基于知识先验和多通道注意力机制的生物活性预测模型

李 梦[1,2],唐文燕[1,2]
1. 重庆工商大学 数学与统计学院,重庆 400067
2. 统计智能计算与监测重庆市重点实验室,重庆 400067

摘 要:目的 针对现有药物研发中分子活性值预测不精、泛化性不高等问题,提出基于知识先验与注意力机制相结合的多通道语义深度神经网络,通过使用分子的 SMILES(Simplified Molecular Input Line Entry System)表达式,预测雌激素受体 α 亚型(ERα)的 pIC50 生物活性值。**方法** 该网络采用两阶段特征提取策略,在语义层设计了将知识先验与迁移学习结合的语义分析网络,它将分子 SMILES、描述符和图表征的关键信息定位,通过在 Erα 数据集中微调参数,得到综合的分子 SMILES 表征信息;在通道层,基于高效通道注意力(Efficient Channel Attention, ECA)机制,设计了 1D-ECA 算法,将其嵌入 CNN 子模块中,构成多通道深度神经 1D-ECA-CNN 模块,实现分子表征的特征再提取,并减少分子表示学习过程中的信息损失;最后将语义层和通道层相结合形成 KBAC(Knowledge-BERT-1D-ECA-CNN)深度神经网络,实现 pIC50 生物活性值的回归预测。**结果** 实验结果表明:所提出的框架在 4 个评估指标上均表现优异,MAE 可达 0.091,MSE 可达 0.014,RMSE 可达 0.117,R^2 可达 0.993,相对于 4 个具有代表性的模型有较为明显的提升,说明所提模型具有更高的预测精度。**结论** 该两阶段特征提取过程使其能够获取更为全面的分子特征,从而帮助筛选治疗疾病的候选药物。

关键词:SMILES;Knowledge-BERT;多通道注意力机制;KBAC;pIC50 预测
中图分类号:R9 **文献标识码:**A doi:10.16055/j.issn.1672-058X.2026.0002.006

A Bioactivity Prediction Model Integrating Knowledge Priors and Multi-Channel Attention

LI Meng[1,2], TANG Wenyan[1,2]
1. School of Mathematics and Statistics, Chongqing Technology and Business University, Chongqing 400067, China
2. Chongqing Key Laboratory of Statistical Intelligent Computing and Monitoring, Chongqing Technology and Business University, Chongqing 400067, China

Abstract: Objective To tackle the problems of inaccurate prediction of molecular activity values and low generalizability in current drug R&D, a multi-channel semantic deep neural network based on the combination of knowledge prior and attention mechanism is proposed. By using the simplified molecular input line entry system (SMILES) expressions of molecules, this network can predict the pIC50 bioactive values of estrogen receptor α isoform(ERα). **Methods** This network adopted a two-stage feature extraction strategy. At the semantic layer, a semantic analysis network that combined knowledge prior and transfer learning was designed. It located the key information of molecular SMILES, descriptors, and graph representations. By fine-tuning the parameters in the Erα dataset, comprehensive molecular SMILES representation information was obtained. At the channel layer, based on the efficient channel attention (ECA) mechanism, a 1D-ECA algorithm was designed and embedded into the CNN sub-module. This formed a multi-channel deep neural 1D-ECA-CNN module, which re-extracted the features of molecular representation and reduced the information loss during the molecular representation learning process. Finally, by combining the semantic layer and the channel layer, a Knowledge-BERT-

收稿日期:2024-03-22 **修回日期:**2024-05-23 **文章编号:**1672-058X(2026)02-0043-11
基金项目:重庆市自然科学基金面上项目资助(CSTC2020JCYJ-MSXMX0162).
作者简介:李梦(1973—),女,四川开江人,博士,教授,从事大数据分析及应用研究.
通信作者:唐文燕(1999—),女,重庆人,硕士研究生,从事大数据分析及应用研究. Email: 997376054@qq.com.
引用格式:李梦,唐文燕. 基于知识先验和多通道注意力的生物活性预测模型[J]. 重庆工商大学学报(自然科学版),2026,43(2):43-53.
LI Meng, TANG Wenyan. A bioactivity prediction model integrating knowledge priors and multi-channel attention[J]. Journal of Chongqing Technology and Business University (Natural Science Edition), 2026, 43(2): 43-53.

1DECA-CNN (KBAC) deep neural network was formed to achieve the regression prediction of the pIC50 biological activity value. **Results** Experimental results demonstrated the superior performance of the proposed framework across four evaluation metrics. The proposed network achieved a mean absolute error (MAE) of 0.091, a mean squared error (MSE) of 0.014, a root mean squared error (RMSE) of 0.117, and a coefficient of determination (R^2) of 0.993. These results represent a significant improvement over four representative baseline models, indicating higher prediction accuracy of the proposed network. **Conclusion** The two-stage feature extraction process enables the acquisition of more comprehensive molecular features, facilitating the screening of candidate drugs for treating diseases.

Keywords: SMILES; Knowledge-BERT; multi-channel attention mechanism; KBAC; pIC50 prediction

在抗乳腺癌药物研发中,雌激素受体 α(ERα)的 pIC50 生物活性预测,对于分析药物分子性质,分解药物分子和疾病的内在联系,设计基于治疗靶点的 AI 药物和研发癌症潜在治疗药物具有重要作用。但由于分子性质和结构的复杂多样,使得传统的 pIC50 生物活性检测依赖于化学家的专业知识和启发式方法,以及各种药物实验和人工特征提取,往往存在着周期长、成本高、副作用大等问题,并因其分子性质的不可预见性,导致检测结果不精或失败,造成研发成本的增加。因此,具备高效自动化能力的机器学习和深度学习等计算机辅助方法,在药物研发中备受青睐。

生物活性的预测主要通过分析药物分子描述符、指纹[1-4]、图[5-10]和 SMILES[11-15]表征等,去挖掘分子性质和结构的特征信息,从而实现后续任务的精准预测。如 Teng 等[16]提出基于多种分子指纹的分子指纹图 Transformer 框架,进行特征学习和毒性预测。这类方法使用数值描述符表示分子的特征,具有易于特征提取和建模的优势,但其严重依赖于专家知识的指导来生成描述符和分子指纹。Li 等[17]对分子图进行图级特征提取,进而预测药物性质以及相互作用。这类方法通过分析分子的图结构来实现特征提取,拥有清晰的结构信息和上下文感知,但其结果取决于样本数据的大小,以及分子结构的复杂性和变异性。Ross 等[18]提出使用旋转位置嵌入和线性注意力机制训练 Transformer 编码模型,捕获分子结构和化学信息用于分子性质预测。这类方法使用 SMILES 字符串中的原子类型、键连接等信息来预测分子的性质,具有简洁的表示和高效的数据处理优势,但其缺乏明确的化学信息,如分子描述符、指纹和分子图。因此,上述方法不能从语义角度充分理解分子表征和结构的复杂性,从而影响后续任务的预测效果。

另一方面,为进一步分析分子的化学含义,基于语义的模型也被提出用于分子表征学习。如 Wu 等[19]提出名为 Knowledge-BERT(K-BERT)的框架,将分析描述符/指纹、图和 SMILES 的方法结合起来,用于 3 个任务阶段的分子表示学习,并在约 180 万个数量的大规模分子数据集上进行了预训练,使其具有显著的自然语言处理能力。这种综合学习策略旨在提高模型对分子特征的理解和表示能力,但忽略了分子重要特征的特别关注,使其对药物分子性质的理解仍然缺乏泛化性。

为应对这些问题,本文基于知识先验与注意力机制,提出多通道语义深度神经网络模型 KBAC,以实现抗乳腺癌药物分子 pIC50 生物活性值的准确预测。所提模型包括两阶段的特征提取:第一阶段,基于语义模型分析分子和原子层面上的特征,并在 ERα 数据集上微调参数,使其能够综合理解药物分子的化学性质;第二阶段,基于高效通道注意力机制,构建一维多通道注意力子模块(1D-ECA),并将其嵌入到 CNN 子模块中形成 1D-ECA-CNN 模块,通过第二次提取获取药物分子的全局-局部特征,并丰富 SMILES 分子表征信息;然后,将 K-BERT 模块和 1D-ECA-CNN 模块结合,提出用于预测药物分子活性的 KBAC 网络。与传统方法不同,该方法通过知识先验和多通道注意力机制实现分子 SMILES 的两阶段不同角度特征提取,对于解释分子属性并实现准确预测具有重要意义。

本文主要贡献如下:

(1) 提出基于知识先验和多通道注意力机制的 KBAC 深度学习框架,用于预测 ERα 分子的 pIC50 生物活性值,通过两阶段特征提取来提高模型的泛化能力和全面提取特征的能力。

(2) 在 KBAC 框架中引入 K-BERT 模块,利用已有领域的知识经验,指导 K-BERT 框架对 ERα 数据集进行迁移学习和分析,从而理解药物分子性质,并提取综合特征。

(3) 基于多通道注意力机制,设计 1D-ECA 子模块,使得一维数据能够在多通道之间进行信息交互,并获取分子的全局特征,而后将 1D-ECA 嵌入到 CNN 子模块中形成 1D-ECA-CNN 模块,引导网络二次提取分子特征,从而获得药物分子的全局-局部特征。

(4) KBAC 框架利用不同模块从不同角度对分子表征进行多次特征提取,提高模型对 SMILES 表征的特征识别和预测精度,并为虚拟筛选和药物设计研究提供有用工具。实验结果说明了模型的有效性。

1 相关工作

1.1 基于描述符和指纹的方法

基于描述符和指纹的方法通过将分子表示为固定长度的数字向量来编码各种分子的性质和结构特征。

如 Morgan 指纹和扩展连通性指纹(Extended Connectivity Finger-prints,ECFP),在分子性质分类、回归和相似性等搜索任务中被广泛使用。Hunt 等[20]通过结合半经验量子力学描述符和机器学习方法,去捕获分子、原子和化学键的性质,进一步预测化合物的酸碱解离常数;Papa 等[21]在基于遗传算法选择的不同理论分子描述符上使用 QSAR 分类方法建模,从而预测人类细胞的 PHA 致突变性;Kumari 等[22]利用分子描述符和指纹构建哺乳动物的雷帕霉素靶激酶抑制剂的预测模型;Teng 等[16]提出基于多种分子指纹技术的分子指纹图 Transformer 框架,用于特征学习和毒性预测。这些基于描述符和指纹的预测方法拥有特征表达简单、解释性强、数据处理便利、可迁移性强等优点,但描述符及其分子指纹的生成需要大量的专家知识指导,生成的过程可能会丢失细粒度的拓扑信息,从而导致预测结果不佳。

1.2　基于图的方法

基于图的方法将分子表示为以原子为节点,边为键的图结构形式进行处理和分析。图卷积网络(Graph Convolutional Networks,GCNs)是最经典的图方法。它利用神经网络架构直接处理图结构的分子数据,通过在图上执行卷积操作,汇总分子局部邻域的信息来生成新的节点特征,从而捕获邻域信息,学习分子结构的层次表示。如 Li 等[17]提出 MPG 学习框架用于药物发现任务,对分子图进行图级特征提取,进而预测药物性质以及相互作用等;Yu 等[23]提出包含基序节点和分子节点的异构基序图框架,用于分子表征学习;Xia 等[24]将知识图谱、基因表达谱和结构信息融合起来,用于预测药物靶标相互作用;Li 等[25]提出双视图框架,通过迭代使用局部和全局表示学习模块来预测药物间的相互作用。基于图的方法可以学习分子结构的层次表示,并有效建模邻域信息。但模型结果依赖于样本数据量的大小,且分子结构的复杂多变使得图的方法在预测中易于过拟合,导致泛化性能差等问题。

1.3　基于 SMILES 的方法

SMILES 表达式[26]是一种用于描述分子结构的文本表示方法,通过把原子和键用特定的符号表示,将分子结构转换为易于理解和处理的字符串形式。因此,可应用自然语言处理技术,从表达式中涵盖的原子类型、键连接等信息提取分子特征。如 SMILES2Vec,它将每个 SMILES 字符串视为由表示原子和键的“单词”组成的“句子”,进而学习分布式表示。Ross 等[18]提出使用旋转位置嵌入和线性注意力机制训练 Transformer 编码模型从而获得分子嵌入,有效捕获分子结构和化学信息用于化学分子的性质预测;Hua 等[27]提出多功能鲁棒(MFR-DTA)模型,用于预测蛋白质分子和药物的结合区域;Shao 等[28]提出通过将 SMILES 转换为药物向量来预测抗 HBV 活性的网络;Zhao 等[29]提出基于序列模型,通过注意力机制来预测药物与靶标的亲和性。SMILES 表示法的简单性和通用性使得它在分子建模中备受青睐。但由于其不像分子描述符、指纹以及分子图等有明确的化学信息和分子结构,因此需要更大的数据量以及更深层次的特征提取能力。

尽管这些方法在使用各类分子表征提取特征方面表现出良好的性能,但仍然依赖于样本数据的大小和分子结构的复杂性,从而导致在分子性质的准确预测任务中,产生一定的泛化限制。因此,本文提出基于知识先验和多通道注意力机制的 KBAC 框架,用于预测 pIC50 的生物活性值,并提高模型的泛化能力和准确预测能力。

2　系统结构设计与算法实现

2.1　整体框架

本章基于知识先验与注意力机制,设计多通道语义深度神经网络模型(KBAC),用于 ERα 化合物的两阶段特征学习和 pIC50 的生物活性预测,其整体架构如图 1 所示。

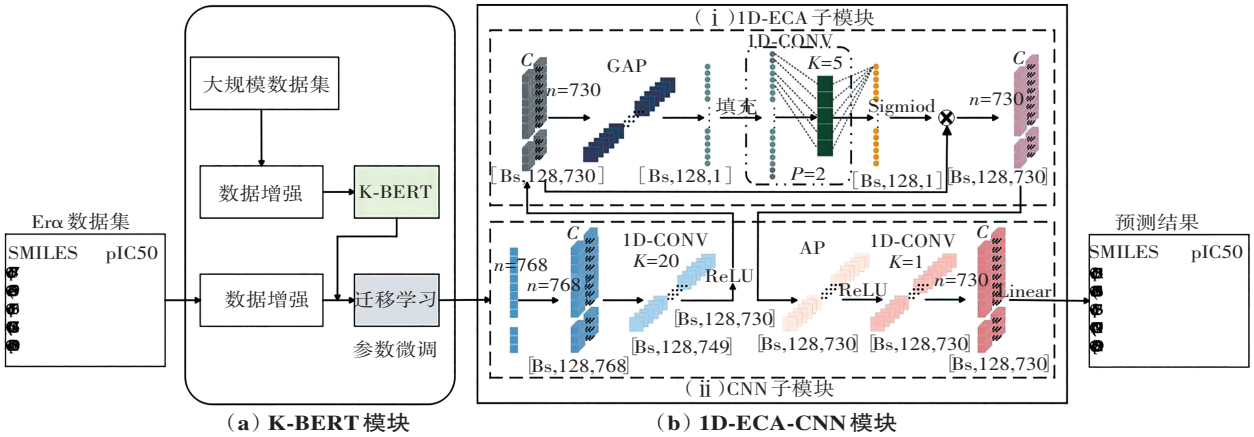

图 1　KBAC 整体框架图

Fig. 1　Overall framework of KBAC

如图 1 所示,KBAC 框架的构建主要包括以下部分:基于知识先验的分子特征提取和基于多通道注意力机制的二次特征提取。基于知识先验的分子特征提取,引入 K-BERT 网络和迁移学习策略,将基于描述符、图形和 SMILES 的方法结合起来,以获取综合的 SMILES 分子表征信息。基于多通道注意力机制的模块中,设计的 1D-ECA-CNN 模块,通过二次提取来获取 ERα 分子化合物的全局-局部特征,其中包括 1D-ECA 子模块和 CNN 子模块。

2.2 基于知识先验的分子特征提取

药物分子的 SMILES 表达式将分子结构转换为易于理解和处理的文本字符串形式,以简单直观的方式描述了分子化合物的原子和键。通过分析 SMILES 表达式中的原子和基团特征,例如氨基的 SMILES 表达式:"[NH2]",表示一个氮原子和两个氢原子;苯环的 SMILES 表达式:"c1ccccc1",表示一个芳香环,每个碳原子之间有一个共轭双键;氟乙酰基的 SMILES 表达式:"FC(=O)",表示一个氟原子连接在一个碳原子上,其中碳原子通过双键与一个氧原子相连。利用自然语言处理技术对 SMILES 表达式进行分析,能够精准解读其中蕴含的复杂分子结构信息,从而有效提取分子特征的关键要素,有助于提高药物研发和分子设计的效率和准确性。

2.2.1 K-BERT 子模块

K-BERT 模型[19] 由 Wu 等提出,该网络将基于描述符/指纹、图和 SMILES 的方法结合起来,利用大规模分子数据集训练,并学习其中包含的专业化学知识,能够更好地分析分子的全局和局部特征信息。本节采用 K-BERT 自然语言处理模型,将其在大规模分子数据集上得到的预训练结果作为分子性质的化学知识先验,从分子的 SMILES 表达式中高效完成第一次特征提取,为分子生物活性值的预测提供综合特征信息。K-BERT 网络由 6 个 Transformer 编码层组成,其中分子表征学习的关键在于以下 3 个预训练任务,具体如图 2 所示。

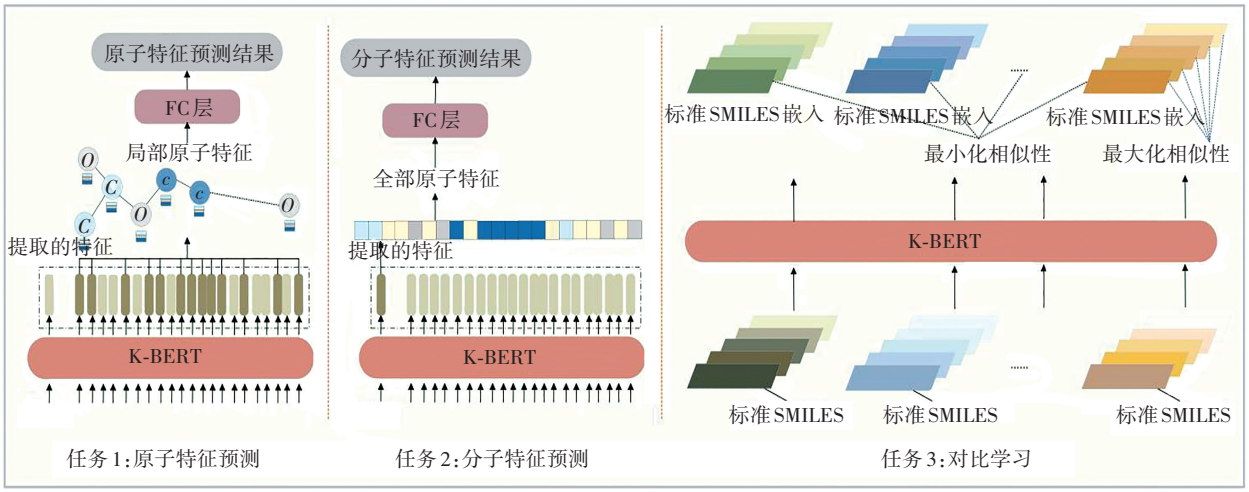

图 2 K-BERT 框架图

Fig. 2 Framework chart of K-BERT

第一阶段采用图的方法对 SMILES 表达式中每个重原子进行学习并提取原子特征信息。

第二阶段采用分子描述符方法或指纹方法学习,对 SMILES 表达式中每个重原子进行学习并提取分子特征信息。

第三阶段在前两个阶段的基础上,将基于描述符/指纹和基于图的方法中人工生成的信息作为知识先验,进一步对同一分子的不同 SMILES 表达式进行对比学习。目的是最大化同一分子的不同 SMILES 表达式之间嵌入的余弦相似性,同时最小化不同分子之间嵌入的相似性。

通过 3 个方法的结合使用,帮助模型更好地理解分子 SMILES 表达式的特征信息,使得模型能够从药物分子的 SMILES 表达式中获取更全面的特征信息,提升 K-BERT 模型的泛化能力,为药物研发工作提供帮助。

2.2.2 损失函数

为使得同一化合物分子的不同 SMILES 表达式的嵌入变得更加相似,在对比学习任务中采用如下损失函数。

$$L_{\mathrm{CL}}=\sum_{n=1}^{N}\sum_{d\in D_n}\frac{1}{2}(1-\mathrm{COS}(\boldsymbol{E}_{n,c},\boldsymbol{E}_{n,d}))+\sum_{n=1}^{N}\sum_{m\in\beta_n}\mathrm{COS}(\boldsymbol{E}_{n,c},\boldsymbol{E}_{n,m}) \tag{1}$$

$$\mathrm{COS}(\boldsymbol{E}_a,\boldsymbol{E}_b)=\frac{\boldsymbol{E}_a\cdot\boldsymbol{E}_b}{\|\boldsymbol{E}_a\|\ \|\boldsymbol{E}_b\|} \tag{2}$$

$$\|\boldsymbol{E}_a\|=\sqrt{E_{a,1}^2+E_{a,2}^2+\cdots+E_{a,p}^2}\ |^2 \tag{3}$$

其中,n 代表当前的分子,N 表示这个批次中的分子数量,D_n 表示由分子 n 的标准 SMILES 表达式产生的 4 种

不同的 SMILES 表达式,B_n 表示这个批次中除了分子 n 以外的其他分子,$\boldsymbol{E}_{n,c}$ 表示由 K-BERT 模块生成的分子 n 的标准 SMILES 的嵌入,$\boldsymbol{E}_{n,d}$ 表示由 K-BERT 模块生成的分子 n 的 4 个增强 SMILES 表达式的嵌入,$\boldsymbol{E}_{n,m}$ 表示由 K-BERT 模块产生的分子 m 的标准 SMILES 表达式的嵌入,COS( )是衡量两个嵌入之间相似性的余弦相似度函数,$\| \boldsymbol{E}_a \|$ 是嵌入 $\boldsymbol{E}_a=(E_{a,1},E_{a,2},\cdots,E_{a,p})$ 的欧氏范数。

2.2.3　参数微调

在获取药物分子 SMILES 表达式特征信息的同时,还需加速机器学习和深度学习任务的完成,减少数据和计算资源的需求,帮助目标域模型更快地收敛,提高模型的泛化性能。

为此,本文将在 K-BERT 网络上针对 ERα 数据集进行参数微调及重新训练,通过 6 个 Transformer 编码层进行迁移学习,捕捉分子 SMILES 表达式中的特征和关系。Transformer 编码器层由自注意力机制子层和前馈神经网络子层交替组成,子层之间包含残差连接和层归一化操作,能够学习输入序列的复杂特征,并促进信息传递。针对迁移学习任务,首先,固定前 5 个 Transformer 层的参数,将其作为冻结层,保留它们在先前任务中学习得到的知识和权重参数,并将其迁移应用于当前 ERα 数据集的特征提取新任务。这种做法有助于减少训练时间和资源消耗,同时保持较高的性能。其次,为充分发挥模型的潜力并使其适应当前数据集的分子特征提取任务,选择让最后一个 Transformer 层从头开始训练,该层的参数将完全重新初始化,并根据新数据集和预测任务进行优化和更新学习,得到适应当前数据集的网络参数值。通过这种方式,模型可以更好地适应新的输入数据。过程如图 3 所示。

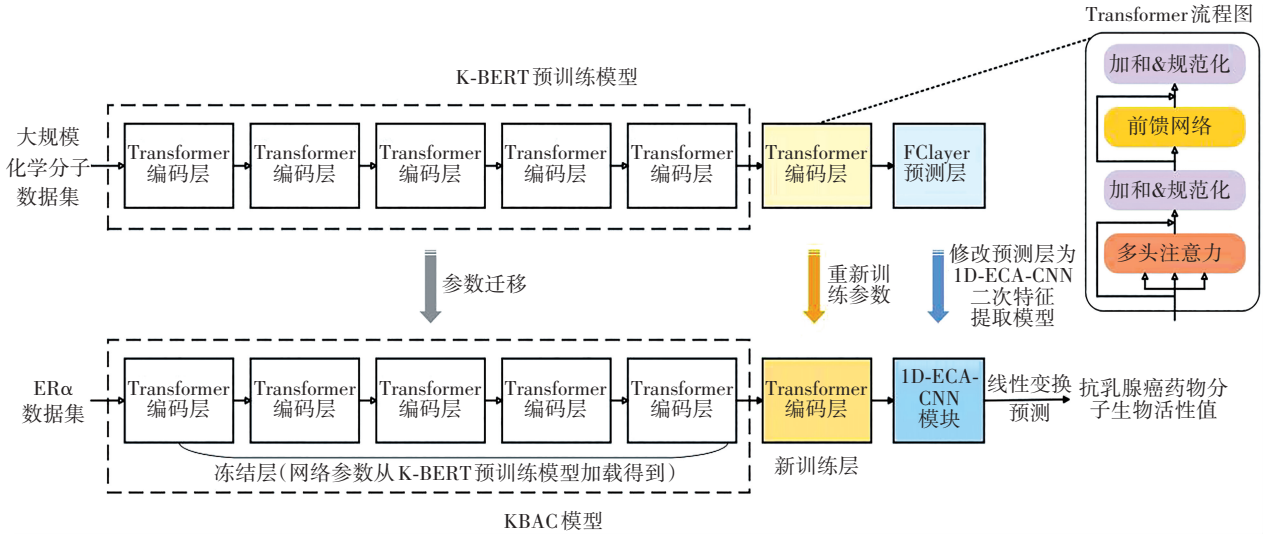

图 3　迁移学习流程图

Fig. 3　Flow chart of transfer learning

2.3　基于多通道注意力机制的二次特征提取模块(1D-ECA-CNN 模块)

为进一步提高分子表征的准确度和稳定性,本节先设计基于多通道注意力机制的 1D-ECA 子模块,再将其嵌入到 CNN 网络,形成 1D-ECA-CNN 模块,对 ERα 数据集的分子 SMILES 表达式进行第二次特征提取。

2.3.1　1D-ECA 子模块

为使模型能够更加关注分子 SMILES 表达式中有关生物活性特征信息的部分,以及平衡模型表现性能和复杂度之间的关系,本节设计 1D-ECA 子模块,用于从一维分子数据中提取全局特征。具体过程如图 1(b)(i)所示。

ECA 算法[30]是一种轻量级注意力机制,可用于捕获二维数据通道之间的依存关系。本节将改变其数据维度,使其适用于当前提取的一维 SMILES 分子特征信息。1D-ECA 子模块的过程包括以下步骤:

首先,将 ECA 算法与从 K-BERT 先验和对 ERα 数据集进行参数微调中导出的一维分子特征信息相集成,并将一维数据的通道维度增加到 768,作为 CNN 子模块的输入;然后,进行全局平均池化以获取聚合特征;接下来,使用大小为 k 的快速一维卷积生成通道权重;最后,通过激活函数,逐元素地将输入数据与通道权重相乘。

1D-ECA 首先采用上一节 K-BERT 子模块进行迁移学习提取的一维分子特征数据作为输入,并将一维数据的通道维度增加到 768,作为 CNN 子模块的输入;再进行全局平均池化获得聚集特征,然后执行核大小为 k 的快速一维卷积来生成通道权重;最后,通过激活函数,逐元素地将输入数据与通道权重相乘。

其中,k 值通过通道维数 C 的映射来自适应地确定,填充通过对 k 值整除以 2 来确定,而通道维数 C 通常为 2 的幂次方,因此可以通过式(4)、式(5)推出:

$$C=\varphi(k)=2^{(\gamma*k-b)} \tag{4}$$

$$k=\varphi(C)=\left|\frac{\log_2 C}{\gamma}+\frac{b}{\gamma}\right|_{\text{odd}} \tag{5}$$

其中,$|t|_{\text{odd}}$ 表示最接近 t 的奇数,在本文中,γ 和 b 的取值分别为 2 和 1,由此可得 k 为 5,填充为 2。对于捕获局部跨信道交互的方法,旨在同时保证效率和效果,可以通过如下矩阵 $\boldsymbol{w}^k$ 来学习通道注意力。

$$\boldsymbol{w}^k=\begin{bmatrix} w^{1,1} & \cdots & w^{1,k} & 0 & 0 & \cdots & \cdots & 0 \\ 0 & w^{2,2} & \cdots & w^{2,k+1} & 0 & \cdots & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & 0 & 0 & \cdots & w^{C,C-k+1} & \cdots & w^{C,C} \end{bmatrix} \tag{6}$$

$$w_i=\sigma\left(\sum_{j=1}^{k} w^j y_i^j\right), y_i^j \in \Omega_i^k \tag{7}$$

其中,y_i 的权重只考虑 y_i 与其 k 个邻居之间的相互作用,Ω_i^k 表示 y_i 的 k 个相邻通道的集合。但这种策略可以通过一个核尺寸为 k 的一维卷积来快速实现。

$$\boldsymbol{w}=\sigma(\text{CONV}_k(\text{GAVG}(\boldsymbol{y}))) \tag{8}$$

因此,1D-ECA 模块公式具体如下:

$$\boldsymbol{M}_{\text{ECA}}=\text{Sigmoid}(\text{CONV}(\text{GAVG}(\boldsymbol{X})))*\boldsymbol{X} \tag{9}$$

式(9)中,$\boldsymbol{X}$ 表示在卷积神经网络预测模块中经过了核大小为 20,通道数为 128 的一维卷积以及 ReLU 激活函数得到的值;GAVG 表示全局平均池化;CONV 表示一维卷积;Sigmoid 表示激活函数。

2.3.2 1D-ECA-CNN 模块

上一节引入了多通道注意力机制,将其应用于深度神经网络模型,可以更加聚焦于对生物活性值预测有贡献的分子特征,提高模型的性能和泛化能力。

本节将 1D-ECA 算法嵌入 CNN 子模块,设计 1D-ECA-CNN 模块,目的是丰富药物分子的特征信息,降低高层次信息的丢失。详细流程见图 1(b)。

由图 1(b)可知,先在 CNN 子模块中,设计卷积核大小为 20,通道数为 128 的一维卷积层;然后采用 ReLU 激活函数学习 SMILES 表达式中的非线性关系,再输入到 1D-ECA 子模块中,对上一阶段的特征信息进行通道信息交互和全局特征提取;之后将提取出来的特征输入到 CNN 子模块中,进行局部平均池化,并使用 ReLU 激活函数进行操作;在网络的最后一个阶段,由一个卷积核大小为 1 的一维卷积层将通道数降为 1,再通过一个全连接层计算输出最后的预测结果,形成 1D-ECA-CNN 模块,其计算公式如下:

$$\boldsymbol{M}_1=\boldsymbol{M}_{\text{ECA}}(\text{ReLU}(\text{CONV1}(\boldsymbol{X}))) \tag{10}$$

$$\boldsymbol{M}_2=\text{LN}(\text{CONV2}(\text{ReLU}(\text{AVG}(\boldsymbol{M}_1)))) \tag{11}$$

其中,$\boldsymbol{X}$ 表示经过迁移学习得到的分子 SMILES 表达式的特征信息;CONV1 表示卷积核大小为 20,通道数为 128 的一维卷积层;ReLU 表示当前使用的激活函数;M_{ECA} 表示 1D-ECA 模块;AVG 表示平均池化函数;CONV1、CONV2 表示卷积核大小为 1,通道数为 128 的一维卷积层;LN 表示线性层。

1D-ECA-CNN 模块利用多通道注意力机制获取分子表征信息,并通过卷积层操作使得不同通道中的信息融合,从而获取更全面更有效的分子特征信息。该模块实现了分子特征的再次提取,避免了仅使用 K-BERT 特征提取导致特征信息不充分的问题,提高了所提 KBAC 模型的预测精度。

2.3.3 损失函数

为衡量预测的准确性,对于 1D-ECA-CNN 回归预测模块,采用均方误差(MSE)损失函数。该损失函数在真实值与预测值之间的误差较大(两者差值>1)的情况下,MSE 会对模型给予更大的惩罚,在误差较小(两者差值<1)的情况下,给予偏小的惩罚,从而使得模型会更加倾向于惩罚较大的情况,对其赋予更大的权重值。损失函数公式如下:

$$L_{\text{R}}=\frac{1}{m}\sum_{i=1}^{m}(y_i-\hat{y}_i)^2 \tag{12}$$

其中,m 表示分子的数量,y_i 表示当前分子的 pIC50 真实值,$\hat{y}_i$ 表示当前分子的 pIC50 预测值。

对于整体的网络框架 KBAC,总体损失函数由对比学习损失函数部分和回归预测损失部分组成,其表达式为

$$L_{\text{T}}=L_{\text{CL}}+L_{\text{R}} \tag{13}$$

其中,L_{CL} 表示对比学习阶段的损失函数,L_{R} 表示 1D-ECA-CNN 回归预测阶段的损失函数。

2.4 算法实现

本文所提 KBAC 框架的训练算法如表 1 所示。

表 1 KBAC 算法流程

Table 1 Flow of KBAC algorithm

算法 1 KBAC 训练步骤
输入:药物分子 SMILES 表达式 $\boldsymbol{X}$;药物分子 pIC50 值 $\boldsymbol{Y}$。
输出:药物分子 pIC50 预测值。
(1) Data augmentation($\boldsymbol{X}$) //数据增强
(2) $\boldsymbol{T}$=Fine-tuning K-BERT($\boldsymbol{X}$) //采用微调策略进行迁移学习,利用 K-BERT 提取分子的全局特征
(3) $\boldsymbol{C}_1$=Relu(Conv1D($\boldsymbol{T}$))
(4) $\boldsymbol{W}$=Sigmoid(Conv1D(GAVG($\boldsymbol{C}_1$)))
(5) $\boldsymbol{A}=\boldsymbol{C}1*\boldsymbol{W}$ // 通道维加权融合
(6) $\boldsymbol{C}_2$=Conv1D(ReLU(AVG($\boldsymbol{A}$)))// 3—6 步是利用 1D-ECA-CNN 神经网络实现分子表征的特征再提取
(7) $\boldsymbol{Y}$_prediction=Linear ($\boldsymbol{C}_2$) //预测

3　仿真实验与结果分析

本节通过设计消融实验、对比实验和神经网络参数选择实验,来分析不同方面的工作,评估所提 KBAC 框架的有效性。所提模型均由 Python 软件和 Pytorch 深度学习网络框架进行实现,所有的实验都在 Google Colaboratory 平台上使用 GPU 执行完成。

3.1　数据来源与数据增强

本文使用的数据集来自阿尔伯塔大学的 drugbank 药物分子数据库,从中收集到 1 974 个化合物 SMILES 表达式、描述符以及对应的 ERα 的 IC50 值或者 pIC50 值。IC50(半抑制浓度)的值越小,说明生物的活性越大,对于抑制 ERα 的活性就越有效,而 pIC50 由 IC50 取负对数得到,和 IC50 相反,pIC50 的值越大,说明生物的活性越大,抑制 ERα 的活性也就越有效,本文将 pIC50 作为因变量进行预测。

本文采用 ERα 数据对生物的活性值进行预测,仅有 1 974 个化合物的 SMILES 表达式。如果直接进行采用,会存在过拟合,不利于在数据非线性的情况下进行学习,为此需要对数据的 SMILES 表达式进行增强。对于同一化合物分子,在基于本身标准的 SMILES 基础上,使用 Python 软件中的 RDKit 库进行计算,随机再增加 4 个不同的 SMILES 表达式,将其扩展成 5 个不同的 SMILES 表达式。将数据集按 8∶1∶1 随机划分成训练集、验证集和测试集,部分数据见表 2。

表 2　增强后的 SMILES 表达式数据表

Table 2　Augmented SMILES expression data

SMILES	pIC50	增强 SMILES_1	增强 SMILES _2	增强 SMILES _3	增强 SMILES _4
Oc1ccc2O[C…	8. 602	C1CCC(C1)[…	Oc1cc2c(cc1)…	c1c(OCCN2C…	c1c(cc2S[C@…
CC\C(=C(/c1…	7. 409	c1ccc(/C(=C…	C(/C(c1ccccc…	OC(=O)/C=C/…	c1c(/C(=C(\C…
Oc1ccc(cc1…	8. 367	c1(ccc(C2C3(…	c1c(ccc(c1)C1…	Oc1ccc(cc1)C…	c1(C2C3(Cc4…
CN1CCN(CC…	9	C1N(CCN(C1…	C1CN(c2ccc([…	c12cc(O)ccc2…	C1CN(CCN1C…
CN(CCCc1cc…	6. 62	C1NCCN(c2n…	c1(O)ccc(cc1)…	C(Cc1ccc(cc…	n1c(nc(nc1N…

3.2　实验参数说明

实验中将预训练模型的隐藏单元数设置为 768,注意力头个数设置为 12,最大词元序列长度设置为 201。在后续的下游任务中,将 batch size 设为 32,学习率设置为 3e-5,epoch 最大次数设为 100,除了将最后一个阶段的一维卷积层的通道数和卷积核大小设为 1 以外,1D-ECA-CNN 模块中其他层的通道数全部设为 128,卷积核大小全部设为 20。同时,在训练过程中,还运用到了 early stop strategy,防止模型过拟合以及减少一些计算成本,并将 patience 设为 20。当验证集在多次迭代的情况下都没有变化时,模型就会停止这个 epoch 的迭代,提早结束。此外,还引入 MASK 策略和 Pos_weigh 参数来提高模型的拟合性。

3.3　评价指标

用于评估模型的指标主要有 4 个:R^2、MAE、MSE、RMSE。R^2 反映了模型预测值与真实值之间的线性相关性,描述了模型的拟合程度,可以作为判断模型好坏的依据,最大值为 1。R^2 的绝对值越高并且越接近于 1,预测值与真实值之间的线性相关性越强,模型的拟合性就越好,回归预测得到的效果也就越好。

$$R^2 = 1 - \frac{\sum_{i=1}^{m}(y_i - y_i)^2}{\sum_{i=1}^{m}(\overline{y_i} - y_i)^2} \tag{14}$$

MAE(Mean Absolute Error,其值用 M_{MAE} 表示)是预测值与真实值两者之差取绝对值后的平均值,其值越低并且越接近于 0,说明两者之间的偏差就越小,模型的预测性能也就越好。RMSE(Root Mean Square Error,其值用 M_{RMSE} 表示)表示预测值与真实值之间的偏差程度,是在 MSE(Mean Square Error,其值用 M_{MSE} 表示)的基础上取根号得到的数值,同 MAE 一样,RMSE 和 MSE 的值越低越接近于 0,说明两者之间的偏差就越小,模型的预测性能越好。

$$M_{\mathrm{MAE}} = \frac{1}{m}\sum_{i=1}^{m} \mid y_i - \hat{y}_i \mid \tag{15}$$

$$M_{\mathrm{MSE}} = \frac{1}{m}\sum_{i=1}^{m}(y_i - \hat{y}_i)^2 \tag{16}$$

$$R_{\mathrm{RMSE}} = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(y_i - \hat{y}_i)^2} \tag{17}$$

3.4　实验结果比较

3.4.1　消融实验分析

为说明 KBAC 的有效性,本节将使用 ERα 数据集的训练集、验证集、测试集进行一系列消融实验。为此,基于 1D-ECA-CNN 模块设计如下两个变体模型:KBL(Knowledge-BERT-Linear)和 KBC(Knowledge-BERT-CNN),具体情况如表 3 所示。

表 3 KBAC 及其变体结构信息

Table 3 Structure information of KBAC and its variants

模 型	CNN	1D-ECA	K-BERT
KBL	×	×	✓
KBC	✓	×	✓
KBAC	✓	✓	✓

表 4 是本文模型在消融情况下 KBL、KBC、KBAC 3 个模型的实验结果比较,选择 MAE、RMSE、MSE 作为评价指标,表中黑体数字表示最优值。由表 4 可以看出:移除 1D-ECA 子模块和 CNN 子模块的 KBL 模型,在各个数据集以及评价指标上,性能均差于 KBC 和 KBAC 模型;移除 1D-ECA 子模块的 KBC 模型,在各个数据集以及评价指标上,性能优于 KBL,但都不及 KBAC 模型。主要是因为 1D-ECA 子模块能够学习输入数据中不同通道之间的相关性,使得模型可以更好地理解和捕捉输入数据的特征以及序列中的长距离依赖关系。模型缺少 1D-ECA 模块,则不能通过通道的权重调整使得模型能够更好地聚焦于对任务有用的特征。CNN 子模块可以通过卷积操作捕捉输入数据的局部上下文信息,模型缺少 CNN 子模块,可能无法有效地利用上下文信息来进行 pIC50 的生物活性预测,从而影响模型的性能。

由表 4 可知:本文所提 KBAC 模型在 3 个数据集以及 3 个评价指标上均取得了最优表现,主要是因为模型在 K-BERT 阶段第一次特征提取的基础上,再将 1D-ECA 子模块嵌入到 CNN 子模块,用于分子表征信息的特征再提取。本文模型充分利用了多通道注意力和 CNN 网络的优点,因此能够更好地提取分子 SMILES 表达式的特征,提高模型的 pIC50 生物活性预测精度。

表 4 消融实验对比表

Table 4 Comparison of ablation experiments

模 型	训练集			验证集			测试集		
	M_{MAE}	M_{RMSE}	M_{MSE}	M_{MAE}	M_{RMSE}	M_{MSE}	M_{MAE}	M_{RMSE}	M_{MSE}
KBL	0.247	0.334	0.112	0.481	0.651	0.423	0.550	0.743	0.552
KBC	0.104	0.141	0.020	0.448	0.616	0.380	0.527	0.736	0.542
KBAC	**0.091**	**0.117**	**0.014**	**0.447**	**0.605**	**0.366**	**0.498**	**0.686**	**0.471**

图 4 是本文模型在消融情况下,KBL、KBC 和 KBAC 模型关于药物分子 SMILES 表达式的 pIC50 真实值和预测值对比图。

由图 4 可知:KBAC 模型在整体上明显优于其他变体方法,在训练集、测试集和验证集中,真实值与预测值之间的差异最小;相反,KBL 模型可能因为无法有效利用上下文信息来捕获分子特征,导致其真实值与预测之间的差异最大,且预测精度较其他两个变体模型偏低;没有 1D-ECA 模块的 KBC 模型缺乏调整通道权重的能力,这使得该模型无法聚焦于与任务相关的特征。虽然 KBC 模型的预测精度相比 KBL 模型有所提高,但仍然略逊于 KBAC 模型。意味着 KBAC 框架在 ERα 数据集上具有很大的性能优势,进一步证实所提 KBAC 深度神经网络的合理性。

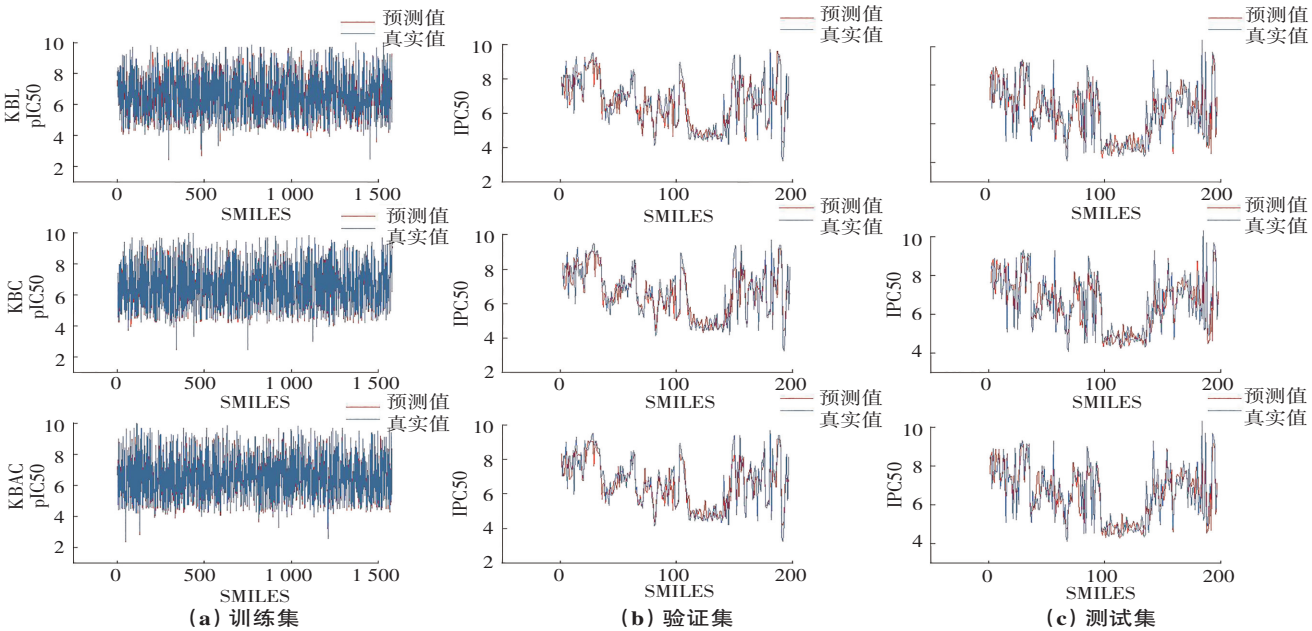

图 4 消融实验下 pIC50 的真实值-预测值对比图

Fig. 4 Comparison of true versus predicted pIC50 ralues under ablation experiment

3.4.2　对比实验分析

为进一步说明所提模型的有效性,选选取当前有代表性的 4 个模型进行比较,即 XGBoost[31]、SVM、FraGAT[32]、Attentive FP[33]。其中 XGBoost、SVM 为基于描述符的方法;FraGAT、Attentive FP 为基于 GNN 的方法。利用 MAE、MSE、RMSE、$R^2$4 个评价指标进行分析,对比结果见表 5,其中黑体数字表示最优值。

表 5　相关模型对比实验表

Table 5　Comparison of metrics of related models

方　法	训练集				测试集			
	M_{MAE}	M_{MSE}	M_{RMSE}	R^2	M_{MAE}	M_{MSE}	M_{RMSE}	R^2
XGBoost[31]	0.231	0.095	0.308	0.952	1.038	1.561	1.249	0.219
SVM	0.520	0.511	0.715	0.750	0.555	0.565	0.752	0.688
FraGAT[32]	0.491	0.408	0.638	0.773	0.502	0.445	0.667	0.747
Attentive FP[33]	0.190	0.069	0.262	0.952	0.512	0.537	0.732	0.700
KBAC	**0.091**	**0.014**	**0.117**	**0.993**	**0.498**	**0.471**	**0.686**	**0.779**

由表 5 可知:所提 KBAC 方法在整体上优于其他比较方法。相比之下,XGBoost[31]的性能最差,而 Attentive FP[33]次之。具体而言,KBAC 在训练集和测试集上的 MAE 值在所有解决方案中均表现最佳,尤其是在训练集上,分别比第二名(Attentive FP[33])低了 0.190 和 0.014。此外,KBAC 在训练集和测试集上获得的 MSE 分别为 0.014 和 0.471,比第二名低了 0.055 和 0.066。而 KBAC 在训练集和测试集上获得的 R^2 分别为 0.993 和 0.779,比第二名高了 0.041 和 0.079,说明本文所提方法在 pIC50 生物活性值的预测中表现出很好的性能优势。主要是因为 KBAC 模型既考虑了结合描述符、图和 SMILES 方法的 K-BERT 子模块,又考虑了结合多通道注意力机制的 CNN 网络结构,使得模型能够更好地理解分子 SMILES 表达式,并融合多通道特征,提取分子表征的全局和局部特征。

3.4.3　神经网络中卷积核和通道数选择比较分析

网络中需要人工干预的超参数包括网络通道数、卷积核大小等,这些超参数的设置直接影响了模型的性能和效果,因此需要进行精确的调整以获取最佳参数组合。本部分的实验将基于 ERα 数据集进行,主要探究卷积神经网络的通道数尺寸以及卷积核尺寸大小对于所提模型生物活性值预测效果和性能的影响。本文通过固定卷积核大小探索网络通道数对模型预测效果的影响,以及固定网络通道数探索卷积核大小对模型预测效果的影响,分析卷积核的最优参数选择和最优通道数选择。本部分将网络的通道数分别设为 16、32、64、128,卷积核大小分别设为 10、20、30、40 进行网络参数对比实验。以 MAE、MSE 和 RMSE 为评价指标展示不同尺寸的通道数和卷积核情况下的预测精度和效果。具体情况如表 6 所示。

表 6　神经网络参数对比实验表

Table 6　Comparison of metrics of neural network parameters

卷积核	通道数 16			通道数 32			通道数 64			通道数 128		
	M_{MAE}	M_{MSE}	M_{RMSE}	M_{MAE}	M_{MSE}	M_{RMSE}	M_{MAE}	M_{MSE}	M_{RMSE}	M_{MAE}	M_{MSE}	M_{RMSE}
10	0.100	0.016	0.125	0.111	0.024	0.154	0.150	0.042	0.200	0.089	0.015	0.121
20	0.093	0.017	0.129	0.129	0.029	0.171	0.150	0.040	0.199	0.091	0.014	0.117
30	0.142	0.036	0.189	0.130	0.030	0.180	0.162	0.042	0.205	0.094	0.015	0.124
40	0.127	0.030	0.173	0.136	0.035	0.187	0.132	0.030	0.174	0.114	0.023	0.153

表 6 是所提 KBAC 模型在 MAE、MSE 和 RMSE 指标下卷积核大小与网络通道数变换对模型预测效果的影响。由表 6 可以看出:在固定卷积核相同的情况下,随着通道数的增加,RMSE、MAE 和 MSE 值均呈现先增加后减少的状态,在通道数为 128 时,KBAC 模型达到最佳性能,说明通过选取适当的卷积通道数大小,能够增强回归模型的学习能力,优化预测结果;在固定网络通道数的情况下,随着卷积核大小逐步提升,RMSE、MAE 和 MSE 值基本上呈现增长趋势,在卷积核大小为 20 时,KBAC 模型达到最佳性能,说明通过选取适当的

卷积核大小,模型能够有效获取邻域特征信息,从而提高模型的判别能力和预测精度。为确保网络具有足够的复杂性和表达能力,本文实验采用的网络通道数为128,卷积核大小为 20。

3.4.4 网络速度收敛分析

图 5 展示了卷积核大小与网络通道数变换对模型收敛速度(Epoch)的影响。由图 5 可以看出:随着通道数和卷积核大小的逐渐增加,模型收敛的速度越来越快,所需要消耗的计算资源也随之减少。

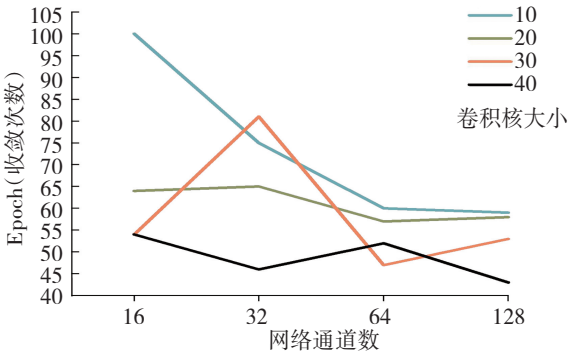

图 5 网络速度收敛分析实验图

Fig. 5 Experimental results on the convergence behavior of network speed

因此,综合所有情况下的结果表明:当通道数为128,卷积核大小为 20 时,所提 KBAC 深度网络可以达到最优训练效果和预测精度,并有较好的收敛速度。

4 结 论

为解决乳腺癌治疗候选药物的筛选问题,本文改进了多通道注意力机制,提出基于知识先验和多通道注意力神经网络相结合的 KBAC 框架,用于药物分子的 pIC50 生物活性值预测。由于模型既考虑了基于描述符、图和 SMILES 方法的 K-BERT 知识先验框架,又构造了基于多通道注意力机制的深度网络,实现了从整体到全局-局部的两次特征提取以及多通道特征融合,使得模型具有泛化性强、特征信息提取全面等优点,能够更好地预测药物分子生物活性值。比较其他具有代表性的方法,所提 KBAC 框架在 MAE、RMSE、MSE 和 R^2 上均具有较高的评价指标,说明了模型的有效性。

深度学习的“黑匣子”性质和生物活性的不可知性,使得基于深度学习的生物活性值预测还有很多问题需要研究。下一步工作将继续探索更多基于知识先验和注意力机制的深度学习模型,并结合多种预测方法的优点,进行多模态特征信息提取,以进一步提高分子预测任务的精度和解释性,为药物研发提供更多的

支持与帮助。

参考文献(References):

[1] WANG T, SUN J, ZHAO Q. Investigating cardiotoxicity related with hERG channel blockers using molecular fingerprints and graph attention mechanism[J]. Computers in Biology and Medicine, 2023, 153(1): 106464-106470.

[2] SHERIDAN R P, WANG W M, LIAW A, et al. Extreme gradient boosting as a method for quantitative structure-activity relationships[J]. Journal of Chemical Information and Modeling, 2016, 56(12): 2353-2360.

[3] GERTRUDES J C, MALTAROLLO V G, SILVA R A, et al. Machine learning techniques and drug design[J]. Current Medicinal Chemistry, 2012, 19(25): 4289-4297.

[4] ZHONG C, AI J, YANG Y, et al. Small molecular drug screening based on clinical therapeutic effect[J]. Molecules, 2022, 27(15): 4807-4827.

[5] ZHU Z, YAO Z, ZHENG X, et al. Drug-target affinity prediction method based on multi-scale information interaction and graph optimization[J]. Computers in Biology and Medicine, 2023, 167(1): 107621-107633.

[6] YANG K, SWANSON K, JIN W, et al. Analyzing learned molecular representations for property prediction[J]. Journal of Chemical Information and Modeling, 2019, 59(8): 3370-3388.

[7] KOROLEV V, MITROFANOV A, KOROTCOV A, et al. Graph convolutional neural networks as “general-purpose” property predictors: the universality and limits of applicability[J]. Journal of Chemical Information and Modeling, 2020, 60(1): 22-28.

[8] RATHI P C, LUDLOW R F, VERDONK M L. Practical high-quality electrostatic potential surfaces for drug discovery using a graph-convolutional deep neural network[J]. Journal of Medicinal Chemistry, 2020, 63(16): 8778-8790.

[9] WANG Y, WANG J, CAO Z, et al. Molecular contrastive learning of representations via graph neural networks[J]. Nature Machine Intelligence, 2022, 4(3): 279-287.

[10] YANG Z, ZHONG W, ZHAO L, et al. MGraphDTA: Deep multiscale graph neural network for explainable drug-target binding affinity prediction[J]. Chemical Science, 2022, 13(3): 816-833.

[11] ZHANG X, WU C, YI J, et al. Pushing the boundaries of molecular property prediction for drug discovery with multitask learning BERT enhanced by SMILES enumeration[J]. Research, 2022, 2022(1): 0004-0017.

[12] WANG S, GUO Y, WANG Y, et al. SMILES-BERT: Large scale unsupervised pre-training for molecular property

prediction[C]//Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics. New York: ACM, 2019: 429-436.

[13] YUAN W, CHEN G, CHEN C Y C. FusionDTA: Attention-based feature polymerizer and knowledge distillation for drug-target binding affinity prediction[J]. Briefings in Bioinformatics, 2022, 23(1): 506-518.

[14] HUGO G, SINAGA V L, DINATA I M, et al. Graph attention network on extracting feature from simplified molecular-input line-entry system for HIV classification[C]//Proceedings of the International Conference on Electrical and Information Technology. Piscataway: IEEE Press, 2022: 398-401.

[15] AHMADI S, MORADI Z, KUMAR A, et al. SMILES-based QSAR and molecular docking study of xanthone derivatives as α-glucosidase inhibitors[J]. Journal of Receptor and Signal Transduction Research, 2022, 42(4): 361-372.

[16] TENG S, YIN C, WANG Y, et al. MolFPG: Multi-level fingerprint-based graph transformer for accurate and robust drug toxicity prediction[J]. Computers in Biology and Medicine, 2023, 164(1): 106904-106912.

[17] LI P, WANG J, QIAO Y, et al. An effective self-supervised framework for learning expressive molecular global representations to drug discovery[J]. Briefings in Bioinformatics, 2021, 22(6): 109-122.

[18] ROSS J, BELGODERE B, CHENTHAMARAKSHAN V, et al. Large-scale chemical language representations capture molecular structure and properties[J]. Nature Machine Intelligence, 2022, 4(12): 1256-1264.

[19] WU Z, JIANG D, WANG J, et al. Knowledge-based BERT: A method to extract molecular features like computational chemists[J]. Briefings in Bioinformatics, 2022, 23(3): 131-143.

[20] HUNT P, HOSSEINI-GERAMI L, CHRIEN T, et al. Predicting pKa using a combination of semi-empirical quantum mechanics and radial basis function methods[J]. Journal of Chemical Information and Modeling, 2020, 60(6): 2989-2997.

[21] PAPA E, PILUTTI P, GRAMATICA P. Prediction of PAH mutagenicity in human cells by QSAR classification[J]. SAR and QSAR in Environmental Research, 2008, 19(1-2): 115-127.

[22] KUMARI C, ABULAISH M, SUBBARAO N. Exploring molecular descriptors and fingerprints to predict mTOR kinase inhibitors using machine learning techniques[J]. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2021, 18(5): 1902-1913.

[23] YU Z, GAO H. Molecular representation learning via heterogeneous motif graph neural networks[C]//International Conference on Machine Learning. PMLR, 2022: 25581-25594.

[24] XIA X, ZHU C, ZHONG F, et al. MDTips: A multimodal-data-based drug-target interaction prediction system fusing knowledge, gene expression profile, and structural data[J]. Bioinformatics, 2023, 39(7): 411-419.

[25] LI Z, ZHU S, SHAO B, et al. DSN-DDI: An accurate and generalized framework for drug-drug interaction prediction by dual-view representation learning[J]. Briefings in Bioinformatics, 2023, 24(1): 597-608.

[26] WEININGER D, WEININGER A, WEININGER J L. SMILES. 2. Algorithm for generation of unique SMILES notation[J]. Journal of Chemical Information and Computer Sciences, 1989, 29(2): 97-101.

[27] HUA Y, SONG X, FENG Z, et al. MFR-DTA: A multi-functional and robust model for predicting drug-target binding affinity and region[J]. Bioinformatics, 2023, 39(2): 56-64.

[28] SHAO J, GONG Q, YIN Z, et al. S2DV: Converting SMILES to a drug vector for predicting the activity of anti-HBV small molecules[J]. Briefings in Bioinformatics, 2022, 23(2): 593-605.

[29] ZHAO Q, DUAN G, YANG M, et al. AttentionDTA: Drug-target binding affinity prediction by sequence-based deep learning with attention mechanism[J]. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2022, 20(2): 852-863.

[30] WANG Q, WU B, ZHU P, et al. ECA-net: Efficient channel attention for deepconvolutional neural networks[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2020: 11531-11539.

[31] CHEN T, GUESTRIN C. XGBoost: A scalable tree boosting system[C]//Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2016: 785-794.

[32] ZHANG Z, GUAN J, ZHOU S. FraGAT: A fragment-oriented multi-scale graph attention model for molecular property prediction[J]. Bioinformatics, 2021, 37(18): 2981-2987.

[33] XIONG Z, WANG D, LIU X, et al. Pushing the boundaries of molecular representation for drug discovery with the graph attention mechanism[J]. Journal of Medicinal Chemistry, 2020, 63(16): 8749-8760.

责任编辑:李翠薇