# 基于图卷积神经网络的杀菌剂活性预测研究

### 杜长委,凌晨阳,贾利峰,陈园园\*

南京农业大学理学院, 江苏 南京

收稿日期: 2025年6月28日; 录用日期: 2025年7月21日; 发布日期: 2025年7月30日

# 摘要

本研究提出了一种基于图神经网络(GCN)的杀菌剂活性预测模型。该模型通过直接解析分子SMILES字符 串构建分子图结构,并利用GCN进行端到端的图表示学习。这种方法避免了传统方法中依赖手工特征工 程的局限性,能够更全面且自动化地捕捉分子的结构信息。为评估模型性能,我们在相同杀菌剂活性数 据集上,将GCN模型与基于分子描述符的支持向量机(SVM)、随机森林(RF)及深度神经网络(DNN)模型进 行了系统对比。实验结果表明,GCN模型显著提高了真阳性率(TPR),有效提高了筛选效率;同时,其在 整体预测性能上全面超越了SVM、RF和DNN模型。本研究证实了图神经网络在直接从分子结构预测生物 活性方面的强大能力,为高效发现新型杀菌剂候选分子提供了一种更具潜力的计算工具。

#### 关键词

杀菌剂,图卷积神经网络,支持向量机,随机森林,深度神经网络

# **Research on Predicting Fungicidal Activity Based on Graph Convolutional Network**

#### Changwei Du, Chenyang Ling, Lifeng Jia, Yuanyuan Chen\*

College of Sciences, Nanjing Agricultural University, Nanjing Jiangsu

Received: Jun. 28<sup>th</sup>, 2025; accepted: Jul. 21<sup>st</sup>, 2025; published: Jul. 30<sup>th</sup>, 2025

#### Abstract

This study proposes a graph neural network (GCN)-based model for predicting fungicide activity. The model directly parses molecular SMILES strings to construct molecular graph structures and utilizes GCN for end-to-end graph representation learning. This approach overcomes the limitations of traditional methods that rely on manual feature engineering, enabling a more comprehensive and automated capture of molecular structural information. To evaluate model performance, we conducted a systematic comparison between the GCN model and descriptor-based models—Support Vector Ma-

\*通讯作者。

chine (SVM), Random Forest (RF), and Deep Neural Network (DNN)—on the same fungicide activity dataset. Experimental results demonstrate that the GCN model significantly improves the true positive rate (TPR), effectively enhancing screening efficiency. Concurrently, it surpassed the SVM, RF, and DNN models in overall predictive performance. This work validates the strong capability of graph neural networks in predicting bioactivity directly from molecular structures and provides a more promising computational tool for the efficient discovery of novel fungicide candidate molecules.

# **Keywords**

Fungicides, Graph Convolution Neural Network, Support Vector Machine, Random Forest, Deep Neural Network

Copyright © 2025 by author(s) and Hans Publishers Inc. This work is licensed under the Creative Commons Attribution International License (CC BY 4.0). http://creativecommons.org/licenses/by/4.0/

# 1. 引言

作为保障农业生产的关键投入品,农药在病虫害防控中发挥着重要作用。其中,杀菌剂作为农药的 重要类别,在减少作物病害损失方面扮演着不可或缺的角色。然而,传统杀菌剂的研发是一个漫长、昂 贵且充满不确定性的过程,严重依赖于大规模的实验筛选。这不仅耗费巨大资源,且成功率较低[1]。高 效、精准地预测候选分子的杀菌活性,对于加速新型杀菌剂的发现、降低研发成本具有重大现实意义。 因此,开发先进的计算预测模型,以建立分子结构与其生物活性之间的可靠关联,已成为现代农药研发 领域的关键技术需求。

近年来,深度学习技术凭借其强大的特征自动提取和复杂模式识别能力,在药物分子属性预测领域 取得了显著成功,展现出超越传统计算方法的巨大潜力。这些模型能够从海量的分子数据中学习深层次 的构效关系,为预测未知分子的生物活性提供了新的强大工具。研究者们已探索了包括卷积神经网络、 循环神经网络以及图神经网络在内的多种深度学习架构,用于处理不同形式的分子表示,并在多个药物 发现任务上验证了其有效性。

针对杀菌剂活性预测这一特定任务,本研究创新性地采用图卷积神经网络(GCN)作为核心模型[2], 直接将分子的 Simplified Molecular Input Line Entry System (SMILES)字符串转化为其分子图结构表示,并 在此基础上进行信息传递与特征学习。这种端到端的方式使得 GCN 能够直接从原始分子结构中捕获原 子间的拓扑连接、局部化学环境以及潜在的官能团信息,无需依赖预先定义的分子指纹或人工挑选的描 述符。为了全面评估 GCN 的性能,我们将其预测结果与基于分子描述符的传统机器学习模型支持向量机 (SVM)和随机森林(RF),以及深度神经网络(DNN)进行了系统比较。实验结果表明,在相同的杀菌剂活性 数据集上,直接处理分子图的 GCN 模型在预测性能上全面超越了基于分子描述符的 SVM、RF 与 DNN。 具体而言,GCN 在关键评价指标如准确率(Accuracy)、ROC 曲线下面积(AUC)和精准率(PRE)上均表现最 优。其中,对于最重要的评价指标 PRE, SVM、RF 与 DNN 的平均值分别为 0.858、0.761、0.857,而 GCN 的平均值达到了 0.962。这一性能优势充分证明了 GCN 模型学习复杂杀菌活性模式的有效性,并为未来 智能化杀菌剂设计与虚拟筛选提供了更强大的计算工具。

# 2. 数据与方法

# 2.1. 数据集

本研究从英国作物保护委员会(BCPC, https://www.bcpc.org/)网站下载了杀菌剂分子,作为正样本。负

样本数据则通过以下流程获取:首先下载了 ZINC 数据库(网址)中的分子;然后基于农药分子的理化性质,筛选满足 Hao's Rule (分子量  $\leq$  435、ClogP  $\leq$  6、氢键受体数  $\leq$  6、氢键供体数  $\leq$  2、可旋转键数  $\leq$  9、芳香键数  $\leq$  17) [3]的分子;最后随机选取与正样本数量相等的分子作为负样本。

所有分子均以 SMILES 格式表示。在数据预处理阶段,我们采用以下筛选标准:首先确保所有 SMILES 字符串能被 RDKit 工具包正确解析且长度不超过 100 个字符;其次去除因原子遍历顺序不同导致的重复 SMILES 表达式;最后排除包含"."符号的条目,以保证每个样本均为单一分子结构。最后保留正样本 与负样本数据各 372 项。

此外,本文通过在 PubMed 数据库中进行文献检索,收集了经实验验证具有杀菌活性的分子(不同于 训练数据中的杀菌剂分子),形成了一个补充数据集。该数据集作为独立测试集的正样本,并从 ZINC 数 据库选择同等数量的分子作为负样本,以便进一步对模型性能进行测试。

#### 2.2. 分子结构图

RDKit 是一个用于化学信息学的开源工具包,能够将 SMILES 字符串自动转化为精确的 2D 分子结构图。其功能核心在于解析 SMILES 文本描述的化学键类型、环结构、立体化学信息等,构建分子拓扑模型,并通过智能算法优化原子空间坐标排布,确保双环系统、官能团及手性中心的合理布局与准确表达。最终生成标准结构图以直观呈现分子空间构型。这些分子结构图将作为 GCN 的学习对象,从中提取关键信息。

以苯霜灵为例,其 SMILES 为 CC1=C(C(=CC=C1)C)N(C(C)C(=O)OC)C(=O)CC2=CC=C2,描述了包含双取代苯环、酰胺键及手性碳的复杂结构。RDKit 能够识别苯环连接位点、酯基(C(=O)OC)等关键特征,保持酰胺键(N-C(=O))的平面构型,并准确表达手性碳的四面体构象。生成的具体分子图如图 1 所示:



Figure 1. Benalaxyl molecular structure diagram 图 1. 苯霜灵分子结构图

此过程输出的分子结构图完整呈现苯霜灵的拓扑连接。也说明了 RDKit 能够将分子式进行可视化结构转换,这为药物分子分析提供了关键技术支撑。

### 2.3. 分子描述符与特征处理

分子描述符是通过数学方法定量表征分子结构特征的参数。本研究采用 Padel-Descriptor 软件将化合物的 SMILES 表达式转换为数值型特征,共获得 1875 个分子描述符,包括原子数目(nAtom)、正辛醇-水分配系数(AlogP)等结构特征。这些描述符将作为机器学习模型 SVM、RF 以及深度学习模型 DNN 的输入特征与 GCN 的预测结果进行比较分析。

## 2.3.1. 数据标准化

采用 Min-Max 标准化方法:

$$Z_{\rm i} = \frac{X_{\rm i} - X_{\rm min}}{X_{\rm max} - X_{\rm min}} \tag{1}$$

对所有特征进行归一化处理。该处理可消除特征间的量纲差异,提高模型收敛速度并优化模型性能。

#### 2.3.2. 方差筛选

为降低模型过拟合风险并去除冗余特征[4],首先对正负样本的特征矩阵进行方差筛选。通过设定方 差阈值,剔除方差小于 0.05 的低信息量特征,保留具有显著差异性的特征变量。

#### 2.3.3. 缺失值填补

在 Padelpy 计算得到的分子描述符中,部分特征存在缺失值。本研究采用 K 近邻算法[5] (KNN)进行 缺失值填补,KNN 基于欧氏距离度量分子相似性,即选取最邻近的 k 个分子,以其特征均值填补缺失值,以此有效保持数据集的整体分布特征。

#### 2.3. 模型简介

#### 2.3.1. 图卷积神经网络

图卷积神经网络(GCN)是一种专门用于处理图结构数据的深度学习模型,属于图神经网络(GNN)的重要分支,为本文的核心模型。与传统卷积神经网络(CNN)不同,GCN 能够直接对非欧几里得空间的图数 据进行特征学习,因此能够直接从分子图中学习特征,减少对人工描述符的依赖。

本文中所使用的 GCN 框架基于注意增强图卷积模块设计,以分子图结构数据为输入,包括原子特征 矩阵和邻接矩阵。分子图经过四次连续的图卷积模块运算来提取局部化学环境信息。随后,融合全局分 子表征的特征聚合层生成固定维嵌入向量(256 维)。然后该载体通过两层全连接层进行处理,并通过 Sigmoid 函数激活输出杀菌剂活性预测的概率分数。

模型的图卷积模块采用了多头注意力机制,集成了四个并行的全连接层和自注意力计算单元。卷积 层输出的节点特征经加权求和聚合后,通过非线性映射生成表征分子全局特性的 256 维图级向量。训练 过程中引入 Dropout 层与权重衰减策略,抑制复杂结构噪声对模型泛化能力的影响。完整框架如图 2 所示:



Figure 2. The complete framework of GCN 图 2. GCN 完整框架 GCN 模型使用正负样本集进行训练,并通过二元交叉熵损失

$$Loss = -\left[y\log(p) + (1-y)\log(1-p)\right]$$
<sup>(2)</sup>

对农药/非农药分类任务进行优化。原子特征包括基本的化学属性,如元素类型、价环境和芳香标志, 而键连通性被编码为邻接矩阵。

#### 2.3.2. 支持向量机

支持向量机[6] (SVM)是一种基于统计学习理论的监督学习方法,其核心原理是结构风险最小化。该 算法通过构造一个最优分离超平面来实现分类,这个超平面不仅要能正确划分训练数据,还要使两类样 本到超平面的最小距离最大化。可以将其转化为一个凸二次规划问题:在约束条件

$$y_i(\omega x_i + b) \ge 1 \tag{3}$$

下,最小化权值向量 ω 的范数 🔟 ,从而确保决策边界的几何间隔最大化。

对于非线性可分数据, SVM 采用核方法将原始特征空间映射到高维空间。通过引入满足 Mercer 条件的核函数

$$K(x_i, x_j) = \varphi(x_i) \cdot \varphi(x_j) \tag{4}$$

可以在高维特征空间中隐式地计算内积,避免了显式的高维映射。常用的核函数包括线性核、多项 式核和高斯径向基核。为处理噪声和异常点,SVM引入松弛变量和惩罚参数C,构造软间隔分类器,在 保证分类性能的同时允许少量样本出现在间隔区域内。最终确定决策函数

$$f(x) = sign\left(\sum_{i=1}^{n} \alpha_i y_i K(x_i, x_j) + b\right)$$
(5)

其中α<sub>i</sub>表示拉格朗日乘子,非零α<sub>i</sub>对应的样本即为支持向量,决定了分类边界[7]。

#### 2.3.3. 随机森林

随机森林[8] (RF)是一种基于集成学习的监督算法,通过构建多棵决策树并综合其预测结果来提高模型的准确性和鲁棒性。其核心方法为 Bootstrap 重采样,即从原始数据集中有放回地随机抽取多个子样本,为每个子样本独立训练一棵决策树,最终通过投票或回归的方式整合所有树的输出。此外,随机森林在每棵树的节点分裂时仅考虑随机选取的部分特征而非全部特征,这种双重随机性能有效降低模型方差,防止过拟合。

在随机森林的训练过程中,首先,从训练集中通过 Bootstrap 采样生成 k 个不同的子数据集;然后,为每个子数据集构建决策树,在树生长过程中的每个节点分裂时,从全部 m 个特征中随机选取  $m_i$  个候选特征(通常  $m_i = \sqrt{m}$ )进行最优分割;最终,将 k 棵树的预测结果进行集成。由于每棵树独立训练且采用随机性策略,随机森林天然具有并行计算优势,能高效处理高维数据,并对缺失值和异常值表现出较强的容忍度,使其成为处理复杂分类和回归问题的有力工具。

#### 2.3.4. 深度神经网络

深度神经网络[9] (DNN)是由多层非线性处理单元构成的深度学习模型,其通过层级化特征变换实现 高维数据的表示学习。每一层通过可学习的权重矩阵对输入进行线性投影,并施加非线性激活函数生成 高阶特征表达。网络的深度结构使其能够逐层解耦数据的复杂模式。训练过程基于反向传播算法优化损 失函数,利用梯度下降调整权重以最小化预测误差。

本文建立了一个全连接层神经网络,其中包括5个隐藏层和1个 Softmax 层作为最后的输出,优化

器选择 Adam 优化器。对于每个隐藏层的激活函数,我们进行了批归一化处理(BN)。BN 可以使训练过程 更加稳定,使用更大的学习率进行训练,提高训练速度,并且从一定程度上抑制了梯度消失。具体公式 如下:

$$BN(x) = \gamma \left(\frac{x_i - \widehat{\mu_B}}{\widehat{\sigma_B}}\right) + \beta$$
(6)

同时我们也对部分层引入了 Dropout,从而减少过拟合现象。本文选择采用二分类交叉熵损失函数,具体公式如下:

$$c = -\frac{1}{n} \sum_{x} \left[ y \ln a + (1 - y) \ln (1 - a) \right]$$
(7)

完整的框架如图3所示:



Figure 3. The complete framework of DNN 图 3. DNN 完整框架

# 2.3.5. 模型评价指标

为了降低偶然性,本文采用十折交叉验证,使用准确率(ACC)、ROC 曲线下面积(AUC)、精准率(PRE)、 召回率(REC)、F1 分数(F1)作为评价指标。下面对于以上评价指标进行详细说明。

首先,对于二分类预测问题,我们做以下定义:

TP (True Positive),表示模型正确预测为正例的样本数量;TN (True Negative),表示模型正确预测为 负例的样本数量;FP (False Positive),表示模型错误预测为正例的样本数量;FN (False Negative),表示模型错误预测为负例的样本数量。

基于以上基础定义,可以得到各评价指标的具体计算公式:

准确率(ACC):模型预测正确的样本占总样本的比例,用于衡量模型整体预测的准确性,计算方式见 公式(8)

$$ACC = \frac{TP + TN}{TP + TN + FP + FN}$$
(8)

ROC 曲线下面积(AUC): 衡量模型区分正负例能力的综合指标。ROC 曲线描绘了模型在不同分类阈 值下真正例率和假正例率的变化关系。而 AUC 是 ROC 曲线下的面积,通常通过数值积分计算,AUC 值 越接近 1,模型区分能力越好。

精准率(PRE):模型预测为正例的样本中实际为正例的比例,用于衡量预测结果的精确性,计算方式 见公式(9)

$$PRE = \frac{TP}{TP + FP} \tag{9}$$

召回率(REC):实际为正例的样本中,被模型正确预测为正例的比例,用于衡量模型找出所有正例的 能力,计算方式见公式(10)

$$REC = \frac{TP}{TP + FN} \tag{10}$$

F1 分数(F1): 精准率(PRE)和召回率(REC)的调和平均数,计算方式见公式(11)

$$F1 = \frac{2(PRE * REC)}{PRE + REC}$$
(11)

#### 3. 实验结果与分析

## 3.1. 模型参数设置

本研究中支持向量机(SVM)与随机森林(RF)模型基于 Scikit-learn 机器学习库构建,分子描述符采用 Padelpy 计算工具生成。输入数据经标准化处理、低方差特征过滤及缺失值填补等预处理步骤后,输入模 型进行训练与性能评估。模型对应的优化超参数分别见表 1 与表 2:

#### Table 1. SVM parameter configuration 表 1. SVM 参数设置

参数	取值
核函数	RBF
惩罚系数 C	1.0
Probablity	True

#### Table 2. RF parameter configuration 表 2. RF 参数设置

参数	取值
最大深度	2
树的数量	10

本研究通过 PyTorch 构建并训练 DNN 和 GCN 模型。DNN 模型的输入为分子描述符, GCN 模型的 输入则是分子的 SMILES 字符串,经由 RDKit 化学信息学工具包转换为原子-键分子图。为缓解模型过 拟合问题,在网络层中引入了 Dropout 正则化技术,其丢弃率(Dropout rate)设定为 0.2。模型的详细超参 数配置汇总于表3和表4。

#### Table 3. DNN parameter configuration 表 3. DNN 参数设置

参数	取值
隐藏层数	5
Dropout	0.2
学习率	0.01

Table 4. GCN parameter configuration 表 4. GCN 参数设置

参数	取值
卷积层数	4
Dropout	0.2
学习率	0.01

# 3.2. 参数优化

对于深度学习模型,训练时 Epoch 数是一个十分重要的参数,合适的 Epoch 值不仅可以减少计算资源的浪费,还能减少过拟合的影响,提升模型的表现。因此,本文对 GCN 模型的模型损失函数值与 Epoch 值的关系进行探究,以选择合适的 Epoch 值(见图 4)。



Figure 4. Relationship between GCN loss function value and Epoch 图 4. GCN 损失函数值与 Epoch 的关系

容易发现,当 Epoch 到达约 350 时,模型的损失函数值到达较低的值,并且后续值变化不大。因此,综合模型的表现以及训练效率,选择 Epoch 值为 350 较为合理,下文也用该参数训练模型。

# 3.3. 模型性能对比

本文采用十折交叉验证测试各模型的性能,具体结果如图 5 所示。从箱线图中可以看出,在四个模型中,GCN模型表现最好,其次是 DNN 和 SVM, RF 表现最差。另外,GCN模型不仅在各项评价指标中都明显优于其他三个模型,且得分的波动也最小,表明模型具有良好的稳定性。



特别地, GCN 模型具有较高的精准率 PRE, 平均值达到了 0.962, 远高于其他三个模型。这说明 GCN 判断为正样本的分子真正是正样本的比例较高, 从而显著降低了假阳性率。在杀菌剂筛选工作中, 这有

Metrics

利于后续筛选时降低时间和金钱的消耗,展现了模型实际应用的潜力。

Figure 6. Performance comparison of models on independent test sets 图 6. 模型在独立测试集上的性能

最后,为了进一步对比模型间的性能,在模型训练完成后,本文使用独立测试集对模型的性能进行进一步测试,结果如图 6 所示。可以看出,独立测试集上的结果与前文的结果相符,展现了 GCN 模型优秀的泛化能力。

# 4. 模型局限性与潜在改进方案

# 4.1. 模型局限性

尽管 GCN 模型在杀菌剂活性预测任务中展现出显著优势,但仍存在一些局限性。

模型对训练集的数据质量要求较高,数据集中若存在分子结构重复或正负样本分布不均衡等问题, 可能导致模型学习到错误的分子结构与活性关联模式,进而影响预测的准确性与泛化能力。

模型预测结果呈现较为明显的两极化现象,即相似性评分往往趋近于极端值(接近1或0),而处于中间区间的评分较少。这种情况可能是由于模型在训练过程中对决策边界的学习过于尖锐,使得其倾向于将样本明确划分为"是杀菌剂"或"不是杀菌剂"两类,而对介于两者之间、具有中等潜在活性的分子识别能力不足。这种评分特性在实际应用中可能存在不合理性,可能会遗漏一些具有中等活性、但经过结构优化后有望成为高效杀菌剂的候选分子,也可能对评分处于中间区域的分子的实际活性评估带来偏差。

此外,作为深度学习模型,GCN 在训练过程中存在计算资源耗费较大的问题。模型的图卷积运算、 多层神经网络结构以及大规模分子图数据的处理,均需要较高的硬件配置和较长的训练时间,这在一定 程度上限制了模型的广泛应用和快速迭代优化。

# 4.2. 改进方案

针对上述局限性,可考虑以下潜在改进方案。为提升模型对数据质量的适应性,可引入数据增强技术,如对分子结构进行合理的官能团替换、键级调整等操作,扩充训练集的多样性。同时,结合主动学 习策略,优先选择对模型性能提升最显著的样本进行人工标注,提高训练数据的质量和有效性。

为缓解预测结果两极化问题,可对模型的损失函数进行优化,例如引入平滑项或梯度惩罚项,促使 模型学习更平滑的决策边界,增加中间评分样本的比例。或采用多任务学习框架,将杀菌剂活性预测与 其他相关分子性质预测模型结合,引导模型从更全面的角度学习分子特征,减少对极端分类的过度依赖。 最后可探索轻量化模型架构设计,以减少计算资源的消耗。如采用图注意力机制等更高效的图神经 网络层,从而减少模型参数数量。又或是利用模型蒸馏技术,将复杂 GCN 模型的知识迁移到轻量级模型 上,在保持预测性能的同时降低计算成本。

# 5. 结语

本研究成功构建了一种基于图卷积神经网络(GCN)的杀菌剂活性预测模型,通过将 SMILES 字符串 直接转化为分子图结构进行端到端学习,显著提升了预测性能。实验表明:在相同数据集上,GCN 模型 在关键指标(准确率、AUC 和精准率)上全面超越传统机器学习方法(SVM 和 RF)及深度神经网络(DNN)。 特别在精准率(PRE)指标上,GCN 达到 0.962 的平均值,较 SVM (0.858)、RF (0.761)和 DNN (0.857)具有 显著优势。这种优势源于 GCN 直接捕获分子拓扑结构和化学环境的能力,有效降低了传统方法对人工描 述符的依赖。独立测试集验证进一步证实了模型优异的泛化性能。该成果为杀菌剂虚拟筛选提供了高效 计算工具,有望大幅缩短研发周期、降低实验成本,推动智能化农药设计的发展。

# 基金项目

本研究由中央高校基本科研业务费专项资金(YDZX2025015)和国家大学生科研训练计划 (202410307085Z)资助。

# 参考文献

- Kola, I. and Landis, J. (2004) Can the Pharmaceutical Industry Reduce Attrition Rates? *Nature Reviews Drug Discovery*, 3, 711-716. <u>https://doi.org/10.1038/nrd1470</u>
- [2] Lee, K., Jang, J., Seo, S., et al. (2022) Drug-Likeness Scoring Based on Unsupervised Learning. Chemical Science, 13, 554-565. <u>https://doi.org/10.1039/D1SC05248A</u>
- [3] Hao, G., Dong, Q. and Yang, G. (2011) A Comparative Study on the Constitutive Properties of Marketed Pesticides. *Molecular Informatics*, 30, 614-622. <u>https://doi.org/10.1002/minf.201100020</u>
- [4] 薛琦, 罗鄂湘. 基于机器学习的银行个人信用风险评估研究[J]. 建模与仿真, 2023, 12(4): 3747-3755.
- [5] 吴信东, 库玛尔. 数据挖掘十大算法[M]. 北京: 清华大学出版社, 2013.
- [6] 张学工.关于统计学习理论与支持向量机[J].自动化学报,2000,26(1):32-42.
- [7] 方逸雯, 刘媛华. 基于机器学习的员工生产效率预测[J]. 建模与仿真, 2023, 12(2): 830-840.
- [8] Breiman, L. (1996) Bagging Predictors. Machine Learning, 24, 123-140. <u>https://doi.org/10.1023/A:1018054314350</u>
- [9] Aldahdooh, A., Hamidouche, W., Fezza, S.A., et al. (2022) Adversarial Example Detection for DNN Models: A Review and Experimental Comparison. Artificial Intelligence Review, 55, 4403-4462. https://doi.org/10.1007/s10462-021-10125-w