

无监督学习探索高介电常数的 ABO_3 型钙钛矿材料刘润林¹, 李长姣², 王 建², 刘韩星¹, 沈忠慧¹

(1. 武汉理工大学材料科学与工程国际化示范学院, 武汉 430070;

2. 武汉理工大学材料科学与工程学院, 武汉 430070)

摘 要: 机器学习已成为新材料研发的重要变革性手段, 但材料数据样本量少、噪音高等特点为数据驱动的研发模式带来巨大挑战。本工作将无监督学习应用于挖掘高介电常数的钙钛矿材料。针对标签数据少的问题, 通过聚类学习的方法不断优化迭代来缩小搜索空间, 最终筛选出了 $BaHfO_3$ 和 $BiFeO_3$ 等 20 种具有高介电常数潜力的钙钛矿材料, 并通过降维分析等手段从元素种类、晶体结构和容忍因子等方面展开规律分析, 挖掘钙钛矿材料结构与介电常数之间的关联。该方法为解决材料性能数据标签的缺失提供了一种思路, 可应用于筛选和挖掘其他新型功能材料。

关键词: 介电常数; 机器学习; 钙钛矿材料; 电介质材料

中图分类号: TM934.33 文献标志码: A 文章编号: 0454-5648(2023)02-0367-06

网络出版时间: 2023-01-18

Discovering ABO_3 -Type Perovskite with High Dielectric Constant via Unsupervised LearningLIU Runlin¹, LI Changjiao², WANG Jian², LIU Hanxing¹, SHEN Zhonghui¹

(1. International School of Materials Science and Engineering, Wuhan University of Technology, Wuhan 430070, China;

2. School of Materials Science and Engineering, Wuhan University of Technology, Wuhan 430070, China)

Abstract: Machine learning has become an important transformative method to explore novel materials, but the small sample size and high noise of material data bring a great challenge to data-driven research and development. To address the challenge, unsupervised learning was applied to discover perovskite materials with a high dielectric constant. Twenty perovskite materials with a high dielectric constant (*i.e.*, $BaHfO_3$ and $BiFeO_3$) were screened out *via* iterative clustering. We performed dimensionality reduction analysis and descriptors analysis including elements, crystal structure and tolerance factors to find the underlying trend and the relationship between ABO_3 structure and dielectric constant. This method can provide an idea for solving the lack of material data labels, which can be also applied to screen other novel functional materials.

Keywords: dielectric constant; machine learning; perovskite materials; dielectrics

电介质材料因其优异的电气绝缘和电荷存储能力, 在微电子、微波、超级电容器和传感器行业中发挥着重要的作用。作为目前研究最多、应用最广的电介质材料体系之一, ABO_3 型钙钛矿材料表现出较高的介电常数(ϵ), 并被广泛应用于储能设备^[1]、太阳能电池设备^[2]、陶瓷电容器^[3]等领域, 如钛酸钡($BaTiO_3$)^[4]、钛酸锶($SrTiO_3$)^[5]等。伴随着电子元器

件向小型化和轻量化的发展趋势, 人们对电介质材料高介电常数的需求越来越高。因此, 探索高介电常数的新电介质材料成为该领域的重要研究方向之一。

随着大数据与人工智能等技术的飞速发展, 机器学习方法成为了数据驱动新材料研发的重要工具, 并在材料领域引起了广泛的关注。例如, 监督

收稿日期: 2022-09-30。 修订日期: 2022-11-24。

基金项目: 国家自然科学基金青年科学基金项目(52002300); 国家自然科学基金重大研究计划培育项目(92066103); 中国科协青年人才托举工程(2019QNRC001)。

第一作者: 刘润林(2001—), 男, 硕士研究生。

通信作者: 沈忠慧(1991—), 男, 博士, 副教授。

Received date: 2022-09-30. Revised date: 2022-11-24.

First author: LIU Runlin (2001—), male, Master candidate.

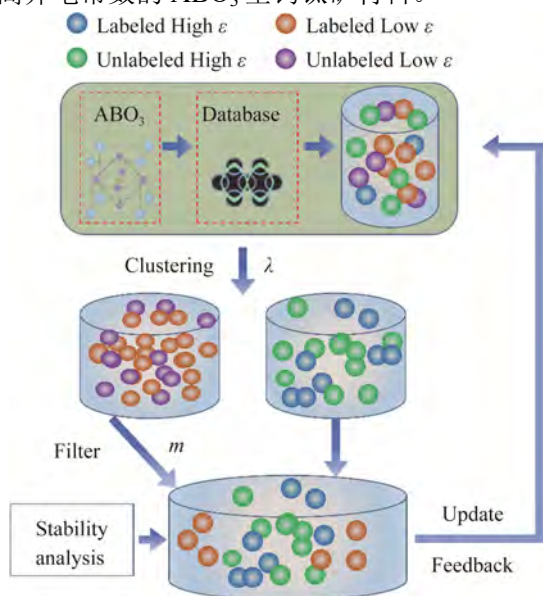
E-mail: rony_l@whut.edu.cn

Correspondent author: SHEN Zhonghui (1991—), male, Ph.D, Associate Professor.

E-mail: zhshen@whut.edu.cn

学习方法对钙钛矿材料带隙^[6]、Curie 温度^[7]、压电系数^[8]的预测。此外,机器学习在评估和预测钙钛矿材料的介电性能中也有着广泛的应用,如 Kim 等^[9]通过机器学习结合密度泛函微扰理论对 ABO₃ 型钙钛矿材料的介电常数进行了预测;Lin 等^[10]通过梯度增强回归(GBR)成功预测 431 组多晶钙钛矿材料介电常数的电子贡献值。然而,高介电常数的电介质材料筛选仍面临着性能数据标签少、搜索空间大等诸多挑战。作为一种从无标记属性数据中提取隐藏信息的技术,无监督学习方法在特征提取、模式识别和新材料发现等领域发挥着重要的作用,并被应用到了高电导率锂离子电容器的发现^[11]、半 Hessele 热电材料的搜索^[12]以及电子结构描述符提取^[13]等。

本工作以 Materials Project 材料数据库(MP, <https://materialsproject.org/>)中 2 398 种 ABO₃ 型钙钛矿结构化合物为原始数据,采用无监督学习的方法对具有高介电常数潜力的 ABO₃ 型钙钛矿材料进行筛选,其中含有介电常数性能标签的数据为 104 种。工作流程如图 1 所示,基于少量有标签的材料数据,在原始数据空间利用聚类方法将低介电常数的钙钛矿材料进行过滤后,通过多次更新迭代,挖掘出具有高介电常数的 ABO₃ 型钙钛矿材料。



ϵ —Dielectric constant; λ —Goodness of clustering; m —Clustering threshold

图 1 无监督学习迭代筛选具有高介电常数的钙钛矿材料流程图

Fig. 1 Workflow of iterative screening of materials with high dielectric constant by unsupervised learning

1 研究方法

1.1 数据集和特征工程

以 ABO₃ 型的钙钛矿材料为研究对象,借助

Pymatgen 包^[14]从 MP 数据库中收集到了 2 398 种钙钛矿材料原始数据,其中 104 种材料的介电常数被标记,相对 ϵ 介于 1~341 之间。结合钙钛矿材料的结构特点,本工作初步选取了 156 种描述符,分为以下 3 类:1) 元素信息:包括元素主族数、周期数、离子半径、电荷量、原子质量、电负性、容忍因子、八面体因子、空间电子结构等^[15],随后对每一种特征计算其 4 个统计量,即最大值、最小值、平均值以及标准差,以确保每个样本具有相同数量的特征;2) 结构信息:空间角(α 、 β 、 γ)、A-O、B-O、A-B 离子距离、体积(V)、密度(ρ)、熔点、空间群数等,其中磁化强度(M_N)也被考虑在其中;3) 能量特征:每个原子的能量(E)、形成能、带隙等。为了删除冗余信息、纠正数据中存在的错误,并保证数据一致性,首先对原始数据集(图 2a)中的异常值(如半径、体积为负值)与缺失值(如带隙、能量不存在)进行过滤处理,同时对所有描述符进行 Min-Max 归一化处理,随后利用 Pearson 相关系数以及距离相关系数去除相关性过高的描述符。当 2 个描述符的相关系数超过 80%,则只随机保留其中 1 个描述符,最终数据清洗后的维度为 1 030×87。图 2b 是特征选择后的 Pearson 相关系数热力图,表明特征之间没有明显的相关性。

1.2 模型选择

为找到合适的聚类算法,使用了 8 种不同的聚类方法并进行了对比,分别为 K 均值^[16]、综合层次聚类(Birch)^[17]、谱聚类(Spectral Clustering)^[18]、均值偏移(MeanShift)、亲和力聚类(Affinity Propagation)、双聚类(SpectralCoclustering)、聚合层次聚类(Agglomerative)、Gaussian 混合模型(GaussianMixture)以及小批量 K 均值(MinibatchKmeans)^[19]。聚类的好坏用聚类优劣度(λ)来评估,其定义为轮廓系数(S)与 Davies Bouldin 指数(D)^[20]的差值, λ 越接近于 1 说明聚类效果越好。轮廓系数越大代表簇内距离越近以及簇间距离越远; D 越小代表了簇内距离之和与簇间距离之和的比值越小。 λ 结合了 2 种系数各自的优势,能够避免聚类算法进入局部最优。计算公式如下:

$$\begin{cases} S = \sum_{i=1}^M \frac{b_i - a_i}{\max\{b_i, a_i\}}, \\ D = \frac{1}{N} \sum_{i=1}^N \max \left(\frac{(\bar{S}_i + \bar{S}_j)}{\|p_i - p_j\|} \right) \\ \lambda = S - D, \end{cases} \quad (1)$$

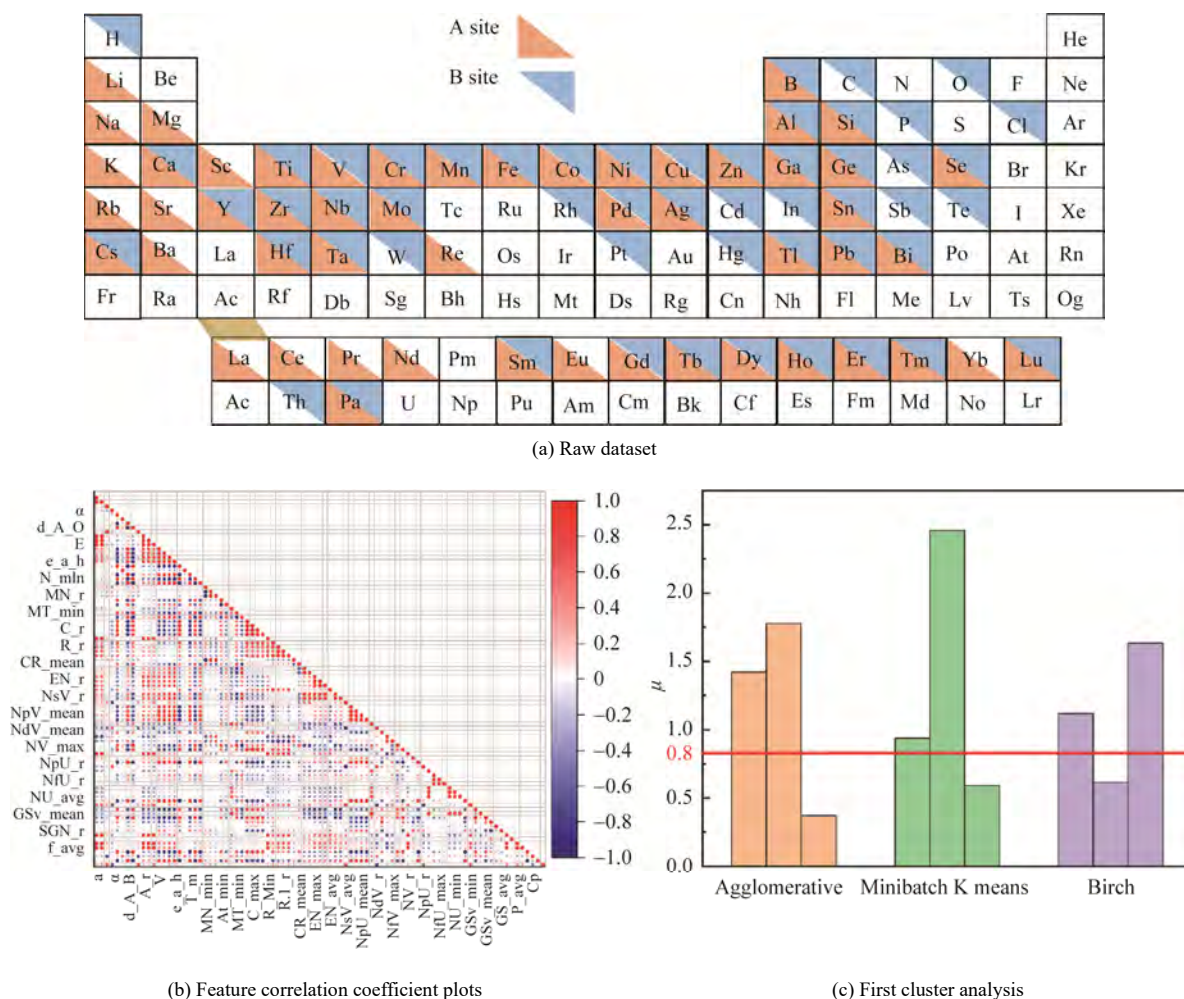


图 2 原始数据集以及数据清洗
Fig. 2 Raw dataset and data cleaning

其中: b_i 为样本 i 到同簇其他样本的平均距离; a_i 为样本 i 到其他簇的所有样本的平均距离; M 代表样本总数; N 代表簇数; $\bar{S}_i$ 代表簇内数据到簇内质心的平均距离; p_i 代表第 i 个簇类的质心。

为寻找高介电常数簇类, 需将有标签样本进行划分, 设定介电常数高于 30 的样本为高介电材料, 其余为中低介电材料。在每次聚类后, 选择 λ 较高的聚类方法进行分析来过滤材料。根据不同聚类结果, 定义

$$\mu(i) = \frac{\omega_h(i) - \omega_k(i)}{\omega_a(i)} \quad (2)$$

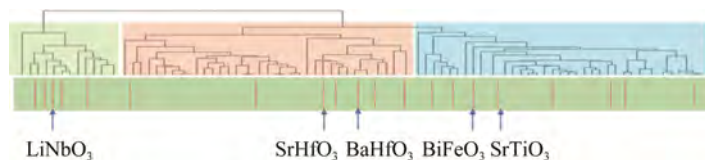
其中: $\omega_h(i)$ 表示 i 类已知标签的高介电钙钛矿在所有高介电钙钛矿的占比; $\omega_k(i)$ 表示第 i 类已知标签的钙钛矿在所有已知钙钛矿的占比; $\omega_a(i)$ 表示第 i 类钙钛矿在所有钙钛矿中的占比; $\mu(i)$ 为一个比例系数, 其越高则说明第 i 类里的高介电钙钛矿材料越多, 即为高介电簇类。设定聚类阈值 m , 当 $\mu(i)$ 超

过 m 则认为第 i 簇类为高介电簇类。通过对 m 进行分析, 发现当其在 0.75~0.90 附近进行扰动时, 数据变化的总体趋势一致, 模型的结果不会因此受到较大的影响, 具有一定的鲁棒性。然而, 当 m 继续减小到 0.70, 会导致聚类结果数据量增大, 难以收敛; 当 m 继续增长到 1.00, 会导致聚类结果收敛较快, 可能过滤掉一些有价值的数据, 因此本工作将阈值 m 设置为 0.80。

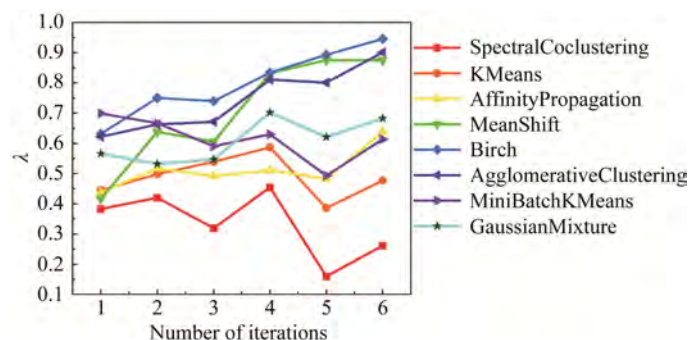
以下是聚类筛选的过程: 首先利用聚类算法对钙钛矿材料空间进行聚类, 选出 λ 最大的 3 种或 4 种聚类方法, 因为此时聚类的效果最优。随后对这几组聚类方法进行分析, 并根据 m 值判断低介电簇类, 过滤掉 2 次都被判定为低介电簇类的数据, 再重新聚类分析。循环迭代 6 次, 从而找出高介电常数簇类。例如, 第 1 次聚类结果如图 2c 所示, 认为 Agglomerative、MinibatchMeans 的第 3 簇类和 Birch 第 2 簇类为低介电簇类, 将 312 组满足以上条

件样本进行过滤。循环迭代 5 次后, 聚类趋于稳定。随后再次筛选出能形成化合物, 且与稳定相形成能的差值不超过 0.15 eV 的样本。最终, 从最后一次聚类结果过滤出 20 种具有高介电常数潜力的钙钛矿材料, 其中部分材料的高介电性能已经在实验中得到验证^[21–23], 例如 LiNbO_3 、 BiFeO_3 等。图 3a 是聚

类结果在原始聚类系统树图中的分布图, 红线代表 20 种具有高介电常数潜力的钙钛矿材料, 实验中已经发现的高介电材料已标出。图 3b 为聚类优劣度 λ 随迭代次数的变化曲线, 发现前 3 种算法聚类优劣度接近甚至超过 90%, 说明筛选出的数据逐渐聚集, 同簇内的材料与有标签高介电性能材料具备相似的特征。



(a) Distribution of the clustering results in the original dendrogram



(b) Curve of λ change with the number of iterations

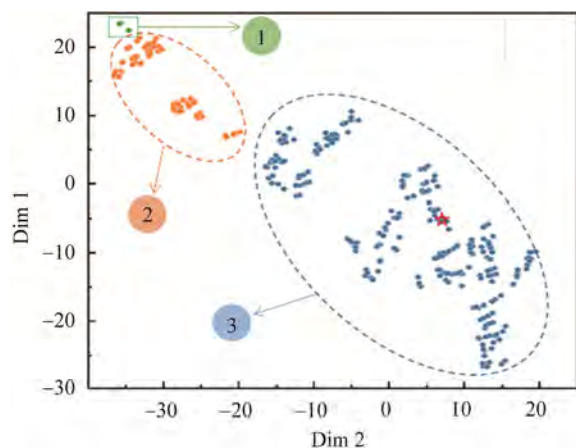
图 3 聚类分析迭代结果

Fig. 3 Results of iterative clustering analysis

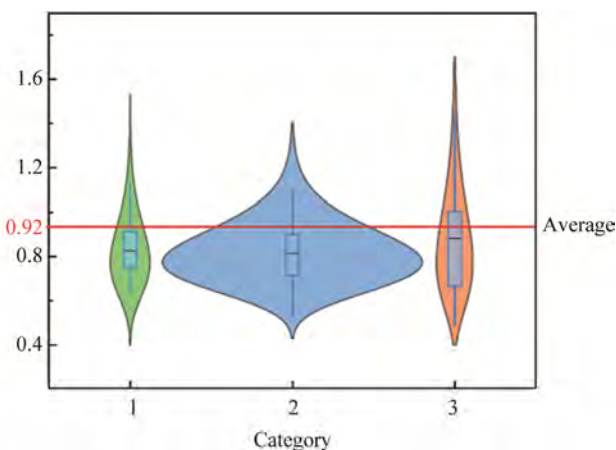
2 结果与讨论

对最后一次聚类未过滤的样本进行 t-SNE 流形学习降维并可视化后发现, 数据分布呈现一定的规

律。如图 4a 所示, 样本被明显分为 3 类, 分别代表着 Ni 基钙钛矿材料、Hf 基钙钛矿与 Ti、Bi、La 基钙钛矿材料, 表明钙钛矿材料的介电常数与组成元素种类存在一定的关联。



(a) Spatial distribution after the data dimension reduction



(b) Comparison of the tolerance factor violin diagram of the three types of materials

Note: Red line represents the average tolerance factor of the original data.

图 4 聚类结果的元素、容忍因子(t)分布

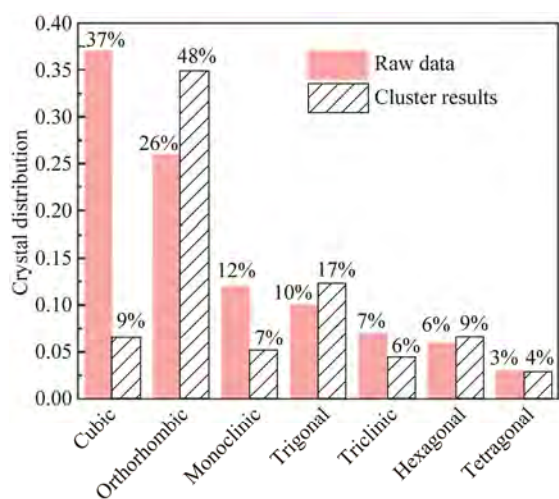
Fig. 4 Elemental, tolerance factor (t) distribution of the clustering results

图 4b 表示不同簇类钙钛矿材料容忍因子 t 的分布, 结果表明大多数钙钛矿材料的容忍因子位于

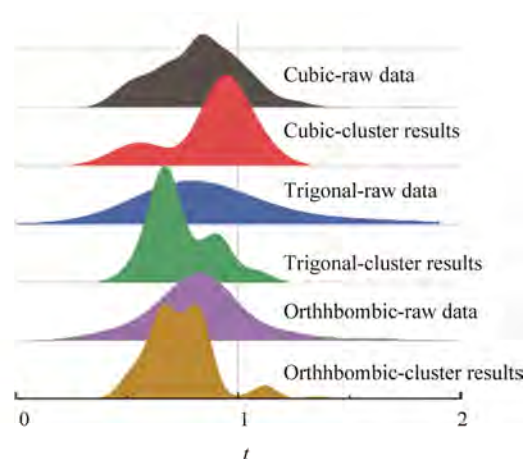
0.7~1.1 之间, 符合形成稳定钙钛矿结构的条件^[24]。与此同时, 筛选出的样本的容忍因子相较于原始数

据的平均容忍因子 0.92 有一定的下降趋势。值得注意的是, 由于数据或者实验误差, 尽管一些材料的容忍因子值不位于其经验稳定区间, 但仍能制备出稳定的钙钛矿结构并呈现较高的介电常数, 如 ScBiO₃ 与 TmBiO₃, 其容忍因子均仅在 0.7 附近并表现出较高的介电常数。图 5a 为原始数据空间的晶体结构和聚类后的晶体结构占比分布图。从图 5a 可以看出, 聚类后获得的具有高介电常数潜力的钙钛矿

材料中, 其立方相比比例相对原始数据下降明显, 而正交相和三角相的比例大幅提高。为了更好地理解晶体结构与介电常数之间的关联, 本工作尝试从结构描述符容忍因子 t 的角度来进行统计分析。图 5b 表明在逐步聚类后, 立方相容忍因子呈上升趋势, 而三角相和正交相的容忍因子呈整体下降趋势, 这恰与晶系分布的变化相对应, 说明介电常数与晶体结构间存在一定的联系^[6]。



(a) Distribution of crystal system before and after clustering



(b) Tolerance factor distribution of cubic, triangular, orthogonal phases

图 5 聚类前后的晶体结构分布

Fig. 5 Crystal structure distribution before and after clustering

3 结论

基于材料标签稀疏的数据特点, 本工作采用迭代的聚类学习方法对 ABO₃ 型钙钛矿材料进行高通量筛选, 最终预测了 BaHfO₃ 和 BiFeO₃ 等 20 种具有高介电常数潜力的 ABO₃ 型钙钛矿材料, 部分体系已在实验中获得验证。进而, 通过将聚类结果与原始数据进行统计分析, 分别从元素种类、容忍因子、晶系分布等特征分析钙钛矿材料结构-介电常数之间的关系, 发现含 Ni、Hf、Ti、Bi、La 元素的钙钛矿材料更具有高介电常数的潜力, 同时发现在具有高介电常数潜力的钙钛矿材料中, 正交相和三角相的比例大幅提高, 容忍因子也有所降低。

参考文献:

- [1] XU B, ÍÑIGUEZ J, BELLAICHE L. Designing lead-free antiferroelectrics for energy storage[J]. Nat Commun, 2017, 8(1): 15682.
- [2] LI W, ZHENG J, HU B, et al. High-performance solar flow battery powered by a perovskite/silicon tandem solar cell[J]. Nat Mater, 2020, 19(12): 1326–1331.
- [3] LI J, SHEN Z, CHEN X, et al. Grain-orientation-engineered multilayer ceramic capacitors for energy storage applications[J]. Nat Mater, 2020,

- 19(9): 999–1005.
- [4] ACOSTA M, NOVAK N, ROJAS V, et al. BaTiO₃-based piezoelectrics: Fundamentals, current status, and perspectives[J]. Appl Phys Rev, 2017, 4: 041305.
- [5] URBAN J J, YUN W S, GU Q, et al. Synthesis of single-crystalline perovskite nanorods composed of barium titanate and strontium titanate[J]. J Am Chem Soc, 2002, 124(7): 1186–1187.
- [6] LI C, HAO H, XU B, et al. A progressive learning method for predicting the band gap of ABO₃ perovskites using an instrumental variable[J]. J Mater Chem C, 2020, 8(9): 3127–3136.
- [7] ZHAI X, CHEN M, LU W. Accelerated search for perovskite materials with higher Curie temperature based on the machine learning methods[J]. Comput Mater Sci, 2018, 151: 41–48.
- [8] HU J, SONG Y. Piezoelectric modulus prediction using machine learning and graph neural networks[J]. Chem Phys Lett, 2022, 791: 139359.
- [9] KIM E, KIM J, MIN K. Prediction of dielectric constants of ABO₃-type perovskites using machine learning and first-principles calculations[J]. Phys Chem Chem Phys, 2022, 24(11): 7050–7059.
- [10] LIN X, LI C, HAO H, et al. Accelerated search for ABO₃-type the electronic contribution of polycrystalline dielectric constants by machine learning[J]. Comput Mater Sci, 2021, 193: 110404.
- [11] ZHANG Y, HE X, CHEN Z, et al. Unsupervised discovery of solid-state lithium ion conductors[J]. Nat Commun, 2019, 10(1): 5260.
- [12] JIA X, DENG Y, BAO X, et al. Unsupervised machine learning for discovery of promising half-Heusler thermoelectric materials[J]. npj Comput Mater, 2022, 8(1): 34.
- [13] ESTERHUIZEN J A, GOLDSMITH B R, LINIC S. Uncovering

- electronic and geometric descriptors of chemical activity for metal alloys and oxides using unsupervised machine learning[J]. Chem Catal, 2021, 1(4): 923–940.
- [14] ONG S P, RICHARDS W D, JAIN A, et al. Python Materials Genomics (pymatgen): A robust, open-source python library for materials analysis[J]. Comput Mater Sci, 2013, 68: 314–319.
- [15] Tilley J D. The ABX₃ Perovskite Structure[M]. John Wiley & Sons, Ltd, 2016: 1–41.
- [16] Macqueen J. Some methods for classification and analysis of multivariate observations[J]. Proc Symp Math Statist and Probability, 1967, 5th: 1.
- [17] 张蓉, 钟艳. 基于 BIRCH 算法的模糊集数据库挖掘算法[J]. 科技通报, 2014, 30(4): 47–49.
ZHANG Rong, ZHONG Yan. Bull Sci Technol (in Chinese), 2014, 30(4): 47–49.
- [18] 谢娟英, 丁丽娟, 王明钊. 基于谱聚类的无监督特征选择算法[J]. 软件学报, 2020, 31(4): 1009–1024.
XIE Juanying, DING Lijuan, WANG Mingzhao. J Software (in Chinese), 2020, 31(4): 1009–1024
- [19] 孙吉贵, 刘杰, 赵连宇. 聚类算法研究[J]. 软件学报, 2008(1): 48–61.
SUN Jigui, LIU Jie, ZHAO Lianyu. J Software (in Chinese), 2008(1): 48–61.
- [20] WANG J-M, ZHANG W, HUA T, et al. Unsupervised learning of topological phase transitions using the Calinski-Harabaz index[J]. Phys. Rev, 2021, 3(1): 013074.
- [21] KIM Y M, PARK C, HA T, et al. High-k perovskite gate oxide BaHfO₃[J]. APL Mater, 2017, 5(1): 016104.
- [22] 沈宗洋, 李敬峰. (Na, K)NbO₃ 基无铅压电陶瓷的研究进展[J]. 硅酸盐学报, 2010, 38(3): 510–520.
SHEN Zongyang, LI Jingfeng. J Chin Soc Ceram, 2010, 38(3): 510–520.
- [23] 唐明阳. BiFeO₃ 基无铅陶瓷介电及储能性能研究[D]. 西南大学, 2021.
TANG Mingyang. The research of dielectric and energy storage properties of bismuth ferrite based lead-free ceramics(dissatation, in Chinese), Chongqing: Southwest University, 2021.
- [24] 贾琳蔚, 李晓云, 丘泰, 等. 微波介质陶瓷分类及各体系研究进展[J]. 材料导报, 2008, 22(4): 10–13, 29.
JIA Linwei, LI Xiaoyun, QIU Tai, et al. Mater Rep (in Chinese), 2008, 22(4): 10–13, 29.

※ ※ ※ ※ ※ ※ ※ ※ ※ ※ ※ ※



沈忠慧 武汉理工大学材料科学与工程国际化示范学院副教授, 博士生导师。任中国硅酸盐学会青年工作委员会委员等, 入选第五届中国科协青年人才托举工程, 入选湖北省高层次人才计划(青年项目), 主持国家自然科学基金青年基金项目 and 重大研究计划培育项目等, 并作为研究骨干参与国家 973 计划、国家重点研发专项和军科委基础加强计划重点研究项目等多个科研项目。近五年在包括 Nature Materials, Nature Communications, Advanced Materials, Advanced Energy Materials, Materials Today, npj Computational Materials 等期刊发表 SCI 论文 40 余篇, 作邀请报告二十余次。研究工作关注于非均质材料结构-性能的关联, 通过智能手段来解析、预测和设计复合材料中电、力、热等性能的集成与耦合, 开展功能复合材料的结构生成-性能预测-材料设计的一体化工作, 为实验上提供了前瞻性的理论指导。



刘润林 武汉理工大学材料学硕士研究生, 研究方向为铁电和介电材料的智能化研发。所在课题组研究项目包括: 有机/无机复合电介质材料的跨尺度模拟; 数据驱动的储能电介质材料的智能研发; 极端冲击下储能电介质材料及电容器的结构演变和动态性能模拟研究。