

上海大学学报(自然科学版)  
*Journal of Shanghai University(Natural Science Edition)*  
ISSN 1007-2861,CN 31-1718/N

## 《上海大学学报(自然科学版)》网络首发论文

题目: 尖晶石氧化物能量和结构的第一性原理计算和机器学习  
作者: 李一航, 肖斌, 唐宇超, 刘馥, 王小梦, 刘轶  
收稿日期: 2020-08-09  
网络首发日期: 2020-09-15  
引用格式: 李一航, 肖斌, 唐宇超, 刘馥, 王小梦, 刘轶. 尖晶石氧化物能量和结构的第一性原理计算和机器学习. 上海大学学报(自然科学版).  
<https://kns.cnki.net/kcms/detail/31.1718.N.20200914.1748.002.html>



**网络首发:** 在编辑部工作流程中, 稿件从录用到出版要经历录用定稿、排版定稿、整期汇编定稿等阶段。录用定稿指内容已经确定, 且通过同行评议、主编终审同意刊用的稿件。排版定稿指录用定稿按照期刊特定版式(包括网络呈现版式)排版后的稿件, 可暂不确定出版年、卷、期和页码。整期汇编定稿指出版年、卷、期、页码均已确定的印刷或数字出版的整期汇编稿件。录用定稿网络首发稿件内容必须符合《出版管理条例》和《期刊出版管理规定》的有关规定; 学术研究成果具有创新性、科学性和先进性, 符合编辑部对刊文的录用要求, 不存在学术不端行为及其他侵权行为; 稿件内容应基本符合国家有关书刊编辑、出版的技术标准, 正确使用和统一规范语言文字、符号、数字、外文字母、法定计量单位及地图标注等。为确保录用定稿网络首发的严肃性, 录用定稿一经发布, 不得修改论文题目、作者、机构名称和学术内容, 只可基于编辑规范进行少量文字的修改。

**出版确认:** 纸质期刊编辑部通过与《中国学术期刊(光盘版)》电子杂志社有限公司签约, 在《中国学术期刊(网络版)》出版传播平台上创办与纸质期刊内容一致的网络版, 以单篇或整期出版形式, 在印刷出版之前刊发论文的录用定稿、排版定稿、整期汇编定稿。因为《中国学术期刊(网络版)》是国家新闻出版广电总局批准的网络连续型出版物(ISSN 2096-4188, CN 11-6037/Z), 所以签约期刊的网络版上网络首发论文视为正式出版。

# 尖晶石氧化物能量和结构的第一性原理计算 和机器学习

李一航<sup>1</sup>, 肖斌<sup>2</sup>, 唐宇超<sup>1</sup>, 刘馥<sup>1</sup>, 王小梦<sup>1</sup>, 刘轶<sup>\*2,1</sup>

(1. 上海大学理学院物理系, 上海 200444; 2. 上海大学材料基因组工程研究院, 上海 200444)

**摘要:** 正尖晶石氧化物  $AB_2O_4$  结构可通过对 A 和 B 位点分别置换 73 种合金元素产生 5329 种原子构型。本工作首先利用高通量第一性原理方法计算了这 5329 种正立方尖晶石结构的形成能和晶格常数。为了发展高效的机器学习方法, 本工作提出“中心-环境”(Center-Environment, CE) 特征模型, 以构建同时包含局部成分和结构信息的特征作为机器学习算法的输入变量。结合第一性原理计算数据, 我们使用随机森林算法开发了机器学习模型, 可以准确地预测尖晶石的形成能和晶格常数。通过比较机器学习预测的假想结构与实验结构的形成能, 我们预测出 361 种比研究的已知实验结构更稳定的新尖晶石氧化物结构。我们讨论了与尖晶石稳定性有关的“好”与“坏” A/B 位元素, 有助于指导实验合成新的稳定尖晶石氧化物。

**关键词:** 密度泛函理论; 机器学习; 特征工程; “中心-环境”特征模型; 尖晶石氧化物

**中图分类号:** O469

**文献标志码:** A

## First-principles computation and machine learning study of the energies and structures of spinel oxides

LI Yihang<sup>1</sup>, XIAO Bin<sup>2</sup>, TANG Yuchao<sup>1</sup>, LIU Fu<sup>1</sup>, WANG Xiaomeng<sup>1</sup>, LIU Yi<sup>\*2,1</sup>

(1. Department of Physics, Shanghai University, Shanghai 200444, China;

2. Materials Genome Institute, Shanghai University, Shanghai 200444, China)

**Abstract:** The formal spinel oxides  $AB_2O_4$  can have 5329 configurations by substituting A and B sites with 73 elements, respectively. This work first applied the first-principles method to calculate the formation energies and lattice parameters of 5329 spinel oxides. To develop efficient machine learning (ML) methods, this work proposed “Center-Environment” (CE) feature models to construct the input variables to ML algorithm containing local composition and structure information. Based on the first-principles computational data, we used random forest algorithm to develop a ML model to predict the formation energies and lattice parameters of spinel oxides. By comparing the formation energies of hypothetical and experimental structures predicted by ML, we discovered 361 new spinel oxides more stable than the studied known experimental structures. We discussed the “good” and “bad” A/B stabilization elements, helping to guide the experimental synthesis of novel stable spinel oxides.

**Keywords:** Density functional theory; Machine learning; Feature engineering; “Center-environment” feature model; Spinel oxide

收稿日期: 2020-08-09 录用日期: 2020-09-09

基金项目: 省部共建高品质特殊钢冶金与制备国家重点实验室、上海市钢铁冶金新技术开发应用重点实验室自主课题(SKLASS 2019-Z024); 上海市科学技术委员会资助项目(19DZ2270200);

通信作者: 刘轶(1971—), 男, 教授, 博士生导师, 博士, 研究兴趣为高性能合金和纳米材料的计算、实验和机器学习研究. E-mail: yiliu@shu.edu.cn

高通量第一性原理计算<sup>[1-2]</sup>方法可用于对大量的材料体系和构型进行系统的研究,但庞大的计算量使得其成本昂贵耗时。机器学习(machine learning, ML)<sup>[3-4]</sup>方法可以从计算或实验材料数据中进行数据挖掘或机器学习,获得其中隐含的关联规律,并用于未知材料性能的高效预测。本研究基于高通量第一性原理计算构建机器学习模型,可加快新材料的能量和结构预测,指导新材料实验合成。

尖晶石材料因其庞大的种类、灵活可调的组分以及优异的物理化学性质而广泛应用于锂电池正极<sup>[5-6]</sup>、光催化剂材料<sup>[7-8]</sup>、铁氧体磁性材料<sup>[9]</sup>和半金属材料<sup>[10]</sup>等领域。尖晶石氧化物可分为正尖晶石结构、反尖晶石结构和混合尖晶石结构<sup>[11-13]</sup>。本工作研究的正尖晶石氧化物化学通式为  $AB_2O_4$ , 空间群为  $Fd\bar{3}m$ , 其中 A、B 阳离子<sup>[14]</sup>分别处于氧离子形成的四面体间隙(T 位点)和八面体间隙(O 位点)中,如图 1 所示。本研究用 73 种元素(图 2)分别替换尖晶石结构的 T 和 O 位点原子,获得 5329 ( $73^2=5329$ ) 个可能的尖晶石氧化物。

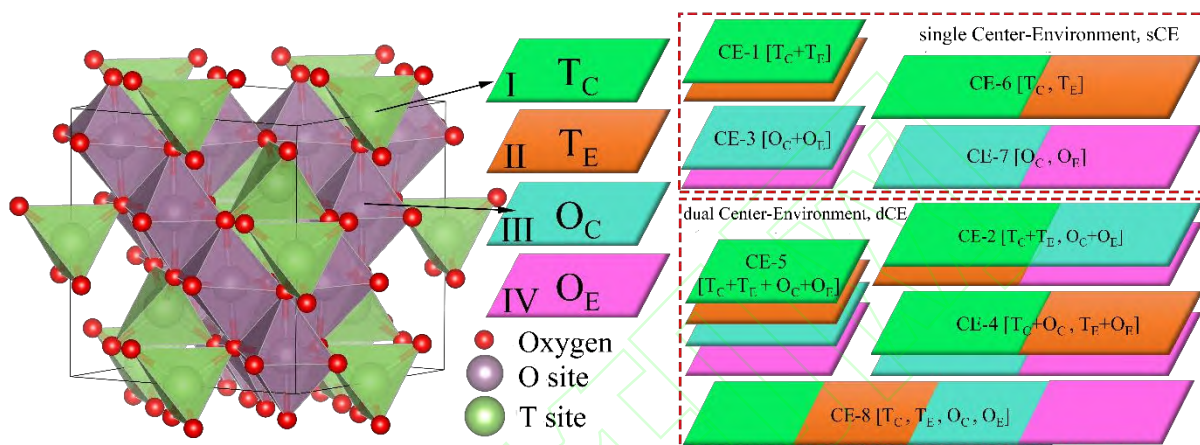


图 1 正尖晶石惯用单胞(其中 T 位点和 O 位点分别为四面体间隙和八面体间隙)。通过多种方式结合 T/O 位点中心原子 T<sub>C</sub> (I)/O<sub>C</sub> (III)和 T/O 位点环境原子 T<sub>E</sub> (II)/O<sub>E</sub> (IV)特征构建的中心-环境特征模型示意图(CE-1 至 CE-8)。

Fig. 1 Conventional cell of the normal spinel type structure and schematic illustration of center-environment feature models (CE-1...CE-8) constructed by combining the features of the center atoms T<sub>C</sub> (I)/O<sub>C</sub> (III) of the T/O site and those of the environment atoms T<sub>E</sub> (II)/O<sub>E</sub> (V) of the T/O site in various manners.

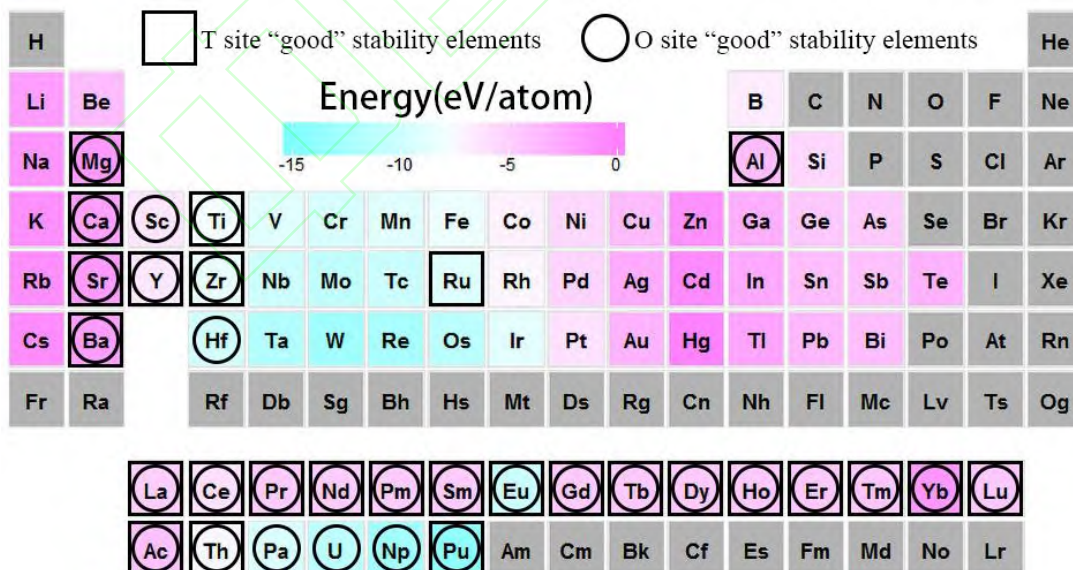


图 2 本工作研究的正尖晶石氧化物的组成元素(分别在 T 和 O 位点共有 73 个,其中灰色背景的元素没有研究),图中色标从蓝色到粉红色作为对应的单原子基态总能量的函数。正方形和圆形代表机器学习预测的“好”的 T 位点和 O 位点的稳定性元素。

Fig. 2 Constitution elements of the normal spinel oxides  $AB_2O_4$  studied in this work (73 in total at the T and the O sites, respectively. The elements with grey background were not studied in this work.), with color bars from blue to pink as a function of the ground state total energy of the corresponding single atom shown in the periodic table. The squared and circled elements represent the “good” T and O site stabilizing elements, respectively, while the others are the “bad” T and O site stabilizing elements predicted by the machine learning

models.

在计算设计预测新材料时，首先要求新材料是热力学稳定或亚稳定的，形成能<sup>[15,16]</sup>的大小是评估新材料是否热力学稳定的重要判据之一。为了得到准确的能量，需要对预测结构进行充分弛豫。弛豫后的计算结构可用于未来的实验验证结构的理论参考依据。在本研究中，我们使用高通量第一性原理计算方法研究了 5329 个 AB<sub>2</sub>O<sub>4</sub> 尖晶石结构的形成能和晶格常数。本工作提出了一种“中心-环境”（Center-Environment, CE）特征模型，以构建同时包含成分和结构信息的特征，作为机器学习算法的输入变量。基于 DFT 计算数据，我们使用随机森林算法开发了机器学习模型，对尖晶石结构的 DFT 计算数据进行了学习训练，获得的机器学习模型可以准确有效地预测尖晶石氧化物的形成能和晶格常数。

## 1 研究方法

### 1.1 第一性原理计算

本工作采用基于赝势投影缀加波的第一性原理计算方法，通过 VASP 软件包<sup>[17]</sup>进行计算。交换关联泛函选取 GGA-PBE，平面波截断能为 520 eV，结构优化过程中保持能量的收敛精度为  $1 \times 10^{-5}$  eV/atom，力的收敛精度为 0.02 eV/Å。通过收敛性测试，布里渊区采用 Monkhorst-Pack 的  $7 \times 7 \times 7$  的网格。本工作中尖晶石形成能的计算公式为：

$$E_f = \frac{1}{N} (E_{\text{spinel}} - n_T E_T - n_O E_O - \frac{1}{2} N_O E_{O_2}) \quad (1)$$

其中， $E_{\text{spinel}}$  是尖晶石晶胞的基态总能， $E_T$ 、 $E_O$  和  $E_{O_2}$  分别为单个 T 位点原子、单个 O 位点原子和氧分子的基态总能量。 $n_T$ 、 $n_O$  和  $N_O$  分别为晶胞中 T 位点原子、O 位点原子以及氧原子的数量。 $N$  为单胞中的总原子数。

### 1.2 “中心-环境”（CE）特征模型

本工作结合元素/单质基本性质和尖晶石氧化物的组成和结构信息来构造复合特征，每个构成元素选取了元素和相应单质的 55 个基本性质<sup>[18]</sup>（表 1），如原子半径、氧化态、电负性等。

基于特征工程理念，本工作构造了“中心-环境”（CE）特征模型，通过将基本属性投影到由组分和结构信息组成的基组上来构造特征。首先定义一个由中心原子和周围环境原子组成的 CE 原子集，其中环境原子包括距离中心原子第 1 近邻到第 N 近邻的原子（ $N=1-10$ ）。如果只有一个中心原子及其对应的环境原子，本文称之为单中心 CE 模型（single Center-Environment, sCE），而双中心 CE 模型（dual Center-Environment, dCE）包括两个中心原子及其对应的环境原子。

CE 模型的基本思想是分别定义区分中心原子和环境原子，然后将环境原子的基础性质以与中心原子的距离倒数作为权重进行线性加和，从而实现将化学成分和结构信息编码到机器学习特征中。机器学习训练集 D 包括特征  $F_i (i=1-55, \text{性质序数})$  和预测目标 T，具体定义如下：

$$D = [F_1, F_2, \dots, F_i, \dots, F_{55}, T] \quad (2)$$

特征  $F_i (i=1-55)$  包括中心原子特征  $f_{C,i}$  和环境原子特征  $f_{E,i}$ ，其中，C 和 E 分别代表中心原子和环境原子：

$$F_i = [f_{C,i}, f_{E,i}], i = 1, 2, \dots, 55 \quad (3)$$

中心原子基本性质  $p_{C,i}$  直接作为该中心原子的特征：

$$f_{C,i} = p_{C,i} \quad (4)$$

环境原子的特征是根据每个环境原子  $j$  的基本性质  $p_{j,i}$  及其到中心原子距离的倒数加权求和而成的复合属性特征：



$$f_{E,i} = \sum_j \omega_j p_{j,i} \quad (5)$$

环境  $j$  原子与中心原子的距离为  $r_j$ ，则求和权重  $\omega_j$  为：

$$\omega_j = \frac{1/r_j}{\sum_j 1/r_j} \quad (6)$$

**表 1** 作为基本性质的元素和单质结构性质  
**Table 1** Atomic and structural properties as fundamental features

Property	Property	Property	Property
Atomic weight	Distance from core electron (Schubert)	Electronegativity (Pauling)	Oxidation state first
Boiling temperature	Distance from valence electron (Schubert)	Electronegativity absolute	Compression modulus
Radii metal (Waber)	Electrochemical weight equivalent	Energy cohesive (Brewer)	Mendeleev d-t start left
Radii covalent	Electronegativity (Alfred-Rochow)	Energy of ionization first	Mendeleev d-t start right
Density	Electronegativity (Martynov&Batsanov)	Energy of ionization second	Mendeleev H d-t start left
Spectral lines n0	Mass attenuation coefficient for MoK $\alpha$	Energy of ionization third	Mendeleev H d-t start right
Entropy of solid	Radii pseudo-potential (Zunger)	Enthalpy of atomization	Mendeleev H t-d start left
Group number	Volume of atom (Villars, Daams)	Enthalpy of melting	Mendeleev H t-d start right
Molar heat capacity	Atomic electron scattering factor at 0.5	Enthalpy of vaporization	Mendeleev Pettifor regular
Mendeleev Pettifor	Magnetic frequency of nuclei	Valence electron number	Mendeleev t-d start left
Spin nuclei	Thermal neutron capture cross section	Moment nuclear magnetic	Mendeleev t-d start right
Magnetic resonance	Charge nuclear effective (Clementi)	Nuclear charge effective (Slater)	Melting temperature
Quantum number	Mendeleev chemists sequence	Periodic number start counting left bottom, left-right sequence	
Periodic number start counting top right, right-left sequence		Atomic environment number (Villars, Daams)	
Atomic number start counting left top, left-right sequence			

本工作以单和双中心、标量和矢量形式等不同方式将中心原子和环境原子特征组合构建成多种复合 CE 特征模型。双中心 CE 模型分别选取尖晶石结构的非等效 T 位点和 O 位点作为中心。T<sub>C</sub> 和 T<sub>E</sub> 代表 T 位点的中心原子及其周围环境，如图 1 中 I 和 II 所示。同样，O<sub>C</sub> 和 O<sub>E</sub> 代表 O 位点的中心原子及其周围环境，如图 1 中 III 和 IV 所示。如图 1 所示，CE-1、CE-3、CE-6 和 CE-7 为四个 sCE 模型 CE-5、CE-2、CE-4 和 CE-8 为四个 dCE 模型。在 CE-1 到 CE-8 CE 模型中，分别将 T 或 O 位点定义为中心原子，并将其第 1 近邻至第 10 近邻的原子分别考虑为环境原子。弛豫后的原型镁铝尖晶石 MgAl<sub>2</sub>O<sub>4</sub> 几何构型信息被用于构建本文研究的尖晶石结构的 CE 特征时的基础结构。使用固定的原型结构信息构造 CE 特征可使机器学习模型更易于预测和迁移，而不受个别结构弛豫后几何构型变化的影响，从而避免对新未知结构进行昂贵的第一性原理计算。

### 1.3 机器学习预测形成能和晶格常数

本研究采用随机森林算法训练机器学习模型。采用前述的 CE 特征为机器学习模型的输入变量，预测性能包括尖晶石结构的形成能和晶格常数等。选取所有数据的 80% 作为训练集，采用 5 折交叉验证，通过网格搜索确定最佳超参数。剩余 20% 数据作为完全独立的性能测试集，并不参与模型训练。以决定系数 R<sup>2</sup> (Coefficient of determination)、平均绝对误差 MAE (Mean Absolute Error) 和均方根误差 RMSE (Root Mean Squared Error) 作为机器学习模型性能的评价指标，具体定义分别如下：

$$R^2 = 1 - \frac{\sum_{i=1}^N (y_i - \hat{y}_i)^2}{\sum_{i=1}^N (y_i - \bar{y})^2} \quad (7)$$

$$MAE = \frac{1}{N} \sum_{i=1}^N |(y_i - \hat{y}_i)| \quad (8)$$

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2} \quad (9)$$

其中， $\hat{y}_i$  是单个样本  $y_i$  的预测值， $\bar{y}$  是所有样本的预测值的平均值， $y_i$  是实验或计算获得的目标值。

## 2 结果与讨论

### 2.1 不同 CE 模型的预测结果

本工作中研究了 8 个基于不同组合方式构建的 CE 机器学习模型，其中包括四个 sCE 模型 CE-1、CE-3、CE-6、CE-7 和四个 dCE 模型 CE-2、CE-4、CE-5、CE-8（图 1）。这些 CE 模型对形成能预测的  $R^2$  和 RMSE 如表 2 所示。

**表 2** 基于 CE-1 到 CE-8 的八种特征模型的随机森林机器学习方法预测尖晶石氧化物形成能测试集的相关系数  $R^2$  和 RMSE。环境原子选取了从中心原子的第一近邻到第十近邻。收敛的结果以粗体显示。

**Table 2** Correlation coefficient  $R^2$  and RMSE (eV/atom) of the prediction of formation energies in the testing data sets by the eight machine learning methods using Random Forest based on CE-1 to CE-8 feature models where the environment atoms defined as from the nearest-neighbor to the 10<sup>th</sup> nearest neighbor atoms of the center T and O sites of spinel oxides. The converged results are shown in bold fonts.

$R^2$	1	2	3	4	5	6	7	8	9	10
CE-1	0.129	<b>0.913</b>	<b>0.895</b>	<b>0.881</b>	<b>0.884</b>	<b>0.881</b>	<b>0.897</b>	<b>0.876</b>	<b>0.877</b>	<b>0.876</b>
CE-2	<b>0.971</b>	<b>0.973</b>	<b>0.974</b>	<b>0.974</b>	<b>0.975</b>	<b>0.975</b>	<b>0.975</b>	<b>0.975</b>	<b>0.975</b>	<b>0.976</b>
CE-3	0.771	0.768	0.771	<b>0.962</b>	<b>0.962</b>	<b>0.961</b>	<b>0.961</b>	<b>0.962</b>	<b>0.961</b>	<b>0.961</b>
CE-4	0.784	<b>0.972</b>	<b>0.972</b>	<b>0.963</b>	<b>0.962</b>	<b>0.963</b>	<b>0.968</b>	<b>0.969</b>	<b>0.962</b>	<b>0.966</b>
CE-5	0.789	<b>0.963</b>	<b>0.961</b>	<b>0.931</b>	<b>0.927</b>	<b>0.924</b>	<b>0.938</b>	<b>0.946</b>	<b>0.927</b>	<b>0.932</b>
CE-6	0.123	<b>0.971</b>	<b>0.97</b>	<b>0.97</b>	<b>0.97</b>	<b>0.971</b>	<b>0.975</b>	<b>0.975</b>	<b>0.969</b>	<b>0.968</b>
CE-7	0.769	0.772	0.77	<b>0.971</b>	<b>0.972</b>	<b>0.972</b>	<b>0.973</b>	<b>0.97</b>	<b>0.971</b>	<b>0.972</b>
CE-8	<b>0.971</b>	<b>0.97</b>	<b>0.97</b>	<b>0.975</b>	<b>0.975</b>	<b>0.974</b>	<b>0.976</b>	<b>0.977</b>	<b>0.974</b>	<b>0.976</b>
RMSE	1	2	3	4	5	6	7	8	9	10
CE-1	0.797	<b>0.252</b>	<b>0.274</b>	<b>0.294</b>	<b>0.29</b>	<b>0.294</b>	<b>0.272</b>	<b>0.3</b>	<b>0.297</b>	<b>0.3</b>
CE-2	<b>0.145</b>	<b>0.14</b>	<b>0.137</b>	<b>0.137</b>	<b>0.134</b>	<b>0.134</b>	<b>0.135</b>	<b>0.135</b>	<b>0.134</b>	<b>0.132</b>
CE-3	0.408	0.41	0.406	<b>0.166</b>	<b>0.166</b>	<b>0.167</b>	<b>0.166</b>	<b>0.166</b>	<b>0.167</b>	<b>0.168</b>
CE-4	0.396	<b>0.142</b>	<b>0.142</b>	<b>0.163</b>	<b>0.166</b>	<b>0.164</b>	<b>0.151</b>	<b>0.149</b>	<b>0.166</b>	<b>0.158</b>
CE-5	0.391	<b>0.164</b>	<b>0.169</b>	<b>0.222</b>	<b>0.229</b>	<b>0.235</b>	<b>0.211</b>	<b>0.196</b>	<b>0.23</b>	<b>0.221</b>
CE-6	0.794	<b>0.145</b>	<b>0.148</b>	<b>0.146</b>	<b>0.147</b>	<b>0.145</b>	<b>0.134</b>	<b>0.135</b>	<b>0.151</b>	<b>0.152</b>
CE-7	0.408	0.407	0.407	<b>0.144</b>	<b>0.143</b>	<b>0.143</b>	<b>0.14</b>	<b>0.147</b>	<b>0.144</b>	<b>0.143</b>
CE-8	<b>0.144</b>	<b>0.147</b>	<b>0.148</b>	<b>0.135</b>	<b>0.135</b>	<b>0.136</b>	<b>0.133</b>	<b>0.13</b>	<b>0.136</b>	<b>0.132</b>

通过对 4 个单中心 sCE 模型的预测性能进行比较，发现标量特征模型 CE-1[ $T_C + T_E$ ] 预测的 RMSE 在第 2 近邻处收敛到 0.252 eV/atom，而 CE-3[ $O_C + O_E$ ] 预测的 RMSE 在第 4 近邻处收敛于 0.166 eV/atom。表明 T 位点标量特征模型的收敛速度比 O 位点标量特征模型快，但是预测精度后者较高。二维矢量特征模型 CE-6[ $T_C, T_E$ ] 的 RMSE 收敛于第 2 近邻，而 CE-7[ $O_C, O_E$ ] 在第 4 近邻处收敛，两者 RMSE 均约为 0.145 eV/atom。对于矢量特征形式，T 位点模型的收敛速度仍然比 O 位点模型快，但两者达到了相似的精度，且均优于标量特征模型。综合考虑收敛性和准确性，CE-6 是最佳的单中心 sCE 模型。

通过对 4 个双中心 dCE 模型的预测性能进行比较，可见标量特征模型 CE-5[ $T_C + T_E + O_C + O_E$ ] 的 RMSE 在第 2 近邻处收敛于 0.164 eV/atom，矢量特征模型 CE-2、CE-4 和 CE-8 均达到了约 0.145 eV/atom 的预测精度，均优于标量特征模型。二维矢量特征模型 CE-2[ $T_C + O_C, T_E + O_E$ ] 收敛于第 1 近邻，而 CE-4[ $T_C + T_E, O_C + O_E$ ] 收敛于第 2 近邻。表明将中心原子与环境原子分离的收敛速度比将 T 位原子与 O 位原子分离的速度更快。四维矢量 dCE 特征模型 CE-8 [  $T_C, T_E, O_C, O_E$  ] 收敛于第一近邻，类似于 CE-2 [  $T_C + O_C, T_E + O_E$  ]。因此，根据收敛性和准确性结果，CE-2 [  $T_C + O_C, T_E + O_E$  ] 和 CE-8 [  $T_C, T_E, O_C, O_E$  ] 在所有 8 个特征模型中均具有相似的最佳性能。但是，四维特征模型 CE-8 [  $T_C, T_E, O_C, O_E$  ] 同时区分了中心原子和环境原子以及 T 和 O 位点，其优点是可以避免由于交换 T 和 O 位点异种元素产生的特征简并。但在二维特征模型 CE-2 [  $T_C + O_C, T_E + O_E$  ] 中会出现这种特征兼并情况。

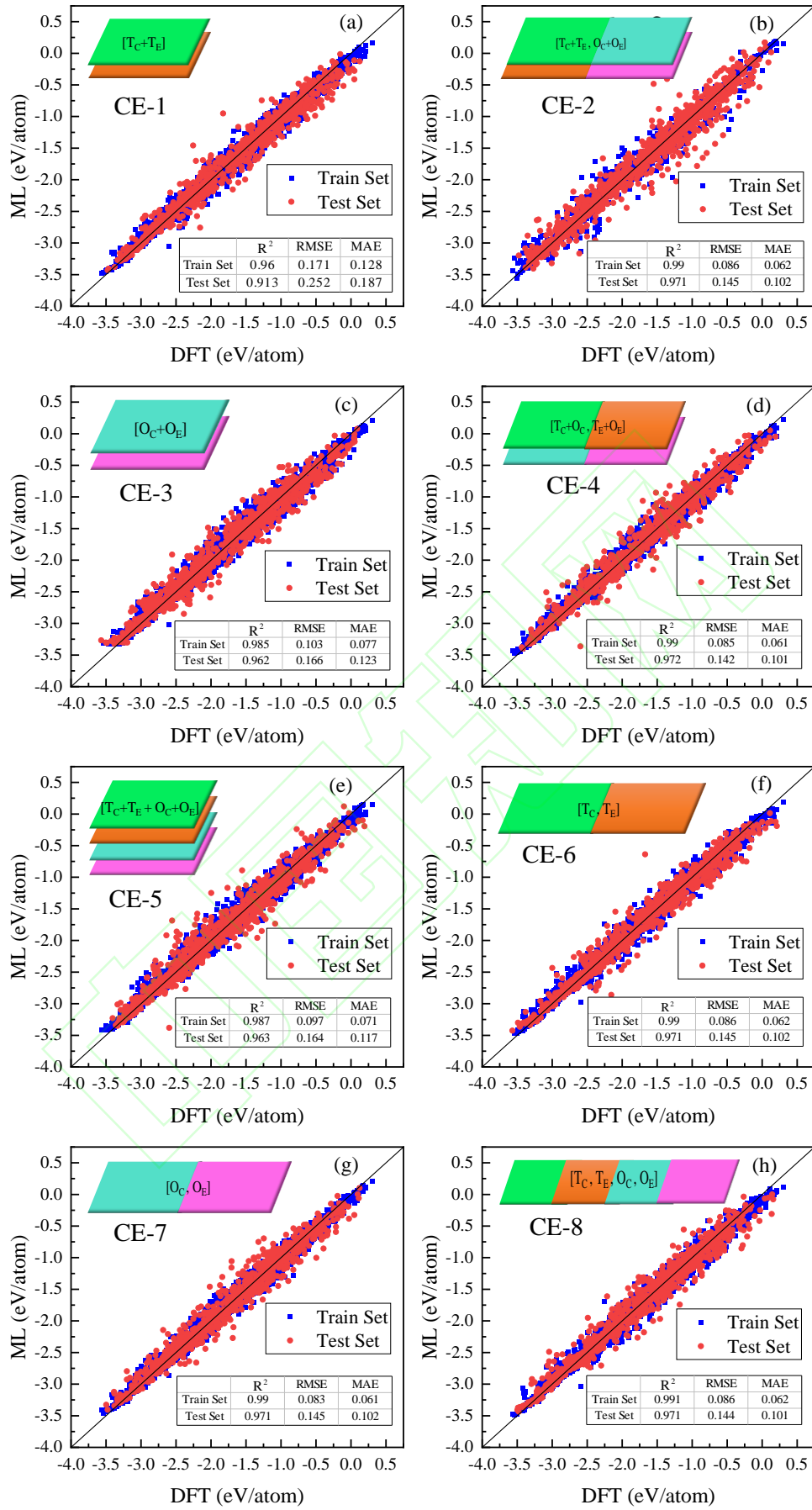


图 3 基于 CE-1 至 CE-8 特征的随机森林机器学习模型预测的形成能与 DFT 计算结果的比较。

**Fig. 3** Comparison of formation energy between the DFT calculations and machine learning predictions by Random Forest algorithm using CE-2 to CE-8 CE feature models, respectively.

本工作中，四维和二维矢量 CE 特征模型具有相似的预测精度，这可能是由于很多 T 和 O 位点占据元素的稳定性具有近似交换不变的对称性。如果这种位点占据稳定性的交换对称被破坏，四维矢量特征模型应该比二维矢量特征模型具有更准确的描述能力和预测准确性。另一方面，如果训练数据数量有限，比二维矢量特征模型更复杂的四维矢量特征模型可能更容易导致过拟合。因此，CE 模型的实际应用中应根据研究结构的复杂性与可用数据集的大小进行折衷选择矢量特征模型的维数。如果结构复杂并且数据集很大，高维 CE 模型在准确性和可迁移性方面表现会更好，否则，低维 CE 模型就可能达到足够高的预测精度，而且避免过拟合。

所有 8 种 CE 特征模型对晶格常数预测的收敛行为类似于形成能，如表 3 所示。但是，对于晶格常数大于 0.7 nm 的结构的预测精度不如其它结构，这可能是由于晶格常数大于 0.7 nm 的结构数量较少，从而导致机器学习模型不能对这类结构进行很好的学习。综合考虑预测精度和收敛速度，CE-8 特征模型预测的 RMSE 在最近邻处达到了 0.087 Å，在研究的晶格常数预测模型中表现最佳。

**表 3** 基于 CE-1 到 CE-8 的八种特征模型的随机森林机器学习方法预测尖晶石氧化物晶格常数测试集的相关系数  $R^2$  和 RMSE。环境原子选取了从中心原子的第一近邻到第十近邻。收敛的结果以粗体显示  
**Table 3** Correlation coefficient  $R^2$  and RMSE (eV/atom) of the prediction of lattice constant in the testing data sets by the eight machine learning methods using Random Forest based on CE-1 to CE-8 feature models where the environment atoms defined as from the nearest-neighbor to the 10<sup>th</sup> nearest neighbor atoms of the center T and O sites of spinel oxides. The converged results are shown in bold fonts

$R^2$	1	2	3	4	5	6	7	8	9	10
CE-1	0.153	<b>0.84</b>	<b>0.818</b>	<b>0.808</b>	<b>0.79</b>	<b>0.777</b>	<b>0.816</b>	<b>0.806</b>	<b>0.796</b>	<b>0.79</b>
CE-2	<b>0.95</b>	<b>0.95</b>	<b>0.952</b>	<b>0.946</b>	<b>0.947</b>	<b>0.949</b>	<b>0.948</b>	<b>0.948</b>	<b>0.945</b>	<b>0.947</b>
CE-3	0.758	0.756	0.754	<b>0.929</b>	<b>0.929</b>	<b>0.925</b>	<b>0.926</b>	<b>0.924</b>	<b>0.927</b>	<b>0.924</b>
CE-4	0.782	<b>0.949</b>	<b>0.949</b>	<b>0.935</b>	<b>0.938</b>	<b>0.937</b>	<b>0.943</b>	<b>0.942</b>	<b>0.937</b>	<b>0.94</b>
CE-5	0.778	<b>0.939</b>	<b>0.933</b>	<b>0.905</b>	<b>0.897</b>	<b>0.89</b>	<b>0.913</b>	<b>0.926</b>	<b>0.894</b>	<b>0.907</b>
CE-6	0.151	<b>0.951</b>	<b>0.95</b>	<b>0.941</b>	<b>0.945</b>	<b>0.945</b>	<b>0.948</b>	<b>0.945</b>	<b>0.94</b>	<b>0.939</b>
CE-7	0.759	0.755	0.756	<b>0.948</b>	<b>0.947</b>	<b>0.949</b>	<b>0.949</b>	<b>0.941</b>	<b>0.946</b>	<b>0.945</b>
CE-8	<b>0.949</b>	<b>0.951</b>	<b>0.951</b>	<b>0.948</b>	<b>0.945</b>	<b>0.947</b>	<b>0.948</b>	<b>0.947</b>	<b>0.951</b>	<b>0.95</b>
RMSE	1	2	3	4	5	6	7	8	9	10
CE-1	0.358	<b>0.155</b>	<b>0.165</b>	<b>0.168</b>	<b>0.177</b>	<b>0.181</b>	<b>0.165</b>	<b>0.17</b>	<b>0.174</b>	<b>0.177</b>
CE-2	<b>0.086</b>	<b>0.087</b>	<b>0.085</b>	<b>0.09</b>	<b>0.089</b>	<b>0.087</b>	<b>0.089</b>	<b>0.088</b>	<b>0.091</b>	<b>0.088</b>
CE-3	0.191	0.19	0.191	<b>0.103</b>	<b>0.103</b>	<b>0.105</b>	<b>0.105</b>	<b>0.106</b>	<b>0.104</b>	<b>0.105</b>
CE-4	0.179	<b>0.087</b>	<b>0.086</b>	<b>0.098</b>	<b>0.097</b>	<b>0.097</b>	<b>0.092</b>	<b>0.092</b>	<b>0.098</b>	<b>0.094</b>
CE-5	0.182	<b>0.095</b>	<b>0.1</b>	<b>0.119</b>	<b>0.124</b>	<b>0.127</b>	<b>0.114</b>	<b>0.104</b>	<b>0.125</b>	<b>0.118</b>
CE-6	0.357	<b>0.085</b>	<b>0.086</b>	<b>0.093</b>	<b>0.09</b>	<b>0.091</b>	<b>0.088</b>	<b>0.09</b>	<b>0.094</b>	<b>0.095</b>
CE-7	0.19	0.19	0.19	<b>0.087</b>	<b>0.089</b>	<b>0.088</b>	<b>0.087</b>	<b>0.093</b>	<b>0.089</b>	<b>0.09</b>
CE-8	<b>0.087</b>	<b>0.085</b>	<b>0.086</b>	<b>0.087</b>	<b>0.09</b>	<b>0.089</b>	<b>0.088</b>	<b>0.089</b>	<b>0.085</b>	<b>0.086</b>



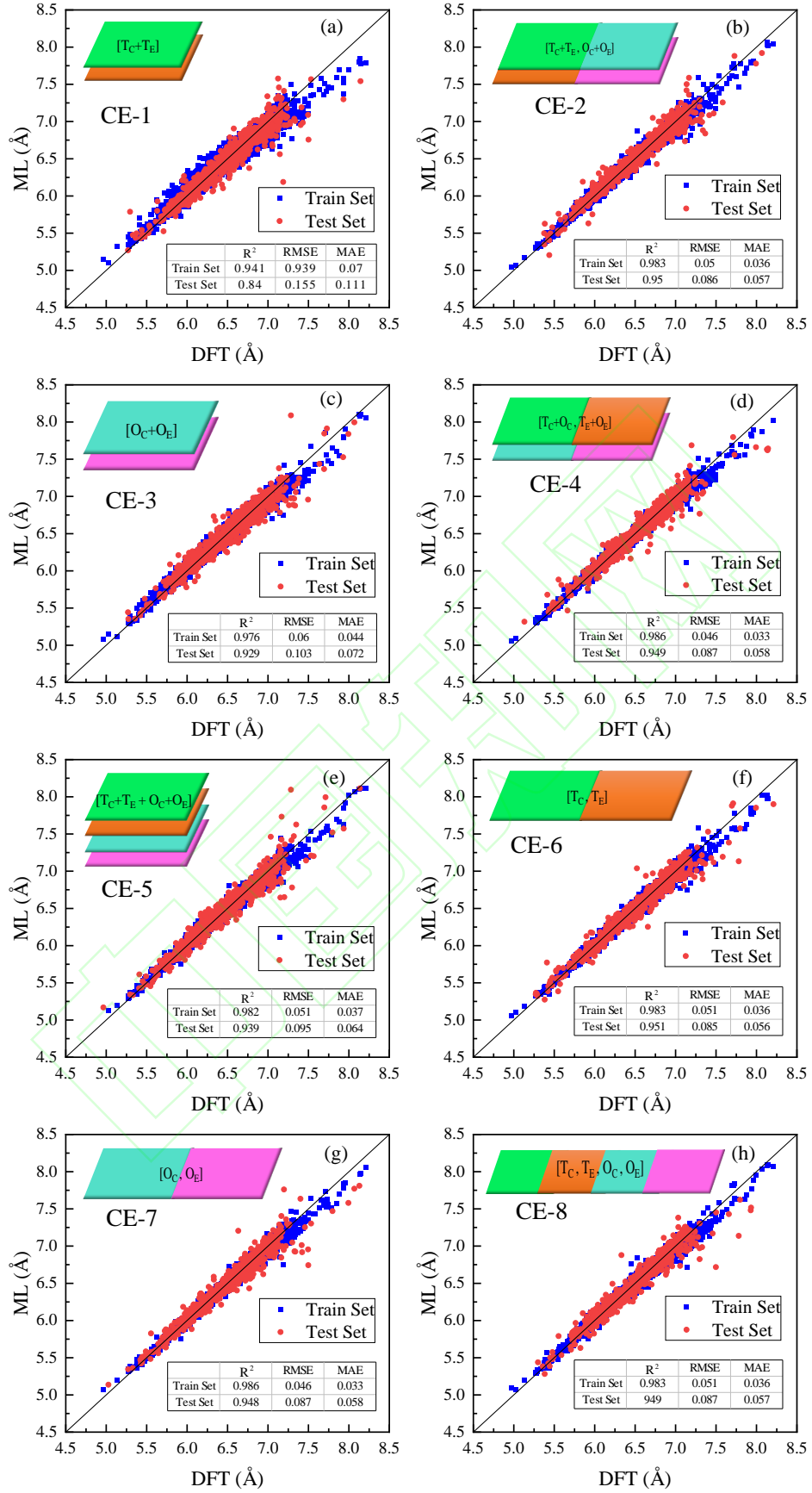


图 4 基于 CE-1 至 CE-8 特征的随机森林机器学习模型预测的晶格常数与 DFT 计算结果的比较。

Fig. 4 Comparison of lattice constants between the DFT calculations and machine learning predictions by Random Forest algorithm using CE-2 to CE-8 CE feature models, respectively.

## 2.2 预测结果统计直方图

综合考虑晶格常数预测结果的精度和收敛速度, CE-8 特征模型在最近邻处预测结果最好。先前的工作中我们对 5329 种尖晶石氧化物结构的形成能和晶格常数的 DFT 计算结果进行了统计分析, 本工作将 CE-8 模型在最近邻处的机器学习预测结果统计如图 5 所示。5329 种尖晶石氧化物结构中有 141 个文献报道的已经被实验合成的结构<sup>[19]</sup>。本工作对这些实验结构进行 DFT 计算, 发现其形成能在  $-2.999$  eV/atom 和  $-0.633$  eV/atom 之间。因此将本工作研究的尖晶石结构分为三类: 形成能大于  $-2.999$  eV/atom 的较稳定(More Stable, MS)结构、形成能介于  $-2.999$  eV/atom 与  $-0.633$  eV/atom 之间的稳定(Stable, S)结构以及形成能小于  $-0.633$  eV/atom 的较不稳定(Less Stable, LS)结构。如图 5(a)所示, 本工作研究的结构可划分成 361 个 MS 结构、4347 个 S 结构和 466 个 LS 结构。此外, 有 155 个结构几何弛豫不收敛, 这些结构主要是含有活性金属 Cs 和重稀土元素, 如 Eu、Dy、Er 等。形成能衡量物质结合强度, 可作为判断物质是否热力学稳定的重要判据之一。本工作 DFT 计算预测形成能最大、热力学最稳定的前 20 个尖晶石氧化物按稳定性降低顺序分别是  $\text{YbLu}_2\text{O}_4$ 、 $\text{YbTm}_2\text{O}_4$ 、 $\text{YbEr}_2\text{O}_4$ 、 $\text{YbSc}_2\text{O}_4$ 、 $\text{CaLu}_2\text{O}_4$ 、 $\text{YbHo}_2\text{O}_4$ 、 $\text{YbY}_2\text{O}_4$ 、 $\text{CaTm}_2\text{O}_4$ 、 $\text{YbDy}_2\text{O}_4$ 、 $\text{SrLu}_2\text{O}_4$ 、 $\text{CaEr}_2\text{O}_4$ 、 $\text{YbTb}_2\text{O}_4$ 、 $\text{HfYb}_2\text{O}_4$ 、 $\text{CaHo}_2\text{O}_4$ 、 $\text{SrTm}_2\text{O}_4$ 、 $\text{ThYb}_2\text{O}_4$ 、 $\text{CaSc}_2\text{O}_4$ 、 $\text{EuHo}_2\text{O}_4$ 、 $\text{YbGd}_2\text{O}_4$  和  $\text{CaY}_2\text{O}_4$ 。这些 DFT 计算预测的最稳定尖晶石结构的形成能范围是  $(-3.57$  eV/atom,  $-3.38$  eV/atom), 比本工作研究的实验结构的形成能范围  $(-2.999$  eV/atom,  $-0.633$  eV/atom) 大很多, 所以这些预测的稳定尖晶石结构更有可能被实验合成。本工作预测结果为将来实验合成新型的稳定尖晶石结构指明了方向。

机器学习预测的尖晶石氧化物的晶格常数[图 5(b)]也呈正态分布, 绝大多数结构的晶格常数处于  $0.6$  nm 与  $0.7$  nm 之间。

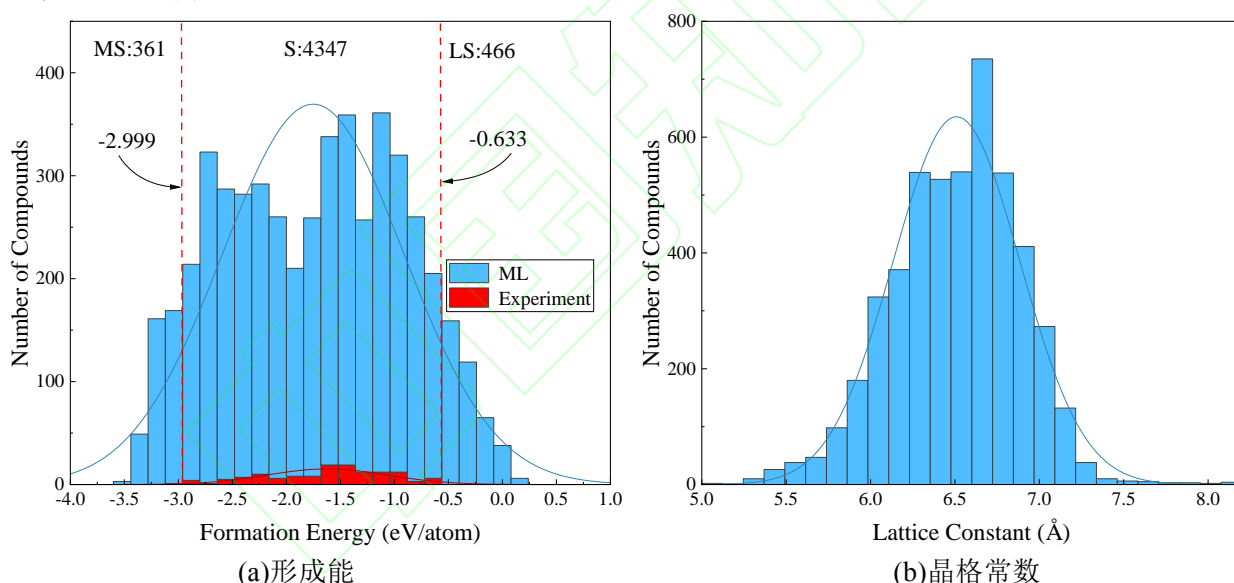


图 5 尖晶石氧化物形成能和晶格参数的机器学习预测结果的统计分布直方图。141 个已知的实验结构的形成能在 (a) 中以红色表示。

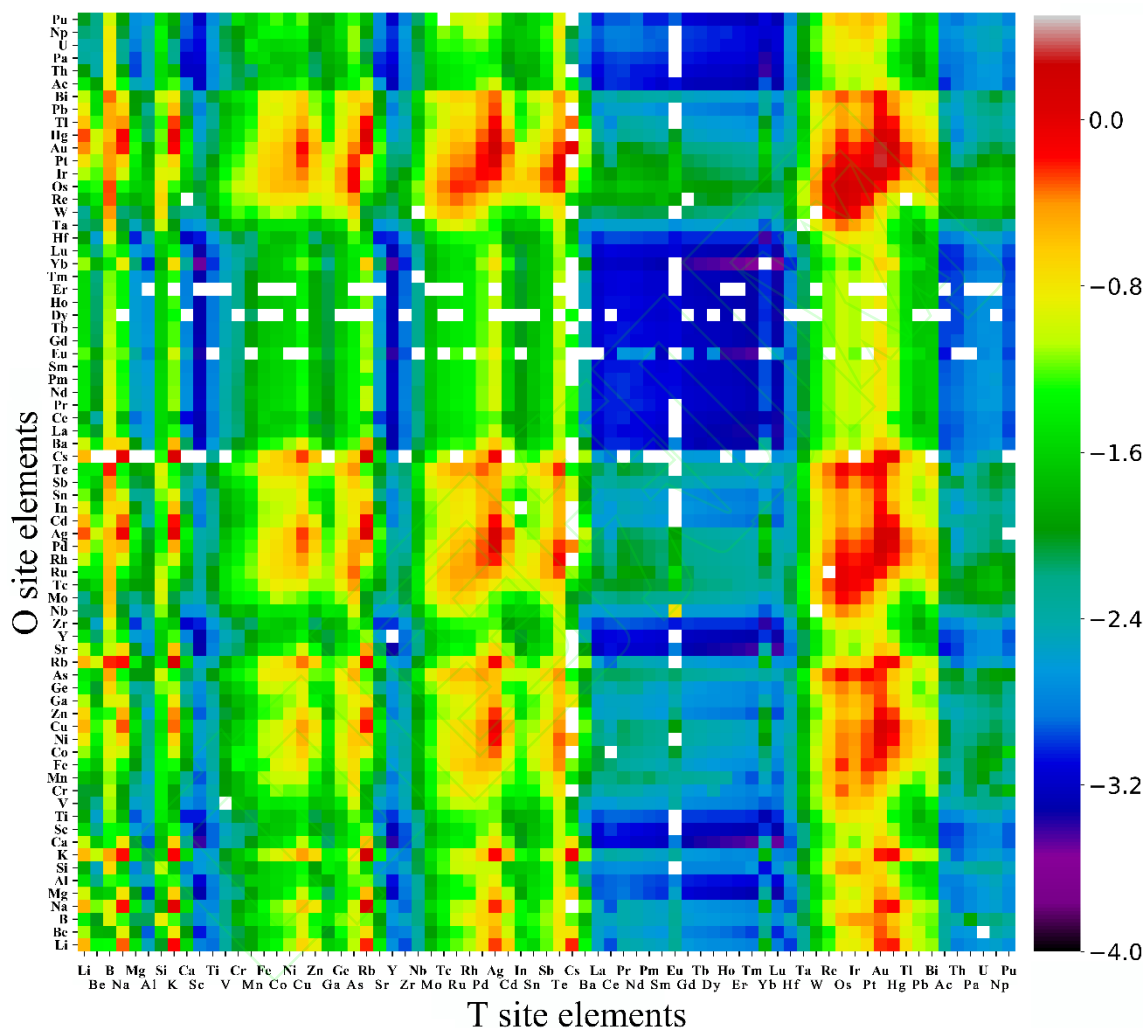
Fig. 5 Statistical distribution histogram of the machine learning prediction results of spinel oxide formation energy and lattice parameters. The formation energies of the 141 known experiment structures are shown in red in (a)

## 2.3 机器学习预测的形成能和晶格常数热图

机器学习预测的形成能分布如图 6 所示, 横轴和纵轴按照 T 位点和 O 位点组成元素的原子序数排序。根据形成能的范围定义的 MS、S 和 LS 结构在热图中分别大致对应蓝色、绿色和橘红色区域。这三种区域呈现出局部聚集的块状, 而非稀疏的点样分布, 表明具有相近原子序数的结构元素通常形成具有相似稳定性的尖晶石氧化物。此外, 除少数情况外, 大多数聚集的块呈对角对称分布, 这表明将 T 或 O 位点元素替换为邻近的元素可以形成稳定性相似的结构, 即稳定性的位点互换不变对称性。这可以解释为什么会有存在

很多稳定的反尖晶石和混合尖晶石氧化物。

为便于指导实验合成新型稳定的尖晶石，本工作分别针对 T 和 O 位点将 73 个置换元素定性地划分为与稳定性有关的“好”和“坏”元素（图 2）。如图 6 所示，T 和 O 位点的稳定性“好”元素组合可形成更稳定的(MS)结构，显示为蓝色区域；在橘红色区域，T 和 O 位点的稳定性“坏”元素组合形成了更不稳定的(LS)结构；如果在 T 和 O 位点分别存在稳定性“好”元素和“坏”元素，则形成一般的稳定(S)结构，对应绿色区域，其稳定性与已知的实验结构相当。如图 2 和图 6 所示，稳定性“好”和“坏”元素都呈现出 T/O 位点的对称稳定性，即大多数元素对于 T 和 O 位点同时都是“好”的或是“坏”的稳定性元素。但也有一些少数例外，例如，Ru 是“好”的 T 位点稳定性元素，但却是“坏”的 O 位点稳定性元素。另一方面，Sc、Hf、Pa、U 和 Np 是 O 位点的稳定性“好”元素，但却是 T 位点的稳定性“坏”元素。



**图 6** 机器学习（CE-8）预测的尖晶石结构的形成能热图，其中 T 位和 O 位点的 73 个组成元素按照原子序数排序。图中根据形成能的大小，较稳定(MS)、稳定(S)和较不稳定(LS)结构分别大致对应蓝色、绿色和橘红色区域。白色区域是计算不收敛的结构

**Fig. 6** Heat map of formation energies of spinel oxide structures predicted by ML, containing 73 constitution elements at the T sites and O sites, respectively, sorted by atom number. The more stable (MS), stable (S), and less stable (LS) structures are roughly shown in blue, green, and red regions, respectively, according to the formation energies. The white squares are the structures that were not converged computationally

机器学习预测的点阵常数分布如图 7 所示。从图 7 中可知 IA 族的碱金属元素，如 Li、Na、K、Rb、Cs，相比于同周期的其他元素通常会形成相对较大晶格参数的结构。这些碱性元素在同周期元素中具有最

小的离子尺寸和电负性，因此在尖晶石结构中它们的电子更容易被邻近的氧原子得到。带更大负电性的氧原子间的斥力使得碱金属尖晶石氧化物晶格膨胀。含有原子质量较小的碱金属元素尖晶石氧化物同时具有较大的晶格参数，说明其密度比较低，有作为轻质材料的应用潜力。此外，碱金属尖晶石氧化物较大的晶格参数可以促进碱金属离子的快速扩散，是作为二次离子电池电极的良好候选。确实， $\text{LiMn}_2\text{O}_4$  和  $\text{NaMn}_2\text{O}_4$  尖晶石氧化物已经作为正极材料在锂离子和钠离子电池中获得了应用<sup>[20]</sup>。

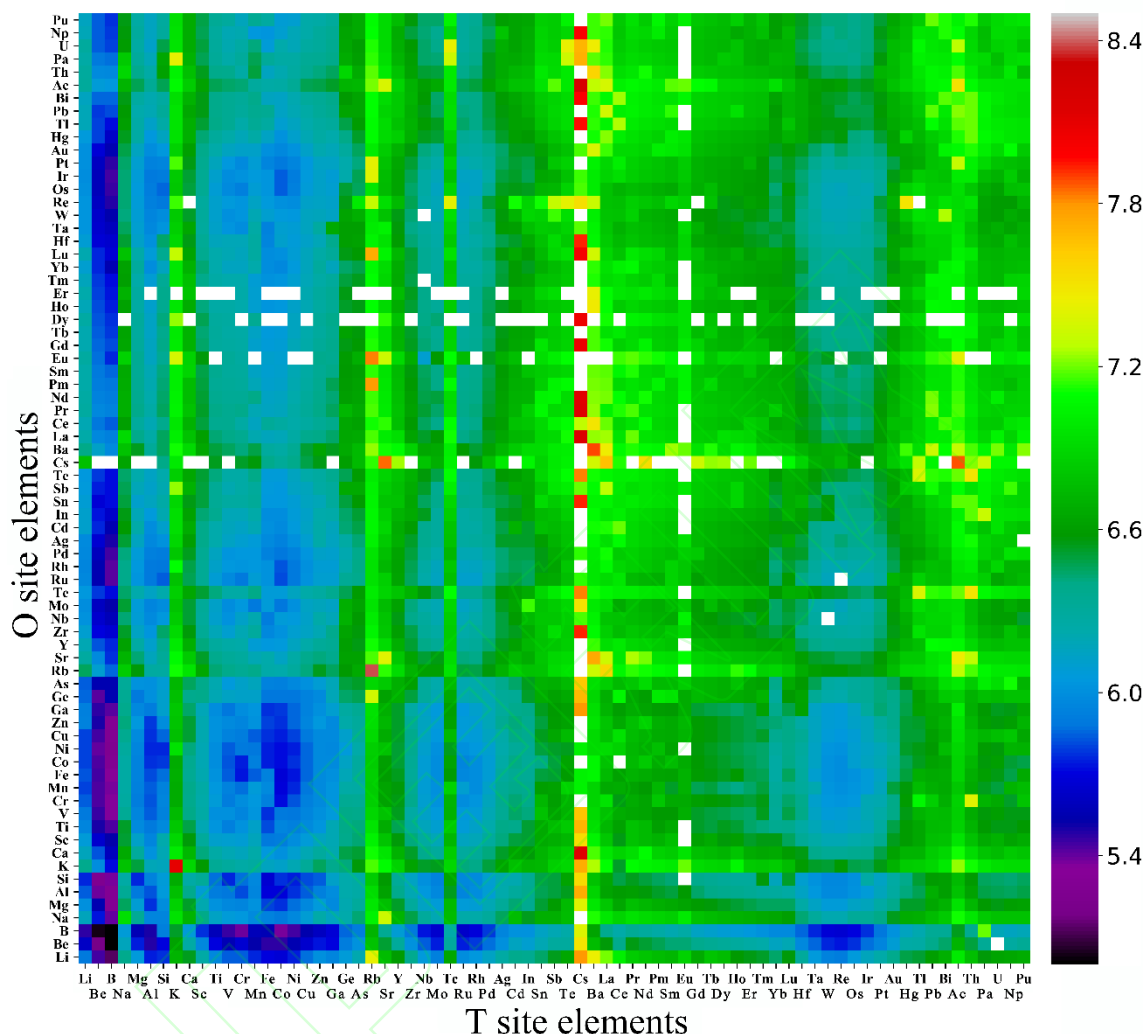


图 7 机器学习（CE-8）预测的尖晶石结构的晶格常数分布热图，其中 T 位和 O 位点的 73 个组成元素按照原子序数排序。白色区域是计算不收敛的结构。

**Fig. 7** Heat map of lattice constant of spinel oxide structures predicted by ML, containing 73 constitution elements at the T sites and O sites, respectively, sorted by atom number. The white squares are the structures that were not converged computationally.

## 2.4 计算和机器学习预测结果与被实验合成的尖晶石氧化物的性质比较

为了进一步独立测试验证本工作结果准确性，我们将文献报道的被实验合成的 141 个尖晶石氧化物结构的晶格常数的实验值分别与 DFT 计算值和随机森林机器学习预测值进行比较，结果如图 8 所示。通过与实验点阵常数比较[图 8(a)]，我们发现机器学习预测点阵常数的 RMSE 误差为  $0.04 \text{ \AA}$ ，比 DFT 计算的 RMSE 误差  $0.09 \text{ \AA}$  更低，进一步证实了机器学习模型的预测有效性。表 4 和表 5 分别是部分已知尖晶石氧化物结构的形成能和带隙的实验值、DFT 计算值以及机器学习预测值。DFT 计算值与部分实验值的偏差可能来源于交换相关密度泛函的近似。由于 DFT 计算值作为目标值进行机器学习训练，所以机器学习和 DFT 计算结果比较一致。



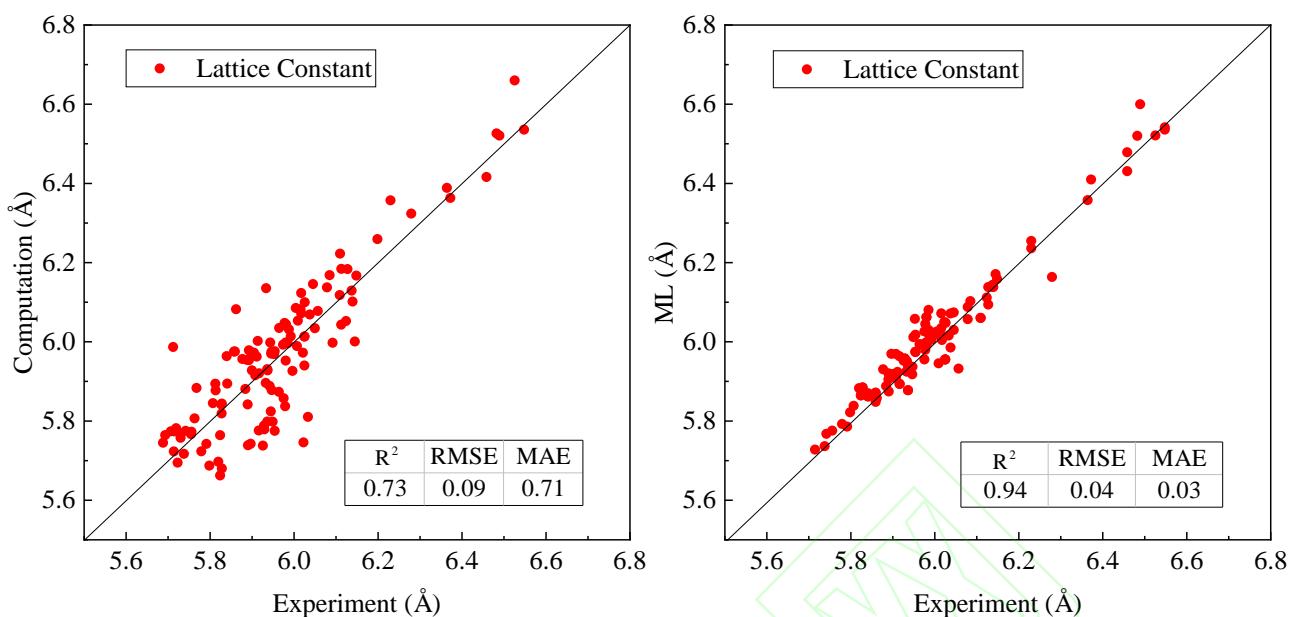


图 8 实验合成的 141 个尖晶石结构的晶格常数的实验值与(a)DFT 计算值和(b)机器学习预测值比较

Fig. 8 Experimental lattice constants of 141 spinel structures compared with the results calculated by DFT (a) and predicted by machine learning (b), respectively

表 4 部分已知尖晶石氧化物形成能的实验值、DFT 计算值和机器学习预测值(eV)

Table 4 Experimental value, DFT calculation value, and machine learning prediction value (eV) of formation energy of spinel oxides

Structure	Experiment	DFT	ML
MoLi <sub>2</sub> O <sub>4</sub>	-2.25	-2.20	-1.97
SiMg <sub>2</sub> O <sub>4</sub>	-0.05	-2.78	-1.87
CoCo <sub>2</sub> O <sub>4</sub>	-0.57	-1.01	-1.03
FeFe <sub>2</sub> O <sub>4</sub>	-1.66	-1.29	-1.26

表 5 部分已知尖晶石氧化物带隙的实验值、DFT 计算值和机器学习预测值(eV)

Table 5 Experimental value, DFT calculation value, and machine learning prediction value (eV) of band gap of spinel oxides

Structure	Experiment	DFT	ML
TiMg <sub>2</sub> O <sub>4</sub>	0.25	4.41	3.95
MgAl <sub>2</sub> O <sub>4</sub>	5.55	5.13	4.53
ZnGa <sub>2</sub> O <sub>4</sub>	2.59	2.30	1.89
MgIn <sub>2</sub> O <sub>4</sub>	3.36	1.93	2.03

### 3 结论

本研究通过第一性原理计算和机器学习方法,系统地研究了包含 73 种 A/B 构成元素的 5329 种 AB<sub>2</sub>O<sub>4</sub> 正尖晶石氧化物。根据 141 个已知实验结构的形成能范围 (-3.0 eV/atom, -0.6 eV/atom),我们将预测结构分为 361 个更稳定 (MS) 结构、4347 个稳定 (S) 结构和 466 个较不稳定 (LS) 结构。为指导实验成分设计,我们提出了与尖晶石氧化物的稳定性定性相关的“好”和“坏”的构成元素。常见的“好”的稳定性元素是 II, III, IV 族元素和稀土元素,并大部分具有对称的占位稳定性。含 IA 族碱金属元素的

尖晶石晶体结构通常具有较大的晶格常数。本工作预测的稳定性尖晶石结构对进一步实验合成新型尖晶石氧化物具有指导意义。

本工作中我们研究了包含成分和结构信息的“中心-环境”(CE)机器学习特征模型,该模型能够高效预测尖晶石氧化物的能量和结构特性,并获得与第一性原理方法相当的精度。通过考察不同的特征构建方式,我们发现双中心CE模型(dCE)的预测性能通常优于单中心CE模型(sCE),并且应综合考虑结构的复杂性和可用数据的数量来选择复合特征向量的维数。如果结构复杂、对称性很低并且具有大量数据,双中心或多中心CE特征模型的矢量形式将比单中心CE模型更加准确可靠。本工作中的“中心-环境”(CE)特征模型可推广用于其它凝聚态材料的能量和结构的机器学习预测。

#### 参考文献:

- [1] XU Y L, WANG X M, LI X, et al. New materials band gap prediction based on the high-throughput calculation and the machine learning [J]. SCIENTIA SINICA Technological, 2018, 49(1): 44-54.
- [2] JAIN A, VOZNY O, SARGENT E H. High-throughput screening of lead-free perovskite-like materials for optoelectronic applications [J]. The Journal of Physical Chemistry C, 2017, 121(13): 7183-7187.
- [3] SUN S J, NOOR T P, Hartono, ZEKUN D R, et al. Accelerated development of perovskite-inspired materials via high-throughput synthesis and machine-learning diagnosis [J]. Joule, 2019, 3(6): 1437-1451.
- [4] LI G Z, HU Z, H F, et al. Machine learning enabled high-throughput screening of hydrocarbon molecules for the design of next generation fuels [J]. Fuel, 2020, 265(1):116968.
- [5] ABE O, HOSONO Y, SANO T, et al. A mechanochemical route to synthesize  $\text{LiMn}_2\text{O}_4$  as a cathode material for lithium batteries [J]. Transactions of the Materials Research Society of Japan, 2008, 33(4):941-944.
- [6] CUI X L, FENG H X, LIU J L, et al. Porous  $\text{LiMn}_2\text{O}_4$  nano-microspheres as durable high power cathode materials for lithium ion batteries. Russian Journal of Electrochemistry [J], 2019, 55(5): 351-357.
- [7] SUELLEN BATTISTON, CAROLINE RIGO, ERIC DA CRUZ SEVERO, et al. Synthesis of zinc aluminate ( $\text{ZnAl}_2\text{O}_4$ ) spinel and its application as photocatalyst [J]. Materials Research, 2014, 17(7): 734-738.
- [8] ZHENG T, ZHU H B, CROUÉ, et al. Production of sulfate radical from peroxydisulfate induced by a magnetically separable  $\text{CuFe}_2\text{O}_4$  spinel in water: efficiency, stability, and mechanism [J]. Environmental Science & Technology, 2013, 47(6): 2784-2791.
- [9] SUZUKI Y, VAN DOVER R B, GYORGY E M, et al. Structure and magnetic properties of epitaxial spinel ferrite thin films [J]. Applied Physics Letters, 1996, 68(5): 714-716.
- [10] JIANG W, HUANG H Q, LIU F, et al. Magnetic Weyl semimetals with diamond structure realized in spinel compounds [J]. Physical Review B, 2020, 101(12):121113-6.
- [11] 向勇, 谢道华. 尖晶石结构功能材料的新进展 [J]. 磁性材料及器件, 2001, 32(3): 21-26.
- [12] 郑弘泰. 第一原理理论计算作为科学发现的工具:磁铁矿(magnetite)之电荷-轨道秩序(charge-orbital ordering)和Verwey相变 [J]. 物理双月刊(中国台湾), 2005, 27(4):5.
- [13] 戴道生, 钱乾坤. 铁磁学 [M]. 北京: 科学图书出版社, 2000: 32-200.
- [14] KURT E. SICKAFUS, JOHN M. WILLS. Structure of spine [J]. Journal of the American Ceramic Society, 1999, 82(12):3279-3292.
- [15] ROBERT E. WARBURTON, HAKIM IDDIR, LARRY A. CURTISS, et al. Thermodynamic stability of low- and high-index spinel  $\text{LiMn}_2\text{O}_4$  surface terminations [J]. ACS Applied Materials & Interfaces, 2016, 8(17):11108-21.
- [16] GABRIEL R. SCHLEDER, CARLOS M A, ADALBERTO F. Exploring two-dimensional materials thermodynamic stability via machine learning [J]. ACS Applied Materials & Interfaces, 2020, 12(18):20149-20157.
- [17] KRESSE G, FURTHMÜLLER J. Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set [J]. Physical Review B, 1996, 54(16):11169-11186.
- [18] BAIKOV A A. Database on properties of chemical elements. Institute of metallurgy and materials science [Z]. RUSSIAN ACADEMY OF SCIENCES, 2020. <http://phases.imet-db.ru/elements/mendel.aspx?main=1>.
- [19] VILLARS P, CENZUAL K, OKAMOTO H, et al. Materials Platform for Data Science [Z]. Switzerland: Villars P, 2019. <https://mpds.io>.
- [20] LIU X Z, WANG X, AKIRA L, et al. High stable post-spinel  $\text{NaMn}_2\text{O}_4$  cathode of sodium ion battery [J]. Journal of Materials Chemistry A, 2014, 2(36):14822-14826.