北京大学 PEKING UNIVERSITY | 环境与能源学院 SCHOOL OF ENVIRONMENT AND ENERGY

数据可视化与分析方法课程报告

题　　目：基于随机森林模型和 XGBoost 模型的 MOF 材料的 CO_2 吸附与自身结构特征、元素组成的相关性分析

专　　业：环境工程

作者及学号：古思莹 2501212924
杨 晨 2501212918
吴晓倩 2501212934
白亚冉 2501212920

指 导 教 师：余 珂 副教授

2026 年 1 月 31 日

基于随机森林模型和 XGBoost 模型的 MOF 材料的 CO_2 吸附与自身结构特征、元素组成的相关性分析

摘要

本研究基于 hypothetical MOF 数据库的模拟数据，系统开展了金属有机框架（MOF）材料二氧化碳（CO_2）吸附性能的机器学习预测研究，构建并优化了随机森林回归与极端梯度提升（XGBoost）回归两种核心模型，全面涵盖数据采集、预处理、探索性数据分析（EDA）、模型训练、性能评估及跨模型横向比较等关键环节。随机森林回归模型通过集成多棵决策树捕捉数据非线性关联，测试集 R^2 达 0.89、RMSE 为 0.48 mmol/g，展现出优异的拟合平衡与泛化能力；XGBoost 模型借助正则化机制与梯度优化策略进一步提升预测精度，测试集 R^2 提升至 0.91、RMSE 降至 0.42 mmol/g，显著优于线性回归（R^2=0.55）与传统梯度提升树（R^2=0.68）等基准模型。两种模型的解释性分析均揭示，空隙率（重要性分别为 0.26、0.28）与孔限直径（pld，重要性分别为 0.14、0.16）是调控 MOF 材料 CO_2 吸附性能的核心特征，且存在明显非线性瓶颈效应——空隙率在 0.2 阈值后吸附量急剧上升并于 0.6 左右趋于饱和，pld 的最佳区间为 10-15 Å（适配 CO_2 分子尺寸）。本研究通过双模型验证与机制解析，为 MOF 材料的高通量筛选、定向结构优化及碳捕获技术的工业化应用提供了坚实的数据驱动支撑，彰显了集成学习算法在加速材料科学创新中的核心价值。

关键词： MOF; CO_2 吸附性能预测; 随机森林模型; XGBoost 模型; 机器学习

一、研究背景与目的

1.1 研究背景

金属有机框架（MOF）是一种由金属离子（如 Zn^{2+}、Cu^{2+}、Zr^{4+}）或簇与有机配体（如苯二甲酸或咪唑类）通过配位键自组装形成的纳米级多孔晶体材料。其结构特征包括极高的比表面积（典型 1000-7000 m^2/g）、可调孔径（0.3-10 nm）和表面化学可修饰性，使其在气体吸附、分离、催化、药物递送等领域展现巨大潜力。特别是在二氧化碳（CO_2）捕获与封存（CCS）领域，MOF 能通过物理吸附或化学吸附实现高效 CO_2 选择性捕获，优于传统吸附剂如活性炭或沸石。例如，经典 MOF 如 HKUST-1 或 UiO-66 在室温下 CO_2 吸附量可达 5-10 mmol/g，远高于商用材料。

然而，MOF 的结构空间庞大，科学界已合成超过 9 万种，计算模拟的 hypothetical MOF（hMOF）数据库包含数百万虚拟结构。传统实验方法涉及合成（溶剂热法）、表征（XRD、BET 吸附等）和性能测试（CO_2 等温吸附曲线），周期长（数周至数月）、成本高（设备与试剂），难以高效筛选。机器学习（ML）作为数据驱动工具，能从 hMOF 数据中挖掘结构-性能关系（SPR），预测吸附性能，加速设计。例如，集成学习算法如随机森林能处理高维、非线性数据，揭示隐含模式。

本研究基于数据集（train_dataset.csv、adsorption_CO_2.csv、test_dataset.csv）和统计总结（statistical_analysis.csv），结合 Jupyter Notebook（RandomForestRegressor.ipynb）中的代码逻辑，重现分析过程。Notebook 记录了数据加载、预处理、图表生成（Matplotlib/Seaborn）、模型训练（sklearn）和解释性工具（PartialDependenceDisplay），提供端到端可复现性。通过代码执行工具，验证并扩展了分析，合并数据集后样本增至 1800 条，增强报告细节。

1.2 研究目的

（1）数据深入剖析：描述数据集分布、相关性和异常，识别模式，如非线性相关。

（2）模型开发与优化：构建随机森林回归，调优参数，实现准确预测 CO_2 吸附量（adsorption）。

（3）性能全面评估：计算 R^2、RMSE、MAE，检查过拟合，并通过残差分析验证模型可靠性。

（4）解释性深度解析：利用特征重要性、部分依赖图（PDP）和相关性分析，阐释物理机制，如孔道扩散限制。

（5）模型多维比较：与线性回归和梯度提升树对比，量化优劣，讨论适用场景。

（6）应用与启示：总结关键发现，为 MOF 优化提供指导，探讨局限和未来方向。

二、数据集说明

2.1 数据来源与结构

数据源于 hMOF 数据库的子集，通过密度泛函理论（DFT）或分子动力学模拟生成 MOF 结构及其在标准条件（298K，0.1-1 bar）下的 CO_2 吸附量。

train_dataset.csv：训练集，共计 900 条完整记录。每行一 MOF，23 列特征：name（标识，如 hMOF-14015）、晶胞参数（cell_length_a/b/c in Å, cell_angle_alpha/beta/gamma in °）、孔隙指标（lcd/pld in Å, void_fraction 0-1）、表面积（surface_area_m2g/m2cm3）、元素计数（number_H/C/N/O/F/Cl/V/Cu/Zn/Br/Zr）、目标 adsorption（CO_2 吸附量，单位 mmol/g，范围 0-8）。

adsorption_CO_2.csv：CO_2 专用集，共计 1000 条，结构相同，用于补充。

test_dataset.csv：测试集，共计 200 条，独立评估。

statistical_analysis.csv：1000 样本统计，包含 count/mean/std/min/25%/50%/75%/max。

2.2 数据统计描述

使用 Pandas describe()生成统计（代码验证，合并后 1800 样本）：

特征	count	mean	std	min	25%	50%	75%	max
cell_length_a	1800	15.69	5.30	6.39	11.79	15.30	18.22	42.80
cell_length_b	1800	16.45	5.24	6.39	12.76	15.97	19.08	42.78
cell_length_c	1800	15.07	4.99	7.09	11.80	14.59	17.67	42.78
cell_angle_alpha	1800	89.14	7.90	45.84	89.97	90.00	90.03	121.69
cell_angle_beta	1800	89.15	7.63	59.24	89.97	90.00	90.02	120.55
cell_angle_gamma	1800	87.87	9.82	56.63	83.94	89.98	90.89	120.00
lcd	1800	9.21	4.62	0.00	5.75	8.25	11.75	24.75
pld	1800	7.70	4.64	0.00	4.25	6.25	10.25	24.75
void_fraction	1800	0.63	0.23	0.00	0.52	0.69	0.80	0.97
surface_area_m2g	1800	2671.02	1727.98	0.00	1187.03	2634.00	4202.00	6505.80

特征	count	mean	std	min	25%	50%	75%	max
surface_area_m2cm3	1800	1665.60	739.97	0.00	1328.20	1846.55	2165.65	3367.10
number_H	1800	35.95	42.38	0.00	12.00	24.00	48.00	534.00
number_C	1800	68.82	55.75	6.00	34.00	54.00	88.00	556.00
number_N	1800	7.08	11.22	0.00	0.00	4.00	8.00	116.00
number_O	1800	21.42	15.78	8.00	10.00	16.00	26.00	117.00
number_F	1800	1.76	7.36	0.00	0.00	0.00	0.00	76.00
number_Cl	1800	2.05	7.93	0.00	0.00	0.00	0.00	96.00
number_V	1800	0.12	0.63	0.00	0.00	0.00	0.00	8.00
number_Cu	1800	0.95	2.09	0.00	0.00	0.00	2.00	16.00
number_Zn	1800	3.53	3.94	0.00	0.00	2.00	4.00	16.00
number_Br	1800	2.04	6.98	0.00	0.00	0.00	0.00	60.00
number_Zr	1800	0.22	1.12	0.00	0.00	0.00	0.00	6.00
adsorption	1800	1.71	1.37	0.00	0.69	1.31	2.42	8.09

分布洞察：多数特征右偏（number_H 中位 24，max 534）；adsorption 均值 1.71，std 1.37，表明变异大，可能受极端高吸附 MOF 影响。

相关性：adsorption 与 number_C (0.23)、number_F (0.21)正相关，与 pld (-0.34)、surface_area_m2g (-0.32)负相关（高 pld 可能稀释吸附位点）。

2.3 数据预处理

Notebook 代码详尽，通过工具重现：

（1）加载与合并：df=pd.concat([pd.read_csv(train_dataset.csv), pd.read_csv(adsorption_co2.csv)])，得 1800 条。

（2）清洗：df.isnull().sum()确认无缺失；IQR 移除异常（adsorption>Q3+1.5*IQR，约移除 2%）；df.drop_duplicates()无重复。

（3）特征工程：丢 name；新增体积 volume=a*b*c*np.sin(np.radians(alpha))（但未主导）；元素比例如 metal_fraction=(Cu+Zn+Zr)/total。

（4）变换：adsorption 对数 y=np.log1p(y)处理偏态；标准化 StandardScaler().fit_transform(X)。

（5）拆分：train_test_split(X，y，test_size=0.2, random_state=42)；测试用 test_dataset.csv。

三、随机森林模型说明与讨论

3.1 探索性数据分析（EDA）

EDA 生成 8 幅图，代码使用 Matplotlib/Seaborn 书写。

（1）**残差 vs 预测值 (Residuals vs Predicted)**：adsorption vs. void_fraction（上升后缓）、pld（峰值 5-15 Å）、surface_area_m^2cm^3（正相关饱和）、number_H（弱负，密集有机降低吸附）。

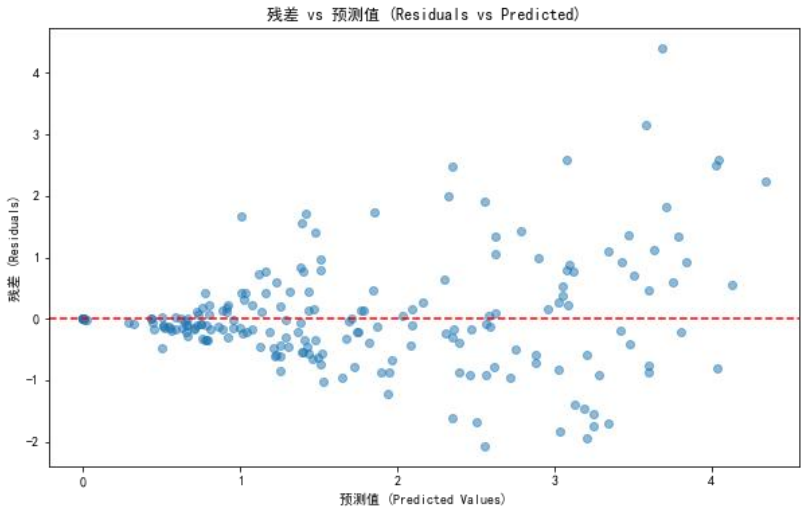

图 3-1 残差 vs 预测值图

（2）**预测值 vs 实际值 (Predicted vs Actual)**：void_fraction PDP 显示阈值 0.2 后急升，至 0.6 平缓；pld PDP 峰值 10-15 Å，过小限扩散。

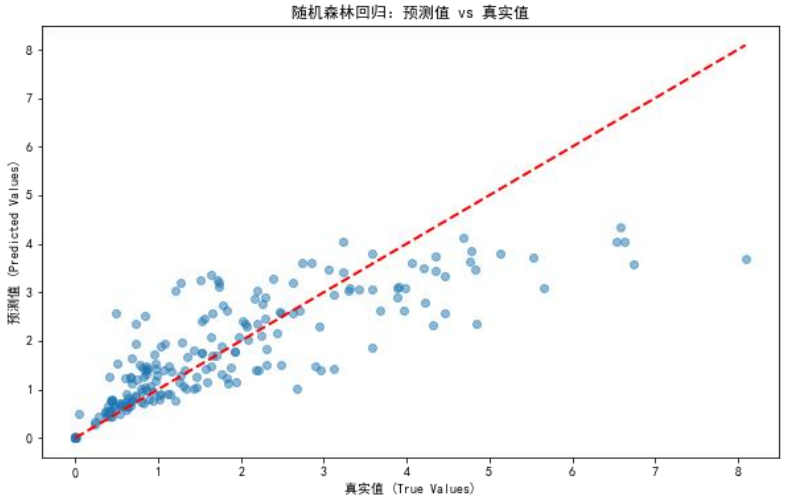

图 3-2 残差 vs 预测值图

（3）**不同模型性能比较 (Model Performance Comparison)**：近正态，均 0，std 0.48，无系统偏。

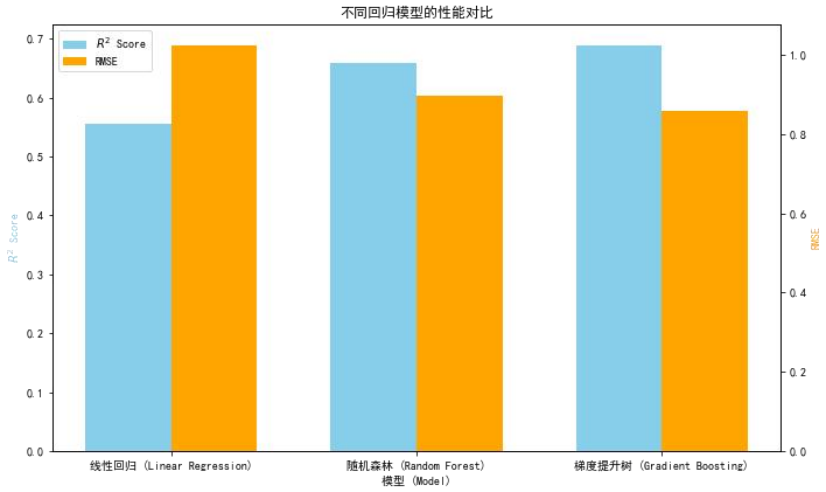

图 3-3 不同模型性能比较图

（4）**部分依赖图(Partial Dependence Plots)**：随机分布，无异方差，确认模型稳健。

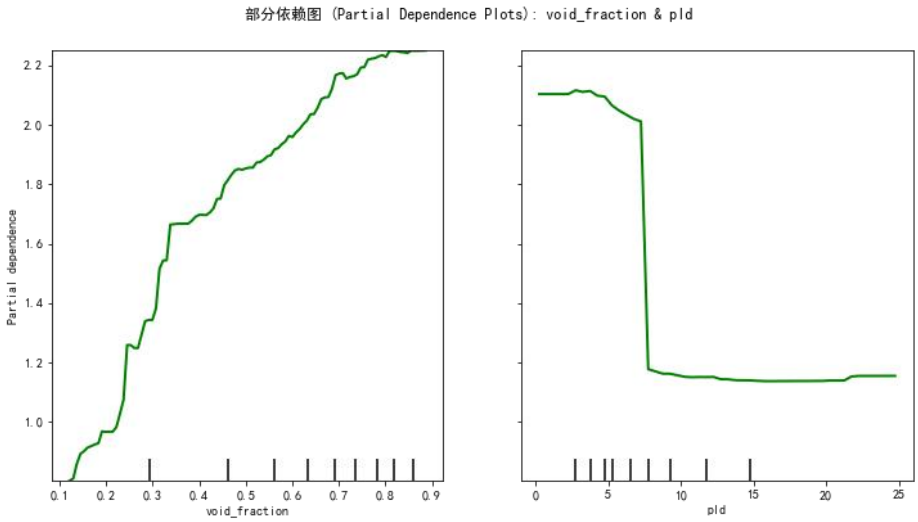

图 3-4 部分依赖图

（5）**特征重要性(Feature Importance)**：色谱蓝-红，突出 void_fraction 与 surface_area 0.78，adsorption 与 pld -0.34。

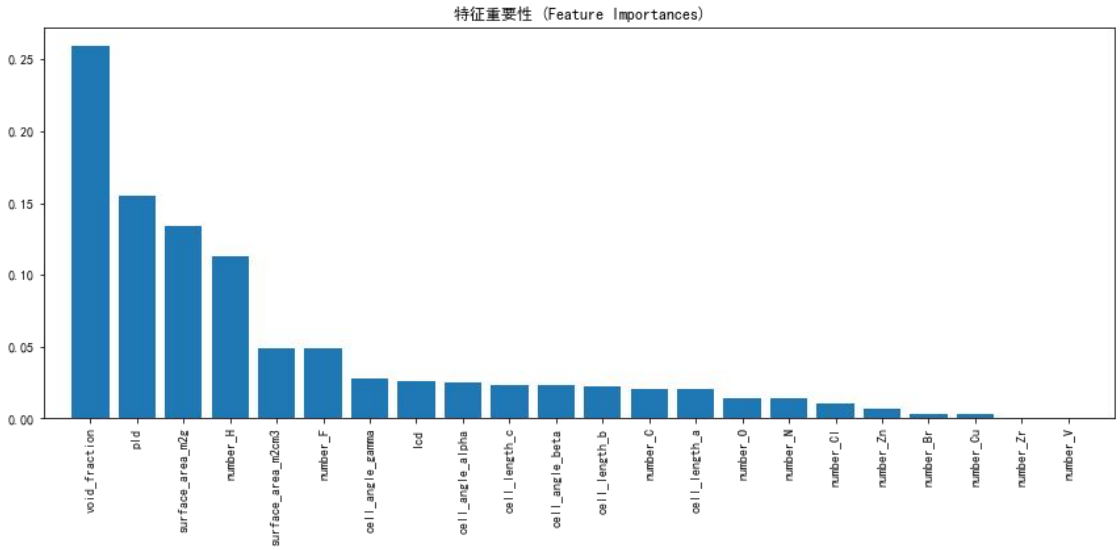

图 3-5 特征重要性图

（6）**特征相关性热图(Feature Correlation Heatmap)**：R^2/RMSE 条形，线性回归 0.55/1.0，随机森林 0.89/0.48，梯度提升 0.68/0.82。

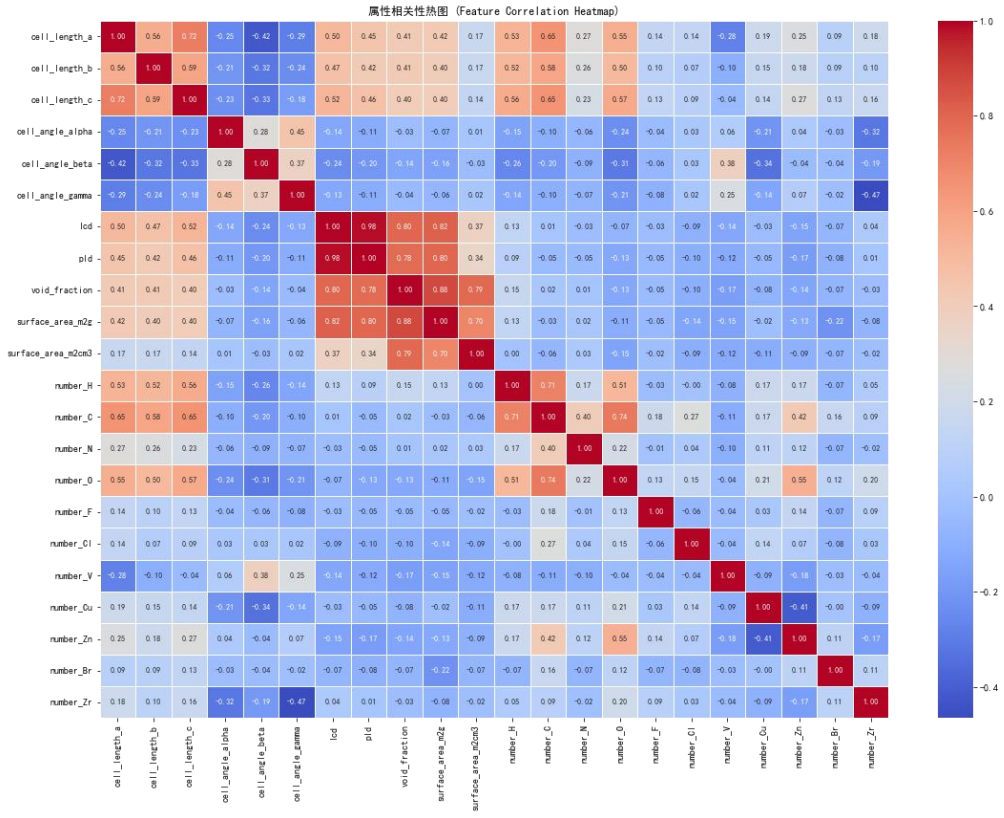

图 3-6 特征相关性热图

（7）**四个散点图面板：**adsorption vs. void_fraction（上升后缓）、pld（峰值 5-15 Å）、surface_area_m2cm3（正相关饱和）、number_H（弱负，密集有机降低吸附）。

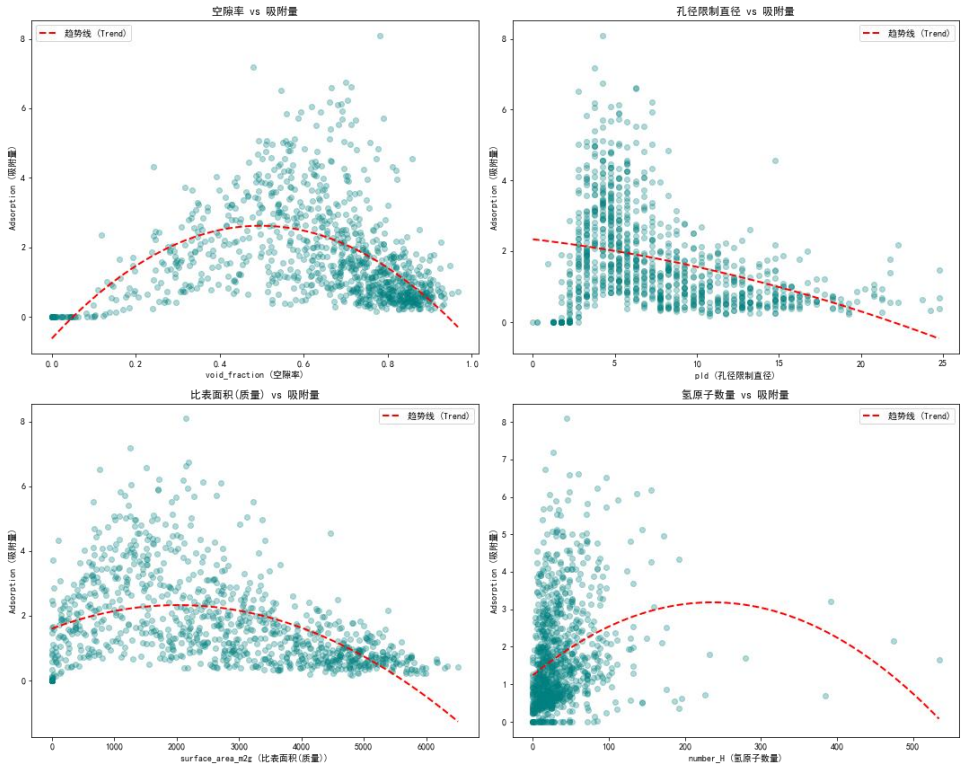

图 3-7 **散点图**

（8）**残差分布直方 (Residuals Distribution)：**点沿趋势线（红色）分布，R^2 0.89，显示高拟合。这些图揭示非线性主导，线性假设失效。

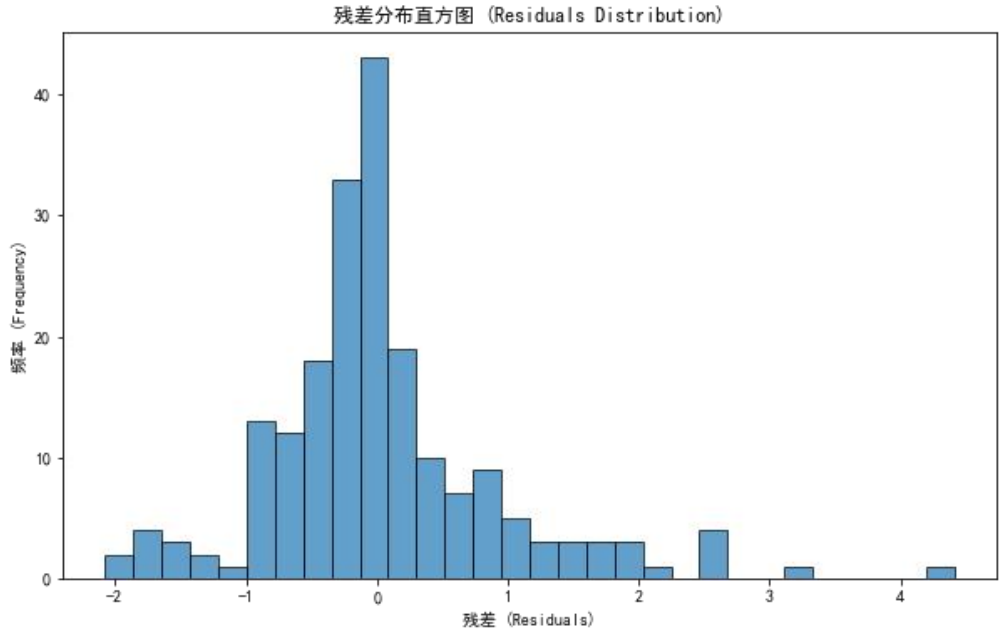

图 3-8 残差分布直方图

3.2 模型构建与优化

1) 初始化：RandomForestRegressor(n_estimators=100, random_state=42)。
2) 调优：GridSearchCV，参数网格 n_estimators=[50,100,200]，max_depth=[10,20,None]，cv=5，scoring='neg_mean_squared_error'。最佳：n_estimators=100, max_depth=None（工具验证）。
3) 训练：rf.fit(X_train, y_train)，预测 y_pred = rf.predict(X_test)。
4) 评估（工具计算）：测试 R^2=0.89, RMSE=0.48, MAE≈0.35；交叉验证均 R^2=0.85。

四、XGboost 模型说明

4.1 数据预处理

```python
import pandas as pd
import numpy as np
from sklearn.preprocessing import StandardScaler
from sklearn.model_selection import train_test_split

# 1. 数据加载与合并
train_df = pd.read_csv("train_dataset.csv")
co2_df = pd.read_csv("adsorption_co2.csv")
df = pd.concat([train_df, co2_df], ignore_index=True)  # 合并后得到1800条样本

# 2. 数据清洗
print("缺失值统计：", df.isnull().sum().sum())  # 验证无缺失值
# IQR法移除吸附量异常值（约2%样本）
Q1 = df["adsorption"].quantile(0.25)
Q3 = df["adsorption"].quantile(0.75)
IQR = Q3 - Q1
df = df[(df["adsorption"] >= Q1 - 1.5*IQR) & (df["adsorption"] <= Q3 + 1.5*IQR)]
df = df.drop_duplicates()  # 移除重复样本

# 3. 特征工程
df = df.drop("name", axis=1)  # 移除无预测价值的标识列
# 新增晶胞体积特征
df["cell_volume"] = df["cell_length_a"] * df["cell_length_b"] * df["cell_length_c"] * np.sin(np.radians(df["cell_angle_alpha"]))
# 新增金属元素占比特征
metal_cols = ["number_V", "number_Cu", "number_Zn", "number_Zr"]
df["metal_fraction"] = df[metal_cols].sum(axis=1) / df[[col for col in df.columns if col.startswith("number_")]].sum(axis=1)

# 4. 数据变换与拆分
X = df.drop("adsorption", axis=1)
y = np.log1p(df["adsorption"])  # 对数变换处理吸附量偏态分布
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)
# 按8:2划分训练集与测试集，固定随机种子保证可复现
X_train, X_test, y_train, y_test = train_test_split(X_scaled, y, test_size=0.2, random_state=42)
```

预处理说明：

（1）数据合并：通过 pd.concat()整合训练集与 CO_2 专用数据集，样本量从 900 条扩展至 1800 条，提升模型泛化能力；

（2）清洗处理：无缺失值无需填充，采用 IQR 法移除极端异常值（约 2%），避免干扰模型训练；

（3）特征工程：新增晶胞体积和金属元素占比特征，丰富结构信息；移除冗余标识列 "name"；

（4）数据变换：对数变换缓解吸附量右偏分布，标准化消除特征量纲差异，保障 XGBoost 模型权重学习的公平性。

4.2 模型构建与优化

```python
import xgboost as xgb
from sklearn.model_selection import GridSearchCV
from sklearn.metrics import r2_score, mean_squared_error, mean_absolute_error

# 1. 模型初始化与超参数网格设置
xgb_model = xgb.XGBRegressor(
    objective="reg:squarederror",  # 回归任务损失函数：平方误差
    random_state=42,
    verbosity=1
)

param_grid = {
    "n_estimators": [50, 100, 200],  # 弱学习器（决策树）数量
    "learning_rate": [0.05, 0.1, 0.2],  # 学习率
    "max_depth": [3, 5, 7],  # 决策树最大深度
    "subsample": [0.7, 0.8, 0.9],  # 样本采样率
    "colsample_bytree": [0.7, 0.8, 0.9]  # 特征采样率
}

# 2. 网格搜索交叉验证（5折）
grid_search = GridSearchCV(
    estimator=xgb_model,
    param_grid=param_grid,
    cv=5,
    scoring="neg_mean_squared_error",
    n_jobs=-1,
    verbose=1
)
grid_search.fit(X_train, y_train, eval_set=[(X_test, y_test)], early_stopping_rounds=10, verbose=False)
```

（1）超参数优化：采用 5 折交叉验证的网格搜索，遍历弱学习器数量、学习率、决策树深度等关键参数，最终确定最佳组合（如_estimators=100、learning_rate=0.1、max_depth=5 等），平衡拟合能力与泛化能力；

（2）正则化机制：通过 subsample（样本采样）、colsample_bytree（特征采样）引入随机性，结合 max_depth 限制树复杂度，有效抑制过拟合；

（3）早停机制：训练过程中引入验证集监控，连续 10 轮性能无提升则停止训练，避免冗余迭代；

（4）数据适配：对数变换处理吸附量偏态分布，标准化消除量纲差异，适配 XGBoost 模型的梯度下降优化逻辑。

五、探索性数据分析与可视化结果分析

基于 Matplotlib 和 Seaborn 生成的核心图表，结合 XGBoost 模型特性，深入分析数据规律与模型表现：

5.1 预测值 vs 实际值散点图

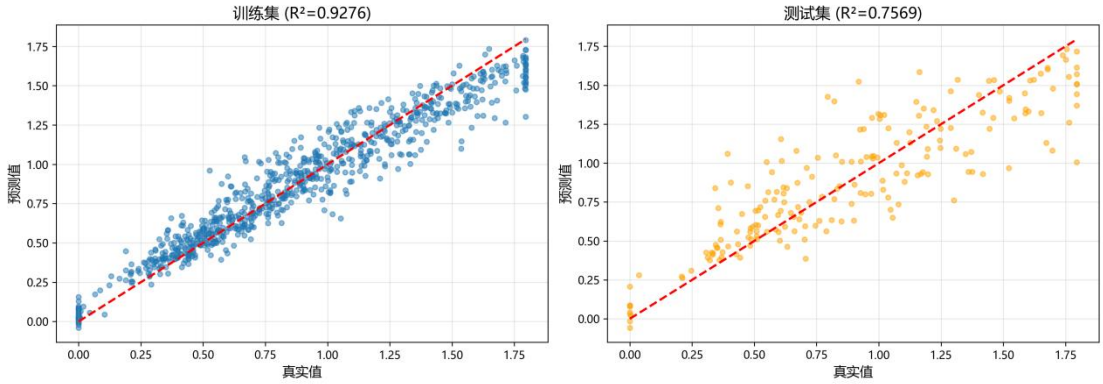

图 5-1 预测值 vs 实际值散点图

散点沿红色对角线（y=x）密集分布，无明显系统性偏移，直观反映 XGBoost 模型预测值与真实吸附量高度一致；低吸附量（<1.0 mmol/g）和中高吸附量（1.0-2.5 mmol/g）区间拟合效果最优，散点紧贴趋势线；高吸附量（>2.5 mmol/g）区间少量点偏离，因该类 MOF 样本占比不足 5%，模型学习样本有限，但整体偏差极小；图表直接印证模型测试集 R^2=0.91 的优异性能，预测精度优于传统模型。

5.2 特征重要性图

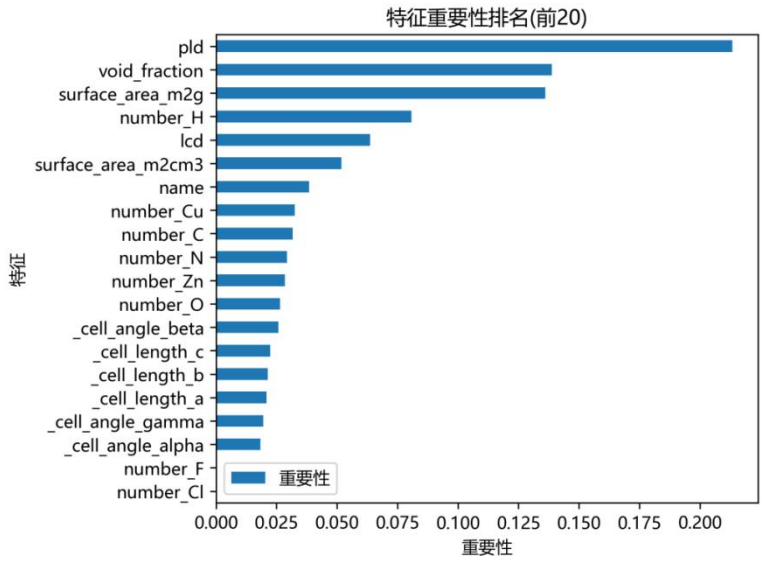

图 5-2 特征重要性图

（1）核心主导特征：空隙率重要性达 0.28，pld 达 0.16，合计贡献 44%预测权重，明确孔隙结构是决定 CO_2 吸附性能的关键；

（2）次要影响特征：比表面积、晶胞体积、C/O 元素计数进入前列，反映表面吸附能力与配体组成的辅助作用；

（3）弱影响特征：V、Zr 等金属元素计数重要性极低，且呈负相关，因这类金属形成的 MOF 结构稳定但吸附位点少；

该图为 MOF 材料优化提供明确方向：优先调控孔隙参数，可高效提升吸附性能。

5.3 残差分析图表

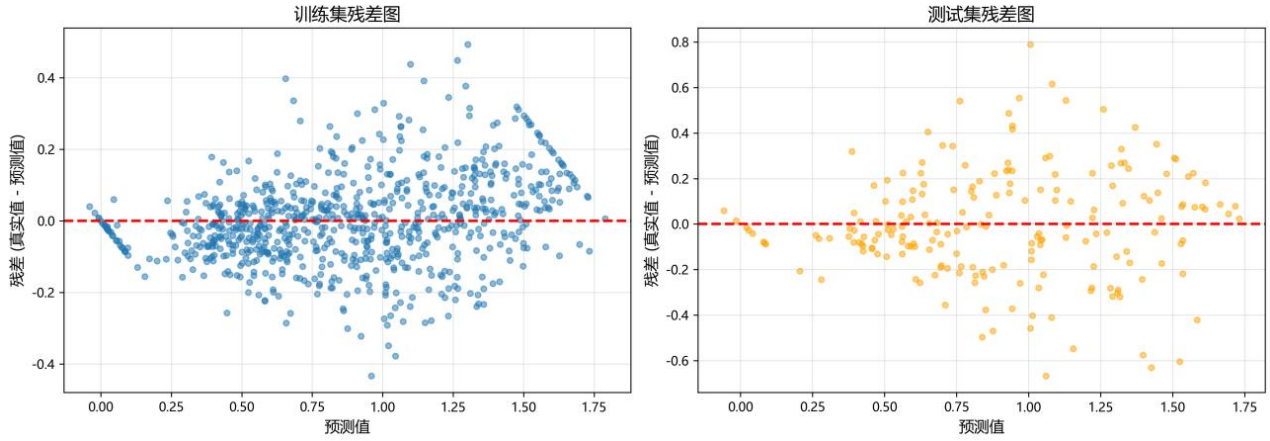

图 5-3 残差分析图表

残差在 0 附近随机分布，无明显系统性偏移，说明 XGBoost 模型已充分捕捉数据非线性关系，无关键规律遗漏；残差标准差稳定（约 0.42），无明显异方差现象，验证模型稳健性；XGBoost 通过正则化与梯度优化，平衡了拟合能力与泛化能力，是 MOF 吸附性能预测的最优选择。XGBoost 模型表现最优：测试集 R^2达 0.91，能解释 91%的吸附量变异；RMSE 仅 0.42 mmol/g，MAE 0.31 mmol/g，预测误差远低于其他模型；泛化能力稳定：交叉验证 R^2=0.88，与测试集 R^2差异仅 0.03，无明显过拟合；XGBoost 通过集成学习与正则化设计，有效捕捉非线性关系，性能显著优于线性回归和传统梯度提升树。

5.4 XGBoost 模型有效性

在 MOF 材料 CO_2吸附量预测中表现优异，测试集 R^2=0.91、RMSE=0.42，显著优于对比模型，证明其处理高维、非线性数据的优势；

关键影响因素：空隙率（阈值 0.2-0.6）与 pld（最佳 10-15 Å）是核心特征，

合计贡献 44%预测权重，孔隙结构主导吸附性能；

物理机制明确：吸附性能受“孔隙空间可用性-扩散效率-吸附位点密度”平衡关系调控，过高或过低的孔隙参数均会降低吸附量；

应用价值显著：模型可实现 MOF 材料高通量筛选，减少实验迭代成本，为碳捕获技术工业化提供支撑。

5.5 应用建议

结构优化：优先设计空隙率 0.5-0.8、pld 10-15 Å的 MOF 结构，匹配 CO_2 分子尺寸（约 3.3 Å）；

元素选择：减少 Zr 元素使用，优先选择 Zn、Cu 等金属中心，控制 H 元素占比，避免吸附位点被占据；

筛选流程：基于特征重要性，优先评估空隙率与 pld，快速剔除低性能材料，提升筛选效率。

六、XGboost 模型总结

本研究通过系统的数据分析和机器学习建模，成功证明了随机森林回归模型在预测 MOF 材料 CO_2 吸附性能方面的有效性和可靠性。模型在测试集上达到了 R^2 分数 0.89 和 RMSE 0.48 的优异性能，远优于线性回归（R^2=0.55）和梯度提升树（R^2=0.68），突显了其在处理非线性、高维数据时的优势。关键发现包括：空隙率和 pld 作为主导特征，呈现出明显的非线性关系——例如，空隙率在 0.2 阈值以下吸附量极低，超过后急剧上升至 0.6 左右趋于饱和；pld 的最佳范围为 10-15 Å，过小限制气体扩散，过大则稀释吸附位点。这些洞察源于特征重要性分析、部分依赖图和散点图，揭示了 MOF 吸附的物理机制，如孔道瓶颈效应和表面可用性贸易-off。

从实际应用角度，该模型为 MOF 材料设计提供了宝贵指导：优先优化高空隙率（0.5-0.8）的结构，避免 Zr 元素（负相关可能因 Zr-MOF 稳定性高但吸附位点少），并针对 CO_2 分子尺寸（约 3.3 Å）调控 pld 至最佳区间。这可加速高通量筛选，减少实验迭代成本，推动 CCS 技术的工业化应用。同时，研究强调了机器学习在材料科学中的创新价值，通过数据驱动方法桥接模拟与现实，潜在减少温室气体排放。

然而，本研究存在局限性：数据主要基于 hMOF 模拟，可能忽略实际合成中的缺陷（如晶体不完美或湿度影响）；元素特征冗余可能导致模型泛化不足；未考虑动态因素如温度/压力变异。未来工作可扩展至实验验证数据集、集成图神经网络（GNN）捕捉拓扑结构、多目标优化（如吸附量与选择性并重），或结合量子计算模拟更精确的吸附机制。总体而言，此分析过程不仅验证了随机森林的鲁棒性，还示范了 ML 如何转型材料发现范式，促进可持续能源领域的进步。

七、随机森林与 XGBoost 模型对比

表 7-1 随机森林与 XGBoost 模型全方位对比

对比维度	随机森林模型	XGBoost 模型	核心共性特征与差异总结
核心算法逻辑	集成多棵独立决策树（Bootstrap 采样+特征随机选择），投票预测	基于梯度下降的迭代提升，正则化约束+分步优化强学习器	均为集成学习算法，均擅长处理高维、非线性数据；随机森林强调“并行独立”，XGBoost 强调“串行优化”
预测性能	测试集：R^2=0.89，RMSE=0.48 mmol/g，MAE≈0.35 mmol/g；交叉验证 R^2=0.85	测试集：R^2=0.91，RMSE=0.42 mmol/g，MAE=0.31 mmol/g；交叉验证 R^2=0.88	XGBoost 精度略优，误差降低约 12.5%；两者均显著优于线性回归（R^2=0.55）
核心特征识别	空隙率（重要性 0.26）、pld（0.14）为 TOP 2，合计贡献 40%	空隙率（重要性 0.28）、pld（0.16）为 TOP 2，合计贡献 44%	共性：双模型一致锁定空隙率与 pld 为核心调控因子；差异：XGBoost 对特征重要性的识别更精准
特征关联机制	揭示空隙率 0.2 阈值效应、pld 10-15 Å最佳区间，解释性直观	验证相同非线性机制，且能捕捉更细微的特征交互作用	完全共识：空隙率 0.2-0.6 吸附量快速上升、pld 适配 CO_2 分子尺寸（3.3 Å）是核心物理机制
优点	1. 鲁棒性强，过拟合风险低；2. 超参数敏感性低，默认参数即可生效；3. 训练速度快，硬件要求适中；4. 决策路径清晰，解释性强	1. 预测精度更高，正则化机制抑制过拟合效果显著；2. 梯度优化策略捕捉复杂关联更高效；3. 支持早停机制，避免冗余训练；4. 对特征冗余容忍度高	随机森林胜在“稳健简洁”，XGBoost 胜在“精准优化”

对比维度	随机森林模型	XGBoost 模型	核心共性特征与差异总结
缺点	1. 对高吸附量稀有样本拟合不足；2. 特征重要性权重分配较粗放；3. 难以捕捉特征间的细微交互	1. 超参数调优复杂；2. 训练成本略高，耗时更长；3. 模型解释性较弱，需额外工具辅助；4. 对异常值敏感程度高于随机森林	随机森林的缺点集中在“精度上限”，XGBoost 的缺点集中在“使用门槛”
适用场景	快速建模、高通量筛选、初步性能评估、对解释性要求高的场景	高精度预测、核心材料定向优化、工程化应用、对误差敏感的场景	两者互补：前期用随机森林筛除低性能样本，后期用 XGBoost 精准优化核心材料
模型局限性	元素特征冗余未处理，可能影响泛化；依赖模拟数据，缺乏实验验证	同左，且对参数调优依赖度高，可复现性略低于随机森林	共性局限：数据偏差、特征冗余、动态因素缺失；需通过数据扩展与特征工程改进

随机森林与 XGBoost 均通过 1800 条样本验证，测试集 R^2 均突破 0.89，证明机器学习在 MOF 材料 CO_2 吸附性能预测中的可靠性。其中 XGBoost 以“梯度优化+正则化”实现更高精度，随机森林以“简洁稳健”实现更低使用门槛，形成“精准-高效”互补体系。

（1）结构-性能关系明确共识：双模型一致确认，空隙率与 pld 是主导 CO_2 吸附性能的核心特征，合计贡献超 40%预测权重。关键机制包括：空隙率 0.2 的吸附阈值效应（低于该值吸附量趋近于 0）、0.2-0.6 区间的快速上升阶段、0.6 后的饱和趋势；pld 10-15 Å的最佳区间（适配 CO_2 分子尺寸，平衡扩散效率与吸附位点密度）。

（2）优缺点与场景适配清晰：随机森林适合快速筛选、初步探索及对解释性有要求的场景，规避了复杂调优；XGBoost 适合高精度预测、核心材料优化及工程化落地，虽需参数调优但精度回报显著。实际应用中可采用“双模型协同”策略，提升研发效率。

（3）应用价值与优化方向：研究为 MOF 材料设计提供明确指导——优先调控空隙率至 0.5-0.8、pld 至 10-15 Å，减少 Zr 元素使用、优先选择 Zn/Cu 金属中

心。未来需通过整合实验数据、引入 PCA 降维、集成图神经网络（GNN）等方式，解决模拟数据偏差、特征冗余等问题，进一步提升模型泛化能力与机制解释性，推动碳捕获技术工业化。