machine learning & knowledge extraction

MDPI

Article

High-Throughput Ensemble-Learning-Driven Band Gap Prediction of Double Perovskites Solar Cells Absorber

Sabrina Djeradi [1], Tahar Dahame [1], Mohamed Abdelilah Fadla [2], Bachir Bentria [1], Mohammed Benali Kanoun [3] and Souraya Goumri-Said [4,*]

[1] Laboratoire de Physique des Matériaux, Université Amar Telidji de Laghouat, BP 37G, Laghouat 03000, Algeria; sa.djeradi@lagh-univ.dz (S.D.)
[2] School of Mathematics and Physics, Queen's University Belfast, Belfast BT7 1NN, UK; m.fadla@qub.ac.uk
[3] Department of Mathematics and Sciences, College of Humanities and Sciences, Prince Sultan University, P.O. Box 66833, Riyadh 11586, Saudi Arabia; mkanoun@psu.edu.sa
[4] Department of Physics, College of Science and General studies, Alfaisal University, P.O. Box 5092, Riyadh 11533, Saudi Arabia
* Correspondence: sosaid@alfaisal.edu; Tel.: +966-1-1215-8984

Abstract: Perovskite materials have attracted much attention in recent years due to their high performance, especially in the field of photovoltaics. However, the dark side of these materials is their poor stability, which poses a huge challenge to their practical applications. Double perovskite compounds, on the other hand, can show more stability as a result of their specific structure. One of the key properties of both perovskite and double perovskite is their tunable band gap, which can be determined using different techniques. Density functional theory (DFT), for instance, offers the potential to intelligently direct experimental investigation activities and predict various properties, including band gap. In reality, however, it is still difficult to anticipate the energy band gap from first principles, and accurate results often require more expensive methods such as hybrid functional or GW methods. In this paper, we present our development of high-throughput supervised ensemble learning-based methods: random forest, XGBoost, and Light GBM using a database of 1306 double perovskites materials to predict the energy band gap. Based on elemental properties, characteristics have been vectorized from chemical compositions. Our findings demonstrate the efficiency of ensemble learning methods and imply that scientists would benefit from recently employed methods in materials informatics.

Keywords: double perovskite; ensemble learning; band gap; Bayesian optimization; materials informatics

Citation: Djeradi, S.; Dahame, T.; Fadla, M.A.; Bentria, B.; Kanoun, M.B.; Goumri-Said, S. High-Throughput Ensemble-Learning-Driven Band Gap Prediction of Double Perovskites Solar Cells Absorber. *Mach. Learn. Knowl. Extr.* **2024**, *6*, 435–447. https://doi.org/10.3390/make6010022

Academic Editor: Luke E. K. Achenie

Received: 7 January 2024
Revised: 11 February 2024
Accepted: 13 February 2024
Published: 16 February 2024

1. Introduction

Perovskite materials have received a lot of attention in recent years due to their diverse composition, readily accessible synthetic conditions, and appealing features [1]. These features include strong optical absorption, long carrier diffusion length, low recombination, and high defect tolerance. These advantages are responsible for the outstanding power conversion efficiency of perovskite solar cells, which has increased from 3.8% to 25.7% in a few years. In contrast to silicon-based conventional solar cells, this impressive progress has marked perovskite solar cells as the start of the photovoltaic industry. Double perovskites AA'BB'X6 compounds have become a popular research topic due to their higher structural stability compared to halide perovskites, which are very sensitive to environmental stress factors like temperature, humidity, and oxygen [2–4]. The crystal structure of double perovskite is composed of two different cations, A and B, arranged in an alternating pattern with anion X in the middle (usually halogen or oxygen). The ionic radius of A is larger than that of B, however, X has a narrow radius, which results in a specific order of the cations in the crystal lattice [1,5]. The crystal structure of double perovskite materials is highly symmetrical and has been studied extensively to gain insight into its properties [6].

It consists of two octahedrons, namely BX6 and B′X6 [1]. This arrangement creates a unique three-dimensional network of atoms that has many important applications in materials science.

Traditional methods for predicting novel perovskite materials and identifying their underlying structure-property correlations involve first-principles computations, which yield optimal crystal structures, surfaces/interface structures, and optoelectronic properties [7–10]. However, the first-principles calculations typically require expensive computing resources, especially for a specific task such as accurate band gap, defect chemistry, interface properties, etc. Machine learning techniques have been developed in addition to the conventional trial-and-error experimental and computational routes to speed up the discovery of novel perovskite materials and reveal their hidden linkages. Due to the exponential growth of data, it is critical to adopt data-driven approaches like machine learning to extract helpful information from databases to accelerate the materials design process [8–15]. Accurate features that have a high correlation with the targeting properties are crucial for the machine learning process [16–19]. Utilizing the feature engineering method, the features—or occasionally descriptors—can be chosen. Once the input and output data for double perovskite materials are properly represented numerically [20–23], the mapping between inputs and outputs might be carried out using a machine learning method [2]. The energy band gap is a crucial feature that influences a perovskite material's ability to capture light and governs the performance of several optoelectronic devices, including solar cells [11]. It is extremely advantageous to be able to predict the band gaps of double perovskites rapidly and correctly for a range of practical applications. Although ML has the potential to replace quantum mechanical computations for high-fidelity band gaps since it requires dramatically less processing time [12].

In this paper, we introduce a high-throughput supervised machine learning model designed to predict the band gap energy of double perovskite materials. Leveraging a comprehensive database comprising 1306 materials, our model employs ensemble learning techniques to generate highly precise predictions. Our contributions extend beyond mere prediction precision. We have implemented a sophisticated approach that involves extracting features from the chemical composition of the materials using various featurization methods. These methods enable us to capture the nuanced attributes of each material, enhancing the robustness and accuracy of our prediction models. Moreover, our model incorporates relevance rankings for each material attribute, providing valuable insights into the factors influencing band gap energy. By elucidating the relative importance of different chemical attributes, our approach offers a deeper understanding of the underlying principles governing band gap behavior in double perovskite materials. The remaining sections are arranged as follows: in Section 2, we give our methodology providing our enormously accelerated machine learning development. Focusing and detailing next the ensemble learning models that have been used in our work. The simulation environment development employed to predict the energy band gap of double perovskite materials is shown in Section 3, along with an explanation of the results. Section 4 then brings this paper to a close.

2. Methodology

2.1. Machine Learning

Machine learning (ML) is a burgeoning interdisciplinary field, seamlessly blending computer science, statistics, mathematics, and engineering. Its primary function lies in constructing statistical models for data analysis and prediction, fundamentally altering the landscape of uncovering hidden relationships without explicit human programming [1]. Within ML, various techniques thrive, including reinforcement, semi-supervised, supervised, and unsupervised learning. Supervised learning relies on input and labeled output training data, while unsupervised learning adeptly processes unlabeled data, pushing the boundaries of autonomous learning [2].

To harness the expansive potential of databases for meaningful insights, the identification of descriptors related to targeted qualities, performances, and applications become paramount [2]. Notably, machine learning exhibits remarkable efficiency in rapidly discovering potential candidates for solar cell materials [13]. In our specific focus on predicting material band gaps, machine learning emerges as a reliable tool, surpassing quantum mechanical computations in fidelity [12].

The essence of our study lies in crafting a model through supervised machine learning that establishes a discernible relationship between input features and target properties [24–27]. This approach enables the accurate prediction of values for previously unknown materials, marking a pivotal step toward advancing material discovery and innovation. Our objective is clear: to unravel hidden patterns and correlations within the data, allowing us to predict new values for materials based on the established relationships. This not only streamlines the materials discovery process but also opens avenues for unprecedented advancements in various scientific and technological domains. As we navigate the intricate landscape of machine learning, our pursuit is not merely predictive accuracy but a deeper understanding of the underlying principles that govern material behavior—a quest that holds immense promise for the future of materials science and technological innovation.

2.2. Ensemble Learning Models

2.2.1. Random Forest

Random Forest (RF), a widely embraced ensemble learning method introduced by Leo Breiman [28], draws inspiration from the groundwork laid by Amit and Geman [29]. Devised as a countermeasure to boosting, Random Forests extend Breiman's bagging theory [30]. In its training process, RF meticulously constructs an array of decision trees, ultimately outputting the class that represents the mode of the classes—this applies to both continuous responses, termed "regression," and categorical responses referred to as "classification". The central aim of employing RF is to amalgamate numerous decision trees into a cohesive model, mitigating overfitting concerns and elevating the overall accuracy of the model [31,32]. By harnessing the collective power of diverse decision trees, RF not only fortifies the model against overfitting but also enhances its predictive prowess, making it a robust choice for various applications. Breiman's innovative approach with Random Forests has significantly influenced the landscape of ensemble learning, offering a versatile and effective tool for predictive modeling in both regression and classification scenarios.

2.2.2. XGBoost

XGBoost, an abbreviation for extreme gradient boosting, stands out as an exceptionally potent and user-friendly optimized distributed gradient boosting tool [33]. It was proposed by Chen and Guestrin in 2016, showcasing a remarkable evolution in the realm of gradient boosting machines (GBM) [34]. Rooted in the gradient boosting machine technique, XGBoost is a versatile approach widely employed for developing predictive models in both regression and classification tasks [35]. The core principle of boosting revolves around enhancing model accuracy by amalgamating multiple low-accuracy trees. Each iteration introduces a new tree to the model, a process guided by the gradient boosting machine, a renowned method for generating these trees, initially conceptualized by Friedman [35]. XGBoost's prowess lies in its ability to sequentially refine and augment the predictive capabilities of the model through an ensemble of trees. This iterative process, guided by gradient boosting principles, not only bolsters accuracy but also enables the model to capture intricate patterns within the data. Chen and Guestrin's introduction of XGBoost has significantly advanced the landscape of gradient boosting, providing an accessible yet powerful tool for predictive modeling that resonates across diverse applications in both regression and classification scenarios.

The term "eXtreme Gradient Boosting" refers to a technique that was created by [36] and is frequently implemented to classification and regression applications. Classification and regression trees (CARTs) form the approach's core. The strategy relies on an ensemble

of decision trees, where each tree gains knowledge from the residuals of the previous tree. The total of all trained classification and regression trees (CARTs) yields the final prediction, $\hat{y}_i$:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i),\ f_k \in F \tag{1}$$

The model's objective involves a loss function $L(\theta)$ and a penalty term $\Omega(\theta)$ to account for model complexity, where K denotes the number of trees, $f_k(x_i)$ is the output of a single tree, and F refers to the set of all potential CARTs. The loss function computes the discrepancy between the predicted and actual values by summing up $\sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t)}\right)$. The tree's complexity is determined by the term $\omega(f)$, which depends on the number of leaves (T), the leaf scores (ω), the leaf penalty coefficient (γ), and a regularization parameter (λ) that limits the maximum size of the leaf scores.

$$L(\theta) = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t)}\right) \tag{2}$$

$$\omega(\theta) = \sum_{i=1}^{n} \omega(f_i) \tag{3}$$

$$\omega(f) = \Upsilon T + \frac{1}{2}\lambda \sum_{j=1}^{T} \omega_j^2 \tag{4}$$

$$Obj = L(\theta) + \Omega(\theta) \tag{5}$$

The model's processing is additive in nature, whereby the t-th prediction $\hat{y}_i^{(t)}$ is the result of adding the prediction of the new tree $f_t(x_i)$ to the prediction from the previous iteration, $\hat{y}_i^{(t-1)}$.

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + f_t(x_i) \tag{6}$$

The objective that incorporates the updates described in Equations (5) and (6) can be expressed as follows:

$$Obj = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)\right) + \Omega(\theta) \tag{7}$$

The first and second derivatives represented by g_i and h_i. Respectively, can be utilized in Equation (9) to approximate the objective through the second-order Taylor expansion of the loss function.

$$Obj = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t-1)} + g_i f_t(x_i) + \frac{1}{2}\right) + \Omega(\theta) \tag{8}$$

$$g_i = \partial \hat{y}_i^{(t-1)}\, l\left(y_i, \hat{y}_i^{(t-1)}\right) \tag{9}$$

$$h_i = \partial^2 \hat{y}_i^{(t-1)}\, l\left(y_i, \hat{y}_i^{(t-1)}\right) \tag{10}$$

In the process of building each decision tree, a gain parameter called "*Gain*" is computed to determine whether a leaf node should be split into multiple sub-nodes. The calculation is performed as follows.

$$Gain = \frac{1}{2}\left[\frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} + \frac{(G_L + G_R)^2}{H_L + H_R + \lambda}\right] - \gamma \tag{11}$$

$$G = \sum_{i \in I} g_i \tag{12}$$

$$H = \sum_{i \in I} h_i \tag{13}$$

In the given equation, G_L and H_L correspond to the outputs for the left leaves, while G_R and H_R represent the outputs for the right leaves. λ is a regularization parameter that is used. Additionally, the tree is pruned based on the value of parameter γ. If the calculated gain for a leaf is less than γ, then the leaf is not divided any further.

2.2.3. Light GBM

Light GBM, a versatile decision tree-based iterative approach, showcases its prowess in both classification and regression through the gradient boosting decision tree (GBDT). What sets it apart is its distinctive leaf-wise tree construction methodology, diverging from the conventional level-wise approach employed by other boosting frameworks like XGBoost. This deviation proves advantageous when dealing with large datasets and feature spaces with numerous dimensions, as it optimizes scalability and efficiency. In contrast to level-wise construction that tends to focus on all potential feature splits, the leaf-wise strategy adopted by Light GBM centers on the most critical features. This not only simplifies the model but also enhances its effectiveness by prioritizing key attributes. The traditional approach of enumerating all potential feature splits and sorting feature values can be resource-intensive in terms of time and memory consumption. LightGBM introduces an innovative solution to this challenge through an enhanced histogram algorithm [34]. This technique significantly streamlines the process, ensuring a more efficient utilization of resources and enabling the model to handle large and complex datasets with heightened agility. By incorporating the leaf-wise construction strategy and the enhanced histogram algorithm, Light GBM offers a compelling solution for boosting frameworks. Its scalability, efficiency, and focus on crucial features make it particularly well suited for applications dealing with substantial datasets and intricate feature spaces, solidifying its position as a valuable tool in both classification and regression scenarios. The innovative approach adopted by Light GBM marks a significant contribution to the optimization of decision tree-based methods in the realm of machine learning.

3. Results and Discussions

3.1. Data Acquisition

This research study was primarily driven by the ambition to advance the prediction of solid-state band gaps, focusing specifically on ensemble learning models. The dataset employed consisted of 1306 double perovskite band gaps, computed using the GLLBSC (Gritsenko, van Leeuwen, van Lenthe, and Baerends solid correlation) functional [37]. To unravel the complex relationship between chemical composition and energy band gap, we employed a composition-based featurizer that converted chemical formulas into numerical values, enabling the application of machine learning algorithms. The development process is succinctly depicted in Figure 1, illustrating the intricate flowchart of creating machine-learning models dedicated to predicting energy band gaps. Given the inherently unlearnable nature of composition data, the featurizer played a pivotal role in transforming these chemical nuances into comprehensible numerical inputs. To ensure robust model training and evaluation, we randomly partitioned the dataset into an 80% training set and a 20% testing set. Each model underwent training on the designated training dataset and subsequent evaluation on the independent testing dataset. This meticulous approach aimed at establishing reliable predictive models for energy band gaps in double perovskite materials, paving the way for more accurate and efficient predictions. The utilization of ensemble learning techniques, coupled with a thoughtful dataset split and featurization strategy, underscores the commitment to precision and reliability in predicting the critical solid-state band gaps.

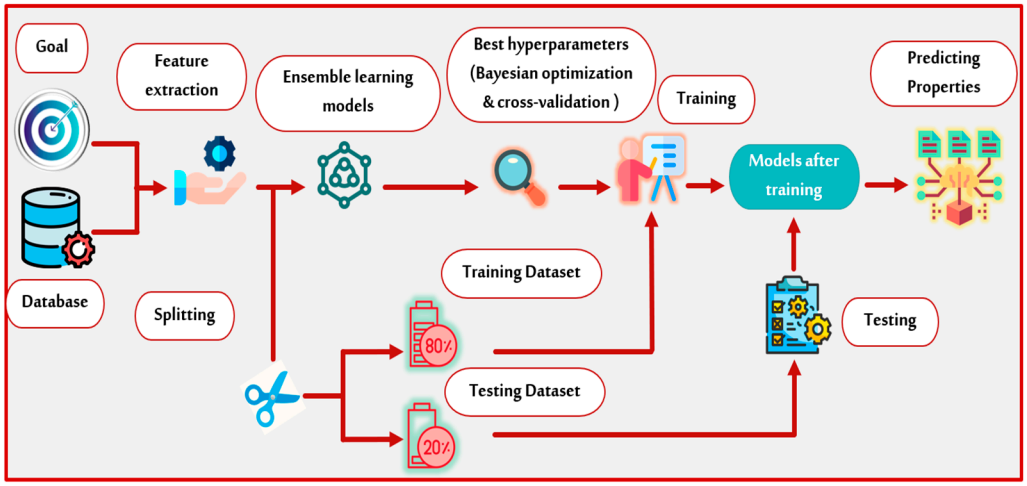


Figure 1. Flowchart of the development of the machine-learning algorithms for the prediction of the energy band gap.

As a first step, we will go through the specifics of the double perovskite bandgap database utilized here to train, verify, and evaluate the machine learning models that we have developed in this study. The Computational Materials Repository (CMR) was the source of the data utilized in the present research. The 53 stable cubic perovskite oxides that were discovered in order to possess a finite bandgap in a prior screening using single perovskites were merged to create the double perovskite structures presented in this dataset. The chemical species of these 53 parents single perovskites are presented in Table 1, highlighting the cations present in the A-site and/or the B-site.

Table 1. Chemical species of the 53 parents' single perovskites.

A-Site Cations	B-Site Cations	A- & B-Site Cations
Ag, Ba, Ca, Cs, K, La, Li, Mg, Na, Pb, Rb, Sr, Tl, Y	Al, Hf, Nb, Sb, Sc, Si, Ta, Ti, V, Zr	Ga, Ge, In, Sn

3.2. Features Extraction

In order to create alternative features that are more important, feature extraction modifies the existing features in some way. Indeed, a vector of elements and fractions is used to originally express compositions. The fractional vector holds the corresponding fractions for every element inside the compound, whereas the element vector holds the chemical's specific atomic numbers. These vectors have been encrypted in a way that makes it possible for machine learning to employ them [38]. Featurization is an important area of study that allows us to create a composition-based feature vector (CBFV) and unable to depict materials depending on carefully selected element attributes [39]. In this work, we combined three powerful methods: Jarvis [40], magpie [41], and oliynyk [42] from composition-based feature vector (CBFV) [43–45] to create element vectors based on every elemental characteristic as shown in Figure 2. Once the feature descriptor set was constructed and determined, we have deleted the features characterized with null and outlier values. In addition, feature descriptors with such a variance of 0, are mentioned as needing to be removed too. Then, for the purpose of identifying the characteristics with such a strong correlation and subsequently getting rid of many collinearities between the features, Pearson correlation coefficients are also computed. It should be highlighted here that we have dropped the feature descriptors with a correlation of 1.

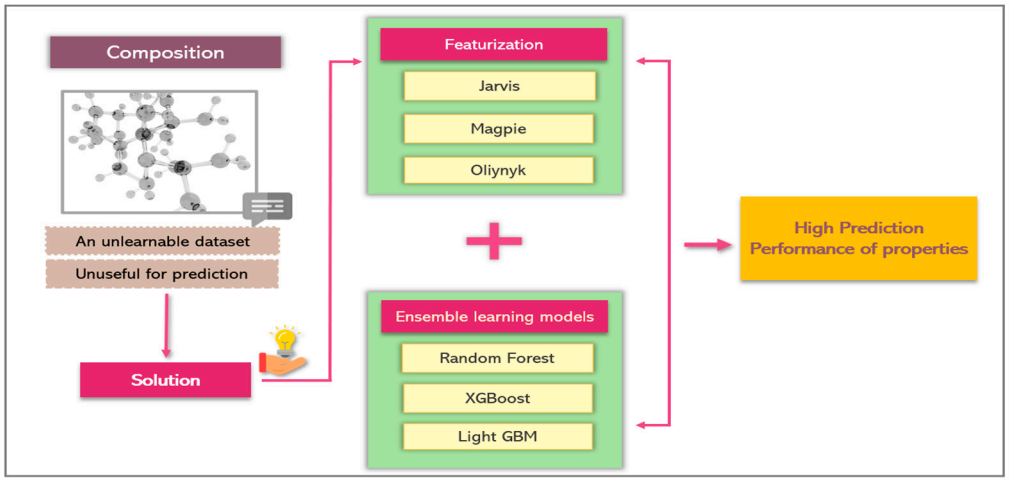


Figure 2. Diagram of proposed methodology of feature extraction and model selection for the prediction of the energy band gap.

Featurization is a procedure by which the training data is changed into learnable data that can differentiate between various materials (usually in the form of vectors or tensors). These numbers are commonly called descriptors, characteristics, or fingerprints. Hence, the process of choosing acceptable input features is necessary to develop a meaningful ML model and obtain good prediction performance. Finding the important traits that are closely related to the target properties is essential before model creation [13].

3.3. Model Selection

A crucial factor in performance prediction is the choice of a suitable ML model, which has a big impact on the results. In order to prove the efficiency of ensemble learning methods, we employed three methods for the prediction of the energy band gap of double perovskites which are: Random Forest, XGBoost, and Light GBM. Moreover, to prevent both overfitting and underfitting, the model algorithm's inner hyper-parameters must typically be optimized. Determining the best hyperparameters and the accuracy of the model and the datasets is considered an important step before making predictions of properties [1]. For this purpose, we have used Bayesian optimization and identified the best hyperparameters for both models and data. Drawing a learning curve helped to identify the appropriate parameters for further optimizing the models with the strongest generalization model. Table 2 below presents the results of the hyperparameters tuning for the three selected machine learning models.

Table 2. Hyperparameters tuning for the three selected machine learning models.

Models	Random Forest	XGBoost	Light GBM
Hyperparameters	• random_state = 24, • n_estimators = 11,990, • max_features = 255, • min_samples_leaf = 1, • min_samples_split = 2, • max_depth = 100	• random_state = 24, • n_estimators = 123, • max_features = 89, • max_depth = 34, • gamma = 1, learning_rate = 0.33	• random_state = 24 • learning_rate = 0.09, • min_child_samples = 59, • n_estimators = 5933, • num_leaves = 91, • reg_alpha = 2.0, • reg_lambda = 4.0, • subsample = 0.66

These parameters were required since the decision tree showed erratic behavior in order to guarantee consistency between runs' output. The generalization potential of the ensemble learning models has been enhanced by hyperparameter optimization.

3.4. Model Developing

Following the construction of the learning curve and after the hyperparameters tuning, we obtained the prediction results for each model on the energy band gap for double perovskites materials. The results were depicted in Figure 3. The Random Forest regression model achieved a cross-validation accuracy of 92.1% on the constructed data set with a low mean absolute error (MAE) equal to 0.320 eV. Considering the gradient-boosting regression models, the best results were obtained by Light GBM, which had a high R2 of 0.934 and a low mean absolute error (MAE) of 0.302 eV. However, the XGBoost model offered a high prediction performance of 0.911 with a low mean absolute error (MAE) of 0.350 eV.

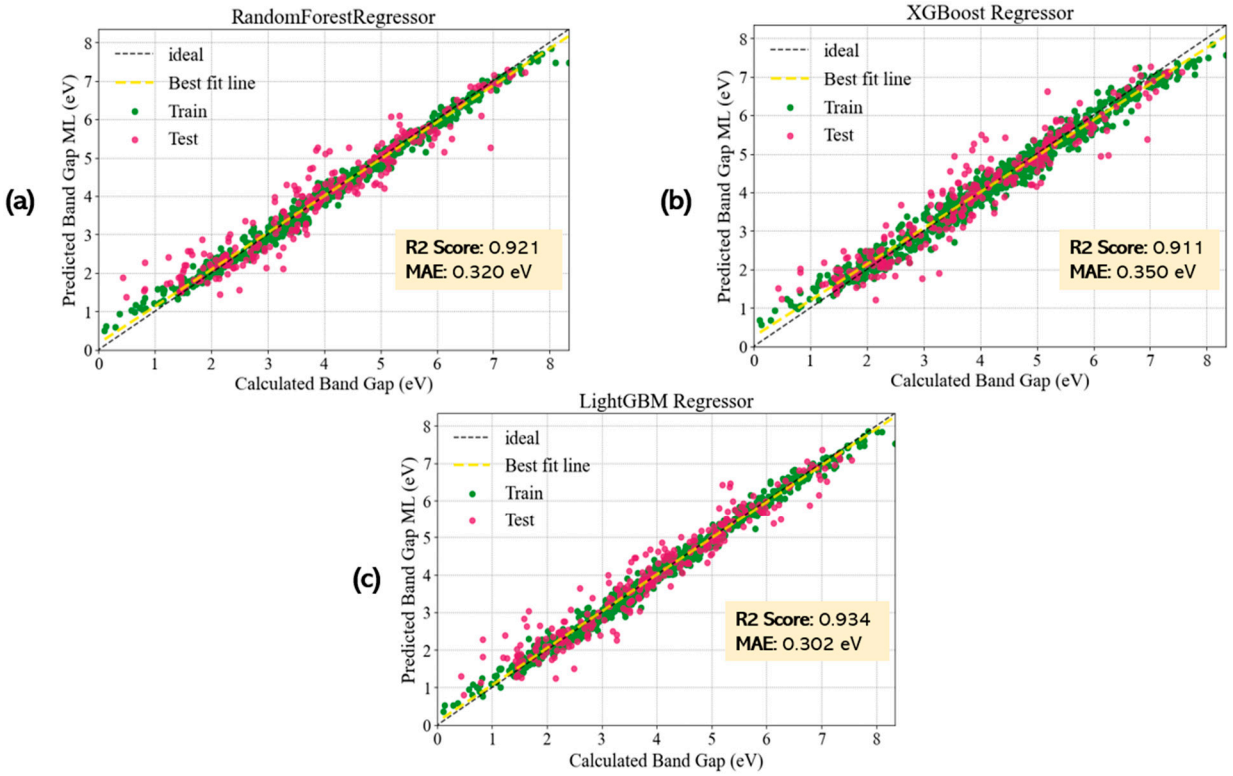


Figure 3. Prediction results of the energy band gap for double perovskite materials for the three ensemble learning models. The training dataset is represented by green-filled circles, whereas the test dataset is represented by pink-filled circles. The reference line that corresponds to the perfect line is indicated by a yellow dashed line. (**a**) Random Forest regression model, (**b**) XGBoost, and (**c**) Light GBM.

Many factors can be considered to explain why predictive models have such high accuracy. One of the main differences is the nonlinearity of ensemble learning compared to other models. Additionally, the use of multiple materials' structural compositions in the dataset is another factor for such high accuracy [46].

In predictive models, feature significance scores are crucial since they offer insight into the data and the model. They are also important for dimensionality reduction and feature selection, which can increase the efficiency of the predictive model. The fundamental benefit of the decision tree-based methods is that they produce feature importance scores, which are useful to extract important features that determine the target attributes and aid in the understanding of the results in addition to producing extremely accurate predictive models. Figure 4 shows the top 10 feature importance scores of ensembles learning models of the energy band gap of double perovskite materials. The importance score is subtly different but still falls within a range that is acceptable since the feature importance of various models can indeed be calculated in different ways.

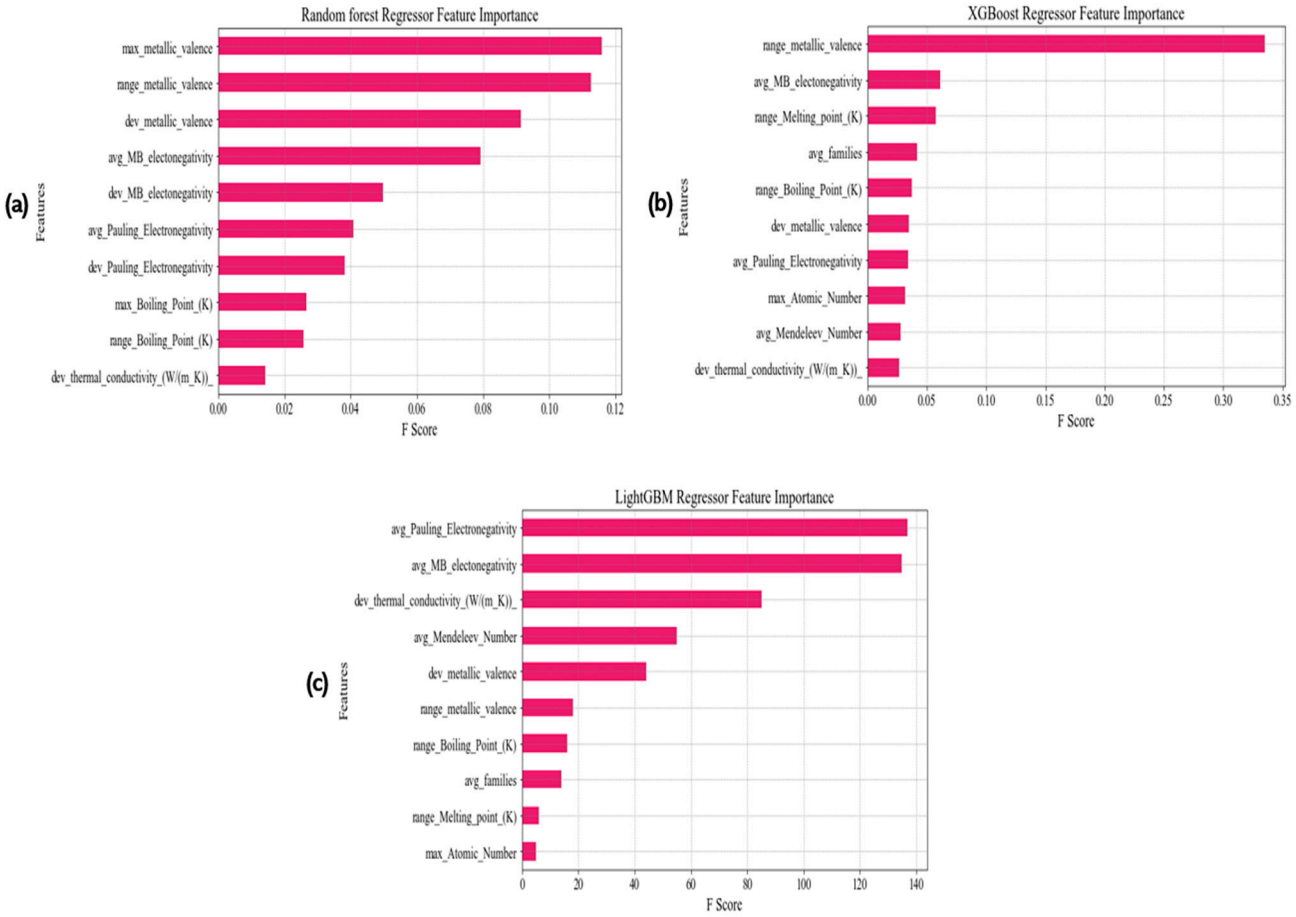


Figure 4. Feature importance contributed to the three ensemble learning models of the energy band gap for double perovskite materials as found by (**a**) Random Forest regression model, (**b**) XGBoost, and (**c**) Light GBM.

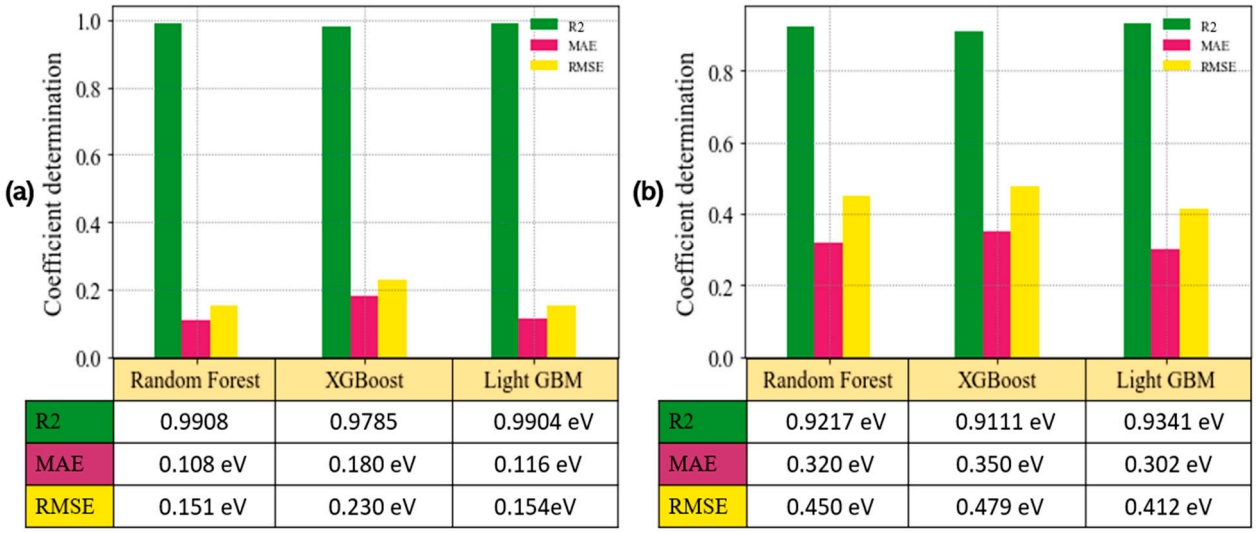
The distribution of feature importance utilized by the Random Forest model is more erratic (Figure 4a). This suggests a more complicated link between the energy band gap and the material properties. The ensuing features, such as range_metallic_valence, dev_metallic_valence, Avg_MB_electronegativity, dev_MB_electronegativity, Avg_pauling_electronegativity, or dev_pauling_electronegativity, are not insignificant, despite the fact that max_metallic_valence represents the most significant feature. The same thing for the XGBoost model in (Figure 4b). However, range_ metallic_ valence is the most relevant feature, the ensuing features, including Avg_MB_electronegativity, range_Melting_point_(K), avg_Families, range_Boiling_point_(K), dev _metallic _valence, or Avg _Pauling _electronegativity, are not inconsequential.

In contrast, there are two most crucial features including the Light GBM model, Avg_pauling_electronegativity and Avg_MB_electronegativity, are made clear (Figure 4c) Surprisingly, the relevance scores of features Avg_pauling_electronegativity and Avg_MB_electronegativity are more than twice as high as that of feature dev_thermal_conductivity_(W/(m·K)), which is rated third. The relevance scores for the features after the initial three sharply decline, showing that only a few material features are important to the energy band gap.

3.5. Model Evaluation

The primary goal of machine learning is to accurately predict unknown subsets based on already-known data. It is inevitable that there will be certain statistical errors in the computation, and these faults should be examined logically and assessed in the

model evaluation procedure for the model's application afterward. Independent testing, cross-validation, and bootstrapping are three approaches for evaluating models that are frequently employed [6]. After training, all ensemble learning models were validated using the cross-validation approach with a 10-fold on 80 percent of the total dataset increase to make sure they weren't overfitting. Three evaluation indicators - RMSE, R2, and MAE - are represented by a histogram. Indeed, the prediction performances of ensemble learning models on the train set and test set for the energy band gap of double perovskite materials are represented in Figure 5.



	Random Forest	XGBoost	Light GBM
R2	0.9908	0.9785	0.9904 eV
MAE	0.108 eV	0.180 eV	0.116 eV
RMSE	0.151 eV	0.230 eV	0.154eV

	Random Forest	XGBoost	Light GBM
R2	0.9217 eV	0.9111 eV	0.9341 eV
MAE	0.320 eV	0.350 eV	0.302 eV
RMSE	0.450 eV	0.479 eV	0.412 eV

Figure 5. Feature importance of the energy band gap for double perovskite materials. (**a**) On the training set. (**b**) On the testing set.

When the three models in the testing set were compared, the Light GBM model performance was found to be the best, achieving the top high predictive accuracy which makes us say that is a good proposal for the prediction of the energy band gap properties. Also, the Random Forest model worked well providing good results. Likewise, in the training set, here the two recent models in the testing set have switched roles. We can see that the Random Forest model was the best and the top corresponding to the highly predictive results. However, the Light GBM didn't work as well as the testing set. In contrast, we could define the top predictive model to help researchers guide, especially in the prediction of this type of target property, to pick the best model since it is the key to proving the efficiency of models and has a big impact on the results. We could explore the possibility of using similar ensemble learning methods to predict other important material properties beyond band gaps, such as electronic and optical properties. This could provide a more comprehensive understanding of material behavior.

4. Conclusions

The band gap energy is one of the most important features that influence a perovskite material's ability to capture light. In this work, we built and developed high-throughput supervised ensemble learning-based methods such as random forest, XGBoost, and Light GBM that accurately and rapidly predict the band gap of double perovskite materials based on their chemical composition. By utilizing these featurization tools based on elemental properties (provided by Jarvis, magpie, and oliynyk), we gained valuable insights into the composition of a material and used them to make predictions. It is clear that ensemble learning methods have the capacity to predict the band gap of double perovskite compounds accurately, quickly, and broadly using their elemental information data as inputs. Comparing the efficiency of bagging ensemble learning methods, such as Random Forest, with the two boosting ensemble learning methods, such as XGBoost and Light GBM, the

energy band gap was most accurately predicted using the Light GBM approach. In addition to having benefits when dealing with data sparsity and weighting instances, Light GBM makes good use of both distributed and parallel processing. These findings demonstrate the efficiency of ensemble learning methods and how machine learning can rapidly and effectively be used for materials discovery only based on chemical composition. To advance this work, it would be valuable to extend the models to predict other material properties beyond band gaps, incorporating structural information, and exploring transfer learning for diverse datasets. Integrating uncertainty quantification into predictions, developing user-friendly interactive tools, and considering environmental factors could enhance the practicality of these models for materials scientists. Collaboration with experimentalists to validate predictions and exploring data augmentation techniques for limited datasets would further refine the models' accuracy. Additionally, experimenting with alternative machine learning architectures or hybrid models may offer avenues for improvement, ultimately contributing to a more comprehensive and efficient approach to materials discovery based on chemical composition.

Author Contributions: Conceptualization, S.D., T.D., M.A.F., B.B. and S.G.-S.; methodology, S.D., T.D., M.A.F., B.B., M.B.K. and S.G.-S.; software, S.D., T.D., M.A.F., B.B., M.B.K. and S.G.-S.; validation, S.D., T.D., M.A.F., B.B., M.B.K. and S.G.-S.; formal analysis, S.D., T.D., M.A.F., B.B., M.B.K. and S.G.-S.; investigation, S.D., T.D., M.A.F., B.B., M.B.K. and S.G.-S.; resource, S.D., T.D., M.A.F., B.B., M.B.K. and S.G.-S.; data curation, S.D., T.D., M.A.F., B.B., M.B.K. and S.G.-S.; writing—original draft preparation writing—review and editing, S.D., T.D., M.A.F., B.B., M.B.K. and S.G.-S. visualization, S.D., T.D., M.A.F., B.B., M.B.K. and S.G.-S.; supervision, S.D., T.D., M.A.F., B.B., M.B.K. and S.G.-S.; project administration, S.D., T.D., M.A.F., B.B., M.B.K. and S.G.-S.; funding acquisition, S.D., T.D., M.A.F., B.B., M.B.K. and S.G.-S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The data used to support the findings of this study are available from the corresponding authors upon request.

Acknowledgments: Author S.G.-S. thanks Alfaisal University Research Office for supporting her research.

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

References

1. Tao, Q.; Xu, P.; Li, M.; Lu, W. Machine learning for perovskite materials design and discovery. *npj Comput. Mater.* **2021**, *7*, 23. [CrossRef]
2. Zhang, L.; He, M.; Shao, S. Machine learning for halide perovskite materials. *Nano Energy* **2020**, *78*, 105380. [CrossRef]
3. Anand, D.V.; Xu, Q.; Wee, J.; Xia, K.; Sum, T.C. Topological feature engineering for machine learning based halide perovskite materials design. *npj Comput. Mater.* **2022**, *8*, 203. [CrossRef]
4. Yılmaz, B.; Yıldırım, R. Critical review of machine learning applications in perovskite solar research. *Nano Energy* **2021**, *80*, 105546. [CrossRef]
5. Travis, W.; Glover, E.N.K.; Bronstein, H.; Scanlon, D.O.; Palgrave, R.G. On the application of the tolerance factor to inorganic and hybrid halide perovskites: A revised system. *Chem. Sci.* **2016**, *7*, 4548–4556. [CrossRef]
6. Saha-Dasgupta, T. Magnetism in Double Perovskites. *J. Supercond. Nov. Magn.* **2013**, *26*, 1991–1995. [CrossRef]
7. Azam, S.; Khan, S.A.; Goumri-Said, S.; Kanoun, M.B. Predicted Thermoelectric Properties of the Layered XBi4S7 (X = Mn, Fe) Based Materials: First Principles Calculations. *J. Electron. Mater.* **2017**, *46*, 23–29. [CrossRef]
8. Alhashmi, A.; Kanoun, M.B.; Goumri-Said, S. Machine Learning for Halide Perovskite Materials ABX_3 (B = Pb, X = I, Br, Cl) Assessment of Structural Properties and Band Gap Engineering for Solar Energy. *Materials* **2023**, *16*, 2657. [CrossRef]
9. Kanoun, M.B.; Goumri-Said, S. Insights into the impact of Mn-doped inorganic CsPbBr3 perovskite on electronic structures and magnetism for photovoltaic application. *Mater. Today Energy* **2021**, *21*, 100796. [CrossRef]
10. Fadla, M.A.; Bentria, B.; Benghia, A.; Dahame, T.; Goumri-Said, S. Insights on the opto-electronic structure of the inorganic mixed halide perovskites γ-CsPb(I1-xBrx)3 with low symmetry black phase. *J. Alloys Compd.* **2020**, *832*, 154847. [CrossRef]
11. Gladkikh, V.; Kim, D.Y.; Hajibabaei, A.; Jana, A.; Myung, C.W.; Kim, K.S. Machine Learning for Predicting the Band Gaps of ABX_3 Perovskites from Elemental Properties. *J. Phys. Chem. C* **2020**, *124*, 8905–8918. [CrossRef]
12. Pilania, G.; Mannodi-Kanakkithodi, A.; Uberuaga, B.P.; Ramprasad, R.; Gubernatis, J.E.; Lookman, T. Machine learning bandgaps of double perovskites. *Sci. Rep.* **2016**, *6*, 19375. [CrossRef]

13. Chen, C.; Zuo, Y.; Ye, W.; Li, X.; Deng, Z.; Ong, S.P. A Critical Review of Machine Learning of Energy Materials. *Adv. Energy Mater.* **2020**, *10*, 1903242. [CrossRef]
14. Zebarjadi, M.; Esfarjani, K.; Dresselhaus, M.S.; Ren, Z.F.; Chen, G. Perspectives on thermoelectrics: From fundamentals to device applications. *Energy Environ. Sci.* **2011**, *5*, 5147–5162. [CrossRef]
15. Curtarolo, S.; Hart, G.L.; Nardelli, M.B.; Mingo, N.; Sanvito, S.; Levy, O. The high-throughput highway to computa-tional materials design. *Nat. Mater.* **2013**, *12*, 191. [CrossRef] [PubMed]
16. Mounet, N.; Gibertini, M.; Schwaller, P.; Campi, D.; Merkys, A.; Marrazzo, A.; Sohier, T.; Castelli, I.E.; Cepellotti, A.; Pizzi, G.; et al. Two-dimensional materials from high-throughput computational exfoliation of experimentally known compounds. *Nat. Nanotechnol.* **2018**, *13*, 246–252. [CrossRef] [PubMed]
17. 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]
18. Kirklin, S.; Saal, J.E.; Meredig, B.; Thompson, A.; Doak, J.W.; Aykol, M.; Rühl, S.; Wolverton, C. The open quantum materials database (oqmd): Assessing the accuracy of dft formation energies. *npj Comput. Mater.* **2015**, *1*, 15010. [CrossRef]
19. Khan, W.; Goumri-Said, S. Exploring the optoelectronic structure and thermoelectricity of recent photoconductive chalcogenides compounds, $CsCdInQ_3$ (Q = Se, Te). *RSC Adv.* **2015**, *5*, 9455–9461. [CrossRef]
20. Azam, S.; Khan, S.A.; Goumri-Said, S. Optoelectronic and Thermoelectric Properties of Bi2OX2 (X = S, Se, Te) for Solar Cells and Thermoelectric Devices. *J. Electron. Mater.* **2018**, *47*, 2513–2518. [CrossRef]
21. Azam, S.; Goumri-Said, S.; Khan, S.A.; Ozisik, H.; Deligoz, E.; Kanoun, M.B.; Khan, W. Electronic structure and related optical, thermoelectric and dynamical properties of Lilianite-type Pb7Bi4Se13: Ab-initio and Boltzmann transport theory. *Materialia* **2020**, *10*, 100658. [CrossRef]
22. Goumri-Said, S. Probing Optoelectronic and Thermoelectric Properties of Lead-Free Perovskite $SnTiO_3$: HSE06 and Boltzmann Transport Calculations. *Crystals* **2022**, *12*, 1317. [CrossRef]
23. 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]
24. Wang, Z.; Yang, M.; Xie, X.; Yu, C.; Jiang, Q.; Huang, M.; Algadi, H.; Guo, Z.; Zhang, H. Applications of machine learning in perovskite materials. *Adv. Compos. Hybrid Mater.* **2022**, *5*, 2700–2720. [CrossRef]
25. 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]
26. Feng, Y.; Chen, D.; Niu, M.; Zhong, Y.; Ding, H.; Hu, Y.; Wu, X.-F.; Yuan, Z.-Y. Recent progress in metal halide perovskite-based photocatalysts: Physicochemical properties, synthetic strategies, and solar-driven applications. *J. Mater. Chem. A* **2023**, *11*, 22058–22086. [CrossRef]
27. Zhang, L.; Mei, L.; Wang, K.; Lv, Y.; Zhang, S.; Lian, Y.; Liu, X.; Ma, Z.; Xiao, G.; Liu, Q.; et al. Advances in the Application of Perovskite Materials. *Nano-Micro Lett.* **2023**, *15*, 177. [CrossRef]
28. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
29. Amit, Y.; Geman, D. Shape Quantization and Recognition with Randomized Trees. *Neural Comput.* **1997**, *9*, 1545–1588. [CrossRef]
30. Breiman, L. Bagging predictors. *Mach. Learn.* **1996**, *24*, 123–140. [CrossRef]
31. Asselman, A.; Khaldi, M.; Aammou, S. Enhancing the prediction of student performance based on the machine learning XGBoost algorithm. *Interact. Learn. Environ.* **2021**, *31*, 3360–3379. [CrossRef]
32. Cutler, A.; Cutler, D.R.; Stevens, J.R. Random forests. In *Ensemble Machine Learning*, 2nd ed.; Zhang, C., Ma, Y.Q., Eds.; Springer: New York, NY, USA, 2012; pp. 157–175.
33. Chen, T.; He, T.; Benesty, M.; Khotilovich, V.; Tang, Y.; Cho, H.; Chen, K.; Mitchell, R.; Cano, I.; Zhou, T. Xgboost: Extreme Gradient Boosting. Package Version-0.4-1.4. 2015. Available online: https://xgboost.ai/ (accessed on 15 May 2023).
34. 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; ACM: New York, NY, USA, 2016; pp. 785–794. [CrossRef]
35. Zhang, L.; Zhan, C. Machine Learning in Rock Facies Classification: An Application of XGBoost. In Proceedings of the International Geophysical Conference, Qingdao, China, 17–20 April 2017; Society of Exploration Geophysicists: Qingdao, China, 2017; pp. 1371–1374. [CrossRef]
36. Xgboost Developers. XGboost Parameter Documentation. 2023. Available online: https://xgboost.readthedocs.io/en/stable/parameter.html (accessed on 9 April 2023).
37. Song, J.; Liu, G.; Jiang, J.; Zhang, P.; Liang, Y. Prediction of Protein–ATP Binding Residues Based on Ensemble of Deep Convolutional Neural Networks and LightGBM Algorithm. *Int. J. Mol. Sci.* **2021**, *22*, 939. [CrossRef] [PubMed]
38. Gritsenko, O.; van Leeuwen, R.; van Lenthe, E.; Baerends, E.J. Self-consistent approximation to the Kohn-Sham exchange potential. *Phys. Rev. A* **1995**, *51*, 1944–1954. [CrossRef] [PubMed]
39. Falkowski, A.R.; Kauwe, S.K.; Sparks, T.D. Optimizing Fractional Compositions to Achieve Extraordinary Properties. *Integrating Mater. Manuf. Innov.* **2021**, *10*, 689–695. [CrossRef]
40. Murdock, R.J.; Kauwe, S.K.; Wang, A.Y.-T.; Sparks, T.D. Is Domain Knowledge Necessary for Machine Learning Materials Properties? *Integr. Mater. Manuf. Innov.* **2020**, *9*, 221–227. [CrossRef]

41. Choudhary, K.; DeCost, B.; Tavazza, F. Machine learning with force-field-inspired descriptors for materials: Fast screening and mapping energy landscape. *Phys. Rev. Mater.* **2018**, *2*, 083801. [CrossRef] [PubMed]
42. Ward, L.; Agrawal, A.; Choudhary, A.; Wolverton, C. A general-purpose machine learning framework for predicting properties of inorganic materials. *npj Comput. Mater.* **2016**, *2*, 16028. [CrossRef]
43. Oliynyk, A.O.; Antono, E.; Sparks, T.D.; Ghadbeigi, L.; Gaultois, M.W.; Meredig, B.; Mar, A. High-Throughput Machine-Learning-Driven Synthesis of Full-Heusler Compounds. *Chem. Mater.* **2016**, *28*, 7324–7331. [CrossRef]
44. Kauwe, S.K.; Welker, T.; Sparks, T.D. Extracting Knowledge from DFT: Experimental Band Gap Predictions Through Ensemble Learning. *Integrating Mater. Manuf. Innov.* **2020**, *9*, 213–220. [CrossRef]
45. Graser, J.; Kauwe, S.K.; Sparks, T.D. Machine Learning and Energy Minimization Approaches for Crystal Structure Predictions: A Review and New Horizons. *Chem. Mater.* **2018**, *30*, 3601–3612. [CrossRef]
46. Im, J.; Lee, S.; Ko, T.-W.; Kim, H.W.; Hyon, Y.; Chang, H. Identifying Pb-free perovskites for solar cells by machine learning. *npj Comput. Mater.* **2019**, *5*, 37. [CrossRef]