Engineered Science
DOI: https://dx.doi.org/10.30919/es1561

Discovery of Quaternary Perovskites via Stacked Machine Learning and Generative Models

Ruichen Tian,[1,2,3,*] Aldrin D. Calderon[1,2,*] and Xiaoyu Liu[1,2]

Abstract

This study explores machine learning-driven prediction of perovskite cathode materials for solid oxide fuel cells (SOFCs) by analyzing formation energy and bandgap properties. Addressing experimental challenges in material discovery, we developed a predictive framework using 602,000 compounds with feature-engineered atomic properties. Four algorithms (random forest, support vector machine, gradient boosting, and decision tree) and a generative adversarial network/variational autoencoders (GAN/VAE)-enhanced stacking model were evaluated through 100 iterations of 70-30 train-test splits. The stacking model achieved superior performance with 96% (±0.09%) tenfold cross-validation accuracy, significantly outperforming previous 60% benchmarks. Feature importance analysis identified key atomic properties guiding material selection. Through generative adversarial networks, we improved bandgap and formation energy predictions, ultimately screening 1,981 promising quaternary perovskites. This data-driven approach demonstrates substantial potential for accelerating SOFC cathode development by reducing experimental trial cycles through reliable computational pre-screening. The methodology establishes a robust foundation for machine learning applications in high-temperature material discovery, particularly in optimizing complex multi-property requirements for energy conversion technologies.

Keywords: Perovskites; Materials design; Materials properties prediction; Machine learning; Stacking model; Variational autoencoders.
Received: 11 January 2025; Revised: 28 February 2025; Accepted: 02 April 2025.
Article type: Research article.

1. Introduction

Solid oxide fuel cells (SOFCs) are considered a promising green energy conversion device.[1] It directly converts the chemical energy of the fuel into electrical energy, which is an effective way to achieve high-efficiency conversion and clean utilization.[2] The energy power density in an SOFC is chiefly determined by the surface area and materials used in the cathode, anode and electrolytes.[3] However, due to the high operating temperatures, SOFCs have a poor durability. There are many active research domains for SOFCs including finding more suitable cathode and anode materials and improving its stability and reliability.[4]

An SOFC cathode typically falls into the class of perovskite compounds. Perovskites have the general form which contains two cation sites. The A cation are typically large and are from the rare earth, alkaline earth, alkali metal groups while the smaller B site cation typically is a transition metal. Cations with a wide range of ionic radius and valence can enter one or another position in the perovskite structure, and this dramatically impacts the perovskites physical and chemical properties. For use as SOFC cathode materials, a high concentration of oxide vacancies is desirable and leads to high ionic conductivity. This ability to maintain large amounts of oxygen vacancies makes them good candidates for electrodes in SOFCs which can accelerate the discovery of the new materials.[5] Although many elements can be used the A and B cations, so far only 1,000 perovskites have been developed for different industries through experiments. This

[1] *School of Mechanical, Manufacturing and Energy Engineering, Mapua University, Muralla Street, Intramuros, Manila, 1004, Philippines*
[2] *School of Graduate Studies, Mapua University, Muralla Street, Intramuros, Manila, 1004, Philippines*
[3] *School of Equipment maintenance, Hunan Defense Industry Polytechnic, NO.9, Xueshi Road, Yuhu District, Xiangtan City, Hunan Province, 411100, China*
**Email:* rtian@mymail.mapua.edu.ph (R. C. Tian); adcalderon@mapua.edu.ph (A. D. Calderon)

shows that perovskite still has great development and exploration areas.[6,7]

The traditional material development method is based on trial-and-error method, when the performance of the material reaches the target, it is need continuous synthesis and continuous trial. This method requires long-term research and complex procedures, so it is a time-consuming and expensive work.[8] In order to overcome the shortcomings and improve efficiency, density functional theory (DFT) can directly obtain some key characteristics of the material without experimental synthesis. DFT is based on the laws of quantum mechanics to predict the bond energy between atoms, allowing scientists to predict the properties of hundreds of molecules and materials, such as the electronic structure, density, hardness, optical properties, and reactions of composites sex. It has had an irreplaceable position in the field of predictive materials in the past 20 years. However, most calculation methods are only for specific systems, and some theoretical methods still cannot meet the requirements for a quantitative description of material properties. In addition, DFT requires high calculation costs and professional skills. At this time, machine learning with the advantages of low cost, quick and accurate to be widely used for predict the properties of the material.[6]

Machine learning is often used to build statistical models for data analysis and prediction. With the continuous development of materials science, machine learning continues to make major breakthroughs in materials science and can draw relationships and trend graphs through dataset without experiments.on the other hand, discovery of reverse materials also can use machine learning.[9,10]

2. Experimental process

This workflow of this project is illustrated in Fig. 1. Initially, applied feature engineering to select out the importance feature for generation new perovskites. Then, four machine learning methods, random forest (RF), support vector machine (SVM), gradient boosting classifier (GBC), decision tree (DT) and a stacking model to combine all the four methods have been introduced to select out the optimal classifier. Then, applied generative adversarial networks (GAN) to improve the prediction performance of bandgap and formation energy prediction. Moreover, VAE network has been introduced to visualize the feature map of perovskites which can be applied to calculate the similarity between two or more perovskites. Last but not least, the structure formula and other information can be predicted in combination with VAE fingerprint, feature engineering and the optimal machine learning classifier.[13]

2.1 Programming language and library

In this work, Python is the code language, which includes collecting the dataset (Mendeleev dataset),[12] data preprocess and feature engineering, as shown in Table 1.

2.2 Dataset and data preprocessing

The screening dataset used in this thesis which is from Ihalage *et al.*'s work.[13] Dataset can be found here: https://figshare.com/articles/dataset/all_generated_compounds_csv/13033262. The dataset used for feature engineering the

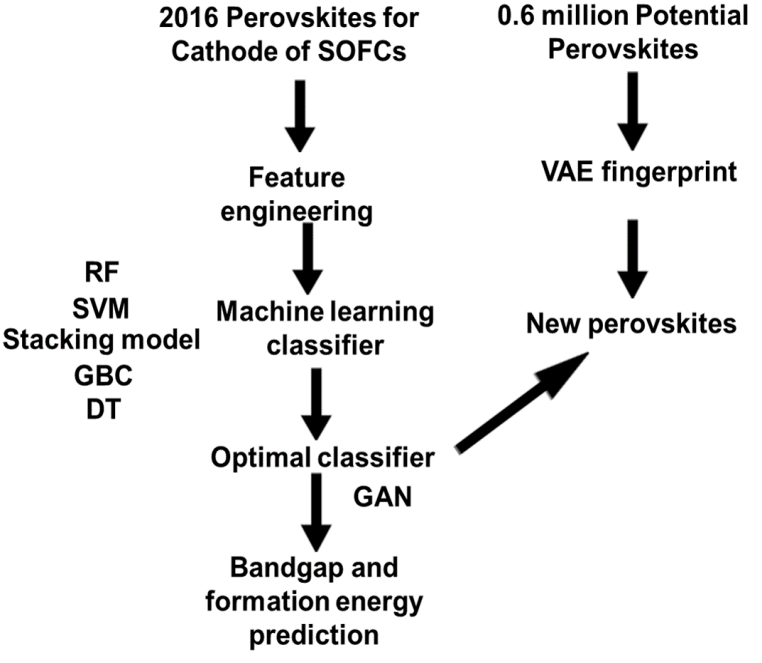

Fig. 1: The workflow of analogue discovery of perovskite.

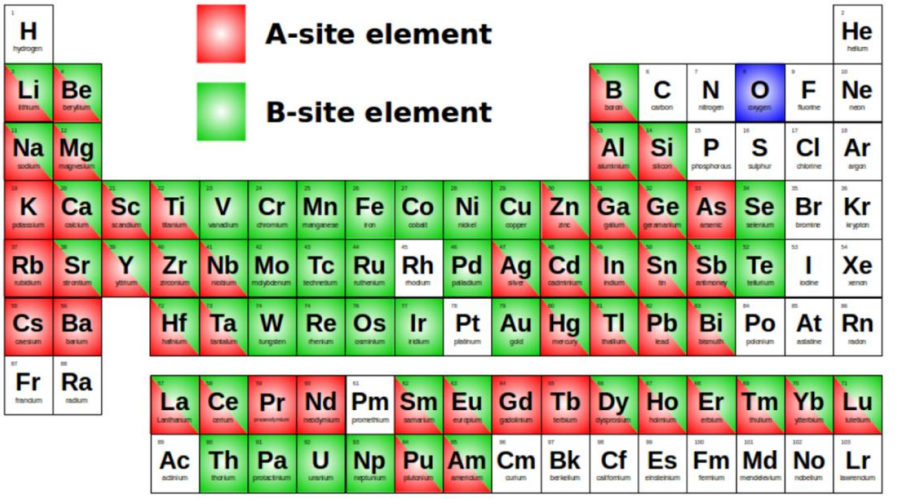

Fig. 2: The A site and the B site elements.[13]

Table 1: Packages of Python used in this work.

Library	Description
Numpy	NumPy is a Python library that provides a wide range of array applications and matrix operations.
Pandas	Pandas is an open source data structure and data analysis tool that provides excellent performance and ease of use.
Matplotlib	Plotting library, using visualization results and data
Mendeleev	Application programming interface is provided to access various attributes of the elements in the periodic table.
Pymatgen	It is an efficient open source Python library for material analysis. Its main feature is to express Element, Site, Molecule, Structure, highly flexible electronic structure analysis, such as density of state and band structure.
Scikit-Learn	A Python machine learning library aiming to speed up ML operations by making ML algorithms easier to implement. It can assist in the division of data into training and test sets, as well as the training of models, make predictions, and evaluate models.
Pytorch	ML library focusing on deep learning.
Tensorflow	TensorFlow is a Python library for fast numerical computing and large-scale ML.
Mlxtend	mlxtend as a supplement and auxiliary tool for sklearn.

dataset gathered by Emery and the Mendeleev dataset.[12] There are processes of how to build the screening and feature engineering dataset. By choosing 49 elements for A and 62 elements for B respectively, compile all possible candidate pools by taking the combination of A and B cations and increasing x from 0.05 to 0.95 at 0.05 intervals. The database was subject to the follow rules: The radius of A should bigger than B. Basic chemistry rules - charge neutrality and Pauling's valency rule.[13]

2.2.1 Feature screening database-Inorganic crystal structure database (ICSD)

The ICSD contains known inorganic crystal structures published since 1913. The 1958 perovskites were simply obtained by searching in this database and removing duplicates. The database has detailed features, which not only including atomic coordinates, specific composition, but also chemical names, molecular formulas, unit cells, space groups, and completeness, atomic parameters.[13]

2.2.2 Normalization

Normalization processing is a fundamental data mining task. Different dimensions and dimensional units are frequently used in different evaluation indications. This condition will have an impact on the data analysis outcomes. To get rid of the dimension between indicators, to tackle the problem of data indicator comparability, data standardization is required. The indicators are in the same order of magnitude once the raw

data is processed via data standardization, making them ideal for extensive comparative comparison. Two commonly used normalizing procedures are as follows:[14]

A) Min-Max Normalization

It is known as dispersion standardization, it is a linear transformation of the source data that maps the outcomes value to [0-1]. The conversion function is as follows in Eq. (1):

$$x' = \frac{x - Min}{Max - Min} \quad (1)$$

B) Z-score standardization method

This method gives the mean and standard deviation of the original data to standardize the data. The processed data conforms to the standard normal distribution, that is, the mean is 0, the standard deviation is 1. This work uses the Minmaxscaler.

2.2.3 One-hot categorical feature encoding

One-hot encoding was an option considered for encoding the crystal structure categorical feature numerically. This would have been done using the OneHotEncoder() preprocessing function from scikit-learn.

2.2.4 Dataset split: training and testing data

By using train_test_split , 70 : 30 split, where 30% of the dataset was set aside for testing. 100 separate iterations were trained and validated by splitting each subdatabase into 80% training and 20% test sets.

2.3 Feature engineering

Screen out the characteristics that determine whether the material is a perovskite material, so as to narrow the forecast range in the future. SISSO is a systematic approach to discover material attribute descriptors in a dimensionality reduction framework based on compressed sensing. Processes large and related feature Spaces and converges from feature combinations related to material properties of interest to optimal solutions.

In addition, SISSO can give stable results even when the training set is small. The method is based on a quantitative prediction of ground state enthalpy of 8-byte binary materials (using ab initio data) and applied to a demonstration example of predicting binary metal/insulator classification (using experimental data). The SISSO algorithm can not only predict the properties of materials, but also give physical interpretable descriptors. It is mainly divided into two parts: feature space construction and searching for optimal descriptors.[15] There are 55 elemental features, 11 compositional features, 5 feature combinations, In the theory, can construct many descriptors, but in this work, only need to find the 10 SISSO descriptors which have the better performance than others, as shown in Tables 2-5.

2.4 Web scraper

Implemented an automatic web scraping tool to search the results material on the Internet, any mention of these materials will be deleted. The aim of Automated mining of the web to ensure the novelty of predicted perovskites. For example, the ICSD does not contain any material in the quaternary system Sr-Pb-Ti-O, however, the predicted composition $(Sr_{0.7}Pb_{0.3})TiO_3$ was found on the internet. Therefore, it is essential delete all the materials of Sr-Pb-Ti-O in format from the predicted dataset. Please find the code in Fig. S1.[13]

Table 2: 55 elemental features.

Index	Feature	Index	Feature	Index	Feature	Index	Feature
1	N_A^{atom}	15	R_O^{atom}	29	$V_{B'}$	43	a_B
2	$N_{A'}^{atom}$	16	X_A	30	V_O	44	a_B
3	N_B^{atom}	17	$X_{A'}$	31	λ_A	45	a_O
4	$N_{B'}^{atom}$	18	X_B	32	$\lambda_{A'}$	46	Grp_A
5	N_O^{atom}	19	$X_{B'}$	33	λ_B	47	Grp_A
6	N_A^{men}	20	X_O	34	$\lambda_{B'}$	48	Grp_B
7	$N_{A'}^{men}$	21	M_A	35	λ_O	49	$Grp_{B'}$
8	N_B^{men}	22	$M_{A'}$	36	P_A	50	Grp_O
9	$N_{B'}^{men}$	23	M_B	37	$P_{A'}$	51	Row_A
10	N_O^{men}	24	$M_{B'}$	38	P_B	52	$Row_{A'}$
11	R_A^{atom}	25	M_O	39	$P_{B'}$	53	Row_B
12	$R_{A'}^{atom}$	26	V_A	40	P_O	54	$Row_{B'}$
13	R_B^{atom}	27	$V_{A'}$	41	a_A	55	Row_O
14	$R_{B'}^{atom}$	28	V_B	42	$a_{A'}$		

Table 3: 11 compositional features.

Index	Feature	Index	Feature
56	x_A	62	$\bar{n}_{B-site}$
57	$x_{A'}$	63	$\bar{n}_O$
58	x_B	64	$\bar{R}^{ion}_{A-site}$
59	$x_{B'}$	65	$\bar{R}^{ion}_{B-site}$
60	x_O	66	R^{ion}_{O-site}
61	$\bar{n}_{A-site}$		

Table 4: 5 feature combinations.

Index	Feature
67	$\bar{R}^{ion}_{A-site} / \bar{R}^{ion}_{B-site}$
68	$\bar{R}^{ion}_{A-site} / \bar{R}^{ion}_{O-site}$
69	$\bar{R}^{ion}_{B-site} / \bar{R}^{ion}_{O-site}$ (octahedral factor)
70	$\bar{R}^{ion}_{A-site} - \bar{R}^{ion}_{B-site}$
71	t

Table 5: 10 SISSO descriptors.

Index	Feature	Index	Feature
72	$\left\|(R^{atom}_{A1})^3 - \|X_{B1} - P_{B1}\|\right\|$	77	$(Row_{A1} + P_{A1}) - \left\|\bar{R}^{ion}_{A-site} - \bar{R}^{ion}_{B-site} - \frac{\bar{R}^{ion}_{A-site}}{\bar{R}^{ion}_{B-site}}\right\|$
73	$\left\|(V_{A1} + R^{atom}_{A1}) - \|Grp_{B1} - P_{B1}\|\right\|$	78	$(Row_{A1} + P_{A1}) - \left\|\frac{\bar{R}^{ion}_{A-site}}{\bar{R}^{ion}_{B-site}} - t\right\|$
74	$(Row_{A1} + R^{atom}_{A1}) - \|X_{A1} - X_{B1}\|$	79	$\sin(R^{atom}_{B1}) + (Row_{A1} + P_{A1})$
75	$\frac{\|R^{atom}_{A1} - \lambda_{B2}\|}{(N^{men}_{A1})^3}$	80	$\frac{1}{(N^{men}_{A1})} - \cos(t)$
76	$\left\|(V_{A1} + R^{atom}_{A1}) - \|X_{B1} - P_{B1}\|\right\|$	81	$\frac{1}{(N^{men}_{A1})} + \left\|\bar{R}^{ion}_{A-site} - \bar{R}^{ion}_{B-site} - \frac{\bar{R}^{ion}_{A-site}}{\bar{R}^{ion}_{B-site}}\right\|$

2.5 Machine learning method

2.5.1 K-nearest neighbors

Similar things are assumed to exist nearby by the KNN algorithm. To put it another way, things that are similar are close to one other.

There is already known classified data. When new data is entered, it first calculates the distance from each point in the training data, then selects the k points closest to the training data to determine which category these points belong to, and then classifies the new data using the minority obeys the majority principle. This work uses Euclidean distance as a distance metric as follows in Eq. (2).

$$SE(p, q)=1/(1+dE(p ,q)) \quad (2)$$

where p and q are two composition points and dE(p, q) is the Euclidean distance between p and q,[13] and in this work, k=5 is chosen.

2.5.2 Gradient boosting classifiers

GBC is a specific type of algorithm used for classification tasks, which combines many weak learning models to create a strong prediction model. The steps to implement a gradient boosting classifier are as follows:

(1) Fit the model using Scikit-Learn.

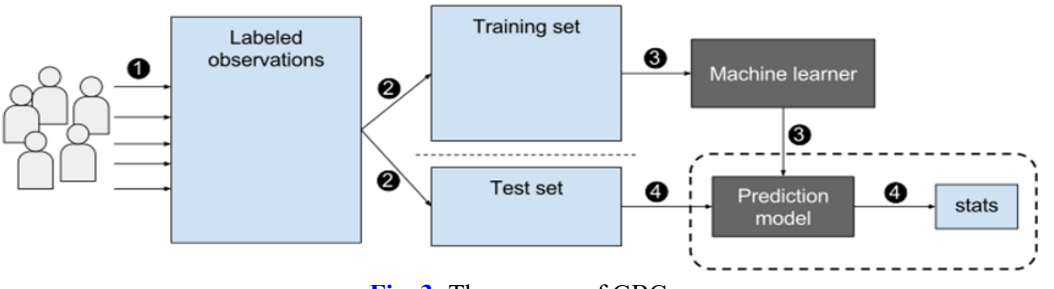

Fig. 3: The process of GBC.

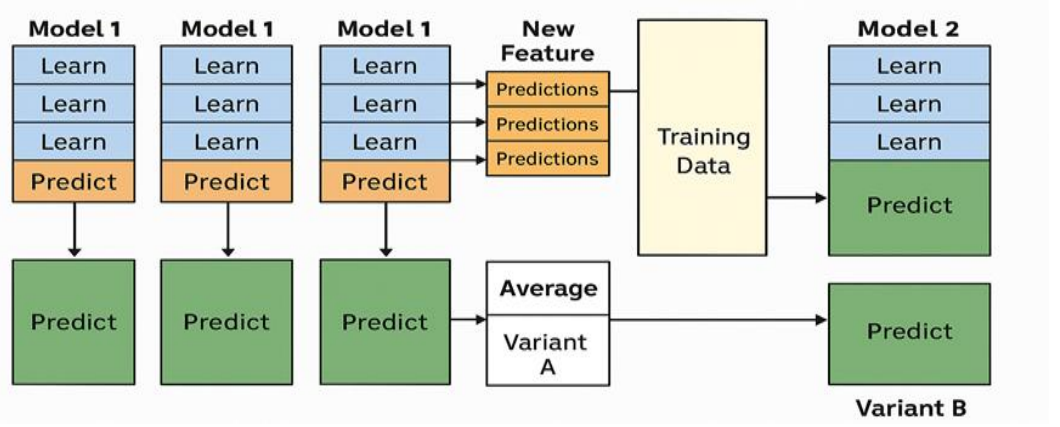

Fig. 4: Process of stacking model.

(2) Adjust the parameters and hyperparameters of the model and try to obtain the best accuracy of the model by adjusting various parameters.
(3) Make predictions. Use this function after fitting the classifier by "predict" to make predictions in Scikit-Learn, and then compare the predicted results with the actual labels.
(4) Interpret the results.

2.5.3 Stacking

For each round of 5-fold, model 1 has to do 5 times of training and prediction. Every fold will generate a small train and a small test. Use Model 1 to train the small train, and then predict the small test. This kind of prediction is performed 5 times, and the generated prediction value is exactly the same as the length of the train data. The prediction value is generated by Model 1 and exists first. Then Model 1 trained by the fold small train every 1 time has to predict all our Test Data, (because Test Data does not add 5-fold, so every time it is all), this action can get the predicted value after 5 times. This work uses it to predict the average value. The Model 1 on the first layer has fulfilled its mission. There will be other Models on the first layer, such as Model 2, 3, and walk the same way. It's on the second floor. The predicted value from 5-fold is used as Train Data to train the second layer model (Fig. S2). The parameters of stacking model are shown in Table 6.

2.5.4 Generative adversarial network

The main structure of GAN includes a generator G (Generator) and a discriminator D (Discriminator). This work applied GAN in the process of predicting the bandgap of new perovskites, which can improve the performance of the machine learning classifier, thereby largely enhancing the classification and prediction ability of the machine learning classifier. Please find the code in Fig. S1, and the parameters of GAN model are shown in Table 7.

2.5.5 Variational autoencoder network

Here are process of variational autoencoder network (VAE).

Table 6: Parameters of stacking model.

ML model	SVR	Linear Model	XGBoost	KNN	ExtraTree	DT
Base Learners	✓	✓	✓	✓	✓	✓
Meta_regressor		Linear Model				
Use_features_in_secondary	TRUE					

Table 7: Parameter of GAN.

Parameter	Value			
Learning rate	1.00E-03	3.00E-03	1.00E-02	
Layer Depth	7	13	16	
Loss criterion	Mean squared error			
Optimisation funcion	SGD	Adam	Adaptive moment estimation with decoupled weight decay (AdamW)	
Activation function	Tanh	Rectified linear unit (ReLu)	Sigmoid	LeakyReLu

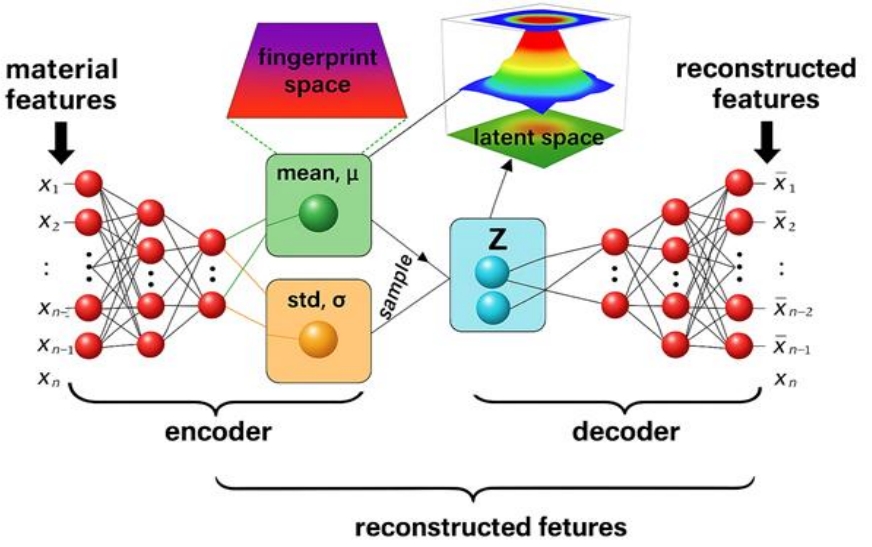

Fig. 5: The working process of VAE.[13]

(1) Import the necessary modules. In this method, Numpy, Matplolib and TensorFlow functions need to be used.
(2) Define the Variational Autoencoder class. Use __init__ class methods to define hyperparameters, such as learning rate, batch size, placeholders for input, encoder and decoder network weights.
(3) Create encoder and decoder. The first layer of the encoder receives input and generates a decreasing latent representation of the input; the second layer maps the input to a Gaussian distribution.
(4) Train and test the model. Convert features into vectors, construct material fingerprints through these multi-dimensional vectors, and compare the similarity of materials by calculating the sum of all the resounding distance differences of each fingerprint—evaluate whether the potential perovskite materials screened out can be used as SOFCs electrode.[16]

The VAE vector is sent to the Stacking model for analysis, and then through ICSD query, experimental results and crystal structure verification, toxicity and production cost analysis, the final screening pool can be used as a potential perovskite material for SOFCs.[13] Eqs. (3) and (4) are as follows:

$$\text{For } \left(A_{1-x}A'_{x}\right)BO_3,\ \mu = R_B/R_O,\ t = \frac{[(1-x)R_A+R_o+xR_{A'}]}{\sqrt{2}(R_B+R_O)} \quad (3)$$

$$\text{For } A\left(B_{1-x}B'_{x}\right)O_3,\ \mu = \left[(1-x)R_B + xR_{B'}\right]/R_O,\ t = \frac{R_o+R_A}{\sqrt{2}(1-x)R_B+R_o+xR_{B'}} \quad (4)$$

2.6 Evaluation metrics

By evaluating which machine learning model is the optimal classifier to discover new perovskites for SOFCs, this work applied the following evaluation metrics.

(1) Precision is a measure of quality which determines what fraction of selected items is relevant. It is formulated as Eq. (5):

$$\text{Precision} = \frac{\text{True positives}}{\text{True posotives+False positives}} \quad (5)$$

(2) Recall is a measure of quantity that indicates what fraction of relevant items that was actually selected. It is postulated as Eq. (6):

$$\text{Recall} = \frac{\text{True positives}}{\text{True posotives+False negatives}} \quad (6)$$

(3) F1-score determines the model's accuracy on a dataset by combining precision and recall values.
When precision and recall are very different, the F1-score tends towards the lower figure. The F1-score is formulated as Eq. (7):

$$\text{F1} - \text{score} = \frac{2*\text{precision}*\text{recall}}{\text{precision+recall}} \quad (7)$$

In order to further confirm the prediction accuracy, this work used root mean square error (RMSE), average absolute error (MAE), and R^2 as the evaluation metrics, which are described as follows:

At the same time, for all the following calculation formulas, n represents the sample size, y_i represents the true value of the i sample, and $\hat{y}_i$ represents the predicted value of the i sample.

The RMSE is a typical indicator of a regression model. It is used to indicate how much error the model will produce in the prediction. For larger errors, the weight is higher. The smaller the RMSE, the better. The formulation of RMSE is

displayed as follows in Eq. (8):

$$\mathrm{RMSE} = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(y_i - \hat{y}_i)^2} \quad (8)$$

The MAE is used to measure the average absolute error between the predicted value and the true value. The smaller the MAE, the better the model. It is defined as shown in Eq. (9):

$$\mathrm{MAE} = \frac{1}{m}\sum_{i=1}^{m}|y_i - \hat{y}_i|^2 \quad (9)$$

K-learn uses indicators by default when implementing linear regression. The larger the model, the better. It is defined as shown in Eq (10):

$$R^2 = 1 - \frac{\sum_i^2(\hat{y}_i - y_i)^2}{\sum_i^2(\bar{y}_i - y_i)^2} \quad (10)$$

The advantage is that the results are normalized, making it easier to see the gap between the models.

2.7 Receiver operating characteristic (ROC) with cross validation

The ROC curve depicts the characteristics of the false positive rate (FPR) on the X axis and the true positive rate (TPR) on the Y axis. Among them, the false positive rate FPR=FP/(FP+TN), that is, the number of negative samples predicted to be positive/the actual total number of negative samples, and the true rate TPR=TP/(TP+FN), that is, the positive samples predicted to be positive Quantity/actual total number of positive samples. The higher the true rate, the better, and the lower the false positive rate, the better, so the top left corner of the graph canvas is the most ideal result, with FPR being 0 and TPR being 1. This means that the larger the area under the ROC curve (AUC), the better (Fig. 6).

3. Results and discussion

As is illustrated in the workflow of this work, the architecture can be divided into five parts, feature engineering for feature learning and dimension process of the perovskites for SOFCs is depicted in 3.1; classification model comparison to select out the optimal model for new perovskites discovery is performed in 3.2; unsupervised VAE fingerprint for a clearer representation of every perovskite and potential perovskites in 3.3; GAN-based property prediction for higher precision in phase identification and bandgap prediction in 3.4 and new perovskite discovery for cathode of SOFCs via the machine in 3.5. This work selects 21 physical features of the perovskites as the input to conduct feature engineering, machine learning classification and VAE fingerprint. Then this work applied the results mentioned above to predict the bandgap and formation energy assisted by GAN and obtained the structured formula of 1980 new perovskites for cathode of SOFCs by machine learning classifier.

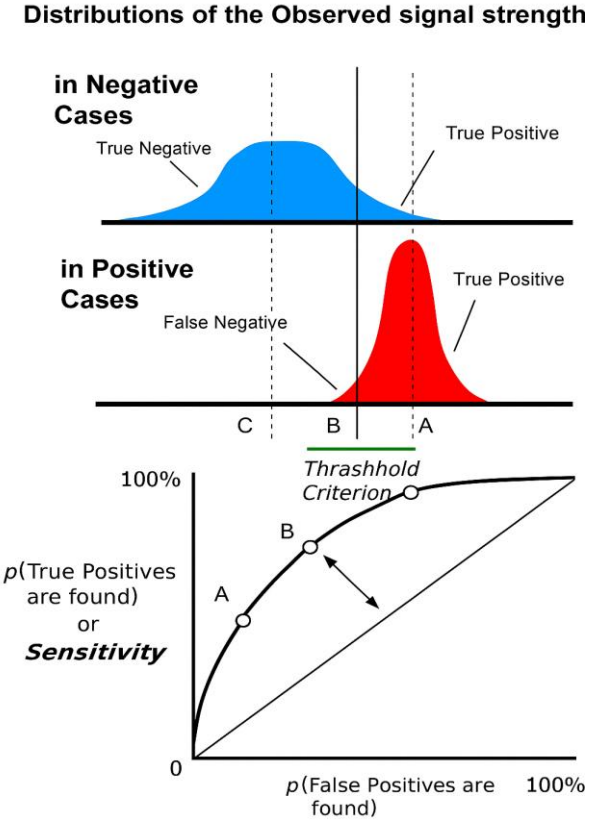

Fig. 6: How the points in the ROC coordinate system move as the threshold is adjusted.

3.1 Feature engineering

It is of great significance to establishing a T a solid, accurate and sufficient collection of candidate compounds for learning experimental material features or classification of perovskites materials for SOFC cathodes or classification of perovskites materials for cathode in SOFCs. In Python, this work utilized pymatgen, an open source python library, to analyze data in database. Pymatgen can use ICSD statistics to guess the most likely oxidation state of an element in a compound. This may result in a partial oxidation state, which must be enlarged in order to estimate the element's effective ionic radius and relative crystal position in the mean cell in this work of a mixed valence component element. This work aims to solve the valences of individual ions by exploiting the common oxidation states of the mixed valence elements in question by first considering the lowest feasible super monomer associated with the integer formula. If it can't find a solution, it may have oxidation states that aren't as prevalent. When the algorithm discovers that the same fractional oxidation state can be produced by many valence configurations, the minimum number of oxidation states and the minimum selection ion

radius change. This work expect that perovskite structures are more likely to be preserved under these conditions, because if there are large component cations with different sizes, octahedral units share the face, making the perovskite lattice more distorted. Mixed valence state is highly dependent on elements, composition, vacancy, etc, and requires experimental characterization to carry out accurate research. This algorithm provides systematic estimates of the mean ionic radii of these elements, allowing us to compute t, μ and develop feature spaces. The two descriptors t and μ of $(A1-xA'x)BO3$ type compositions are computed as follows in Eqs. (11) and (12):

$$t = \frac{[(1-x)R_A + xR_{A'} + R_O]}{\sqrt{2}(R_B + R_O)} \quad (11)$$

$$\mu = R_B/R_O \quad (12)$$

Similarly, for ABO_3 type perovskites, Eqs. (13) and (14) are as follows:

$$t = \frac{(R_A + R_O)}{\sqrt{2}[(1-x)R_B + xR_{B'} + R_O]} \quad (13)$$

$$\mu = [(1-x)R_B + xR_{B'}]/R_O \quad (14)$$

where R_A, $R_{A'}$, R_B, $R_{B'}$, and R_o are the Shannon's ionic radius of A, A′, B, B′ cations and oxygen anion, respectively.

For mixed valence elements, the mean ionic radius is employed. This work obtained 21 features for feature engineering including elemental and compositional features, and part of them can construct t and μ. SISSO is known as a data-driven screening approach. Specifically, this work has 21 primary features. This work identified 10 excellent descriptors characterizing perovskite formability by training the SISSO algorithm on roughly 100 randomly generated compositions from the database and considering the size of subspace.[13]

3.2 Classification model comparison

In order to choose a suitable and accurate classifier to classify the range of the candidate compounds, this work tested 5 different model and using a stacking model to find the best model. These include DT, RF, SVM and GBC. Besides, this work applied model stacking method to improve the overall prediction accuracy which can compass the all advantages of the four machine learning algorithms mentioned above.[13] Similar to individual models, were assessed by using sub-databases created by combining 227 non-perovskites with 250 randomly picked perovskites from a total of 2016 perovskites. The purpose of doing so is to maintain category balance in a single training set, minimizing potential overfitting to a single category. It also guarantees that all available peroxides are used rather than evaporating. By separating each sub-database into a 70% training set and a 30% test set, all five algorithms are trained and verified in 100 separate iterations. The stacking model was chosen to proceed because it offered the greatest results, with tenfold CV and test mean classification accuracies of 96% (±0.09%). Table 8 provides more information on these measures. As is shown in Fig. 7, confusion matrix of DT, RF, SVM, GBC and stacking model have been listed and receiver ROC curves in Fig. 8, respectively, as obtained by tenfold CV. Moreover, Fig. 9 show the top 10 most important features, and Fig. S3 shows the full feature importance. The comparison of 5 different methods. So, it is clear that the stacking model has the best performance have used it to classify the 600,000 datasets.

3.3 Unsupervised material fingerprinting

The similarity and feature relativity can be integrated into a fingerprint map for better comparison. In this part, this work constructed a material fingerprint based on variant autoencoder (VAE). According to the fingerprint as designed, two or more compounds, specifically perovskites here will get close which is contributive to select out the optimal potential new perovskites for cathode in SOFCs, by contrast, perovskites with components which are quite different will be distant and almost has no relation. VAE is utilized here for an unsupervised technique to sear material fingerprint as mentioned above, which refers to the code itself. VAE consist of an encoder and a decoder, the encoder can compress

Table 8: Performance of five models.

Model	GBC	RF	DT	SVM	Stacking model
Accuracy	0.939 (±0.009)	0.907 (±0.007)	0.917 (±0.013)	0.924 (±0.007)	0.960 (±0.0009)
Precision	0.947 (±0.019)	0.902 (±0.092)	0.943 (±0.003)	0.917 (±0.035)	0.950 (±0.0006)
Recall	0.946 (±0.021)	0.881 (±0.114)	0.943 (±0.003)	0.914 (±0.042)	0.950 (±0.0018)
F1-score	0.946 (±0.001)	0.879 (±0.013)	0.943 (±0.001)	0.913 (±0.003)	0.950 (±0.0006)

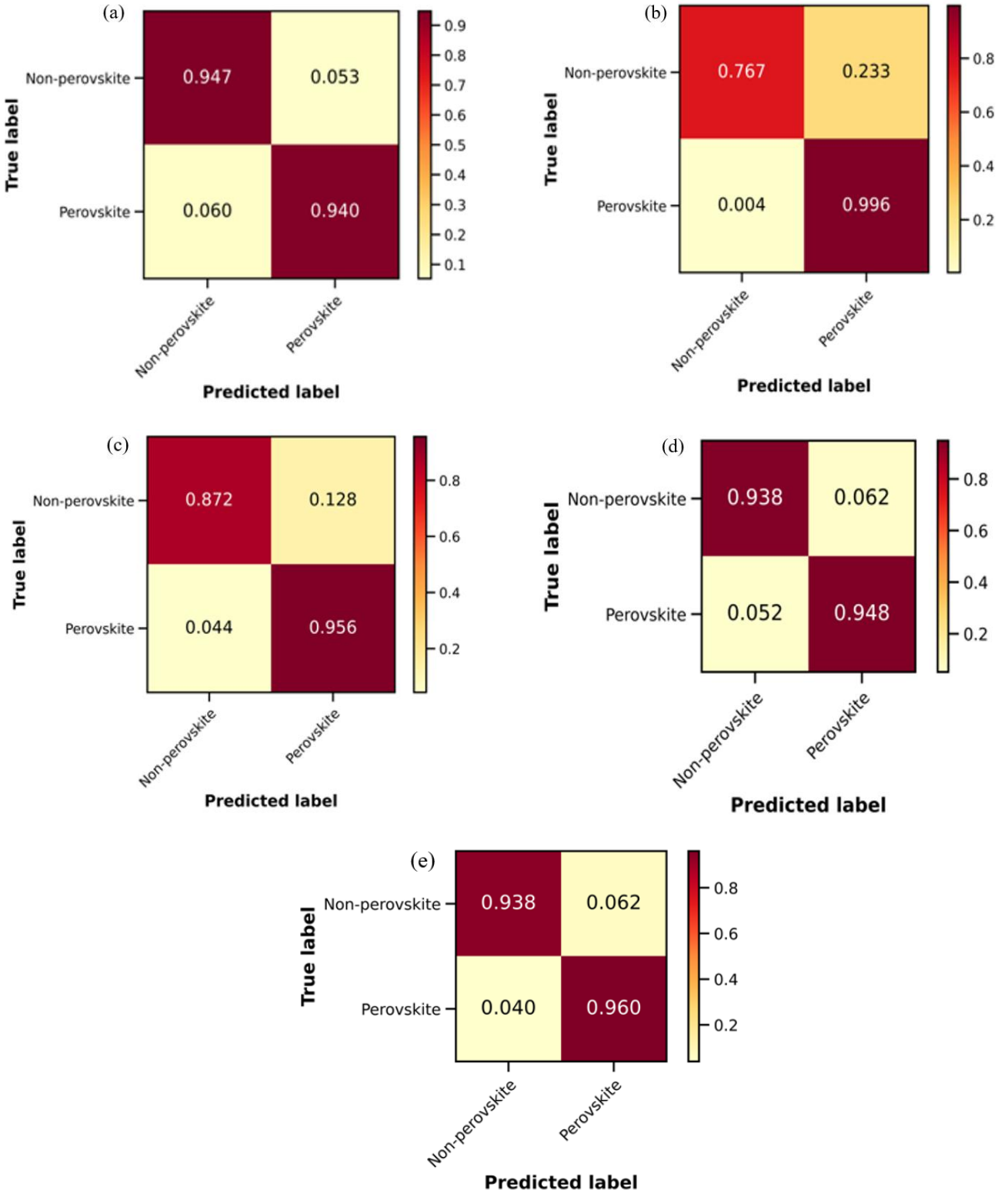

Fig. 7: Confusion matrix of different methods: (a) DT, (b) RF, (c) SVM, (d) GBC, (e) Stacking model.

discrete features of materials into potential vectors. However, a fundamental problem with this strategy, however, is that potential Spaces can form blind spots that may have nothing to do with the real material when decoding.

Compared with common since the encoder, the variational since the encoder (VAEs) produced by forcing encoder and a group of average mu sigma related vector to solve this problem, this group of mean mu and sigma is very similar to the standard gaussian distribution (Z ~ N (0, 1)), then carries on the sampling to get vector potential, makes the latent space is continuous. This indicates that the latent code of the material is not unique, but randomly dispersed around a center (Z ~ N

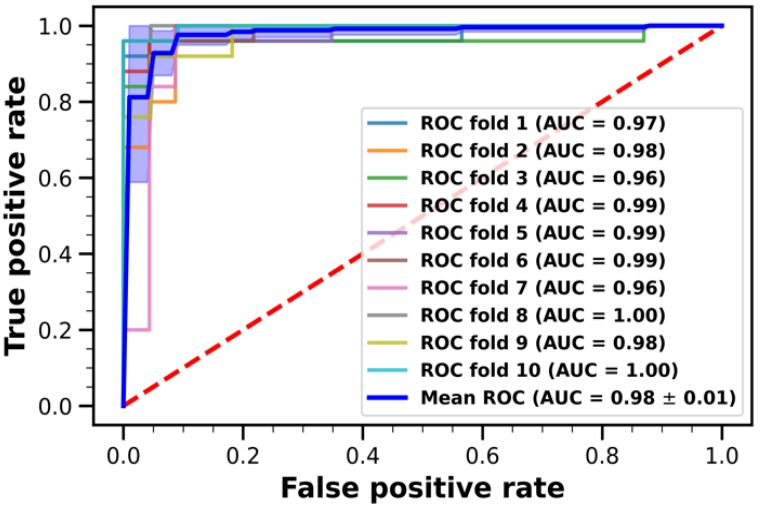

Fig. 8: 10-fold CV ROC curve.

(μ, σ^2)) determined by the mean vector, allowing the decoder to observe the entire latent space during training and allowing the encoder to produce interpretable representations. Dynamic fingerprinting is resilient owing to the unpredictability of the latent embedding T, but due to its non-uniqueness, it is not suitable for material fingerprinting. Thus, the mean vector (μ) generated by the encoder is used as the material fingerprint.[13]

3.4 GAN-based phase identification and bandgap prediction

A few of perovskites exhibit temperature-dependent phase transitions. Coexisting phases can appear at the same temperature. Duplicate chemical compositions are not allowed to appear in the database,[17] in order to the elimination of additional phases. It is found that a strong relationship between at least two distinct symmetries, as revealed by low Euclidean distances (SE > ~ 95%), may indicate the presence of potential phase transitions in the material. The described method can predict 2 or more phases as shown in Fig. 10.[13]

This work used GAN as mentioned above to improve the prediction performance of the bandgap prediction. The material properties of band gap and phase transition details can be recovered by proximity analysis of the fingerprint topology of the VAE assembly. Thus, this work applied bandgap prediction to ensure the uniqueness of each predicted perovskite. After utilizing GAN, the R^2 has been improved from 0.955 to 0.970 as illustrated in Fig. 11 in bandgap prediction, which manifests that GAN can improve the overall prediction performance in bandgap prediction.

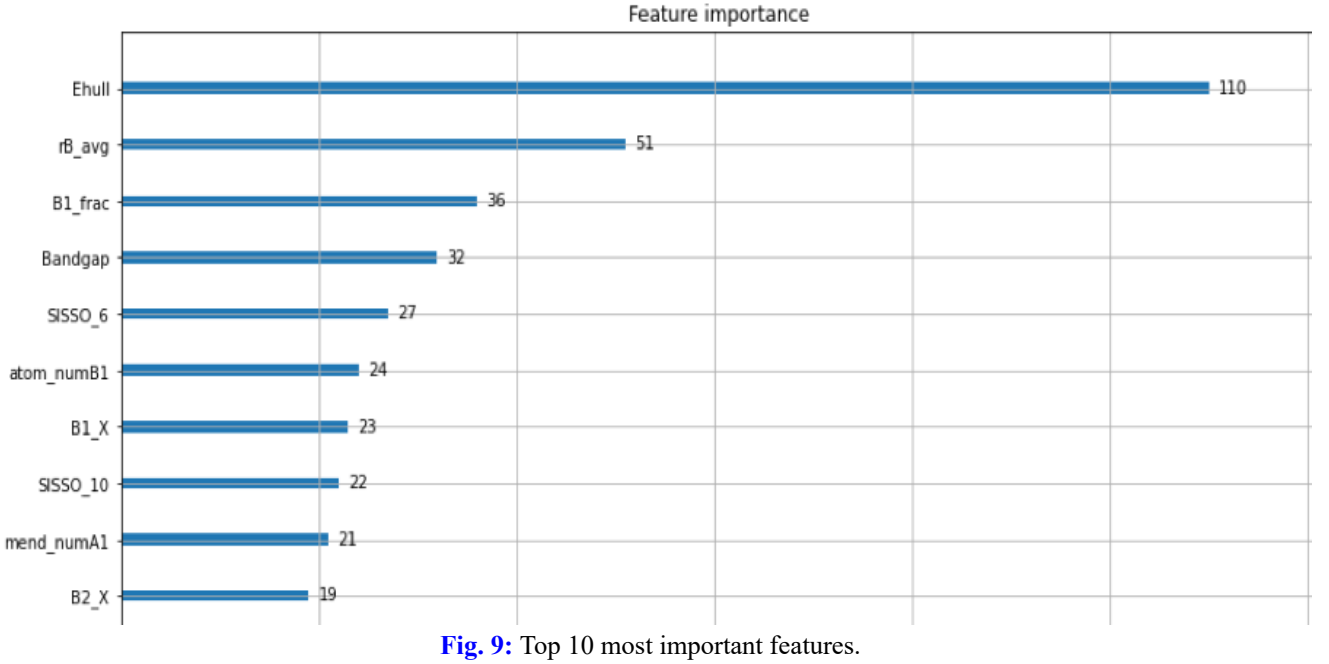

Fig. 9: Top 10 most important features.

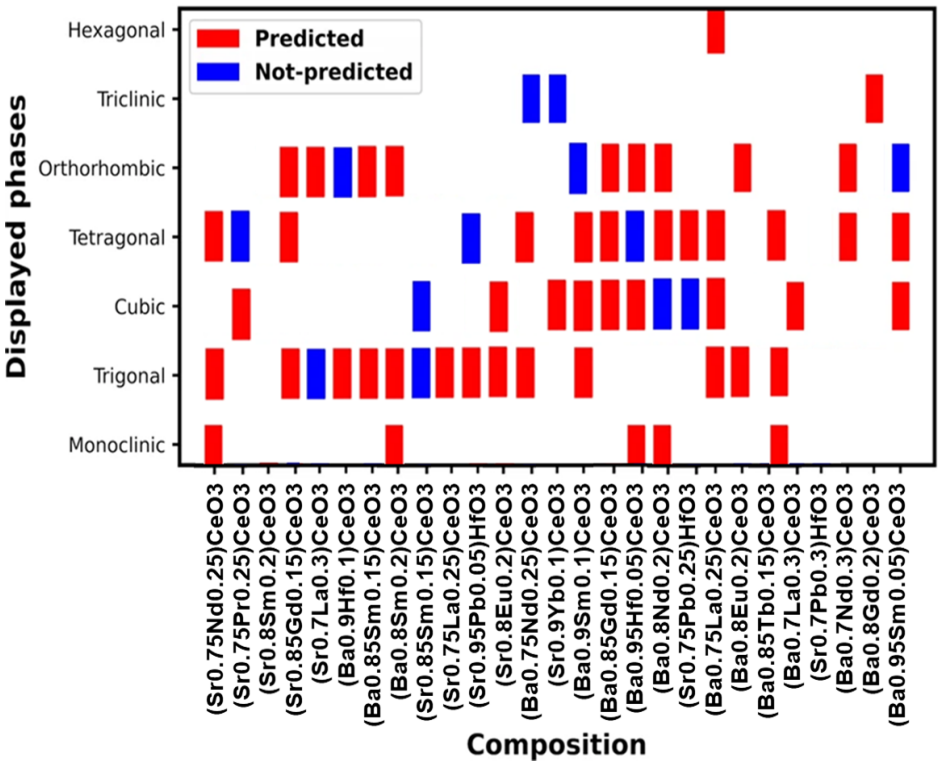

Fig. 10: Phase identification of perovskites via VAE.

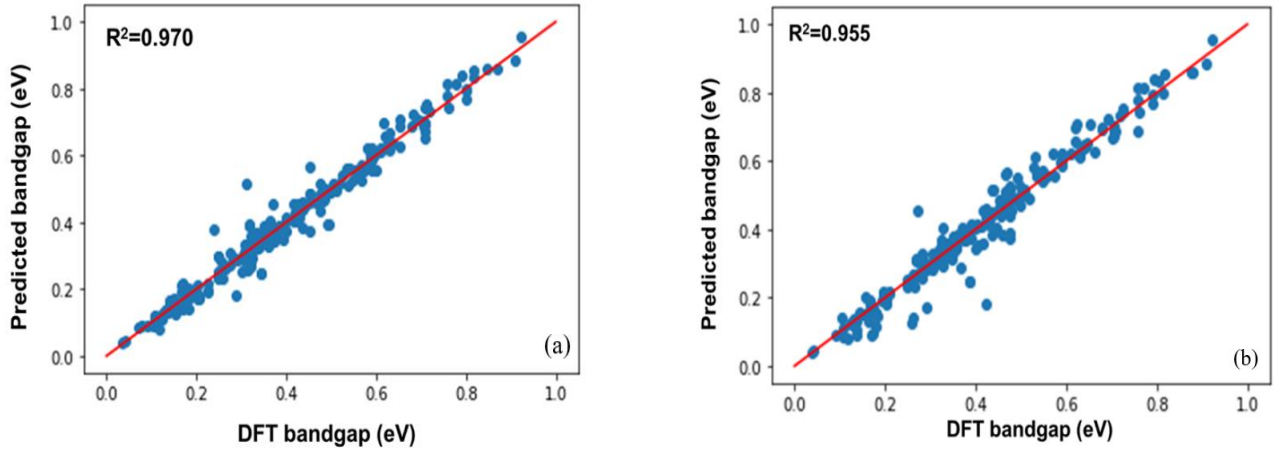

Fig. 11: Bandgap prediction results: (a) GAN enhanced and (b) No GAN.

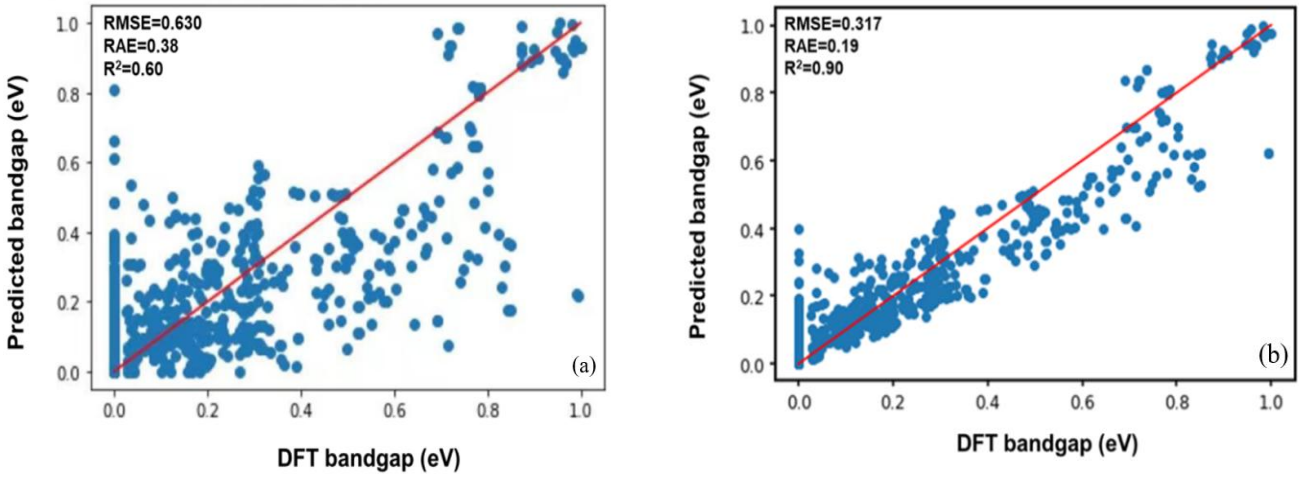

Fig. 12: Bandgap prediction results: (a) SVM and (b) Stacking model and VAE.

This work subsequently evaluated the prediction performance of bandgap and the results can be seen in Fig. 12. A much better regression can be achieved after applying the optimal stacking model for classifier in combination with VAE fingerprint. By contrast, the R^2 of SVM and stacking model in combination with VAE are 0.61 and 0.90, respectively, which shows the necessity to introduce stacking model for classification, bandgap prediction and VAE fingerprint for feature visualization.

Moreover, this work compared the formation energy of the following four models, SVM, stacking model, stacking model and VAE, stacking model and GAN as is shown in Fig. 13. From the R^2 value we can know that when in combination with stacking model & GAN, the optimal regression can be obtained, which manifests that GAN generator and discriminator are effective in prediction various properties of perovskites.

In summary, the results show the significance of GAN, stacking model and VAE in perovskite property prediction which are validated in bandgap and formation energy prediction. The application of GAN can realize a significantly large improvement in regression performance which can contribute to the prediction accuracy.

As shown in Table 9, the stacking model combined with GAN significantly outperformed other models in predicting the formation energy of perovskite materials, achieving the lowest RMSE (0.032), RAE (0.020), and the highest R^2 value (0.97), as shown in Fig. S4. When predicting bandgap, the stacking model with VAE integration also improved predictive accuracy compared to traditional SVM, reducing RMSE from 0.630 to 0.317 and increasing R^2 from 0.61 to 0.90. These results validate the effectiveness of combining stacked machine learning with generative models in enhancing the prediction performance for complex material properties, as demonstrated in this study.

3.5 New perovskites discovery for cathode in SOFCs

In this part, the fingerprint of predicted perovskites has been extracted by sending associated features to VAE fingerprint generator. This brings candidate components and experimental observations to a common basis, the fingerprint space, in which unexplored perovskites falling near the target experimental material are identified as similar components of

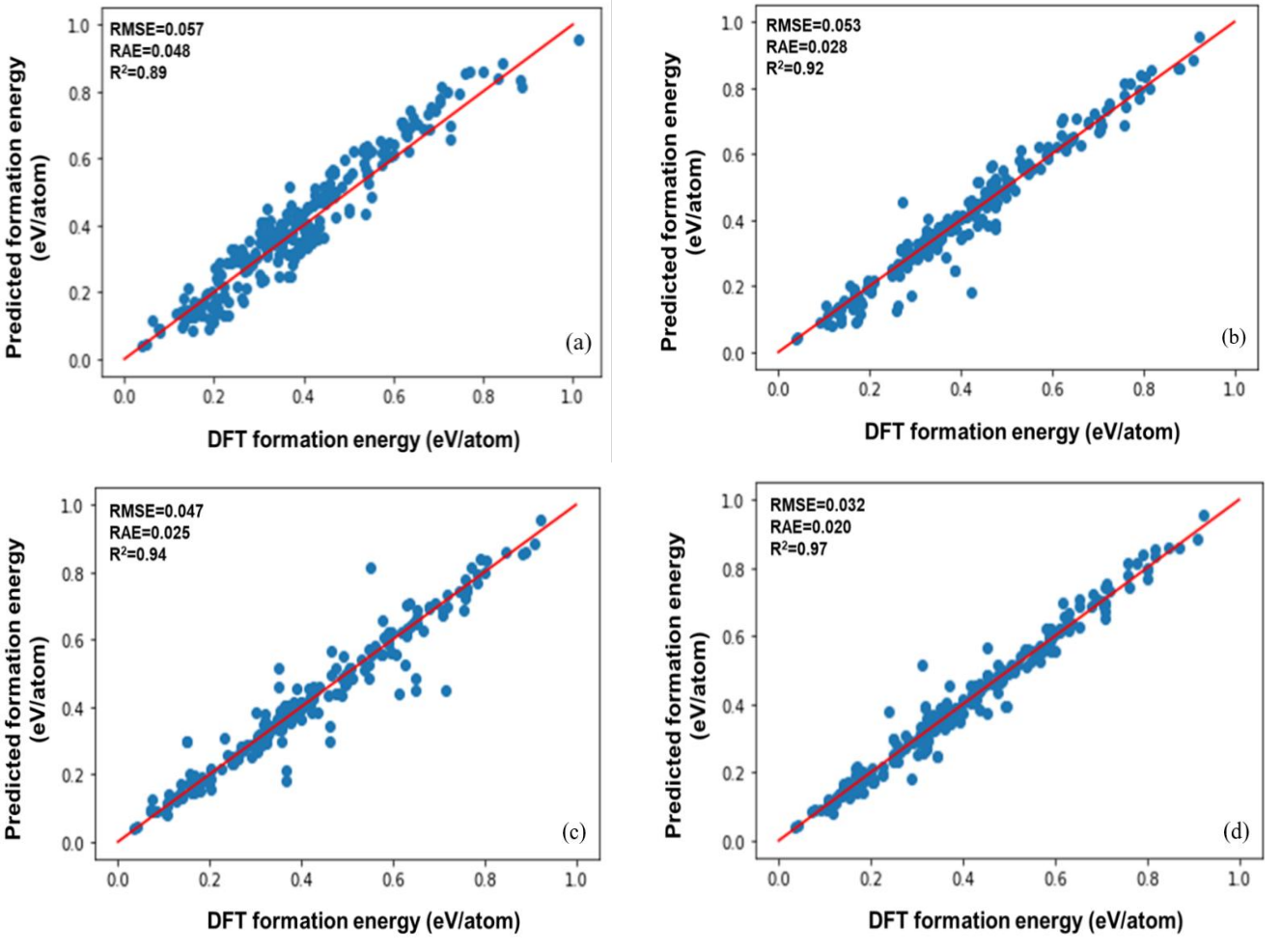

Fig. 13: Formation energy results: (a) SVM, (b) Stacking model, (c) Stacking model and VAE, and (d) Stacking model and GAN.

Table 9: The summary of prediction results.

Methods	SVM	Stacking model and VAE	Stacking model and VAE	SVM	Stacking model	Stacking model and GAN
Prediction of target	Bandgap	Bandgap	Formation energy	Formation energy	Formation energy	Formation energy
RMSE	0.630	0.317	0.047	0.057	0.053	0.032
RAE	0.380	0.190	0.025	0.048	0.028	0.020
R^2	0.61	0.90	0.94	0.89	0.92	0.97

that material. This work first tested the validity of a set of fingerprints for predicting perovskites. Finally, this work obtained 1981 new potential perovskites for cathode in SOFCs, and some of them are shown in Table 10. The complete list of new perovskites is attached in Supporting information.

As is shown in Table 10, StructuredFormula is the structure formula of perovskites, Mean_classification_prob shows the probability of the predicted perovskites to be the candidate of cathode in SOFCs, Euclidean Distance represents the distance between VAE fingerprint, and the most similar experimental compounds contains five most similar compounds with experimental data that have been discovered.

Table 10: Examples for predicted perovskites.

ID	Structured formula	Mean_classification_prob	Euclidean distance	Most similar experimental compounds
0	$(Sr_{0.6}Pb_{0.4})CeO_3$	0.982963516	0.189451039	('$(Ga_{0.75}Al_{0.25})FeO_3$', '$(Y_{0.6}Cd_{0.4})(VO_3)$', '$(Ga_{0.5}Al_{0.5})FeO_3$', '$(Ga_{0.9}Cr_{0.1})FeO_3$', '$(Ga_{0.9}Mn_{0.1})FeO_3$')
1	$(Sr_{0.65}Pb_{0.35})CeO_3$	0.982963516	0.191398636	('$(Ga_{0.75}Al_{0.25})FeO_3$', '$(Ga_{0.5}Al_{0.5})FeO_3$', '$(Ga_{0.9}Cr_{0.1})FeO_3$', '$(Ga_{0.9}Mn_{0.1})FeO_3$', '$(Mn_{0.9}Zn_{0.1})TiO_3$')
2	$(Sr_{0.55}Pb_{0.45})CeO_3$	0.98268341	0.192191884	('$(Y_{0.6}Cd_{0.4})(VO_3)$', '$(Ga_{0.75}Al_{0.25})FeO_3$', '$(Ga_{0.5}Al_{0.5})FeO_3$', '$(Ga_{0.9}Mn_{0.1})FeO_3$', '$(Ga_{0.9}Cr_{0.1})FeO_3$')
3	$(Sr_{0.7}Pb_{0.3})CeO_3$	0.982840592	0.197940052	('$(Ga_{0.75}Al_{0.25})FeO_3$', '$(Ga_{0.5}Al_{0.5})FeO_3$', '$(Ga_{0.9}Cr_{0.1})FeO_3$', '$(Mn_{0.9}Zn_{0.1})TiO_3$', '$(Ga_{0.9}Mn_{0.1})FeO_3$')
4	$(Sr_{0.5}Pb_{0.5})CeO_3$	0.982085954	0.199303493	('$(Y_{0.6}Cd_{0.4})(VO_3)$', '$(Ga_{0.75}Al_{0.25})FeO_3$', '$(Ga_{0.9}Mn_{0.1})FeO_3$', '$(Ga_{0.5}Al_{0.5})FeO_3$', '$(Ga_{0.9}Cr_{0.1})FeO_3$')
5	$(Sr_{0.45}Pb_{0.55})CeO_3$	0.98193895	0.210227445	('$(Y_{0.6}Cd_{0.4})(VO_3)$', '$(Ga_{0.75}Al_{0.25})FeO_3$', '$(Ga_{0.9}Mn_{0.1})FeO_3$', '$(Ga_{0.9}Cr_{0.1})FeO_3$', '$(Ga_{0.5}Al_{0.5})FeO_3$')
6	$(Sr_{0.75}Pb_{0.25})CeO_3$	0.981688051	0.210459322	('$(Ga_{0.75}Al_{0.25})FeO_3$', '$(Ga_{0.5}Al_{0.5})FeO_3$',

ID	Structured formula	Mean_classification_prob	Euclidean distance	Most similar experimental compounds
7	$(Sr_{0.4}Pb_{0.6})CeO_3$	0.981794429	0.22440818	'$(Ga_{0.9}Cr_{0.1})FeO_3$', '$(Mn_{0.9}Zn_{0.1})TiO_3$', '$(Ga_{0.9}Mn_{0.1})FeO_3$') ('$(Y_{0.6}Cd_{0.4})(VO_3)$', '$(Ga_{0.75}Al_{0.25})FeO_3$', '$(Ga_{0.9}Mn_{0.1})FeO_3$', '$(Ga_{0.9}Cr_{0.1})FeO_3$', '$(Ga_{0.5}Al_{0.5})FeO_3$')
8	$(Sr_{0.8}Pb_{0.2})CeO_3$	0.981651627	0.227259889	('$(Ga_{0.75}Al_{0.25})FeO_3$', '$(Ga_{0.5}Al_{0.5})FeO_3$', '$(Mn_{0.9}Zn_{0.1})TiO_3$', '$(Ga_{0.9}Cr_{0.1})FeO_3$', '$(Ni_{0.5}Mn_{0.5})TiO_3$')
9	$(Sr_{0.85}Bi_{0.15})CeO_3$	0.982094538	0.229945004	('$(Ga_{0.75}Al_{0.25})FeO_3$', '$(Ga_{0.9}Mn_{0.1})FeO_3$', '$(Ga_{0.9}Cr_{0.1})FeO_3$', '$(Ga_{0.5}Al_{0.5})FeO_3$', '$(Mn_{0.9}Zn_{0.1})TiO_3$')
10	$(Sr_{0.8}Bi_{0.2})CeO_3$	0.981044297	0.238486543	('$(Ga_{0.9}Mn_{0.1})FeO_3$', '$(Ga_{0.75}Al_{0.25})FeO_3$', '$(Ga_{0.9}Cr_{0.1})FeO_3$', '$(Y_{0.6}Cd_{0.4})(VO_3)$', '$(Ga_{0.5}Al^{0.5})FeO_3$')

4. Discussion

4.1 Limitation

The dataset is based on quaternary perovskite, and other perovskites, such as simple perovskite, also can be potential martials used for SOFC. The accuracy of the classifier still needs to be improved, it is better that combine machine learning and DFT. By using DFT to verify, we can achieve a more reliable result. The experimental data were obtained close to room temperature, so the temperature dependence of the crystal structure prediction was not carefully checked. All forecasts are based only on assumptions of environmental conditions.[13]

4.2 Application of machine learning in materials engineering

In materials science, large-scale, high-dimensional data sets are often generated. Machine learning can provide a scalable method to identify patterns in large data sets and extract patterns and trends in the data, which provides for the reverse design of materials A way of thinking. At present, machine learning has been applied to the research and design of many materials, not only the perovskite mentioned in this article, but also other materials. Including metal organic framework materials, biological materials, lithium ion battery materials, thermoelectric materials, catalytic materials, and carbon materials. Not only can it effectively speed up the design and development of new materials, but machine learning can also optimize and improve existing material theoretical calculation methods.[18]

Machine learning can also introduce the concept of analog material discovery. It is possible to find alternatives to certain perovskites that may be toxic, expensive, unstable, or have an impact on the environment. By excluding expensive and toxic chemical components, it can be predicted that alternatives to certain materials. At the same time, this concept and the method of machine learning to predict analog materials can also be applied to more materials in the future, not just limited to perovskites and SOFCs.[13]

4.3 Future study

Machine learning methods still have certain limitations, mainly because of the strong dependence of machine learning on data. In the current material science, with limited data and small data sets, machine learning methods are prone to overfitting, which greatly reduces the generalization ability of the methods.

Since the application of machine learning in materials research and development is still in its initial stage, there is still many areas should be explored in the future.

Here are some suggestions for further study: (1) To

establish a reliable large database containing structural, functional and characterization information by combining theory and experiment; (2) To develop faster and more accurate learning algorithms; (3) To combine machine learning with other physics methods such as density generalization theory, molecular dynamics and Monte Carlo methods.[6,18]

5. Conclusion

Combining multiple machine learning models and generative models to improve prediction efficiency and accuracy. Classification accuracy of 96% with stacked models. Meanwhile, optimize performance prediction by GAN to provide reliable basis for experimental validation. GAN-enhanced bandgap prediction improves R^2 from 0.955 to 0.970. Lastly, using VAE to generate material fingerprints for material similarity analysis and new structure prediction, successfully predicted 1,981 potential quaternary perovskite materials.

Acknowledgments

This work was supported by the Mapua university.

Conflict of Interest

There is no conflict of interest.

Supporting Information

Applicable.

Reference

[1] G. Meng, W. Liu, D. Peng, New solid state fuel cells - green power source for 21st century, *Ionics*, 1998, **4**, 451-462, doi: 10.1007/BF02375890.

[2] S. B. Boshoman, O. S. Fatoba, T. C. Jen, Transition metal oxides as electrocatalytic material in fuel cells: a review, *Engineered Science*, 2023, **25**, 948, doi: 10.30919/es948.

[3] M. K. Skakov, S. K. Kabdrakhmanova, K. Akatan, A. M. Zhilkashinova, E. Shaimardan, M. M. Beisebekov, K. Nurgamit, V. V. Baklanov, Y. T. Koyanbayev, A. Z. Miniyazov, I. A. Sokolov, N. M. Mukhamedova, La_2CuO_4 electrode material for low temperature solid oxide fuel cells, *ES Materials & Manufacturing*, 2023, **22**, 969, doi: 10.30919/es948.

[4] S. B. Adler, Sources of cell and electrode polarisation losses in SOFCs, High-Temperature Solid Oxide Fuel Cells for the 21st Century, Academic Press, Boston, 2016, 357-381, ISBN: 978-0-12-410453-2.

[5] I. R. de Larramendi, N. Ortiz-Vitoriano, I. B. Dzul-Bautista, T. Rojo, Designing perovskite oxides for solid oxide fuel cells, Perovskite Materials - Synthesis, Characterisation, Properties, and Applications, InTech, Rijeka, 2016, ISBN: 978-953-51-2245-6.

[6] Q. Tao, P. Xu, M. Li, W. Lu, Machine learning for perovskite materials design and discovery, *NPJ Computational Materials*, 2021, **7**, 23, doi: 10.1038/s41524-021-00495-8.

[7] Q. Xu, Z. Li, M. Liu, W. Yin, Rationalizing perovskite data for machine learning and materials design, *The Journal of Physical Chemistry Letters*, 2018, **9**, 6948-6954, doi: 10.1021/acs.jpclett.8b03232.

[8] J. Graser, S. K. Kauwe, T. D. Sparks, Machine learning and energy minimization approaches for crystal structure predictions: a review and new horizons, *Chemistry of Materials*, 2018, **30**, 3601-3612, doi: 10.1021/acs.chemmater.7b05304.

[9] N. A. Jalil, H. J. Hwang, N. M. Dawi, Machines learning trends, perspectives and prospects in education sector, *Proceedings of the 3rd International Conference on Education and Multimedia Technology*, 2019, 201-205, doi: 10.1145/3345120.3345147.

[10] M. Sudhi, V. K. Shukla, D. K. Shetty, V. Gupta, A. S. Desai, N. Naik, B. Z. Hameed, Advancements in bladder cancer management: a comprehensive review of artificial intelligence and machine learning applications, *Engineered Science*, 2023, **26**, 1003, doi: 10.30919/es1003.

[11] Z. Yang, Y. Liu, Y. Zhang, L. Wang, C. Lin, Y. Lv, Y. Ma, C. Shao, Machine learning accelerates the discovery of light-absorbing materials for double perovskite solar cells, *The Journal of Physical Chemistry C*, 2021, **125**, 22483-22492, doi: 10.1021/acs.jpcc.1c07262.

[12] A. A. Emery, C. Wolverton, High-throughput DFT calculations of formation energy, stability and oxygen vacancy formation energy of ABO_3 perovskites, *Scientific Data*, 2017, **4**, 170153, doi: 10.1038/sdata.2017.153.

[13] A. Ihalage, Y. Hao, Analogical discovery of disordered perovskite oxides by crystal structure information hidden in unsupervised material fingerprints, *NPJ Computational Materials*, 2021, **7**, 75, doi: 10.1038/s41524-021-00536-2.

[14] R. Jacobs, T. Mayeshiba, J. Booske, D. Morgan, Material discovery and design principles for stable, high activity perovskite cathodes for solid oxide fuel cells, *Advanced Energy Materials*, 2018, **8**, 1702708, doi: 10.1002/aenm.201702708.

[15] R. Ouyang, S. Curtarolo, E. Ahmetcik, M. Scheffler, L. M. Ghiringhelli, SISSO: a compressed-sensing method for identifying the best low-dimensional descriptor in an immensity of offered candidates, *Physical Review Materials*, 2018, **2**, 083802, doi: 10.1103/physrevmaterials.2.083802.

[16] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, *Nature*, 2015, **521**, 436-444, doi: 10.1038/nature14539.

[17] M. Li, X. Tang, S. Zeng, Q. Liu, Y. Jiang, T. Zhang, W. Li, Large electrocaloric effect in lead-free $Ba(Hf_xTi_{1-x})O_3$

ferroelectric ceramics for clean energy applications, *ACS Sustainable Chemistry & Engineering*, 2018, **6**, 8920-8925, doi: 10.1021/acssuschemeng.8b01277.

[18] W. Wu, Q. Sun, Applying machine learning to accelerate new materials development, *Scientia Sinica Physica, Mechanica & Astronomica*, 2018, **48**, 107001, doi: 10.1360/SSPMA2018-00073.