

DOI: 10.11949/0438-1157.

机器学习辅助5-羟甲基糠醛的制备及工艺条件优化

陈大洋¹, 张艺濒¹, 陈奕铭¹, 杜玮辰², 赵怀远^{1,2}, 侯昭胤^{1,2}

(¹ 浙江大学化学系, 浙江省全省高值化学品低碳合成重点实验室, 浙江 杭州 310058; ² 浙江恒逸石化研究院有限公司, 浙江 杭州 311200)

摘要: 高质量标准化数据集匮乏是制约催化反应机器学习和建模的主要障碍。以5-羟甲基糠醛(5-Hydroxymethylfurfural, HMF)的制备为目标, 首先借助大语言模型的文本处理能力开展了大规模、选择性的文献数据提取; 然后结合机器学习进行反应原料、溶剂及相关催化剂的筛选; 并在机器学习的指导下有针对性地制备了具有多级孔结构的有机聚合物固体酸(hierarchical porous Sulfonated Polymeric BCMBP and HBSA, hSPBCH)催化剂; 最后通过人工智能辅助进一步优化了反应的工艺条件以提升HMF的产率。机器学习的研究结果表明: 以果糖为原料时得到的HMF平均产率最高, 表明其在产率导向的HMF制备体系中具有明显优势; 采用磺酸基固体酸催化剂及纯水溶剂, HMF的产率最高可达70.0%, 引入有机溶剂后可将HMF产率进一步提高至79.2%。经过机器学习后的优化和实验验证, 发现在自制的hSPBCH催化剂作用下、以果糖为原料、二甲基亚砜(dimethyl sulfoxide, DMSO)为溶剂、155℃下反应2h后, HMF产率提升至92.8%, HMF的时空产率最高可达45.6 g-HMF/g-cat/h。本研究提出的方法有助于推动机器学习在HMF制备及催化剂研制领域的发展与应用。

关键词: 5-羟甲基糠醛; 神经网络; 有机聚合物固体酸; 生物质; 催化; 工艺优化

中图分类号: O643.3

文献标志码: AA

文章编号: 0438-1157 (XXXX) XX-0001-13

Machine learning-assisted preparation of HMF and process optimization

CHEN Dayang¹, ZHANG Yibin¹, CHEN Yiming¹, DU Weichen², ZHAO Huaiyuan^{1,2}, HOU Zhaoyin^{1,2}

(¹ Zhejiang Key Laboratory of Low-Carbon Synthesis of Value-Added Chemicals, Department of Chemistry, Zhejiang University, Hangzhou 310058, Zhejiang, China; ² Zhejiang Hengyi Petrochemical Research Institute Co., Ltd., Hangzhou 311200, Zhejiang, China)

Abstract: The scarcity of high-quality standardized datasets is a major obstacle constraining machine learning and modeling in catalytic reactions. This study focuses on the preparation of HMF(5-hydroxymethylfurfural, HMF) via machine learning. First, leveraging the text-processing capabilities of large language models, large-scale and selective extraction of literature data was conducted. Subsequently, machine learning was employed to screen the raw materials, solvents, and related catalysts for HMF production. Based on the machine learning results, a hierarchically porous organic polymer solid acid catalyst (hierarchical porous Sulfonated Polymeric BCMBP and HBSA, hSPBCH) was purposefully prepared. Finally, artificial intelligence-assisted optimization of the reaction process conditions was carried out to enhance the yield of HMF. The machine learning results indicate that fructose

收稿日期: 2026-04-18 修回日期: 2026-05-13

通信作者: 侯昭胤(1968—),男,博士,教授,zyhou@zju.edu.cn

第一作者: 陈大洋(2002—),男,硕士研究生,1542199982@qq.com

基金项目: 中央高校基本科研业务费专项(26-2024-00060)

引用本文: 陈大洋, 张艺濒, 陈奕铭, 杜玮辰, 赵怀远, 侯昭胤. 机器学习辅助5-羟甲基糠醛的制备及工艺条件优化[J]. 化工学报, XXXX, XX(X): 1–13

Citation: CHEN Dayang, ZHANG Yibin, CHEN Yiming, DU Weichen, ZHAO Huaiyuan, HOU Zhaoyin. Machine learning-assisted preparation of HMF and process optimization[J]. CIESC Journal, XXXX, XX(X): 1–13

as the raw material yields the highest average HMF production, indicating that it has a distinct advantage in yield-oriented HMF preparation. Using sulfonic acid-based solid acid catalysts and pure water as the solvent, the yield of HMF can reach 70.0%, while the introduction of organic solvents can increase the HMF yield to 79.2%. After machine learning-guided optimization and experimental validation, it was found that under the action of the self-prepared hSPBCH catalyst, using fructose as the raw material and dimethyl sulfoxide (DMSO) as the solvent, the yield of HMF was increased to 92.8% at 155°C, with the highest space-time yield of HMF reaching 45.6 g-HMF/g-cat/h. The methodology proposed in this study can contribute to promoting the development and application of machine learning in the field of HMF preparation and catalytic reactions.

Keywords: HMF; neural networks; organic polymer solid acid; biomass; catalysis; process optimization

引 言

5-羟甲基糠醛(5-hydroxymethylfurfural, HMF)含有羟甲基和醛基官能团,是一种重要的生物质平台

化合物。HMF可以通过聚合、卤化、酯化、水解、氢化、氧化脱氢等反应转化为高附加值产品,如2,5-呋喃二甲酸、2,5-二甲基呋喃等,在聚合物、燃料和精细化学品合成等领域具有广阔的应用前景^[1,2]。因此,从可再生的生物质资源出发高效地制备HMF具有重要的意义和应用前景。

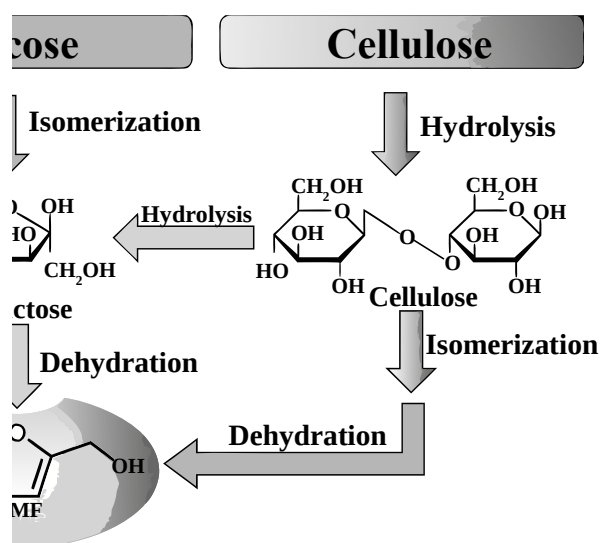


图1 果糖、葡萄糖、纤维素制备HMF的反应路径图

Fig.1 Reaction pathway diagrams for the preparation of HMF from fructose, glucose, and cellulose

目前,HMF的制备主要是在酸催化剂作用下,通过C6糖的脱水反应实现^[3,4],常用的原料有果糖、葡萄糖、纤维素等(图1)。在溶剂中,部分果糖以呋喃结构存在,其脱水反应的活性最高,可高效转变为HMF^[5]。早期的果糖脱水反应主要采用液体酸催化剂,如无机酸(H_2SO_4 、 HCl 、 H_3PO_4)和金属盐类(AlCl_3 、 CrCl_3)等^[6,7]。然而,均相催化剂存在设备腐蚀、回收困难、废液量大等问题,限制了其工业化应用^[8]。近年来,固体酸催化剂因其易于分离回收、可重复使用、环境友好等优势,正逐渐成为果糖脱水反应的研究热点^[9,10]。

尽管在制备HMF研究中已取得了显著的进展,但传统的催化剂研制和工艺优化仍面临诸多挑战。首先,基于试错法的实验策略耗时费力,难以在庞大的参数空间中高效寻找最优的反应条件^[3,11]。其次,不同研究团队的实验数据分散在海量的文献中,这些宝贵的数据资源很难得到充分利用^[3]。此外,副反应的发生(如HMF水解生成乙酰丙酸、部分脱水中间体缩合生成腐殖质等)会降低HMF的选择性和产率,需要系统地理解多因素间的复杂相互作用^[2,12]。近年来,人工智能与机器学习技术在材料科学和化学工程领域展现出巨大潜力,为解决上述挑

战提供了全新的思路^[13-15]。机器学习能够从大量实验数据中自动找寻复杂的非线性关系,识别关键影响因素,并预测最优实验条件,从而可以显著加快新催化剂的发现和工艺条件的优化^[16, 17]。例如,Butler等^[18]利用机器学习成功预测了新型钙钛矿材料的性能;Chen等^[19]通过数据驱动方法优化了CO₂电还原催化剂的设计。然而,机器学习在催化领域的应用仍面临一个关键瓶颈——高质量标准化数据集的匮乏^[20, 21]。传统的人工数据提取方式不仅效率低,而且难以保证数据的一致性和完整性,这一问题严重制约了数据驱动方法在催化剂筛选中的应用。大语言模型(Large Language Models, LLMs)的快速发展为文献数据的自动化提取提供了新机遇^[22, 23]。LLMs具有强大的自然语言理解能力和信息抽取能力,能够从非结构化的科学文献中准确识别和提取关键信息,包括反应条件、催化剂性能、产率数据等^[24, 25]。这为构建大规模、高质量的催化反应数据集创造了有利条件,有望打破数据获取的瓶颈,推动机器学习在催化科学中的应用。

基于上述背景,本研究提出了一种融合大语言模型文献挖掘、机器学习建模、工艺智能优化及实验验证复现的完整研究范式,旨在探索HMF制备过程中的数据驱动反应体系筛选与工艺优化的方法。首先,借助大语言模型,从近60年(1966–2025)的4013篇文献中自动化提取并构建了包含821个高质量样本的标准化数据集;随后,采用XGBoost(eXtreme Gradient Boosting, XGBoost)机器学习模型并结合SHAP(SHapley Additive exPlanations, SHAP)可解释性分析,系统筛选出最优反应体系(包括原料、溶剂、催化剂类型);基于机器学习的指导,理性设计并合成了一种新型多级孔磺酸化有机聚合物固体酸催化剂(hierarchical porous Sulfonated Polymeric BCMBP and HBSA, hSPBCH);最后,通过迁移学习和贝叶斯优化算法进一步优化反应工艺参数,并通过实验验证对产率进行了评估。

1 计算与实验细节

本研究的策略如图2所示,包含5个阶段:(一)数据集收集和预处理;(二)模型搭建与开发;(三)用超参数优化提升模型性能;(四)模型可解释性分析,采用SHAP分析确定特征对预测值的贡献度,从而理解模型挖掘数据集的隐含模式,为催化体系优化

提供依据;(五)开展反应条件优化以进一步提高HMF产率。

1.1 数据集构建

以ScienceDirect数据库中的期刊论文为核心数据来源,以5-hydroxymethylfurfural/HMF为关键词批量获取文献元数据及可用的全文接口标识(DOI/PII)。每条记录以DOI为主键去重,仅保留唯一条目,以便后续读取正文,并提前构造可直接通过接口访问的地址。基于DOI调用出版商内容接口获取结构化全文(XML格式),从XML文本中抽取标题、期刊及发表信息、摘要、关键词、分节正文与参考文献等内容,清洗其中的空白字符与格式符号。调用Google语言模型接口先判断文章主题是否为生物质原料制备HMF,对通过判定的文章执行结构化信息抽取,主要涵盖反应条件、催化剂信息及活性数据等描述变量。反应条件采用温度、时间、压力、底物浓度、溶剂体系(包括单相、双相、离子液体、深共熔溶剂等)及助剂等参数描述;催化剂相关描述包含催化剂类型、名称、化学式、组成元素、活性位点、载体及负载量等信息;对于活性数据,每篇文献仅保留其中出现的最高产率及对应的反应条件与催化剂信息,以降低数据的处理难度。提取的结构化数据需经过单位统一化与表述规范化处理,生成汇总表用于后续模型输入。最终,从近60年(1966–2025)的4013篇文献中提取了821个符合要求的数据样本。

1.2 反应体系的筛选和优化

研究选取反应温度、反应时间、生物质原料类型、反应溶剂类型、催化剂组成等作为输入特征,以HMF产率作为输出特征。删除缺失率超过30%的特征列,并对其余缺失值采用众数进行填充。反应物和溶剂采用One-Hot编码,催化剂元素组成采用Multi-Hot编码。为避免过拟合和降低计算复杂度,对稀有特征值进行过滤,移除低频次的特征值。最终完整的特征向量可表示为:

$$X = [\mathbf{x}_{\text{num}}, \mathbf{x}_{\text{cat}}, \mathbf{e}, \mathbf{m}] \#(1)$$

式中, $\mathbf{x}_{\text{num}} = [T, t]$: 反应温度、反应时间; $\mathbf{x}_{\text{cat}}$: 反应原料、溶剂; $\mathbf{e}$: 催化剂组成; $\mathbf{m}$: 酸中心类型。

将文献中的HMF产率归并成三类:

$$Y_{\text{class}} = \begin{cases} \text{低产率} & \text{if } 0 \leq Y < 40\% \\ \text{中产率} & \text{if } 40\% \leq Y < 60\% \\ \text{高产率} & \text{if } 60\% \leq Y \leq 100\% \end{cases}$$

选择XGBoost做为预测模型,并基于梯度提升框架,通过迭代多个弱学习器来构建强学习器:

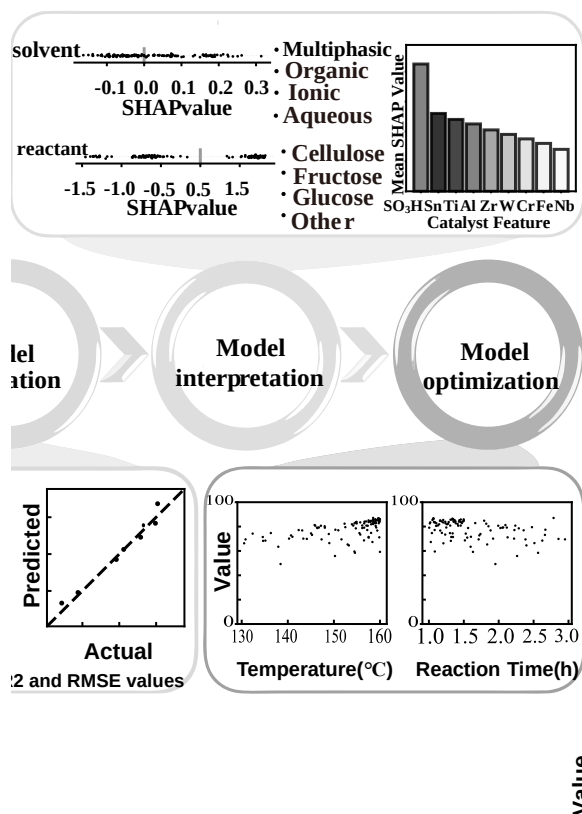


图2 本研究的流程框架

Fig.2 The process framework of this study

$$F(x) = F_0(x) + \alpha_1 h_1(x) + \alpha_2 h_2(x) + \cdots + \alpha_n h_n(x) \quad (2)$$

式中, $F(x)$ 是最终预测函数; $F_0(x)$ 表示初始的强学习器; $h_i(x)$ 是第 i 个弱学习器; α_i 是第 i 个弱学习器的权重。

目标函数定义为:

$$L(\phi) = \sum_{i=1}^n l(y_i, \hat{y}_i) + \sum_{k=1}^K \Omega(f_k) \quad (3)$$

式中, $l(y_i, \hat{y}_i)$ 为损失函数, y_i 为真实产率类别, $\hat{y}_i$ 为模型预测类别; $\Omega(f_k) = \gamma T + \frac{1}{2} \lambda \|w\|$ 为正则化项, T 为叶子节点数, w 为叶子节点权重向量, γ 和 λ 为正则化项的超参数。

采用分层抽样确保训练集和测试集各类别比例与原始数据集一致, 避免类别不平衡导致的评估偏差, 其中 80% 为训练集, 20% 为测试集。采用 5 折交叉验证评估模型的稳定性:

$$CV - Score = \frac{1}{K} \sum_{k=1}^K Accuracy_k \quad (4)$$

式中, $Accuracy_k$ 表示第 k 折的预测准确率。

采用以下指标评估模型性能:

准确率(Accuracy)

$$Acc_c = \frac{TP + TN}{TP + TN + FP + FN} \quad (5)$$

精确率(Precision)

$$Prec_c = \frac{TP_c}{TP_c + FP_c} \quad (6)$$

召回率(Recall)

$$Rec_c = \frac{TP_c}{TP_c + FN_c} \quad (7)$$

F1 分数

$$F1_c = 2 \times \frac{Prec_c \times Rec_c}{Prec_c + Rec_c} \quad (8)$$

式中, c 表示类别; TP、TN、FP、FN 分别表示真阳性、真阴性、假阳性、假阴性。

除此之外, 对数据集还进行了 SHAP 分析, 以理解包括催化剂材料在内的各描述符对模型预测的影响。该方法有助于确定描述符的相对重要性及对模型预测的负面或正面影响, 为 XGBoost 提供一定的可解释性。

1.3 优化后催化剂的制备

根据前面的优化结果制备了系列含有磺酸基的有机聚合物 hSPBCH 催化剂(hierarchical porous Sulfonated Polymeric BCMBP and HBSA), 催化剂的

制备工艺及结构如图3所示。制备催化剂所用的原料、操作步骤和表征方法及结果见附件(supporting

information, SI)。

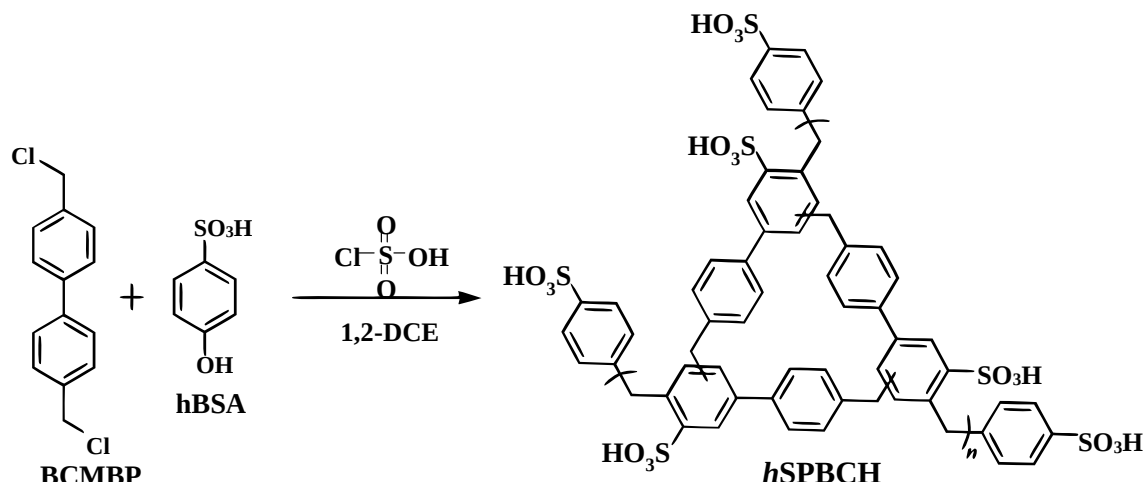


图3 催化剂hSPBCH的制备过程示意图

Fig.3 Schematic diagram of the preparation process for catalyst hSPBCH

1.4 验证反应

在典型实验中,将果糖(0.5 g)、hSPBCH 催化剂(0.025 g)和溶剂(10 mL)置于不锈钢高压反应釜中。随后用 N_2 吹扫高压反应釜将空气排尽并加压至1 MPa。接着将高压反应釜预热至所需温度后,立即开启搅拌并在搅拌条件下(1000 rpm)保持规定时间。反应结束后,将高压釜冷却至室温。对反应混合物进行离心去除固体催化剂,使用配备30 m毛细管柱(DB-WAX 52CB, USA)和火焰离子化检测器的气相色谱仪(Shimadzu 14B)进行检测,并采用内标法进行定量分析。果糖的转化率、产物选择性及产率按以下公式计算。

$$\text{Conversion}(\%) = \left(\frac{n_{\text{fructose, converted}}}{n_{\text{fructose, added}}} \right) \times 100\% \quad (9)$$

$$\text{HMF yield}(\%) = \left(\frac{n_{\text{HMF}}}{n_{\text{fructose, added}}} \right) \times 100\% \quad (10)$$

$$\text{HMF selectivity}(\%) = \left(\frac{n_{\text{HMF}}}{n_{\text{fructose, converted}}} \right) \times 100\% \quad (11)$$

1.5 工艺参数优化

以果糖为原料、采用hSPBCH催化剂和二甲基亚砜(dimethyl sulfoxide, DMSO)溶剂,选取4因素4水平的正交实验数据(见SI中表S1)及其它参比催化剂实验数据作为训练样本,反应温度($^{\circ}\text{C}$)、反应时间(h)、果糖浓度(g/10 mL)、催化剂用量(g)作为输入特征,以HMF产率(mol%)作为模型输出特征。为充分利用其它催化剂的实验数据,采用迁移学习策略开展模型训练:先在源数据域(即其它催化剂实验数

据)上完成预训练,再在目标数据域(hSPBCH实验数据)上进行微调。采用多层前馈神经网络构建单目标回归模型,超参数优化使用贝叶斯方法进行参数空间搜索,以提升模型性能。工艺参数优化采用Optuna的TPE(Tree-structured Parzen Estimator, TPE)采样器进行序贯贝叶斯优化。TPE采样器基于已完成试验的参数组合及其预测产率构建概率模型,并据此选择新的候选参数组合。每组候选条件的HMF产率由训练好的神经网络模型预测。优化过程在达到100次迭代后终止,并以预测产率最高的参数组合作为候选最优工艺条件。形式化表达如下:

$$\max_{x \in B} f(x)$$

2 结果与讨论

2.1 数据分析

图4a显示了在机器检索到的821篇有效文献中制备HMF的原料分布情况,可以发现果糖是最常见的原料(45%),其次是葡萄糖(27%)、纤维素(17%)和其它底物(11%)。图4b表明有效文献中DMSO和 H_2O 是两种最常见的溶剂,使用次数分别为147次和140次;其中,DMSO的高使用频率源于其可以稳定反应中间体((4R,5R)-4-羟基-5-羟甲基-4,5-二氢呋喃-2-甲醛)和抑制不期望的副反应^[26]。与此同时,水+有机溶剂的双溶剂体系也常被用做溶剂,该体系可以利用原位萃取策略,将产物从反应区域(水

相中)连续移除至有机相来提高HMF选择性,防止其进一步转化。虽然离子液体也可以用作反应溶剂,但是高成本和高粘度限制了其大规模工业应用^[27]。图4c定量描述了不同原料类型下对应的催化剂组成元素的出现频次分布。其中含铝的混合

氧化物为最常见的催化剂组分,在以果糖和葡萄糖为原料的文献中分别出现53次和66次,总计119次应用,这可归因于催化剂中 Al^{3+} 所表现出的Lewis酸性和低成本等。此外,Cr、Zr、W、Sn等酸性金属氧化物也是常用的催化剂组分。

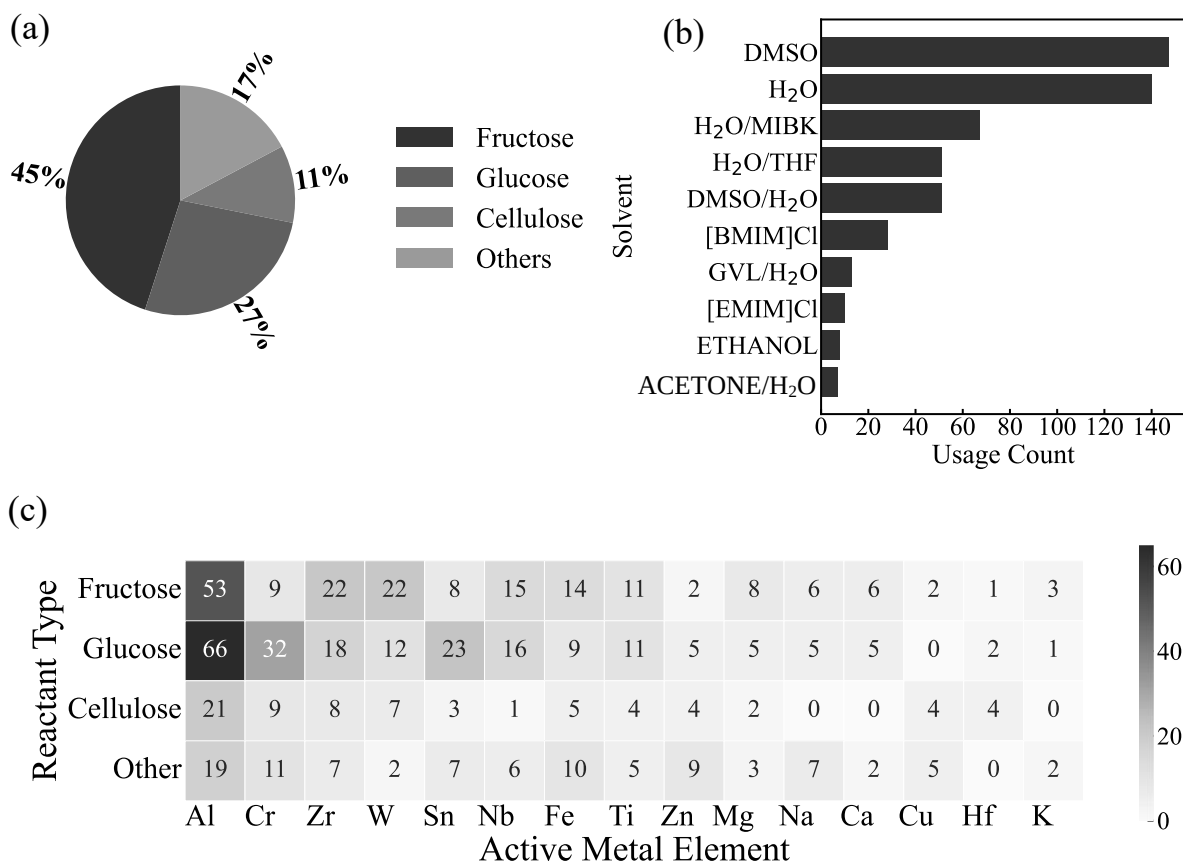


图4 制备HMF的数据集分析(a)原料分布、(b)溶剂使用频次、(c)原料-金属关联图

Fig.4 Analysis of the HMF preparation dataset: (a) feedstock distribution, (b) solvent usage frequency, (c) feedstock-metal correlation diagram

图5a是文献中采用不同反应原料制备HMF时的反应温度分布情况,可以发现以果糖为原料时反应温度较低,主要集中在120~150℃范围内,这可能是由于果糖具有开链的酮糖结构,无需克服糖苷键水解的能垒,反应活化能较低;相比之下,葡萄糖需要在较高温度下完成C1-C2的Lobry de Bruyn-Alberda van Ekenstein重排^[28],这导致其最佳反应温度比以果糖为原料时提高了约20℃(150~170℃);基于同样的原因及高聚合度,以纤维素为原料时所需的反应温度进一步升高至170~200℃。与此同时,图5b显示以果糖为原料时,HMF的产率多数处于65~95%范围内,而以葡萄糖为原料时,HMF的产率降低至40~65%。这种明显的差距主要是因为果糖仅需一步脱水即可得到HMF,而以葡萄糖为原料

时需要经过异构(至甘露糖和果糖)、逆羟醛缩合、进一步降解等过程,这明显增加了反应的难度^[29]。由于纤维素需要更多的中间过程才能得到HMF,以纤维素为原料时,HMF的产率降低至30~55%。文献结果还显示含有磺酸基的催化剂对所有的原料都具有较好的效果,HMF的中位产率约为79%(图5c),这可以归因于此类催化剂具有较强的Brønsted酸性,能够高效质子化糖分子的羟基,促进分子内脱水及异构化;而含Al、Cr和Zr的催化剂酸性位点以Lewis酸为主,决定了其可能主要作用于葡萄糖异构化步骤,但对脱水步骤的效率不及Brønsted酸^[30,31]。图5d则进一步显示在DMSO溶剂中,HMF的中位产率约为79%,较其他溶剂能更好地促进HMF的生成,这可能是由于DMSO能与HMF中的羟

基形成适当配位,可以有效抑制 HMF 的再水合和腐殖质形成^[32]。

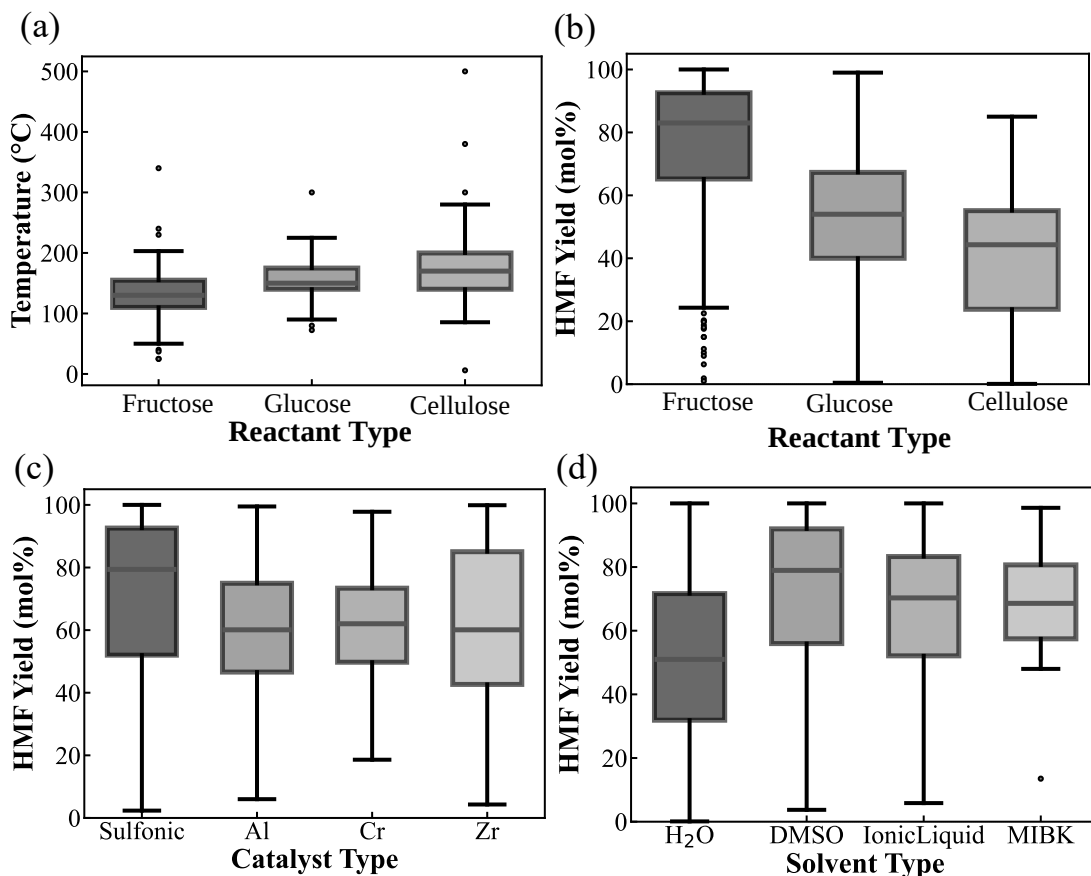


图5 (a)不同原料的反应温度分布、(b)不同原料的产率分布、(c)不同催化剂类型的产率分布、(d)不同溶剂类型的产率分布
Fig.5 (a) Reaction temperature distribution for different feedstocks, (b) yield distribution for different feedstocks, (c) yield distribution for different catalyst types, (d) yield distribution for different solvent types

2.2 模型性能评估

从表1可以看出,本文构建的XGBoost模型在高产率样本上的表现最为优异,F1分数达到0.85,显著高于中等产率(0.70)和低产率(0.59)。具体而言,高产率的召回率高达0.89,表明模型能有效识别89%的真实高产率催化体系,能够高效地指导实际催化剂筛选。虽然低产率类别的精确率达到0.70,但F1分数和召回率分别仅为0.59和0.51,说明该模型仅能识别约一半的真实低产率样本,这可能和低产率样本训练数量较少有关。

为了深入理解XGBoost模型的决策过程与预测

结果,本研究用SHAP方法对模型进行了全面的可解释性分析。SHAP值基于博弈论中的Shapley理论,能够量化每个特征对模型预测结果的边际贡献。图6为溶剂-反应物-催化剂SHAP分析。有机溶剂主要分布在SHAP值的正向区域(0.1~0.3),表明有机溶剂(如DMSO、DMF(N,N-dimethylformamide,DMF)等)可以有效促进HMF的生成。水相体系呈现双相分布特征,说明水作为溶剂时的效果高度依赖于其它反应条件(如催化剂类型、反应温度等)。离子液体集中分布在接近零值的区域,显示出偏中性的贡献。多相体系分布较为分散,反映了多相反应体系的复杂性和高度可调性。

对于反应原料,果糖集中分布在SHAP值>0.5的区域,显示出最强的正向贡献。这主要是因为果糖在溶液中部分以呋喃型结构存在,使其在脱水过程中更易转化为HMF,转化路径更为直接。葡萄糖的SHAP值分布在-0.5~0的区间,呈现中等偏负的

表1 模型性能表

Table 1 Model Performance Table

Catalyst Yield Categories	Precision	Recall	F1-Score
Low Yield	0.70	0.51	0.59
Medium Yield	0.67	0.72	0.70
High Yield	0.81	0.89	0.85

贡献。这可以归因于葡萄糖需要先异构化为果糖再脱水生成HMF,这一额外步骤降低了整体产率和选择性。纤维素(红色点)主要集中在强负值区域(-1.0~-1.5),对HMF产率有显著的负面影响。这一现象可以从两个层面解释:首先,纤维素的高度聚合结构导致其在催化体系中的溶解能力差,反应活性低;其次,纤维素需要经过多步水解和异构化才能转化为HMF,中间步骤增加了中间产物的产率和副反应发生的可能性(图6)。其他原料的SHAP值也处于-0.5~0区间,不是制备HMF的适合原料。

对催化剂活性位点的分析可以看到,磺酸基催化剂的SHAP值显著超越所有的金属氧化物催化剂,主要是因为1.磺酸基催化剂能在反应体系中提供丰富且强度均匀的酸性位点;2.强Brønsted酸能高效质子化糖分子的羟基,促进脱水反应的进行;3.磺酸基对C5-C6糖的呋喃环脱水具有高度的选择性,抑制副反应的生成^[33]。金属氧化物基催化剂SHAP值主要集中在0.02~0.03,对HMF的生成有一定程度的促进作用。

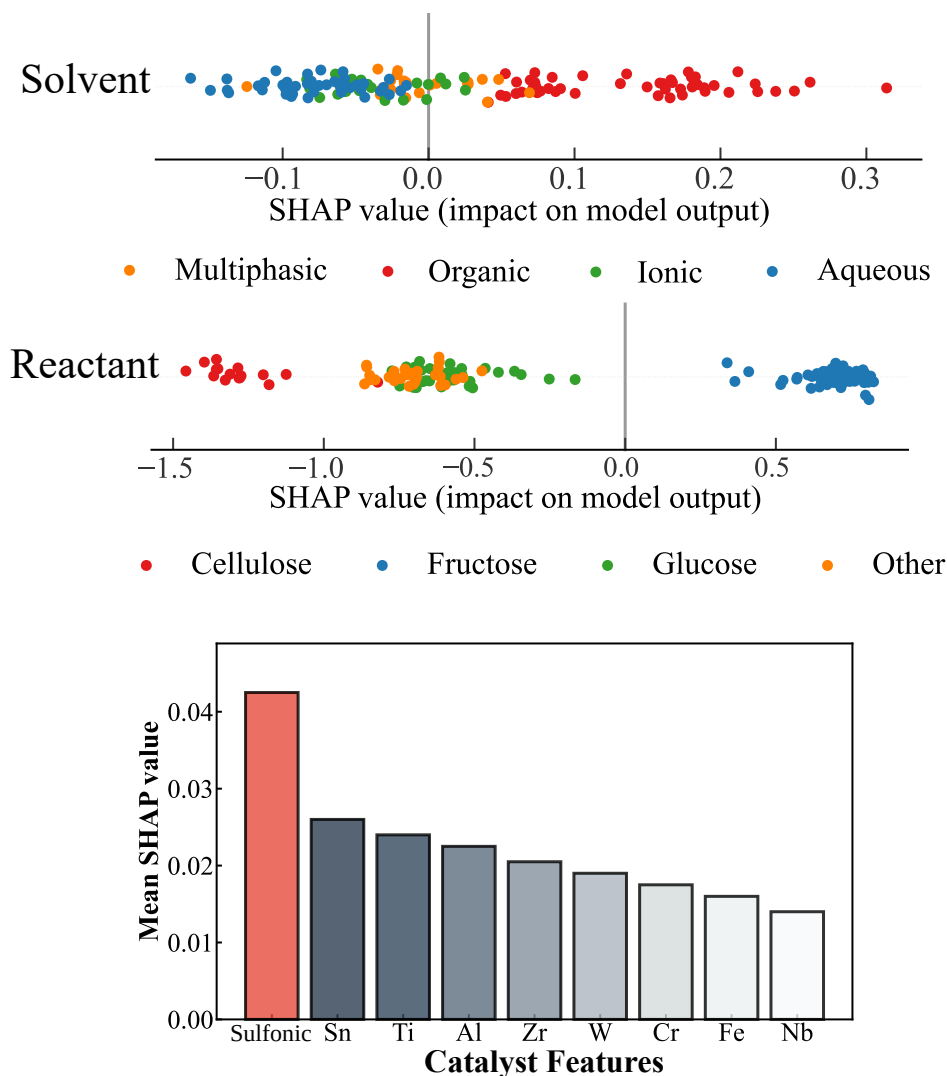


图6 溶剂-反应物-催化剂SHAP分析

Fig.6 SHAP analysis of solvent-reactant-catalyst interactions

XGBoost模型分析结果为以生物质为原料制备HMF提供了清晰的策略指导:优先选择果糖作为原料,在有机溶剂(如DMSO)中进行反应,采用磺酸基功能化催化剂,这一组合在以HMF产率为主要目标

的模型评价中表现出较高潜力。

2.3 工艺参数优化

因此,根据前面的机器学习结果,本研究设计并制备了一种具有多级孔结构的磺酸基聚合物固

体酸,以用于果糖制备HMF反应。热重-差热分析结果发现大量磺酸基的引入使得hSPBCH催化剂的分解更加平稳(见图S1),同时XPS(X-ray Photoelectron Spectroscopy, XPS)能谱和红外光谱证实了hSPBCH催化剂中含有密度极高的磺酸基团(见图S2和S3);氮气吸脱附结果(见图S4)表明, hSPBCH催化剂表面积高达644 m²/g,介孔孔容达到0.321 cm³/g,且存在大量介孔(孔径主要集中在5.3 nm),明显大于果糖和HMF的分子直径,有利于反应物进入孔道并促进产物及时扩散脱出,提高孔道内酸性位点的利用效率。以上表征结果说明,本文制备的hSPBCH催化剂具有大比表面积、多级孔结构和丰富的磺酸基位点等特点。

初步实验发现(见表2),相比于传统的固体酸催化剂(分子筛、酸性树脂等),hSPBCH催化剂更有利于HMF的生成,在150 °C、1 h、0.025 g hSPBCH、0.25 g 果糖条件下, HMF产率为90.2%, TOF(Turnover Frequency, TOF)值达到14.8 h⁻¹,这主要得益于催化剂中磺酸基活性位点(3.45 mmol/g)及多级孔结构(图S4)。

表2 不同催化剂在果糖脱水制HMF反应中的活性^①

Table 2 Activity of different catalysts in the dehydration reaction of fructose to produce HMF^①

Entry	catalyst	HMF 产率 (%)	acidity (mmol/g)	TOF (h ⁻¹) ^⑦
1	hSPBCH	90.2	3.45	14.8
2	hPB	8.2	0.95	5.0
3	hPBH	69.3	2.40	16.8
4	hSPB	72.8	2.85	14.9
5	HY ^②	59.3	2.60 ^⑤	13.3
6	HZSM-5 ^③	51.4	0.39 ^⑤	76.8
7	HBEA ^④	70.8	1.80 ^⑤	22.9
8	Amberlyst-45	49.2	3.00 ^⑤	9.6
9	Amberlyst-15	48.5	4.65 ^⑥	6.1

注: ① Reaction conditions: 150 °C, 0.025 g catalyst, 10 mL DMSO, 1 h, 0.25 g Fructose. ② Si/Al = 2.4; pretreated at 500 °C for 4 h. ③ Si/Al = 36; pretreated at 500 °C for 4 h. ④ Si/Al = 25; pretreated at 500 °C for 4 h. ⑤ From reference^[34]. ⑥ From reference^[35]. ⑦ TOF, defined as (mol of converted fructose)/(mol of the acid sites)/(reaction time).

图7为hSPBCH催化剂的重复使用性能。可以发现, hSPBCH催化剂经过5次循环使用后,果糖转化率由第一次的82.3%逐渐下降至第五次的61.4%, HMF选择性由64.5%下降至55.8%,这可能是因为循环使用过程中催化剂流失和酸性下降导

致的。

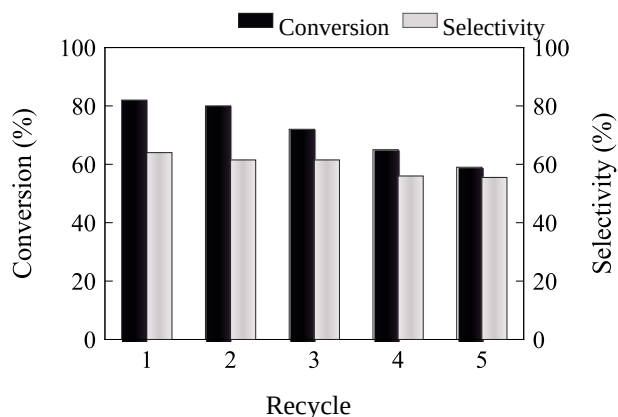


图7 hSPBCH催化剂的重复使用性能

Fig.7 Reusability of hSPBCH catalyst. Reaction conditions: 145 °C, 0.05 g catalyst, 10 mL DMSO, 1.5 h, 1 g Fructose.

接着以果糖为原料, DMSO为溶剂, 通过改变催化剂用量、反应温度及时间等参数构建实验数据集, 采用多层前馈神经网络对HMF产率进行工艺条件的深度优化。图8展示了源域数据(old)和目标域数据(new)在特征空间较为充分的反应规律的t-SNE(t-Distributed Stochastic Neighbor Embedding, t-SNE)降维分布。源域样本分布范围较广, 而目标域样本主要集中在相对局部的区域。说明两类数据有一定的相似性, 通过源域数据为模型提供较为充分的反应规律, 而目标数据用于微调模型, 使其更适应新催化体系下的HMF产率预测。

图9比较了迁移学习模型与直接训练模型对HMF产率的预测效果。从训练集结果来看, 迁移学

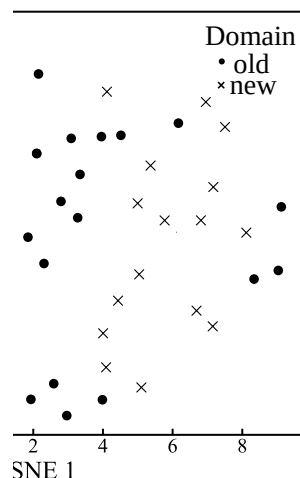


图8 源域与目标域反应条件特征的t-SNE可视化

Fig.8 t-SNE visualization of reaction condition features in source and target Domains

习模型表现出更高的拟合精度,其训练集 $R^2 = 0.9684$ (见图9a),直接训练模型的训练集 $R^2 = 0.9533$ (见图9c)。在测试集的预测中,迁移学习在测试集达到 $R^2 = 0.9661$ (见图9b),直接训练的模型的测试集 $R^2 = 0.8061$ (见图9d),说明迁移学习模型能够利用原域数据中的反应规律,为目标预测提供更充分的先验信息,在独立测试数据上表现出卓越的预测性能(模型参数见SI中表S3)。

为了系统地探索最优反应条件,本研究采用贝叶斯优化算法对反应参数空间进行智能搜索。图10展示了优化过程中四个关键参数(反应温度、反应时间、催化剂用量和果糖浓度)与目标函数(HMF产率)之间的关系。图中散点的颜色梯度反映了优化迭代的时间顺序,浅色代表早期探索阶段,深色代表后期优化阶段。

温度参数展现出最为明显的优化收敛特征(见图10a),高温区域($155\sim 160^\circ\text{C}$)形成了清晰的高产率

聚集带,深色散点几乎完全集中在此范围内;相比之下,低温区域($130\sim 145^\circ\text{C}$)的试验点较为分散,目标值范围跨度大,表明低温条件下反应性能受其他参数影响更为显著,这一结果充分证明了适当提高反应温度可以促进HMF的生成。从反应时间的分布可以看出(见图10b),较短的反应时间内密集分布着高目标值点,目标值主要集中在80~90%范围内,这表明短时间反应内即可得到较高的HMF产率;延长反应至2~3 h后,点分布较为稀疏且值偏低,这可能归因于HMF在长时间、高温条件下可能发生了进一步的降解或缩合反应(生成不溶性腐殖质等副产物)^[36]。催化剂用量的优化结果呈现出显著的浓度依赖性(图10c),低比例催化剂区域明显富集了高产率试验点,深色散点主要集中于此范围,说明贝叶斯算法成功识别出该最优区间。当催化剂比例超过0.1后,HMF的产率逐渐下降并呈现更大的波动。与上述三个因素相比,果糖浓度对HMF的产率影响

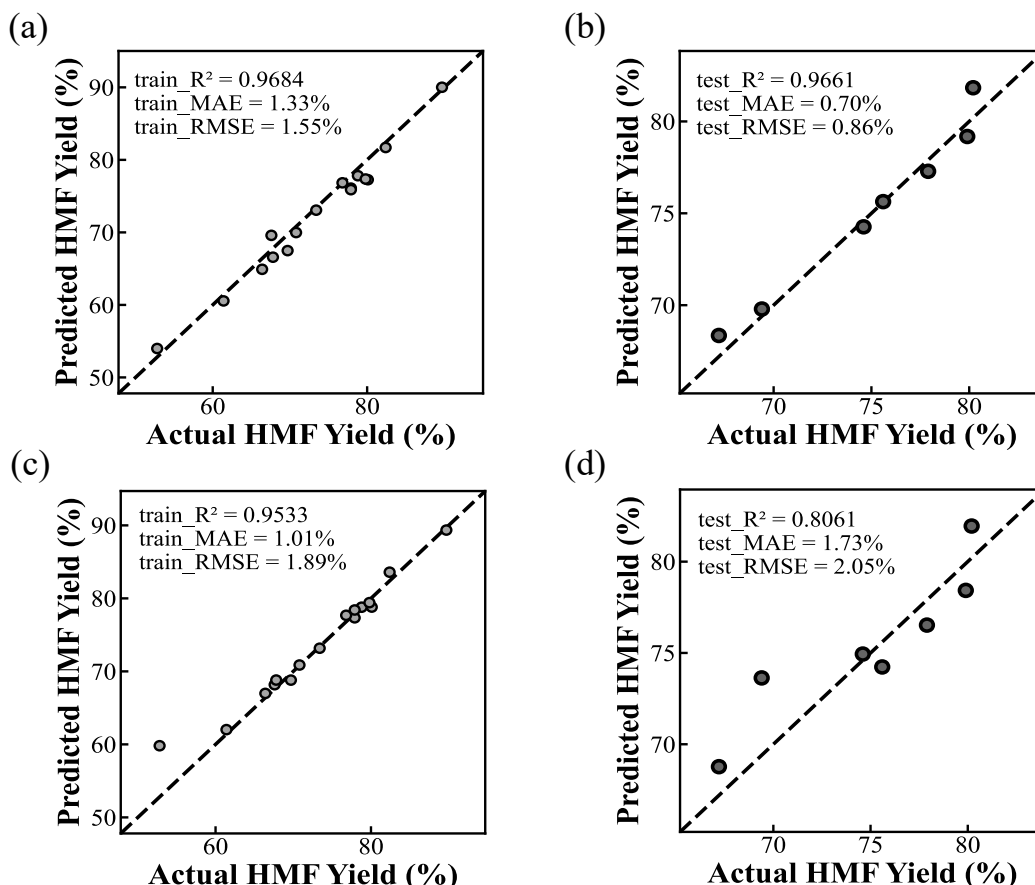


图9 迁移学习模型与直接训练模型的HMF产率预测性能对比: (a) 迁移学习模型训练集, (b) 迁移学习模型测试集, (c) 直接训练模型训练集, (d) 直接训练模型测试集

Fig. 9 Comparison of HMF yield prediction performance between the transfer learning model and the directly trained model: (a) training set of the transfer learning model, (b) test set of the transfer learning model, (c) training set of the directly trained model, and (d) test set of the directly trained model

呈现出复杂的非线性特征(见图 10d),低浓度范围出现了大量高产率点;但是,随着浓度增加至 0.5 g/10 mL 以上,HMF 的产率分布变得更加分散,部分点仍能维持较高产率,而另一部分则显著下降,这种分散性表明高浓度条件下 HMF 的产率对其他工艺参数的协同作用非常敏感。总之,贝叶斯优化算法展

现出优异的自适应学习能力,从颜色梯度的演变可以观察到,早期阶段在整个参数空间进行广泛探索,而后期逐步向高性能区域收敛,最终锁定在高温、短时间、低催化剂用量和适中果糖浓度的参数组合,这可以显著提高参数筛选效率。

图 10 贝叶斯优化过程中反应参数对 HMF 产率的影响: (a)反应温度、(b)反应时间、(c)催化剂和果糖质量比、(d)果糖浓度
Fig.10 Influence of reaction parameters on HMF yield during Bayesian optimization: (a) reaction temperature, (b) reaction time, (c) catalyst-to-fructose mass ratio, (d) fructose concentration

最终,采用机器学习得到的最优反应条件(155 °C、2 h、0.014 g hSPBCH、0.25 g 果糖)下,可以得到的最高 HMF 产率达到 93.6%,实验验证得到的产率为 92.8%(见 SI 中表 S4),高于正交实验所得的 90.2% 产率。如果以催化剂的时空产率为目标进行优化,发现在 120 °C、0.5 h、0.014 g hSPBCH、1 g 果糖的条件下,机器学习优化的时空产率可达 45.6 g-HMF/g-cat/h,高于正交实验所得的 37.1 g-HMF/g-cat/h(见 SI 中表 S5、S6),充分彰显了机器学习的巨大优势。

3 结论

本研究成功构建了一套从文献数据挖掘、机器学习建模到工艺优化与实验验证的完整数据新范式,并成功将该模型用于 HMF 制备中。首先利用大语言模型从 4013 篇文献中自动化提取、清洗和结构化数据,成功构建了一个包含 821 个高质量样本的催化反应数据集。基于该数据集,采用 XGBoost 分类模型和 SHAP 可解释性分析明确指出了实现高 HMF 产率的核心策略:以果糖为原料,采用磺酸基功能化催化剂和有机溶剂。根据机器学习的指导,设计并合成了一种新型多孔有机聚合物固体酸催

化剂 hSPBCH, 在实验中表现出优异的催化活性。结合迁移学习和贝叶斯优化算法, 构建的神经网络模型优化并精准预测了 hSPBCH 催化体系的反应条件。最终, 在预测的最优条件(155 °C, 2 h, 0.25 g 果糖, 0.014 g hSPBCH)下, 实验获得的 HMF 产率高达 92.8%, 不仅与模型预测值(93.6%)高度吻合, 也显著优于传统正交实验优化所得的 90.2%。

总之, 本研究提出的文献数据挖掘-机器学习指导-智能优化-实验验证的方法论, 为加速新型催化材料的发现和化学反应工艺优化提供了强有力的工具。该范式不仅在生物质转化领域展现出了巨大潜力, 也为其他复杂化学反应体系的研究提供可借鉴的思路和方法。

参考文献

- [1] Nasrollahzadeh M, Bidgoli N S S, Shafiei N, et al. Biomass valorization: Sulfated lignin-catalyzed production of 5-hydroxymethylfurfural from fructose[J]. International Journal of Biological Macromolecules, 2021, **182**: 59-64.
- [2] Gan L H, Lyu L, Shen T R, et al. Sulfonated lignin-derived ordered mesoporous carbon with highly selective and recyclable catalysis for the conversion of fructose into 5-hydroxymethylfurfural[J]. Applied Catalysis A: General, 2019, **574**: 132-143.
- [3] Shimizu K I, Uozumi R, Satsuma A. Enhanced production of hydroxymethylfurfural from fructose with solid acid catalysts by simple water removal methods[J]. Catalysis Communications, 2009, **10**(14): 1849-1853.
- [4] Wang S G, Zhang Z H, Liu B. Catalytic conversion of fructose and 5-hydroxymethylfurfural into 2, 5-furandicarboxylic acid over a recyclable Fe_3O_4 - CoO_x magnetite nanocatalyst[J]. ACS Sustainable Chemistry & Engineering, 2015, **3**(3): 406-412.
- [5] Román-Leshkov Y, Chheda J N, Dumesic J A. Phase modifiers promote efficient production of hydroxymethylfurfural from fructose[J]. Science, 2006, **312**(5782): 1933-1937.
- [6] Qu Y S, Huang C P, Zhang J, et al. Efficient dehydration of fructose to 5-hydroxymethylfurfural catalyzed by a recyclable sulfonated organic heteropolyacid salt[J]. Bioresource Technology, 2012, **106**: 170-172.
- [7] Salak Asghari F, Yoshida H. Acid-catalyzed production of 5-hydroxymethyl furfural from d-fructose in subcritical water[J]. Industrial & Engineering Chemistry Research, 2006, **45**(7): 2163-2173.
- [8] Qi X H, Watanabe M, Aida T M, et al. Catalytic conversion of fructose and glucose into 5-hydroxymethylfurfural in hot compressed water by microwave heating[J]. Catalysis Communications, 2008, **9**(13): 2244-2249.
- [9] Teong S P, Yi G S, Zhang Y G. Hydroxymethylfurfural production from bioresources: past, present and future[J]. Green Chemistry, 2014, **16**(4): 2015-2026.
- [10] Wang T F, Nolte M W, Shanks B H. Catalytic dehydration of C6 carbohydrates for the production of hydroxymethylfurfural (HMF) as a versatile platform chemical[J]. Green Chemistry, 2014, **16**(2): 548-572.
- [11] Portillo Perez G, Mukherjee A, Dumont M J. Insights into HMF catalysis[J]. Journal of Industrial and Engineering Chemistry, 2019, **70**: 1-34.
- [12] Li M F, Zhang Q T, Luo B, et al. Lignin-based carbon solid acid catalyst prepared for selectively converting fructose to 5-hydroxymethylfurfural[J]. Industrial Crops and Products, 2020, **145**: 111920.
- [13] Jordan M I, Mitchell T M. Machine learning: Trends, perspectives, and prospects[J]. Science, 2015, **349**(6245): 255-260.
- [14] Butler K T, Davies D W, Cartwright H, et al. Machine learning for molecular and materials science[J]. Nature, 2018, **559**(7715): 547-555.
- [15] Raccuglia P, Elbert K C, Adler P D F, et al. Machine-learning-assisted materials discovery using failed experiments[J]. Nature, 2016, **533**(7601): 73-76.
- [16] Janet J P, Kulik H J. Resolving transition metal chemical space: feature selection for machine learning and structure - property relationships[J]. The Journal of Physical Chemistry A, 2017, **121**(46): 8939-8954.
- [17] Gómez-Bombarelli R, Wei J N, Duvenaud D, et al. Automatic chemical design using a data-driven continuous representation of molecules[J]. ACS Central Science, 2018, **4**(2): 268-276.
- [18] Butler K T, Frost J M, Walsh A. Ferroelectric materials for solar energy conversion: photoferroics revisited[J]. Energy & Environmental Science, 2015, **8**(3): 838-848.
- [19] Chen A, Zhang X, Chen L T, et al. A machine learning model on simple features for CO_2 reduction electrocatalysts[J]. The Journal of Physical Chemistry C, 2020, **124**(41): 22471-22478.
- [20] Segler M H S, Preuss M, Waller M P. Planning chemical syntheses with deep neural networks and symbolic AI[J]. Nature, 2018, **555**(7698): 604-610.
- [21] Coley C W, Green W H, Jensen K F. Machine learning in computer-aided synthesis planning[J]. Accounts of Chemical Research, 2018, **51**(5): 1281-1289.
- [22] Brown T B, Mann B, Ryder N, et al. Language models are few-shot learners[C]//Proceedings of the 34th International Conference on Neural Information Processing Systems. December 6 - 12, 2020, Vancouver, BC, Canada. ACM, 2020: 1877-1901.
- [23] Devlin J, Chang M W, Lee K, et al. BERT: pre-training of deep bidirectional transformers for language understanding[C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Minneapolis, Minnesota Association for Computational Linguistics, 2019: 4171-4186.
- [24] Krallinger M, Rabal O, Lourenço A, et al. Information retrieval and text mining technologies for chemistry[J]. Chemical Reviews, 2017, **117**(12): 7673-7761.
- [25] Swain M C, Cole J M. ChemDataExtractor: a toolkit for automated extraction of chemical information from the scientific literature[J]. Journal of Chemical Information and Modeling, 2016, **56**(10): 1894-1904.
- [26] Zhang J, Das A, Assary R S, et al. A combined experimental and computational study of the mechanism of fructose dehydration to 5-hydroxymethylfurfural in dimethylsulfoxide using Amberlyst

- 70, PO4³⁻/niobic acid, or sulfuric acid catalysts[J]. Applied Catalysis B: Environmental, 2016, **181**: 874–887.
- [27] Baruah J, Medhi U, Nath B K, et al. Enhancements in the production of 5-HMF from glucose: a review on recent advances in heterogeneous catalysts and solvent effects[J]. Next Energy, 2025, **9**: 100420.
- [28] Liu T, Niu Y M, Cheng K X, et al. Exploring the formation pathway and antioxidant properties of the sugar-smoking pigment 5-GGMF[J]. Food Chemistry, 2024, **442**: 138406.
- [29] Marianou A A, Michailof C M, Pineda A, et al. Effect of Lewis and Brønsted acidity on glucose conversion to 5-HMF and lactic acid in aqueous and organic media[J]. Applied Catalysis A: General, 2018, **555**: 75–87.
- [30] Zhao Y, Lu K F, Xu H, et al. A critical review of recent advances in the production of furfural and 5-hydroxymethylfurfural from lignocellulosic biomass through homogeneous catalytic hydrothermal conversion[J]. Renewable and Sustainable Energy Reviews, 2021, **139**: 110706.
- [31] 董喜恩, 蒋锋, 罗根祥. 糖类脱水生成5-羟甲基糠醛的研究进展[J]. 化工中间体, 2010, **6**(2): 17–22.
- Dong X E, Jiang F, Luo G X. Dehydration of Carbohydrates into 5-Hydroxymethylfurfural in the presence of double micro-porous zeolite catalyst[J]. Chemical Intermediate, 2010, **6**(2): 17–22.
- [32] Ren Q H, Ma H, Wang W B, et al. Catalytic dehydration of fructose to 5-hydroxymethylfurfural over mesoporous Nb–W oxide solid acid catalyst[J]. Biomass and Bioenergy, 2023, **174**: 106860.
- [33] Shi H, Ramasamy K K, Ma R S, et al. Nanoporous catalysts for biomass conversion[M]//Nanoporous Materials for Molecule Separation and Conversion. Amsterdam: Elsevier, 2020: 387–440.
- [34] Jiang Y Y, Zhou R R, Ye B Y, et al. Acetalization of glycerol over sulfated UiO-66 under mild condition[J]. Journal of Industrial and Engineering Chemistry, 2022, **110**: 357–366.
- [35] Sonsiam C, Kaewchada A, Pumrod S, et al. Synthesis of 5-hydroxymethylfurfural (5-HMF) from fructose over cation exchange resin in a continuous flow reactor[J]. Chemical Engineering and Processing – Process Intensification, 2019, **138**: 65–72.
- [36] Shen Y, Sun J K, Wang B, et al. Catalytic approaches to the production of furfural and levulinates from lignocelluloses[M]//Biomass as Renewable Raw Material to Obtain Bioproducts of High-Tech Value. Amsterdam: Elsevier, 2018: 235–269.