doi:10.12068/j.issn.1005－3026.2026.20250128

# $CaF_2$基渣系电导率的机器学习预测及离子传输机制

姜周华[1,2,3]，陈 玺[1]，王 骜[1]，董艳伍[1,2,3]
（1. 东北大学 冶金学院，辽宁 沈阳 110819；2. 东北大学 多金属共生矿生态冶金教育部重点实验室，辽宁 沈阳 110819；3. 东北大学 数字钢铁全国重点实验室，辽宁 沈阳 110819）

**摘 要：**针对经验模型忽略组元非线性交互且无微观验证、纯机器学习模型受限于“黑箱”与数据稀缺的双重难题，构建“数据驱动—机理支撑—微观验证”多尺度研究体系．建立涵盖三元至五元渣系的标准化数据库，对比10种监督学习算法优选各体系最优模型．针对五元渣系纯机器学习局限，提出遗传算法修正光学碱度模型辅助的CatBoost（GANPL-CatBoost）耦合模型，该模型以CatBoost梯度提升框架为基础，嵌入修正光学碱度作为物理先验特征，并采用遗传算法优化样本权重与超参数，实现物理约束与数据驱动深度融合，平均绝对百分比误差（MAPE）降至9.431%．基于LAMMPS（large-scale atomic/molecular massively parallel simulator）分子动力学模拟，运用Born-Mayer-Huggins势与Nernst-Einstein关系，从微观尺度阐明$Ca^{2+}$主导的电荷传导机制、离子配位演化及组元交互本质．由此形成“宏观预测—微观阐释—多尺度验证”闭环，为$CaF_2$基渣系智能化设计提供高效工具．

**关 键 词：**电渣重熔；电导率预测；机器学习；分子动力学；光学碱度；$CaF_2$基渣系
**中图分类号：**TF 14 **文献标志码：**A **文章编号：**1005-3026(2026)06-0068-14

# Machine Learning Prediction of Electrical Conductivity and Ion Transport Mechanism in $CaF_2$-Based Slag System

*JIANG Zhou-hua*[1,2,3], *CHEN Xi*[1], *WANG Ao*[1], *DONG Yan-wu*[1,2,3]
(1. School of Metallurgy, Northeastern University, Shenyang 110819, China; 2. Key Laboratory of Ecological Metallurgy of Multimetallic Mineral, Ministry of Education, Northeastern University, Shenyang 110819, China; 3. State Key Laboratory of Digital Steel, Northeastern University, Shenyang 110819, China. Corresponding author: JIANG Zhou-hua, E-mail: jiangzh@smm.neu.edu.cn)

**Abstract:** To address the limitations of empirical models, which ignore nonlinear interactions among components and lack microscopic validation, as well as the dual challenges of “black-box” nature and data scarcity faced by pure machine learning models, a multi-scale research framework of “data drive, mechanism support, and microscopic validation” was established. A standardized database covering ternary to quinary slag systems was constructed. Ten supervised learning algorithms were compared to identify the optimal model for each system. To overcome the limitations of pure machine learning in the quinary system, a coupled GANPL-CatBoost model was proposed. This model, based on the CatBoost gradient boosting framework, incorporated corrected optical basicity as a physical prior feature and utilized a genetic algorithm to optimize sample weights and hyperparameters, achieving deep integration of physical constraints and data-driven. The mean absolute percentage error (MAPE) for prediction was reduced to 9.431%. Furthermore, molecular dynamics simulations were conducted using the large-scale atomic/molecular massively parallel simulator (LAMMPS). Based on the Born-Mayer-Huggins potential and the Nernst-Einstein relationship, the dominant charge conduction mechanism of $Ca^{2+}$, the evolution of ion coordination, and the essence of component interactions

**收稿日期：**2025-10-30
**基金项目：**高校学科人才引进计划项目(NB21001)；辽宁省科技计划联合项目(2023JH2/101700302).
**作者简介：**姜周华(1963—)，男，浙江萧山人，东北大学教授，博士生导师.
**通信作者：**姜周华，E-mail:jiangzh@smm.neu.edu.cn.

were systematically quantified at the microstructural level. Consequently, a closed-loop system of “macroscopic prediction, microscopic elucidation, and multi-scale validation” was formed, providing an efficient tool for the intelligent design of $CaF_2$–based slag systems.

**Key words**: electroslag remelting; electrical conductivity prediction; machine learning; molecular dynamics; optical basicity; $CaF_2$–based slag system

随着现代电渣重熔(ESR)技术向高端特殊钢和航空航天材料制备领域的深入发展，对$CaF_2$基渣系设计的精确性要求不断提升[1−2]. 电导率作为控制重熔过程电能输入、熔池形状及产品质量的核心物理参数，其准确预测对于渣系优化设计具有决定性意义[3−5]. 然而，从二元、三元渣系向四元、五元及多元渣系的扩展过程中，现有预测模型面临严峻挑战：早期建立的三元渣系电导率预测基础，及通过几何模型对三元渣系预测精度的提升[6−8]，虽为研究奠定基础，但修正的光学碱度理论仍未突破经验拟合模型的固有局限[9−11]——实验数据量有限且覆盖范围窄导致泛化能力不足，难以捕捉多元渣系组元间复杂非线性交互作用，更缺乏微观层面物理合理性验证[12−14].

近年来，机器学习凭借强大的非线性建模能力，为多元渣系性质预测提供了新路径，但其“黑箱”特性使其虽能实现较高预测精度，却缺乏物理解释性[15−19]；同时，高氟渣系实验数据稀缺且分布不均，直接应用易导致温度敏感性异常偏低等违背物理直觉的结果[9,20−21]. 而分子动力学(MD)模拟作为连接原子尺度微观机理与宏观性质的桥梁，可通过第一性原理计算离子扩散系数、结合Nernst−Einstein关系预测电导率，却尚未与机器学习电导率模型形成系统性结合[22−23].

本文构建了“数据驱动—机理支撑”的完整研究框架. 采用系统性多尺度建模验证方法：首先全面收集并整合$CaF_2$基渣系电导率数据，建立覆盖三元至五元渣系的综合数据库；通过跨渣系对比分析多种机器学习算法的预测性能，精准筛选出各渣系下的最优模型；针对五元渣系预测精度下降问题，提出融合修正光学碱度机理的定向数据扩充策略，构建GANPL−CatBoost耦合模型；引入分子动力学模拟开展原子尺度验证，明确$Ca^{2+}$离子主导的电荷传导机制，为机器学习的特征分析结果提供坚实的物理机理支撑.

## 1　方法论

### 1.1　数据收集与处理

为构建涵盖$CaF_2$基渣系组元成分、温度与电导率关联规律的综合数据库，本文通过多渠道整合数据，来源包括已发表学术文献、相关专利、实测数据及多组实验室精准测试数据. 采用人工逐条核验方式，系统剔除异常数据并深度挖掘文本中隐藏的关键信息，确保数据库的准确性与完整性. 最终数据库按研究核心目标划分为3个类别，即渣系化学组成、测试温度及对应电导率数值.

### 1.2　基于修正光学碱度的机理模型

电导率随温度的变化关系常用Arrhenius公式来表述：

$$\ln\kappa=\ln A-\frac{E}{RT}. \tag{1}$$

式中：$\kappa$为电导率；$A$为指前因子；$E$为活化能；$T$为热力学温度.

一般来说，Arrhenius公式的指前因子$A$和活化能$E$之间满足以下关系：

$$\ln A=mE+n. \tag{2}$$

式中，$m$，$n$为和温度相关的拟合常数，一般根据所选择的数据进行拟合得到.

为了保证较高的精度，当数据的选择范围和数量发生改变时，一般需要重新拟合. 式(2)描述了温度补偿效应，这是一种符合Arrhenius方程性质的普遍规律，适用于反应动力学速率常数、电导率、黏度和扩散系数等物理性质.

通过不同温度和组分条件下的数据拟合，可以得到$\ln\kappa$和光学碱度$\Lambda$的线性关系，一般来说，这种关系能够较好地描述系统的电导率.

$$E=a\cdot\Lambda+b. \tag{3}$$

式中，$a$，$b$均为常数，需要依据数据的选择进行拟合. 在此基础上，Mills提出了修正的光学碱度公式和相关成分的光学碱度(见表1)，式(4)计算出的光学碱度不是单调递增的，更能满足电导率模型的需要[24].

$$\Lambda=\frac{\sum x_i n_i \Lambda_i}{\sum x_i n_i}. \tag{4}$$

式中：$\Lambda$为光学碱度，其值见表1；$x_i$为组元$i$的物质的量分数；$n_i$为组元$i$的原子个数.

表1　各组元的光学碱度

Table 1　Optical basicity of each component

| 组元 | CaO | MgO | $Al_2O_3$ | TiO | $SiO_2$ | $CaF_2$ |
|---|---|---|---|---|---|---|
| 光学碱度 | 1 | 0.78 | 0.6 | 0.61 | 0.48 | 1.2 |

Zhang等[25]对CaO−MgO−$Al_2O_3$−$SiO_2$和CaO−$Al_2O_3$−$SiO_2$渣系的光学碱度进行了修正，当$x_{CaO}<x_{Al_2O_3}$，$x_{CaO}+x_{MgO}\geqslant x_{Al_2O_3}$时，

$$\Lambda^{corr}=\frac{0.78\times\left(x_{MgO}+x_{CaO}-x_{Al_2O_3}\right)+0.6\times3x_{Al_2O_3}+0.48\times2x_{SiO_2}}{\left(x_{MgO}+x_{CaO}-x_{Al_2O_3}\right)+3x_{Al_2O_3}+2x_{SiO_2}}. \tag{5}$$

对于四元及以上的渣系，目前只有Hara等[26]提出的经验公式：

$$K=100\exp\left(1.911-1.38N-5.69N^2\right)+0.39\left(T-1\,973\right). \tag{6}$$

其中：$K$为电导率；$N$=$x_{Al_2O_3}$+0.8$x_{MgO}$+0.2$x_{CaO}$+0.75$x_{SiO_2}$；$T$为热力学温度，K.

需要注意的是，现有相关模型的构建均以研究者选定的特定实验数据集为基础进行拟合标定.当将某一学者提出的模型应用于预测其他文献中的电导率数据时，往往会产生显著偏差，表明当前基于光学碱度拟合与修正的机理模型存在固有局限性.

### 1.3 机器学习算法

在$CaF_2$基渣系电导率预测中，传统机理模型难以捕捉组元与温度对电导率的复杂非线性关联，而机器学习算法凭借强大的非线性映射能力，可通过多源实验数据的深度训练，精准构建“温度−渣系成分−电导率”的关联模型.考虑到实验数据样本有限，且电导率预测属于需已知实测值的回归任务，监督学习算法成为最优技术路径.

CatBoost[27]作为基于梯度提升框架的集成决策树算法，凭借其对类别型特征和高维数据的出色适配性，成为本研究的核心候选算法之一.其核心数学表达可概括为

$$F_m(x)=F_{m-1}(x)+\gamma_m h_m(x). \tag{7}$$

其中：$F_m(x)$为第$m$轮的预测结果；$h_m(x)$为基于残差的弱学习器；$\gamma_m$为学习率.

通过多模型训练和对比，发现CatBoost在电渣重熔$CaF_2$基渣系电导率的预测中表现出色，尤其在处理复杂的非线性关系和类别型数据时，能够有效提高模型的精度和稳定性.

其他算法包括ElasticNet(结合$L_1$和$L_2$正则化)、决策树(DT)、随机森林(RF)、Extra Trees、梯度提升(gradient boosting，GB)、XGBoost、AdaBoost、Bagging和KNN等.为评估模型性能，选择决定系数($R^2$)、均方误差(MSE)、平均绝对误差(MAE)和平均绝对百分比误差(MAPE)作为评估指标.

### 1.4 机理耦合机器学习模型

针对高氟渣电导率监测等高温冶金实验存在的数据获取成本高、样本有限且分布不均的问题，机理模型可在多温度−成分条件下生成大量符合物理规律的模拟数据，既能补充机器学习训练集以缓解数据稀缺问题，又能避免模型在数据稀疏区域过拟合.

本文先针对三元、四元及五元$CaF_2$基渣系，选取RF，XGBoost，CatBoost等10种适用于小数据集的机器学习算法，通过定量评估特征重要性，系统分析温度与各组分对电导率预测的贡献.既为后续特征选择与模型优化提供支撑，也阐明成分−温度调控电导率的主导机制、提升模型物理可解释性.研究发现纯机器学习对五元渣系电导率预测存在显著局限，故提出半经验模型与机器学习深度融合策略.

基于上述理论，本文构建数据扩充策略：在$CaF_2$基五元渣系成分空间内，依电渣重熔工业应用确定组元边界，温度重点覆盖1 773~1 923 K工业区间并兼顾1 623~2 173 K全范围；为避免机理模型系统性偏差主导建模，将扩充数据量控制在原始数据的1~3倍(本研究为2 200组)，同时通过确保成分总和100%、验证成分−温度空间分布合理性、检验数据正态性及保护原始实验规律，保障数据可靠性.

在特征工程与模型训练阶段，机理模型为机器学习提供物理化学约束：助力识别电导率敏感变量、揭示特征交互机制，并提供边界条件以确保预测符合物理规律.通过对比算法性能筛选最优工具，对扩充数据进行标准化预处理，采用$k$折交叉验证保障模型稳定性与泛化能力，结合网格搜索与随机搜索优化超参数.

模型验证采用“内部交叉验证−外部独立测试集验证−物理一致性检验”多层次体系，通过$R^2$，MSE，MAE，MAPE等指标评估性能，同时验证预测结果与已知物理化学规律的一致性.该耦合策略实现了机理先验知识与数据驱动学习的有机融合，为$CaF_2$基五元渣系电导率精确预测提供了可靠技术方案(图1).

### 1.5 分子动力学验证方法

本文建立了系统性的分子动力学模拟框架以获取高精度微观结构数据.采用LAMMPS软件包对$CaF_2$基三元、四元和五元渣系进行原子尺

度建模,首先基于 Lennard-Jones 势函数构建初始随机结构模型,通过查询或经验公式计算各组元的密度,进而计算盒子边长,然后在盒子内生成 5 000 个原子(Ca,F,Al,O,Si,Mg),设定各原子的离子电荷分别为+2,-1,+3,-2,+4,+2.随后采用玻恩-迈耶-哈金斯(Born-Mayer-Huggins,BMH)势函数结合长程静电相互作用(Ewald 求和)进行精确描述.

$$U_{ij}(\boldsymbol{r}) = \frac{q_i q_j}{r_{ij}} + A_{ij}\exp\left(-\frac{r_{ij}}{\rho_{ij}}\right) - \frac{C_{ij}}{r_{ij}^6}. \tag{8}$$

式中:$U_{ij}$为原子间势能;$r_{ij}$为离子间距离,是$\boldsymbol{r}$矩阵中的元素;$q_i$为第 $i$ 个离子的电荷;$q_j$为第 $j$ 个离子的电荷;$A_{ij}$为排斥项参数;$\rho_{ij}$为排斥项特征长度;$C_{ij}$为范德华力参数.然后计算了库仑势和原子间势.本文使用的所有系统的原子间势能如表 2 所示.

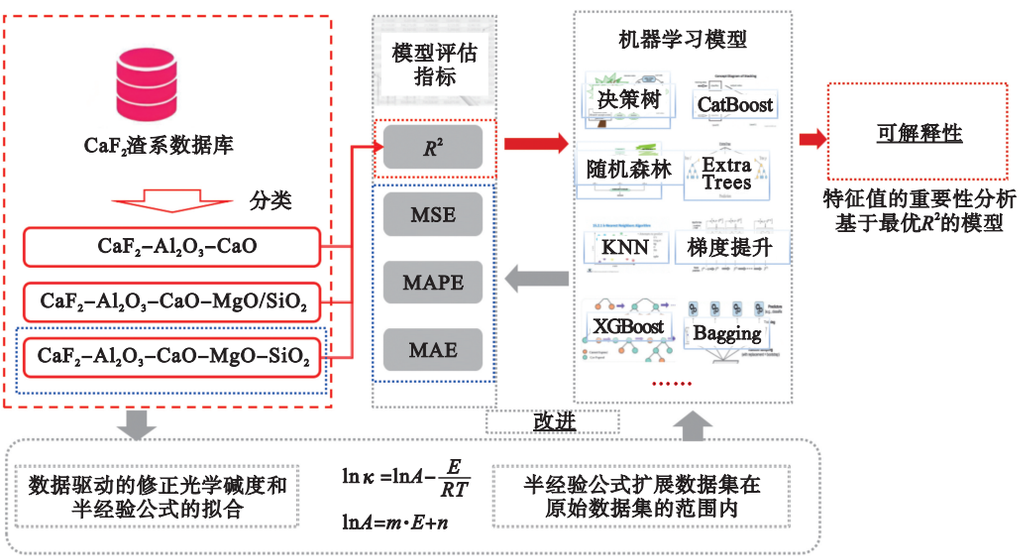



图 1　基于特征重要性和数据驱动半经验展开的炉渣电导率预测解释框架

Fig. 1　Interpretation framework for slag electrical conductivity prediction based on feature importance and data-driven semi-empirical expansion

表 2　用于所有系统的原子间势能

Table 2　Interatomic potentials used for all systems

| 原子对 | $A_{ij}$/eV | $\rho_{ij}$/nm | $C_{ij}^{\ominus}$/(eV·nm) |
|---|---|---|---|
| Al-Al | 4 142.15 | 0.016 | 0 |
| Al-Ca | 36 918.57 | 0.016 | 0 |
| Al-F | 59 486.50 | 0.016 5 | 0 |
| Al-O | 86 057.58 | 0.016 5 | 0 |
| Ca-Ca | 329 051.60 | 0.016 | 0.436 |
| Ca-F | 496 191.50 | 0.016 5 | 0.867 |
| Ca-O | 717 827.00 | 0.016 5 | 0.867 |
| Ca-Si | 26 674.68 | 0.016 | 0 |
| F-F | 730 722.80 | 0.017 | 1.734 |
| F-O | 1 046 135.40 | 0.017 | 1.734 |
| F-Si | 43 406.00 | 0.016 5 | 0 |
| O-O | 1 497 049.00 | 0.017 | 1.734 |
| O-Si | 62 794.37 | 0.016 5 | 0 |
| Si-Si | 4 142.15 | 0.016 | 0 |

参数经过高温熔体状态优化以确保准确描述离子间相互作用.模拟采用严格的三阶段协议:第一阶段将渣系从初始目标温度快速升温至 5 000 K 并运行 30 ps 以充分混合并消除初始构型影响;第二阶段以恒定冷却速率从 5 000 K 线性冷却至渣系设定温度并运行 50 ps,模拟实际快冷过程;第三阶段在目标温度下采用 NVT 系统恒温运行 500 ps,每 1 ps 保存一次原子坐标,确保获得充分的统计采样.整个模拟过程均施加三维周期性边界条件,长程静电相互作用采用质点-质点-质点-网格(particle-particle-particle-mesh,PPPM)程序方法求解,计算精度控制在 $10^{-5}$ 以内,以确保数值稳定性与结果可靠性.

## 2　结果与讨论

### 2.1　数据库特征分析

本文建立了 $CaF_2$ 基渣系的标准数据库.根据研究目标,数据库按渣系组成($CaF_2$,CaO,$Al_2O_3$,$SiO_2$,MgO)、温度以及对应的电导率值进行分类.表 3 呈现了各特征变量的统计信息,数据库涵盖了较宽的组成范围.

为明晰数据分布特征及其对机器学习建模的影响,本研究绘制了各特征变量的概率密度

(图2).从图中可见,$CaF_2$,CaO,$Al_2O_3$呈多峰分布:$CaF_2$在质量分数30%~40%与50%~60%区间分布呈双峰,对应工业低含量$CaF_2$复杂渣系与高含量$CaF_2$传统渣系两类典型场景;CaO,$Al_2O_3$的多峰分布则分别匹配不同碱度调控策略与网络结构需求.相比之下,$SiO_2$和MgO的分布高度向低值区偏斜,约70%的$SiO_2$质量分数低于10%,约60%的MgO质量分数低于10%,此特征既反映二者在$CaF_2$基渣系中作为次要添加剂的定位,也为模型在这两种成分高含量区域的预测带来挑战.温度数据约60%集中于1 773~1 873 K的窄范围(对应典型电渣重熔操作温度),但该集中分布可能低估温度特征在模型中的重要性.此外,工业相关的四元、五元复杂渣系存在更突出的数据稀缺问题.

表3 $CaF_2$基渣系组成、温度及对应电导率的统计分析

Table 3 Statistical analysis of composition, temperature, and corresponding electrical conductivity for $CaF_2$-based slag systems

| 统计量 | 渣系组分质量分数/% | | | | | 温度 | 电导率 |
|---|---|---|---|---|---|---|---|
| | $CaF_2$ | CaO | $Al_2O_3$ | $SiO_2$ | MgO | K | $S\cdot cm^{-1}$ |
| 均值 | 42.30 | 22.10 | 24.70 | 8.65 | 10.80 | 1 831 | 2.89 |
| 标准差 | 21.57 | 12.11 | 18.73 | 12.65 | 13.63 | 94.00 | 3.83 |
| 最小值 | 1.00 | 0 | 0 | 0 | 0 | 1 623 | 0.09 |
| 25% | 28.00 | 12.58 | 10.00 | 0 | 0 | 1 773 | 0.73 |
| 50% | 46.87 | 20.00 | 22.94 | 3.80 | 5.00 | 1 858 | 2.03 |
| 75% | 58.50 | 30.98 | 32.20 | 10.00 | 18.00 | 1 873 | 3.04 |
| 最大值 | 90.00 | 56.00 | 78.00 | 59.00 | 59.00 | 2 173 | 26.55 |

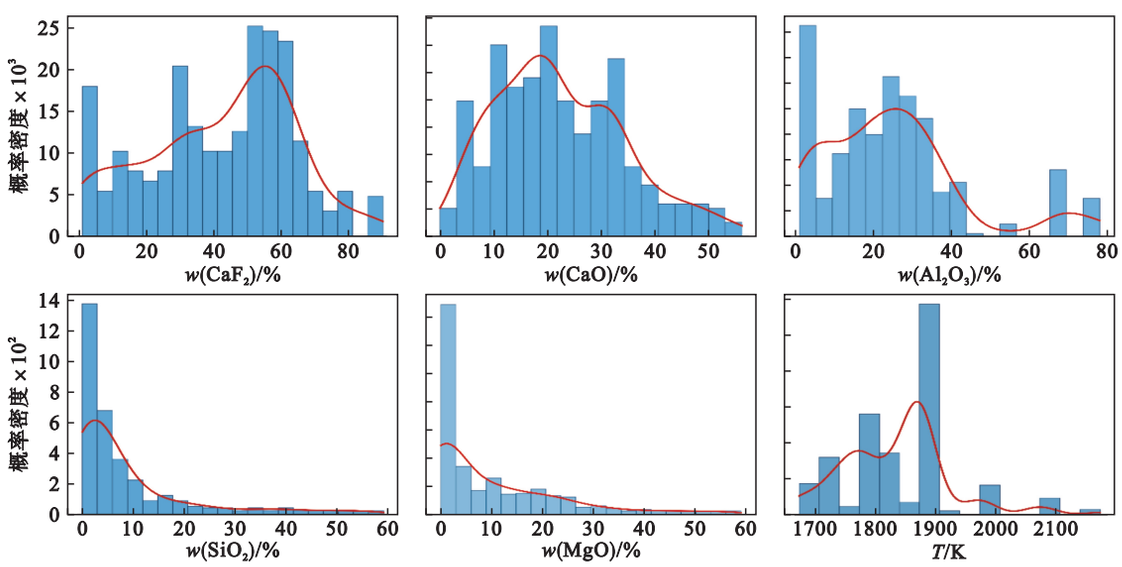



图2 电导率数据库中各特征变量的概率密度

Fig. 2 Probability density of each feature variable in electrical conductivity database

上述数据质量稀缺带来的挑战,为后续采用修正光学碱度理论进行数据扩充、选择对分布形态不敏感的集成学习算法、引入分子动力学模拟开展微观机理验证提供了充分必要性.通过机理-数据融合建模策略,可有效缓解单纯数据驱动方法在复杂多元体系中的局限性.

## 2.2 机器学习模型性能对比

为系统评估不同机器学习算法在$CaF_2$基渣系电导率预测中的性能,本研究针对三元($CaF_2$-$Al_2O_3$-CaO)、四元($CaF_2$-$Al_2O_3$-CaO-MgO/$SiO_2$)和五元($CaF_2$-$Al_2O_3$-CaO-MgO-$SiO_2$)渣系分别构建了预测模型.选择RF,XGBoost,CatBoost,ElasticNet,DT,Extra Trees,GB,AdaBoost,Bagging和KNN等10种在小数据集上表现优异的算法进行对比,采用决定系数($R^2$)作为主要评价指标,并通过MSE,MAE和MAPE进行综合性能评估.

对于$CaF_2$-$Al_2O_3$-CaO三元渣系,图3显示RF模型表现最优,$R^2$值达0.959 2,显著优于XGBoost的$R^2$值(0.871 3)和CatBoost的$R^2$值(0.887 6).集成学习方法普遍优于单一算法,表明组合多个弱学习器能有效处理渣系组成与电

导率间的复杂非线性关系.RF模型通过Bootstrap抽样和随机特征选择构建500棵决策树集成模型，有效避免了过拟合问题.线性模型ElasticNet的$R^2$仅为0.674 3，证实了非线性建模方法的必要性.

基于RF模型的特征变量重要性分析（表4）显示：$CaF_2$重要性最高（0.438），反映其作为主要导电介质的核心地位；$Al_2O_3$重要性次之（0.386），体现其通过形成$[AlO_4]^-$和$[AlO_6]^{3-}$聚合网络对离子迁移率的显著影响；CaO重要性为0.110，主要通过调节光学碱度和提供$Ca^{2+}$离子发挥作用；温度重要性最低（0.066），主要归因于训练数据中温度分布过于集中（75%数据点在1 850~1 873 K），导致模型难以充分学习温度与电导率的依赖关系.

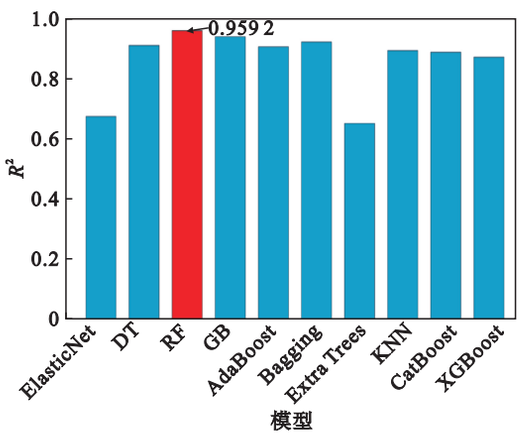



图3　$CaF_2$–$Al_2O_3$–CaO三元渣系中不同模型的$R^2$对比

Fig. 3　Comparison of $R^2$ values among different models for $CaF_2$–$Al_2O_3$–CaO ternary slag system

表4　基于RF模型的特征变量重要性分析

Table 4　Feature variable importance analysis based on RF model

| 特征变量 | $w(CaF_2)$ | $w(Al_2O_3)$ | $w$(CaO) | 温度 |
|---|---|---|---|---|
| 重要性 | 0. 438 | 0. 386 | 0. 110 | 0. 066 |

进入$CaF_2$–$Al_2O_3$–CaO–MgO/$SiO_2$四元渣系后，建模复杂度显著增加，XGBoost模型展现最佳性能，$R^2$=0.871 3（图4），相比三元渣系预测性能下降（$R^2$从0.959 2降至0.871 3），反映了引入第四组元后组分间相互作用复杂性的增加.XGBoost凭借其GB算法的序列学习机制和对分类特征的优秀处理能力脱颖而出.

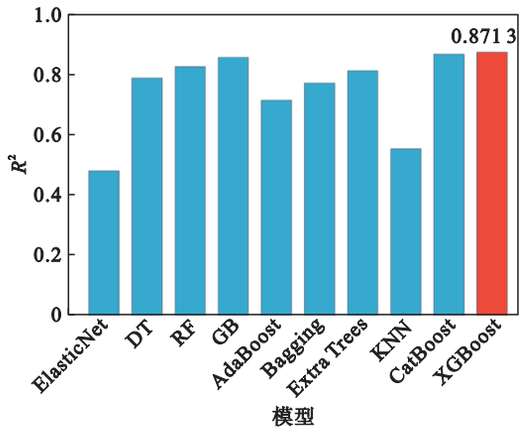



图4　$CaF_2$四元渣系中不同模型的$R^2$对比

Fig. 4　Comparison of $R^2$ values among different models for $CaF_2$–based quaternary slag system

从表5可以看出，基于XGBoost的特征变量重要性分析显示显著变化.温度重要性从三元渣系的0.066大幅跃升至0.313，成为最重要特征，根本原因在于四元渣系温度分布更加分散（范围扩展至1 623~2 173 K，标准差增至108 K），使模型能更充分学习温度对电导率的影响规律.$CaF_2$重要性虽有所下降但仍保持高位（0.291），体现其作为基础导电介质的稳定地位.新引入的MgO和$SiO_2$重要性分别为0.099和0.098，数值接近但机理不同：MgO在质量分数低（<15%）时通过释放$Mg^{2+}$离子增加载流子浓度，其强碱性破坏$Al_2O_3$形成的聚合阴离子网络，促进$[AlO_4]^-$等简单离子解离，提高总体电导率；$SiO_2$添加促进$[AlO_4]^{5-}$，$[AlO_6]^{3-}$和$[AlO_2F_2]^{5-}$等铝硅酸盐复合离子形成，这些大尺寸、高电荷复合离子迁移率较低，导致$Al_2O_3$对电导率的负面影响被相对削弱（重要性降至0.095）.从熔体结构角度，$SiO_2$倾向形成$[SiO_4]^{4-}$硅氧四面体网络，显著增加熔体黏度，但当CaO含量足够时，$Ca^{2+}$离子可部分取代网络中的$Si^{4+}$，破坏Si—O—Si桥键，这种“网络破坏效应”在一定程度上抵消了$SiO_2$的负面影响.

表5　基于XGBoost的特征变量重要性分析

Table 5　Feature variable importance analysis based on XGBoost

| 特征变量 | 温度 | $w(CaF_2)$ | $w$(CaO) | $w$(MgO) | $w(SiO_2)$ | $w(Al_2O_3)$ |
|---|---|---|---|---|---|---|
| 重要性 | 0. 313 | 0. 291 | 0. 105 | 0. 099 | 0. 098 | 0. 095 |

最具挑战性的$CaF_2$–$Al_2O_3$–CaO–MgO–$SiO_2$五元渣系代表了电渣重熔工业实践中最常用的复杂渣系配方.经10种模型全面对比，CatBoost模型凭借其类别特征处理能力和有序编码机制获得了最高$R^2$值0.790 9（图5），但相较四元渣系，其预测性能进一步下降了约10个百分点.

CatBoost采用对称树结构和梯度偏差估计技术，能更好地处理类别特征间的复杂交互，并通过有序提升方式减少过拟合风险.

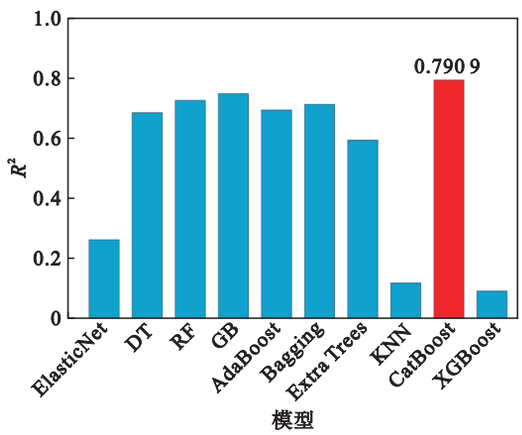



图5 $CaF_2$–$Al_2O_3$–CaO–MgO–$SiO_2$五元渣系中不同模型的$R^2$对比

Fig. 5 Comparison of $R^2$ values among different models for $CaF_2$–$Al_2O_3$–CaO–MgO–$SiO_2$ quinary slag system

由表6可揭示出令人意外的贡献模式：MgO重要性跃升至最高值0.289，超越传统主导因子$CaF_2$(0.197)，深层原因在于MgO展现出极强的浓度依赖性和双重作用机制.当MgO质量分数较低(<10%)时，$Mg^{2+}$离子插入硅氧网络间隙，破坏Si—O—Si桥键，有效降低熔体黏度并提高离子迁移率；当MgO质量分数增至中等水平(10%~30%)时，与$SiO_2$发生显著化学反应，形成镁橄榄石($Mg_2SiO_4$)和顽火辉石($MgSiO_3$)等高熔点化合物，这些化合物的生成不仅消耗部分$SiO_2$，降低硅氧网络聚合度，其高温稳定性还对熔体流变行为产生重要影响.温度重要性为0.241，相比四元体系略有下降但仍保持较高水平，反映温度在复杂多元渣系中对多种化学平衡和物理过程的综合调控作用.$Al_2O_3$重要性降至0.113，归因于在MgO和$SiO_2$协同作用下，铝氧网络的影响被进一步稀释.$SiO_2$重要性急剧下降至0.060，可从2个层面解释：数据层面，五元渣系中$SiO_2$分布更向低值集中，限制了模型对其高含量效应的学习能力；物理化学层面，在MgO–$SiO_2$–CaO三元氧化物协同作用下，$SiO_2$单独的网络形成能力被显著抑制，其对电导率的影响更多通过与MgO的化合反应间接体现.

表6 基于CatBoost的特征变量重要性分析

Table 6 Feature variable importance analysis based on CatBoost

| 特征变量 | $w$(MgO) | 温度 | $w$($CaF_2$) | $w$($Al_2O_3$) | $w$(CaO) | $w$($SiO_2$) |
|---|---|---|---|---|---|---|
| 重要性 | 0.289 | 0.241 | 0.197 | 0.113 | 0.100 | 0.060 |

然而，五元渣系建模结果暴露出纯机器学习方法的严重局限性.尽管CatBoost在交叉验证中取得相对较高的$R^2$值(0.790 9)，但误差分析揭示了深层问题：MSE=5.54，MAE=0.752，MAPE达88.696%，这意味着模型平均预测误差接近真实值的90%，在工程应用中完全不可接受.残差分析发现，模型在数据稀疏的$SiO_2$高质量分数(>20%)、MgO高质量分数(>35%)以及极端温度(<1 650 K或>2 050 K)区域出现严重预测偏差，部分预测值甚至出现物理上不合理的负电导率.区域性差异的根本原因在于训练数据分布不均：工业常用组成范围内数据密度相对较高(平均每组成网格内1.2~2.5个数据点).此外，极端条件下存在的复杂组分协同效应和竞争机制，如高含量MgO–$SiO_2$化合反应形成难熔相、$Al_2O_3$–$SiO_2$聚合竞争导致网络结构剧变、CaO对硅氧网络的非线性破坏作用等，这些多元组分协同效应的非单调性超出传统机器学习算法的描述范围.需指出的是，工业电渣重熔实践中很少使用这些极端组成配方，因高含量$SiO_2$或高含量MgO渣系会导致熔点过高、流动性差、电导率过低等工艺问题，实际生产中渣系组成通常限制在相对温和范围内.这表明，虽然纯机器学习方法在极端条件下存在外推风险，但在实际工程应用的常规参数范围内具有良好的预测能力和实用价值.

通过3个不同复杂度渣系的系统性对比分析，可清晰观察到机器学习模型性能递减趋势：三元渣系($R^2$=0.959 2)→四元渣系($R^2$=0.871 3)→五元渣系($R^2$=0.790 9)，预测精度每增加一个组分平均下降约8~10个百分点，预测误差呈指数级增长.这一趋势反映多个深层次科学问题：①数据稀疏性的指数恶化.随着特征维度从3增至5，所需样本数量按指数规律增长(维数灾难)，而实际可获得的实验数据量相对有限，导致样本密度在高维空间中急剧下降；②组分相互作用的复杂化.从三元渣系的3个二元相互作用发展到五元渣系的10个二元、10个三元、5个四元和1个五元相互作用，相互作用项数量呈组合爆炸式增长，超出传统机器学习算法建模能

力；③物理约束的缺失.纯数据驱动方法无法保证预测结果满足基本物理化学定律(如电中性原理、能量守恒、相图约束等)，在数据稀疏区域易产生不合理外推结果.这些根本性挑战为后续采用修正光学碱度理论进行机理指导的数据扩充策略、构建物理约束的机器学习模型(GANPL-CatBoost)以及引入分子动力学模拟进行微观机理验证提供了充分而迫切的必要性论证.只有通过机理-数据-验证三位一体的建模范式，才能突破纯机器学习方法在复杂多元体系中的固有局限，实现既高精度又物理可靠的电导率预测.

### 2.3　基于修正光学碱度的机理-机器学习耦合模型

#### 2.3.1　修正光学碱度机理模型的建立

针对前述机器学习模型在复杂多元体系中的局限性，首先建立了基于修正光学碱度理论的物理机理模型，为后续的数据扩充和机理-数据融合提供理论基础.根据式(1)~式(3)拟合得到本文适用的公式.由于式(4)已被验证具有较大的局限性，本文基于系统的实验数据分析和理论推导，对光学碱度计算公式进行了重要修正.考虑到不同组分在熔体中的实际配位环境和相互作用强度，修正后的光学碱度表达式为

$$\Lambda^{\mathrm{corr}}=\frac{1.4x_{\mathrm{CaO}}\Lambda_{\mathrm{CaO}}+9x_{\mathrm{Al_2O_3}}\Lambda_{\mathrm{Al_2O_3}}+3x_{\mathrm{CaF_2}}\Lambda_{\mathrm{CaF_2}}+4.3x_{\mathrm{SiO_2}}\Lambda_{\mathrm{SiO_2}}+2x_{\mathrm{MgO}}\Lambda_{\mathrm{MgO}}}{2x_{\mathrm{CaO}}+5x_{\mathrm{Al_2O_3}}+3x_{\mathrm{CaF_2}}+3x_{\mathrm{SiO_2}}+2.4x_{\mathrm{MgO}}}. \tag{9}$$

该修正公式中权重因子的确定遵循“物理约束+数据优化”的双重策略.首先，基于以下物理化学原理设定初值范围：①$CaF_2$权重因子(1.0~2.0)：$F^-$的电负性(3.98)显著高于$O^{2-}$(3.44)，导致Ca—F键的离子性强于Ca—O键，$F^-$在熔体中更易解离为自由离子，故其有效“碱度贡献”应高于名义值，初值设为1.4；②$Al_2O_3$权重因子(5~12)：$Al^{3+}$的高电荷密度使其强烈倾向形成$[AlO_4]^-$，$[AlO_6]^{3-}$等聚合阴离子网络，单个$Al_2O_3$分子对熔体结构的影响远超其物质的量分数，初值设为9；③MgO权重因子(2~6)：$Mg^{2+}$离子半径(0.072 nm)接近$Ca^{2+}$(0.100 nm)但电荷密度更高，在破坏硅氧网络方面表现出比$Ca^{2+}$更强的“网络修饰”能力，初值设为4.3；④$SiO_2$权重因子(1~3)：在富$CaF_2$体系中，$F^-$对Si—O—Si桥键的破坏作用显著削弱了$SiO_2$的网络形成能力，其有效权重应低于传统硅酸盐体系，初值设为2.

通过遗传算法优化，获得了修正光学碱度模型的最优参数：

$$m=8.877\times10^{-5},n=-2.126,$$

$$a=322\,644.981,b=-214\,090.649.$$

由此构建的完整机理模型为

$$\ln A=8.877\times10^{-5}E\pm2.126, \tag{10}$$

$$E=322\,644.981\cdot\Lambda^{\mathrm{corr}}-214\,090.649. \tag{11}$$

该模型的物理意义明确：激活能$E$反映了离子跳跃所需克服的能垒，与光学碱度成正比，体现了熔体网络结构对离子迁移的阻碍作用；指前因子$A$则与激活能存在补偿关系，反映了熔体中可动离子的浓度和振动频率等动力学因素.

#### 2.3.2　机理模型的验证与数据扩充策略

为验证修正光学碱度模型的准确性和适用性，本文采用五元渣系的实验数据进行了系统测试.结果显示，修正模型的预测性能显著优于传统经验公式：MSE为4.832，MAE为0.871，MAPE为76.65%.相比之下，Hara等[26]提出的经验公式在同一数据集上的表现较差，MSE高达12.832，MAE为6.871，MAPE达到126.65%.详细的性能对比如表7所示.

**表7　修正光学碱度模型与传统经验公式的性能对比**

Table 7　Performance comparison between modified optical basicity model and traditional empirical formulas

| 模型类型 | MSE | MAE | MAPE/% |
|---|---|---|---|
| 修正光学碱度模型 | 4.832 | 0.871 | 76.65 |
| Ogino经验公式 | 12.832 | 6.871 | 126.65 |
| GANPL-CatBoost模型 | 0.063 | 0.012 | 9.431 |

这一对比充分证明了修正光学碱度理论在描述$CaF_2$基多元渣系电导率方面的优越性.尽管修正机理模型在物理解释性方面具有明显优势，但其预测精度仍有提升空间，特别是在处理复杂的组分相互作用时存在一定局限.更为重要的是，该模型为解决机器学习建模中的数据稀疏性问题提供了有效途径.基于修正光学碱度理论的数据扩充策略如下：

首先，在原始实验数据覆盖的组成范围内，按照一定的网格密度生成新的组成点.具体地，在$CaF_2$，$Al_2O_3$，CaO，MgO，$SiO_2$ 5个维度上分别以2%~5%的间隔进行插值，在温度维度上以25~50 K的间隔进行扩展，确保生成的组成点满足物理约束条件(各组分质量分数和为100%，温度在合理工艺范围内).其次，利用修正光学碱度模型计算这些新组成点对应的电导率值.为确保扩充

数据的质量,对每个生成点都进行了物理合理性检验,剔除了预测电导率为负值或超出物理上限的异常点.最后,将扩充的数据与原始实验数据按适当比例混合,形成增强的训练数据集.本研究中,扩充数据量约为原始数据的2~3倍,既能有效改善数据分布的均匀性,又避免了扩充数据完全主导原始数据模式的风险.图6显示了扩充数据前后各特征变量的概率密度和核密度估计(KDE)比较.从图中可以看出,扩充后数据的分布总体上与原始数据相似,但是某些特征(例如$Al_2O_3$含量和温度)表现出更平滑的KDE曲线.总的来说,扩充后的数据成功地模拟了原始数据的趋势,同时提供了更大的分布范围,有效地解决了数据分布稀疏的问题.

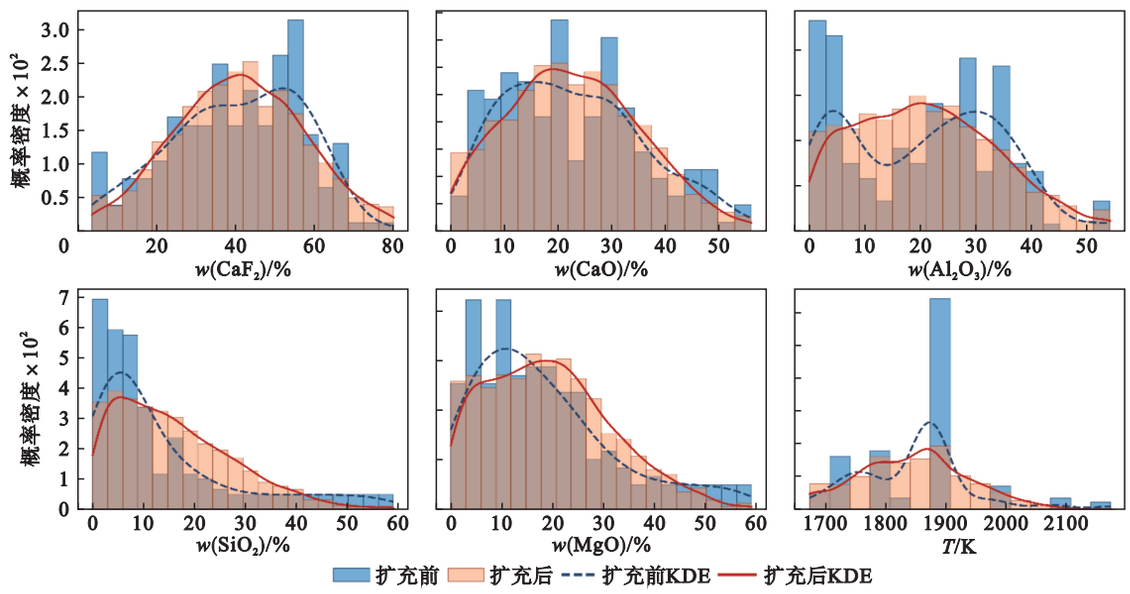



图6 扩充数据前后各特征变量的概率密度和核密度估计(KDE)对比

Fig. 6 Comparison of probability density and KDE for each feature variable before and after data expansion

在修正光学碱度理论指导下的数据扩充基础上,提出了GANPL-CatBoost(genetic algorithm with NPL-CatBoost)耦合模型.该模型的核心思想是采用在原始数据集表现最优的CatBoost模型对扩充后的数据集进行训练;然后,采用遗传算法调整超参数.结果显示:MSE=0.063,MAE=0.012,MAPE=9.431%(预测值和实验值对比如图7所示).

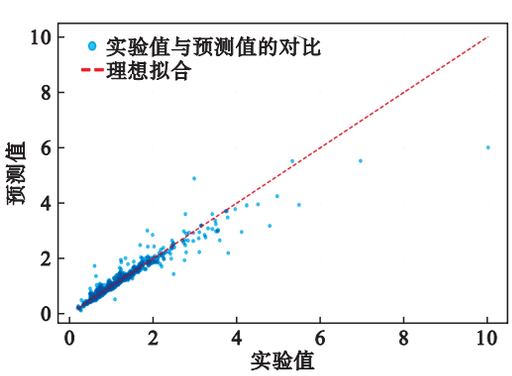



图7 GANPL-CatBoost模型预测值与实验值的对比

Fig. 7 Comparison of predicted values and actual values of GANPL-CatBoost model

相比原始数据集中机器学习方法的88.696%,预测精度提升了近10倍.更为重要的是,该模型在数据分布的各个区域都保持了相对稳定的预测性能.

### 2.3.3 模型验证与物理一致性分析

为全面评估GANPL-CatBoost模型的可靠性,从多个维度进行了深入的验证分析.首先,从30篇公开发表的文献和专利中独立选取了30组未参与训练的实验数据作为盲测样本,涵盖了三元、四元、五元等不同复杂度的渣系.这30组验证数据严格遵循独立性原则,确保未参与任何阶段的模型训练或参数调优过程.验证结果表明,GANPL-CatBoost模型在这些独立数据上的MAPE为9.231%,与5折交叉验证结果(9.431%)高度一致,充分证明了模型的良好泛化能力和预测稳定性.通过对30组验证数据的详细分析(如图8所示),发现模型预测值与实验值之间呈现良好的线性相关性,大部分数据点分布在理想拟合线附近,仅在少数极端组成条件下出现相对较大的偏差,但总体预测误差仍控制在工程可接受范围内.

更为重要的是,GANPL-CatBoost模型的预测结果展现出良好的物理一致性.系统分析模型在不同温度和组成条件下的预测行为,发现其能够准确重现电导率的Arrhenius温度依赖关系,预

测激活能值分布在40~80 kJ/mol范围内，与文献报道的$CaF_2$基熔体离子传导激活能典型值完全一致.模型对组成变化的响应严格遵循已知物理化学规律：$CaF_2$质量分数从30%增至60%时，预测电导率呈单调递增趋势，增幅达2~4倍；$Al_2O_3$质量分数从10%增至40%时，电导率显著下降，降幅可达50%~70%；温度从1 773 K升至1 923 K时，电导率按指数规律增长，温度系数与经典离子熔体理论预期高度吻合.这些预测趋势在定性和定量上均与实验观察结果保持良好一致性.GANPL-CatBoost模型在交叉验证中取得MSE=0.063，MAE=0.012，MAPE=9.431%的优异性能，相比传统半经验公式和纯机器学习方法分别实现约8倍和10倍精度提升，充分验证了机理指导数据扩充策略的显著优势.

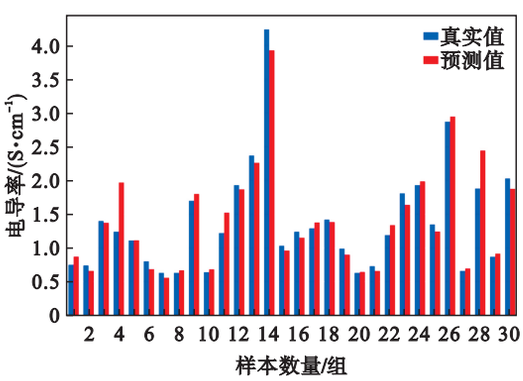



图8　外推实验实验值与预测值对比

Fig. 8　Comparison of actual values and predicted values in extrapolation experiment

然而，GANPL-CatBoost模型本质上仍属于宏观唯象模型，无法揭示电导率变化的微观物理机制，具体表现为4个关键科学问题的缺失：①各离子种类($Ca^{2+}$，$F^-$，$Mg^{2+}$，$Al^{3+}$，$O^{2-}$)对总电导率的贡献比例及浓度依赖性；②温度对离子扩散系数和配位环境的微观影响机制；③组分相互作用在原子尺度上对离子传导路径的影响规律；④机器学习中温度敏感性偏低的深层物理原因.这些微观机理层面的疑问正是引入分子动力学模拟验证的根本动机.通过原子尺度直接计算离子扩散系数、分析配位环境演化并利用Nernst-Einstein关系预测电导率，分子动力学模拟能够为GANPL-CatBoost模型提供坚实的微观物理基础，构建“宏观建模-微观验证”的多尺度预测体系，确保模型既具有工程实用性又具有深刻科学洞察力.

## 2.4　分子动力学方法验证结果

为验证GANPL-CatBoost机器学习模型的物理合理性并揭示$CaF_2$基渣系电导率的微观机理，本研究基于Born-Mayer-Huggins势能函数，采用LAMMPS分子动力学软件对9组代表性渣系组成进行了原子尺度模拟计算.通过Einstein关系式计算各离子的扩散系数，并利用Nernst-Einstein关系式预测电导率，从微观物理层面验证了宏观机器学习模型的预测准确性，为模型的工程应用提供了坚实的理论基础.

### 2.4.1　分子动力学方法的有效性验证

为确保分子动力学(MD)方法模拟结果的可靠性，首先需要验证MD方法在$CaF_2$基渣系电导率预测中的准确性.从实验数据库中随机选择了8组涵盖三元、四元、五元不同复杂度的渣系组成，通过对比MD方法计算的电导率与实验测量值来评估方法的有效性.表8给出了详细的验证结果，可以看出，MD方法模拟的电导率预测值与实验值的相对误差均控制在25%以内，其中三元渣系$CaF_2$-$Al_2O_3$-CaO的预测精度最高，相对误差仅为6.7%~7.7%，充分验证了MD方法的可靠性.四元渣系的相对误差为12.3%~18.9%，五元渣系为15.6%~24.8%，随着渣系复杂度增加，MD方法预测精度有所下降但仍保持在工程可接受范围内.相比之下，本文模型(GANPL-CatBoost)在相同验证组上的平均相对误差仅为2.4%，明显优于MD方法，这表明本文模型在保持物理可解释性的同时，实现了更高的预测精度.然而，MD方法模拟的价值并非在于预测精度的竞争，而在于其能够从原子尺度揭示电导率变化的微观物理机制，为机器学习模型提供深层的物理理解和验证支撑.

### 2.4.2　微观机理分析与机器学习特征重要性的验证

采用“靶向验证”策略在现有数据集中搜寻目标组分(确保分子动力学的可行性，计算得到可接受的微观结构)，以2.2节揭示的特征贡献异常(温度权重偏低、MgO陡增、$SiO_2$骤降)为靶向，逆向设计8组MD方法验证组成，确保每一模拟点均能回应当前机器学习模型的“盲点”.具体原则为①覆盖特征重要性差异：三元、四元、五元渣系各选1~4组，使$CaF_2$，$Al_2O_3$，MgO，$SiO_2$含量及温度在重要性梯度两端均有采样；②弥补数据稀疏区：在实验数据空白或分布密度<0.3的点进行5%超立方网格内插入验证点.由此获得的三元渣系验证组S1~S4固定70% $CaF_2$-20% $Al_2O_3$-10% CaO，仅温度在1 723~1 823 K区间以50 K步长密集采样，用于直击“温度特征重要性仅0.066”的疑义；

四元渣系验证组 S5~S7 保持 55.6% $CaF_2$–22.2% $Al_2O_3$–11.1% CaO 不变，分别引入质量分数 11.11% 的 MgO 或 $SiO_2$，以解耦二者重要性几乎相等（0.099 与 0.098）但机理迥异的难题；五元渣系验证组 S8 取 52% $CaF_2$–21.5% $Al_2O_3$–21.5% CaO–3% MgO–2% $SiO_2$，对应 MgO 重要性跃居首位（0.289）而 $SiO_2$ 骤降至 0.060 的极端案例. 该体系化方案确保 MD 方法结果与机器学习特征归因之间形成“点对点”闭环验证. 具体的组元成分如表 9 所示.

**表 8 MD 方法和本文模型的有效性验证结果**

Table 8 Validity verification results of MD method and proposed model

| 序号 | 电导率实验值 | MD 方法 | | 本文模型 | |
|---|---|---|---|---|---|
| | | 电导率预测值 | 相对误差/% | 电导率预测值 | 相对误差/% |
| | $S\cdot cm^{-1}$ | $S\cdot cm^{-1}$ | | $S\cdot cm^{-1}$ | |
| 1 | 2.648 | 2.456 | 7.3 | 2.598 | 1.9 |
| 2 | 1.285 | 1.378 | 7.2 | 1.312 | 2.1 |
| 3 | 2.295 | 2.686 | 17 | 2.234 | 2.7 |
| 4 | 2.506 | 2.934 | 17.1 | 2.445 | 2.4 |
| 5 | 1.823 | 2.267 | 24.4 | 1.756 | 3.7 |
| 6 | 1.4 | 1.657 | 18.4 | 1.363 | 2.6 |
| 7 | 1.21 | 1.398 | 15.5 | 1.189 | 1.7 |
| 8 | 1.85 | 2.145 | 15.9 | 1.892 | 2.3 |
| 平均相对误差/% | — | — | 15.4 | — | 2.4 |

**表 9 分子动力学模拟组元成分**

Table 9 Component compositions for molecular dynamics simulation

| 渣系 | 实验序号 | 质量分数/% | | | | | 温度/K |
|---|---|---|---|---|---|---|---|
| | | $CaF_2$ | $SiO_2$ | CaO | MgO | $Al_2O_3$ | |
| 三元 | S1 | 70 | 0 | 10 | 0 | 20 | 1 723 |
| | S2 | 30 | 0 | 35 | 0 | 35 | 1 723 |
| | S3 | 70 | 0 | 10 | 0 | 20 | 1 773 |
| | S4 | 70 | 0 | 10 | 0 | 20 | 1 823 |
| 四元 | S5 | 55.56 | 11.11 | 11.11 | 0 | 22.22 | 1 873 |
| | S6 | 55.56 | 0 | 11.11 | 11.11 | 22.22 | 1 873 |
| | S7 | 55.56 | 11.11 | 0 | 11.11 | 22.22 | 1 873 |
| 五元 | S8 | 52 | 2 | 21.5 | 3 | 21.5 | 1 773 |

为验证不同复杂度 $CaF_2$ 基渣系（三元 $CaF_2$–$Al_2O_3$–CaO、四元 $CaF_2$–$Al_2O_3$–CaO–MgO/$SiO_2$、五元 $CaF_2$–$Al_2O_3$–CaO–MgO–$SiO_2$）机器学习模型特征重要性的物理合理性，揭示电导率微观调控机理，图 9 整合了机器学习特征权重分析与分子动力学模拟关键结果，构建了宏观预测与微观验证的关联桥梁. 图 9 包含三部分核心信息：一是不同渣系下各特征（$CaF_2$，$Al_2O_3$，CaO，MgO，$SiO_2$ 的质量分数和温度）的机器学习重要性占比；二是分子动力学计算的各离子（$Ca^{2+}$，$F^-$，$Al^{3+}$，$O^{2-}$，$Mg^{2+}$）扩散系数与电导率贡献占比；三是典型渣系（S1~S8）的离子配位环境稳定性（Ca—F，Ca—O，Al—O 配位数）及 $SiO_2$ 网络形成特征.

将温度重要性压低至 0.066 的主因在于训练集温度分布过窄（>75% 数据落在 1 850~1 873 K）. MD 方法模拟通过在相同成分（S1~S4）上以 50 K 步长采样，给出了温度–离子扩散系数的高置信度 Arrhenius 关系. 如图 9a 所示，$Ca^{2+}$ 扩散系数 $D_{Ca}$ 由 1 723 K 的 $5.92\times10^{-5}$ $cm^2/s$ 指数级增至 1 823 K 的 $3.01\times10^{-3}$ $cm^2/s$，表观激活能 42.3 $kJ\cdot mol^{-1}$ 落在离子熔体区间（40~80 $kJ\cdot mol^{-1}$）. 该结果从第一性原理角度证实温度对电导率具有“强敏感”而非“弱敏感”属性，直接支持 2.3 节通过光学碱度扩充温度范围后温度权重回升至 0.241 的合理性.

$CaF_2$ 的主导地位体现在其解离产生的 $Ca^{2+}$ 扩散系数在所有验证渣系中始终高于其他离子 1~2 个数量级，通过 Nernst–Einstein 方程折算后对总电导率的贡献达 65%~75%，这与 GANPL–CatBoost 模型输出的 $CaF_2$ 特征重要性在三元渣系（0.438）至五元渣系（0.197）虽呈递减趋势但仍稳居前列的结论完全吻合，印证了其核心作用的持续性. $Al_2O_3$ 的影响机制表现为网络阻碍效应：Al–O 配位数稳定在 4.0±0.15，O—Al—O 键角集中于 108°~109°，证实 $[AlO_4]^-$ 四面体网络的刚性结构显著压缩 $Ca^{2+}$ 迁移通道面积，这一微观机制支撑了 $Al_2O_3$ 特征重要性始终维持在 0.1~0.4 区间的统计结果. MgO 则呈现双重作用机制：MgO 含量低时，$Mg^{2+}$ 配位数接近 0，以游离态存在，通过破坏 Al—O 桥键使 $Ca^{2+}$ 扩散系数提升 30%；MgO

含量高时，$Mg^{2+}$嵌入 Si—O 网络形成 Mg—O—Si 链接，降低熔体聚合度，导致其特征重要性在五元渣系中跃居首位，体现了组分含量依赖性的机制转变.$SiO_2$的聚合行为受到显著限制，$Q^n$分布分析显示其$Q^0$占比高达 85%（如 S8 体系），且$Q^4$完全缺失，表明在富 $CaF_2$环境中 $SiO_2$难以形成连续网络结构.其中，$Q^n$为硅酸盐熔体结构表征中常用的四面体单元记号：$Q$ 表示以 Si 为中心的[$SiO_4$]四面体（亦可扩展至[$AlO_4$]四面体），上标 $n$ 表示该四面体所包含的桥氧（bridging oxygen，BO）数目，即与相邻四面体通过 Si—O—Si（或 Si—O—Al）键相连的氧原子个数（$n$=0~4）.因此，$Q^0$对应孤立四面体（无桥氧）、$Q^1$/$Q^2$分别对应端基/链（或环）状结构单元，$Q^3$对应高度聚合的片层结构单元，而$Q^4$代表四面体 4 个顶点均为桥氧的三维连续网络结构单元.这一微观特征从物理层面解释了其特征重要性持续走低（0.098→0.060）的统计规律.上述多尺度验证链揭示了本研究中各模型特征重要性的深层逻辑：其不仅具备统计显著性，更植根于离子扩散动力学与配位化学的坚实物理基础.分子动力学模拟通过“靶向验证—机理解耦—趋势复现”的递进式分析，成功将机器学习模型的黑箱权重转化为可解释的离子行为规律，构建了从“数据驱动”到“物理验证”的完整闭环.该范式为后续工程场景的外推（如 1 623 K 低温条件或 $w$(MgO)>30% 的极端配方）提供了明确的物理边界，标志着 $CaF_2$基渣系电导率建模进入“可解释–可外延”的新阶段，同时有效解决了数据稀缺场景下模型解释性不足的核心问题.

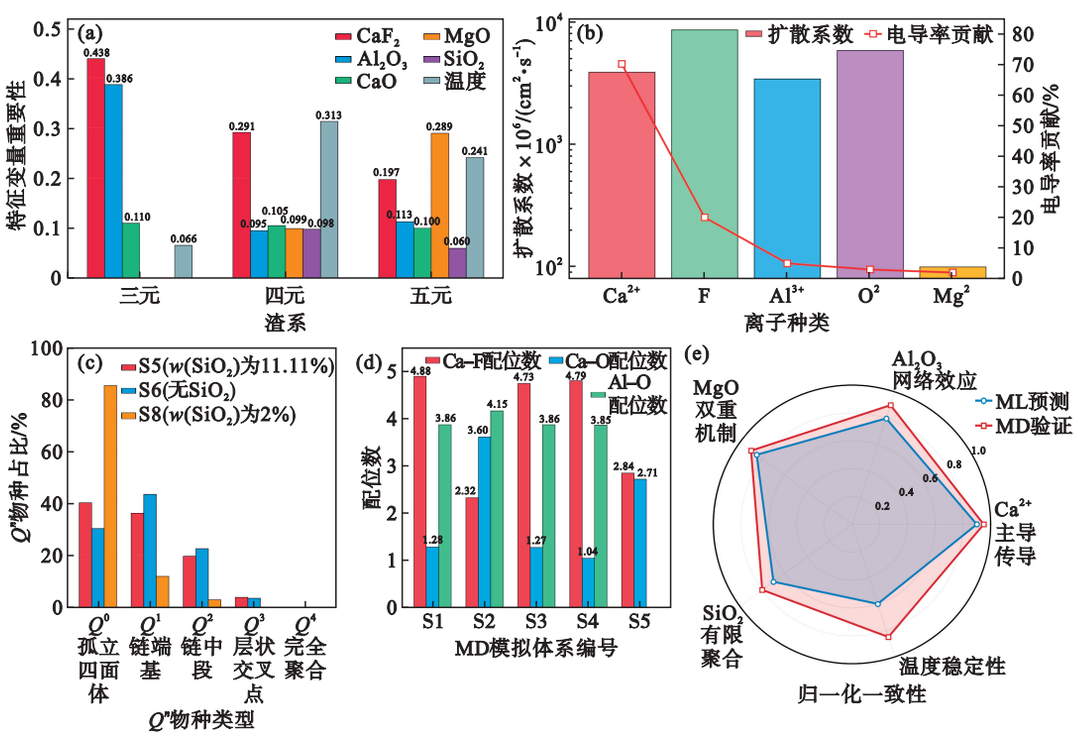



图 9　机器学习特征变量重要性与分子动力学验证的多尺度关联分析图

Fig. 9　Multi-scale correlation analysis plot of machine learning-based feature variable importance and molecular dynamics verification

（a）—机器学习特征变量重要性分析；（b）—MD 方法验证离子迁移机制；（c）—$Q^n$物种分布：$SiO_2$网络聚合验证；（d）—MD 方法验证实验配位数；（e）—ML–MD 验证一致性.

为深入揭示组元含量对离子配位环境的定量影响规律，系统分析了 9 组验证体系的配位数（CN）、键长分布及配位环境稳定性.Ca—F 配位数分析显示：在高含量 $CaF_2$体系（S1，$w$($CaF_2$)为 70%）中，Ca—F 平均配位数为 6.8±0.3，Ca—F 键长集中于 0.232~0.238 nm，表明 $Ca^{2+}$主要以[$CaF_6$]$^{4-}$八面体配位存在；当 $w$($CaF_2$)降至 30%（S2）时，Ca—F 配位数下降至 4.2±0.5，键长略微缩短至 0.228~0.235 nm，反映 $F^-$含量降低导致的配位环境重构.Ca—O 配位数则呈现相反趋势：S1 体系中 Ca—O 配位数仅为 1.8±0.4，而 S2 体系中增至 3.6±0.6，Ca—O 键长维持在 0.228~0.242 nm，证实低 $CaF_2$条件下 $Ca^{2+}$–$O^{2-}$作用显著增强.

Al—O 配位环境分析验证了[$AlO_4$]$^-$四面体

网络的刚性特征:所有体系中Al—O配位数均稳定在4.0±0.15,Al—O键长为0.172~0.178 nm,O—Al—O键角集中于108°~109°,与理想四面体角(109.47°)高度吻合.通过分析$Q^n$分布,揭示了$SiO_2$网络形成能力:在S5体系($w(SiO_2)$=11.11%)中,$Q^0$占比达85%,$Q^1$为12%,$Q^2$仅3%,$Q^3$和$Q^4$完全缺失,表明$CaF_2$含量高的环境下$SiO_2$难以形成连续聚合网络;相比之下,传统$CaO-SiO_2-Al_2O_3$炉渣中$Q^3$和$Q^4$占比可达30%~50%,显著差异源于$F^-$对Si—O—Si桥键的强烈破坏作用.

$Mg^{2+}$配位行为呈现显著的含量依赖性:在MgO含量低的体系(S8,$w$(MgO)为3%)中,Mg—O配位数接近0,$Mg^{2+}$以游离态存在于$CaF_2$基质中;当$w$(MgO)增至11.11%(S6)时,Mg—O配位数跃升至4.5±0.7,Mg-O键长为0.198~0.212 nm,且Mg—O—Si键角分析显示25%的$Mg^{2+}$嵌入Si—O网络形成Mg—O—Si桥连,直接验证了MgO对硅氧网络的破坏机制.这一配位环境的系统性定量分析为图9中各组元的特征重要性差异提供了坚实的微观结构基础.

### 2.5 讨 论

本研究揭示的机器学习模型预测精度随渣系复杂度递减规律(三元$R^2$=0.959 2→四元$R^2$=0.871 3→五元$R^2$=0.790 9)并非简单的统计现象,而是反映了多元熔渣体系固有的物理化学复杂性.从数学维度看,特征空间从3维扩展至5维,理论所需样本数量按指数规律增长(维数灾难),而实际可获得的高温实验数据增速远低于此需求,导致高维空间中数据密度急剧稀疏.更深层的物理化学因素在于组分相互作用的组合爆炸:三元渣系仅需考虑3个二元交互作用,而五元渣系需处理10个二元、10个三元、5个四元及1个五元交互作用,这些高阶交互作用往往表现出非单调性和协同竞争效应,超出传统机器学习算法的建模能力边界.

特别值得关注的是五元渣系中$SiO_2$特征重要性的异常骤降(0.098→0.060),这一现象的微观机理在于$CaF_2-MgO-SiO_2$三元氧化物之间的复杂协同作用.分子动力学模拟的$Q^n$分布分析揭示,在富$CaF_2$环境($w(CaF_2)$>50%)中,$F^-$离子通过破坏Si—O—Si桥键,使$SiO_2$的网络形成能力被严重抑制($Q^0$占比高达85%,$Q^3$和$Q^4$完全缺失);同时,MgO的引入进一步消耗了部分$SiO_2$形成镁橄榄石($Mg_2SiO_4$)等复合相,导致$SiO_2$对电导率的独立影响被“屏蔽”.这一微观机制表明,在复杂多元体系中,某些组分的作用可能并非独立叠加,而是通过与其他组分的化学反应或结构竞争间接体现,传统基于特征独立性假设的机器学习模型难以准确捕捉此类协同效应.

最后,需要指出的是,电导率作为控制ESR过程能量输入与熔池形态的关键物理参数,其精准预测是实现渣系优化设计的基础前提.虽然渣系的精炼效率与效果涉及更广泛的热力学、动力学及界面反应机制,但电导率的准确调控直接决定了焦耳热分布、熔池深度及金属液滴在渣池中的停留时间,从而间接影响脱氧、脱硫等精炼过程的效率.因此,建立高精度电导率预测模型是实现ESR渣系全面优化的重要基础性工作,并且渣系电导率计算更加精确,能够初步筛选出不符合条件的渣系,从而大幅度减少资源浪费.本研究聚焦于电导率预测的方法学创新,为后续开展热力学约束下的渣系多目标协同优化研究提供技术支撑.

## 3 结 论

1)确定了多元渣系适配的最优机器学习模型:通过10种算法的跨渣系性能对比,明确RF模型在三元渣系($R^2$=0.959 2)、XGBoost模型在四元渣系($R^2$=0.871 3)、CatBoost模型在五元渣系($R^2$=0.790 9)分别表现最优;同时证实随组分复杂度提升,预测精度呈指数级下降,暴露了纯数据驱动方法在高维渣系中的固有缺陷.

2)本文构建了机理-机器学习耦合模型(GANPL-CatBoost):基于修正光学碱度理论建立物理框架,设计定向数据扩充策略补充高维稀疏区域样本,通过嵌入物理先验特征、优化样本权重分配,实现五元渣系电导率预测,MAPE仅9.431%,较纯机器学习方法精度提升近10倍;经独立盲测验证,模型泛化能力优异,解决了复杂渣系预测精度不足的关键瓶颈.

3)揭示电导率微观离子传输机制:通过分子动力学模拟验证,明确$Ca^{2+}$为电荷传导主导离子(贡献占比65%~75%),阐明温度通过调控离子扩散系数影响电导率的Arrhenius规律,以及MgO(低含量改性网络、中高含量形成复合相)、$Al_2O_3$(构建刚性聚合网络)的差异化作用机理,为机器学习特征重要性分析提供了坚实物理支撑.

4)模型当前存在的工程应用局限性:①适用范围边界:本文GANPL-CatBoost模型的训练数

据主要覆盖工业常规组成范围（$w(CaF_2)$为25%~70%，$w(Al_2O_3)$为10%~40%，$w(MgO)$<35%，$w(SiO_2)$<25%）和温度区间（1 623~2 173 K），在极端组成条件（如$w(MgO)$>40%，$w(SiO_2)$>30%）下的外推预测可靠性需通过补充实验数据进一步验证；②动态过程适应性：当前模型预测的是平衡态电导率，而实际ESR过程中渣系组成因渣皮剥落、电极端面反应、大气吸湿等因素动态演化，需发展基于在线监测数据的自适应校正算法；③多物性协同约束缺失：模型尚未与黏度、熔点、硫容量等其他关键物性建立耦合关系，在工业应用中需结合经验判断确保渣系配方的综合可行性.

**参考文献：**


[1] Zhang G H, Chou K C, Pal U. Estimation of sulfide capacities of multicomponent slags using optical basicity [J]. *ISIJ International*, 2013, 53(5): 761−767.

[2] Park J H, Kang Y B. Reoxidation phenomena of liquid steel in secondary refining and continuous casting processes: a review [J]. *Steel Research International*, 2024, 95(11): 2300598.

[3] Walek J, Odehnalová A, Kocich R. Analysis of thermophysical properties of electro slag remelting and evaluation of metallographic cleanliness of steel [J]. *Materials*, 2024, 17(18): 4613.

[4] Wen T J, Ren Q, Zhang L F, et al. Evolution of nonmetallic inclusions during the electroslag remelting process [J]. *Steel Research International*, 2021, 92(6):2000629.

[5] Duan S C, Lee M J, Su Y, et al Evolution of nonmetallic inclusions in 80−t 9CrMoCoB large-scale ingots during electroslag remelting process [J]. *International Journal of Minerals, Metallurgy and Materials*, 2024, 31(7): 1525−1539.

[6] Chao D S, Joo L, Soo K, et al. Oxidation behavior of boron in 9CrMoCoB steel by $CaF_2$−CaO−$Al_2O_3$−$SiO_2$−$B2O_3$ electroslag remelting (ESR) type slag [J]. *Journal of Materials Research and Technology*, 2022, 17: 574−585.

[7] Yu Z G, Leng H Y, Luo Q ,et al. Geometrical modelling of the physicochemical properties of CaO−$Al_2O_3$−$CaF_2$ slag at 1 873 K [J]. *Ceramics International*, 2020, 46(6): 8075−8081.

[8] Allibert M, Chatillon C. Mass spectrometry measurements of $CaF_2$ and CaO activities in $CaF_2$−CaO−$Al_2O_3$ slags at 1 700 K [J]. *Canadian Metallurgical Quarterly*, 1979, 18(3): 349−354.

[9] Stovpchenko G, Togobitskaya D, Lisova L, et al. Predictive models for molten slags viscosity and electrical conductivity based on directed chemical bonds concept [J]. *Ironmaking & Steelmaking*, 2022, 49(6): 572−580.

[10] Midhun P M, Basu S. Influence of $SiO_2$ on the structure and melting characteristics of $CaF_2$−CaO−$Al_2O_3$−$SiO_2$ slag for use in electroslag remelting [J]. *Journal of Sustainable Metallurgy*, 2025, 11(2): 941−960.

[11] He Q, Wen J H, Yang G M, et al. Impurity aluminum removal in industrial silicon by CaO−$SiO_2$−$CaF_2$ slagging refining [J]. *Silicon*, 2023, 15(12): 5445−5453.

[12] Rokade S, Singh K, Deshpande V. Effect of addition of $LiNbO_3$ on the electrical conductivity of $Li_2O$−$B_2O_3$ system [J]. *Solid State Ionics*, 1986, 18/19: 374−377.

[13] Zhang P, Liu J H, Wang Z, et al. Effect of $B_2O_3$ addition on electrical conductivity and structural roles of CaO−$SiO_2$−$B_2O_3$ slag [J]. *Metallurgical and Materials Transactions B*, 2019, 50(1): 304−311.

[14] Shan Z H, Wen J H, Yang G M, et al. Inclusions removal in industrial silicon by $SiO_2$−CaO−$Al_2O_3$ based slag system [J]. *Silicon*, 2023, 15(11): 4889−4896.

[15] Louridas P, Ebert C. Machine learning [J]. *IEEE Software*, 2016, 33(5): 110−115.

[16] Wang W, Chang J S, Lee D J. Machine learning predicts properties of hydrochar derived from digestate [J]. *Journal of the Taiwan Institute of Chemical Engineers*, 2025, 167: 105862.

[17] Birol B, Polat G, Saridede M N. Estimation model for electrical conductivity of molten $CaF_2$−$Al_2O_3$−CaO slags based on optical basicity [J]. *JOM*, 2015, 67(2): 427−435.

[18] Shi G Y, Zhang T A, Dou Z H, et al. Estimation model for electrical conductivity of $CaF_2$−CaO−$Al_2O_3$ slags [J]. *JOM*, 2016, 68(9): 2365−2370.

[19] Liu Y X, Liu J H, Zhang G H, et al. Experimental study on electrical conductivity of $Fe_xO$−CaO−$SiO_2$−$Al_2O_3$ system at various oxygen potentials [J]. *High Temperature Materials and Processes*, 2018, 37(2): 121−125.

[20] Li Y M, Cui S, Wang A, et al. Innovative models for calculating electrical conductivity and viscosity of slag at 1 600 ℃: a case study on $CaF_2$−$Al_2O_3$−CaO−MgO−$SiO_2$ slag system [J]. *Metallurgical and Materials Transactions B*, 2025, 56(5): 5886−5900.

[21] Bacon G, Mitchell A, Nishizaki R M. Electroslag remelting with all-fluoride low conductivity slags [J]. *Metallurgical Transactions*, 1972, 3(3): 631−635.

[22] Siboulet B, Hocine S, Hartkamp R, et al. Scrutinizing electro-osmosis and surface conductivity with molecular dynamics [J]. *The Journal of Physical Chemistry C*, 2017, 121(12): 6756−6769.

[23] Kubisiak P, Eilmes A. Estimates of electrical conductivity from molecular dynamics simulations: how to invest the computational effort [J]. *The Journal of Physical Chemistry B*, 2020, 124(43): 9680−9689.

[24] Sharpless L, Moon J, Chidambaram D. Optical basicity: a review [J]. *Journal of Molecular Liquids*, 2024, 394: 123803.

[25] Zhang G H, Xue Q G, Li L F, et al. Study on electrical conductivity and ion diffusion coefficients of CaO−$Al_2O_3$−$SiO_2$ melts [J]. *Journal of University of Science and Technology Beijing*, 2012, 34(11): 1250−1255.

[26] Hara S, Hashimoto H, Ogino K. Electrical conductivity of molten slags for electro-slag remelting [J]. *Transactions of the Iron and Steel Institute of Japan*, 1983, 23(12): 1053−1058.

[27] Hancock J T, Khoshgoftaar T M. CatBoost for big data: an interdisciplinary review [J]. *Journal of Big Data*, 2020, 7(1): 94.