研究论文

DOI：10.11949/0438−1157.20251294

数据与智能算法驱动的苯乙烯氢甲酰化催化剂制备与反应条件优化

马志[1,2]，朱吉钦[2]，柳萍[2]，许昊翔[1,2,3]，程道建[1,2,3]

（[1]北京化工大学有机无机复合材料全国重点实验室，北京 100029；[2]北京化工大学化学工程学院，北京 100029；[3]深度智化实验科技（北京）有限公司，北京 100071）

摘要： 烯烃氢甲酰化是制备高附加值精细化学品的重要反应，但其非均相催化剂配方的理性设计仍具挑战。为此，本研究构建并验证了一套数据与智能算法驱动的闭环优化框架，以加速催化剂筛选与性能预测。该框架基于75条苯乙烯氢甲酰化的文献数据，融合了特征工程、自动化超参数优化、主动学习与全局寻优算法。结果表明，凭借集成学习算法的鲁棒性，全量特征策略在模型性能上（Test R^2 = 0.9144）优于特征提取与特征选择方法。进一步引入主动学习进行以数据为中心的优化，使正异构比（$y_{n/i}$）代理模型提升至Test R^2 = 0.9347。SHAP可解释性分析表明，反应温度、配体/金属比和Rh质量分数是影响选择性的关键因素。最终，结合高精度的正异构比代理模型与并行的转化率（y_{cov}）代理模型（XGBoost， Test R^2 = 0.833），通过两阶段差分进化算法实现多目标全局寻优，获得可同时实现 $y_{n/i}$ = 1.813 与 y_{cov} = 95.28% 的最优配方。本工作提出的机器学习研究范式为小样本、多噪声条件下高性能催化剂的理性设计提供了有效途径。

关键词： 非均相催化；整体优化；代理模型；氢甲酰化；机器学习；可解释性分析

中图分类号： TQ 426　　**文献标志码：** A

文章编号： 0438−1157（2026）06−3404−20

Data−driven and machine learning−assisted optimization of catalyst preparation and reaction conditions for styrene hydroformylation

MA Zhi[1,2], ZHU Jiqin[2], LIU Ping[2], XU Haoxiang[1,2,3], CHENG Daojian[1,2,3]

([1] *State Key Laboratory of Organic−Inorganic Composites, Beijing University of Chemical Technology, Beijing* 100029, *China*; [2] *College of Chemical Engineering, Beijing University of Chemical Technology, Beijing* 100029, *China*; [3] *Deep Intelligence Experiment Technology (Beijing) Co., Ltd, Beijing* 100071, *China*)

Abstract: Olefin hydroformylation is an important reaction for the preparation of high−value−added fine chemicals, but the rational design of heterogeneous catalyst formulations remains challenge. To address this, this study constructs and validates a data− and intelligent algorithm−driven closed−loop optimization framework to accelerate catalyst screening and performance prediction. Leveraging a dataset of 75 literature−mined instances of styrene hydroformylation, the framework integrates feature engineering, automated hyperparameter optimization, active learning, and global optimization. Our results demonstrate that, underpinned by the robustness of ensemble learning, a full−feature strategy (Test R^2 = 0.9144) outperforms feature extraction and selection methods. The implementation of active learning for data−centric optimization further enhanced the surrogate model for the n/i

收稿日期： 2025−11−22　　**修回日期：** 2026−01−30

通信作者： 许昊翔(1991—)，男，博士，教授，xuhx@mail.buct.edu.cn；柳萍(1995—)，女，博士，讲师，liuping@buct.edu.cn

第一作者： 马志(2002—)，男，硕士研究生，2819236769@qq.com

基金项目： 国家自然科学基金项目(U23A20110, 22522802, 22278014)

引用本文： 马志, 朱吉钦, 柳萍, 许昊翔, 程道建. 数据与智能算法驱动的苯乙烯氢甲酰化催化剂制备与反应条件优化[J]. 化工学报, 2026, 77(6): 3404−3423

Citation: MA Zhi, ZHU Jiqin, LIU Ping, XU Haoxiang, CHENG Daojian. Data−driven and machine learning−assisted optimization of catalyst preparation and reaction conditions for styrene hydroformylation[J]. CIESC Journal, 2026, 77(6): 3404−3423

(normal-to-iso) ratio to a Test R^2 of 0.9347. SHAP analysis identified reaction temperature, ligand-to-metal ratio, and Rh mass fraction as the critical factors governing selectivity. Finally, by coupling the high-fidelity n/i surrogate with a parallel conversion model (XGBoost, Test R^2 = 0.833), a two-stage differential evolution algorithm performed multi-objective global optimization. This identified an optimal formulation predicted to simultaneously achieve an n/i ratio of 1.813 and 95.28% conversion. This work presents a machine learning paradigm that provides an effective pathway for the rational design of high-performance catalysts, particularly under the constraints of small and noisy datasets.

Keywords: catalysis; global optimization; surrogate model; hydroformylation; machine learning; interpretability analysis

引　言

氢甲酰化反应，也称为“羰基合成”，是工业上最重要的催化反应之一，其核心价值在于将烯烃和合成气（CO/H_2）转化为高附加值的醛类化合物[1]。这些醛是生产增塑剂、醇类以及多种精细化学品的关键中间体。比如苯乙烯氢甲酰化的支链产物2-苯基丙醛，是合成药物（如布洛芬）和香料的重要前体[2]。然而，传统的均相催化剂（如Rh/Co配合物）存在难以分离和回收的问题，导致高成本和潜在的环境污染。因此，开发高效稳定、易于分离回收的非均相催化剂是实现该过程绿色化和工业化的关键路径[3]。非均相催化剂的设计与优化面临着巨大的挑战。其催化性能不仅受活性中心（如金属纳米颗粒或单原子）的影响，更高度依赖于一系列高维度的制备条件（如载体类型、金属负载量、焙烧温度、还原气氛）和反应条件（如温度、压力、溶剂、原料配比）。这些参数之间存在复杂的耦合与协同效应，构成了一个高维且非线性的参数空间。在过去，催化剂的优化主要依赖于“试错法”和基于化学家直觉的“理性设计”。这种方法耗时耗力，传统的一次一变量的实验方法很难捕捉到复杂体系的全貌，导致优化过程往往陷入局部最优解。

智能算法能够分析稀疏且昂贵的实验数据，构建可靠的“配方-性能”预测模型，并智能地指导下一轮实验，从而以最少的实验次数逼近全局最优解。这种数据驱动的策略已经在多个催化领域展现了其强大的能力，典型的应用案例包括：基于描述符预测催化性能[4-5]、筛选催化剂配方或优化工艺参数[6-7]、用于提升催化性能的酶工程[8-9]以及辅助（微）动力学模型确定反应网络和机理[10-11]。尽管数据驱动方法在催化领域已初见成效，但将其应用于非均相催化时，面临着独特的困境。用于训练代理模型的数据主要来源于已发表的文献和历史数据库，这类数据具有“小数据”和“异构性”的典型特征——实验条件不统一、报道规范不一致。在这类数据特征下，如何构建一个能够准确捕捉“制备-结构-性能”关系的机器学习模型，并避免高维参数优化陷入过拟合，是本领域亟待解决的难题。此外，从非结构化的文献数据中提取有意义的、标准化的催化剂特征工程，并建立专门的催化数据库[12-14]，是构建可靠预测模型的关键瓶颈与前提。

本工作流程始于对现有文献数据的标准化处理与特征工程。随后，系统地比较包括随机森林、梯度提升在内的多种机器学习算法，并利用主动学习策略对训练集进行迭代优化，解决数据稀缺问题。同时，利用Optuna[15]自动机器学习框架对模型超参数进行精细调优，构建最优的“黑箱”预测模型。更进一步地，将此最优模型作为“代理模型”，嵌入到一个进化优化算法的优化框架中。该平台能够智能地分析包含高维制备条件和反应条件的复杂参数空间，其最终目标是实现多目标最大化，即优化产物正异构比和转化率，从而快速定位出全局最优的工艺条件组合。

1　方法

本研究旨在预测并优化非均相催化剂在苯乙烯氢甲酰化反应中的性能。为此，本文构建了一套数据驱动的闭环优化框架，如图1所示。该框架以集成多种机器学习算法的代理模型为核心，通过主动学习策略优化训练集构建，并最终实现催化剂性能的全局寻优。整体工作流程包含以下6个关键步骤：数据准备、特征工程、模型开发、主动学习策略、可解释性分析及全局寻优。

1.1　数据准备与标准化

本研究的数据来源于Web of Science (WoS) 学

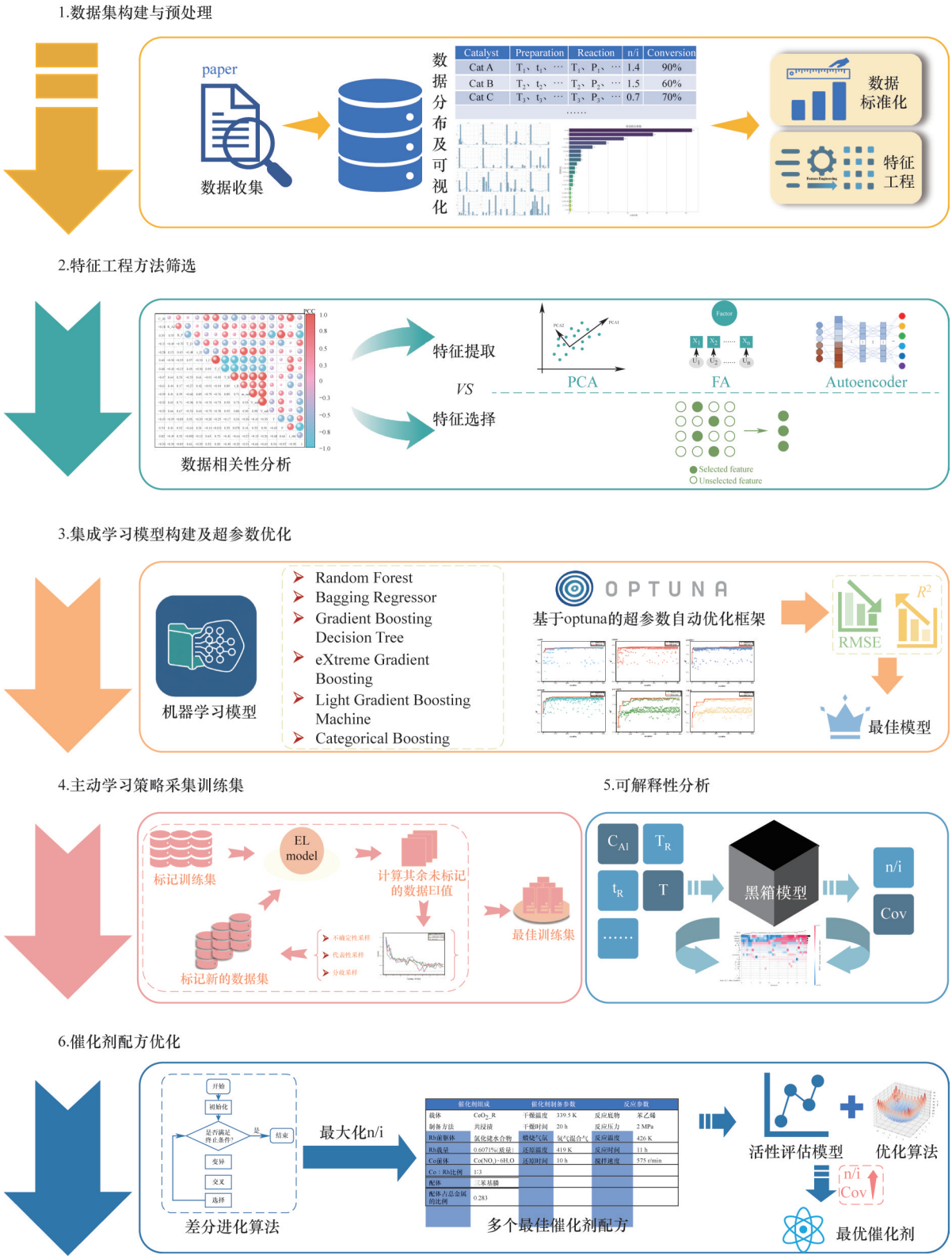


图1 数据与智能算法驱动的催化剂配方优化工作流程图

Fig.1 Workflow of machine learning-driven catalyst recipe optimization

术数据库，检索关键词为“supported rhodium-based catalyst”、“styrene”、“hydroformylation”及“heterogeneous catalysis”。在数据库选择上，本研究优先选取WoS作为核心数据源，主要基于对数据质量与特征提取可行性的考量。一方面，WoS收录期刊经过严格筛选，能有效减少因数据源异构性带来的噪声，这对

本研究中小样本主动学习建模的鲁棒性至关重要；另一方面，本工作涉及包含详细制备条件（如干燥、焙烧、还原参数）在内的38维复杂特征提取，WoS提供的完善全文索引和结构化元数据，为实现深度人工筛选与清洗提供必要的基础。通过严格的人工筛选与数据清洗，最终获得包含38个输入特征维度的数据集（表1）。其中，载体类型涵盖15种不同材料，包括氧化物、改性氧化物等。为了确保特征工程中制备参数的一致性与可比性，本数据集严格限定收录基于湿法浸渍路线制备的催化剂样本。据此，本模型的适用范围明确界定为：负载型Rh单金属及Rh-M (M=Co, Pd, Ru 等)双金属/合金纳米颗粒体系。对于单原子催化剂或经由特殊多步工艺控制的严格核壳结构，除非其通过常规浸渍参数（如低负载量控制或特定热处理）即可获得并在特征空间中有所体现，否则不作为本模型的主要预测对象。

苯乙烯氢甲酰化反应相关数据的收集主要聚焦于4个方面，即催化剂属性、制备条件、反应参数与催化性能。其中，催化剂属性、制备条件及反应参数作为机器学习模型的输入变量，催化性能则作为输出变量。值得说明的是，输入变量主要来源于原始文献中的文本描述，而催化性能数据则通过WebPlotDigitizer软件从图表中提取获得。在对数据进行系统整理与一致性校验后，最终构建了包含75条有效实验记录的人工数据集，为后续的机器学习建模与性能预测提供了可靠的数据基础。

为了确保数据的一致性并消除异构性对模型训练的干扰，对原始数据执行了严格的预处理流程，首先，对来自不同文献的数据进行审查，剔除明显错误或关键信息缺失的数据点。随后，将所有物理量（如压力、温度）统一为国际标准单位，以消除量纲差异。

为确保数据的完整性与模型训练的准确性，对缺失值进行了系统处理。具体策略如下。

（1）温度相关变量（如煅烧温度、还原温度）：当文献提及进行了热处理操作（煅烧或还原）但未提供具体温度时，以该变量在数据集中的均值进行填充；若文献明确指出未进行热处理，则以室温（296.75 K）替代，以符合催化剂在未处理状态下处于环境温度的物理事实。该策略并非简单的数值填补，而是旨在还原真实的物理状态，同时确保温度变量与对应的处理时间（赋值为0）在特征空间中保持逻辑自洽性，有助于模型准确区分“煅烧”与“未煅烧”的工艺差异。

（2）热处理时间变量（如煅烧时间）：当文献中提及进行了煅烧但未报告具体时间时，采用该变量的均值进行填充；若文献明确未进行煅烧，则将该变量赋值为0。

（3）其他参数：对于未明确报道的其他特征参数，依据其数据类型与分布特征，采用常规的统计方法（如均值或众数填充）进行合理插补。

此外，由于苯乙烯氢甲酰化数据集包含多维输入特征，不同特征间的数值量级存在显著差异（例如，温度与催化剂负载量的数量级差距较大），若直接用于模型训练，可能导致部分特征在学习过程中占据主导权重，从而影响模型的整体预测性能。为

表1 用于开发苯乙烯氢甲酰化反应机器学习模型的输入与输出特征

Table 1 Input and output features used to develop the machine learning model for the styrene hydroformylation reaction

类别	特征的详细描述	缩写	单位
催化剂属性	活性金属1理论负载量	C_A_1	%（质量）
	活性金属2负载比例	R_A_2	mol/mol
	助剂负载比例	R_P	mol/mol
	活性金属1-Rh	A_1_Rh	—
	活性金属2-Co	A_2_Co	—
	活性金属2-Pd	A_2_Pd	—
	活性金属2-Ru	A_2_Ru	—
	助剂-P	P_P	—
	助剂-S	P_S	—
	载体类型	S_type	—
制备条件	干燥温度	T_D	K
	干燥时间	t_D	h
	焙烧温度	T_C	K
	焙烧时间	t_C	h
	还原温度	T_R	K
	还原时间	t_R	h
	还原剂类型	R	—
反应参数	催化剂用量	m_cat	mg
	底物用量	V_sub	ml
	溶剂用量	V_sol	ml
	反应温度	T	K
	反应压力	P	MPa
	搅拌速率	v_stir	r/min
	反应时间	t	h
催化性能	正异比	n/i	—
	转化率	cov	%

此，在模型训练前需对输入特征进行标准化处理。

考虑到本研究数据来源于多源独立文献，不同实验室的操作习惯差异可能引入非物理的系统性偏移或离群值。传统的Z-score标准化依赖于均值和标准差，极易受此类异常值影响而导致分布失真。因此，本研究采用了基于scikit-learn机器学习库[16]实现的鲁棒标准化（RobustScaler）方法，对全部特征数据进行归一化处理，该方法利用数据的中位数和四分位距进行缩放，其计算公式如式（1）所示。

$$X' = \frac{X - \mathrm{Median}(X)}{\mathrm{IQR}(X)} \tag{1}$$

式中，X为原始特征值；X'为标准化后的特征值；Median(X)表示特征X的中位；IQR(X)表示该特征的四分位距，即上四分位数与下四分位数之差。通过引入该方法，可有效压缩因文献来源差异产生的极端异常值对特征整体分布的干扰，在保留数据物理结构的同时，显著提高了模型在异构数据上的训练稳定性与收敛速度。

1.2 特征工程

特征工程是机器学习建模过程中至关重要的环节，其核心目标在于通过对原始数据的系统分析与优化处理，最大程度地挖掘和表达数据中所蕴含的有效信息，从而提升模型的学习效率与预测性能[17]。

本研究针对苯乙烯氢甲酰化数据集中多维特征较多、特征间相关性较强的特点，分别采用特征选择与特征提取2种方式以降低输入特征数量，并对比分析2种方法在模型性能与可解释性方面的优势与局限性。从催化剂配方优化的角度出发，特征工程不仅需要提升模型的预测精度，更应尽可能保持特征的物理化学可解释性，以实现数据驱动结果与实验现象之间的有机联系。因此，除特征选择与特征提取外，本研究还设置了第3种对照方式——在保持全部原始特征的前提下，仅进行数据标准化处理，以保证模型训练效果的同时，不对特征空间进行额外降维。

通过上述3种方案的系统比较，可以全面评估不同特征处理策略对模型性能、可解释性及其在催化体系优化中的适用性，为后续催化剂性能预测与组成的优化提供方法学参考。

1.2.1 相关性分析 在进行特征降维与建模之前，为避免多重共线性对模型训练造成干扰，本研究首先对输入变量之间的相关性进行了系统分析。通过计算皮尔逊相关系数，定量评估各输入特征之间以及输入特征与输出变量之间的线性相关程度。皮尔逊相关系数（PCC）的计算如式（2）所示。

$$\mathrm{PCC} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{2}$$

式中，x_i与y_i分别为变量X与Y的第i个样本值；$\bar{x}$与$\bar{y}$为其均值；n为样本总数。系数PCC的取值范围为[−1,1]，其绝对值越接近1，表示变量间线性相关性越强。

在实际分析中，本研究分别计算了各输入特征间的两两相关性矩阵，并绘制热力图以辅助识别显著相关的特征对。若某两特征间的相关系数绝对值超过设定阈值（如0.85），则判定其存在较强的线性相关性。在后续建模中，对此类高相关特征进行适度筛除或合并，以降低信息冗余，防止模型在训练过程中出现参数偏倚和过拟合现象。

1.2.2 特征提取 为进一步降低数据维度并提取具有代表性的信息特征，本研究采用3种特征提取方法进行对比分析[18]，分别为主成分分析（PCA）、因子分析（FA）以及自编码器（AE）。3种方法均旨在将高维输入变量映射至低维潜在空间，以减少特征冗余并提高模型训练的效率与稳定性。

其中，PCA通过对特征协方差矩阵进行特征值分解，提取能够最大化样本方差的正交主成分，从而在保留主要信息的同时有效压缩特征维度；FA侧重于挖掘潜在共性因子以解释变量间的内在关联，适用于探讨催化体系中实验参数的耦合关系；AE则基于神经网络结构，通过非线性映射自动学习特征的低维表示，能够捕捉复杂的非线性特征模式，在信息重构与模式识别方面具有更高的表达能力。

为系统评估3种方法的特征提取效果，本研究采用均方误差（MSE）、方差解释率（EVR）和信噪比（SNR）3项指标进行性能比较。其中，MSE反映特征重构的精度，EVR衡量降维后保留的有效信息比例，SNR则用于评价特征空间的稳健性与噪声抑制能力。通过综合这3项指标，可全面比较不同特征提取方法在信息保持、模型训练稳定性及物理可解释性方面的表现，为后续的模型优化与方法选择提供依据。

1.2.3 特征选择 除特征提取方法外，本研究还采用特征选择策略以实现输入特征的降维与优化。与特征提取通过构建新特征空间不同，特征选择旨

在直接从原始特征集中筛选出对模型预测最具贡献的关键变量，从而在保留数据物理可解释性的同时，减少冗余信息并提升模型的泛化能力。

在具体实现上，本文采用递归特征消除交叉验证（recursive feature elimination with cross-validation, RFECV）方法[19]。基于模型对各特征的重要性评分，逐步移除贡献较低的特征，并在每一步中通过交叉验证评估模型性能。模型性能的评估指标包括均方根误差（RMSE）和决定系数（R^2）等。最终选取使模型性能最优的特征组合作为输入变量，用以构建预测模型。

RFECV的优点在于能在保证模型预测精度的同时显著减少冗余特征，从而提高模型泛化能力与训练效率。此外，特征选择所得变量均具有明确的物理意义，能够与催化反应参数、制备条件及性能指标形成直接对应关系，为后续实验优化提供理论支持。综合特征提取与特征选择2类方法的结果，本文系统比较不同特征处理策略在模型性能提升与可解释性保持方面的适用性与局限性，为催化剂数据驱动优化提供方法论参考。

1.3 机器学习模型的开发与优化

本研究构建并对比了6种常用的回归型机器学习模型，包括随机森林（random forest, RF）[20]、装袋回归器（bagging regressor, BR）[21]、梯度提升决策树（gradient boosting decision tree, GBDT）[22]、轻量梯度提升机（light gradient boosting machine, LGBM）[23]、极限梯度提升（eXtreme gradient boosting, XGB）[24]以及类别提升（catBoost, CB）[25]模型。上述模型均基于集成学习原理，通过组合多个基学习器以提升整体模型的预测性能与鲁棒性。

1.3.1 *集成模型的开发* 集成学习模型在催化性能预测领域具有显著优势。与单一模型相比，集成方法通过“多模型决策”机制可有效降低单模型的偏差与方差，实现更优的泛化能力。其中，RF与BR模型代表典型的并行集成策略（bagging），通过对样本的重采样与多棵决策树结果的平均，降低模型的方差并提升稳定性；而GBDT、LGBM、XGB与CB则属于序列集成策略（boosting），通过逐步拟合前一模型的残差以不断提升预测精度。

此外，不同集成算法在处理特征非线性关系、噪声数据以及特征重要性评估等方面各具特点，为模型解释性与优化方向提供了互补性基础。因此，本研究系统比较了多种集成模型在催化剂性能预测中的表现，以获得兼顾预测精度与可解释性的最优方案。

1.3.2 *模型的超参数优化* 机器学习模型的性能在很大程度上依赖于超参数的合理设定[26]。传统的网格搜索与随机搜索虽可实现超参数调优，但搜索效率较低且难以在高维空间中获得全局最优解。为此，本研究引入基于Optuna框架[15]的自动化超参数优化策略。

Optuna通过贝叶斯优化与自适应采样机制动态选择搜索空间内的高潜力区域，实现超参数的智能搜索与迭代优化。与传统方法相比，Optuna不仅显著降低了计算成本，还能在有限样本条件下获得更优的模型性能。在优化过程中，研究以交叉验证的平均均方根误差为目标函数，自动寻优模型的关键参数（如树深度、学习率、子采样率等），确保不同模型在相同数据条件下均处于最优性能状态。

1.3.3 *模型评估* 为系统评估6种基于特征提取与选择的机器学习模型的预测精度，本研究采用RMSE与R^2作为主要性能评估指标[27]。RMSE反映模型的预测误差，值越小表示模型拟合效果越好；而R^2则衡量模型对样本数据方差的解释程度，值越接近1表示模型能够更好地捕捉数据的变化规律。具体的计算如式（3）和式（4）所示。

$$\mathrm{RMSE}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\widehat{y_i})^2} \tag{3}$$

$$R^2=1-\frac{\sum_{i=1}^{n}(y_i-\hat{y})^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2} \tag{4}$$

式中，n为样本数量；y_i与$\hat{y}$分别表示第i个样本的真实值与预测值；$\bar{y}$为数据集的均值。此外，为进一步直观展示各模型的预测性能，本研究还采用回归拟合图对模型的预测结果与实际值进行可视化，帮助分析不同模型在样本数据上的拟合趋势和误差分布。通过比较各模型的RMSE与R^2值，可以全面评估不同算法在预测精度、稳定性以及泛化能力等方面的表现。模型的残差分布与拟合趋势进一步验证了其在苯乙烯氢甲酰化反应数据集上的应用效果，为模型选择与优化提供理论依据。

1.4 主动学习策略的开发

在小样本学习中，数据的质量与模型的训练效果密切相关。由于样本量较小，数据中的噪声和异常值可能会对模型的性能产生显著影响，因此，在模型训练初期，如何有效选择与标注最具代表性或

信息量最大的样本，成为提升模型预测精度的关键[28]。传统的机器学习方法通常依赖于大规模标注数据，而在小样本学习的情况下，获取更多标注数据不仅困难且成本高昂，因此，通过主动学习策略优化训练集，选择最具代表性的样本进行标注，能够显著提高模型的训练效率和性能。

本研究提出一种基于主动学习策略的训练集优化方法，旨在减少小样本数据中的噪点对整体模型性能的负面影响。通过主动选择并标注最有价值的数据点，能够构建一个既能充分代表数据分布，又能最大化模型预测能力的高效训练集。为此，采用3种常见的主动学习策略：不确定性采样、代表性采样和基于分歧的采样。

不确定性采样方法关注模型在预测时的“不确定”样本，尤其是那些位于决策边界附近的样本。通过优先标注这些样本，可以优化决策边界的准确性，有效减少模型在小样本情况下受到噪声影响的风险，进而提升模型的稳定性和泛化能力[29]；代表性采样方法从数据分布的全局角度出发，选择那些能够代表整个数据集分布的样本。该策略有效避免了数据中的冗余和偏差，有助于减少因噪声数据或局部偏差导致的模型过拟合，增强模型对全局模式的捕捉能力[30]；基于分歧的采样方法利用多个模型间的预测差异来选择最具争议的样本。这些样本通常位于模型的认知盲区，标注这些样本能够有效帮助模型避免局部最优解，探索那些难以被单一模型捕捉的复杂模式。这一方法特别适用于小样本学习中的样本不足问题，能够促进模型的更全面学习[31]。

通过这3种主动学习策略，能够在局部、全局以及模型协作层面优化训练集，全面提升模型的性能与泛化能力。在小样本学习中，主动学习的应用能够有效减少噪声数据对模型训练的影响，避免过拟合，并使得有限的样本数据在模型学习中发挥最大价值。这些策略不仅减少了冗余数据的影响，还能够显著提升训练效率，避免“黑箱”干扰，同时保持数据的可解释性和实验指导性。

1.5 可解释性分析

随着机器学习方法在催化剂性能预测中的广泛应用，如何解释复杂的黑箱模型变得尤为重要。本研究采用了2种主流的可解释性分析方法：全局解释的SHAP（shapley additive explanations）[32-33]方法和局部解释的PDP（partial dependence plot）[22]方法。SHAP是一种基于博弈论的特征重要性评估方法，通过计算每个特征对模型输出的边际贡献来量化各特征在模型决策中的作用。SHAP值能够为每个数据点提供一致且精确的特征贡献评分，使得复杂的集成模型具有更高的透明度和解释性，帮助研究者理解哪些特征对催化剂性能预测具有重要影响。相比之下，PDP方法则侧重于局部解释，旨在揭示单一特征（或特征组合）与预测结果之间的关系。通过绘制特征变化与模型输出的依赖图，PDP能够帮助我们分析特定变量如何在不同范围内影响催化剂性能的预测。结合SHAP和PDP 2种方法，能够从全局和局部2个层面全面评估模型的预测机制，提供深入的可解释性分析，进一步提升模型的可信度和实验设计的指导价值。

1.6 催化剂配方、制备条件与反应参数优化

在催化剂优化阶段，本研究采用了前期构建的最优机器学习模型（基于最佳特征策略与优化训练集），将其作为代理模型，并结合差分进化算法（DE）[34]实现催化剂配方、制备条件及反应参数的全局优化。差分进化算法是一种基于种群进化的全局随机搜索方法，能够在高维复杂空间中有效寻优，其核心机制是通过个体间差分变异与选择迭代不断逼近全局最优解。催化剂优化过程被设计为一个双阶段优化框架，目标在于平衡产物选择性与反应转化率两方面的性能表现。整体优化问题可形式化表示为式（5）。

$$\max_{\boldsymbol{x}\in R^n} f\{\boldsymbol{x}\} = \{y_{\mathrm{n/i}}, y_{\mathrm{cov}}\} \tag{5}$$

式中，$\boldsymbol{x}$表示输入变量向量，涵盖催化剂配方参数（如金属组分及负载量）、制备条件（如煅烧温度、还原温度与时间）及反应条件（如反应温度、压强与气氛组成）；$y_{\mathrm{n/i}}$与y_{cov}分别表示代理模型预测的正构/异构醛比与转化率。

本研究的第一阶段是正构/异构醛比优化，该阶段以训练好的“正异构比”预测模型为核心，通过差分进化算法在高维参数空间中全局搜索，目标为最大化$y_{\mathrm{n/i}}$即：

$$\max_{\boldsymbol{x}} f_1\{\boldsymbol{x}\} = y_{\mathrm{n/i}}(\boldsymbol{x}) \tag{6}$$

此阶段的优化结果用于识别能实现最高产物选择性的催化剂配方及其对应的制备与反应参数组合。

本研究的第二阶段是转化率优化，在第一阶段获得的最优结果集中，筛选出正构/异构醛比排名前100的参数组合，形成候选子空间。在此基础上，以

训练好的“转化率”预测模型为代理，再次利用差分进化算法进行二次优化，目标为最大化y_{cov}即：

$$\max_{\boldsymbol{x}\in S_{top100}} f_2\{\boldsymbol{x}\} = y_{cov}(\boldsymbol{x}) \tag{7}$$

式中，S_{top100}表示由第一阶段筛选出的高选择性样本子集。通过该两阶段优化框架，本研究在优先确保产物选择性的前提下，进一步最大化反应效率，从而获得兼顾选择性与活性的最优催化体系。该策略为后续实验验证与新型催化剂设计提供了高精度的理论依据与优化路径。

2 结果与讨论

2.1 数据可视化分析

在构建机器学习模型之前，首先对所收集的75条实验数据进行了探索性数据分析（EDA），以深入理解数据集的内在特征与分布规律。这有助于揭示数据集中存在的潜在问题，如数据不平衡、偏态分布或异常值，从而为后续的数据预处理和特征工程提供依据。图2展示了数据集中主要类别特征与数值特征的总体分布情况，图2(a)为类别特征的频数统计，图2(b)为代表性数值特征（活性金属1理论负载量、反应压力、催化剂用量、反应时间）的直方图及核密度估计曲线。其余特征的分布结果见图S1。

从类别特征分布来看，数据集呈现出明显的不均衡性。活性金属铑与载体二氧化硅是最常见的组合，分别出现63次和53次，构成了数据集的主体；而其他载体与配体类型出现频次较低，部分类别仅有1~2条记录。这种样本分布的稀疏性使得模型在训练过程中可能难以充分学习稀有类别与催化性能之间的潜在关联规律，从而增加模型泛化的不确定性。

数值特征的分布同样显示出较强的非均一性。图2(b)的结果表明，多数数值特征偏离正态分布。其中，反应压力呈现明显的双峰特征，可能反映了文献中常见的两类实验条件（2.0 MPa与3.0 MPa）；而特征活性金属1负载量和催化剂用量则表现出右偏态分布，说明大多数实验集中在较低取值区间，同时存在少量较高取值的极端样本，提示数据中存在一定的异常值。

综合上述可视化分析结果可以看出，数据集在类别分布与数值特征上均呈现出明显的不均衡特征，这为后续特征预处理方法与建模策略的选择提供了直接依据。针对多数连续特征存在的偏态分布及离群值问题，本章节明确采用RobustScaler作为统一的预处理方法。与传统的Z-score标准化相比，该方法基于中位数和四分位距进行缩放，能够有效降低极端值对特征尺度的影响，从而更稳定地保留不同样本在数值空间中的相对差异。

进一步地，EDA结果显示，活性金属与载体等关键类别特征在样本中的分布存在显著不均衡，部分类别仅占据较小比例。这一特征意味着依赖单一建模手段可能在训练过程中对低频样本关注不足，从而影响模型的整体泛化性能。基于这一认识，本文在后续建模过程中优先选择了对类别特征分布不敏感、且具备内生正则化机制的CatBoost集成模型（见2.3与2.4节）。同时，针对样本信息密度差异较大的问题，2.5节进一步引入主动学习策略，通过不确定性采样引导模型优先学习信息量较高或预测不稳定的样本，以缓解数据不均衡对模型泛化能力可能带来的影响。

2.2 特征相关性分析

在对数据总体分布特征进行初步了解后，为进一步探究各输入变量之间的相互关系，并为后续的特征降维提供依据，本研究计算了所有数值型输入变量的皮尔逊相关系数（Pearson correlation coefficient, PCC），结果如图3所示。该分析的核心目的在于识别可能导致多重共线性问题的高度相关特征对，从而为模型构建与特征筛选提供数据支撑。

从图3所示的热力图可以看出，数据集中存在明显的共线性关系，尤其在催化剂制备参数与反应体系参数之间表现突出。首先，热处理相关变量之间的强耦合性尤为显著。煅烧时间（t_C）与还原时间（t_R）的PCC高达0.94，同时t_C还与煅烧温度（T_C, PCC = 0.89）和还原温度（T_R, PCC = 0.91）高度相关，而T_C与T_R之间的相关系数亦达到0.90。这一现象符合催化剂制备过程的物理化学规律：高温煅烧样品往往伴随高温还原处理，且两者的处理时间通常同步延长，反映出实验设计中的参数联动特征。

此外，反应体系参数之间同样存在较强的相关性。溶剂用量（V_sol）与底物用量（V_sub）之间的PCC达到0.96，表明研究者在实验设计中普遍保持底物浓度恒定，即通过按比例调整溶剂与底物用量来维持一致的反应体系条件。类似地，反应压力与

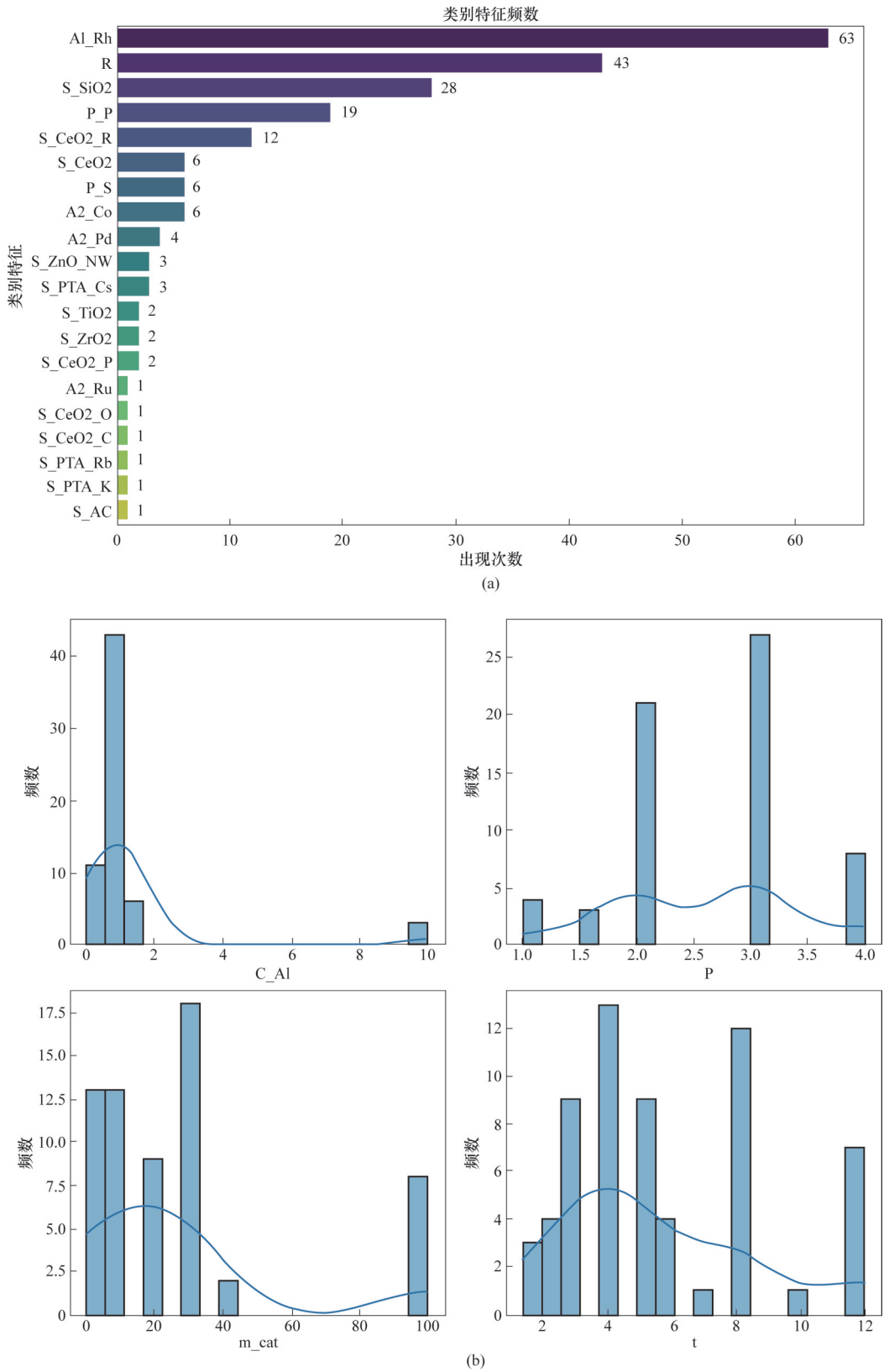


图2 数据集的数据分布情况:(a) 类别特征的频数统计,包括活性金属理论负载量、反应压力、催化剂用量、反应时间等;(b) 数值型特征的分布,包括活性金属理论负载量、反应压力、催化剂用量、反应时间等

Fig.2 Overview of the data distribution: (a) frequency distribution of categorical features, including theoretical active metal loading, reaction pressure, catalyst amount, and reaction time; (b) distribution of numerical features, including theoretical active metal loading, reaction pressure, catalyst amount, and reaction time

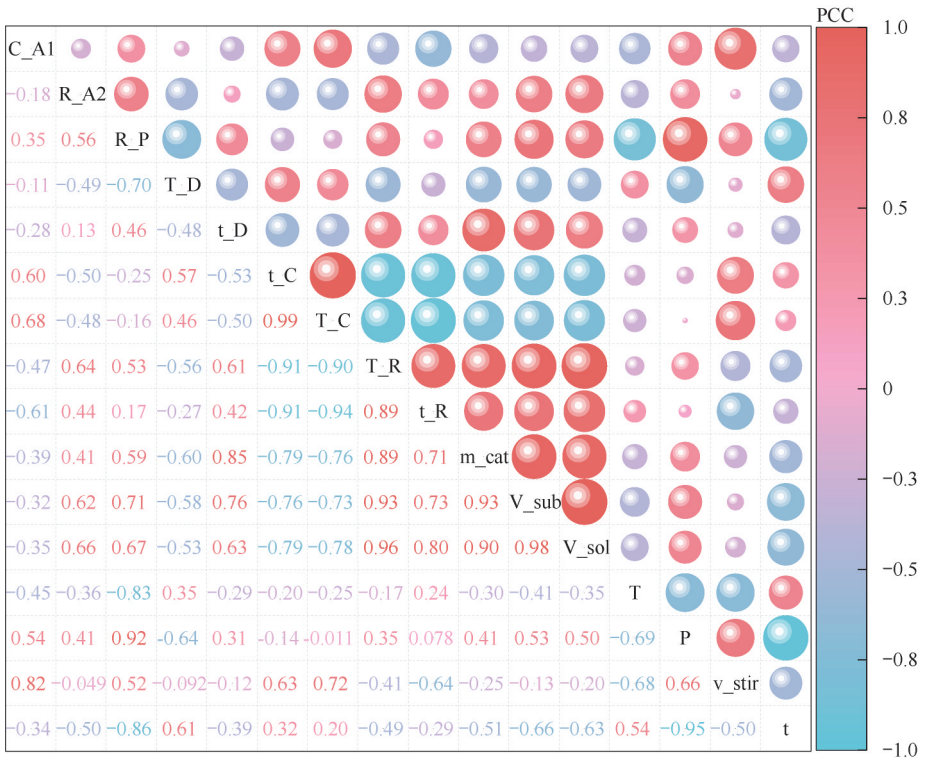


图3 输入特征的皮尔逊相关系数热力图

Fig.3 Pearson correlation coefficient heatmap of the input features

V_sub(PCC = 0.93)以及V_sol(PCC = 0.90)均呈现强正相关,暗示在高压条件下通常伴随底物投料量与体系体积的相应增加,以保证体系稳定性与可比性。

综合来看,图3中存在大量PCC绝对值超过0.85的特征对,说明数据集中存在显著的信息冗余。如果不加以处理直接建模,多重共线性不仅会增加模型计算复杂度,还可能导致参数估计不稳定,并干扰对特征重要性的准确评估,使模型在解释层面出现偏差。

因此,特征相关性分析的结果强有力地印证了1.2节中所提出特征工程步骤的必要性。针对高度共线的特征,应通过适当的降维策略(如1.2.2节的特征提取方法或1.2.3节的特征选择方法)来削弱或消除共线性影响,从而构建结构更精简、稳定性更高且具有更强可解释性的机器学习预测模型。

2.3 特征工程策略

为构建具有高精度与可解释性的机器学习模型,本研究针对原始的38个输入特征,系统比较了3种特征工程策略:(1)特征提取(以主成分分析PCA为代表),(2)特征选择(递归特征消除交叉验证RFECV),(3)全量特征策略(仅进行标准化处理,不进行降维)。本节旨在通过对比6种机器学习模型:RF、BR、GBDT、XGB、LGBM、CB,在不同特征策略下的性能表现,筛选出最佳的特征功策略,用于后续的可解释性分析与全局寻优。

2.3.1 降维策略的适用性评估 研究首先从特征提取方法展开探索。表S1对比了3种常见的特征提取技术:PCA、FA、AE。结果表明,PCA在特征重构精度(最低MSE)和信噪比(最高SNR)两项指标上均表现最佳,因此,被选作后续降维分析的代表方法。

然而,当基于PCA特征训练各类模型时(见图S2),模型整体性能表现不佳。6种机器学习模型在PCA处理后的特征集上均未表现出理想预测精度,即使是性能最优的CB模型[图S2(f)],其交叉验证R^2仅为0.629。该结果表明,PCA作为无监督线性变换方法,虽然能有效压缩数据方差,但在保留与催化性能高度相关的关键信息方面存在明显不足。

相较之下,监督式特征选择方法RFECV展现出更优的建模表现。通过交叉验证确定最优特征子集(图S3),RFECV在特征数量与模型性能间取得了良好平衡。例如,在GBDT模型下[图S3(c)],当特征数减少至34个时,其交叉验证R^2达到0.742,明显优于任意PCA方案下的结果。

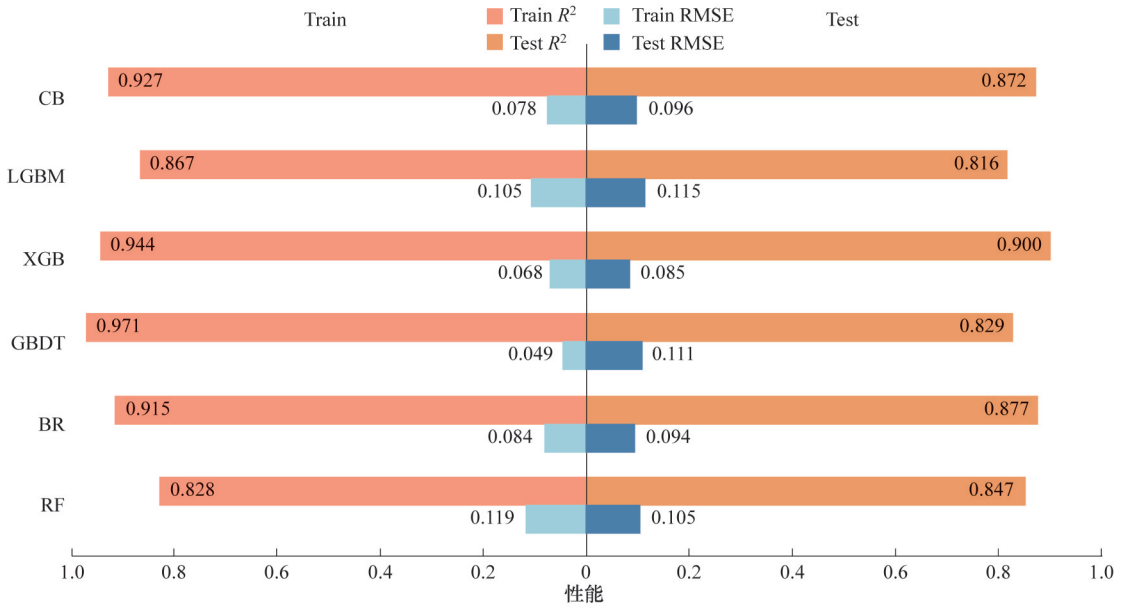
这一对比结果充分说明,以模型预测精度为导

向的监督式“选择”方法，相较于基于方差压缩的无监督“提取”策略，能更有效地保留与反应性能密切相关的信息，并有效剔除噪声与冗余变量。

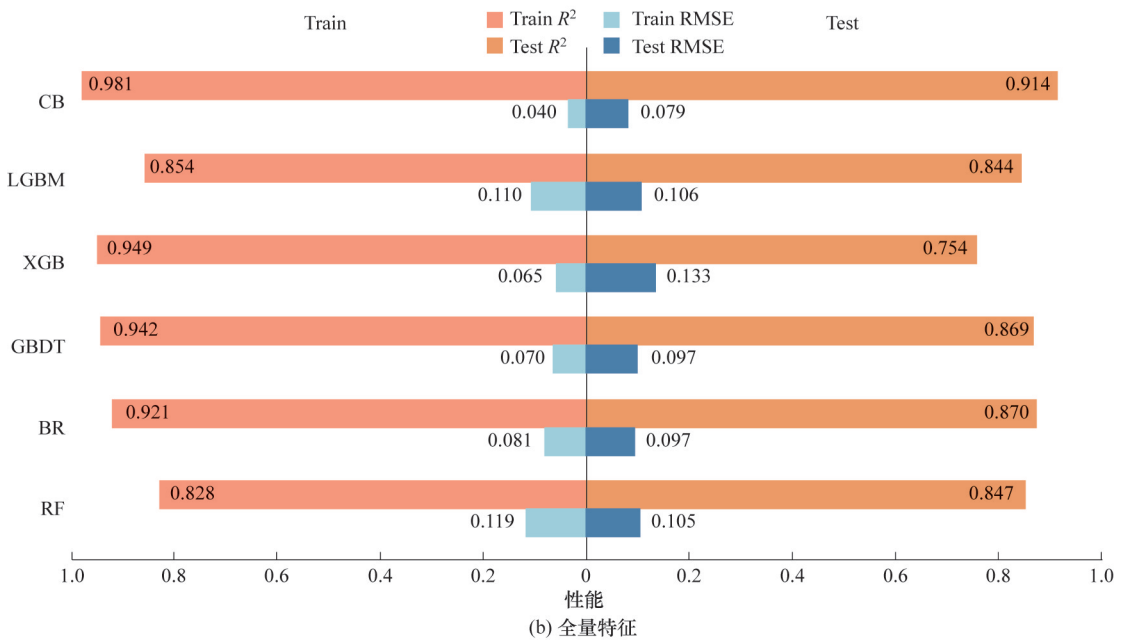
2.3.2 最终特征策略的确定 鉴于RFECV在降维策略中的优越表现，研究进一步将其与“全量特征策略”进行对比分析（图4）。结果显示，两种策略下模型性能差异并不显著。无论是RF、GBDT、XGB还是CB模型，基于全量特征与基于RFECV特征集的R^2和RMSE均高度接近，测试集预测精度均维持在R^2 = 0.87~0.91内。

这一现象揭示出现代集成学习算法所具备的内在鲁棒性，其内置的正则化机制与分裂准则能够在训练过程中自动削弱共线性特征的干扰，从而显著降低对外部降维预处理的依赖。换言之，当采用具有良好泛化能力与正则约束的模型体系时，过度的特征筛选反而可能损失潜在有用信息。

基于上述分析，本研究最终确定采用全量特征策略作为后续建模与可解释性分析的统一输入方案。这一选择的根本依据在于：保留全部38个原始特征，能够最大限度地维持“催化剂配方-制备工



(a) RFECV策略



(b) 全量特征

图4 各机器学习模型在使用两种不同特征集下的性能对比

Fig.4 Performance comparison of the models using two different feature sets

艺-反应条件 → 催化性能”之间的复杂非线性映射关系完整性，为后续可解释性分析及全局优化提供更丰富的变量空间。

同时，保留全特征输入还能更直观地揭示各参数在模型决策中的相对贡献，从而实现对催化剂设计的定量认知与策略指导，而非仅局限于经特征筛选后的局部变量解释。

2.4 模型超参数优化与最优代理模型确定

在2.3节的特征工程研究中，本研究确定“全量特征策略”为兼顾预测性能与信息完整性的最优方案。基于该策略，为进一步挖掘不同集成学习模型的潜在性能，并确定后续可解释性分析与全局寻优所依托的唯一最优代理模型，本节对六种集成学习模型进行了系统的超参数优化与性能评估。

2.4.1 *超参数优化方法与设定* 本研究采用1.3.2节所述的Optuna自动化调参框架，结合贝叶斯优化与自适应采样机制，对各模型在全量特征集上的关键超参数进行全局搜索与精细调优。优化目标函数设定为交叉验证的均方根误差（RMSE）最小化，以确保模型在泛化性能与稳定性间取得平衡。

优化过程中，Optuna框架基于树结构的Parzen估计器实现探索与利用的动态平衡，从而在有限计算成本下实现高效寻优。各模型的优化迭代历史与超参数重要性排序见图S5、图S6。最终确定的各模型最优超参数组合见表S3。

2.4.2 *模型性能对比与评估结果* 图5展示了6种集成学习模型在各自最优超参数配置下，于训练集和测试集上的拟合与预测表现。整体来看，各模型在全量特征集下均表现出较高的拟合精度与良好的泛化能力，充分验证了前述特征工程与标准化策略的合理性。6种模型的预测性能普遍优异，测试集R^2均超过0.84，说明模型能够有效捕捉催化体系中主要的非线性关联特征。然而，不同模型之间仍存在一定的性能差异。例如，XGB模型[图5(d)]在训练集上表现出极高的拟合度（R^2 = 0.9493），但在测试集上性能显著下降（R^2 = 0.7542），表明其存在一定程度的过拟合问题。相较之下，CB模型[图5(f)]在训练集与测试集上均展现出优异且稳定的表现，其训练集决定系数达到0.9807，而测试集的R^2高达0.9144，同时对应的RMSE最低，仅为0.0785。预测值与真实值在对角线附近高度重合，数据点分布紧密，显示出该模型在复杂非线性催化体系中的卓越拟合能力与强大泛化性能。

综上所述，基于全量特征集、并经过Optuna框架优化的CB模型，被确定为本研究的最优代理模型。该模型将作为后续可解释性分析及催化剂配方全局寻优的核心工具。

2.5 基于主动学习的数据集优化与模型性能强化

在上一节中，我们通过全面的模型对比与超参数调优，确定了基于全量特征的CB模型是性能最优的代理模型（Test R^2=0.9144, RMSE=0.0785）。然而，该模型是建立在n=63的完整、但可能包含噪声的小样本数据集之上的。正如1.4节所述，在小样本学习中，数据质量对模型泛化能力的掣肘可能比算法本身更为严重。

为解决这一问题，本研究并未止步于模型调优，而是进一步引入了主动学习（AL）策略。其核心目标是从“以模型为中心”转向“以数据为中心”，通过算法自动筛选出一个信息密度最高、噪声影响最小的“黄金”训练子集，从而进一步强化最优CB模型的性能上限。需要特别指出的是，本研究引入的主动学习策略并非用于构建传统的“预测-实验-反馈”闭环（即指导新一轮湿法实验），而是作为一种基于“以数据为中心”范式的训练集智能优化机制。针对文献数据固有的异构性与潜在噪声（如实验室间操作误差或非标准描述），该策略旨在从有限的原始数据池中，动态筛选出高信息量样本进行优先学习，从而在算法层面自动过滤低置信度的干扰噪声，实现训练数据的“提纯”。

图6(a)展示了主动学习策略的迭代优化过程。采用了不确定性采样、代表性采样和基于分歧的采样策略，从原始数据集中迭代选择最有价值的样本。如图所示，随着训练集规模的增加，模型的RMSE（以交叉验证评估）迅速下降并趋于收敛。这表明AL策略成功地识别并优先纳入了高信息量的数据点，同时规避了可能对模型产生误导的噪声或异常值。

图6(b)直观地对比了模型在AL优化前后的最终性能。优化前：模型的测试集R^2为0.9144，RMSE为0.0785；AL优化后：基于AL筛选出的最优训练集重新训练CB模型后，模型性能实现了显著的“极限突破”，测试集R^2提升至0.9347，RMSE则进一步降低至0.0686。这种在未引入外部新数据前提下的性能跃升，有力证实了主动学习策略在挖掘存量文献数据时的独特价值，即通过重构训练样本的分布，算法成功抑制了原始数据集中阻碍模型泛化的“异

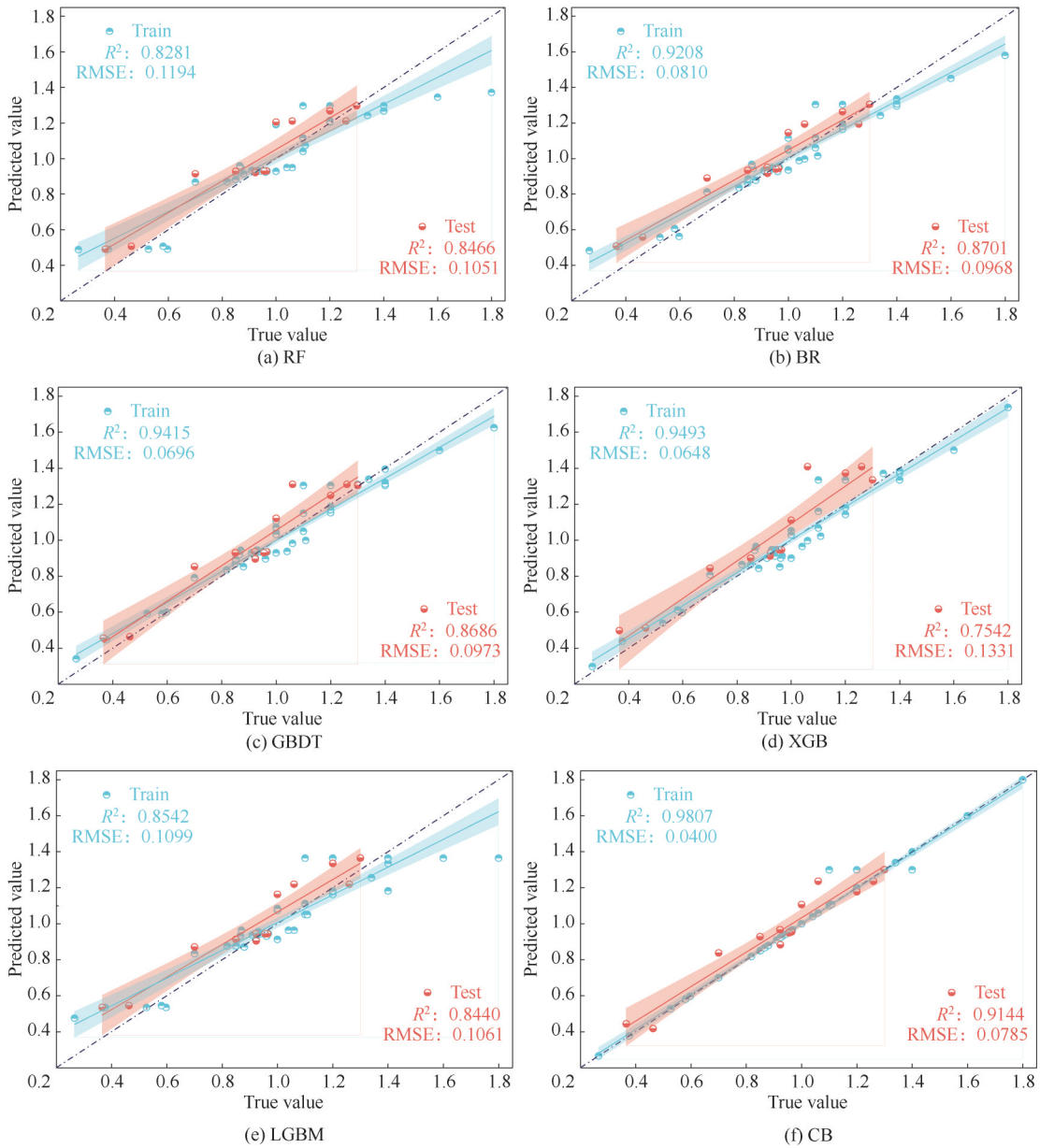


图5 各机器学习模型使用完整特征集的训练集/测试集拟合效果

Fig.5 Fitting results of the training and testing subsets using the full feature set

构噪声”干扰。这使得模型能够更聚焦于潜在的物理化学核心规律，从而在现有数据条件下挖掘出模型预测精度的内在极限潜力。

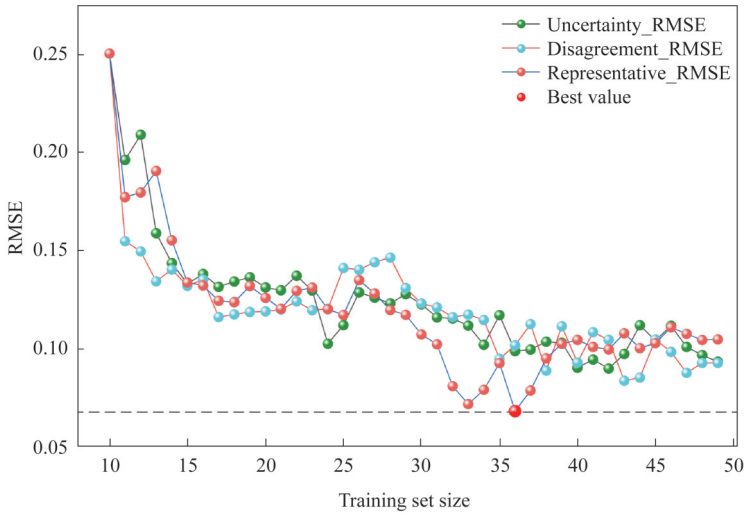
图6(c)展示了经过AL优化的最终CB模型的拟合结果。与优化前的图5(f)相比，测试集（红色）数据点更紧密地收束在$y=x$对角线周围，表现出更低的预测误差和更强的泛化能力。

这一结果强有力地证明，在小样本研究中，数据中心的优化策略（如主动学习）与模型中心的调优（如超参数优化）相结合，是实现最佳预测性能的关键。因此，本研究后续所有分析（包括可解释性分析和全局寻优）都将基于这个性能最强、鲁棒性最高的最终代理模型，即AL优化后的CB模型进行。

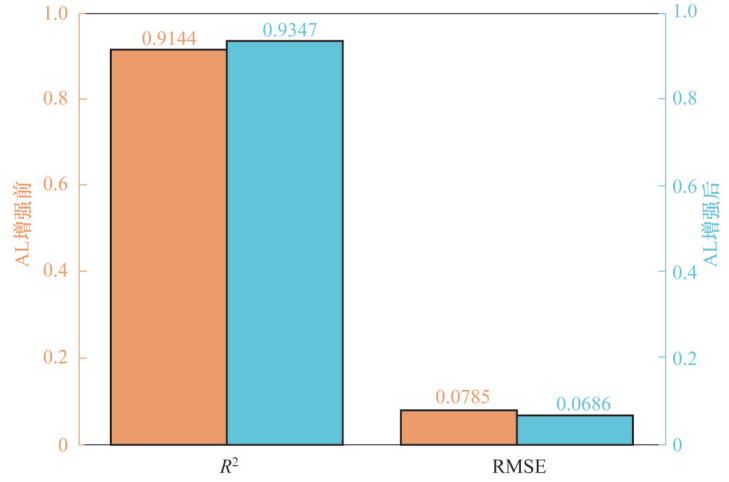
2.6 可解释性分析

为验证最优代理模型并非简单“记忆”训练样本，而是基于可泛化的物理化学规律做出预测，本研究在独立测试集上对模型输出开展了SHAP可解释性分析，并进一步用部分依赖图（PDP）检验关键特征的定量依赖关系。图7给出了模型在测试集上的SHAP结果[图7(a)全局蜂群图，展示了特征影响的分布情况；图7(b)局部贡献热力图]，图8则展示了3个最重要特征之间的两两PDP交互图谱。

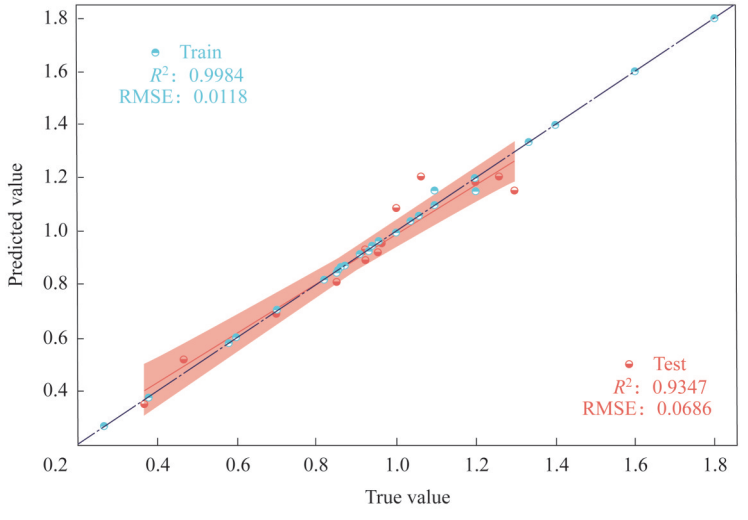
从全局视角看[图7(a)]，模型对测试集的预测高度依赖若干关键变量，其中反应温度T的影响最



(a) 主动学习过程的优化历史



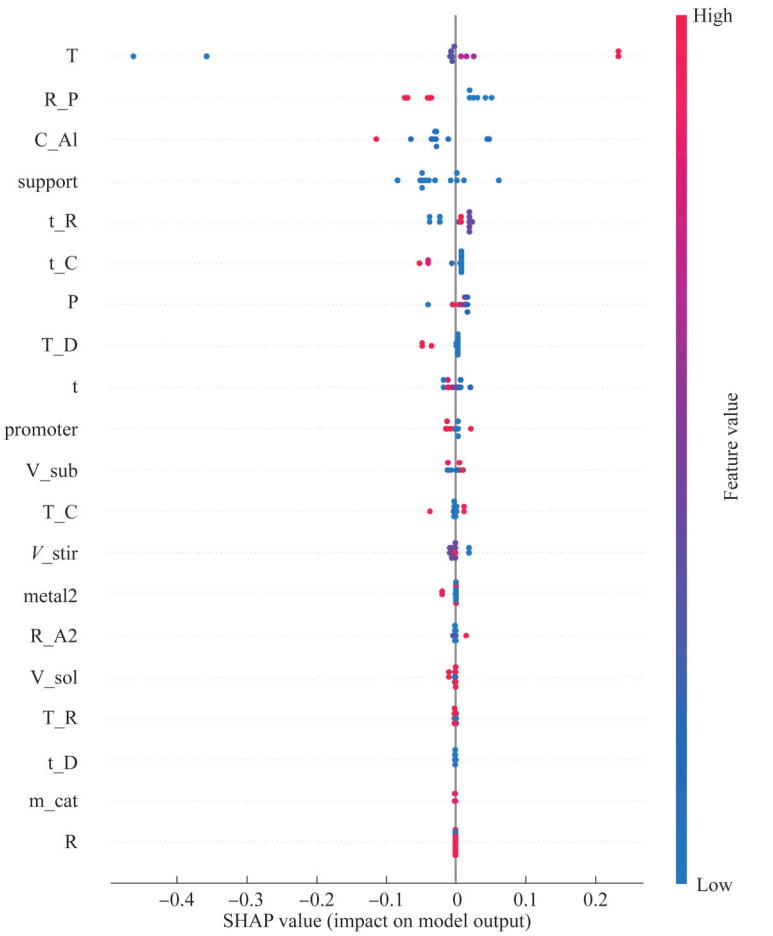
(b) 模型在主动学习优化前后的性能对比



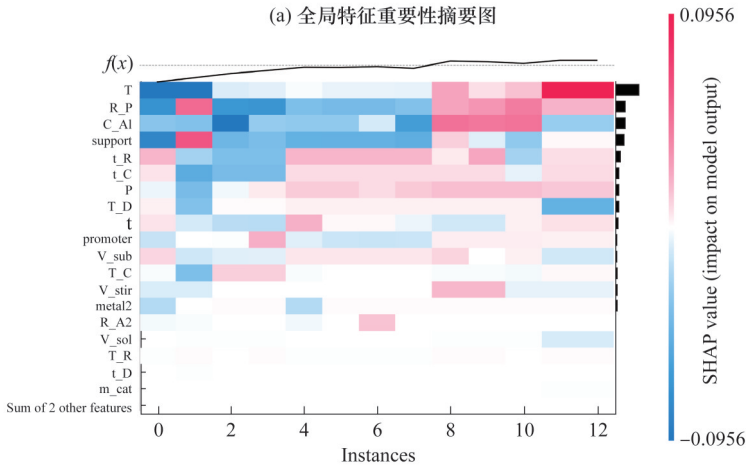
(c) 优化后CB模型的拟合结果

图6 基于主动学习策略对CB模型的优化与性能分析

Fig.6 Optimization and performance of the CB model using an active learning strategy



(a) 全局特征重要性摘要图



(b) 多个样本的局部特征贡献热力图

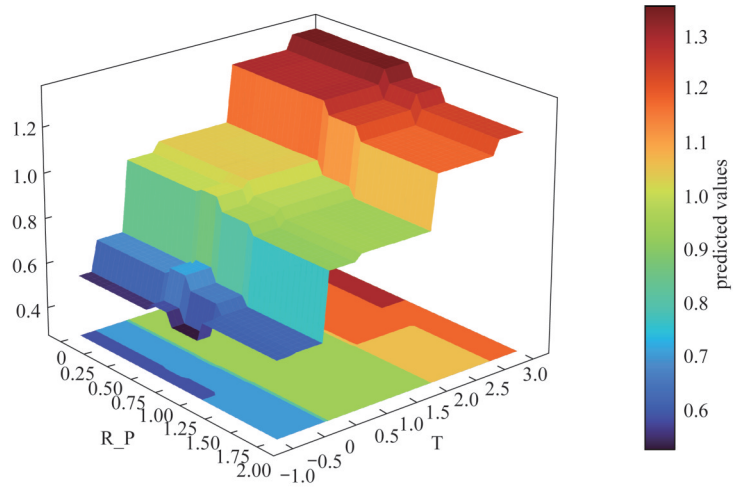
图7 模型的SHAP特征重要性分析

Fig.7 SHAP feature importance analysis for the model

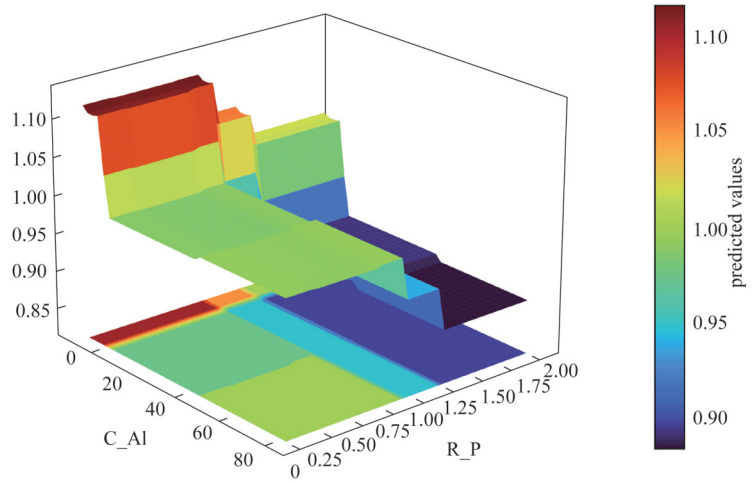
为显著，其SHAP值与标准化后的特征值呈明显正相关（高温对应较高的SHAP值），表明模型学习到了“升高温度利于提升目标性能”的基本动力学趋势。配体/金属比R_P与某组分的质量分数C_A1也列为高贡献特征，其中R_P与性能呈正相关，表明更高的配体配比在本体系中倾向于增强预测性能；而C_A1则表现出负相关趋势，高C_A1值普遍对应负向SHAP贡献，提示在本数据范围内较低的该组分含量反而有利于目标指标的提升。载体（support）、还原时间t_R等制备参数在全局上显示为中等重要性，意味着制备工艺对性能有辅助性影响但并非主导因素。总体而言，SHAP的全局排序

与催化化学的直觉一致，支持模型并非通过学习偶然的伪相关性来做出预测。

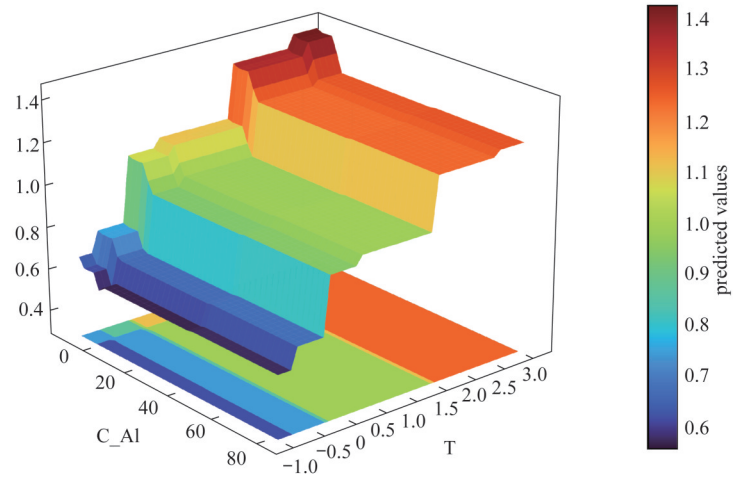
局部解释[图7(b)]进一步证实了上述全局结论的一致性。图7(b)所示的多个测试样本中，模型输



(a) 反应温度(T)与配体金属比(R_P)的交互作用



(b) Rh质量分数(C_A1)与配体金属比(R_P)的交互作用



(c) Rh质量分数(C_A1)与反应温度(T)的交互作用

图8 关键特征交互作用的PDP分析

Fig.8 Partial Dependence Plot analysis of key feature interactions

出(顶部$f(x)$黑线)的升降能够被具体特征的正负贡献直观解释。例如,某一低性能样本(Instance 1)对应的预测下降主要由R_P与C_A1的强负向贡献驱动;而一组高性能样本(Instances 8-12)则显示出R_P、C_A1与support的共同正向推动。将图7(a)、(b)综合解读可见,C_A1在全局呈现负相关,但在样本层面其贡献方向与特征取值一致:低C_A1常对应正向贡献,高C_A1则对应负向贡献,这一交叉验证进一步证明模型在未见数据上的决策逻辑具有内在一致性与可迁移性。

为了定量刻画关键特征的数值依赖及交互效应,进一步绘制了部分依赖图(图8),用于解析T、R_P与C_A1三者的两两交互关系。需要强调的是,图8中的坐标轴均基于训练前对所有特征施加的鲁棒标准化,因此显示为标准化后的数值区间(例如T的坐标可能呈现[-1.0, 3.0]),而非原始物理单位;下述对图谱的解读均基于该标准化尺度下的趋势性判断。图8(a)显示了T与R_P之间的显著协同效应:高性能区间集中在高温与高R_P的交叉区域(标准化尺度下T与R_P均处于较高值域),而任一参数处于低位时均会显著抑制预测性能。图8(b)表明R_P在多数条件下具有主导作用:当R_P处于高位时,模型预测的性能维持在较高平台且对C_A1的敏感性降低;但当R_P低于某一阈值时,C_A1的变化将无法挽回性能的下降。这一显著的交互效应揭示了Rh质量分数影响反应区域选择性的化学本质,即活性中心微观结构演变与配体调控能力的竞争。具体而言,在配体相对不足(低R_P)的条件下,单纯增加Rh质量分数不仅导致"配体/金属"有效摩尔比降低,更易诱导Rh物种发生团聚,形成大量未被配体充分修饰或空间位阻效应较弱的大颗粒金属位点。这些位点往往缺乏足够的立体选择性控制,倾向于生成热力学稳定的混合产物,从而拉低整体正异构比。相反,当配体处于过量(高R_P)区间时,强配位作用不仅维持了活性中心的饱和配位状态,还能有效抑制Rh物种的烧结与团聚,使其维持在高选择性的分散状态。此时,Rh负载量的数值波动便不再是制约选择性的关键因素。图8(c)则再次凸显了温度的决定性地位:高温条件下模型普遍预测较高性能,且C_A1的影响相对减弱;低温区间则整体对应较低性能,不利于实现目标指标提升。

综合SHAP与PDP的分析可得出明确的实验指导意义:反应温度与配体/金属比是决定体系性能的首要变量,且二者存在强烈的正向协同作用;Rh质量分数虽为重要参数,但在T与R_P均处于优良区间时,其影响相对次要。更重要的是,全局与局部解释的一致性表明,代理模型在独立测试集上的预测基于可泛化的物理化学关系,而非对训练集的机械记忆。由此,本节的可解释性分析不仅增强了对代理模型可靠性的信任,也为基于该模型开展的全局寻优提供了明确而有物理意义的调控方向。

2.7 催化剂配方、制备条件及反应参数的全局寻优

本节展示了本研究数据驱动框架的最终价值与应用成果。为实现1.6节中提出的"选择性-转化率"双目标优化,本研究构建了2个高性能且相互独立的机器学习代理模型。其中,正异构比($y_{n/i}$)代理模型作为核心模型,基于2.4~2.6节所开发并经主动学习强化的CB结构,在测试集上实现了优异的预测性能(Test R^2=0.9347)。与此同时,为准确预测反应活性,本研究采用与$y_{n/i}$模型完全一致的建模框架(包括全量特征集与Optuna超参数优化),并行构建了一个独立的y_{cov}代理模型。模型筛选结果表明,经过优化的XGB模型在转化率预测任务中表现最佳(Test R^2=0.833),因此,被确定为该任务的最终代理模型。完整的模型筛选与性能评估结果见图S7。基于已确定的最优代理模型(CatBoost-AL用于$y_{n/i}$,XGBoost用于y_{cov}),本研究结合差分进化(DE)算法,实施了两阶段的全局寻优,以探索超越文献数据记录的潜在最优催化体系。

第一阶段聚焦于产物选择性最大化,即通过优化$y_{n/i}$实现高选择性催化性能。依据1.6节式(6)定义,DE算法以CB代理模型为核心驱动函数,在38维高维参数空间中开展全局搜索。图9(a)展示了这一迭代优化过程:密集的蓝色散点表示算法在超过10万次迭代中探索的参数组合,而红色实线则描绘了每轮迭代中获得的当前最优值。结果表明,算法在初始阶段便快速突破原始数据集性能上限(数据集中$y_{n/i}$平均值为1.0 ~ 1.2),并最终收敛至稳定最优区域,对应的理论最优预测值约为$y_{n/i}$=1.813。这一结果证明,基于机器学习代理的差分进化优化能够有效捕捉复杂催化体系中的非线性结构-性能关联,显著拓展了实验数据的性能边界。在第二阶段优化中,研究进一步考察了反应活性(y_{cov}),以实现选择性与活性的协同提升。首先从阶段一所得结果中筛选出$y_{n/i}$排名前100的参数组合,构建"高选择性候选集";随后利用XGBoost代理模型对这100组

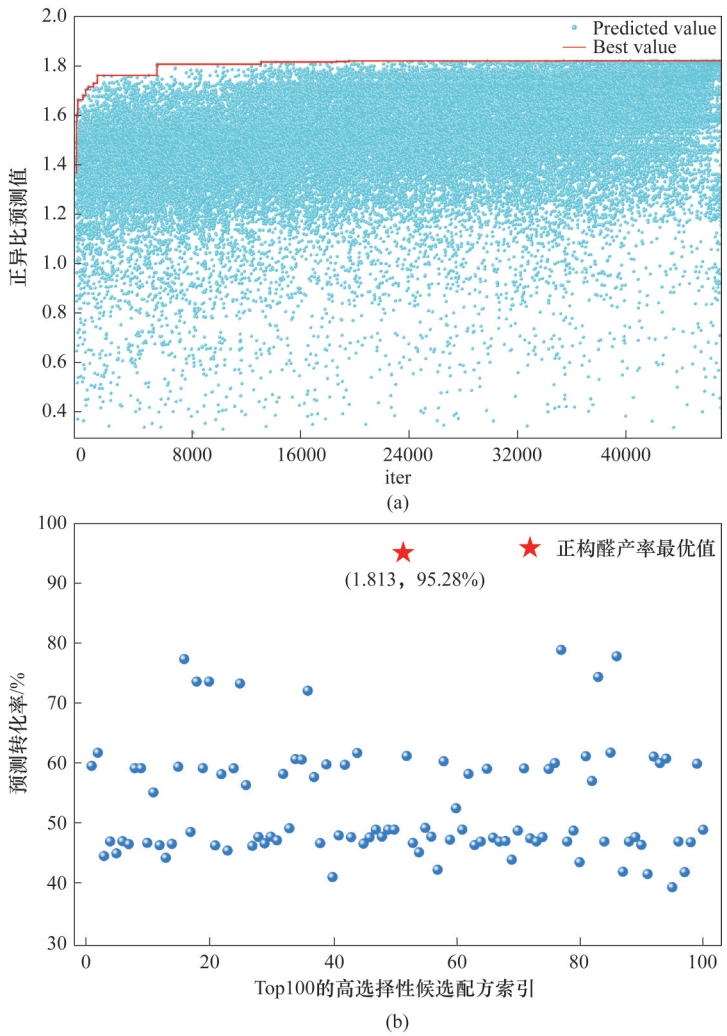


图9 基于机器学习与差分进化算法的两阶段配方优化：(a) 结合差分进化算法搜索最优正异比（选择性）的迭代过程；(b) 第一阶段筛选出的Top100高选择性配方的转化率预测结果。横坐标代表按预测选择性排序的候选配方索引编号，并确定多目标最优解（正异比 1.813，转化率 95.28%）

Fig.9 Two-stage recipe optimization based on machine learning and differential evolution: (a) The iterative process of searching for the optimal n/i ratio (selectivity) using the DE algorithm combined with the ML model; (b) Conversion rate prediction for the Top 100 high-selectivity recipes screened from Stage 1. The *x*-axis represents the index of the candidate recipes sorted by their predicted selectivity, and the determination of the multi-objective optimal solution (n/i ratio 1.813, Conversion 95.28%)

样本的转化率进行预测，结果如图9(b)所示。可以看到，即便在高选择性候选集中，转化率仍呈现出显著的分布差异，范围由约40%~95%以上。这一结果揭示了催化体系中选择性与活性之间的非线性竞争关系，也进一步印证了多目标优化策略的必要性。最终，图9(b)中由红色星标标示的点被识别为多目标最优解（表S6），该参数组合在保持第一阶段识别出的最高选择性（$y_{n/i}$=1.813）的同时，实现了候选集中最高的反应转化率（y_{cov}=95.28%）。

综上所述，本研究通过差分进化算法驱动的两阶段全局寻优，成功实现了对催化剂配方、制备条件与反应参数的系统性优化，获得了一组具有实验可验证性的理论最优解。该最优处方在选择性与活性两方面均显著优于现有文献报道水平，展现了机器学习在揭示复杂催化体系内在规律与加速高性能催化剂开发方面的强大潜力。这一优化结果不仅为后续实验设计提供了高价值的方向性指导，也为实现数据驱动的“从模型到实验”的闭环催化开发模式奠定了坚实基础。需要指出的是，受限于当前实验条件，本研究尚未对该预测出的最优配方进行合成确证。然而，该结果的可靠性主要由两方面支撑，一是模型在完全独立的测试集文献数据上表现出极高的预测精度（R^2=0.9347），证明了其对未见实验体系的泛化能力；二是前述SHAP与PDP分

析证实模型捕捉的特征规律与催化化学机理高度一致，排除了“黑箱”模型学习虚假相关性的风险。因此，该理论最优解具有较高的可信度，待具备条件后将开展针对性的实验验证。

3 结 论

利用机器学习模型开展非均相催化剂的精准设计受到小样本、高维度与噪声数据的制约。本研究以苯乙烯氢甲酰化作为探针体系，构建了一套数据与智能算法驱动的闭环优化框架，以验证机器学习与全局寻优在催化剂理性设计中的适用性。基于75条文献数据，该框架集成特征工程、集成学习、主动学习与差分进化算法，有效提升了建模效率与预测准确性。在现代集成模型的鲁棒性加持下，全量特征策略获得最高预测性能（Test R^2 = 0.9144），优于特征选择与降维方法。进一步引入主动学习筛选高价值训练样本，使模型性能提升至Test R^2 = 0.9347，体现了小样本场景下数据质量优先的关键作用。通过SHAP与PDP分析，反应温度、助剂/金属比（R_P）与Rh质量分数被识别为主导选择性的关键因素，其作用趋势与氢甲酰化化学规律高度一致。在此基础上，构建选择性与转化率的双代理模型（其中XGBoost模型转化率Test R^2 = 0.833），并结合两阶段差分进化实现配方寻优，最终获得性能优于文献报道的最优配方：预测正异构比1.813，转化率95.28%。该结果证明了数据驱动方法在催化剂探索中的有效性。尽管当前模型界定在浸渍法范畴，但本研究提出的闭环优化框架仍具有显著的推广潜力。针对原子层沉积或磁控溅射等其他制备工艺，只需建立与之适配的特定工艺特征描述符（如原子层沉积的脉冲周期与窗口温度等），并以此重构特征空间，即可将该“数据-模型”驱动范式迁移至更广泛的非均相催化剂开发中。未来，该闭环优化框架有望被推广应用于更广泛和复杂的非均相催化体系设计中。

增强出版说明：本文支撑材料（图S1~图S7，表S1~表S6）仅在网络版发布，纸质印刷版不予收录。读者可通过《化工学报》官网及中国知网网络版查阅。

参考文献

[1] Bohnen H W, Cornils B. Hydroformylation of Alkenes: An Industrial View of the Status and Importance[M]. Advances in Catalysis. Amsterdam: Elsevier, 2002: 1-64.

[2] Cuny G D, Buchwald S L. Practical, high-yield, regioselective, rhodium-catalyzed hydroformylation of functionalized. α-olefins [J]. Journal of the American Chemical Society, 1993, **115**(5): 2066-2068.

[3] Marras F, Wang J, Coppens M O, et al. Ordered mesoporous materials as solid supports for rhodium-diphosphine catalysts with remarkable hydroformylation activity[J]. Chemical Communications, 2010, **46**(35): 6587-6589.

[4] Ahneman D T, Estrada J G, Lin S S, et al. Predicting reaction performance in C-N cross-coupling using machine learning[J]. Science, 2018, **360**(6385): 186-190.

[5] Ma X F, Li Z, Achenie L E K, et al. Machine-learning-augmented chemisorption model for CO_2 electroreduction catalyst screening [J]. The Journal of Physical Chemistry Letters, 2015, **6**(18): 3528-3533.

[6] Shields B J, Stevens J, Li J, et al. Bayesian reaction optimization as a tool for chemical synthesis[J]. Nature, 2021, **590**(7844): 89-96.

[7] Kim M, Ha M Y, Jung W B, et al. Searching for an optimal multi-metallic alloy catalyst by active learning combined with experiments[J]. Advanced Materials, 2022, **34**(19): e2108900.

[8] Lu H Y, Diaz D J, Czarnecki N J, et al. Machine learning-aided engineering of hydrolases for PET depolymerization[J]. Nature, 2022, **604**(7907): 662-667.

[9] Wu Z, Jennifer Kan S B, Lewis R D, et al. Machine learning-assisted directed protein evolution with combinatorial libraries[J]. Proceedings of the National Academy of Sciences of the United States of America, 2019, **116**(18): 8852-8858.

[10] Li F R, Yuan L, Lu H Z, et al. Deep learning-based k_{cat} prediction enables improved enzyme-constrained model reconstruction[J]. Nature Catalysis, 2022, **5**(8): 662-672.

[11] Ulissi Z W, Medford A J, Bligaard T, et al. To address surface reaction network complexity using scaling relations machine learning and DFT calculations[J]. Nature Communications, 2017, **8**: 14621.

[12] Chanussot L, Das A, Goyal S, et al. Open catalyst 2020 (OC20) dataset and community challenges[J]. ACS Catalysis, 2021, **11** (10): 6059-6072.

[13] Yano J, Gaffney K J, Gregoire J, et al. The case for data science in experimental chemistry: examples and recommendations[J]. Nature Reviews Chemistry, 2022, **6**(5): 357-370.

[14] Kearnes S M, Maser M R, Wleklinski M, et al. The open reaction database[J]. Journal of the American Chemical Society, 2021, **143** (45): 18820-18826.

[15] Akiba T, Sano S, Yanase T, et al. Optuna: a next-generation hyperparameter optimization framework[C]//Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. Anchorage AK USA. ACM, 2019: 2623-2631.

[16] Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in Python[PP/OL]. V4. arXiv(2018-06-05) [2026-01-30].https://doi.org/10.48550/arXiv1201.0490.

[17] Domingos P. A few useful things to know about machine learning [J]. Communications of the ACM, 2012, **55**(10): 78-87.

[18] Jia W K, Sun M L, Lian J, et al. Feature dimensionality reduction: a review[J]. Complex & Intelligent Systems, 2022, **8**(3): 2663-

2693.

[19] Bai X J, Tao T, Gao L Y, et al. Wind turbine blade icing diagnosis using RFECV-TSVM pseudo-sample processing[J]. Renewable Energy, 2023, **211**: 412-419.

[20] Breiman L. Random forests[J]. Machine Learning, 2001, **45**(1): 5-32.

[21] Breiman L. Bagging predictors[J]. Machine Learning, 1996, **24**(2): 123-140.

[22] Friedman J H. Greedy function approximation: a gradient boosting machine[J]. The Annals of Statistics, 2001, **29**(5): 1189-1232.

[23] Ke G L, Meng Q, Finley T, et al. LightGBM: a highly efficient gradient boosting decision tree[C]//Neural Information Processing Systems., 2017

[24] Chen T Q, Guestrin C. XGBoost: a scalable tree boosting system [C]//Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco California USA. ACM, 2016: 785-794.

[25] Ostroumova L, Gusev G, Vorobev A, et al. CatBoost: unbiased boosting with categorical features[C]//Neural Information Processing Systems., 2017

[26] Zhu X Y, Shen X Y, Chen K L, et al. Research on the prediction and influencing factors of heavy duty truck fuel consumption based on LightGBM[J]. Energy, 2024, **296**: 131221.

[27] Namdeo S, Srivastava V C, Mohanty P. Machine learning implemented exploration of the adsorption mechanism of carbon dioxide onto porous carbons[J]. Journal of Colloid and Interface Science, 2023, **647**: 174-187.

[28] Aryandoust A, Patt A, Pfenninger S. Enhanced spatio-temporal electric load forecasts using less data with active deep learning[J]. Nature Machine Intelligence, 2022, **4**(11): 977-991.

[29] Jayarathna R, Onsree T, Drummond S, et al. Experimental discovery of novel ammonia synthesis catalysts *via* active learning [J]. Journal of Materials Chemistry A, 2024, **12**(5): 3046-3060.

[30] Xiao Y S, Chang Z, Liu B. An efficient active learning method for multi-task learning[J]. Knowledge-Based Systems, 2020, **190**: 105137.

[31] Li Z C, Zhang S, Li H Q, et al. On-line transfer learning for multi-fidelity data fusion with ensemble of deep neural networks[J]. Advanced Engineering Informatics, 2022, **53**: 101689.

[32] Lundberg S M, Lee S I. A unified approach to interpreting model predictions[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. December 4 - 9, 2017, Long Beach, California, USA. ACM, 2017: 4768-4777.

[33] Lundberg S M, Erion G G, Lee S I. Consistent individualized feature attribution for tree ensembles[PP/OL]. V3. arXiv(2019-03-07)[2026-01-30].https://doi.org/10.48550/arXiv.1802.03888.

[34] Storn R, Price K. Differential evolution-a simple and efficient heuristic for global optimization over continuous spaces[J]. Journal of Global Optimization, 1997, **11**(4): 341-359.