PIBB 生物化学与生物物理进展
Progress in Biochemistry and Biophysics
2025,52(3):783~794
www.pibb.ac.cn

基于Bert+GCN多模态数据融合的药物分子属性预测*

闫效莺[1)**] 靳艳春[1)] 冯月华[1)] 张绍武[2)**]
([1)] 西安石油大学计算机学院，西安 710065；[2)] 西北工业大学自动化学院，信息融合教育部重点实验室，西安 710072）

摘要 **目的** 药物研发成本高、周期长且成功率低。准确预测分子属性对有效筛选药物候选物、优化分子结构具有重要意义。基于特征工程的传统分子属性预测方法需研究人员具备深厚的学科背景和广泛的专业知识。随着人工智能技术的不断成熟，涌现出大量优于传统特征工程方法的分子属性预测算法。然而这些算法模型仍然存在标记数据稀缺、泛化性能差等问题。鉴于此，本文提出一种基于Bert+GCN的多模态数据融合的分子属性预测算法（命名为BGMF），旨在整合药物分子的多模态数据，并充分利用大量无标记药物分子训练模型学习药物分子的有用信息。**方法** 本文提出了BGMF算法，该算法根据药物SMILES表达式分别提取了原子序列、分子指纹序列和分子图数据，采用预训练模型Bert和图卷积神经网络GCN结合的方式进行特征学习，在挖掘药物分子中“单词”全局特征的同时，融合了分子图的局部拓扑特征，从而更充分利用分子全局-局部上下文语义关系，之后，通过对原子序列和分子指纹序列的双解码器设计加强分子特征表达。**结果** 5个数据集共43个分子属性预测任务上，BGMF方法的AUC值均优于现有其他方法。此外，本文还构建独立测试数据集验证了模型具有良好的泛化性能。对生成的分子指纹表征（molecular fingerprint representation）进行t-SNE可视化分析，证明了BGMF模型可成功捕获不同分子指纹的内在结构与特征。**结论** 通过图卷积神经网络与Bert模型相结合，BGMF将分子图数据整合到分子指纹恢复和掩蔽原子恢复的任务中，可以有效地捕捉分子指纹的内在结构和特征，进而高效预测药物分子属性。

关键词 Bert预训练，注意力机制，分子指纹，分子属性预测，图卷积神经网络
中图分类号 TP391 **DOI**：10.16476/j.pibb.2024.0299 **CSTR**：32369.14.pibb.20240299

利用各种实验方法对分子的属性进行预测，是新药发现中的一个重要环节。准确可靠地预测分子属性，包括物理化学、生物活性以及吸收、分配、代谢、排泄和毒性等，进而寻找具有理想属性的药物是药学领域的一个长期目标[1]。然而如何对药物分子进行有效表征是分子属性预测的关键步骤[2]。传统的分子属性预测方法通常是基于专家手动设计的描述符或分子指纹，如扩展连接指纹（extended connectivity fingerprints，ECFP）等，作为药物分子特征输入到机器学习模型中，如支持向量机、随机森林、逻辑回归等[3]，最后输出分子属性预测结果。然而分子描述符设计繁琐，其准确性很大程度上依赖于专家的生物信息背景[4-5]。近年来，能够从原始数据中自动学习并提取高阶特征的深度学习（deep learning，DL）方法因其强大的自学习能力成为各领域的热点研究对象，并在自然语言处理[6-7]、计算机视觉[8-9]等领域取得突破性进展。

简化分子线性输入规范（simplified molecular input line entry system，SMILES）序列和分子图是药物分子的两种常见表征方式。其中SMILES序列采用一串ASCII字符描述药物分子的组成和化学结构。作为一种文本，一系列文本处理算法如卷积神经网络（convolutional neural network，CNN）和循环神经网络（recurrent neural network，RNN），长

* 国家自然科学基金（62173271），陕西省自然科学基础研究计划（2023-JC-YB-591），西安石油大学研究生创新与实践能力培养计划（YCS23213171）和西安石油大学研究生精品案例库建设（2024-X-YAL-003）资助项目。
** 通讯联系人。
闫效莺 Tel：029-81469729，E-mail：xiaoying_yan@126.com
张绍武 Tel：029-88431308，E-mail：zhangsw@nwpu.edu.cn
收稿日期：2024-07-06，接受日期：2024-12-02

短期记忆（long short-term memory，LSTM）网络等基于序列的模型可用于药物表征学习[10-13]。然而，SMILES序列无法直接编码重要的拓扑信息。分子图将药物表示为以原子为节点，键为边的图。许多最近的工作将图神经网络（graph neural network，GNN）用于药物分子的特征学习[14-20]，如Abbassi等[14]将SMILES序列转换为分子图，采用消息传递神经网络（message passing neural networks，MPNN）进行特征提取，最后使用MLP构建分类器预测分子化合物的性质。Xiong等[15]通过引入原子水平和分子水平的注意力机制，提出一种基于注意力机制的GNN模型用于分子特征学习。Han等[16]采用分层级的GNN模型和基于Transformer局部增强模型，对Atom级和motif级图数据分别学习拓扑表征，并通过集成基于Atom和motif的化学结构信息来增强motif特征的表征能力，最后通过多层感知机（multilayer perceptron，MLP）模块进行分子属性预测。此外，Liu等[17]结合化合物中的大量文本知识，提出一种联合化合物化学结构和文本描述的自监督学习策略用于分子属性预测。由于分子的三维（3D）几何空间表征包含丰富的空间信息，对理解分子性质至关重要，基于此，近年来，Son等[18]融合化合物的1D序列特征、2D图特征和3D空间特征到一个共享子空间，得到一个融合的特征表征。模型首先将3类特征分别进行特征嵌入映射，得到序列网络、拓扑网络和立体网络特征；之后采用融合转换模块，将3类特征和3类特征的两两组合特征分别进行基于Transformer模块的处理；最后融合交叉熵损失和组合损失进一步优化分子特征表征。Ma等[19]采用4种基本的深度学习框架，从分子指纹、2D分子图、3D分子图和分子图像特征中分别提取不同类型的特征，最后采用融合模块整合4个特征向量，经过预测层进行分子属性预测。

整合多种分子特征数据，可以提高分子特征表征的准确性，然而，标记数据稀缺是DL模型在化合物分子属性预测中面临的一个巨大难题，由此催生了大量自监督学习模型。自监督学习方法有效地利用丰富的未标记的分子数据，自动构建有意义的特征表征，已成为一个强大的战略。在这类方法中，基于预训练框架和对比学习框架的特征表征学习是最流行的技术[21-26]。其中，基于预训练框架的分子特征表征学习方法，通过DL模型在大规模的分子数据集上训练，学习到的特征表征具有通用性和迁移性，能够广泛应用于各类下游任务，可提高模型的泛化性和效率。基于对比学习框架的分子特征表征学习方法，则是需要训练一个模型，以最大限度地提高来自相同分子但表征不同的分子数据对之间的相似性，从而增强对有区别的分子模式的识别，使得分子表征具有鲁棒性。如，Zhang等[21]提出一种基于预训练的多表征融合网络用于分子属性预测（命名为PremuNet）。PremuNet模型包含PremuNet-L和PremuNet-H两个分支，其中PremuNet-L利用分子指纹，SMILES Transformer和GNN提取低维分子特征，PremuNet-H则是通过融合2D图形信息和3D几何信息提取分子的高维特征，设置了3个预训练任务，以探索分子的高质量特征表征。此外，广泛应用于自然语言处理领域的双向编码器表征法（Bert）模型[22-23]，使用大规模无标签语料来预训练模型，预训练完成后保存模型参数，并加入一个输出层，在不同的下游任务上进行微调，可以实现对各种数据规模小且有标签下游任务的预测。如果将一个药物分子的分子指纹或原子序列类比成一个自然语言语句，那么药物分子的子结构或原子则相当于是语句中的一个单词。基于此，Wang等[24]对药物SMILES中提取的原子序列进行掩蔽操作，并以Transformer模型为骨干网络进行掩蔽恢复任务的预训练，之后通过微调完成了分子属性预测任务。由于SMILES序列表征的原子字典词汇量较小，故恢复掩蔽原子作为预训练任务，无法学习信息丰富的分子表征，而且恢复掩蔽原子任务只关注到分子序列的上下文信息，基本没有考虑分子图的子结构特征。Li等[25]提出的Mol-BERT模型中，通过RDKit提取SMILES序列的摩根指纹，获取原子级和子结构级特征，对指纹进行掩蔽，并应用于下游分子属性预测任务。该方法引入相对较大的分子指纹词汇表，增加了模型恢复掩蔽的难度，使得模型能够更精细地描述分子的结构和特性。最近，Wang等[26]基于对比学习框架，对药物分子图分别进行原子级别，键级别和子图级别的图增强操作，通过最大化来自同一个分子的多个增强图，并最小化不同分子图之间的一致性，对药物分子进行特征学习，最后在不同属性预测任务中进行微调。考虑化合物分子中原子数量相对小，而且原子集合中元素数量极其不平衡，Xia等[27]引入了一种变体的VQ-VAE方法[28]对同一种原子根据周边环境编码为不同的ID，在扩大原子词典的同时，消除了不同原子之间的数量差异，采用对比

学习通过图神经网络进行分子特征表征学习。

不同类型的分子表征方法可以从不同角度刻画化合物，它们包含了化合物的各种不同信息。尽管从分子的指纹序列、原子序列和分子图表征来学习分子特征表征已经取得了很好的结果，然而当前的研究并不能很好地从多方面整合利用药物分子这些的原始特征。基于此，本文提出一种基于Bert+图卷积神经网络（graph convolutional networks，GCN）的多模态多任务数据融合策略预测分子属性，设计了双损失函数通过Bert模型同时对来自SMILES的原子序列和分子指纹序列进行掩蔽恢复的特征学习，设计了基于GCN的分子图表征学习，将Bert的全局特征提取能力和GCN的局部特征提取能力结合，增强Bert中的注意力机制，提高药物分子表征的质量。

1　材料与方法

1.1　数据来源

预训练阶段的数据集，来自于ChEMBL[29]和ZINC[30]数据库，其中，ChEMBL[29]是一个“化学基因组学”数据库，汇集了百万数量级的化学、生物活性和基因组数据，ZINC是一个免费的商业化合物数据库，包含了2.3亿化合物的详细数据信息，为虚拟筛选提供便利。本文从ChEMBL数据库中随机选择1 000万种化合物，从ZINC数据库随机选择200万种化合物，将获取的1 200万种无标签化合物用于预训练模型。在下游任务微调阶段，为验证模型性能，我们在MoleculeNet[3]的5个分类数据集上进行了分子属性预测实验，包括43个特定任务，囊括了各种常见的分子特性以及关键药物动力学方法（absorption，distribution，metabolism，excretion，and toxicity，ADMET）端点，数据详情见表1。

由于分子的SMILES序列长度从几个到100多个不等，为保证模型训练一致性，通常需要将一个批次中不同长度的SMILES序列补齐到相同长度。然而，过多填充字符可能会对模型的性能产生负面影响。因此，本文采用了一种优化策略。具体地，设输入N个化合物，根据分子SMILES序列长度的不同范围（如0~5、5~10、10~15等）将化合物分成M组，组内打乱化合物的顺序。由此确保了每个批次中的化合物在原子数量上基本一致，同时保证了一个化合物序列最多只需填充5个字符。这不仅提高了数据处理的效率，而且显著减少了填充字符对模型性能的不良影响，从而提升了模型的稳定性和预测准确性。

Table 1　The downstream task dataset used in BGMF

Dataset	Number of molecules	Task type	Number of tasks
BBBP	2 053	Binary	1
SIDER	1 427	Multilabel	27
HIV	41 127	Binary	1
Tox21	8 014	Multilabel	12
ClinTox	1 491	Multilabel	2

1.2　预训练阶段

Bert是由Google在2018年提出的预训练语言模型，它是基于Transformer架构的深度双向编码器[22]。与传统的词向量模型不同，Bert模型利用大规模文本数据进行预训练，其输入是由单词组成的句子序列，通过学习语言的深层表征，使模型能够更好地理解语义和上下文关系，之后，在微调部分对预训练学习到的语言表示进行实际应用。本文将药物分子分别表征为由原子组成的原子序列和由分子指纹组成的指纹序列。基于此，本文提出基于Bert+GCN的多模态数据融合策略预测分子属性的算法模型，其模型框架如图1所示。

1.2.1　分子语句生成

分子指纹语句生成。给定化合物的SMILES序列，使用摩根（Morgan）算法可以得到该化合物的一组分子子结构，即分子指纹（fingerprint）表征。每个子结构用一个固定长度的二进制向量表征。本文通过RDKit工具中的Morgan算法相关函数获取化合物中以每个原子为中心的半径为0及半径为1的子结构和在指纹词典中的对应位置索引，并将其按照SMILES序列中原子顺序和半径大小顺序，构建分子指纹语句。如化合物氟胞嘧啶的SMILES序列（NC1=NC(=O)NC=C1F）中包含9个原子，其对应的分子指纹序列为$(A_1^0, A_1^1, \cdots, A_9^0, A_9^1)$，其中$(A_1^0 - A_9^0)$分别对应序列中原子半径为0的子结构，$(A_1^1 - A_9^1)$分别对应原子半径为1的子结构。由Morgan算法获取到的半径为0及半径为1的子结构对应的分子指纹词典包含13 321个分子指纹。但是，在实际编码中除了具体指纹外，还有4种常见的特殊标记符，［CLS］、［PAD］、［UNK］和［MASK］，其中［CLS］用于表示分子的超节点（通常在指纹语句的第一个单词前加入），［PAD］用于表示填充补齐，［UNK］用

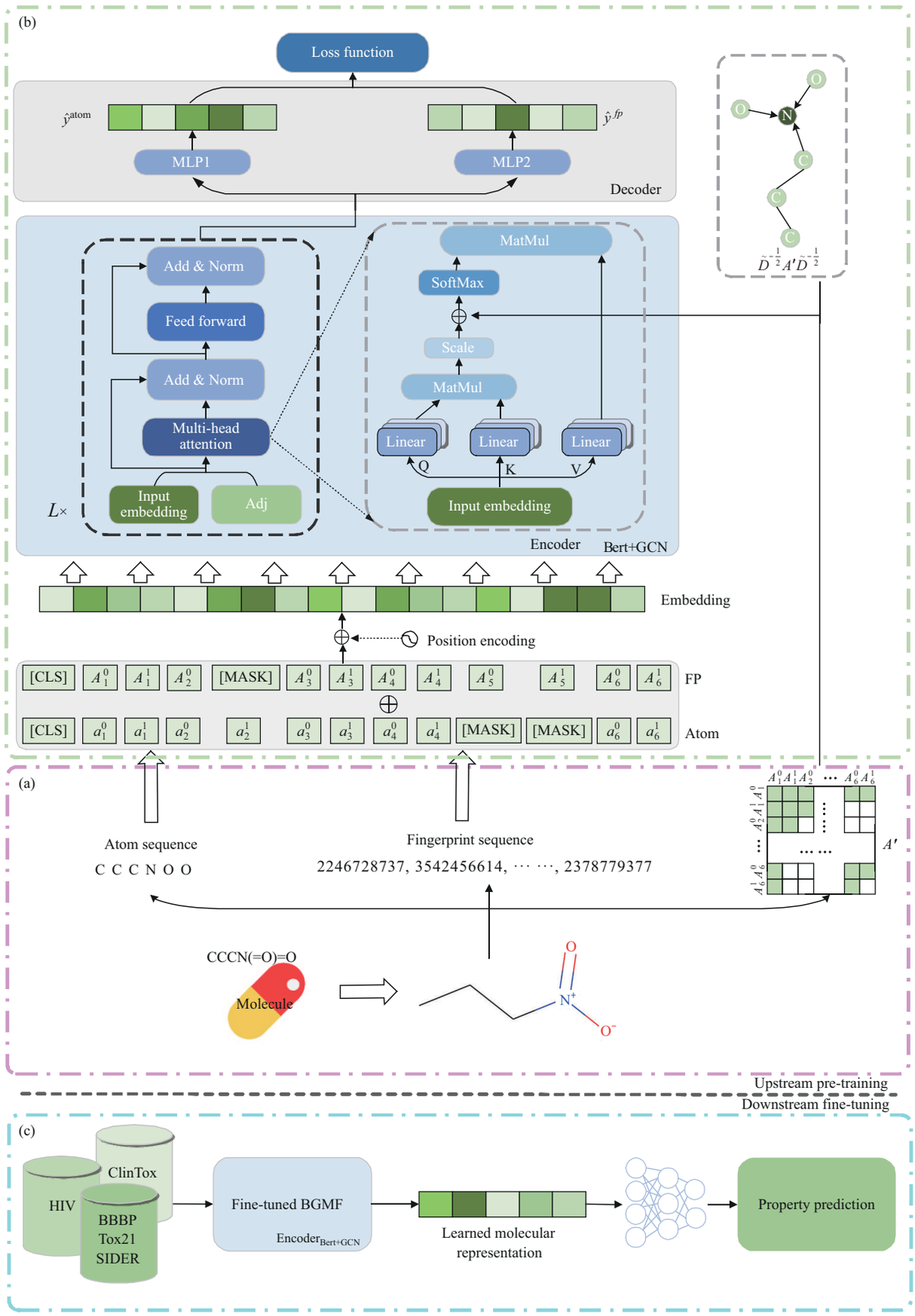


Fig. 1　The Overall framework of BGMF

(a) Molecular feature extraction; (b) Bert-GCN based pre-training; (c) fine-tuning.

于表示指纹词典中不存在的原子，［MASK］表示掩盖预测的词。最终的分子指纹词典一共包含13 321个子结构和4个特殊标记符。考虑在数据处理中一个批次的SMILES序列长度不一，需要补齐到相同长度，设一个批次中最长序列的药物化合物有n个原子，则该批次中原子数量不达n个的药物，需要使用［PAD］填充到n个，每个原子对应两个分子指纹（半径为0的子结构和半径为1的子结构），故该批次中药物d_i的分子指纹语句表示为$X_i^{fp}=\{e_{i,1}^{fp},\cdots,e_{i,l}^{fp},\cdots,e_{i,2n+1}^{fp}\}$，其中第一个为特殊标记符［CLS］，$e_{i,l}^{fp}$的初始编码为13 325维的独热编码（one-hot encoding）。

原子序列语句生成。给定SMILES序列，还可以用原子序列语句表征该化合物。分子指纹语句和原子序列语句将同时作为Bert模型的输入，因此，分子指纹语句和原子序列语句的“单词”数量应保持一致性。而在分子指纹语句中，对每个原子采用了半径为0和半径为1的两个子结构指纹，为了语句长度对齐，在原子序列语句中，对同一个原子a_1，设置了a_1^0和a_1^1两个“单词”。在随后掩码策略中，对应同一原子的这两个单词将同时进行掩蔽操作，以避免信息冗余。如，对具有9个原子的药物SMILES序列，可将其原子序列语句表示为$(a_1^0,a_1^1,\cdots,a_9^0,a_9^1)$，语句中$a_1^0$~$a_9^0$的值分别对应原子在原子词典中的位置索引，需要注意的是a_i^0和a_i^1对应同一个原子，故a_i^0和a_i^1对应相同的位置索引id。本文只考虑常用原子，原子词典共包含25个原子，与分子指纹语句类似，在原子序列编码中，除了原子本身，还包含了［CLS］、［PAD］、［UNK］和［MASK］4类常见的特殊标记符，故原子词典共包含29个单词。考虑一个批次中药物化合物SMILES序列长度的补齐，药物d_i的原子序列语句表示为$X_i^{\text{atom}}=\{e_{i,1}^{\text{atom}},\cdots,e_{i,l}^{\text{atom}},\cdots,e_{i,2n+1}^{\text{atom}}\}$，其中$n$为一个批次中最长序列药物的原子数目，$e_{i,l}^{\text{atom}}$的初始编码为29维的独热编码。

对药物的分子指纹编码和原子编码分别进行线性变换，转换得到相同维度的特征，公式如下：

$$h_{i,l}^{fp}=ReLU(W_{fp}^{(0)}e_{i,l}^{fp}) \tag{1}$$

$$h_{i,l}^{\text{atom}}=ReLU(W_{\text{atom}}^{(0)}e_{i,l}^{\text{atom}}) \tag{2}$$

这里的$ReLU$为激活函数，$e_{i,l}^{fp}$和$e_{i,l}^{\text{atom}}$分别对应化合物d_i编码中的第l个指纹和第l个原子的初始特征，其维度分别为13 325维和29维，$W_{fp}^{(0)}$和$W_{\text{atom}}^{(0)}$为可学习的权重参数，$h_{i,l}^{fp}$和$h_{i,l}^{\text{atom}}$为映射后的指纹特征和原子特征，其维数为H。

掩码策略。本文采用Bert的掩码语言模型任务作为预训练模型。具体来讲，从分子指纹序列随机抽取序列总长度的15%，对抽取到的分子指纹在80%的情况下用掩码标识符替换，以10%的概率随机从分子指纹词库中抽词替换，剩余10%的分子指纹不做处理；同样地，原子序列掩码也遵循上述方法。对于一些只有几个原子或分子指纹的语句，我们确保至少有一个单词将被选择用于掩蔽。

位置信息生成。分子指纹语句和原子序列语句中的“单词”都按照SMILES标准化后的原子顺序排序，因此化合物分子指纹语句中每一个“单词”的位置均具有独特含义。由于Transformer模型本身不具有处理序列中单词位置信息的能力，因此需要位置嵌入来提供这一信息。本文使用正余弦函数来编码分子指纹/原子序列语句中“单词”的位置，其公式定义为：$PE_{(pos,2i)}=\sin(pos/10\,000^{\frac{2i}{H}})$，$PE_{(pos,2i+1)}=\cos(pos/10\,000^{\frac{2i}{H}})$，其中$pos$表示分子指纹/原子序列语句中“单词”的位置，$H$为化合物的分子指纹/原子序列编码的特征维度，$i$对应特征中的第$i$个维度，最终的位置嵌入表征矩阵$X^{pos}\in R^{(2n+1)\times H}$。化合物$d_i$的位置嵌入表征为$X_i^{pos}=\{e_{i,1}^{pos},\cdots,e_{i,l}^{pos},\cdots,e_{i,2n+1}^{pos}\}$。

将分子指纹嵌入表征、原子序列嵌入表征和位置嵌入表征融合后作为化合物的分子语句特征。其融合方式为：$X_i=Tanh(W_1\cdot X_i^{fp}+W_2\cdot X_i^{\text{atom}}+W_3\cdot X_i^{pos})$，其中$W_1$，$W_2$，$W_3$为可学习的训练参数矩阵，$Tanh$为非线性激活函数。此时，输入模型的第$i$个化合物的特征表示为$X_i^{(0)}=[e_{i,1},\cdots,e_{i,l},\cdots,e_{i,2n+1}]$。

1.2.2　基于Bert+GCN的分子表征学习

Bert实质上是由多个Transformer编码器层堆叠而成的。每个编码器层都包含自注意力机制和前馈神经网络，允许模型捕捉输入序列中的复杂依赖关系。自注意力机制是Bert编码层的主要模块。具体地：

$$\text{Attention}(Q,K,V)=\text{soft max}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{3}$$

这里，Q，K和V是对输入特征矩阵分别进行线性转换得到，$Q=XW^Q$，$K=XW^K$，$V=XW^V$，W^Q，W^K和$W^V\in R^{H\times H}$是可学习的权重参数矩阵，d_k是Q和K的维度，即H，T表征矩阵的转置。

我们采用多头注意力机制，这种机制将多个缩放点积注意力层堆叠在一起，并允许它们并行运

行，能够组合来自不同子空间的信息，增强其鲁棒性并捕捉不同的模式[31]，具体来说，假设多头注意力机制拥有h个并行运行的注意力层，计算过程如下：

$$\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V) \quad (4)$$

$$\text{MultiHead}(X) = \text{Concat}(\text{head}_1, \cdots, \text{head}_h)W^o \quad (5)$$

这里的Q_i，K_i和$V_i \in R^{H \times \frac{H}{h}}$分别对应第$i$个注意力头的$Q$，$K$和$V$矩阵，$W^o \in R^{H \times H}$是可学习的参数矩阵。

GCN通过同时捕获图结构拓扑特征和节点特征进行特征表征与学习[32]。设图$G=(V, E)$，$V=\{v_1, \cdots, v_i, \cdots, v_n\}$是$n$个节点的集合，$E$是节点之间边的集合，其对应的邻接矩阵记作$A \in \{0, 1\}^{n \times n}$，节点特征矩阵记为$X \in R^{n \times p}$，GCN模型定义为$H^{l+1}=f(\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}H^lW^l)$，$H^l$和$W^l$分别为$l$层的节点表征和映射权重矩阵，$\tilde{D}_{ii}=\sum_j \tilde{A}_{ij}$为度矩阵，$\tilde{A}=A+I_n$，其中$I_n$为对角线为1的单位矩阵，初始节点表征矩阵$H^0$为$X$。

Bert模型中的自注意力机制可以提取到药物分子中单词（对应原子）的全局特征，得到语境化的词向量，而GCN旨在提取药物分子图中原子节点的局部特征信息。前者有助于理解分子在整体环境下的行为，而后者则揭示了分子内部的精细结构和相互作用。融合这两种特征信息，使得模型更好地理解分子的复杂性和多样性。第l层的融合过程为：

$$\text{fused}(X^l) = \text{MultiHead}(X^l) + h(X^l) \quad (6)$$

更具体地，描述为：

$$\text{fused}(X^l) = \text{softmax}(\frac{Q_iK_i^T}{\sqrt{d_{ik}}} + \tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}})V \quad (7)$$

这里的$Q_i = X^lW_i^Q$，$K_i = X^lW_i^K$，$V_i = X^lW_i^V$，d_{ik}是第i头中Q_i和K_i的维度，即$\frac{H}{h}$。

需要注意的是，由化合物分子的SMILES序列（假设由n个原子组成）得到的分子图，其对应的邻接矩阵维度为$n \times n$，分别对应$(a_1 - a_n)$原子之间的连接关系。为方便与Bert模型中化合物分子中单词的全局特征进行融合，需要将邻接矩阵A扩展为$(2n+1) \times (2n+1)$维度的邻接矩阵A'，其对应的原子信息为$(CLS, a_1^0, a_1^1, \cdots a_n^0, a_n^1)$，具体方法是：将邻接矩阵$A$中对应原子$a_i$的连接关系，复制到$A'$中对应的$a_i^0$和$a_i^1$所在的行和列中，超节点$CLS$与其他所有原子节点都有连接关系，即其所在行和列值均为1。

编码模块由多个具有相同结构的编码层堆叠而成，将上一层的输出加上残差作为下一层的输入。具体表示为：

$$\tilde{X}^l = \text{LayerNorm}(X^l + \text{fused}(X^l)) \quad (8)$$

经过L层编码后得到分子语句的编码表征记为Z。

1.2.3 解码器与损失计算

由于药物分子表征学习中使用到了原子序列和分子指纹序列两类特征，本文在解码器模块中设计了双解码器，将上述得到的分子语句编码表征Z分别解码得到初始的原子序列和分子指纹序列特征。考虑在基于Bert模型的非对称编码器-解码器设计中，恢复掩码仅在轻量级解码器中处理，而较浅的解码器不仅不会影响算法的整体性能，而且可以显著减少模型训练的计算量和内存消耗[33]，故将Bert编码器得到的药物分子语句的特征表征Z，分别经过一个多层感知机模块，映射为原子序列和分子指纹词典大小的特征，并对其进行归一化处理，得到原子和分子指纹的预测得分，见公式(9)和(10)。并采用多分类交叉熵损失计算目标值和预测值之间的损失，见公式(11)~(14)。

$$\hat{y}_i^{\text{atom}} = \text{MLP}_1(Z) \quad (9)$$

$$\hat{y}_i^{fp} = \text{MLP}_2(Z) \quad (10)$$

$$\text{Loss} = -\frac{1}{N_1}\sum_{i}^{N_1}\sum_{c=1}^{C_1} y_{ic}\log\hat{y}_{ic} \quad (11)$$

$$\text{Loss}_{\text{atom}} = \frac{1}{N}\sum_{i=1}^{N}\text{Loss}(\hat{y}_i^{\text{atom}}, y_i) \quad (12)$$

$$\text{Loss}_{fp} = \frac{1}{N}\sum_{i=1}^{N}\text{Loss}(\hat{y}_i^{fp}, y_i) \quad (13)$$

$$\text{Loss}_{\text{pretrain}} = (\text{Loss}_{fp} + \text{Loss}_{\text{atom}})/2 \quad (14)$$

其中，N_1表示药物分子指纹语句/原子序列语句中“单词”的数量，C_1表示分子指纹字典/原子序列字典的大小，N表示预训练集中样本的数量，y_{ic}是该分子指纹语句/原子序列语句中“单词”的真实标签，$\hat{y}_{ic}$是语句中第i个“单词”属于第c类的预测概率。

1.3 下游任务微调阶段

与预训练部分不同，下游任务的微调阶段不再掩蔽原子序列和分子指纹序列。我们在基于Bert+GCN特征提取器的基础上，添加了池化层、随机初始化的MLP和交叉熵损失完成特定的分子属性预测任务。其中，在池化层中本文实现了平均池化（MP）和最大池化（GP），并与序列输入中超级节点［CLS］的特征进行融合，最终的药物分子特征

记为D，公式如下：

$$D = Concat(GP(Z), MP(Z), Z_{[CLS]}) \quad (15)$$

$$y_i = ReLU(W_3(W_2(W_1(D)))) \quad (16)$$

$$Loss_{\text{fine-tune}} = -\frac{1}{M}\sum_{i=1}^{M}\sum_{c=1}^{C_2}[y_{ic}\cdot\log(\hat{y}_{ic}) + (1-y_{ic})\cdot \log(1-\hat{y}_{ic})] \quad (17)$$

这里的$Concat(\cdot)$表示特征串联操作，$ReLU$为激活函数，W_1，W_2和W_3为线性映射中可学习的参数矩阵，M是下游微调任务中药物样本数量，C_2表示下游微调数据集的分子属性标签数量，y_{ic}是药物样本i的第c类分子属性标签，$\hat{y}_{ic}$是药物样本i属于第c类分子属性的预测概率。

2　实验结果与分析

2.1　实验设置

使用PyCharm集成开发环境，Pytorch 1.7.1框架。在预训练阶段，将预训练数据集按9∶1比例随机分割为训练集和验证集。在下游任务微调阶段，即分子属性预测数据集上，使用scaffold分割法，按8∶1∶1比例将数据分为训练集、验证集和测试集。对测试集中每个样本分别进行预测，并将预测结果与实际标签进行对比。使用ROC曲线下面积AUC表征模型预测性能。AUC值越大，表示算法性能越好。通过寻找损失函数的最小值，对模型参数进行网格搜索确定最优参数。最后设定的预训练阶段参数为：学习率=1e−5，batch size=16，dropout rate=0.1，编码器层数为L=6，注意力头数h=6，特征维度H=300，使用Adam优化器运行5个epoch。微调阶段采用Adam优化器，经10次50个epoch的微调，并报告10次运行的平均测试结果，以获得测试集上ROC-AUC的平均值，并使用该值评价模型的预测性能。对不同的微调数据集，最优超参数取值不同，其中学习率为｛3e−6，3e−4｝，批大小为｛8，16，32｝，dropout［0.0~0.5］，其余超参数与预训练阶段一致。

2.2　与其他方法比较

为了全面评估算法的性能，我们将BGMF模型的预测结果与当前8种先进的基线方法：MoleculeSTM[17]、FTMMR[18]、SMILES-BERT[24]、Mol-BERT[25]、MolCLR[26]、Mole-BER[27]、MPNN[34]和FP2VEC[35]进行对比（图2）。BGMF在5个数据集中的4个展现了最优性能。在BBBP数据集上，BGMF模型与其他基线方法相比，AUC值提高了0.81%~23.03%，在Tox21数据集上，提高了1.63%~15.31%；在SIDER数据集上，提高了1.99%~13.34%；HIV数据集上，AUC提高了2.31%~6.26%。但是，在ClinTox数据集上BGMF方法的ROC-AUC值略低于Mol-BERT[25]、MPNN[34]、FTMMR[18]和MoleculeSTM[17]。为分析原因，我们对模型所使用的五个数据集进行分析（文档S1第**1**节，图S1），发现5个数据集中，BBBP和HIV是2个二分类数据集，各自对应一个任务，Sider、Tox21和ClinTox三个数据集对应有多个任务。Sider和Tox21中各个任务的正样本数量基本符合正态分布，而ClinTox中的2个任务（临床药物是否具有毒性信息和是否通过FDA批准）的正样本数量占比分别为7.56%和93.72%。因而推测，ClinTox数据集上BGMF方法的ROC-AUC值略低于MoleculeSTM等4种方法的原因可能与训练数据集的任务样本不均衡分布有关。鉴于MoleculeSTM[17]方法在ClinTox数据集上的ROC-AUC值最高，在ClinTox数据集上通过计算AUPR值比较BGMF和MoleculeSTM[17]两种方法的性能（文档S1第**2**节，表S1），从表S1可以看出，虽然BGMF的ROC-AUC值低于MoleculeSTM[17]方法，但AUPR值却高于MoleculeSTM[17]。

值得注意的是，与同样采用Bert模型的基线方法（SMILES-BERT和Mol-BERT）相比，BGMF模型在4个数据集上均取得了最佳性能。这一结果不仅体现了我们对Bert模型改进策略的有效性，也进一步证明了我们的模型在分子表征学习方面的卓越能力。

为了验证模型泛化性能，在模型训练的5个数据集之外，从MoleculeNet提供的分子属性数据集Bace中，选取与BBBP训练数据集相似性极低的300个化合物作为独立测试样本，测试模型的泛化性能，具体实施步骤见文档S1第**3**节。表S2的结果显示BGMF模型的泛化性能仍拥有最好的表现，具有良好推广性能。

2.3　消融实验

为了深入剖析BGMF模型中原子序列、分子指纹序列、基于GCN的分子图模块以及有无预训练模块对模型性能的影响，在消融实验中我们设计了4个BGMF的变体模型：第一个是将BGMF模型中的原子序列删除掉，记为BGMF（w/o Atom）；第二个是将模型中的分子指纹序列删除掉，记为

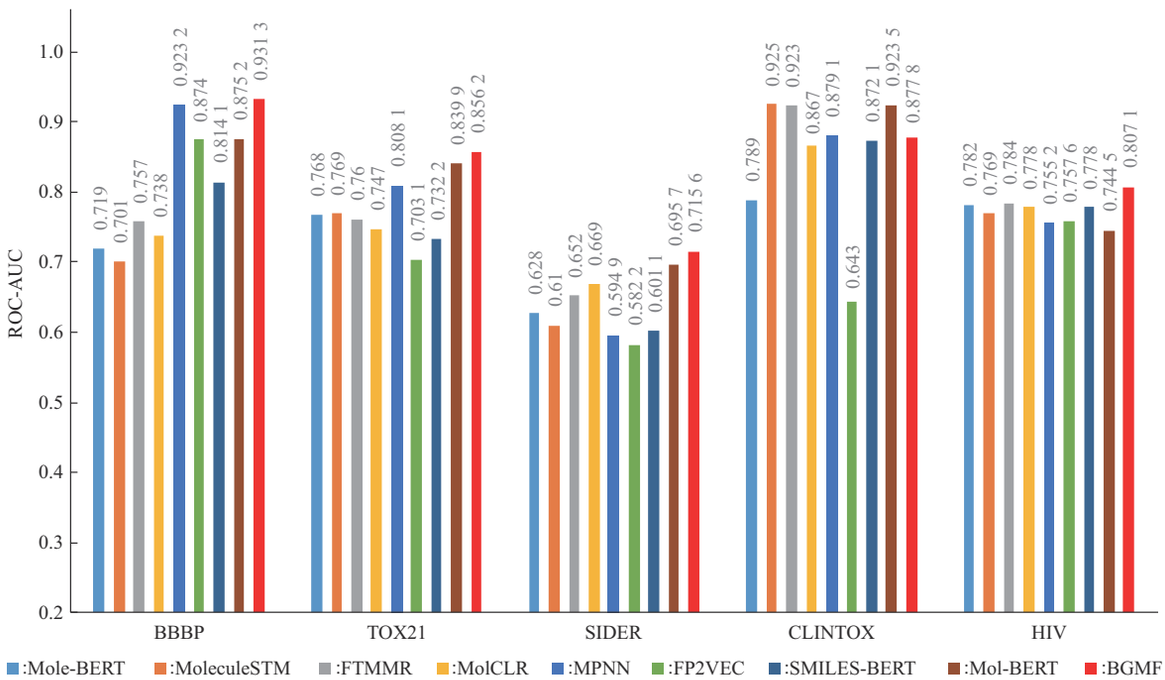


Fig. 2　The performance comparison of BGMF and other methods

BGMF（w/o FP）；第三个是将基于GCN的分子图模块删除掉，记为BGMF（w/o GCN）；第四个是将预训练模块删除掉，记为BGMF（w/o Pretrain）。将BGMF和4个变体模型在5个数据集上分别进行测试。由文档S1第**4**节、图S2可见：a. 同时集成原子序列、分子指纹序列和基于GCN的分子图模块的BGMF具有最优的预测性能；b. 删除模型中的原子序列、删除模型中的指纹序列、删除基于GCN的分子图模块和删除预训练模块均会导致模型性能的下降。其中，对于SIDER数据集，删除GCN模块的Bert模型性能最差，而其他4个数据集中，删除分子指纹模块的BGMF（w/o FP）性能最差，由此可见，分子指纹序列中子结构全局特征学习和基于GCN的分子图局部拓扑特征学习对提高分子属性预测的重要性。在5个数据集中，删除预训练模块均对模型有很大的影响。

为了深入剖析模型在预训练阶段捕获特征的有效性，本文将模型生成的分子指纹表征进行可视化。具体来说将从预测数据集Tox21中随机选取的药物分子作为研究样本。将600个分子输入BGMF模型，经过编码层后得到指纹嵌入向量，之后利用t分布式随机邻居嵌入（t-distributed stochastic neighbor embedding，t-SNE）算法[36]，将指纹嵌入向量映射到二维空间，以便直观地分析模型的学习性能。部分分子指纹表征的可视化结果如图3所示，图中的每一种颜色代表一种独特的分子指纹，各种颜色所代表的指纹及其序号对应图4。图3中的序号4对应［MASK］标记的可视化表征。由图3可见，各个指纹簇内的点紧密聚集，而不同指纹簇之间的点则相对分散，边界清晰。由此表明，本模型可以成功捕获不同分子指纹的内在结构与特征。

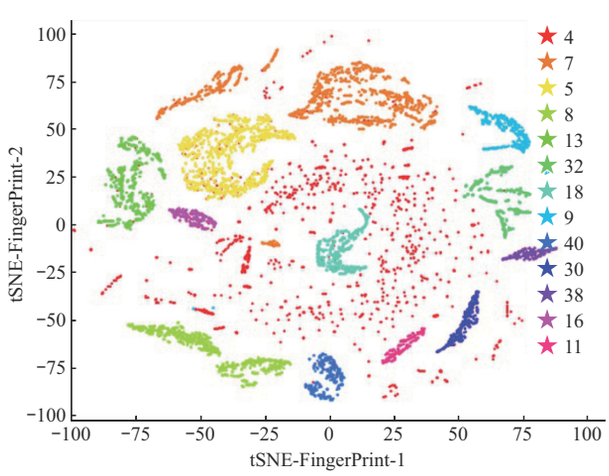


Fig. 3　Visualization of partial molecular fingerprint representation

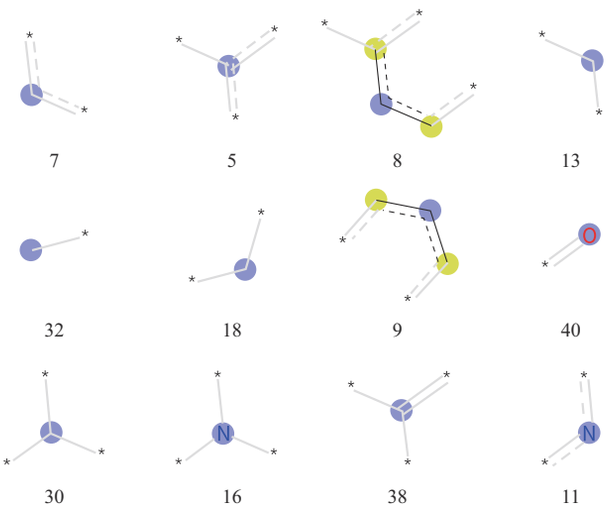


Fig. 4　Visualization of partial molecular fingerprint

3　结　　论

本文基于药物分子的原子序列、分子指纹序列和分子图，提出了一种联合Bert和GCN模型的分子属性预测BGMF算法。该算法包括预训练和微调两部分，预训练数据集整合了ChEMBL和ZINC数据库中的化合物，通过Bert模型的掩蔽操作和双解码器模块设计，学习原子序列和分子指纹序列的特征表征，在多头注意力机制部分集成了基于GCN的分子图拓扑特征，可同时捕获分子全局序列上下文语义关系和局部拓扑特征。下游任务微调中包含池化层、MLP层和交叉熵损失模块3部分。冻结预训练模块后，在5个经典分子属性数据集43个任务上进行药物分子属性预测，与其他方法的对比实验结果表明，BGMF具有较高的预测精度。对BGMF所采用的4个策略分别进行消融实验，结果表明，分子指纹序列中子结构全局特征学习和基于GCN的分子图局部拓扑特征学习对提高分子属性预测具有重要作用。同时，通过使用t-SNE算法进行可视化展示发现，使用BGMF模型学习到的分子指纹表征可以成功捕获不同分子指纹的内在结构与特征，具有各个指纹簇内的点紧密聚集，而不同指纹簇之间的点则相对分散，边界清晰的特点。

尽管提出的BGMF模型在药物分子属性预测方面取得了成功，但是仍然有改进的空间，模型在独立样本测试中的性能有待提高，由于模型包括预训练和下游任务微调两个模块，参数较多，训练时间较长。此外，我们的分子指纹特征和原子特征之间没有进行特征交互学习的融合操作，只有一种初步的融合。在将来的研究中，将进一步探索特征的相关性等融合策略，并考虑采用参数共享机制解决这些问题。同时希望可以增加更多的化学知识，如化合物的3D结构信息，原子性质等在模型中的应用，并进一步探索模型在药物分子属性预测回归任务中的应用。

附件　见本文网络版（http://www.pibb.ac.cn, http://www.cnki.net）：

PIBB_20240299_Table S1.pdf

PIBB_20240299_Table S2.pdf

PIBB_20240299_Figure S1.pdf

PIBB_20240299_Figure S2.pdf

PIBB_20240299_Document S1.pdf

参　考　文　献

[1] Zheng Z, Tan Y, Wang H, *et al*. CasANGCL: pre-training and fine-tuning model based on cascaded attention network and graph contrastive learning for molecular property prediction. Brief Bioinform, 2023, **24**(1): bbac566

[2] Liyaqat T, Ahmad T, Saxena C. Advancements in molecular property prediction: a survey of single and multimodal approaches. arXiv, 2024: 2408.09461. https://arxiv.org/abs/2408.09461v2

[3] Wu Z, Ramsundar B, Feinberg E N, *et al*. MoleculeNet: a benchmark for molecular machine learning. Chem Sci, 2018, **9**(2): 513-530

[4] Puzyn T, Leszczynski J, Cronin M T. Recent Advances in QSAR Studies, Methods and Applications. New York: Springer, 2010: 20-102

[5] Todeschini R, Consonni V. Handbook of Molecular Descriptors. Weinheim, Germany: John Wiley & Sons, 2008

[6] Vaswani A, Shazeer N, Parmar N, *et al*. Attention is all you need// Curran Associates Inc. Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, California, USA: Curran Associates, 2017: 6000-6010

[7] Floridi L, Chiriatti M. GPT-3: its nature, scope, limits, and consequences. Mines Mach, 2020, **30**(4): 681-694

[8] Szegedy C, Vanhoucke V, Ioffe S, *et al*. Rethinking the inception architecture for computer vision//IEEE/CVF. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ, USA: IEEE, 2016: 2818-2826

[9] He K, Zhang X, Ren S, *et al*. Identity mappings in deep residual networks//Leibe B, Matas J, Sebe N, *et al*. European Conference on Computer Vision. Cham, Switzerland: Springer International Publishing, 2016: 630-645

[10] Xu Z, Wang S, Zhu F, *et al*. Seq2seq Fingerprint: an unsupervised deep molecular embedding for drug discovery//ACM SIGBio.

Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics. Boston, Massachusetts, USA: Association for Computing Machinery, 2017: 285-294

[11] Lv Q, Chen G, Zhao L, *et al*. Mol2Context-vec: learning molecular representation from context awareness for drug discovery. Brief Bioinform, 2021, **22**(6): bbab317

[12] Chen J H, Tseng Y J. Different molecular enumeration influences in deep learning: an example using aqueous solubility. Brief Bioinform, 2021, **22**(3): bbaa092

[13] Karpov P, Godin G, Tetko I V. Transformer-CNN: Swiss knife for QSAR modeling and interpretation. J Cheminform, 2020, **12**(1): 17

[14] Abbassi O, Ziti S, Belhiah M, *et al*. GMPP-NN: a deep learning architecture for graph molecular property prediction. Discov Appl Sci, 2024, **6**(7): 352

[15] Xiong Z, Wang D, Liu X, *et al*. Pushing the boundaries of molecular representation for drug discovery with the graph attention mechanism. J Med Chem, 2020, **63**(16): 8749-8760

[16] Han S, Fu H, Wu Y, *et al*. HimGNN: a novel hierarchical molecular graph representation learning framework for property prediction. Brief Bioinform, 2023, **24**(5): bbad305

[17] Liu S, Nie W, Wang C, *et al*. Multi-modal molecule structure-text model for text-based retrieval and editing. Nat Mach Intell, 2023, **5**: 1447-1457

[18] Son Y H, Shin D H, Kam T E. FTMMR: fusion transformer for integrating multiple molecular representations. IEEE J Biomed Health Inform, 2024, **28**(7): 4361-4372

[19] Ma M, Lei X. A deep learning framework for predicting molecular property based on multi-type features fusion. Comput Biol Med, 2024, **169**: 107911

[20] Bouritsas G, Frasca F, Zafeiriou S, *et al*. Improving graph neural network expressivity *via* subgraph isomorphism counting. IEEE Trans Pattern Anal Mach Intell, 2023, **45**(1): 657-668

[21] Zhang H, Wu J, Liu S, *et al*. A pre-trained multi-representation fusion network for molecular property prediction. Inf Fusion, 2024, **103**: 102092

[22] Devlin J, Chang M W, Lee K, *et al*. BERT: Pre-training of deep bidirectional transformers for language understanding// Association for Computational Linguistics. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Minneapolis, Minnesota: Association for Computational Linguistics, 2019: 4171-4186

[23] Zhang X C, Wu C K, Yang Z J, *et al*. MG-BERT: leveraging unsupervised atomic representation learning for molecular property prediction. Brief Bioinform, 2021, **22**(6): bbab152

[24] Wang S, Guo Y, Wang Y, *et al*. SMILES-BERT: large scale unsupervised pre-training for molecular property prediction// ACM. Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics. Niagara Falls, NY, USA: Association for Computing Machinery, 2019: 429-436

[25] Li J, Jiang X. Mol-BERT: an effective molecular representation with BERT for molecular property prediction. Wirel Commun Mob Comput, 2021, **2021**(1):1-7

[26] Wang Y, Wang J, Cao Z, *et al*. Molecular contrastive learning of representations *via* graph neural networks. Nat Mach Intell, 2022, **4**: 279-287

[27] Xia J, Zhao C, Hu B, *et al*. Mole-BERT: rethinking pre-training graph neural networks for molecules//IEEE Information Theory Society. Proceedings of the International Conference on Learning Representations. KigaliRwanda: ICLR, 2023. https://api.semanticscholar.org/CorpusID:259298531

[28] Oord A V D, Vinyals O, Kavukcuoglu K. Neural discrete representation learning//Curran Associates Inc. Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, California, USA: Curran Associates, 2017: 6309-6318

[29] Mendez D, Gaulton A, Bento A P, *et al*. ChEMBL: towards direct deposition of bioassay data. Nucleic Acids Res, 2019, **47**(D1): D930-D940

[30] Irwin J J, Sterling T, Mysinger M M, *et al*. ZINC: a free tool to discover chemistry for biology. J Chem Inf Model, 2012, **52**(7): 1757-1768

[31] Vilar S, Friedman C, Hripcsak G. Detection of drug-drug interactions through data mining studies using clinical sources, scientific literature and social media. Brief Bioinform, 2018, **19**(5): 863-877

[32] Kipf T N, Welling M. Semi-supervised classification with graph convolutional networks. arXiv, 2016: 1609.02907. https://arxiv.org/abs/1609.02907v4

[33] He K, Chen X, Xie S, *et al*. Masked autoencoders are scalable vision learners//IEEE CS. Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2022. OrleansNew, United States: IEEE Computer Society, 2022: 15979-15988

[34] Gilmer J, Schoenholz S S, Riley P F, *et al*. Neural message passing for quantum chemistry. arXiv, 2017: 1704.01212. https://arxiv.org/abs/1704.01212v2

[35] Jeon W, Kim D. FP2VEC: a new molecular featurizer for learning molecular properties. Bioinformatics, 2019, **35**(23): 4979-4985

[36] Van Der Maaten L, Hinton G. Visualizing data using t-SNE. J Mach Learn Res, 2008, **9**: 2579-2605

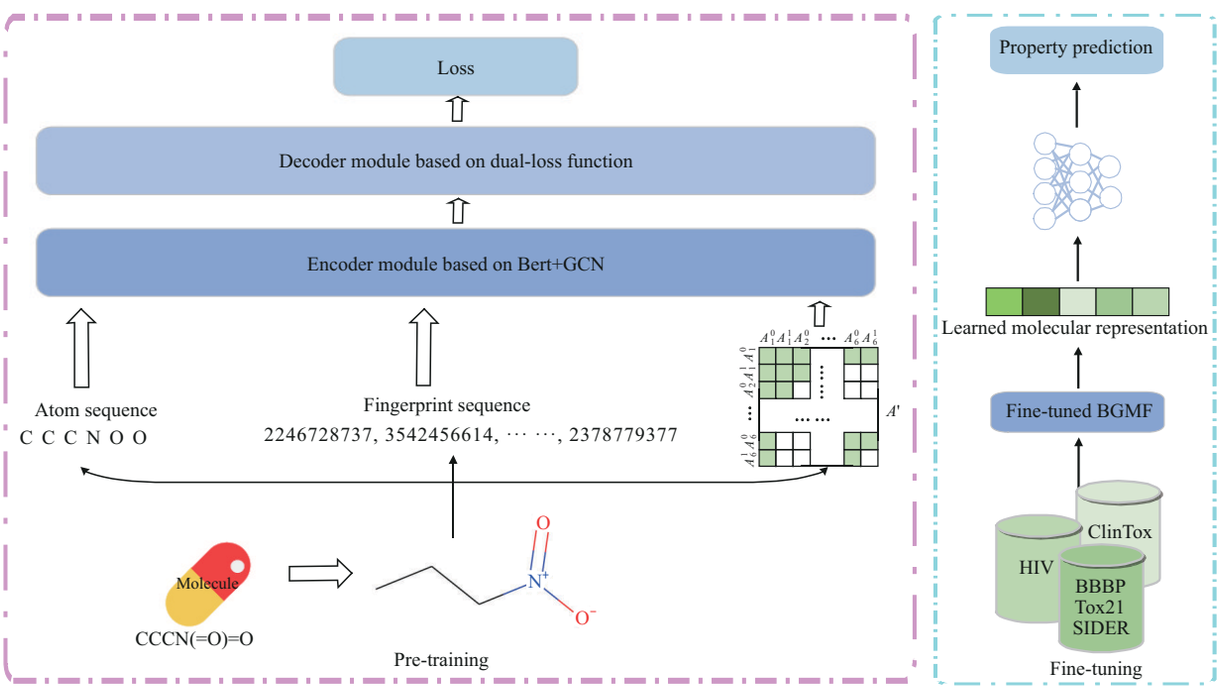
A Multimodal Fusion Drug Molecular Attribute Prediction Method Based on Bert and GCN*

YAN Xiao-Ying[1)**], JIN Yan-Chun[1)], FENG Yue-Hua[1)], ZHANG Shao-Wu[2)**]

([1)]*College of Computer Science, Xi'an Shiyou University, Xi'an* 710065, China;

[2)]*School of Automation, Key Laboratory of Information Fusion Technology of Ministry of Education, Northwestern Polytechnical University, Xi'an* 710072, China)

Graphical abstract



Abstract　Objective　Molecular property prediction plays a crucial role in drug development, especially in virtual screening and compound optimization. The advancement of artificial intelligence (AI) technologies has led to the emergence of numerous deep learning-based methods, which have demonstrated significant potential in improving molecular property prediction. Nonetheless, acquiring labeled molecular data can be both costly and time-consuming. The scarcity of labeled data poses a substantial challenge for supervised machine learning models to effectively generalize across the vast chemical space. In order to overcome the above limitations, in this work, we proposed a novel Bert and GCN-based multimodal fusion method (called BGMF) to predict molecular property. **Methods**　BGMF can extract comprehensive molecular representation from atomic sequences, molecular fingerprint sequences, and molecular graph data and combine them through pre-training and fine-

* This work was supported by grants from The National Natural Science Foundation of China (62173271), Shaanxi Natural Science Basic Research Program (2023-JC-YB-591), the Postgraduate Innovation and Practical Ability Training Program of Xi'an Shiyou University (YCS23213171), and the Postgraduate High-quality Case Database Construction Project of Xi'an Shiyou University (2024-X-YAL-003).

** Corresponding author.

YAN Xiao-Ying. Tel：86-29-81469729，E-mail：xiaoying_yan@126.com

ZHANG Shao-Wu. Tel：86-29-88431308，E-mail：zhangsw@nwpu.edu.cn

Received：July 6，2024　Accepted：December 2，2024

tuning. Specifically, our method consists of the following three main parts. (1) Molecular feature extraction; (2) Bert-GCN based pre-training; (3) fine-tuning. During molecular feature extraction, the Morgan algorithm is employed to generate the molecular fingerprints, transforming input SMILES strings of drugs into molecular fingerprint sentences. Simultaneously, atom sentences are created based on the atom indices within the molecule, Consequently, drug molecule are represented as both molecular fingerprint sentences and atom sentences. In the pre-training section, BGMF utilizes a self-supervised learning strategy, specifically masked molecular fingerprint and masked atom recovery, on a large dataset of unlabeled data using the Bert model. Here, molecular graph data is incorporated by merging graph convolutional neural networks with the Bert model, effectively combining the global “word” features of drug molecules with the local topological features of molecular graphs. We have also developed a dual decoder for atomic and molecular fingerprints to amplify molecular feature expression. Finally, in the fine-tuning stage, the addition of a pooling layer and task-specific fully connected neural networks allows the pre-trained module to be applied to a variety of downstream tasks for molecular property prediction. **Results** To validate the effectiveness of our BGMF, we conduct several experiments on 43 molecular attribute prediction tasks across 5 datasets. In comparison with other recent state-of-the-art methods, our BGMF achieves the best results in terms of area under the ROC curve (AUC). We also verified the generalization performance of the BGMF model by constructing independent test dataset, showing that the BGMF model has the best generalization performance. Additionally, we conduct the ablation studies to demonstrate the effect of atomic sequence, molecular fingerprint sequence, GCN based molecular graph module, and pre-training module on the overall performance of the model. **Conclusion** In this paper, we propose a novel method for drug molecular attribute prediction named BGMF which integrating the molecular graph data into tasks of molecular fingerprint recovery and masked atom recovery by combining graph convolutional neural network with the Bert model. The molecular fingerprint representations generated by BGMF were visualized using t-SNE, revealing that the BGMF model effectively captures the intrinsic structure and features of molecular fingerprints.

Key words Bert pretraining, attention mechanism, molecular fingerprint, molecular attribute prediction, graph convolutional neural network

DOI：10.16476/j.pibb.2024.0299 **CSTR**：32369.14.pibb.20240299