

## Predicting the formation of fractionally doped perovskite oxides by a function-confined machine learning method

Ximei Zhai<sup>1,3</sup>, Fei Ding<sup>2,3</sup>, Zeyu Zhao<sup>1</sup>, Aaron Santomauro<sup>1</sup>, Feng Luo<sup>2</sup>✉ & Jianhua Tong<sup>1</sup>✉

Fractionally doped perovskites oxides (FDPOs) have demonstrated ubiquitous applications such as energy conversion, storage and harvesting, catalysis, sensor, superconductor, ferroelectric, piezoelectric, magnetic, and luminescence. Hence, an accurate, cost-effective, and easy-to-use methodology to discover new compositions is much needed. Here, we developed a function-confined machine learning methodology to discover new FDPOs with high prediction accuracy from limited experimental data. By focusing on a specific application, namely solar thermochemical hydrogen production, we collected 632 training data and defined 21 desirable features. Our gradient boosting classifier model achieved a high prediction accuracy of 95.4% and a high F1 score of 0.921. Furthermore, when verified on additional 36 experimental data from existing literature, the model showed a prediction accuracy of 94.4%. With the help of this machine learning approach, we identified and synthesized 11 new FDPO compositions, 7 of which are relevant for solar thermochemical hydrogen production. We believe this confined machine learning methodology can be used to discover, from limited data, FDPOs with other specific application purposes.

<sup>1</sup>Department of Materials Science and Engineering, Clemson University, Clemson, SC, USA. <sup>2</sup>School of Computing, Clemson University, Clemson, SC, USA. <sup>3</sup>These authors contributed equally: Ximei Zhai, Fei Ding. ✉email: [luofeng@clemson.edu](mailto:luofeng@clemson.edu); [jianhut@clemson.edu](mailto:jianhut@clemson.edu)

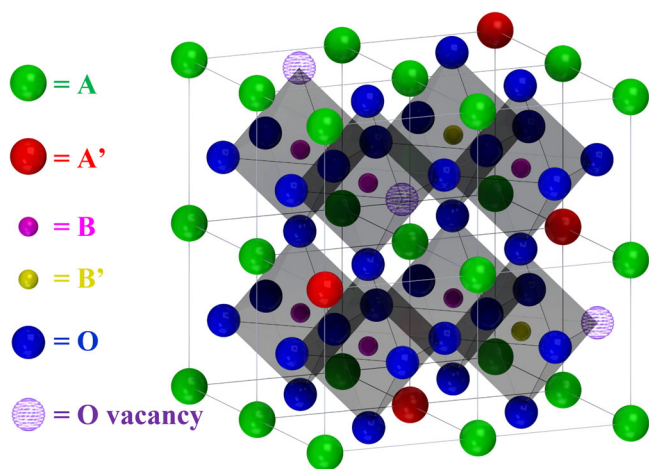
Perovskite materials derived from the mineral of  $\text{CaTiO}_3$  have the general formula  $\text{ABX}_3$ , in which A and B are cations, and X is an anion. Perovskites have a typical crystal structure shown in Fig. 1. The smaller B-site cation has six-fold coordination and forms a  $\text{BX}_6$  octahedron. Eight  $\text{BX}_6$  octahedra share corners to form a dodecahedron interstitial space for accommodating a twelve-fold coordinated larger A-site cation. The X is usually an oxygen, halogen or chalcogen anion, which can tolerate many vacancies. Perovskites have demonstrated ubiquitous applications to energy conversion/storage/harvesting, catalysis, sensor, superconductor, ferroelectric, piezoelectric, magnetic, and luminescence due to the numerous unique properties caused by their versatile compositions and flexible crystal structure symmetries.

Hybrid organic-inorganic perovskites (HOIPs) and inorganic perovskites (IPs) are the two main perovskites. HOIPs have a positively charged organic group (e.g., methylammonium, MA; formamidinium, FA) in A site, a metal cation (e.g., Pb, Sn) in B site, and a halogen anion (e.g., I, Cl, Br) in X site<sup>1</sup>. As semiconductors, HOIPs showed excellent optical and electrical properties and found extensive applications to solar cells<sup>2–5</sup>, photodetectors<sup>6</sup>, photocatalysis<sup>7</sup>, and light-emitting<sup>8</sup>. IPs usually denote perovskite oxides with metal cations in A and B sites and oxygen anions in the X site. Perovskite oxides allow fractionally doping multiple metals in both A and B sites and forming crystal structures with ~15 symmetries, enabling almost infinite compositions with unique properties. Instead of the simple  $\text{ABO}_3$  perovskite oxides, the fractionally doped perovskite oxides (FDPOs) with a generic formula of  $\text{A}_{x_1}^1\text{A}_{x_2}^2\cdots\text{A}_{x_m}^m\text{B}_{y_1}^1\text{B}_{y_2}^2\cdots\text{B}_{y_n}^n\text{O}_3$  ( $\text{A}^m$  and  $\text{B}^n$  are metal cations,  $m \geq 1$ ,  $n \geq 1$ ,  $x_1 + x_2 + \cdots + x_m = 1$ ,  $y_1 + y_2 + \cdots + y_n = 1$ ) have demonstrated pervasive applications. Mn and Sr doped  $\text{LaAlO}_3$  was the first experimentally proved perovskite oxide to efficiently split water and carbon dioxide to produce hydrogen and carbon monoxide based on the high-temperature and low-temperature solar thermochemical redox cycles<sup>9</sup>.  $\text{Ba}_{0.5}\text{Sr}_{0.5}\text{Co}_{0.8}\text{Fe}_{0.2}\text{O}_{3-\delta}$  showed extremely high performance as an oxygen-permeable membrane and the cathode of oxygen-ion conducting solid oxide fuel cells (O-SOFCs)<sup>10,11</sup>. Perovskite oxides of  $\text{BaCe}_{0.7}\text{Zr}_{0.1}\text{Y}_{0.1}\text{Yb}_{0.1}\text{O}_{3-\delta}$  and  $\text{BaCo}_{0.4}\text{Fe}_{0.4}\text{Zr}_{0.1}\text{Y}_{0.1}\text{O}_{3-\delta}$  demonstrated promising performance as an electrolyte and an oxygen electrode for protonic ceramic fuel/electrolysis cells (PCFCs and PCECs)<sup>12,13</sup>. Doped lanthanum gallate oxide ( $\text{La}_{1-x}\text{Sr}_x\text{Ga}_{1-y}\text{Mg}_y\text{O}_{3-\delta}$ ) showed high ionic conductivity at intermediate temperatures (~600 °C) as an O-SOFC electrolyte

material<sup>14</sup>.  $\text{YBa}_2\text{Cu}_3\text{O}_7$  has become a state-of-the-art superconductor for many years as one perovskite-derived structure<sup>15</sup>.  $\text{BaTiO}_3$  and  $\text{BaNi}_{0.5}\text{Nb}_{0.5}\text{O}_3$  showed promising ferroelectric and piezoelectric properties<sup>16,17</sup>. Perovskite manganites with a general formula  $\text{R}_{1-x}\text{A}_x\text{MnO}_3$  ( $\text{R}$  = rare-earth cation,  $\text{A}$  = divalent alkaline-earth cation) attracted much attention due to their colossal magnetoresistance and magnetocaloric effect<sup>18,19</sup>. Therefore, discovering new FDPOs with desired geometrical symmetries and properties for targeted applications has become one of the most important research topics in many materials science fields. However, FDPOs allow nearly 90% of the elements from the periodic table to come into the perovskite structures and fractionally doping, resulting in almost infinite complex compositions, which may be the targeted perovskite oxides. The prerequisite for discovering new perovskite oxides is to make quick and accurate predictions to determine if the designed compositions are perovskite or not before the time-consuming and expensive theoretical computation, experimental synthesis, and property characterization.

Establishing a link descriptor between composition and structure via basic elemental and simple oxide properties has attracted a significant effort to predict new perovskite oxides. Historically, simple geometrical factors were in use for predicting the formability of perovskite oxides. The Goldschmidt tolerance factor calculated from the ionic radii of A and B cations and oxygen anion ( $0.75 \leq t = \frac{(r_A + r_O)}{\sqrt{2}(r_B + r_O)} \leq 1.05$ ) was the most commonly used criterion to determine if a new composition was a perovskite or not<sup>20–24</sup>. Considering that the stable octahedron  $\text{BO}_6$  was the basic unit for perovskite structure, the octahedral factor ( $0.414 \leq \mu = r_B/r_O \leq 0.732$ ) defined as the ratio between B-site cation and oxygen anion was a prerequisite to forming the perovskite structure<sup>25,26</sup>. The combination of the octahedral factor and the tolerance factor showed improvement in the prediction accuracy<sup>21</sup>. Most recently, a new tolerance factor ( $\tau = \frac{r_O}{r_B} - n_A(n_A - \frac{r_A/r_B}{\ln(r_A/r_B)}) < 4.18$ ) (where  $n_A$  is the oxidation state of A,  $r_i$  is the ionic radius of ion  $i$ , and  $r_A > r_B$  by definition) integrated A-site cation oxidation state information with the geometry properties demonstrated an improved prediction accuracy for simple ( $\text{ABO}_3$ ) and double perovskite oxides ( $\text{A}_2\text{B}^1\text{B}^2\text{O}_6$ )<sup>27</sup>. To further improve the prediction accuracy, based on two-dimensional structure map technology, a series of scientific works also studied the effect of different feature pairs on the formability of simple perovskite oxides, which involved the ionic radii of A and B ions<sup>21</sup>, the bond length between A or B site and oxygen atom<sup>23</sup>, the tolerance factor and octahedral factor<sup>21,26</sup>. However, the two-dimensional structure map only utilized one feature pair to predict, which did not simultaneously consider more than two features to improve prediction accuracy further. Besides, first-principles density functional theory (DFT) calculations were also adopted to predict the formability of perovskite oxides. Emery et al. calculated the formation energies for 5,329  $\text{ABO}_3$  combinations generated by inserting 73 metals and semi-metals into A and B sites, respectively, which quickly predicted 395 thermodynamically stable simple perovskite oxides<sup>28</sup>. Although DFT calculation showed a promising future for predicting perovskite oxides, it still has not shown very successful predicting performance for the infinite FDPOs. Computing many complicated compositions needs more complex unit cells related to time-consuming and expensive DFT computations.

In recent years, the machine learning method has emerged as a powerful tool to accelerate new materials discovery. Several research groups applied machine learning techniques with DFT calculated results as training data to identify promising simple perovskite and double perovskite candidates<sup>29–32</sup>. Balachandran et al. used the random forest and gradient tree boosting classifiers



**Fig. 1** The schematic drawing of the crystal structure of fractionally doped perovskites. The chemical formula of fractionally doped perovskites is  $\text{A}_x\text{A}'_{1-x}\text{B}_y\text{B}'_{1-y}\text{O}_{3-\delta}$ , where  $1 \geq x \geq 0$ ,  $1 \geq y \geq 0$  and  $3-\delta$  represents oxygen nonstoichiometry.

to correctly classify 90% of their database as perovskites or non-perovskites using DFT as a source of training data<sup>33</sup>. Ye et al. utilized deep neural networks to predict the formation energy of ABO<sub>3</sub> perovskites using DFT training data<sup>34</sup>. However, most machine learning prediction work focused on the simple ABO<sub>3</sub> or double perovskites calculated from the DFT method. Utilizing the machine learning tool to predict new complex FDPOs based on experimental perovskite structure data has not been reported yet. The more reliable perovskite and non-perovskite experimental data are not ready yet. The extensive selection of features for predicting perovskite structures still needs significant effort. In the current work, we confined the FDPO study to a specific application of the solar thermochemical hydrogen (STCH) production for ensuring high-quality machine learning using a limited amount of data, serving as a case study for other functional FDPOs. We established a reasonable experimental database, selected 21 basic features, and screened six models to predict FDPOs. The verification based on literature data and the synthesis by solid-state reaction and modified Pechini method and XRD characterization proved the successful prediction of FDPOs using our model.

## Results

**Data collection.** The prerequisite to using machine learning to predict FDPOs is to collect enough reliable data for the given compositions with identified crystal structures, either perovskite or non-perovskite. Instead of the data derived from DFT calculation, the experimental data from Inorganic Crystal Structure Database (ICSD) and sporadic literature served as the crucial data sources. However, all data must be manually examined and compared to the original literature by the person with experience in FDPOs to ensure data reliability. Collecting a large quantity of data (e.g., more than 1000) takes too long or costs too much to perform efficiently. Furthermore, sometimes we don't have enough data available to train the model. Therefore, we confined the FDPO study to a specific application of the solar thermochemical hydrogen (STCH) production for ensuring high-quality machine learning using a limited amount of data, serving as a case study for other functional FDPOs.

We searched the ICSD database by inputting key STCH elements, setting the total element numbers, and downloaded eligible FDPO data experimentally obtained at room temperature under atmospheric pressure. Since the first experimental report of Sr<sub>x</sub>La<sub>1-x</sub>Al<sub>y</sub>Mn<sub>1-y</sub>O<sub>3-δ</sub> as excellent STCH materials with promising oxygen exchange capacity and high hydrogen production yield, FDPOs containing La, Sr, Mn, and O have become one of the most popular STCH material systems<sup>9</sup>. With La, Sr, Mn, and O as STCH key elements and 4, 5, 6, and 7 as the total element numbers, we downloaded 578 perovskite data. Furthermore, the most recent studies (e.g., CaTi<sub>1-x</sub>Fe<sub>x</sub>O<sub>3</sub><sup>35</sup> and BaCe<sub>x</sub>Mn<sub>1-x</sub>O<sub>3</sub><sup>36</sup>) indicated that some Ca, Ba, or Ti-doped FDPOs exhibited fast reaction kinetics as the potential STCH materials. So we also searched the ICSD database with Ca, Mn, and O as key STCH elements and 4, 5, and 6 as total element numbers, resulting in another 661 eligible perovskite data. In addition, using Ba and Ti as crucial elements and a total element number of 5, we got 166 eligible perovskite data. In summary, we downloaded around 1400 experimental FDPO data from the ICSD database by confining the specific application of STCH materials.

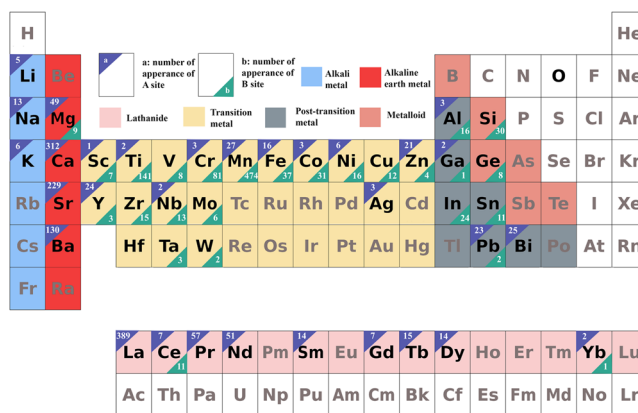
We established several rules to thoroughly clean up the raw data to improve the data's reliability. First, many repeated data in our data set resulted from the overlapped searching and the repeated reports. We manually checked the data set and removed all the repeated data points. Secondly, we were interested in the FDPOs, which were stable under room temperature and

atmospheric pressure, which urged us to rule out some thermodynamically metastable compounds from our data set. Furthermore, some compounds did not show the desired unity for A and B metal sites, which did not qualify them for reliability. We also removed some unrealistic perovskite data involved in radioactive elements or expensive rare earth elements by considering the safety and economic viability. While looking into original literature to verify the data reliability during the cleaning process, we found some perovskite oxides containing one or more elements among La, Sr, Mn, Ca, Ba, and Ti elements did not show in our data set. We treated these as reliable data and added them to our data set. Finally, we obtained a data set consisting of 516 FDPOs (Supplementary Data 1).

The collection of the non-perovskite compounds for the negative data point for machine learning was not a trivial work. We first searched the ICSD database for pyroxene, ilmenite, or other non-perovskite structures with a doped ABO<sub>3</sub> formula. The stability at room temperature under atmospheric pressure was also additional confinement. The inputting elements of these compositions were the same as the chosen perovskites, e.g., Ca, Ba, Mn, and Ti. We also double-checked these non-perovskites by tracing back to the original literature. Some mislabeled non-perovskite compounds in the ICSD database, which were confirmed to be perovskite by further studying literature, were removed from our non-perovskite data set. We obtained 96 non-perovskite data points from the ICSD database. The non-perovskites with mixed phases commonly occurred during the experimental synthesis, which did not show in the ICSD database. We extensively searched sporadic literature and collected another 20 non-perovskites with mixed phases. Finally, we obtained 116 non-perovskite data in our data set (Supplementary Data 1).

Figure 2 summarizes the chemical space of these compounds in our data set. Our STCH FDPO data set contains 31 elements (viz. Li, Na, Mg, Al, K, Ca, Sc, Ti, Cr, Mn, Fe, Co, Ni, Zn, Ga, Sr, Y, Nb, Ag, Ba, La, Ce, Pr, Nd, Sm, Gd, Tb, Dy, Yb, Pb, and Bi) in A site, 26 elements (viz. Mg, Al, Si, Sc, Ti, V, Cr, Mn, Fe, Co, Ni, Cu, Zn, Ga, Ge, Y, Zr, Nb, Mo, In, Sn, Ce, Yb, Ta, W, and Pb) in B site, and 41 elements in total. The numbers at the corner of each element show the perovskite compound numbers, which contain that element. The chemical space significantly expanded the key STCH elements we used to search perovskite compounds in the ICSD database, which provides us more opportunities to establish new potential PDFOs to serve as STCH materials for future study.

**Features selection and models selection.** The feature selection was initialized by summarizing the popular perovskite structure



**Fig. 2 The chemical space of the database.** The background colors indicate different subcategories and numbers exhibit the appearance frequency of different elements in A or B site.

descriptors commonly reported in the literature. Historically, the widely used descriptors for predicting the formability of perovskites were the Goldschmidt tolerance factor ( $t$ ) and the octahedral factor ( $\mu$ ) defined by the radii of the metal and oxygen ions<sup>21,37,38</sup>. Recently, the new tolerance factor ( $\tau$ ) further included the information of the oxidation state of A ( $n_A$ ) besides the ionic radii to improve the predictability<sup>27</sup>. These three features are closely related to the A and B elements' ionic radii and oxidation states. In addition, although it has not been studied extensively as a perovskite formation descriptor, the ratio between A-site cation to oxygen anion ( $r_A/r_O$ ) is a parallel factor to the octahedral factor. Therefore, we selected Goldschmidt tolerance factor ( $t$ ), octahedral factor ( $\mu$ ),  $r_A/r_O$ , new tolerance factor ( $\tau$ ), ionic radii ( $r_A$  and  $r_B$ ), and oxidation states ( $O_A$  and  $O_B$ ) of A site and B site elements as our features. The previous prediction of simple ABO<sub>3</sub> perovskite oxides by machine learning indicated that, besides the ionic radii of A-site and B-site cations, the tolerance factor and the octahedral factor, the A-O and B-O bond lengths and the Villars' Mendeleev numbers also played a critical role to decide if a known composition could form perovskite structure or not<sup>30</sup>. Therefore, we also chose the  $d_{AO}$ ,  $d_{BO}$ ,  $M_A$ , and  $M_B$  as our features for predicting FDPOs.

Besides the 12 commonly used features ( $r_A$ ,  $r_B$ ,  $O_A$ ,  $O_B$ ,  $t$ ,  $\mu$ ,  $r_A/r_O$ ,  $\tau$ ,  $d_{AO}$ ,  $d_{BO}$ ,  $M_A$ , and  $M_B$ ) for predicting the formability of simple perovskites, we further introduced another 9 features for emphasizing the contribution of multiple dopant ions in both A and B sites of FDPOs. The multiple dopants introduced the size discrepancy, resulting in phase separation of perovskite or the transformation to other structures. The size variance factor ( $\sigma^2 = \sum_i y_i r_i^2 - (\sum_i y_i r_i)^2$ )<sup>39</sup>, which described the mismatch in A-site ion radii in perovskite, was correlated to phase separation of perovskites. We adapted the size variance factors for both A-site and B-site ions as our new features. Furthermore, the differences in the size scales of different atoms pairs (i.e.,  $r_{A \max}/r_O$ ,  $r_{B \max}/r_O$ ,  $r_{A \min}/r_O$ ,  $r_{B \min}/r_O$ ,  $r_{A \max}/r_{A \min}$ , and  $r_{B \max}/r_{B \min}$ ) were also defined as new features further to investigate the size effect on the formability of FDPOs. Moreover, the multiple dopants commonly broke electroneutrality in FDPOs, relating to the formation of oxygen nonstoichiometry in the perovskite structures. However, the large nonstoichiometric oxygen or the big offset of electroneutrality can break the perovskite structure. Therefore, we defined a new factor of  $(O_A + O_B)/6$  to include the electroneutrality effect. As summarized in Supplementary Table 1, we selected 21 features of  $r_A$ ,  $r_B$ ,  $O_A$ ,  $O_B$ ,  $t$ ,  $\tau$ ,  $r_B/r_O(\mu)$ ,  $r_A/r_O$ ,  $d_{AO}$ ,  $d_{BO}$ ,  $M_A$ ,  $M_B$ ,  $r_{A \max}/r_O$ ,  $r_{B \max}/r_O$ ,  $r_{A \min}/r_O$ ,  $r_{B \min}/r_O$ ,  $r_{A \max}/r_{A \min}$ ,  $r_{B \max}/r_{B \min}$ ,  $\sigma^2(A)$ ,  $\sigma^2(B)$ , and  $(O_A + O_B)/6$  to perform our model training. Here, the fractionally doped ABO<sub>3</sub> involved multiple A-site and B-site cations. Therefore, we took the average using a weighted linear combination based on the fractions of the multiple elements to get the features for each fractionally doped ABO<sub>3</sub> composition (including both positive and negative data). Here, we simplified the expression by using the original symbols to express the average features.

The feature calculation started from the basic element properties such as ionic radius, stable oxidation state, bond length, and Mendeleev number for the individual primary elements from an easily accessible database and literature (Supplementary Tables 2 and 3). Then we either took average or did calculations based on the feature definition equations on getting the values for the 21 features. We adapted the ionic radii and oxidation states for all the A-site and B-site elements shown in Fig. 2. Typically, coordination number (CN), oxidation state, electronic spin, etc. affect the ionic radius for each element<sup>40</sup>. The CNs for A-site and B-site ions in the perovskite structure are 12 and 6, respectively. Usually, the 12-fold coordinated A-site ions have ionic radii

much bigger than B-site ions, relating to lower oxidation states (e.g., +1, +2, and +3). The smaller 6-fold coordinated B-site ions have higher and sometimes multiple oxidation states (e.g., +4, +3, +2). It is easy to find the Shannon ionic radii for the 6-fold coordinated B-site cations. However, some A-site elements don't have the Shannon ionic radii available with a CN of 12. Instead of directly deriving the ionic radii from the Shannon database, we adapted the newly complemented Shannon ionic radii by using the data from Ouyang's work<sup>41</sup>. The data-driven method SISSO extrapolated the Shannon radii to unusual oxidation states and arbitrary CNs, allowing the easy access of 12-coordinated A-site ions. As for the bond length, an empirical correlation  $s_{ij} = \exp(\frac{d_0 - d_{ij}}{B})$  was used to describe the relationship between the bond length  $d_{ij}$  and the bond valence  $s_{ij}$  in the bond-valence theory, where  $d_0$  and  $B$  were constants whose values were determined empirically for each pair of atoms that form bonds<sup>42,43</sup>. Assuming  $s_{ij}$  equal to the ideal valence of the atom, the quotient of the natural valence divided by the number of the nearest oppositely charged ions (CN), we estimated the ideal bond length  $d_{ij}$ . In our case, atoms in the A site and B site were both bonded to oxygen atoms. Then we could get the ideal bond length  $d_{AO}$  and  $d_{BO}$ . As for the Mendeleev numbers, Villars et al. found an expression involving the Mendeleev number to predict the formability of compounds for any binary, ternary, or quaternary system<sup>44</sup>. Many researchers have adopted the Mendeleev number for the prediction of simple perovskite oxides<sup>30,33,45</sup>.

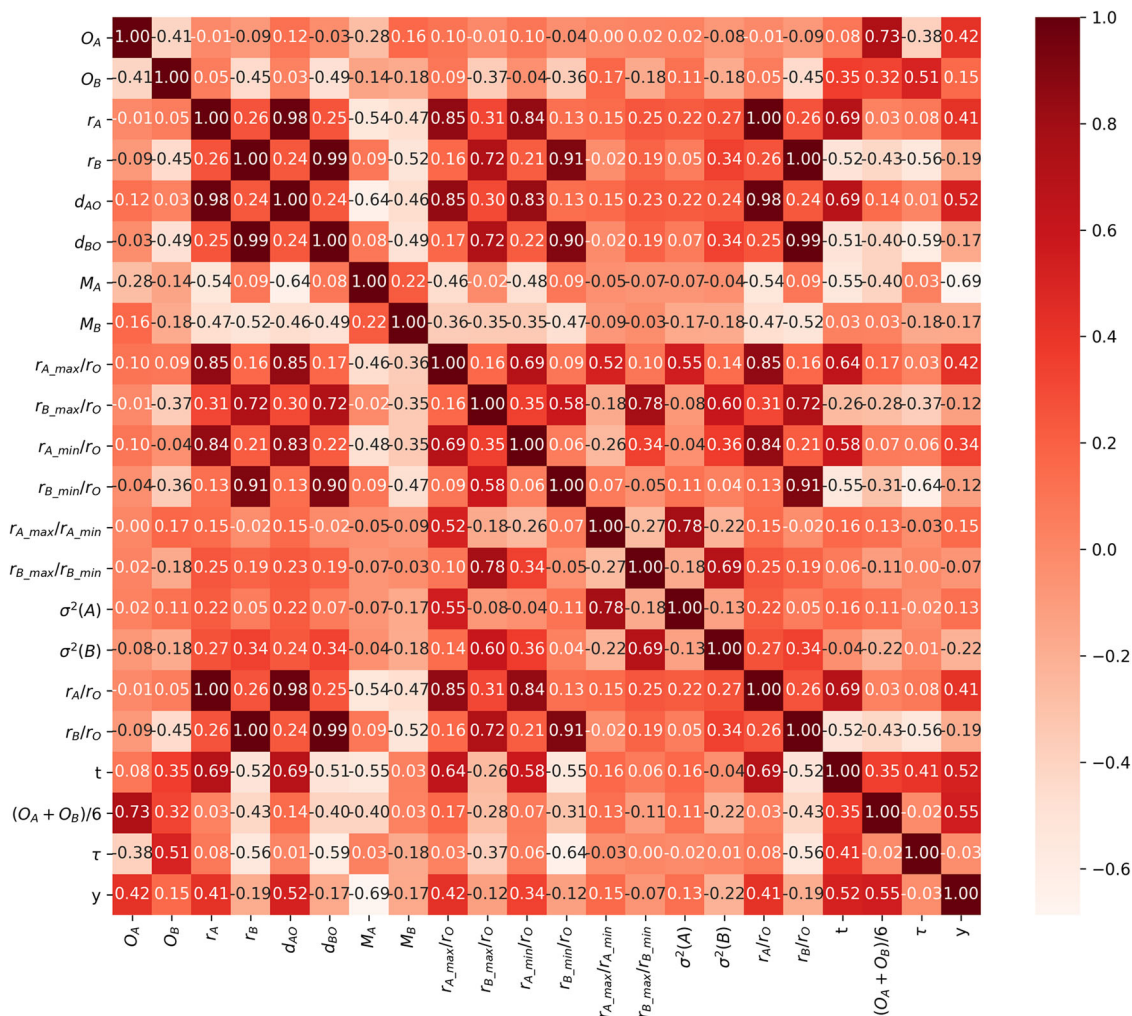
Based on the primary element properties of ionic radius, stable oxidation state, bond length, and Mendeleev number for both A-site and B-site elements, we could obtain the feature values of the average  $r_A$ ,  $r_B$ ,  $O_A$ ,  $O_B$ ,  $d_{AO}$ ,  $d_{BO}$ ,  $M_A$ , and  $M_B$  for each FDPOs. After that, we could calculate the values for the features of  $t$ ,  $\tau$ ,  $r_B/r_O(\mu)$ ,  $r_A/r_O$ , and  $(O_A + O_B)/6$  from the above calculated average values. By comparing the ionic radii for the A-site cations and B-site cations in each composition, we could get  $r_{A \max}$ ,  $r_{B \max}$ ,  $r_{A \min}$ , and  $r_{B \min}$ . Therefore, we could quickly get  $r_{A \max}/r_O$ ,  $r_{B \max}/r_O$ ,  $r_{A \min}/r_O$ ,  $r_{B \min}/r_O$ ,  $r_{A \max}/r_{A \min}$ ,  $r_{B \max}/r_{B \min}$ . Following the definition equations for  $\sigma^2(A)$ ,  $\sigma^2(B)$ , we could calculate their values easily from the ionic radii of multiple A-site and B-site cations. All the values of the 21 features for the 632 perovskite and non-perovskite data used for our model training are summarized in the Supplementary Data 1.

We evaluated the performances of the six models using the data (632 compositions with 21 features). Table 1 provides the average accuracy, F1 score, and their corresponding standard deviations for the trained six different models. We can see that the gradient boosting classifier exhibits the best accuracy and the best F1 score among all models, which indicates the gradient boosting classifier is the most suitable model for this classification problem. Consequently, we chose the gradient boosting classifier to study prediction of FDPOs further.

We have found out that the gradient boosting classifier with 21 features has demonstrated higher accuracy and F1 score during

**Table 1 A summary of average accuracy, average F1 score, and their corresponding standard deviations of six models.**

Classifier	Average accuracy ( $\pm$ std)	Average F1 score ( $\pm$ std)
SVM	0.924 $\pm$ 0.017	0.855 $\pm$ 0.035
Random Forest	0.954 $\pm$ 0.025	0.916 $\pm$ 0.050
AdaBoost	0.933 $\pm$ 0.022	0.886 $\pm$ 0.039
XGBoost	0.953 $\pm$ 0.017	0.913 $\pm$ 0.035
Gradient Boosting	0.954 $\pm$ 0.018	0.921 $\pm$ 0.037
MLP	0.930 $\pm$ 0.022	0.865 $\pm$ 0.047



**Fig. 3** Pearson correlation map for 21 features based on the dataset. The number represents the linear correlation between two variables.

the training process. However, a large number of features usually increases the model complexity<sup>46</sup>. It is desirable to identify and employ the most relevant features to develop a relatively simple model to predict the FDPOs efficiently and accurately. We employed the Pearson correlation coefficient (R), feature importance, and univariate feature selection to select necessary features. Figure 3 provides the Pearson correlation matrix showing the correlation for each pair of the 21 features used in this study. The features of  $r_A$ ,  $r_A/r_O$ ,  $r_{Amin}/r_O$ , and  $d_{AO}$  strongly correlate. The features of  $r_B$ ,  $r_B/r_O$ ,  $r_{Bmin}/r_O$ , and  $d_{BO}$  strongly correlate. The strong linear correlations for the two groups of the A-site cation and B-site cation-related features make it possible to shrink the necessary features. We may only use two features to represent the eight features for our model development.

The feature importance given by the model was adopted to guarantee the rationality and correctness of the feature selection overall. Table 2 indicates that the top 10 features for each fivefold cross-validation experiment are not entirely the same. However, the features of  $d_{AO}$ ,  $d_{BO}$ ,  $M_A$ ,  $(O_A + O_B)/6$ ,  $\tau$  and  $\sigma^2(B)$  showed up in all five experiments, indicating that these six features are the most important ones for achieving good performance. The  $r_{Amax}/r_{Amin}$  feature showed up in four experiments, while  $M_B$  and  $r_{Bmin}/r_O$  features showed up in three experiments. Compared with other features, these 9 features have a more significant influence on the prediction results.

Furthermore, we performed univariate feature selection with F-test to select the top-n most significant features on the whole

data set. First, we computed the ANOVA (Analysis of Variance) F-value between features and labels to see any statistically significant relationship between them. Then we selected features according to the n highest F-values. Here we set n from 9 to 4 (top 9 features to top 4 features). Aiming at further down-selecting features, we explored the performance of the gradient boosting classifier by employing different feature combinations among nine features, respectively.

Supplementary Table 4 showed the specific different feature combinations among top nine features of the gradient boosting classifier. Table 3 summarizes the performance of different feature combination choices by evaluating their average accuracy, average F1 score, and corresponding standard deviations. From Table 3, we can see that the model with 5 features ( $d_{AO}$ ,  $d_{BO}$ ,  $M_A$ ,  $(O_A + O_B)/6$  and  $\sigma^2(B)$ ) has the highest average accuracy and average F1 score. Based on the Pearson correlation map in Fig. 3, the absolute Pearson correlation coefficient of any pairwise features of these 5 features is much smaller than 0.85, indicating no highly linear correlation among these 5 features. And the importance of these 5 features is consistent with the previous theoretical and experimental researches<sup>30,45,47,48</sup>. Thus, using the gradient boosting classifier with these 5 features is reasonable and trustworthy to predict the unknown compounds' formability in the target chemical space. It is noteworthy that adopting as many features as possible would not guarantee higher validation accuracy here<sup>49</sup> because the presence of uncorrelated or nonsignificant features is likely to make the prediction

**Table 2 Top 10 crucial features given by the gradient boosting classifier with 5-fold cross-validation.**

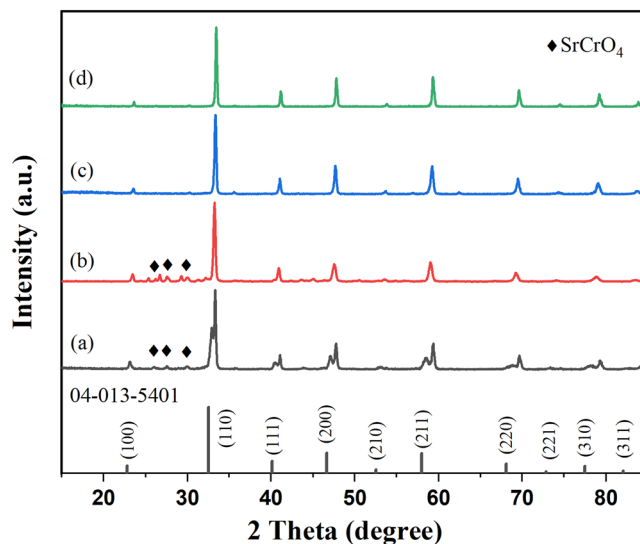
Feature importance	Experiment 1	Experiment 2	Experiment 3	Experiment 4	Experiment 5
Top 1	$M_A$	$d_{AO}$	$d_{AO}$	$d_{AO}$	$d_{AO}$
Top 2	$d_{BO}$	$d_{BO}$	$M_A$	$M_A$	$d_{BO}$
Top 3	$d_{AO}$	$r_{Amax}/r_{Amin}$	$r_{Amax}/r_{Amin}$	$r_{Amax}/r_{Amin}$	$\sigma^2(A)$
Top 4	$\tau$	$M_A$	$d_{BO}$	$d_{BO}$	$\tau$
Top 5	$(O_A + O_B)/6$	$\tau$	$t$	$\sigma^2(B)$	$r_{Amax}/r_{Amin}$
Top 6	$M_B$	$r_A$	$\tau$	$\tau$	$r_{Bmin}/r_O$
Top 7	$t$	$M_B$	$\sigma^2(B)$	$(O_A + O_B)/6$	$(O_A + O_B)/6$
Top 8	$\sigma^2(B)$	$r_{Bmin}/r_O$	$(O_A + O_B)/6$	$O_A$	$M_A$
Top 9	$r_{Amax}/r_O$	$\sigma^2(B)$	$r_{Bmin}/r_O$	$M_B$	$\sigma^2(B)$
Top 10	$\sigma^2(A)$	$(O_A + O_B)/6$	$r_A/r_O$	$r_{Bmax}/r_{Bmin}$	$r_{Bmax}/r_{Bmin}$

**Table 3 A summary of average accuracy, average F1 score, and corresponding standard deviations of the gradient boosting classifier with different feature combinations.**

Feature numbers	Average accuracy ( $\pm$ std)	Average F1 score ( $\pm$ std)
9	0.953 $\pm$ 0.016	0.917 $\pm$ 0.028
8	0.956 $\pm$ 0.012	0.923 $\pm$ 0.021
7	0.957 $\pm$ 0.023	0.924 $\pm$ 0.044
6	0.952 $\pm$ 0.019	0.916 $\pm$ 0.037
5	0.959 $\pm$ 0.023	0.926 $\pm$ 0.045
4	0.943 $\pm$ 0.016	0.899 $\pm$ 0.034

performance of models degraded. It indicates that selecting high-quality features is crucial to utilize machine learning for predicting new material discovery.

**Verification of the established machine learning method.** Using the machine learning method established from the gradient boosting classifier, the five features, and 632 perovskite and non-perovskite data, we predicted the perovskite structure formation for 47 compositions. Among them, 36 were reported in the literature, and 11 were newly synthesized in this work (Supplementary Table 5). Among the 36 compositions with known crystal structures from the literature, we obtained 34 accurate predictions. Although the verification data came from the FDPOs for extensive applications (i.e., STCH, SOFC electrolyte, SOFC electrode, and dual-phase perovskite materials), the established method achieved a prediction accuracy as high as 94.4%. Therefore, our method can predict potential FDPOs extensively without considering the chemical space too strictly. Among the 11 new compositions, 9 were predicted to be perovskite, while two were non-perovskite. We synthesized the powder samples using the conventional solid-state reaction method and characterized their crystal structure using XRD for all 11 compositions. Supplementary Figures 1-4 provide the XRD patterns for all 11 samples of Sr-La-Ti-Fe-O, Sr-La-Cr-Mn-O, Ca-La-Cr-Mn-O, and Ba-Ce-Co-Fe-O groups. For Sr-La-Ti-Fe-O group in Supplementary Figure 1, the Miller indices of the main reflection planes were identified according to the diffraction files of International Centre for Diffraction Data (ICDD) (04-021-6619). And the presence of impurities was observed in  $Sr_{0.4}La_{0.6}Ti_{0.8}Fe_{0.2}O_3$  and  $Sr_{0.6}La_{0.4}Ti_{0.8}Fe_{0.2}O_3$  samples. For Sr-La-Cr-Mn-O group in Supplementary Figure 2,  $Sr_{0.1}La_{0.9}Cr_{0.5}Mn_{0.5}O_3$  were well assigned to ICDD standard data card (04-013-5401). It was seen that the rest diffraction peaks of the  $Sr_{0.7}La_{0.3}Cr_{0.5}Mn_{0.5}O_3$  and  $Sr_{0.9}La_{0.1}Cr_{0.5}Mn_{0.5}O_3$  samples of the impurities exist. All three compositions of Ca-La-Cr-Mn-O group in Supplementary Figure 3 were assigned to ICDD standard data card (04-007-6214). As for  $BaCe_{0.5}Co_{0.5}O_3$  and  $BaCe_{0.3}Y_{0.2}Fe_{0.3}Co_{0.2}O_3$ ,

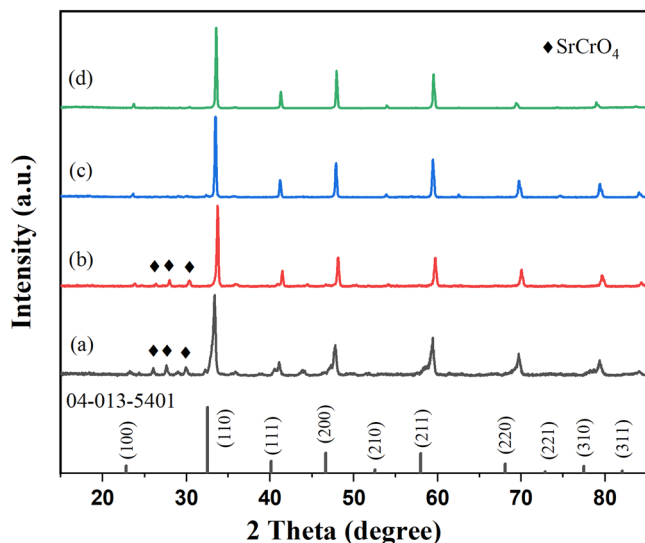


**Fig. 4 XRD patterns of  $Sr_{0.7}La_{0.3}Cr_{0.5}Mn_{0.5}O_3$  in different experimental conditions. a** calcination at 1350 °C for 12 hours via the solid-state method, **b** calcination at 1100 °C for 16 hours via the modified Pechini method, **c** calcination at 1350 °C for 12 hours via the modified Pechini method, and **d** calcination at 1500 °C for 2 hours via the modified Pechini method.

secondary phases clearly existed except the perovskite structure in XRD patterns of Supplementary Figure 4.

The XRD patterns indicate that 4 samples ( $Sr_{0.7}La_{0.3}Cr_{0.5}Mn_{0.5}O_3$ ,  $Sr_{0.9}La_{0.1}Cr_{0.5}Mn_{0.5}O_3$ ,  $Sr_{0.4}La_{0.6}Ti_{0.8}Fe_{0.2}O_3$ , and  $Sr_{0.6}La_{0.4}Ti_{0.8}Fe_{0.2}O_3$ ) synthesized via the solid-state methods did not form pure perovskite. However, our model predicted that they should be able to form a perovskite structure. Interestingly, the 2 samples of  $Sr_{0.7}La_{0.3}Cr_{0.5}Mn_{0.5}O_3$  and  $Sr_{0.9}La_{0.1}Cr_{0.5}Mn_{0.5}O_3$  contain many Mn and Cr, which could easily vaporize and make the solid-state reaction synthesis difficult. The Mn and Cr loss at the high-temperature calcination might cause the wrong structure results.

We suspected that  $Sr_{0.7}La_{0.3}Cr_{0.5}Mn_{0.5}O_3$  and  $Sr_{0.9}La_{0.1}Cr_{0.5}Mn_{0.5}O_3$  synthesized by solid-state reaction might encounter the Cr or Mn problem, causing non-perovskite structure and inconsistency with the predicted result (perovskite structure). After XRD characterization, XRD patterns in Figs. 4a and 5a exhibited the presence of  $SrCrO_4$  phase, one common impurity for similar compositions<sup>50</sup>, in two compositions synthesized by the solid-state method. Besides, the XRF test results of the two samples synthesized by the solid-state reaction method in Supplementary Table 6 and Supplementary Table 7 showed apparent Cr loss and Mn loss (For  $Sr_{0.7}La_{0.3}Cr_{0.5}Mn_{0.5}O_3$ , the Mn loss is apparent, and for  $Sr_{0.9}La_{0.1}Cr_{0.5}Mn_{0.5}O_3$ , the Mn loss is negligible). The fitted Cr spectra in Supplementary Figure 5 show that  $Cr^{6+}$  ions resulting from  $SrCrO_4$  impurity exist in  $Sr_{0.7}La_{0.3}Cr_{0.5}Mn_{0.5}O_3$  and

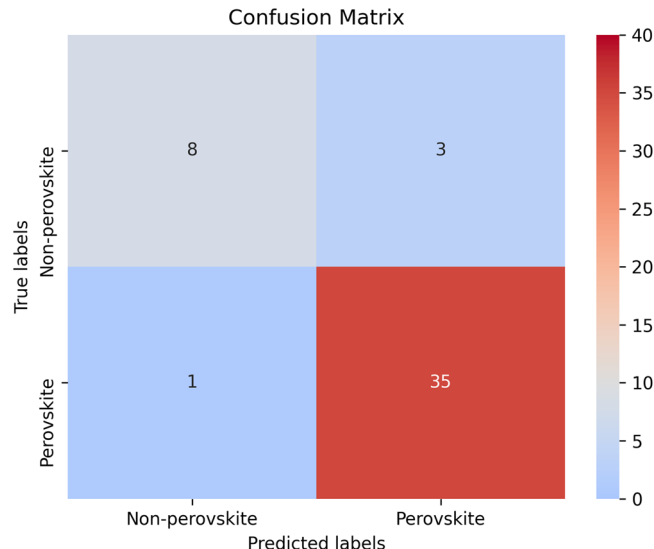


**Fig. 5** XRD patterns of  $\text{Sr}_{0.9}\text{La}_{0.1}\text{Cr}_{0.5}\text{Mn}_{0.5}\text{O}_3$  in different experimental conditions. **a** calcination at 1350 °C for 12 hours via the solid-state method, **b** calcination at 1350 °C for 12 hours via the modified Pechini method, **c** calcination at 1500 °C for 2 hours via the modified Pechini method, and **d** calcination at 1500 °C for 12 hours via the modified Pechini method.

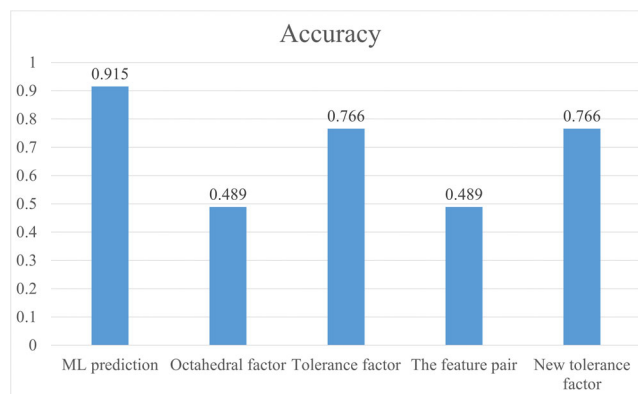
$\text{Sr}_{0.9}\text{La}_{0.1}\text{Cr}_{0.5}\text{Mn}_{0.5}\text{O}_3$ , which matched with XRD results well. XPS peak fitting for Cr 2p peaks was determined based on published literature<sup>51,52</sup>. Therefore, we further synthesized the compositions of  $\text{Sr}_{0.7}\text{La}_{0.3}\text{Cr}_{0.5}\text{Mn}_{0.5}\text{O}_3$  and  $\text{Sr}_{0.9}\text{La}_{0.1}\text{Cr}_{0.5}\text{Mn}_{0.5}\text{O}_3$  using the modified Pechini method and did XRD characterization again for the as-synthesize powders. We also optimized the calcination conditions to ensure the synthesis quality. Finally, as shown in Figs. 4d and 5d, both  $\text{Sr}_{0.7}\text{La}_{0.3}\text{Cr}_{0.5}\text{Mn}_{0.5}\text{O}_3$  and  $\text{Sr}_{0.9}\text{La}_{0.1}\text{Cr}_{0.5}\text{Mn}_{0.5}\text{O}_3$  formed phase-pure perovskite structures. We can conclude that the prediction can even help correct the wrong synthesis process to achieve the correct crystal structure.

As for  $\text{Sr}_{0.4}\text{La}_{0.6}\text{Ti}_{0.8}\text{Fe}_{0.2}\text{O}_3$  and  $\text{Sr}_{0.6}\text{La}_{0.4}\text{Ti}_{0.8}\text{Fe}_{0.2}\text{O}_3$ , it is not easy to synthesize Ti-containing oxides using wet chemistry because of the strong hydrolysis effect of  $\text{Ti}^{4+}$  in an aqueous solution. We have done multiple-time grinding and calcination to improve the synthesis quality based on solid-state reaction synthesis. However, we still did not get a phase-pure compound as the model predicted. Therefore, we counted these compositions as wrong predictions. We are developing new wet chemistry to synthesize the Ti-containing FDPOs and hope to clarify the composition in a future publication.

In summary, the machine learning prediction guided us to improve the experimental synthesis results for some synthesis-history-sensitive compounds. The model obtained 9 correct predictions out of 11 new compounds. The prediction accuracy for new compounds is 81.8%. Figure 6 shows the visualization result of the confusion matrix for the best prediction on a total of 47 experimental data. The confusion matrix is a common way to evaluate and visualize the performance of binary classification problems. The confusion matrix is a table with two rows and two columns, where rows report the true classification and columns report the predicted classification. There are four areas in the confusion matrix. All correct predictions where prediction labels are the same with true labels are in the diagonal of the table (a red area and a grey area). The accuracy of this model is 0.915 (43 true predictions of the total 47 data points), which means excellent reliability in predicting the formability of FDPOs.



**Fig. 6** Confusion matrix of the gradient boosting classifier with 5 features on 47 experimental data. The rows represent true labels, and the columns represent the predicted labels.



**Fig. 7** The comparison in accuracy between traditional methods and machine learning prediction with five features. The feature pair refers to using both octahedral factor and tolerance factor for the prediction.

**Comparison with traditional approaches.** For comparison, we performed predictions for the 47 compositions using conventional tolerance factor ( $t$ ), octahedral factor ( $\mu$ ), and new tolerance factor ( $\tau$ ). Figure 7 provides the prediction accuracy results for the five prediction methods. The model has a much better prediction accuracy (91.5%) than the prediction accuracies (76.6%, 48.9%, 48.9%, and 76.6%) of tolerance factor ( $t$ ), octahedral factor ( $\mu$ ), the feature pair ( $t$  and  $\mu$ ), and new tolerance factor ( $\tau$ ). Furthermore, both tolerance factor and new tolerance factor predicted that all the 47 compositions could form perovskites. Therefore, the most popular conventional method could not identify the compositions with complicated elemental compositions. The simple octahedral factor shows the worst prediction accuracy (48.9%), consistent with a simple consideration of only one feature. Compared with the prediction results using the single octahedral factor or the tolerance factor, using the feature pair does not guarantee the better prediction accuracy of the formability of FDPOs. All the values of these features for the 47 compositions used for our model training are summarized in the Supplementary Data 2.

## Conclusions

A function-confined facile machine learning method was discovered to predict new fractional and multicomponent perovskite oxides from the basic properties of metal elements and their corresponding simple oxides using typical classification models and a small number of experimental data. As a case study for discovering new perovskite oxides for solar thermochemical hydrogen production by water splitting, we obtained 632 experimental training data (516 are perovskites and 116 are non-perovskites), selected 12 available features, and created 9 new features. Based on high prediction accuracy (0.954), high F1 score (0.921), and the corresponding low standard errors (0.018 and 0.037), we selected the gradient boosting classifier to predict perovskite formation. The further verification of the derived machine learning algorithm using 47 data not included in the training data set (36 are from the literature, and 11 newly synthesized data) showed that the prediction accuracy reached 91.5%, much higher than the prediction performance of the conventional tolerance factor (76.6%), octahedral factor (48.9%), and new tolerance factor (76.6%). Furthermore, the machine learning prediction helped correct two wrong solid-state reaction syntheses. For future research, our model also can be expanded to discover other functional perovskite oxides.

## Methods

### Machine Learning

**Model training.** We applied models to classify input compositions into perovskites or non-perovskites using the above-selected 21 features and 632 perovskite and non-perovskite data. A feature vector (21 features) described each sample and labeled it with  $y$  (1 or 0, 1: perovskites, 0: non-perovskites). We employed six different known models in our classification work. Three models among them were often used for prediction in the material field: SVM (support vector machine)<sup>30,53</sup>, random forest<sup>30,53</sup>, and gradient boosting classifier<sup>30,45</sup>. The XGBoost<sup>54</sup>, AdaBoost<sup>55</sup>, and MLP (multilayer perceptron) classifier<sup>56</sup> were widely used models in classification tasks and showed excellent performance. Based on the nonlinear algorithm, SVM constructed a hyperplane with the maximum-margin having the largest separation between the nearest training data set of two classes in high dimensional space, guaranteeing accurate prediction results on unseen data<sup>53</sup>. The random forest, gradient boosting, XGBoost, and AdaBoost were ensemble learning methods used with the decision-tree algorithm. The different paths from the root to leaf represented classification rules<sup>57</sup>. The random forest classifier got the final prediction based on the majority voting of classification results from the decision-tree algorithm by learning several classification rules from the data set. Gradient boosting classifier trained the decision tree models sequentially, and new models corrected the previous ones' errors and improved the performance by training. AdaBoost could learn from the weak classifiers sequentially and give feedback to build a predictive classifier. XGBoost was an open-source library, providing an implementation of the regularizing gradient boosting algorithm. MLP was an artificial neural network and could model nonlinear relationships between targets and features because of the multiple layers and nonlinear activation.

We split the entire data set into a training set and a testing set. Models only trained by the training data are easily overfitting and overoptimistic. The testing set helped us check how models performed on unseen examples. We used 5-fold cross-validation to objectively estimate different models' performance. We divided the total data set into 5 subsets of the same size and chose one subset as the testing set, making the remaining subsets the training set. We performed 5 iterations for every algorithm. Over 5 iterations, the models were trained on 4 subsets and used to predict the held-out subset.

**Model evaluation.** Our data classification using models resulted in four possible cases, including true positive (TP), true negative (TN), false positive (FP), and false negative (FN). The commonly used metric to evaluate classification involves the accuracy ( $\frac{TP+TN}{TP+TN+FP+FN}$ ) parameter. The accuracy parameter equals the number of correctly predicted data points divided by the total number of all data points in the data set, which is one essential criterion to evaluate the performance of classification prediction. The model with higher accuracy has more accurate predictions on unseen data. Since our data set includes more positive data points than negative data points, our task is an imbalanced classification problem. Therefore, the typical F1 score ( $\frac{2TP}{2TP+FP+FN}$ ) was used to assess the imbalanced classification problem for further evaluating our models' performance. Since we performed the 5-fold cross-validation over five times and got five sets of accuracy and F1 score values, we adapted the average values to demonstrate the models' performance.

**Features selection.** We employed the Pearson correlation coefficient (R), feature importance, and univariate feature selection to select necessary features. The R with a value ranging from  $-1$  to  $+1$  measures the linear correlation between two variables. The R with a value of zero indicates that the studied two variables have no linear correlation. The positive and negative values represent positive linear correlation and negative linear correlation, respectively. The high absolute value of R indicates a strong correlation between the two variables. The closer to 1 the absolute value of R is the stronger correlation the variables have. Usually, we can regard the two features are highly correlated when  $|R| > 0.85$ <sup>46</sup>.

The feature importance given by the model was adopted to guarantee the rationality and correctness of the feature selection overall. The value of each feature importance is between 0 and 1. The total sum of all feature importance equals 1.0. The bigger the value of the feature importance is, the more essential the feature is for the prediction.

**Model validation and prediction of FDPOs.** After finding the best model from the mentioned six models and selecting the most important features out of the 21 features, we validated the performance using 47 unseen experimental data (Supplementary Table 5) from recently reported literature and newly synthesized in this work (not included in our training data set). The validation data consisted of FDPOs used for STCH production, solid oxide fuel cell electrolytes, mixed ionic and electronic conducting electrode materials, dual-perovskite composites, and some non-perovskites. The selection of FDPOs with extensive applications rather than STCH materials only is to find out the effectiveness of the established method for extensive study of the FDPO formation. Supplementary Table 5 indicated that the verification data includes seven materials groups (Sr-La-Mn-Al-O, Sr-La-Ti-Fe-O, Sr-La-Cr-Mn-O, Ca-La-Cr-Mn-O, Ba-Ce-Fe-O, Ba-Ce-Co-Fe-O, and Ba-Ce-Zr-Y-O). 36 compositions are the materials reported in the literature, and the rest 11 compositions were the newly designed and synthesized ones for discovering new potential STCH or SOFC materials. We also compared our prediction performance with those obtained using commonly used descriptors, the tolerance factor, the new tolerance factor, and the octahedral factor.

### Experimental section

**Synthesis of the predicted FDPOs.** We first synthesized all the 11 proposed compositions using the conventional solid-state reaction method. The stoichiometric amounts of raw materials of carbonates and oxides calculated from the proposed FDPO formula were weighed and mixed by ball-milling in isopropanol with 3 mm YSZ ball grinding media for 48 h. The ball-milled slurry was dried at 90 °C for 24 h in the oven to remove the solvent. After calcining the well-mixed powders at 1100–1500 °C for 12–48 h in stagnant air in a box furnace, we got the final powders for structure characterization.

For the compositions synthesized by the solid-state reaction method, which have shown opposite results to the prediction, we adapted a more powerful wet-chemistry synthesis method of the modified Pechini method to synthesize them again to ensure the synthesis accuracy. Using the synthesis of  $\text{Sr}_{0.7}\text{La}_{0.3}\text{Cr}_{0.5}\text{Mn}_{0.5}\text{O}_3$  as an example, we can briefly describe the modified Pechini synthesis process. The stoichiometric amounts of water-soluble materials such as nitrates calculated from the proposed FDPO formula were weighed and dissolved into distilled water. The chelating organic compounds of ethylenediaminetetraacetic acid (Alfa Aesar 99.4%) and citric acid monohydrate (ACROS Organics 99.5%) with a mole ratio of 1.5: 1.5: 1 for EDTA: citric acid: total metal ions were added to the metal salt aqueous solution. Then the pH of the obtained solution was adjusted to be  $\sim 9$  by adding ammonium hydroxide or nitric acid. After slow water vaporization at  $\sim 80$ – $90$  °C, the solution turned into a homogeneous viscous gel. Any solid precipitation during the evaporation process should be avoided by controlling the vaporization temperature, initial pH, and relative amount of chelates and metal ions. The bouffant charcoal-like powders obtained after heating the viscous gel in the oven at 150 °C for 48 h were further calcined at 600 °C for 5 hours in stagnant air to burn out organic substances. The final calcination at 900–1500 °C for 2–16 h allows us to obtain the final powders.

**Characterization.** The crystal structure of as-synthesized powders was characterized by powder X-ray diffraction using Cu K $\alpha$  radiation. The XRD patterns were collected by Rigaku Ultima IV diffractometer at the voltage of 40 kV and current of 15 mA. The XRD patterns were recorded in the  $2\theta$  range of 15–85° with a step size of 0.02°. XRD pattern analysis was conducted using HighScore software.

The elemental compositions for some samples were characterized by an X-ray fluorescence (XRF) spectrometer (Thermo ARL Perform'x wavelength-dispersive spectrometer). The powder materials are grinded in a puck mill, and they are all mixed with cellulose binder, then pressed into a disk for analysis. The cellulose was added in a ratio of 1.5 g of Cellulose to 12 g of sample (0.125: 1). The OXSAS software by ThermoFisher commands the instrument to go to the correct detector angle and counts the intensity for each element. Then the UniQuant software was adapted to convert each intensity into a mass percent.

X-ray photoelectron spectroscopy (XPS) spectra were recorded by a PHI VersaProbe III spectrometer using monochromatic Al K $\alpha$  x-ray source ( $h\nu = 1486.7$  eV) at an accelerating voltage of 15 kV and output power of 25 W. The high-resolution spectra for Cr 2p were collected at 90 degrees with a pass energy of 69 eV, 0.125 eV step size, and a



dwell time of 50 ms. To calibrate the binding energy scale, the C 1 s level was assumed to be 284.8 eV. Ar ion etching at 3 kV was used to characterize chemical information of deeper position. Etch steps were 5 min with the etching speed of 136.5623 Å min<sup>-1</sup>. Peaks were fitted using Shirley baseline with the peak fitting parameters of GL(60) and a FWHM of 1.4 eV for peaks within CasaXPS.

### Data availability

The authors declare that the main data supporting the findings of this study are available within the article and its Supplementary Information files. The main data is also available at this link (<https://github.com/TheLuoFengLab/perovskite-classification>). All other relevant data are available from the corresponding authors upon reasonable request.

### Code availability

All source code for the machine learning methods is available at this link (<https://github.com/TheLuoFengLab/perovskite-classification>).

Received: 6 December 2021; Accepted: 21 June 2022;

Published online: 06 July 2022

### References

- Saparov, B. & Mitzi, D. B. Organic-Inorganic Perovskites: Structural Versatility for Functional Materials Design. *Chem. Rev.* **116**, 4558–4596 (2016).
- Jung, E. H. et al. Efficient, sfi and scalable perovskite solar cells using poly(3-hexylthiophene). *Nature* **567**, 511–515 (2019).
- Green, M. A., Ho-Baillie, A. & Snaith, H. J. The emergence of perovskite solar cells. *Nat. Photonics* **8**, 506–514 (2014).
- Schmidt-Mende, L. et al. Roadmap on organic-inorganic hybrid perovskite semiconductors and devices. *APL Mater.* **9**, 109202 (2021).
- Jacobsson, T. J. et al. An open-access database and analysis tool for perovskite solar cells based on the FAIR data principles. *Nat. Energy* **7**, 107–115 (2022).
- Tan, Z. et al. Two-Dimensional (C<sub>4</sub>H<sub>9</sub>NH<sub>3</sub>)<sub>2</sub>PbBr<sub>4</sub> Perovskite Crystals for High-Performance Photodetector. *J. Am. Chem. Soc.* **138**, 16612–16615 (2016).
- Huang, H., Pradhan, B., Hofkens, J., Roeffaers, M. B. J. & Steele, J. A. Solar-Driven Metal Halide Perovskite Photocatalysis: Design, Stability, and Performance. *ACS Energy Letters* **5**, 1107–1123 (2020).
- Yuan, Z. et al. One-dimensional organic lead halide perovskites with efficient bluish white-light emission. *Nat. Commun.* **8**, 109202 (2017).
- McDaniel, A. H. et al. Sr- and Mn-doped LaAlO<sub>3-δ</sub> for solar thermochemical H<sub>2</sub> and CO production. *Energy Environ. Sci.* **6**, 2424–2428 (2013).
- Shao, Z. et al. Investigation of the permeation behavior and stability of a Ba<sub>0.5</sub>Sr<sub>0.5</sub>Co<sub>0.8</sub>Fe<sub>0.2</sub>O<sub>(3-δ)</sub> oxygen membrane. *J. Memb. Sci.* **172**, 177–188 (2000).
- Shao, Z. & Haile, S. M. A high-performance cathode for the next generation of solid-oxide fuel cells. *Mater. Sustain. Energy A Collect. Peer-Reviewed Res. Rev. Artic. from Nat. Publ. Gr 3*, 255–258 (2010).
- Zhu, H., Ricote, S., Duan, C., O’Hayre, R. P. & Kee, R. J. Defect chemistry and transport within dense BaCe<sub>0.9</sub>Zr<sub>0.1</sub>Y<sub>0.1</sub>O<sub>3-δ</sub>(BCZYYb) proton-conducting membranes. *J. Electrochem. Soc.* **165**, F845–F853 (2018).
- Xia, C. et al. Shaping triple-conducting semiconductor BaCo<sub>0.4</sub>Fe<sub>0.4</sub>Zr<sub>0.1</sub>Y<sub>0.1</sub>O<sub>3-δ</sub> into an electrolyte for low-temperature solid oxide fuel cells. *Nat. Commun.* **10**, 1–9 (2019).
- Morales, M. et al. Correlation between electrical and mechanical properties in La<sub>1-x</sub>Sr<sub>x</sub>Ga<sub>1-y</sub>Mg<sub>y</sub>O<sub>3-δ</sub> ceramics used as electrolytes for solid oxide fuel cells. *J. Power Sources* **246**, 918–925 (2014).
- King, A. H. & Zhu, Y. Twin-corner disclinations in YBa<sub>2</sub>Cu<sub>3</sub>O<sub>7-δ</sub>. *Philos. Mag. A Phys. Condens. Matter, Struct. Defects Mech. Prop.* **67**, 1037–1044 (1993).
- Cohen, R. E. Origin of ferroelectricity in perovskite oxides. *Nature* **358**, 136–138 (1992).
- Ali, A. I. & Hassen, A. Synthesis, characterization, ferroelectric, and piezoelectric properties of (1-x)BaTiO<sub>3</sub>-x(BaNi<sub>0.5</sub>Nb<sub>0.5</sub>O<sub>3</sub>) perovskite ceramics. *J. Mater. Sci. Mater. Electron.* **32**, 10769–10777 (2021).
- Sheng, L., Xing, D. Y., Sheng, D. N. & Ting, C. S. Theory of colossal magnetoresistance in R<sub>1-x</sub>A<sub>x</sub>MnO<sub>3</sub>. *Phys. Rev. Lett.* **79**, 1710–1713 (1997).
- Balli, M., Jandl, S., Fournier, P. & Kedous-Lebouc, A. Advanced materials for magnetic cooling: Fundamentals and practical aspects. *Appl. Phys. Rev.* **4**, 021305 (2017).
- Fedorovskiy, A. E., Drigo, N. A. & Nazeeruddin, M. K. The role of Goldschmidt’s tolerance factor in the formation of A<sub>2</sub>BX<sub>6</sub> double halide perovskites and its optimal range. *Small Methods* **4**, 1–6 (2020).
- Li, C., Soh, K. C. K. & Wu, P. Formability of ABO<sub>3</sub> perovskites. *J. Alloys Compd.* **372**, 40–48 (2004).
- Li, C. et al. Formability of ABX<sub>3</sub> (X = F, Cl, Br, I) halide perovskites. *Acta Crystallogr. Sect. B Struct. Sci.* **64**, 702–707 (2008).
- Zhang, H., Li, N., Li, K. & Xue, D. Structural stability and formability of ABO<sub>3</sub>-type perovskite compounds. *Acta Crystallogr. Sect. B Struct. Sci.* **63**, 812–818 (2007).
- Tidrow, S. C. Mapping comparison of goldschmidt’s tolerance factor with perovskite structural conditions. *Ferroelectrics* **470**, 13–27 (2014).
- Wang, Z. L. & Kang, Z. C. Functional and Smart Materials. *Funct. Smart Mater.* **72**, 3264–3266 (1998).
- Feng, L. M. et al. Formability of ABO<sub>3</sub> cubic perovskites. *J. Phys. Chem. Solids* **69**, 967–974 (2008).
- Bartel, C. J. et al. New tolerance factor to predict the stability of perovskite oxides and halides. *Sci. Adv.* **5**, 1–10 (2019).
- Emery, A. A. & Wolverton, C. High-throughput DFT calculations of formation energy, stability and oxygen vacancy formation energy of ABO<sub>3</sub> perovskites. *Sci. Data* **4**, 1–10 (2017).
- Pilania, G. et al. Machine learning bandgaps of double perovskites. *Sci. Rep.* **6**, 1–10 (2016).
- Liu, H. et al. Screening stable and metastable ABO<sub>3</sub> perovskites using machine learning and the materials project. *Comput. Mater. Sci.* **177**, 109614 (2020).
- Talapatra, A., Uberuaga, B. P., Stanek, C. R. & Pilania, G. A machine learning approach for the prediction of formability and thermodynamic stability of single and double perovskite oxides. *Chem. Mater.* **33**, 845–858 (2021).
- Li, W., Jacobs, R. & Morgan, D. Predicting the thermodynamic stability of perovskite oxides using machine learning models. *Comput. Mater. Sci.* **150**, 454–463 (2018).
- Balachandran, P. V. et al. Predictions of new ABO<sub>3</sub> perovskite compounds by combining machine learning and density functional theory. *Phys. Rev. Mater.* **2**, 1–18 (2018).
- Ye, W., Chen, C., Wang, Z., Chu, I. H. & Ong, S. P. Deep neural networks for accurate predictions of crystal stability. *Nat. Commun.* **9**, 1–6 (2018).
- McDaniel, A. H. et al. Nonstoichiometric perovskite oxides for solar thermochemical H<sub>2</sub> and CO production. *Energy Procedia* **49**, 2009–2018 (2014).
- Barcellos, R. D., Sanders, M. D., Tong, J., McDaniel, A. H. & O’Hayre, R. P. BaCe<sub>0.25</sub>Mn<sub>0.75</sub>O<sub>3-δ</sub>-a promising perovskite-type oxide for solar thermochemical hydrogen production. *Energy Environ. Sci.* **11**, 3256–3265 (2018).
- Kumar, A., Verma, A. S. & Bhardwaj, S. R. Prediction of formability in perovskite-type oxides. *Open Appl. Phys. J.* **1**, 11–19 (2008).
- Li, W., Ionescu, E., Riedel, R. & Gurlo, A. Can we predict the formability of perovskite oxynitrides from tolerance and octahedral factors? *J. Mater. Chem. A* **1**, 12239–12245 (2013).
- Rodriguez-Martinez, L. M. & Attfield, J. P. Cation disorder and size effects in magnetoresistive manganese oxide perovskites. *Phys. Rev. B - Condens. Matter Mater. Phys.* **54**, R15622–R15625 (1996).
- Shannon, R. D. Revised effective ionic radii and systematic studies of interatomic distances in halides and chalcogenides. *Acta Crystallographica Section A: crystal physics, diffraction, theoretical and general crystallography* **32**, 751–767 (1976).
- Ouyang, R. Exploiting ionic radii for rational design of halide perovskites. *Chem. Mater.* **32**, 595–604 (2020).
- Brown, I. D. & Altermatt, D. Bond-valence parameters obtained from a systematic analysis of the inorganic crystal structure database. *Acta Crystallographica Section B: Structural Science* **41**, 244–247 (1985).
- Brown, I. D. Recent developments in the methods and applications of the bond valence model. *Chem. Rev.* **109**, 6858–6919 (2009).
- Villars, P. et al. Binary, ternary and quaternary compound former/nonformer prediction via Mendelev number. *J. Alloys Compd.* **317–318**, 26–38 (2001).
- Pilania, G., Balachandran, P. V., Gubernatis, J. E. & Lookman, T. Classification of ABO<sub>3</sub> perovskite solids: A machine learning study. *Acta Crystallogr. Sect. B Struct. Sci. Cryst. Eng. Mater.* **71**, 507–513 (2015).
- Sharma, V., Kumar, P., Dev, P. & Pilania, G. Machine learning substitutional defect formation energies in ABO<sub>3</sub> perovskites. *J. Appl. Phys.* **128**, 034902 (2020).
- Pilania, G., Balachandran, P. V., Kim, C. & Lookman, T. Finding new perovskite halides via machine learning. *Front. Mater.* **3**, 1–7 (2016).
- Li, K., Wang, Y., Lin, J. & Li, Z. Phase relations of BaCoO<sub>3-δ</sub>-BaInO<sub>2.5</sub> and size variation effect of B-site cations on the phase transitions. *Solid State Ionics* **183**, 7–15 (2011).
- Liu, W., Ma, X., Ren, S., Lei, X. & Liu, L. Tunable phase transition in (Bi<sub>0.5</sub>Na<sub>0.5</sub>)<sub>0.94</sub>Ba<sub>0.06</sub>TiO<sub>3</sub> by B-site cations. *Appl. Phys. A Mater. Sci. Process.* **126**, 1–10 (2020).
- Ding, X., Liu, Y., Gao, L. & Guo, L. Synthesis and characterization of doped LaCrO<sub>3</sub> perovskite prepared by EDTA – citrate complexing method. *J. Alloys Compd.* **458**, 346–350 (2008).

51. Biesinger, M. C., Brown, C., Mycroft, J. R., Davidson, R. D. & McIntyre, N. S. X-ray photoelectron spectroscopy studies of chromium compounds. *Surf. Interface Anal.* **36**, 1550–1563 (2004).
52. Biesinger, M. C. et al. Resolving surface chemical states in XPS analysis of first row transition metals, oxides and hydroxides: Cr, Mn, Fe, Co and Ni. *Appl. Surf. Sci.* **257**, 2717–2730 (2011).
53. Agiorgousis, M. L., Sun, Y. Y., Choe, D. H., West, D. & Zhang, S. Machine learning augmented discovery of chalcogenide double perovskites for photovoltaics. *Adv. Theory Simulations* **2**, 1–9 (2019).
54. Chen, T. & Guestrin, C. XGBoost: a scalable tree boosting system. *Proceedings of the 22<sup>nd</sup> ACM sigkdd international conference on knowledge discovery and data mining*. 785–795 (2016).
55. Freund, Y. & Schapire, R. E. A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Syst. Sci.* **55**, 119–139 (1997).
56. Hinton, G. E. Connectionist learning procedures. in *Machine Learning* 555–610 (Morgan Kaufmann Publishers, Inc., 1990).
57. Li, J. et al. AI Applications through the whole life cycle of material discovery. *Matter* **3**, 393–432 (2020).

## Acknowledgements

This material is based upon work supported by the US Department of Energy's Office of Energy Efficiency and Renewable Energy (EERE) under the Fuel Cell Technologies Office Award Number DE-EE0008428. This work was supported in part by the US National Science Foundation (NSF) under Grant ABI-1759856 and MTM2-2025541 to FL. This work was supported in part by the National Aeronautics and Space Administration (NASA) under Grant #80NSSC20M0233 (NASA) to JT. The authors also acknowledge Dr. Nathaniel Huygen in the National Brick Research Center at Clemson University for the XRF testing and Dr. Kelliann Koehler in the Clemson University Electron Microscopy Lab for the XPS testing.

## Author contributions

X.Z., F.D., F.L., and J.T. contributed the intellectual concept, method design, and regular discussion/analysis of this work. X.Z. collected and analyzed the training data and experimental verification data. F.D. wrote the codes and did the machine learning. Z.Z. helped with the collection of data. A.S. helped with the figures drawing. X.Z. and J.T. led

the writing of this manuscript and all other coauthors helped the discussion and edition during the writing of the manuscript. F.L. and J.T. oversight and managed the project.

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1038/s43246-022-00269-9>.

**Correspondence** and requests for materials should be addressed to Feng Luo or Jianhua Tong.

**Peer review information** *Communications Materials* thanks Ghanshyam Paliana, Lazar Rakočević and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Primary Handling Editor: Aldo Isidori. Peer reviewer reports are available.

**Reprints and permission information** is available at <http://www.nature.com/reprints>

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2022