



基于多算法机器学习的铁单原子催化剂设计及其对新污染物降解效能的预测

李明¹, 罗鑫¹, 胡春¹, 吕来^{1,2,✉}

1. 广州大学大湾区环境研究院, 珠江三角洲水质安全与保护教育部重点实验室, 广州 510006; 2. 广州大学环境科学与工程学院, 广州 510006

摘要 为高效评估铁单原子掺杂沸石咪唑酯骨架材料-8 (iron-single-atom Zeolitic ZIF8, IZ) 对新污染物的降解性能, 本研究采用了机器学习 (ML) 方法建立了 IZ 对新兴污染物的降解性能预测模型。首先, 对含有 46 个样本的 IZ 降解性能数据集进行了相关性分析, 以确定 ML 模型的 8 个输入特征。随后, 使用此数据集训练、优化和评估 9 个 ML 模型, 包括自适应提升 (AdaBoost)、决策树 (DT)、 k 近邻 (KNN)、套索回归模型 (Lasso)、多元线性回归模型 (MLR)、随机森林 (RF)、岭回归 (Ridge)、极端梯度提升 (XGBoost) 和梯度提升决策树 (GBDT)。然后, 利用沙普利加性解释 (SHAP) 方法计算了预测精度最高的黑盒模型的特征重要性, 并探讨了特征与降解性能之间的关系。结果表明, RF 模型的决定系数 (R^2) 超过了 0.75, 其预测值的残差主要分布在 -0.005~0.005。RF 模型为铁的添加量 and 水的添加量赋予了最高的特征重要度, 因为这些因素对降解性能有显著影响。在一定范围内增加铁的添加量、水的添加量、甲醇的添加量、锌的添加量、煅烧温度、pH 和过氧化物的投加量均可提升 IZ 的降解性能。基于 RF 模型, 本研究预测了 IZ 最佳的设计工艺参数组合, 为利用 ML 预测和优化 IZ 对新兴污染物的降解性能提供了见解。

关键词 机器学习; 铁单原子; 沸石咪唑酯骨架材料-8; 新污染物; 材料预测; 降解性能

水体自净化依赖于水体内源性物质 (如溶解氧、天然有机物、盐分离子等) 之间复杂相互作用, 主要通过氧化还原反应实现污染物的降解。然而, 这一自然过程往往缓慢且效率低下, 污染物的完全降解可能需要长达数十年的时间^[1-2]。随着工业化进程的加速和人们生活方式的变化, 水体中涌现了大量新型污染物, 如药物及个人护理产品、内分泌干扰物、持久性有机污染物等。这些新污染物因其难降解性、生物累积性和长距离传输能力, 对生态环境和人类健康构成了严峻威胁^[3]。

金属有机框架材料, 特别是沸石咪唑酯骨架材料 (ZIFs), 因其独特的结构特征和物理化学性质, 在催化领域展现出巨大潜力^[4]。其中, ZIF8 因具有优异的化学稳定性和热稳定性, 以及可调节的孔径和表面性质, 成为催化剂设计和性能优化的理想选择^[5-7]。近年来, 铁单原子催化剂因其高活性、高选择性和良好的稳定性, 在多种催化反应中表现出色^[8-11]。将铁单原子引入 ZIF8 中, 有望结合两者的优势, 开发出具有高效降解新污染物能力的新型催化剂。然而, 催化剂的设计与性能优化是一个复杂且耗时的过程, 涉及多个关键因素, 如催化剂的结构特征、合成条件、反应介质和反应条件等。传统的实验方法往往依赖于大量的试错实验, 不仅成本高昂, 而且难以全面探索催化剂性能与结构之间的复杂关系^[12]。因此, 开发一种高效、系统的催化剂性能预测方法, 以指导催化剂的设计与优化, 对于加速新型催化剂的研发进程具有重要意义。机器学习作为一种强大的数据分析和预测工具, 在材料科学和化学工程等领域得到了广泛应用^[13-16]。通过收集和分析大量实验数据, 机器学习算法能够揭示催化剂性能与结构之间的内在联系, 建立准确的预测模型。特别是在催化剂性能预测方面, 机器学习技术已展现出显著优势, 能够显著提高预测精度和效率, 缩短催化剂的研发周期^[12,14]。

收稿日期: 2025-02-13 录用日期: 2025-04-03

基金项目: 广东省重点领域研发计划项目纳米科技专项 (2023B0101200004); 国家自然科学基金资助项目 (面上 52470069)

第一作者: 李明 (2000—), 男, 硕士研究生, 研究方向为多相催化废水处理技术原理与应用, 2112204018@e.gzhu.edu.cn

✉通信作者: 吕来 (1990—), 男, 博士, 教授, 研究方向为废水深度处理新技术新原理及其工程应用, lyulai@gzhu.edu.cn

因此,本研究利用机器学习技术,基于已有文献数据构建数据集^[8,17-27],对 IZ 降解效率进行理论预测。通过深入探讨合成过程和降解过程中不同因素对催化剂性能的影响机制,为催化剂的设计与优化提供理论指导,这不仅有助于深入理解 IZ 催化机理,还为开发高效降解新污染物的新型催化剂提供了新思路和方法。

1 材料与方法

1.1 数据集

为了确保数据集的一致性,文献来源要求对于 IZ 的合成方法主要分为 3 个步骤:1)前驱体 ZIF8 自组装;2)Fe 掺杂至前驱体;3)高温煅烧。其次,合成后的 IZ 主要用于降解含羟基的芳香类新污染物。基于上述标准,从相关文献中收集了 46 个关于 IZ 的数据样本。数据集共 12 个特征作为模型的输入,分别为锌的添加量、铁的添加量、水的添加量、甲醇添加量、搅拌温度、搅拌时间、烘干温度、烘干时间、煅烧温度、催化剂投加量、过氧化物投加量和 pH。动力学降解常数(k)则作为模型的输出。对所有特征进行标准化变换以消除量纲、数量级的影响,并提升 ML 模型训练速度。如式 (1) 所示。

$$x'_i = \frac{x_i - u_i}{\sigma_i} \quad (1)$$

式中: x_i 、 x'_i 分别为第 i 个原始、标准化变换后的特征; u_i 、 σ_i 分别为原始特征的平均值、标准差。

1.2 特征筛选

递归特征消除方法(recursive feature elimination, RFE)能够解决高维数据分析中的特征筛选问题。其原理是于初始特征全集构建基准预测模型;实施递归式特征排除策略,每轮迭代移除当前特征重要性排序末位的特征;重复执行模型重构与特征排除的递归操作直至最后一个特征,在此过程中,特征排除顺序构成逆向重要性排序,最终得到最优特征子集。RFE 保证模型解释性的同时,降低了过拟合风险,为后续模型拟合提供可靠的数据基础。

评估特征间及特征-目标变量的线性关联强度是构建预测模型的重要步骤。为此,采用皮尔森相关系数(pearson correlation coefficient, PCC)系统评估特征间及特征-目标变量的线性关联强度,以明确其线性关系的强弱。PCC 的含义如下:取值范围为-1~1,越接近-1 或 1,代表两个数据之间有负或正线性相关性,越接近于 0 则表明两个数据之间线性相关性越弱。

1.3 模型选用、训练与评估

本研究选用 AdaBoost^[28]、DT^[29]、KNN^[30]、Lasso^[31]、MLR^[32]、RF^[33]、Ridge^[34]、XGBoost^[35]、GBDT^[36]这 9 个 ML 模型进行训练,以预测 IZ 的性能。具体而言,AdaBoost、XGBoost、RF 和 GBDT 作为集成学习方法的代表,通过结合多个基模型(即弱学习器),有效降低了预测的方差或偏差,从而显著提升了整体的预测性能。其中,RF 通过并行训练多棵决策树,并利用投票或平均机制,显著增强了模型的泛化能力与抗噪性;而 AdaBoost、XGBoost 及 GBDT 则采用串行训练方式,AdaBoost 通过迭代调整样本权重,优化弱分类器的组合策略,以提升分类精度;XGBoost 则优化了梯度提升框架,能高效处理结构化数据,并支持正则化处理以增强模型稳定性。此外,MLR、Lasso 回归及 Ridge 回归作为线性模型的代表,基于线性假设构建预测模型,并通过正则化方法有效约束系数复杂度,从而有效防止了过拟合现象的发生。DT 模型则基于特征分割规则,递归地划分数据集以进行预测,展现了树模型在处理复杂数据关系方面的独特优势。而 KNN 模型则作为一种基于实例的学习方法,根据样本邻域的多数类别或均值进行分类或回归预测,为处理非线性关系提供了有力工具。

在数据集划分方面,严格遵循 8:2 的比例,将 80% (37 个)的数据用于机器学习模型的交叉验证,以科学训练模型并确定各模型的最优超参数组合;而剩余的 20% (9 个)数据则作为测试集,用于客观评估各机器学习模型的泛化性能。

采用决定系数(coefficient of determination, R^2)、均方误差(mean squared error, MSE)、平均绝对误差(mean absolute error, MAE)作为模型训练与测试时的预测精确度评估指标,其计算方法如式 (2)~式 (4) 所示。

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_{t,i} - y_{p,i})^2}{\sum_{i=1}^n (\bar{y} - y_{p,i})^2} \quad (2)$$

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (y_{t,i} - y_{p,i})^2 \quad (3)$$

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |y_{t,i} - y_{p,i}| \quad (4)$$

式中： n 为训练集或测试集的数据样本数； $y_{t,i}$ 、 $y_{p,i}$ 分别为实验值、预测值； $\bar{y}$ 为数据集中实验值的平均值。 R^2 的常见取值范围为 0~1，模型预测精确度越高， R^2 越接近于 1；MAE 和 MAPE 的取值范围为 0~ $+\infty$ ，模型预测精确度越高，其取值越接近于 0。此外，还计算了各预测值的残差（Res）和相对百分比误差（RPE），计算公式如式（5）和式（6）所示。

$$\text{Res} = y_{p,i} - y_{t,i} \quad (5)$$

$$\text{PRE} = \frac{y_{p,i} - y_{t,i}}{y_{t,i}} \times 100\% \quad (6)$$

1.4 模型解释

本研究选取了 DT、MLR、Lasso、Ridge 以及 KNN 这 5 种典型的白盒模型进行分析，白盒模型结构和决策过程完全透明，可通过数学公式、规则或可视化路径明确解释预测逻辑，具有很高可解释性。但是白盒模型在提供高可解释性的同时，也面临着模型复杂度相对较低的问题。这可能导致模型在处理复杂数据关系时难以充分捕捉数据中的潜在信息，进而引发欠拟合的风险。白盒模型的预测函数可以表示为线性形式，如式（7）和式（8）所示。

$$y = \mathbf{A}\mathbf{X} + b \quad (7)$$

$$\mathbf{A} = [a_1, a_2, a_3, \dots, a_i] \quad (8)$$

式中： $\mathbf{X}$ 、 y 分别为特征矩阵和预测值向量； $\mathbf{A}$ 为特征的权重向量； a_i 为第 i 特征的权重系数； a_i 的绝对值越大，代表该特征的重要度越高； b 为白盒模型的截距。

黑盒模型的拟合能力往往强于白盒模型，本研究涉及的 RF、AdaBoost、XGBoost 和 GBDT，以其强大的拟合能力而著称。这些模型在处理复杂、非线性数据时表现出色，能够捕捉到数据中的深层次特征。然而，由于其内部机制复杂或高度非线性，预测逻辑难以直接解释，通常依赖于输入-输出映射关系。为了克服这一局限，引入了外部解释方法以协助理解模型并挖掘其内在价值。

在模型训练完成后，采用了 SHAP 这一基于博弈论的解释方法，对具有最高预测精确度的黑盒机器学习模型进行了解释。SHAP 方法的核心思想在于通过计算特征的 Shapley 值来量化特征对模型预测结果的影响程度，从而明确哪些特征在预测 IZ 对污染物降解性能时起到了关键作用。这一方法不仅提供了对模型预测逻辑的深入理解，而且为优化模型性能、提升预测精度提供了有力支持。Shapley 值的计算公式如式（9）所示。

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F|-|S|-1)!}{|F|!} [f_{S \cup \{i\}}(x_{S \cup \{i\}}) - f_S(x_S)] \quad (9)$$

式中： ϕ_i 为特征 i 的 Shapley 值， ϕ_i 越大，代表该特征对预测值影响越大（该特征越重要）； F 为包含所有特征的特征集； S 为 F 的子集； $f_{S \cup \{i\}}(x_{S \cup \{i\}})$ 为包含 S 和第 i 个特征的特征子集的 ML 模型预测值； $f_S(x_S)$ 为包含 S 但不包含第 i 个特征的特征子集的 ML 模型预测值。此外，还可使用 SHAP 获取特征对预测目标影响的趋势信息。

1.5 建模工具

数据处理与建模流程主要由 Python 编程语言完成，数据处理由 Pandas 库和 numby 库完成，绘图由 matplotlib 库实现，ML 模型基于 Scikit-learn 库实现。

2 结果与讨论

2.1 特征选择

采用了 RFE 对全部特征通过逐步移除对模型贡献最小的特征来实现特征筛选, 最终选出了 8 个特征作为模型的输入, 分别为锌的添加量、铁的添加量、水的添加量、甲醇添加量、煅烧温度、催化剂投加量、过氧化物投加量和 pH。然后, 采用 PCC 系统评估特征间及特征-目标变量的线性关联强度, 以明确其线性关系的强弱。如图 1 所示, 通过量化分析发现特征之间 PCC 值域集中于-0.50~0.60, 揭示了特征空间的低共线性特征。但是烘干温度和烘干时间的相关性是-0.9, 主要是因为在大部分实际实验中, 科研人员烘干的主要目的是去除材料中的水分, 对于水分的烘干条件总是在 60 ℃ 烘干 12 h 或者在 80 ℃ 烘干 6 h, 所以呈现极高的负相关性。而锌的添加量与烘干温度和烘干时间的相关性分别为 0.96 和-0.95, 在 RFE 中可以得知烘干温度和烘干时间的贡献度不高, 所以剔除烘干温度和烘干时间两个特征, 保留锌的添加量这一特征。水的添加量和甲醇的添加量的相关性为-1, 是因为在 ZIF8 自组装过程中溶液环境主要是甲醇和水, 当主要溶剂是甲醇时, ZIF8 合成缓慢, 形成的主要形状为十二面体^[37]; 当主要溶剂是水时, ZIF8 合成迅速, 形成的主要形状为方钠石形和菱形^[38]。可以通过调整甲醇与水的比例从而调整 ZIF8 前驱体的尺寸和形状。所以特征主要选择为锌的添加量、铁的添加量、水的添加量、甲醇添加量、煅烧温度、催化剂投加量、过氧化物投加量和 pH。特征-目标变量之间的 PCC 为集中在-0.50~0.50, 说明目标变量与预测因子间存在显著非线性关联机制, IZ 对新污染物的降解性能是各个特征共同影响的结果。这一发现为复杂系统建模中的模型选择提供了关键判别依据, 后续应主要考虑非线性 ML 模型。

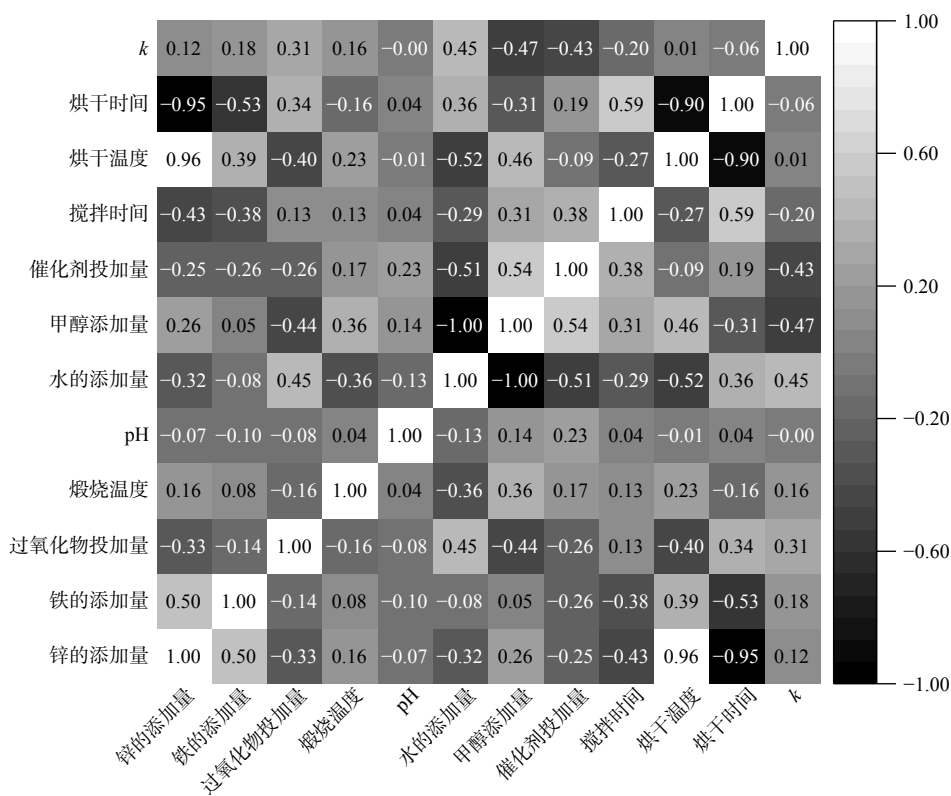


图 1 数据集的皮尔森相关系数矩阵

Fig. 1 Pearson correlation coefficient matrix of the dataset

2.2 模型预测精确度

首先, 根据数据集的特点调整各个模型的超参数以确保最大程度拟合数据集, 超参数的选择如表 1 和表 2 所示。通过计算各模型在训练集上的预测精确度, 系统地检查了模型是否存在欠拟合问题, 如图 2 所

表 1 线性模型的超参数选择

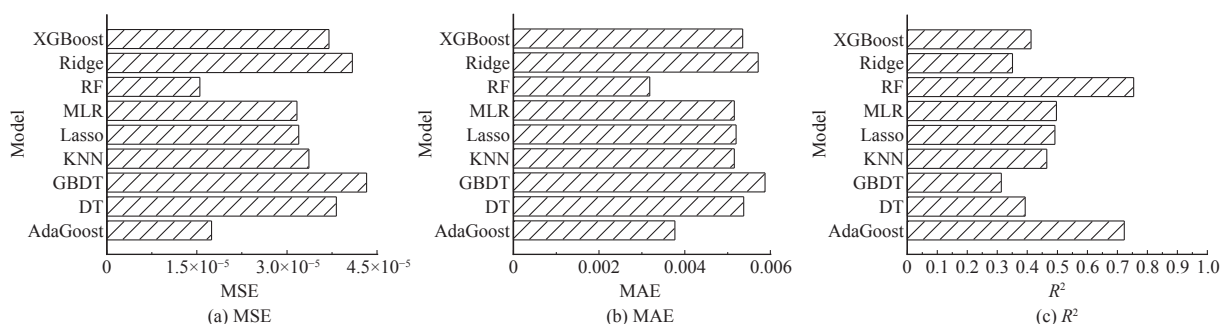
Table 1 Hyperparameter selection of linear models

线性模型	DT	Lasso	Ridge	KNN
超参数	max_depth': 3 'max_features': 'sqrt' 'min_samples_leaf': 10 'min_samples_split': 2 'random_state': 42	'alpha': 0.001	'alpha'=80	knn_n_neighbors': 10 'knn_p': 1 'knn_weights': 'uniform'

表 2 非线性模型的超参数选择

Table 2 Hyperparameter selection of non-linear models

非线性模型	RF	AdaBoost	GBDT	XGBoost
超参数	n_estimators': 7 'max_depth': 8 'min_samples_split': 8 'min_samples_leaf': 1 'max_features': 6 'bootstrap': True 'random_state': 42	base_estimator_max_depth': 1 'learning_rate': 0.2 'loss': 'linear' 'n_estimators': 20 'random_state': 42	learning_rate': 0.05 'max_depth': 7 'max_features': 'sqrt' 'min_samples_leaf': 3 'min_samples_split': 8 'n_estimators': 10 'random_state': 42	colsample_bytree': 0.8 'learning_rate': 0.01 'max_depth': 5 'n_estimators': 300 'subsample': 0.6

图 2 各模型的 MSE、MAE 和 R^2 Fig. 2 MSE、MAE and R^2 of various ML models

示,所有模型的 R^2 均大于 0.3,说明对于数据集具有一定的可解释能力,其中 RF 和 AdaBoost 的 R^2 均大于 0.7,在预测精度上显著优于其他模型。

然后,通过评估模型在测试集上的预测精确度,量化了模型的泛化性能。全部模型 MSE 和 MAE 都分别低于 4.5×10^{-5} 和 6.0×10^{-3} ,其中 RF 模型 MSE 和 MAE 分别为 1.5×10^{-5} 和 3.1×10^{-3} ,AdaBoost 模型 MSE 和 MAE 分别为 1.74×10^{-5} 和 3.7×10^{-3} ,RF 模型 MSE 和 MAE 均为所有模型中的最低值,最终选择 RF 模型作为预测 IZ 降解性能的最优模型。

图 3 直观展示了每个模型的预测值与实验值的对比情况,数据点越靠近对角线就说明模型预测值与实验值接近。当实际 k 值 < 0.01 时,预测 k 值总是大于实际 k 值;而当实际 k 值 > 0.01 时,预测 k 值总是小于实际 k 值,可能是因为 $k < 0.01$ 的样本数量明显少于高 k 值样本。样本中的 $k < 0.01$ 的数据主要来源于不含金属的对照催化剂的实验结果以及不加过氧化物的对照试验结果,其 k 值均比正常实验结果要低至少一个数量级,导致模型可能对低 k 区间的学习不足,模型倾向于向均值方向收缩预测值,最终导致各个模型对于低 k 预测值都偏高,而对高 k 预测值都偏低。

此外,各模型在测试集上的残差 (Res) 分布情况如图 4 所示。所有模型的 Res 分布区间集中在 $-0.02 \sim 0.02$,说明预测结果相对可靠。其中,RF 模型预测值的 Res 分布在 $-0.01 \sim 0.01$,而且 90% 的数据点分布在 $-0.005 \sim 0.005$ 之间,表明其预测结果高度可靠。这一发现进一步强调了 RF 模型在预测 IZ 降解性能方面的优势。

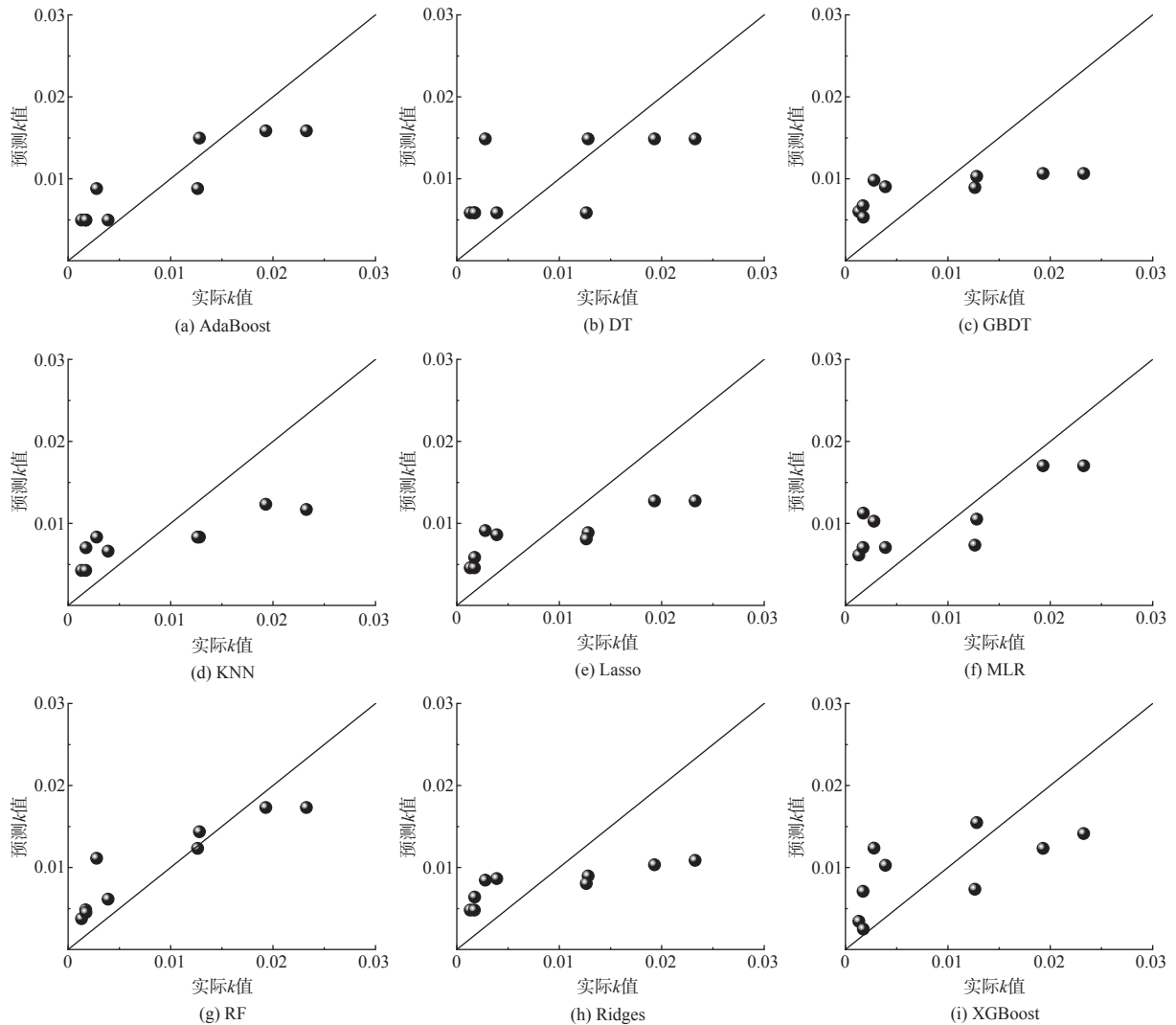
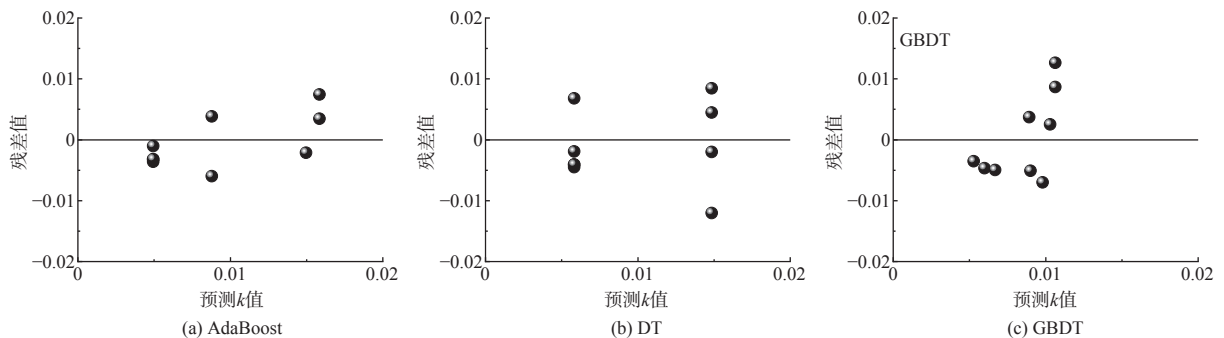


图3 实验值与 ML 模型预测值的对比

Fig. 3 Comparison of experimental values and ML model predicted values

总体而言, RF 模型在 R^2 (>0.75)、MSE ($<1.5 \times 10^{-5}$)、MAE ($<3.1 \times 10^{-3}$) 等指标上表现出色, 相较于其他模型具有最佳的总体性能。此外, RF 模型通常对离群值表现出较高的稳健性, 使得其对数据中的异常值不敏感。RF 模型通过构建多个决策树并取平均值或投票的方式来预测, 这种集成学习方法能够有效地捕捉数据中的复杂关系, 并处理非线性问题。这种灵活性使得 RF 模型特别适合处理具有复杂数据结构的数据集。因此, 选择 RF 模型作为预测 k 值的最佳模型。



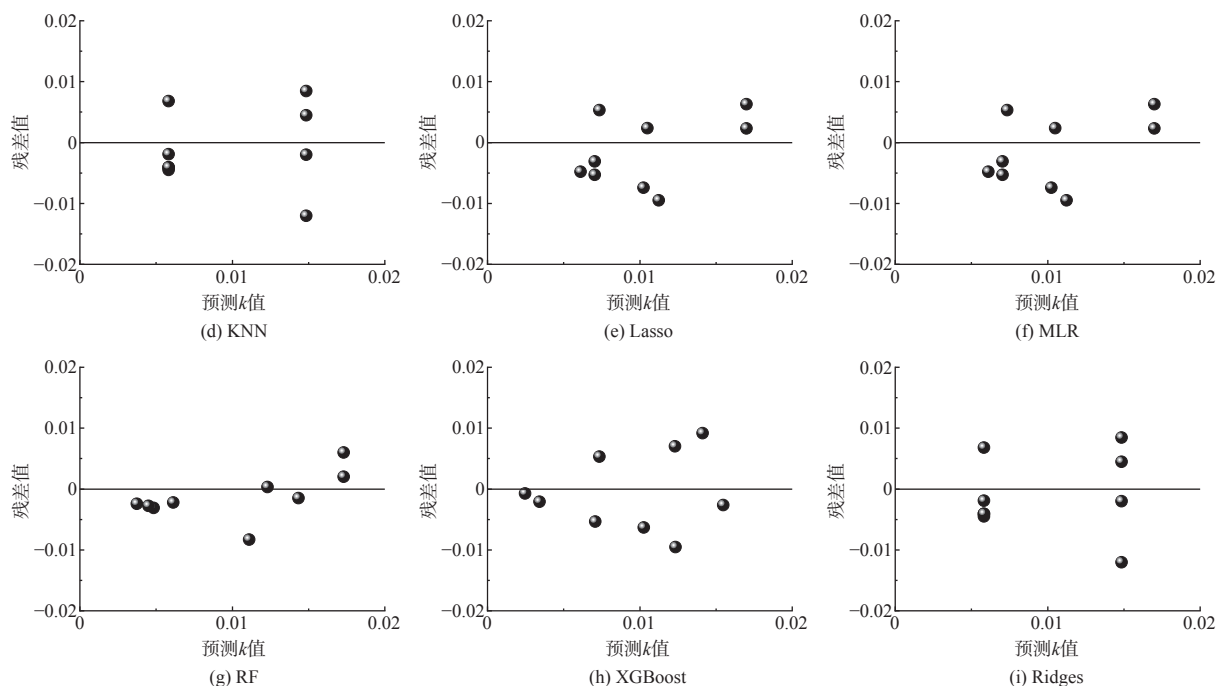


图 4 实验值与 ML 模型预测值的残差

Fig. 4 Residual of experimental values and ML model predicted values

2.3 RF 模型解释

基于各模型在测试集上的 R^2 进行排序, 选择排名第一的 RF 模型进行解释 (模型预测精确度高, 解释结果的可信赖度更高)。为了更全面地理解 RF 模型中 IZ 特征与目标变量之间的复杂关系, 采用了 SHAP 揭示特征对模型预测结果的影响, 从而识别出影响 IZ 降解性能的关键因素。如图 5 所示, y 轴上显示的特征排名表示预测模型中每个描述符的重要性, 在 x 轴上表示的 SHAP 值用作量化模型内特定描述符的影响的统一度量, 每一行的小提琴图说明了描述符对模型结果的归因, 其中红色部分表示高特征值, 蓝色部分对应于低特征值。通过 SHAP 分析发现, 铁的添加量的 SHAP 值最高, 表明铁的添加量是调控降解性能的主导因素。这一结论与铁

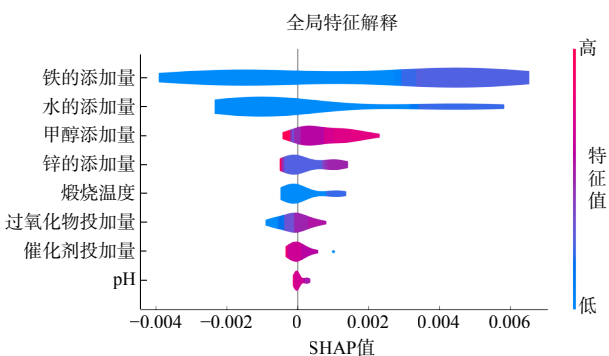


图 5 RF 模型的特征值—SHAP 值小提琴图

Fig. 5 Violin plot of SHAP values for the RF model

的 SHAP 值最高, 表明铁的添加量是调控降解性能的主导因素。这一结论与铁的强电负性诱导的电子重构机制一致: 铁掺杂促使电荷从碳基质向铁原子转移从而有效调控催化剂的电子结构, 提升催化剂导电性^[39-40]。同时通过形成 Fe-N_x 配位结构, 促进活性位点的单原子级分散^[41]。在一定范围内, 铁的添加量与 k 呈正相关性, 是因为当铁含量 < 10 wt.% 时, 铁能够均匀分散在催化剂表面并通过特定的金属-有机络合作用构建形成高活性的铁单原子^[42]。当铁的含量增加时, 铁单原子之间的距离逐渐减小, Fe-N_x 配位之间的强相互作用会改变电子结构从而增强活性。排名第二和第三的特征分别是水的添加量和甲醇的添加量。其中, 水的添加量的 SHAP 值大于甲醇的添加量的 SHAP 值主要是因为在水作为溶剂的液体环境中, 2-甲基咪唑主要是依靠饱和效应与 Zn²⁺ 发生自组装作用, 容易形成聚集态十二面体、方钠石形和菱形的 ZIF8 前驱体^[38], 容易构建多尺度的极性催化剂, 所以在一定范围内, 与 k 呈正相关性。而在甲醇作为溶剂的液体环境中, 2-甲基咪唑主要是依靠去质子效应与 Zn²⁺ 发生自组装作用, 作用时间缓慢, 容易形成十二面体的 ZIF8 前驱体^[37], 所以甲醇的添加量在一定范围内与 k 呈正相关性。但是当甲醇的添加量过多时, 反应动力学速率会下降, 容易导致配体之间反应不充分, 导致催化剂成核过大, 降低 ZIF8 的比表面积, 所以甲醇的添加量在一定范围内与

k 呈负相关性^[43]。其余特征的 SHAP 值主要集中在-0.001~0.002, 说明对 IZ 的降解性能有一定程度的影响。在一定范围内, 锌的添加量与 k 先呈负相关性, 后呈正相关性, 锌作为 ZIF8 模板剂, 其作用主要是构建 ZIF8 的框架和调控尺寸, 有利于其他金属的掺杂和提高催化剂的稳定性。在高温煅烧过程 (>600 °C) 中低沸点金属锌会蒸发, 同时有利于铁单原子的形成, 进一步提高了传质过程^[44-45]。适量提高锌的含量有利于高效 IZ 的形成, 但是过量锌的添加会形成锌氧化物, 从而阻碍其他金属掺杂在 ZIF8 上, 导致活性位点减少。煅烧温度与 k 呈现正相关的原因是因为当煅烧温度升高时, ZIF8 会碳化成为具有石墨相结构的多孔碳。超过 800 °C 时, 碳基质石墨化程度提升从而形成含更多 sp^2 杂化 σ 键的结构, 同时锌的挥发有助于形成稳定的 Fe-N_x 配位。由于催化剂的 π 键处于半充满状态, 这使 π 键周围的电子束缚性降低, 导致催化剂中的电子活性增加, 有利于电子传递^[46]。过氧化物的投加量与 SHAP 值呈明显的正相关性, 其作用主要是在于激活反应体系从而产生活性自由基, 其次能够直接参与降解反应^[2,47-48]。此外, 催化剂的投加量和 pH 与 SHAP 值都呈现负相关性, 但是其影响都相对较小, 因为 IZ 的活性依赖于高度分散的铁单原子位点, 当催化剂的投加量过多时, 可能会诱发铁原子的迁移和团聚, 形成低活性的纳米颗粒或团簇, 导致部分活性位点被覆盖^[49]。在类芬顿体系中, 酸性环境下的降解性能总是高于碱性条件, 是因为酸性条件下存在更多的 H⁺, 有利于金属位点的还原, 提高对过氧化物的利用率, 保证反应的高效进行^[50]。RF 模型的 SHAP 值的绝对值均值排序与各特征的 SHAP 值一致 (图 6), 进一步验证了结论的准确性。

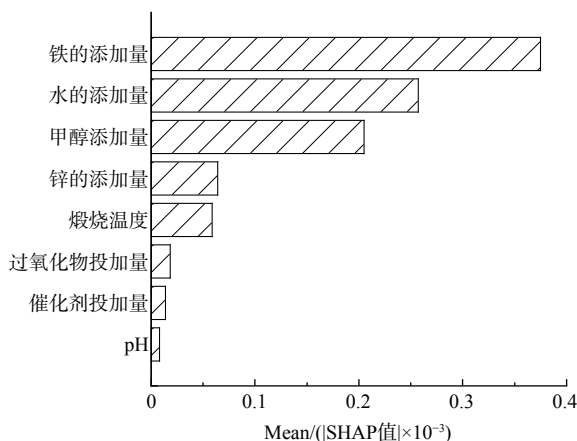


图 6 RF 模型的特征重要度排名

Fig. 6 Feature importance ranking of the RF model

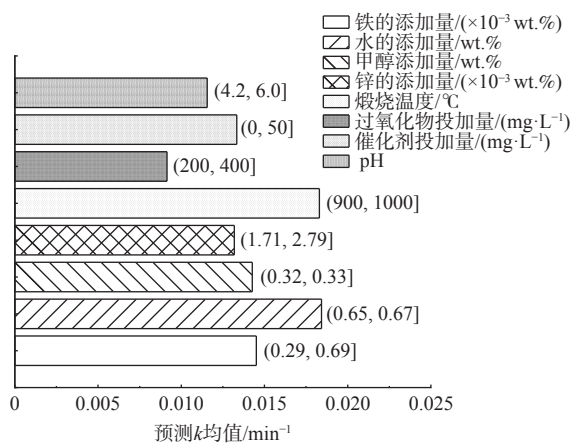
采用分段分布方法探究了 IZ 设计的最优取值范围, 旨在为设计具备高效降解性能的 IZ 提供科学依据。如图 7 所示, 铁的添加量应控制在 $0.29 \times 10^{-3} \sim 0.69 \times 10^{-3}$ wt.%, 水的添加量则为 $0.65 \times 10^{-3} \sim 0.67 \times 10^{-3}$ wt.%, 甲醇添加量为 $0.32 \times 10^{-3} \sim 0.33 \times 10^{-3}$ wt.%, 煅烧温度的最优区间为 900~1000 °C, 过氧化物的投加量在 200~400 mg·L⁻¹, 催化剂的投加量在 0~50 mg·L⁻¹, 体系 pH 保持在 4.2~6.0, 体系的降解性能能够达到了最优水平。

RF 模型通过数据学习到了 IZ 合成因素与降解性能之间的潜在关系, 并成功识别出了影响降解性能的关键因素。作为数据驱动模型, RF 模型的识别能力受到输入特征变化范围、数据集中度等因素的影响。然而, 这并不影响其在学术领域和实践应用中的重要价值。通过 RF 模型的解析, 不仅能够更深入地理解 IZ 降解性能的内在机制, 还能够为材料设计提供科学依据和优化方向。

3 结论

1) 成功建立多个 ML 模型以预测 IZ 的降解性能, 各类 ML 模型的 R^2 均大于 0.30。其中, RF 模型的 R^2 超过了 0.75, 而且预测误差主要集中在-0.05~0.05 之间, 表明 RF 模型在预测 IZ 降解性能方面具有较高的准确性和可靠性。

2) 基于 RF 模型和 SHAP 方法, 明确了特征对降解性能影响的趋势: 在一定范围内, 增加铁的添加

图 7 基于 RF 模型预测最优 k 值的特征范围分布图Fig. 7 Histogram of feature range distribution for predicting optimal k values based on RF model

量、水的添加量、甲醇添加量和锌的添加量以及提高煅烧温度能有效形成高性能的 IZ。然后，在一定的酸性条件下，确定适量的催化剂用量以及适当增加过氧化物的投加量都会增强降解性能。

3) 基于 RF 模型预测了 IZ 的最优工艺参数组合：铁的添加量为 $0.29 \times 10^{-3} \sim 0.69 \times 10^{-3}$ wt.%, 水的添加量为 $0.65 \times 10^{-3} \sim 0.67 \times 10^{-3}$ wt.%, 甲醇添加量为 $0.32 \times 10^{-3} \sim 0.33 \times 10^{-3}$ wt.%, 锌的添加量为 $1.71 \times 10^{-3} \sim 2.79 \times 10^{-3}$ wt.%, 煅烧温度为 900~1000 °C。同时，过氧化物的投加量在 200~400 mg·L⁻¹，催化剂的投加量在 0~50 mg·L⁻¹，体系 pH 保持在 4.2~6.0。

本研究利用 ML 技术不仅构建了高性能的预测模型，为 IZ 降解性能的预测提供了科学依据，还通过 SHAP 方法深入探讨了特征对催化剂性能的影响机制，为催化剂的设计与优化提供了理论支撑。但是本研究存在一定的局限性，预测模型的精度和泛化能力仍有待提升，这需要进一步收集更多高质量的实验数据，优化数据集质量和特征选择方法。同时，探索更先进的机器学习算法和技术，也是提高预测性能的重要途径。未来将开展实际应用研究，将理论预测结果转化为实际催化剂设计与优化方案。

参考文献

- [1] SCHWARZENBACH R P, ESCHER B I, FENNER K, et al. The challenge of micropollutants in aquatic systems[J]. *Science*, 2006, 313(5790): 1072-1077.
- [2] LU C, HU C, WU J M, et al. Endogenous substances utilization for water self-purification amplification driven by nonexpendable H₂O₂ over a micro-potential difference surface[J]. *Environmental Science & Technology*, 2024, 58(52): 23241-23250.
- [3] DAUGHTON C G, TERNES T A. Pharmaceuticals and personal care products in the environment: Agents of subtle change?[J]. *Environmental Health Perspectives*, 1999, 107(suppl 6): 907-938.
- [4] KOUSER S, HEZAM A, KHADRI M J N, et al. A review on zeolite imidazole frameworks: Synthesis, properties, and applications[J]. *Journal of Porous Materials*, 2022, 29(3): 663-681.
- [5] BANERJEE R, PHAN A, WANG B, et al. High-throughput synthesis of zeolitic imidazolate frameworks and application to CO₂ capture[J]. *Science*, 2008, 319(5865): 939-943.
- [6] ZHU Q L, XU Q. Metal-organic framework composites[J]. *Royal Society of Chemistry Chemical Society Reviews*, 2014, 43(16): 5468-5512.
- [7] LIAO W, ZHOU S Y, WANG Z C, et al. Stress induced to shrink ZIF-8 derived hollow Fe-NC supports synergizes with Pt nanoparticles to promote oxygen reduction electrocatalysis[J]. *Journal of Materials Chemistry A*, 2022, 10(40): 21416-21421.
- [8] ZHANG S Q, YIN Z H, LU Z C, et al. A tandem dual-pathway non-radical water purification simultaneously induced by single-molecule peroxymonosulfate with a catalyst of coexisting Cu single atoms and Fe₃C nanoparticles[J]. *Chemical Engineering Journal*, 2024, 500: 157136.
- [9] ZHANG S Q, LU Z C, HU C, et al. Understanding the distance effect of the single-atom active sites in fenton-like reactions for efficient water remediation[J]. *Advanced Science*, 2024, 11(12): 2307151.
- [10] QIAO B T, WANG A Q, YANG X F, et al. Single-atom catalysis of CO oxidation using Pt/FeO[J]. *Nature Chemistry*, 2011, 3(8): 634-641.
- [11] XIE X Y, PENG L S, YANG H Z, et al. MIL-101-derived mesoporous carbon supporting highly exposed Fe single-atom sites as efficient oxygen reduction reaction catalysts[J]. *Advanced Materials*, 2021, 33(23): 2101038.
- [12] RAMPRASAD R, BATRA R, PILANIA G, et al. Machine learning in materials informatics: Recent applications and prospects[J]. *npj Computational Materials*, 2017, 3: 54.
- [13] CECHINEL M A P, NEVES J, FUCK J V R, et al. Enhancing wastewater treatment efficiency through machine learning-driven effluent quality prediction: A plant-level analysis[J]. *Journal of Water Process Engineering*, 2024, 58: 104758.
- [14] FU H Y, LI K, ZHANG C F, et al. Machine-learning-assisted optimization of a single-atom coordination environment for accelerated fenton catalysis[J]. *ACS Nano*, 2023, 17(14): 13851-13860.
- [15] ZHU M Y, WANG J W, YANG X, et al. A review of the application of machine learning in water quality evaluation[J]. *Eco-Environment & Health*, 2022, 1(2): 107-116.
- [16] ZHANG H Z, REN R, GAO X, et al. Synchronous monitoring agricultural water qualities and greenhouse gas emissions based on low-cost internet of things and intelligent algorithms[J]. *Water Research*, 2025, 268: 122663.
- [17] AN S R, WANG L N, JIN Q. ZIF-8 derived Fe single atom catalysts prepared by low-temperature pyrolysis for efficient peroxymonosulfate activation[J]. *Journal of Water Process Engineering*, 2024, 58: 104907.
- [18] CHAI Y, DAI H, DUAN X, et al. Elucidation of the mechanistic origin of spin-state-dependent P-doped Fe single-atom catalysts for the oxidation of organic pollutants through peroxymonosulfate activation[J]. *Applied Catalysis B: Environmental*, 2024, 341: 123289.
- [19] CHEN B, CHEN L, ZU L, et al. Zero-strain high-capacity silicon/carbon anode enabled by a MOF-derived space-confined single-atom catalytic strategy for lithium-ion batteries[J]. *Advanced Materials*, 2022, 34(21): 2200894.
- [20] HE J J, WAN Y, ZHOU W J. ZIF-8 derived Fe-N coordination moieties anchored carbon nanocubes for efficient peroxymonosulfate activation via non-radical pathways: Role of Fe_N sites[J]. *Journal of Hazardous Materials*, 2021, 405(5): 124199.
- [21] HOU H R, FANG L, LIU L J, et al. Efficient iron single-atom materials for environmental pollutants removal from aqueous solutions: A review[J]. *Journal of Cleaner Production*, 2023, 426: 139150.
- [22] HUANG Z, YU H, WANG L, et al. Ferrocene doped ZIF-8 derived Fe-N-C single atom catalyst to active peroxymonosulfate for removal of bisphenol A[J]. *Separation and Purification Technology*, 2023, 305(15): 122402.
- [23] LI H C, SHAN C, PAN B C. Fe(III)-doped g-C₃N₄ mediated peroxymonosulfate activation for selective degradation of phenolic compounds via high-valent

- iron-oxo species[J]. *Environmental Science & Technology*, 2018, 52(4): 2197-2205.
- [24] LI Y, YANG T, QIU S H, et al. Uniform N-coordinated single-atomic iron sites dispersed in porous carbon framework to activate pms for efficient bpa degradation via high-valent iron-oxo species[J]. *Chemical Engineering Journal*, 2020, 389: 124382.
- [25] YANG T, FAN S S, LI Y, et al. Fe-NC single-atom catalysts with high density of Fe-N_x sites toward peroxymonosulfate activation for high-efficient oxidation of bisphenol a: Electron-transfer mechanism[J]. *Chemical Engineering Journal*, 2021, 419: 129590.
- [26] YU Y, WANG Y, YANG F, et al. Meso/microporous single-atom catalysts featuring curved Fe-N₄ sites boost the oxygen reduction reaction activity[J]. *Angewandte Chemie-international Edition*, 2024, 64(3): e202415691.
- [27] ZHOU Q, LV N, WANG J Y, et al. The enhanced peroxymonosulfate activation ability and mechanism on low-coordinated Fe-N₃ single sites for organic pollutant degradation in wastewater[J]. *Applied Surface Science*, 2023, 640: 158333.
- [28] FREUND Y, SCHAPIRE R E. A decision-theoretic generalization of on-line learning and an application to boosting[J]. *Journal of Computer and System Sciences*, 1997, 55(1): 119-139.
- [29] MOORE II D H. Classification and regression trees, by Leo Breiman, Jerome H. Friedman, Richard A. Olshen, and Charles J. Stone. Brooks/Cole publishing, monterey, 1984, 358 pages, \$27.95[J]. *Cytometry*, 1987, 8(5): 534-535.
- [30] JI L, ZHAO Y W, LI W J, et al. An improved KNN method for antenna optimization[C]//Institute of Electrical and Electronics Engineers. International Conference on Microwave and Millimeter Wave Technology. Qingdao, 2023: 1-3.
- [31] OBUCHI T, KABASHIMA Y. Cross validation in lasso and its acceleration[J]. *Journal of Statistical Mechanics: Theory and Experiment*, 2016, 2016(5): 053304.
- [32] FREEDMAN D A. The bootstrap. In: Statistical models: Theory and practice[M]. Cambridge: Cambridge University Press, 2009: 41-60.
- [33] BREIMAN L. Random forests[J]. *Machine Learning*, 2001, 45(1): 5-32.
- [34] TAO Q L, XU P C, LI M J, et al. Machine learning for perovskite materials design and discovery[J]. *npj Computational Materials*, 2021, 7: 23.
- [35] CLIFT A K, DODWELL D, LORD S, et al. Development and internal-external validation of statistical and machine learning models for breast cancer prognostication: Cohort study[J]. *British Medical Journal*, 2023, 381: e073800.
- [36] FRIEDMAN J H. Greedy function approximation: A gradient boosting machine[J]. *The Annals of Statistics*, 2001, 29(5): 1189-1232.
- [37] BUSTAMANTE E L, FERNANDEZ J L, ZAMARO J M. Influence of the solvent in the synthesis of zeolitic imidazolate framework-8 (ZIF-8) nanocrystals at room temperature[J]. *Journal of Colloid and Interface Science*, 2014, 424(15): 37-43.
- [38] KIDA K, OKITA M, FUJITA K, et al. Formation of high crystalline ZIF-8 in an aqueous solution[J]. *Crystengcomm*, 2013, 15(9): 1794-1801.
- [39] ZHAO Y, WANG Q, HU R, et al. Oxygen reduction reaction performance of Fe-N-C catalyst with dual nitrogen source[J]. *Front Energy*, 2024, 18(6): 841-849.
- [40] GENG D S, YANG S L, ZHANG Y, et al. Nitrogen doping effects on the structure of graphene[J]. *Applied Surface Science*, 2011, 257(21): 9193-9198.
- [41] 李志文, 陈从标, 王俊刚, 等. 氮掺杂介孔炭负载 FeCu 双金属催化剂及其 CO 加氢性能研究[J]. *燃料化学学报*, 2019, 47(6): 709-717.
- [42] LIANG X, FU N H, YAO S C, et al. The progress and outlook of metal single-atom-site catalysis[J]. *Journal of the American Chemical Society*, 2022, 144(40): 18155-18174.
- [43] FREUND R, LANZA A E, CANOSSA S, et al. Understanding the chemistry of metal oxide to metal-organic framework reactions for morphology control[J]. *Chemistry of Materials*, 2023, 35(5): 1891-1900.
- [44] WANG Z H, JIN H H, MENG T, et al. Fe, Cu-coordinated ZIF-derived carbon framework for efficient oxygen reduction reaction and zinc-air batteries[J]. *Advanced Functional Materials*, 2018, 28(39): 1802596.
- [45] ZHAO J J, QUAN X, CHEN S, et al. Cobalt nanoparticles encapsulated in porous carbons derived from core-shell ZIF67@ZIF8 as efficient electrocatalysts for oxygen evolution reaction[J]. *ACS Applied Materials & Interfaces*, 2017, 9(34): 28685-28694.
- [46] 刘欣, 张雅欣, 陈滢, 等. 石墨烯和氧化石墨烯制备技术及应用研究进展[J]. *陶瓷学报*, 2023, 44(2): 217-235.
- [47] LU C, HU C, LYU L. Nonexpendable H₂O₂ induces rapid molecular oxygen activation over micropotential Cu-modulated zinc oxide[J]. *Cell Reports Physical Science*, 2025, 6(2): 102431.
- [48] SUN Y T, HU C, LYU L. H₂O₂ triggering electron-directed transfer of emerging contaminants over asymmetric nano zinc oxide surfaces for water self-purification expansion[J]. *Journal of the American Chemical Society*, 2024, 146(1): 271-280.
- [49] 邱汉迅, 李文露. 非贵金属基催化剂用于催化降解有机污染物的研究进展[J]. *有色金属材料与工程*, 2022, 43(2): 1-10.
- [50] LU C, FANG Q, HU C, et al. Sustainable micro-activation of dissolved oxygen driving pollutant conversion on Mo-enhanced zinc sulfide surface in natural conditions[J]. *Fundamental Research*, 2023, 3(3): 422-429.

(责任编辑:陶雪)

Iron monoatomic catalyst design and its performance prediction for emerging contaminants degradation based on multi-algorithm machine learning

LI Ming¹, LUO Xin¹, HU Chun¹, LYU Lai^{1,2,*}

1. Key Lab for Water Quality and Conservation of the Pearl River Delta, Ministry of Education, China, Institute of Environmental Research at Greater Bay, Guangzhou University, Guangzhou 510006, China; 2. School of Environmental Science and Engineering, Guangzhou University, Guangzhou 510006, China

*Corresponding author, E-mail: lyulai@gzhu.edu.cn

Abstract To efficiently evaluate the degradation performance of Iron-single-atom ZIF8 (IZ) towards emerging contaminants, this work employs machine learning (ML) to establish a predictive model for the degradation capability of IZ. Firstly, a correlation analysis was conducted on the degradation performance dataset of IZ comprising 46 samples to determine 8 input features for ML models. The dataset was utilized to train, optimize, and evaluate 9 ML models, including AdaBoost, Decision Tree (DT), *K*-Nearest Neighbors (KNN), Lasso Regression, Multiple Linear Regression (MLR), Random Forest (RF), Ridge Regression, eXtreme Gradient Boosting (XGBoost), and Gradient Boosting Decision Tree (GBDT). The feature importance of the black-box model with the highest prediction accuracy was calculated using the SHapley Additive exPlanations (SHAP) method, and the relationship between the features and degradation performance was explored. The results indicated that the RF model achieved a coefficient of determination (R^2) exceeding 0.75, with the residuals of its predictions primarily distributed between -0.005 and 0.005. The RF model assigned the highest importance to Fe addition and H₂O addition, as these factors significantly influenced the degradation performance. Furthermore, within certain ranges, increasing Fe addition, H₂O addition, methanol addition, Zn addition, calcination temperature, peroxide addition and pH enhanced the degradation performance of IZ. Based on the RF model, this work predicted the optimal combination of design process parameters for IZ, providing insights into the application of ML for predicting and optimizing the degradation performance of IZ towards emerging pollutants.

Keywords machine learning; iron single atom; zeolitic imidazolate framework-8 (ZIF8); emerging contaminants; material prediction; degradation performance