MRS Communications (2023) 13:754–763

https://doi.org/10.1557/s43579-023-00433-3

MRS Communications

MRS 50th Anniversary Prospective

Evolution of artificial intelligence for application in contemporary materials science

Vishu Gupta, Wei-keng Liao, Alok Choudhary, Ankit Agrawal, Department of Electrical and Computer Engineering, Northwestern University, Evanston, USA

Address all correspondence to Ankit Agrawal at **ankitag@eecs.northwestern.edu**

(Received 30 April 2023; accepted 2 August 2023; published online: 16 August 2023)

Abstract

Contemporary materials science has seen an increasing application of various artificial intelligence techniques in an attempt to accelerate the materials discovery process using forward modeling for predictive analysis and inverse modeling for optimization and design. Over the last decade or so, the increasing availability of computational power and large materials datasets has led to a continuous evolution in the complexity of the techniques used to advance the frontier. In this Review, we provide a high-level overview of the evolution of artificial intelligence in contemporary materials science for the task of materials property prediction in forward modeling. Each stage of evolution is accompanied by an outline of some of the commonly used methodologies and applications. We conclude the work by providing potential future ideas for further development of artificial intelligence in materials science to facilitate the discovery, design, and deployment workflow.

Introduction

Materials science, like any other field in science and technology, constitutes of four paradigms that are empirical, theoretical, computational, and data-driven science.[1–4] Over the last couple of decades, the increasing availability of advanced computational resources and the generation of big data[5–9] using the first three paradigms have shifted our approach from traditional methods to data-driven methods for data analysis (Fig. 1). Traditional methods involve designing empirical formulations and computational methods with chemical intuition and performing trial and error-based hands-on experimentation and/or simulations. However, with a near-infinite space of possible candidate materials, trying to discover new materials with desirable properties and performance using traditional methods becomes extremely costly and time-consuming. Hence, data-driven methods have become extremely popular for screening purposes, which can significantly reduce the cost and development time compared to hands-on experiments and simulations. Data-driven methods use artificial intelligence (AI) techniques that have been employed and improved upon by people in their respective fields of research for various applications.[10–16] These data-driven AI techniques have also been used to help solve various tasks in the field of materials science, which can be broadly categorized into forward modeling for property prediction analysis and inverse modeling for process optimization and materials design and have helped materials scientists better understand the underlying correlations and advance the frontier of knowledge.[17–29]

Some of the learning methodologies used to perform forward and inverse modeling include reinforcement learning, active learning, generative modeling, genetic algorithms, scientific machine learning, and transfer learning. Reinforcement learning involves decision-making tasks where the agent is trained to make optimal decisions within an environment by trial and error to obtain maximum reward.[30,31] Active learning aims to efficiently label or acquire new data by iteratively selecting the most informative samples from an unlabeled dataset.[32,33] Generative modeling is used to train models which learn the underlying relation and patterns within the input dataset and use that information to generate new samples that have similar characteristics to that of the input data.[34,35] A genetic algorithm is an optimization technique designed to search and find near-optimal solutions to complex problems with a large solution space or non-linearity of the objective function.[36,37] Scientific machine learning focuses on creating models that incorporate constraints based on scientific knowledge and physical principles when training the model.[38,39] Transfer learning is a technique that involves leveraging pre-learned knowledge from one task or domain to improve performance on a different task or domain.[40–42] All these methods are increasingly gaining interest and applicability in materials science to accelerate material discovery, property prediction, and optimization, leading to the development of new materials with tailored properties and enhanced performance.

In this brief Review, we provide a high-level overview of the evolution of AI in contemporary materials science for the task of materials property prediction in forward modeling. The three stages of evolution discussed in this work include 'Traditional Machine Learning,' 'Conventional Deep Learning,' and 'Graph Neural Networks.' Each stage of evolution is accompanied by an outline of some of the commonly used methodologies and/or network architectures and general applications. We conclude

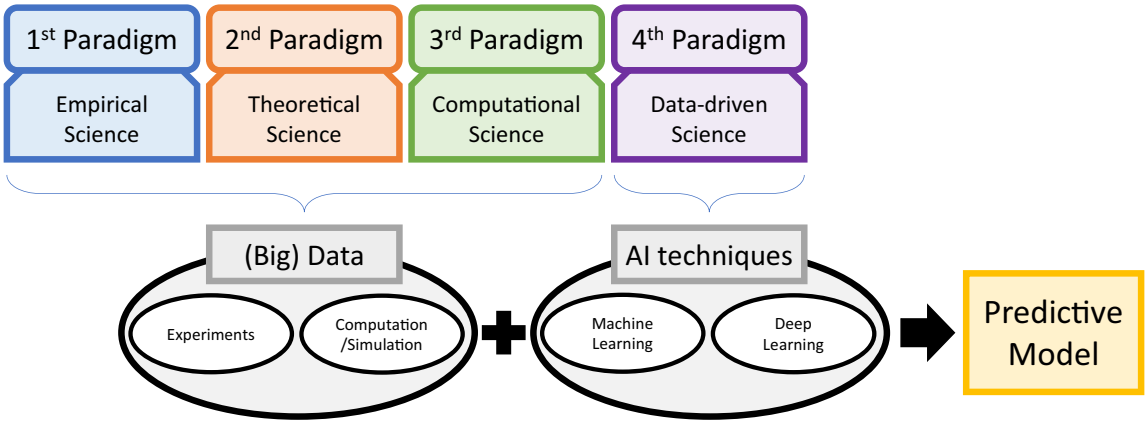

Figure 1. The four paradigms of science and its application toward predictive modeling.

the work by providing some possible future ideas for further development of artificial intelligence in materials science to facilitate the discovery, design, and deployment workflow.

Traditional machine learning

The advent of AI in materials science was accompanied by the application of traditional machine learning (ML), which fundamentally consists of algorithms that learn from structured data and build a (usually somewhat easily interpretable) model to make predictions. Traditional ML algorithms have been widely used for classification, regression, and clustering tasks in materials science.[43–47] To construct an effective and efficient ML model, one has to choose the algorithm used for model training and perform feature engineering to develop a suitable representation for the input data. Some of the traditional machine learning algorithms commonly used for predictive analysis for both classification and regression tasks are shown in Table I.

Applications

Most works involving traditional ML in materials science put emphasis on the input representation obtained via feature engineering of the unstructured data based on domain knowledge.[43,47,53–61] The general workflow for the data-driven approach that incorporates traditional machine learning for training predictive models in materials science is shown in Fig. 2.

The workflow comprises the following steps: (1) Obtain raw input files from the first three paradigms of materials science, i.e., empirical science, theoretical science, and computational science; (2) Generate ML-friendly features from a featurizer that uses the domain knowledge to obtain attributes that represent a sufficiently diverse range of physical/chemical properties for a given composition and/or structure; (3) Feed the ML-friendly features into the traditional ML technique of the user's choice, depending on the input representation and application, in order to maximize the model performance; and (4) Use the trained model in further analysis involving materials property prediction for materials discovery, design, and deployment. A few examples of solving materials science problems with traditional ML techniques trained on materials data with a brief overview are as follows. Work in[59] used thousands of descriptors obtained via domain knowledge-based feature engineering containing combinations of elemental properties such as the atomic number and ionization potential to analyze the tendency for materials to form different crystal structures. Meredig et al.[43] used the fraction of each element present and various intuitive factors, such as the maximum difference in electronegativity as the materials representation to perform predictive modeling for the formation energy of ternary compounds.

Table I. Traditional machine learning algorithms for predictive analysis.

Modeling algorithm	Brief description
K-Nearest neighbors (KNN)[48]	Uses proximity information to perform classifications/regression of class/value of a given data point
Support vector machine (SVM)[49]	Construct a hyperplane in multidimensional space which maximizes the distance between different categories/values to distinctly classify/predict the data points
Decision tree[50]	Model structured in the form of a tree, derived from the independent variables in the dataset, with each node having a condition over a feature
Random forest[51]	Combines the output of multiple decision trees with a fixed set of parameters to reach the final result
XGBoost[52]	Sequentially built shallow decision trees whose parameters adjust by itself iteratively and produce a stronger prediction

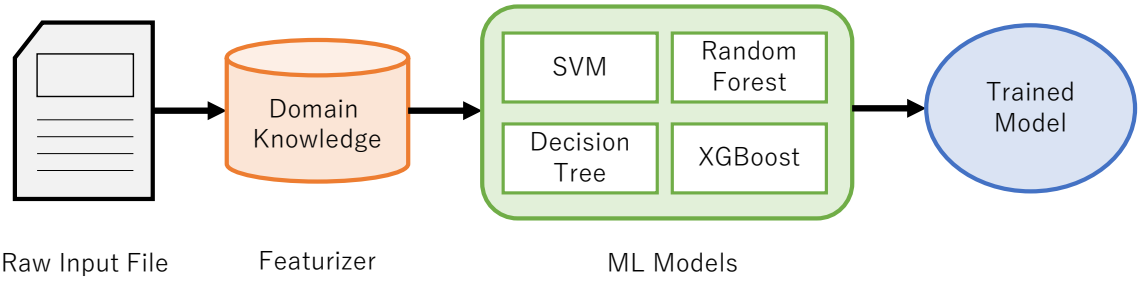

Figure 2. The general workflow for the data-driven approach that incorporates traditional machine learning for training predictive models in materials science.

Work in[58] performed ML algorithm-based predictive modeling for different materials properties using a generalized set of composition based consisting of stoichiometric attributes (e.g., number of elements present in the compound, several L^p norms of the fractions), elemental property statistics (e.g., mean, mean absolute deviation, range, minimum, maximum and mode of different elemental properties), electronic structure attributes (e.g., average fraction of electrons from the *s*, *p*, *d*, and *f* valence shells between all present elements), and ionic compound attributes (e.g., whether it is possible to form an ionic compound based on[62]) as materials representations. Faber et al.[60] took 3938 entries from Materials Project and used Coulomb matrix (CM)-based representation to train the ML model to obtain a low prediction error in cross-validation. Work in[47] used representation based on four different kinds of structural descriptors to create a model for cohesive energy from 18,903 entries consisting of compounds based on a select set of structures and elements. Schütt et al.[61] predicts the density of states at the Fermi level using an ML model with a representation based on the partial radial distribution function and demonstrate that the model can be used to predict the property value for crystal structures outside of the original training set. Work in[63] uses 126 features derived from the local environment of each atom within a crystal structure known as Voronoi tessellation (e.g., effective coordination number, structural heterogeneity attributes, chemical ordering attributes, maximum packing efficiency, local environment attributes) of a material to perform predictive analysis. These sets of descriptors generated via feature engineering tend to be more effective toward a specific materials property only making them less generalizable. Moreover, in general, ML algorithms are less scalable with an increase in the number of data points.

Conventional deep learning

Using unstructured data as model input for traditional ML algorithms is challenging as the user has to first perform manual or domain knowledge-based feature engineering and then select a desirable algorithm to train the ML model. This makes the whole workflow costly, time-consuming, and difficult to scale with the ever-increasing data. In such scenarios, deep learning (DL) algorithms—which are ML algorithms based on deep neural networks—have emerged as a powerful tool for performing predictive analysis. Given a large materials dataset available for training the model, DL techniques can automatically and efficiently extract features from those unstructured data and build accurate models for different materials properties, often surpassing traditional ML techniques. There are different types of deep neural networks that can be used for training the model depending on the input representation of the unstructured data, some of which are shown in Table II.

Applications

Deep learning offers an alternative route for accelerating the production of predictive models by being able to excel on raw inputs and therefore reducing the need for designing physically relevant features using manual or domain knowledge-based feature engineering. There have been several works that used

Table II. Well-known conventional deep learning algorithms for predictive analysis.

Modeling algorithm	Brief description
Feedforward neural network[64] (FNN)	A network of a straight line from input to output connected via hidden layers of neurons. They are used for a wide variety of tasks when dealing with fixed length tabular data
Convolutional neural network[65] (CNN)	A type of neural network that is specifically designed for image processing tasks. Their ability to learn the spatial relationships between pixels in an image make them effective for tasks, such as image classification and object detection
Recurrent neural network[66] (RNN)	A type of neural network that is able to process sequential data by learning long-term dependencies between data points. They are often used for tasks, such as natural language processing and speech recognition

deep neural networks to perform forward modeling for the task of materials property prediction.[67–76] The general workflow for the data-driven approach that incorporates conventional deep learning for training predictive models in materials science is shown in Fig. 3.

The workflow comprises the following steps: (1) Obtain raw input files from the first three paradigms of materials science, i.e., empirical science, theoretical science, and computational science; (2) Obtain DL-friendly features directly from the raw input files without going through the featurizer that uses domain knowledge to obtain composition- and/or structure-based attributes incorporating a sufficiently diverse range of physical/chemical properties; (3) Feed the DL-friendly features into the conventional DL technique of the user's choice, depending on the input representation and application, in order to maximize the model performance as given in Table II; and (4) Use the trained model in further analysis involving materials property prediction for materials discovery, design, and deployment. It is advisable to apply conventional DL methods on a given dataset when dealing with a large dataset, as conventional DL methods are consistently shown to help improve the performance of the trained model as compared to traditional ML techniques in large dataset scenarios. A few examples of solving materials science problems with conventional DL techniques trained on materials data with a brief overview are as follows. Harvard Energy Clean Project by Pyzer-Knapp et al.[77] used a three-layer network for predicting the power conversion efficiency of organic photovoltaic materials. Montavon et al.[68] predicted multiple electronic ground-state and excited-state properties using a model trained on a four-layer network on a database of around 7000 organic compounds. Zhou et al.[67] used high-dimensional vectors learned using Atom2Vec along with a fully connected network with a single hidden layer to predict formation enthalpy. CheMixNet[70] and Smiles2Vec[78] applied deep learning methods to learn molecular properties from the molecular structures of organic materials. ElemNet[69] used a 17-layered architecture to learn formation enthalpy from elemental fractions but has shown performance degradation beyond that depth (Fig. 3). Work in[71] used a combination of principal component analysis and convolutional neural networks to predict the stress–strain behavior of binary composite. Zheng et al.[79] uses multi-channel input for the deep convolution neural networks to improve the prediction accuracy as compared to single input channels. Nazarova et al.[80] use recurrent neural networks along with a series of optimization strategies to achieve high learning speeds and sufficient accuracy for the task of polymer property prediction. Yang et al.[72] used convolution recurrent neural networks to learn and predict several microstructure evolution phenomena of different complexities. IRNet[26,73] introduced the concept of deeper neural network architecture in materials science, where they build 17-, 24-, and 48-layered architecture with residual connections to learn different materials properties from the composition and structure information of a crystal without degrading the performance. Branched Residual Network (BRNet) and Branched Network (BNet)[74] introduce the concept of branching in neural network architecture to perform materials properties prediction from the composition-based attributes for improved performance under parametric constraints (Fig. 3).

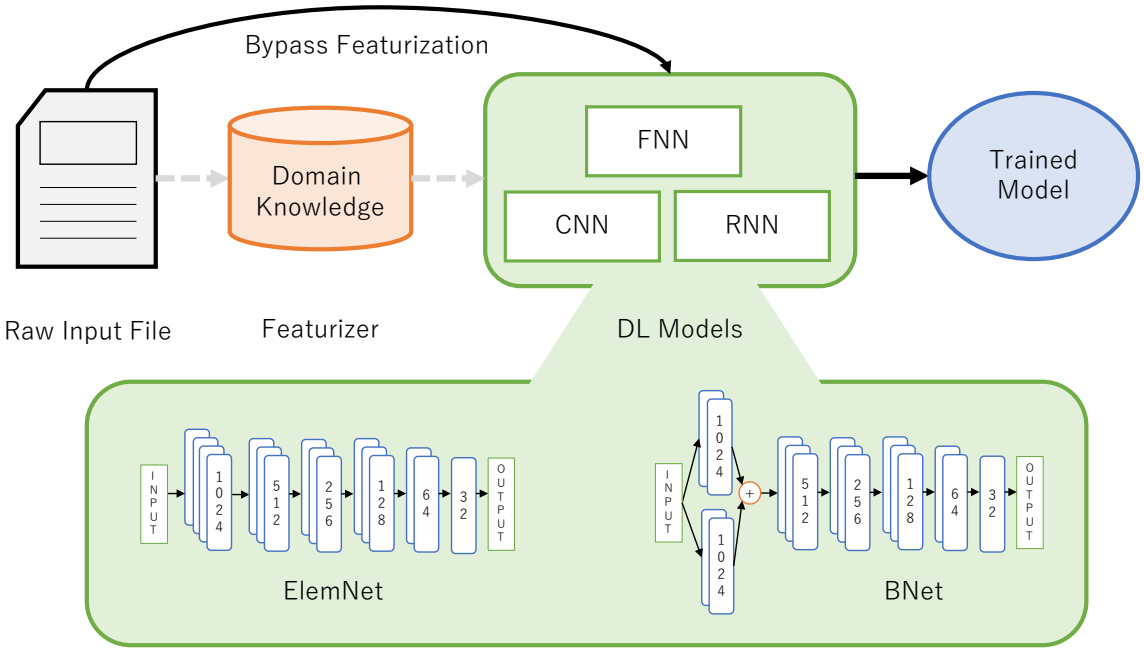

Figure 3. The general workflow for the data-driven approach that incorporates conventional deep learning for training predictive models in materials science.

Graph neural networks

Conventional DL is used to perform predictive analysis when dealing with input representation based on Euclidean datasets comprised fixed forms (such as images, text, and numerical tables). These datasets also tend to work on the fundamental assumption that every instance is independent of each other. However, applying conventional DL algorithms becomes challenging when presented with more complex data represented as graphs (non-Euclidean) without a fixed form and comprised intricate interactions between the instances inside the graph. Graph neural networks (GNNs) are a category of deep learning algorithms used to handle and perform inference on the complex data represented as graphs. GNNs work by iteratively updating the representation of each node in the graph based on the representations of its neighbors. This allows GNNs to learn about the local and global structure of the graph, which can be used to perform predictive analysis for various applications. Some of the common examples of GNNs based on how they learn the representations of nodes are shown in Table III.

Table III. Common examples of GNNs based on how they learn the representations of nodes in a graph.

Modeling algorithm	Main component
Graph convolutional networks (GCN)[81]	Convolutional layers
Graph attention networks (GAT)[82]	Attention mechanisms
GraphSAGE[83]	Sampling-based approach
Graph autoencoders (GAE)[84]	Autoencoder architecture
PinSAGE[85]	Pooling-based approach

Applications

As GNNs are able to capture the complex relationships between the nodes and edges in graphs, they have been used to learn the atomic interaction or the material embeddings from the crystal structure and composition.[61,86–95] The general workflow for the data-driven approach that incorporates graph neural networks for training predictive models in materials science is shown in Fig. 4.

The workflow comprises the following steps: (1) Obtain raw input files from the first three paradigms of materials science, i.e., empirical science, theoretical science, and computational science; (2) Obtain GNN-friendly features from the raw input files, which are usually represented in a graph form with intricate interactions between the instances inside the graph. Additionally, in some cases, atom-, bond-, or angle-based embeddings containing the pre-defined knowledge are also provided as a form of input to the model in order to aid the training process; (3) Feed the GNN-friendly features with/without embeddings into the graph neural network of

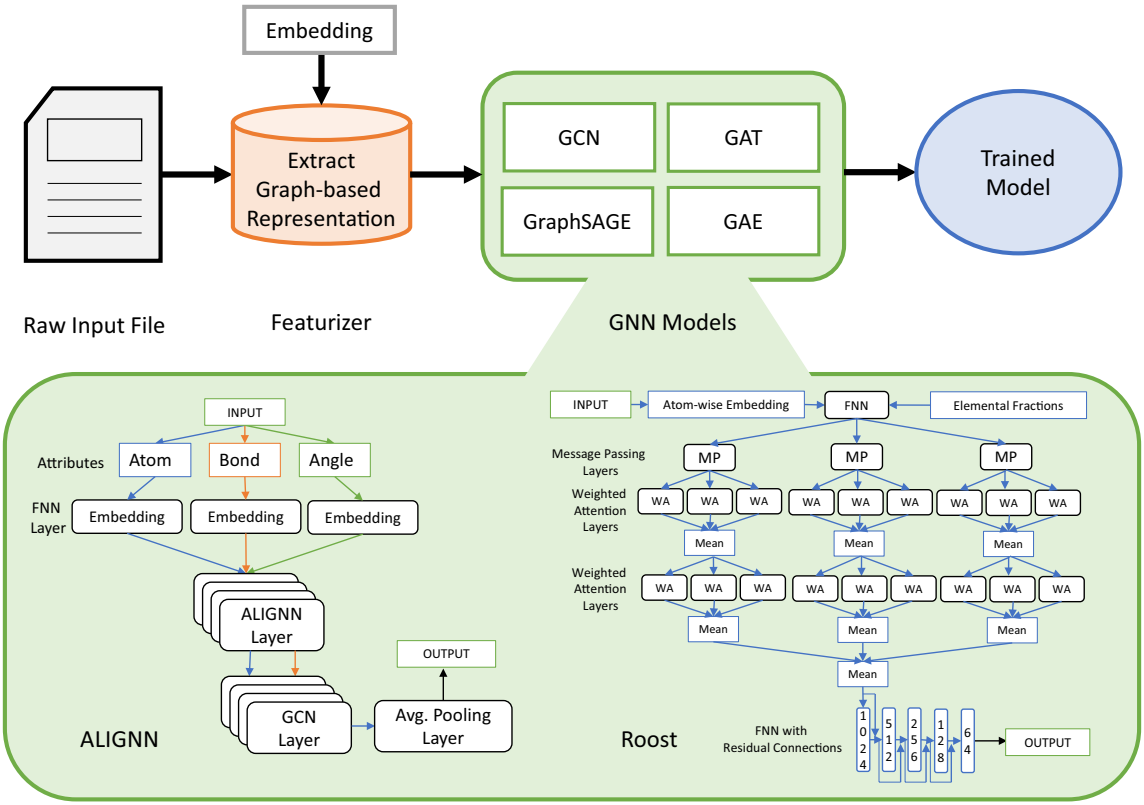

Figure 4. The general workflow for the data-driven approach that incorporates graph neural networks for training predictive models in materials science.

the user's choice, depending on the input representation and application, in order to maximize the model performance as given in Table III; and (4) Use the trained model in further analysis involving materials property prediction for materials discovery, design, and deployment. In a real-life scenario, a given compound is always represented in a graphical form with intricate interactions between the atoms. Moreover, for a given compound, it is possible to have various structure types or polymorphs with completely different materials property values, which is difficult for traditional ML and conventional DL techniques to distinguish. Hence, graph neural networks tend to perform better compared to other techniques due to their ability to excel in such scenarios. A few examples of solving materials science problems with graph neural networks trained on materials data with a brief overview are as follows. Crystal graph convolution neural networks (CGCNN)[88] directly learn material properties via the connection of atoms in the crystal structure of the crystalline materials, providing an interpretable representation, which was then improved in[89] by incorporating Voronoi-tessellated crystal structure information, explicit 3-body correlations of neighboring constituent atoms, and optimize chemical representation of interatomic bonds in the crystal graph. OGCNN[92] incorporates orbital–orbital interaction and topological characteristics information in the CGCNN model to improve the performance of the model. A-CGCNN[93] introduces an attention mechanism and normalizing node features in the network architecture to improve the prediction accuracy of the CGCNN model. SchNet[61] incorporated continuous filter convolutional layers to model quantum interactions in molecules for the total energy and interatomic forces which was then extended in[86] where the authors used an edge update network to allow for neural message passing between atoms for better property prediction for molecules and materials. MatErials Graph Network (MEGNet)[87] was developed as a universal model for materials property prediction of different crystals and molecules, which uses temperature, pressure, and entropy as global state inputs. Goodall and Lee[90] developed an architecture called Representation Learning from Stoichiometry (Roost) that takes elemental fraction-based stoichiometric attributes as input features along with embedding obtained via material science literature using advanced natural language processing algorithms known as matscholar embedding to learn appropriate materials descriptors from data. The architecture uses a graph neural network that takes matscholar embeddings and the elemental fraction of each element present in the compound, which is passed through a series of parallelly stacked message-passing layers, weighted attention layers, and fully connected layers with residual connections before making a prediction (Fig. 4). Directional Message Passing Neural Network (DimeNet)[96] and DimeNet++[97] use the directional information by transforming messages based on the angle between the atoms along with spherical Bessel functions and spherical harmonics to achieve better performance than the Gaussian radial basis representations with latter model being faster as compared to the former model. Geometric Message Passing Neural Network (GemNet)[98] was developed as a universal approximator for molecule predictions that is invariant to translation and equivariant to permutation and rotation using directed edge embeddings and two-hop message passing in its architecture. Atomistic Line Graph Neural Network (ALIGNN)[91] combines different structure-based features, including atom, bond, and angle information of the materials, to perform materials property prediction and obtain high-accuracy models for improved materials property prediction. ALIGNN architecture consists of embedding layers for each of the input types, followed by the ALIGNN layer and GCN layer, each containing two edge-gated graph convolution layers[99] and one edge-gated graph convolution layer, respectively, and finally an average pooling layer before making a prediction (Fig. 4). ALIGNN was then improvised as ALIGNN-d in[100] where they introduced dihedral angles along with other information as the model input. DeeperGATGNN[94] constructed based on GATGNN[95] combines residual connections and global attention mechanism with differentiable group normalization to address the over-smoothing issue and improves the prediction accuracy of crystal properties when dealing with large datasets. Graphormer[101] uses a self-attention mechanism in the GNN to achieve significantly improved performance in the prediction of crystal and molecular properties in the OGB[102] and OC20[103] challenges. Crystal Edge Graph Attention Neural Network (CEGANN)[104] learns unique feature representations using graph attention-based architecture and performs classification of materials across multiple scales and diverse classes.

Future directions

The widespread use and development of various AI-based models in materials science inspired by standard practices in the computer science community have led to the utilization of advanced algorithms with tailor-made input representations for application in materials science. With how fast materials science is catching up with the state-of-the-art methodology in computer science, it is just a matter of time before researchers formulate a method to design a generalized workflow to extract input representation which can then be used to train the next generation of neural networks. Hence, in this section, we would like to give a brief overview of a new class of neural networks known as the graph matching networks (GMNs)[105], which might be the next class of neural networks that can help advance and boost the model accuracy closer to the chemical accuracy if modified and implemented for the materials science community.

Graph matching networks

GMNs[105] is a class of neural networks that is used to perform supervised learning by processing the similarity between a pair of graphs given as input. As the name suggests, they are particularly well suited for tasks where the input data are structured in the form of a graph. As GMNs take a pair of graphs as the model input and jointly compute the similarity score on

the pair, they tend to be potentially more powerful than the embedding models, which independently map each graph to a vector. The network architecture comprises encoders (one for each graph) and a cross-graph attention mechanism. First, the encoder is used to produce vector representation for each node in the graph given as model input. Then, these vector representations are passed to the cross-graph attention mechanism to compute a similarity score between each pair of nodes. Finally, the similarity score between each pair of nodes is used to compute a final similarity score between the two graphs. Compared to the embedding-based graph models, the matching model can potentially change the vector representation of the graphs on the basis of the other graph used for comparison. This way, if the two graphs do not match, the model will modify the vector representation of the graph to be comparatively more different from the two graphs that match.

GMNs have been shown to be a powerful tool for both graph–graph classification and graph–graph regression tasks, as it is able to learn the complex relationships between the nodes and edges of graphs more effectively as compared to the traditional methods. They also generalize well to new graphs that have not been seen during training by learning robust vector representations of graphs. Although GMNs have been shown to outperform state-of-the-art models in several studies, they are computationally expensive to train, difficult to interpret, and highly sensitive to the choice of hyperparameters. Overall, GMN is a new approach that has shown to be promising when utilizing graph-structured objects for performing graph–graph classification and regression[106–109] and is expected to improve the robustness and accuracy of the predictive modeling for different scientific domains, including materials science.

Conclusion

AI has grown to become an important and flexible tool with various applications for materials discovery, design, and deployment. In this section, we discuss some other facets of AI in the context of materials informatics, which are important considering the growing interest, applicability, and impact of data-driven approaches in materials science.

Limitations and challenges

There still exist a wide variety of limitations and challenges that need to be worked upon to leverage the maximum potential of AI-based models in the materials science community. Some of these limitations and challenges (Fig. 5) include reliability and quality assessment of datasets, uncertainty quantification of the deployed model, conversion of the raw data to tailor-made input representation, explainability/interpretability of the trained model for prediction tasks, reproducibility, transferability, and usability of the complex models. Moreover, for most of the datasets, it is only feasible to use traditional machine learning or convectional deep learning techniques due to the lack of tools and information to convert the raw data into a suitable input representation that can be fed into the more advanced

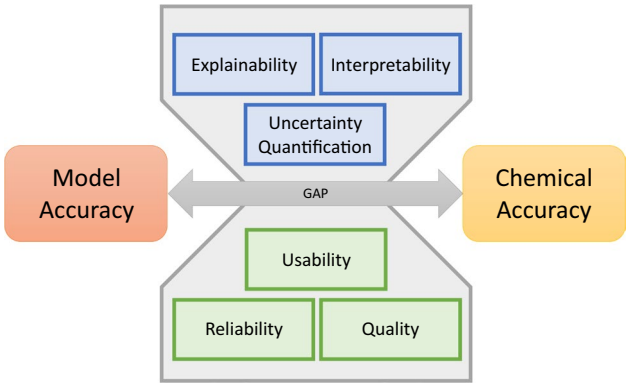

Figure 5. Limitations and challenges of AI-based models in the materials science community.

methods, such as graph neural networks. Although there have been ongoing efforts to address the challenges associated with the application of AI in materials science for materials discovery, design, characterization, and performance prediction, which incorporates various techniques,[110–113] these areas of research are still in their nascent stage. Hence, more research on filling the gap in the knowledge between AI-based models and their application to the materials science problem will help to better understand underlying correlations, create an easy-to-use pipeline for raw data to tailor-made input representation conversion, potentially determine physical laws and knowledge that are currently unknown, and bring down the model errors to resemble the chemical accuracy and eventually contributing to scientific understanding and progress with minimal human input.

Ethical considerations

Like any other scientific field, materials science must look into various ethical considerations when incorporating AI in its research, development, and applications. These ethical considerations are essential to ensure responsible and sustainable progress and prevent unintended negative consequences. Some of the key ethical considerations in materials science include (1) potential impacts on jobs: we need to recognize the potential negative effect of new materials and technologies on the job market and come up with proactive measures to mitigate negative impacts on workers, such as retraining programs, reskilling initiatives, and social safety net; (2) recognizing AI-generated content: reducing the risk of spreading possible misinformation through data generated using AI (such as generating and spreading materials properties and structures information obtained from generative modeling) via rigorous validation and testing using transparent and reproducible practices; and (3) bias in AI models: mitigating bias in AI models when dealing with training data that are unrepresentative or contain inherent biases to ensure fair and unbiased predictions by trying to use diverse and inclusive datasets and employing bias detection and correction techniques to minimize potential

biases. Addressing these considerations and establishing ethical guidelines and codes of conduct can help guide responsible research and innovation in materials science.

Collaborative efforts

Successful multidisciplinary collaborations have played a significant role in advancing AI in materials science. These collaborations bring together experts from various fields to tackle complex challenges, create innovative solutions, and open new possibilities. Some examples of how experts from different fields can work together to advance the field include: (1) Computer scientists provide expertise in algorithm development and optimization by developing AI models and algorithms tailored for materials data analysis and prediction; (2) Data scientists offer insights into handling large and complex materials datasets by ensuring data quality and accessibility for AI-driven analyses; (3) Computational materials scientists contribute their knowledge in developing efficient simulation methods by incorporating high-performance computing into the workflows; (4) Experimental materials scientists provide guidance on relevant material properties and structures and insights into material processing and performance evaluation; (5) Chemists contribute their expertise in chemical synthesis and offer domain-specific knowledge on material properties and molecular structures; and (6) Industry partners provide real-world testing and validation, ensuring the practical relevance of AI models and materials discoveries. Fostering an inclusive and collaborative research environment with experts from different disciplines can collectively advance the field, leading to transformative discoveries and developments in materials science.

Acknowledgments

This work was performed under the following financial assistance award 70NANB19H005 from U.S. Department of Commerce, National Institute of Standards and Technology as part of the Center for Hierarchical Materials Design (CHiMaD). Partial support is also acknowledged from NSF award CMMI-2053929 and DOE awards DE-SC0019358, DE-SC0021399, and Northwestern Center for Nanocombinatorics.

Author contributions

VG conceptualized the structure of this review article with input and guidance from AA, AC, and WL. VG and AA wrote the manuscript. All authors discussed and reviewed the manuscript.

Declarations

Conflict of interest

The authors declare that they have no competing interests.

Open Access

References

1. A. Agrawal, A. Choudhary, Perspective: Materials informatics and big data: Realization of the fourth paradigm of science in materials science. APL Mater. **4**, 053208 (2016)
2. A. Agrawal, A. Choudhary, Deep materials informatics: applications of deep learning in materials science. MRS Commun. **9**, 779–792 (2019)
3. K. Choudhary et al., Recent advances and applications of deep learning methods in materials science. NPJ Comput. Mater. **8**, 59 (2022)
4. K. Choudhary, et al. Large scale benchmark of materials design methods. arXiv preprint arXiv:2306.11688 (2023)
5. S. Kirklin et al., The open quantum materials database (oqmd): assessing the accuracy of dft formation energies. NPJ Comput. Mater. **1**, 15010 (2015)
6. S. Curtarolo et al., AFLOWLIB.ORG: A distributed materials properties repository from high-throughput ab initio calculations. Computational Materials Science **58**, 227–235 (2012) http://linkinghub.elsevier.com/retrieve/pii/S0927025612000687
7. A. Jain et al., The materials project: a materials genome approach to accelerating materials innovation. APL Mater. **1**, 011002 (2013) http://link.aip.org/link/AMPADS/v1/i1/p011002/s1&Agg=doi
8. K. Choudhary, et al. JARVIS: an integrated infrastructure for data-driven materials design (2020). 2007.01831
9. NoMaD. http://nomad-repository.eu/cms/
10. A. Abugabah, A.A. AlZubi, F. Al-Obeidat, A. Alarifi, A. Alwadain, Data mining techniques for analyzing healthcare conditions of urban space-person lung using meta-heuristic optimized neural networks. Clust. Comput. **23**, 1781–1794 (2020)
11. R. Collobert et al., Natural language processing (almost) from scratch. J. Mach. Learn. Res. **12**, 2493–2537 (2011)
12. G. Hinton et al., Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups. IEEE Signal Process. Mag. **29**, 82–97 (2012)
13. T. Ohki, V. Gupta, M. Nishigaki, Efficient spoofing attack detection against unknown sample using end-to-end anomaly detection. In 2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), 224–230 (IEEE, 2019)
14. Z. Jiang, S. Gao, An intelligent recommendation approach for online advertising based on hybrid deep neural network and parallel computing. Clust. Comput. **23**, 1987–2000 (2020)
15. A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, 1097–1105 (2012)
16. I. Goodfellow, Y. Bengio, A. Courville, Y. Bengio, *Deep learning*, vol. 1 (MIT press, Cambridge, 2016)
17. D.P. Tabor et al., Accelerating the discovery of materials for clean energy in the era of smart automation. Nat. Rev. Mater. **3**, 5–20 (2018)

18. K.T. Butler, D.W. Davies, H. Cartwright, O. Isayev, A. Walsh, Machine learning for molecular and materials science. Nature **559**, 547–555 (2018)
19. B. Sanchez-Lengeling, A. Aspuru-Guzik, Inverse molecular design using machine learning: generative models for matter engineering. Science **361**, 360–365 (2018)
20. G. Pilania, Machine learning in materials science: from explainable predictions to autonomous design. Comput. Mater. Sci. **193**, 110360 (2021)
21. D. Morgan, R. Jacobs, Opportunities and challenges for machine learning in materials science. Annu. Rev. Mater. Res. **50**, 71–103 (2020)
22. A. Mannodi-Kanakkithodi, M.K. Chan, Computational data-driven materials discovery. Trends Chem. **3**, 79–82 (2021)
23. P. Friederich, F. Häse, J. Proppe, A. Aspuru-Guzik, Machine-learned potentials for next-generation matter simulations. Nat. Mater. **20**, 750–761 (2021)
24. R. Pollice et al., Data-driven strategies for accelerated materials design. Acc. Chem. Res. **54**, 849–860 (2021)
25. J. Westermayr, M. Gastegger, K.T. Schütt, R.J. Maurer, Perspective on integrating machine learning into computational chemistry and materials science. J. Chem. Phys. **154**, 230903 (2021)
26. D. Jha et al., Enabling deeper learning on big data for materials informatics applications. Sci. Rep. **11**, 1–12 (2021)
27. D. Jha, V. Gupta, W.-K. Liao, A. Choudhary, A. Agrawal, Moving closer to experimental level materials property prediction using AI. Sci. Rep. **12**, 1–9 (2022)
28. V. Gupta et al., Mppredictor: an artificial intelligence-driven web tool for composition-based material property prediction. J. Chem. Inf. Model. **63**, 1865–1871 (2023)
29. C.B. Wahl et al. Machine learning enabled image classification for automated data acquisition in the electron microscope (2023)
30. T. Pereira, M. Abbasi, B. Ribeiro, J.P. Arrais, Diversity oriented deep reinforcement learning for targeted molecule generation. J. Cheminform. **13**, 21 (2021)
31. R. Mercado et al., Graph networks for molecular design. Machine Learn.: Sci. Technol. **2**, 02502025023 (2021)
32. T. Lookman, P.V. Balachandran, D. Xue, R. Yuan, Active learning in materials science with emphasis on adaptive sampling using uncertainties for targeted design. NPJ Comput. Mater. **5**, 21 (2019)
33. C. Kim, A. Chandrasekaran, A. Jha, R. Ramprasad, Active-learning and materials design: the example of high glass transition temperature polymers. Mrs Commun. **9**, 860–866 (2019)
34. C. Zang, F. Wang, Moflow: an invertible flow model for generating molecular graphs. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, 617–626 (2020)
35. M. Sacha et al., Molecule edit graph attention network: modeling chemical reactions as sequences of graph edits. J. Chem. Inf. Model. **61**, 3273–3284 (2021)
36. P.C. Jennings, S. Lysgaard, J.S. Hummelshøj, T. Vegge, T. Bligaard, Genetic algorithms for computational materials discovery accelerated by machine learning. NPJ Comput. Mater. **5**, 46 (2019)
37. C. Kim, R. Batra, L. Chen, H. Tran, R. Ramprasad, Polymer design using genetic algorithm and machine learning. Comput. Mater. Sci. **186**, 110067 (2021)
38. H. Chan et al., Rapid 3d nanoscale coherent imaging via physics-aware deep learning. Appl. Phys. Rev. **8**, 021407 (2021)
39. G.P. Pun, R. Batra, R. Ramprasad, Y. Mishin, Physically informed artificial neural networks for atomistic modeling of materials. Nat. Commun. **10**, 2339 (2019)
40. D. Jha et al., Enhancing materials property prediction by leveraging computational and experimental data using deep transfer learning. Nat. Commun. **10**, 1–12 (2019)
41. V. Gupta et al., Cross-property deep transfer learning framework for enhanced predictive analytics on small materials data. Nat. Commun. **12**, 1–10 (2021)
42. V. Gupta, W.K. Liao, A. Choudhary, A. Agrawal, Pre-activation based representation learning to enhance predictive analytics on small materials data. In 2023 International Joint Conference on Neural Networks (IJCNN), IEEE, 2023 Jun 18, pp. 1–8
43. B. Meredig et al., Combinatorial screening for new materials in unconstrained composition space with machine learning. Phys. Rev. B **89**, 094104 (2014)
44. D. Xue et al., Accelerated search for materials with targeted properties by adaptive design. Nature commun. **7**, 1–9 (2016)
45. F.A. Faber, A. Lindmaa, O.A. Von Lilienfeld, R. Armiento, Machine learning energies of 2 million elpasolite (a b c 2 d 6) crystals. Phys. Rev. Lett. **117**, 135502 (2016)
46. R. Ramprasad, R. Batra, G. Pilania, A. Mannodi-Kanakkithodi, C. Kim, Machine learning in materials informatics: recent applications and prospects. NPJ Comput. Mater. **3**, 54 (2017). https://doi.org/10.1038/s41524-017-0056-5
47. A. Seko, H. Hayashi, K. Nakayama, A. Takahashi, I. Tanaka, Representation of compounds for machine-learning prediction of physical properties. Phys. Rev. B **95**, 144110 (2017)
48. L.E. Peterson, K-nearest neighbor. Scholarpedia **4**, 1883 (2009)
49. M.A. Hearst, S.T. Dumais, E. Osuna, J. Platt, B. Scholkopf, Support vector machines. IEEE Intell. Syst. Their Appl. **13**, 18–28 (1998)
50. A.J. Myles, R.N. Feudale, Y. Liu, N.A. Woody, S.D. Brown, An introduction to decision tree modeling. J. Chemom.: A J. Chemom. Soc. **18**, 275–285 (2004)
51. L. Breiman, Random forests. Mach. learn. **45**, 5–32 (2001)
52. T. Chen, C. Guestrin, Xgboost: a scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, 785–794 (2016)
53. Y. Saad et al., Data mining for materials: computational experiments with a b compounds. Phys. Rev. B **85**, 104104 (2012)
54. K. Fujimura et al., Accelerated materials design of lithium superionic conductors based on first-principles calculations and machine learning algorithms. Adv. Energy Mater. **3**, 980–985 (2013)
55. A. Seko, T. Maekawa, K. Tsuda, I. Tanaka, Machine learning with systematic density-functional theory calculations: application to melting temperatures of single-and binary-component solids. Phys. Rev. B **89**, 054303 (2014)
56. A. Seko et al., Prediction of low-thermal-conductivity compounds with first-principles anharmonic lattice-dynamics calculations and bayesian optimization. Phys. Rev. Lett. **115**, 205901 (2015)
57. J. Lee, A. Seko, K. Shitara, K. Nakayama, I. Tanaka, Prediction model of band gap for inorganic compounds by combination of density functional theory calculations and machine learning techniques. Phys. Rev. B **93**, 115104 (2016)
58. L. Ward, A. Agrawal, A. Choudhary, C. Wolverton, A general-purpose machine learning framework for predicting properties of inorganic materials. NPJ Comput. Mater. **2**, 16028 (2016). https://doi.org/10.1038/npjcompumats.2016.28
59. L.M. Ghiringhelli, J. Vybiral, S.V. Levchenko, C. Draxl, M. Scheffler, Big data of materials science: critical role of the descriptor. Phys. Rev. Lett. **114**, 105503 (2015)
60. F. Faber, A. Lindmaa, O.A. von Lilienfeld, R. Armiento, Crystal structure representations for machine learning models of formation energies. Int. J. Quantum Chem. **115**, 1094–1101 (2015)
61. K. Schütt et al., How to represent crystal structures for machine learning: towards fast prediction of electronic properties. Phys. Rev. B **89**, 205118 (2014)
62. W.D. Callister, D.G. Rethwisch et al., *Materials science and engineering: an introduction*, vol. 7 (Wiley, New York, 2007)
63. L. Ward et al., Including crystal structure attributes in machine learning models of formation energies via Voronoi tessellations. Phys. Rev. B **96**, 024104 (2017)
64. D. Svozil, V. Kvasnicka, J. Pospichal, Introduction to multi-layer feed-forward neural networks. Chemom. Intell. Lab. Syst. **39**, 43–62 (1997)
65. S. Albawi, T.A. Mohammed, S. Al-Zawi, Understanding of a convolutional neural network. In 2017 International conference on engineering and technology (ICET), IEEE, pp. 1–6 (2017)
66. L.R. Medsker, L. Jain, Design and Applications. Recurrent Neural Netw **5**, 64–67 (2001)
67. Q. Zhou et al., Learning atoms for materials discovery. Proc. Natl. Acad. Sci. **115**, E6411–E6417 (2018)

68. G. Montavon et al., Machine learning of molecular electronic properties in chemical compound space. New J. Phys. Focus Issue Novel Mater. Discov. **15**, 095003 (2013)
69. D. Jha et al., ElemNet: deep learning the chemistry of materials from only elemental composition. Sci. Rep. **8**, 17593 (2018)
70. A. Paul et al. CheMixNet: mixed DNN architectures for predicting chemical properties using multiple molecular representations. In Workshop on Molecules and Materials at the 32nd Conference on Neural Information Processing Systems (2018)
71. C. Yang, Y. Kim, S. Ryu, G.X. Gu, Prediction of composite microstructure stress-strain curves using convolutional neural networks. Mater. Des. **189**, 108509 (2020)
72. K. Yang et al., Self-supervised learning and prediction of microstructure evolution with convolutional recurrent neural networks. Patterns **2**, 100243 (2021)
73. D. Jha, et al. IRNet: A general purpose deep residual regression framework for materials discovery. In 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2385–2393 (2019)
74. V. Gupta, W.-k. Liao, A. Choudhary, A. Agrawal, Brnet: Branched residual network for fast and accurate predictive modeling of materials properties. In Proceedings of the 2022 SIAM International Conference on Data Mining (SDM), SIAM, pp. 343–351 (2022)
75. V. Gupta, A. Peltekian, W.-K. Liao, A. Choudhary, A. Agrawal, Improving deep learning model performance under parametric constraints for materials informatics applications. Sci. Rep. **13**, 9128 (2023)
76. Y. Mao et al., An AI-driven microstructure optimization framework for elastic properties of titanium beyond cubic crystal systems. NPJ Comput. Mater. **9**, 111 (2023)
77. E.O. Pyzer-Knapp, K. Li, A. Aspuru-Guzik, Learning from the harvard clean energy project: the use of neural networks to accelerate materials discovery. Adv. Func. Mater. **25**, 6495–6502 (2015)
78. G.B. Goh, N.O. Hodas, C. Siegel, A. Vishnu, SMILES2Vec: An interpretable general-purpose deep neural network for predicting chemical properties. arXiv preprint arXiv:1712.02034 (2017)
79. X. Zheng, P. Zheng, L. Zheng, Y. Zhang, R.-Z. Zhang, Multi-channel convolutional neural networks for materials properties prediction. Comput. Mater. Sci. **173**, 109436 (2020)
80. A.L. Nazarova et al., Dielectric polymer property prediction using recurrent neural networks with optimizations. J. Chem. Inf. Model. **61**, 2175–2186 (2021)
81. T.N. Kipf, M. Welling, Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907 (2016)
82. P. Velickovic et al., Graph attention networks. Stat. **1050**, 10–48550 (2017)
83. W. Hamilton, Z. Ying, J. Leskovec, Inductive representation learning on large graphs. Advances in neural information processing systems **30** (2017)
84. S. Pan, et al. Adversarially regularized graph autoencoder for graph embedding. arXiv preprint arXiv:1802.04407 (2018)
85. R. Ying, et al. Graph convolutional neural networks for web-scale recommender systems. In Proceedings of the 24th ACM SIGKDD International Conference on knowledge discovery & data mining, 974–983 (2018)
86. K.T. Schütt, H.E. Sauceda, P.-J. Kindermans, A. Tkatchenko, K.-R. Müller, Schnet-a deep learning architecture for molecules and materials. J. Chem. Phys. **148**, 241722 (2018)
87. C. Chen, W. Ye, Y. Zuo, C. Zheng, S.P. Ong, Graph networks as a universal machine learning framework for molecules and crystals. Chem. Mater. **31**, 3564–3572 (2019)
88. T. Xie, J.C. Grossman, Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. Phys. Rev. Lett. **120**, 145301 (2018). https://doi.org/10.1103/PhysRevLett.120.145301
89. C.W. Park, C. Wolverton, Developing an improved crystal graph convolutional neural network framework for accelerated materials discovery. Phys. Rev. Materials **4**, 063801 (2020) https://doi.org/10.1103/PhysRevMaterials.4.063801
90. R.E. Goodall, A.A. Lee, Predicting materials properties without crystal structure: Deep representation learning from stoichiometry. arXiv preprint arXiv:1910.00617 (2019)
91. K. Choudhary, B. DeCost, Atomistic line graph neural network for improved materials property predictions. NPJ Comput. Mater. **7**, 1–8 (2021)
92. M. Karamad et al., Orbital graph convolutional neural network for material property prediction. Phys. Rev. Mater. **4**, 093801 (2020)
93. B. Wang, Q. Fan, Y. Yue, Study of crystal properties based on attention mechanism and crystal graph convolutional neural network. J. Phys.: Condens. Matter **34**, 195901 (2022)
94. S.S. Omee et al., Scalable deeper graph neural networks for high-performance materials property prediction. Patterns **3**, 100491 (2022)
95. S.-Y. Louis et al., Graph convolutional neural networks with global attention for improved materials property prediction. Phys. Chem. Chem. Phys. **22**, 18141–18148 (2020)
96. J. Gasteiger, J. Groß, S. Günnemann, Directional message passing for molecular graphs. arXiv preprint arXiv:2003.03123 (2020)
97. J. Gasteiger, S. Giri, J.T. Margraf, S. Günnemann, Fast and uncertainty-aware directional message passing for non-equilibrium molecules. arXiv preprint arXiv:2011.14115 (2020)
98. J. Gasteiger, F. Becker, S. Günnemann, Gemnet: universal directional graph neural networks for molecules. Adv. Neural. Inf. Process. Syst. **34**, 6790–6802 (2021)
99. V.P. Dwivedi, C.K. Joshi, T. Laurent, Y. Bengio, X. Bresson, Benchmarking graph neural networks. arXiv preprint arXiv:2003.00982 (2020)
100. T. Hsu et al., Efficient and interpretable graph network representation for angle-dependent properties applied to optical spectroscopy. NPJ Comput. Mater. **8**, 151 (2022)
101. C. Ying et al., Do transformers really perform badly for graph representation? Adv. Neural. Inf. Process. Syst. **34**, 28877–28888 (2021)
102. W. Hu, et al. Ogb-lsc: A large-scale challenge for machine learning on graphs. arXiv preprint arXiv:2103.09430 (2021)
103. L. Chanussot et al., Open catalyst 2020 (oc20) dataset and community challenges. ACS Catal. **11**, 6059–6072 (2021)
104. S. Banik et al., Cegann: crystal edge graph attention neural network for multiscale classification of materials environment. NPJ Comput. Mater. **9**, 23 (2023)
105. Y. Li, C. Gu, T. Dullien, O. Vinyals, P. Kohli, Graph matching networks for learning the similarity of graph structured objects, in *International Conference on Machine Learning* (PMLR, 2019), pp. 3835–3845
106. X. Ling et al., Multilevel graph matching networks for deep graph similarity learning. IEEE Transact. Neural Netw. Learn. Syst. (2021). https://doi.org/10.1109/TNNLS.2021.3102234
107. Z. Zhang, et al., H2MN: graph similarity learning with hierarchical hypergraph matching networks, in *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining* (2021), pp. 2274–2284
108. R. Wang, J. Yan, X. Yang, Neural graph matching network: learning lawler's quadratic assignment problem with extension to hypergraph and multiple-graph matching. IEEE Trans. Pattern Anal. Mach. Intell. **44**, 5261–5279 (2021)
109. M. Soldan, M. Xu, S. Qu, J. Tegner, B. Ghanem, VLG-Net: video-language graph matching network for video grounding, in *Proceedings of the IEEE/CVF International Conference on Computer Vision*(2021), pp. 3224–3234
110. B. Meredig et al., Can machine learning identify the next high-temperature superconductor? Examining extrapolation performance for materials discovery. Mol. Syst. Des. Eng. **3**, 819–825 (2018)
111. P. Friederich, M. Krenn, I. Tamblyn, A. Aspuru-Guzik, Scientific intuition inspired by machine learning-generated hypotheses. Mach. Learn.: Sci. Technol. **2**, 025027 (2021)
112. F. Oviedo, J.L. Ferres, T. Buonassisi, K.T. Butler, Interpretable and explainable machine learning for materials science and chemistry. Acc. Mater. Res. **3**, 597–607 (2022)
113. V. Korolev, I. Nevolin, P. Protsenko, A universal similarity based approach for predictive uncertainty quantification in materials science. Sci. Rep. **12**, 14931 (2022)