Check for updates

Machine learning-enabled prediction of the electronic band-edge shapes and properties of 2D transition metal dichalcogenide alloys

Tarvir Anjum Aditto, [iD] † Vivek Chowdhury, [iD] † Hafiz Imtiaz [iD] * and Ahmed Zubair [iD] *

Conventional *ab initio* calculations based on density functional theory (DFT), though accurate, are computationally expensive and time-consuming, paving the way for the emergence of data-driven models. While most existing works are confined to scalar electronic property estimation, such as band gaps or stability, whole-band structure prediction has remained relatively unexplored. Here, we derived a dataset from DFT computations of monolayer transition metal dichalcogenide (TMD) alloys composed of tungsten (W), molybdenum (Mo), and chalcogen (X) atoms, namely sulfur (S), selenium (Se), and tellurium (Te), and developed a decision tree based model – extra trees, for predicting the conduction and valence band structures for binary and ternary TMDs. The model illustrated the dependency of the band structure on crucial features, such as compositional ratios. We trained separate models for the conduction and valence bands to capture non-linear feature–target relationships. We note that we performed a comparative analysis of different types of machine learning models, and found that the extra trees model performed the best. Low mean squared error (MSE < 0.001) and high correlation values (Pearson correlation coefficient ≫ 0.99) with respect to DFT simulated band energies for all the TMD alloys of the dataset validate the excellent predictive power of the model. Additionally, the model made excellent predictions for compositions beyond the training dataset, showcasing the robustness of the model as a substitute for computationally heavy DFT simulation. We emphasize that the proposed whole band model is more informative than the existing band gap-only models, since it provides the band gaps and enables the derivation of secondary electronic properties, such as effective carrier mass by curvature fitting. As such, we argue that our work in this study will benefit materials research in nanoscience by providing a faster and insightful alternative framework to computer-based simulations.

1 Introduction

Two-dimensional transition metal dichalcogenides (TMDs) have emerged as a pivotal class of materials in nanoscience, distinguished by properties that complement and overcome the limitations of graphene and other nanomaterials.[1] Platinum is the benchmark catalyst for hydrogen production in fuel cells; however, TMDs offer a promising, low-cost, and earth-abundant alternative.[2] Another vibrant application of TMD research is in sensors, where the goal is to exploit the material's high surface sensitivity and tunable electrical properties for detecting gases, chemicals, and biomolecules.[3] Intriguingly, 2D material TMDs can be used in nanomedicine as a carrier for targeted drug

delivery.[4] These TMDs, commonly named as MX_2 compounds (where M is a transition metal and X is a chalcogen), consist of covalently bonded atomic planes stacked by weak van der Waals forces, enabling exfoliation into atomically thin layers.[5] Today, new opportunities are arising from using nanostructured and few-layer TMDs to improve the capacity, rate performance, and cycling stability of electrodes.[6] In contrast to graphene, many semiconducting TMDs exhibit sizable band gaps in the visible to infrared range. Notably, monolayer TMDs such as MoS_2, $MoSe_2$, WS_2, and WSe_2 undergo an indirect to direct band-gap crossover at the monolayer limit, yielding strong light–matter coupling and robust photoluminescence.[5,7] These properties have enabled a range of nanoelectronic and photonic devices: for example, TMD monolayers have been used to make high-performance field-effect transistors, photodetectors, solar cells, light-emitting diodes, and lasers.[5,7–9] In particular, group-VI TMDs (with M = Mo, W and X = S, Se, Te) combine direct band gap behavior with high carrier mobility and efficient switching, addressing

Department of Electrical and Electronic Engineering, Bangladesh University of Engineering and Technology, Dhaka 1205, Bangladesh.
E-mail: hafizimtiaz@eee.buet.ac.bd, ahmedzubair@eee.buet.ac.bd
† These authors contributed equally.

graphene's lack of an intrinsic gap.[7,8] Among the 2H-phase semiconducting TMDs, the Mo- and W-based dichalcogenides are especially prominent.[7,10] In single-layer form, these materials have optical band gaps on the order of 1–2 eV, placing them in the visible and near-infrared spectrum.[7,10]

Layered transition-metal dichalcogenides exhibit tunable electronic properties that depend on their metal and chalcogen composition. In particular, mixing different metals and chalcogenides produces ternary and quaternary TMD alloys with continuously adjustable band gaps. Susarla et al. reported that quaternary Mo–W–S–Se monolayers can achieve band gaps tunable from roughly 1.60 to 2.03 eV.[11] Such complex alloys are of great interest for electronics and optoelectronics; however, their full electronic band structures must be known to accurately predict electronic transport and optical behavior. Traditionally, density functional theory (DFT) has been used to compute TMD band structures with high accuracy, but this approach is computationally expensive. A full DFT band-structure calculation requires solving the Kohn–Sham equations on dense k-point grids (often including spin–orbit coupling for heavy metals), leading to very large computational workloads. This high cost is evident in large-scale studies; for instance, Muller et al. performed plane-wave DFT relaxations for 672 ternary TMD compositions (over 50 000 individual relaxation steps) to build an electronic structure dataset.[12] In general, high-throughput DFT screening of broad TMD composition spaces is prohibitive without massive computing resources.[13,14]

Since computer-based first-principles calculations are computationally expensive, particularly in the case of multinary alloys, there is an increasing trend to shift towards data-driven machine learning (ML) models that can accelerate property prediction by learning structure–property relationships from a small number of samples. Previous studies that integrated computer-based simulation and machine learning have explored a wide range of applications and demonstrated success across various material classes. Yuan et al.[15] predicted band gap opening in graphene-semiconductor systems, while Nguyen et al.[16] estimated band gap, electron affinity, and ionization potentials for polycyclic aromatic hydrocarbons (PAHs). Yi-Hsuan Liu et al.[17] modeled local electronic properties such as on-site electron number and double occupation for a disordered correlated electron system, and Di Liu et al.[18] combined EXtreme gradient boosting (XGBoost) and convolutional neural network for density of states prediction for MoB/ Si$_3$N$_4$ superlattice heterojunctions. A similar data-driven approach has been applied to investigate the stability of atomic configurations like γ-Al$_2$O$_3$[19] and for formation energy and band gaps for perovskites such as ABX$_3$ and A$_2$BB'X$_6$.[20]

Despite these advances in electronic property predictions, few studies have explored ML applications in TMD alloys. Wang et al.[21] predicted layer-dependent electronic properties like band gaps for binary TMDs using ML, while Xu et al.[22] employed a graph crystal neural network to predict band gaps and band alignment types in TMDs/2D-LHP heterostructures. Siddiqui et al. applied equivariant neural network to predict vibrational properties of quaternary TMD alloys of the form

Mo$_{1-x}$W$_x$S$_{2-2y}$Se$_{2y}$. Other notable research includes that of Gao et al.,[23] who used DFT generated data of binary to quaternary TMD alloys to train neural networks for band gap prediction, a faster alternative to intensive DFT simulations; Kumar et al.[24] who developed decision tree and support vector machine (SVM) models to classify phase stability in TMDs (2H vs. 1T); and Jia et al.[25] who focused on binding energy and charge transfer in transition metal doped TMDs applying decision tree-based models like random forest and SVM. Wang et al.[5] used a transfer learning approach to predict the band gap of mono-layer materials. This involves applying knowledge from a pre-trained model based on similar tasks. Although deep neural networks (DNNs) show promise, they require large amounts of training data to generalize effectively.

While these studies demonstrated the potential of machine learning in electronic property prediction, they are primarily confined to scalar quantities, such as band gaps and phase stability. To the best of our knowledge, no previous research has attempted to explicitly predict the whole band structure shapes or extract secondary features, such as effective carrier mass, from data-driven models.

In this paper, we constructed a dataset derived from DFT calculations that captures both structural and electronic descriptors of different binary and ternary TMD alloys. We expanded the application of machine learning to predict electronic band structure by learning the conduction and valence band energies at multiple k-points with high fidelity. In our experiments, we revealed that extra trees, a decision tree-based model, achieved the best performance. The model was used to predict band structure shapes of compositions beyond the training dataset. Our models were not only capable of reconstructing the band structure but also derived other electronic properties, such as effective mass and band gap. Effective masses, for example, can be derived directly through curve fitting at the band edges, and band gaps can be calculated using only the conduction-band minimum and valence-band maximum. Thus, predicting band gaps is more versatile compared to band-gap-only models, offering deeper insights into electronic properties and opening up opportunities for the rapid exploration of alloy compositions for diverse electronic, optoelectronic, and spintronic applications.

2 Computational details

In this section, we describe the dataset generation procedure, the machine learning models, and the evaluation metrics used here. The overall workflow of this study is presented in Fig. 1.

2.1 Density functional theory calculations

The first-principles calculations were performed using the QUANTUM ESPRESSO simulation package.[26] The exchange–correlation functional was treated within the generalized-gradient approximation of Perdew–Burke–Ernzerhof (GGA-PBE). As hybrid functionals, such as HSE, GW, and PBE+U are computationally more expensive and do not change the
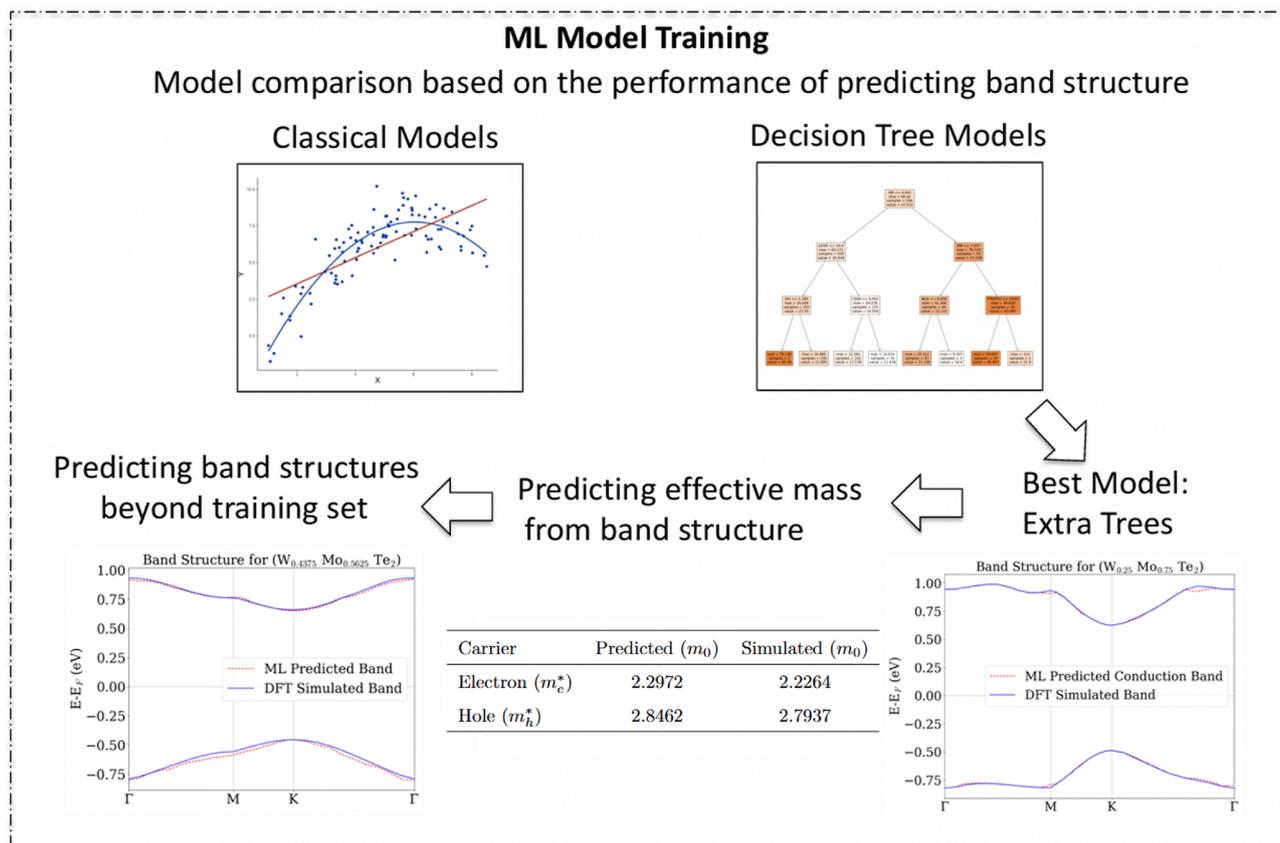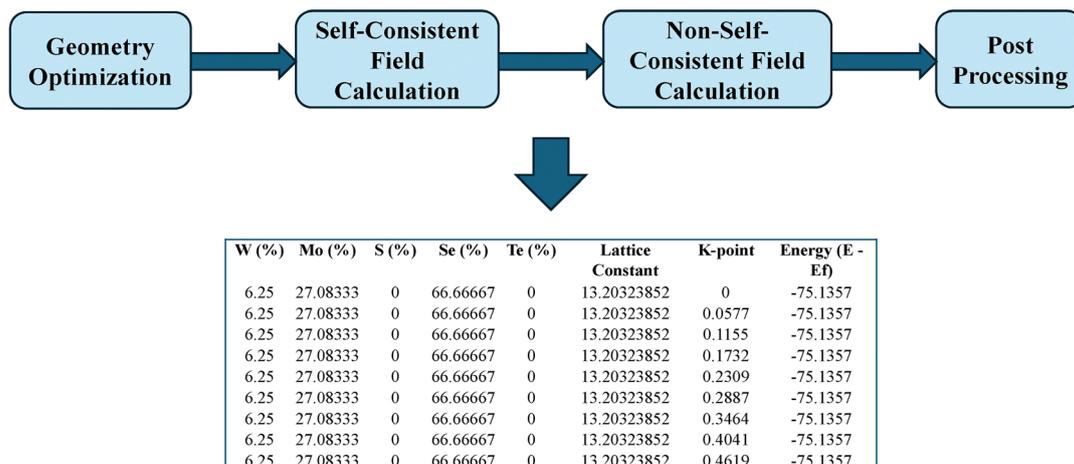
Dataset Generation

W (%)	Mo (%)	S (%)	Se (%)	Te (%)	Lattice Constant	K-point	Energy (E - Ef)
6.25	27.08333	0	66.66667	0	13.20323852	0	-75.1357
6.25	27.08333	0	66.66667	0	13.20323852	0.0577	-75.1357
6.25	27.08333	0	66.66667	0	13.20323852	0.1155	-75.1357
6.25	27.08333	0	66.66667	0	13.20323852	0.1732	-75.1357
6.25	27.08333	0	66.66667	0	13.20323852	0.2309	-75.1357
6.25	27.08333	0	66.66667	0	13.20323852	0.2887	-75.1357
6.25	27.08333	0	66.66667	0	13.20323852	0.3464	-75.1357
6.25	27.08333	0	66.66667	0	13.20323852	0.4041	-75.1357
6.25	27.08333	0	66.66667	0	13.20323852	0.4619	-75.1357

ML Model Training
Model comparison based on the performance of predicting band structure

Carrier	Predicted (m_0)	Simulated (m_0)
Electron (m_e^*)	2.2972	2.2264
Hole (m_h^*)	2.8462	2.7937

Fig. 1 A workflow diagram representing dataset generation and the ML methodology of this study. Lattice constants and band energies for different compositions were calculated using DFT simulations, which produced the training/testing datasets for machine learning models. The performance of different ML models was assessed using metrics such as mean squared error and correlation analysis. The best-performing model, extra trees, was then applied to predict the conduction and valence band structures across different compositions, even for the compositions that were unseen during training.

band shape, we focused on GGA-PBE functionals.[27,28] Scalar-relativistic projector-augmented-wave (PAW) pseudopotentials were employed for W, Mo, S, Se, and Te. The kinetic-energy cutoff for plane-wave basis sets was set to 500 eV for the wavefunctions, with a charge-density cutoff of 5000 eV. A Gaussian smearing of width 0.05 eV was used to handle partial electronic occupations. All calculations were performed in the non-spin-polarized scheme.

2.1.1 Supercell construction and alloying. To model controlled alloying between Mo and W atoms, we constructed a $(4 \times 4 \times 1)$ supercell of the parent dichalcogenide structure as can be seen in Fig. 2. In this scheme, one chalcogen atom (S, Se, or Te) was fixed, while the ratio of Mo:W atoms was systematically varied in the transition-metal layer. The $(4 \times 4 \times 1)$ supercell (a 16-fold replication of the primitive cell) was adopted in this work primarily to enable a systematic construction and counting of distinct binary and ternary alloy configurations involving W, Mo, S, Se, and Te within a single consistent computational setup.

We first built the parent $(4 \times 4 \times 1)$, which is supercell $Mo_{16}X_{32}$. We then randomly substituted $16y$ (where y is a fraction of a particular atom, which can be a minimum of 0.0625 and a maximum of 1) Mo atoms with W atoms to achieve the desired composition. For instance, $W_2Mo_{14}S_{32}$ was obtained by replacing two Mo sites in $Mo_{16}S_{32}$ with W. Binary alloys were generated analogously by substituting only one sublattice at a time while maintaining the TMD stoichiometry.

The alloy composition, y for the transition metal was evaluated by,

$$y = \frac{N_{specified}}{N_{total}}, \quad (1)$$

where

$$N_{total} = N_{unit-cell} \times N_x \times N_y \times N_z. \quad (2)$$

Here, $N_{specified}$ is the number of atoms of the selected element in the supercell, and $N_{unit-cell}$ is the number of transition metal atoms in the primitive unit cell. The integers N_x, N_y, and N_z denote the number of repetitions of the primitive unit cell along the lattice vectors $\mathbf{a}_1$, $\mathbf{a}_2$, and $\mathbf{a}_3$, respectively, so that the supercell size is $(N_x \times N_y \times N_z)$ and the total replication factor is $N_xN_yN_z$. For monolayer calculations, $N_z = 1$, hence, $N_{total} = N_{unit-cell} \times N_x \times N_y$. For monolayer calculations, $N_z = 1$, and the supercell is therefore $(N_x \times N_y \times 1)$.

2.1.2 Structural relaxation. The equilibrium structures were obtained using the Broyden–Fletcher–Goldfarb–Shanno (BFGS) quasi-Newton algorithm.[29] Both the ionic positions and in-plane cell vectors were optimized with the following convergence thresholds: 1.0×10^{-8} eV for electronic self-consistency, 1.0×10^{-3} eV Å^{-1} for atomic forces, 0.5 kbar for the residual pressure, and 5.3×10^{-7} eV as the ionic total-energy threshold between consecutive steps.

A maximum of 100 ionic steps was allowed during optimization. Charge-density mixing was performed using a mixing factor of 0.4, with up to 200 electronic iterations per ionic step. To eliminate spurious interlayer interactions, a vacuum slab of at least 18–20 Å was added along the out-of-plane direction.

2.1.3 Brillouin-zone sampling. Since a larger real-space supercell corresponds to a smaller reciprocal-space Brillouin zone, we used a Monkhorst–Pack mesh of $(2 \times 2 \times 1)$ with no shift for Brillouin-zone integration during the self-consistent calculations. A less dense k mesh is usually sufficient since a larger real-space supercell corresponds to a smaller reciprocal-space Brillouin zone.
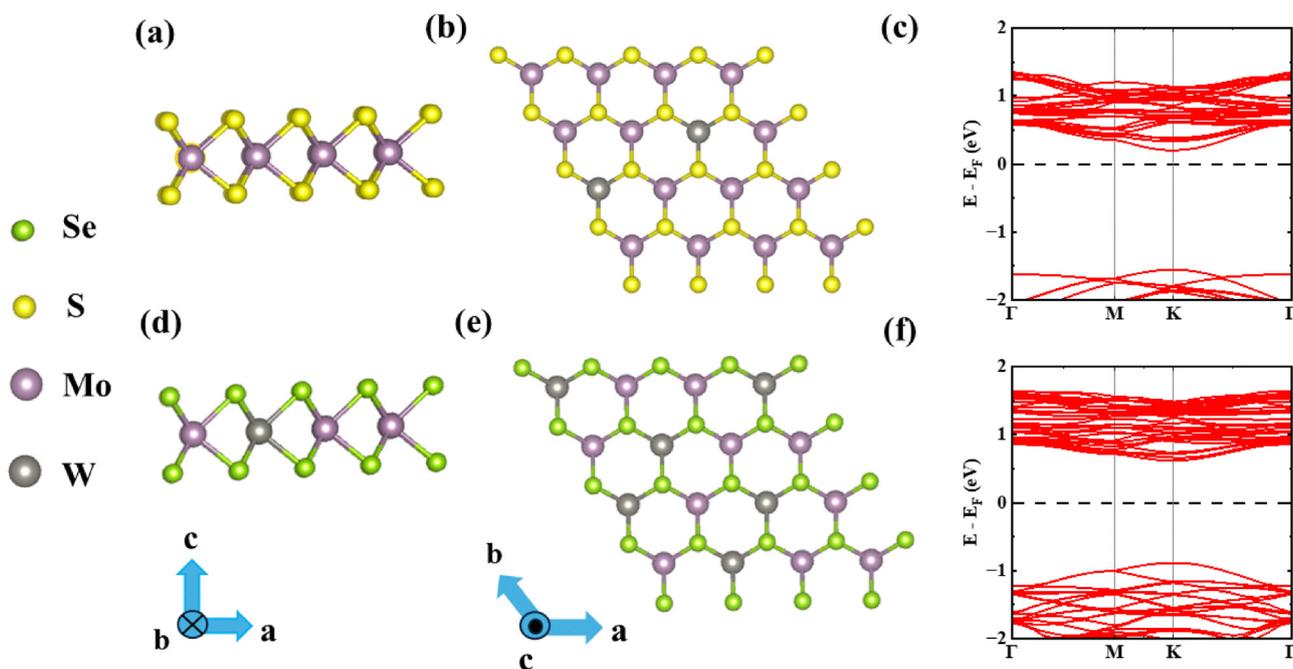

Fig. 2 (a) Side view of $W_{0.125}Mo_{0.875}S_2$, (b) top view of $W_{0.125}Mo_{0.875}S_2$, (c) *ab initio* band structure of $W_{0.125}Mo_{0.875}S_2$, (d) side view of $W_{0.375}Mo_{0.625}Se_2$, (e) top view of $W_{0.375}Mo_{0.625}Se_2$, and (f) *ab initio* band structure of $W_{0.375}Mo_{0.625}Se_2$.

Therefore, the mentioned k-point density ensured sufficient convergence of the total energy and forces while keeping computational cost manageable (see Section S5 of the SI).

2.1.4 Band structure calculations. Electronic band structures were computed using the self-consistent charge density obtained from the relaxed structures. The Kohn–Sham eigenvalues were evaluated along the high-symmetry path $\Gamma \rightarrow M \rightarrow K \rightarrow \Gamma$ of the hexagonal Brillouin zone. Each segment of the path was sampled using 10 equally spaced k-points, giving a total of 40 k-points. The resulting band dispersions provided direct insight into the effect of Mo–W alloying and chalcogen substitution on the electronic properties of the material.

2.2 Machine learning models

ML models rely on two key components: the algorithm and the material descriptors (or features) set, which define their performance. A good feature set must characterize the uniqueness of the material, be sensitive to the target property, and minimize redundancy among the features.[30,31] We used the presence of S/Se/Te, the percentage of W and Mo, and k-points as features for predicting conduction and valence band energy. The dataset can be found in this GitHub repository (https://github.com/ahmedzubair003/ML_Based_Bands_Prediction). Since the target properties, which were the conduction and valence band energies, are continuous-valued, we considered regression models. We started with linear models, and then investigated decision tree-based ensemble models. We noted that the intricate non-linear relation among the features and target variables was better captured in non-linear models.

2.2.1 Linear models. Linear regression assumes a linear relationship among the input features and target values:[32,33]

$$z = \mathbf{w}^T \mathbf{x} + b, \quad (3)$$

by minimizing the cost function:

$$J(\mathbf{w}, b) = \frac{1}{n} \sum_{i=1}^{n} \left(z_i - \left(\mathbf{w}^T \mathbf{x}_i + b\right)\right)^2. \quad (4)$$

Here, z is the target value, $\mathbf{x}$ is the input feature vector, $\mathbf{w}$ is the weight (parameters) vector that the model learns, and $\mathbf{b}$ is the bias term (intercept) of the model.

Ridge regression introduces an extra L_2 weight regularization penalty term to minimize overfitting as,[32]

$$J_{\text{ridge}}(\mathbf{w}, b) = \frac{1}{n} \sum_{i=1}^{n} \left(z_i - \left(\mathbf{w}^T \mathbf{x}_i + b\right)\right)^2 + \lambda \|\mathbf{w}\|^2, \quad (5)$$

where λ is the regularization parameter. $\lambda = 0$ for unregularized linear regression. Both models leverage the ordinary least squares (OLS) solution:

$$\hat{\mathbf{w}} = \left(\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I}\right)^{-1} \mathbf{X}^T z, \quad (6)$$

which minimizes the sum of the squared differences between the observed dependent variable in the input dataset and the output of the independent variable's (linear) function. Although efficient, the problem with linear and ridge regression is that they assume linearity, which may not capture the complex material-property relationship.[30]

2.2.2 Tree-based ensemble models. To address non-linear dependencies, we employed decision tree-based regressors; in fact, decision trees are a popular method for various machine learning tasks, especially when data are in tabular format.[34] We used three different versions of decision trees: random forest regressor (RFR), gradient boosting regressor (GBR), and extra trees (ET). Random forests, also known as random decision forests, is an ensemble learning method for classification, regression, and other problems that generates a large number of decision trees throughout the training process. For regression problems, the output of a random forest is the mean of the trees' predictions. In the random forest algorithm, a subsample of the training dataset is taken through bootstrapping, and the optimal split (threshold) is chosen to maximize impurity reduction, such as mean squared error (MSE) reduction for regression.[35]

$$\hat{z} = \frac{1}{M} \sum_{m=1}^{M} h_m(\mathbf{x}) \quad (7)$$

where $h_m(\mathbf{x})$ is the prediction of the mth tree. Extra trees differs in that it utilizes the entire dataset and randomly selects split thresholds within the feature range, then chooses the best one.[36]

The gradient boosting model leverages the basic idea of decision trees; however, it is sequential in nature, that is, trees are built sequentially, and each tree corrects the errors of the previous one using gradient descent. At each iteration m, the ensemble model is updated as,

$$F_m(\mathbf{x}) = F_{m-1}(\mathbf{x}) + \rho_m h_m(\mathbf{x}). \quad (8)$$

Here, $F_m(\mathbf{x})$ denotes the ensemble prediction function after the mth boosting stage, ρ_m is the optimal step size determined by minimizing the loss function at iteration m, and $h_m(\mathbf{x})$ represents the weak learner fitted to the negative gradient of the loss.[37]

Importantly, we can rank feature importance from all these tree-based models, enabling the identification of the most influential features for prediction.[35,36]

2.2.3 Model evaluation. For the evaluation of our models, we primarily used the MSE calculated by

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^{n} \left(z_i - \hat{z}_i\right)^2, \quad (9)$$

computed on the training and testing datasets. Here, the ML model predicted values are $\hat{z}$, and the DFT-simulated values are z, which are assumed to be the true values. To further evaluate the performance of the model, we measured the correlation between $\hat{z}$ and z. Correlation analysis quantifies how strongly two variables are linearly related to each other. The Pearson correlation coefficient r is commonly used and is defined as,

$$r = \frac{\sum_{i=1}^{n} \left(z_i - \bar{z}\right)\left(\hat{z}_i - \bar{\hat{z}}\right)}{\sqrt{\sum_{i=1}^{n} \left(z_i - \bar{z}\right)^2} \sqrt{\sum_{i=1}^{n} \left(\hat{z}_i - \bar{\hat{z}}\right)^2}} \quad (10)$$

where $\bar{z}$ and $\bar{\hat{z}}$ are means of the true and predicted values, respectively.

$r \approx +1$: strong positive correlation (predictions closely match the true values), $r \approx 0$: little to no correlation, $r \approx -1$: strong negative correlation. A high positive correlation coefficient indicates that the model is able to capture the characteristics of the DFT data effectively.

3 Results and discussion

3.1 Band structure calculated from DFT

We considered ternary TMDs composed of tungsten (W), molybdenum (Mo), and chalcogen (X) atoms, namely sulfur (S), selenium (Se), and tellurium (Te). Two transition metal atoms and one chalcogen atom were combined to construct the ternary TMD configurations. The *ab initio* band structure of $W_{0.125}Mo_{0.875}S_2$ is presented in Fig. 2(c), while that of $W_{0.375}Mo_{0.625}Se_2$ is shown in Fig. 2(f).

From the DFT calculations, the electronic band gaps for $W_2Mo_{14}S_{32}$ and $W_6Mo_{10}Se_{32}$ were obtained as 1.75 eV and 1.51 eV, respectively. As a consequence of band folding arising from the use of a $(4 \times 4 \times 1)$ supercell, a higher density of bands was observed in the band structure plots. In both cases, the band gap appears at the k point along the high-symmetry path $\Gamma \rightarrow M \rightarrow K \rightarrow \Gamma$.

3.2 Feature statistics and target values

The dataset, generated using DFT computations of monolayer TMD alloys ($W_yMo_{1-y}X_2$), comprises of a total of 51 unique alloy compositions, including binary and ternary systems, which were considered. For each composition, conduction and valence band energies were computed at 31 k-points.

Each data sample consists of the following input features: (i) one-hot encoded indicators for the chalcogen species (S, Se, and Te), (ii) the fractional composition of Mo and W, and (iii) the sampled k-points. These descriptors were selected to capture both chemical identity and structural/electronic characteristics. The corresponding target values are the conduction band (CB) energies and valence band (VB) energies at the specified k-points.

From Fig. 3, we can see that the range of values is very different across features. Therefore, prior to model training, features were normalized using the scikit-learn library[38] to ensure balanced weighting across the features. The dataset was randomly split into training and test subsets, with 70% of the samples used for training unless otherwise noted.

3.3 Train-test protocol

The dataset was split between training and testing subsets using a 70:30 split. As illustrated in Fig. 4, this ratio was chosen for maintaining a balance between the model's efficiency and generalization: a smaller training ratio resulted in significant performance loss; on the other hand, larger training ratios can affect generalization *i.e.*, ML models may overfit the
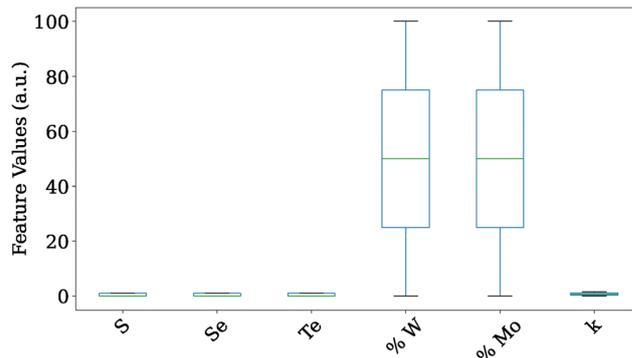

Fig. 3 Boxplots of the input features (in arbitrary units) used in training the machine learning models. The plots illustrate the natural distribution and variability of each feature without normalization, including one-hot encoded chalcogen presence (S, Se, and Te), percentage composition of W and Mo, and k-points. The boxplots provide insight into the raw dataset characteristics prior to modeling. The difference in ranges across features highlights the need for normalization.

training data. To ensure robustness, the results were averaged over 5 random splits with fixed seeds for reproducibility.

3.4 Model architecture

We used ML models, diversifying from simple linear regression approaches to tree-based ensembles, focusing on how architecture and hyperparameters affect the predictions of target values.

Linear regression is a baseline model that tries to predict the output using a straight line, eqn (3), while ridge regression adds L_2 regularization ($\alpha = 0.01$) to the loss function (5) to improve stability by keeping the weights regularized to reduce overfitting of the training data. To capture non-linear relationships, we used ensembled decision tree models. Random forest was employed with 100 shallow trees (depth = 2), where the depth refers to how many layers each tree has. Shallow trees were less likely to overfit, hence balancing variation with generalization. A gradient boosting model was constructed using 100 sequential trees (depth = 2) with squared error loss. In this sequential
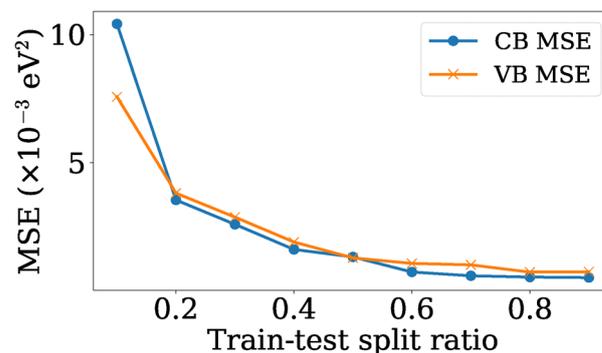

Fig. 4 MSE *vs.* training set size ratio for CB and VB energy predictions represented by the blue line with solid circle markers and the orange line with cross markers, respectively, for the extra trees model. A smaller training ratio negatively affects the performance of the model, while a larger ratio can overfit the training data. 70:30 appears to be an optimal choice for balancing model efficiency and generalization.

Table 1 Hyperparameters used for conduction/valence band energy prediction models

Model	Hyperparameters
Ridge regression	$\alpha = 0.01$
Random forest	$n_{estimators} = 100$, $depth_{max} = 2$
Gradient boosting	$n_{estimators} = 100$, $depth_{max} = 2$
Extra trees	$n_{estimators} = 100$, $depth_{max} = None$

Table 2 Performance of different models for conduction band and valence band prediction

Model	Train MSE		Test MSE	
	CB	VB	CB	VB
Ridge regression	0.0282	0.0165	0.0275	0.0160
Random forest	0.0256	0.0185	0.0264	0.0168
Gradient boosting	0.0029	0.0045	0.0029	0.0047
Extra trees	0.0000	0.0000	0.0005	0.0009

approach, each tree was trained to correct the error of previous trees, enabling the model to capture subtle residual patterns. Lastly, the extra trees model also utilized 100 trees but with randomized splits and no depth limit, utilizing randomness for robustness.

These settings, summarized in Table 1, were chosen based on exploratory trials and common practice. Fig. 5 illustrates the architecture of an extra trees model.

3.5 Machine learning model validation

To validate the performance of the different machine learning models, we trained separate models for conduction band and valence band energy prediction. Training and test errors were calculated using mean squared error, and the results from different regression models are summarized in Table 2. Band structures predicted from different models are shown in Fig. 6 and 7 along with the structures obtained from DFT calculations.

Across all the methods, tree-based ensembles consistently outperformed linear models, validating the existence of non-linear dependencies among the features and targets. Extra trees achieved the lowest MSE for both CB and VB predictions, making it the best model, followed by gradient boosting. In contrast, linear regression and ridge regression showed relatively higher

errors, demonstrating their limitations in capturing the non-linear dependencies in the dataset.

As shown in Fig. 6(a), ridge regression predicted a linear band structure for both the VB and CB, whereas the actual structures were not linear at all. Though random forest predicted structures that were visibly similar to the original one (Fig. 6(b)), it failed to capture the curvature properly, resulting in high MSE for both the VB and CB. The gradient boosting model exhibited slightly better results than random forest, but still struggled with accurately predicting the curvature, particularly around the CB minima and VB maxima, as we can see in Fig. 7(a). Failure in accurately predicting the curvature, especially around the CB minima and VB maxima, can cause significant error in calculating the band gap and effective carrier mass from the band structure.

In contrast, extra trees provided the most accurate predictions and captured the curvature especially around the CB minima and VB maxima (Fig. 7(b)), making it the superior one among all the models. The MSEs for both conduction and valence band energy were calculated using 5-fold cross-validation with the results showing minimal deviation (shown in Table 3), indicating that the model is consistent. A qualitative comparison between extra trees predicted and DFT-simulated band structures is
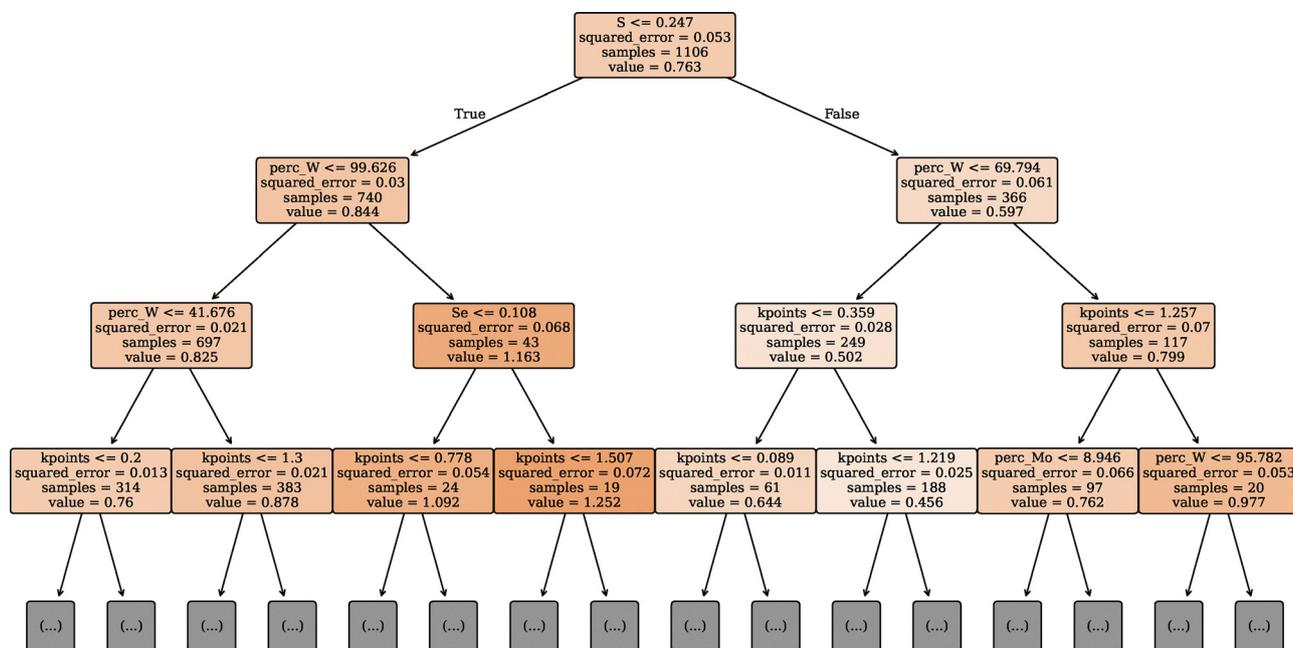
Fig. 5 A part of a decision tree from the extra trees ensemble model for CB. The tree is limited to a depth of 3 for clarity (see Section S2 of the SI for the full architecture).
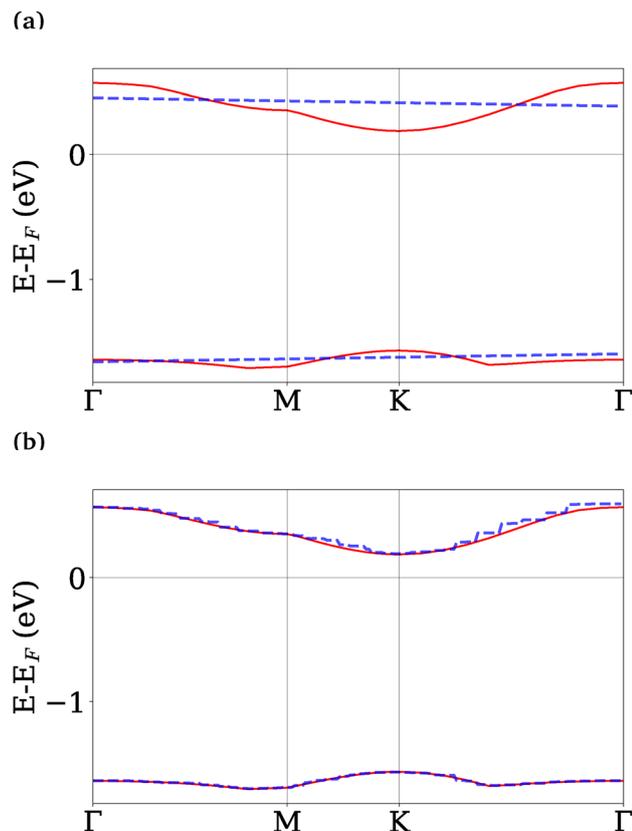
(a)

(b)

Fig. 6 Comparison of ML-predicted and DFT-simulated band structures for $W_{0.0625}Mo_{0.9375}S_2$: (a) ridge regression and (b) random forest. The continuous and broken lines represent ground truth and ML-predicted structures, respectively.
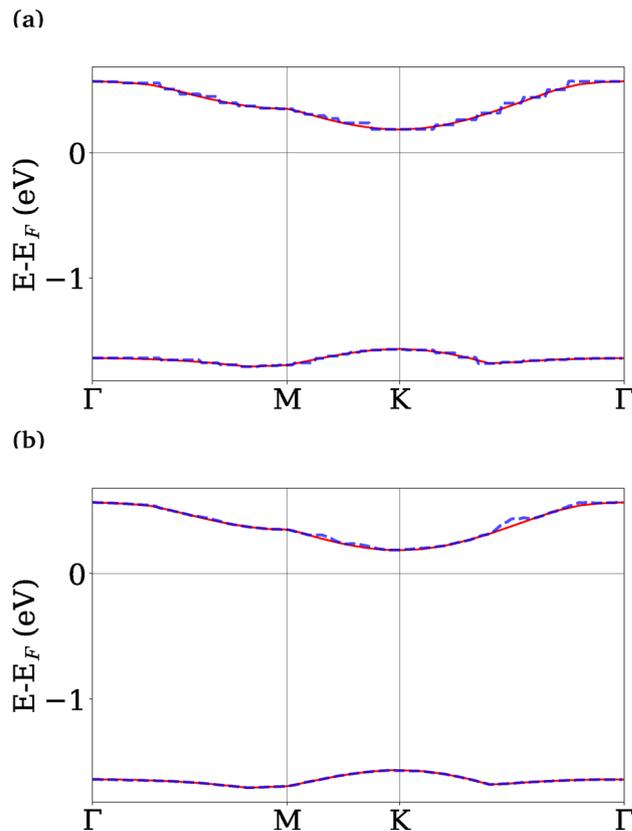
(a)

(b)

Fig. 7 Comparison of ML-predicted and DFT-simulated band structures for $W_{0.0625}Mo_{0.9375}S_2$: (a) gradient boosting and (b) extra trees. Solid and dashed lines represent DFT ground truth and ML predictions, respectively.

presented in Fig. 8 for 6 different compositions. The predicted band structure from the best model almost superimposes on the DFT-simulated structure. This highlights that the ML models, particularly extra trees, are capable of not only achieving numerically accurate results but also capturing the physical shape of the band structure across k-space, which can be beneficial in the successful calculation of secondary properties like effective carrier mass and band gap.

3.6 Feature importance

Fig. 9 and 10 present an analysis of feature importance for conduction band and valence band energy, respectively. For CB energy prediction, Fig. 9(a) shows the feature importance for the random forest model, where the most important features are the presence of S as well as elemental percentages of W and Mo. Fig. 9(b) indicates a slightly different ranking for gradient boosting, with the percentage of W and Mo still holding significant weight. Fig. 9(c) illustrates that, unlike the other two models, extra trees distributes the importance among the features more evenly. This balanced feature importance could be one of the reasons why extra trees exhibited less error and generalizes well.

In Fig. 10(a–c), we see that the percentage of W and Mo had less impact on VB energy prediction, while the chalcogen

Table 3 5-Fold cross-validation MSE for conduction and valence band energy predictions using the extra trees model

Fold	CB MSE	VB MSE
Fold 1	0.0005	0.0008
Fold 2	0.0004	0.0011
Fold 3	0.0007	0.0012
Fold 4	0.0005	0.0007
Fold 5	0.0006	0.0007
Mean MSE	0.0005	0.0009
Std. Dev.	0.0001	0.0002

presence had considerably higher impact compared to the CB energy prediction. This pattern makes physical sense: the valence band maximum in the TMD alloys showed sensitivity to chalcogen substitution, but the transition metal composition had a greater influence on the conduction band. Thus, the model's learnt feature importance is consistent with existing theoretical understanding.

Fig. 9(d) and 10(d) present the MSE *vs.* number of trees for the extra trees model. We found that the test MSE stabilized after a certain number of trees, indicating that the model performance reaches its peak with an optimal number of trees. Based on this observation, we chose 100 trees as the optimal number for both CB and VB energy prediction.
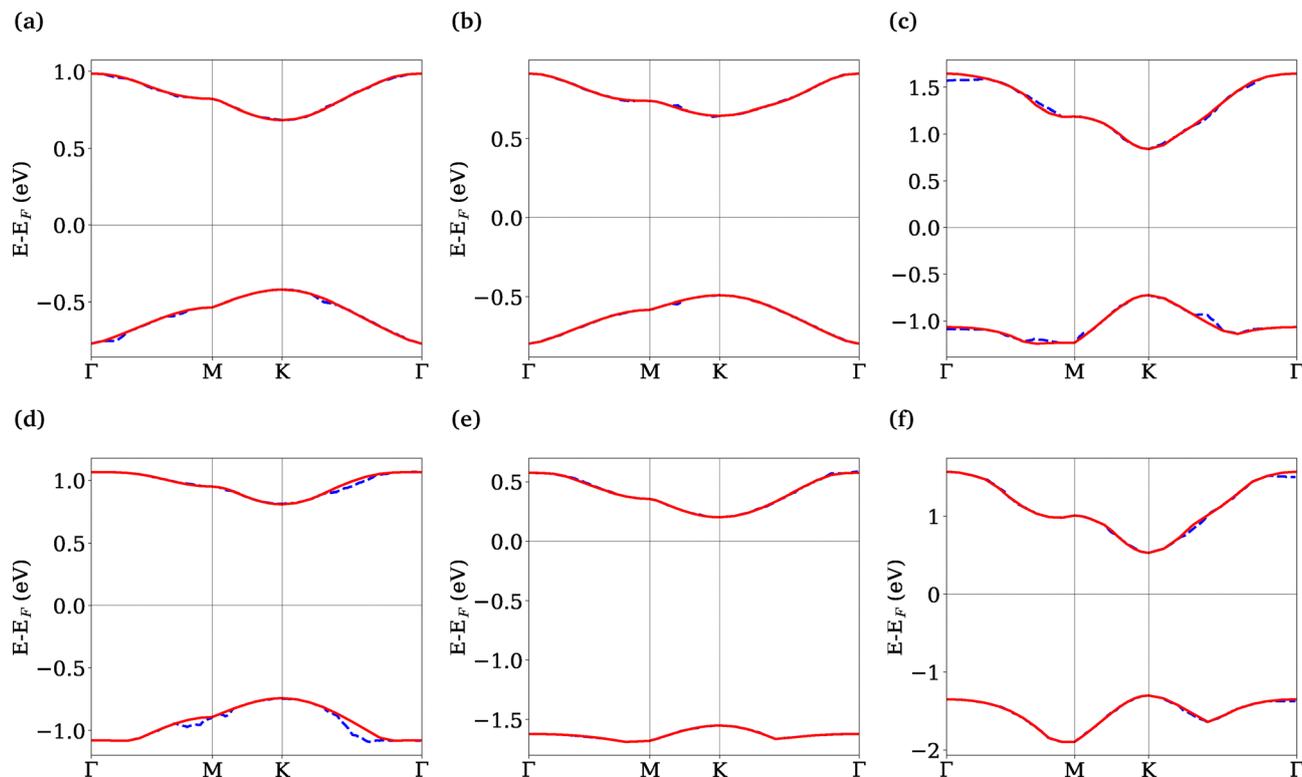
Fig. 8 Band structure comparison for six different atomic configurations predicted by the extra trees model alongside DFT simulations. The compositions $W_yMo_{1-y}X_2$ correspond to: (a) $y = 0.625$, X = Te; (b) $y = 0.1875$, X = Te; (c) $y = 1$, X = Se; (d) $y = 0.9375$, X = Se; (e) $y = 0.125$, X = S; (f) $y = 1$, X = S. Solid and dashed lines represent DFT ground truth and ML predictions, respectively.
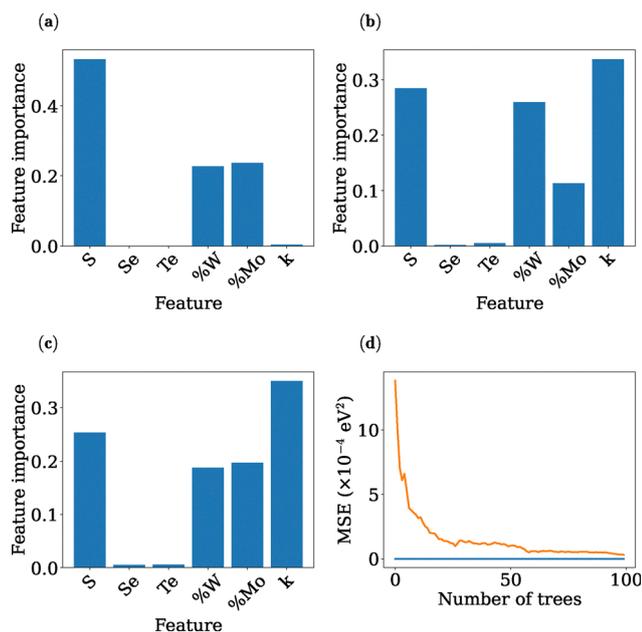

Fig. 9 Feature-importance analysis and error *versus* number of trees for conduction band prediction. Feature importance for (a) random forest, (b) gradient boosting, and (c) extra trees. (d) MSE (blue line for MSE of training data and orange line for MSE of test data) *versus* number of trees for extra trees.
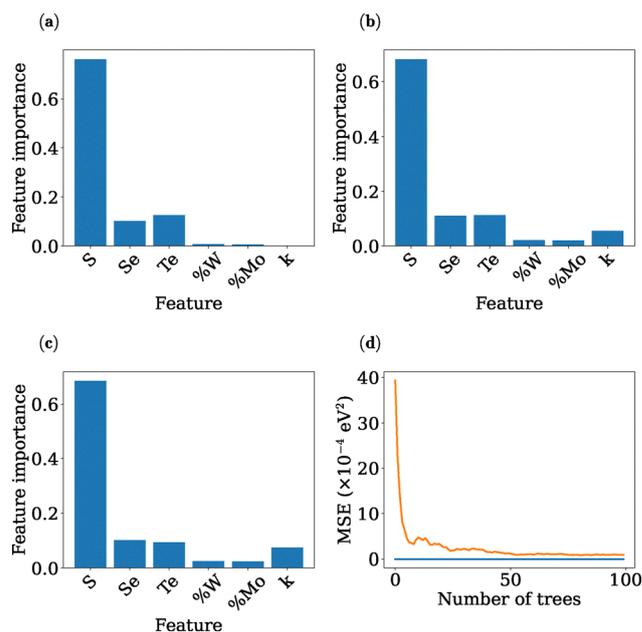

Fig. 10 Feature-importance analysis and error *versus* number of trees for valence band prediction. Feature importance for (a) random forest, (b) gradient boosting, and (c) extra trees. (d) MSE (blue line for MSE of training data and orange line for MSE of test data) *versus* number of trees for extra trees.
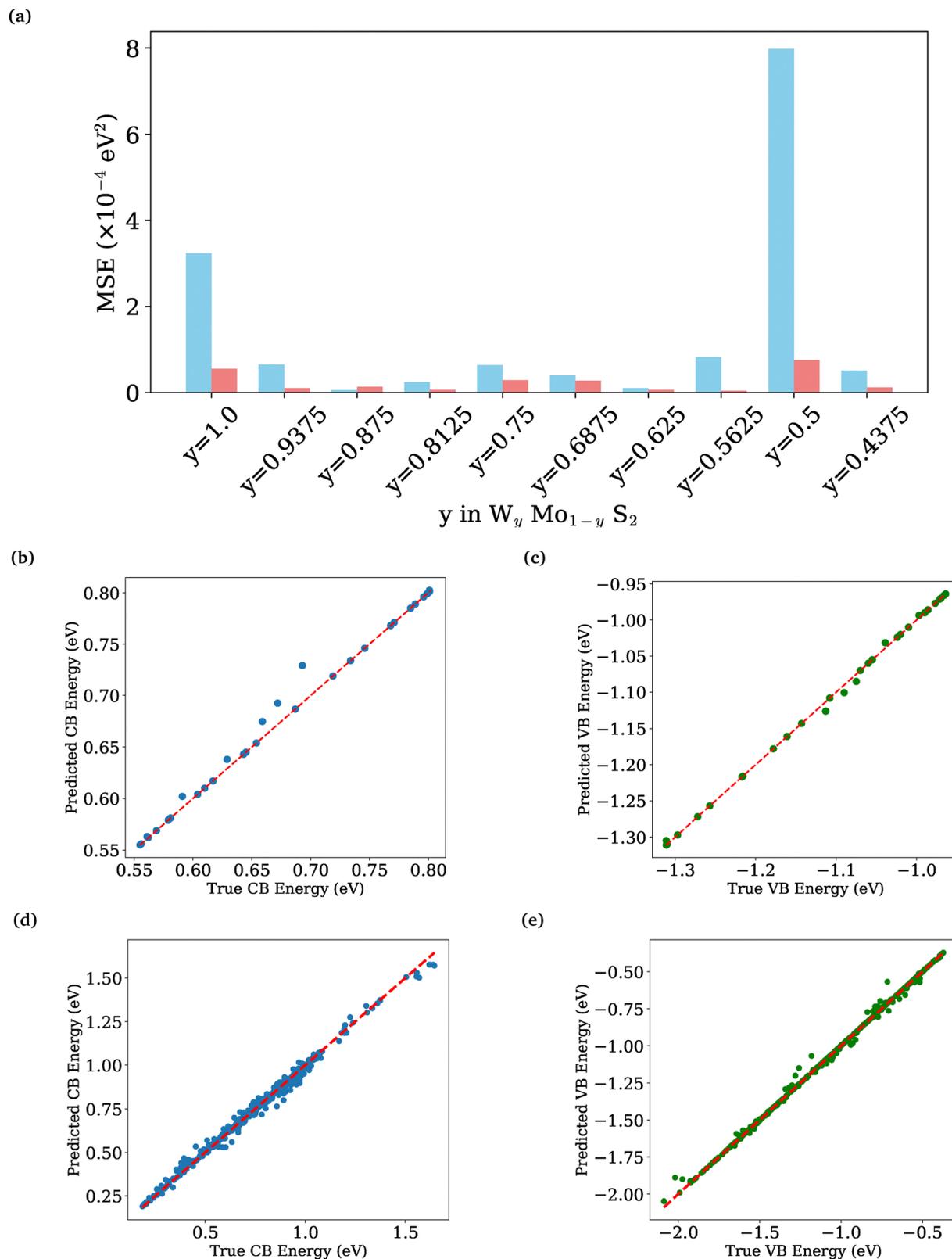
Fig. 11 (a) MSE for conduction and valence band energy predictions across 10 different alloy compositions represented by blue and orange bars, respectively. (b)–(e) Correlation plots between ML-predicted and DFT-simulated band energies using the extra trees model: (b) and (c) show conduction and valence bands for $W_{0.0625}Mo_{0.9375}Se_2$, while (d) and (e) illustrate the results for the entire test set. The near-unity correlations (represented by broken lines) demonstrate the high fidelity and predictive accuracy of the model.

3.7 Error and correlation analysis

To further evaluate the robustness of the extra trees model, we examined prediction errors as a function of alloy composition (Fig. 11(a)). While extra trees regression generally achieves very low errors across most alloys, certain compositions exhibited slightly higher MSE, likely due to the limited representation of those compositions in the training set, as we randomly selected samples for training and testing. Nevertheless, the error magnitude remained below 10^{-3} eV^2, indicating strong generalization.

MSE alone may not be enough to evaluate the performance of predicting band structure. For example, as shown in Table 2, ridge regression and random forest are close in terms of MSE, but in reality, ridge regression performed poorly compared to random forest in terms of capturing the curvature of the band structure, as illustrated in Fig. 6(a). This is where correlation metrics, such as Pearson correlation coefficient (eqn (10)), become valuable. The Pearson correlation coefficient allowed us to cross-check the similarity between the ground truth and predicted data distributions. Thus, it was essential to analyze the correlation between the ML model-predicted band structure and the ground truth band structure to properly assess the predictive fidelity of our model.

Firstly, correlation plots between ML-predicted and DFT-calculated band energies were generated for both the conduction and valence bands for a specific composition, $W_{0.0625}Mo_{0.9375}Se_2$ (Fig. 11(b) and (c)), and near-unity correlation was obtained. Then, correlation plots were generated for the whole test dataset (Fig. 11(d) and (e)). The extra trees model demonstrated near-perfect alignment with the DFT reference values across the test set. For conduction band energies, the Pearson correlation coefficient was $r = 0.9984$, while the valence band energies achieved $r = 0.9994$. These results implied that the ML framework did not merely interpolate, but effectively generalized the physical trends across different alloy compositions and k-points. Combined with the low mean squared errors reported in Table 2, the correlation analysis validated that the model can serve as a reliable alternative for first-principles calculations in band structure prediction.

3.8 Generalization across compositions beyond training data

To assess the robustness of the extra trees model, we conducted an experiment where two alloy compositions were entirely excluded from the training dataset. The model was trained on the remaining compositions and then tested on these unseen compositions.

The prediction errors for these unseen compositions, reported in Table 4, were slightly higher than those obtained when the model was trained on all compositions. However, these errors are still within the same order of magnitude as the test set errors reported earlier, confirming that the model retains its predictive capability even for unseen alloy configurations, as shown in Fig. 12.

This outcome highlights the ability of the extra trees regressor to capture the underlying fundamental physical relationships influencing the band structures in TMD alloys. It also demonstrates the model's ability to generalize to new

Table 4 Mean squared error and correlation values for TMD alloys beyond training data

	CB		VB	
Composition	MSE	Corr.	MSE	Corr.
$W_{0.4375}Mo_{0.5625}Te_2$	0.0005	0.9762	0.0036	0.9500
$W_{0.625}Mo_{0.375}Te_2$	0.0003	0.9974	0.0001	0.9965

compositions, a crucial feature for materials discovery, where new compositions are continually explored.

A rigorous generalization test was conducted by training the model exclusively on sulfide and selenide alloys and testing on telluride alloys. To enable this test, the original one hot encoding of chalcogen species was replaced with the atomic number of the chalcogen. This change was necessary because the model would never encounter Te during training under a one-hot representation and thus would fail to learn the dependence on telluride presence. Under this separated training-testing
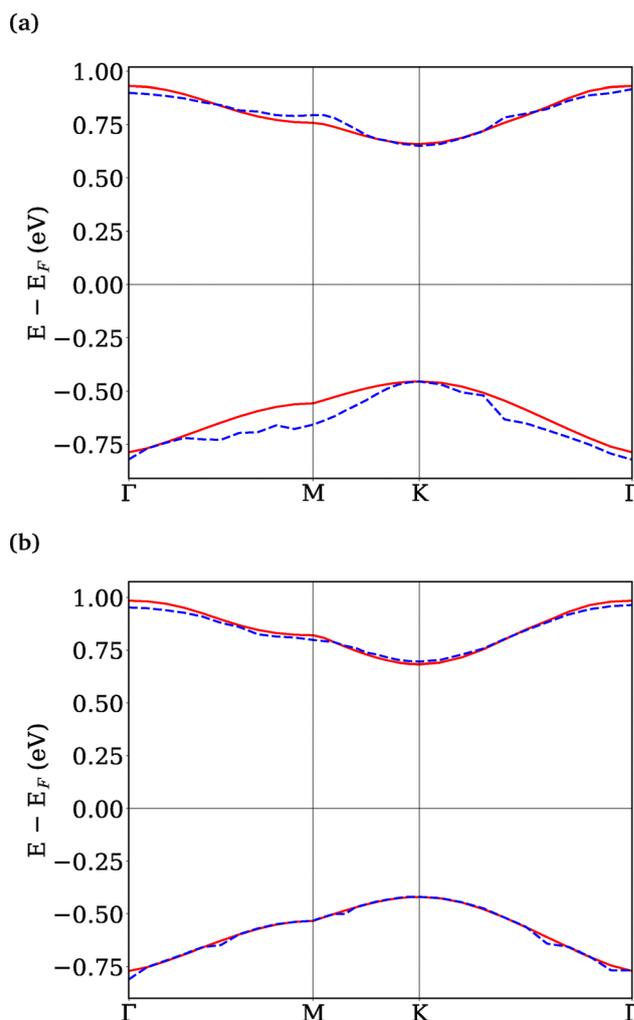

Fig. 12 ML-predicted and DFT-simulated band structures for the alloy compositions beyond training data. (a) $W_{0.4375}Mo_{0.5625}Te_2$ and (b) $W_{0.625}Mo_{0.375}Te_2$. The continuous and broken lines represent the ground truth and ML-predicted structures, respectively.
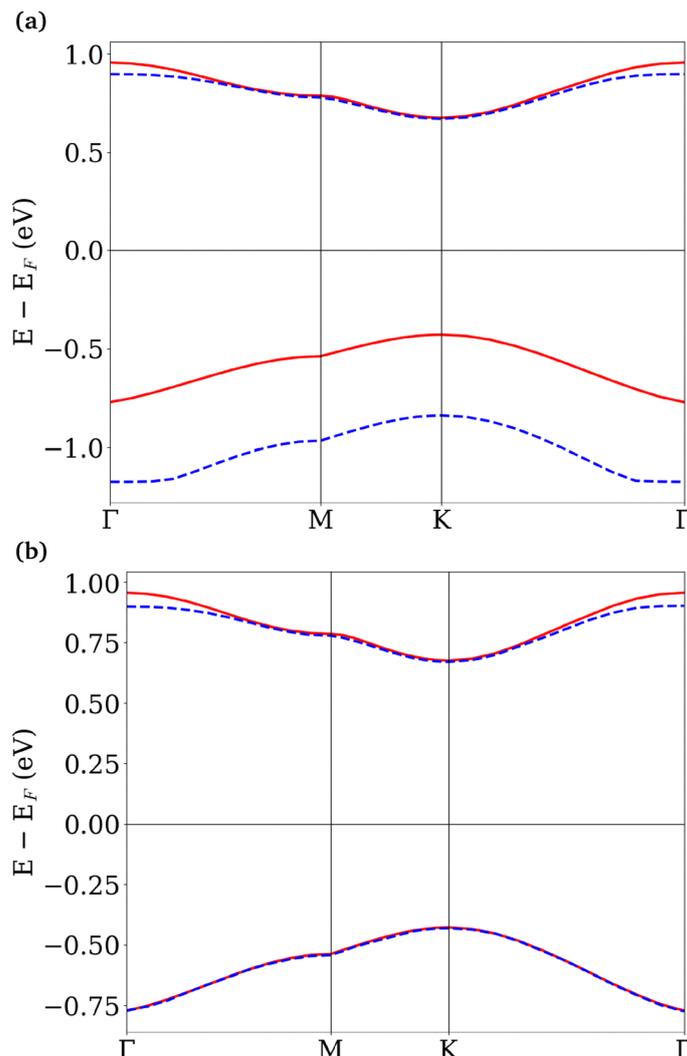
Fig. 13 Generalization performance of the model for telluride alloy $W_{0.5625}Mo_{0.4375}Te_2$. (a) Band-structure prediction when the model is trained exclusively on sulfide and selenide alloys and tested on telluride alloys. (b) Band-structure prediction when five telluride compositions are included in the training dataset and the remaining tellurides are used for testing. The continuous and broken lines represent ground truth and ML-predicted structures, respectively.

protocol, the model successfully approximated the overall band curvature, as reflected by high correlation coefficients (greater than 0.9). However, MSE is a lot higher, especially in VB prediction (0.004 for CB to 0.2 for VB), which is also evident in Fig. 13(a). This behavior highlights a limitation of data-driven models; accurate prediction for an unseen material class requires representative training data from that material. To mitigate this issue, we further performed a generalization test by including only five telluride compositions in the training dataset and evaluated the performance on the remaining tellurides. In this case, the MSE dropped significantly (to approximately 0.001 for CB and 0.005 for VB) compared to the previous protocol where there was no representation of telluride alloys, and the DFT-simulated and predicted band structures almost overlap, as shown in Fig. 13(b). Importantly, this outcome highlights that the model does not merely interpolate but learns structure–property relationships, but remains constrained by the coverage of the training dataset. In the context of new material

discovery, we can simulate a few number of compositions and use that for the prediction of properties for new alloys within the same composition regime.

3.9 Effective mass and band gap calculation from predicted band structure

The effective mass of charge carriers, electrons and holes, can be calculated from the curvature of the conduction and valence band structures using the parabolic band approximation. The effective mass is related to the second derivative of the energy (E) with respect to the wave vector (k) at the band extrema, following the expression:

$$m^* = \frac{\hbar^2}{\left(\frac{d^2 E}{dk^2}\right)} \tag{11}$$

where m^* is the effective mass, and $\hbar$ is the reduced Planck's constant. The effective mass provides insights into the mobility of

Table 5 Comparison of band gaps and effective masses calculated from DFT-simulated and extra trees (ET) predicted electronic band structures for selected TMD alloy compositions (band structures shown in Fig. 8)

Composition	Band gap (eV)			Effective mass of electrons (m_0)			Effective mass of holes (m_0)		
	DFT	ET	Error (%)	DFT	ET	Error (%)	DFT	ET	Error (%)
$W_{0.625}Mo_{0.375}Te_2$	1.1039	1.1046	0.06	2.2264	2.2935	3.01	2.7937	2.4343	12.87
$W_{0.1875}Mo_{0.8125}Te_2$	1.1330	1.1280	0.44	3.0021	2.5703	14.38	3.4482	3.3014	4.26
WSe_2	1.5641	1.5667	0.16	0.4709	0.5317	12.90	0.5818	0.6290	8.12
$W_{0.9375}Mo_{0.0625}Se_2$	1.5519	1.5569	0.32	1.7635	1.9443	10.25	2.1469	2.2822	6.30
$W_{0.125}Mo_{0.875}S_2$	1.7535	1.7543	0.05	2.0076	1.9961	0.57	2.4242	2.4396	0.63
WS_2	1.8281	1.8297	0.09	0.3721	0.4396	18.15	0.4992	0.5194	4.04

charge carriers, with larger effective masses referring to lower mobility.

To calculate the effective masses, the curvature of the conduction and valence bands was computed using a parabolic fit over a selected range of k-points around the band extrema. The effective masses were then derived from the second derivative of the energy with respect to k at the K point.

The band gap can be calculated by taking the difference between the conduction band minima and the valence band maxima. We validated the strength of our model by comparing the band gap and effective mass calculated from the predicted band structure and the true band structure (DFT simulated).

The calculated effective masses for electrons and holes and band gaps from both ML-predicted and DFT-simulated band structures for 6 different structures of Fig. 8 are shown in Table 5. The results illustrated that from the predicted band structure, we can derive the secondary properties (band gap and effective mass) with reasonable accuracy. This validated the model's reliability in capturing the essential features of the band structure precisely. See the SI Section S3 for the comparison results of other TMD alloys.

4 Conclusions

In this paper, DFT simulations were conducted to analyze monolayer TMD alloys with varying compositions of W, Mo, and chalcogen atoms (S/Se/Te). Conduction and valence band energies were computed for each alloy at multiple k points. No study has reported such a comprehensive analysis of band structures yet. To enable ML model training, compositional descriptors were derived for 51 different alloys. Representing the atomic configuration as a percentage of W and Mo, chalcogens provide a simple yet generalized feature representation framework for similar types of materials. The model trained on this feature set can predict the band energies at any k point of the TMD alloys. A thorough investigation of different ML models, ranging from linear regression to decision tree-based ensembling, was performed to identify the most accurate model for conduction and valence band structure estimation. The extra trees algorithm emerged as the best performing model, achieving low MSE and high correlation with respect to DFT calculated values for both the conduction (MSE = 0.0005, correlation = 0.9984) and valence band (MSE = 0.0009, correlation = 0.9994), showcasing its excellent predictive power.

Subsequently, this model was applied to predict the CB and VB of alloys beyond the training dataset, and minimal errors validated the robustness of the model. This work is the first demonstration of an ML model capable of predicting the whole band structure of any material given just structural features. Unlike conventional band gap-only prediction models, our model provides richer insights as we can extract the band gaps by the CB minima and VB maxima, and also the effective mass by band curvature fitting. Good agreement between DFT simulated values and predicted values of the extra trees model implied that by deploying our model, one can save a great deal of computational resources. Both the CB and VB models took ≈ 1 second to train, and the inference time for predicting the CB and VB of all 51 compositions is ≈ 0.22 s, while DFT took $\gg 1$ h to calculate only the self-consistent field calculation of a single composition.

Nonetheless, the accuracy of the prediction of electronic structures is limited by the learning capacity of the algorithm and the inherent limitations of the density functional theory approximations used to generate the training dataset. Despite this limitation, our model achieved a favorable balance between accuracy and computational efficiency. This overall framework can be extended to predict band structures beyond the conduction and valence band, across diverse classes of materials, and to other optical and electronic property prediction tasks.

Author contributions

Tarvir Anjum Aditto: formal analysis, methodology, visualization, data curation, investigation, validation, and writing – original draft. Vivek Chowdhury: formal analysis, methodology, visualization, software, investigation, validation, and writing – original draft. Hafiz Imtiaz: conceptualization, methodology, visualization, resources, funding acquisition, writing – original draft, writing – review and editing, and supervision. Ahmed Zubair: conceptualization, methodology, visualization, resources, funding acquisition, writing – original draft, writing – review and editing, and supervision.

Conflicts of interest

There are no conflicts to declare.

Data availability

Data for this article are available at GitHub at https://github.com/ahmedzubair003/ML_Based_Bands_Prediction.git.

Supplementary information: full ML models and additional results are provided. See DOI: https://doi.org/10.1039/d5ma01485a.

Acknowledgements

References

1 C. Patra, S. Mondal, R. Mukherjee and Y. Nandakishora, *ACS Mater. Au*, 2025, **5**, 745–766.

2 J. Xu, G. Shao, X. Tang, F. Lv, H. Xiang, C. Jing, S. Liu, S. Dai, Y. Li and J. Luo, *et al.*, *Nat. Commun.*, 2022, **13**, 2193.

3 J. Wang, W. Zeng and Q. Zhou, *Front. Chem.*, 2022, **10**, 950974.

4 K. Mahmud, T. Yashir and A. Zubair, *Nanoscale Adv.*, 2024, **6**, 2447–2458.

5 Q. H. Wang, K. Kalantar-Zadeh, A. Kis, J. N. Coleman and M. S. Strano, *Nat. Nanotechnol.*, 2012, **7**, 699–712.

6 K. Yang, B. Li, Z. Li, Y. Wang, H. Chen, C. Wang, C. Wu, F. Guo and G. Zhang, *Mater. Res. Bull.*, 2025, 113490.

7 H. S. Kang, J. H. Kang, S. Lee, K. Lee, D. H. Koo, Y.-S. Kim, Y. J. Hong, Y.-J. Kim, K. Kim and D. Lee, *et al.*, *NPG Asia Mater.*, 2022, **14**, 90.

8 J. Yu, S. Wu, X. Zhao, Z. Li, X. Yang, Q. Shen, M. Lu, X. Xie, D. Zhan and J. Yan, *Nanomaterials*, 2023, **13**, 2843.

9 T. Mueller and E. Malic, *npj 2D Mater. Appl.*, 2018, **2**, 29.

10 A. Chaves, J. G. Azadani, H. Alsalman, D. Da Costa, R. Frisenda, A. Chaves, S. H. Song, Y. D. Kim, D. He and J. Zhou, *et al.*, *npj 2D Mater. Appl.*, 2020, **4**, 29.

11 S. Susarla, A. Kutana, J. A. Hachtel, V. Kochat, A. Apte, R. Vajtai, J. C. Idrobo, B. I. Yakobson, C. S. Tiwary and P. M. Ajayan, *Adv. Mater.*, 2017, **29**, 1702457.

12 S. E. Muller, M. P. Prange, Z. Lu, W. S. Rosenthal and J. A. Bilbrey, *Sci. Data*, 2023, **10**, 336.

13 X. Jiang, G. Liu, L. Zhang and Z. Hu, *Catalysts*, 2025, **15**, 309.

14 Y. Zhang, W. Xu, G. Liu, Z. Zhang, J. Zhu and M. Li, *PLoS One*, 2021, **16**, e0255637.

15 Y. Yuan, J. Ren, H. Xue, J. Li, F. Tang, P. La and X. Lu, *ACS Appl. Mater. Interfaces*, 2023, **15**, 12462–12472.

16 T. H. Nguyen, L. H. Nguyen and T. N. Truong, *ACS Omega*, 2022, **7**, 22879–22888.

17 Y.-H. Liu, S. Zhang, P. Zhang, T.-K. Lee and G.-W. Chern, *Phys. Rev. B*, 2022, **106**, 035131.

18 D. Liu, *Mater. Data Energy*, 2024, **9**, 4025947.

19 Z. Bu, Y. Xue, X. Zhao, G. Liu, Y. An, H. Zhou and J. Chen, *ACS Appl. Mater. Interfaces*, 2024, **16**, 60458–60471.

20 E. T. Chenebuah, *Mater. Data Energy*, 2021, **3**, 1004542.

21 T. Wang, X. Tan, Y. Wei and H. Jin, *Nanoscale*, 2022, **14**, 2511–2520.

22 C. Xu, X. Deng and P. Yu, *ACS Appl. Mater. Interfaces*, 2024, **16**, 55970.

23 C. Gao, X. Yang, M. Jiang, L. Chen, Z. Chen and C. V. Singh, *Phys. Chem. Chem. Phys.*, 2022, **24**, 4653–4665.

24 P. Kumar, V. Sharma, S. N. Shirodkar and P. Dev, *Phys. Rev. Mater.*, 2022, **6**, 094007.

25 X. Jia, Y. Zhang, J. Liu and H. Wang, *J. Mater. Chem. A*, 2025, **13**, 3457–3468.

26 P. Giannozzi, S. Baroni, N. Bonini, M. Calandra, R. Car, C. Cavazzoni, D. Ceresoli, G. L. Chiarotti, M. Cococcioni and I. Dabo, *et al.*, *J. Phys.: Condens. Matter*, 2009, **21**, 395502.

27 V. Chowdhury and A. Zubair, *arXiv*, 2025, preprint, arXiv:2511.06069, DOI: 10.48550/arXiv.2511.06069.

28 K. S. Hoque and A. Zubair, *ACS Omega*, 2022, **7**, 36184–36194.

29 D. C. Liu and J. Nocedal, *Math. Prog.*, 1989, **45**, 503–528.

30 A. Seko, A. Togo and I. Tanaka, *Nanoinformatics*, Springer, Singapore, 2018, pp. 3–23.

31 G. W. Kim, M. Lee, J. Bae, J. Han, S. Park and W. Shim, *Nano Convergence*, 2024, **11**, 8.

32 C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer, New York, 2006.

33 T. Hastie, R. Tibshirani and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer, 2nd edn, 2009.

34 H. Blockeel, *Front. Artif. Intell.*, 2023, **6**, 1124553.

35 L. Breiman, *Mach. Learn.*, 2001, **45**, 5–32.

36 P. Geurts, D. Ernst and L. Wehenkel, *Mach. Learn.*, 2006, **63**, 3–42.

37 J. H. Friedman, *Ann. Stat.*, 2001, **29**, 1189–1232.

38 F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and É. Duchesnay, *J. Mach. Learn. Res.*, 2011, **12**, 2825–2830.