[4]. Despite their abundance and importance, only a very small portion of them is determined experimentally due to difficulty in obtaining well-ordered crystals. Such

H. Al-Mubaid, T. Aldwairi and O. Eulenstein (eds.), BICOB-2024 (EPiC Series in Computing, vol. 101), pp. 1–10

difficulty might hinder the determination of their structures[5]. The need for developing computational tools to determine their structure, therefore, becomes essential. However, compared to globular proteins, computationally determining TM proteins' structure can be harder due to the limited availability of high-resolution X-ray structures to be used as templates for structural modeling and prediction. Inter-helical residue contact is one of the most successful computational approaches to reduce the TM protein fold search space and generate an acceptable 3D structure[6]. In this work, we use scores extracted from an unsupervised pre-trained language model, MSA Transformer[7], and incorporate information from the neighborhood of a residue pair to predict TM protein's inter-helical residue contact. To the best of our knowledge, this is the first time that scores extracted from a pre-trained language model are used to predict TM protein's inter helical residue contacts and, subsequently, their 3D structure. Results show that our model can produce a more accurate contact map than the current models and outperforms the state-of-the-art inter-helical TM protein residue contacts predictor. In addition, we show that the resulting contact map can be further used to generate a better 3D structure for helical TM proteins, which can help better understand their structure and functions.

2 Materials and Methods

2.1 Dataset

In this study, we adopt the dataset used in[8], which consists of 222 α-helical TM proteins extracted from the Protein Data Bank (PDB) with a resolution better than 3.5 Å and transmembrane helices ranging from 2 to 17 helices. Some proteins used in this dataset have since become obsolete in the new version of PDB (PDB 2023), and hence are excluded from this study. The total number of remaining proteins is 187 α-helical TM proteins divided into two sets TRAIN set (136 proteins) and TEST set (51 proteins).

Protein Contact Map is a 2D square matrix representation of a protein where each position in the matrix i, j can take a binary value of zero (no contact) and one (contact) [9]. It is considered an intermediate step in order to reduce the protein's folding search space and ultimately predict its 3D structure[6]. In this study, we follow the contact definition in [8] where a pair of residues are in contact if the distance between their heavy atoms is less than 5.5 Å and the sequence separation between these pair is not less than 5 positions. The contact map thus constructed from the PDB structures will be used as ground truth in training and evaluating the test results.

2.2 Multiple Sequence Alignment

For each single protein sequence in the dataset, Multiple Sequence Alignment (MSA) is extracted using the iterative searching tool Hhblits[10] against Uniprot20 E-value cutoff 0.001. While most studies aim to generate the largest possible number of multiple sequence alignments for their pipeline, which can be computationally expensive, only the top 128 hit MSAs are used in our pipeline due to resource limitations. Even with this limitation, as shown in the results section, our method outperforms both DeepHelicon and MSA Transformer. Also, the advantage of using a small number of MSAs could reduce the computational time considerably, especially for large-size proteins.

2.3 Extracting coevolutionary features

The fundamental concept of protein Co-evolutionary features is the utilization of statistical techniques capable of discerning direct connections between pairs of columns within a multiple sequence alignment and differentiating them from those pairs merely correlated. Co-evolutionary features have proved to be discriminative in predicting whether or not a pair of residues in a protein are

in contact. For example, according to[11], the Co-evolutionary feature is ranked first in predicting protein contacts compared to other less important features, such as amino acid composition and evolutionary conservation. Many statistical and machine learning approaches have been applied to measure the co-evolutionary score between a pair of residues. For example, Evfold[12] uses a global maximum entropy model to calculate the residue coupling score while PSICOV[13] uses sparse covariance matrix inversion. CCMpred[14], plmDCA[15], and GREMLIN[16] learn the direct couplings as parameters of a Markov random field by maximizing its pseudo-likelihood. SVM_{con} [17] and $TSVM_{ES}$ [18] use inductive and transductive support vector machines, subsequently, to predict protein contact map. However, recently, there has been evidence that using large language models and approaches, inherited from the natural language processing field, can assist in solving various bioinformatics-related problems. For example, in [19], [20], authors show that biological structure and function emerge from using a deep contextual language model with unsupervised learning on a large database of protein sequences. ESM[20] shows that some proteins' atomic-level structure started to appear by scaling their model up to 15 billion parameters. The MSA Transformer[7] applies an unsupervised language model with an axial attention mechanism over the rows and columns of the multiple sequence alignment and shows that the extracted features can be useful in some protein-related tasks.

2.4 MSA Transformer

The MSA Transformer is a large unsupervised pre-trained language model that has over 100M parameters. The model was trained on a large database containing 26 million multiple-sequence alignments[7]. It takes multiple sequence alignment as an input, uses axial attention across the rows and columns of the MSA.

$$\sum_{m=1}^{M} \frac{Q_m K_m^T}{\lambda(M,d)} \tag{1}$$

where Q and K are the query and key metrices subsequently and $\lambda(M, d)$ is a normalization function. It applies the unsupervised masked language modeling as an object for the training.

$$\mathcal{L}_{MLM}(x;\theta) = \sum_{(m,i)\in mask} \log\, p(x_{mi}|\tilde{\boldsymbol{x}};\theta) \tag{2}$$

Where x is the input MSA, $\tilde{\boldsymbol{x}}$ is the masked MSA, and θ represents the model parameters or weights that are being optimized during training. The model has proven its success in surpassing Potts model which can be used to infer the direct coupling analysis (co-evolutionary score) between two amino acid residues in a protein. MSA Transformer outputs an L*L scoring matrix, where L is the length of the protein. Every position in the matrix contains a score between 0 and 1. This value can be used to infer the coupling strength between residues i, j in the protein.

2.5 Extracting MSA transformer's Scores and TM protein's inter-helical residues

First, we extract the residues' coupling score generated by the MSA Transformer for each sequence in the dataset (TRAIN and TEST). The input to the MSA transformer model is a multiple sequence alignment of a protein and the output is a scoring matrix $M = R^{L*L}$. However, since MSA Transformer produces a score for any residue pair in a protein sequence, we are only interested in predicting contacts between residues located in the transmembrane domain.

Many tools have been proposed to annotate each residue in a protein's sequence as being inside or outside the membrane. While previous studies rely on TMHMM[4] to determine the inter-helical

residues of each TM protein, DeepTMHMM[21] has been proven to be more accurate and thus used in this study. We use the online version of DeepTMHMM (available at: https://dtu.biolib.com/DeepTMHMM) to annotate each residue as I (located inside the membrane) or O (located outside)

After the annotation, for each coupling's scoring matrix M, we create a binary matrix N of the same dimension as M where each position takes a value of 1(if the residue is located in the transmembrane region) or 0 (if not). Then, the Hadamard product $M \circ N$ is applied to extract the transmembrane residues' coupling scores of each TM protein in the dataset.

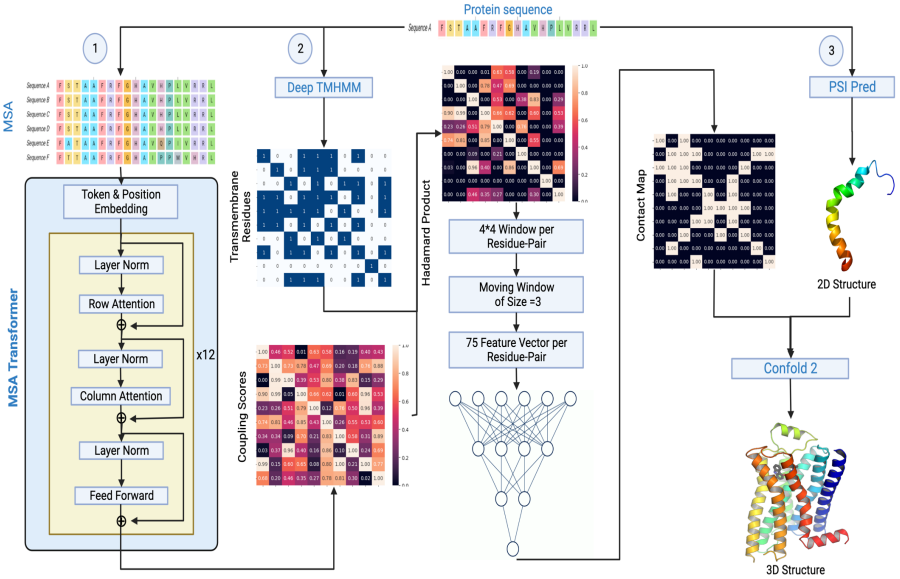

Figure 1. An overview of the TMHC-MSAT model. First, (1) the model extracts the coupling scores using the MSA transformer model. Then, (2) it uses Deep TMHMM to determine the transmembrane residues. The Hadamard product is applied to the coupling scores matrix and the transmembrane residues matrix to produce the coupling scores matrix of Transmembrane residues. Subsequently, a 4*4 window and a moving window of size 3 are applied to the output matrix resulting in a 75-feature vector per residue pair. A feed-forward neural network is then used to the predict contact probability of the residue pair. The resulting contact map can be used alongside the 2D structure produced by a model like PSI Pred (3) as an input to the Confold2 model to generate the predicted 3D structure of the target protein.

2.6 Classifier and computational pipeline

After extracting the TM protein's inter-helical coupling scores for each residue pair, we use a window of size $4 * 4$ to take into consideration the neighboring residues when predicting the contact between i, j residues. Incorporating information about residue neighbors can help in reducing false-positive predictions since true contacts often exhibit a pattern of correlated mutations that is consistent across their neighborhood, making them more reliable predictions[22]. In our pipeline, we apply the same window used in[8]. Specifically, for a residue pair at position (i, j) located in different helices, the following residues are considered $(\mathrm{i}+\mathrm{x},\mathrm{j}+\mathrm{y}),(\mathrm{i}+\mathrm{x},\mathrm{j}-\mathrm{y}),(\mathrm{i}-\mathrm{x},\mathrm{j}-\mathrm{y})$ *and* $(\mathrm{i}-\mathrm{x},\mathrm{j}+\mathrm{y})$ *where* $(\mathrm{x},\mathrm{y}) \in \{(0,0),(0,1),(0,3),(0,4),(1,0),(3,0),(3,4),(4,0),(4,3),(4,4)\}$. This results in a 25 features vector per residue pair. In addition, to capture more information, a moving window of size 3 over residue i is used $(i-1,j),(i,j)$ *and* $(i+1,j)$, which results in a vector of size 75 per residue pairs. We found experimentally that increasing the moving window beyond 3 or using an additional

moving window over j has a very minor impact on the performance and hence windows of size higher than 3 are not used here.

The aforementioned features are then used to train a fully connected feedforward neural network classifier in order to calculate the contact probability of any residue pairs in the TM protein's inter-helical domain. The network consists of an input layer, an output layer and six dense hidden layers. Relu activation is used in all layers except the output layer where sigmoid is being used. The total number of trainable parameters is 13807 parameters with 10 epochs. The batch size is 250 and Adam optimization is used as an optimizer. The input to the model is a vector of length 75 per residue-pair and the output is a value between 0 and 1 representing the contact probability between residue i and j.

The binary cross entropy loss is used to train our model.

$$Binary\ Cross\ Entropy\ Loss = -\frac{1}{N}\sum_{i=1}^{N} y_i \cdot \log\big(p(y_i)\big) + (1 - y_i) \cdot \log(1 - p(y_i)) \quad (3)$$

A depiction of the pipeline of our method is shown in Figure 1.

3 Results and Discussions

3.1 TMHC-MSAT can produce a more accurate contact map.

To measure the performance of our model in predicting the contact maps, we use the L/n precision score, which is a well-known metric used in the proteins' residues contact prediction field. To calculate the L/n score, we first rank the predicted contacts descending by probability and look at the first L/n contacts, where L refers to the length of the sequence. Then the top n contact probabilities are predicted as 1 (contact) and others as 0 (none-contact). In this study, we use L, L/2, L/5, and L/10 to measure the accuracy of our model and compare it with other methods. The results show that our model outperforms other state-of-art models in Transmembrane protein contact prediction in almost all L/n average precision, recall, and F1 score metrics and can deliver a more accurate contact map. Table.1 shows a comparison between TMHC-MSAT versus DeepHelicon and MSA Transformer models.

Model	L/n score	Precision	Recall	F1
DeepHelicon	L	0.6135	0.4618	0.5087
	L/2	0.7475	0.2865	0.4012
	L/5	0.8364	0.1329	0.2243
	L/10	0.8754	0.0722	0.1315
MSA Transformer	L	0.6586	0.5073	0.5526
	L/2	0.7849	0.3098	0.4296
	L/5	0.8613	0.1397	0.2351
	L/10	0.8992	**0.0758**	**0.1377**
TMHC-MSAT	L	**0.6830**	**0.5225**	**0.5713**
	L/2	**0.8127**	**0.3203**	**0.4442**
	L/5	**0.8915**	**0.1454**	**0.2442**
	L/10	**0.9076**	0.0753	0.1371

Table 1: Contact Map prediction accuracy of the MHC-MSAT Vs. DeepHelicon and MSA Transformer models. These results represent the average performance across all protein sequences in the test set.

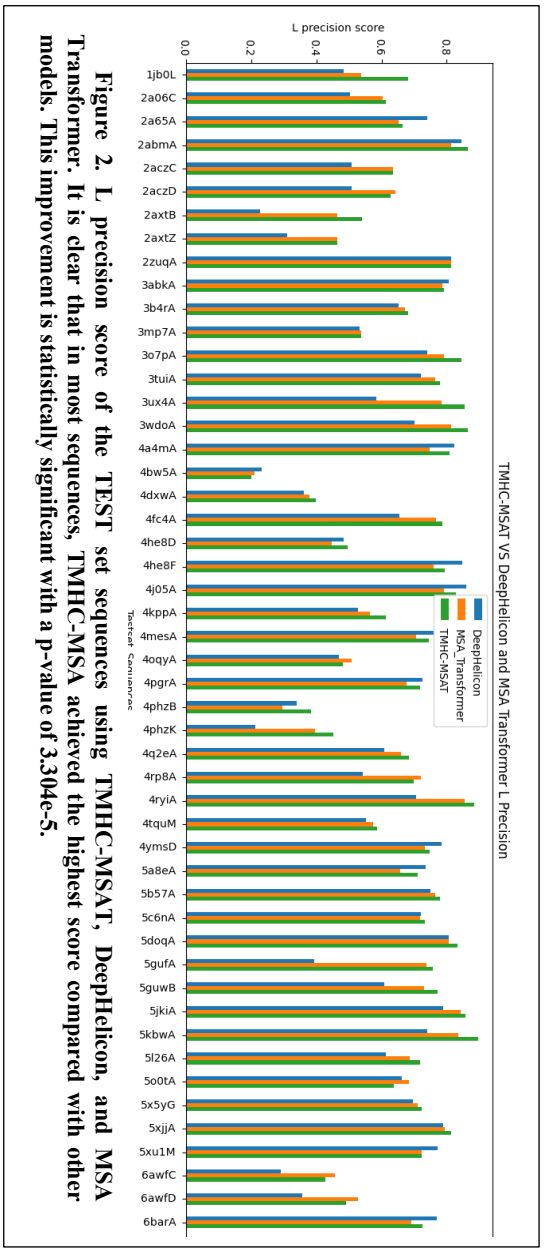

Figure 2. L precision score of the TEST set sequences using TMHC-MSAT, DeepHelicon, and MSA Transformer. It is clear that in most sequences, TMHC-MSA achieved the highest score compared with other models. This improvement is statistically significant with a p-value of 3.304e-5.

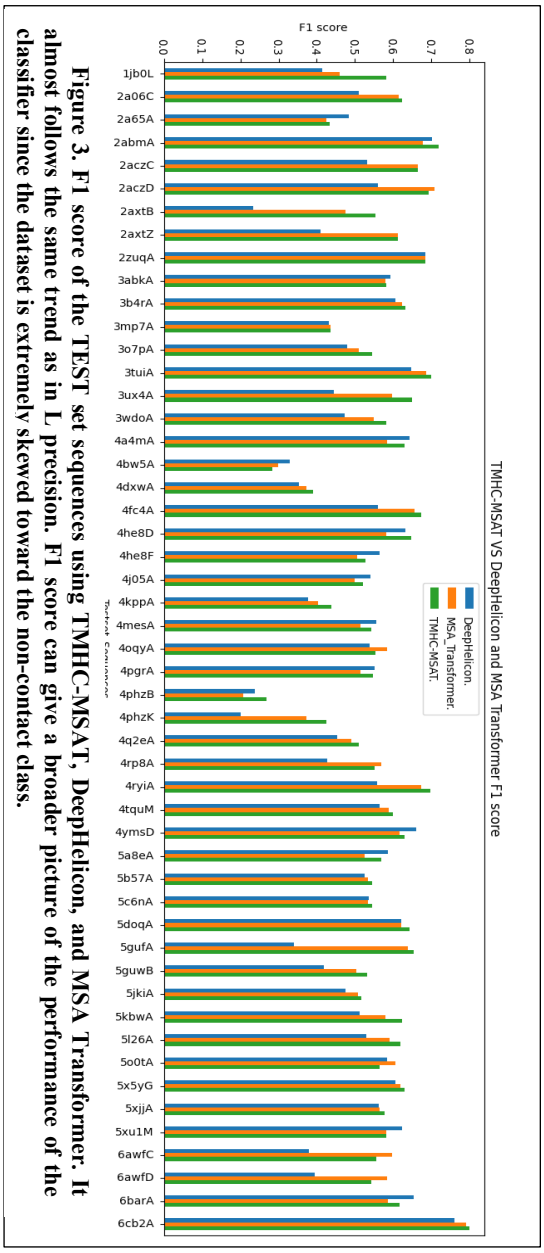

Figure 3. F1 score of the TEST set sequences using TMHC-MSAT, DeepHelicon, and MSA Transformer. It almost follows the same trend as in L precision. F1 score can give a broader picture of the performance of the classifier since the dataset is extremely skewed toward the non-contact class.

From Table.1 it is clear that our method TMHC-MSAT surpasses all other models including the state of art in the field DeepHelicon and even the MSA Transformer itself. In terms of L precision, TMHC-MSAT outperforms DeepHelicon by 6 to 7% and MSA Transformer by 2.5 % on average. In the L/10 score, THMC-MSAT lacks just a few fractions of a point behind the MSA Transformer in terms of recall and F1 score. Note that the results shown in table.1 are averaged over all the sequences in the test set, making the usual relationship between higher precision leading to higher recall inapplicable here. Regarding the number of features, DeepHelicon uses a total of 728 features vector per residue pair, which can result in a large feature space and consequently slow training time. On the

other side, TMHC-MSAT uses only 75 features which makes it considerably faster than DeepHelicon and other large models in the field. To investigate the performance of TMHC-MSAT in the residue pair contact prediction task, we compare our model with DeepHelicon and MSA Transformer on each sequence of the test set individually in term of L precision (Figure2.) and F1 scores (Figure3.) As it can be seen, in most sequences, our model outperforms DeepHelicon and does better or as good as MSA Transformer. To ensure that the improvement led by our model is statistically significant, we conducted the t-test on L precision and f1 scores. The p-value is $3.304e-5$ for L, $1.8252e-25$ for L/2, $3.538e-46$ for L/5, and $6.1853e-51$ for L/10. In a few sequences, DeepHelicon scores better than the other models. This might be a result of the deep residual neural network it uses, which is able to capture more distinctive features for those sequences. However, in general, TMHC-MSAT is still superior to DeepHelicon. For example, Figure4. shows the ROC curve of TMHC-MSAT vs. DeepHelicon. It is obvious that our model has a better true/false positive rate and thus a better ROC curve.

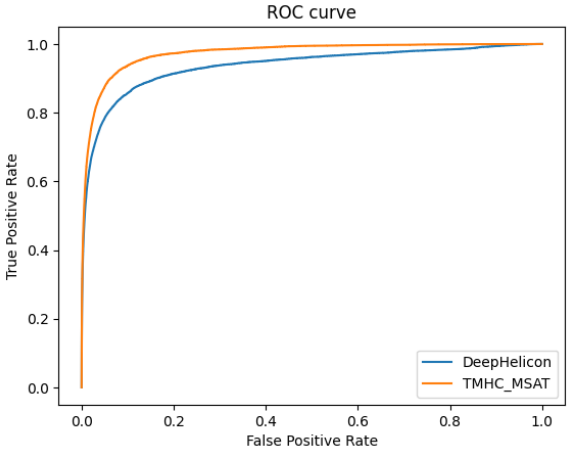

Figure 4. ROC curve of the TMHC-MSAT Vs. DeepHelicon Model.

3.2 TMHC-MSAT produced contact Map can deliver a better 3D structure.

To further evaluate the quality of the produced contact map by our model, we generate a 3D structure of a random protein (PDB: 4YRI) from the test set using the contact-map driven 3D structure tool CONFOLD2[23]. CONFOLD2 is a tool for building three-dimensional protein models using the predicted contact map and secondary structures. The output structure is then compared to the 3D structure of the same protein with a contact map produced by another model (here DeepHelicon). Using the RMSD score, we measure the accuracy of the two predicted structures by comparing them to the ground truth PDB structure.

$$RMSD = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(||X_i - Y_i||\right)^2}$$

Where N is the number of atoms or data points being compared. X_i and Y_i represent the coordinates of the i^{th} atom or data point in two structures being compared. ||.|| denotes the Euclidean distance between corresponding atoms or data points.

T. cruzi Histidyl-tRNA (PDB:4YRI) is chosen to compare the two 3D structures produced by two different contact maps, TMHC and DeepHelicon. "T. cruzi" refers to Trypanosoma cruzi, a parasitic protozoan that causes Chagas disease in humans[24]. Histidyl-tRNA synthetase in T. cruzi (TcHisRS) is an enzyme specific to this parasite and is essential for its survival[24]. Inhibiting or disrupting the function of this enzyme has been explored as a potential target for drug development against Chagas disease[24]. Figure5. Shows two 3D structures of the same protein (T. cruzi Histidyl-tRNA) produced using two different contact maps. The structure on the left is generated using the DeepHelicon contact map while the one on the right is generated using TMHC-MSAT. In Figure5. It is obvious that in the generated DeepHelicon 3D structure (Blue), one helix significantly deviates from the true PDB structure (Green) causing a high RMSD score of 3.029. This can be a direct effect of the poor contact map produced by DeepHelicon. On the other side, the more accurate contact map generated by TMHC-MSAT led to a better 3D structure with an RMSD score of 2.040. Notice that the two structures use the same PSI pred model [25] to acquire the 2D structure of the protein.

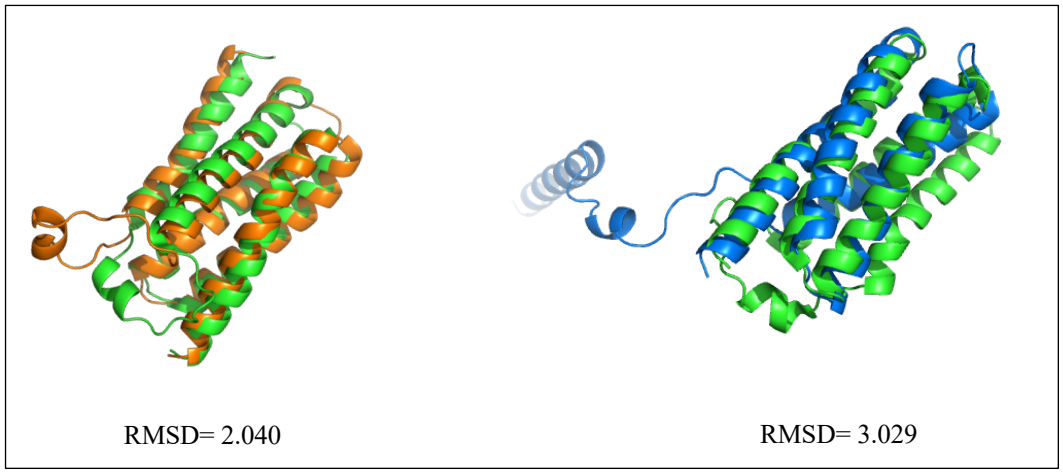

Figure 5. 4ryi Protein chain A. Comparison between the TMHC-MSAT predicted structure (Orange) and DeepHelicon predicted structure (Blue) vs the original PDB structure (Green)

4 Conclusion

In this study, we developed a deep neural network, TMHC-MSA, to predict residues' contacts in alpha-helical transmembrane proteins. The model uses novel features extracted from a protein language model trained on a large protein database, where unsupervised masking is used as an objective function. Besides the extracted features, TMHC-MSA incorporates information about residue neighborhood which can help in reducing false-positive predictions and enhance the quality of the produced contact map. The results from a widely adopted benchmark dataset show that our proposed model surpasses the state-of-art model in alpha-helical transmembrane protein contacts prediction significantly and is able to provide a better contact map that can be used to generate a high-quality 3D structure.

Acknowledgment

The authors would like to thank the National Science Foundation (NSF-MCB1820103) and Delaware Bioscience Center for Advanced Technology, which partly supported this research.

References

[1] J. G. Almeida, A. J. Preto, P. I. Koukos, A. M. J. J. Bonvin, and I. S. Moreira, "Membrane proteins structures: A review on computational modeling tools," *Biochimica et Biophysica Acta (BBA) - Biomembranes*, vol. 1859, no. 10, pp. 2021–2039, 2017, doi: https://doi.org/10.1016/j.bbamem.2017.07.008.

[2] W. Stillwell, "Chapter 6 - Membrane Proteins," in *An Introduction to Biological Membranes (Second Edition)*, W. Stillwell, Ed., Elsevier, 2016, pp. 89–110. doi: https://doi.org/10.1016/B978-0-444-63772-7.00006-3.

[3] L. E. Hedin, K. Illergård, and A. Elofsson, "An Introduction to Membrane Proteins," *J Proteome Res*, vol. 10, no. 8, pp. 3324–3331, Aug. 2011, doi: 10.1021/pr200145a.

[4] A. Krogh, B. Larsson, G. von Heijne, and E. L. L. Sonnhammer, "Predicting transmembrane protein topology with a hidden markov model: application to complete genomes11Edited by F. Cohen," *J Mol Biol*, vol. 305, no. 3, pp. 567–580, 2001, doi: https://doi.org/10.1006/jmbi.2000.4315.

[5] E. P. Carpenter, K. Beis, A. D. Cameron, and S. Iwata, "Overcoming the challenges of membrane protein crystallography," *Curr Opin Struct Biol*, vol. 18, no. 5, pp. 581–586, Oct. 2008, doi: 10.1016/j.sbi.2008.07.001.

[6] M. Torrisi, G. Pollastri, and Q. Le, "Deep learning methods in protein structure prediction," *Computational and Structural Biotechnology Journal*, vol. 18. Elsevier B.V., pp. 1301–1310, Jan. 01, 2020. doi: 10.1016/j.csbj.2019.12.011.

[7] R. Rao *et al.*, "MSA Transformer," 2021. [Online]. Available: https://github.com/facebookresearch/

[8] J. Sun and D. Frishman, "DeepHelicon: Accurate prediction of inter-helical residue contacts in transmembrane proteins by residual neural networks," *J Struct Biol*, vol. 212, no. 1, Oct. 2020, doi: 10.1016/j.jsb.2020.107574.

[9] M. Vendruscolo, E. Kussell, and E. Domany, "Recovery of protein structure from contact maps," *Fold Des*, vol. 2, no. 5, pp. 295–306, 1997, doi: https://doi.org/10.1016/S1359-0278(97)00041-2.

[10] M. Remmert, A. Biegert, A. Hauser, and J. Söding, "HHblits: Lightning-fast iterative protein sequence searching by HMM-HMM alignment," *Nat Methods*, vol. 9, no. 2, pp. 173–175, Feb. 2012, doi: 10.1038/nmeth.1818.

[11] K. Stahl, M. Schneider, and O. Brock, "EPSILON-CP: using deep learning to combine information from multiple sources for protein contact prediction," *BMC Bioinformatics*, vol. 18, no. 1, p. 303, 2017, doi: 10.1186/s12859-017-1713-x.

[12] D. S. Marks *et al.*, "Protein 3D structure computed from evolutionary sequence variation," *PLoS One*, vol. 6, no. 12, Dec. 2011, doi: 10.1371/journal.pone.0028766.

[13] D. T. Jones, D. W. A. Buchan, D. Cozzetto, and M. Pontil, "PSICOV: Precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments," *Bioinformatics*, vol. 28, no. 2, pp. 184–190, Jan. 2012, doi: 10.1093/bioinformatics/btr638.

[14] S. Seemayer, M. Gruber, and J. Söding, "CCMpred—fast and precise prediction of protein residue–residue contacts from correlated mutations," *Bioinformatics*, vol. 30, no. 21, pp. 3128–3130, Nov. 2014, doi: 10.1093/bioinformatics/btu500.

[15] M. Ekeberg, T. Hartonen, and E. Aurell, "Fast pseudolikelihood maximization for direct-coupling analysis of protein structure from many homologous amino-acid sequences," *J Comput Phys*, vol. 276, pp. 341–356, 2014, doi: https://doi.org/10.1016/j.jcp.2014.07.024.

[16] H. Kamisetty, S. Ovchinnikov, and D. Baker, "Assessing the utility of coevolution-based residue–residue contact predictions in a sequence- and structure-rich era," *Proceedings of the National Academy of Sciences*, vol. 110, no. 39, pp. 15674–15679, Sep. 2013, doi: 10.1073/pnas.1314045110.

[17] J. Cheng and P. Baldi, "Improved residue contact prediction using support vector machines and a large feature set," *BMC Bioinformatics*, vol. 8, no. 1, p. 113, 2007, doi: 10.1186/1471-2105-8-113.

[18] B. Almalki, A. Sawhney, and L. Liao, "Transmembrane Protein Inter-Helical Residue Contacts Prediction Using Transductive Support Vector Machines," pp. 35–22. doi: 10.29007/3ztg.

[19] A. Rives *et al.*, "Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences," *Proceedings of the National Academy of Sciences*, vol. 118, no. 15, p. e2016239118, Apr. 2021, doi: 10.1073/pnas.2016239118.

[20] Z. Lin *et al.*, "Evolutionary-scale prediction of atomic-level protein structure with a language model," *Science (1979)*, vol. 379, no. 6637, pp. 1123–1130, Mar. 2023, doi: 10.1126/science.ade2574.

[21] J. Hallgren *et al.*, "DeepTMHMM predicts alpha and beta transmembrane proteins using deep neural networks," *bioRxiv*, p. 2022.04.08.487609, Jan. 2022, doi: 10.1101/2022.04.08.487609.

[22] S. Ahmad and K. Mizuguchi, "Partner-Aware Prediction of Interacting Residues in Protein-Protein Complexes from Sequence Data," *PLoS One*, vol. 6, no. 12, p. e29104, Dec. 2011, doi: 10.1371/journal.pone.0029104.

[23] B. Adhikari and J. Cheng, "CONFOLD2: improved contact-driven ab initio protein structure modeling," *BMC Bioinformatics*, vol. 19, no. 1, p. 22, 2018, doi: 10.1186/s12859-018-2032-6.

[24] C. Y. Koh *et al.*, "A binding hotspot in Trypanosoma cruzi histidyl-tRNA synthetase revealed by fragmentbased crystallographic cocktail screens," *Acta Crystallogr D Biol Crystallogr*, vol. 71, pp. 1684–1698, Aug. 2015, doi: 10.1107/S1399004715007683.

[25] L. J. McGuffin, K. Bryson, and D. T. Jones, "The PSIPRED protein structure prediction server," *Bioinformatics*, vol. 16, no. 4, pp. 404–405, Apr. 2000, doi: 10.1093/bioinformatics/16.4.404.