Journal of Energy Chemistry 60 (2021) 351–359

Contents lists available at ScienceDirect

Journal of Energy Chemistry

journal homepage: www.elsevier.com/locate/jechem

http://www.journals.elsevier.com/journal-of-energy-chemistry/

Machine learning aided design of perovskite oxide materials for photocatalytic water splitting

Qiuling Tao [a], Tian Lu [b], Ye Sheng [b], Long Li [a], Wencong Lu [a,b,*], Minjie Li [a,*]

[a] Department of Chemistry, College of Sciences, Shanghai University, Shanghai 200444, China
[b] Materials Genome Institute, Shanghai University, Shanghai 200444, China

ARTICLE INFO

Article history:
Received 23 September 2020
Revised 19 January 2021
Accepted 19 January 2021
Available online 11 February 2021

Keywords:
Perovskite
Machine learning
Online web service
Photocatalytic water splitting
Bandgap
Hydrogen production rate

ABSTRACT

Suffering from the inefficient traditional trial-and-error methods and the huge searching space filled by millions of candidates, discovering new perovskite visible photocatalysts with higher hydrogen production rate (R_{H_2}) still remains a challenge in the field of photocatalytic water splitting (PWS). Herein, we established structural-property models targeted to R_{H_2} and the proper bandgap (E_g) via machine learning (ML) technology to accelerate the discovery of efficient perovskite photocatalysts for PWS. The Pearson correlation coefficients (R) of leave-one-out cross validation (LOOCV) were adopted to compare the performances of different algorithms including gradient boosting regression (GBR), support vector regression (SVR), backpropagation artificial neural network (BPANN), and random forest (RF). It was found that the BPANN model showed the highest R values from LOOCV and testing data of 0.9897 and 0.9740 for R_{H_2}, while the GBR model had the best values of 0.9290 and 0.9207 for E_g. Furtherly, 14 potential PWS perovskite candidates were screened out from 30,000 ABO_3-type perovskite structures under the criteria of structural stability, E_g, conduction band energy, valence band energy and R_{H_2}. The average R_{H_2} of these 14 perovskites is 6.4% higher than the highest value in the training data set. Moreover, the online web servers were developed to share our prediction models, which could be accessible in http://materials-data-mining.com/ocpmdm/material_api/ahfga3d9puqlknig (E_g prediction) and http://materials-data-mining.com/ocpmdm/material_api/i0ucuyn3wsd14940 (R_{H_2} prediction).

1. Introduction

Photocatalytic water-splitting (PWS) is one of the most promising exploitations of utilizing solar energy to generate hydrogen and oxygen [1]. The semiconductor as photocatalyst in PWS plays a key role, which should meet the following demands: (1) Efficient optical absorptions from sunlight to produce electron-hole pairs; (2) energy level matching between the photocatalyst band level and redox potential of water splitting reaction; (3) effective separation of photoexcited electrons and holes [2–6].

Since Fujishima *et al.* [7] firstly realized the water splitting on TiO_2 under ultraviolet light, various semiconductor materials have been reported as photocatalysts. Due to the superior properties of low price, adaptability, compositional diversity and thermal stability, ABO_3-type perovskite oxides are considered as one of the most promising photocatalytic materials [8,9]. However, most of the existing stable perovskite photocatalysts are semiconductors with a broad bandgap energy (E_g) over 3 eV, of which the effective sunlight absorption are not guaranteed [10]. That is one of the main obstacles to increase the hydrogen production of perovskite photocatalyst. Various strategies have been developed to expand its light absorption to improve hydrogen production. For example, Li et al. [11] used furfural alcohol-derived polymerization-oxidation (FAPO) process and polymerized complex (PC) method to prepare $NaNbO_3$, respectively. The $NaNbO_3$ synthesized by the FAPO route crystallized a cubic system, while the $NaNbO_3$ prepared through the PC method forms an orthorhombic system. The cubic $NaNbO_3$ shows a narrower E_g (3.29 eV) than the orthorhombic $NaNbO_3$ (3.45 eV), and its hydrogen production rate (R_{H_2}) (127 μmol h^{-1}) is higher than orthorhombic $NaNbO_3$ (72.3 μmol h^{-1}). Zhang et al. [12] synthesized a novel hollow nanostructure $BaZrO_3$ photocatalyst with visible light response for hydrogen evolution by substituting Fe (III) into Zr sites. Compared with pure $BaZrO_3$ with band energy of 4.96 eV, the visible light absorption of Fe-doped perovskite is enhanced to 436.6 nm (E_g = 2.84 eV), and the highest R_{H_2} is 9.45 μmol g^{-1} h^{-1}. These work improved the photocatalytic performance of perovskite in PWS to a certain extent. Although

* Corresponding authors.
E-mail addresses: wclu@shu.edu.cn (W. Lu), minjieli@shu.edu.cn (M. Li).

https://doi.org/10.1016/j.jechem.2021.01.035

many methods can be used to extend the light absorption range of ABO_3-type perovskites to further improve their photocatalytic activity, the currently reported perovskite oxides still exhibit low photocatalytic performance under visible light [1]. Therefore, discovering new perovskite materials, especially visible-light photocatalysts with higher hydrogen production is still the major issue in the study of PWS.

The synthesis and development of functional materials have been relying on traditional trial-and-error experimental, which requires major recourses like time, material, equipment, and manpower. This method severely limits the search for high-performance functional materials [13,14]. To accelerate the materials design, advanced theoretical techniques have emerged such as density functional theory (DFT) and machine learning (ML) [15–18]. Materials properties and model processes could be simulated via DFT at the atomic level to help guide experiments and avoid the extra expenses of synthesis. Howbeit, the calculation costs of DFT method scale exponentially with the sizes of basis set and atom number, often taking hours even days to obtain a few targeted properties for a certain chemical system [19–21]. On the contrary, ML could be used to build mathematic models, relying on the existing materials sample set, to predict the targeted properties efficiently and accurately, which avoids the complex quantum mechanics compared to DFT method [22,23]. Recently, ML technique has successfully applied to design new materials, such as binary alloys [24–26], metal–organic frameworks [27–29], perovskite materials [30–32]. It should be noted that many materials designed by ML techniques have been successfully synthesized and shown excellent performance [15,30,33]. Furthermore, the corresponding structure files of new candidate materials can obtain based on existing inorganic materials via ML [34]. ML technology provides some new ideas in the field of material design on perovskites. Despite that, it rarely reported utilizing ML technology to analyze existing PWS data and design perovskite materials with high photocatalytic properties. Therefore, it is a meaningful work to design new perovskite photocatalysts with higher R_{H_2} for PWS based on ML with limited data.

This work is divided into three main steps, schematically shown in Fig. 1. Firstly, the ML models targeted to E_g and R_{H_2} are built up for ABO_3-type perovskite materials. The atomic parameters and experimental conditions are regarded as the variables for the model constructions. Different algorithms are taken into considerations, including SVR, RF, GBR and BPANN. GBR and BPANN are selected out for the targets of E_g and R_{H_2}, respectively. Secondly, 30,000 virtual samples of ABO_3-type perovskite structures are generated and filtered with the target values predicted by the established models. Thirdly, online web server is developed to share the models for the further researches. We are firmly believed that our work could help accelerate the discovery of new perovskite photocatalysts with higher performance and lower cost.

2. Methods

2.1. Backpropagation artificial neural network

Artificial neural network (ANN) [35–39] is an algorithm built by simulating the information processing method of human brain. The algorithm possesses a strong capability to describe nonlinear relationships, good self-learning adaptability, and high tolerance to outliers. It has a good performance in solving pattern recognition, clustering, prediction, optimization, and other problems. ANN often adopts multi-layer perceptron (MLP) structure. The typical single hidden layer ANN structure is shown in Fig. 2, which is composed of input layer, hidden layer and output layer connected to each other. The input data of each node is operated by activation function, and the operation results are output. A connection between two different nodes corresponds to a weight that passes through the connection signal. The output of the network varies according to the network connection mode, weight, and activation function.

Backpropagation (BP) is an algorithm for supervised learning using multi-layer feedforward networks. It is trained by error back-propagation algorithm and is one of the most widely used artificial neural networks. The basic idea of the back-propagation learning algorithm is to repeatedly apply chain rule to calculate the influence of each weight in the network on the arbitrary error function. The calculation formula of the error function is:

$$\frac{\partial E}{\partial \omega_{ij}^{(k)}} = \frac{\partial E}{\partial s_i^{(k)}} \frac{\partial s_i^{(k)}}{\partial net_i^{(k)}} \frac{\partial net_i^{(k)}}{\partial \omega_{ij}^{(k)}}, \quad (1)$$

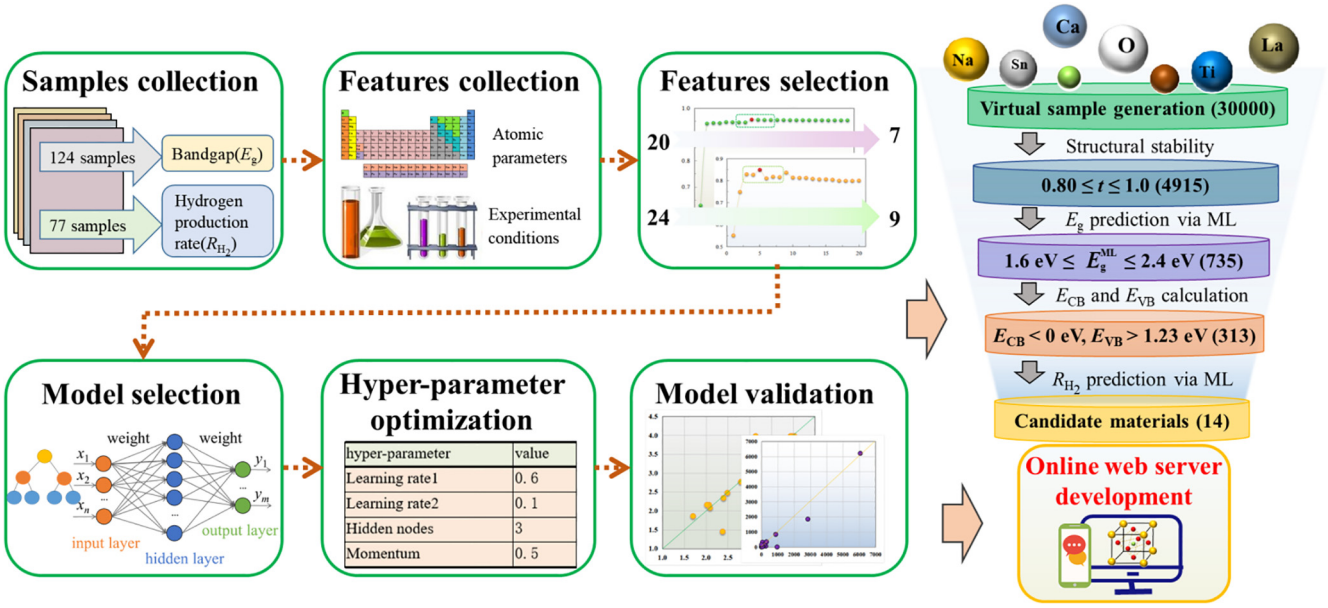

Fig. 1. The flowchart of materials ML in this work.

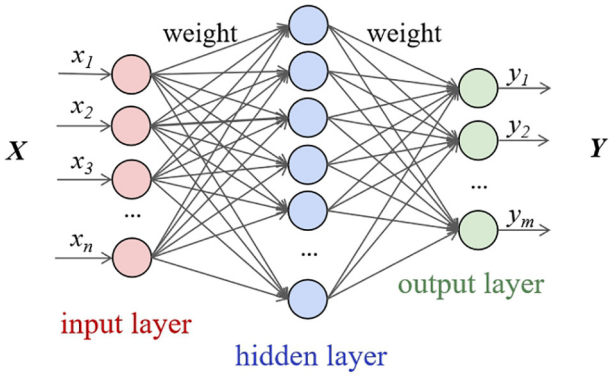

Fig. 2. Structure diagram of ANN.

where $\partial\omega_{ij}$ is the weight from neuron j to neuron i, ∂s_i is the output, ∂net_i is the weighted sum of the input of neuron i, and the k can indicate the kth layer of ANN. Once the partial derivative of each weight is known, the purpose of minimizing the error function is achieved by performing simple gradient descent:

$$\omega_{ij}(t+1) = \omega_{ij}(t) - \varepsilon\frac{\partial E}{\partial \omega_{ij}}, \tag{2}$$

where ε is the learning rate, how to set ε is an important issue.

2.2. Gradient boosting regression

Gradient boosting regression (GBR) [40–42] was originally proposed by Friedman. It is a powerful machine learning technology with good practical application effect. The basic idea is to use the value of the negative gradient of the loss function under the current model as the approximation of the residual of the model's training result this time, and use this value as the target of the next training. The output of the model will move in the direction of loss function reduction. The basic principle of gradient boosting method with regression tree as weak learner can be expressed as follows:

For the sample space $N = \{(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)\}$, the goal is to find a prediction function $F(x)$ that minimizes the loss function $L(y,F(x))$ under all x to y mappings. The prediction function is expressed as:

$$F(x) = \sum_{m=1}^{M}\beta_m h(x; a_m), \tag{3}$$

where $h(x; a_m)$ is the m_{th} subtree of the weak learner, $m = 1, 2, \ldots, m$; a_m is the parameter of the mth subtree; β_m is the weight of the subtree.

If the prediction function generated by the first m training weak learner is $F_m(x)$, the optimization problem is equivalent to finding the parameters of the new subtree (β_m, a_m),

$$(\beta_m, a_m) = argmin\sum_{i=1}^{N} L(y_i, F_{m-1}(x_i) + \beta_m h(x; \alpha_m)). \tag{4}$$

For the above conditions, the update process of the entire gradient boosting model is:

Step 1: initialize the first regression tree:

$$F_0 = argmin\sum_{i=1}^{N} L(y_i, h_0(x; \alpha_m)). \tag{5}$$

Step 2: For $m = 1,2,3, \ldots M$, the negative gradient of the loss function is:

$$y_{im} = -\left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\right]_{F(x)=F_{m-1}(x)}. \tag{6}$$

Fit a new subtree with $\tilde{y}_{im}$ as the training target, and calculate the relevant parameters of the subtree, then get the area of the leaf node:

$$a_m = argmin\sum_{i=1}^{N}\left[y_{im} - \beta_m h(x_i; \alpha_m)\right]^2, \tag{7}$$

$$\beta_m = argmin\sum_{i=1}^{N} L(y_i, F_{m-1}(x_i) + \beta_m h(x; \alpha_m)). \tag{8}$$

Update the prediction function:

$$F_m(x) = F_{m-1}(x) + v\beta_m h(x; a_m), \tag{9}$$

where v is the step size to control the learning rate. The smaller the v is, the more training times are required to achieve the required prediction accuracy; but if the v is set too large, the higher prediction accuracy may not be achieved.

2.3. Support vector regression

Support vector regression (SVR) [43–45] is an effective method for solving problems in nonlinear regression. It is also a supervised learning algorithm that has been widely applied to various fields. It considers the balance between empirical risk and expected risk, and then makes computational model have the good prediction and generalization performances. The biggest advantage of SVR is that it only needs a small number of samples to achieve good results.

Set the sample set as: $(y_1, x_1), \ldots, (y_l, x_l)$, the nonlinear regression function is expressed by the following regression function:

$$f(x) = \sum(\alpha_i^* - \alpha_i)K(x_i, x_j). \tag{10}$$

The Lagrange multipliers α_i^*, α_i and the kernel function $K(x_i, x_j)$ can be calculated according to the following equation:

$$\max_{\alpha,\alpha^*} W(\alpha, \alpha^*) = \max_{\alpha,\alpha^*}\left\{\begin{array}{c} -0.5\times\sum_{i=1}^{l}\sum_{j=1}^{l}(\alpha_i - \alpha_i^*)\left(\alpha_j - \alpha_j^*\right)K(x_i, x_j) \\ +\sum_{i=1}^{l}[\alpha_i(y_i - \varepsilon) - \alpha_i^*(y_i + \varepsilon)] \end{array}\right\}.. \tag{11}$$

The constraint conditions of the Lagrange are:

$$0 \le \alpha_i \le C,\ i = 1, \cdots, l,$$

$$0 \le \alpha_i^* \le C,\ i = 1, \cdots, l,$$

$$\sum_{i=1}^{l}(a_i^* - \alpha_i) = 0, \tag{12}$$

here ε is the deviation value; parameter C is a regularized constant determining the trade-off between the training error and the model flatness.

$K(x_i, x_j)$ can be replaced with an appropriate kernel function. In the study, the SVR model applies Gaussian kernel function (RBF) shown as following:

$$K(x_i, x_j) = e^{-\frac{\|x_i - x_j\|^2}{2\sigma^2}}, \tag{13}$$

where σ is the function parameter.

2.4. Random forest

Random forest (RF) [46–49] is an ensemble learning algorithm based on classification tree proposed by Breinman, which is an important method in ML. The basic principle of RF algorithm is similar to classification and regression tree (CART) algorithm. The RF algorithm possesses the advantages of improving prediction accuracy, reducing overfitting, and being insensitive to missing value and multicollinearity.

In RF regression algorithm, the final prediction results are based on the predicted values of all subtrees. The average of the predicted values of all subtrees is used as the final prediction. The construction steps of the random forest regression model can be summarized as follows:

(1) With the bootstrap sampling method, N training sample subsets is drawn from the training data set to form the training set $S_i\{i = 1,2,\ldots,N\}$
(2) For each training set above, the corresponding subtree $CART_1$,$CART_2$,...,$CART_N$ is generated. At each node, the best split is selected among a randomly selected subset of $M_{feature}$ (rather than all) features. The tree is grow to the maximum size and not pruned back.
(3) Using the test data set to test the performance of the previous models, the predicted values of $CART_1$(Test),$CART_2$(Test),...,$CART_N$(Test) from the subtree output is obtained.
(4) The predicted values of the outputs of N decision trees were counted. Then the average predicted values of the outputs of all subtrees were inversely normalized to the final predicted values.

3. Results and discussion

3.1. The data analysis

In this work, we collected 160 ABO_3-type perovskite photocatalytic data from 43 experimental literatures. The data set can be seen in Table S1 of the Supporting Information. The E_g dataset and the R_{H_2} dataset contain 124 and 77 samples, respectively.

In the PWS, the E_g is a crucial parameter to determine the light absorption capacity of a photocatalyst. Thus, the first ML model is constructed to predict the E_g of perovskite materials. And our goal is to design new perovskites with high R_{H_2} and satisfactory E_g values. Thereupon, the second ML model is to predict the R_{H_2} for PWS. Since the majority of the solar spectrum consists of visible light (approximately 50%), the primary goal of PWS is to efficiently and fully utilize the visible light [50–53]. In the two models, the variables of the ML models are the atomic parameters and experimental conditions, while the targets are the corresponding E_g and R_{H_2} of the perovskite materials. We randomly divided the data sets into two subsets at a ratio of 4:1, namely the training set and the testing set, which were respectively used to train the models and test the quality of the built models.

3.2. Feature engineering

For any ML method with pre-specified targets, it usually depends on a certain number of features. There may be many factors that affect the targeted property of materials, but the best strategy is to choose features that can excellently represent the properties of materials. The number of features must be reasonable and should be less than the number of samples in the input dataset [24,54]. Furthermore, data reliability and comparability is extremely crucial for a reasonable machine learning model. To reduce the impact of data differences from different kinds of literature on the ML model, the corresponding experimental conditions of these samples were used as the original feature variables. Other original feature variables are generated based on atomic parameters, which are shown in Table 1. For the E_g model, we collected 20 incipient features, including 17 atomic parameters generated from online computation platform for materials data mining (OCPMDM) [55] and 3 experimental conditions extracted from references (see the Supporting Information). And for the R_{H_2} model, there are 24 incipient features, including 18 atomic parameters and 6 experimental conditions.

Before developing the models, it is necessary to delete the redundant or irrelevant features. The existence of these features will significantly affect the performances of the models, leading to the risk of overfitting. Eliminating these features can reduce the dimension of feature space without losing critical information, and further promote the prediction ability and generalization performance of the models [56–58]. In this work, Max Relevance Min Redundancy (mRMR) [59] approach was employed to screen the optimal feature subset for SVR and BPANN models. As for GBR and RF models, the embedded method [60,61] was used for feature screening. The method of mRMR uses information relevance or similarity score to select feature variables and minimizes redundancy among selected variables in the variable selection process. The method of embedded feature selection is to integrate the feature selection process with the learner training process and automatically select the features during the learner training process. Based on the features score ranking given by mRMR, the feature variables were deleted one by one from back to front. Each time a feature was deleted, a model was built with the remaining feature variables, thereby obtaining a series of models. The Pearson correlation coefficients (R) of the leave-one-out cross-validation (LOOCV) for each model are shown in Fig. 3. The best feature sets were chosen based on their largest R values. The optimal feature sets for each algorithm are listed in Table 2.

3.3. Model selection

To build the two models for E_g and R_{H_2}, we use ML tools from ExpMiner (Data mining software package) and OCPMDM developed in our laboratory. The free version of the ExpMiner can be downloaded on the website of Laboratory of Materials Data Mining in Shanghai University: http://chemdata.shu.edu.cn:8080/MyLab/Lab/download.jsp. The OCPMDM can be accessible at the web address: http://materials-data-mining.com/ocpmdm/

In ML technology, it is important to choose the appropriate ML algorithm. In this work, we selected the best algorithm from the four widely used algorithms of BPANN, SVR, GBR and RF to build the models. In order to select the optimal regression models, LOOCV was performed to evaluate the performance of each ML model. Pearson correlation coefficient (R) and root mean square error (*RMSE*) were used as evaluation indexes. *RMSE* and R were used to evaluate the error and correlation between predicted and experimental values, respectively. In general, a small *RMSE* and a large R value indicate good prediction results. As shown in Table 3, the BPANN and the GBR models have the best prediction ability for R_{H_2} and E_g, respectively. The *RMSE* values are quite large in the R_{H_2} models, which can be explained by many factors. On the one hand, the different experimental conditions lead to large experimental errors. For instance, $LaFeO_3$ reported in the same literature has different R_{H_2} values due to different calcination temperatures in the preparation process. The R_{H_2} values of $LaFeO_3$ were 5466.7 and 8600 μmol g^{-1} h^{-1}, respectively [62]. Accordingly, the mean relative error (*MRE*) of experimental results is between 36.43% and 57.32%, and *RMSE* is 3133.3. On the other hand, the range of data

Table 1
The features include atomic parameters and experimental conditions. Nos. 1–24 are used for the R_{H_2} model and Nos. 6–25 are used for the E_g models.

No.	Meanings	Features	No.	Meanings	Features
1	Light intensity	*LI*	14	Electronegativity (Pauling) of B-site	χ_{pb}
2	Photocatalyst dose (g/L)	*PD*	15	Ionization energy of A-site (eV)	I_{1a}
3	Reaction solution	*RS*	16	Ionization energy of B-site (eV)	I_{1b}
4	Co-catalyst	*CL*	17	Electron affinity of A-site (kJ/mol)	EA_a
5	The enthalpy of fusion of A-site (kJ/mol)	$\Delta_{fus}H_a$	18	Electron affinity of B-site (kJ/mol)	EA_b
6	Preparation method	*PM*	19	Melting point of A-site (°C)	*A_Tm*
7	Calcination temperature (K)	*CT*	20	Melting point of B-site (°C)	*B_Tm*
8	Ionic radius of A-site (pm)	R_a	21	Boiling point of A-site (°C)	*A_Tb*
9	Ionic radius of B-site (pm)	R_b	22	Boiling point of B-site (°C)	*B_Tb*
10	Ratio of ionic radius of A-site to B-site	R_a/R_b	23	Density of A-site (g/cm^3)	ρ_a
11	Tolerance factor	*t*	24	Density of B-site (g/cm^3)	ρ_b
12	Molecular mass (g/mol)	*M*	25	Calcination time (h)	*AH*
13	Electronegativity (Pauling) of A-site	χ_{pa}			

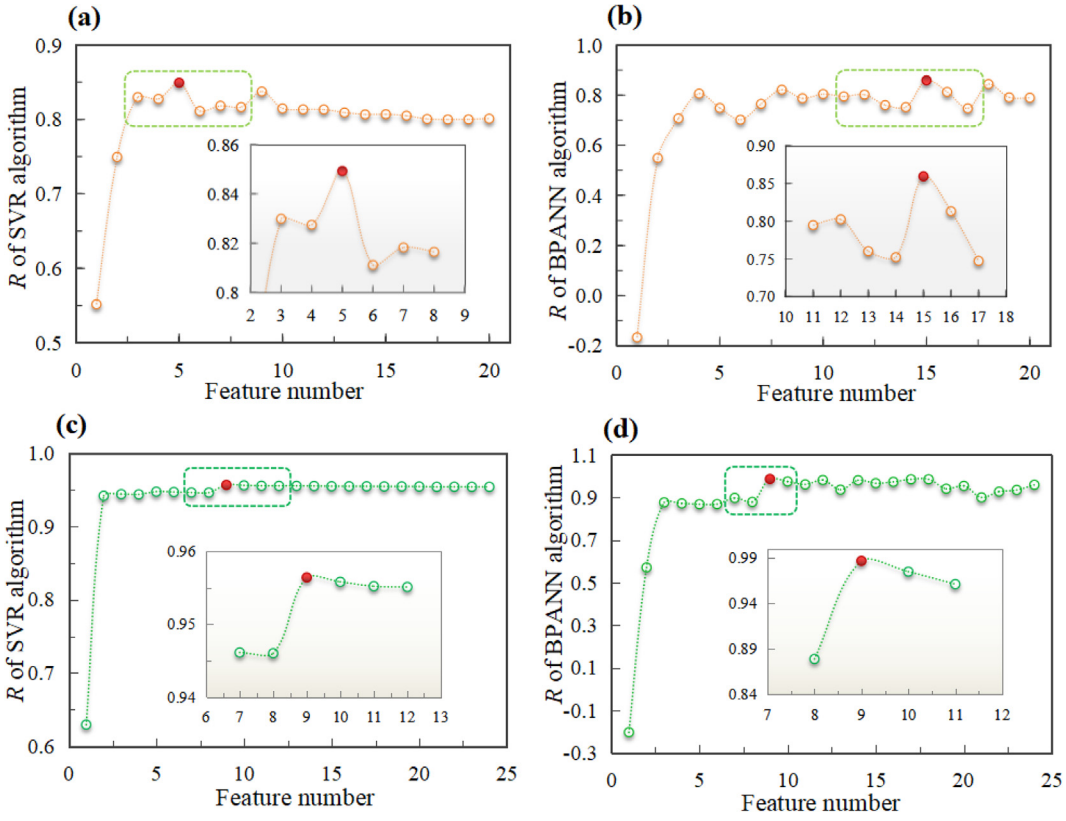

Fig. 3. *R* of the LOOCV of the ML model in each selection process. The position of the red point is the maximum value of *R*. (a) *R* of SVR in the E_g model. (b) *R* of BPANN in the E_g model. (c) *R* of SVR in the R_{H_2} model. (d) *R* of BPANN in the R_{H_2} model.

Table 2
The results of feature selection.

Target	Algorithm	The selected features
R_{H_2}	SVR BPANN	*PM*, *PD*, *A_Tb*, EA_b, *LI*, $\Delta_{fus}H_a$, χ_{pa}, *t*, *CT*
	GBR	*PM*, *CT*, *LI*, *A_Tm*, $\Delta_{fus}H_a$, ρ_a
	RF	*PM*, *CT*, *LI*, *A_Tm*, *A_Tb*, $\Delta_{fus}H_a$, ρ_a
E_g	SVR	*B_Tb*, I_{1b}, EA_b, ρ_b, I_{1a}
	BPANN	*B_Tb*, I_{1b}, EA_b, ρ_b, I_{1a}, *CT*, *A_Tb*, *B_Tm*, *AH*, *PM*, ρ_a, R_b, *A_Tm*, *M*, χ_{pa}
	GBR	R_b, χ_{pb}, I_{1b}, EA_b, *A_Tb*, ρ_a, ρ_b
	RF	R_b, EA_b, *A_Tb*, ρ_a, ρ_b

values is enormous in the dataset (from 1.25 to 8600 μmol g^{-1} h^{-1}), which results in a larger *RMSE*.

To evaluate the stability and generalization of the best models, the data set was divided randomly into the training set and the testing set at a ratio of 4:1 for 100 times and rebuilt the models based on the selected features, respectively. The average *R* values of LOOCV in 100 training models were 0.9010 (E_g) and 0.9805 (R_{H_2}), respectively, and the *R* values of the testing set were 0.9125 (E_g) and 0.9543 (R_{H_2}), respectively (Table S4). The results guarantee the robustness and generalization of the models.

3.4. Hyper-parameter optimization

Most ML algorithms require appropriate parameter adjustments to improve the efficiency and generalization performance

Table 3
The results of different ML algorithms for the prediction of the E_g of perovskite and the prediction of the R_{H_2} based on perovskite.

ML algorithms	Bandgap prediction		Hydrogen production rate prediction	
	R value	*RMSE*	*R* value	*RMSE*
SVR	0.8493	0.3964	0.9564	926.8851
RF	0.8840	0.3490	0.8915	846.2868
GBR	0.9217	0.2879	0.9158	770.4925
BPANN	0.8594	0.4181	0.9869	290.5684

of the model. After selecting the ML algorithm, BPANN and GBR models were best for E_g and R_{H_2}, respectively. Therefore, we optimize the hyper-parameters of the BPANN and GBR algorithms. Grid search method was employed to optimize hyper-parameters BPANN and GBR algorithms to improve the performance of the models. Here, six hyper-parameters (loss function, estimators, learning rate, max depth, min samples split and min samples leaf) in the GBR algorithm are optimized. For the BPANN algorithm, hyper-parameters optimized, there are hidden layer nodes, momentum, learning rate1 (input layer to hidden layer), learning rate2 (hidden layer to output layer). The results of hyper-parameter optimization are shown in Table 4. After hyperparameter optimization, the performance of the models improved (Fig. 4).

3.5. Model validation

In order to further test the prediction ability of the models obtained, we used the remaining 20% of the data as a testing set to validate our models. The results were shown in Fig. 5. Both *R* and *RMSE* values of the testing results agree to those of training results, indicating that our models have good predictive ability.

Table 4
The optimal hyper-parameters of BPANN and GBR.

GBR hyper-parameter	Value	BPANN hyper-parameter	Value
Estimators	200	Learning rate1	0.6
Learning rate	0.1	Learning rate2	0.1
Max depth	3	Hidden nodes	3
Min samples leaf	1	Momentum	0.5
Min samples split	3		

3.6. Virtual screening

In a nutshell, the GBR and BPANN models exhibit excellent performance in predicting E_g and R_{H_2} values, respectively. Thus, these models can be put to use for virtual screening the candidate perovskite materials.

We extracted 10 A-site cations and 24B-site cations from the dataset and other literature (Fig. 6). The new perovskite materials are designed in the formula of $A1_x(A2_mA3_{1-x-m})B1_y(B1_nB2_{1-y-n})O_3$, where x, y are in the range of 0.6–1.0 with the interval of 0.01, $m \leq 0.4$, $n \leq 0.4$. Thence, hundreds of millions of potential combinations could be generated, where the searching space is so enormous that only a few of them have been reported. We generated a total of 30,000 candidates of perovskite oxides based on the designed formula. The tolerance factor (t), defined as $t = (r_A + r_O)/[\sqrt{2}(r_B + r_O)]$, where r_A, r_B, and r_O are ionic radius of ions A, B, and O, respectively, used for clarify the perovskite formability of candidate materials. In general, perovskite can be formed in the range of $0.8 < t < 1.0$ [63,64]. After considering the t, 4915 compounds were retained for the further virtual screening.

One requirement for PWS is that the conduction band energy (E_{CB}) and valence band energy (E_{VB}) for a semiconductor material should straddle the redox potentials for H^+/H_2 and H_2O/O_2. The conduction band bottom-edge must be more negative than the reduction potential of H^+ to H_2 (0 V vs. NHE at pH = 0), while the valence band top-edge should be more positive than the oxidation potential of H_2O to O_2 (1.23 V vs. NHE at pH = 0) [65–69]. Thus, the E_g should be greater than 1.23 eV to meet the thermodynamic requirements for the decomposition of H_2O into H_2 and O_2. In practice, due to the energy losses at the solid–liquid junctions, E_g should be restrained in 1.6–2.4 eV [70]. Consequently, when choosing semiconductors for PWS, the width of the bandgap and the band edge potential are regarded as the parameters that should be considered first.

For that reason, we employ the GBR model to predict the E_g of the 4915 perovskite oxides. Then eliminate the compounds with E_g values out of the range (1.6–2.4 eV). To determine the band edge positions of the valence-band maxima and the conduction-band minima, we used the empirical equation proposed by Xu and Schoonen [71]. In this equation, the E_{CB} and E_{VB} are derived from the geometric averages of the absolute electronegativities of neutral atoms and E_g. The formulae [72] are as follows:

$$E_{CB} = \left(\chi^a_{(A)} \cdot \chi^b_{(B)} \cdot \chi^o_{(O)}\right)^{\frac{1}{(a+b+o)}} - \frac{1}{2}E_g^{ML} + E_0 \quad (14)$$

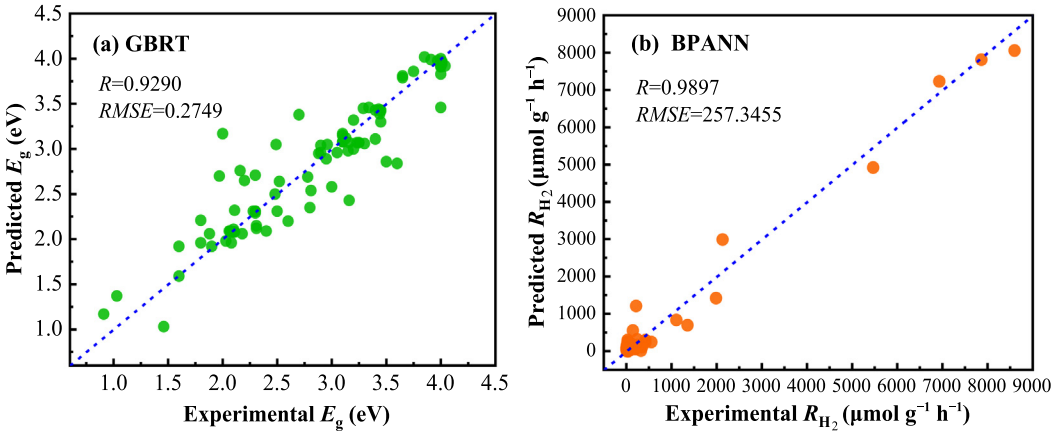

Fig. 4. Performance of training models after hyperparameter optimization. (a) GBR model of E_g. (b) BPANN model of R_{H_2}.

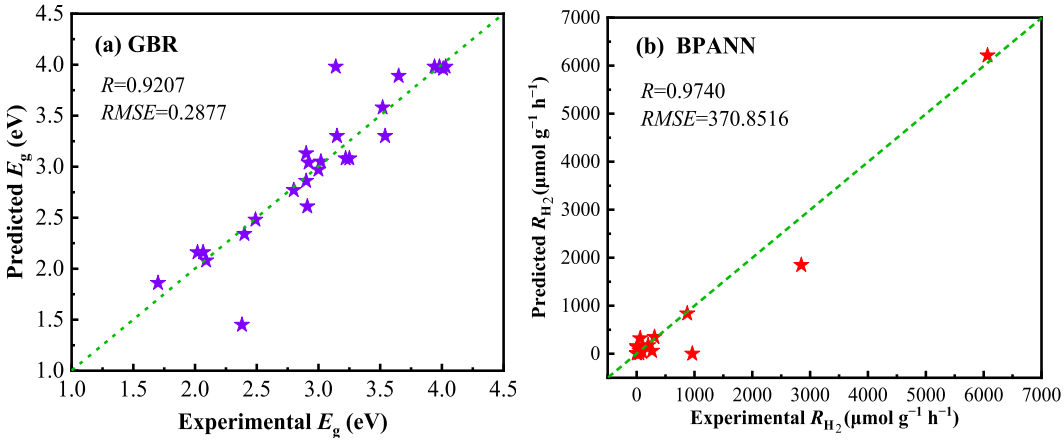

Fig. 5. Performances of the model by the GBR and BPANN algorithms. (a) Predicted E_g vs. experimental E_g. (b) Predicted R_{H_2} versus experimental R_{H_2}.

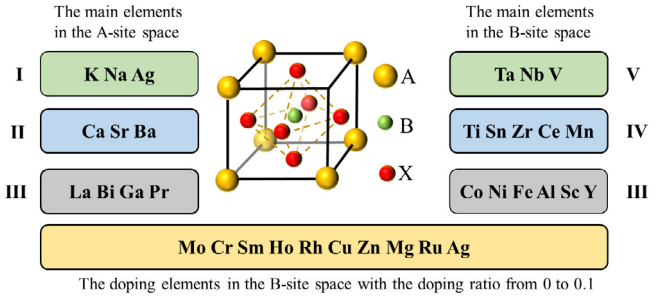

Fig. 6. ABO_3-type perovskite material composition for virtual screening, in which ABO_3-type perovskite is represented as a cage-like structure. The Roman numerals indicate the valence of elements.

$$E_{VB} = E_{CB} + E_g^{ML} \quad (15)$$

Here, $\chi_{(A)}$, $\chi_{(B)}$, and $\chi_{(O)}$ are the absolute electronegativities of constituent atoms A, B, and O, respectively; E_g^{ML} is the bandgap energy via the ML model prediction, and E_0 is a scale factor relating the reference electrode redox level to the vacuum level (E_0 = −4.5 eV) [73]. The absolute electronegativities were estimated from the arithmetic mean of the first ionization potential and electron affinity [74]. After excluding perovskites with E_{CB} greater than 0 eV and E_{VB} less than 1.23 eV, 300 compounds were remained.

Next, the BPANN model was used to predict the R_{H_2} of these 313 ABO_3-type perovskites, and perovskite molecules with R_{H_2} more than 8900 μmol g^{-1} h^{-1} were screened out. Finally, 14 potential candidate perovskite photocatalysts for visible-light water splitting were obtained (Table 5). The average R_{H_2} of these 14 candidate perovskites was 9154.18 μmol g^{-1} h^{-1}, 6.4% higher than the highest value of 8600 μmol g^{-1} h^{-1} in the dataset. 14 candidate perovskite photocatalysts were obtained by doping ions into $BaSnO_3$ or $SrSnO_3$. Both $BaSnO_3$ and $SrSnO_3$ have E_{CB} and E_{VB} edge positions that satisfy the H_2O redox conditions. The E_{CB} and E_{VB} of $BaSnO_3$ are −0.59 and 2.51 eV, while those of $SrSnO_3$ are −0.89 and 2.96 eV [75]. Although $BaSnO_3$ and $SrSnO_3$ are semiconductors with broad E_g, various studies have proved that ion doping can change the E_{CB} and E_{VB} edge positions of perovskites reducing the E_g [76–79]. Furthermore, existing reports have demonstrated that $BaSnO_3$ and $SrSnO_3$ possess photocatalytic activity in PWS [80–83]. That indicates the reliability of the 14 candidate perovskite photocatalysts we proposed. In addition, the E_g of $BaSn_{0.375}Ti_{0.625}O_3$ with Pm__3m space group was calculated by DFT. The calculation details are shown in calculation details of DFT of the Supporting Information. The E_g values of HSE06 and PBE + U are 3.92 and 1.34 eV, respectively. The data are different from the predicted result (2.26 eV) via the ML model, which could be caused by the space group. When there are enough samples with the specific space group in the future, we will explore the influence of space groups on E_g.

3.7. Online prediction

The online web servers are developed to predict the E_g and R_{H_2} of ABO_3-type perovskite based on the established models. It is convenient for the experimental scientists to use the models to design

Table 5
The potential candidate ABO_3-type perovskites with higher R_{H_2} in visible-light water splitting reaction screened out by using the model available.

No.	Molecular formula	R_{H_2} (μmol g^{-1} h^{-1})	Tolerance factor	E_g^{ML} (eV)	E_{CB}^{Cal} (eV)	E_{VB}^{Cal} (eV)
1	$BaSn_{0.375}Ti_{0.625}O_3$	9195.28	0.96	2.26	−0.42	1.84
2	$BaSn_{0.68}Ti_{0.32}O_3$	9230.30	0.95	2.30	−0.37	1.93
3	$Ba_{0.9}Ca_{0.1}Sn_{0.6}Ti_{0.4}O_3$	9205.45	0.94	2.26	−0.37	1.89
4	$Ba_{0.6}Ca_{0.4}Sn_{0.6}Ti_{0.4}O_3$	9175.73	0.90	2.32	−0.43	1.89
5	$Ba_{0.66}Ca_{0.34}Sn_{0.68}Ti_{0.32}O_3$	9203.42	0.90	2.20	−0.35	1.85
6	$Ba_{0.9}Ca_{0.1}Sn_{0.7}Ti_{0.28}Cr_{0.02}O_3$	9227.77	0.93	2.36	−0.40	1.96
7	$Ba_{0.88}Ca_{0.12}Sn_{0.68}Ti_{0.3}Cr_{0.02}O_3$	9223.86	0.93	2.31	−0.38	1.93
8	$Ba_{0.9}Ca_{0.1}Sn_{0.62}Ti_{0.37}Sm_{0.01}O_3$	9208.49	0.93	2.26	−0.37	1.89
9	$Ba_{0.6}Ca_{0.24}Sr_{0.16}Sn_{0.6}Ti_{0.4}O_3$	9146.37	0.91	2.29	−0.43	1.86
10	$Ba_{0.84}Sr_{0.16}Sn_{0.68}Ti_{0.32}O_3$	9195.25	0.94	2.31	−0.40	1.91
11	$Ba_{0.92}Sr_{0.08}Sn_{0.66}Ti_{0.32}Ru_{0.02}O_3$	9213.25	0.94	2.35	−0.40	1.95
12	$Sr_{0.7}Ba_{0.3}Sn_{0.61}Mn_{0.39}O_3$	8955.69	0.91	2.37	−0.52	1.85
13	$Sr_{0.82}Ba_{0.18}Sn_{0.69}Mn_{0.31}O_3$	8981.64	0.90	2.33	−0.51	1.82
14	$Sr_{0.63}Ca_{0.02}Ba_{0.35}Sn_{0.6}Mn_{0.4}O_3$	8977.15	0.91	2.40	−0.52	1.88

Fig. 7. An example of online prediction for the E_g of ABO_3-type perovskite.

the new ABO_3-type perovskites for visible-light PWS. In the process of applying the models, the user needs to input the experimental conditions, while the atomic parameters can be generated automatically via the OCPMDM. Fig. 7 illustrates an example of online prediction for the E_g of ABO_3-type perovskite. The predicted E_g values of ABO_3-type perovskite could be easily obtained by inputting the designed chemical formula. These online web servers might accelerate the experimental researches of the promising ABO_3-type perovskite with proper E_g in photocatalysis field. The online Web servers can be accessible at the web address: http://materials-data-mining.com/ocpmdm/material_api/ahfga3d9puqlknig (E_g prediction); http://materials-data-mining.com/ocpmdm/material_api/i0ucuyn3wsd14940 (R_{H_2} prediction).

4. Bandgap classification model

To determine the E_g of the candidates is direct or indirect, a classification model is established. The classification model contains 70 data, of which 80% is used for training and 20% for testing. There are a total of 20 feature variables, which are consistent with the E_g regression model. Given the superiority of the GBR algorithm on the E_g model, the gradient boosting classification (GBC) algorithm is used to construct the E_g classification model. After feature screening by embedding method combined with GBC algorithm, six features were retained, including the preparation method, the radius of A and B ions, the electronegativity of A, tolerance factor, and the electron affinity of A. The accuracy of LOOCV and testing set are 94.64% and 92.86%, respectively (Table S5). Furthermore, to evaluate the stability of the model, the data set was randomly divided into the training set and testing set with the ratio 4:1 for 100 times and rebuilt the model with the same features that are selected in the constructed model. The average accuracy of LOOCV and testing set for 100 models was 92.32% and 91.07%, respectively (Table S6). The average accuracy values are slightly lower than the constructed models, considering the existence of extreme splitting situations, *e.g.,* all the noisy samples are split into testing data. The results guarantee the reliability and robustness of the model. This model can predict whether E_g is direct or indirect.

5. Conclusions

This work is focused on the predictions of E_g and R_{H_2} of ABO_3-type perovskite materials in the PWS. For E_g, the LOOCV results show that GBR has stronger prediction ability and higher prediction accuracy than other models. The R of LOOCV and testing set validation are 0.9290 and 0.9207, respectively. In the R_{H_2} model, BPANN model has better prediction performance, the corresponding R values of LOOCV and testing set validation are 0.9897 and 0.9740, respectively. Through virtual screening, obtain 14 potential perovskite photocatalysts with higher R_{H_2}. The average R_{H_2} of these 14 perovskites is 6.4% higher than the highest value in the training data set. We also developed the user-friendly and publicly accessible web servers to predict the E_g and R_{H_2} of ABO_3-type perovskite materials, which could provide benefit guidance for accelerating material design and optimization. The method demonstrated in the research can also be extended to materials design and controllable synthesis of other compounds, and further improve the research in the field of the theoretical materials design using the ML technique.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

Financial support to this work from the National Key Research and Development Program of China (No. 2016YFB0700504) and the Science and Technology Commission of Shanghai Municipality (18520723500) is gratefully acknowledged.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jechem.2021.01.035.

References

[1] K. Sawada, T. Nakajima, APL Mater. 6 (2018) 101103.
[2] S. Fang, Y.H. Hu, Int. J. Energy Res. 43 (2019) 1082–1098.
[3] H. Huang, B. Pradhan, J. Hofkens, M.B.J. Roeffaers, J.A. Steele, ACS Energy Lett. 5 (2020) 1107–1123.
[4] S. Kim, N. Nguyen, C. Bark, Appl. Sci. 8 (2018) 1526.
[5] B.L. Phoon, C.W. Lai, J.C. Juan, P.L. Show, G.T. Pan, Int. J. Hydrogen Energy 44 (2019) 14316–14340.
[6] W. Wang, M. Xu, X. Xu, W. Zhou, Z. Shao, Angew. Chem. Int. Ed. 59 (2020) 136–152.
[7] A. Fujishima, K. Honda, Nature 238 (1972) 37–38.
[8] Y. Huang, J. Liu, Y. Deng, Y. Qian, X. Jia, M. Ma, C. Yang, K. Liu, Z. Wang, S. Qu, Z. Wang, J. Semicond. 41 (2020) 011701.
[9] I.M. Nassar, S. Wu, L. Li, X. Li, ChemistrySelect 3 (2018) 968–972.
[10] L. Lu, S. Ni, G. Liu, X. Xu, Int. J. Hydrogen Energy 42 (2017) 23539–23547.
[11] P. Li, S.X. Ouyang, G.C. Xi, T. Kako, J.H. Ye, J. Phys. Chem. C 116 (2012) 7621–7628.
[12] J. Meng, X. Fu, K. Du, X. Chen, Q. Lin, X. Wei, J. Li, Z. Zhang, Int. J. Hydrogen Energy 43 (2018) 9224–9232.
[13] D. Jain, S. Chaube, P. Khullar, S. Goverapet Srinivasan, B.R. Rai, Phys. Chem. Chem. Phys. 21 (2019) 19423–19436.
[14] J. Li, B. Pradhan, S. Gaur, J. Thomas, Adv. Energy Mater. 9 (2019) 1901891.
[15] S. Chakraborty, W. Xie, N. Mathews, M. Sherburne, R. Ahuja, M. Asta, S.G. Mhaisalkar, ACS Energy Lett. 2 (2017) 837–845.
[16] D. Liu, Q. Li, J. Hu, H. Jing, K. Wu, J. Mater. Chem. C 7 (2019) 371–379.
[17] N. Mounet, M. Gibertini, P. Schwaller, D. Campi, A. Merkys, A. Marrazzo, T. Sohier, I.E. Castelli, A. Cepellotti, G. Pizzi, N. Marzari, Nat. Nanotechnol. 13 (2018) 246–252.
[18] J. Varignon, M. Bibes, A. Zunger, Nat. Commun. 10 (2019) 1658.
[19] P.V. Balachandran, Comput. Mater. Sci. 164 (2019) 82–90.
[20] V. Gladkikh, D.Y. Kim, A. Hajibabaei, A. Jana, C.W. Myung, K.S. Kim, J. Phys. Chem. C 124 (2020) 8905–8918.

[21] C. Li, H. Hao, B. Xu, G. Zhao, L. Chen, S. Zhang, H. Liu, J. Mater. Chem. C 8 (2020) 3127–3136.
[22] S. Lu, Q. Zhou, L. Ma, Y. Guo, J. Wang, Small Methods 3 (2019) 1900360.
[23] S. Lu, Q. Zhou, Y. Ouyang, Y. Guo, Q. Li, J. Wang, Nat. Commun. 9 (2018) 3405.
[24] X. Cao, Y. Zhang, J. Li, H. Chen, Mater. Res. Express 7 (2020) 046506.
[25] K. Sasikumar, H. Chan, B. Narayanan, S.K.R.S. Sankaranarayanan, Chem. Mater. 31 (2019) 3089–3102.
[26] H. Zhang, Z. Guo, H. Hu, G. Zhou, Q. Liu, Y. Xu, Q. Qian, D. Dai, Modell. Simul. Mater. Sci. Eng. 28 (2020) 035002.
[27] P.Z. Moghadam, S.M.J. Rogge, A. Li, C.M. Chow, J. Wieme, N. Moharrami, M. Aragones Anglada, G. Conduit, D.A. Gomez Gualdron, V.V. Speybroeck, D. Fairen Jimenez, Matter 1 (2019) 219–234.
[28] Z. Shi, W. Yang, X. Deng, C. Cai, Y. Yan, H. Liang, Z. Liu, Z. Qiao, Mol. Syst. Des. Eng. 5 (2020) 725–742.
[29] Y. Wu, H. Duan, H. Xi, Chem. Mater. 32 (2020) 2986–2997.
[30] B.C. Weng, Z.L. Song, R.L. Zhu, Q.Y. Yan, Q.D. Sun, C.G. Grice, Y.F. Yan, W.J. Yin, Nat. Commun. 11 (2020) 3513.
[31] Z. Li, Q. Xu, Q. Sun, Z. Hou, W.J. Yin, Adv. Funct. Mater. 29 (2019) 1807280.
[32] W.A. Saidi, W. Shadid, I.E. Castelli, npj Comput, Mater. 6 (2020) 36.
[33] K. Kaufmann, D. Maryanovsky, W.M. Mellor, C. Zhu, A.S. Rosengarten, T.J. Harrington, C. Oses, C. Toher, S. Curtarolo, K.S. Vecchio, npj Comput Mater. 6 (2020) 42.
[34] G. Hautier, C. Fischer, V. Ehrlacher, A. Jain, G. Ceder, Inorg. Chem. 50 (2011) 656–663.
[35] M.A.N. Dewapriya, R.K.N.D. Rajapakse, W.P.S. Dias, Carbon 163 (2020) 425–440.
[36] W. Gao, C. Su, J. Comput. Appl. Math. 380 (2020) 112991.
[37] H.J. Hwang, J.W. Baek, J.Y. Kim, C.S. Kim, Eng. Struct. 198 (2019) 109535.
[38] E. Matel, F. Vahdatikhaki, S. Hosseinyalamdary, T. Evers, H. Voordijk, Int. J. Constr. Manage. (2019) 1–14.
[39] Y. Wang, C. Ling, H. Yin, W. Liu, Z. Tang, Z. Li, Sol. Energy 204 (2020) 667–672.
[40] J.H. Friedman, Ann. Stat. 29 (2001) 1189–1232.
[41] X. He, J. Luo, P. Li, G. Zuo, J. Xie, Water Resour. Manage. 34 (2020) 865–884.
[42] S. Liao, Z. Liu, B. Liu, C. Cheng, X. Jin, Z. Zhao, Hydrol. Earth Syst. Sci. 24 (2020) 2343–2363.
[43] N.D. Hoang, K.W. Liao, X.L. Tran, J. Civ. Struct. Health 8 (2018) 431–442.
[44] Y. Liu, J. Wu, G. Yang, T. Zhao, S. Shi, Sci. Bull. 64 (2019) 1195–1203.
[45] X. Zhai, M. Chen, W. Lu, Comput. Mater. Sci. 151 (2018) 41–48.
[46] l. Breiman, Mach. Learn. 45 (2001) 5–32.
[47] Y. Li, C. Zou, M. Berecibar, E. Nanini-Maury, J.C.W. Chan, P. van den Bossche, J. Van Mierlo, N. Omar, Appl. Energy 232 (2018) 197–210.
[48] Z. Xu, J. Lian, L. Bin, K. Hua, K. Xu, H.Y. Chan, Water 11 (2019) 228.
[49] W. Zhang, C. Wu, Y. Li, L. Wang, P. Samui, Georisk: Assessment and Management of Risk for Engineered Systems and Geohazards (2019) 1–14.
[50] J. Kundu, B.K. Satpathy, D. Pradhan, Ind. Eng. Chem. Res. 58 (2019) 22709–22717.
[51] M. Umar, N. Mahmood, S.U. Awan, S. Fatima, A. Mahmood, S. Rizwan, RSC Adv. 9 (2019) 17148–17156.
[52] W. Wang, M.O. Tade, Z. Shao, Chem. Soc. Rev. 44 (2015) 5371–5408.
[53] X. Yu, J. Zhao, J. Huang, J. Zhao, Y. Guo, Y. Tang, X. Ma, Z. Li, Q. Guo, J. Zhao, J. Colloid Interface Sci. 572 (2020) 141–150.
[54] C. Chen, Y. Zuo, W. Ye, X. Li, Z. Deng, S.P. Ong, Adv. Energy Mater. 10 (2020) 1903242.
[55] Q. Zhang, D. Chang, X. Zhai, W. Lu, Chemom. Intell. Lab. Syst. 177 (2018) 26–34.
[56] J. Schmidt, M.R.G. Marques, S. Botti, M.A.L. Marques, npj Comput Mater. 5 (2019) 83.
[57] H. Wang, Y. Ji, Y. Li, WIREs Comput. Mol. Sci. 10 (2019) e1421.
[58] X. Zhai, M. Chen, W. Lu, D. Chang, J. Math. Chem. 56 (2018) 1744–1758.
[59] H. Peng, F. Long, C. Ding, IEEE Trans. Pattern Anal. Mach. Intell. 27 (2005) 1226–1238.
[60] V.F. Rodriguez Galiano, J.A. Luque Espinar, M. Chica Olmo, M.P. Mendes, Sci. Total Environ. 624 (2018) 661–672.
[61] M.H.M. Yusof, M.R. Mokhtar, A.M. Zain, C. Maple, Int. J. Adv. Comput. Sci. Appl. 9 (2018) 509–517.
[62] K.M. Parida, K.H. Reddy, S. Martha, D.P. Das, N. Biswal, Int. J. Hydrogen Energy 35 (2010) 12161–12168.
[63] V.M. Goldschmidt, Naturwissenschaften 14 (1926) 477–485.
[64] Q.D. Sun, W.J. Yin, J. Am. Chem. Soc. 139 (2017) 14905–14908.
[65] M. Ahmed, G. Xinxin, Inorg. Chem. Front. 3 (2016) 578–590.
[66] M.A. Bin Adnan, K. Arifin, L.J. Minggu, M.B. Kassim, Int. J. Hydrogen Energy 43 (2018) 23209–23220.
[67] W. Fang, W. Shangguan, Int. J. Hydrogen Energy 44 (2019) 95–912.
[68] Y. Li, J. Li, W. Yang, X. Wang, Nanoscale Horiz. 5 (2020) 1174–1187.
[69] X. Ning, G. Lu, Nanoscale 12 (2020) 1213–1223.
[70] P. Zhang, J. Zhang, J. Gong, Chem. Soc. Rev. 43 (2014) 4395–4422.
[71] Y. Xu, M.A.A. Schoonen, Am. Mineral. 85 (2000) 543–556.
[72] C. Pan, T. Takata, K. Kumamoto, S.S. Khine Ma, K. Ueda, T. Minegishi, M. Nakabayashi, T. Matsumoto, N. Shibata, Y. Ikuhara, K. Domen, J. Mater. Chem. A 4 (2016) 4544–4552.
[73] O.A. Carrasco Jaim, J.M. Mora Hernandez, L.M. Torres Martínez, E. Moctezuma, J. Photoch. Photobio. A 371 (2019) 98–108.
[74] P.G. Ralph, Inorg. Chem. 27 (1988) 734–740.
[75] T. Alammar, I.I. Slowing, J. Anderegg, A.V. Mudring, ChemSusChem 10 (2017) 3387–3401.
[76] J. John, M. Dhananjaya, S. Suresh, S. Savitha Pillai, M. Sahoo, O.M. Hussain, R. Philip, V.P. Mahadevan Pillai, J. Mater. Sci.: Mater. Electron. 31 (2020) 11159–11176.
[77] Y. Kumar, R. Kumar, R.J. Choudhary, A. Thakur, A.P. Singh, Ceram. Int. 46 (2020) 17569–17576.
[78] K.F. Moura, L. Chantelle, D. Rosendo, E. Longo, I.M.G. Santos, Mater. Res. 20 (2017) 317–324.
[79] T.T. Zhang, X.D. Gao, Y.Q. Wu, J.N. Yang, X.M. Li, Chem. Phys. 522 (2019) 91–98.
[80] T. Alammar, I. Hamm, V. Grasmik, M. Wark, A.V. Mudring, Inorg. Chem. 56 (2017) 6920–6932.
[81] M. Kim, B. Lee, H. Ju, J.Y. Kim, J. Kim, S.W. Lee, Adv. Mater. 31 (2019) 1903316.
[82] S. Omeiri, B. Hadjarab, A. Bouguelia, M. Trari, J. Alloys Compd. 505 (2010) 592–597.
[83] F. Zhong, H. Zhuang, Q. Gu, J. Long, RSC Adv. 6 (2016) 42474–42481.