

基于物理信息机器学习的玻璃杨氏模量预测模型构建与机理分析

钱兆青¹, 曹 毅², 康泽玉¹, 卢亚东¹, 高文凯¹, 康俊峰^{1,3}

(1. 济南大学材料科学与工程学院, 济南 250022; 2. 济南大学信息科学与工程学院, 济南 250022;

3. 山东省高性能玻璃纤维重点实验室, 山东 泰安 271000)

摘 要: 针对玻璃杨氏模量公开数据库数据一致性差、传统清洗难以剔除异常样本、纯数据驱动模型泛化与外推能力受限等问题, 提出一种融合物理先验与机器学习的玻璃杨氏模量预测方法。将 Makishima-Mackenzie(MM)半经验模型嵌入梯度提升机(GBM)建模流程, 依据实测值大于 MM 计算值的物理规律完成数据清洗, 构建受物理约束的高质量数据集; 以 MM 模型计算残差为预测目标, 通过残差修正与 MM 计算值加和实现杨氏模量精准预测。结合 SHAP 可解释性分析与物理化学信息描述符建模, 揭示 MM 模型系统误差的产生机制。结果表明, 所构建的物理信息机器学习(PIML)模型测试集决定系数 R^2 达 0.962, 中位绝对误差 MedAE 为 0.77, 均方误差 MSE 为 7.04, 预测精度、鲁棒性与泛化能力显著优于纯数据驱动模型; 外推测试验证模型在训练集域外样本仍保持稳定低误差。该方法为高模量玻璃配方快速开发提供了兼具高精度、可解释性与强外推能力的性能预测新途径。

关键词: 物理信息机器学习; Makishima-Mackenzie 模型; 高模量玻璃; SHAP 分析

中图分类号: TQ171.1 文献标志码: A 文章编号: 0454-5648(2026)10-3593-11

网络出版时间: 2026-07-29



随着航空航天^[1]、风力发电^[2]等领域的飞速发展, 作为关键增强材料的高模量玻璃需求量快速增长, 新型高性能高模量玻璃配方的开发迫在眉睫。随着人工智能技术的快速发展, 机器学习成为继试错实验法与计算机模拟技术后的又一全新研究范式^[3]。作为一种典型的数据驱动方法, 机器学习可从已有的研究成果中学习并挖掘隐含的数据规律并作出预测。凭借高效、精准、低能耗等突出优势, 该技术已被广泛应用于玻璃杨氏模量的研究中^[4, 5]。

现有研究证明, 玻璃杨氏模量预测模型普遍存在性能瓶颈, 且预测精度显著低于其他性能的预测。Qin 等^[6]基于 INTERGLAD 8.3 数据库建模, 使用 2977 个样本训练的热膨胀系数预测 R^2 达 0.968, 而使用 5088 个样本的杨氏模量预测 R^2 仅 0.816; Kang 等^[3]基于 SciGlass 数据库建模, 密度模型的 R^2 达 0.937, 杨氏模量模型的 R^2 仅 0.880。造成该现象的原因主要有 2 点: 1) 杨氏模量易受制备热历史、测试

条件影响, 数据库相关数据一致性不佳; 2) 由于缺乏有效剔除异常样本的清洗策略, 模型易受异常样本干扰导致过拟合, 泛化性能受限, 弱化模型性能。

Liu 等^[7, 8]提出了纯数据驱动模型在材料领域应用的三大主要矛盾与数据治理框架, 并发现将材料领域知识嵌入机器学习全流程能够有效提升模型性能^[9-12]。并且 Liu 等认为, 数据集的质量对模型性能起决定性作用^[13], 将材料领域知识引入数据预处理, 建立数据异常检测 workflow, 有效地提升了模型预测精度^[14, 15]。不仅如此, 相关研究还发现, 使用“分而治之”法^[16-21], 利用领域知识将主任务分解为多阶段子任务, 能够给出兼具精准度与可解释性的预测模型^[16, 18]。

Wang 等^[22]的研究中, 使用物理信息机器学习(Physics-informed machine learning, PIML)方法, 将物理先验与神经网络(Neural networks, NN)模型的损失函数结合建立物理信息神经网络(Physics-informed

收稿日期: 2026-05-13。 修订日期: 2026-06-08。

基金项目: 山东省自然科学基金面上项目(ZR2025MS774); 国家自然科学基金面上项目(52172019); 泰安市岱岳区科技创新项目(CXXM2025002)。

第一作者: 钱兆青(2001—), 男, 硕士研究生。

通信作者: 康俊峰(1988—), 男, 博士, 副教授。

Received date: 2026-05-13. Revised date: 2026-06-08.

First author: QIAN Zhaoqing (2001—), male, Master candidate.

E-mail: 864897665@qq.com

Correspondent author: KANG Junfeng (1988—), male, Ph.D., Associate Professor.

E-mail: mse_kangjf@ujn.edu.cn

neural networks, PINN)模型,有效抑制了过拟合,所建立的预测玄武岩纤维杨氏模量预测模型性能大幅提升。PIML 通过将物理先验融入机器学习建模流程,让模型在数据驱动的同时遵循物理约束,提升预测的泛化性与外推能力。

Makishima-Mackenzie(MM)模型是氧化物玻璃杨氏模量预测的经典半经验模型,其根据离子堆积密度与氧化物的离解能拟合出了一个易于计算且物理意义清晰的计算公式,是理想的物理先验知识^[23]。但 Shi 等^[24]的研究指出,MM 模型计算值与实测值存在系统性偏差,会显著低估玻璃的杨氏模量。

Willard 等^[25]提到,残差建模(Residual modeling)是一种有效的解决物理模型缺陷的机器学习建模方法。基于此,本工作针对玻璃杨氏模量的纯数据驱动模型精度受限、可解释性与外推能力不足的问题,将物理先验 MM 模型与机器学习模型深度融合,利用 MM 模型的固有规律^[24]剔除异常样本,构建高质量数据集,将预测目标变更为 MM 模型计算残差,构建 PIML 模型。通过 SHAP 分析与物理化学信息

描述符建模揭示 MM 模型误差的来源,最终得到兼具高预测精度、可解释性与强外推能力的玻璃杨氏模量预测模型。

1 计算方法

研究流程如图 1 所示,基于 SciGlass 数据库筛选清洗构建原始数据集,并基于原始数据集训练梯度提升机(Gradient boosting machine, GBM)、随机森林(Random forest, RF)、极端提升树(eXtreme gradient boosting, XGB)、人工神经网络(Artificial neural network, ANN)和支持向量回归(Support vector regression, SVR)5 种数据驱动机器学习模型并进行优选。将物理先验 MM 模型与最优模型结合,利用实测值大于 MM 计算值的特征^[24]开展数据清洗,构建受物理约束的高质量数据集,进而建立以 MM 计算残差为预测目标的 PIML 模型。通过物理化学信息描述符与 SHAP 分析揭示 MM 模型误差机制,并采用训练集范围外样本开展外推测试,验证构建的 PIML 模型兼具高预测精度、可解释性与强外推能力。

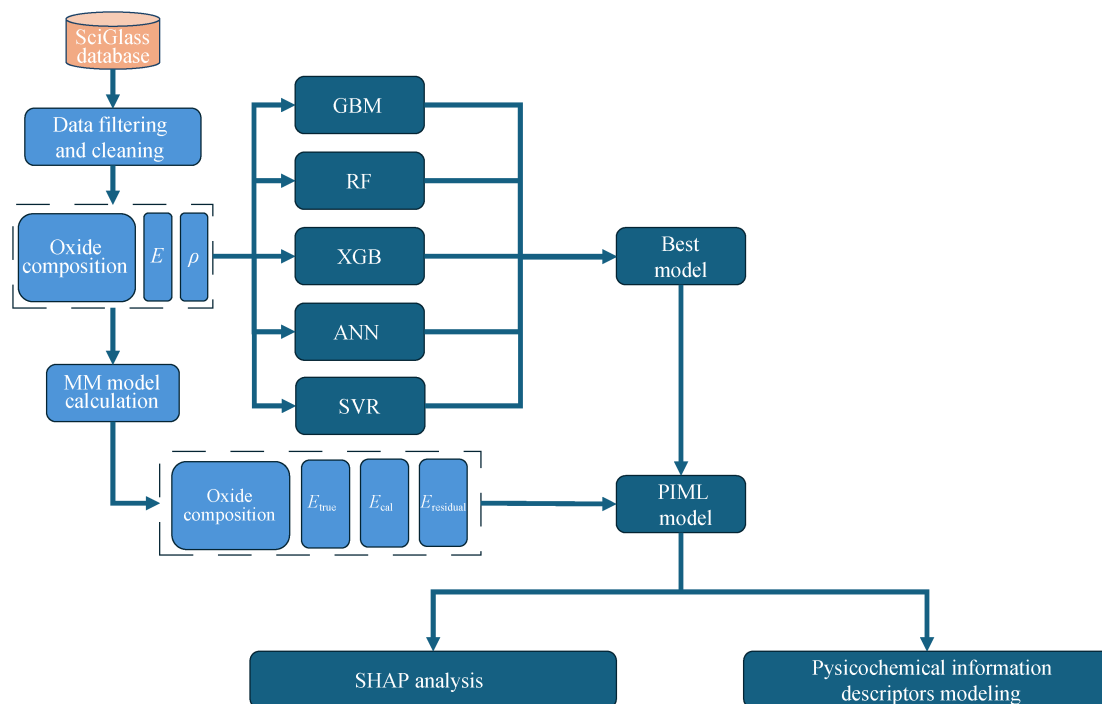


图 1 研究流程图

Fig. 1 Research flowchart

1.1 数据集建立

数据集依托 Glasspy^[26]软件包调取 SciGlass 数据库构建。筛选 15 种核心氧化物(SiO_2 、 Al_2O_3 、 Li_2O 、 Na_2O 、 K_2O 、 MgO 、 CaO 、 SrO 、 BaO 、 ZnO 、 TiO_2 、

ZrO_2 、 Y_2O_3 、 La_2O_3 、 Sc_2O_3)作为特征;同时提取杨氏模量与密度数据,分别作为模型预测目标及 MM 模型的计算参数,清洗工作主要包括:1) 去除成分之和 < 0.999 或 > 1.001 配方不完整样本;2) 去除杨

氏模量和密度为空值、零值、负值的异常样本; 3) 剔除重复配方, 每组保留杨氏模量中位数及其对应密度, 合并为单一样本。

计算各样本的 MM 模型计算值。式(1)为 MM 模型公式^[23]:

$$E = 2V_i \sum_i G_i X_i \quad (1)$$

其中: E 为杨氏模量; V_i 为离子堆积密度; G_i 为组分 i 的离解能; X_i 为组分 i 的摩尔百分比。其中 V_i 的计算公式如式(2)所示:

$$V_i = \frac{\rho}{M} \sum_i V_i X_i \quad (2)$$

其中: ρ 为玻璃的密度; M 为玻璃的摩尔质量。对于化学式为 $A_x B_y$ 的组分, 其 V_i 计算公式如式(3)所示:

$$V_i = 6.022 \times 10^{23} \times \frac{4}{3} \pi X r_A^3 + Y r_B^3 \quad (3)$$

根据上述公式、文献报道的各氧化物的离解能 G_i ^[23]以及数据集的组分与密度, 计算得到杨氏模量的 MM 模型计算值 E_{cal} , 得到原始数据集。

MM 模型对玻璃杨氏模量存在显著的系统性低估^[24], 因此, 若某一样本的实测值 E_{true} 低于计算值 E_{cal} , 则该样本大概率为异常样本。据此进一步清洗原始数据集: 根据式(4)计算 MM 模型的预测残差 $E_{residual}$, 剔除 $E_{residual}$ 为负的异常样本。构建受物理约束的高质量数据集。

$$E_{residual} = E_{true} - E_{cal} \quad (4)$$

1.2 数据驱动模型构建

以常规数据集的组分作为特征、实测值作为标签, 分别输入 GBM、RF、XGB、ANN、SVR 模型中进行训练, 建立常规模型。数据集以随机种子 42 按照 8:2 划分为训练集与测试集。其中 RF、XGB、ANN、SVR 为参考文献[3, 6]中常用的、预测精度较高的机器学习模型。而 GBM 与 XGB 同属梯度提升型模型, 因此同样作为备选模型之一。采用贝叶斯优化搜索与五折交叉验证优化模型的超参数。使用的模型基于 scikit-learn^[27]和 XGBoost^[28]软件包实现, 超参数的贝叶斯优化搜索基于 hyperopt 包实现^[29]。

1.3 PIML 模型构建

以高质量数据集的组分作为特征, $E_{residual}$ 作为标签, 输入最优数据驱动模型中进行训练。训练配置和软件包与数据驱动模型相同。

1.4 模型评估指标

采用 R^2 ^[30]、MSE^[30]、MedAE^[31]3 种评估指标,

综合衡量不同预测模型的性能, 以此作为预测精度高低的判据。以下是 3 种指标的计算公式与定义:

式(5)为决定系数(R^2)的计算公式:

$$R^2(y, \hat{y}) = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (5)$$

其中: $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$; y_i 为真实值; $\hat{y}_i$ 为预测值。

R^2 表示模型解释数据波动的能力, 用于评估模型对数据的拟合程度, R^2 越接近 1, 说明模型拟合的程度越好, 模型的解释能力越强。

式(6)为均方误差(MSE)的计算公式:

$$MSE(y, \hat{y}) = \frac{1}{n_{samples}} \sum_{i=0}^{n_{sample}-1} (y_i - \hat{y}_i)^2 \quad (6)$$

MSE 为真实值与预测值误差平方的算术平均值, 它表示了模型预测的准确程度, MSE 越低越接近 0, 说明模型预测值与真实值的平均差异程度越小, 模型的预测越准确。

式(7)为中位绝对误差(MedAE)的计算公式:

$$MedAE(y, \hat{y}) = \text{median}(|y_1 - \hat{y}_1|, \dots, |y_n - \hat{y}_n|) \quad (7)$$

MedAE 为真实值与预测值之差的绝对值的中位数, 反映模型预测误差的典型水平。MedAE 数值越低, 说明模型对异常值拥有更强的鲁棒性, 模型抗噪声数据能力强。

以 MSE 作为贝叶斯优化目标。模型优选综合考虑 R^2 、MSE、MedAE 3 种指标。

1.5 模型解释

由于机器学习模型属于典型的黑箱模型, 其决策过程缺乏可解释性。而 SHAP 是基于博弈论的事后模型解释方法^[32], 通过为各特征分配 SHAP 值量化其对预测结果的影响方向与幅度, 可直观反映单特征在模型决策过程中的贡献度。

因此, 基于 Python 平台调用 SHAP Python 程序包^[33], 解析模型中特征对预测结果的贡献程度, 揭示模型决策机制的物理可靠性。

2 结果与讨论

2.1 数据驱动模型训练集构建与模型训练

从 SciGlass 数据库中筛选并清洗得到共计 2861 组数据作为原始数据集。由图 2a 中各氧化物的频数分布图可见, 训练集样本主要由铝硅酸盐玻璃构成, 与研究目标及常见高模量玻璃配方相符, 并且图 2b

实测值 E_{true} 频数分布直方图表明, 数据集中 E_{true} 主要分布于 70~90 GPa 区间内, 符合铝硅酸盐玻璃的分布特征。数据集内 E_{true} 连续分布, 无断层离群数据, 可以用于后续的研究工作。

输入 GBM、RF、XGB、ANN、SVR 模型中进行训练, 5 种模型的预测性能图如图 3 所示, 测试集 R^2 分别为 0.910、0.892、0.906、0.910、0.856, 模型可以实现良好的拟合效果, 但数据集中包含的异常值仍对模型产生了相当的误导, 限制了模型性能的进一步提升。

2.2 PIML 模型训练集构建与模型训练

依据 MM 模型显著低估实际模量的物理规律^[24], 剔除原始数据集中 $E_{\text{residual}} < 0$ 的异常样本, 构建受物理约束的高质量数据集, 共计 2463 组配方, 清洗前后 E_{true} 与 E_{cal} 的频数分布对比如图 4 所示。原始数据集与高质量数据集的最密分布区间均为 70~90 GPa, 并且数据连续分布, 无断层、离群样本, 分布范围较原始数据集更窄, 说明基于 MM 模型物理规律的清洗策略在未破坏数据集主体分布特征的前提下, 有效剔除了异常数据。

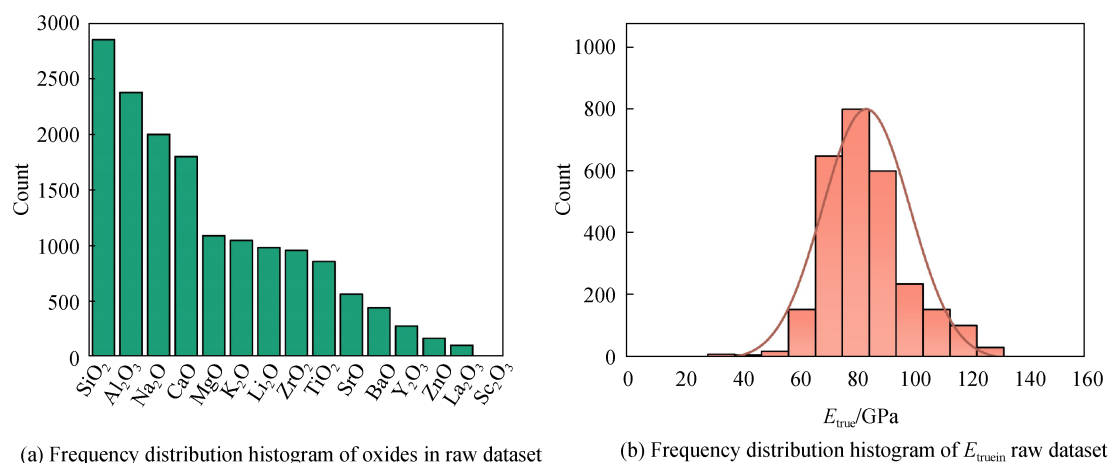


图 2 原始数据集中氧化物频数分布和实测值 E_{true} 频数分布直方图

Fig. 2 Frequency distribution histogram of oxides and measured values E_{true} in the raw dataset

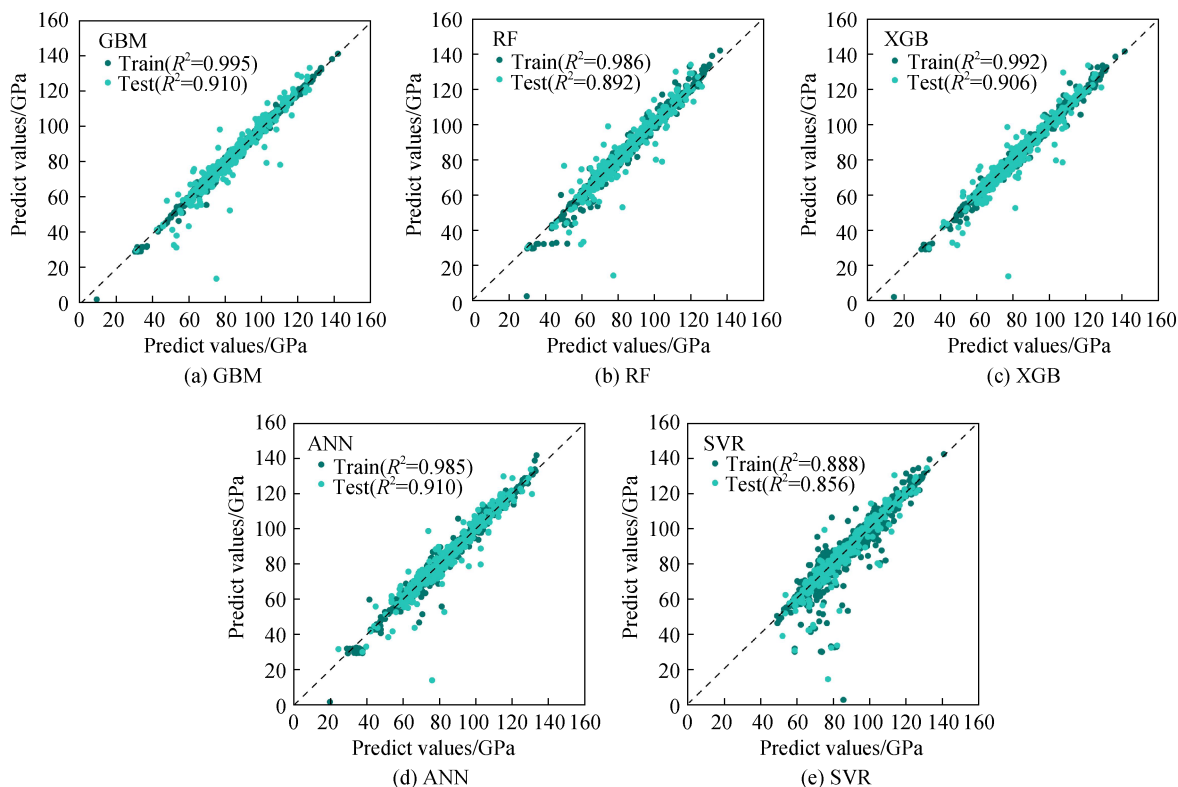


图 3 5 种模型的预测性能图

Fig. 3 Prediction performance plots of five models

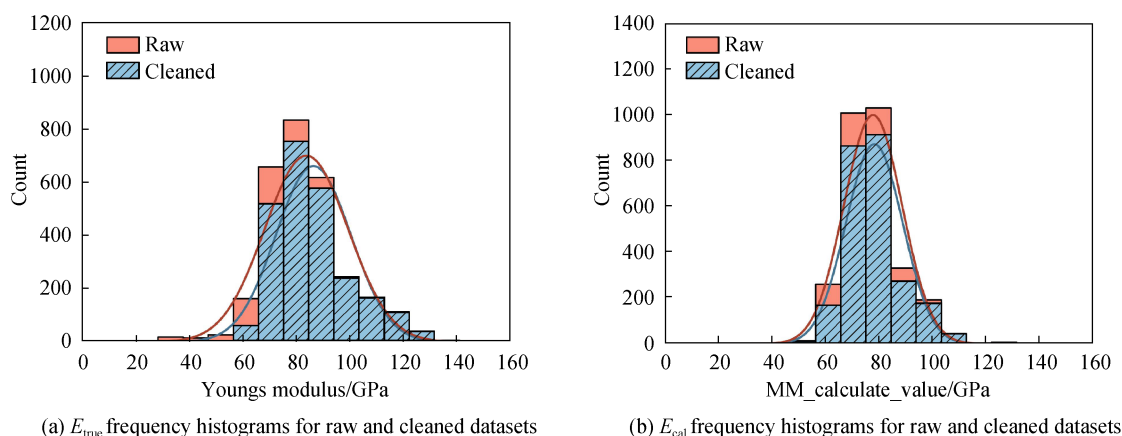
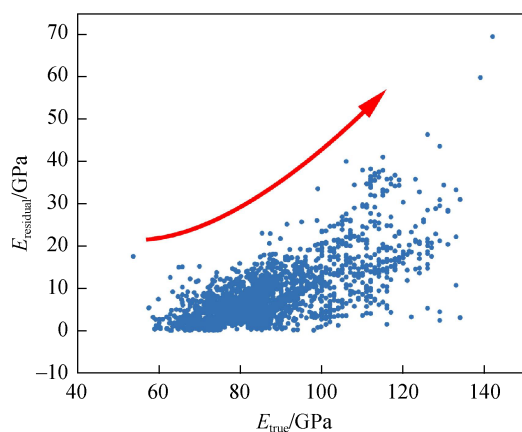
图 4 清洗前后数据集中 E_{true} 与 E_{cal} 的频数分布直方图Fig. 4 Frequency distribution histograms of E_{true} and E_{cal} in datasets before and after cleaning

图 5 为 MM 模型计算残差 $E_{residual}$ 随实测值 E_{true} 变化的趋势图。由图 5 可见, $E_{residual}$ 与 E_{true} 呈现高度正相关, 说明 MM 模型的计算残差并非随机误差, 而是有规律的系统性偏差。随着玻璃杨氏模量的增加, MM 模型计算残差同步上升, 说明 MM 模型过度简化了玻璃结构与杨氏模量之间的内在关联, 无法精准表征高模量玻璃体系的力学性能。因此, 相较于直接学习组分-模量映射关系的纯数据驱动模型, PIML 模型可以通过学习更简单的规律, 修正已有的物理模型, 做出准确的预测。

图 5 $E_{residual}$ 随 E_{true} 变化趋势图Fig. 5 Variation trend plot of $E_{residual}$ with E_{true}

选取纯数据驱动模型中表现较好的 GBM 模型与物理先验结合, 使用该模型预测 MM 模型的计算残差, 构建 GBM_PIML 模型。图 6a 为残差模型预测性能图, 其测试集 $R^2=0.866$, 说明模型能拟合 MM 模型产生误差的规律。将残差预测值与 MM 模型计算值相加, 得到杨氏模量预测值, 并重新计算模型指标。得到测试集 R^2 达 0.962, 模型对杨氏模

量拟合能力显著提升, 突破了数据驱动模型的拟合瓶颈。表 1 为 4 种模型的指标对比, PIML 模型的中位误差 MedAE 降低约 40%, 均方误差 MSE 降低约 70%, 模型精度显著提升, 极端误差样本大幅减少。说明被剔除的 $E_{residual}<0$ 的样本的确是导致模型性能受限的异常样本, 构建的受物理先验约束的高质量数据集有效地提升了模型的预测精度。证明物理先验的加入改变了预测目标, 降低特征学习难度, 最终得出更准确的预测效果。

GBM_PIML 模型的 SHAP 分析图如图 7 所示。由于模型的预测目标为 MM 模型计算残差, 因此, SHAP 分析所揭示的是各氧化物令 MM 模型算不准的程度。图 7a 为各氧化物特征的重要性图, 重要性前 5 的氧化物分别依次为: SiO_2 、 Al_2O_3 、 CaO 、 Y_2O_3 、 ZrO_2 。图 7b 为 SHAP 蜂群图, SHAP 分析表明, 其中 Y_2O_3 、 La_2O_3 含量的增加会显著增加 MM 模型的计算残差, 与被 MM 模型忽视的稀土金属氧化物相互作用生成提升玻璃杨氏模量的高配位铝的理论^[34]相符。同时, MgO 、 Li_2O 含量的增加也会显著增加计算残差, 此类高场强修饰体氧化物会积聚玻璃中的非桥氧键, 改变局部结构^[35-37], 而 MM 模型却将玻璃视为均匀体系, 这种差异导致了计算残差增加。

2.3 物理化学信息描述符建模与分析

GBM_PIML 模型的优异性能说明其内部决策过程能复现导致 MM 模型出现误差的物理机制, 因此, 设计了用于表征玻璃内部结构与氧化物间相互作用的 4 种物理化学信息描述符作为特征, 以残差值作为标签建立物理化学信息描述符预测模型并进行 SHAP 分析。表 2 为 4 种描述符的计算公式与作用。

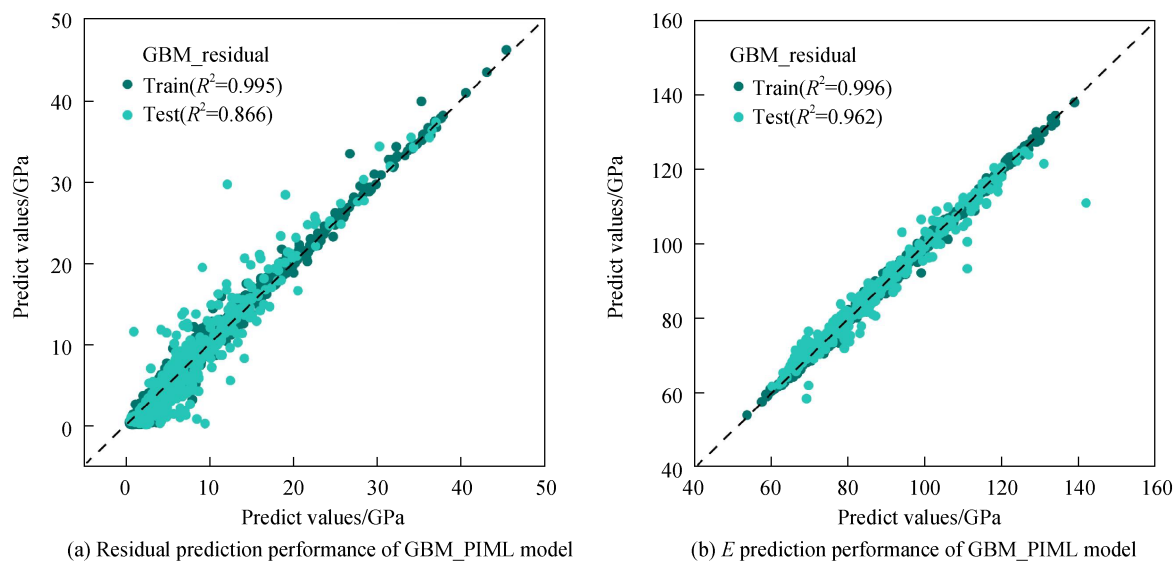


图6 使用高质量数据集的GBM_PIML模型预测性能图

Fig. 6 Prediction performance plot of GBM_PIML model based on high-quality dataset

表1 数据驱动模型与PIML模型指标对比

Table 1 Performance comparison of data-driven models and PIML model

Model	R^2	MedAE	MSE
GBM	0.910	1.31	22.04
RF	0.892	1.23	26.36
XGB	0.906	1.30	22.90
ANN	0.910	1.34	22.11
SVR	0.856	1.09	35.53
GBM_PIML	0.962	0.77	7.04

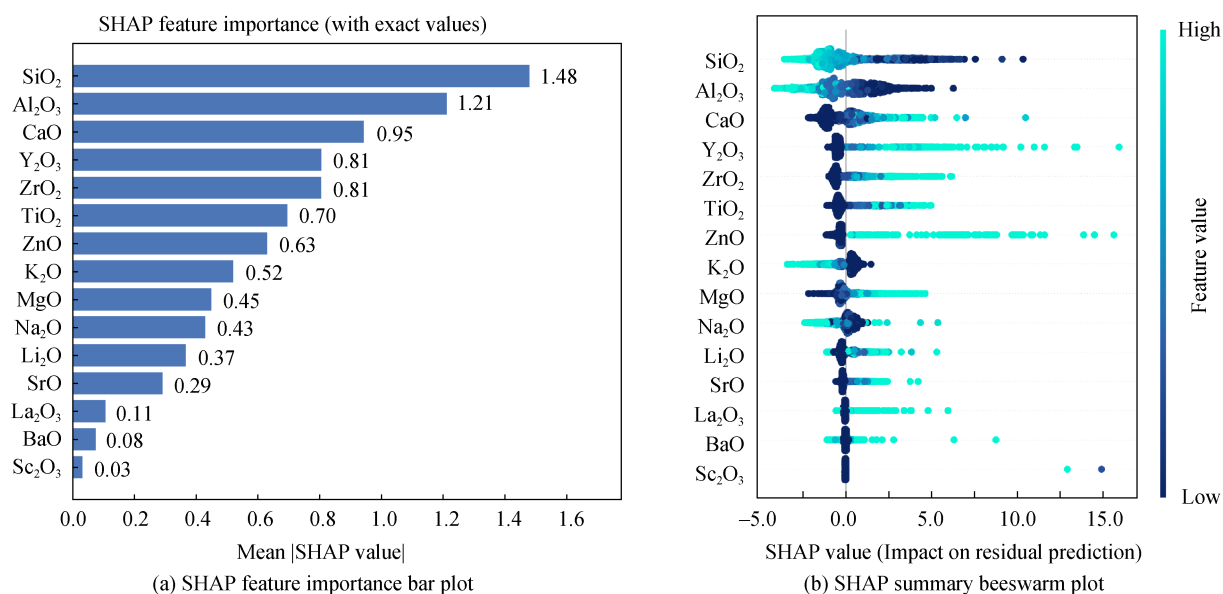


图7 GBM_PIML模型的SHAP分析图

Fig. 7 SHAP analysis diagram of GBM_PIML model

表 2 各物理化学描述符的含义与作用

Table 2 Definition and function of each physicochemical descriptor

Description	Calculation equation	Function
Ma	$Ma = \frac{\sum X_{Acid} + \sum X_{Amphoteric}}{\sum X_{basic}} \quad [22]$ Where X_{Acid} , $X_{Amphoteric}$, and X_{basic} are the molar contents of acidic, amphoteric, and basic oxides, respectively	Characterizes the degree of network polymerization
ACFS	$ACFS = \frac{\sum X_{NWM} CFS_{NWM}}{\sum X_{NWM}}$ Where X_{NWM} and CFS_{NWM} are the content and cation field strength of modifier cations (including alkali metals, alkaline earth metals, and rare earth metal ions)	Characterizes the aggregation effect of high-field-strength ions on non-bridging oxygen (NBO) ^[36]
Re/Al	$Re / Al = \frac{\sum X_{Re}}{\sum X_{Al}}$ Where X_{Re} and X_{Al} are the amounts of rare earth metal ions and aluminum ions in 1 mol of glass, respectively	Characterizes the interaction between rare earth elements and aluminum ions ^[35]
IMO/Si	$IMO / Si = \frac{\sum X_{IMO}}{X_{Si}}$ Where X_{IMO} and X_{Si} are the amounts of intermediate ions (Al^{3+} and Ti^{4+}) and Si^{4+} in 1 mol of glass, respectively	Characterizes the degree to which intermediate oxides act as modifiers

采用方差膨胀因子(VIF)对 4 种物理化学信息描述符进行共线性分析^[26], 由图 8 可见, 4 种描述符的 VIF 值均小于 5, 表明各变量间不存在显著的多重共线性问题。说明设计的 4 种物理化学描述符满足线性回归模型的无关性假设。

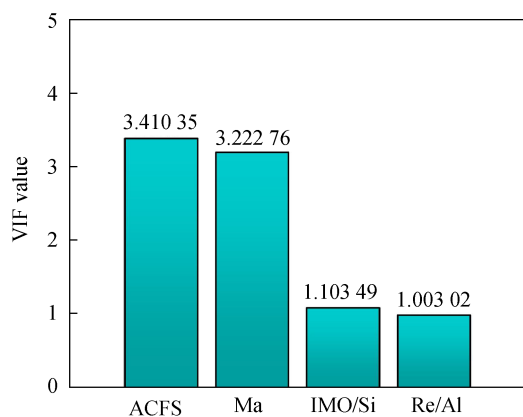


图 8 物理化学信息描述符共线性分析

Fig. 8 Multicollinearity analysis of physicochemical information descriptors

图 9 为物理化学信息描述符预测模型预测性能图, 构建的模型测试集 $R^2=0.932$, 表明所建模型能够有效学习物理化学描述符与 MM 模型计算残差间的内在关联, 保证了 SHAP 分析的可信度。

SHAP 分析结果如图 10 所示。当代表网络聚合度的 Ma 降低与平均场强 ACFS 增加时 MM 模型计算残差随之增大, 与 MM 模型简化高场强离子对非桥氧的局部积聚作用^[35-37], 将玻璃内部结构理想化为均匀体系相符。同时, 计算残差随 IMO/Si 的增加

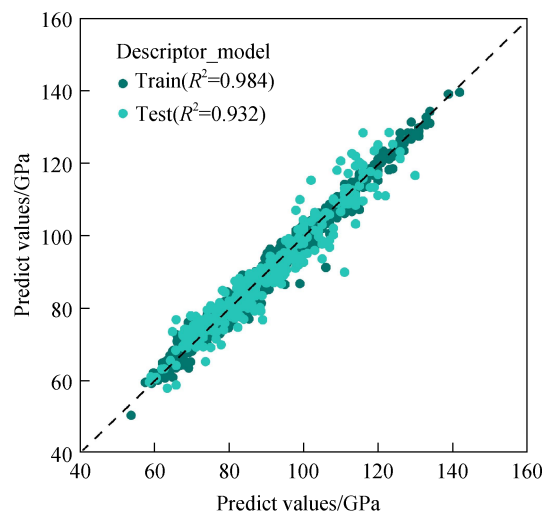


图 9 物理化学信息描述符预测模型性能图

Fig. 9 Performance plot of prediction model with physicochemical information descriptors

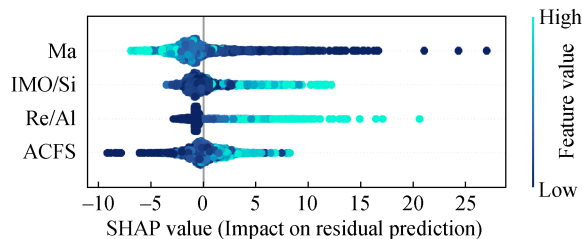


图 10 物理化学信息描述符模型 SHAP 分析图

Fig. 10 SHAP analysis plot of physicochemical information descriptor model

而升高。因为当 IMO/Si 增大时, 多余的网络中间体氧化物转为高配位修饰体进入网络, 但 MM 模型未区分中间体氧化物不同存在形式的离解能, 带来计

算偏差, 与 SHAP 分析相符。 Re/Al 的增加意味更多四配位铝与稀土金属离子相互作用^[34], 体系中高配位铝增加, 而 MM 模型未区分四配位铝与高配位铝的结构差异, 导致计算精度降低, 与 SHAP 分析中计算残差随 Re/Al 的增加而升高一致。

图 11 为测试集中 PIML 模型预测显著优于 MM 模型的样品的氧化物频数分布直方图, 其中绝大多数样本为铝硅酸盐玻璃, 这与 MM 模型忽视不同配位铝的结构差异的结论相符。Shi 等^[24]的研究明确区分了不同配位铝对堆积密度、解离能的不同影响, 优化了 MM 模型, 并显著提升了计算准确度。然而, 目前获取不同配位铝含量尚无比 NMR 更简便的方法, 而 NMR 的测试成本甚至高于直接测量杨氏模量, 限制了优化公式的应用。因此, 有必要收集铝配位体含量相关数据, 建立简便的预测模型, 从而降低获取铝配位信息的门槛。

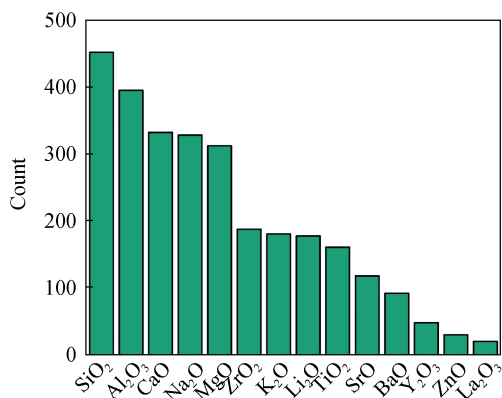


图 11 PIML 模型预测精度优于 MM 模型计算的样本中氧化物频数分布直方图

Fig. 11 Frequency distribution histogram of oxides in samples where PIML model prediction outperforms MM model calculation

2.4 PIML 模型外推能力

外推能力是指模型预测位于训练集范围之外的样本的能力, 纯数据驱动的机器学习模型由于仅能学习到训练数据内部的统计规律, 无法保证模型拟合出的规律仍适用于超出训练数据范围的样本, 因此无法稳定地给出可靠的预测结果。

由于 PIML 模型将物理先验与机器学习建模流程融合, 模型内部预测流程不仅要遵守训练集得出的数据规律, 还受物理规律的约束。当面对位于训练集范围之外的样本时, 即使数据规律失准, 仍可以依靠物理规律得到相对准确的结果, 提升机器学习模型的外推能力。

图 12 为训练模型所用数据集中修饰体氧化物的分布范围。由图 12 可见, 各氧化物含量已逼近玻

璃形成能力的极限。选取 Li_2O 、 Y_2O_3 、 La_2O_3 和 ZrO_2 4 种常用于提升玻璃模量的且在训练集分布范围较小的氧化物用于外推测试。采用其位于训练集范围外的 29 个域外样本进行外推测试。将样本分别输入 GBM 与 GBM_PIML 2 个模型中进行预测并计算预测误差, 得到 2 种模型外推预测误差分布图如图 13 所示。GBM_PIML 模型的预测误差显著低于纯数据

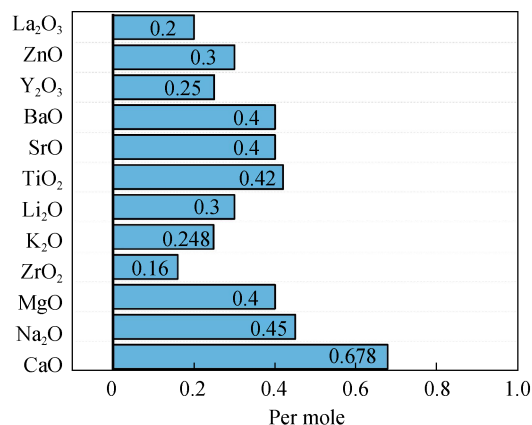
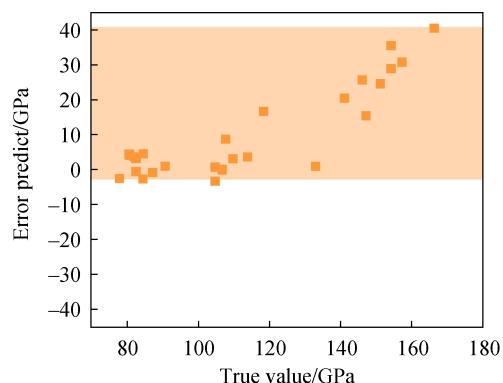
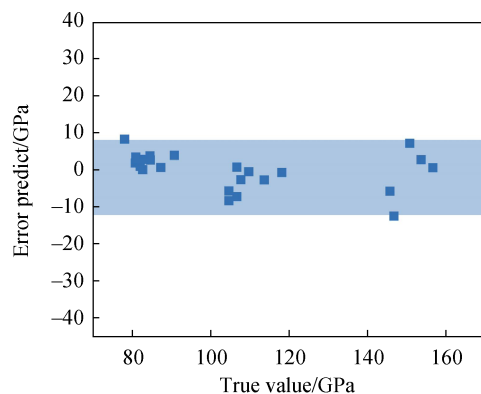


图 12 数据集中各氧化物含量范围图

Fig. 12 Content range diagram of each oxide in the dataset



(a) GBM extrapolation error distribution histogram



(b) GBM_PIML extrapolation error distribution histogram

图 13 GBM 与 GBM_PIML 模型外推预测误差分布图
Fig. 13 Extrapolation prediction error distribution plot of GBM and GBM_PIML models

拟合的 GBM 模型, 并且前者的误差不随样本真实模量的提升而出现剧烈变化, 在高模量区域仍能提供稳定的预测结果。这说明物理信息的引入有效提升了模型的外推能力。

3 结论

1) 将 MM 物理模型与 GBM 机器学习结合, 构建以残差为预测目标的物理信息机器学习(PIML)模型, 测试集 R^2 达 0.962, MedAE=0.77、MSE=7.04, 模型拟合能力、鲁棒性、预测精度显著优于传统纯数据驱动模型。

2) SHAP 分析表明, MM 模型的系统误差主要源于未考虑高场强离子、氧化物相互作用对杨氏模量的贡献。

3) 外推测试表明, PIML 模型在训练集域外样本上仍保持低误差与高稳定性, 优于数据驱动模型。

4) 构建的 PIML 模型为高模量玻璃的组分开发与性能预测提供了一种兼具高预测精度、可解释性与强外推能力的方法。

(第二十八届中国科协年会学术论文)

参考文献:

- [1] LIU G, ZHANG X F, CHEN X L, et al. Additive manufacturing of structural materials[J]. *Mater Sci Eng R Rep*, 2021, 145: 100596.
- [2] WAN J T, GAN B, LI C, et al. A sustainable, eugenol-derived epoxy resin with high biobased content, modulus, hardness and low flammability: Synthesis, curing kinetics and structure-property relationship[J]. *Chem Eng J*, 2016, 284: 1080–1093.
- [3] KANG Z Y, WANG L J, LI X Y, et al. Interpretable machine learning accelerates development of high-specific modulus glass[J]. *Comput Mater Sci*, 2025, 246: 113482.
- [4] BISHNOI S, SINGH S, RAVINDER R, et al. Predicting Young's modulus of oxide glasses with sparse datasets using machine learning[J]. *J Non Cryst Solids*, 2019, 524: 119643.
- [5] HU Y J, ZHAO G, ZHANG M F, et al. Predicting densities and elastic moduli of SiO₂-based glasses by machine learning[J]. *npj Comput Mater*, 2020, 6: 25.
- [6] QIN J C, ZHANG F Q, MA M S, et al. Multiobjective optimization of dielectric, thermal, and mechanical properties of inorganic glasses utilizing explainable machine learning and genetic algorithm[J]. *Mater Genome Eng Adv*, 2025, 3(2): e70005.
- [7] LIU Y, YANG Z W, ZOU X X, et al. A general framework to govern machine learning oriented materials data quality[J]. *Mater Sci Eng R Rep*, 2025, 166: 101050.
- [8] LIU Y, GUO B R, ZOU X X, et al. Machine learning assisted materials design and discovery for rechargeable batteries[J]. *Energy Storage Mater*, 2020, 31: 434–450.
- [9] LIU Y, WU J M, AVDEEV M, et al. Multi-layer feature selection incorporating weighted score-based expert knowledge toward modeling materials with targeted properties[J]. *Adv Theory Simul*, 2020, 3(2): 1900215.
- [10] LIU Y, ZOU X X, MA S C, et al. Feature selection method reducing correlations among features by embedding domain knowledge[J]. *Acta Mater*, 2022, 238: 118195.
- [11] 刘悦, 邹欣欣, 杨正伟, 等. 材料领域知识嵌入的机器学习[J]. *硅酸盐学报*, 2022, 50(3): 863–876.
LIU Yue, ZOU Xinxin, YANG Zhengwei, et al. *J Chin Ceram Soc*, 2022, 50(3): 863–876.
- [12] 刘悦, 马舒畅, 杨正伟, 等. 面向材料领域机器学习的数据质量治理[J]. *硅酸盐学报*, 2023, 51(2): 427–437.
LIU Yue, MA Shuchang, YANG Zhengwei, et al. *J Chin Ceram Soc*, 2023, 51(2): 427–437.
- [13] LIU Y, YANG Z W, ZOU X X, et al. Data quantity governance for machine learning in materials science[J]. *Natl Sci Rev*, 2023, 10(7): 229–245.
- [14] LIU Y, MA S C, YANG Z W, et al. Domain knowledge-assisted materials data anomaly detection towards constructing high-performance machine learning models[J]. *J Materiomics*, 2025, 11(6): 101066.
- [15] SHI S Q, SUN S Y, MA S C, et al. Detection method on data accuracy incorporating materials domain knowledge[J]. *J Inorg Mater*, 2022, 37(12): 1311.
- [16] 杨正伟, 刘子图, 张玉颖, 等. 材料领域知识嵌入的分而治之机器学习[J]. *硅酸盐学报*, 2026, 54(1): 14–24.
YANG Zhengwei, LIU Zitu, ZHANG Wangying, et al. *J Chin Ceram Soc*, 2026, 54(1): 14–24.
- [17] LIU Y, WANG S Y, YANG Z W, et al. Auto-MatRegressor: Liberating machine learning alchemists[J]. *Sci Bull*, 2023, 68(12): 1259–1270.
- [18] LIU Y, WU L H, YANG Z W, et al. Descriptors divide-and-conquer enables multifaceted and interpretable materials structure-activity relationship analysis[J]. *Adv Funct Mater*, 2025, 35(26): 2421621.
- [19] LIU Y, WU J M, WANG Z C, et al. Predicting creep rupture life of Ni-based single crystal superalloys using divide-and-conquer approach based machine learning[J]. *Acta Mater*, 2020, 195: 454–467.
- [20] LIU Y, WU J M, YANG G, et al. Predicting the onset temperature (T_g) of Ge_xSe_{1-x} glass transition: A feature selection based two-stage support vector regression method[J]. *Sci Bull*, 2019, 64(16): 1195–1203.
- [21] LIU Y, ZHAO T L, YANG G, et al. The onset temperature (T_g) of As_xSe_{1-x} glasses transition prediction: A comparison of topological and regression analysis methods[J]. *Comput Mater Sci*, 2017, 140: 315–321.
- [22] WANG X M, KAN Q H, PETRU M, et al. Study on the composition-property relationships of basalt fibers based on symbolic regression and physics-informed neural network[J]. *Compos Part A Appl Sci Manuf*, 2024, 185: 108324.
- [23] MAKISHIMA A, MACKENZIE J D. Direct calculation of Young's modulus of glass[J]. *J Non Cryst Solids*, 1973, 12(1): 35–45.
- [24] SHI Y, TANDIA A, DENG B H, et al. Revisiting the Makishima-Mackenzie model for predicting the Young's modulus of oxide glasses[J]. *Acta Mater*, 2020, 195: 252–262.
- [25] WILLARD J, JIA X W, XU S M, et al. Integrating scientific knowledge with machine learning for engineering and environmental systems[J]. *ACM Comput Surv*, 2023, 55(4): 1–37.
- [26] CASSAR D R. GlassNet: A multitask deep neural network for predicting many glass properties[J]. *Ceram Int*, 2023, 49(22): 36013–36024.
- [27] PEDREGOSA F, VAROQUAUX G, GRAMFORT A, et al. Scikit-learn: Machine Learning in Python[J]. *J Mach Learn Res*, 2011, 12: 2825–2830.

- [28] CHEN T Q, GUESTRIN C. XGBoost: A scalable tree boosting system[C]//Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco California USA. ACM, 2016: 785-794.
- [29] BERGSTRA J, YAMINS D, COX D D. Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures[C]//International Conference on Machine Learning, 2013
- [30] CHICCO D, WARRENS M J, JURMAN G. The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation[J]. Peerj Comput Sci, 2021, 7: e623.
- [31] JĚKABSON S G, LAVENDELS J, SITIKOV S V. Model evaluation and selection in multiple nonlinear regression analysis[J]. Math Model Anal, 2007, 12(1): 81–90.
- [32] BHATTOO R, BISHNOI S, ZAKI M, et al. Understanding the compositional control on electrical, mechanical, optical, and physical properties of inorganic glasses with interpretable machine learning[J]. Acta Mater, 2023, 242: 118439.
- [33] LUNDBERG S M, ERION G, CHEN H, et al. From local explanations to global understanding with explainable AI for trees[J]. Nat Mach Intell, 2020, 2(1): 56–67.
- [34] FLORIAN P, SADIKI N, MASSIOT D, et al. ²⁷Al NMR study of the structure of lanthanum- and yttrium-based aluminosilicate glasses and melts[J]. J Phys Chem B, 2007, 111(33): 9747–9757.
- [35] SEN S, STEBBINS J F, KROEKER S, et al. Evidence for mixed Mg coordination environments in silicate glasses: Results from ²⁵Mg NMR spectroscopy at 35. 2 T[J]. J Phys Chem B, 2023, 127(49): 10659–10666.
- [36] GANISETTI S, GADDAM A, KUMAR R, et al. Elucidating the formation of Al–NBO bonds, Al–O–Al linkages and clusters in alkaline-earth aluminosilicate glasses based on molecular dynamics simulations[J]. Phys Chem Chem Phys, 2019, 21(43): 23966–23977.
- [37] LEIGHT C J, YTSMA C, MCCANTA M C, et al. Compositional characterization of glassy volcanic material from VSWIR and MIR spectra using partial least squares regression models[J]. Earth Space Sci, 2024, 11(6): e2023EA003439.

作者贡献声明:

钱兆青: 论文构思、设计研究方法和模型算法、撰写论文、分析/综合数据并推导结论

曹 毅: 设计实验方案、分析/综合数据并推导结论

康泽玉: 参与结论推导

卢亚东: 提出修改意见

高文凯: 提出修改意见

康俊峰: 监督指导、审阅与修订

Construction and Mechanism Analysis of Glass Young's Modulus Prediction Model Using Physics-Informed Machine Learning

QIAN Zhaoqing¹, CAO Yi², KANG Zeyu¹, LU Yadong¹, GAO Wenkai¹, KANG Junfeng^{1,3}

(1. School of Materials Science and Engineering, University of Jinan, Jinan 250022, China; 2. School of Information Science and Engineering, University of Jinan, Jinan 250022, China 3. Shandong Key Laboratory of High Performance Glass Fiber, Taian 271000, Shandong, China)

Extended Abstract

Introduction As a critical reinforcing material, high-modulus glass has experienced a rapidly growing demand in aerospace, wind power and other high-tech industries. This trend makes the development of novel high-performance, high-modulus glass formulations an urgent and significant task for materials science.

Machine learning has emerged as a transformative new research paradigm following traditional trial-and-error experiments and computer simulations. It offers remarkable advantages in efficiency, accuracy, and low energy consumption, and has been widely used to predict various glass properties.

However, purely data-driven machine learning models face severe performance bottlenecks when predicting glass Young's modulus (YM), showing significantly lower accuracy than models for other glass properties. This limitation mainly stems from two factors: poor data consistency in public databases due to the sensitivity of YM to preparation-related thermal history and testing conditions, and the lack of effective strategies to eliminate abnormal samples, which leads to model overfitting and limited generalization and extrapolation capabilities.

Liu and co-workers showed that pure data-driven models face three major contradictions in materials science, and proposed using data governance and domain knowledge embedding to improve model performance. Data quality is decisive for model performance, and combining anomaly detection with a divide-and-conquer strategy can yield accurate and interpretable models.

The Makishima–Mackenzie (MM) model is a classic semi-empirical model for predicting the YM of oxide glass with clear physical significance, but it exhibits a well-documented systematic underestimation of actual YM values. Residual modeling has been proven to be an effective approach to address the deficiencies of physical models. Therefore, this study aims to develop a physics-informed machine learning (PIML) model by integrating the MM model with machine learning to overcome the limitations of pure data-driven models.

Methods The original dataset was constructed from the SciGlass database, which contained 2861 samples, with 15 core oxides as features and YM and density as target variables. Basic data cleaning was first performed to remove samples with incomplete

compositions, missing values, zero values, negative values, and duplicate formulations.

Five pure data-driven models, namely Gradient Boosting Machine (GBM), Random Forest (RF), eXtreme Gradient Boosting (XGB), Artificial Neural Network (ANN) and Support Vector Regression (SVR), were trained on the original dataset. All models were optimized using Bayesian optimization and five-fold cross-validation, with the dataset split into training and test sets in an 8:2 ratio.

Subsequently, the MM model was used to calculate the theoretical YM values for all the samples. Based on the physical principle that the MM model systematically underestimates glass YM, samples with negative residuals (measured YM minus MM-calculated YM) were eliminated. This procedure yielded a high-quality dataset of 2463 samples constrained by physical principles.

The GBM model, which exhibited the best performance among the three pure data-driven models, was selected for constructing the PIML model. Instead of directly predicting YM, the PIML model used oxide compositions as input features and the MM model residuals as the prediction target, and was trained on the high-quality dataset. The final YM prediction was obtained by adding the predicted residual to the MM-calculated value.

Model performance was evaluated using three comprehensive metrics: the coefficient of determination (R^2), mean squared error (MSE), and median absolute error (MedAE). SHapley Additive exPlanations (SHAP) analysis and physicochemical-descriptor modeling were employed to reveal the error mechanism of the MM model. Extrapolation capability was tested using 29 samples with Li_2O , Y_2O_3 , La_2O_3 and ZrO_2 contents outside the training set range.

Results and discussion The pure data-driven GBM model achieved a test-set R^2 of 0.910 on the original dataset, representing the best performance among the three models. However, its performance was still substantially limited by the abnormal samples present in the dataset.

In contrast, the developed GBM-PIML model trained on the physically constrained high-quality dataset achieved a markedly improved test set R^2 of 0.962. Compared with the pure GBM model, its MedAE was reduced by approximately 40% and its MSE was reduced by approximately 70%. These results demonstrated that the physical-constraint-based data cleaning effectively removed misleading abnormal samples, and the residual learning strategy reduced the learning difficulty of the model by allowing the model to learn the systematic error pattern of the MM model rather than the entire complex composition-YM relation.

SHAP analysis revealed that increases in Y_2O_3 , La_2O_3 , MgO , and Li_2O contents significantly increased the MM model residuals. This result was attributed to the fact that the MM model treated glass as a homogeneous system and ignored two key effects: the accumulation of non-bridging oxygen by high-field-strength modifiers and the interaction between rare earth metals and aluminum ions.

A physicochemical descriptor model was further constructed using four descriptors (Ma, ACFS, Re/Al, IMO/Si) that characterize internal glass structure and oxide interactions and achieved a test set R^2 of 0.932. SHAP analysis of this model confirmed that a lower degree of network polymerization, higher average cation field strength, higher Re/Al ratio and higher IMO/Si ratio all contributed to larger MM model residuals, which was consistent with the underlying physical mechanism.

Extrapolation tests showed that the GBM-PIML model exhibited significantly lower prediction errors than the pure GBM model in the high-modulus region. Most importantly, its error did not increase sharply as YM increased, indicating strong extrapolation stability.

Conclusions This study constructs a residual-learning-based PIML model by integrating the MM physics-based model with the GBM machine-learning algorithm. The model achieves a test set R^2 of 0.962, a MedAE of 0.77, and an MSE of 7.04, and demonstrates superior prediction accuracy, robustness and generalization capability compared to conventional purely data-driven models trained on the same original dataset.

The combination of SHAP analysis and physicochemical descriptor modeling provides deep insight into the error mechanism of the MM model. The results reveal that the systematic error of the MM model mainly originates from its neglect of the effects of high-field-strength ions and oxide interactions on glass YM.

Extrapolation tests confirm that the incorporation of physical prior knowledge significantly enhances the model's extrapolation ability, especially in the high-modulus region. The proposed PIML model provides a new paradigm for the rapid development of high-modulus glass formulations with high prediction accuracy, interpretability, and strong extrapolation capability.

Keywords physics-informed machine learning; Makishima-Mackenzie model; high-modulus glass; SHAP analysis