科学通报 2026 年 第 71 卷 第 23 期: 5627 ~ 5637
人工智能-能源材料交叉前沿专题 论 文
《中国科学》杂志社 SCIENCE CHINA PRESS

# 物理可信软标签驱动的钙钛矿本征击穿强度预测与可解释建模

李长姣[1], 彭奥[1], 王琳[1*], 高文强[2], 刘韩星[3], 周伟家[2*]

1. 济南大学信息科学与工程学院, 山东省泛在智能计算重点实验室, 济南 250022
2. 济南大学前沿交叉科学研究院, 济南 250022
3. 武汉理工大学材料科学与工程学院, 材料复合新技术全国重点实验室, 武汉 430070
* 联系人, E-mail: ise_wanglin@ujn.edu.cn; ifc_zhouwj@ujn.edu.cn

2025-12-29 收稿, 2026-03-03 修回, 2026-03-04 接受, 2026-03-05 网络版发表
国家自然科学基金(61872419)、济南大学学科交叉融合建设项目(XKJC-202402)和济南大学优秀博士项目(XBS2515)资助

**摘要** 钙钛矿材料因其独特的结构稳定性与组成可调性, 成为高电场耐受介电材料极具潜力的候选体系. 本征击穿强度($F_b$)作为表征其击穿理论极限的参数, 存在计算成本高昂、标签数据严重匮乏等问题, 严重阻碍了该体系电场耐受机制研究与高理论击穿极限材料筛选. 针对上述问题, 本文提出一种物理可信软标签驱动的$F_b$预测建模框架. 基于现有高质量$F_b$标签的绝缘体数据集构建$F_b$高精度概率化预测模型后, 将其迁移至钙钛矿材料体系, 以生成含有不确定性量化的软标签. 基于此构建了监督可信度加权的$F_b$准确预测模型(验证集$R^2$=0.8952), 并通过单特征主效应分析、特征间协同作用解析等方法, 对模型内部决策逻辑进行可视化分析, 确保预测结果的物理可信性. 基于该模型, 本文构建了基于可信度加权的符号回归方法, 将重要特征对$F_b$值的影响规律映射为可显式计算的数学关系式. 该关系式在真实钙钛矿样本上的预测误差较已发表$F_b$值计算公式降低1~2个数量级, 且对高不确定性样本的预测偏差缩小60%~80%. 本文所提出的软标签构建策略与基于可信度加权的符号回归方法, 不仅显著缓解了因高质量标签数据稀缺所导致的建模局限问题, 更为广泛材料体系在标签数据匮乏场景下的本征物性预测提供了一种可迁移的建模范式.

**关键词** 钙钛矿材料, 本征击穿强度, 软监督建模, 物理可解释性

介电电容器凭借高功率密度与快速充放电响应特性, 已成为脉冲功率系统、混合动力系统等先进应用领域的核心部件, 并为新能源、先进电子电力等新兴产业的发展提供了关键技术支撑. 随着电子设备持续向微型化、高集成化方向发展, 对具备更高电场耐受能力的新型电介质材料需求日益迫切. 本征击穿强度(intrinsic dielectric breakdown strength, $F_b$)作为反映电介质材料击穿理论极限的核心参数, 决定了材料在高电场条件下的服役边界、能量存储上限及长期工作稳定性[1~7]. 因此, 系统探索介电材料高$F_b$的内在物理机制, 设计并筛选具有优异电场耐受能力的新型介电材料, 对推动光伏、电力电子等相关领域的持续创新发展具有重要意义[8~10].

钙钛矿结构($ABX_3$)材料因其结构稳定性与组成元素的高度可调性, 在光电[11~13]、催化[14~16]、电学[17,18]等众多领域表现出优异的性能与设计潜力. 该特性使其可通过精准的成分设计与结构调控, 实现介电性能关键参数的协同优化与突破, 被视为当前高电场耐受介电材料研发中极具潜力的候选体系[18~22]. 然而, 目前先进的实验及表征技术仍无法精确剥离和量化各外

**引用格式:** 李长姣, 彭奥, 王琳, 等. 物理可信软标签驱动的钙钛矿本征击穿强度预测与可解释建模. 科学通报, 2026, 71: 5627–5637
Li C, Peng A, Wang L, et al. Physics-informed soft label-driven prediction and interpretable modeling of intrinsic dielectric breakdown strength for perovskites (in Chinese). Chin Sci Bull, 2026, 71: 5627–5637, doi: 10.1360/CSB-2025-5998

© 2026《中国科学》杂志社

在因素(如缺陷、杂质、界面等)在电介质击穿过程中的具体影响, 使得基于电子-声子散射理论进行量子力学计算的方法成为当前获取$F_b$值最为有效的途径之一. 但该方法存在计算成本高昂、周期长等固有缺陷, 致使该体系高质量$F_b$数据长期处于匮乏状态, 从而使得系统探索钙钛矿材料高$F_b$的物理根源, 并在其广阔设计空间内进行高效的材料筛选与设计, 面临巨大挑战[23,24].

近年来, 机器学习(machine learning, ML)方法凭借其在复杂非线性关系建模方面的优势, 逐渐成为挖掘材料数据规律的有力工具, 为在广阔的钙钛矿材料设计空间内发现高$F_b$的候选体系提供了新的思路. 然而, 在钙钛矿材料$F_b$性能(以及更广泛的材料本征物性)预测中, 高质量标签数据获取困难、可用样本规模严重不足的问题尤为突出, 严重影响了模型的预测准确性、泛化能力及可靠性[25,26], 也成为制约ML方法在高电场耐受钙钛矿材料研发中的关键瓶颈. 为应对该问题, Kim等人[27]利用ML方法, 以高精度杂化泛函HSE06计算带隙值$E_{g(HSE06)}$和最大声子截止频率($\omega_{max}$)为关键描述符, 基于多种绝缘材料构建了$F_b$的拟合计算公式[8,28,29], 并将其用于计算钙钛矿材料的$F_b$标签值. 然而, 该工作在跨体系迁移过程中, 无法量化其计算结果的可信程度, 且依赖于高精度描述符$E_{g(HSE06)}$和$\omega_{max}$, 极大增加了计算成本. Li等人[3]利用XGBoost算法构建了钙钛矿材料$F_b$分步预测框架, 利用低计算成本描述符构建$\omega_{max}$预测模型后, 基于$\omega_{max}$预测值与$E_{g(HSE06)}$真实值建立了$F_b$预测模型. 该方法有效解决了关键描述符$\omega_{max}$计算成本高昂的问题. 但其$F_b$训练阶段由低质量描述符引入的系统误差, 以及模型本身的黑盒特性, 使其预测结果的可信度和可解释性存在明显局限. 因此, 面对钙钛矿材料庞大的设计空间, 如何在高质量$F_b$标签数据匮乏、关键描述符计算成本高昂的双重约束下, 构建同时具备物理一致性、模型可解释性与跨体系可迁移性的$F_b$性能预测模型, 已成为该领域的核心挑战.

针对上述挑战, 本文提出一种基于物理可信软标签的钙钛矿材料$F_b$预测框架. 以具备高质量$F_b$标签的二元绝缘体数据集为基础, 构建概率化$F_b$预测模型并迁移至钙钛矿材料体系, 为钙钛矿样本生成了同时包含预测值与不确定性估计的软标签, 从而有效缓解了$F_b$标签数据缺失带来的挑战. 在此基础上, 系统引入了元素属性、化学键以及结构畸变参数等多层次、易获取的特征描述符, 并结合基于低计算成本广义梯度近似泛函PBE (Perdew-Burke-Ernzerhof)的带隙值$E_{g(PBE)}$, 逐步降低模型对高精度$E_{g(HSE06)}$与$\omega_{max}$的直接依赖. 进一步地, 通过构建基于软标签的钙钛矿$F_b$预测模型并筛选获得支撑模型性能的核心特征集, 在保证预测精度的同时显著降低特征计算成本. 随后, 综合运用模型无关的可解释分析方法与符号回归技术, 在验证预测结果物理可信的同时, 系统性地揭示了钙钛矿材料获得高$F_b$的潜在物理机制, 并构建了映射影响规律的高精度数学关系式. 本研究在标签数据缺失的条件下实现了钙钛矿材料$F_b$高效预测与规律挖掘, 也为广泛材料本征物性在高质量标签数据缺失情况下的准确预测提供了一种可迁移的建模范式.

## 1 方法

### 1.1 跨体系软标签数据集构建

现有的$F_b$量子力学计算方法通常基于电子-声子散射理论. 在该框架下, 介电材料的本征击穿行为由载流子能量积累与晶格能量耗散之间的竞争所主导. 反映电子能量积累能力的$E_{g(HSE06)}$与表征晶格对高能载流子能量耗散上限的$\omega_{max}$, 被广泛认为是描述$F_b$的核心物理量. 因此, 本文采用高斯过程回归(Gaussian Process Regression, GPR), 通过在$E_{g(HSE06)}$-$\omega_{max}$特征空间中对绝缘体体系下$F_b$与物理描述符之间的映射关系进行推断. 基于文献报道的82个拥有高质量$F_b$标签的绝缘体样本训练GPR模型后(验证集上$R^2$=0.90, 远超已报道$F_b$预测工作$R^2$=0.72[27]), 将该模型迁移至AFLOWlib (http://www.aflowlib.org)数据库内的203个钙钛矿样本, 给出可作为$F_b$标签值的预测均值($\mu$)的同时, 提供预测标准差($\sigma$)用于不确定性估计, 以量化由有限训练样本中特征空间覆盖不足所引入的外推误差. 最终生成同时包含预测均值与样本级不确定性估计的$F_b$软标签, 构建钙钛矿材料$F_b$软标签数据集. 模型参数及软标签生成细节见补充文件附录1. 结果如图1所示, 多数钙钛矿样本位于GPR模型预测不确定性较低的区间内, 表明这些样本的预测均值具有较高的信息价值. 该策略在有效解决了真实$F_b$标签匮乏问题的同时, 通过引入外推不确定性量化机制, 对监督信息的质量施加了约束, 为后续钙钛矿材料$F_b$性能预测与可解释建模提供了可靠的数据基础.

### 1.2 多层次特征集构建

为在不引入高计算成本物性描述符的前提下, 充

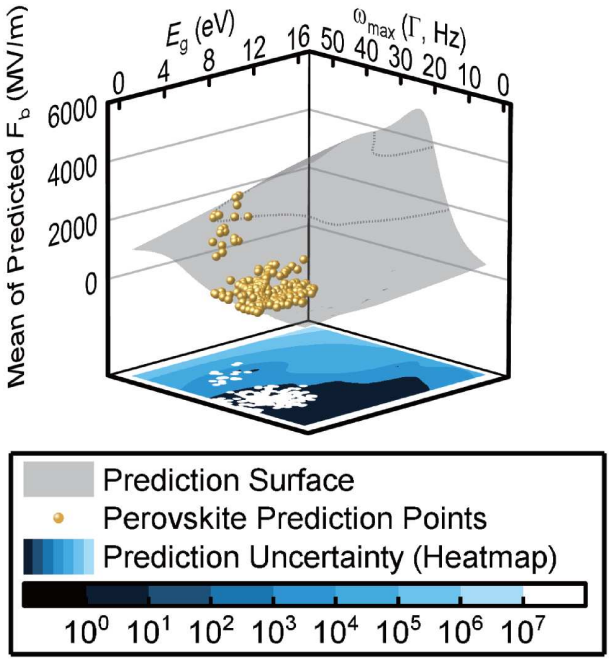



**图 1** GPR模型对钙钛矿材料的预测均值及其不确定性表征
**Figure 1** GPR-based prediction mean and uncertainty characterization of perovskite materials

分表征钙钛矿材料在化学组成与晶体结构上的多样性, 本工作基于软标签数据集系统构建了多层次、易获取的特征集. 在元素层面, 引入钙钛矿A位、B位及X位元素的原子半径、电负性、第一电离能等基本物理化学参数. 在化学键与晶格层面, 引入B-X位离子成键键长、晶格常数等几何结构参数. 此类特征(如表1所示)借助Pymatgen[30]工具对晶体结构解析获取. 鉴于$E_{g(HSE06)}$值计算成本高昂不利于大规模预测应用, 本工作选用低计算成本的$E_{g(PBE)}$对钙钛矿材料电子特性进行描述. 由此, 系统构建了包含元素基本属性、化学键键长、结构畸变参数及电子结构特性等50个多层次、易获取信息构成的特征集, 在降低对高计算成本物性描述符$E_{g(HSE06)}$与$\omega_{max}$的直接依赖的同时, 以低成本元素属性与结构信息对材料的局域成键环境、整体结构特征及电子结构响应进行协同描述, 为后续软监督建模提供具有明确物理意义的输入空间.

### 1.3 软监督建模与关键特征筛选

为充分利用现有样本并避免低可信度样本对模型训练过程产生的干扰, 本文在钙钛矿多层次特征与$F_b$建模阶段, 引入标签值的不确定性构建基于可信度加权的软监督学习框架. 利用XGBoost模型的自定义加权损失函数特性及其在小规模数据下的优秀泛化性能, 本文采用如下加权均方误差损失函数(式1)进行XGBoost模型训练:

**表 1** 钙钛矿材料多层次特征集
**Table 1** Multi-level feature set of perovskite materials

| 特征符号 | 物理意义 | 特征符号 | 物理意义 |
|---|---|---|---|
| | 组成元素信息 | | |
| $M_{(A)}$, $M_{(B)}$, $M_{(X)}$ | 元素(平均)原子质量 | $V_{mol(A)}$, $V_{mol(B)}$, $V_{mol(X)}$ | 元素(平均)摩尔体积 |
| $R_{a(A)}$, $R_{a(B)}$, $R_{a(X)}$ | 元素(平均)原子半径 | $R_{i(A)}$, $R_{i(B)}$, $R_{i(X)}$ | 元素(平均)离子半径 |
| $R_{vdw(A)}$, $R_{vdw(B)}$, $R_{vdw(X)}$ | 元素(平均)范德华半径 | $R_{c(A)}$, $R_{c(B)}$, $R_{c(X)}$ | 元素(平均)共价半径 |
| $X_{(A)}$, $X_{(B)}$, $X_{(X)}$ | 元素(平均)鲍林电负性 | $EA_{(A)}$, $EA_{(B)}$, $EA_{(X)}$ | 元素(平均)电负性 |
| $FIE_{(A)}$, $FIE_{(B)}$, $FIE_{(X)}$ | 元素(平均)第一电离能 | $B_{v(A)}$, $B_{v(B)}$, $B_{v(X)}$ | 元素(平均)键价矢量和 |
| $S_{ij(A)}$, $S_{ij(B)}$, $S_{ij(X)}$ | 元素(平均)键价值 | $t_{BV(Viox)}$, $t_{BV(Vical)}$, $t_{BV(Sij)}$ | 由键价理论计算的容限因子 |
| | 化学键信息 | | |
| $d_{(B-X)}$ | B-X位离子平均成键键长 | $\sigma<d>_{(B-X)}$ | B-X位离子成键键长标准差 |
| $V_{(B-X)}$ | B-X键平均最近邻耦合强度 | $B_{(B-X)}$ | B-X位离子平均键能 |
| $D_i$ | 整体结构的平均差异因子 | $\sigma<D_i>$ | 整体结构的差异因子标准差 |
| GII | 整体结构的全局不稳定系数 | | |
| | 晶体结构整体信息 | | |
| Volume | 晶胞体积 | Density | 晶胞密度 |
| $a$, $b$, $c$, $\alpha$, $\beta$, $\gamma$ | 结构参数 | | |
| | 电子密度信息 | | |
| $E_{g(PBE)}$ | PBE泛函计算的带隙 | | |

$$L=\frac{1}{N}\sum_{i=1}^{N}\omega_i(\hat{y}_i-\mu_i)^2, \tag{1}$$

其中, $\omega_i=\frac{1}{\sigma_i^2+\varepsilon}$, 表示模型预测值, $\mu_i$为软标签值, $\sigma_i$为标签值的不确定性, $\varepsilon$为数值平滑项. 通过将该加权均方误差损失函数融入XGBoost模型, 实现钙钛矿$F_b$值回归建模时, 模型对不同标签可信度样本权重的调节. 即样本$F_b$标签值的不确定性越高(即$\sigma_i$值越大), 该类样本在模型训练过程中的权重将会通过损失函数降低. 相反, $F_b$标签值的不确定性越低(即$F_b$标签值越可靠), 其在训练中的权重越高. 利用该机制, 模型会在训练过程中优先拟合高可信度样本, 并在采纳低可信度样本信息的同时, 削弱其标签值误差对模型学习准确性的影响. 以此解决现有迁移学习工作中, 将迁移模型预测结果直接作为新体系样本标签值时, 由特征值空间分布不同而引起的外推样本预测结果误差大的问题. 本工作XGBoost模型软监督建模细节见补充文件附录2.

基于此软监督策略, 将203个钙钛矿样本划分为163个训练样本与40个验证样本进行模型评估. 结果表明(如图2所示), 模型在训练集与验证集上均表现出稳定的收敛趋势(训练集$R^2$=0.9806, 验证集$R^2$=0.8819), 证明在样本软标签不确定性差异显著的情况下, 引入可信度加权机制可维持可靠的预测性能.

由于初始特征主要来源于组成元素与晶体结构信息, 各特征间存在显著的相关性与信息重叠, 极易削弱模型在不同钙钛矿体系间的泛化能力与预测一致性, 并放大软标签中固有噪声对模型学习过程的不利影响. 为系统消除冗余信息并提取关键特征子集, 本工作采用置换特征重要性(permutation feature importance, PFI)评估方法, 对特征集中各描述符的重要性进行量化评估. 通过依次置换单一特征取值并监测模型预测误差的相对变化, 判断置换特征对模型决策的重要程度. 在采用基于PFI贡献度由低至高的逐步特征剔除策略后(如图3(a)所示), 得到仅包含四个关键描述符的特征子集: $E_{g(PBE)}$、$d_{(B\text{-}X)}$、$\beta$以及$R_{a(B)}$, 其中各特征PFI如图3(b)所示. 基于该关键特征子集, 模型在验证集上的预测精度反而提高($R^2$=0.8952). 表明钙钛矿材料$F_b$的物理规律可由少量耦合元素、晶体结构与电子性质的物理量进行有效表征.

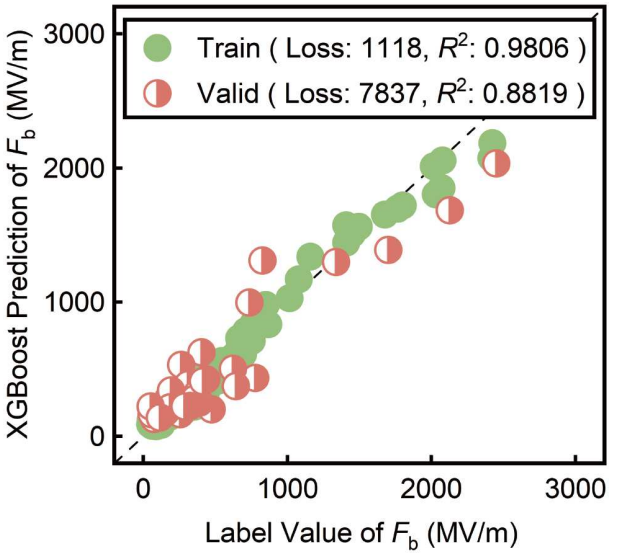



图 2 基于软监督训练的XGBoost拟合效果

**Figure 2** Fitting performance of XGBoost trained under soft supervision

## 2 结果与讨论

为揭示钙钛矿材料$F_b$预测模型的内部决策逻辑, 并将所构建的模型转化为可应用于钙钛矿数据集的高

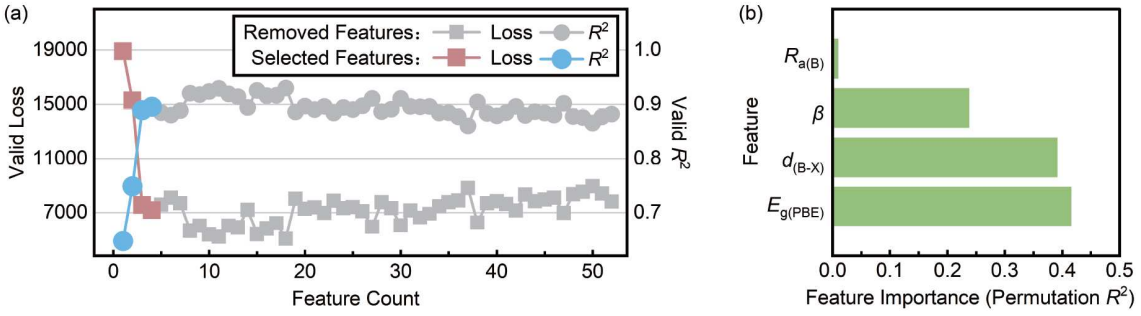



图 3 基于PFI的关键特征筛选. (a) 模型性能随特征剔除的变化曲线; (b) 重要特征的置换特征重要性

**Figure 3** Key feature selection results based on PFI method. (a) Variation curves of model performance with feature elimination; (b) permutation feature importance of the finally selected core features

效筛选工具，本文建立了包含单特征主效应分析、特征间协同调控作用解析及显式关联表达式构建的多层次可解释性分析框架. 基于该框架, 本文对模型所捕捉的$F_b$预测规律与现有物理规律间的一致性进行系统验证, 以进一步增强预测结果的可信度. 同时, 系统提炼出$F_b$关于关键特征的表达式, 并应用于现有钙钛矿材料数据的快速评估与筛选.

## 2.1 特征调控机制解析

为系统揭示模型对单一特征的响应规律，本文首先采用SHAP(SHapley Additive exPlanations)与累积局部效应(accumulated local effects, ALE)相结合的方法, 从局部贡献解释与全局效应趋势两个维度，将各特征值变化对$F_b$预测结果的影响趋势进行可视化，结果如图4所示. 其中，钙钛矿$d_{B\text{-}X}$与$F_b$呈显著负相关，而$E_{g(PBE)}$则与$F_b$呈显著正相关. 该物理趋势与钙钛矿材料介电击穿的主导机制相吻合. 较长的B-X键通常对应较弱的键合强度，在强电场作用下易成为局域电场集中的薄弱区域，从而诱发击穿通道的早期形成[31~33]. 而更宽的带隙提高了载流子跃迁势垒，抑制电子在高场环境下的能量积累, 从而增强材料的介电稳定性[34]. 除此之外，$\beta$的ALE曲线呈明显的分段特征(如图4(d)所示). 在60°~80°区间内，$\beta$值变化对$F_b$的影响相对平缓. 而当$\beta$值趋近90°时，$F_b$预测值出现明显跃降. 该非线性特性与钙钛矿材料中适度结构畸变有助于保持极化弛豫能力，而过度对称则可能削弱介电稳定性的设计规律相吻合[35,36]. 相比之下，$R_{a(B)}$对$F_b$的直接影响较弱(如图4(e)所示), 其作用更主要通过调节B-X键长、键角等局部配位环境间接实现.

为深入探究$F_b$预测中的多因素耦合规律，本工作采用二维偏依赖图(2D partial dependence plots, 2D-PDP)对特征间的非线性交互效应进行可视化分析，结果如图5所示. $E_{g(PBE)}$-$d_{(B\text{-}X)}$及$E_{g(PBE)}$-$\beta$两个特征组合对$F_b$值的调控趋势基本呈现独立叠加的现象，其中某一特征对$F_b$的作用趋势未因另一特征的变化而发生显著偏移, 表明在模型所学习的关联模式中, 电子结构特征与局部几何结构特征分别主导着相对独立的介电击穿调控路径, 其耦合交互效应较弱[8,37,38]. 然而, 在$d_{(B\text{-}X)}$与$\beta$的组合中观察到显著的协同调控特征. 如图5(c)所示, 当B-X位离子键长较大(即键合较弱)时，晶格常数$\beta$对$F_b$表现出近似单调的负向影响. 与之相反，当B-X位离子键长较小(即键合较强)时，晶格常数$\beta$对$F_b$值的影响不再呈现单调规律，其响应曲线表现为随$\beta$值变化呈波动式响应，表明在该区间内晶格畸变对性能的调控方

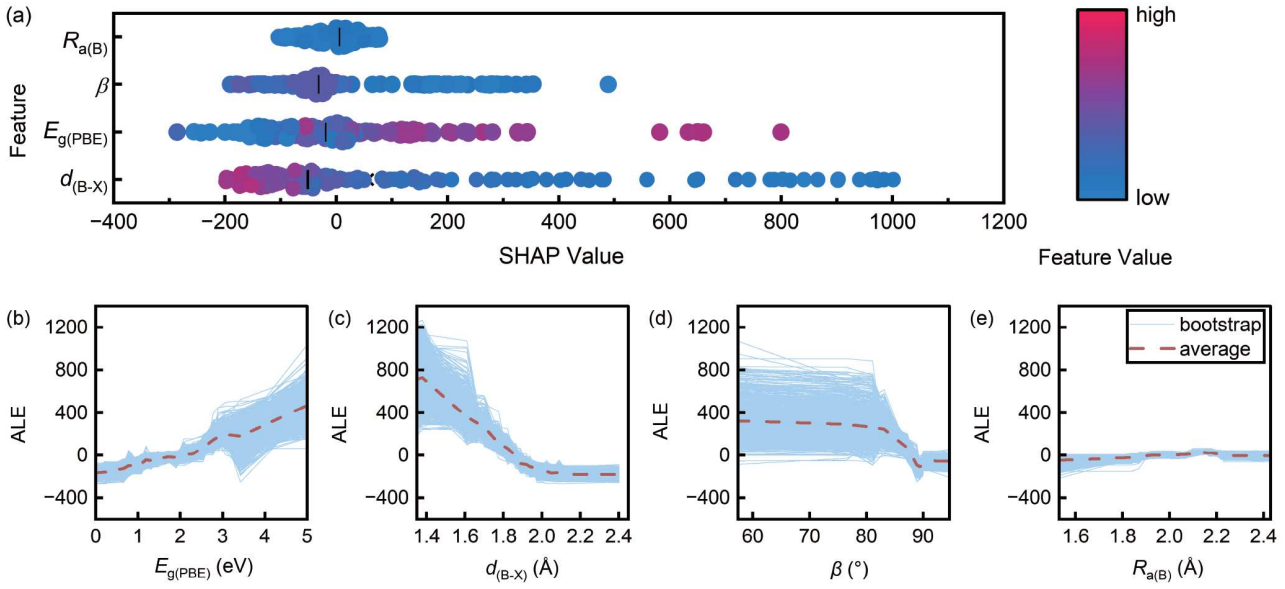



**图 4** 基于XGBoost模型的可解释性分析: 核心特征对$F_b$的作用规律. (a) SHAP分布; (b) ALE曲线: $d_{(B\text{-}X)}$; (c) ALE曲线: $E_{g(PBE)}$; (d) ALE曲线: $\beta$; (e) ALE曲线: $R_{a(B)}$

**Figure 4** Interpretability analysis based on the XGBoost model: Effects of key features on $F_b$. (a) SHAP distribution; (b) ALE curve: $d_{(B\text{-}X)}$; (c) ALE curve: $E_{g(PBE)}$; (d) ALE curve: $\beta$; (e) ALE curve: $R_{a(B)}$

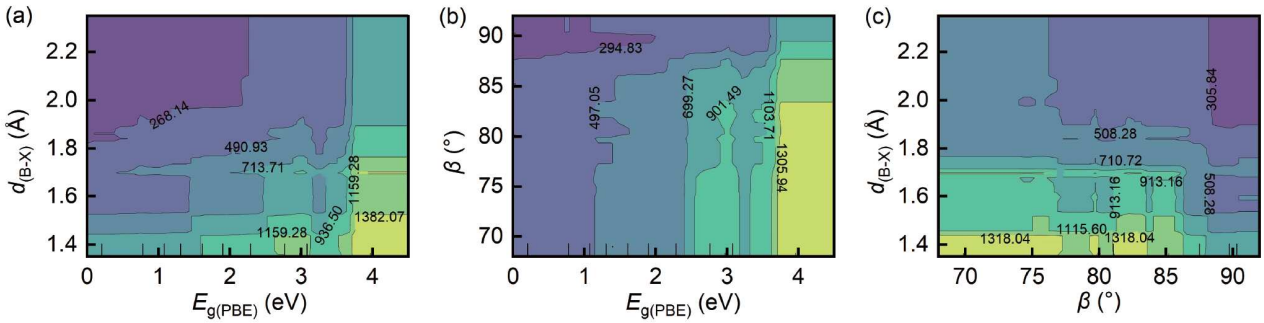



**图 5** 核心特征对$F_b$的协同作用规律(二维PDP图). (a) $E_{g(PBE)}$与$d_{(B-X)}$的协同作用; (b) $E_{g(PBE)}$与$\beta$的协同作用; (c) $d_{(B-X)}$与$\beta$的协同作用

**Figure 5** Synergistic effect of key features on $F_b$ (2D-PDP). (a) Synergistic effect of $E_{g(PBE)}$ and $d_{(B-X)}$; (b) synergistic effect of $E_{g(PBE)}$ and $\beta$; (c) synergistic effect of $d_{(B-X)}$ and $\beta$

式受到键合强度的显著调制[35,39].

综上所述，本文基于软监督方法构建的$F_b$预测模型不仅在单特征效应层面提取出与物理认知一致的响应趋势，更在特征间协同作用维度识别出符合材料机制的协同作用模式，证实所筛选的四维核心特征足以有效描述钙钛矿材料$F_b$值变化的关键调控规律.

## 2.2 显式关联表达式构建

为构建可脱离软监督模型、可直接应用于钙钛矿体系快速筛选的预测工具，本工作在已识别的四维关键特征子集基础上，引入不确定性指导的符号回归建模策略，将特征-性能映射关系压缩为可显式计算的数学表达式. 符号回归算法利用PySR[40]软件包实现，其能够在无预设函数形式条件下自动生成结构简洁且具有物理可解释性的解析表达式，适用于本研究中小样本条件下的建模需求.

鉴于PySR框架尚不支持样本级权重直接输入，为保持软标签可信度信息在表达式构建阶段的传递，本文采用一种等效的不确定性引导的样本复制加权策略. 依据软标签的不确定性$\sigma$, 对样本权重$1/(\sigma^2+\varepsilon)$进行归一化. 在此基础上, 依据归一化权重大小按比例复制高可信度样本，从而构建一个反映标签可信度分布特征的加权扩充训练数据集，使符号回归在结构搜索和参数优化过程中优先拟合高可信度样本所约束的稳定特征-性能关系.

在符号回归模型构建与参数选择阶段，数据按约8:2划分为训练集与验证集，根据模型在训练集构建的显示表达式在测试集上的拟合误差，对符号回归过程的主要参数进行超参数寻优. 详细符号算符集合、复杂度约束及搜索参数设置见补充文件附录3. 随后利用上述确定的最优核心参数，在全部203个钙钛矿样本上开展符号回归建模. 构建的四维重要特征对$F_b$值影响规律的显式关联表达式如下所示:

$$F_b = \left(0.94 \times E_{g(PBE)} - 8.10 \times d_{(B-X)} + 19.98\right)^3 + \left(5.85 \times R_{a(B)} - 0.2 \times \beta + 9.80\right)^3, \quad (2)$$

该表达式在数据集中203个钙钛矿样本的拟合效果如图6所示. 其中，图6(a~d)曲线均为固定其他值为均值后，根据式(2)拟合所得的单特征对$F_b$值的平均影响规律. 结果表明, 该显式关联表达式能较好地拟合数据的整体分布趋势. 图6(e)为关联表达式在训练集上的整体拟合表现，由公式计算所得$F_b$值与数据集样本$F_b$值之间呈现出良好的一致性($R^2$达0.8950). 结果表明本工作提取的该显式关联表达式能够有效反映钙钛矿重要特征对$F_b$值的影响规律，具备作为钙钛矿材料本征击穿性能筛选工具的精度基础.

为进一步验证所构建显式关联表达式的准确性，本工作将其应用于文献报道的6种具有量子力学计算$F_b$标签的钙钛矿样本[27]，并与已有文献中报道的经验公式预测结果进行了对比[27,41]. 如表2及图7所示, 尽管本文所建立的表达式计算成本显著降低，但其在钙钛矿样本上与真实$F_b$值间的整体预测误差(MSE)较已有文献经验公式降低了1~2个数量级. 尤其对于在软标签构建阶段被识别为高不确定性的难预测样本，已有经验公式存在较大偏差. 而本工作的关联表达式(式(2))表现出最为显著的误差收敛效果，其预测偏差较文献报道公式降低60%~80%. 上述结果表明，该显式关联表达式不仅能够以极低的计算成本实现钙钛矿$F_b$的快速估算，且其在具有高不确定性的复杂预测场景中仍能保持稳健的推断性能，可被直接应用于高理论击穿

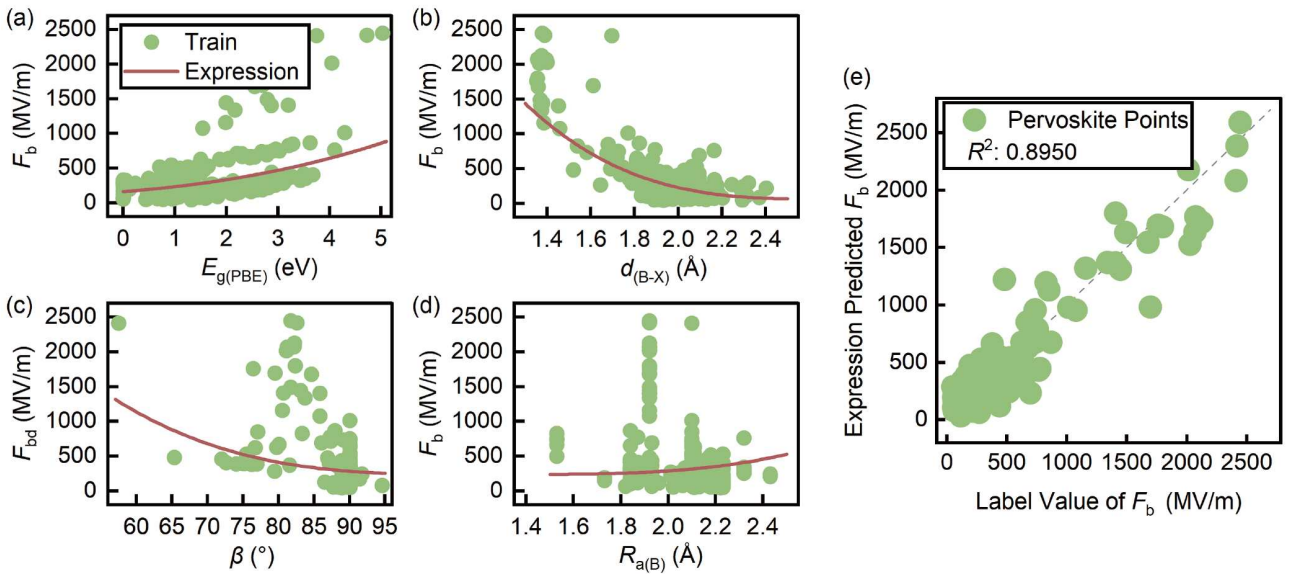



**图 6** 符号回归表达式的拟合效果. (a) $d_{(B-X)}$对$F_b$的拟合曲线; (b) $E_{g(PBE)}$对$F_b$的拟合曲线; (c) $\beta$对$F_b$的拟合曲线; (d) $R_{a(B)}$对$F_b$的拟合曲线; (e) 公式计算$F_b$值与数据集样本$F_b$值之间的数据关系

**Figure 6** Fitting performance of the symbolic regression expression. (a) Fitting curve of $d_{(B-X)}$ vs $F_b$; (b) fitting curve of $E_{g(PBE)}$ vs $F_b$; (c) fitting curve of $\beta$ vs $F_b$; (d) fitting curve of $R_{a(B)}$ vs $F_b$; (e) the data relationship between formula calculated $F_b$ value and sample $F_b$ value

**表 2** 各公式在真实数据集上的预测精度对比

**Table 2** Comparison of prediction accuracy of different empirical formulas on real datasets

| 经验公式 | $R^2$ | MSE |
|---|---|---|
| 本文构建的公式(式(2)) | 0.5883 | $7.82\times10^5$ |
| $F_b=\sqrt{\omega_{max}E_{g(HSE06)}}$ | 0.3577 | $6.78\times10^6$ |
| $F_b=1360\times(0.25\times E_{g(HSE06)})^3$ | 0.3972 | $1.29\times10^7$ |

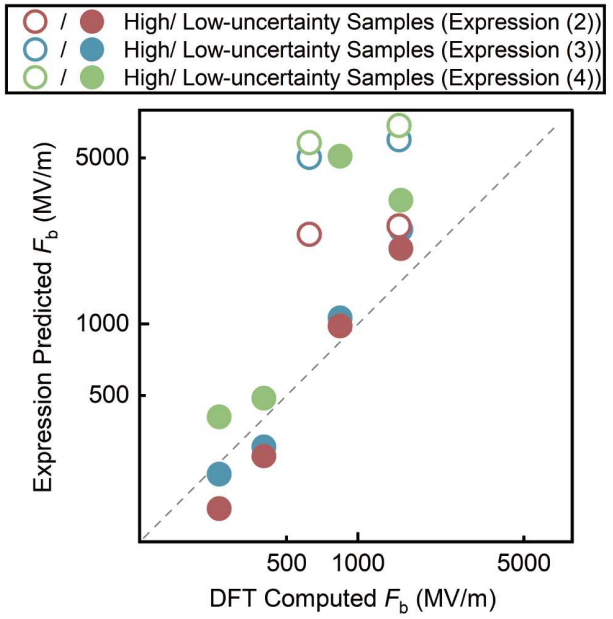



**图 7** 各公式在真实数据集上的拟合效果

**Figure 7** Fitting performance of each formula on real data set

极限钙钛矿体系的快速识别与筛选.

## 2.3 高$F_b$体系快速筛选与元素分布规律统计

本节将得到的显式关联表达式(式(2))独立应用于现有钙钛矿材料数据的快速评估与筛选, 对Materials Project[42](MP, https://materialsproject.org)数据库内4113个$ABX_3$型钙钛矿样本进行高通量计算与筛选后, 共识别出102个$F_b$值超2000 MV/cm的高电场稳定候选材料. 进一步对候选材料A/B位元素组成进行统计分析发现, 高$F_b$值候选材料的B位元素明显集中于低周期元素, 且随着B位元素周期数的增加, 其对应的平均$F_b$值呈现下降趋势. 结果表明较小的有效离子半径通常伴随更强的局域成键作用, 有利于提升钙钛矿材料在高电场下的介电稳定性, 与本文前述特征调控机制解析结果相互印证; 进一步说明构建稳定且结构紧凑的八面体框架, 是获得高电场稳定钙钛矿材料的重要结构基础. 相比之下, A位元素变化对$F_b$的直接影响较弱, 更多体现为对晶格稳定性和结构容忍度的调节作用. 综上所述, 高$F_b$钙钛矿材料的形成通常遵循低畸变、强键合的结

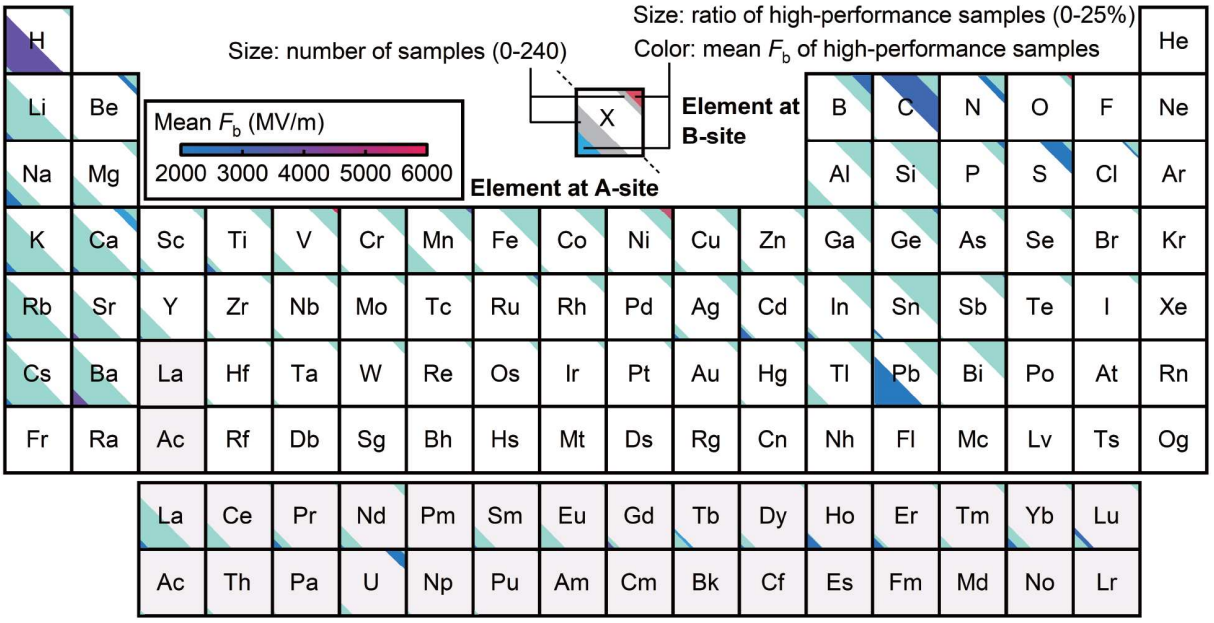



**图 8** A/B位元素分布及高性能样本统计
**Figure 8** Distribution of A/B-site elements and statistics of high-performance samples

构特征. 其核心设计规律在于优先选取具有较强局域成键能力的B位元素, 以构建稳定且致密的$BX_6$八面体骨架, 并在此基础上通过合理调控晶格倾斜度与电子结构, 实现对材料高电场稳定性的协同优化.

## 3 总结

本研究针对钙钛矿材料$F_b$高质量标签数据匮乏、关键描述符计算成本高昂的核心问题, 构建了物理可信软标签驱动的$F_b$预测建模框架, 有效缓解了传统高精度建模方法在稀缺标签条件下面临的适用性限制. 基于高质量$F_b$标签的二元绝缘体数据集, 通过GPR算法构建了$F_b$的高精度概率化预测模型, 将其迁移至钙钛矿材料体系以生成含有不确定性度量的软标签后, 构建了基于软标签监督可信度加权的钙钛矿$F_b$准确预测模型, 实现了在软标签不确定性差异显著条件下的稳健建模. 结合通过单特征主效应分析与特征间协同作用解析等方法, 对模型内部决策逻辑进行了可视化分析, 并确保了预测结果的物理可信性. 在此基础上, 本文采用基于可信度加权的符号回归方法, 提取了重要特征关于$F_b$值的显式表达式, 该表达式较已有文献中报道的经验公式在带有真实$F_b$标签钙钛矿样本上的预测误差减少了1~2个数量级. 结果表明, 本文的软标签构建策略与不确定性指导的可解释建模方法, 不仅能够显著缓解因高质量标签匮乏所导致的建模局限, 更以钙钛矿$F_b$研究为例, 为广泛材料本征物性的可靠预测提供了一种可迁移的建模范式.

## 参考文献


1 Kim J, Saremi S, Acharya M, et al. Ultrahigh capacitive energy density in ion-bombarded relaxor ferroelectric films. Science, 2020, 369: 81–84
2 Zeng X, Lin J, Dong G, et al. Polymorphic relaxor phase and defect dipole polarization co-reinforced capacitor energy storage in temperature-monitorable high-entropy ferroelectrics. Nat Commun, 2025, 16: 1870
3 Li J, Peng Y, Zhao L, et al. Machine-learning-assisted discovery of perovskite materials with high dielectric breakdown strength. Mater Adv, 2022, 3: 8639–8646
4 Zheng W, Liang T, Liu Y, et al. Enhancing energy storage performance of dielectric capacitors: tailoring CaO-SrO-$Na_2O$-$Nb_2O_5$-$SiO_2$ glass-ceramics through controlled crystallization for pulse power applications. J Eur Ceramic Soc, 2025, 45: 116949
5 Yue D, Yin J, Zhang W, et al. Computational simulation for breakdown and energy storage performances with optimization in polymer dielectrics.

Adv Funct Mater, 2023, 33: 2300658
6 Cao W, Wu Y, Yang X, et al. Breaking polarization-breakdown strength paradox for ultrahigh energy storage density in NBT-based ceramics. Nat Commun, 2025, 16: 6228
7 Liu Y, Tian G, Sun Y, et al. A double-gradient all-organic dielectric polymer film achieving superior breakdown strength and energy density. Small, 2025, 21: 2411304
8 Liu S, Xiang H, Gong X G, et al. Intrinsic breakdown strength: theoretical derivation and first-principles calculations. Phys Rev Lett, 2025, 134: 086301
9 Song D, Jeong M, Kim J, et al. High-*k* perovskite gate oxide for modulation beyond $10^{14}$ $cm^{-2}$. Sci Adv, 2022, 8: eabm3962
10 Fan Y, Qu W, Qiu H, et al. High entropy modulated quantum paraelectric perovskite for capacitive energy storage. Nat Commun, 2025, 16: 3818
11 Zhang Z, Li M, Li R, et al. Suppressing ion migration by synergistic engineering of anion and cation toward high-performance inverted perovskite solar cells and modules. Adv Mater, 2024, 36: 2313860
12 Yuan Y, Yan G, Dreessen C, et al. Shallow defects and variable photoluminescence decay times up to 280 μs in triple-cation perovskites. Nat Mater, 2024, 23: 391–397
13 Liu Z, Lin R, Wei M, et al. All-perovskite tandem solar cells achieving >29% efficiency with improved (100) orientation in wide-bandgap perovskites. Nat Mater, 2025, 24: 252–259
14 Tian G, Li Z, Zhang C, et al. Upgrading $CO_2$ to sustainable aromatics via perovskite-mediated tandem catalysis. Nat Commun, 2024, 15: 3037
15 Li Z, Mao X, Feng D, et al. Prediction of perovskite oxygen vacancies for oxygen electrocatalysis at different temperatures. Nat Commun, 2024, 15: 9318
16 Andrei V, Roh I, Lin J A, et al. Perovskite-driven solar $C_2$ hydrocarbon synthesis from $CO_2$. Nat Catal, 2025, 8: 137–146
17 Ye J, Mondal N, Carwithen B P, et al. Extending the defect tolerance of halide perovskite nanocrystals to hot carrier cooling dynamics. Nat Commun, 2024, 15: 8120
18 Zhang M, Zhu M, Chang Z, et al. Achieving excellent energy storage performance of $K_{1/2}Bi_{1/2}TiO_3$-based ceramics via multi-phase boundary and bandgap engineering. Chem Eng J, 2023, 473: 145314
19 Wang Z, Hao H, Jiang S, et al. Anti-reduction $BaTiO_3$-based dielectric ceramics for multi-layer ceramic capacitors (in Chinese). J Chin Ceram Soc, 2025, 53: 808−815 [王震, 郝华, 蒋松松, 等. 用于多层陶瓷电容器的抗还原$BaTiO_3$基介质陶瓷. 硅酸盐学报, 2025, 53: 808−815]
20 Song S, Jiao Y, Chen F, et al. Ultrahigh electric breakdown strength, excellent dielectric energy storage density, and improved electrocaloric effect in Pb-free $(1–x)Ba(Zr_{0.15}Ti_{0.85})O_{3–x}NaNbO_3$ ceramics. Ceramics Int, 2022, 48: 10789–10802
21 Wang X, Hao H, Li D, et al. Optimization of electric energy storage performance of $Sr_{0.7}Bi_{0.2}TiO_3$ modified $Bi_{0.5}Na_{0.5}TiO_3$-based films (in Chinese). J Chin Ceram Soc, 2024, 52: 1374−1383 [王鑫, 郝华, 李东旭, 等. $Sr_{0.7}Bi_{0.2}TiO_3$改性$Bi_{0.5}Na_{0.5}TiO_3$基薄膜电储能特性的优化. 硅酸盐学报, 2024, 52: 1374−1383]
22 Qiao W, Mei J, Bai M, et al. Microstructural and dielectric characteristics of A-site high-entropy oxide ceramics with a perovskite structure. Ceramics Int, 2024, 50: 51812–51819
23 Xu N, Qi X, Shen Z, et al. Point defects in metal halide perovskites. Nat Rev Phys, 2025, 7: 554–564
24 Ma X, Chen H, He R, et al. Active learning of effective Hamiltonian for super-large-scale atomic structures. npj Comput Mater, 2025, 11: 70
25 Xu P, Ji X, Li M, et al. Small data machine learning in materials science. npj Comput Mater, 2023, 9: 42
26 Liu Y, Yang Z, Zou X, et al. Data quantity governance for machine learning in materials science. Natl Sci Rev, 2023, 10: nwad125
27 Kim C, Pilania G, Ramprasad R. Machine learning assisted predictions of intrinsic dielectric breakdown strength of $ABX_3$ perovskites. J Phys Chem C, 2016, 120: 14575–14580
28 Wang W, Li S. Research status and development of insulation breakdown in engineering solid dielectrics (in Chinese). Chin Sci Bull, 2020, 65: 3461–3474 [王威望, 李盛涛. 工程固体电介质绝缘击穿研究现状及发展趋势. 科学通报, 2020, 65: 3461–3474]
29 Kim C, Pilania G, Ramprasad R. From organized high-throughput data to phenomenological theory using machine learning: the example of dielectric breakdown. Chem Mater, 2016, 28: 1304–1311
30 Ong S P, Richards W D, Jain A, et al. Python Materials Genomics (pymatgen): a robust, open-source python library for materials analysis. Comput Mater Sci, 2013, 68: 314–319
31 Xu B, Cooper V R, Singh D J, et al. Relationship between bond stiffness and electrical energy storage capacity in oxides: density functional calculations for *h*-$La_2O_3$, MgO, and BeO. Phys Rev B, 2011, 83: 064115
32 Wang J, Shen Z H, Li W, et al. Dynamic atomic-scale electron avalanche breakdown in solid dielectrics. Nat Commun, 2025, 16: 6465
33 Tan T L, Gan C L, Du A Y, et al. Delamination-induced dielectric breakdown in Cu/low-*k* interconnects. J Mater Res, 2011, 23: 1802–1808
34 Luo C, Chen Y, Wu X, et al. Dielectric molecular-bridges enable durable inverted perovskite solar cells with 26.60% efficiency and a high reverse breakdown voltage. Energy Environ Sci, 2026, 19: 338–348
35 Jin T, Niu G. Research progress on perovskite structural distortion and performance regulation (in Chinese). J Synth Cryst, 2025, 54: 1160−1174

[金童, 牛广达. 钙钛矿结构畸变与性能调控研究进展. 人工晶体学报, 2025, 54: 1160−1174]
36 Huang F T, Gao B, Kim J W, et al. Topological defects at octahedral tilting plethora in bi-layered perovskites. npj Quant Mater, 2016, 1: 16017
37 Ji Y, Lin P, Ren X, et al. Geometric and electronic structures of $Cs_2BB'X_6$ double perovskites: the importance of exact exchange. Phys Rev Res, 2024, 6: 033172
38 Mohanty S, Mukherjee S. Effect of Jahn-Teller distortion on microstructural and dielectric properties of La based double perovskites. J Alloys Compd, 2021, 892: 162204
39 Bai Y, Meng Y, Yang M, et al. Lattice stabilization and strain homogenization in Sn-Pb bottom subcells enable stable all-perovskite tandems solar cells. Nat Commun, 2025, 16: 7344
40 Tonda A. Review of PySR: high-performance symbolic regression in Python and Julia. Genet Program Evolvable Mach, 2025, 26: 7
41 Slobodyan O, Flicker J, Dickerson J, et al. Analysis of the dependence of critical electric field on semiconductor bandgap. J Mater Res, 2022, 37: 849–865
42 Jain A, Ong S P, Hautier G, et al. The Materials Project: a materials genome approach to accelerating materials innovation. APL Mater, 2013, 1: 011002

## 补充材料

**附录1** 软标签生成: GPR模型训练与预测细节
**附录2** XGBoost软监督建模细节
**附录3** 基于PySR软件的符号回归建模细节

本文以上补充材料见网络版csb.scichina.com. 补充材料为作者提供的原始数据, 作者对其学术质量和内容负责.

Summary for “物理可信软标签驱动的钙钛矿本征击穿强度预测与可解释建模”

# Physics-informed soft label-driven prediction and interpretable modeling of intrinsic dielectric breakdown strength for perovskites

Changjiao Li[1], Ao Peng[1], Lin Wang[1*], Wenqiang Gao[2], Hanxing Liu[3] & Weijia Zhou[2*]

[1] *School of Information Science and Engineering, Shandong Key Laboratory of Ubiquitous Intelligent Computing, University of Jinan, Jinan 250022, China*
[2] *Institute for Advanced Interdisciplinary Research, University of Jinan, Jinan 250022, China*
[3] *School of Materials Science and Engineering, State Key Laboratory of Advanced Technology for Materials Synthesis and Processing, Wuhan University of Technology, Wuhan 430070, China*
* Corresponding authors, E-mail: ise_wanglin@ujn.edu.cn; ifc_zhouwj@ujn.edu.cn

Perovskite materials, characterized by their exceptional structural stability and compositional tunability, have emerged as promising candidates for high dielectric breakdown strength ($F_b$) applications. However, accurately predicting $F_b$ remains challenging due to the scarcity of experimental data and the high computational cost associated with traditional quantum mechanical calculations. This study introduces a novel physics-informed framework that leverages soft labels for the efficient prediction of $F_b$ in perovskite materials.

The proposed approach begins with the construction of a high-precision probability model for $F_b$ prediction based on a high-quality insulator dataset. This model is then transferred to the perovskite materials system, generating soft labels that include both predicted values and associated uncertainty estimates. These soft labels, which quantify prediction uncertainty, are critical for overcoming the data scarcity problem and allow for more reliable predictions in the absence of abundant experimental labels. Using these soft labels, we then develop a supervised learning model that incorporates confidence-weighted predictions. The model employs XGBoost, a gradient-boosting machine learning algorithm, which is tailored with a custom loss function that prioritizes samples with higher confidence, resulting in improved model robustness and accuracy. The model achieves an impressive $R^2$ value of 0.8952 on a validation set, demonstrating its effectiveness in making reliable predictions with limited data.

Furthermore, we employ a symbolic regression approach to extract explicit mathematical expressions for the key features influencing $F_b$. This method, implemented through the PySR software package, generates concise, interpretable equations that map the relationship between critical material descriptors and $F_b$. These expressions are not only easy to interpret but also facilitate fast predictions without the need for complex machine learning models, making them highly valuable for large-scale screening of perovskite materials. The symbolic regression model also addresses the issue of noisy or uncertain data by incorporating uncertainty guidance into the regression process, which improves the stability and accuracy of the derived expressions.

Through this work, we identify the key structural and electronic features that significantly influence the dielectric breakdown strength of perovskites, including the bond lengths, electronic structure, and lattice distortions. Our model reveals that shorter B-X bond lengths and larger band gaps are associated with higher $F_b$ values, consistent with the fundamental mechanisms of dielectric breakdown. The results from this study offer a highly scalable and transferable methodology that can be applied to a broad range of material systems, providing a reliable framework for the high-throughput screening of materials with superior dielectric properties.

Overall, the methodology presented in this paper not only addresses the challenges of data scarcity but also offers a robust, interpretable, and computationally efficient model for the prediction of intrinsic material properties across different material systems.

**perovskite materials, intrinsic breakdown strength, soft-supervised modeling, physical interpretability**

doi: 10.1360/CSB-2025-5998