

柔性-环增强图神经网络用于中药分子复杂结构的性质预测

贺 怀¹, 陈治瑜², 任 琪², 陈双扣^{1*}, 钟 节², 陶梦瑶², 程榆淞², 代浚豪², 周寰宇², 刘 曾², 戴传云^{2*}

1. 重庆科技大学化学化工学院, 重庆 401331

2. 重庆中医药学院中药学院, 制药过程数字化重庆市重点实验室, 重庆 402760

摘要: 目的 针对中药 (traditional Chinese medicine, TCM) 分子结构复杂、环系结构多样等特点, 优化图同构神经网络 (graph isomorphism network, GIN), 提升其对中药分子关键性质的预测能力。**方法** 基于 GIN 模型, 在分子图中引入表示键长张力和角度张力的边描述符, 并结合多个优势构象构建多图数据, 从而构建了一种柔性-环增强图神经网络 (flexibility-ring enhanced graph neural network, FRGNN)。在包含 37 822 个分子的 2 个中药数据库中, 评估了该模型在 7 项关键分子属性上的预测性能, 并与 3 种当前最优 (state-of-the-art, SOTA) 图神经网络模型及 2 种基础图神经网络模型进行了对比。**结果** 与表现次优的模型相比, FRGNN 在 7 项性质预测上的均方根误差平均降低了 8.63%; 对于含多环及大环结构的分子, 其预测均方根误差进一步降低 10.04%。**结论** FRGNN 模型在中药分子关键性质预测任务中表现出优于现有小分子预测 SOTA 模型的性能, 为复杂天然化合物的性质预测提供了新的有效方法。

关键词: 人工智能; 图神经网络; 中药分子; 中药数据库; 性质预测

中图分类号: R284.1 文献标志码: A 文章编号: 0253-2670(2026)16-6260-13

DOI: 10.7501/j.issn.0253-2670.2026.16.005

Property prediction of complex traditional Chinese medicine molecular structures using a flexibility-ring enhanced graph neural network

HE Huai¹, CHEN Zhiyu², REN Qi², CHEN Shuangkou¹, ZHONG Jie², TAO Mengyao², CHENG Yusong², DAI Junhao², ZHOU Huanyu², LIU Zeng², DAI Chuanyun²

1. School of Chemistry and Chemical Engineering, Chongqing University of Science and Technology, Chongqing 401331, China

2. Chongqing Key Laboratory of Digitalization of Pharmaceutical Processes, College of Traditional Chinese Medicine, Chongqing University of Chinese Medicine, Chongqing 402760, China

Abstract: Objective To address the complex molecular structures and diverse ring systems characteristic of traditional Chinese medicine (TCM) compounds, this study aims to optimize the graph isomorphism network (GIN) to enhance its predictive capability for key molecular properties of TCM molecules. **Methods** Based on the GIN architecture, edge descriptors representing bond length strain and angle strain were introduced into the molecular graph, and multi-graph data were constructed by incorporating multiple favorable conformations, thereby developing a flexibility-ring enhanced graph neural network (FRGNN). The predictive performance of the proposed model on seven key molecular properties was evaluated using two TCM databases containing 37 822 molecules, and compared against three state-of-the-art (SOTA) graph neural network models and two basic graph neural network models. **Results** Compared with the second-best performing model, FRGNN achieved an average reduction of 8.63% in root mean square error across the seven property prediction tasks, for molecules containing polycyclic and macrocyclic structures, the predictive root mean square error was further reduced by 10.04%. **Conclusion** The proposed FRGNN model demonstrates superior performance over existing state-of-the-art models for small-molecule property prediction in the task of predicting key properties of TCM compounds, providing

收稿日期: 2026-04-06

基金项目: 国家自然科学基金面上项目 (22578039); 重庆市璧山区中医药领航联合专项重大项目 (BSZYHLH001); 重庆市自然科学基金创新发展联合基金 (CSTB2023NSCQ-LZX0060)

作者简介: 贺 怀 (1999—), 男, 硕士研究生, 研究方向为化学场景下的人工智能应用。E-mail: 2727889123@qq.com

*通信作者: 陈双扣 (1976—), 男, 博士, 教授, 研究方向为药物设计及药物递送。E-mail: cskcn@yeah.net

戴传云 (1976—), 男, 博士, 教授, 研究方向为干湿结合场景下的人工智能在药剂学中的应用。E-mail: cydai@126.com

a novel and effective approach for the property prediction of complex natural products.

Key words: artificial intelligence; graph neural networks; traditional Chinese medicine molecules; Chinese medicine datasets; property prediction

准确获得中药化学成分的理化性质与生物活性是阐明其药效物质基础的关键环节。这对于揭示中药作用机制、优化制剂工艺以及推动中药现代化进程具有重要意义^[1-2]。然而,中药化学成分复杂多样^[3-4],性质各异^[5-6],且绝大多数天然产物的骨架和结构具有高度多样性和复杂性。传统的实验方法如药理实验^[7]、气相色谱^[8]以及液相色谱^[9-10]等,虽能在一定程度上揭示中药成分的性质,但面对中药庞大而复杂的成分体系,这些方法往往耗时费力,难以全面覆盖所有成分^[11]。因此,亟需开发一种高效、准确的中药成分性质预测方法。

机器学习作为人工智能的重要分支,凭借其在数据拟合与模式识别方面的强大能力,已在有机合成^[12-15]、新材料研究^[16]以及药物发现^[17-21]等领域取得突破性进展。这些进展为解析中药复杂体系下的活性成分、阐释多靶点作用机制以及开展分子水平的表征与优化,提供了全新的研究方法与技术支撑^[22]。其中,能够有效处理图结构的图神经网络(graph neural network, GNN)^[23],因其出色的数据拟合性能,在分子性质预测领域持续受到广泛关注。

然而,传统 GNN 模型将分子视为仅由化学键连接的拓扑结构,忽略了键长与键角等几何信息^[24-25]。部分研究尝试引入键长、键角作为几何特征^[26-27],并在 MoleculeNet^[28]等小分子数据集上取得了较好的预测性能,但中药来源的天然产物往往包含多元环、螺环、大环等复杂结构,具备多构象和柔性特征^[29],这些几何特征对其生物活性与理化性质至关重要,如青蒿素的过氧桥键的笼状骨架^[30]以及雷帕霉素的构象互变现象^[31]。现有方法未能显式感知环状结构带来的分子张力,也缺乏环张力明确定义与编码,因而难以实现对该类复杂分子的准确预测。

在多构象建模方面,OG-GNN^[32]将环作为独立节点,但仅编码环的存在性而忽略环内应变;SCAGE 等^[33-34]法通过大规模预训练或注意力机制实现构象感知,在通用小分子数据上表现良好,但存在以下局限:(1)依赖海量预训练数据和复杂注意力机制,在中药样本有限时易发生过拟合;(2)多构象融合引入大量可学习参数,模型复杂度高;(3)缺乏对环张力等中药核心化学特征的高效表征。因

此,本研究提出一种基于分子内部环-柔性增强的图神经网络(flexibility-ring enhanced graph neural network, FRGNN),用于建模环结构与多构象对分子性质的影响。该模型不引入复杂的注意力机制,也无需大规模预训练。与现有工作的本质差异体现在以下 2 个维度:其一为多构象简约融合。基于中药分子多低能构象贡献化学等价的先验^[35],在 GNN 框架内对多构象隐藏特征进行节点级加和融合,不引入额外可学习参数;其二为环张力显式编码。通过构建无环参考分子体系,将键长张力和角度张力作为独立能量型先验特征直接输入网络,而非依赖模型从键角数值中隐式推断。

为了验证所构建的模型的有效性,本研究选取 2 大中药数据库 TCMSP^[36]和 HERB2^[37],共包含 37 822 个分子及 7 项去重后的分子性质,构建训练、验证和测试集,比较 FRGNN 与多种基线模型的预测性能。此外,本研究设计了针对含环化合物的挑战性预测任务,即从上述数据库中筛选出富含多种环结构的化合物,分布采用几种现有最优(state-of-the-art, SOTA)模型和 FRGNN 模型进行比较。综上所述,本研究提出的 FRGNN 模型通过增强对环与柔性结构的表征能力,为中药分子性质预测、活性成分的高通量虚拟筛选、吸收(absorption)、分布(distribution)、代谢(metabolism)、排泄(excretion)和毒性(toxicity)性质预评估,以及含环中药分子的构象-活性关系解析提供了新方法,也为中医药现代化研究提供了新的计算范式。

1 材料与方法

1.1 计算机硬件与系统

Linux 操作系统为 CentOS Linux release 7.9.2009 (Core),CPU 架构为 X86_64, GPU 采用 NVIDIA RTX 3090 (显存 20 GB)。计算语言为 Python 3.8,所有模型均于 CUDA 12.6 环境下运行。

1.2 数据集与预测任务

为了评估模型在中药数据集上的预测性能,下载了 2 个数据集 TCMSP^[36]和 HERB2^[37]。

TCMSP 数据集涵盖 29 384 个分子,每个分子均提供多项理化与药动学属性包括相对分子质量(molecular weight, M_w)、辛醇-水分配系数(octanol-

water partition coefficient, $AlogP$)、氢键供体数 (hydrogen bond donor, H_{dom})、氢键受体数 (hydrogen bond acceptor, H_{acc})、口服生物利用度 (oral bioavailability, OB)、Caco-2 细胞通透性 (P_{Caco-2})、血脑屏障穿透性 (blood-brain barrier, BBB)、类药性 (drug-likeness, DL)、带负部分电荷的所有原子的水可及表面积分数 (fractional water accessible surface area of all atoms with negative partial charge, FASA)、可自由旋转的键的数量 (rotatable bonds number, RBN) 以及拓扑极性表面积 (topological polar surface area, TPSA)。由于 $AlogP$ 、OB、 P_{Caco-2} 、BBB、FASA、DL 以及 TPSA 7 个性质对中药的研究过程至关重要, 且实验测定成本较高, 本研究将其选定为预测任务, 并筛选得到 6 904 个分子。

HERB2 数据集包含 49 258 个分子, 每个分子所含原子的平均数量为 29.97, 涵盖了诸如 DL、OB、 $AlogP$ 、RBN、 H_{dom} 、 H_{acc} 等属性。依据上述原则, 选取 DL、OB 以及 $AlogP$ 3 项性质作为预测任务 (表 1), 并筛选得到 30 918 个分子。

表 1 数据集详细属性
Table 1 Details of datasets

数据集名称	分子总数	分子的平均原子数	标签数
TCMSP	6 904	21.42	7
HERB2	30 918	29.97	3

1.3 数据集划分

按骨架划分^[38] (Scaffold split) 是一种基于分子骨架的数据集划分方法, 为了准确评估模型性能, 本研究采用 Scaffold split 对所有数据集进行划分,

按 80%、10%、10% 划分为训练集、验证集和测试集 (表 2), 其对应的损失函数分别称为训练损失 (train loss)、验证损失 (valid loss) 和测试损失 (test loss)。

表 2 划分数据集的详细信息
Table 2 Details of each split dataset

数据集名称	数据集属性	训练集	验证集	测试集
TCMSP	分子数量	5 524	690	690
	分子平均原子数	22.51	18.62	15.50
	标签值均值	12.411	11.830	13.682
	标签值标准差	14.221	14.961	17.632
HERB2	分子数量	24 734	3 092	3 092
	分子平均原子数	20.31	21.03	30.69
	标签值均值	7.523	7.956	15.20
	标签值标准差	7.201	7.110	7.440

1.4 特征提取

特征提取方式如图 1 所示, 首先将分子的简化分子线性输入规范 (simplified molecular input line entry system, SMILES)^[39] 进行多构象建模, 再进行能量最低优化, 选择能量最低的几个作为优势构象, 随后对选取的优势构象分别做节点和边的特征计算, 最后计算分子内部各个化学键的长度张力和角度张力, 从而完成特征提取。

1.4.1 多构象构建 通过 rdkit (<https://www.rdkit.org/>) 中的 EmbedMultipleConfs 模块生成大量分子的初始构象, 设置最大构象数为 10, 其余设置均为默认值, 即使用基于实验扭转角知识距离几何 (experimental torsion knowledge distance geometry, ETKDG) 算法, 这是 RDKit 用于生成三维构象的

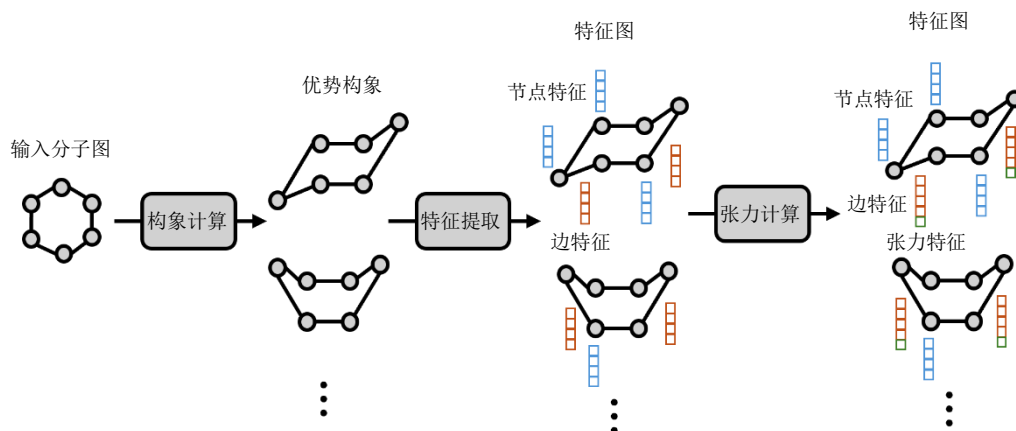


图 1 特征提取流程

Fig. 1 Process of feature extraction

算法,通过距离几何和来自晶体数据库的扭转角偏好来构建初始三维结构,以充分覆盖构象空间。随后使用 MMFFOptimizeMoleculeConfs 函数对各个构象进行优化[该模块的优化器、收敛条件以及最大迭代步数均为默认值,即使用 Limited-memory Broyden-Fletcher-Goldfarb-Shanno(L-BFGS),它是一种拟牛顿优化算法,通过近似 Hessian 矩阵来高效寻找函数的极小值,优化器收敛条件为力场默认的最大受力 1×10^{-8} kcal/(mol·Å) ($1 \text{ Å} = 0.1 \text{ nm}$),最大迭代步数 200],再通过 Merck Molecular Force

Field 94 (MMFF94) 对构象进行优化,其通过键伸缩、角度弯曲、二面角扭转、范德华及静电等相互作用项计算构象的总势能,选择能量最低的几个构象作为后续特征提取的优势构象。

1.4.2 节点和边的特征计算 使用 rdkit 内置的 Getatoms 函数将分子中的原子转换为包含原子序数、杂化形式、是否在环中等性质的向量,再使用 Getbonds 将化学键被转换成键长、键级、键角以及二面角等包含几何结构的向量。具体的特征描述符如表 3 所示。

表 3 特征描述符详细信息
Table 3 Details of feature descriptors

特征级别	特征类型	特征描述符	数据类型	大小
原子级	原子特征	原子半径、电负性等	独热向量	92
化学键级	键方向	无方向、实心楔形键、虚线楔形键等	独热向量	7
	键级	单键、双键、三键、芳香键	独热向量	4
	键长	化学键的长度	浮点数	1
	是否在环中	化学键是否在环中	独热向量	2
	键角	键角	浮点数	1
	二面角	二面角	浮点数	1
	长度张力	长度张力	浮点数	1
	角度张力	角度张力	浮点数	1

1.4.3 键长、键角张力构建 键长张力的构建同样采用 RDKit 模块及其内部模块。首先通过构建与目标化学键具有相同原子类型和键级的无环开链参考分子(如碳碳双键,以乙烯作为参考分子),并使用 AddHs 函数增加氢原子,再采用内置 ETKDG 算法通过 EmbedMolecule 函数生成初始三维构象,随后利用 RDKit 内置的通用力场(universal force field,UFF)对其进行能量极小化几何优化。从优化后的构象中提取原子坐标,计算目标键的 2 原子间欧氏距离作为无张力键长特征。最后计算“1.4.1”项下构建的构象的所有键长与无张力键长相减得到键长张力。

无张力参考键角的构建同样基于 RDKit 模块实现。首先定位目标中心原子(即具有 2 个及以上相邻原子的非氢原子),通过构建与目标化学键角度具有相同原子类型和键级的无环开链参考分子(如 -CH-O-CH- 结构,以甲醚作为参考分子)。使用 AddHs 方法添加显式氢原子,采用 ETKDG 算法通过 EmbedMolecule 方法生成初始三维构象,利用 RDKit 内置的通用力场 UFF 对其进行能量极小化几何优化。从优化后的构象中,以目标中心原子为顶点、2 个相邻原子为端点构成 2 个向量,由向量

内积直接计算键角的余弦值。

特别的是,对于芳香体系,本研究采取六元环作为参考分子,尽管六元环分子存在环状结构,但由于其是标准的正六边形,所以其内部不存在(或极小)环张力,且芳香化合物的键长与普通的离域体系也有不同,所以采取六元环分子作为参考分子。

将无张力化学键长度和分子中所有的化学键的长度做差来建模长度张力,按公式(1)具体计算。角度张力的构建如公式(2)所示

$$T_i^{\text{bl}} = |L_i^s - L_i|, i \in B \quad (1)$$

B 代表分子中的化学键的集合, T_i^{bl} 代表化学键 i 的键长偏差, L_i^s 代表第 i 个化学键处于无张力状态下的长度, L_i 为第 i 个化学键的长度

$$T_i^{\text{ba}} = \sum_{j=1}^{j \in N_i} |a_{ij}^s - a_{ij}| / A, i \in B, j \in N_i \quad (2)$$

T_i^{ba} 代表化学键 i 所受到的键角偏差, a_{ij}^s 代表化学键 i 和 j 不存在环张力情况下的键角, a_{ij} 代表化学键 i 和 j 的键角, N_i 代表的是与化学键 i 相邻的所有化学键, A 表示以节点 i 为起点的化学键的总数

所有的角度偏差均取余弦值作为角度偏差的数字特征。最后使用 pg1[40]模块的 Graph 函数将原子、键包装为图数据完成特征提取。

1.5 模型构建

如图 2 所示, 首先多图特征经过 RBF (radial basis function) 进行快速升维, 随后经过一个 GIN 层, 节点在 GIN 层内通过相邻节点的逐元素相加, 再经过 2 个归一化层 LayerNorm^[41]和 GraphNorm^[42]完成归一化。再经过激活函数 Leaky ReLU 完成激活, 再经历 1 个 dropout 层随机丢弃部分特征防止过拟合。与该层最初的特征直接相加完成残差连接, 然后逐元素相加完成多图数据的融合, 最后传入下一层。其中威斯费勒-莱曼 (Weisfeiler-Lehman, WL) 核迭代次数即模型层数。在完成训练后将得到的隐藏特征输出至一个多层感知机 (multilayer perceptron, MLP) 进行降维, 读出函数内置在 MLP 中, 所采用的方法为平均池化。MLP 共 3 层, 其中前 2 层为隐藏层, 由线性层、Dropout 层和 Leaky ReLU 依次构成; 第 3 层为输出层, 由线性层组成, 将隐藏特征直接映射到目标属性数量。对于分类任务, 在线性层后再增加 1 个 Sigmoid 函数将输出值映射到 0~1。模型在训练的过程中采用 Adam 优化器^[43]自动调整模型参数。

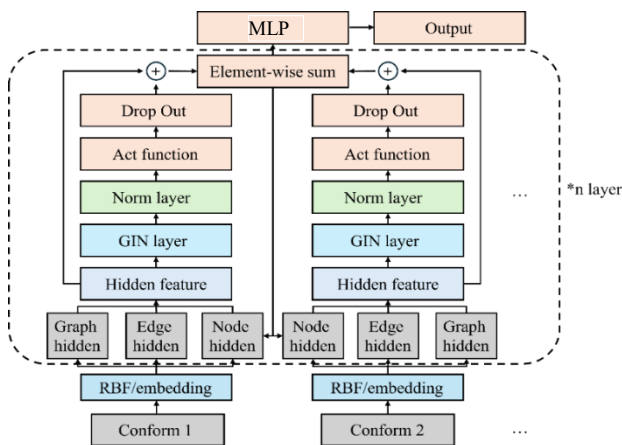


图 2 模型架构

Fig. 2 Architecture of model

1.6 评价指标

由于平均绝对误差 (mean absolute error, MAE) 对误差相对不敏感, 不会因为个别的极端样本而影响全局的优化方向, 故选择 MAE 作为训练的损失函数。其计算公式如式 (3) 所示。

$$\text{MAE} = \sum_{i=1}^n (y_i - \bar{y}_i) / n \quad (3)$$

n 为数据集子集样本的数量, y_i 为第 i 个样本的标签值, $\bar{y}_i$ 为模型对第 i 个样本的预测值

均方根误差 (root mean square error, RMSE)

是用来计算预测值和真实值之间的差距, RMSE 越大, 则代表模型预测的越不准确, 反之则代表预测准确。由于其能够直观反映模型的性能, 所以本研究采用 RMSE 作为验证集和测试集的评价指标, 按公式 (4) 计算。在训练过程中, 首先记录每个轮次的训练集和验证集的 RMSE, 最后选择验证集上 RMSE 最小的轮次作为最佳模型, 然后使用测试集对最佳模型进行 RMSE 计算, 得到模型测试集 RMSE。

$$\text{RMSE} = [\sum_{i=1}^n (y_i - \bar{y}_i)^2 / n]^{1/2} \quad (4)$$

n 为数据集子集样本的数量, y_i 为第 i 个样本的标签值, $\bar{y}_i$ 为模型对第 i 个样本的预测值, 式中为取均值操作

模型的提升比例计算方法是 FRGNN 与次优模型的 RMSE 做差, 再与次优模型的 RMSE 做商, 得到 FRGNN 的提升比例。计算公式如式 (5) 所示。

$$\text{提升比例} = (\text{SOTA-FRGNN}) / \text{SOTA} \quad (5)$$

1.7 超参数设置

由于数据集的样本数量充足, 所以本研究参考 Fang 等^[26]和 Li 等^[27]所使用的超参数空间搜索的方式对模型进行优化。具体的超参数名称及取值如表 4 所示。每个模型都在如下超参数空间下独立 100 个轮次, 独立运行 4 次取均值和标准差作为最终结果。

表 4 超参数设置

Table 4 Settings of hyperparameters

超参数名称	值
丢弃率 (dropout rate)	0.2, 0.5
学习率 (learning rate)	1e-4, 1e-3
模型层数 (layers)	1, 3, 5, 7, 8, 9, 11, 13
隐藏层大小 (hidden dim)	32, 64, 128, 256
优势构象数 (conformers)	1, 2, 3, 4, 5

2 结果与讨论

2.1 数据集属性

2.1.1 标签值分布 图 3 为 TCMSP 数据库 7 个属性的标签值, 结果表明 TPSA、AlogP、 $P_{\text{Caco-2}}$ 以及 BBB 4 个标签值的总体分布较为集中, 但存在一些极端值, 其他标签值的分布较为宽泛。

图 4 为 HERB2 数据集的 3 个属性的标签值, 其 AlogP 标签值较为集中, 但其存在的极端标签值比 TCMSP 数据集中的 AlogP 标签值更大, 这是因为 HERB2 数据集所包含的分子更多, 所以性质差异也更大; 其余的标签值则同样比较宽泛, 这是因为中药组分所包含的分子结构和性质差异较大, 所以导致标签值的分布相比于其他的数据库如 MoleculeNet^[28]比较宽。

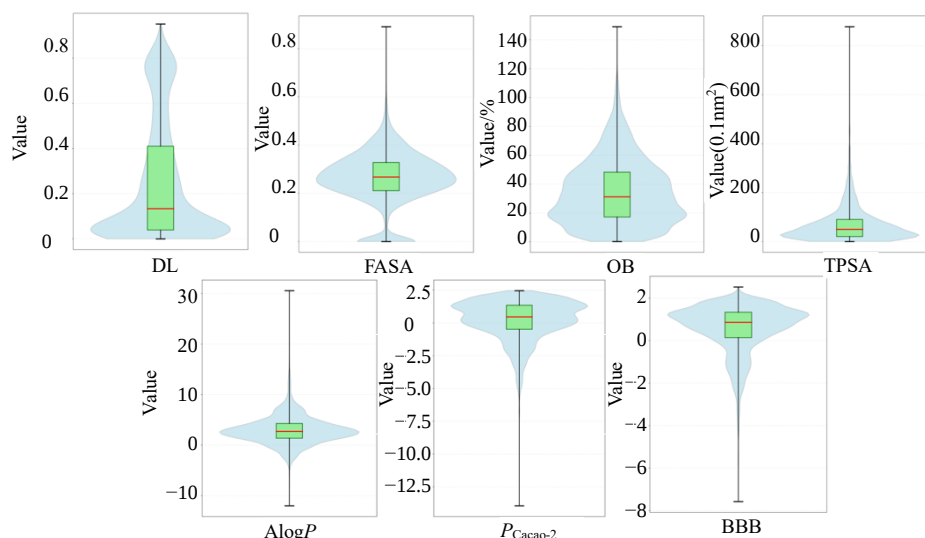


图 3 TCMSP 数据集各个任务标签值分布
Fig. 3 Distribution of TCMSP each task label value

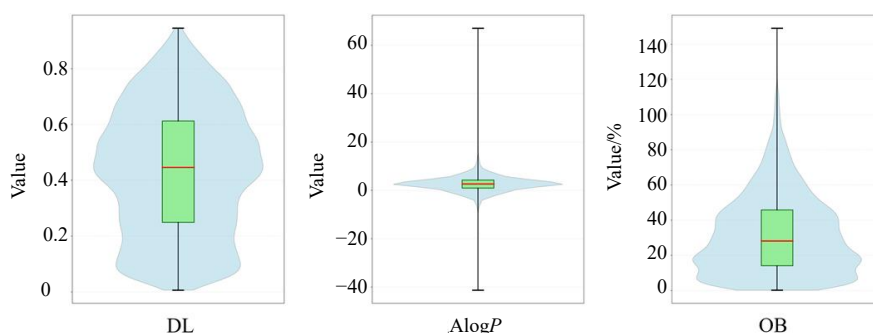


图 4 HERB2 数据集各个任务标签值分布
Fig. 4 Distribution of HERB2 each task label value

2.1.2 数据集含环化合物的结构分布信息 为了详细比较中药数据集和通用小分子数据集中分子的环结构差异,本研究分别统计了中药数据集 TCMSP、HERB2 以及通用数据库 MoleculeNet 上生物活性数据集的分子结构,如表 5~7 所示。在中药数据集中,桥环和螺环化合物比例远高于通用数据集上的比例。同时,在环大小上看,中药数据集的小环和大环化合物比例也高于通用数据集。在环数量上看,通用数据集中的分子更多集中于在含 3 个环附近,而中药数据集中的分子则是在各个数量的环中都分布比较均匀。这说明中药数据集的分子相比于通用数据集中的分子的环数量更多、结构更复杂、环张力更大。

2.2 超参数调整

2.2.1 学习率和丢弃率敏感性 学习率控制着模型的训练行为,过小则会陷入局部最优^[44],学习率

表 5 各数据集环类型分布

数据库集	环类别/%		
	含环化合物	桥环化合物	螺环化合物
TCMSP+HERB2	89.52	11.78	11.20
MoleculeNet	96.87	3.12	2.35

表 6 各数据集环大小分布

数据库集	环大小/%					
	3 元环	4 元环	5 元环	6 元环	7 元环	8 元及以上
TCMSP+	5.60	1.42	42.04	82.22	4.60	5.45
HERB2						
MoleculeNet	1.60	1.12	57.22	92.52	2.85	1.16

过大会导致模型跳过全局最优解而欠拟合,而丢弃率控制着模型在训练过程中所能获得的特征,丢弃率过小会导致模型学习到更多无关的特征而导致

表 7 各数据集环数量分布

Table 7 Distribution of ring number across datasets

数据库集	环数量/%					
	1	2	3	4	5	≥6
TCMSP+	12.88	13.84	18.43	16.13	11.53	16.71
HERB2						
MoleculeNet	8.65	24.37	32.80	21.31	6.73	3.01

表 8 模型验证及测试损失函数随丢弃率和学习率的变化

Table 8 Variations in model valid loss and test loss with learning rate and dropout rate

数据集名称	Head lr	Encoder lr	Dropout rate	valid loss	test loss
TCMSP	0.001 0	0.001 0	0.2	4.036±0.023	3.845±0.040
	0.000 1	0.001 0	0.2	5.073±0.021	3.344±0.032
	0.001 0	0.000 1	0.2	4.004±0.026	3.252±0.031
	0.000 1	0.000 1	0.2	4.230±0.031	4.064±0.021
	0.001 0	0.001 0	0.5	4.615±0.020	3.891±0.043
	0.000 1	0.001 0	0.5	5.231±0.019	5.985±0.021
	0.001 0	0.000 1	0.5	3.980±0.019	4.690±0.013
	0.000 1	0.000 1	0.5	5.364±0.013	5.566±0.020
HERB2	0.001 0	0.001 0	0.2	0.115±0.001	1.344±0.002
	0.000 1	0.001 0	0.2	0.180±0.001	1.120±0.002
	0.001 0	0.000 1	0.2	0.122±0.001	0.639±0.002
	0.000 1	0.000 1	0.2	0.182±0.001	0.797±0.003
	0.001 0	0.001 0	0.5	0.120±0.001	0.621±0.002
	0.000 1	0.001 0	0.5	0.205±0.001	0.805±0.002
	0.001 0	0.000 1	0.5	0.130±0.001	0.326±0.002
	0.000 1	0.000 1	0.5	0.188±0.001	0.527±0.002

Encoder lr=0.000 1, Dropout rate=0.5, 在 HERB2 上的丢弃率更大是因为数据量更充足, 所以带来的特征也更多, 为了不过拟合所以需要更大的丢弃率^[27]。

2.2.2 隐藏层大小敏感性 隐藏层大小控制着模型的复杂度, 隐藏层越高, 就能将分子的特征映射到越高的维度, 会使得模型对复杂数据具备更好的分析能力, 但是隐藏层的提升也会导致模型的参数增加, 运行时间变慢, 所以需要在运行时间和模型性能之间做平衡。模型均在隐藏层维度为 128 时表现最好, 这也是图神经网络在分子性质预测方面的 1 个通用值^[26-27], 符合过往深度学习研究的规律和经验。模型性能随隐藏层大小变化的具体信息见表 9。

2.2.3 模型层数敏感性 模型层数控制着模型的深度, 越深的模型越能够捕捉到更加细致的特征, 但是模型过深也会导致梯度平滑或者梯度爆炸, 为了避免这一问题, 本研究在模型中增加了残差连接防止梯度平滑和梯度爆炸^[29]。模型层数对性能的影响如表 10 所示, 模型在层数为 8 时表现出最佳性

过拟合, 过大会导致模型学习不到足够的对编码器学习率 (encoder learning rate, Encoder lr)、输出头学习率 (head learning rate, Head lr) 和丢弃率 (dropout rate) 进行了详细的搜索。结果如表 8 所示, FRGNN 在 TCMSP 和 HERB2 数据集上的最佳学习率-丢弃率组合分别为 Head lr=0.001, Encoder lr=0.000 1, Dropout rate=0.2 和 Head lr=0.001,

表 9 模型预测性能随隐藏层大小的变化

Table 9 Variations in model predictive performance with hidden size

隐藏层 大小	TCMSP		HERB2	
	valid loss	test loss	valid loss	test loss
32	8.817±0.033	6.909±0.033	0.172±0.001	2.176±0.003
64	6.693±0.029	4.340±0.029	0.132±0.002	1.194±0.002
128	4.004±0.026	3.252±0.031	0.130±0.001	0.326±0.002
256	4.682±0.023	4.782±0.036	0.135±0.002	2.136±0.002

能, 这与以 GIN 模型为基础的模型的超参数设置也是一致的, 符合过往图神经网络模型的超参数设置规律^[26-27]。

2.2.4 构象数敏感性 本研究比较了多图构建时的优势构象数量对模型的影响, 结果如表 11 所示, 结果表明, 选取能量最低的 2 个优势构象时, 模型能够达到最佳预测性能, 这表明过多的构象数量会导致模型将某些含量极低的构象作为学习对象, 从而带来不必要的冗余信息, 导致预测性能的下降。

表 10 模型预测性能随模型层数的变化

Table 10 Variations in model predictive performance with number of layers

模型层数	TCMSP		HERB2	
	valid loss	test loss	valid loss	test loss
1	6.644±0.023	5.666±0.004	0.170±0.002	1.573±0.002
3	5.330±0.017	3.623±0.005	0.215±0.002	1.053±0.002
5	5.129±0.015	5.409±0.004	0.617±0.002	1.204±0.003
7	5.161±0.013	3.992±0.003	0.215±0.006	1.708±0.004
8	4.004±0.026	3.252±0.031	0.130±0.001	0.326±0.002
9	6.259±0.026	5.009±0.004	0.201±0.003	0.997±0.002
11	5.659±0.022	3.906±0.003	0.215±0.003	0.708±0.002
13	4.820±0.018	5.167±0.004	0.795±0.002	0.566±0.001

表 11 模型预测性能随构象数量的变化

Table 11 Variations in model predictive performance with number of conforms

优势构象数	TCMSP		HERB2	
	valid loss	test loss	valid loss	test loss
1	4.674±0.027	3.511±0.032	0.097±0.002	0.384±0.003
2	4.004±0.026	3.252±0.031	0.130±0.001	0.326±0.002
3	4.315±0.033	3.411±0.020	0.102±0.002	0.335±0.002
4	4.476±0.042	3.672±0.022	0.102±0.003	0.353±0.006
5	4.554±0.051	3.851±0.047	0.111±0.005	0.361±0.004

2.3 模型性能比较

本研究将 FRGNN 和其他 8 个模型的总体预测性能进行了比较, 其中, GIN^[24,27]、GCN^[25]为 2 种基础的 GNN 模型, GEM^[26]、SMPT^[27]为 2 种针对分子空间结构进行改进的模型, KA-GNN^[45]是一种对模型计算过程进行改进的模型, $\mathcal{O}$ -GNN^[32]和

RSE^[46]是为专门针对环结构改进的模型, geom2vec^[47]是经过空间预训练的模型。为确保公平对比, 所有对比测试中的模型均在其已提出的超参数优化方法进行优化^[24-27,32,45-47]。结果如表 12、13 所示, 下划线为次最优模型结果, 以 RMSE 作为评价指标, FRGNN 在 2 个数据集的各项预测任务中均取得了最小的 RMSE, 且与次最优模型相比平均实现 8.63%的性能提升, 这是由于 FRGNN 能够很好的捕捉到中药分子内部的环状、多构象以及具备柔性的特点, 所以表现出比其他模型更好的预测性能^[23-30], 其能够很好的完成中药分子的性质预测。 t 检验结果也表明, 除 TCMSP 中的 DL 这一属性外其余属性预测的 P 值都小于 0.05, 说明 FRGNN 的性能提升相比于次优模型是显著的。而 DL 性质预测的不显著, 这可能是因为 TCMSP 的数据量小,

表 12 FRGNN 与其他最先进模型在 TCMSP 数据集上预测性能比较

Table 12 Comparison of predictive performance between FRGNN and SOTA models on TCMSP dataset

模型/ 统计量	AlogP	OB	P_{Caco2}	BBB	DL	FASA	TPSA
GIN	0.645±0.031	22.4±0.2	0.237±0.005	0.510±0.004	0.082±0.003	0.123±0.006	15.7±2.1
GCN	0.660±0.024	20.0±0.1	<u>0.192</u> ±0.001	0.453±0.003	0.090±0.002	0.094±0.003	16.5±2.4
GEM	0.542±0.027	19.2±0.2	0.193±0.001	0.387±0.002	0.063±0.002	0.098±0.002	12.3±0.3
SMPT	0.574±0.022	18.8±0.2	0.212±0.001	<u>0.357</u> ±0.002	<u>0.061</u> ±0.005	<u>0.088</u> ±0.002	20.6±0.3
KA-GNN	0.624±0.019	19.5±0.2	0.236±0.001	0.391±0.005	0.069±0.003	0.105±0.006	22.6±0.3
$\mathcal{O}$ -GNN	0.752±0.095	20.4±0.6	0.224±0.009	0.542±0.061	0.077±0.003	0.107±0.003	13.9±0.8
RSE	<u>0.531</u> ±0.010	18.3±2.2	0.215±0.058	0.362±0.063	0.068±0.009	0.088±0.002	<u>11.5</u> ±0.7
geom2vec	0.713±0.022	<u>17.9</u> ±0.4	0.260±0.023	0.418±0.014	0.101±0.004	0.092±0.004	19.6±4.6
FRGNN	0.492±0.021	17.2±0.2	0.174±0.001	0.292±0.004	0.058±0.002	0.081±0.001	10.2±0.1
提升比例	7.34%	3.91%	9.38%	18.2%	4.92%	7.95%	11.3%
t 检验 P 值	<0.05	<0.05	<0.05	<0.05	0.29	<0.05	<0.05

表 13 FRGNN 与其他最先进模型在 HERB2 数据集上预测性能比较

Table 13 Comparison of predictive performance between FRGNN and SOTA models on HERB2 dataset

模型/统计量	AlogP	DL	OB
GIN	0.953±0.021	0.053±0.001	10.60±0.21
GCN	0.953±0.017	0.098±0.001	12.00±0.19
GEM	0.423±0.018	0.052±0.001	<u>9.60</u> ±0.24
SMPT	<u>0.361</u> ±0.020	<u>0.049</u> ±0.001	11.40±0.33
KA-GNN	0.371±0.022	0.053±0.001	10.20±0.23
$\mathcal{O}$ -GNN	1.170±0.161	0.098±0.004	22.50±0.33
RSE	0.391±0.017	0.052±0.001	17.70±0.37
geom2vec	0.741±0.053	0.063±0.022	18.50±0.37
FRGNN	0.330±0.011	0.045±0.001	9.01±0.19
提升比例	8.59%	8.16%	6.54%
t 检验 P 值	<0.05	<0.05	<0.05

表 14 FRGNN 与其他最先进模型对含不同环大小化合物预测性能的比较

Table 14 Comparison of predictive performance between FRGNN with SOTA models in predicting different ring size compounds

模型名称	三元环	四元环	五元环	六元环	七元环及以上
GEM	4.015±0.026	3.975±0.019	4.128±0.028	3.854±0.027	3.851±0.028
SMPT	<u>3.922</u> ±0.028	<u>3.857</u> ±0.027	<u>3.855</u> ±0.026	<u>3.749</u> ±0.032	<u>3.810</u> ±0.027
KA-GNN	4.125±0.025	4.021±0.025	4.175±0.027	3.952±0.034	3.877±0.022
$\mathcal{O}$ -GNN	8.808±0.060	7.051±0.080	8.117±0.056	8.374±0.083	8.113±0.067
RSE	7.287±1.900	11.371±8.700	9.392±1.000	10.282±0.330	8.786±1.400
geom2vec	7.911±0.290	5.180±0.642	10.593±0.413	10.860±1.700	17.240±5.400
FRGNN	3.375±0.015	3.343±0.029	3.421±0.024	3.500±0.035	3.521±0.021

表 15 FRGNN 与其他最先进模型对含不同环类型化合物预测性能的比较

Table 15 Comparison of predictive performance between FRGNN with SOTA models in predicting different ring types compounds

模型名称	螺环体系	桥环体系	稠环体系
GEM	6.28±0.01	9.46±0.01	7.81±0.01
SMPT	5.96±0.01	7.33±0.01	7.23±0.01
KA-GNN	7.32±0.08	9.25±0.08	8.57±0.08
$\mathcal{O}$ -GNN	7.92±0.05	8.04±0.07	8.14±0.07
RSE	10.60±3.40	11.52±0.48	10.44±0.73
geom2vec	12.34±2.60	12.31±1.70	11.63±2.20
FRGNN	5.23±0.01	6.50±0.01	6.86±0.01

相比于其他模型, FRGNN 取得了最佳预测性能, 平均性能提升为 10.04%, 同时在复杂环体系以及多环体系的预测中, FRGNN 也取得了最佳预测性能。但随着环的增大和结构的更加复杂, FRGNN 的预测性能与其他模型的差距不断减小, 这是因为随着环的增大, 环张力逐渐减小, 但分子内部存在

且 DL 属性更多的是对碳骨架的相似性描述, 与环张力的相关性较小。

2.4 含环化合物性质的预测

为了进一步考察模型对含环化合物的预测性能, 本研究设计了 1 个新的实验, 先从数据集中筛选了 5 234 个含环化合物, 其中含 3、4、5、6、7 元及以上环的分子占比分别为 22.20%、17.77%、15.57%、70.69%、43.91%, 含单环和多环分子占比分别为 10.36%和 89.64%, 螺环、桥环和稠环化合物分别占比 15.87%、17.12%以及 67.01% (需要指出的是, 由于分子可能包含不止 1 个环, 所以总比例相加会大于 100%, 因为存在重复计算)。然后比较各个模型预测性能, 结果如表 14~16 所示, 下划线为次最优模型结果。

的柔性结构能够被 FRGNN 捕捉, 导致其仍具备优势, 而环体系的复杂化可能会导致除了环张力之外的作用力的凸显如氢键^[48], 导致其优势被削弱。

2.5 误差来源分析

为了明确各个模型在中药数据集上预测的偏差来源, 本研究统计了各个模型中预测误差最高前 5%分子的平均含环数量、大环比例、小环比例和螺碳原子和桥头碳原子数, 如表 17 所示。相比其他 SOTA 模型 FRGNN 在预测误差最高的部分分子中, 含环的比例和复杂程度都小于其他模型。这说明 FRGNN 对于环状结构的预测能力强于其他 SOTA 模型, 也证明这些 SOTA 模型主要的误差来源正是这些含有复杂环结构的分子。GEM^[26]和 SMPT^[27]由于引入了空间结构, 其中隐式的包含了分子内部的张力, 因此在一定程度上学习到了分子的环张力特征。

2.6 消融实验

为了量化评价每一项改动对于模型性能的影响

表 16 FRGNN 与其他最先进模型对含不同环数量化合物预测性能的比较

Table 16 Comparison of predictive performance between FRGNN with SOTA models in predicting different ring number compounds

模型名称	含 1 个环	含 2 个环	含 3 个环	含 4 个环	含 5 个环	含 6 个环以上
GEM	10.124±0.009	11.142±0.006	10.026±0.011	9.661±0.009	11.492±0.012	11.949±0.015
SMPT	9.473±0.007	9.501±0.010	9.012±0.007	9.840±0.009	10.178±0.009	11.263±0.006
KA-GNN	9.471±0.072	9.480±0.094	9.157±0.150	8.896±0.093	8.881±0.113	9.143±0.132
O-GNN	9.851±0.074	9.748±0.081	10.044±0.093	10.904±0.172	10.463±0.097	10.262±0.083
RSE	10.701±3.236	10.791±1.847	10.908±2.486	11.514±1.949	12.623±2.229	10.742±2.910
geom2vec	11.063±3.542	11.451±4.704	13.323±2.134	12.605±3.449	10.954±3.127	14.162±3.692
FRGNN	8.860±0.021	8.813±0.003	8.752±0.004	7.859±0.014	8.123±0.007	7.852±0.005

表 17 误差来源分析

Table 17 Analysis of error sources

模型名称	平均含环数	平均大环数	平均小环数	平均螺碳原子数	平均桥头碳原子数
GEM	2.36	2.89	3.02	0.17	0.21
SMPT	2.15	2.57	2.66	0.18	0.26
KA-GNN	2.83	3.02	3.54	0.47	0.94
O-GNN	1.78	0.97	0.84	0.32	0.75
RSE	4.56	4.76	4.62	0.51	0.99
geom2vec	4.27	4.72	4.23	0.54	1.00
FRGNN	1.23	0.22	0.31	0.04	0.08

响,本研究将全部的模型改进逐步的添加到基础模型中,并比较各项改进后模型的预测性能,结果如表 18 所示,键长张力、角度张力、多构象的添加对模型性能的提升幅度分别为 2.05%、4.53%和 6.43%,针对性的 3 项改进都对模型的预测性能起到了正向的作用,且多构象的引入对模型的性能提升最为明显,这表明多构象的引入使得模型能够通过多个构象图学习不同的特征,而非从单一构型学习某些特定的特征^[49]。

2.7 鲁棒性检验

为进一步考察模型鲁棒性,本研究参考了 Boiko 等^[49]的鲁棒性检验方法尝试减少用于训练的数据量,从 80%起始,以 20%为梯度逐步递减。表 19 展示了各模型在 TCMSP 数据集(数据量更少)上不同训练比例的变化情况。随后使用采用 Bootstrap^[50]重采样(2 000 次)估计 RMSE 的 95%置信区间,如图 5 所示。结果表明,FRGNN 在减少训练数据比例的情况下始终取得最佳的模型性能,FRGNN 的均方根误差(RMSE)均值均低于 GEM,且其 95%置信区间上限始终低于 GEM 的置信区间下限。具体而言,在训练比例为 80%时,FRGNN 的 RMSE 为 3.252 (95% CI: 3.218 5~3.254 5),GEM 为 3.371

表 18 FRGNN 的消融实验

Table 18 Ablation study of FRGNN

消融分组	TCMSP	HERB2
所有改进去除	4.187±0.037	0.343±0.040
键长张力添加	4.031±0.050	0.341±0.004
角度张力添加	3.832±0.039	0.340±0.004
多构象添加	3.703±0.063	0.338±0.005
FRGNN	3.252±0.031	0.326±0.002

表 19 鲁棒性检验

Table 19 Robustness check

模型名称	训练比例			
	0.8	0.6	0.4	0.2
GEM	3.37±0.09	4.33±0.10	4.26±0.12	4.50±0.17
SMPT	3.94±0.11	4.21±0.11	4.23±0.11	5.06±0.14
KA-GNN	3.54±0.08	4.68±0.01	4.81±0.11	5.10±0.10
O-GNN	7.92±0.16	8.64±0.35	9.82±0.35	9.70±0.30
RSE	6.05±0.12	10.31±0.30	10.62±0.09	11.05±0.13
geom2vec	6.43±0.14	11.27±0.22	11.59±0.52	12.38±0.35
FRGNN	3.25±0.03	3.51±0.05	3.77±0.06	4.43±0.06

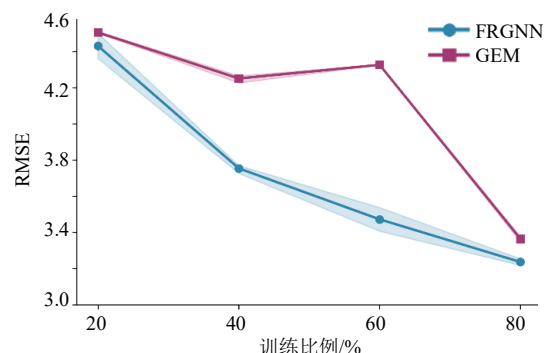


图 5 置信度分析

Fig. 5 Confidence level analysis

(95% CI: 3.348 5~3.377 8); 当训练比例降至 20%时,FRGNN 的 RMSE 为 4.431 (95% CI: 4.361~

4.501), GEM 为 4.503 (95% CI: 4.501 5~4.512 5)。2 组置信区间在 4 个比例下均无重叠,表明 FRGNN 相对于 GEM 的鲁棒性具有统计显著性^[49],能够以比较小的训练数据完成更好的预测效果。

3 讨论

本研究使用 2 大中药数据库 TCMSP 和 HERB2 作为数据来源,通过对模型添加环张力描述和多构象建模完成了 GIN 模型的改进,构建了 FRGNN。其能够快速、准确的预测中药这类结构复杂分子的各项性质,有效改善了以往模型存在的过拟合风险高、参数复杂、环张力表征不足的缺点。但此研究仍存在局限性,首先,对于更加复杂的体系而言,其他如氢键等作用力对分子性质的影响可能会更占据主导地位,仅依靠环张力无法完成对此类情况的模拟。未来可以借助分子动力学模拟软件进一步的细化复杂体系内的相互作用力,更全面的表示复杂体系内部的化学效应。其次,环张力的构建依赖于无张力分子的对比,这与分子内部真实的环张力情况可能存在偏差。这一点同样也可以借助分子动力学模拟软件来实现全面的表示。最后,整个实验所涵盖的数据不足,仅在 2 个数据集上验证了模型的优越性,且只对中药分子做了预测,其他天然产物是否有同样的预测效果还有待进一步的验证。未来将进一步扩大数据集的覆盖范围,纳入更多结构多样化的分子数据;引入更高精度的量子化学方法改进张力特征,同时开发更多可解释的方法,详细比较不同模型在训练过程中的学习过程。

综上,本研究为图神经网络在中药分子性质预测方面提供了切实可行的改进方向,也为未来的分子性质预测方法提供了方法学上的参考,为后续中药有效成分的筛选、作用机制的研究以及新药的开发提供技术支持。

利益冲突 所有作者均声明不存在利益冲突

参考文献

- [1] 白雪,谭天阳. 雷公藤生殖及肝肾毒理机制研究进展 [J]. 中国药物警戒, 2024, 21(10): 1192-1199.
- [2] 熊贤招,祝瑶露,汪新婳,等. 中药质量分析方法研究进展 [J]. 中国医药科学, 2025, 15(1): 35-38.
- [3] Karschnia P, Nelson T A, Dietrich J. Mechanisms and treatment of cancer therapy-induced peripheral and central neurotoxicity [J]. *Nat Rev Cancer*, 2025, 25(12): 887-909.
- [4] 赵宁,韩著,简颖琳,等. 中药多糖结构表征及质量评价研究进展 [J]. 中草药, 2024, 55(21): 7491-7506.
- [5] 赵崇军,贾哲,汤阳,等. 基于中药“毒-效”关系探讨中药毒性特征 [J]. 环球中医药, 2024, 17(8): 1602-1606.
- [6] Zhou T T. Unveiling secrets of traditional Chinese medicine: Cutting-edge techniques in component analysis [J]. *Chin Herb Med*, 2025, 17(3): 484-487.
- [7] Li X Y, et al. Integrated toxicology of *Aconitum carmichaelii* Debx.: Bridging traditional toxicity-efficacy understanding and modern toxicology [J]. *Front Pharmacol*, 2025, 16: 1667059.
- [8] 唐玺,刘韶,何鸽飞. 基于声表面波气相色谱技术和顶空-固相微萃取-气相-色谱质谱联用对不同产地苍术的鉴别和挥发油成分分析 [J]. 中南药学, 2026, 24(1): 41-47.
- [9] 葛薇. 超高效液相色谱-串联质谱法测定中药饮片中黄色素类染色剂的研究 [J]. 实验室检测, 2025, 3(18): 44-46.
- [10] 王帅. 基于液相色谱技术的中药制剂质量控制标准研究 [J]. 品牌与标准化, 2025(5): 133-135.
- [11] 巩亦萍,顾志荣,郭引兄,等. 中药药效物质基础研究方法与思路的进展与展望 [J]. 中成药, 2025, 47(4): 1211-1216.
- [12] Votkina D, Petunin P, Miliutina E, et al. Uncovering the role of chemical and electronic structures in plasmonic catalysis: The case of homolysis of alkoxyamines [J]. *ACS Catal*, 2023, 13(5): 2822-2833.
- [13] Burés J, Armstrong A, Blackmond D G. Explaining anomalies in enamine catalysis: “Downstream species” as a new paradigm for stereocontrol [J]. *Acc Chem Res*, 2016, 49(2): 214-222.
- [14] Richter W E. The increased Diels-Alder reactivity of umpolung tropone: Analysis of individual atoms and bonds using QTAIM and IQA along complete IRC paths [J]. *J Phys Org Chem*, 2024, 37(7): e4572.
- [15] Hui B Y K, Le Osmund Chin K, Lim J J L, et al. Strategies to enhance the electrochromic properties of conjugated polymers bearing pyromellitic diimide acceptors [J]. *Chem*, 2024, 19(12): e202400236.
- [16] Ree N, Wollschläger J M, Göller A H, et al. Atom-based machine learning for estimating nucleophilicity and electrophilicity with applications to retrosynthesis and chemical stability [J]. *Chem Sci*, 2025, 16(13): 5676-5687.
- [17] Ghidini A, Serra E, et al. On free energy calculations in drug discovery [J]. *Acc Chem Res*, 2025, 58(20): 3137-

- 3145.
- [18] Kim G, Grams R J, *et al.* Advancing covalent ligand and drug discovery beyond cysteine [J]. *Chem Rev*, 2025, 125(14): 6653-6684.
- [19] Gaudry A, Pagni M, Mehl F, *et al.* A sample-centric and knowledge-driven computational framework for natural products drug discovery [J]. *ACS Cent Sci*, 2024, 10(3): 494-510.
- [20] Adediwura V A, Koirala K, Do H N, *et al.* Understanding the impact of binding free energy and kinetics calculations in modern drug discovery [J]. *Expert Opin Drug Discov*, 2024, 19(6): 671-682.
- [21] Qian R T, Xue J, Xu Y, *et al.* Alchemical transformations and beyond: Recent advances and real-world applications of free energy calculations in drug discovery [J]. *J Chem Inf Model*, 2024, 64(19): 7214-7237.
- [22] Boulougouri M, Vanderghenst P, Probst D. Molecular set representation learning [J]. *Nat Mach Intell*, 2024, 6(7): 754-763.
- [23] Corso G, Stark H, Jegelka S, *et al.* Graph neural networks [J]. *Nat Rev Meth Primers*, 2024, 4: 17.
- [24] Xu K, Hu W, Leskovec J, *et al.* How powerful are graph neural networks? [A] // 7th International Conference on Learning Representations [C]. New Orleans: ICLR, 2019: 1-17.
- [25] Kipf T N, Welling M. Semi-supervised classification with graph convolutional networks [A] // Proceedings of the 5th International Conference on Learning Representations [C]. Toulon: ICLR, 2017: 1-14.
- [26] Fang X M, Liu L H, Lei J Q, *et al.* Geometry-enhanced molecular representation learning for property prediction [J]. *Nat Mach Intell*, 2022, 4(2): 127-134.
- [27] Li Y S, Wang W, Liu J, *et al.* Pre-training molecular representation model with spatial geometry for property prediction [J]. *Comput Biol Chem*, 2024, 109: 108023.
- [28] Wu Z Q, Ramsundar B, Feinberg E N, *et al.* MoleculeNet: A benchmark for molecular machine learning [J]. *Chem Sci*, 2017, 9: 513-530.
- [29] Chupakhin E, Babich O, Prosekov A, *et al.* Spirocyclic motifs in natural products [J]. *Molecules*, 2019, 24(22): 4165.
- [30] Nakase I, Lai H, Singh N P, *et al.* Anticancer properties of artemisinin derivatives and their targeted delivery by transferrin conjugation [J]. *Int J Pharm*, 2008, 354(1/2): 28-33.
- [31] Crull E B, Jain A N, Hawkins P C D, *et al.* Unmasking the true identity of rapamycin's minor conformer [J]. *J Nat Prod*, 2023, 86(7): 1862-1869.
- [32] Zhu J H, Wu K, Wang B H, *et al.* O-GNN: Incorporating ring priors into molecular modeling [A] // Proceedings of the 11th International Conference on Learning Representations [C]. Kigali: ICLR, 2023: 1-19.
- [33] Qiao J B, Jin J R, Wang D, *et al.* A self-conformation-aware pre-training framework for molecular property prediction with substructure interpretability [J]. *Nat Commun*, 2025, 16: 4382.
- [34] Luo Y Z, Yang K, Hong M, *et al.* MolFM: A multimodal molecular foundation model [EB/OL]. (2023-06-06) [2026-05-20]. <https://arxiv.org/abs/2307.09484>.
- [35] Axelrod S, Gómez-Bombarelli R. Molecular machine learning with conformer ensembles [J]. *Mach Learn Sci Technol*, 2023, doi: 10.1088/2632-2153/acefa7.
- [36] Ru J L, Li P, Wang J N, *et al.* TCMSp: A database of systems pharmacology for drug discovery from herbal medicines [J]. *J Cheminf*, 2014, 6(1): 13.
- [37] Gao K, *et al.* HERB 2.0: An updated database integrating clinical and experimental evidence for traditional Chinese medicine [J]. *Nucleic Acids Res*, 2025, 53(D1): D1404-D1414.
- [38] Ramsundar B, Eastman P, Walters P, *et al.* *Deep learning for the life sciences: Applying deep learning to genomics, microscopy, drug discovery, and more* [M]. Sebastopol: O'Reilly Media, 2019: 39-58.
- [39] Weininger D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules [J]. *J Chem Inf Comput Sci*, 1988, 28: 31-36.
- [40] Zhu Z, Li L, Wang Z, *et al.* PGL: A graph learning library for large-scale heterogeneous graphs [A] // Proceedings of the 30th ACM International Conference on Information & Knowledge Management [C]. New York: ACM, 2021: 4884-4888.
- [41] Ba J L, Kiros J R, Hinton G E. Layer normalization [EB/OL]. (2016-07-21) [2026-02-10]. <https://arxiv.org/abs/1607.06450>.
- [42] Chen Y H, Tang X, Qi X B, *et al.* Learning graph normalization for graph neural networks [EB/OL]. (2020-09-24) [2026-02-10]. <https://arxiv.org/abs/2009.11746>.
- [43] Kingma D P, Ba J. Adam: A method for stochastic optimization [EB/OL]. (2014-12-22) [2026-02-10]. <https://arxiv.org/abs/1412.6980>.

- [44] Croitoru F A, Ristea N C, Ionescu R T, *et al.* Learning rate curriculum [J]. *Int J Comput Vis*, 2025, 133(1): 291-314.
- [45] Li L L, Zhang Y P, Wang G H, *et al.* Kolmogorov-Arnold graph neural networks for molecular property prediction [J]. *Nat Mach Intell*, 2025, 7(8): 1346-1354.
- [46] Liu Z, Vinskus J, *et al.* Fast and accurate ring strain energy predictions with machine learning and application in strain-promoted reactions [J]. *JACS Au*, 2025, 5(10): 4750-4761.
- [47] Pengmei Z, Lorpaiboon C, Guo S C, *et al.* geom2vec: Pretrained GNNs as geometric featurizers for conformational dynamics [EB/OL]. (2024-09-30) [2026-05-20]. <https://arxiv.org/abs/2307.09484>.
- [48] Appavoo S D, Huh S, Diaz D B, *et al.* Conformational control of macrocycles by remote structural modification [J]. *Chem Rev*, 2019, 119(17): 9724-9752.
- [49] Boiko D A, Reschützegg T, Sanchez-Lengeling B, *et al.* Advancing molecular machine learning representations with stereoelectronics-infused molecular graphs [J]. *Nat Mach Intell*, 2025, 7(5): 771-781.
- [50] Efron B. Bootstrap methods: Another look at the jackknife [J]. *Ann Stat*, 1979, 7(1): 1-26.

[责任编辑 王文倩]