doi: 10. 3969 / j. issn. 1674 – 1242. 2022. 04. 003

一种基于胶囊网络外膜蛋白拓扑结构预测方法

宋世龙，赵雨桐，王茜，王晗
（东北师范大学信息科学与技术学院计算生物研究所，吉林长春 130117）

【摘要】 **目的**　采用计算手段探索在当前外膜蛋白小样本的条件下提升外膜蛋白拓扑结构预测精度的深度学习方法。**方法**　首先，选取和构建适用于预测外膜蛋白拓扑结构的数据集；其次，经特征筛选和对比实验确定模型的最优输入；再次，构建和优化基于胶囊网络的拓扑结构预测模型 TopOMP-capsnet；最后，通过对比同类方法评估和验证模型性能。**结果和结论**　拓扑结构预测模型 TopOMP-capsnet 与同类方法相比，性能有所提升，证明深度学习技术能够在有限样本条件下识别相应序列模式，有助于外膜蛋白结构和功能的大规模分类及筛选。**创新之处**　拓扑结构预测模型 TopOMP-capsnet 的三态预测准确率（Q3）达到 87.7%，优于传统机器学习方法。

【关键词】 胶囊网络；外膜蛋白；拓扑结构；深度学习
【中图分类号】Q811.4　　【文献标志码】A　　文章编号：1674-1242(2022)04-0207-12

A Method for Predicting the Topology of Outer Membrane Proteins Based on Capsule Network

SONG Shilong, ZHAO Yutong, WANG Xi, WANG Han
(Institute of Computational Biology, College of Information Science and Technology, Northeast Normal University, Changchun, Jilin 130117, China)

【**Abstract**】 **Objective**　Employing computational means to explore an efficient deep learning method for improving the prediction accuracy of outer membrane protein topology under the current conditions of small sample size of outer membrane proteins. **Methods**　First, selecting and constructing data sets is suitable for the prediction of outer membrane protein topology; Second, this paper determining the optimal input of the model through feature screening and comparative experiments; Third, building and optimizing a topology prediction model the TopOMP-capsnet is on the basis of capsule network; Finally, the model performance is evaluated and verified by comparative congenic methods. **Results and conclusions**　Topology prediction model the TopOMP-capsnet has better performance compared with similar methods, which proves that deep learning technology can identify corresponding sequence patterns under limited sample conditions, and is helpful for large scale classification and screening of outer membrane protein structure and function. **Innovation**　Topology prediction model the TopOMP-capsnet has a three-state prediction accuracy (Q3) of 87.7%, which is superior to traditional machine learning methods.

【**Key words**】 Capsule Network; Outer Membrane Proteins; Topological Structure; Deep Learning

1　引言

外膜蛋白作为跨膜蛋白的一个重要类别，在生物细胞基本生理过程中扮演着举足轻重的角色[1-5]。外膜蛋白具有在外膜通道传输物质等的生物功能，在疫苗和抗生素等药物设计[6-8]研究中具有重要意义。基于“序列决定结构，结构决定功能”这一普遍共识，有关外膜蛋白结构的研究成为了解该

收稿日期：2022-11-22
作者简介：宋世龙，E-mail：slsong@nenu.edu.cn
通信作者：王晗，E-mail：wangh101@nenu.edu.cn

蛋白生物功能并应用于生物工程的主要途径。当前，生物测序技术难以大规模、低成本且有效地测定外膜蛋白结构，因此利用计算手段探索和优化蛋白结构预测方法成为蛋白功能研究的主要途径之一。深度学习在蛋白结构研究中不断地被尝试并表现优异[9-16]，如 Fang 等[17]提出了一种结合氨基酸理化性质、HHblits 蛋白质图谱和蛋白序列位置特异性矩阵 PSSM 等特征的深度神经网络架构 Deep3I 预测蛋白二级结构；Madeo 等[18]提出了一种结合语法约束隐性条件随机场（GRHCRFs）和长短期记忆网络模型（LSTM）的 BetAware-Deep 方法预测外膜蛋白的拓扑结构，引入深度学习方法有望在外膜蛋白拓扑结构预测工作中取得新进展。AlphaFold 2[19, 20]方法应用于大规模蛋白结构预测，确定了几乎覆盖整个人类蛋白质组的蛋白质结构。AlphaFold 2 等方法具备很高的预测水平，但是没有针对跨膜蛋白做出调整[21]，尚未有研究表明这些蛋白结构预测方法在跨膜蛋白上与水溶蛋白达到了相同的预测水平，跨膜蛋白的结构预测仍需继续探索。本文利用胶囊向量在小样本学习任务中的优势构建的外膜蛋白拓扑结构预测模型 TopOMP-capsnet，在交叉验证和对比实验中的识别效果良好，其中三态预测准确率（Q3）达到 87.7%。

2　关键研究要素分析

2.1　外膜蛋白结构特征分析

外膜蛋白是广泛存在于生物细胞和细胞器外膜上的蛋白质。在特殊的生存环境的作用下，外膜蛋白的蛋白结构会产生特异性变化。因此，外膜蛋白与水溶蛋白在理化性质和残基分布方面均有显著区别，与 α 螺旋跨膜蛋白在结构和功能上也存在差异。外膜蛋白跨膜区以 β 链的形式构成桶状的跨膜通道，组成外膜蛋白桶状结构的 β 链以反向平行的方式在其相邻两条链之间首尾相接。通常情况下，细胞周质一侧的 β 链末端的氨基酸连接相邻两条 β 链的氨基酸形成短的卷曲结构，而细胞膜外一侧的 β 链末端形成较长的卷曲结构连接相邻两条 β 链。这些卷曲结构分布在膜外，流动性大且易变性较强。外膜蛋白在空间结构特征方面具有特异性，可以由蛋白序列模式表示。

2.2　结构样本资源

在难以获得相关研究所需的高分辨率晶体结构的背景下，当前条件下的膜蛋白数据库规模十分有限。目前，被广泛使用的膜蛋白数据库有 Mptopo（Membrane protein topology database）、OPM（Orientations of Proteins in Membranes Database）、PDBTM（Protein Data Bank of Transmembrane Proteins）、OMPdb（a database of β-barrel outer membrane proteins from Gram-negative bacteria）及 TOPDB（Topology Data Bank of Transmembrane Proteins）。上述数据库总结见表 1。通过对现有各膜蛋白数据库信息的对比分析，最终选定数据量更丰富、结构数据更清晰和最准确的 PDBTM 数据库中的数据作为本文实验的数据集。

表 1　膜蛋白数据库总结

Tab. 1　Summary of protein data bank

数据库名称	数据库地址
OPM[22]	https://opm.phar.umich.edu/
PDBTM[23]	http://pdbtm.enzim.hu/
Mptopo[24]	https://blanco.biomol.uci.edu/mptopo/
OMPdb[25]	http://bioinformatics.biol.uoa.gr/OMPdb
TOPDB[26]	http://topdb.enzim.hu

3　实验方案及数据处理

3.1　实验方案

跨膜蛋白的拓扑结构是该类膜蛋白在结构上的高度近似描述。本实验的目标是对任给一条外膜蛋白序列中的跨膜、非跨膜片段做相应预测。具体内容是由被最终确定的模型预测出给定的序列上每一个残基的标签都为“I”（膜内），或“M”（跨膜），或“O”（膜外）。模型预测流程示意图如图 1 所示。模型预测得到的结果表示成拓扑序列字符串形式。

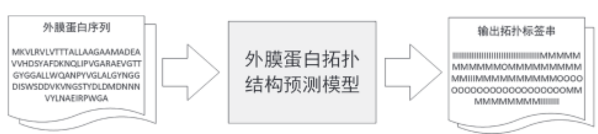

图 1　模型预测流程示意图

Fig.1　Schematic of model prediction flow

3.2　数据集的构建

3.2.1　训练数据集

通过前述对比选取 PDBTM 数据库中所有外膜蛋白数据作为本实验的数据来源。对于训练数据集，选取 PRED-TMBB$_{HNN}$[27]训练数据集的 49 个蛋白作为模型训练和验证的数据集。选择与文献［27］

同样的方式去除结构不规范、序列过短和冗余的数据，并根据基于 Pfam 分类[28]的 OMPdb 数据库[25]中的家族分类将实验数据进行交叉验证划分，划分结果见表 2。其中，对于冗余数据采取 Hobohm 等提出的第二种算法[29]并使用 BLAST 工具[30]剔除，删去序列相似性超过 30% 的数据。为避免过短的数据包含的信息量不足所带来的消极影响，删去序列长度低于 80 的数据。

表 2　实验训练数据集数据

Tab. 2　Experimental Training Dataset Data

交叉验证 ID	PDB_ID CHAIN	UNIPROT_ AC	OMPdb 数据库中所属家族分类	跨膜片段数
1	4fuv_A	G2JM62	The Carbapenem resistance-associated outer membrane protein (carO) Family	8
2	3gp6_A	P37001	The Antimicrobial peptide resistance and lipid A acylation protein (PagP) Family	8
3	2erv_A	Q9HVD1	The Outer membrane-localized lipid A 3-O-deacylase (PagL) Family	8
4	2x27_X	Q9HWW1	The OmpW Family	8
5	3dzm_A	Q72JD8	The Thermus thermophilus HB27 TtoA Family	8
6	1qj8_A	P0A917	The Outer Membrane Protein beta-barrel domain Family	8
6	2lhf_A	Q51486	The Outer Membrane Protein beta-barrel domain Family	8
7	1qjp_A	P0A910	The OmpA Family	8
8	1p4t_A	Q9RP17	The Neisserial Surface Protein A (NspA) Family	8
9	1i78_A	P09169	The Omptin (OmpT) Family	10
10	1k24_A	Q51227	The Opacity (OpcA) Family	10
11	4e1s_A	P43261	The Intimin/Invasin Famil	12
12	1qd5_A	P0A921	The Outer Membrane Phospholipase (OMPLA) Family	12
13	3fid_A	Q8ZPT3	The systemic factor protein A (SfpA/LpxR) Family	12
14	1tly_A	P0A927	The Nucleoside-specific Channel-forming Outer Mem-brane Porin (Tsx) Family	12
15	2wjr_A	P69856	The Oligogalacturonate-specific Porin (KdgM) Family	12
15	|4fqe_A	Q934G3	The Oligogalacturonate-specific Porin (KdgM) Family	12
16	1uyn_X	Q8GKS5	The Autotransporter (AT) Family	12
16	3kvn_X	O33407	The Autotransporter (AT) Family	12
16	3qq2_A	Q45340	The Autotransporter (AT) Family	12
16	3slt_A	Q7BSW5	The Autotransporter (AT) Family	12
17	2f1c_X	P76045	The OmpG Porin (OmpG) Family	14
18	3dwo_X	Q9HVJ6	The FadL Outer Membrane Protein (FadL) Family	14
18	3bry_A	Q9RBW8	The FadL Outer Membrane Protein (FadL) Family	14
18	3pgu_A	P10384	The FadL Outer Membrane Protein (FadL) Family	14
19	2o4v_A	P05695	The Pseudomonas OprP Porin (POP) Family	16
20	3wi4_A	P30690	The General Bacterial Porin (GBP-1) Family 1	16
20	1osm_A	Q48473	The General Bacterial Porin (GBP-1) Family 1	16
21	2por_A	P31243	The General Bacterial Porin (GBP-4) Family 4	16
21	2fgq_X	P24305	The General Bacterial Porin (GBP-4) Family 4	16
21	1prn_A	P39767	The General Bacterial Porin (GBP-4) Family 4	16
22	2qdz_A	P35077	The Two-Partner Secretion (TPS) Family	16
23	4gey_A	A5VZA8	The Glucose-selective OprB Porin (OprB) Family	16
24	4c00_A	P0ADE4	The Outer Membrane Protein Insertion Porin (Omp-IP/Omp85) Family	16
24	4k3c_A	Q93PM2	The Outer Membrane Protein Insertion Porin (Omp-IP/Omp85) Family	16
25	4afk_A	P18895	The Alginate Export Porin (algE) Family	18
26	2ynk_A	Q8GNN6	The wzi Family	18
27	1a0s_P	P22340	The Sugar Porin (SP) Family	18
27	1mpr_A	P26466	The Sugar Porin (SP) Family	18
28	3syb_A	Q9HVS0	The Outer Membrane Porin (OprD) Family	18
29	2iah_A	P48632	The Outer Membrane Receptor (OMR-TonB Dependent Receptor) Family	22
29	3fhh_A	P72412	The Outer Membrane Receptor (OMR-TonB Dependent Receptor) Family	22
29	4b7o_A	Q841A2	The Outer Membrane Receptor (OMR-TonB Dependent Receptor) Family	22
29	3v8x_A	Q9K0U9	The Outer Membrane Receptor (OMR-TonB Dependent Receptor) Family	22
29	1fep_A	P05825	The Outer Membrane Receptor (OMR-TonB Dependent Receptor) Family	22
29	2grx_A	P06971	The Outer Membrane Receptor (OMR-TonB Dependent Receptor) Family	22
29	1kmo_A	P13036	The Outer Membrane Receptor (OMR-TonB Dependent Receptor) Family	22
29	2guf_A	P06129	The Outer Membrane Receptor (OMR-TonB Dependent Receptor) Family	22
30	3rfz_B	P30130	The Outer Membrane Fimbrial Usher Porin (FUP) Family	24

3.2.2　基准数据集

为了更好地将最终确定的模型性能同其他预测方法进行对比，选择 Hayat 等提出的方法 BOCTOPUS2[31] 中包含 42 条外膜蛋白序列在内的训练集作为本实验的基准测试集，将完整蛋白序列数据整理为表 3。

表 3　BOCTOPUS2 基准测试集

Tab. 3　BOCTOPUS2 benchmark test set

boctopus2 数据集 (42)					
4Q35_A	3FID_A	2X4M_A	3SLJ_A	3CSL_A	3DZM_A
1TLY_A	3RBH_A	4GEY_A	1UYN_X	2IAH_A	2MLH_A
2MPR_A	1QD5_A	2O4V_A	3KVN_A	3DWO_X	1K24_A
3EMN_X	3SYB_A	3PRN_A	4MEE_A	4AIP_A	3QRA_A
3RFZ_B	4C00_A	3POX_A	2X9K_A	3V89_A	2K0L_A
2YNK_A	3NJT_A	3A2S_A	1FEP_A	4FUV_A	2LHF_A
3GP6_A	4K3B_A	4E1T_A	1KMO_A	2ERV_A	2X27_X

3.2.3　独立测试集

独立测试集选取 2021 年 12 月 10 日版本 PDBTM 数据库中全部 477 个外膜蛋白。经 psi-cd-hit-protein 工具[32] 去冗余、去重和去除过短序列处理得到 7 条外膜蛋白序列，见表 4。

表 4　实验独立测试集

Tab. 4　Experimental independent test set

PDB_ID CHAIN	UNIPROT_AC	跨膜片段数量 / 个
4Y25_A	P69434	16
5O67_A	C4IN73	12
5LDV_A	Q659I5	18
6QAM_A	Q00595	8
5DL8_B	A0A0D5YI36	18
5MDO_A	L0RVU0	16
6H3I_F	Q5I6C7	14

3.3　外膜蛋白先验序列特征

3.3.1　氨基酸理化属性特征

跨膜蛋白内的跨膜片段具有特殊的理化属性，如极性和疏水性。拓扑结构预测的前期研究中，运用跨膜蛋白的疏水性及“正电荷居内”法则有效地使 α 螺旋跨膜蛋白的拓扑结构预测性能得到较大限度的提升。参照跨膜蛋白结构与理化性质，实验中使用表 5 中所列举的 30 种从 AAindex[33] 数据库中筛选出来的氨基酸理化属性组合（phys30）和包括极性、疏水性、极化率、电荷性、二级结构、溶剂可及性和范德华规范化体积在内的 7 维理化性质（phys7）作为序列特征对预测模型进行了训练，经训练发现该特征对模型预测精度提升效果不够理想。

表 5　30 种从 AAindex 数据库中筛选的理化属性

Tab. 5　The physicochemical attributes of 30 variants screened from the AAindex database

AAindex	理化属性	属性范围
BULH740101	将自由能转移到表面	[−2.46,0.16]
BULH740102	表观部分比体积	[0.558,0.842]
PONP800102	周围疏水性的平均增益	[5.72,10.93]
PONP800104	α – 螺旋中的周围疏水性	[10.98,15.36]
PONP800105	β – 折叠中的周围疏水性	[11.79,16.49]
PONP800106	周围疏水性	[9.93,15]
MANP780101	平均周围疏水性	[10.85,15.71]
EISD840101	共识归一化疏水性尺度	[−1.76,0.73]
JOND750101	疏水性	[0,3.77]
HOPT810101	亲水值	[−3.4,3]
PARJ860101	高效液相色谱参数	[−10,10]
JANJ780101	平均可接触表面积	[15.5,103]
PONP800107	无障碍减少率	[1.79,7.69]
CHOC760102	折叠蛋白中残留的可及表面积	[18,97]
ROSG850101	平均埋置面积	[62.9,224.6]
ROSG850102	平均分数面积损失	[0.52,0.91]
BHAR880101	平均柔韧性指数	[0.295,0.544]
KARP850101	无刚性邻域的柔度参数	[0.925,1.169]
KARP850102	一个刚性邻域的柔度参数	[0.862,1.089]
KARP850103	两刚性邻域的柔度参数	[0.803,1.057]
JANJ780102	掩埋残留物的百分比	[3,74]
JANJ780103	暴露残留物的百分比	[5,85]
LEVM780101	α – 螺旋的归一化频率，带权重	[0.52,1.47]
LEVM780102	β – 折叠的归一化频率，带权重	[0.64,1.49]
LEVM780103	反向转弯的标准化频率，带权重	[0.41,1.91]
GRAR740102	极性	[4.9,13]
GRAR740103	体积	[3,170]
MCMT640101	折射率	[0,42.53]
PONP800108	周围残基的平均数	[4.88,7.86]
KYTJ820101	亲水指数	[−4.5,4.5]

3.3.2　氨基酸进化保守性特征

氨基酸进化保守性是蛋白质结构预测研究中的重要特征之一，通常由序列比对计算得到。当前存在多种搜索同源性序列比对的方法，如 jackhmmer[34]、PSI-BLAST[30]、HHblits[35] 等。本文分别选择 PSI-BLAST 和 HHblits 两种方法得

到的结果作为氨基酸进化保守性特征。

位置特异性打分矩阵 PSSM[30] 是蛋白质和生物信息学中的一个重要统计量，通过衡量不同氨基酸在蛋白质上某个特定序列位置上出现的概率来表示这个位置的氨基酸进化保守性。针对获取的 PSSM 矩阵中的特征产生的范围差异性问题，采用 sigmoid 函数（式 1）进行归一化处理。

$$f(x)=\frac{1}{1+e^{-x}} \tag{1}$$

Remmert 等提出了一种能够提高蛋白质序列比对分析效率的序列比对工具[35] HHblits。针对差异性问题，同 PSSM 矩阵一样对特征矩阵进行归一化处理。

3.3.3　蛋白质结构信息特征

本文的蛋白质结构特征选自 Zhou 等[36] 提出的方法 Frag1D 的预测结果。针对任意给出的一个蛋白质序列，Frag1D 都能够得到三态二级结构以及三态和八态形状串。截取部分结果得到的片段见表 6，由于 Frag1D 结果本身经预测得到等原因，该特征对模型预测精度提升效果不够理想。

表 6　Frag1D 预测结果片段

Tab. 6　Frag1D prediction result fragments

序号	氨基酸	三态二级结构	可信度	八态形状字符串	预测可信度	三态形状字符串	预测可信度
1	M	R	0.874	A	0.333	H	0.726
2	H	R	0.695	R	0.500	S	0.582
3	E	R	0.766	S	0.556	S	0.726
4	T	R	0.731	A	0.333	H	0.726
5	K	R	0.769	A	0.667	H	0.726
6	Q	R	0.737	A	0.579	H	0.741
7	G	R	0.856	G	0.348	T	0.676
8	G	R	0.911	K	0.296	H	0.503
9	E	R	0.661	A	0.437	H	0.582
10	K	R	0.701	A	0.645	H	0.707
11	R	R	0.623	A	0.323	H	0.652
12	F	R	0.562	S	0.344	S	0.744
13	T	R	0.636	S	0.344	S	0.717
14	G	R	0.537	S	0.344	S	0.448
15	A	H	0.587	R	0.625	S	0.717
16	I	S	0.611	S	0.625	S	0.690
17	C	H	0.636	A	0.613	H	0.707
18	R	H	0.611	A	0.677	H	0.735
19	C	H	0.555	A	0.514	H	0.693
20	S	R	0.589	A	0.472	H	0.654
21	H	R	0.600	A	0.343	H	0.742
22	R	R	0.719	A	0.455	H	0.595
23	Y	R	0.623	S	0.406	S	0.636
24	N	R	0.573	A	0.500	H	0.668
25	S	H	0.573	A	0.300	H	0.640
26	M	H	0.521	A	0.462	H	0.582
27	E	R	0.576	A	0.476	H	0.603
28	V	R	0.593	A	0.722	H	0.773
29	K	R	0.569	A	0.467	H	0.611
30	M	R	0.611	A	0.750	H	0.869
31	A	R	0.537	A	0.444	H	0.726
32	A	R	0.583	K	0.667	H	0.726
……	……	……	……	……	……	……	……

3.4　胶囊网络拓扑结构预测模型

3.4.1　胶囊网络预测模型构建

本文提出了一种基于胶囊网络的外膜蛋白拓扑预测模型 TopOMP-capsnet，该模型整体结构见图 2。该模型由 3 个一维卷积层（ConvA，ConvB 和 PrimaryCaps）与一个完全连接层（BetaCaps）构成。模型的输入为经 PSI-Blast 处理的 PSSM 特征矩阵和由蛋白序列得到的 HHblits 图谱特征，把上述特征分别输入 ConvA、ConvB 层。这两个卷积层将特征从初始表示转换处理成为中间级特征，

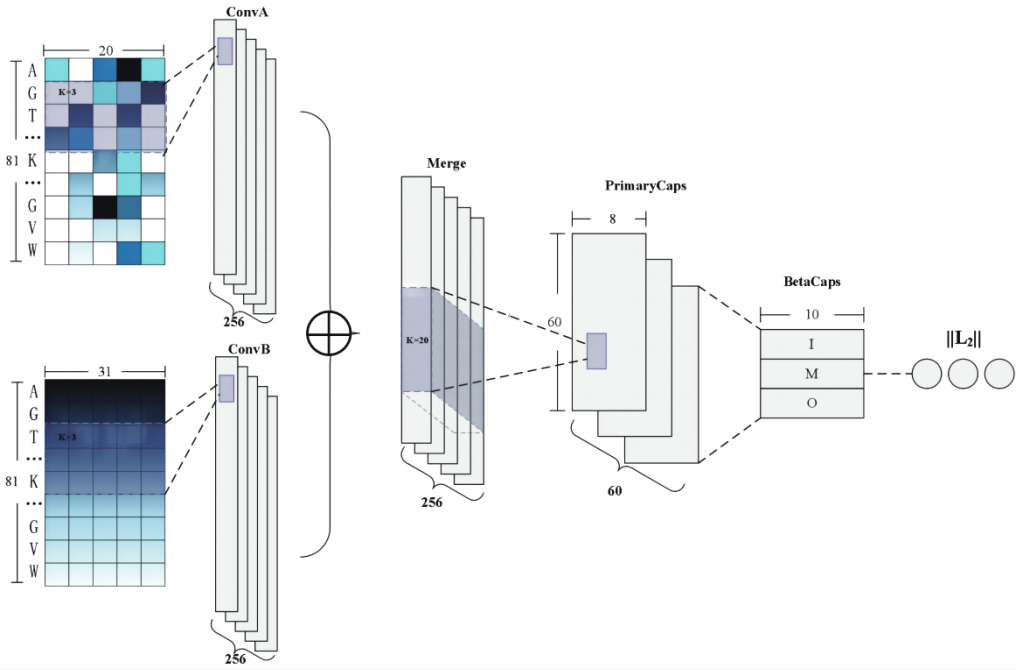

图 2　网络结构示意图

Fig.2　Schematic of network structure

再输入 PrimaryCaps，BetaCaps 层对特征做下一步的提取工作。

ConvA、ConvB 层具有 256 个一维卷积内核，每个卷积内核大小都为 3，步长设置为 1，激活函数为 ReLU 函数。ConvA、ConvB 层的一维卷积内核作为“特征过滤器”汇总氨基酸特征图谱的特征，使之成为单个标量特征，特征融合后输入 PrimaryCaps 层。

PrimaryCaps 层是一个包含 60 个卷积胶囊通道的一维卷积胶囊层。该层内胶囊均含有 8 个卷积单元，每个卷积单元都是大小为 20 的一维卷积核的结果，步长设置为 1。ConvA、ConvB 层的第一个有效维数是 79；而 Merge、PrimaryCaps 层的第一个有效维度分别为 79 和 60。挤压函数（式 2）把胶囊长度放缩至区间［0,1］。

$$\boldsymbol{V}_j=\frac{\|\boldsymbol{S}_j\|^2}{1+\|\boldsymbol{S}_j\|^2}\frac{\boldsymbol{S}_j}{\|\boldsymbol{S}_j\|} \tag{2}$$

式中，$\boldsymbol{S}_j$ 是胶囊 j 的输入，$\boldsymbol{V}_j$ 是矢量输出。另外，压缩激活功能应用于 BetaCaps 层中的胶囊。

BetaCaps 层用来接收从 PrimaryCaps 层中全部胶囊输出的输入。图 3 展示了 PrimaryCaps 层和 BetaCaps 层之间的计算过程。BetaCaps 层包含 3 个胶囊（$\boldsymbol{V}_j$，$j\in$［1,2,3］），其中 $\boldsymbol{V}_j$ 为 10 维向量。经由 PrimaryCaps 层再通过挤压函数（式 3）获取所有输出 $\hat{\boldsymbol{\mu}}_{j|i}$ 上的加权和 $\boldsymbol{S}_j$。

$$\boldsymbol{S}_j=\sum_i c_{i,j}\hat{\boldsymbol{\mu}}_{j|i},\qquad \hat{\boldsymbol{u}}_{j|i}=\boldsymbol{W}_{i,j}\boldsymbol{\mu}_i \tag{3}$$

式中，$c_{i,j}$ 为耦合系数，在 BetaCaps 层的 3 个胶囊求和结果为 $c_{i,1}+c_{i,2}+c_{i,3}=1$；$\boldsymbol{\mu}_i,i\in$［1,3600］为 PrimaryCaps 层的 8 维封装；$\boldsymbol{W}_{i,j}$ 为进行仿射变换的权重矩阵。迭代过程由动态路由[37]确定。

BetaCaps 中的 3 个胶囊的长度值代表预测拓扑结构的概率。分别对每个胶囊载体的 L2 标准正则化进行计算，再计算损失函数。损失函数见式（4）：

$$\boldsymbol{L}_k=\boldsymbol{T}_k max(0,m^+-\|\boldsymbol{v}_k\|)^2+\lambda(1-\boldsymbol{T}_k)max(0,\|\boldsymbol{v}_k\|-m^-)^2 \tag{4}$$

式中，$\boldsymbol{T}_k$=1，表示属于其中某一种分类。其他参数按照动态路由[37]设置，λ=0.5，m^+=0.9，m^-=0.1。

模型预测期间，首先经计算获得 3 个胶囊 I、M、O 的长度，再将其中长度值最大者的标签分配给所预测的残基，从而得到预测的拓扑序列字符串。

3.4.2　模型训练

预测模型在每个实验中均使用完全一致的训练策略进行训练。为科学分析和检验模型的拟合效果，模型通过 30 折交叉验证得到相应验证结果见图 4。

TopOMP-capsnet 模型架构由开源的 Keras 和 Tensorflow 实现，网络初始权重参数选定为 Keras

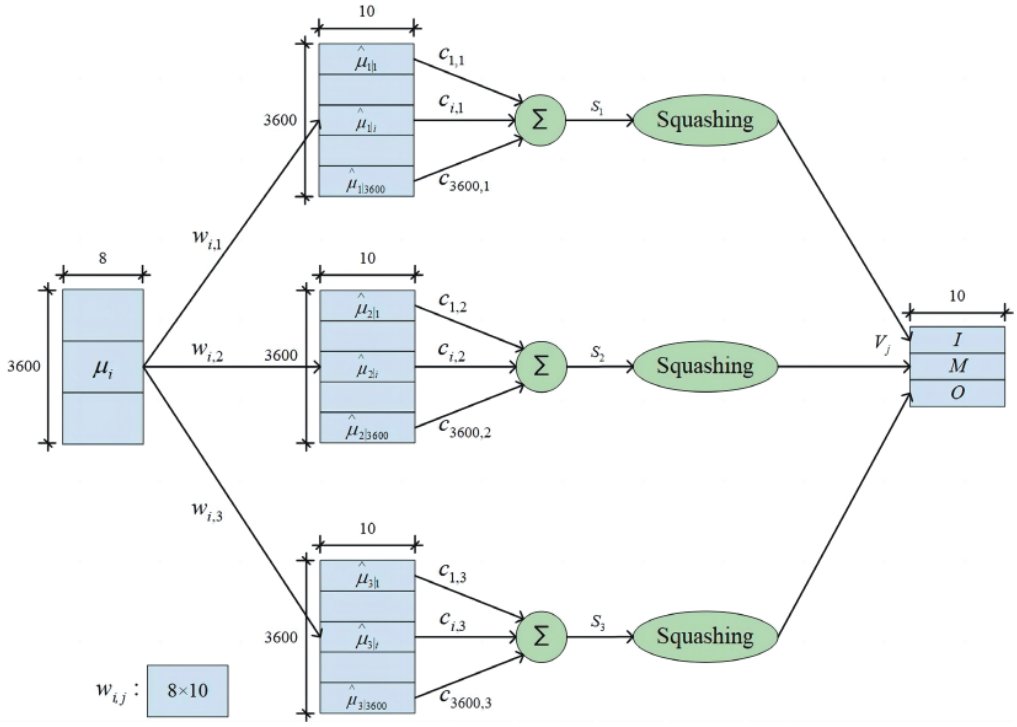

图 3　PrimaryCaps 层和 BetaCaps 层之间的计算过程

Fig.3　Calculation between PrimaryCaps and BetaCaps

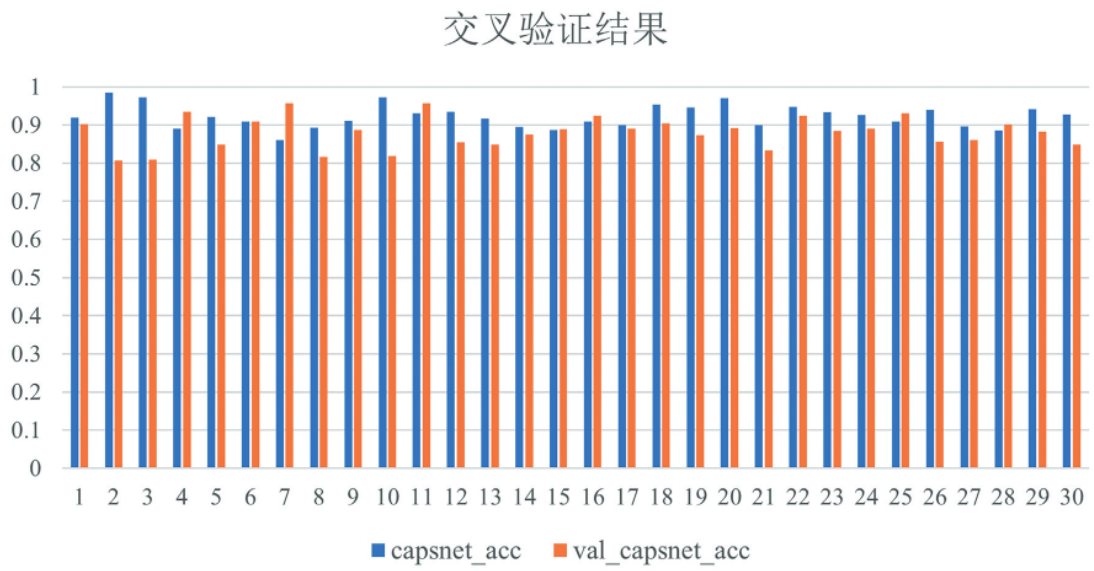

图 4　30 折交叉验证结果

Fig.4　Thirty-fold cross validation results

中的默认值。实验选择前述原始动态路由流程对胶囊网络进行训练，ConvA、ConvB 层的参数 Dropout 值设置为 0.7，PrimaryCaps 层的 Dropout 值设置为 0.2。在训练迭代期间，任意一次送入网络中的 BatchSize 都被设置为 50。同时，在训练预测模型期间采取 Adam 随机优化方法[38]和早期停止策略，相关参数值如学习率被设置为 0.001，第一时刻估计的指数衰减率为 0.9，第二时刻估计

的指数衰减率为 0.999。

3.4.3　模型预测流程

模型预测流程图如图 5 所示，实验步骤简单介绍为以下 4 点内容。

（1）从膜蛋白数据库中获取所需的序列信息并进行数据预处理工作。参照蛋白家族分类，对获取得到的数据集进行交叉验证分组。

（2）将得到的序列进行特征化编码，采用相同尺寸进行滑动窗口处理，并选取模型的最优输入。

（3）建立外膜蛋白拓扑结构预测模型，并对预测模型中的超参数、网络结构进行调整和优化。在训练集上使用交叉验证确认最终模型。

（4）模型 TopOMP-capsnet 对蛋白数据通过预测得到拓扑结构标签，最后对模型性能进行评价。

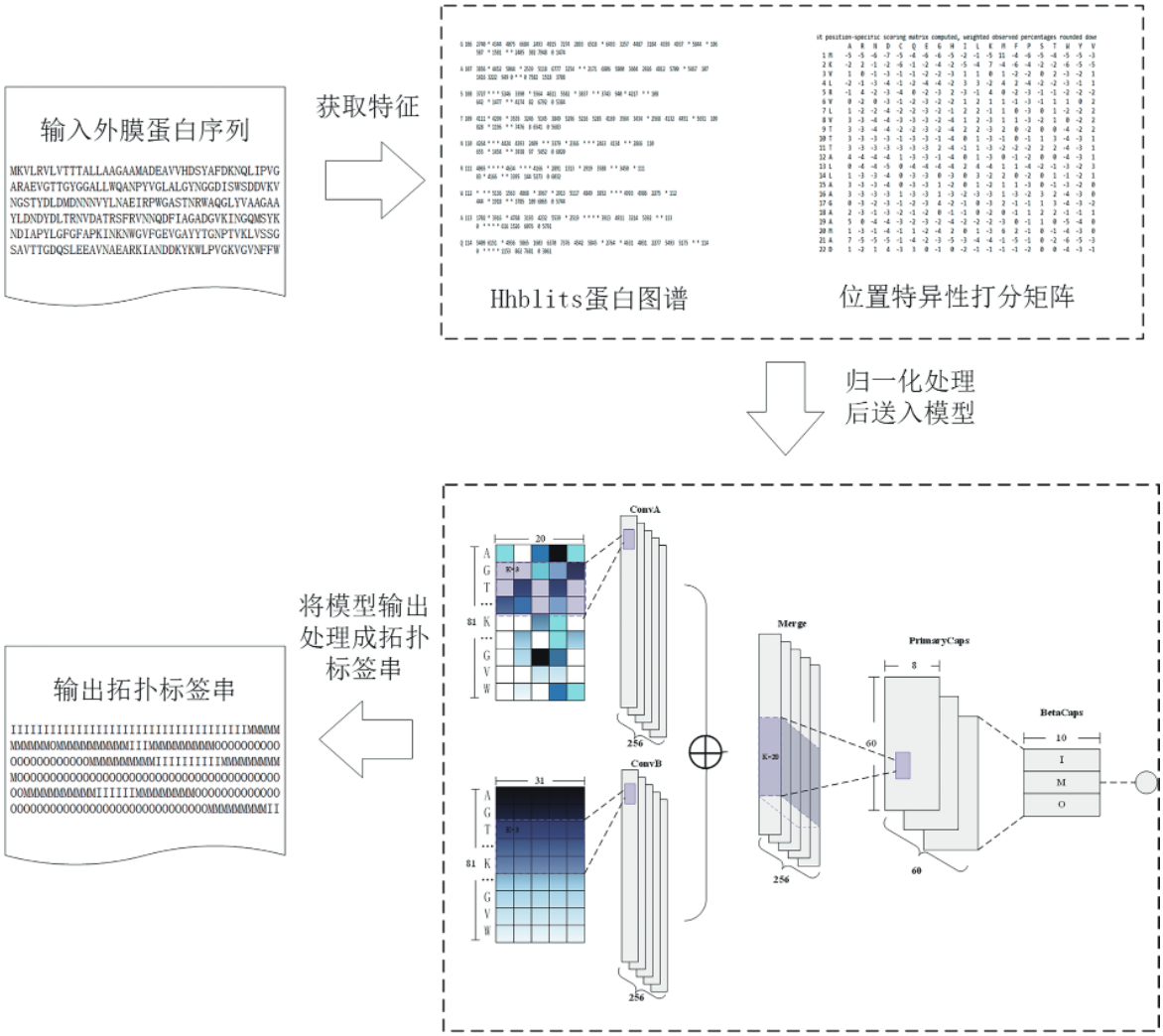

图 5　模型预测流程图

Fig.5　Flowchart of model prediction

3.4.4　模型评价指标

本文选择本类实验普遍使用的 4 项指标[39]（Q3，SOV，Correct #TM，Correct top）从残基、跨膜片段和膜蛋白整体拓扑结构三方面对模型的性能做整体评估。

Q3 值表示模型的预测准确度。Q3 值越大表示被正确预测的残基数量越多，表明模型对蛋白残基的预测效果越好。一般情况是把蛋白拓扑结构主要分为 I（膜内）、M（跨膜）和 O（膜外）3 类，所以将 Q3 值的计算表述为式（5）：

$$Q3=\frac{num_r(I)+num_r(M)+num_r(O)}{len(SS)} \tag{5}$$

式中，num_r 是正确预测的数目，而 $len(SS)$ 是蛋白质长度。

段重叠分数（SOV）用于衡量预测的跨膜片段和真实的跨膜片段之间的平均重合度。预测的跨膜片段与真实的跨膜片段越接近，识别正确率越高。式（6）中，S_1 和 S_2 分别表示真实的序列和预测的序列，而 S_0 表示两者所有状态相同片段。$max(S_1, S_2)$ 表示两种序列的并集，$min(S_1, S_2)$ 表示两种序列的交集，$length(S_1)$ 表示 S_1 的长度。

$$SOV=\frac{100}{N_{SOV}}\sum_{S_0}\left[\frac{min(S_1,S_2)+\delta(S_1,S_2)}{max(S_1,S_2)}length(S_1)\right] \quad (6)$$

式中，δ 的作用是适应蛋白结构边缘处片段的变化，$\delta(S_1, S_2)$ 取值符合以下定义：

$$\delta(S_1,S_2)=min\left\{\begin{matrix}(max(S_1,S_2)-min(S_1,S_2))\\ min(S_1,S_2)\\ int[length(S_1)\div 2]\\ int[length(S_2)\div 2]\end{matrix}\right\} \quad (7)$$

Correct #TM 代表跨膜链数预测正确的蛋白质数，数值越大，表明被正确预测的跨膜片段越多。

$$\text{Correct \#TM}=\sum_{i=1}^{N}x_i;\ x_i=\begin{cases}1, if\ PRE_{Hi}=REC_{it}=1\\ 0, else\end{cases} \quad (8)$$

Correct top 表示具有正确拓扑结构的蛋白质数。Correct top 的值每增加 1，就表示有一个蛋白的基本拓扑结构被正确预测出来。

$$\text{Correct top}=\sum_{i=1}^{N}y_i;\ y_i=\begin{cases}1, if\ PREHi=RECHi=TOP_i=1\\ 0, else\end{cases} \quad (9)$$

4　实验结果与讨论

4.1　特征选取

本文选用 Q3 值评估模型预测性能并单独选取不同的特征进行实验。该实验在获取到交叉验证结果后再求平均值且与不同特征作对比，对比情况如表 7 所示。

表 7　采用不同单一特征时模型的 Q3 预测值

Tab. 7　Q3 predictive value of models with different single features

特征名称	Q3
Phys30	68.7
Phys7	69.3
PSSM	83.2
HHblits	**84.8**
Frag1D	75.4

同样，采用 Q3 值作为模型评估指标，实验组合选取不同特征进行，交叉验证结果同样求平均值作预测精度对比。不同组合对比情况见表 8。

表 8　采用不同组合特征时模型的 Q3 预测值

Tab. 8　Q3 predictive values of models with different combinations of features

特征名称	Q3
Phys30+HHblits	73.2
Phys7+HHblits	74.3
PSSM+HHblits	**86.1**
Frag1D+HHblits	78.1
Phys7+PSSM+HHblits	79.3
Frag1D+PSSM+HHblits	81.9
Phys7+Frag1D+PSSM+HHblits	82.4

通过表 7 和表 8 可以看出，单一特征结果中 HHbilts 蛋白图谱这一特征的结果最优。然而组合实验结果中的第三组，即 PSSM 矩阵和 HHblits 图谱在全部实验中的预测精度最优，所以能够容易得出组合特征 HHblits 图谱加上 PSSM 矩阵是模型的最佳最小输入的结论。

4.2　滑窗尺寸选择

本实验采用 Q3 值作为模型评估指标，选取不同的窗口尺寸进行实验后得到的交叉验证结果同样分别求平均值再互相作对比。对比情况见表 9。

表 9　采用不同窗口大小时模型的 Q3 预测值

Tab. 9　Q3 predictive values for models with different window sizes

滑动窗口大小	Q3
20	84.3
30	86.9
40	**87.1**
50	86.1

经对比，实验确定最终选择窗口尺寸值为 40，数据处理后的特征输入维度是 81×51。

4.3　预测性能交叉验证结果

模型在训练数据集上的交叉验证结果如表 10 所示。

表 10　训练数据集的交叉验证结果（49 蛋白质）

Tab. 10　Comparison of cross validation results on the training dataset (49 proteins)

Method	Q3	Correct #TM	Correct top	SOV
TopOMP-capsnet	0.877	30	28	0.823

模型验证结果表明，通过实验确定的拓扑预测模型 TopOMP-capsnet 的三态预测准确率为 87.7%，SOV 值为 82.3%。其中能够被正确预测出拓扑结构的外膜蛋白条数为 30，跨膜链数预测正确的蛋白数为 28。

4.4　同类方法性能对比

为了对模型的预测性能做更深入的检验，实验

参照与本文蛋白质家族划分种类一致并进行了交叉验证的 PRED-TMBB2 方法，选择相同的数据集和交叉验证方法，对模型的预测性能做整体上的评价和讨论。表 11 展示了该实验的对比情况。

表 11　与 PRED-TMBB2 方法对比

Tab. 11　Versus PRED-TMBB2 method

Method	Q3	Correct #TM	Correct top	SOV
PRED-TMBB2-MSA [40]	0.817	29	26	0.841
PRED-TMBB2-single [40]	0.750	20	15	0.728
TopOMP-capsnet	0.877	30	28	0.838

表 7 中展示的结果全部是同样的 30 折交叉验证方法取得的平均值结果。从表 7 中的实验数据能够看出，TopOMP-capsnet 的预测性能中的 Q3 值高于其他方法，证明该模型能够正确预测出更多的残基，而模型预测片段整体的连续性与准确度方面则有待提升。

为进一步提升模型预测性能的可信度，选择 Hayat 等提出的 BOCTOPUS2 预测方法中的训练集用作基准测试集，目的在于严谨、公平地与现存的外膜蛋白拓扑最佳预测效果的两个方法 BOCTOPUS2、PRED-TMBB2 以及其他外膜蛋白结构的预测模型作对比。交叉验证集的基准结果以及与其他预测模型在 BOCTOPUS2 数据集上的比较如表 12 所示。

表 12　基准数据集结果对比

Tab. 12　Contrasting benchmark dataset results

Method	Q3	Correct #TM	Correct top	SOV
PRED-TMBB2-MSA [40]	0.892	39	32	0.905
PRED-TMBB2-single [40]	0.868	22	14	0.840
BOCTOPUS2 [31]	0.914	35	29	0.925
PROFtmb [41]	0.840	24	18	0.751
PRED-TMBB [42]	0.826	21	12	0.678
BetAware [43]	0.851	23	10	0.725
TMBETAPRED-RBF [44]	0851	19	8	0.559
TopOMP-capsnet	0.920	40	33	0.907

表 13 展示了模型与新的方法 PRED-TMBB2$_{HNN}$ 和 BetAware-Deep 在前文提出的包含 7 条外膜蛋白的独立测试集上的预测性能对比。为了更客观地展示不同方法的预测性能，采用的独立测试集不包含在前期的训练集中，并与两种新方法的训练集进行去冗余处理。表 13 中 PRED-TMBB2$_{HNN}$ 方法没有表现出良好的效果，说明隐马尔可夫模型和神经网络组合训练会导致模型过于依赖训练数据，而在陌生数据集上不够理想。

表 13　独立测试集结果对比

Tab. 13　Contrasting independent dataset results

Method	Q3	Correct #TM	Correct top	SOV
PRED-TMBB2$_{HNN}$ [27]	0.695	2	1	75.92
BetAware-Deep [18]	0.807	6	6	87.74
TopOMP-capsnet	0.854	6	6	94.58

从表 12 和表 13 中的数据信息能够再次看出，模型 TopOMP-capsnet 能够更多地预测出准确的残基状态，模型预测片段在整体的连续性和准确度上表现得不够优秀。该方法能够对外膜蛋白单个残基的特征进行有效的学习，而对于其前后残基关联性与整体结构的连续性特征的学习仍有进步空间。

5　总结

本文提出了基于胶囊网络的外膜蛋白拓扑结构预测模型 TopOMP-capsnet，分别对模型的构建、训练、测试和验证评估的过程做了介绍。该预测模型的准确率达到 87.7%，与其他预测方法相比在蛋白结构预测的精度上有所提升。胶囊网络作为近年来提出的适用于小样本数据特征学习的方法，鲜有领域内研究人员在预测外膜蛋白拓扑结构这一方向做相关尝试。本文设计的实验检验了胶囊网络在膜蛋白结构预测研究中对小样本数据学习的优秀表现。实验结论证明深度学习技术可以很好地对特征进行抽象并学习到残基的特征，进而有效地学习到外膜蛋白的序列模式，再一次表明深度学习技术在膜蛋白结构预测方向具有一定的可行性。

参考文献

[1] VITALI D G, DRWESH L, CICHOCKI B A, *et al*. The Biogenesis of Mitochondrial Outer Membrane Proteins Show Variable Dependence on Import Factors [J]. **iScience**, 2020, 23(1): 100779.

[2] TEHRANI S S, JAHANGIRI A, TAHERIANGANEH M, *et al*. Designing an Outer Membrane Protein (Omp-W) Based Vaccine for Immunization against Vibrio and Salmonella: An in silico Approach [J]. **Recent Pat Biotechnol**, 2020, 14(4): 312-324.

[3] 王磊，柯跃华，栗业，等. 细菌外膜蛋白 A 通用抗原表位预测及其抗体磁珠捕获病原菌效果的初步评价 [J]. **军事医学**, 2021, 45(5): 5.
WANG Lei, KE Yuehua, LI Ye, *et al*. Prediction of universal antigen epitope of bacterial outer membrane protein A and preliminary evaluation of the effect of antibody magnetic beads on capturing pathogens [J]. **Military Medicine**, 2021, 45 (5): 5

[4] SANDOZ K M, MOORE R A, BEARE P A, *et al*. β-Barrel

proteins tether the outer membrane in many Gram-negative bacteria [J]. **Nat Microbiol**, 2021, 6(1): 19-26.

[5] ZHENG J, LI L, JIANG H. Molecular pathways of mitochondrial outer membrane protein degradation [J]. **Biochem Soc Trans**, 2019, 47(5): 1437-1447.

[6] KAUR H, JAKOB R P, MARZINEK J K, *et al.* The antibiotic darobactin mimics a β-strand to inhibit outer membrane insertase [J]. **Nature**, 2021: 1-5.

[7] ANWAR M, MUHAMMAD F, AKHTAR B, *et al*. Outer Membrane Protein-Coated Nanoparticles as Antibacterial Vaccine Candidates [J]. **International Journal of Peptide Research and Therapeutics**, 2021, 27(3): 1689-1697.

[8] SOLANKI V, SHARMA S, TIWARI V. Subtractive Proteomics and Reverse Vaccinology Strategies for Designing a Multiepitope Vaccine Targeting Membrane Proteins of Klebsiella Pneumoniae [J]. **International Journal of Peptide Research and Therapeutics**, 2021, 27(2): 1177-1195.

[9] HANSON J, PALIWAL K, LITFIN T, *et al*. Improving prediction of protein secondary structure, backbone angles, solvent accessibility and contact numbers by using predicted contact maps and an ensemble of recurrent and residual convolutional neural networks [J]. **Bioinformatics**, 2019, 35(14): 2403-2410.

[10] FANG C, SHANG Y, XU D. Improving Protein Gamma-Turn Prediction Using Inception Capsule Networks [J]. **Scientific Reports**, 2018, 8(1): 15741.

[11] LIU Z, GONG Y, BAO Y, *et al*. TMPSS: A Deep Learning-Based Predictor for Secondary Structure and Topology Structure Prediction of Alpha-Helical Transmembrane Proteins [J]. **Front Bioeng Biotechnol**, 2020, 8: 629937.

[12] WANG H, YANG Y, YU J, *et al*. DMCTOP: Topology Prediction of Alpha-Helical Transmembrane Protein Based on Deep Multi-Scale Convolutional Neural Network [C]. 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2019.

[13] YANG Y, YU J, LIU Z, *et al*. An Improved Topology Prediction of Alpha-Helical Transmembrane Protein Based on Deep Multi-Scale Convolutional Neural Network [J]. **IEEE/ACM Trans Comput Biol Bioinform**, 2022, 19(1): 295-304.

[14] HEFFERNAN R, YANG Y, PALIWAL K, *et al*. Capturing non-local interactions by long short-term memory bidirectional recurrent neural networks for improving prediction of protein secondary structure, backbone angles, contact numbers and solvent accessibility [J]. **Bioinformatics**, 2017, 33(18): 2842-2849.

[15] WANG S, SUN S, LI Z, *et al*. Accurate De Novo Prediction of Protein Contact Map by Ultra-Deep Learning Model [J]. **PLoS Comput Biol**, 2017, 13(1): e1005324.

[16] WANG D, LIANG Y, XU D. Capsule network for protein post-translational modification site prediction [J]. **Bioinformatics**, 2019, 35(14): 2386-2394.

[17] FANG C, SHANG Y, XU D. MUFOLD-SS: New deep inception-inside-inception networks for protein secondary structure prediction [J]. **Proteins: Structure, Function, and Bioinformatics**, 2018, 86(5): 592-598.

[18] MADEO G, SAVOJARDO C, MARTELLI P L, *et al*. BetAware-Deep: An Accurate Web Server for Discrimination and Topology Prediction of Prokaryotic Transmembrane beta-barrel Proteins [J]. **Journal of Molecular Biology**, 2021, 433(11): 166729.

[19] JUMPER J, EVANS R, PRITZEL A, *et al*. Highly accurate protein structure prediction with AlphaFold [J]. **Nature**, 2021, 596(7873): 583-589.

[20] TUNYASUVUNAKOOL K, ADLER J, WU Z, *et al*. Highly accurate protein structure prediction for the human proteome[J]. **Nature,** 2021, 596(7873): 590-596.

[21] HEGEDŰS T, GEISLER M, LUKÁCS G L, *et al*. Ins and outs of AlphaFold2 transmembrane protein structure predictions[J]. **Cellular and Molecular Life Sciences**, 2022, 79(1): 1-12.

[22] LOMIZE M A, POGOZHEVA I D, JOO H, *et al*. OPM database and PPM web server: resources for positioning of proteins in membranes [J]. **Nucleic Acids Res**, 2012, 40(Database issue): D370-D376.

[23] KOZMA D, SIMON I, TUSNÁDY G E. PDBTM: Protein Data Bank of transmembrane proteins after 8 years [J]. **Nucleic Acids Research**, 2012, 41(D1): D524-D529.

[24] JAYASINGHE S, HRISTOVA K, WHITE S H. MPtopo: A database of membrane protein topology [J]. **Protein Sci**, 2001, 10(2): 455-458.

[25] TSIRIGOS K D, BAGOS P G, HAMODRAKAS S J. OMPdb: a database of β-barrel outer membrane proteins from Gram-negative bacteria [J]. **Nucleic Acids Research**, 2010, 39(suppl_1): D324-D331.

[26] TUSNÁDY G E, KALMÁR L, SIMON I. TOPDB: topology data bank of transmembrane proteins [J]. **Nucleic Acids Res**, 2008, 36(Database issue): D234-D239.

[27] TAMPOSIS I A, SARANTOPOULOU D, THEODOROPOULOU M C, *et al*. Hidden neural networks for transmembrane protein topology prediction [J]. **Comput Struct Biotechnol J**, 2021, 19: 6090-6097.

[28] FINN R D, COGGILL P, EBERHARDT R Y, *et al*. The Pfam protein families database: towards a more sustainable future[J]. **Nucleic Acids Res**, 2016, 44(D1): D279-D285.

[29] Hobohm U, Scharf M, Schneider R, *et al*. Selection of representative protein data sets [J]. **Protein Sci**, 1992, 1(3): 409-417.

[30] ALTSCHUL S F, MADDEN T L, SCHÄFFER A A, *et al*. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs [J]. **Nucleic Acids Research**, 1997, 25(17): 3389-3402.

[31] HAYAT S, PETERS C, SHU N, *et al*. Inclusion of dyad-repeat pattern improves topology prediction of transmembrane beta-barrel proteins [J]. **Bioinformatics**, 2016, 32(10): 1571-1573.

[32] LI W, GODZIK A. Cd-hit: a fast program for clustering and

comparing large sets of protein or nucleotide sequences [J]. **Bioinformatics**, 2006, 22(13): 1658-1659.

[33] KAWASHIMA S, KANEHISA M. AAindex: amino acid index database [J]. **Nucleic Acids Res**, 2000, 28(1): 374.

[34] FINN R D, CLEMENTS J, ARNDT W, *et al*. HMMER web server: 2015 update [J]. **Nucleic Acids Res**, 2015, 43(W1): W30-W38.

[35] REMMERT M, BIEGERT A, HAUSER A, *et al*. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment [J]. **Nat Methods**, 2011, 9(2): 173-175.

[36] ZHOU T, SHU N, HOVMOLLER S. A novel method for accurate one-dimensional protein structure prediction based on fragment matching [J]. **Bioinformatics**, 2010, 26(4): 470-477.

[37] SABOUR S, FROSST N, HINTON G E. Dynamic routing between capsules [C]. Proceedings of the 31st International Conference on Neural Information Processing Systems, 2017: 3859–3869.

[38] KINGMA D, BA J. Adam: A Method for Stochastic Optimization [J]. **Computer Science**, 2014: 1-14.

[39] BAGOS P G, LIAKOPOULOS T D, HAMODRAKAS S J. Evaluation of methods for predicting the topology of β-barrel outer membrane proteins and a consensus prediction method[J]. **BMC Bioinformatics**, 2005, 6(1): 7.

[40] TSIRIGOS K D, ELOFSSON A, BAGOS P G. PRED-TMBB2: improved topology prediction and detection of beta-barrel outer membrane proteins [J]. **Bioinformatics**, 2016, 32(17): i665-i671.

[41] BIGELOW H R, PETREY D S, LIU J, *et al*. Predicting transmembrane beta-barrels in proteomes [J]. **Nucleic Acids Research**, 2004, 32(8): 2566-2577.

[42] BAGOS P G, LIAKOPOULOS T D, SPYROPOULOS I C, *et al*. A Hidden Markov Model method, capable of predicting and discriminating beta-barrel outer membrane proteins [J]. **BMC Bioinformatics**, 2004, 5: 29.

[43] SAVOJARDO C, FARISELLI P, CASADIO R. BETAWARE: a machine-learning tool to detect and predict transmembrane beta-barrel proteins in prokaryotes [J]. **Bioinformatics,** 2013, 29(4): 504-505.

[44] OU Y Y, CHEN S A, GROMIHA M M. Prediction of membrane spanning segments and topology in beta-barrel membrane proteins at better accuracy [J]. **Journal of Computational Chemistry**, 2010, 31(1): 217-223.