计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
2026,35(3):32−43 [doi: 10.15888/j.cnki.csa.010103] [CSTR: 32024.14.csa.010103]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

T-SeGAT: 面向不平衡数据的分子性质及 CPI 预测模型①

石全宾[1], 吴　萌[2], 何芊平[1], 王亚琪[1]

[1](西安建筑科技大学 信息与控制工程学院, 西安 710055)
[2](西安建筑科技大学 交叉创新研究院, 西安 710055)
通信作者: 吴　萌, E-mail: wumeng@xauat.edu.cn

摘　要: 分子性质预测和化合物-蛋白质相互作用 (CPI) 预测是药物发现中的关键环节, 但传统图卷积网络 (GCN) 受限于局部感受野, 难以充分捕捉化学结构复杂性、分子构象动态变化以及长程电子相互作用等信息, 预测性能存在瓶颈. 为解决这一问题, 本文提出了一种深度学习模型 T-SeGAT, 用于提升分子性质和 CPI 预测的准确性与泛化能力. 该模型融合了 ESM-2 蛋白语言模型、ChemBERTa 分子语言模型以及基于图注意力网络 (GAT) 与 Set2Set 的图神经网络, 实现从序列到结构的多层次特征提取与融合. 同时, 针对实验数据的不平衡问题, 模型在数据加载、损失计算和预测决策这 3 个层面引入加权随机采样、平衡/焦点/自适应损失函数以及动态阈值搜索机制, 并结合基于 AUC 差值的过拟合抑制方法、早停策略和学习率调度, 提升训练稳定性与泛化能力. 本文在 BACE、P53 和 hERG 数据集上进行分子性质预测实验, 在 Human 和 C. elegans 数据集上进行 CPI 预测实验, 均采用分层 5 折交叉验证进行性能评估. 实验结果表明, T-SeGAT 在所有数据集上均优于现有基线模型, 其中在 BACE 和 hERG 数据集上, AUC 和精确率分别较次优模型提升 0.022、0.010 和 0.004、0.022, 在 Human 数据集上的精确率提升 0.013. 综合实验结果表明, T-SeGAT 在精度、稳定性和实用性方面表现出显著优势, 为药物发现过程中的分子性质预测与 CPI 预测提供了有力支持.

关键词: 图神经网络; 分子性质预测; 化合物-蛋白质相互作用; 不平衡数据学习; 多头注意力机制; 深度学习

引用格式: 石全宾,吴萌,何芊平,王亚琪.T-SeGAT: 面向不平衡数据的分子性质及 CPI 预测模型.计算机系统应用,2026,35(3):32–43. http://www.c-s-a.org.cn/1003-3254/10103.html

T-SeGAT: Molecular Property and CPI Prediction Model for Imbalanced Data

SHI Quan-Bin[1], WU Meng[2], HE Qian-Ping[1], WANG Ya-Qi[1]

[1](College of Information and Control Engineering, Xi'an University of Architecture and Technology, Xi'an 710055, China)
[2](Institute for Interdisciplinary Innovation Research, Xi'an University of Architecture and Technology, Xi'an 710055, China)

Abstract: Molecular property prediction and compound-protein interaction (CPI) prediction are key steps in drug discovery. However, traditional graph convolutional network (GCN) are limited by local receptive fields and cannot fully capture the complexity of chemical structures, dynamic changes of molecular conformation, and long-range electronic interactions, which causes bottlenecks to the prediction performance. To this end, this study proposes a deep learning model, T-SeGAT, designed to improve the accuracy and generalization ability of molecular property and CPI prediction. T-SeGAT integrates the ESM-2 protein language model, ChemBERTa molecular language model, and a graph neural network based on graph attention network (GAT) and Set2Set, thereby enabling multi-level feature extraction and fusion from sequence to structure. Meanwhile, to handle the imbalance of experimental data, the model introduces weighted

① 基金项目: 广东省区域联合基金重点项目 (2022B1515120075)
收稿时间: 2025-09-02; 修改时间: 2025-09-22; 采用时间: 2025-10-20; csa 在线出版时间: 2026-01-19
CNKI 网络首发时间: 2026-01-20

random sampling, balanced/focal/adaptive loss functions, and a dynamic threshold search mechanism at the levels of data loading, loss calculation, and prediction decision-making. Furthermore, it combines an AUC difference-based overfitting suppression method, early stopping strategy, and learning rate scheduling to enhance training stability and generalization ability. Experiments are conducted on the BACE, P53, and hERG datasets for molecular property prediction, and on the Human and C. elegans datasets for CPI prediction, with stratified five-fold cross-validation adopted for performance evaluation. The results show that T-SeGAT consistently outperforms existing baseline models on all datasets. Among them, on the BACE and hERG datasets, the AUC and precision improved by 0.022, 0.010 and 0.004, 0.022 respectively compared with the second-best model, while on the Human dataset, precision increases by 0.013. In conclusion, T-SeGAT demonstrates clear advantages in precision, stability, and practicality, providing powerful support for molecular property and CPI prediction in drug discovery.

Key words: graph neural network (GNN); molecular property prediction; compound-protein interaction (CPI); imbalanced data learning; multi-head attention (MHA); deep learning

随着人口老龄化加剧与慢性病负担加重, 公众对优质医疗资源的需求日益迫切, 但新药研发周期长、成功率低的问题正严重影响着民生福祉[1]. 传统药物发现流程高度依赖实验筛选, 往往耗时十余年、耗资数十亿美元, 候选药物的整体失败率高达 96%[2]. AI 药物筛选通过高效分析海量数据、精准预测化合物活性及安全性, 显著缩短研发周期并降低成本, 突破传统临床试验的效率瓶颈. 可通过计算方法快速评估化合物库中小分子与靶标 (如酶或受体蛋白) 的结合亲和力, 有效减少了实验筛选的化合物数量, 将阳性率提高至 30%[3], 成为当今最具潜力的药物开发工具之一.

在药物虚拟筛选与发现中, 分子性质预测与化合物-蛋白质相互作用 (compound-protein interaction, CPI) 决定了小分子的 ADMET (吸收、分布、代谢、排泄、毒性) 特征, CPI 模型可帮助筛选出对靶标具有生物活性的苗头化合物 (hit compound), 并为多重药理学预测和药物重定位提供基础[4]. 提高这两项任务的计算预测性能, 不仅能显著加快先导化合物的筛选速度, 还可减少成本和失败风险.

早期方法主要依赖专家设计的分子描述符 (如拓扑指纹、结构片段) 和传统机器学习算法 (SVM[5]、随机森林[6]、贝叶斯优化[7]) 进行 QSAR[8]/QSPR 分析[9]. Mahé等人[10]提出的药效团核 (pharmacophore kernel) 就是该思路的典型代表. 然而, 这些方法过度依赖人工特征, 难以捕捉分子内部的复杂结构关系.

为了解决这一瓶颈, 图神经网络 (graph neural network, GNN) 被引入分子预测领域[11]. 分子本质上是一种图结构, 原子对应节点, 化学键对应边. Gilmer 等人[12]提出的神经消息传递网络 (message passing neural network, MPNN) 奠定了分子图学习的基石; 随后, Chen 等人[13]结合全局状态变量与 GNN, 实现了对分子与晶体属性的精准预测. Ma 等人[14]设计的多视角 GNN (MV-GNN) 通过自注意力和一致性损失进一步提升图表示力. 而 Jiang 等人[15]利用神经架构搜索 (NAS) 自动优化 MPNN 聚合方式, 在 QM9 与 MoleculeNet 数据集上均取得新高. 尽管如此, GNN 模型往往被视为“黑箱”, 可解释性不足. 为此, Wu 等人[16]提出了支持可解释化学设计的 SME 方法, 以增强 GNN 的透明度.

与此同时, 自然语言处理 (NLP) 技术也在分子性质预测中大放异彩. Bjerrum[17]利用 SMILES 枚举法结合 LSTM 网络进行预测. Jaeger 等人[18]提出的 Mol2Vec 模型, 则借鉴 Word2Vec 实现了分子嵌入学习. 最近, ChemBERTa[19]的预训练检查点 ChemBERTa-77M-MLM 在小分子回归任务上表现出更高的稳健性.

在 CPI 预测领域, 早期工作如 Wen 等人[20]基于深度置信网络 (DBN) 的方法提取原始特征以预测新型药物-靶标相互作用; DeepDTA[21]则用双卷积网络端到端学习 SMILES 与蛋白 FASTA 序列, 实现亲和力回归与分类; TransformerCPI[22]利用 Transformer 架构融合原子多特征与蛋白 Word2Vec 嵌入, 进一步提升了模型的可解释性和泛化能力. 另一类基于分子图的方法, 如 BindingSiteDTI[23]和 GraphDTA[24], 则通过 GCN 模型对 RdKit 转换而来的分子图进行节点信息聚合, 并结合蛋白序列或结构特征进行预测[25]. 尽管这些深度学

习模型展示了强大的预测能力[26], 它们在 3D 结构信息利用[27]、跨蛋白家族迁移[28]与可解释性[29]方面仍存在不足.

基于以上背景, 本研究提出了一个新的深度学习模型 T-SeGAT, 该模型将 ESM-2 蛋白语言模型、ChemBERTa 分子语言模型与 GAT+Set2Set 图神经网络相结合, 并以多层感知器为预测器, 对分子性质与 CPI 任务进行统一建模. 相较于现有方法, 该模型在端到端特征学习、多不平衡处理策略、过拟合自适应缓解以及可解释性增强等方面均做出了改进, 并通过在多个公开数据集 (包括 P53、hERG、BACE、Human、C. elegans) 上的系统评估, 验证了所提方法的优越性与泛化能力. 具体而言, T-SeGAT 的主要工作如下.

1) 将 ESM-2 蛋白语言模型、ChemBERTa 分子语言模型和 GAT+Set2Set 图神经网络融合集成, 打通序列和结构两类表征, 实现高效的特征提取与融合.

2) 在数据层、损失层和决策层分别引入加权随机采样、多种可切换的平衡/焦点/自适应损失函数以及动态阈值搜索, 全面提升不平衡样本的识别能力.

3) 设计了基于训练/验证 AUC 差距的自适应过拟合缓解机制, 配合早停与学习率调度, 有效提高模型的稳定性和泛化性能.

4) 通过在 P53、hERG、BACE、Human 和 C. elegans 这 5 个公开数据集上进行分层 5 折交叉验证、详尽的可视化分析与消融实验, 系统量化了各模块的独立贡献, 并验证了所提方法在多项指标上的显著优势.

1 方法

1.1 T-SeGAT 概述

如图 1 所示, T-SeGAT 主要由以下 4 个模块组成: 分子序列编码模块、分子图编码模块、蛋白质序列编码模块和预测器模块.

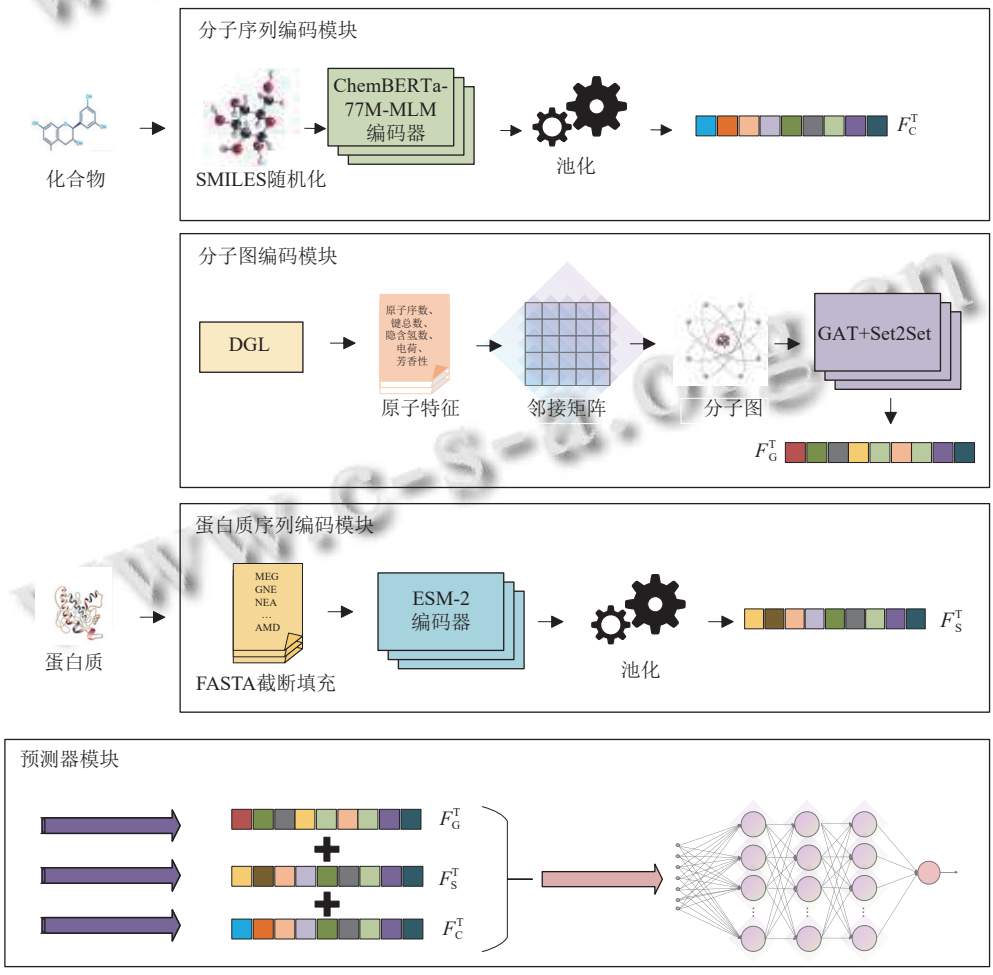

图 1 T-SeGAT 模型框架流程图

T-SeGAT 模型首先对数据进行预处理, 对每条分子 SMILES 序列多次随机化, 生成等价但表示不同的 SMILES 序列, 对蛋白质 FASTA 序列进行最大长度裁剪 (如 1 024 残基) 并在末端填充, 目的是提高模型的泛化能力和防止过拟合, 随后将处理后的数据输入各个模块进行训练.

分子序列编码模块提取化合物的原始特征, 然后输出分子语义向量 (即 F_C^T).

分子图编码模块对分子图特征进行提取, 输出图向量 (即 F_G^T).

蛋白质序列编码模块提取上下文嵌入向量, 最后输出池化特征作为蛋白质全局表示 (即 F_S^T).

预测器模块对上述得到的图级向量、蛋白质向量、分子向量进行拼接, 形成融合特征. 随后输入预测器中进行预测.

1.2 数据预处理

数据预处理是为了将化合物和蛋白质序列表示为能够被编码器接受和处理的格式.

化合物的 SMILES 和 FASTA 蛋白质序列是模型的基本输入格式, 它们是一种用字符串描述化合物和蛋白质分子结构的特殊序列, 包含了大量的化学信息和生物信息, 其表现形式如表 1 所示. 本文主要使用 Python 包 DGL-LifeSci[30]、ChemBERTa-77M-MLM[31]、RDKit[25]和 ESM-2[32]进行处理. 提取的基于分子的原始特征见表 2.

表 1 化合物蛋白质序列的表现形式

项目	表示
化合物	咖啡因
SMILES	CN1C=NC2=C1C(=O)N(C(=O)N2C)C
蛋白质	溶菌酶
FASTA序列	—

注: “—”表示不存在 FASTA 序列

表 2 分子的原始特征

原子特征	表示
原子序数	—
隐含氢原子数	0–3
芳香性	是否有芳香性
总键数	1–4
形式电荷	—

注: “—”表示特征表示比较复杂, 可忽略

1.3 序列编码模块

T-SeGAT 设计的序列编码模块由分子序列编码模块和蛋白质序列编码模块组成, 它们分别学习化合物和蛋白质的高级特征. 在对分子性质进行预测时, 分别使用图注意网络和分子序列编码器学习分子图结构和分子序列特征. 在预测化合物-蛋白质相互作用时, 通过图注意网络和蛋白质序列编码器学习分子图结构和蛋白质序列特征. 如图 1 所示, 分子序列编码模块和蛋白质序列编码模块分别输入化合物的 SMILES 序列和蛋白质 FASTA 序列, 随后学习序列的深度隐藏特征, 其维度在学习过程中确定. 下面将对这两个模块进行详细介绍.

1.3.1 分子序列编码模块

T-SeGAT 设计的分子序列编码模块用的是 ChemBERTa 编码器. ChemBERTa 是一种基于 BERT 架构的预训练语言模型, 专门设计用于处理化学文本和分子序列, 它通过自监督学习任务 (如掩码语言建模) 学习分子序列的语义表示. 在预测分子性质时, 通过 ChemBERTa 对随机化后的 SMILES 序列进行 BERT 风格的 Masked LM 嵌入, 池化后得到分子语义向量. ChemBERTa 编码过程包括 3 个操作: token 嵌入、多头自注意力层和输出池化.

(1) token 嵌入定义如下:

$$E = \left[e_{s_1}, e_{s_2}, \cdots, e_{s_{L_m}}\right] \tag{1}$$

其中, $S_m = (s_1, s_2, \cdots, s_{L_m})$: 长度为 L_m 的 SMILES token 序列, e_{s_i}: token s_i 的嵌入向量, $e_{s_i} \in R^{d_{\text{emb}}}$, d_{emb}: 嵌入维度.

(2) 多头自注意力层定义如下:

$$\begin{cases} Q = EW^Q, K = EW^K, V = EW^V \\ Attention(Q,K,V) = Softmax\left(\dfrac{QK^{\mathrm{T}}}{\sqrt{d_k}}\right)V \end{cases} \tag{2}$$

其中, $W^Q, W^K, W^V \in R^{d_{\text{hid}}}$: 线性变换权重, d_k: 每个 head 的键, *Softmax* 函数在序列长度 (token) 维度上对注意力得分进行归一化, 从而得到各位置的注意力权重分布 (和为 1).

(3) 输出池化定义如下:

[CLS]token 向量: 如果取第 1 个位置向量 $h_{\text{mol}} = h_i^{(L)} \in R^{d_{\text{hid}}}$, 则均值池化为:

$$h_{\text{mol}} = \frac{1}{L_m}\sum_{i=1}^{L_m} h_i^{(L)} \tag{3}$$

其中, $h_i^{(l)} \in R^{d_{\text{hid}}}$ 表示第 l 层第 i 个 token 的隐藏层向量, L 为 Transformer 层数, d_{hid} 为 Transformer 隐藏层维度.

1.3.2 蛋白质序列编码模块

蛋白质序列编码模块所用的是 ESM-2 编码器, ESM-2 是 Facebook AI Research 开发的一种蛋白质语

言模型, 基于自监督学习能够捕捉蛋白质序列的长期依赖关系和结构信息. 在预测化合物-蛋白质相互作用时, ESM-2 对输入的蛋白质 K-mer 序列提取上下文嵌入向量. 仅保留最后一层输出的平均池化特征作为蛋白质全局表示. 关键公式与 ChemBERTa 类似, 不同的是针对蛋白质序列进行了优化.

1.4 分子图编码模块

在提出的模型中引入图注意力网络 (GAT), 因为它是一种专门处理分子图的深度神经网络模型, 通过引入注意力机制来解决图结构数据中节点特征聚合的问题. 具体是通过注意力机制为不同邻居节点分配不同的贡献权重, 使关键邻居在信息聚合中占更大比重, 从而提升对节点间复杂关系的建模能力.

图注意力网络 (GAT) 输入 DGL 构造的带自环的图, 然后将分子结构视为无向图 $G=(V, E)$, 节点 V 为原子, 边 E 为化学键. 其中每个节点 i 有输入特征向量 $x_i \in R^{d_{\text{in}}}$ (如原子序数、度数、形式电荷、隐式氢数、芳香性等). 输出为每个节点的隐藏表示 $h_i^{\text{out}} \in R^{H\times d_{\text{hid}}}$ (多头拼接后的维度), 其中 H 代表多头数, d_{hid} 为每个注意力头的输出维度.

对于第 h 个注意力头, 节点 i 和其邻居 $j \in N(i)$ 的更新流程分为以下几步.

(1) 线性变换

将节点 i 和 j 的输入特征映射到隐藏空间:

$$z_i^{(h)} = W^{(h)^{\mathrm{T}}}, x_i = R^{d_{\text{hid}}};\ z_j^{(h)} = W^{(h)^{\mathrm{T}}}, x_j = R^{d_{\text{hid}}}$$

其中, $x_i \in R^{d_{\text{in}}}$ 为节点 i 原始特征, $W^{(h)} \in R^{d_{\text{in}}\times d_{\text{hid}}}$ 为第 h 个头的可学习投影矩阵, $z_j^{(h)} \in R^{d_{\text{hid}}}$ 为线性变换后的隐藏表示.

(2) 计算未归一化注意力系数

将 $z_i^{(h)}$ 和 $z_j^{(h)}$ 拼接后, 通过一个可学习向量 $a^{(h)}$ 计算注意力分数:

$$e_{ij}^{(h)} = LeakyReLU\left(a^{(h)^{\mathrm{T}}}\left[z_i^{(h)} \| z_j^{(h)}\right]\right) \tag{4}$$

其中, $z_i^{(h)}, z_j^{(h)} \in R^{d_{\text{hid}}}$, $\left[z_i^{(h)} \| z_j^{(h)}\right] \in R^{2d_{\text{hid}}}$ 表示向量拼接, $a^{(h)} \in R^{2d_{\text{hid}}}$ 为第 h 个头的注意力权重向量.

(3) 注意力权重归一化

对同一个中心节点 i, 将与各邻居 $j \in N(i)$ 的注意力分数做 *Softmax* 归一化:

$$\alpha_{ij}^{(h)} = \frac{\exp\left(e_{ij}^{(h)}\right)}{\sum_{k\in N(i)} \exp\left(e_{ik}^{(h)}\right)} \tag{5}$$

其中, $\alpha_{ij}^{(h)} \in [0,1]$ 为节点 i 对邻居 j 的注意力权重, 满足 $\sum_{k\in N(i)} \alpha_{ij}^{(h)} = 1$.

(4) 消息聚合与激活

对所有邻居 $j \in N(i)$ 按注意力加权求和, 并经过激活函数操作:

$$\tilde{h}_i^{(h)} = \sigma\left(\sum_{k\in N(i)} \alpha_{ij}^{(h)} z_j^{(h)}\right) \tag{6}$$

其中, $\tilde{h}_i^{(h)}$ 为第 h 头输出的节点 i 新表示, $\tilde{h}_i^{(h)} \in R^{d_{\text{hid}}}$, $\sigma(\cdot)$ 为激活函数, 常用 ELU 或 ReLU.

(5) 多头拼接

将 H 个头的输出在维度上拼接, 得到最终的节点表示:

$$h_i^{\text{out}} = \Big\|_{h=1}^{H} \tilde{h}_i^{(h)} \in R^{H\times d_{\text{hid}}} \tag{7}$$

其中, $h_i^{\text{out}} \in R^{H\times d_{\text{hid}}}$ 表示拼接后每个节点的最终隐藏表示.

在图注意力网络 (GAT) 计算每一层的注意力头之后, 还有一步操作 Set2Set 池化[33]. Set2Set 池化是一种基于 LSTM 的可学习读出 (readout) 方法, 适用于对无序节点集生成全局图表示. 其核心在于通过若干轮“LSTM 查询+注意力加权”迭代, 动态地从节点集中提取信息.

节点表示为 $\{h_i\}_{i=1}^{N}$, LSTM 状态 (q^t, c^t) 初始化为 0, 迭代轮数为 K. 其中, $h_i = h_i^{\text{out}} \in R^d$, $d = Hd_{\text{hid}}$, $q^t \in R^{d_{\text{L}}}$, $c^t \in R^{d_{\text{L}}}$, $t = 0, 1, \cdots, T_{d_{\text{L}}}$ 为 Set2Set 中 LSTM 的隐藏维度.

查询更新 (LSTM)、注意力得分与归一化权重、读出向量 (加权聚合)、输出拼接 (全局图表示) 的核心公式分别为:

$$(q^t, c^t) = LSTM(q^{t-1}, c^{t-1}) \tag{8}$$

$$e_i^t = h_i^{\mathrm{T}} q^t,\ a_i^t = \frac{\exp\left(e_i^t\right)}{\sum_{j=1}^{N} \exp\left(e_j^t\right)} \tag{9}$$

$$r^t = \sum_{i=1}^{N} a_i^t h_i \in R^d \tag{10}$$

$$h_{\text{graph}} = \left[r^1 \| r^2 \| \cdots \| r^K\right] \tag{11}$$

其中, q^t 为查询向量, c^t 为 LSTM 单元状态, $e_i^t \in (0,1)$, N 为节点总数, r^t 称为第 t 步的“读出 (readout)”或“聚合向量”, $h_{\text{graph}} \in R^{Kd}$.

1.5 预测器模块

预测器模块由多层感知器组成. 在预测时, 隐藏层数和每层的神经元数量由超参数传入. 该模块定义

如下:

$$h^{(\text{cat})} = \left[h^{(\text{mol}_{\text{smi}})} \middle\| h^{(\text{mol}_{\text{seq}})} \middle\| h^{(\text{prot})} \right] \tag{12}$$

$$u^{(l)} = W^{(l)} z^{(l-1)} + b^{(l)} \tag{13}$$

$$h^{(l)} = \varphi\left(u^{(l)}\right),\ h^{(l)} \in R^{d^{(l)}} \tag{14}$$

$$z^{(l)} = Dropout\left(h^{(l)}\right) \tag{15}$$

$$\sigma = W^{(\text{out})} z^{(L)} + b^{(\text{out})} \tag{16}$$

$$p_1 = \frac{\exp(\sigma_1)}{\exp(\sigma_0) + \exp(\sigma_1)},\ p_0 = 1 - p_1 \tag{17}$$

$$\hat{p} = \sigma(\sigma_1 - \sigma_0),\ \sigma(x) = \frac{1}{1 + \mathrm{e}^{-x}} \tag{18}$$

其中, $h^{(\text{cat})}$ 是特征拼接公式, $h^{(\text{cat})} \in D_{\text{cat}} = d_m + d_g + d_p$, $h^{(\text{mol}_{\text{smi}})}$、$h^{(\text{mol}_{\text{seq}})}$、$h^{(\text{prot})}$ 分别是分子序列编码、蛋白质编码和分子图编码, "||"表示拼接操作. $u^{(l)}$ 为第 l 层隐藏层的线性变换公式, $W^{(l)} \in R^{d^{(l)} \times d^{(l-1)}}$, $b^{(l)} \in R^{d^{(l)}}$, $h^{(l)}$ 为激活函数, 这里常用 $\varphi(x) = \max(0, x)$ 对向量逐元素作用. $z^{(l)}$ 为第 l 层的输出, 并输入第 l+1 层. 当完成 L 个隐藏层后, 得到 $z^{(L)} \in R^{d^{(L)}}$, 输出层将其映射为 C 维的 logits 向量, 其中 C 为类别数. 在本实验中令 C=2, 得到 $\sigma = [\sigma_0, \sigma_1]^{\mathrm{T}}$. p_1 和 $\hat{p}$ 为二分类 (C=2) 的两种预测输出概率. 且 $W^{(\text{out})} \in R^{C \times d^{(L)}}$, $b^{(\text{out})} \in R^{C}$, $\sigma \in R^{C}$, $\hat{p} \in (0, 1)$.

1.6 注意力权重分布分析

为增强 T-SeGAT 模型的可解释性, 我们在 P53 数据集上系统分析了多头注意力机制的功能. 具体而言, 从图神经网络 (GAT) 及分子/蛋白序列编码器 (ChemBERTa 与 ESM-2) 的注意力矩阵中提取各头的权重分布, 并计算多样性 (熵、稀疏度、多跳相关) 及子结构富集度 (归因相关、ICC 稳定性) 等指标, 其结果如表 3 所示. 结果表明, 不同注意力头在功能上呈现明显分化: Head 1/3/6 表现为低熵、低稀疏度的"局部聚焦型"头, 其注意力集中在芳香环与氮杂环等关键化学基序上, 并在富集检验中达到高度显著, 同时在归因一致性与跨折稳定性方面表现最佳. 相较之下, Head 2/5/7 的熵值较高, 注意力分布更分散, 对 3 跳以上拓扑路径具有更强敏感性, 显示其偏向捕获分子结构的长程依赖. 多头注意力机制在 T-SeGAT 中不仅提升了预测性能, 更通过"低熵局部基序头"与"高熵长程依赖头"的互补作用形成了稳定的表征分工. 该机制能够同时关注药物分子的关键功能团与整体拓扑结构, 在解释性与泛化性能之间实现平衡, 为后续药物-蛋白相互作用预测提供了坚实的理论支撑.

表 3 T-SeGAT 模型多头注意力功能综合分析表

Head	熵	稀疏度	多跳相关	归因相关	ICC 稳定性	芳香环	氮杂环	羰基	HBA	HBD
1	1.21±0.06	0.18±0.02	0.12	0.46	0.82	2.31, <1E–4	1.88, <1E–3	1.41, 0.012	1.33, 0.028	1.07, 0.41
3	1.25±0.05	0.20±0.03	0.09	0.41	0.79	2.12, <1E–4	1.79, <1E–3	1.55, 0.006	1.29, 0.035	1.05, 0.48
6	1.28±0.07	0.22±0.03	0.11	0.39	0.77	1.95, <5E–4	1.52, 0.009	1.38, 0.018	1.21, 0.07	0.98, 0.71
2	1.63±0.08	0.37±0.04	0.28	0.18	0.61	0.96, 0.63	1.03, 0.58	1.07, 0.41	1.44, 0.011	1.36, 0.019
5	1.58±0.09	0.35±0.05	0.26	0.16	0.59	—	—	—	—	—
7	1.67±0.10	0.39±0.05	0.3	0.15	0.57	—	—	—	—	—

1.7 模型超参数

训练模型时, 超参数的设置对模型优化有很大影响. 众所周知, 超参数的调优过程是深度学习模型训练中一个非常重要且耗时的过程. 我们综合考虑数据集规模、模型复杂度和 GPU 显存与单次训练时间预算等因素, 经过大量实验并且使用 5 折交叉验证最终确定了超参数的最优范围, 如表 4 所示.

1.8 交叉验证

在进行分子性质和化合物-蛋白质相互作用 (CPI) 预测时, 数据集通常很小, 而且有的数据集正负样本极不平衡 (例如 P53), 这会极大地影响预测的准确率. 为了克服这一弊端, 采用加权采样使少数类样本被抽到的概率高于多数类, 从而平衡各类别在梯度更新中的影响力, 并且使用 5 折交叉验证评估模型的稳定性与泛化能力, 在数据划分上既保证训练、验证分布一致, 又充分利用数据做多次训练和验证.

表 4 超参数范围

超参数	范围
GAT_IN_DIM	5
GAT_HIDDEN_DIM	[128, 256]
GAT_OUT_DIM	[128, 256]
GAT_LAYERS	[2, 4]
GAT_HEADS	[4, 8]
MLP_DIMS	[128, 64, 32]
DROPOUT	[0.2, 0.3]
BATCH_SIZE	[32, 64]
LEARNING_RATE	[1E–4, 3E–5]
WEIGHT_DECAY	[1E–5, 5E–4]
OPTIMIZER	[Adam, AdamW, RMSpro]

1.9 评估指标

本文用到的评价指标为准确率 (*Accuracy*)、精确

率 (*Precision*)、召回率 (*Recall*)、接受者操作特征曲线下面积 (receiver operating characteristic curve, ROC) 曲线、Matthews 相关系数 (Matthews correlation coefficient, *MCC*) 以及精确率-召回率 (precision-recall, PR) 曲线. 下面将详细描述如何计算各项指标.

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \tag{19}$$

$$Precision = \frac{TP}{TP+FP} \tag{20}$$

$$Recall = \frac{TP}{TP+FN} \tag{21}$$

$$TPR = \frac{TP}{TP+FN} \tag{22}$$

$$FPR = \frac{FP}{FP+TN} \tag{23}$$

$$MCC = \frac{TP\times TN - FP\times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \tag{24}$$

其中, *TP* 是真实为正、预测也为正的样本数, *TN* 是真实为负、预测也为负的样本数, *FP* 和 *FN* 分别为真实为负、预测为正的样本数以及真实为正、预测为负的样本数. ROC 曲线是以假阳率 *FPR* 为 *x* 轴、真阳率 *TPR* 为 *y* 轴构成的曲线, AUC 就是 ROC 曲线下的面积值, 越接近 1 表示分类性能越好.

一般而言, *Accuracy* 是最直观的衡量指标, 被广泛用于各种任务中; 对于二元分类问题, 尤其是常见的非平衡数据集, *Precision* 和 *Recall* 重点关注正样本, 因此具有很好的评估效果. ROC 曲线和 PR 曲线计算得到的 AUC 和 AUPR (area under precision-recall curve) 也是两个非常重要的综合性衡量指标. Matthews 相关系数 (*MCC*) 综合考虑 *TP*、*TN*、*FP*、*FN* 这 4 项指标, 对不平衡数据具有更稳定的评价性能.

2 实验与结果分析

2.1 所用数据集

为了全面评估本文方法, 我们使用了多个数据集, 包括 BACE、P53、hERG、Human、C. elegans. 表 5 详细列出了每个数据集的详细信息. 其中 BACE、P53 和 hERG 这 3 个数据集用于分子性质预测任务, Human 和 C. elegans 用于化合物-蛋白质相互作用预测任务.

在分子性质预测任务中, 使用了以下 3 个数据集.

BACE 数据集[34]来自 NCBI PubChem BioAssay (Assay AID 411279), 包含约 1 500 个化合物的 β-分泌酶 1 (BACE-1)[35]抑制活性数据, 用于评估模型对 BACE-1 抑制效果的预测性能.

P53 数据集[36]来自 IARC 和 NCI 的 TP53 突变数据库, 包含约 2.5–3.4 万条在人类肿瘤或细胞系中观察到的 P53 突变, 广泛用于蛋白质功能改变相关研究.

hERG 数据集[37]整合了电生理实验结果 (PubChem Assay AID 588834、ChEMBL、Zenodo), 共有几千至几万条化合物的 IC_{50} 值和阻滞标签, 是药物安全性 (心脏毒性) 预测的重要标准数据集.

在 CPI 预测任务中, 进一步采用了两个基准数据集: Human[26]与 C. elegans[38]数据集为公开的 CPI 基准数据集, 其负样本依据 Liu 等人[38]提出的高可信筛选框架构建, 并保持正负样本比例为 1:1. 数据采集自 DrugBank 与 Matador 等资源, 后被 Tsubaki 等人[26]复用于多项深度学习研究中, 已成为跨物种 CPI 建模的重要对照集.

Human 数据集包含人类基因组相关数据库, 涵盖人类基因信息, 包括基因序列、功能、表达等多方面数据, 是生物信息学研究的重要资源.

C. elegans 数据集包含秀丽隐杆线虫的基因组、转录组、蛋白质组等数据, 为研究线虫生物学及基因功能提供重要支持.

表 5 数据集详细信息

数据集	靶点	化合物	阳性	阴性
BACE	—	1 507	691	822
P53	—	7 460	529	6 968
hERG	—	9 804	4 668	5 136
Human	2 001	1 767	3 364	3 364
C. elegans	1 876	2 111	3 893	3 893

2.2 基线方法

为了评估本模型的性能, T-SeGAT 将分别与多种传统的机器学习方法以及多个目前已公开发表的、基于深度学习的前沿工作进行比较. 方法分别为 K 近邻 (KNN)[39]、随机森林 (RF)[6]、梯度提升[40]、GCN[41]、MPNN[12]、MG-S[42]、CPI-GNN[26]、Transformer-CPI[22]、DeepCPI[43]、GraphCPI[44].

2.3 实验结果

2.3.1 分子性质预测结果评估

表 6 展示了 T-SeGAT 和基准方法在 BACE 数据集上的结果. 可以看出 T-SeGAT 在 BACE 数据集上的 AUC、*Precision*、*Recall*、*MCC* 指标分别为 0.887、0.789、0.834、0.664. 在 6 个基准模型中, T-SeGAT 在

AUC 指标上取得了最好结果, 梯度提升在 *Precision*、*Recall*、*MCC* 这 3 个指标上取得了最好成绩. 即本文方法在 *Precision*、*Recall* 和 *MCC* 这 3 项指标上整体优于除梯度提升之外的基准模型, 整体性能也优于基准性能平均水平, 这说明本模型和最新提出的前沿研究相比具有相当的特征学习能力.

表 6 BACE 数据集结果

方法	AUC	*Precision*	*Recall*	*MCC*
KNN	0.815±0.017	0.748±0.030	0.784±0.024	0.563±0.048
RF	0.787±0.043	0.746±0.029	0.784±0.011	0.589±0.084
梯度提升	0.881±0.026	0.830±0.044	0.859±0.034	0.706±0.054
GCN	0.865±0.012	0.777±0.030	0.810±0.045	0.617±0.125
MPNN	0.793±0.031	0.720±0.015	0.760±0.030	0.620±0.068
MG-S	0.865±0.023	0.779±0.042	0.882±0.026	0.621±0.057
本方法	0.887±0.006	0.789±0.052	0.834±0.036	0.644±0.034

表 7 给出了 T-SeGAT 与各基准方法在 P53 数据集上的对比结果. T-SeGAT 在 *Precision* 指标上取得最优表现, 相较次优方法提升约 1 个百分点. 在 AUC、*Recall* 与 *MCC* 这 3 项指标上, T-SeGAT 与次优方法差异较小、表现相当; 同时, 在其余评价指标上整体优于其他基准模型. 上述结果表明, 本文方法在 P53 数据集上具有良好的综合性能, 与现有先进方法相比具备较强竞争力, 这说明本方法在 P53 数据集上同其他先进方法相比也极具竞争力.

表 7 P53 数据集结果

方法	AUC	*Precision*	*Recall*	*MCC*
KNN	0.953±0.006	0.837±0.014	0.990±0.005	0.815±0.027
RF	0.789±0.012	0.690±0.010	0.760±0.027	0677±0.028
梯度提升	0.950±0.010	0.863±0.010	0.953±0.012	0.813±0.025
GCN	0.955±0.004	0.891±0.010	0.919±0.005	0.780±0.105
MPNN	0.941±0.007	0.870±0.011	0.941±0.011	0.826±0.090
MG-S	0.989±0.003	0.943±0.007	0.993±0.002	0.925±0.044
本方法	0.988±0.007	0.950±0.015	0.976±0.017	0.921±0.030

表 8 展示了 T-SeGAT 和基准方法在 hERG 数据集上的结果. T-SeGAT 在 AUC 和 *Precision* 上均取得了最优结果, 在 *Recall*、*MCC* 指标上也优于基线模型平均水平. 总体来说, T-SeGAT 在 hERG 数据集上取得了最高性能.

表 8 hERG 数据集结果

方法	AUC	*Precision*	*Recall*	*MCC*
KNN	0.808±0.016	0.735±0.014	0.734±0.008	0.466±0.024
RF	0.707±0.012	0.621±0.007	0.710±0.020	0.429±0.022
梯度提升	0.793±0.014	0.716±0.008	0.728±0.012	0.434±0.015
GCN	0.843±0.019	0.742±0.019	0.744±0.014	0.585±0.045
MPNN	0.840±0.023	0.770±0.016	0.771±0.019	0.578±0.060
MG-S	0.845±0.006	0.768±0.007	0.755±0.007	0.547±0.034
本方法	0.849±0.008	0.768±0.029	0.743±0.057	0.540±0.025

2.3.2 化合物-蛋白质相互作用预测结果评估

表 9 展示了 T-SeGAT 和基准方法在 Human 数据集上的结果. 结果显示, 同所有基线方法相比, T-SeGAT 在 AUC、*Precision*、*MCC* 均是最高水准, 其中, *Precision* 与次优模型差异为 0.013, 虽然 AUC 和 *MCC* 与次优模型基本持平, 但是从总体性能来说, T-SeGAT 在 Human 数据集上取得的性能均优于基线模型.

表 9 Human 数据集结果

方法	AUC	*Precision*	*Recall*	*MCC*
KNN	0.927±0.009	0.887±0.012	0.876±0.011	0.758±0.024
RF	0.836±0.019	0.723±0.019	0.800±0.009	0.495±0.046
梯度提升	0.936±0.010	0.880±0.014	0.882±0.018	0.761±0.019
GCN	0.968±0.005	0.914±0.011	0.905±0.007	0.798±0.045
MPNN	0.950±0.012	0.883±0.011	0.879±0.016	0.807±0.037
MG-S	0.977±0.003	0.933±0.005	0.932±0.015	0.866±0..038
CPI-GNN	0.970	0.918	0.923	0.825
TransformerCPI	0.973±0.002	0.916±0.006	0.925±0.006	0.869
DeepCPI	0.955	—	—	—
GraphCPI	0.946	—	—	—
本方法	0.977±0.002	0.946±0.006	0.917±0.001	0.866±0.009

表 10 展示了 T-SeGAT 和基准方法在 C. elegans 数据集上的结果. 可以看出 T-SeGAT 在 C. elegans 数据集上的 AUC、*Precision*、*Recall*、*MCC* 指标分别为 0.982、0.960、0.932、0.890. T-SeGAT 并非最佳, 可能是因为 C. elegans 数据集数据量较小, 而且少数类标签更稀疏, 使加权采样和加权损失难以完全补偿, 影响了整体性能. 即便如此, T-SeGAT 在 C. elegans 数据集上的性能仍与次优模型基本持平且高于基准模型的平均水平.

表 10 C. elegans 数据集结果

方法	AUC	*Precision*	*Recall*	*MCC*
KNN	0.953±0.008	0.911±0.010	0.911±0.021	0.829±0.018
RF	0.888±0.012	0.863±0.019	0.700±0.025	0.628±0.025
梯度提升	0.967±0.010	0.941±0004	0.919±0.013	0.862±0.010
GCN	0.985±0.003	0.955±0.008	0.946±0.004	0.881±0.033
MPNN	0.976±0.007	0.929±0.010	0.930±0.015	0.909±0.044
MG-S	0.989±0.002	0.961±0.004	0.963±0.003	0.918±0.027
CPI-GNN	0.978	0.938	0.929	0.816
TransformerCPI	0.988±0.002	0.952±0.006	0.953±0.005	0.906
DeepCPI	0.943	—	—	—
本方法	0.982±0.002	0.960±0.013	0.932±0.019	0.890±0.014

综上所述, T-SeGAT 在分子性质预测方面展现了最高性能, 并且可以持续提供好的结果. 在 CPI 预测任务中, T-SeGAT 在 Human 数据集上取得了最佳性能, 与先进模型相比, 尽管没有在全部指标上取得最好成绩, 综合表现也在全部基准模型中排在前列, 证明在

CPI 预测中也具有很强的竞争性. 从侧面也反映出 T-SeGAT 具有较强的泛化性和鲁棒性. 在整体表现上, 基于深度学习的方法要好于基于机器学习的方法, 这表明深度学习相较于传统机器学习方法更适合解决分子性质预测和 CPI 预测问题.

2.4 消融实验

本节在 P53 数据集和 Human 数据集上进行消融实验, 探究 T-SeGAT 模型中不同模块对模型整体性能的影响. 由此, 本节设置了 4 个 T-SeGAT 的消融模型: 1) T-SeGAT-A 模型, 不使用 ESM-2 预训练嵌入, 蛋白序列用随机噪声代替, 用来评估蛋白编码作用; 2) T-SeGAT-B 模型, 取消使用 ChemBERTa, 分子序列直接用全 0 向量或随机噪声代替, 考察序列编码贡献; 3) T-SeGAT-C 模型, 用简单的全局求和池化代替 GAT+Set2Set, 考察图注意力与 Set2Set 的增益; 4) T-SeGAT-D 模型, 将原 MLP 替换为单层线性, 去掉隐藏层与非线性激活, 仅做一次线性映射. 每次实验只应用一项修改, 其余模块和超参数保持配置不变.

图 2 是 T-SeGAT 模型在 P53 数据集与 Human 数据集上的消融实验结果. (a)、(b) 为 4 项评价指标 AUC、*Precision*、*Recall* 与 *MCC* 的柱状对比, 用于展示不同消融设置的指标绝对值差异. (c)、(d) 为与 (a)、(b) 相同数据的折线趋势图, 用于突出各指标随消融设置变化的整体趋势. 纵轴为指标值, 取值范围为 0–1.

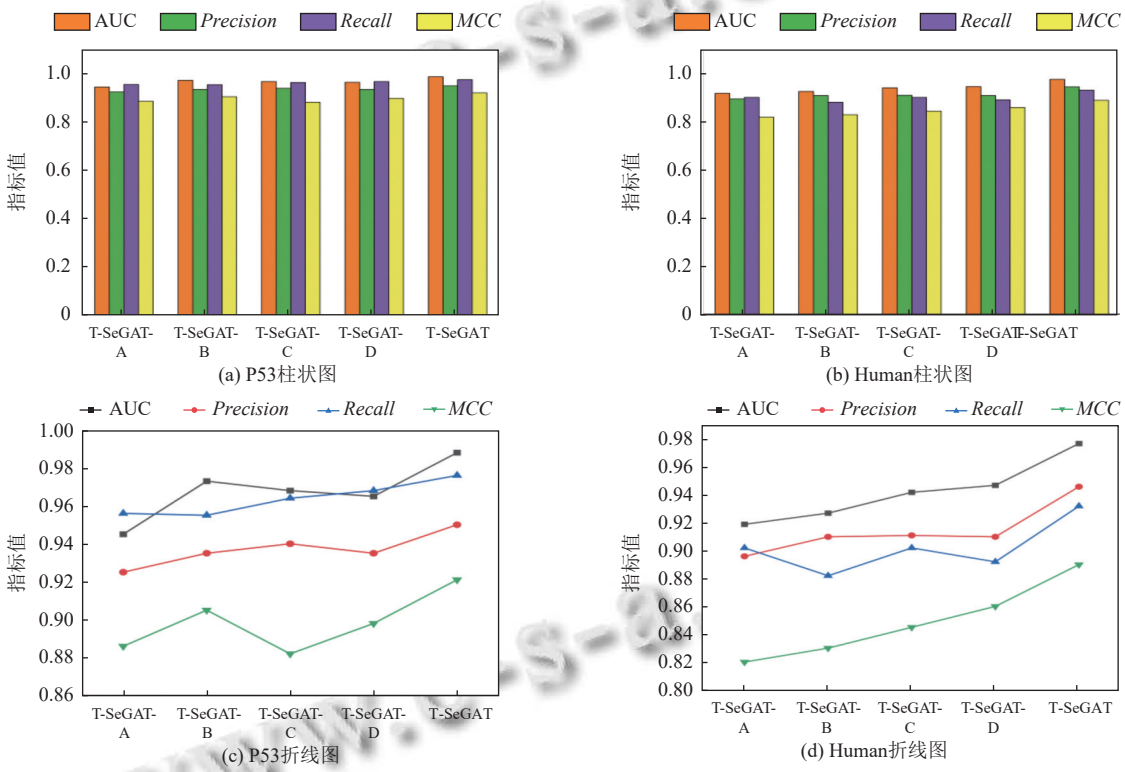

图 2 T-SeGAT 模型在 P53 数据集和 Human 数据集上的消融实验结果

从图 2 中可以看出, 在 P53 数据集和 Human 数据集上, 4 个消融模型的表现具有大致相同的趋势. 取消了 ESM-2 模块的 T-SeGAT-A 模型在 4 个消融实验模型中的性能表现最差, 表明 ESM-2 模块对 T-SeGAT 模型的贡献最大; T-SeGAT-B 模型对比 T-SeGAT 模型也有不小的性能损失, 尤其是 *Recall* 指标, 说明拥有捕捉复杂分子模式的能力同样重要; T-SeGAT-C 模型在 4 个指标上也有不同程度的下降, 表明加入合适的池化模块对于提高模型的性能也很重要; T-SeGAT-D 模型对比 T-SeGAT 模型同样有不同程度的下降, 这表明 MLP 的非线性转换对捕捉化合物相互作用的复杂模式占有重要地位. 综上所述, 每个模块对于 T-SeGAT 模型都有举足轻重的地位并且发挥着不可或缺的作用.

2.5 时间、空间复杂度、训练效率

在时间复杂度、空间复杂度、训练效率上选择与次优模型 MG-S 进行对比, 为了公平起见, 这两个方法我们均在 Linux 环境中运行实验, 利用一张型号为 NVIDIA RTX 3090 的 GPU 来加速模型训练过程. 结果

如表 11 所示.

结果显示, T-SeGAT 在时间复杂度、空间复杂度与训练效率方面均显著优于对照模型 MG-S. 首先, 从总训练时间来看, 本方法在所有任务上均大幅缩短: 在 BACE 和 P53 数据集上加速比分别高达 15.8 倍和 7.94 倍, 而在 Human、C. elegans 与 hERG 数据集上也保持在 1.5–3 倍之间, 体现出跨数据规模的稳定优势. 其次, 在空间复杂度方面, 本方法的显存占用更低, 5 个数据集的峰值范围为 0.9–2.3 GB, 始终低于 MG-S 的 1.4–3.1 GB, 同时参数量也更小, 表明模型结构更加轻量. 再次, 在训练效率维度, 本方法的单 epoch 平均耗时普遍更短, 且所需收敛轮次显著减少, 这使得在整体训练时间上优势更加明显. 综上所述, T-SeGAT 在不同数据集和任务规模下均展现出“更快收敛、更低显存、更短训练时间”的综合优势, 实现了性能与资源消耗之间的高效平衡, 体现出较强的实用性与可扩展性.

表 11 时间、空间复杂度、训练效率结果展示

数据集	实验用时 (h)		加速比 (MG-S/本方法)	峰值显存 (GB)		平均耗时 (s/epoch)		收敛轮次	
	本方法	MG-S		本方法	MG-S	本方法	MG-S	本方法	MG-S
Human	6.7	10.2	1.52×	2.1	2.8	28.4	31.5	42	65
C. elegans	9.1	26.7	2.93×	2.3	3.1	31.0	39.8	53	88
BACE	0.33	5.2	15.8×	0.9	1.4	12.5	21.0	10	42
hERG	1.73	4.67	2.70×	1.4	2.0	25.8	30.5	24	55
P53	1.7	13.5	7.94×	1.3	2.2	27.0	36.5	22	74

3 结论

本文提出了一种面向不平衡数据的分子性质与化合物-蛋白质相互作用预测模型 T-SeGAT. 该模型融合 ESM-2 蛋白语言模型、ChemBERTa 分子语言模型与 GAT+Set2Set 图神经网络, 实现了序列与结构特征的联合建模. 在应对数据不平衡问题时, 模型引入加权采样、多种损失函数以及动态阈值策略, 有效提升了少数类样本的识别能力; 同时, 结合基于 AUC 差异的过拟合抑制机制、早停与学习率调度, 显著增强了训练的稳健性与泛化性.

在 BACE、P53、hERG、Human 和 C. elegans 这 5 个数据集上的实验结果表明, T-SeGAT 在 AUC、*Precision*、*Recall* 和 *MCC* 等指标上均优于主流基线模型, 展现出更高的准确性和稳定性. 消融实验进一步验证了各组成模块的独立贡献, 表明序列建模、图结构建模与注意力机制在整体性能提升中均发挥了关键作用. 同时, T-SeGAT 在时间复杂度、空间复杂度与训练效率方面也表现出明显优势, 实现了预测性能与资源消耗之间的良好平衡.

综上所述, T-SeGAT 不仅在分子性质预测与 CPI 预测任务中取得了优异的实验结果, 也为复杂生物数据的多模态深度学习提供了可行的研究思路与有效的实践框架.

参考文献

1 药智网数据团队. 重磅! 中国化学新药「研发成功率」公布. 新华社客户端. https://news.yaozh.com/archive/45244.html. (2025-04-02).

2 Gorgulla C, Boeszoermenyi A, Wang ZF, *et al*. An open-source drug discovery platform enables ultra-large virtual screens. Nature, 2020, 580(7805): 663–668. [doi: 10.1038/s41586-020-2117-z]

3 Du BX, Qin Y, Jiang YF, *et al*. Compound-protein interaction prediction by deep learning: Databases, descriptors and models. Drug Discovery Today, 2022, 27(5): 1350–1366. [doi: 10.1016/j.drudis.2022.02.023]

4 Keserű GM, Makara GM. Hit discovery and hit-to-lead approaches. Drug Discovery Today, 2006, 11(15-16): 741–748. [doi: 10.1016/j.drudis.2006.06.016]

5 Suthaharan S. Machine learning models and algorithms for big data classification: Thinking with examples for effective learning. Boston: Springer, 2016. 207–235.

6 Biau G, Scornet E. A random forest guided tour. Test, 2016, 25(2): 197–227. [doi: 10.1007/s11749-016-0481-7]

7 Frazier PI. A tutorial on Bayesian optimization. arXiv: 1807.02811, 2018.

8 Muratov EN, Bajorath J, Sheridan RP, *et al*. QSAR without borders. Chemical Society Reviews, 2020, 49(11): 3525–3564. [doi: 10.1039/D0CS00098A]

9 Karelson M, Lobanov VS, Katritzky AR. Quantum-chemical descriptors in QSAR/QSPR studies. Chemical Reviews, 1996, 96(3): 1027–1044. [doi: 10.1021/cr950202r]

10 Mahé P, Ralaivola L, Stoven V, *et al*. The pharmacophore kernel for virtual screening with support vector machines. Journal of Chemical Information and Modeling, 2006, 46(5): 2003–2014. [doi: 10.1021/ci060138m]

11 Wieder O, Kohlbacher S, Kuenemann M, *et al*. A compact

review of molecular property prediction with graph neural networks. Drug Discovery Today: Technologies, 2020, 37: 1–12. [doi: 10.1016/j.ddtec.2020.11.009]

12 Gilmer J, Schoenholz SS, Riley PF, *et al*. Neural message passing for quantum chemistry. Proceedings of the 2017 International Conference on Machine Learning. PMLR, 2017. 1263–1272.

13 Chen C, Ye WK, Zuo YX, *et al*. Graph networks as a universal machine learning framework for molecules and crystals. Chemistry of Materials, 2019, 31(9): 3564–3572. [doi: 10.1021/acs.chemmater.9b01294]

14 Ma HH, Bian YT, Rong Y, *et al*. Multi-view graph neural networks for molecular property prediction. arXiv:2005.13607, 2020.

15 Jiang S, Balaprakash P. Graph neural network architecture search for molecular property prediction. Proceedings of the 2020 IEEE International Conference on Big Data (Big Data). Atlanta: IEEE, 2020. 1346–1353.

16 Wu ZX, Wang JK, Du HY, *et al*. Chemistry-intuitive explanation of graph neural networks for molecular property prediction with substructure masking. Nature Communications, 2023, 14(1): 2585. [doi: 10.1038/s41467-023-38192-3]

17 Bjerrum EJ. SMILES enumeration as data augmentation for neural network modeling of molecules. arXiv:1703.07076, 2017.

18 Jaeger S, Fulle S, Turk S. Mol2vec: Unsupervised machine learning approach with chemical intuition. Journal of Chemical Information and Modeling, 2018, 58(1): 27–35. [doi: 10.1021/acs.jcim.7b00616]

19 Chithrananda S, Grand G, Ramsundar B. ChemBERTa: Large-scale self-supervised pretraining for molecular property prediction. arXiv:2010.09885, 2020.

20 Wen M, Zhang ZM, Niu SY, *et al*. Deep-learning-based drug-target interaction prediction. Journal of Proteome Research, 2017, 16(4): 1401–1409. [doi: 10.1021/acs.jproteome.6b00618]

21 Öztürk H, Özgür A, Ozkirimli E. DeepDTA: Deep drug-target binding affinity prediction. Bioinformatics, 2018, 34(17): i821–i829. [doi: 10.1093/bioinformatics/bty593]

22 Chen LF, Tan XQ, Wang DY, *et al*. TransformerCPI: Improving compound-protein interaction prediction by sequence-based deep learning with self-attention mechanism and label reversal experiments. Bioinformatics, 2020, 36(16): 4406–4414. [doi: 10.1093/bioinformatics/btaa524]

23 Pan F, Yin C, Liu SQ, *et al*. BindingSiteDTI: Differential-scale binding site modelling for drug-target interaction prediction. Bioinformatics, 2024, 40(5): btae308. [doi: 10.1093/bioinformatics/btae308]

24 Nguyen T, Le H, Quinn TP, *et al*. GraphDTA: Predicting drug-target binding affinity with graph neural networks. Bioinformatics, 2021, 37(8): 1140–1147. [doi: 10.1093/bioinformatics/btaa921]

25 Bento AP, Hersey A, Félix E, *et al*. An open source chemical structure curation pipeline using RDKit. Journal of Cheminformatics, 2020, 12(1): 51. [doi: 10.1186/s13321-020-00456-1]

26 Tsubaki M, Tomii K, Sese J. Compound-protein interaction prediction with end-to-end learning of neural networks for graphs and sequences. Bioinformatics, 2019, 35(2): 309–318. [doi: 10.1093/bioinformatics/bty535]

27 Li SY, Wan FP, Shu HT, *et al*. MONN: A multi-objective neural network for predicting compound-protein interactions and affinities. Cell Systems, 2020, 10(4): 308–322. e11.

28 Li M, Lu ZL, Wu YF, *et al*. BACPI: A bi-directional attention neural network for compound-protein interaction and binding affinity prediction. Bioinformatics, 2022, 38(7): 1995–2002. [doi: 10.1093/bioinformatics/btac035]

29 Zhao MH, Yuan M, Yang YN, *et al*. CPGL: Prediction of compound-protein interaction by integrating graph attention network with long short-term memory neural network. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2023, 20(3): 1935–1942. [doi: 10.1109/TCBB.2022.3225296]

30 Li MF, Zhou JJ, Hu JJ, *et al*. DGL-LifeSci: An open-source toolkit for deep learning on graphs in life science. ACS Omega, 2021, 6(41): 27233–27238. [doi: 10.1021/acsomega.1c04017]

31 bioRxiv. Gromova AA, Maida AS. ChemBERTaDDI: Transformer driven molecular structures and clinical data for predicting drug-drug interactions. https://www.biorxiv.org/content/biorxiv/early/2025/06/11/2025.01.22.634309.full.pdf. (2025-06-11).

32 Cordoves-Delgado G, García-Jacas CR. Predicting antimicrobial peptides using ESMFold-predicted structures and ESM-2-based amino acid features with graph deep learning. Journal of Chemical Information and Modeling, 2024, 64(10): 4310–4321. [doi: 10.1021/acs.jcim.3c02061]

33 Vinyals O, Bengio S, Kudlur M. Order matters: Sequence to sequence for sets. Proceedings of the 4th International Conference on Learning Representations. San Juan: OpenReview.net, 2015.

34 Wang YL, Bolton E, Dracheva S, *et al*. An overview of the PubChem BioAssay resource. Nucleic Acids Research, 2010, 38(S1): D255–D266.

35 Vassar R. BACE 1: The β-secretase enzyme in Alzheimer's disease. Journal of Molecular Neuroscience, 2004, 23(1-2): 105–113. [doi: 10.1385/JMN:23:1-2:105]

36 娄文加, 陈青, 刘立, 等. miR-34 家族——肿瘤抑制蛋白 p53 高度相关的 microRNA. 遗传, 2010, 32(5): 423–430.

37 Kalyaanamoorthy S, Barakat KH. Development of safe drugs: The hERG challenge. Medicinal Research Reviews, 2018, 38(2): 525–555. [doi: 10.1002/med.21445]

38 Liu H, Sun J, Guan J, *et al*. Improving compound-protein interaction prediction by building up highly credible negative samples. Bioinformatics, 2015, 31(12): i221–i229. [doi: 10.1093/bioinformatics/btv256]

39 Guo GD, Wang H, Bell D, *et al*. KNN model-based approach in classification. Proceedings of the 2003 OTM Confederated International Conferences on the Move to Meaningful Internet Systems: CoopIS, DOA, and ODBASE. Catania: Springer, 2003. 986–996.

40 Friedman JH. Greedy function approximation: A gradient boosting machine. Annals of Satistics, 2001, 29(5): 1189–1232. [doi: 10.1214/aos/1013203451]

41 Kipf TN. Semi-supervised classification with graph convolutional networks. arXiv:1609.02907, 2016.

42 Ma J, Zhang RS, Li TF, *et al*. A deep learning method for predicting molecular properties and compound-protein interactions. Journal of Molecular Graphics and Modelling, 2022, 117: 108283. [doi: 10.1016/j.jmgm.2022.108283]

43 Wan FP, Zhu Y, Hu HL, *et al*. DeepCPI: A deep learning-based framework for large-scale in silico drug screening. Genomics, Proteomics & Bioinformatics, 2019, 17(5): 478–495.

44 Quan Z, Guo Y, Lin X, *et al*. GraphCPI: Graph neural representation learning for compound-protein interaction. Proceedings of the 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). San Diego: IEEE, 2019. [doi: 10.1109/BIBM47256.2019.8983267]

(校对责编: 张重毅)