OPEN ACCESS
IOP Publishing
Journal of Physics D: Applied Physics
J. Phys. D: Appl. Phys. **58** (2025) 073005 (14pp)
https://doi.org/10.1088/1361-6463/ad9933

Topical Review

Recent advances in machine learning and deep learning-enabled studies on transition metal dichalcogenides

Shivani Bhawsar[1] **and Eui-Hyeok Yang**[1,2,*]

[1] Department of Mechanical Engineering, Stevens Institute of Technology, Hoboken, NJ, United States of America
[2] Center for Quantum Science and Engineering, Stevens Institute of Technology, Hoboken, NJ, United States of America

E-mail: eyang@stevens.edu

Received 8 July 2024, revised 15 August 2024
Accepted for publication 2 December 2024
Published 11 December 2024

Abstract

The machine learning and deep learning (ML/DL) techniques have significantly advanced the understanding and utilization of transition metal dichalcogenides (TMDs) by enabling efficient analysis, prediction, and optimization of their properties. ML/DL methods permit rapid screening, optimization and analysis of two-dimensional (2D) material candidates, potentially accelerating the discovery and development of TMDs with desired electronic, optoelectronic, and energy storage properties. This review provides a comprehensive review of ML/DL methods to enhance 2D materials research via the optimization of synthesis conditions, interpretation of complex data sets, and the use of generative adversarial networks and variational autoencoders for innovative material design and image processing tasks. Furthermore, it highlights the potential of ML/DL techniques in predicting and tailoring the electronic, optical, and mechanical properties of 2D materials to meet specific application requirements.

Keywords: machine learning, transition metal dichalcogenides, deep learning, characterization, generative AI, properties, 2D materials

1. Introduction

Machine learning and deep learning (ML/DL) methods are revolutionizing material science by providing powerful methods for analyzing and predicting the properties and behaviors of materials. These computational techniques involve training algorithms on large datasets to recognize patterns, make predictions, and optimize processes, thus accelerating research and development in the field. The ML/DL techniques utilize a sub-domain of artificial intelligence that can recognize patterns in the data, enabling an efficient identification and characterization of materials [1–3]. ML and DL models can analyze vast amounts of experimental and computational data to identify relationships and trends that are not easily discernible through traditional methods. This capability allows researchers to more efficiently discover new materials with specific desired properties. To this end, ML/DL methods have been used in studying transition metal dichalcogenides (TMDs), a class of atomically thin two-dimensional (2D) materials composed of a transition metal and a chalcogen, characterized by their layered structure and unique properties. Typical characterization and data collection are performed by analyzing spectroscopic measurements and optical images

* Author to whom any correspondence should be addressed.

of individual TMD flakes. ML/DL methods can enhance 2D materials research by optimizing synthesis conditions and interpreting complex data sets using generative adversarial networks (GANs) and variational autoencoders (VAEs) [4, 5]. Furthermore, these techniques enable high-throughput screening of 2D material properties, significantly speeding up the discovery process for materials with optimal characteristics [6, 7]. The integration of ML/DL with computational simulations and experimental data can transform the landscape of 2D material research, enabling the rapid development of materials for applications in electronics [8, 9], optoelectronics [10], and energy storage [11, 12].

Several ML/DL studies have identified correlations between design and optical characteristics and properties of TMDs, including analysis of physical [2, 9], optical (thickness identification [3, 13], detection of materials [14, 15], segmentation [14, 16]), electronic (bandgap, electron affinity) [8, 9, 17] properties, and point defects [18–22]. ML/DL models can optimize the conditions for material synthesis, such as temperature [7], pressure [23, 24], and chemical composition [6], to achieve the best performance. This optimization reduces the need for costly and time-consuming trial-and-error experiments. Recently, generative AI has utilized cutting-edge methods, such as GANs and VAEs in material science. These advanced ML methods can expedite the discovery of novel materials by generating novel chemical structures and compositions, creating synthetic data to augment small datasets, and improving the training of other ML models. Additionally, image-to-image translation [25], using conditional GANs (cGANs), has received notable attention in various applications, including translating optically sectioned structured illumination microscopy (SIM) images, semantic segmentation [14, 16], and image processing [25, 26].

Here, we review the recent advancements in ML/DL strategies applied to predicting and optimizing TMD properties. We delve into data processing methods and modeling techniques to characterize TMDs, encompassing their physical, electronic, and optical properties. We also discuss the inherent limitations and challenges facing ML/DL methods in the context of TMD research, such as the need for large and diverse datasets, the interpretability of complex models, and computational resources for training. Despite these challenges, ML/DL offers promising avenues for accelerating material discovery and optimization processes. Future advancements in algorithmic design, data availability, and interdisciplinary collaboration are pivotal for overcoming these obstacles and further harnessing the potential of ML/DL in advancing TMD research.

2. Basic principles and procedures

Figure 1 shows the workflow of synthesis, characterization, and predictive modeling of properties of TMDs. The ML/DL techniques require attaining initial data to train models and predict pertinent properties of 2D materials. Before training, selecting and processing the data is essential to ensure compatibility with the model.

2.1. Database and preprocessing

The ML/DL studies on 2D materials require sufficient initial data to predict material properties Several standard material-based databases have been established to explore various ML and DL methods in material science [27, 28]. The computational database of 2D materials contains structural, thermodynamic, elastic, electronic, magnetic, and optical properties of around four thousand 2D materials distributed over 40 different crystal structures [27, 29]. The inorganic crystal structure database is the world's largest, containing more than 299 000 fully identified inorganic crystal structures [30, 31]. The quantum point defects in the 2D materials database consist of more than 1900 defect systems comprising various charge states of 503 intrinsic point defects (vacancies and antisites) in 82 different 2D semiconductors and insulators [20]. Monolayer TMDs contain calculated structural and electronic properties of various 2D materials. Table 1 shows a list of open-source databases of 2D materials. These datasets encompass optical images, physical and chemical properties, structural details, temperature measurements, scanning electron microscopy (SEM) images, and spectroscopy data. Existing databases, while carefully curated, still require additional preprocessing and preparation such as data cleaning, normalization, feature selection, oversampling, undersampling, and data augmentation to ensure their suitability for specific tasks in ML and DL.

The data cleaning involves removing inconsistencies, errors, or missing values from the dataset. Normalizing the data involves scaling the features to a similar range, typically between 0 and 1 or -1, to ensure that all features contribute equally to the model training process and prevent features with larger scales from dominating the learning process. In some cases, the dataset may contain a large number of features, some of which may not be relevant to or informative for the ML/DL task. Therefore, feature selection involves choosing the most relevant features for the task at hand, while feature extraction involves transforming the original features into a new set of features that better represent the underlying patterns in the data. In datasets where one class is significantly more prevalent than others, imbalanced data can lead to biased model performance. Techniques such as oversampling, undersampling, or generating synthetic samples balance the dataset and improve model performance. Data augmentation involves generating additional training examples by applying transformations such as rotation, scaling, or flipping to the existing data. This technique helps increase the diversity of the training data and improve the generalization ability of the ML/DL models. The dataset is typically divided into training, validation, and test sets. The training set is used to train the ML/DL models, the validation set is used to tune hyperparameters and evaluate model performance during training,

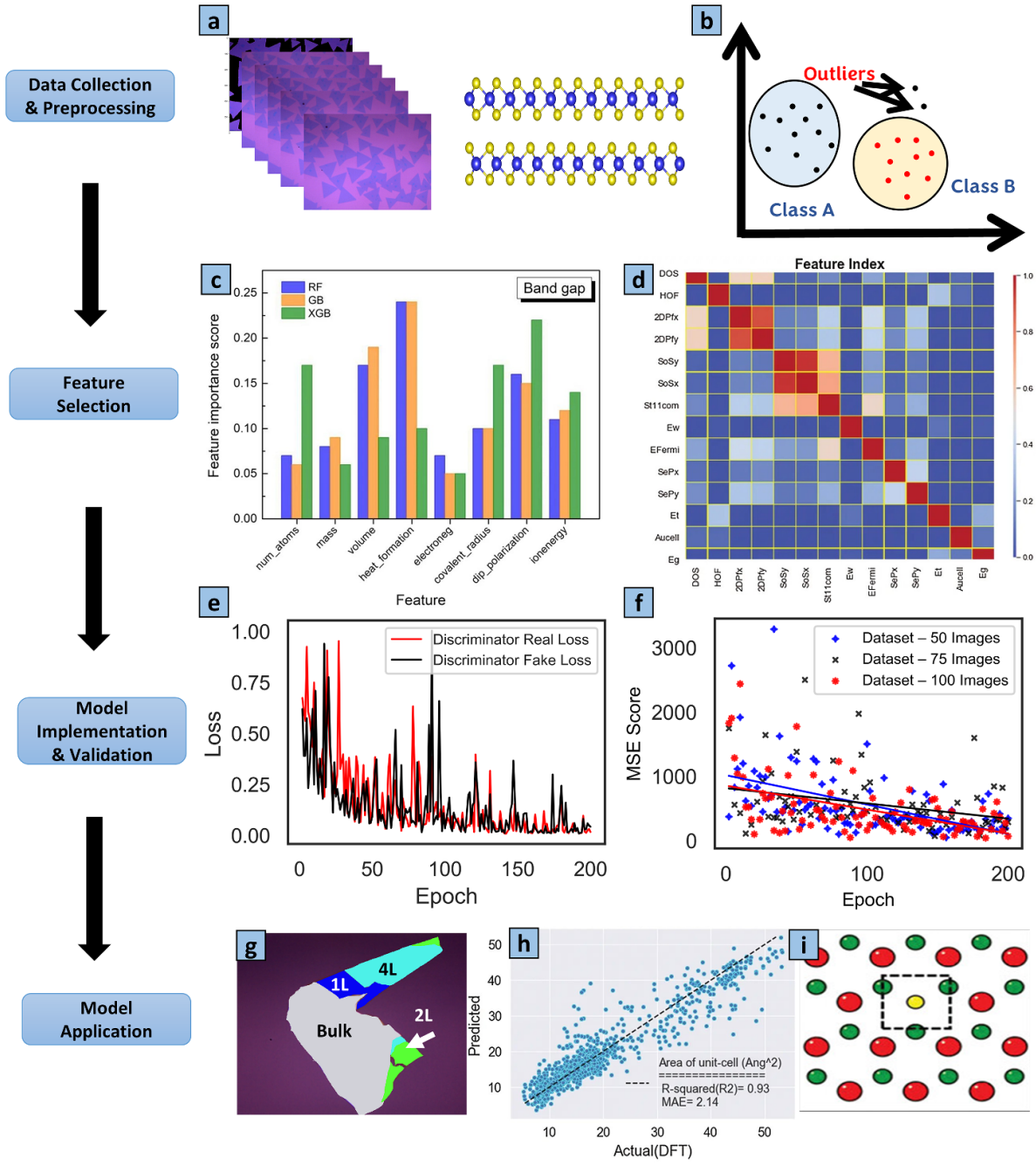

Figure 1. Workflow of data collection and preprocessing, feature selection, model implementation and validation, and model application (a) optical images of TMDs flakes and crystal structure (b) outlier detection (c) feature importance score. Adapted from [32]. CC BY 4.0. (d) Heat map of the Pearson correlation coefficient matrix. Adapted from [9]. CC BY 4.0. (e) Loss curves depicting the training progression of the discriminator in a Pix2Pix model (f) evaluation metrics to compare the generated result with the ground truth using the mean square error (MSE). (g) Model generated results for monolayers, bilayers, four layers, and bulk MoS_2. Adapted from [33]. CC BY 4.0. (h) Predicted vs. actual values for the regression gradient boosted model. Adapted from [9]. CC BY 4.0. (i) Atomic sites prediction from the annular dark-field scanning transmission electron microscopy (ADF STEM) image. Adapted from [21]. CC BY 4.0.

and the test set is used to evaluate the final performance of the trained models. For datasets containing optical images or SEM images of 2D materials, preprocessing may involve tasks such as resizing, cropping, and converting images to a standardized format (e.g., grayscale or RGB) suitable for input to the ML/DL models.

2.2. Data analysis

Data analysis is crucial for uncovering underlying patterns and relationships that influence TMDs' properties, aiding in the optimization and prediction. Pearson correlation [39] method is a statistical method used to measure the strength

Table 1. List of open-source databases of 2D materials.

Database	Web URL	Name	Description	References
C2DB	https://cmr.fysik.dtu.dk/c2db/c2db.html	Computational 2D materials database	4000 2D materials distributed over 40 different crystal structures	[27, 29]
ICSD	https://icsd.products.fiz-karlsruhe.de/	Inorganic crystal structure database	299 000 inorganic crystal structures	[30, 31]
IMP2D	https://cmr.fysik.dtu.dk/qpod/qpod.html	Quantum point defects in 2D materials (QPOD) database	500 distinct intrinsic point defects in 2D materials	[34]
MC2D	www.materialscloud.org/discover/mc2d/dashboard/ptable	Materials cloud two-dimensional crystals database	258 two-dimensional crystal structures, exfoliated from three-dimensional experimental crystal structures	[35, 36]
2DMatPedia	http://2dmatpedia.org/about	2D materials encyclopedia	2940 were obtained by exfoliation of existing layered materials ('top–down' approach) and 3409 were obtained by chemical substitution from 2D materials ('bottom–up' approach)	[37]
V2DB	https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/SNCZF4	Virtual 2D materials database	316 505 stable materials with AI predicted key properties (energy, electronic, and magnetic)	[38]
QPOD	https://cmr.fysik.dtu.dk/qpod/qpod.html	Quantum point defects in 2D materials (QPOD) database	500 distinct intrinsic point defects in 2D materials	[20]

and direction of the linear relationship between two variables. In the context of 2D materials, this technique can be applied to assess the correlation between different material properties, including electrical conductivity and thermal stability

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}} \quad (1)$$

where x_i and y_i represent the values of the two variables, and $\bar{x}$ and $\bar{y}$ represent the mean of the values of the variables. The Pearson correlation coefficient ranges from -1 to 1, where 1 indicates a perfect positive linear relationship, -1 indicates a perfect negative linear relationship, and 0 indicates no linear relationship.

Principal component analysis [40] (PCA) is a dimensionality reduction technique that simplifies the complexity in high-dimensional datasets while preserving as much variance as possible. For 2D materials, PCA can reduce the number of variables by transforming them into a set of principal components, which helps visualize the data, detect patterns, and identify the most significant features that contribute to material properties.

Cluster analysis is a method of grouping similar data points into clusters based on their features. In TMD studies, cluster analysis can be used to categorize materials with similar properties or behaviors, which can help understand how different materials relate to each other and identify distinct material classes that may exhibit unique characteristics. Regression analysis involves modeling the relationship between a dependent variable and one or more independent variables. Heatmaps use color coding to represent the strength of correlations between pairs of variables, facilitating quick identification of strong or weak relationships.

2.3. Data modeling

Once the data has been collected and preprocessed, the next step in the workflow is data modeling. The choice of model depends on the specific task and dataset characteristics. For predicting material properties of 2D materials, various models such as regression, classification, clustering, and neural networks can be employed. Regression models may be suitable for predicting continuous properties like bandgap, while classification models like support vector machines or neural networks can classify materials based on properties. Support vector machines [41] are widely used in classification and regression tasks and operate by transforming the input data into a high-dimensional space and finding the hyperplane that best separates the classes while maximizing the margin between them. This margin represents the distance between the hyperplane and the nearest data points, known as support vectors, from each class. K-nearest neighbors [42] algorithm classifies or predicts properties of materials based on the majority vote or averaging of the K nearest data points in the feature space. In contrast, K-means clustering assigns each data point to the cluster with the nearest centroid, effectively partitioning the dataset into distinct groups [43]. Figure 2(b)

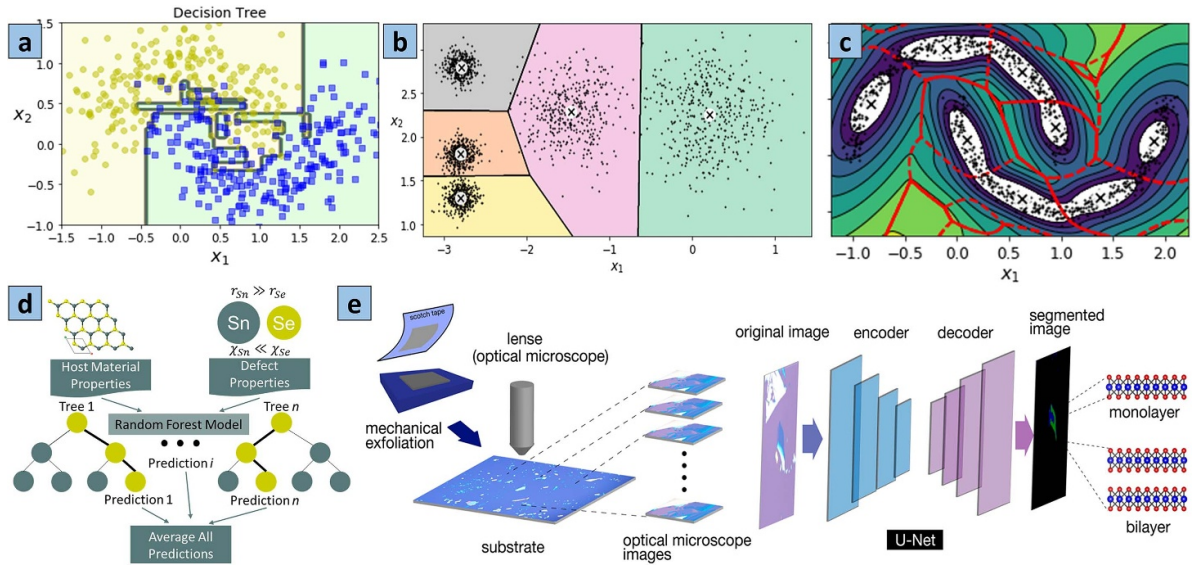

Figure 2. Schematics of ML/DL models (a) decision trees (hierarchical decision boundaries) (b) K-means clustering (centroid-based clustering of data points) (c) Bayesian Gaussian mixture (probabilistic clustering with Gaussian distributions) (d) random forests (predicting output by averaging host material and defect properties inputs). Reprinted (adapted) with permission from [18]. Copyright (2020) American Chemical Society. (e) Convolutional neural networks (prediction of monolayer and bilayer in MoS_2). Adapted from [60]. CC BY 4.0.

illustrates the classification of five clusters using K-means clustering. Decision trees (DT) are powerful tools in material science for classification, regression, and pattern recognition [44]. DT partition the feature space into regions based on the values of input features, effectively creating a hierarchical structure of decision rules. This structure enables interpretation of the relationships between material descriptors and properties. Figure 2(a) illustrates the classification of two classes using a DT classifier. Random forests (RFs), an ensemble learning method, operate by constructing a multitude of DT during training and outputs the class, that is, the mode of the classes (classification) or mean prediction (regression) of the individual trees [45]. RFs mitigate overfitting by training each tree on a random subset of the data and a random subset of the features. This ensemble approach enhances predictive accuracy and provides insights into feature importance, aiding in interpreting complex material datasets. Figure 2(d) illustrates the prediction of defects using RFs, utilizing host material and defect properties as input.

Gradient-boosted trees (GB) enhance model performance compared to DTs and RFs by sequentially focusing on the errors of previous trees, thereby reducing residuals and improving accuracy. Unlike DTs, which operate independently, GB iteratively refines predictions and performs intrinsic feature selection, leading to better classification and regression results while optimizing computational efficiency and data quality. Bayesian Gaussian mixture models (BGMMs) [46] offer a sophisticated approach to uncovering intricate patterns and clusters within complex datasets. By seamlessly integrating Bayesian inference with Gaussian mixture modeling, BGMM allows for automatically determining optimal cluster configurations and quantifying uncertainty in cluster assignments. This capability is particularly valuable in 2D material applications, where datasets often contain diverse and multifaceted information about material properties, phases, and compositions. BGMM enables researchers to identify distinct material phases, classify materials based on their properties, and explore the underlying structure of materials datasets with unprecedented depth and precision. Figure 2(c) illustrates the detection of 8 clusters using BGMM. Convolutional neural networks (CNNs) [47] enable advanced analysis of complex data such as microscopy images, spectroscopic data, and crystal structures, which are used for tasks, including image classification, object detection, and image segmentation. For instance, CNNs can identify and classify different phases or defects in microscopy images of 2D materials, facilitating materials characterization and defect analysis. Similarly, CNNs can analyze spectral data to predict material properties or classify materials based on their chemical composition. Figure 2(e) demonstrates the prediction of monolayer and bilayer in MoS_2 using CNN. Recently, generative models offer several advantages in material science, including the ability to generate novel and diverse materials candidates, accelerate materials discovery processes, and augment limited experimental datasets with synthetic data. One popular type of generative model is the GAN [48], which consists of two neural networks—a generator to generate synthetic data samples and a discriminator to distinguish between real and fake samples. Through iterative training, the generator learns to produce increasingly realistic samples while the discriminator becomes more adept at distinguishing real from fake. GANs have been applied in material science for generating novel molecular structures, designing new materials with desired properties, and generating synthetic microscopy images for training and data augmentation purposes. Another type of generative model

Table 2. List of ML/DL models along with their applications.

ML/DL model	Category	Application	References
Random forests and k-nearest neighbor	Classification	Classify 2D materials as monolayer MoS_2 and non-monolayer MoS_2	[6]
Decision trees, gradient-boosted decision trees, and random forests	Identification	Identify mechanically exfoliated 2D materials	[2]
DL-based generative model combined with a random-forest model	Discovery	Discover new 2D materials	[49]
DL-based graph convolutional neural networks	Discovery	Discover HER catalysts from a 2D database	[50]
Random forest (RF), gradient-boosted decision tree (GBDT), support vector regression (SVR), and multilayer perceptron (MLP)	Determination of electronic property	Predict the band gap of 2D materials	[17]
Random forest (RF), gradient boosting (GB), extreme gradient boosting (XGB)	Determination of electronic property	Predict the band gap and work function of 2D materials	[32]
Gradient-boosted (GB) algorithm	Determination of electronic property	Predict electronic band gap (E_g), work function, total energy, unit-cell area, Fermi level and density of states	[9]
DL-based convolutional neural network	Determination of electronic property	Identify 2D materials with flat electronic bands	[8]
Deep transfer learning and first-principles calculations	Determination of material defects	Predict properties of point defects in 2D materials and generate descriptions of defect structures	[18]
DL-based fully convolutional neural network	Determination of material defects	Detect and classify defects and dopants in TMDs	[21]
Monte Carlo tree search (MCTS)	Determination of material defects	Identify optimal defect arrangements in 2D materials	[22]
DL-based encoder–decoder semantic segmentation network	Layer number classification	Characterize and extract deep graphical features such as color, contrast, edges, shapes, and flake sizes	[51]
3D convolutional neural network	Layer number classification	Identify and segment MoS_2 flakes with mono, bi, tri and multilayers	[3]
DenseNet, U-Net, and mask-region convolutional neural network	Layer number classification	Classify, segment, and detect microscopic images of 2D materials for automated atomic layer mapping	[14]

is the VAE [4], which captures a latent representation of the input data and generates new samples by sampling from this latent space. VAEs have been used for materials discovery to generate new molecular structures or predict material properties based on learned latent representations. Table 2 shows a list of ML/DL models along with their applications in material science.

2.4. Data validation

Evaluation metrics are crucial for evaluating the performance of ML and DL models. These metrics provide insights into the performance of the model and help researchers make informed decisions about model selection, tuning, and deployment. Accuracy [52] can be calculated as the ratio of correct predictions to the total number of predictions. Precision [53] measures the proportion of true positive predictions out of all positive predictions, indicating the model's ability to avoid false positives and is calculated as

$$\text{Precision} = \text{TP}/(\text{TP}+\text{FP}) \tag{2}$$

where TP is the number of true positives and FP is the number of false positives. Recall [54] measures the proportion of true positive predictions out of all actual positive instances. It indicates the model's ability to capture all positive instances and is calculated as

$$\text{Recall} = \text{TP}/(\text{TP}+\text{FN}) \tag{3}$$

where FN is the number of false negatives. The F1-score is the harmonic mean of precision and recall, providing a balanced measure of a model's performance [55]

$$\text{F1}-\text{score} = 2*(\text{precision}*\text{recall})/(\text{precision}+\text{recall}). \tag{4}$$

The receiver operating characteristic (ROC) curve plots the true positive rate against the false positive rate at various threshold settings [56]. The area under the curve (AUC) summarizes the ROC curve, representing the model's ability

to distinguish between classes. AUC values closer to 1 indicate better performance. A confusion matrix provides a detailed breakdown of model predictions versus actual labels, showing the number of true positives, true negatives, false positives, and false negatives. Mean squared error (MSE) [57] measures the average squared difference between predicted and actual values, calculated as the average squared differences between predicted and actual values. Root MSE [58] is the square root of MSE, measuring the average deviation between predicted and actual values in the original units. Mean absolute error measures the average absolute difference between predicted and actual values, providing a more interpretable measure of model performance than MSE. R-squared (R^2) [59] represents the proportion of the variance in the dependent variable explained by the independent variables. It ranges from 0 to 1, with higher values indicating a better model fit to the data.

3. Identification and determination of TMD properties

3.1. Discovery and identification of TMDs

In recent years, ML/DL methods have been widely used to enhance synthesis and to discover and identify synthesized materials. An ML-based approach has been studied to predict the optimum parameters for the CVD growth of MoS_2 [6]. RFs and k-nearest neighbor models were trained on various parameters, including substrate type, precursor characteristics, spectroscopy data, growth parameters, and CVD method, to classify materials as either monolayer MoS_2 or non-monolayer MoS_2. The Kursa and Rudnicki method has been employed to identify important features by creating shadow descriptors for each input variable, allowing the RF model to distinguish relevant features by comparing their importance against randomly shuffled versions of the input variables. The RF model was chosen because of its ability to highlight the key growth variables that influence the formation of MoS_2 monolayers. The k-NN model was used as a baseline for performance comparison with the RF approach. This approach began with an unsupervised strategy and later transitioned to a supervised approach. Figure 3(a) illustrates a box plot depicting the importance of parameters using the RF model, clearly indicating that the Mo precursor temperature holds the highest importance among the other parameters.

Tree-based ML algorithms, including DT, GB DT, and RFs, have been proposed to identify mechanically exfoliated 2D materials with a limited training dataset [2]. Tree-based algorithms identify 2D materials by selecting features based on inherent physical attributes like color contrast, allowing their decision-making process to be visualized through a single DT. These algorithms outperform CNNs by leveraging physical image attributes, such as color contrast, to accurately identify 2D atomic crystals without the risks of overfitting and high dataset demands. Figure 3(b) illustrates the data processing workflow for $MoSe_2$, where a color quantization technique has been implemented on optical images of $MoSe_2$ to remove unnecessary noise and prepare them for training.

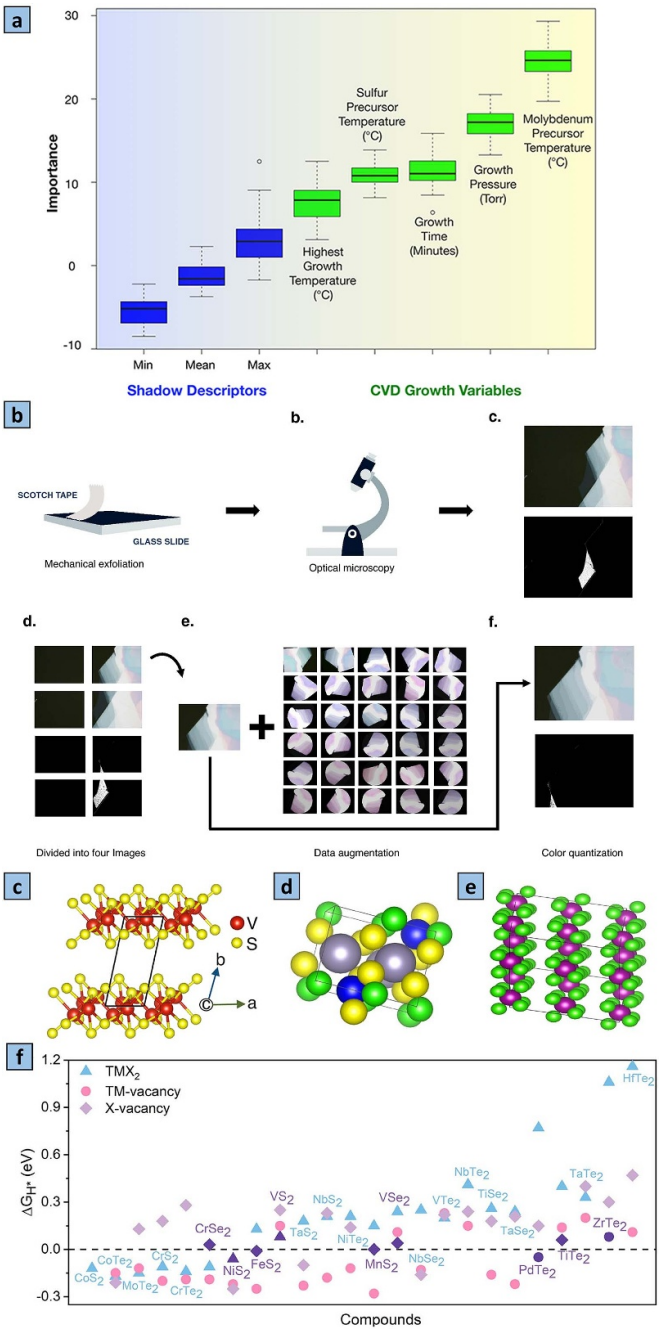

Figure 3. Discovery and identification of TMDs. (a) CVD parameters importance plot. Adapted from [6], with the permission of AIP Publishing. (b) Data processing workflow of $MoSe_2$. Adapted from [2]. CC BY 4.0. (c) Triclinic crystal V_2S_3. (d), (e) Predicted 2D crystal structure of $SnAsClS_2$ and $MnClYI_3$. Reprinted (adapted) with permission from [49]. Copyright (2021) American Chemical Society. (f) ΔG_{H^*} (hydrogen adsorption free energy) of metallic phase 2D-TMDs. Reprinted (adapted) with permission from [1]. Copyright (2021) American Chemical Society.

Furthermore, a DL-based generative model combined with a RF model has been proposed to discover new 2D materials [49]. Approximately 267 489 hypothetical 2D materials were

identified from 2.6 million samples using a GAN model called MatGAN. The RF model identified 2D materials using the Magpie feature set, which computed 132 composition-based descriptors, such as atomic number and radii, to predict material probabilities. RF enhances prediction accuracy without substantially increasing computational demands and remains relatively robust against unbalanced data. In this approach, a template-based structure prediction framework was implemented to predict new formula structures, where the crystal structure prediction network tool was utilized to predict the most similar space group from the generated 2D formulas. Figure 3(c) illustrates the structure of a triclinic crystal, V_2S_3, which is a stable 2D material. Figures 3(d) and (e) showcases multiple newly discovered 2D materials and their density functional theory (DFT)-calculated formation energies.

Another study based on ML has been proposed to screen metallic 2D TMDs for their potential as high-performance hydrogen evolution reaction (HER) electrocatalysts [1]. MoS_2 was used as a reference to determine if these materials belong to the lowest-energy phase among existing phases (1H, 2H, 1T, 1T′, 1T″). A 2D material exhibiting the lowest-energy phase and zero band gap was identified as a high-activity HER catalyst. Figure 3(F) displays the hydrogen adsorption free energy (ΔG_{H^*}) of metallic phase 2D TMDs with different defect structures at zero potential. To determine HER electrocatalysts, ΔG_{H^*} was used as a target feature in a linear regression model, followed by PCA and gradient boosting (GB) to filter out the features affecting ΔG_{H^*}. Pearson correlation was applied next to eliminate highly correlated features, ultimately refining the selection to five key features: nearest neighbor local electronegativity (LEf), average valence electron number of TM-X (Vtmx), first ionization energy of transition metal (EItm), covalent radius of the transition-metal atom (Rtm), and next-nearest neighbor local electronegativity (LEs). Ultimately, a new catalytic descriptor expression was derived as

$$\Delta GH* = 0.093 - (0.195 \cdot \mathrm{LEf} + 0.205 \cdot \mathrm{LEs}) - 0.15\mathrm{Vtmx} \quad (5)$$

where LEf and LEs represent local structure electronegativity, and Vtmx represents the valence electron number. Similarly, another method utilizing DL-based graph CNNs was proposed to discover HER catalysts from a 2D database with an accuracy close to 95% [50]. A total of 38 HER catalysts were screened from 6531 2D materials. In contrast to traditional models, this method could consider all potential active sites on the catalyst surface and analyze the structures' characteristics.

3.2. Determination of electronic and structural properties of TMDs

3.2.1. Electronic properties. The electronic properties of 2D materials refer to characteristics related to the behavior of electrons within these materials, including band gap (E_g), work function (E_w), density of states (DOS), and Fermi level (EFermi). While the standard method for calculating electronic properties is DFT, ML/DL methods have been widely employed to expedite DFT calculations and predict electronic properties in TMDs. A set of four ML-based models, including RF, GBDT, support vector regression (SVR), and multilayer perceptron (MLP), have been modeled to predict the band gap of 2D materials utilizing the computational 2D materials database (C2DB) [17]. The C2DB [27] is a standard open-source database containing various structural, thermodynamic, elastic, electronic, magnetic, and optical properties of around 1500 2D materials distributed over more than 30 distinct crystal structures. Feature selection was based on Spearman correlation analysis and feature importance scores from RF and GBDT models, revealing that the DOS at the Fermi energy (Dosef), heat of formation (Hform), and gap without spin–orbit coupling (SOC) (Gap_nosoc) significantly influenced model performance. An accuracy of more than 90% and RMSE of 0.24 and 0.27 eV was achieved using GBDT and RF models, respectively, while SVR and MLP gave an accuracy of 70% with RMSE of 0.41 and 0.43 eV, respectively. Incorporating bandgap calculations without SOC significantly improves all models, with RMSEs reducing to 0.09–0.17 eV and R^2 values exceeding 94%, integrating physical characteristics and demonstrating the effectiveness of these models in providing precise predictions for 2D materials.

Furthermore, a tree-based ML model has been proposed to build material descriptors to predict the band gap and work function of 2D materials, in which a total of 8 descriptors were employed, including the number of atoms, cell volume, molecular mass and formation heat, electronegativity, covalent radius, dipole polarizability, and ionization energy [32]. Feature selection involved creating hybrid descriptors by combining vectorized property matrices with empirical functions, significantly improving model accuracy and reducing overfitting for predicting the band gap and work function of 2D materials. ML-based models, including RF, GB, and extreme GB (XGB), were trained to perform the inference on the band gap and the work function of 2D materials in the C2DB dataset. Gradient boosting models (GB and XGB) outperformed the RF model due to their greater sensitivity to newly designed features. Figures 4(a) and (b) display the predicted bandgap and work function of 2D materials, respectively, where the filled region represents the data from the literature, while the curves represent the results from trained RF, GB, and XGB models. An R-squared (R^2) of 0.95, 0.98, and MAE of 0.16 and 0.10 eV were obtained using XGB for the bandgap and work–function predictions, respectively.

A GB ML-based method was proposed to predict electronic band gap (E_g), work function, total energy, unit-cell area, Fermi level, and DOS [9]. A selective dataset containing 981 stable materials with 15 features was considered. The features were analyzed, and the histogram shows that the band gap (E_g) and DOS are left-skewed, the area of the unit cell is bi-modal, and the total energy is right-skewed. The relative importance of the predictor variables for the Fermi level shows that the work function is the most relevant predictor for its prediction. Based on the final predictions of the Fermi level, the *R*-squared value is calculated as 0.93 and the MAE as 0.26, which shows promising results in this method.

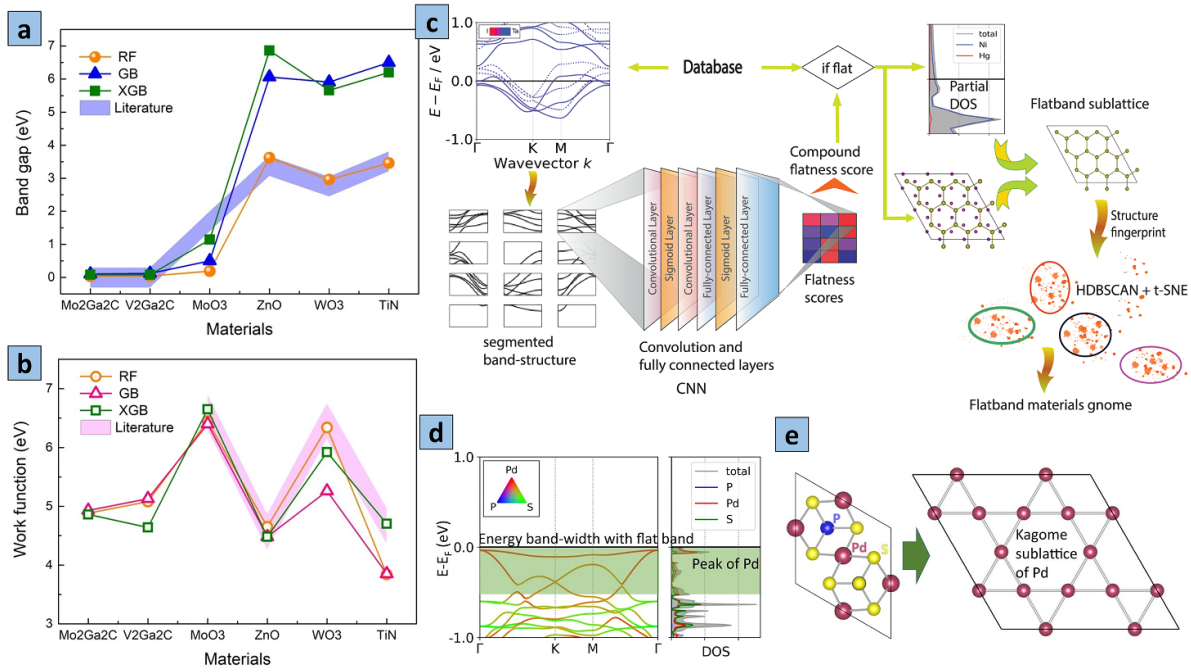

Figure 4. Prediction and identification of electronic properties of TMDs (a), (b) predicted bandgap and work function of 2D materials respectively. Adapted from [32]. CC BY 4.0. (c) Schematic of the workflow of DL-based method for identification of 2D materials with flat electronic bands [8] (d) band structure of an identified flat band material $Pd_3P_2S_8$ with projected density of states (DOS) [8]. (e) Kagome sublattice of Pd [8]. Adapted from [8]. CC BY 4.0.

A DL-based CNN was employed to identify 2D materials with flat electronic bands [8]. These are materials where the energy E_k remains independent of momentum k, forming a flat band where E_k remains constant, facilitating electron-electron interactions. CNNs for identifying band structures in images allow for high-throughput and accurate detection of band features from extensive databases, leveraging their ability to effectively capture complex patterns and relationships in image data. Figure 4(c) shows the architecture of supervised and unsupervised algorithms used in this method. This approach trained a CNN model using band structure images from an open 2D materials database known as 2D Materials Encyclopedia to detect 2D materials with flat band gaps. Subsequently, a symmetry-based analysis using an unsupervised ML algorithm was employed to classify the identified lattices based on their structural fingerprints. Figure 4(d) displays the band structure of an identified flat band material, $Pd_3P_2S_8$, where Palladium (Pd) and its sublattice depict the maximum projected DOS. Figure 4(e) shows the Kagome sublattice of Pd after stripping P and S elements from the $Pd_3P_2S_8$ structure.

3.2.2. Materials defect. Defect engineering plays a significant role in impacting material properties, where defects are intentionally designed and arranged in 2D materials, including vacancies, substitutional atoms, and grain boundaries, to achieve specific functionality and induce changes in properties like single-photon emission and resistive switching [18, 20, 21]. Several ML/DL methods have been used to study the atomic-scale defect-property relationships, control the changes, and detect and classify the defects.

An ML-based approach integrating deep transfer learning and first-principles calculations was proposed to predict properties of point defects in 2D materials and generate descriptions of defect structures within the materials [18]. Feature selection involved using physics-informed descriptors based on chemical and structural information to represent defect structures, which were then mapped to defect properties using ensemble ML models, reducing the need for extensive DFT calculations. The models selected utilized accessible descriptors to capture defect physics, enabling accurate predictions of defect properties such as formation energies and defect level positions without requiring electronic structure calculations. This method identified approximately 100 unexplored dopant defect structures across layered metal chalcogenides, hexagonal nitrides, and metal halides, suitable for quantum emission and neuromorphic computing applications. Utilizing the C2DB [27] comprising nearly 4000 2D materials alongside transfer learning, the model was trained to screen promising host materials for quantum emission and resistive switching without DFT calculations. Following the screening of potential host materials, a classifier was employed to predict the presence of deep center defects, while a regressor was developed to forecast defect formation energies. Figure 5(b) illustrates the schematic of a memory device featuring a MoS_2 layer positioned between two metallic electrodes, where M_s denotes a substitution defect (sulfur vacancy) in the MoS_2 layer.

A DL-based fully convolutional neural (FCN) network was proposed to automate detecting and classifying defects and dopants in TMDs with single-atom precision, achieving a detection limit of 1×10^{12} cm^{-2} and an accuracy of approximately 98% [21]. Figure 5(a) shows the architecture of FCN

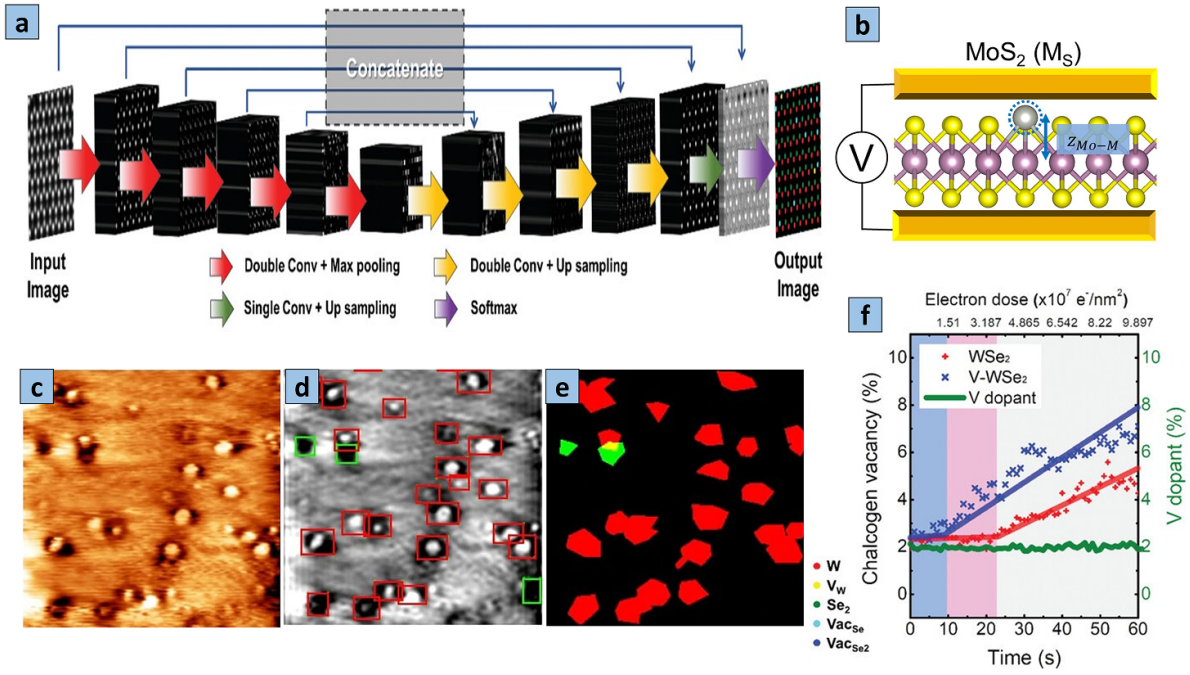

Figure 5. Defect engineering in TMDs using ML/DL methods (a) DL-based fully convolutional neural network model architecture for quantification analysis of the annular dark-field mode of scanning transmission electron microscopy images (a), (f) Adapted from [21]. CC BY 4.0. (b) Schematic of a memory device with a MoS_2 layer sandwiched between two metallic electrodes where M_s represents a substitution defect (sulfur vacancy) in MoS_2 [18]. Reprinted (adapted) with permission from [19]. Copyright (2020) American Chemical Society. (c), (d), (e) scanning tunneling microscope MoS_2 image, ground truth image, YOLOV4 model detected image respectively. Adapted from [19]. CC BY 4.0. (f) Electron beam-induced generation of chalcogen vacancies in WSe_2 and V doped WSe_2

model. The FCN model was chosen for its ability to segment image features, addressing the challenges of noise and distortion in high-resolution ADF images. This DL approach facilitates reliable quantification of atomic defects and dopants in 2D TMDs, overcoming the limitations of traditional imaging methods. An image restoration algorithm based on CNN was modeled to reduce the noise and enhance the contrast of STEM images, followed by a FCN to quantify the dopants and defects in TMDs monolayers and their doped forms, including WSe_2, MoS_2, V-doped WSe_2, and V-doped MoS_2, achieving single-atom precision and a 1200-fold increase in efficiency for site classification [21]. Figure 5(f) displayes the electron-beam-induced generation of chalcogen vacancies in WSe_2 and V doped WSe_2.

A reinforcement learning model employing a Monte Carlo tree search (MCTS) with a delayed rewards approach was introduced to efficiently explore the defect configurational space and identify optimal defect arrangements in 2D materials [22]. MCTS is chosen over evolutionary algorithms for its ability to efficiently navigate complex, high-dimensional search spaces and adaptively balance exploration and exploitation. Unlike evolutionary algorithms, MCTS excels in achieving faster convergence and making decisions based on past interactions with the environment, making it well-suited for materials design and discovery tasks. MCTS dynamically explores and optimizes defect configurations based on learned rewards, effectively managing feature selection and transformation without explicit manual feature engineering. This method begins with an initial distribution of randomly placed S point defects (or vacancies) across the chalcogen layer of MoS_2 and navigates the search space of various extended defect configurations to identify the lowest energy optimal defect configuration. The results were compared with those obtained using a genetic algorithm, revealing that MCTS could predict lower energy configurations with fewer evaluations. A DL-based atomic defect detection framework was explicitly developed for detecting atomic defects in MoS_2 and then generalized to apply to other TMDs [19]. Within this framework are a data augmentation module to generate additional pseudo data for training the model, a color preprocessing module and a noise filtering module to improve data quality, and a U-Net-based detection model to accurately locate and identify the atomic defect. MoS_2 and WS_2 were considered testing samples, achieving an F1-score of 0.89 for impurities and 0.80 for voids in MoS_2 when trained with approximately 70 images. Furthermore, using the same model without retraining, an F1-score of 0.94 for voids was achieved in WS_2. Figure 5(c), (d), and (e) shows the scanning tunneling microscope image, ground truth image, and YOLOV4 model detected image of MoS_2 respectively.

3.2.3. Layer number classification. Standard methods to identify the number of layers are optical microscopy, Raman spectroscopy, photoluminescence spectroscopy, and atomic force microscopy. ML/DL algorithms use a high generalization approach to recognize and interpret images, enabling an efficient identification of the optical properties of materials. These methods involve image processing tasks, including

classification, segmentation, and detection of material flakes from optical images.

A DL-based encoder–decoder semantic segmentation network, named 2D material optical identification neural network, was proposed to characterize and extract deep graphical features such as color, contrast, edges, shapes, and flake sizes [51]. DL-based models can identify various 2D materials in real-time from optical microscopy images, irrespective of variations in optical setups, significantly reducing workload. Additionally, the model's ability to correlate optical images with the physical properties of 2D materials makes it valuable for predicting the properties of new, uncharacterized 2D crystals. A total of 817 optical images and 100 background images of mechanically exfoliated flakes collected from 13 different 2D materials have been used for training and testing the model. A random rotation data augmentation method was employed to introduce random positioning, and a stochastic gradient descent using the momentum method was utilized to train the weights. This method successfully characterized flakes with different thicknesses (1 l for monolayer, 2–6 l for bilayer to 6-layer, and >6 for more than 6 layers).

A 3D CNN, deep-learning-enabled atomic layer mapping (DALM), has been studied to identify and segment MoS_2 flakes with mono, bi, tri and multilayers [3]. This method involves merging RGB images for profile information and hyperspectral imaging microscopy, which combines spectroscopy and imaging techniques to provide both spatial and spectral information on the region of interest. This fusion enabled precise identification of layer numbers, addressing the challenges of merging 2D and 3D data for comprehensive material characterization. The dice similarity coefficient (DSC), Hausdorff distance, and confusion matrix were utilized to analyze the performance of the prediction results quantitatively. Results were compared with a single-stream U-Net (S-U-Net) model trained on RGB images. Based on the calculated median DSC values, the DALM model achieves higher DSC value (>90%) for monolayers and background, while S-U-Net only achieves 22%, 20%, and 4% for bi-, tri-, and multilayers, respectively.

The three DL architectures, DenseNet, U-Net, and Mask-region CNN, have been studied [14] to classify, segment, and detect microscopic images of 2D materials for automated atomic layer mapping. The three models DenseNet, U-Net, and Mask-RCNN were chosen for their suitability in handling distinct computer vision tasks, which are multilabel material classification, material segmentation, and material detection, respectively. The performance of these models was evaluated based on RGB optical contrast differences and the CIE 1931 color space. Figures 6(a)–(c) show the RGB optical image, manually annotated optical image and final predicted optical image showing thicknesses as monolayer (1 l), bilayer (2 l), and bulk, respectively. Figure 6(d) shows the CIE 1931 color space of all categories (background, monolayer, bilayer, trilayer, multilayer, and bulk) with the different contrast sampling index ($\gamma R \neq \gamma G \neq \gamma B$ in the range of [0.8, 1.2]). In addition to the CNN approaches above, image-to-image translation using cGAN has been studied in computer vision. A cGAN model termed Pix2Pix has been proposed containing a U-net generator plus a discriminator to translate spatially aligned pairs of images [26]. It has been utilized in various applications, including optical sectioning to translate wide-field fluorescence microscopy images to optically sectioned SIM images [14], semantic segmentation [16], and image processing.

4. ML-based applications

ML accelerates the development process through high-throughput screening and automated design, enabling the rapid identification of new compounds with desirable properties. In optoelectronics [10], ML refines the design of devices such as solar cells and LEDs by predicting performance outcomes and optimizing material configurations. For energy storage [11, 12], ML models predicted the efficiency and longevity of battery materials and supercapacitors, driving advancements in energy density and performance. A high-throughput computational framework has been developed to evaluate 2D materials for lithium-ion batteries and supercapacitors [12]. By analyzing adsorption configurations and predicting material performance, this framework speeded the discovery of effective electrode materials, paving the way for advancing next-generation energy storage systems. In catalysis, ML was used to optimize catalytic reactions and discover new catalysts, improving industrial processes and environmental sustainability [50]. For structural materials, ML predicts mechanical properties and failure risks, aiding in the design of durable and reliable components. ML was combined with advanced imaging techniques to accurately identify material layers and predict their properties, leading to innovations in electronics and photonics.

5. Discussions

In exploring the recent advances in ML and DL methods applied to TMD research, their accuracy heavily relies on the quantity and quality of the initial data that must be obtained manually via optical and spectroscopic measurements. Insufficient data can lead to overfitting or underfitting, resulting in less accurate predictions. This limitation is particularly pronounced in TMD studies, where datasets may be limited or fragmented. Another concern is the black-box nature of many ML/DL models, where the decision-making process is opaque, making it challenging to interpret the underlying mechanisms driving the predictions. This lack of interpretability can inhibit the ability to derive meaningful physical insights from the models, limiting their utility in guiding experimental studies and advancing our understanding of TMD properties. To address these limitations, several strategies can be adopted. Firstly, data curation efforts should be prioritized to ensure the availability of high-quality, standardized datasets, which require close collaborations between researchers and data repositories to aggregate and annotate relevant data effectively. By fostering an open-source culture and sharing both data and code, researchers can accelerate progress and

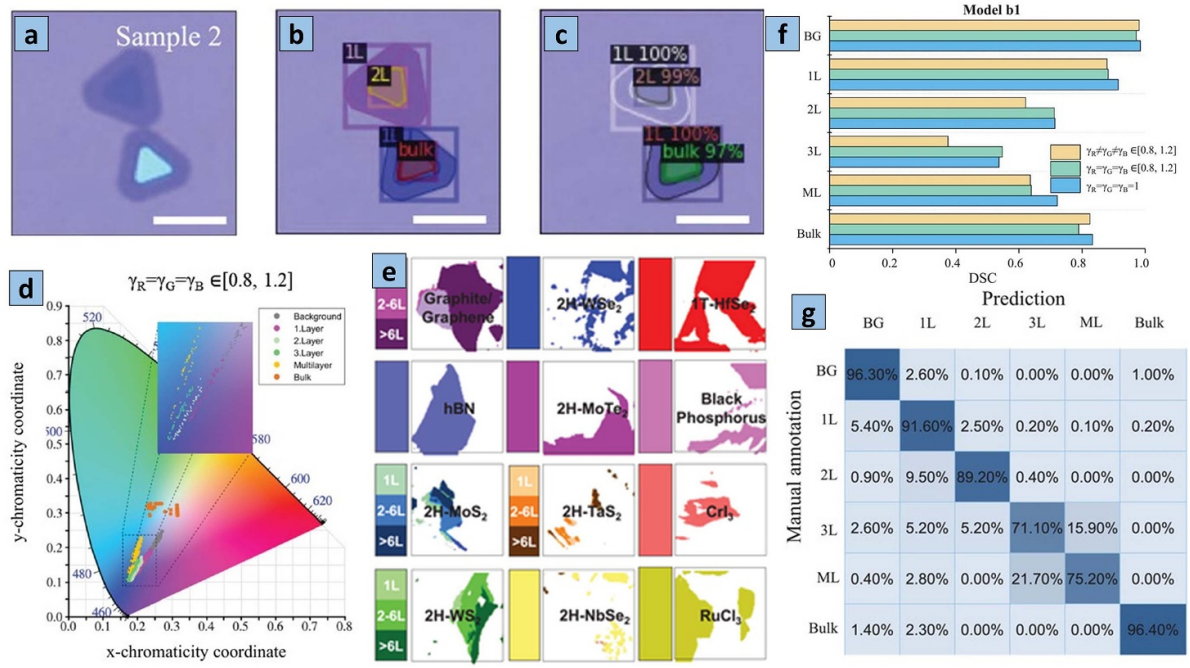

Figure 6. Identification of thickness in TMDs (a), (b), (c) RGB optical image, manually annotated optical image and final predicted optical image showing thicknesses as monolayer (1 l), bilayer (2 l), and bulk, respectively (d) The CIE 1931 color space of all categories (background, monolayer, bilayer, trilayer, multilayer, and bulk) with the different contrast sampling index ($\gamma R \neq \gamma G \neq \gamma B$ in the range of [0.8, 1.2]) (e) Labeled images of graphite, hBN, 2 H–MoS_2, 2 H–WS_2, 2 H–WSe_2, 2 H–$MoTe_2$, 2 H–TaS_2, 2 H–$NbSe_2$, 1 T–$HfSe_2$, black phosphorus, CrI_3, and $RuCl_3$ with material identities and thicknesses (f) Dice similarity coefficient values of Mask-RCNN model with different gamma contrast sampling indices (g) Confusion matrix with six classes as substrate, mono-, bi-, tri-, multilayer, and bulk. (a)–(d), (f), (g) Adapted from [14]. CC BY 4.0. (e) [51] John Wiley & Sons. © 2020 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim.

facilitate the development of more robust ML/DL models for TMD studies. Data augmentation techniques can also play a crucial role in mitigating data scarcity issues by synthetically generating additional training samples. Augmentation methods such as rotation, flipping, and adding noise can help diversify the dataset and improve the robustness of the trained models.

Furthermore, recent advancements in ML/DL models, such as attention mechanisms and generative models, offer promising avenues for enhancing model interpretability and performance. Attention mechanisms enable the model to focus on relevant features while disregarding noise, thereby improving prediction accuracy and extracting meaningful insights. On the other hand, generative models facilitate the creation of synthetic data that closely resembles real-world observations, aiding in data augmentation efforts and expanding the diversity of training samples. Integrating generative models into the ML/DL framework can enhance the robustness and generalization capabilities of the models, contributing to more accurate predictions and a deeper understanding of TMD properties. Integrating feedback from experimental techniques into ML/DL models can enhance their predictive capabilities and promote synergy between computational and experimental approaches. However, despite these advancements, challenges remain in fully realizing the potential of ML/DL models for TMD research. One significant challenge is the interpretability of complex models, especially deep neural networks, which are often treated as black boxes [61]. Attention mechanisms may sometimes focus on features that are not physically meaningful, leading to potential misinterpretations. Similarly, generative models, while useful for creating diverse datasets, may introduce artifacts that do not correspond to the real physical nature of TMDs. Several attempts are being made to address these inefficiencies. First, integrating domain-specific knowledge into the model architecture and training process can help improve interpretability by ensuring that the model's focus aligns with known physical principles. Second, the use of explainable AI techniques, such as feature importance analysis and model-agnostic interpretation methods, can provide additional layers of transparency by allowing researchers to analyze the model's decision-making process.

By iteratively refining the model based on experimental observations, researchers can improve its accuracy and ensure its relevance. By addressing these challenges through improved data curation, augmentation, model transparency, collaboration, and experimental feedback, we can harness the full potential of ML/DL in TMD research and pave the way for future breakthroughs in this exciting field.

6. Conclusions

In this article, we have reviewed ML and DL methods in identifying and characterizing TMDs. We have examined

studies that employ ML/DL methods to characterize various aspects of TMDs, including the determination of the number of layers, crystal structure, band gap energy, and detection of defects. Through these investigations, we have witnessed the potential of ML/DL techniques to streamline and enhance the analysis of TMD materials, offering insights that can inform both theoretical understanding and practical applications. Additionally, we have reviewed the challenges and perspectives associated with this emerging research direction. Issues such as data scarcity, model interpretability, and integration with experimental techniques present ongoing challenges that must be addressed to fully harness the potential of ML/DL in TMD research. Looking ahead, the continuous advancement of ML/DL methodologies, coupled with interdisciplinary collaborations and data-sharing initiatives, holds potential for unlocking new opportunities for materials innovation. Addressing the identified challenges and embracing emerging research directions can pave the way for transformative breakthroughs in 2D material studies, with far-reaching implications for various technological applications.

Data availability statement

The data cannot be made publicly available upon publication because they are owned by a third party and the terms of use prevent public distribution. The data that support the findings of this study are available upon reasonable request from the authors.

ORCID iDs

Shivani Bhawsar https://orcid.org/0009-0005-1824-9917
Eui-Hyeok Yang https://orcid.org/0000-0003-4893-1691

References

[1] Ran N, Sun B, Qiu W, Song E, Chen T and Liu J 2021 Identifying metallic transition-metal dichalcogenides for hydrogen evolution through multilevel high-throughput calculations and machine learning *J. Phys. Chem. Lett.* **12** 2102–11
[2] Zichi L, Liu T, Drueke E, Zhao L and Xu G 2023 Physically informed machine-learning algorithms for the identification of two-dimensional atomic crystals *Sci. Rep.* **13** 6143
[3] Dong X *et al* 2021 3D deep learning enables accurate layer mapping of 2D materials *ACS Nano* **15** 3139–51
[4] Vahdat A and Kautz J 2020 NVAE: a deep hierarchical variational autoencoder *Advances in Neural Information Processing Systems* vol 33 (Curran Associates, Inc.) 19667–79
[5] Zhuge H, Summa B, Hamm J and Brown J Q 2021 Deep learning 2D and 3D optical sectioning microscopy using cross-modality Pix2Pix cGAN image translation *Biomed. Opt. Express* **12** 7526
[6] Costine A, Delsa P, Li T, Reinke P and Balachandran P V 2020 Data-driven assessment of chemical vapor deposition grown MoS_2 monolayer thin films *J. Appl. Phys.* **128** 235303
[7] Ryu B, Wang L, Pu H, Chan M K Y and Chen J 2022 Understanding, discovery, and synthesis of 2D materials enabled by machine learning *Chem. Soc. Rev.* **51** 1899–925
[8] Bhattacharya A, Timokhin I, Chatterjee R, Yang Q and Mishchenko A 2023 Machine learning approach to genome of two-dimensional materials with flat electronic bands *npj Comput. Mater.* **9** 101
[9] Alibagheri E, Mortazavi B and Rabczuk T 2021 Predicting the electronic and structural properties of two-dimensional materials using machine learning *Comput. Mater. Contin.* **67** 1287–300
[10] Cheng Z *et al* 2021 2D materials enabled next-generation integrated optoelectronics: from fabrication to applications *Adv. Sci.* **8** 2003834
[11] Yang H and He Z 2023 Reshaping the material research paradigm of electrochemical energy storage and conversion by machine learning *EcoMat.* **5** e12330
[12] Kabiraj A and Mahapatra S 2022 High-throughput assessment of two-dimensional electrode materials for energy storage devices *Cell Rep. Phys. Sci.* **3** 100718
[13] Yang J and Yao H 2020 Automated identification and characterization of two-dimensional materials via machine learning-based processing of optical microscope images *Extreme Mech. Lett.* **39** 100771
[14] Dong X *et al* 2022 Deep-learning-based microscopic imagery classification, segmentation, and detection for the identification of 2D semiconductors *Adv. Theory Simul.* **5** 2200140
[15] Sanchez-Juarez J, Granados-Baez M, Aguilar-Lasserre A A and Cardenas J 2022 Automated system for the detection of 2D materials using digital image processing and deep learning *Opt. Mater. Express* **12** 1856
[16] Masubuchi S, Watanabe E, Seo Y, Okazaki S, Sasagawa T, Watanabe K, Taniguchi T and Machida T 2020 Deep-learning-based image segmentation integrated with optical microscopy for automatically searching for two-dimensional materials *npj 2D Mater. Appl.* **4** 3
[17] Zhang Y, Xu W, Liu G, Zhang Z, Zhu J, Li M and Tan M L P 2021 Bandgap prediction of two-dimensional materials using machine learning *PLoS One* **16** e0255637
[18] Frey N C, Akinwande D, Jariwala D and Shenoy V B 2020 Machine learning-enabled design of point defects in 2D materials for quantum and neuromorphic information processing *ACS Nano* **14** 13406–17
[19] Chen F-X R, Lin C-Y, Siao H-Y, Jian C-Y, Yang Y-C and Lin C-L 2023 Deep learning based atomic defect detection framework for two-dimensional materials *Sci. Data* **10** 91
[20] Bertoldo F, Ali S, Manti S and Thygesen K S 2022 Quantum point defects in 2D materials—the QPOD database *npj Comput. Mater.* **8** 56
[21] Yang S *et al* 2021 Deep learning-assisted quantification of atomic dopants and defects in 2D materials *Adv. Sci.* **8** 2101099
[22] Banik S, Loeffler T D, Batra R, Singh H, Cherukara M J and Sankaranarayanan S K R S 2021 Learning with delayed rewards—a case study on inverse defect design in 2D materials *ACS Appl. Mater. Interfaces* **13** 36455–64
[23] Lu M, Ji H, Zhao Y, Chen Y, Tao J, Ou Y, Wang Y, Huang Y, Wang J and Hao G 2023 Machine learning-assisted synthesis of two-dimensional materials *ACS Appl. Mater. Interfaces* **15** 1871–8
[24] Priya P, Nguyen T C, Saxena A and Aluru N R 2022 Machine learning assisted screening of two-dimensional materials for water desalination *ACS Nano* **16** 1929–39
[25] Isola P, Zhu J-Y, Zhou T and Efros A A 2016 Image-to-image translation with conditional adversarial networks *2017 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)* pp 5967–76 (arXiv:1611.07004)
[26] Abdelmotaal H, Abdou A A, Omar A F, El-Sebaity D M and Abdelazeem K 2021 Pix2pix conditional generative adversarial networks for scheimpflug camera color-coded

corneal tomography image generation *Transl Vis. Sci. Technol.* **10** 21
[27] Haastrup S *et al* 2018 The computational 2D materials database: high-throughput modeling and discovery of atomically thin crystals *2D Mater.* **5** 042002
[28] Rasmussen F A and Thygesen K S 2015 Computational 2D materials database: electronic structure of transition-metal dichalcogenides and oxides *J. Phys. Chem.* C **119** 13169–83
[29] Gjerding M N *et al* 2021 Recent progress of the computational 2D materials database (C2DB) *2D Mater.* **8** 044002
[30] Belsky A, Hellenbrandt M, Karen V L and Luksch P 2002 New developments in the inorganic crystal structure database (ICSD): accessibility in support of materials research and design *Acta Crystallogr.* B **58** 364–9
[31] Bergerhoff G, Hundt R, Sievers R and Brown I D 1983 The inorganic crystal structure data base *J. Chem. Inf. Comput. Sci.* **23** 66–69
[32] Dau M T, Al Khalfioui M, Michon A, Reserbat-Plantey A, Vézian S and Boucaud P 2023 Descriptor engineering in machine learning regression of electronic structure properties for 2D materials *Sci. Rep.* **13** 5426
[33] Bhawsar S, Fang M, Sarkar A S, Chen S and Yang E-H 2024 Deep learning-based multimodal analysis for transition-metal dichalcogenides *MRS Bull.* **49** 1021–31
[34] Davidsson J, Bertoldo F, Thygesen K S and Armiento R 2023 Absorption versus adsorption: high-throughput computation of impurities in 2D materials *npj 2D Mater. Appl.* **7** 26
[35] Campi D, Mounet N, Gibertini M, Pizzi G and Marzari N 2022 The materials cloud 2D database (MC2D) *Mater. Cloud* **2022** 84
[36] Mounet N *et al* 2018 Two-dimensional materials from high-throughput computational exfoliation of experimentally known compounds *Nat. Nanotechnol.* **13** 246–52
[37] Zhou J *et al* 2019 2DMatPedia, an open computational database of two-dimensional materials from top-down and bottom-up approaches *Sci. Data* **6** 86
[38] Sorkun M C, Astruc S, Koelman J M V A and Er S 2020 An artificial intelligence-aided virtual screening recipe for two-dimensional materials discovery *npj Comput. Mater.* **6** 106
[39] Sedgwick P 2012 Pearson's correlation coefficient *BMJ* **345** e4483
[40] Maćkiewicz A and Ratajczak W 1993 Principal components analysis (PCA) *Comput. Geosci.* **19** 303–42
[41] Hearst M A, Dumais S T, Osuna E, Platt J and Scholkopf B 1998 Support vector machines *IEEE Intell. Syst. Their Appl.* **13** 18–28
[42] Guo G, Wang H, Bell D, Bi Y and Greer K 2003 KNN model-based approach in classification *On the Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE* ed R Meersman, Z Tari and D C Schmidt (Springer) pp 986–96
[43] Sinaga K P and Yang M-S 2020 Unsupervised K-means clustering algorithm *IEEE Access* **8** 80716–27
[44] Quinlan J R 1986 Induction of decision trees *Mach. Learn.* **1** 81–106
[45] Breiman L 2001 Random forests *Mach. Learn.* **45** 5–32
[46] Roberts S J, Husmeier D, Rezek I and Penny W 1998 Bayesian approaches to Gaussian mixture modeling *IEEE Trans. Pattern Anal. Mach. Intell.* **20** 1133–42
[47] O'Shea K and Nash R 2015 An introduction to convolutional neural networks (arXiv:1511.08458) (Accessed 26 November 2015)
[48] Creswell A, White T, Dumoulin V, Arulkumaran K, Sengupta B and Bharath A A 2018 Generative adversarial networks: an overview *IEEE Signal Process. Mag.* **35** 53–65
[49] Song Y, Siriwardane E M D, Zhao Y and Hu J 2021 Computational discovery of new 2D materials using deep learning generative models *ACS Appl. Mater. Interfaces* **13** 53303–13
[50] Wu S, Wang Z, Zhang H, Cai J and Li J 2023 Deep learning accelerates the discovery of two-dimensional catalysts for hydrogen evolution reaction *Energy Environ. Mater.* **6** e12259
[51] Han B *et al* 2020 Deep-learning-enabled fast optical identification and characterization of 2D materials *Adv. Mater* **32** 2000953
[52] Liu B and Udell M 2020 Impact of accuracy on model interpretations (arXiv:2011.09903) (Accessed 17 November 2020)
[53] Michaud E J, Liu Z and Tegmark M 2023 Precision machine learning *Entropy* **25** 175
[54] Fränti P and Mariescu-Istodor R 2023 Soft precision and recall *Pattern Recognit. Lett.* **167** 115–21
[55] Sokolova M, Japkowicz N and Szpakowicz S 2006 Beyond accuracy, F-score and ROC: a family of discriminant measures for performance evaluation *Adv. Artif. Intell. Lecture Notes Comput. Sci.* **4304** 1021
[56] Hajian-Tilaki K 2013 Receiver operating characteristic (ROC) curve analysis for medical diagnostic test evaluation *Casp. J. Intern. Med.* **4** 627–35
[57] Hodson T O, Over T M and Foks S S 2021 Mean squared error, deconstructed *J. Adv. Model. Earth Syst.* **13** e2021MS002681
[58] Hodson T O 2022 Root-mean-square error (RMSE) or mean absolute error (MAE): when to use them or not *Geosci. Model. Dev.* **15** 5481–7
[59] Chicco D, Warrens M J and Jurman G 2021 The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation *PeerJ Comput. Sci.* **7** e623
[60] Saito Y, Shin K, Terayama K, Desai S, Onga M, Nakagawa Y, Itahashi Y M, Iwasa Y, Yamada M and Tsuda K 2019 Deep-learning-based quality filtering of mechanically exfoliated 2D crystals *npj Comput. Mater.* **5** 1–6
[61] Dobson J E 2023 On reading and interpreting black box deep neural networks *Int. J. Digit. Humanit.* **5** 431–49