Journal of Physics and Chemistry of Solids 171 (2022) 110973

Contents lists available at ScienceDirect

Journal of Physics and Chemistry of Solids

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

Bandgap analysis of transition-metal dichalcogenide and oxide via machine learning approach

Upendra Kumar [a], Km Arti Mishra [b], Ajay Kumar Kushwaha [c], Sung Beom Cho [a,d,*]

[a] Virtual Engineering Center, Technology Convergence Division, Korea Institute of Ceramic Engineering and Technology (KICET), Jinju 52851, South Korea
[b] Department of Electrical Engineering, Faculty of Engineering and Technology, Rama University, Kanpur, UP 209217, India
[c] Department of Metallurgy Engineering and Materials Science, Indian Institute of Technology Indore, Khandwa Road, Simrol, Indore 453552, India
[d] Department of Materials Science and Engineering, Ajou University, Suwon 16499, South Korea

ARTICLE INFO

Keywords:
Transition-metal dichalcogenides and oxides
Machine learning
Compressive sensing
Regression
Classification

ABSTRACT

Predicting bandgap is a crucial topic in materials informatics, however, it is still difficult when the available dataset is limited and unbalanced. Here, we applied a machine learning approach to construct a prediction model for transition metal dichalcogenides and oxides. Using an oversampling technique and atomistic feature engineering, we successfully constructed the machine learning model and analyzed the correlation with other physical properties. Furthermore, we also utilized the model to obtain a compressive sensing model based on physical quantities for analytic interpretation and quick prediction.

1. Introduction

Graphene is immensely popular due to its many unique features, but its absence of an electronic bandgap has motivated the pursuit of 2D materials with semiconducting properties [1,2]. Transition metal dichalcogenides (TMDCs) becomes an alternative to graphene. The compositional formula of TMDCs is MX_2, where M is a transition metal atom (like Mo or W) and X is a chalcogen atom (like S, Se, or Te) [3], consist both semiconductors and metals [4] properties. When the thickness of MoS_2, $MoSe_2$, WS_2, and WSe_2 is reduced down to a single layer, they are demonstrated to convert from indirect to direct bandgap materials [5]. There is a presence of many unique properties in TMDCs such as direct bandgap, strong spin–orbit coupling [6] and favorable electronic and mechanical properties [7]. Due to such interesting properties, TMDCs become crucial in application of spintronics [8], optoelectronics [4], energy harvesting [9], flexible electronics [10], DNA sequencing and personalized medicine [11]. There is demonstration of the first transistor [12] and strong photoluminescence in MoS_2 monolayers makes 2D TMDCs a lucrative material [13]. Moreover, the absence of inversion symmetry in 2H monolayer structures causes a spin–orbit driven splitting of the valence band, allowing valley-selective charge carrier excitation [14]. This feature creates the hope of tuning the electronic properties with the help of strain, dielectric screening, electrostatic gating and nanostructuring.

Transition metal oxides (TMOs) are likely one of the most fascinating solid classes, with a wide range of structures and characteristics [15]. Metal–oxygen bonding may range from almost ionic to strongly covalent (metallic) in character. Therefore, TMOs are remarkable for their extraordinary spectrum of electrical and magnetic characteristics. At one end of the spectrum, we find oxides with metallic qualities (e.g. RuO_2, ReO_3, $LaNiO_3$) and oxides with extremely insulating characteristics (e.g. $BaTiO_3$). The p–n heterojunction built on transition-metal oxides shows potential in a variety of domains, including H_2 evolution, CO_2 reduction, overall water splitting, photoreforming, and photodegradation of hazardous pollutants [16].

The bandgap is an inherent feature of a material, it is estimated by experiments such as UV spectroscopy [17] or differential and cyclic voltage measurements [18]. These experiments need huge, costly equipment and are time-consuming to carry out. Some materials are challenging, especially because of the use of specialized settings to conduct these experiments. Faster computers have made it possible to use density-functional-theory-based (DFT-based) [19] methods to quickly calculate the bandgap. In spite of this, the fundamental bandgaps computed using the local-density or generalized-gradient approximation (LDA or GGA) are significantly underestimated. The underestimate of the bandgap could be rectified by using GW technique based on the many-body perturbation theory [20]. However, calculating the correct bandgap of material becomes a time-consuming (GW technique) and costly approach (experiment) [20,21]. The precise calculation of bandgaps cannot be accomplished by the use of an equation or a simple formula. Therefore, material physicists are still grappling with the challenge of precisely calculating bandgap. Data science and machine

* Corresponding author at: Virtual Engineering Center, Technology Convergence Division, Korea Institute of Ceramic Engineering and Technology (KICET), Jinju 52851, South Korea.
E-mail address: csb@ajou.ac.kr (S.B. Cho).

https://doi.org/10.1016/j.jpcs.2022.110973
Received 23 May 2022; Received in revised form 16 August 2022; Accepted 18 August 2022
Available online 31 August 2022
0022-3697/© 2022 Elsevier Ltd. All rights reserved.

learning (ML) techniques have been around for a long time, but their use in subjects like material science has only just begun. Over the last several years, ML has emerged as an absolutely superb approach for predicting the structures [22] and characteristics of numerous types of materials [23]. Using ML approaches, it is possible to forecast exactly the characteristics of materials such as cohesive energy, lattice thermal conductivity [24], bandgap [25] and entropy as well as free energy [26] and heat capacity [27].

The bandgap of double perovskites can be accurately predicted using a KRR model up to RMSE of 0.36 eV [28]. Additionally, a technique based on neural networks was employed to estimate the bandgap for chalcopyrite, group III–V and II–VI binary semiconductors [25] and group I–III–VI_2 and II–IV–V_2 ternary semiconductors [29]. It has also been possible to estimate the bandgap of binary and ternary semiconductors using artificial neural networks [30] or support vector regression [31], however both methods rely on a very limited training set of compounds (around 30). A co-kriging regression model consisting of a multi Gaussian process was used to estimate bandgap for a large number of elpasolite compounds [32]. The AFLOW project's predicted electronic structure has been utilized as a training set for machine learning, a technique that extends beyond particular crystal systems [33]. This method combines universal fragment descriptors with a gradient boosting decision tree to offer reliable estimates of the bandgap for inorganic materials as well as other thermochemical parameters such as heat capacity, Debye temperature, and elastic moduli. In the same way, another method that uses crystal graph convolutional neural networks based on how atoms connect in a crystal can reach the level of accuracy of DFT calculations after being trained with DFT bandgap [34]. All of these approaches may be used to rapidly estimate features like bandgap. Therefore, we are trying to apply a similar methodology for finding the bandgap of TMDCs and TMOs.

In addition, we evaluated our prediction accuracy with recently published research. Yu Zhang et al. have predicted the bandgap of two-dimensional materials with root mean square error (RMSE) of 0.24 eV and 0.27 eV by using gradient boosted decision trees and random forest model [35]. The accurate bandgap prediction of perovskites has been performed with RMSE of 0.055 eV [36], which agrees with our compressive sensing result (SISSO). There is an accurate bandgap prediction of solids assisted by machine learning with RMSE of 0.28 eV [37]. The bandgap of quaternary thermoelectric materials has been predicted within RMSE range of 0.25 eV [38]. Support vector machine model has been utilized for prediction of the polymer bandgap with RMSE of 0.485 eV [39]. A decision-tree-based ML model is used for the prediction of G_0W_0 corrections in the PBE band structure of a 2D semiconductor [40]. The computational 2D materials database has been utilized for prediction of HSE06 band gap, the PBE heat of formation(ΔH), the exciton binding energy, static polarizability and Voigt modulus via ML approach [41]. Thus, our bandgap prediction accuracy in TMDCs and TMOs materials is often within the error widely described in the literature, indicating that we have built a functioning ML model with the optimum set of hyperparameters.

Here, we are studying the bandgap behavior of TMDCs and TMOs by using a supervised ML approach. There are two steps have been performed in the supervised machine-learning process. In the first step, classification and regression have been carried out to determine if a material is a semimetal (zero bandgap) or semiconductor (non-zero bandgap). In the second step, it is shown how to derive the formula of bandgaps that are categorized as non-metals (semiconductor) with the help of compressive sensing i.e. SISSO [42]. Various regression and classification algorithms have been examined, including the linear (ordinary least square, partial least square, etc.) and non-linear models (gradient boosting, support vector machine, random forest regression, etc.). Details of the ML model have been given in the supplementary information. The random forest model beat all others ML model in terms of performance. To prevent bias in the final predictions, the dataset must be balanced. The synthetic minority oversampling technique (classification) [43] and synthetic minority oversampling technique for regression with Gaussian noise (regression) [44] have been used for removing behavior of the imbalance dataset. Finally, it can be said, it is tried to find out the best ML model for the prediction of bandgap with a limited number of the dataset of TMDCs and TMOs.

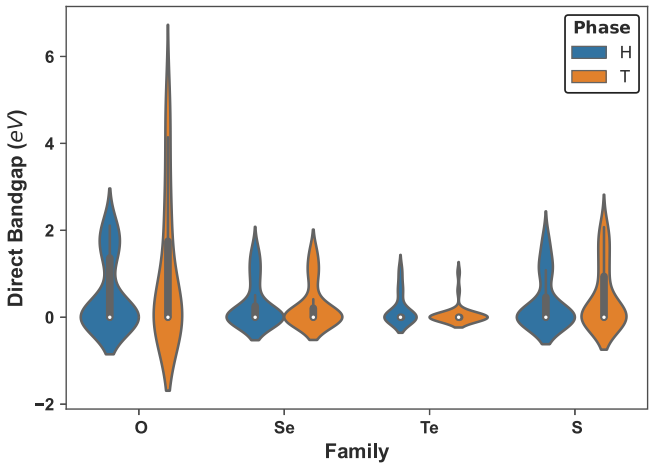

Fig. 1. The nature of the bandgap (violin plot) in various TMDCs and TMOs families with respect to phase. Here, H stands for a hexagonal and T stands for a trigonal phase of the unit cell. The shape of the violin shows behavior kernel density estimate (KDE) plot, a histogram-like approach for showing a dataset's distribution of observations.

2. Computational details and data information

The dataset used for the ML model is taken from the article by Rasmussen et al. [45]. In this dataset [45], the bandgap of monolayer TMDs and TMOs has been calculated, which consists of 2H and 1T structures based on 27 different materials. The projector augmented wave approach, as implemented in the GPAW code, was used for all computations [46]. A structural relaxation employing the Perdew–Burke–Ernzerhof (PBE) exchange–correlation (Xc) function has been used to calculate the lattice constants of the 216 monolayer TMDs and TMOs. The PBE pseudopotential with a 750 eV energy cutoff has been applied. The density of states computations is performed using the 18 × 18 × 1 Monkhorst–Pack k-point mesh. There is a gap of 20 Å between periodically repeated layers. For both cases of unit cell phase, i.e., 2H and 1T, the lattice constant of the minimal unit cell and the vertical positions of the oxygen or chalcogen atoms have been relaxed until all forces are equal to or less than 0.01 eV/Å. The dataset has 216 entries. We performed 20-fold cross-validation in our ML models. The data was split into two parts: 20% for testing and 80% for training the ML model. The variation of bandgap data in various materials is depicted in Fig. 1. The details of distribution are given in the Table 1. In the predictions of the direct bandgap of TMDCs and TMOs, we include various properties of the compositional elements as features. The details of the elemental and structural features are shown in the Table 2. There are two types of elements in these TMD compounds. The first is the transition metal (TM) element and the second is from the chalcogen (Chalco.) family. The features which are correlated with more than 0.7 values have been removed from the descriptor, as shown in Fig. 2. We also found group number of Chalco. element is a constant feature, so we removed it. The final features in our descriptors are **(1)** lattice constant **(2)** magnetic moment **(3)** formation energy **(4)** phase and TM element **(5)** atomic number **(6)** group number **(7)** atomic radius **(8)** Van der Waals radius **(9)** fusion heat **(10)** first ionization energy.

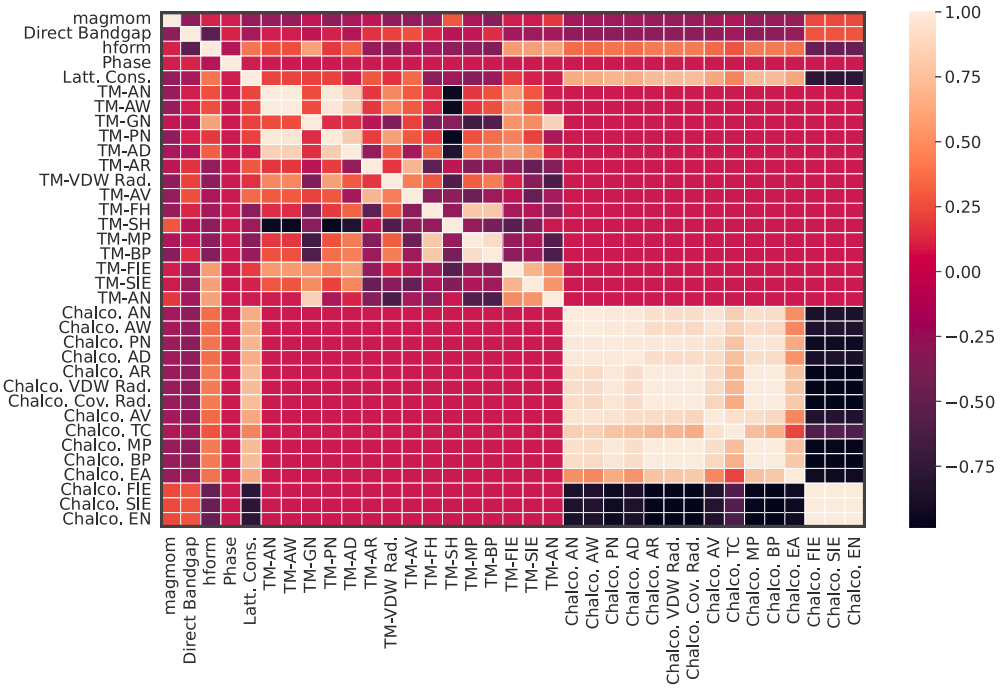

Fig. 2. Pearson correlation of direct bandgap with various elemental features.

Table 1
Nature of bandgap data used in ML model, std is abbreviation of standard deviation.

Family	Count	Phase	Min	Max	Mean	std
O	27	H	0	2.11	0.51	0.82
	27	T	0	5.04	1.01	1.63
Se	27	H	0	1.55	0.28	0.51
	27	T	0	1.49	0.27	0.50
Te	27	H	0	1.08	0.17	0.34
	27	T	0	1.03	0.06	0.22
S	27	H	0	1.81	0.33	0.59
	27	T	0	2.07	0.45	0.72

Table 2
Features used in ML models.

Elemental	Structural
Atomic Number (AN)	Magnetic Moment (magmom)
Atomic Weight (AW)	Formation Energy (hform)
Group Number (GN)	Phase
Period Number (PN)	Lattice Constant of the supercell (Latt. Cons.)
Atomic Density (AD)	
Atomic Radius (AR)	
Van der Waals radius (VDW Rad.)	
Atomic Volume (AV)	
Fusion Heat (FH)	
Specific Heat (SH)	
Melting Point (MP)	
Boiling Point (BP)	
First Ionization Energy (FIE)	
Second Ionization Energy (SIE)	
Allen Electronegativity (AEN)	
Covalent Radius (CR)	
Electron Affinity (EA)	

3. Result and discussion

3.1. Classification analysis

The goal of this work is to classify zero (class 0) and non-zero bandgaps (class 1) in the dataset [45]. When imbalanced classification is present in the ML model, i.e. one class has a very higher number of samples and the other class has a very lower number of samples (minority class) compared to earlier, a problem of decision-making boundary in the ML model classifier exists. Such a type of problem can be resolved by performing oversampling of the minority class, i.e., simply duplicating examples from the minority class but due to the absence of any new information, ML model predictability becomes very poor. To increase ML model predictability, there will be a synthesization of new samples from the minority class. *The Synthetic Minority Oversampling Technique (SMOTE)*, developed by Nitesh Chawla et al. [43], is one of the most effective synthesizing technique. Initially, the dataset [45] used in ML model has 55 bandgap with class 1 (non-zero bandgap) and 161 bandgap with class 0 (zero bandgap), shown in Fig. 3(a). After applying SMOTE technique, both classes have 161 numbers [Fig. 3(b)] and samples trained ML model for performing best predictability. The comparison of ML model predictability without SMOTE and with SMOTE is shown in Tables 3 and 4, respectively. It can be seen, that there is a lot of modification that comes to ML model test accuracy and support score after applying the SMOTE technique.

After applying various ML models, the classification of zero and non-zero bandgap has been performed. All the details of the ML models used for classification is given in supplementary information section(I). It is found that random forest gives the best performance among all ML models. So we have considered it and used it in further classification analysis. The feature importance found by the random forest classifier is shown in Fig. 4.

It is found that magnetic moment and heat of formation energy are very important features, which can be used for classification. We have used these features as a classifier of bandgap, shown in Fig. 5. It has already been shown that the initial magnetic moments have a major impact on the band structure and pave the way for tuning the bandgap dynamically [47]. The insulator–metal transition of Nd_2CoFeO_6 has been performed by the local increment in the magnetic moment of Co^{3+} ions and shown that there is a variation of the bandgap, which is a function of the average local Co magnetic moment [48]. The term "heat of formation" refers to the amount of heat absorbed or emitted during the production of one mole of a compound from its component

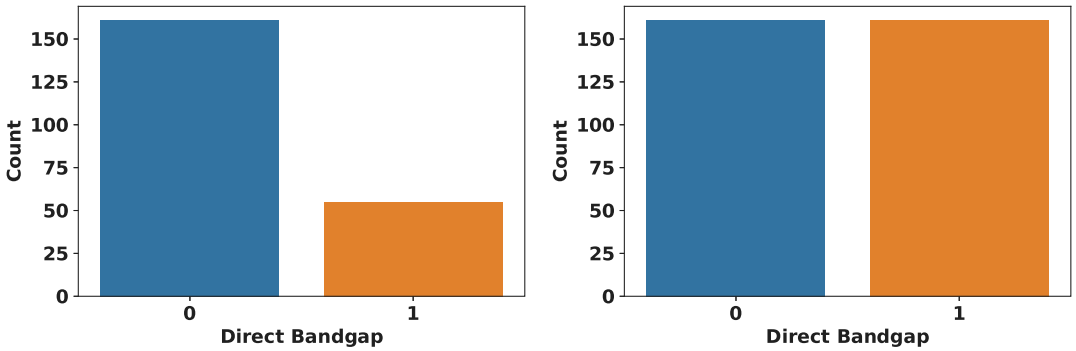

Fig. 3. (a) Class imbalance of direct bandgap. **(b)** Balancing direct bandgap class after SMOTE. The zero direct bandgap is showing semimetal behavior, on the other hand, non-zero bandgap shows semiconductor behavior of TMDCs and TMOs.

Table 3
Classification without SMOTE.

Classification method	Test accuracy score	Binary classifier	Precision	Recall	f_1 Score	Support
Decision tree	0.88	0	0.97	0.89	0.93	36
		1	0.64	0.88	0.74	8
GB	0.91	0	0.97	0.91	0.94	35
		1	0.73	0.89	0.80	9
K-NN	0.82	0	0.94	0.84	0.89	37
		1	0.45	0.71	0.56	7
Gaussian NB	0.65	0	0.58	0.95	0.72	20
		1	0.91	0.42	0.57	24
SVM	0.75	0	1.00	0.75	0.86	44
		1	0.00	0.00	0.00	0
Random forest	0.89	0	0.94	0.91	0.93	34
		1	0.73	0.80	0.76	10

A comparison of various ML classification models. The abbreviation GB for gradient boosting, K-NN for k-nearest neighbors, NB for naive Bayes and SVM for support vector machine.

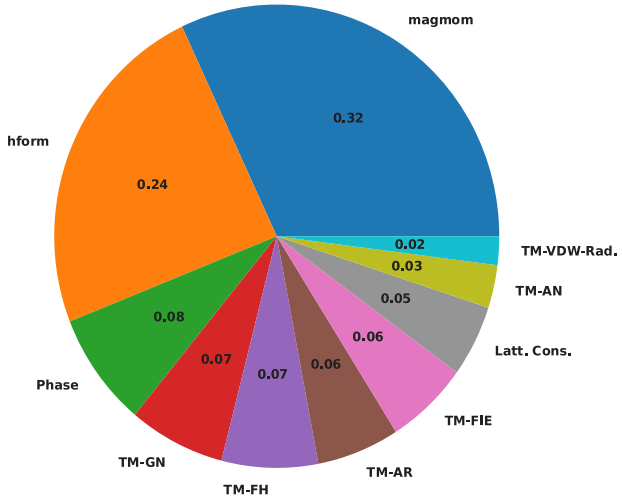

Fig. 4. Feature importance of the RFR Classifier. TM is an abbreviation of transition metal.

parts. So it is directly proportional to the temperature. The bandgap of the material is also directly proportional to the temperature [49]. Lattice constant and atomic volume is complementary of each other. It is already found that the lattice constant is inversely proportional to the bandgap of material [50].

The above-mentioned feature explanation can be seen in our classification model. All systems with magnetism ($\geq 0.01\mu_B$) are semimetal. A small variation in the magnetic moment changes the bandgap behavior and it becomes semimetallic, as depicted in Fig. 5(a). The heat of formation energy becomes a proxy for temperature for the classification of the bandgap. *So from our ML model, it can be seen, that when the heat of formation energy becomes more positive (≥ 0.07 eV), TMD behaves like semimetal*, shown in Fig. 5(b). The lattice constant also plays a crucial role in the classification of the bandgap. In the case of a higher lattice constant (≥ 4.02 Å) only zero bandgap exists, shown in Fig. 5(c). *From Fig. 4, it can also be seen transition metal properties have more importance than chalcogen for deciding the class of bandgap.* Therefore, it can be said, the transition metal decides the nature of the bandgap in TMDCs and TMOs.

Receiver operating characteristic (ROC): The area under the ROC curve (AUC) is a two-dimensional measurement of the complete two-dimensional area beneath the full ROC curve from (0,0) to (1,1). The AUC is a measure of overall performance for all possible classifications. The x-axis is 1-specificity (= false-positive fraction), the y-axis is sensitivity (= true positive fraction). By using the ROC curve, the performance of our ML classification model can be justified. There are two parameters, i.e., true positive rate and false positive rate, plotted in such a curve [51]. True-positive rate is defined as the sensitivity, recall, or probability of detection of an ML model. On the other hand, the false-positive rate is defined as the probability of a false alarm. The area under the curve (AUC) describes the probability, i.e., a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one. A perfect model has an AUC equal to 1. After using all the ML models, it is found that the random forest classifier has the best performance in comparison to other models in ROC analysis, as shown in Fig. 5**(d)**. So, random forest is the best ML model among all for classifying bandgap data.

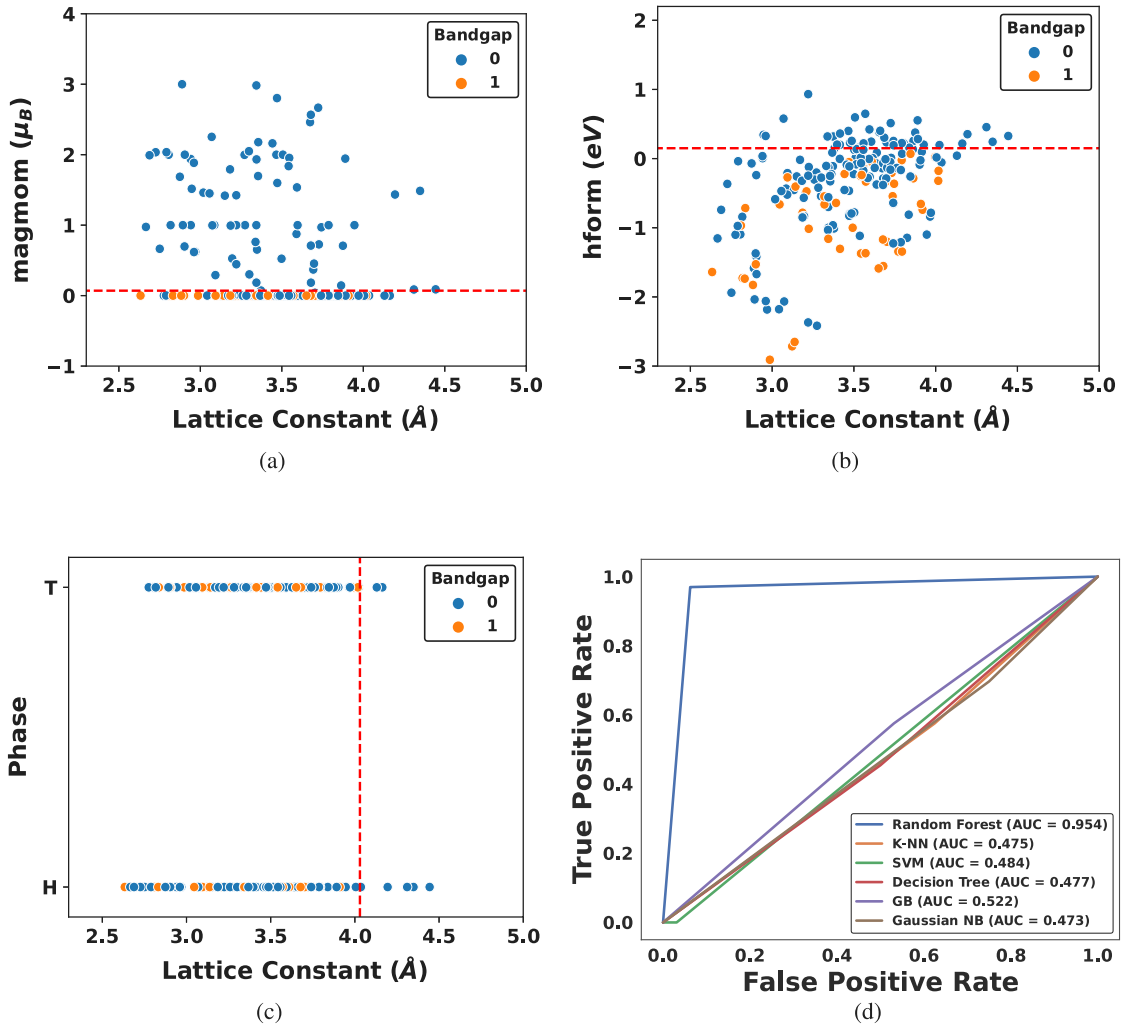

Fig. 5. The plot between **(a)** magnetic moment (magmom) and lattice constant, **(b)** heat of formation energy and lattice constant **(c)** phase and lattice constant. We have classified 0 as a zero bandgap (semimetal), i.e., the conduction band and valence band touch each other, and 1 as a non-zero band gap. **(d)** The curve of the receiver operating characteristic of various ML classification models.

Table 4
Classification after applying SMOTE.

Classification method	Test accuracy score	Binary classifier	Precision	Recall	f_1 Score	Support
Decision tree	0.91	0	0.94	0.88	0.91	34
		1	0.88	0.94	0.91	31
GB	0.93	0	0.91	0.97	0.94	30
		1	0.97	0.91	0.94	35
K-NN	0.90	0	0.81	1.00	0.90	26
		1	1.00	0.85	0.92	39
Gaussian NB	0.77	0	0.55	1.00	0.71	18
		1	1.00	0.68	0.81	47
SVM	0.48	0	0.97	0.48	0.65	64
		1	0.00	0.00	0.00	1
Random forest	0.95	0	0.94	0.97	0.95	31
		1	0.97	0.94	0.96	34

A comparison of various ML classification models. The abbreviation GB for gradient boosting, K-NN for k-nearest neighbors, NB for naive Bayes, and SVM for support vector machine.

3.2. Regression analysis

Imbalanced domains are a significant issue that has mostly been investigated in the context of classification tasks [52]. In various predicting situations, such as regression tasks, data streams, or time-series forecasting, unbalanced domains do arise [53]. Nonetheless, there is a scarcity of research into novel solutions that are appropriate for these jobs because of **(i)** the user's non-uniform preferences throughout the scope of the target variable **(ii)** the limited depiction of the most important examples to the user in the accessible data. So the learner's predictive performance has been hindered by these two factors. There is a new method for solving the problem of imbalanced regression, called *Synthetic Minority Over-Sampling Technique for Regression with Gaussian Noise* (SMOGN) [44]. In SMOGN, there is a utilization of the

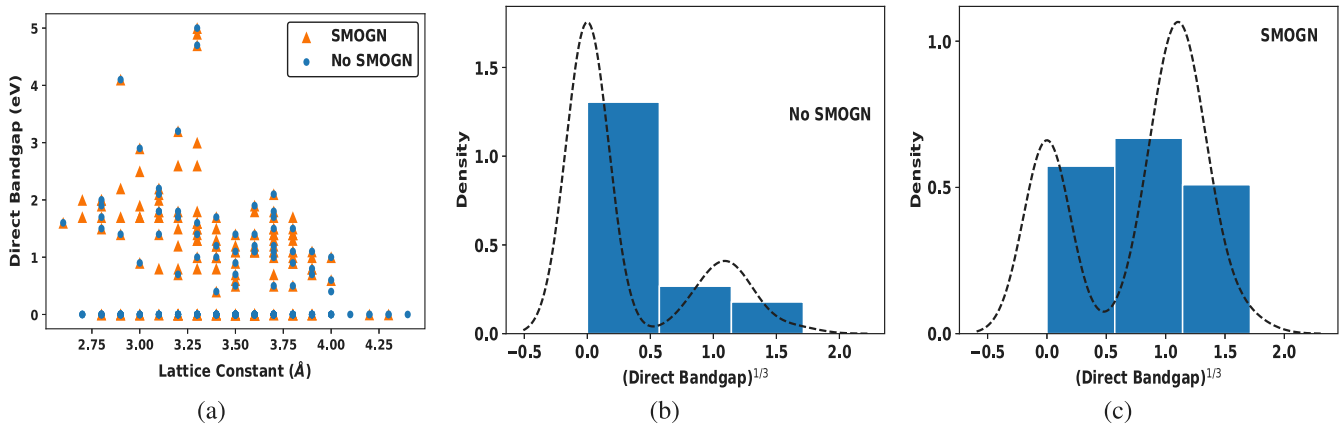

Fig. 6. The plot **(a)** between lattice constant and direct bandgap of TMD, showing new data synthesize with help of SMOGN. Kernel density estimation plot of **(b)** direct bandgap without SMOGN (`skewness = 1.29, kurtosis = -0.05`) **(c)** direct bandgap with SMOGN (`skewness = -0.50, kurtosis = -1.39`).

Synthetic Minority Over-Sampling Technique for Regression (SMOTER) with traditional interpolation and the introduction of Gaussian Noise (SMOTER-GN). If K-Nearest Neighbour (KNN) distance between samples is small, SMOTER is applied. On the other hand, if KNN distances between samples are large SMOTER-GN is applied. SMOGN also becomes an alternative to log transform for a skewed response variable. After applying the SMOGN technique, the data shape is reduced and the new dataset has only 165 direct bandgaps. After applying SMOGN technique, there is a presence of 54 zeros, which was 161 before balancing, depicted in Fig. 6**(a)**. After applying SMOGN, there is a drastic change in skewness and kurtosis of original dataset, shown in Fig. 6**(b)** and **(c)**. For reducing kurtosis and skewness in the bandgap data, a cube root transformation has been performed. The detailed behavior can be seen in supplementary section (II).

There are various linear [e.g. Ordinary Least Square (OLS), Partial Least Square (PLS), Ridge and least absolute shrinkage and selection operator (Lasso)] and non-linear [e.g. Gradient Boosting Regression (GBR), Kernel Ridge Regression (KRR), Random Forest Regression (RFR), Support Vector Machine (SVM), XgBoost (XGBoost) [54]] ML models. In this work, these ML models have been applied for regression of the direct bandgap dataset. The `Scikit-Learn` library [55] has been used as a tool for the application of these ML models. All the details of hyperparameters used in these models are given in the supplementary information section(II). The feature vectors have been normalized and applied in ML models. The Monte Carlo cross-validation method is applied to the evolution of the ML model predictability [56]. The prediction power of various ML models is assessed by using the root mean square error (RMSE), R-Squared (R^2) value, and mean absolute error (MAE) of the test dataset, shown in Fig. 7. The PLS, Ridge, Lasso, GBR, and SVM models have poor performance, i.e R^2 test has very small value. In the case of the XGBoost model, the training RMSE is negligible and the testing RMSE is high. It shows a kind of overfitting in the model, so we do not consider it in bandgap regression. Only RFR is giving good accuracy for RMSE, R^2 and MAE value. So it is used for further analysis of the dataset. The detailed parameter of RMSE, standard deviation, R^2 and MAE are given in the supplementary information section(II).

The individual state depicting accuracy of RFR, shown in Fig. 7**(d)**. The feature importance of RFR is described in Fig. 8**(a)**. From Fig. 8**(a)**, it can be said *magmon* and *hfrom* are most important features for deciding bandgap. The physical significance of these quantities has already been described in the classification result and discussion Section 3.1. The lattice constant is also among the most important feature, as shown in Fig. 8**(a)**. There is an empirical relationship between energy gap E_0 (bandgap) and lattice constant a_0 in the cubic semiconductors [50] i.e. $E_0 \propto 1/a_0^2$. Therefore, it can be said bandgap decreases with increasing lattice constant, shown in Fig. 6**(a)**. It is showing a kind of agreement between the ML model and experimental prediction. In the case of regression, we are working with a very small dataset i.e. 165 direct bandgap values. The model performance increases with increasing training data size. But if we give training data size more than 80% (testing less than 20%) over-fitting problem comes into the picture. The detail of learning curve is shown in Fig. 8**(b)**.

3.3. Compressive sensing

To predict the direct bandgap, we also utilized equation-based machine learning i.e. compressive sensing, which is giving a relation by using physical constraints, known as SISSO [42]. In SISSO, there is some analytical formula found by utilizing model prediction i.e. the prediction has been expressed in terms of physical quantities with help of some algebraic operations like addition and exponentiation. The SISSO method utilizes a sure-independence screening (SIS) method and the sparse-solution algorithm with the application of sparsifying operators (SO) in tandem. When a new feature space has been created, the SIS method takes a subspace of features, which has the largest linear correlation to the target property (direct bandgap). After that, the SO step assesses all potential feature combinations from the SIS subspace, producing the best least-squares solution and residual. Because the feature space is so large, each SO step's combinatorial optimization relies on L_0 regularization, which penalizes the number of nonzero coefficients. Each feature is utilized to produce one predictive model when combined with one numerical prefactor, which is fitted by using available data. The prediction of superconductor's critical temperature [57] and adsorption energy [58] have already been performed by using SISSO.

With the help of a three-dimensional descriptor, taking the top two features of RFR described in Fig. 8 and the atomic volume of the chalcogen family, the mathematical formula found for the prediction of direct bandgap has the form:

$$\begin{aligned}(\text{Direct Bandgap})^{1/3} = &-0.044(\text{Atomic Volume Chalco.})\\ &-67\,345(\text{magmom}) - 0\,003(\text{hform})^3\\ &+1.793.\end{aligned}$$

SISSO is able to predict direct bandgap with a RMSE: 0.056 eV, as shown in Fig. 9**(a)**. We have considered zero direct bandgaps as an outlier and removed them from the SISSO input dataset. Now, we have 101 direct bandgap datasets. There are 79 SISSO predicted values that have a difference of ≤ 0.1 i.e. nearly agree with the DFT value. All the details of SISSO's predicted direct bandgap are given in the supplementary information section(III). Therefore, the SISSO model used in the dataset shows a robust prediction. Only direct bandgap material has photovoltaic application, so we focus on it as a target for

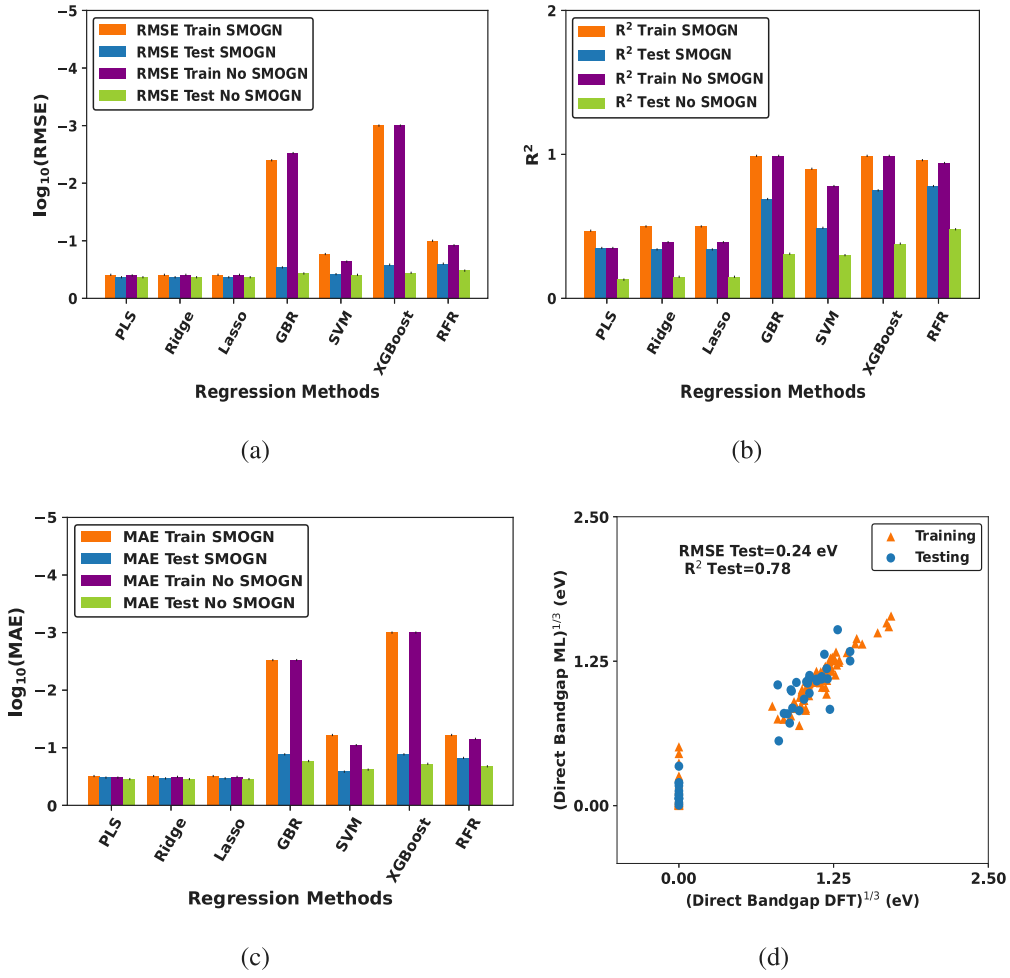

Fig. 7. The **(a)** root mean square error (RMSE), **(b)** R^2 and **(c)** mean absolute error (MAE) of various ML models for the prediction of TMD bandgap. **(d)** the plot between DFT and ML predicted bandgap of the TMD by using RFR.

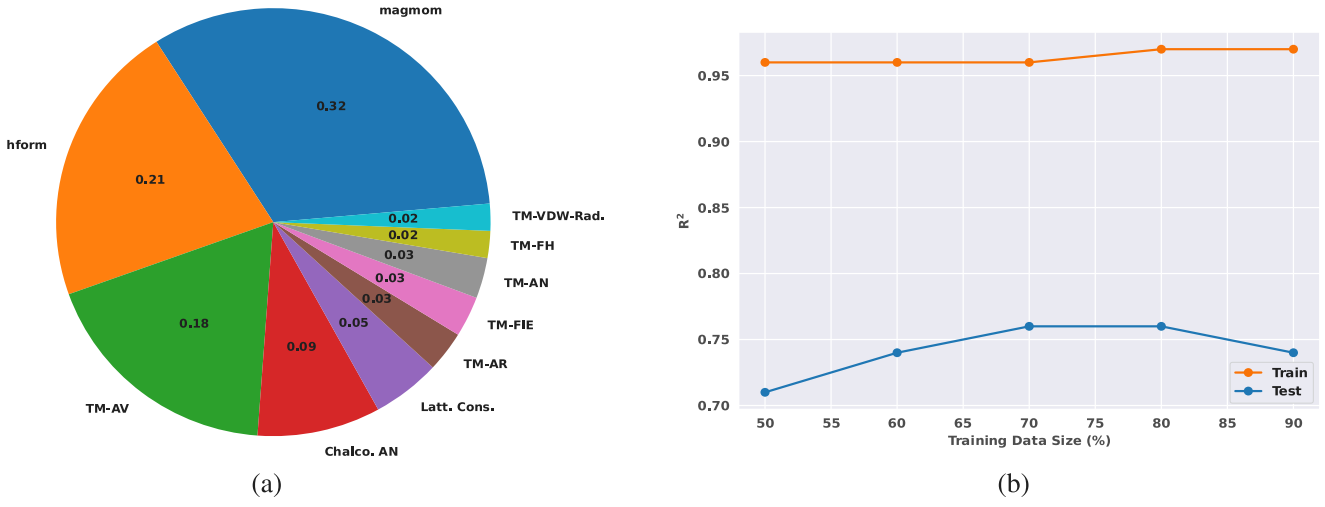

Fig. 8. **(a)** The feature importance by RFR model. **(b)** The plot between training data and R^2 i.e. learning curve. The testing data is complementary to training data.

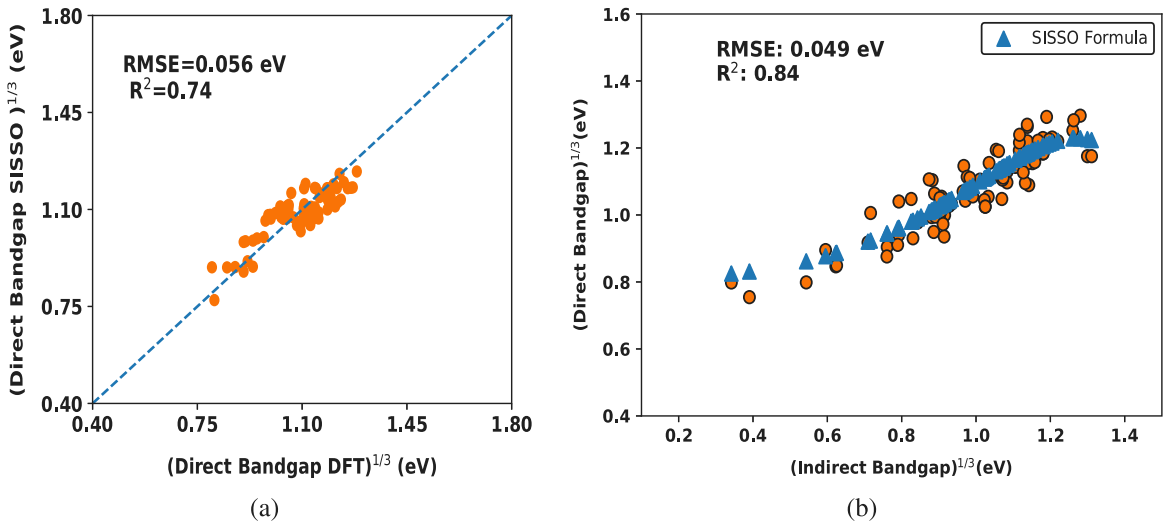

Fig. 9. **(a)** The plot between direct bandgap calculated by DFT and SISSO, **(b)** The plot between direct and indirect bandgap of the TMD.

our ML model. The direct and indirect bandgaps are highly correlated features. The relationship between direct and indirect bandgap has form:

$$(\text{Direct Bandgap})^{1/3} = 0.304(\text{Indirect Bandgap}) - 0.024(\text{Indirect Bandgap})^3 + 0.813$$

Above SISSO formula has the capability of predicting bandgap with 84% accuracy, as shown in Fig. 9(**b**).

4. Conclusion

In this work, we perform a study of the direct bandgap behavior of the TMDCs and TMOs via the ML approach. The value of the bandgap highly depends on the magnetic moment, the heat of formation energy and the atomic volume of these compounds. By considering the bandgap family in the two classes, i.e., 0 (zero bandgap) and 1 (non-zero bandgap), the classification has been performed with 95% accuracy. The bandgap value of these materials can also be predicted (regression ML approach) with 78% accuracy. A compressed sensing method, SISSO has been applied to get the analytical formula of the bandgap. Such a formula predicts bandgap with an accuracy of 74%. Overall, this study will help in the prediction of the bandgap of new materials in the TMDCs and TMOs families. Students and researchers may utilize this model in studies involving unidentified bandgaps or novel substances in the family of TMDCs and TMOs.

CRediT authorship contribution statement

Upendra Kumar: Conceived the idea of utilizing this open-source data and performed the all machine learning model related calculation, Wrote the manuscript. **Km Arti Mishra:** Compressive sensing in prospective of used dataset, Wrote the manuscript. **Ajay Kumar Kushwaha:** Wrote the manuscript and involved in verifying the key calculations. **Sung Beom Cho:** Supervision.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Sung Beom Cho reports financial support was provided by National Research Foundation of Korea.

Data availability

The authors do not have permission to share data.

Acknowledgments

We gratefully acknowledge support from the National Research Foundation of Korea, South Korea (2020M3H4A3081867 and 2022R1F1A1063060). The computations were carried out using resources from Korea Supercomputing Center (KSC-2021-RND-0025). Upendra Kumar is profoundly thankful to Dr. Seung-Cheol Lee (Director at Indo-Korea Science and Technology Center and a Principal Research Scientist, Center for Electronic Materials Research, Korea Institute of Science and Technology) for giving him the motivation to work in machine learning.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jpcs.2022.110973.

References

[1] A.K. Geim, K.S. Novoselov, The rise of graphene, in: Nanoscience and Technology: A Collection of Reviews from Nature Journals, World Scientific, 2010, pp. 11–19.
[2] Enamullah, V. Kumar, U. Kumar, G.S. Setlur, Quantum Rabi oscillations in graphene, J. Opt. Soc. Amer. B 31 (2014) 484.
[3] S. Manzeli, D. Ovchinnikov, D. Pasquier, O.V. Yazyev, A. Kis, 2D transition metal dichalcogenides, Nature Rev. Mater. 2 (2017) 1.
[4] Q.H. Wang, K. Kalantar-Zadeh, A. Kis, J.N. Coleman, M.S. Strano, Electronics and optoelectronics of two-dimensional transition metal dichalcogenides, Nature Nanotechnol. 7 (2012) 699.
[5] K.F. Mak, C. Lee, J. Hone, J. Shan, T.F. Heinz, Atomically thin mos_2: a new direct-gap semiconductor, Phys. Rev. Lett. 105 (2010) 136805.
[6] S.C. de la Barrera, M.R. Sinko, D.P. Gopalan, N. Sivadas, K.L. Seyler, K. Watanabe, T. Taniguchi, A.W. Tsen, X. Xu, D. Xiao, et al., Tuning ising superconductivity with layer and spin–orbit coupling in two-dimensional transition-metal dichalcogenides, Nature Commun. 9 (2018) 1.
[7] A. Kuc, T. Heine, A. Kis, Electronic properties of transition-metal dichalcogenides, MRS Bull. 40 (2015) 577.
[8] N. Zibouche, A. Kuc, J. Musfeldt, T. Heine, Transition-metal dichalcogenides for spintronic applications, Ann. Phys. 526 (2014) 395.
[9] D. Bhattacharya, S. Bayan, R.K. Mitra, S.K. Ray, Flexible biomechanical energy harvesters with colossal piezoelectric output (2.07 v/kpa) based on transition metal dichalcogenides-poly (vinylidene fluoride) nanocomposites, ACS Appl. Electron. Mater. 2 (2020) 3327.
[10] L. Zheng, X. Wang, H. Jiang, M. Xu, W. Huang, Z. Liu, Recent progress of flexible electronics by 2d transition metal dichalcogenides, Nano Res. (2021) 1.
[11] Z. Li, S.L. Wong, Functionalization of 2d transition metal dichalcogenides for biomedical applications, Mater. Sci. Eng. C 70 (2017) 1095.
[12] B. Radisavljevic, A. Radenovic, J. Brivio, V. Giacometti, A. Kis, Single-layer mos_2 transistors, Nature Nanotechnol. 6 (2011) 147.

[13] A. Splendiani, L. Sun, Y. Zhang, T. Li, J. Kim, C.-Y. Chim, G. Galli, F. Wang, Emerging photoluminescence in monolayer mos2, Nano Lett. 10 (2010) 1271.
[14] T. Cao, G. Wang, W. Han, H. Ye, C. Zhu, J. Shi, Q. Niu, P. Tan, E. Wang, B. Liu, et al., Valley-selective circular dichroism of monolayer molybdenum disulphide, Nature Commun. 3 (2012) 1.
[15] N. Arunadevi, S.J. Kirubavathy, Metal oxides: Advanced inorganic materials, Inorg. Anticorrosive Mater. (2022) 21.
[16] V. Soni, P. Singh, A.A.P. Khan, A. Singh, A.K. Nadda, C.M. Hussain, Q. Van Le, S. Rizevsky, V.-H. Nguyen, P. Raizada, Photocatalytic transition-metal-oxides-based p–n heterojunction materials: synthesis, sustainable energy and environmental applications, and perspectives, J. Nanostruct. Chem. (2022) 1.
[17] K.J. Hamam, M.I. Alomari, A study of the optical band gap of zinc phthalocyanine nanoparticles using uv–vis spectroscopy and dft function, Appl. Nanosci. 7 (2017) 261.
[18] C.A. Thomas, K. Zong, K.A. Abboud, P.J. Steel, J.R. Reynolds, Donor-mediated band gap reduction in a homologous series of conjugated polymers, J. Am. Chem. Soc. 126 (2004) 16440.
[19] W. Kohn, L.J. Sham, Self-consistent equations including exchange and correlation effects, Phys. Rev. 140 (1965) A1133.
[20] F. Aryasetiawan, O. Gunnarsson, The gw method, Rep. Progr. Phys. 61 (1998) 237.
[21] J.M. Crowley, J. Tahir-Kheli, W.A. Goddard III, Resolution of the band gap prediction problem for materials design, J. Phys. Chem. Lett. 7 (2016) 1198.
[22] A.F. Bialon, T. Hammerschmidt, R. Drautz, Three-parameter crystal-structure prediction for sp-d-valent compounds, Chem. Mater. 28 (2016) 2550.
[23] G. Pilania, C. Wang, X. Jiang, S. Rajasekaran, R. Ramprasad, Accelerating materials property predictions using machine learning, Sci. Rep. 3 (2013) 1.
[24] A. Seko, H. Hayashi, K. Nakayama, A. Takahashi, I. Tanaka, Representation of compounds for machine-learning prediction of physical properties, Phys. Rev. B 95 (2017) 144110.
[25] 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 (2016) 115104.
[26] F. Legrain, J. Carrete, A. van Roekeghem, S. Curtarolo, N. Mingo, How chemical composition alone can predict vibrational free energies and entropies of solids, Chem. Mater. 29 (2017) 6220.
[27] O. Isayev, C. Oses, C. Toher, E. Gossett, S. Curtarolo, A. Tropsha, Universal fragment descriptors for predicting properties of inorganic crystals, Nature Commun. 8 (2017) 1.
[28] G. Pilania, A. Mannodi-Kanakkithodi, B. Uberuaga, R. Ramprasad, J. Gubernatis, T. Lookman, Machine learning bandgaps of double perovskites, Sci. Rep. 6 (2016) 1.
[29] P. Dey, J. Bible, S. Datta, S. Broderick, J. Jasinski, M. Sunkara, M. Menon, K. Rajan, Informatics-aided bandgap engineering for solar materials, Comput. Mater. Sci. 83 (2014) 185.
[30] Z. Zhaochun, P. Ruiwu, C. Nianyi, Artificial neural network prediction of the band gap and melting point of binary and ternary compound semiconductors, Mater. Sci. Eng. B 54 (1998) 149.
[31] T. Gu, W. Lu, X. Bao, N. Chen, Using support vector regression for the prediction of the band gap and melting point of binary and ternary compound semiconductors, Solid State Sci. 8 (2006) 129.
[32] G. Pilania, J.E. Gubernatis, T. Lookman, Multi-fidelity machine learning models for accurate bandgap predictions of solids, Comput. Mater. Sci. 129 (2017) 156.
[33] S. Curtarolo, W. Setyawan, G.L. Hart, M. Jahnatek, R.V. Chepulskii, R.H. Taylor, S. Wang, J. Xue, K. Yang, O. Levy, et al., Aflow: An automatic framework for high-throughput materials discovery, Comput. Mater. Sci. 58 (2012) 218.
[34] T. Xie, J.C. Grossman, Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties, Phys. Rev. Lett. 120 (2018) 145301.
[35] Y. Zhang, W. Xu, G. Liu, Z. Zhang, J. Zhu, M. Li, Bandgap prediction of two-dimensional materials using machine learning, PLoS One 16 (2021) e0255637.
[36] 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.
[37] T. Wang, X. Tan, Y. Wei, H. Jin, Accurate bandgap predictions of solids assisted by machine learning, Mater. Today Commun. 29 (2021) 102932.
[38] Z. Wan, Q.-D. Wang, D. Liu, J. Liang, Machine learning prediction of the optimal carrier concentration and band gap of quaternary thermoelectric materials via element feature descriptors, Int. J. Quantum Chem. 121 (2021) e26752.
[39] P. Xu, T. Lu, L. Ju, L. Tian, M. Li, W. Lu, Machine learning aided design of polymer with targeted band gap based on dft computation, J. Phys. Chem. B 125 (2021) 601.
[40] N.R. Knøsgaard, K.S. Thygesen, Representing individual electronic states for machine learning gw band structures of 2d materials, Nature Commun. 13 (2022) 1.
[41] M.N. Gjerding, A. Taghizadeh, A. Rasmussen, S. Ali, F. Bertoldo, T. Deilmann, N.R. Knøsgaard, M. Kruse, A.H. Larsen, S. Manti, et al., Recent progress of the computational 2d materials database (c2db), 2D Mater. 8 (2021) 044002.
[42] R. Ouyang, S. Curtarolo, E. Ahmetcik, M. Scheffler, L.M. Ghiringhelli, Sisso: A compressed-sensing method for identifying the best low-dimensional descriptor in an immensity of offered candidates, Phys. Rev. Mater. 2 (2018) 083802.
[43] N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, Smote: synthetic minority over-sampling technique, J. Artificial Intelligence Res. 16 (2002) 321.
[44] P. Branco, L. Torgo, R.P. Ribeiro, Smogn: a pre-processing approach for imbalanced regression, in: First International Workshop on Learning with Imbalanced Domains: Theory and Applications, PMLR, 2017, pp. 36–50.
[45] F.A. Rasmussen, K.S. Thygesen, Computational 2d materials database: electronic structure of transition-metal dichalcogenides and oxides, J. Phys. Chem. C 119 (2015) 13169.
[46] J. Enkovaara, C. Rostgaard, J.J. Mortensen, J. Chen, M. Dułak, L. Ferrighi, J. Gavnholt, C. Glinsvad, V. Haikola, H. Hansen, et al., Electronic structure calculations with gpaw: a real-space implementation of the projector augmented-wave method, J. Phys.: Condens. Matter 22 (2010) 253202.
[47] L. Zheng, Q. Yao, H. Wang, H. Zhan, W. Cai, Y. Zhou, J. Kang, Band structure regulation in fe-doped mgzno by initial magnetic moments, RSC Adv. 11 (2021) 3209.
[48] L.S. de Oliveira, F.P. Sabino, D.Z. de Florio, A. Janotti, G.M. Dalpian, J.A. Souza, Insulator–metal transition in the nd2cofeo6 disordered double perovskite, J. Phys. Chem. C 124 (2020) 22733.
[49] W. Bludau, A. Onton, W. Heinke, Temperature dependence of the band gap of silicon, J. Appl. Phys. 45 (1974) 1846.
[50] R. Dalven, Empirical relation between energy gap and lattice constant in cubic semiconductors, Phys. Rev. B 8 (1973) 6033.
[51] J.N. Mandrekar, Receiver operating characteristic curve in diagnostic test assessment, J. Thorac. Oncol. 5 (2010) 1315.
[52] H. He, E.A. Garcia, Learning from imbalanced data, IEEE Trans. Knowl. Data Eng. 21 (2009) 1263.
[53] P. Branco, L. Torgo, R.P. Ribeiro, A survey of predictive modeling on imbalanced domains, ACM Comput. Surv. 49 (2016) 1.
[54] 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.
[55] L. Buitinck, G. Louppe, M. Blondel, F. Pedregosa, A. Mueller, O. Grisel, V. Niculae, P. Prettenhofer, A. Gramfort, J. Grobler, R. Layton, J. VanderPlas, A. Joly, B. Holt, G. Varoquaux, API design for machine learning software: experiences from the scikit-learn project, in: ECML PKDD Workshop: Languages for Data Mining and Machine Learning, 2013, pp. 108–122.
[56] Q.-S. Xu, Y.-Z. Liang, Monte Carlo cross validation, Chemometr. Intell. Lab. Syst. 56 (2001) 1.
[57] S. Xie, G. Stewart, J. Hamlin, P. Hirschfeld, R. Hennig, Functional form of the superconducting critical temperature from machine learning, Phys. Rev. B 100 (2019) 174513.
[58] S. Nayak, S. Bhattacharjee, J.-H. Choi, S.C. Lee, Machine learning and scaling laws for prediction of accurate adsorption energy, J. Phys. Chem. A 124 (2019) 247.