基于序列与跨模态对齐的蛋白质功能预测模型

徐 敏,胡春玲*,胡 婷,张芳芳,代相龙
(合肥大学人工智能与大数据学院,安徽合肥 230031)

摘 要: 蛋白质功能预测是生物信息学核心任务之一. 现有方法虽能实现蛋白质多模态特征的融合,但仍存在预测准确率不足、依赖有限的实验数据导致适用范围受限等问题. 为解决此类问题,本研究提出基于序列与跨模态对齐的蛋白质功能预测模型(Sequence-based and Cross-Modal Alignment Model for Protein Function Prediction,SCMAGO),以蛋白质序列作为唯一输入,通过主流工具 AlphaFold2、InterProScan 分别预测三级结构和家族结构域信息;使用蛋白质大语言模型(Evolutionary Scale Model Cambrian,ESMC)实现序列嵌入,并采用几何向量感知机图神经网络(Geometric Vector Perceptron Graph Neural Network,GVP-GNN)提取三级结构特征,再通过广播嵌入方法获取家族结构域表示;模型 SCMAGO 设计两步跨模态对齐方法:基于双向交叉注意力,在残基层面对序列和结构特征进行对齐;结合图注意力池化方法,进一步融合家族结构域特征. 实验结果表明,SCMAGO 在 Swiss-Prot 数据集上的性能优于现有的基准方法,在生物过程(Biological Process,BP)、分子功能(Molecular Function,MF)和细胞组分(Cellular Component,CC)三方面的 F_{max} 分别为 0.487、0.739 和 0.736,AUPR 则分别达到 0.507、0.760、0.800. 此外,对序列一致性低于 40% 的蛋白质,仍能保持稳定的预测性能.

关键词: 蛋白质功能预测;多模态融合;注意力机制;Gene Ontology
基金项目: 国家自然科学基金(No.62306100)
中图分类号: TP18;Q51 **文献标识码:** A **文章编号:** 0372-2112(2025)11-4022-13
电子学报 URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20250851

Sequence-Based and Cross-Modal Alignment Model for Protein Function Prediction

XU Min, HU Chun-ling*, HU Ting, ZHANG Fang-fang, DAI Xiang-long
(*School of Artificial Intelligence and Big Data*, *Hefei University*, *Hefei*, *Anhui* 230031, *China*)

Abstract: Protein function prediction is one of the core tasks in bioinformatics. Although existing methods can fuse multimodal features of proteins, they still suffer from issues such as insufficient prediction accuracy and limited application scope due to reliance on limited experimental data. To address these problems, this study proposes a sequence- and cross-modal alignment-based protein function prediction model (SCMAGO), which takes protein sequences as the sole input. Specifically, it predicts tertiary structure and family domain information using the mainstream tools AlphaFold2 and InterProScan, respectively. It employs the protein large language model (Evolutionary Scale Model Cambrian, ESMC) to achieve sequence embedding, uses the geometric vector perceptron graph neural network (GVP-GNN) to extract tertiary structure features, and further obtains family domain representations through the broadcast embedding method. The SCMAGO model is designed with a two-step cross-modal alignment approach: first, it aligns sequence and structure features at the residue level based on bidirectional cross-attention; second, it further fuses family domain features by combining the graph attention pooling method. Experimental results show that SCMAGO outperforms existing benchmark methods on the Swiss-Prot dataset. Its F_{max} values for biological process (BP), molecular function (MF), and cellular component (CC) are 0.487, 0.739 and 0.736, respectively, while the corresponding AUPR values reach 0.507, 0.760 and 0.800. Furthermore, SCMAGO still maintains stable prediction performance for proteins with sequence identity below 40%.

Key words: protein function prediction; multimodal fusion; attention mechanism; Gene Ontology
Foundation Item(s): National Natural Science Foundation of China (No.62306100)

收稿日期:2025-09-27;录用日期:2025-11-10;责任编辑:覃怀银
*通讯作者:胡春玲

1　引言

蛋白质是生命活动的主要承担者和执行者,精确解析其功能在破译生物分子机制、揭示疾病病理以及开发新药物等方面至关重要. 然而,通过实验手段(如基因敲除、组学技术等)解析蛋白质功能不仅周期长,而且成本高昂. 随着高通量测序技术的快速发展,海量蛋白质序列数据的产出加剧了序列同功能注释产出能力之间的不平衡. 截至2025年,UniProt(Universal Protein)[1]数据库包含了约2.46亿条蛋白质序列记录,但其中经过专家验证(Swiss-Prot证据代码如EXP、IDA等)具有注释信息的蛋白质仅231 709条,占比不到0.1%. 因此,对蛋白质功能注释的预测成为生物信息学中一项重要任务.

尽管基于深度学习的蛋白质功能预测方法已产出大量成果,在预测性能上有一定突破,但仍有可优化的方向:其一,对于缺乏多生物数据库信息支持的蛋白质,现有模型难以实现有效预测;其二,在蛋白质多模态特征的高效融合方法上面临挑战. 为解决上述两个问题,本研究提出新型蛋白质功能预测模型SCMAGO(Sequence-based and Cross-Modal Alignment Model for Protein Function Prediction),本文工作如下.

(1)多模态信息获取. 以蛋白质序列为基础,使用AlphaFold2[2]预测三级结构、InterProScan[3]预测家族结构域信息,通过这种仅依赖序列即可拓展多模态信息的方式,降低模型对实验测定的多生物数据库的依赖;首次引入蛋白质预训练大语言模型ESMC(Evolutionary Scale Model Cambrian)[4],完成对蛋白质序列的初始嵌入.

(2)模型框架设计. 基于三种编码器模块提取蛋白质序列、三级结构、家族结构域特征;创新提出两步跨模态对齐方法,精准挖掘不同模态信息间的内在关联,构建从蛋白质序列到功能预测的端到端映射.

(3)实验验证. 在Swiss-Prot数据集上开展大量实验,结果表明SCMAGO方法优于现有的同类型基线方法;进一步验证和分析了ESMC模型的引入、序列编码器及两步跨模态对齐方法的设计有效性.

2　相关工作

2.1　基于序列的方法

早期蛋白质功能预测依赖序列相似性推断,通过比对目标蛋白质与已知功能蛋白质的序列同源性,实现已知功能信息的跨物种或跨家族迁移,如BLAST[5]、PSI-BLAST(Position Specific Iterated BLAST)[6]等序列比对工具,该类方法在面对新型蛋白质或远缘同源蛋白质时具有局限性,难以准确推断其功能. 深度学习的兴起为研究蛋白质功能预测带来了突破性的进展. 通过多层非线性网络,构建从蛋白质序列到功能语义的映射,能够捕捉序列比对技术无法识别的隐式特征(如氨基酸上下文依赖等),显著提升了功能预测的准确性. DeepGO[7]是最早使用深度学习预测蛋白质功能的方法之一,通过序列特征结合蛋白质相互作用网络实现对功能的预测. 此后,一系列基于单一序列特征的深度学习模型相继提出,例如ProLanGO[8]将序列转化为基于k-mer片段的"ProLan"语言,通过循环神经网络构建机器翻译模型实现序列到功能的映射. DEEPred[9]通过比较k-mer片段、位置特异性评分矩阵(Position-Specific Scoring Matrix,PSSM)等序列嵌入方法,筛选出最佳特征以优化神经网络模型,并在CAFA2[10]和CAFA3[11](Critical Assessment of Functional Annotation,CAFA)大规模基准数据集上验证了深度学习模型的实用性. DeepGOPlus[12]在基于序列比对的基础上,结合卷积神经网络(Convolutional Neural Networks,CNN)[13]来提取蛋白质序列特征. GODoc[14]以PSI-LAST生成的PSSM为基础构建序列特征,通过主成分分析法(Principal Component Analysis,PCA)与新型K-近邻算法(K-Nearest Neighbor,KNN)展现出优异性能.

2.2　基于多模态特征的方法

DeepFRI[15]是早期融合序列和三级结构特征的方法,在来自Pfam[16]数据库的大约1 000万个蛋白质序列语料库上训练长短期记忆网络大语言模型(long short-term memory language model)[17],用于对序列特征的嵌入,再使用图卷积网络(Graph Convolutional Network,GCN)[18]提取来自PDB(Protein Data Bank)数据库[19]中的三级结构信息. GAT-GO[20]和Struct2Graph[21]使用结构信息构建二维蛋白质残基接触图作为模型输入. 此类方法虽然通过残基接触图实现对蛋白质结构信息的建模,但丢失了残基在三维空间中的相对位置、空间距离等关键信息.

GVP-GNN[22]方法主要以蛋白质结构设计为目标,将三维空间特性应用于蛋白质结构建模,采用几何向量感知机(Geometric Vector Perceptron,GVP)对蛋白质三级结构中的等变性特征进行建模,这种架构突破了难以将蛋白质结构映射到三维空间的问题. PredGO[23]以GVP-GNN框架为基础,结合蛋白质与蛋白质相互作用网络信息进行预测,取得了优异成效. DeepMind团队开发的蛋白质结构预测模型AlphaFold2能基于氨基酸序列精准预测三维空间结构,有效解决了PDB数据库中实验解析结构稀缺的难题. PANDA-3D[24]借鉴GVP-GNN架构,在AlphaFold2预测的三级结构中筛选出置信度指标预测局部距离差异测试(predicted Local Distance Difference Test,pLDDT)高于0.9的残基片段作为模型输入,对蛋白质功能进行预测. 此外,蛋白质大语言模型(如ESM-1b[25]、ESM2[26]等)凭借对蛋白质序列

深层语义信息的捕捉，为蛋白质功能预测带来了重要支撑．ATGO[27]、SPROF-GO[28]、TransFew[29]、DPFunc[30]等方法通过蛋白质大模型对序列进行初始嵌入，以此为基础来提升模型的预测性能．其中，DPFunc从生物数据库InterProDB[31]中获取可用信息，创新性地将结构域特征应用到模型中，为蛋白质功能预测提供了有效思路．

3 模型设计与分析

3.1 模型概述

SCMAGO模型框架如图1所示，由三级结构编码器a、序列编码器b、家族结构域编码器c，以及两步特征对齐模块d和e构成．结构编码器主要由数据预处理和GVP-GNN模块组成；序列编码器先将序列经过ESMC生成初级特征，再通过卷积模块进一步提取高阶特征；家族结构域编码器借助嵌入层和广播机制生成特征矩阵；序列-结构对齐模块基于双向交叉注意力机制，实现序列和结构特征在残基层面的对齐融合；序列-结构的融合特征与家族结构域特征对齐模块由图神经网络和图注意力池化方法组成，最终完成对蛋白质跨模态特征的深度整合．

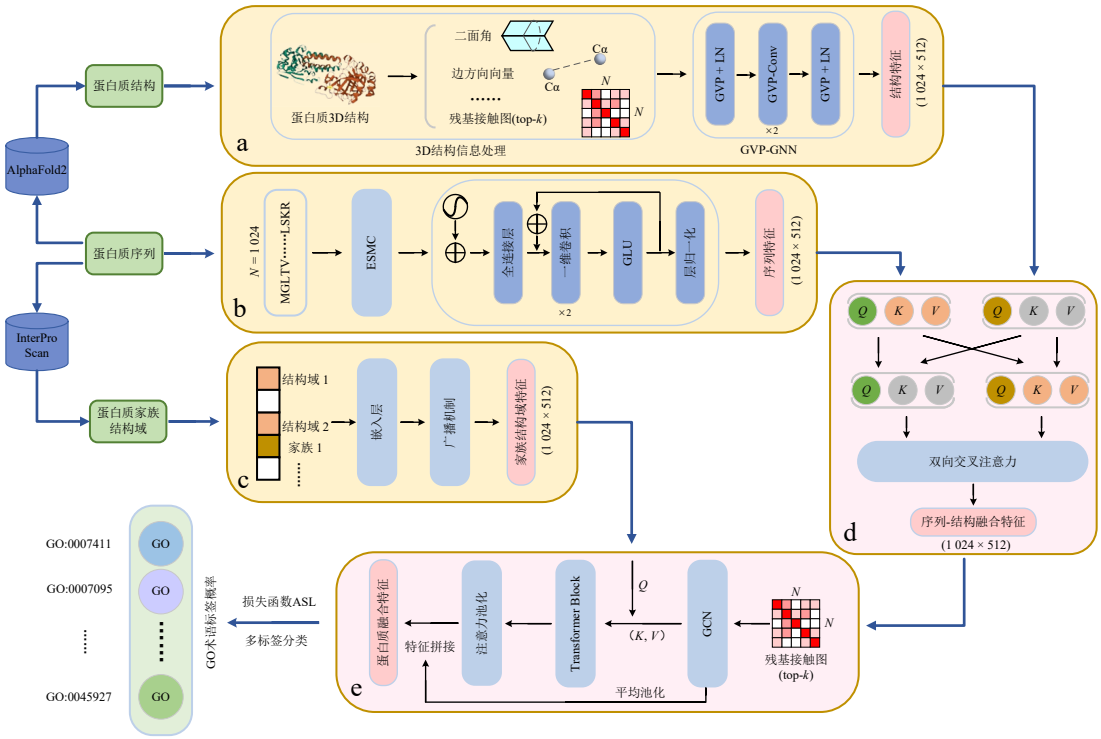


图1 SCMAGO模型框架图

3.2 三级结构编码器

模块a为三级结构编码器．在AlphaFold2预测的三级结构数据中提取氨基酸残基中主链原子Cα、C和N的坐标．基于Cα之间的原子距离，使用KNN方法建立残基接触图．将残基主链原子之间的二面角大小、欧几里得距离以及位置编码等具有旋转不变性的特征，处理为接触图中节点的标量特征s；将原子间相对方向、图中边的方向等具有等变性的特征，处理为接触图中节点的向量特征$\boldsymbol{v}$．使用GVP-GNN框架完成对蛋白质结构的三维建模．

在GVP-GNN框架中，GVP模块接收由标量特征$s\in\mathbb{R}^n$和向量特征$\boldsymbol{v}\in\mathbb{R}^{v\times 3}$组成的元组，输出更新后的标量特征$s'\in\mathbb{R}^m$和向量特征$\boldsymbol{v}'\in\mathbb{R}^{v\times 3}$．如式(1)所示：

$$\mathrm{GVP}(s,\boldsymbol{v})=(s',\boldsymbol{v}') \tag{1}$$

式(2)为邻居节点j向节点i的消息传递过程：

$$h_m^{j\to i}=\mathrm{GVP}\left(\mathrm{concat}\left(h_v^j,h_e^{j\to i}\right)\right) \tag{2}$$

其中，h_v^j为节点j的特征；$h_e^{j\to i}$为边(j,i)的特征；concat(·)表示特征拼接操作．

式(3)描述在蛋白质残基接触图中，通过聚合所有邻居节点j的特征更新节点i的过程：

$$h_v^i=\mathrm{LN}\left(h_v^i+\frac{1}{k'}\mathrm{Dropout}\left(\sum_{j:e_{j\to i}\in\varepsilon}h_m^{j\to i}\right)\right) \tag{3}$$

其中，k'表示节点i的邻居节点个数；Dropout(·)表示丢弃层；LN(·)表示层归一化．

最后，使用GVP模块将残基节点i的标量特征与向量特征继续更新，得到节点i的三级结构特征，如式(4)所示：

$$h_v^i = \mathrm{LN}\left(h_v^i + \mathrm{Dropout}\left(\mathrm{GVP}\left(h_v^i\right)\right)\right) \tag{4}$$

3.3　序列编码器

模块b为序列编码器. 将蛋白质序列长度截断至1 024后,输入ESMC模型进行初步的特征嵌入,使每个氨基酸获得1 152维的特征表示. 本研究设计了结合一维卷积与位置编码的特征提取方法,用于从初始序列嵌入中挖掘具有判别性的高阶特征. 实现方法如下.

(1)对ESMC模型生成的初始序列嵌入特征引入位置编码,以保留蛋白质序列中每个氨基酸的位置信息,避免模型丢失序列的时序关联性;

(2)采用全连接层对融合了位置编码后的特征进行维度压缩,降低模型计算复杂度的同时,实现特征的初步筛选;

(3)通过一维卷积在序列维度将特征长度扩展至氨基酸数量N的两倍,得到输出$Y=[AB]\in\mathbb{R}^{2N\times d}$($N$为序列长度,$d$为残基特征长度),再引入门控线性单元(Gated Linear Unit, GLU)[32]作为激活函数,如式(5)所示:

$$\mathrm{GLU}\left(\left[AB\right]\right) = A\odot\sigma\left(B\right) \tag{5}$$

其中,$A\in\mathbb{R}^{N\times d}$,$B\in\mathbb{R}^{N\times d}$;$\sigma(\cdot)$表示sigmoid激活函数;$\odot$表示向量逐元素乘法. 通过GLU门控机制,实现对序列特征的选择性传递,有效保留序列中的关键信息,抑制冗余特征和噪声干扰.

3.4　家族结构域编码器

模块c为蛋白质家族结构域编码器. 本研究统计出Swiss-Prot数据集中所有蛋白质在InterProDB中对应的家族结构域信息,总共得到29 270个唯一ID(如结构域、蛋白质家族、功能位点等),使用独热编码嵌入后作为家族结构域特征的初始表示. 随后通过嵌入层将这些独热编码特征映射到512维稠密向量空间,以此捕捉不同蛋白质之间的功能与进化关联. 最后采用广播机制,将表示家族结构域特征的稠密向量扩展到蛋白质的每个残基上.

3.5　序列与结构对齐模块

模块d为序列与结构对齐模块,本研究提出一种基于双向交叉注意力[33]的残基层面对齐机制. 通过构建两个独立的注意力路径,允许序列特征和结构特征相互引导信息的传递. 本模块的输入为蛋白质序列特征$\mathrm{seq}\in\mathbb{R}^{N\times d}$与三级结构特征$\mathrm{str}\in\mathbb{R}^{N\times d}$. 设计如下.

序列到结构的注意力路径如式(6)所示,以序列特征seq作为Query,三级结构特征str作为Key和Value,使序列的每个残基能够选择性地关注对应的结构特征:

$$\mathrm{Attn}_{\mathrm{seq}\to\mathrm{str}} = \mathrm{softmax}\left(\frac{Q_{\mathrm{seq}}K_{\mathrm{str}}^{\mathrm{T}}}{\sqrt{d_k}}\right)V_{\mathrm{str}} \tag{6}$$

结构到序列的注意力路径如式(7)所示,以残基的三级结构特征str为Query,序列特征seq作为Key和Value,使每个残基的结构特征能关注蛋白质序列的局部和全局语义:

$$\mathrm{Attn}_{\mathrm{str}\to\mathrm{seq}} = \mathrm{softmax}\left(\frac{Q_{\mathrm{str}}K_{\mathrm{seq}}^{\mathrm{T}}}{\sqrt{d_k}}\right)V_{\mathrm{seq}} \tag{7}$$

最后通过式(8)获得序列-结构融合特征Fusion:

$$\mathrm{Fusion} = \mathrm{FC}\left(\mathrm{concat}\left(\mathrm{Attn}_{\mathrm{seq}\to\mathrm{str}}, \mathrm{Attn}_{\mathrm{str}\to\mathrm{seq}}\right)\right) \tag{8}$$

其中,concat(·)表示特征拼接;FC(·)为全连接层.

3.6　序列-结构与家族结构域对齐模块

模块e为序列-结构融合特征与家族结构域特征的对齐模块. 聚焦于解决传统图平均池化方法的不足,通过构建更优的特征对齐与融合方式,实现蛋白质的全局表示.

为获取蛋白质的全局表示,传统图平均池化方法通常是对所有残基特征进行求和后取平均,如式(9)所示:

$$y_{\mathrm{mean}} = \frac{1}{N}\sum_{i=1}^{N} h_i \tag{9}$$

其中,y_{mean}是蛋白质全局表示;h_i表示残基i的序列-结构融合特征.

然而,上述方式无法区分序列中不同残基的重要程度. 本模块采用图注意力池化策略,通过注意力机制,根据每个残基对家族结构域特征的贡献度大小,为不同残基节点分配对应权重. 方法为:以残基接触图为基础架构,将序列-结构融合特征作为图的初始节点特征,经过多层GCN的聚合后,作为图注意力池化中的Key和Value;再将家族结构域特征作为Query,计算每个残基的注意力权重α_i,如式(10)所示:

$$\alpha_i = \mathrm{softmax}\left(\frac{QK_i^{\mathrm{T}}}{\sqrt{d_k}}\right) \tag{10}$$

其中,Q表示家族结构域特征;K_i为第i个残基经过多层GCN更新后的序列-结构融合特征;softmax(·)为激活函数.

蛋白质的最终全局特征表示y_{final}计算方法如式(11)所示:通过每个残基的注意力权重权α_i对序列-结构融合特征V_i进行加权求和,再拼接平均池化结果y_{mean},从而保留蛋白质的全局和局部关键特征,即

$$y_{\mathrm{final}} = \mathrm{concat}\left(y_{\mathrm{mean}}, \sum_{i=1}^{N}\alpha_i V_i\right) \tag{11}$$

其中,concat(·)表示特征拼接.

3.7　损失函数

蛋白质功能预测为多标签分类任务,样本中正标签的数量一般远少于负标签数量. 以蛋白质“P35213”

在分子功能(Molecular Function,MF)方面的基因本体(Gene Ontology,GO)[34]术语为例,其正标签对应的GO术语有"GO:0003714"、"GO:0003712"、"GO:0140110"、"GO:0008022"、"GO:0005515"、"GO:0003674"和"GO:0005488"七种,而数据集中MF标签总种类达441个.正标签占比仅约1.59%,其余均为负标签.因此,本研究采用非对称损失函数(ASymmetric Loss,ASL)[35].

在传统多标签分类任务中,样本正负标签的不平衡性会主导优化过程,致使模型在训练期间往往会忽略来自正标签的梯度信息,从而降低预测的准确性.ASL通过不对称聚焦和概率偏移机制,有效缓解了样本正负标签不平衡主导模型优化过程的问题,定义如式(12)所示:

$$\mathrm{ASL}=\begin{cases}\left(1-p\right)^{\gamma^+}\log_2 p\\ \max\left(p-m,0\right)^{\gamma^-}\log_2\left(1-\max\left(p-m,0\right)\right)\end{cases}\tag{12}$$

其中,γ^+为正标签的聚焦参数,用于强化少量正标签的梯度贡献;γ^-为负标签的聚焦参数,通过设置相对较大的值,可以有效抑制负样本的整体损失,使正负标签损失达到动态平衡;m为概率偏移参数,对于预测概率小于m的负标签(即高置信度预测为负的标签),其预测概率p接近于0,对分类能力提升无实际意义,但占用大部分计算资源并产生梯度冗余.ASL通过$\max(p-m,0)$,可直接忽略这类标签的梯度贡献,进一步提升训练效率与精确率.

4 实验结果与分析

4.1 基准数据集

本研究从Swiss-Prot数据库[36]下载发布于2022年5月25日且经过人工审查的蛋白质数据(证据代码:EXP、IDA、IMP、IGI、IEP、TAS或IC),包括序列和GO功能注释术语.将每个蛋白质的GO术语按照生物过程(Biological Process,BP)、MF、细胞组分(Cellular Component,CC)三个方面分类后,根据2022年7月1日发布的GO定义文件[37],将每个GO术语沿着有向无环图传递到根节点,作为该蛋白质在功能预测任务中的标签.本研究从AlphaFoldDB[38]获取预测的蛋白质三级结构数据,对于数据库中未收录的蛋白质,使用在线AlphaFold2工具预测其结构.由于AlphaFold2对蛋白质主链的预测具有高度可靠性,为保障最终训练模型具备多场景适用性,不再区分该数据的评估指标pLDDT,使模型在不同置信度的蛋白质预测结构上均能有效预测,避免因依赖高可靠性结构而局限模型的应用范围.蛋白质家族结构域信息来自InterProScan工具的预测结果.

为保证本研究方法与基线方法的公平对比,数据做以下统一处理:首先采用序列比对工具CD-HIT[39]对测试集进行去冗余处理,设定序列一致性阈值为95%,以排除高度相似样本对预测性能评估的干扰;其次对GO术语标签进行过滤,剔除数据集中出现频次小于50的GO术语标签,这样既能通过减少低频术语降低计算复杂度,又可以通过聚焦高频、高代表性术语,提升模型对主要功能的预测精度;但同时也存在局限性,无法对低频GO术语实现有效识别与预测.

基于上述数据来源和处理流程,最终筛选出64 070条高质量蛋白质数据,按照BP、MF、CC三个方面以8:1:1的比例划分为训练集、验证集和测试集,如表1所示.

表1 Swiss-Prot数据集

GO术语类型	蛋白质数量	GO标签数量
BP	50 864	3 149
MF	36 794	441
CC	42 716	412

4.2 评价指标

本研究采用最大$F1$分数($F_{\max}$)、最小语义距离($S_{\min}$)和精准率-召回率曲线下面积(AUPR)来评价模型预测结果.其中$F_{\max}$和$S_{\min}$是CAFA竞赛中的主要评价指标;AUPR适用于类别不平衡,特别在正样本极少的分类任务中,评估模型在不同召回率水平下维持高精确率的能力.

$F_{\max}$在置信度阈值t下的计算如式(13)~(17)所示.其中$p(t)$和$r(t)$为测试蛋白质的平均精确率和平均召回率;$m(t)$表示至少预测出一个GO术语的蛋白质数量;$p_i(t)$和$r_i(t)$是第i个蛋白质的精确率和召回率;n为测试集蛋白质总数量;$T_i(t)$为第i个蛋白质GO术语预测分数大于阈值t的标签集合;Y_i表示该蛋白质的真实标签集合.

$$F_{\max}=\max\left(\frac{2p(t)r(t)}{p(t)+r(t)}\right)\tag{13}$$

$$p(t)=\frac{1}{m(t)}\sum_{i=1}^{m(t)}p_i(t)\tag{14}$$

$$r(t)=\frac{1}{n(t)}\sum_{i=1}^{n(t)}r_i(t)\tag{15}$$

$$p_i(t)=\frac{\left|T_i(t)\cap Y_i\right|}{\left|T_i(t)\right|}\tag{16}$$

$$r_i(t)=\frac{\left|T_i(t)\cap Y_i\right|}{\left|Y_i\right|}\tag{17}$$

$S_{\min}$的计算方法基于GO术语标签类别的信息含量(Information Content,IC),如式(18)~(21)所示.IC(·)表示GO术语的IC值,Fa(·)表示术语的父术语标签集合,P(·)表示条件概率.对于功能描述越精细的GO术语,其出现频率越低,则IC值越大;ru(t)为阈值t下未预测到

真实正例的代价,mi(t)为阈值t下错误预测为正例的代价;n为数据集中蛋白质总数,其中$I(\cdot)$为指示函数,当条件满足时为1,否则为0.

$$\mathrm{IC}(c)=-\log\left(P\left(c|\mathrm{Fa}(c)\right)\right) \tag{18}$$

$$S_{\min}=\min_t\sqrt{\mathrm{ru}(t)^2+\mathrm{mi}(t)^2} \tag{19}$$

$$\mathrm{ru}(t)=\frac{1}{n}\sum_{i=1}^{n}\sum_{c}\mathrm{IC}(c)\times I\left(c\notin T_i(t)\wedge c\in Y_i\right) \tag{20}$$

$$\mathrm{mi}(t)=\frac{1}{n}\sum_{i=1}^{n}\sum_{c}\mathrm{IC}(c)\times I\left(c\in T_i(t)\wedge c\notin Y_i\right) \tag{21}$$

4.3　实验参数

模型代码基于PyTorch、ESMC、GCN和Transformer等框架. 序列编码器中卷积层设为2,卷积核大小设为7;结构编码器采用2层GVP-GNN;序列-结构对齐模块使用8个注意力头;家族结构域对齐模块包含4个注意力头和1层GCN;损失函数ASL的正例聚焦系数设置为1,负例聚焦系数设置为5,裁剪系数为0.05;模型使用Adam优化器,初始学习率设为1×10^{-4},结合ReduceLROnPlateau学习率调度器:以验证集$F_{\max}$指标为监控目标,衰减因子设为0.5,参数patience设为4,实现学习率动态调整,保障模型收敛的稳定性;批次大小设为16;Dropout层的丢弃率统一设为0.2;基础训练轮次设为80次.

4.4　对比实验

本研究基于Swiss-Prot数据集,采用蛋白质功能预测的评估方法,选取两类方法作为基线进行对比:一是CAFA竞赛中广泛采用的经典方法Naïve与BLAST;二是领域内代表性的深度学习方法DeepGOCNN、DeepFRI、PANDA-3D及DPFunc. 这些基线方法在特征输入与设计上各有特点,但部分方法依赖实验测定的蛋白质信息,或采用高置信度的预测结构,这种信息获取方式在数据稀缺场景下限制了模型的泛化能力,为预测精度提升留下空间.

表2为对比实验结果,显示七种方法在$F_{\max}$、$S_{\min}$以及AUPR上的性能指标,最优结果以粗体显示.

SCMAGO模型在BP、MF、CC三方面的预测中均表现出最佳效果:BP方面,$F_{\max}$(0.487)和AUPR(0.507)较第二名方法PANDA-3D分别提升1.6%和6.2%;MF方面,$F_{\max}$(0.739)和AUPR(0.760)比第二名方法DPFunc分别高8.2%和5.3%;CC方面,$F_{\max}$(0.736)和AUPR(0.800)较第二名方法PANDA-3D分别提升3.1%和3.4%. 图2呈现了SCMAGO与同类型深度学习方法DeepGOCNN、DeepFRI、PANDA-3D、DPFunc的AUPR曲线及$F_{\max}$点,直观表现出在蛋白质功能预测任务上,SCMAGO相较于其他模型更具有优势.

表2　SCMAGO方法与六种基线方法在Swiss-Prot测试集上的性能对比

方法	BP			MF			CC		
	$F_{\max}$	$S_{\min}$	AUPR	$F_{\max}$	$S_{\min}$	AUPR	$F_{\max}$	$S_{\min}$	AUPR
Naïve	0.311	49.404	0.227	0.322	11.381	0.197	0.605	12.728	0.525
BLAST	0.407	49.211	0.305	0.567	9.269	0.498	0.534	12.566	0.465
DeepGOCNN	0.387	46.861	0.336	0.469	9.834	0.444	0.658	11.635	0.690
DeepFRI	0.352	48.079	0.257	0.435	9.894	0.305	0.477	12.560	0.377
PANDA-3D	0.471	43.601	0.445	0.642	7.290	0.645	0.705	10.027	0.766
DPFunc	0.419	42.988	0.420	0.657	6.591	0.707	0.673	11.228	0.672
SCMAGO	**0.487**	**38.662**	**0.507**	**0.739**	**5.199**	**0.760**	**0.736**	**9.248**	**0.800**

注:粗体表示对比实验的最优结果.

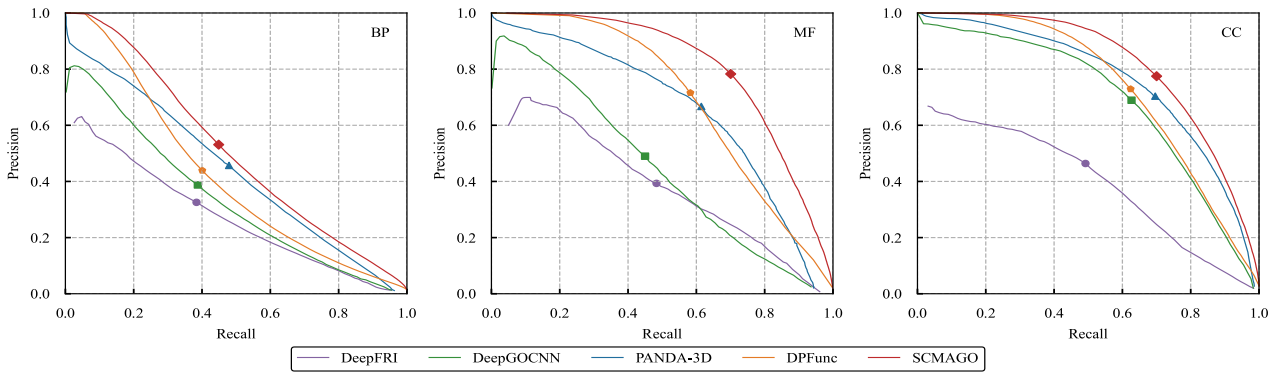


图2　SCMAGO方法与四种同类型方法DeepFRI、DeepGOCNN、PANDA-3D、DPFunc的AUPR曲线和$F_{\max}$点

本研究评估了SCMAGO模型对低同源性蛋白质功能的预测性能，采用CD-HIT工具对测试集中蛋白质序列进行去冗余，序列一致性阈值分别设置为40%、50%、60%和70%，去冗余后的测试集蛋白质数量见表3. 模型预测结果如表4所示，当阈值从95%变为40%时，F_{max}和AUPR在BP方面分别降低0.8%和0.7%；MF方面分别降低2.4%和2.6%；CC方面的两项指标几乎无下降. 图3反映在不同阈值下，SCMAGO模型在BP、MF、CC三个方面的预测性能变化，左、右子图分别为F_{max}和AUPR指标，横轴表示序列一致性阈值，纵轴表示对应指标的预测值. 实验结果表明，即使在面对低同源性蛋白质功能的预测任务中，SCMAGO模型仍能维持稳定的预测准确性，未出现显著性能衰减.

表3 Swiss-Prot测试集蛋白质在不同序列一致性下的数量

GO术语类型	阈值	蛋白质数量
BP	40%	2 571
	50%	3 037
	60%	3 354
	70%	3 627
MF	40%	1 891
	50%	2 255
	60%	2 494
	70%	2 669
CC	40%	2 143
	50%	2 476
	60%	2 716
	70%	2 928

4.5 消融实验

为验证模型设计中各模块对蛋白质功能预测性能的贡献，本研究设计四组消融实验，分别替换或移除关键模块以对比性能差异，消融实验结果如表5所示. 图4使用分组柱状图的形式，系统量化各关键模块对模型整体性能的贡献差异，进一步佐证ESMC、序列编码器及两步跨模态对齐模块在模型预测精度提升中的核心支撑作用. 各组消融方案及分析如下.

表4 SCMAGO方法在Swiss-Prot测试集（不同序列一致性）上的性能

阈值	BP			MF			CC		
	F_{max}	S_{min}	AUPR	F_{max}	S_{min}	AUPR	F_{max}	S_{min}	AUPR
0.4	0.479	35.583	0.500	0.715	5.438	0.734	0.735	9.019	0.799
0.5	0.482	36.375	0.503	0.726	5.238	0.747	0.739	9.057	0.802
0.6	0.485	36.586	0.505	0.731	5.150	0.751	0.739	8.906	0.802
0.7	0.485	36.734	0.506	0.735	5.096	0.756	0.738	8.930	0.802
0.95	0.487	38.662	0.507	0.739	5.199	0.760	0.736	9.248	0.800

（1）蛋白质序列预训练大模型的替换. 将ESMC模型替换为领域内主流的ESM-1b模型，F_{max}在BP方面降低了10.0%，MF方面降低了9.7%，CC方面降低了6.5%；S_{min}各方面均有升高，AUPR各方面同步下降，整体性能下降. ESMC通过扩展训练数据多样性、优化Transformer架构以增强长程依赖建模能力等，显著强化了对序列的语义嵌入能力. 本研究验证了ESMC在序列表征能力和下游任务适配性上的优异表现.

（2）序列编码器的替换. 将序列编码器中的卷积模块替换为两层全连接层组合来进行高阶特征提取与降维，F_{max}在BP方面降低5.6%，MF方面降低8.8%，CC方面降低1.4%；S_{min}呈现一定程度升高，AUPR出现不同幅度下降. 实验结果表明，本研究设计的序列编码器通过一维卷积设计，能精准捕捉连续多个氨基酸构成的功能基序（motif），GLU门控机制进一步实现有效特征筛选，抑制冗余信息干扰，在高阶序列特征提取上优势显著.

（3）序列与结构对齐模块的替换. 将序列与结构对齐模块换成特征拼接方法，F_{max}在BP方面降低2.2%，MF方面降低2.4%，CC方面降低5.9%；S_{min}有小幅上升，

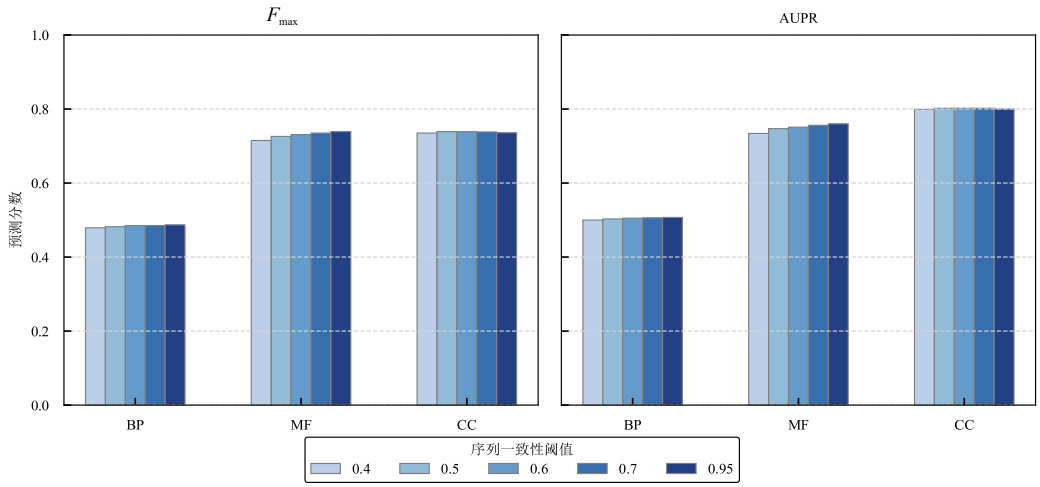


图3 SCMAGO方法在不同序列一致性阈值下的F_{max}和AUPR性能

AUPR对应下降. 这表明序列与结构对齐模块能有效挖掘模态特征关联,融合效果优于无交互的特征拼接策略. 本研究设计的模块采用双向交叉注意力机制,实现序列残基特征与三维结构原子空间位置特征的精准对齐,协同编码蛋白质折叠稳定性等关键信息;特征拼接无法建立残基与结构特征的关联,导致预测准确率下降. CC方面主要依赖三维结构特征,因此降幅达5.9%.

(4)移除家族结构域特征对齐模块. F_{max}在BP方面降低3.2%,MF方面降低3.8%,CC方面降低1.2%;S_{min}略有升高,AUPR也相应下降,且BP和CC方面变化显著. 这表明蛋白质家族结构域信息的引入,对提升模型预测性能具有重要意义. 家族结构域蕴含跨物种的进化保守信息,移除该模块后,模型无法关注与功能相关的核心残基,导致功能特征定位精度下降. MF方面注释更依赖进化保守的功能关联性,在失去家族结构域约束后,预测准确性下降最显著(3.8%);CC方面注释更多依赖三级结构的空间特征,对家族结构域信息依赖度较低,因此降幅最小(1.2%).

表5　消融实验结果

模块	BP			MF			CC		
	F_{max}	S_{min}	AUPR	F_{max}	S_{min}	AUPR	F_{max}	S_{min}	AUPR
w/o ESMC	0.387	44.989	0.373	0.642	7.025	0.680	0.671	11.655	0.738
w/o SeqEncoder	0.431	43.936	0.420	0.651	6.902	0.592	0.722	9.999	0.679
w/o SeqStructFusion	0.465	41.788	0.443	0.715	5.708	0.624	0.677	11.201	0.707
w/o Family&Domain	0.455	42.411	0.465	0.701	5.985	0.737	0.724	9.748	0.746
SCMAGO	0.487	38.662	0.507	0.739	5.199	0.760	0.736	9.248	0.800

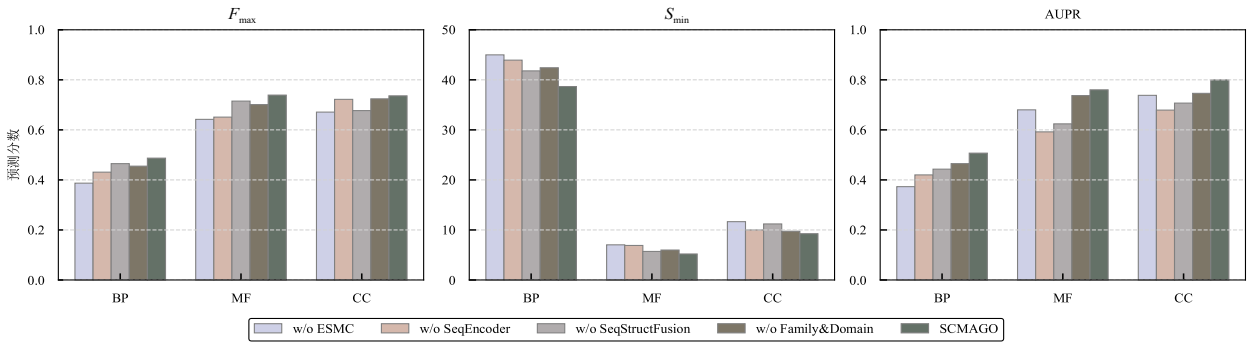


图4　SCMAGO方法中各模块对模型性能的贡献度对比

除模块设计的影响外,蛋白质序列截断长度是模型的关键参数,对基于Transformer架构的模型性能与预测结果影响显著. Transformer架构的时间复杂度为$O(N^2\times d)$. 序列长度变化会通过平方级复杂度制约模型推理效率,因此需在计算效率与预测精度间实现平衡.

图5显示,Swiss-Prot数据集在BP、MF、CC方面的蛋白质序列长度分布均呈以500左右为峰值的右偏分布,仅少量序列长度超过1 500. 基于该分布特征,选取三种截断长度(512、1 024、1 536)开展消融实验. 从表6实验结果可见,1 024长度下模型性能显著优于512长度;而在1 536截断长度下,虽然BP、CC方面的F_{max}和AUPR略有提升,但MF方面的F_{max}降低了1.6%. 因此,综合考虑预测精度与计算效率,选择1 024为序列截断长度.

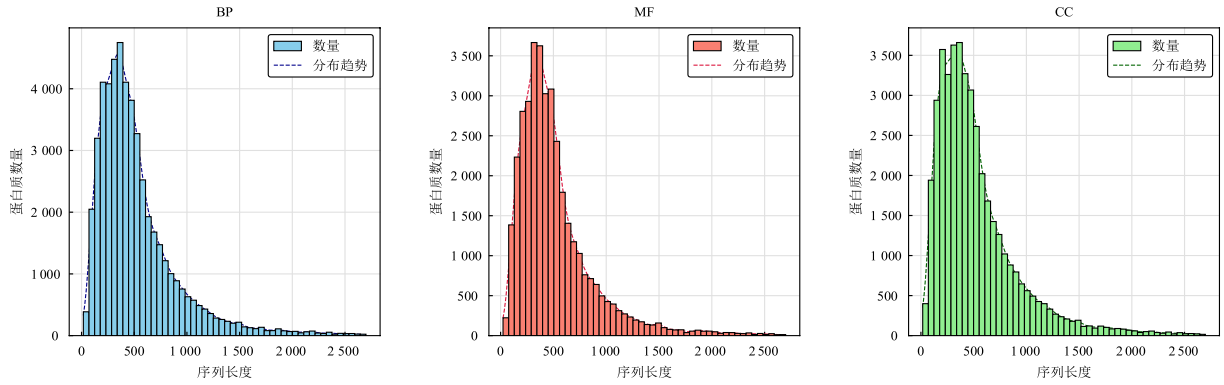


图5　Swiss-Prot数据集在BP、MF、CC方面的蛋白质序列长度分布

表6 SCMAGO方法在不同序列截断长度下的实验结果

序列截断长度	BP			MF			CC		
	F_{max}	S_{min}	AUPR	F_{max}	S_{min}	AUPR	F_{max}	S_{min}	AUPR
512	0.457	42.431	0.466	0.685	6.236	0.750	0.707	10.574	0.782
1 024	0.487	38.662	0.507	0.739	5.199	0.760	0.736	9.248	0.800
1 536	0.494	38.132	0.516	0.723	5.879	0.764	0.739	9.197	0.814

4.6 结果分析

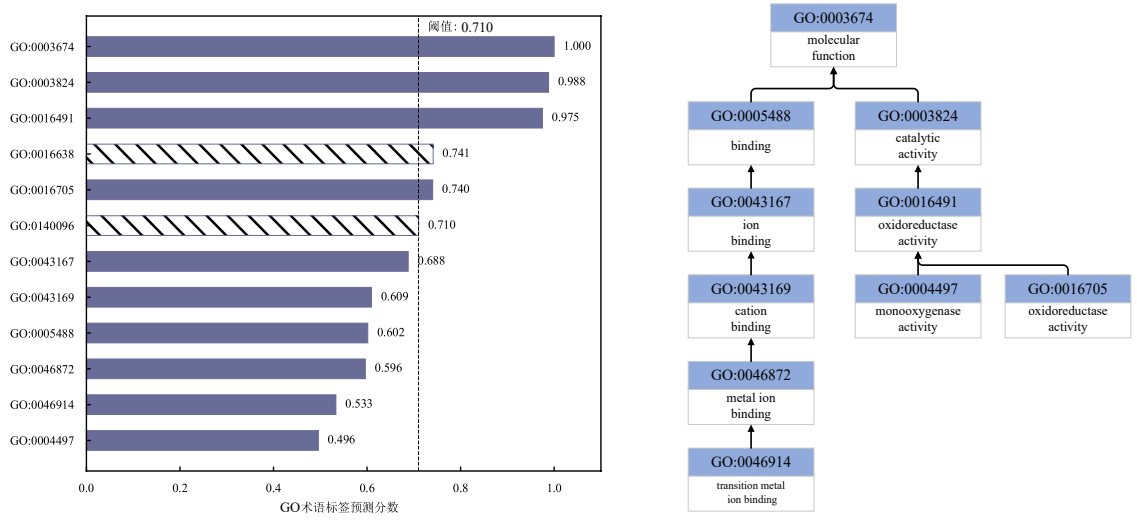
在Swiss-Prot测试集中随机选取蛋白质P15101(多巴胺β-羟化酶)为例,分析模型对该蛋白质在MF方面的预测过程.P15101蛋白质的序列长度为610,表7为InterProScan预测的家族结构域详细信息:

图6(a)表示SCMAGO模型对蛋白质P15101的预测结果,其中虚线柱表示本研究模型已预测到,但是标签中不存在的GO术语;实线柱表示蛋白质真实标签的GO术语及其分数,阈值0.71为模型训练过程中,通过评估验证集性能得到的最佳阈值.图6(b)为标签的有向无环图,可以看出,SCMAGO模型对该蛋白质"结合功能"(binding)的分支预测错误,对其"催化活性"(catalytic activity)的预测较为准确,说明本研究模型能够预测出部分较深层次的GO术语(如"GO:0016705"表示氧化还原酶活性功能,预测分数为0.74).这一结果与蛋白质P15101的家族结构域信息高度契合:该蛋白质中"N端单加氧酶"(IPR000323)等属性决定其核心功能是催化多巴胺发生羟基化反应,生成去甲肾上腺素,这一过程属于氧化还原反应[40].

表7 P15101蛋白质家族结构域信息

类型	InterProID	名称
Domain	IPR000323	N端单加氧酶
	IPR005018	DOMON结构域
	IPR024548	C端单加氧酶
	IPR045266	铜依赖单加氧酶的DOMON结构域
Family	IPR000945	多巴胺β-羟化酶家族
	IPR028460	酪胺/多巴胺β-羟化酶家族
Conserved Site	IPR014783	组氨酸簇2保守位点
	IPR020611	组氨酸簇1保守位点
Homologous Superfamily	IPR008977	PHM/PNGase F超家族
	IPR014787	C端单加氧酶样超家族
	IPR036939	N端单加氧酶超家族

此外,为探究本研究模型的预测效果,从测试集中随机选取蛋白质Q9H4A6、P0A955、P31271及P35327在MF方面的预测结果,如图7所示,图中虚线柱表示标签为0的预测,实心柱表示标签为1的预测;阈值0.71由实验确定,预测分数高于阈值的标签为模型的预测结果.这四个蛋白质在MF方面的功能注释中,分别包含7、9、8和12个GO术语标签.由图7可见,本模型能够预测到各蛋白质的大部分真实标签,但存在一定的假阳性预测:例如Q9H4A6预测到"GO:0043167"和"GO:0043168";P31271则被预测到5个额外标签;对于蛋白质P0A955,模型仅预测到除根节点外的唯一真实标签.结合表8中四个案例蛋白质在MF方面正标签预测置信度数据可见:除蛋白质P0A955外,其余3个蛋白质的正标签平均预测分数均超过0.83,高于阈值的正标



(a) P15101标签预测结果 (b) P15101标签的有向无环图

图6 蛋白质P15101在MF方面的预测结果

签占比达91.7%以上,表明模型对多数蛋白质正标签的预测具有较高置信度.

表8　案例蛋白质在MF方面的正标签预测置信度数据

蛋白质ID	正标签数量	正标签平均预测分数	高于阈值的正标签数量	高于阈值的正标签占比
Q9H4A6	7	0.839	7	100%
P0A955	9	0.615	2	22.2%
P31271	8	0.972	8	100%
P35327	12	0.877	11	91.7%

为进一步验证该结论,本研究对Swiss-Prot测试集进行全面统计,结果如表9所示,分析如下:

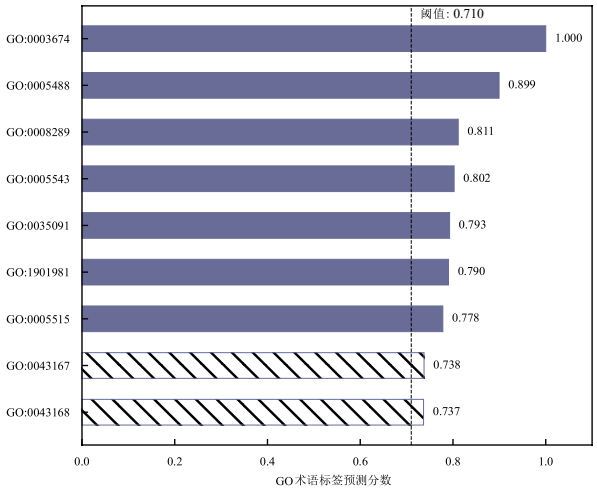
(1)MF和CC两方面的正标签平均预测分数均超过0.75,高于阈值的正标签占比分别为65.14%、61.57%,说明本模型对这两类蛋白质的正标签预测具有稳定的高置信度;

(2)BP方面平均预测分数较低(0.603),主要因为BP方面GO术语标签类别更为复杂,多标签预测难度增加,但仍能保持一定预测精度.

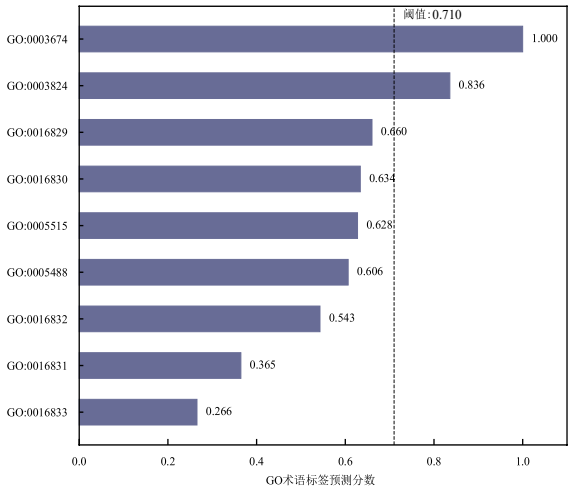
表9　Swiss-Prot测试集上蛋白质的正标签置信度数据

GO术语类型	正标签总数	正标签平均预测分数	高于阈值的正标签数量	高于阈值的正标签占比
BP	191 171	0.603	73 341	38.36%
MF	26 947	0.755	17 552	65.14%
CC	41 772	0.753	25 718	61.57%

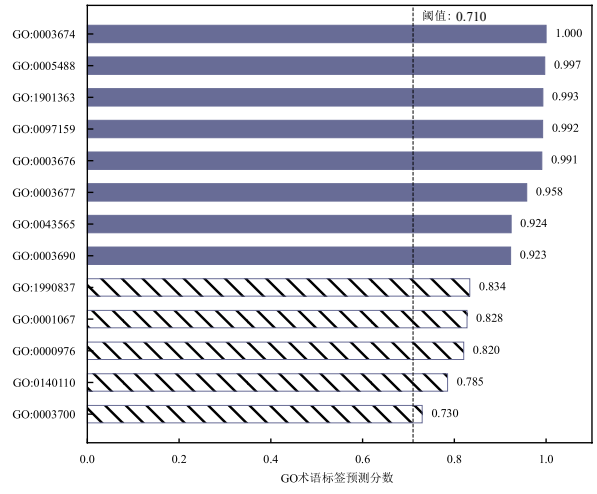
图8中有向无环图为SCMAGO与同类型方法PANDA-3D和DPFunc对四个案例蛋白质在MF方面的预测结果对比,图中字母P、D、S分别表示PANDA-3D、DPFunc和SCMAGO方法,GO术语上方的字母标注表示正确预测到该术语的方法. 术语"GO:0003674"为MF方面的根节点,有向无环图的叶节点为蛋白质最精细的功能注释,因此越接近叶节点,模型对其预测难度越高.



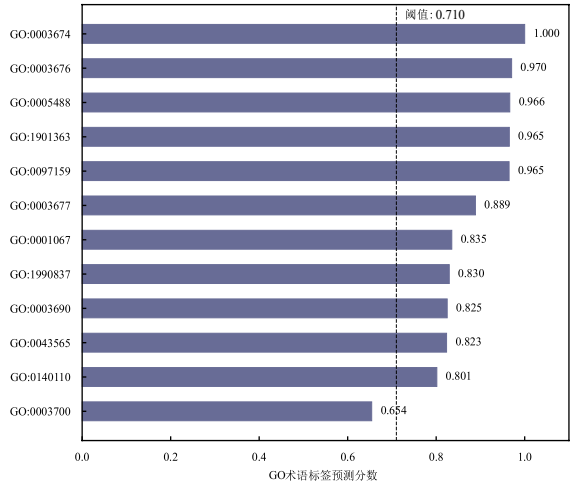
(*a*) Q9H4A6预测结果



(*b*) P0A955预测结果



(*c*) P31271预测结果



(*d*) P35327预测结果

图7　蛋白质Q9H4A6、P0A955、P31271、P35327在MF方面的预测结果

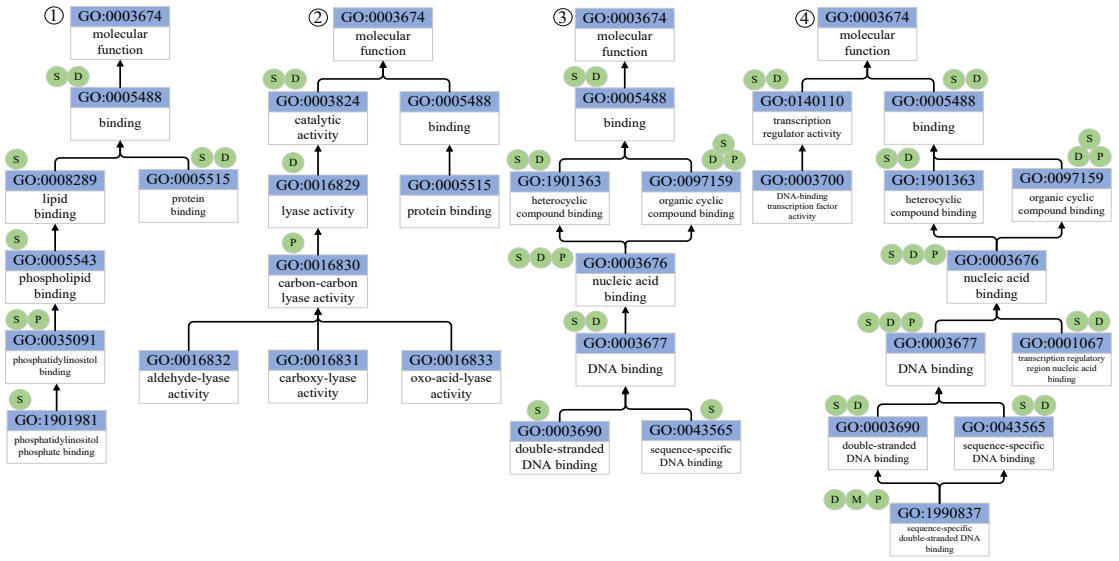


图 8 蛋白质①Q9H4A6、②P0A955、③P31271、④P35327 在 MF 方面标签的有向无环图

对于图 8 中蛋白质① Q9H4A6 和③ P31271，SCMAGO 能预测到全层级功能，而 PANDA-3D 和 DPFunc 方法仅能预测到除叶节点外的节点；针对蛋白质② P0A955，PANDA-3D 和 DPFunc 方法在深层节点预测上的表现优于 SCMAGO；在蛋白质④P35327 的预测上，三种方法效果较为接近，且均未预测到术语“GO:0003700”. 结果表明，SCMAGO 表现出最佳的预测性能，尤其在叶节点预测方面具有优势，这表明本研究提出的特征处理与对齐融合模块能够提取到蛋白质更精细的特征，进而提升对蛋白质复杂功能的识别能力.

5 结束语

本研究提出了一种用于蛋白质功能预测的新型深度学习模型 SCMAGO，通过可靠预测工具 AlphaFold2 和 InterProScan 获取蛋白质的多模态信息，针对异质信息设计不同的特征提取方法和跨模态对齐策略，实现对蛋白质功能的精准预测. 在大规模数据集 Swiss-Prot 的非冗余蛋白质样本上的测试结果表明：SCMAGO 在无需完全依赖实验测定数据的前提下，能够保持 GO 术语预测精度高于现有的 6 种同类型方法. 进一步分析表明，SCMAGO 的性能优势主要来源于三点改进：第一，ESMC 模型的引入，强化了对序列语义的嵌入能力；第二，针对蛋白质的不同特征，设计独特的特征提取模块；第三，在残基层级实现多种特征的精细对齐，使模型充分挖掘各模态特征中与蛋白质功能相关的信息，从而提升预测的准确性和可靠性.

尽管 SCMAGO 模型在预测性能上较现有方法已实现一定突破，但仍存在提升空间，后续工作将围绕两方面重点展开：一是着重提升模型的整体预测精度，突破当前精度水平对算法应用价值的限制；二是增强对低频次 GO 术语的预测能力，这类术语能够定位到更精细的蛋白质功能，其预测准确性对疾病标志物筛选、药物靶点识别等应用场景至关重要.

参考文献

[1] UniProt: The universal protein knowledgebase in 2025[J]. Nucleic Acids Research, 2025, 53(D1): D609-D617.

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

[3] MULDER N, APWEILER R. InterPro and interproscan: Tools for protein sequence classification and comparison[M]// Comparative Genomics. Totowa: Humana Press, 2007: 59-70.

[4] TEAM ESM. ESM Cambrian: Revealing the mysteries of proteins with unsupervised learning[EB/OL]. (2024-12-04)[2025-10-10]. https://www.evolutionaryscale.ai/blog/esm-cambrian.

[5] ALTSCHUL S F, GISH W, MILLER W, et al. Basic local alignment search tool[J]. Journal of Molecular Biology, 1990, 215(3): 403-410.

[6] BHAGWAT M, ARAVIND L. Psi-Blast tutorial[M]//Comparative Genomics. Totowa: Humana Press, 2007: 177-186.

[7] KULMANOV M, KHAN M A, HOEHNDORF R, et al. DeepGO: Predicting protein functions from sequence and interactions using a deep ontology-aware classifier[J]. Bioinformatics, 2018, 34(4): 660-668.

[8] CAO R Z, FREITAS C, CHAN L, et al. ProLanGO: Protein function prediction using neural machine translation based on a recurrent neural network[J]. Molecules, 2017,

22(10): 1732.

[9] SUREYYA RIFAIOGLU A, DOĞAN T, JESUS MARTIN M, et al. DEEPred: Automated protein function prediction with multi-task feed-forward deep neural networks[J]. Scientific Reports, 2019, 9: 7344.

[10] JIANG Y X, ORON T R, CLARK W T, et al. An expanded evaluation of protein function prediction methods shows an improvement in accuracy[EB/OL]. (2016-01-03)[2025-10-10]. https://arXiv.org/abs/1601.00891.

[11] ZHOU N H, JIANG Y X, BERGQUIST T R, et al. The CAFA challenge reports improved protein function prediction and new functional annotations for hundreds of genes through experimental screens[J]. Genome Biology, 2019, 20(1): 244.

[12] KULMANOV M, HOEHNDORF R. DeepGOPlus: Improved protein function prediction from sequence[J]. Bioinformatics, 2020, 36(2): 422-429.

[13] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. Imagenet classifycation with deep convolutional neural networks[C]//Proceedings of the 26th International Conference on Neural Information Processing Systems. New York: Curran Associates Inc, 2012: 1097-1105.

[14] LIU Y W, HSU T W, CHANG C Y, et al. GODoc: High-throughput protein function prediction using novel k-nearest-neighbor and voting algorithms[J]. BMC Bioinformatics, 2020, 21(Suppl 6): 276.

[15] GLIGORIJEVIĆ V, RENFREW P D, KOSCIOLEK T, et al. Structure-based protein function prediction using graph convolutional networks[J]. Nature Communications, 2021, 12: 3168.

[16] MISTRY J, CHUGURANSKY S, WILLIAMS L, et al. Pfam: The protein families database in 2021[J]. Nucleic Acids Research, 2021, 49(D1): D412-D419.

[17] SHERSTINSKY A. Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network[J]. Physica D: Nonlinear Phenomena, 2020, 404: 132306.

[18] KIPF T N, WELLING M. Semi-supervised classification with graph convolutional networks[EB/OL]. (2017-02-22)[2025-10-10]. https://arXiv.org/abs/1609.02907.

[19] BERMAN H M, WESTBROOK J, FENG Z, et al. The protein data bank[J]. Nucleic Acids Research, 2000, 28(1): 235-242.

[20] LAI B Q, XU J B. Accurate protein function prediction via graph attention networks with predicted structure information[J]. Briefings in Bioinformatics, 2022, 23(1): bbab502.

[21] BARANWAL M, MAGNER A, SALDINGER J, et al. Struct2Graph: A graph attention network for structure based predictions of protein-protein interactions[J]. BMC Bioinformatics, 2022, 23(1): 370.

[22] JING B W, EISMANN S, SURIANA P, et al. Learning from protein structure with geometric vector perceptrons[EB/OL]. (2021-05-16)[2024-10-11]. https://arXiv.org/abs/2009.01411.

[23] ZHENG R T, HUANG Z J, DENG L. Large-scale predicting protein functions through heterogeneous feature fusion[J]. Briefings in Bioinformatics, 2023, 24(4): bbad243.

[24] ZHAO C G, LIU T, WANG Z. PANDA-3D: Protein function prediction based on AlphaFold models[J]. NAR Genomics and Bioinformatics, 2024, 6(3): lqae094.

[25] RIVES A, MEIER J, SERCU T, et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences[J]. Proceedings of the National Academy of Sciences of the United States of America, 2021, 118(15): 1-12.

[26] LIN Z M, AKIN H, RAO R, et al. Language models of protein sequences at the scale of evolution enable accurate structure prediction[EB/OL]. (2022-07-21) [2025-10-15]. https://www.biorxiv.org/content/10.1101/2022.07.20.500902v1 utm_campaign=Weekly%20Life%20Science%20Informatics%20News&utm_medium=email&utm_source=Revue%20newsletter.

[27] ZHU Y H, ZHANG C X, YU D J, et al. Integrating unsupervised language model with triplet neural networks for protein gene ontology prediction[J]. PLoS Computational Biology, 2022, 18(12): e1010793.

[28] YUAN Q M, XIE J J, XIE J C, et al. Fast and accurate protein function prediction from sequence through pretrained language model and homology-based label diffusion[J]. Briefings in Bioinformatics, 2023, 24(3): bbad117.

[29] BOADU F, CHENG J L. Improving protein function prediction by learning and integrating representations of protein sequences and function labels[J]. Bioinformatics Advances, 2024, 4(1): vbae120.

[30] WANG W K, SHUAI Y Y, ZENG M, et al. DPFunc: Accurately predicting protein function via deep learning with domain-guided structure information[J]. Nature Communications, 2025, 16: 70.

[31] PAYSAN-LAFOSSE T, BLUM M, CHUGURANSKY S, et al. InterPro in 2022[J]. Nucleic Acids Research, 2023, 51(D1): D418-D427.

[32] DAUPHIN Y N, FAN A, AULI M, et al. Language mod-

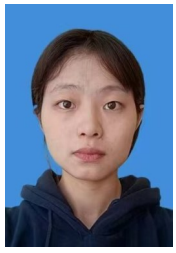
eling with gated convolutional networks[EB/OL]. (2017-09-08)[2025-10-10]. https://arXiv.org/abs/1612.08083.
[33] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[EB/OL]. (2023-08-02)[2025-10-20]. https://arXiv.org/abs/1706.03762.
[34] ALEKSANDER S A, BALHOFF J, CARBON S, et al. The gene ontology knowledgebase in 2023[J]. Genetics, 2023, 224(1): iyad031.
[35] RIDNIK T, BEN-BARUCH E, ZAMIR N, et al. Asymmetric loss for multi-label classification[C]//2021 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE, 2022: 82-91.
[36] UniProt: The universal protein knowledgebase in 2023[J]. Nucleic Acids Research, 2023, 51(D1): D523-D531.
[37] ASHBURNER M, BALL C A, BLAKE J A, et al. Gene Ontology: Tool for the unification of biology[J]. Nature Genetics, 2000, 25(1): 25-29.
[38] VARADI M, ANYANGO S, DESHPANDE M, et al. AlphaFold protein structure database: Massively expanding the structural coverage of protein-sequence space with high-accuracy models[J]. Nucleic Acids Research, 2022, 50(D1): D439-D444.
[39] FU L M, NIU B F, ZHU Z W, et al. CD-HIT: Accelerated for clustering the next-generation sequencing data[J]. Bioinformatics, 2012, 28(23): 3150-3152.
[40] EVANS J P, AHN K, KLINMAN J P. Evidence that dioxygen and substrate activation are tightly coupled in dopamine beta-monooxygenase: Implications for the reactive oxygen species[J]. The Journal of Biological Chemistry, 2003, 278(50): 49691-49698.

作者简介

徐 敏 男,1991年7月出生于安徽省芜湖市. 现为合肥大学人工智能与大数据学院硕士研究生. 主要研究方向为人工智能、生物信息学.
E-mail: xumin@stu.hfuu.edu.cn

胡春玲 女,1970年1月出生于安徽省铜陵市. 现为合肥大学人工智能与大数据学院教授、硕士生导师. 主要研究方向为人工智能、生物信息学.
E-mail: huchunling@hfuu.edu.cn

胡 婷 女,2000年6月出生于安徽省安庆市. 现为合肥大学人工智能与大数据学院硕士研究生. 主要研究方向为药物与靶标亲和力预测.
E-mail: 24085403019@stu.hfuu.edu.cn



张芳芳 女,2002年1月出生于山东省菏泽市. 现为合肥大学人工智能与大数据学院硕士研究生. 主要研究方向为人工智能、蛋白质与RNA相互作用.
E-mail: 24085404032@stu.hfuu.edu.cn

代相龙 男,1999年11月出生于安徽省亳州市. 现为合肥大学人工智能与大数据学院硕士研究生. 主要研究方向为人工智能、蛋白质与RNA相互作用.
E-mail: daixianglong@stu.hfuu.edu.cn