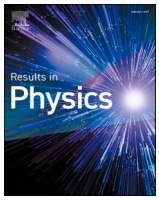
Results in Physics 65 (2024) 107978

Contents lists available at ScienceDirect

Results in Physics

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



# Interpretable machine learning methods to predict the mechanical properties of $ABX_3$ perovskites

S.B. Akinpelu [a,*], S.A. Abolade [b], E. Okafor [c], D.O. Obada [b,d,f,**], A.M. Ukpong [d,e], S. Kumar R. [a,b], J. Healy [a], A. Akande [a,b,***]

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

ARTICLE INFO

*Keywords:*
Perovskites
Density functional theory
Machine learning
Features
Elastic constants
Bulk modulus
Shear modulus
Young modulus

ABSTRACT

This paper proposes the utility of interpretable ensemble learning models for predicting the mechanical properties (bulk, shear and Young moduli) of $ABX_3$ perovskite compounds with the A, B, and X referring to the 3 elements that make the cubic 3-dimensional framework of the perovskite compounds. These models consist of 3 ensemble learning techniques namely CatBoost, Random Forest, and XGBoost. To expand the feature space, robust first-principles density functional theory calculations were used to generate some of the input features, namely elastic constants, density, volume per atom, and ground state energy per atom. The order of the input feature ranking that influences the machine learning (ML) model decisions was then determined. For this, we performed correlation analysis on the multi-dimensional input feature space, suppressed features with high collinearity, and selected features with limited correlation. We trained the three ensemble learning techniques on the desired vectorial input feature representation to predict the mechanical properties. Furthermore, we employed the Shapley Additive Explanations (SHAP) algorithm for analysing the intrinsic decision-making rationality of the ensemble learning models. We measured the performance in the context of the error metrics and coefficient of determination, $R^2$. The results show that XGBoost outperforms other approaches when predicting the shear modulus or Young modulus of the perovskite compounds yielding the least error metrics and the highest $R^2$ value (0.97) in the testing phase. However, both CatBoost and Random Forest outperformed XGBoost when attempting to predict the bulk modulus in the testing phase. The deficiency of the XGBoost in predicting the bulk modulus can be ascribed to an overfitting problem which can occur when the ML model gives accurate predictions for training data but not for test data. Furthermore, the SHAP algorithm provides an insight into the order of feature importance (from highest to lowest). Additionally, we conducted a post-analysis using a holistic ranking to analyse the relative importance of the SHAP feature impact comprehension for the examined ensemble learning techniques. Our findings indicate that the elastic constants are the most important input features influencing the predictive decision of the ensemble learning models.

* Corresponding author at: Modelling & Computation for Health And Society (MOCHAS), Atlantic Technological University, Ash Lane, Ballytivnan, Sligo, F91 YW50, Ireland.
** Corresponding author at: 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 at: Modelling & Computation for Health And Society (MOCHAS), Atlantic Technological University, Ash Lane, Ballytivnan, Sligo F91 YW50, Ireland.
*E-mail addresses:* babatunde.akinpelu@research.atu.ie (S.B. Akinpelu), david.obada1636@associate.atu.ie (D.O. Obada), akinlolu.akande@atu.ie (A. Akande).

https://doi.org/10.1016/j.rinp.2024.107978
Received 1 June 2024; Received in revised form 24 August 2024; Accepted 15 September 2024
Available online 18 September 2024

## Introduction

Halide perovskites have emerged as a highly significant category of materials, owing to their high absorption coefficient, direct bandgap, extended carrier lifetime, elevated balanced hole and electron mobility, cost-effectiveness, and easily implementable deposition techniques, among other attributes. Consequently, there has been an increase in their potential for applications in photovoltaics, light emitting devices, Lasers, Gas sensors, X-ray detectors, Transistors, memory devices, etc. [1–3]. These compounds have a unique $ABX_3$ stoichiometry where A could be an alkaline, alkaline-earth, or lanthanide cation, B could be a transitional metal, and X could be a halogen or chalcogen. This perovskite structure has a cubic 3-dimensional framework with corner-sharing $BX_6$ octahedral molecular geometry [4]. They are classes of compounds that have been utilised for energy applications, such as photovoltaics and thermoelectrics, to mention a few [5]. The determination of important properties, such as absorption coefficients and bandgaps, and suitable mechanical properties of these materials for various applications, is primarily achieved through a combination of theoretical and experimental methodologies. For these materials to be viable for energy applications, properties like electronic, optical, and mechanical properties are crucial in understanding their performance for practical applications [6]. Some properties, such as mechanical properties, depend on the crystal structures and composition of the compounds, affecting stress, strain, and pressure.

Theoretically, the mechanical behaviour of perovskite materials, which reveals valuable insights into the response of the materials to stress and deformation, is usually determined by altering the elemental contribution of different prototypes or structures [7]. However, a quick assessment of the mechanical performance of many perovskite compounds through experimental and computational approaches is quite time-consuming and rigorous. To mitigate the substantial effort and labour associated with this approach, scientists tend to employ a comprehensive prediction and screening methodology supported by statistical learning techniques. In the recent past, these techniques, known as machine learning (ML), have proven to be effective within the domain of material informatics. These methods have demonstrated their ability to forecast material properties and enhance the accuracy of density functional theory (DFT) predictions. This approach uses a dataset of the material's structural information like the elemental composition, band gaps, bulk modulus, and formation energy as inputs, and then predicts the desired outputs [4,8,9].

ML has persistently been utilised in the advancement of inorganic materials. Zhuang et al [10] worked on the improved elastic models of amorphous alloy design using ML. In their study, effectively optimised, generated, and interpreted elastic predictive models with composition and structure descriptors were used as inputs with Lasso models showing a high level of accuracy with $R^2$ values of 0.97 and 0.98 for shear modulus and bulk modulus, respectively. The experimental results show a strong correlation with the predicted results. Their study indicates that molar volume and Pauling electronegativity are crucial structural descriptors that affect both the shear and bulk moduli of amorphous alloys. Lee et al [11], conducted a study to identify the crucial features for predicting two elastic moduli, namely the bulk and shear moduli. They analysed 17,051 datasets of inorganic solids, using 4399 optimal descriptors and four feature selection methods. The Light Gradient Boosting Model (LGBM) produced the best performance with an $R^2$ value of 0.89 ($\pm$0.02) for the shear modulus and 0.91($\pm$0.03) for the bulk modulus. The study made use of four feature selection methods which include Pearson's correlation coefficient-based feature selection (PFS) amongst others. The study concluded that the identified features like X-ray diffraction powder pattern and element properties, in addition to the method of feature selection (importance-based feature selection) could be useful for the prediction of bulk and shear moduli of these compounds, and for future developments in inorganic materials.

In other related studies, Bishnoi et al. [12] predicted the mechanical properties of inorganic glass using a Gaussian process regression (GPR) model with a dataset of over 100,000 glass compositions. They demonstrated that the models built in their study outperform the current cutting-edge ML models. Furthermore, they showed that the GPR models can accurately represent composition-dependent physics, even in areas with limited training data. In a study conducted by Xie and Grossman [13], they created a framework called crystal graph convolutional neural networks (CGCNN) that allows the direct determination of material attributes (in some cases, shear, and bulk moduli of perovskite materials) by analysing the connections between atoms in a crystal. This framework provided a comprehensive and easily understandable description of crystalline materials. Wang et al. [14] used extremely randomised trees and deep neural networks to construct a hierarchical model for predicting ternary properties using ML models trained on binary data. The model focused on the elastic properties of bulk and shear moduli. The model achieved mean absolute errors of 0.56 GPa and 1.49 GPa in the bulk and shear modulus predictions, respectively. Their study highlighted the importance of compositions, temperature, and ordering effects in the predictions. An ML model based on the support vector regression algorithm was used in the work of Mansouri Tehrani et al. [15] to predict elastic moduli distribution. The approach screened 118,287 crystal structure databases for materials with the highest bulk and shear moduli. Complementary experimental studies confirmed the prediction accuracy and ultra-incompressibility, demonstrating the effectiveness of ML techniques in identifying functional inorganic materials.

To further enhance the generalisation of ML models, feature selection is crucial for improving prediction accuracy and reducing computational costs. It helps identify relevant features for target properties. Studies have shown that reducing features without compromising model performance can improve prediction accuracy. For instance, a study on perovskite oxides reduced 791 features to 70, achieving a better $R^2$ of 0.89 [16]. In a study conducted by Revi et al. [17], the variation inflation factor demonstrated that reducing the number of features in a model can still improve the prediction of the elastic properties of multicomponent alloys. Additionally, features in automaticity can be more efficient for large-scale data predictions [18].

Despite the progress made in this area of scientific importance, studies that have utilised ML methods for predicting the mechanical properties of selected materials are scarce. In this study, we aim to further the development of supervised learning (SL) models by predicting the elastic characteristics of perovskite compounds across a wide range of chemical compositions. Once the model's generalization capability has been established, then such a model finds relevance in solving specific design challenges, which are essential parts of the broader process of finding new perovskite materials in high dimensional composition space. The main objective of this study is to train/develop ML models that can accurately predict important elastic characteristics. Large-scale finite element models, which are used to analyse deformation and microstructure changes of perovskite compounds, can benefit from these elastic characteristics. These attributes mainly encompass characteristics such as elasticity constants, bulk, shear and Young moduli, hardness, and fracture toughness. An essential addition to this study is the first principles DFT calculations of the bulk, shear, and Young moduli for the 199 perovskite compounds used as the target output.

**Contributions:** The mechanical properties based on bulk, shear, and Young moduli of 199 perovskite compounds with the $ABX_3$ stoichiometry were predicted using three ensemble learning techniques. Element descriptors, elastic constants, and physical descriptors were used as the input features. Feature engineering was carried out to analyse and screen the strongly correlated descriptors and consequently passed to ensemble learning algorithms. The SHapley Additive exPlanations (SHAP) was employed for analysing the relative feature importance or impact to predict the mechanical properties (bulk, shear, and Young moduli) of the perovskite compounds. To gain further insight into the

post-model interpretability, a computationally efficient holistic ranking method proposed by Obada et al. [8] was used to rank the input features accurately and globally concerning the different ML algorithms. Furthermore, the impact of these features on the moduli predictions from the physics standpoint was discussed.

## Methodology

### *Data Collection and DFT calculations*

The dataset used in this study was an extension of the datasets initially reported from first principles calculations of Korbel et al. [19]. This was further utilised by Obada et al. [8] to predict the indirect and direct bandgaps of 199 cubic perovskite compounds. The additions to the previously used datasets are twofold:

1. DFT calculations were conducted at the PBE GGA level to determine ground state energy per atom, volume per atom, and density for the 199 perovskite compounds.
2. Additionally, the stiffness matrix for these perovskite materials was obtained using a first principle-based stress–strain method implemented in the Vienna Ab-initio Simulation Package (VASP) [20]. An energy cut-off of 420 eV and a Monkhorst-Pack K-points mesh (4*4*4) were applied during the calculations. The Voigt-Reuss-Hill approximation was employed to translate elastic tensors into elastic moduli, including bulk, shear, and Young moduli. The study aimed to provide a comprehensive understanding of the mechanical properties of the investigated perovskite materials.

### *Data Description*

Various descriptors such as electronegativity, covalent radius, first ionization energy, and periodic table row, along with elastic constants of $C_{11}$, $C_{12}$, and $C_{44}$ (mainly for cubic materials) and physical descriptors (density, volume per atom, ground state energy per atom), were used as input features for each element in the $ABX_3$ compound space. The dataset employed comprises 23 input features per compound, with the calculated DFT bulk, shear, and Young moduli set as the target output.

In this investigation, the mechanical properties of the cubic perovskite solid materials using key mechanical parameters such as bulk, shear, and Young's moduli, which depend on the elastic constants $C_{11}$, $C_{12}$, and $C_{44}$ were evaluated. The dimensional strength of a cubic crystal system is determined by the interplay between these three specific elastic constants. Collectively, these constants define the fundamental characteristics of the material's behaviour. Specifically, $C_{11}$ represents the material's ability to withstand strain, $C_{12}$ indicates its tendency towards shear stress, and $C_{44}$ demonstrates its resistance to shear deformation[21]. Understanding the bulk modulus of a material is crucial for assessing its ability to withstand volume changes caused by compression. Materials with higher bulk modulus values exhibit greater resistance to volume changes compared to those with lower values. The shear modulus is essential for evaluating a materials' strength and resistance to plastic deformation. A higher shear modulus indicates increased hardness and a greater ability to resist deformation. Young's modulus measures the inherent stiffness of a material, with higher values suggesting greater rigidity[22].

Table 1 shows the list of input features, these include; elastic constants ($C_{11}$, $C_{12}$, and $C_{44}$), atomic radius of element A (AR(A)), atomic radius of element B (AR(B)), atomic radius of element X (AR(X)), Density (DST), volume per atom(VPA), ground state energy per atom(GSA), indirect band gap(IBG), direct bandgap (DBG), row of element A (Row (A)), row of element B (Row(B)), row of element X (Row(X)), first ionization energy of element A (FIE(A)), first ionization energy of element B (FIE(B)), first ionization energy of element X (FIE(X)), covalent radius of element A (CR(A)), covalent Radius of element B(CR(B)), covalent radius of element X (CR(X)), Pauling electronegativity of element A (PE(A)), Pauling electronegativity of element B (PE(B)), Pauling electronegativity of element X (PE(X)).

**Table 1**
List of input features and their abbreviation.

| Features | Meaning |
|---|---|
| $C_{11}$, $C_{12}$, $C_{44}$ | Elastic Constants |
| AR(A) | Atomic Radius of Element A |
| AR(B) | Atomic Radius of Element B |
| AR(X) | Atomic Radius of Element X |
| DST | Density |
| VPA | Volume per atom |
| GSA | Ground State Energy per Atom |
| DBG | Direct Band Gap |
| IBG | Indirect Band Gap |
| ROW(A) | Row of Element A |
| ROW(B) | Row of Element B |
| ROW(X) | Row of Element X |
| FIE(A) | First Ionization Energy of Element A |
| FIE(B) | First Ionization Energy of Element B |
| FIE(X) | First Ionization Energy of Element X |
| CR(A) | Covalent Radius of Element A |
| CR(B) | Covalent Radius of Element B |
| CR(X) | Covalent Radius of Element X |
| PE(A) | Pauling Electronegativity of Element A |
| PE(B) | Pauling Electronegativity of Element B |
| PE(X) | Pauling Electronegativity of Element X |

### *Data Splitting*

Due to the scarcity in the amount of available data and the need to foster a machine learning model with good generalisation capability, preliminary experiments were conducted using the bulk modulus to investigate three forms of data splits presented in the following ratios: 95 %: 5 %, 90 %: 10 %, and 80 %: 20 % for the training set and testing set, respectively. The early insight drawn from the supervised learning models trained on 95 % of the used data yielded the best performance across the evaluated metrics relative to the assessment of the remaining data distribution. The result displayed in Table 2 shows that as the proportion of the training dataset decreased from 95 % to 90 % and 80 %, the errors (RMSE, MAE) for each ML models decreases respectively during the training phase. Also, the results further indicate a decrease in RMSE values during the testing phase as the training data size increases and suggest that the model's performance on unseen data improved with a larger training dataset. This excellent generalisation prowess is a critical factor in our decision-making process and influenced our choice of 95 % and 5 % for the training and testing sets. Four more random splits were performed while ensuring no overlap in the testing and training sets, resulting in a five-fold cross-validation.

However, other techniques are available, such as the stratified split [24], which maintains the same proportion of classes in each subset and is commonly used for classification tasks. Another method is the temporal split, which involves dividing the data based on time, ensuring that the training set consists of data from an earlier period while the test set contains data from a later period[25]. Furthermore, cross-validation involves partitioning the data into multiple folds, allowing the model to be trained and validated iteratively on different subsets[26]. Additionally, Leave-One-Out Cross-Validation (LOOCV) uses each sample once as a test set while utilising the remaining data for training. However, this approach is computationally expensive due to the high number of iterations required[27]. Therefore, Random Splitting is preferred due to its ability to provide a representative distribution of the data, minimise bias, and support the development of robust and generalisable machine learning models[23].

### *Feature engineering*

Feature engineering is a crucial preprocessing step before creating

**Table 2**
The splitting ratio of the datasets into 95%: 5%, 90%: 10%, and 80%: 20% to assess its influence on the models' performance.

| Moduli | Technique | Splitting percentage | Train | | | Test | | |
|---|---|---|---|---|---|---|---|---|
| | | | MAE | RMSE | $R^2$ | MAE | RMSE | $R^2$ |
| Bulk Modulus | CatBoost | 95 %: 5 % | 0.9286 | 1.1499 | 0.9995 | 9.6610 | 16.5475 | 0.9541 |
| | | 90 %: 10 % | 0.7308 | 0.9094 | 0.9999 | 18.1095 | 31.1554 | 0.8858 |
| | | 80 %: 20 % | 0.5903 | 0.7253 | 0.9999 | 17.3841 | 28.8656 | 0.9086 |
| | XGBoost | 95 %: 5 % | 0.0033 | 0.0048 | 1.0000 | 9.6877 | 17.8521 | 0.9474 |
| | | 90 %: 10 % | 0.0020 | 0.0029 | 1.0000 | 23.0971 | 48.4216 | 0.7242 |
| | | 80 %: 20 % | 0.0020 | 0.0028 | 1.0000 | 18.1363 | 31.4554 | 0.8915 |
| | RANDOM FOREST | 95 %: 5 % | 4.3657 | 9.3795 | 0.9898 | 7.5909 | 13.0009 | 0.9648 |
| | | 90 %: 10 % | 4.0107 | 9.3941 | 0.9894 | 13.2298 | 22.3424 | 0.9413 |
| | | 80 %: 20 % | 3.6360 | 8.3572 | 0.9915 | 15.1460 | 24.3747 | 0.9349 |

ML models. To ensure that ML algorithms are trained on independent input features as well as reducing data redundancy and memory usage, relevant subsets of features that are independent were selected using the Pearson Correlation Coefficient (PCC) approach[28,29]. The formula for deriving the PCC between variables $x_i$ (input features) and $y_i$ (output features) is shown in Equation (1):

$$r_{x,y} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}} \quad (1)$$

The greater the absolute value of $r_{x,y}$, the stronger the correlation. The PCC value ranges from + 1 to −1. Here, $r_{x,y} = +1$ represents a completely positive linear correlation, while $r_{x,y} = -1$ stands for a completely negative linear correlation. On the other hand, $r_{x,y} =$ 0 means no linear correlation. In this study, the correlation between the input features was assessed, and the graphical representation is shown in Fig. 1. A threshold of 0.85 taken from a related study [30] was set to eliminate one of the two input features strongly correlated with each other based on the target outputs. Since the PCC plot follows the same profile for the three target outputs namely, bulk, shear, and Young moduli, one PCC plot displayed in Fig. 1 represents the three. After conducting this screening, the initial datasets of twenty-three (23) input features were reduced to fifteen (15). Eight features ($C_{12}$, AR(X), IBG, VPA, FIE (X), Row(X), AR(A), and FIE(A)) were dropped due to the same value of correlation with the other selected input features. For instance, DBG and IBG have the same number of correlations with each other with correlated values greater than the set threshold of 0.85 (Fig. 1).

The reduced input features distribution was transformed before passing the compelling input features and corresponding labels to train SL models with potential generalisation capability [31]. The performance and accuracy of the ML models were enhanced by normalising the input features to a scale [0,1]. This was done using the min–max normalisation scheme, as presented in Equation (2).

$$x_{norm} = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (2)$$

Where $x$ represents the raw input features. $x_{min}$ and $x_{max}$ are the minimum and maximum values of $x$ in the dataset. The normalised input features can be expressed as $x_{norm}$. Finally, the SL algorithms were fed with the effective normalised input feature.

*Ensemble learning algorithms*

Ensemble learning methods have been widely recognised in the field of data mining and machine learning in the past decade. They combine numerous models into a single entity, typically surpassing the accuracy of its components. As such, brief explanations of the model used in this study are provided. It is worth noting that models perform better when trained on larger datasets. Nevertheless, it is also essential to note that the choice of algorithms significantly impacts the model's accuracy. For

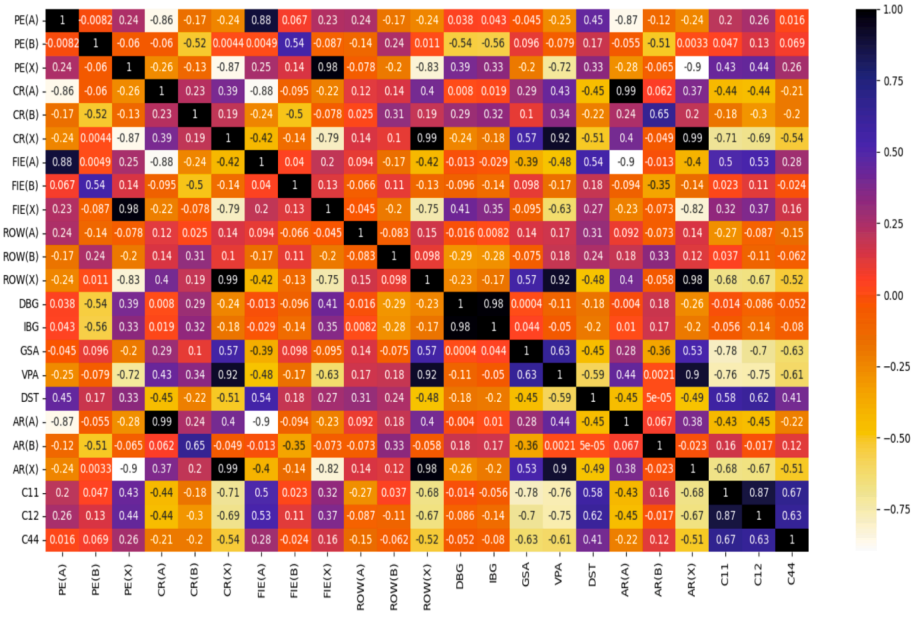



**Fig. 1.** Correlation plot heatmap of the input feature relationship of $ABX_3$ perovskites for the respective mechanical properties: bulk, shear, and Young moduli. $C_{11}$, $C_{12}$ and $C_{44}$: Elastic constants.

instance, algorithms like Random Forest, XGBoost, and CatBoost, utilized for this task, are adept at handling machine learning with smaller datasets [32]. These ML models can minimise overfitting through techniques such as regularisation and ensembling. Overfitting happens when a model performs well on the training data but poorly on the test data, indicating that it has captured the noise and intricacies of the training data rather than generalising to new, unseen data. Ensemble methods combine the predictions of multiple models to enhance robustness and generalisation. For example, XGBoost and CatBoost [33,34] utilise gradient boosting with regularisation techniques to penalise complex models and prevent them from closely fitting the training data. Conversely, Random Forest averages the predictions from many decision trees to reduce variance and enhance generalisation[35]. For all the ensemble learning techniques, default hyperparameter values were employed.

*CatBoost*

CatBoost is an ensemble learning-based SL algorithm that employs a decision tree for performing classification and regression tasks[36,37]. The algorithm relies on two primary attributes: the capacity to handle data categorically, and the utilisation of gradient boosting (base learners) for learning sequentially from input data. The paradigm of gradient boosting involves the computation of the weighted sum of multiple decision trees. The accuracy of subsequent trees is enhanced by using the outcomes of preceding trees. The CatBoost algorithm improves upon the original gradient-boosting approach to expedite its implementation [36]. CatBoost is a distinctive approach based on decision trees that streamlines the process of data pre-processing. The model can efficiently process a mixture of categorical and non-categorical variables by employing ordered encoding to substitute categorical characteristics. One distinguishing characteristic of CatBoost is the utilisation of symmetric trees, wherein decision nodes at each level of depth employ identical split conditions. This methodology exhibits enhanced computational efficiency compared to alternative methods such as XGBoost, while simultaneously preserving crucial attributes like cross-validation, regularisation, and support for missing values from previous algorithms. Furthermore, CatBoost demonstrates strong performance across datasets of varying sizes, including both small and large datasets [38].

*XGBoost*

The XGBoost algorithm is a popular ensemble learning method for gradient-boosting machines. Its efficient problem-solving capabilities require minimal feature engineering [39], support distributed computing, tree pruning, is characterised by the prowess to handle missing data, and can be used to compute relative importance for each feature when making a certain prediction. It is a popular choice in industry due to its high performance and generalisation prowess [40]. It is important to highlight that to achieve high predictive performance, there is a need to consider proper tuning of hyperparameters.

*Random Forest*

Random Forest is an ensemble learning algorithm used for executing either classification or regression tasks. This algorithm depends on the principle of combining various base learners to tackle complex predictive decisions while improving the model's performance [41]. This method uses multiple decision trees, each trained on a different subset of a given dataset, to enhance its predictive accuracy. Unlike a single decision tree, which may be prone to overfitting, the Random Forest algorithm aggregates the predictions from all the trees and uses the majority vote to generate the final output. This approach enables it to handle large datasets with high dimensionality and enhances the robustness and accuracy of the model [42]. During the experiments, a random forest ($Y_{rf}$) comprising of 10,000 base estimators was utilized.

*Performance metrics*

To assess the effectiveness of each ensemble learning method, three performance metrics were utilized during the experiments: mean absolute error (MAE), coefficient of determination ($R^2$), and root mean square error (RMSE). It is assumed that the dataset contains N samples, $\widehat{y}_i$ is the predicted value for the $i-th$ data point, $y_i$ is the known value for the $i-th$ datapoint, $\bar{\boldsymbol{y}}$ is the mean of the predicted value of $y_i$ and these metrics are defined as follows in these equations (3–5):

$$MAE = \frac{\sum_{i=1}^{N} |y_i - \widehat{y}_i|}{N} \tag{3}$$

$$R^2 = 1 - \frac{\frac{\sum_{i=1}^{N} (y_i - \widehat{y_i})^2}{N}}{\frac{\sum_{i=1}^{N} (y_i - \bar{y})^2}{N}} \tag{4}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} (y_i - \widehat{y}_i)^2}{N}} \tag{5}$$

To demonstrate the accuracy of a model, it is important to consider the RMSE, and MAE values. If these values are close to zero, it suggests that the model has a perfect fit and is highly reliable. Additionally, the $R^2$ value is another important metric to evaluate the model's performance. A value close to 1 indicates that the model is highly accurate and dependable [8].

*Shapley Additive exPlanations (SHAP) and the novel holistic ranking method*

ML methods are black box models created directly from data by an algorithm, meaning that it is difficult to understand how variables are being combined to make predictions. This is why researchers use explainable artificial intelligence (XAI), which helps in understanding the features contributing the most during ML predictions [43]. One example of an XAI model is called SHAP. This is a tool employed for interpreting complex models. The SHAP algorithm zooms into each feature and decides how much it contributes to the final prediction [44]. This approach makes the predicted outcomes easy to understand. In this study, two different explainers were used, tree-based and kernel-based to interpret the ensemble learning techniques. The mathematical details of the approach can be found in a previous study [8].

*Experimental Setup*

For all the experiments, Python 3.9.13 was used on a standard PC with a 12th Gen Intel(R) Core (TM) i5-1235U processor running at 1.30 GHz and 16.0 GB of RAM. The Scikit-learn framework was used, and all codes were executed on a Jupyter Notebook server version 6.4.12.

## Results

The performance metrics on the training and test dataset using fivefold cross-validation reveal distinct patterns of predictive performance as shown in Table 3. Each performance is recorded for each cross-validation and the average was computed. The MAE, RMSE and $R^2$ values show that the training and test data for each cross-validation have a good model fit. Analysis of the three ensemble learning algorithms—CatBoost, XGBoost, and Random Forest—used in this study in predicting the bulk, shear, and Young moduli of the perovskite compounds was conducted.

For bulk modulus, XGBoost exhibited accuracy with a remarkably low MAE of 0.003, RMSE of 0.005, and $R^2$ of 1.00, outperforming CatBoost and Random Forest during the training phase. Similar trends were observed for shear and Young moduli, where XGBoost consistently displayed the lowest errors and highest $R^2$ values. CatBoost demonstrated

**Table 3**
Performance metrics of the ensemble learning methods for the average values for bulk, shear, and Young moduli.

| Moduli | Technique | Train | | | Test | | |
|---|---|---|---|---|---|---|---|
| | | MAE | RMSE | $R^2$ | MAE | RMSE | $R^2$ |
| **Bulk Modulus** | CATBOOST | 0.92859 | 1.14993 | 0.99949 | 9.66104 | 16.54747 | 0.95408 |
| | XGBOOST | 0.00330 | 0.00482 | 1.00000 | 9.68772 | 17.85206 | 0.94736 |
| | RANDOM FOREST | 4.36572 | 9.37954 | 0.98978 | 7.59087 | 13.00090 | 0.96480 |
| **Shear Modulus** | CATBOOST | 0.34733 | 0.44742 | 0.99991 | 4.58824 | 8.76823 | 0.94950 |
| | XGBOOST | 0.00186 | 0.00268 | 0.99999 | 3.51651 | 6.83419 | 0.96719 |
| | RANDOM FOREST | 1.78640 | 3.21031 | 0.99562 | 4.33551 | 7.64518 | 0.96024 |
| **Young Modulus** | CATBOOST | 0.98519 | 1.25778 | 0.99989 | 11.52757 | 21.30316 | 0.95379 |
| | XGBOOST | 0.00380 | 0.00541 | 0.99990 | 9.20291 | 15.65350 | 0.97439 |
| | RANDOM FOREST | 4.28718 | 8.40932 | 0.99520 | 10.19858 | 18.87071 | 0.96436 |

commendable performance across all modulus types, striking a balance between accuracy and computational efficiency. While Random Forest proved effective, higher errors were noticed as compared to other models. The exceptionally high $R^2$ values across all models indicate, in addition to the effectiveness of the feature engineering process, a significant capture of variance in the data. Overall, these findings contribute to understanding the predictive capabilities of ML models in the realm of materials science, with the XGBoost model particularly promising, comparatively, for accurate modulus predictions. The graphical illustrations of the training phase for all the models considered in this study are shown in Fig. 2. It is worth noting that during the training phase, negative values were noticed in the correlated plot for the moduli, respectively. This is attributed to negative elastic constants in some of the examined materials such as $RbCuBr_3$, $TlMnF_3$, $RbMnCl_3$, and $RbSnCl_3$. Nevertheless, negative moduli suggest that a surface constraint is needed in materials to attain overall stability[45]. For instance, a negative bulk modulus is because of a restrained lattice of crystalline materials, while a negative Young modulus is the result of elastic stored energy at equilibrium and is not influenced by resonance or other inertial effects. Furthermore, the behaviour of ferroelastic materials near phase transitions further suggests a negative shear modulus. These negative parameters noticed in some of the investigated cubic $ABX_3$ perovskites have also been observed in other cubic crystal structures[45–47].

For the test phase, the graphical illustrations for all the models are shown in Fig. 3. For the bulk modulus, CatBoost demonstrates competitive metrics with a MAE of 9.66 and an $R^2$ of 0.95, while Random Forest returned the lowest RMSE at 13.00 and lowest MAE at 7.59 as noticed in

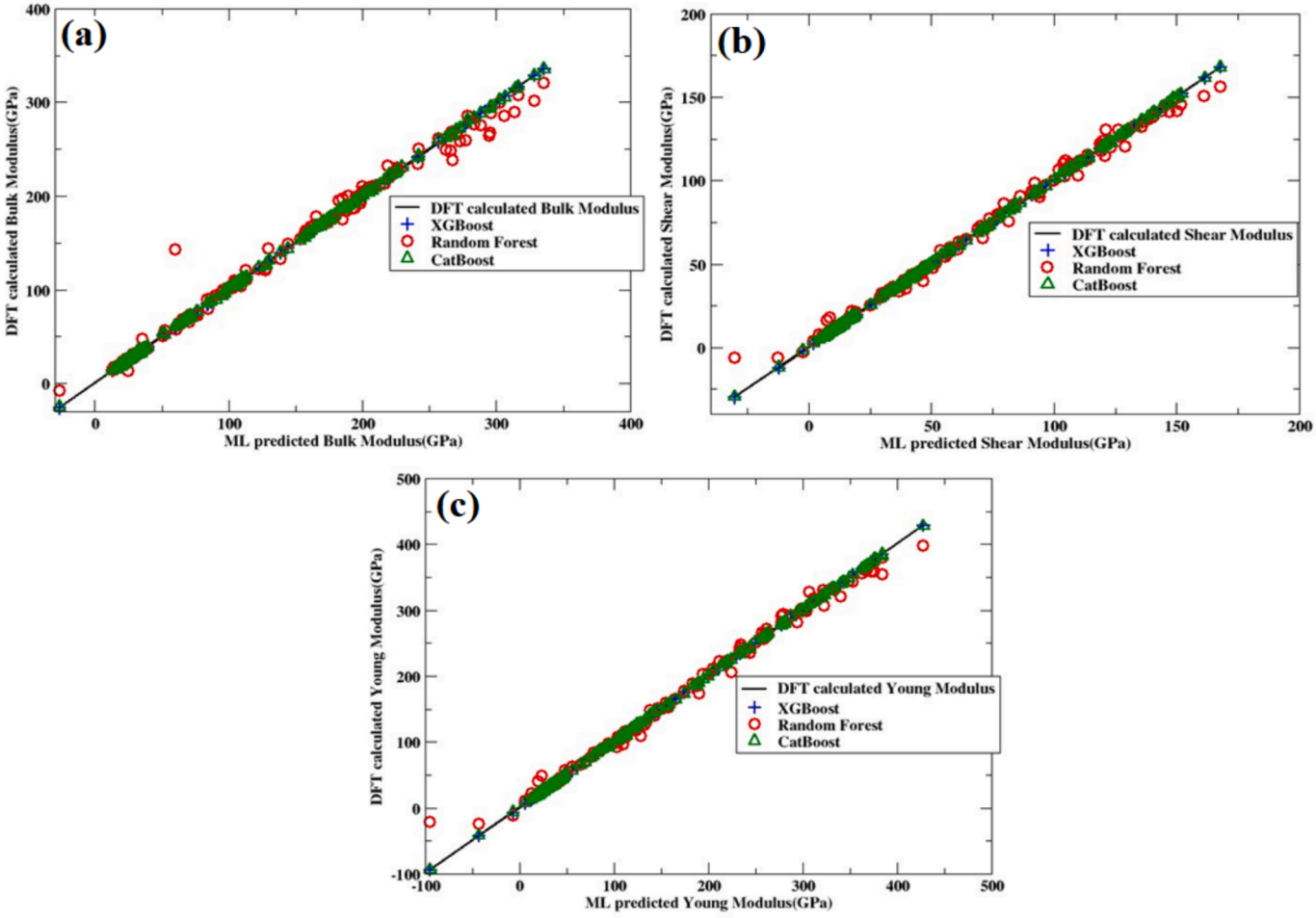



**Fig. 2.** Correlation plots for the training phase for each of the ensemble learning methods. (a) Bulk modulus; (b) Shear modulus; (c) Young modulus.

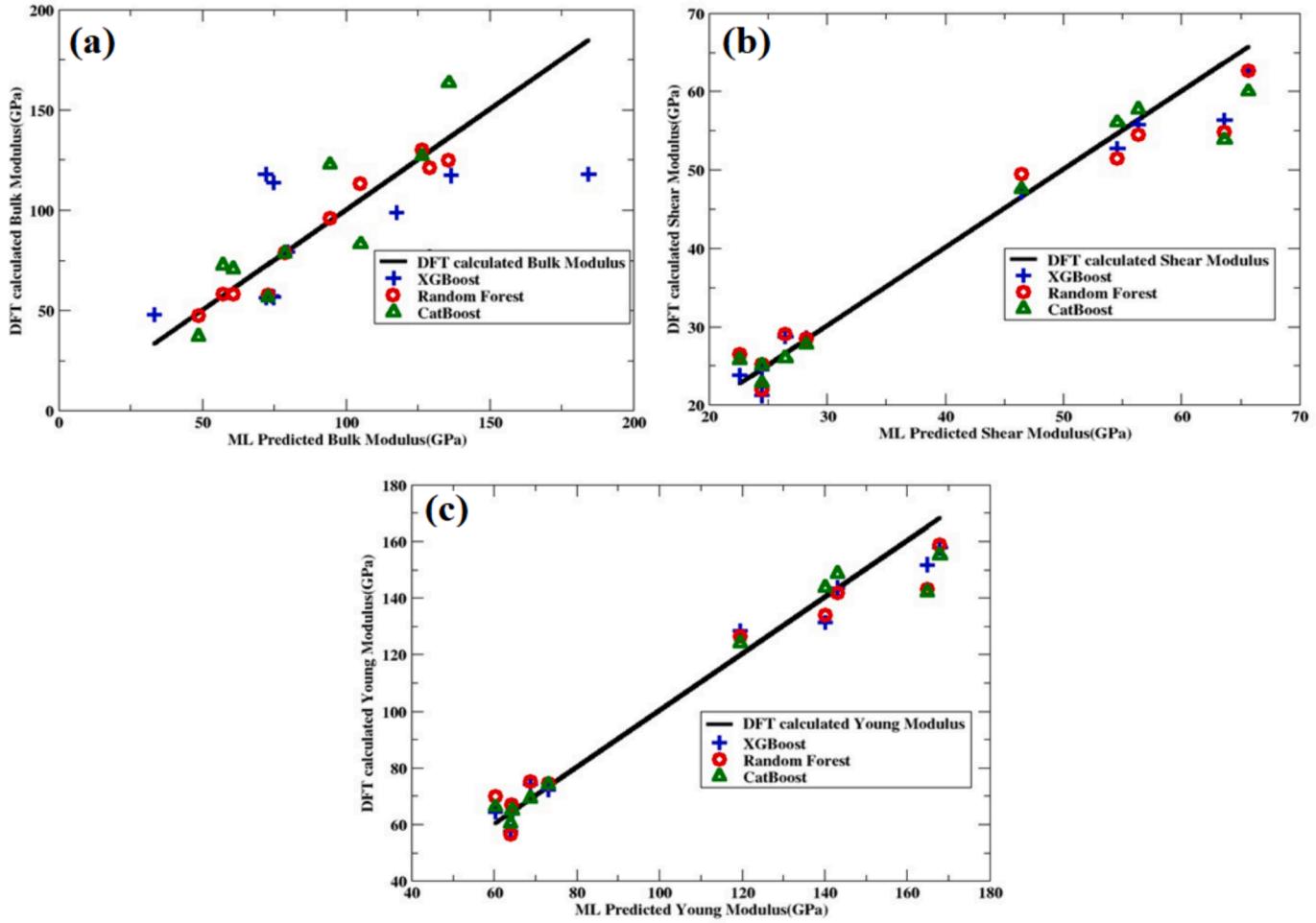



**Fig. 3.** Correlation plots for the testing phase for each of the ensemble learning methods. (a) Bulk modulus; (b) Shear modulus; (c) Young modulus.

Table 3. The shear modulus shows XGBoost as the standout performer with the lowest MAE of 3.52, and RMSE of 6.83, with a remarkable $R^2$ of 0.97. In the context of the Young modulus, XGBoost maintained its superior performance with MAE of 9.20 and an outstanding $R^2$ of 0.97, while Random Forest remained competitive. It was noticed that the RMSE and MAE of the training dataset are lower than that of the test though with a good $R^2$ value close to one. It is a well-known fact that if a linear regression model fits the data perfectly, then the RMSE and MAE values should be 0, and if the fit is less than perfect, then the values should be positive. Similarly, the $R^2$ is expected to have a value of 1 if the linear regression model fits the data perfectly, which means that the Mean Square Error (MSE) = 0. However, interpreting the sole values of MSE and RMSE can be difficult since they have an upper bound of $+\infty$. For example, an MSE value of 0.7 does not provide much information about the overall quality of a regression model. The value could represent both an excellent and a poor regression model, and this cannot be determined unless the maximum MSE value for the regression task is provided or if the distribution of all the ground truth values is known. The same concept applies to other rates having an upper bound of $+\infty$, such as RMSE [48]. However, the test errors are higher because of the number of datasets used for the testing phase and the statistical variance that occurs within the test datasets. This has been observed in some ML works [49,50]. Above all, XGBoost consistently outperformed, demonstrating its efficacy throughout the predicting process. The analysis of all the ensemble learning techniques viz-a-viz the performance metric provides a better understanding of the strengths and weaknesses of each model, facilitating an informed decision in choosing the most suitable ML algorithms.

*Explainability analysis*

The SHAP concept which originates from the game theory[51], offers an explanatory framework for understanding the importance of the input features for different ML methods. This approach was employed to provide a theoretical explanation of the influence of each feature used in predicting the bulk, shear, and Young moduli of the perovskite compounds. The SHAP graphs are displayed in Figs. 4, 5, and 6. The top-ranked feature is situated at the apex, while the hierarchy of features descends along the feature axis in accordance with their respective significance as shown in the SHAP plots. The mean magnitude represents the average impact of a particular feature on the output of the model. A greater average magnitude indicates a more pronounced influence on the projected outcome. The Top performing input feature contributing to the prediction of the moduli includes the elastic constants, GSA, PE(X) and DST while the least performing are the ROW(A), ROW(B), CR(A), FIE(B). The explanation of how these aforementioned input features contributed to the prediction of the moduli is further explained in the next paragraphs.

Bulk modulus, a physical property that quantifies the ability of a substance to withstand bulk compression, provides an insight and explanation for the SHAP plots. From a physics standpoint, bulk modulus can be defined as the quantitative measure of the ratio between the applied pressure and the resulting strain exerted on a given material. The bulk modulus depends on the ionic properties of materials, such as the bond length, due to the symmetric and antisymmetric portions in the unit cell[52]. The variation in the crystal structure inevitably impacts the bulk modulus. In general, smaller AR and CR tend to form stronger

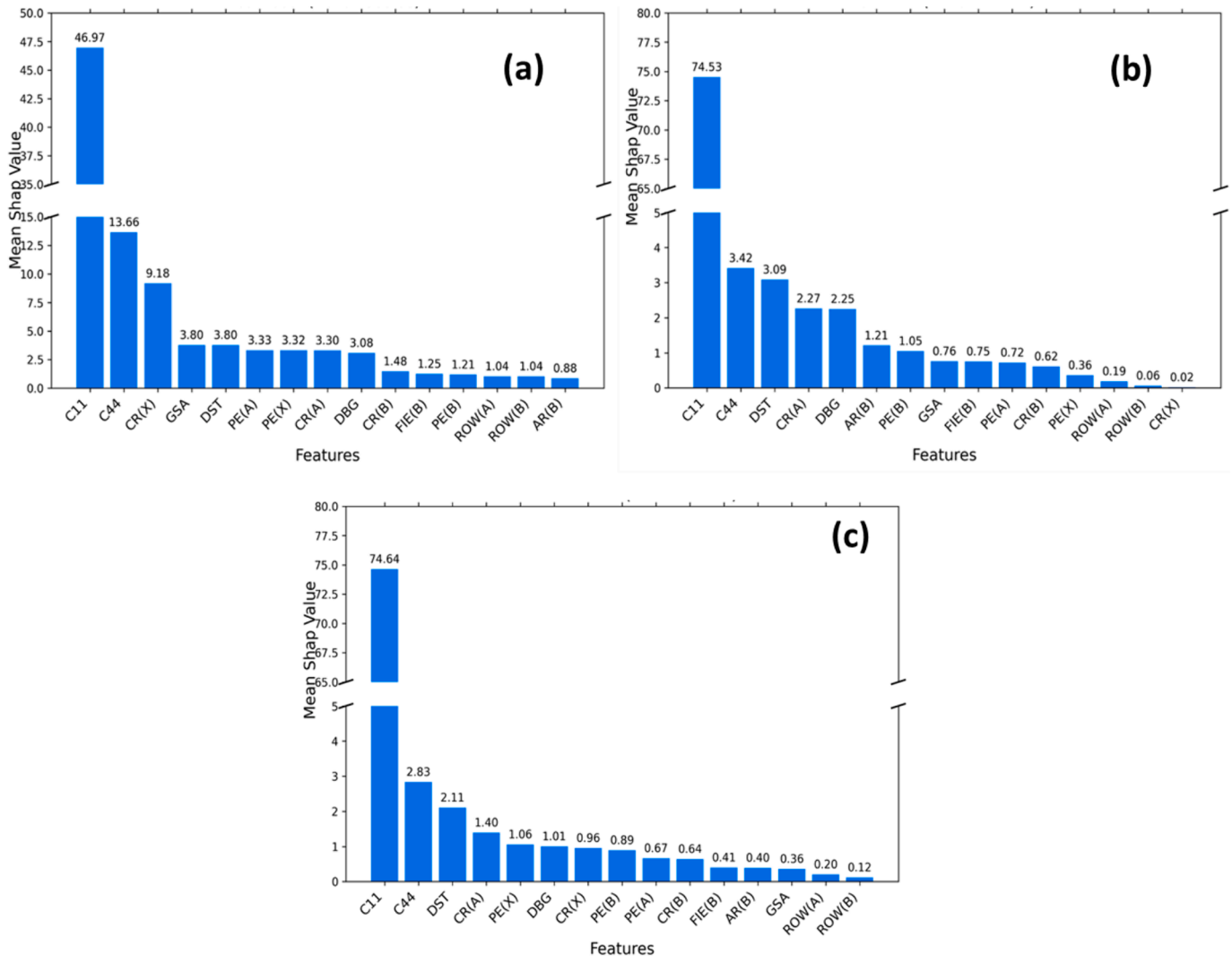



**Fig. 4.** Explainability of the supervised learning model prediction of the bulk modulus while revealing the feature importance influencing the model prediction. CatBoost: (a); XGBoost (b); Random Forest (c).

covalent bonds and result in higher bulk modulus. This is because smaller atoms can establish closer bonding distances and stronger bond formation, leading to resistance to compression. This can be confirmed from Fig. 4 which represents the SHAP plots for the bulk modulus for the three ensemble learning methods, respectively. It was also noticed that DBG influences the bulk modulus of this class of material. This is because the bulk modulus depends on the interatomic bonding and atomic arrangement within the material which is influenced by the electronic crystal structure. The predictions of the bulk modulus also reveal that DST, CR(A), and PE(B) as important features. DST represents how resistant a material is to changing volume under compression applied on all sides of the material which can affect their ductility and brittleness. A-sites are the alkali metal ions that act as a spacer in the compounds' composition and stabilise the compound while element X is the halides and oxides. The PE 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 these perovskite compounds would typically involve the interaction of electrons within the elemental composition which can strengthen the materials and improve their mechanical properties.

Similarly, the shear modulus is the ratio of the shear stress to the shear strain, while Young modulus is the ratio of the tensile stress to the tensile stain. From the SHAP analysis in Figs. 5 and 6, it was observed that the PE(X), and CR(X) are top-ranked features in addition to the elastic constants and DST. PE is a measure of the tendency of an atom to attract a bonding pair of electrons. It is a relevant factor in understanding chemical bonding and shows a significant influence on the shear and Young moduli of the $ABX_3$ perovskites under investigation. The CR impacts bond lengths between atoms, affecting the distances between the A and B cations and the X anion. This is crucial when examining the mechanical stability of materials. The bond length contributes to the lattice parameters, influencing the overall structure's stability and strength. The strength of the covalent bonds is influenced by the size of the participating atoms, especially the X site, which belongs to element X in the cubic 3-dimensional framework of the perovskites. The larger the covalent radius the weaker the bonds. However, the specific impact of these features depends on the site (A, B, and X) of the elements involved. As evident in the SHAP plots (Figs. 4, 5, and 6), the bulk, shear, and Young moduli are highly dependent on the elastic constants.

Other important input features noticed for the prediction of the shear and Young modulus are the PE(X), CR(X), and GSA. PE reflects the relative ability to attract electrons when two different atoms are chemically bonded. This generally results in a stronger bond between the atoms and in turn increases their mechanical properties. CR is defined as half the internuclear distance (bond length) between two bonded atoms. The longer the bond, the softer the material. The closer the bond, the stronger the material. Meanwhile, compounds with lower

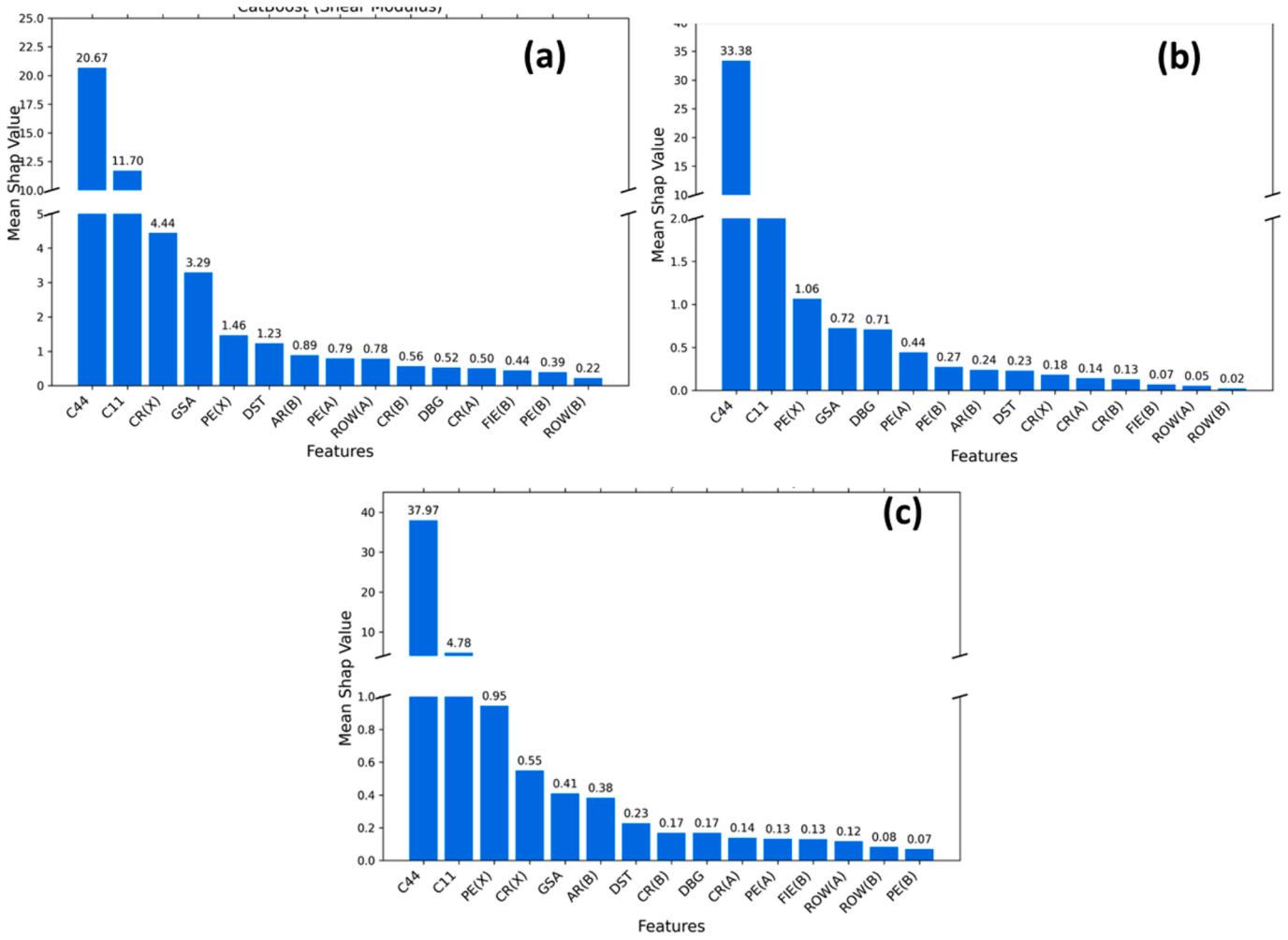



**Fig. 5.** Explainability of the supervised learning model prediction of the shear modulus while revealing the feature importance influencing the model prediction. CatBoost: (a); XGBoost (b); Random Forest (c).

ground state energy per atom are characterised by smaller values of bond length which in turn results into high bond strength thereby, increasing the mechanical properties. From this insight, it can be observed that the halides and oxides play significant roles in the mechanical properties because they form a closer bond with both the A and B site cations.

*Holistic ranking of the input features*

To provide a theoretical explanation for the underlying reasoning behind the superiority of one ensemble learning model over the other, we investigate the SHAP plots to evaluate the significance of the features in predicting the bulk, shear, and Young moduli of the perovskite compounds using the novel holistic ranking method. SHAP features analysis from Figs. 4, 5, and 6 was used as a basis for applying the holistic ranking method [8]. The formula for the holistic ranking is expressed in equation (6). The $F_i^p$ and $R_s$ are recorded in Table 4. The most important feature will have the lowest cumulative score while the least important will have the highest cumulative score.

$$R_s(F_i^p) = \sum_i count(F_i^p) \times p \tag{6}$$

where $R_s$ is the sum of effective ranking per feature for all the learning models, $F_i^p$ is the data input features, and $p$ is the positional ranking of features in each of the explained machine learning.

The global ranking of these features in terms of their position on the SHAP plots was considered for bulk, shear, and Young moduli of the perovskite materials across all explained SL models. The findings regarding the relevance of the features and the overall ranking of features for all the models are presented in Table 4, for bulk shear and Young moduli, respectively. Based on the findings of the SHAP analysis, it was observed that the most significant features are the elastic constants: $C_{11}$ and $C_{44}$. This is because the feature consistently ranks first or second across all the analysed methods.

The insights from the holistic feature ranking of the testing phase for bulk, shear, and Young moduli reveal that the elastic constants $C_{11}$, and $C_{44}$ are the most important features in predicting all the moduli. This is because these features consistently ranked first among all the methods analysed and have the lowest cumulative sum of ranking as indicated in Table 4. The row of element B contributed the least to the prediction tasks for all the mechanical properties.

The determination of elastic constants is of utmost importance in comprehending the mechanical characteristics of materials and their reaction to external forces. These offer valuable insights into the structural, thermo-mechanical, and chemical properties of materials [53,54]. The elastic response of a cubic structure is solely determined by the elastic constants ($C_{11}$, $C_{12}$, and $C_{44}$) which are derived from the stiffness matrix. In this work, the $C_{11}$ and $C_{44}$ parameters were exclusively examined. The $C_{11}$ coefficient characterises the ability of a material to withstand changes in length resulting from the application of axial stress along the crystal plane. The parameter $C_{44}$ represents the ability to resist deformation when subjected to tangential shear stress on the crystal

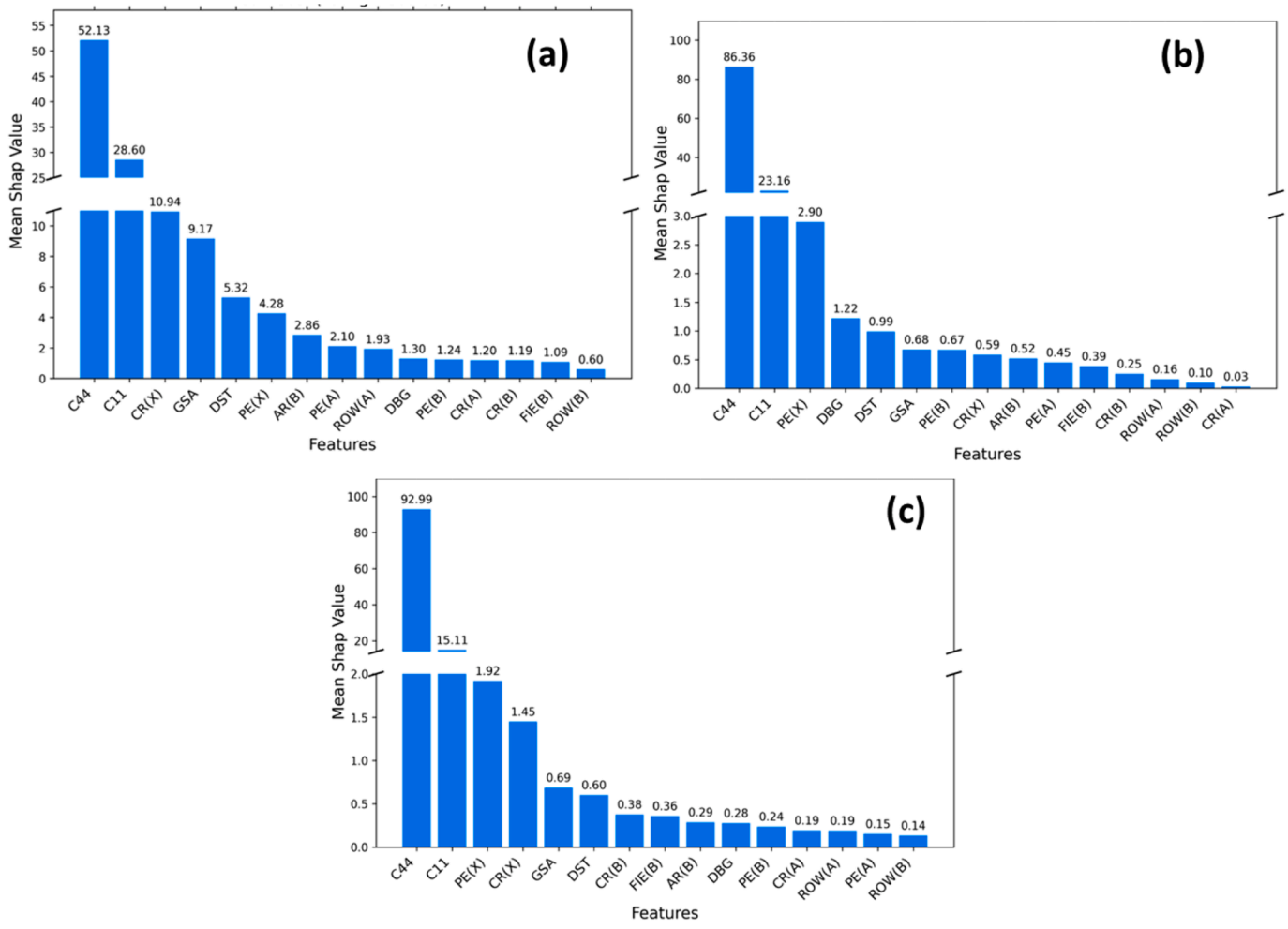



**Fig. 6.** Explainability of the supervised learning model prediction of the Young modulus while revealing the feature importance influencing the model prediction. CatBoost: (a); XGBoost (b); Random Forest (c).

**Table 4**
Holistic ranking features using all the outcomes from the explained machine models after predicting the bulk, shear, and Young moduli.

| | Methods | F1 | F2 | F3 | F4 | F5 | F6 | F7 | F8 | F9 | F10 | F11 | F12 | F13 | F14 | F15 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Bulk Modulus | CatBoost | $13^1$ | $14^2$ | $5^3$ | $10^4$ | $11^5$ | $0^6$ | $2^7$ | $3^8$ | $9^9$ | $4^{10}$ | $6^{11}$ | $1^{12}$ | $7^{13}$ | $8^{14}$ | $12^{15}$ |
| | XGBoost | $13^1$ | $14^2$ | $11^3$ | $3^4$ | $9^5$ | $12^6$ | $1^7$ | $10^8$ | $6^9$ | $0^{10}$ | $4^{11}$ | $2^{12}$ | $7^{13}$ | $8^{14}$ | $5^{15}$ |
| | Random Forest | $13^1$ | $14^2$ | $11^3$ | $3^4$ | $2^5$ | $9^6$ | $5^7$ | $1^8$ | $0^9$ | $4^{10}$ | $6^{11}$ | $12^{12}$ | $10^{13}$ | $7^{14}$ | $8^{15}$ |
| | Best Feature Ranking | 7 | 10 | 6 | 4 | 11 | 7 | 11 | 14 | 15 | 5 | 7 | 3 | 13 | 1 | 2 |
| | | PE (A) | PE (B) | PE (X) | CR (A) | CR (B) | CR (X) | FIE (B) | Row (A) | Row (B) | DBG | GSA | DST | AR (B) | $C_{11}$ | $C_{44}$ |
| Shear Modulus | CatBoost | $14^1$ | $13^2$ | $5^3$ | $10^4$ | $2^5$ | $11^6$ | $12^7$ | $0^8$ | $7^9$ | $4^{10}$ | $9^{11}$ | $3^{12}$ | $6^{13}$ | $1^{14}$ | $8^{15}$ |
| | XGBoost | $14^1$ | $13^2$ | $2^3$ | $10^4$ | $9^5$ | $0^6$ | $1^7$ | $12^8$ | $11^9$ | $5^{10}$ | $3^{11}$ | $4^{12}$ | $6^{13}$ | $7^{14}$ | $8^{15}$ |
| | Random Forest | $14^1$ | $13^2$ | $5^3$ | $10^4$ | $11^5$ | $2^6$ | $12^7$ | $0^8$ | $7^9$ | $9^{10}$ | $1^{11}$ | $3^{12}$ | $4^{13}$ | $6^{14}$ | $8^{15}$ |
| | Best Feature Ranking | 7 | 10 | 4 | 13 | 13 | 5 | 14 | 10 | 15 | 10 | 3 | 6 | 7 | 2 | 1 |
| | | PE (A) | PE (B) | PE (X) | CR (A) | CR (B) | CR (X) | FIE (B) | Row (A) | Row (B) | DBG | GSA | DST | AR (B) | $C_{11}$ | $C_{44}$ |
| Young Modulus | CatBoost | $14^1$ | $13^2$ | $5^3$ | $10^4$ | $11^5$ | $2^6$ | $12^7$ | $0^8$ | $7^9$ | $9^{10}$ | $1^{11}$ | $3^{12}$ | $4^{13}$ | $6^{14}$ | $8^{15}$ |
| | XGBoost | $14^1$ | $13^2$ | $2^3$ | $9^4$ | $11^5$ | $10^6$ | $1^7$ | $5^8$ | $12^9$ | $0^{10}$ | $6^{11}$ | $4^{12}$ | $7^{13}$ | $8^{14}$ | $3^{15}$ |
| | Random Forest | $14^1$ | $13^2$ | $2^3$ | $5^4$ | $10^5$ | $11^6$ | $4^2$ | $6^8$ | $12^9$ | $9^{10}$ | $1^{11}$ | $3^{12}$ | $7^{13}$ | $0^{14}$ | $8^{15}$ |
| | Best Feature Ranking | 10 | 9 | 3 | 14 | 10 | 4 | 12 | 13 | 15 | 7 | 4 | 6 | 8 | 2 | 1 |
| | | PE (A) | PE (B) | PE (X) | CR (A) | CR (B) | CR (X) | FIE (B) | Row (A) | Row (B) | DBG | GSA | DST | AR (B) | $C_{11}$ | $C_{44}$ |

plane in a particular direction [55]. The elastic constants play a crucial role in the theoretical calculation of a material's mechanical properties as they provide information about the tensor stresses in various magnitudes and directions. The top three features are highlighted in green, while the bottom three are in red as shown on the holistic ranking table in Table 4.

## Conclusion

In this study, interpretable ensemble learning models were used for predicting the mechanical properties (bulk, shear, and Young moduli) of $ABX_3$ perovskite compounds. The results show that Random Forest outperforms other ensemble learning techniques in predicting the bulk modulus, while XGBoost performed better comparatively for predicting the shear and Young moduli. Specifically, the Random Forest technique showed a correlation score of $R^2$ of 0.97 during the testing phase for the bulk modulus, while an $R^2$ of 0.97 and $R^2$ of 0.97 was observed as the correlation score when XGBoost was utilized for the prediction of the shear and Young moduli. The ensemble learning models used in this study are promising for practical applications. Furthermore, the SHAP analysis sheds light on the features that have the most impact on the predictions. Among these features, elastic constants $C_{11}$ and $C_{44}$ in the cubic perovskite compounds are found to be the most important in predicting the bulk, shear, and Young moduli. Additionally, it was also deduced that CR(X) shows a significant influence in predicting all the moduli, respectively. Holistic ranking approach for globally ranking all the input features across different ML models was employed and the topmost ranked features were explained from a physics standpoint. These findings are particularly useful for designing materials in a vast materials discovery space and for identifying pressure-resistant perovskites. The strategic employment of ML algorithms can facilitate the intentional exploration of novel $ABX_3$ perovskites possessing favourable mechanical properties. By doing so, the need for extensive laboratory trial and error experiments and ab initio calculations can be considerably decreased. This approach aligns with the fundamental goal of materials informatics which strives to expedite the material design and selection process.

## CRediT authorship contribution statement

**S.B. Akinpelu:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **S.A. Abolade:** Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation. **E. Okafor:** Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **D.O. Obada:** Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **A.M. Ukpong:** Validation, Investigation, Formal analysis. **S. Kumar R.:** Writing – review & editing, Visualization, Validation, Supervision, Project administration, Funding acquisition, Formal analysis, Data curation. **J. Healy:** Writing – review & editing, Writing – original draft, Supervision, Resources, Methodology, Investigation, Formal analysis. **A. Akande:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, 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.

## Acknowledgements

SBA appreciates the Modelling & Computation for Health and Society (MOCHAS) Group of Atlantic Technological University, Ireland for financial support. SAA acknowledges Atlantic Technological University, Sligo, President Bursary Award for financial assistance. DOO acknowledges the funding support of the Irish Research Council granted with Project ID GOIPD/2021/28. The Trinity Centre for High Performance Computing (TCHPC) Kelvin cluster was used for most of these calculations (Project codes: HPC_21_01219, HPC_22_01290 and HPC_22_01254). This cluster was supported by funding from the Higher Education Authority's PRTLI programme. The Irish Centre for High-End Computing (ICHEC) (Project codes: isphy006c, atphy001c and isphy005c) is also acknowledged by the authors for providing computing resources.

## References

[1] Babu R, Giribabu L, Singh SP. Recent Advances in Halide-Based Perovskite Crystals and Their Optoelectronic Applications. Cryst Growth Des 2018;18:2645–64. https://doi.org/10.1021/acs.cgd.7b01767.
[2] Kim H, Han JS, Choi J, Kim SY, Jang HW. Halide Perovskites for Applications beyond Photovoltaics. Small Methods 2018;2:1700310. https://doi.org/10.1002/smtd.201700310.
[3] Ahmadi M, Ziatdinov M, Zhou Y, Lass EA, Kalinin SV. Machine learning for high-throughput experimental exploration of metal halide perovskites. Joule 2021;5:2797–822. https://doi.org/10.1016/j.joule.2021.10.001.
[4] Thoppil GS, Alankar A. Predicting the formation and stability of oxide perovskites by extracting underlying mechanisms using machine learning. Comput Mater Sci 2022;211:111506. https://doi.org/10.1016/j.commatsci.2022.111506.
[5] Chen C, Zuo Y, Ye W, Li X, Deng Z, Ong SP. A Critical Review of Machine Learning of Energy Materials. Adv Energy Mater 2020;10:1903242. https://doi.org/10.1002/aenm.201903242.
[6] Dwivedi N, Balasubramanian K, Sahu R, Manna S, Banik S, Dhand C, et al. Unusual High Hardness and Load-Dependent Mechanical Characteristics of Hydrogenated Carbon-Nitrogen Hybrid Films. ACS Appl Mater Interfaces 2022;14:20220–9. https://doi.org/10.1021/acsami.2c01508.
[7] S. Banik, K. Balasubramanian, S. Manna, S. Derrible, S. Sankaranarayananan, Machine Learning for Elastic Properties of Materials: A predictive benchmarking study in a domain-segmented feature Space, (2023). https://doi.org/10.26434/chemrxiv-2023-07vcr.
[8] Obada DO, Okafor E, Abolade SA, Ukpong AM, Dodoo-Arhin D, Akande A. Explainable machine learning for predicting the band gaps of ABX3 perovskites. Mater Sci Semicond Process 2023;161:107427. https://doi.org/10.1016/j.mssp.2023.107427.
[9] Wang J, Yang X, Zeng Z, Zhang X, Zhao X, Wang Z. New methods for prediction of elastic constants based on density functional theory combined with machine learning. Comput Mater Sci 2017;138:135–48. https://doi.org/10.1016/j.commatsci.2017.06.015.
[10] Li Z, Long Z, Lei S, Tang Y. Machine learning driven rationally design of amorphous alloy with improved elastic models. Mater Des 2022;220:110881. https://doi.org/10.1016/j.matdes.2022.110881.
[11] Lee M, Kim M, Min K. Evaluation of principal features for predicting bulk and shear modulus of inorganic solids with machine learning. Mater Today Commun 2022;33:104208. https://doi.org/10.1016/j.mtcomm.2022.104208.
[12] S. Bishnoi, R. Ravinder, H. Singh Grover, H. Kodamana, N.M. Anoop Krishnan, Scalable Gaussian processes for predicting the optical, physical, thermal, and mechanical properties of inorganic glasses with large datasets, Mater. Adv. 2 (2021) 477–487. https://doi.org/10.1039/D0MA00764A.
[13] Xie T, Grossman JC. Crystal Graph Convolutional Neural Networks for an Accurate and Interpretable Prediction of Material Properties. Phys Rev Lett 2018;120:145301. https://doi.org/10.1103/PhysRevLett.120.145301.
[14] Wang R, Zeng S, Wang X, Ni J. Machine learning for hierarchical prediction of elastic properties in Fe-Cr-Al system. Comput Mater Sci 2019;166:119–23. https://doi.org/10.1016/j.commatsci.2019.04.051.
[15] A. Mansouri Tehrani, A.O. Oliynyk, M. Parry, Z. Rizvi, S. Couper, F. Lin, L. Miyagi, T.D. Sparks, J. Brgoch, Machine Learning Directed Search for Ultraincompressible, Superhard Materials, J. Am. Chem. Soc. 140 (2018) 9844–9853. https://doi.org/10.1021/jacs.8b02717.
[16] Li W, Jacobs R, Morgan D. Predicting the thermodynamic stability of perovskite oxides using machine learning models. Comput Mater Sci 2018;150:454–63. https://doi.org/10.1016/j.commatsci.2018.04.033.
[17] Revi V, Kasodariya S, Talapatra A, Pilania G, Alankar A. Machine learning elastic constants of multi-component alloys. Comput Mater Sci 2021;198:110671. https://doi.org/10.1016/j.commatsci.2021.110671.

[18] Dunn A, Wang Q, Ganose A, Dopp D, Jain A. Benchmarking materials property prediction methods: the Matbench test set and Automatminer reference algorithm. Npj Comput Mater 2020;6:1–10. https://doi.org/10.1038/s41524-020-00406-3.
[19] Körbel S, Marques MAL, Botti S. Stability and electronic properties of new inorganic perovskites from high-throughput ab initio calculations. J Mater Chem C 2016;4:3157–67. https://doi.org/10.1039/C5TC04172D.
[20] Kresse G, Joubert D. From ultrasoft pseudopotentials to the projector augmented-wave method. Phys Rev B 1999;59:1758–75. https://doi.org/10.1103/PhysRevB.59.1758.
[21] Mouhat F, Coudert F-X. Necessary and sufficient elastic stability conditions in various crystal systems. Phys Rev B 2014;90:224104. https://doi.org/10.1103/PhysRevB.90.224104.
[22] Ayyaz A, Murtaza G, Usman A, Umer M, Shah MQ, Ali HS. First principles insight on mechanical stability, optical and thermoelectric response of novel lead-free Rb2ScCuBr 6 and Cs2ScCuBr6 double perovskites. Mater Sci Semicond Process 2024;169:107910. https://doi.org/10.1016/j.mssp.2023.107910.
[23] Purba M, Ermatita E, Abdiansah A, Noprisson H, Ayumi V, Setiawan H, et al. Effect of Random Splitting and Cross Validation for Indonesian Opinion Mining using Machine Learning Approach. Int J Adv Comput Sci Appl 2022;13. https://doi.org/10.14569/IJACSA.2022.0130917.
[24] May RJ, Maier HR, Dandy GC. Data splitting for artificial neural networks using SOM-based stratified sampling. Neural Netw 2010;23:283–94. https://doi.org/10.1016/j.neunet.2009.11.009.
[25] Ha YJ, Yoo M, Lee G, Jung S, Choi SW, Kim J, et al. Spatio-Temporal Split Learning for Privacy-Preserving Medical Platforms: Case Studies With COVID-19 CT, X-Ray, and Cholesterol Data. IEEE Access 2021;9:121046–59. https://doi.org/10.1109/ACCESS.2021.3108455.
[26] D. Berrar, Cross-validation., (2019). http://berrar.com/resources/Berrar_EBCB_2nd_edition_Cross-validation_preprint.pdf.
[27] Cheng J, Dekkers JCM, Fernando RL. Cross-validation of best linear unbiased predictions of breeding values using an efficient leave-one-out strategy. J Anim Breed Genet 2021;138:519–27. https://doi.org/10.1111/jbg.12545.
[28] Pearson K. Contributions to the Mathematical Theory of Evolution. Philos Trans r Soc Lond A 1894;185:71–110.
[29] F. Sustainability, Analyzing meteorological parameters using Pearson correlation coefficient and implementing machine learning models for solar energy prediction in Kuching, Sarawak | Future Sustainability, (2024). https://fupubco.com/fusus/article/view/154 (accessed March 19, 2024).
[30] Nwafor O, Okafor E, Aboushady AA, Nwafor C, Zhou C. Explainable Artificial Intelligence for Prediction of Non-Technical Losses in Electricity Distribution Networks. IEEE Access 2023;11:73104–15. https://doi.org/10.1109/ACCESS.2023.3295688.
[31] R. Saidi, W. Bouaguel, N. Essoussi, Hybrid Feature Selection Method Based on the Genetic Algorithm and Pearson Correlation Coefficient, in: A.E. Hassanien (Ed.), Mach. Learn. Paradig. Theory Appl., Springer International Publishing, Cham, 2019: pp. 3–24. https://doi.org/10.1007/978-3-030-02357-7_1.
[32] Xu P, Ji X, Li M, Lu W. Small data machine learning in materials science. Npj Comput Mater 2023;9:1–15. https://doi.org/10.1038/s41524-023-01000-z.
[33] Hancock JT, Khoshgoftaar TM. CatBoost for big data: an interdisciplinary review. J Big Data 2020;7:94. https://doi.org/10.1186/s40537-020-00369-8.
[34] Ahn JM, Kim J, Kim K. Ensemble Machine Learning of Gradient Boosting (XGBoost, LightGBM, CatBoost) and Attention-Based CNN-LSTM for Harmful Algal Blooms Forecasting. Toxins 2023;15:608. https://doi.org/10.3390/toxins15100608.
[35] A. Anghel, N. Papandreou, T. Parnell, A. De Palma, H. Pozidis, Benchmarking and Optimization of Gradient Boosting Decision Tree Algorithms, arXiv.Org (2018). https://arxiv.org/abs/1809.04559v3.
[36] A.V. Dorogush, V. Ershov, A. Gulin, CatBoost: gradient boosting with categorical features support, (2018). http://arxiv.org/abs/1810.11363 (accessed August 29, 2023).
[37] A. Ustimenko, A. Beliakov, L. Prokhorenkova, Gradient Boosting Performs Gaussian Process Inference, (2023). http://arxiv.org/abs/2206.05608 (accessed January 15, 2024).
[38] Prokhorenkova L, Gusev G, Vorobev A, Dorogush AV, Gulin A. CatBoost: unbiased boosting with categorical features. Adv. Neural Inf. Process. Syst., Curran Associates Inc; 2018. accessed August 29, 2023.
[39] Tamayo D, Silburt A, Valencia D, Menou K, Ali-Dib M, Petrovich C, et al. A machine learns to predict the stability of tightly packed planetary systems. Astrophys J Lett 2016;832:L22. https://doi.org/10.3847/2041-8205/832/2/L22.
[40] Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. In: Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. New York, NY, USA: Data Min., Association for Computing Machinery; 2016. p. 785–94. https://doi.org/10.1145/2939672.2939785.
[41] Breiman L. Random Forests. Mach Learn 2001;45:5–32. https://doi.org/10.1023/A:1010933404324.
[42] Fawagreh K, Gaber MM, Elyan E. Random forests: from early developments to recent advancements. Syst Sci Control Eng 2014;2:602–9. https://doi.org/10.1080/21642583.2014.956265.
[43] Lundberg SM, Lee S-I. A Unified Approach to Interpreting Model Predictions. Adv. Neural Inf. Process. Syst., Curran Associates Inc; 2017. accessed August 29, 2023.
[44] S.M. Lundberg, G.G. Erion, S.-I. Lee, Consistent Individualized Feature Attribution for Tree Ensembles, arXiv.Org (2018). https://arxiv.org/abs/1802.03888v3 (accessed August 29, 2023).
[45] Gulzar F, Siddique S, Gillani SSA, Abbas N, Zeba I. Doping induced modulation in structural, electronic, optical, elastic and mechanical properties of RbPbF3: Insights from DFT computation. Mater Sci Eng B 2024;305:117435. https://doi.org/10.1016/j.mseb.2024.117435.
[46] Wang YC, Lakes RS. Composites with Inclusions of Negative Bulk Modulus: Extreme Damping and Negative Poisson's Ratio. J Compos Mater 2005;39:1645–57. https://doi.org/10.1177/0021998305051112.
[47] Kindler B, Finsterbusch D, Graf R, Ritter F, Assmus W, Lüthi B. Mixed-valence transition in ${\mathrm{YbInCu}}_{4}$. Phys Rev B 1994;50:704–7. https://doi.org/10.1103/PhysRevB.50.704.
[48] Chicco D, Warrens MJ, Jurman G. The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. PeerJ Comput Sci 2021;7:e623.
[49] Li C, Lu X, Ding W, Feng L, Gao Y, Guo Z. Formability of ABX3 (X = F, Cl, Br, I) halide perovskites. Acta Crystallogr B 2008;64:702–7. https://doi.org/10.1107/S0108768108032734.
[50] Pilania G, Balachandran PV, Kim C, Lookman T. Finding New Perovskite Halides via Machine Learning. Front Mater 2016;3. https://doi.org/10.3389/fmats.2016.00019.
[51] Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, et al. From local explanations to global understanding with explainable AI for trees. Nat Mach Intell 2020;2:56–67. https://doi.org/10.1038/s42256-019-0138-9.
[52] Gupta R, Varshney P, Pravesh M, Lal D, Kumar K, Singh AV. Mechanical stability parameters of chalcogenides and pnictides based optoelectronic materials. Chalcogenide Lett 2023;20:101–12. https://doi.org/10.15251/CL.2023.202.101.
[53] Khanzadeh M, Alahyarizadeh G. A DFT study on pressure dependency of TiC and ZrC properties: Interconnecting elastic constants, thermodynamic, and mechanical properties. Ceram Int 2021;47:9990–10005. https://doi.org/10.1016/j.ceramint.2020.12.145.
[54] Duan YH, Sun Y, Peng MJ, Zhou SG. Anisotropic elastic properties of the Ca–Pb compounds. J Alloys Compd 2014;595:14–21. https://doi.org/10.1016/j.jallcom.2014.01.108.
[55] Rahman MA, Rahaman MZ, Rahman MA. The structural, elastic, electronic and optical properties of MgCu under pressure: A first-principles study. Int J Mod Phys B 2016;30:1650199. https://doi.org/10.1142/S021797921650199X.