

机器学习方法预测含硼材料能隙

李珺卿 宋千禧 刘子义 王东琪*

(大连理工大学精细化工国家重点实验室 辽宁省碳资源催化转化重点实验室 化学学院 化工学院 大连 116024)

摘要 近年来,含硼材料在新能源、催化等领域日益受到重视,然而,对于高附加值的含硼材料发展还存在很高的技术壁垒.因此,亟需深入研究含硼材料微观性质间的关联关系,推动高端含硼材料的研发.本工作面向材料研究从传统的试错法向数据驱动的研究范式转变的需求,通过特征选择、网格搜索优化以及特征重要性分析,探索了多种重要的机器学习算法在含硼材料能隙预测中的应用.结果表明,采用随机森林算法的能隙预测模型决定系数(R^2)可达 0.84,并发现含硼材料的总磁化强度(total magnetization)特征与能隙存在显著的负相关关系,即材料的总磁化强度越小,其能隙越大.本工作表明机器学习方法可用于定向设计具有特定能隙的含硼材料.同时,结果也表明,作为一种集成学习模型,随机森林具有较好的学习能力与稳定的预测性能,可以应用到其它类型材料体系的能隙以及其它材料属性的预测,加速材料性能的设计与优化过程,对新型功能材料的快速筛选与高性能预测具有重要的科学意义.

关键词 含硼材料; 能隙; 总磁化强度; 机器学习

Machine Learning for Predicting Band Gap in Boron-containing Materials

Li, Junqing Song, Qianxi Liu, Ziyi Wang, Dongqi*

(State Key Laboratory of Fine Chemistry, Key Laboratory of Catalytic Conversion of Carbon Resources, School of Chemistry, School of Chemical Engineering, Dalian University of Technology, Dalian, Liaoning 116024, China)

Abstract New materials are an important driving force for social development. In recent years, attention on boron-containing materials is growing in the fields of new energy and catalysis, and calls for compelling need for in-depth study of their structure-property relationship to contribute to the research and development of boron-containing materials. In this work, on the aware of the shift of materials research from traditional trial-and-error paradigm to data-driven research paradigm, we explored the application of ten important machine learning algorithms in the prediction of band gaps of boron-containing materials through feature selection (Pearson correlation analysis), grid search-based optimization (Model optimal parameters), and feature importance analysis (interpretability analysis of the model). The results show that the band gap prediction model using the Random Forest algorithm outperforms the other models with a 84% prediction accuracy, and the total magnetization of boron-containing materials is identified to significantly correlate negatively with the band gap, i.e. the smaller the total magnetization of the material, the larger its band gap. The advantage of the Random Forest algorithm over other models is that it is better able to capture correlations between features. For example, the linear model is unable to detect the importance of the total magnetization of boron-containing materials from the material features, thus leading to a lower model prediction performance. This work shows that machine learning methods can be used to guide the design of boron-containing materials with specific band gap. Meanwhile, the results also show that, as an integrated learning model, Random Forest has good learning ability and stable prediction performance, and can be applied to the prediction of band gap and other material properties of other types of material systems, accelerating the design and optimization process of material properties, and is of great scientific significance for the rapid screening and high-performance prediction of new functional materials.

Keywords boron-containing materials; band gap; total magnetization; machine learning

1 引言

硼元素的三价电子构型导致其轨道数多于价电子数,在与 p 区元素形成的硼化物中易形成多中心键^[1].多中心键对体系的稳定性和电子性质起着决定性的作

用,这也使得硼元素成为了化学性质最复杂的化学元素之一.尽管硼元素的研究历史悠久,但由于其复杂的化学性质,硼元素仍然是一种神秘的元素.近些年的理论计算证明了许多纳米尺度的硼结构,例如纳米线、纳米带、纳米片和纳米管的存在^[2-6].

* E-mail: wangdq@dlut.edu.cn

Received October 27, 2023; published January 5, 2024.

Supporting information for this article is available free of charge via the Internet at <http://sioc-journal.cn>.

Project supported by the State Key Research and Development Program (2021YFA1500301), the Fundamental Research Funds for the Central Universities (DUT20RC(3)081), and the LiaoNing Revitalization Talents Program (XLYC2002015).

项目受科技部重点研发专项(2021YFA1500301)、中央高校基础研究基金项目(DUT20RC(3)081)和辽宁兴辽英才计划(XLYC2002015)资助.

在周期表中,硼元素距离碳元素最近,二者的电子性质相似.这种与碳相似的基本性质使得硼在清洁能源^[7-9]、催化^[6,10-11]和生物大分子^[12]等方面显示出广泛的应用前景.硼、碳元素结合可产生多种具有丰富的结构和电子性质的纳米材料,是优异性能的纳米器件的候选材料^[13-14].同样地,硼与氮、磷、氧等元素可以合成出多种具有高比表面积、高机械强度和高热导率的二维含硼材料^[11,15-16],在烷烃活化和光水解等催化领域展现了优异的性能^[6,11,17-18].金属含硼材料也是受到广泛关注的领域.过渡金属含硼材料^[19-20]具有高导热性和导电性,低磨损率和高热稳定性等特点被应用于高温材料和催化领域.稀土含硼材料^[21]具有功函数低、电导率高等特点,是电子发射材料、热电材料和超导材料领域的潜在候选材料,此外在海水淡化和污水净化等环境领域起到了重要的作用.

尽管含硼材料在多个材料领域表现出了优异的性能,但与氧化物、氮化物等其他材料家族相比,对含硼材料的认识仍显著不足,含硼材料的发现和设计仍有广阔的空间.认识并理解已知的含硼材料的电子性质有助于设计和筛选新的含硼材料.其中材料的能隙是影响半导体性能的决定因素,对评估材料在能源等领域的应用十分重要,需要深入研究,揭示能隙与材料组分和结构的关系^[22].然而,含硼材料种类多,现有的能隙实验数据有限,高通量数据表征也很难实现.因此,能隙的准确预测与计算对靶向设计和筛选含硼材料具有重要的意义.

高通量^[23]技术加速了新材料的设计和发现,在此基础上,研究人员建立了相关的材料性质的数据库^[24-26].随着材料基因组计划的开展^[27],数据驱动的第四研究范式也逐渐日益受到重视.该研究范式仰赖的三个要素是数据、计算能力和算法,通过对一些宏量材料数据进行数据挖掘,提取相关信息并对目标信息做出预测,减少实验的盲目性.

机器学习作为一种人工智能技术,已广泛应用于计算机视觉和自然语言等领域.近年来,机器学习方法开始被用于材料研究,特别是物理化学性质预测和材料高通量搜索中的预筛选.例如, Hansen 等^[28]根据团簇中原子的结构信息(原子直角坐标、核电荷等),利用线性回归算法建立了团簇中原子的结构信息与团簇能量的关系模型,进而预测新团簇能量. Tong 等^[29]使用核岭回归(kernel ridge regression)算法预测出的势能面加速 CALYPSO^[30]对硼团簇的结构预测,其速度比密度泛函理论(DFT)计算快了1~2个数量级. Jia 等^[31]以元素属性作为特征,使用支持向量机算法预测了氮化物的能隙,发现随着阳离子类型数量的增加,能隙逐渐变窄,并且能带的偏移量趋于增加. Dey 等^[32]基于28个已获得实验能隙的化合物,使用化合物元素特征描述符,用最小二乘回归(least squares regression)、稀疏偏最小二乘回归

(sparse partial least squares regression)和套索(lasso)算法预测了200种三元黄铜矿(CuFeS₂等ABC₂类)化合物的直接能隙,其中A、B、C分别代表金属元素(A、B)和非金属元素(C),结果表明B类(Fe)和C类(S)元素的原子序数以及C类元素单质的熔点为能隙的主要影响因素.这种以熔点作为描述符的策略,可用于挖掘应用于太阳能电池的黄铜矿化学材料. Yang 等^[33]使用集成学习算法发现了类金刚石化合物的电负性差与化合物中阳离子平均原子质量对预测热电材料类金刚石化合物的能隙值具有重要影响,对新型功能材料的快速筛选具有科学意义.

这些研究工作表明了机器学习方法在预测材料性质研究中的潜力,虽然在各自的研究方向取得了较好的预测效果,但是预测性能仍有很大的提升空间($R^2 \approx 0.77$ ^[33]).另外,已发表的应用机器学习算法研究材料性质的工作多采用材料的宏观物理和化学性质作为体系的描述符,这相当于将可能富含微观作用机制的材料作为一个“黑箱”处理,不利于深入认识材料的微观结构和材料的宏观性质之间的关联关系.对于含硼材料尤其是如此,相关研究很少,缺少含硼材料能隙预测模型和基于含硼材料性质的描述符选择等预测含硼材料性质相关的工作.

为了评估机器学习方法预测含硼材料性质,助力含硼材料的理性设计,设计并发现与含硼材料最相关的描述符是非常必要的.本文展示了特征选取的过程,使用机器学习算法建立了一种稳定有效的含硼材料能隙预测模型,并找出与含硼材料能隙最相关的材料特征以及对能隙的影响.

2 结果与讨论

本文的目标是利用已知的含硼材料能隙数据预测出新含硼材料的能隙值,并与通过高精度量子化学计算获得的结果进行对比,验证预测结果的准确性,评估训练得到的含硼材料能隙预测模型的可靠性.

2.1 数据集分析

通过 python materials genomics (pymatgen)^[34]从 materials project^[25]数据库中获得含硼材料数据,共计3180种含硼材料,其中包含二元含硼材料105种,三元含硼材料703种,四元含硼材料1702种,五元含硼材料547种,六元及以上含硼材料123种.

数据集中共存在81种非B元素(图1a),共有1294种元素组合,其中O(2508)、Li(1013)、H(589)、Fe(488)元素占比最高,数据集中78.8%的化合物都含有O元素(普通氧化物69.2%,氢氧化物8.3%,超氧化物0.9%,臭氧化物0.3%以及过氧化物0.1%).对数据集中的多元含硼材料的能隙的高斯核密度分析图如图2和图S22(支持信息)所示,三元和四元含硼材料存在较多的峰,表

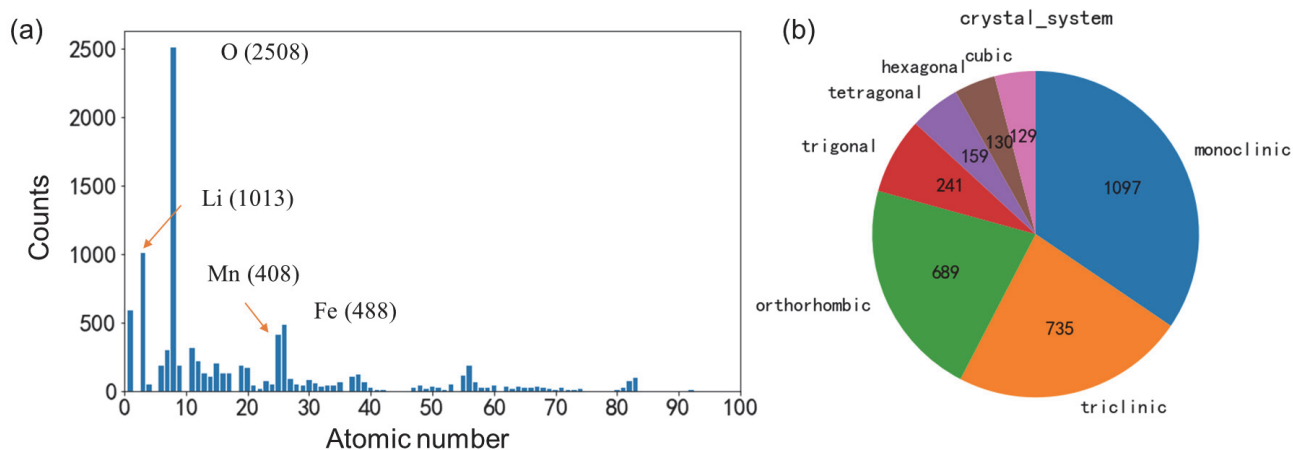


图1 数据集分析: (a)含硼材料数据集中非B元素统计, (b)含硼材料晶系统计

Figure 1 Analysis of the dataset: (a) statistics of non-B elements in the dataset of boron-containing materials, (b) statistics of crystal systems of boron-containing materials

明相比于其他多元含硼材料, 三元、四元和五元含硼材料的能隙分布具有较显著的离散的特征, 但三元和四元含硼材料的不同能隙分布区域之间的数量比例更为均衡.

含硼数据集的晶系统计结果如图 1b 所示. 根据数据集中各材料的结构特征, 可划分为 7 种晶系, 32 种点群, 158 种空间群. 其中单斜晶系的材料占比最高(34%), 其次为三斜和正交晶系, 三种晶系的材料加起来占全部数据的 80%. 在机器学习模型训练中, 将数据集以

70%:30%的比例随机划分为训练集和测试集.

2.2 特征分析

在训练回归模型之前, 通过特征选择, 构建对目标变量能隙具有有效预测能力的输入特征空间, 可以显著地提高模型预测的准确性, 增强模型的泛化能力. 为了更加清楚地了解特征对训练样本的结果影响, 本文分别对训练样本的全训练样本特征和经相关性分析筛选后的样本特征进行模型的搭建.

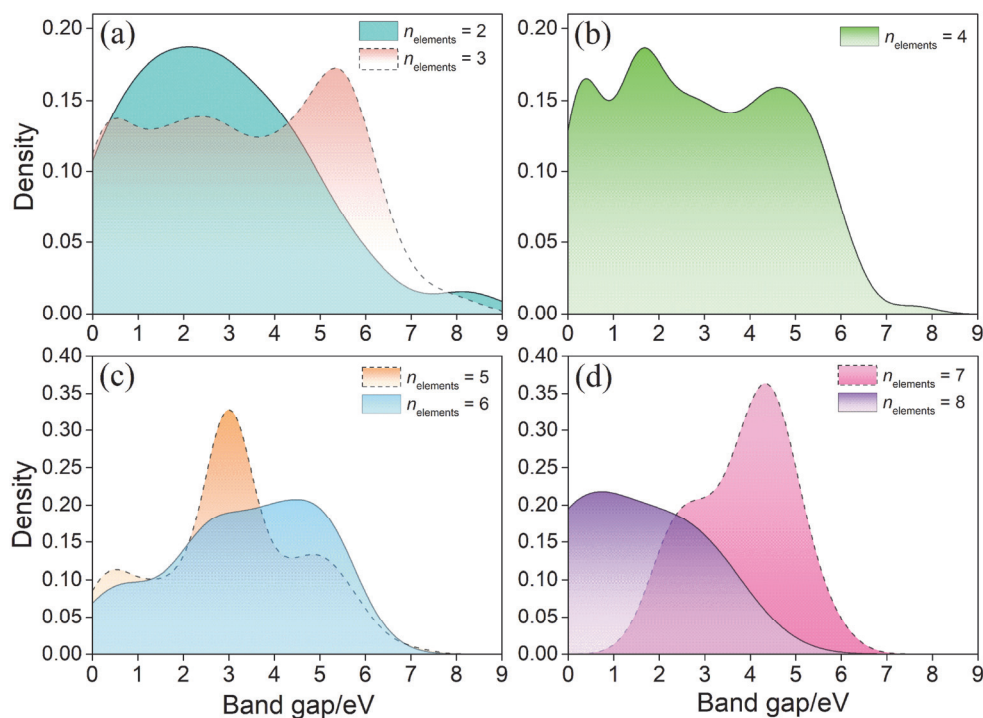


图2 2~8元含硼材料的能隙的高斯核密度估计图(二元、三元(a), 四元(b), 五元、六元(c), 七元、八元(d))

Figure 2 Gaussian kernel density estimates for the band gap of boron-containing materials from 2 to 8 elements (binary, ternary (a); quaternary (b); quinary, senary (c); septenary, octary (d))

2.2.1 全训练特征选取

含硼材料数据集中共有 163 个特征, 其中存在文本信息特征(非结构化数据特征), 经筛选有 145 个特征可通过回归模型训练的特征(基于晶体结构和元素组成的基础特征, 包括电负性、质量分数、原子序数、熔点等), 这里我们视为本研究的含硼材料数据集的“全特征”。

2.2.2 特征相关性分析

对筛选出的 145 种特征进行了皮尔逊相关性分析^[35], 以分析并识别出具有高度相关性的特征, 消除特征之间的多重共线性, 避免因特征之间可能存在的高度相关关系而导致回归模型预测失真或难以预测准确。这种现象在线性模型中尤其明显。因此, 有必要对于具有强相关性的多个特征进行甄别, 筛除冗余特征, 保留一个有效特征。根据分析结果绘制的热力图如图 S1(支持信息)所示。经过皮尔森相关性分析将 0.5 以上的相关性进行简化, 最终共筛选出 15 个特征, 其中特征的描述列于表 1。这 15 个特征的相关性热力图如图 3 所示。

2.3 回归结果分析

2.3.1 全特征预测结果

首先, 以上述 145 种材料特征作为描述符进行训练

表 1 筛选后的特征及其描述

Table 1 Selected features and their descriptions

Features	Descriptions
total magnetization	total magnetizing strength of the material
dens	density of the material
energy	energy of the material
energy per atom	atomic energy per unit of energy
formation energy per atom	unit atomic formation energy
e above hull	phase stability energy (difference in energy between new phase and equilibrium phase)
minimum Number	minimum atomic number
minimum Mendeleev Number	minimum value of Mendeleev number of elements in compounds
maximum Melting T	compound melting temperature maximum
maximum Nd Valence	maximum number of d-valence orbitals filled
mean Np Unfilled	mean number of unfilled p-valence orbitals
maximum Nd Unfilled	maximum number of unfilled d-valence orbitals
maximum Nf Unfilled	maximum number of unfilled f-valence orbits
maximum Space Group Number	maximum space group of the ground state structure at 0 K
minimum oxidation state	minimum oxidation state

可预测能隙的模型。结果如表 2 所示, 不同的回归模型对材料特征的(非)线性表达能力和核映射空间变化能力

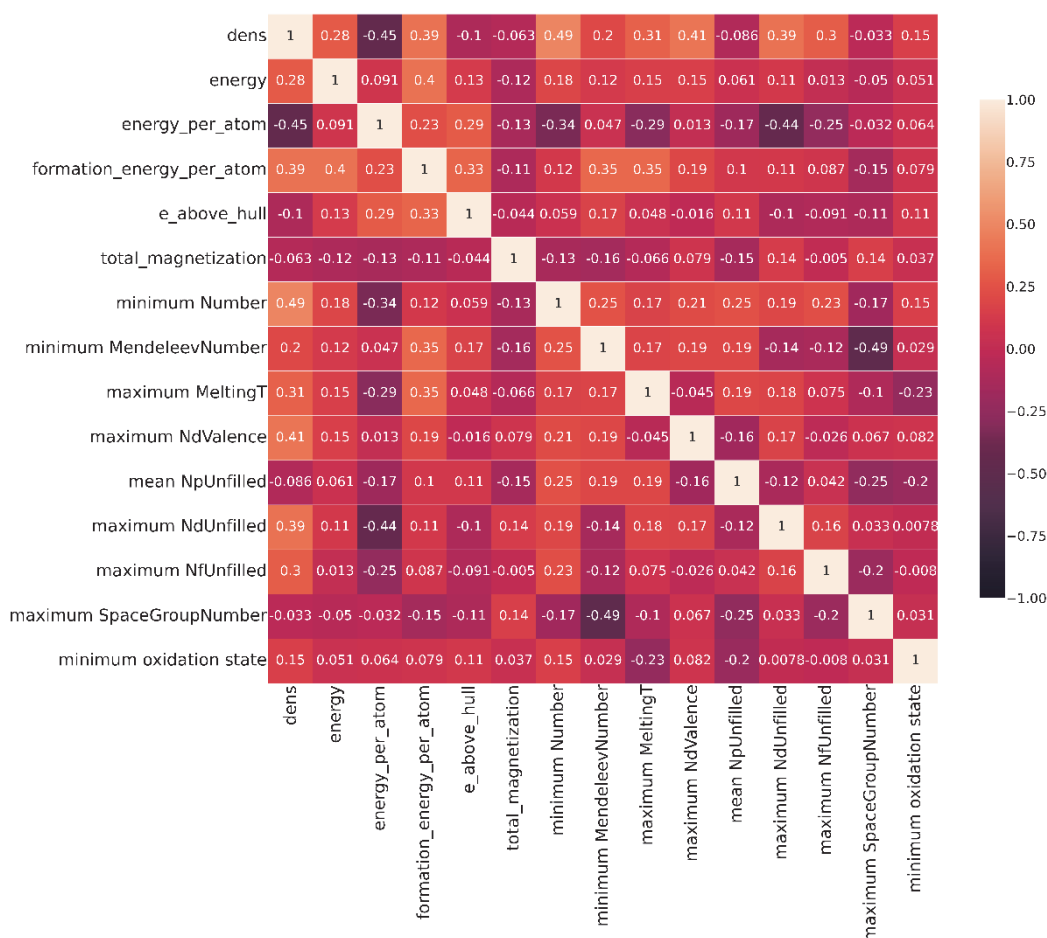


图 3 含硼材料的 15 种特征的皮尔森相关系数热图

Figure 3 Heatmap of Pearson correlation coefficients for 15 features of boron-containing materials

表2 使用 145 种材料特征和 15 种材料特征作为训练样本的九种回归模型的测试集结果

Table 2 Test set results for nine regression models using 145 material features and 15 material features as training samples

Models	145 features			15 features		
	MAE	MSE	R^2	MAE	MSE	R^2
Lasso	1.24	2.38	0.31	1.38	2.65	0.20
Bayesian Ridge	0.87	1.48	0.57	1.10	1.96	0.41
Elastic Net	1.17	2.18	0.37	1.34	2.57	0.22
Support Vector Regression	1.38	2.92	0.15	1.50	3.25	0.02
Decision Tree Regression	0.67	1.22	0.65	0.65	1.02	0.69
Gradient Boosting Regression	0.58	0.66	0.81	0.61	0.70	0.79
AdaBoost Regression	0.91	1.21	0.65	0.89	1.22	0.63
Random Forest Regression	0.51	0.58	0.83	0.51	0.57	0.84
Extra Tree Regression	0.67	1.20	0.65	0.63	1.07	0.68

的评估均存在差异. 其中, Lasso、Elastic Net 和 Linear Regression 是经典的线性模型, 支持向量机回归(SVR)是基于核映射空间变化的模型, 这些模型的决定系数 R^2 值都相对较低.

其中线性回归模型通过 10 倍交叉验证后, 预测的能隙会存在较大的离群值, 影响模型整体的预测性能, 大大降低了模型的泛化能力(支持信息图 S2). 模型的预测性能如图 4 所示, 随机森林回归是经典的非线性模型, 训练得到的随机森林模型在测试集中的预测能力最佳(图 4c 所示), 得到了最高的 R^2 值为 0.83. 其他的非线性模型算法的 R^2 值在 0.6~0.81 之间(图 4a, 4b, 4d).

线性预测模型与非线性模型对测试集的预测性能的显著差异表明含硼化合物数据集集中的材料特征与目标变量能隙之间存在复杂的非线性关系. SHAP^[36]特征重要性的评估有助于了解影响含硼化合物能隙的关键材料特征属性, 并为定向设计含硼材料提供方向性的指导.

对线性模型预测结果的分析发现线性回归(LR)模型无法找到特征与能隙之间的关联关系, 表明线性回归模型不适用于材料能隙的预测. SHAP 分析表明该模型把熔点温度识别为与能隙相关的重要特征, 这与其他模型不一致(支持信息图 S3). 其他线性模型(Lasso, Bayesian Ridge, Elastic Net)也均具有较高的 MAE 和 MSE, 表明这些模型具有较差的预测含硼材料能隙的能力. 对这些模型的 SHAP 分析(支持信息图 S4~S6)将能量和共价半径识别为与能隙较为相关的量, 但模型识别的特征重要性的顺序不同. 基于支持向量机算法的 SVR 模型(支持信息图 S7)也未能捕捉到特征与能隙之间的非线性的关系, 导致预测分数 R^2 仅为 0.15. 线性回归和支持向量机的低预测能力说明了两个模型对数据中的噪声较为敏感.

对随机森林模型进行特征重要性分析发现, 总磁化强度“total magnetization”和“e above hull”是两个与材料能隙最相关的量(图 5). 这两个物理量都与材料的电子结构性质密切相关. 前者定义为材料永久或诱发的偶极磁矩的矢量场, 用于描述材料的磁性强弱程度; 后者是衡量材料热力学稳定性的一个能量参数, 是吉布斯自由能中焓项的反映, 用于评估合成该构型材料的难易程度. 其他的集成学习算法也与随机森林算法相近, 均捕捉到了总磁化强度和“e above hull”是预测含硼材料能隙较为关键的因素. 这些结果表明非线性模型对含硼

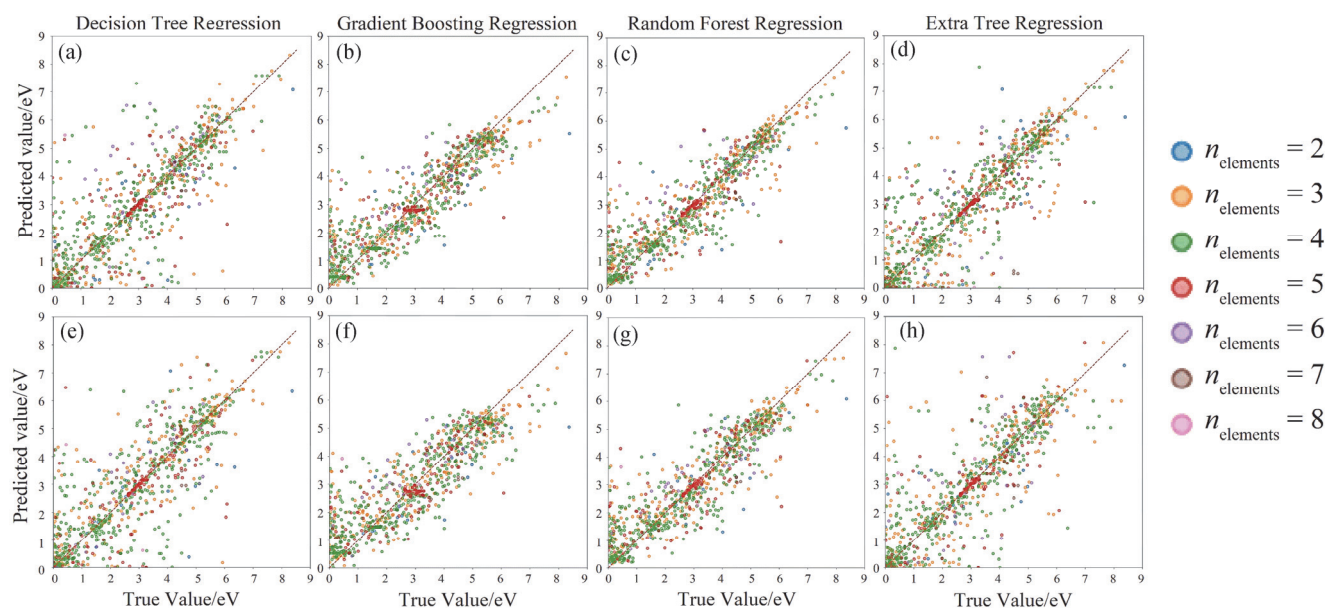


图 4 四种集成学习模型的带隙预测性能: (a~d) 145 种特征, (e~h) 15 种特征

Figure 4 The band gap prediction performance of the four ensemble learning models: (a~d) 145 features, (e~h) 15 features

材料能隙的预测性能远高于线性模型和支持向量机模型.

2.3.2 筛选特征分析结果

为了考察冗余特征对模型预测性能的影响, 对特征描述符进行相关性分析. 经过皮尔森相关性分析后筛选出 15 种材料特征, 使用上述十种回归模型基于训练集进行模型训练, 并应用得到的模型对测试集中的数据进行材料能隙的预测, 预测结果的决定系数 R^2 如表 3 所示. 结果表明线性回归模型的性能提升明显, 这是由于特征数量的减少使得非线性相关的复杂程度降低导致的(支持信息图 S12). Lasso 的性能却降低, 这是由于算法中的 L1 正则化会惩罚掉非重要影响的特征. SHAP 分析发现 Lasso 模型(支持信息图 S14)仅认为 6 种特征对能隙预测具有相关性, 这也证实了 L1 正则化惩罚掉了其

余特征其中包括“e above hull”项. Elastic Net 模型结合了 L1 和 L2 正则化来选择重要的特征, 与 Lasso 一样预测的能力略有损失(支持信息图 S16). 支持向量机模型仍无法准确预测含硼材料的能隙(支持信息图 S12, S17). 梯度增强树(图 4f 所示)和自适应增强(支持信息图 S12)中预测性能有略微下降, 通过特征分析, 发现在新模型预测结果中特征“e above hull”的相关程度降低. 决策树(图 4e 所示)、随机森林(图 4g 所示)和极端树(图 4h 所示)回归的性能有所提升.

这些结果表明随机森林模型仍然是预测能力最佳的回归模型(网格优化参数如表 4 所示). 其中 MSE 的值有所降低, 决定系数 R^2 有所提升, 并且在随机森林模型中, 预测的准确性与数据集的数量成正比(表 3).

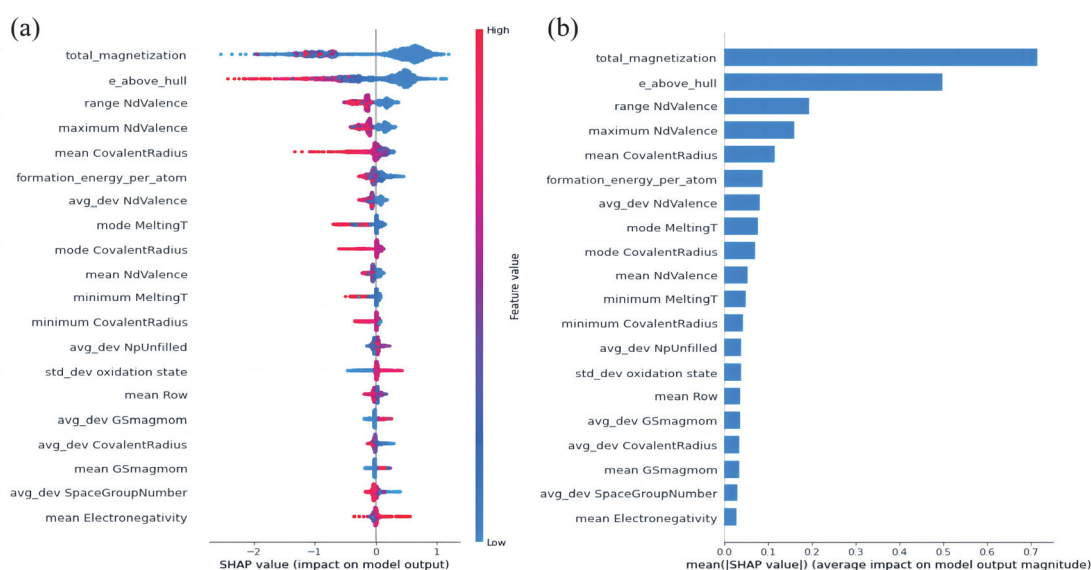


图 5 随机森林模型(145 个特征)的 SHAP 特征重要性分析: SHAP 特征的总体特征图(a), SHAP 特征的重要性值(b)

Figure 5 SHAP feature importance analysis of the Random Forest model (145 features) for predicting band gap, with the overall feature map of SHAP features (a) and the importance values of SHAP features (b)

表 3 十种模型中对于多元含硼材料的能隙预测的 R^2 值(使用 15 种材料特征作为训练样本的十种回归模型的测试集结果)

Table 3 R^2 values for ten models (test set results of ten regression models using 15 material features as training samples)

Models	Number of elemental species						
	2	3	4	5	6	7	8
Linear Regression	0.42	0.44	0.46	0.26	0.33	-0.01	-2.15
Lasso	-0.07	0.16	0.23	-0.06	0.05	-1.14	-1.2
Bayesian Ridge	0.42	0.45	0.46	0.25	0.33	0	-2.09
Elastic Net	-0.11	0.2	0.28	-0.1	0.1	-1.52	-0.83
Support Vector Regression	-0.14	0.02	0.01	-0.18	-0.25	-0.72	-1.43
Decision Tree Regression	0.55	0.76	0.78	0.47	0.5	0.6	-4.75
Gradient Boosting Regression	0.64	0.81	0.81	0.73	0.57	0.07	-2.48
AdaBoost Regression	0.45	0.62	0.66	0.53	0.43	-0.65	-2.65
Random Forest Regression	0.72	0.84	0.86	0.73	0.68	0.37	-1.9
Extra Tree Regression	0.25	0.77	0.63	0.42	-0.29	0.84	-0.46

表4 随机森林回归模型的训练参数

Table 4 Training parameters of Random Forest Regression Model

Parameters
bootstrap: False, max_depth: 30, max_features: log2, min_samples_leaf: 1, min_samples_split: 2, n_estimators: 300

对新的随机森林模型进行 SHAP 和特征重要性分析发现与之前的全特征样本训练得到的结果类似(如图 6 所示), 总磁化强度和“e above hull”是两个与材料能隙最相关的量, 并且总磁化强度的重要程度是第三重要特征(填满的 d 价轨道数最大值)的两倍. 从 SHAP 总体特征图可知总磁化强度越小, 材料的能隙就可能越大. 磁化强度描述物质被磁化的程度, 当施加外磁场于物质时, 物质的内部会被磁化, 会出现很多微小的磁偶极子. 物质被磁化会产生磁偶极矩, 磁偶极矩的一种起源是原子内部的电子在外磁场作用下产生的, 因此磁化强度跟晶体中的自由电子存在相关性. 在能带理论中紧束缚近似下将在一个原子附近的电子看作受该原子势场的作用为主, 其他原子势场的作用看作微扰, 从而可以得到电子的原子能级和晶体中能带之间的相互关系. 而机器学习方法挖掘出它们之间隐含的潜在关系. 对于具有相同化学组成的结构, 相对于最稳定结构构型, 具有较高相对能量(高“e above hull”)的构型处于亚稳态, 电子处于较高的能级, 该体系的能隙较小, 导致较低的稳定性, 即材料的“e above hull”的数值越大其能隙值越小. 两种特征与含硼材料能隙呈现负相关的趋势, 对于新的材料结构, 可以方便地通过第一性原理计算得到. 总磁化强度与能隙之间的密度图如图 7 所示, 根据总磁化强度可以分为两个部分, 在高磁化强度区间, 含硼材料的能隙值偏低, 较为集中; 在低磁化强度区域, 能隙较为集

中在 4~6 eV 的高能隙区间.

3 结论

新材料的研究是推动社会发展的重要驱动力. 本文研究工作面向材料研究从传统的试错法向数据驱动的研究范式转变的需求, 通过特征选择、网格调优等一系列手段, 探索了 Linear Regression、Lasso、Bayesian Ridge、Elastic Net、Support Vector Regression、Decision Tree Regression、Gradient Boosting Regression、AdaBoost Regression、Random Forest Regressor 和 Extra Tree Regression 等十种重要的机器学习算法在含硼材料能隙预测中的应用, 其中随机森林算法获得了具有决定系数 $R^2=0.84$ 的集成学习能隙预测模型, 并发现了含硼材料的总磁化强度“total magnetization”特征对预测目标变量能隙值具有重要影响, 存在总磁化强度越小, 材料的能隙越大的趋势, 即总磁化强度与含硼材料能隙存在负相关的关系. 对定向设计特定能隙的含硼材料提供了理论指导. 随机森林这种集成学习模型具有较好的学习能力与稳定的预测性能, 可以应用到其它类型材料体系的能隙以及其它材料属性的预测过程中, 加速材料性能的设计与优化过程, 对新型功能材料的快速筛选与高性能预测具有重要的科学意义.

4 实验部分

4.1 特征选择

在传统的机器学习(ML)方法的基本组成部分中(描述每个材料的特征以及目标性质的数据集以及将特征映射到目标性质的 ML 算法), 考虑到特征对模型预测能力的影响, 本工作调用 Matminer^[37]库中的 Magpie 特

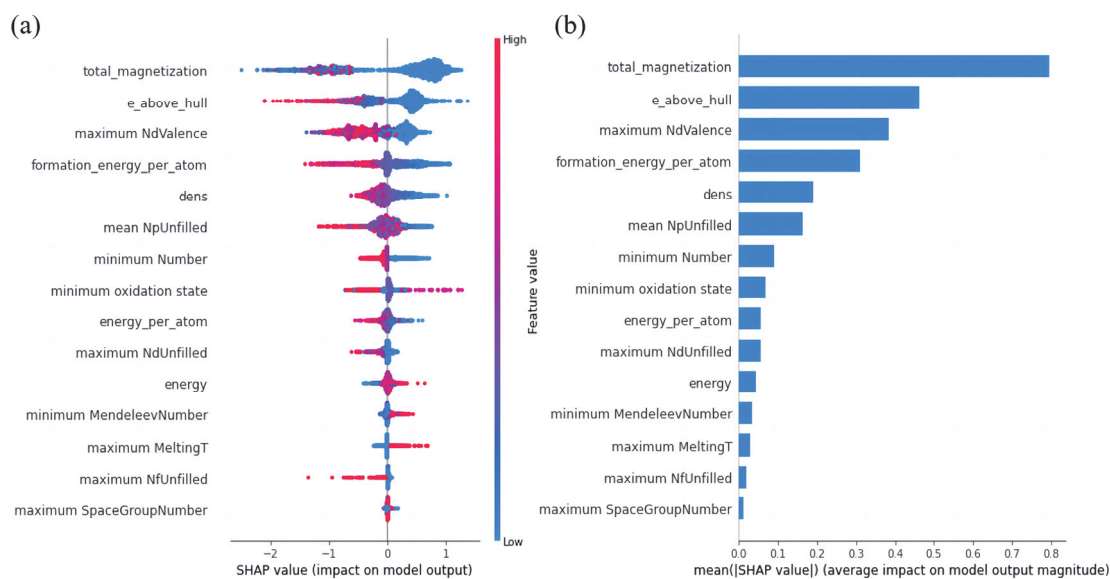


图6 随机森林模型(15个特征)的 SHAP 特征重要性分析: SHAP 特征的总体特征图(a), SHAP 特征的重要性值(b)

Figure 6 SHAP feature importance analysis of the Random Forest model (15 features) for predicting band gap, with the overall feature map of SHAP features (a) and the importance values of SHAP features (b)

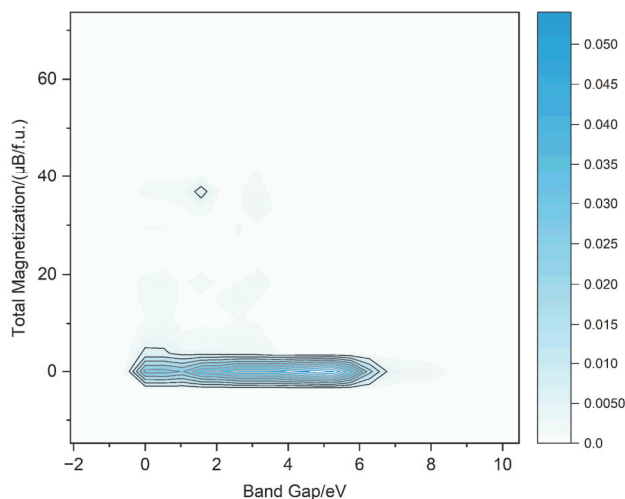


图7 总磁化强度和能隙的高斯核密度估计图

Figure 7 Contour map of the distribution of total magnetization as a function of band gap

征描述符来建立一个基于元素组成和晶体结构的综合特征集。通过添加 Magpie^[38](Materials Agnostic Platform for Informatics and Exploration)特征描述符(包括基础特征、组分特征和结构特征等共 132 个), 将材料数据集的特征数量扩充至 163 个。经筛选后保留 145 个特征用于回归模型训练(模型训练过程中采用的网格搜索参数见支持信息表 S1)。该特征集为每种材料创建一个独特的定量表示, 并可以在含硼材料的各种研究中重复使用, 而不受结构多样性的限制。

由于材料性质特征间可能存在的强相关性会造成特征的冗余, 影响训练模型的预测能力, 故需要进行各性质间的相关性分析, 过滤冗余特征, 这也可以降低学习模型的训练时间, 节约计算成本。本文使用皮尔森相关系数(Pearson correlation coefficient)表征各材料特征间的相关性, 两变量之间的皮尔森相关系数 $\rho_{X,Y}$ 定义为两个变量之间的协方差和标准差的商:

$$\rho_{X,Y} = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{E((X-E(X))(Y-E(Y)))}{\sigma_X \sigma_Y} \quad (1)$$

4.2 评价指标

构建一个能准确预测材料目标性质的预测模型是将机器学习应用到材料化学领域的目标之一, 通常以泛化能力来衡量模型的好坏, 即模型的预测能力的可通用性。在实际情况中, 通常通过测试误差来评价模型的泛化能力。回归分析中常用的误差主要是平均绝对误差(Mean Absolute Error, MAE)、均方误差(Mean Square Error, MSE)和决定系数(coefficient of determination, R^2)。

MAE 通过计算模型的绝对误差的平均值来衡量预测值误差(式(2))的实际情况。MSE 是真实值与预测值的差值的平方, 然后求和的平均, 一般用来检测模型的预测值和真实值之间的偏差(式(3))。决定系数 R^2 反映的是模型拟合数据的准确程度, 取值范围在 0 到 1 之间。其

值越接近 1, 表明模型对真实值的解释能力越强, 即模型对数据拟合的也较好(式(4))。对于不同的回归模型产生的预测结果, 本文采用上述评估标准来观察并测量每个模型在测试集上的预测精度, 比较不同模型之间性能差异。简而言之, MAE 与 MSE 越小且越接近于 0, R^2 越大且越接近于 1, 说明模型的性能越好, 预测准确率越高。以下是几种类型误差求解计算公式。为了避免出现随机误差, 本文的所有模型均使用了 10 折交叉检验。

$$MAE = \frac{1}{N} \sum_{i=1}^N |y_i - \hat{y}_i| \quad (2)$$

$$MSE = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2 \quad (3)$$

$$R^2 = 1 - \frac{\sum_{i=1}^N (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^N (y_i - \bar{y})^2} \quad (4)$$

References

- [1] Fujimori, M.; Nakata, T.; Nakayama, T.; Nishibori, E.; Kimura, K.; Takata, M.; Sakata, M. *Phys. Rev. Lett.* **1999**, *82*, 4452.
- [2] Shen, Y. F.; Xu, C.; Huang, M.; Wang, H. Y.; Cheng, L. J. *Prog. Chem.* **2016**, *28*, 1601 (in Chinese). (沈艳芳, 徐畅, 黄敏, 王海燕, 程龙玖, 化学进展, **2016**, *28*, 1601.)
- [3] Yang, X. Q.; Hu, Y.; Zhang, J. L.; Wang, Y. Q.; Pei, C. M.; Liu, F. *Acta Physica Sinica* **2014**, *63*, 048102 (in Chinese). (杨秀清, 胡亦, 张景路, 王艳秋, 裴春梅, 刘飞, 物理学报, **2014**, *63*, 048102.)
- [4] Rubio, A.; CORKILL, J. L.; Cohen, M. L. *Phys. Rev. B* **1994**, *49*, 5081.
- [5] Feng, B.; Zhang, J.; Zhong, Q.; Li, W.; Li, S.; Li, H.; Cheng, P.; Meng, S.; Chen, L.; Wu, K. *Nat. Chem.* **2016**, *8*, 563.
- [6] Li, P.; Zhang, X.; Wang, J.; Xue, Y.; Yao, Y.; Chai, S.; Zhou, B.; Wang, X.; Zheng, N.; Yao, J. *J. Am. Chem. Soc.* **2022**, *144*, 5930.
- [7] Hao, K. R.; Yan, Q. B.; Su, G. *Phys. Chem. Chem. Phys.* **2020**, *22*, 709.
- [8] Cheng, Z. S.; Zhang, X. M.; Zhang, H.; Liu, H. Y.; Yu, X.; Dai, X. F.; Liu, G. D.; Chen, G. F. *J. Phys. Chem. C* **2022**, *126*, 21542.
- [9] Zhan, C.; Zhang, P. F.; Dai, S.; Jiang, D. E. *ACS Energy Lett.* **2016**, *1*, 1241.
- [10] Qiu, B.; Lu, W. D.; Gao, X. Q.; Sheng, J.; Ji, M.; Wang, D. Q.; Lu, A. H. *J. Catal.* **2023**, *417*, 14.
- [11] Grant, J. T.; Carrero, C. A.; Goeltl, F.; Venegas, J.; Mueller, P.; Burt, S. P.; Specht, S. E.; McDermott, W. P.; Chieragato, A.; Hermans, I. *Science* **2016**, *354*, 1570.
- [12] Lu, X.; Li, K.; Xie, Y.; Qi, S.; Shen, Q.; Yu, J.; Huang, L.; Zheng, X. *J. Biomed. Mater. Res. Part A* **2019**, *107*, 12.
- [13] Liu, L.; Zhao, Z.; Yu, T.; Zhang, S.; Lin, J.; Yang, G. *J. Phys. Chem. C* **2018**, *122*, 6801.
- [14] Gao, Y.; Ma, Y. *J. Phys. Chem. C* **2019**, *123*, 23145.
- [15] Tian, X. X.; Xuan, X. Y.; Yu, M.; Mu, Y. W.; Lu, H. G.; Zhang, Z. H.; Li, S. D. *Nanoscale* **2019**, *11*, 11099.
- [16] Xu, L.; Wang, A.; Li, B.; Zhao, J.; Zeng, H.; Zhang, S. *J. Phys. Chem. Lett.* **2022**, *13*, 6455.
- [17] Yun, J.; Zhang, Y.; Xu, M.; Yan, J.; Zhao, W.; Zhang, Z. *J. Mater. Sci.* **2017**, *52*, 10294.
- [18] Xu, J.; Wan, Q.; Anpo, M.; Lin, S. *J. Phys. Chem. C* **2020**, *124*, 6624.
- [19] Chung, H. Y.; Weinberger, M. B.; Levine, J. B.; Cumberland, R. W.; Kavner, A.; Yang, J. M.; Tolbert, S. H.; Kaner, R. B. *Science* **2007**, *316*, 436.
- [20] Yao, Y.; Zhang, Z.; Jiao, L. *Energy Environ. Mater.* **2021**, *5*, 470.
- [21] Gabani, S.; Flachbart, K.; Siemensmeyer, K.; Mori, T. *J. Alloys Compd.* **2020**, *821*, 153201.
- [22] Yan, X.; Jin, Q.; Jiang, Y.; Yao, T.; Li, X.; Tao, A.; Gao, C.; Chen,

- C.; Ma, X.; Ye, H. *ACS Appl. Mater. Interfaces* **2022**, *14*, 36875.
- [23] Curtarolo, S.; Hart, G. L. W.; Nardelli, M. B.; Mingo, N.; Sanvito, S.; Levy, O. *Nat. Mater.* **2013**, *12*, 191.
- [24] Draxl, C.; Scheffler, M. *J. Phys. Mater.* **2019**, *2*, 036001.
- [25] Jain, A.; Ong, S. P.; Hautier, G.; Chen, W.; Richards, W. D.; Dacek, S.; Cholia, S.; Gunter, D.; Skinner, D.; Ceder, G.; Persson, K. A. *APL Mater.* **2013**, *1*, 011002.
- [26] Mehl, M. J.; Hicks, D.; Toher, C.; Levy, O.; Hanson, R. M.; Hart, G.; Curtarolo, S. *Comput. Mater. Sci.* **2017**, *136*, S1-S828.
- [27] De Pablo, J. J.; Jackson, N. E.; Webb, M. A.; Chen, L.-Q.; Moore, J. E.; Morgan, D.; Jacobs, R.; Pollock, T.; Schlom, D. G.; Toberer, E. S.; Analytis, J.; Dabo, I.; Delongchamp, D. M.; Fiete, G. A.; Grason, G. M.; Hautier, G.; Mo, Y.; Rajan, K.; Reed, E. J.; Rodriguez, E.; Stevanovic, V.; Suntivich, J.; Thornton, K.; Zhao, J.-C. *npj Comput. Mater.* **2019**, *5*, 41.
- [28] Hansen, K.; Montavon, G.; Biegler, F.; Fazli, S.; Rupp, M.; Scheffler, M.; Von Lilienfeld, O. A.; Tkatchenko, A.; Müller, K.-R. *J. Chem. Theory Comput.* **2013**, *9*, 3404.
- [29] Wei, X. H.; Zhou, C. B.; Shen, X. X.; Liu, Y. Y.; Tong, Q. C. *Journal of Jilin University (Engineering and Technology Edition)*, **2021**, *51*, 667 (in Chinese). (魏晓辉, 周长宝, 沈笑先, 刘圆圆, 童群超, 吉林大学学报(工学版), **2021**, *51*, 667.)
- [30] Wang, Y.; Lv, J.; Zhu, L.; Ma, Y. *Comput. Phys. Commun.* **2012**, *183*, 2063.
- [31] Huang, Y.; Yu, C.; Chen, W.; Liu, Y.; Li, C.; Niu, C.; Wang, F.; Jia, Y. *J. Mater. Chem. C* **2019**, *7*, 3238.
- [32] Dey, P.; Bible, J.; Datta, S.; Broderick, S.; Jasinski, J.; Sunkara, M.; Menon, M.; Rajan, K. *Comput. Mater. Sci.* **2014**, *83*, 185.
- [33] Xu, Y. L.; Wang, X. M.; Li, X.; Xi, L. L.; Ni, J. Y.; Zhu, W. H.; Zhang, W.; Yang, J. *Sci. Sin. Tech.* **2019**, *49*, 44 (in Chinese). (徐永林, 王香蒙, 李鑫, 席丽丽, 倪剑樾, 朱文浩, 张武, 杨炯, 中国科学:技术科学, **2019**, *49*, 44.)
- [34] Ong, S. P.; Richards, W. D.; Jain, A.; Hautier, G.; Kocher, M.; Cholia, S.; Gunter, D.; Chevrier, V. L.; Persson, K. A.; Ceder, G. *Comput. Mater. Sci.* **2013**, *68*, 314.
- [35] Hauke, J.; Kossowski, T. *Quaest. Geogr.* **2011**, *30*, 87.
- [36] Lundberg, S. M.; Erion, G.; Chen, H. *Nat. Mach. Intell.* **2020**, (2), 56.
- [37] Ward, L.; Dunn, A.; Faghaninia, A.; Zimmermann, N. E. R.; Bajaj, S.; Wang, Q.; Montoya, J.; Chen, J.; Bystrom, K.; Dylla, M.; Chard, K.; Asta, M.; Persson, K. A.; Snyder, G. J.; Foster, I.; Jain, A. *Comput. Mater. Sci.* **2018**, *152*, 60.
- [38] Ward, L.; Agrawal, A.; Choudhary, A.; Wolverton, C. *npj Comput. Mater.* **2016**, *2*, 16028.

(Cheng, B.)