

基于机器学习和符号回归筛选具有优异光电性能的卤化物双钙钛矿材料*

刘研博 刘佳冬 熊青 张梧桐 张照胜†

(河北大学, 化学与材料科学学院, 保定 071002)

摘要

为了满足高效率光电器件日益增长的需求, 加速筛选具备卓越光电性能的卤化物双钙钛矿 $A_2B'B''X_6$ (halide double perovskites, HDPs) 势在必行。本研究基于包含 3624 组 HDPs 的数据集, 系统比较了多种机器学习与符号回归算法在形成能 (E_f) 和带隙 (E_g) 预测中的性能。所有模型均在 MATLAB 平台实现, 包括极限梯度提升 (MXGBoost)、轻量梯度提升 (MLightGBM)、随机森林 (random forest)、残差门控神经网络 (Residual Gated Neural Network, RGNN) 以及符号回归方法 GPTIPS 和 SISSO (mGPTIPS 和 mSISSO)。结果表明, MXGBoost 模型在 E_f 和 E_g 预测中表现最佳, 其测试集决定系数 (R^2) 分别达到 0.984 和 0.816。进一步结合符号回归分析, 揭示了 X 位阴离子电负性 (χ_X) 以及 B 位阳离子电负性 ($\chi_{B''}$ 、 $\chi_{B'}$) 是描述 E_f 和 E_g 与晶体结构关系的关键特征。基于上述模型筛选, 获得了两种热力学稳定的直接带隙半导体材料 Na_2HgZnF_6 和 Tl_2PbSnF_6 。第一性原理计算表明, 相较于 Na_2HgZnF_6 , Tl_2PbSnF_6 具有更强的电荷局域化特征和更低的深能级缺陷浓度, 从而有利于抑制非辐射复合过程。此外, 其光谱极限最大效率 (SLME) 达到 28.01%, 表现出优异的光电性能。该研究构建了兼具高效性与可解释性的材料筛选框架, 为新型卤化物双钙钛矿材料的设计与性能预测提供了重要理论依据。

关键词: 钙钛矿, 机器学习, 符号回归, 电子结构, 本征缺陷

PACS: 07.05.Mh, 02.60.Gf, 71.15 Mb, 71.55.-i

基金: 国家自然科学基金 (No. 22573026)、河北省自然科学基金 (No. B2024201051)、河北省教育厅 (No. BJK2024094) 和河北大学 (No. DC2026275) 资助的课题。

† 通信作者.E-mail: misaraty@163.com

第一作者.E-mail: 16632001711@163.com

1 引言

卤化物双钙钛矿 (halide double perovskite, HDPs) 因其结构稳定性^[1]和优异的

光电性能^[2], 在太阳能电池^[3,4]、热电领域^[1]、发光二极管(LED)^[5]和光电探测器^[6]等多种领域具有良好的应用潜力。自 2016 年 McClure 等^[7]首次成功合成 $\text{Cs}_2\text{AgBiBr}_6$ 和 $\text{Cs}_2\text{AgBiCl}_6$ 以来, 其凭借高缺陷容忍性和可调带隙等特点逐渐成为热门研究材料。HDPs 结构通式为 $\text{A}_2\text{B}'\text{B}''\text{X}_6$, 目前报道最多且性能稳定的构型是空间群为 $Fm\bar{3}m$ 的面心立方结构^[7,8](如图 1 所示)。一般情况下, B' 和 B'' 位阳离子是半径较小的主族金属离子、过渡金属离子或稀土离子, 可由一价与三价金属阳离子的组合或者两种二价金属阳离子组合构成。它们分别与卤素离子 X^- 形成 $\text{B}'\text{X}_6$ 和 $\text{B}''\text{X}_6$ 八面体, 并占据八面体中心。A 位阳离子半径较大, 填充在八面体间隙中, 对晶格形成和结构稳定性具有重要影响^[9]。同时, B' 和 B'' 位阳离子的电子构型和能级分布决定了材料的禁带宽度和载流子性质^[10,11], 而 X 位阴离子的变化也会影响材料的能带结构和光学性质^[12]。为了进一步挖掘 HDPs 的性能, 探索不同的离子组合亟待进行, 但可选的元素范围和组合种类十分庞大, 采用传统的实验合成或第一性原理计算筛选高性能材料因其较高的成本成为性能挖掘的阻碍。

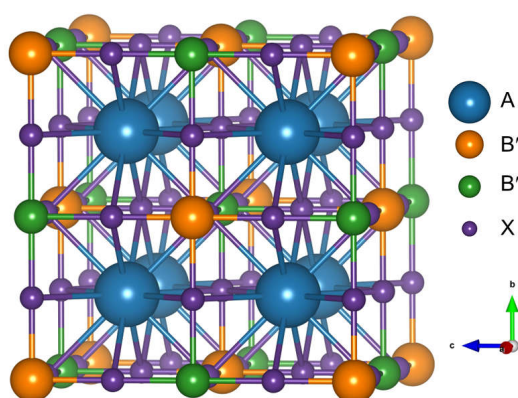


图 1 $\text{A}_2\text{B}'\text{B}''\text{X}_6$ 的晶体结构

Fig.1. Crystal structure of $\text{A}_2\text{B}'\text{B}''\text{X}_6$

为了加速新型 HDPs 材料的高效发现, 研究者们提出将材料的结构与组成特征转化为可量化的描述子, 并利用机器学习算法, 对大量候选材料进行快速性能预测和筛选^[8,13-16]。2017 至 2019 年, 密度泛函理论(DFT)^[17]高通量计算相关的

HDPs 研究一定程度上为机器学习提供了特征数据支持^[8,13]。2022 年, Kim 等^[14]基于外部材料数据库构建机器学习模型, 对 13542 个全无机 HDPs 候选材料进行形成能(E_f)与带隙(E_g)预测, 初始决定系数(R^2)分别仅为 0.77 和 0.427, 并通过主动学习策略逐步提升模型精度。2025 年, Luo 等^[15]基于 680 组实测数据构建梯度提升树模型, 同步实现 E_f 回归、 E_g 回归和 E_g 分类, 预测精度显著提升, 并成功实验合成 $\text{Cs}_2\text{RbBiI}_6$ 应用于光探测器和光催化领域, 验证了机器学习在高通量筛选和实验指导中的实用性。然而该研究数据集规模相对有限, 其模型的泛化能力仍有待提升。近期, Wang 等^[16]构建包含 1181 个 HDPs 的数据库, 系统对比了随机森林(random forest)、梯度提升回归、极限梯度提升(eXtreme Gradient Boosting, XGBoost)及深度神经网络(DNN)等多种模型, 并采用 DNN 实现对 E_g 的精准预测, 同时结合 DFT 计算与器件模拟筛选出多类高性能 HDPs 材料, 为光伏应用提供了重要参考。该研究主要聚焦于 E_g 预测, 未对 E_f 这一关键稳定性指标尚未进行系统建模。

综上, 现有研究在数据集规模、预测指标及模型系统性对比方面仍存在进一步完善的空间。基于此, 本研究引入多种经典机器学习与深度学习算法, 对 E_f 与 E_g 开展高精度预测。XGBoost 基于 Boosting 方法, 通过迭代拟合残差可精确描述复杂的非线性关系, 已广泛应用于钙钛矿太阳能电池筛选^[18]、电极材料设计^[19]、二维材料性质预测^[20]及合金材料筛选^[21]等领域。类似地, 轻量梯度提升模型(Light Gradient Boosting Machine, LightGBM)^[22]凭借高效训练和直方图优化等特点, 在半导体材料^[23]、催化剂^[24]和有机骨架材料^[25]的结构筛选和性质预测中表现出优异性能, 并拓展至原子核质量表探究^[26]等跨学科任务。在此基础上, 本研究将 XGBoost 与 LightGBM 算法应用于更大规模的全无机 HDPs 材料数据集, 并与 random forest 及残差门控神经网络(Residual Gated Neural Network, RGNN)进行系

统对比,旨在构建材料结构描述符与 E_f 、 E_g 之间的高精度预测模型,同时结合符号回归方法揭示模型预测的内在机制,并通过第一性原理计算筛选具有高热力学稳定性和优异光电性能的候选材料,为 HDPs 的高效开发与精准设计提供可靠的数据驱动方案。

2 计算方法

2.1 算法模型

本研究采用Kim等^[14]公开的13542个全无机卤化物双钙钛矿(HDPs)数据集作为研究对象。该数据集均为 $A_2B'B''X_6$ 型结构,由高通量元素组合生成,其中A位包含约30种一价金属阳离子,B'/B''位包含约55种金属阳离子,X位为四种卤素阴离子(F⁻、Cl⁻、Br⁻和I⁻)。所有候选结构均经过容忍因子判据与价态平衡规则初步筛选。经数据预处理后,最终得到包含3624组有效数据的规整数据库,数据预处理流程及数据集组成元素分布如图S1和图S2所示。离子半径与电负性作为描述结构与电子性质关系的常用物理描述符^[23]。本研究选取各位点的离子半径(r_A 、 $r_{B'}$ 、 $r_{B''}$ 、 r_X)及电负性(χ_A 、 $\chi_{B'}$ 、 $\chi_{B''}$ 、 χ_X)作为描述子输入,通过机器学习与符号回归算法开展建模与机理分析,建立 E_f 、 E_g 与结构之间的定量关联,流程如图2所示。为了将算法融合为统一的计算框架,便于跨平台的模拟计算,所有算法均使用MATLAB语言实现。本文实现的机器学习算法包括梯度提升树(MXGBoost和MLightGBM)、random forest和RGNN,它们与 E_f 和 E_g 构建了高精度的预测框架。平均绝对误差(MAE)、均方根误差(RMSE)和 R^2 三项指标被用来综合评估模型性能。为保证结果的可重复性,所有模型训练均采用固定随机种子。考虑到数据集规模有限,采用五折交叉验证评估模型性能。数据集被划分为数量相近的五个子集,每次选取一个子集作为测试集,其余作为训练集,并在训练集内部进行交叉验证以完成模型训练与超参数优化。重复五次后取测试结果平均值作为最终性能

指标，且各折测试数据均未参与训练与参数优化过程。

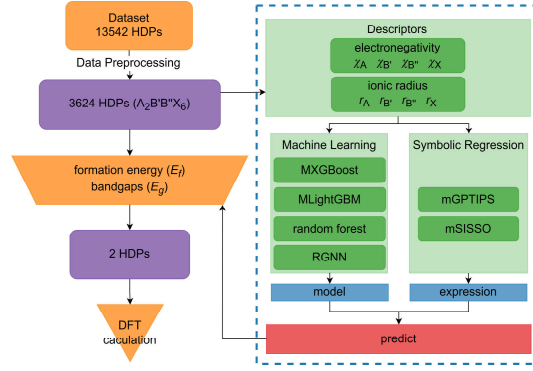


图2 基于机器学习与符号回归算法筛选3624 HDPs的流程图

Fig.2. Flowchart of screening 3624 HDPs using machine learning and symbolic regression algorithms

2.2 第一性原理计算

第一性原理计算均在维也纳从头算模拟软件包(Vienna Ab initio simulation package, VASP)^[27]的DFT框架下实现，并使用投影缀加波方法(projector augmented wave, PAW)^[28]描述电子与离子核之间的相互作用。电子交换关联效应采用广义梯度近似(GGA)^[29]的Perdew-Burke-Ernzerhof(PBE)泛函^[30]处理。对于电子结构计算，我们采用以 Γ 点为中心的 $4 \times 4 \times 4$ Monkhorst-Pack型 k 点网格进行了初始结构弛豫，并选取更密集的 $6 \times 6 \times 6$ Monkhorst-Pack型 k 点网格进行态密度计算，最后通过在高对称 k 点路径中使用五点插值采样法计算了能带结构。另外，能带结构的电子交换关联效应还选用包含了25%的Hartree-Fock交换能混合泛函(Heyd-Scuseria-Ernzerhof, HSE06)^[31]处理。为确保能量与波函数收敛，所有计算的平面波截断能设定为500 eV，离子弛豫力需收敛至0.01 eV/Å内，能量收敛精度要求为 1×10^{-5} eV。

为确保结构稳定不产生AX、B'X₂或B''X₂等分解相，相关化学势约束条件如式S1–S4所示。缺陷 D 在电荷态 q 下的 E_f 定义^[32]如式1所示。

$$\Delta H_{D,q}(E_F, \mu) = [E_{D,q} - E_H] + \sum_i n_i \mu_i + qE_F + E_{corr}, \quad (1)$$

其中， $\Delta H_{D,q}(E_F, \mu)$ 表示缺陷形成所需要的能量， $E_{D,q}$ 和 E_H 分别表示含有缺陷结构

的总能量和完美晶胞的总能量， μ_i 表示从完美晶胞中移除($n_i>0$)或添加($n_i<0$)原子 i 形成缺陷时体系的化学势， E_F 表示费米能级， E_{corr} 表示带电计算中的有限尺寸相关修正项。本研究对点电荷相互作用误差进行校正采用的是Kumagai-Oba (KO) 方案^[33]。

3 结果与讨论

3.1 模型分析

3.1.1 机器学习

图 3(a)–3(h)展示的是机器学习算法对 3624 种 HDPs 的 E_f 和 E_g 的回归预测奇偶图，图 S3–S5 展示了四种算法在训练集、验证集及测试集上的奇偶图。结合表 1 和表 S1、S2 性能分析可知，基于梯度提升策略的 MXGBoost 和 MLightGBM 在两类性质预测中均表现出最优性能，尤其是在 E_f 预测任务中，两者的 R^2 均达到 0.984，且 MAE 和 RMSE 均维持在较低水平，表明其对结构稳定性相关特征具有较强的拟合能力。在 E_g 预测中，MXGBoost 的性能略优于 MLightGBM，显示出其在处理更复杂非线性关系时具有一定优势。相比之下，random forest 在 E_f 和 E_g 测试集上的 R^2 分别为 0.971 和 0.802，略逊于梯度提升树模型。这主要源于其基于 Bagging 的独立建树机制缺乏残差迭代优化过程，对复杂特征与响应关系的表达能力相对有限。而梯度提升树通过逐步拟合残差，可更有效捕捉高阶非线性特征，从而获得更高预测精度。RGNN 在 E_f 预测中取得了与 random forest 相当的性能($R^2=0.972$)，但在 E_g 预测任务中表现相对较差($R^2=0.723$)。这一现象表明，在当前数据规模和特征维度条件下，神经网络模型的表达能力未能得到充分发挥，反而更容易受到数据稀疏性与噪声的影响。相比之下，基于决策树的模型在处理中低维结构描述符时表现出更好的稳健性与泛化能力。从物理本质上看， E_f 主要由原子排列和键能决定，其与结构描述符之间关系相对明确，因此更易被

机器学习模型捕捉；而 E_g 不仅受到结构因素影响，还与电子结构、轨道耦合及周期性势场等多种因素密切相关，表现出更强的非线性与复杂性。这也解释了各类模型在 E_g 预测中整体性能低于 E_f 的现象。

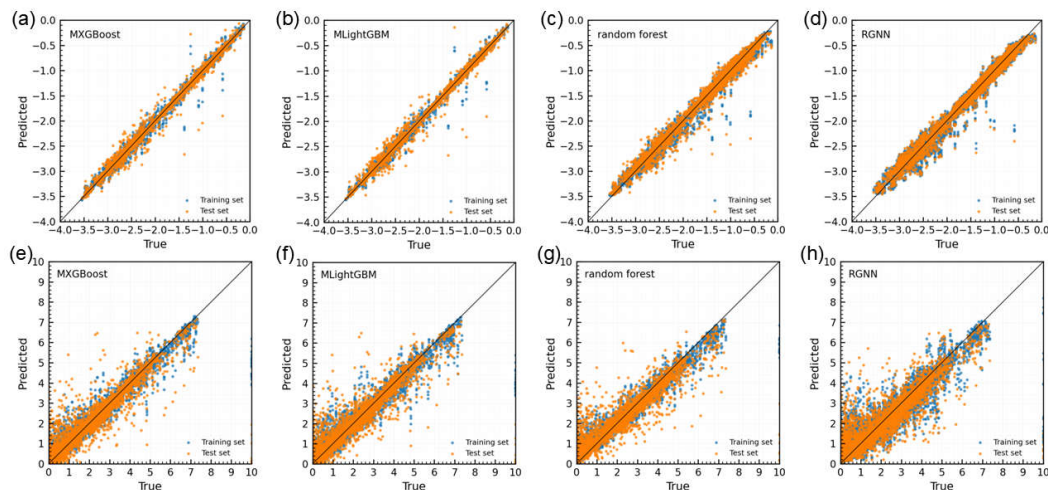


图 3 基于五折交叉验证的机器学习模型对 3624 种 HDPs 的 E_f (a-d)和 E_g (e-h)的预测奇偶图。

蓝色点表示训练数据的平均预测结果，橙色点表示测试数据的平均预测结果

Fig.3. Parity plots of predicted E_f (a-d) and E_g (e-h) for 3624 HDPs, generated by machine learning models based on five-fold cross-validation. Blue points represent the mean predicted results for the training data, while orange points represent the mean predicted results for the test data

表 1 机器学习与符号回归算法在 3624 种 HDPs 测试数据上的性能评估

Table 1. Prediction performance of machine learning and symbolic regression algorithms on the test data of 3624 HDPs

Algorithms	E_f			E_g		
	MAE / eV	RMSE / eV	R^2	MAE / eV	RMSE / eV	R^2
MXGBoost	0.038	0.094	0.984	0.318	0.679	0.816
MLightGBM	0.035	0.094	0.984	0.328	0.679	0.814
random forest	0.074	0.130	0.971	0.357	0.707	0.802
RGNN	0.077	0.127	0.972	0.498	0.839	0.723
mGPTIPS	0.083	0.111	0.980	0.728	0.978	0.634
mSISSO	0.120	0.164	0.954	0.855	1.133	0.480

3.1.2 符号回归

为进一步揭示机器学习模型预测结果的内在物理机制，本文进一步采用 mGPTIPS 和 mSISSO 两种主流的符号回归方法，对 3624 种 HDPs 的 E_f 和 E_g 进行了解析建模，从而获得明确的数学表达关系。其中，mGPTIPS 通过遗传操作在较自由的函数空间中搜索非线性表达式，适用于识别复杂的变量交互关系^[34]；而 mSISSO 则通过构建候选描述符空间，并结合独立筛选与稀疏优化，获得低维、紧凑且具有较好可解释性的解析表达式^[35]。本文借助二者不同搜索机制的互补性，对关键描述符进行相互印证。图 4(a)、图 4(b)、图 4(d)、图 4(e)和图 S6 分别展示了两种模型训练集、测试集和验证集的预测奇偶图。结合表 1 和表 S1、S2 的性能分析可得，两种符号回归模型在 E_f 数据集上均取得了较高精度(mGPTIPS 和 mSISSO 的测试集 R^2 分别为 0.980 和 0.954)。相比之下，在 E_g 预测任务中，两种模型的性能明显下降(R^2 分别为 0.634 和 0.480)，反映出带隙与结构描述符之间存在更复杂的非线性关系，难以通过低维解析表达式进行充分刻画。尽管如此，符号回归在揭示关键影响因素方面仍具有重要优势。mSISSO 提供的解析式如式 (2)和式(3)所示。由 E_f 解析式可以看出，B 位和 X 位电负性是影响结构稳定性的主要因素。 $\left[1/\chi_A - (\log\chi_{B'})^2\right]$ 的系数最大， $\chi_{B''}$ 表现出较强的非线性调控， $\chi_X^2 / \chi_{B'}^{1/4}$ 项系数较小，但 χ_X 的平方项放大了 X 位电负性的影响，其可通过增强 B-X 键强度显著降低形成能。 $|\chi_X - \chi_{B''}^2|$ 虽系数极小，但也表明了 B''位和 X 位电负性差越大，B-X 键极化作用越强，结构越稳定。此外， $(1/r_{B'} + 1/r_{B''})$ 表明较小的 B'位和 B''位半径会使结构更加紧密，是影响结构稳定性的次要因素。在 E_g 解析式中，涉及 B'位和 B''位阳离子的电负性及其组合项的系数较大，说明其在 E_g 调控中起主导作用。尽管部分半径相关项的系数较小，但其通过与电负性的交互项参与模型，对 E_g 调控具有次要贡献。

$$y_{E_f} = 0.983 - 0.227 \cdot \chi_X^2 / \chi_{B'}^{1/4} - 0.686 \cdot \left[1 / \chi_A - (\log \chi_{B''})^2 \right] + 0.125 \cdot (1 / r_{B'} + 1 / r_{B''}) - 0.090 \cdot |\chi_X - \chi_{B''}^2|, \quad (2)$$

$$y_{E_g} = 0.290 - 1.270 \cdot (\chi_{B'} - r_{B''} - \chi_X) - 0.018 \cdot f_1(r_{B'}) - 0.015 \cdot f_2(r_{B''}) + 1.845 \cdot f_3(r_{B'}, \chi_{B''}), \quad (3)$$

为进一步验证特征重要性, 本研究结合 MXGBoost 模型对 mSISSO 筛选出的特征进行了夏普利加性解释(SHAP)分析^[36], 其特征重要性排序如图 4(c)和图 4(f)所示。在 E_f 预测中, χ_X 具有最高的重要性, 其平均 SHAP 值约为 0.440。而在 E_g 预测中, $\chi_{B''}$ 、 $\chi_{B'}$ 的平均 SHAP 值分别约为 0.428 和 0.412, 表现出相近且占主导的贡献。这一结果与符号回归得到的解析关系具有良好的一致性, 说明 B 位与 X 位电负性通过调控 B-X 键的极性与强度, 进一步影响 BX_6 八面体结构及其电子结构, 从而共同决定材料的热力学稳定性与光电性质。

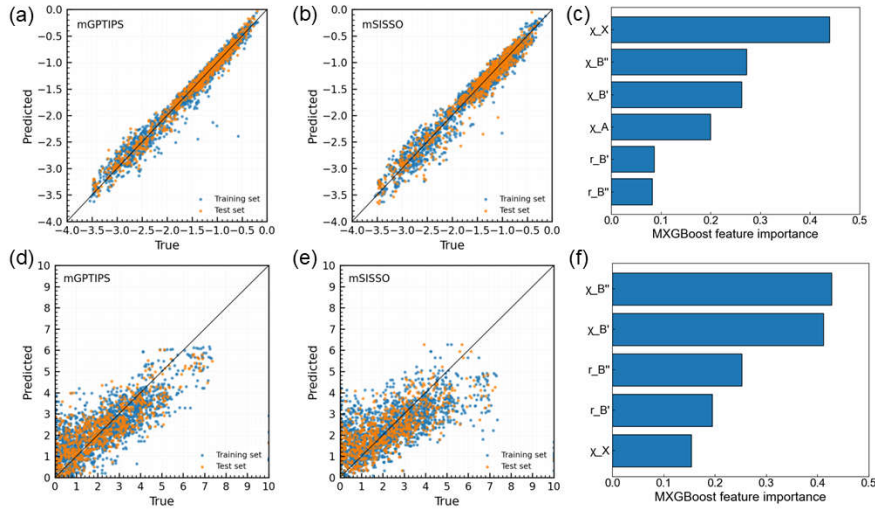


图 4 mGPTIPS(a, d)和 mSISSO(b, e)对 3624 种 HDPs 的 E_f 和 E_g 预测奇偶图, 蓝色点表示训练数据的平均预测结果, 橙色点表示测试数据的平均预测结果, 图(c, f)展示了利用 MXGBoost 模型预测的特征重要性顺序

Fig.4. Parity plots of E_f and E_g of 3624 HDPs predicted by mGPTIPS (a, d) and mSISSO (b, e). Blue points represent the predicted results for the training data, while orange points represent the predicted results for the test data. (c, f) show the feature importance rankings predicted using the MXGBoost model

3.2 计算验证

3.2.1 结构优化

上述机器学习与符号回归相结合的预测流程有效提升了材料性质筛选的效率与可解释性，为高通量发现具有优异光电性能的候选材料提供了可靠工具。在候选材料筛选过程中，我们首先基于光伏器件的 Shockley-Queisser(SQ)极限以及卤化物钙钛矿材料的实际光吸收特性，确定目标光学带隙范围约为 1.3–1.4 eV。考虑到 PBE 泛函对带隙存在系统性低估误差，为避免潜在高性能材料的遗漏，我们将筛选窗口适当放宽至 1.0–1.5 eV。在此基础上，结合机器学习预测结果，对材料按 E_f 由低到高进行初筛，以优先保留热力学稳定性较高的候选体系。随后，对满足 E_g 与稳定性筛选条件的材料开展 DFT 计算，进一步精确计算其电子结构性质，包括能带结构、带隙数值以及带隙类型，以验证机器学习预测的可靠性。最终，通过综合热力学稳定性与光学带隙要求，在筛选范围内确定了 $\text{Na}_2\text{HgZnF}_6$ 与 $\text{Tl}_2\text{PbSnF}_6$ 两种同时满足 $E_g=1.0\text{--}1.5$ eV 且具有直接带隙特征的优选 HDP 候选材料，并以总原子数为 40、晶格常数 $a=b=c=11.49$ Å、 $\alpha=\beta=\gamma=90^\circ$ 的结构作为初始弛豫模型进行结构优化。为了验证晶体结构的热稳定性，我们计算了八面体因子 μ 、Goldschmidt 容差因子 t 和新型容差因子 τ ^[37,38]，结果如表 2 所示。优化后 $\text{Tl}_2\text{PbSnF}_6$ 的晶格常数(9.49 Å)大于 $\text{Na}_2\text{HgZnF}_6$ (8.57 Å)，主要是由于 TI^+ 、 Pb^{2+} 和 Sn^{2+} 离子半径较大导致晶胞膨胀。两者晶格角度未发生偏离，保持了高对称性，同时 μ 和 t 均在钙钛矿稳定范围内($0.41<\mu<0.90$, $0.80<t<1.00$)， τ 均小于 4.18，表明两种晶体均为稳定的钙钛矿结构。相比之下， $\text{Tl}_2\text{PbSnF}_6$ 较大的 μ 、 t 和较小的 τ ，表明其晶胞匹配更均匀、更接近理想立方结构且局部几何扭曲较小，因此晶体结构比 $\text{Na}_2\text{HgZnF}_6$ 更稳定。

表 2 晶体优化后的晶格参数及稳定性指标

Table 2. Optimized lattice parameters and stability metrics of the crystals

Structures	$a/b/c$ /Å	$\alpha/\beta/\gamma$ /°	μ	t	τ
$\text{Na}_2\text{HgZnF}_6$	8.57	90	0.66	0.87	3.97

3.2.2 电子结构

图 5 展示了 Na₂HgZnF₆ 和 Tl₂PbSnF₆ 的电子结构, 包括能带结构、投影态密度(projected density of states, PDOS)及能带附近的电荷密度。如图 5(a)和图 5(e)所示, Na₂HgZnF₆ 的 PBE E_g 为 1.41 eV, Tl₂PbSnF₆ 为 1.02 eV, 且两者价带顶(valence band maximum, VBM)和导带底(conduction band minimum, CBM)均位于 Γ 点, 表明二者均为直接带隙半导体材料, 且 E_g 接近最佳光学带隙范围。HSE06 计算得到的 E_g 分别为 3.40 eV 和 1.73 eV。由图 5(b)和图 5(f)的 PDOS 数据分析可得, Na₂HgZnF₆ 的 VBM 由 F 和 Hg 原子轨道做主要贡献, CBM 由 Hg 原子轨道做主要贡献。Tl₂PbSnF₆ 的 VBM 由 F 和 Sn 原子轨道做主要贡献, CBM 由 Tl 原子轨道做主要贡献。为了进一步分析 VBM 和 CBM 电荷的集中分布特征, 我们计算了两种材料的电荷密度, 并采用逆参与比(inverse participation ratio, IPR)^[39]量化了室温下的电荷局域化程度, IPR 的计算公式如式(4)所示。

$$\text{IPR} = 1 / N \cdot \left(\sum_i k_i^2 \right)^2 / k_i^4, \quad (4)$$

其中 N 代表采样网格点数, k_i 代表每个网格点的电荷密度, IPR 取值范围为 0 到 1, 数值越小代表电荷分布越局域。图 5(c)和图 5(g)显示了 2 种材料沿 y 轴方向 VBM 电荷的平面投影, 图 5(d)和图 5(h)显示了 CBM 电荷分布情况, 蓝紫色区域代表电荷密度较低, 反之越接近红色代表电荷密度越高。结果表明, Tl₂PbSnF₆ 的 VBM 和 CBM 电荷分布均比 Na₂HgZnF₆ 更为局域, IPR 分别从 Na₂HgZnF₆ 的 0.137、0.304 降至 0.057、0.224。根据 Franck-Condon 原理^[40], 电荷更高的局域化特征会降低电子与空穴的波函数有效重叠, 抑制非辐射复合, 有利于延长载流子寿命并提升光电性能。

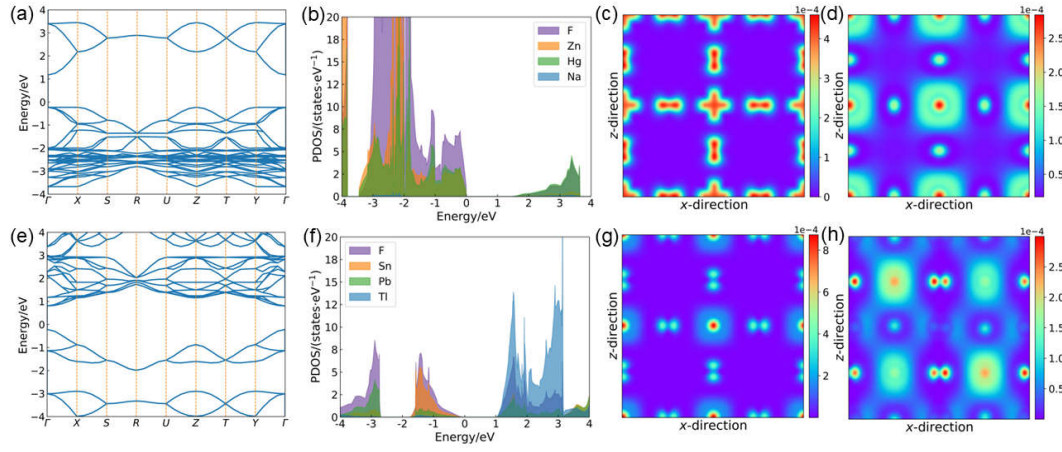


图 5 (a-d) $\text{Na}_2\text{HgZnF}_6$ 和 (e-h) $\text{Tl}_2\text{PbSnF}_6$ 的电子特征: (a, e) 能带结构; (b, f) PDOS; (c, g) VBM 的电荷密度; (d, h) CBM 的电荷密度

Fig.5. Electronic properties of (a-d) $\text{Na}_2\text{HgZnF}_6$ and (e-h) $\text{Tl}_2\text{PbSnF}_6$: (a, e) band structures; (b, f) PDOS, (c, g) charge density of the VBM, (d, h) charge density of the CBM. The Fermi level was set at 0 eV

3.2.3 本征缺陷

为探究不同阳离子组合对 $\text{Na}_2\text{HgZnF}_6$ 和 $\text{Tl}_2\text{PbSnF}_6$ 缺陷性能的影响, 本研究从空位缺陷、间隙缺陷和反位缺陷三个方面选取若干代表性本征点缺陷进行比较。在钙钛矿反位缺陷中, 已知 A 位替换 B 位形成的缺陷(A_B)具有较低的缺陷形成能, 更容易在环境热激励下自发形成^[41]。因此, 我们分别选取了 $\text{Na}_2\text{HgZnF}_6$ 和 $\text{Tl}_2\text{PbSnF}_6$ 中的八种本征点缺陷进行分析, 包括三种空位缺陷($\text{V}_{\text{Na}}/\text{V}_{\text{Tl}}$ 、 $\text{V}_{\text{Hg}}/\text{V}_{\text{Pb}}$ 和 $\text{V}_{\text{Zn}}/\text{V}_{\text{Sn}}$), 三种间隙缺陷($\text{Na}_\text{i}/\text{Tl}_\text{i}$ 、 $\text{Hg}_\text{i}/\text{Pb}_\text{i}$ 和 $\text{Zn}_\text{i}/\text{Sn}_\text{i}$)以及两种反位缺陷($\text{Na}_{\text{Hg}}/\text{Tl}_{\text{Pb}}$ 和 $\text{Na}_{\text{Zn}}/\text{Tl}_{\text{Sn}}$)。相应的缺陷形成能和跃迁能级如图 6(a)-(f)所示。在空位缺陷中, $\text{Na}_2\text{HgZnF}_6$ 的 V_{Zn} 和 $\text{Tl}_2\text{PbSnF}_6$ 的 V_{Pb} 的(-/0)、(-2/0)跃迁能级分别高于 VBM 0.28 eV 和 0.29 eV, 表现为受主型(*p* 型)深能级缺陷, 但由于 V_{Pb} 的缺陷形成能(3.42 eV)高于 V_{Zn} (2.78 eV), $\text{Tl}_2\text{PbSnF}_6$ 空位深能级缺陷浓度较低。在间隙缺陷中, Na_i 和 Hg_i 在 $\text{Na}_2\text{HgZnF}_6$ 中的缺陷形成能分别为-0.43 eV 和-0.05 eV, 其跃迁能级(0/+), (0/+2)高于 VBM 0.50 eV 和 0.27 eV, 呈现出易于形成的显著 *p* 型深能级特征, 可在材料内部捕获载流子, 形成非辐射复合中心, 进而严重影响器件性能。相比之

下, $\text{Ti}_2\text{PbSnF}_6$ 中存在的施主型(n 型)深能级缺陷 Sn_i , 其跃迁能级($0/+$)低于 CBM 0.24 eV, 但缺陷形成能较大(2.10 eV), 对材料性能影响有限。 Ti_i 和 Pb_i 则表现为 p 型浅能级缺陷, 跃迁能级($0/+$)均高于 VBM 0.01 eV。在反位缺陷类型中, $\text{Na}_2\text{HgZnF}_6$ 的 Na_{Hg} 和 Na_{Zn} 缺陷形成能分别为 0.89 eV 和 2.48 eV, 其跃迁能级($-/0$)分别位于 VBM 上方 0.49 eV 和 0.47 eV 处, Na_{Hg} 由于深能级和低缺陷形成能成为主导缺陷, 可显著加速载流子损耗。 $\text{Ti}_2\text{PbSnF}_6$ 中对应位点的 Ti_{Pb} 跃迁能级($-/0$)位于 VBM 上方 0.12 eV 处, 缺陷形成能为 1.21 eV, 表现为高浓度 p 型浅能级反位缺陷。综上, $\text{Na}_2\text{HgZnF}_6$ 更易形成深能级缺陷, 具有较高的深能级浓度, 容易加剧载流子的非辐射复合; 而 $\text{Ti}_2\text{PbSnF}_6$ 表现出更高的 p 型浅能级浓度, 可有效增加材料中的空穴浓度, 增强器件的导电性。

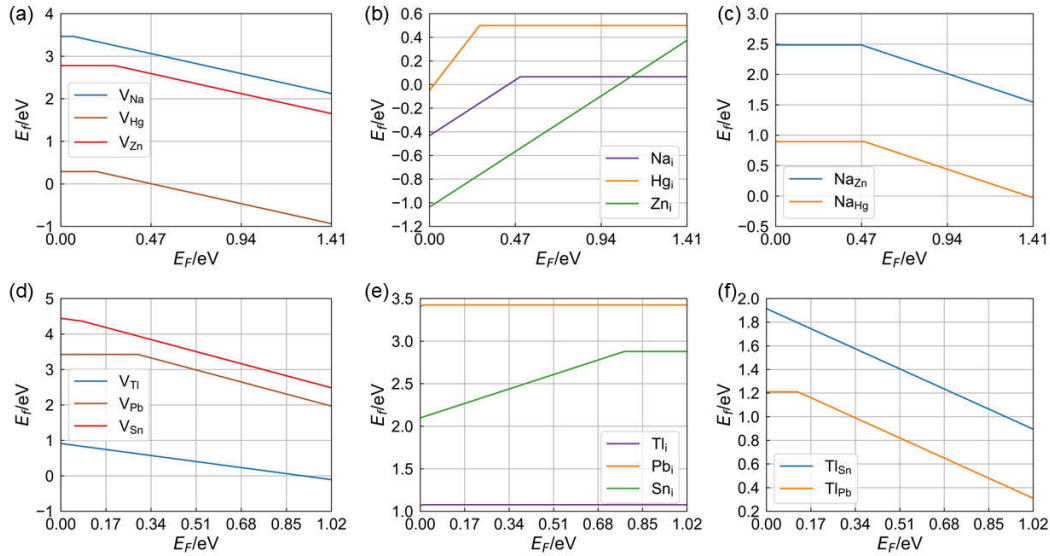


图 6 (a–c) $\text{Na}_2\text{HgZnF}_6$ 和(d–f) $\text{Ti}_2\text{PbSnF}_6$ 的缺陷形成能: (a, d)空位缺陷; (b, e)间隙缺陷; (c, f)反位缺陷

Fig.6. (a–c) $\text{Na}_2\text{HgZnF}_6$ and (d–f) $\text{Ti}_2\text{PbSnF}_6$ defect formation energies: (a, d) vacancy defects; (b, e) interstitial defects; (c, f) antisite defects

3.2.4 光学性质

基于上述缺陷特征分析, 我们进一步计算了这两种材料的吸收光谱图来评估其光吸收特性与应用潜力。光吸收系数($\alpha(\omega)$)是能最直观表达材料光学性质的重

要参数, 可以通过复介电函数求解^[42], 如式(5)和式(6)所示。

$$\varepsilon(\omega) = \varepsilon_1(\omega) + i\varepsilon_2(\omega), \quad (5)$$

$$\alpha(\omega) = \sqrt{2}\omega \cdot \left[\left(\varepsilon_1^2(\omega) + \varepsilon_2^2(\omega) \right)^{1/2} - \varepsilon_1(\omega) \right]^{1/2}, \quad (6)$$

式中 $\varepsilon(\omega)$ 代表复介电函数, $\varepsilon_1(\omega)$ 和 $\varepsilon_2(\omega)$ 分别代表实部和虚部, i 代表虚数单位。

ω 代表光的角频率。基于计算的光吸收谱数据, 我们进一步评估了材料的光谱极限最大效率(Spectroscopic Limited Maximum Efficiency, SLME)^[43]。相比于传统的 SQ 极限, SLME 综合考虑了材料的带隙类型、 $\alpha(\omega)$ 和薄膜厚度对材料性能的影响, 能够更真实地反映材料的光伏潜力。图 7(a)展示了 $\alpha(\omega)$ 随光子能量的变化曲线, $\text{Ti}_2\text{PbSnF}_6$ 的吸收系数在 1.70 eV 附近迅速升高至第一个峰值, 而 $\text{Na}_2\text{HgZnF}_6$ 的吸收增幅则在 3.40 eV 附近明显增大。这一特征与其 HSE06 计算的 E_g 高度吻合, 体现了电子从 VBM 跃迁至 CBM 的本征激发过程。在 1.65–3.10 eV 可见光区范围内, $\text{Ti}_2\text{PbSnF}_6$ 的 $\alpha(\omega)$ 保持在 10^5 数量级以上, 而 $\text{Na}_2\text{HgZnF}_6$ 的 $\alpha(\omega)$ 相对较低, 直至紫外光区才显著增加。如图 7(b)所示, 在厚度分别为 0.82 μm 和 1.24 μm 时, $\text{Ti}_2\text{PbSnF}_6$ 和 $\text{Na}_2\text{HgZnF}_6$ 的最大 SLME 分别为 28.10% 和 1.70%。这些结果表明, $\text{Ti}_2\text{PbSnF}_6$ 在可见光区具有优异的光吸收能力, 并在较薄的厚度下表现出较高的光转化效率, 显示了其作为高性能吸光材料的巨大潜力。

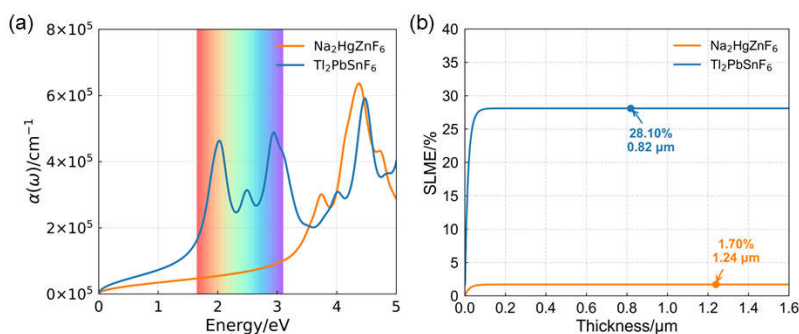


图 7 $\text{Na}_2\text{HgZnF}_6$ 和 $\text{Ti}_2\text{PbSnF}_6$ 的(a)吸收光谱; (b) SLME

Fig.7. Absorption spectra (a) and SLME (b) of $\text{Na}_2\text{HgZnF}_6$ and $\text{Ti}_2\text{PbSnF}_6$

4 结 论

本研究通过使用机器学习和符号回归算法，系统预测了 3624 种 HDPs 的 E_f 和 E_g 与结构的关系，得到了高精度的预测模型和明确的数学表达式，其中，MXGBoost 基于 Boosting 的迭代集成策略、损失函数优化与正则化能力，在 E_f 和 E_g 预测中均表现出优异的准确性(R^2 分别为 0.984 和 0.816)和泛化性能。结合 mSISSO 符号回归和 SHAP 值分析表明，X 位阴离子的电负性(χ_X)是影响材料稳定性的关键特征，而 B 位阳离子的电负性($\chi_{B''}$ 、 $\chi_{B'}$)则是决定禁带宽度的主要因素。通过第一性原理计算验证筛选出的新型稳定直接带隙 HDPs 材料中， $\text{Tl}_2\text{PbSnF}_6$ 相较于 $\text{Na}_2\text{HgZnF}_6$ 能带附近的电荷分布更加局域(IPR 分别为 0.057 和 0.224)，且本征点缺陷表现出更低的深能级浓度及更高的浅能级浓度。此外， $\text{Tl}_2\text{PbSnF}_6$ 在可见光区的光吸收能力显著增强($\alpha(\omega)$ 达到 10^5 数量级)，在较薄的器件厚度下($0.82\ \mu\text{m}$)实现了较高的理论光电转换效率(SLME 为 28.01%)，显示出优异的缺陷容忍性和光电性能。综上所述，本研究构建了完整、高效且可解释的机器学习预测框架，并成功筛选出性能卓越的 $\text{Tl}_2\text{PbSnF}_6$ 材料，为探索新型高性能光电子材料提供了理论指导。

数据集、机器学习代码及运行结果：available free of charge via the internet at https://github.com/LYB-maker/scripts/tree/main/ML_RF_outputs.

参考文献

- [1] Faizan M, Zhao G Q, Zhang T X, Wang X Y, He X, Zhang L J 2024 *Acta Phys. -Chim. Sin.* **40** 2303004 (in Chinese) [Muhammad Faizan, 赵国琪, 张天旭, 王啸宇, 贺欣, 张立军 2024 物理化学学报 **40** 2303004]
- [2] Xiang W J, Cronk E, Wall J, Li L, Zhu K, Berry J J, Lad R J, Yu L, Yan F 2024 *ACS Appl. Mater. Interfaces* **16** 44988

- [3] Wang Y R, Tian H M, Zhang D Q, Liu W L, Ma X L 2024 *Acta Phys. Sin.* **73** 028802 (in Chinese) [王月荣, 田汉民, 张登琪, 刘维龙, 马旭蕾 2024 物理学报 **73** 028802]
- [4] Wang J W, Tian H M, Wang Y R, Cao R, Xu W 2025 *Acta Phys. Sin.* **74** 038802 (in Chinese) [王纪伟, 田汉民, 王月荣, 曹蕊, 许武 2025 物理学报 **74** 038802]
- [5] Bai T X, Wang X C, Wang Z Y, Ji S J, Meng X, Wang Q J, Zhang R L, Han P G, Han K-L, Chen J S, Liu F, Yang B 2023 *Angew. Chem. Int. Ed.* **62** e202213240
- [6] Guo H W, He M M, Jiang Y, Li H, Zhang J Y, Lian M, Cui T 2025 *Acta Phys. Sin.* **74** 178401 (in Chinese) [郭宏伟, 贺苗苗, 姜云, 李会, 张金彦, 连敏, 崔田 2025 物理学报 **74** 178401]
- [7] McClure E T, Ball M R, Windl W, Woodward P M 2016 *Chem. Mater.* **28** 1348
- [8] Cai Y, Xie W, Teng Y T, Harikesh P C, Ghosh B, Huck P, Persson K A, Mathews N, Mhaisalkar S G, Sherburne M, Asta M 2019 *Chem. Mater.* **31** 5392
- [9] Yu Q L, Xu Y F, Zhang P J, Hao S W, Zhu C Q, Yang C H 2025 *Chinese J. Inorg. Chem.* **41** 1058 (in Chinese) [于麒麟, 徐怡菲, 张彭军, 郝树伟, 朱崇强, 杨春晖 2025 无机化学学报 **41** 1058]
- [10] Novikov S A, Valueva A D, Klepov V V 2024 *Dalton Trans.* **53** 12442
- [11] Ban J F, Li Z H, Zhang X L, Liu C L, Zhao X S, Zhou H W, Jiang X K 2026 *Acta Phys. Sin.* **75** 030816 (in Chinese) [班季峰, 李忠辉, 张兴隆, 刘长立, 赵星胜, 周恒为, 蒋小康 2026 物理学报 **75** 030816]
- [12] Chen X Y, Xie J J, Wang W, Yuan H H, Xu D, Zhang T, He Y L, Shen H J 2019 *Acta Chim. Sinica* **77** 9 (in Chinese) [陈薪羽, 解俊杰, 王伟, 袁慧慧, 许頔, 张焘, 何云龙, 沈沪江 2019 化学学报 **77** 9]

- [13] Jain A, Voznyy O, Sargent E H 2017 *J. Phys. Chem. C* **121** 7183
- [14] Kim J, Min K 2022 *Adv. Theor. Simul.* **5** 2200068
- [15] Luo W, Xian X Y, Zhu J, Shen Y Y, Cao L F, Chen F F, Pu Y Y, Qi F, Zhang N, Tang X S, Huang Q 2025 *ACS Appl. Mater. Interfaces* **17** 18609
- [16] Wang Z J, Chen C C, Yun X F, Han Z, Tuo Y L, Du Y X, Zhang X H, Zhang C L, Guan X N, Xie J Z, Liu G, Lu P F 2026 *Acta Phys. Sin.* **75** 030810 (in Chinese)
[王正君, 陈长城, 云雄飞, 韩昭, 拓娅莉, 杜钰玺, 张欣会, 张春玲, 关晓宁, 谢江舟, 刘刚, 芦鹏飞 2026 物理学报 **75** 030810]
- [17] Orio M, Pantazis D A, Neese F 2009 *Photosynth. Res.* **102** 443
- [18] Akbar B, Tayara H, Chong K T 2024 *iScience* **27** 109200
- [19] Wang N, Yuan B Y, Zheng F Y, Mo S Y, Zhang X H, Du L, Xing L X, Meng L, Zhao L, Aoki Y, Tang C M, Ye S Y 2024 *Adv. Funct. Mater.* **34** 2309855
- [20] Wang Y T, Sorkun M C, Brocks G, Er S 2024 *J. Phys. Chem. Lett.* **15** 4983
- [21] Sun C H, Goel R, Kulkarni A R 2024 *Langmuir* **40** 3691
- [22] Ke G L, Meng Q, Finley T, Wang T F, Chen W, Ma W D, Ye Q W, Liu T-Y 2017 *Adv. Neural Inform. Process. Syst.* **30** 3149
- [23] Zhang Z S, Xiong Q, Liu S J 2024 *J. Phys. Chem. Lett.* **15** 9795
- [24] Peng P, Liu F, Long S, Zhang C, Wan X, Tang A, Yu Z, Pan F 2025 *Int. J. Hydrogen Energy* **177** 151616
- [25] Peng T Z, Shen J K, Guo S Y, Xia X X, Li W 2026 *Acta Chim. Sinica* **84** 305 (in Chinese) [彭天资, 沈嘉克, 郭淑雅, 夏潇潇, 李炜 2026 化学学报 **84** 305]
- [26] Liu Y Q, Li Z L, Wang Y J, Li Q F, Ma, C W 2026 *Acta Phys. Sin.* **75** 95 (in Chinese) [刘亚琪, 李志龙, 王永佳, 李庆峰, 马春旺 2026 物理学报 **75** 95]

- [27] Kresse G, Furthmüller J 1996 *Comput. Mater. Sci.* **6** 15
- [28] Blöchl P E 1994 *Phys. Rev. B* **50** 17953
- [29] Perdew J P, Burke K, Ernzerhof M 1996 *Phys. Rev. Lett.* **77** 3865
- [30] del Campo J M, Gázquez J L, Trickey S B, Vela A 2012 *J. Chem. Phys.* **136** 104108
- [31] Heyd J, Scuseria G E, Ernzerhof M 2003 *J. Chem. Phys.* **118** 8207
- [32] Goyal A, Gorai P, Peng H, Lany S, Stevanović V 2017 *Comput. Mater. Sci.* **130** 1
- [33] Kumagai Y, Oba F 2014 *Phys. Rev. B* **89** 195205
- [34] Li Z 2022 *Comput. Environ. Urban. Syst.* **96** 101845
- [35] Gandomi A H, Atefi E 2020 *GENET PROGRAM EVOL M* **21** 273
- [36] Ouyang R, Curtarolo S, Ahmetcik E, Scheffler M, Ghiringhelli L M 2018 *Phys. Rev. Mater.* **2** 083802
- [37] Bartel C J, Sutton C, Goldsmith B R, Ouyang R, Musgrave C B, Ghiringhelli L M, Scheffler M 2019 *Sci. Adv.* **5** eaav0693
- [38] Macias-Pinilla D F, Giannici F 2025 *ACS Appl. Mater. Interfaces* **17** 20501
- [39] Justo J F, de Brito Mota F, Fazzio A 2002 *Phys. Rev. B* **65** 073202
- [40] Lax M 1952 *J. Chem. Phys.* **20** 1752
- [41] Xu N, Qi X R, Shen Z Q, Hu L H, Lv J, Zhong Y F, Wang B, Zou Z G 2025 *Nat. Rev. Phys.* **7** 554
- [42] Li J Q, Shi J, Shi Z X, Wang J 2026 *Acta Phys. Sin.* **75** 1 (in Chinese) [李进前, 时军, 史忠祥, 王晶 2026 物理学报 **75** 1]
- [43] Yu L, Zunger A 2012 *Phys. Rev. Lett.* **108** 068701

Machine Learning and Symbolic Regression Guided Screening of Halide Double Perovskites with Superior Optoelectronic Properties^{*}

LIU Yanbo LIU Jiadong XIONG Qing ZHANG Wutong ZHANG Zhaosheng[†]

(College of Chemistry and Materials Science, Hebei University, Baoding 071002, China)

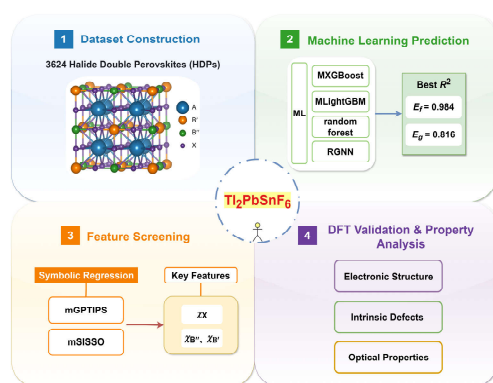
Abstract

To meet the growing demand for high-efficiency optoelectronic devices, it is imperative to accelerate the screening of halide double perovskites (HDPs)—specifically the $A_2B'B''X_6$ family—that possess exceptional optoelectronic properties. In this work, we systematically investigate the predictive performance of multiple machine learning and symbolic regression approaches for formation energies (E_f) and band gaps (E_g) based on a dataset comprising 3624 HDPs. All models were implemented within a unified MATLAB framework, including eXtreme Gradient Boosting (MXGBoost), Light Gradient Boosting Machine (MLightGBM), Random Forest, and a Residual Gated Neural Network (RGNN), together with symbolic regression methods GPTIPS and SISSO (mGPTIPS and mSISSO). The machine learning models were trained and evaluated using five-fold cross-validation to ensure robustness and generalization capability. Among all tested algorithms, MXGBoost exhibits the best predictive performance, achieving coefficients of determination (R^2) of 0.984 and 0.816 for E_f and E_g on the test sets, respectively. Furthermore, symbolic regression analysis provides explicit mathematical relationships between structural descriptors and target properties, revealing that the electronegativities of X-site anions (χ_X) and B-site cations ($\chi_{B''}$ and $\chi_{B'}$) play dominant roles in governing both thermodynamic stability and electronic structure. Based on the combined machine learning and symbolic regression framework, two thermodynamically stable direct band gap semiconductors, Na_2HgZnF_6 and Tl_2PbSnF_6 , are identified. Subsequent first-principles calculations are carried out to validate their electronic structures, intrinsic defect properties, and optical responses. The results demonstrate that Tl_2PbSnF_6 exhibits more localized charge distributions near the band edges and a significantly reduced

concentration of deep-level defects compared with $\text{Na}_2\text{HgZnF}_6$, which is beneficial for suppressing nonradiative recombination. In addition, $\text{Tl}_2\text{PbSnF}_6$ achieves a high spectroscopic limited maximum efficiency (SLME) of 28.01%, indicating excellent optoelectronic performance. Overall, this study establishes an efficient and interpretable computational framework that integrates machine learning, symbolic regression, and first-principles calculations, providing valuable guidance for the rational design and discovery of high-performance HDPs for optoelectronic applications.

Keywords: Perovskites, Machine Learning, Symbolic Regression, Electronic Structure,

Intrinsic Defects



* Project supported by the National Natural Science Foundation of China (Grant No. 22573026), the Hebei Natural Science Foundation (Grant No. B2024201051), and the Science Research Project of Hebei Education Department (Grant No. BJK2024094), and the Research Foundation of Hebei University (Grant No. DC2026275).

† Corresponding author. E-mail: misaraty@163.com
 The first author. E-mail: 16632001711@163.com