

# 铅基钙钛矿铁电晶体高临界转变温度的机器学习研究\*

杨自欣 高章然 孙晓帆 蔡宏灵 张凤鸣 吴小山†

(南京大学物理学院, 固体微结构国家实验室, 南京 210093)

(2019年6月18日收到; 2019年8月20日收到修改稿)

铁电材料由铁电相转化为顺电相的临界温度被称为居里温度, 是铁电材料的一个关键指标. 本文使用固溶体组成元素的基本物理性质等特征对不同组分和配比的铅基钙钛矿铁电固溶体进行了统一的描述, 采用岭回归、支持向量回归、极端随机森林回归等机器学习方法对铅基钙钛矿铁电固溶体的居里温度进行了学习. 使用交叉验证的方法对学习效果进行验证, 得到上述机器学习方法对材料居里温度的预测值与实验值之间的平均误差分别为 14.4, 14.7, 16.1 K, 集成三种回归方法优化的模型在交叉验证中测得的平均误差为 13.9 K. 在此基础上对超过 20 万种铅基钙钛矿的居里温度进行了预测, 给出了两种可能具有高居里温度的铁电材料.

**关键词:** 机器学习, 铁电, 居里温度, 钙钛矿

**PACS:** 05.70.Fh, 02.60.Pn, 43.38.Fx

**DOI:** 10.7498/aps.68.20190942

## 1 引言

近年来, 具有优秀压电和热电性质的弛豫铁电体受到了广泛关注. 其中以固溶体  $\text{Pb}(\text{Mg}_{1/3}\text{Nb}_{2/3})\text{O}_3$ - $\text{PbTiO}_3$  (PMN-PT) 为代表的铅基钙钛矿铁电体由于其高压电系数而备受关注. 然而 PMN-PT 的居里温度较低 ( $\sim 130$ — $170$  °C), 限制了它在高温场景下的应用. 因此, 为了解决这类铁电体高温应用的难题, 研究者在实验中尝试将  $\text{Pb}(\text{In}_{1/2}\text{Nb}_{1/2})\text{O}_3$ ,  $\text{Pb}(\text{Lu}_{1/2}\text{Nb}_{1/2})\text{O}_3$ ,  $\text{Pb}(\text{Yb}_{1/2}\text{Nb}_{1/2})\text{O}_3$  (PIN, PLN, PYN) 等不同的简单钙钛矿与  $\text{PbTiO}_3$  结合在一起形成二元或三元的铅基钙钛矿固溶体来提高铅基钙钛矿铁电体的热稳定性.

Liu 等<sup>[1]</sup> 研究者制备了一系列 PLN-PT 二元固溶体材料, 其测得的居里温度范围为 145—365 °C. Wu 等<sup>[2]</sup> 研究者制备了一系列 PIN-PMN-PT 三元

固溶体, 测得的居里温度范围为 188—231 °C.

在实验上类似的尝试还有很多. 然而将十几种简单铅基钙钛矿与  $\text{PbTiO}_3$  组合成二元或三元固溶体拥有超过 100 种可能的组合, 再考虑到其中各组分比例的变化, 可以看出该类体系的搜索空间是巨大的, 通过实验的方法穷尽所有可能的组合来寻找高温铁电体是非常困难的. 在实验方法之外, 第一性原理计算的方法可以对材料的诸多电学性质进行模拟计算<sup>[3-5]</sup>, 然而目前该方法尚不能用来计算材料的居里温度.

近年来, 机器学习方法在物理、材料领域发挥了越来越重要的作用. 在机器学习方法的指引下, 人们可以快速地探索、挖掘出具有目标性质的材料<sup>[6-9]</sup>. 在本研究之前, 就有研究者曾使用机器学习方法探索钙钛矿铁电材料的居里温度. 例如 Balachandran 等<sup>[10]</sup> 针对  $(1-x)\text{Pb}(\text{B}_1, \text{B}_2)\text{O}_3$ - $x\text{PbTiO}_3$  体系从 117 种不同组分的数据中训练得到机

\* 国家自然科学基金 (批准号: 118742007) 资助的课题.

† 通信作者. E-mail: xswu@nju.edu.cn

机器学习模型,用以预测该体系的居里温度,他们使用交叉验证的方法进行模型评估,发现其最优模型的预测值与实验值之间的平均绝对误差 (mean average error, MAE) 为 30.2 K; Zhai 等<sup>[11]</sup> 从 47 种钙钛矿材料数据中训练得到了多个预测钙钛矿居里温度机器学习模型,其表现最优的模型为支持向量回归 (support vector regression, SVR) 模型,交叉验证得到预测模型的 MAE 为 21.2 K, 预测值与实验值之间相关系数为 0.85.

为了进一步得到适用于铅基钙钛矿铁电晶体的、具有更高预测准确度和更具有鲁棒性的居里温度预测模型,本文通过利用元素属性构建特征,从 205 种不同铅基钙钛矿固溶体的数据中学习,得到三种预测居里温度的机器学习模型. 并且通过模型集成的方法对习得的模型进行集成,得到预测误差更小的机器学习模型. 据此对大量复杂铅基钙钛矿的居里温度进行了预测,以供实验研究者参考.

## 2 方法

本文构建机器学习模型来预测复杂铅基钙钛矿的居里温度主要包含以下 3 个步骤:

1) 获取训练数据,构建用来预测居里温度的特征;

2) 通过交叉验证的方法选取合适的机器学习模型并进行其超参数的调整;将多个模型进行集成,构建最终的机器学习模型;

3) 使用机器学习模型学习训练数据,对未知的复杂铅基钙钛矿材料的居里温度进行预测.

本工作使用了 python 库 scikit-learn<sup>[12]</sup> 来进行模型的训练和评估.

### 2.1 数据集及构建特征

我们从十余篇已发表的实验论文中收集了 205 种不同的铅基钙钛矿的居里温度数据<sup>[1,2,13–28]</sup>. 其居里温度的范围为  $-120\text{--}490\text{ }^{\circ}\text{C}$ . 其中涉及到 18 种简单铅基钙钛矿和 45 种前者组合而成的复杂钙钛矿. 在这些铅基钙钛矿中,钙钛矿结构的  $A$  位离子都为 Pb 离子,  $X$  位离子都为 O 离子,其居里温度的差异是由  $B$  位离子的差异所导致的. 在我们收集的训练数据中既有结构较简单的  $\text{PbTiO}_3$ ,  $\text{PbZrO}_3$  等  $B$  位仅有一种离子的钙钛矿,也有 PMN-PFN-PZT 等  $B$  位有 5 种离子的复杂

钙钛矿,这保证了学习得到的模型对于不同复杂程度的钙钛矿固溶体均有良好的泛化能力.

由于本研究的体系之间的差异仅在于钙钛矿的  $B$  位离子的不同,因此针对铅基钙钛矿铁电体的  $B$  位离子构建了以下特征来描述不同的材料: 1) 材料中不同种类  $B$  位离子所对应的化学计量占比、不同离子种类数目; 2) 对于材料中的  $B$  位元素,引入其离子半径、电离能等 39 种的物理化学性质,该数据取自 Balachandran 等<sup>[10]</sup> 的研究,针对材料中各元素的属性及配比计算材料中  $B$  位元素化学计量占比加权平均得到的平均属性和各种属性的方差; 3) 采用第一性原理计算和机器学习方法得到的钙钛矿中  $B$  位离子偏移氧八面体中心的距离,取自文献<sup>[29, 30]</sup>,计算其加权平均及方差. 根据上述元素级别性质,对每种材料构建了 86 项特征用来描述不同材料之间的差异,并提供给机器学习模型用作建模的依据. 所有特征的数值都进行了缩放,将特征数值缩放到均值为 0、方差为 1 的范围内,这样可以确保所有的特征数值在相同的尺度上,从而避免由于特征尺度不同而对机器学习模型带来误导. 得到特征向量  $\mathbf{x}$ , 作为后续不同机器学习模型的输入. 本文中采用的具体特征指标及取值在补充文件 ([online](#)) 中给出.

### 2.2 机器学习模型

如果仅研究某种特定组分的钙钛矿固溶体不同构成比例的材料居里温度,可以通过开展一系列实验,制备出若干组分占比不同的材料,然后通过线性回归等简单方法进行拟合,并分析该组分的钙钛矿固溶体的居里温度随配比的变化. 这种不依赖于额外物理知识和物理规律的统计方法难以推广至各种不同组分的钙钛矿固溶体. 仅能分析特定组分固溶体的居里温度与配比的关系,而无法触类旁通得对其他组分的材料居里温度进行判断.

为了克服这一缺点,发挥机器的“智能”来解决这一难题,构建了上述特征来对铅基钙钛矿铁电晶体进行描述,使用机器学习的方法从其基本的物理性质出发挖掘不同材料之间深层的相似性,将不同种类固溶体的知识纳入统一的模型,达到“触类旁通”的效果. 基于这些基本物理性质和学到的不同固溶体居里温度的知识,对完全未知的材料的居里温度做出预言.

本文尝试了多种机器学习模型来对该问题进

行建模, 其中效果较好的模型为: SVR<sup>[31]</sup>、极端随机森林回归 (extremely randomized trees regression, ETR)<sup>[32]</sup>、岭回归 (kernel ridge regression, KRR) (使用径向基函数 (radial basis function, rbf) 核)<sup>[33]</sup>.

### 2.2.1 KRR (使用 rbf 核)<sup>[33,34]</sup>

KRR 是线性回归模型的拓展. 线性回归模型基于一系列的假设, 例如: 目标值与特征之间呈现线性关系; 特征之间应该相互独立等.

为了将线性回归推广至更一般的场景, 引入了 KRR: 通过在线性回归的损失函数中添加 L2 正则项来解决潜在的特征共线性的问题; 引入了核技巧来使得模型可以拟合目标变量与特征之间的非线性关系. 使用核技巧的 KRR 模型具有如下数学形式和损失函数<sup>[34]</sup>:

$$y = \mathbf{w}^T \Phi(\mathbf{x}), \quad (1)$$

$$C = \frac{1}{2} \sum_i (y_i - \mathbf{w}^T \Phi(\mathbf{x}_i))^2 + \frac{1}{2} \lambda \|\mathbf{w}\|^2, \quad (2)$$

其中  $\mathbf{x}$  为输入的特征向量;  $\lambda$  为正则项的权重, 是为了平衡结构风险与经验风险而预先设定的超参数;  $\Phi(\mathbf{x})$  为将特征向量映射到高维希尔伯特空间的映射函数, 其映射方法取决于核函数. 本文采用的核函数  $K$  为机器学习中较为常用的 rbf 核, 其对应了特征从原始空间到无穷维空间映射, 核函数中的  $\gamma$  为预先设定的超参数<sup>[35]</sup>.

$$K(\mathbf{x}_i, \mathbf{x}) = \exp\left(-\gamma \|\mathbf{x}_i - \mathbf{x}\|_2^2\right), \quad (3)$$

$$\phi(x) = e^{-\gamma x^2} \left[ 1, \sqrt{\frac{2\gamma}{1!}} x, \sqrt{\frac{(2\gamma)^2}{2!}} x^2, \sqrt{\frac{(2\gamma)^3}{3!}} x^3, \dots \right]^T. \quad (4)$$

模型在映射空间中使用线性模型进行数据建模, 线性模型的权重  $\mathbf{w}$  通过优化损失函数而得到. 据此推导出使得损失函数最小的  $\mathbf{w}$ :

$$\mathbf{w} = (\lambda \mathbf{I}_d + \Phi \Phi^T)^{-1} \Phi \mathbf{y} = \Phi (\Phi^T \Phi + \lambda \mathbf{I}_n)^{-1} \mathbf{y}, \quad (5)$$

式中定义  $\Phi = \Phi_{ai}$  为数据集映射到高维空间后的特征矩阵,  $\mathbf{y} = \mathbf{y}_i$  为目标向量. 将该  $\mathbf{w}$  代回 (1) 式中, 并使用核函数代替高维向量的内积, 得到

$$\begin{aligned} y &= \mathbf{w}^T \Phi(\mathbf{x}) = \mathbf{y} (\Phi^T \Phi + \lambda \mathbf{I}_n)^{-1} \Phi^T \Phi(\mathbf{x}) \\ &= \mathbf{y} (\mathbf{K} + \lambda \mathbf{I}_n)^{-1} \kappa(\mathbf{x}), \end{aligned} \quad (6)$$

其中  $\mathbf{K}_{ij} = \Phi(\mathbf{x}_i)^T \Phi(\mathbf{x}_j)$ ,  $\kappa(\mathbf{x}) = \mathbf{K}(\mathbf{x}_i, \mathbf{x})$ . 如果

将核函数  $\mathbf{K}(\mathbf{x}_i, \mathbf{x})$  视为计算新样本  $\mathbf{x}$  与训练样本  $\mathbf{x}_i$  之间的相似度, 那么可以认为使用核函数的 KRR 的预测结果是基于计算新样本与训练样本之间的相似度的方法得到的. 与训练样本相似的新样本具有与之相似的元素级别特征, 因而也会根据实验已知的结论给出相似的预测结果. 而与训练数据较不同的样本, 则通过全局考察其相似性, 来给出合理的预测值.

### 2.2.2 SVR<sup>[31]</sup>

SVR 与 KRR 类似, 具有线性模型的回归形式. 在 SVR 模型中, 同样采用核函数的方法为模型引入非线性拟合能力. 而该方法与 KRR 不同的是, SVR 对权重  $\mathbf{w}$  进行优化时容许预测值与实验值之间有  $\varepsilon$  的偏差, 因此它不需要模型完美地满足所有的数据点, 可以减少模型过拟合的风险. 具体地, 它的损失函数如下<sup>[31]</sup>:

$$\min_{\mathbf{w}, b} \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^m \ell_\varepsilon(f(\mathbf{x}_i) - y_i), \quad (7)$$

$$\ell_\varepsilon(z) = \begin{cases} 0, & |z| \leq \varepsilon, \\ |z| - \varepsilon, & \text{otherwise,} \end{cases} \quad (8)$$

其中  $C$  为权衡结构风险与经验风险的超参数;  $\ell_\varepsilon$  为  $\varepsilon$ -不敏感函数, 用以忽略数据与模型之间的较小偏差. 在本研究中 SVR 的核函数选为径向核函数. 为了使得其损失函数最小化, 通过构造拉格朗日函数并解其对偶问题, 最终可以将 SVR 模型表示为

$$f(\mathbf{x}) = \sum_{i=1}^m (\hat{\alpha}_i - \alpha_i) \kappa(\mathbf{x}, \mathbf{x}_i) + b. \quad (9)$$

SVR 与 KRR 具有类似的形式, 但是在 SVR 中只有对应的  $\hat{\alpha}_i - \alpha_i$  不为 0 的训练样本才被用来预测未知的铁电体, 而这些样本就被称为“支持向量”. SVR 的思想与实验科学非常切合, 研究者在实验测得的居里温度本身就存在一定的随机误差, 在损失函数中引入  $\varepsilon$ -不敏感函数可以在一定程度上防止将这种随机误差纳入模型. 此外使用该方法可以防止对训练样本的过拟合, 同时还可以使模型更快地进行预测.

### 2.2.3 ETR 模型<sup>[32]</sup>

为了尝试使机器学习模型获得更好的可解释性, 还使用了 ETR 模型对复杂铅基钙钛矿的居里温度进行建模. ETR 将若干个分类与回归树

(classification and regression trees, CART) 集成在一起. 其中每棵 CART 在构建时采用随机挑选属性和分裂值的分裂方法, 从而构建出若干个完全随机的树. 在 ETR 模型中, 通过控制每棵树的深度“depth”、划分节点时最少的样本数“min\_samples”来限制树的规模, 通过控制总共集成树的数量“ $n$ ”来调节模型的拟合能力.

## 2.3 模型超参数的优化

### 2.3.1 交叉验证策略

为了调整模型的超参数和检验模型的泛化效果, 采用 5 次 10 折交叉验证的策略. 在一次 10 折交叉验证中, 将样本集随机划分为 10 等份, 轮流使用其中的 1 份作为测试集而其余 9 份作为训练集. 机器学习训练集的数据而对测试集中的样本的居里温度进行预测, 通过比较测试样本居里温度的实验值与预测值之间的 MAE、均方跟误差 (root-mean-square error, RMSE) 来衡量模型对于未知样本的预测能力. 为了减少单次随机划分样本集带来的偶然性, 以上过程重复 5 次, 取结果的均值作为最终的评价依据.

由于本文研究对象分为简单铅基钙钛矿和复杂铅基钙钛矿两种, 而简单铅基钙钛矿的种类有限且研究较为详尽, 同时本研究的主要目标——复杂铅基钙钛矿种类繁多, 相对而言已研究确认的材料较少, 因此在交叉验证的过程中将数据集中  $B$  位元素种类小于 3 的 24 项简单钙钛矿的数据作为“教师”, 只出现在训练集中用来模型训练, 不出现在测试集中参加模型性能的评估.

### 2.3.2 超参数的优化

模型的超参数直接影响了模型的拟合能力和拟合策略, 选取一组合适的超参数对模型的构建非常重要. 在定义了交叉验证的策略之后, 在一定的超参数区间内通过随机搜索<sup>[36]</sup>和网格搜索的方式选取若干组超参数并使用该超参数构建模型. 通过之前定义的交叉验证策略对模型效果进行检验, 最终选取使得交叉验证结果最理想的一组超参数作为模型最终的超参数.

例如, 当优化 SVR 中的超参数  $\varepsilon$  时, 在 0—10 的取值范围内以 0.5 为步长对  $\varepsilon$  进行网格搜索. 如图 1 所示, 对比了  $\varepsilon$  取不同值时模型的测试误差, 同时将该参数模型应用于全部数据集时的支持

向量数目也一并列出, 以作参考. 当  $\varepsilon$  取值小于 4 时模型的误差较低; 当  $\varepsilon$  取值进一步增大时模型的误差出现快速上升. 另一方面随着  $\varepsilon$  取值的增大, 模型中的支持向量的数目也在减少. 在泛化能力相似的情况下, 倾向于选择较简单的模型以避免过拟合, 因此  $\varepsilon$  的值最终取为 4. 此时, 模型经过完整地学习训练集中所有 205 项数据后, 其模型中的支持向量数目为 147 项, 也就是说模型仅使用其中 147 项材料的数据对未知材料进行预测.

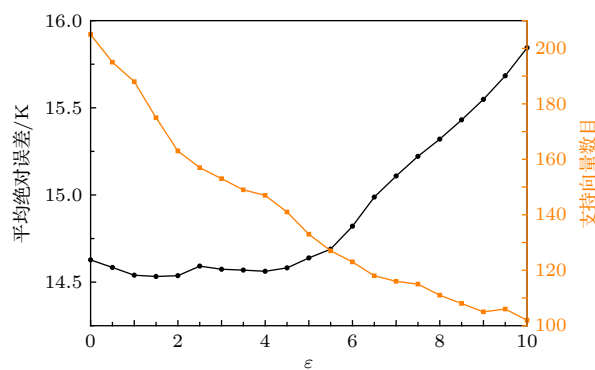


图 1 支持向量回归中的超参数  $\varepsilon$  的优化及支持向量数目分析

Fig. 1. Optimization of hyperparameters in support vector regression and the analysis of the number of support vectors.

依据上述方法, 同样对其他超参数做了类似的搜索和优化, 最终得到的超参数如表 1 所列.

表 1 本文三种机器学习方法所采用的超参数

Table 1. Hyperparameters of the three machine learning methods in this study.

模型	超参数
KRR	$\lambda = 0.001, \gamma = 0.005$
SVR	$\gamma = 0.02, \varepsilon = 4, C = 900$
ETR	$n = 108, \text{depth} = 20, \text{min\_samples} = 2$

## 2.4 模型的集成

上述三种模型, 从不同的角度出发对钙钛矿铁电材料的居里温度进行了学习, 均取得了良好的成效. 为了进一步提高预测的准确度, 通过集成的方法, 将上述三种回归模型融合为一个统一的模型. 集成模型同时训练上述三种机器学习模型, 并使用三种模型分别对居里温度进行预测. 将三种机器学习模型的预测均值作为最终的输出.

为了检验集成的效果, 也对集成模型使用交叉验证的方法进行了评估. 另外, 也尝试了赋予三个

模型其他合理的权重,如图2所示,当将三种模型融合在一起时,往往可以得到比最优单模型更好的效果.测得的最优融合比例为0.6:0.2:0.2(KRR:SVR:ETR),其MAE为13.7 K.等权重模型融合测得的MAE为13.9 K(图2中的红色点).出于保持模型简单性的考虑,直接采用了等权重的方法对模型进行融合.

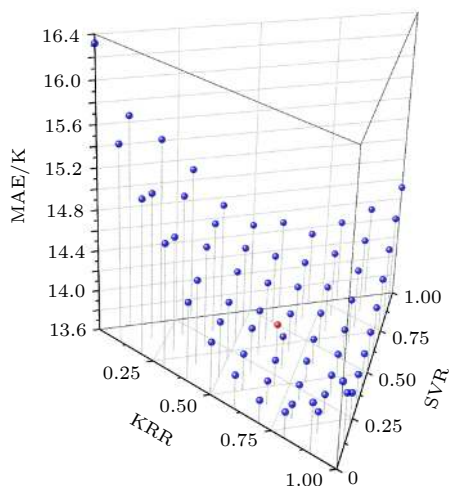


图2 不同模型权重的融合实验结果  
Fig. 2. Performance of ensemble model with different base model weight.

### 2.5 使用机器学习方法对未知材料探索

对数据集中出现的20种不同组分的简单铅基钙钛矿相互组合,得到了大量的包含 $\text{PbTiO}_3$ 组分的二元或三元的复杂钙钛矿固溶体,共有190种可能的组合方式,而在这些复杂钙钛矿固溶体中,使其中简单钙钛矿组分从2%开始,以2%为步长变化,从而研究不同组分的复杂铅基钙钛矿.至此,得到了超过20万种铅基钙钛矿材料.使用上文中的方法对它们构建特征,并使用机器学习的方法对其中所有材料的居里温度进行预测.

## 3 结果与讨论

### 3.1 模型效果验证

在5轮的交叉验证中,使用了MAE, RMSE和相关系数三项指标对三种机器学习模型和集成后模型的预测结果进行了评估.评估结果列于表2,在三种学习模型中KRR给出了最低的MAE,为14.4 K,同时RMSE为最低的22.5 K.所有模型的在交叉验证中对于未学习过的材料的预测值与材

料的实验值之间均有较大的相关系数.而将三种机器学习模型进行集成后,集成模型的表现比三种机器学习模型更为优秀.同时在表2中也将本文机器学习模型与其他使用机器学习方法研究钙钛矿居里温度的工作中的最优模型进行了比较.

表2 本文所使用的机器学习方法的评估及与其他研究者工作的对比

Table 2. Evaluation of machine learning methods in this paper and the comparison with other works.

	模型				对比	
	KRR	SVR	ETR	集成模型	文献[10]	文献[11]
MAE/K	14.4	14.7	16.1	13.9	30.2	21.2
RMSE/K	22.5	23.4	23.8	21.4	—	28.7
相关系数	0.96	0.96	0.96	0.97	—	0.85

为了给读者更直观地展示各种机器学习模型对于未知材料居里温度的预测效果,给出了5轮交叉验证过程中不同材料的预测均值与实验值,如图3所示.图3中不同的橘色点对应不同的材料,其纵坐标为该材料的居里温度实验值,横坐标为本文机器学习模型对该材料居里温度的预测均值.蓝色点划线 $y = x$ 代表着预测值与实验值完全吻合的理想情况.从图3可以看出在交叉验证过程中对于大部分未学习的铅基钙钛矿,机器学习模型对其居里温度的预测值都落在其实验值的附近,预测值与实验值之间具有较强的相关性.

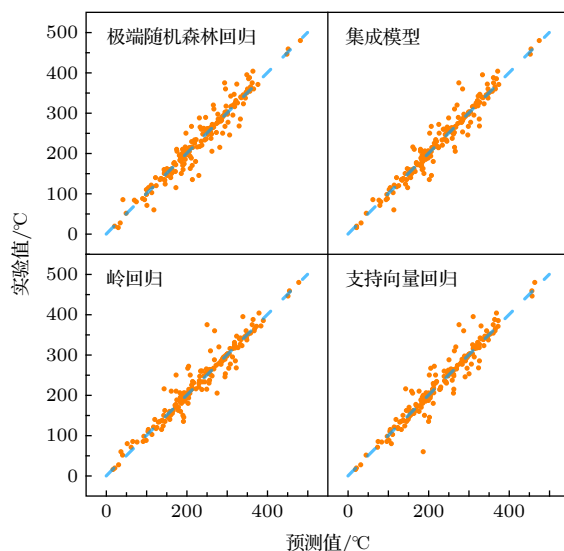


图3 三种机器学习模型及其集成模型对材料的预测值与实验值的比较

Fig. 3. Comparison of prediction and experimental values of three machine learning models and their ensemble models.

此外,在图3中对比三种机器学习模型不难看出,这三种机器学习模型一致得给出了不错的预测成绩,但是其对于特定材料的预测结果却不尽相同,这是由模型之间的差异所导致的.如上文所述,KRR模型与SVR模型具有相同的广义线性回归形式,相同的核函数形式.但是其不同的损失函数赋予了两种模型不同的“价值观”,在学习过程中它们遵从各自的“价值观”各自寻找其最优参数,最后得到的结果必然是不同的.ETR模型基于CART,与前两种模型形式显著不同,在交叉验证中,其MAE与RMSE也略逊于前两种模型,但是其树的形式更利于人类理解.同时如图2所示,融合更多的模型也有助于提高模型的预测能力,因此将这三种模型融合在一起,相互取长补短,得到了预测能力更优的集成模型.例如支持向量模型对于实验值50 K附近的某材料的预测值发生了明显的偏差,经过模型融合之后,该偏差得到了有效的缓解.

### 3.2 模型成果

本文的产出主要分为两方面:本研究中整理的铅基钙钛矿的居里温度数据和本文中构建的机器学习模型;对超过20万种铅基钙钛矿材料的居里温度进行了预测,并挑选了两种可能具有高居里温度的材料以供实验研究人员参考.

#### 3.2.1 数据与模型

在本研究中收集的包含205种铅基钙钛矿的居里温度数据集在补充文件(online)中给出,可以供其他相关领域的理论或实验研究者参考和使用.利用该数据集,本文构建了三种机器学习模型:KRR,SVR,ETR.与多数研究者不同的是,还将这三种机器学习模型集成为一个模型.

为了检验单个机器学习模型和集成机器学习模型的可靠性,本文通过交叉验证的方法对模型进行测试.在交叉验证的过程中KRR,SVR,ETR单模型的MAE分别为14.4 K,14.7 K,16.1 K.集成后的机器学习模型在交叉验证中的MAE为13.9 K.而在其他研究者对钙钛矿铁电体的居里温度进行预测的而建立的模型,其交叉验证的MAE为30.2 K<sup>[10]</sup>和21.2 K<sup>[11]</sup>.该交叉验证误差的差异可能来源于模型构建过程中特征构建的方式不同等因素,因此本文中元素属性的均值和方差构建特征的方法拥有一定的泛化意义.

此外,使用ETR模型学习后,可以根据CART在不同特征上进行分裂时的方差变化来量化不同特征对于材料居里温度的重要性.具体而言,ETR模型是多棵CART的总和,CART在其每个节点中以某个特征的取值作为分割点将节点中的样本分割至两个子节点(一个节点中的样本对应特征取值均大于等于该值,另一个节点中的样本对应特征取值均小于该值).本文采用节点中样本居里温度的方差来刻画节点中样本的离散性,采用分裂前后的加权方差变化来刻画该分裂带来的聚合增益.如果依据某特征值将样本分为两类时带来了较大的聚合增益,也就是说依据样本该特征的值可以将节点中具有较高居里温度的样本和居里温度较低的样本更好地地区分开来,那么该特征的取值对于判断居里温度显然是重要的.将所有节点上采用某特征的取值作为分割条件时的方差变化以其节点中的样本数量作为权重加权求和,得到该特征对于预测材料居里温度的重要性.该指标表示了在使用ETR模型预测居里温度时该特征的重要程度<sup>[37]</sup>.

将所有特征重要性进行了归一化,并将重要性最大的五种特征表示在图4中.图4所示为B位元素(或该元素最常见单质)的热导率均值、电导率均值、比热容方差、元素序数均值、钙钛矿中离子偏移均值,五种特征在使用ETR模型预测材料的居里温度中起到了最重要的作用.

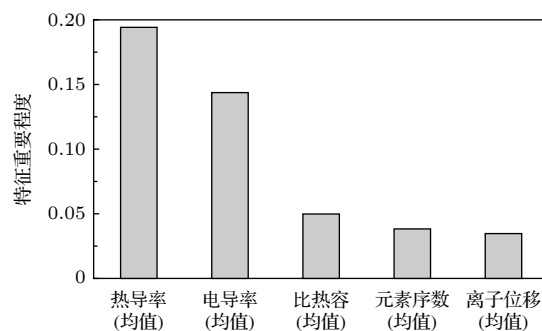


图4 ETR模型中最重要的5项特征  
Fig. 4. Five most important features in ETR.

#### 3.2.2 使用机器学习寻找高温铅基铁电体

使用集成的机器学习模型,对数据集以外的超过20万种铅基钙钛矿材料的居里温度进行了预测,并挑选出了0.02PbMn<sub>1/2</sub>Nb<sub>1/2</sub>O<sub>3</sub>-0.98PbTiO<sub>3</sub> (0.02PMN-0.98PT), 0.02PbGa<sub>1/2</sub>Nb<sub>1/2</sub>O<sub>3</sub>-0.02PbMn<sub>1/2</sub>Nb<sub>1/2</sub>O<sub>3</sub>-0.96PbTiO<sub>3</sub> (0.02PGN-0.02PMN-

0.96PT) 两种具有较高居里温度的铅基钙钛矿固溶体, 其对应的居里温度分别为 481 °C 和 466 °C. 集成机器学习模型对于 PGN-PMN-PT 体系中不同组分权重所形成的一系列固溶体进行的预测结果如图 5 所示.

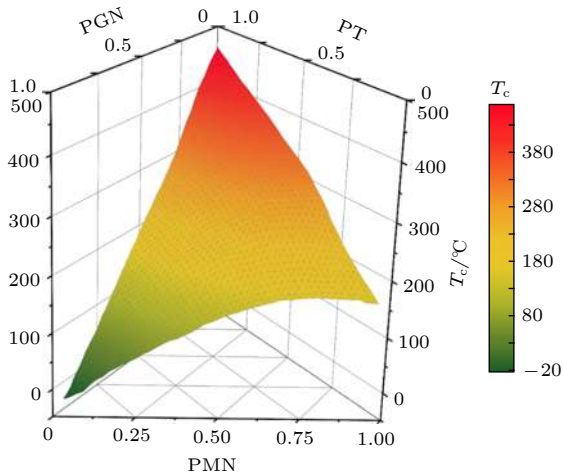


图 5 集成机器学习模型对 PGN-PMN-PT 固溶体的居里温度的预测

Fig. 5. Prediction of Curie temperature of PGN-PMN-PT solid solution by ensemble machine learning model.

此外, 在补充材料 (online) 中还提供了全部 20 余万种钙钛矿固溶体的居里温度预测数据来为实验研究提供更详尽的参考.

## 4 结 论

本文在不同组分和配比的铅基钙钛矿铁电晶体所形成的庞大搜索空间中寻找具有高居里温度的铁电体, 利用元素基本物理性质等信息构建了一系列特征, 得以对各种不同组分的铅基钙钛矿铁电晶体进行统一的描述, 从而使得机器可以从基本物理性质出发对固溶体的居里温度进行建模和预测. 分别使用了 KRR, SVR, ETR 三种机器学习方法对铅基钙钛矿铁电固溶体的居里温度进行了学习, 并使用了等比例融合的方法构建了集成学习模型.

采用交叉验证的方法对模型的可靠性进行了评估, 其中集成机器学习模型在交叉验证中测得 MAE 为 13.9 K. 使用该方法对超过 20 万种铅基钙钛矿铁电体的居里温度进行了预测, 并挑选出了两种具有潜在高居里温度的铅基钙钛矿铁电晶体: 0.02PMN-0.98PT, 0.02PGN-0.02PMN-0.96PT.

## 参考文献

- [1] Liu Y, Li X, Wang Z, He C, Li T, Ai L, Chu T, Pang D, Long X 2013 *CrystEngComm* **15** 1643
- [2] Wu J, Chang Y, Yang B, Zhang S, Sun Y, Guo F, Cao W 2015 *J. Mater. Sci.: Mater. Electron* **26** 1874
- [3] Zhao Y Q, Ma Q R, Liu B, Yu Z L, Yang J, Cai M Q 2018 *Nanoscale* **10** 8677
- [4] Ding Y F, Zhao Q Q, Yu Z L, Zhao Y Q, Liu B, He P B, Zhou H, Li K, Yin S F, Cai M Q 2019 *J. Mater. Chem. C* **7** 7433
- [5] Yu Z L, Ma Q R, Liu B, Zhao Y Q, Wang L Z, Zhou H, Cai M Q 2017 *J. Phys. D* **50** 465101
- [6] Zhang T, Jiang Y, Song Z, Huang H, He Y, Fang Z, Weng H, Fang C 2019 *Nature* **566** 475
- [7] Umehara M, Stein H S, Guevarra D, Newhouse P F, Boyd D A, Gregoire J M 2019 *NPJ Comput. Mater.* **5** 34
- [8] Iwasaki Y, Takeuchi I, Stanev V, et al. 2019 *Sci. Rep.* **9** 2751
- [9] Aykol M, Hegde V I, Hung L, Suram S, Herring P, Wolverton C, Hummelshøj J S 2019 *Nat. Commun* **10** 1
- [10] Balachandran P V, Kowalski B, Sehirlioglu A, Lookman T 2018 *Nat. Commun.* **9** 1668
- [11] Zhai X, Chen M, Lu W 2018 *Comput. Mater. Sci.* **151** 41
- [12] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay É 2011 *J. Mach. Learn. Res.* **12** 2825
- [13] Chen H, Fan C 2010 *Mater. Lett.* **64** 654
- [14] Fang B, Shan Y, Tezuka K, Imoto H 2005 *J. Mater. Sci.* **40** 6445
- [15] Gao Y Q, Xu H Q, Wu Y J, He T H, Xu G S, Luo H S 2001 *Jpn. J. Appl. Phys.* **40** 4998
- [16] Guo Y P, Xu H Q, Luo H S, Xu G S, Yin Z W 2001 *J. Cryst. Growth* **226** 111
- [17] Hao H, Zhang S, Liu H, Shrout T R 2009 *J. Appl. Phys.* **105** 024104
- [18] He C, Li X, Wang Z, Liu Y, Shen D, Li T, Long X, Ye Z G 2012 *CrystEngComm* **14** 4407
- [19] Wang L, Liang R, Mao C, Du G, Wang G, Dong X 2013 *Ceram. Int.* **39** 8571
- [20] Yasuda N, Ohwa H, Hasegawa D, Hayashi K, Hosono Y, Yamashita Y, Iwata M, Ishibashi Y 2000 *Jpn. J. Appl. Phys.* **39** 5586
- [21] Zhang S, Lebrun L, Jeong D Y, Randall C A, Zhang Q, Shrout T R 2003 *J. Appl. Phys.* **93** 9257
- [22] Bing Y, Ye Z 2002 *Proceedings of the 13th IEEE International Symposium on Applications of Ferroelectrics*, Nara, Japan, June 1, 2002 pp447-450
- [23] Li X, Wang Z, He C, Liu Y, Long X, Han S, Pan S 2015 *Mater. Lett.* **143** 88
- [24] Rajan K K, Zhang M J, Lim L C 2005 *Jpn. J. Appl. Phys.* **44** 264
- [25] Salak A N, Shilin A D, Bushinski M V, Olekhovich N M, Vyshatko N P 2000 *Mater. Res. Bull.* **35** 1429
- [26] Salak A N, Vyshatko N P, Ferreira V M, Olekhovich N M, Shilin A D 2003 *Mater. Res. Bull.* **38** 453
- [27] Yamashita Y, Harada K 1997 *Jpn. J. Appl. Phys.* **36** 6039
- [28] Wang J, Tang X G, Chan H L W, Choy C L, Luo H 2005

- Appl. Phys. Lett.* **86** 152907
- [29] Grinberg I, Rappe A M 2007 *Phase Trans.* **80** 351
- [30] Balachandran P V, Shearman T, Theiler J, Lookman T 2017 *Acta Crystallogr. Sect. B* **73** 962
- [31] Zhou Z H 2016 *Machine Learning* (Beijing: Tsinghua University Press) pp133–137 (in Chinese) [周志华 2016 机器学习 (北京: 清华大学出版社) 第133—137页]
- [32] Geurts P, Ernst D, Wehenkel L 2006 *Machine Learning* **63** 3
- [33] Hoerl A E, Kennard R W 1970 *Technometrics* **12** 55
- [34] Welling M [https://www.ics.uci.edu/~welling/classnotes/papers\\_class/Kernel-Ridge.pdf](https://www.ics.uci.edu/~welling/classnotes/papers_class/Kernel-Ridge.pdf) [2019-6-7]
- [35] Lin C J [https://www.csie.ntu.edu.tw/~cjlin/talks/kuleuven\\_svm.pdf](https://www.csie.ntu.edu.tw/~cjlin/talks/kuleuven_svm.pdf) [2019-6-7]
- [36] Bergstra J, Bengio Y 2012 *J. Mach. Learn. Res.* **13** 281
- [37] Louppe G, Wehenkel L, Suter A, Geurts P 2013 *Understanding Variable Importances in Forests of Randomized Trees* (Advances in Neural Information Processing Systems) pp431–439



# High critical transition temperature of lead-based perovskite ferroelectric crystals: A machine learning study\*

Yang Zi-Xin   Gao Zhang-Ran   Sun Xiao-Fan   Cai Hong-Ling

Zhang Feng-Ming   Wu Xiao-Shan<sup>†</sup>

(*Collaborative Innovation Center of Advanced Microstructures, Laboratory of Solid State Microstructures,*

*School of Physics, Nanjing University, Nanjing 210093, China*)

( Received 18 June 2019; revised manuscript received 20 August 2019 )

## Abstract

Ferroelectrics undergoes a reversible structural phase from the ferroelectric phase to the paraelectric phase when its temperature exceeds the critical temperature namely Curie temperature  $T_c$ . As ferro-paraelectric phase transition is always accompanied by heat-flow, dielectric and pyroelectric anomaly, the value of  $T_c$  is extremely important for ferroelectrics. In this paper, the Curie temperature of lead-based perovskite ferroelectric solid solution is studied by machine learning methods including kernel ridge regression (KRR), support vector regression (SVR) and extremely randomized trees regression (ETR). We collect the  $T_c$  values of 205 different lead-based perovskites from published experimental papers, both simple perovskites with only one type of  $B$  site ion and complex perovskites with up to 5 kinds of ions in  $B$  position such as PMN-PFN-PZT are gathered. The diversity of our dataset is guaranteed for the good generalization of our model in perovskite solid solution of different complexity. The features are constructed from the physical and chemical properties of the  $B$  site elements in corresponding materials. The weighted-average and variance of the elemental properties are calculated and fed to machine learning models. We use the 5 runs of ten fold cross-validation method to evaluate the machine learning models. The hyperparameters are also chosen carefully with the cross-validation to avoid over fitting. The radial basis function kernel is used in both KRR and SVR. The insensitive error in the SVR is set to be 4 which is comparable to the random error in experiment. From our cross-validation, we find that the mean average errors (MAEs) between the predicted and experimental values of the machine learning methods are 14.4 K, 14.7 K, and 16.1 K, respectively. And the root-mean-square errors (RMSEs) are 22.5 K, 23.4 K, 23.8 K, respectively. After the optimization and the evaluation, our three machine learning models are stacked together by averaging the output of each regression model and thus building an ensemble model. The MAE of the ensemble model is 13.9 K. The RMSE of the ensemble model is 21.4 K. The predicted values keep a correlation coefficient of 0.97 with the experimental values. From the variance reduction in ETR, we derive the importance of our features when determining the Curie temperatures. The five most important factors in our ETR model are “weighted-average thermal conductivity”, “weighted-average conductivity”, “variance of specific heat capacity”, “weighted-average element number”, and “weighted-average relative atomic displacement”. We predict the Curie temperatures higher than those of 200000 types of lead-based perovskites after being trained. Now, we provide two ferroelectric materials that may have high Curie temperatures:  $0.02\text{PbMn}_{1/2}\text{Nb}_{1/2}\text{O}_3\text{-}0.98\text{PbTiO}_3$  (0.02PMN-0.98PT) and  $0.02\text{PbGa}_{1/2}\text{Nb}_{1/2}\text{O}_3\text{-}0.02\text{PbMn}_{1/2}\text{Nb}_{1/2}\text{O}_3\text{-}0.96\text{PbTiO}_3$  (0.02PGN-0.02PMN-0.96PT). The predicted Curie temperatures of them are 481 °C and 466 °C, respectively.

**Keywords:** machine learning, ferroelectric, Curie temperature, perovskite

**PACS:** 05.70.Fh, 02.60.Pn, 43.38.Fx

**DOI:** 10.7498/aps.68.20190942

\* Project supported by the National Natural Science Foundation of China (Grant No. 118742007).

<sup>†</sup> Corresponding author. E-mail: xswu@nju.edu.cn