

基于 Voronoi 图结构描述的有机无机杂化钙钛矿带隙机器学习

王径舟, 欧阳润海

(上海大学材料基因组工程研究院, 上海 200444)

摘 要: 有机-无机杂化钙钛矿因其优异的物理化学性质在太阳能电池及其他光电器件中有重要应用。其中, 带隙是材料设计中必须考虑的一个关键物理量。利用机器学习建立准确的带隙模型可以帮助快速预测新材料。本工作通过研究基于 Voronoi 图方法的结构特征对机器学习的重要性发现, 相比于只有元素特征的机器学习, 通过加入 Voronoi 图结构描述特征, 符号回归(VS-SISSO)、人工神经网络(ANN)和随机森林(RF)建立的带隙模型精度均有提升。由于符号回归模型相对简单, Voronoi 图结构特征的重要性在该方法中体现最高。以上 3 种机器学习方法给出的模型预测能力接近, 但符号回归模型具有明确的数学表达式, 透明度相对更高。特征重要性分析表明, Voronoi 图结构特征在所有输入特征中的重要性最高。本工作展示了 Voronoi 图结构特征在提升机器学习模型准确度及可解释性方面的重要价值。

关键词: 有机-无机杂化钙钛矿; 带隙; Voronoi 图; 机器学习; 符号回归

中图分类号: TM914.4; TP181 文献标志码: A 文章编号: 0454-5648(2023)02-0397-08

网络出版时间: 2023-01-17



Machine Learning for the Bandgap of Organic-Inorganic Hybrid Perovskites with Voronoi Structure Representation

WANG Jingzhou, OUYANG Runhai

(Materials Genome Institute, Shanghai University, Shanghai 200444, China)

Abstract: Organic-inorganic hybrid perovskites have promising applications in solar cells and other optoelectronic devices due to their superior physical and chemical properties. A band gap is a key physical characterization that is related to the efficiency of solar energy conversion. We performed machine learning for the band gap of hybrid perovskites, and investigated the influence of structural features based on the Voronoi method on the model accuracy. The results show that compared to machine learning with only element features as an input, the accuracy of the band-gap models can be improved when the Voronoi structural feature is included in all the three methods of symbolic regression (VS-SISSO), artificial neural network (ANN) and random forest (RF). In particular, the Voronoi structural feature is of vital importance in the VS-SISSO model for its nature of being simple explicit expressions. The models from the three machine learning methods have comparable prediction accuracies, and the VS-SISSO model has better transparency and interpretability. According to the feature importance analysis, the Voronoi structural feature is the most important among all the input features.

Keywords: hybrid organic-inorganic perovskites; band gap; Voronoi tessellation; machine learning; symbolic regression

光伏发电是一种将太阳能转换为电能的可再生能源技术, 目前硅基无机材料在商业中广泛应用, 但其制造成本比较昂贵^[1]。有机-无机杂化钙钛矿(HOIPs)卤化物用于太阳能电池材料具有较高的能量转换效率^[2]。HOIPs 的化学式为 ABX₃, A 位是有机

阳离子, B 位是无机阳离子, X 为卤素阴离子。通过调节 A、B、X 的元素组成可获得巨大的化学组成空间。其中, HOIPs 带隙是决定光电效率的重要特征^[3], 一般可通过实验或第一性原理密度泛函理论(DFT)计算获得。在 DFT 计算中, 精确计算带隙需

收稿日期: 2022-11-02。 修订日期: 2022-12-05。

基金项目: 国家自然科学基金面上项目(22173058)。

第一作者: 王径舟(1997—), 男, 硕士研究生。

通信作者: 欧阳润海(1984—), 男, 博士, 副研究员。

Received date: 2022-11-02. Revised date: 2022-12-05.

First author: WANG Jingzhou (1997—), male, Master candidate.

E-mail: cnwangjingzhou@163.com

Correspondent author: OUYANG Runhai (1984—), Ph.D., Associate Professor.

E-mail: rouyang@shu.edu.cn

要使用杂化泛函^[4-5]、Green 函数屏蔽 Coulombic 相互作用(GW)近似^[6]等方法, 计算成本较高, 限制了对巨大材料空间的探索。所以, 如何准确而快速的筛选和发现新材料成为重要研究课题。

机器学习是一种从经验数据中学习规律建立模型的统计学习算法。在众多机器学习方法中, 朴素 Bayes、决策树、人工神经网络、卷积神经网络等都已经广泛应用于材料领域^[7-8]。比如, 人工神经网络用于钙钛矿的形成能、带隙的预测^[9-10], 神经网络等“黑箱”方法建立的模型往往形式复杂, 模型本身一般没有物理意义。另一类方法中, 比如符号回归, 其要求模型具有明确的相对简单的数学表达式, 故其所需数据量较小, 相对来说便于函数分析和模型预测能力的判断, 易于用户使用, 因此在可解释性方面具有独特优势。其中, 确定独立筛选和稀疏算子(SISSO)^[11-12]是一种近期开发的结合压缩感知和符号回归的机器学习方法, 已经在材料等领域有重要应用^[13-14]。给定输入特征和数据集, SISSO 方法便从巨大的表达式空间(或称特征空间)中找出准确度高复杂度低的符号回归模型。相比于其他方法, SISSO 需要较大计算机内存, 导致其能接受的特征数量往往局限在 20 多个。郭震等^[15]对 SISSO 方法进行了拓展, 建立了基于变量选择的 SISSO 方法(VS-SISSO), 使得可接受的输入特征数量大大增加, 比如几十甚至几百个, 提升了符号回归方法应对各种复杂场景的应用能力。

为了建立准确的结构-性质-性能关系(构效关系), 机器学习往往需要输入材料的结构信息。描述结构的方法很多, 比如结构图、Coulombic 矩阵、拓扑描述符、衍射指纹、正弦矩阵、多体张量表示、原子中心对称函数、原子轨道的平滑重叠(SOAP)等^[16-17]。其中结构图具有简单、灵活、可解释性高等优点, 可以与几何工具, 如 Voronoi 图^[18]结合使用。比如, Ward 等^[19]使用 Voronoi 图方法对晶体结构的三维空间进行划分, 获得结构相关特征; 然后使用决策树预测晶体形成能, 获得较准确的模型预测能力。

本工作使用了 VS-SISSO、RF、ANN 3 种机器学习方法研究结构特征对 HOIPs 带隙模型准确度的影响。考虑到 Voronoi 图结构特征具有较好的可解释性, 采纳 Voronoi 图方法提取结构特征。具体地, 先基于对 DFT 弛豫后的晶胞的 Voronoi 图划分构造大量结构特征。然后, 对比和分析 VS-SISSO、RF、ANN 共 3 种机器学习方法的模型特点和预测能力。研究结果发现, 随着 Voronoi 结构特征的加入, 3 种

机器学习模型预测精度都有明显提升, 且 3 种机器学习模型都给出了同等精度的预测能力, 说明了结构特征对机器学习准确预测材料性质的重要性。

1 方法

1.1 Voronoi 图结构描述

Voronoi 图广泛应用于有限元模拟^[20]、图像处理^[21]、路径规划^[22]等领域中。具体地, 以图 1 所示二维空间为例, 每个点(原子)与其近邻点线段间可构造一条中垂线(橙色); 该点周围所有中垂线封闭形成的空间为该点对应的一个凸多边形。每个点对应的凸边形构成了整个二维空间, 即 Voronoi 图。类似地, Voronoi 图方法可将材料三维晶体结构中每个原子对应于一个凸多面体。基于 Voronoi 图多面体提取的材料结构特征具有以下优势^[17, 19]: 不依赖于划分对象是原胞还是惯用晶胞; 对晶胞体积的放缩不敏感; 在结构弛豫中, 当结构细微变化时, Voronoi 图特征变化连续; 旋转和平移的不变性; 计算成本低。基于 Voronoi 图多面体定义的结构特征^[19]如下:

1) 有效配位数 N_{eff} :

$$N_{\text{eff}} = \frac{(\sum_i A_i)^2}{\sum_i A_i^2} \quad (1)$$

其中: N_{eff} 为有效配位数; A_i 为 Voronoi 图多面体中第 i 个面的面积。一般地, 最近邻原子对配位数的影响最大。

2) 平均键长 $\bar{l}_t$, 及键长的偏差 $\hat{l}_t$:

$$\bar{l}_t = \frac{\sum_i A_i \cdot \|r_i - r_t\|_2}{\sum_i A_i} \quad (2)$$

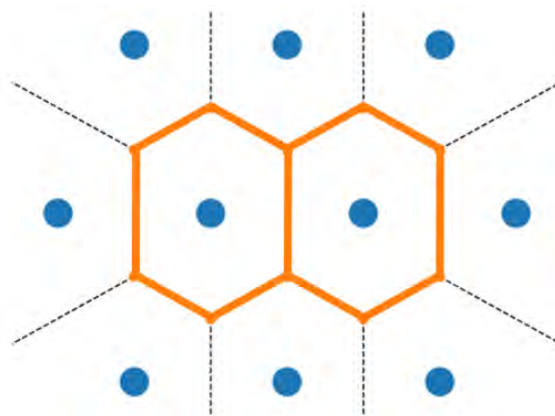


图 1 Voronoi 图划分示意图

Fig. 1 Schematics of the Voronoi tessellations

$$\hat{t}_i = \frac{\sum_i A_i \cdot \|\bar{r}_i - r_t\|_2 - \bar{t}_i}{\bar{t}_i \sum_i A_i} \quad (3)$$

其中: $\bar{r}_i$ 为中心原子的位置; $\bar{r}_i$ 为第 i 个面对应的原子位置。每个中心原子与周围原子的平均键长是以多面体的面积做加权平均的, 中心原子的最近邻原子对平均键长的影响最大。

3) 第一近邻序参数:

$$\alpha(t) = 1 - \frac{\sum_i A_i \cdot \delta_{t,t_i}}{x_t \sum_i A_i} \quad (4)$$

其中: δ 为 Dirac 函数; t 为中心原子; t_i 为第 i 个面对应的近邻原子; x_t 为 t 原子在晶胞中的浓度。序参数表示晶体的无序程度; 当原子有序排列时(如 NaCl), 序参数为 1; 否则, 无序状态序参数趋近于 0。

4) 以 Voronoi 图多面体的中心原子为球心的最大内接球的半径, 等于中心原子到多面体的面的距离的最小值。

5) 原子局域环境差异 Δ_p , 表示为以 Voronoi 图面积加权的相邻 2 个原子的物理量差值:

$$\Delta_p = \frac{\sum_i A_i |p_i - p|}{\sum_i A_i} \quad (5)$$

其中: p 为中心原子的物理量; p_i 为与 i 面原子对应的物理量。每个原子物理量的局域环境差异以多面体的各面面积加权平均。同样, 离中心原子最近的原子对 Δ_p 的影响最大。本工作考虑的物理量包括原子序数、Mendeleev 数、价电子轨道半径(s 、 p 、 d 、 f)、化学标度、原子量、共价半径、电负性、熔沸点、电离能等 25 种基础物理量。

因为晶胞中原子个数众多, 且数量不一定相等, 所以以上针对每个中心原子生成的 5 类结构特征不能直接用于训练模型, 需要进一步处理。针对晶胞中的所有原子, 计算了每种特征的平均值、最大值、最小值、平均绝对偏差作为模型训练的特征。结构相关特征共计 126 个, 其中包括根据 Voronoi 图生成的 124 个特征和 2 个晶胞体积相关特征。在 126 个结构特征中, 有 5 个在所有样本中为常数, 故不用于模型训练。

1.2 符号回归方法 VS-SISSO

符号回归是根据用户输入的变量和数学算符寻找拟合数据最佳的数学表达式的机器学习方法。SISSO^[11-12]是一种新的符号回归方法, 其穷举所有给定复杂度内的数学表达式, 然后利用压缩感知原理, 通过特征空间降维和稀疏回归找出最佳模型。但由于内存需求量较大, 输入特征的数量往往限制

在 20 多以内, 否则可能会导致内存不足而无法计算。为了解决该问题, 郭震等^[15]开发了变量选择协助的 SISSO 符号回归(VS-SISSO)。该算法将 SISSO 符号回归与迭代变量选择相结合, 从大量输入特征中自动搜索重要变量。具体工作流程(图 2)如下:

1) 从初始特征中随机产生特征子集 S_a (优先选择未被访问的特征);

2) 提取上一次迭代的模型中包含的变量 S_b (S_b 第一次迭代为空集), 合并 S_a 与 S_b 为 S ;

3) S 作为 SISSO 的输入获得当前步骤的数学表达式和训练误差;

4) 如果训练误差为历史最小值, 则更新 S_b ;

5) 如果训练误差收敛或迭代达到最大值, 程序停止, 否则返回第 1 步。

在 VS-SISSO 中影响收敛速度的超参数为 S_a 的大小, S_a 越大, VS-SISSO 会经过越少的迭代次数收敛, 但相应的 S 也越大, 单步 SISSO 的计算时间更长。本工作统一使用 $S_a = 4$ 。SISSO 中可调整的超参数包括特征复杂度和描述符维度。描述符维度定义为线性模型有多少项, 特征复杂度定义为每一项中包含的数学运算符个数, 本工作中描述符维度为 2, 特征复杂度为 7。特征构造中用到的数学操作符为: $+$, $-$, $\times$, $\div$, $\exp(x)$, $\exp(-x)$, x^{-1} , x^2 , x^3 , $x^{1/2}$, $x^{1/3}$, $\ln(x)$, $|x_1 - x_2|$ 。

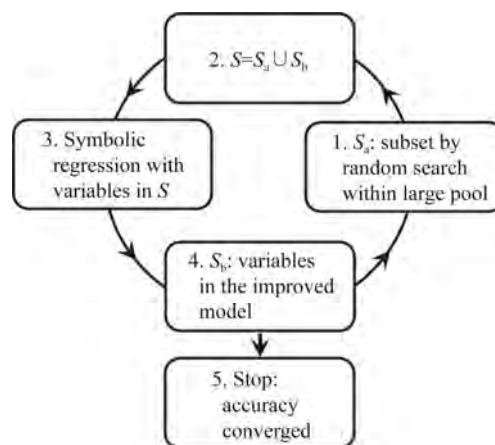


图 2 VS-SISSO 流程图^[15]

Fig. 2 Flow chart of the VS-SISSO^[15]

1.3 人工神经网络(ANN)

ANN^[23]是模拟生物神经网络结构的数学模型, 其结构包括输入层、隐藏层和输出层。每一层包含很多节点, 每个节点称为神经元。输入层神经元接收训练集的输入特征, 经过隐藏层的处理后由输出

层输出。隐藏层中每个节点的值等于将前一层所有节点值权重线性加和后再通过激活函数做非线性变换得到。常见的激活函数包括 Sigmoid 函数、双曲正切函数、线性整流函数等。基本框架用公式表达即为:

$$y_m^n = f_m^n(x_m^n) = f_m^n\left(b_m^n + \sum_{i=1}^{N_{n-1}} x_i^{n-1} w_{im}^{n-1,n}\right) \quad (6)$$

其中: y_m^n 为第 n 层的第 m 个神经元; f_m^n 为激活函数在第 n 层的第 m 个神经元的输出值; 激活函数的变量 x_m^n 等于前一层($n-1$)所有神经元的权重加和; b_m^n 为偏置; 所有 w 和 b 都是待拟合系数。给定训练数据, 神经网络一般可通过反向传播算法高效求解, 本工作的人工神经网络回归基于 SciKit-Learn^[24], 用到的参数包括神经网络隐藏层层数、每一个隐藏层中神经元的个数和 $L2$ 正则化参数。

1.4 随机森林(RF)

RF^[25]是通过构建多个决策树模型, 并对每个模型做出的决策取平均(或投票表决)的集成学习算法。随机森林在建立每棵决策树过程中, 对训练数据和特征同时进行有放回的随机抽样, 以降低决策树模型之间的相关性, 获得最大预测性能。在面对分类任务时采用投票方式, 在众多子模型中得票最多的决策作为结果; 在面对回归任务时, 使用众多子模型预测的均值作为结果。本工作中随机森林回归基于 SciKit-Learn 库实现, 具体参数包括决策树数量和决策树的最大深度。

1.5 数据集

为了探索有机-无机杂化钙钛矿的构效关系, Kim 等^[26]通过结合 Minima hopping 的结构搜索和密度泛函理论计算了 1346 个有机无机杂化钙钛矿的结构和带隙, 其中包括 16 种 A 位有机阳离子、3 种第 4 主族 B 位阳离子(Ge、Sn、Pb)和 4 种卤化物阴离子(F、Cl、Br、I), 共计 192 种化学式。对于每种成分都存在多个构型, Teunissen 等^[10]提取了最接近立方结构的 192 个, 并发现 NH_4PbF_3 和 $\text{NH}_3\text{OH}\text{SnBr}_3$ 不具有钙钛矿结构。因此, 使用 190 个立方钙钛矿作为数据集, 目标量为基于 GGA 泛函计算的带隙(E_{GGA})。用于模型训练的特征总数为 159, 其中包括 38 个不包含结构信息的体相特征(表 1)、119 个基于三维 Voronoi 图生成的结构特征(表 2, 其中 5 个方差为零的特征不算在内), 以及晶胞体积 V 和晶胞体积的立方根 $V^{1/3}$ 。取总样本数的 90% (171)作为训练集, 10% (19)作为测试集。

表 1 38 个不包含 Voronoi 图结构描述的特征

Table 1 38 features without Voronoi structural description

Feature	A	B	X
Atomic weight ^[27]	AW_A	AW_B	AW_X
Electronegativity (Pauling) ^[27]		χ_B^{Pauling}	χ_X^{Pauling}
Electronegativity (Allred-Rochow) ^[27]		χ_B^{AR}	χ_X^{AR}
Radius (Ionic) ^[28]	r_A	r_B	r_X
Radius (Van der Waals) ^[27]		r_B^{VDW}	r_X^{VDW}
Radius (Covalent) ^[27]		r_B^{CV}	r_X^{CV}
Radius (Valence s-orbital) ^[27]		r_B^s	r_X^s
Radius (Valence p-orbital) ^[27]		r_B^p	r_X^p
Radius (Valence d-orbital) ^[27]		r_B^d	r_X^d
Density of solid ^[27]		ρ_B	ρ_X
Chemical scale ^[29]		X_B	X_X
Electron affinity ^[27]		E_B^A	E_X^A
Ionization energies (1st) ^[27]		E_B^{1st}	E_X^{1st}
Ionization energies (2nd) ^[27]		E_B^{2nd}	E_X^{2nd}
Melting point ^[27]		MP_B	MP_X
Boiling point ^[27]		BP_B	BP_X
Bond enthalpy of B-X ^[30]			E^{B-X}
Octahedral factor ^[31]		μ	
Goldschmidt's tolerance factor ^[32]		t	
New tolerance factor ^[13]		τ	

2 结果与分析

为了对比 Voronoi 图结构特征的加入对模型预测能力的影响, 本工作对比了输入特征中不包含结构特征(特征数 $n=38$)和包含结构信息(特征数 $n=159$)时, VS-SISSO、ANN 和 RF 构建的机器学习模型的训练误差(图 3a)和测试误差(图 3b 和图 3c)。使用均方根误差(RMSE)衡量拟合值与真实值之间的偏差。影响 ANN 模型复杂度的主要超参数为隐藏层层数、每一个隐层中神经元的个数和 $L2$ 正则化参数。为了保证 ANN 超参数的全局最优, 超参数的确定均通过包含五折交叉验证的网格搜索。对于 $n=38$ 和 $n=159$ 时优化后的 ANN 主要超参数如表 3 所示。当输入特征数从 38 增加到 159 时, ANN 模型的测试误差从 0.343 eV 降低到 0.244 eV, 结果表明, 结构特征的加入可以提升 ANN 模型精度。但值得注意的是, 当 $n=159$ 时 ANN 的测试误差高于训练误差 0.15 eV, 表明 ANN 模型具有一定的过拟合风险。

影响 RF 模型复杂度的主要超参数为决策树数量和决策树的最大深度, 对于 $n=38$ 和 $n=159$ 时优化后的 RF 主要超参数如表 3 所示。当输入特征数从 38 增加到 159 时, RF 模型的测试误差从 0.318 eV

表 2 124 个 Voronoi 图结构特征
Table 2 124 Voronoi structural features

Feature	Average	Maximum	Minimum	MAE
Effective coordination number	CN^{avg}	CN^{max}	CN^{min}	CN^{mac}
Average bond length	$\bar{l}^{avg}$	$\bar{l}^{max}$	$\bar{l}^{min}$	$\bar{l}^{mac}$
Variance of average bond length	$\hat{l}^{avg}$	$\hat{l}^{max}$	$\hat{l}^{min}$	$\hat{l}^{mac}$
Ordering parameter	ω^{avg}	$\omega^{max} *$	ω^{min}	ω^{mac}
Packing efficiency	ε^{avg}	ε^{max}	ε^{min}	ε^{mac}
Packing radius	$r_{packing}^{avg}$	$r_{packing}^{max}$	$r_{packing}^{min}$	$r_{packing}^{mac}$
Volume of Voronoi cell	$V_{Voronoi}^{avg}$	$V_{Voronoi}^{max}$	$V_{Voronoi}^{min}$	$V_{Voronoi}^{mac}$
Local environment attributes, Δ				
Atomic weight ^[27]	Δ_{AW}^{avg}	Δ_{AW}^{max}	Δ_{AW}^{min}	Δ_{AW}^{mac}
Electronegativity (Pauling) ^[27]	Δ_{Zp}^{avg}	Δ_{Zp}^{max}	Δ_{Zp}^{min}	Δ_{Zp}^{mac}
Electronegativity (Allred-Rochow) ^[27]	Δ_{ZAR}^{avg}	Δ_{ZAR}^{max}	Δ_{ZAR}^{min}	Δ_{ZAR}^{mac}
Radius (Van der Waals) ^[27]	$\Delta_{r_{VDW}}^{avg}$	$\Delta_{r_{VDW}}^{max}$	$\Delta_{r_{VDW}}^{min}$	$\Delta_{r_{VDW}}^{mac}$
Radius (Covalent) ^[27]	$\Delta_{r_{CV}}^{avg}$	$\Delta_{r_{CV}}^{max}$	$\Delta_{r_{CV}}^{min}$	$\Delta_{r_{CV}}^{mac}$
Radius (Valence s-orbital) ^[27]	$\Delta_{r_s}^{avg}$	$\Delta_{r_s}^{max}$	$\Delta_{r_s}^{min}$	$\Delta_{r_s}^{mac}$
Radius (Valence p-orbital) ^[27]	$\Delta_{r_p}^{avg}$	$\Delta_{r_p}^{max}$	$\Delta_{r_p}^{min}$	$\Delta_{r_p}^{mac}$
Radius (Valence d-orbital) ^[27]	$\Delta_{r_d}^{avg}$	$\Delta_{r_d}^{max}$	$\Delta_{r_d}^{min}$	$\Delta_{r_d}^{mac}$
Density of solid ^[27]	Δ_{ρ}^{avg}	Δ_{ρ}^{max}	Δ_{ρ}^{min}	Δ_{ρ}^{mac}
Chemical scale ^[29]	Δ_X^{avg}	Δ_X^{max}	Δ_X^{min}	Δ_X^{mac}
Electron affinity ^[27]	$\Delta_{E_A}^{avg}$	$\Delta_{E_A}^{max}$	$\Delta_{E_A}^{min}$	$\Delta_{E_A}^{mac}$
Ionization energies (1st) ^[27]	$\Delta_{E_{1st}}^{avg}$	$\Delta_{E_{1st}}^{max}$	$\Delta_{E_{1st}}^{min}$	$\Delta_{E_{1st}}^{mac}$
Ionization energies (2nd) ^[27]	$\Delta_{E_{2nd}}^{avg}$	$\Delta_{E_{2nd}}^{max}$	$\Delta_{E_{2nd}}^{min}$	$\Delta_{E_{2nd}}^{mac}$
Boiling point ^[27]	Δ_{BP}^{avg}	Δ_{BP}^{max}	Δ_{BP}^{min}	Δ_{BP}^{mac}
Melting point ^[27]	Δ_{MP}^{avg}	Δ_{MP}^{max}	Δ_{MP}^{min}	Δ_{MP}^{mac}
Mendeleev number	Δ_{MN}^{avg}	Δ_{MN}^{max}	Δ_{MN}^{min}	Δ_{MN}^{mac}
s valence electrons	$\Delta_{n_s}^{avg}$	$\Delta_{n_s}^{max}$	$\Delta_{n_s}^{min}$	$\Delta_{n_s}^{mac}$
p valence electrons	$\Delta_{n_p}^{avg}$	$\Delta_{n_p}^{max}$	$\Delta_{n_p}^{min}$	$\Delta_{n_p}^{mac}$
d valence electrons	$\Delta_{n_d}^{avg} *$	$\Delta_{n_d}^{max} *$	$\Delta_{n_d}^{min} *$	$\Delta_{n_d}^{mac} *$
Row	Δ_{Row}^{avg}	Δ_{Row}^{max}	Δ_{Row}^{min}	Δ_{Row}^{mac}
Column	Δ_{Column}^{avg}	Δ_{Column}^{max}	Δ_{Column}^{min}	Δ_{Column}^{mac}
Atomic number	Δ_N^{avg}	Δ_N^{max}	Δ_N^{min}	Δ_N^{mac}
Ground state volume ^[33-34]	$\Delta_{V_{GS}}^{avg}$	$\Delta_{V_{GS}}^{max}$	$\Delta_{V_{GS}}^{min}$	$\Delta_{V_{GS}}^{mac}$
Ground state bandgap ^[33-34]	$\Delta_{E_{GS}}^{avg}$	$\Delta_{E_{GS}}^{max}$	$\Delta_{E_{GS}}^{min}$	$\Delta_{E_{GS}}^{mac}$

* Feature with zero variance is not included in modeling.

降低到 0.213 eV，结果表明，结构特征的加入可以提升 RF 模型精度。但值得注意的是，当输入特征分别为 38 和 159 时，RF 的测试误差高于训练误差分别为 0.20 eV 和 0.10 eV，表明 RF 模型也具有一定的过拟合风险。分析 $n=159$ 时的基尼重要性，排名前 10 的特征中有 9 个是基于 Voronoi 图生成的结构特征，表明结构信息能够提升 RF 模型的精度。

对于 VS-SISSO，本工作做了 3 次 $n=38$ 的模型训练。3 次训练收敛时误差的极差为 0.011 eV，且均在平均值(0.270 eV)附近，表明 3 次训练结果接

近。当 $n=159$ 时，3 次 VS-SISSO 训练收敛时误差的极差为 0.009 eV，且均在平均值(0.230 eV)附近，同样表明该方法的稳定性。输入特征数从 38 增加到 159，VS-SISSO 模型的测试误差从 0.269 eV 降低到 0.224 eV。关于过拟合风险， $n=38$ 时的测试误差最低为 0.269 eV，低于训练误差 0.014 eV； $n=159$ 测试误差最低为 0.224 eV，低于训练误差 0.001 eV。2 种特征输入下 VS-SISSO 的测试误差均接近测试误差，表明 VS-SISSO 模型的过拟合风险相对较低。

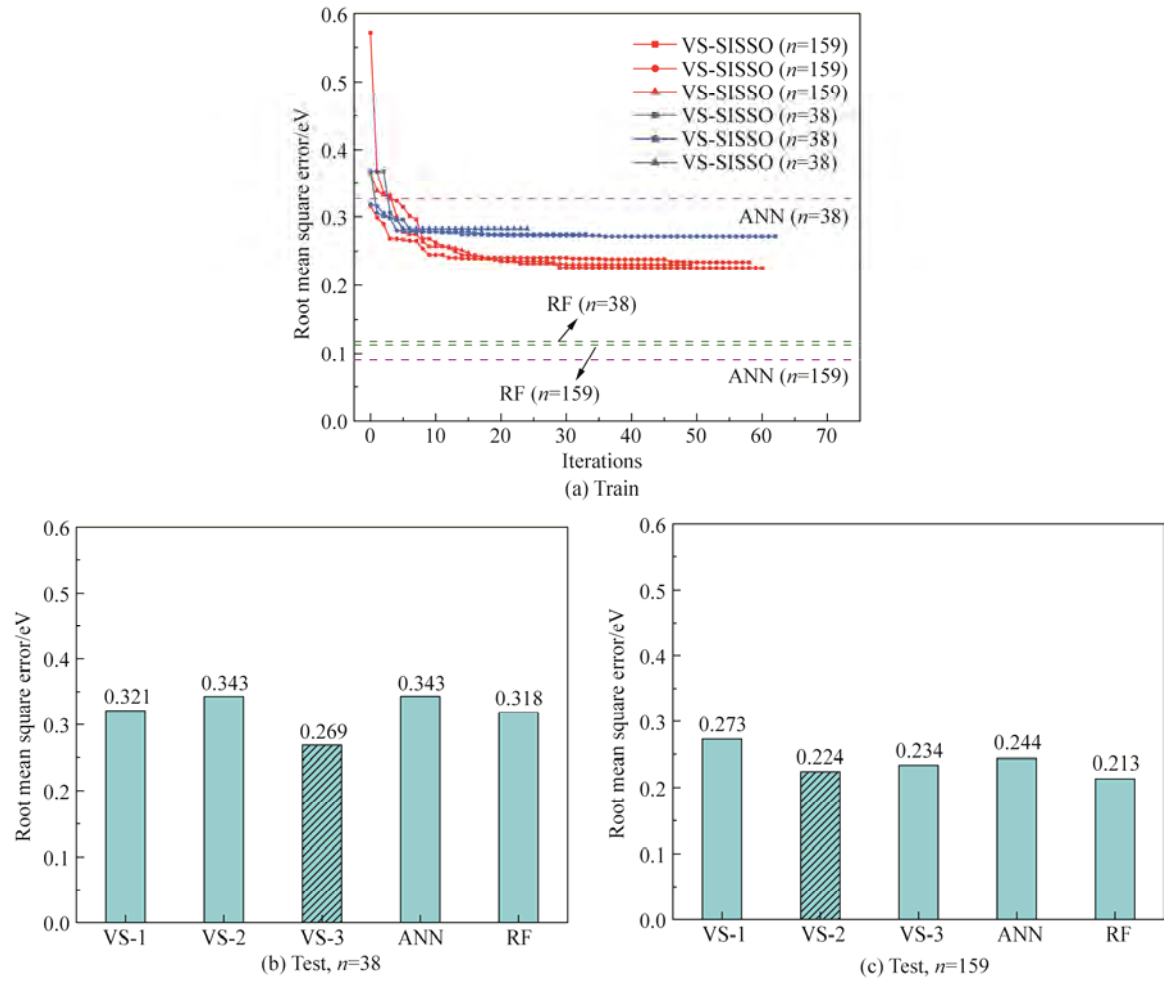


图3 VS-SISSO (运行3次)、人工神经网络(ANN)和随机森林(RF)的训练结果和测试结果的比较。(a) 特征数量 $n=38$ 和 $n=159$ 的训练结果; (b) $n=38$ 的测试结果; (c) $n=159$ 的测试结果

Fig. 3 Comparison of the training and testing results between VS-SISSO (repeated 3 times), artificial neural network (ANN) and random forest (RF). (a) training results with the number of input features $n=38$ and $n=159$, respectively; (b) testing results of $n=38$; (c) testing results of $n=159$

表3 优化后的 ANN 和 RF 模型主要超参数

Table 3 Optimized hyper-parameters of ANN and RF models					
n	ANN			RF	
	Number of hidden layers	Number of neurons of a single hidden layer	L2 regularization parameter	Number of decision trees	Maximum depth of decision trees
38	2	20	0.000 1	100	10
159	4	40	0.000 1	100	10

可见,当输入变量包含 Voronoi 图结构特征后,VS-SISSO、ANN 和 RF 的训练误差和测试误差均有下降,说明 Voronoi 图结构特征中包含了更精确描述 HOIPs 带隙的关键变量。虽然 VS-SISSO 的训练误差高于 ANN 和 RF,但测试误差与二者相近。综上所述,本工作中的 VS-SISSO 与 RF 和 ANN 相比具有更低的过拟合风险,其原因是:1) SISSO 作为符号回归方法,建立的模型的复杂程度低于 ANN 和 RF;2) VS-SISSO 通过迭代筛选获得关键特征,

规避了对提升模型精度无益的冗余特征;3) 用于训练的样本数过少,ANN 和 RF 的复杂模型过拟合相对较高。此外,以上所有模型的预测精度都在 0.2 以上。若要进一步降低误差,除了考虑更先进的机器学习算法和结构描述符之外,还需要考虑计算数据本身的准确度问题。众所周知,DFT 计算采用 GGA 泛函不能准确描述材料的带隙,与实验或 GW 方法相比往往存在较大差距,比如 0.5 eV;且其差距可能随具体材料而异。因此,改进数据集本身的质

量有望进一步降低模型的误差。

针对 VS-SISSO 模型做进一步分析。以包含 Voronoi 图结构特征且测试误差最低的模型(VS-2)为例, 其函数形式为:

$$E_{\text{GGA}}(\text{eV}) = 4.34 + 2.18 \frac{r_{\text{B}}^{\text{VDW}} - r_{\text{B}}^{\text{CV}} - r_{\text{X}} + \Delta_{\text{r}^{\text{VDW}}}^{\text{avg}}}{\frac{r_{\text{X}}^{\text{P}}}{r_{\text{packing_max}}} + \frac{r_{\text{B}}}{\epsilon_{\text{packing_mac}}}} + 0.88 \left| (r_{\text{B}}^{\text{s}} + r_{\text{B}}^{\text{VDW}}) e^{\Delta_{\text{r}^{\text{s}}}^{\text{min}}} - (r_{\text{X}}^{\text{P}} + r_{\text{X}}^{\text{VDW}}) \ln(V_{\text{Voronoi}}^{\text{min}}) \right| \quad (7)$$

其中: r 、 r^{CV} 、 r^{VDW} 、 r^{s} 、 r^{P} 分别表示 Shannon

离子半径、共价半径、van der Waals 半径、s 轨道半径和 p 轨道半径, r^{packing} 、 $\epsilon^{\text{packing}}$ 分别表示 Voronoi 图多面体内接球的半径和填充率, Δ 表示局部环境差异, $V_{\text{Voronoi}}^{\text{min}}$ 表示晶胞经过 Voronoi 图划分后形成多个多面体中的最小体积。将模型中特征的重要性定义为: 对模型中出现的特征逐一后向删除, 训练误差上升越大的表明特征越重要。图 5 展示了式(7)中涉及到的 12 个特征的重要性, $V_{\text{Voronoi}}^{\text{min}}$ 为最重要的特征。该结果表明 Voronoi 图结构特征能够有效提升模型精度, 也说明了原子的局域环境在建立材料构效关系中的重要作用。

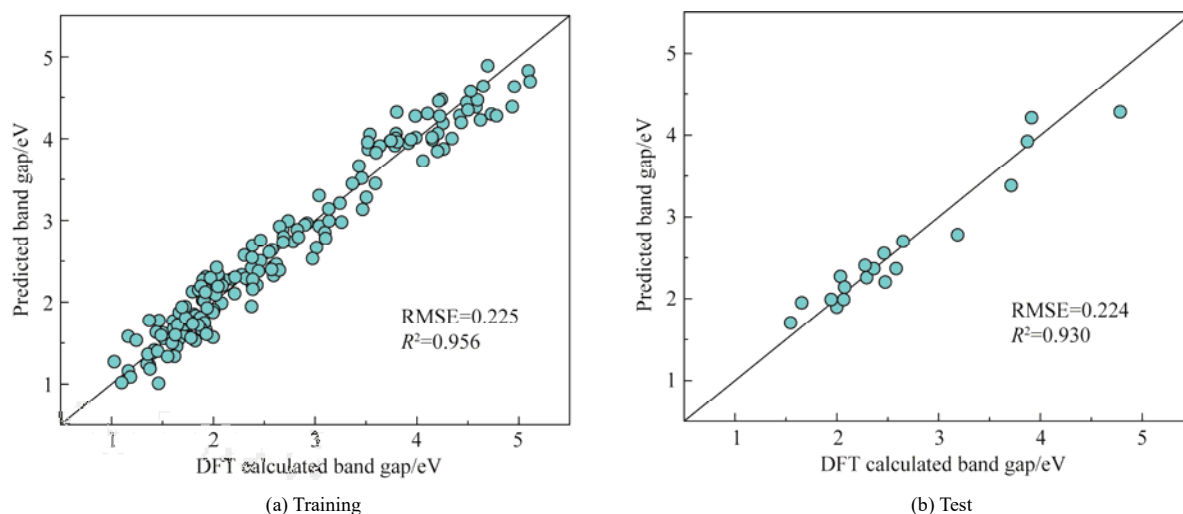


图 4 VS-SISSO 模型拟合 HOIPs 带隙与 DFT 数据对比图

Fig. 4 Comparison between the prediction by VS-SISSO and the DFT data on the bandgap of HOIPs

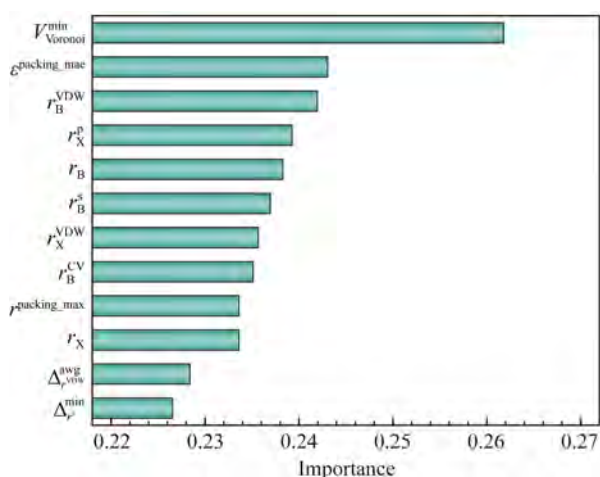


图 5 式(7)中特征的重要性分析(后向删除法)

Fig. 5 Importance of features in Equation (7)

郭震等^[15]也曾使用 VS-SISSO 研究无机卤化钙钛矿的带隙, 带隙与 X 原子最高占据能级负相关。在 HOIPs 带隙的 VS-SISSO 模型中, 带隙与 X 离子半径负相关。随着 X 位原子从 F 到 I 的变化, 原子

最高占据能级增加, 离子半径增加, 表明 X 元素对 HOIPs 和无机卤化钙钛矿的带隙均的影响趋势相同。同时, 在郭震等的 VS-SISSO 模型中 a/b 为最重要特征, $V_{\text{Voronoi}}^{\text{min}}$ 为最重要特征, 这都表明了结构特征对于准确描述钙钛矿带隙的必要性。相较于郭震等的研究中结构特征仅包含 a/b (晶格参数)和原子堆积因子(体积信息), 本工作使用的 Voronoi 图方法可以更全面的描述晶体结构, 包括原子配位、键长、无序度等局域原子环境信息。该结构描述方法可以应用到任何材料体系, 且因为每个特征都具备较明确的物理意义, 特别适用于可解释机器学习建模问题。

3 结论

基于 DFT-GGA 计算的 HOIPs 材料结构和带隙, 先构造了 Voronoi 图结构特征, 然后并分别采用 VS-SISSO、ANN、RF 3 种机器学习方法建立了较为准确的机器学习模型。结果表明, 在 3 种机器学习

习方法中,相比于只含元素特征的机器学习,加入描述原子局域结构环境的 Voronoi 图结构特征的会明显提升模型的精度,说明了结构信息在材料构效关系中的重要作用。对比 3 种机器学习模型,他们都给出了同等精度的预测能力,但 VS-SISSO 可以通过迭代筛选获得关键特征,得到了相对简洁的可解释性较好的数学公式。根据对 VS-SISSO 模型的特征重要性分析, $V_{\text{Voronoi}}^{\text{min}}$ 重要性最高,进一步证实了结构信息对机器学习建模的重要性。使用的 Voronoi 图方法不局限于特定材料体系,可以较全面地描述晶体结构,包括原子配位、键长、无序度等局域原子环境信息;且因为每个特征都具备较明确的物理意义,对提升模型准确度及可解释性具有重要价值。

参考文献:

- [1] NA G R. A first-principles study on two-dimensional hybrid organic-inorganic halide perovskite optoelectronic materials[D]. Jilin: Jilin University, 2021.
- [2] GLADKIKH V, KIM D Y, HAJIBABAEI A, et al. Machine learning for predicting the band gaps of ABX_3 perovskites from elemental properties[J]. *Phys Chem C*, 2020, 124(16): 8905–8918.
- [3] ZHOU Y, GAME O S, PANG S, et al. Microstructures of organometal trihalide perovskites for solar cells: Their evolution from solutions and characterization[J]. *J Phys Chem Lett*, 2015, 6(23): 4827–4839.
- [4] BECKE A D. A new mixing of Hartree-Fock and local density functional theories[J]. *J Chem Phys*, 1993, 98(2): 1372–1377.
- [5] PEVERATI R, TRUHLAR D G. Improving the accuracy of hybrid meta-GGA density functionals by range separation[J]. *J Phys Chem Lett*, 2011, 2(21): 2810–2817.
- [6] ARYASETIWAN F, GUNNARSSON O. The GW method[J]. *Rep Progress Phys*, 1998, 61(3): 237.
- [7] WEI J, CHU X, SUN X Y, et al. Machine learning in materials science[J]. *InfoMat*, 2019, 1(3): 338–358.
- [8] MORGAN D, JACOBS R. Opportunities and challenges for machine learning in materials science[J]. *Annual Rev Mater Res*, 2020, 50: 71–103.
- [9] WU T, WANG J. Global discovery of stable and non-toxic hybrid organic-inorganic perovskites for photovoltaic systems by combining machine learning method with first principle calculations[J]. *Nano Energy*, 2019, 66: 104070.
- [10] TEUNISSEN J L, DA PIEVE F. Molecular bond engineering and feature learning for the design of hybrid organic-inorganic perovskite solar cells with strong noncovalent halogen-cation interactions[J]. *J Phys Chem C*, 2021, 125(45): 25316–25326.
- [11] OUYANG R, CURTAROLO S, AHMETCIK E, et al. SISSO: A compressed-sensing method for identifying the best low-dimensional descriptor in an immensity of offered candidates[J]. *Phys Rev Mater*, 2018, 2(8): 083802.
- [12] OUYANG R, AHMETCIK E, CARBOGNO C, et al. Simultaneous learning of several materials properties from incomplete databases with multi-task SISSO[J]. *J Phys: Materials*, 2019, 2(2): 024002.
- [13] BARTEL C J, SUTTON C, GOLDSMITH B R, et al. New tolerance factor to predict the stability of perovskite oxides and halides[J]. *Sci Adv*, 2018, 5(2): eaav0693.
- [14] OUYANG R. Exploiting ionic radii for rational design of halide perovskites[J]. *Chem Mater*, 2019, 32(1): 595–604.
- [15] GUO Z, HU S B, HAN Z K, et al. Improving symbolic regression for predicting materials properties with iterative variable selection[J]. *J Chem Theory Comput*, 2022, 18(8): 4945–4951.
- [16] LI S, LIU Y, CHEN D, et al. Encoding the atomic structure for machine learning in materials science[J]. *Wiley Interdiscipl Rev: Comput Molecul Sci*, 2022, 12(1): e1558.
- [17] HIMANEN L, JÄGER M O J, MOROOKA E V, et al. DScribe: Library of descriptors for machine learning in materials science[J]. *Computer Phys Commun*, 2020, 247: 106949.
- [18] AURENHAMMER F. Voronoi diagrams—A survey of a fundamental geometric data structure[J]. *ACM Comput Surveys (CSUR)*, 1991, 23(3): 345–405.
- [19] WARD L, LIU R, KRISHNA A, et al. Including crystal structure attributes in machine learning models of formation energies via Voronoi tessellations[J]. *Phys Rev B*, 2017, 96(2): 024104.
- [20] GHOSH S, MALLETT R L. Voronoi cell finite elements[J]. *Comput Struct*, 1994, 50(1): 33–46.
- [21] WANG J, JU L, WANG X. An edge-weighted centroidal Voronoi tessellation model for image segmentation[J]. *IEEE Transact Image Process*, 2009, 18(8): 1844–1858.
- [22] BHATTACHARYA P, GAVRILOVA M L. Roadmap-based path planning-using the voronoi diagram for a clearance-based shortest path[J]. *IEEE Robot Automn Magazine*, 2008, 15(2): 58–66.
- [23] CIREGAN D, MEIER U, SCHMIDHUBER J. Multi-column deep neural networks for image classification[C]//2012 IEEE conference on computer vision and pattern recognition. IEEE, 2012: 3642–3649.
- [24] PEDREGOSA F, VAROQUAUX G, GRAMFORT A, et al. Scikit-learn: Machine learning in Python[J]. *J Machine Learn Res*, 2011, 12: 2825–2830.
- [25] HO T K. Proceedings of 3rd international conference on document analysis and recognition[J]. IEEE, 1995, 1: 278–282.
- [26] KIM C, HUAN T D, KRISHNAN S, et al. A hybrid organic-inorganic perovskite dataset[J]. *Scient Data*, 2017, 4(1): 1–11.
- [27] WebElements, <https://www.webelements.com> [2022-11-02].
- [28] SHANNON R D. Revised effective ionic radii and systematic studies of interatomic distances in halides and chalcogenides[J]. *Acta Cryst Section A*, 1976, 32(5): 751–767.
- [29] PETTIFOR D G. A chemical scale for crystal-structure maps[J]. *Solid State Commun*, 1984, 51(1): 31–34.
- [30] COTTRELL T L. The strengths of chemical bonds[M]. Butterworths Scientific Publications, 1958.
- [31] LI C, SOH K C K, WU P. Formability of ABO_3 perovskites[J]. *J Alloys Compounds*, 2004, 372(1/2): 40–48.
- [32] GOLDSCHMIDT V M. Die gesetze der krystallochemie[J]. *Naturwissenschaften*, 1926, 14(21): 477–485.
- [33] SAAL J E, KIRKLIN S, AYKOL M, et al. Materials design and discovery with high-throughput density functional theory: the open quantum materials database (OQMD)[J]. *J Materiomics*, 2013, 65(11): 1501–1509.
- [34] KIRKLIN S, SAAL J E, MEREDIG B, et al. The open quantum materials database (OQMD): Assessing the accuracy of DFT formation energies[J]. *NPJ Comput Mater*, 2015, 1(1): 1–15.