

DOI: 10.20189/j.cnki.CN/61-1527/E.202603011

基于轻量级梯度提升机的共晶形成可解释预测模型

刘钰琨, 刘延飞, 杜子昂, 刘 博
(火箭军工程大学, 西安 710025)

摘 要: 针对含能材料共晶筛选效率低、现有预测模型可解释性不足且难以适配不平衡数据集的问题, 提出了一种基于轻量级梯度提升机(light gradient boosting machine, LightGBM)的共晶形成可解释预测模型, 从而实现对含能材料共晶形成的高精度可解释预测。该模型采用Morgan分子指纹作为输入特征, 以实现对分子局部化学环境的精准表征; 同时, 采用融合直方图离散化、带深度限制的Leaf-wise分裂及单边梯度采样的LightGBM算法进行共晶预测, 并引入改进的对数损失函数来平衡正负样本训练权重, 通过贝叶斯优化算法确定树数量与最大深度的最优超参数组合; 最后, 基于SHAP理论构建可解释性分析模块, 得到各指纹位对预测结果的贡献度, 筛选出关键Morgan指纹位并反向映射可视化分子局部环境。实验结果表明: 在完全隔离的独立测试集上, 模型平衡准确率为95.15%、准确率为97.16%、F1分数为98.35%; 经3折、5折、10折交叉验证后指标波动极小, 泛化能力与鲁棒性能均优于其他机器学习模型; 可解释性分析明确了芳香环、氢键供受体、疏水区域等关键结构特征及其作用机制, 为共晶组分定向筛选与结构修饰提供理论参考。

关键词: 共晶; 可解释预测; LightGBM; 分子指纹; SHAP

中图分类号: TP181

文献标志码: A

开放科学标识码(OSID):

文章编号: 2097-2741(2026)03-0131-10



听语音
与作者互动
聊科研

共晶是两种或多种分子通过氢键、 π - π 堆积、范德华力等非共价相互作用而有序组装形成的多组分晶体, 其凭借在调控物质理化性质方面所具有的独特优势, 在药物研发、材料科学、化工生产等领域展现出重要的应用价值^[1-3]。特别是在含能材料领域, 传统单一组分含能材料普遍存在能量密度与安全性能无法兼顾的问题, 而共晶技术可兼顾能量密度与安全性, 其性能优劣直接决定了武器装备的毁伤效能、航天运载火箭的推进能力以及民用爆破工程的作业效率。然而, 共晶的形成大多依赖于组分分子间的结构互补性与非共价相互作用的协同匹配, 其形成规律复杂且难以通过理论直接推导。因此, 共晶形成的预测一直

是一项复杂且具有挑战性的任务。

传统的共晶筛选主要依赖于试错法与高通量实验筛选等手段。试错法通过反复试错检验分子对能否形成共晶, 这种方法操作繁琐、耗时耗力, 且成功率不高^[4]。高通量筛选虽能通过自动化设备提升实验效率, 但仍需消耗大量样品与试剂, 成本高昂, 且难以覆盖众多潜在的分子组合^[5-6]。随着大数据和人工智能技术的发展, 数据驱动的机器学习预测模型在跨学科领域中得到广泛应用^[7-8]。与传统的计算筛选方法相比, 机器学习模型具有更加强大的高维数据处理能力、复杂的非线性关系挖掘能力和高效的模式识别特性。机器学习模型不依赖于简化的物理化学假设或经验参数, 可

收稿日期: 2026-01-28

修回日期: 2026-03-12

录用日期: 2026-04-15

基金项目: 智剑实验室开放基金(2024-ZJSYS-KF02-01)

第一作者: 刘钰琨(2001—), 男, 硕士研究生, 主要从事基于机器学习的共晶预测研究。

引用格式: 刘钰琨, 刘延飞, 杜子昂, 等. 基于轻量级梯度提升机的共晶形成可解释预测模型[J]. 火箭军工程大学学报, 2026, 40(3): 131-140. LIU Yukun, LIU Yanfei, DU Ziang, et al. Interpretable prediction model for cocrystal formation based on a light gradient boosting machine[J]. Journal of Rocket Force University of Engineering, 2026, 40(3): 131-140.

以直接从大规模已知共晶数据中自主学习分子对之间的构效关系。现有的研究通过构建不同规模的共晶数据集,运用多种分子表征方式和模型训练策略,初步验证了机器学习在共晶预测任务中的可行性,推动了共晶筛选从理论探索向实际应用方向发展^[9-12]。

尽管目前关于共晶形成预测的研究已从传统实验方法转向了机器学习方法,但仍存在一些不足之处。首先,分子表征依赖于全局分子描述符,这些描述符主要反映分子整体的物理和化学性质,难以精准捕捉对共晶形成至关重要的局部化学环境信息;其次,现有机器学习模型的预测过程无法清晰解释哪些分子结构特征主导了预测结果,也无法揭示模型针对特定分子的决策逻辑,这种可解释性的缺失可能导致研究人员对预测结果的不信任或盲目信任,也无法建立预测结果与共晶形成机制之间的联系。上述问题使得现有模型难以满足共晶设计的实际需求,因此迫切需要构建具有高预测精度和强可解释性的预测模型。

针对上述问题,本文构建了基于轻量级梯度提升机(light gradient boosting machine, LightGBM)的共晶形成可解释预测模型,即分子指纹-轻量级梯度提升机(molecular fingerprint LightGBM, MF-LightGBM)。该模型以 Morgan 分子指纹作为特征载体,融合 2 种组分分子的结构信息与相互作用潜力,采用 LightGBM 算法进行预测,并针对共晶预测任务进行适配改进,从而实现高维特征处理效率与模型泛化能力之间的相互平衡。同时,本文引入 SHAP 理论进行可解释性分析,对

关键作用点位的分子局部环境进行了可视化处理。

1 MF-LightGBM 模型

1.1 模型整体框架

MF-LightGBM 模型结构如图 1 所示,包含特征提取、改进 LightGBM、可解释性分析 3 个模块,能够实现从分子结构信息到共晶形成结果的端到端映射,有效解决传统机器学习模型存在的“黑箱”问题。首先,特征提取模块将 2 种共晶组分分子的 SMILES 化学符号转化为能够表征分子局部环境的 Morgan 指纹;然后,将该指纹输入改进 LightGBM 模块进行二分类预测;最后,借助可解释性分析模块得到各 Morgan 指纹位对预测结果的贡献度,筛选出贡献度排名前 3 的 Morgan 位。

1.2 特征提取模块

共晶能否形成与分子的拓扑结构、官能团的类型以及原子连接方式密切相关。Morgan 指纹是一种基于分子局部环境的表征方法,通过考虑每个原子及其一定半径范围内的邻接原子来捕捉分子局部结构信息^[13-15],能够精确描绘分子中的特定官能团、氢键供体、受体位点和芳香环系统的局部化学环境,囊括了影响共晶形成的主要物理和化学因素。因此,本模块采用 Morgan 分子指纹作为特征载体,以表征分子的局部结构信息。

Morgan 指纹的生成是基于以原子为中心的圆形扫描机制,其具体流程如下:1) 对输入的 SMILES 字符串进行解析,构建分子的二维拓扑结构,明确原子的类型、价态、成键方式以及邻

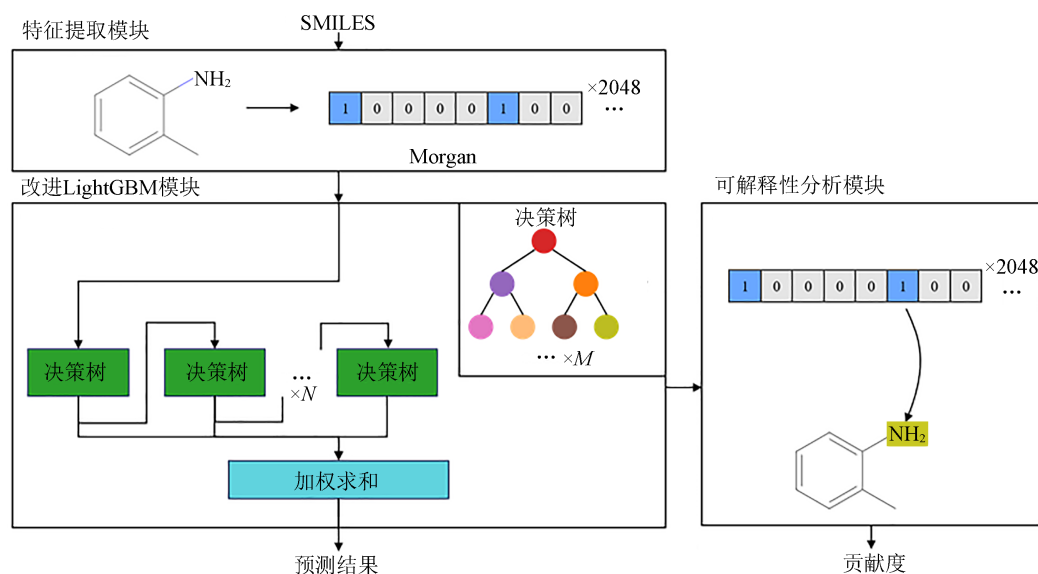


图 1 MF-LightGBM 模型整体框架

Fig. 1 Overall framework of the MF-LightGBM model

接原子信息; 2) 对于每个原子, 遍历其指定半径范围内的所有原子和化学键, 构建分子局部环境的子集; 3) 对每个局部环境子集进行哈希编码, 将化学结构信息转换为固定长度的二进制向量, 其中每个位置对应特定局部化学环境的存在状态。

设 Morgan 指纹向量分别为 $\mathbf{F}_1 \in \{0, 1\}^d$ 和 $\mathbf{F}_2 \in \{0, 1\}^d$ 。其中, d 为指纹维度。为同时表征 2 种分子的结构信息及相互作用潜力, 将 $\mathbf{F}_1$ 与 $\mathbf{F}_2$ 进行横向拼接, 得到最终输入特征向量 $\mathbf{F} = \mathbf{F}_1 \oplus \mathbf{F}_2 \in \{0, 1\}^{2d}$ 。其中, 符号 “ $\oplus$ ” 表示向量拼接操作, 特征向量 $\mathbf{F}$ 可表示为

$$\mathbf{F} = [f_{11}, f_{12}, \dots, f_{1d}, f_{21}, f_{22}, \dots, f_{2d}]^T \quad (1)$$

式中: f_{ij} 为第 i 个分子的 Morgan 指纹向量中第 j 位取值, 具体为 0 或 1, 分别对应应该位置所代表的局部化学环境不存在或存在。本文设 $d = 2048$, 该维度为 Morgan 指纹的常用标准维度, 可在完整覆盖含能分子常见官能团、芳香环、氢键位点等局部化学环境的同时, 避免维度冗余所导致的模型计算效率下降与过拟合风险, 可兼顾特征完整性与模型效率。

1.3 改进 LightGBM 模块

LightGBM 是基于梯度提升决策树 (gradient boosting decision tree, GBDT) 的集成学习算法, 通过直方图优化、带深度限制的 Leaf-wise 叶子生长策略、单边梯度采样等改进策略, 能够有效提升模型训练效率与高维特征处理能力, 在小样本、高维稀疏特征的分类任务中表现出色^[16]。但原始 LightGBM 并未针对共晶数据集正负样本不平衡问题进行优化, 且默认的超参数难以适配分子指纹的高维稀疏特性。因此, 本文对 LightGBM 进行多维度改进, 以适用于共晶预测任务。

1.3.1 直方图算法

直方图算法通过对连续特征进行离散化处理来优化节点分裂过程, 解决了传统 GBDT 算法分裂点搜索计算成本高的问题。该算法首先将连续特征取值映射至有限数量的区间 (即直方图 bin), bin 边界采用分位数离散化方法确定, 以确保各区间样本分布相对均衡, 其数学表达为

$$b_j = \left\{ x \mid q_{(j-1) \cdot \frac{1}{K}} \leq x < q_{j \cdot \frac{1}{K}} \right\} \quad j = 1, 2, \dots, K \quad (2)$$

式中: $q_{(j-1) \cdot \frac{1}{K}}$ 和 $q_{j \cdot \frac{1}{K}}$ 分别为特征的第 $(j-1) \cdot \frac{1}{K}$ 与 $j \cdot \frac{1}{K}$ 分位数; K 为直方图 bin 的数量, 本文设 $K = 255$, 该值对应 8 位二进制数所能表示的最大十进

制数, 是 LightGBM 原算法针对二值稀疏特征优化的默认最优值, 可在保证离散化精度的同时, 最大限度降低内存占用与计算成本, 并适配 Morgan 指纹的 0-1 二值特征。

随后构建存储样本数量及梯度统计信息的特征直方图, 以 bin 边界作为候选分裂点来计算分裂增益, 最终选取增益最大的分裂点执行分裂。分裂增益采用信息增益率 G_{split} 评价, 表达式为

$$G_{\text{split}} = \frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{(G_L + G_R)^2}{H_L + H_R + \lambda} - \gamma \quad (3)$$

式中: λ 为 L2 正则化系数, 用于防止模型过拟合; γ 为叶子节点复杂度惩罚系数, 用于控制决策树的生长规模; G_L 、 G_R 分别为左、右子树样本的梯度之和; H_L 、 H_R 分别为左、右子树样本的 Hessian 之和, 其定义为样本损失函数对预测值的二阶导数之和, 表达式为

$$H_i = \sum_{x \in \text{leaf}_i} \frac{\partial^2 \text{Loss}(y, \hat{y})}{\partial \hat{y}^2} \quad (4)$$

式中: $\text{Loss}(\cdot)$ 为模型的损失函数; y 为样本的真实标签值; $\hat{y}$ 为模型对样本的预测值; leaf_i 表示决策树的第 i 个叶子节点; x 为落入该叶子节点的单个样本。直方图算法的离散化处理大幅提升了模型训练速度, 同时还能增强模型对噪声特征的鲁棒性。

1.3.2 带深度限制的 Leaf-wise 算法

Leaf-wise 算法能够优先对增益值最高的叶节点进行分裂, 从而快速聚焦于高价值样本区域, 并生成精简且准确的决策树, 但同时也极易出现过度分裂的情况, 从而导致模型产生过拟合。因此, 本文引入深度限制机制, 当树的深度达到阈值时, 所有叶子节点的分裂都会终止。在每次迭代中, 只有具有最高增益且不超过深度阈值的叶节点才会被分裂, 这个过程会一直持续, 直到没有节点能够满足分裂条件为止。带有深度限制的 Leaf-wise 算法保留了原算法精准分裂的优势, 同时通过深度约束抑制过拟合, 确保了模型的泛化能力。

1.3.3 单边梯度采样算法

单边梯度采样算法通过基于梯度的样本采样来优化模型训练效率, 重点采样核心样本, 减少了计算资源消耗。在梯度提升过程中, 样本对参数更新的贡献度与梯度的绝对值呈正相关。高梯度的样本对模型调整有显著影响, 而低梯度的样本贡献有限。阈值 θ 的定义为

$$\theta = \text{quantile}(|g_i|, 0.8) \quad (5)$$

式中： $|g_i|$ 为单个样本的梯度绝对值； $\text{quantile}(\cdot, 0.8)$ 表示梯度绝对值的 80 百分位数。单边梯度采样算法首先计算所有样本梯度绝对值 $|g_i|$ ，选择 $|g_i| \geq \theta$ 的样本作为高梯度样本集（占比为 20%）并保留所有这些样本，然后从剩余低梯度样本中随机抽取 10% 来补偿采样偏差。该方法既能保证训练的准确性，又能显著提高效率。

1.3.4 改进的对数损失函数

针对共晶预测数据集中普遍存在的正负样本不平衡问题，采用改进的对数损失函数，通过引入权重因子 L 来平衡正负样本的训练权重，其表达式为

$$L = - \sum_{i=1}^n [\omega \cdot y_i \cdot \log(p_i) + (1 - y_i) \cdot \log(1 - p_i)] \quad (6)$$

式中： y_i 为样本真实标签，正样本取 1，负样本取 0； p_i 为预测结果为正样本的概率； ω 为负样本数量与正样本数量之比； n 为样本总数。该损失函数通过提升正样本的训练权重，避免模型因负样本占比过高而偏向多数类，有效提升了模型对少数类（正样本）的识别能力。

1.3.5 基于贝叶斯优化的超参数寻优

机器学习模型性能很大程度上取决于超参数设置。原始 LightGBM 的默认超参数难以适配共晶预测任务中高维稀疏的分子指纹特征，因此本文以验证集的平衡准确率最大化为目标，采用贝叶斯优化算法对 MF-LightGBM 模型的树数量 N 和最大深度 M 进行迭代搜索，以得到最优的超参数组合。贝叶斯优化算法是一种智能的黑箱函数优化算法，能够在有限的评估次数内高效找到全局最优解，避免了网格搜索和随机搜索的盲目性与低效性。

1.4 可解释性分析模块

尽管 LightGBM 模型具有出色的预测性能，但其决策过程却难以进行解释，无法揭示分子结构特征与共晶形成之间的内在关系。可解释性分析模块基于 SHAP 理论构建，通过计算每个 Morgan 指纹位的 SHAP 值来量化其对共晶形成预测结果的贡献度，并对关键特征相对应的分子局部环境进行可视化处理，从而实现模型预测的可解释性^[17-19]。

SHAP 值是基于博弈论中的 Shapley 值推导而来的，其核心思想是将模型的预测结果分解为每个特征的贡献之和，且每个特征的贡献程度必须满足公平性和一致性。第 j 个特征的 Shapley 值 ϕ_j 可表示为

$$\phi_j = \frac{1}{|U|!} \sum_s |S|! \cdot (|U| - |S| - 1)! \cdot [f(S \cup \{j\}) - f(S)] \quad (7)$$

式中： U 为所有特征的集合； S 为不含第 j 个特征的特征子集， $|\cdot|$ 表示集合中元素的个数； $f(\cdot)$ 为 LightGBM 模型的预测函数。式 (7) 通过将所有特征子集的边际贡献相加，公平地量化了单个特征对预测结果的影响。

在可解释性分析模块中，SHAP 值将模型的预测结果分解为基准值以及每个指纹位贡献度之和，数学表达式为

$$P(\mathbf{x}) = P_0 + \sum_{j=1}^{2d} \text{SHAP}_j(\mathbf{x}) \quad (8)$$

式中： $P(\mathbf{x})$ 为模型对输入样本 $\mathbf{x}$ 的共晶形成概率预测值； $\mathbf{x}$ 为分子对的拼接 Morgan 指纹向量； P_0 为所有样本的平均预测值（即基准值）； $\text{SHAP}_j(\mathbf{x})$ 为第 j 个 Morgan 指纹位对于该样本预测结果的 SHAP 值，该值越大，其贡献度也就越大。通过式 (8) 可以清晰地确定单个样本中每个分子结构特征对预测结果的具体影响。

对于筛选出的关键 Morgan 指纹位，通过反向映射追溯其对应的分子局部环境。假设关键指纹位对应的哈希编码为 h ，其反向映射至分子局部环境的数学表达为

$$E = \{ (A_i, B_{ij}) \mid \text{hash}(A_i, B_{ij}) = h \} \quad (9)$$

式中： E 为哈希编码 h 对应的分子局部环境集合； A_i 为中心原子； B_{ij} 为中心原子半径范围内的邻接原子及化学键类型； $\text{hash}(\cdot)$ 为指纹生成阶段的哈希函数。

1.5 训练过程

为确保模型训练的稳定性、预测结果的可靠性以及对不平衡数据的适配性，设计了一套严格的训练流程，涵盖数据集划分、正负样本不平衡处理、超参数调优、模型训练及评估等多个关键步骤^[20-23]，每个步骤紧密相连，以实现模型性能和泛化能力的双重保障。

共晶数据集采用三级划分策略，以防止数据泄露并确保模型评估的客观性，具体划分方式可分为以下 2 个步骤：1) 将整个数据集按 1:9 的比例随机划分为独立测试集与训练-验证集。独立测试集仅在模型训练完成后用于最终的性能测试，不参与训练和超参数调整过程，以真实反映模型对未知数据的泛化能力。2) 将训练-验证集进一步以 8:2 的比例划分为训练集与验证集。训练集用于模型参数的迭代更新，而验证集用于在模型

训练过程中监测过拟合情况, 并指导超参数调整和模型提前停止。在数据集划分过程中, 采用了分层抽样策略, 以确保每个数据集中的正样本和负样本的比例与原始数据集一致, 从而避免由于样本分布偏差而导致模型在训练时偏向多数类样本的情况发生。

模型训练阶段, 使用训练集数据拟合 MF-LightGBM 模型, 同时在验证集上计算性能指标。最终训练好的最优模型在完全独立的测试集上进行性能测试, 所得到的指标作为模型真实泛化能力的评估依据。

2 实验与结果分析

2.1 数据集构建

构建规模充足且质量优良的数据集是实现机器学习可靠预测的基础^[24-26]。通过系统采集与筛选, 本文构建了包含 6 700 个正样本和 1 000 个负样本的高质量共晶数据集。正样本来源于剑桥晶体学数据库^[27], 筛选条件为仅保留由 2 个有机分子组成、不含非中性分子、溶剂分子及原子位置无序的晶体记录; 负样本来源于已发表文献中明确报道的未能形成共晶的实验结果^[28-32], 相较于计算模拟生成的负样本, 其可靠性与代表性更强。所有样本中的分子均以标准 SMILES 字符串形式存储。

按照上述 1.5 节的三级划分策略, 最终得到包含 5 544 个样本的训练集、包含 1 386 个样本的验证集以及包含 770 个样本的测试集, 各子集无数据重叠, 功能明确, 数据集划分及正负样本分布情况如图 2 所示。

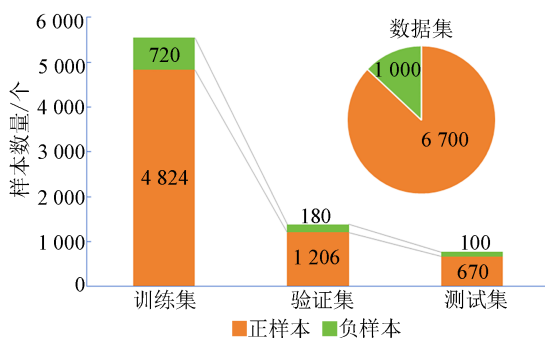


图 2 数据集划分及正负样本分布情况

Fig. 2 Dataset division and distribution of positive and negative samples

2.2 实验设置

所有计算都在 Windows 10 操作系统上完成, CPU 为 2.40 GHz、4 核 8 线程的 Intel Core i5-

10200H, GPU 为 6 GB 显存的 NVIDIA GeForce RTX 2060, 编程语言为 Python 3.10。为确保实验可复现, 所有实验均固定随机种子为 42。

针对共晶形成预测的二分类任务, 采用 6 项关键指标全面评估模型性能, 涵盖总体分类效果、类别平衡适应性及正负样本识别能力。具体包括: 1) 准确率 A_{Accuracy} , 用于衡量整体预测正确的比例; 2) 平衡准确率 B_{BACC} , 更客观地反映模型对两类样本的均衡识别能力, 适用于正负样本数量差异较大的数据集; 3) 精确率 $P_{\text{Precision}}$, 关注模型预测为共晶的样本中实际为共晶的比例; 4) 召回率 R_{Recall} , 实际为共晶的样本中被模型正确识别的比例; 5) 特异度 $S_{\text{Specificity}}$, 实际为非共晶的样本中被模型正确识别的比例; (6) F1 分数 $F_{\text{F1-Score}}$, 作为精确率与召回率的调和平均数, 是综合评估模型在不平衡数据集上性能的关键指标。上述性能指标的数学表达式分别为

$$A_{\text{Accuracy}} = \frac{N_{\text{TP}} + N_{\text{TN}}}{N_{\text{TP}} + N_{\text{TN}} + N_{\text{FP}} + N_{\text{FN}}} \quad (10)$$

$$B_{\text{BACC}} = \frac{1}{2} \left(\frac{N_{\text{TP}}}{N_{\text{TP}} + N_{\text{FN}}} + \frac{N_{\text{TN}}}{N_{\text{TN}} + N_{\text{FP}}} \right) \quad (11)$$

$$P_{\text{Precision}} = \frac{N_{\text{TP}}}{N_{\text{TP}} + N_{\text{FP}}} \quad (12)$$

$$R_{\text{Recall}} = \frac{N_{\text{TP}}}{N_{\text{TP}} + N_{\text{FN}}} \quad (13)$$

$$S_{\text{Specificity}} = \frac{N_{\text{TN}}}{N_{\text{TN}} + N_{\text{FP}}} \quad (14)$$

$$F_{\text{F1-score}} = 2 \times \frac{P_{\text{Precision}} \times R_{\text{Recall}}}{P_{\text{Precision}} + R_{\text{Recall}}} \quad (15)$$

式中: N_{TP} 为真阳性, 代表实际为共晶且预测为共晶的样本数; N_{TN} 为真阴性, 代表实际非共晶且预测为非共晶的样本数; N_{FP} 为假阳性, 代表实际非共晶但预测为共晶的样本数; N_{FN} 为假阴性, 代表实际为共晶但预测为非共晶的样本数。

2.3 超参数调优

本文采用贝叶斯优化算法, 以验证集平衡准确率最大化为目标, 对 MF-LightGBM 最关键的 2 个超参数 (树数量 N 和最大深度 M) 进行系统调优, 并使用独立测试集监测模型泛化性能以辅助评估不同超参数组合的优劣。超参数搜索范围及贝叶斯优化确定的最优组合如表 1 所示。

2.4 模型对比

为客观评估 MF-LightGBM 模型的性能, 将其与决策树 (decision tree, DT)、逻辑回归 (logistic regression, LR)、高斯朴素贝叶斯 (Gaussian naive Bayes, GNB)、随机森林 (random for-

est, RF)、GBDT 以及 K 最近邻 (K-nearest neighbor, KNN) 6 种主流机器学习模型进行对比, 所有模型使用相同的特征输入与数据集。不同模型在相同独立测试集上的性能对比结果如表 2 所示。

表 1 贝叶斯优化的超参数搜索范围及最佳超参数
Table 1 Hyperparameter search range and optimal hyperparameters for Bayesian optimization

超参数	搜索范围	最佳超参数
<i>N</i>	50~800	300
<i>M</i>	3~20	15

表 2 不同模型在相同测试集上的性能对比
Table 2 Performance comparison of different models on the same test set %

模型	平衡准确率	准确率	精确率	召回率	特异度	F1 分数
DT	89.01	94.20	97.14	96.13	81.90	96.63
LR	94.03	95.23	98.77	95.68	92.38	97.20
GNB	88.95	92.01	97.50	93.14	84.76	95.27
RF	86.80	95.23	96.21	98.36	75.24	97.27
GBDT	87.39	89.30	97.42	90.01	84.76	93.57
KNN	90.50	93.30	97.84	94.34	86.67	96.05
MF-LightGBM	95.15	97.16	98.80	97.91	92.38	98.35

注：粗体表示每列最优值。

由表 2 可得, 相比其他 6 种主流机器学习模型, MF-LightGBM 模型的综合表现最好, 在相同测试集上的平衡准确率、准确率、精确率、特异度及 F1 分数均达到最优。

为进一步降低由数据集随机划分所带来的性能评估偏差, 本文采用 3 折、5 折、10 折交叉验证方法对各模型性能进行测试。在交叉验证过程中, 每次将数据集按对应折数均分, 以其中 1 份作为测

试集, 其余按 8 : 2 的比例划分为训练集与验证集, 迭代完成模型训练与测试, 最终取多轮验证结果的均值作为性能指标。

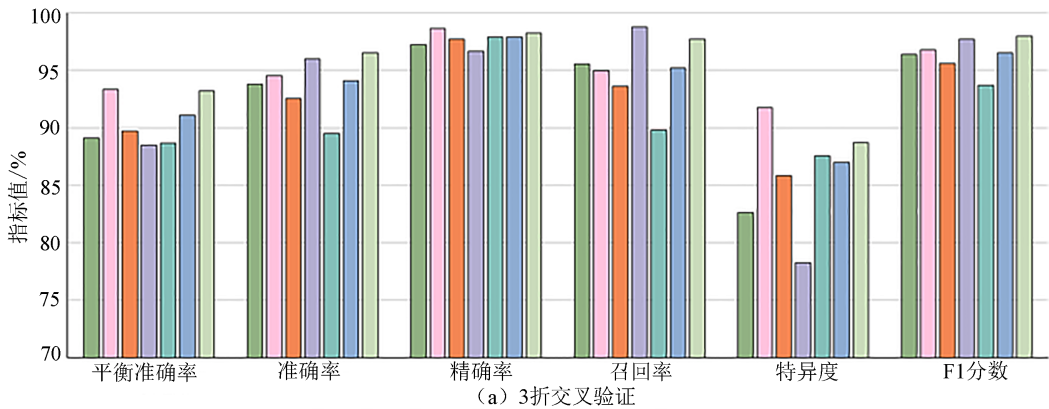
不同模型交叉验证的核心指标变化情况如图 3 所示。从图中可以看出, MF-LightGBM 模型的平衡准确率、准确率、F1 分数在 3 折、5 折、10 折交叉验证中均为最高, 且柱状图高度变化极小, 呈现出最稳定的性能表现; LR 模型的交叉验证稳定性次之, 10 折测试集平衡准确率与 MF-LightGBM 接近; DT、RF 与 KNN 模型的指标值随折数增加有小幅上升, 但整体显著低于 MF-LightGBM 与 LR; GNB 与 GBDT 模型的各项指标在所有折数下均处于较低水平。

综合来看, MF-LightGBM 在各项评估指标、泛化能力与稳定性上均表现最优, 这得益于直方图算法、带深度限制的 Leaf-wise 算法及单边梯度采样算法对高维稀疏特征的高效处理能力, 同时本文引入的改进对数损失函数有效适配了正负样本不平衡场景, 贝叶斯优化得到的最优超参数组合进一步提升了模型的拟合精度与泛化能力。

2.5 可解释性分析

为从分子结构层面揭示共晶形成的微观机制, 本文基于 SHAP 框架对 MF-LightGBM 模型进行了可解释性分析。选取涵盖氢键、 π - π 堆积、疏水作用等不同非共价作用类型的 5 组共晶正样本, 对关键 Morgan 指纹位进行反向映射与分子局部环境可视化处理, 结果如图 4 所示。其中, 黄色圆圈标记了各样本中 2 个分子贡献度排名前 3 的关键 Morgan 位对应的分子局部环境。

图 4 (a) 中, 黄色圆圈标记出了分子一中的芳香环系统和氢键受体区域, 在分子二中标记出了芳香环区域和氢键供体基团, 体现出经典的 π - π 堆积与氢键互补模式; 图 4 (b) 中, 样本二的分子一具有柔性极性链, 分子二为刚性平面结构,



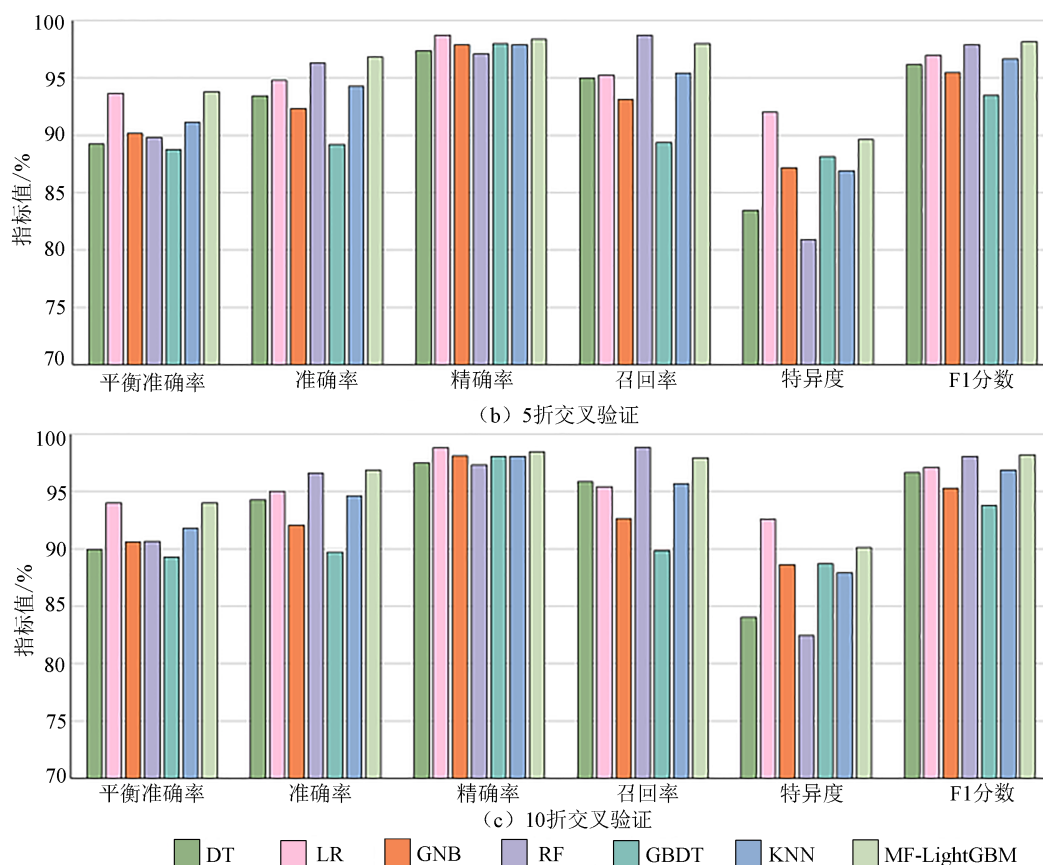


图3 不同模型交叉验证在测试集上的效果对比

Fig. 3 Performance comparison of different models on the test set using

二者通过极性相互作用和空间匹配共同促进共晶形成; 图4(c)显示, 样本三的大疏水片段为共晶形成的高贡献位点, 通过疏水相互作用或范德

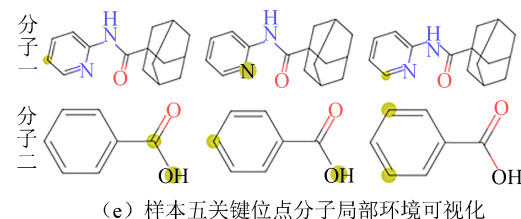
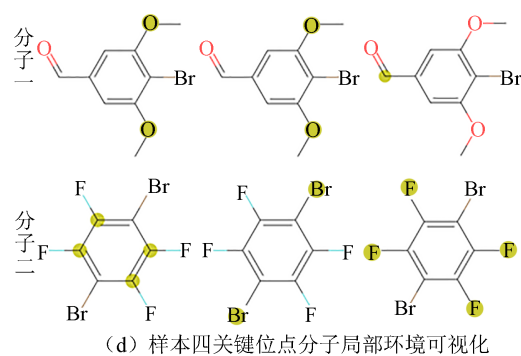
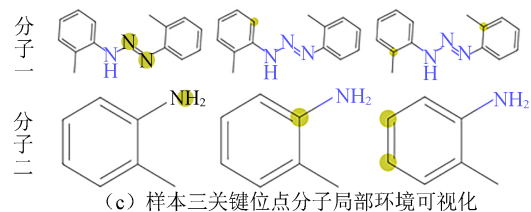
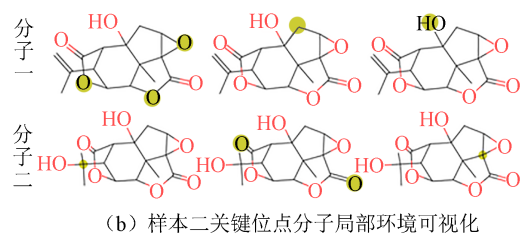
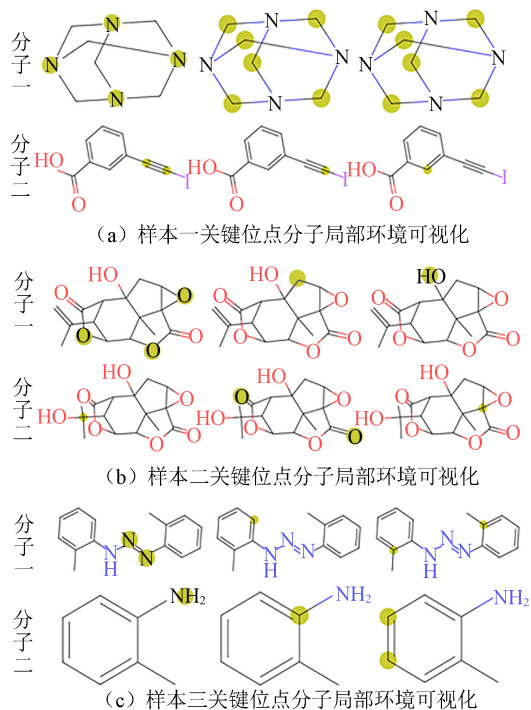


图4 代表性样本的可解释性分析结果

Fig. 4 Interpretability analysis results of representative samples

德华力促进紧密分子间堆积; 图4(d)显示, 在样本四中, 分子一的多官能团区域与分子二上相对简单的互补基团形成多重氢键网络; 图4(e)显示, 样本五的2个分子均通过多个弱作用位点的累

积效应实现共晶形成。这些可视化结果共同表明, MF-LightGBM 模型能够准确识别并量化符合超分子化学原理的关键结构片段, 如芳香环、氢键供受体、疏水区域及柔性/刚性骨架等。该分析框架提供了直接的基于结构的指导, 明确指出哪些分子局部子结构可能促进共晶形成, 从而为共晶体配体的合理设计或筛选提供依据。

3 结论

本文针对含能材料共晶形成预测中存在的筛选效率低、模型可解释性不足及数据集不平衡等难题, 构建了基于改进 LightGBM 的共晶形成可解释预测模型, 主要结论如下。

1) 在独立测试集上, 本文模型的平衡准确率达 95.15%、准确率达 97.16%、F1 分数达 98.35%, 各项核心指标均优于其他 6 种主流机器学习模型。

2) 模型泛化能力与鲁棒性突出。经 3、5、10 折交叉验证, 本文模型指标无明显波动, 对不平衡数据集具备良好的适配性。

3) 模型可解释性显著。基于 SHAP 理论的可解释性分析得到了各 Morgan 指纹位对预测结果的贡献度, 通过反向映射实现了关键作用位点的分子局部环境可视化, 明确了芳香环、氢键供受体、疏水区域等结构特征是驱动共晶形成的核心因素, 揭示了氢键互补、 π - π 堆积、疏水相互作用等非共价键作用的协同机制, 打破了传统机器学习模型的“黑箱”局限。

参考文献

- [1] Roshni J, Karthick T. A comprehensive review on theoretical screening methods for pharmaceutical cocrystals[J]. *Journal of Molecular Structure*, 2025, 1321: 139868.
- [2] Muravyev N V, Fershtat L, Zhang Qinghua. Synthesis, design and development of energetic materials: Quo Vadis?[J]. *Chemical Engineering Journal*, 2024, 486: 150410.
- [3] Li Tingting, Fu Siyao, Ding Shuaishuai, et al. Advancing room-temperature magnetic semiconductors with organic radical charge transfer cocrystals[J]. *Advanced Materials*, 2025, 37(10): 2414719.
- [4] Hasa D, Schneider R G, Voinovich D, et al. Cocrystal formation through mechanochemistry: from neat and liquid-assisted grinding to polymer-assisted grinding[J]. *Angewandte Chemie International Edition*, 2015, 54(25): 7371-7375.
- [5] Wu Di, Zhang Bei, Yao Qichao, et al. Evaluation on cocrystal screening methods and synthesis of multicomponent crystals: a case study[J]. *Crystal Growth & Design*, 2021, 21(8): 4531-4546.
- [6] Mohammad M A, Alhalaweh A, Velaga S P. Hansen solubility parameter as a tool to predict cocrystal formation[J]. *International Journal of Pharmaceutics*, 2011, 407(1/2): 63-71.
- [7] 王世杰, 杨杰, 马硕, 等. 机器学习在功能梯度材料设计-制备中的应用综述[J]. *材料导报*, 2023, 37(21): 111-119.
Wang Shijie, Yang Jie, Ma Shuo, et al. A review: applications of machine learning in design-fabrication of functionally graded materials[J]. *Materials Reports*, 2023, 37(21): 111-119.
- [8] 侯腾跃, 孙炎辉, 孙舒鹏, 等. 机器学习在材料结构与性能预测中的应用综述[J]. *材料导报*, 2022, 36(6): 165-176.
Hou Tengyue, Sun Yanhui, Sun Shupeng, et al. A review of the application of machine learning in material structure and performance prediction[J]. *Materials Reports*, 2022, 36(6): 165-176.
- [9] Wang Dingyan, Yang Zeen, Zhu Bingqing, et al. Machine-learning-guided cocrystal prediction based on large data base[J]. *Crystal Growth & Design*, 2020, 20(10): 6610-6621.
- [10] Jayachandiran R, Akkarapakam S J, Thangavel K. Virtual screening of pharmaceutical cocrystals using machine learning algorithms[J]. *Crystal Growth & Design*, 2025, 25(9): 3100-3109.
- [11] Yang Dezhi, Wang Li, Yuan Penghui, et al. Cocrystal virtual screening based on the XGBoost machine learning model[J]. *Chinese Chemical Letters*, 2023, 34(8): 107964: 424-429.
- [12] Jiang Yuanyuan, Yang Zongwei, Guo Jiali, et al. Coupling complementary strategy to flexible graph neural network for quick discovery of cofomer in diverse co-crystal materials[J]. *Nature Communications*, 2021, 12: 5950.
- [13] Cereto-Massagué A, Ojeda M J, Valls C, et al. Molecular fingerprint similarity search in virtual screening[J]. *Methods*, 2015, 71: 58-63.
- [14] Muegge I, Mukherjee P. An overview of molecular fingerprint similarity search in virtual screening[J]. *Expert Opinion on Drug Discovery*, 2016, 11(2): 137-148.
- [15] Janela T, Takeuchi K, Bajorath J. Introducing a chemically intuitive core-substituent fingerprint designed to explore structural requirements for effective similarity searching and machine learning[J]. *Molecules*, 2022, 27(7): 2331.

- [16] Ke Guolin, Meng Qi, Finley T, et al. LightGBM: a highly efficient gradient boosting decision tree[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. New York, USA: ACM, 2017: 3149–3157.
- [17] Mandler H, Weigand B. A review and benchmark of feature importance methods for neural networks[J]. ACM Computing Surveys, 2024, 56(12): 1–30.
- [18] Wu Kailang, Gu Jie, Meng Lu, et al. An explainable framework for load forecasting of a regional integrated energy system based on coupled features and multi-task learning[J]. Protection and Control of Modern Power Systems, 2022, 7(1): 349–362.
- [19] Zheng Guozhong, Zhang Yuqin, Yue Xuhui, et al. Interpretable prediction of thermal sensation for elderly people based on data sampling, machine learning and SHapley Additive exPlanations (SHAP)[J]. Building and Environment, 2023, 242: 110602.
- [20] 罗云, 徐嘉骏, 彭明阳, 等. 忠实性增强的国防政策检索增强生成方法[J]. 火箭军工程大学学报, 2026, 40(1): 132–142.
- Luo Yun, Xu Jiajun, Peng Mingyang, et al. Faithfulness-enhanced retrieval-augmented generation method for national defense policies[J]. Journal of Rocket Force University of Engineering, 2026, 40(1): 132–142.
- [21] 廖语诗, 苏娟, 马得草. 基于改进StarGAN网络的多时段红外图像反演算法[J]. 火箭军工程大学学报, 2025, 39(1): 21–30.
- Liao Yushi, Su Juan, Ma Decao. Improved StarGAN-based multi-period infrared image inversion algorithm[J]. Journal of Rocket Force University of Engineering, 2025, 39(1): 21–30.
- [22] 杨育婷, 李玲玲, 刘旭, 等. 基于多尺度-多方向Transformer的图像识别[J]. 计算机学报, 2025, 48(2): 249–265.
- Yang Yuting, Li Lingling, Liu Xu, et al. Multi-scale and multidirectional transformer-based image recognition[J]. Chinese Journal of Computers, 2025, 48(2): 249–265.
- [23] 孟小峰, 郝新丽, 马超红, 等. 科学发现中的机器学习方法研究[J]. 计算机学报, 2023, 46(5): 877–895.
- Meng Xiaofeng, Hao Xinli, Ma Chaohong, et al. Research on machine learning for scientific discovery[J]. Chinese Journal of Computers, 2023, 46(5): 877–895.
- [24] 张峰干, 何芳, 肖磊, 等. 基于拉普拉斯矩阵图的高光谱图像异常目标检测方法[J]. 火箭军工程大学学报, 2025, 39(1): 1–12.
- Zhang Fenggan, He Fang, Xiao Lei, et al. Laplacian matrix graph-based anomaly target detection method for hyperspectral graphs[J]. Journal of Rocket Force University of Engineering, 2025, 39(1): 1–12.
- [25] 王明, 鲜勇, 潘晨辉, 等. 基于大型语言模型与先验知识的兵棋决策方法[J]. 火箭军工程大学学报, 2025, 39(2): 79–88.
- Wang Ming, Xian Yong, Pan Chenhui, et al. Decision-making method for wargaming based on large language models and prior knowledge[J]. Journal of Rocket Force University of Engineering, 2025, 39(2): 79–88.
- [26] 申通, 王思宇, 杨小冈, 等. 多模态巡飞器时敏目标检测数据集构建[J]. 火箭军工程大学学报, 2026, 40(1): 68–78.
- Shen Tong, Wang Siyu, Yang Xiaogang, et al. Construction of an aircraft multimodal time-sensitive object-detection dataset[J]. Journal of Rocket Force University of Engineering, 2026, 40(1): 68–78.
- [27] Cambridge Crystallographic Data Centre. Cambridge Structural Database[DB/OL]. [2025–11–30]. <https://www.ccdc.cam.ac.uk/>.
- [28] Sarkar N, Mitra J, Vittengl M, et al. A user-friendly application for predicting the outcome of co-crystallizations[J]. CrystEngComm, 2020, 22: 6776–6779.
- [29] Khalaji M, Potrzebowski M J, Dudek M K. Virtual cocrystal screening methods as tools to understand the formation of pharmaceutical cocrystals a case study of linezolid, a wide-range antibacterial drug[J]. Crystal Growth & Design, 2021, 21(4): 2301–2314.
- [30] Vriza A, Canaj A B, Vismara R, et al. One class classification as a practical approach for accelerating $\pi-\pi$ co-crystal discovery[J]. Chemical Science, 2021, 12(5): 1702–1719.
- [31] Devogelaer J J, Charpentier M D, Tijink A, et al. Cocrystals of praziquantel: discovery by network-based link prediction[J]. Crystal Growth & Design, 2021, 21(6): 3428–3437.
- [32] Vriza A, Sovago I, Widdowson D, et al. Molecular set transformer: attending to the co-crystals in the Cambridge structural database[J]. Digital Discovery, 2022, 1(6): 834–850.

Interpretable Prediction Model for Cocrystal Formation Based on a Light Gradient Boosting Machine

LIU Yukun, LIU Yanfei, DU Ziang, LIU Bo

(Rocket Force University of Engineering, Xi'an 710025, China)

ABSTRACT: This study aimed to address the issues of low screening efficiency for energetic material cocrystals, insufficient interpretability of existing prediction models, and poor adaptability to imbalanced data-sets. An interpretable prediction model for cocrystal formation based on a lightweight gradient boosting machine (Light-GBM) was proposed, achieving high-precision and interpretable prediction of energetic material cocrystal formation. The model employed Morgan molecular fingerprints as input features to accurately characterize the local chemical environment of molecules; a LightGBM algorithm integrating histogram discretization, depth-constrained Leaf-wise splitting; and gradient-based one-side sampling was adopted for cocrystal prediction. An improved logarithmic loss function was introduced to balance the training weights of positive and negative samples, and the Bayesian optimization algorithm was utilized to determine the optimal hyperparameter combination of tree quantity and maximum depth. Furthermore, an interpretability analysis module was constructed based on the SHapley Addictive exPlanation (SHAP) theory to quantify the contribution of each fingerprint bit to the prediction results, screen out key Morgan fingerprint bits, and realize the visualization of the local molecular environment through reverse mapping. The results demonstrate that on a completely isolated independent test set, the model achieves a balanced accuracy of 95.15%, an accuracy of 97.16%, and an F1-score of 98.35%. After three-, five-, and ten-fold crossvalidation, the evaluation metrics exhibit minimal fluctuations, indicating superior generalization ability and robustness compared with other machine learning models. The interpretability analysis clarifies key structural features, such as aromatic rings, hydrogen bond donors and acceptors, hydrophobic regions, and their corresponding mechanisms, providing theoretical reference for the targeted screening and structural modification of cocrystal components.

KEYWORDS: cocrystal; interpretable prediction; lightweight gradient boosting machine; molecular fingerprint; SHapley Addictive exPlanation

(编辑: 刘建民)