

引用格式: 叶序韬, 图尔逊娜依·艾力, 冯鑫, 等. 基于定制门控系统的活性肽多任务预测模型[J]. 浙江农林大学学报, 2026, 43(X): 1-11. YE Xutao, Tuerxunnayi Aili, FENG Xin, *et al.* A multi-task prediction model for bioactive peptides optimized by a custom gating system[J]. *Journal of Zhejiang A&F University*, 2026, 43(X): 1-11.

基于定制门控系统的活性肽多任务预测模型

叶序韬^{1,2}, 图尔逊娜依·艾力¹, 冯鑫¹, 曹鉴华³, 周飒^{1,2}, 仲飞亮^{1,2}, 罗学刚^{1,2}

(1. 天津科技大学 生物工程学院 工业发酵微生物教育部重点实验室, 天津 300457; 2. 天津科技大学 生物工程国家级实验教学示范中心, 天津 300457; 3. 天津科技大学 人工智能学院, 天津 300457)

摘要: 【目的】针对传统活性肽鉴定筛选周期长、研发成本高, 现有采用单一数据嵌入方式机器学习的肽活性预测模型难以区分同一功能下不同作用靶点的肽段特征等问题, 研究一种组合定制门控单元、对比学习和交叉注意力机制功能的活性肽多任务预测模型。【方法】将蛋白质序列、结构预训练模型、多肽物理化学预测模型与 Transformer 编码器相结合, 并采用专家门控系统优化模型架构、综合靶点结构、序列及物理化学性质信息, 进行活性肽-靶点分类优化和预测。【结果】建立一种新型活性肽多任务预测模型 Biopepformer, 该模型在降血脂生物活性肽靶点分类任务中显示出高整体准确率 (0.899) 与加权 F1 分数 (0.902); 在结合预测二分类任务中 F1 分数=0.958、平均准确率=0.959 和平均曲线下面积 (AUC)=0.952, 性能显著优于传统机器学习模型的效果。进一步消融实验表明, 多任务学习框架与门控多专家单元是模型性能提升的关键。将预训练模型迁移至降血糖生物活性肽数据集进行微调, 在模型靶点结合预测中, F1 分数=0.969、整体准确率=0.969 和平均 AUC=0.976, 在靶点分类任务中, 平均准确率=0.863 和加权 F1 分数=0.884, 展现出跨靶点泛化的潜力。【结论】Biopepformer 可为不同靶点的活性肽挖掘和设计提供高效虚拟筛选工具, 也为功能活性肽的生物活性功能挖掘提供了新的思路。图 6 表 1 参 25

关键词: 活性肽; 多任务深度学习; 蛋白质语言模型; 定制门控系统

中图分类号: TP391; TQ464 **文献标志码:** A **文章编号:** 2095-0756(2026)00-0001-11

A multi-task prediction model for bioactive peptides optimized by a custom gating system

YE Xutao^{1,2}, Tuerxunnayi Aili¹, FENG Xin¹, CAO Jianhua³, ZHOU Sa^{1,2}, ZHONG Feiliang^{1,2}, LUO Xuegang^{1,2}

(1. Key Laboratory of Industrial Fermentation Microbiology of the Ministry of Education, College of Biotechnology, Tianjin University of Science and Technology, Tianjin 300457, China; 2. National Demonstration Center for Experimental Bioengineering Education, Tianjin University of Science and Technology, Tianjin 300457, China; 3. College of Artificial Intelligence, Tianjin University of Science & Technology, Tianjin 300457, China)

Abstract: [Objective] Aiming at the problems of long identification and screening cycles and high research and development costs of traditional bioactive peptides, as well as the limitation that existing machine learning-based peptide activity prediction models adopting a single data embedding method struggle to distinguish the sequence features of peptides with different action targets under the same biological function, this study develops a multi-task bioactive peptide prediction model integrating a customized gating unit, contrastive learning and cross-attention mechanism. [Method] This study combined pre-trained models for protein sequence and structure and peptide physicochemical property prediction models with a Transformer encoder,

收稿日期: 2026-03-16; 修回日期: 2026-06-03

基金项目: 国家级大学生创新训练计划 (202510057002); 国家重点研发计划项目 (2022YFF1103305)

作者简介: 叶序韬 (ORCID: 0009-0008-7979-227X), 从事人工智能交叉的生物应用研究。E-mail: 23044311@mail.tust.edu.cn。通信作者: 罗学刚 (ORCID: 0000-0002-7975-0722), 教授, 博士, 从事微生态健康功能及益生制品研究。E-mail: luoxuegang@tust.edu.cn

adopted an expert gating system to optimize the model architecture, and integrated comprehensive information including target structure, peptide sequence and physicochemical properties to optimize and predict bioactive peptide-target classification. **[Result]** A novel multi-task bioactive peptide prediction model named Biopepformer was developed in this study. In the peptide target classification task, the model achieved an overall accuracy of 0.899 and a weighted F1 score of 0.902 on the hypolipidemic bioactive peptide dataset. In the binary binding prediction task, it obtained an F1-score of 0.958, an average accuracy of 0.959 and an average area under curve (AUC) of 0.952, with its performance significantly outperforming that of traditional machine learning models. Further ablation experiments confirmed that the multi-task learning framework and the gated mixture-of-experts unit were the key contributors to the model's performance improvement. After transferring fine-tuning of the pre-trained model to the hypoglycemic bioactive peptide dataset, the model achieved an F1 score of 0.969, an overall accuracy of 0.969 and an AUC of 0.976 in the target binding prediction task, as well as an average accuracy of 0.863 and a weighted F1-score of 0.884 in the target classification task, demonstrating its potential for cross-target generalization. **[Conclusion]** Biopepformer can serve as an efficient virtual screening tool for the mining and design of bioactive peptides targeting different targets, and also provides a novel approach for the bioactivity exploration of functional bioactive peptides. [Ch, 6 fig. 1 tab. 25 ref.]

Key words: bioactive peptides; multi-task deep learning; protein language model; customized gating system

活性肽是具有特定营养与生理功能的天然活性物质,在功能食品和生物医药领域具有广泛的应用前景^[1-2]。但传统活性肽筛选存在需大量湿实验验证、筛选周期长、研发成本高等问题,而基于机器学习的预测模型可以有效提高功能性肽的筛选效率。如 ESMR4FBP 通过优化 LSTM 模型可提升预测抗氧化肽的准确率^[3]; pLM4ACE 采用多种嵌入方式搭配不同的模型,可确认筛选血管紧张素转换酶抑制肽的最佳模型^[4]; UmamiYYDS 可通过机器学习预测 7 种风味肽并输出相关的分数^[5]; NeuroCL 可将对比学习与交叉注意力结合,用于神经肽识别,性能显著优于传统模型^[6]。在各类模型架构中,多任务预测凭借其独特的优势得到了广泛应用,其核心是在单个模型框架内同步学习多个相关的预测任务,通过共享特征表示实现任务间的信息互补,相较于单任务模型可有效提升模型的泛化能力与学习效率^[7],已成为当前生物信息学领域处理复杂生物数据的主流方法之一。如 GRU4ACE 模型便通过多任务学习框架结合门控单元实现了血管紧张素转化酶抑制肽的高效识别,其平衡准确率可达 0.948,验证了该架构在活性肽预测场景的有效性^[8]。这些模型的应用,有力推动了特定场景下的生物活性肽虚拟筛选,减轻了湿实验负担。随着活性肽研发的深入,同一功能下的多靶点分类成为领域核心需求,即对具备同类生理功能的活性肽,细分其具体作用靶点。以降血脂活性肽为例,其均以调节脂质代谢为核心功能,但作用靶点涵盖 3-羟基-3-甲基戊二酰辅酶 A(3-hydroxy-3-methylglutaryl-coenzyme A reductase, HMG-CoA) 还原酶、前蛋白转化酶枯草杆菌蛋白酶/kexin 9 型(proprotein convertase subtilisin/kexin type 9, PCSK9)、过氧化物酶体增殖物激活受体 α (peroxisome proliferator-activated receptor α , PPAR- α)、胰脂肪酶等多种类型^[9-10];精准的靶点分类可明确活性肽的作用机制,避免表型筛选的机制模糊问题,为功能食品、低副作用候选药物研发提供精准依据,缩小湿实验筛选范围。

精准的活性预测与靶点分类依赖对生物分子的全面特征表征,多模态数据融合为此提供了可行方案。本研究中的多模态指来自不同数据源、具备不同生物学语义的特征类型,包括多肽序列模态、靶点结构模态与多肽理化性质模态;多模态融合可突破单一特征的表征局限,更全面地刻画多肽-靶点配对^[11]。近年来,预训练蛋白语言模型的发展为多模态特征提取与融合提供了技术支撑,多模态模型也已在蛋白功能预测等领域展现出显著优于单模态模型的性能^[12-14]。现有活性肽分析模型在特定场景虽然表现良好,但仍存在标签设置模糊、难以适配同一功能下不同靶点的筛选场景等问题,且这种不同靶点的冗杂数据会导致模型泛化能力下降。同时,单一的输入特征无法涵盖肽的特征信息、作用靶点结构实现数据

融合, 限制了模型精度的提高^[12]。目前尚无研究将多模态融合应用于活性肽靶点分类任务, 也未开发适配同功能多靶点细分筛选的专用模型, 这一空白限制了活性肽虚拟筛选技术的发展。

为了提升生物活性肽的筛选精度, 本研究基于多任务学习的深度学习框架, 集成了门控单元、对比学习和 Transformer 编码器, 开发了 Biopepformer 深度学习模型。与现有的生物活性肽预测模型不同, Biopepformer 具有通过多任务架构实现多模态学习的优势, 融合了多肽序列、作用靶点结构和物理化学性质多维度信息, 可更精准地界定生物活性肽的作用范围, 为生物活性肽功能探究提供新思路。

1 材料与方法

1.1 数据集的选择与处理

采用多肽序列-多肽物理化学性质-靶点蛋白三元组数据, 开展 Biopepformer 的作用预测数据集构建。降血脂生物活性肽数据集通过 BioPep-UWM 数据库检索及文献人工整理获得, 含 5 类共 267 条序列^[15-18]; 靶点蛋白质结构从 PDB 数据库中搜集。为评估模型泛化能力, 通过手动收集了降血糖生物活性肽数据集 (3 类共 591 条序列)^[19]。降血脂肽生物活性包括 HMG-CoA 还原酶抑制剂、PCSK9 抑制剂、PPAR α 激动剂、胰脂肪酶抑制剂和胆固醇酯酶抑制剂^[20]。降血糖肽生物活性包括二肽基肽酶 4(dipeptidyl peptidase 4, DPP4) 抑制剂, 胰高血糖素样肽 1(glucagon-like peptide 1, GLP-1) 类似肽, 和麦芽糖酶-葡糖淀粉酶 (maltase-glucoamylase, MGA) 抑制剂^[21]。所有数据集均采用 5 折分层交叉验证, 数据集详细信息见 https://github.com/XutaoYe/peptide_prediction。

1.2 模型输入计算

使用预训练蛋白质语言模型 (protein language model, pLM) ProtT5, 蛋白质结构嵌入模型 ProTreck, 以及 Toxinpred 对多肽物理化学性质进行预测^[22-23]。其中序列嵌入为 $1 \times 1\,024$ 维向量, 结构嵌入为 $1 \times 1\,024$ 维向量, 多肽性质使用 Toxinpred 预测的 10 类理化性质。同时为了表示生理作用, 在物理化学性质中增加作用维度, 用 0 表示拮抗, 用 1 表示激活。

1.3 模型开发与评估

1.3.1 任务自适应门控多专家特征提取单元 (task adaptive gated multi-expert unit, TA-GMEU) 为解决多任务学习的特征冲突与负迁移问题, 将 TA-GMEU 作为核心特征提取模块, 在多肽、靶点支路各设置 1 个独立单元, 每个单元含 1 个全局共享专家和 2 个任务专属专家, 引入 L2 正则化缓解过拟合, 采用 LeakyReLU 激活函数避免梯度消失^[24]。

1.3.1.1 专家层设计 参考 CACLENS 的多任务专家门控以设计专家层^[12]。对于输入特征 $x \in R^{\text{d}_{\text{model}}}$, 单个专家网络 $\mathcal{E}(x)$ 的计算过程为:

$$h_1 = \text{LayerNorm}(W_1 x + b_1);$$

$$h_2 = \text{LeakyReLU}(h_1, \alpha = 0.01);$$

$$\mathcal{E}(x) = \text{LayerNorm}(W_2 h_2 + b_2)。$$

其中, $W_1, W_2 \in R^{512 \times 512}$ 、 $b_1, b_2 \in R^{512}$ 为可学习参数, $\alpha = 0.01$ 为 LeakyReLU 的负斜率。对 3 个专家网络进行分工, 分别关注不同的特征以实现多任务学习: 全局共享专家学习所有任务通用底层特征; 分类任务专属专家服务于靶点分类任务; 结合预测任务专属专家服务于结合预测核心任务。

1.3.1.2 门控网络与任务自适应权重偏置 引入固定权重偏置强化核心任务的专家特征贡献。基础权重计算公式为:

$$w^{\text{cls}}(x) = \text{Softmax} \left[W_g^{\text{cls}} \text{LeakyReLU} \left(W_{g1}^{\text{cls}} x + b_{g1}^{\text{cls}} \right) + b_g^{\text{cls}} \right]。$$

其中, $W_{g1}^{\text{cls}} \in R^{256 \times 512}$ 、 $W_g^{\text{cls}} \in R^{3 \times 256}$ 为可学习权重参数, $w^{\text{cls}} = [w_{\text{shared}}, w_{\text{cls}}, w_{\text{bind}}]^T$ 为分类任务的基础权重向量, 满足 $w_{\text{shared}} + w_{\text{cls}} + w_{\text{bind}} = 1$ 。引入固定权重偏置, 强化共享专家与分类专属专家的贡献, 调整后的权重为:

$$w_{\text{adj}}^{\text{cls}} = \text{Softmax} \left(w^{\text{cls}} \odot b_{\text{cls}} \right)。$$

其中, $b_{\text{cls}} = [1.2, 0.5, 0.5]$ 为分类任务专属权重偏置, $\odot$ 为哈达玛积。针对靶点预测任务, 门控网络对所

有专家分配权重,通过固定偏置强化共享专家与结合预测任务专属专家的贡献,基础权重计算公式为:

$$w^{\text{bind}}(x) = \text{Softmax} \left[W_g^{\text{bind}} \text{LeakyReLU} \left(W_{gl}^{\text{bind}} x + b_{gl}^{\text{bind}} \right) + b_g^{\text{bind}} \right]。$$

其中, $W_{gl}^{\text{bind}} \in R^{256 \times 512}$ 、 $W_g^{\text{bind}} \in R^{3 \times 256}$ 为可学习权重参数, $w^{\text{bind}} = [w_{\text{shared}}, w_{\text{cls}}, w_{\text{bind}}]^T$ 为结合预测任务的基础权重向量。

结合预测任务同样引入任务自适应权重偏置,调整后的权重为:

$$w_{\text{adj}}^{\text{bind}} = \text{Softmax} \left(w^{\text{bind}} \odot b_{\text{bind}} \right)。$$

其中, $b_{\text{bind}} = [0.3, 1.2, 0.5]^T$ 为结合预测任务专属权重偏置。多肽支路与靶点支路分别通过独立的 TA-GMEU, 最终输出两组任务特异性特征: ①多肽支路。靶点分类任务特征 $H_{\text{pep}}^{\text{cls}}$ 、结合预测任务特征 $H_{\text{pep}}^{\text{bind}}$ 。②靶点支路。靶点分类任务特征 $H_{\text{tar}}^{\text{cls}}$ 、结合预测任务特征 $H_{\text{tar}}^{\text{bind}}$ 。

模型的关键超参数基于预实验结果与相关文献经验进行设置^[12], 具体取值如下: 批次大小为 32, 隐藏层维度为 512, 初始学习率设为 $1e-4$, 采用 AdamW 优化器并设置权重衰减为 $1e-5$; 监督对比损失温度系数 (temperature) 为 0.07, 难样本比例为 0.2; 训练轮次设为 100, 采用基于验证集 F1 分数的早停策略 (patience=10, min delta=0.001)。

1.3.2 基于 Transformer 的双向交叉注意力交互模块 构建了双向交叉注意力交互模块。模块采用两个独立的多头注意力层分别学习两个方向的交互模式, 同时补充了残差连接、层归一化与前馈网络。将 TA-GMEU 单元输出的结合特征扩展序列维度, 构建交叉注意力的查询 (query, Q)、键 (key, K)、值 (value, V) 矩阵。

1.3.2.1 缩放点积注意力 采用 Transformer 标准缩放点积注意力计算, 计算公式为:

$$\text{Attention}(Q, K, V) = \text{Softmax} \left(\frac{QK^T}{\sqrt{d_k}} \right) V。$$

其中, d_k 为单头注意力的维度, 注意力头数 $h=8$ 。

1.3.2.2 多头双向交叉注意力 多头注意力将 Q 、 K 、 V 分别映射到 h 个独立的特征子空间, 分别计算注意力后拼接输出, 实现不同维度交互特征的并行提取, 计算公式为:

$$\text{MHCA}(Q, K, V) = \text{Concat}(\text{head}_1, \text{head}_2, \dots, \text{head}_h) W^O。$$

其中, $W_i^Q, W_i^K, W_i^V \in R^{512 \times 64}$ 、 $W^O \in R^{512 \times 512}$ 为可学习参数。

基于多头注意力机制, 构建双向交叉注意力计算, 分别得到多肽感知靶点的注意力特征, 与靶点感知多肽的注意力特征:

$$H_{\text{pep}}^{\text{attn}} = \text{MHCA}_{\text{pep2tar}}(Q_{\text{pep}}, K_{\text{tar}}, V_{\text{tar}});$$

$$H_{\text{tar}}^{\text{attn}} = \text{MHCA}_{\text{tar2pep}}(Q_{\text{tar}}, K_{\text{pep}}, V_{\text{pep}})。$$

其中, $\text{MHCA}_{\text{pep2tar}}$ 与 $\text{MHCA}_{\text{tar2pep}}$ 为 2 个独立的多头注意力层, 分别学习 2 个方向的交互模式, 避免参数共享导致的特征学习冲突。

1.3.3 多任务学习框架与损失函数设计

1.3.3.1 带多靶点规避的难例挖掘监督对比损失 针对多靶点多肽场景, 在负样本构建中自动规避多肽的所有真实作用靶点, 对于批次内 N 个样本, 首先对特征进行 L2 归一化, 计算样本间的相似度矩阵:

$$\text{Sim}(z_i, z_j) = \frac{z_i z_j^T}{\tau}。$$

其中, z_i, z_j 为 L2 归一化后的特征向量, τ 为温度系数, 本研究设置 $\tau=0.07$ 。

构建正负样本掩码。对于样本 i , 正样本掩码 $\text{pos}_{\text{mask}}(i, j)=1$, 当且仅当样本 i 与样本 j 属于同一靶点类别且 $i \neq j$; 负样本掩码 $\text{neg}_{\text{mask}}(i, j)=1$, 当且仅当样本 i 与样本 j 属于不同类别靶点, 且对应靶点不为多肽的真实作用靶点。引入了难例挖掘策略, 对每个样本 i 选取相似度最高的前 M 个负样本作为难负样本, $M = \max[1, \text{int}(N \times r)]$, 其中, r 为难负样本比例, 本研究设置 $r=0.2$ 。最终的监督对比损失计算公式为:

$$\mathcal{L}_{scl} = -\frac{1}{N_{pos}} \sum_{i=1}^N \sum_{j: pos_{mask}(i,j)=1} \ln \left(\frac{\exp[\text{Sim}(z_i, z_j)]}{\sum_{k: pos_{mask}(i,k)=1} \exp[\text{Sim}(z_i, z_k)] + \sum_{k \in \text{hard}_{neg}(i)} \exp[\text{Sim}(z_i, z_k)]} \right)$$

其中, N_{pos} 为批次内正样本对的总数, $\text{hard}_{neg}(i)$ 为样本 i 的难负样本集合。本研究分别对多肽支路的分类特征 H_{pep}^{cls} 计算多肽分类监督对比损失 $\mathcal{L}_{scl}^{pep}$, 对靶点支路的分类特征 H_{tar}^{cls} 计算靶点分类监督对比损失 $\mathcal{L}_{scl}^{tar}$ 。

总损失为 3 个子任务损失的加权求和, 重点优化结合预测核心任务, 同时通过对比损失辅助提升特征表征能力, 计算公式为:

$$L_{total} = \lambda_{pep} \mathcal{L}_{scl}^{pep} + \lambda_{tar} \mathcal{L}_{scl}^{tar} + \lambda_{bind} L_{lsce}$$

其中, λ_{pep} 、 λ_{tar} 、 λ_{bind} 为损失权重, 本研究分别设置为 0.25、0.25、0.50, 平衡 3 个任务的优化梯度。

1.3.4 模型评估 采用 5 个核心分类指标对模型性能进行 5 折交叉验证评估。这些指标包括准确率, 精确率, 召回率, 加权 F1 分数, ROC-AUC。将 Biopepformer 与现有肽预测模型进行比较, 选取主流的肽性质预测机器学习模型 (XGBoost、随机森林、支持向量机 SVM、LSTM, CNN, RNN, LightGBM) 和目前在肽分类任务上表现较好的 SOTA 模型 (pLM4ACE、peptideBERT) 作为对比模型。其中传统的机器学习模型输入特征为 1024 维序列向量拼接 11 维多肽性质, pLM4ACE 输入特征为基于 ESM2 嵌入的 320 维向量 [13], peptideBERT 使用其训练权重进行微调。

1.3.4.1 消融实验 通过逐一移除模型中的多模态特征、TA-GMEU 单元、监督对比学习模块、双向交叉注意力模块, 对比各模块移除后的模型性能变化, 验证各模块对模型性能的贡献。

1.3.4.2 模型的泛化能力评估 采用降血糖生物活性肽数据集作为迁移学习的对象, 将在降血脂肽数据集上完成预训练的 Biopepformer 模型, 迁移至降血糖生物活性肽数据集进行微调与测试, 测试模型的跨靶点泛化能力。

1.4 应用程序开发

利用 Python 的 Streamlit 库开发了一款用户友好型的肽生物活性预测交互页面, 该应用程序采用了 Biopepformer 模型。这款用户友好型程序允许研究人员上传表格文件进行批量预测, 程序能够预测肽的降血糖活性和降血脂活性, 并且输出可能作用的靶点信息。该文件可在 https://github.com/XutaoYe/peptide_prediction 获取。

2 结果与分析

2.1 Biopepformer 的开发与数据集构建

图 1 为这 2 个数据集的分布情况。为解决负样本数据缺失的问题, 以多肽真实结合靶点为正样本, 将多肽与非其真实靶点的其他靶点配对作为负样本, 通过设定正负样本 1:4 生成负样本。为避免多靶点多肽带来的数据标注错误, 在生成负样本的时候排除了多靶点多肽的所有靶点标签。

Biopepformer 同时完成 2 项任务的端到端训练, 包括多肽-靶点结合活性二分类预测、多肽作用靶点多分类辅助任务。模型的输入为多肽 ProtT5 序列嵌入、基于 Toxinpred 预测的多肽理化特征和靶点蛋白 ProTreck 结构嵌入组成的三元组数据。通过特征融合层完成输入维度的统一映射后, 经 2 条独立的 TA-GMEU 单元分别完成多肽与靶点特征的独立提取, 实现通用生物学特征与靶点预测任务的分离学习。针对靶点预测任务, 通过双向交叉注意力交互模块深度建模多肽与靶点的互作模式, 最终输出靶点分类特征与结合活性预测结果。为避免多任务学习中的负迁移, 采用的 TA-GMEU 单元通过“全局共享专家+任务专属专家”的并行结构, 配合任务门控权重机制, 平衡不同任务的优化梯度。

构建了活性肽高通量筛选工作流: 先将待筛选多肽序列、候选靶点蛋白结构与功能信息输入 Biopepformer; 随后靶点分类塔预测多肽对应的作用靶点, 对候选靶点进行筛选; 结合预测塔评估多肽与靶点的结合活性可行性 (图 2)。

2.2 Biopepformer 在降血脂活性肽预测任务的基准性能

采用 5 折分层交叉验证策略对 Biopepformer 的性能进行无偏评估。模型靶点预测结果如图 3A 所示。在仅正样本的 5 靶点多分类任务中, Biopepformer 整体准确率达 0.899, 加权 F1 分数达 0.902, 展现

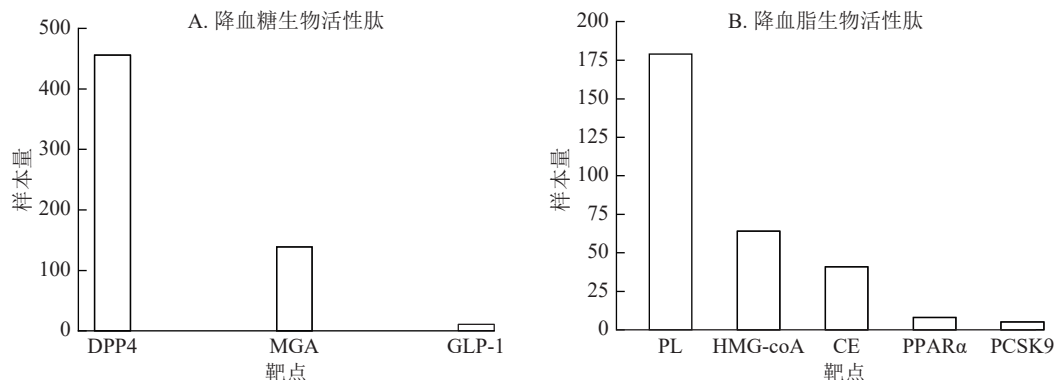
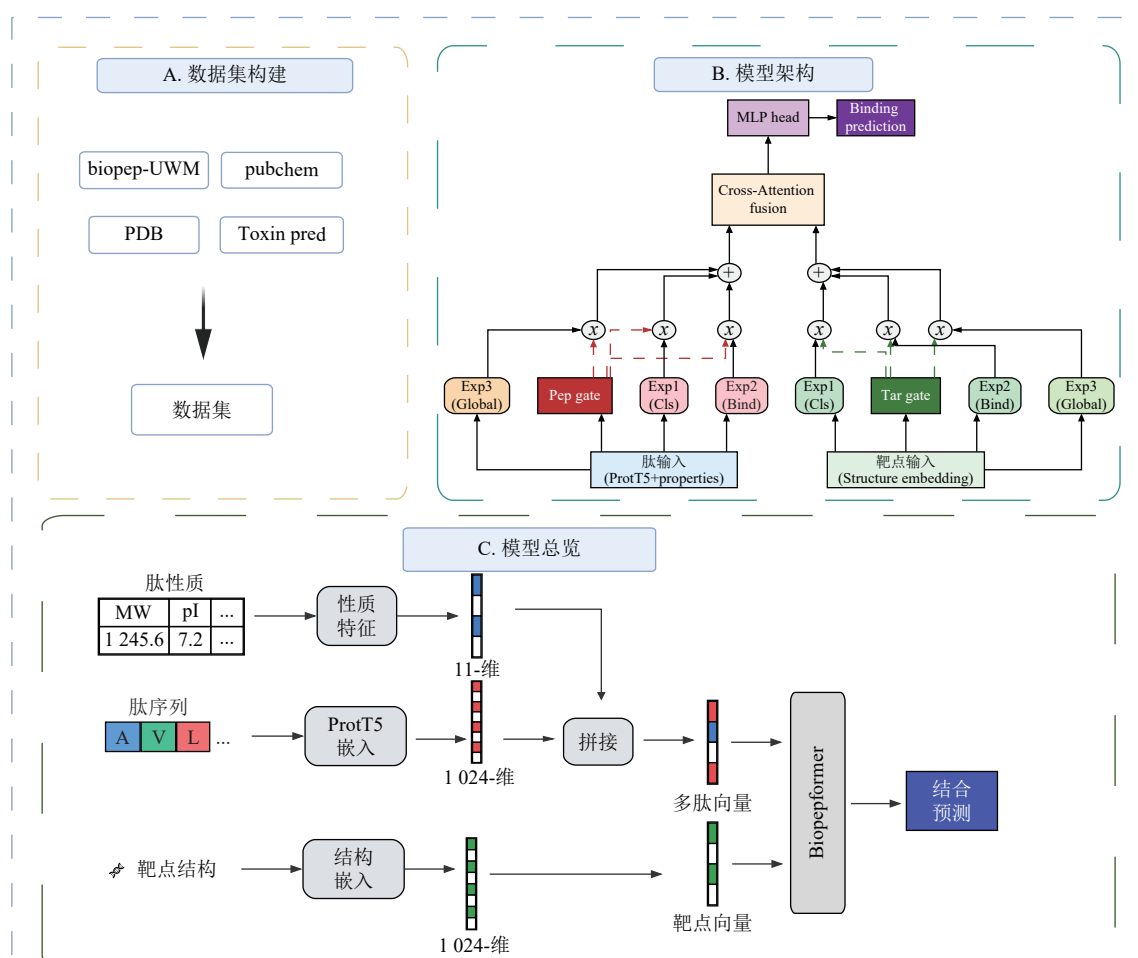


图 1 不同生物活性肽数据集按靶点的分布

Figure 1 Distribution of different bioactive peptide datasets by target



A. 数据集构建, 包含多肽序列, 靶点结构, 和多肽物理化学性质。B. Biopepformer 的多任务框架, 包含门控专家网络和交叉注意力机制。C. Biopepformer 的输入和输出。在蛋白质序列编码后, 将多肽的序列嵌入特征和物理化学性质特征进行拼接。模型同时学习靶点嵌入向量和生物活性肽特征向量。

图 2 Biopepformer 的构建流程

Figure 2 Construction process of Biopepformer

出优异的靶点分类能力 (图 3C)。进一步统计了 5 个靶点的单靶点结合预测准确率, 结果如图 3D 所示。在结合活性预测的二分类任务框架下, 模型整体预测准确率达 0.959, 加权 F1 分数达 0.958, 高于纯序列多分类任务的性能, 证实了多任务学习框架对靶点分类能力的增益。与纯序列多分类结果一致, 模型对样本量充足、序列特征显著的 PL、HMG-coA 靶点表现出最优的预测性能, 而对小样本的 PCSK9 靶点仍保持了稳定的分类能力, 验证了模型在小样本场景下的鲁棒性。各折次损失曲线快速收敛且无明显训练-验证差距, 最终平均验证损失低至 0.381; 验证准确率稳定在 0.940~0.970; 5 折平均 AUC 达 0.952,

表明模型对活性肽-靶点的良好匹配能力 (图 3B、E、F)。

2.3 Bioepformer 在多肽-靶点匹配预测上的性能对比

Bioepformer 在所有评估指标上均显著优于所有基线模型 (图 4)。其性能优势主要源于基于 Transformer 的架构设计。与传统机器学习模型依赖手工提取的理化特征不同, Bioepformer 通过大规模预训练学习到了多肽序列与靶点结构之间的深层交互模式, 其自注意力机制能够有效捕捉多肽与靶点的关键信息, 从而更精准地建模关联信息。相比之下, 传统机器学习模型 (如随机森林、SVM) 受限于单

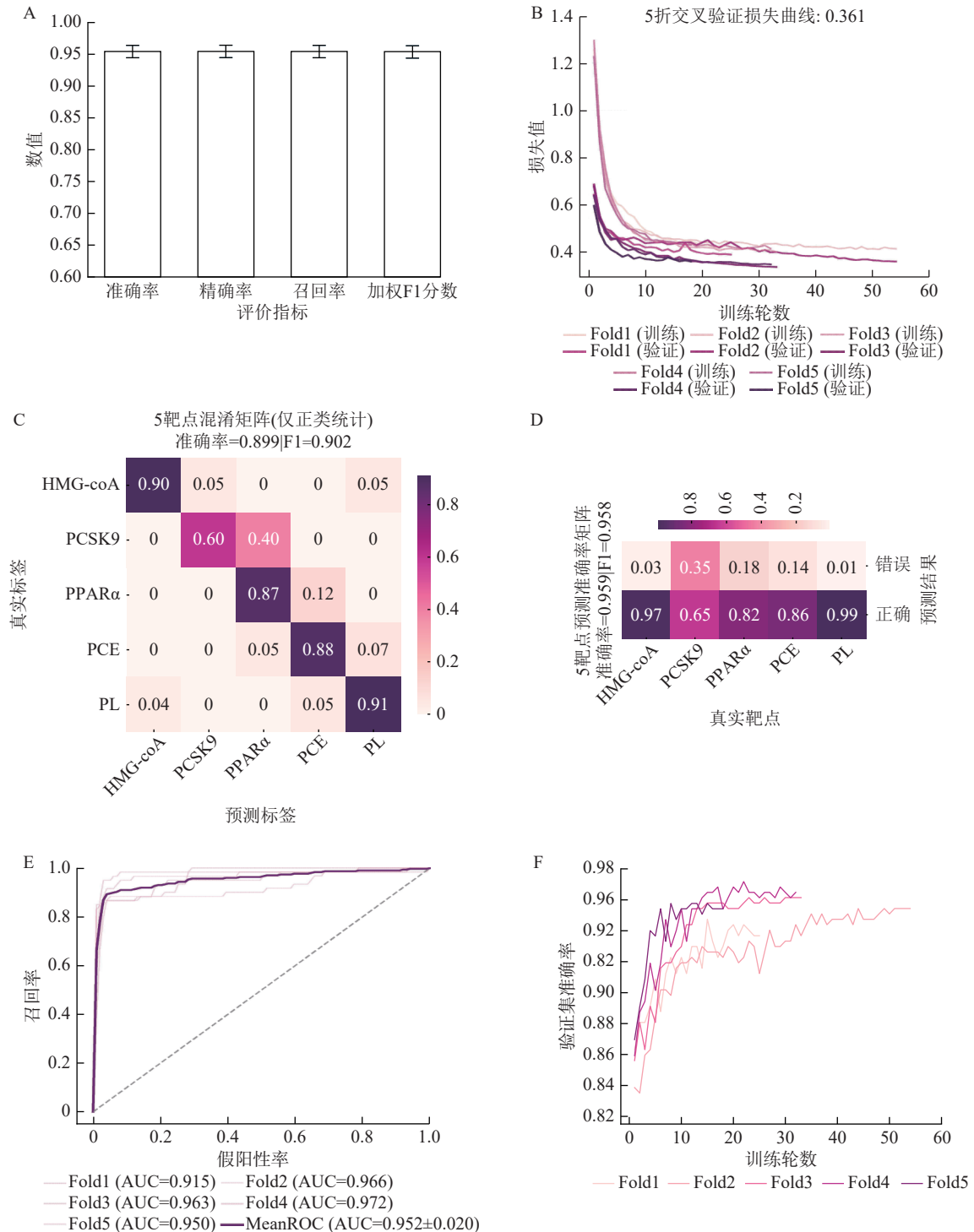


图 3 模型在 5 折交叉验证下对模型的性能评估

A. 模型核心评价指标柱状图, 包含准确率、精确率、召回率与加权F1分数。B. 5折交叉验证的训练与验证损失曲线。C. 仅正样本的5靶点多分类混淆矩阵。D. 靶点预测准确率矩阵。E. 5折交叉验证的ROC曲线, 横轴为假阳性率, 纵轴为真阳性率, 灰色虚线为随机分类基线, 其中AUC指曲线下面积(area under curve)。F. 5折交叉验证的验证准确率曲线。

图 3 模型在 5 折交叉验证下对模型的性能评估

Figure 3 Performance evaluation of the model under 5-fold cross-validation

一特征嵌入,难以充分挖掘复杂的生物分子交互信息;而深度学习模型(如 LSTM、CNN)在处理长程依赖关系和全局特征提取方面仍存在不足。目前的 SOTA 模型, pLM4ACE 仅使用 ESM2 对序列进行嵌入,输入到传统的机器学习模型中进行下游分类任务。最优模型逻辑回归仅针对二分类场景的线性边界优化,无法建模多靶点对应的高维非线性类别边界,在多分类任务中出现严重的特征区分度不足。而 PeptideBERT 虽然是基于 Transformer 架构构建,依赖大规模微调数据才能避免过拟合、发挥性能优势。本研究的降血脂肽数据集仅 300 余条样本,属于典型的小样本场景。PeptideBERT 在微调过程中极易出现过拟合,无法有效学习多类决策边界。

2.4 Biopepformer 核心模块的消融实验分析

为量化 Biopepformer 各模块对多肽-靶点结合活性预测性能的贡献,进行了消融实验。检验了交叉验证中测试集与训练集的分布情况,以避免数据泄露。特征工程消融实验表明,多模态融合策略是靶点配对任务的关键,将靶点蛋白结构信息与多肽信息同时传入模型参与损失,可以显著提升靶点分类性能(表 1)。模块消融实验表明:多任务学习框架和 TA-GMEU 为性能核心驱动因素(图 5)。移除多任务学习框架(MultiTask-Off)后,模型平均加权 F1 分数降至 0.932,为所有消融变体最低,证明该框架引导模型学习更具生物学意义的通用特征。移除 TA-GMEU(GME-Off)后,模型平均加权 F1 分数从 0.936 降至 0.935,性能标准差从 0.0141 升至 0.0236,训练稳定性下降,证实该单元通过独立学习通用和任务特异性特征,解决多任务学习的特征冲突与负迁移问题。值得注意的是,移除监督对比学习模块(SupCon-Off)后,模型的性能反而上升。初步推测是因为降血脂肽数据集为小样本多分类场景,该模块移除后模型可聚焦核心任务优化。整体消融实验明确多模态融合、独立特征提取与多任务联合学习,是提升现有生物活性肽预测模型性能瓶颈的关键。

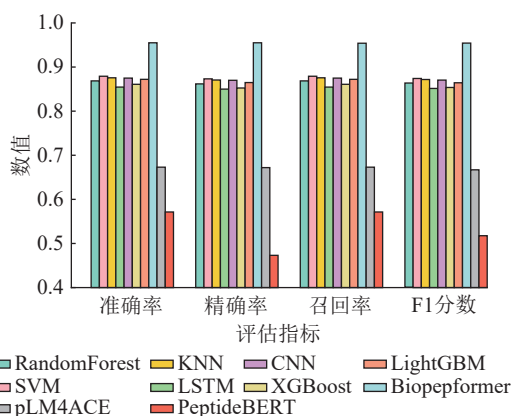


图 4 Biopepformer 与目前主流的生物活性肽判别器的性能对比

Figure 4 Performance comparison between Biopepformer and current mainstream bioactive peptide discriminators

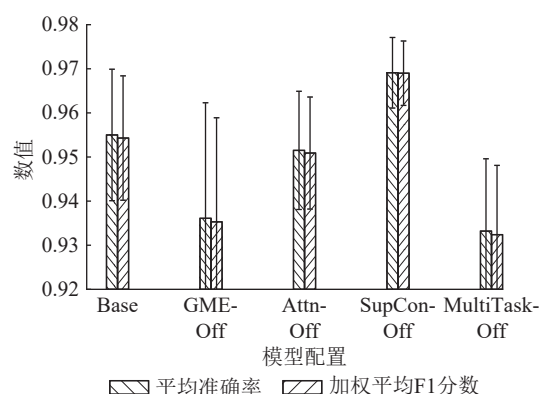


图 5 Biopepformer 核心模块消融实验性能对比

Figure 5 Performance comparison of ablation experiments on the core modules of Biopepformer

表 1 特征工程消融实验结果

Table 1 Results of feature engineering ablation experiments

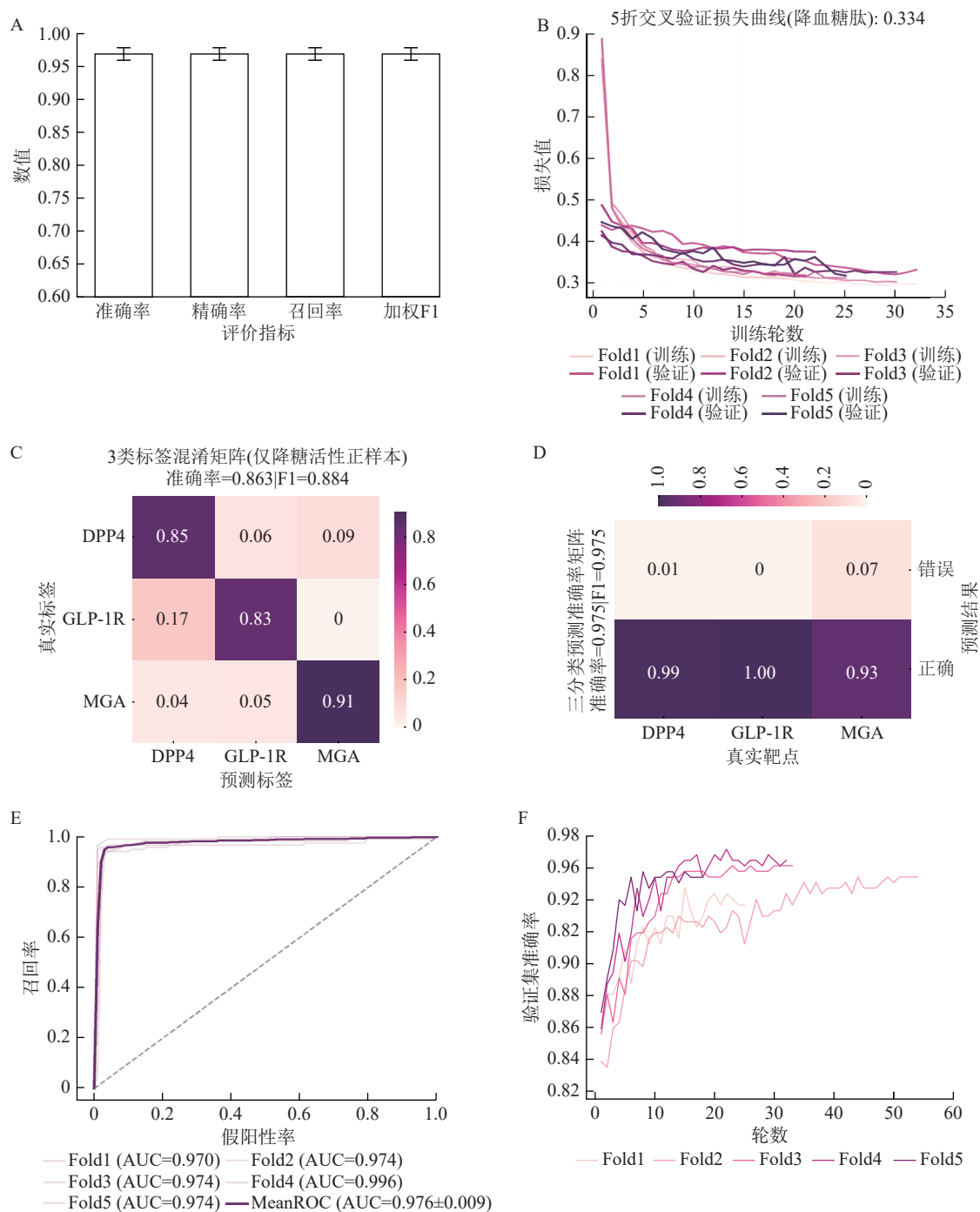
实验组	任务类型	准确率	F1分数	精确率	召回率
无物化性质	结合预测	0.927±0.016	0.926±0.016	0.926±0.016	0.927±0.016
无ProT5序列嵌入	结合预测	0.920±0.024	0.918±0.026	0.919±0.0256	0.920±0.024
无靶点嵌入	靶点预测	0.576±0.036	0.556±0.035	0.542±0.036	0.576±0.036

说明: 数值为平均值±标准差。

2.5 模型在其他生物活性肽数据集上的泛化能力评估

为评估 Biopepformer 的泛化能力,将其迁移至样本量更大的降血糖活性肽数据集进行微调。结果表明模型在该数据集上的准确率、精确率、召回率与加权 F1 分数均有提升(图 6A)。5 折交叉验证显示训

训练损失与验证准确率平稳收敛(图 6B、F)。针对 DPP4、GLP-1 受体、MGA 等 3 类靶点, 模型多类别分类表现优异。整体分类准确率达 0.863, 加权 F1 分数为 0.884, 靶点结合预测任务的准确率和 F1 分数均为 0.975(图 6C、D)。ROC 曲线进一步证实模型具备极强的样本判别能力, 5 折平均 AUC 高达 0.976 ± 0.009 (图 6E), 初步验证了模型在更大生物活性肽数据集上的良好泛化潜力。



A. 模型核心评价指标柱状图, 包含准确率、精确率、召回率与加权F1分数。B. 5折交叉验证的训练与验证损失曲线。C. 仅正样本的5靶点多分类混淆矩阵。D. 靶点预测准确率矩阵。E. 5折交叉验证的ROC曲线, 横轴为假阳性率, 纵轴为真阳性率, 灰色虚线为随机分类基线, 其中AUC为折线下面积(area under curve)。F. 5折交叉验证的验证准确率曲线。

图 6 Biopeformer 在降血糖生物活性肽数据集的 5 折交叉验证性能评估

Figure 6 Performance evaluation of Biopeformer on the hypoglycemic bioactive peptide dataset using 5-fold cross-validation

3 讨论

靶向同一生物功能但作用于不同分子靶点的活性肽细分筛选, 是阐明作用机制并降低脱靶风险的关键

键前提。已有研究多聚焦于单一靶点的二分类预测或基于序列的多类识别,未整合靶点结构信息,导致模型难以区分功能相似但结合口袋不同的肽段特征^[4]。多任务学习与多模态融合为突破这一局限提供了可能,但现有架构面临特征冲突与负迁移挑战^[12]。本研究通过构建任务自适应门控多专家单元与双向交叉注意力机制,首次实现了多肽序列、靶点结构与理化性质三模态信息的联合学习,并在同一框架下同步完成结合活性预测与靶点分类任务^[25]。

本研究首次将多肽序列、靶点结构与理化性质三类模态信息纳入统一的多任务学习框架,同步完成靶点分类与结合活性预测。在降血脂活性肽数据集上,模型靶点分类整体准确率达 0.899、加权 F1 分数为 0.902,结合预测 F1 分数为 0.958,平均 AUC 为 0.952。消融实验证实,多模态融合是该性能的核心支撑。移除靶点结构嵌入后靶点分类准确率骤降至 0.576,表明仅凭序列信息无法有效分离同功能异靶点的肽段特征,而结构信息的引入让模型得以捕捉与靶点之间的潜在关联。

任务自适应门控多专家单元通过“共享专家+任务专属专家”的解耦设计,缓解了多任务学习中的特征冲突。移除该单元后模型训练稳定性下降。监督对比损失在小样本多靶点场景下被移除后性能反而小幅上升,提示严格对比可能放大类内噪声,在数据稀缺时需谨慎设计辅助目标。与现有方法相比,传统机器学习模型受限于一维嵌入而无法学习多靶点高维决策边界。Biopepformer 则凭借多模态信息互补与自适应任务平衡,在所有指标上大幅领先,表明了合理架构对抑制小样本过拟合的决定性作用。

本研究建立了多任务分类模型 Biopepformer。该模型整合了多任务学习框架、TA-GMEU 单元等核心模块,实现了多肽与靶点多模态数据融合,同步完成了结合活性预测与靶点分类任务,在降血脂、降血糖数据集上表现优异。这些结果共同表明,多模态融合与门控多专家多任务架构是突破同功能多靶点精细分类瓶颈的有效策略,为功能活性肽的高通量虚拟筛选和作用机制解析提供了新的技术路线。

4 结论

本研究基于多任务学习框架,集成 TA-GMEU 单元、双向交叉注意力与对比学习策略,开发了 Biopepformer 多任务预测模型,实现了多肽多模态信息融合及双任务同步预测。Biopepformer 在降血脂肽数据集上性能显著优于主流模型。消融实验证实,多模态数据融合是模型性能优异的关键。迁移学习结果表明,该模型具有在更大数据集上的泛化潜力。综上所述,多模态数据融合与独立特征提取是提升活性肽预测精度的有效策略,可为活性肽虚拟筛选及功能研究提供新的思路。虽然 Biopepformer 展现出优异的鲁棒性,但其预测精度仍受限于较小的数据集规模,尤其对小样本靶点,且尚未整合分子对接或蛋白质相互作用网络等更深层信息。未来研究将扩展多肽-靶点互作数据库,纳入额外的生物学模态并探索预训练策略,以进一步提升模型的泛化性与可解释性。

5 参考文献

- [1] DA Jiayi, *et al.* Postbiotic-mediated obesity reduction via bioactive peptide ETLVK and gut microbiota in HFD mice[J]. *Journal of Agricultural and Food Chemistry*, 2025, **73**(36): 22436–22447. DOI: 10.1021/acs.jafc.5c07319.
- [2] LAFIA A T, GOMES W P C, MELCHERT W R. Bioactive peptides from whey proteins: production, purification, identification, mechanism of action and biofunctionality in foods[J]. *International Dairy Journal*, 2026, **177**: 106595. DOI: 10.1016/j.idairyj.2026.106595.
- [3] ZHANG Ruihao, LI Yonghui, JIANG Qinbo, *et al.* ESMR4FBP: a pLM-based regression prediction model for specific properties of food-derived peptides optimized multiple bionic metaheuristic algorithms[J]. *Food Chemistry*, 2025, **464**: 141840. DOI: 10.1016/j.foodchem.2024.141840.
- [4] DU Zhenjiao, DING Xingjian, HSU W, *et al.* pLM4ACE: a protein language model based predictor for antihypertensive peptide screening[J]. *Food Chemistry*, 2024, **431**: 137162. DOI: 10.1016/j.foodchem.2023.137162.
- [5] CUI Zhiyong, ZHANG Zhiwei, ZHOU Tianxing, *et al.* A TastePeptides-Meta system including an umami/bitter classification model Umami_YYDS, a TastePeptidesDB database and an open-source package Auto_Taste_ML[J]. *Food Chemistry*, 2023, **405**: 134812. DOI: 10.1016/j.foodchem.2022.134812.
- [6] LIU Jian, GENG Aoyun, CUI Feifei, *et al.* NeuroCL: a deep learning approach for identifying neuropeptides based on

- contrastive learning[J]. *Analytical Biochemistry*, 2025, **705**: 115920. DOI: [10.1016/j.ab.2025.115920](https://doi.org/10.1016/j.ab.2025.115920).
- [7] MA Chang, LIU Runcheng, LU Jiarui, *et al.* PEER: a comprehensive and multi-task benchmark for protein sequence understanding[C]//*Advances in Neural Information Processing Systems 35*. Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2022: 35156–35173. DOI: [10.52202/068431-2548](https://doi.org/10.52202/068431-2548).
- [8] AHMED S, SCHADUANGRAT N, CHUMNANPUEN P, *et al.* GRU4ACE: enhancing ACE inhibitory peptide prediction by integrating gated recurrent unit with multi-source feature embeddings[J]. *Protein Science*, 2025, **34**(6): e70026. DOI: [10.1002/pro.70026](https://doi.org/10.1002/pro.70026).
- [9] BALLANTYNE C M, NORATA G D. The evolving landscape of targets for lipid lowering: from molecular mechanisms to translational implications[J]. *European Heart Journal*, 2025, **46**(44): 4737–4750. DOI: [10.1093/eurheartj/ehaf606](https://doi.org/10.1093/eurheartj/ehaf606).
- [10] MALICK W A, CHOI D, LANGSTED A, *et al.* Challenges in achieving LDL cholesterol targets and novel approaches to lipid lowering[J]. *European Journal of Preventive Cardiology*, 2025, **32**(13): 1136–1144. DOI: [10.1093/eurjpc/zwaf123](https://doi.org/10.1093/eurjpc/zwaf123).
- [11] HU Fan, HU Yishen, ZHANG Weihong, *et al.* A multimodal protein representation framework for quantifying transferability across biochemical downstream tasks[J]. *Advanced Science*, 2023, **10**(22): 2301223. DOI: [10.1002/advs.202301223](https://doi.org/10.1002/advs.202301223).
- [12] YI Xilong, TAN Yingzhu, LIN Huikang, *et al.* CACLENS: a multitask deep learning system for enzyme discovery[J]. *Advanced Science*, 2026, **13**(9): e18063. DOI: [10.1002/advs.202518063](https://doi.org/10.1002/advs.202518063).
- [13] LIN Zeming, AKIN H, RAO R, *et al.* Evolutionary-scale prediction of atomic-level protein structure with a language model[J]. *Science*, 2023, **379**(6637): 1123–1130. DOI: [10.1126/science.ade2574](https://doi.org/10.1126/science.ade2574).
- [14] ELNAGGAR A, HEINZINGER M, DALLAGO C, *et al.* ProtTrans: toward understanding the language of life through self-supervised learning[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022, **44**(10): 7112–7127. DOI: [10.1109/TPAMI.2021.3095381](https://doi.org/10.1109/TPAMI.2021.3095381).
- [15] MINKIEWICZ P, IWANIAK A, DAREWICZ M. BIOPEP-UWM database of bioactive peptides: current opportunities[J]. *International Journal of Molecular Sciences*, 2019, **20**(23): 5978. DOI: [10.3390/ijms20235978](https://doi.org/10.3390/ijms20235978).
- [16] LI Jianqiang, BOLLATI C, D'ADDUZIO L, *et al.* Food-derived peptides with hypocholesterolemic activity: production, transepithelial transport and cellular mechanisms[J]. *Trends in Food Science & Technology*, 2024, **143**: 104279. DOI: [10.1016/j.tifs.2023.104279](https://doi.org/10.1016/j.tifs.2023.104279).
- [17] KIM S, CHEN Jie, CHENG Tiejun, *et al.* PubChem in 2021: new data content and improved web interfaces[J]. *Nucleic Acids Research*, 2021, **49**(D1): D1388–D1395. DOI: [10.1093/nar/gkaa971](https://doi.org/10.1093/nar/gkaa971).
- [18] TOMBLING B, ZHANG Yuhui, HUANG Yenhua, *et al.* The emerging landscape of peptide-based inhibitors of PCSK9[J]. *Atherosclerosis*, 2021, **330**: 52–60. DOI: [10.1016/j.atherosclerosis.2021.06.903](https://doi.org/10.1016/j.atherosclerosis.2021.06.903).
- [19] HU Gui'e, QIN Yebi, QU Mei, *et al.* Hypoglycemic effects and mechanisms of animal-derived blood peptides in C2C12 myotubes and alloxan-induced diabetic mice[J]. *Journal of Functional Foods*, 2026, **136**: 107134. DOI: [10.1016/j.jff.2025.107134](https://doi.org/10.1016/j.jff.2025.107134).
- [20] HAMADOU M. Bioactive peptides and metabolic health: a mechanistic review of the impact on insulin sensitivity, lipid profiles, and inflammation[J]. *Applied Food Research*, 2025, **5**(2): 101056. DOI: [10.1016/j.afres.2025.101056](https://doi.org/10.1016/j.afres.2025.101056).
- [21] SU Jingqian, LUO Yingsheng, HU Shan, *et al.* Advances in research on type 2 diabetes mellitus targets and therapeutic agents[J]. *International Journal of Molecular Sciences*, 2023, **24**(17): 13381. DOI: [10.3390/ijms241713381](https://doi.org/10.3390/ijms241713381).
- [22] RATHORE A S, CHOUDHURY S, ARORA A, *et al.* ToxinPred 3.0: an improved method for predicting the toxicity of peptides[J]. *Computers in Biology and Medicine*, 2024, **179**: 108926. DOI: [10.1016/j.compbimed.2024.108926](https://doi.org/10.1016/j.compbimed.2024.108926).
- [23] SU Jin, ZHOU Xibin, ZHANG Xuting, *et al.* ProTrek: Navigating the protein universe through Tri-Modal contrastive learning[J]. *bioRxiv*, 2024. DOI: [10.1101/2024.05.30.596740](https://doi.org/10.1101/2024.05.30.596740).
- [24] TANG Hongyan, LIU Junning, ZHAO Ming, *et al.* Progressive layered extraction (PLE): a novel multi-task learning (MTL) model for personalized recommendations[C]//*Fourteenth ACM Conference on Recommender Systems*. ACM, 2020: 269–278. DOI: [10.1145/3383313.3412236](https://doi.org/10.1145/3383313.3412236).
- [25] SU Liangcai, PAN Junwei, WANG Ximei, *et al.* STEM: unleashing the power of embeddings for multi-task recommendation[J]. *Proceedings of the 38th AAAI Conference on Artificial Intelligence*, 2024: 9002–9010. DOI: [10.1609/aaai.v38i8.28749](https://doi.org/10.1609/aaai.v38i8.28749).