華東理工大學学报 自然科学版 Natural Science Edition
JOURNAL OF EAST CHINA UNIVERSITY OF SCIENCE AND TECHNOLOGY

基于图神经网络的单步逆合成集成研究

吴和兴 任付越 彭雨 张鹏鹏 卢静宜

Graph Neural Network-Based Ensemble Methods for Single-Step Retrosynthesis

WU Hexing, REN Fuyue, PENG Yu, ZHANG Pengpeng, LU Jingyi

在线阅读 View online: http://doi.org/10.14135/j.cnki.1006-3080.20260204001

您可能感兴趣的其他文章

Articles you may be interested in

基于多通道图神经网络和CNN-BiLSTM的漏洞检测方法

Vulnerability Detection Method Based on Multi-Channel Graph Neural Network and CNN-BiLSTM

华东理工大学学报（自然科学版）. 2025, 51(6): 835-842 http://doi.org/10.14135/j.cnki.1006-3080.20250317001

基于多模态特征融合的药物靶标亲和力预测

Drug-Target Affinity Prediction Based on Multi-Modal Feature Fusion

华东理工大学学报（自然科学版）. 2026, 52(1): 118-128 http://doi.org/10.14135/j.cnki.1006-3080.20250418001

融合建模的图神经网络会话推荐模型

Graph Neural Network Session Recommendation Model with Fusion Modeling

华东理工大学学报（自然科学版）. 2025, 51(6): 827-834 http://doi.org/10.14135/j.cnki.1006-3080.20250225002

基于子空间多尺度特征融合的试卷语义分割

Semantic Segmentation of Test Papers Based on Subspace Multi-Scale Feature Fusion

华东理工大学学报（自然科学版）. 2023, 49(3): 429-438 http://doi.org/10.14135/j.cnki.1006-3080.20220117001

基于多粒度网络预测增强子-启动子相互作用

Enhancer-Promoter Interaction Prediction Based on Multi-Granularity Network

华东理工大学学报（自然科学版）. 2024, 50(1): 106-113 http://doi.org/10.14135/j.cnki.1006-3080.20221213001

基于融合路径监督的多波段图像语义分割

Multi-Spectral Image Semantic Segmentation Based on Supervised Feature Fusion

华东理工大学学报（自然科学版）. 2021, 47(2): 233-240 http://doi.org/10.14135/j.cnki.1006-3080.20191216002

扫码关注公众号，获取更多信息！

文章编号：1006-3080(2026)05-0713-12 DOI：10.14135/j.cnki.1006-3080.20260204001

基于图神经网络的单步逆合成集成研究

吴和兴[1]， 任付越[1]， 彭 雨[1]， 张鹏鹏[2]， 卢静宜[1]
（1. 华东理工大学信息科学与工程学院, 上海 200237； 2. 浙江大学药学院, 杭州 310058）

摘要：化学逆合成分析是从目标产物分子反向推导至可商业获取前体分子的过程，为了解决现有单步化学逆合成方法在泛化能力与预测精度方面的局限，本文提出了一种融合分子指纹的图神经网络集成预测模型，解决了单一模型扩展性和可解释性不高的问题。该方法通过构建多维分子表征并动态选择最优预测策略，以提升逆合成预测的准确性与稳定性。模型基于消息传递神经网络(Message Passing Neural Network，MPNN)、全局自注意力机制及 AttentiveFP 网络提取分子结构特征，并引入扩展连接指纹(Extended Connectivity Fingerprints，ECFP)实现局部结构信息与全局分子特征的联合建模，构建了 MPNN+FPS (Fingerprints) 与 AttentiveFP+FPS 两种融合架构，在 USPTO-50K 和天然产物数据集中进行了大量实验，本文方法在预测精度与泛化能力方面均表现出显著优势，为复杂分子与天然产物的逆合成分析提供了一种新的技术路径。

关键词：逆合成反应预测；图神经网络；分子指纹；产物分类；特征融合；动态策略选择

中图分类号：TP181 **文献标志码：**A

化学逆合成分析是一种从目标分子出发，反向推导其可行前体分子的过程。逆合成的概念最早在 20 世纪 60 年代被提出，推动了计算机辅助综合规划(Computer-Aided Synthesis Planning, CASP)的发展，随着计算机技术的发展，CASP 也成为解决有机化学合成的重要工具。早期的逆合成设计完全依赖化学家的专业知识和经验，过程繁琐且耗时。近年来，随着人工智能(AI)技术的快速发展，AI 在逆合成分析中的应用逐渐变得热门，并带来了革命性的变化，使得逆合成预测更加高效和准确。

当前单步逆合成预测方法主要可以分为三类：基于模板、无模板或半模板的方法。基于模板的方法是将逆合成任务建模为与模板匹配问题或模板选择问题，通过将目标分子与预先计算好的模板进行匹配来预测可能的反应路径，这些模板中包含了特定类型的化学反应的基本特征。Coley 等[1] 计算给定产品与语料库中存在的分子之间的分子指纹相似度，对候选模板进行排序。Segler 等[2] 利用深度神经网络学习多分类器进行模板的选择。Dai 等[3-4] 提出了 GLN(Graph Logic Network)方法，构建了一个条件图逻辑网络来学习模板和反应物的条件联合概率以选择模板。Chen 等[5]提出了 LocalRetro 方法，对目标分子的局部原子或键模板的适用性进行了评估，并通过全局自注意力机制将非局部效应纳入化学反应中。Ishida 等[6] 提出了 ReTRek 方法，开发了一个数据驱动的 CASP 程序，将逆合成的知识引入到评估体系中，使用聚合断开分数、可获得物质分数、环断开分数、选择性转换分数等 4 个分数进行综合评判。Xie 等[7] 提出了 RetroKNN 方法，在推理时通过 KNN 从预先构建的原子模板和键模板存储库中检索，并结合神经网络的预测生成最终输出；同时提出了一种轻量级适配器网络来自适应地调整神经网络和 KNN(K Near Neighbor)的权重，从而提升预测的准确性。基于模板的方法虽然提供了反应的可解释性和

收稿日期： 2026-02-04
作者简介： 吴和兴(2001—)，男，浙江人，硕士生，主要研究方向为化学反应逆合成研究。E-mail: wuhexing02@163.com
通信联系人： 卢静宜，E-mail: jylu_cise@ecust.edu.cn
引用本文： 吴和兴, 任付越, 彭 雨, 等. 基于图神经网络的单步逆合成集成研究 [J]. 华东理工大学学报(自然科学版), 2026, 52(5): 713-724.
Citation: WU Hexing, REN Fuyue, PENG Yu, *et al*. Graph Neural Network-Based Ensemble Methods for Single-Step Retrosynthesis[J]. Journal of East China University of Science and Technology, 2026, 52(5): 713-724.

分子的有效性,但是在未提供模板的天然产物的预测上,泛化性和可扩展性受限,因而影响实际应用。

自从反向传播算法提出之后[8],多层神经网络能够通过梯度下降进行有效训练,从而推动了多层感知机等模型的发展[9]。为了解决图结构复杂的拓扑结构问题,图神经网络(Graph Neural Network, GNN)也被提出,旨在将深度学习方法推广到图结构数据中,Gori 等[10]首次提出 GNN 框架,通过递归方式在节点之间传播信息,实现图结构数据的学习表示,Kipf 等[11]提出了经典的 GCN(Graph Convolutional Network)模型,通过对谱图卷积进行一阶近似,使图卷积运算能够高效应用于大规模图数据,从而极大地推动了 GNN 的发展。2017 年,Transformer 模型的提出也加强了人们对于端到端生成式单步逆合成模型的研究[12],随着深度学习的发展,越来越多的深度学习方法逐渐被应用于化学逆合成中,为计算机辅助合成规划和逆合成预测等任务提供了新的研究思路。

基于无模板方法是利用生成模型来生成反应物分子,而不是依赖于预定义的模板。现有的大多数方法将该任务转化为序列到序列的翻译问题,化学分子式使用简化分子线性输入规范(Simplified Molecular Input Line Entry System, SMILES)[13]表示。Liu 等[14]最早提出了利用长短期记忆(Long Short-Term Memory, LSTM)网络的序列到序列模型将产物的 SMILES 表示转换为反应物的 SMILES 表示。Karpov 等[15]进一步提出了基于 Transformer 的序列到序列方法用于逆合成。Schwaller 等[16]在 Transformer 模型的基础上预测了序列的概率分布以估计结果的置信度。Zheng 等[17]提出了 SCROP(Self-Corrected Retrosynthesis Predictor)的方法,在 Transformer 模型中加入了一个语法校正网络对预测出的不正确结果进行修正,旨在解决序列到序列方法中普遍存在的语法失效问题。Tetko 等[18]在 Transformer 模型的基础上对数据进行了增强,例如生成随机 SMILES 增加数据多样性,使用正向和逆向的混合反应数据,在推理阶段使用束搜索方法等,提高了模型预测的准确性。Wan 等[19]引入了局部-全局注意力头,可以同时编码分子的序列和图结构信息,并在局部反应区域和全局反应背景之间高效交换信息。Zhong等[20]提出了基于根原子对齐的 SMILES 方法,为反应物和产物指定相同的根原子,使得输入/输出的 SMILES 表示之间有着严格的一对一映射,减少了编辑距离,提高了预测的准确性。Fang 等[21]利用共同保留的子结构进行逆合成的预测,通过预测和提取子结构,在 Transformer 模型上进行子结构水平的序列到序列预测,最后合并预测片段和产物得到反应物结果。Zhang 等[22]提出了 RetroWISE 的方法,利用基础模型从非配对数据生成计算机模拟反应,用作模型训练的额外数据,提高了罕见变换的预测准确性。由于无模板方法完全由数据驱动,因此在分子可解释性和化学有效性上存在不足,虽然可以通过加入一些化学知识的信息来弥补这些缺失,然而,针对天然化学分子产物,采用无模板方法依然会得到一些不够准确的结果。

不同类型的逆合成方法各具优势:模板方法在化学合理性与可解释性上表现突出,无模板方法在覆盖率与灵活性方面更优。实际应用中,不同分子在两类模型中的预测效果差异显著,因此融合不同类型的逆合成模型以动态选择最优预测策略成为提升总体预测性能的重要方向。Levin 等[23]聚焦于酶促反应与传统有机反应的协同融合,从数据库中提取所有酶促反应,应用模板方法训练出一个酶促反应的逆合成模型,并与传统逆合成模型结合应用于多步逆合成的路径搜索中。Kim 等[24]先将序列到序列模型与图到序列模型融合,然后经 softmax 归一化,并合并排序,利用序列到序列模型的高单步准确率与图到序列模型的高多步适配性,提升单步预测的全面性,但最终结果由一个简单的模型进行训练,丢失了分子的有效信息。Lin 等[25]提出了一种基于“能量”的重排序模型,将不同单步逆合成模型输出依据“能量”的信息进行一个重排序。Li 等[26]则是训练了一个单独的模型 RetroRanker,对单个逆合成模型的输出结果进行重排序。MAZIAZ 等[27]提出了 RetroChimera模型,通过基于学习的集成策略整合来自两个具有不同归纳偏差的新模型的预测结果,从而为不同模型类别的优势提供了关键视角。

综上,现有的逆合成融合研究存在 3 个紧密关联的难点:

(1)多层级分子结构表征的完备性问题:现有模型多聚焦于原子或局部反应中心,忽略了分子全局结构。如何融合原子、键、分子指纹等多层级信息以全面刻画产物分子,是实现模型泛化与分类判断的关键。

(2)多模型集成的动态判别机制问题。当前多模型融合多为静态规则(如投票、置信度排序),难以针对不同分子特征自适应选择最优预测路径。如何将单步逆合成预测转化为模型选择分类任务,使分类器基于分子特征学习各类模型的适用性,是融合策略的核心问题。

(3)预测置信度与失败样本识别问题。在逆合

成预测中, 对高难度或结构异常分子的失败检测尤为重要。如何通过分类模型输出的不确定性或得分信息, 评估样本的可合成性, 为多步逆合成提供前置信号, 是提升整体系统可靠性的重要挑战。

针对上述问题, 本文提出了一种融合分子指纹的 GNN 单步逆合成分类模型, 实现不同单步逆合成方法的有效集成。模型基于消息传递网络(Message Passing Neural Network, MPNN), 结合扩展连接指纹(Extended Connectivity Fingerprints, ECFP), 并引入全局自注意力机制以实现局部与全局特征的联合建模。具体而言, 模型根据目标产物分子的多尺度分子表示特征, 预测其在不同单步逆合成算法下的适配程度, 并据此动态选择最优逆合成策略。本文的主要创新点包括:

(1)提出融合原子、键与分子指纹的多层级特征的分子结构表征方法, 显著提高分子描述的全面性与区分度。

(2)将单步逆合成预测转化为分子分类任务, 以学习驱动的方式实现多模型的动态集成, 替代传统投票或置信度加权机制, 提升预测效率与准确性。

(3)在模型分类结果中引入可合成性的判别能力, 为高难度分子提供不确定性估计, 从而在逆合成预测之前识别潜在的不可合成分子或高风险样本, 提高整体逆合成决策的稳定性与安全性。

1　数据准备及多模型融合架构设计

为充分融合有模板与无模板逆合成方法的优势, 本文设计了一种基于分子结构特征分类的动态策略选择框架。首先, 针对目标产物分子, 通过 GNN 提取其原子维度特征以及键维度的特征表征, 包括原子类型、电荷类型、键类型和原子度数等特征, 同时融合分子指纹特征以增强模型对分子化学特性的捕捉能力。随后, 构建一个分子分类模型, 对不同的目标产物分子进行策略判别, 自动判断其更适合采用有模板方法还是无模板方法进行逆合成预测。通过这种基于特征驱动的自适应选择机制, 模型能够充分发挥两类方法的互补优势, 从而在整体上提升逆合成分析的准确性与泛化能力。

1.1　数据集与数据准备

选择化学反应预测领域中最常用的 USPTO-50K 数据集, 作为模型训练与评估的基础数据来源。USPTO-50K 数据集是由美国专利商标局发布的一系列数据集[28], 这些数据经过专家或算法处理后, 提取出标准化的化学反应形式, 因此, 被广泛应用于反应预测与逆合成方向的研究中。

USPTO-50K 数据集包含约 50 000 条具有准确原子映射的反应数据。每条数据均由一组反应物与产物组成, 并以标准的 SMILES 表示形式存储。本文使用与文献 [4] 相同的数据划分方式, 以 8∶1∶1 的数据比例分别划分为训练集、验证集、测试集。

此外, 为了验证所提出方法在天然产物数据上的适用性与泛化性能, 本文基于 Reaxys 数据库构建了一个天然产物单步逆合成数据集。Reaxys[29] 是由 Elsevier 维护的综合性化学信息数据库, 其整合了文献报道与专利中的化学反应、物性及合成路线信息, 被广泛用于有机合成设计与反应数据挖掘研究。该数据集包含大约 100 000 条训练数据、10 000 条验证数据以及 10 000 条测试数据, 相比 USPTO-50K, 该数据集规模更大、覆盖的反应类型更加多样化, 但同时也包含更多的噪声与数据不一致问题。这一数据集能够更好地模拟真实化学反应预测任务中数据分布复杂、反应多样的情况。

在实验过程中, 分别使用目前性能表现优秀的基于模板的单步逆合成模型与无模板的单步逆合成模型在数据集上进行训练与测试。对于测试集中的每一条反应, 使用两种不同的方法分别进行逆合成预测, 并计算 Top-*K* 准确率。然后比较两种方法在同一反应上的预测表现, 以确定哪种方法在该反应上表现更优。

具体而言, 我们将每条反应样本赋予一个新的标签, 用以标识在该反应中两种方法的表现差异。该标签的设计也考虑到了两种预测方法均成功与均失败的情况。标签类型与含义如表 1 所示。

表 1　逆合成模型集成分类任务的标签定义

Table 1　Label definition for inverse synthetic model ensemble classification task

Label	Description
0	Template-based method performs better
1	Template-free method performs better
2	Both methods give the same result; Choose the template-based method
3	Both methods give the same result; Choose the template-free method
4	Neither method predicts the correct result

这种标签体系的设计使得分类模型不仅能够学习到“哪种方法更好”的信息, 还能捕捉更细粒度的差异特征, 例如在结果相同的情况下对模型的偏好倾向, 从而为后续的模型融合或自适应方法的选择提供依据。通过对这一分类任务的训练, 模型能够根

据目标分子的结构特征，自动判断最合适的逆合成分析路径，实现多种单步逆合成方法的高效集成协同。

1.2 多模型集成架构

为进一步分析不同逆合成预测方法的适用场景，并实现与目标分子对应的最优逆合成方法的自动化选择，本文构建了一种融合分子指纹特征的 GNN 分类模型，如图 1 所示。该模型以产物分子的结构特征为输入，能够精准预测逆合成方法性能的优劣，为逆合成路线的高效设计提供技术支撑。该模型主要包括分子表示层、特征编码层和分类决策层。

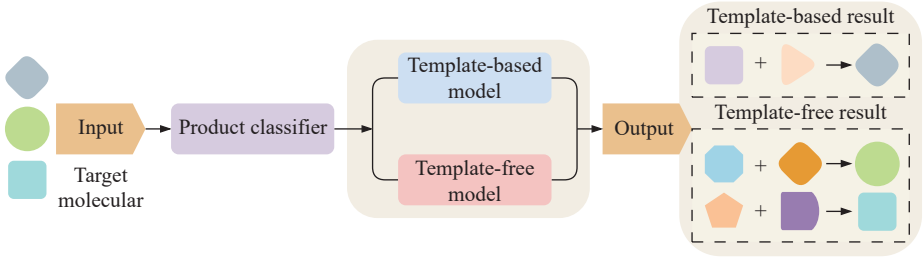


图 1 产物分子分类模型结构图

Fig. 1 Structure diagram of product molecular classification model

在化学信息学任务中，如何有效表示分子的结构信息是模型性能的关键因素之一。本文采用两种互补的分子表示方式：图结构表示与分子指纹表示。其中图结构表示将每个化学分子视为无向图，每个节点携带原子的特征向量，包括原子类型、电荷、价态以及是否在环中等，每条边携带键类型特征，包括单键、双键、芳香键等信息，分子指纹采用 ECFP4 分子指纹对分子进行拓扑结构编码，得到定长的高维向量表示，用于捕捉局部结构模式与官能团信息。两种表示方式能够从不同层面反映分子的化学特性：图结构有助于学习局部信息与全局拓扑信息，而分子指纹表示有助于快速捕获高频子结构与分子模式，从而形成互补特征。

为了充分挖掘分子图的结构特征，采用 GNN 作为主要的编码框架。具体而言，选用 MPNN 和 AttentiveFP 两种典型结构进行比较与融合。MPNN 模块通过多轮消息传递机制，在分子图中迭代聚合邻居节点的信息，从而得到每个原子的上下文感知表示。AttentiveFP 模块在 MPNN 的基础上引入了注意力机制，通过学习不同邻居节点的重要性权重，增强模型对关键化学子结构的捕捉能力，其核心思想是让模型能够更加关注对反应起决定性作用的原子或官能团，从而提升分类的准确性。此外，将分子指纹表示通过全连接层映射至与图表示相同的向量维度，并与 GNN 输出的分子嵌入进行拼接融合，从而形成兼具局部与全局特征的统一分子表征。

融合后的分子特征向量将被输入至分类决策模块。该模块由全连接网络构成，并在层间加入归一化与 Dropout 操作以防止过拟合。输出层使用 ReLU 激活函数将特征向量映射至五维概率分布，对应 5 种标签类别。预测结果即为模型在当前分子上选择的最优逆合成方法的类型。损失函数采用交叉熵损失，并通过 Adam 优化器进行参数更新。为提升模型的泛化性能，在训练过程中引入早停策略与学习率衰减机制，以防止模型在噪声数据上的过拟合。

整体模型结构如图 2 所示。输入为产物分子，之后融合 GNN 的结构建模能力与分子指纹的局部结构信息，同时引入全局自注意力机制，对分子整体结构进行建模，并加入分子指纹进行模型的分类训练。输出为在给定一个目标分子的条件下，预测该分子更适合使用哪种逆合成分析方法。在逆合成策略选择方面，本文选用了两个在 USPTO-50K 数据集上表现优异的有模板方法与无模板方法，两者的结合不仅能够覆盖更广泛的反应类型，而且能更充分地发挥模型融合的优势，从而在准确率上超越单一模型的表现。

1.3 融合分子指纹的 GNN 分类模型

在本文中，产物分子首先被表示为一个分子图 $G=(N,B)$，其中 N 表示原子节点集合，B 表示化学键集合，每个原子和化学键被初始化为固定维度的特征向量。具体来说，每个原子被表示为一个 80 维向量，包含了原子类型、价态、电荷、杂化状态等基本化学属性；每个键被初始化为一个 13 维的向量，涵盖了键类型、是否为共轭键、是否处于环中等特征。这一特征初始化过程由 DGL(Deep Graph Library)提供的分子图构建与特征化工具完成，从而将化学分子转换为适合深度学习处理的图结构输入。

在得到分子图特征后，再使用两种典型的网络结构(MPNN 与 AttentiveFP架构)对产物分子进行编码。MPNN 是一种通用的 GNN 框架，适用于分子图

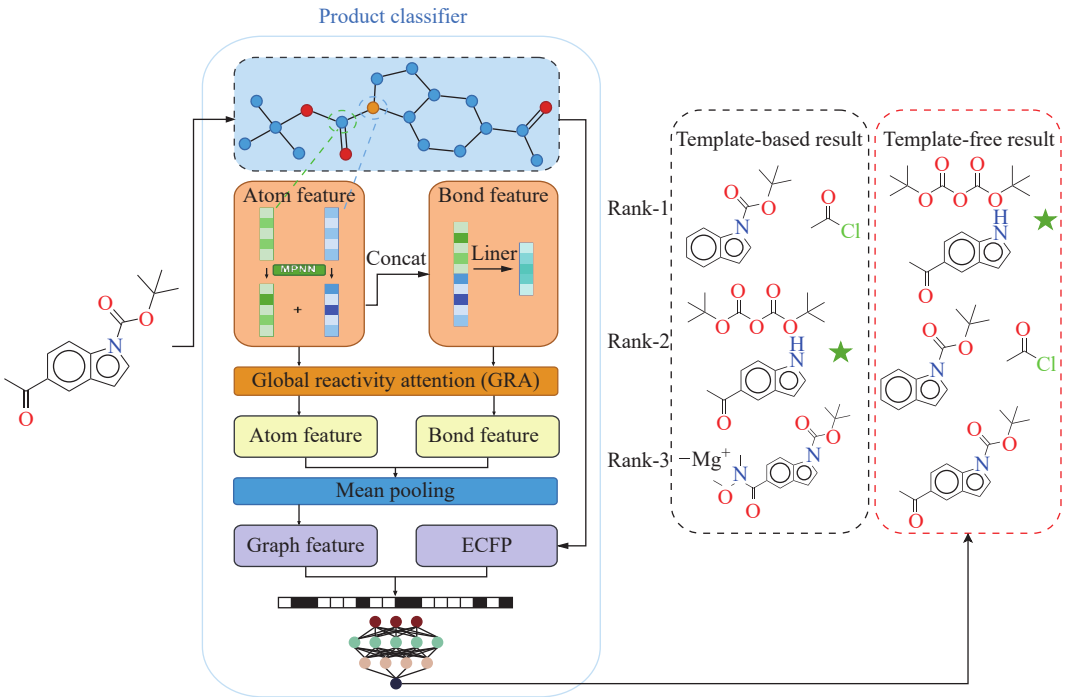


图 2　融合分子指纹的 GNN 分类模型 MPNN+FPS 架构图

Fig. 2　MPNN+FPS architecture diagram of graph neural network classification model integrating molecular fingerprints

的表征学习, 通过消息传递机制, 在每一轮迭代中更新分子图节点之间的特征, 并通过聚合函数将邻居信息整合; 再结合自身的状态更新节点表示, 经过多轮迭代之后, 节点的表示能够融合来自局部邻域的结构信息; 最后通过全局池化函数将整个图的节点表示编码为分子级别的嵌入向量, 具有较强的灵活性和可扩展性。AttentiveFP 是一种基于图注意力机制(Graph Attention Network, GAT)的分子图表示方法, 可以捕获图结构中的全局信息, 对比传统的 GNN, AttentiveFP 引入注意力机制使得其可以更加关注长距离的信息而不是局部特征, 此外, AttentiveFP 采用了多层迭代的传播, 并结合了 GRU (Gated Recurrent Unit)结构对节点表示进行逐步更新, 可以有效缓解长距离信息丢失的问题。在本文中, 我们在 MPNN 的基础上也加上了注意力机制来确保其拥有更加强大的特征编码能力。

为了更加清晰地展示本文提出的动态策略选择框架, 图 3 给出了整体流程图。该框架主要包括三个阶段: 分子结构编码、模型分类选择以及逆合成预测。首先将输入产物分子转换为分子图, 并通过 MPNN 和 AttentiveFP 两种 GNN 进行结构编码, 同时结合 ECFP4 分子指纹获得多层级分子表示。随后将融合特征输入分类模型, 预测最适合该分子的逆合成预测模型类别, 最后根据分类结果动态调用对应的单步逆合成模型生成反应物预测结果。

1.3.1　特征提取以及特征融合　首先使用消息传递网络 MPNN 来更新每个原子的特征表示。通过多轮消息传递, MPNN 使得每个原子可以聚合邻居节点的结构信息, 从而获得上下文感知的嵌入表示。假设节点特征在第 t 步时的状态向量为 $\boldsymbol{h}_u^t$, 邻居节点在第 t 步的状态向量为 $\boldsymbol{h}_v^t$, 边的特征为 $\boldsymbol{e}_{uv}$, 则消息传递与节点更新过程可表示为:

$$\boldsymbol{m}_u^{t+1} = \sum_{v \in N(u)} M_t(\boldsymbol{h}_u^t, \boldsymbol{h}_v^t, \boldsymbol{e}_{uv}) \tag{1}$$

$$\boldsymbol{h}_u^{t+1} = U_t(\boldsymbol{h}_u^t, \boldsymbol{m}_u^{t+1}) \tag{2}$$

其中, M_t 表示消息函数, U_t 表示节点更新函数, $\boldsymbol{N}_u$ 表示节点 u 的邻居节点, 在所有消息传递轮数结束后, 对所有节点的状态向量进行图级聚合, 以获得图的全局表示, 状态聚合函数 R 可以采用求和、平均或者最大值的方法进行实现:

$$\boldsymbol{N}_u = R(\{\boldsymbol{h}_v^t | v \in G\}) \tag{3}$$

其中 G 表示化合物所有节点的图。在获取了原子级别的特征表示后, 进一步构建键的特征表示。键的特征表示由其两端连接的原子特征拼接得到, 并且拼接好的向量需要进行线性变化, 具体公式如下:

$$\boldsymbol{B}_{uv} = w([\boldsymbol{N}_u, \boldsymbol{N}_v]) + b \tag{4}$$

其中, $[\boldsymbol{N}_u, \boldsymbol{N}_v]$ 表示将两个原子特征向量进行拼接, $\boldsymbol{w}$ 和 b 分别表示线性全连接层的权重和偏置参数。

为了进一步提升模型对全局结构信息的建模能力, 将原子特征与键特征进行了全局自注意力(Self-

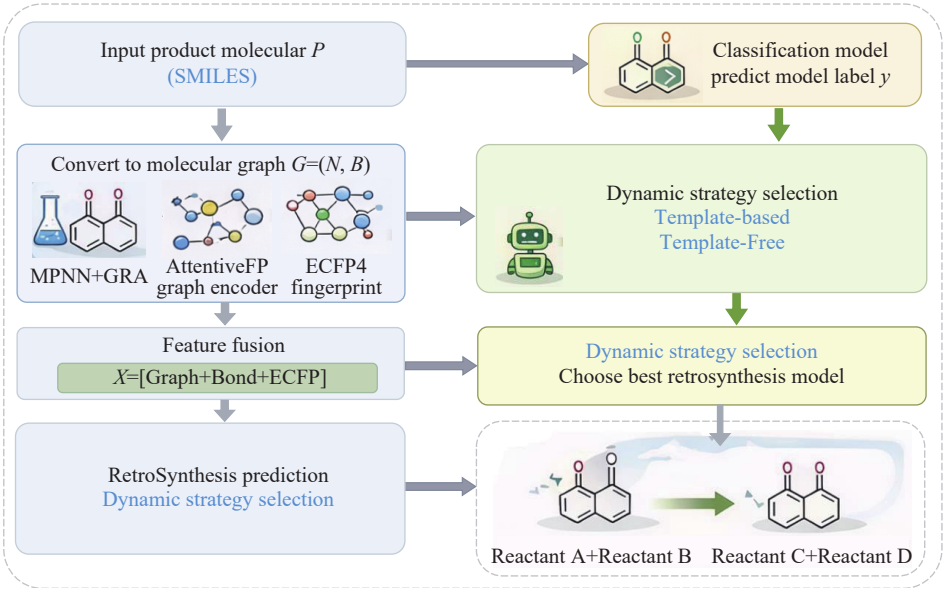


图 3 动态策略选择框架整体流程图

Fig. 3 Overall flowchart of dynamic strategy selection framework

Attention)机制的处理,引入了 Transformer 模型中的自注意力机制的模块,通过学习给定特征的键、查询和值来获得原子和键之间的上下文关系,该机制允许模型在计算特征表示时动态地考虑其他节点或边对当前节点或边的影响。我们将原子特征和键特征分别作为输入,通过 GRA(Global Reactivity Attention)进行上下文建模,生成增强后的全局表示,然后对所有注意力增强后的节点和边表示进行平均池化,以获得最终的分子图表示。用 GRA 表示全局自注意力函数,Mean 表示平均池化函数,则结果表示为:

$$\boldsymbol{N}_u' = \mathrm{GRA}(\boldsymbol{N}_u, \{\boldsymbol{N}_u | u \in N\}, \{\boldsymbol{B}_{uv} | (u,v) \in B\}) \tag{5}$$

$$\boldsymbol{N}_u'' = \mathrm{Mean}(\boldsymbol{N}_u' | u \in G) \tag{6}$$

$$\boldsymbol{B}_{uv}' = \mathrm{GRA}(\boldsymbol{B}_{uv}, \{\boldsymbol{N}_u | u \in N\}, \{\boldsymbol{B}_{uv} | (u,v) \in B\}) \tag{7}$$

$$\boldsymbol{B}_{uv}'' = \mathrm{Mean}(\boldsymbol{B}_{uv}'' | u \in G) \tag{8}$$

其中,$\boldsymbol{N}_u'$ 为 GRA 增强后的节点 $\boldsymbol{N}_u$ 的特征表示,$\boldsymbol{N}_u''$ 为对所有节点的 $\boldsymbol{N}_u'$做 Mean 后得到的全局图表示,B 为所有边的集合,$\boldsymbol{B}_{uv}$ 为原始的边 (u,v) 的特征表示,$\boldsymbol{B}_{uv}'$ 为 GRA 增强后的边 (u,v) 的特征表示,$\boldsymbol{B}_{uv}''$ 为对所有边的 $\boldsymbol{B}_{uv}'$ 做 Mean 后得到的全局边表示。同时,通过 AttentiveFP 进行产物分子编码,AttentiveFP通过引入 GAT 的结构,在每轮迭代中利用注意力机制动态分配权重,关注与当前节点或边相关性更高的邻居信息,从而生成更具全局感知的特征表示。

1.3.2 策略分类 在 AttentiveFP 中,每个节点的更新过程可以分为两个步骤:消息聚合与状态更新。在消息聚合阶段,中心节点 u 从其邻居节点集合 $\boldsymbol{N}(u)$ 中接收信息,与 MPNN 不同,AttentiveFP 使用可学习的注意力权重来加权邻居节点的贡献。其计算过程可表示为:

$$\boldsymbol{m}_u^{(t)} = \sum_{v \in N(u)} \alpha_{uv} \cdot \boldsymbol{h}_v^{(t-1)} \tag{9}$$

其中,α_{uv} 为注意力系数,计算方式为:

$$\alpha_{uv} = \mathrm{softmax}(\mathrm{ReLu}(\boldsymbol{a}^{\mathrm{T}}[\boldsymbol{W}\boldsymbol{h}_u^{(t)}, \boldsymbol{W}\boldsymbol{h}_v^{(t)}, \boldsymbol{e}_{uv}])) \tag{10}$$

其中,$\boldsymbol{a}$ 表示注意力参数向量,$\boldsymbol{W}$ 为共享的线性变换矩阵,$\boldsymbol{e}_{uv}$ 为边特征。

在状态更新阶段,GRU(Gated Recurrent Unit)用于逐步更新节点表示,可以有效缓解长距离信息丢失问题。节点更新公式为:

$$\boldsymbol{h}_u^{(t)} = \mathrm{GRU}(\boldsymbol{h}_u^{(t-1)}, \boldsymbol{m}_u^{(t)}) \tag{11}$$

在图级读出阶段,AttentiveFP 通过全局池化生成分子图的最终表示:

$$\boldsymbol{h}_G = \mathrm{Pool}(\{\boldsymbol{h}_u^{(T)} | u \in G\}) \tag{12}$$

式(11)和式(12)中,$\boldsymbol{h}_u^{(t)}$ 表示第 t 轮阶段节点的状态,$\boldsymbol{h}_u^{(T)}$ 表示节点在 T 轮后的状态。除了 GNN 生成的图结构特征外,本文引入了基于分子指纹的信息来辅助模型学习,使用扩展连接指纹 ECFP4 作为分子的结构指纹,这是一种基于原子邻域拓扑结构的离散表示方法,由 Rogers 等[30] 提出,能够有效捕捉分子的局部子结构特征。设产物分子的分子指纹表示为 ECFP(P),最终将图特征、键特征与分子指纹向量进

行拼接, 得到完整的特征编码:

$$\boldsymbol{X} = [\boldsymbol{N}_u'', \boldsymbol{B}_{uv}'', \text{ECFP}(P)] \tag{13}$$

最后, 使用该特征向量 $\boldsymbol{X}$ 输入到一个分类模型中, 预测产物分子的标签 $\hat{y}$, 从而选择最合适的逆合成预测模型:

$$\hat{y} = \boldsymbol{W}_2\text{Drop}(\text{GELU}(\text{BN}(\boldsymbol{W}_1\boldsymbol{X} + b_1))) + b_2 \tag{14}$$

其中, GELU 表示激活函数, BN 表示批量归一化层, $\boldsymbol{W}$ 和 b 分别为全连接层的权重和偏置。

1.3.3　*算法流程结构*　AttentiveFP+FPS 架构的整体示意图如图 4 所示, 从输入产物分子中提取原子特征与键特征, 然后经过图注意力机制, 学习到局部与全局的图特征, 最后与 ECFP4 分子指纹特征结合, 输出得到产物分子编码结果。

为了更加清晰地描述本文提出的动态策略算法步骤, MPNN+FPS 架构的整体执行流程如表 2 所示。该流程首先对输入产物分子的原子特征进行编码得到原子特征, 然后通过原子特征的线性变化得到键特征, 并融合分子指纹的多元特征; 随后通过分类模型预测最适合该分子的逆合成预测模型类别; 最后根据预测结果动态选择对应的单步逆合成模型生成反应物。

本文综合采用了 MPNN+GRA 与 AttentiveFP 两种方法对分子图进行编码。前者偏向于从局部邻域逐步聚合信息, 并在后处理阶段通过全局注意力机制建模整体依赖关系; 而后者则在消息传递的每一步中都引入了注意力权重和 GRU 更新, 能够更有效地捕获远程依赖与全局信息, 为后续分类任务建模

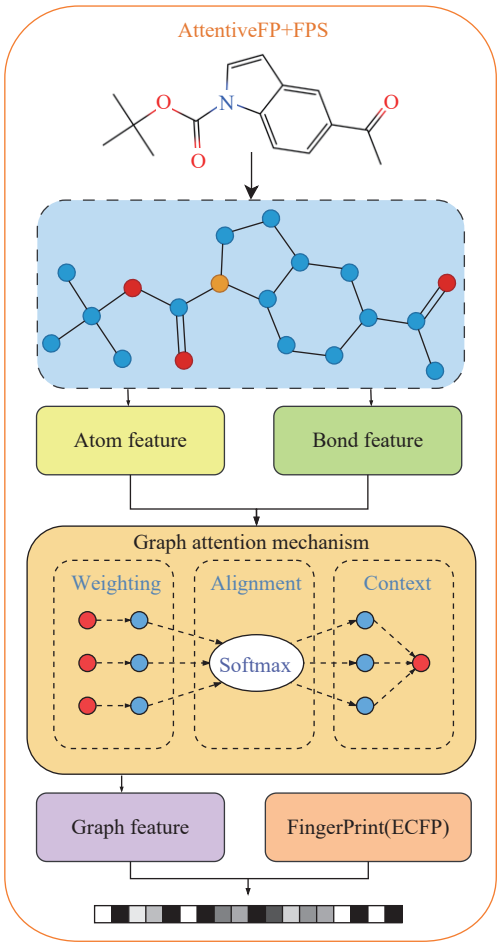


图 4　产物分子 AttentiveFP+FPS 编码结构图

Fig. 4　Structure diagram of product molecule AttentiveFP+FPS coding

表 2　MPNN+FPS 架构模型训练及预测流程

Table 2　MPNN+FPS architecture model training and prediction process

Algorithm
Input: Target product molecule P
Output: Predicted reactant set R
1 Convert the input product molecule P into a molecular graph $G = (\boldsymbol{N}, \boldsymbol{B})$
2 Initialize the atom feature matrix $\boldsymbol{N}$, including atom type, valence, and other atomic properties
3 Initialize the bond feature matrix $\boldsymbol{B}$, including bond type, conjugation information
4 Encode the molecular graph using a MPNN to obtain node representations $\boldsymbol{N}_u$
5 Construct bond feature representations based on atom features, where bond representations $\boldsymbol{B}_{uv}$
6 Introduce the Global Reactivity Attention (GRA) mechanism to capture long-range dependencies within the molecule, producing enhanced feature representations $\boldsymbol{N}_u''$ and $\boldsymbol{B}_{uv}''$
7 Compute the ECFP4 molecular fingerprint of the molecule to obtain structural fingerprint features F
8 Fuse the multi-source features to obtain the final feature representation $\boldsymbol{X} = \text{concat}(\boldsymbol{N}_u'', \boldsymbol{B}_{uv}'', F)$
9 Input the fused feature $\boldsymbol{X}$ into a classification model to predict the optimal retrosynthesis model category y for the given molecule
10 Select the corresponding retrosynthesis prediction model M from the model pool according to the predicted category y
11 Perform single-step retrosynthesis prediction on the target molecule P using the selected model M
12 Output the predicted reactant set R

提供了更强大的分子图表示。

2 结果与讨论

2.1 在 USPTO-50K 上的结果

本节采用 Top-*K* 的准确率作为评价指标，示出了模型在 USPTO-50K 数据集上结合模板方法与非模板方法的性能。Top-*K* 准确率是逆合成预测任务中常用的评价指标，用于衡量模型在给定目标产物分子时，其预测结果列表中前 *K* 个候选反应是否包含真实反应。具体而言，当真实反应物组合出现在模型预测的前 *K* 个结果中时，即认为该样本在 Top-*K* 条件下预测正确。该指标能够反映模型在多候选预测场景下的实际可用性，因此被广泛应用于单步逆合成方法的性能评估。其中，模板方法采用 LocalRetro，非模板方法采用 R-SMILES，充分融合了两者的优势。如表 3所示，在 USPTO-50K 数据集上，本文方法在 Top-*K* 准确率上均表现出优越的性能，MPNN+FPS 在 Top-1 上获得了 59.1% 的准确率，明显优于 LocalRetro 的 53.4% 和 R-SMILES 的 56.3% 的准确率，并且也超过了文献基线结果。在更高的 *K* 值下，该优势依旧保持，例如在 Top-5 上，MPNN+FPS 的准确率达到 87.2%，高于 LocalRetro 的 85.9% 以及 R-SMILES 的 86.2%。这表明模型能够有效学习两种方法之间的差异，并结合其正确部分。此外，在 Top-3 和 Top-5 中，模型同样也表现出色，优于两种基线方法 LocalRetro 和 R-SMILES，具有较强的鲁棒性。

表 3 不同方法在 USPTO-50K 数据集上的 Top-*K* 准确率

Table 3 Top-*K* accuracy of different models on USPTO-50K dataset

Model	Accuracy/%					
	Top-1	Top-3	Top-5	Top-10	Top-20	Top-50
RetroSim	37.3	54.7	63.3	74.1	82.0	85.3
GLN	52.5	69.0	75.6	83.7	89.0	92.4
GraphRetro	53.7	68.3	72.2	75.5	—	—
LocalRetro	53.4	77.5	85.9	**92.4**	—	**97.7**
R-SMILES	56.3	79.2	86.2	91.0	93.1	94.6
Ours(MPNN+FPS)	**59.1**	**80.9**	**87.2**	92.4	**95.0**	96.6
Ours(AttentiveFP+FPS)	58.9	80.9	87.0	92.3	94.9	96.8

此外，本文引入 MaxFrag 指标以评估模型在逆合成最大片段预测任务中的性能。MaxFrag 指标关注模型在生成的候选反应中，是否能够正确预测出与真实反应中结构最为接近的最大反应物片段。具体而言，当模型预测结果中包含与真实反应物具有最大公共子结构或最大分子片段一致的候选时，即认为该样本在 MaxFrag 指标下预测正确。该指标能够更细致地反映模型对关键断键位置及核心结构片段的识别能力，尤其适用于评估复杂分子和多反应物场景下的逆合成预测性能。在 MaxFrag 指标下，集成模型再次显示出了优势。如表 4 所示，在 MPNN+FPS 模型结构和 AttentiveFP+FPS 结构上，Top-1 的 MaxFrag准确率分别为 61.9% 和 62.5%，优于基线方法；同时，AttentiveFP+FPS 的Top-3 和 Top-5 分别达到了 84.3% 和 90.1% 的正确率，均优于其他方法，这表明模型在识别结构相关前体的准确性上有所提升。另外，AttentiveFP+FPS 在 Top-1 的 MaxFrag 准确率上高于 MPNN+FPS，表明每轮迭代中的注意力机制能够更好地在融合模型中增强分子片段的选择性。

表 4 不同方法在 USPTO-50K 数据集上的 MaxFrag 准确率

Table 4 MaxFrag accuracy of different models on USPTO-50K dataset

Model	Accuracy/%					
	Top-1	Top-3	Top-5	Top-10	Top-20	Top-50
MEGAN	54.2	75.7	83.1	89.2	—	95.1
Aug.Transformer	58.5	73.0	85.4	90.0	—	—
Graph2Edits	56.8	80.3	87.5	92.8	—	95.4
LocalRetro	57.8	82.1	89.7	**95.0**	—	**98.4**
R-SMILES	61.0	82.5	88.5	92.8	94.6	95.7
Ours(MPNN+FPS)	61.9	83.9	89.8	93.9	95.8	96.9
Ours(AttentiveFP+FPS)	**62.5**	**84.3**	**90.1**	94.5	**96.5**	97.5

综上所述，本文融合方法不仅在精确匹配的准确率上优于现有基线模型，还在分子片段级别的预测上展现出更高的鲁棒性。这意味着在实际的合成规划中，本文方法能够以更可靠的方式对候选合成路径进行初步筛选，从而提升新分子合成的成功率。

2.2 高难度合成反应的分类结果

第 4 类反应是 USPTO 数据集中的特殊子类，主要包括高难度合成反应与潜在不可合成反应，如多步环化反应和不稳定中间体的反应，此类反应在实际有机合成中往往面临较高的不确定性和失败的风险，对模型的可合成性预测能力提出了更高的要求。准确识别并预测这类高难度反应的可合成性具有重要的研究与应用价值。一方面，它能够帮助模型在复杂反应空间中区别可行与不可行的反应路径，从而提升反应预测与设计的整体鲁棒性；另一方

面, 该预测能力可直接应用于多步逆合成规划中, 用作反应可行性约束或路径筛选准则, 为反应路线搜索提供可合成性先验, 从而有效减少不切实际或低概率路径的探索, 提升合成规划的整体可行性与效率。

在实验设计中, 本文将第 4 类反应的可合成性预测转化为二分类任务: 可合成性分子(代表了表 1 中前 4 类结果的总和)与不可合成性分子。由于 USPTO-50K 中第 4 类反应的数量较少, 因此我们将 USPTO-MIT 的第 4 类样本(12 000 条)与 USPTO-50K 数据集进行混合训练, 评估不同特征架构的分类性能, 结果如表 5 所示。AttentiveFP+FPS 准确率最高, 为 75.8%, 其次是 MPNN+FPS(75.7%)以及 FPS (75.5%)。结果表明, 将 GNN 中的注意力机制与分子指纹特征相结合, 能够更充分地捕捉复杂反应体系中的关键结构信息与反应位点差异, 从而在可合成性预测中表现出更强的判别能力。

表 5　第 4 类数据的预测模型准确率

Table 5　Accuracy of predicted model for the fourth category of data

Method	Accuarcy/%	Precision/%	Recall/%	F1/%
MPNN	71.6	64.6	53.7	58.7
FPS	75.5	70.7	52.5	60.3
AttentiveFP	69.1	63.6	44.1	52.1
MPNN+FPS	75.7	72.9	**60.8**	**66.3**
AttentiveFP+FPS	**75.8**	**73.3**	53.2	61.7
AttentiveFP+MPNN	72.9	61.4	50.3	55.3

基于注意力的融合模型更高的准确率表明其更好地捕捉了细微的反应模式, 值得在多分类扩展中进行进一步探索, 这一创新显著提升了模型的实用性。在多步逆合成规划中, 首先通过分类模型判断分子是否可合成, 若预测为不可合成, 则直接排除该分子, 避免后续逆合成分析的无效计算; 若预测为可合成, 则进一步使用融合模型选择最优逆合成方法。这种“先筛选后预测”的流程能显著提升合成规划的效率。

此外, 为了进一步分析模型在第 4 类反应识别中的表现, 将分类结果中的误判样本进行了统计与分析。具体而言, 误判主要可以分为两类: 假阳性(False Positive)和假阴性(False Negative)。其中, 假阳性样本指模型将实际不可合成或高难度反应错误地预测为可合成反应; 假阴性样本则指模型将实际可合成反应误判为不可合成反应。

通过对部分误判样本的结构特征进行分析, 假阳性样本通常具有较为复杂的分子骨架或高度取代的环结构, 这类分子在拓扑结构上与常见可合成分子具有较高相似性, 因此模型在结构表示层面容易将其误判为可合成分子。这说明当前模型主要依赖分子结构特征进行判断, 在缺乏反应条件、试剂信息或反应机理知识的情况下, 对复杂反应可行性的识别仍存在一定局限。另一方面, 假阴性样本则多出现在结构较为复杂但在实际化学中已有成熟合成路径的反应。综合来看, 这些误判现象表明, 尽管融合 GNN 与分子指纹特征的模型在整体分类性能上表现良好, 但在处理结构高度复杂或反应模式稀有的分子体系时仍存在一定局限。

2.3　在天然产物数据集上的结果

为了进一步验证模型在天然产物数据集上的效果, 从 Reaxys 平台上收集了部分天然产物单步反应数据集, 数据集一共有 12 万条单步逆合成反应数据, 按照 8∶1∶1 划分为训练集、验证集和测试集, 将数据集在 LocalRetro 方法与 R-SMILES 方法上进行训练与测试, 得到两种单步逆合成方法的结果, 使用融合分子指纹的 GNN 产物分子分类模型, 不同方法在天然产物数据集上的 Top-*K* 准确率和 MaxFrag 准确率结果分别如表 6 与表 7 所示。

表 6　不同方法在天然产物数据集上的 Top-*K* 准确率

Table 6　Top-*K* accuracy of different models on natural product dataset

Model	Accuracy/%					
	Top-1	Top-3	Top-5	Top-10	Top-20	Top-50
LocalRetro	45.4	58.2	63.1	68.6	72.5	75.6
R-SMILES	45.8	60.5	65.5	70.7	74.6	77.9
MPNN+FPS	48.0	61.7	66.5	72.0	75.7	79.0
AttentiveFP+FPS	**49.6**	**63.2**	**68.2**	**74.0**	**77.8**	**80.9**

表 7　不同方法在天然产物数据集上的 MaxFrag 准确率

Table 7　MaxFrag accuracy of different models on natural product dataset

Model	Accuracy/%					
	Top-1	Top-3	Top-5	Top-10	Top-20	Top-50
LocalRetro	49.1	61.9	66.8	72.5	76.6	79.7
R-SMILES	49.8	64.9	69.7	74.9	79.1	82.5
MPNN+FPS	51.6	65.4	70.2	75.6	79.6	82.9
AttentiveFP+FPS	**53.5**	**67.1**	**72.0**	**77.8**	**81.5**	**84.8**

在天然产物的数据集上, 模型整体性能全面领先, 预测准确率优势显著: 在 Top-*K* 准确率和 MaxFrag

准确率两大评价指标下,两种融合架构在所有评价维度均表现最优。在 Top-K 准确率上,AttentiveFP+FPS 模型结构在 Top-1 上获得了 49.6% 的准确率,相较于原始的 R-SMILES 方法提高了 3.8 个百分点。在更高的 K 值下,该优势依旧保持,例如在 Top-50 上,AttentiveFP+FPS 准确率达到 80.9%,高于 LocalRetro(75.6%)和 R-SMILE(77.9%)。并且 MPNN+FPS 的模型架构相较于两种基准方法也均有提升,表明基于分子指纹的 GNN 多模型集成的方法在天然产物上对不同单步逆合成的集成能力表现依旧突出。在 MaxFrag 准确率上,本文模型同样也表现出色,优于两种基线方法,AttentiveFP+FPS 的 Top-1 准确率达到了 53.5%,相较于原始的 R-SMILES方法提高了 3.7 个百分点,Top-50 的准确率也相较于原始的 R-SMILES 方法提高了 2.3 个百分点,显示出其在集成优化方面的鲁棒性。融合分子指纹的 GNN 分类模型的两种架构均能有效提升逆合成方法预测性能,其中 AttentiveFP+FPS 架构在天然产物数据集上表现最优,MPNN+FPS 架构则在传统 USPTO-50K 数据集上突出,实现了性能与计算复杂度的平衡,两种架构均可为天然产物的逆合成方法筛选提供高效、可靠的技术支撑。

综上所述,本文集成模型不仅在通用数据集上表现优秀,在天然产物数据集上的表现也同样突出,通过结合模板与非模板策略,以及引入分子指纹结构和注意力机制,该方法显著提高了逆合成预测的准确性和可解释性。尤其值得一提的是,对第 4 类高难分子合成成功率的判断可以为多步逆合成研究提供新的范式。未来工作可以探索实时部署以及在未知反应空间中的扩展应用。

2.4 模型训练与特征分析

在模型训练阶段,本文基于 PyTorch 与 DGL(Deep Graph Library)框架实现了所提出的分子图分类模型。所有实验均在支持 CUDA 的 GPU 环境下进行训练,并使用 PyTorch 深度学习框架实现模型训练与推理过程。

在数据输入阶段,将产物分子转换为分子图结构,并通过 DGL 构建图数据。每个原子节点使用 80 维特征向量表示,其中包括原子类型、价态、电荷、杂化状态等基本化学属性;每条化学键使用 13 维特征向量表示,包含键类型、共轭信息以及是否处于环结构等特征,这些特征包含了化学分子的绝大部分信息,可以进行有效的分类训练。模型通过多步消息传递更新节点表示,使每个原子节点不仅包含自身的化学属性,还融合了其邻域结构的信息,从而能够有效识别潜在的反应中心和关键官能团,并结合分子指纹特征完成最终的分类预测。

在训练过程中,模型采用 Adam 优化器进行参数更新,初始学习率设置为 0.001,权重衰减设置为 0.000 01,用于缓解模型过拟合问题。为了进一步提高训练稳定性,在反向传播过程中对梯度进行了梯度裁剪,最大梯度范数设置为 20。学习率策略使用余弦退火方式动态调整,重启周期参数设置为 10,在训练批次设置方面,模型采用 batch size = 16 进行训练,模型最大训练轮次设置为 50 个 epoch。同时,为了防止模型过拟合,在训练过程中引入 Early Stopping 策略,当验证集性能在连续 5 个 epoch 内没有提升时停止训练,并保存最佳模型参数。

3 结束语

本文针对现有逆合成方法在泛化能力与化学有效性方面的局限,提出了一种融合分子指纹的 GNN 分类模型,旨在通过动态策略选择机制集成不同单步逆合成方法的优势,从而提升逆合成预测的准确率与泛化性能。研究中首先设计了一种多源特征融合的分子编码方案,将 MPNN 消息传递机制、AttentiveFP 网络与 ECFP4 分子指纹的信息有机结合,构建了同时刻画分子局部结构细节与全局关联特征的高质量分子表征,为后续分类决策模型提供了有效输入。在此基础上,以 LocalRetro 与 R-SMILES 作为基础逆合成模型,构建动态策略选择的分类框架,通过学习分子结构特征与不同方法适配性的内在关联,实现对目标分子最优逆合成策略的自动选择,实验结果表明该模型对分子核心片段具有较强的识别与判别能力。在天然产物数据集上的实验也进一步验证了模型的泛化性与实用价值,在高难度的第 4 类反应可合成性分类任务中,模型能够有效识别潜在不可合成分子,为实际合成实验提供可靠的决策支持。然而,本文仍存在一定局限性,例如基础模型类型尚未覆盖半模板方法,分子表示也主要依赖二维结构信息,未显式引入三维构象及反应环境特征。未来工作可从扩展基础模型体系、引入三维结构与反应条件信息等方面进一步完善模型能力,以提升其在复杂分子与真实反应场景中的适用性。随着模型体系的持续优化,该方法有望发展为有机合成设计中的重要辅助工具,推动合成化学向数据驱动与智能决策方向进一步演进。

参考文献:

[1] COLEY C W, ROGERS L, GREEN W H, *et al*. Computer-

assisted retrosynthesis based on molecular similarity[J]. ACS Central Science, 2017, 3(12): 1237-1245.

[2] SEGLER M H S, WALLER M P. Neural symbolic machine learning for retrosynthesis and reaction prediction[J]. Chemistry:A European Journal, 2017, 23(25): 5966-5971.

[3] CHEN B, LI C, DAI H, *et al*. Retro*: Learning retrosynthetic planning with neural guided A* search[C]//Proceedings of the 37th International Conference on Machine Learning. Vienna, Austria: PMLR, 2020: 1608-1616.

[4] DAI H, LI C, COLEY C, *et al*. Retrosynthesis prediction with conditional graph logic network[J]. Advances in Neural Information Processing Systems, 2019, 32: 8870-8880.

[5] CHEN S, JUNG Y. Deep retrosynthetic reaction prediction using local reactivity and global attention[J]. JACS Au, 2021, 1(10): 1612-1620.

[6] ISHIDA S, TERAYAMA K, KOJIMA R, *et al*. AI-driven synthetic route design incorporated with retrosynthesis knowledge[J]. Journal of Chemical Information and Modeling, 2022, 62(6): 1357-1367.

[7] XIE S, YAN R, GUO J, *et al*. Retrosynthesis prediction with local template retrieval[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Washington, DC, USA: AAAI Press, 2023: 5330-5338.

[8] LILLICRAP T P, SANTORO A, MARRIS L, *et al*. Backpropagation and the brain[J]. Nature Reviews Neuroscience, 2020, 21(6): 335-346.

[9] CAMACHO M T, PAEGELOW M, MAS J F, *et al*. Geomatic Approaches for Modeling Land Change Scenarios[M]. Cham: Springer, 2018: 1-8.

[10] 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.

[11] KIPF T N, WELLING M. Semi-superivised classification with graph convolutional networks [EB/OL]. (2016-09-09). https://arXiv.org/abs/1609.02907.

[12] VASWANI A, SHAZEER N, PARMAR N, *et al*. Attention is all you need[C]//Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017. Long Beach, CA, USA: Curran Associates Inc, 2017: 5998-6008.

[13] WEININGER D. SMILES: A chemical language and information system. Introduction to methodology and encoding rules[J]. Journal of Chemical Information and Computer Sciences, 1988, 28(1): 31-36.

[14] LIU B, RAMSUNDAR B, KAWTHEKAR P, *et al*. Retrosynthetic reaction prediction using neural sequence-to-sequence models[J]. ACS Central Science, 2017, 3(10): 1103-1113.

[15] KARPOV P, GODIN G, TETKO I V. A transformer model for retrosynthesis[C]//International Conference on Artificial Neural Networks. Cham: Springer International Publishing, 2019: 817-830.

[16] SCHWALLER P, LAINO T, GAUDIN T, *et al*. Molecular transformer: A model for uncertainty calibrated chemical reaction prediction[J]. ACS Central Science, 2019, 5(9): 1572-1583.

[17] ZHENG S, RAO J, ZHANG Z, *et al*. Predicting retrosynthetic reactions using self-corrected transformer neural networks[J]. Journal of Chemical Information and Modeling, 2019, 60(1): 47-55.

[18] TETKO I V, KARPOV P, VAN DEURSEN R, *et al*. State-of-the-art augmented NLP transformer models for direct and single-step retrosynthesis[J]. Nature CommuniCations, 2020, 11(1): 5575.

[19] WAN Y, HSIEH C Y, LIAO B, *et al*. Retroformer: Pushing the limits of end-to-end retrosynthesis transformer[C]//Proceedings of the 39th International Conference on Machine Learning. Baltimore, USA: PMLR, 2022: 22475-22490.

[20] ZHONG Z, SONG J, FENG Z, *et al*. Root-aligned SMILES: A tight representation for chemical reaction prediction[J]. Chemical Science, 2022, 13(31): 9023-9034.

[21] FANG L, LI J, ZHAO M, *et al*. Single-step retrosynthesis prediction by leveraging commonly preserved substructures[J]. Nature Communications, 2023, 14(1): 2446.

[22] ZHANG X, MO Y, WANG W, *et al*. Retrosynthesis prediction enhanced by in-silico reaction data augmentation[EB/OL]. (2024-01-31). https://arxiv.org/abs/2402.00086.

[23] LEVIN I, LIU M, VOIGT C A, *et al*. Merging enzymatic and synthetic chemistry with computational synthesis planning[J]. Nature Communications, 2022, 13(1): 7747.

[24] KIM T, LEE S, KWAK Y, *et al*. READRetro: Natural product biosynthesis predicting with retrieval-augmented dual-view retrosynthesis[J]. New Phytologist, 2024, 243(6): 2512-2527.

[25] LIN M H, TU Z, COLEY C W. Improving the performance of models for one-step retrosynthesis through re-ranking[J]. Journal of Cheminformatics, 2022, 14(1): 15.

[26] LI J, FANG L, LOU J G. RetroRanker: leveraging reaction changes to improve retrosynthesis prediction through re-ranking[J]. Journal of Cheminformatics, 2023, 15(1): 58.

[27] MAZIARZ K, LIU G, MISZTELA H, *et al*. Chemist-aligned retrosynthesis by ensembling diverse inductive bias models[EB/OL]. (2024-12-09). https://arvix.org/abs/2412.05269.

[28] MARCO A C, MYERS A, GRAHAM S J H, *et al*. The USPTO patent assignment dataset: Descriptions and analysis[R]. Alexandria, VA, USA: U.S. Patent and Trademark Office, 2015: 1-53.

[29] GOODMAN J. Computer software review: Reaxys[J]. Journal of Chemical Information and Modeling, 2009, 49(12): 2897-2898.

[30] ROGERS D, HAHN M. Extended-connectivity fingerprints[J]. Journal of Chemical Information and Modeling, 2010, 50(5): 742-754.

Graph Neural Network-Based Ensemble Methods for Single-Step Retrosynthesis

WU Hexing[1], REN Fuyue[1], PENG Yu[1], ZHANG Pengpeng[2], LU Jingyi[1]

(1. School of Information Science and Engineering, East China University of Science and Technology, Shanghai 200237, China; 2. College of Pharmaceutical Sciences, Zhejiang University, Hangzhou 310058, China)

Abstract: Retrosynthetic analysis is the process of deriving commercially available precursor molecules from target product molecules. To address the limitations of existing single-step chemical retrosynthetic methods in terms of generalization ability and prediction accuracy, this paper proposes an ensemble prediction model integrating molecular fingerprints and graph neural networks, which addresses the poor scalability and interpretability of individual models. This method improves the accuracy and stability of retrosynthetic prediction by constructing multidimensional molecular representations and dynamically selecting the optimal prediction strategy. The model is based on message-passing neural networks (MPNNs), global self-attention mechanisms, and AttentiveFP networks to extract molecular structural features, and introduces extended connectivity fingerprints (ECFP) to achieve joint modeling of local structural information and global molecular features, constructing two fusion architectures: MPNN+FPS (Fingerprints) and Attentive FP+FPS. Extensive experiments on the USPTO-50K and natural-product datasets validate the effectiveness of the method. The proposed ensemble retrosynthetic prediction method demonstrates significant advantages in both prediction accuracy and generalization ability, providing a new technical approach for retrosynthetic analysis of complex molecules and natural products.

Key words: retrosynthetic reaction prediction; graph neural networks; molecular fingerprint; product classification; feature fusion; dynamic strategy selection

（责任编辑：张欣）