Science Research
2026, Vol. 14, No. 6, pp. 457–463
https://doi.org/10.11648/j.sr.20261406.20

SciencePG
Science Publishing Group

Research Article

Interpretable Prediction of CO_2 Uptake Using Multi-Source Steel Slag Data and Transfer Learning Across Feedstocks

Xin Wei, Yi Zhang, Lingfeng Xia, Mengjiao He, Jingyi Meng*

Zhong Cun·Big·Data Technology Company Limited,·Beijing, China

Abstract

CO_2 uptake through the carbonation of industrial solid wastes represents an important technological pathway toward carbon neutrality, with performance jointly influenced by feedstock composition and process conditions. However, differences in data sources and feedstock characteristics complicate reliable prediction and limit model applicability across materials. This study integrated 818 CO_2 uptake records covering steel slag, magnesium slag, and recycled fine powder to establish a random forest framework incorporating grouped cross-validation, nested feature selection, and cross-feedstock transfer comparisons. Steel-slag models trained on literature data, compiled experimental data, and their combination were compared using common input features and consistent test groups. Connected groups based on feedstock composition and known literature sources were used to reduce information leakage between training and test sets. Joint training using steel-slag literature and experimental data achieved a mean absolute error (MAE) of 6.198 and a root mean squared error (RMSE) of 8.515. Permutation importance analysis and nested feature selection identified particle size, carbonation time, and reaction temperature as consistently retained predictors, although the optimal feature combination varied across validation folds. Feature-space comparisons revealed differences in feedstock composition and process conditions, motivating the evaluation of multiple transfer strategies using eight common features. With literature-derived steel-slag data as the source domain, joint training for magnesium slag and direct transfer for recycled fine powder reduced group-weighted MAE, with equal weight assigned to each group, by 26.7% and 32.3%, respectively, relative to random forest models trained solely on target-domain data. These findings demonstrate that steel-slag data can support CO_2 uptake prediction for other solid wastes, while transfer benefits depend on the target feedstock and training strategy. The proposed framework provides a methodological basis for evaluating multisource data integration, identifying predictive features, and assessing model transferability across solid-waste feedstocks.

Keywords

Steel Slag, CO_2 Uptake, Interpretable Machine Learning, Transfer Learning Across Feedstocks

多源钢渣固碳数据的可解释预测及跨原料迁移研究

魏鑫，张翼，夏凌风，贺梦蛟，蒙景怡*

中存大数据科技有限公司，北京，中国

*Correspondence: Meng Jingyi (mengjingyi@ctc.ac.cn)

Received: 2 October 2026; **Accepted:** 26 October 2026; **Published:** 10 November 2026

Copyright: © The Author(s), 2026. Published by Science Publishing Group. This is an **Open Access** article, distributed under the terms of the Creative Commons Attribution 4.0 License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution and reproduction in any medium, provided the original work is properly cited.

摘要

工业固废固碳是助力工业领域碳中和的重要技术途径，其固碳性能受原料组成与工艺条件共同影响。然而，模型对不同原料的适用性尚不明确，限制了其跨原料应用。本研究整合 818 条固碳数据，构建随机森林预测模型，并进行分组验证、嵌套特征筛选和跨原料迁移对比。结果表明，钢渣文献与实验汇总数据联合训练的 MAE 和 RMSE 分别为 6.198 和 8.515。粒径、碳化时间和反应温度在各轮筛选表现出预测价值。以文献钢渣为源域时，镁渣联合训练与再生微粉直接迁移分别取得较好的预测表现，表明迁移效果随原料与策略而异。本研究为固废固碳率预测及跨原料模型迁移提供了方法参考。

关键词

钢渣，CO_2 矿化固碳，可解释机器学习，跨原料迁移学习

1. 引言

在“双碳”战略全面推进的背景下，众多高耗能行业整着力推动低碳技术的创新与应用。在钢铁冶炼、水泥生产及供热领域，相关研究致力于将工业固废用于制备绿色建材、人工骨料以及高性能透水砖[1-4]。钢渣、镁渣、再生微粉均属于富含钙镁组分的碱性工业固废，均可通过矿物碳化实现CO_2封存，同时制备低碳建材[5-7]。然而，固碳性能难以仅依据单一组成指标或工艺参数进行判断，需要综合考虑多种因素及其关联，以建立不同实验条件的预测模型。

机器学习为固碳性能的多因素预测提供了新的研究方法。He等收集318组钢渣固碳数据，以12项材料与工艺参数为输入，建立多层感知机、随机森林及支持向量回归模型，并结合SHAP分析不同特征对预测结果的贡献，展示了机器学习在钢渣固碳预测与影响因素分析中的应用潜力[8]。Guo等进一步基于613条钢渣数据开展多模型比较与参数优化，并将固碳预测与经济可行性分析相结合，拓展了数据驱动方法在钢渣碳化评价中的应用[9]。在上述研究基础上，仍有必要进一步研究：当模型输入受不同数据来源的共同特征限制时，其预测能力如何变化；补充实验数据能否改善模型对未参与训练的数据组的预测；以及钢渣数据所提供的信息能否用于其他固废原料。

除同类原料的预测外，跨原料应用也是固碳模型值得研究的方向。迁移学习能够利用相关源域的信息辅助目标域建模，但其收益取决于源域与目标域之间的关联程度；当源域信息与目标任务不匹配时，引入源域数据也可能降低预测性能，形成负迁移[10, 11]。在开展迁移之前，有必要比较不同原料的组成分布、工艺范围及特征空间重叠情况，并以仅使用目标域数据建立的模型为基线，在相同测试条件下判断源域信息是否产生实际收益[12]。

基于此，本研究整合钢渣、镁渣及再生微粉的818条固碳数据，构建结合分组验证、可解释性分析、嵌套特征筛选与跨原料迁移比较的随机森林预测框架。首先，以钢渣为研究对象，比较完整特征与共同特征条件下的模型表现，并在一致测试组条件下对比文献数据、实验汇总数据及二者联合训练的预测效果；依据原料组成与已知文献来源进行分组，降低关联记录跨训练集与测试集分布的风险。其次，通过置换重要性与嵌套筛选分析特征的预测贡献及稳定性。最后，对比不同原料的特征空间，在固定8个共同特征的条件下，比较直接迁移、预测融合、联合训练及相似度加权策略，并采用MAE、RMSE及组等权MAE评估预测性能。本研究旨在明确多源数据融合的作用及钢渣模型的跨原料适用性，为固废固碳率预测与模型迁移提供方法依据。

2. 研究方法

2.1. 数据来源和预处理

本研究整合文献数据[8]与实验数据，经来源核对与去重，保留818个样本，文献与实验数据分别为317个和501个，占38.7%和61.3%。其中，钢渣样本量551个、镁渣样本量173个、再生微粉样本量94个。以CaO含量、MgO含量、粒径、CO_2浓度、反应时间、液固比、反应温度和CO_2分压共8个作为共同输入特征，以固碳率为预测目标，文献钢渣数据另保留SiO_2、AlO_3、Fe_2O_3和MnO含量，用于分析额外成分信息对性能预测的影响。

进行标准化处理，对粒径、反应时间、液固比和CO_2分压进行对数变换，缺失值采用训练折中位数插补。依据原料组成及已知文献来源进行分组，以组为单位进行交叉验证，避免同一组样本同时进入训练集与测试集。

2.2. 模型构建

本研究采用随机森林（random forest, RF）[13]构建钢渣相关体系的固碳预测模型。固定决策树数量设为100，叶节点最小样本数设为3，节点分裂考虑全部输入，随机种子固定。设置两组对照：在相同文献数据划分下比较12与8特征模型；对8个共同特征，分别采用文献、实验及二者合并数据训练，在相同实验测试集上评价。

同一原料类别内，将CaO和MgO含量相同或来自同一文献的样本作为一组，通过该关系的传递形成连通分组，并为每个样本赋予组标识。采用四折分组交叉验证（GroupKFold）[14]，依次留出一折进行测试，保证同组样本不跨训练与测试集。实验数据评价进一步剔除文献训练数据中与实验测试是同一组的记录。

特征数量及迁移融合权重通过外层训练数据内部三折分组交叉验证得到，以组等权平均绝对误差作为选择依据。外层各折预测完成后，将所有留出记录的预测结果汇总为折外预测，用于计算总体性能指标。模型性能采用平均绝对误差（MAE）和均方根误差（RMSE）评价，计算式如下：

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}$$

式中，N为折外评价记录数，y_i和$\hat{y}_i$分别为固碳性能的实测值和预测值。由于各组包含的样本数量不同，进一步计算组等权MAE，以反映模型在不同组上的平均预测误差：

$$MAE_{group} = \frac{1}{G}\sum_{g=1}^{G}(\frac{1}{n_g}\sum_{i\in g}|y_i - \hat{y}_i|)$$

式中，G为评价组数，n_g为第g组的记录数。误差指标均以固碳率的百分点表示。

2.3. 可解释性分析

为识别原料组成及工艺参数对模型预测的影响，本研究采用置换重要性分析评价各输入特征的预测贡献[15]。针对包含12个特征的文献钢渣模型，随机打乱待评价特征的取值，以置换前后MAE的变化评价该特征的重要性。此外，为评价额外成分信息的作用，在12个特征的模型基础上，分别移除SiO_2、AlO_3、Fe_2O_3和MnO含量，在相同的数据划分和参数下评价单特征消融结果。

在每个外层训练集中，通过三折分组验证比较特征筛选方案：先根据内层训练数据的置换重要性对特征排序，再分别保留前4、6和8个特征建立RF模型，以内层验证的组等权MAE选择特征数量。随后，在完整外层训练集上重新排序、筛选并训练模型，在外层测试集上得到各折入选特征及数量。

2.4. 跨原料迁移策略

基于上述RF模型的预测结果，进一步探究钢渣固碳数据对其他原料预测任务可迁移性，以钢渣数据为源域，镁渣和再生微粉分别作为目标域。设置两种源域方案：一是仅使用文献钢渣数据，二是合并文献与已有实验中的钢渣数据。各方案均采用8个共同特征及统一RF参数，分别开展直接预测、预测融合、联合训练和相似度加权对照。

在迁移建模前，采用经验累积分布、目标域观测值超出钢渣源域取值范围的比例，描述不同原料的特征分布差异。进一步开展主成分分析（PCA）[16]。标准化参数和PCA投影均利用钢渣数据拟合，再将其他原料映射至同一空间。

直接预测将源域数据训练模型直接应用于目标域。预测融合分别构建源域模型与目标域模型，并对二者的预测结果进行加权组合。候选权重通过目标训练数据内部的三折分组交叉验证，以组等权MAE确定，再用完整目标训练集建立模型。联合训练将源域数据与目标训练数据合并，重新训练RF模型，源域和目标训练域各50%的总样本权重。

相似度加权根据源域样本与目标训练数据的特征距离调整源域样本的训练权重。距离计算使用CaO含量、MgO含量、CO_2浓度、反应时间和反应温度5个描述符。计算每个源域样本至目标训练集中最近样本的距离d_i，进一步采用带下限的高斯核构建相似度权重：

$$s_i = 0.1 + 0.9\exp(-\frac{d_i^2}{2h^2})$$

式中，带宽h取源域样本到目标域训练样本的非零距离中位数，没有时取1。描述符不完整的源域样本赋予0.1的相似度权重。相似度权重与基础样本权重相乘后，在各来源内重新归一化，并据此重新训练源域RF模型。

3. 结果与讨论

3.1. 模型训练与性能评估

基于预处理后的数据，对钢渣文献数据中不同特征组合进行比较，并在相同的实验测试组上对比文献训练、实验训练和二者联合训练的效果。各模型均采用相同参数的RF模型，减少算法差异对结果解释的干扰。采用MAE和RMSE评价样本整体预测误差，同时计算组等权MAE，比较不同组之间的平均预测结果。不同特征组合的验证结果见表1。

表 1 钢渣文献数据不同特征组合的预测性能。

特征组合	MAE	RMSE	组等权 MAE
12 个特征	8.483	11.243	9.290
8 个共同特征	9.196	11.958	10.413
共同特征内嵌套筛选模型	8.833	11.567	10.106

基于12个特征建立的RF模型MAE和RMSE均低于8个共同特征模型，说明文献数据中额外提供的化学组成信息具有一定预测价值。将输入限制为8个共同特征后，MAE由8.483增加至9.196，RMSE由11.243增加至11.958。结果说明，统一不同来源的数据提高了模型的适用范围，但可能伴随有效信息损失。

在8个共同特征内开展嵌套筛选后，MAE降低至8.833，组等权MAE由10.413降低至10.106。结合不同外层折所选特征数量的变化，这一结果尚不支持确定一个适用于全部数据来源的唯一最优特征组合。

进一步以实验汇总中的 234 条钢渣数据作为共同评价对象，对应 23 个组，比较仅使用文献数据、文献与实验数据联合，以及仅使用实验数据三种训练方案。各方案使用 8 个共同特征作为模型输入，对比在相同钢渣实验测试集上的预测性能（表 2）。

表 2 不同训练方案的预测性能。

训练集	MAE	RMSE	组等权 MAE
文献数据	7.079	9.327	6.822
文献与实验数据	6.198	8.515	6.752
实验数据	6.890	9.361	8.495

联合训练获得了三种训练方案中最低的MAE和RMSE，其MAE相对于仅文献数据、仅实验数据分别降低约12.4%和10.0%。结果表明，在共同特征和相同测试组的条件下，两类数据提供了具有互补性的预测信息。比较组等权MAE，联合训练相对于仅文献训练的改善较小，差值为0.070，相对于仅实验训练，差值为1.743。

此外，预测值和实测值的对比显示（图1），高固碳率存在低估现象。对实测固碳率不低于40%的8条记录进行检查，三种训练方案均表现出低估，相比联合训练有所缓解，但也未能消除这一偏差。这说明模型在高固碳率条件下的预测能力仍需补充数据验证。

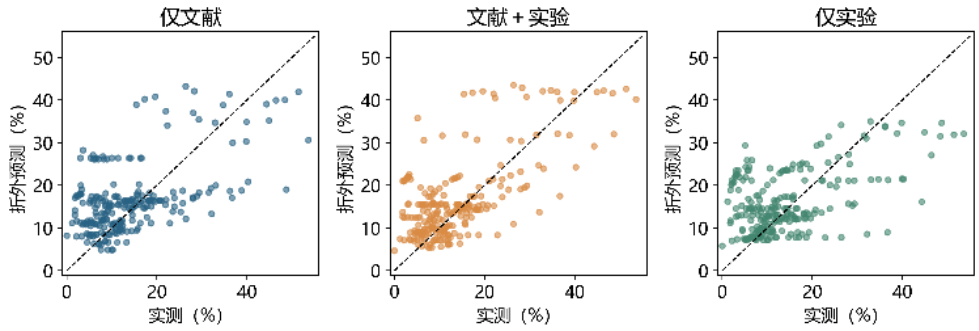


图 1 不同训练集钢渣固碳率的预测结果。

3.2. 特征空间分析

进一步，本研究从特征贡献及跨原料分布差异两个方面开展分析。在钢渣文献的12特征模型中，折外置换重要性表明，Al_2O_3、反应温度、碳化时间、粒径和液固比提供了相对较多的预测信息。其中，打乱Al_2O_3后，MAE平均增加0.924个百分点，跨折标准差为1.411个百分点；反应温度、碳化时间、粒径和液固比对应的平均MAE增量分别为0.533、0.315、0.306和0.282个百分点。删除Al_2O_3后，模型MAE由8.483增加至9.177，与置换分析得到的结果一致。

结果表明，Al_2O_3在当前文献数据中具有较高的预测贡献，但其贡献随验证折变化较大。同样，部分特征的置换重要性接近0或出现负值，说明模型在当前数据和其他输入条件下未稳定利用该变量。

共同特征的嵌套筛选结果进一步显示，四次外层验证中分别保留8、6、4和8个特征，其中粒径、碳化时间和反应温度均被保留。这说明上述变量在不同训练子集中具有较一致的预测价值。考虑到特征筛选仅带来有限改善，且不存在完全一致的精简组合，后续跨原料对比仍采用统一的8个共同特征，以保持模型输入的一致性。

以合并钢渣数据的取值范围作为参照，镁渣在部分变量上与源域存在重叠，但仍有一定比例的记录超出源域范

围。其中，7.5%的CaO记录高于钢渣源域上限，18.5%的粒径记录低于源域下限，27.2%的液固比记录高于源域上限。这表明钢渣至镁渣的迁移包含一定程度的范围外预测，不能仅依据原料均含有Ca、Mg组分就判断其完全处于相同特征空间。

再生微粉的组成差异更为突出。在86条具有CaO和MgO记录的样本中，64.0%的CaO取值高于钢渣源域上限，88.4%的MgO取值低于源域下限。结合其粒径和CO_2分压的大量缺失，可以看出，再生微粉同时存在组成分布偏移和工艺信息不足的问题。

为降低大量插补对空间分布展示的影响，主成分分析仅采用CaO、MgO、CO_2浓度、碳化时间和反应温度五个共同完整字段，并在钢渣数据上确定标准化规则并建立PCA变换。第一、第二主成分分别解释约27.0%和24.8%的钢渣特征方差，累计解释率为51.8%。

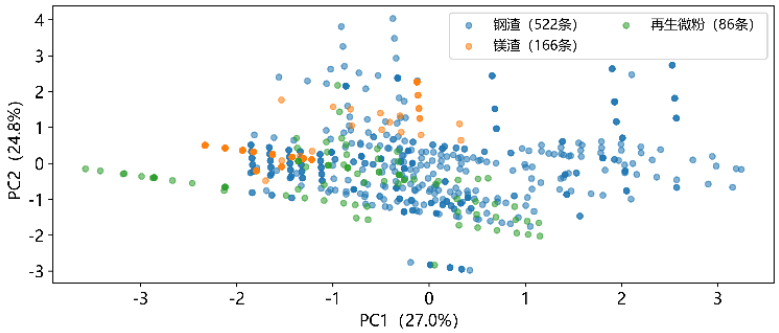


图 2 主成分分析。

主成分结果显示，镁渣和再生微粉与钢渣分布均存在局部重叠，同时也分布于钢渣样本相对稀疏的区域。但不足以直接判定迁移是否有效。一方面，二维投影仅保留约一半的源域特征方差；另一方面，该分析未纳入粒径、液固比及CO_2分压，也未包含矿物相等信息。

3.3. 跨原料迁移效果

基于上述特征空间差异，分别采用钢渣的文献数据和合并数据作为源域，对镁渣和再生微粉比较直接迁移、预测加权适配、源域目标域联合训练及相似度加权四种策略。仅使用目标域训练数据的随机森林作为基线。不同策略使用一致的目标域外层测试组，目标域测试集不参与模型拟合或适配参数选择。

镁渣任务中，文献钢渣与目标域联合训练取得最低的组等权MAE，为6.706，相对于仅目标域随机森林降低约26.7%。合并钢渣数据联合训练也优于目标域随机森林。说明钢渣源域数据在镁渣任务中具有可利用的预测信息。

不同评价指标对应的最优方案并不一致。文献数据相似度加权模型取得最低MAE，合并数据联合训练取得最低RMSE，而文献数据联合训练的组等权MAE最低。这种差异与目标域样本分布不均有关：镁渣两个最大的组包含159条记录，占总记录数的91.9%，另外还有多个样本量很小的组。样本等权MAE和RMSE更易受到大组预测表现的影响，组等权MAE则给予各组相同权重。因此，迁移策略的比较需要同时关注总体记录误差和跨组表现。

相似度加权的效果也随源域构成而变化。对文献数据进行相似度加权后，相对于直接迁移的组等权MAE减小了0.884；而对合并数据实施相同策略后，误差反而增加了0.516。

表 3 不同迁移策略的预测性能。

数据来源与策略	镁渣			再生微粉		
	MAE	RMSE	组等权 MAE	MAE	RMSE	组等权 MAE
仅目标域随机森林	10.415	11.990	9.153	17.274	18.925	16.437
文献数据：直接迁移	7.016	9.193	8.908	12.138	13.446	11.129
文献数据：预测加权适配	9.114	10.849	8.756	13.622	15.799	13.084

数据来源与策略	镁渣			再生微粉		
	MAE	RMSE	组等权 MAE	MAE	RMSE	组等权 MAE
文献数据：联合训练	6.860	8.376	6.706	12.786	15.174	14.101
文献数据：相似度加权	6.266	8.077	8.024	12.504	13.765	11.484
合并数据：直接迁移	6.277	8.229	8.575	12.519	13.895	11.401
合并数据：预测加权适配	6.878	8.564	8.752	13.890	16.004	13.237
合并数据：联合训练	6.539	7.734	7.213	14.280	16.269	13.845
合并数据：相似度加权	6.395	8.316	9.091	11.834	14.085	11.413

再生微粉任务中，文献数据直接迁移取得最低的RMSE和组等权MAE。其组等权MAE相对于仅目标域随机森林降低5.309。引入目标域训练信息并未进一步改善再生微粉预测。文献数据预测加权适配的组等权MAE增加至13.084，相对于直接迁移的减少了1.955。合并数据联合训练相对于直接迁移的组等权MAE增加2.444。因此，这些适配方案虽然仍可能优于仅目标域随机森林，但相对于直接迁移出现了性能下降。

结合特征空间分析，再生微粉较明显的组成偏移及关键工艺字段缺失，可能限制了目标域训练数据对模型的有效修正。合并数据相似度加权虽然取得最低的MAE，但其组等权MAE未优于文献数据直接迁移。因此，现阶段不能认定增加源域记录或采用更复杂的适配方式能够稳定提高再生微粉的预测性能。

4. 结论

本研究基于钢渣文献与实验数据，构建了结合分组验证、特征筛选和跨原料迁移的随机森林固碳率预测框架。在共同特征和一致测试组条件下，联合训练的MAE和RMSE分别为6.198和8.515。粒径、碳化时间和反应温度在各轮筛选中均被保留，表现出较一致的预测价值，而不同轮次的最优特征组合存在差异。跨原料分析显示，不同目标域在组成分布、工艺范围方面存在差异，迁移效果也随原料和训练策略变化。本研究不仅为钢渣固碳率预测与关键特征识别提供了有效方法，也为跨原料预测模型迁移及固废碳化工艺优化提供了方法参考。

基金项目

本文为国家重点研发计划项目《固废高效碳捕集及固碳建材高性能制备关键技术研究与应用》(2024YFE0210400)的阶段性成果之一。

参考文献

[1] MA Y, LUO Y, MA H, et al. Upcycling steel slag in producing eco-efficient iron-calcium phosphate cement [J]. Journal of Cleaner Production, 2022, 371: 133688. https://doi.org/10.1016/j.jclepro.2022.133688

[2] SONG Q, GUO M Z, WANG L, et al. Use of steel slag as sustainable construction materials: A review of accelerated carbonation treatment [J]. Resources, Conservation and Recycling, 2021, 173: 105740. https://doi.org/10.1016/j.resconrec.2021.105740

[3] YANG Y, XU W, WANG Y, et al. Progress of CCUS technology in the iron and steel industry and the suggestion of the integrated application schemes for China [J]. Chemical Engineering Journal, 2022, 450: 138438. https://doi.org/10.1016/j.cej.2022.138438

[4] JIANG Y, LING T C. Production of artificial aggregates from steel-making slag: Influences of accelerated carbonation during granulation and/or post-curing [J]. Journal of CO_2 Utilization, 2020, 36: 135-144. https://doi.org/10.1016/j.jcou.2019.11.009

[5] HUIJGEN W J J, WITKAMP G J, COMANS R N J. Mineral CO2 sequestration by steel slag carbonation [J]. Environmental Science & Technology, 2005, 39(24): 9676-9682. https://doi.org/10.1021/es050795f

[6] 刘松辉, 张程, 管学茂. CO_2 协同镁渣基固废制备纤维板及其性能与微结构 [J]. 硅酸盐学报, 2023, 51(9): 2166-2178. https://doi.org/10.14062/j.issn.0454-5648.20230107

[7] 吴扬, 周鹏程, 仇建国, 等. 基于加速固碳活化的再生微粉水泥稳定淤泥性能研究 [J]. 土木工程与绿色建筑, 2025, 1(2): 96-99.

[8] HE B, ZHU X, CANG Z, et al. Interpretation and prediction of the CO_2 sequestration of steel slag by machine learning [J]. Environmental Science & Technology, 2023, 57(46): 17940-17949. https://doi.org/10.1021/acs.est.2c06133

[9] GUO H, JIAO W, ZHANG M, et al. Prediction of CO_2 sequestration rate of steel slag and economic feasibility analysis [J/OL]. Journal of Environmental Sciences, 2026: 105712. https://doi.org/10.1016/j.jes.2026.105712

[10] DEVI R, BUTLER K T, SAI GAUTAM G. Leveraging transfer learning for accurate estimation of ionic migration barriers in solids [J]. npj Computational Materials, 2026, 12: 88. https://doi.org/10.1038/s41524-026-01972-8

[11] PAN S J, YANG Q. A survey on transfer learning [J]. IEEE Transactions on Knowledge and Data Engineering, 2010, 22(10): 1345-1359. https://doi.org/10.1109/TKDE.2009.191

[12] WANG Z, DAI Z, PÓCZOS B, et al. Characterizing and avoiding negative transfer [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 11293-11302. https://doi.org/10.1109/CVPR.2019.01155

[13] BREIMAN L. Random forests [J]. Machine Learning, 2001, 45(1): 5-32. https://doi.org/10.1023/A:1010933404324

[14] AGUIAR L G, KASEMODEL M C. Application of random forest regression in modeling the adsorption of methylene blue onto clays [J]. Neural Computing and Applications, 2026, 38: 496. https://doi.org/10.1007/s00521-026-12200-1

[15] FISHER A, RUDIN C, DOMINICI F. All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously [J]. Journal of Machine Learning Research, 2019, 20(177): 1-81. https://doi.org/10.48550/arXiv.1801.01489

[16] JOLLIFFE I T, CADIMA J. Principal component analysis: A review and recent developments [J]. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 2016, 374(2065): 20150202. https://doi.org/10.1098/rsta.2015.0202