Review | Received 28 August 2025; Revised 25 November 2025; Accepted 14 January 2026; Published 24 March 2026
https://doi.org/10.55092/aimat20260003

Advanced artificial intelligence algorithms and hardware acceleration techniques applied to material structure design

Jiqun Zhang[1], Juhong Yu[2,*], Nianxiang Qiu[3], Yong Liu[4], Yangyang Song[5], Liang Zhang[5] and Shiyu Du[1,2,6,*]

[1] Qingdao Institute of Software College of Computer Science and Technology, China University of Petroleum (East China), Qingdao 266580, China
[2] College of Materials Science and Engineering, China University of Petroleum (East China), Qingdao 266580, China
[3] Yangtze Delta Region Institute (Huzhou), University of Electronic Science and Technology, Huzhou 313001, China
[4] Department of Chemistry, University of Colorado Denver, Denver, Colorado 80217-3364, USA
[5] Putuo People's Hospital, Frontier Science Center for Stem Cell Research, School of Life Sciences and Technology, Tongji University, Shanghai 200092, China
[6] Milky-Way Sustainable Energy Ltd., Zhuhai 519000, China

* Corresponding authors; E-mails: 20230031@upc.edu.cn (J.Y.); dushiyu@nimte.ac.cn (S.D.).

Highlights:

- Deep learning accurately predicts and optimizes material properties and structures.
- Generative models accelerate novel material discovery.
- Natural language processing automates knowledge extraction from materials literature.
- AI-driven structural optimization enhances material performance.
- Hardware acceleration enhances efficiency for AI-driven materials science.

Abstract: The ability of AI-based algorithms to reflect the physical properties of training data and accurately predict the properties of undeveloped materials has made artificial intelligence (AI) algorithms an important tool in the domain of materials science. Material structure design, as a multi-step and multi-scientific task, involves many aspects from the determination of design objectives to material selection, preparation methods, sample testing and performance evaluation, *etc.* Traditional experiment-driven, theory-driven and algorithm-driven approaches have accumulated a large amount of textual material text data. With the development of technology, the data-driven approach of "Big Data+AI" has accelerated the development of material structure design, and in particular, deep learning (DL) as a branch of machine learning has become the fastest growing topic in materials science because of its powerful capabilities to analyze unstructured data and automatically identify features. In this paper, we focus on the common deep learning methods in materials research, and then review the material property prediction, material structure optimization, material discovery and information extraction from materials

literature in the context of deep learning-based material structure design. At the same time, we introduce the hardware acceleration technologies based on deep learning. Finally, the summary and future directions of deep learning in future materials research are discussed.

Keywords: artificial intelligence; hardware acceleration; material structure design; machine learning; deep learning; data-driven

1. Introduction

In the past two decades, the research on materials has been rapidly developed, experimental analysis methods and theory-based algorithm methods have been widely used in material design. However, traditional experimental analysis methods such as trial-and-error methods [1] have problems such as tedious experimental process, long development cycle and insufficient measurement accuracy. Traditional theory-driven approaches, such as the first-principles theoretical calculation [2], form the foundation of computational materials science. However, their application in exploring vast material spaces typically necessitates a large number of discrete computational tests, which incurs prohibitive time and resource costs. With the emergence of large-scale material systems and the demand for materials with higher performance, these approaches face limitations in solving complex problems like property prediction. They struggle to fully account for the simultaneous influence of multiple factors, making it difficult to excavate the internal relationship between material characteristics and properties, which hinders the actual development of material structures [3]. The field has progressed by leveraging a suite of computational simulation methods, with theory-driven approaches as its core, which include finite element analysis (FEA) [4], density functional theory [5], molecular dynamics (MD) [6] and phase field method [7], *etc.* Based on theoretical calculation and simulation, promising candidate materials can be predicted, the experimental scope can be narrowed, and finally verified by experiment. It can greatly improve the research efficiency of materials science, but there is still a problem that the computational cost is too high when facing larger and more complex systems and space and time scales, and the computational simulation method cannot meet the requirements of quantitative characterization of material properties. Therefore, it is necessary to develop new methods to guide the design and development of materials.

Materials research and development has accumulated a large amount of data after three stages of experiment-driven, theory-driven and algorithm-driven approaches in the early stage. Data-driven, powered by machine learning techniques and large-scale datasets, marked by “big data+AI”, has become the fourth paradigm for the development of materials science [8]. The proposal of the fourth paradigm has opened a new data-driven research model in the field of materials science. Machine learning (ML) based on material database extracts hidden variables between data, builds a model of specific material properties, guides the design of material structures, and reduces the research and development cycle of materials. With the progress of computing power, deep learning provides novel insights and tools for the production and processing of massive data, and has become a cutting-edge direction and hot spot in materials research.

Based on this, this paper reviews the application of deep learning in material structure design and hardware acceleration technology based on deep learning. It has been shown that deep learning outperforms traditional computational methods in studying materials with multi-dimensional geometric

structure, complex internal structure and multi-type and multi-scale defects [9]. Based on this, the proposed applications in this paper are mainly based on the dielectric loss of polycrystalline materials over a wide frequency and temperature range, the mechanical properties of composite materials, the service life of alloy materials, the large-scale screening and optimization of porous materials, and the thermoelectric properties of inorganic materials.

The chapters of this paper are arranged as follows: The second section introduces the basic concepts of machine learning. Section 3 focuses on deep learning architectures applied to materials, including convolutional neural networks, graph neural networks, recurrent neural networks, Transformer, and generative models. The fourth section introduces the application of deep learning in material structure design, including (1) material performance prediction, which aims to provide the basis for material structure design. At the same time, the performance of the designed structure in practical application can be tested and evaluated and fed back into the performance prediction model to further optimize the prediction algorithm and improve the accuracy of structural design; (2) Material structure optimization, which aims to improve the material structure to meet specific performance requirements. In the design process, the optimization algorithm can help to find the best material structure, and the designed structure is improved through the iterative optimization process, and the optimization results guide the new structure design. (3) Material discovery, which aims to reveal new material structures, discover new material properties, and promote the development of the field of materials science; (4) Information extraction in material texts aims to extract and integrate knowledge related to material structure from a large number of literatures and reports, provide data support for material structure design, and assist in verifying the rationality of material structure design. The fifth section introduces the hardware acceleration technology based on deep learning. Section 6 discusses the summary and future directions of deep learning in future materials research. Commonly used artificial intelligence methods in the field of materials science, as well as the relationships between various learning methods, are illustrated in Figure 1.

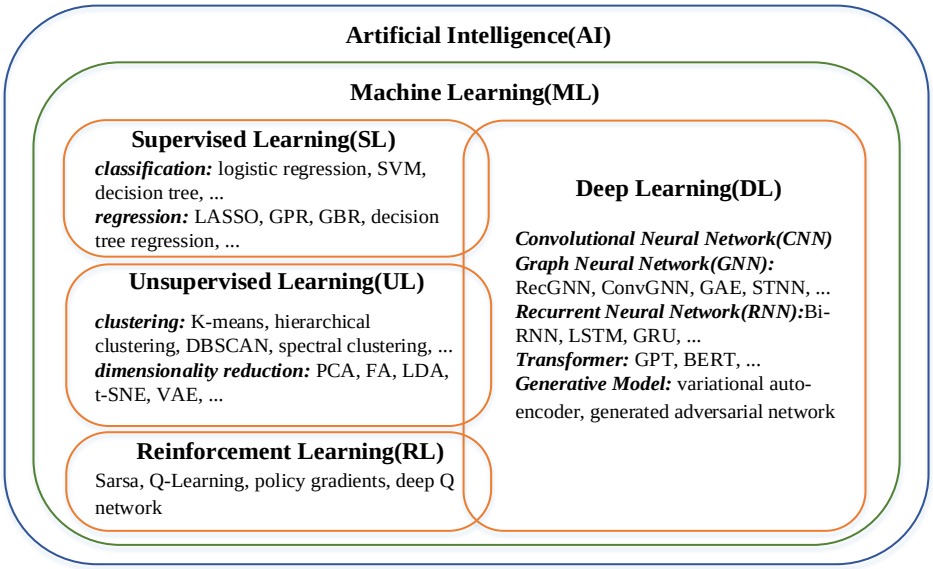

Figure 1. The commonly used artificial intelligence methods in materials science and the relationships between AI, ML and DL, as well as among various learning approaches.

2. Machine learning

Machine learning has become a pivotal tool in materials science, enabling the discovery of complex patterns within high-dimensional data that are often intractable for traditional theoretical or experimental methods. The application of ML in materials science typically falls into three paradigms, each suited to different data types and design objectives.

2.1. Supervised learning

Supervised learning (SL) is extensively used for establishing mappings between material descriptors and target properties. Based on the type of target variable, it can be divided into classification and regression. The classification algorithm aims to predict the class of a samples (e.g., whether fission occurs), and the target variable is discrete. Typical classification algorithms include logistic regression, support vector machine (SVM) and decision tree.

The regression task predicts the real output corresponding to the sample (e.g., temperature, pressure), and its target variable is continuous. The regression algorithm includes least absolute shrinkage and selection operator (LASSO) [10], Gaussian process regression (GPR) [11], gradient boosting regression (GBR) [12] and decision tree regression [13]. For example, Chandran *et al.* [14] adopted LASSO algorithm to predicted the energies of different configurations; Khatavkar *et al.* [15] adopted GPR algorithm to establish the relationship between vickers hardness, microstructure and composition parameters.

2.2. Unsupervised learning

Unsupervised learning (UL) extracts information and finds hidden patterns from unlabeled data. A common approach is clustering, which groups data points with similar characteristics. UL is valuable for extracting knowledge from unlabeled data, which is abundant in materials science. It is primarily used for clustering and dimensionality reduction.

Clustering algorithms such as K-means [16], hierarchical clustering [17], DBSCAN [18], spectral clustering [19], which group similar material structures or compositions, aiding in the identification of new material phases. Dimensionality reduction methods such as principal component analysis (PCA) [20], factor analysis (FA) [21], linear discriminant analysis (LDA) [22], t-distributed stochastic neighbor embedding (t-SNE) [23], variational auto-encoder (VAE) [24], which help visualize high-dimensional material data in lower-dimensional spaces, revealing underlying patterns and correlations in complex datasets like phase mappings or spectroscopic data.

2.3. Reinforcement learning

Reinforcement learning (RL) refers to a class of problems that constantly learn from interactions and methods to solve such problems.

The objects of interaction include agent and environment. Agents are used to perceive external states and feedback reward, and make learning and decisions. The environment is all the things outside the agent, and is affected by the agent's actions to change its state, and feedback to the agent the corresponding reward. In materials science, RL has shown promise in guiding multi-step processes such

as experimental synthesis and structural optimization. Common reinforcement learning methods include sarsa [25], Q-Learning [26], policy gradients [27], deep Q network [28].

3. Brief introduction to deep learning

3.1. Convolutional neural network

Convolutional neural network (CNN) [29] is a deep learning model that processes data with a grid-like structure. For example, such as material microstructure images from microscopy (SEM, TEM) or voxelized 3D atomic structures. CNN is mainly composed of convolutional layer, pooling layer and fully connected layer. The convolutional layer consists of multiple feature maps. Each feature map consists of multiple neurons. Different input features can be extracted by depthwise convolution, group convolution, transposed convolution and other convolution operations. The features are down sampled through maximum pooling, mean pooling, and random pooling, and the resolution of Feature Map is reduced to obtain spatially invariant features. In materials science, the standard convolutional, pooling, and fully-connected layers are often adapted to capture physically relevant features. Convolutional neural networks perform well in many applications. In material structure design, CNN can be used to identify and classify the microstructure of materials, predict the properties of materials, and enable reverse engineering of materials.

3.2. Graph neural network

The CNNs mentioned above are often used to extract features from Euclidean data, and the core assumption of such deep learning algorithms is that the data samples are independent of each other. However, in real life data, such as social networks, image fragments, word vectors, recommendation systems, and atomic/molecular structures, are often non-Euclidean, so graph-based data structures become particularly important.

A graph can be represented by $G = (V, E)$, where V is a set of vertices or nodes and E is a set of edges, *i.e.* if there are nodes in the graph $v_i, v_j \in V$, $e_{ij} = (v_i, v_j) \in E$ represents an edge from v_i to v_j. $N(v) = \{V|(v, u) \in E\}$ indicates the neighborhood of node v. For a graph, each data sample in the graph will have edges related to other real data samples in the graph, and this information can be used to capture the interdependencies between the samples. The adjacency matrix A is an n × n matrix, if $e_{ij} \in E$ then $A_{ij} = 1$; If $e_{ij} \notin E$, then $A_{ij} = 0$. The node attribute X^l of the graph, where $X \in R^{n \times d}$ is a node eigenmatrix and $x_v \in R^d$ represents the eigenvector of node v; The edge property X^e of the graph, where $X^e \in R^{m \times c}$ is the edge eigenmatrix and $x^e_{v,u} \in R^c$ represents the eigenvector of the edge (v, u).

Existing studies have divided graph neural network (GNN) [30] into four categories: recurrent graph neural network (RecGNN) [31], convolutional graph network (ConvGNN) [32], graph auto-encoder (GAE) [33] and spatial-temporal neural network (STNN) [34]. GNN inputs graph structure and node information and determines the output according to different tasks. RecGNN and ConvGNN can extract high-level node representations via information propagation or graph convolution. Using the multilayer perceptron (MLP) or softmax layer as the output layer, GNN is able to perform classification or regression tasks in an end-to-end method; Using the hidden representation of two nodes, similarity functions or neural networks can be utilized for edge classification or prediction tasks; By combining GNN with pooling

and readout operations, graph representations can be obtained and graph classification tasks can be realized. At present, GNN has been used to predict the properties of various materials based on structural information. The value and challenges of GNNs in materials science coexist. Their strength lies in the native ability to process atom-bond graph structures, directly capturing key interactions to correlate with material properties. The challenges, however, center on adapting models to satisfy three core domain-specific characteristics: crystal periodicity, physical symmetries, and complex many-body interactions. This has shifted the research focus toward deeply embedding such domain knowledge into model architectures, driving the development of more accurate and reliable next-generation methods.

3.3. Recurrent neural network

Recurrent neural network (RNN) [35] is mainly used for tasks that require processing sequence information, such as time series prediction, task-based conversations, *etc.*, which can take into account the order and context information about each word and produce an output of arbitrary length. Because the hidden layer of the RNN is cyclic, that is, the output value of the next moment of the RNN is affected by the input value of the previous moment, the nodes between the hidden layers are connected. This structure allows RNNs to store, remember, and process complex signals from the long past. RNNs can map the input sequence to the output sequence at the current time step and predict the output at the next time step.

The traditional RNN model has the problem of gradient exploding and gradient vanishing, which leads to the accumulation of errors and the difficulty of gradient transmission in long sequence, and cannot save remote context information. Generally, a threshold is set to intercept the gradient exceeding the threshold to solve the gradient explosion problem, and RNNs with other structures are used to solve the problem of gradient disappearance, such as bidirectional recurrent neural network (Bi-RNN) [36], which calculates from the forward and reverse directions, and simultaneously learns the information of the previous moment and the following moment. Long short-term memory networks (LTSM) use gated units and memory mechanisms to alleviate long-term dependence. Further, the gated recurrent unit (GRU) is improved on the basis of LSTM. The GRU uses the update gate to control how much information can be brought into the current state from the previous state, and uses the reset gate to control how much information can be written into the current state from the previous state. At present, LSTM has been successfully applied to various named entity recognition tasks in the field of materials.

3.4. Transformer

Transformer [37] is an encoder-decoder architecture model based on self-attention mechanism, which mainly includes attention layer, feedforward network (FFN), residual connection and LayerNorm, positional encoding.

Inspired by Transformer, some researchers have focused on sequencer-to-sequence models based on Transformer architecture. The pre-trained language model generative pre-trained transformer (GPT) [38] proposed by OpenAI team adopts Transformer-based Decoder structure to learn text representations through large-scale unsupervised learning. It provides pre-trained language model for various natural language processing tasks. The BERT [39] model proposed by Google in 2018, which is a bidirectional encoder representation from Transformers, the attention mechanism makes the model allow

simultaneous access to all previous words without the need for an explicit time step. This mechanism facilitates parallelization and also better preserves long-term context. natural language processing (NLP) methods in the field of materials have been applied to information extraction and search, synthetic prediction, and material discovery. Over the past few years, deep learning methods based on the Transformer model have been used to extract various categories of information from materials science texts.

3.5. Generative model

Generative model refers to a series of models used to randomly generate observable data, a capability directly leveraged for the inverse design of novel material structures. Specifically, in a continuous or discrete high level space χ, where a random vector X follows the data distribution $p_r(x), x \in \mathrm{X}$, the generating model uses the existing sample $x_1, x_2, \ldots, x_N$ of known material configurations to learn parameterized model $p_\theta(x)$ to approximate the unknown distribution $p_r(x)$, and the generated sample using model $p_\theta(x)$ can approximate the real sample, effectively producing new candidate materials with targeted properties.

With the wide application of deep learning, a complex distribution $p_r(x)$ can be modeled by using the ability of deep neural network to approximate any function. Deep generation models integrate neural networks with probabilistic graphical models to approximate complex probability distributions, such as those governing material microstructures or molecular geometries. Currently, the commonly used deep generation models include VAE [40] and Generative Adversarial Network [41].

3.5.1. Variational auto-encoder

Variational auto-encoders are designed to use neural networks to model two conditional probability density functions, the inference network and the generation network, respectively. Variational auto-encoders can effectively solve the problem that the posterior distribution is difficult to estimate in probabilistic models with hidden variables. While VAE was usually applied to image generation problems, Kim *et al.* [42] used VAE to generate a continuous microstructure space based on synthetic microstructural images.

3.5.2. Generative adversarial network

The generated adversarial network (GAN) is learned by means of adversarial training, aiming to make the samples generated by the generator network conform to the real data distribution.

In GANs, there are two networks engaged in adversarial training. One is the discriminator network, the network that has the goal of accurately distinguishing whether a sample originates from real data or is produced by the generator network; The other is the generator network, which aims to generate as many samples as possible that the discriminator network cannot distinguish from the source. These two networks with opposing objectives are continuously alternating training until the discriminator network cannot determine the source of a sample, indicating that the generator network can produce samples that conform to the real data distribution.

The generative adversarial network breaks through the limitation that previous probabilistic models must learn parameters through maximum likelihood estimation. In the specific application of generative adversarial networks, Dan *et al.* [43] proposed MatGAN to inverse design inorganic materials. Long *et al.* [44] focused on generate multi-component crystal structures and proposed the CCDCGAN, the energy

required for their formation can be optimized within the latent space by utilizing reversible crystal images that offer a continuous representation.

4. Application around the design of material structure

4.1. Material property prediction

Material performance prediction can provide performance targets and optimization guidance for material structure design. Compared with traditional methods such as finite element method, deep learning can predict material properties more accurately [45−48], accelerate research process and reduce research costs. Deep learning models such as CNN-based and ConvLSTM-based architectures, as illustrated in Figure 2, have been utilized to predict a wide range of properties of crystalline and molecular materials.

Dong *et al.* [49] employed three different CNNs: VGG16, residual convolutional network and concatenate convolutional network to predict the band gap of hybrid boron-nitrogen-graphene. Firstly, density functional theory (DFT) calculation was used to generate bandgap values with random configurations, construct a dataset that include the structural information of graphene and its corresponding bandgap values for training a CNN model. They described the structure of graphene or boron nitride sheets through a two-dimensional matrix, where "0" represented carbon-carbon bond and "1" represented boron-carbon bond. The design of this structure and band gap-related material descriptor can assist CNN models in achieving more accurate predictions. Experiments show that compared with SVM, the predicted values of the three CNNs have strong linear correlation with the calculated values of DFT. In most cases, the relative error of the band gap prediction of the three CNNs for the 4×4 supercell system is less than 10%, and the accuracy is greater than 80%.

Ma *et al.* [50] proposed a novel CNN to predict defective graphene and extend it to molybdenum disulfide (Mos2). Used the chemical bond between atoms as the description unit, a three-dimensional structure description matrix is constructed by voxelization. The chemical structure parameter matrix includes the bond position matrix (BPM), bond length matrix (BLM) and bond angle matrix (BAM). By encoding the chemical bond information, it is possible to comprehensively describe the deviation of surrounding atoms from their original positions due to defects.

The multi-layer descriptor contains key position information and can identify single vacancy (SV) and double vacancy (DV) defects. It contains key length and key angle information to identify stone-wales (SW) defects. Based on the above description, they developed a CNN model with descriptors as input samples to predict the formation energy of defective graphene and MoS2. The training dataset includes the data with the distance between the defect combination and the surrounding graphene sheet greater than 15 ° to ensure the randomness and diversity of the structure, and the second-generation data obtained by translation and rotation of the structure description matrix along a certain lattice axis and the different defect concentration graphene data conforming to the Gaussian distribution are obtained. The CNN-based model reduces the huge dimension brought by multi-layer descriptors through operations such as convolution and pooling, and finally can accurately predict the formation energy of defect MoS2 system.

Yu *et al.* [51] proposed a convolutional long short-term memory model (ConvLSTM) for predicting crack paths in crystalline materials. First, molecular dynamics simulations of tensile tests on crystalline materials are performed using Lennard-Jones potentials to record crack propagation patterns at different crystal orientations and strain levels. Then, the molecular dynamics simulation results are converted into

image data as a training dataset, and the geometric matrix of the initial crack and crystal structure information is taken as input. The geometric features in the image are extracted through the convolution layer, and the LSTM layer learns the sequence relationship of crack propagation. Finally, the parameters of fracture mode, fracture toughness and crack length are accurately predicted.

Similarly, Lew *et al.* [52] proposed a ConvLSTM-based model to predict the crystalline Lennard-Jones material fracture pattern. Firstly, various graphene systems were simulated by tensile tests using MD. The results were then converted into image-based data in order to learn spatio-temporal information about the crack propagation of each graphene system. The model is composed of convolutional layer, LSTM layer and dense layer, among which there are two 1D convolutional layer, which are composed of 64 filters with kernel sizes of 60 and 61 respectively, and used to extract the geometric features of crack slices. The LSTM layer learns the sequence relationship along crack propagation, and the dense layer serves as the output layer to predict crack propagation of graphene with complex sets and shapes. The fractional dimension of the crack path predicted by the final model is slightly lower than that of the MD simulation results, but the difference between the two is less than 0.1, indicating that the proposed model can predict the fracture path of graphene under various defects and geometric shapes. Because mechanical properties are not used as input to the model, and thus proved that the model can learn the relationship between graphene fracture behavior and mechanical properties, compared with traditional MD simulation methods, the proposed model alleviates the large computational costs required to deal with the complex combination of various defects and facilitates the design of new graphene. However, 1D ConvLSTM can obtain limited spatial information, which limits the predictive performance of the model.

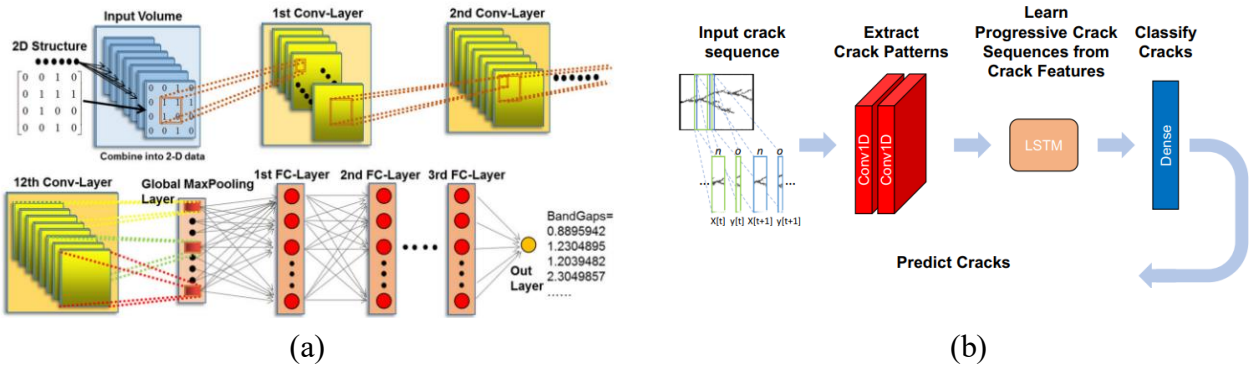

Figure 2. Schematic illustrations of material property prediction using deep learning models. **(a)** The prediction of bandgaps of 2D doped graphene systems using CNN-based models [49]. Reprinted with permission. Copyright 2019 Nature. **(b)** The prediction of fracture paths in crystalline materials using a ConvLSTM-based model [52]. Reprinted with permission. Copyright 2021 Nature.

To address the above problem, Yu *et al.* [53] introduced a 2D ConvLSTM layer to extract more space features. Specifically, the model was composed of a ConvLSTM layer, a pooling layer, two convolutional layers and a dense layer. The ConvLSTM layer is used to capture the dynamic process of crack propagation, max pooling is used to enhance the features of cracks and defects, and the convolution layer further captures the overall spatial features. In particular, compared to previous work, the geometric matrix was extended to the regions in front and behind the crack tip to better predict the crack growth

path of defective graphene. In the end, the graphene fracture path was predicted with 98% accuracy across different defect sets.

In order to predict the crack path quickly and accurately while avoiding the high computational cost of molecular dynamics simulation, Elapolu *et al.* [54] proposed a model based on CNN and Bi-RNN to predict the crack propagation path of brittle fracture of polycrystalline graphene under tensile load. CNN is adopted to extract space features including grain orientation and grain boundary, and Bi-RNN is adopted to capture sequence features of crack location and microstructure details.

Shishir *et al.* [55] proposed a deep CNN to predict Young's modulus and fracture properties of polycrystalline graphene. Based on centroidal voronoi tessellation (CVT), generate polycrystalline graphene sheets with realistic atomic structures. For the randomness of atomic structure and grain boundary orientation, 10 different atomic structures were created for each grain, and more samples were generated through mirroring and rotation operations. Finally, a dataset with 2000 grayscale images of chips were generated. The stress-strain curve was obtained by MD simulation, and the average Young's modulus and fracture stress were calculated by 5 different velocity seeds. The model can extract the average grain size, Young's modulus and fracture stress of the polycrystalline wafer. The final experiment shows that the predicted error of the proposed deep CNN model is within the standard deviation of MD simulation. The microstructure of the polycrystalline wafer will affect the fracture stress and fracture mechanism of the graphene sheet, and the grain size will affect the fracture stress and Young's modulus. The fracture strain is independent of grain size, eliminating the computational cost of traditional MD simulation methods.

The material-performance mapping of two-dimensional materials is complex, and multiple defects may coexist to affect the performance of materials. Shen *et al.* [56] proposed a model based on CNN. Each convolution operation block includes a convolution layer, a batch normalization layer and a ReLU layer, which are used to perform convolution operations on adjacent pixels of RGB images. The encoded image represents the spatial information of the object, and the final fully connected layer is used to predict the tensile strength and Young's modulus of hexagonal boron nitride (h-BN) containing vacancy defects and substitution doping defects. The results show that the R^2 value of the predicted Young modulus is 0.986 and the R^2 value of the predicted tensile strength is 0.894, which proves that the proposed model has high prediction accuracy and is helpful to the defect engineering design of h-BN.

Yang *et al.* [57] designed single-layer and multi-layer neural networks to predict the interface thermal resistance (R) between graphene and h-BN, in which both single-layer neural networks contain only one hidden layer containing 10 (ANN-10) or 20 (ANN-20) neurons; The two multi-layer neural networks each contain two hidden layers, each composed of 10 (DNN-10-10) or 20 (DNN-20-20) neurons. The training data of the model are obtained through high-throughput computing (HTC). Combined with the influence of system temperature, coupling strength and tensile strain, the trained model can predict R. The results show that neural networks perform better than machine learning models such as linear regression, polynomial regression, decision trees and random forests.

The original graphene is gapless, which limits its application in graphene-based semiconductor devices, and the nano-level pores created in graphene can not only regulate the electronic properties of graphene, but also have a significant impact on the heat transport of graphene. Previous studies have shown that the density and spatial distribution of graphene pores affect thermal conductivity. Wan *et al.* [58] proposed to learn and predict the thermal conductivity of porous graphene based on CNN, reverse design

to select the structure with the lowest predicted thermal conductivity, and conduct MD simulation to obtain the actual thermal conductivity. The newly obtained structure and its thermal conductivity were added to the training data set to retrain CNN. Until the best porous graphene structure with the lowest thermal conductivity is found.

Liu *et al.* [59] proposed a deep neural network to predict the thermal conductivity of stacked graphene structures under mechanical tensile action. Specifically, the model consists of an input layer, a hidden layer and an output layer. The hidden layer adopts CNN, and the output layer includes a fully connected layer and a regression layer. The training data of the model includes the grayscale image of the stacked graphene structure and the corresponding thermal conductivity value. In order to obtain the fingerprint information, the stacked graphene structure is projected onto the x-y plane and divided into several sub-regions, the geometric center of the sub-region is used to determine the number of graphene sheets in the sub-region, and then the matrix is constructed by the physical information pixel value (PIPV). The geometric features of graphene are captured and the final grayscale image of the graphene structure is presented. The corresponding thermal conductivity values were simulated by MD. The model is trained by minimizing the root mean square error (RMSE) between the predicted thermal conductivity and the true thermal conductivity. The results show that the accuracy of the proposed deep neural network can reach 94% when the proportion of training dataset is 12.5%. At the same time, they used stacked graphene and corresponding thermal conductivity obtained from MD simulations and deep neural network predictions to build a comprehensive database storing geometric characteristics such as the number of pieces, total area, and design domain size of stacked graphene, which can be used to search for stacked graphene structures and their corresponding thermal conductivity, helping to find stacked graphene structures with specific thermal conductivity.

Gu *et al.* [60] used CNN to predict the mechanical properties of composite materials and designed a cell battery made of hard materials and soft materials, which was composed of three different cell combinations in terms of microstructure. The different stiffness of the cell in the x and y directions resulted in symmetrical and asymmetric mechanical behaviors. In order to enable the model to learn the characteristics of these microstructure, different cells are numbered and transformed into a matrix. Meanwhile, material properties are used as part of the input to train the CNN model so that it can learn the relationship between microstructure and mechanical properties. Further, reinforcement learning is used to optimize the model. Helps design compliant materials with higher performance. The normalized root-mean-square deviation (NRMSD) of the final model was 0.2978 and 0.4926 for the training data and the testing data, respectively. By 3D printing the structure predicted by CNN to be the best value into a homogeneous sample, the best composite material designed by the model is about 25 times harder than the hard material and about 40 times harder than the soft material, indicating that the proposed model can learn from the training data and produce a better design than the training data.

Lrencezitnick *et al.* [61] adopted graph neural networks, including SchNet [62], CGCNN [63] and DimeNet [64]. The graph is represented by a set of vertices and edges, each vertex represents an atom, and the edge represents the interaction between atoms. By updating the hidden characterization of each atom so that the vertices can pass information to each other, this process amounts to simulating the interactions between atoms and can ultimately be used to predict the catalytic behavior of materials. Based on the open Catalyst project dataset OC20 [65], Set the S2EF Structure to Energy and Forces, IS2RS-Initial Structure to Relaxed Structure and IS2RE-Initial Structure to Relaxed Energy three child

tasks, to predict the energy of each atom's force and structure energy (S2EF), relaxation structure (IS2RS) and relaxation energy (IS2RE). The proposed deep learning model is used to efficiently approximate quantum mechanical calculations (e.g. DFT) to find more efficient and cost-effective electrocatalysts.

Larmuseau *et al.* [66] proposed a deep learning model based on CNN to extract features from scanning electron microscope SEM images and predict the hardness and composition of complex martensitic steel. Specifically, they split the data into two sets, the first consisting of 26 different four-element Fe-C-Mn-Si alloys with no hardness information and 34 FEG SEM images for each material. The second set of data consisted of 52 different quaternary Fe-C-Mn-Si alloys containing brinell hardness HBW, with 40 FEG SEM images for each material. First, the CNN classifier is trained using the first set of data with the aim of identifying different materials in the data set. Then, the trained CNN network is used to extract the intermediate output features of each SEM image in the second dataset, and the feature vector is reduced by PCA. Finally, the Gaussian process regression model is trained using the dimensionality reduction features to effectively predict the composition and hardness of the material, which shows that the model is able to capture the relationship between the microstructure information and the material properties.

Ren *et al.* [67] proposed a CNN-based model to predict the tensile and yield strength (YS) of dual-phase (DP) steel. First, a database containing six kinds of DP steel microstructure images and related tensile properties were established, and then the microstructure-attribute relationship was constructed based on CNN. They directly took the microstructure images as model inputs to capture more comprehensive microstructure information and predict the tensile properties. Furthermore, the gradient-class activation map (GradCAM) visualization method was used to draw the thermal maps, which helped to find the key organizational features affecting the tensile properties. The interpretability of the model is improved. For YS prediction, the average absolute error value under the traditional model is 12.5 MPa, while the MAE based on the CNN model is only 6.8 MPa, which highlights the superiority of the proposed CNN model.

Conventional methods are difficult to describe the irregular microstructure of cast alloys, resulting in the inaccuracy of the established microstructure-property relationship (MPR). Ma *et al.* [68] proposed a novel CNN model for predicting the impact toughness value of materials. CNN is used for image feature extraction. Different from traditional CNN, the proposed model selects the best feature map channel in the convolution layer and converts the feature map into 10×10 feature units through the adaptive average pooling layer to retain the spatial information of the irregular tissue connectivity and spatial distribution of the cast metal. Finally, a fully connected layer was used to map the properties to the impact toughness values, and the relationship between microstructure and toughness of cast austenitic steel at different temperatures was established. The results show that the accuracy of the proposed CNN model is 0.82, which is 0.08 higher than that of VGG. The model is trained with stochastic gradient descent and Adam optimizer. Cross-validation shows that the error between the predicted value and the actual test impact toughness is only ± 2 J, indicating that the model has good generalization ability. In order to prove that the model can effectively capture information such as volume fraction, morphology and distribution of irregular ferrites, the features extracted by the proposed novel CNN are visualized, and the microstructure such as volume fraction, morphology and distribution of irregular ferrites that have the greatest impact on the impact toughness is revealed, which increases the interpretability of the model.

Heidenreich *et al.* [69] proposed CNN-FCNN model to predict the yield surface of porous media. Firstly, two parallel CNNs are used to extract the geometric features of the microstructure. The 100 × 100 pixel image can be compressed into a vector containing 10 features through multiple convolutional layers and pooling layers. Then the FCNN decoder predicts the radial coordinates of the yield surface according to the CNN-based encoded features and loading direction. Finally, the CNNFCNN model is trained by minimizing the mean square error between the predicted value and the true value, so as to learn the relationship between the microstructure and the yield surface. Furthermore, the CNN-FCNN model can not only process two-dimensional information similar to grayscale images, but also process three-dimensional geometric information including multi-layer images from three-dimensional microscopic structures. Tasks, datasets, models, and metrics for deep learning in material property prediction are summarized in Table 1.

While deep learning models have demonstrated remarkable accuracy in predicting material properties, their black-box nature often hinders the extraction of actionable scientific insights. To address this limitation, explainable AI (XAI) techniques are increasingly being integrated into the workflow. For instance, gradient-based visualization methods such as Grad-CAM have been employed to identify critical microstructural regions that govern mechanical properties in dual-phase steels [67]. Similarly, adaptive feature map analysis has elucidated how the morphology and distribution of irregular ferrite phases impact the fracture toughness of cast alloys [68]. By providing a window into the model's decision-making process, XAI not only enhances the trustworthiness of the predictions but also facilitates a reverse mapping from target properties back to fundamental structural design principles. This transforms data-driven models from mere predictive tools into powerful engines for hypothesis generation and scientific discovery.

Table 1. Deep learning for material property prediction: tasks, datasets, models and metrics.

Tasks	Datasets	SOTA Models	Metrics
bandgap predict [49]	DFT calculation	CNN	MAE, R^2
mechanical property prediction [51−56,60,66−69]	DFT calculation, MD simulations, Finite Element Analysis, Aachen-Heerlen annotated steel microstructure dataset	ANN, CNN, RNN	R^2, NRMSD, Binary Cross-Entropy, Accuracy
thermal conductivity prediction [57,58]	MD simulations	ANN, DNN, CNN	RMSE, Accuracy
energy prediction [50,61]	Open Catalyst Project OC20, DFT calculation	CNN, GNN	MAE, EFwT

4.2. Material structure optimization

Material optimization task refers to the process of improving material performance, reducing cost and improving reliability by adjusting the microstructure or macrostructure of materials. Deep learning can assist the optimal design of material structures. By learning the relationship between material structures and properties, design parameters can be automatically adjusted to achieve the optimal performance indicators. However, the core challenge in inverse design lies in navigating the massive material design space and resolving multi-objective optimization conflicts, where generative models offer distinct advantages.

In a topological optimization problem, the goal is to determine the optimal material distribution that produces the desired properties. Abueidda *et al.* [70] proposed a CNN-based material model to predict the optimal design of elastic structures given a set of boundary conditions, loads, and optimization constraints. Three kinds of CNN models are designed for linear elastic response of materials with stress constraints, linear elastic response of materials without stress constraints and nonlinear elastic response of Neo-Hookean materials. First of all, according to the different tasks to the angle of the applied force, location of the applied force and volume constraint to randomization operating parameters; Then, use ABAQUS or MATLAB for finite element analysis to create parameter optimization problems and find the corresponding optimization design; Finally, the optimized material distribution and corresponding boundary condition, load and volume constraint are used as datasets to train the CNN model and predict the optimal material distribution. In order to prove the validity of the proposed model, experiments were carried out on the test set, and the results showed that the dice similarity coefficient (DSC) value was 0.958, that is, the results of the actual optimal design and the predicted optimal design were almost the same, which indicated that the proposed CNN model could accurately predict the optimal design. Compared to traditional finite element methods that require iterative simulations, this deep learning approach achieves a favorable trade-off by reducing computational cost from hours to minutes while maintaining high accuracy, with a DSC exceeding 0.95.

The goal of parameter optimization is to find a set of optimal parameter values, so that the designed system or model can achieve the predetermined performance goal under the premise of satisfying all constraints. These performance goals can be to maximize some performance indicator (e.g. maximum bulk modulus, maximum shear modulus, *etc.*) or to minimize some undesired indicator (e.g. minimum Poisson's ratio). Kollmann *et al.* [71] proposed a deep learning model based on CNN for predicting optimal 2D metamaterial structure design. Firstly, the data required for model training is generated by solving a large number of inverse homogenization boundary value problems. Then, three parameters of filter radius, design constraint and design target are set to optimize CNN. Ultimately, optimal metamaterial designs that predict maximized bulk modulus, maximized shear modulus, or minimized Poisson's ratio are achieved. The experimental results show that the mean square error and mean DSC on the test set are 0.00794 and 0.970 respectively, which proves that the model performs well on the unseen data. At the same time, through the case analysis, comparing the actual image and the predicted image, it is found that the similarity between them is high. Generative models like VAEs and GANs fundamentally address the massive design space challenge by learning compact latent representations, enabling efficient sampling of novel structures rather than exhaustive search. For multi-objective conflicts, they can incorporate weighted loss functions or conditional inputs to balance competing properties.

4.3. Material discovery

The aim of material discovery is to find new materials suitable for specific purposes and provide new material basis and research direction for material structure design. Previous studies have experimentally synthesized and characterized material structure or function, and finding molecular structures that meet multiple target conditions at the same time requires a lot of labor and time costs, so the use of AI algorithms with prior learning to infer the desired molecules can help accelerate the material discovery process, the specific example is shown in Figure 3. Generative models are particularly valuable for

addressing the inverse design challenge in material discovery, where the goal is to identify structures satisfying multiple target properties within vast chemical spaces.

In recent years, researchers have tried to learn the structure of molecules in neural networks. To apply neural network to molecular structure, it needs to be transformed into structured data. For the graph-like structure composed of atoms and connecting bonds such as crystal inorganic materials and organic molecules, the graph representation method is difficult to capture the symmetry and long-distance dependence of molecular structure, and the data in sequence form is easier to process. Currently, there are studies on the representation of molecular structure in sequence form. At the same time, natural language processing model has been widely used in molecular sequence generation. SMILES has been used to convert molecular diagrams into string representations and then generate new molecules using RNN-based methods. For example, Amabilino [72] and Kotsias [73] use RNN-based models, input selected molecular descriptors as conditions, and output new molecular structures matching the conditions. The disadvantage of the above model is that it is difficult to generate molecules that satisfy multiple target properties.

In order to solve the above problems, inspired by the GPT model, Rothchild *et al.* [74] proposed an organic molecule generation model C5T5 based on Transformer, which learns the semantic relationship between names (molecular properties) and molecular structure information through self-supervised method, so that molecular editing can be done according to the target in the case of zero samples. Improve the generalization ability of the model and can be used to optimize multiple molecular properties (logP, logD, PSA and refractive index); The model can choose to replace molecular fragments and formulate target property values, so as to achieve fine control of molecular optimization. The resulting molecular structures are presented in the form of IUPAC names, which contribute to the generation of interpretable organic molecules. Further, Fu *et al.* [75] trained seven models: GPT, GPT-2, GPT-J, GPT-NEO, RoBERTa, BART and BLMM on the data of extended formulas in ICSD, OQMD and Materials Projects databases. Taking full advantage of Transformer's ability to overcome remote dependencies and learn the context of chemical syntax, the results demonstrate that a causal language model-based material Transformer can effectively generate chemically valid material combinations.

Current research combined with VAE model to generate molecules, Kim *et al.* [76] proposed a model of generative chemical transformer (GCT), the use of the Transformer in the attention mechanism to identify chemical language semantics and grammar, Combined with conditional variational auto-encoder (CVAE), make the model to better understand the molecular geometry, overcome the semantics of chemical language itself is not continuity, generated at the same time satisfy multiple prerequisite for high performance, Moreover, a variety of high-performance molecules are generated under the same set of conditions to meet different requirements. VAEs address the massive design space challenge by learning continuous latent spaces that enable smooth interpolation and constrained sampling, while their conditional variants explicitly handle multi-property optimization through targeted sampling.

Researchers have also used GAN to generate molecules. Although the previous SeqGAN framework [77] can generate high-quality sequences, it cannot guarantee that the generated sequences meet the goals of specific domains. Based on the SeqGAN framework, the objective-reinforced generative adversarial networks (ORGAN) proposed by Guimaraes *et al.* [78] introduces domain-specific indicators as part of the reward function and trains the generator through reinforcement learning. To generate samples with

specific properties. The model introduces a penalty mechanism to reduce the reward for repeated sequences, and uses Wasserstein distance as a loss function, which is conducive to the generation of diversified samples. GANs tackle multi-objective conflicts through adversarial training that implicitly learns complex property distributions, while reinforcement learning integration allows explicit optimization of domain-specific objectives. The principal advantage of these deep learning-based generative models over traditional computational methods like DFT and MD is their ability to rapidly propose promising candidate materials from a vast design space at a fraction of the computational cost, shifting the focus from exhaustive simulation to efficient, intelligent exploration.

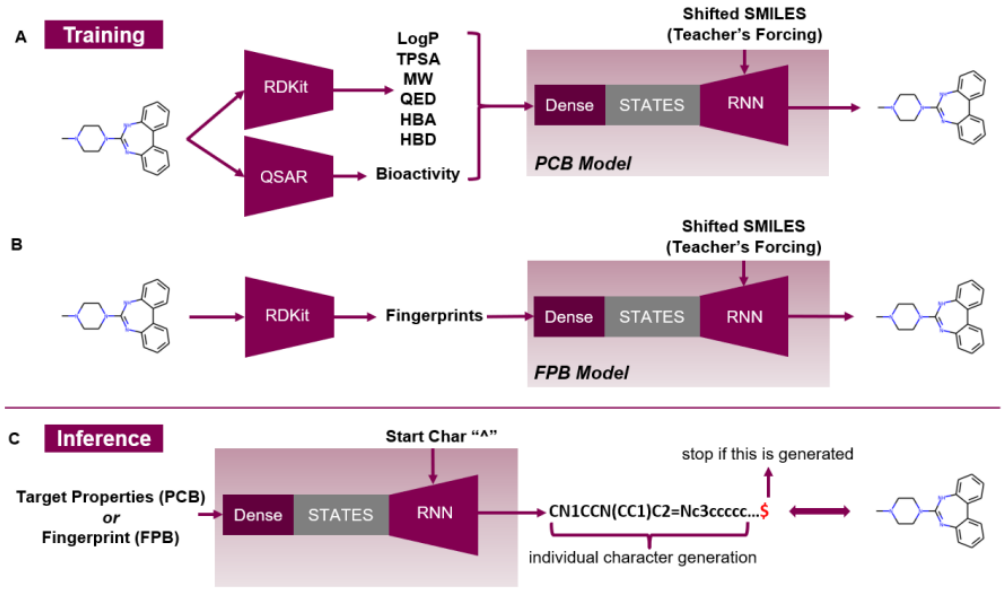

Figure 3. Schematic illustrations of material discovery using deep learning models. The RNN-based model on different conditions, combining physchem and fingerprint inference [73]. Reprinted with permission. Copyright 2020 Springer Nature.

4.4. Information extraction in material text

Researchers often store materials science knowledge in text form, and the explosive growth in published literature has made the direct analysis of documents costly. Finding literature that matches the desired performance of material systems has become very difficult. Due to the scarcity of high-quality labeled data in materials science, researchers are attempting to apply information extraction techniques from NLP to materials science texts. Information extraction aims to convert unstructured texts into structured knowledge, which holds tremendous potential in the materials science field. It can assist researchers in accessing and utilizing published literature more effectively, and in developing NLP tools to understand and process texts, as well as to construct materials corpora. Furthermore, can provide insights for tasks such as the discovery of new materials, structural design and synthesis.

Currently, information extraction in the materials field typically involves the following two subtasks: (1) named entity recognition (NER) [79] aims to identify the unstructured text has a specific meaning or refer to a strong entity and its classification for the predefined entity type; (2) relation extraction (RE) [80] aims to extract structured knowledge from unstructured texts and identify certain semantic relationships between entities.

4.4.1. Named entity recognition

The goal of the Materials NER task is to extract and identify entities such as materials, material properties and descriptors from materials science texts. There have been researches on extracting inorganic materials or material properties based on chemical NER system [81−83]. For the study of materials NER systems, previous research often utilized tools such as ChemDataExtractor [84], ChemicalTagger [85], *etc.*, for NER. Due to the diverse forms of entity writing involved in the material text, entity normalization is also important in the NER task of the material text. Based on this, researchers created expert annotation datasets for extracting non-numerical entities such as material and attribute names. As shown in Figure 4 Weston *et al.* [86] adopted the LSTM model based on supervised machine learning to realize named entity recognition. The purpose is to convert the abstract unstructured text of inorganic material articles into structured data, and to build a material science literature information extraction and analysis platform MATScholar. Specifically, Bi-LSTM will be trained using 800 abstract data labeled in BIO format, then conditional random fields (CRF) will act as a classifier to extract entities, and entity normalization operations will be performed after NER model, converting material names to standard chemical formulas (e.g. converting "titanium dioxide" to "O2Ti"). Normalized synonyms of different forms of representation (e.g. converting "chemical vapor deposition" and "CVD" to "chemical vapor deposition").

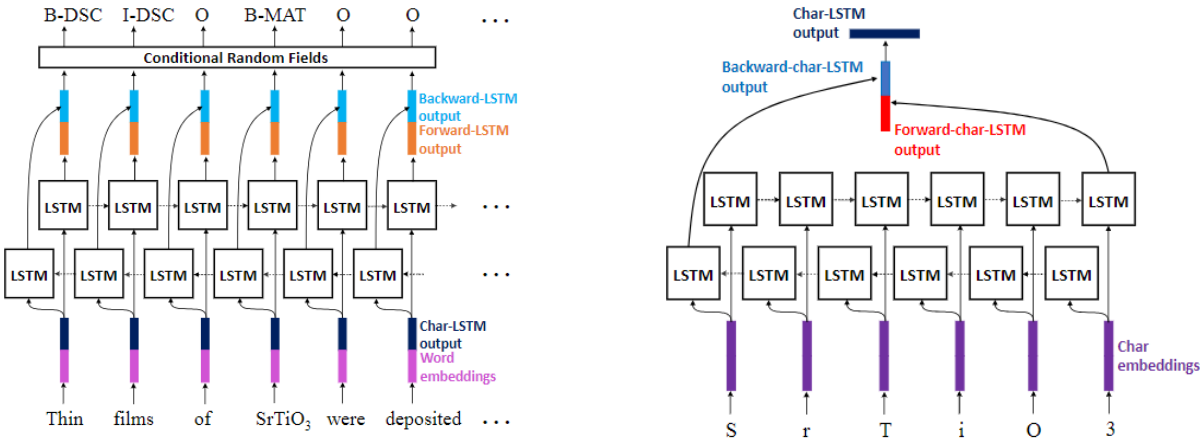

Figure 4. LSTM network architecture for named entity recognition [86]. The word-level bi-directional LSTM, that takes as input a sequence of words and returns a sequence of entity tags in IOB format (left). The character-level LSTM. This model takes a single word such as SrTiO3 and runs a bi-directional LSTM over each character to encode the morphological properties of each word (right). Reprinted with permission. Copyright 2024 Spring Nature.

Further, researchers have expanded the identification scope to identify numerical values as entities. For example, Friedrich *et al.* [87] published a SOF-EXP database containing 45 open academic papers, focusing on extracting information related to solid hydride fuel cells (SOFC). They proposed a novel labeling scheme that includes the type of material, the values and their units, and the type of equipment used in the experiment. Based on this labeling scheme, NER is implemented using BiLSTM-CRF and BERT.

With the success of large language models (LLMs) on general texts, the development of domain-specific pre-trained language models has been further promoted [88−91], a specific example is shown in Figure 5.

Beltagy *et al.* [92] conducted unsupervised training on a large number of scientific texts based on pre-trained language model BERT, and the obtained SciBERT improved the performance of downstream scientific NLP tasks. Further, Gupta *et al.* [93] based on SciBERT, MatSciBERT, a material perception language model trained by computer science, biomedical corpus and knowledge from the material field, further improved the performance of entity recognition and other tasks in scientific texts. Walker *et al.* [94] proposed the MatBERT, which uses text data from materials science journals for pre-training, and can well realize material science term extraction and paragraph level scientific reasoning. Yamaguchi *et al.* [95] focused on the field of superconducting materials, used the SciBERTNER to automatically label 9000 abstracts on the manually labeled SC-CoMIcs database, and created a term retrieval tool based on word vector similarity, which could find superconducting terms related to the query terms in the specified named entity category. Shetty *et al.* [96] took PubMedBERT as a starting point and fine-tuned on 400,000 material science abstracts to obtain the MaterialsBERT, which could accurately and automatically extract the entity (POLYMER_CLASS), single unit (MONOMER) and other entities from a large number of polymer literatures.

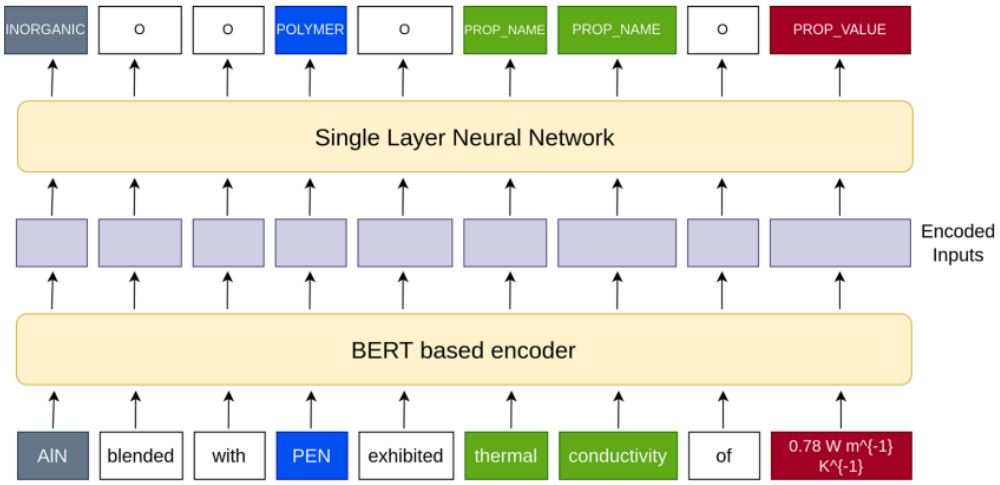

Figure 5. BERT-based architecture used for named entity recognition [96]. First, embedding each token by a BERT-based encoder. Then input the embedding vector to a single-layer neural network. The output of the neural network is the entity type of the input token. Reprinted with permission. Copyright 2023 Nature.

4.4.2. Relation extraction

Through the entity analysis of the relationship between material structural characteristics and properties described in the abstract, the key structural factors affecting material properties can be revealed and the scientific basis for how to optimize performance by adjusting structure can be provided. For example, the process-structure-property design diagram (PSPP) [97] is proposed by extracting relationships in the material library. The reciprocity of PSPP can be used to control the microstructure of materials by changing the process in the process of material design and manufacturing, and then affect the macroscopic properties of materials.

Currently, RE can be divided into three categories: rule-based methods, methods based on pre-trained language models and graph-based methods. The idea of the rule-based method is that entities closer to each other are more likely to form relationships. As shown in Figure 6, Onishi *et al.* [98] proposed a CNN-based

model on weakly supervised learning to identify factor pair relationships. Specifically, 104 factor pairs and their binary relationships were collected from the design chart as training data. Then, the CNN is used to extract sentence features. Finally, outputs the probability distribution of the two relationships in each sentence and can determine whether there is a relationship between the two scientific concepts in the material design and the nature of the relationship (positive or negative correlation). The method based on the pre-trained language model makes use of the powerful representation ability of the pre-trained model for entities, and combines the relational embedding with simple operations such as sum in series to make further predictions, a specific example is shown in Figure 7. The PURE model proposed by Zhong *et al.* [99] is based on BERT. A special "entity marker" is inserted around the entities in the candidate relation to complete the binary relation extraction. Further, Tiktinsky *et al.* [100] made a variation on the PURE model and proposed a Pure-SUM model to implemented the N-ary extraction task using the method of summing embedding. Similarly, Mullick *et al.* [101] proposed a material science relation extractor MatSciRE based on the encoder-decoder framework of pointer networks, aiming to extract triplet containing entities and relations from the scientific texts related to battery materials. For example, the sentence "The energy density based on AC and nanowire $Na_{0.35}$ MnO_2 is 42.6 Wh kg^{-1} at a power density of 129.8 Wh kg^{-1}.", There are entity "$Na_{0.35}$ MnO_2" and "42.6 Wh kg^{-1}", and their corresponding relationship is "Energy", MatSciRE can extraction triples ($Na_{0.35}$ MnO_2, Energy, 42.6 Wh kg^{-1}). The graph-based method can also well realize the N-ary extraction task.

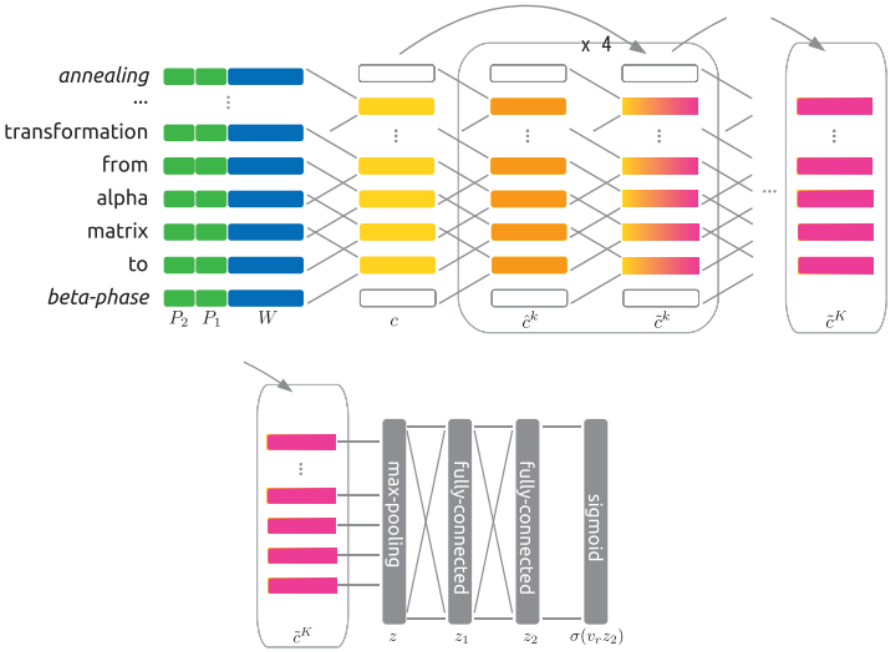

Figure 6. The rule-based method to implement relation extraction using a CNN-based model architecture [98]. The model consists of convolutional layers, max pooling, and two fully connected layers, ultimately obtaining a binary probability distribution through the sigmoid function. Reprinted with permission. Copyright 2018 Physical Sciences.

With the development of large language models, LLM-based models have been applied to relation extraction. As shown in Figure 8, Cheung *et al.* [102] randomly selected a subset of samples from the training dataset as a few shot instance and sent them to GPT-3.5-turbo and GPT-4 as direct prompts, further improving the performance of relational extraction tasks.

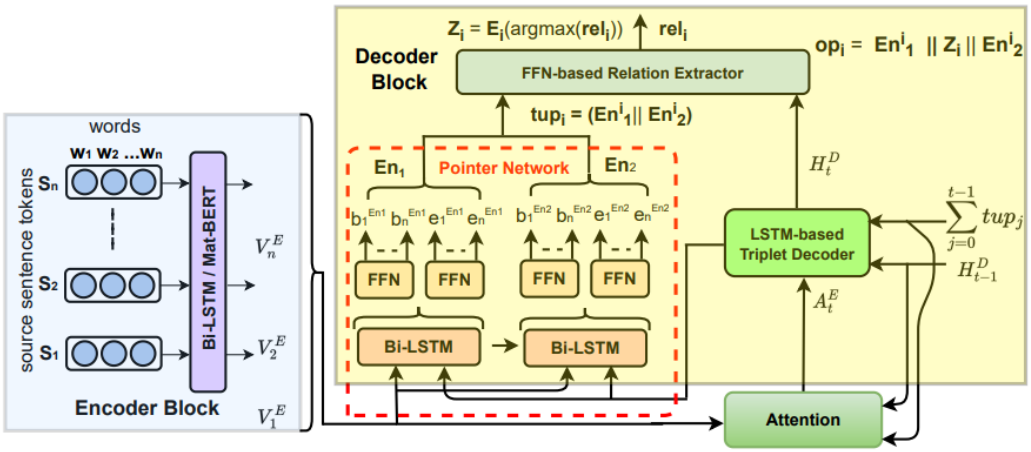

Figure 7. A method based on the pre-trained language model to implement relation extraction. Mullick *et al.* [101] proposed encoder-decoder based pointer network model, to jointly extract entities and relations from material science articles as a triplet (entity1, relation, entity2) in battery materials datasets. Reprinted with permission. Copyright 2024 Elsevier.

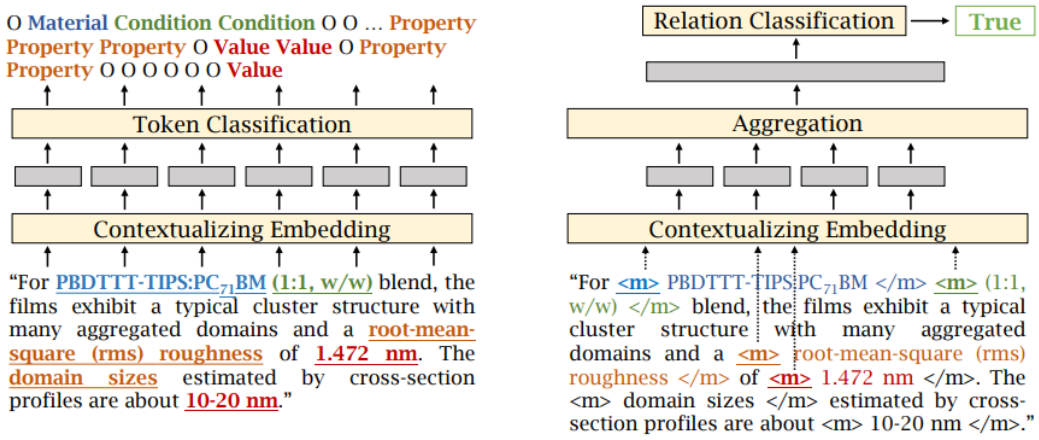

Figure 8. Model architecture for named entity recognition (left) and N-ary relation extraction (right) [102]. This architecture can study the performance of LLM including GPT-3.5 and GPT-4 on NER and RE. Reprinted with permission. Copyright 2024 ACL.

5. Hardware acceleration based on deep learning

Deep learning has been applied in fields including materials and medicine, and designing high-performance, low-power, and low latency deep learning hardware accelerators has become a hot topic in the field of architecture, with convolution operations being the focus of deep learning acceleration.

There are already many hardware platforms that can accelerate deep learning, and CPU, as the most widely used general-purpose processor, can balance general-purpose performance while accelerating computation; graphics processing units (GPUs) have parallel computing capabilities and high computational efficiency, making them the preferred choice for deep learning. Coates *et al.* [103] implemented scalable robot object detection learning based on GPU hardware, which is 90 times faster than previous optimal software versions and can be trained on millions of instances. The current GPU achieves the operation of large-scale deep learning models by increasing the number of parallel

processing channels and expanding the number of computing units, but there are still significant resource and energy consumption shortcomings.

As an embedded programmable platform, field programmable gate array (FPGA) can be directly programmed for hardware through hardware description language (HDL). Generally, edge computing hardware devices [104,105] and hardware joint optimization design [106] can be used to accelerate in-depth learning. However, writing HDL code for machine learning models is expensive. Open computing language (OpenCL) based on C/C++enables cross platform parallel programming and is compatible with various hardware devices. By using custom encoders (e.g. Intel FPGA SDK for OpenCL), attention can no longer be focused on FPGA details, reducing hardware development time and facilitating rapid deployment of complex programs on FPGA. An *et al.* [107] proposed an FPGA based CNN acceleration framework using OpenCL as the development language. They designed a selective shift quantization scheme with a differentiable threshold, which converts floating-point weights into one or two sums of powers of 2 and transforms all multiplication operations into shift operations, thereby improving the performance of the accelerator. The method of combining hardware and software acceleration based on FPGA is also commonly applied in cloud based deep learning models. Lyu *et al.* [108] proposed a scalable CNN architecture ChipNet, which uses FPGA to develop a reusable and efficient 3D convolution block for processing real-time LiDAR data. Zheng *et al.* [109] designed a fast and low-power FPGA accelerator O-Pointnet, which optimized the nonlinear implementation, multi-layer sensing layer, and max pooling layer of PointNet. For graph convolutional networks, Jamali *et al.* [110] designed a set of independent modular accelerators based on LOP computation graph isolation, which enables the constructed FPGA architecture to be used for various learning models without the need to reconfigure structure and parameters, achieving hardware accelerated deployment of dynamic graph convolutional neural network DGCNN.

In order to address the low efficiency of general accelerators, researchers have begun to focus on the development of efficient deep learning specific hardware. Currently, research has shown that using dedicated integrated circuit ASICs, such as Cambrian series [111], Eyeriss [112], TPU series accelerators [113], *etc.*, can achieve higher energy efficiency than CPU/GPU/FPGA when processing deep learning networks. Farabet *et al.* [114] used a layered approach, combining shallow convolutional computing accelerators and neural network acceleration chips, to achieve high-precision computing. Through the compiler LuaFlow, the trained convolutional neural network model was compiled into machine code that runs on NeuFlow and can be deployed on FPGA or generated into ASIC for tasks such as face recognition and scene segmentation. Furthermore, Pham *et al.* [115] evaluated the FPGA implemented NeuFlow using IBM's 45 nm SOI process library and concluded that if NeuFlow is implemented using ASIC, its performance to power ratio will reach 490 GOPs/W, far greater than the 14.7 GOPs/W achieved by FPGA and the 1.8 GOPs/W achieved by GPU.

The point cloud-based neural network processing unit (PNNPU) proposed by Kim *et al.* [116] realizes the deployment of PNNPU at the chip level in the cloud. It adopts PMMU, a page-based point-block memory management unit, combined with LLPT, which is based on linked list, to reduce the chip memory consumption. Layered block farthest point sampling (HFPS) and block skip ball query (BSBQ) for fast and efficient point processing, skip based maximum pool prediction (SMPP) for improved throughput. The PNNPU was manufactured under a 65 nm CMOS process and evaluated on 3D object inspection (3DOD) applications. The results show that it achieves a frame rate of 84.8 fps with a power consumption of

266.8 mW and an energy efficiency of 6.6–11.9 TOPS/W. However, ASics lack system-level optimization considerations when working with other general-purpose chips.

Bio-imitative pulse neural network chip has also been concerned by the industry in recent years. Merolla *et al.* [117] proposed the pulsed neural network chip TrueNorth. They used crossbar-structured SRAM to build a chip containing 5.4 billion transistors with 4096 synaptic cores connected to each other through an on-chip network. The network integrates 1 million programmable pulsing neurons and 256 million configurable synapses, which are efficient in real-time operation while using very low power. Similarly, using "brain-inspired computing" technology to mimic biological nerve cells, Kumar *et al.* [118] introduced Zeroth's processor, which aims to implement deep learning while being suitable for low-power platforms. The proposed novel neural processing unit (NPU) can realize bioheuristic learning, perceive the world and anticipate user needs, and share perceptions. Although biologically inspired neural networks are very close to real neurons, their low precision on machine learning tasks and the complexity of the process make it difficult for them to be used more widely in the current industry. A comparison of hardware acceleration platforms for deep learning is summarized in Table 2.

Table 2. A comparison of hardware acceleration platforms for deep learning.

Hardware	Model Class	Throughput/Energy
CPU	General-purpose DL Models	Baseline Reference: provides general-purpose computation, balances versatility with acceleration
GPU	Large-scale DL Models	High Throughput: 90× faster training, capable of training on millions of instances
FPGA	Customized & Medium-Scale Models	Customizable Efficiency: performance gains via shift operations, real-time LiDAR processing, low-power point cloud processing
ASIC	Fixed-Function & High-Efficiency Models	Peak Energy Efficiency: NeuFlow (ASIC): 490 GOPs/W, PNNPU: 6.6–11.9 TOPS/W
Neuromorphic Chip	Spiking Neural Networks & Bio-inspired Models	Ultra-Low Power: real-time operation at very low power, designed for low-power platforms

6. Summary and future directions

In recent years, the emergence of deep learning has brought unprecedented opportunities to investigate materials, and has far outpaced the processing capacities of conventional experimental and computational approaches. The training datasets can be collected from material databases, experimental results, and simulation computations. The dataset is then used to train various deep learning models, establishing a mapping relationship between input features and target outputs. The trained models can perform property prediction, structure optimization, material discovery and information extraction. To a certain extent, the limitations of traditional experimental and computational methods are overcome, and the research efficiency of materials structure design is greatly improved.

There have been a lot of works focused on information extraction in material texts, to build larger shareable material data sets, but establishing sizable, high-quality, openly accessible datasets can offer crucial support for deep learning research in the domain of materials. The future direction is to aggregate and share large data sets of materials to facilitate the training of more accurate and efficient models.

With the development of large language models, intuitively adding more prior knowledge and constraints to the model can improve the reliability and interpretability of the model. Future work could allow grand prediction models to fully learn prior knowledge and constraints to more accurately reflect patterns and features of the real world, resulting in more theoretically sound predictions.

To realize this vision and ensure the robust application of these powerful models, it is crucial to address their current limitations. Despite the progress, key limitations such as inadequate uncertainty quantification and poor out-of-distribution (OOD) generalization remain major hurdles. Current models often provide overconfident predictions for novel materials outside their training domain, limiting their reliability in practical discovery. Future efforts must prioritize developing robust uncertainty estimates and embedding physical constraints to enhance model trustworthiness and generalizability across the vast, unexplored chemical space.

Declaration of generative AI and AI-assisted technologies

The authors did not use generative AI or AI-assisted technologies in the writing of this manuscript.

Acknowledgments

The authors acknowledge the support of the National Key R&D Program of China (No.2024YFB3817300), the National Natural Science Foundations of China (Grant Nos. 52250005, 21875271, U20B2021), the support of the Key R & D Projects of Zhejiang Province (No. 2022C01236, 2019C01060), the National Key Laboratory of Nuclear Reactor Technology (Grant No. STRFML-2023-06), Nuclear Power Institute of China, the Entrepreneurship Program of Foshan National Hi-tech Industrial Development Zone, the Major Project of the Ministry of Science and Technology of China (Grant No. 2015ZX06004-001), Natural Science Foundation of Shanghai Science and Technology Commission (25ZR1401353), the Fundamental Research Funds for the Central Universities (22120250339), the Fundamental Research Funds for the Central Universities (22120250374), Peak Disciplines (Type IV) of Institutions of Higher Learning in Shanghai, Ningbo Natural Science Foundations (Grant Nos. 2014A610006, 2016A610273 and 2019A610106).

Authors' contribution

Jiqun Zhang: writing—original draft, conceptualization, methodology, investigation. Juhong Yu: writing—review & editing, software. Nianxiang Qiu: writing—review & editing. Yong Liu: formal analysis, validation. Yangyang Song: visualization, data curation. Liang Zhang: writing—review & editing, data curation. Shiyu Du: supervision, funding acquisition, project administration. All authors have read and agreed to the published version of the manuscript.

Conflicts of interest

Shiyu Du holds the position of Editor-in-Chief for *AI & Materials* and has not peer reviewed or made any editorial decisions for this paper.

References

[1] Lee JA, Park J, Sagong MJ, Ahn SY, Cho JW, *et al.* Active learning framework to optimize process parameters for additive-manufactured Ti-6Al-4V with high strength and ductility. *Nat. Commun.* 2025, 16(1):931.

[2] Jiang L, Wang B, Zhou Y, Jiang Y, Zhang Z, *et al.* First-principles investigations to evaluate FeN2 as an electrocatalyst to improve the performance of Li–S batteries. *Chem. Phys. Impact* 2025, 10:100785.

[3] Pham MV, Nguyen MN, Bui TQ. A staggered local damage model for fracture analysis in bi-material structures. *Vietnam J. Mech.* 2024, 46(3):217–228.

[4] Yalcin HC. Finite element analysis of evolut transcatheter heart valves: effects of aortic geometries and valve sizes on post-TAVI wall stresses and deformations. *J. Clin. Med. Res.* 2025, 14(3):850.

[5] Nedyalkova M, Heredia D, Barroso-Flores J, Lattuada M. Comparative analysis of p*K*a predictions for arsonic acids using density functional theory-based and machine learning approaches. *ACS Omega* 2025, 10(3):3128–3140.

[6] Arabie M, Toghraie D, Samani MR, Haratian M, Aghadavoudi F. Thermal performance of octadecane as phase change materials in circular tube applying molecular dynamics simulation: the effect of initial temperature. *Eur. Phys. J. Plus* 2025, 140(4):1–12.

[7] Tan Y, Sivak JT, Almishal SSI, Maria JP, Sinnott SB, *et al.* Phase-field study of precipitate morphology in epitaxial high-entropy oxide films. *Acta Mater.* 2025, 286:120721.

[8] Jiang X, Xue D, Bai Y, Wang W, Liu J, *et al.* AI4Materials: transforming the landscape of materials science and engineering. *Rev. Mater. Res.* 2025, 1(1):100010.

[9] Zhao K, Li Q. Recent advances and applications of graph convolution neural network methods in materials science. *Adv. Appl. Sci.* 2024, 9(2):17–30.

[10] Wang J, Yin Y, Zheng J, Liu L, Yao Z, *et al.* Least absolute shrinkage and selection operator-based prediction of collision cross section values for ion mobility mass spectrometric analysis of lipids. *Analyst* 2022, 147(6):1236–1244.

[11] Yoon J, Lee J, Ryu S, Park J. Enhanced energy harvesting in rotational triboelectric nanogenerator via Gaussian process regression-based Bayesian optimization. *Nano Energy* 2025, 135:110653.

[12] Dong G, Li X, Zhao J, Su S, Misra RDK, *et al.* Machine learning guided methods in building chemical composition-hardenability model for wear-resistant steel. *Mater. Today Commun.* 2020, 24:101332.

[13] Rad D, Cuc LD, Croitoru G, Gomoi BC, Mazuru L, *et al.* Modeling investment decisions through decision tree regression—a behavioral finance theory approach. *Electronics* 2025, 14(8):1505.

[14] Chandran M, Lee SC, Shim JH. Machine learning assisted first-principles calculation of multicomponent solid solutions: estimation of interface energy in Ni-based superalloys model. *Simul. Mater. Sci. Eng.* 2018, 26(2):025010.

[15] Khatavkar N, Swetlana S, Singh AK. Accelerated prediction of Vickers hardness of Co-and Ni-based superalloys from microstructure and composition using advanced image processing techniques and machine learning. *Acta Mater.* 2020, 196:295–303.

[16] Pramod CP, Pillai GN. K-means clustering based Extreme Learning ANFIS with improved interpretability for regression problems. *Knowledge-Based Syst.* 2021, 215:106750.

[17] Hong X, Wang J, Qi G. Comparison of spectral clustering, K-clustering and hierarchical clustering on e-nose datasets: application to the recognition of material freshness, adulteration levels and pretreatment approaches for tomato juices. *Chemom. Intell. Lab. Syst.* 2014, 133:17–24.

[18] Tu X, Qin T, Ji X, Wang Z, Chen J, *et al.* DBSCAN clustering model for parameter inversion using laser cutting edge morphology characteristic in Zr-4 alloy. *Opt. Laser Technol.* 2025, 184:112461.

[19] Seo J, Choi S, Paek E. NovoRank: refinement for de novo peptide sequencing based on spectral clustering and deep learning. *J. Proteome Res.* 2025, 24(2):903–910.

[20] Abdi H, Williams LJ. Principal component analysis. *Wiley Interdiscip. Rev. Comput. Stat.* 2010, 2(4):433–459.

[21] Xu F, Li X, Yang Z, C Zhu. Spatiotemporal characteristics and driving factor analysis of embodied CO2 emissions in China's building sector. *Energy Policy* 2024, 188:114085.

[22] Hu M, Ming W, An Q, Chen M. Tool wear monitoring in milling of titanium alloy Ti–6Al–4 V under MQL conditions based on a new tool wear categorization method. *Int. J. Adv. Manuf. Technol.* 2019, 104(9):4117–4128.

[23] Saito Y, Itakura K, Ohtake N, Hasegawa H. Classification of soybean chemical characteristics by excitation emission matrix coupled with t-SNE dimensionality reduction. *Spectrochim. Acta, Part A* 2024, 322:124785.

[24] Lee IH, Chang KJ. Crystal structure prediction in a continuous representative space. *Comput. Mater. Sci.* 2021, 194:110436.

[25] Sutton RS. Generalization in reinforcement learning: successful examples using sparse coarse coding. *Adv. Neural Inf. Process. Syst.* 1996, 8:1038–1044.

[26] Werbos P. Advanced forecasting methods for global crisis warning and models of intelligence. *Gen. Syst. Yearb.* 1977, 22:25–38.

[27] Sutton RS, McAllester D, Singh S, Mansour Y. Policy gradient methods for reinforcement learning with function approximation. *Adv. Neural Inf. Process. Syst.* 1999, 12:1057–1063.

[28] Mnih V, Kavukcuoglu K, Silver D, Graves A, Antonoglou I, *et al.* Playing Atari with deep reinforcement learning. *arXiv* 2013, arXiv:1312.5602.

[29] Gertsvolf D, Horvat M, Aslam D, Khademi A, Berardi U. A U-net convolutional neural network deep learning model application for identification of energy loss in infrared thermographic images. *Appl. Energy* 2024, 360:122696.

[30] Nandy S, Jose KVJ. Directed electrostatics strategy integrated as a graph neural network approach for accelerated cluster structure prediction. *J. Chem. Theory Comput.* 2025, 21(2):978–990.

[31] Sun H, Li S, Huang J, Li H, Jing G, *et al.* Dynamic spatial-temporal graph neural network for cooling capacity prediction in HVDC systems. *Energies* 2025, 18(2):313.

[32] Wang J, Guo Y, Yang L, Wang Y. Binary graph convolutional network with capacity exploration. *IEEE Trans. Pattern Anal. Mach. Intell.* 2024, 46(5):3031–3046.

[33] Wang W, Liang H, Liang S, Liu D, Zhang H, *et al.* MDGAE-DTI: drug-target interactions prediction based on multi-information integration and graph auto-encoder. In *International Conference on Intelligent Computing*, Singapore, August 5–8, 2024, pp. 232–242.

[34] Deng Z, Xu J, Feng Y, Dong L, Zhang Y. MAVGAE: a multimodal framework for predicting asymmetric drug–drug interactions based on variational graph autoencoder. *Comput. Methods Biomech. Biomed. Eng.* 2025, 28(7):1098–1110.

[35] Das D, Teixeira ES, Morales JA. Recurrent neural network/machine learning predictions of reactive channels in $H^+ + C_2H_4$ at E_{Lab} = 30 eV: a prototype of ion cancer therapy reactions. *J. Comput. Chem.* 2025, 46(5):e70033.

[36] Wang G, Feng H, Cao C. BiRNN-DDI: a drug-drug interaction event type prediction model based on bidirectional recurrent neural network and Graph2Seq representation. *J. Comput. Biol.* 2025, 32(2):198–211.

[37] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, *et al.* Attention is all you need. *Adv. Neural Inf. Process. Syst.* 2017, 30:5998–6008.

[38] Radford A, Narasimhan K, Salimans T, Sutskever I. Improving language understanding by generative pre-training. 2018.

[39] Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Minneapolis, USA, June 2–7, 2019, pp. 4171–4186.

[40] Moreno Haro LM, Oliveira-Filho A, Agard B, Tahan A. Failure detection in sensors via variational autoencoders and image-based feature representation. *Sensors* 2025, 25(7):2175.

[41] Guimaraes GL, Sanchez-Lengeling B, Outeiral C, Farias PLC, Aspuru-Guzik A. Objective-reinforced generative adversarial networks (ORGAN) for sequence generation models, *arXiv* 2017, arXiv:1705.10843.

[42] Lee IH, Chang KJ. Crystal structure prediction in a continuous representative space. *Comput. Mater. Sci.* 2021, 194:110436.

[43] Dan Y, Zhao Y, Li X, Li S, Hu M, *et al.* Generative adversarial networks (GAN) based efficient sampling of chemical composition space for inverse design of inorganic materials. *npj Comput. Mater.* 2020, 6(1):84.

[44] Long T, Zhang Y, Fortunato NM, Shen C, Dai M, *et al.* Inverse design of crystal structures for multicomponent systems. *Acta Mater.* 2022, 231:117898.

[45] Wang Z, Wang Q, Han Y, Ma Y, Zhao H, *et al.* Deep learning for ultra-fast and high precision screening of energy materials. *Energy Storage Mater.* 2021, 39:45–53.

[46] Choudhary K, Decost B, Chen C, Jain A, Tavazza F, *et al.* Recent advances and applications of deep learning methods in materials science. *npj Comput. Mater.* 2022, 8(1):59.

[47] Tagade PM, Adiga SP, Pandian S, Park MS, Hariharan KS, *et al.* Attribute driven inverse materials design using deep learning Bayesian framework. *npj Comput. Mater.* 2019, 5(1):127.

[48] Zeng S, Zhao Y, Li G, Wang R, Wang X, *et al.* Atom table convolutional neural networks for an accurate prediction of compounds properties. *npj Comput. Mater.* 2019, 5(1):84.

[49] Dong Y, Wu C, Zhang C, Liu Y, Cheng J, *et al.* Bandgap prediction by deep learning in configurationally hybridized graphene and boron nitride. *npj Comput. Mater.* 2019, 5(1):26.

[50] Ma Y, Lu S, Zhang Y, Zhang T, Zhou Q, *et al.* Accurate energy prediction of large-scale defective two-dimensional materials via deep learning. *Appl. Phys. Lett.* 2022, 120(21):213103.

[51] Hsu YC, Yu C, Buehler MJ. Using deep learning to predict fracture patterns in crystalline solids. *Matter* 2020, 3(1):197.

[52] Lew AJ, Yu C, Hsu YC, Buehler MJ. Deep learning model to predict fracture mechanisms of grapheme. *npj 2D Mater. Appl.* 2021, 5(1):48.

[53] Yu C, Wu C, Buehler MJ. Deep learning based design of porous graphene for enhanced mechanical resilience. *Comput. Mater. Sci.* 2022, 206:111270.

[54] Elapolu MS, Shishir MIR, Tabarraei A. A novel approach for studying crack propagation in polycrystalline graphene using machine learning algorithms. *Comput. Mater. Sci.* 2022, 201:110878.

[55] Shishir MIR, Elapolu MSR, Tabarraei A. A deep learning model for predicting mechanical properties of polycrystalline graphene. *Comput. Mater. Sci.* 2023, 218:111924.

[56] Shen Y, Zhu S. Machine learning mechanical properties of defect-engineered hexagonal boron nitride. *Comput. Mater. Sci.* 2023, 220:112030.

[57] Yang H, Zhang Z, Zhang J, Zeng X. Machine learning and artificial neural network prediction of interfacial thermal resistance between graphene and hexagonal boron nitride. *Nanoscale* 2018, 10(40):19092.

[58] Wan J, Jiang J, Park HS. Machine learning based design of porous graphene with low thermal conductivity. *Carbon* 2020, 157:262–269.

[59] Liu Q, Gao Y, Xu B. Transferable, deep-learning driven fast prediction and design of thermal transport in mechanically stretched graphene flakes. *ACS Nano* 2021, 15(10):16597–16606.

[60] Gu X, Chen C, Richmond DJ, Buehler MJ. Bioinspired hierarchical composite design using machine learning: Simulation, additive manufacturing, and experiment. *Mater. Horiz. J.* 2018, 5(5):939–945.

[61] Zitnick CL, Chanussot L, Das A, Goyal S, Heras-Domingo J, *et al.* An introduction to electrocatalyst design using machine learning for renewable energy storage, *arXiv* 2020, arXiv:2010.09435.

[62] Schütt K T, Sauceda H E, Kindermans P J A, *et al.* Schnet–a deep learning architecture for molecules and materials. The Journal of Chemical Physics, 2018, 48(24):241722.

[63] Xie T, Grossman JC. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Phys. Rev. Lett.* 2018, 120(14):145301.

[64] Klicpera J, Groß J, Günnemann S. Directional message passing for molecular graphs. *arXiv* 2020, arXiv:2003.03123.

[65] Chanussot L, Das A, Goyal S, Lavril T, Shuaibi M, *et al.* Open catalyst 2020 (OC20) dataset and community challenges. *ACS Catal.* 2021, 11(10):6059–6072.

[66] Larmuseau M, Theuwissen K, Lejaeghere K, Duprez L, Dhaene T, *et al.* Towards accurate processing-structure-property links using deep learning. *Scr. Mater.* 2022, 211:114478.

[67] Ren D, Wei X, Wang C, Xu W. Deep learning-based method for microstructure property linkage of dual-phase steel. *Comput. Mater. Sci.* 2023, 227:112285.

[68] Ma B, He J, Ramazani A, Fehlemann N, Wang X, *et al.* Irregular microstructure-property linkage for cast alloys by a novel deep learning approach: application on cast austenitic stainless steel. *Mater. Today Commun.* 2023, 35:105979.

[69] Heidenreich JN, Gorji MB, Mohr D. Modeling structure-property relationships with convolutional neural networks: yield surface prediction based on microstructure images. *Int. J. Plast.* 2023, 163:103506.

[70] Abueidda DW, Bakir M, Al-Rub RKA, Bergström JS, Sobh NA, *et al.* Mechanical properties of 3D printed polymeric cellular materials with triply periodic minimal surface architectures. *Mater. Des.* 2017, 122(15):255–267.

[71] Kollmann HT, Abueidda DW, Koric S, Guleryuz E, Sobh NA. Deep learning for topology optimization of 2D metamaterials. *Mater. Des.* 2020, 196:109098.

[72] Amabilino S, Pogany P, Pickett SD, Green DVS. Guidelines for recurrent neural network transfer learning-based molecular generation of focused libraries. *J. Chem. Inf. Model.* 2020, 60(12):5699–5713.

[73] Kotsias PC, Arús-Pous J, Chen H, Engkvist O, Tyrchan C, *et al.* Direct steering of de novo molecular generation with descriptor conditional recurrent neural networks. *Nat. Mach. Intell.* 2020, 2(5):254–265.

[74] Rothchild D, Tamkin A, Yu J, Misra U, Gonzalez J. C5T5: controllable generation of organic molecules with transformers. *arXiv* 2021, arXiv:2108.10307.

[75] Fu N, Wei L, Song Y, Li Q, Xin R, *et al.* Material transformers: deep learning language models for generative materials design. *Mach. Learn.: Sci. Technol.* 2023, 4(1):015001.

[76] Kim H, Na J, Lee WB. Generative chemical transformer: neural machine learning of molecular geometric structures from chemical language via attention. *J. Chem. Inf. Model.* 2021, 61(12):5804–5814.

[77] Yu L, Zhang W, Wang J, Yu Y. Seqgan: sequence generative adversarial nets with policy gradient. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, San Francisco, USA, February 4–9, 2017, pp. 2852–2858.

[78] Guimaraes GL, Sanchez-Lengeling B, Outeiral C, Farias PLC, Aspuru-Guzik A, *et al.* Objective-reinforced generative adversarial networks (ORGAN) for sequence generation models. *arXiv* 2017, arXiv:1705.10843.

[79] Rau LF. Extracting company names from text. In *Proceedings of the 7th IEEE Conference on Artificial Intelligence Application*, Miami Beach, USA, February 24–28, 1991, pp. 29–32.

[80] Yangarber R, Grishman R. Description of the Proteus/PET system as used for MUC-7 ST. In *Proceedings of the 7th Message Understanding Conference (MUC-5)*, Fairfax, USA, April 29–May 1, 1998, pp. 1–7.

[81] Hiszpanski A, Gallagher B, Chellappan K, Li P, Liu S, *et al.* Nanomaterials synthesis insights from machine learning of scientific articles by extracting, structuring, and visualizing knowledge. *J. Chem. Inf. Model.* 2020, 60(6):2876–2887.

[82] Kim E, Huang K, Jegelka S, Olivetti E. Virtual screening of inorganic materials synthesis parameters with deep learning. *npj Comput. Mater.* 2017, 3(1):53.

[83] Wang X, Hu V, Song X, Garg S, Xiao J, *et al.* ChemNER: fine-grained chemistry named entity recognition with ontology-guided distant supervision. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Punta Cana, Dominican Republic, November 7–11, 2021, pp. 5227–5240.

[84] Swain MC, Cole JM. ChemDataExtractor: a toolkit for automated extraction of chemical information from the scientific literature. *J. Chem. Inf. Model.* 2016, 56(10):1894–1904.

[85] Hawizy L, Jessop DM, Adams N, Murray-Rust P. ChemicalTagger: a tool for semantic text-mining in chemistry. *J. Cheminf.* 2011, 3(1):17.

[86] Weston L, Tshitoyan V, Dagdelen J, Kononova O, Trewartha A, *et al.* Named entity recognition and normalization applied to large-scale information extraction from the materials science literature. *J. Chem. Inf. Model.* 2019, 59(9):3692–3702.

[87] Friedrich A, Adel H, Tomazic F, Hingerl J, Benteau R, *et al.* The SOFC-exp corpus and neural approaches to information extraction in the materials science domain. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, July 5–10, 2020, pp. 1255–1268.

[88] Song Y, Miret S, Liu B. MatSci-NLP: evaluating scientific language models on materials science language tasks using text-to-schema modeling. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, Toronto, Canada, July 9–14, 2023, pp. 3621–3639.

[89] Gu Y, Tinn R, Cheng H, Lucas M, Usuyama N, *et al.* Domain-specific language model pretraining for biomedical natural language processing. *ACM Trans. Comput. Healthcare* 2021, 3(1):1–23.

[90] Lee J, Yoon W, Kim S, Kim D, Kim S, *et al.* BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* 2019, 36(4):1234–1240.

[91] Huang S, Cole JM. Batterybert: a pretrained language model for battery database enhancement. *J. Chem. Inf. Model.* 2022, 62(24):6365–6377.

[92] Beltagy I, Lo K, Cohan A. SciBERT: a pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Hong Kong, China, November 3–7, 2019, pp. 3615–3620.

[93] Gupta T, Zaki M, Krishnan NM, Mausam. Matscibert: a materials domain language model for text mining and information extraction. *npj Comput. Mater.* 2022, 8(1):1–11.

[94] Walker N, Trewartha A, Huo H, Lee S, Cruse K, *et al.* The impact of domain-specific pre-training on named entity recognition tasks in materials science. *SSRN Electronic Journal* 2021.

[95] Yamaguchi K, Asahi R, Sasaki Y. SC-CoMIcs: a superconductivity corpus for materials informatics. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, Marseille, France, May 11–16, 2020, pp. 6753–6760.

[96] Shetty P, Rajan AC, Kuenneth C, Gupta S, Panchumarti LP, *et al.* A general-purpose material property data extraction pipeline from large polymer corpora using natural language processing. *npj Comput. Mater.* 2023, 9(1):52.

[97] Olson GB. Genomic materials design: the ferrous frontier. *Acta Mater.* 2013, 61(3):771–781.

[98] Onishi T, Kadohira T, Watanabe I. Relation extraction with weakly supervised learning based on process-structure-property-performance reciprocity. *Sci. Technol. Adv. Mater.* 2018, 19(1):649–659.

[99] Zhong Z, Chen D. A frustratingly easy approach for entity and relation extraction. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, June 6–11, 2021, pp. 50–61.

[100] Tiktinsky A, Viswanathan V, Niezni D, Azagury DM, Shamay Y, *et al.* A dataset for N-ary relation extraction of drug combinations. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Seattle, USA, July 10–15, 2022, pp. 3190–3203.

[101] Mullick A, Ghosh A, Chaitanya GS, Ghui S, Nayak T, *et al.* MatSciRE: leveraging pointer networks to automate entity and relation extraction for material science knowledge-base construction. *Comput. Mater. Sci.* 2024, 233:112659.

[102] Cheung J, Zhuang Y, Li Y, Shetty P, Zhao W, *et al.* POLYIE: a dataset of information extraction from polymer material scientific literature. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Mexico City, Mexico, June 16–21, 2024, pp. 2370–2385.

[103] Coates A, Huval B, Wang T, Wu D, Catanzaro B, *et al.* Deep learning with COTS HPC systems. In *International conference on machine learning*, Atlanta, USA, June 16–21, 2013, pp. 1337–1345.

[104] Loh J, Dudchenko L, Viga J, Gemmeke T. Towards hardware supported domain generalization in DNN-based edge computing devices for health monitoring. *IEEE Trans. Biomed. Circuits Syst.* 2024, 19(1):5–15.

[105] He F, Ding K, Yan D, Li J, Wang J, *et al.* A novel quantization and model compression approach for hardware accelerators in edge computing. *Comput. Mater. Continua* 2024, 80(2):3021.

[106] Krestinskaya O, Fouda ME, Eltawil A, Salama KN. Towards efficient IMC accelerator design through joint hardware-workload co-optimization. In *2025 IEEE International Symposium on Circuits and Systems (ISCAS)*, London, UK, May 25–28, 2025, pp. 1–5.

[107] An J, Zhang D, Xu K, Wang D. An OpenCL-based FPGA accelerator for faster R-CNN. *Entropy* 2022, 24(10):1346.

[108] Lyu Y, Bai L, Huang X. ChipNet: real-time LiDAR processing for drivable region segmentation on an FPGA. *IEEE Trans. Circuits Syst. I Regul. Pap.* 2018, 66(5):1769–1779.

[109] Zheng X, Zhu M, Xu Y, Li Y. An FPGA based parallel implementation for point cloud neural network. In *2019 IEEE 13th International Conference on ASIC (ASICON)*, Chongqing, China, October 29–November 1, 2019, pp. 1–4.

[110] Jamali Golzar S, Karimian G, Shoaran M, Shoaran M, Fattahi Sani M. DGCNN on FPGA: acceleration of the point cloud classifier using FPGAS. *Circuits Syst. Signal Process.* 2023, 42(2):748–779.

[111] Kim D, Choh SJ, Liu W, Zhang X, Hong J. Cambrian Series 2 calcimicrobial crust-cement boundstone in the Yangtze Block, China: a distinctive bioconstruction as a legacy of Precambrian reef evolution. *Sediment. Geol.* 2025, 477:106804.

[112] Chen Y, Emer J, Sze V. Eyeriss: a spatial architecture for energy-efficient dataflow for convolutional neural networks. *ACM SIGARCH Comput. Archit. News* 2016, 44(3):367–379.

[113] Reidy B, Mohammadi M, Elbtity ME, Zand R. Efficient deployment of transformer models on edge TPU accelerators: a real system evaluation. In *Architecture and System Support for Transformer Models (ASSYST@ ISCA 2023)*, Orlando, USA, June 17, 2023.

[114] Farabet C, Martini B, Corda B, Akselrod P, Culurciello E, *et al.* NeuFlow: a runtime reconfigurable dataflow processor for vision. In *Proceedings of Computer Vision and Pattern Recognition Workshops*, Colorado Springs, USA, June 20–25, 2011, pp. 109–116.

[115] Pham PH, Jelaca D, Farabet C, Martini B, LeCun Y, *et al.* NeuFlow: dataflow vision processing system-on-a-chip. In *Proceedings of the 55th International Midwest Symposium on Circuits and Systems*, Boise, USA, August 5–8, 2012, pp. 1044–1047.

[116] Kim S, Lee J, Im D, Yoo HJ. PNNPU: a 11.9 TOPS/W high-speed 3D point cloud-based neural network processor with block-based point processing for regular DRAM access. In *2021 Symposium on VLSI Circuits*, Kyoto, Japan, June 13–19, 2021, pp. 1–2.

[117] Merolla PA, Arthur JV, Alvarez-Icaza R, Cassidy AS, Sawada J, *et al.* A million spiking-neuron integrated circuit with a scalable communication network and interface. *Science* 2014, 345(6197):668–673.

[118] Kumar S. Introducing Qualcomm zeroth processors: brain-inspired computing. Qualcomm OnQ Blog. 2013. Available: https://www.qualcomm.com/news/onq/2013/10/introducing-qualcomm-zeroth-processors-brain-inspired-computing (accessed on 24 January 2016).