Materials Science in Semiconductor Processing 161 (2023) 107427

Contents lists available at ScienceDirect

Materials Science in Semiconductor Processing

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

Explainable machine learning for predicting the band gaps of ABX_3 perovskites

David O. Obada [a,b,c,d,*], Emmanuel Okafor [e,**], Simeon A. Abolade [a], Aniekan M. Ukpong [b,f], David Dodoo-Arhin [g], Akinlolu Akande [a,***]

[a] Mathematical Modelling and Intelligent Systems for Health and Environment Research Group, School of Science, Atlantic Technological University, Ash Lane, Ballytivnan, Sligo, F91 YW50, Ireland
[b] Theoretical and Computational Condensed Matter and Materials Physics Group (TCCMMP), School of Chemistry and Physics, University of KwaZulu-Natal, Pietermaritzburg, 3201, South Africa
[c] Multifunctional Materials Laboratory, Shell Office Complex, Department of Mechanical Engineering, Ahmadu Bello University, Zaria, 810222, Nigeria
[d] Africa Centre of Excellence on New Pedagogies in Engineering Education, Ahmadu Bello University, Zaria, 810222, Nigeria
[e] SDAIA-KFUPM Joint Research Center for Artificial Intelligence, King Fahd University of Petroleum and Minerals, 31261, Saudi Arabia
[f] National Institute for Theoretical and Computational Sciences (NITheCS), Pietermaritzburg, 3201, South Africa
[g] Department of Materials Science and Engineering, University of Ghana, Legon, LG 25, Ghana

ARTICLE INFO

Keywords:
Ensemble learning
Neural networks
Band gaps
Explainable artificial intelligence

ABSTRACT

In this study, we trained and compared explainable machine learning algorithms for predicting the band gaps of perovskite materials that have the formula ABX_3 containing both zero and non-zero band gaps. Six supervised learning models: 5 ensemble learning methods and 1 neural network (CompoundNet) were employed to study the non-linear relationship that exists between the band gap and the characteristics of its constituent elements such as electronegativity, covalent radius, first ionization energy, and row in the periodic table. The machine learning (ML) models were trained on datasets obtained from density functional theory (DFT) calculations. The results show that CatBoost and XGBoost models yielded the least predictive errors and the highest coefficient of determination of $R^2 \geq 88\%$ than other approaches in the testing phase. Furthermore, the Shapley Additive Explanation (SHAP) was used for explaining the model based on the elemental composition of each perovskite compound from the physics standpoint, and a novel holistic feature ranking of the explained models was proposed. One key insight gained from the SHAP analysis is that the Pauling electronegativity of the B site cation in the cubic perovskites which characteristically plays an important role in the electronic properties of this class of materials is the feature that contributes most to the prediction of the band gaps. These results reveal the potential of ML to predict materials properties quickly and accurately with datasets useful in the engineering of efficient solar cell devices.

1. Introduction

It is well-known that perovskite compounds show a remarkable variety of mechanical, electrical, optical, magnetic, and transport properties [1–3]. Typically, the structure of an ideal perovskite with the general formula ABX_3 consists of A cations which occupy 12-fold coordination sites, B cations in the center, and corner-sharing octahedra of X anions. Over the years, perovskite solar cells (PSC) have attracted attention in the field of photovoltaics because of their performance potentials, simple fabrication processes amongst others [4,5], and the power conversion efficiency (PCE) of PSC has reached $\geq 25\% - 25.7\%$ [6–8]. The performance of PSCs is determined by some factors such as, band gap, electrical conductivity, high carrier mobility, remarkable energy level alignment, and low density of defects at the interfaces.

* Corresponding author. Mathematical Modelling and Intelligent Systems for Health and Environment Research Group, School of Science, Atlantic Technological University, Ash Lane, Ballytivnan, Sligo, F91 YW50, Ireland.
** Corresponding author. SDAIA-KFUPM Joint Research Center for Artificial Intelligence, King Fahd University of Petroleum and Minerals, 31261, Saudi Arabia.
*** Corresponding author. Mathematical Modelling and Intelligent Systems for Health and Environment Research Group, School of Science, Atlantic Technological University, Ash Lane, Ballytivnan, Sligo, F91 YW50, Ireland.
E-mail addresses: david.obada@atu.ie (D.O. Obada), emmanuel.okafor@kfupm.edu.sa (E. Okafor), akinlolu.akande@atu.ie (A. Akande).

https://doi.org/10.1016/j.mssp.2023.107427
Received 8 November 2022; Received in revised form 1 February 2023; Accepted 27 February 2023
Available online 24 March 2023

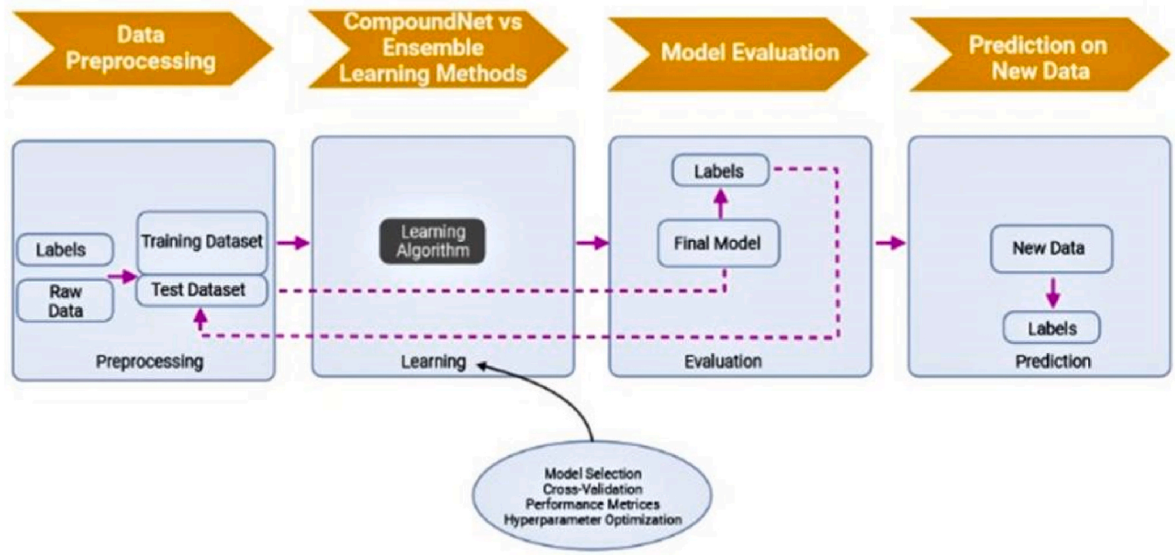

Fig. 1. Block diagram illustrating the learning processes for the supervised learning methods.

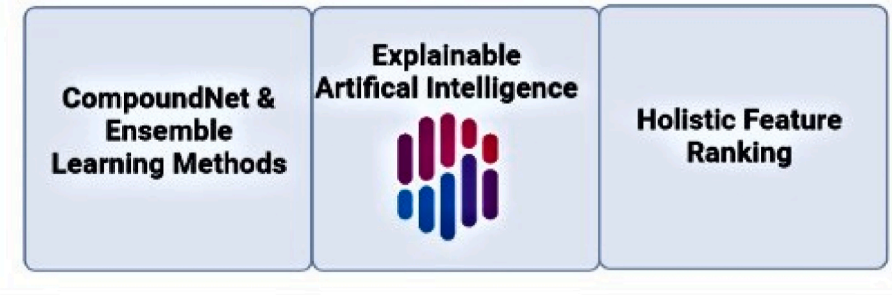

Fig. 2. Block diagram illustration of the system learning pipeline.

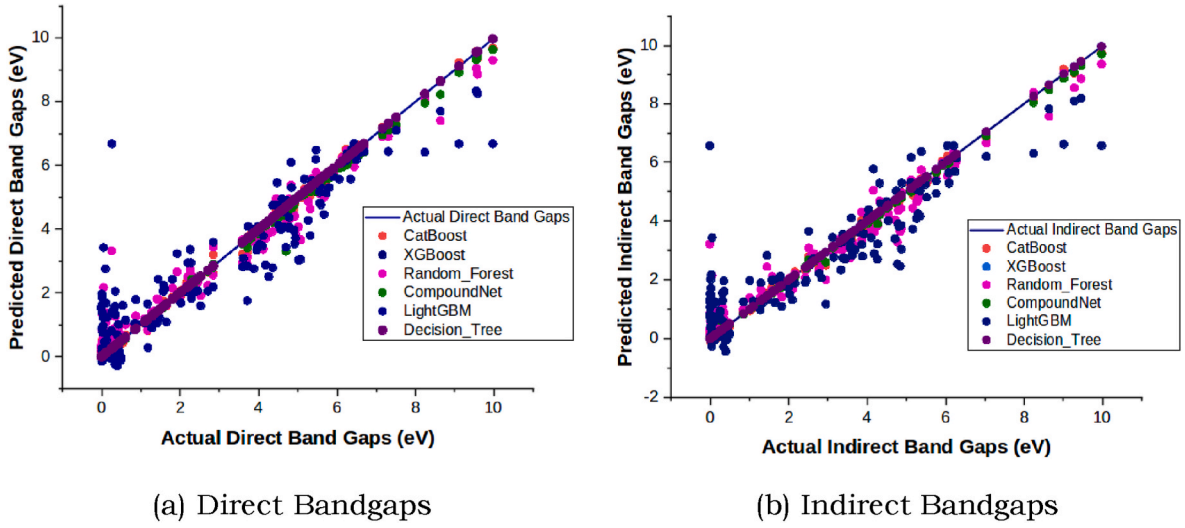

Fig. 3. Training phase correlation plot for each of the supervised learning models prediction on either direct or indirect band gaps compared with their respective actual values for 95% of the entire data.

Amongst all these factors, the design of the band gap of the perovskite layer is crucial and it directly determines its response to the solar spectrum [9,10]. Perovskite semiconductors create a unique opportunity in the engineering of materials because of the large design space. This is possible because the variety of cation-anion combinations has band gaps that span the visible spectrum. Given the complex design space of ABX_3 compounds, it is difficult to explore all the possible combinations both theoretically and experimentally. From a theoretical standpoint, the bandgaps of semiconductors using traditional DFT simulations are underestimated within the generalized gradient

Table 1
Performance evaluation metric comparison of the different machine learning techniques prediction of direct band gaps of ABX_3 perovskites on 95%–5% data splits.

Techniques	Train			Test		
	MAE	RMSE	R^2	MAE	RMSE	R^2
CATBOOST	0.149 ± 0.005	0.205 ± 0.006	0.993 ± 4.658×10^{-4}	0.633 ± 0.112	0.867 ± 0.119	0.887 ± 0.004
XGBOOST	0.005 ± 0.001	0.008 ± 0.001	0.999 ± 2.910×10^{-6}	0.661 ± 0.213	0.871 ± 0.246	0.880 ± 0.056
RANDOM FOREST	0.324 ± 0.004	0.508 ± 0.004	0.959 ± 0.001	0.798 ± 0.235	0.966 ± 0.224	0.860 ± 0.034
COMPOUNDNET	0.074 ± 0.029	0.129 ± 0.051	0.997 ± 0.002	0.749 ± 0.234	1.090 ± 0.347	0.818 ± 0.074
DECISION TREE	0.000 ± 0.000	0.000 ± 0.000	1.000 ± 0.000	0.765 ± 0.194	1.070 ± 0.308	0.813 ± 0.094
LIGHTGBM	0.678 ± 0.012	0.941 ± 0.027	0.860 ± 0.009	0.920 ± 0.342	1.140 ± 0.349	0.801 ± 0.078

Table 2
Performance evaluation metric comparison of the different machine learning techniques prediction of indirect band gaps of ABX_3 perovskites on 95%–5% data splits.

Techniques	Train			Test		
	MAE	RMSE	R^2	MAE	RMSE	R^2
CATBOOST	0.137 ± 0.003	0.185 ± 0.004	0.994 ± 2.329×10^{-4}	0.572 ± 0.150	0.776 ± 0.227	0.906 ± 0.028
XGBOOST	0.005 ± 0.001	0.007 ± 0.002	0.999 ± 4.942×10^{-6}	0.588 ± 0.225	0.773 ± 0.304	0.904 ± 0.050
RANDOM FOREST	0.313 ± 0.006	0.497 ± 0.006	0.956 ± 0.001	0.697 ± 0.264	0.857 ± 0.324	0.884 ± 0.053
COMPOUNDNET	0.068 ± 0.041	0.126 ± 0.072	0.996 ± 0.003	0.595 ± 0.194	0.827 ± 0.279	0.869 ± 0.089
LIGHTGBM	0.640 ± 0.013	0.902 ± 0.026	0.854 ± 0.007	0.808 ± 0.357	1.002 ± 0.441	0.837 ± 0.092
DECISION TREE	0.000 ± 0.000	0.000 ± 0.000	1.000 ± 0.000	0.634 ± 0.110	1.068 ± 0.235	0.803 ± 0.097

approximation (GGA) [11] but can be overcome by using hybrid functionals or many-body perturbation theory (GW) [12,13]. Nonetheless, these more accurate theoretical approaches are more computationally expensive and can be difficult to implement on a vast array of materials. A possible approach to overcome this limitation is the application of ML models to enhance the predictions. Typically, in the ML approach, the properties of the materials under investigation are calculated by DFT or obtained from laboratory experiments for a small sample size which is then used to train statistical ML models. The model learns the trends in the data distribution during the training phase. Thereafter, the model is used to predict the properties of new datasets based on the trends it has learned. Contextually, when accurate high performing ab-initio calculations are performed on a set of ABX_3 compounds, the results can be used to train ML models to obtain predictions for the remaining materials in the large design space.

Several studies have focused on using ML models to predict the band gaps of ABX_3 type compounds [14–18]. To highlight a few, Lee et al. focused on using the support-vector regression ML technique to predict the band gaps of 156 binary compounds using descriptors which include the band gaps obtained from DFT calculations using GGA and obtained a root mean squared error (RMSE) of 180 meV [19]. Gladkikh et al., [20] studied non-linear mappings that exist between band gap and the properties of the elements using the Alternating Conditional Expectations ML technique (a method useful for small datasets) and compared the results with other ML methods. They concluded that the best ML methods which successfully studied the linear mappings were Kernel Ridge Regression and Extremely Randomized Trees. Liu et al. [7], used ML to predict the experimental band gaps of 227 perovskites obtained from 1254 recent publications with a bid to identify 4 models from 24 kinds of ML models. The models achieved high accuracy with an RMSE of down to 0.55. In addition, explainability ML was used to explain the effect of each chemical composition for their proposed models, and this further established the potential of ML to accurately predict the band gaps of perovskite materials. In a study conducted by Huang et al., [21], the band gaps of 300 wurtzite nitride semiconductors were calculated using DFT. These datasets were then used to train many ML models for predicting the band gaps. From all ML models tested, the best performance was achieved using support-vector regression. Pilania et al., [22] used the kernel-ridge regression technique and 16 sets of the element-specific descriptor to predict the band gaps of 1306 double perovskites and obtained an RMSE of 80 meV. Rath et al., [23] classified ABX_3 type perovskites into direct and indirect band gap materials using the XGBoost classifier with datasets of 1528 ABX_3 compounds and obtained an average accuracy of about 72.8%. In a recent article published by Lyu et al. [24], it was shown that machine-learning models could be useful for low-dimensional organic-inorganic halide perovskites.

To allow for a critical investigation of the models, interpretable machine learning using Shapley additive explanations (SHAP) was adopted. The SHAP analysis was performed to determine which of the descriptors used in the prediction was most important. They concluded that one major finding from the SHAP analysis is that the absence of transition metals increased the probability of the perovskite having a direct band gap. SHAP which originated from the cooperative game theory have been used to interpret many complex ML models which are so-called black box models. In 2017, Lundberg and Lee [25] proposed the SHAP value to explain various models for better interpretation. Before the adoption of SHAP, feature importance has been used to explain ML models. Although these reflect the importance of the features directly with emphasis on the impact of the features on the final model, it has proven deficient in judging the relationship between the features and the prediction of the results. More so, in the context of using several ML algorithms, a global ranking of the input features for all the models may not be feasible. Therefore, there is a need to propose computationally efficient methods for accurate global ranking of input features if several ML algorithms are used.

To the best of our knowledge, this is the first work that holistically employs the use of explainable ML models for regressive prediction of direct and indirect bandgaps of inorganic perovskite compounds. Another contribution is the use of the explainable model (SHAP) in attempting to assess the feature importance or rationale behind how these ML models can accurately predict the band gaps of the inorganic perovskite materials. To assess the importance of these features, we propose a novel holistic ranking method for identifying the most prominent feature from the explained models.

In this study, the band gaps of 199 perovskites having a formula ABX_3 were modeled per the element-specific descriptors of individual elements viz: electronegativity, covalent radius, first ionization energy, and row in the periodic table. We compared the new ANN model called CompoundNet to other ensemble machine learning techniques: Decision Trees, Random Forest, CatBoost, XGBoost, and LightGBM. Finally, the results are explained from the physics standpoint using the Shapley Additive Explanations (SHAP) to evaluate the effect of the specific

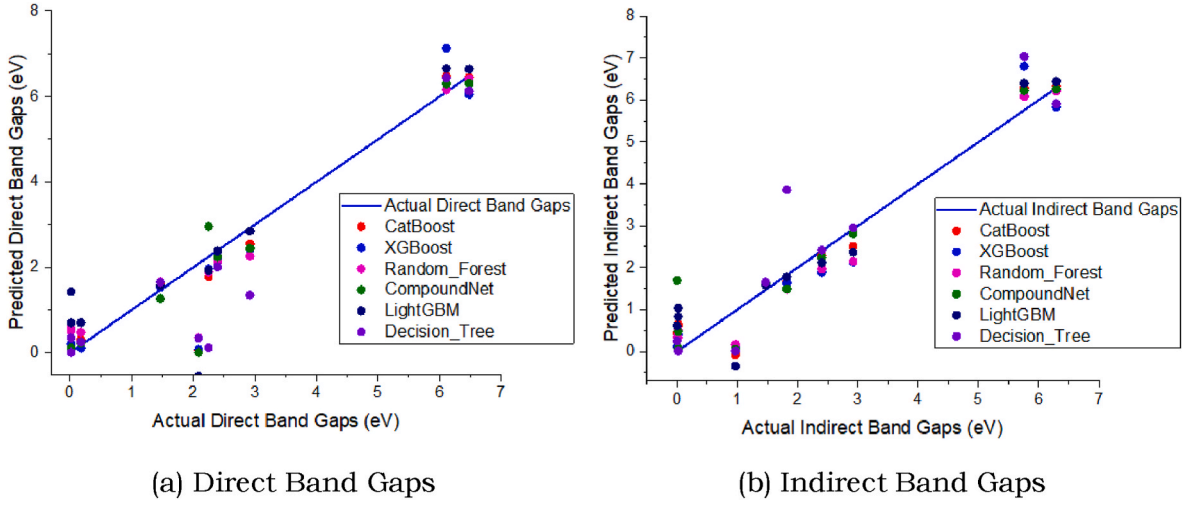

Fig. 4. Testing phase correlation plot for each of the supervised learning models prediction on either direct or indirect band gaps compared with their respective actual values for 5% of the entire data.

Table 3
Performance evaluation metric comparison of the different machine learning techniques prediction of direct band gaps of ABX_3 perovskites on 80%–20% data splits.

Techniques	Train			Test		
	MAE	RMSE	R^2	MAE	RMSE	R^2
CATBOOST	0.124 ± 0.003	0.166 ± 0.004	0.996 ± 3.656×10^{-4}	0.845 ± 0.094	1.390 ± 0.186	0.697 ± 0.099
XGBOOST	0.003 ± 0.001	0.004 ± 0.001	0.999 ± 1.264×10^{-6}	0.810 ± 0.126	1.460 ± 0.177	0.664 ± 0.116
RANDOM FOREST	0.342 ± 0.011	0.507 ± 0.020	0.959 ± 0.004	0.963 ± 0.175	1.450 ± 0.264	0.665 ± 0.150
LIGHTGBM	0.720 ± 0.025	0.991 ± 0.066	0.843 ± 0.021	1.130 ± 0.106	1.610 ± 0.170	0.599 ± 0.101
COMPOUNDNET	0.065 ± 0.027	0.129 ± 0.050	0.997 ± 0.002	0.991 ± 0.124	1.680 ± 0.252	0.557 ± 0.141
DECISION TREE	0.000 ± 0.000	0.000 ± 0.000	1.000 ± 0.000	0.993 ± 0.142	1.790 ± 0.245	0.505 ± 0.124

Table 4
Performance evaluation metric comparison of the different machine learning techniques prediction of indirect band gaps of ABX_3 perovskites on 80-20 data splits.

Techniques	Train			Test		
	MAE	RMSE	R^2	MAE	RMSE	R^2
CATBOOST	0.116 ± 0.006	0.155 ± 0.008	0.996 ± 4.873×10^{-4}	0.795 ± 0.103	1.301 ± 0.226	0.699 ± 0.125
RANDOM FOREST	0.329 ± 0.012	0.495 ± 0.025	0.957 ± 0.005	0.909 ± 0.182	1.386 ± 0.310	0.650 ± 0.189
XGBOOST	0.003 ± 0.006	0.004 ± 0.001	0.999 ± 2.006×10^{-6}	0.791 ± 0.127	1.397 ± 0.276	0.647 ± 0.178
LIGHTGBM	0.687 ± 0.029	0.965 ± 0.073	0.835 ± 0.025	1.073 ± 0.109	1.529 ± 0.192	0.591 ± 0.131
COMPOUNDNET	0.053 ± 0.018	0.087 ± 0.037	0.998 ± 0.001	0.879 ± 0.146	1.521 ± 0.366	0.576 ± 0.237
DECISION TREE	0.000 ± 0.000	0.000 ± 0.000	1.000 ± 0.000	0.812 ± 0.170	1.574 ± 0.365	0.557 ± 0.195

descriptors on the band gap predictions. In what follows, we present the data description, distribution and pre-processing in section 2, the methodology in section 3, the results and discussion in section 4, and we conclude in section 5.

2. Data description, distribution and preprocessing

In this study, we have used datasets obtained from the work of Korbel et al., [26]. The authors used the more accurate hybrid HSE06 exchange-correlation function to calculate the band gaps of 199 compounds. The numerical values for the adopted 199 compounds are outlined in Table ESI-1 in the Supplementary Information of Ref. [26]. Hence in this study, we used the obtained HSE06 band gaps to train our models. We calculated the tolerance and octahedral factors to further establish the formation of perovskites and the stability of the compounds. The Goldschmidt tolerance factor was used by assessing the ionic radius of the values as compiled by Shannon [27]. We also performed first principle calculations with the Vienna Ab-initio Simulation Package-VASP at PBE GGA functional level (results not shown) on randomly selected cubic perovskites from the 199 compounds to ensure that the datasets we have adopted are reproducible [28,29].

Different features have been proposed as descriptors for the properties of materials [30,31]. In this study, a simple set of element-specific descriptors have been used. For each of the elements in the ABX_3 compound, the electronegativity, covalent radius, first ionization energy, and row in the periodic table were used because from the physics perspective, the selected descriptors have an influence on the band gaps of the compounds, and these descriptors can improve the training of machine learning models leading to better performance and greater accuracy. This gave 12 features in total per compound. The 12-dimensional feature space has proven to be effective in the prediction of the magnitude of band gaps when regression techniques are used [32].

The calculated band gap examples from the ABX_3 perovskites data as obtained from Ref. [24], were partitioned in the ratio 95%: 5% for the training set and testing set, respectively. Furthermore, the input features for the used dataset were normalized in the scale [0, 1] using the

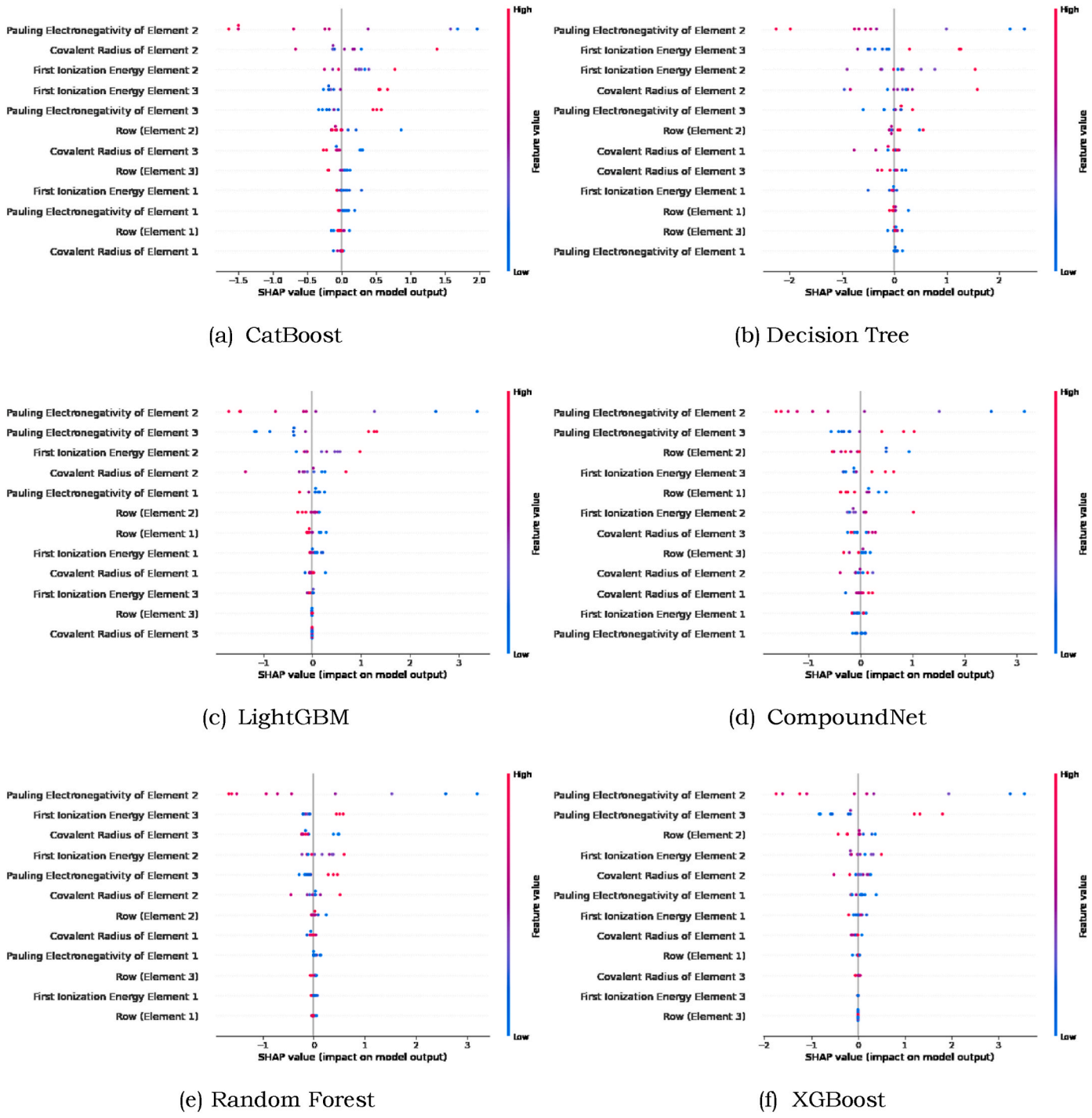

Fig. 5. Explainability of the supervised learning model prediction of the direct band gap while revealing the feature importance influencing the model prediction.

expression in equation (1).

$$X_n = \frac{X_j - min(X_j)}{max(X_j) - min(X_j)} \tag{1}$$

where the X_j denotes the raw input features and the normalized input features can be represented as X_n. The effective normalized input features with the corresponding continuous output labels for each of the aforementioned datasets were passed to the supervised learning algorithms.

3. Methods

In this section, we briefly describe the supervised learning algorithms and the corresponding explainable artificial intelligence (XAI) tools used in this study for a better understanding of the theoretical background.

3.1. Supervised learning algorithms

3.1.1. CompoundNet

The CompoundNet is a feedforward artificial neural network that consists of three main units; input unit, hidden unit, and output unit. The CompoundNet is a multilayer perceptron (MLP) which is an example of a

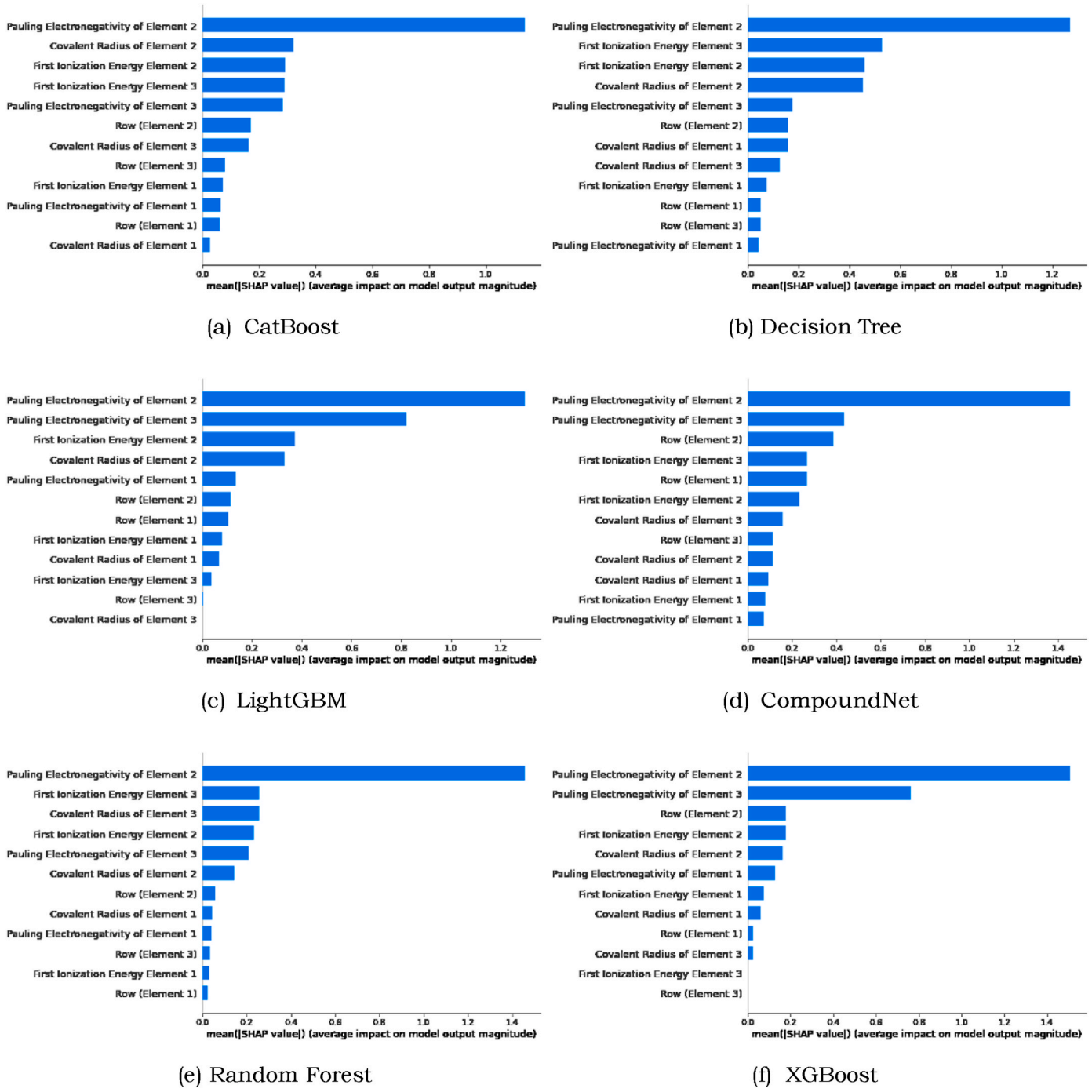

Fig. 6. Explainability of the supervised learning model prediction of the direct band gap while revealing the feature importance influencing the model prediction (mean SHAP value).

supervised learning algorithm that can be used for performing classification or regression tasks. In our experiment, the hidden unit consists of five neural network layers; whereby each layer contains 64 network nodes. The rationale behind the choice of 64 network nodes is based on an intuitive design philosophy whereby we attempt to explore the uniform nodal distribution based on the formula: $2^6 = 64$. Each network node within the hidden and output layers generate feature maps as defined in the expression;

$$Y^{l}_{mlp} = b^{l}_{j} + \sum_{k} (W^{l-1}_{kj} \times X^{l-1}_{k}) \tag{2}$$

here the hypothetic model output denoted by Y^{l}_{mlp} compute the effective addition of the sum of weighted inputs $\sum_k (W^{l-1}_{kj} \times X^{l-1}_{k})$ and the corresponding bias b^{l}_{j} in $R^{1\times1}$ dimensional space. Note that the variable W_{kj} and X^{l-1}_{k} account for the input weights and input features, respectively based on the dimension $R^{1\times12}$ per each example in the perovskite material composition. Hence the predictive error (cost function) was computed using the information from the actual output and predictive outcome from the hypothetic model. We used the Adam optimizer [33] in optimizing the predictive error via backpropagation to yield optimal weights needed for the CompoundNet model to predict direct or indirect band gap from the perovskites examples in the testing phase.

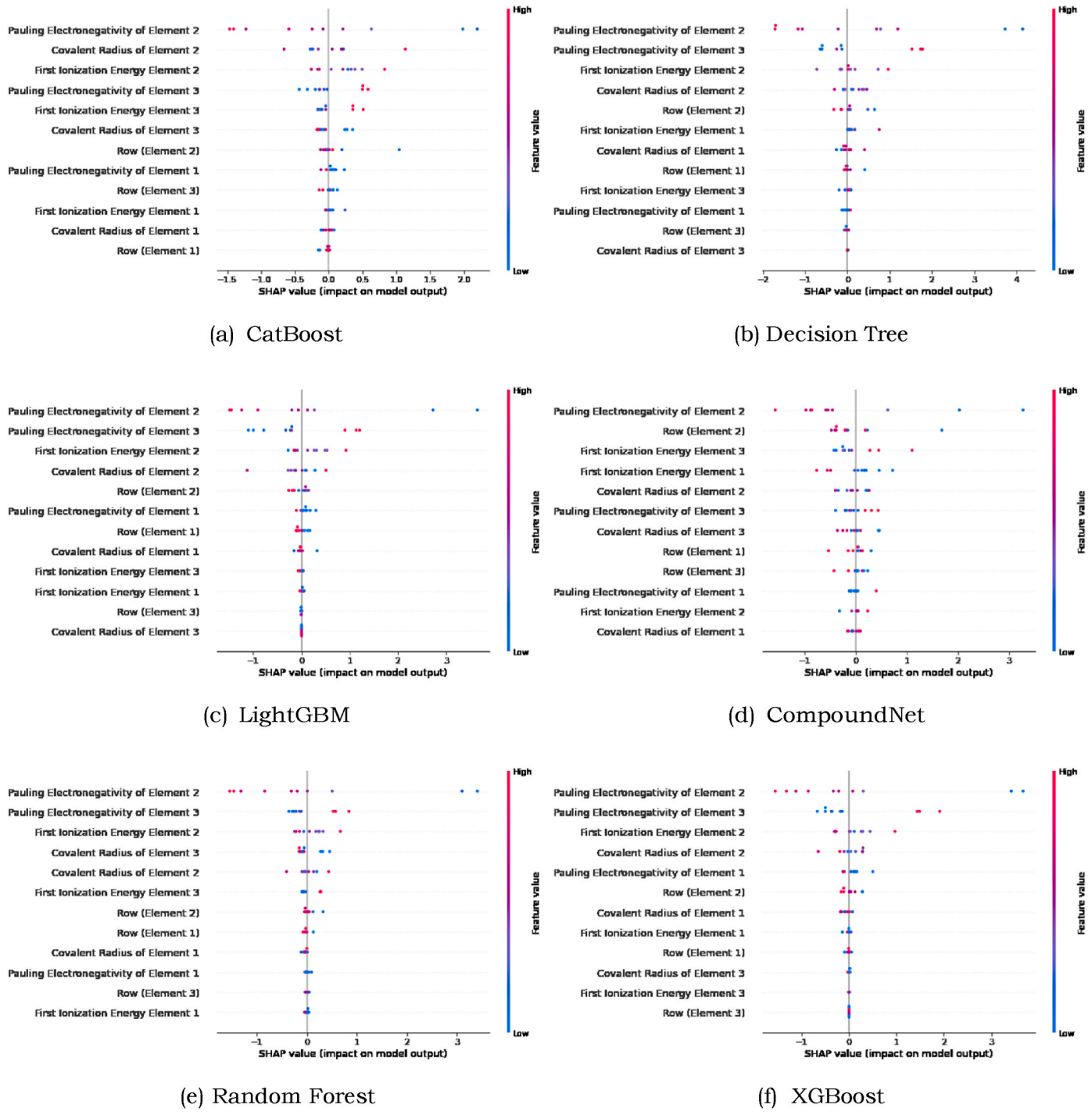

Fig. 7. Explainability of the supervised learning model prediction of the indirect band gap while revealing the feature importance influencing the model prediction.

3.1.2. Random forest

Random Forest (RF) [34] is one of the traditional ensemble learning techniques that was originally derived from the bagging aggregation principle. This method can be created by integrating several instances of de-correlated estimator trees [35]. This method computes the average from the aggregation of several base learners before determining the most likely continuous output (estimating an average score from the base learners). For training example given as $D_j = (X_j, Y_j)\ \forall\ J \in N$, it should be noted that the variable X_j represent the actual input features and Y_j denotes the actual output from the original dataset. For a given normalized input feature $X_n \in [0,1]^k$, our goal is to calculate a regression function $Y_{rf}(x) = E[Y|X = X_j]$ within the dataset D_j.

$$\overline{Y_{rf}}(D_j) = E_W[Y_j(W_k, X \in D_j)] \tag{3}$$

The variable E_W denotes the estimated output from the random forest. An RF is a kind of predictive estimator that collects a set of randomized base predictive estimator regression trees $\{Y_{rf}(W_k, X \in D_j), k \geq 2\}$, where the weight variable can be denoted $W = \{W_1, W_2, \ldots W_k\}$ as a randomly distributed variable. The random output decision trees were integrated to generate aggregation of several regression trees. We used random forest (Y_{rf}) containing 10,000 base estimators when conducting our experiments.

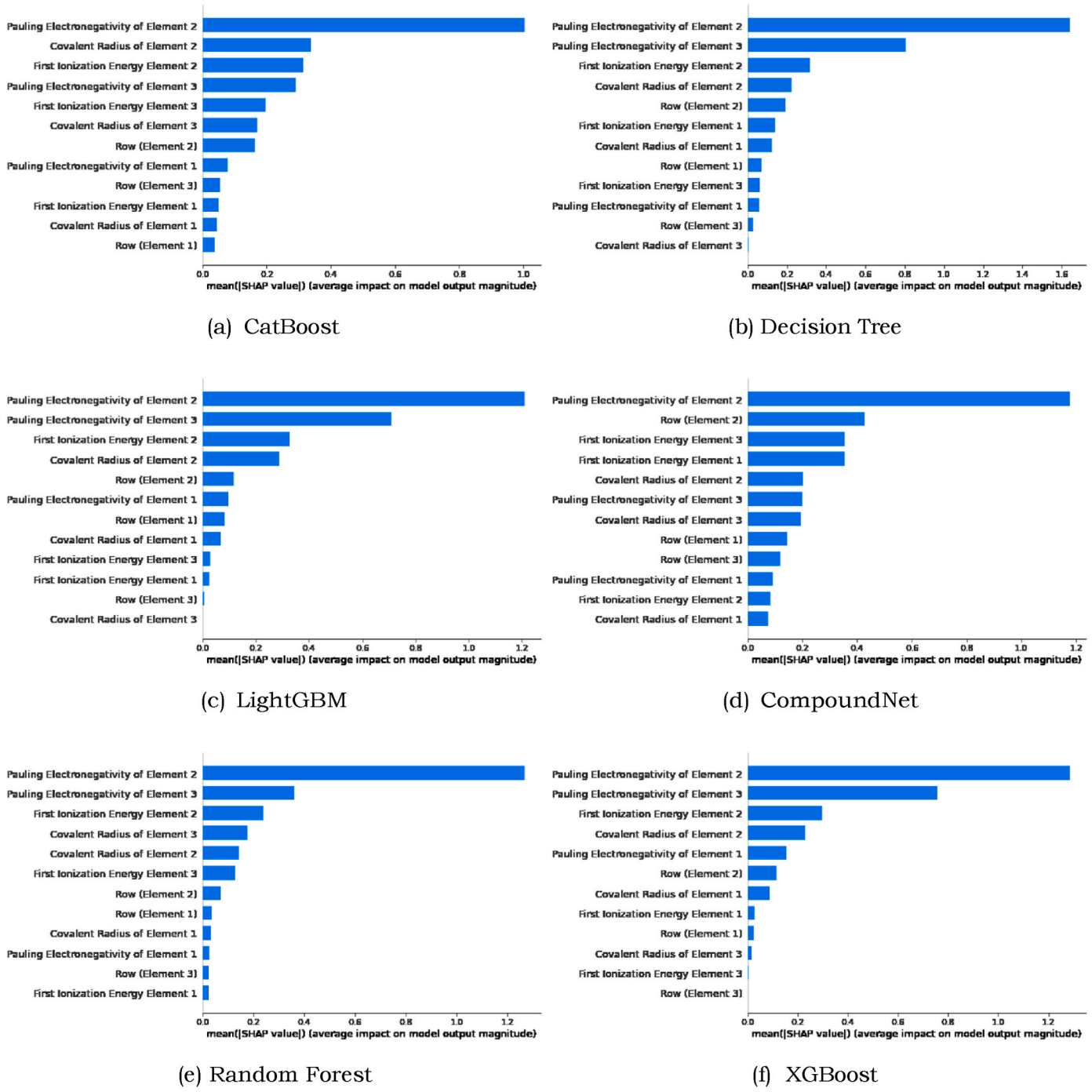

Fig. 8. Explainability of the supervised learning model prediction of the indirect band gap while revealing the feature importance influencing the model prediction (mean SHAP value).

Table 5
Explainablity model feature importance ranking in the testing phase; the power index represents the feature positional ranking, when considering direct bandgap prediction.

Methods	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	F11	F12
CATBOOST	1^{1}	4^{2}	7^{3}	8^{4}	2^{5}	10^{6}	5^{7}	11^{8}	6^{9}	0^{10}	9^{11}	3^{12}
XGBOOST	1^{1}	2^{2}	10^{3}	7^{4}	4^{5}	0^{6}	6^{7}	3^{8}	9^{9}	5^{10}	8^{11}	11^{12}
RANDOM FOREST	1^{1}	8^{2}	5^{3}	7^{4}	2^{5}	4^{6}	10^{7}	3^{8}	0^{9}	11^{10}	6^{11}	9^{12}
COMPOUNDNET	1^{1}	2^{2}	10^{3}	8^{4}	9^{5}	7^{6}	5^{7}	11^{8}	4^{9}	3^{10}	6^{11}	0^{12}
LIGHTGBM	1^{1}	2^{2}	7^{3}	4^{4}	0^{5}	10^{6}	9^{7}	6^{8}	3^{9}	8^{10}	11^{11}	5^{12}
DECISION TREE	1^{1}	8^{2}	7^{3}	4^{4}	2^{5}	10^{6}	3^{7}	5^{8}	6^{9}	9^{10}	11^{11}	0^{12}

Table 6
Holistic feature ranking using all the outcomes from the explained machine learning models after predicting direct bandgap.

Features	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	F11	F12
Sum of ranking R_s	6	21	23	30	31	33	47	54	54	54	55	60
Best feature ranking order	1	2	7	4	10	8	5	0	3	9	6	11
Feature importance	Pauling Electronegativity of Element 2	Pauling Electronegativity of Element 3	First Ionization energy of Element 2	Covalent Radius of Element 2	Row (Element 2)	First Ionization Energy Element 3	Covalent Radius of Element 3	Pauling Electronegativity of Element 1	Covalent Radius of Element 1	Row (Element 1)	First Ionization Energy Element 1	Row (Element 3)

3.1.3. Decision tree

A decision tree is an example of a supervised learning technique mainly used for solving classification or regression tasks [36,37]. Hence, given an input feature space, the decision trees operate based on the principle associated with entropy and information gain in the formation of a supervised learning model.

3.1.4. XGBoost

EXtreme Gradient Boosting (XGBoost) method is a scalable tree boosting technique [38]; this method relies on a sparse-aware learning paradigm that allows multiple base-tree learners to predict sparse and clustered data. The main design philosophy of an XGBoost is that it factors in data compression, cache accessibility, and sharding for creating a more scalable decision tree predictive system.

3.1.5. CatBoost

The Catboost [39] is an example of the ensemble learning algorithm. The name CatBoost was derived from the compound words; "categorical boosting". A typical CatBoost relies on base learners by ordering and employing an innovative learning algorithm for operating categorical features. The main merit of CatBoost is that it has the prowess to address prediction shifting arising from output target leakage. This method is one of the most competitive state-of-the-art ensemble learning method.

3.1.6. LightGBM

The light gradient boosting method (LightGBM) [40] is another competitive ensemble learning method that depends on decision trees that employ two main algorithm paradigms; gradient-based one-side sampling and an exclusive feature bundling. This method is often used for solving classification and regression tasks. The block diagram describing the learning process for each of the described methods is shown in Fig. 1. The learning process starts with data preprocessing, then followed by the actual learning process using the ML models. The model evaluation and the prediction on the new data completes the process.

3.2. Explainable artificial intelligence

Many classical machine learning and deep learning techniques are often considered black-box as a result of limited internal information about the rationale behind their model interpretability [41]. Based on the recent advances in AI, It has become pertinent to explore explainable Artificial Intelligence (XAI) and its relevance in understanding the feature importance that influences a certain machine learning model prediction. An example of an XAI algorithm is SHapley Additive exPlanations (SHAP). SHAP is an explainability tool that relies on the unification of frameworks that allow researchers or experts to gain insightful interpretation of complex predictive models. The core unit of a SHAP algorithm involves identifying a novel class by assessing additive feature relevance and finding the unique solution of the new class based on a collection of desirable attributes. Overall, the SHAP estimation approach aligns effectively with human intuition. We considered two forms of SHAP explainers; Tree-based explainer was used for interpreting the ensemble learning models, while the sampling-based explainer was used for interpreting the CompoundNet model. A block diagram illustration of the developed system pipeline is shown in Fig. 2.

3.3. Performance metrics

The generalization capacity of the trained models can be measured using the following performance metrics;

1. Coefficient of Determination: The coefficient of determination commonly known as R^2 is a metric tool employed for determining the degree of correlation existing between two or more sets of variables. The R^2 can also be described as the goodness of fit. An R^2 can be

Table 7
Explainablity model feature importance ranking in the testing phase; the power index represents the feature positional ranking when considering indirect bandgap prediction.

Methods	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	F11	F12
CATBOOST	1^{1}	4^{2}	7^{3}	2^{4}	8^{5}	5^{6}	10^{7}	0^{8}	11^{9}	6^{10}	3^{11}	9^{12}
XGBOOST	1^{1}	2^{2}	7^{3}	4^{4}	0^{5}	10^{6}	3^{7}	6^{8}	9^{9}	5^{10}	8^{11}	11^{12}
RANDOM FOREST	1^{1}	2^{2}	7^{3}	5^{4}	4^{5}	8^{6}	10^{7}	9^{8}	3^{9}	0^{10}	11^{11}	6^{12}
COMPOUNDNET	1^{1}	10^{2}	8^{3}	6^{4}	4^{5}	2^{6}	5^{7}	9^{8}	11^{9}	0^{10}	7^{11}	3^{12}
LIGHTGBM	1^{1}	2^{2}	7^{3}	4^{4}	10^{5}	0^{6}	9^{7}	3^{8}	8^{9}	6^{10}	11^{11}	5^{12}
DECISION TREE	1^{1}	2^{2}	7^{3}	4^{4}	10^{5}	6^{6}	3^{7}	9^{8}	8^{9}	0^{10}	11^{11}	5^{12}

defined within the range {0, 1}. If $R^2 = 1$, the model is said to have a perfect fit and is highly reliable. However, if a model yields $R^2 = 0$, then the hypothetical model can be described as yielding poor correlation with weak generalization potential. The mathematical formula R^2 can be described as:

$$R^2 = 1 - \frac{\sum_i (Y_j - Y_i^p)^2}{\sum_j (Y_j - \overline{Y_j})^2} \tag{4}$$

where Y_j accounts for the target values and Y_j^p denotes the predicted outputs from the described supervised learning algorithms. The variable $\overline{Y_j}$ denotes mean of Y_j.

2. Root Mean Square Error (RMSE): The RMSE measures the effective difference between the actual experimental target output and the predicted model output. Furthermore, RMSE determines or measures the goodness of fit from the generated generalized regression model. An RMSE can be defined as;

$$RMSE = \left(\frac{1}{n}\sum_j (Y_j - Y_j^p)^2\right)^{\frac{1}{2}} \tag{5}$$

3. Mean Absolute Error (MAE): MAE is another evaluation metric for calculating the absolute difference between target output and the predicted model output (continuous variables);

$$MAE = \frac{1}{n}\sum_j |Y_j - Y_j^p| \tag{6}$$

4. Results and discussion

In this section, we provide the computational results obtained, and discussed the research findings for the investigated supervised learning models. We start with the supervised learning models and then discuss our findings on the explainability ML.

4.1. Supervised learning performance evaluation

In Fig. 3, we show correlation plots of all the method predictions for both direct (Fig. 3a) and indirect band gaps (Fig. 3b) for one-fifth $(\frac{1}{5})$ of the total experimental runs within the training phase. Using larger training examples which translate to (95%) of the effective data was used for generating each of the supervised learning models. The summarized evaluation performance results obtained from the supervised learning model prediction of the direct and indirect band gaps in the training and testing phases are reported in Table 1 and Table 2 respectively.

From Tables 1 and 2, we report that most of the ensemble learning techniques (CatBoost, XGBoost, and Decision Tree), and neural network method (CompoundNet) yielded a coefficient of determination $R^2 > 0.99$ and lower predictive error when compared with LightGBM and Random Forest in the training phase. The R^2 of the superior ensemble learning models indicates that there is a strong degree of correlation between the elemental composition of each ABX_3 type perovskites and the band gaps.

However, we observe that the LightGBM and Random Forest experienced underfitting relative to other techniques in the training phase. To further inspect the generalization potential for each of the methods, it is pertinent to explore the overall performance metrics for these methods in the testing phase. The observation of one-fifth of the total experimental runs showing the testing phase correlation plots of all the methods predicting direct or indirect bandgaps is shown in Fig. 4. Overall, the best methods (CatBoost and XGBoost) yielded the least predictive errors: $MAE \leq 0.66$ or $RMSE \leq 0.87$ and the highest $R^2 \geq 0.88$ for the prediction of both direct and indirect band gap. We generally observe that the Decision Tree technique suffers from an overfitting problem.

For fewer amount of training examples (80%) of the entire data: the summary of the supervised learning performance index for both training and testing sets were presented in Table 3 and Table 4, respectively. From the described tables, we report an excellent R^2 and minimal predictive errors in CatBoost, XGBoost, and Decision Trees in the training phase, but these methods were unable to generalize very well in the testing phase due to overfitting problems. This often emanates due to decision tree depth and specificity drawn in the testing phase that is unable to assume the model distribution in the testing phase. However, in the testing phase, we report that CatBoost when compared with other methods yielded the best $R^2 \geq 0.697$ and least predictive errors ($MAE \geq 0.795$ and $RMSE \geq 1.30$) in the testing phase. Hence we can draw an inference from our evaluations that CatBoost has the best generalization potential and was capable to learn the multivariate input feature space. The remaining ensemble learning techniques and CompoundNet outperform Decision Tree across all the evaluated metrics. These observations are the same for models generated using large and fewer training examples.

4.2. Explainability analysis and the proposed holistic feature ranking

To hypothetically explain the model rationale with respect to feature relevance behind the goodness of one method relative to the other approaches, we examine the SHAP algorithm for assessing the impact of the feature importance of the ABX_3 perovskites when the supervised learning algorithms are used in predicting the indirect and direct band gaps. The SHAP graphs are shown in Figs. 5–8. Highlighting the importance of the feature graphs, the most important feature can be found at the top, and the importance of the other features is ranked in descending order. As shown in Figs. 5–8, the Y-axis depicts the feature nomenclature, while the X-axis shows the corresponding mean of the magnitude of the SHAP values. The mean of the magnitude reveals the average impact the feature has on the model output. When the mean of the magnitude is high, it means the impact on the predicted value is high. The color of all the points represents the value of the corresponding feature. The red color represents high values while the blue color represents low values. It should be noted that in Figs. 5 and 7, the points of the feature on the right-hand side depicts that the SHAP analysis contributes positively in the model prediction, while the left-hand-side contributes negatively during the model predictions.

Based on the observation of the feature ranking from Figs. 5–8, we employed our novel holistic feature ranking method to determine the

Table 8
Holistic feature ranking using all the outcomes from the explained machine learning models after predicting indirect bandgap.

Features	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	F11	F12
Sum of ranking R_s	6	18	24	26	32	43	49	50	51	52	54	63
Best feature ranking order	1	2	4	7	10	8	0	6	5	9	3	11
Feature importance	Pauling Electronegativity of Element 2	Pauling Electronegativity of Element 3	Covalent Radius of Element 2	First Ionization energy of Element 2	Row (Element 2)	First Ionization Energy Element 3	Pauling Electronegativity of Element 1	First Ionization Energy Element 1	Covalent Radius of Element 3	Row (Element 1)	Covalent Radius of Element 1	Row (Element 3)

global feature ranking of the ABX_3 perovskites across all the explained supervised learning models. A summary of our findings on the feature importance and the holistic feature ranking for all the methods were reported on Table 5 to Table 8, respectively. From the latter, we report that the most important feature is the "Pauling Electronegativity of Element 2". This is because the feature appears as the first ranked across all the examined methods and has the least sum of ranks as shown in Table 6 and Table 8.

From Figs. 5–8, we report each of the methods' feature importance ranking on Tables 5 and 7. Then we developed a hypothesis formulation to assess each of the methods by counting the frequency of the feature ranking across all the methods to determine which of the feature contributes the most from the trained supervised learning model predictions.

$$R_s(F_i^p) = \sum_i count(F_i^p) \times p \tag{7}$$

where R_s is the sum of effective ranking per feature $\forall$ the learning models, F_i^p represents the input feature having a code in the range $\{0-11\}$, the index variable p denotes the feature positional or ranking value, and i is the number of entries per each of the features. Suppose an input is given as $F_i^p = 2$, the positional value is $p = \{2, 5\}$, the frequency of occurrence of the input feature $F_i^p = 2$ is given as $count(F_i^p) = \{3, 3\}$ and by performing a calculation on equation (7) yields a value $R_s = 21$. By extending the same principle to the remaining features, a summarized best feature ranking is reported in Tables 6 and 8 for the prediction of the direct and indirect bandgaps, respectively.

Based on the insights drawn from the holistic feature ranking in the testing phase for both direct and indirect band gaps, we used a scatter plot relationship (not shown) of the Pauling electronegativity of element 2 (most important feature) and row of element 3 (least important feature) versus band gap predictions for each of the supervised learning models. The investigation revealed that electronegativities between 1.8 and 2.0 contributes the most to the band gaps between 0 and 2 eV (a consideration that aligns with the Shockley–Queisser limit of band gaps). On the other hand, the investigation showed that the anions (element 3) in row 2 contributes the least to the band gaps between 0 and 2 eV. Therefore, the synergistic effect of varied electronegativities in the crystal system and the rows of the elements can be used to tune the band gaps of perovskite compounds for effective light-harvesting.

From Tables 6 and 8, based on the holistic ranking done for the prediction of both direct and indirect band gaps, the Pauling electronegativity of element 2 and 3, and the first ionization energy of element 2 are the most important features. A reflection on this assertion can be made: the Pauling electronegativity is based on the energies of dissociation and cannot be regarded as a property of individual atoms, but of atoms that are bonded. Therefore, the energy of the semiconducting perovskite compounds would typically involve the transfer of an electron from the valence band to the conduction band. Since the valence band is characterized primarily by the orbitals of the anion, and the conduction band is primarily characterized by the orbitals of the cation, then it is expected that some numerical parameters like the ionization energy and the electronegativity of the anion and cation will correlate strongly with the energy band gaps [42,43]. It has been established that there is a strong relationship between the offset of the conduction band and ionicity (difference between the metal and oxygen electronegativities of the compounds) [44].

The inferences from the SHAP analysis as described are important from a physics standpoint. These inferences can enable scientists with the selection of constituent elements to maximize the probabilities of obtaining the predictions of direct and indirect band gaps of ABX_3 perovskite. This can help in tailoring the search for direct band gap materials for specific industrial applications. The SHAP analysis reveal that the range of electronegativity for the B cation of all the compounds is the most important feature which determines the band gaps and this is

in agreement with work carried out by Gladkikh et al., [20], and Rath et al., [23].

5. Conclusion

In this study, we have demonstrated that the CatBoost model has a more predictive power for the determination of direct and indirect band gaps of ABX_3 perovskites with a correlation score of $R^2 \geq 0.88$ in the testing phase when models are generated from large data samples, which is fascinating for the practical exploitation of the algorithms. The SHAP analysis yielded the features and their impact on the predictions. The electronegativities for the B cation in the cubic perovskite compounds were showcased as the most important feature in the prediction of the band gaps. The insights are crucial when designing materials in a large materials discovery space and when synthesizing light-harvesting perovskites.

The robust implementation of the ML algorithms can aid a deliberate discovery of new ABX_3 perovskites with suitable band gaps for opto-electronic applications. This can invariably reduce trial and error experiments in the laboratories and also reduce the number of ab initio DFT calculations needed. This is therefore in resonance with the overall objective of materials informatics which accelerates the design and selection of materials. Furthermore, this study demonstrated that the newly proposed holistic ranking feature provides a simple and efficient global ranking across all the investigated methods.

Additionally, some of the ensemble learning methods outperformed the CompoundNet; however, there were instances the CompoundNet was better than some ensemble learning techniques during the prediction of the direct or indirect band gaps. Future work can explore using $1 - dimensional$ deep learning architecture involving a $1 \times k$ moving kernel convolving with the input feature of the ABX_3 perovskites to generate informative feature-maps that may help in yielding a possible improvement in the prediction of the band gaps.

Funding

The authors wish to thank the Irish Research Council for funding granted to David O. Obada with Project ID GOIPD/2021/28. Most of the calculations were performed on the Kelvin cluster maintained by the Trinity Centre for High Performance Computing. This cluster was funded through grants from the Higher Education Authority, through its PRTLI program. The authors also wish to acknowledge the Irish Centre for High-End Computing (ICHEC) for the provision of computational facilities and support.

CRediT authorship contribution statement

David O. Obada: Writing – review & editing, Writing – original draft, Validation, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Emmanuel Okafor:** Writing – review & editing, Writing – original draft, Validation, Resources, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Simeon A. Abolade:** Methodology, Formal analysis, Data curation. **Aniekan M. Ukpong:** Supervision. **David Dodoo-Arhin:** Formal analysis, Data curation. **Akinlolu Akande:** Writing – review & editing, Validation, Supervision, Software, Resources, Project administration, Methodology, Funding acquisition, Conceptualization.

Declaration of competing interest

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

Data availability

Data will be made available on request.

References

[1] A. Kojima, K. Teshima, Y. Shirai, T. Miyasaka, Organometal halide perovskites as visible-light sensitizers for photovoltaic cells, J. Am. Chem. Soc. 131 (17) (2009) 6050–6051.
[2] J. Wang, J. Neaton, H. Zheng, V. Nagarajan, S. Ogale, B. Liu, D. Viehland, V. Vaithyanathan, D. Schlom, U. Waghmare, et al., Epitaxial bifeo3 multiferroic thin film heterostructures, Science 299 (5613) (2003) 1719–1722.
[3] S. Aharon, A. Dymshits, A. Rotem, L. Etgar, Temperature dependence of hole conductor free formamidinium lead iodide perovskite based solar cells, J. Mater. Chem. A 3 (17) (2015) 9171–9178.
[4] L. Meng, J. You, Y. Yang, Addressing the stability issue of perovskite solar cells for commercial applications, Nat. Commun. 9 (1) (2018) 1–4.
[5] E.H. Jung, N.J. Jeon, E.Y. Park, C.S. Moon, T.J. Shin, T.-Y. Yang, J.H. Noh, J. Seo, Efficient, stable and scalable perovskite solar cells using poly (3-hexylthiophene), Nature 567 (7749) (2019) 511–515.
[6] Y. Zhao, F. Ma, Z. Qu, S. Yu, T. Shen, H.-X. Deng, X. Chu, X. Peng, Y. Yuan, X. Zhang, et al., Inactive (pbi2) 2rbcl stabilizes perovskite films for efficient solar cells, Science 377 (6605) (2022) 531–534.
[7] Y. Liu, W. Yan, H. Zhu, Y. Tu, L. Guan, X. Tan, Study on bandgap predications of abx3-type perovskites by machine learning, Org. Electron. 101 (2022), 106426.
[8] L. Chu, S. Zhai, W. Ahmad, J. Zhang, Y. Zang, W. Yan, Y. Li, High-performance large-area perovskite photovoltaic modules, Nano Res. Energy 1 (2) (2022), e9120024.
[9] X.-X. Gao, W. Luo, Y. Zhang, R. Hu, B. Zhang, A. Züttel, Y. Feng, M.K. Nazeeruddin, Stable and high-efficiency methylammonium-free perovskite solar cells, Adv. Mater. 32 (9) (2020), 1905502.
[10] K.-G. Lim, S. Ahn, Y.-H. Kim, Y. Qi, T.-W. Lee, Universal energy level tailoring of self-organized hole extraction layers in organic solar cells and organic–inorganic hybrid perovskite solar cells, Energy Environ. Sci. 9 (3) (2016) 932–939.
[11] P. Mori-Sánchez, A.J. Cohen, W. Yang, Localization and delocalization errors in density functional theory and implications for band-gap prediction, Phys. Rev. Lett. 100 (14) (2008), 146401.
[12] J. Heyd, J.E. Peralta, G.E. Scuseria, R.L. Martin, Energy band gaps and lattice parameters evaluated with the heyd-scuseria-ernzerhof screened hybrid functional, J. Chem. Phys. 123 (17) (2005), 174101.
[13] M. Shishkin, G. Kresse, Self-consistent g w calculations for semiconductors and insulators, Phys. Rev. B 75 (23) (2007), 235102.
[14] X. Cai, F. Liu, A. Yu, J. Qin, M. Hatamvand, I. Ahmed, J. Luo, Y. Zhang, H. Zhang, Y. Zhan, Data-driven design of high-performance masnxpb1-xi3 perovskite materials by machine learning and experimental realization, Light Sci. Appl. 11 (1) (2022) 1–12.
[15] G.S. Thoppil, A. Alankar, Predicting the formation and stability of oxide perovskites by extracting underlying mechanisms using machine learning, Comput. Mater. Sci. 211 (2022), 111506.
[16] M. Del Cueto, C. Rawski-Furman, J. Arago, E. Orti, A. Troisi, Data-driven analysis of hole-transporting materials for perovskite solar cells performance, J. Phys. Chem C 126 (31) (2022) 13053–13061.
[17] V. Venkatraman, The utility of composition-based machine learning models for band gap prediction, Comput. Mater. Sci. 197 (2021), 110637.
[18] O. Allam, C. Holmes, Z. Greenberg, K.C. Kim, S.S. Jang, Density functional theory–machine learning approach to analyze the bandgap of elemental halide perovskites and ruddlesden-popper phases, ChemPhysChem 19 (19) (2018) 2559–2565.
[19] J. Lee, A. Seko, K. Shitara, K. Nakayama, I. Tanaka, Prediction model of band gap for inorganic compounds by combination of density functional theory calculations and machine learning techniques, Phys. Rev. B 93 (11) (2016), 115104.
[20] V. Gladkikh, D.Y. Kim, A. Hajibabaei, A. Jana, C.W. Myung, K.S. Kim, Machine learning for predicting the band gaps of abx3 perovskites from elemental properties, J. Phys. Chem. C 124 (16) (2020) 8905–8918.
[21] Y. Huang, C. Yu, W. Chen, Y. Liu, C. Li, C. Niu, F. Wang, Y. Jia, Band gap and band alignment prediction of nitride-based semiconductors using machine learning, J. Mater. Chem. C 7 (11) (2019) 3238–3245.
[22] G. Pilania, A. Mannodi-Kanakkithodi, B. Uberuaga, R. Ramprasad, J. Gubernatis, T. Lookman, Machine learning bandgaps of double perovskites, Sci. Rep. 6 (1) (2016) 1–10.
[23] S. Rath, G.S. Priyanga, N. Nagappan, T. Thomas, Discovery of direct band gap perovskites for light harvesting by using machine learning, Comput. Mater. Sci. 210 (2022), 111476.
[24] R. Lyu, C.E. Moore, T. Liu, Y. Yu, Y. Wu, Predictive design model for low-dimensional organic–inorganic halide perovskites assisted by machine learning, J. Am. Chem. Soc. 143 (32) (2021) 12766–12776.
[25] S.M. Lundberg, B. Nair, M.S. Vavilala, M. Horibe, M.J. Eisses, T. Adams, D. E. Liston, D.K.-W. Low, S.-F. Newman, J. Kim, et al., Explainable machine-learning predictions for the prevention of hypoxaemia during surgery, Nat. Biomed. Eng. 2 (10) (2018) 749–760.
[26] S. Körbel, M.A. Marques, S. Botti, Stability and electronic properties of new inorganic perovskites from high-throughput ab initio calculations, J. Mater. Chem. C 4 (15) (2016) 3157–3167.

[27] R.D. Shannon, Revised effective ionic radii and systematic studies of interatomic distances in halides and chalcogenides, Acta Crystallogr. Sect. A Cryst. Phys. Diffr. Theor. Gen. Crystallogr. 32 (5) (1976) 751–767.
[28] G. Kresse, J. Furthmüller, Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set, Phys. Rev. B 54 (16) (1996), 11169.
[29] J.P. Perdew, K. Burke, M. Ernzerhof, Generalized gradient approximation made simple, Phys. Rev. Lett. 77 (18) (1996) 3865.
[30] L. Ward, A. Agrawal, A. Choudhary, C. Wolverton, A general-purpose machine learning framework for predicting properties of inorganic materials, npj Comput. Mater. 2 (1) (2016) 1–7.
[31] F.A. Faber, A. Lindmaa, O.A. Von Lilienfeld, R. Armiento, Machine learning energies of 2 million elpasolite (a b c 2 d 6) crystals, Phys. Rev. Lett. 117 (13) (2016), 135502.
[32] L. Weston, C. Stampfl, Machine learning the band gap properties of kesterite i 2- ii- iv- v 4 quaternary compounds for photovoltaics applications, Phys. Rev. Mater. 2 (8) (2018), 085407.
[33] D. P. Kingma, J. Ba, Adam: A method for Stochastic Optimization, arXiv preprint arXiv:1412.6980.
[34] L. Breiman, Random forests, Mach. Learn. 45 (1) (2001) 5–32.
[35] G. Biau, Analysis of a random forests model, J. Mach. Learn. Res. 13 (1) (2012) 1063–1095.
[36] X. Wu, V. Kumar, J. Ross Quinlan, J. Ghosh, Q. Yang, H. Motoda, G.J. McLachlan, A. Ng, B. Liu, P.S. Yu, et al., Top 10 algorithms in data mining, Knowl. Inf. Syst. 14 (1) (2008) 1–37.
[37] O.Z. Maimon, L. Rokach, Data Mining with Decision Trees: Theory and Applications, vol. 81, World scientific, 2014.
[38] T. Chen, C. Guestrin, Xgboost: a scalable tree boosting system, in: Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining, ACM, 2016, pp. 785–794.
[39] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, A. Gulin, Catboost: unbiased boosting with categorical features, Adv. Neural Inf. Process. Syst. 31.
[40] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, T.-Y. Liu, Lightgbm: a highly efficient gradient boosting decision tree, Adv. Neural Inf. Process. Syst. 30.
[41] S. M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, Adv. Neural Inf. Process. Syst. 30.
[42] J. Duffy, Trends in energy gaps of binary compounds: an approach based upon electron transfer parameters from optical spectroscopy, J. Phys. C Solid State Phys. 13 (16) (1980) 2979.
[43] K. Dagenais, M. Chamberlin, C. Constantin, Modeling energy band gap as a function of optical electronegativity for binary oxides, J. Young Invest. 25 (2013) 1–6.
[44] R. Ruh, V.A. Patel, Proposed phase relations in the hfo 2-rich portion of the system hf–hfo 2, J. Am. Ceram. Soc. 56 (11) (1973) 606–607.