


Review

# Feature Selection in Machine Learning for Perovskite Materials Design and Discovery

Junya Wang <sup>1</sup>, Pengcheng Xu <sup>2</sup>, Xiaobo Ji <sup>3</sup>, Minjie Li <sup>3,\*</sup> and Wencong Lu <sup>3,4,5,\*</sup> <sup>1</sup> Department of Mathematics, College of Sciences, Shanghai University, Shanghai 200444, China<sup>2</sup> Materials Genome Institute, Shanghai University, Shanghai 200444, China<sup>3</sup> Department of Chemistry, College of Sciences, Shanghai University, Shanghai 200444, China<sup>4</sup> Zhejiang Laboratory, Hangzhou 311100, China<sup>5</sup> Key Laboratory of Silicate Cultural Relics Conservation (Shanghai University), Ministry of Education, Shanghai 200444, China

\* Correspondence: minjieli@shu.edu.cn (M.L.); wclu@shu.edu.cn (W.L.)

**Abstract:** Perovskite materials have been one of the most important research objects in materials science due to their excellent photoelectric properties as well as correspondingly complex structures. Machine learning (ML) methods have been playing an important role in the design and discovery of perovskite materials, while feature selection as a dimensionality reduction method has occupied a crucial position in the ML workflow. In this review, we introduced the recent advances in the applications of feature selection in perovskite materials. First, the development tendency of publications about ML in perovskite materials was analyzed, and the ML workflow for materials was summarized. Then the commonly used feature selection methods were briefly introduced, and the applications of feature selection in inorganic perovskites, hybrid organic-inorganic perovskites (HOIPs), and double perovskites (DPs) were reviewed. Finally, we put forward some directions for the future development of feature selection in machine learning for perovskite material design.

**Keywords:** perovskites; materials design; machine learning; feature selection



**Citation:** Wang, J.; Xu, P.; Ji, X.; Li, M.; Lu, W. Feature Selection in Machine Learning for Perovskite Materials Design and Discovery. *Materials* **2023**, *16*, 3134. <https://doi.org/10.3390/ma16083134>

Academic Editor: Lucia Nasi

Received: 19 March 2023

Revised: 11 April 2023

Accepted: 13 April 2023

Published: 16 April 2023



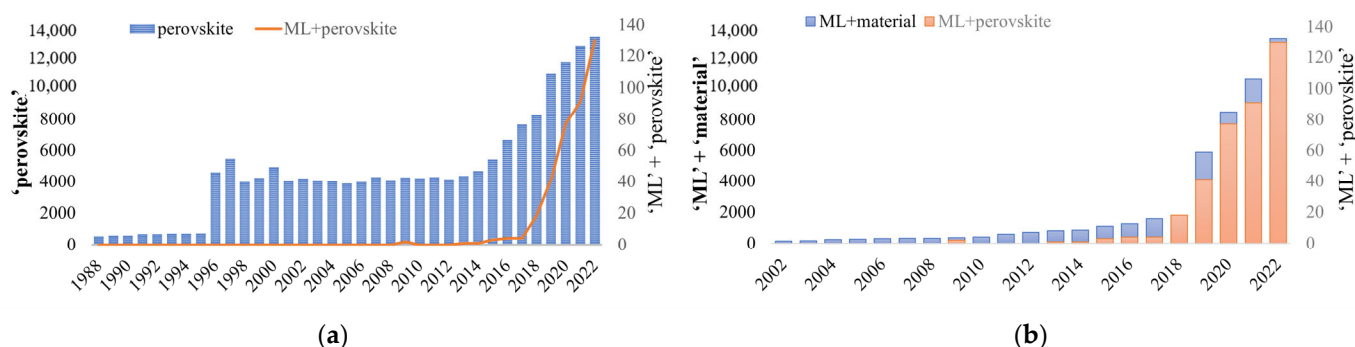
**Copyright:** © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

Machine learning (ML), as an interdisciplinary technique covering multiple fields of statistics, computer science, and mathematics, has been widely used in the medical, bioinformatics, financial, and agriculture fields [1–5]. Especially in the materials field, ML technology has accelerated the design and discovery of new materials by constructing models for the prediction of their properties [6,7]. In recent years, perovskite materials have drawn the attention of many scholars due to their excellent properties, such as excellent electrical conductivity, ferroelectricity, superconductivity, longer carrier diffusion lengths, a tunable bandgap ( $E_g$ ), and high light absorption that can be applied in solar cells, light-emitting diodes, lasers, and photocatalysis materials fields [8–11]. Figure 1a demonstrates the overall growth pattern in the number of papers searched on the website ‘web of science’ with the key words ‘perovskite’ and ‘machine learning and perovskite’ from 1988 to 2022. Especially since 2013, after the breakthrough in the applications of perovskite materials in solar cells, there has been a spurt of research results, indicating that perovskite materials have always been a research hotspot. Figure 1b shows that ML technology has become a powerful tool in materials science in recent years, and its applications in the field of perovskite materials have been increasing year by year since 2013, indicating that ML has played an increasingly important role in the research of perovskite materials.

Data is the cornerstone of ML, and high-quality data allows ML to capture the hidden patterns in the data to make the correct predictions about the research objects. The data of materials are generally divided into target variables reflecting material properties and features associated with the target variables, which can also be described as variables,

descriptors, or fingerprints in the material field. For perovskite materials, property data such as  $E_g$ , formability, thermodynamic stability, specific surface area (SSA), and Curie temperature ( $T_c$ ) are commonly employed as target variables, and the associated features usually involve elemental components, atomic parameters, structural parameters, experimental parameters, etc. [12–15], which usually have the characteristics of high feature dimensionality. The high feature dimensionality would not only lead to limitations due to overfitting and computational inefficiency but also cause difficulty in exploring the physical meaning of features [16,17]. Thus, it is crucial to pick an appropriate method for reducing feature dimensionality. The two commonly used methods for dimensionality reduction are feature extraction and feature selection [18]. Feature extraction transforms the feature space by transformation or mapping, thus effectively reducing the dimensionality of features [19]. Feature selection preserves the original information of features by selecting a valid subset from the original feature set and removing redundant and irrelevant features. Feature extraction may generally lack interpretability, while feature selection methods are numerous. Therefore, it is necessary to select an appropriate feature selection method to approximate the upper limit of the performance of the trained model as much as possible. Reviewing the progress of feature selection methods in ML for perovskite materials and providing an outlook on future work will help further the development of perovskite material design.



**Figure 1.** Number of published papers. (a) On the key words 'perovskite' and 'machine learning and perovskite' (from 1988 to 2022). (b) On the key words 'machine learning and material' and 'machine learning and perovskite' (from 2002 to 2022).

In this review, we discuss the applications and importance of feature selection in the ML workflow for perovskite materials. In Section 2, the basic workflow of ML in the field of materials science is outlined. In Section 3, we present the different types of perovskite materials and their associated features. Section 4 is an introduction to feature selection methods, including filter, wrapper, and embedded. In Section 5, the applications of feature selection methods in the study of inorganic perovskites, hybrid organic-inorganic perovskites (HOIPs), and double perovskites (DPs) are introduced. In Section 6, some of the current challenges and opportunities encountered in the applications of feature selection in ML to perovskite design and discovery are briefly discussed. Our work will help researchers better deal with the feature selection problems involved when using ML as a tool to study perovskite materials.

## 2. Workflow of Materials Machine Learning

As shown in Figure 2, the workflow of ML in materials could be divided into four steps: data preparation, feature engineering, model evaluation and selection, and model application [20].

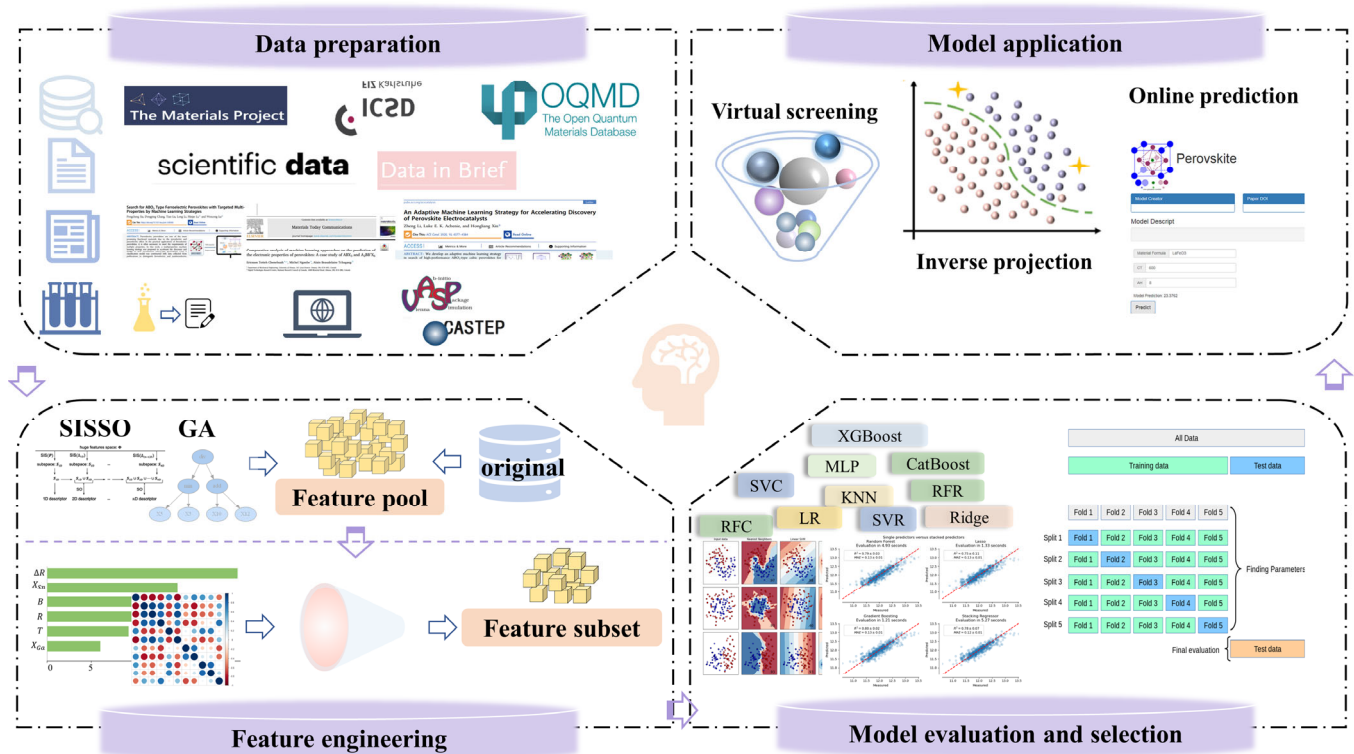


Figure 2. The general workflow of materials ML.

Data preparation includes data collection and data preprocessing. Materials data could generally be obtained through publicly available materials databases, published papers, experimental data of the same standard, data journals, and density functional theory (DFT) calculations [13,21–24]. The latest data can be obtained by searching the literature, but it is time-consuming and laborious. Data from data journals and databases can be obtained in a short time, but the latest data are generally not available in a timely manner. *Scientific Data* by Springer Nature and *Data in Brief* by Elsevier are the more representative data journals. Table 1 lists the commonly used material databases, including perovskites. Experimental data may be a good source of data, but it is costly. DFT calculations are susceptible to material systems, which may lead to the doubling of time and computing resources. Data preprocessing is essential due to the characteristics of multi-source data and the high noise of the material data. To ensure the availability of data, common preprocessing operations include filling in missing values, removing duplicates and outliers, dimensionless processing, treating data imbalances, and rationally dividing data [25,26].

Feature engineering, including feature construction and feature selection, is an extremely important part of the ML workflow. In most ML processes, the quality of the data related to the sample size and feature dimensionality, as well as the validity of the features, determines the upper limit of the model's performance. In general, a high ratio of sample size to feature dimension would lead to better model performance. When the existing features do not contain enough valid information to cause low model performance, new features can be either constructed based on domain knowledge or generated by simple mathematical transformation of existing features through algorithms such as the Sure Independence Screening Sparsifying Operator (SISSO) and genetic algorithm (GA) to improve model performance [27,28]. The properties of materials are influenced by their composition, structure, experimental conditions, and environmental factors, but there may be weakly correlated, uncorrelated, or redundant features in the data. For the original set of features in the data, feature selection can remove the redundant features and keep the key features that are easily accessible and have a significant impact on the target variable to further improve the model's performance while increasing the computational efficiency.

**Table 1.** Commonly used materials databases, including perovskites.

Name	URL	Data Type
The Perovskite Database Project (PDP)	<a href="https://www.perovskitedatabase.com">https://www.perovskitedatabase.com</a> (accessed on 19 March 2023)	Exp.
Open Quantum Materials Database (OQMD)	<a href="http://www.oqmd.org/">http://www.oqmd.org/</a> (accessed on 19 March 2023)	Comp.
Materials Project (MP)	<a href="https://materialsproject.org/">https://materialsproject.org/</a> (accessed on 19 March 2023)	Comp.
Computational Materials Repository (CMR)	<a href="https://cmr.fysik.dtu.dk/">https://cmr.fysik.dtu.dk/</a> (accessed on 19 March 2023)	Comp.
The Inorganic Crystal Structure Database (ICSD)	<a href="https://icsd.fiz-karlsruhe.de/index.xhtml">https://icsd.fiz-karlsruhe.de/index.xhtml</a> (accessed on 19 March 2023)	Exp.
Materials Platform for Data Science (MPDS)	<a href="https://mpds.io/#modal/menu">https://mpds.io/#modal/menu</a> (accessed on 19 March 2023)	Comp. and Exp.
Automatic-FLOW for Materials Discovery (AFLOW)	<a href="http://www.aflowlib.org/">http://www.aflowlib.org/</a> (accessed on 19 March 2023)	Comp.

Before building models, it is necessary to confirm the type of models corresponding to classification and regression models when the target variables are discrete and continuous, respectively. There are many ML algorithms, but no perfect algorithm exists. Although for a specific classification or regression task, the researchers could choose linear, nonlinear, or ensemble algorithms preliminary based on their understanding or guessing of the potential “structure-property relationship” of the materials. It is still difficult to determine the most suitable algorithm based on the limited data volume. Even with the same data and algorithm, the trained model will not be the same with the different hyperparameters. Therefore, it is necessary to evaluate a series of models to select the relatively optimal one. Model performance and model complexity are the key factors that determine model selection. Model performance can be measured by evaluation metrics calculated based on the true and predicted values of the target variable. Common evaluation metrics for regression tasks include coefficient of determination ( $R^2$ ), correlation coefficient (R), mean square error (MSE), root mean squared error (RMSE), mean absolute error (MAE), and average relative error (MRE), while common evaluation metrics for classification tasks include accuracy (ACC), area under the curve (AUC), recall, precision, and  $F_1$  score. To ensure the reliability of the results, the hold-out method and cross-validation method are generally used to evaluate the models after the evaluation metrics are determined. Common methods of cross validation are 5-fold cross validation (5-fold CV), 10-fold cross validation (10-fold CV), and leave-one-out cross validation (LOOCV). Furthermore, we tend to choose the model with better performance and lower model complexity. After selecting a specific ML algorithm, hyperparameter optimization is usually performed to further improve the performance of the model, and the final model is determined after the determination of hyperparameters. Contemporary hyperparametric optimization algorithms can be mainly classified into various types, including grid-search, Bayesian-based optimization algorithms, gradient-based optimization, and population-based optimization.

The final aim of ML is to predict the target variables of unknown samples based on the trained model. The three major scenarios of model application are high-throughput screening (HTS), inverse design, and the development of online prediction programs. HTS uses the constructed model to predict the target variables of a huge number of virtual samples in order to filter out samples with high performance potential and guide experimental synthesis [29,30]. The inverse design can be used to obtain the features of designed samples via the inverse projection method, which is an effective way to realize

the material from properties to composition [31,32]. The prediction of designed samples helps screen out candidates with breakthrough performance and improves computational efficiency. The development of an online prediction program makes it possible to quickly achieve the prediction of target properties by simply inputting the necessary information, such as a chemical formula, on the input page, which facilitates the extension of model application and effectively realizes model sharing [33].

### 3. The Structure and Features of Perovskite

Named after Russian geologist Perovski, perovskite can be divided into narrow sense perovskite, referring to the specific compound  $\text{CaTiO}_3$ , and broad sense perovskite, referring to the  $\text{ABX}_3$ -type compound with the same structure as the compound  $\text{CaTiO}_3$  [34]. The cations at A-site and B-site can be replaced by ions with approximate radii or certain groups due to the tunable ionic structure of perovskite materials, leading to the emergence of a lot of perovskite derivatives. The common perovskites generally can be subdivided into inorganic perovskites, HOIPs, and DPs [8]. The  $\text{ABX}_3$  inorganic perovskites have been widely used in solar cells, solid oxide fuel cells, magnetic refrigeration, and photocatalysis for their multiple properties such as catalytic activity, strong flexibility, outstanding stability, and low cost [8,35]. The HOIPs have been widely applied in the fields of solar cells, light-emitting diodes, X-ray or  $\gamma$ -ray detectors, lasers, and photodetectors due to their longer charge diffusion lengths, high absorption coefficients, high defect tolerance, high carrier mobility, and tunable  $E_g$  [36–38]. Because of the adjustable photoelectric performance and good stability, the DPs have demonstrated promising applications in photocatalysis as well as in functional devices including solar cells, light-emitting diodes, scintillators, and photodetectors [39,40].

#### 3.1. Inorganic Perovskites

For  $\text{ABX}_3$ -type inorganic perovskites, A-site and B-site are cations of alkaline earth or rare earth metal with a larger radius and transition metal with a smaller radius, respectively, and X is usually an anion of oxygen or halogen [41]. As shown in Figure 3a, the ideal structure of  $\text{ABX}_3$  perovskites has cubic symmetry with space group  $\text{Pm}\bar{3}\text{m}$ , and the cations at the A-site and B-site are coordinated to the X-site anion via 12 and 6, respectively [42].  $\text{ABX}_3$  inorganic perovskites can be further divided into oxide perovskites and halide perovskites when X refers to oxygen ions and halide ions, respectively. The  $\text{ABO}_3$  perovskite oxides are one of the most common and widely studied structures in materials. Given that not all compounds with  $\text{ABX}_3$  stoichiometry are necessarily perovskite materials, geometric structural features such as the octahedral factor ( $\mu$ ), Goldschmidt's tolerance factor ( $t$ ), and a modified tolerance factor ( $\tau$ ) (Equations (1)–(3)) are used in the study of perovskite materials by ML for the determination of perovskite formability and stability [43–45]. In addition, the structural features of A-X and B-X bond lengths based on bond valence have also been used to indicate the formability and stability of inorganic perovskites [46]. For  $\text{ABX}_3$ -type inorganic perovskites, the features are generally dominated by atomic parameters indicating the properties of the elements in the A/B sites, such as atomic radius, electronegativity, ionization energy, highest occupied molecular orbital (HOMO) energy, and lowest unoccupied molecular orbital (LUMO) energy, etc. It is worth noting that when the elements at the A-site or B-site of the  $\text{ABX}_3$  perovskites are doped, the general formula can be expressed as  $\text{A}_{1-x}\text{A}'_x\text{B}_{1-y}\text{B}'_y\text{X}_3$  in which the features of the A/B positions are generally calculated by taking a weighted average of the properties of the doped elements at the corresponding positions (Equations (4) and (5)) [47,48]. Commonly used atomic parameters are publicly available from the Villars database [49], Mendeleev package [50], and RDKit [51] and can also be obtained by direct populating through online calculation platforms or software [33]. The models based on 21 features including structural and atomic parameters of the materials populated by the OCPMDM platform developed in our laboratory have yielded good results in the prediction of target variables such as SSA and  $E_g$  of  $\text{ABO}_3$ -type perovskite materials [13]. In addition, the SISO method can be used to generate new key features



based on features that are directly accessible. Equation (3) is a new tolerance factor obtained by Bartel et al. based on the SISSO method of ionic radii, which has an ACC of 92% in determining the formability and stability of  $ABX_3$  perovskites [45].

$$\mu = \frac{r_B}{r_X}. \quad (1)$$

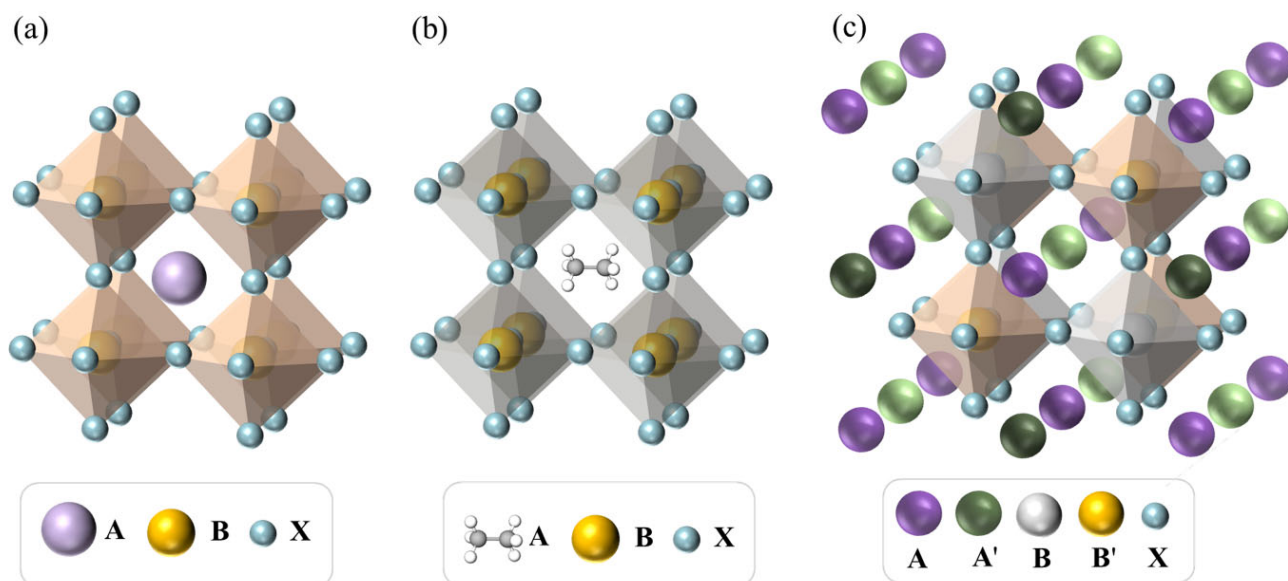
$$t = \frac{r_A + r_X}{\sqrt{2}(r_B + r_X)}. \quad (2)$$

$$\tau = \frac{r_X}{r_B} - n_A \left( n_A - \frac{\frac{r_A}{r_B}}{\ln\left(\frac{r_A}{r_B}\right)} \right). \quad (3)$$

$$f_{\bar{A}} = (1 - x) * f_A + x * f_{A'}. \quad (4)$$

$$f_{\bar{B}} = (1 - y) * f_B + y * f_{B'}. \quad (5)$$

where  $r_A$ ,  $r_B$ , and  $r_X$  are the ionic radii of  $ABX_3$  perovskites, respectively;  $n_A$  is the oxidation state of the A-site ion;  $(1 - x)$  and  $x$  are the percentages of A-site doped elements,  $(1 - y)$  and  $y$  are the percentages of B-site doped elements;  $f_A$  and  $f_{A'}$  are the respective features of A-site doped elements, and  $f_{\bar{A}}$  is the weighted average feature of the A-site;  $f_B$  and  $f_{B'}$  are the respective features of the B-site doped elements, and  $f_{\bar{B}}$  is the weighted average feature of the B-site.



**Figure 3.** Crystal structures of different perovskites. (a)  $ABX_3$ , an inorganic perovskite structure. (b)  $ABX_3$ , a hybrid organic-inorganic perovskite structure. (c)  $AA'BB'X_6$ , double perovskite structure.

### 3.2. Hybrid Organic-Inorganic Peroxites

As shown in Figure 3b, the A-site of  $ABX_3$ -type HOIPs could be replaced by an organic cation such as methylammonium  $CH_3NH_3^+$  or formamidinium  $CH(NH_2)_2^+$  compared to  $ABX_3$  inorganic perovskites [52]. The features involved in the inorganic part of the HOIPs are still dominated by atomic parameters, but the organic molecular features have few parts in common with the atomic features due to the complexity of the organic cation at the A-site, which requires additional calculations of the features of the organic structure [53,54]. The basic properties of an A-site ion, such as its first and second ionization energies, electron affinity, molecular volume, molecular radius, and chemical potential, can be estimated

based on theoretical methods, and Multiwfn and Gaussian are commonly used calculation software [31,53]. In addition to using the radius of organic ions as a feature, the anisotropy of organic cations can also be considered. Chen et al. improved the ACC of  $E_g$  models by using three geometric parameters, namely length, width, and height, as features [55].

### 3.3. Double Perovskites

The structural general formula of DPs could be expressed as  $AA'BB'X_6$ , where A and A' are more commonly the same or different cations, and B and B' are different cations that alternate with the X site ions to form the  $BX_6$  and  $B'X_6$  octahedrons (Figure 3c) [56–58]. Similarly, not all materials satisfying the chemical formula  $AA'BB'X_6$  are perovskite structures. The tolerance factor  $t$  is proposed for single perovskite materials, but the formability of perovskites is essentially all based on geometric criteria derived from ion radii or bond distances. By using arithmetic or weighted averages of ion radii or bond lengths, the concept of tolerance factors can be extended to DPs with more complex compositions [12]. The generalized octahedral factor has also been introduced as a judgment of perovskite formability [39]. For DPs, the common features are similar to those of the  $ABX_3$  type, which are generally based on atomic parameters. And similar to single perovskites containing doped elements, the features of the A/B sites can be treated by common methods including arithmetic averaging and geometric averaging [59]. It is also noteworthy that  $AA'BB'X_6$ ,  $A'ABB'X_6$ , and  $A'AB'BX_6$  are all unified systems because the exchange of two A-site cations as well as two B-site cations does not affect the structure of perovskite. The features are treated symmetrically when considering the inclusion of structural symmetry into the model [56].

Furthermore, experimental conditions are also quite important features, and the gradient boosting regression (GBR) model for the crystallite size (CS) of  $ABO_3$  perovskite materials developed by Tao et al. indicates the high importance of two experimental conditions: the preparation method (PM) and the calcination temperature (CT) [15]. If possible, it is encouraged to use the experimental conditions as features to build predictive models for the target properties.

Notably, perovskite materials are widely used in solar cells and photodetectors in the form of thin films [36]. Especially in the field of solar cells, the power conversion efficiency (PCE) of perovskite solar cells (PSCs) has surpassed 25% within just 10 years, which is comparable to crystalline silicon solar cells [60]. Research has revealed that high-quality thin films are one of the crucial factors influencing the performance of PSCs. The methods to fabricate perovskite films include several techniques such as solution processing, vacuum deposition, physical vapor deposition, vapor-assisted solution processing, and scalable deposition [61–66]. Thin film properties such as grain size, morphology, crystallinity, defect density, and surface coverage may vary under different preparation methods, leading to differences in the quality of the thin film [61]. Experiments have shown that various process parameters such as stoichiometry, thermal treatment, substrate temperature, solvent engineering, additives, and environmental control have a great influence on the quality of perovskite thin films [61–66].

## 4. The Methods of Feature Selection

According to whether the evaluation criteria are independent of the learning algorithm, the feature selection methods could be generally classified into filter, wrapper, and embedded [67,68]. The filter methods are independent of the ML algorithm, using an evaluation criterion based on statistical theory or information theory to select a subset of features after ranking the features [19,69]. In the process of feature selection, the wrapper methods use the performance of the evaluator as the criterion to select the optimal feature subset [70]. The embedded methods can be used to realize feature selection in the modeling by combining the training of the evaluator and the processes of feature selection into a single optimization process [17]. The filter methods are computationally efficient and generalize well. However, due to the lack of interaction with the evaluator, the model performance of feature subsets selected based on the filter methods is generally less effec-

tive than the wrapper and embedded methods, which are relatively less computationally efficient [71].

#### 4.1. Filter

The filter feature selection methods include the chi-square test ( $\chi^2$ ), analysis of variance (ANOVA), Pearson correlation coefficient (PCC), distance correlation coefficient (DCC), max-relevance and min-redundancy (mRMR), maximal information coefficient (MIC), and Relief, etc.

The  $\chi^2$  and ANOVA are correlation measure methods based on hypothesis testing, with the former for testing the independence between discrete variables and the latter for testing the independence between discrete and continuous variables [72,73]. Hypothesis testing generally includes four steps: (1) proposing the null hypothesis and alternative hypothesis; (2) designing the hypothesis testing statistic according to the hypothesis; (3) getting the  $p$ -value according to the distribution after calculating the current value of the statistic; and (4) considering the acceptance or overturning of the null hypothesis according to the  $p$ -value and drawing the final conclusion. The smaller the  $p$ -value of the output, the smaller the probability that the null hypothesis holds, and the more likely it is that the two features are not independent. Features with significant associations can be screened out when the  $p$ -value is less than  $\alpha$  referring to the significance level. It is worth noting that, generally, the smaller the  $p$ -value usually means the larger the value of the statistic, which can be equated to the feature score. In specific usage scenarios, the user can select features based on the ranking of features according to the value of the statistic [74].

PCC generally measures the linear correlation between continuous variable pairs  $(x, y)$  by Equation (6), where  $n$  is the number of samples in the dataset,  $x_i$  and  $y_i$  are the  $i$ th sample point, and  $\bar{x}$  and  $\bar{y}$  are the means of the samples [75,76]. The range of PCC values from  $-1$  to  $1$  indicates that the relationship between variables changes from a completely negative correlation to a completely positive correlation. Additionally, the closer the PCC is to zero, the weaker the linear correlation will be. In a practical ML process, PCC can indicate the linear correlation between the target variable and features to represent the degree of association and a linear correlation between any feature pairs to represent the redundancy among feature pairs:

$$PCC = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}. \quad (6)$$

The accuracy of PCC may not be guaranteed when there is a nonlinear correlation between the variables. The DCC is an alternate correlation coefficient that does not have this weakness, which defines the independence between variables:  $dCor(x, y) = 0$  if and only if  $x$  and  $y$  are independent, where  $dCov(x, y)$  is the sample distance covariance (Equation (7)) [77]. The DCC takes a value in the range  $[0, 1]$ ; the larger the value, the stronger the correlation:

$$dCor(x, y) = \begin{cases} \frac{dCov(x, y)}{\sqrt{dCov(x) dCov(y)}}, & dCov(x) dCov(y) > 0 \\ 0, & dCov(x) dCov(y) = 0 \end{cases}. \quad (7)$$

The measure of correlation based on mutual information is a non-parametric approach, and the essence of mutual information is the extent to which two variables explain each other, which can be understood in terms of the consistency of the distribution and the amount of information contained in each other. Meanwhile, mutual information can identify arbitrary relationships between any type of variable.



mRMR based on mutual information theory attempts to select the features with the maximum relevance to the target variable and the minimum redundancy among the features [78]. It is supposed that there are a total of  $F$  features in the dataset, and  $S_m$  denotes the set of  $m$  features that have been selected; the importance of the  $(m + 1)$ th feature is defined in Equation (8), where  $I(x_j, y)$  is the mutual information between variables  $x_j$  and  $y$ . Additionally, the mutual information of any variable pair  $(x, y)$  could be calculated by Equation (9), where  $p(x)$ ,  $p(y)$ , and  $p(x, y)$  are their probabilistic density functions. Then the scoring function  $\max_{x_j \in (F - S_m)} [f^{mRMR}(x_j)]$  can be used to select the  $(m + 1)$ th feature from the remaining set of features  $(F - S_m)$  to join  $S_m$ . Therefore, mRMR is actually a stepwise method where, at each step of the feature selection process, the feature with the highest feature importance will be added to the subset until the number of features in the subset reaches the user requirement:

$$f^{mRMR}(x_j) = I(x_j, y) - \frac{1}{m} \sum_{x_i \in S_m} I(x_j, x_i), \quad (8)$$

$$I(x, y) = \iint p(x, y) \log \frac{p(x, y)}{p(x)p(y)} dx dy. \quad (9)$$

The solution of joint probabilities is often difficult, and MIC overcomes this shortcoming of mutual information. The MIC belonging to the nonparametric method can provide an effective measure of linear and nonlinear relationships between the variables, as well as nonfunctional dependencies [79]. The values of MIC between the features and the target variable are regarded as the scores of each feature in the feature selection process. The features can be ranked based on the sizes of the MIC values, and the features are then chosen based on the threshold value or the predetermined number of features.

Relief is a feature weighting method used to handle binary classification, where features are given different weights according to the relevance of each feature to the category, which is based on the ability of the feature to discriminate between nearby samples, and features with weights less than a certain threshold are removed [69]. According to regression and classification tasks, the ReliefF and RReliefF methods were proposed, which support multi-class classification and regression problems, respectively [69].

#### 4.2. Wrapper

Wrapper methods for feature selection include greedy sequential searches such as sequential feature selection (SFS) and sequential backward selection (SBS), as well as more complex ones like recursive feature elimination (RFE) and evolutionary and swarm intelligence algorithms such as GA [80–82].

The SFS method takes the empty set as the starting point of the search and selects one feature at a time that makes the objective function generally optimal, referring to the cross-validation score of an estimator to join the feature set  $S$ . The SFS selection method is an iterative selection process that involves only adding features. In contrast to SFS, the SBS method starts with the full set of features and then continuously discards features from the feature set to optimize the objective function value. Both methods stop searching when a set number of features is reached.

RFE is a feature selection method based on model performance that continuously removes the least important features through recursion. The basic execution steps of RFE are: (1) training on the current feature set  $S_1$  and calculating the importance of each feature according to the given evaluator; (2) eliminating the least important feature to obtain the feature subset  $S_2$ , and then training the model again to calculate the importance of the remaining features; and (3) repeating step two until the number of features is equal to the value manually set. The recursive feature addition (RFA), the opposite method, iteratively adds features [83]. RFE and RFA are often used in conjunction with the RF algorithm [83,84].

GA as one of the representatives of intelligent algorithms is proposed based on the core idea of biological evolutionary theory, where each solution is encoded as a ‘chromosome’ or an individual to constitute a population (a subset of all possible solutions) when solving a problem [85]. The general steps of GA include: (1) generating an initial population representing potential solutions to the problem randomly; (2) selecting an appropriate fitness function to evaluate individuals; (3) then applying genetic operations such as selection, crossover, and mutation to generate new populations; and (4) repeating steps 2–3 until the termination condition of the iterative calculation is met [86,87]. Binary coding is adopted when using GA to solve the problem of feature selection, where a binary value of ‘1’ indicates that a feature at the corresponding position is selected, so that a genetic individual consisting of a fixed-length binary string represents a subset of features [87]. In other words, the realization of feature selection based on GA is to find an optimal binary code which represents the optimal feature subset.

#### 4.3. Embedded

Embedded methods can be broadly classified into those based on regularized models and those based on tree models. Many ML models introduce regularization terms such as  $L_1$ -penalty or  $L_2$ -penalty in their loss functions to prevent overfitting problems. Regularization terms such as least absolute shrinkage and selection operator (LASSO), ridge regression (RR), and support vector machine (SVM) can effectively shrink the coefficients of certain features to zero, thus enabling feature selection [88–90]. A major branch of ML is tree-based ML models such as random forest (RF), GBR, and extreme gradient boosting (XGBoost), etc. [91–93]. These tree models record how each feature progressively reduces the model error in the bifurcation of the tree nodes during the process of modeling and generally use feature importance to indicate the degree of feature contribution to the current model.

In addition, SHapley Additive exPlanations (SHAP) method, which can be used in nesting with different ML algorithms, serves as a unifying framework for interpreting black box models, and the SHAP value also indicates how much the feature contributes to the model’s prediction. Since global importance is required, the average of the absolute Shapley values for each feature is used as the SHAP feature importance. Then feature selection can be performed after ranking the features according to SHAP feature importance [94].

## 5. Feature Selection in Machine Learning for Perovskite Materials

### 5.1. Feature Selection for Inorganic Perovskites

In the research of inorganic perovskite materials, a single feature selection method was sometimes employed. Priyanga G et al. [95] used ML methods to predict the nature of  $E_g$  of  $ABO_3$  perovskite oxides. Datasets were obtained from various databases and experimental research papers, with the features generated using Matminer. After preprocessing, 5276 samples consisting of ‘direct  $E_g$ ’ and ‘indirect  $E_g$ ’ were obtained to construct the classification model for predicting the nature of  $E_g$  in perovskite materials. The highly correlated features were removed based on the PCC matrix, retaining the six features, including the ionic radius of the A-site ( $R_A$ ), the ionic radius of the B-site ( $R_B$ ), the electronegativity of element A ( $E_{NA}$ ), the electronegativity of element B ( $E_{NB}$ ), the electronegativity difference with radius ( $E_{NR}$ ), and the average ionic character of A and B (avg ionic char [95]). Logistic regression (LR), decision tree (DT), RF, k-nearest neighbors (KNN), light gradient boosting machine (Light GBM), XGBoost, and support vector clustering were used to build the classification models, and the RF model was optimal with an ACC of 91%. A feature importance analysis of the RF model revealed that the most important features in the  $E_g$  classification of perovskite materials are avg ionic char  $E_{NA}$ ,  $E_{NB}$ , and  $E_{NR}$ . Additionally, the tendency to obtain direct  $E_g$  is higher as the average ionic character increases, while the tendency to obtain indirect  $E_g$  increases as the average ionic character decreases. Zhang et al. [96] developed a model for the automatic identification of perovskite crystal structures. Firstly, 1647  $ABX_3$ -type perovskites data containing seven crystal systems, 40 space groups, and lattice parameters were extracted from the MP database, and the initial features include

24 elemental and structural descriptors. The recursive feature descriptor method was used in the feature engineering process to eliminate weakly and unreliably correlated atomic parameters while maintaining the same level of model accuracy. Ten features, including the number of atoms, bond-valence vector sum (BVVS), and atomic number ( $Z$ ), were ultimately kept. The SVM, RF, gradient boosting trees (GBDT), and XGBoost algorithms were used in combination with the selected features to build classification and regression models, respectively. The RF model did the best when the models were first built using a subset of features without BVVS. Subsequently, the RF was used to build classification and regression models based on a subset of features containing BVVS. Additionally, the ACC of the crystal systems classification model increased from 0.915 to 0.974, and the  $R^2$  of the lattice constant model increased from 0.710 to 0.887, indicating that the addition of BVVS can more accurately reflect the structural properties of crystals.

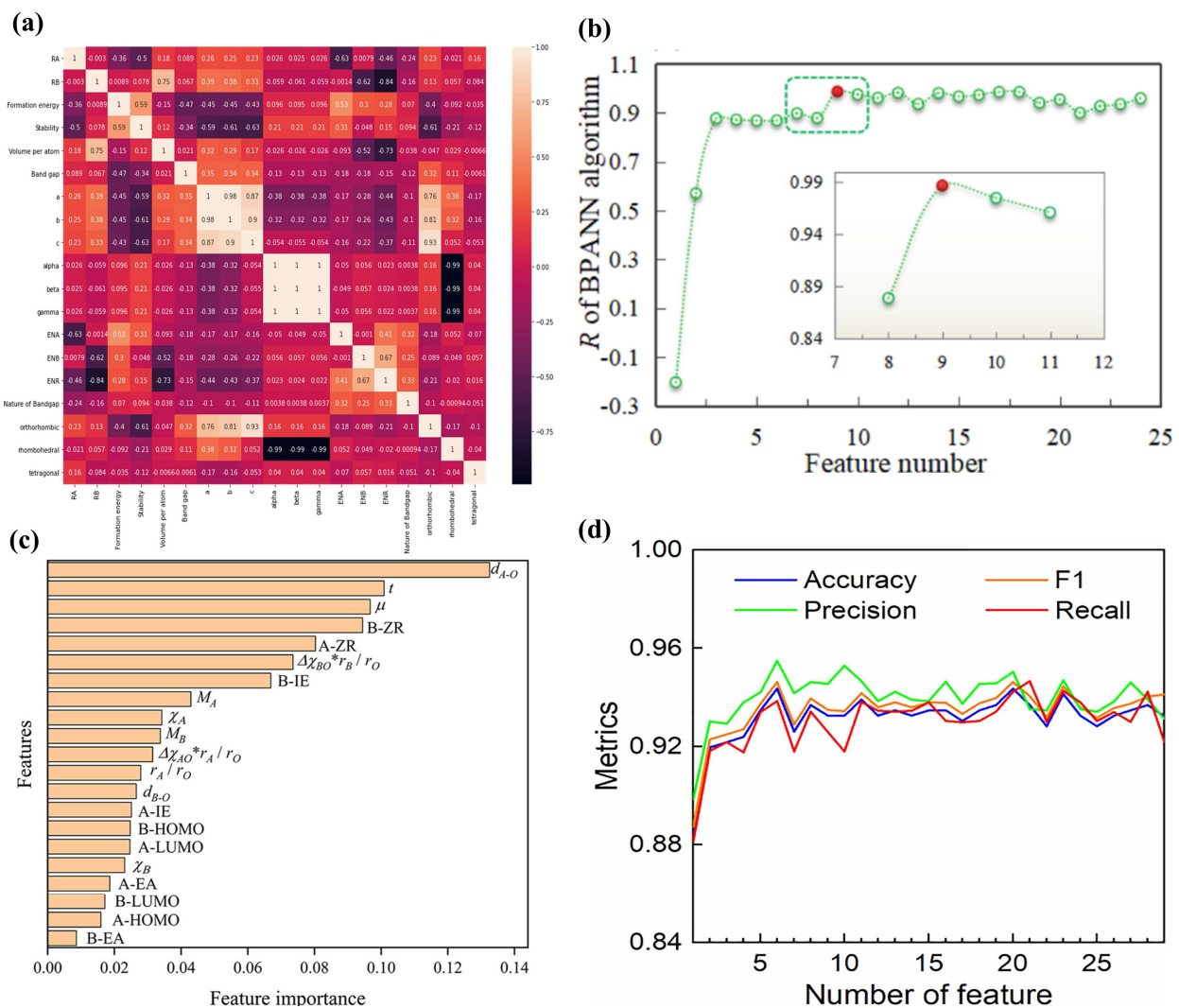
*PCC* as the 'star method' for feature selection is also often used as a feature primary screen or mixed with other feature selection methods such as mRMR, RFE, and embedded methods. In addition, there are other different feature selection methods that are mixed or used step by step. Zhao et al. [46] used the ML method to screen formable and stable perovskite oxides from unexplored  $ABO_3$  combinations. The input data for the ML model consisted of 343 known  $ABO_3$ -type perovskites and 21 initial features. Feature selection was performed based on feature correlation and importance to remove redundant and less important features. Feature correlation was measured by the *PCC* method, and paired Pearson correlation coefficients (*PCCs*) were calculated for the 21 features. Feature importance was obtained from the results of 100 RF models for formability and stability prediction. The importance of the features demonstrates that the formability of perovskites depends mainly on the structural features of the A- and B-site elements, while the properties of the B-site element are the key factor to predict the stability of perovskites. Finally, 16 features were retained for training the formability and stability prediction models of perovskites by analyzing their correlation and importance. For comparison, the RFE method was also used to evaluate the importance of 21 features, and 17 features were retained. The prediction models for formability based on 21, 16, and 17 features, respectively, were denoted as models 1–3, and model two had the highest ACC, precision,  $F_1$  score, and AUC with 0.988, 0.983, 0.992, and 0.999, respectively. Additionally, 21, 16, and 17 features were combined with E-hull to train the stability prediction models, which were denoted as models 4–6, and model five had the best overall results with an AUC as high as 0.983. Li et al. [59] also studied the formability of perovskites based on ML. First, 576  $ABX_3$ -type compounds, including 314 perovskites and 262 non-perovskites, were collected from publications. The initial features were 53 physicochemical parameters. In the step of feature engineering, the initial screening of features was first performed based on the *PCC* method, and the number of feature dimensions was reduced to 29 by using 0.9 as the selecting threshold. For further feature selection, the RFE method was applied to the 29 features, and finally six features ( $\tau$ ,  $\mu$ ,  $t$ ), the ratio of A ion radius to B ion radius ( $R_A/R_B$ ), Pauling electronegativity (EP\_A), and dipole polarizability of the B-site (DP\_B), were retained. Subsequently, five ML algorithms, including RF, DT, SVM, KNN, and LR, were used to construct the classification models, of which the RF model was optimal and the ACC of the model after hyperparameter optimization reached 94.85%. Moreover, it was found that the RF model also correctly predicted whether the compounds could form DPs after testing. The importance of the features of the model shows that  $\tau$  plays a decisive role in the classification model to distinguish between perovskites and non-perovskites. Tao et al. [30] accelerated the discovery of new high-performance and low-cost perovskite photocatalysts in the field of photocatalytic hydrolysis (PWS) by building ML models for hydrogen production rate ( $R_{H_2}$ ) and  $E_g$ . First, 160  $ABO_3$  perovskite photocatalyst data were collected from the experimental literature, of which the  $R_{H_2}$  and  $E_g$  datasets contain 77 and 124 samples, respectively. For the  $E_g$  model, the initial features are 17 atomic parameters and three experimental conditions, while there are 18 atomic parameters and six experimental conditions for the  $R_{H_2}$  model. Four algorithms, including GBR, support vector regression (SVR), backpropagation artificial

neural network (BPANN), and RF, were used to construct the regression models. The mRMR method was used to select the best subset of features for the SVR and BPANN models, while the embedded method was used for the GBR and RF models. The BPANN and GBR models performed optimally for  $R_{H_2}$  and  $E_g$  prediction, which correspond to feature subset dimensions of 9 (Figure 4b) and 7, respectively, while the R of LOOCV reached 0.9869 and 0.9217. Subsequently, Tao et al. [15] proposed a stepwise design strategy for multi-objective optimizations to accelerate the design of potential  $ABO_3$  perovskites with high photocatalytic activity. Data were obtained from the published experimental literature, where the sample sizes used to build models for  $E_g$ , SSA, and CS were 170, 172, and 117, respectively, and the features included 20 atomic parameters and three experimental conditions. Preliminary feature selection was performed by combining PCC and mRMR methods to remove highly correlated features. Firstly, the features of  $E_g$ , SSA, and CS were ranked using the mRMR method. Then the PCCs of any feature pairs were calculated, and if the value of the PCC was greater than 0.9, the features with a lower ranking of mRMR were removed. After the initial selection,  $E_g$  model retained 19 features, while the SSA and CS models both retained 20 features. GBR, SVR, BPANN, and multiple linear regression (MLR) were used to construct the models. The results of LOOCV indicated that GBR was the optimal model with an R of 0.8869 and 0.8733 for predicting  $E_g$  and CS, while SVR was the optimal model with an R of 0.8461 for predicting SSA. In further feature selection, the embedded and mRMR methods were used to select the best features for the GBR and SVR models, respectively, and the final number of retained features was 6, 10, and 9 for  $E_g$ , SSA, and CS, respectively. The SHAP analysis of the retained features showed that the boiling point of the B site showed a significant positive correlation with  $E_g$  and contributed the most to the GBR model; the CT and electron affinity of the B site were key features for the SVR model of SSA; and for the CS model, the CT showed a significant positive correlation with CS, which is consistent with the actual experimental conclusion that the higher the CT, the larger the CS formed.

Some researchers had used a particular feature selection method as a tool to determine whether the initial feature subset was valid and then taken other measures to construct other, more useful features. Liu et al. [28] collected 3430 samples to predict the formation of the oxygen vacancy defect in perovskites. The target variable is the oxygen vacancy formation energy, which is defined as a dichotomous problem of whether an oxygen vacancy defect is likely to form or not by using 0.5 eV as the cutoff, and the initial features are 16 structural parameters containing ionic radius, ionic chemical valence, electronegativity, lattice parameters, tolerance factor, and octahedral factor. In the feature engineering, after drawing the correlation coefficient heat map of the features and the target variable, it was found that no feature was significantly correlated with the target variable; therefore, symbolic classification is used to discover the hidden underlying physical relationships. Since the parsimony coefficient can change the complexity of the corresponding formulas of the generated new structural features, a parsimony coefficient of 0.01 was chosen after weighing, and a simple and effective new structural descriptor,  $n_a(r_a/E_{na} - r_b)$ , was obtained, with the  $n_a$ ,  $r_a$ , and  $E_{na}$  meaning the valence, radius, and electronegativity of the a-site ion, respectively, and  $r_b$  being the b-site ion radius. After modeling with the newly constructed descriptor, the AUC of the interpretable model could reach 0.797. Talapatra et al. [12] constructed their ML model to predict the formability and thermodynamic stability of perovskites. Firstly, a database  $D_F$  of formability and a database  $D_S$  of thermodynamic stability of perovskite were established.  $D_F$  consists of experimentally known  $ABO_3$  and  $AA'BB'O_6$  types of perovskites collected from the literature, including 1187 perovskites and 318 non-perovskites.  $D_S$  contains 3469 samples from their own, independently constructed, basic chemically compatible dataset  $D_C$ . It was found that 1501 perovskites are thermodynamically stable, while the remaining 1955 are thermodynamically unstable after being calculated by DFT. Structural and chemical features were initially used. These features are associated with the A- and B-site atoms of single perovskites, the A-, A'-, B-, and B'-site atoms, and the symmetric and antisymmetric compound features of DPs. The

RFE method was used for feature selection. It was found that atomic features, electron affinity, and geometric features had significant effects on formability and stability. 24 constructed symmetric and antisymmetric compound features based on the first six features and 4 geometric features, including  $t$ ,  $\mu$ , and mismatch factors ( $\overline{\mu_B}$  and  $\overline{\mu_A}$ ) were finally retained. For the formation and thermal stability of perovskites, RF classification models were constructed based on these 28 features, respectively, and the average classification ACC reached 94.01% and 94.09%, respectively. The analysis of feature importance reveals that not only the traditional  $t$  and  $\mu$  contribute very highly to formability, but also many elemental features at the B-site, such as the Zunger pseudopotential radius, electronegativity, and LUMO, are important features to distinguish perovskites from non-perovskites. For the stability classification model, the symmetry features of B-site, such as HOMO, LUMO, ionization energy, and pseudopotential radius, are key features, and the  $t$  is the most important among the geometric features. There is an interesting phenomenon that RFE is the most common feature selection method in the ML workflow for predicting formability and stability. In some application scenarios, GA is also a more effective feature selection method. Xu et al. [13] proposed a multi-properties ML strategy to accelerate the discovery and design of ABO<sub>3</sub>-type ferroelectric perovskites. The data were obtained from publications, including classification data containing 86 ferroelectric perovskites and 61 non-ferroelectric perovskites and regression data containing 95 SSA, 185  $E_g$ , 110  $T_c$ , and 29 dielectric loss ( $\tan\delta$ ) samples. A total of 21 atomic parameters were selected as initial features, and seven features were retained using GA combined with the support vector classification (SVC) model for feature selection. The prediction ACC of LOOCV of the SVC model after hyperparameter optimization was increased from 85.59% to 87.29%. Regression models for SSA,  $E_g$ ,  $T_c$ , and  $\tan\delta$  were built based on the ML workflow and SISSO method, respectively. The SSA,  $E_g$ ,  $T_c$ , and  $\tan\delta$  models by ML workflow all used GA and SVR to select features, and the number of retained features were 13, 16, 16, and 2, respectively. The LASSO models are constructed by using new features selected by the SISSO method. The analysis results indicated that SSA,  $E_g$ , and  $T_c$  tended to be built as regression models by the ML workflow, which had higher R values of 0.935, 0.891, and 0.971, respectively, while a better  $\tan\delta$  model was obtained when using the SISSO method with an R value of 0.931. It could be speculated that the SISSO method may perform better in the case of small datasets. SHAP analysis of the retained features revealed that the three models for SSA,  $E_g$ , and  $T_c$  contained nine common features, including six features associated with the A-position, two features associated with the B-position, and molecular mass. The A-site atomic density showed a strong negative association with SSA and  $E_g$ , and the B-site atomic density demonstrated a negative correlation with all three target variables, according to the Pearson correlation analysis based on the nine features and target variables.



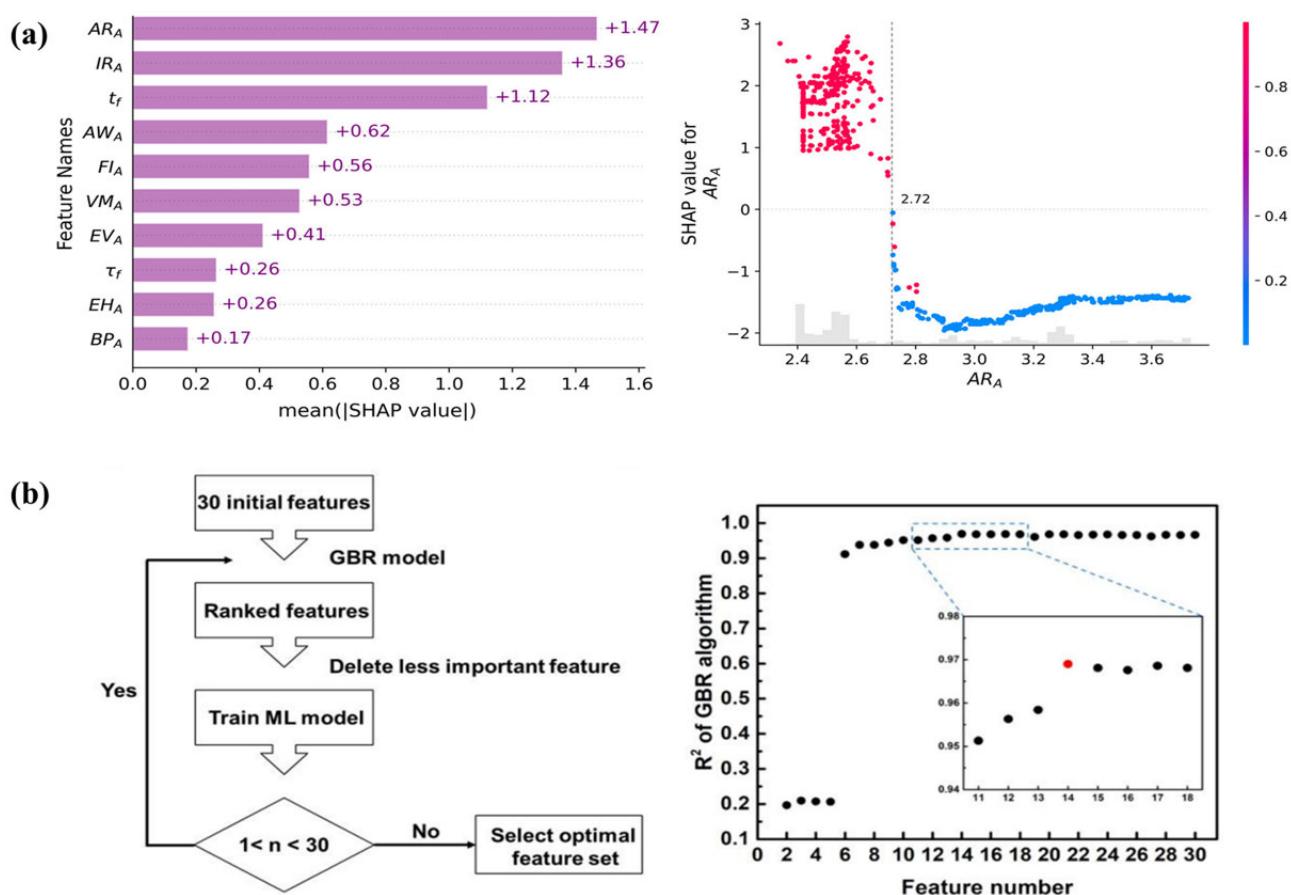


**Figure 4.** Applications of feature selection in inorganic perovskites. (a) A heat map of the correlation between features. Reprinted with permission from ref. [95]. Copyright 2022 Elsevier. (b) For  $R$  of the LOOCV in the feature selection process of the  $R_{H_2}$  model, the position of the red point is the maximum value of  $R$ . Reprinted with permission from ref. [30]. Copyright 2021 Elsevier. (c) The feature importance of the 21 features in predicting the formability. Reprinted with permission from ref. [46]. Copyright 2022 The Authors. (d) An evaluation index vs. feature number of  $ABX_3$  compounds based on the recursive elimination method. Reprinted with permission from ref. [59]. Copyright 2021 Elsevier.

### 5.2. Feature Selection for Hybrid Organic-Inorganic Perovskites

In the study of HOIPs and double HOIPs using ML methods, feature selection by a combination of *PCC* and embedded methods seems to be common. Chen et al. [55] achieved the accelerated discovery of double HOIPs (DHOIPs) by combining ML techniques, HTS, and DFT calculations. The two input datasets consist of 11,161 DHOIPs or HOIPs with  $E_g$  as the target property and 26 initial features, considering the anisotropy of the organic cations at the A-site as well as the HOMO-LUMO gaps and the rotational temperatures. Feature selection was performed based on *PCC* and feature importance from the GBR model, which measured permutation importance and the mean decrease in impurity (MDI). The correlations show that the  $R_A$  and length ( $L_a$ ) of the A-site cations are highly correlated, and the HOMO-LUMO gap is negatively correlated with the cation size. The GBR model was based on 26 initial features, where both the MDI and permutation importance of the  $E_g$  model with the total dataset as input indicate that the features of B-site play a key role

in predicting  $E_g$ . Additionally, the accuracy of the model fitted using only the second dataset was very high, with a MAE of only 0.09 eV. Taking PCC and feature importance into consideration together, the length  $L_A$  of the A site and the number of f electrons in the B site were finally removed, and 24 features were retained. Lu et al. [26] predicted the experimental formability of HOIPs via imbalanced learning. A total of 539 HOIPs and 24 non-HOIPs were obtained from reported literature as a dataset, while 129 features were created based on the Python package for fast-machine-learning. A total of 43 features were kept after the initial feature selection process, which eliminated constant and strongly correlated features. Nine sampling methods and 10 algorithms were used to handle the imbalanced problem and build the classification models, respectively, and it was found that both combinations of SMOTEENN-CAT and SMOTEENN-SVC achieved 100% ACC and precision of LOOCV after a comparative analysis. The CAT model was nested with the SHAP method to achieve further feature selection, and the highest ACC was achieved for both LOOCV and the test set with 100% and 95.5%, respectively, when the number of features was 28. After analyzing the SHAP feature importance and the relationship between the feature values and the corresponding SHAP values, it is found that perovskite is more likely to be formed when the values of the A site atomic radii ( $AR_A$ ) are in the range of 2.30–2.72 Å, which can be confirmed by the existing perovskites (Figure 5a). It is also found that both larger  $R_A$  and  $t$  contribute negatively to the formability of HOIPs.



**Figure 5.** (a) Feature importance extracted via the SHAP method, the scatter plot of  $AR_A$ , and its SHAP value. Reprinted with permission from ref. [26]. Copyright 2022 American Chemical Society. (b) The workflow of 'last-place elimination',  $R^2$  of the GBR model, in each selection process. Reprinted with permission from ref. [97]. Copyright 2018 Springer Nature.

Moreover, the combination of PCC and recursive methods is also very popular among researchers. Zhang et al. [54] predicted the formability of HOIPs using an interpretable

ML strategy. A total of 44 HOIPs and 58 non-HOIPs were collected from publications, and raw features consisted of the three structural parameters  $t$ ,  $\tau$ , and  $\mu$  as well as features obtained from the Mendeleev library and Villars database. A two-step method was used to perform the feature engineering, and the first step used the filter method. The number of features was reduced from 339 to 45 after the removal of features with missing values and relatively unimportant features in feature pairs where the  $PCC$  values exceeded 0.95. Recursive feature addition (RFA) is used in the second stage of the feature selection process to screen out the key features by evaluating the performance of models constructed by the top 2–20 features, which are in the specified feature importance order. For the different algorithms, the specified feature importance is obtained based on the SHAP and mRMR methods, respectively, where the former corresponds to the XGBoost and gradient boosting classifier (GBC) and the latter corresponds to the SVC and the KNN. The optimal prediction ACC under LOOCV was 0.94, 0.91, 0.90, and 0.83 for XGBoost, GBC, SVC, and KNN models with six, four, four, and three features, respectively. SHAP analysis revealed that the  $R_B$  was most important for the formability of HOIP. Wu et al. [98] combined ML techniques and first-principles calculations to achieve rapid screening of mixed double HOIPs (MDHOIPs) for solar cells. Structure-formability classification,  $E_g$  classification, and  $E_g$  regression models were trained based on the reported data of 2274 DHOIPs, with the initial feature set consisting of 87 features related to ion radius, electronegativity, and ion polarizability. Last-place elimination was used to perform feature selection, based on which the relative importance ranking of features can be obtained. For the structure-formability of perovskites, the performance of the classification model was no longer improved when the number of features was greater than 16. The GBC model with an AUC value of 94.3% was trained using the best 16 features, where the ion radius significantly influences the formability of DHOIPs. For the classification and regression models of  $E_g$ , the seventh and eighth most important features were selected, respectively. The  $E_g$  classification model had an AUC value of 97.8%, and the GBR model had an  $R^2$  of 0.974. Both types of models together revealed the importance of the B/B' site ion, and the GBR model demonstrated that the  $E_g$  value was also influenced by the interaction between the B/B' site ion and the X site ion. Cai et al. [99] hastened the discovery of novel lead-free hybrid organic-inorganic DPs with excellent stability, a high Debye temperature, and a suitable  $E_g$  for high-performance PSCs based on DFT and ML techniques. The dataset includes 4456 hybrid organic-inorganic DPs obtained by DFT calculation and 95 features that can be obtained from the periodic table. Among them, 425 compounds with direct  $E_g$  validated by PBE-DFT calculations were extracted to construct the  $E_g$  model. The features were chosen by combining the feature importance of the GBR model with the last-place elimination method, and the  $R^2$ , MSE, and MAE tended to be stable and reached the relative optimal value at 32 features. Analysis of the top 10 features revealed that B/B' and X sites play a key role in  $E_g$  formation. The  $PCC$ s of the 32 retained features were then calculated, and the features with lower feature importance were deleted when the correlation coefficients between any two features were greater than 0.8. Eventually, 14 features were retained. As a side note, the last-place elimination method (Figure 5b) is found to be RFE in essence, and it seems that researchers tend to use it in conjunction with the GBR model.

### 5.3. Feature Selection for Double Perovskites

The feature selection methods used in the study of DPs also include a single method and a combination of different methods. Wang et al. [40] collected 1747 known DP structures with calculated  $E_g$  values obtained from the MP database to predict the  $E_g$  for rapid screening out suitable DPs. Additionally, based on  $E_g$  values, the target variable was classified into three categories:  $E_g$  less than 1.0 eV, between 1.0 and 2.0 eV, and greater than 2.0 eV. A total of 14 descriptors, including isolated elemental properties and differences between properties, were used as initial features to build the GBDT classification model, and the last-place elimination method was used for feature selection. The top ( $N - 1$ ) features were selected to perform the next training at the end of each modeling. After

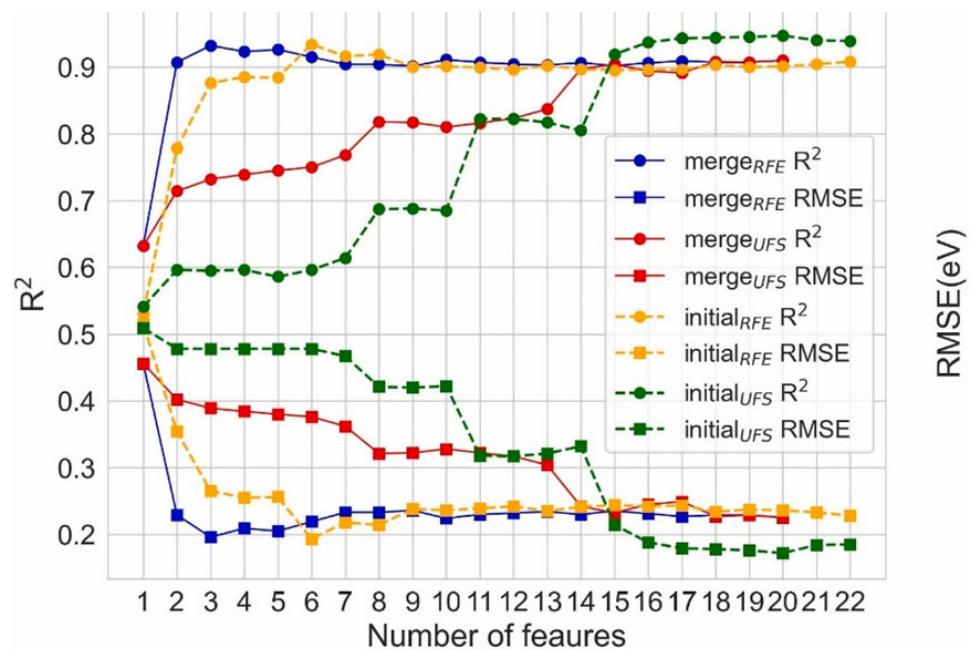
visualization of the relationship between the ACC of the model and the number of features, it was found that the ACC of the model reached its optimal value when nine features were selected, with an ACC of ~92%. The important analysis of the features leads to the inference that the design of the B- and B'-site cation combinations has a significant impact on the value of  $E_g$  for DPs. Liang et al. [39] developed ML models based on the energy above the convex hull ( $E_{\text{hull}}$ ) to screen thermodynamically stable lead-free halide DPs. The dataset was assembled from 469  $A_2B'BX_6$ -type halide DPs with known labels and  $E_{\text{hull}}$  values, containing 112 stable compounds with  $E_{\text{hull}} \leq 0$  and 357 unstable compounds with  $E_{\text{hull}} > 0.24$  elemental features combined with six algorithms were used to build classification models for stable/unstable perovskites as well as the regression model of  $E_{\text{hull}}$ . Based on the SHAP method for feature selection, the XGBoost classification model was optimal when the top 13 features of the SHAP importance ranking were selected for modeling, with an AUC of 0.9551 under a 10-fold CV. For the regression model, the  $R^2$  of the XGBoost regression model constructed based on the top 13 features was 0.83, which was only 0.01 lower than when all features were used for modeling. After analysis of the importance of the retained features, it can be inferred from the SHAP summary plot that perovskites with lower Shannon's ionic radii of X and B'-site atoms as well as higher Shannon's ionic radii of A and B-site atoms tend to have higher stability. The conclusions of the classification and regression models are consistent.

Gao et al. [100] proposed a search strategy combining ML and DFT calculations to screen lead-free inorganic DPs with suitable  $E_g$  and high stability. The dataset consists of 481  $A_2B(\text{I})B(\text{III})X_6$  DPs and 264  $A_2B(\text{II})B(\text{II})X_6$  DPs with a target property of  $E_g$  and 28 chemical properties associated with the  $E_g$  as initial features. The PCC method and the feature importance from the XGBoost algorithm were used together to select features. If the absolute value of PCC for a feature pair is greater than 0.8, a feature with lower feature importance will be deleted. A total of 13 features were finally retained for constructing the models, among which the XGBoost model had the best  $R^2$  of 0.956. The number of valence electrons at the B-site ranks first, and the B'-site polarizability and the B'-X bond energy are relatively important features. The importance ranking of the top 3 features is reliable, which has been confirmed by published papers or could be reasonably explained based on existing theories. Yang et al. [14] discovered potential oxide DPs with narrow  $E_g$  based on the ML method. Firstly, 79  $A_2B'B''O_6$ -type oxide DPs and 75 non-perovskites with  $E_g$  values were collected from the experimental literature. A total of 64 atomic parameters and two process conditions were applied as initial features to the classification model of DPs and the  $E_g$  regression model. To perform feature selection, first the mRMR and PCC methods were used to rank the initial features and measure the correlation between features, respectively. The lower-ranking features were eliminated if the PCC of a feature pair scored higher than 0.95, and finally 49 and 46 features were retained for building classification and regression models, respectively. Further feature selection was then performed for the retained features in combination with classification and regression algorithms to visualize the relationship between the number of features and the evaluation metrics of models including ACC and R. It was found that the highest prediction ACC under LOOCV for the SVC model was 0.959 when the top six features were selected, and when the top 11 features were selected, R under LOOCV for the SVR model reached a peak of 0.916. Further calculations of the PCC between the 11 features and  $E_g$  revealed the same conclusion, in agreement with the results of existing studies, that the  $E_g$  of the oxide DPs is mainly influenced by the ions at the B' and B'' sites.

An interesting case is combining different initial feature sets with different feature selection methods. Liu et al. [101] collected 236 perovskite oxides containing experimental  $E_g$  values from peer-reviewed publications to predict and screen out double perovskite oxides with suitable  $E_g$ . There were two feature sets, including the set of initial features, which consists of 42 component features, and the set of merge features, which consists of 20 new features produced from the weighted average of A- or B-site doped element features. The classical nonlinear regression algorithm RF was chosen considering that



the  $PCCs$  between each feature and the  $E_g$  less than 0.5. The univariate feature selection (UFS) and RFE method based on the RF model (RF-RFE) were used for feature selection. Additionally, the feature set and feature selection methods were combined in two ways, i.e., for both the initial feature set and the merge feature set, different numbers of features were selected for modeling using the UFS and RF-RFE methods, respectively, and the optimal models obtained from different combinations were noted as  $M_1$ ,  $M_2$ ,  $M_3$ , and  $M_4$ , respectively. When using the RFE method, the prediction performance of the model improves rapidly to the optimal level for both feature sets in the ranges of 1–6 and 1–3, respectively, with an  $R^2$  of 0.932 and a RMSE of 0.196 eV for a merge feature number of three (Figure 6). Unlike the RFE, when using the UFS method, the prediction performance of the model improves slowly as the number of features increases, with the RMSE for both feature sets achieving the minimum value when the feature dimension was 20. It was found that the A-site ions contribute particularly significantly to the model based on the importance scores of the features in the  $M_1$ ,  $M_2$ ,  $M_3$ , and  $M_4$  models, and the effect of the A-site ions on the  $E_g$  has been confirmed in studies. According to the  $PCCs$  between features, it was also found that a feature with a small importance score may not mean including less information because the other selected features contain similar information.



**Figure 6.** Computed test set RMSE and  $R^2$  with  $n$ -dimension features, with  $n$  ranging from 1 to 22 for initial features but from 1 to 20 for merge features. Reprinted with permission from ref. [101]. Copyright 2022 Elsevier.

A point worth pondering is that in the above cases, the feature selection methods chosen for predicting the  $E_g$  of DPs were different, which may be due to the difference in sample size and feature dimensions that led to the different choices finally made after trying different methods.

Here are a few cases of PSCs. Liu et al. [102] used a ML approach to intelligently screen passivation materials that help improve the PCE of PSCs. The dataset had a total of 105 samples, each of which included the interface materials used for the perovskite/hole transport layer (HTL) and the corresponding values of PCE. Feature sets are three types of features extracted from interface materials, perovskites, and the performance of standard devices, including electrotopological-state indexes and cheminformatics, ion ratios in precursor solutions, ion types, and control device performance (C\_PCE). The prediction performance of RF models constructed based on different combinations of features showed



that the above three types of features played a key role in model performance. Considering that the feature dimension exceeds 300, the 15 most critical features were selected using SHAP and *PCC* methods. The PCE of the modified device and the C\_PCE have a high positive correlation with a *PCC* value of 0.84. Additionally, based on the correlation matrix, it can be inferred that excess  $\text{Pb}^{2+}$  ions in the precursor solution could lead to the high PCE. Four ML algorithms, including linear regression, RF, XGBoost, and neural networks (NN), were used for modeling to map the relationships between the PCE and the 15 selected features. The RF model with the best performance was used for feature importance analysis, and the results showed that C\_PCE was the most essential feature for determining PCE, in agreement with the analysis of *PCC*. She et al. [103] used a two-step ML method to predict high-efficiency PSCs with doped electron transport layers (ETLs). The 2006 samples of PSCs were collected from the published literature, and two datasets, which include 1820 and 186 samples, respectively, were constructed for the two-step ML. Additionally, the first dataset was the PCE data of PSCs with undoped ETL, while the second dataset was the efficiency improvement rate (EIR) of PSCs with doped ETL, of which 90 PSCs are doped-SnO<sub>2</sub>-based and 96 are doped-TiO<sub>2</sub>-based. Initial features include the doping element and concentration, the physicochemical properties of dopant elements, and the optoelectronic properties of ETL after doping. The feature engineering of the second dataset was performed based on PSCs of doped-SnO<sub>2</sub>-based and doped-TiO<sub>2</sub>-based, respectively. The RF regression model was first built using all features, and the 16 features were ordered by feature importance. Then, the *PCC* of any feature pairs was calculated, and if the absolute value of the *PCC* was higher than 0.8, the one feature with lower importance in the feature pair was deleted, and the features of doped-SnO<sub>2</sub>-based and doped-TiO<sub>2</sub>-based were finally reduced to 10 and 11. Among the top five features, the Fermi level, CBM,  $E_g$ , and conductivity are common features to both SnO<sub>2</sub> and TiO<sub>2</sub>, as well as the generally accepted factors for ETLs to achieve high PCE of PSCs. Since the *PCC*s between any two retained features are mostly below 0.5, it can be inferred that the redundant features have been successfully removed. Modeling based on the retained features, the RMSE values for SnO<sub>2</sub> in the training and test sets are 0.05 and 0.04, respectively, while the values of  $R^2$  are 0.90 and 0.92, which are better than the performance of TiO<sub>2</sub>.

In addition, it should be noted that all the perovskites in the above literature review section are 3D perovskites. The low-dimensional perovskite materials include 0D, 1D, and 2D perovskites, which are classified depending on the spatial arrangement of octahedra in the form of 0D dots, 1D chains, and 2D layers, respectively [104]. Low-dimensional perovskites have also been widely used in solar cells, light-emitting diodes, and photodetectors due to their flexible structures, excellent photovoltaic properties, and higher stability [105,106]. Among them, 2D perovskites have attracted a lot of attention due to the wide tunability of their photovoltaic properties and excellent stability [107,108]. The (100)-oriented 2D perovskites are the most common, especially the Ruddlesden–Popper (RP) and Dion–Jacobson (DJ) phases [109]. Therefore, a few cases of applications of feature selection in 2D perovskite materials are also briefly described below.

Lyu et al. [110] reported an ML-assisted method to investigate how the dimensionality of lead iodide perovskites was impacted by the structure of organic cations. The dataset is derived from 86 amines reported in the literature for low-dimensional lead iodide perovskites, which were classified according to the dimensionality of the perovskites as “2D” and “non-2D”. A total of 40 initial features were generated by descriptor functions, and 21 features were finally retained after using 0.95 as the threshold for *PCC*s to remove highly correlated features. LR, SVM, KNN, and DT were used to build the classification models, and the LR model with a prediction ACC of  $0.82 \pm 0.08$  on the test set was used in the follow-up study. Feature selection was performed based on the feature coefficients with the  $L_1$  penalty in the LR model, and four features were finally selected to construct the prediction model. Additionally, it was found that the topological and geometric properties of ammonium cations played a key role in determining the dimensionality. The primary amine with a smaller steric effect index (STEI) is more likely to form 2D

perovskites. Due to the eccentricity ( $E_c$ ) having a feature coefficient of 1.922, it is possible to determine that octylammonium is predicted to form 2D perovskite more readily than cyclooctylammonium. According to the largest ring size (LRS) with a negative feature coefficient, molecules with a bigger ring are likely to produce lower-dimensional perovskite. Hu et al. [111] obtained the adsorption energy of 640 ion/perovskites by first-principles calculations to assess the interaction between 2D  $A_2BX_4$  halide perovskites and ions in energy storage applications. The *PCC* method was used for feature selection, and only appropriate features were retained when the *PCCs* of the feature pairs were greater than 0.8 or less than  $-0.8$ . A total of 13 features were finally selected from 73 original features. After calculating and ranking the *PCCs* of these 13 features with the adsorption energy, it was found that ion density, melting point, and shell layer had higher rankings, which emphasized the major contributions made by the types of ion adsorbates. A total of six ML algorithms—KNN, Kriging, RF, Rpart, SVM, and XGBoost—were used to build the models, and the XGBoost model had the highest  $R$  and  $R^2$  of 0.968 and 0.93, respectively. Meanwhile, to avoid the bias caused by the *PCC* method, 14 feature ranking methods were selected to comprehensively assess the importance of ion density, ion radius, and first ionization of B-site elements. The different ranking methods consistently show the importance of ion density on the adsorption energy, but the *PCC* method is slightly biased in assessing the importance of atomic radii. Zhang et al. [109] applied the ML method to accelerate the synthetic development of (100)-oriented 2D lead halide perovskites (LHPs). The dataset was derived from 264 crystal structures containing PR and DJ phases in the existing literature, and the feature pool consists of nine features, including the number of protonated nitrogen atoms ( $q$ ), the radius of the halide ion ( $r(X)$ ), the distortion of the  $PbX_6$  octahedral bond length ( $\lambda$ ), etc. The Spearman correlation coefficient (SCC) was used to perform univariate feature selection, and the linear correlation coefficient between  $\lambda$  and  $r(X)$  was found to exceed 0.8, up to 0.91.  $r(X)$  was removed because  $\lambda$  contained more information, and eight features were finally retained. A total of 26 ML classification models were selected, of which the XGBoost model had the best ACC at 84.4%. The importance of features in the XGBoost model showed that  $q$  is the dominant feature. Overall, the electronic, topologic, and geometric properties of the organic amine cations have a significant impact on the crystal structures of 2D LHPs. Using the SHAP method for further feature analysis, it was found that low octahedral bond angle distortion, small inorganic layer spacing, and high octahedral bond length distortion have a significant negative contribution on forming the RP/nRP-phase. It is easy to see that the *PCC* method is still the preferred method, but the comparison results with other ranking methods also show that the *PCC* method sometimes has bias while the SCC method is less common.

Generally speaking, feature selection reduces the dimensionality of the features while maintaining or improving the performance of the model in almost all of the scenarios mentioned above, fully demonstrating the importance of feature selection. In terms of the choice of feature selection methods, *PCC* is the most frequently used method for perovskite materials, but the threshold value selected for filtering highly correlated features varies in different usage scenarios. The mixed feature selection methods are also a common screening strategy. When selecting feature selection methods for one's own research object, one can first try to use the method with a relatively high frequency, but it should be clear that the effectiveness of the feature selection methods is also closely related to the data quality and the selected algorithm, etc.

## 6. Conclusions and Outlook

In conclusion, feature selection is an essential part of the materials ML workflow. This review briefly introduces the common structures of perovskite materials and the generic descriptor types, as well as the common feature selection methods in the filter, wrapper, and embedded methods. Some of the applications of feature selection in the discovery and design process of perovskite materials based on ML methods are reviewed. It is found that *PCC* in the filter method, RFE in the wrapped method, and tree modeling in the embedded

method appear more frequently, whether they are used singly or in combination. From this review, we found that an appropriate feature selection method can reduce model complexity and improve model interpretability to a great extent. Although feature selection has been successfully applied in the materials ML workflow, there is still much room for progress. Here, we tend to propose the following directions for the subsequent application of feature selection in the design and discovery of perovskite materials:

- (1) The establishment and improvement of the perovskite materials database: Data is the 'hardware' for performing ML, and the quantity and quality of data are the keys to model performance. Compared with other fields, data in the materials field is usually characterized by small size and multiple sources. However, a sample size in a large proportion of materials research articles is less than 1000 or even less than 500. For perovskite materials, a dedicated perovskite database platform to collect data of various excellent properties and perovskite device parameters can be established and made available in a form that adheres to FAIR (findable, accessible, interoperable, and reusable) data principles;
- (2) Descriptor construction and sharing: To maximize the accuracy of the model and to avoid situations where the ML results contradict the domain expert knowledge, the descriptors can be constructed manually by combining the material domain knowledge. At the same time, for researchers in non-specialized fields, new descriptors can be constructed automatically by means of SISSO and symbolic regression methods. In addition, to break the professional barriers of different fields and further promote the discovery and design of materials, it is also necessary to establish an online access platform of descriptors corresponding to the database, which can make the professional people focus on doing the professional things to provide a greater possibility for the breakthrough of material properties. Taking perovskite thin film as an example [62–66], we encourage researchers to record more detailed process parameters for preparing high-quality thin films in manuscripts and construct a relevant database of process parameters. The key parameters affecting film quality could be selected by employing suitable feature selection methods based on the database. Then an ML model for quantitative analysis of process parameters and film quality can be constructed, offering the possibility of accelerating the optimization of process parameters and guiding the experimental synthesis of high-quality thin films;
- (3) Evaluation and development of feature selection methods: In the application of materials ML workflow, researchers have mostly only objectively stated which methods were used for feature selection, and then model construction and selection based on the selected feature subsets were performed. The selection of methods is, in essence, serving the current data. The input of different feature subsets is the result of different selection methods, so the evaluation and comparison of feature selection methods in conjunction with ML algorithms is also quite an important topic. The development of new feature selection methods for material data can also be considered. Based on some practical experience, the ensemble idea can be used to develop ensemble feature selection methods applicable to materials data, which can ensure the stability of feature subsets and thus have stronger generality.

In summary, with the increase of material requirements and demands as well as the rapid development of intelligent methods, ML will continue to be an important tool for other materials. The feature selection, as a key part of the ML workflow, will also receive more attention in the discovery and design process of perovskite materials via ML.

**Author Contributions:** Writing—original draft preparation, J.W.; writing—review and editing, P.X., M.L., and W.L.; funding acquisition, M.L. and X.J.; supervision, X.J. 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 (52102140) and the Shanghai Pujiang Program (21PJJD024).

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** All the data of the examples could be obtained from the corresponding references.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Jordan, M.I.; Mitchell, T.M. Machine learning: Trends, perspectives, and prospects. *Science* **2015**, *349*, 255–260. [[CrossRef](#)] [[PubMed](#)]
2. Shehab, M.; Abualigah, L.; Shambour, Q.; Abu-Hashem, M.A.; Shambour, M.K.Y.; Alslibi, A.I.; Gandomi, A.H. Machine learning in medical applications: A review of state-of-the-art methods. *Comput. Biol. Med.* **2022**, *145*, 105458. [[CrossRef](#)]
3. Henrique, B.M.; Sobreiro, V.A.; Kimura, H. Literature review: Machine learning techniques applied to financial market prediction. *Expert Syst. Appl.* **2019**, *124*, 226–251. [[CrossRef](#)]
4. Liakos, K.G.; Busato, P.; Moshou, D.; Pearson, S.; Bochtis, D. Machine Learning in Agriculture: A Review. *Sensors* **2018**, *18*, 2674. [[CrossRef](#)] [[PubMed](#)]
5. Larranaga, P.; Calvo, B.; Santana, R.; Bielza, C.; Galdiano, J.; Inza, I.; Lozano, J.A.; Armananzas, R.; Santafe, G.; Perez, A.; et al. Machine learning in bioinformatics. *Brief. Bioinform.* **2006**, *7*, 86–112. [[CrossRef](#)] [[PubMed](#)]
6. Butler, K.T.; Davies, D.W.; Cartwright, H.; Isayev, O.; Walsh, A. Machine learning for molecular and materials science. *Nature* **2018**, *559*, 547–555. [[CrossRef](#)]
7. Schmidt, J.; Marques, M.R.G.; Botti, S.; Marques, M.A.L. Recent advances and applications of machine learning in solid-state materials science. *NPJ Comput. Mater.* **2019**, *5*, 83. [[CrossRef](#)]
8. Tao, Q.; Xu, P.; Li, M.; Lu, W. Machine learning for perovskite materials design and discovery. *NPJ Comput. Mater.* **2021**, *7*, 23. [[CrossRef](#)]
9. Min, K.; Cho, E. Accelerated discovery of potential ferroelectric perovskite via active learning. *J. Mater. Chem. C* **2020**, *8*, 7866–7872. [[CrossRef](#)]
10. Gok, E.C.; Yildirim, M.O.; Haris, M.P.U.; Eren, E.; Pegu, M.; Hemasiri, N.H.; Huang, P.; Kazim, S.; Uygun Oksuz, A.; Ahmad, S. Predicting Perovskite Bandgap and Solar Cell Performance with Machine Learning. *Sol. RRL* **2021**, *6*, 2100927. [[CrossRef](#)]
11. Yin, W.-J.; Weng, B.; Ge, J.; Sun, Q.; Li, Z.; Yan, Y. Oxide perovskites, double perovskites and derivatives for electrocatalysis, photocatalysis, and photovoltaics. *Energy Environ. Sci.* **2019**, *12*, 442–462. [[CrossRef](#)]
12. 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](#)]
13. Xu, P.; Chang, D.; Lu, T.; Li, L.; Li, M.; Lu, W. Search for ABO<sub>3</sub> Type Ferroelectric Perovskites with Targeted Multi-Properties by Machine Learning Strategies. *J. Chem. Inf. Model.* **2022**, *62*, 5038–5049. [[CrossRef](#)]
14. Yang, X.; Li, L.; Tao, Q.; Lu, W.; Li, M. Rapid discovery of narrow bandgap oxide double perovskites using machine learning. *Comput. Mater. Sci.* **2021**, *196*, 110528. [[CrossRef](#)]
15. Tao, Q.; Chang, D.; Lu, T.; Li, L.; Chen, H.; Yang, X.; Liu, X.; Li, M.; Lu, W. Multiobjective Stepwise Design Strategy-Assisted Design of High-Performance Perovskite Oxide Photocatalysts. *J. Phys. Chem. C* **2021**, *125*, 21141–21150. [[CrossRef](#)]
16. Liu, Y.; Wu, J.M.; Avdeev, M.; Shi, S.Q. Multi-Layer Feature Selection Incorporating Weighted Score-Based Expert Knowledge toward Modeling Materials with Targeted Properties. *Adv. Theory Simul.* **2020**, *3*, 1900215. [[CrossRef](#)]
17. Yao, G.; Hu, X.; Wang, G. A novel ensemble feature selection method by integrating multiple ranking information combined with an SVM ensemble model for enterprise credit risk prediction in the supply chain. *Expert Syst. Appl.* **2022**, *200*, 117002. [[CrossRef](#)]
18. Hira, Z.M.; Gillies, D.F. A Review of Feature Selection and Feature Extraction Methods Applied on Microarray Data. *Adv. Bioinform.* **2015**, *2015*, 198363. [[CrossRef](#)]
19. Zhang, X.; Yu, L.; Yin, H.; Lai, K.K. Integrating data augmentation and hybrid feature selection for small sample credit risk assessment with high dimensionality. *Comput. Oper. Res.* **2022**, *146*, 105937. [[CrossRef](#)]
20. Xu, P.; Chen, H.; Li, M.; Lu, W. New Opportunity: Machine Learning for Polymer Materials Design and Discovery. *Adv. Theory Simul.* **2022**, *5*, 2100565. [[CrossRef](#)]
21. Zhou, Q.; Lu, S.; Wu, Y.; Wang, J. Property-Oriented Material Design Based on a Data-Driven Machine Learning Technique. *J. Phys. Chem. Lett.* **2020**, *11*, 3920–3927. [[CrossRef](#)] [[PubMed](#)]
22. Belsky, A.; Hellenbrandt, M.; Karen, V.L.; Luksch, P. New developments in the Inorganic Crystal Structure Database (ICSD): Accessibility in support of materials research and design. *Acta Crystallogr. Sect. B-Struct. Sci. Cryst. Eng. Mat.* **2002**, *58*, 364–369. [[CrossRef](#)]
23. Saal, J.E.; Kirklin, S.; Aykol, M.; Meredig, B.; Wolverton, C. Materials Design and Discovery with High-Throughput Density Functional Theory: The Open Quantum Materials Database (OQMD). *JOM* **2013**, *65*, 1501–1509. [[CrossRef](#)]
24. 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](#)]



25. Dong, Y.; Zhang, Y.; Ran, M.; Zhang, X.; Liu, S.; Yang, Y.; Hu, W.; Zheng, C.; Gao, X. Accelerated identification of high-performance catalysts for low-temperature NH<sub>3</sub>-SCR by machine learning. *J. Mater. Chem. A* **2021**, *9*, 23850–23859. [[CrossRef](#)]
26. Lu, T.; Li, H.; Li, M.; Wang, S.; Lu, W. Predicting Experimental Formability of Hybrid Organic-Inorganic Perovskites via Imbalanced Learning. *J. Phys. Chem. Lett.* **2022**, *13*, 3032–3038. [[CrossRef](#)] [[PubMed](#)]
27. Ouyang, R.; Curtarolo, S.; Ahmetcik, E.; Scheffler, M.; Ghiringhelli, L.M. SISSO: A compressed-sensing method for identifying the best low-dimensional descriptor in an immensity of offered candidates. *Phys. Rev. Mater.* **2018**, *2*, 083802. [[CrossRef](#)]
28. Liu, S.; Wang, J.; Duan, Z.; Wang, K.; Zhang, W.; Guo, R.; Xie, F. Simple Structural Descriptor Obtained from Symbolic Classification for Predicting the Oxygen Vacancy Defect Formation of Perovskites. *ACS Appl. Mater. Interfaces* **2022**, *14*, 11758–11767. [[CrossRef](#)]
29. Mai, J.; Lu, T.; Xu, P.; Lian, Z.; Li, M.; Lu, W. Predicting the maximum absorption wavelength of azo dyes using an interpretable machine learning strategy. *Dyes Pigment.* **2022**, *206*, 110647. [[CrossRef](#)]
30. Tao, Q.; Lu, T.; Sheng, Y.; Li, L.; Lu, W.; Li, M. Machine learning aided design of perovskite oxide materials for photocatalytic water splitting. *J. Energy Chem.* **2021**, *60*, 351–359. [[CrossRef](#)]
31. Lu, T.; Li, H.; Li, M.; Wang, S.; Lu, W. Inverse Design of Hybrid Organic-Inorganic Perovskites with Suitable Bandgaps via Proactive Searching Progress. *ACS Omega* **2022**, *7*, 21583–21594. [[CrossRef](#)] [[PubMed](#)]
32. Yang, C.; Ren, C.; Jia, Y.; Wang, G.; Li, M.; Lu, W. A machine learning-based alloy design system to facilitate the rational design of high entropy alloys with enhanced hardness. *Acta Mater.* **2022**, *222*, 117431. [[CrossRef](#)]
33. Shi, L.; Chang, D.; Ji, X.; Lu, W. Using Data Mining To Search for Perovskite Materials with Higher Specific Surface Area. *J. Chem. Inf. Model.* **2018**, *58*, 2420–2427. [[CrossRef](#)] [[PubMed](#)]
34. Wang, Y.; Lv, Z.; Zhou, L.; Chen, X.; Chen, J.; Zhou, Y.; Roy, V.A.L.; Han, S.-T. Emerging perovskite materials for high density data storage and artificial synapses. *J. Mater. Chem. C* **2018**, *6*, 1600–1617. [[CrossRef](#)]
35. Žužić, A.; Ressler, A.; Macan, J. Perovskite oxides as active materials in novel alternatives to well-known technologies: A review. *Ceram. Int.* **2022**, *48*, 27240–27261. [[CrossRef](#)]
36. Tian, W.; Zhou, H.; Li, L. Hybrid Organic-Inorganic Perovskite Photodetectors. *Small* **2017**, *13*, 170210. [[CrossRef](#)]
37. Zuo, T.; He, X.; Hu, P.; Jiang, H. Organic-Inorganic Hybrid Perovskite Single Crystals: Crystallization, Molecular Structures, and Bandgap Engineering. *ChemNanoMat* **2019**, *5*, 278–289. [[CrossRef](#)]
38. Kumar, A.; Rana, N.K.; Rani, S.; Ghosh, D.S. Toward all-inorganic perovskite solar cells: Materials, performance, and stability. *Int. J. Energy Res.* **2022**, *46*, 14659–14695. [[CrossRef](#)]
39. Liang, G.-Q.; Zhang, J. A machine learning model for screening thermodynamic stable lead-free halide double perovskites. *Comput. Mater. Sci.* **2022**, *204*, 111172. [[CrossRef](#)]
40. Wang, Z.; Han, Y.; Lin, X.; Cai, J.; Wu, S.; Li, J. An Ensemble Learning Platform for the Large-Scale Exploration of New Double Perovskites. *ACS Appl. Mater. Interfaces* **2022**, *14*, 717–725. [[CrossRef](#)]
41. Wang, H.; Zhang, Q.; Qiu, M.; Hu, B. Synthesis and application of perovskite-based photocatalysts in environmental remediation: A review. *J. Mol. Liq.* **2021**, *334*, 116029. [[CrossRef](#)]
42. Wang, W.; Tadé, M.O.; Shao, Z. Research progress of perovskite materials in photocatalysis- and photovoltaics-related energy conversion and environmental treatment. *Chem. Soc. Rev.* **2015**, *44*, 5371–5408. [[CrossRef](#)]
43. Tai, Q.; Tang, K.-C.; Yan, F. Recent progress of inorganic perovskite solar cells. *Energy Environ. Sci.* **2019**, *12*, 2375–2405. [[CrossRef](#)]
44. Liu, X.; Li, J.; Cui, X.; Wang, X.; Yang, D. Strategies for the preparation of high-performance inorganic mixed-halide perovskite solar cells. *RSC Adv.* **2022**, *12*, 32925–32948. [[CrossRef](#)]
45. 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*, eaav0693. [[CrossRef](#)] [[PubMed](#)]
46. Zhao, J.; Wang, X. Screening Perovskites from ABO<sub>3</sub> Combinations Generated by Constraint Satisfaction Techniques Using Machine Learning. *ACS Omega* **2022**, *7*, 10483–10491. [[CrossRef](#)] [[PubMed](#)]
47. Fu, M.; Wang, L.; Ma, T.; Wu, J.; Dai, S.; Chang, Z.; Zhang, Q.; Xu, H.; Li, X. Chemical formula input relied intelligent identification of an inorganic perovskite for solar thermochemical hydrogen production. *Inorg. Chem. Front.* **2021**, *8*, 2097–2102. [[CrossRef](#)]
48. Zhai, X.; Ding, F.; Zhao, Z.; Santomauro, A.; Luo, F.; Tong, J. Predicting the formation of fractionally doped perovskite oxides by a function-confined machine learning method. *Commun. Mater.* **2022**, *3*, 42. [[CrossRef](#)]
49. Villars, P. Materials Platform for Data Science. 2019. Available online: <https://mpds.io/> (accessed on 10 March 2023).
50. Mentel, L.M. Mendeleev—A Python Resource for Properties of Chemical Elements, Ions and Isotopes. 2014. Available online: <https://github.com/lmmentel/mendeleev> (accessed on 10 March 2023).
51. Landrum, G. RDKit: Open Source Cheminformatics. 2012. Available online: <http://www.rdkit.org/> (accessed on 10 March 2023).
52. Basavarajappa, M.G.; Nazeeruddin, M.K.; Chakraborty, S. Evolution of hybrid organic-inorganic perovskite materials under external pressure. *Appl. Phys. Rev.* **2021**, *8*, 041309. [[CrossRef](#)]
53. Lu, T.; Li, M.; Lu, W.; Zhang, T.-Y. Recent progress in the data-driven discovery of novel photovoltaic materials. *J. Mater. Inform.* **2022**, *2*, 7. [[CrossRef](#)]
54. Zhang, S.; Lu, T.; Xu, P.; Tao, Q.; Li, M.; Lu, W. Predicting the Formability of Hybrid Organic-Inorganic Perovskites via an Interpretable Machine Learning Strategy. *J. Phys. Chem. Lett.* **2021**, *12*, 7423–7430. [[CrossRef](#)] [[PubMed](#)]
55. Chen, J.; Xu, W.; Zhang, R.  $\Delta$ -Machine learning-driven discovery of double hybrid organic-inorganic perovskites. *J. Mater. Chem. A* **2022**, *10*, 1402–1413. [[CrossRef](#)]



56. Pilania, G.; Mannodi-Kanakthodi, A.; Uberuaga, B.P.; Ramprasad, R.; Gubernatis, J.E.; Lookman, T. Machine learning bandgaps of double perovskites. *Sci. Rep.* **2016**, *6*, 19375. [[CrossRef](#)] [[PubMed](#)]
57. Halder, A.; Ghosh, A.; Dasgupta, T.S. Machine-learning-assisted prediction of magnetic double perovskites. *Phys. Rev. Mater.* **2019**, *3*, 084418. [[CrossRef](#)]
58. Nair, S.S.; Krishnia, L.; Trukhanov, A.; Thakur, P.; Thakur, A. Prospect of double perovskite over conventional perovskite in photovoltaic applications. *Ceram. Int.* **2022**, *48*, 34128–34147. [[CrossRef](#)]
59. Li, L.; Tao, Q.; Xu, P.; Yang, X.; Lu, W.; Li, M. Studies on the regularity of perovskite formation via machine learning. *Comput. Mater. Sci.* **2021**, *199*, 110712. [[CrossRef](#)]
60. Zhu, W.; Wang, S.; Zhang, X.; Wang, A.; Wu, C.; Hao, F. Ion Migration in Organic-Inorganic Hybrid Perovskite Solar Cells: Current Understanding and Perspectives. *Small* **2022**, *18*, 2105783. [[CrossRef](#)]
61. Song, T.-B.; Chen, Q.; Zhou, H.; Jiang, C.; Wang, H.-H.; Yang, Y.; Liu, Y.; You, J.; Yang, Y. Perovskite solar cells: Film formation and properties. *J. Mater. Chem. A* **2015**, *3*, 9032–9050. [[CrossRef](#)]
62. Costa, J.C.S.; Azevedo, J.; Araújo, J.P.; Santos, L.M.N.B.F.; Mendes, A. High purity and crystalline thin films of methylammonium lead iodide perovskites by a vapor deposition approach. *Thin Solid Films* **2018**, *664*, 12–18. [[CrossRef](#)]
63. Saki, Z.; Byranvand, M.M.; Taghavinia, N.; Kedia, M.; Saliba, M. Solution-processed perovskite thin-films: The journey from lab-to large-scale solar cells. *Energy Environ. Sci.* **2021**, *14*, 5690–5722. [[CrossRef](#)]
64. Xu, F.; Li, Y.; Yuan, B.; Zhang, Y.; Wei, H.; Wu, Y.; Cao, B. Large-area CsPbBr<sub>3</sub> perovskite films grown with effective one-step RF-magnetron sputtering. *J. Appl. Phys.* **2021**, *129*, 245303. [[CrossRef](#)]
65. Alanazi, T.I. Current spray-coating approaches to manufacture perovskite solar cells. *Results Phys.* **2023**, *44*, 106144. [[CrossRef](#)]
66. Swartwout, R.; Hoerantner, M.T.; Bulović, V. Scalable Deposition Methods for Large-area Production of Perovskite Thin Films. *Energy Environ. Mater.* **2019**, *2*, 119–145. [[CrossRef](#)]
67. Chandrashekar, G.; Sahin, F. A survey on feature selection methods. *Comput. Electr. Eng.* **2014**, *40*, 16–28. [[CrossRef](#)]
68. Remeseiro, B.; Bolon-Canedo, V. A review of feature selection methods in medical applications. *Comput. Biol. Med.* **2019**, *112*, 103375. [[CrossRef](#)]
69. Urbanowicz, R.J.; Meeker, M.; La Cava, W.; Olson, R.S.; Moore, J.H. Relief-based feature selection: Introduction and review. *J. Biomed. Inform.* **2018**, *85*, 189–203. [[CrossRef](#)]
70. Zhang, J.; Xiong, Y.; Min, S. A new hybrid filter/wrapper algorithm for feature selection in classification. *Anal. Chim. Acta* **2019**, *1080*, 43–54. [[CrossRef](#)]
71. Pudjihartono, N.; Fadason, T.; Kempa-Liehr, A.W.; O’Sullivan, J.M. A Review of Feature Selection Methods for Machine Learning-Based Disease Risk Prediction. *Front. Bioinform.* **2022**, *2*, 927312. [[CrossRef](#)]
72. Saeys, Y.; Inza, I.; Larranaga, P. A review of feature selection techniques in bioinformatics. *Bioinformatics* **2007**, *23*, 2507–2517. [[CrossRef](#)]
73. Venkatesh, B.; Anuradha, J. A Review of Feature Selection and Its Methods. *Cybern. Inf. Technol.* **2019**, *19*, 3–26. [[CrossRef](#)]
74. Wang, Y.H.; Zhang, Y.F.; Zhang, Y.; Gu, Z.F.; Zhang, Z.Y.; Lin, H.; Deng, K.J. Identification of adaptor proteins using the ANOVA feature selection technique. *Methods* **2022**, *208*, 42–47. [[CrossRef](#)] [[PubMed](#)]
75. Biesiada, J.; Duch, W. Feature Selection for High-Dimensional Data—A Pearson Redundancy Based Filter. In *Computer Recognition Systems 2*; Kurzynski, M., Puchala, E., Wozniak, M., Zolnierak, A., Eds.; Springer: Berlin/Heidelberg, Germany, 2007; pp. 242–249.
76. Liu, Y.; Mu, Y.; Chen, K.; Li, Y.; Guo, J. Daily Activity Feature Selection in Smart Homes Based on Pearson Correlation Coefficient. *Neural Process. Lett.* **2020**, *51*, 1771–1787. [[CrossRef](#)]
77. Edelman, D.; Móri, T.F.; Székely, G.J. On relationships between the Pearson and the distance correlation coefficients. *Stat. Probab. Lett.* **2021**, *169*, 108960. [[CrossRef](#)]
78. Peng, H.; Long, F.; Ding, C. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans. Pattern Anal. Mach. Intell.* **2005**, *27*, 1226–1238. [[CrossRef](#)]
79. Reshef, D.N.; Reshef, Y.A.; Finucane, H.K.; Grossman, S.R.; McVean, G.; Turnbaugh, P.J.; Lander, E.S.; Mitzenmacher, M.; Sabeti, P.C. Detecting Novel Associations in Large Data Sets. *Science* **2011**, *334*, 1518–1524. [[CrossRef](#)]
80. Bommert, A.; Welchowski, T.; Schmid, M.; Rahnenfuhrer, J. Benchmark of filter methods for feature selection in high-dimensional gene expression survival data. *Brief. Bioinform.* **2022**, *23*, bbab354. [[CrossRef](#)]
81. Almaghthawi, Y.; Ahmad, I.; Alsaadi, F.E. Performance Analysis of Feature Subset Selection Techniques for Intrusion Detection. *Mathematics* **2022**, *10*, 4745. [[CrossRef](#)]
82. Xue, B.; Zhang, M.; Browne, W.N.; Yao, X. A Survey on Evolutionary Computation Approaches to Feature Selection. *IEEE Trans. Evol. Comput.* **2016**, *20*, 606–626. [[CrossRef](#)]
83. Jablonka, K.M.; Ongari, D.; Moosavi, S.M.; Smit, B. Big-Data Science in Porous Materials: Materials Genomics and Machine Learning. *Chem. Rev.* **2020**, *120*, 8066–8129. [[CrossRef](#)]
84. Granitto, P.M.; Furlanello, C.; Biasioli, F.; Gasperi, F. Recursive feature elimination with random forest for PTR-MS analysis of agroindustrial products. *Chemom. Intell. Lab. Syst.* **2006**, *83*, 83–90. [[CrossRef](#)]
85. Tsai, C.-F.; Eberle, W.; Chu, C.-Y. Genetic algorithms in feature and instance selection. *Knowl. Based Syst.* **2013**, *39*, 240–247. [[CrossRef](#)]
86. Tan, F.; Fu, X.; Zhang, Y.; Bourgeois, A.G. A genetic algorithm-based method for feature subset selection. *Soft Comput.* **2008**, *12*, 111–120. [[CrossRef](#)]

87. Yang, J.W.; Wang, S.L.; Chen, Y.Y.; Lu, S.K.; Yang, W.Z. Feature Subset Selection Based on the Genetic Algorithm. *Adv. Mater. Res.* **2013**, *774*, 1532–1537. [[CrossRef](#)]
88. Ai, C. A Method for Cancer Genomics Feature Selection Based on LASSO-RFE. *Iran. J. Sci. Technol. Trans. A Sci.* **2022**, *46*, 731–738. [[CrossRef](#)]
89. Chen, H.; Shang, Z.; Lu, W.; Li, M.; Tan, F. A Property-Driven Stepwise Design Strategy for Multiple Low-Melting Alloys via Machine Learning. *Adv. Eng. Mater.* **2021**, *23*, 2100612. [[CrossRef](#)]
90. Jiménez-Cordero, A.; Morales, J.M.; Pineda, S. A novel embedded min-max approach for feature selection in nonlinear Support Vector Machine classification. *Eur. J. Oper. Res.* **2021**, *293*, 24–35. [[CrossRef](#)]
91. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [[CrossRef](#)]
92. Otchere, D.A.; Ganat, T.O.A.; Ojero, J.O.; Tackie-Otoo, B.N.; Taki, M.Y. Application of gradient boosting regression model for the evaluation of feature selection techniques in improving reservoir characterisation predictions. *J. Pet. Sci. Eng.* **2022**, *208*, 109244. [[CrossRef](#)]
93. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794.
94. Lundberg, S.M.; Lee, S.I. A Unified Approach to Interpreting Model Predictions. *arXiv* **2017**, arXiv:1705.07874.
95. Priyanga, G.S.; Mattur, M.N.; Nagappan, N.; Rath, S.; Thomas, T. Prediction of nature of band gap of perovskite oxides (ABO<sub>3</sub>) using a machine learning approach. *J. Mater.* **2022**, *8*, 937–948. [[CrossRef](#)]
96. Zhang, L.; Zhuang, Z.; Fang, Q.; Wang, X. Study on the Automatic Identification of ABX<sub>3</sub> Perovskite Crystal Structure Based on the Bond-Valence Vector Sum. *Materials* **2022**, *16*, 334. [[CrossRef](#)]
97. Lu, S.; Zhou, Q.; Ouyang, Y.; Guo, Y.; Li, Q.; Wang, J. Accelerated discovery of stable lead-free hybrid organic-inorganic perovskites via machine learning. *Nat. Commun.* **2018**, *9*, 3405. [[CrossRef](#)]
98. Wu, Y.; Lu, S.; Ju, M.G.; Zhou, Q.; Wang, J. Accelerated design of promising mixed lead-free double halide organic-inorganic perovskites for photovoltaics using machine learning. *Nanoscale* **2021**, *13*, 12250–12259. [[CrossRef](#)]
99. Cai, X.; Zhang, Y.; Shi, Z.; Chen, Y.; Xia, Y.; Yu, A.; Xu, Y.; Xie, F.; Shao, H.; Zhu, H.; et al. Discovery of Lead-Free Perovskites for High-Performance Solar Cells via Machine Learning: Ultrabroadband Absorption, Low Radiative Combination, and Enhanced Thermal Conductivities. *Adv. Sci.* **2022**, *9*, 2103648. [[CrossRef](#)]
100. Gao, Z.; Zhang, H.; Mao, G.; Ren, J.; Chen, Z.; Wu, C.; Gates, I.D.; Yang, W.; Ding, X.; Yao, J. Screening for lead-free inorganic double perovskites with suitable band gaps and high stability using combined machine learning and DFT calculation. *Appl. Surf. Sci.* **2021**, *568*, 150916. [[CrossRef](#)]
101. Liu, H.; Feng, J.; Dong, L. Quick screening stable double perovskite oxides for photovoltaic applications by machine learning. *Ceram. Int.* **2022**, *48*, 18074–18082. [[CrossRef](#)]
102. Liu, W.; Lu, Y.; Wei, D.; Huo, X.; Huang, X.; Li, Y.; Meng, J.; Zhao, S.; Qiao, B.; Liang, Z.; et al. Screening interface passivation materials intelligently through machine learning for highly efficient perovskite solar cells. *J. Mater. Chem. A* **2022**, *10*, 17782–17789. [[CrossRef](#)]
103. She, C.; Huang, Q.; Chen, C.; Jiang, Y.; Fan, Z.; Gao, J. Machine learning-guided search for high-efficiency perovskite solar cells with doped electron transport layers. *J. Mater. Chem. A* **2021**, *9*, 25168–25177. [[CrossRef](#)]
104. Zhang, Z.; Wang, S.; Liu, X.; Chen, Y.; Su, C.; Tang, Z.; Li, Y.; Xing, G. Metal Halide Perovskite/2D Material Heterostructures: Syntheses and Applications. *Small Methods* **2021**, *5*, 2000937. [[CrossRef](#)]
105. Wang, H.P.; Li, S.; Liu, X.; Shi, Z.; Fang, X.; He, J.H. Low-Dimensional Metal Halide Perovskite Photodetectors. *Adv. Mater.* **2021**, *33*, 2003309. [[CrossRef](#)]
106. Misra, R.K.; Cohen, B.-E.; Iagher, L.; Etgar, L. Low-Dimensional Organic-Inorganic Halide Perovskite: Structure, Properties, and Applications. *ChemSusChem* **2017**, *10*, 3712–3721. [[CrossRef](#)] [[PubMed](#)]
107. Li, S.; Zhang, Y.; Yang, W.; Liu, H.; Fang, X. 2D Perovskite Sr<sub>2</sub>Nb<sub>3</sub>O<sub>10</sub> for High-Performance UV Photodetectors. *Adv. Mater.* **2020**, *32*, 1905443. [[CrossRef](#)]
108. Li, X.; Hoffman, J.M.; Kanatzidis, M.G. The 2D Halide Perovskite Rulebook: How the Spacer Influences Everything from the Structure to Optoelectronic Device Efficiency. *Chem. Rev.* **2021**, *121*, 2230–2291. [[CrossRef](#)] [[PubMed](#)]
109. Zhang, Z.-Z.; Guo, T.-M.; Li, Z.-G.; Gao, F.-F.; Li, W.; Wei, F.; Bu, X.-H. Machine learning assisted synthetic acceleration of Ruddlesden-Popper and Dion-Jacobson 2D lead halide perovskites. *Acta Mater.* **2023**, *245*, 118638. [[CrossRef](#)]
110. Lyu, R.; Moore, C.E.; Liu, T.; Yu, Y.; Wu, Y. Predictive Design Model for Low-Dimensional Organic-Inorganic Halide Perovskites Assisted by Machine Learning. *J. Am. Chem. Soc.* **2021**, *143*, 12766–12776. [[CrossRef](#)] [[PubMed](#)]
111. Hu, W.; Zhang, L.; Pan, Z. Designing Two-Dimensional Halide Perovskites Based on High-Throughput Calculations and Machine Learning. *ACS Appl. Mater. Interfaces* **2022**, *14*, 21596–21604. [[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.