Coordination Chemistry Reviews 529 (2025) 216436

Contents lists available at ScienceDirect

Coordination Chemistry Reviews

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

Deep learning-assisted methods for accelerating the intelligent screening of novel 2D materials: New perspectives focusing on data collection and description

Yuandong Lin [a,b], Ji Ma [a,c], Yong-Guang Jia [d], Chongchong Yu [e], Jun-Hu Cheng [a,b,*]

[a] School of Food Science and Engineering, South China University of Technology, Guangzhou 510641, China
[b] Guangdong Provincial Key Laboratory of Intelligent Food Manufacturing, College of Food Science and Engineering, Foshan University, Foshan 528225, Guangdong, China
[c] State Key Laboratory of Luminescent Materials and Devices, Center for Aggregation-Induced Emission, South China University of Technology, Guangzhou 510640, China
[d] Center for Advanced Materials Research, Beijing Normal University, Zhuhai 519085, China
[e] Key Laboratory of Industrial Internet and Big Data, China National Light Industry, Beijing Technology and Business University, Beijing 100048, China

ARTICLE INFO

Keywords:
2D materials
Deep learning
Data collections
Data descriptions
Material screenings

ABSTRACT

Since the isolation of graphene, the interest in two-dimensional (2D) materials has been steadily growing thanks to their unique chemical and physical properties, as well as their potential for various applications. Deep learning (DL), currently one of the most sophisticated machine learning (ML) models, is emerging as a highly effective tool for intelligently investigating and screening 2D materials. The utilization of abundant data sources, appropriate descriptors, and neural networks enables the prediction of the structural and physicochemical properties of undiscovered 2D materials based on DL. Specifically, high-quality and well-described data plays a crucial role in effective model training, accurate predictions, and the discovery of new 2D materials. It also promotes reproducibility, collaboration, and continuous improvement within this field. This tutorial review is dedicated to an examination of the characterization, prediction, and discovery of 2D materials facilitated by various DL techniques. It focuses on the perspective of data collection and description, aiming to provide a clearer understanding of underlying principles and predicting outcomes. In addition, it also offers insights into future research prospects. The growing acceptance of DL is set to accelerate and transform the study of 2D materials.

1. Introduction

Since the discovery of graphene, numerous studies have been conducted with the objective of exploring and synthesizing functional graphene, due to the intriguing fundamental physics that emerges in reduced dimensions [1]. To date, in addition to graphene, various emerging 2D materials, including hexagonal boron nitride (h-BN), graphitic carbon nitride (g-C_3N_4), black phosphorus (BP), transition metal dichalcogenides (TMDs), MXenes, and metal-organic frameworks (MOFs), have been extensively researched and applied in energy storage, environmental pollution, biomedical engineering, and analytical testing fields due to their excellent mechanical, electrical, and optical characteristics [2,3]. Although 2D materials exhibit intriguing properties and can now be synthesised feasibly, they encounter a number of challenges that hinder their industrial wider applications. These challenges include the absence of a band gap in graphene [4], low carrier mobility in TMDs [5], the need for scalable production of h-BN [6], and the chemical instability of BP [7]. The conventional approach of experimental synthesis to address these challenges primarily relies on chemical intuition and serendipitous discoveries. This often results in a lengthy process, typically taking a decade or more from initial research to commercial applications. As a result, discovering new 2D materials can be a time-consuming and expensive undertaking [8]. Consequently, the exploration of all conceivable 2D configurations presents significant challenges.

As is the case in other fields, the advancement of 2D materials has shifted from the conventional approach of experimental findings to the contemporary approach of data mining. Theoretical methods, including

* Corresponding author at: School of Food Science and Engineering, South China University of Technology, Guangzhou 510641, China
E-mail address: chengjunhu1229@163.com (J.-H. Cheng).

https://doi.org/10.1016/j.ccr.2025.216436
Received 12 October 2024; Received in revised form 27 December 2024; Accepted 2 January 2025
Available online 16 January 2025

ab initio calculations, molecular dynamics (MD) simulations, and Monte Carlo simulations, have been employed to examine the micro-level system behaviour of 2D materials and inform future working efforts [9,10]. The advancements in theoretical calculation tools and methodologies have markedly accelerated the process of screening and designing 2D materials. This approach has not only resulted in a reduction in both time and cost, but it has also emerged as a highly effective instrument for the research and development of novel 2D materials. Nevertheless, it is important to acknowledge that these simulations, which are based on existing theories and have limited experimental validation, continue to be heavily reliant on traditional ways of thinking and working. As an illustration, a substantial obstacle persists for 2D-MOFs, given that the extensive range of potential MOFs that can be generated through the adaptable combination of metal ions and ligands exceeds the computational capacity to conduct a comprehensive analysis of them with the currently available resources [11]. In order to achieve the desired 2D materials, it is necessary to develop a novel and efficient method that is affordable, rapid, and precise at the theoretical calculation level.

In comparison to conventional computing techniques, machine learning (ML) represents a superior option. By analysing a subset of data and generating a model, ML can rapidly, accurately, and cost-effectively predict the performance of 2D materials, thereby elucidating the relationships between structure and activity [12,13]. Furthermore, deep learning (DL), a branch of ML (Fig. 1A), has demonstrated notable progress in the prediction, characterization, and discovery of 2D materials, having the ability to automatically, effectively, and precisely learn the characteristics and patterns within the data. As shown in Fig. 1B, the

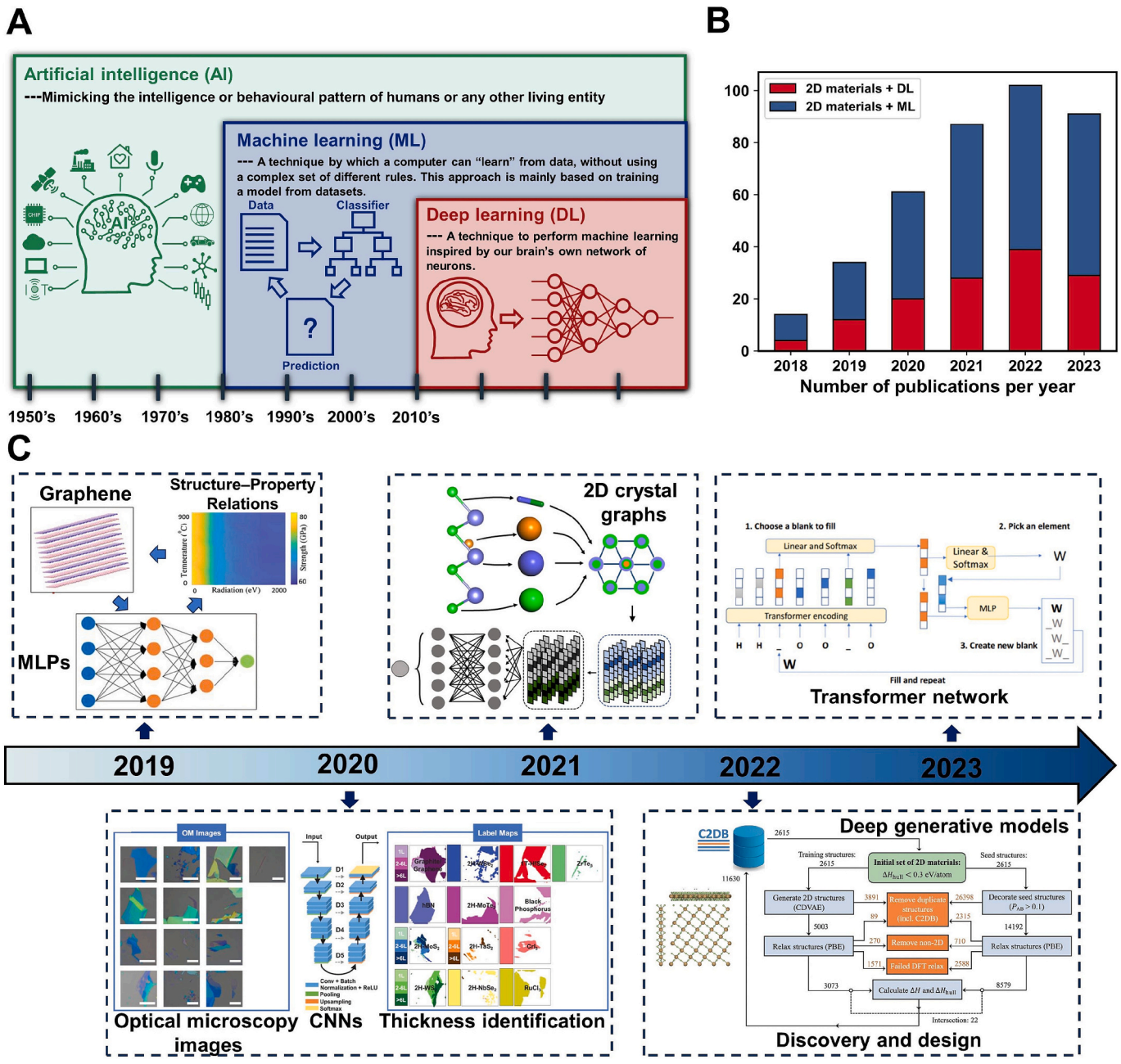

Fig. 1. Advancements in deep learning research on 2D materials. (A) Exploring the evolution and distinctions between artificial intelligence (AI), machine learning (ML), and deep learning (DL). (B) Quantifying the rise in publications utilizing ML and DL for research on 2D materials (as of May 2024). (C) Major improvements in 2D materials achieved through DL-based approaches. Reproduced with permission from Refs. [14–16]. Copyright 2019, 2020, and 2022 Wiley-VCH. Reproduced under the terms of Creative Commons Attribution 4.0 International (CC BY 4.0) license from Ref. [17, 18] published by Springer Nature and Wiley-VCH.

number of articles published annually on the subject of ML and DL in 2D materials has exhibited a consistent upward trend, with projections indicating that this trend will continue. In contrast to conventional ML techniques, which require the manual extraction of features, DL employs a hierarchical framework that enables the automatic extraction of abstract features from raw data. This approach is characterised by enhanced data-driven capabilities, resulting in notable improvements in accuracy for tasks such as classification, recognition, and prediction. Specifically, as illustrated in Fig. 1C, Hundi and Shahsavari emphasised the potential of DL techniques, such as multilayer perceptron (MLP), in accelerating the discovery of new materials by facilitating the development of processing-structure-property relationships for 2D materials [14]. The utilization of DL has the potential to reduce the necessity for simulations by a significant margin, with a reduction of up to 45 % compared to a solely MD-based approach. This reduction in the number of simulations can facilitate a considerable acceleration in the pace at which new 2D materials are understood, resulting in substantial savings in time and cost. Alternatively, DL can be employed to analyse the morphological or spectroscopic data about 2D materials. In one study, convolutional neural networks (CNNs) were developed with the objective of accelerating the processing and preliminary assessment of 2D materials with exceptional precision. It was demonstrated that CNNs were capable of extracting intricate visual features, including contrast, color, edges, shapes, segment sizes, and distributions, from optical microscopic images [15]. Furthermore, graph neural networks (GNNs) are effective methods for representing the crystal structure of 2D materials, as they are designed to handle input structured as graphs by utilizing message-passing mechanisms. These mechanisms facilitate the continuous updating of node representations by incorporating information from adjacent nodes and edges. This distinctive capacity of GNNs renders them particularly well-suitable for investigating the characteristics of 2D materials and enabling inverse design [16]. Moreover, the utilization of generative adversarial networks (GAN) has the potential to extend the scope of feasible 2D materials. This is achieved by representing lattice structures, atomic arrangements, and elemental compositions as a matrix of features in both real and reciprocal space, which can evolve dynamically. For example, Lyngby and Thygesen developed a sophisticated generative model capable of producing over 8500 distinct 2D crystals with optimal formation energies. Moreover, all of these structures are available in the C2DB database. Furthermore, the composition of 2D materials can be represented as a sequence, which allows for the construction of a DL model that excels in both sequential learning and sequence generation [17]. In 2023, Dong et al. presented a novel generative design process for the discovery of 2D materials. Their approach employed a transformer-based generator, specifically designed for the creation of 2D material compositions [18]. It was observed that this transformer generator was capable of effectively capturing composition preferences, thereby enabling the generation of chemically valid potential 2D materials. In summary, the capacity of DL can significantly improve the efficiency, streamline the 2D materials development process, and reduce the costs associated with the discovery of new 2D materials undoubtedly making it a promising avenue for future research. In addition, it is evident that the aforementioned studies have concentrated on the collection of data and the identification of suitable descriptive methods for the training of common and effective DL models, thereby ensuring the accuracy and generalization capability of the predictive models. The utilization of high-quality and well-described data not only enables DL models to learn accurately and generalize effectively, but also facilitates the implementation of rapid, automated, and innovative 2D material discovery processes [19]. Additionally, understanding the methods used to obtain data and the types of data descriptions can facilitate the rapid familiarization of researchers in the field of materials with data-driven material property prediction [20].

To the best of our knowledge, several published reviews on the subject matter in question have been conducted, each focusing on the different aspects [8,12,21,22]. However, there is a pressing need for a comprehensive and up-to-date overview, with accurate references, particularly in the field of data preparation and description of 2D materials, in order to provide a more robust foundation for the training of various DL models. The objective of this review is to deeply summarize the studies utilizing DL for predicting, characterizing, and discovering properties of 2D materials, highlighting the methods employed for dataset preparation and the establishment of descriptors. Progress in 2D materials screening can be accelerated by using rich and large volumes of datasets obtained from published databases, theoretical calculations, scientific literatures, and experimental data, especially suitable for the DL models that may encounter data hunger problems. In addition, due to the impact on the performance and understanding of the DL models, different levels of data description methods regarding 2D materials need to be given attention and summarised, mainly including gross-level, molecular-fragment level, and atomistic-based level descriptors. Besides that, several DL architectures commonly used for the characterization, discovery, and properties prediction of 2D materials are introduced. The current challenges and potential future prospects for DL-based investigations of 2D materials are outlooked.

2. The workflow of deep learning

As illustrated in Fig. 2, the conventional approach involves an exhaustive examination of the existing literature to identify the optimal properties, structures, and synthesis conditions of the desired 2D materials. The findings are then subjected to a process of trial-and-error and expert knowledge, which is an extremely time-consuming endeavour [12]. In contrast, the data-driven approach involves the creation of a ML model that utilizes a range of data sources, including automatic materials databases, scientific literatures, theoretical calculations, and experimental results. By incorporating data from various sources, the ML model can enhance its predictions of properties, structures, and synthesis conditions, resulting in a significantly more efficient 2D prediction cycle. In general, the construction of a ML model typically comprises six principal stages. These include the collection and preprocessing of data, the selection of features, the choice of an appropriate ML algorithm, the optimization of parameters, and the performance of requisite calculations. For further details, please refer to Lu et al. [21]. It is worth noting that DL, especially CNNs and recurrent neural networks (RNNs), have the ability to automatically learn hierarchical representations of features from unprocessed data, such as images, spectra, and sequences. In comparison ML, DL techniques eliminate the necessity for manual feature engineering, which can be a time-consuming process and may result in the omission of pertinent information in complex datasets [22]. Consequently, DL provides a more powerful toolkit for accelerating the discovery, optimization, and prediction of properties of 2D materials. The essential steps of DL are reviewed in detail in the following sections, including data collection, data description, and the selection of DL algorithms.

2.1. Data collection for 2D materials

The effective collection of data is the foundation of successful data-driven decision-making based on DL method, influencing the accuracy, robustness, and ethical considerations of the models generated. At present, data collection of materials, whether experimental or computational, can be sourced from public databases or generated internally [23]. These data can be classified as discrete (e.g., texts), continuous (e.g., vectors and tensors), or in the form of weighted graphs. This section briefly introduces four primary data sources for 2D materials, including material databases, scientific literatures, theoretical calculations, and experimental findings, as depicted in Fig. 3. Additionally, it demonstrates effective methodologies for accessing and collecting 2D material datasets throughout the discovery-to-application process.

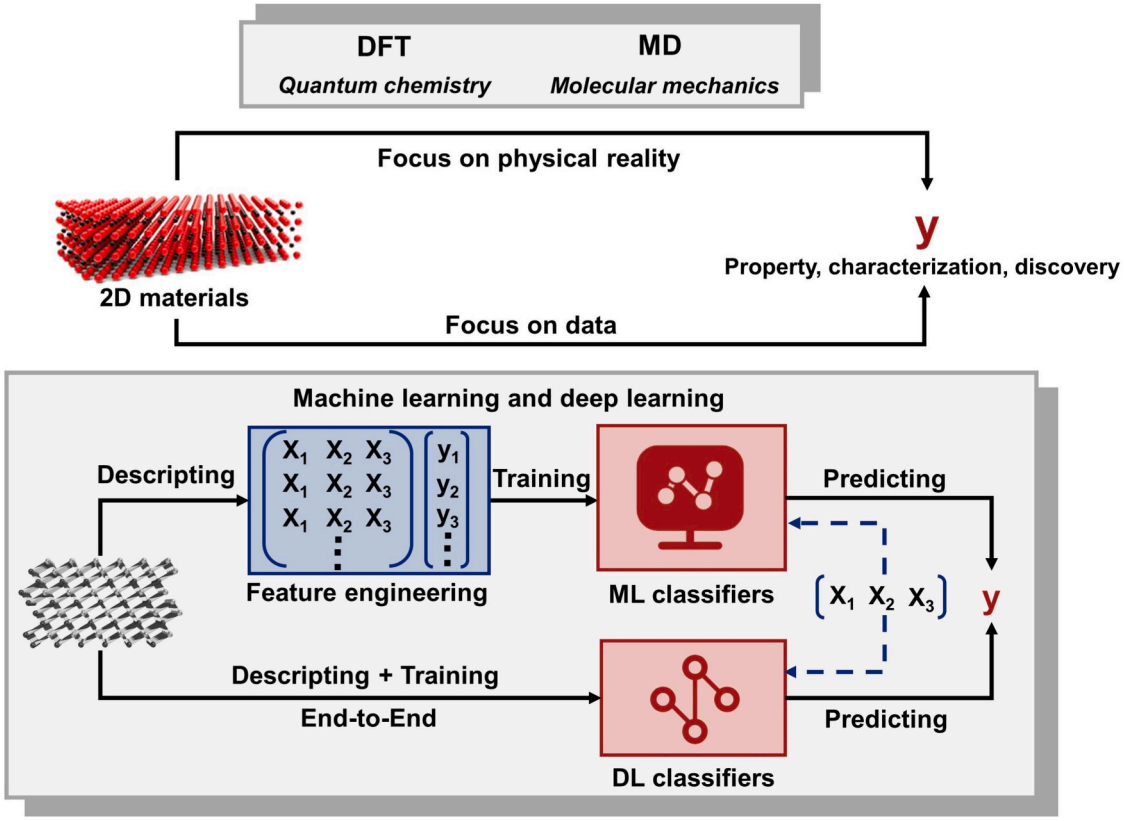

Fig. 2. The differences between simulated calculation, machine learning (ML), and deep learning (DL).

2.1.1. Materials databases for 2D materials

Public databases developed by various countries, including the National Institute of Standards and Technology (NIST) Materials Data Facility, the Materials Project (MP), the Open Quantum Materials Database (OQMD), and the Materials Genome Engineering Database (MGED), play a crucial role in tackling the worldwide scarcity of material data. These databases are of great importance in the development of new materials and are widely used in the field [24]. These databases encompass a diverse range of material types, including both pure and composite substances, as well as their respective crystal structures [25]. As listed in Table 1, the MP database is a leading repository for 2D materials databases, offering a comprehensive range of computed data on known and predicted materials. Its scope encompasses a diverse array of materials, including 2D materials, providing valuable insights into their properties and potential applications [26]. For example, Elrashidy et al. developed three GNNs using 1190 magnetic monolayers with energy above the convex hull (E_{hull}) (<0.3 eV/atom), sourced from the MP database [27]. In contrast, the 2D Materials Encyclopedia (2DMatPedia) is a comprehensive resource dedicated exclusively to 2D materials. It provides detailed information on their structural, electronic, and topological properties, catering specifically to researchers interested in this emerging class of materials [28]. Another noteworthy database is the Computational 2D Materials Database (C2DB), which places a strong emphasis on computational predictions and characterizations of 2D materials. It serves as a repository of theoretical data for the discovery and design of new materials. In addition, the published databases encompass a wide range of computational simulations, including structural, electronic, optical, mechanical, and thermal properties, facilitating research and innovation through computational methods [27–29]. Besides that, Materials Cloud, Crystallographic Open Database (COD), Inorganic Crystal Structure Database (ICSD), and Open Quantum Materials Database (OQMD) are other popular open-source databases of 2D materials that can be consulted for further information from Table 1. The collective resources of these databases provide researchers with invaluable tools for investigating, analysing, and designing 2D materials for a multitude of technological applications, thereby facilitating advancements in materials science and engineering. Moreover, the DL models that are trained using the aforementioned database can be employed as source models for fine-tuning and feature selection-based transfer learning. This approach can significantly enhance the predictive performance when training on smaller datasets [29].

In addition to parametrizing the features, using the image data from published databases represents a viable approach for identifying the optimal 2D materials. Bhattacharya et al. proposed an unconventional yet more inclusive methodology, utilizing band structure images from 2DMatPedia instead of the parameterised band themselves, with the objective of determining 2D flat band materials through the application of a CNN model [28]. It was observed that extant databases lack a comprehensive description of wave functions, which is necessary for the systematic categorization of electronic bands using DL techniques that rely on characteristics such as topology. Nevertheless, the ready availability of band structure images provides a convenient option for the recognition of flat features in the bands. This enables the utilization of these 2D databases even at the initial stages, as the advancement of data science further enhances their information available.

Over the course of several decades, the materials database has undergone continuous development in order to effectively store data pertaining to various materials. However, there are still impediments to the creation of this database. These include a low rate of utilization, high costs associated with collecting, updating, and maintaining the data, a lack of standardised and comprehensive classification standards for materials, as well as methods for evaluating the quality of the data, and concerns regarding intellectual property. Despite the aforementioned challenges, considerable progress has been made in building the 2D material database. The rapid retrieval of material data from the database has been instrumental in overcoming the challenge of limited data and

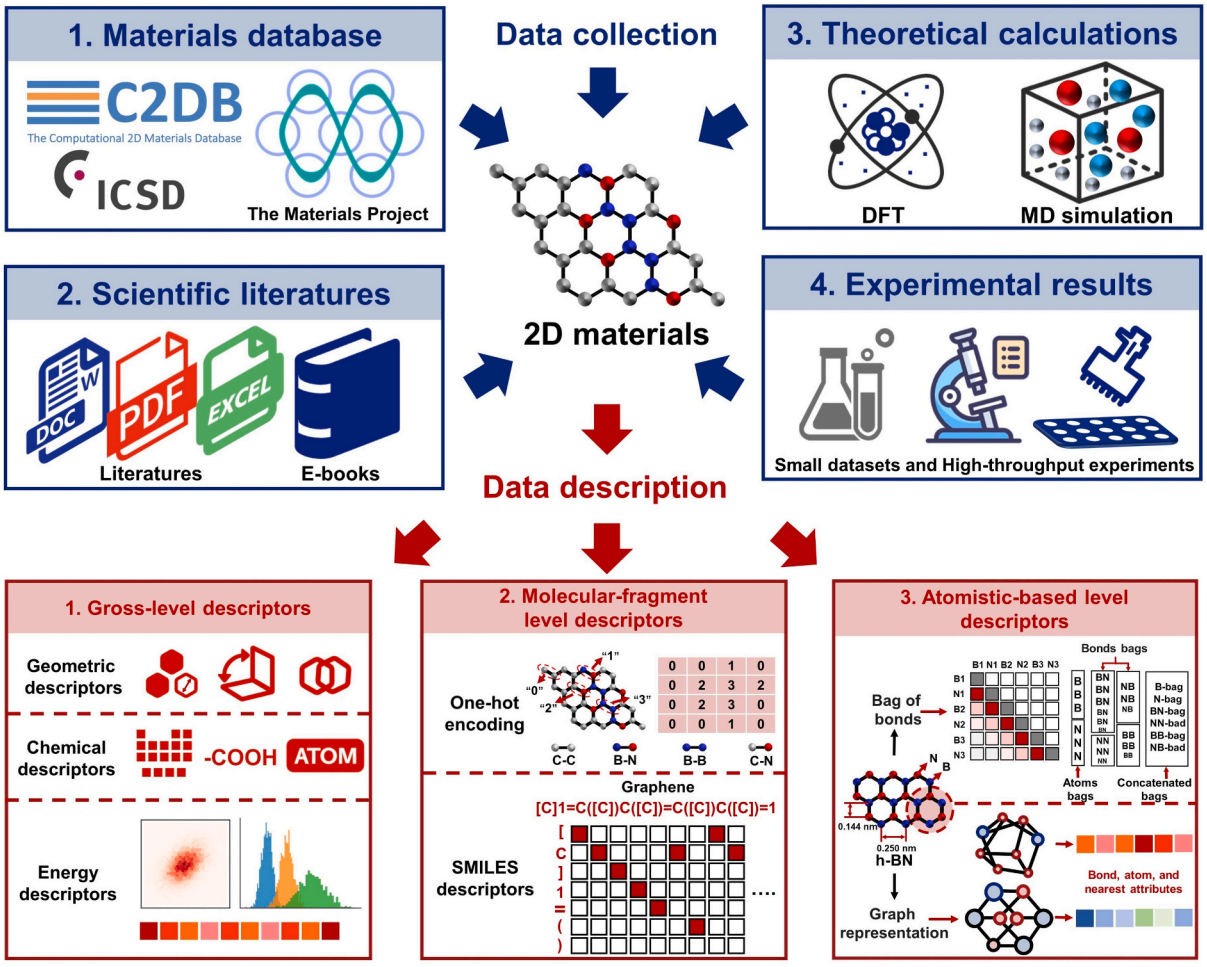

Fig. 3. The data collection from various sources and descriptions from diverse levels for 2D materials.

has become an essential method for collecting information on 2D materials created by DL [12].

2.1.2. *Scientific literatures for 2D materials*

The literature constitutes a further significant data source for 2D materials. The near-exponential increase in the number of published papers provides materials scientists with an extensive resource for constructing datasets [30]. On the one hand, these studies offer valuable insights into structures, properties, synthesis, processing conditions, and other origin data of 2D materials. Additionally, they frequently involve analyses of the connections and relationships between relevant properties [31]. This enhanced comprehension, in contrast to the solely quantitative information and structural data typically found in 2D material databases, offers materials scientists profound insights into the relationships between properties, structures, and descriptors. This, in turn, facilitates the identification of the most effective descriptors for representing 2D materials [32].

At present, the predominant methodology for extracting scientific literature data with 2D materials-related keywords is manual extraction using the Elsevier Science Direct API, CrossRef REST API, and PubMed API [33]. To guarantee the training data accurately reflects the heterogeneous range of optical microscopy (OM) images of 13 distinct exfoliated 2D materials, Han et al. amassed a total of 917 OM images [15]. These images in question were generated by a minimum of 30 users from 8 research groups, utilizing 6 distinct optical microscopes, over a period of 10 years. Following the implementation of color normalization, image resizing, and random rotation, a training dataset comprising 3825 RGB images and a test dataset containing 2550 images were generated from the original 917 OM images. In addition, for the purpose of identifying two-dimensional materials, it is advisable to consider the option of labelling OM images on a pixel-by-pixel basis. Furthermore, the process of data labelling represents a pivotal stage in the context of supervised learning for object detection, with the degree of accuracy associated with the annotations exerting a significant influence on the outcomes of DL models [34]. Roboflow, which has been developed independently, is the most commonly used tool for annotating OM images [15]. It is noted that the images obtained from other research projects should be re-annotated to ensure uniformity in experimental results. Additionally, the primary difficulties in utilizing spectral data (X-ray diffraction (XRD), Raman, and infrared spectra) from different sources for 2D materials are the restricted availability and uneven distribution. Despite the availability of extensive databases such as the Open Raman database, XRD and Chemistry database, RRUFF Raman, and the SOP spectra library, the data remains constrained by inherent limitations and inconsistencies [35]. The effective resolution of inconsistencies in spectral data from disparate sources can be achieved through the utilization of data normalization and standardization, baseline correction, spectral alignment, spectral smoothing, and the combination of chemometric techniques [36].

In recent years, a plethora of accessible and specialised datasets have been published in academic literature. Notable examples include the 2D Materials Defect datasets [37] and the Quantum Point Defect (QPOD) [38]. Besides the organised and accessible datasets, there is a considerable quantity of data distributed across a multitude of literature sources, the majority of which are unorganised and underutilised. A multitude of research initiatives have been initiated with the objective of

Table 1
List of popular structure and property databases used in 2D materials screening.

Name	Descriptions	Sizes	Data sources	URL/Reference
Computational 2D Materials Database (C2DB)	Specifically includes a wide range of computational properties for 2D materials, covering metals, semiconductors, insulators, and more.	2D materials:16 k	Density functional theory (DFT) and many-body perturbation theory	https://www.cmr.fysik.dtu.dk/c2db/c2db.html
Materials Cloud two-dimensional crystals database (MC2D)	Results from screening known 3D crystal structures finding those that can be computationally exfoliated, producing 2D materials candidates.	2D materials: 3 k	Experiments and theoretical calculations calculation	https://www.materialscloud.org/discover/mc2d/dashboard/ptable
aNANt	Sharing the structures and electronic properties of computationally designed 2D layered materials in a single platform.	MXene: 23 k Octahedral 2D materials: 3 k	Density functional theory (DFT) calculation	https://anant.mrc.iisc.ac.in/
2D Materials Encyclopedia (2DMatPedia)	A web interface that provides computational data of 2D materials from top-down and bottom-up approaches.	2D materials: 6 k	Experiments, theoretical calculations, and literature reviews	http://www.2dmatpedia.org/
Stacking of two-dimensional (2D) materials	A comprehensive and systematic overview of 8451 unit cell commensurate vdW homobilayers created by combining 1052 monolayers in various stacking configurations.	Stable bilayer systems: 2.5 k	Density functional theory (DFT) calculations and experiment validations	[162]
Van der Waals bilayer materials	A computational bilayer materials dataset containing 760 structures with their structural, electronic, and transport properties.	Bilayer materials: 760	High throughput Density Functional Theory (DFT) calculations	[163]
Materials Project	Containing the charge densities, EXAFS, XANES, tensor properties, density of states, band structures, molecules, crystal structures materials database.	Materials: 169 k; Molecules: 577 k	Experiments and theoretical calculations	https://www.materialsproject.org
JARVIS-DFT	Containing the various properties of materials, including the formation energies, bandgaps, elastic, piezoelectric, dielectric constants, and magnetic moments, exfoliation energies for van der Waals bonded materials, and so on.	Materials: > 80 k	Density functional theory (DFT) predictions	https://jarvis.nist.gov/jarvisdft
Inorganic Crystal Structure Database (ICSD)	Containing experimental inorganic structures, experimental metal-organic structures, and theoretical inorganic structures.	Crystal structures: 300 k	Scientific publications, crystallographic databases, experimental studies, and curated datasets	https://www.psds.ac.uk/icsd
Crystallography Open Database (COD)	Open-access collection of crystal structures of organic, inorganic, metal-organic compounds and minerals, excluding biopolymers.	Crystal structures: 520 k	Research publications	https://www.crystallography.net
Cambridge Structural Database (CSD)	Providing a comprehensive repository of experimentally determined molecular crystal structures.	Molecular crystal structures >1 M	Experiment X-ray crystallography, electron diffraction or neutron diffraction	http://www.ccdc.cam.ac.uk/
Aflow	Maintaining an ample repository of quantum, thermal, structural and elastic properties of inorganic compounds, the high-throughput open-source code used to characterize them, and the online post-processing tools and infrastructure to analyse the data.	Material compounds >3.5 M	High throughput computing	https://www.aflowlib.org

gathering data from literature sources for the purpose of data mining [39]. Nevertheless, extracting information manually from this narrative is a time-consuming and error-prone task, especially considering the way factual data is presented in a test format. With the rapid increase in unstructured and disorganised data on the internet, natural language processing (NLP) has become a valuable tool for automatically extracting and organizing information from large text datasets [30]. Numerous openly accessible NLP toolkits have been developed for the purpose of parsing and extracting chemical information from scientific documents, including ChemcialTagger [40], ChemSpot [41], and LeadMine [42]. However, many of these toolkits have been created to target only certain parts of a scientific paper. ChemDataExtractor represents a departure from other toolkits in that it employs physical-quantity models to analyse scientific text and populate a user-defined ontology of chemical information. This approach results in a well-organised database of materials and their properties, making it easier to analyse using data science techniques [43]. ChemDataExtractor can not only extract data from text but also parse tables to retrieve information. It utilizes a specialised version of the natural language processing pipeline in order to extract information from tables. In this process, each row represents a specific chemical entity, and each column describes the attributes of that entity. By treating each cell in the table as a concise and highly structured sentence, it can achieve the successful extraction of information [44]. To date, ChemDataExtractor has been successfully applied in automatically generating chemical data from scientific documents, including studies on thermoelectric properties [45], yield strength and grain size [46], and a database on semiconductor band gaps database [47]. Despite the availability of reliable data sources, the prevailing approach for obtaining information from published materials remains manual. This is mainly due to the difficulties posed by various factors, such as inconsistent material naming conventions, complex chemical compositions, the use of multiple languages, and the abundance of specialised terminology. These challenges make it hard to apply NLP and text-mining technology for automated data extraction from publications. It would be advantageous to investigate methods that facilitate semantic compatibility of data across text, tables, and figures in the future. Furthermore, the creation of standardised formats for data presentation could significantly enhance the efficiency of data extraction and analysis [30].

2.1.3. Theoretical calculations for 2D materials

The databases and literature data are currently available to constitute an invaluable foundation for DL models. Nevertheless, it is imperative to conduct supplementary experiments or calculations to reinforce the data, particularly when there is a paucity of information regarding a specific property of interest [48]. Theoretical calculations based on DFT calculation and MD simulation have been pivotal in the rapid development of 2D materials [9,49]. These studies usually focus on one of the subsequent areas: (1) inherent physical characteristics, (2) adjusting properties by applying strain, doping, defects, or chemical functional groups, (3) examining the interaction between 2D materials and other substances, and (4) exploring potential applications [50]. The utilization of DFT calculations offers a valuable methodology to investigate and verify the magnetic, optical, and electronic characteristics of predicted materials through the analysis of electron energy states in accordance with quantum mechanical principles. On the other hand, MD

simulations offer detailed information on the thermodynamic and kinetic properties of systems at the atomic level, which may be challenging or impossible to obtain through experimental methods based on classical Newtonian mechanics. These tools can be regarded as a supplementary means of corroborating experimental outcomes and establishing a theoretical basis for observed phenomena [51]. So far, numerous studies using theoretical calculations have been carried out for 2D materials where properties such as mechanical properties [52], atomistic physical fields [53], and band structure parameter prediction [54].

When DFT calculations, users are faced with critical decisions regarding the approximations to be employed. Firstly, the incorporation of a suitable exchange and correlation function, including the local density approximation (LDA), the generalised gradient approximation (GGA), and hybrid functions, enables the resolution of complexities that arise from the quantum behaviour of electrons through the application of DFT calculations [10,48,49,55,56]. Secondly, the user needs to choose a basis set, which consists of functions utilised for expanding the Kohn-Sham orbitals. The selection of a basis set involves a balance between computational cost and accuracy [48]. In the study of 2D materials, researchers frequently employ a range of basis sets. These include the plane wave basis set [57] and the projector-augmented-wave pseudopotential or ultrasoft (US) pseudopotential [58]. In addition, there are numerical atomic-orbital basis set plus p-polarization functions (DNP) or d-polarization functions (DND) [59]. Furthermore, the Perdew-Wang (PW91) and Perdew-Burke-Ernzerhof (PBE) functions are the most commonly employed exchange and correlation functions with the GGA [60], while the Ceperley Alder (CA) or Troullier-Martins (TM) functions are frequently utilised in the LDA [61]. In general, DFT calculations are capable of providing accurate qualitative and, in some cases, quantitative information on the structural geometries, vibrational properties, and electronic structures of molecules, crystals, and surface systems. Nevertheless, the precision of this approach is limited by the inherent inaccuracies of exchange and correlation functions and the difficulty in accurately representing dispersion interactions, such as van der Waals forces [55]. Until now, DL methods have demonstrated the capability to accurately predict the oxidation and spin states. Moreover, DL techniques prove effectively in overseeing geometry optimizations, ensuring the convergence towards meaningful resultant structures.

The computational cost of DFT and other quantum mechanical techniques can become prohibitive for systems with large numbers of atoms. As an alternative, MD simulation uses classical Newtonian dynamics to simulate the interactions between atoms and molecules. These interactions are characterised by a potential energy function, which determines the forces experienced by the particles [62]. In order to accurately model 2D materials, the potential energy function must accurately replicate their lattice structures and phonon dispersion relations, ideally matching those observed in experiments or calculated using DFT. In some cases, forces can be derived directly from electron energy states obtained through ab initio calculations. While DFT calculations typically involve systems of fewer than one hundred atoms, MD simulations are capable of handling significantly larger systems with thousands to tens of millions of atoms [49]. A variety of multiple potential energy functions or parameter sets are available for specific materials. To model 2D nanomaterials, some generalised force fields, including AMBER potential, GROMACS, and COMPASS, have been employed. Nevertheless, the general nature of these potentials renders the accuracy of material property predictions uncertain [63]. The Stillinger-Weber (SW), Tersoff, and Reactive empirical bond order (REBO) potentials are the most commonly used and extensively developed potentials for studying 2D materials [64]. Although the modified embedded atom method (MEAM), reactive force field (ReaxFF), and valence force field (VFF) have been used to model certain 2D materials, their accuracies have not been thoroughly tested and are limited to studying specific properties of these materials [65,66]. The detailed information and description of these potential energy functions can be referred to by Liu and Zhou [50]. In addition, MD methods can also aid in database screening to assess the water and thermal stability of 2D materials and can be trained to predict various other properties of interest [52,67].

Furthermore, developments in high-performance computing and computational simulation tools have enabled the use of first-principles-based high-throughput calculations (HTC) to efficiently discover and design 2D materials. MatCloud, Aflow, FireWorks, the joint automated repository for various integrated simulations (JARVIS), ASE, AiiDA, and SEHC are the common toolkits for HTC in materials science [68]. Taking Aflow as an example, Aflow is an automated HTC materials discovery framework developed by Duke University, which can be used to perform HTC on the properties of alloys, inorganic crystal structures, and other materials, even for the 2D materials [69,70]. Moreover, JARVIS is a comprehensive platform designed to expedite the process of discovering and designing materials. It encompasses electronic structure methods, classical force fields, ML techniques, quantum computation algorithms, and experiments [71]. This integrated infrastructure plays a crucial role in minimizing the cost and time required for the discovery, optimization, and deployment of 2D materials. Moving forward, the next frontier in 2D materials goes beyond simply refining simulations to gain a deeper understanding of these materials' behaviour under various conditions along with the incorporation of innovative DL techniques and advanced computational resources.

2.1.4. Experimental results for 2D materials

A recent surge in research has focused on the utilization of DL with datasets obtained from numerous experiments to prepare and characterize target 2D materials with greater precision. While NLP methods discussed in Section 2.1.2 can obtain experimental datasets from scientific literatures, the sensitivity and diversity of 2D materials to synthetic processes and characterization methods can lead to inconsistencies in data sources, potentially impacting data accuracy. Furthermore, it is common practice for literature to report only successful outcomes, with any data indicating failure being deliberately omitted. In this context, experimental data represents a vital source of information for the creation of datasets for DL models. Advanced microscopy and/or spectroscopy, as essential tools for 2D materials characterization, can offer valuable insights into the synthesis mechanism, crystallography, chemical compositions, and physicochemical properties [72,73]. One study reported by Masubuchi et al., an autofocus microscope was developed to screen for 2D materials on SiO_2/Si substrates with varying layer thickness [74]. The microscope utilised a motorized XY scanning stage to capture approximately 2000 OM images including graphene, MoS_2, WTe_2, and h-BN flakes. Subsequently, these OM images were classified into three categories using an online website labelling tool before being inputted into the networks. Interestingly, the prediction results remained consistent even when the illumination conditions were changed, as the deep neural network (DNN) was able to identify 2D crystals based on the high-dimensional and hierarchical characteristics of the images. Furthermore, the utilization of the data augmentation technique is extensively employed in DNN networks to enhance the accuracy of learning and mitigate the issue of overfitting. To illustrate, Saito et al. effectively augmented the training data by randomly applying a series of image processing techniques, including cropping, flipping, rotating, and altering the color of the initial images [75]. As a result, the training dataset grew significantly, expanding to include 960 data points sourced from 24 OM images of MoS_2. Similar methods can be referred to in Mahjoubi et al. [73] and Ramezani et al. [34].

Although OM has been frequently employed to differentiate between monolayer and few-layer 2D flakes, accurately mapping the layers solely based on RGB images is a formidable challenge, largely due to the visual similarities between few-layer flakes. It is thus possible to obtain supplementary dimensional data by employing alternative imaging techniques, such as hyperspectral imaging, to differentiate specimens of varying thickness. One instance of combining spectroscopy and imaging

techniques is hyperspectral imaging microscopy, which offers both spatial and spectral information about a measured region. The rich spectral data can lead to more accurate identification of layer numbers compared to traditional microscope imaging. By combining the detailed profile information from RGB images with the abundant spectral information from hyperspectral reflection images, accurate mapping of atomic layers in 2D materials can be achieved [76]. Nevertheless, the advancement of effective 3D semantics learning and the integration of images of disparate dimensions into a unified training process is impeded by two principal challenges. Therefore, it is crucial to identify effective DL techniques to overcome these obstacles. Furthermore, Raman spectroscopy is a rapid and non-invasive technique that provides distinctive benefits in the structural analysis of 2D materials [77]. It is essential to acknowledge that the spectral peaks of 2D materials can exhibit considerable variability across different experimental setups. This is due to the fact that each material may undergo analysis in a multitude of states, which can lead to discrepancies in the observed spectral characteristics [78]. In practical applications, it can be difficult to differentiate the faint signals of trace substances from the background noise when using Raman spectra analysing methods. This obstacle impedes the detection of the desired substance signal and restricts the amount of spectral data gathered in experiments [32]. To address this problem, DL methods provide an alternative approach to traditional spectral data analysis. Rather than interpreting the spectra as being convoluted with distinct peak positions and intensities, DL architecture view the spectra data as a continuous pattern, much like an image. This strategy has been successful in overcoming the challenges mentioned above [32,76].

Currently, the absence of a complete database that includes synthesis parameters for 2D materials is a major obstacle for DL in researching ways to control the production of 2D materials. In the future, high-throughput experimental (HTE) methods combined with machine robots enable multiple chemical experiments to be conducted simultaneously through automated processes and various routine chemical workflows, which can effectively address the challenge in small datasets. The combination of DL and HTE offers numerous opportunities for exploring the chemical space. In addition, it is essential to have an electronic laboratory notebook for researchers to effectively plan, describe, store, and manage their daily work during the utilization of HTE methods.

2.2. Data descriptors for 2D materials

Descriptors, also called factors, features, or fingerprints, represent material properties from a particular perspective. The heterogeneous nature of data storage, with data residing in disparate databases and formats, presents a significant challenge in the integration of data from multiple sources [23]. Additionally, the required data format depends on the DL used. Therefore, it is essential to standardize data formats and select appropriate data representations for DL algorithms during data processing. In summary, from the perspective of descriptor forms, the descriptors used in the analysis of 2D materials can be categorised into three levels, namely gross-level, molecular fragment level, and atomic-based level descriptors (Fig. 3). These descriptors characterize the properties of 2D materials at various physical scales, including the entire material, fragments, and individual atoms [19,79]. As the amount of descriptive information decreases, so does the complexity of interpreting the descriptors also decreases, which poses a challenge for researchers in direct interpretation. This means that more descriptors are needed to meet the information requirements, making the process less efficient. Therefore, a key challenge in the development of DL-assisted 2D materials lies in effectively utilizing the gross-level, molecular fragment level, and atomic-based level descriptors to construct models that are both accurate, rational, and easy to understand.

2.2.1. Gross-level descriptors

Gross-level descriptors, being reliable and intuitive, provide an overview of the 2D material characteristics, including factors such as geometric structure, elemental properties, and energy. Based on that, gross-level descriptors can be classified into chemical descriptors, geometric descriptors, and energy descriptors. Among these, chemical descriptors play a crucial role in summarizing the 2D chemical characteristics of materials. These descriptors include various aspects such as the fraction and number of elements [80], element mass or thermochemical electronegativity [81], band structure parameters, interlayer asymmetry, chemical potential, charge density, as well as functional groups and active sites [79]. The selection of functional forms for creating interatomic potentials and fitting potential energy surfaces depends on choosing a representation of atomic neighborhoods' that closely aligns with the properties of 2D materials. For example, the mechanical properties of graphene oxide (GO) are affected by the total amount and relative ratio of hydroxyl and epoxide groups due to the numerical and distributional advantages in the graphene basal plane (GBP). Thus, the number and distribution of hydroxyl and epoxide groups as chemical descriptors in GBP can be used to investigate the toughness of GO by using the one-hot encoded method as input to the trained DL model [82]. Furthermore, a CNN model was created to effectively forecast the formation energy of defective graphene and MoS_2 on a large scale by incorporating the chemical bond parameters into 2D materials [83].

Geometric descriptors offer a comprehensive representation of the aperture structure of 2D materials. They are especially useful in detecting defects that result in percentage variations in properties when compared to the original bulk material, such as alterations in the average atomic radius [84]. Utilizing percent differences as features, instead of absolute values, allows for easier comparison of defect structures among material systems with varying values. Common geometric descriptors for 2D materials descriptions include translation vectors, radial distribution functions, bond lengths, bond angles, lattice parameters, surface area, shape and symmetry, and interatomic distances [85,86]. In addition, a research study employed the structural factors (SF) found within inorganic octahedron layers as a means to describe the structure of 2D hybrid lead-halide perovskites. SF serves not only as a feature to depict the connection between structure and performance, but also as a tool to explore the entire range of material structures [87]. In addition, topological data analysis (TDA) methods enable the effective geometric description and deduction of topological information related to the complex micropore environments found in porous materials. These methods allow for the characterization of micropore typology, morphology, and connectivity [88,89]. However, there are still difficulties in developing featurization schemes for crystalline porous materials such as graphene. These challenges arise from the need to consider both the representation of element atomic monomers and the periodicity/symmetry of the crystalline frameworks.

Energy descriptors are made up of specific energy values or various energy distributions combined together. The level of detail provided by each descriptor can vary greatly depending on the type of energy being described. Some energy descriptors have a clear and significant physical significance on their own, while others may require multiple descriptors to fully convey all the necessary information [23]. For instance, to evaluate the thermodynamic stability of the materials produced, a standard method involves calculating the energy of the convex hull (E_{hull}) based on the Gibbs free energy at 0 K and 0 atm pressure. This is achieved by creating a E_{hull} set of the normalised formation energy relative to the number of atoms, offering a more advantageous way to assess the energetic favourability compared to utilizing the formation energy per atom (E_f) [27]. Therefore, E_{hull} has become a widely used metric in various data-driven methods for evaluating the thermodynamic stability of generated materials. Furthermore, another common energy descriptor is formation energy, a widely used measure of energy that represents the change in energy when a compound is formed from

its constituent elements in their reference states. This measure is usually expressed in terms of energy per atom or energy per formula unit and serves as a key determinant of the thermodynamic stability of a 2D material. When used as an input feature, formation energy can help in building models that predict other properties, such as electronic band structure, mechanical strength, and chemical reactivity. The Gibbs free energy as the thermodynamic can be used as one of the hydrogen adsorption descriptors [90]. Furthermore, the fracture strain, fracture strength, Young's modulus, shear modulus, Poisson's ratio, and tensile strength as mechanical energy descriptors are applied to characterize 2D materials in predicting mechanical properties [67,82,91]. A study used temperatures, strain rates, vacancies, directions, and velocity seeds as gross-level descriptors generated by MD simulation to obtain the fracture strain, fracture strength, and Young's modules post-processed by MATLAB script [67]. Lastly, for 2D electronic properties description using energy-based level, electronic-density-of-states, band structure, the projected densities of states (PDOS), and the Kohn-Sham band-grp were shown to improve the electronic properties prediction of 2D materials [86,92,93]. Certainly, the incorporation of 2D materials properties that necessitate tensorial representations, such as stiffness, heat conduction, and susceptibility tensors used for predicting mechanical, heat transport, and magnetic/electronic properties, poses a challenge in parameterizing and implementing them in DL.

2.2.2. Molecular-fragment level descriptors

Molecular-fragment descriptors are generated by identifying the building blocks of 2D materials. For instance, graphene is a crystalline form of carbon consisting of layers of carbon atoms arranged in a hexagonal lattice. Because each carbon atom is bonded to three other carbon atoms in a flat structure, molecular-fragment descriptors can be calculated by considering the contributions of each group of atoms that form the repeating unit. In this way, creating blocks and topological frameworks are employed to analyse molecular fragments in 2D materials [23]. Generally, the molecular fragment descriptors are mainly represented in the form of encoding. This means that the creating blocks or topologies are depicted in binary form, specifically through the use of one-hot encoding. In this encoding, the presence of these building blocks or topologies in a specific 2D material is indicated by a value of 1, while their absence is indicated by a value of 0. To illustrate, when predicting or calculating the physical field associated with deformations on atoms or bonds, the non-atom pixels were assigned a value of 0 (black), while the on-atom pixels represented the normalised field values. For atoms, a square block of 3×3 pixels was used and centred on the location of the atom. Similarly, for bonds (such as bond length or hopping energy), the centres of these 3×3 pixels were positioned at the centres of the bonds. [53] Similarly, Dong et al. proposed a new method for characterizing hybridised graphene structures by representing them as a 2D matrix [94]. In this matrix, "0" and "1" represented C—C and B—N pairs, respectively. The interactions between neighboring atoms in these structures play a crucial role in determining the bandgaps of the entire structure. This simplified representation, akin to a 2D image, greatly streamlines the learning process for CNNs and ensures accurate bandgap predictions not only for hybridised graphene but also for other 2D materials. The interactions between atoms within these structures collectively determine the bandgaps of the entire material, highlighting the importance of considering the influence of each atom on its neighboring atoms. Similarly, Dong et al. labelled the optical RGB images of h-BN with 0 for the background pixels, 1 for the monolayer pixels, and 2 for the bilayer pixels [76]. Following this procedure, the labels were transformed into one-hot images to generate binary target masks within the network (Fig. 3).

Other molecular-fragment level descriptors, such as Ewald sum matrix [95], Smooth Overlap of Atomic Positions (SOAP) [96], Many-body Tensor Representation (MBTR) [97], Coulomb matrices [98], and CrystalNNFingerprint [99], typically result in a fixed length vector representing the structure of 2D materials. Take CrystalNNFingerprint as an illustration, it excels at preserving a greater amount of information when there is a large number of atomic sites. This makes it particularly well-suited for detecting local coordination patterns in crystals, even when there are minor lattice distortions. Specifically, CrystalNNFingerprint utilizes a neighbor-finding algorithm based on Voronoi decomposition to determine the local environment for each atomic site. This coordination pattern is compared to different coordination templates known as Local Structure Order Parameters (LoStOPs). In this way, each atomic site is assigned a 61-dimensional fingerprint vector through comparison. In addition, through mathematical statistics, other values of each local CrystalNNFingerprint within the structure are then used to create a global fingerprint representing the overall structure. This ultimate structure characteristic is located in a vector space with 244 dimensions [28]. In addition to these descriptors, the simplified molecular input line entry system (SMILES) descriptors are a way of representing a chemical structure using a string of characters, allowing for a concise and unambiguous description of the structure of molecules [20]. 2D materials, such as graphene, h-BN, and TMDs, have unique properties and structures that can be described using SMILES strings. However, due to their extended, often periodic nature, the representation of 2D materials using SMILES can be challenging [100]. Based on that, extending SMILES for periodicity is an effective method to address this issue. For example, by using a notation that explicitly denotes repeating units, or by applying techniques similar to those used in crystallographic information files (CIF). In addition, using recursive SMILES to represent repeating units and their connection can encapsulate the periodic nature of 2D materials with a hierarchical structure.

2.2.3. Atomistic-based level descriptors

Descriptors at the atomistic level focus on the specific physical properties of individual atoms, including their characteristics and their positions within materials. Because of the constraints of the small scale, each individual microphysical quantity carries minimal characteristics. In order to meet all requirements with the limited data available, it is necessary to have hundreds of millions of descriptors. Complicated inputs may not always improve efficiency, and imperfect characteristics can lead to unreliable models. Atomistic-based level descriptors play a crucial role in efficiently describing atomic systems with minimal loss of information. These descriptors are carefully designed to be elegant, providing a unique representation of structures while remaining unaffected by rotation, translation, and alignment. Thanks to their exceptional clarity in conveying information, atomistic-based level descriptors offer unparalleled accuracy compared to other descriptors. However, their application in 2D materials research has been somewhat restricted due to the need for a certain level of expertise. At present, many researchers are still unsure of the reasoning behind comparing multiple atomic-based level descriptors at once, despite some studies doing so. In this part, we will examine the commonly used atomic-based level descriptors in 2D materials, with a primary focus on utilizing bag-of-bonds and graph representation methods.

The bag-of-bonds (BoB) method is a representation technique used in materials informatics to describe the local atomic environment of materials for ML applications [101]. The BoB method transforms the complex atomic structure of a material into a fixed-length feature vector by considering pairwise atomic interactions (bonds) within a certain cutoff radius. Each bag contains counts or properties of specific types of bonds, such as bond lengths or bond angles [102]. h-BN has a hexagonal lattice structure similar to graphene, with boron (B) and nitrogen (N) atoms alternating in a 2D plane. Using the BoB method to represent h-BN can be summarised as follows. Identifying relevant atom pairs with cutoff radius (B—N, B—B, and N—N), calculating features like bond lengths, creating histograms for each type of interaction, and combining these histograms into a features vector. Although the BoB method provides a simple, efficient, and flexible to represent 2D materials, the high dimensionality, choice of cutoff radius, data sparsity, and complex bonding environments are the main challenges that can affect its

effectiveness and applicability.

To our knowledge, crystal structures can be represented as graphs where atoms are represented by vertices and chemical bonds are represented by edges [103]. Taking h-BN and MoS_2 as examples, they can be represented graphically by considering the atomic numbers of the chemical elements involved and the spatial distance between atoms. In this way, fragment descriptors define subgraphs within the complete 2D molecular network. Any invariant of a molecular graph can be expressed as a distinct linear combination of fragment descriptors. These descriptors have numerous benefits compared to other chemical descriptors, such as ease of computation, storage, and understanding. Nonetheless, they also have some drawbacks. When faced with new fragments that were not included in their training, models based on fragment descriptors exhibit poor performance. Moreover, typical fragments are created using only the characterization of single atomic elements like C, N, and Na. This limited representation is inadequate to accurately predict the intricate chemical interactions occurring within materials. Considering that, Isayev et al. introduced a novel concept of creating a fragments graph labelled with properties, which proved to be highly beneficial for the topological analysis of MOFs, molecules, and inorganic crystals [104]. Specifically, the crystal structure was analysed for atomic neighbors via Voronoi tessellation, which defined the connectivity within the material meeting several criteria. After obtaining the total list of connections, the full graph and its corresponding adjacency matrix were created. The global structure of a given system, which includes interatomic bonds and contacts within the crystal, is represented by this adjacency matrix. Additionally, the complete graph was divided into smaller subgraphs that correspond to individual fragments. The materials fragments with property labels were categorised based on local reference properties, which included general properties, measured properties, and derived properties. In order to include information about the shape, size, and symmetry of the crystal unit cell, the following crystal-wide properties were taken into account, including lattice parameters, their ratios, angles, density, volume, number of atoms, number of species (atom types), lattice type, and space group [105]. Despite the increasing studies and publications on the development and application of atomistic-based level descriptors, they are still considered to be expert tools, requiring expert knowledge to better describe the states, structures, and attributes of materials. It is crucial for reliable quantitative comparisons of future developments to have widespread use and continued improvement of standard benchmarks.

2.3. Deep learning models

After defining a descriptor vector for each chemical structure, the design matrix and kernel matrix can be constructed for a set of 2D material structures. These matrices can serve as inputs for ML modelling. DL has capitalised on the exponential growth of data and the continuously increasing computational power. One key distinction between DL and ML approaches lies in the flexibility of neural network architecture. Single-layer neural networks have been traditionally used in quantitative structure-activity relationships. However, with the growth of data size and computational capabilities, there has been a natural progression towards the adoption of multilayer feed-forward networks for physicochemical property prediction of 2D materials. In recent years, the utilization of RNNS in de novo design using chemical element sequences has a somewhat unexpected development. Moreover, with adoption of advancing optical microscopy imaging equipment, CNNs have gained remarkable success in computer vision and have become a good choice for 2D material image processing. Furthermore, GNN and GAN models have taken advantage of the effective representation and inverse design of 2D materials. In this section, various DL architectures that have been in use so far in the characterization, design, and discovery of 2D materials are discussed, as shown in Table 2.

2.3.1. Multilayer perceptron (MLPs)

A drawback of deep neural networks is that they must be trained by fitting millions of parameters, which can result in slow training times. Conversely, MLPs only need to fit a few thousand parameters and can be trained in a matter of minutes [8]. As shown in Fig. 4A, it consists of a layered architecture, with neurons organised into an input layer, one or more hidden layers, and an output layer. Importantly, the neurons in the hidden and output layers employ non-linear activation functions such as ReLU (Rectified Linear Unit), Sigmoid, or Tanh. These functions introduce non-linearity into the model, enabling it to capture intricate patterns in the data [106]. Moreover, MLPs are trained using the backpropagation method, in which the error between the prediction and the actual output is calculated. This error is propagated backward through the network to compute gradients of the error with respect to each weight and bias. These gradients are used to update the weights and biases using an optimization algorithm like gradient descent, reducing the error iteratively. By utilizing non-linear activation functions and multiple hidden layers, MLPs can capture complex patterns and relationships in data that simpler models might miss [52]. However, MLPs are limited to working with low-dimensional representations of data. These representations must be computationally simple yet capable

Table 2
The characteristics of different deep learning frameworks for predicted 2D materials.

DL framework	Primary focus	Key features	Advantages	Limitations	Examples
Graph neural networks (GNNs)	2D material structure-property prediction	Captures relationships between atoms and bonds, suitable for crystal graphs	Accurate representation of atomic interactions, suitable for diverse 2D materials	Computationally intensive for large-scale systems	Predicting band gaps, adsorption energy [27]
Convolutional neural networks (CNNs)	Image-based 2D material analysis	Works with material microstructure images or spectroscopic data	Efficient in recognizing spatial patterns and features	Requires extensive labelled image datasets	Analysing TEM/SEM images, phase identification [15]
Recurrent neural networks (RNNs)	Sequential data modelling	Processes time-dependent or sequential data like molecular dynamics trajectories	Effective for sequential property predictions	Limited ability to handle non-sequential features	Predicting thermal stability in dynamic systems [119]
Transformers	Global attention mechanisms	Handles long-range dependencies in data	High accuracy for large datasets, scalable	Computationally demanding, needs large datasets	Analysing textual descriptions or atomic sequences [18]
Variational autoencoders (VAEs)	Generative modelling	Learns probabilistic representations of materials	Generates new 2D material candidates, models uncertainties	May generate physically unrealistic structures	Predicting novel 2D material candidates [110]
Generative adversarial Networks (GANs)	Generative design	Generates new 2D material structures by training on known data	Capable of creating innovative 2D material designs	Challenging to train and evaluate stability	Creating hypothetical 2D materials for screening [139]
Hybrid Frameworks	Multimodal data fusion	Combines features from multiple models (e.g., CNN + GNN)	Captures diverse material properties simultaneously	Complex design and integration challenges	Integrating spectroscopic and structural data [18]

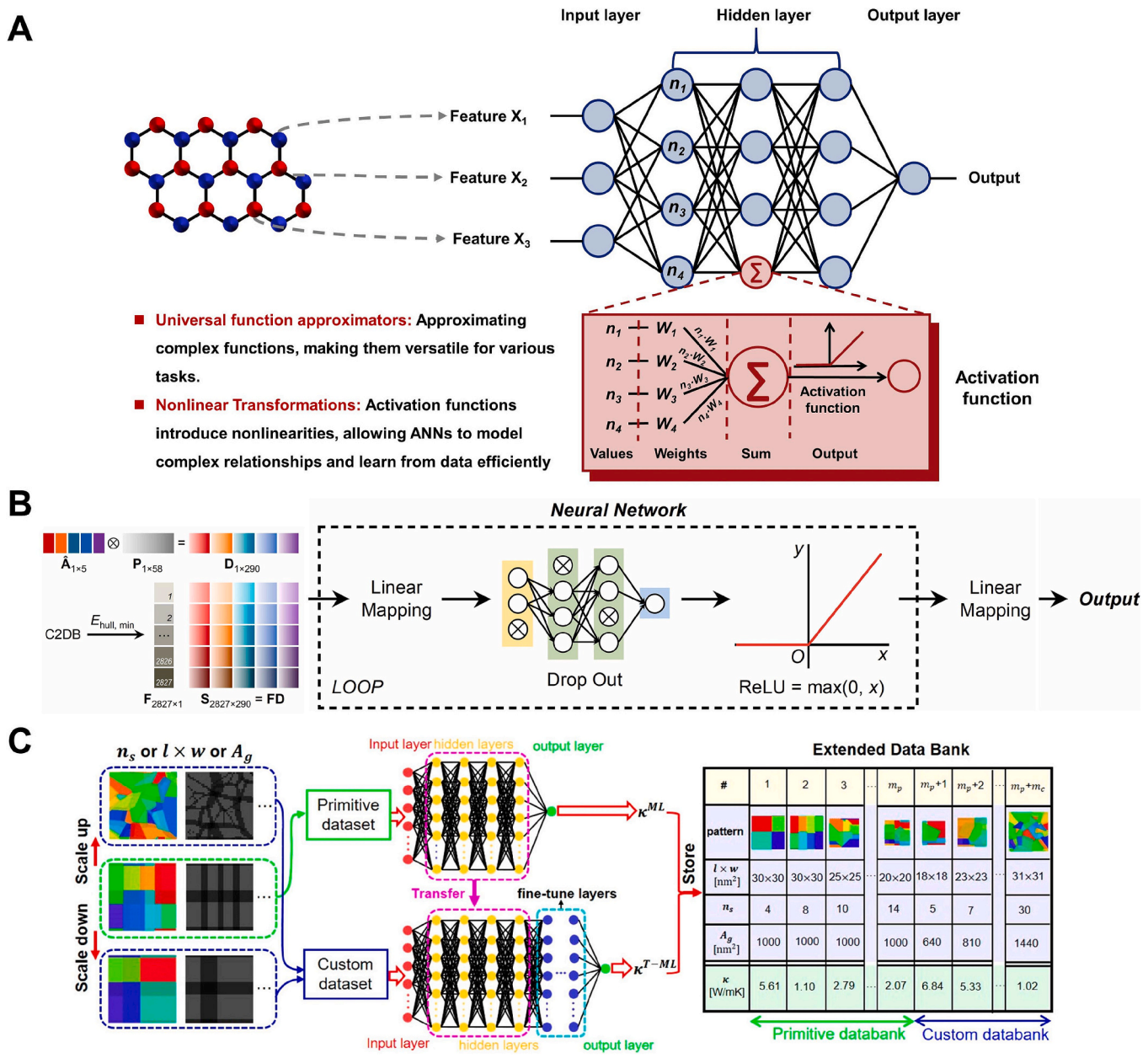

Fig. 4. Multilayer perceptron (MLPs) and their application in 2D materials screening. (A) Basic framework and principles. (B) The process of obtaining the descriptors and datasets for crystal system and space group prediction using MLPs. Reproduced with permission from Ref. [79]. Copyright 2023 Elsevier. (C) Transfer learning of the MLPs model for the thermal conductivity prediction of piled graphene structures. Reproduced with permission from Ref. [107]. Copyright 2021 American Chemical Society.

of capturing the maximum variation among different training data points. Based on that, Hundi and Shahsavari used N-point statistics and correlation functions to process the microstructure of graphene in low dimensional physical description to construct MLPs for structure-property relations [14]. Dewapriya et al. used MLPs to investigate the fracture stress of graphene with different defects based on temperature, vacancy concentration, strain rate, and loading direction as input. They showed that MLPs possessed great extrapolation capability and were useful for limited datasets [52]. Moreover, as shown in Fig. 4B, Che, Wang, Lv and Wu [79] created a sophisticated deep MLPs model trained by a 2827 (including 6 crystal systems, 21 point groups, and 54 space groups) × 290 (5 chemical elements × 58 gross-level descriptors) dataset. This model was capable of accurately predicting crystal systems and space groups of 2D materials in AA stacking based on their chemical formulas, achieving an accuracy of 74.56 %. Interestingly, their findings revealed that even with only the top 80 descriptors, reasonably accurate predictions can still be generated, closely resembling the results obtained using all 290 descriptors. This suggested that there were differences in the relative importance of these descriptors in determining the outcome. In addition, due to the non-linear activation functions and multiple-layer architecture, MLPs have good robustness and transferability. As illustrated in Fig. 4C, Liu et al. adjusted the size and scale to create a unique dataset consisting of various stacked graphene structures with more diverse geometric and dimensional characteristics compared to the original dataset to verify the robustness of MLPs in predicting the thermal properties of graphene [107]. In contrast to starting the MLPs training from scratch, they incorporate fine-tuned layers to transfer the knowledge gained from the initial dataset to update the MLPs with the geometric characteristics of stacked graphene structures from a specialised data set. The overall R^2 and RMSE values of the trained MLPs with

the original dataset met the accuracy requirements. Certainly, the prediction accuracy of MLPs is easily constrained by the limited data and uneven distribution of space groups. In the future, expanding the database size, enhancing predictions for low-symmetry and low-data-volume space groups, and uncovering additional correlations between descriptors and targets will be all effective methods.

2.3.2. Autoencoders

Autoencoder networks are a type of feed-forward neural network that aims to reconstruct the input data at the output layer, consisting of multiple hidden layers [108]. As shown in Fig. 5A, the encoder in an autoencoder framework is tasked with understanding high-dimensional input data and transforming it into a lower-dimensional representation known as latent space. On the other hand, the decoder has to reconstruct the original input based on this compressed representation. In particular, similar to principal component analysis (PCA), autoencoders reduce the dimensionality of the input data by compressing it into a smaller-dimensional code space in the hidden layer. This smaller hidden layer size allows the original data to be reconstructed at the output layer. As a result, autoencoders can provide bi-directional mappings, enabling conversion between the data and code space [106]. Unlike traditional autoencoders, which map input data to a single point in the latent vector, variation autoencoders (VAE) map data to a distribution over the latent space. This characteristic of VAEs enhances the generative process. VAEs are trained using a probabilistic method that focuses on maximizing the likelihood of the input data given the latent representation that has been learned. The data generated by a VAE is determined by the distribution of the latent space, allowing for the manipulation and creation of various types of synthetic data. VAEs possess the ability to generate diverse and realistic synthetic data due to their probabilistic nature and the use of variational inference. This allows them to produce novel data that deviates from the training dataset, making them highly valuable. Based on that, Xie et al. utilised a diffusion model in conjunction with VAE to create the crystal diffusion variational autoencoder (CDVAE) for the prediction of 3D periodic materials [109]. CDVAE is able to create new materials by studying existing crystalline structures in the training dataset. Rather than relying on predetermined chemical formulas, the model learns to generate stable crystals by recognizing energy minimums and bonding preferences between different atom types within the dataset. Therefore, the compositions and elemental ratios in the generated data are indirectly determined by the training data. Furthermore, the stability of CDVAE is also ensured by the use of a specific decoder with a noise conditional score network. This decoder employs a harmonic force field to determine the forces acting on atoms when they are out of equilibrium, adding a crucial physical-inductive bias to create stable materials. In order to discover new 2D materials in infinite chemical space and to overcome the challenge of non-periodicity in one direction, as shown in Fig. 5B, Chen et al. implemented artificial periodicity based on existing methods. They adjusted the lattice vector size in the non-periodic direction to be ten times larger than in the periodic direction [110]. Additionally, the cutoff radius for the GNN in the decoder was increased from 7 Å to 10 Å. The modification guarantees that the GNNs connect only atoms within the 2D layer, enabling CDVAE to effectively learn how to generate 2D materials. Similarly, Elrashidy et al. used a CDVAE model to generate 1190 magnetic monolayers with E_{hull} values below 0.3 eV/atom. This was achieved through the training of over 15,000 monolayers with diverse properties sourced from C2DB [27]. They also indicated that the use of CDVAE to decorate lattices with chemically similar elements not only achieves stability and generates new crystal structures, but also shows the potential of employing deep generative models based on autoencoder architecture in the discovery of 2D materials. While the basic VAE is already more powerful than simple autoencoders, there is still potential for improvement by extending the architecture. β-VAE [111],

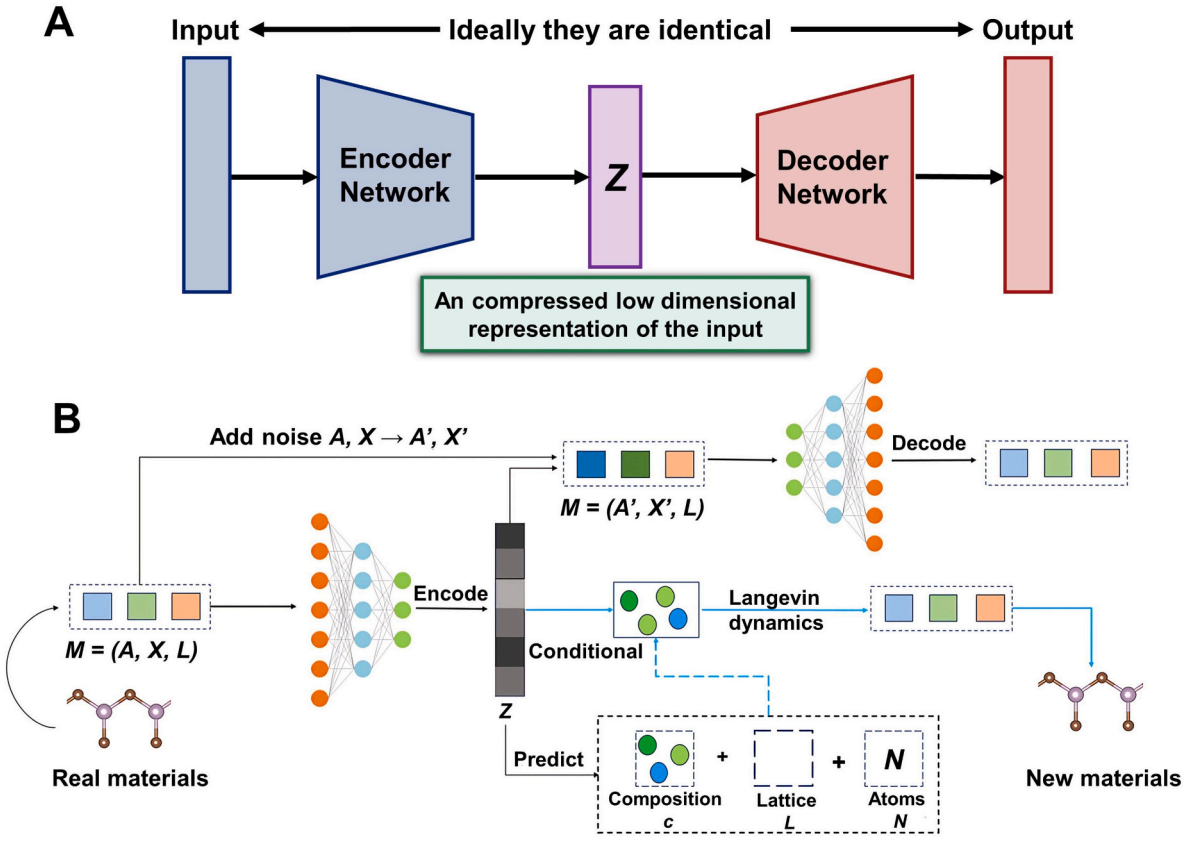

Fig. 5. Schematic architecture of (A) autoencoders and (B) crystal diffusion variational autoencoders (CDVAE) for the inverse design of new 2D BCP monolayer. Reproduced with permission from Ref. [110]. Copyright 2024 Elsevier.

adversarial autoencoder (AAE) [112], and two-stage VAEs [113] are useful in expanding the capabilities of VAEs in 2D and 3D data processing [114].

2.3.3. Convolution neural networks (CNNs)

CNNs stand as the predominant DL algorithm utilised for image analysis. In contrast to MLPs, which process 1D arrays as input, CNNs are specifically designed to handle 2D arrays, such as pixels in images, or even 3D arrays, such as voxels in 3D models. This unique capability allows CNNs to learn and extract hierarchical features directly from the geometric features of 2D materials in either 2D or 3D space. This is achieved by processing the input through a series of convolutional and pooling layers [115]. The typical structure of CNNs, as depicted in Fig. 6A, consists of convolutional layers (Conv), pooling layers (Pool), and fully connected layers (FC). Conv utilize convolutional kernels to extract features from the input 2D or 3D array, while Pool reduce the dimensionality of the output obtained from the convolutional layers. Early layers typically capture low-level features (e.g., edges), while deeper layers capture high-level features (e.g., object parts). Finally, FC achieve the ultimate classification or regression task. At present, CNNs have been extensively employed in image analysis and are known for their strong performance and the need for minimal pre-processing. Therefore, CNNs excel in their capacity to process inputs with a high number of dimensions [15,22]. For instance, as shown in Fig. 6B, the damaged microstructure of h-BN represented by a simple series of matrix (nearly 11,000 atoms) without any further processing was used to train a pre-trained CNNs model as input [14]. Recently, a new method utilizing a fully convolutional network (FCN) encoder-decoder has been developed to detect point defects in 2D materials. This process involved two main steps. To begin, the input image underwent convolution with a set of kernels to identify features like background, host atoms, and defect sites. Subsequently, the image was deconvoluted using these

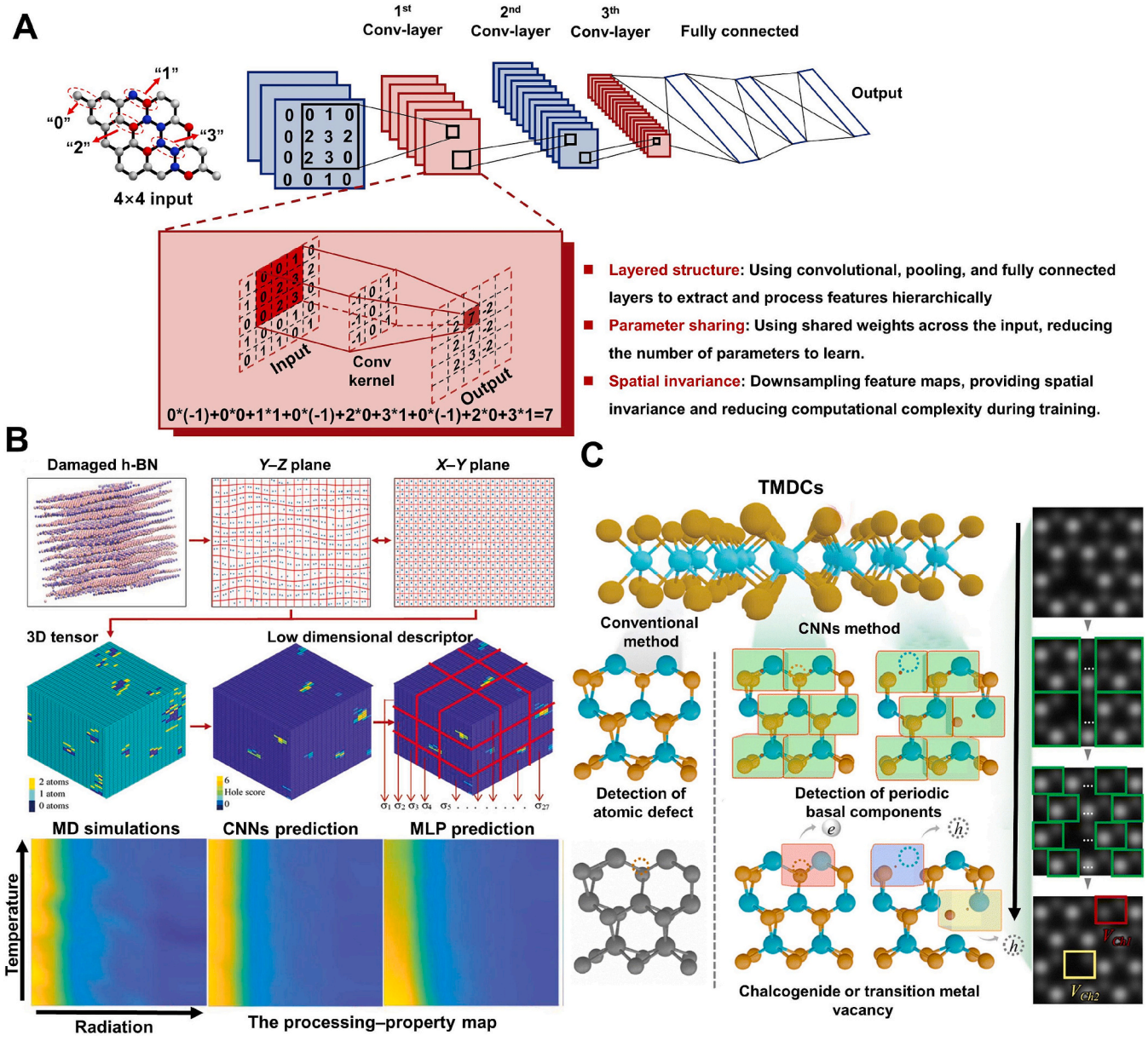

Fig. 6. Convolution neural networks (CNNs) and their application in 2D materials characterizing. (A) The schematic architecture of CNNs. (B) The physical descriptors of damaged h-BN for CNNs input and the processing-property maps as obtained from CNNs model compared with others results. Reproduced with permission from Ref. [14]. Copyright 2019 Wiley-VCH. (C) The workflow of unit cell detection and point defect classification using CNNs models. Reproduced with permission from Ref. [116]. Copyright 2024 Royal Society of Chemistry.

features to restore it to its original dimensions. By utilizing the FCN technique for analysing point defects in 2D materials, scientists can investigate the various atomic contrasts and arrangements of transition metals and chalcogens to detect these defects as depicted in Fig. 6C [116]. During the training process, data augmentation techniques, including flip, rotation, scaling, cropping, translation, and contrast, can be used to enlarge the original data sets to improve the robustness of the CNNs and prevent overfitting problem. Despite their numerous advantages, CNNs also have several disadvantages and limitations that are important to consider. Firstly, CNNs are sensitive to various hyperparameters, such as the learning rate, number of filters, filter size, and network depth. Finding the optimal set of hyperparameters often requires extensive experimentation and fine-tuning. Moreover, without proper regularization techniques (e.g., dropout, weight decay), CNNs can overfit the training data, especially when the model is very deep and the dataset is not sufficiently large or diverse, as is often the case with small dataset of 2D materials [117].

2.3.4. *Sequential neural network architectures*

Sequential neural network architectures encompass models designed to process sequential data, capturing dependencies and patterns within the data over time or across locations. These architectures include recurrent neural networks (RNNs), long short-term memory (LSTM) networks, and Transformers [118,119]. As depicted in Fig. 7A, RNNs process sequential data by iteratively passing information from one step to the next, allowing them to maintain a memory of past inputs. LSTM networks improve upon RNNs by introducing gated mechanisms that better capture long-range dependencies and mitigate the vanishing gradient problem. Transformers, on the other hand, rely on self-attention mechanisms to capture global dependencies in the input sequence, enabling parallel processing of the entire sequence. These sequential neural network architectures have applications in various domains, including natural language processing, time series analysis, speech recognition, and sequential data generation [120]. In comparison to natural language texts, the composition sequences of inorganic materials are subject to strict constraints among elements to form

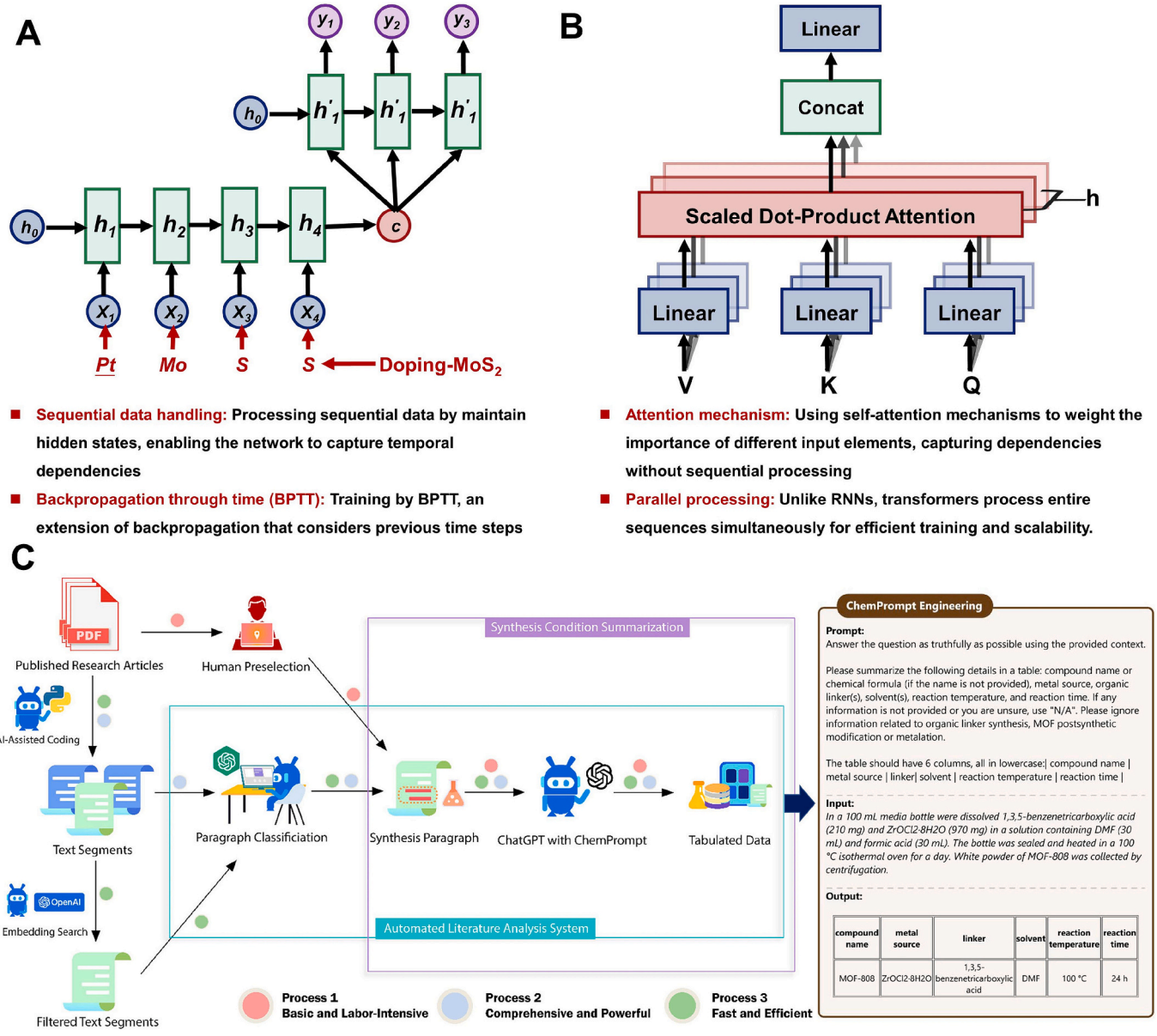

Fig. 7. Schematic architecture of (A) recurrent neural networks (RNNs) and (B) transformer for predictive modelling. (C) Schematics of the ChatGPT chemistry assistant for efficient text mining and summarization of MOF synthesis conditions. Reproduced with permission from Ref. [128]. Copyright 2023 American Chemical Society.

structures with chemical activity and stable periodicity. This necessitates intricate atomic interactions involving ionic or covalent bonds and the oxidation states of the elements involved. Successful models for creating material compositions must be able to understand and consider intricate local and long-distance relationships, as well as the context in which the generation occurs. Moreover, sequential neural networks are particularly adept at recognizing and modelling this complexity.

Many highly successful algorithms, such as Word2Vec in NLP field, utilize unsupervised learning to generate efficient representations of objects within the system. While combining vectors through pooling operations to create representations of systems composed of parts (such as sentences from words) is a common practice in NLP, it seems to be largely unexplored in materials informatics [121]. The analogy they investigated was that atoms are like compounds in the same way that words are like sentences. Their findings showed that compound representations can be effectively created by combining the vector representations of individual atoms. At present, the use of sequential neural network architectures for predicting the properties of 2D materials still faces several challenges. Efficiently representing and encoding the complex and diverse properties of 2D materials for input into neural networks is one of the primary obstacles. For RNNs and LSTMs, these models typically require sequential data. Encoding the structural, electronic, and mechanical properties of 2D materials into an appropriate sequential format can be complex. Transformer, on the other hand, excels in capturing spatial relationships and interactions within 2D materials due to its attention mechanism, which provides flexibility with various input types (Fig. 7B). Fu et al. conducted training on seven well-known transformer models (GPT-2, GPT-Neo, GPT-J, BLMM, BART, and RoBERTa) to generate designs for inorganic material compositions [122]. They utilised the extended formulas from ICSD, OQMD, and Materials Projects databases for this purpose. Their motivation stemmed from the belief that by adhering to a specific sequence, the composition or formula of materials could be converted into a unique arrangement of elements. For example, $SrTiO_3$ can be decomposed into Sr, Ti, O, O, and O based on increasing electronegativity. While their results showed that the transformer-based model could produce chemically valid hypothetical material compositions, it remained uncertain whether these generated compositions could be realised with thermally stable structures. This uncertainty is particularly pertinent when the generative model generates compositions with a high number of elements. In addition, the availability of crystal structures for composition generators using sequential neural network architectures must also be taken into account. In the future, crystal structure prediction can be accomplished using various methods, such as template-based crystal structure prediction (TCSP) algorithms [123,124], DL-based models [125], and global optimization-based tools [126,127]. These techniques can be employed to forecast the crystal structures of hypothetical compositions generated by material transformer models.

Recently, the emergence of generative AI, particularly advancements in large language models (LLMs), has marked a significant shift in the capacity to leverage unstructured data across various fields, including material science. Considering that ChatGPT can effectively extract and organize a wide range of materials information cohesively using natural language without the need for coding skills. As shown in Fig. 7C, Zheng et al. recently introduced a new method called ChemPrompt, using ChatGPT to streamline the synthesis conditions of MOFs by converting text into tables and providing summaries of scientific literature [128]. They also developed a reliable MOF chatbot that utilizes data to respond to inquiries about chemical reactions and synthesis methods. However, this method was ineffective in extracting organised representations of intricate hierarchical entity connections and did not extend beyond the constraints of the pretraining dataset. To tackle this problem, Ekuma proposed a LLMS-based system named Property Extractor, which employed a blended dynamic zero-shot-few-shot in-context learning approach to autogenerate a thickness database for 2D materials [33]. Leveraging this model and a trained ML algorithm, they successfully analysed over 8000 materials and predicted their thickness. This method accelerated the characterization process, simplifying the screening and selection of materials for specific applications, thereby expediting the development of next-generation technologies. Additionally, optimization of this method could be enhanced through the utilization of tools such as Google Gemini Pro and OpenAI GPT-4.

2.3.5. Graph neural networks (GNNs)

Graph neural networks (GNNs) have emerged as a promising solution to address the challenge of representing materials, which often encounter limitations in traditional DL structures such as vectors and matrices. Despite the success of many high-throughput computing (HTC) and DL techniques in forecasting material properties and discovering new ones, accurately representing crystalline materials remains a challenge. GNNs offer a new approach by leveraging the inherent graph structure of materials, allowing for more effective representation and analysis [27]. Unlike traditional neural networks designed for grid-like data, GNNs can operate seamlessly on irregular, non-Euclidean structures such as social networks, molecular structures, or citation networks [129]. In DL methods, an atomic structure is represented as a point cloud, which is a collection of multiple points in 3D or higher dimensional space. Each point is linked to a vector or property, including the atomic number and potentially additional physics-based characteristics like radius and valence electrons. Similar concepts can be applied to detect defects in 2D materials. As shown in Fig. 8A, Kazeev et al. treated MoS_2 as a point cloud of defects rather than atoms. Through sparse representation, each point in the structure is characterised by two parameters: the atomic number of the atom in the original structure and in the structure with the defects or vacancies assigned an atomic number of 0. Based on this representation, GNNs are a suitable choice for generalizing materials [130]. Specifically, as illustrated in Fig. 8B, GNNs leverage graph convolution layers to capture features from the neighborhood of each node, enabling them to capture the intricate relationships and dependencies inherent in graph data [131]. These networks have garnered significant attention across various fields for their versatility and effectiveness in tasks such as node classification, link prediction, and graph generation. They exhibit exceptional properties with applications in exploring 2D material properties and inverse design.

Recent advancements have shown remarkable progress in utilizing GNN models for predicting and characterizing properties of 2D materials. Notable examples include Crystal Graph Convolutional Neural Networks (CGCNN) [105], MatErials Graph Network (MEGNet) [132], OrbNet [133], atomistic line graph neural network (ALIGNN) [134], Open Catalyst benchmark [130], and similar derivative structures [135]. In these models, a molecule or crystal material is represented as a graph, where each atom is depicted as a node and the interatomic bonds are represented as edges. These models typically utilize fundamental characteristics as node attributes and may incorporate interatomic distances or bond valences as edge attributes. By continuously adjusting node characteristics based on their adjacent chemical environment using graph convolution layers, these models can accurately capture intricate many-body interactions [105]. ALIGNN, a line graph neural network developed by Choudhary and DeCost introduces an innovative approach to incorporate angular information, creating highly accurate models for 2D materials [134]. This is crucial as the properties of these materials, particularly electronic properties like band gaps, are greatly influenced by structural characteristics such as band angles and local geometric distortions. The ALIGNN model conducts edge-gated graph convolution message updates on two types of graphs, namely the atomistic bond graph (where atoms represent nodes and bonds represent edges) and its line graph (where bonds represent nodes and edges connect bond pairs with a common atom). This variant of edge-gated graph offers a unique advantage in adjusting both node and edge characteristics (Fig. 8C). In practical applications aiming to reduce the computational cost of DFT simulation for determining thermal stability and

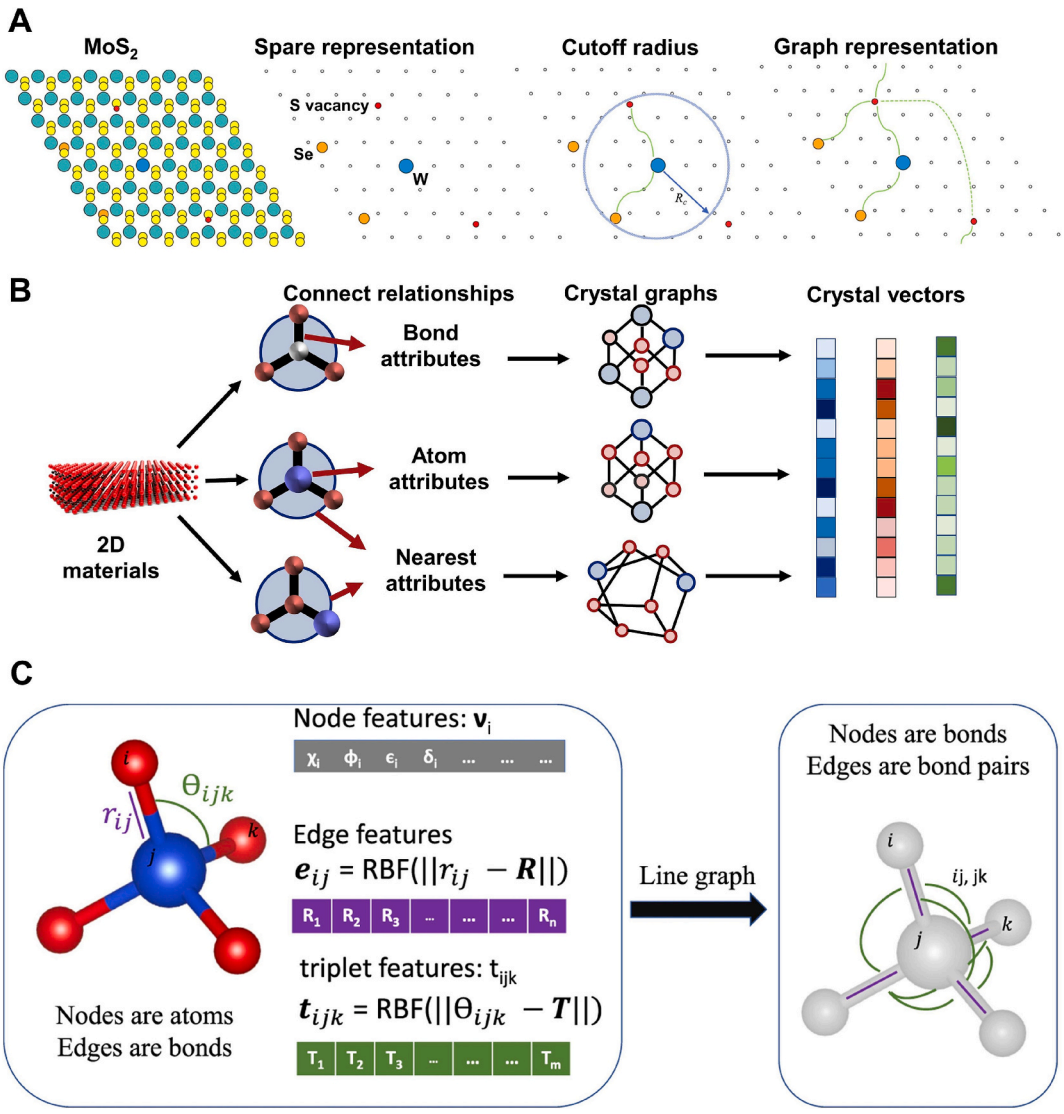

Fig. 8. A graph-based method for representing 2D materials. (A) The sparse and graph representation of MoS_2 with point defects. Reproduced under the terms of Creative Commons Attribution 4.0 International (CC BY 4.0) license from Ref. [130] published by Springer Nature (B) 2D material can be represented as a graph, where atoms serve as nodes and interatomic details such as distance, bond type, and solid angle act as edges. Subsequently, these attributes of the graph are converted into vectors, incorporating diverse levels of information extracted from the 2D materials. (C) The ALIGNN convolution layer alternates between message passing on the bond graph (Top) and its line graph (or bond adjacency, bottom). Reproduced under the terms of Creative Commons Attribution 4.0 International (CC BY 4.0) license from Ref. [134] published by Springer Nature.

magnetic properties, Elrashidy et al. developed and trained two ALIGNN architectures as filters to predict the E_{hull} of magnetic monolayers. They achieved a mean absolute error (MAE) of 0.039 eV/atom and a classification accuracy of 93 % by training the models on 154,718 materials from the MP database [27]. These models approached the performance of DFT calculations but at significantly reduced computational expense.

Additionally, M3GNet is a versatile GNN architecture designed as an all-purpose inverse atomic potential (IAP) tool for predicting material properties. Constructed on a comprehensive dataset that includes energies, forces, and stress parameters from the Materials Project database, the network accurately predicts both individual atom characteristics and overall crystal structures. Significantly, M3GNet is notable for its implementation of three interacting modules, allowing the network to effectively capture long-range interactions without the need to increase the cutoff radius during bond construction. Elemental features of each atom are represented by embedding vectors within the architecture. This design allows for straightforward extension to handle materials with multiple components and chemistries [136]. Currently, M3GNet functions as a pre-structure optimizer capable of efficiently approaching the vicinity of potential energy surface minima, significantly reducing the computational cost compared to DFT, before conducting full relaxations. This is achieved through the application of both tensile and compressive stress parallel to the plane of the monolayers [27]. In summary, using message-passing mechanisms, GNNs have the ability to continuously enhance node representations by integrating information from neighboring nodes and edges. These capabilities make GNNs well-suited for 2D material characterization and inverse engineering. However, it is important to acknowledge that GNNs can be computationally intensive, particularly when handling large graphs with numerous nodes and edges. Efficient training and inference processes require significant computational resources [131].

2.3.6. *Generative adversarial networks (GANs)*

The potential of generative DL algorithms extends beyond the traditional lattice decoration paradigm, enabling exploration of a broader range of candidate 2D materials [17,137]. Generally, GANs,

consisting of a generator and a discriminator, train concurrently through an adversarial process. The generator's goal is to produce data that resembles real samples, while the discriminator aims to distinguish between generated (fake) and real data. Training continues until equilibrium is reached, at which point the generator successfully deceives the discriminator. The generator, typically an artificial neural network, accepts randomly selected low-dimensional noise variables, usually from a Gaussian distribution, and produces high-dimensional output (such as images of 2D materials). On the other hand, the discriminator, another neural networks, is provided with input in a high-dimensional space from training set and generator, following by classifying them as either real or fake [138]. Developing a generative model with robust performance for periodic 2D materials has proven to be a challenging task. One of the primary obstacles is designing representations of the lattice, atomic coordinates, and chemical composition that are invariant under translations and rotations, yet also reversible. Additionally, the vast chemical space of elements represent in inorganic crystals adds another layer of complexity to the design of these representations. As a result, previous efforts to develop generative models for periodic 2D materials have often been constrained by either focusing on a specific group of chemical elements or a limited range of crystal structures. However, a recent breakthrough has introduced a general invertible representation that encodes materials using a matrix of true and interacting spatial characteristics. It should be noted that unlike the CDVAE model mentioned in Section 2.3.2, this representation is not invariant under translations and rotations. They employed an equivariant GNN to ensure invariance, particularly equivariance, and assessed the quality and variety of the resulting crystals through qualitative assessments including charge neutrality and minimum bond distance [27]. In contrast, Lyngby and Thygesen performed a thorough quantitative analysis using full DFT-based relaxation and stability analysis on the new 2D structures [17]. The structures generated by the CDVAE exhibited comparable formation energies to those in the training set but displayed great differences in both chemical compositions and crystal structures. Additionally, MatGAN, a typical GNN model, is employed to acquire implicit patterns in chemical compositions such as electronegativity balance and charge neutrality, from three datasets (OQMD, ICSD, and Materials Project) [139]. The MatGAN generator has the ability to generate virtual formulas that satisfy both electronegativity and charge neutrality. At the same time, the discriminator can distinguish between generated formulas that closely resemble real ones. Importantly, MatGAN has shown a remarkable ability to create new materials not present in the original training data, with 84.5 % of the materials generated meeting the criteria for electronegativity balance and charge neutrality. Some studies have identified new materials with wide bandgaps or thermodynamically stable 2D structures by analysing the hypothetical inorganic formulas produced by MatGAN [137,140]. For more detailed information on MatGAN, refer to the study by Dan et al. [139], Furthermore, in order to address the issue of limited data availability and achieve precise defect segmentation in classical 2D material systems and MoS_2/WS_2 heterostructures at the atomic level, AtomGAN was developed. This model includes a class encoder, generator, and discriminator to learn features in both domains, enabling the detection of crystallographic defects and the rapid generation of TEM images that closely resemble real electron microscope images. AtomGAN's innovative class encoder branch prevents non-convergence loss during GAN training and facilitates quick conversion of output to the target domain.

3. Characterization of 2D materials

Understanding the physical and chemical properties of 2D materials, such as defect, thickness, and morphology, is crucial for unlocking their potential in various applications. In order to represent the structure of 2D materials, researchers have utilised both experimental and theoretical methods [141]. Experimental techniques include OM, TEM, scanning transmission electron microscopy (STEM), AFM, Raman spectroscopy, and reflection contrast spectroscopy [15]. On the other hand, theoretical methods involve MD simulations and DFT calculations. Nevertheless, conventional characterization approaches mentioned above possess certain drawbacks, including the need for expensive computational resources and the potential for biases introduced through manual analysis [142]. By integrating DL with these traditional methods, it is possible to overcome their inherent limitations and tackle certain obstacles (Table 3). This section will introduce the characterization of defects and thickness in 2D materials [22,85].

3.1. Defects identification for 2D materials

The properties of 2D materials are greatly affected by defects such as vacancies, doping, and edge defects. Controlling these different defect types in lattice structures can alter the electronic, optical, mechanical, and magnetic properties of the materials [84]. Accurately detecting defects at the atomic level is crucial for understanding and manipulating these properties. High-resolution imaging and characterization techniques, such as AFM and TEM, have significantly advanced, allowing for detailed microscopic analysis of materials and their defects [143]. Advanced characterizations can uncover the dynamics of material characteristics at the atomic level with sub-second temporal resolution. However, despite the improved ability to collect high-resolution material data, the analysis of 2D material properties from these images is still constrained by manual methods [83,116]. As a result, qualitative research primarily relies on these high-quality data. However, depending solely on manual analysis poses challenges in efficiently and accurately extracting all relevant features from the images. Therefore, a significant amount of data is overlooked and filtered. It is clear that the widespread adoption of advanced characterization techniques is hindered by the limitations of manual analysis. Hence, there is an urgent requirement for automatic and intelligent methods to extract additional information regarding the dynamics and interactions of single defects from high-resolution micro images.

DL, especially for CNNs, would be a suitable option for analysing large-scale 2D defective systems with a high number of atoms, various defect types, and intricate defect distribution, utilizing image pixel and one-hot encoding to create 2D tensor descriptors. Point defects are common in 2D materials and are often associated with various physical phenomena. One potential method to utilize structure-modulated 2D materials is by directly observing these point defects and conducting statistical analysis [84]. As shown in Fig. 9A, Yang et al. developed a 2DIP-Net based on CNNs to detect point defects in monolayer 2H-$MoTe_2$ at pixel-by-pixel level [116]. They utilised experimental results from TEM datasets and theoretical calculations from STEM datasets, focusing on unit cell unit detection and classification. To relieve sample linear drift-related image distortion, they scanned a single STEM image over a short dwell time of 5–6 μs per pixel. It was noteworthy that 2DIP-Net, integrating faster-region-based (*R*)-CNN and ResNet-18, successfully automated the analysis of point defects in 2H-$MoTe_2$ using STEM data. Faster R-CNN was utilised to detect the hexagonal cell, as shown in Fig. 9A(a), followed by cropping of the unit cell. Subsequently, ResNet-18 was employed to classify the different types of defects within the unit cell, as depicted in Fig. 9A(b). This approach effectively addressed the challenge of extracting detailed image features from HADDF-STEM images. Moreover, DL has shown remarkable effectiveness in restoring images from raw data with high levels of statistical noise or blurring, particularly for STEM data of radiation-sensitive 2D materials that necessitate observation under low-dose conditions. In addition to image classification, CNNs can function as restoration algorithm to reduce statistical noise and enhance contrast in STEM images during the preprocessing stage. Utilizing a dilated convolutional kernel, as depicted in Fig. 9B, proved effective in enhancing the model's capacity without the need to increase the network depth. This approach enabled the extraction of contextual information for image restoration, significantly reducing computational expenses [143]. Although STEM can image

Table 3
Summary of utilizing deep learning in the diverse fields of 2D materials.

2D materials	Research area	Data sources	Descriptors	Hyperparameter settings	DL algorithms	Results	References
Identification and characterization							
MoS_2/WS_2	Point and line defects identification	TEM and simulation images	256 × 256 × 1 pixels with different brightness values	Batch size = 4; Optimizer: Adam; Epoch = 30,000; Learning rate = 10^{-4}	AtomGAN	Precision = 96.9 %; Recall = 92.0 %; F1 = 94.4 %	[31]
Graphene, h-BN, TMDs	Material and thickness identification	OM images	224 × 224 × 3 pixels with different RGB values	Optimizer: SGD; Epoch = 20–250; Learning rate = 10^{-2}	2DMOINet	Acc = 96.89 %; IoU = 58.78 %	[15]
Graphene	Thickness identification	OM images	256 × 256 × 3 pixels with different RGB values	Batch size = 8; Optimizer: SGD; Learning rate = 0.1; Loss function: Cross-entropy	Hierarchical CNNs UNet++ DeepLabv3+	Acc = 81.6 %; F1 = 91.8 %; IoU = 71.7 %	[73]
BP, graphene, MoS_2, WSe_2	Recognition and classification of 2D materials	Raman spectra	1 × 57 vectors with different intensity values	Batch size = 32; Optimizer: Adam; Learning rate = 2 × 10^{-4}; Epoch = 100	DDPM-CNNs	Acc = 98.8 % Precision = 94.5 %; Recall = 93.7 %	[32]
h-BN	Thickness identification	OM images	256 × 256 × 1 pixels with different brightness values	Learning rate = 0.0025; Epoch =20;	DetectoRS	Precision = 90.0 %.	[34]
TMDs ($MoTe_2$)	Point defect detection	STEM images and simulations	256 × 256 × 1 pixels with different point defect concentrations	Optimizer: Adam; Learning rate = 10^{-4} and 2 × 10^{-6}; Epoch =40; Loss function: Cross-entropy	2DIP-Net	Acc = 97.9 % IoU = 0.691	[116]
h-BN, WS_2, MoS_2 and their heterostructures	Identification and characterizations	OM images	1D vectors of six color channels	Optimizer: SGD Learning rate = 10^{-3}; Epoch = 500 Loss function: Mean squared error	MLPs	Acc > 90 %	[141]
Properties prediction							
Graphene	Atomistic physical fields prediction based on strain and defect engineering	MD simulations	256 × 256 pixels with different magnitudes of field values	Batch size = 64; Epoch = 300; Loss function: Cross-entropy	GAN	$R^2 > 0.97$	[53]
WSe_2, h-BN, MoS_2	Total energy, band gap, and Fermi energy prediction	C2DB database; DFT calculations	Graph representation	Optimizer: Adam; Learning rate = 0.001; Epoch = 4000	MEGNet	$R^2 > 0.73$ MAE < 0.36 eV	[84]
Defective graphene	Fracture stress	Scientific literature; MD simulations	256 × 256 pixels 5 × 5 feature matrix of chemical descriptors	Batch size = 40; Optimizer: SGD; Learning rate = 10^{-4}	MLPs, CNNs	RMSE = 1.98 MAE = 1.55	[52]
Graphene and boron nitride	Bandgap prediction	DFT calculations	4 × 4, 5 × 5, and 6 × 6 supercells	Activation function: Exponential linear units	CNNs, RCN, and VCN	Acc > 0.90 RMSE ≈ 0.1 eV; MAE < 0.15 eV	[94]
2D magnets	Exploration and prediction of magnetic property	Materials Project and C2DB database	Chemical and graph descriptors	Optimizer: SGD Batch size = 100;	CDVAE, ALIGNN, and M3GNet	Accuracy = 93 % MAE = 0.039 eV/atom	[27]
2D materials	Property prediction	Materials Projects, JARVIS-2D, and HOPV databases	Graph descriptors	Batch size = 64; Optimizer: Adam Learning rate = 10^{-4}; Epoch = 200	ALIGNN	MAE < 0.0235 eV/atom	[29]
Trilayer graphene	Band structure parameter prediction	MD calculations	1 × 4 feature vectors of chemical descriptors	Batch size = 512; Optimizer: Adam; Learning rate = 10^{-3};	DNNs	MAE < 1.54 RMSE <1.82 R2 > 0.999	[54]
Graphene oxide	Mechanical property prediction	Reactive MD simulation	One-hot encoded vectors	Activation function: ReLu; Learning rate = 10^{-3}; Optimizer: Adam;	Deep RL-MLPs	Over two standard deviations	[82]
2D materials	Flat electronic bands prediction	Materials Project and 2Dmatpedia database	96 × 96 pixels	Optimizer: Adam; Learning rate = 5 × 10^{-5}; L2-regularization = 0.003	CNNs, t-SNE	—	[28]
Discovery and design							
Graphene, h-BN	Microbial induced corrosion data generation	Electrochemical impedance spectroscopy	1 × 7 feature vectors of chemical descriptors	Activation function: ReLU; Epoch = 50;	GAN, VAE, MLPs	Accuracy >83 %; ROC >0.825	[117]
2D BCP monolayer	Inverse design of BCP monolayer materials	Scientific literature	Graph representations	Optimizer: Adam; Learning rate = 0.001; Loss function: Cross-entropy and mean square error	CDVAE, M3GNet	Finding five novel BCP monolayer materials	[110]

(*continued on next page*)

Table 3 (*continued*)

2D materials	Research area	Data sources	Descriptors	Hyperparameter settings	DL algorithms	Results	References
2D materials	Inverse design	C2DB, MC2D, 2DMatPedia, and V2DB databases	Sequence vectors and graph descriptors	Optimizer: Bayesian optimization; Vocabulary size = 130; Sequence length < 205; Training steps = 200,000;	BLMM, TCSP, CSPML, BOWSR, and M3GNET	Finding four stable 2D materials	[18]
h-BN and graphene	Structure–Property Relations	MD simulations	3D tensors based on voxelization	Optimizer: Adam; Learning rate = 10^{-3}; Activation function: ReLU;	CNNs, MLPs	$R^2 \approx 95$ % $E_{max} = 3.2$ % ~ 7.1 %	[14]
2D materials	topology optimization	DFT calculations	128 × 128 pixels assigned numeric identifier	Batch size = 128; Learning rate = 0.001; Optimizer: Adam;	CNNs	MSE = 0.00794 DSC = 0.970,	[164]
2D materials	Inverse design	C2DB database	Chemical and geometric descriptors	Learning rate = 0.001; Epoch = 30;	CDVAE	Finding 2004 ideal 2D materials	[17]
2D materials	Discovery of new hypothetical 2D materials.	2Dmaterials and Material Project databases	8 × 85 sparse matrix with 0/1 cell values	Learning rate = 10^{-3}; Batch size = 1024; Optimizer: Adam;	MatGAN	Probability scores >0.95	[137]
MXenes	Screening of MXenes for hydrogen storage	aNANt database DFT calculations	Graph descriptors	Optimizer: Adam; Batch size = 64; Learning rate = 0.001;	ALIGNN	Hydrogen storage capacity of 5.7 wt% at 230 K and 100 bar	[165]

Note: Acc = Accuracy; ALIGNN = Atomistic line graph neural networks; AtomGAN = Atom generative adversarial networks; BP = Black phosphorus; BLMM = Blank language models for materials; BOWSR = Bayesian optimization with symmetry relaxation algorithm; C2DB = The computational 2D materials database; CDVAE = Crystal diffusion variational autoencoder; CNNs = Convolution neural networks; CSPML = A machine learning-based crystal structural prediction method using a machine learning model to select template; DALM = Deep-learning-enabled atomic layer mapping; DDPM = Denoising diffusion probabilistic models; DFT = Density functional theory; DetectoRS = A multi-stage supervised object detection algorithm; DNN = Deep neural network; Deep RL = Deep reinforcement learning; DSC = Dice similarity coefficient; E_{max} = The maximum percentage error among the predictions; F1 = F1 score; GAN = Generative adversarial networks; h-BN = Hexagonal boron nitride; IoU = Intersection over union; MAE = Mean absolute error; MD = Molecular dynamic; MEGNet = MatErials graph network; M3GNet = Materials graph with three-body interaction neural networks; MLPs = Multilayer perceptron; OM = Optical microscopy; PSO = Particle swarm optimization; R^2 = Coefficient of determination; RCN = Residual convolutional network; RMSE = Root mean square error; SGDM = Stochastic gradient descent; TCSP = A template-based crystal structure prediction algorithm based on oxidation state patterns; TEM = Transmission electron microscopy; TMDs = 2D transition metal dichalcogenides; t-SNE = t-Distributed stochastic neighbor embedding; STEM = Screening transmission electron microscopy; VAE = Variation autoencoder; VCN = VGG16 convolutional network; 2DIP-Net = A CNN-based analytic platform; 2DMOINet = 2D material optical identification neural network.

individual atoms, its accuracy is compromised by the signal to noise ratio (SNR) ranging from 1.1 to 2.2. High-does radiation allows accurate detection of atomic positions, but ionization effects can alter defective 2D structures. On the other hand, low-dose irradiation results in images with high noise levels, making it difficult to quantitatively evaluate atomic defects. To address this issue, a DL approach involved taking a range of STEM images taken with low-dose irradiation as input and employing a CNN architecture to enhance and de-noise the STEM images, ultimately improving the signal-to-noise ratio [143].

Generally, supervised DL approaches that treat defect detection as a pixel-by-pixel object detection task face limitations due to the lack of sufficient labelled and balanced training datasets, which requires significant processing time. As a result, unsupervised learning methods, particularly GANs, have gained popularity for identifying defects. Recently, Cheng et al. developed an unsupervised AtomGAN model related to atom-scale defect detection, capable of segmenting different defect types in MoS_2/WS_2 simultaneously [31]. The AtomGAN, utilizing unique class encoding and multichannel output, can accurately segment point defects and line defects with 96.9 % precision, even on very blurry images. As shown in Fig. 9C, the output image from AtomGAN increasingly resembled the target segmentation results after training 10,000 epochs. Additionally, the cyclical structure of AtomGAN enables the rapid generation of a large number of simulated electron microscope images, effectively addressing the problems of small and unbalanced dataset. In addition to generative models, transfer learning can also be applied to small datasets. By utilizing a pre-trained network, it is possible to identify the layer numbers of a new 2D material produced through different synthesis method. In summary, a DL-based approach for material characterization is highly effective, speeding up the synthesis and initial determination of 2D materials and other nanomaterials. This has the potential to accelerate the discovery of novel materials [15,29].

3.2. Thickness characterization for 2D materials

Chemical vapor deposition (CVD) and mechanical exfoliation are common methods for preparing 2D materials. During these processes, layers of 2D sheets of varying thickness are haphazardly placed onto a substrate [15]. However, different atomic layer thicknesses in 2D materials result in significant variations in their properties [15,144]. The combination of DL and OM enables the automatic acquisition of elaborate features from images, facilitating the efficient and cost-effective determination of 2D materials on a large scale. This approach does not rely on expensive equipment and offers exceptional scalability, rendering it a valuable tool for researchers and industry professionals [34].

DL techniques applied to the analysis of 2D material thickness can streamline the process of detecting flakes with minimal human involvement, while also mitigating the risk of losing valuable insights from outdated data. To accommodate images captured with various microscope setting, as shown in Fig. 10A, Ramezani et al. implemented a new DL pipeline based on DetectoRs as an object detector to classify crystallites of h-BN [34]. This model integrated additional feedback connections from feature pyramid networks into the bottom-up backbone layers on a larger scale, as well as convolved information with varying atrous rates and aggregated the outcomes using switch functions at a smaller scale. This approach can be applied to different microscope configurations and is resilient to variations in color or substrate background. Furthermore, to tackle the challenge of images with diverse backgrounds, a hierarchical CNNs was developed to accurately detect and categorize the thickness of exfoliated graphene flakes from OM images. This advanced model not only has the capability to learn from

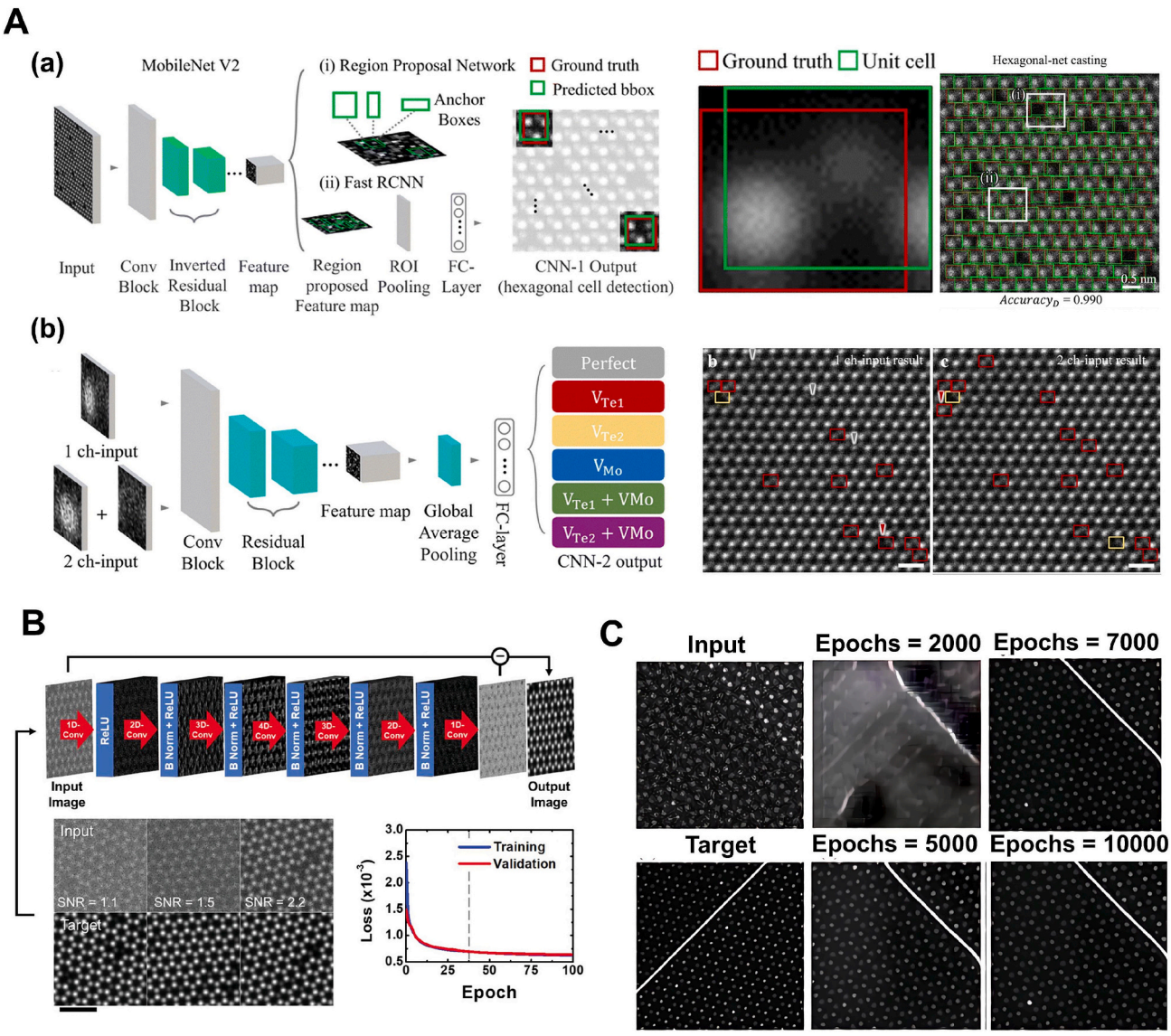

Fig. 9. Applications in 2D materials defect detection using DL methods. (A) Full automation of point defect detection in TMD based on two CNNs with different functions, mainly focusing on (a) unit cell detection and (b) point defect classification. Reproduced with permission from Ref. [116]. Copyright 2024 Royal Society of Chemistry. (B) Quantification of atomic dopants and defects in 2D materials assisted by CNNs. Reproduced with permission from Ref. [143]. Copyright 2021 Wiley-VCH. (C) The training results of AtomGAN obtained from 2000, 5000, 7000, and 10,000 epochs compared with the target image. Reproduced with permission from Ref. [31]. Copyright 2023 Springer Nature.

new OM images, but also retains knowledge gained from previous images, unlike conventional CNNs. Additionally, techniques such as weak learning, data augmentation, iterative stratification, and weighted cross-entropy loss were employed to enhance the accuracy and generalization ability of the model [73]. A 2DMOINet employing CNNs, as shown in Fig. 10B(a), accurately determines the material types and thicknesses of individual 2D material flakes [15]. This study also revealed that the graphical features extracted by CNNs correlate with the optical properties (Fig. 10B(b)) and mechanical properties (Fig. 10B(c)) of the 2D materials. The optical response of the 2D material, depicted in Fig. 10C (a), has a significant impact on the contrast/color and edge features during the process of mechanical exfoliation. There responses are the reflections of electronic band structure and flake thickness of the material. Additionally, the mechanical properties of the material, including crystal symmetry, mechanical anisotropy, and exfoliation energy, play a significant role in determining the usual variations in flake shapes and sizes. As shown in Fig. 10C(b), U-Net has been utilised in previous studies to distinguish between monolayer and bilayer MoS_2 and graphene flakes by analysing optical microscopy images [75]. Compared to full convolutional methods, region-based CNNs used for object detection and instance segmentation tasks were selected to further improve performance for 2D materials automatically searching, including graphene, h-BN, MoS_2, and WTe_2. By incorporating a mask prediction branch, region-based CNNs offers more accurate object localization and instance segmentation. Furthermore, to address the accurate identification of similar few-layer flaks using only RGB images, Dong et al. introduced a new method called 3D deep-learning-enabled atomic layer mapping (DALM). This method combines 3D hyperspectral reflection images with high spectral resolution and 2D RGB images with high spatial resolution to accurately identify and segment MoS_2 flakes of different thickness (mono-, bi-, tri-, and multilayer) [76]. To address the dimensional disparities, the researchers employed a U-shape architecture and applied squeezing the 2D convolution (3×3) to compress the dimensions of the 3D features. This was done to maintain uniformity in the size of the 2D features, which was accomplished by using a feature fusion block. Other related applications are shown in

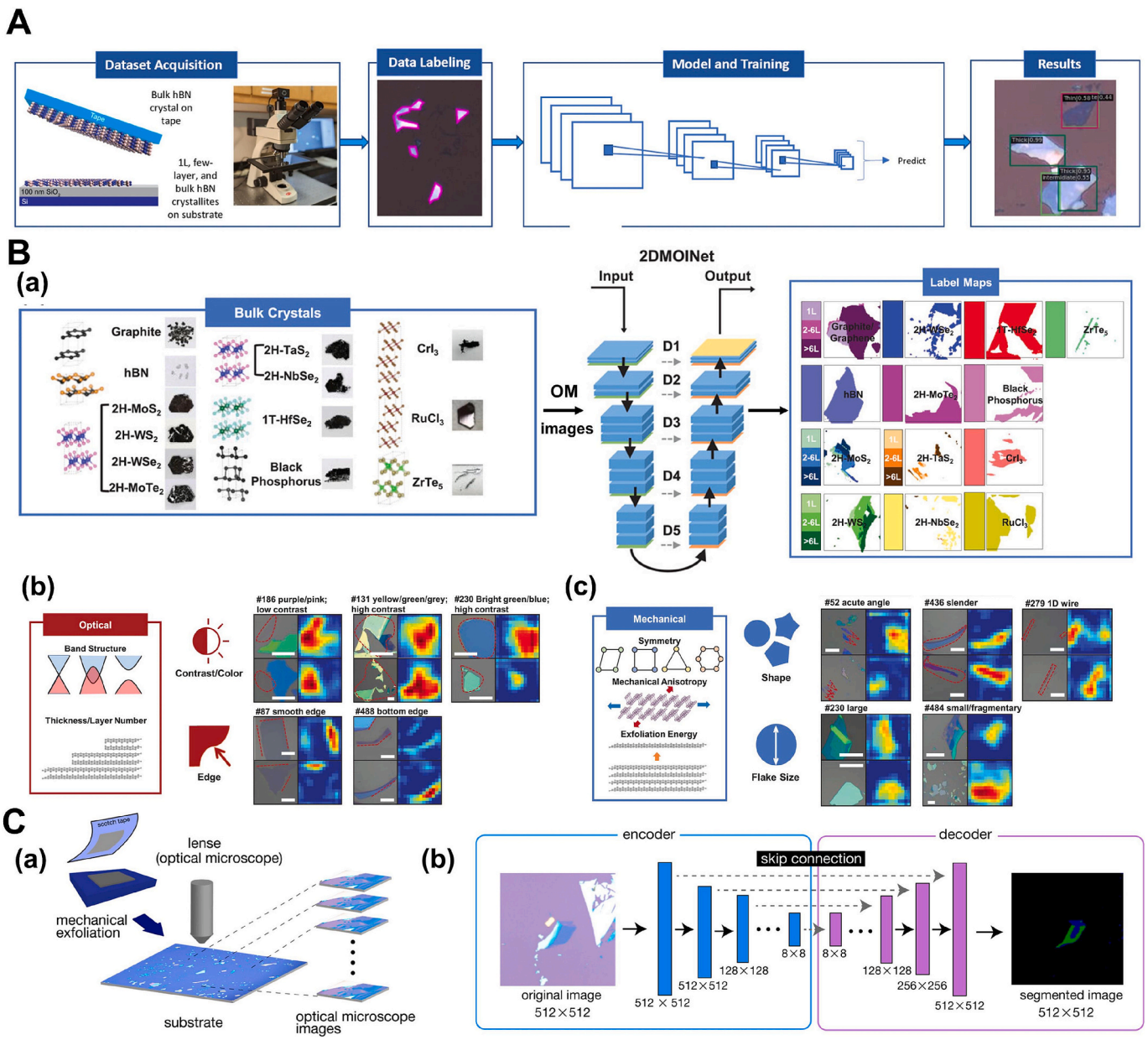

Fig. 10. Applications in thickness identification of 2D materials based on DL models. (A) The common workflow of DL models applied in 2D materials identification based on optical microscopy images. Reproduced under the terms of Creative Commons Attribution 4.0 International (CC BY 4.0) license from Ref. [34] published by Springer Nature. (B) The workflow of 2DMOInNet in capturing the deep graphical features (a), and schematics of the physical properties that determine (b) the optical response and (c) mechanical exfoliation of the 2D flakes. Reproduced with permission from Ref. [15]. Copyright 2020 Wiley-VCH. (C) (a) Mechanically exfoliating MoS_2 crystals on SiO_2/Si substrate, and (b) the U-Net encoder utilizes convolution and pooling layers to extract a small feature map from the input image, while the decoder expands it back to the original image size using convolution and upsampling layers. Reproduced under the terms of Creative Commons Attribution 4.0 International (CC BY 4.0) license from Ref. [75] published by Springer Nature.

Table 3.

4. Prediction of 2D materials properties

Compared to 3D bulk materials, 2D materials consist of only one or a few layers of atoms. Their physicochemical characteristics can be finely adjusted through changes in composition, introduction of defects, surface doping, phase transitions, thickness variations, and chemical modifications. As a result, 2D materials offer a wide range of possibilities for exploration [9,21]. The utilization of DL techniques in forecasting the characteristics of 2D materials has the potential to expedite research advancements and minimize associated expenses. Here, we present an overview of the advancements made in combining DL with theoretical calculations to forecast the electronic and mechanical properties of 2D materials (Table 3).

4.1. Electrical properties

The electronic properties of 2D materials are highly significant in determining their applications, particularly concerning their bandgap. Graphene, for example, exhibits exceptional electrical conductivity but lacks a bandgap, thereby restricting its utility in tasks like switching within digital circuits, semiconductors, and optoelectronic devices [145]. On the other hand, MoS_2 is considered an excellent 2D semiconductor due to its direct bandgap of approximately 1.89 eV. In contrast, h-BN exhibits characteristics of a wide bandgap

semiconductor, with a bandgap measuring around 5.9 eV [146]. Consequently, h-BN finds extensive application as an insulating layer in semiconductor equipment [147]. Hence, it is crucial to accurately and swiftly predict the bandgap of 2D materials in order to effectively utilize them in different industries.

The combination of DL and theoretical calculations has been shown to enable accurate bandgap prediction at low cost [94]. As our knowledge, the electronic properties of 2D materials can be greatly influenced by the atomic and nanoscale arrangement of dopants [19]. The interactions between atoms in a structure collectively determine the bandgaps of the material, as each atom influences its neighboring atoms. The convolutional kernels of CNNs can extract key features not just from input data elements, but also from their neighbors. This capability allows the model to effectively capture the features of configurationally hybridised graphene using various supercell systems, as shown in Fig. 11A, both qualitatively and quantitatively [94]. Importantly, DFT calculations were employed to investigate the nitrogen doping in graphene and to predict a class of novel 2D carbon nitrides. It is widely accepted that CNN methods outperform non-CNN machine learning methods in extracting features for problems involving spatial structures, as the flattening process of non-CNN methods may result in the loss of crucial spatial features that are essential for determining bandgaps. This is mainly attributed to the convolution processing utilised in CNN methods. Consequently, many studies employing CNN model for predicting 2D material properties have focused on single-layer datasets rather than considering multiple layers [52,91]. In the future, analysing

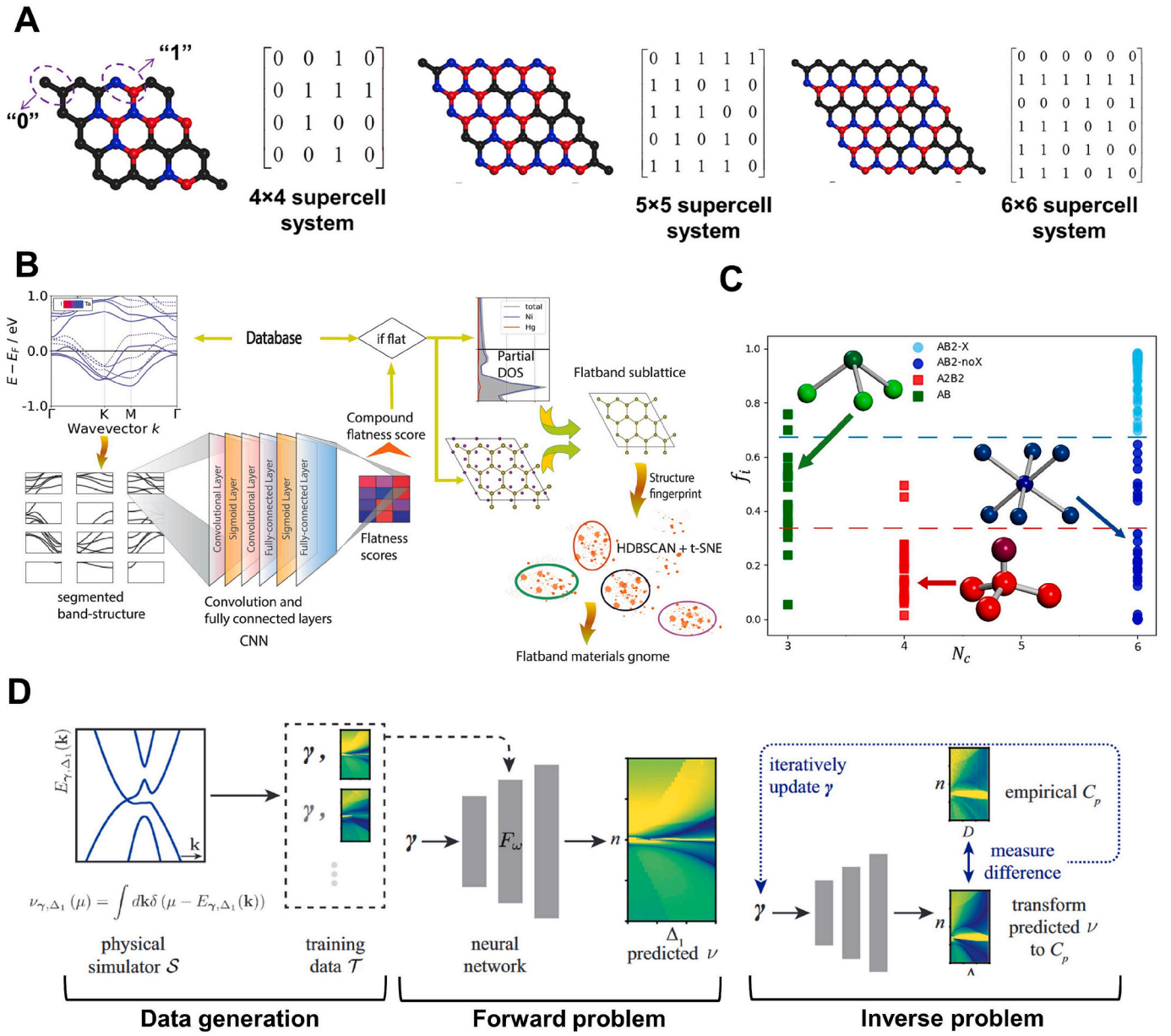

Fig. 11. Applications in the electronic properties of 2D materials based on DL methods. (A) Descriptors of 2D doped graphene supercell systems for CNN training. Reproduced under the terms of Creative Commons Attribution 4.0 International (CC BY 4.0) license from Ref. [94] published by Springer Nature. (B) CNN trained by segmented band structure images was applied to identify the flat band of 2D materials. Reproduced under the terms of Creative Commons Attribution 4.0 International (CC BY 4.0) license from Ref. [28] published by Springer Nature. (C) Structure type separation according to the coordinate number and ionicity. Reproduced with permission from Ref. [153] Copyright 2019 American Chemical Society. (D) Band structure parameters obtained from a physical simulator was used to train deep neural networks (DNNs), providing a more effective solution to the forward problem. Reproduced with permission from Ref. [54] Copyright 2024 American Physical Society.

images as multiple 2D numerical matrices, akin to RGB images classification, presents an effective approach to tackle this problem. Furthermore, when employing CNN models, it is crucial to thoroughly examine the layer-to-layer interaction, which can lead to intricate physical and chemical complexities.

Electron-electron correlations are crucial in condensed matter physics, influencing phenomena such as superconductivity and magnetism, as well as various technological applications. 2D materials with flat electronic bands offer a unique opportunity to study interaction-driven physics due to their localised electrons [148]. As shown in Fig. 11B, to address the problem of band crossings, Bhattacharya et al. utilised a CNN to detect the genome of flat band structure images on 2Dmatpedia. This approach enabled the identification of materials with flat bands extending throughout the Brillouin zone [28]. Unlike traditional methods that rely on parameterised bands and predefined bandwidth, the researchers proposed a unique and comprehensive approach. Instead of analysing parametrised bands directly, they utilised band structure images from a database. Furthermore, readily available band structure images provide a convenient method for identifying flat features in the bands, enabling the early-stage compilation of 2D materials databases through the integration of DL techniques. Beside utilizing structural information like lattice coordination patterns to differentiate various flat band materials, exploring electronic information such as the similarity of electronic states holds promise for future research. This approach could open new avenues for the discovery of novel 2D materials.

The electronic band structure of a crystalline solid is a vital and fundamental characteristic, holding great significance. It establishes a connection linking the energy levels of an electron in a solid to its momentum within the crystal. This relationship serves as the foundation for explaining and comprehending various material properties [149]. Consequently, accurately predicting the electronic band structure is a fundamental challenge in computational condensed matter physics. The typical approach for addressing the electronic structure dilemma of 2D materials from first principles involves using DFT with semi-local exchange-correlation functionals [150]. However, the accuracy of the DFT single-particle energies in modelling the electronic band structure is generally limited. Instead, the GW self-energy method is considered the most dependable approach for calculating the band structure. It accurately determines the quasiparticle (QP) band structure by incorporating exchange and many-body screening effects. A study compared the accuracy of calculating the bandgap in ten simple semiconductors and insulators using two different methods. The mean absolute error (MAE) for DFT-LDA was found to be 2.05 eV, while for non-self-consistent G_0W_0@LDA it was 0.31 eV, both compared to experimental reference values [151]. The improved accuracy of the GW method necessitates a more complex methodology and significantly greater computational resources. In practical applications, this restricts GW calculations to conducting small-scale investigations on relatively uncomplicated materials. However, the use of DL to predict material properties without relying on costly quantum mechanical calculations has been of growing interest. For example, del Rio et al. introduced a MLPs architecture that effectively learns the input-output patterns of the Kohn-Sham Eq. [80]. This model can rapidly and accurately predict the electronic density of states, a key outcome of DFT calculations. Furthermore, it has shown the capability to adapt and enhance its predictive accuracy when exposed to novel and diverse atomic configurations. While the fingerprints employed in this study yielded excellent results, the reliance on handcrafted features can introduce bias, potentially restricting the mapping between structure and density of states (DOS). By removing these constraints and enabling neural networks to autonomously discover the optimal mapping, there is an expectation of substantial improvements in accuracy, versatility, and transferability. Effective approaches include employing spherical and icosahedral CNNs, which maintain permutation, translation, and rotation invariance of the atomic structure.

For bandgap prediction, Rajan et al. employed a range of regression techniques to determine the bandgaps of MXenes crystals, using a dataset that included 76 G_0W_0 bandgaps along with representation-encoded atomic and structural characteristics [152]. Liang et al. employed atomic ionicity descriptors as representations to forecast GW bandgaps of various 2D materials, as illustrated in Fig. 11C [153]. In all prior research, the ML models were trained solely to predict the bandgap size instead of the full *k*-resolved band structure. Consequently, crucial features were overlooked such as the determination of whether the bandgap is direct or indirect, the curvature of the valence and conduction bands at critical points, and the position and characteristics of additional bands beyond the bandgap. Certainly, forecasting the complete band structure solely from the atomic arrangement is indeed a challenging endeavour. While theoretically possible, achieving this goal demands advanced DL models and substantial training data [150].

Currently, aligning band structure parameters derived from experimental data with theoretical calculations poses a significant challenge since the intricate nature of 2D material band structures. In other words, an efficient method to the forward problem does not ensure a rapid means to the inverse issue, which involves identifying the physical parameters in accordance with a given set of empirical data. Based on that, Fig. 11D presents a DNN-based approach that automated the comparison between numerical simulations and experimental data, facilitating the determination of the physical parameters of the band structure corresponding to specific experimental datasets with minimal human intervention [54]. This method was successfully employed to a high-quality experimental dataset of ABA graphene using the penetration capacitance method, and their results showed no significant differences compared to the reference model. In the future, to thoroughly explore the connection between the simulated quantities and experimental measurements, it will be necessary to take into account more factors when analysing penetration field capacitance. These factors include the non-trivial screening of the electric field, the presence of parasitic capacitance, and the geometric capacitances of the gates [154].

4.2. Mechanical properties

One of the most significant applications of 2D materials is their incorporation into mechanical, civil, and aerospace reinforcement structures. Nanocomposites based on 2D materials utilize these building blocks arranged in layered structures, providing enhanced stiffness, strength, and energy dissipation capabilities [155]. The mechanical strength of the 2D nanosheets can be quantified using various parameters like Young's modulus, bending rigidity, ultimate fracture strength, fracture strain, etc. [156]. In recent years, as shown in Table 3, the integration of DL algorithms and HTC has accelerated the discovery of the mechanical properties of 2D materials [8].

To consider several factors when investigating the mechanical properties of graphene using the MLPs model, Zhang et al. used MD simulation to collect a dataset comprising 1440 data points, involving four parameters (system temperature, strain rate, single vacancy defect, and chirality) and three target values (fracture strain, fracture strength, and Yong's modulus) [67]. Once trained, the MLPs model achieved rapid predictions of the mechanical properties of graphene within milliseconds (Fig. 12A). However, computational constraints limited the analysis to four impact scenarios and single-point defects. Addressing multiple factors through more robust DL models will necessitate substantial future research efforts. Moreover, shallow networks such as MLPs lack the capability to accurately forecast how the distribution of defects impacts the fracture stress of 2D materials, primarily because of the complex nonlinear stress-strain relationships and anisotropic fracture properties. To solve that, Dewapriya et al. developed CNNs trained by graphene images containing random distributions of vacancy defects to predict the fracture stress [52]. In order to collect data using MD simulation and generate different spatial distributions of defects, it is necessary to ensure that the length and width of the simulated graphene sheets are at least ten times larger than half the initial crack length. This is done to prevent the finite dimensions of the sheets from affecting the

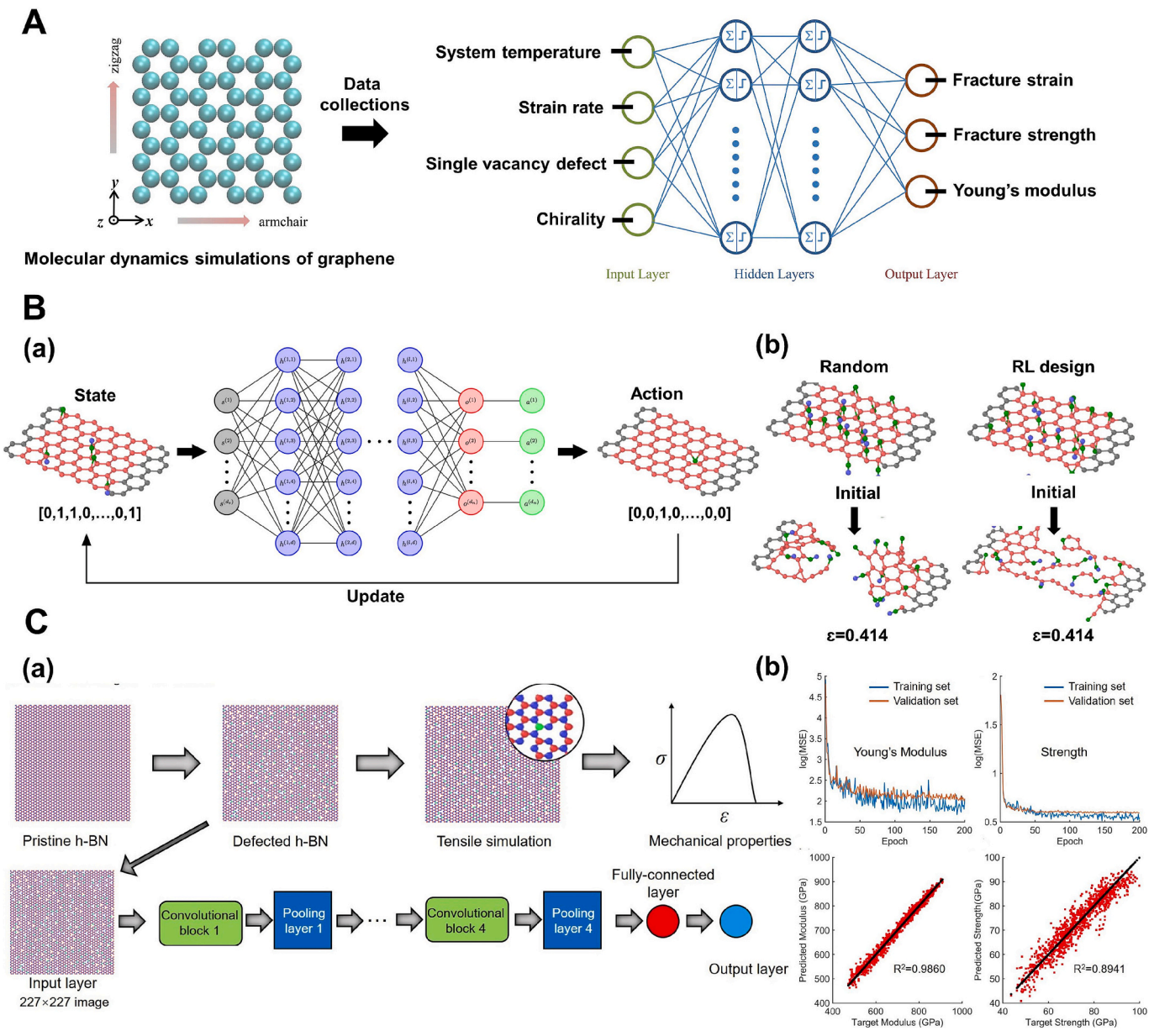

Fig. 12. Applications in the mechanical properties of 2D materials based on DL methods. (A) Accelerated discoveries of mechanical properties of graphene using molecular simulation (MD) and MLPs. Reproduced with permission from Ref. [67]. Copyright 2019 Elsevier. (B) (a) The state transition process involving policy network and action for the optimization of GO in toughness design and (b) the rupturing comparison between a random GO and an RL-designed GO. Reproduced under the terms of Creative Commons Attribution 4.0 International (CC BY 4.0) license from Ref. [82] published by Springer Nature (C). (a) The workflow for the investigation of defect-engineered mechanical properties of h-BN based on MD and CNNs, and (b) the predicted results of Young's modulus and strength. Reproduced with permission from Ref. [91]. Copyright 2023 Elsevier.

results of the MD simulation. In addition, they also indicated that transfer learning using pre-trained CNNs such as AlexNet, originally trained for a classification task, could have some inherent limitations in the property regression prediction of 2D materials due to the differences in model parameters and dataset.

At present, based on DL and HTC techniques, some studies tried to use the functional group distribution information for predicting the mechanical properties of 2D material when given a specific chemical composition combined with DL methods. For example, due to the composition of the graphene basal plane (GBP), the mechanical property of graphene oxide (GO) can be affected by the types, number, and distribution of oxygen-containing functional groups such as hydroxyl (C—OH), epoxide (C—O—C), and carboxyl (O═C—OH). Fig. 12B(a) illustrates the application of policy-gradient RL models in designing mechanically robust GOs by adjusting the distribution of functional groups. Rather than approaching the optimization problem as a single decision to choose the best functional group distribution, researchers viewed the assignment of functional groups as a sequential decision-making process and employed RL to address it. Moreover, through comparing the molecular structure and detailed failure behaviour, the GO designed by RL contained more atoms that help absorb energy, leading to a higher toughness as depicted in Fig. 12B(b) [82]. However, the investigation remains incomplete as defect-engineered 2D materials typically consist of multiple types of defects, including vacancies, grain boundaries, and substitutional atoms. These defects interact with each other, leading to complex effects on the material's structure-properties relationships. Moreover, the presence of multiple types of defects can result in a multitude of potential configurations that need to be

investigated, further complicating the issue. To address the above problems, as illustrated in Fig. 12C(a), Shen and Zhu introduced a novel approach using CNNs to accurately forecast the mechanical properties of defect-engineered h-BN [91]. By analysing RGB images encoded by defected atomistic configuration, they were able to gain valuable understanding into the impact of coexisting defect types on the material's mechanical properties based on CNNs. As shown in Fig. 12C(b), the CNNs had the ability to predict Young's modulus and tensile strength with high R^2 reaching 0.986 and 0.894, respectively. Through extensive MD simulations, this study not only shed light on the relationship between defect types and mechanical properties in 2D materials, but also demonstrated the potential of CNN in quickly predicting the mechanical properties of defect-engineered 2D materials by analysing the image which encodes the defected atomistic configuration. However, it is worth noting that the thermodynamics of the designed GOs were not considered in this study, which raises concerns about the thermal stability of the resulting GO configurations. Applying a thermodynamic criterion and modifying the reward in the RL algorisms can be beneficial for selecting the thermally stable graphene oxides in future works. In addition, according to Table 3, it was not difficult to find that advanced DNNs can quickly evaluate a large number of configurations with specific defect conditions, making it easier to screen configurations with desired properties. By efficiently analysing configurations created with a certain vacancy or doping rate, the desired properties are ranked. This rapid screening capability is essential for creating algorithms for accelerated design, such as neural network-based methods for designing graphene kirigami with high stretchablity [157] or porous graphene with low thermal conductivity [158].

5. Discovery and design of 2D materials

2D Materials can be classified in different ways depending on factors like their composition, which includes the type and amount of atoms present, whether they have stoichiometric or non-stoichiometric elements, and their structural characteristics like crystallography, nanostructure, and microstructure [8]. In optics, mechanics, and electronics, 2D materials exhibit unique characteristics due to differences in their atomic compositions, stoichiometries, and structures. The traditional approach to 2D material design usually starts with determining the material parameters in order to achieve the desired properties. The process of designing new materials typically includes various stages such as molecular [12]. These phases involve conducting multiple experiments and simulations, leading to a considerable investment of time and resources, as well as significant trial-and-error costs. One solution to these challenges is HTC screening, which uses first-principles calculations. By using combinatorial enumeration, this approach creates a virtual chemical library that simplifies the process of screening potential candidates for future chemical synthesis. As a result, the efficiency of material design and discovery is greatly improved. Nevertheless, when it comes to addressing significant challenges, the time and computational costs increase exponentially due to the immense size of the chemical search space [159]. Moreover, the development of a virtual chemical library greatly depends on the knowledge and intuition of materials scientists. On the other hand, inverse design begins with a material that exhibits specific desired functions and then works backwards to deduce its chemical makeup and arrangement. This method helps identify the best material design parameters based on desired properties, essentially creating an inverse pathway from performance goals to design criteria [160]. In recent years, data-driven DL has emerged as the fourth paradigm in materials science, driven by the growing volume of experimental and simulation data [21,22].

Deep generative models have been widely used in inverse design, particularly in the extraction of latent material design knowledge and principles from complex data. These models enable the creation of innovative materials with customised functionalities, eliminating the subjective influence of researchers' experience or intuition [18]. At present, generative models, such as VAEs, GANs, reinforcement learning, and RNNs are considered optimal inverse design methods for tackling the computational challenges associated with searching. For the systematic design of 2D materials, a challenge still exists due to the constrained pool of fewer than 100 experimentally synthesised 2D materials. Based on that, Lyngby and Thygesen showed that a CDVAE can create 2D materials with a wide range of chemical and structural variations, resulting in formation energies that closely match those of the original training structures [17]. To compare, they initially utilised the lattice decoration protocol (LDP) to replace atoms in the initial structure with atoms of similar chemical characteristics as illustrated in Fig. 13A. They screened all possible seed structures with single and double substitutions based on these substitution relationships. For instance, the seed structure MoS_2 produces six MX_2 structures with M = Mo, W and X = O, S, Se. After eliminating duplicates, a total of 14,192 distinct 2D crystals with the same simplified formula and space group were selected for relaxation using DFT. Moreover, as shown in Fig. 13B, by training 2615 2D materials obtained from C2DB as seed structures for lattice decoration, 11,630 materials were generated using CDVAE model such as Cu_2F_6 and YBrSe (Fig. 13D). Furthermore, as shown in Fig. 13C, t-SNE analysis was used to obtain a global overview of the data sets obtained from CDVAE and LDP methods, and they found that the generative model and lattice decoration method complement each other, leading to materials that possess similar stability properties despite their distinct crystal structures and chemical compositions. As a result, the CDVAE, serving as a proficient and dependable crystal generation machine, greatly broadened the range of 2D materials available.

Before systematically studying and designing 2D materials, formation energy as a crucial feature can be used to identify whether the predicted materials are stable initially. Grain boundaries (GBs), defects, and lateral heterostructures (LHS) in 2D materials have a great impact on the formation energy. Unlike the conventional evolutionary search methods using genetic algorithm (GA), Zhang et al. combined GNN and an evolutionary algorithm to design novel lateral interfaces of blue phosphorene by identifying GB structures [161]. The GNN models trained by Tersoff formalism (a ML bond order potential) and DFT results can predict structural energy under 5 % mean absolute error (MAE) on the DFT energy hypersurface. Significantly, in order to create a precise DFT surrogate, the total energy of each structure serves as the target value for the GNN model, significantly making it a graph-level regression task. As shown in Fig. 14A(a), in the final stage of the model, the node embedding vectors were input into a global pooling layer to produce a vector that represented the entire graph, and the data within this vector were transformed into a single overall energy predicted by MLPs. Furthermore, the GNN model utilised multi-objective GA searching to forecast the structures of 2D lateral interfacing nanosheets containing GB defects (Fig. 14A(b)).

Recently, DL algorithms based on sequential neural networks have shown excellence in sequence learning and sequence generation using the material composition. As illustrated in Fig. 14B(a), Dong et al. introduced a generative material design pipeline called material transformer generator (MTG), which utilised self-learning neural language models based on transformers. This model successfully identified four new DFT-validated stable 2D materials, demonstrating its potential in the discovery of novel 2D materials [18]. During the data preparation stage, they trained a series of blank language models for materials (BLMM) composition generators using known 2D formulas sourced from various open datasets (C2DB, MC2D, 2DMatPedia, and V2DB) to generate new 2D formulas. Fig. 14B(c) displays the distribution of the top 50 element pairs in the 2DMatPedia and generated datasets, which can assess the composition generation capabilities of BLMM. However, only the H—O element pair is represented in the top five of generation results. It was noteworthy that these two datasets share only two element pairs in common among the top ten most frequent pairs. Furthermore, t-SNE was also used to verify the newly generated compositions. Each point depicted in Fig. 14B(d) corresponds to an

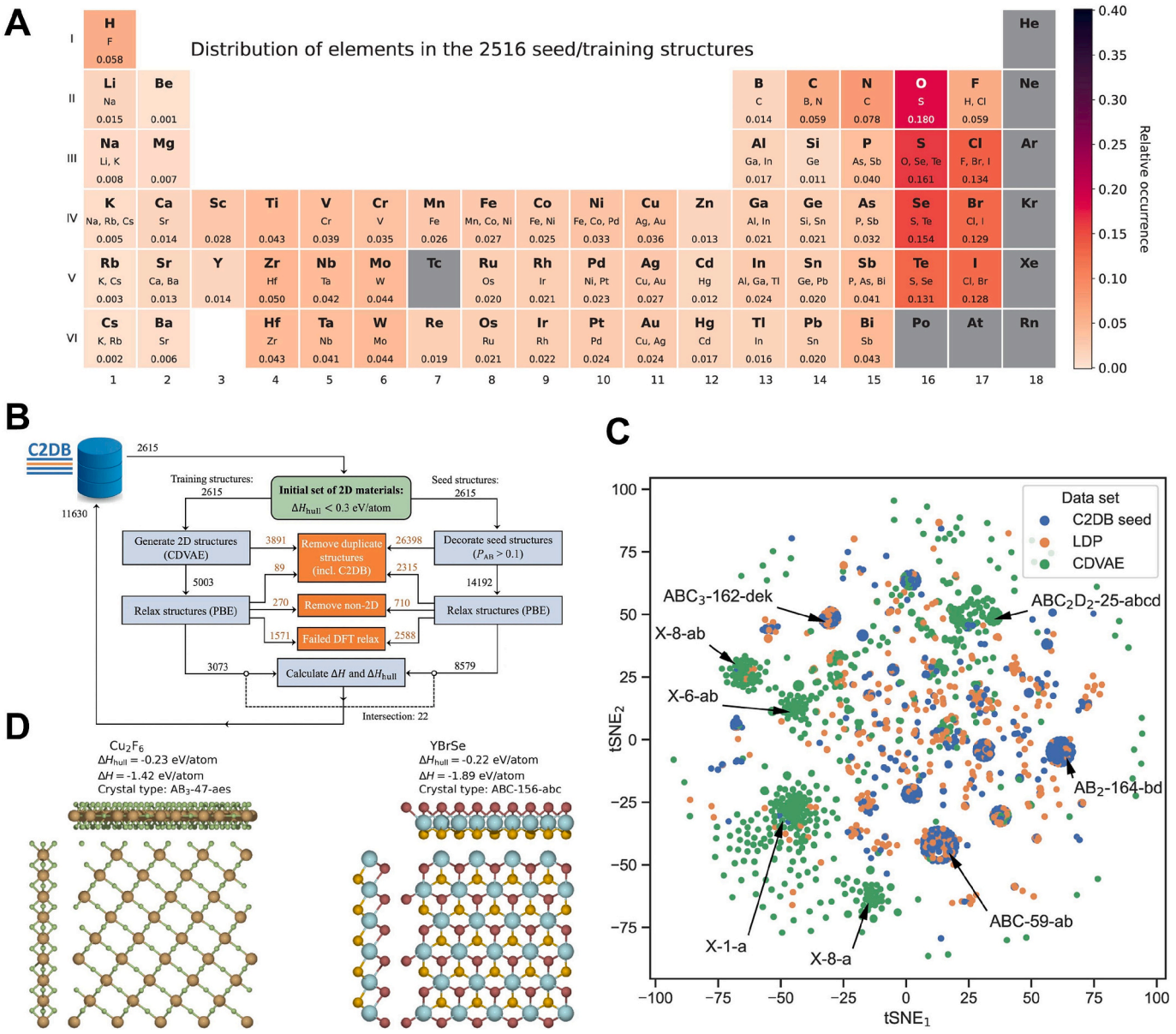

Fig. 13. Data-driven discovery of 2D materials by deep generative models. (A) Heat map of the relative occurrence of each element in the 2D materials used to train the CDVAE model. (B) The workflow diagram of CDVAE. (C) The t-SEN visualization of the structural diversity using CDVAE model. (D) The structures of Cu_2F_6 and YBrSe generated by CDVAE. Reproduced with permission from Ref. [17]. Copyright 2022 Springer Nature.

individual formula, with the colors indicating the respective formation energy levels. As a result, the newly created samples are consistently situated near the existing samples, and this result indicated that the BLMM model had strong interpolation capability when generating new samples by filling in blanks in existing materials. This means that the newly generated samples are consistently located near known samples. As a result, GaBrO and $CuBR_3$ were generated as new 2D materials using TSCP and CSPML algorithm based on the templated 2D structures (Fig. 14B(b)). This method significantly improved the performance in machine learning potential-based relaxation and DFT-based relaxation procedures.

In addition, in the inverse design process, a sampling or optimization procedure is employed to choose a subset of the entire design space in order to identify the optimal design candidate. Typically, the sampling process is guided by a global optimization algorithm, which is complemented by evaluating the performance of functions through atomic simulation. The global optimization techniques used in inverse design can be classified into various categories, such as gradient-based and gradient-free methods, generative approaches using neural networks, and Bayesian methods utilizing surrogate evaluation models. In comparison to traditional gradient-based search operators or genetic search operators using crossover and random mutation, active learning with VAE networks can actively explore the design space to search for candidates with desired properties or identify informative samples for constructing properties predicting models. This allows for the training of optimal prediction models using existing data and a minimal number of newly labelled samples. Based on that, Xin et al. introduced a novel approach to generative inverse design by combining autoencoder and GAN model, aiming to identify new materials with desired properties across the entire chemical design space [140]. To describe material compositions easily, they used one-hot encoding method to describe each material as a sparse matrix with 0/1 cell values as shown in Fig. 14C(a). As a result, the predicted $SrClF_3$ with a wide band gap showed a low formation energy (−2.55 eV/atom) and stable semiconductor with *P1* space-group symmetry after DFT calculations (Fig. 14C(b)). To further accelerate the screening of 2D materials, it is

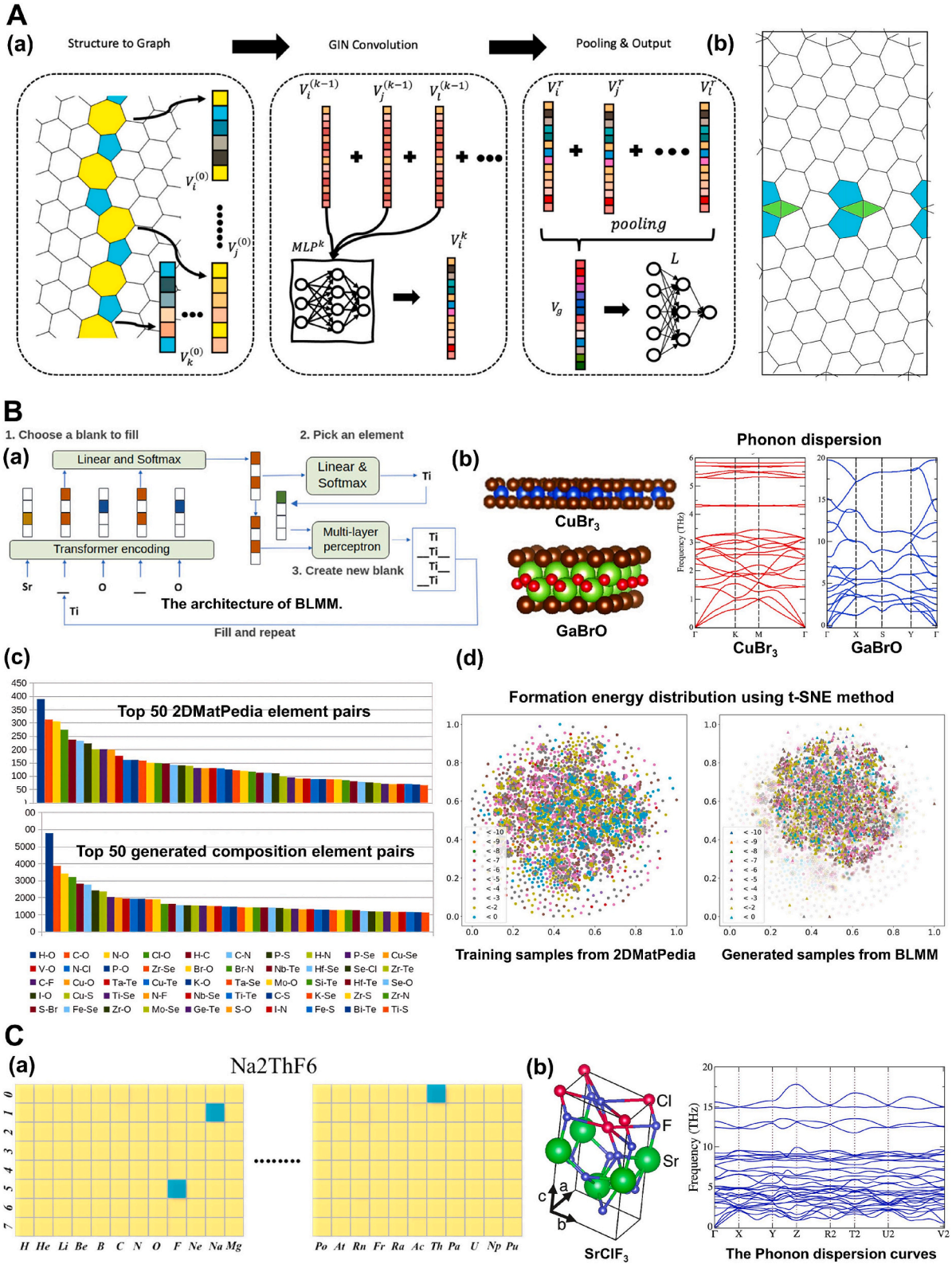

Fig. 14. Data-driven discovery of 2D materials using DL model. (A) (a) Schematic architecture of the GNN model and (b) The structure verified by multi-objective genetic algorithm (MOGA) and the proposed GNN model using the Tersoff potential. Reproduced under the terms of Creative Commons Attribution 4.0 International (CC BY 4.0) license from Ref. [161] published by American Chemical Society. (B) The architecture of BLMM. Using (a) transformer to encode the given formula sequence with blanks. After embedding every element and blank, choosing one single blank to predict the most suitable element for this blank based on the context information and filling the original blank. (b) The new 2D structure of $CuBr_3$ and GaBrO discovered by MTG pipeline with 0 and $<$ 0.05 eV E_{hull} energy, respectively. (c) Elements distribution in training and generating samples. (d) Formation energy distributions in the training (2DMatPedia) and generated samples based on MTG. Reproduced with permission from Ref. [18]. Copyright 2023 Wiley-VCH. (C) (a) One-hot encoding of material compositions based on chemical elements for the active generative model training. (b) The structure and phonon dispersion of $SrClF_3$. Reproduced with permission from Ref. [140]. Copyright 2021 American Chemical Society.

more effective to utilize active learning to enhance the DL model with minimal expensive sample annotation (such as DFT), especially when the number of labelled training samples is limited.

6. Challenge and solution strategies

DL methods have rapidly emerged as crucial and adaptable tool in the fields of materials science and chemistry. Significant progress has been made in developing specialised methods to meet the unique needs of these disciplines and in applying DL for tasks such as characterizing, discovering, and designing 2D materials. However, despite their promising potential, several challenges must be overcome to effectively use DL in this area.

Firstly, in the past, data was perceived simply as a collection of observable facts aiding knowledge acquisition. However, looking forward, data will be recognised as information reflecting intricate interactions among various factors. This shift in perspective will necessitate more advanced data management techniques. Data management involves the various stages of collecting, storing, screening, labelling, annotating, enhancing, evaluating, removing, and virtualizing data. It's a complex and ongoing process that requires global collaboration between researchers and governments. In the past, datasets were static, and DL was primarily applied to these specific datasets in previous materials science research. However, there is now a strong emphasis on data iteration. In particular, iterative methods like active learning are increasingly used to improve model performance. This change also necessitates a more systematic approach to data management. Concurrently, the creation of extensive, high-quality datasets that are openly accessible can significantly bolster DL research in studying 2D materials. Compiling and sharing comprehensive datasets on 2D materials makes it easier to train more accurate and efficient models.

Secondly, researchers generally rely on keyword searches in publication databases and manually extract relevant information, a process hindered by the sheer volume of publications. To address this inefficiency, DL-based text extraction methods, particularly NLP techniques, are increasingly utilised to mine chemical information from the literature. These methods offer robust batch extraction capabilities, significantly enhancing the efficiency of data retrieval from publications. Nevertheless, it is crucial to thoroughly evaluate the dependability of the data published in literature sources. Variability in data reliability may arise due to differences in measurement techniques, particularly in the field of catalysis, where various factors such as the shape of the reactor and stirring speed can influence results. Moreover, discrepancies may exist in how catalytic reactions are measured, further complicating data compatibility. To address these challenges, it is vital for materials scientists, catalyst scientists, and computer scientists to work closely together. Moreover, it is important to establish a standardised format for recording data and implement strong protocols to assess the credibility of published data. These steps are crucial in generating DL models with good performance. In addition, a variety of state-of-the-art DL layers and models are available from libraries such as PyTorch, TensorFlow, Keras, and Cntk, which can accelerate the development of screening 2D material. These open-source frameworks make DL-based 2D material prediction models easy to implement, adapt, transfer, train, and use and are suitable for beginners contacting and learning DL algorithms.

Thirdly, small datasets often lead to challenges such as imbalanced data and overfitting or underfitting of models. These issues arise due to the limited scale of the data and the presence of either excessively high or low feature dimensions. Addressing these challenges has long been a significant concern in materials deep learning. To address that, besides data augmentation methods for 2D material images datasets, transfer learning, active learning, and model generation techniques are the novel methods to address the problems of small datasets. Transfer learning is the process of acquiring knowledge from solving a particular problem and utilizing it to solve a different yet related problem. On the other hand, active learning involves prioritizing the labelling of data in order to maximize the impact on training DL models. Model generation techniques, such as GANs, offer a novel solution to the small sample size problem by synthesizing new data instances that closely resemble real 2D materials. By leveraging these novel methodologies synergistically, researchers can overcome the limitations posed by limited data availability and develop more accurate and reliable predictive models for various applications in 2D materials science and engineering.

Fourthly, it is important to be cautious when interpreting and validating predictions made using DL methods. To do so, experiments or computational simulations should be conducted, as the predictions from DL are not based on a deep understanding of the underlying physics of 2D materials. Instead, DL relies on identifying correlations between input features and desired outcomes to make predictions. Taking neural networks as an example, they create intricate connections between multiple nodes in hidden layers, disregarding any theoretical knowledge about 2D materials. Consequently, this approach may lead to incorrect correlations from a theoretical standpoint. Essentially, DL methods are often seen as black-box algorithms. Their predictions are not easily understood, and their internal mechanisms and logic are unclear, making them challenging to interpret. In the near future, challenges in DL for the study of 2D materials may be overcome as researchers globally work together to create extensive and precise libraries of 2D materials. Additionally, cutting-edge DL studies are integrating physics-based constraints into their models, allowing for more theoretically sound predictions.

Lastly, the combination of digital light processing and robotics offers a promising avenue for the development of automated systems capable of efficiently producing a diverse range of 2D materials and complex heterogeneous structures. This technological advancement holds significant potential for the intelligent preparation of 2D materials and the design of devices. Ongoing research in 2D materials primarily remains partially independent, and a notable technical hurdle lies in establishing a self-contained and automated process for materials experimentation. The advent of autonomous robotic scientists will bring about a substantial transformation in the current approach of human-machine collaboration. Furthermore, there is considerable potential for further exploration into the utilization of DL for the investigation of 2D materials. Further research is required in the field of inverse engineering 2D materials to develop materials with specific thermal conductivity and mechanical properties, except for materials with the desired band gaps. To screen 2D materials with good performance, it is crucial to consider not only their electronic, mechanical, and thermal properties, but also their optical properties, superconductivity, and toxicity. The incorporation of DL into the investigation of 2D materials is propelling the field of 2D materials science forward by addressing the challenges of conventional experimental and theoretical calculation techniques. Despite some advancements, there are still many obstacles to overcome in the future and this review hopes to provide some insight into the challenges mentioned above.

7. Conclusions

In the past few years, there has been a significant growth in the number of 2D materials and their heterostructures. This growth has outpaced the capabilities of traditional experimental and computational methods. However, a new tool called DL has emerged as a valuable addition to these conventional approaches. DL provides intelligent study opportunities and complements the existing methods effectively. The characteristics of 2D materials, gathered from materials databases, experimental observations, and computational data, are used as input features for training different DL models such as CNNs, RNNs, and GANs. By analysing the complex relationships between input features and correlating them with desired outputs, these trained DL models offer valuable insights and accurate predictions, facilitating the exploration, discovery, and creation of 2D materials. In addition, to further improve the performance of DL models, effective data description methods and

perspectives of 2D materials need to be considered, which are beneficial for understanding models and enabling researchers to quickly grasp interdisciplinary knowledge. Furthermore, by combining theoretical and experimental approaches in a complementary manner, the suitability of deep learning-predicted 2D materials can be thoroughly validated, and their feasibility in practical applications can be assessed. The potential for advancements in the intersection of DL and 2D materials is limitless, and this review discussed here is just the beginning of an exciting journey.

Declaration of competing interest

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

Data availability

Data will be made available on request.

Acknowledgements

This work was supported by the National Key Research and Development Program of China [grant number 2023YFD2101000 and 2023YFD2101002]; the National Natural Science Foundation of China [grant number 32272466 and U23A20267); the Guangdong Basic and Applied Basic Research Foundation [grant number 2024A1515011498]; Outstanding Youth Funding of Guangdong Province Science and Technology Innovation (NYQN2024003) and the TCL Young Scholar Foundation.

References

[1] K.S. Novoselov, A.K. Geim, S.V. Morozov, D. Jiang, Y. Zhang, S.V. Dubonos, I. V. Grigorieva, A.A. Firsov, Electric field effect in atomically thin carbon films, Science 306 (5696) (2004) 666–669, https://doi.org/10.1126/science.1102896.
[2] P.V. Shinde, A. Tripathi, R. Thapa, C. Sekhar Rout, Nanoribbons of 2D materials: a review on emerging trends, recent developments and future perspectives, Coord. Chem. Rev. 453 (2022) 214335, https://doi.org/10.1016/j.ccr.2021.214335.
[3] F. Chen, Q. Tang, T. Ma, B. Zhu, L. Wang, C. He, X. Luo, S. Cao, L. Ma, C. Cheng, Structures, properties, and challenges of emerging 2D materials in bioelectronics and biosensors, InfoMat 4 (5) (2022) e12299, https://doi.org/10.1002/inf2.12299.
[4] J.H. Jørgensen, A.G. Čabo, R. Balog, L. Kyhl, M.N. Groves, A.M. Cassidy, A. Bruix, M. Bianchi, M. Dendzik, M.A. Arman, L. Lammich, J.I. Pascual, J. Knudsen, B. Hammer, P. Hofmann, L. Hornekaer, Symmetry-driven band gap engineering in hydrogen functionalized graphene, ACS Nano 10 (12) (2016) 10798–10807, https://doi.org/10.1021/acsnano.6b04671.
[5] H.G. Ji, P. Solís-Fernández, D. Yoshimura, M. Maruyama, T. Endo, Y. Miyata, S. Okada, H. Ago, Chemically tuned p- and n-type WSe2 monolayers with high carrier mobility for advanced electronics, Adv. Mater. 31 (42) (2019) 1903613, https://doi.org/10.1002/adma.201903613.
[6] A.E. Naclerio, P.R. Kidambi, A review of scalable hexagonal boron nitride (h-BN) synthesis for present and future applications, Adv. Mater. 35 (6) (2023) 2207374, https://doi.org/10.1002/adma.202207374.
[7] H. Kwon, S.W. Seo, T.G. Kim, E.S. Lee, P.T. Lanh, S. Yang, S. Ryu, J.W. Kim, Ultrathin and flat layer black phosphorus fabricated by reactive oxygen and water rinse, ACS Nano 10 (9) (2016) 8723–8731, https://doi.org/10.1021/acsnano.6b04194.
[8] B. Ryu, L. Wang, H. Pu, M.K.Y. Chan, J. Chen, Understanding, discovery, and synthesis of 2D materials enabled by machine learning, Chem. Soc. Rev. 51 (6) (2022) 1899–1925, https://doi.org/10.1039/D1CS00503K.
[9] J. Pan, S. Lany, Y. Qi, Computationally driven two-dimensional materials design: what is next? ACS Nano 11 (8) (2017) 7560–7564, https://doi.org/10.1021/acsnano.7b04327.
[10] R.X. Yang, C.A. McCandler, O. Andriuc, M. Siron, R. Woods-Robinson, M. K. Horton, K.A. Persson, Big data in a nano world: a review on computational, data-driven design of nanomaterials structures, properties, and synthesis, ACS Nano 16 (12) (2022) 19873–19891, https://doi.org/10.1021/acsnano.2c08411.
[11] H. Demir, H. Daglar, H.C. Gulbalkan, G.O. Aksu, S. Keskin, Recent advances in computational modeling of MOFs: from molecular simulations to machine learning, Coord. Chem. Rev. 484 (2023) 215112, https://doi.org/10.1016/j.ccr.2023.215112.
[12] H. He, Y. Wang, Y. Qi, Z. Xu, Y. Li, Y. Wang, From prediction to design: recent advances in machine learning for the study of 2D materials, Nano Energy 118 (2023) 108965, https://doi.org/10.1016/j.nanoen.2023.108965.
[13] J.A. Keith, V. Vassilev-Galindo, B. Cheng, S. Chmiela, M. Gastegger, K.-R. Müller, A. Tkatchenko, Combining machine learning and computational chemistry for predictive insights into chemical systems, Chem. Rev. 121 (16) (2021) 9816–9872, https://doi.org/10.1021/acs.chemrev.1c00107.
[14] P. Hundi, R. Shahsavari, Deep learning to speed up the development of structure–property relations for hexagonal boron nitride and graphene, Small 15 (19) (2019) 1900656, https://doi.org/10.1002/smll.201900656.
[15] B. Han, Y. Lin, Y. Yang, N. Mao, W. Li, H. Wang, K. Yasuda, X. Wang, V. Fatemi, L. Zhou, J.I.J. Wang, Q. Ma, Y. Cao, D. Rodan-Legrain, Y.-Q. Bie, E. Navarro-Moratalla, D. Klein, D. MacNeill, S. Wu, H. Kitadai, X. Ling, P. Jarillo-Herrero, J. Kong, J. Yin, T. Palacios, Deep-learning-enabled fast optical identification and characterization of 2D materials, Adv. Mater. 32 (29) (2020) 2000953, https://doi.org/10.1002/adma.202000953.
[16] S. Wu, Z. Wang, H. Zhang, J. Cai, J. Li, Deep learning accelerates the discovery of two-dimensional catalysts for hydrogen evolution reaction, Energy Environ. Mater. 6 (1) (2023) e12259, https://doi.org/10.1002/eem2.12259.
[17] P. Lyngby, K.S. Thygesen, Data-driven discovery of 2D materials by deep generative models, npj Comput. Mater. 8 (1) (2022) 232, https://doi.org/10.1038/s41524-022-00923-3.
[18] R. Dong, Y. Song, E.M.D. Siriwardane, J. Hu, Discovery of 2D materials using transformer network-based generative design, Adv. Intell. Syst. 5 (12) (2023) 2300141, https://doi.org/10.1002/aisy.202300141.
[19] M.T. Dau, M. Al Khalfioui, A. Michon, A. Reserbat-Plantey, S. Vézian, P. Boucaud, Descriptor engineering in machine learning regression of electronic structure properties for 2D materials, Sci. Rep. 13 (1) (2023) 5426, https://doi.org/10.1038/s41598-023-31928-7.
[20] J. Wei, X. Chu, X.-Y. Sun, K. Xu, H.-X. Deng, J. Chen, Z. Wei, M. Lei, Machine learning in materials science, InfoMat 1 (3) (2019) 338–358, https://doi.org/10.1002/inf2.12028.
[21] B. Lu, Y. Xia, Y. Ren, M. Xie, L. Zhou, G. Vinai, S.A. Morton, A.T.S. Wee, W.G. van der Wiel, W. Zhang, P.K.J. Wong, When machine learning meets 2D materials: a review, Adv. Sci. 11 (13) (2024) 2305277, https://doi.org/10.1002/advs.202305277.
[22] X. Meng, C. Qin, X. Liang, G. Zhang, R. Chen, J. Hu, Z. Yang, J. Huo, L. Xiao, S. Jia, Deep learning in two-dimensional materials: characterization, prediction, and design, Front. Phys. 19 (5) (2024) 53601, https://doi.org/10.1007/s11467-024-1394-7.
[23] J. Lin, Z. Liu, Y. Guo, S. Wang, Z. Tao, X. Xue, R. Li, S. Feng, L. Wang, J. Liu, H. Gao, G. Wang, Y. Su, Machine learning accelerates the investigation of targeted MOFs: performance prediction, rational design and intelligent synthesis, Nano Today 49 (2023) 101802, https://doi.org/10.1016/j.nantod.2023.101802.
[24] L. Himanen, A. Geurts, A.S. Foster, P. Rinke, Data-driven materials science: status, challenges, and perspectives, Adv. Sci. 6 (21) (2019) 1900808, https://doi.org/10.1002/advs.201900808.
[25] Z. Wang, Z. Sun, H. Yin, X. Liu, J. Wang, H. Zhao, C.H. Pang, T. Wu, S. Li, Z. Yin, X.-F. Yu, Data-driven materials innovation and applications, Adv. Mater. 34 (36) (2022) 2104113, https://doi.org/10.1002/adma.202104113.
[26] A. Jain, S.P. Ong, G. Hautier, W. Chen, W.D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, K.A. Persson, Commentary: the materials project: a materials genome approach to accelerating materials innovation, APL Mater. 1 (1) (2013) 011002, https://doi.org/10.1063/1.4812323.
[27] A. Elrashidy, J. Della-Giustina, J.-A. Yan, Accelerated data-driven discovery and screening of two-dimensional magnets using graph neural networks, J. Phys. Chem. C 128 (14) (2024) 6007–6018, https://doi.org/10.1021/acs.jpcc.3c07246.
[28] A. Bhattacharya, I. Timokhin, R. Chatterjee, Q. Yang, A. Mishchenko, Deep learning approach to genome of two-dimensional materials with flat electronic bands, npj Comput. Mater. 9 (1) (2023) 101, https://doi.org/10.1038/s41524-023-01056-x.
[29] V. Gupta, K. Choudhary, B. DeCost, F. Tavazza, C. Campbell, W.-K. Liao, A. Choudhary, A. Agrawal, Structure-aware graph neural network based deep transfer learning framework for enhanced predictive analytics on diverse materials datasets, npj Comput. Mater. 10 (1) (2024) 1, https://doi.org/10.1038/s41524-023-01185-3.
[30] H.M. Sayeed, W. Smallwood, S.G. Baird, T.D. Sparks, NLP meets materials science: quantifying the presentation of materials data in literature, Matter 7 (3) (2024) 723–727, https://doi.org/10.1016/j.matt.2023.12.032.
[31] D. Cheng, W. Sha, Z. Xu, S. Li, Z. Yin, Y. Lang, S. Tang, Y.-C. Cao, AtomGAN: unsupervised deep learning for fast and accurate defect detection of 2D materials at the atomic scale, Sci. China Inf. Sci. 66 (6) (2023) 160410, https://doi.org/10.1007/s11432-022-3757-x.
[32] Y. Qi, D. Hu, Z. Wu, M. Zheng, G. Cheng, Y. Jiang, Y. Chen, Deep learning assisted Raman spectroscopy for rapid identification of 2D materials, in: ArXiv Physics, 2023, https://doi.org/10.48550/arXiv.2312.01389. ArXiv:2312.01389.
[33] C.J.A.P.A. Ekuma, Computational toolkit for predicting thickness of 2D materials using machine learning and autogenerated dataset by large language model, Arxiv Preprint, 2024, https://doi.org/10.48550/arXiv.2405.15131. Arxiv: 2405.15131.
[34] F. Ramezani, S. Parvez, J.P. Fix, A. Battaglin, S. Whyte, N.J. Borys, B. M. Whitaker, Automatic detection of multilayer hexagonal boron nitride in optical images using deep learning-based computer vision, Sci. Rep. 13 (1) (2023) 1595, https://doi.org/10.1038/s41598-023-28664-3.
[35] K. Choudhary, B. DeCost, C. Chen, A. Jain, F. Tavazza, R. Cohn, C.W. Park, A. Choudhary, A. Agrawal, S.J.L. Billinge, E. Holm, S.P. Ong, C. Wolverton,

Recent advances and applications of deep learning methods in materials science, npj Comput. Mater. 8 (1) (2022) 59, https://doi.org/10.1038/s41524-022-00734-6.
[36] P. Mishra, A. Biancolillo, J.M. Roger, F. Marini, D.N. Rutledge, New data preprocessing trends based on ensemble of multiple preprocessing techniques, TrAC Trends Anal. Chem. 132 (2020) 116045, https://doi.org/10.1016/j.trac.2020.116045.
[37] P. Huang, R. Lukin, M. Faleev, N. Kazeev, A.R. Al-Maeeni, D.V. Andreeva, A. Ustyuzhanin, A. Tormasov, A.H. Castro Neto, K.S. Novoselov, Unveiling the complex structure-property correlation of defects in 2D materials based on high throughput datasets, npj 2D Mater. Appl. 7 (1) (2023) 6, https://doi.org/10.1038/s41699-023-00369-1.
[38] F. Bertoldo, S. Ali, S. Manti, K.S. Thygesen, Quantum point defects in 2D materials - the QPOD database, npj Comput. Mater. 8 (1) (2022) 56, https://doi.org/10.1038/s41524-022-00730-w.
[39] M.E. Saleh, Y.M. Wazery, A.A. Ali, A systematic literature review of deep learning-based text summarization: techniques, input representation, training strategies, mechanisms, datasets, evaluation, and challenges, Expert Syst. Appl. 252 (2024) 124153, https://doi.org/10.1016/j.eswa.2024.124153.
[40] L. Hawizy, D.M. Jessop, N. Adams, P. Murray-Rust, ChemicalTagger: a tool for semantic text-mining in chemistry, J. Chemother. 3 (1) (2011) 17, https://doi.org/10.1186/1758-2946-3-17.
[41] M. Khabsa, C.L. Giles, Chemical entity extraction using CRF and an ensemble of extractors, J. Chemother. 7 (1) (2015) S12, https://doi.org/10.1186/1758-2946-7-S1-S12.
[42] D.M. Lowe, R.A. Sayle, LeadMine: a grammar and dictionary driven approach to entity recognition, J. Cheminf. 7 (1) (2015) S5, https://doi.org/10.1186/1758-2946-7-S1-S5.
[43] C.J. Court, A. Jain, J.M. Cole, Inverse design of materials that exhibit the magnetocaloric effect by text-mining of the scientific literature and generative deep learning, Chem. Mater. 33 (18) (2021) 7217–7231, https://doi.org/10.1021/acs.chemmater.1c01368.
[44] J. Mavračić, C.J. Court, T. Isazawa, S.R. Elliott, J.M. Cole, ChemDataExtractor 2.0: autopopulated ontologies for materials science, J. Chem. Inf. Model. 61 (9) (2021) 4280–4289, https://doi.org/10.1021/acs.jcim.1c00446.
[45] O. Sierepeklis, J.M. Cole, A thermoelectric materials database auto-generated from the scientific literature using ChemDataExtractor, Sci. Data 9 (1) (2022) 648, https://doi.org/10.1038/s41597-022-01752-1.
[46] P. Kumar, S. Kabra, J.M. Cole, Auto-generating databases of yield strength and grain size using ChemDataExtractor, Sci. Data 9 (1) (2022) 292, https://doi.org/10.1038/s41597-022-01301-w.
[47] Q. Dong, J.M. Cole, Auto-generated database of semiconductor band gaps using ChemDataExtractor, Sci. Data 9 (1) (2022) 193, https://doi.org/10.1038/s41597-022-01294-6.
[48] F. Formalik, K. Shi, F. Joodaki, X. Wang, R.Q. Snurr, Exploring the structural, dynamic, and functional properties of metal-organic frameworks through molecular modeling, Adv. Funct. Mater. (2023) 2308130, https://doi.org/10.1002/adfm.202308130.
[49] Y. Qian, H. Yang, Computational insight into the bioapplication of 2D materials: a review, Nano Today 53 (2023) 102007, https://doi.org/10.1016/j.nantod.2023.102007.
[50] B. Liu, K. Zhou, Recent progress on graphene-analogous 2D nanomaterials: properties, modeling and applications, Prog. Mater. Sci. 100 (2019) 99–169, https://doi.org/10.1016/j.pmatsci.2018.09.004.
[51] R. Xu, X. Zou, B. Liu, H.-M. Cheng, Computational design and property predictions for two-dimensional nanostructures, Mater. Today 21 (4) (2018) 391–418, https://doi.org/10.1016/j.mattod.2018.03.003.
[52] M.A.N. Dewapriya, R.K.N.D. Rajapakse, W.P.S. Dias, Characterizing fracture stress of defective graphene samples using shallow and deep artificial neural networks, Carbon 163 (2020) 425–440, https://doi.org/10.1016/j.carbon.2020.03.038.
[53] J. Chang, S. Zhu, Deep learning on atomistic physical fields of graphene for strain and defect engineering, Adv. Intell. Syst. 6 (4) (2024) 2300601, https://doi.org/10.1002/aisy.202300601.
[54] P. Henderson, A. Ghazaryan, A.A. Zibrov, A.F. Young, M. Serbyn, Deep learning extraction of band structure parameters from density of states: a case study on trilayer graphene, Phys. Rev. B 108 (12) (2023) 125411, https://doi.org/10.1103/PhysRevB.108.125411.
[55] G. Hai, H. Wang, Theoretical studies of metal-organic frameworks: calculation methods and applications in catalysis, gas separation, and energy storage, Coord. Chem. Rev. 469 (2022) 214670, https://doi.org/10.1016/j.ccr.2022.214670.
[56] G. Hu, V. Fung, J. Huang, P. Ganesh, Work function engineering of 2D materials: the role of polar edge reconstructions, J. Phys. Chem. Lett. 12 (9) (2021) 2320–2326, https://doi.org/10.1021/acs.jpclett.1c00278.
[57] J.R. Reimers, A. Sajid, R. Kobayashi, M.J. Ford, Convergence of defect energetics calculations, J. Phys. Chem. C 124 (38) (2020) 21178–21183, https://doi.org/10.1021/acs.jpcc.0c06445.
[58] K.R. Abidi, P. Koskinen, Optimizing density-functional simulations for two-dimensional metals, Phys. Rev. Mater. 6 (12) (2022) 124004, https://doi.org/10.1103/PhysRevMaterials.6.124004.
[59] G. Miceli, J. Hutter, A. Pasquarello, Liquid water through density-functional molecular dynamics: plane-wave vs atomic-orbital basis sets, J. Chem. Theory Comput. 12 (8) (2016) 3456–3462, https://doi.org/10.1021/acs.jctc.6b00271.
[60] F. Tran, J. Doumont, L. Kalantari, P. Blaha, T. Rauch, P. Borlido, S. Botti, M.A. L. Marques, A. Patra, S. Jana, P. Samal, Bandgap of two-dimensional materials: thorough assessment of modern exchange–correlation functionals, J. Chem. Phys. 155 (10) (2021) 104103, https://doi.org/10.1063/5.0059036.
[61] O.V. Gritsenko, Ł.M. Mentel, E.J. Baerends, On the errors of local density (LDA) and generalized gradient (GGA) approximations to the Kohn-Sham potential and orbital energies, J. Chem. Phys. 144 (20) (2016) 204114, https://doi.org/10.1063/1.4950877.
[62] S. Lin, C.-J. Shih, V. Sresht, A. Govind Rajan, M.S. Strano, D. Blankschtein, Understanding the colloidal dispersion stability of 1D and 2D materials: perspectives from molecular simulations and theoretical modeling, Adv. Colloid Interfac. 244 (2017) 36–53, https://doi.org/10.1016/j.cis.2016.07.007.
[63] H. Heinz, T.-J. Lin, R. Kishore Mishra, F.S. Emami, Thermodynamically consistent force fields for the assembly of inorganic, organic, and biological nanostructures: the interface force field, Langmuir 29 (6) (2013) 1754–1765, https://doi.org/10.1021/la3038846.
[64] K. Xu, A.J. Gabourie, A. Hashemi, Z. Fan, N. Wei, A.B. Farimani, H.-P. Komsa, A. V. Krasheninnikov, E. Pop, T. Ala-Nissila, Thermal transport in MoS2 from molecular dynamics using different empirical potentials, Phys. Rev. B 99 (5) (2019) 054303, https://doi.org/10.1103/PhysRevB.99.054303.
[65] S. Rajabpour, Q. Mao, N. Nayir, J.A. Robinson, A.C.T. van Duin, Development and applications of ReaxFF reactive force fields for group-III gas-phase precursors and surface reactions with graphene in metal–organic chemical vapor deposition synthesis, J. Phys. Chem. C 125 (19) (2021) 10747–10758, https://doi.org/10.1021/acs.jpcc.1c01965.
[66] N. Nayir, Y. Wang, S. Shabnam, D.R. Hickey, L. Miao, X. Zhang, S. Bachu, N. Alem, J. Redwing, V.H. Crespi, A.C.T. van Duin, Modeling for structural engineering and synthesis of two-dimensional WSe2 using a newly developed ReaxFF reactive force field, J. Phys. Chem. C 124 (51) (2020) 28285–28297, https://doi.org/10.1021/acs.jpcc.0c09155.
[67] Z. Zhang, Y. Hong, B. Hou, Z. Zhang, M. Negahban, J. Zhang, Accelerated discoveries of mechanical properties of graphene using machine learning and high-throughput computation, Carbon 148 (2019) 115–123, https://doi.org/10.1016/j.carbon.2019.03.046.
[68] A. Al-Maeeni, M. Lazarev, N. Kazeev, K.S. Novoselov, A. Ustyuzhanin, Review on automated 2D material design, 2D Mater. 11 (3) (2024) 032002, https://doi.org/10.1088/2053-1583/ad4661.
[69] S. Curtarolo, W. Setyawan, G.L.W. Hart, M. Jahnatek, R.V. Chepulskii, R. H. Taylor, S. Wang, J. Xue, K. Yang, O. Levy, M.J. Mehl, H.T. Stokes, D. O. Demchenko, D. Morgan, AFLOW: an automatic framework for high-throughput materials discovery, Comput. Mater. Sci. 58 (2012) 218–226, https://doi.org/10.1016/j.commatsci.2012.02.005.
[70] H. Zhao, C.I. Ezeh, W. Ren, W. Li, C.H. Pang, C. Zheng, X. Gao, T. Wu, Integration of machine learning approaches for accelerated discovery of transition-metal dichalcogenides as Hg0 sensing materials, Appl. Energy 254 (2019) 113651, https://doi.org/10.1016/j.apenergy.2019.113651.
[71] D. Wines, R. Gurunathan, K.F. Garrity, B. DeCost, A.J. Biacchi, F. Tavazza, K. Choudhary, Recent progress in the JARVIS infrastructure for next-generation data-driven materials design, Appl. Phys. Rev. 10 (4) (2023), https://doi.org/10.1063/5.0159299.
[72] M. Ragone, R. Shahabazian-Yassar, F. Mashayek, V. Yurkiv, Deep learning modeling in microscopy imaging: a review of materials science applications, Prog. Mater. Sci. 138 (2023) 101165, https://doi.org/10.1016/j.pmatsci.2023.101165.
[73] S. Mahjoubi, F. Ye, Y. Bao, W. Meng, X. Zhang, Identification and classification of exfoliated graphene flakes from microscopy images using a hierarchical deep convolutional neural network, Eng. Appl. Artif. Intell. 119 (2023) 105743, https://doi.org/10.1016/j.engappai.2022.105743.
[74] S. Masubuchi, E. Watanabe, Y. Seo, S. Okazaki, T. Sasagawa, K. Watanabe, T. Taniguchi, T. Machida, Deep-learning-based image segmentation integrated with optical microscopy for automatically searching for two-dimensional materials, npj 2D Mater. Appl. 4 (1) (2020) 3, https://doi.org/10.1038/s41699-020-0137-z.
[75] Y. Saito, K. Shin, K. Terayama, S. Desai, M. Onga, Y. Nakagawa, Y.M. Itahashi, Y. Iwasa, M. Yamada, K. Tsuda, Deep-learning-based quality filtering of mechanically exfoliated 2D crystals, npj Comput. Mater. 5 (1) (2019) 124, https://doi.org/10.1038/s41524-019-0262-4.
[76] X. Dong, H. Li, Z. Jiang, T. Grünleitner, İ. Güler, J. Dong, K. Wang, M.H. Köhler, M. Jakobi, B.H. Menze, A.K. Yetisen, I.D. Sharp, A.V. Stier, J.J. Finley, A.W. Koch, 3D deep learning enables accurate layer mapping of 2D materials, ACS Nano 15 (2) (2021) 3139–3151, https://doi.org/10.1021/acsnano.0c09685.
[77] K. Shevchuk, A. Sarycheva, C.E. Shuck, Y. Gogotsi, Raman spectroscopy characterization of 2D carbide and carbonitride MXenes, Chem. Mater. 35 (19) (2023) 8239–8247, https://doi.org/10.1021/acs.chemmater.3c01742.
[78] X. Zhao, Z. Li, S. Wu, M. Lu, X. Xie, D. Zhan, J. Yan, Raman spectroscopy application in anisotropic 2D materials, Adv. Electron. Mater. 10 (2) (2024) 2300610, https://doi.org/10.1002/aelm.202300610.
[79] Y. Che, D. Wang, H. Lv, X. Wu, Crystal system and space group prediction of two-dimensional materials from chemical formula via deep neural networks, Mater. Today Chem. 33 (2023) 101667, https://doi.org/10.1016/j.mtchem.2023.101667.
[80] B.G. del Rio, C. Kuenneth, H.D. Tran, R. Ramprasad, An efficient deep learning scheme to predict the electronic structure of materials and molecules: the example of graphene-derived allotropes, J. Phys. Chem. A 124 (45) (2020) 9496–9502, https://doi.org/10.1021/acs.jpca.0c07458.
[81] Y. Zhong, L. Zhang, J.-H. Park, S. Cruz, L. Li, L. Guo, J. Kong, E.N. Wang, A unified approach and descriptor for the thermal expansion of two-dimensional transition metal dichalcogenide monolayers, Sci. Adv. 8 (46) (2022) eabo3783, https://doi.org/10.1126/sciadv.abo3783.

[82] B. Zheng, Z. Zheng, G.X. Gu, Designing mechanically tough graphene oxide materials using deep reinforcement learning, npj Comput. Mater. 8 (1) (2022) 225, https://doi.org/10.1038/s41524-022-00919-z.
[83] Y. Ma, S. Lu, Y. Zhang, T. Zhang, Q. Zhou, J. Wang, Accurate energy prediction of large-scale defective two-dimensional materials via deep learning, Appl. Phys. Lett. 120 (21) (2022), https://doi.org/10.1063/5.0091994.
[84] N.C. Frey, D. Akinwande, D. Jariwala, V.B. Shenoy, Machine learning-enabled design of point defects in 2D materials for quantum and neuromorphic information processing, ACS Nano 14 (10) (2020) 13406–13417, https://doi.org/10.1021/acsnano.0c05267.
[85] M. Tohidi Vahdat, K. Varoon Agrawal, G. Pizzi, Machine-learning accelerated identification of exfoliable two-dimensional materials, Mach. Learn.: Sci. Technol. 3 (4) (2022) 045014, https://doi.org/10.1088/2632-2153/ac9bca.
[86] M.N. Gjerding, A. Taghizadeh, A. Rasmussen, S. Ali, F. Bertoldo, T. Deilmann, N. R. Knøsgaard, M. Kruse, A.H. Larsen, S. Manti, T.G. Pedersen, U. Petralanda, T. Skovhus, M.K. Svendsen, J.J. Mortensen, T. Olsen, K.S. Thygesen, Recent progress of the computational 2D materials database (C2DB), 2D Mater. 8 (4) (2021) 044002, https://doi.org/10.1088/2053-1583/ac1059.
[87] A. Chen, Z. Wang, J. Gao, Y. Han, J. Cai, S. Ye, J. Li, A data-driven platform for two-dimensional hybrid lead-halide perovskites, ACS Nano 17 (14) (2023) 13348–13357, https://doi.org/10.1021/acsnano.3c01442.
[88] Y. Lee, S.D. Barthel, P. Dłotko, S.M. Moosavi, K. Hess, B. Smit, Quantifying similarity of pore-geometry in nanoporous materials, Nat. Commun. 8 (1) (2017) 15396, https://doi.org/10.1038/ncomms15396.
[89] B. Xu, Z. Gong, J. Liu, Y. Hong, Y. Yang, L. Li, Y. Liu, J. Deng, J.Z. Liu, Tunable ferroelectric topological defects on 2D topological surfaces: complex strain engineering skyrmion-like polar structures in 2D materials, Adv. Funct. Mater. 34 (26) (2024) 2311599, https://doi.org/10.1002/adfm.202311599.
[90] T. Yang, J. Zhou, T.T. Song, L. Shen, Y.P. Feng, M. Yang, High-throughput identification of exfoliable two-dimensional materials with active basal planes for hydrogen evolution, ACS Energy Lett. 5 (7) (2020) 2313–2321, https://doi.org/10.1021/acsenergylett.0c00957.
[91] Y. Shen, S. Zhu, Machine learning mechanical properties of defect-engineered hexagonal boron nitride, Comput. Mater. Sci. 220 (2023) 112030, https://doi.org/10.1016/j.commatsci.2023.112030.
[92] M. Kuban, S. Rigamonti, M. Scheidgen, C. Draxl, Density-of-states similarity descriptor for unsupervised learning from materials data, Sci. Data 9 (1) (2022) 646, https://doi.org/10.1038/s41597-022-01754-z.
[93] C. Ben Mahmoud, A. Anelli, G. Csányi, M. Ceriotti, Learning the electronic density of states in condensed matter, Phys. Rev. B 102 (23) (2020) 235130, https://doi.org/10.1103/PhysRevB.102.235130.
[94] Y. Dong, C. Wu, C. Zhang, Y. Liu, J. Cheng, J. Lin, Bandgap prediction by deep learning in configurationally hybridized graphene and boron nitride, npj Comput. Mater. 5 (1) (2019) 26, https://doi.org/10.1038/s41524-019-0165-4.
[95] Z. Zheng, T. Xu, D. Legut, R. Zhang, High-throughput informed machine learning models for ultrastrong B-N solids, Comput. Mater. Sci. 215 (2022) 111789, https://doi.org/10.1016/j.commatsci.2022.111789.
[96] M.O.J. Jäger, E.V. Morooka, F. Federici Canova, L. Himanen, A.S. Foster, Machine learning hydrogen adsorption on nanoclusters through structural descriptors, npj Comput. Mater. 4 (1) (2018) 37, https://doi.org/10.1038/s41524-018-0096-5.
[97] T. Felser, S. Notarnicola, S. Montangero, Efficient tensor network Ansatz for high-dimensional quantum many-body problems, Phys. Rev. Lett. 126 (17) (2021) 170603, https://doi.org/10.1103/PhysRevLett.126.170603.
[98] A. Raja, A. Chaves, J. Yu, G. Arefe, H.M. Hill, A.F. Rigosi, T.C. Berkelbach, P. Nagler, C. Schüller, T. Korn, C. Nuckolls, J. Hone, L.E. Brus, T.F. Heinz, D. R. Reichman, A. Chernikov, Coulomb engineering of the bandgap and excitons in two-dimensional materials, Nat. Commun. 8 (1) (2017) 15251, https://doi.org/10.1038/ncomms15251.
[99] N.E.R. Zimmermann, A. Jain, Local structure order parameters and site fingerprints for quantification of coordination environment and crystal structure similarity, RSC Adv. 10 (10) (2020) 6063–6081, https://doi.org/10.1039/C9RA07755C.
[100] H. Jintoku, D.N. Futaba, Machine learning-assisted exploration and identification of aqueous dispersants in the vast diversity of organic chemicals, ACS Appl. Mater. Interfaces 16 (9) (2024) 11800–11808, https://doi.org/10.1021/acsami.3c18612.
[101] M. Krykunov, T.K. Woo, Bond type restricted property weighted radial distribution functions for accurate machine learning prediction of atomization energies, J. Chem. Theory Comput. 14 (10) (2018) 5229–5237, https://doi.org/10.1021/acs.jctc.8b00788.
[102] K. Hansen, F. Biegler, R. Ramakrishnan, W. Pronobis, O.A. von Lilienfeld, K.-R. Müller, A. Tkatchenko, Machine learning predictions of molecular properties: accurate many-body potentials and nonlocality in chemical space, J. Phys. Chem. Lett. 6 (12) (2015) 2326–2331, https://doi.org/10.1021/acs.jpclett.5b00831.
[103] S. Lu, Q. Zhou, Y. Guo, Y. Zhang, Y. Wu, J. Wang, Coupling a crystal graph multilayer descriptor to active learning for rapid discovery of 2D ferromagnetic semiconductors/half-metals/metals, Adv. Mater. 32 (29) (2020) 2002658, https://doi.org/10.1002/adma.202002658.
[104] O. Isayev, C. Oses, C. Toher, E. Gossett, S. Curtarolo, A. Tropsha, Universal fragment descriptors for predicting properties of inorganic crystals, Nat. Commun. 8 (1) (2017) 15679, https://doi.org/10.1038/ncomms15679.
[105] T. Xie, J.C. Grossman, Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties, Phys. Rev. Lett. 120 (14) (2018) 145301, https://doi.org/10.1103/PhysRevLett.120.145301.
[106] F. Strieth-Kalthoff, F. Sandfort, M.H.S. Segler, F. Glorius, Machine learning the ropes: principles, applications and directions in synthetic chemistry, Chem. Soc. Rev. 49 (17) (2020) 6154–6168, https://doi.org/10.1039/C9CS00786E.
[107] Q. Liu, Y. Gao, B. Xu, Transferable, deep-learning-driven fast prediction and design of thermal transport in mechanically stretched graphene flakes, ACS Nano 15 (10) (2021) 16597–16606, https://doi.org/10.1021/acsnano.1c06340.
[108] Y. Lin, J. Ma, Q. Wang, D.-W. Sun, Applications of machine learning techniques for enhancing nondestructive food quality and safety detection, Crit. Rev. Food Sci. 63 (12) (2023) 1649–1669, https://doi.org/10.1080/10408398.2022.2131725.
[109] T. Xie, X. Fu, O.-E. Ganea, R. Barzilay, T. Jaakkola, Crystal diffusion variational autoencoder for periodic material generation, ArXiv. abs/2110.06197, 2021, https://doi.org/10.48550/arXiv.2110.06197.
[110] C. Chen, J. Zheng, C. Chu, Q. Xiao, C. He, X. Fu, An effective method for generating crystal structures based on the variational autoencoder and the diffusion model, Chinese Chem. Lett. 241 (2024) 109739, https://doi.org/10.1016/j.cclet.2024.109739.
[111] I. Higgins, L. Matthey, A. Pal, C.P. Burgess, X. Glorot, M.M. Botvinick, S. Mohamed, A.J.I. Lerchner, Beta-VAE: learning basic visual concepts with a constrained variational framework, in: 5th International Conference on Learning Representations, 2017.
[112] L. Mi, M. Shen, J.J.A. Zhang, A probe towards understanding GAN and VAE models, Arxiv Preprint, 2018, https://doi.org/10.48550/arXiv.1812.05676. Arxiv:1812.05676.
[113] B. Dai, D. Wipf, Diagnosing and enhancing VAE models, in: 7th International Conference on Learning Representations, 2019, https://doi.org/10.48550/arXiv.1903.05789. ArXiv:1903.05789.
[114] S. Molnár, L. Tamás, Variational autoencoders for 3D data processing, Artif. Intell. Rev. 57 (2) (2024) 42, https://doi.org/10.1007/s10462-023-10687-x.
[115] A. Krizhevsky, I. Sutskever, G.E. Hinton, ImageNet classification with deep convolutional neural networks, ACM Comput. Surv. 60 (6) (2017) 84–90, https://doi.org/10.1145/3065386.
[116] D.-H. Yang, Y.-S. Chu, O.F.N. Okello, S.-Y. Seo, G. Moon, K.H. Kim, M.-H. Jo, D. Shin, T. Mizoguchi, S. Yang, S.-Y. Choi, Full automation of point defect detection in transition metal dichalcogenides through a dual mode deep learning algorithm, Mater. Horiz. 11 (3) (2024) 747–757, https://doi.org/10.1039/D3MH01500A.
[117] C. Allen, S. Aryal, T. Do, R. Gautum, M.M. Hasan, B.K. Jasthi, E. Gnimpieba, V. Gadhamshetty, Deep learning strategies for addressing issues with small datasets in 2D materials research: microbial corrosion, Front. Microbiol. 13 (2022) 1059123, https://doi.org/10.3389/fmicb.2022.1059123.
[118] S. Islam, H. Elmekki, A. Elsebai, J. Bentahar, N. Drawel, G. Rjoub, W. Pedrycz, A comprehensive survey on applications of transformers for deep learning tasks, Expert Syst. Appl. 241 (2024) 122666, https://doi.org/10.1016/j.eswa.2023.122666.
[119] A. Sherstinsky, Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network, Phys. D 404 (2020) 132306, https://doi.org/10.1016/j.physd.2019.132306.
[120] P. Ma, S. Tsai, Y. He, X. Jia, D. Zhen, N. Yu, Q. Wang, J.K.C. Ahuja, C.-I. Wei, Large language models in food science: innovations, applications, and future, Trends Food Sci. Technol. 148 (2024) 104488, https://doi.org/10.1016/j.tifs.2024.104488.
[121] S. Chen, Y. Zhang, Q. Yang, Multi-task learning in natural language processing: an overview, ACM Comput. Surv. 56 (12) (2024) 295, https://doi.org/10.1145/3663363.
[122] N. Fu, L. Wei, Y. Song, Q. Li, R. Xin, S.S. Omee, R. Dong, E.M.D. Siriwardane, J. Hu, Material transformers: deep learning language models for generative materials design, Mach. Learn.: Sci. Technol. 4 (1) (2023) 015001, https://doi.org/10.1088/2632-2153/acadcd.
[123] M. Kusaba, C. Liu, R. Yoshida, Crystal structure prediction with machine learning-based element substitution, Comput. Mater. Sci. 211 (2022) 111496, https://doi.org/10.1016/j.commatsci.2022.111496.
[124] L. Wei, N. Fu, E.M.D. Siriwardane, W. Yang, S.S. Omee, R. Dong, R. Xin, J. Hu, TCSP: a template-based crystal structure prediction algorithm for materials discovery, Inorg. Chem. 61 (22) (2022) 8431–8439, https://doi.org/10.1021/acs.inorgchem.1c03879.
[125] S. Kim, J. Noh, T. Jin, J. Lee, Y. Jung, A structure translation model for crystal compounds, npj Comput. Mater. 9 (1) (2023) 142, https://doi.org/10.1038/s41524-023-01094-5.
[126] I.A. Kruglov, A.V. Yanilkin, Y. Propad, A.B. Mazitov, P. Rachitskii, A.R. Oganov, Crystal structure prediction at finite temperatures, npj Comput. Mater. 9 (1) (2023) 197, https://doi.org/10.1038/s41524-023-01120-6.
[127] J. Hu, W. Yang, R. Dong, Y. Li, X. Li, S. Li, E.M.D. Siriwardane, Contact map based crystal structure prediction using global optimization, CrystEngComm 23 (8) (2021) 1765–1776, https://doi.org/10.1039/D0CE01714K.
[128] Z. Zheng, O. Zhang, C. Borgs, J.T. Chayes, O.M. Yaghi, ChatGPT chemistry assistant for text mining and the prediction of MOF synthesis, J. Am. Chem. Soc. 145 (32) (2023) 18048–18062, https://doi.org/10.1021/jacs.3c05819.
[129] S. Abadal, A. Jain, R. Guirado, J. López-Alonso, E. Alarcón, Computing graph neural networks: a survey from algorithms to accelerators, ACM Comput. Surv. 54 (9) (2021) 1–38, https://doi.org/10.1145/3477141.
[130] N. Kazeev, A.R. Al-Maeeni, I. Romanov, M. Faleev, R. Lukin, A. Tormasov, A. H. Castro Neto, K.S. Novoselov, P. Huang, A. Ustyuzhanin, Sparse representation for machine learning the properties of defects in 2D materials, npj Comput. Mater. 9 (1) (2023) 113, https://doi.org/10.1038/s41524-023-01062-z.

[131] G. Corso, H. Stark, S. Jegelka, T. Jaakkola, R. Barzilay, Graph neural networks, Nat. Rev. Methods Primers 4 (1) (2024) 17, https://doi.org/10.1038/s43586-024-00294-7.
[132] C. Chen, W. Ye, Y. Zuo, C. Zheng, S.P. Ong, Graph networks as a universal machine learning framework for molecules and crystals, Chem. Mater. 31 (9) (2019) 3564–3572, https://doi.org/10.1021/acs.chemmater.9b01294.
[133] Z. Qiao, M. Welborn, A. Anandkumar, F.R. Manby, T.F. Miller III, OrbNet: Deep learning for quantum chemistry using symmetry-adapted atomic-orbital features, J. Chem. Phys. 153 (12) (2020), https://doi.org/10.1063/5.0021955.
[134] K. Choudhary, B. DeCost, Atomistic line graph neural network for improved materials property predictions, npj Comput. Mater. 7 (1) (2021) 185, https://doi.org/10.1038/s41524-021-00650-1.
[135] K.T. Schütt, H.E. Sauceda, P.-J. Kindermans, A. Tkatchenko, K.-R. Müller, SchNet – a deep learning architecture for molecules and materials, J. Chem. Phys. 148 (24) (2018) 241722, https://doi.org/10.1063/1.5019779.
[136] C. Chen, S.P. Ong, A universal graph deep learning interatomic potential for the periodic table, Nat. Comput. Sci. 2 (11) (2022) 718–728, https://doi.org/10.1038/s43588-022-00349-3.
[137] Y. Song, E.M.D. Siriwardane, Y. Zhao, J. Hu, Computational discovery of new 2D materials using deep learning generative models, ACS Appl. Mater. Interfaces 13 (45) (2021) 53303–53313, https://doi.org/10.1021/acsami.1c01044.
[138] D. Saxena, J. Cao, Generative adversarial networks (GANs): challenges, solutions, and future directions, ACM Comput. Surv. 54 (3) (2021) 63, https://doi.org/10.1145/3446374.
[139] Y. Dan, Y. Zhao, X. Li, S. Li, M. Hu, J. Hu, Generative adversarial networks (GAN) based efficient sampling of chemical composition space for inverse design of inorganic materials, npj Comput. Mater. 6 (1) (2020) 84, https://doi.org/10.1038/s41524-020-00352-0.
[140] R. Xin, E.M.D. Siriwardane, Y. Song, Y. Zhao, S.-Y. Louis, A. Nasiri, J. Hu, Active-learning-based generative design for the discovery of wide-band-gap materials, J. Phys. Chem. C 125 (29) (2021) 16118–16128, https://doi.org/10.1021/acs.jpcc.1c02438.
[141] L. Zhu, J. Tang, B. Li, T. Hou, Y. Zhu, J. Zhou, Z. Wang, X. Zhu, Z. Yao, X. Cui, K. Watanabe, T. Taniguchi, Y. Li, Z.V. Han, W. Zhou, Y. Huang, Z. Liu, J.C. Hone, Y. Hao, Artificial neuron networks enabled identification and characterizations of 2D materials and van der Waals heterostructures, ACS Nano 16 (2) (2022) 2721–2729, https://doi.org/10.1021/acsnano.1c09644.
[142] X. Liu, M.C. Hersam, Interface characterization and control of 2D materials and heterostructures, Adv. Mater. 30 (39) (2018) 1801586, https://doi.org/10.1002/adma.201801586.
[143] S.-H. Yang, W. Choi, B.W. Cho, F.O.-T. Agyapong-Fordjour, S. Park, S.J. Yun, H.-J. Kim, Y.-K. Han, Y.H. Lee, K.K. Kim, Y.-M. Kim, Deep learning-assisted quantification of atomic dopants and defects in 2D materials, Adv. Sci. 8 (16) (2021) 2101099, https://doi.org/10.1002/advs.202101099.
[144] P. Li, Z. Kang, Z. Zhang, Q. Liao, F. Rao, Y. Lu, Y. Zhang, In situ microscopy techniques for characterizing the mechanical properties and deformation behavior of two-dimensional (2D) materials, Mater. Today 51 (2021) 247–272, https://doi.org/10.1016/j.mattod.2021.10.009.
[145] J. Kim, Y. Lee, M. Kang, L. Hu, S. Zhao, J.-H. Ahn, 2D materials for skin-mountable electronic devices, Adv. Mater. 33 (47) (2021) 2005858, https://doi.org/10.1002/adma.202005858.
[146] B. Wang, Y. Sun, H. Ding, X. Zhao, L. Zhang, J. Bai, K. Liu, Bioelectronics-related 2D materials beyond graphene: fundamentals, properties, and applications, Adv. Funct. Mater. 30 (46) (2020) 2003732, https://doi.org/10.1002/adfm.202003732.
[147] X. Chen, Z. Zhou, B. Deng, Z. Wu, F. Xia, Y. Cao, L. Zhang, W. Huang, N. Wang, L. Wang, Electrically tunable physical properties of two-dimensional materials, Nano Today 27 (2019) 99–119, https://doi.org/10.1016/j.nantod.2019.05.005.
[148] B. Lowe, B. Field, J. Hellerstedt, J. Ceddia, H.L. Nourse, B.J. Powell, N. V. Medhekar, A. Schiffrin, Local gate control of Mott metal-insulator transition in a 2D metal-organic framework, Nat. Commun. 15 (1) (2024) 3559, https://doi.org/10.1038/s41467-024-47766-8.
[149] T. Dutta, N. Yadav, Y. Wu, G.J. Cheng, X. Liang, S. Ramakrishna, A. Sbai, R. Gupta, A. Mondal, Z. Hongyu, A. Yadav, Electronic properties of 2D materials and their junctions, Nano Mater. Sci. 6 (1) (2024) 1–23, https://doi.org/10.1016/j.nanoms.2023.05.003.
[150] N.R. Knøsgaard, K.S. Thygesen, Representing individual electronic states for machine learning GW band structures of 2D materials, Nat. Commun. 13 (1) (2022) 468, https://doi.org/10.1038/s41467-022-28122-0.
[151] F. Hüser, T. Olsen, K.S. Thygesen, Quasiparticle GW calculations for solids, molecules, and two-dimensional materials, Phys. Rev. B 87 (23) (2013) 235132, https://doi.org/10.1103/PhysRevB.87.235132.
[152] A.C. Rajan, A. Mishra, S. Satsangi, R. Vaish, H. Mizuseki, K.-R. Lee, A.K. Singh, Machine-learning-assisted accurate band gap predictions of functionalized MXene, Chem. Mater. 30 (12) (2018) 4031–4038, https://doi.org/10.1021/acs.chemmater.8b00686.
[153] J. Liang, X. Zhu, Phillips-inspired machine learning for band gap and exciton binding energy prediction, J. Phys. Chem. Lett. 10 (18) (2019) 5640–5646, https://doi.org/10.1021/acs.jpclett.9b02232.
[154] A.A. Zibrov, P. Rao, C. Kometter, E.M. Spanton, J.I.A. Li, C.R. Dean, T. Taniguchi, K. Watanabe, M. Serbyn, A.F. Young, Emergent dirac gullies and gully-symmetry-breaking quantum hall states in ABA trilayer graphene, Phys. Rev. Lett. 121 (16) (2018) 167601, https://doi.org/10.1103/PhysRevLett.121.167601.
[155] M. Dong, Y. Sun, D.J. Dunstan, R.J. Young, D.G. Papageorgiou, Mechanical reinforcement from two-dimensional nanofillers: model, bulk and hybrid polymer nanocomposites, Nanoscale 16 (28) (2024) 13247–13299, https://doi.org/10.1039/D4NR01356E.
[156] G. Wang, H. Hou, Y. Yan, R. Jagatramka, A. Shirsalimian, Y. Wang, B. Li, M. Daly, C. Cao, Recent advances in the mechanics of 2D materials, Int. J. Extrem. Manuf. 5 (3) (2023) 032002, https://doi.org/10.1088/2631-7990/accda2.
[157] P.Z. Hanakata, E.D. Cubuk, D.K. Campbell, H.S. Park, Accelerated search and design of stretchable graphene Kirigami using machine learning, Phys. Rev. Lett. 121 (25) (2018) 255304, https://doi.org/10.1103/PhysRevLett.121.255304.
[158] J. Wan, J.-W. Jiang, H.S. Park, Machine learning-based design of porous graphene with low thermal conductivity, Carbon 157 (2020) 262–269, https://doi.org/10.1016/j.carbon.2019.10.037.
[159] J. Lee, D. Park, M. Lee, H. Lee, K. Park, I. Lee, S. Ryu, Machine learning-based inverse design methods considering data characteristics and design space size in materials design and manufacturing: a review, Mater. Horiz. 10 (12) (2023) 5436–5456, https://doi.org/10.1039/D3MH00039G.
[160] A. Zunger, Inverse design in search of materials with target functionalities, Nat. Rev. Chem. 2 (4) (2018) 0121, https://doi.org/10.1038/s41570-018-0121.
[161] J. Zhang, A. Koneru, S.K.R.S. Sankaranarayanan, C.M. Lilley, Graph neural network guided evolutionary search of grain boundaries in 2D materials, ACS Appl. Mater. Interfaces 15 (16) (2023) 20520–20530, https://doi.org/10.1021/acsami.3c01161.
[162] S. Pakdel, A. Rasmussen, A. Taghizadeh, M. Kruse, T. Olsen, K.S. Thygesen, High-throughput computational stacking reveals emergent properties in natural van der Waals bilayers, Nat. Commun. 15 (1) (2024) 932, https://doi.org/10.1038/s41467-024-45003-w.
[163] R.K. Barik, L.M. Woods, High throughput calculations for a dataset of bilayer materials, Sci. Data 10 (1) (2023) 232, https://doi.org/10.1038/s41597-023-02146-7.
[164] H.T. Kollmann, D.W. Abueidda, S. Koric, E. Guleryuz, N.A. Sobh, Deep learning for topology optimization of 2D metamaterials, Mater. Des. 196 (2020) 109098, https://doi.org/10.1016/j.matdes.2020.109098.
[165] J. Cheng, T. Li, Y. Wang, A.H. Ati, Q. Sun, High-throughput screening of MXenes for hydrogen storage via graph neural network, Appl. Surf. Sci. 641 (2023) 158560, https://doi.org/10.1016/j.apsusc.2023.158560.