

低碳熟料强度预测的数据增强建模及模型可解释性分析

翟牧楠¹, 郅晓¹, 叶家元¹, 任雪红¹, 吴春丽², 张洪滔¹, 崔文娟¹, 颜景华¹, 张文生¹

(1. 中国建筑材料科学研究总院有限公司, 北京 100024; 2. 中存大数据科技有限公司, 北京 100024)

摘 要: 针对低碳模拟熟料体系样本量小、组成空间稀疏, 导致代理模型跨体系预测能力不足的问题, 本工作提出一种自迭代数据增强框架, 用于优化其 28 d 抗压强度预测模型。该框架在组成约束条件下生成虚拟组成, 并由上一轮模型赋予伪标签以逐轮自训练。采用测试集全程隔离、验证集仅包含实验数据的锚定评估策略, 抑制由伪标签引入造成的误差传递。结果表明: 迭代增强后的代理模型测试误差降低, 跨体系泛化能力提升。在第 2 轮迭代中, 以模式 B 采样训练得到的代理模型在体系内拟合与体系外泛化之间取得较优平衡, 其总体均方误差(MSE)为 9.8, 体系内、外 MSE 分别为 6.9、11.3, 相关系数 R^2 为 0.79, 平均绝对误差与平均绝对百分比误差分别为 2.36 与 5.11%。在此基础上, 进一步结合受约束的置换特征重要性、累积局部效应与沙普利加性解释开展分析, 识别关键矿相的全局敏感性、平均响应关系与样本级贡献, 为模型可靠性审查、关键矿相筛选及后续模拟熟料组成优化提供可检验的参考依据。

关键词: 低碳熟料; 28 d 抗压强度; 预测模型; 自迭代; 数据增强; 可解释性

中图分类号: TQ172.11 文献标志码: A 文章编号: 0454-5648(2026)08-2627-17

网络出版时间: 2026-07-14



新材料研发普遍面临小样本约束的难题^[1]。高通量计算与实验(制备/表征)可在一定程度上缓解数据短缺的情况, 尤为适用于既有体系内的组成搜索与优化^[2, 3]。然而, 其有效实施依赖流程标准化、样本可扩展性与数据自洽性。对水泥与混凝土而言, 制样与测试流程的标准化程度较高, 高通量实验具有较强可操作性, 但针对水泥熟料体系, 由于其煅烧过程涉及高温固-液反应与多相生成, 实现多元矿相组成的可控合成与快速扩展仍面临挑战^[4]。此外, 熟料性能并排取决于单一矿相的活性, 而是由矿相间协同水化等多过程耦合影响, 因而导致矿相组成与强度、耐久性等宏观性能之间呈现强非线性映射关系。上述因素共同导致实验迭代难以覆盖足够的组成空间, 从而限制了对潜在新型低碳熟料的系统挖掘。

机器学习为解决该问题提供了一条新的技术路径, 通过构建代理模型可学习材料组成及其性能间的映射关系, 将探索模式由逐点试配转变为模型引导的筛选与搜索, 并进一步服务于面向目标的组

成设计^[5-7]。在迭代成本可控的场景下, 主动学习等闭环策略通过模型推荐候选组分、实验反馈更新模型的机制, 显著提高数据利用效率并逐步逼近最优解目标^[8, 9]。然而, 熟料体系的合成-养护-测试链条周期长、一致性控制难度大, 限制了闭环迭代的频率, 导致新型熟料组成数据难以快速累积并外扩覆盖。现有公开研究与成熟体系的累积数据往往集中于少数既有体系^[10, 11], 采用基因表达式编程、多表达式编程等算法可实现这些单一体系的性能拟合与优化, 尤其在未知目标表达式复杂度时, 这类算法在搜索组成关系和构建力学性能和耐久性能表达式方面具有一定优势^[12-14]。然而, 这类研究通常仍难以支撑对新体系或新矿相组合的系统性外推预测。因此, 挖掘新型低碳水泥熟料的瓶颈并不在于算法缺失, 而在于缺乏能够有效扩展材料组成、支撑可靠外推的训练信息。

在此背景下, 以有限的既有实验数据为基础, 通过扩展训练信息提升代理模型的可用性, 是更现实的改进路径。数据增强可在满足组成约束的前

收稿日期: 2026-03-09。 修订日期: 2026-04-11。

基金项目: “十四五”国家重点研发计划(2022YFC3803101); 国家自然科学基金双碳专项项目(52341202)。

第一作者: 翟牧楠(1994—), 男, 博士研究生。

通信作者: 叶家元(1980—), 男, 博士, 教授。

Received date: 2026-03-09. Revised date: 2026-04-11.

First author: ZHAI Munan (1994—), male, Doctoral candidate.

E-mail: zmunann@163.com

Correspondent author: YE Jiayuan (1980—), male, Ph. D., Professor.

E-mail: yejiayuan@cbma.com.cn

提下生成额外候选组成点,通过构造新样本来扩大训练分布的覆盖范围,并进一步通过半监督学习中的伪标签机制将其转化为训练样本增量,从而为样本稀疏区域补充监督信号^[15,16]。已有研究表明,在小样本场景下,合理的数据增强、对抗神经网络、自训练等策略可显著改善模型的稳健性与泛化表现^[17–19]。但必须关注的是,伪标签噪声可能随迭代累积而放大模型误差,引入新的可靠性风险^[20]。因此,增强策略需配合严格的评估机制,以规避由伪标签引入导致的评估偏移^[21,22]。与此同时,熟料体系面向工程应用,仅有预测精度不足以支撑组成决策,需揭示模型如何利用各矿相信息,识别其对抗压强度预测的主要驱动因素与响应趋势^[23–25]。因此,还应将可解释性分析作为增强建模可靠性的必要补充。

本工作提出一种适用于样本规模有限、组成空间稀疏的小样本模拟熟料体系自迭代数据增强框架,旨在扩展有效训练信息并提升代理模型的泛化稳定性。同时,为了尽量减少模型差异所造成的分析误差,本工作选用人工神经网络(ANN)作为统一代理模型开展后续模型迭代与评估、解释性分析等研究工作。该自迭代数据增强框架在组成约束下生成候选组成点,并通过逐轮自训练将其转化为训练样本增量。在评价方法上,本工作采用锚定策略,即测试集全程隔离,验证集仅含真实数据,并以验证集结果作为模型选择与迭代停止的判据,从而有效抑制伪标签引入造成的评估误差。在此基础上,本工作进一步开展模型可解释性分析,从全局重要性、局部响应与贡献分解 3 个维度揭示矿相变化对抗压强度预测的影响规律,为低碳模拟熟料组成的探索提供可检验的模型依据与可复用的线索。

1 实验

1.1 数据样本来源

本工作将熟料单矿视为最小特征单元,并将由多种矿相按比例干混得到的模拟熟料视为一个完整样本。所用熟料单矿按文献[4]的合成路线实验室制得,各单矿纯度均高于 98% (质量分数,下同), f -CaO 含量低于 0.5%,比表面积控制在 (400 ± 20) m^2/kg 。模拟熟料中所用二水石膏为国药集团有限公司提供的分析纯试剂。样本 28 d 抗压强度参考 GB/T 17671—2021《水泥胶砂强度检验方法》^[26]制样与养护。考虑到本工作样本批量制备对原料单矿消耗较

大,为降低单矿用量,试件尺寸采用 20 mm×20 mm×20 mm 的净浆试块,并以 6 个试块的平均值表示样本 28 d 抗压强度。实验样本 28 d 抗压强度的整体分布如图 1 所示,其抗压强度范围为 0~70.3 MPa,平均值为 32.2 MPa,中位数为 32.5 MPa。需要指出,该尺寸会引入一定的尺寸效应,且可能隐性传递至模型中。但由于本工作所有样本均在同一条件下制样与测试,可视为在相同实验基准上对模拟熟料体系的规律进行挖掘与建模。此外,各单矿的标准稠度用水量要求存在显著差异,若采用单一固定水胶比(W/B)可能导致部分样本工作性不足、难以成型。因此,本工作采用 2 种 W/B 控制策略,即当试样标准稠度需水量小于 0.35 时,W/B 固定为 0.35;当其标准稠度需水量大于 0.35 时,通过调整 W/B 使样本净浆的表观黏度与标准稠度下波特兰水泥(P·I)净浆的表观黏度一致^[27,28],该表观黏度通过数字旋转黏度计测量^[29]。该策略可以在保证工作性的同时限制 W/B 差异对模型预测误差的影响。此外,未将 W/B 作为模型输入的原因在于,本工作旨在讨论同一工作性能条件下模拟熟料矿相组成与 28 d 抗压强度间的构效关系,并将以此构建的预测模型用于后续组成筛选与设计。但仍需指出,微小的 W/B 差异仍可能是产生模型预测误差的原因之一。

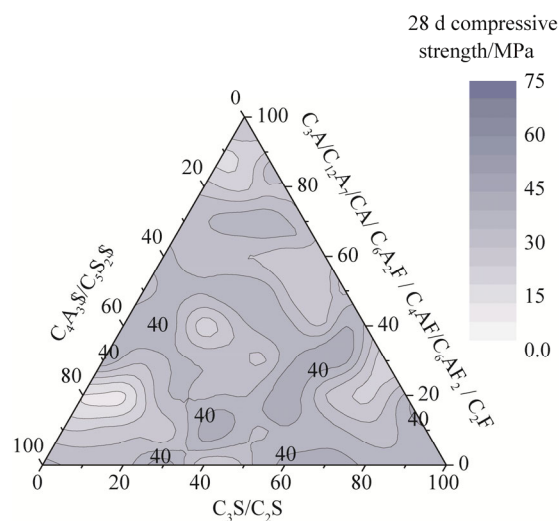


图1 实验样本 28 d 抗压强度分布图

Fig. 1 Distribution of the 28 d compressive strength of the experimental samples

1.2 自迭代数据增强策略

针对实验数据规模有限、组成空间分布稀疏导致初始代理模型泛化能力不足的问题,本工作提出

自迭代数据增强策略^[30], 扩充训练样本并提升代理模型的预测精度与跨体系泛化性能。该策略可视为基于半监督学习的自训练框架: 首先利用实验数据训练得到初始代理模型, 随后对尚未测得 28 d 抗压强度的虚拟模拟熟料组成进行预测, 生成伪标签, 并将伪标签样本与实验样本混合用于再训练, 从而实现逐轮迭代。整体流程见图 2。

该策略主要由“数据增强-模型迭代”和“数据集划分-锚定评价策略”2 部分构成。在“数据增强-模型迭代”中, 首先以实验数据训练获得初始 ANN 代理模型($f_{\theta}^{(1)}$)。在第 t 轮迭代($t \geq 2$)中, 按照既定组成约束生成一批虚拟模拟熟料组成(生成规则见 1.3 节与表 3), 并由上一轮模型 $f_{\theta}^{(t-1)}$ 预测其 28 d 抗压强度获得伪标签。随后将该批虚拟样本与实验样本合并训练, 得到模型 $f_{\theta}^{(t)}$ 。为避免测试集信息泄露, 各轮模型超参数选择与迭代停止判据均基

于验证集结果——当验证集误差连续 2 轮不再下降时停止迭代, 并输出验证误差最优的模型用于后续测试集评估与模型可解释性分析。

由于伪标签来源于模型预测并可能引入系统误差, 需设置相应的伪标签质量控制措施对虚拟样本进行限制^[31]。因此, 本工作除在虚拟样本生成阶段保持组成可行性外, 还采用“数据集划分-锚定评价策略”在模型迭代层面对虚拟样本的使用进行限制, 以此降低伪标签误差在模型中的隐性传递。具体来说, 预先从实验样本中留出 36 个样本作为独立测试集, 测试集在模型训练、迭代增强与超参数选择的全过程中完全隔离, 仅用于最终模型性能评价。剩余实验样本用于构建训练/验证集, 并且在任意迭代轮次中, 验证集始终仅由实验样本组成。虚拟样本仅进入训练集而不进入验证集与测试集, 从而避免自我验证造成的评估偏移。

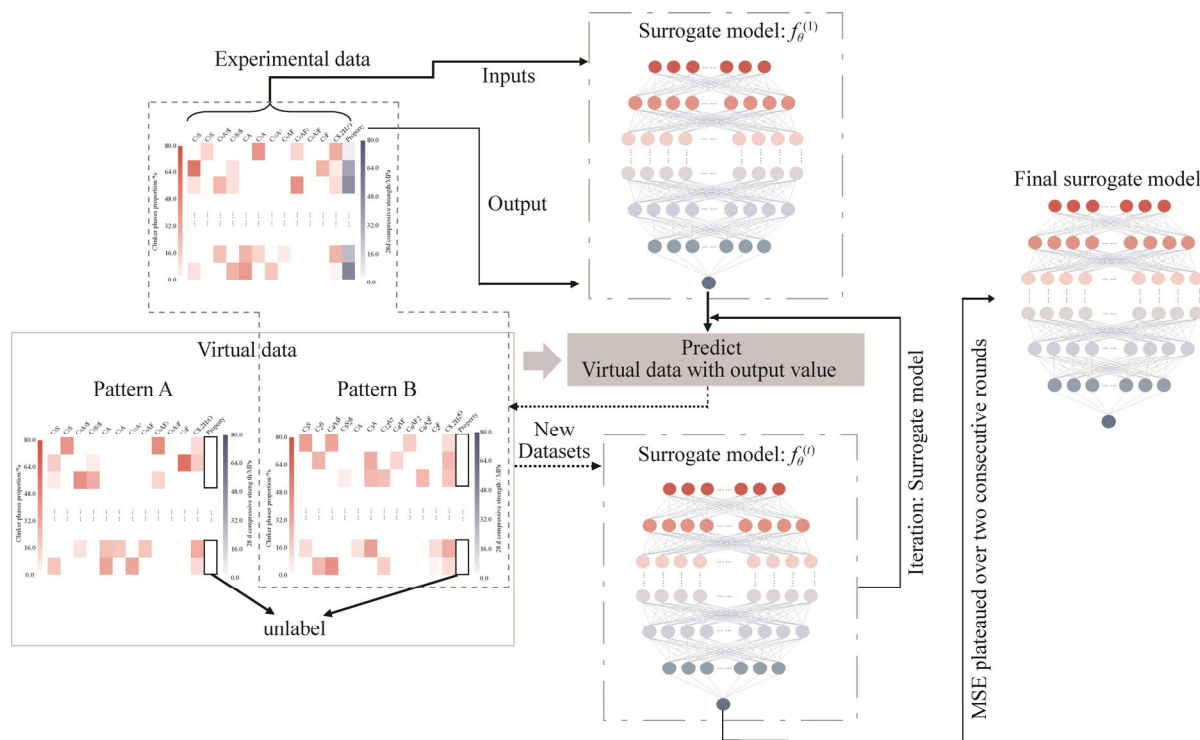


图 2 自迭代数据增强模型迭代整体流程图

Fig. 2 Overall workflow of the model iteration process under the self-iterative data augmentation strategy

1.3 数据集构建准则

实验数据采用受约束的三阶段抽样策略, 以确定各模拟熟料体系、体系内矿相与二水石膏占比。第 1 阶段, 从 11 种矿相中选取 3 或 4 种矿相定义一个模拟熟料体系。鉴于铁相 C_2F 、 C_6AF_2 、 C_4AF 与 C_6A_2F 为连续固溶体^[32], 每个体系中约束为至多包含一种铁相。第 2 阶段, 以 10% 为步长随机分配

各矿相质量分数, 并约束总和为 100%。为避免出现 0, 任一入选矿相的最小含量设定为 10%, 该阶段所得样本记为 x 。该采样策略旨在提高数据多样性, 并减少主观选择导致的数据聚类, 有利于建立模型跨体系泛化能力。第 3 阶段, 在 x 中掺入一定比例二水石膏作为调凝组分, 加入二水石膏后对整体组成进行归一化, 按式(1)和式(2)计算得到掺石膏

样本, 记为 X 。

$$X_i = \frac{x_i}{100 + \max(0, \dots, \sum \alpha_i x_i)} \times 100\% \quad (1)$$

$$X_g = \frac{\max(0, \dots, \sum \alpha_i x_i)}{100 + \max(0, \dots, \sum \alpha_i x_i)} \times 100\% \quad (2)$$

式中: X_i 为 X 中第 i 种熟料矿相占比; $\max(0, \dots, \sum \alpha_i x_i)$ 为 x 中应掺入二水石膏量, 其中 x_i 为 x 中第 i 种熟料矿相的占比; α_i 为第 i 种熟料矿相对应的石膏系数, 由表 1 提供。 X_g 为 X 中二水石膏的占比。由各熟料矿相与二水石膏构成的 12 维数据 X 作为模型输入, 28 d 抗压强度作为模型输出。实验数据集包含 264 个模拟熟料样本, 具体信息见表 2。

为提升小样本条件下代理模型对模拟熟料组成空间的表征能力, 并在组成空间覆盖扩展与伪标签噪声累积之间取得平衡, 用于数据增强的虚拟样本采用逐轮递增、分批引入的方式进行, 其矿相组成按上述约束规则生成, 28 d 抗压强度由上一轮代理模型预测获得伪标签。虚拟样本的各矿相分配步长调整为 1%, 并采用两类不同模式生成。模式 A (Pattern A) 在实验样本已覆盖的三元与四元子体系内加密采样, 用于系统性拓展实验样本中的组成区域。模式 B (Pattern B) 为分布式扩展策略。每轮按 1 : 1 生成体系内(实验样本中出现过的模拟熟料体系)

与体系外虚拟数据, 以增强训练数据多样性并抑制模型对局部组成空间的过拟合。第 II~IV 轮(Iteration II~IV)虚拟样本与实验样本的数量比分别为 1 : 1、2 : 1、3 : 1, 样本分布情况如图 3 所示。

1.4 数据集划分准则与模型评价指标

为保证不同迭代模型性能评价具有可比性, 并真实反映代理模型跨体系泛化能力, 本工作在建模之初即从实验样本中预先剥离测试集, 仅用于最终性能评估。其余实验样本与之后生成的虚拟样本用于构建训练集和验证集, 开展代理模型超参数优化研究。鉴于样本规模有限且需尽可能保持组成空间覆盖度, 本工作采用受约束的重复留出策略确定代理模型超参数, 即每次划分以体系均衡为核心原则, 从每个三元与四元熟料体系中随机抽取固定数量样本进入验证集, 避免验证误差被样本量较大的少数体系主导。以 Iteration I (仅含真实数据) 为例, 共有 39 个样本进入验证集(三/四元熟料体系中各抽取 1 个样本)、189 个进入训练集。对于 Iteration II~IV, 虚拟样本与实验样本采用相同权重参与模型训练, 以便直接评估虚拟样本作为监督信息增量的作用。为避免伪标签干扰验证判断, 虚拟样本严格限定仅进入训练集用于更新代理模型, 不作为模型验证样本。Iteration II~IV 的训练集与验证集比例分别为 378 : 69、567 : 99、747 : 138, 其余数据集划分信息见表 3。

表 1 矿相对应的石膏系数
Table 1 Gypsum coefficients of each clinker phase

C ₃ S	C ₂ S	C ₄ A ₃ \$	C ₅ S ₂ \$	CA	C ₁₂ A ₇	C ₃ A	C ₂ F	C ₆ AF ₂	C ₄ AF	C ₆ A ₂ F
0	0	0.224 ^a	-0.358 ^b	0.363 ^c	0.031 ^c	0.637 ^c	0.317 ^c	0.227 ^c	0.354 ^c	0.246 ^c

Notes: ^a The dosage of gypsum for C₄A₃\$ is based on an M value of 0.8, where M is defined as the ratio of the amount of SO₃ supplied by gypsum (by mass) to the amount of SO₃ required for Aft formation during the hydration of C₄A₃\$; ^b The gypsum coefficient of C₅S₂\$ is negative since it does not form Aft or AFm during hydration but instead produces gypsum; ^c Gypsum dosage was determined stoichiometrically by assuming complete hydration of aluminum and ferrite phases in model-clinker and maximizing the potential formation of AFm.

表 2 初始代理模型所用数据集的数据构建准则与数量
Table 2 Selection rules and quantity of model-clinker in dataset for Iteration I

Iteration	Belong to	Model-clinker system	Clinker phase selection rules	Number of systems and samples selected	Total
I	Training & validation sets	Ternary system	Select 3 phases from 11 clinker phases	9 systems/3 samples	228
		Quaternary system	Select 4 phases from 11 clinker phases	30 systems/6 samples	
		Known clinker	PC and BYT	PC/1 sample, BYT/2 samples	
		Unit clinker phase	Pure-phase clinker phases	11 samples	
		Phase-pure clinker phases + gypsum	pure-phase clinker phases with gypsum	7 samples	
	Test set	Ternary system & Quaternary system	In-system (included in the training and validation sets) Out-of-system	12 samples 24 samples	36

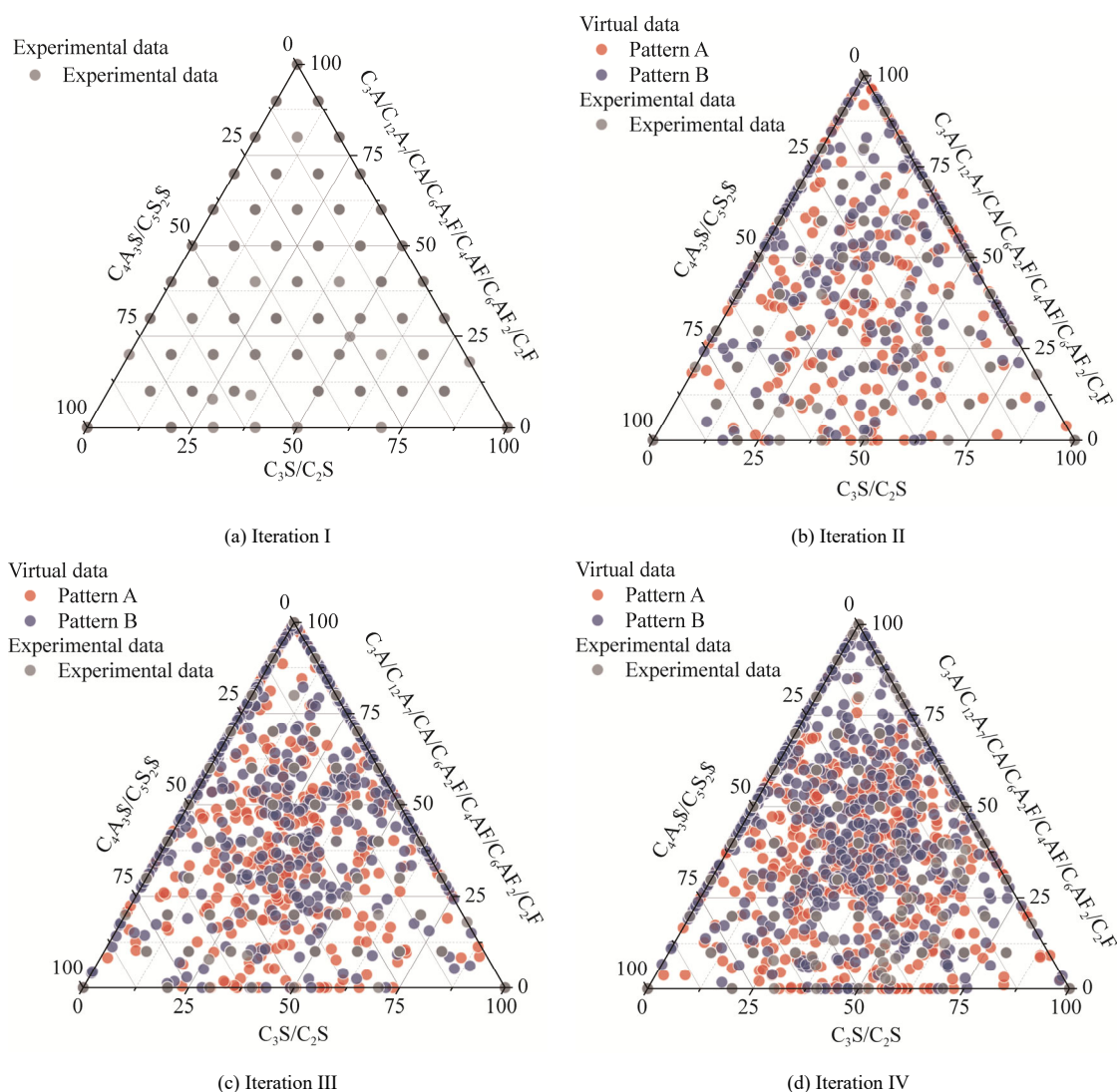


图 3 模型迭代所用训练集/验证集中实验样本和虚拟样本的矿相组成分布情况(x 样本空间)

Fig. 3 Phase composition distributions of the experimental and virtual samples in the training and validation sets used for each model iteration (represented in x)

采用均方误差(MSE)作为代理模型超参数选择的评价指标。当代理模型超参数组合确定后,在合并后的训练集与验证集上重新训练模型,并最终在测试集上评估其稳定性与泛化性能。为量化测试过程中预测值与实验值之间的偏差,本工作以 MSE、相关系数(R^2)、平均绝对误差(MAE)与平均绝对百分比误差(MAPE)作为评价指标,各指标的计算公式参考文献[15]。

1.5 代理模型超参数优化

在代理模型迭代过程中,训练集的样本数量增加、组成空间覆盖扩展、真实/虚拟样本比例变化以及潜在伪标签噪声水平变化,均会影响模型所需容量与正则化强度。因此,采用固定单套超参数难以在各迭代轮次保持最优。本工作对每轮代理模型进

行独立的超参数优化,在第 t 轮对应的数据集上确定该轮代理模型超参数组合,并以验证集 MSE 为判据。每轮超参数选择采用基于树结构 Parzen 估计器的贝叶斯优化策略(TPE),以最小化验证集 MSE 为目标,执行 1000 次超参数组合搜索迭代。表 4 列出了待优化的超参数及其搜索范围,其余超参数使用 Python 库默认值。为降低数据稀疏导致的数据集划分异质性影响,每个候选超参数组合在 5 个固定随机种子对应的数据集划分上分别训练,最终以 5 次验证集 MSE 的平均值作为该组合的判断值。当多个候选超参数组合表现接近时,优先选择平均 MSE 低且标准偏差小的组合作为最终代理模型超参数,以降低单次数据集划分带来的偶然波动。

表 3 第 II~IV 轮迭代中代理模型所用数据集的数据构建准则与数量
Table 3 Selection rules and quantity of model-clinker in dataset for Iteration II–IV

Iteration	Belong to	Experimental or virtual data	Pattern	Selection of clinker system rules	Number of systems and samples selected	Each/Set	Total
II	Training & validation sets	Experimental	A	In Iteration I training & validation sets	228 samples	228	447
				The systems in Iteration I: 9 ternary system, 30 quaternary system, PC and BYT	9 systems/3 samples 30 systems/6 samples PC and BYT/6 samples	219	
		Virtual	B	The systems in Iteration I: 9 ternary system, 30 quaternary system, PC and BYT	9 systems/2 samples 30 systems/3 samples PC and BYT/3 samples	219	
				The systems not included in Iteration I	105 samples		
				In Iteration I test sets	36 samples	36	
				In Iteration I training & validation sets	228 samples	228	
III	Test set	Experimental	A	The systems in Iteration I: 9 ternary system, 30 quaternary system, PC and BYT	9 systems/6 samples 30 systems/12 samples PC and BYT/12 samples	438	666
				The systems in Iteration I: 9 ternary system, 30 quaternary system, PC and BYT	9 systems/3 samples 30 systems/6 samples PC and BYT/6 samples	438	
		Virtual	B	The systems in Iteration I: 9 ternary system, 30 quaternary system, PC and BYT	9 systems/3 samples 30 systems/6 samples PC and BYT/6 samples	438	
				The systems not included in Iteration I	219 samples		
				In Iteration I test sets	36 samples	36	
				In Iteration I training & validation sets	228 samples	228	
IV	Training & validation sets	Experimental	A	The systems in Iteration I: 9 ternary system, 30 quaternary system, PC and BYT	9 systems/9 samples 30 systems/18 samples PC and BYT/18 samples	657	885
				The systems in Iteration I: 9 ternary system, 30 quaternary system, PC and BYT	9 systems/5 samples 30 systems/9 samples PC and BYT/9 samples	657	
		Virtual	B	The systems in Iteration I: 9 ternary system, 30 quaternary system, PC and BYT	9 systems/5 samples 30 systems/9 samples PC and BYT/9 samples	657	
				The systems not included in Iteration I	324 samples		
				In Iteration I test sets	36 samples	36	
				In Iteration I training & validation sets	228 samples	228	
	Test set	Experimental		In Iteration I test sets	36 samples	36	36

表 4 每轮代理模型的所需优化超参数及其范围

Table 4 Hyperparameters to be optimized for each surrogate model and its respective selection ranges

Hidden_dim	Dropout	Num_layers	lr	Weight_decay	Optimizer	Batch_size	Epoch
4–256, step=4	0–0.3	2–20, step=2	e^{-4} – e^{-2}	e^{-3} – $5e^{-1}$	“Adam”, “AdamW”	16–128, step=16	2000

1.6 代理模型可解释性分析

1.6.1 置换特征重要性 在确定最终代理模型后，为量化各矿相作为输入对预测结果的全局贡献，本工作采用受约束的置换特征重要性(Constrained permutation feature importance, Constrained PFI)进行评估研究，组成约束规则同 1.3 节所述。需要指出，若直接在 12 维输入 X 上逐维独立置换，将破坏矿相与二水石膏的耦合关系，易产生不可行样本并引入重要性偏置。因此，本工作仅对 x 样本中的矿相组成进行置换。对每次置换后的 x 样本，按 1.3 节规则反推二水石膏掺量并归一化生成对应的 12 维输入 X ，以确保置换样本始终位于可行域内。将验证集与测试集合并的数据集定义为解释集(Exp set)。首先计算未置换时模型在 Exp set 上的基线均方误差， MSE_0 。随后针对每个矿相分别重复置换

300 次，得到对应误差 MSE_i ，并按式(3)计算该矿相的平均置换重要性。

$$PFI_i = \frac{\sum (MSE_i - MSE_0)}{300} \quad (3)$$

式中： PFI_i 为置换 i 矿相的置换重要性， MSE_i 为每次置换 i 矿相后均方误差。同样，该部分亦需考虑到数据划分的随机性，以上过程在 1.5 节中 5 个不同划分的数据集上重复执行，取 5 次结果平均值作为最终平均置换重要性。

1.6.2 累积局部效应分析 为进一步刻画代理模型对矿相含量变化的响应趋势，采用累积局部效应(Accumulated local effects, ALE)对 11 种矿相组成变量进行解释分析，以此作为 Constrained PFI 的补充，识别平均影响方向、非线性响应及可能的阈值区间。ALE 的局部扰动在满足 1.3 节定义的组成

约束下进行。针对样本在零值附近分布不均的问题,对目标矿相(X_i)采用 0 值独立成箱以及正值分位数分箱相结合方式确定区间边界。对落入某分箱区间的样本,将该样本的目标矿相含量分别设置为该区间上/下边界,其余矿相在 x 样本空间按比例缩放,并按规则反推二水石膏掺量,归一化后得到对应 X ,并据此计算模型输出差分。分箱局部效应定义为该分箱内模型输出差分的样本均值,并沿分箱方向累积获得 ALE 曲线。考虑数据稀疏区域的外推风险,允许分箱为空,其分箱局部效应记为缺失。为评估趋势的统计稳健性,对训练集与 Exp set 分别进行自助法(Bootstrap)重采样(500 次)并重复计算 ALE 曲线,构建 95%置信区间(95% CI)。

1.6.3 沙普利加性解释 采用基于 SHapley 值(Φ)的加性解释框架(SHAP)对代理模型输出进行归因分解,计算各矿相对单样本预测值的贡献,并以 Exp set 上 $|\Phi_i|$ 的均值表征全局重要性,以反映矿相的贡献强度及其驱动方向。该框架的解释空间仍定义在 x 样本,在 SHAP 调用预测函数时,仍按照前述规则对于组成进行约束。本工作采用模型无关的解释器(KernelExplainer)进行 SHAP 求解,背景分布取训练集样本的估计基线值,对 Exp set 中样本进行全局解释。

2 结果与讨论

2.1 数据增强结果分析

图 4 展示了自迭代数据增强过程中,虚拟样本在组成空间与强度标签空间的分布演化过程。由图 4a 可见,迭代由 Iteration I 推进至 Iteration IV 时,组成空间相似度分布曲线的峰位与宽度变化较小,但低相似度样本对(<0.3)占比下降、中等相似度样本对区间($0.3\sim 0.7$)占比上升。该结果表明,在不破坏组成可行域约束的前提下,引入虚拟样本提高了原实验数据集(Iteration I)在组成空间中的覆盖密度,降低了稀疏邻域所占比例。同时,分布仍保留少量高相似度样本对,且该占比未出现明显变化,说明增强过程并未产生大量近相似重复样本,数据集整体异质性仍然存在。进一步比较同一体系在不同迭代轮次下 2 种采样模式的相似度分布(图 4b~图 4d),2 种模式均在组成分布层面引入了信息增量。体系内样本对相似度上移,表明组成空间获得更连续的密集覆盖。基于此,结果表明 Pattern A 通过体系内加密采样缩短了组成空间中的局部外推间隙,而 Pattern B 还在此基础上进一步

引入体系外样本,扩大体系覆盖范围,从而降低由覆盖不足引起的外推风险。在强度标签空间方面,28 d 抗压强度分布(图 4e)表明,虚拟样本并非对原始强度分布的简单复制,而是对强度区间覆盖进行扩展与一定程度的均衡化。更均衡的目标空间覆盖可避免训练监督信号长期由少数强度区间主导,缓解低分布区间出现系统性偏差与误差放大的风险,提升模型对宽强度范围样本的稳健拟合能力。因此,图 4 所示的组成空间与目标空间分布变化共同证明,该自迭代增强策略在保持可行域约束的同时有效填补了稀疏区域,暗示了代理模型预测性能改善或许并非仅源于样本数量增加,更关键的是其改变了训练数据在组成空间中的覆盖方式及信息分布。

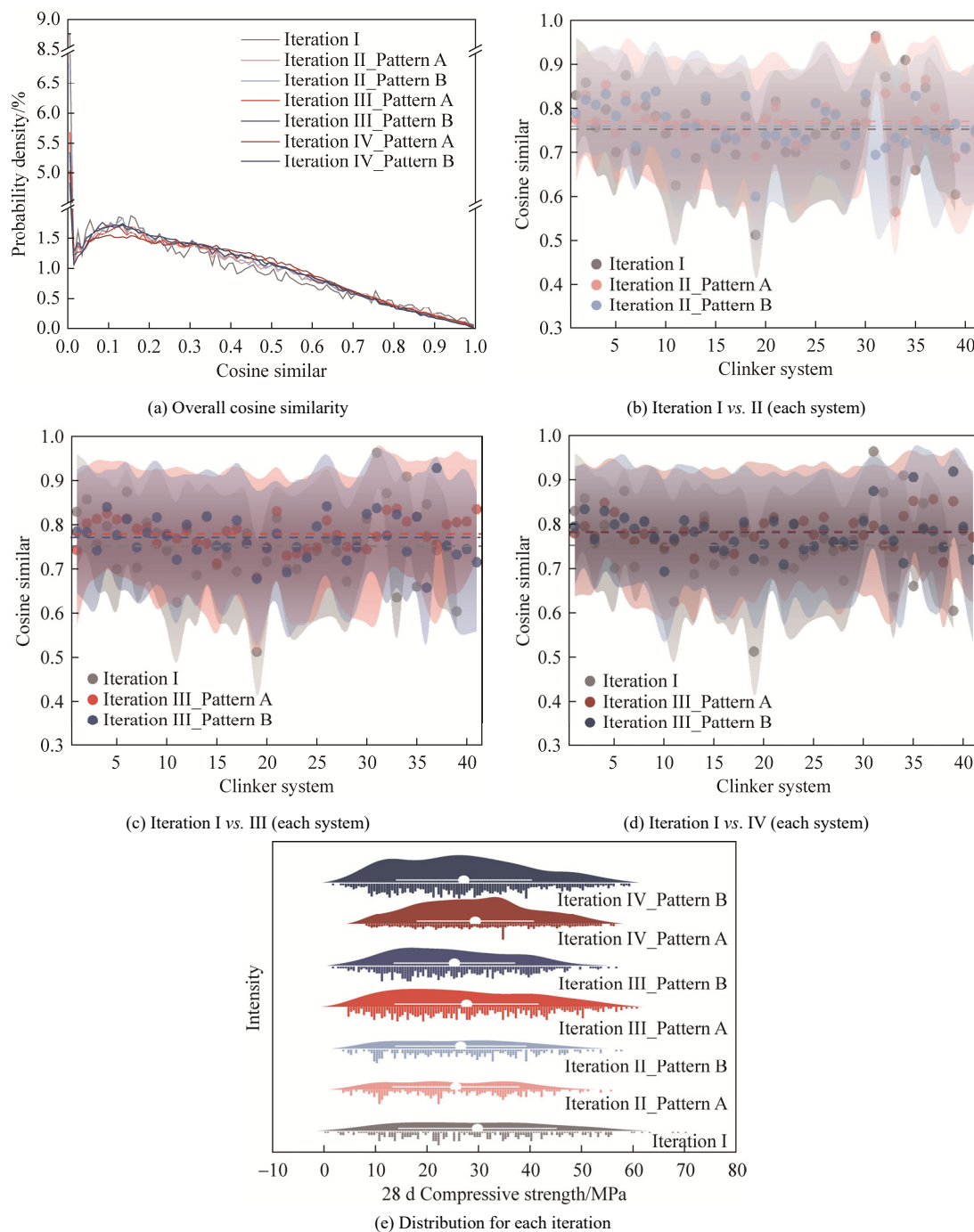
2.2 模型超参数及验证集 MSE

图 5 汇总了自迭代增强过程中代理模型的最优超参数及其对应验证集 MSE 的变化规律。随着代理模型的迭代与训练样本规模的扩大,代理模型最优网络层数与隐藏层维度呈上升趋势,验证集 MSE 在 Iteration II 和 Iteration III 阶段下降并在 Iteration IV 阶段出现回升,表明增强策略在增加样本信息的同时,其边际收益逐步递减,并开始受到伪标签误差积累的影响。这一变化在一定程度上反映了图 4 所示组成空间覆盖的扩展过程,即随着训练数据由局部稀疏转向相对连续,模型能够利用更多有效样本信息降低预测误差,但与此同时,伪标签的累积也将会逐步引入额外噪声风险。

由图 5a 可见,模型容量随样本量增加而提高,反映虚拟样本引入了更丰富的组成-强度非线性信息,需要更强的函数表达能力来降低误差。模型迭代初期(Iteration II),Pattern A 倾向于选择更高容量配置,而 Pattern B 则保持相对更小的网络规模(2 层,88 个神经元)。这一差异归因于 Pattern A 选择在已覆盖体系内加密采样,新增样本主要提供局部连续监督信号,提高容量有利于刻画微小响应。相比之下,Pattern B 通过体系外扩展提升多样性,但同时引入更显著的输入分布漂移与潜在伪标签偏差,高容量模型更易对伪标签噪声产生过拟合,从而放大误差传播风险。因此在该阶段,超参数搜索更倾向于容量受控的配置以维持验证误差的稳定性。随着迭代继续推进,2 种模式均逐步选择更高容量,说明当样本信息累积到一定程度后,误差下降带来的收益开始超过容量增大引入的方差代价。

图 5b 显示验证集 MSE 随迭代呈现先降后升的变化特征。Iteration II 和 Iteration III 阶段验证集 MSE 下降,说明伪标签样本在总体上提供了有效的监督并改善了对稀疏组成区域的函数逼近。迭代至 Iteration IV 阶段,验证集 MSE 出现回升,提示新增伪标签的边际信息增益已不足以抵消噪声累积带来的泛化损失。该回升在 Pattern B 下更为显著,

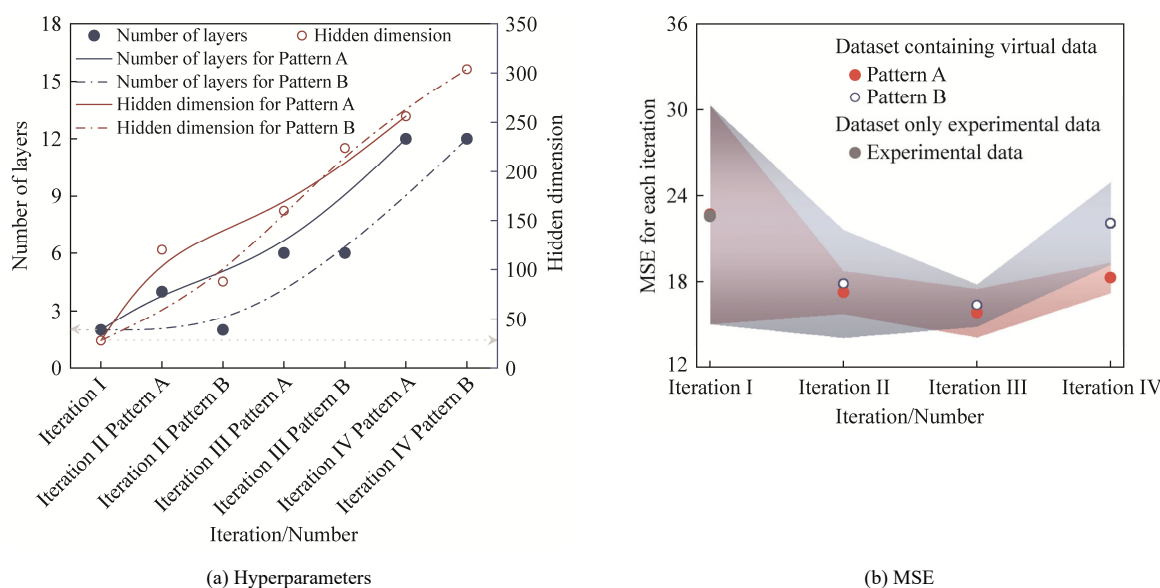
表明体系外样本更容易加速伪标签偏差在训练过程中的隐性传递,更早触发泛化误差反弹。此外,随着虚拟样本量的增加,验证集 MSE 的波动幅度显著降低[由 Iteration I 的 ± 7.63 降至 Iteration IV 的 ± 1.07 (Pattern A)和 ± 2.88 (Pattern B)],说明模型对数据划分随机性的敏感性得到抑制,验证评估信号更稳定。



Note: Shaded bands represent the corresponding standard deviations of the cosine similarity values for all sample pairs within each system. Wider shaded bands indicate greater in-system variability, whereas narrower bands suggest higher consistency among samples.

图 4 数据集中样本间余弦相似性及 28 d 抗压强度分布情况

Fig. 4 Cosine similarity among samples in the dataset and the distribution of 28 d compressive strength.



Note: Shaded band represents the corresponding standard deviation of the validation set MSE

图 5 代理模型超参数及其在验证集上的 MSE 变化

Fig. 5 Variation of surrogate model hyperparameters and the MSE on the validation set

2.3 模型预测准确性分析

如图 6 所示, 自迭代增强显著提升了代理模型在测试集上的预测精度与跨体系泛化能力。图 6a 将测试集 MSE 分解为总体、体系内与体系外误差。Iteration I 中体系外误差显著高于体系内, 说明初始模型误差主要来自未覆盖体系所导致的外推偏差。随着 Iteration II 和 Iteration III 阶段引入虚拟样本并扩展模拟熟料组成覆盖范围, 总体 MSE 降低且体系外 MSE 降幅更大, 表明增强策略补足了稀疏区域的监督信号, 使模型在体系外样本上的偏差得到抑制。需要注意, Iteration IV 阶段 2 种采样模式呈现不同变化, Pattern A 下各子集 MSE 仍保持较低误差且置信区间波动受控, 而 Pattern B 虽进一步降低体系内误差, 却伴随体系外误差增加。这表明在真实数据支撑有限的条件下, 体系外伪标签易削弱外推稳定性。当伪标签边际信息增益递减且噪声累积占优时, 迭代可能出现过增强现象。测试集中样本预测值与实验值的散点图进一步验证了上述趋势(图 6b~图 6d)。相较 Iteration I 和 Iteration II 阶段散点整体向对角线收敛, 落入 $\pm 15\%$ 误差带的样本数量显著增加, 说明模型对组成-强度映射的主要趋势拟合更稳定。然而在 Iteration IV 阶段, 即便对角线附近仍保持较高密度, 高/低强度端仍可观察到一定系统性偏差及少量越界点, 暗示极端强度区间的有效监督仍不足。结合图 4 与图 5 可知, 测试集误差的下降是由组成空间覆盖扩展、目标空间监督信号均衡化以及相应模型容量调整共同驱动的, 而

Iteration IV 阶段中部分指标反弹, 则说明一味增加虚拟样本熟料, 扩展组成覆盖范围未必会持续转化为提升泛化性能。因此, 对于高/低强度区间预测的改进可能更依赖实验样本补充, 或引入更严格的伪标签筛选机制。

图 6e 和图 6f 以雷达图汇总了 4 项核心指标的迭代演化过程。Pattern A 和 Pattern B 均呈现从 Iteration I 到 Iteration II 和 Iteration III 阶段的显著改善, 模型 R^2 提升, MSE、MAE、MAPE 同步降低, 说明伪标签总体上提供了有效监督信息并增强了代理模型的泛化性能。综合总体误差与体系内/外误差分解结果(图 6a)以及多评价指标(图 6e 和图 6f), Iteration II 阶段的 Pattern B 在测试集上取得最低总体 MSE (9.8), 且体系内与体系外 MSE 分别为 6.9 与 11.3, 同时 R^2 达到 0.79, MAE 与 MAPE 分别为 2.36 与 5.11%。值得注意的是, 最优模型 R^2 为 0.79, 表明在当前实验数据规模下, 即使开展了数据增强工作, 预测模拟熟料 28 d 抗压强度仍存在一定误差。该部分误差一方面是由实验样本对组成空间覆盖有限造成, 另一方面则可能与自迭代过程中伪标签的误差积累有关。此外, 本工作中现有的映射关系还跨越了模拟熟料晶体结构、水化微结构、水化产物组成等多个尺度的因果链, 这也是限制模型解释能力的重要原因之一。因此, 本工作模型更适用于小样本条件下模拟熟料组成的趋势筛选和相对比较, 不宜替代实验用于低碳熟料体系的高精度绝对预测。后续仍需在筛选出模拟熟料组成的基础

上,开展煅烧制度、水化反应动力学等更加详尽的验证研究工作。综上所述,本工作在精度与跨体系

泛化间取最优折中,选取 Iteration II 阶段的 Pattern B 模型作为最终代理模型用于后续可解释性分析。

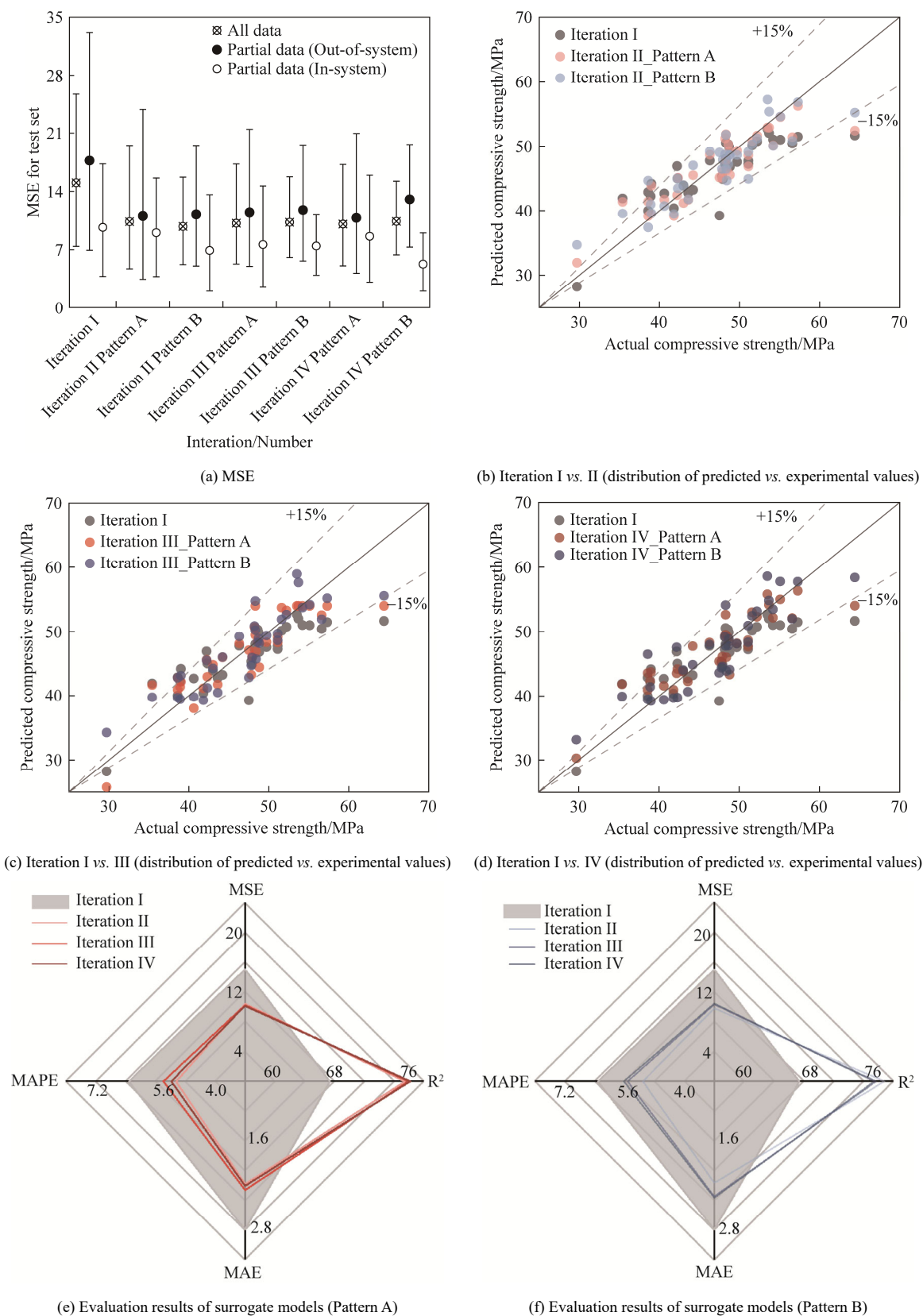


图6 迭代过程中各代理模型的预测性能评估

Fig. 6 Evaluation of the predictive performance of individual surrogate models during the iterative process

2.4 代理模型可解释性

2.4.1 置换特征重要性分析 图 7 量化了各矿相特征对代理模型预测误差的全局敏感性, 并比较了不同 Exp set 上的稳健性。Constrained PFI 分布呈现显著的稀疏驱动特征, 仅置换少数矿相即引起预测误差陡增, 而多数变量的边际增量接近于零, 表明模型倾向于将有效判别信息主要集中于有限的关键矿相。

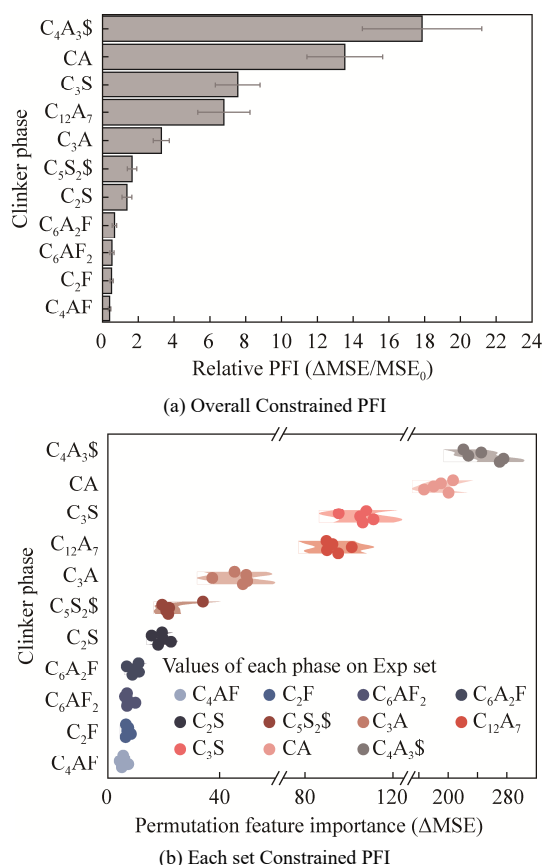


图 7 置换特征重要性的全局排序、解释集差异与相对对比
Fig. 7 Global ranking, explanatory-set differences, and relative comparisons of permutation feature importance

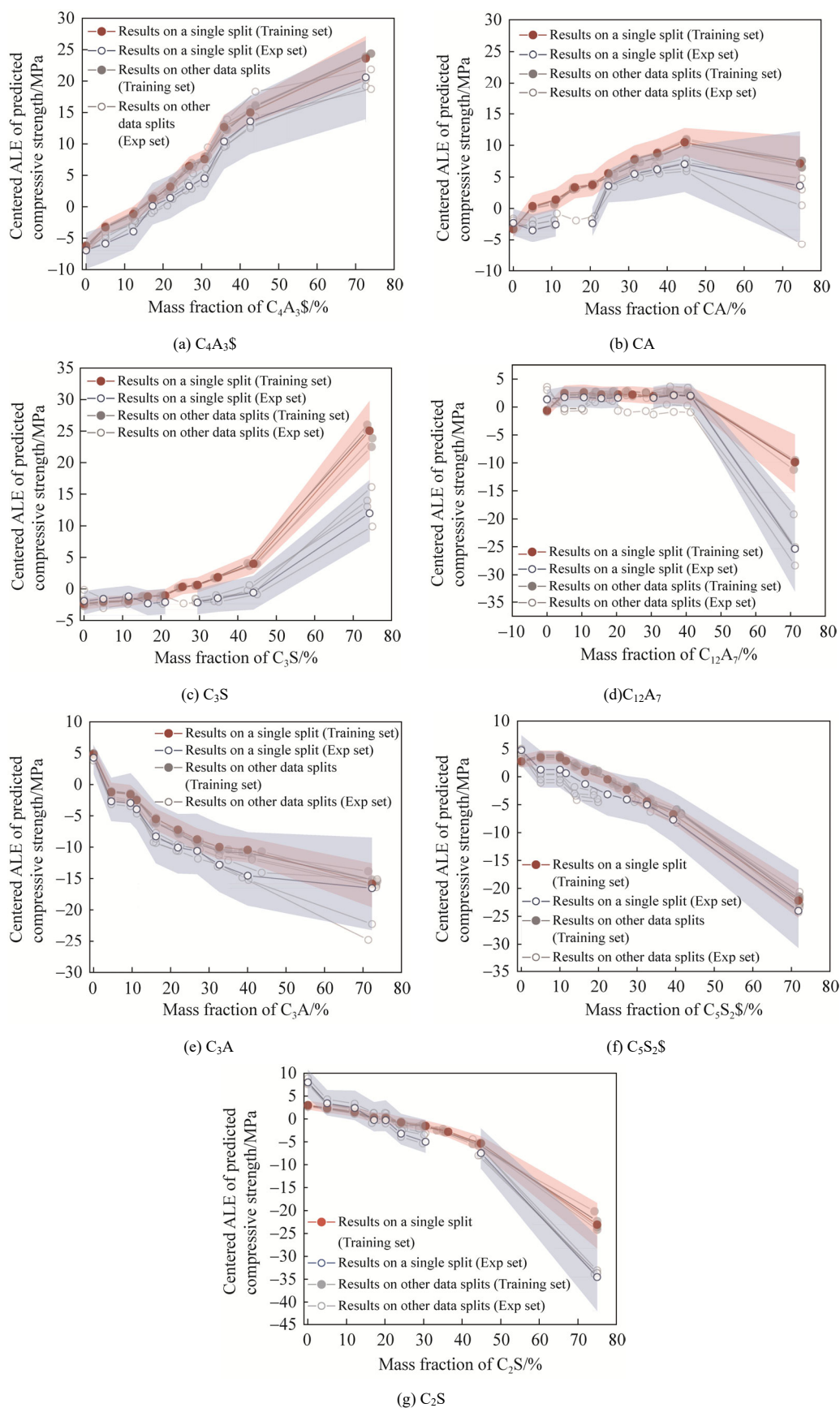
从全局排序结果可见(图 7a), 置换 $C_4A_3\$$ 与 CA 会导致 $\Delta MSE/MSE_0$ 显著高于其他矿相, C_3S 与 $C_{12}A_7$ 次之。相比之下, C_3A 、 $C_5S_2\$$ 、 C_2S 的 $\Delta MSE/MSE_0$ 小于 5, 约为 $C_4A_3\$$ 与 CA 的 20%, 而铁相置换后 $\Delta MSE/MSE_0$ 更是小于 1。该结果表明, 在当前数据分布与组成约束条件下, $C_4A_3\$$ 与 CA 等矿相对模型预测信息具有较强的不可替代性。需明确, 矿相置换不仅会改变该相的占比信息, 还会改变二水石膏掺量, 并将这一变化传导至 X 样本空间中。因此, $C_4A_3\$$ 、CA 和 $C_{12}A_7$ 的扰动, 往往会引发更强的误差敏感性。而对于 C_3S , 扰乱其占比会削弱模型判别 28 d 抗压强度水平的能力, 表现出较强且相对稳定的敏感性。同时, 中低 Constrained

PFI 值并不意味着该矿相对强度无关。由于组成约束下各矿相之间存在相关、互补与冗余, 置换单一矿相所造成的信息缺失可能被其他矿相部分补偿, 或被二水石膏联动配平引起的缩放效应所抵消, 从而在全局平均意义上表现为较低的边际敏感性。

进一步观察不同 Exp set 上的重要性分布可知(图 7b), $C_4A_3\$$ 、CA、 C_3S 与 $C_{12}A_7$ 的 ΔMSE 始终保持同一量级且波动较小(相对误差 < 10%), 说明其重要性在跨数据集划分上仍具备稳健性。相对而言, 其他矿相在不同 Exp set 上的 ΔMSE 离散度更大, 提示其边际作用对样本组成波动与局部稀疏性更为敏感。综上, Constrained PFI 识别了维系模型跨体系泛化稳定性的核心驱动矿相, 并揭示了代理模型对矿相信息的非均等依赖性。但 Constrained PFI 作为全局扰动敏感性指标, 无法给出方向性、阈值与交互细节, 需结合 ALE 与 SHAP 进一步解析。

2.4.2 累积局部效应分析 图 8 展示了 Constrained PFI 排名前 7 的矿相在训练集与 Exp set 上的 ALE 曲线及其 95% CI。ALE 扰动同样是在组成约束下实现, 提高或降低某一矿相必然相应改变其他组分, 因而其实质含义应为组成约束下的平均响应趋势而非单一矿相的独立贡献。同时, 由于采用分箱局部扰动并允许空箱跳过, 曲线断点对应于该区间缺少样本支撑。ALE 曲线结果显示, 各矿相在不同数据划分下呈现一致性, 即 $C_4A_3\$$ 、CA、 C_3S 和 $C_{12}A_7$ 的累积效应以正向为主, 而 C_3A 、 $C_5S_2\$$ 、 C_2S 表现为稳定的负向响应。对于正向驱动矿相, $C_4A_3\$$ 和 C_3S 在样本覆盖的含量区间内近似单调上升(图 8a 和图 8c), 其中 C_3S 在低含量端边际增益较弱, 而进入中高含量区(约 > 25%)后出现明显的增强, 提示模型更倾向于在一定含量水平以上, 将 C_3S 的提升映射表现为明显的抗压强度增益效果。相较之下, CA 和 $C_{12}A_7$ 则呈现明显的阈值特征(图 8b 和图 8d)。在低、中含量范围内其贡献仍为正向累积, 但当其占比超过 40% 后, 增益效果递减, 甚至转变为强负响应。结合其矿相的水化机制^[33], 或可理解为二者增加必然会挤压其他矿相并联动改变石膏掺量, 在模型层面上将这种组合变化映射为对 28 d 强度的惩罚项。因此, CA 和 $C_{12}A_7$ 的响应阈值可为下一步组成设计提供搜索的软边界。

对负向响应矿相而言, $C_5S_2\$$ 和 C_2S 的 ALE 曲线随含量增加呈持续下降趋势, 表明模型学习到二者占比升高会稳定拉低预测强度, 体现为强度惩罚趋势, 这一趋势与其相对缓慢的水化反应过程、对



Note: Shaded band represents the corresponding 95% CI

图 8 Constrained PFI 中排列前 7 的特征 ALE 曲线

Fig. 8 ALE curves of the top seven features ranked by Constrained PFI

28 d 抗压强度贡献有限的认识一致。 C_3A 则表现为平台段后(超过约 10%)加速下降, 提示其负向效应具有阈值放大行为, 因而在配比空间搜索中需对其采用更严格的上限控制。此外, 训练集与 Exp set 在低/中含量区间的曲线形态总体一致, 表明模型对矿相的响应

方向及组成与强度间非线性映射具有跨划分稳定性。高含量端多数矿相的 95% CI 明显加宽, 主要由该区间样本覆盖不足(甚至空分箱)所致。

2.4.3 SHAP 加性解释分析 在 Constrained PFI 与 ALE 的基础上, 图 9 进一步采用 SHAP 框架对

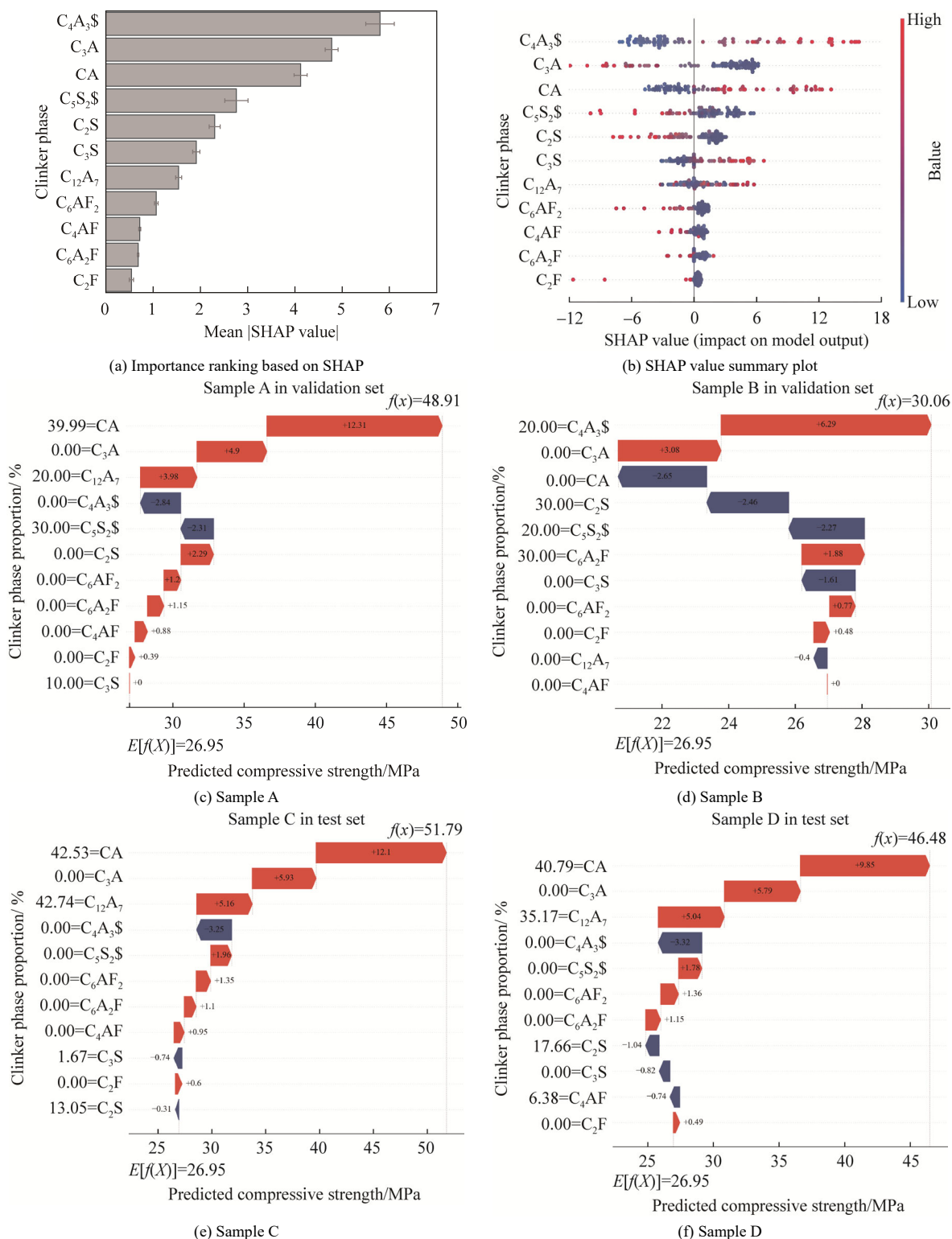


图 9 基于 SHAP 加性解释的模型分析

Fig. 9 Model interpretation based on SHAP additive explanations

代理模型输出进行加性归因,将单一样本的预测分解为各矿相相对改变基线期望值的贡献。由图 9a 可见, C_4A_3S 的归因幅度最大,而铁相整体贡献有限。部分矿相的排序与 2.4.1 节 Constrained PFI 结果存在差异,这是由于两者描述对象不同。例如 C_3A 在 Constrained PFI 中排名靠后而在 SHAP 中位列前列,说明其对预测值的抑制幅度显著(SHAP 分析结果),但其信息可在强相关组成体系中被其他矿相部分补偿(Constrained PFI 分析结果)。这一现象反而表明代理模型学习到的并非单矿相的孤立效应,而是矿相间在配比约束下的协同变化。因此,SHAP 排序阐明的是矿相对预测值的提升或抑制作用,而 Constrained PFI 则是揭示矿相对模型误差的影响,两者相互补充。

进一步从方向性看,图 9b 显示 C_4A_3S 含量较高的样本主要对应正 SHAP 值、较低的样本对应负 SHAP 值,说明 C_4A_3S 对强度预测呈稳定正向驱动。 CA 同样表现为高质量占比带来正向贡献,与 ALE 中 CA 的总体正响应区间一致。 C_3S 总体分布偏向正 SHAP 值,显示其对 28 d 抗压强度提供稳定增益,但贡献强度低于 C_4A_3S/CA 。相反, C_3A 呈现清晰的反向分布,表明其在当前体系与 28 d 抗压强度指标下, C_3A 对抗压强度增长表现为稳定抑制作用。同样, C_5S_2S 与 C_2S 的高值点大多分布在负 SHAP 区,指示其含量升高会拉低 28 d 抗压强度预测,与 ALE 中两者负向边际效应一致。值得注意的是,同一矿相在不同样本上的 SHAP 值离散幅度较大,说明矿相贡献并非常数,而是容易受到组成约束与矿相协同变化的耦合影响,这也解释了为何单一矿相比例的提升并不必然对应抗压强度的线性提升。

图 9c~图 9f 给出了验证集与测试集中各 2 组样本局部解释示例,分别依次记为 Sample A、Sample B、Sample C、Sample D。Sample A 与 Sample B 为验证集中样本,其抗压强度从基线期望值分别上升至 48.9 MPa 和 30.1 MPa。前者的主要推升来自 C_4A_3S 的强正贡献,并叠加 CA 的正向增益,而 C_3A 与 C_5S_2S 提供主要负向抵消,说明高强度来自正向叠加且必须同步抑制负向相关矿相占比,以避免净增益被抵消。在 Sample B 中,尽管 C_4A_3S 仍表现出显著正贡献,但其正向相关矿相(CA 与 C_3S 等)在该样本的特征取值为 0,使得负向项的累计抵消占优,从而将预测强度拉低,该样本说明高 C_4A_3S 含量并不足以单独保证高强度,模型实质学习到的映射规律是受组成约束的耦合响应。在测试集上, Sample

C(图 9e)达到 51.8 MPa,其推升路径表现为 CA 主导的正向贡献并叠加其他矿相的小幅修正,该样本中 C_3A 取值为 0,使得无明显负向抵消项,从而更有利于实现高强度输出。Sample D(图 9f)为 46.5 MPa, CA 依旧提供主要正向贡献,但 C_4A_3S 与 C_5S_2S 等矿相的负向抵消更强,导致最终强度低于 Sample C。对比结果进一步印证了图 9b 所揭示同一矿相的贡献幅度受其他矿相配比与约束条件影响的规律。

需要强调的是,本工作中 Constrained PFI、ALE 和 SHAP 所揭示的是最优代理模型在当前模拟熟料数据分布及组成约束条件下学习到的统计关联和平均响应趋势,而非对模拟熟料的水化机理进行直接证明,上述分析结果仅作为机理研究与配比筛选的线索,但不能直接替代针对真实熟料体系的独立实验验证。

3 结论

1) 提出并验证了一种面向小样本模拟熟料体系的自迭代数据增强框架,适用于样本规模有限且组成空间稀疏的模拟熟料体系。在组成约束下生成候选组成并以伪标签进行逐轮自训练,同时采用“测试集全程隔离、验证集仅含真实数据”的评价策略抑制评估偏移,从而在数据稀疏条件下有效扩展了模拟熟料体系组成与强度间映射的监督信号覆盖范围。

2) 自迭代增强显著提升了代理模型在独立测试集上的预测精度与跨体系泛化能力。综合多指标对比, Iteration II 阶段(Pattern B)在精度与外推稳定性之间取得最优折中,其测试集总体 MSE 为 9.8,体系内/外 MSE 分别为 6.9 与 11.3, R^2 为 0.79, MAE 与 MAPE 分别为 2.36 与 5.11%。进一步迭代虽可降低部分体系内误差,但在体系外样本上出现误差反弹,表明伪标签边际信息增益递减与噪声累积对泛化性能形成约束。

3) 综合 Constrained PFI、ALE 与 SHAP 三类模型的解释结果可知, SHAP 与 ALE 在方向性与非线性特征上形成相互印证。 C_4A_3S 与 CA 是 28 d 抗压强度提升的主要正向驱动因子, C_3A 、 C_5S_2S 与 C_2S 为主要负向因子, $C_{12}A_7$ 体现阈值效应。同时, SHAP 与 Constrained PFI 排序差异提示,在组成强相关体系中高归因不等同于高不可替代性。

4) 若进一步开展模拟熟料体系的组成逆向设计,可优先利用 Constrained PFI 分析中对模型泛化

最敏感的关键矿相(如 $C_4A_3\$$ 、CA)并结合 SHAP 与 ALE 给出的方向性与阈值信息, 对 C_3A 、 $C_5S_2\$$ 、 C_2S 及 $C_{12}A_7$ 设置组成约束与搜索边界, 从而实现 28 d 抗压强度的可解释与可控优化。

参考文献:

- [1] XU P C, JI X B, LI M J, et al. Small data machine learning in materials science[J]. *npj Comput Mater*, 2023, 9: 42.
- [2] 梁叔全, 郭珊, 曹鑫鑫, 等. 二次电池正极材料能带高通量计算与电化学特性关联分析[J]. *硅酸盐学报*, 2025, 53(7): 1786–1800. LIANG Shuquan, GUO Shan, CAO Xinxin, et al. *J Chin Ceram Soc*, 2025, 53(7): 1786–1800.
- [3] 刘怿泓, 毕文柱, Mohamed Ait Tamer, 等. 智能模型高通量筛选无机钠固态电解质[J]. *硅酸盐学报*, 2025, 53(7): 1801–1808. LIU Yihong, BI Wenzhu, TAMERD M A, et al. *J Chin Ceram Soc*, 2025, 53(7): 1801–1808.
- [4] ZHAI M N, YE J Y, REN X H, et al. Kilogram-scale synthesis of phase-pure clinker phases with hydration reactivity benchmarks and CO_2 emissions constraints for machine-learning-assisted low-carbon clinker design[J]. *J Build Eng*, 2026, 120: 115395.
- [5] KE X Y, DUAN Y. A Bayesian machine learning approach for inverse prediction of high-performance concrete ingredients with targeted performance[J]. *Constr Build Mater*, 2021, 270: 121424.
- [6] CHEN W G, LIANG L, JIANG F M, et al. New opportunity: Materials genome strategy for engineered cementitious composites (ECC) design[J]. *Cem Concr Compos*, 2025, 159: 106009.
- [7] 杨嘉轩, 王仕奇, 卓峰谊, 等. 基于集成学习与多目标优化的石灰石煅烧黏土水泥智能配比设计[J]. *硅酸盐学报*, 2026, 54(3): 970–981. YANG Jiaxuan, WANG Shiqi, ZHUO Fengyi, et al. *J Chin Ceram Soc*, 2026, 54(3): 970–981.
- [8] CAI J M, HAN M X, YAN X R, et al. A process-synergistic active learning framework for high-strength Al-Si alloys design[J]. *npj Comput Mater*, 2025, 11: 228.
- [9] LIN M Y, SEVERSON K, GRANDGEORGE P, et al. Closed-loop optimization using machine learning for the accelerated design of sustainable cements incorporating algal biomatter[J]. *Matter*, 2025, 8(9): 102267.
- [10] GONÇALVES J P, HAN T H, SANT G, et al. Toward smart and sustainable cement manufacturing process: Analysis and optimization of cement clinker quality using thermodynamic and data-informed approaches[J]. *Cem Concr Compos*, 2024, 147: 105436.
- [11] BOIGER R, XI B, MIRON G D, et al. Machine learning-accelerated discovery of green cement recipes[J]. *Mater Struct*, 2025, 58(5): 173.
- [12] ASHRAF M, IQBAL M F, RAUF M, et al. Developing a sustainable concrete incorporating bentonite clay and silica fume: Mechanical and durability performance[J]. *J Clean Prod*, 2022, 337: 130315.
- [13] IQBAL M F, JAVED M F, RAUF M, et al. Sustainable utilization of foundry waste: Forecasting mechanical properties of foundry sand based concrete using multi-expression programming[J]. *Sci Total Environ*, 2021, 780: 146524.
- [14] IQBAL M F, LIU Q F, AZIM I, et al. Prediction of mechanical properties of green concrete incorporating waste foundry sand based on gene expression programming[J]. *J Hazard Mater*, 2020, 384: 121322.
- [15] 姚翊珩, 任雪红, 叶家元, 等. 基于半监督生成对抗网络的水泥熟料游离氧化钙数据增强方法及其有效性评价[J]. *硅酸盐学报*, 2025, 53(11): 3127–3135. YAO Xinyue, REN Xuehong, YE Jiayuan, et al. *J Chin Ceram Soc*, 2025, 53(11): 3127–3135.
- [16] WANG L J, CHEN H H, YANG L Q, et al. Single-atom catalysts property prediction via Supervised and Self-Supervised pre-training models[J]. *Chem Eng J*, 2024, 487: 150626.
- [17] LIU J L, WANG S Q, CHEN K Y, et al. Synthetic data augmentation and Integrated prediction framework for low-carbon recycled concrete: Leveraging CTGAN and multiple ML models stacking[J]. *J Clean Prod*, 2026, 543: 147599.
- [18] YANG J L, FAN Y C, HANG Z Y, et al. Data augmentation-aided machine learning prediction of 28-day compressive strength of CNT/cement composites[J]. *Neural Comput & Applic*, 2025, 37(5): 4009–4033.
- [19] YAO X Y, REN X H, YE J Y, et al. Prediction of f-CaO content in cement clinker using a GRU-based deep learning model with masked-attention mechanism for incomplete DCS data[J]. *J Sustain Cem Based Mater*, 2025, 14(7): 1413–1426.
- [20] QU H L, HU T Y, QU J T, et al. A machine learning model based on GAN-ANN data augmentation for predicting the bond strength of FRP-reinforced concrete under high-temperature conditions[J]. *Compos Struct*, 2025, 369: 119321.
- [21] NGUYEN H A T, PHAM D H, AHN Y. Effect of data augmentation using deep learning on predictive models for geopolymer compressive strength[J]. *Appl Sci*, 2024, 14(9): 3601.
- [22] HAN Y X, LIU Y H, CHEN Q C. Data augmentation in material images using the improved HP-VAE-GAN[J]. *Comput Mater Sci*, 2023, 226: 112250.
- [23] SHEN W K, JIANG C X, JIN Z Q, et al. Can concrete pumping pressure be predicted from mixture composition Insights from machine learning on full-scale pumping experiments[J]. *Cem Concr Compos*, 2026, 167: 106458.
- [24] LYNGDOH G A, ZAKI M, KRISHNAN N M A, et al. Prediction of concrete strengths enabled by missing data imputation and interpretable machine learning[J]. *Cem Concr Compos*, 2022, 128: 104414.
- [25] QI C C, WU M T, ZHENG J S, et al. Rapid identification of reactivity for the efficient recycling of coal fly ash: Hybrid machine learning modeling and interpretation[J]. *J Clean Prod*, 2022, 343: 130958.
- [26] 中国建筑材料联合会. GB/T 17671—2021. 水泥胶砂强度检验方法 (ISO法)[S]. 北京: 中国标准出版社, 2021.
- [27] YUAN Y, WANG X Y, CHEN X, et al. Mathematical models of apparent viscosity as a function of water–cement/binder ratio and superplasticizer in cement pastes[J]. *Sci Rep*, 2023, 13: 22301.
- [28] KIM Y J, CHO B Y, LEE S J, et al. Investigation of rheological properties of blended cement pastes using rotational viscometer and dynamic shear rheometer[J]. *Adv Mater Sci Eng*, 2018, 2018: 6303681.
- [29] HUANG Y M, ZHANG J F, TZE ANN F, et al. Intelligent mixture

design of steel fibre reinforced concrete using a support vector regression and firefly algorithm based multi-objective optimization model[J]. *Constr Build Mater*, 2020, 260: 120457.

- [30] JIANG C Y, XIN Y J, LIU N, et al. Generation and selection: A self-iterative two-stage data augmentation method for automated ECG classification[J]. *IEEE J Biomed Health Inform*, 2026, 30(4): 3473–3484.
- [31] CHEN X, YANG H S, ZHANG H C, et al. Uncertainty-aware pseudo-label filtering for source-free unsupervised domain adaptation[J]. *Neurocomputing*, 2024, 575: 127190.
- [32] WU S, YAO X L, REN C Z, et al. Effect of iron on the preparation of iron-rich calcium sulfoaluminat cement using gypsum as the sole calcium oxide source and its incorporation into mineral phases[J]. *Constr Build Mater*, 2021, 290: 123214.
- [33] DAMIDOT D, RETTEL A. Effect of gypsum on CA and C₁₂A₇

hydration at room temperature[C]//Proceedings of the 11th International Congress on the Chemistry of Cement (ICCC), Durban, South Africa. 2003: 11–16.

作者贡献声明:

翟牧楠: 设计研究方法/实验方案/模型算法、撰写初稿、绘制图表

邱晓: 论文构思、提供资源、监督指导

叶家元: 论文构思、调查研究、审阅与修订

任雪红: 分析/综合数据并推导结论、获取资助

吴春丽: 软件使用、提供资源

张洪滔: 调查研究、分析/综合数据并推导结

崔文娟: 绘制图表、项目管理

颜景华: 验证、数据管理

张文生: 审阅与修订、获取资助、监督指导

Data-Augmented Modeling and Model Interpretability Analysis for Strength Prediction of Low-Carbon Clinker

ZHAI Munan¹, ZHI Xiao¹, YE Jiayuan¹, REN Xuehong¹, WU Chunli², ZHANG Hongtao¹, CUI Wenjuan¹, YAN Jinghua¹, ZHANG Wensheng¹

(1. China Building Materials Academy Co., Ltd., Beijing, 100024, China; 2. ZhongCun Big Data Technology Co., Ltd, Beijing 100024, China)

Extended Abstract

Introduction The clinker calcination process involves high-temperature solid/liquid reactions, and the controllable preparation of multiple clinker phase compositions, which further affect subsequent synergistic hydration reactions. The relationship between composition and 28 d compressive strength is not simply linear, but rather a complex mapping governed by the coupling of multiple processes. Conventional point-by-point experimental screening makes it difficult to cover a sufficiently broad compositional space within a limited timeframe. This restricts the exploration efficiency of novel low-carbon clinker. Machine learning models can be used to learn the composition and property relationships between clinker phase and strength, and to support the screening of candidate clinker compositions. However, under small data conditions, these models are susceptible to insufficient training distribution coverage, resulting in a unstable cross-system predictive performance. Therefore, this study was to develop a self-iterative data augmentation framework to address the above-mentioned issues. This framework was to expand effective supervised information to optimize the model's predictive performance in sparsely sampled compositional regions. Subsequently, model interpretability analysis was conducted to identify the key clinker phases affecting 28 d compressive strength prediction and their response patterns. This could provide a testable model basis for the optimization of low-carbon clinker compositions.

Methods This study defined each simulated clinker as one sample. The dataset contained 264 simulated clinker samples, with the compressive strengths ranging from 0 to 70.3 MPa. The model inputs consisted of 12 features, including 11 clinker phases and gypsum, while the output was the 28 d compressive strength. In the self-iterative data augmentation strategy, an artificial neural network (ANN) model was first trained using the experimental data. Virtual data were then generated under compositional constraints, and their 28 d compressive strengths were predicted by the model from the previous iteration to assign pseudo-labels. The virtual and experimental samples were subsequently combined for training, and the surrogate model was updated iteratively. Two sampling patterns were adopted for virtual sample generation, *i.e.*, 1) Pattern A: dense sampling within the experimental samples' compositional systems, and 2) Pattern B: generating virtual samples at a ratio of 1 : 1 from the in- and out-of-system regions of the experimental samples' compositional space. Before training, 36 experimental samples were set aside as a test set, which remained isolated throughout training, iteration and hyperparameter selection. The validation set consisted exclusively of real experimental samples, whereas virtual samples were included only in the training set. Model hyperparameters were determined based on the Bayesian optimization, with the validation-set mean squared error (MSE) as the selection criterion. Finally, the model predictive performance was evaluated by MSE, the coefficient of determination (R^2), mean absolute error (MAE), and mean absolute percentage error (MAPE). After the final surrogate model was determined, multilevel model interpretation was further performed *via* combining constrained permutation feature importance, accumulated local effects and SHapley additive explanations (SHAP).

Results and discussion The self-iterative augmentation strategy effectively fills sparsely sampled regions, while maintaining feasible-domain constraints. This provides richer training information for subsequent accuracy evaluation and interpretability analysis of the surrogate model. Meanwhile, the data augmentation process does not generate a large number of near-duplicate samples, and the training set retains the necessary heterogeneity. The inclusion of virtual samples reduces the model's sensitivity to randomness in data partitioning. Moreover, the overall test set error and the in-/out-of-system errors decrease progressively over successive iterations, indicating an improved cross-system generalization performance. Pattern B in Iteration II achieves a favorable balance between predictive accuracy and extrapolation stability. Its overall test set MSE is 9.8, with in-system and out-of-system MSE values of 6.9 and 11.3, respectively. The R^2 , MAE, and MAPE are 0.79, 2.36 and 5.11%, respectively. These results indicate that, under the current small data conditions, data augmentation can substantially improve the model's ability to identify performance trends. Subsequent interpretability analysis of the optimal surrogate model shows that different clinker phases can be classified as positive or negative driving phases. The analysis further indicates that when C_3A content exceeds approximately 10%, its positive driving effect on strength decreases rapidly. Although CA and $C_{12}A_7$ act as positive driving phases, their benefits diminish when their contents exceed 40%, and even shift to strongly negative responses.

Conclusions The self-iterative data augmentation framework proposed could expand a supervised training information for low-carbon clinker systems characterized by small datasets and sparse compositional spaces. The evaluation bias introduced by pseudo-labels could be also suppressed under a rigorous evaluation strategy. The results showed that data augmentation significantly improved the prediction accuracy and cross-system generalization capability of the ANN surrogate model for 28 d compressive strength. In Iteration II, Pattern B achieved the optimum overall performance under the current data conditions. Although further iterations reduced some in-system errors, an error rebound was observed for out-of-system samples. This indicated that diminishing marginal information could gain from pseudo-labels constrained the model's generalization performance. The combined use of constrained PFI, ALE, and SHAP revealed the decision-making basis of the surrogate model in terms of global sensitivity, average response trends, and sample-level contributions. The results indicated that C_4A_3 and CA were the main positive driving clinker phases for compressive strength enhancement, whereas C_3A , C_5S_2 and C_2S were the main negative factors. Moreover, $C_{12}A_7$ exhibited a pronounced threshold effect. These findings could provide a testable reference for subsequent composition screening, constraint-boundary definition and inverse design of low-carbon clinker.

Keywords low-carbon clinker; 28 d compressive strength; prediction model; self-iterative; data augmentation; model interpretability