

Article

Studying the Thermodynamic Phase Stability of Organic–Inorganic Hybrid Perovskites Using Machine Learning

Juan Wang ¹, Xinzhong Wang ¹, Shun Feng ¹ and Zongcheng Miao ^{2,*}

¹ Xi'an Key Laboratory of Advanced Photo-Electronics Materials and Energy Conversion Device, School of Electronic Information, Xijing University, Xi'an 710123, China; wangjuan@xijing.edu.cn (J.W.); fengshun@xijing.edu.cn (S.F.)

² School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an 710072, China

* Correspondence: miaozongcheng@nwpu.edu.cn

Abstract: As an important photovoltaic material, organic–inorganic hybrid perovskites have attracted much attention in the field of solar cells, but their instability is one of the main challenges limiting their commercial application. However, the search for stable perovskites among the thousands of perovskite materials still faces great challenges. In this work, the energy above the convex hull values of organic–inorganic hybrid perovskites was predicted based on four different machine learning algorithms, namely random forest regression (RFR), support vector machine regression (SVR), XGBoost regression, and LightGBM regression, to study the thermodynamic phase stability of organic–inorganic hybrid perovskites. The results show that the LightGBM algorithm has a low prediction error and can effectively capture the key features related to the thermodynamic phase stability of organic–inorganic hybrid perovskites. Meanwhile, the Shapley Additive Explanation (SHAP) method was used to analyze the prediction results based on the LightGBM algorithm. The third ionization energy of the B element is the most critical feature related to the thermodynamic phase stability, and the second key feature is the electron affinity of ions at the X site, which are significantly negatively correlated with the predicted values of energy above the convex hull (E_{hull}). In the screening of organic–inorganic perovskites with high stability, the third ionization energy of the B element and the electron affinity of ions at the X site is a worthy priority. The results of this study can help us to understand the correlation between the thermodynamic phase stability of organic–inorganic hybrid perovskites and the key features, which can assist with the rapid discovery of highly stable perovskite materials.

Keywords: organic–inorganic hybrid perovskites; thermodynamic phase stability; machine learning; LightGBM algorithms



Citation: Wang, J.; Wang, X.; Feng, S.; Miao, Z. Studying the Thermodynamic Phase Stability of Organic–Inorganic Hybrid Perovskites Using Machine Learning. *Molecules* **2024**, *29*, 2974. <https://doi.org/10.3390/molecules29132974>

Academic Editor: Hiroshi Sakiyama

Received: 20 May 2024

Revised: 5 June 2024

Accepted: 11 June 2024

Published: 22 June 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Organic–inorganic hybrid perovskites, comprising both organic and inorganic components, exhibit promising prospects for applications in photovoltaic power generation, luminescence, ferroelectricity, optical detection, and other fields [1,2]. Particularly in the field of solar cells, they offer advantages such as low production costs, minimal environmental impact, and high photoelectric conversion efficiency [3,4]. However, the commercialization process of perovskite solar cells has been impeded by their inherent instability [5,6]. The stability of these cells is primarily influenced by the structural stability of perovskite materials [7]. The structure of organic–inorganic hybrid perovskite materials is susceptible to damage in high-temperature or high-humidity environments, resulting in a layered perovskite structure [8]. Some perovskite materials are metastable at room temperature, and if they are exposed to moisture, heat, light, or polar solvents, the phase transition of perovskites can be accelerated, which prevents their application in solar cells [9,10].

Consequently, developing highly stable organic–inorganic perovskites holds significant potential for facilitating the commercial application of perovskite solar cells. However, identifying stable perovskite materials from the thousands of available perovskite materials presents significant challenges. Their stability is influenced by numerous complex factors, which cannot be accurately described using traditional theoretical models or empirical rules [11]. Therefore, identifying key factors that affect the stability of perovskite materials can provide crucial guidance for the research and development of highly stable perovskites.

The stability of perovskites has mainly been studied in terms of their structural stability and thermodynamic phase stability at present. The structural stability of perovskites can be researched using the tolerance factor. In [12], Bartel et al. developed an accurate, physically interpretable, and one-dimensional tolerance factor to predict the stability of perovskite oxides and halides, where the stability refers to the stability of the perovskite lattice structure. The thermodynamic phase stability is a key parameter that broadly governs whether the material is expected to be synthesizable, and whether it may degrade under certain operating conditions [13]. The energy above the convex hull (E_{hull}) provides a direct measure of the thermodynamic phase stability [14]. Meanwhile, Wanjian Yin found that there is no significant quantitative relationship between the thermodynamic stability of cubic perovskite and the tolerance factor [15]. Works on the evaluation of perovskite thermodynamic stability are important for validating the results recorded when researching perovskite stability according to the tolerance factor [16]. The E_{hull} value can be calculated using density functional theory (DFT). A greater positive value of E_{hull} indicates a lower stability [17]. However, the huge computational cost limits the use of DFT in large chemical spaces [18]. Compared with DFT calculations, the utilization of machine learning in material development offers significant advantages. It can effectively reduce the expenditure on research and development, expedite the process of material development, and enhance material performance, thereby presenting a promising prospect for the advancement of novel materials [19,20]. The machine learning (ML) method relies on experimental and computational data as the basis for learning [21]. By analyzing a large amount of test data, machine learning can identify key factors affecting the stability of materials and provide a sensitivity analysis to help optimize the preparation process of materials [22]. For example, machine learning models are trained using data on the structure and performance of known materials, which allows them to learn and establish relationships between the structure and performance of materials so as to predict and screen the materials with excellent properties [23]. Machine learning can simulate and calculate the materials of different components, analyze their advantages and disadvantages, and realize the optimization of components [24]. It can also analyze the optimal preparation conditions through various aspects of the material preparation process, which can guide the acquisition of optimal materials [25,26].

Since machine learning shows great potential for applications in material development, more and more scholars have applied it in the prediction of perovskite materials. Jie Zhao et al. [27] constructed classification and regression models to predict the thermodynamic stability and energy above the convex hull. They found that the highest occupied molecular orbital energy and the elastic modulus of the B-site elements for perovskite oxides are the top two features for stability prediction. Liu et al. [28] successfully classified 37 thermodynamically stable materials and 13 metastable oxide perovskites based on the E_{hull} value. Talapatra et al. [29] focused on the detailed chemical space of mono-oxide perovskites and di-oxide perovskites and optimally explored this space to identify stable compounds that might form using a random forest generation ML model. Yunlai Zhu et al. [30] developed an optimal model for predicting the thermodynamic stability of perovskites using machine learning based on a data set of 2877 ABX_3 and found that the extreme gradient-boosting regression technique showed the best performance in predicting the E_{hull} value, with an RMSE of 0.108 and an R^2 of 0.93. Bhattacharya et al. [31] revealed the key structural features that govern ABO_3 -type oxide perovskites and the extent of correlation between the stability and formability of these compounds using machine learning.

So, ML has applicability in the prediction of perovskite stability and has great potential to help accelerate the screening of high-stability perovskite materials.

For organic–inorganic hybrid perovskites (HOIPs), scholars have studied the band gap prediction of organic–inorganic perovskites [32], and the classification of two-dimensional and three-dimensional organic–inorganic perovskites [33] used in ML methods. Zhang et al. [19] proposed an interpretive ML strategy to speed up the discovery of potential HOIPs, which employed 102 samples as a training set. Yiqiang Zhan et al. pursued superior efficiency and thermal stability by combining machine learning with high-throughput computing to accelerate the discovery of hybrid organic–inorganic perovskites $A_2BB'X_6$ from a large number of candidates [34]. However, there are relatively few research projects on predicting the thermodynamic phase stability of organic–inorganic perovskites based on machine learning and exploring the relationship between thermodynamic phase stability and the features of elements in perovskites.

So, in this work, four methods, random forest regression (RFR), support vector machine regression (SVR), XGBoost regression, and LightGBM regression, were used to establish the prediction model of organic–inorganic hybrid perovskites to predict the energy above the convex hull (E_{hull}) value. At the same time, the Shapley Additive Explanation (SHAP) model was used to analyze the prediction results to help understand the correlation between the thermodynamic phase stability prediction and key features of organic–inorganic hybrid perovskites.

2. Results and Discussion

2.1. Feature Engineering

2.1.1. Feature Processing

The data set was preprocessed by the exploratory data analysis. By utilizing MinMaxScaler for feature scaling, the disparity in scales among features was mitigated, thereby promoting a more equitable distribution of weights across different features and enhancing both the performance and stability of the model [35]. The scaling of features can expedite the convergence process of the optimization algorithm and enhance the efficiency of model training. The MinMaxScaler operated based on the following principles:

- (1) Found the minimum (*min*) and maximum (*max*) for each feature in the training set;
- (2) Calculated, for each sample, the normalization result of its eigenvalue by the formula:

$$X_{\text{normalize}} = (x - \min) / (\max - \min); \quad (1)$$

- (3) The value of each feature was mapped to the interval of [0, 1].

The data are visualized and analyzed using a box plot and density map. The distribution and outlier of the data can be judged by observing the position and size of the box, the length of the longitudinal line, and the outlier points. Figure 1a shows the box plot of stability data. It can be observed that there are obvious outliers in the data of seven descriptors: crystal length, standard deviation of proportion for B and X atoms, percentage standard deviation of adjacent atoms in X position, sphericity of the B atom, standard deviation of the volume for atom at X position, and crystal structure at B element. The removal of these anomalous data is necessary during the process of feature manipulation. The density diagram of the data set about stability is shown in Figure 1b. The distribution and skewness of the original data in the data set can be assessed by examining the density curve. The majority of the density curves exhibit a unimodal symmetric distribution, implying a uniform data distribution. The density curves of these two descriptors, crystal structure at B element and standard deviation of the volume for an atom at X position, show left or right skew, indicating that there is skew in the data. The density curves of the radius of the B-site ion, the radius of the A-site ion, and the third ionization energy of the B element have multi-peak distributions, suggesting the presence of diverse data distributions.

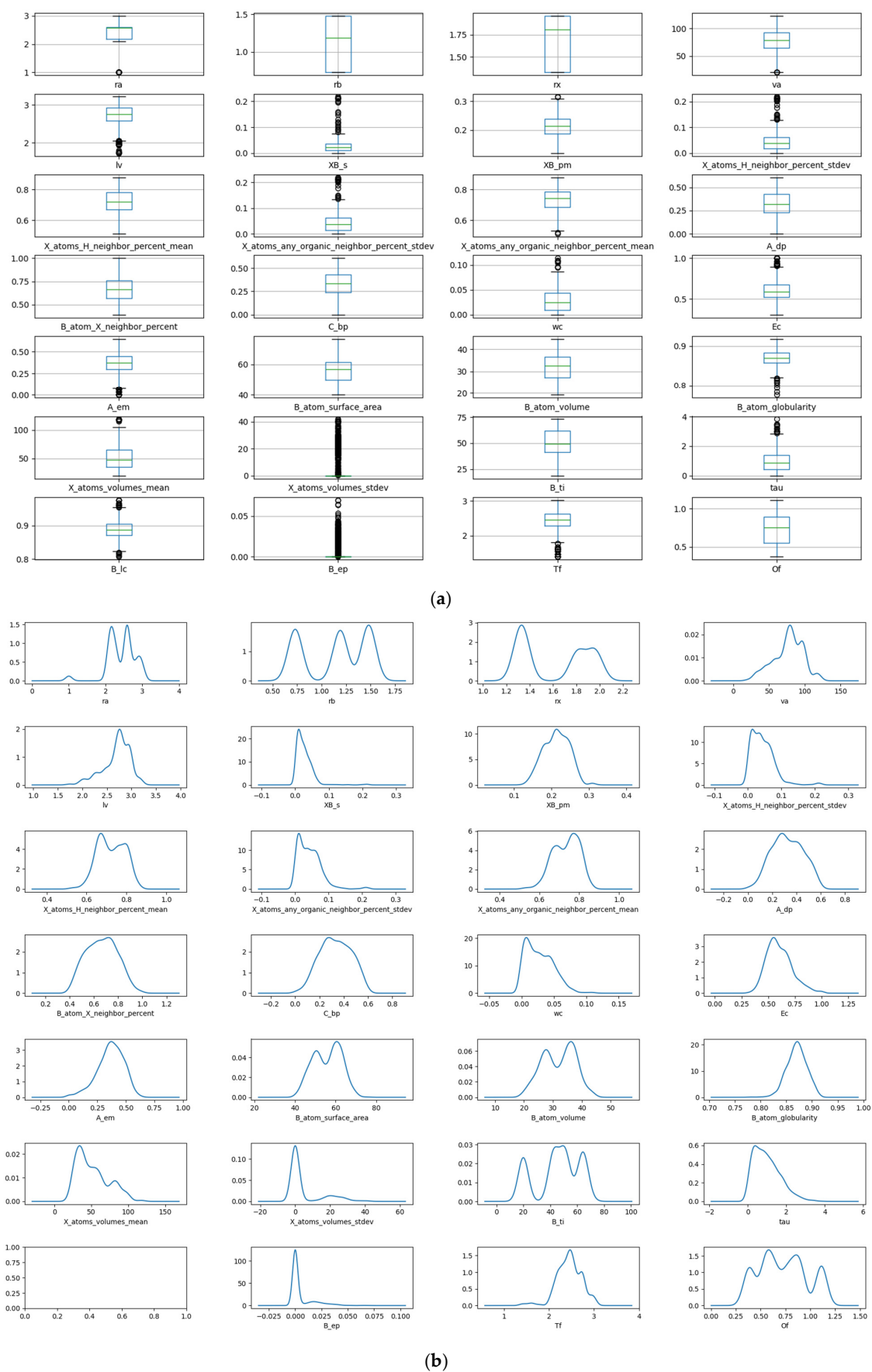


Figure 1. Box plot (a) and density map (b) of features about E_{hull} .

2.1.2. Feature Correlation Analysis

The technique of feature selection plays a pivotal role in the field of machine learning. The main purpose is to select the most relevant or representative subset of features from the raw data in order to effectively reduce the dimension of data and improve the performance of the model. It can help weed out useless or redundant features, make the model easier to understand and interpret, and improve the accuracy and generalization of the model. In feature selection, the Pearson correlation coefficient [36] is used to evaluate the correlation between each feature and the target variable, and the features that exhibit a strong correlation with the target variable are selected. The Pearson correlation coefficient is a statistic that measures the strength of the linear relationship between two variables, and it is often used to assess the correlation between two features. Its definition is shown in Formula (2).

$$\rho_{X,Y} = \frac{\text{cov}(X,Y) = 1/(n-1) \sum_{i=1}^n (X_i - X^-)(Y_i - Y^-)}{\sigma_X \sigma_Y} \quad (2)$$

The value of the Pearson correlation coefficient ranges from -1 to 1 , where -1 means a complete negative correlation, 1 means a complete positive correlation, and 0 means no linear correlation. Firstly, unnecessary features are eliminated to reduce the dimension of the data set. In order to eliminate the influence of unnecessary features on the accuracy of the prediction model, 28 features whose correlation coefficient is close to 0 were excluded. The data set was initially screened by the derived correlation coefficient. The correlation coefficient among the ra, rb, B_ep, Tf, Of, X_atoms_volumes_mean, X_atoms_volumes_stdev, B_atom_globularity, XB_s, and rx is all in the range of $0-0.2$, indicating that these features are not correlated. Therefore, the above features are not taken into consideration in predicting E_{hull} . We removed irrelevant features and retained the related features of lv, A_em, tau, A_dp, va, B_atom_X_neighbor_percent, B_atom_surface_area, B_atom_volume, B_ti, B_lc, XB_pm, and X_atoms_H_neighbor_percent_stdev, X_atoms_volumes_stdev, X_atoms_H_neighbor_percent_mean, X_atoms_any_organic_neighbor_percent_stdev, X_atoms_any_organic_neighbor_percent_mean, C_bp, wc, and Ec. The Pearson correlation coefficient of these features is shown in Figure 2. The gradient color bar on the right corresponds to the magnitude of the correlation coefficient. The color red indicates a positive correlation, while blue signifies a negative correlation; furthermore, the intensity of the color corresponds to the strength of the correlation.

As shown in Figure 2, when the Pearson correlation coefficient between two features is close to 1 , it means that when one feature increases, the other feature also increases in almost the same proportion, which usually causes the problem of multicollinearity. The presence of multicollinearity can have a detrimental impact on the performance of linear models that rely on estimated coefficients. In this case, the model's stability may be compromised, leading to unreliable estimates of coefficients and diminished interpretability. For example, the correlation coefficient between va and lv is as high as 0.99 , indicating that the two features have a high linear correlation. In order to reduce the dimension of data, consider removing one of these highly correlated features. The method of deleting highly correlated features is as follows: when the Pearson correlation coefficient of two features A and B is close to 1 , delete any one of the features, construct the feature set separately, and perform the prediction. Compare the R^2 values of the prediction model between the two cases of feature A deleted and feature B deleted. Select the case with a high R^2 . At the same time, MSE and MAE should also be considered. After deleting one of the highly correlated features, MSE is used to determine whether the model would be over-fitting or under-fitting. In addition, it is also necessary to consider key characteristics affecting thermodynamic stability, such as va. Although va and lv are highly correlated, the chemical lattice constant va is strongly related to the stability of the compound. So, the va feature is preserved.

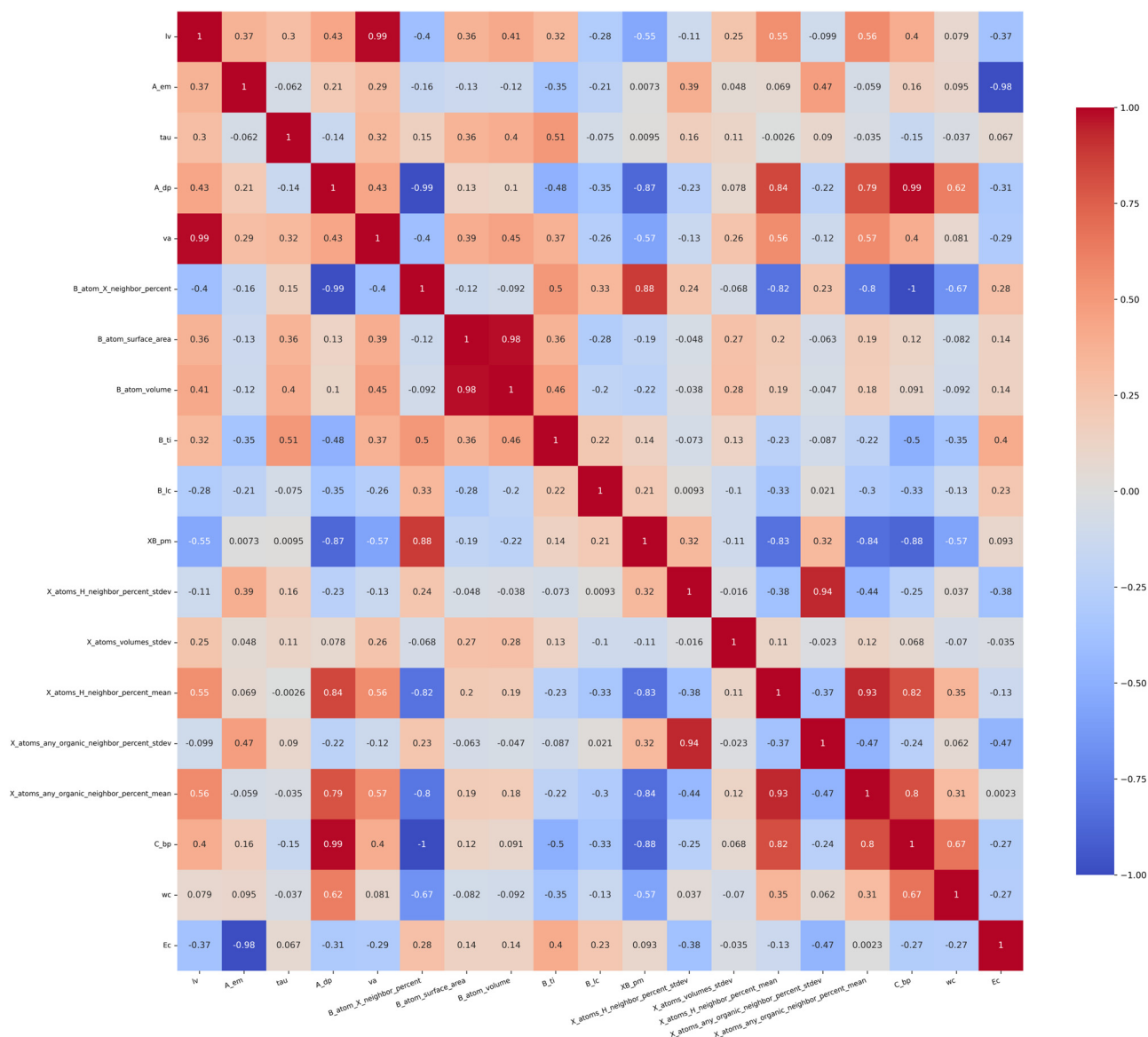


Figure 2. Pearson correlation coefficients of features about E_{hull} .

After the Pearson correlation analysis, redundant features with particularly high linear correlations were screened out and scatter plots were drawn between crystal length, crystal volume, and target variables, as shown in Figure 3. The diagram shows the relationship between the two variables and E_{hull} . In the scatter plot, each data point represents a sample, blue is the point of the training set, red is the point of the test set, and its x -axis and y -axis coordinates represent the values of the two variables, respectively. By observing the distribution and trend of data points in the scatter plot, it is found that the relationship between va and lv and E_{hull} has a nonlinear relationship.

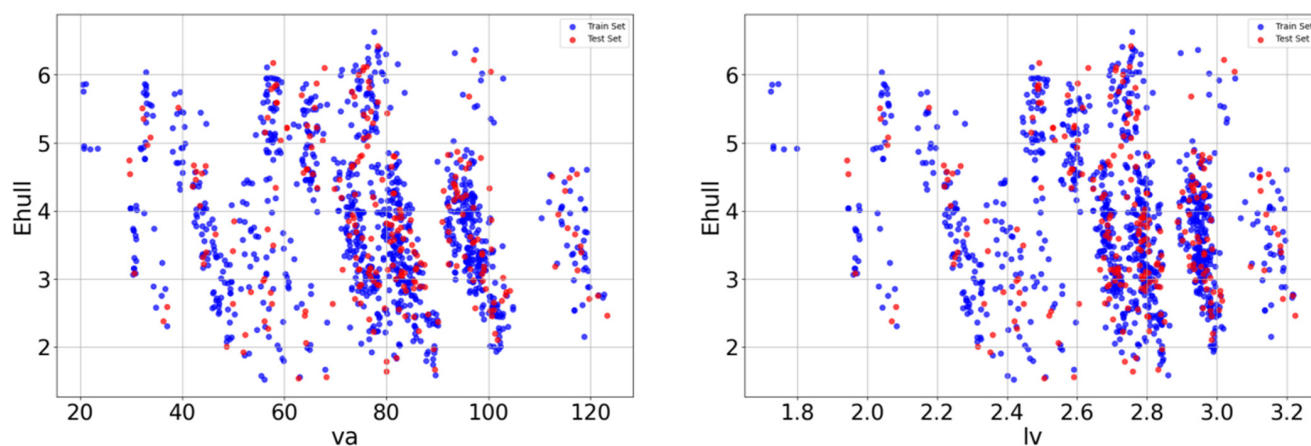


Figure 3. Relationship between v_a and l_v characteristics and E_{hull} .

2.1.3. Feature Screening

Feature screening is a continuous iterative process, which needs to consider the correlation between data features and target variables, and between features. If multiple features are highly correlated, consider selecting only one of them to reduce redundant information. This can reduce the dimension of the input feature matrix, help screen out important features related to the thermodynamic phase stability of organic-inorganic hybrid perovskites, and shorten the training data time. Four feature selectors were established based on the GBR algorithm, the RFR algorithm, the ETR algorithm, and the AdaBoost algorithm, which were used for feature screening. The worst feature was discarded each time. The score of the four algorithms on the number of features in the data set is shown in Figure 4. In the AdaBoost regression model, the optimal R^2 value is 0.905936, and the number of features is 17, which includes rx , v_a , l_v , XB_s , XB_{pm} , $X_{atoms_any_organic_neighbor_percent_stdev}$, $X_{atoms_any_organic_neighbor_percent_mean}$, A_{dp} , $B_{atom_X_neighbor_percent}$, C_{bp} , wc , Ec , A_{em} , $B_{atom_surface_area}$, B_{atom_volume} , $B_{atom_globularity}$, $X_{atoms_volumes_mean}$, B_{ti} , τ , Tf , and Of . In the ETR model, the optimal R^2 value is 0.951990, and the number of features is 17, which includes rx , l_v , XB_s , $X_{atoms_H_neighbor_percent_stdev}$, $X_{atoms_any_organic_neighbor_percent_stdev}$, $X_{atoms_any_organic_neighbor_percent_mean}$, A_{dp} , C_{bp} , wc , Ec , A_{em} , $B_{atom_surface_area}$, B_{atom_volume} , $B_{atom_globularity}$, $X_{atoms_volumes_mean}$, B_{ti} , and Of . In the GBR model, the optimal R^2 value is 0.943720, and the optimal number of features is 14, which includes rx , v_a , l_v , $X_{atoms_any_organic_neighbor_percent_stdev}$, $X_{atoms_any_organic_neighbor_percent_mean}$, C_{bp} , wc , Ec , A_{em} , and B_{atom_volume} , $B_{atom_globularity}$, $X_{atoms_volumes_mean}$, B_{ti} , and Of . In the RFR model, the optimal R^2 value is 0.94585, and the number of features is 12, which includes rx , l_v , $X_{atoms_any_organic_neighbor_percent_stdev}$, $X_{atoms_any_organic_neighbor_percent_mean}$, C_{bp} , wc , Ec , and $B_{atom_surface_area}$, $B_{atom_globularity}$, $X_{atoms_volumes_mean}$, B_{ti} , and Of . Among these algorithms, the RFR regression algorithm has the highest optimal R^2 in the feature screening. Among the above four algorithms, the optimal R^2 value of ETR is the highest, but the complexity of the ETR model is small, which is prone to underfitting problems. At the same time, the feature subset after ETR screening is input into the prediction model of LightGBM. The MSE and MAE of the prediction model are, respectively, 0.0681 and 0.1886, which are significantly higher than those obtained by using the features selected by random forest as inputs. So, the 12 features of the RFR algorithm are selected as feature subsets, and optimally divide the data set into an 8:2 test set and a training set.

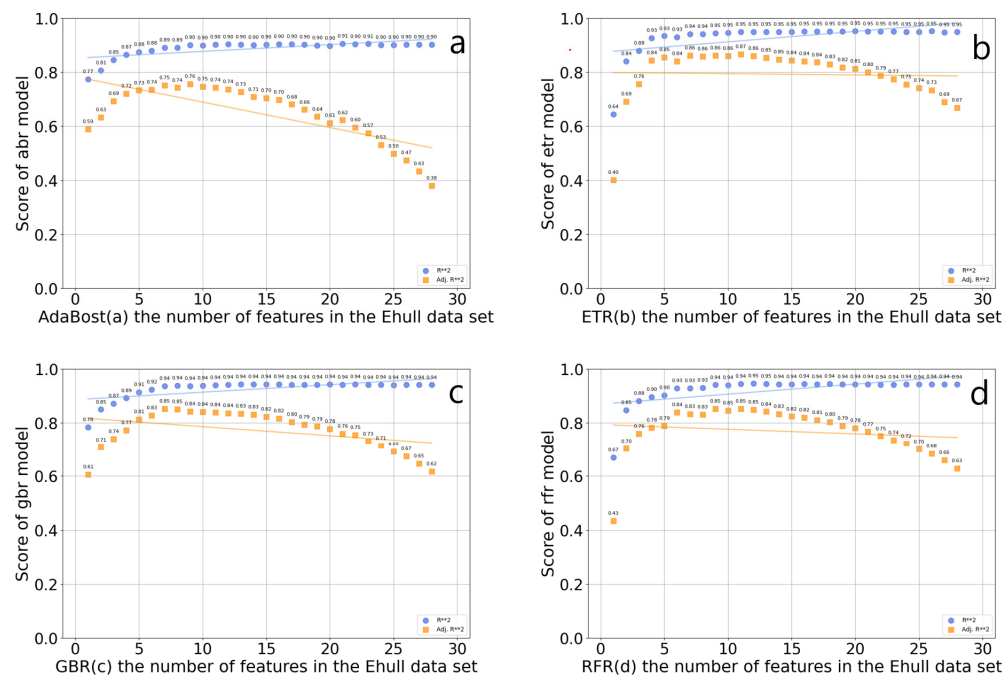


Figure 4. Scores of the four algorithms AdaBoost (a), ETR (b), GBR (c), and RFR (d) on the number of features in the E_{hull} data set.

2.2. Model Prediction Results

To mitigate dimensional disasters and ensure effective training of the model without over-fitting, the number of descriptors will be constrained to a quantity smaller than the data set's sample size. Simultaneously, during the hyperparameter tuning process, the model's performance is evaluated across varying ratios of training and test set, while 10-fold cross-validation is to mitigate issues of under-fitting and over-fitting.

The performance of the prediction model utilizing four machine learning algorithms, RFR, SVR, XGBoost, and LightGBM, is shown in Table 1. The R^2 of the LightGBM regression model is 0.953914, which measures the proportion of variance between the predicted value and the true value. The R^2 value ranges from 0 to 1, and the closer it approaches 1, the higher the predictive power of the model. The MAE value is more sensitive to small prediction errors, while the MSE value is more sensitive to large prediction errors. The MAE value of the LightGBM regression is 0.1664, which is the minimum value among the four algorithms. And the MSE value of the LightGBM regression is 0.0531, which is also the smallest among the four models and is closest to 0. The MAE and MSE values indicate that the LightGBM regression model exhibits a smaller prediction error.

Table 1. Comparison of performance outcomes from regression models.

Algorithm	Evaluating Indicator		
	MAE	MSE	R^2
RFR	0.1884	0.0662	0.942552
SVR	0.1713	0.0614	0.946759
XGBoost	0.1839	0.0689	0.940207
LightGBM	0.1664	0.0531	0.953914

The training set was trained using four machine learning algorithms, RFR, SVR, XGBoost, and LightGBM. In Figure 5, the horizontal coordinate is the real values of E_{hull} , which are obtained through DFT calculations and derived from the data set. The vertical coordinate is the E_{hull} value predicted using machine learning. The prediction of the E_{hull} value of organic–inorganic hybrid perovskite based on the four machine learning

algorithms of RFR, SVR, XGBoost, and LightGBM tends to be a straight line with the E_{hull} value of organic–inorganic hybrid perovskite calculated by DFT. It can be concluded that these four algorithms are good at predicting and fitting the E_{hull} value of organic–inorganic hybrid perovskites. The predictive model established by the LightGBM algorithm is better than the SVR algorithm, the RFR algorithm, and the XGBoost algorithm, and the accuracy of the LightGBM model is also greater. The LightGBM regression prediction model emerges as the optimal choice for accurately predicting the E_{hull} values of organic–inorganic hybrid perovskites.

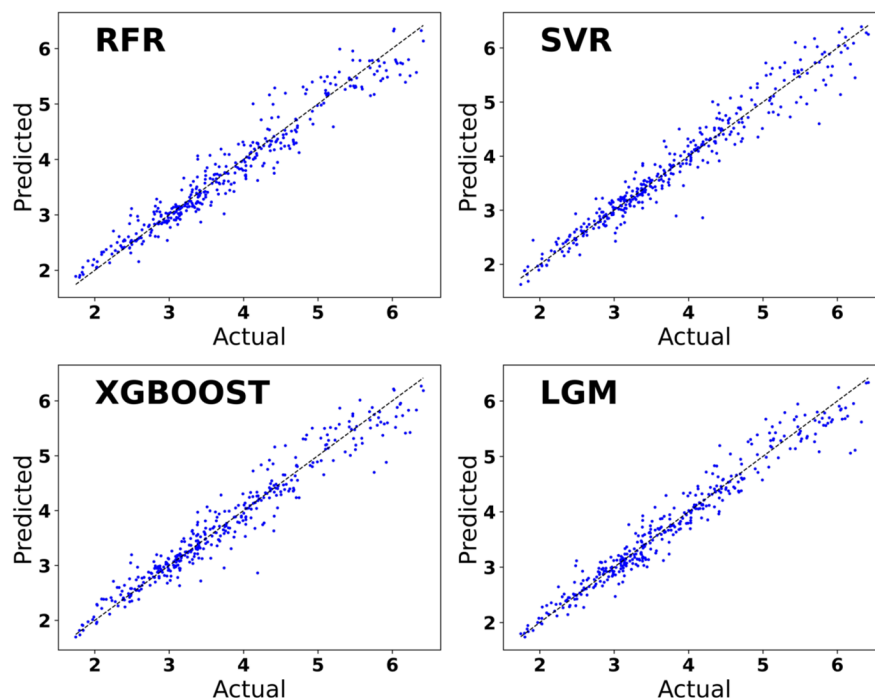


Figure 5. Prediction of E_{hull} used regression models based on the machine learning algorithm.

Since the LightGBM prediction model has achieved the best performance in the prediction of E_{hull} for organic–inorganic hybrid perovskites, only the first 20 samples of its test results are recorded in Table 2 for the sake of space.

Table 2. Prediction results of the LightGBM model for E_{hull} on the test set.

HOIP Order Number	E_{hull} (eV/Oatom)	ML_ E_{hull} (eV/Oatom)	Prediction Error
1	3.087	3.070	0.016
2	5.6831	5.309	0.373
3	4.3624	4.366	−0.003
4	4.6596	5.407	−0.747
5	3.6561	3.704	−0.048
6	5.039	5.743	−0.704
7	3.5853	3.764	−0.179
8	5.8538	5.801	0.051
9	3.196	3.727	−0.531
10	4.0485	3.933	0.115
11	2.9874	3.007	−0.020
12	5.8294	5.823	0.006
13	4.622	4.419	0.202
14	3.5053	3.530	−0.025
15	4.9335	4.754	0.179
16	2.6572	2.531	0.125

Table 2. Cont.

HOIP Order Number	E_{hull} (eV/Oatom)	ML_ E_{hull} (eV/Oatom)	Prediction Error
17	5.1533	4.809	0.344
18	2.7138	2.542	0.171
19	2.9218	3.237	−0.315
20	1.9719	2.0300	−0.058

2.3. Interpretation of Prediction Model by SHAP Method

The SHAP method is invaluable for model interpretation and feature selection. It can also serve to validate the accuracy and reliability of the model, identify the weaknesses of the model, and propose improvement schemes. From the interpretation diagram of the SHAP model, the contribution degree of each feature to the prediction results, and the direction and degree of influence of the feature value can be understood. The input feature subsets of the prediction model established by the LightGBM algorithm are rx, lv, X_atoms_any_organic_neighbor_percent_stdev, X_atoms_any_organic_neighbor_percent_mean, C_bp, wc, Ec, B_atom_surface_area, B_atom_globularity, X_atoms_volumes_mean, B_ti, and Of, which are picked out by the RFR algorithm. The SHAP method decomposes the contribution degree of each feature to the prediction result into the sum of the contribution degree of different feature subsets.

In Figure 6, the SHAP values of 12 features are listed in the order of the importance of features from high to low. The horizontal coordinate represents the SHAP value, which is the contribution of each feature to the model's prediction results. A negative value indicates that the feature is negatively correlated with the prediction results of E_{hull} , while a positive value indicates that the feature is positively correlated with the prediction results of E_{hull} . The significance of features can be assessed based on the absolute magnitude of the SHAP value. The scatter points in the data set samples represent different observations, with blue dots indicating smaller eigenvalues and red dots representing larger eigenvalues. From Figure 6, the third ionization energy of the B element is the most critical feature in the prediction of E_{hull} , and the second key feature is the electron affinity of ions at the X site. They are significantly negatively correlated with the predicted values of E_{hull} . The octahedral factors, the average percentage of atoms in the X position adjacent to any atom in the A position, crystal length, and mean volume of the atom at the X position are negatively correlated with predicted E_{hull} . The radius of the atom at the X position, atomic weight at the X position, and surface area of the atom at the B position are positively correlated with the predicted E_{hull} . In addition, the sphericity of the B atom and boiling point at the X position are also positively correlated with the predicted E_{hull} . The magnitude of the SHAP value indicates the relative importance of the corresponding feature, signifying its dominant role in predicting E_{hull} . A large value of E_{hull} indicates that the organic–inorganic perovskites are less stable. So, when the values of the third ionization energy of the B element, the electron affinity of ions at X position, the average percentage of atoms in X position adjacent to any atom in A position, crystal length, and mean volume of the atom at X position are larger, the thermodynamic phase stability of perovskites is higher. When the values of the radius of the atom at the X position, atomic weight at the X position, the surface area of the atom at the B position, the sphericity of the B atom, and boiling point at the X position are smaller, the stability of organic–inorganic perovskites is higher. In physics, the third ionization energy is the energy at which an atom's electrons are separated from its nucleus between two ionization levels. It significantly indicates chemical properties, including the tendency of high-energy ion reactions, molecular stability, and conformational transformations. The higher the ionization energy of an atom, the greater its structural stability. Among the results of prediction, the third ionization energy of the B element and electron affinity of ions at the X site are the two most critical characteristics. It is indicated that LightGBM regression is feasible to predict the thermodynamic stability of perovskites, and the prediction of E_{hull} can assist in verifying the structural stability obtained by tolerance factor analysis.

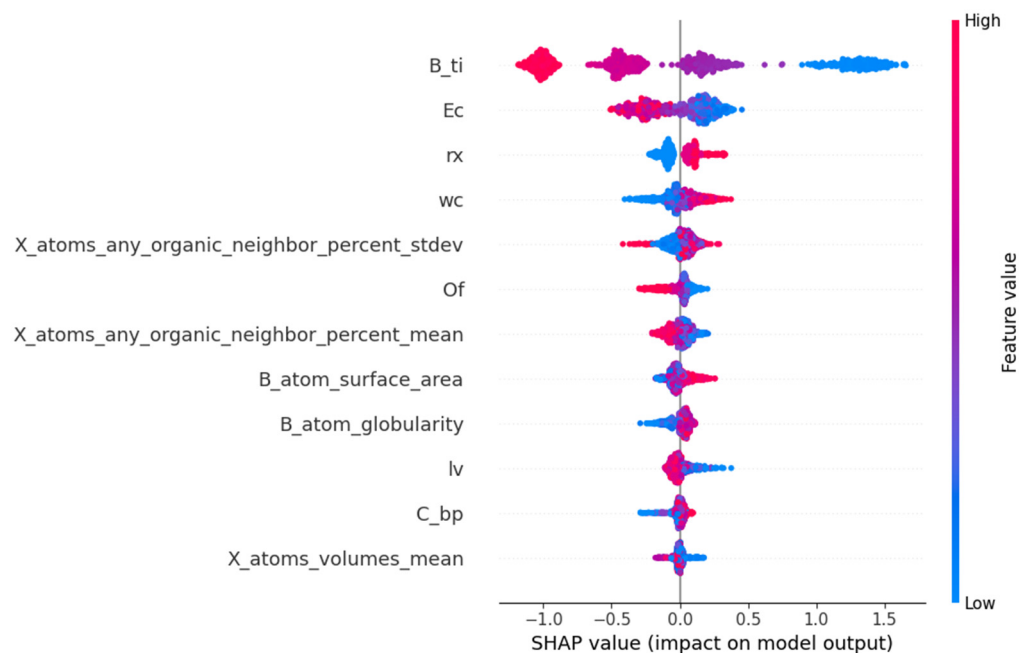


Figure 6. The feature importance ranking of stability for organic–inorganic hybrid perovskites obtained by SHAP value.

3. Data and Methods

3.1. Data Sources

Raw data sets include 1254 raw data, which come from the Materials Project, an open access materials database (established by Lawrence Berkeley National Laboratory and Carnegie Mellon University), and some of the literature [30,37]. The structure of perovskites is typically characterized by octahedral arrangements, as illustrated in Figure 7. The ions at the A site, B site, and X site contained in the data set are detailed below. The ions at the A site are organic amine cations; the ions at the B site are metal cations of the elements Ge, Pb, and Sn; and the ions at the X site are anions of the elements C, Br, I, and Cl. The feature descriptors of the thermodynamic phase stability prediction for organic–inorganic hybrid perovskites are seen in Table 3. The data set was randomly divided into an 80% training set and a 20% validation set to train and validate the machine learning model.

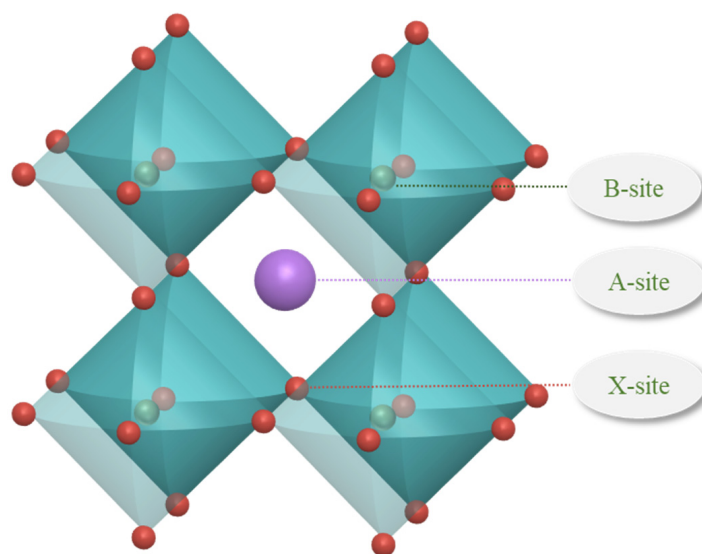


Figure 7. Structure diagram of perovskites.

Table 3. Feature descriptors of E_{hull} for organic–inorganic hybrid perovskites.

Feature	Unit	Description
A site	None	Chemical formula for the ion at A site
B site	None	Chemical formula for ions at B site
X site	None	Chemical formula for ions at X site
rb	Å	Radius of the ion at B site
ra	Å	Radius of the ion at A site
tf	None	Tolerance factor
lv	Å	Crystal length
XB_s	Å	Proportion of B and X atoms standard deviation
XB_pm	None	Average percentage of adjacent X and B atoms
X_atoms_H_neighbor_percent_stdev	None	Percentage standard deviation of adjacent atoms in X position
X_atoms_volumes_stdev	None	Standard deviation of volume for X atom
X_atoms_H_neighbor_percent_mean	Å	Average percentage of adjacent X atom
X_atoms_any_organic_neighbor_percent_stdev	None	Percentage standard deviation of X and A positions
X_atoms_any_organic_neighbor_percent_mean	Å	Average percentage of atoms in X position adjacent to any atom in A position
A_dp	C/m ²	Dipolarizability
B_atom_X_neighbor_percent	Å	Percentage of atoms at B position to X position
C_bp	°C	Boiling point at X position
wc	Å	Atomic weight at X position
Ec	eV	Electron affinity of ions at X position
A_em	Å	Lattice constant of ions at A position
B_atom_surface_area	Å ³	Surface area of atom at B position
B_atom_volume	Å ³	Volume of atom at B position
B_atom_globularity	°	Sphericity of the B atom
X_atoms_volumes_mean	Å	Mean volume of the atom at X position
X_atoms_volumes_stdev	Å ³	Standard deviation of the volume for atom at X position
of	None	Octahedral factors
rx	Å	Radius of atom at X position
tau	None	New tolerance factor
B_ti	eV	Third ionization energy of B element
B_lc	nm	Lattice constant of B element
B_ep	None	Crystal structure at B element

3.2. Machine Learning Algorithms and Model Evaluation

3.2.1. Gradient-Boosting Regression

Gradient-boosting regression (GBR) is an ensemble learning algorithm. It combines multiple weaker learning algorithms, typically decision trees, to create a more powerful model for improving prediction accuracy [38]. The algorithm flow of GBR is as follows: through continuous iteration, multiple weak regression models are combined to form a strong model.

Let the training data set be $\{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$, where $x_i \in R^m$, $y_i \in R$, m is the characteristic number. First, initialize the model according to the following function:

$$f_0(x) = \operatorname{argmin}_{\gamma} \sum_{i=1}^n L(y_i, \gamma) \quad (3)$$

For each iteration step $t = 1, 2, \dots, T$, calculate the negative gradient by Formula (4); by fitting the residual, the leaf node region of the t -th tree is obtained, and the output value of each leaf node is calculated by Formula (5). The model is then updated, as shown in Formula (6), to return the final model $f_T(x)$.

$$r_{it} = -\left[\frac{\partial L(y_i, f(x_i))}{\partial f(y_i)}\right]_{f(x)=f_{t-1}(x)} \quad (4)$$

$$\gamma_{jt} = \operatorname{argmin}_{\gamma} \sum_{x_i \in R_{jt}} L(y_i, f_{t-1}(x_i) + \gamma) \quad (5)$$

$$f_t(x) = f_{t-1}(x) + \sum_{j=1}^J \gamma_{jt} I(x \in R_{jt}) \quad (6)$$

where $L(y, f(x))$ is the loss function, $I(x \in R_j)$ is the indication function. In the GBR algorithm, the square error loss function is usually used.

3.2.2. Random Forest Regression

Random forest regression (RFR) is also an ensemble learning algorithm. It constructs multiple decision trees to generate predictions and subsequently combines the predictions from each tree to form the final outcome [39]. The random forest regression algorithm can effectively deal with the over-fitting problem and improve the prediction ability of the model. The proposed algorithm exhibits superior performance in handling high-dimensional and nonlinear data compared to conventional regression methods. The proposed method simultaneously achieves high accuracy and robustness, thereby effectively shortening the development cycle and reducing costs in the material development process.

The main steps of the random forest regression algorithm are as follows. (1) Select p samples from the training set by a “self-help method” to form a training set subset as a new training set. (2) When constructing the decision tree, if each sample has K attributes, then K (k less than K) attributes are randomly selected from these K attributes during the splitting process of each node. Then, according to the variance of these k attributes, one is selected as the best split attribute of the node. (3) During the generation of the decision tree, the splitting of each node is performed iteratively according to step 2 until further splitting becomes infeasible. The decision tree is trained by using this subset, and no pruning is performed on this decision tree. (4) Build a large number of decision trees according to steps 1–3 until m decision trees are trained. (5) For each test sample, regression prediction is performed individually by every decision tree, followed by statistical analysis of the prediction results from all decision trees for the same sample. The final prediction value is obtained by calculating the average of these results.

3.2.3. Extra Trees Regression

The extra trees regression algorithm (ETR) is mainly used to solve regression problems [40]. The core idea of the algorithm is to obtain the final predicted value by constructing multiple decision trees and a weighted average of the results of each decision tree. In the process of constructing each decision tree, the ETR regression algorithm randomly selects a part of the samples and a feature subset of the training set for training to minimize the error. It can process the missing data, and optimize the model parameters by using the self-help method and cross-validation, so as to improve the accuracy and generalization ability of the model.

3.2.4. Support Vector Regression

Support vector regression (SVR) [41] is a regression algorithm based on support vector machines. In contrast to conventional regression algorithms, SVR does not directly predict the value of the target variable; instead, it optimizes the margin between training data points to identify a function that can yield superior predictions for unknown data. According to the actual problem, the algorithm can improve the generalization ability of the model by adjusting the parameters of the kernel function and penalty function. Support vector regression algorithm performs better than traditional machine learning algorithms in handling experimental or observational data with low noise, especially in machine learning with small sample data.

3.2.5. Adaptive Boosting

The AdaBoost algorithm (AdaBoost) is an ensemble learning method based on the boosting strategy, which trains weak learners sequentially in a highly adaptive way [42]. Its core idea is to adjust the weight of error samples, and then upgrade iteratively.

AdaBoost’s regression algorithm process is as follows.

(1) Initialize the weights of each sample (all weights are equal) by the Formula (7).

(2) Calculate the error rate. According to the calculation formula of the error rate (see Formula (8)), construct the weak learner with the smallest error rate.

(3) Adjust the weight of the weak learner to obtain the strong learner of the m iteration.

(4) Increase the weight of incorrectly classified samples and reduce the weight of correctly classified samples, so that more accurate classification can be achieved later.

(5) Iterate over the above process until the error score reaches a threshold or the number of iterations reaches a set maximum. The expression of the final strong learner obtained after several iterations is shown in Formula (9).

$$w_{mi} = \frac{1}{N} (i = 1, 2, \dots, N) \quad (7)$$

$$e_m = \sum_{i=1}^N w_{mi} I(F_m(x_i) \neq y_i) \quad (8)$$

$$\text{sign}[f_M(x)] = \text{sign}\left[\sum_{i=1}^M a_i F_i(x)\right] \quad (9)$$

where the error rate e_m is the sum of the weights of the misclassified samples; the predicted value $F_m(x_i)$ refers to the classification of the sample i predicted by the weak learner $F_m(x)$; w_{mi} is the weight of the sample; y_i is the actual value; $I(F_m(x_i) \neq y_i)$ is the indicating function, whose value is 1 if the prediction fails and the condition in parentheses holds, and, otherwise, whose value is 0. And where $\text{sign}(x)$ is the symbolic function, as shown in the following formula.

$$\text{sign}(x) = \begin{cases} 1, & x > 0 \\ 0, & x = 0 \\ -1, & x < 0 \end{cases} \quad (10)$$

3.2.6. Extreme Gradient Boosting (XGBoost)

The extreme gradient-boosting (XGBoost) algorithm [43] is an algorithm implementation based on the boosting framework. It conforms to the form of a model function, and the output of the model can be expressed as the sum of the output of K weak learners. The basic idea of the algorithm is to reduce the bias of the model by constantly generating new trees, each tree learning based on the difference between the previous tree and the target value to reduce the bias of the model. The output of the final model result is shown in the Formula (11). That is, the results of all trees add up to what the model predicts for a sample.

$$\hat{y}_i^{(t)} = \sum_{k=1}^i f_k(x_i) = \hat{y}_i^{(t-1)} + f_t(x_i) \quad (11)$$

where $\hat{y}_i^{(t)}$ is the prediction result of the sample i after the t -th iteration, $\hat{y}_i^{(t-1)}$ is the prediction result of the front $t - 1$ tree, and $f_t(x_i)$ is the model of the t -th tree.

The objective function consists of the loss function of the model and the regular term that inhibits the complexity of the model, and the formula is as follows.

$$\text{Obj} = \sum_{i=1}^n l(y_i, \hat{y}_i) + \sum_{i=1}^t \Omega(f_i) \quad (12)$$

where $\sum_{i=1}^n l(y_i, \hat{y}_i)$ is to sum the complexity of all trees and it is added to the objective function as a regularization term to prevent the model from over-fitting.

3.2.7. Lightweight Gradient Lifting Algorithm (LightGBM)

The lightweight gradient lifting algorithm (LightGBM) is a machine learning algorithm based on a gradient lifting decision tree (GBDT) developed by the Microsoft team [44]. The LightGBM uses a histogram-based decision tree algorithm and mutually exclusive feature bundling to process a large number of data instances and a large number of features, which improves the training efficiency and accuracy of the model. Compared with the traditional GBDT algorithm, the LightGBM has higher accuracy and faster training speed.

Suppose the data set is $H = \{(x_i, y_i)\}_{i=1}^n$, use the LightGBM algorithm to find the proximate value $f(x)$ of a function $\hat{f}(x)$, and minimize the loss function $L(y, f(x))$ through the function. To judge the quality of the model fitting data, observe the size of the loss function, whose optimization function can be expressed as follows:

$$\hat{f} = \arg \min_f E_{y,X} L(y, f(x)) \quad (13)$$

Meanwhile, the LightGBM model integrates k regression trees to fit the final model, which can be expressed as follows:

$$f_k(x) = \sum_{i=1}^k f_i(H) \quad (14)$$

3.3. Evaluation Indicators of Model

The performance of regression models is evaluated using the following metrics: mean squared error (MSE), mean absolute error (MAE), and R^2 . The MSE measures the sum of squares of the difference between the predicted value and the true value, and the MAE is the average of the difference between the predicted value and the true value, which are indicators used to measure forecasting error [45,46]. But, the difference is that MSE is more sensitive to larger prediction errors, while MAE is more sensitive to smaller prediction errors [47]. The R^2 value is another commonly used metric that measures how well the model explains changes in the data, representing the degree of linear correlation between the regression value and the true value. The value of R^2 ranges from 0 to 1, and a higher R^2 indicates a better model explaining changes in the data.

Their formulas are as follows:

$$MSE = \frac{1}{m} \sum_{i=1}^m (f_i - y_i)^2 \quad (15)$$

$$MAE = \frac{1}{m} \sum_{i=1}^m |y_i - \hat{y}_i| \quad (16)$$

$$R^2 = 1 - \frac{\sum_{i=0}^{n-1} (y_i - f_i)^2}{\sum_{i=0}^{n-1} (y_i - \bar{y}_i)^2} \quad (17)$$

where m is the number of samples; f_i is the true value; y_i is the predicted value.

4. Conclusions

In this work, the original data set of E_{hull} for organic–inorganic hybrid perovskites were constructed using crystal structure data and material composition, and the data were preprocessed by exploratory data analysis. And the Pearson correlation coefficient analysis was employed to compare the original features, select effective features as feature subsets, and optimally divide the data set into an 8:2 ratio for testing and training purposes. Various machine learning prediction models were established using the four kinds of machine learning algorithms: RFR, SVR, XGBoost, and LightGBM, to predict the E_{hull} of organic–inorganic hybrid perovskites. The results show that the prediction model of the LightGBM algorithm has the best prediction effect with a MAE value of 0.1664, a MSE value of 0.0531, and an R^2 value of 0.953914. The model exhibits a low prediction error, which indicates that the input features fit the prediction model well. Compared with the original input features and default parameters, the performance of the model is greatly improved. The prediction results of the LightGBM model were analyzed by the SHAP value, and it was found that the third ionization energy of B element and electron affinity of ions at X site contributed more to the stability prediction and they are significantly negatively correlated with the prediction results of E_{hull} value. A large value of E_{hull} indicates that the

organic–inorganic perovskites are less stable. The stability of perovskites is closely related to its crystal structure. The higher values for the third ionization energy of the B element and the electron affinity of ions at the X site may help maintain the stability of the crystal structure. In the screening of organic–inorganic perovskites with high stability, priority can be given to the third ionization energy of B element and electron affinity of ions at the X site. The results are expected to provide guidance for the design and synthesis of novel organic–inorganic hybrid perovskites and provide references for machine learning prediction research of other materials.

In the future, various methods should be tested to reduce multicollinearity, such as principal component analysis (PCA), and data enhancement techniques for limited data sets will be explored, in order to improve the accuracy of the machine learning model. Moreover, machine learning and DFT methods will be used to select candidate element combinations with high thermodynamic phase stability.

Author Contributions: J.W.: methodology of experiment, data curation, and writing—original draft; X.W.: data curation; S.F.: data curation; Z.M.: data curation and writing—review and editing. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China (No. 52173263), the Shaanxi Association for Science and Technology Youth Talent Support Program (No. 20230520), the Xi'an Science Technology Plan Project of Shaanxi Province (No. 2023JH-ZCGJ-0068), and the Qinchuangyuan High-level Talent Project of Shaanxi (No. QCYRCXM-2022-219).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data are contained within the article.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Ni, H.; Ye, L.; Zhuge, P.; Hu, B.; Lou, J.; Su, C.; Zhang, Z.; Xie, L.; Fu, D.; Zhang, Y. A nickel(ii)-based one-dimensional organic-inorganic halide perovskite ferroelectric with the highest Curie temperature. *Chem. Sci.* **2023**, *14*, 1781–1786. [[CrossRef](#)] [[PubMed](#)]
2. Li, C.; Loh, K.P.; Leng, K. Organic-inorganic hybrid perovskites and their heterostructures. *Matter* **2022**, *5*, 4153–4169. [[CrossRef](#)]
3. Wang, G.; Liao, L.; Elseman, A.M.; Yao, Y.; Lin, C.; Hu, W.; Liu, D.; Xu, C.; Zhou, G.; Li, P.; et al. An internally photoemitted hot carrier solar cell based on organic-inorganic perovskite. *Nano Energy* **2020**, *68*, 104383. [[CrossRef](#)]
4. Zhang, X.; Wei, M.; Qin, W. Magneto-open-circuit voltage in organic-inorganic halide perovskite solar cells. *Appl. Phys. Lett.* **2019**, *114*, 033302. [[CrossRef](#)]
5. Leijtens, T.; Eperon, G.E.; Pathak, S.; Abate, A.; Lee, M.M.; Snaith, H.J. Overcoming ultraviolet light instability of sensitized TiO₂ with meso-superstructured organometal tri-halide perovskite solar cells. *Nat. Commun.* **2013**, *4*, 2885. [[CrossRef](#)] [[PubMed](#)]
6. Kahandal, S.S.; Tupke, R.S.; Bobade, D.S.; Kim, H.; Piao, G.; Sankapal, B.R.; Said, Z.; Pagar, B.P.; Pawar, A.C.; Kim, J.M.; et al. Perovskite Solar Cells: Fundamental Aspects, Stability Challenges, and Future Prospects. *Prog. Solid State Chem.* **2024**, *74*, 100463. [[CrossRef](#)]
7. Li, X.; Tang, J.; Zhang, P.; Li, S. Strategies for achieving high efficiency and stability in carbon-based all-inorganic perovskite solar cells. *Cell Rep. Phys. Sci.* **2024**, *5*, 101842. [[CrossRef](#)]
8. Zhang, D.; Zheng, L.L.; Ma, Y.Z.; Wang, S.F.; Bian, Z.Q.; Huang, C.H.; Gong, Q.H.; Xiao, L.X. Factors influencing the stability of perovskite solar cells. *Acta Phys. Sin.* **2015**, *64*, 038803. [[CrossRef](#)]
9. Cakir, H.Y.; Yalcinkaya, Y.; Demir, M.M. Ligand engineering for improving the stability and optical properties of CsPbI₃ perovskite nanocrystals. *Opt. Mater.* **2024**, *152*, 115420. [[CrossRef](#)]
10. Dutta, A.; Pradhan, N. Phase-stable red-emitting CsPbI₃ nanocrystals: Successes and challenges. *ACS Energy Lett.* **2019**, *4*, 709–719. [[CrossRef](#)]
11. Jordan, M.I.; Mitchell, T.M. Machine learning: Trends, perspectives, and prospects. *Science* **2015**, *349*, 255–260. [[CrossRef](#)]
12. Bartel, C.J.; Sutton, C.; Goldsmith, B.R.; Ouyang, R.; Musgrave, C.B.; Ghiringhelli, L.M.; Scheffler, M. New tolerance factor to predict the stability of perovskite oxides and halides. *Sci. Adv.* **2019**, *5*, eaav0639. [[CrossRef](#)] [[PubMed](#)]
13. Li, W.; Jacobs, R.; Morgan, D. Predicting the thermodynamic stability of perovskite oxides using machine learning models. *Comput. Mater. Sci.* **2018**, *150*, 454–463. [[CrossRef](#)]
14. Tao, Q.; Xu, P.; Li, M.; Lu, W. Machine learning for perovskite materials design and discovery. *NPJ Comput. Mater.* **2021**, *7*, 23. [[CrossRef](#)]

15. Sun, Q.; Yin, W. Thermodynamic stability trend of cubic perovskites. *J. Am. Chem. Soc.* **2017**, *139*, 14905–14908. [[CrossRef](#)]
16. Maddah, H.A.; Berry, V.; Behura, S.K. Cuboctahedral stability in Titanium halide perovskites via machine learning. *Comput. Mater. Sci.* **2020**, *173*, 109415. [[CrossRef](#)]
17. Jain, A.; Ong, S.P.; Hautier, G.; Chen, W.; Richards, W.D.; Dacek, S.; Cholia, S.; Gunter, D.; Skinner, D.; Ceder, G.; et al. Commentary: The Materials Project: A materials genome approach to accelerating materials innovation. *APL Mater.* **2013**, *1*, 011002. [[CrossRef](#)]
18. Zhong, H.; Wu, Y.; Li, X.; Shi, T. Machine learning and DFT coupling: A powerful approach to explore organic amine catalysts for ring-opening polymerization reaction. *Chem. Eng. Sci.* **2024**, *292*, 119955. [[CrossRef](#)]
19. Chen, M.; Yin, Z.; Shan, Z.; Zheng, X.; Liu, L.; Dai, Z.; Zhang, J.; Liu, S.; Xu, Z. Application of machine learning in perovskite materials and devices: A review. *J. Energy Chem.* **2024**, *94*, 254–272. [[CrossRef](#)]
20. Mishra, S.; Boro, B.; Bansal, N.K.; Singh, T. Machine learning-assisted design of wide bandgap perovskite materials for high-efficiency indoor photovoltaic applications. *Mater. Today Commun.* **2023**, *35*, 106376. [[CrossRef](#)]
21. Allam, O.; Cho, B.W.; Kim, K.C.; Jang, S.S. Application of DFT-based machine learning for developing molecular electrode materials in Li-ion batteries. *RSC Adv.* **2018**, *8*, 39414–39420. [[CrossRef](#)] [[PubMed](#)]
22. Qin, C.; Liu, J.; Yu, Y.; Xu, Z.; Du, J.; Jiang, G.; Zhao, L. Prediction of thermodynamic stability of actinide compounds by machine learning model. *Ceram. Int.* **2024**, *50*, 1220–1230. [[CrossRef](#)]
23. Chibani, S.; Coudert, F.X. Machine learning approaches for the prediction of materials properties. *APL Mater.* **2020**, *8*, 080701. [[CrossRef](#)]
24. Liu, R.; Kumar, A.; Chen, Z.; Agrawal, A.; Sundararaghavan, V.; Choudhary, A. A predictive machine learning approach for microstructure optimization and materials design. *Sci. Rep.* **2015**, *5*, 11551. [[CrossRef](#)] [[PubMed](#)]
25. Sun, Y.; Sun, P.; Jia, J.; Liu, Z.; Huo, L.; Zhao, L.; Zhao, Y.; Niu, W.; Yao, Z. Machine learning in clarifying complex relationships: Biochar preparation procedures and capacitance characteristics. *Chem. Eng. J.* **2024**, *485*, 149975. [[CrossRef](#)]
26. Liu, Z.; Rolston, N.; Flick, A.C.; Colburn, T.W.; Ren, Z.; Dauskardt, R.H.; Buonassisi, T. Machine learning with knowledge constraints for process optimization of open-air perovskite solar cell manufacturing. *Joule* **2022**, *6*, 834–849. [[CrossRef](#)]
27. Zhao, J.; Wang, X.; Li, H.; Xu, X. Interpretable machine learning-assisted screening of perovskite oxides. *RSC Adv.* **2024**, *14*, 3909–3922. [[CrossRef](#)] [[PubMed](#)]
28. Liu, H.; Cheng, J.; Dong, H.; Feng, J.; Pang, B.; Tian, Z.; Ma, S.; Xia, F.; Zheng, C.; Dong, L. Screening stable and metastable ABO₃ perovskites using machine learning and the materials project. *Comput. Mater. Sci.* **2020**, *177*, 109614. [[CrossRef](#)]
29. Talapatra, A.; Uberuaga, B.P.; Stanek, C.R.; Pilania, G. A Machine Learning Approach for the Prediction of Formability and Thermodynamic Stability of Single and Double Perovskite Oxides. *Chem. Mater.* **2021**, *33*, 845–858. [[CrossRef](#)]
30. Zhu, Y.; Zhang, J.; Qu, Z.; Jiang, S.; Liu, Y.; Wu, Z.; Yang, F.; Hu, W.; Xu, Z.; Dai, Y. Accelerating stability of ABX₃ perovskites analysis with machine learning. *Ceram. Int.* **2024**, *50*, 6250–6258. [[CrossRef](#)]
31. Bhattacharya, S.; Roy, A. Linking stability with molecular geometries of perovskites and lanthanide richness using machine learning methods. *Comput. Mater. Sci.* **2024**, *231*, 112581. [[CrossRef](#)]
32. Wu, T.; Wang, J. Global discovery of stable and non-toxic hybrid organic-inorganic perovskites for photovoltaic systems by combining machine learning method with first principle calculations. *Nano Energy* **2019**, *66*, 104070. [[CrossRef](#)]
33. Zhu, Q.; Xu, P.; Lu, T.; Ji, X.; Shao, M.; Duan, Z.; Lu, W. Discovery and verification of two-dimensional organic-inorganic hybrid perovskites via diagrammatic machine learning model. *Mater. Des.* **2024**, *238*, 112642. [[CrossRef](#)]
34. Cai, X.; Zhang, Y.; Shi, Z.; Chen, Y.; Xia, Y.; Yu, A.; Xu, Y.; Xie, F.; Shao, H.; Zhu, H.; et al. Solar cells via machine learning: Ultrabroadband absorption, low radiative combination, and enhanced thermal conductivities. *Adv. Sci.* **2021**, *9*, 2103648. [[CrossRef](#)]
35. Li, J.; Cheng, K.; Wang, S.; Morstatter, F.; Trevino, R.P.; Tang, J.; Liu, H. Feature selection: A data perspective. *ACM Comput. Surv.* **2017**, *50*, 1–45. [[CrossRef](#)]
36. Shen, Y.; Wang, J.; Ji, X.; Lu, W. Machine Learning-Assisted Discovery of 2D Perovskites with Tailored Bandgap for Solar Cells. *Adv. Theory Simul.* **2023**, *6*, 2200922. [[CrossRef](#)]
37. Marchenko, E.I.; Fateev, S.A.; Petrov, A.A.; Korolev, V.V.; Mitrofanov, A.; Petrov, A.V.; Goodilin, E.A.; Tarasov, A.B. Database of Two-Dimensional Hybrid Perovskite Materials: OpenAccess Collection of Crystal Structures, Band Gaps, and Atomic Partial Charges Predicted by Machine Learning. *Chem. Mater.* **2020**, *32*, 7383–7388. [[CrossRef](#)]
38. Teimourian, A.; Rohacs, D.; Dimililer, K.; Teimourian, H.; Yildiz, M.; Kale, U. Airfoil aerodynamic performance prediction using machine learning and surrogate modeling. *Heliyon* **2024**, *10*, e29377. [[CrossRef](#)] [[PubMed](#)]
39. Hu, J.; Chen, Z.; Chen, Y.; Liu, H.; Li, W.; Wang, Y.; Peng, L.; Liu, X.; Lin, J.; Chen, X.; et al. Interpretable machine learning predictions for efficient perovskite solar cell development. *Sol. Energy Mater. Sol. Cells* **2024**, *271*, 112826. [[CrossRef](#)]
40. Wang, Z.; Mu, L.; Miao, H.; Shang, Y.; Yin, H.; Dong, M. An innovative application of machine learning in prediction of the syngas properties of biomass chemical looping gasification based on extra trees regression algorithm. *Energy* **2023**, *275*, 127438. [[CrossRef](#)]
41. Liu, Y.; Wang, L.; Gu, K. A support vector regression (SVR)-based method for dynamic load identification using heterogeneous responses under interval uncertainties. *Appl. Soft Comput.* **2021**, *110*, 107599. [[CrossRef](#)]
42. Zhu, Z.; Zhou, M.; Hu, F.; Wang, S.; Ma, J.; Gao, B.; Bian, K.; Lai, W. A day-ahead industrial load forecasting model using load change rate features and combining FA-ELM and the AdaBoost algorithm. *Energy Rep.* **2023**, *9*, 971–981. [[CrossRef](#)]
43. Zhang, Z.; Zhang, Y.; Liu, S. Integrative approach of machine learning and symbolic regression for stability prediction of multicomponent perovskite oxides and high-throughput screening. *Comput. Mater. Sci.* **2024**, *236*, 112889. [[CrossRef](#)]

44. Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T.Y. Lightgbm: A highly efficient gradient boosting decision tree. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 3146–3154.
45. Anand, D.V.; Xu, Q.; Wee, J.; Xia, K.; Sum, T.C. Topological feature engineering for machine learning based halide perovskite materials design. *NPJ Comput. Mater.* **2022**, *8*, 203. [[CrossRef](#)]
46. Djeradi, S.; Dahame, T.; Fadla, M.A.; Bentría, B.; Kanoun, M.B.; Goumri-Said, S. High-Throughput Ensemble-Learning- Driven Band Gap Prediction of Double Perovskites Solar Cells Absorber. *Mach. Learn. Knowl. Extr.* **2024**, *6*, 435–447. [[CrossRef](#)]
47. Ahmed, U.; Mahmood, A.; Tunio, M.A.; Hafeez, G.; Khan, A.R.; Razzaq, S. Investigating boosting techniques' efficacy in feature selection: A comparative analysis. *Energy Rep.* **2024**, *11*, 3521–3532. [[CrossRef](#)]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.