单位代码 11799
学　　号 2022310095

重庆工商大学

硕士学位论文

多模态对比学习深度网络的抗乳腺癌药物分子性质预测

论文作者：唐文燕
所在学院：数学与统计学院
学科专业：统计学
研究方向：大数据分析
指导教师：李梦

提交论文日期：　2025 年　5 月　16　日
论文答辩日期：　2025 年　5 月　16　日

中国•重庆
2025 年 5 月

多模态对比学习深度网络的抗乳腺癌药物分子性质预测

摘 要

乳腺癌的高发病率和高死亡率对女性健康构成了严重威胁，因此，抗乳腺癌药物的研发和筛选已成为医学研究的重点方向。分子性质是评估药物化合物成药性的关键指标，通过研究分子的生物活性以及药代动力学性质和安全性，可以揭示潜在的生物医学机制和特定情况下的物理、化学或生物性质。传统的药物研发工作，由于人力、时间、成本以及实验条件等因素的限制，往往存在周期长、成本高、副作用大等问题。近年来，人工智能技术的迅猛发展极大推动了计算机辅助药物设计领域的进步，通过深度学习算法预测药物分子的生物活性以及ADMET性质（吸收Absorption、分布Distribution、代谢Metabolism、排泄Excretion和毒性Toxicity），能够显著提高药物筛选和研发的效率和成功率，并降低其研发成本。然而，现有方法通常只考虑单模态下药物分子性质的简单结构分析，另有一些文献采用了多模态融合方法学习分子特征，但未能充分考虑多模态数据间固有的异质性、复杂性及其相互关系，从而限制了模型的预测精度和泛化能力。为解决这些问题，本文提出多通道语义深度神经网络和多模态对比学习深度网络，深入研究抗乳腺癌候选药物分子特征的准确提取和有效融合关键技术，以实现高精度的分子性质预测。本文的主要研究内容如下所述：

（1）针对单模态药物分子性质预测所面临的分子生物活性值预测不精、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模型（Knowledge-BERT-1D-ECA-CNN，KBAC），使用分子的单模态SMILES（Simplified Molecular Input Line Entry System）表征预测抗乳腺癌候选药物化合物的pIC50生物活性值，降低药物研发的成本和周期。

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 Knowledge-BERT-1D-ECA-CNN（KBAC）深度神经网络，实现 pIC50 生物活性值回归预测。

（2）针对多模态药物分子性质预测所存在的多模态数据表征不一致、互补信息挖掘不足等问题，提出基于对比学习和自适应权重分配机制的多模态自适应对比融合深度网络（Adaptively multi-modal contrastive fusion network，AMCFNet），使用分子的 SMILES 表征和图结构（Graph）信息预测抗乳腺癌候选药物化合物的 ADMET 性质，提升药物筛选的准确性和效率。

该网络首先构建双分支特征提取模块，设计 K-BERT 网络和多层图神经网络（Graph Neural Network，GNN）分别学习药物分子的一维 SMILES 序列信息（1D-SMILES 特征）和二维图结构信息（2D-graph 特征）；其次，基于对比学习理论，设计自适应对比融合模块，通过共识评分机制动态评估模态内与模态间的一致性，缓解多模态数据间的认知差异问题，同时根据对比学习结果自适应地分配多模态特征权重，从而实现药物分子语义特征与结构信息的深度融合（SMILES-graph 特征）；最后，通过整合 1D-SMILES、2D-graph 以及融合的 SMILES-graph 特征，构建多视角的分子互补特征，用于分子 ADMET 性质预测。

实验结果表明，所提出的 KBAC 和 AMCFNet 框架在相应的评估指标上均表现出显著优势，相比于其他具有代表性的模型有较为明显的提升，说明了所提模型的有效性，能够获取更为全面的分子特征，从而帮助筛选治疗乳腺癌的候选药物，并加速药物研发的进程。

关键词：分子性质预测；知识先验；注意力机制；多模态融合；对比学习

A MULTIMODAL CONTRASTIVE LEARNING DEEP NETWORK FOR PREDICTING DRUG MOLECULAR PROPERTIES IN ANTI-BREAST CANCER

ABSTRACT

The high incidence and mortality rates of breast cancer pose a significant threat to women's health, making the development and screening of anti-breast cancer drugs a critical focus in medical research. Molecular properties serve as crucial indicators for evaluating the drug-likeness of compounds. Investigating molecular bioactivity, pharmacokinetic properties, and safety profiles can reveal potential biomedical mechanisms along with physical, chemical, and biological characteristics under specific conditions. Traditional drug discovery approaches, constrained by human resources, time, costs, and experimental conditions, often suffer from prolonged cycles, high costs, and significant side effects.

Recent advancements in artificial intelligence have substantially propelled progress in computer-aided drug design (CADD). Deep learning algorithms enable the prediction of molecular bioactivity and ADMET properties (Absorption, Distribution, Metabolism, Excretion, and Toxicity), significantly enhancing the efficiency and success rate of drug screening and development while reducing associated costs. However, existing methodologies predominantly focus on simplistic structural analyses of molecular properties under single-modality frameworks. Although some studies employ multimodal fusion approaches to learn molecular features, they inadequately address the inherent heterogeneity, complexity, and interrelationships among multimodal data, thereby limiting model prediction accuracy and generalization capabilities.

To address these challenges, this study proposes a multi-channel semantic deep neural network and a multimodal contrastive learning deep network. These frameworks aim to thoroughly investigate key technologies for accurate extraction and effective

fusion of molecular features in anti-breast cancer drug candidates, ultimately achieving high-precision molecular property prediction. The primary research contributions of this work are outlined as follows:

(1) To address the challenges of imprecise prediction of molecular bioactivity values and limited generalization capability in single-modality molecular property prediction, we propose a Knowledge-BERT-1D-ECA-CNN (KBAC) model that integrates knowledge priors with attention mechanisms. This multi-channel semantic deep neural network employs single-modal SMILES (Simplified Molecular Input Line Entry System) representations to predict the pIC50 bioactivity values of anti-breast cancer drug candidates, thereby reducing drug development costs and timelines.

The network adopts a two-stage feature extraction strategy: to the semantic layer, a semantic analysis network combining knowledge priors with transfer learning is designed to localize critical information from molecular SMILES, descriptors, and graph representations. By fine-tuning parameters on the Erα dataset, it generates comprehensive molecular SMILES representations. To the channel layer, a 1D-ECA algorithm based on the Efficient Channel Attention (ECA) mechanism is developed and embedded into a CNN submodule, forming a multi-channel deep neural 1D-ECA-CNN module. This enhances feature re-extraction of molecular representations while minimizing information loss during molecular representation learning. The integration of these two layers forms the Knowledge-BERT-1D-ECA -CNN (KBAC) framework, enabling regression prediction of pIC50 bioactivity values.

(2) To resolve inconsistencies in multimodal data representation and insufficient exploitation of complementary information in multimodal molecular property prediction, we propose an Adaptively Multi-modal Contrastive Fusion Network (AMCFNet). This framework leverages contrastive learning and adaptive weight allocation mechanisms to predict ADMET properties of anti-breast cancer drug candidates using molecular SMILES representations and graph structural information, thereby improving drug screening accuracy and efficiency.

The network consists of three modules: the K-BERT network and multi-layer GNN are constructed to separately learn 1D-SMILES sequence features and 2D-graph structural features of drug molecules. And the contrastive fusion module is designed based on contrastive learning theory. It dynamically evaluates intra- and inter-modal consistency through a consensus scoring mechanism, mitigating cognitive disparities between multimodal data. Simultaneously, it adaptively allocates multimodal feature weights based on contrastive learning outcomes, achieving deep fusion of molecular semantic and structural features (SMILES-graph features). Furthermore, the framework integrates 1D-SMILES, 2D-graph, and fused SMILES-graph features to construct multi-view and complementary molecular representations for ADMET property prediction.

Experimental results demonstrate that the proposed KBAC and AMCFNet frameworks exhibit significant advantages over baselines across relevant evaluation metrics. The performance improvements validate the effectiveness of these models in capturing comprehensive molecular features, thereby facilitating the screening of breast cancer therapeutic candidates and accelerating drug development processes.

Keywords：Molecular property prediction; Prior knowledge; Attention mechanism; Multi-modal fusion; Contrastive learning

缩 略 词 表

英文缩写	英文名	中文名
ERα	Estrogen Receptor Alpha	雌激素受体 α
ADMET	Absorption, Distribution, Metabolism, Excretion, and Toxicity	药代动力学性质和安全性
SMILES	Simplified Molecular Input Line Entry System	简化分子输入线性条目系统
ML	Machine Learning	机器学习
DL	Deep Learning	深度学习
SVM	Support Vector Machine	支持向量机
RF	Random Forest	随机森林
CNN	Convolutional Neural Networks	卷积神经网络
RNN	Recurrent Neural Network	循环神经网络
GNN	Graph Neural Network	图神经网络
GCN	Graph Convolutional Network	图卷积网络
GAT	Graph Attention Network	图注意力网络
GAP	Global Average Pooling	全局平均池化
GMP	Global Max Pooling	全局最大池化
BERT	Bidirectional Encoder Representations from Transformers	基于 Transformer 的双向编码器表示
MLM	Masked Language Model	掩码语言模型
NSP	Next Sentence Prediction	下一句预测
NLP	Natural Language Processing	自然语言处理
ECA	Efficient Channel Attention	高效通道注意力
MHA	Multi-head Attention	多头注意力
K-BERT	Knowledge-BERT	基于知识先验的 BERT 模型
AP	Average Pooling	平均池化
MAE	Mean Absolute Error	平均绝对误差
MSE	Mean Square Error	均方误差
RMSE	Root Mean Square Error	均方根误差
R^2	R-squared	决定系数
ROC-AUC	Receiver Operating Characteristic - Area Under the Curve	接收者操作特征曲线下面积
ACC	Accuracy	准确率
FC	Fully Connected	全连接层
LN	Layer Normalization	层归一化
ReLU	Rectified Linear Unit	修正线性单元
FFN	Feedforward Network	前馈神经网络
1D	One-Dimensional	一维
2D	Two-Dimensional	二维

目 录

插 图 索 引

表 格 索 引

第1章 绪论

1.1 研究背景及意义

乳腺癌是一种源自乳腺组织的恶性肿瘤，因其恶性细胞异常增生而形成具有侵袭性的肿瘤组织[1-2]。据世界卫生组织和美国癌症协会统计数据显示，2022 年乳腺癌的发病率和死亡率分别占全球恶性肿瘤新发病例和死亡总数的 11.6%和6.9%，在全球恶性肿瘤发病率和死亡率排名中均位居前列[3]。因此，构建涵盖精准诊疗和新型药物研发的完整防治体系，已成为提升乳腺癌患者生存率和治愈质量的关键路径[4]。

在抗乳腺癌药物研发领域，深入分析药物分子的理化性质及其与疾病的内在联系，对发现潜在治疗药物具有决定性意义[5-8]。其中，生物活性与 ADMET 性质（吸收 Absorption、分布 Distribution、代谢 Metabolism、排泄 Excretion 和毒性 Toxicity）[9-10]是评估候选化合物成药性和有效性的关键指标。生物活性值反映了化合物对靶标分子的效力和选择性，能够直接揭示药物的潜在药理机制。ADMET 性质（即药代动力学性质和安全性）则反映了化合物在体内的安全性和有效性，能够揭示药物在体内的作用机制。通过综合分析候选药物的分子性质，研究人员能够筛选出兼具高活性和理想药代动力学特征的候选分子，从而确保药物的疗效和安全性。

传统的药物研发和筛选过程严重依赖于化学生物实验和人工特征提取，需要研发人员具备大量的专业知识和经验，往往存在着周期长、成本高、副作用大、成功率低等问题，同时还面临伦理和法律层面的诸多限制。并因其分子性质的不可预见性，可能导致出现检测结果不精或失败的情况，从而造成研发成本的增加。这些局限性使得传统的药物研发模式难以满足个性化抗乳腺癌药物治疗需求。因此，具备高效自动化能力的机器学习（Machine Learning，ML）和深度学习（Deep Learning，DL）等计算机辅助方法[11-14]，在药物研发中备受青睐。

深度学习算法凭借其强大的高维数据处理、复杂特征提取和非线性建模能力，能够高效捕捉药物分子在一维序列信息、二维拓扑结构以及三维空间构象中的复杂非线性特征，从而减少对人工设计特征的依赖，使其在药物研发和筛选领域中具有显著优势。当前基于深度学习的药物分子性质预测算法主要分为两种：单模

态分子性质预测方法[15-17]和多模态分子性质预测方法[18-19]。单模态方法主要基于药物分子的描述符、指纹、简单化学分子表达式或图结构数据等输入类型，构建相应深度学习算法，如：卷积神经网络（Convolutional Neural Network, CNN）、循环神经网络（Recurrent Neural Network, RNN）[20]、以及图神经网络（Graph Neural Network, GNN）[21]等模型，学习药物分子单一表征信息与其理化性质、生物活性之间的潜在关联，从而实现高精度的分子性质预测。但分子性质通常涉及多种复杂因素的相互作用，单模态方法无法全面反映其内在机制，导致模型的泛化性能不佳；同时，现有方法缺乏对分子重要特征的特别关注，导致其对药物分子性质的理解不足，从而在跨靶标预测任务中表现欠佳。多模态方法则通过简单拼接、加权融合、交叉注意力机制等融合技术，整合分子的多种表征数据，并学习模态间的语义以及结构互补信息，使得模型能够综合理解药物化合物的特征信息，从而提升预测模型的准确性和鲁棒性，加速药物研发的进程。但在药物分子性质预测领域，现有的多模态融合技术大多仅采用简单的特征拼接或加权平均策略，使得模型难以捕捉不同分子模态间的深层关联和相互作用机制；且现有方法往往忽略了分子模态间的异质性，以及不同模态的贡献度程度，导致特征融合的效果不佳，从而影响模型的预测精度和鲁棒性。

因此，构建基于深度学习的抗乳腺癌药物分子性质预测模型，对于提升药物研发效率和精准度具有重要的研究价值和应用前景。本研究围绕药物分子的单模态和多模态表征策略，提出了两种创新的基于深度学习的预测方法，并在多个公开分子数据集上进行了系统性验证。实验结果表明，本研究所提方法在生物活性与 ADMET 性质预测任务中均展现出优越性能。通过对小分子筛选数据的深度挖掘，本研究能够有效识别关于乳腺癌治疗的潜在候选药物，为降低新药研发成本、提升筛选效率提供了有效解决方案，以应对药物设计领域对化合物高效筛选和分子性质精准预测的迫切需求。

1.2 国内外研究现状

1.2.1 单模态分子性质预测

（1）基于描述符的分子性质预测

分子描述符是将分子的结构特征转化为数值形式的表示，通常采用定量构效关系（Quantitative Structure-Activity Relationship, QSAR）[22-24]和机器学习方法[25-26]来建立分子结构与其分子性质之间的数学关系。基于描述符的方法凭借其计算高

效性、结果可解释性和广泛适用性，而被广泛应用于分子性质研究领域[27-28]。如 Hunt P 等[29]通过结合半经验量子力学描述符和机器学习方法，用于捕获分子、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 E 等[30]在基于遗传算法选择的不同分子描述符上，使用 QSAR 分类方法建模，从而预测人类细胞的 PHA 致突变性。Shi L H 等[31]引入集成学习算法，通过分子描述符数据分析分子关键特征，进而预测分子的 ADMET 性质。何冰等[32]基于分子描述符特征筛选和机器学习算法构建分类模型，以筛选具有高亲和力的乳腺癌新型抑制剂。付洺宇等[33]基于分子描述符的重要性特征分析和 GA-XGBoost 算法，构建化合物血脑屏障通透性预测模型。

然而，描述符往往是对分子信息的高度抽象，可能丢失重要的结构信息，导致模型对分子之间细微差异的敏感度降低，且描述符的选择通常依赖专家经验，难以全面捕捉分子的复杂性。传统的基于描述符的方法主要依赖于预先定义的数学模型来捕获分子结构与性质之间的关系，使其表达能力有限，难以处理分子性质之间复杂的非线性关系。

（2）基于指纹的分子性质预测

分子指纹是一种将分子结构转化为二进制格式或数值向量的表示方法，每个指纹位代表特定的结构特征，如化学键、官能团或分子中的其他重要子结构。常结合支持向量机（Support Vector Machine，SVM）[34]、随机森林（Random Forest，RF）[35]和小型神经网络（Neural Networks，NN）等机器学习算法[36]进行研究分析。如：Kumari C 等[37]利用分子描述符和指纹来构建哺乳动物的雷帕霉素靶激酶抑制剂的预测模型。Teng S S 等[38]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。Ding W Z 等[39]将分子指纹的 MDFP、Baseline2D、ECFP4 和 PropertyFP 与机器学习模型相结合，以预测化合物的 hERG 心脏毒性。卢昂等[40]基于黄酮类化合物的子结构指纹，结合卡方检验筛选出与抗氧化活性显著相关的分子指纹，并建立了基于判别分析的 DPPH 自由基清除活性预测模型。于亚运等[41]基于分子子结构的药物和靶点指纹特征和指纹相似度，结合随机森林算法构建药物-靶点相互作用预测模型。

然而，分子指纹作为固定长度的二进制向量表示，可能会丢失一些详细的结构信息。这些基于描述符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，但分子指纹的生成需要大量的专家知识指导，且主要关注分子的局部结构特征，难以捕捉分子之间的全局信息或动态行为，从而导致预

测结果不佳。

（3）基于 SMILES 的分子性质预测

SMILES（Simplified Molecular Input Line Entry System）表达式[42]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。由于其简洁的文本表示，基于自然语言处理的模型在分子性质预测中得到了广泛应用[43-44]，如基于 SMILES 的 Transformer[45-46]模型和基于 SMILES 的 BERT[47-48]模型。Chen Y L 等[49]采用金字塔网络结构和分子 SMILES 表征，融合从多尺度层提取的特征，以预测药物-靶点结合亲和力。Ross J 等[50]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 Y 等[51]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 J S 等[52]提出通过将 SMILES 转换为药物向量来预测抗 HBV 活性的网络。Zhao Q C 等[53]提出基于序列的模型，通过注意机制来预测药物与靶标的亲和性。

SMILES 表征由于其简单性和通用性在分子建模中备受青睐。但由于其不具备分子描述符、指纹以及分子图那样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。

（4）基于 Graph 的分子性质预测

基于图的方法[54-56]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。图神经网络、图卷积网络（Graph Convolutional Networks，GCNs）[57]、图注意力网络（Graph Attention Networks，GAT）[58]和图同质网络（Graph Isomorphism Networks，GIN）[59]是最经典的图结构特征提取网络方法。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 P Y 等[60]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 Z N 等[61]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 X Q 等[62]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 Z M 等[63]提出的双视图框架，通过迭代地使用局部和全局表示学习模块来预测药物间相互作用。

Chen Z D 等[64]将分子图数据划分为多个基于功能组的簇，并设计了分割 MPNN 来学习功能组，以发现结构-性质关系。

基于图的方法可以学习分子结构的层次表示，并有效地建模邻域信息。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。

1.2.2 多模态药物分子性质预测

随着新技术在药物发现领域的不断进步，越来越多的多模态方法[65-66]被引入，以便更全面和深入地理解分子特性和结构的机制。这些多模态方法可以结合来自不同来源的数据，如化学结构信息、分子生物学数据、实验室结果以及临床信息，通过多角度、多维度的分析，揭示了分子行为和生物活性的内在规律。相比传统的单一模态分析，多模态方法能够更有效地捕捉复杂的分子特征，尤其是在处理大型、高维数据集时，能够整合不同类型的信息，从多个角度出发，全面评估候选化合物的生物学活性、药代动力学性质、毒性及其他重要特征，从而优化药物设计过程，缩短研发周期，提高成功率。如，Liu S C 等[67]采用 2D 拓扑结构和 3D 几何视角，通过自监督学习捕捉分子数据中的高级特征。Zhang H H 等[68]定义 PremuNet-L 和 PremuNet-H 两个分支，分别用于提取 1D、2D 和 3D 分子表征的低维单模态特征和高维的几何拓扑融合分子特征，进而预测分子性质。Wu T Y 等[69]使用分子 SMILES 表示和图结构进行分子联合表示学习，通过在 Transformer 中引入键级图表示作为注意力偏差来改进自注意力机制，以加强多模态信息之间的特征对应关系。Lu X H 等[70]利用 Transformer-Encoder、双向门控循环单元和 GCN 构建三模态学习模型，以处理来自化学语言和分子图的三种模态信息，并通过五种机器学习和数值组合方法融合多模态特征。Wang Z Y 等[71]提出 MoleculeKit 框架，结合传统核方法与深度学习下的图结构网络和序列语义网络，学习分子图和序列特征，进行多任务分子性质预测。

尽管这些方法能够有效利用各种技术从不同模态的数据中提取分子特征。然而，在实际场景中，仍然存在一些问题和挑战。1）多模态融合挑战。大多数现有模型只考虑了不同模态之间的简单加权融合，忽略了多模态分子表征中的不一致性、复杂性和内在的关联性，对于探索跨模态的互补信息和融合不够充分。2）预测精度挑战。当前方法对分子性质的深度分析仍然不足，未能充分利用和结合各种单模态方法的优势，忽略了分子重要特征的特别关注，使得模型的泛化能力和

学习能力可能受到限制，导致性能下降。

1.3 研究内容

本文针对抗乳腺癌药物分子的生物活性及其 ADMET 性质预测问题提出了两种不同的算法，分别从分子的单模态和多模态表征角度进行深入探讨，旨在利用深度学习神经网络、注意力机制、统计学知识和对比学习等技术，构建高效、精准的分子性质预测模型。首先，解决医学单模态数据中的高质量特征提取问题是关键，如何准确捕捉单一表征的重要信息，是提高预测精度的基础。同时，多模态数据间的异质性问题也需得到有效处理，如何有效融合这些多源异构数据，进行统一建模，是提升预测效果的另一大挑战。在此基础上，本文进一步从单模态表征的角度展开探索，采用多层次、多通道的建模策略，充分学习与下游任务相关的分子特征，对药物分子生物活性进行精准预测。并以药物分子多模态表征数据建模，从多角度、多空间对分子特征进行深度挖掘，通过深度学习模型的训练和优化，高效提取不同模态间的关联性，从而显著提高药物研发过程中的分子性质预测精度。本文所提出的算法不仅能够加速药物分子筛选的过程，还能有效节省研发时间和资金成本，从而推动药物研发效率的提升，为乳腺癌的精准治疗提供强有力的支持。具体如图 1.1 所示。

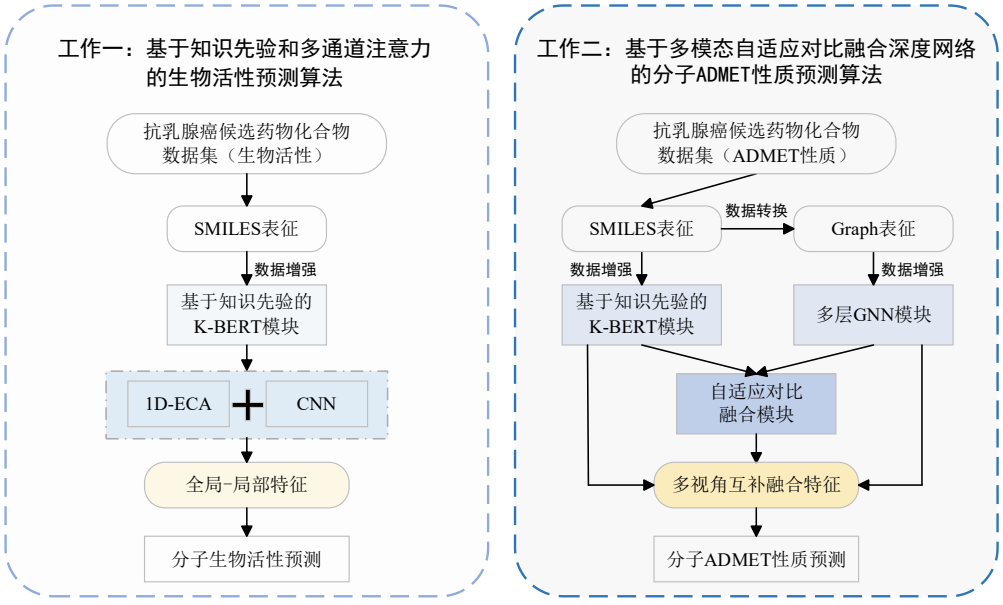


图 1.1 本文主要研究内容示意图

本文主要内容如下：

（1）基于知识先验和多通道注意力的生物活性预测算法

设计基于知识先验和多通道注意力的生物活性预测算法（Knowledge-BERT-1D-ECA-CNN，KBAC），该网络采用两阶段特征提取策略，针对语义层，设计将知识先验与迁移学习结合的语义分析网络，通过在 ERα 数据集中微调参数，得到综合的分子 SMILES 表征信息。针对通道层，基于高效通道注意力（Efficient Channel Attention, ECA）机制，设计 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 Knowledge-BERT-1D- ECA-CNN 深度神经网络，实现抗乳腺癌药物分子的 pIC50 生物活性值回归预测。

（2）基于多模态下自适应对比融合的抗乳腺癌药物分子性质预测网络

设计基于多模态自适应对比融合深度网络的抗乳腺癌药物分子性质预测网络（Adaptively multi-modal contrastive fusion network，AMCFNet），该方法首先构建双分支特征提取模块，引入 K-BERT 网络和多层 GNN 网络分别学习药物分子的 1D 语义特征和 2D 图结构信息；其次基于共识分数设计自适应对比融合模块，对相同分子的不同表征和不同分子之间的同一表征进行对比学习，避免多模态数据中异质性所导致的语义差异，同时，根据对比学习结果自适应地分配权重，融合语义和结构信息，得到多模态互补和交互信息（1D-2D 特征）；最后，通过整合 1D-SMILES、2D-graph 以及融合的 1D-2D 特征，构建多视角互补特征，进行抗乳腺癌药物分子的 ADMET 性质预测。

1.4 论文创新点与组织安排

1.4.1 论文创新点

紧扣国家正面临日益突出的医学与人工智能应用问题，紧密关注国内外药物研发发展动态，利用深度学习和统计学知识解决实际工作，主要创新之处在：

（1）基于知识先验和多通道注意力的生物活性预测算法

本文基于知识先验和多通道注意力，提出的两阶段特征（语义层和通道层）提取 KBAC 深度学习框架，用于预测 ERα 分子的 pIC50 生物活性值。首先，引入K-BERT模块，利用已有领域的知识经验，对ERα数据集进行迁移学习和分析，理解药物分子性质，并提取综合特征；其次，基于多通道注意力，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，并获取分子的局部特征，

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。该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，提高了模型对 SMILES 表征的特征识别和预测精度。

（2）基于多模态自适应对比融合深度网络的抗乳腺癌药物分子性质预测算法

本文基于知识先验、多头注意力机制和对比学习策略，设计用于抗乳腺癌药物分子 ADMET 性质预测的多模态自适应对比融合深度网络（AMCFNet）。首先，针对不同模态特点，设计 K-BERT 和多层 GNN 模块，多角度提取 1D-SMILES 和 2D-graph 分子特征，实现对分子特征的语义信息和结构性质的全面捕捉；其次，基于共识分数构建自适应对比学习融合模块，在多模态数据间进行对比学习，并考虑不同模态之间的不一致性与相似性信息，自适应地融合来自药物分子 SMILES 表征和图表征的语义与结构信息。该网络通过分子多模态数据间的交互作用，实现了模型对分子多视角互补信息的综合学习，以及药物分子 ADMET 的预测精度。

1.4.2 论文组织安排

第一章为绪论。从现实与理论出发，概述本研究的研究背景，阐述本研究的理论价值与实践价值，并介绍分子性质预测方法的国内外研究现状。

第二章为相关技术和理论基础。首先介绍药物研发和分子性质预测的相关理论基础；然后介绍深度学习神经网络中各种常用算法的网络结构和作用。

第三章为基于知识先验和多通道注意力的生物活性预测算法。首先介绍药物分子生物活性值的基础背景；然后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；其次基于多通道注意力机制构建 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；最后根据获得的特征预测药物分子的生物活性值，并对实验结果与性能进行分析。

第四章为基于多模态自适应对比融合深度网络的抗乳腺癌药物分子性质预测网络。首先介绍多模态抗乳腺癌潜在药物分子筛选技术；然后构建深度神经网络有针对性地提取分子不同模态的特征；其次通过对比学习策略学习不同模态下的一致性和相同模态下的异质性，并动态更新权重，自适应融合不同模态信息；最后根据融合特征预测药物分子的 ADMET 性质，并对实验结果与性能进行分析。

第五章为总结与展望。本章总结前述章节的内容，并对抗乳腺癌药物分子性质预测技术做出展望，希望进一步探索出更好的工作。

第2章 相关技术和理论基础

2.1 药物的生物活性与 ADMET 性质

2.1.1 药物的生物活性

药物的生物活性是指药物分子与生物体内的靶标相互作用，从而引发特定生物学效应的能力。这种活性通常通过药物与靶标（如酶、受体、离子通道等）的结合强度、选择性以及引发的功能变化来衡量，是药物发挥治疗作用的基础，也是药物筛选和优化过程中的关键指标。

在药物的生物活性研究中，pIC50（负对数半抑制浓度）是最常用的评估参数[72]，用于量化药物对特定靶标的抑制能力。该参数与 IC50（半数抑制浓度）存在直接数学关联。其中，IC50 是指药物抑制靶标活性 50% 所需的浓度。然而，由于 IC50 值通常跨越多个数量级（可从纳摩尔级至毫摩尔级），这种宽泛的数值分布往往呈现出显著的长尾特征，不利于建立稳健的预测模型。为解决这一问题，研究人员采用负对数转换方法，将 IC50 值转化为 pIC50 值。这种数学处理将原始数据压缩到 1-12 的范围内，显著改善了数据的分布特征，使得模型能够更容易地学习到数据的潜在特征。具体转换公式如下：

$$\mathrm{pIC50} = -\log_{10}(\mathrm{IC50}) \tag{2.1}$$

其中，pIC50 值越大，表示 IC50 值越小，药物的活性越强，即较低的药物浓度就能达到 50%的抑制效果。pIC50 值越小，IC50 值就越大，则药物的活性越弱，需要较高的浓度才能发挥相同的抑制作用。

2.1.2 药物的 ADMET 性质

在药物研发领域，ADMET 性质（吸收 Absorption、分布 Distribution、代谢 Metabolism、排泄 Excretion、毒性 Toxicity）是评估候选药物成药性的核心指标。这些特性可直接影响药物的生物可利用度、安全性、疗效和最终的市场批准，对药物化合物的设计和研发至关重要。

（1）吸收（Absorption）：药物的吸收特性直接影响其进入体内的速度和程度。通过分析化合物的吸收特性，可以评估其进入血液循环的效率，以便确定最佳的给药途径和给药剂型。

（2）分布（Distribution）：药物的分布特性决定了其在体内各组织的浓度分

布。通过分析化合物的分布特性，不仅可以评估药物的疗效和潜在副作用，还可以准确定位药物的有效作用靶点。

（3）代谢（Metabolism）：药物在体内通过酶促反应转化为代谢产物。虽然代谢通常有助于药物排泄，但也可能导致药物失活或产生毒性代谢物。通过分析化合物的代谢特性，可研究其降解动力学和代谢产物，进而预测药物的半衰期。

（4）排泄（Excretion）：药物及其代谢产物从体内清除的过程。通过分析化合物的排泄特性，可以优化其给药剂量和给药频率，防止药物在体内过度蓄积。

（5）毒性（Toxicity）：药物在体内蓄积后引发的副作用或不良反应。通过分析化合物的毒性特性，不仅可以预防不良反应的发生，还能帮助确定药物的安全使用范围。

2.2 分子表征

在药物化学与分子建模研究领域，分子表征技术为理解化合物结构、性质及其与生物靶标的相互作用关系提供了基础支撑。目前，主流的分子表征方法包括：分子描述符（Descriptor）、分子指纹（Fingerprints）、分子 SMILES（Simplified Molecular Input Line Entry System）以及分子图结构表示（Graph）。每种表征方法都有其独特的优势和应用场景，基于这些表征方法，研究人员可以开展分子筛选、药物设计、虚拟筛选以及 QSAR 建模等关键研究。

（1）分子描述符（Descriptor）：分子描述符是通过量化分子结构特征得到的数值化表示。它们通常基于分子的物理化学性质，如分子量、极性、拓扑性质等。分子描述符可以分为结构描述符、拓扑描述符、电子描述符等多种类型。例如，分子量描述了分子的大小，logP（分配系数）则反映了分子在油水之间的分配情况，这对药物的吸收、分布和渗透性至关重要。分子描述符能够为机器学习算法提供有效的输入特征，是药物设计和筛选的重要工具。

（2）分子指纹（Fingerprints）：是一种通过特定算法将分子转化为一系列二进制位的表示形式，每个位代表分子中某一特征的存在或缺失。常见的分子指纹包括 MACCS（Molecular ACCess System）键指纹、ECFP（Extended Connectivity Fingerprints）指纹、Daylight 指纹等。指纹表征方法通过对分子结构的特征进行编码，能够高效地描述分子的构成和相似性，还能在一定程度上降低数据维度，简化计算，非常适用于大规模分子库的筛选。在药物相似性分析、虚拟筛选以及分子对接中，分子指纹被广泛应用于快速匹配和查找具有相似活性的分子。

（3）分子 SMILES（Simplified Molecular Input Line Entry System）：SMILES 是一种以文本形式表示分子结构的简洁方法，它通过字符编码来表示化学分子中的原子和键。SMILES 可以直观地表达分子的结构信息，并且易于计算机处理。如水（H_2O）的 SMILES 表示为[H]O[H]或 O（氢原子通常被隐式表示）。更复杂的分子，如苯乙烯（C_6H_5-CH=CH_2），也可以通过 SMILES 表示：C=CC1=CC=CC=C1，C=C 表示乙烯基部分（CH=CH_2）；C1=CC=CC=C1 表示苯环部分（C_6H_5）；1 是环闭合标记，表示苯环的起点和终点。这种简洁高效的分子表示方法为大规模分子数据库的构建和检索提供了重要支持。

（4）分子图（Graph）：分子图（Graph）表示法是一种基于图论的分子结构表征方法。它将分子中的原子映射为图的节点，而化学键则对应为连接节点的边。这一方法能够全面而精确地捕捉分子的结构特征，如：原子间的连接方式、环系结构以及立体化学信息等。其优势在于，它能够自然地处理分子的结构复杂性，并充分利用图结构中的拓扑信息进行特征学习。通过将分子转化为图结构的形式，研究人员能够有效提取分子内部的复杂拓扑信息，从而显著提升药物性质中的预测精度。

不同的分子表征方法有不同的优势和应用领域，在分子性质研究中扮演着互补的角色。具体而言：分子描述符方法擅长量化分子的理化性质，为传统的统计分析提供可靠的数据基础；分子指纹技术因其计算速度较快，在大规模虚拟筛选和分子相似性评估中展现出优异的表现；分子 SMILES 表示法则以其简洁直观的特点，为分子信息的存储和检索提供了高效解决方案；而分子图表示方法则能够精确捕捉分子的复杂拓扑结构，为基于深度学习的分子分析开辟了新途径。随着人工智能和机器学习算法的创新应用，分子表征方法正在经历显著的技术革新，为药物设计和分子优化提供了更加高效的工具，推动着药物研发向智能化、精准化方向不断迈进。

2.3 深度学习模型

深度神经网络通过模拟人脑神经网络的工作机制，能够从海量数据中自主学习复杂的特征表示。这种网络架构的核心在于通过构建多层人工神经网络，提取层次化的抽象特征。在训练过程中，网络利用梯度下降算法不断优化各层的权重和偏置参数，从而提升模型对数据的拟合能力，最终实现高精度的预测结果。不同于传统机器学习依赖于人工特征工程，深度学习通过堆叠多个隐藏层，能够从

原始数据中自动学习从低阶到高阶的特征表示，并采用反向传播算法进行端到端的模型训练，且为每个神经元都引入了非线性激活函数，大大增强了模型的表达能力。具体计算公式如下所示：

$$z^{[l]} = \mathbf{W}^{[l]} x^{[l-1]} + b^{[l]} \tag{2.2}$$

$$x^{[l]} = f^{[l]}(z^{[l]}) \tag{2.3}$$

其中，$x^{[l-1]} \in \mathbb{R}^{m_{l-1} \times 1}$ 表示第 $l-1$ 层的输出，m_{l-1} 是第 $l-1$ 层神经元的数量；$\mathbf{W}^{[l]} \in \mathbb{R}^{m_l \times m_{l-1}}$ 表示第 l 层的权重矩阵，m_l 是第 l 层的神经元数量；$b^{[l]} \in \mathbb{R}^{m_l \times 1}$ 是第 l 层的偏置项；$z^{[l]} \in \mathbb{R}^{m_l \times 1}$ 是线性变换的结果；$f^{[l]}(z)$ 是第 l 层的激活函数，用于引入非线性变换，如 ReLU（Rectified Linear Unit）：$f(x) = \max(0, x)$；$x^{[l]} \in \mathbb{R}^{m_l \times 1}$ 是第 l 层的输出。

当前主流的深度学习架构包括卷积神经网络、图神经网络和 Transformer 架构等，这种层级化的特征学习机制，使其在计算机视觉、自然语言处理、语音识别等领域取得突破性进展。下述章节将详细介绍本研究所涉及到的深度学习算法。

2.3.1 Efficient Channel Attention Mechanism

ECA（Efficient Channel Attention）注意力机制[73]是一种轻量级的通道注意力机制，旨在高效地捕捉网络通道之间的依赖关系，并优化深度神经网络模型的性能，特别适用于计算资源有限的情况。其核心思想是通过对每个通道的响应进行加权，来学习通道之间的相关性。不同于传统的通道注意力机制，如 SE-Net（Squeeze-and-Excitation Networks）[74]和 CBAM（Convolutional Block Attention Module）[75]等，ECA 通过局部跨通道卷积的方式代替了全连接层操作，从而减少了网络的计算量，并且可以更加高效地捕捉跨通道特征的依赖关系。具体框架见图 2.1 所示。

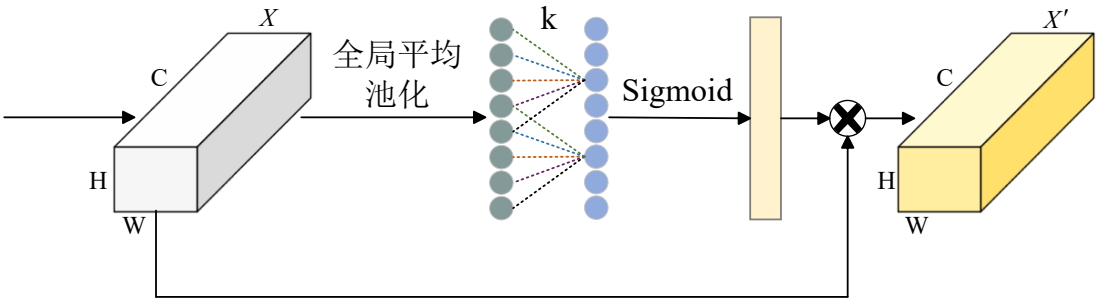


图 2.1 ECA 注意力机制结构图

如图 2.1 所示，当给定输入数据 $X \in \mathbb{R}^{H \times W \times C}$ 时（其中 H 是特征图的高度（空间维度）；W 是特征图的宽度（空间维度）；C 是通道数），ECA 机制首先对每个通道的空间维度（即 $H \times W$ ）进行全局平均池化，为每个通道生成一个标量 z_c，用于表示其全局信息：

$$z_c = \frac{1}{H \times W} \sum_{h=1}^{H} \sum_{w=1}^{W} X_{h,w,c} \tag{2.4}$$

其次，为网络定义一个卷积核大小为 k 的卷积操作，其权重参数则在模型训练过程中通过反向传播自动学习，用于生成通道注意力权重。与传统的卷积方法不同，ECA 机制中的卷积核大小并非固定值，而是通过公式（2.5）自适应确定。这种自适应机制能够根据特征图的通道维度，动态调整感受野大小。具体计算公式如下所示：

$$k = \psi(C) = \left| \frac{\log_2(C)}{\gamma} + \frac{b}{\gamma} \right|_{odd} \tag{2.5}$$

$$s_c = \sigma\left(\mathbf{W}_k * z_c\right) \tag{2.6}$$

$$\sigma(x) = \frac{1}{1+e^{-x}} \tag{2.7}$$

$$\mathbf{W}_k = \begin{bmatrix} w^{1,1} & \cdots & w^{1,k} & 0 & 0 & \cdots & \cdots & 0 \\ 0 & w^{2,2} & \cdots & w^{2,k+1} & 0 & \cdots & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & \cdots & 0 & 0 & \cdots & w^{C,C-k+1} & \cdots & w^{C,C} \end{bmatrix} \tag{2.8}$$

其中，σ 是 Sigmoid 非线性激活函数，用于将输出映射到[0,1]之间；$\mathbf{W}_k$ 是卷积核；*表示卷积操作，为每个通道的生成注意力权重。从公式（2.8）可以看出，ECA 注意力机制采用了局部跨通道交互策略。具体而言，该机制仅考虑当前特征与其 k 个邻近通道间的相互作用，通过这种轻量级的注意力建模方式，在保证模型计算效率的同时，实现了显著的性能提升。最后，生成的通道注意力权重 s_c 会被用来重新标定原始输入特征图 X，得到加权后的输出 X'。

2.3.2 Multi-head Attention Mechanism

多头注意力机制（Multi-Head Attention, MHA）[76]作为深度学习领域的重要突破，在深度网络构建中得到了广泛的应用，尤其是在 Transformer 架构及其衍生模型中，已成为特征学习的核心技术之一。该机制的核心思想是通过并行计算多

个独立的注意力子空间，从不同维度捕捉输入数据间的复杂依赖关系，从而显著提升模型的表征能力和学习效率。

与传统的单一注意力机制相比，多头注意力具有显著优势：传统方法通常只能关注输入数据的某个特定方面，而多头机制通过并行计算 h 个独立的注意力头，使每个头能够专注于数据的不同特征子空间，能够从多个维度捕捉输入数据的复杂特征。例如，在自然语言处理中，不同的注意力头可以分别关注语法依赖、语义关系或上下文信息等不同层面的特征，从而实现对文本内容的全面理解。这种设计不仅增强了模型的表达能力，还提高了对复杂模式的捕捉能力。具体框架见图 2.2 所示。

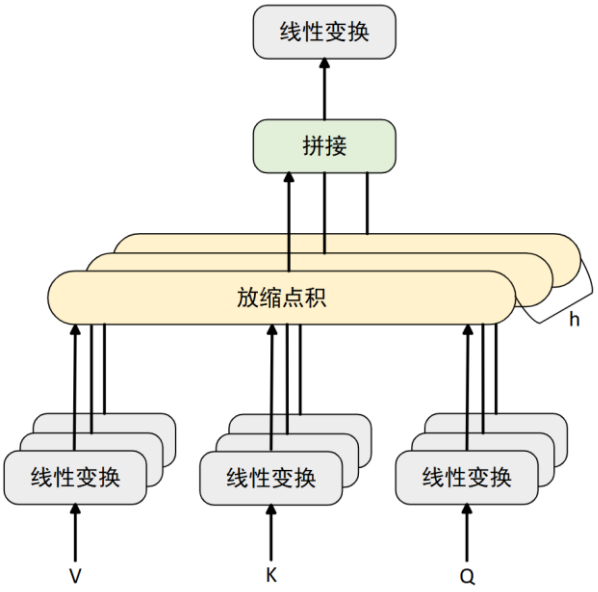


图 2.2 多头注意力机制结构图

如图 2.2 所示，当给定输入序列表示 $X \in \mathbb{R}^{n\times d}$ 时（其中 n 是序列长度；d 是每个词的表示维度），多头注意力机制中的每个注意力头，都会通过三个可学习的权重参数 W_Q，W_K，W_V，将输入序列分别映射为查询（Query）、键（Key）和值（Value）：

$$Q = XW_Q,\ K = XW_K,\ V = XW_V \tag{2.9}$$

其次，通过计算查询 Q 与键 K 的点积，得到未归一化的注意力分数。为了稳定梯度传播，这些分数会经过缩放因子 $\sqrt{d_k}$ 进行调整。最后，对缩放后的分数应用 Softmax 函数，将其转换为概率分布形式的注意力权重：

$$\text{Attention}(Q,K,V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{2.10}$$

$$\text{Softmax}(z_i)=\frac{\exp(z_i)}{\sum_j \exp(z_i)} \tag{2.11}$$

其中，d_k 是键和查询的维度。随后，每个头基于独立的可学习参数 W_Q，W_K，W_V 和查询、键、值矩阵，计算各自的注意力：

$$\text{head}_i=\text{Attention}(QW_i^Q,KW_i^K,VW_i^V) \tag{2.12}$$

最后，将所有头的输出拼接在一起并通过一个线性变换得到最终结果：

$$\text{MultiHead}(Q,K,V)=\text{Concat}\left(\text{head}_1,\ldots,\text{head}_h\right)W_O \tag{2.13}$$

其中，h 是头的数量，$W_O\in\mathbb{R}^{hd_v\times d}$ 是输出的权重矩阵，d_v 是值（Value）的维度。通过多个注意力头的协同工作，降低了模型对单一注意力头的依赖，有效提高了模型的鲁棒性。

2.3.3 Transformer

Transformer 是一种基于自注意力机制（Self-Attention，SA）的深度学习架构，由 Vaswani A 等[76]在 2017 年首次提出。该架构最初被设计用于解决自然语言处理（Natural Language Processing, NLP）领域的序列到序列（sequence-to-sequence, seq2seq）任务，如机器翻译、文本分类等。其核心创新在于利用自注意力机制来建模序列元素间的全局依赖关系，突破了传统 RNN 在处理长距离依赖时的局限性。Transformer 架构主要由编码器（Encoder）和解码器（Decoder）两部分组成，每部分都包含多层相同的子模块。其中，编码器负责将输入序列转换为特征表示，而解码器则根据编码器的输出生成目标序列。这种模块化设计不仅提高了模型的可扩展性，还为处理不同长度的序列数据提供了灵活性。值得注意的是，Transformer 完全摒弃了传统的循环结构，转而采用位置编码（Positional Encoding）来保留序列的顺序信息。这种设计使得模型能够并行处理整个序列，显著提高了训练效率。具体框架见图 2.3 所示。

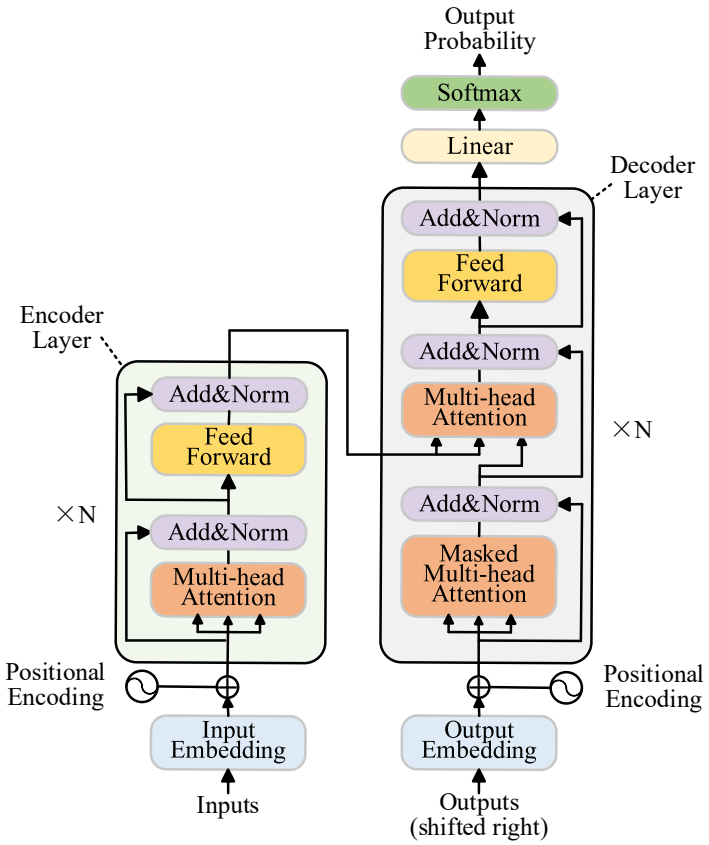


图 2.3 Transformer 结构图

如图 2.3 所示，当给定输入序列表示 $X \in \mathbb{R}^{n \times d}$ 时（其中 n 是序列长度；d 是每个词的表示维度），Transformer 首先通过嵌入层（Embedding Layer）将每个离散的输入符号 x_i 映射为 d 维的连续向量表示。随后，为了保留序列的顺序信息，模型会为每个位置 i 添加相应的位置编码（Positional Encoding，PE）。这种设计使得模型能够在不使用循环结构的情况下，有效地处理序列数据中的顺序依赖关系。最终的输入表示由词嵌入和位置编码相加得到，其计算过程如下：

$$PE_{(pos,2i)} = \sin\left(\frac{pos}{10000^{2i/d_{\text{model}}}}\right) \tag{2.14}$$

$$PE_{(pos,2i+1)} = \cos\left(\frac{pos}{10000^{2i/d_{\text{model}}}}\right) \tag{2.15}$$

$$X_{input} = X + PE \tag{2.16}$$

其中，d_{model} 是词向量的维度，i 是词向量的第 i 维，pos 是当前词的绝对位置。然后，输入序列会依次通过多个编码器层。每个编码器层由两个核心子模块组成：多头注意力机制和前馈神经网络（Feedforward Network, FFN）。其中，MHA 能够

从不同子空间捕捉输入序列的多样化特征，而 FFN 则通过非线性变换进一步整合这些特征。具体而言，对于序列中的每个位置，FFN 模块首先对输入向量进行线性变换和 ReLU 激活；然后，通过另一个线性变换将维度映射回原始大小。这种设计使得每个位置的输出向量都包含了丰富的局部和全局上下文信息，为后续的解码过程提供了高质量的特征表示。具体计算公式如下：

$$\mathrm{FFN}(x)=\max(0,xW_1+b_1)W_2+b_2 \tag{2.17}$$

其中，W_1 和 W_2 是权重；b_1 和 b_2 偏置；$\max(0,x)$ 表示 ReLU 激活函数。此外，在 Transformer 架构中，每个子层（SA 层和 FFN 层）都采用了残差连接（Residual Connection，RC）和层归一化（Layer Normalization，LN）操作，从而缓解深层网络中的梯度消失问题，并加速模型的训练过程。

在解码器部分，其整体架构与编码器相似，但在细节设计上存在一些关键差异。首先，解码器的输入通常是右移一位的目标序列，这种设计确保了模型在训练时只能看到当前位置之前的信息，从而防止信息泄露。其次，解码器采用了带遮蔽的自注意力机制（Masked Self-attention），通过掩码操作限制每个位置只能关注到其左侧的位置，从而保证序列生成的自回归特性。此外，解码器还增加了一个额外的多头注意力层，用于关注编码器的输出表示。经过多层解码器的处理，最终的输出向量会通过一个线性变换层（通常是全连接层）映射到目标词汇表的大小。随后，对每个位置的输出向量应用 Softmax 函数，计算目标词汇的概率分布，模型根据当前的概率分布选择最可能的词作为输出，并将其作为下一个时间步的输入，直到生成序列结束标记。这种自回归生成方式不仅保证了序列生成的连贯性，还能够灵活处理不同长度的输出序列，在自然语言处理领域中展现出卓越的性能。

2.3.4 Bidirectional Encoder Representation from Transformers

BERT（Bidirectional Encoder Representations from Transformers）[77]是一种基于 Transformer 编码器架构的预训练语言模型，在自然语言处理领域得到了广泛应用。该模型的核心创新在于其双向编码器设计和两阶段训练策略。与传统语言模型不同，BERT 通过掩码语言模型（Masked Language Model, MLM）和下一句预测（Next Sentence Prediction, NSP）任务进行预训练，使得模型能够同时捕捉每个单词的前后上下文信息，从而更深入地理解文本的语义关系。这种设计显著提升了模型在多种 NLP 任务中的表现，包括文本分类、问答系统和命名实体识别等。

具体框架见图 2.4 所示。

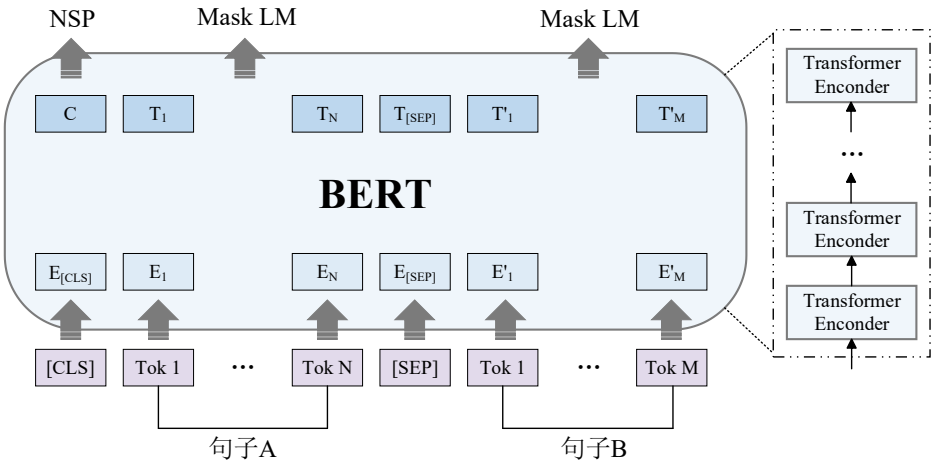


图 2.4 BERT 结构图

如图 2.4 所示，BERT 的架构主要由多个 Transformer 编码器层堆叠而成，但在位置编码的设计上与原始 Transformer 有所不同。具体而言，BERT 采用可学习的位置编码，通过随机初始化并在预训练过程中优化，而不是使用固定的三角函数编码。这种设计使得模型能够自适应地调整位置表示，更好地捕捉文本中的语义关系和相对位置信息。BERT 的预训练过程包含两个核心的自监督学习任务：MLM 和 NSP。

在 MLM 任务中，部分输入单词会被随机掩码掉，并要求模型预测这些被遮蔽的单词。这一策略使得模型能够从部分信息中推测出整个语义，更好地理解单词间的相互关系，从而学习双向上下文表示。假设输入句子为 $X=[x_1,x_2,...,x_n]$，随机掩码部分单词后得到序列 $X'=[x_1,x_2,[\text{Mask}],...,x_i,...,x_n]$，在后续的训练过程中 BERT 通过左侧和右侧的上下文信息学习每个单词的表示来预测被遮蔽的单词，即：

$$P(x_m \mid X'_{-m}) = \text{softmax}(W_1 h_m + b_1) \tag{2.18}$$

其中，x_m 是被掩蔽的词；X'_{-m} 表示去掉 x_m 后的输入上下文；h_m 是经过 Transformer 编码器处理后每个词的上下文特征表示；W_1 和 b_1 是权重矩阵和偏置项，用于映射到词汇表空间。

而 NSP 任务则是判断两个句子是否连续，即判断句子 A 和句子 B 是否在原文中相邻，从而捕捉句子级别的语义关联和逻辑关系。当给定句子 A 和句子 B 时，这两个句子会被拼接成一个长文本输入到 BERT 中，如‘[CLS] Sentence A [SEP] Sentence B [SEP]’，其中，[CLS]是用于表示整个输入句子对的特殊标记；[SEP]

是用于分隔句子 A 和句子 B 的特殊标记。随后，模型会通过 Transformer 的编码器生成一个全局上下文表示，并基于该表示进行计算得到一个二分类概率：

$$P(y=1|A,B)=\text{sigmoid}(W_2[h_A;h_B]+b_2) \tag{2.19}$$

其中，h_A 和 h_B 分别表示句子 A 和句子 B 经过 Transformer 处理后得到的全局表示；W_2 和 b_2 是用于分类的权重和偏置；$[h_A;h_B]$ 表示将句子 A 和句子 B 的表示拼接成一个向量。

通过这两个任务的结合，BERT 不仅能捕捉单词层面的细节，还能处理句子层面的逻辑和语义关系。此外，BERT 的多任务学习策略使其在多种下游任务中展现出卓越的性能。例如，在预训练阶段，BERT 通过大规模文本数据学习通用的语言表示，随后通过少量的微调即可快速适应不同的 NLP 任务。这种预训练-微调框架大大降低了特定任务的训练成本，并提升了模型的泛化能力。

2.3.5 Graph Neural Network

图神经网络（Graph Neural Network, GNN）[21]是一类专门用于处理图结构数据的深度学习模型，其核心组件包括消息传递机制、邻接矩阵和图卷积操作等。与传统神经网络主要处理欧几里得数据（如网格或序列数据）不同，GNN 能够有效处理由节点和边构成的非欧几里得数据结构。这种特性使得 GNN 在社交网络分析、分子性质预测、知识图谱推理和推荐系统等复杂场景中展现出显著优势。GNN 的核心思想是通过消息传递机制，使每个节点的特征表示不仅包含自身信息，还融合了邻居节点的特征信息。这种机制通过多轮迭代的消息传递和特征聚合，逐步学习节点的高阶表示。通过这种方式，GNN 能够同时捕捉图的局部结构信息和全局拓扑特征，为节点分类、链接预测和图分类等任务提供了强大的特征表示能力。具体框架见图 2.5 所示。

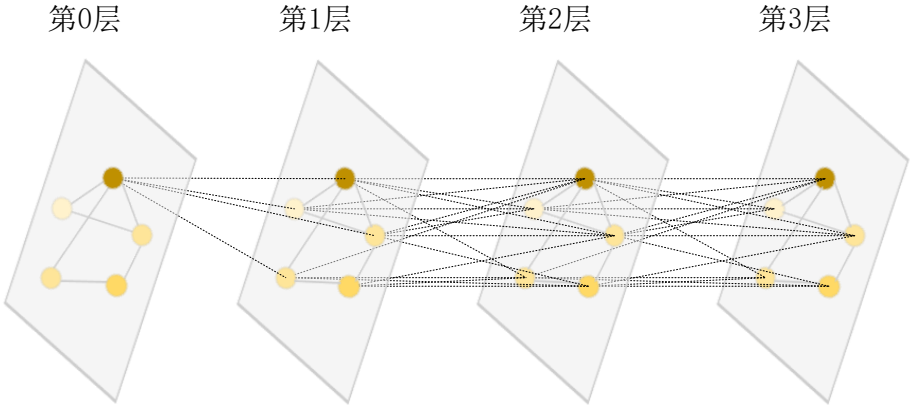


图 2.5 GNN 结构图

如图 2.5 所示，假设图$G=(V,E)$由n个节点和m条边组成，其中V是节点集合，E是边集合。每个节点$v_i \in V$都有一个特征向量$h_v^{(0)} \in \mathbb{R}^d$，其中$d$是特征的维度，将所有节点的特征向量组合成节点特征矩阵$\mathbf{H}^{(0)} \in \mathbb{R}^{n\times d}$。在 GNN 中，节点通过聚合其邻居节点的特征来更新自身特征：

$$h_v^{(l+1)} = \sigma\left(\sum_{u\in \mathcal{N}(v)} \frac{1}{\sqrt{d_v d_u}} \mathbf{W}^{(l)} h_u^{(l)} \right) \tag{2.20}$$

其中，$N(v)$是节点v的邻居集合；d_v和d_u分别是节点v和u的度数；$\mathbf{W}^{(l)}$是第l层的权重矩阵；σ是激活函数。

其变体图卷积网络（Graph Convolutional Network，GCN）[57]通过将节点的特征与其邻居节点的特征进行加权平均或其他形式的聚合，从而生成新的节点表示：

$$\mathbf{H}^{(l+1)} = \sigma\left(\hat{\mathbf{A}} \mathbf{H}^{(l)} \mathbf{W}^{(l)} \right) \tag{2.21}$$

其中，$\mathbf{A}=\mathbf{A}+\mathbf{I}$是邻接矩阵$\mathbf{A}$的归一化版本；$\mathbf{I}$是单位矩阵；$\mathbf{H}^{(l)}$是第$l$层的节点特征矩阵。

而图注意力网络（Graph Attention Network，GAT）[58]通过计算注意力权重来聚合邻居信息：

$$h_v^{(l+1)} = \sigma\left(\sum_{u\in \mathcal{N}(v)} \alpha_{vu}^{(l)} W^{(l)} h_u^{(l)} \right) \tag{2.22}$$

其中，$\alpha_{vu}^{(l)}$是节点v和u之间的注意力权重。

在进行最终的预测任务之前，通常需要对整个图的表示进行池化操作，以提取图的全局特征表示。最常见的池化方法包括全局平均池化（Global Average Pooling, GAP）（式 2.23）和全局最大池化（Global Max Pooling, GMP）（式 2.24）。其中，GAP 通过计算所有节点特征的平均值来获得图的全局表示，从而反映图的整体特征分布；而 GMP 则通过选取每个特征维度的最大值来构建图的表示，从而突出图中最显著的特征。

$$h_{\text{graph}} = \frac{1}{n} \sum_{v\in V} h_v \tag{2.23}$$

$$h_{global} = \max(h_1, h_2, \ldots, h_N) \tag{2.24}$$

2.4 本章小结

本章系统性地阐述了药物研发需要的基本知识和涉及到的深度网络模型。首

先，详细介绍了药物化合物生物活性值和 ADMET 性质的基本概念及其在药物研发中的核心作用；其次，全面梳理了分子表征的核心理论框架，重点阐述了分子描述符、分子指纹、SMILES 和图表示等方法在计算化学中的基础定义和基础应用；最后，深入剖析了多种深度学习算法的核心架构，包括深度神经网络、ECA 注意力机制、多头注意力机制、Transformer、BERT 以及图神经网络等模型，为后续的药物分子性质预测研究奠定了坚实的理论基础和方法论支撑。

第3章 基于知识先验和多通道注意力的抗乳腺癌药物分子生物活性预测算法

本章主要针对抗乳腺癌药物分子的生物活性预测问题进行研究，以 ERα 数据集为研究对象，以药物分子的 SMILES 表征为输入，引入知识先验和多通道注意力机制，构建深度学习预测算法（KBAC）。通过参数对比实验确定模型最佳参数设置，提高模型的精度和鲁棒性。并设计相关消融实验和对比实验对所提模型的有效性进行评估和分析。

3.1 引言

在抗乳腺癌药物研发中，雌激素受体 α（ERα）的 pIC50 生物活性值预测，对于分析药物分子性质，以及揭示药物分子和疾病的内在联系具有重要作用。传统实验筛选方法虽可靠，但因其通量低、成本高等问题，严重限制了药物研发的效率。基于深度学习的单模态药物分子生物活性预测方法凭借其计算高效性与可解释性优势，成为了早期虚拟筛选的关键技术手段。

单模态药物分子生物活性预测方法主要基于分子的单一模态数据（如分子描述符、指纹、图和 SMILES 等），构建相应的 ML 或 DL 模型挖掘药物的分子性质和结构信息，以实现高精度的生物活性预测，从而快速预筛潜在药物化合物，并显著降低实验验证的盲目性。如，基于分子指纹的随机森林模型可通过百毫秒级推理速度预测数万分子的抑制活性；GNN 通过直接解析分子图的拓扑与化学键特征，可实现对复杂结构-活性关系的深层挖掘。

然而，现有单模态方法仍面临诸多挑战和局限性。一方面，分子描述符或简化的二维分子图难以全面表征动态构象变化、溶剂化效应等关键三维化学信息，导致模型对某些靶标（如变构位点）的预测存在较大偏差；另一方面，对单一数据源的依赖性使模型易受数据稀疏性与噪声干扰——据统计，超过 70%的靶标蛋白仅有数百个公开活性数据点；同时，现有方法缺乏对分子重要特征的特别关注，导致其对药物分子性质的理解不足，从而在跨靶标预测任务中表现欠佳。

为应对这些问题，本章基于知识先验与注意力机制，提出了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，在第一阶段，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。在第二阶段，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。最后，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实现了分子 SMILES 的两阶段不同角度的特征提取，对于解释分子属性并实现准确预测具有重要意义。

3.2 数据来源及预处理

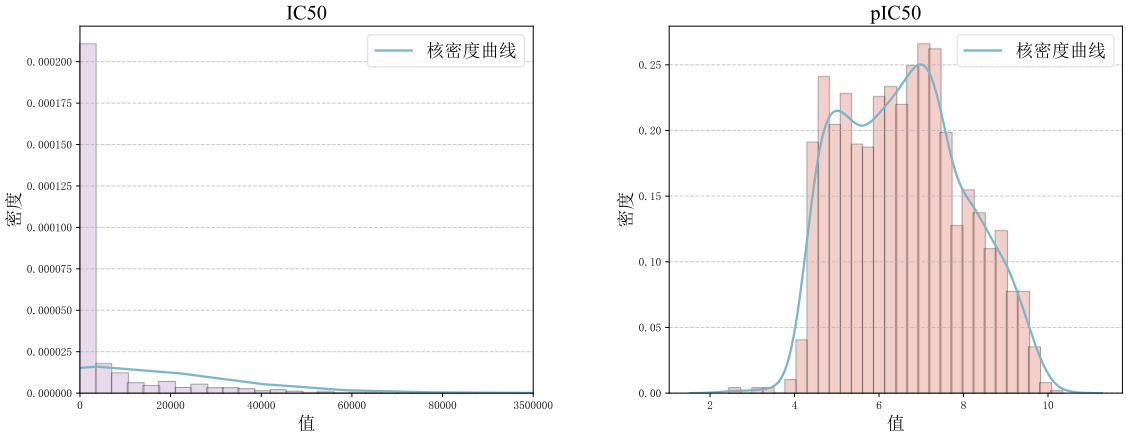
本章研究基于抗乳腺癌 ERα数据集进行，以药物分子的 SMILES 表征作为输入特征，以药物分子的 pIC50 值（生物活性评估参数）作为目标变量，构建抗乳腺癌候选药物化合物的生物活性预测模型。

3.2.1 雌激素受体 α（ERα）数据集

雌激素受体 α（ERα）作为乳腺癌的关键生物标志物之一，对乳腺癌细胞的增殖和生长产生了重要影响。雌激素通过与 ERα 结合，激活受体介导的信号通路，促进乳腺癌细胞的增殖与存活。因此，ERα 被广泛认为是乳腺癌治疗中的一个重要靶点。本研究使用的 ERα 数据集来自阿尔伯塔大学的 drugbank 药物分子数据库，收集了 1974 个抗乳腺癌候选药物化合物的 SMILES 表征、对应的 IC50 值、pIC50 值、以及相应的 ADMET 性质类别标签。其中，IC50 取值范围为 0.046-3500000，活性值主要集中在 0-20000 之间，其分布严重不均，且向左倾斜，不利于模型预测；pIC50 取值范围为 2.456-10.337，活性值主要集中在 4-9 之间，其分布较均匀，符合建模条件。部分数据展示和对应的生物活性值频率分布见表 3.1 和图 3.1。

表 3.1 ERα 数据集（生物活性部分）示意表

SMILES	IC50	pIC50
Oc1ccc2O[C@H]([C@H](Sc2…	2.5	8.602
Oc1ccc2O[C@H]([C@H](Sc2c…	7.5	8.125
Oc1ccc(cc1)[C@H]2Sc3cc(O)cc…	3.1	8.509
Oc1ccc2O[C@H]([C@@H](CC…	3.9	8.409
Oc1ccc2O[C@H]([C@@H](Cc3…	7.4	8.131
Oc1ccc2O[C@H]([C@H](Sc2c1…	490	6.310



(A) IC50值-频率分布直方图和核密度图　　(B) pIC50值-频率分布直方图和核密度图

图 3.1 ERα 数据集的频率分布直方图和核密度图

3.2.2 数据预处理

深度学习模型通常使用大量带标注数据进行训练，以确保其具有高效性和泛化性。然而，在分子生物学和化学领域，由于人力、资金和实验条件的限制，使得研究人员获取大量标注数据较为困难。这一情况直接导致标注数据的稀缺，从而影响模型的泛化能力和过拟合风险。因此，本章研究引入了 SMILES 枚举策略，旨在通过增强分子表征的多样性和数量，从而提高深度学习模型在分子预测任务中的性能。

具体而言，针对同一化合物分子，以其标准的 SMILES 表达式为基础，使用 Python 软件中的 RDKit 库进行分子结构计算和分析，随机生成 4 个不同的 SMILES 表达式，作为原始数据的增强表示。此时，每个分子将对应 5 个不同的 SMILES 表达式，包括一个标准表征和四个增强表征。这种数据增强策略使得模型能够通过不同的 SMILES 表征进行训练，从而更全面地捕捉化学结构中的细微变化，显著提高了模型对分子结构的理解能力和泛化性能。最后将数据集按照 8:1:1 的比例随机划分为训练集、验证集和测试集。部分数据示例见表 3.2。

表 3.2 增强后的 SMILES 表达式数据表

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

3.3 KBAC 模型设计

3.3.1 框架设计

本章基于知识先验与通道注意力机制，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。其整体架构如图 3.2 所示。

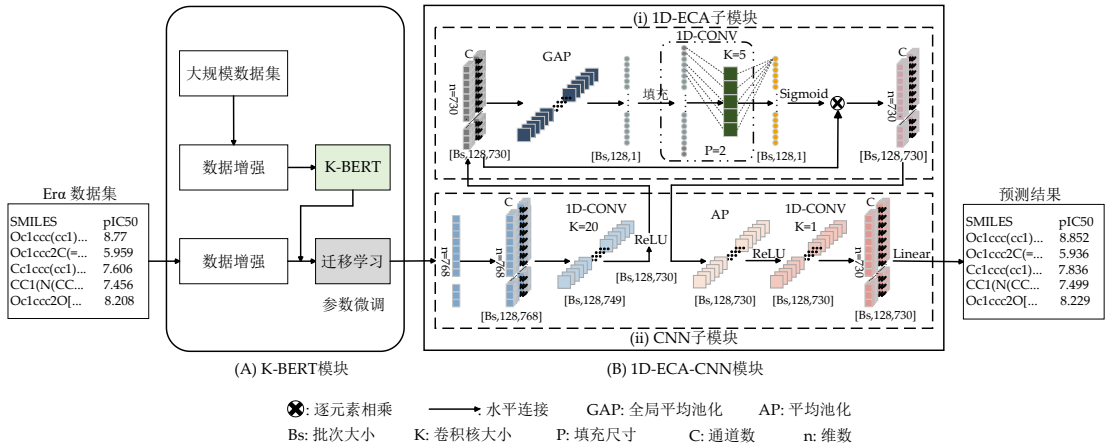


图 3.2 KBAC 整体框架图

如图 3.2 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。具体而言，1) 基于知识先验的分子特征提取：该模块引入 K-BERT 网络和迁移学习策略，将基于描述符、基于图形和基于 SMILES 的方法结合起来，以获取综合的 SMILES 分子表征信息。2) 基于多通道注意力机制的二次特征提取：该部分设计的 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。对应模块的详细信息介绍如下。

3.3.2 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。

（1）K-BERT 子模块

K-BERT 模型由 WU Z X 等[78]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，学习其中包含的专业化学知识。本小节采用 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三个预训练任务，具体如图 3.3 所示。

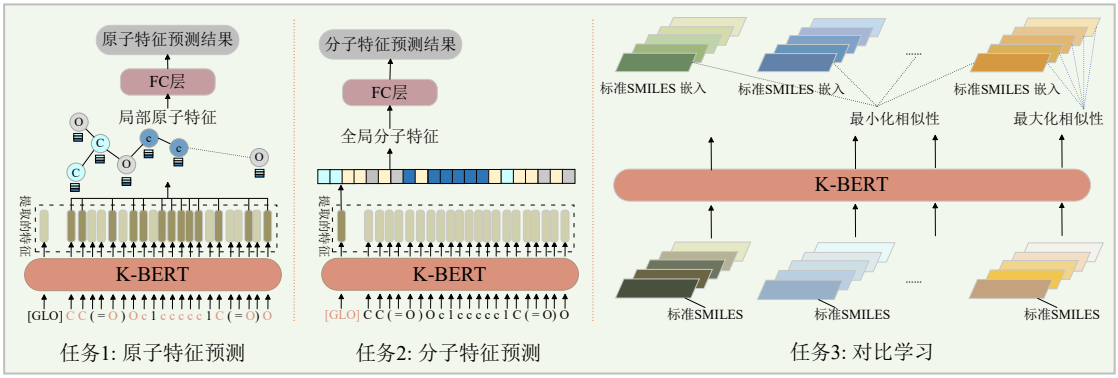


图 3. 3 K-BERT 框架图

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。

通过三个方法的结合使用，帮助模型更好的理解分子 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）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。

$$L_{CL}=\sum_{n=1}^{N}\sum_{d\epsilon D_n}\frac{1}{2}\left(1-\cos\left(\boldsymbol{E}_{n,c},\boldsymbol{E}_{n,d}\right)\right)+\sum_{n=1}^{N}\sum_{m\epsilon B_n}\cos\left(\boldsymbol{E}_{n,c},\boldsymbol{E}_{n,m}\right) \tag{3.1}$$

$$\cos(\boldsymbol{E}_a,\boldsymbol{E}_b)=\frac{\boldsymbol{E}_a\cdot\boldsymbol{E}_b}{\|\boldsymbol{E}_a\|\|\boldsymbol{E}_b\|} \tag{3.2}$$

$$\|\boldsymbol{E}_a\|=\sqrt{E_{a,1}^2+E_{a,2}^2+\cdots+E_{a,p}^2}|^2 \tag{3.3}$$

其中，N 表示这个批次中的分子数量；n 代表当前的分子；m 表示批次中其他分子的标准 SMILES 之一；D_n 表示由分子 n 的标准 SMILES 表达式产生的四种不同的 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；$E_{n,c}$ 表示由 K-BERT 模块生成的分子 n 的标准 SMILES 的嵌入；$E_{n,d}$ 表示由 K-BERT 模块生成的分子 n 的四个增强 SMILES 表达式的嵌入；$E_{n,m}$ 表示由 K-BERT 模块产生的此批次中分子 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；‖‖ 是嵌入的欧氏范数。

（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，本章将在 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 3.4 所示。

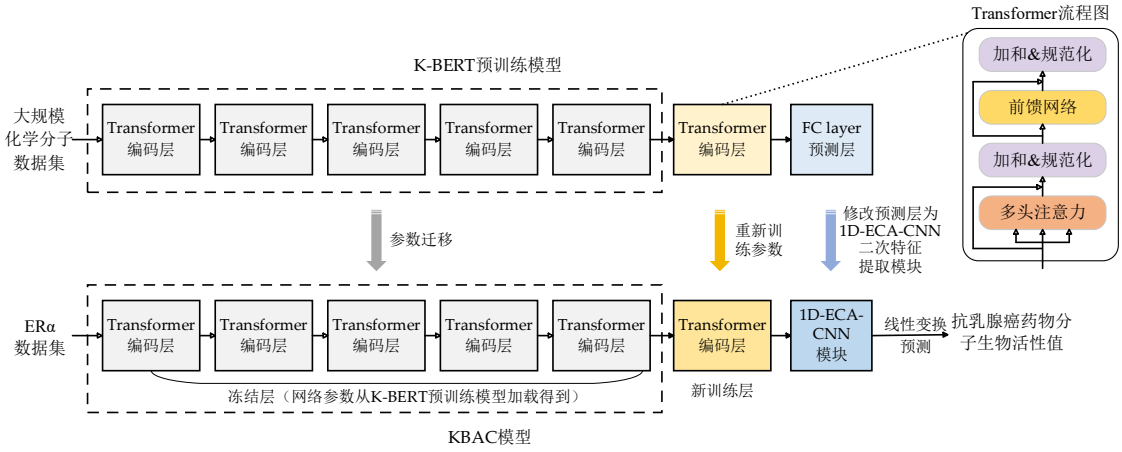


图 3.4 迁移学习流程图

3.3.3 基于多通道注意力机制的二次特征提取

为进一步提高分子表征的准确度和稳定性，本小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。具体过程如图 3.2（B）和图 3.5 所示。

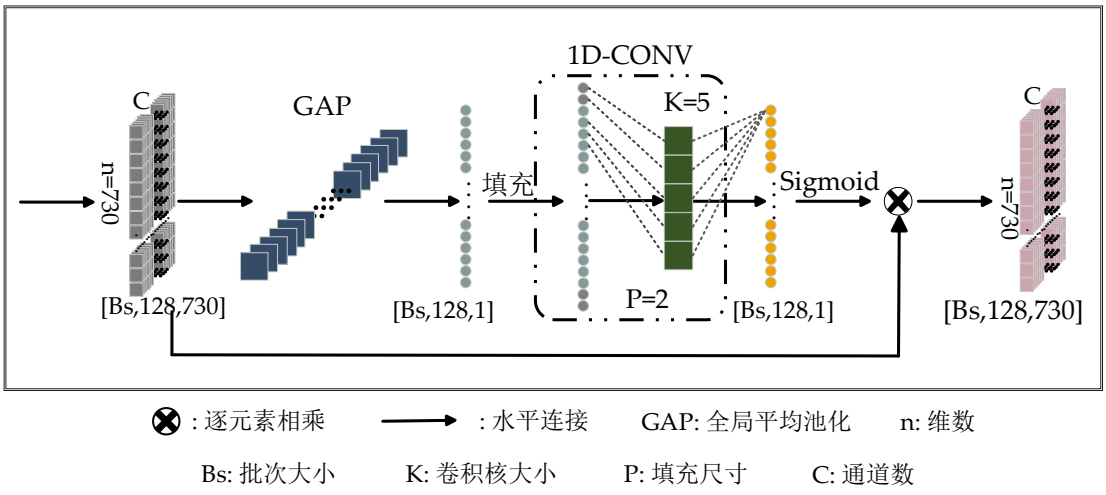


图 3.5 1D-ECA-CNN 流程图

（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，本小节设计 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 3.5 所示。

ECA 算法[73]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。本小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，因此可以通过下式公式推出：

$$C = \phi(k) = 2^{(\gamma * k - b)} \tag{3.4}$$

$$k = \psi(C) = \left| \frac{\log_2(C)}{\gamma} + \frac{b}{\gamma} \right|_{odd} \tag{3.5}$$

其中 $|t|_{odd}$ 表示最接近 t 的奇数，在本章中，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 $\mathbf{w}^{k}$ 来学习通道注意力。

$$\begin{bmatrix} w^{1,1} & \cdots & w^{1,k} & 0 & 0 & \cdots & \cdots & 0 \\ 0 & w^{2,2} & \cdots & w^{2,k+1} & 0 & \cdots & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & \cdots & 0 & 0 & \cdots & w^{C,C-k+1} & \cdots & w^{C,C} \end{bmatrix} \tag{3.6}$$

$$w_i = \sigma\left(\sum_{j=1}^{k} w^j y_i^j\right), y_i^j \in \Omega_i^k \tag{3.7}$$

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。

$$\mathbf{w} = \sigma(\mathrm{CONV}(\mathrm{GAP}(\mathbf{y}))) \tag{3.8}$$

因此，1D-ECA 模块公式具体如下：

$$\mathrm{M}_{\mathrm{ECA}}(\boldsymbol{X}) = \mathrm{Sigmoid}\left(\mathrm{CONV}\left(\mathrm{GAP}(\boldsymbol{X})\right)\right) * \boldsymbol{X} \tag{3.9}$$

其中，$\boldsymbol{X}$ 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 激活函数得到的特征；GAP 表示全局平均池化；CONV 表示一维卷积；Sigmoid 表示激活函数。

（2）1D-ECA-CNN 模块

上一节引入了多通道注意力机制，若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。

本小节将 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特征信息，降低高层次信息的丢失。详细流程见图 3.2（B）。

由图 3.2（B）和图 3.5 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：

$$\boldsymbol{M}_1 = \mathrm{M}_{\mathrm{ECA}}(\mathrm{ReLU}(\mathrm{CONV1}(\boldsymbol{X}))) \tag{3.10}$$

$$\boldsymbol{M}_2 = \mathrm{LN}(\mathrm{CONV2}(\mathrm{ReLU}(\mathrm{AP}(\boldsymbol{M}_1)))) \tag{3.11}$$

其中，$\boldsymbol{X}$ 表示经过迁移学习得到的分子 SMILES 表达式特征信息；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；$\mathrm{M}_{\mathrm{ECA}}$ 表示 1D-ECA 模块；AP 表示平均池化函数；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，提高所提 KBAC 模型的预测精度。

（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：

$$L_R = \frac{1}{m}\sum_{i=1}^{m}(y_i - y_i)^2 \tag{3.12}$$

其中，m 表示分子的数量，y_i 表示当前分子的 pIC50 真实值，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 L_T 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：

$$L_T = L_{CL} + L_R \tag{3.13}$$

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。

3.4 算法实现

本章所提 KBAC 算法利用两阶段特征提取策略，通过语义层和通道层有效学习分子数据的全局-局部特征，从而实现精确的生物活性预测。KBAC 框架的预测训练算法如表 3.3 所示。

表 3.3 KBAC 算法流程表

算法 **1　KBAC 训练步骤**
输入：药物分子 SMILES 表达式 ***X***；药物分子 pIC50 值 *Y*。
（1）Data augmentation(***X***) // 数据增强
（2）***T***=Fine-tuning K-BERT(***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）***C1***=ReLU*(*Conv1D*(****T****))*
（4）***W***=Sigmoid*(*Conv1D*(*GAP*(****C1****)))*
（5）***A***=***C1*W*** // 通道维加权融合
（6）***C2***=Conv1D*(*ReLU*(*AP*(****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）*Y_prediction*=Linear *(****C2****)* // 预测
输出：药物分子 pIC50 预测值。

3.5 实验结果及分析

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，来分析本章不同方面的工作，以评估所提 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。

3.5.1 评价指标

本章用于评估模型回归任务的指标主要有四个：R^2，MAE，MSE，RMSE。R^2（R-squared）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，其最大值为 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。

$$R^2 = 1 - \frac{\sum_{i=1}^{m}(y_i - y_i)^2}{\sum_{i=1}^{m}(\overline{y_i} - y_i)^2} \tag{3.14}$$

MAE（Mean Absolute Error）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）表示预测值与真实值之间的偏差程度，是在 MSE（Mean Square Error）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。

$$M_{MAE} = \frac{1}{m}\sum_{i=1}^{m}| y_i - \hat{y}_i | \tag{3.15}$$

$$M_{MSE} = \frac{1}{m}\sum_{i=1}^{m}\left(y_i - \hat{y}_i \right)^2 \tag{3.16}$$

$$R_{RMSE} = \sqrt{\frac{1}{m}\sum_{i=1}^{m}\left(y_i - \hat{y}_i \right)^2} \tag{3.17}$$

3.5.2 实验设置

本小节将详细描述模型训练时的参数设置。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 设置为 32，学习率设置为 3e-5，epoch 最大次数设置为 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置为 1 以外，1D-ECA-CNN 模块中其他层的通道数全部设置为 128，卷积核大小全部设置为 20。同时，在训练过程中，还运用了 early stop 策略，防止模型过拟合以及减少一些计算成本，并将 patience 设置为 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，还引入了 MASK 策略和

Pos_weigh 参数来提高模型的拟合性。具体参数设置见表 3.4。

表 3.4 训练参数表

Parameter Value	Value
Learning rate	3e-5
Batch size	32
Max epoch	100
Max token	201
Optimizer	Adam
Patience	20

3.5.3 参数对比实验分析

（1）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。本小节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。本小节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 为评价指标展示了不同尺寸的通道数和卷积核情况下的预测精度和效果。具体情况见表 3.5 所示。

表 3.5 神经网络参数对比实验表

卷积	通道数 16			通道数 32			通道数 64			通道数 128		
核	MAE	MSE	RMSE	MAE	MSE	RMSE	MAE	MSE	RMSE	MAE	MSE	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 3.5 是 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的大小，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，在卷积核大小维 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，本章实验采用的网络通道数为 128，卷积核大小为 20。

（2）网络速度收敛分析

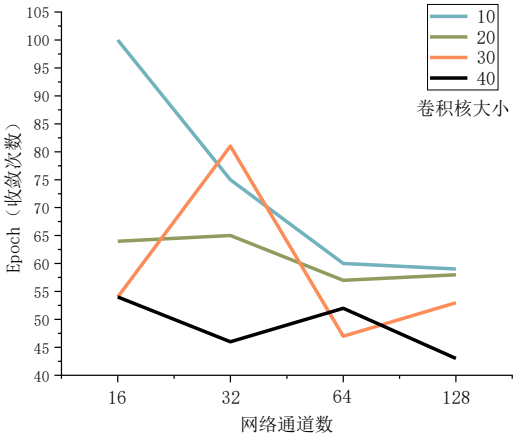


图 3.6 网络速度收敛分析实验图

图 3.6 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 3.6 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将会随之减少。

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，本章所提 KBAC 深度网络算法可以达到最优训练效果和预测精度，并有较好的收敛速度。

3.5.4 消融实验分析

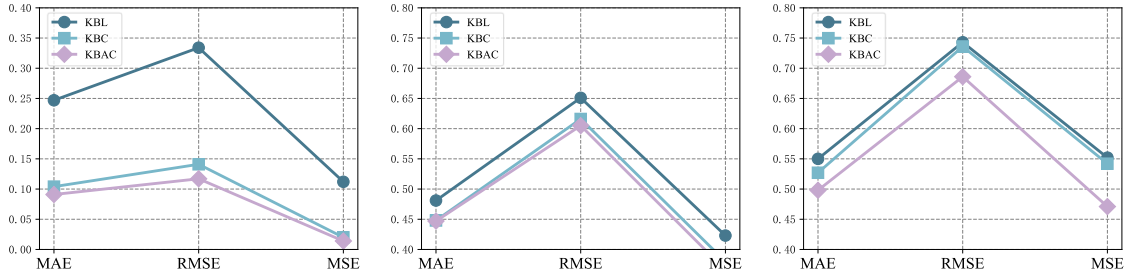
为说明 KBAC 的有效性，本小节将使用 ERα 数据集的训练集、验证集、测试集进行一系列消融实验。为此，基于 1D-ECA- CNN 模块设计做如下两个变体模型：KBL（Knowledge-BERT-Linear）和 KBC（Knowledge-BERT-CNN），具体情况如表 3.6 所示。

表 3.6 KBAC 及其变体结构信息表

模型	CNN	1D-ECA	K-BERT
KBL	×	×	√
KBC	√	×	√
KBAC	√	√	√

表 3.7 消融实验对比表

模型	训练集			验证集			测试集		
	MAE	RMSE	MSE	MAE	RMSE	MSE	MAE	RMSE	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**



(A) 模型在训练集中的性能表现　(B) 模型在验证集中的性能表现　(C) 模型在测试集中的性能表现

图 3.7 消融实验结果实验图

图 3.7 和表 3.7 是本章所提模型在消融情况下，以 MAE、RMSE、MSE 为评价指标，基于 KBL、KBC、KBAC 三个模型的实验结果比较，表中黑体数字表示最优值。由图 3.7 和表 3.7 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，模型性能均差于 KBC 模型和 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，模型性能优于 KBL，但均不及 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其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中的局部上下文信息，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 生物活性预测，从而影响模型的性能。

由图 3.7 和表 3.7 可知，本章所提 KBAC 模型在三个数据集以及三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，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。本章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。

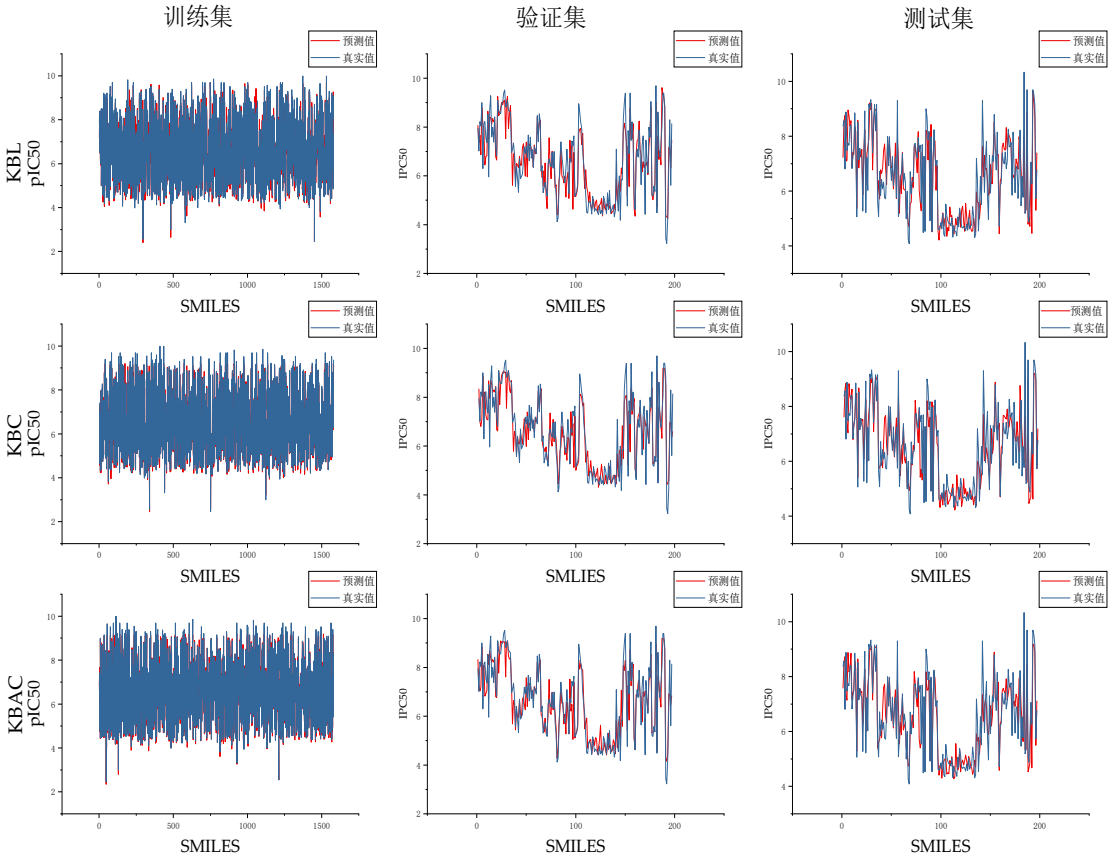


图 3.8 消融实验下 pIC50 的真实值-预测值对比图

图 3.8 是本章模型在消融情况下，KBL、KBC 和 KBAC 模型关于药物分子 SMILES 表达式的 pIC50 的真实值和预测值的对比图。

由图 3.8 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 KBL 模型的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。

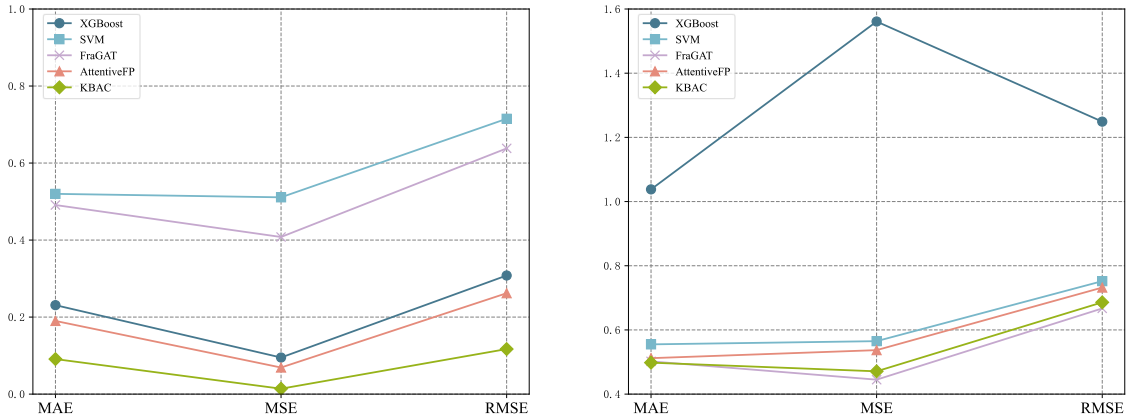
3.5.5 对比实验分析

为进一步说明本章所提模型的有效性，本小节选取当前最具代表性的四个模型进行比较，即：XGBoost[79]、SVM[34]、FraGAT[80]、Attentive FP[81]。其中 XGBoost、

SVM 为基于描述符的方法；FraGAT、Attentive FP 为基于图的方法。利用 MAE、MSE、RMSE、R^2 四个评价指标进行分析，具体对比结果见表 3.8、图 3.9 和图 3.10 所示，其中黑体数字表示最优值。

表 3.8 相关模型对比实验表

方法	训练集				测试集			
	MAE	MSE	RMSE	R^2	MAE	MSE	RMSE	R^2
XGBoost [79]	0.231	0.095	0.308	0.952	1.038	1.561	1.249	0.219
SVM [34]	0.520	0.511	0.715	0.750	0.555	0.565	0.752	0.688
FraGAT [81]	0.491	0.408	0.638	0.773	0.502	**0.445**	**0.667**	0.747
Attentive FP [81]	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**



(A) 模型在训练集中的性能表现　　(B) 模型在测试集中的性能表现

图 3.9 基于 MAE、MSE 和 RMSE 评价指标的对比实验结果图

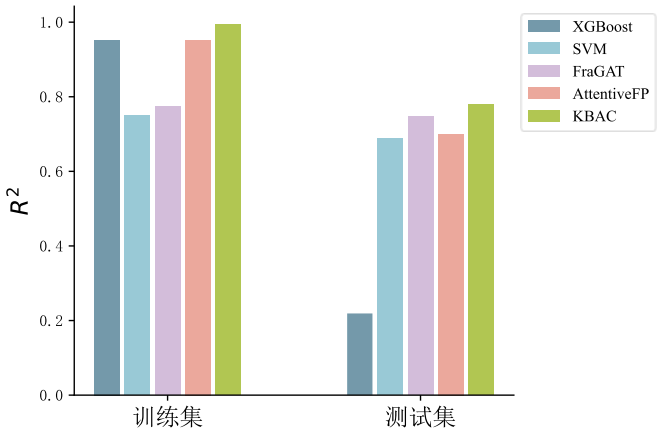


图 3.10 基于 R^2 评价指标的对比实验结果图

由表 3.8、图 3.9 和图 3.10 可知，本章所提 KBAC 方法在整体上优于其他比较方法。相比之下，XGBoost 的性能最差，而 Attentive FP 次之。具体而言，KBAC 在训练集和测试集上的 MAE 值在所有基准方法中均表现最佳，尤其是在训练集

上，分别比第二名（Attentive FP）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，比第二名的表现分别低了 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，比第二名的结果高了 0.041 和 0.079。说明本章所提方法在 pIC50 生物活性值的预测任务中表展现出了显著的性能优势。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，并通过融合多通道特征，来提取分子表征的全局和局部特征，使得模型能够更好的理解分子 SMILES 表达式所包含的化学和生物学性质。

3.6 本章小结

为应对单模态药物分子性质预测所面临的分子生物活性值预测不精、泛化性不高等问题，本章改进多通道注意力机制，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，用于抗乳腺癌药物分子的 pIC50 生物活性值预测。该网络采用了两阶段特征提取策略，所提模型即考虑了基于描述符方法、基于图方法和基于 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，从而实现了从整体到全局-局部的两次分子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的生物活性值。与其他基准模型相比，本章所提 KBAC 框架在 MAE、RMSE、MSE 和 R^2 评价指标上整体表现最佳，获得了高精度的预测结果。在训练集上，KBAC 模型的 MAE 可达 0.091，MSE 可达 0.014，RMSE 可达 0.117，R^2 可达 0.993；在测试集上，KBAC 模型的 MAE 可达 0.498，MSE 可达 0.471，RMSE 可达 0.686，R^2 可达 0.779。以上实验结果说明了 KBAC 模型在药物分子生物活性预测任务中的有效性。该框架不仅能够全面学习药物分子的全局-局部特征，还能综合理解多通道分子特征中包含的化学和生物学特性，为抗乳腺癌候选药物化合物的筛选与生物活性预测提供了帮助和支持。

第4章 基于多模态自适应对比融合深度网络的抗乳腺癌药物分子 ADMET 性质预测算法

上一章工作研究了基于抗乳腺癌候选药物分子 SMILES 单模态表征的生物活性预测算法，本章将在准确提取单模态分子特征的基础上进一步研究多模态表征策略下的分子性质预测模型。本章内容主要基于 ERα 数据集进行，利用药物分子 SMILES 表征和图结构表征，引入对比学习理论和多头注意力机制，构建抗乳腺癌药物分子 ADMET 性质深度学习预测算法（AMCFNet）。

4.1 引言

分子 ADMET 性质是决定药物化合物成药性的关键因素。在乳腺癌药物研发中，通过对药物分子的吸收、分布、代谢、排泄和毒性性质进行综合分析，研究人员能够深入了解候选药物的化学特性以及药物在体内的作用机制，从而最大限度地降低药物研发的失败风险。

近年来，基于计算机辅助的分子 ADMET 预测方法已成为药物研发和筛选的重要工具。其中，单模态方法在分子特征提取方面取得了显著进展，但因其仅针对特定分子表征进行分析，忽略了多模态数据间的协同和互补分子信息，从而限制了模型的泛化性能。而多模态方法通过简单拼接、加权融合、交叉注意力机制等融合技术，整合多视角的分子模态信息，能够从不同维度捕捉 ADMET 性质的复杂机制，从而更全面地理解药物分子的特征信息。如，分子 SMILES 表征提供分子的语义序列特征，分子描述符提供理化性质信息，二者结合可同时捕捉分子的序列模式与理化特性；分子指纹通过固定长度的二进制向量表征分子子结构特征，分子 Graph 数据通过节点和边表征分子的拓扑结构与化学键信息，二者结合可同时利用局部子结构特征与全局分子图信息。

然而，在分子性质预测领域，多模态融合技术仍面临诸多挑战。一方面，现有方法多局限于浅层的特征融合策略，如直接拼接或简单加权，难以有效建模跨模态分子信息间的复杂依赖关系和非线性交互模式；另一方面，不同模态数据通常具有不同的表示形式和统计特性，其异质性增加了特征对齐与融合的难度。

为应对上述问题，本章基于对比学习和自适应权重分配机制，提出了自适应多模态对比融合网络（AMCFNet），通过整合抗乳腺癌药物分子的 SMILES 和 graph

表征预测其 ADMET 性质。针对分子不同模态特征提取问题，构建基于知识先验 K-BERT 模块和多层 GNN 模块的双分支特征提取结构，以学习分子的序列语义和拓扑结构信息；针对多模态特征高效融合问题，设计自适应对比融合模块，以学习模态内与模态间的一致性和异质性，并基于相似性分数对多模态权重进行自适应分配；此外，整合分子的 1D-SMILES、2D-graph 以及融合的 SMILES-graph 特征，以形成多视角的分子互补特征，用于 ADMET 性质预测。通过这种方式，模型能够深入地理解并自适应地融合分子的语义与结构特性的关联和信息，从而有效缓解多模态数据间的异质性问题，以提升模型的预测性能和多模态融合能力。

4.2 数据来源与预处理

本章使用抗乳腺癌 ERα 数据集和五个 ChEMBL 数据集对药物分子 ADMET 性质预测进行研究，其中，分子 SMILES 表征和图结构表征将作为 AMDET 性质预测模型的自变量。

4.2.1 雌激素受体 α（ERα）数据集

本章主要工作仍基于 ERα数据集进行，在 KBAC 工作的基础上，针对抗乳腺癌药物分子的 ADMET 性质进行深入探索和研究。部分数据如表 4.1 所示。

表 4.1 ERα 数据集（ADMET 性质部分）示意表

SMILES	Caco-2	CYP3A4	hERG	HOB	MN
Oc1ccc2O[C@H]([C@H](Sc2…	0	1	1	0	0
Oc1ccc2O[C@H]([C@H](Sc2c…	0	1	1	0	0
Oc1ccc(cc1)[C@H]2Sc3cc(O)cc…	0	1	1	0	1
Oc1ccc2O[C@H]([C@@H](CC…	0	1	1	0	0
Oc1ccc2O[C@H]([C@H](Sc2c1…	0	1	1	0	1

ERα数据集包含了抗乳腺癌候选药物化合物的五种药代动力学性质和毒性：Caco-2、CYP3A4、hERG、HOB 和 MN。每种指标均采用 0-1 数值编码类别属性，其类别标签统计分布和具体指标含义如表 4.2 和表 4.3 所示。

表 4.2 ERα 数据集类别标签统计表

ADMET 指标	样本量	类别标签统计分布	
Caco-2	1974	Y=1：759（38.45%）	Y=0：1215（61.55%）
CYP3A4	1974	Y=1：1461（74.01%）	Y=0：513（25.99%）
hERG	1974	Y=1：1099（55.67%）	Y=0：875（44.33%）
HOB	1974	Y=1：509（25.79%）	Y=0：1465（74.21%）
MN	1974	Y=1：1514（76.70%）	Y=0：460（23.30%）

表 4.3 ERα 数据集 ADMET 指标含义表

ADMET 指标	指标含义
Caco-2	Y=1：化合物的小肠上皮细胞渗透性较好 Y=0：化合物的小肠上皮细胞渗透性较差
CYP3A4	Y=1：化合物能够被代谢 Y=0：化合物不能被代谢
hERG	Y=1：化合物具有心脏毒性 Y=0：化合物不具有心脏毒性
HOB	Y=1：化合物的口服生物利用度较好 Y=0：化合物的口服生物利用度较差
MN	Y=1：化合物具有遗传毒性 Y=0：化合物不具有遗传毒性

4.2.2 ChEMBL 数据集

为进一步验证所提预测模型的有效性和泛化性，本章引入五个与分子性质预测任务相关的公开基准数据集，具体包括：BBBP、BACE、Clintox、ESOL 和 FreeSolv。这些数据集涵盖多种下游任务，能够从不同角度评估模型的表现，其中 BBBP、BACE 和 Clintox 是分类任务数据集；ESOL 和 FreeSolv 是回归任务数据集。数据集的类别标签统计分布和具体指标含义如表 4.4 和表 4.5 所示。

表 4.4 ChEMBL 数据集类别标签统计表

数据集	样本量	类别标签统计分布	
BBBP	2050	Y=1：1567（76.44%）	Y=0：483（23.56%）
BACE	1513	Y=1：691（45.67%）	Y=0：822（54.33%）
Clintox	1484	Y=1：112（7.55%）	Y=0：1372（92.45%）
ESOL	1128	--	--
FreeSolv	642	--	--

表 4.5 ChEMBL 数据集指标含义表

数据集	下游任务	指标含义
BBBP	分类	Y=1：化合物具备血脑屏障渗透性 Y=0：化合物不具备血脑屏障渗透性
BACE	分类	Y=1：化合物具有抑制 β-分泌酶的作用 Y=0：化合物没有抑制 β-分泌酶的作用
Clintox	分类	Y=1：化合物在生物体内有毒性反应 Y=0：化合物在生物体内没有毒性反应
ESOL	回归	分子在水中的溶解度，标签值越高， 表示分子在水中的溶解度越好。
FreeSolv	回归	分子的自由溶解度，标签值越高， 表示分子在溶剂中的溶解度越好。

4.2.3 数据预处理

本小节采用的数据预处理方法，同 3.2.2 小节类似。具体而言，对每一个化合物分子，首先基于其标准 SMILES 表达式，利用 Python 中的 RDKit 工具包将每个分子的 SMILES 表征数量扩展为五个，包括一个标准的 SMILES 表征和四个增强的 SMILES 表征。当面对多模态预测任务时，则需进一步将每个分子的五个 SMILES 表征通过 RDKit 工具包转换为五个分子图结构数据，从而为每个分子化合物生成一组 SMILES-graph 数据对，其中，每组数据包含五个 SMILES-graph 数据对。这些分子图能够有效保留分子内部的结构信息和整体的拓扑结构，为模型增强多视角特征学习能力提供数据支持。最后，将数据集按照 8:1:1 的比例随机划分为训练集、验证集和测试集。

4.3 AMCFNet 模型设计

4.3.1 框架设计

本章基于对比学习理论和自适应权重分配机制，提出多模态自适应对比融合深度网络框架（AMCFNet），用于精准预测抗乳腺癌候选药物分子的 ADMET 性质。具体流程如图 4.1 所示。

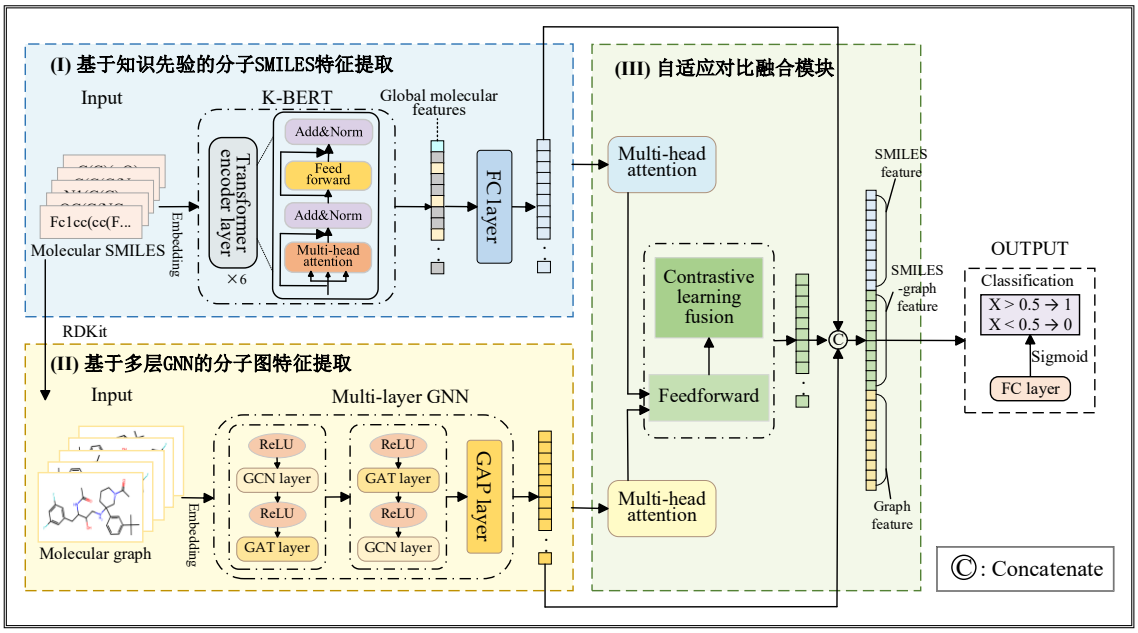


图 4.1 AMCFNet 整体框架图

如图 4.1 所示，AMCFNet 框架的构建主要包括以下部分：知识先验 K-BERT 模块、多层 GNN 模块以及自适应对比融合模块。具体而言，1）知识先验 K-BERT

模块旨在从一维 SMILES 表征中提取全局分子特征（1D-SMILES）；2）多层 GNN 模块则用于从二维分子图表征中提取分子图结构特征（2D-graph）；3）自适应对比融合模块：通过对比学习策略在模态内和模态间进行交互学习，并对分子特征进行自适应权重分配，实现特征的高效融合（SMILES-graph）。最后，分子的 1D-SMILES 特征、2D-graph 特征以及融合的 SMILES-graph 特征将被拼接，以形成多视角的分子互补特征，用于最终的 ADMET 性质预测任务。对应模块的详细信息介绍如下。

4.3.2 基于知识先验的分子 SMILES 特征提取

本小节引入 3.3.2 节提出的基于知识先验的 K-BERT 网络，用于提取分子一维 SMILES 表征中的全局特征，具体流程如图 1(I)所示。假设输入的分子 SMILES 表征为 $Input_S \in \mathbb{R}^{N\times 1}$，其中 N 为分子数量，则分子 SMILES 模态的嵌入机制可定义为：

$$X_{eS} = \mathrm{LN}(\mathrm{tok}_e(Input_S) + \mathrm{pos}_e(pos)),\ pos \in (0, l_S) \tag{4.1}$$

其中，$X_{eS} \in \mathbb{R}^{N\times l_g}$ 是通过嵌入机制得到的嵌入向量；l_S 是 SMILES 序列中词元的长度；$pos \in \mathbb{R}^{N\times l_S}$ 是 SMILES 序列中词元的位置信息；tok_e 是词元嵌入层；pos_e 是位置嵌入层；LN 是层归一化。

但如果直接通过 K-BERT 网络提取分子的 SMILES 特征可能会导致特定领域信息的丢失，并导致模型泛化性能降低。为捕获更多与分子 ADMET 性质预测任务相关的关键特征，本小节引入参数微调策略，在抗乳腺癌分子 ERα 数据集上进行迁移学习。具体训练过程定义如下：

$$X_S = \mathrm{FC}(\text{K-BERT}(X_{eS})) \tag{4.2}$$

$$\text{K-BERT}(X) = \mathrm{I}_{Tf}(\mathrm{F}_{Tf}(X) \circ 5) \tag{4.3}$$

其中，$X_S \in \mathbb{R}^{768\times 1}$ 表示通过 K-BERT 提取的分子 SMILES 特征；FC 表示全连接层；F_{Tf} 表示冻结 Transformer 编码器参数的操作；I_{Tf} 表示初始化 Transformer 编码器参数的操作；∘是操作符。具体而言，模块前五层 Transformer 编码器的参数将被冻结，以保留来自预训练任务的知识；其次，初始化最后一层 Transformer 编码器的参数，让网络在 ERα 数据集上从头开始训练。

本小节通过引入 K-BERT 网络和参数微调策略，使得模型在高效提取分子 1D-SMILES 语义特征的同时，能够显著减少其训练时间和计算资源消耗。

4.3.3 基于多层 GNN 的分子图特征提取

分子图结构数据同时包含了分子的局部和全局信息，使其具有高度的灵活性和多样性。为更有效地学习分子的图结构信息，本小节通过堆叠多个 GNN 网络层和动态调整节点之间的注意力权重，设计层级耦合架构的多层 GNN 网络，旨在实现分子图结构特征的层级化提取，从而捕捉不同输入节点和通道的局部与全局信息，以及分子结构中的连接特性和拓扑特征。具体而言，该模块由两个 GCN 层、两个 GAT 层、四个 ReLU 层和一个 GAP 层组成。多层 GNN 网络的具体流程如图 1（II）所示。假设输入的图结构数据表征为 $Input_g \in \mathbb{R}^{N\times 1}$，则分子图结构模态的嵌入机制可定义为：

$$X_{eg} = E_g(Input_g) \tag{4.4}$$

其中，E_g 是前向嵌入方法，用于获取初始节点信息和边信息 $X_{eg} \in \mathbb{R}^{N\times l_g}$，包括节点特征、边索引和边特征；$l_g$ 是节点特征的数量。为提升模型对邻近节点的差异化关注能力，本小节结合 GCN 与 GAT 算法构建层级耦合架构，以学习节点表征的深层结构信息，具体公式如下：

$$X_{g1} = \mathrm{GAT}(\mathrm{ReLU}(\mathrm{GCN}(\mathrm{ReLU}(X_{eg})))) \tag{4.5}$$

$$X_{g2} = \mathrm{GCN}(\mathrm{ReLU}(\mathrm{GAT}(\mathrm{ReLU}(X_{g1})))) \tag{4.6}$$

$$\mathrm{ReLU} = \max(0, X) \tag{4.7}$$

其中，$X_{g1} \in \mathbb{R}^{l_g\times 1024}$ 和 $X_{g2} \in \mathbb{R}^{l_g\times 1024}$ 表示分子的节点级特征；ReLU 表示非线性激活函数，用于捕捉数据中的复杂模式；GCN 表示图卷积网络层，用于聚合节点的邻居信息并学习每个节点的低维表示；GAT 表示图注意力网络层，每层包含八个注意力头，用于执行局部卷积操作。最后，通过 GAP 方法聚合分子图表征的整体特征信息，从而将节点级特征转换为图级特征 $X_g \in \mathbb{R}^{64\times 1}$：

$$X_g = \mathrm{GAP}(\mathrm{ReLU}(X_{g2})) \tag{4.8}$$

$$\mathrm{GAP} = \frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W} x_{i,j} \tag{4.9}$$

其中，H 为图结构特征的高度；W 为宽度；$x_{i,j}$ 为位置 (i, j) 处的特征值。

本小节设计的基于层级耦合架构的多层 GNN 网络，不仅实现了局部结构特征和全局拓扑信息的精确提取，还能有效聚合来自相邻节点的分子结构特征，使

其能够全面捕捉到药物分子的二维图结构（2D-graph）特征。

4.3.4 自适应对比融合模块

传统的特征拼接融合方法忽略了模态间的交互学习，而现有的大多数深度融合技术没有充分考虑模态间的语义区分和模态一致性，使其难以学习到不同模态间的深层关联信息。针对药物分子 1D-SMILES 序列与 2D-graph 特征的融合问题，本小节基于对比学习和自适应权重分配机制，设计自适应对比融合模块。该模块通过建立模态间的语义对齐机制，在保持模态特异性的同时，能够有效捕获跨模态的共性特征，充分考虑了模态之间的语义区分和模态一致性，从而克服多源异构数据表征不一，以及模型互补信息挖掘不足等难题。如图 4.2 所示，自适应对比融合模块包含三个主要的子模块：MHA 注意力子模块、前馈子模块以及对比学习融合子模块，通过多层次的模态交互实现特征的深度融合与语义对齐。

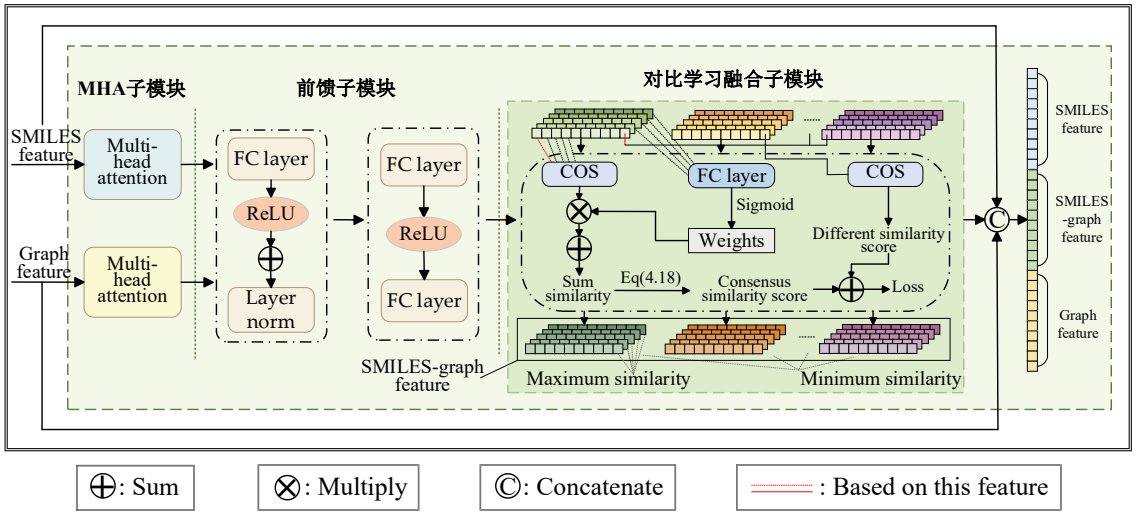


图 4.2 自适应对比融合模块框架图

（1）MHA 注意力子模块

将本章前述工作中，K-BERT 模块和多层 GNN 模块学习得到的分子 1D-SMILES 特征 $X_S \in \mathbb{R}^{768\times1}$ 和 2D-graph 特征 $X_g \in \mathbb{R}^{64\times1}$，作为该模块的输入。为充分挖掘这两种模态的特征表示，本小节为 $X_S \in \mathbb{R}^{768\times1}$ 和 $X_g \in \mathbb{R}^{64\times1}$ 分别引入一个多头注意力模块进行投影变换：

$$X_{SM} = \mathrm{MA}(X_S) \tag{4.10}$$

$$X_{gM} = \mathrm{MA}(X_g) \tag{4.11}$$

$$\mathrm{MHA}(Q,K,V)=\mathrm{Concat}(\mathrm{head}_1,\mathrm{head}_2,...,\mathrm{head}_h)\cdot W^O \tag{4.12}$$

$$\mathrm{head}_i=\mathrm{Attention}(QW_i^Q,KW_i^K,VW_i^V) \tag{4.13}$$

其中 MHA 是多头注意力机制，它利用可学习权重矩阵将 $X_S\in\mathbb{R}^{768\times1}$ 和 $X_g\in\mathbb{R}^{64\times1}$ 特征分别投影到 Q（查询）、K（键）和 V（值）向量空间；$Q_i=XW_i^Q$，$K_i=XW_i^K$，和 $V_i=XW_i^V$；其中 h 是注意力头的数量；W^O，W_i^Q，W_i^K 和 W_i^V 是可学习的矩阵。

本小节为 SMILES 和图结构模态数据设计的八个注意头的 MHA 子模块，能够并行捕捉输入特征中多样化的语义关系，实现对特征空间中不同位置信息的差异化关注，从而提升模型的特征提取能力。

（2）前馈子模块

考虑到不同模态间存在维度和信息分布差异，为更好地实现跨模态特征融合，本小节定义了一个前馈子模块，以 MHA 注意力子模块输出的投影特征 $X_{SM}\in\mathbb{R}^{768\times1}$ 和 $X_{gM}\in\mathbb{R}^{64\times1}$ 作为输入。具体细节如下：

$$X_{N(S-g)}=\mathrm{LN}(\mathrm{ReLU}(\mathrm{FC}(X_{SM}))+\mathrm{ReLU}(\mathrm{FC}(X_{gM}))) \tag{4.14}$$

其中 FC 是全连接层，用于减少模态自身的冗余信息，并对齐模态间的特征维度；ReLU 是非线性激活函数；LN 是层归一化，用于解决深度神经网络中的内部协变量偏移问题，确保每一层的输入具有稳定的均值和方差。具体公式如下：

$$X_{S-g}=\mathrm{FC}(\mathrm{ReLU}(\mathrm{FC}(X_{N(S-g)}))) \tag{4.15}$$

其中 $X_{S-g}\in\mathbb{R}^{128\times1}$ 表示通过两个 ReLU 函数和一个 FC 层得到的前馈特征。

本小节设计的前馈子模块，能够有效解决模态间的特征维度和数据分布差异，从而增强模型捕捉特征内部复杂关系的能力。

（3）对比学习融合子模块

每个模态都具有其独特的信息、性质和表达方式，仅通过简单的拼接融合方式来获取模态间的互补表示，往往导致最后的预测结果不太理想。因此，本小节基于相似性分数，设计了对比学习子模块，旨在实现多模态特征的有效交互和融合。由图 4.2 可以看出，对比学习子模块主要包含两个任务：共识相似性分数计算和差异相似性分数计算。

共识相似性分数旨在量化跨模态特征之间的相似性，强调不同模态特征在语义空间中的对齐程度和表示一致性。具体而言，该子模块首先基于分子的标准

SMILES 表示，通过余弦函数计算相同分子的相似度，其计算公式如下：

$$\cos(E_a, E_b) = \frac{F_a \cdot F_b}{\| F_a \| \ \| F_b \|} \tag{4.16}$$

其中，cos( ) 是余弦函数；F_a 和 F_b 是同一分子的不同表征；$\|\cdot\|$ 是欧几里得范数函数。随后，在子模块中引入 FC 层和 Sigmoid 函数，对分子的增强表示特征进行非线性变换，并将其结果作为初始特征权重值，通过动态权重分配机制，在后续的训练中自适应的更新权重参数，并学习同一分子多模态特征间的一致性。为评估同一分子不同模态的融合特征在共享语义空间中的对齐程度，在对比学习融合子模块中，通过将相似度和权重相乘，设计了加权相似性分数：。

$$Sum\_s = \sum_{d \in D_n} (\cos(F_{n,c}, F_{n,d}) \times \text{Sigmoid}(\text{FC}(F_{n,d}))) \tag{4.17}$$

其中，D_n 表示四种基于增强 SMILES 表示的融合特征；$F_{n,c}$ 表示当前分子的基于标准 SMILES 表示的融合特征；$F_{n,d}$ 表示当前分子的四种基于增强 SMILES 表示的融合特征之一。然后，通过公式（4.18）计算整体的共识相似性分数，以评估多模态表示之间的一致性程度。该评分函数能够为模型的训练提供指导，使其能够在复杂的多模态特征空间中学习到最具一致性的表示，并最小化模型因多模态特征存在噪声或不一致性带来的影响。

$$Consensus\_s = \sum_{n=1}^{N} \frac{4 - Sum\_S_n}{Sum\_S_n} \tag{4.18}$$

其中，$Consensus\_s$ 是一致性相似度评分；N 表示批量大小；n 表示当前分子。同时，基于每个标准 SMILES 通过公式 (4.19) 计算分子间的差异相似性分数。该评分函数通过对比同一分子融合表示与其他分子表示的不同之处，以学习模态特征间的多样性，从而鼓励模型保留部分来自每个模态的独特信息。

$$Different\_s = \sum_{n=1}^{N} \sum_{m \in B_n} \cos(F_{n,c}, F_{n,m}) \tag{4.19}$$

其中，$Different\_s$ 是差异相似性分数；m 表示该批次中其余基于标准 SMILES 表示的融合特征的分子之一；B_n 表示该批次中其余基于标准 SMILES 表示的融合特征的分子；$F_{n,m}$ 表示该批次中基于标准 SMILES 表示的融合特征。

通过上述对比学习过程，模型能够得到对比融合的多模态特征，即 SMILES-graph 融合特征。该模块的损失函数由共识相似性分数和差异相似性分数组成，其具体定义如下：

$$Loss_{CL} = Consensus\_s + Different\_s \tag{4.20}$$

其中，$Loss_{CL}$ 是对比学习损失函数，旨在最大化同一分子不同融合特征之间相似性的同时，最小化不同分子同一融合特征之间的相似性。

本章研究提出的 AMCFNet 网络，通过对比学习损失函数，能够捕捉模态之间的整体一致性及异质性。为实现这一目标，模型中使用了 Sigmoid 函数为相似性分数分配了自适应权重。通过自适应地学习不同模态之间的共享和独特特征，模型能够动态调整每个共识相似度分数的贡献，从而为相似度更高的特征分配更多的权重，使得模型最小化多模态特征存在一致性和异质性带来的影响。

最后，模型将 1D-SMILES（$X_S \in \mathbb{R}^{768\times1}$）特征、2D-graph（$X_g \in \mathbb{R}^{64\times1}$）特征和 SMILES-graph（$X_{F(S-g)} \in \mathbb{R}^{128\times1}$）融合特征进行拼接操作，以构建全面的多视角互补分子特征，即最终的多模态融合特征 $X_g \in \mathbb{R}^{960\times1}$，用于药物分子 ADMET 性质预测任务。

$$X_F = \text{Concat}(X_S, X_g, X_{F(S-g)}) \tag{4.21}$$

本章研究的目标是对抗乳腺癌候选药物分子的 ADMET 性质进行分类预测。因此，分类网络部分将采用二元交叉熵损失函数进行优化，其具体定义如下：

$$Loss_C = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(y_i) + (1-y_i)\log(1-y_i)] \tag{4.22}$$

其中，N 是分子的数量；log 是自然对数函数；y_i 是第 i 个分子的实际二元标签（0 或 1）；y_i 是第 i 个分子属于该类的预测概率。因此，对于整体 AMCFNet 网络框架，其总体损失函数可被定义为：

$$Loss_{Total} = Loss_{CL} + Loss_C \tag{4.23}$$

4.4 算法实现

基于多视角学习的理念，AMCFNet 将 1D-SMILES 特征、2D-graph 特征和融合后的 SMILES-graph 特征进行拼接，通过结合所有的语义和结构信息，形成最终的多视角互补融合特征，这些特征将通过 FC 层和 Sigmoid 函数，用于分子的 ADMET 特性预测。所提算法巧妙地结合了对比融合和多视图学习机制，通过模态内与模态间的交互学习，能够深入理解分子的多模态特征表示，从而实现更精确的预测。AMCFNet 框架的预测训练算法如表 4.6 所示。

表 4.6 AMFCNet 算法流程表

算法 **1　AMCFNet 训练步骤**

输入：药物分子 SMILES 表达式 X；药物分子 ADMET 标签 Y；药物分子样本量 N；训练批量大小 B。

（1）$i \leftarrow 0$

（2）$Input_S, Input_g \leftarrow \{\}$

（3）**while** S is not empty **do**

（4）　**for all** i from i to N **do**

（5）　　$Input_{Si}, Input_{gi} \leftarrow$ Data_Aug($SMILES_i$)　//数据增强

（6）　　$Input_S, Input_g \leftarrow Input_{Si}, Input_{gi}$

（7）　**end for**

（8）**end while**

（9）$Loss \leftarrow 0$

（10）**while** Not Converged **do**

（11）　**for all** each sample i in the batch **do**

（12）　　Compute X_S = K-BERT($Input_S$)　//利用 K-BERT 网络提取分子的一维 SMILES 语义特征

（13）　　Compute X_g = Multi-GNN($Input_g$)　//利用多层 GNN 网络提取分子的二维图结构特征

（14）　　Compute $X_{S\text{-}g}$ = Feedback(Fusion(MHA(X_S, X_g)))　//调整多模态特征的维度和信息分布差异

（15）　　Compute $X_{F(S\text{-}g)}$ = Contrastive learning($X_{S\text{-}g}$)　//通过对比学习策略得到 SMILES-graph 融合特征

（16）　　Compute contrastive learning loss by Equation (19)　//对比学习损失函数

（17）　　$Loss \leftarrow Loss + Loss_{CL}$

（18）　　$X_F \leftarrow$ Concatenate X_S, X_g, and $X_{F(S\text{-}g)}$　//构建多视角互补融合特征

（19）　　Predict molecular ADMET properties　//预测分子 ADMET 性质

（20）　　$Y \leftarrow$ Sigmoid(FC(X_F))

（21）　　Compute total loss by Equation (22) or (23)　//整体网络损失函数

（22）　　$Loss \leftarrow Loss + Loss_C$

（23）　**end for**

（24）**end while**

输出：药物分子 ADMET 预测标签 Y 。

4.5 实验结果及分析

本节将首先介绍评估方法和模型训练的实验设置。随后，通过消融实验、对比实验和泛化实验验证本章所提 AMCFNet 算法的有效性。所有实验均在 Nvidia4090 和 Google Colaboratory 云服务器上采用 Pytorch 深度学习框架实现。

4.5.1 评价指标

本章研究采用 ROC-AUC（Receiver Operating Characteristic-Area Under the

Curve）和准确率（Accuracy，ACC）指标评估所提 AMCFNet 模型以及基准模型在 ERα数据集上的性能表现。此外，在扩展验证环节，针对五个公开基准数据集的不同下游任务，分类任务采用 ROC-AUC 指标进行评估，而回归任务则使用 RMSE 指标进行评估。数据集及评价指标的具体细节见表 4.7。

表 4.7 数据集细节表

数据集	下游任务	主要评价指标
ERα_Caco-2	分类	ROC-AUC(%)/ACC
ERα_CYP3A4	分类	ROC-AUC(%)/ACC
ERα_hERG	分类	ROC-AUC(%)/ACC
ERα_HOB	分类	ROC-AUC(%)/ACC
ERα_MN	分类	ROC-AUC(%)/ACC
BACE	分类	ROC-AUC(%)
BBBP	分类	ROC-AUC(%)
Clintox	分类	ROC-AUC(%)
ESOL	回归	RMSE
FreeSolv	回归	RMSE

准确率表示模型正确分类的样本所占的比例，反映了模型在所有样本中做出正确预测的能力。在类别平衡的情况下，准确率能够很好地反映模型的性能。其具体计算如下：

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{4.24}$$

其中，TP（True Positive）是真阳性；FP（False Positive）是假阳性；TN（True Negative）是真阴性；FN（False Negative）是假阴性。ROC-AUC 指标基于模型的真阳性率（True Positive Rate, TPR）和假阳性率（False Positive Rate, FPR）计算得出：

$$TPR = \frac{TP}{TP + FN} \tag{4.25}$$

$$FPR = \frac{FP}{FP + TN} \tag{4.26}$$

ROC 曲线是模型在不同阈值下以 FPR 为横坐标，TPR 为纵坐标，绘制出来的曲线，比单一的准确率指标更加全面，能更好的适用于不平衡数据集的评估。AUC 则是 ROC 曲线下的面积，表示分类器的性能。AUC 的值范围在 0 到 1 之间，AUC 值越接近 1，表示分类器的性能越好。具体计算公式如下：

$$AUC = \int_0^1 TPR(FPR)dFPR \tag{4.27}$$

4.5.2 实验设置

本小节将详细描述模型训练时的参数设置。实验中，将基于 SMILES 的特征提取模块中的隐藏单元数设置为 768，注意力头数设置为 8，最大词元序列长度设置为 201。在基于图的特征提取模块中，将初始输入维度设置为 1，输出维度分别设置为 64、64、128 和 128，注意力头数设置为 8。在自适应对比融合模块中，将 FC 层的输出维度设置为 128。在后续的下游任务中，将 batch size、学习率和 epoch 最大迭代次数分别设置为 32、3e-5 和 100。本章研究使用了 Adam 优化器进行模型训练。此外，在训练过程中，还运用了 Dropout 技术、early stop 策略、MASK 策略和 Pos_weigh 参数来减轻过拟合并降低计算成本。具体参数设置如表 4.8。

表 4.8 训练参数表

Parameter Value	Value
Learning rate	3e-5
Batch size	32
dropout rate	0.5
Max epoch	100
Fusion dimension	128
Number of attention head	8
Optimizer	Adam
Patience	20

4.5.3 参数对比实验分析

为优化模型性能并确定最佳参数配置，本小节针对自适应对比融合模块中的关键参数进行了对比实验验证：多头注意力机制中的 head 数量（设置为 2、4、8）和特征通道维度（设置为 64、128、256）。实验结果如表 4.9 和图 4.3 所示。

表 4.9 融合模块中的参数对比实验表

Parameter		Dataset				
head	channel	Caco2	CYP3A4	hERG	HOB	MN
2	64	97.38	98.37	95.01	91.70	98.69
4	64	97.43	98.51	95.72	**92.25**	99.05
8	64	96.74	98.25	95.85	90.70	98.04
2	128	97.60	99.17	95.46	90.17	98.97
4	128	97.48	99.02	95.43	89.10	98.87
8	128	**97.88**	**99.25**	95.79	92.14	**99.48**
2	256	97.12	99.14	**96.48**	89.74	98.68
4	256	97.39	98.47	96.39	90.64	99.43
8	256	96.97	98.37	95.62	89.76	98.85

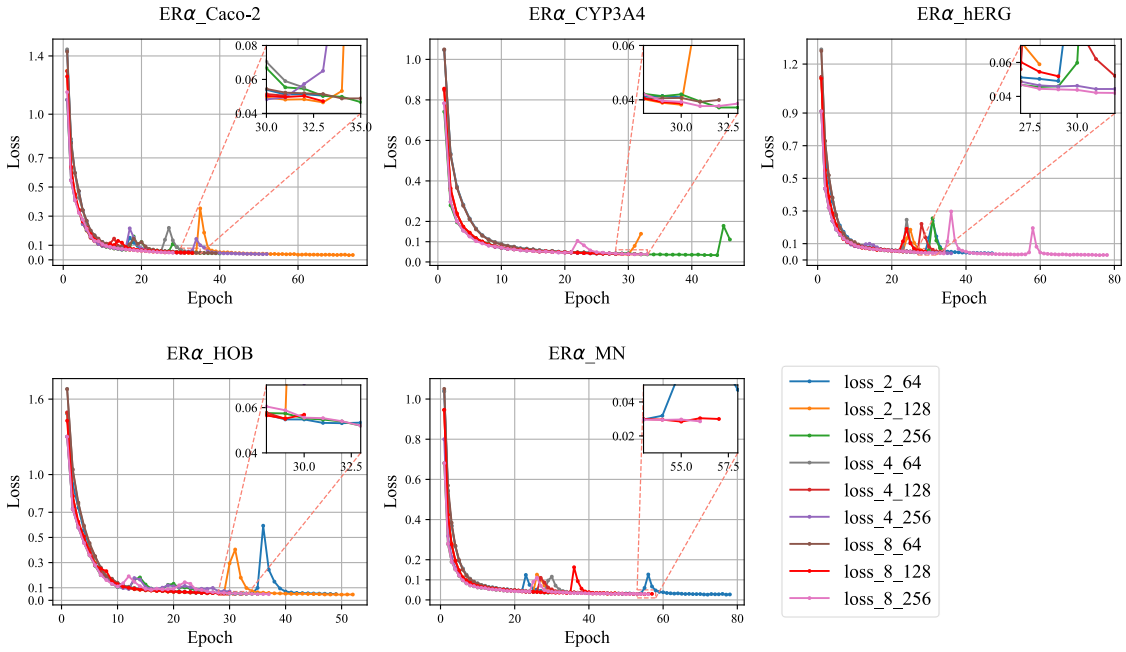


图 4. 3 不同参数组合的 loss-epoch 图

head 参数是自适应对比融合模块中多头注意力机制的参数，其取值直接影响模型的计算复杂度与表达能力之间的平衡。如表 4.9 所示，当 head 值从 2 增加到 8 时，五个分类任务的 ROC-AUC 值均有一定程度的提升。此外，通过分析图 4.3 中的 loss-epoch 曲线可以发现，当 head 值大于 2 时，损失函数的收敛过程更加稳定。因此，head=8 被确定为 AMCFNet 模型的最佳配置。

Channel 参数控制着多模态融合的维度，其作用是平衡一维 SMILES 表征和二维图结构表示之间的特征表达能力。如表 4.9 所示，当通道数从 64 增加到 128 时，ROC-AUC 值呈现显著提升。同时，图 4.3 中的 loss-epoch 曲线表明，当通道数为 128 时，模型展现出最佳的收敛特性。因此，channel=128 被确定为 AMCFNet 模型的最优设置。

基于对不同参数组合的系统性评估，综合考虑模型表达能力、损失函数收敛特性及训练效率等关键指标，实验结果表明：当 head=8 且 channel=128 时，所提 AMCFNet 算法能够实现最优的性能平衡。该参数配置不仅确保了模型具备充分的特征提取能力，同时保持了较高的计算效率，在训练过程中表现出稳定的收敛性和优异的预测精度，为后续的药物分子性质预测任务提供了可靠的模型基础。

4. 5. 4 消融实验分析

为全面验证本章提出的 AMCFNet 框架的有效性，本小节在 ERα数据集上设计并开展了两项消融实验，旨在深入分析不同模块和模态对模型性能的贡献度。

具体而言，消融实验分为两个主要部分：分子表征性能分析和自适应对比融合模块性能分析。其中，分子表征性能分析着重评估不同模态策略对预测结果的影响；而自适应对比融合模块性能分析则重点考察该模块在特征融合和表示学习中的作用。

（1）分子表征性能分析

在本小节中，AMCFNet 模型将被拆解为其基本组成部分：Transformer-SMILES 和 GNNs-graph（如第 4.3.2 和 4.3.3 节所述），分别用于分子 SMILES 表征和图结构表征的特征提取，详细结果如图 4.4 和表 4.10 所示。

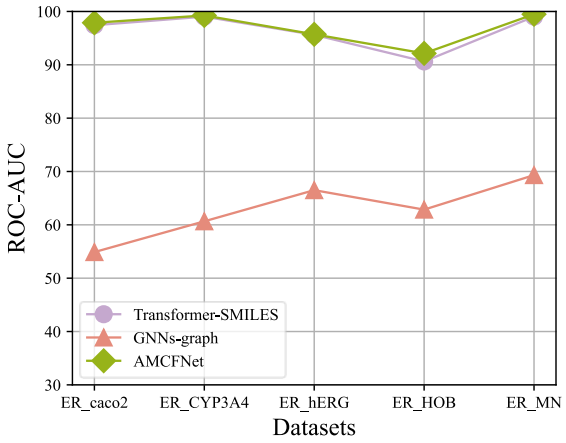


图 4.4 分子表征消融实验图

表 4.10 分子表征消融实验表

Method	SMILES	Graph	Caco2 ROC-AUC	CYP3A4 ROC-AUC	hERG ROC-AUC	HOB ROC-AUC	MN ROC-AUC
Transformer-SMILES	√	×	97.42	99.03	95.62	90.63	99.03
GNNs-graph	×	√	54.89	60.66	66.50	62.85	69.32
AMCFNet	√	√	**97.88**	**99.25**	**95.71**	**92.14**	**99.48**

图 4.4 和表 4.10 中的 Transformer-SMILES 指仅使用 SMILES 表征预测分子性质的方法，而 GNNs-graph 则表示仅使用图结构表征预测分子性质的模型。AMCFNet 是本章所提模型，能够自适应地融合分子的 SMILES 和 graph 表征。如图 4.4 和表 4.10 所示，AMCFNet 的整体性能显著优于其单一模态组件。具体而言，基于分子 SMILES 表征的 Transformer-SMILES 模型表现较为优异，在五个任务上的 ROC-AUC 值分别为 97.42%、99.03%、95.62%、90.63%和 99.03%。相比之下，基于分子图结构表征的模型表现则相对较差，其 ROC-AUC 值分别为 54.89%、60.66%、66.5%、62.85%和 69.32%，这种差异在一定程度上源于图数据的稀疏性特征。而 AMCFNet 模型通过自适应整合 SMILES 和 Graph 两种模态，使其在所

有任务上均实现了最佳性能，其 ROC-AUC 值在 Caco-2、CYP3A4、hERG、HOB 和 MN 五个分类任务上分别达到：97.88%、99.25%、95.71%、92.14%和 99.48%。这种性能提升主要得益于 AMCFNet 能够从多模态信息中提取互补的分子特征，并捕捉更精确的分子结构细节，从而显著提高了预测准确性。其实验结果充分验证了所提模型的有效性和优越性。

（2）自适应对比融合模块性能分析

本小节将评估 AMCFNet 框架及其变体在 ERα数据集上的性能表现，旨在通过消融实验深入分析不同模块对模型预测能力的影响。具体结果见图 4.5 和表 4.11。其中，AMCFNet-WF 表示移除融合子模块的变体模型，AMCFNet-WCL 表示移除对比学习子模块的变体模型，而 AMCFNet-WFCL 表示同时移除融合和对比学习子模块的变体模型。

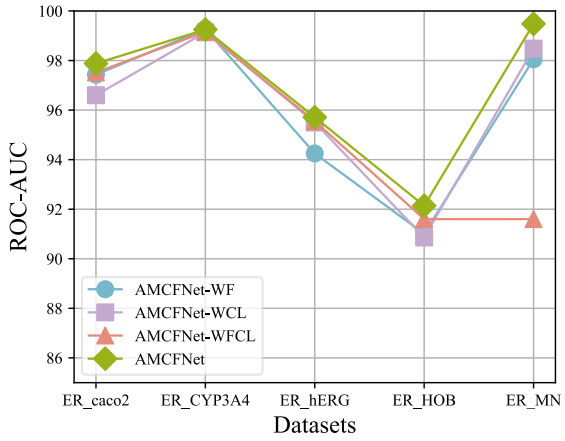


图 4.5 自适应对比融合模块消融实验图

表 4.11 自适应对比融合模块消融实验表

Method	Fusion	Contrastive learning	Caco-2 ROC-AUC	CYP3A4 ROC-AUC	hERG ROC-AUC	HOB ROC-AUC	MN ROC-AUC
AMCFNet-WF	×	√	97.43	**99.27**	94.25	91.01	98.05
AMCFNet-WCL	√	×	96.60	99.14	95.50	90.86	98.48
AMCFNet-WFCL	×	×	97.52	99.16	95.55	91.60	91.60
AMCFNet	√	√	**97.88**	99.25	**95.71**	**92.14**	**99.48**

融合子模块的有效性。融合子模块考虑了 SMILES 和 graph 表征之间的交互信息。当移除融合子模块时，其 AMCFNet-WF 变体模型在所有分子性质预测任务中的表现均较差。这表明融合子模块对预测准确性具有显著贡献，缺乏该模块会阻碍模型捕捉来自不同数据源的互补信息，从而说明有效整合不同分子模态信息的重要性。

对比学习子模块的有效性。对比学习子模块在理解分子表征之间的关联性上

起着重要作用。与 AMCFNet 相比，移除对比学习子模块的 AMCFNet-WCL 变体模型在某些性质预测任务上的表现略低，证明了该子模块在识别分子多模态特征间微妙关系的潜在价值，使得模型能够更好地捕捉分子特征的共性和差异。

融合和对比学习子模块的协同效应。融合子模块与对比学习子模块之间存在一定程度的协同效应。当 AMCFNet 同时移除这两个子模块时，其 AMCFNet-WFCL 变体模型在所有分子性质预测任务中的性能均出现明显下降。这一现象充分凸显了两个子模块的联合重要性。这些组件的缺失会对严重削弱模型捕捉分子多模态数据中互补特征和复杂关系的能力。只有当融合子模块和对比学习子模块协同工作时，完整的 AMCFNet 模型才能充分发挥其多模态学习优势，实现最佳的预测准确性。

本章所提算法在五个分子性质预测任务中的四个任务上实现了最佳性能，其实验结果强调了 AMCFNet 算法在联合理解不同分子模态方面的有效性，从而显著提升了对 ERα相关分子性质的预测能力。

4.5.5 对比实验分析

（1）AMCFNet 在 ERα 数据集上的性能

为评估本章所提 AMCFNet 算法的有效性，本小节引入四个最先进的分子性质预测模型作为基准，针对 ERα数据集中的五个性质预测任务进行全面评估。具体方法包括：XGBoost-ECFP4[79]、Attentive-FP[81]、FraGAT[80]和 ST+MLP（SMILES-Transformer + MLP）[82]。其中，XGBoost-ECFP4 是基于指纹的方法，Attentive-FP 和 FraGAT 是基于图的方法，而 ST+MLP 是基于 SMILES 的方法。为确保实验的公平性，每个方法均遵循相同的数据预处理步骤。本小节采用 ROC-AUC 和 ACC 两个评价指标分析模型表现。具体结果见图 4.6、图 4.7、表 4.12 和表 4.13。

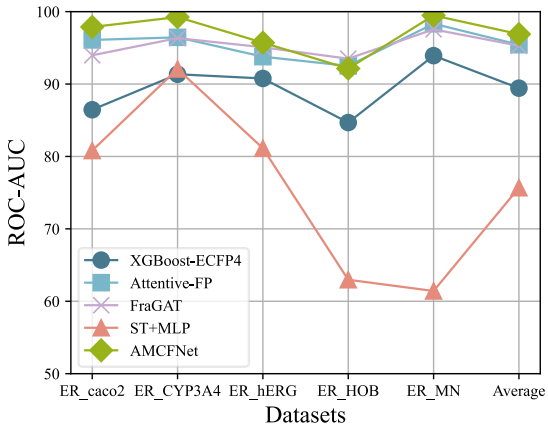


图 4.6 AMCFNet 及基准模型在 ROC-AUC 评价指标上的性能对比图

表 4.12 AMCFNet 及基准模型在 ROC-AUC 评价指标上的性能对比表

Method	Caco-2 ROC-AUC	CYP3A4 ROC-AUC	hERG ROC-AUC	HOB ROC-AUC	MN ROC-AUC	Average ROC-AUC
XGBoost-ECFP4[79]	86.45	91.35	90.77	84.68	93.92	89.43
Attentive-FP[81]	96.09	96.45	93.78	92.50	98.33	95.43
FraGAT[80]	93.97	96.34	95.08	**93.50**	97.44	95.27
ST+MLP[82]	80.81	92.08	81.13	62.97	61.42	75.68
AMCFNet	**97.88**	**99.25**	**95.71**	92.14	**99.48**	**96.89**

图 4.6 和表 4.12 展示了本章所提模型与基准模型，以 ROC-AUC 为评价指标的对比实验结果，表中黑体数字表示最优值。由图 4.6 和表 4.12 可以看出，AMCFNet 在 ERα 数据集的分子 ADMET 性质分类任务中具有最佳性能。具体而言，在五个预测任务中，AMCFNet 在四个任务上取得了最优结果，平均 ROC-AUC 值达到 96.89%。并且，在表现最差的 ERα_hERG 预测任务中，AMCFNet 也取得了优异结果（ROC-AUC 值达到 92.14%），而在 ERα_MN 任务中的 ROC-AUC 值则更是接近 100%。这种显著的性能优势主要是因为基准方法通常只通过特定网络直接学习单一模态的分子特征，导致了重要信息的丢失。相比之下，AMCFNet 通过自适应融合 1D-SMILES 和 2D-graph 表征，能够更全面地捕捉分子的多视图特征，从而显著提升了预测准确性。

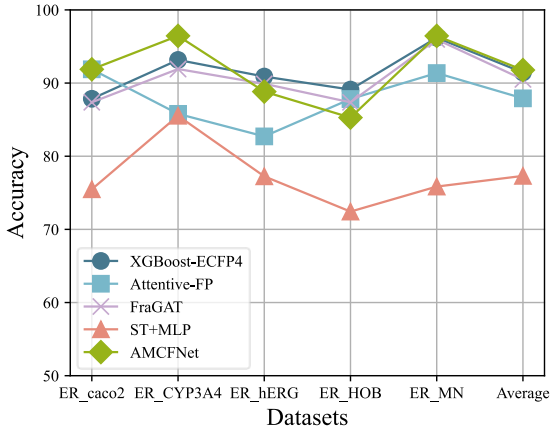


图 4.7 AMCFNet 及基准模型在 Accuracy 评价指标上的性能对比图

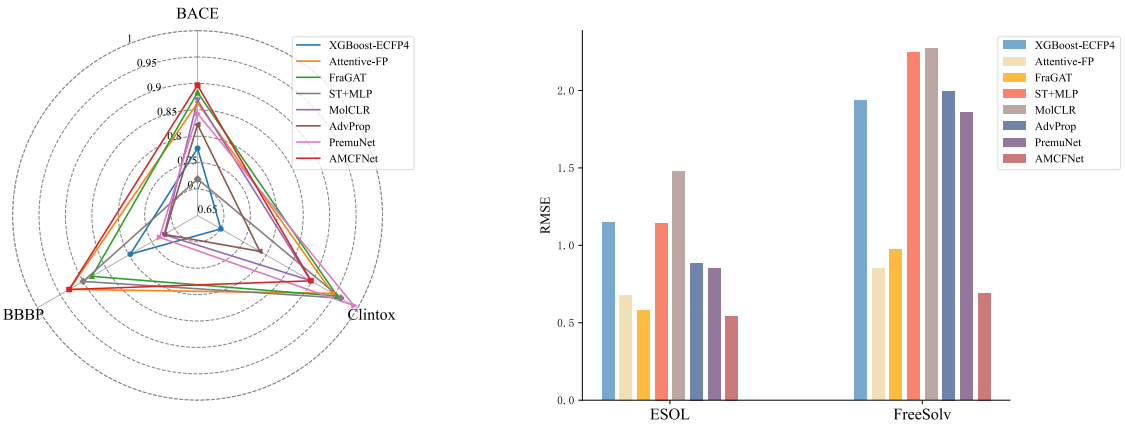
表 4.13 AMCFNet 及基准模型在 Accuracy 评价指标上的性能对比表

Method	Caco-2 Accuracy	CYP3A4 Accuracy	hERG Accuracy	HOB Accuracy	MN Accuracy	Average Accuracy
XGBoost-ECFP4[79]	87.85	93.16	**90.89**	**89.11**	96.20	91.44
Attentive-FP[81]	91.88	85.79	82.71	87.82	91.37	87.91
FraGAT[80]	87.37	91.92	89.90	87.37	95.96	90.50
ST+MLP[82]	75.48	85.53	77.25	72.43	75.85	77.31
AMCFNet	**91.88**	**96.45**	88.83	85.28	**96.45**	**91.78**

为全面评估 AMCFNet 模型的性能，本小节引入准确率作为补充评价指标，具体结果如图 4.7 和表 4.13 所示。实验结果表明，AMCFNet 在 ERα 数据集的分子 ADMET 性质分类中表现最优，其平均准确率达到 91.78%。与表现第二的 XGBoost-ECFP4 方法相比，AMCFNet 在准确率上提升了 0.34%。此外，AMCFNet 在 ERα_Caco-2、ERα_CYP3A4 和 ERα_MN 任务中的表现显著优于所有基准方法，充分证明了其优越性。然而，在 ERα_hERG 和 ERα_HOB 任务中，AMCFNet 的表现则相对欠佳。这种差异主要是因为图数据本身的稀疏性，使其在不同测试环境中导致一定程度的模型性能波动。这一发现促使我们探索更有效的策略来优化 GNN 模块，例如通过图数据增强或改进图表示学习方法来缓解数据稀疏性问题，从而进一步提高 AMCFNet 的鲁棒性和泛化能力。

（2） AMCFNet 在 ChEMBL 数据集上的性能

为进一步验证 AMCFNet 算法在分子性质预测中的有效性和泛化能力，本小节将在五个与分子性质预测相关的公开数据集上进行实验，具体数据集包括 BACE、BBBP、Clintox、ESOL 和 FreeSolv，这些数据集涵盖了三个分类任务和两个回归任务。其中，ROC-AUC 指标用于评估分类任务，RMSE 指标用于评估回归任务。具体实验结果见图 4.8 和表 4.14 所示。图 4.8（A）和图 4.8（B）分别描绘了不同模型在分类任务和回归任务中的性能表现，表 4.14 则展示了 AMCFNet 以及基准方法在 ChEMBL 数据集上的实验结果。针对每个数据集，本小节分别使用当前最先进的基于图的算法（Attentive-FP、FraGAT 和 MolCLR[83]）、基于指纹的算法（XGBoost-ECFP4）、基于 SMILES 的算法（ST+MLP）和多模态融合算法（AdvProp[84]和 PremuNet[68]）进行对比实验分析。



（A）模型在分类任务中的性能表现　　（B）模型在回归任务中的性能表现

图 4.8 AMCFNet 及基准模型在 ChEMBL 数据集上的性能对比图

表 4. 14 AMCFNet 及基准模型在 ChEMBL 数据集上的性能对比表

Method	BACE ROC-AUC	BBBP ROC-AUC	Clintox ROC-AUC	ESOL RMSE	FreeSolv RMSE
XGBoost-ECFP4[79]	77.70	79.70	70.10	1.146	1.934
Attentive-FP[81]	86.30	93.01	94.54	0.679	0.850
FraGAT[80]	88.28	88.08	95.54	0.580	0.973
ST+MLP[82]	71.90	90.00	96.30	1.144	2.246
MolCLR[83]	86.80	72.30	89.70	1.477	2.273
AdvProp[84]	82.20	72.20	78.60	0.886	1.996
PremuNet[68]	84.30	73.30	**99.20**	1.858	0.851
AMCFNet	**89.65**	**93.08**	89.78	**0.545**	**0.691**

由图 4.8 和表 4.14 可以看出，AMCFNet 框架在分子性质预测任务中展现出显著的结果优势。与当前表现优异的 AdvProp 和 PremuNet 多模态融合模型相比，本章研究提出的 AMCFNet 框架在 BACE、BBBP、ESOL 和 FreeSolv 数据集上均取得了最佳性能。对于分类预测任务，其具体结果见表 4.12 的第 1-3 列和图 4.8（A）。AMCFNet 模型在 BACE 和 BBBP 数据集中的 ROC-AUC 值分别达到了 89.65%和 93.08%，相较于其他基线方法有显著提升。然而，在 Clintox 数据集中，由于数据存在严重的类别不平衡问题（原始测试集中 95.75%的数据类别为 0，4.25%的数据类别为 1），导致模型表现欠佳，其 ROC-AUC 为 89.78%。但如果将类别标签按照均匀分配策略重新分配至训练集、验证集和测试集（调整后三个数据集中 91.97%的数据为类别 0，8.03%的数据为类别 1），AMCFNet 在该数据集上的 ROC-AUC 值可提升至 94.6%，与次优模型的表现相当。这一结果体现了数据不平衡问题对所提模型性能的显著影响，在后续工作中，我们将继续优化算法，以解决此类问题。

对于回归预测任务，其具体结果见表 4.14 的第 5 列和第 6 列，以及图 4.8（B）。AMCFNet 在 ESOL 和 FreeSolv 数据集上均表现最优，其 RMSE 值分别为 0.545 和 0.691，说明了所提模型具有更高的预测准确性。

AMCFNet 在五个基准数据集中的四个数据集上均表现出明显的性能优势。主要是因为 AdvProp 和 PremuNet 模型只采用简单拼接方法直接整合多模态特征，忽略了不同模态间多视角表示的相互作用。而本章所提 AMCFNet 算法通过对比学习策略和相似度评分函数对模型进行训练和约束，使其能够有效识别模态内与模态间的关联信息，从而实现多模态特征的自适应融合。

实验结果充分证明了 AMCFNet 算法在分子性质预测中的有效性和泛化能力，展示了其在多种化学任务中的优越表现。所提模型的鲁棒性不仅体现在乳腺癌

ERα数据集的 ADMET 性质预测上，还体现在对其他数据集分子性质的可靠预测能力上。此外，这些结果进一步凸显了 AMCFNet 算法通过自适应对比学习融合技术学习多视图分子信息的能力，为复杂分子性质预测提供了新的解决方案。

4.6 本章小结

为应对多模态药物分子性质预测所存在的多模态数据表征不一致、互补信息挖掘不足等问题，本章提出基于对比学习和自适应权重分配机制的多模态自适应对比融合深度网络（AMCFNet），用于抗乳腺癌药物分子的 ADMET 性质预测。该网络首先构建双分支特征提取结构，引入知识先验 K-BERT 模块和多层 GNN 模块，分别学习分子的一维序列信息和二维图结构特征；此外，基于对比学习策略，设计自适应对比融合模块，进行分子模态间和模态内的交互学习，通过自适应分配权重，融合分子的 1D-SMILES 和 2D-graph 的特征；最后，整合 1D-SMILES、2D-graph 和融合的 SMILES-graph 跨模态特征，以生成多视角的分子互补特征，用于分子 ADMET 性质的预测。所提出的 AMCFNet 模型在相关数据集与最先进的基准方法的测试中，均实现了优异的性能表现。在 ERα数据集上，AMCFNet 的平均 ROC-AUC 值高达 96.89%，准确率可达 91.78%。此外，在五个 ChEMBL 数据集中，AMCFNet 也展现了出色的性能，在 BACE、BBBP 和 Clintox 数据集的分类任务中，分别实现了 89.65%、93.08%和 89.78%的 ROC-AUC 值，在 ESOL 和 FreeSolv 数据集的回归任务中，分别实现了 0.545 和 0.691 的 RMSE 值。实验结果表明，AMCFNet 模型能有效地学习和融合分子多模态特征，从而提升抗乳腺癌候选药物分子 ADMET 性质预测的准确性。

第5章 总结与展望

5.1 本文总结

乳腺癌作为女性健康的首要威胁，其高发病率和高死亡率给医疗系统带来了巨大挑战。随着精准医疗的发展，乳腺癌治疗正逐步从传统模式转向个性化治疗，这对药物研发的效率与精准性提出了更高要求。然而，传统药物研发模式受限于周期长、成本高、成功率低等问题，难以满足快速变化的临床需求。近年来，人工智能技术的快速发展为药物研发带来了新的机遇，尤其是机器学习和深度学习算法在药物分子性质预测中的应用，显著提升了药物化合物的筛选效率与成功率，为抗乳腺癌药物的研发提供了新的突破口。

本文针对药物分子性质预测精度提升的关键问题，从药物分子的单模态和多模态表征训练策略出发，提出两种基于深度学习的分子性质预测网络，旨在为抗乳腺癌药物研发提供更高效、精准的计算工具。本文的主要研究成果总结如下：

（1）本研究提出了基于知识先验和多通道注意力的生物活性预测算法。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 Knowledge-BERT-1D-ECA-CNN（KBAC）深度神经网络，实现 pIC50 生物活性值回归预测。实验结果表明，所提出的 KBAC 框架在四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，相对于四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，其两阶段特征提取策略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。

（2）本研究提出了基于多模态自适应对比融合深度网络的分子 ADMET 性质预测算法。该网络采用分层设计思想，首先构建双分支特征提取模块，引入 K-BERT 模块提取分子的 1D-SMILES 序列的语义特征，捕获分子功能团和化学键的序列信息；同时，设计多层 GNN 模块学习药物分子的 2D-Graph 表示，有效建

立原子的空间连接关系和拓扑结构。其次，设计自适应对比融合模块，动态学习模态内与模态间的一致性和异质性，并根据对比学习结果自适应地分配权重，深度融合药物分子的语义和结构信息。最后，模型整合 1D-SMILES 特征、2D-Graph 特征以及融合的 SMILES-graph 跨模态特征，构建多视角的分子互补融合表示，用于分子 ADMET 性质分类预测。所提出的 AMCFNet 模型在 ERα数据集上表现出色，平均 ROC-AUC 值达到 96.89%，准确率高达 91.78%。此外，在五个 ChEMBL 数据集的分类和回归任务中，AMCFNet 也展现了卓越的性能。实验结果表明，AMCFNet 在分子性质预测任务中显著优于现有的最先进基线模型，充分验证了所提模型的有效性。其自适应对比融合策略能够有效学习分子多模态间的交互信息，从而显著提升分子性质预测的精度。

5.2 不足与展望

尽管本研究在现有研究基础上取得了重要突破，但仍存在一定的改进空间。为进一步提升分子性质预测的性能，未来研究可从以下几个方面展开深入探索：

（1）本研究所提的多模态预测网络仅针对分子级别的多维表示（如一维序列信息、二维拓扑结构以及三维空间构象）的融合和学习，未考虑其他与药物发现相关的重要临床信息，如基因和病理数据。未来的工作将探索更有效且具有广泛适应性的多模态融合方法，将药物分子与病理学和基因特征结合起来，以分析药物治疗研究中的潜在机制，推动其在生物学和医学领域中的应用。

（2）本研究所提模型容易受到数据类别分布不均和数据稀疏性的影响，从而限制了模型在处理此类情况时的特征信息提取能力和泛化能力。未来的工作将进一步探索基于自监督的图神经网络或基于代价敏感学习的方法，通过预训练策略以及损失函数权重调整，优化模型在处理数据类别不平衡和数据稀疏性方面的能力，为分子性质预测提供更可靠的技术支持。

（3）深度学习的“黑匣子”性质和分子性质的不可知性，使得基于深度学习的分子性质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，重点探讨多模态数据间的关联机制和可视化分析方法，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

[1] Łukasiewicz S, Czeczewski M, Forma A, et al. Breast cancer-epidemiology, risk factors, classification, prognostic markers, and current treatment strategies-an updated review[J]. Cancers, 2021, 13(17): 4287.

[2] Vy V P T, Yao M M S, Le N Q K, et al. Machine learning algorithm for distinguishing ductal carcinoma in situ from invasive breast cancer[J]. Cancers, 2022, 14(10): 2437.

[3] Bray F, Laversanne M, Sung H, et al. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries[J]. CA: A Cancer Journal for Clinicians, 2024, 74(3): 229-263.

[4] Barzaman K, Karami J, Zarei Z, et al. Breast cancer: Biology, biomarkers, and treatments[J]. International Immunopharmacology, 2020, 84: 106535.

[5] 花蕊,朱家明.基于深度神经网络对抗乳腺癌候选药物ERα生物活性的预测[J].陕西理工大学学报(自然科学版),2023,39(02):47-53.

[6] 尚雅欣,雷小洁,方子牛,等.基于GA-BP神经网络模型的抗乳腺癌候选药物活性预测[J].数学理论与应用,2024,44(02):103-125.

[7] 秦雅琴,夏玉兰,卢梦媛,等.抗乳腺癌活性化合物的 ADMET 性质预测模型[J].云南大学学报(自然科学版),2022,44(06):1127-1134.

[8] 张襄松,高秀秀.基于随机森林的逻辑回归预测抗乳腺癌药物的ADMET性质[J].内蒙古工业大学学报(自然科学版),2023,42(06):481-487.

[9] Ferreira L L G, Andricopulo A D. ADMET modeling approaches in drug discovery[J]. Drug Discovery Today, 2019, 24(5): 1157-1165.

[10] Sheridan R P, Culberson J C, Joshi E, et al. Prediction accuracy of production ADMET models as a function of version: Activity cliffs rule[J]. Journal of Chemical Information and Modeling, 2022, 62(14): 3275-3280.

[11] 濮澄韬,顾灵茜,陈兴晔,等.基于人工智能的药物人体肠道吸收性质预测[J].中国药科大学学报,2023,54(03):355-362.

[12] 李俊毅,李文青,高升华,等.计算机辅助药物设计在抗病毒药物研究中的应用[J].中国药物化学杂志,2023,33(12):930-947.

[13] 陈柏宇,吕泸楠,徐小迪,等.AIDD与CADD提升药物成功率的思考[J].中国药科大学学报,2024,55(03):284-294.

[14] 左玲,李煜桐,向玲宝,等.计算机辅助药物设计和实验验证探讨柴胡皂苷D诱导膀胱癌细胞凋亡的分子机制[J].中国实验方剂学杂志,2024,30(17):87-94.

[15] 顾耀文,张博文,郑思,等.基于图注意力网络的药物ADMET分类预测模型构建方法[J].数据分析与知识发现,2021,5(08):76-85.

[16] 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, 2019, 60(1): 22-28.

[17] Han S, Fu H T, Wu Y Y, et al. HimGNN: a novel hierarchical molecular graph representation learning framework for property prediction[J]. Briefings in Bioinformatics, 2023, 24(5): bbad305.

[18] Gong X, Liu M T, Liu Q, et al. MDFCL: Multimodal data fusion-based graph contrastive learning framework for molecular property prediction[J]. Pattern Recognition, 2025: 111463.

[19] Xia J, Zhao C S, Hu B Z, et al. Mole-bert: Rethinking pre-training graph neural networks for molecules[J]. 2023.

[20] Zaremba W, Sutskever I, Vinyals O. Recurrent neural network regularization[J]. arXiv preprint arXiv:1409.2329, 2014.

[21] Scarselli F, Gori M, Tsoi A C, et al. The graph neural network model[J]. IEEE transactions on neural networks, 2008, 20(1): 61-80.

[22] 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.

[23] Song X Y, Wen G Y, Chai L. Graph signal processing based nonlinear QSAR/QSPR model learning for compounds[J]. Biomedical Signal Processing and Control, 2024, 91: 106011.

[24] Wu X R, Gong J X, Ren S Y, et al. A machine learning-based QSAR model reveals important molecular features for understanding the potential inhibition mechanism of ionic liquids to acetylcholinesterase[J]. Science of The Total Environment, 2024, 915: 169974.

[25] Zhao Y K, Mulder R J, Houshyar S, et al. A review on the application of molecular descriptors and machine learning in polymer design[J]. Polymer Chemistry, 2023, 14(29): 3325-3346.

[26] Ji B H, Wu Y H, Thomas E N, et al. Predicting anti-SARS-CoV-2 activities of chemical compounds using machine learning models[J]. Artificial Intelligence Chemistry, 2023, 1(2): 100029.

[27] 李文静,齐飞宇,霍晓乾,等.基于化学成分群加和性分子描述符的中药浸膏粉溶化性预测研究[J].中草药,2022,53(22):7029-7038.

[28] 秦传东,廖奥林.基于PSO-BP的抗乳腺癌药物毒性研究[J].计算机仿真,2024,41(04):320-324.

[29] 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.

[30] 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.

[31] Shi L H, Yan F, Liu H H. Screening model of candidate drugs for breast cancer based on ensemble learning algorithm and molecular descriptor[J]. Expert Systems with Applications, 2023, 213: 119185.

[32] 何冰,罗勇,李秉轲,等.基于分子描述符和机器学习方法预测和虚拟筛选乳腺癌靶向蛋白HEC1抑制剂[J].物理化学学报,2015,31(09):1795-1802.

[33] 付洺宇,董逸潇,吴春勇,等.基于机器学习的化合物血脑屏障通透性预测[J].中国药学杂志,2021,56(20):1677-1683.

[34] Hearst M A, Dumais S T, Osuna E, et al. Support vector machines[J]. IEEE Intelligent Systems and Their Applications, 1998, 13(4): 18-28.

[35] Breiman L. Random forests[J]. Machine Learning, 2001, 45: 5-32.

[36] Venkatraman V. FP-ADMET: a compendium of fingerprint-based ADMET prediction models[J]. Journal of cheminformatics,2021,13(1):75-75.

[37] 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, 2020, 18(5): 1902-1913.

[38] Teng S S, Yin C L, 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: 106904.

[39] Ding W Z, Nan Y, Wu J S, et al. Combining multi-dimensional molecular fingerprints to predict the hERG cardiotoxicity of compounds[J]. Computers in Biology and Medicine, 2022, 144: 105390.

[40] 卢昂,陈壮志,巫秀美,等.基于分子指纹对黄酮类化合物及其抗DPPH自由基活性的QSAR

分析[J].化学通报,2022,85(10):1261-1266.

[41] 于亚运,刘勇国,蒋羽,等.基于指纹相似度的药物-靶点相互作用预测[J].中国中药杂志,2017,42(18):3578-3583.

[42] 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.

[43] Wang S, Guo Y Z, Wang Y H, 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. 2019: 429-436.

[44] 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 Receptors and Signal Transduction, 2022, 42(4): 361-372.

[45] Monteiro N R C, Pereira T O, Machado A C D, et al. FSM-DDTR: End-to-end feedback strategy for multi-objective de novo drug design using transformers[J]. Computers in Biology and Medicine, 2023, 164: 107285.

[46] Li H, Zhao D, Zeng J Y. KPGT: knowledge-guided pre-training of graph transformer for molecular property prediction[C]. Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. 2022: 857-867.

[47] Liu Y W, Zhang R S, Li T F, et al. MolRope-BERT: An enhanced molecular representation with rotary position embedding for molecular property prediction[J]. Journal of Molecular Graphics and Modelling, 2023, 118: 108344.

[48] Zhang X C, Wu C K, Yi J C, et al. Pushing the boundaries of molecular property prediction for drug discovery with multitask learning BERT enhanced by SMILES enumeration[J]. Research, 2022: 0004.

[49] Chen Y L, Zhu Y, Zhang Z T, et al. Prediction of drug protein interactions based on variable scale characteristic pyramid convolution network[J]. Methods, 2023, 211: 42-47.

[50] 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.

[51] Hua Y, Song X N, Feng Z H, et al. MFR-DTA: a multi-functional and robust model for predicting drug–target binding affinity and region[J]. Bioinformatics, 2023, 39(2): 056-064.

[52] Shao J S, Gong Q N, Yin Z Y, 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.

[53] Zhao Q C, Duan G H, Yang M Y, 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.

[54] Zhu Z Q, 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.

[55] Wang Y Y, Wang J R, Cao Z L, et al. Molecular contrastive learning of representations via graph neural networks[J]. Nature Machine Intelligence, 2022, 4(3): 279-287.

[56] Yang Z D, Zhong W H, 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.

[57] Kipf T N, Welling M. Semi-supervised classification with graph convolutional networks[J]. arXiv preprint arXiv:1609.02907, 2016.

[58] Velickovic P, Cucurull G, Casanova A, et al. Graph attention networks[J]. stat, 2017, 1050(20): 10-48550.

[59] Xu K Y L, Hu W H, Leskovec J, et al. How powerful are graph neural networks?[J]. arXiv preprint arXiv:1810.00826, 2018.

[60] Li P Y, Wang J, Qiao Y X, 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.

[61] Yu Z N, Gao H Y. Molecular representation learning via heterogeneous motif graph neural networks[C]. International Conference on Machine Learning. PMLR, 2022: 25581-25594.

[62] Xia X Q, Zhu C Y, 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.

[63] Li Z M, Zhu S C, 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.

[64] Chen Z D, Li D Z, Liu M H, et al. Graph neural networks with molecular segmentation for property prediction and structure–property relationship discovery[J]. Computers & Chemical Engineering, 2023, 179: 108403.

[65] Wang T Y, Sun J Q, 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.

[66] Yuan W N, Chen G X, 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.

[67] Liu S C, Wang H C, Liu W Y, et al. Pre-training molecular graph representation with 3D geometry[J]. arXiv preprint arXiv:2110.07728, 2021.

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

[69] Wu T Y, Tang Y, Sun Q Y, et al. Molecular joint representation learning via multi-modal information of SMILES and graphs[J]. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2023.

[70] Lu X H, Xie L X, Xu L, et al. Multimodal fused deep learning for drug property prediction: Integrating chemical language and molecular graph[J]. Computational and Structural Biotechnology Journal, 2024, 23: 1666–1679.

[71] Wang Z Y, Liu M, Luo Y Z, et al. Advanced graph and sequence neural networks for molecular property prediction and drug discovery[J]. Bioinformatics, 2022, 38(9): 2579-2586.

[72] Thakur A, Kumar A, Sharma V, et al. PIC50: An open source tool for interconversion of PIC50 values and IC50 for efficient data representation and analysis[J]. BioRxiv, 2022: 2022.10. 15.512366.

[73] Wang Q L, Wu B G, Zhu P F, et al. ECA-Net: Efficient channel attention for deep convolutional neural networks[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 11534-11542.

[74] Hu J, Shen L, Sun G. Squeeze-and-excitation networks[C]. Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 7132-7141.

[75] Woo S, Park J, Lee J Y, et al. Cbam: Convolutional block attention module[C]. Proceedings of the European conference on computer vision (ECCV). 2018: 3-19.

[76] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.

[77] Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[C]. 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). 2019: 4171-4186.

[78] Wu Z X, Jiang D J, Wang J K, et al. Knowledge-based BERT: a method to extract molecular features like computational chemists[J]. Briefings in Bioinformatics, 2022, 23(3): 131-143.

[79] Chen T Q, Guestrin C. Xgboost: A scalable tree boosting system[C]. Proceedings of the 22nd ACM Sigkdd International Conference on Knowledge Discovery and Data Mining. 2016: 785-794.

[80] Zhang Z Q, Guan J H, Zhou S G. FraGAT: a fragment-oriented multi-scale graph attention model for molecular property prediction[J]. Bioinformatics, 2021, 37(18): 2981-2987.

[81] Xiong Z P, Wang D Y, Liu X H, et al. Pushing the boundaries of molecular representation for drug discovery with the graph attention mechanism[J]. Journal of Medicinal Chemistry, 2019, 63(16): 8749-8760.

[82] Honda S, Shi S, Ueda H R. SMILES transformer: Pre-trained molecular fingerprint for low data drug discovery[J]. arXiv preprint arXiv:1911.04738, 2019.

[83] Wang Y Y, Wang J R, Cao Z L, et al Molecular contrastive learning of representations via graph neural networks[J]. Nature Machine Intelligence, 2022, 4(3): 279-287.

[84] Wang Z Y, Liu M, Luo Y Z, et al. Advanced graph and sequence neural networks for molecular property prediction and drug discovery[J]. Bioinformatics, 2022, 38(9): 2579-2586.