Progress in Energy

ACCEPTED MANUSCRIPT • OPEN ACCESS

A review on machine learning-guided design of energy materials

To cite this article before publication: Seongmin Kim *et al* 2024 *Prog. Energy* in press https://doi.org/10.1088/2516-1083/ad7220

Title: A review on machine learning-guided design of energy materials

Authors: Seongmin Kim[1,3], Jiaxin Xu[1], Wenjie Shang[1], Zhihao Xu[1], Eungkyu Lee[2,*], and Tengfei Luo[1,*]

[1]Department of Aerospace and Mechanical Engineering, University of Notre Dame; Notre Dame, Indiana 46556, United States.
[2]Department of Electronic Engineering, Kyung Hee University; Yongin-Si, Gyeonggi-do 17104, Republic of Korea.
[3]National Center for Computational Sciences, Oak Ridge National Laboratory, Oak Ridge, Tennessee 37830, United States.
*Correspondence: eleest@khu.ac.kr (E. Lee), tluo@nd.edu (T. Luo)

Abstract

The development and design of energy materials are essential for improving the efficiency, sustainability, and durability of energy systems to address climate change issues. However, optimizing and developing energy materials can be challenging due to large and complex search spaces. With the advancements in computational power and algorithms over the past decade, machine learning (ML) techniques are being widely applied in various industrial and research areas for different purposes. The energy material community has increasingly leveraged ML to accelerate property predictions and design processes. This article aims to provide a comprehensive review of research in different energy material fields that employ ML techniques. It begins with foundational concepts and a broad overview of ML applications in energy material research, followed by examples of successful ML applications in energy material design. We also discuss the current challenges of ML in energy material design and our perspectives. Our viewpoint is that ML will be an integral component of energy materials research, but data scarcity, lack of tailored ML algorithms, and challenges in experimentally realizing ML-predicted candidates are major barriers that still need to be overcome.

Keywords: machine learning, energy material, optimization, material design, property prediction

1. Introduction

With challenges brought by climate change and the need for decarbonization, there are significant efforts globally to cut down reliance on conventional energy [1]. International commitments (e.g., the 2016 Paris Accord) exemplify this effort, where countries worldwide are coming together to address these global issues [2-5]. Energy materials are substances or materials that generate, release, convert, or store energy, which can be used in applications like energy storage devices, energy conversion systems, and energy generators. For instance, any materials used in batteries, conductors, photovoltaics, thermoelectric, fuel cells, and hydrogen production are considered energy materials. Such materials are indispensably used in our modern lives, but they potentially contribute to global warming by emitting or producing environmental-damaging materials or CO_2 during their operations or fabrications [6-10]. In response to these challenges, the ongoing evolution and development of energy materials over the past few decades have significantly enhanced their energy conversion efficiency, resulting in less dependence on fossil fuels and their derivatives [11-15]. Therefore, designing and optimizing energy materials become an important part of addressing global environmental issues [16].

The performance of energy materials is dependent on many design factors, such as geometrical features, composition, processing conditions, and environmental factors, leading to large design spaces, which means that there are numerous possible configurations for their optimization [17-19]. Conducting experiments to comprehensively search these large design spaces for finding optimal material states is usually too costly and time-consuming. Hence, researchers have been using simulation tools, such as numerical methods, first-principles calculations, and atomistic simulations, to design materials and calculate their properties [20-27]. Nevertheless, exploring such large spaces with different design parameters using conventional simulation methods can still lead to high computational costs and time. Furthermore, these simulation methods rely on high-fidelity models to accurately mimic the dynamics of materials [28]. However, models constructed for the simulations may not fully capture the complexity of real systems, and simulations may be difficult or impossible in certain fields where established theories are lacking or physical models are too complicated [29-32].

Data-driven approaches, especially machine learning (ML) [24, 33-37], can establish efficient surrogate models, which describe design spaces by approximating the relationship between material states and their performance, [38-45] by learning hidden patterns with data. These surrogate models can be used to predict material properties for given material features (e.g., chemistry, composition, and geometry), and can be leveraged to help design materials with desirable performance (figure 1) [46-48]. Over the past decade, ML algorithms have been actively explored to accelerate material designs. For example, Wan et al. identified optimal electrode structures for redox flow batteries using a framework that couples an ML regression model with a genetic algorithm for multi-objective optimization [49]. Dave et al. [50] used an experimental design scheme that includes Bayesian optimization and robotics to optimize non-aqueous Li-ion battery electrolytes. Li et al. [51] designed high-performance perovskite solar cells using ML techniques (e.g., artificial neural networks) with data collected from the literature. Wu et al. [52] used ML algorithms (linear regression, multinomial logistic regression and boosted regression trees) to accelerate discovering donor/acceptor combinations for high-performance organic solar cell applications. Feng et al. [53] used ML techniques (random forest, support vector machine and neural networks) to design polymer nanocomposites for energy storage applications. These examples demonstrate the potential of ML in designing high-performance energy materials.

In this article, we provide a comprehensive overview of research in energy fields using ML techniques. First, we introduce basic knowledge of ML, including commonly used ML-based design algorithms, aiming to inspire the community to consider applying ML techniques in their material design works. Next, we survey recent successful examples of using ML algorithms in different energy material fields, demonstrating the potential of ML techniques in high-performance energy material design. At last, we close the review by discussing the current challenges of ML and our perspectives.

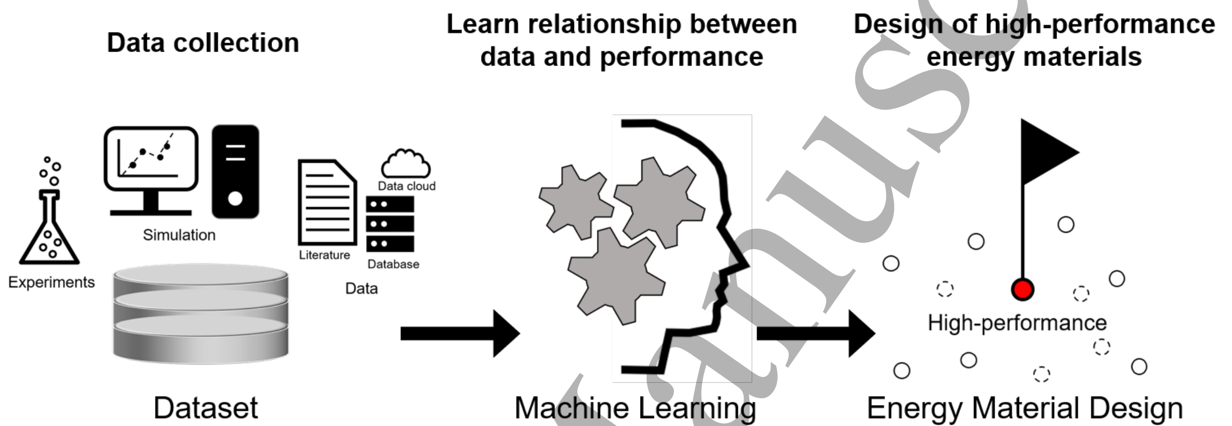

Figure 1. The schematic of a typical workflow to design high-performance energy materials using ML. Researchers need to prepare datasets to train ML models that can learn the underlying knowledge in data. The trained ML models can be used to identify high-performance energy materials in large design spaces with the help of optimization algorithms.

2. Introduction to ML

2.1. ML techniques

ML, which is a subset of artificial intelligence, aims at learning knowledge with data and algorithms to emulate the human learning process, steadily enhancing its accuracy. ML algorithms generate surrogates through training processes to make predictions without explicit physics-based simulations or calculations. Generally, ML algorithms require data to learn knowledge, but physics-informed ML can leverage both data and physical principles, which can be beneficial when collecting data is difficult and expensive [54-58]. With the enhanced surrogate prediction capability, handling a large number of material candidates becomes possible, allowing us to design energy materials with complex characteristics [59-61]. ML may be divided into three categories: supervised learning, unsupervised learning, and semi-supervised learning (figure 2) [60, 62].

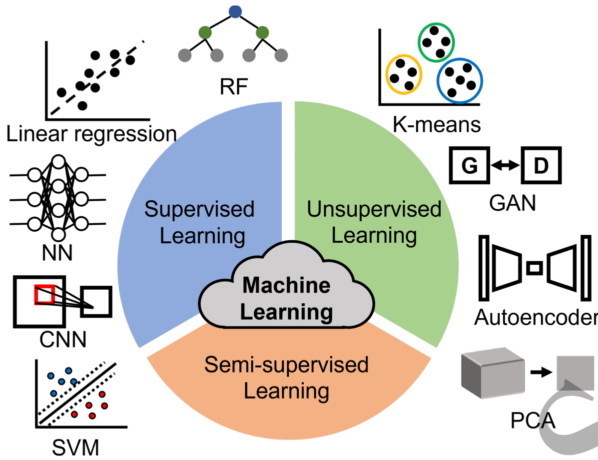

Figure 2. Three main categories of ML algorithms (supervised learning, unsupervised learning, and semi-supervised learning). Examples of supervised learning: random forest (RF), linear regression, neural network (NN), convolutional neural network (CNN), and support vector machine (SVM). Examples of unsupervised learning: k-means clustering, generative adversarial network (GAN), autoencoder, and principal component analysis (PCA).

2.1.1. Supervised learning

Supervised learning algorithms are trained with labeled data, where each piece of data is paired with a known output value, which allows the algorithms to learn the correlation between inputs and their corresponding outputs. The supervised ML models are usually used as surrogates to efficiently calculate the output values of new, unseen input data without the need to perform expensive experiments or physics-based simulations. Such models have been seen in a wide range of applications, such as image recognition, natural language processing, material designs, property prediction, and fraud detection [60]. Some examples of widely used supervised learning algorithms are RF, linear regression, NN, CNN, and SVM. In the materials design domain, such supervised ML models are commonly used to describe the structure-property relationship to quickly evaluate new materials.

Some ML models, such as decision trees and linear regression, are transparent, interpretable, and explainable, offering clear insights into their decision-making processes. For example, Weng et al. [63] used ML regression models to discover new perovskite catalysts that have enhanced oxygen evolution reaction activities, which play important roles in renewable energy production and storage. They used a symbolic regression model to identify a key material descriptor, which enabled them to predict the oxygen evolution reaction activities and discover new catalysts. However, for some complex ML models, the rationale behind the outputs is not readily interpretable and explainable, making such models a "black box". Despite their non-transparent properties, black box models remain highly useful for predicting labels once properly trained. Many complex ML models, such as NN, can be considered black-box models, and they are used for property predictions, material designs, classifications, and recognitions [64, 65]. Strategy like

SHapley Additive exPlanations (SHAP) values, which provide interpretable means to understand the importance of features, can be used as post-analysis to interpret and explain the predictions made by black-box models. Fu et al. [66] employed the SHAP analysis to extract synthetic parameters of catalysts by interpreting the impact of the descriptors of the trained ML model (e.g., k-nearest neighbors, eXtreme gradient boosting, and adaptive boosting).

2.1.2. Unsupervised learning

Unsupervised learning algorithms learn knowledge from unlabeled data that does not have explicit output value. These algorithms discover hidden patterns, structures, or relationships within the given dataset, enabling clustering of similar data points or simplification of datasets to reveal their inherent structures. These ML models are generally used for data exploration, pattern recognition, and feature extraction [62]. Examples of unsupervised learning algorithms are K-means clustering, generative adversarial network (GAN), autoencoder-decoder, and principal component analysis (PCA). Unsupervised learning has also been used in studying energy materials. Liu et al. [67] used an unsupervised classification model to classify whether a given compound has a phonon band gap before conducting transfer learning. Jia et. al. [68] designed high-performing thermoelectric materials by grouping half-Heusler compounds using an iterative unsupervised learning algorithm. Unsupervised learning, however, lacks the ability to predict properties, although it can sometimes be combined with supervised learning to narrow down the candidate space [67].

2.1.3. Semi-supervised learning

Annotating properties for various energy materials can prove to be costly and time-consuming, leading to limitations in collecting sufficient labeled training data for accurate screening. This is especially true for many materials used in energy applications. For example, designing polymers, characterized by their high complexity, remains challenging due to limited datasets. This data insufficiency in energy materials is usually in contrast to other domains where ML has been more active and effective. For instance, datasets such as PubChem [69] and the Open Quantum Materials Database (OQMD) [70] boast large volumes (~million scale) for drug discovery and inorganic compounds, respectively, but polymers suffer from notable data sparsity (~hundred to thousand scale) [71, 72]. This substantial difference in data size poses a significant hurdle for training generalizable ML models. Moreover, properties of interest, such as gas permeabilities of polymeric membranes, are often observed less frequently above satisfactory performance thresholds [72], creating an imbalanced nature in data labels. This imbalance often leads to a false-negative problem in virtual screening, potentially biasing ML models toward materials of lower interest and causing researchers to overlook promising candidates for targeted performance. To address the challenges, semi-supervised learning becomes a promising approach [73], especially given the expense of producing labeled data for energy materials. Semi-supervised learning deals with situations where there are few labeled training data but a large number of unlabeled data, which aligns with the constraints of annotating energy materials. We categorize semi-supervised learning methods into data-centric and model-centric methods. Data-centric methods focus on improving data quantity and quality, while model-centric methods refine the learning of model parameters.

A notable data-centric method is pseudo-labeling [74], a semi-supervised learning approach that assigns pseudo-labels to unlabeled data and incorporates them into the labeled training set. Liu et al. [75] utilized pseudo-labeling in a semi-supervised graph imbalanced regression (SGIR) framework to address sparsity and imbalance issues in polymer permeability data by utilizing the large unlabeled polymer dataset to augment the limited labeled training data. SGIR achieved significant prediction error reduction compared to the conventional vanilla graph neural network (GNN). Challenges in pseudo-labeling include defining confidence scores and improving uncertainty estimation. Future work may explore integrating active learning as a complementary approach and developing sampling strategies for pseudo-labels to balance imbalanced label distributions.

In model-centric methods, self-supervised learning for example, involves fine-tuning learned data representations from unlabeled data with a labeled dataset to solve supervised learning problems [76]. Self-supervised learning transfers knowledge from unlabeled data to labeled data through model parameters. Methods for self-supervised representation learning include predictive tasks and contrastive tasks on unlabeled data, such as masked atom attribute prediction and masked subgraph prediction in graph ML for polymers. Kuenneth et al. [77] introduced polyBERT, a polymer embedding tool inspired by natural language processing concepts, trained through predictive self-supervised learning. The polyBERT model outperformed existing fingerprint schemes in terms of speed and accuracy. However, self-supervised learning methods encounter challenges in cross-domain knowledge transfer, mainly due to differences between unlabeled and labeled data and between self-supervised learning tasks and downstream tasks. Effective leverage of recent self-supervised learning advancements for energy material screening requires specific, larger-scale, high-quality datasets and self-supervised learning tasks relevant to material properties, along with careful examination of potential model bias in labeled datasets.

Over the past decade, these ML techniques have seen increasing use in designing materials and predicting their properties. To statistically analyze trends in ML application within the materials field, we extracted the number of relevant publications from the Web of Science using specific keywords in the 'Topic' search term. The keywords include 'Material', 'Design', 'Property prediction', 'Machine learning', 'Supervised learning', 'Unsupervised learning', and 'Semisupervised (or semi-supervised) learning'. We opted for the keyword 'Material' instead of 'Energy material' to avoid overly narrowing the search index, as many researchers use the broader term. Figures 3a and 3b illustrate the growing number of publications applying ML to material design and property prediction, indicating an active adoption of ML techniques in material research fields.

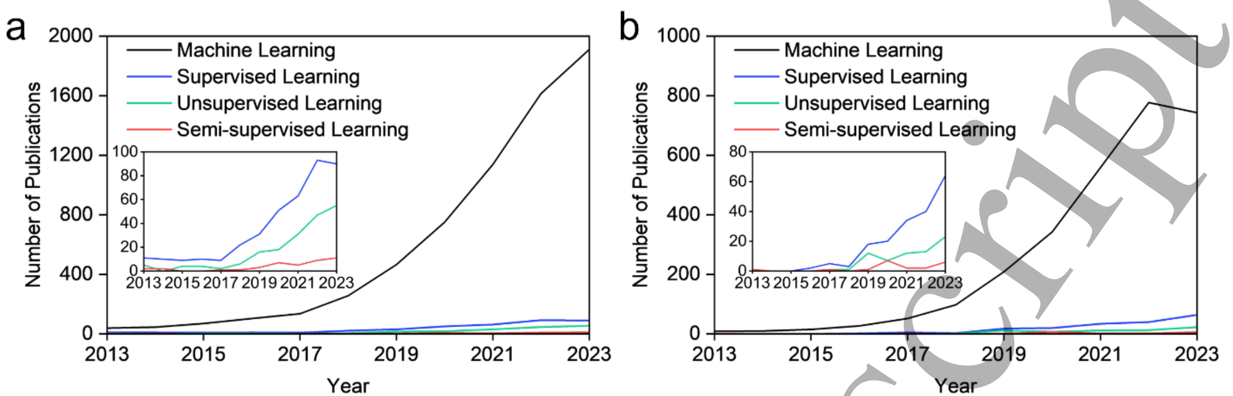

Figure 3. Annual number of publications in the research field of ML for material science. Keywords include 'Material', (a) 'Design', or (b) 'Property prediction', and those in the figure legend.

2.2. ML-facilitated material optimization and inverse design

The forward inferences of ML models can be used to predict the properties of candidate materials using surrogates. However, in many cases, it is required to optimize or inversely design new materials with desirable target properties. Therefore, ML models are also used with different optimization schemes to optimize or design new materials.

Inverse design refers to the process of identifying material structures or compositions that exhibit desired properties or performance characteristics. In traditional design processes, researchers iteratively design and test until they achieve their goals, which might take a long time. In contrast, the inverse design starts with desired outputs (i.e., characteristics, functionalities, or properties), and then works backward to determine the optimal structures that satisfy the predefined objectives (figure 4(a)). In energy materials, ML techniques can be beneficial to constructing reliable inverse design models using various optimization techniques, such as genetic algorithms, Bayesian optimization, and reinforcement learning. These methods explore the vast design space efficiently, guiding the search towards optimal solutions that meet specific property requirements.

Various design models have been used to integrate with ML algorithms, such as active learning, inverse design, and black box models [64]. Collecting a lot of training data to build solid models by training ML algorithms can be costly and challenging, and a lack of training data often leads to suboptimal predictions or classifications. These challenges (i.e., sparsity and imbalance issues in the dataset) generally come from the limited availability of experimental/computational data (compared to the oftentimes large design space). The disproportionate representation of different classes or ranges of values can lead to biased models, resulting in inaccurate predictions. To overcome this challenge, ML-aided active learning algorithms have gained popularity in materials design and optimization. These active learning algorithms iteratively select the most informative samples during an optimization cycle. Hence, the algorithms gradually update their models by selectively incorporating informative data points labeled by an oracle, which is an entity that provides expertise in labeling or evaluating data. The updated dataset is used for the next iteration, guiding further data collection. Active learning enables the iterative improvement of the model's performance with a minimal number of training data; thus, it can reduce optimization costs. Hence, active learning is widely used for the purpose of optimal designs, such as material design and system optimizations [78-81].

Force fields are mathematical models used to estimate the potential energy of a system of atoms or molecules, essential for molecular dynamics simulations and materials modeling [82]. Developing accurate force fields involves parameterizing the model to capture the interactions between atoms accurately [83]. ML techniques have been increasingly applied to force field development, where models are trained on high-quality data from quantum mechanical calculations (figure 4(b)) [84, 85]. This approach enhances the accuracy and transferability of force fields, enabling more reliable simulations of complex material systems [84, 85].

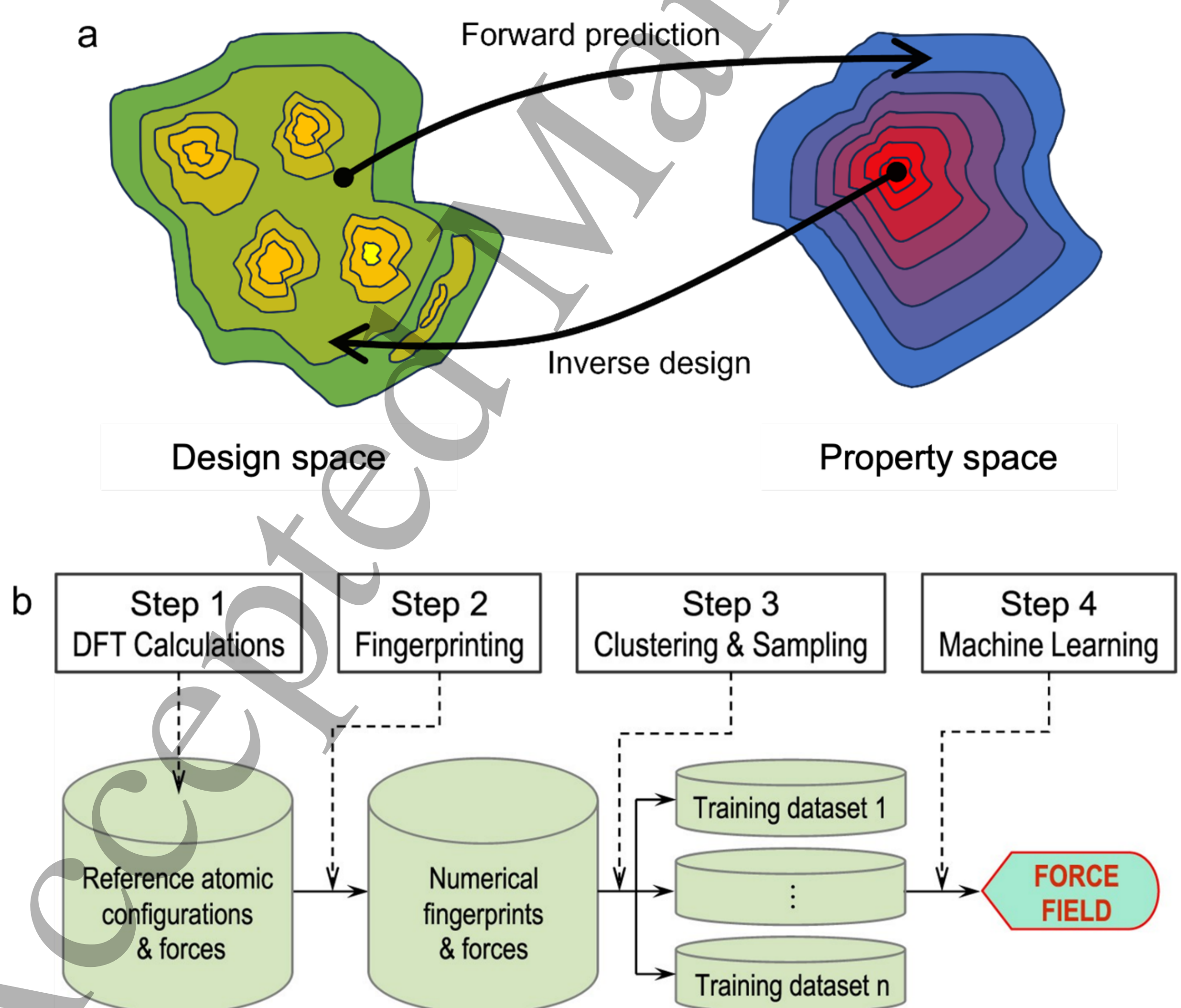

Figure 4. (a) The schematic of inverse design, where inverse design starts with the desired properties to find the optimal design. (b) The schematic of ML force field, reproduced from [84]. CC BY-NC-ND 4.0.

2.3. Data preparation

Data preparation is an important step for ML [86]. Training data used for ML can be collected from experiments, reported results, computations, and databases. Using reported data can minimize costs for generating training data, but it is essential to consider many factors besides the target property of interest in material designs (i.e., experimental conditions, measurement techniques, or design baseline). There often can be large deviations between data from different literature even for the same material. Hence, researchers are increasingly using computational simulations where users can have more control of the data production procedure. Although computations are usually more efficient than experiments, they can still be time-consuming. To address this limitation, researchers have shared data from their experiments and computations in publicly accessible databases, aiming to assist other users with their ML tasks. This is becoming more common with many journals mandating data sharing. However, these data usually have different formats and are not easy to mass download. There are some databases that are for general use or more specialized (e.g., for gas permeability) for material designs. These include: Materials Project (MP), Open Quantum Materials Database (OQMD), Materials Cloud, National Renewable Energy Laboratory Materials, Inorganic Crystal Structure Database (ICSD), superconducting critical temperatures (SuperCon), Harvard Clean Energy Project (HCEP), Materials Commons, Cambridge Structural Databases, Materials Data Facility, Nano-HUB, Pearson Crystal Data, AiiDA, novel materials discovery (NOMAD), AFLOWLIB, computational materials repository, Crystal Open Database, PubChem, Protein Data Bank (PDB), CRYSTMET, Fireworks, PoLyInfo, and MatWeb [29, 87-91].

As ML techniques have been more frequently applied in material science, the importance of data preparation has increased. To statistically analyze trends in data preparation, we retrieved the number of publications from the Web of Science using the keywords ‘Material’, ‘Machine learning’, ‘Experiments’, ‘Simulation’, ‘Database’, ‘Materials Project’, ‘Inorganic crystal structure database’, and ‘Materials Commons’. Figure 5(a) shows that ‘Experiments’, ‘Simulation’, and ‘Database’ have been increasingly utilized to prepare data, highlighting an increasing use of representative databases in figure 5(b).

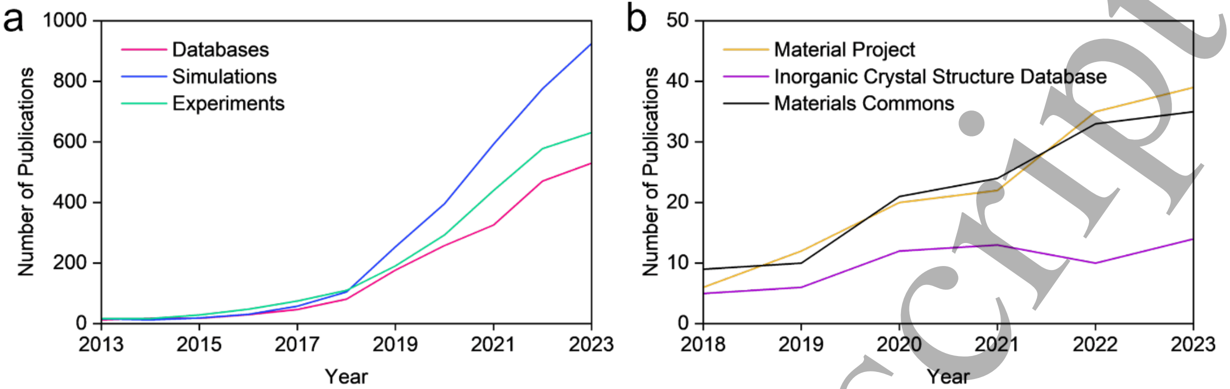

Figure 5. Annual number of publications to prepare data for ML in the material field. Keywords include (a) ‘Material’, ‘Machine learning’, and (b) ‘Material’, ‘Machine learning’, ‘Database’, and those in the figure legend.

Both data quality and quantity are critical to the performance of the trained ML models. Although these databases can support the training of many good ML models, there may be a lack of specific properties of particular interest to certain users. Hence, additional data may be required to further improve the quality and quantity of training data for these cases. If it is challenging to collect a large number of training data because of difficulties in experiments or computations, data augmentation strategies may be applied, which however are more popular for image data [90, 92, 93]. Recently for graph-type data, which can be described by graphs such as molecules [94], polymers [95] and crystals [96], techniques like node feature masking, edge dropping, and subgraph replacement are also emerging for data augmentation [76, 97, 98].

2.4. Training and evaluating ML models

With the data prepared, ML models of choice can be trained. Available datasets are usually split in a certain ratio into training, validation, and testing sets. Training stays largely as an art, which involves experience in hyperparameters (e.g., epoch, batch size, learning rate, momentum, cost function, hidden unit, regularization parameter and iteration) tuning using different techniques (e.g., grid search, random search, or advanced optimization methods) to optimize the model quality [99]. After training, the built models are usually evaluated using a validation set to ensure performance by mitigating underfitting or overfitting problems. Here, hyperparameters can be finely adjusted to further enhance the model performance. Afterward, a test set is employed to test the ML model’s accuracy, estimating the performance of the trained ML model with new and unseen data. The performance can be evaluated by comparing known values with predicted results from the ML model. Several metrics are used to measure the accuracy of the ML models, for instance, accuracy, receiver operating characteristic - area under the curve (ROC-AUC), root mean squared error (RMSE), mean absolute error (MAE), and coefficient of determination (R^2) [100]. Typically, accuracy and ROC-AUC are used for classification tasks:

$$\text{Accuracy} = C/N \tag{1}$$

where C is the number of correct predictions and N is the total number of predictions. The ROC is a graphical curve that illustrates the performance of a classification model by plotting true positive rate against false positive rate at classification threshold settings. The AUC quantifies the two-dimensional area under the ROC curve, serving as an indicator of the model performance.

On the other hand, MAE, RMSE, and R^2 are widely used to evaluate the performance of regression models.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\frac{|\hat{y}_i - y_i|}{y_i} \quad (2)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2} \quad (3)$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{n}(\bar{y} - y_i)^2} \quad (4)$$

where n is the total number of data, y_i presents true value for i^{th} data point, $\hat{y}_i$ presents the predicted value for i^{th} data point, and $\bar{y}$ represents the mean of true values. Lower values for MAE and RMSE (closer to 0) are preferable, indicating better performance of ML models. In contrast, a higher R^2 score (closer to 1) indicates that the ML model fits well.

2.5. ML-aided design models used for energy materials

In this section, we highlight three optimization algorithms that have been used for energy material optimization and design.

2.5.1. Neural network

NNs also known as artificial neural networks (ANNs), are a class of ML algorithms inspired by the structure and functioning of organismic neural networks. The basic unit of ANNs is the artificial neuron and information flows through the network as the weights of connections between neurons are adjusted during a training process. NN can have various architectures and can be generally classified into several categories, which are multi-layer perceptron (MLP), convolutional neural networks (CNNs), recurrent neural networks (RNNs), graph neural networks (GNNs), and attention-based network networks.

In the domain of energy material studies, MLP stands out as a prevalent NN structure, due to the simplicity of the model structure and limited dataset sizes of energy materials. MLP is constructed from perceptron, which is the basic unit that processes the weighted sum of inputs through a chosen activation function to generate an output. Comprising an input layer, one or more hidden layers, and an output layer, the MLP's interconnected neurons allow for customization in terms of the number of hidden layers and neurons, with the activation function determining the linearity or nonlinearity of its operations. Common nonlinear activation functions, such as Sigmoid, Hyperbolic Tangent, and Rectified Linear Unit (ReLU), are widely used, enabling the model's universality [101, 102]. Various loss functions, including cross-entropy (for classification task) and RMSE (for regression task), are used to quantify the disparities between predictions and actual values [103]. The optimization of MLP weights regarding the loss function utilizes various techniques, with gradient descent recognized for its stability and efficiency [104].

Deep neural networks (DNNs) are multi-layer MLPs capable of learning intricate data representations through various levels of abstraction [105]. DNNs have demonstrated diverse capabilities in various domains and can be generally categorized into CNNs for grid-like data, RNNs for sequential information, GNNs for graph-like structures, and attention-based networks for the selective focus on different parts of the data. CNNs utilize convolutional and pooling layers to automatically extract hierarchical features from grid-like data, commonly applied in image-related tasks like classification and recognition [106]. RNNs are designed for processing data points sequentially related across time or space. It incorporates information from previous time steps to capture temporal dependencies. This makes RNNs suitable for handling time-dependent phenomena as well as text-based data [107]. GNNs specialize in analyzing graph-like data by considering the inherent structural relationships between nodes and edges, frequently employed in chemistry, biology, and social network analysis [108]. For example, graph data can represent molecules' structural information where atoms are nodes and bonds are edges, providing a natural and intuitive way to model the complex relationships in molecular and crystalline structures. Here, graph data allows for the identification of functional groups, the detection of cycles and rings, and the analysis of molecular stability and reactivity, showing better predictive performance than traditional fingerprinting methods. Furthermore, in crystalline structures, GNNs help model and predict properties such as conductivity, thermal stability, and heat capacity [109-111]. GNNs generally operate by iteratively updating the representation of each node based on its neighbors' features and the edges connecting them (figure 6). This process allows the network to learn complex interactions within the material structures, making it suitable for predicting the properties and behaviors of materials. Attention-based networks introduce a dynamic and adaptive mechanism that sets them apart from other DNN architectures. Unlike conventional models that process the entire input uniformly, attention-based networks selectively focus on specific elements of the input, assigning varying levels of importance based on their relevance to the task [112]. This makes attention-based networks particularly powerful in scenarios where nuanced attention and context-aware processing are crucial, such as machine translation, sentiment analysis, image captioning, and material science [113-115].

In general, NN excels in capturing intricate patterns in data, making them well-suited for predicting complex material properties and optimizing material structures. They can automatically learn relevant features from the input data, eliminating the need for manual feature engineering. This is advantageous when dealing with high-dimensional and unstructured materials data, thus it has been increasingly utilized in energy material research. For example, Li et al. [116] designed battery thermal management systems using ANN models. Kaya et al. [117] optimized ultra-thin organic solar cells using an NN-based surrogate model. These examples show that NN is useful for energy material design. However, NN also has limitations, for example, it usually requires large amounts of labeled data for training mainly because of the complexity of the model structure, and the quality of predictions heavily depends on the diversity and representativeness of the training dataset. Moreover, the complex, non-linear nature of NN often results in models that are challenging to interpret.

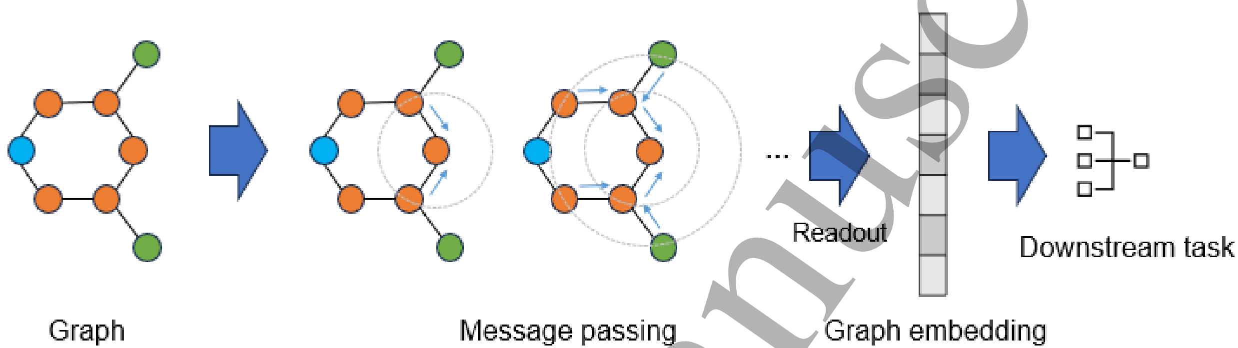

Figure 6. The schematic of GNNs, illustrating how to update the representation of nodes based on the features of its neighboring nodes and the edges connecting them.

2.5.2. Genetic algorithm

Genetic algorithms (GAs) are stochastic search techniques inspired by evolutionary biology, encapsulating procedures such as inheritance, mutation, selection, and crossover to explore the broad regions of the solution space and avoid local minima. After determining the fitness values for all chromosomes, the algorithm selects two elite chromosomes, which exhibit the highest fitness values. These are then subjected to a single-point crossover operation, executed with a crossover probability, to produce offspring. This newly formed offspring subsequently undergoes a uniform mutation, with a mutation probability, resulting in the creation of a modified offspring, which is then incorporated into the new population. The entire process, encompassing selection, crossover, and mutation, is methodically repeated for the current population until the composition of the new population is fully realized. Chromosomes in GAs for energy material design are the objectives in the GA evolutions, which represent key parameters of material structures, such as atomic composition, crystal structure, or structural configuration. The encoding of material structural features usually involves transforming the parameters into a genetic format (i.e., binary encoding, or integer encoding). This encoding process ensures that GAs can effectively manipulate and optimize the material structures through mutation, crossover, and selection. At the end of the optimization, the optimized chromosomes are decoded into corresponding material structures, providing a pathway to discover materials with enhanced energy-related properties.

Benefiting from the outstanding performance in problem domains characterized by complex fitness landscapes, GAs have been widely applied for design problems, which also include the designing of energy materials. Mayer et al. [118] employed GAs to optimize the geometric parameters of flat-plate solar thermal collectors, which led to the maximized solar absorption rate and minimized thermal emissivity with a much lower computational cost. The adaptability of GAs was also shown by Lin et al., who utilized GAs for optimizations of random diffraction gratings in thin-film solar cells [119]. Their findings enhanced the light coupling and trapping effects for a broad range of the solar spectrum, where a 29% improvement over flat cells and 9% improvement over the best periodic gratings were observed. With the development of computational science, researchers have explored the integration of GAs with other advanced techniques to facilitate material design. Patra et al. introduced a novel approach combining NN with GAs [120]. This strategy harnessed the learning capability of NN to guide the evolutionary search of GAs, leading to accelerated material discovery by allowing the algorithms to search as well as learn from the search process. Such a combination was later widely applied to design high-temperature energy capacitors [121], desiccant cooling systems [122], and multilayer microwave radar absorbing material [123]. Zhou et al. [124] developed a molecular-dynamics (MD) based GA to design polyethylene–polypropylene copolymers with high thermal conductivity, indicating the potential of the MD-GA computational framework for accelerating the design of co-polymeric materials. A noteworthy contribution to this domain was the development of the GAMaterial software [125]. This software provides a convenient platform for researchers to apply GAs for material design and discovery.

Generally, GAs are prized for their robustness and ability to handle complex, nonlinear problems, but they also have limitations. Binary representations can lead to intractable string lengths and precision issues, while continuous problems may require specialized crossover and mutation operators to maintain genetic diversity. Moreover, the risk of converging to local optima and the computational cost of simulating many generations can be significant, especially for high-dimensional problems where the time complexity can become prohibitively high.

2.5.3. Bayesian optimization

Gradient-based optimization strategies, suitable for continuous variables and smooth landscapes, can be ineffective in cases involving discrete variables. This is a prevalent issue in material science, where aspects like chemical composition, processing methods, and structural configurations are inherently discrete or categorical. In this context, Bayesian optimization (BO) emerges as a robust and efficient method for navigating these complex and multidimensional spaces. BO is considered a non-derivative algorithm, which uses mechanisms (Bayes' theorem) rather than relying on gradient information to explore solution spaces. Non-derivative algorithms are particularly advantageous for objective functions that are discontinuous, noisy, or have multiple local minima, where gradient information is either unavailable or unreliable. BO, which is a non-derivate and iterative algorithm, uses Bayes' theorem to formulate the parametric space, and employs an acquisition function (e.g., expected improvement) to estimate the best input parameters for the next optimization cycles [126]. The process begins with defining objective functions and decision variables, followed by initiating preliminary experiments using space-filling samples like Latin hypercube designs. The core of BO is updating a Gaussian process (GP) surrogate model, $f(x) \sim GP(m(x), k(x, x'))$, with experimental [127] or computational data [128], which then informs the optimization of an acquisition function, such as Expected Improvement (EI), for selecting the next sampling point. This iterative method continues with experiments and data enhancement until achieving objectives or resource depletion. BO hinges on a probabilistic surrogate model and an acquisition function [129], where the surrogate model encapsulates initial beliefs about an unknown function and data generation, evolving through iterative queries into a more informative posterior. This approach efficiently navigates the multidimensional design spaces (see figure 7 as an example).

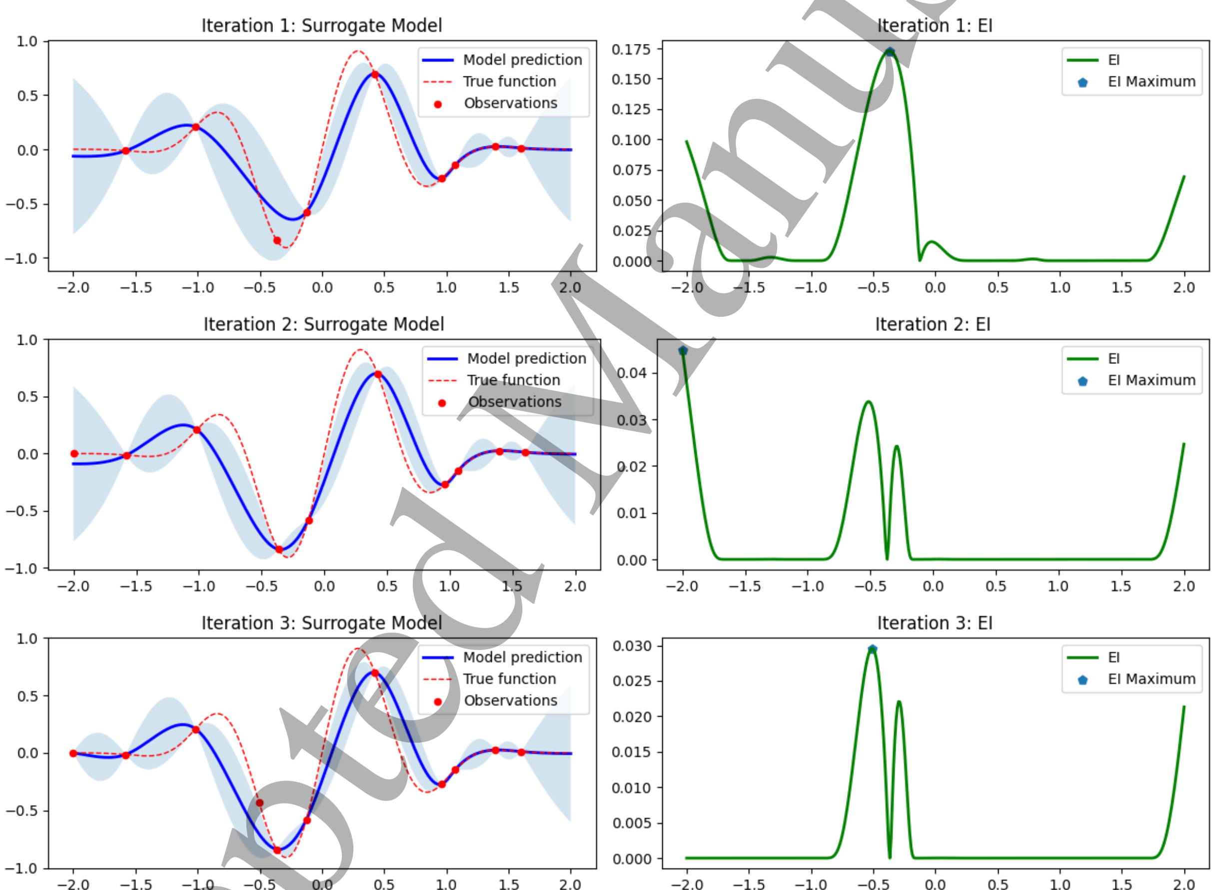

Figure 7. Illustration of BO. BO with EI acquisition function is applied to minimize the test problem $f(x) = sin(5x) * (1 - tanh(x)^2)$ over three iterations. The left column of the plots illustrates the mean and confidence intervals as predicted by the GP model for the objective function. While these plots also display the actual objective function, it is important to note that this function is typically unknown in real-world scenarios. In the right column, the acquisition functions are depicted as green curves. These functions attain high values in regions where the model anticipates a high objective function value, indicating opportunities for exploitation. It is noteworthy that the far-left region remains unexplored in the sampling. This is because, despite its

high uncertainty, the model accurately forecasts minimal improvement in this area compared to the highest observed value so far.

In recent years, BO has emerged as a pivotal tool in the field of energy materials, revolutionizing the way researchers approach optimization and discovery [130, 131]. Shang et al. [127] employed Bayesian Optimization with a hybrid dataset of literature-reported and experimental data to enhance the power factor of AgSe-based thermoelectric materials, achieving double the power factor with approximately ten experimental iterations. Saeidi-Javash and colleagues [132] applied BO to optimize flash sintering parameters for silver-selenide thermoelectric films, considering both continuous variables like voltage and pulse duration, and discrete variables like the number of pulses. Zhang et al. [133] integrated a latent variable GP model with BO, tackling both qualitative and quantitative variables in material design. This approach enhanced optimization in complex material design challenges, such as Hybrid Organic-Inorganic Perovskite design. Each of these studies underscores the diverse and potent applications of BO in energy material science.

These representative design models have been widely employed in material research. Figure 8 shows the growing trend of utilizing these models in material design. Notably, NN has seen rapid growth in use in recent years due to enhanced computational power, which enables the effective handling of large datasets for training.

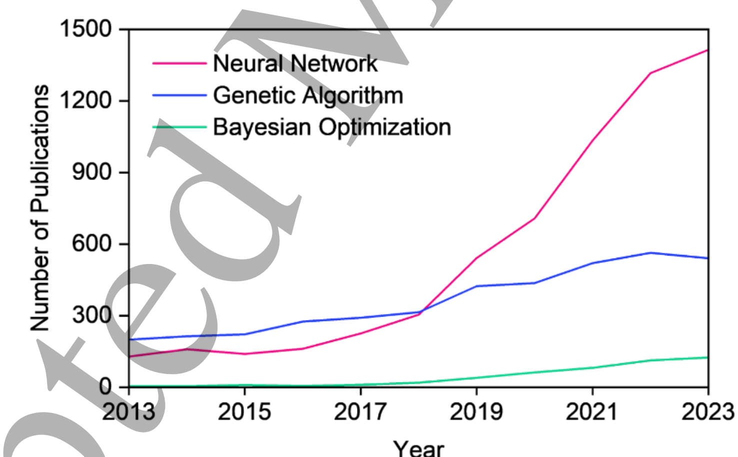

Figure 8. Annual number of publications using ML-aided design models in the material field. Keywords include 'Material', 'Design', and those in the figure legend.

2.6. Quantum annealing-aided active learning for material design

In many energy material design tasks, binary optimization can be an efficient strategy as material states can be described using discrete variables. For example, in the design of optical materials, planar multilayered geometry can be represented as a binary vector by assigning a binary number to each layer according to the corresponding material. Similarly, metasurfaces or stratified gratings geometries can be represented as a binary vector by discretizing the unit cell into pixels and assigning a binary label to each pixel depending on the material. As the material configuration directly determines the material performance, the design task can be transformed into binary optimization (i.e., combinatorial optimization problems). However, increasing the number of variables (e.g., the number of layers or pixels in the material structures) will exponentially increase the total possible combinations, resulting in an explosion of the combinatorial design space. For example, the design space size is 2^{20} (=1,048,576) if there are 20 binary variables for the input vector (assuming each layer or pixel has two options in material choice), while the design space size is 2^{30} (=1,073,741,824) for 30 binary variables. Exploring such large design spaces to find the best input state is extremely challenging or impossible because of computational limitations. To overcome this limitation, one can transform material design tasks into quadratic unconstrained binary optimization (QUBO) problems, where QUBO can be efficiently solved by a quantum computer [134, 135]. In particular, a quantum annealer, which is specially designed for solving combinatorial optimization problems by providing quantum speedup against classical counterparts by taking advantage of quantum physics (quantum tunneling), can efficiently be used to solve QUBO problems [136]. Then, the quantum annealer can find the ground state and the corresponding binary state of the given QUBO within a fraction of a second, even if the problem size is large [137]. A key to leveraging quantum annealing for material optimization is to formulate QUBO models as surrogates to describe the relationship between material states and their corresponding performance metrics since quantum computing is compatible with the QUBO model.

Factorization machine (FM) is a model that can be directly used to formulate the QUBO model (Q) by employing the model parameters after training FM [79]. FM was proposed by Rendle, and can be used as a supervised learning algorithm [138], which is designed to handle sparse and high dimensional data for classification and regression tasks. FM includes linear and factorization models, allowing the capture of the relationships between individual features and target variables (i.e., linear model) as well as interactions between features (i.e., factorization model). FM can learn feature interactions efficiently without explicitly enumerating all possible combinations and can be trained with gradient descent methods, enabling relatively short training times. Owing to these advantages, FM can be widely applicable to real-world problems that have sparse data, enabling us to design energy materials efficiently [138, 139]. Since input vector x is discretized into n variables, FM is suitable for combinatorial optimization problems. Individual features and interactions of FM can be trained with linear and quadratic models as the following equations:

$$\hat{y}(x) = w_0 + \sum_{i=1}^{n} w_i x_i + \sum_{i=1}^{n} \sum_{j=i+1}^{n} < v_i, v_j > x_i x_j \tag{5}$$

where n is the number of variables of x, w_0 is global bias, w is linear coefficients presenting individual features and $< v_i, v_j >$ models the interactions between x_i and x_j of size k. Factorizing the quadratic model $< v_i, v_j >$ can significantly reduce computational complexity (from $O(kn^2)$ to $O(kn)$) by reformulating complex interaction models into linear ones:

$$\sum_{i=1}^{n}\sum_{j=i+1}^{n} < v_i, v_j > x_i x_j = \frac{1}{2}\sum_{f=1}^{k}(\left(\sum_{i=1}^{n} v_{i,j} x_i\right)^2 - \sum_{i=1}^{n} v_{i,f}^2 x_i^2) \quad (6)$$

In a QUBO matrix (Q), diagonal elements are formulated from linear coefficients (w), and off-diagonal elements are formulated from quadratic coefficients (v) of the FM model. Then, quantum computers can be leveraged to find the ground state and corresponding binary state of the given QUBO problem:

$$\hat{y}(x) = x^T Q x \quad (7)$$

where x is the input binary vector, Q is a given QUBO, and $\hat{y}(x)$ is the objective function.

Active learning algorithms that integrate FM with quantum annealing have recently been utilized to design energy materials, such as multi-layered photonic structures, metamaterials for thermal management, and metamaterials for thermophotovoltaic applications [79, 134, 140-142]. These algorithms demonstrate potential in designing complex structures that pose large optimization spaces.

3. Design of energy materials using ML

In the previous section, we have discussed different ML schemes used in energy materials design with examples for each of them. In this section, we discuss several types of energy materials that have seen most ML activities.

3.1. Radiative cooling materials and structures

Passive radiative cooling, emitting thermal radiation into cold space (~3 K) through an atmospheric window (AW; wavelength: 8 to 13 μm), has attracted enormous attention as an efficient solution to reduce cooling energy consumption in response to climate change [143-145]. However, optimal design of radiative cooling materials is challenging as there are multiple design parameters such as dimensions and material composition. ML has been introduced to enable the optimization of such design parameters to achieve high-performance radiative cooling materials. Li et al. [146] optimized material compositions and layer thicknesses for daytime radiative cooler using ML (light gradient boosting machine) and genetic algorithm (figure 9(a)). They demonstrated that time consumption for the optimization could be significantly reduced from 7783.37 s to 115.81 s (~67 times acceleration) by using ML instead of using an analytical method (transfer matrix method). The optimized structure showed high reflectivity in the solar spectrum range and high emissivity in the AW (figure 9(b)), allowing to emit thermal radiation efficiently, leading to high cooling power (~140.38 W/m^2) and daytime temperature reduction (~9.08 °C) compared to the ambient temperature. Guan et al. [147] designed a transmissive colored radiative cooling film by optimizing film structures (layer configuration and thicknesses) with ML techniques (mixed-integer memetic algorithm and tandem NN, figure 9(c)). ML substituted the time-consuming 3D optics simulations, which led to significant acceleration for the optimization. The optimized film presented better visible light transmissivity compared to other colored radiative cooling films. Furthermore, the film showed a high emissivity in the AW (figure 9(d)), yielding a good cooling performance with a cooling power density of 126.6 W/m^2.

ML-aided optimization is getting more challenging as the design space is getting larger. To overcome this computational limitation, Kiati et al. [134] proposed a structural optimization method (called FMQA, figure 9(e)), which incorporates FM and QA. They designed metamaterial to achieve high radiative cooling performance using the FMQA scheme where FM was used to build a QUBO, and QA (D-Wave quantum annealer) was employed to solve the QUBO. They demonstrated a great performance of the proposed FMQA method compared to other optimization methods (GP, random search, and exhaustive search). Moreover, they could successfully design complex metamaterials with large design spaces (total possible configuration: ~2^{50}) thanks to the advantages from QA, and the optimized metamaterial presented near-ideal emissivity in the AW (figure 9(f)). Existing radiative cooling materials are generally reflective to reduce solar absorption and transmission [148]. Although radiative coolers that are transparent in the solar spectrum have been proposed, transmitted ultraviolet (UV) and near-infrared (NIR) lights can still significantly contribute to optical heating, which adversely affects cooling performance [149, 150]. Kim et al. [79] designed planar-multilayered photonic structures for transparent radiative coolers that have selective transmissivity to reduce solar heating by reflecting UV and NIR light while allowing visible light transmission. For multilayered structures, there can be lots of possible configurations (4^{24}), which may be beyond the limits of the computational capability. Hence, they used the FMQA to enable the optimization, and were able to successfully optimize a multi-layered structure within 58 hours, which might take ~89 million years with an exhaustive enumeration (figure 9(g)). The optimized structure showed the best-in-class performance compared to other transparent radiative coolers or energy-saving glasses. Furthermore, they experimentally demonstrated the unique optical characteristics (i.e., selective transmissivity in the visible regime, figure 9(h)) and cooling performance (temperature reduction of 6.1°C and potential cooling energy saving of 86.3 MJ/m^2 compared to normal glass window). This represents the first example of the practical realization of quantum computing designed energy material.

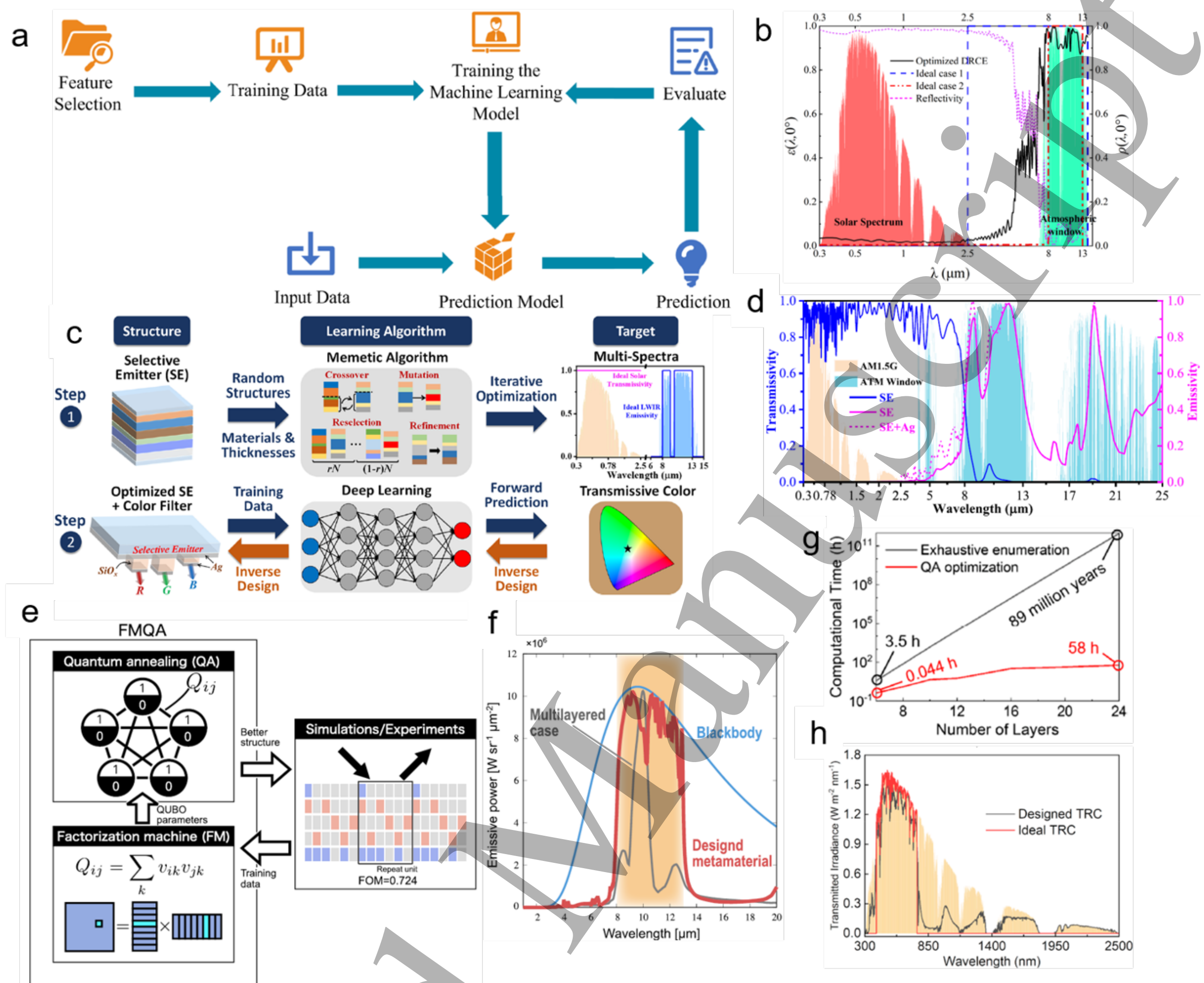

Figure 9. ML design of radiative cooling materials and structures. (a) A ML workflow to design a high-performance radiative cooler, developed in [146]. (b) Optical characteristics of the optimally designed radiative cooler using ML. The designed cooler has high emissivity in the AW to have high cooling performance. Reprinted with permission from [146] © Optical Society of America. (c) ML-assisted inverse design strategy for the design of transmissive colored radiative cooling films in [147]. (d) Optical characteristics of the optimally designed film. This film has high transmissivity in the solar spectrum and high emissivity in the AW, showing visible transparency with high cooling capability. Reproduced from [147]. CC BY 4.0. (e) A workflow of FMQA for automated designs, suggested in [134]. (f) Emissive power of the designed metamaterial in the AW, enabling high cooling performance. Reproduced from [134]. CC BY 4.0. (g) Computational time required for the optimization of a complex system with exhaustive enumeration and FMQA method, studied in [79]. (h) Transmitted irradiance through the designed transparent radiative cooler (TRC). This transparent radiative cooler has high transmissivity in the visible range while having low transmissivity in the ultraviolet and near-infrared ranges, resulting in minimized optical heating from sunlight while keeping visible transparency. Reprinted with permission from [79] © American Chemical Society.

3.2. Batteries

As new technologies, such as electric vehicles, portable electronics (smartphones), and renewable energies, become an integral part of our daily lives, developing high-performance batteries is crucial for providing and storing the energy for them [151]. However, it is also challenging to optimize batteries because of the large design space that comes from many parameters such as material composition, mixing ratio, stoichiometry, mechanical properties, shapes, and sizes. Hence, researchers have utilized ML techniques for the optimization of batteries. Using solid electrolytes to suppress dendrite growth has emerged as a promising strategy for next-generation batteries based on lithium metal anodes. Ahmad et al. [152] employed data-driven ML algorithms (graph convolutional NN, gradient boosting regressor, and kernel ridge regression) to predict the mechanical properties of inorganic solid electrolytes (e.g., shear modulus, Poisson's ratio, and molar volume ratio of solid electrolytes), which are important to determine the stability of the interface by estimating dendrite initiation. They trained their ML algorithms with data in the Material Project database (figure 10(a)) [153], and they were able to find some electrolytes expected to suppress dendrite initiation and growth (e.g., Li_2WS_4, $LiAuI_4$, $Ba_{38}Na_{58}Li_{26}N$). Joshi et al. [154] developed a ML-based algorithm (DNN, SVM, and kernel ridge regression) to predict electrode voltages for metal-ion batteries (figure 10(b)). They also used the Material Project database [153] to train their ML algorithms. Their data-driven ML approach enabled them to overcome computational difficulties to explore large design spaces and provided a fast estimation of the voltages as an alternative to DFT calculations. Their ML models showed high accuracy (figure 10(c)) in predicting voltages of electrode materials (e.g., Li-, Na-, K-, Mg-, Ca-, Zn-, Al-, and Y-ion batteries), thus it could guide the exploration of many different combinations of electrode materials.

Improvements in battery performance include costly and time-consuming work due to the difficulty in accurately formulating the relationships between inputs and outputs of the optimization problem. Dave et al. [155] used BO to autonomously discover novel battery materials (aqueous electrolytes). They demonstrated that the optimized electrolytes increased stability at a low leakage current (24 mV higher in the blend) and suppressed current density (~58% at 2 V, compared to $NaClO_4$ feeder solution). Accurate prediction of battery life is challenging since it requires a comprehensive understanding of battery systems and involves high costs for testing. Kim et al. [156] used ML methods (deep learning with simulation and predictive curve fitting) for early battery life prediction. ML algorithms were well trained with 2-3 weeks of data for the life prediction, and predictions were accurate with errors below 10%, enabling the reduction of costs associated with the prediction of battery performance (figure 10(d)). Although voltage profile images contain lots of information to determine battery performance, capturing subtle changes in images by human eyes is difficult. Chen et al. used a ML algorithm (CNN) pre-trained on ImageNet [157] to predict battery performance by using voltage profile images [158]. They further trained the algorithm on experimental data collected at different experimental conditions, and the resulting ML model showed high accuracy. Battery performance is dependent on historical information, and their ML model trained on historical data could be used to predict future performance such as remaining useful lifetime and general stability.

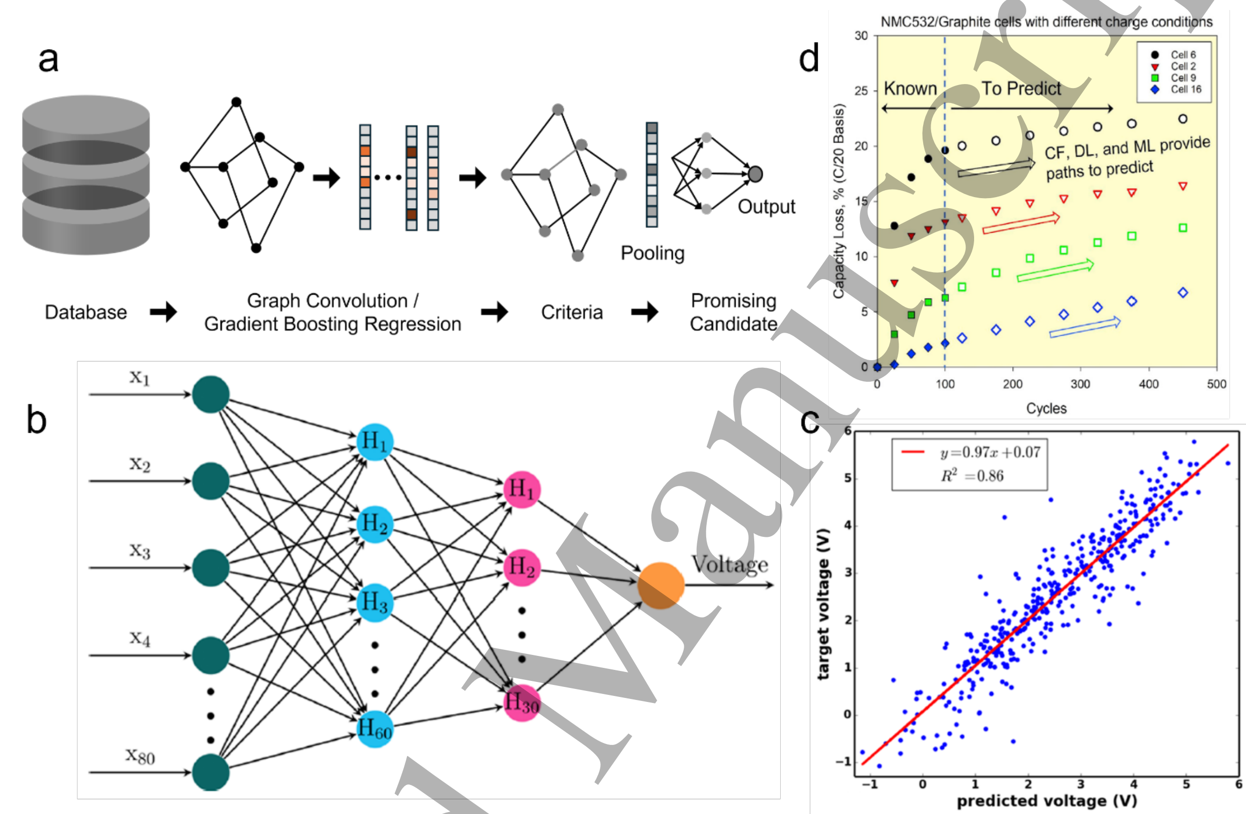

Figure 10. ML design of battery materials. (a) An ML workflow to design high-performance batteries. (b) Architecture of the ML to predict output voltage of metal-ion batteries, developed in [154]. (c) Accuracy of the developed ML model, enabling the prediction of output voltage of batteries. Reprinted with permission from [154] © American Chemical Society. (d) Battery life predictions using early-life data by utilizing ML models, developed in [156]. Reproduced from [156]. CC BY-NC-ND 4.0.

3.3. Photovoltaics

Perovskite materials are promising candidates that can be used in photovoltaics [159-162], which have attracted extremely extensive interest in the scientific community in recent years. However, improving the performance of photovoltaics, such as energy conversion efficiency, durability, and lifespans, poses challenges due to the complexity of optimizations [163, 164]. To overcome those challenges, Yu et al. [165] built ML models to predict relations between chemical-physical properties of amines and their reactivities to organic-inorganic hybrid perovskite ($MAPbI_3$) film (figure 11(a)). They tested various ML algorithms such as logistic regression, SVM, K-nearest neighbors and decision trees, and they achieved the highest score of 86% accuracy (accurate prediction / total prediction) on test data using the SVM with a radial basis function kernel. With the trained ML model, they could predict reactivities of un-trained amines to the hybrid perovskite. Moreover, they could learn chemical insights and knowledge by screening coefficients of the model, guiding new experimental conditions. To enable the rapid discovery of functional materials for ferroelectric photovoltaic perovskites, Lu et al. [166] developed a multistep screening scheme by combining DFT calculations and ML techniques. They successfully trained ML algorithms with collected data from high-throughput first-principles calculations. The trained models could achieve high accuracy (ROC-AUC of ~0.89 for the classification model and R^2 score of ~0.921 for the gradient boosting regression model) and showed accurate prediction for both perovskites and non-perovskites. Using the models, they found some mixed halide perovskites (e.g., $CsGeBr_2I$, $RbGeBr_2I$, $CsGeI_2Br$, $RbSnCl_2I$, and $RbSnI_2Cl$), which were close to the optimal value of single-junction solar cells.

Prediction of material properties is important to design perovskite materials. To predict key properties of perovskite materials, Stanley et al. [167] employed a ML approach (kernel ridge regression) for learning complex relations between material compositions and corresponding properties from a limited number of data. They calculated 344 mixed perovskites using DFT, and used them to train their ML algorithm, resulting in a good model for the prediction (figure 11(b)). Thus, they could rapidly predict several important properties of photovoltaics in the composition space, enabling the suggestion of the rational design of new perovskites (figure 11(c)). She et al. [168] utilized a two-step ML approach with classification and regression models to find highly efficient perovskite solar cells by exploring a vast design space. They used experimental data extracted from the published literature to train the ML algorithm. With the model showing high accuracy, they could successfully extract general underlying knowledge of perovskite solar cells by analyzing important features. In addition, they could discover high-performance perovskite solar cells with doped electron transport layers (e.g., Cs-doped TiO_2 electron transport layers, and S-doped SnO_2 electron transport layers) having high power conversion efficiency of up to 30.47%. Inherent ionic defects in perovskites can lead to damage to their stability, impeding their practical applications, but high computational costs associated with DFT calculations and inaccurate predictions pose challenges to improving the stability of perovskite materials. Yang et al. [169] developed an interatomic potential model by employing a ML algorithm (deep learning) to analyze the ionic defect effects. The model performance was improved by iteratively exploring design space similar to active learning, leading to an efficient model with high-level accuracy close to classical MD calculations (figure 11(d,e)). With their model, they revealed the factors affecting ionic defects.

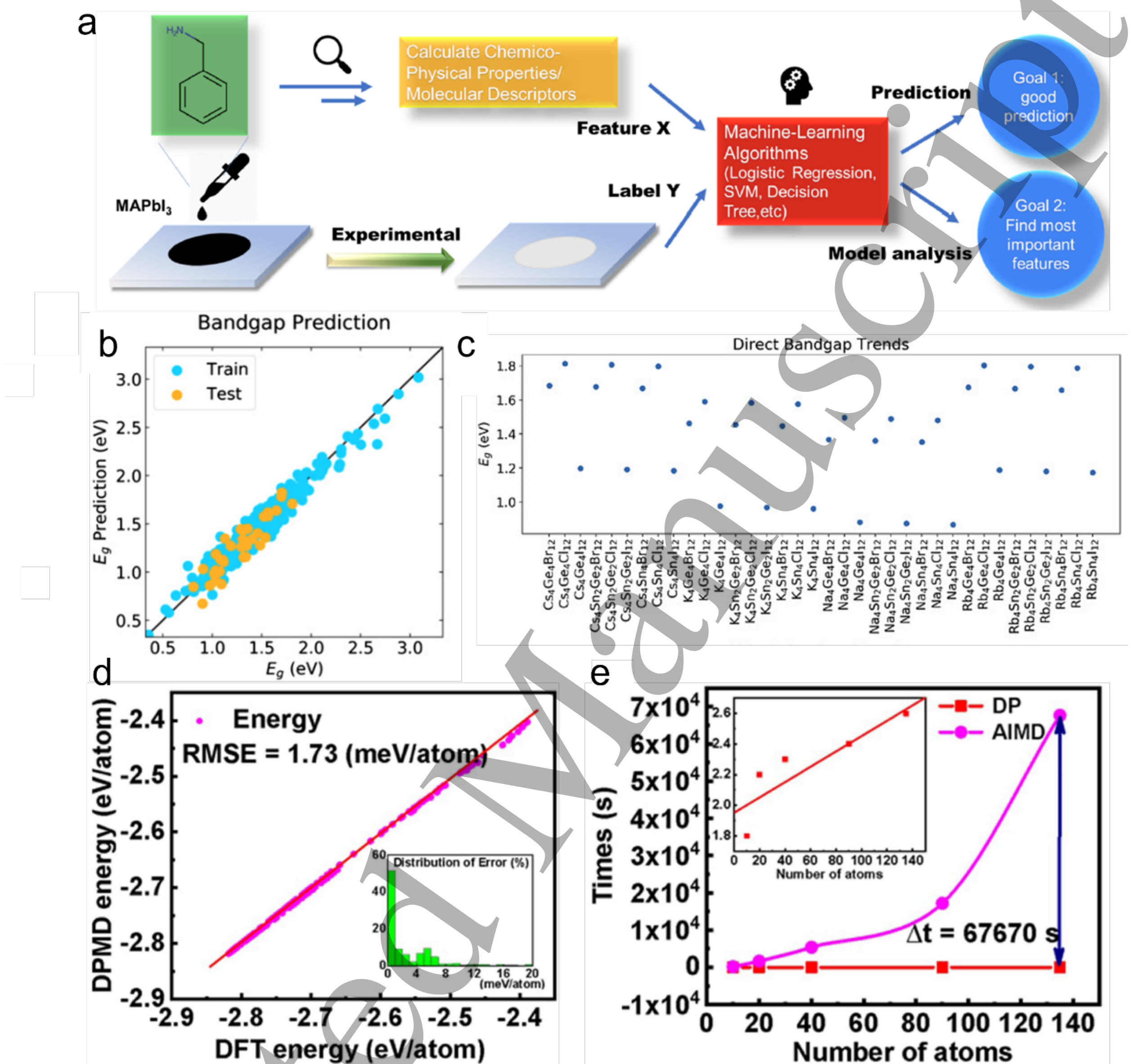

Figure 11. ML for perovskite photovoltaic materials. (a) A workflow of ML-assisted exploration to study the compatibility of organic-inorganic hybrid perovskite film with amines, developed in [165]. Reprinted with permission from [165] © American Chemical Society. (b) Comparison between actual and predicted value of the ML model, showing high accuracy in [167]. (c) ML model can be used to predict characteristics of perovskites. Reproduced from [167]. CC BY 4.0. (d) Comparison between the predicted values from the ML-assisted model and DFT-calculated energy values, studied in [169]. (e) Comparison of time for calculations using the developed model and ab initio molecular dynamics, demonstrating the efficiency of the developed ML-assisted model. Reprinted with permission from [169] © John Wiley & Sons.

3.4. Gas separation materials

The application of membrane technology, especially utilizing polymers for gas separation, has become critical for processes like carbon dioxide capture, hydrogen separation, and natural gas sweetening [170, 171]. While polymeric membranes find widespread use, they encounter challenges such as permeability-selectivity trade-offs, physical aging, and plasticization, limiting their broader utility. To overcome these multi-objective design challenges, the integration of ML techniques has gained some momentum in expediting the screening and design of high-performance polymeric gas separation materials. An early effort in this field traces back to 1994 when Wessling et al. [172] pioneered the use of a NN to model the CO_2 permeability of polymers, utilizing infrared spectra as input features. Despite a limited database size (only 33 polymers), relatively accurate predictions highlighted the substantial potential of ML in quantitative structure-property relationship (QSPR) analysis for polymeric membrane gas separation materials. Subsequent research endeavors have expanded on this foundation, with the accumulation of gas separation data and the advancement in ML algorithms. Zhu et al. [173] utilized GP regression to predict permeability for various gases in a dataset of 315 polymers, employing a hierarchical fingerprinting method based on the chemical structure of the polymer repeating unit. Barnett et al. [174] followed a similar approach, utilizing GP regression and a topological, path-based fingerprint for around 700 polymers, demonstrating the model's ability to predict permeability values for ~ 10,000 unlabeled polymers. In addition to using handcrafted fingerprints or descriptors to represent polymer structural information, recent approaches involve representation learning from deep neural networks. Wilson et al. [175] treated polymer structures as graphs, developing a GNN named PolyID for efficient identification of high-performance biobased polymers. PolyID facilitated the discovery of biobased poly(ethylene terephthalate) analogs with enhanced thermal and gas separation performance.

3.5. Thermoelectric materials

Thermoelectric materials, which can convert thermal energy into electricity, can be a solution to global energy challenges by converting waste heat into useful energy. Due to the large stoichiometry and processing space, physics intuition-based optimization has been slow for thermoelectric materials design and process optimization. To overcome these challenges, researchers have applied ML techniques for the efficient development of thermoelectric materials and the prediction of their properties [176, 177]. Figure of merit (zT) is an important indication for the performance of thermoelectric materials. Hence, researchers have tried to efficiently predict zT and develop thermoelectric materials with high zT. zT is related to a few intercorrelated transport properties as the following equation [178]:

$$zT = S^2 \rho^{-1} \kappa^{-1} T \tag{8}$$

where S, ρ, κ, and T respectively represent the Seebeck coefficient, electrical resistivity, thermal conductivity, and absolute temperature. As can be seen from the zT expression, thermoelectric materials usually benefit from low thermal conductivity which can in turn improve their efficiency. However, prediction of the thermal conductivity of inorganic materials is challenging since only a few portions (5% among 10^5 synthesized inorganic materials) have a low thermal conductivity that is effective for thermoelectric materials. To tackle this challenge, Zhu et al. [179] employed ML techniques (crystal graph convolutional network and RF) for the prediction of the thermal conductivity of all known inorganic materials for thermoelectric applications (figure 12(a,b)). The trained models after including the transfer learning exhibited good accuracy, allowing for accurate predictions of thermal conductivity. Furthermore, they could identify a promising material system for thermoelectrics.

Li et al. [180] used a data-driven light gradient boosting (LGB) model to directly predict the performance (zT) of thermoelectric materials. They trained the model with selected data from the database by the University of California Santa Barbara (UCSB) [181]. The trained model showed a high accuracy (high R^2 value of ~0.96 and low RMSE of ~0.09), resulting in accurate zT value predictions (figure 12(c)). As a result, they could discover some potential materials that have high zT among a large candidate pool (1 million). Furthermore, they could extract feature importance by analyzing the frequency of a feature used as a node (figure 12(d)). Zhan et al. [182] leveraged an ML method to predict thermal boundary resistance, which is one of the keys for the thermal conductivity of thermoelectric materials. They collected data from the literature, and trained their ML models (generalized linear regression, least-absolute shrinkage and selection operator regularization, GP regression, and support vector regression), resulting in some reliable models. They successfully predicted thermal boundary resistance with a model, and they could find the important descriptor (film thickness) to predict the thermal property. Jia et al. [68] used an unsupervised learning method to discover promising materials for thermoelectrics. They trained several unsupervised algorithms (e.g., K-means clustering, Gaussian Mixture, Mean Shift) with data in the Materials Project database [183] for clustering promising materials. They successfully discovered some materials with high performance using their trained ML model.

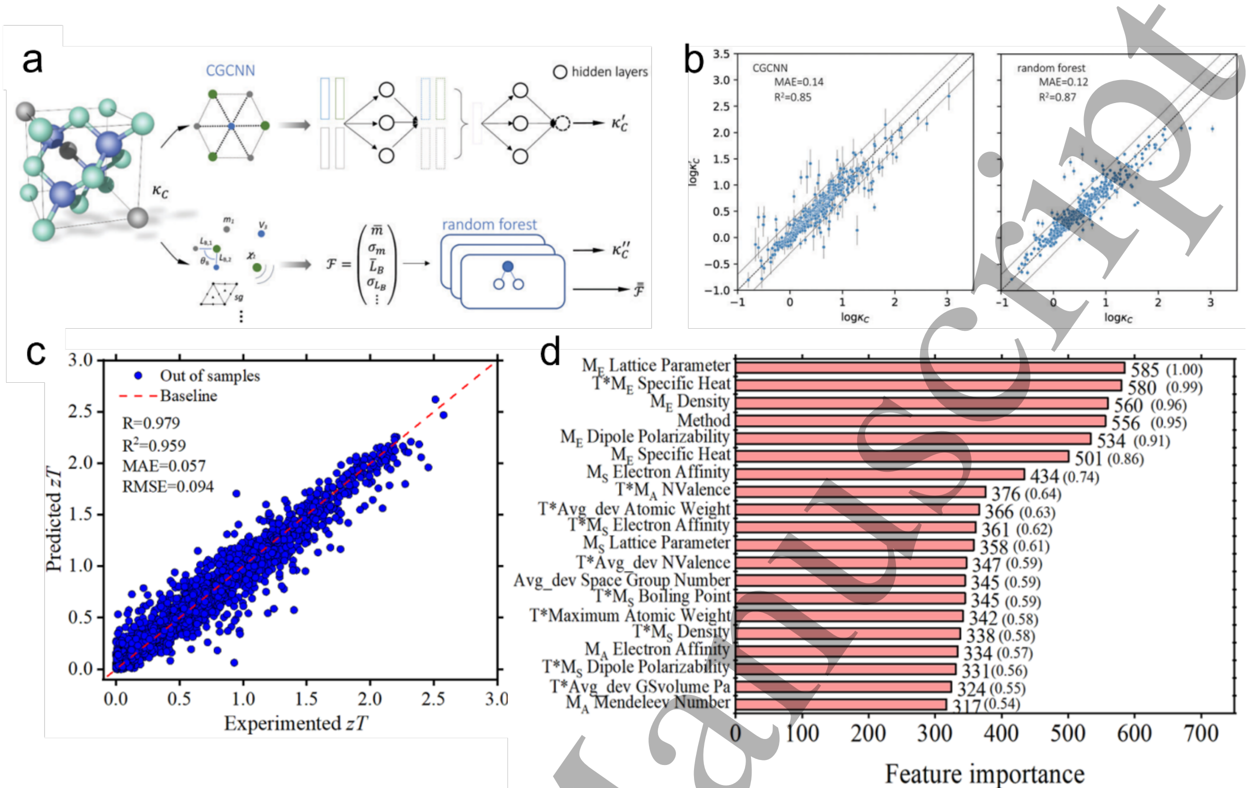

Figure 12. ML for thermoelectric materials. (a) The schematic of the ML models (crystal graph convolutional network (CGCNN) and RF), developed in [179]. (b) Accuracy of the models. k_c and k'_c respectively represent the calculated and predicted thermal conductivity. Reproduced from [179]. CC BY-NC 3.0. (c) Accuracy of the trained ML model used, used in [180]. (d) Extracted feature importance from the ML model. Reprinted with permission from [180] © American Chemical Society.

3.6. Supercapacitors

Designing high-performance supercapacitors, which are energy storage devices, has drawn great attention over the past few decades due to their potential high power density, high specific capacity, and rapid charging/discharging rate [184, 185]. Predicting specific capacity and cyclic stability is important for evaluating the performance of supercapacitors, but it is challenging with first-principles strategies. To address this issue, Ghosh et al. [186] utilized RF and MLP models for the prediction of the capacitance and cyclic stability of supercapacitors. Their ML models successfully predicted these important properties for supercapacitors composed of cerium oxynitride, a promising electrode material. Aqueous supercapacitors have emerged as promising energy storage devices since they exhibit excellent power density and long lifetime cycles. Here, porous carbon materials, which possess large surface area and rich porous structures, can enhance the overall performance of supercapacitors [187]. However, designing these porous structures is difficult and time-intensive [188]. Wang et al. [189] employed an ANN model to identify the critical features of carbon materials by predicting the specific capacitance of hyperporous carbons. They revealed that the ANN model achieved high accuracy when employing Bayesian regularization (figure

13(a)), which led to the successful prediction of the capacitance and cyclic stability. This enabled the discovery of high-performance carbon materials for supercapacitors (figure 13(b)).

3.7. Polymers

Polymers are widely used in energy materials, such as energy storage devices, batteries, and solar cells, making the optimal design of polymers important [190-192]. However, the limited data on polymeric properties and their structural complexity hinder the identification of high-performance polymers. To tackle these challenges, Wu et al. [193] used ML models that combine the Bayesian molecular design framework and transfer learning to predict polymeric properties. They trained the ML model using the database from PoLyInfo, and the trained model achieved high accuracy, as can be seen in figure 13(c). As a result, they could discover promising polymers yielding high thermal conductivities (figure 13(d)). The dielectric constant of polymers is a key parameter for determining the performance of energy materials, but predicting this property using conventional methods, such as density functional perturbation theory or MD simulations, involves time-intensive work with low reliability. To address this challenge, Chen et al. [194] developed an ML-based model that includes a polymer fingerprinting scheme and Gaussian process regression algorithm. They trained their model with data collected from the literature, achieving acceptable prediction accuracy. This led to the successful prediction of the dielectric constant of synthesizable candidate polymers.

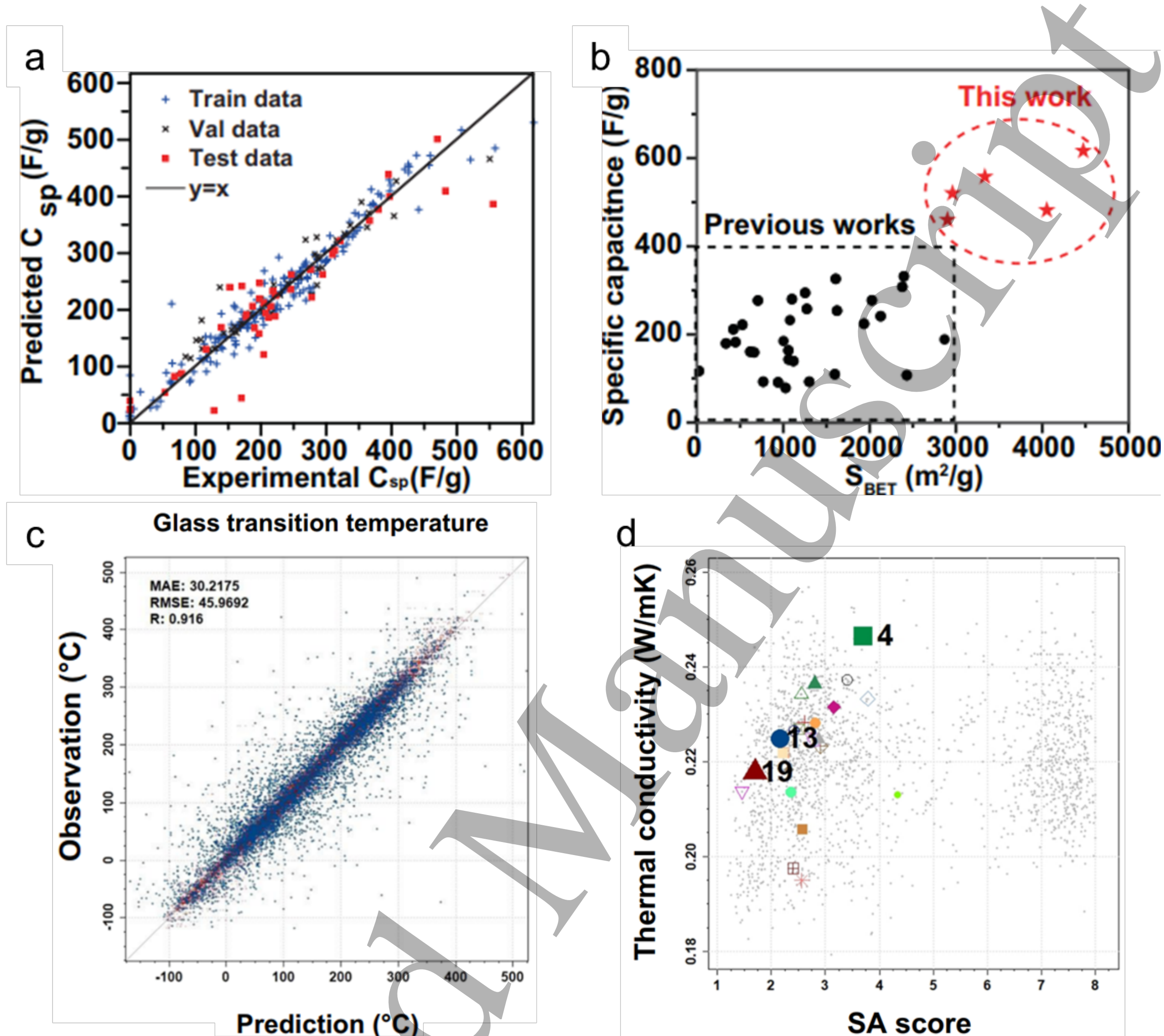

Figure 13. ML for supercapacitors and polymers. (a) Accuracy of the ANN model for predicting capacitance, used in [189]. (b) Comparison of capacitances between previously reported values and those identified in this work, demonstrating potential in discovering high-performance supercapacitors. Reproduced from [189]. CC BY 4.0. (c) Accuracy of the ML model, used in [193]. (d) Predicted thermal conductivity as a function of SA score that indicates synthesizability, demonstrating the capability to the identification of synthesizable polymers with high thermal conductivity. Reproduced from [193]. CC BY 4.0.

4. Summary and perspectives

4.1. Summary

In summary, by reviewing the literature, we have shown that ML approaches have been widely used for the design of energy materials for a wide variety of applications to overcome limitations caused by experimental or computational costs to obtain material properties. Recent progress in computational power and ML algorithms enables users to utilize ML more efficiently in energy material fields to predict material properties, search vast design spaces, and discover optimal design parameters. We have concisely reviewed the basics of ML techniques and surveyed some ML-aided optimization schemes for energy materials. We have shown that the trained ML models can be applied in various research fields for property predictions or inverse design, which have been demonstrated with the examples. Overall, it has been demonstrated that ML techniques can play important roles in guiding the efficient design of high-performing energy materials, although challenges still exist.

4.2. Challenges and Perspectives

A number of major challenges are still present in using ML for energy materials design and optimization. These are discussed in this section.

4.2.1. Low quality and low volume of data for ML

ML training with small, imbalanced or low-quality data can make the models biased and cannot properly cover entire feature spaces, hindering learning complex relationships across the whole design spaces. Hence, the model can be under- or over-fitted, which leads to inaccurate predictions [195]. To mitigate these issues, data augment techniques, such as rotating [93], node feature masking [196], edge dropping [197], and subgraph replacement [198], can be applied. In addition, active learning strategies can allow the model to collect meaningful data, enhancing the model's performance iteratively even starting with a limited amount of data [127, 132].

4.2.2. ML algorithms working with limited and imbalanced data

Large materials databases based on high-accuracy simulations and experiments are the foundation for the applications of advanced ML algorithms, especially deep learning algorithms for material design, and catalyzed the development of materials informatics. However, for many of the properties that are not easy to measure or compute, the lack and imbalance of data remain huge obstacles for researchers to train accurate ML models. Recently, some techniques such as threshold-moving [199], transfer learning (leverages models trained on large datasets to build models on small datasets of different properties) [200-202], multi-fidelity modeling [203], and active learning [129] have been proposed to face the challenges of small and imbalanced data. These techniques allowed for material designs with limited and imbalanced data [204, 205].

4.2.3. Design of synthesizable materials using ML

The synthesizability of materials designed using ML remains one of the greatest challenges for the further development of ML for energy materials and materials in general. Bridging the gap between algorithmically proposed materials and successful laboratory synthesis involves addressing critical factors like possible and optimal experimental conditions. To augment the synthesizability of generated materials, integrating ML-driven retrosynthesis planning with generative algorithms emerges as a promising solution. Retrosynthesis planning falls into template-based and template-free categories [206, 207]. Template-based approaches rely on summarized reaction rules while template-free methods, often utilizing deep learning, predict reactants directly. An example of template-based retrosynthesis planning is presented by Chen et al. [208], who developed a data-assisted tool. However, it has limitations, including neglecting important design factors such as experimental conditions and potential ineffectiveness with new materials. Template-free methods, although potentially more versatile, may require substantial training data. Exploring the potential of large language models for polymer structure generation and optimization, considering retrosynthesis planning, represents an exciting avenue for future research [77, 209, 210].

4.2.4. Multi-objective optimization

Multi-objective optimization in material design often faces conflicts in different properties to be optimized – improvement in one can lead to degradation in others. In this scenario, decision-makers can identify preferred solutions from the Pareto front, which represents optimal trade-offs between conflicting objectives. Approaches to solving these problems fall into two categories: a posteriori and a priori [211]. Posteriori methods aim to discover the entire Pareto front, allowing decision-makers to understand achievable objective values and make decisions based on the trade-offs between each objective. Recently, a noticeable number of works have been developed to reveal the Pareto optimal solutions [212, 213]. However, identifying the preferred solution on the Pareto front can be resource-intensive, particularly with a posteriori methods that require evaluating a large number of objective functions [214]. In a priori multi-objective optimization methods, decision-makers define their preferences upfront, streamlining the process towards specific goals and reducing the need for extensive objective evaluations. One common technique is the use of Achievement Scalarizing Functions, typically formulated as weighted sums of objectives based on the decision-maker's preferences and knowledge. While easy to implement, finding the right weight vectors to achieve Pareto optimal solutions remains a challenge. Another approach is optimizing a single objective subject to constraints on others [215]. Lexicographic methods are also used [211], prioritizing objectives according to an established hierarchy of importance. Each method offers distinct advantages and faces unique challenges, influenced by the optimization problem's complexity. For materials, additional challenges lie in the fact that different properties have various levels of difficulties to acquire computationally or experimentally. Therefore, removing the rate-limiting barrier for materials characterization is also key to ML-assisted energy material design.

4.2.5. Material design with properties outside the range of training data

Designers frequently face situations where the collected data does not adequately represent the domain trends, or in some cases, there is insufficient data to train an optimization model. This is usually known as the out-of-distribution prediction/design problem. This may be partially addressed by leveraging a latent space using encoder/decoder architectures. This strategy allows for the exploration of new material compositions and properties by navigating a lower-dimensional latent space, which enhances computational efficiency. Additionally, interpolation in the latent space may appear to be extrapolation when decoded into the real space, which has a much higher dimension. The latent space has enabled the discovery of novel materials exhibiting properties beyond those presented in the training data [216]. Also, the issue can be addressed through active learning and the utilization of surrogate models. Initially, the surrogate model is assumed to best represent the search space. New data points are actively acquired and integrated into the dataset for subsequent optimization rounds, gradually expanding the property boundaries. However, this approach, focusing only on the predictive mean of the surrogate model, may not effectively balance exploration and exploitation. Advanced methods involve applying BO to probabilistic surrogate models (e.g., GP), considering both uncertainty and predictive mean. This allows for tailored adjustments in the balance between exploration and exploitation, based on prior beliefs. Such an ML manner to data acquisition can help minimize the need for new data in reaching the design target [217, 218].

4.2.6. Other thoughts

Addressing these above challenges will enable ML techniques to be more effective and to yield reliable outcomes in energy material design, allowing for applying them in various research and industrial fields. However, many ML algorithms are black-box, meaning that it is hard to explain their mechanisms. Hence, future development of ML algorithms should focus on building transparent and interpretable models, which will be more broadly applicable for decision-making, predictions, and inverse designs. Opening up the box will also shed light on the fundamental physics governing the material behavior, understanding which will improve the knowledge base and is more generalizable than a dataset or a ML model.

Hyperparameters, which are not learned from data, are crucial components to determining the performance of ML algorithms, but identifying optimal hyperparameters is challenging. Optimization spaces of hyperparameters may be complex, and interactions between hyperparameters may add complexity to the optimization process, making non-convexity of the objective function. This imposes an additional optimization problem on the ML materials optimization task. To tackle these difficulties, many approaches have been proposed to optimize hyperparameters using ML methods. With the optimal hyperparameters, ML can present higher performance for prediction and design in the energy material field.

As can be seen in Kiati and Kim's works [79, 134, 140], quantum computers exhibit notably enhanced computational capabilities to explore optimization spaces. Hence, the integration of ML algorithms and quantum computers will become important for the optimization of energy materials that have complex structures and characteristics. There are still current limitations on quantum computing hardware, such as the limited number of qubits, limited connections between the qubits, and the lack of capability to optimize effectively continuous variables.

Furthermore, in the future, it is expected that quantum ML, leveraging principles from quantum mechanics to address certain computational challenges much more efficiently, will enable us to build better models and identify optimal solutions much faster than classical ML approaches. While these are still limited by quantum computing hardware, these advancements, if realized, will open new avenues in energy material fields for highly complex properties and significantly large optimization spaces, which are difficult for now.

Data availability statement

All data that support the findings of this study are included within the article.

Acknowledgments

TL would like to acknowledge National Science Foundation (2102592, 2332270), DARPA (HR00112390112), DOE (DE-EE0009103) and ONR (N00014-18-1-2429). This research was supported by the Quantum Computing Based on Quantum Advantage Challenge Research (RS-2023-00255442) through the National Research Foundation of Korea (NRF) funded by the Korean government (Ministry of Science and ICT(MSIT)). This research used resources of the Oak Ridge Leadership Computing Facility at the Oak Ridge National Laboratory, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC05-00OR22725.
Notice: This manuscript has in part been authored by UT-Battelle, LLC under Contract No. DE-AC05-00OR22725 with the U.S. Department of Energy. The United States Government retains and the publisher, by accepting the article for publication, acknowledges that the U.S. Government retains a non-exclusive, paid up, irrevocable, world-wide license to publish or reproduce the published form of the manuscript, or allow others to do so, for U.S. Government 15 purposes. The Department of Energy will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan (http://energy.gov/downloads/doe-publicaccess-plan).

References

[1] Díaz S, et al. 2019 Pervasive human-driven decline of life on Earth points to the need for transformative change *Science* **366** eaax3100

[2] Rogelj J, et al. 2016 Paris Agreement climate proposals need a boost to keep warming well below 2 degrees C *Nature* **534** 631-639

[3] Green M A 2019 Photovoltaic technology and visions for the future *Progress in Energy* **1** 013001

[4] Gielen D, Boshell F, and Saygin D 2016 Climate and energy challenges for materials science *Nat Mater* **15** 117-120

[5] Rogelj J, et al. 2019 A new scenario logic for the Paris Agreement long-term temperature goal *Nature* **573** 357-363

[6] Dehghani-Sanij A R, Tharumalingam E, Dusseault M B, and Fraser R 2019 Study of energy storage systems and environmental challenges of batteries *Renew Sustain Energy Rev* **104** 192-208

[7] González M J and García Navarro J 2006 Assessment of the decrease of CO2 emissions in the construction field through the selection of materials: Practical case study of three houses of low environmental impact *Build Environ* **41** 902-909

[8] Han N, et al. 2019 Promises of Main Group Metal–Based Nanostructured Materials for Electrochemical CO2 Reduction to Formate *Adv Energy Mater* **10** 1902338

[9] Piasecka I, Bałdowska-Witos P, Piotrowska K, and Tomporowski A 2020 Eco-Energetical Life Cycle Assessment of Materials and Components of Photovoltaic Power Plant *Energies* **13** 1385

[10] Soleimani Z, et al. 2021 The cradle to gate life-cycle assessment of thermoelectric materials: A comparison of inorganic, organic and hybrid types *Sustain Energy Technol Assess* **44** 101073

[11] Green M A and Bremner S P 2016 Energy conversion approaches and materials for high-efficiency photovoltaics *Nat Mater* **16** 23-34

[12] Cuevas F, et al. 2022 Metallic and complex hydride-based electrochemical storage of energy *Progress in Energy* **4** 032001

[13] Sivula K and van de Krol R 2016 Semiconducting materials for photoelectrochemical energy conversion *Nat Rev Mater* **1** 15010

[14] Yan C, et al. 2018 Cu2ZnSnS4 solar cells with over 10% power conversion efficiency enabled by heterojunction heat treatment *Nat Energy* **3** 764-772

[15] Zhu L, et al. 2022 Single-junction organic solar cells with over 19% efficiency enabled by a refined double-fibril network morphology *Nat Mater* **21** 656-663

[16] Allouhi A, et al. 2015 Energy consumption and efficiency in buildings: current status and future trends *J Clearn Prod* **109** 118-130

[17] Pope M A and Aksay I A 2015 Structural Design of Cathodes for Li-S Batteries *Adv Energy Mater* **5** 1500124

[18] Zhang X, et al. 2020 Multiscale Understanding and Architecture Design of High Energy/Power Lithium-Ion Battery Electrodes *Adv Energy Mater* **11** 2000808

[19] Fayaz H, et al. 2022 Optimization of Thermal and Structural Design in Lithium-Ion Batteries to Obtain Energy Efficient Battery Thermal Management System (BTMS): A Critical Review *Arch Comput Methods Eng* **29** 129-194

[20] Yang C, et al. 2019 Theory-Driven Design and Targeting Synthesis of a Highly-Conjugated Basal-Plane 2D Covalent Organic Framework for Metal-Free Electrocatalytic OER *ACS Energy Lett* **4** 2251-2258

[21] Sahoo S K, et al. 2018 Rational Design of TiC-Supported Single-Atom Electrocatalysts for Hydrogen Evolution and Selective Oxygen Reduction Reactions *ACS Energy Lett* **4** 126-132

[22] Mouchou R T, Jen T C, Laseinde O T, and Ukoba K O 2021 Numerical simulation and optimization of p-NiO/n-TiO2 solar cell system using SCAPS *Mater Today Proc* **38** 835-841

[23] Chavez S, Aslam U, and Linic S 2018 Design Principles for Directing Energy and Energetic Charge Flow in Multicomponent Plasmonic Nanostructures *ACS Energy Lett* **3** 1590-1596

[24] Ma R, et al. 2022 Machine learning-assisted exploration of thermally conductive polymers based on high-throughput molecular dynamics simulations *Mater Today Phys* **28** 100850

[25] Ma R, Zhang H, and Luo T 2022 Exploring High Thermal Conductivity Amorphous Polymers Using Reinforcement Learning *ACS Appl Mater Interfaces* **14** 15587-15598

[26] Xu Z, Huang D, and Luo T 2021 Molecular-Level Understanding of Efficient Thermal Transport across the Silica–Water Interface *J Phys Chem C* **125** 24115-24125

[27] Li R, Lee E, and Luo T 2020 A unified deep neural network potential capable of predicting thermal conductivity of silicon in different phases *Mater Today Phys* **12** 100181

[28] Herrera A, et al. 2012 Applications of finite element simulation in orthopedic and trauma surgery *World J Orthop* **3** 25-41

[29] Gu G H, Noh J, Kim I, and Jung Y 2019 Machine learning for renewable energy materials *J Mater Chem* **7** 17096-17117

[30] Say A C C and Akın H L 2003 Sound and complete qualitative simulation is impossible *Artif Intell* **149** 251-266

[31] Kespe M and Nirschl H 2015 Numerical simulation of lithium-ion battery performance considering electrode microstructure *Inter J Energy Res* **39** 2062-2074

[32] Siddique N A and Liu F 2010 Process based reconstruction and simulation of a three-dimensional fuel cell catalyst layer *Electrochim Acta* **55** 5357-5366

[33] Gvozdetskyi V, Selvaratnam B, Oliynyk A O, and Mar A 2023 Revealing Hidden Patterns through Chemical Intuition and Interpretable Machine Learning: A Case Study of Binary Rare-Earth Intermetallics RX *Chem Mater* **35** 879-890

[34] Zou Q, et al. 2023 Deciphering Alloy Composition in Superconducting Single-Layer FeSe(1-x)S(x) on SrTiO(3)(001) Substrates by Machine Learning of STM/S Data *ACS Appl Mater Interfaces* **15** 22644-22650

[35] Liu Z, Jiang M, and Luo T 2022 Leveraging low-fidelity data to improve machine learning of sparse high-fidelity thermal conductivity data via transfer learning *Mater Today Phys* **28** 100868

[36] Moon G, et al. 2022 Machine learning and its applications for plasmonics in biology *Cell Rep Phys Sci* **3** 101042

[37] Lee J, et al. 2023 Machine Learning Driven Channel Thickness Optimization in Dual-Layer Oxide Thin-Film Transistors for Advanced Electrical Performance *Adv Sci* **10** e2303589

[38] Elzouka M, et al. 2020 Interpretable Forward and Inverse Design of Particle Spectral Emissivity Using Common Machine-Learning Models *Cell Rep Phys Sci* **1** 100259

[39] Tanaka K, et al. 2019 Machine-Learning Analysis to Predict the Exciton Valley Polarization Landscape of 2D Semiconductors *ACS Nano* **13** 12687-12693

[40] Frey N C, et al. 2019 Prediction of Synthesis of 2D Metal Carbides and Nitrides (MXenes) and Their Precursors with Positive and Unlabeled Machine Learning *ACS Nano* **13** 3031-3041

[41] Zhi C, et al. 2023 Machine-Learning-Assisted Screening of Interface Passivation Materials for Perovskite Solar Cells *ACS Energy Lett* **8** 1424-1433

[42] Ma W, Cheng F, and Liu Y 2018 Deep-Learning-Enabled On-Demand Design of Chiral Metamaterials *ACS Nano* **12** 6326-6334

[43] Liang Z, et al. 2022 Machine-learning exploration of polymer compatibility *Cell Rep Phys Sci* **3** 100931

[44] Dong H, Xie J, and Zhao X 2022 Wind farm control technologies: from classical control to reinforcement learning *Progress in Energy* **4** 032006

[45] Yin H, et al. 2021 The data-intensive scientific revolution occurring where two-dimensional materials meet machine learning *Cell Rep Phys Sci* **2** 100482

[46] Siemers F M, Feldmann C, and Bajorath J 2022 Minimal data requirements for accurate compound activity prediction using machine learning methods of different complexity *Cell Rep Phys Sci* **3** 101113

[47] Zhang L, et al. 2022 Fundamentals of hydrogen storage in nanoporous materials *Progress in Energy* **4** 042013

[48] Guan J, et al. 2022 Design and prediction of metal organic framework-based mixed matrix membranes for CO2 capture via machine learning *Cell Rep Phys Sci* **3** 100864

[49] Wan S, et al. 2021 A coupled machine learning and genetic algorithm approach to the design of porous electrodes for redox flow batteries *Appl Energy* **298** 117177

[50] Dave A, et al. 2022 Autonomous optimization of non-aqueous Li-ion battery electrolytes via robotic experimentation and machine learning coupling *Nat Commun* **13** 5454

[51] Li J, Pradhan B, Gaur S, and Thomas J 2019 Predictions and Strategies Learned from Machine Learning to Develop High-Performing Perovskite Solar Cells *Adv Energy Mater* **9** 1901891

[52] Wu Y, Guo J, Sun R, and Min J 2020 Machine learning for accelerating the discovery of high-performance donor/acceptor pairs in non-fullerene organic solar cells *npj Comput Mater* **6** 120

[53] Feng Y, et al. 2021 Machine learning and microstructure design of polymer nanocomposites for energy storage application *High Voltage* **7** 242-250

[54] Deng C, et al. 2020 Integrating Machine Learning with Human Knowledge *iScience* **23** 101656

[55] Mintz Y and Brodie R 2019 Introduction to artificial intelligence in medicine *Minim Invasive Ther Allied Technol* **28** 73-81

[56] Zhou J, Li R, and Luo T 2023 Physics-informed neural networks for solving time-dependent mode-resolved phonon Boltzmann transport equation *npj Comput Mater* **9** 212

[57] Li R, Lee E, and Luo T 2023 Physics-Informed Deep Learning for Solving Coupled Electron and Phonon Boltzmann Transport Equations *Phys Rev Appl* **19** 064049

[58] Li R, Wang J-X, Lee E, and Luo T 2022 Physics-informed deep learning for solving phonon Boltzmann transport equation with large temperature non-equilibrium *npj Comput Mater* **8** 29

[59] Estalaki S M, et al. 2022 Predicting defects in laser powder bed fusion using in-situ thermal imaging data and machine learning *Addit Manufacturing* **58** 103008

[60] Li M, Dai L, and Hu Y 2022 Machine Learning for Harnessing Thermal Energy: From Materials Discovery to System Optimization *ACS Energy Lett* **7** 3204-3226

[61] Leong Y X, et al. 2022 Where Nanosensors Meet Machine Learning: Prospects and Challenges in Detecting Disease X *ACS Nano* **16** 13279-13293

[62] Zhang G-X, et al. 2022 Machine learning-facilitated multiscale imaging for energy materials *Cell Rep Phys Sci* **3** 101008

[63] Weng B, et al. 2020 Simple descriptor derived from symbolic regression accelerating the discovery of new perovskite catalysts *Nat Commun* **11** 3513

[64] Terayama K, Sumita M, Tamura R, and Tsuda K 2021 Black-Box Optimization for Automated Discovery *Acc Chem Res* **54** 1334-1346

[65] Alarie S, et al. 2021 Two decades of blackbox optimization applications *EURO Journal on Computational Optimization* **9** 100011

[66] Fu H, et al. 2023 Machine-Learning-Assisted Optimization of a Single-Atom Coordination Environment for Accelerated Fenton Catalysis *ACS Nano* **17** 13851-13860

[67] Liu Z, Jiang M, and Luo T 2020 Leverage electron properties to predict phonon properties via transfer learning for semiconductors *Sci Adv* **6** eabd1356

[68] Jia X, et al. 2022 Unsupervised machine learning for discovery of promising half-Heusler thermoelectric materials *npj Comput Mater* **8** 34

[69] Kim S, et al. 2023 PubChem 2023 update *Nucleic Acids Res* **51** D1373-D1380

[70] Kirklin S, et al. 2015 The Open Quantum Materials Database (OQMD): assessing the accuracy of DFT formation energies *npj Comput Mater* **1** 15010

[71] Sha W, et al. 2021 Machine learning in polymer informatics *InfoMat* **3** 353-361

[72] Robeson L M 2008 The upper bound revisited *J Membr Sci* **320** 390-400

[73] Yang X, Song Z, King I, and Xu Z 2023 A Survey on Deep Semi-Supervised Learning *IEEE Transactions on Knowledge and Data Engineering* **35** 8934-8954

[74] Lee D-H 2013 Pseudo-Label : The Simple and Efficient Semi-Supervised Learning Method for Deep Neural Networks *Workshop on challenges in representation learning, ICML*

[75] Liu G, et al. 2023 Semi-Supervised Graph Imbalanced Regression *arXiv:2305.12087*
[76] Hu W, et al. 2020 Strategies for Pre-training Graph Neural Networks *arXiv:1905.12265*
[77] Kuenneth C and Ramprasad R 2023 polyBERT: a chemical language model to enable fully machine-driven ultrafast polymer informatics *Nat Commun* **14** 4099
[78] Bassman Oftelie L, et al. 2018 Active learning for accelerated design of layered materials *npj Comput Mater* **4** 74
[79] Kim S, et al. 2022 High-Performance Transparent Radiative Cooler Designed by Quantum Computing *ACS Energy Lett* **7** 4134-4141
[80] Raju L, et al. 2022 Maximized Frequency Doubling through the Inverse Design of Nonlinear Metamaterials *ACS Nano* **16** 3926-3933
[81] Zunger A 2018 Inverse design in search of materials with target functionalities *Nat Rev Chem* **2** 0121
[82] Li Y, et al. 2017 Machine Learning Force Field Parameters from Ab Initio Data *J Chem Theory Comput* **13** 4492-4503
[83] Deringer V L, Caro M A, and Csanyi G 2020 A general-purpose machine-learning force field for bulk and nanostructured phosphorus *Nat Commun* **11** 5461
[84] Huan T D, et al. 2017 A universal strategy for the creation of machine learning-based atomistic force fields *npj Computational Materials* **3**
[85] Unke O T, et al. 2021 Machine Learning Force Fields *Chem Rev* **121** 10142-10186
[86] Song Z, et al. 2020 Machine learning in materials design: Algorithm and application* *Chin Phys B* **29** 116103
[87] Zhou T, Song Z, and Sundmacher K 2019 Big Data Creates New Opportunities for Materials Research: A Review on Methods and Applications of Machine Learning for Materials Design *Engineering* **5** 1017-1026
[88] Liu Y, Esan O C, Pan Z, and An L 2021 Machine learning for advanced energy materials *Energy and AI* **3** 100049
[89] Butler K T, et al. 2018 Machine learning for molecular and materials science *Nature* **559** 547-555
[90] Liu Y, Zhao T, Ju W, and Shi S 2017 Materials discovery and design using machine learning *J Mater* **3** 159-177
[91] Moosavi S M, Jablonka K M, and Smit B 2020 The Role of Machine Learning in the Understanding and Design of Materials *J Am Chem Soc* **142** 20273-20287
[92] Chen C, et al. 2020 A Critical Review of Machine Learning of Energy Materials *Adv Energy Mater* **10** 1903242
[93] Shorten C and Khoshgoftaar T M 2019 A survey on Image Data Augmentation for Deep Learning *Journal of Big Data* **6** 60
[94] Chen C, et al. 2019 Graph Networks as a Universal Machine Learning Framework for Molecules and Crystals *Chem Mater* **31** 3564-3572
[95] Park J, et al. 2022 Prediction and Interpretation of Polymer Properties Using the Graph Convolutional Network *ACS Polym Au* **2** 213-222

[96] Ekström Kelvinius F, Armiento R, and Lindsten F 2022 Graph-based machine learning beyond stable materials and relaxed crystal structures *Phys Rev Mater* **6** 033801

[97] Sun M, et al. 2021 MoCL: Data-driven Molecular Fingerprint via Knowledge-aware Contrastive Learning from Molecular Graph *KDD '21* **2021** 3585-3594

[98] Liu G, et al., *Graph Rationalization with Environment-based Augmentations*, in *KDD '22*. 2022. p. 1069-1078.

[99] Aszemi N M and Dominic P D D 2019 Hyperparameter_Optimization_in_Convolutional_Neural_Network *(IJACSA) International Journal of Advanced Computer Science and Applications* **10** 269-278

[100] Steurer M, Hill R J, and Pfeifer N 2021 Metrics for evaluating the performance of machine learning based automated valuation models *J Prop Res* **38** 99-129

[101] Leshno M, Lin V Y, Pinkus A, and Schocken S 1993 Multilayer Feedforward Networks With a Nonpolynomial Activation Function Can Approximate Any Function *Neural Networks* **6** 861-867

[102] Hornik K, Stinchcombe M, and White H 1989 Multilayer feedforward networks are universal approximators *Neural Networks* **2** 359-366

[103] Wang Q, Ma Y, Zhao K, and Tian Y 2020 A Comprehensive Survey of Loss Functions in Machine Learning *Annals of Data Science* **9** 187-212

[104] Ruder S 2017 An overview of gradient descent optimization algorithms *arXiv:1609.04747v2*

[105] LeCun Y, Bengio Y, and Hinton G 2015 Deep learning *Nature* **521** 436-44

[106] Li Z, et al. 2022 A Survey of Convolutional Neural Networks: Analysis, Applications, and Prospects *IEEE Trans Neural Netw Learn Syst* **33** 6999-7019

[107] Sherstinsky A 2020 Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) network *Physica D: Nonlinear Phenomena* **404**

[108] Zhou J, et al. 2020 Graph neural networks: A review of methods and applications *AI Open* **1** 57-81

[109] Gong S, et al. 2023 Examining graph neural networks for crystal structures: Limitations and opportunities for capturing periodicity *Sci Adv* **9** eadi3245

[110] Park C W and Wolverton C 2020 Developing an improved crystal graph convolutional neural network framework for accelerated materials discovery *Physical Review Materials* **4**

[111] Sun Y and Hu W 2021 Novel machine learning framework for thermal conductivity prediction by crystal graph convolution embedded ensemble *SmartMat* **3** 474-481

[112] Vaswani A, et al. 2017 Attention Is All You Need *Advances in Neural Information Processing Systems 30 (NIPS 2017)*

[113] Zhong X, et al. 2022 Explainable machine learning in materials science *npj Comput Mater* **8**

[114] Lin T, Wang Y, Liu X, and Qiu X 2022 A survey of transformers *AI Open* **3** 111-132

[115] Niu Z, Zhong G, and Yu H 2021 A review on the attention mechanism of deep learning *Neurocomputing* **452** 48-62

[116] Li A, et al. 2022 Integration of Computational Fluid Dynamics and Artificial Neural Network for Optimization Design of Battery Thermal Management System *Batteries* **8**

[117] Kaya M and Hajimirza S 2018 Application of artificial neural network for accelerated optimization of ultra thin organic solar cells *Solar Energy* **165** 159-166

[118] Mayer A, et al. 2014 Multi-objective genetic algorithm for the optimization of a flat-plate solar thermal collector *Opt Express* **22 Suppl 6** A1641-9

[119] Lin A and Phillips J 2008 Optimization of random diffraction gratings in thin-film solar cells using genetic algorithms *Sol Energy Mater Sol Cells* **92** 1689-1696

[120] Patra T K, Meenakshisundaram V, Hung J H, and Simmons D S 2017 Neural-Network-Biased Genetic Algorithms for Materials Design: Evolutionary Algorithms That Learn *ACS Comb Sci* **19** 96-107

[121] Kim C, et al. 2021 Polymer design using genetic algorithm and machine learning *Comput Mater Sci* **186** 110067

[122] Ren H, et al. 2019 Optimal design and size of a desiccant cooling system with onsite energy generation and thermal storage using a multilayer perceptron neural network and a genetic algorithm *Energy Conv Manag* **180** 598-608

[123] Krishna K M, et al. 2022 Development of the Broadband Multilayer Absorption Materials with Genetic Algorithm up to 8 GHz Frequency *Secur Commun Netw* **2022** 1-12

[124] Zhou T, Wu Z, Chilukoti H K, and Muller-Plathe F 2021 Sequence-Engineering Polyethylene-Polypropylene Copolymers with High Thermal Conductivity Using a Molecular-Dynamics-Based Genetic Algorithm *J Chem Theory Comput* **17** 3772-3782

[125] Lourenco M P, et al. 2023 GAMaterial-A genetic-algorithm software for material design and discovery *J Comput Chem* **44** 814-823

[126] Shahriari B, et al. 2016 Taking the Human Out of the Loop: A Review of Bayesian Optimization *Proc IEEE* **104** 148-175

[127] Shang W, et al. 2023 Hybrid Data-Driven Discovery of High-Performance Silver Selenide-Based Thermoelectric Composites *Adv Mater* **35** e2212230

[128] Malakpour Estalaki S, Luo T, and Manukyan K V 2023 Bayesian optimization of metastable nickel formation during the spontaneous crystallization under extreme conditions *J Appl Phys* **133** 215901

[129] Lookman T, Balachandran P V, Xue D, and Yuan R 2019 Active learning in materials science with emphasis on adaptive sampling using uncertainties for targeted design *npj Comput Mater* **5** 21

[130] Nakano K, et al. 2020 Exhaustive and informatics-aided search for fast Li-ion conductor with NASICON-type structure using material simulation and Bayesian optimization *APL Mater* **8** 041112

[131] Chibani S and Coudert F-X 2020 Machine learning approaches for the prediction of materials properties *APL Mater* **8** 080701

[132] Saeidi-Javash M, et al. 2022 Machine learning-assisted ultrafast flash sintering of high-performance and flexible silver–selenide thermoelectric devices *Energy Environ Sci* **15** 5093-5104

[133] Zhang Y, Apley D W, and Chen W 2020 Bayesian Optimization for Materials Design with Mixed Quantitative and Qualitative Variables *Sci Rep* **10** 4924

[134] Kitai K, et al. 2020 Designing metamaterials with quantum annealing and factorization machines *Phys Rev Res* **2** 013319

[135] Kim S, et al. 2024 Quantum annealing-aided design of an ultrathin-metamaterial optical diode *Nano Converg* **11** 16

[136] Su J, Tu T, and He L, *A quantum annealing approach for boolean satisfiability problem*, in *Proceedings of the 53rd Annual Design Automation Conference*. 2016. p. 1-6.

[137] Johnson M W, et al. 2011 Quantum annealing with manufactured spins *Nature* **473** 194-198

[138] Rendle S, *Factorization Machines*, in *2010 IEEE International Conference on Data Mining*. 2010. p. 995-1000.

[139] Rendle S 2012 Factorization Machines with libFM *ACM Trans Intell Syst Technol* **3** 1-22

[140] Kim S, et al. 2023 Design of a High-Performance Titanium Nitride Metastructure-Based Solar Absorber Using Quantum Computing-Assisted Optimization *ACS Appl Mater Interfaces* **15** 40606-40613

[141] Wilson B A, et al. 2021 Machine learning framework for quantum sampling of highly constrained, continuous optimization problems *Appl Phys Rev* **8** 041418

[142] Kim S, et al. 2024 Wide-Angle Spectral Filter for Energy-Saving Windows Designed by Quantum Annealing-Enhanced Active Learning *Cell Rep Phys Sci* 101847

[143] Kousis I, D'Amato R, Pisello A L, and Latterini L 2023 Daytime Radiative Cooling: A Perspective toward Urban Heat Island Mitigation *ACS Energy Lett* **8** 3239-3250

[144] Cho J W, et al. 2023 Directional Radiative Cooling via Exceptional Epsilon-Based Microcavities *ACS Nano* **17** 10442-10451

[145] Wang J, Tan G, Yang R, and Zhao D 2022 Materials, structures, and devices for dynamic radiative cooling *Cell Rep Phys Sci* **3** 101198

[146] Li S, et al. 2023 Daytime radiative cooling multilayer films designed by a machine learning method and genetic algorithm *Appl Opt* **62** 4359-4369

[147] Guan Q, et al. 2023 Machine Learning-Enabled Inverse Design of Radiative Cooling Film with On-Demand Transmissive Color *ACS Photonics* **10** 715-726

[148] Felicelli A, et al. 2022 Thin layer lightweight and ultrawhite hexagonal boron nitride nanoporous paints for daytime radiative cooling *Cell Rep Phys Sci* **3** 101058

[149] Kim M, et al. 2021 Visibly Transparent Radiative Cooler under Direct Sunlight *Adv Opt Mater* **9** 2002226

[150] Lee K W, et al. 2021 Visibly Clear Radiative Cooling Metamaterials for Enhanced Thermal Management in Solar Cells and Windows *Adv Funct Mater* **32** 2105882

[151] Liu L, et al. 2020 Layered ternary metal oxides: Performance degradation mechanisms as cathodes, and design strategies for high-performance batteries *Prog Mater Sci* **111** 100655
[152] Ahmad Z, et al. 2018 Machine Learning Enabled Computational Screening of Inorganic Solid Electrolytes for Suppression of Dendrite Formation in Lithium Metal Anodes *ACS Cent Sci* **4** 996-1006
[153] Materials Project *https://next-gen.materialsproject.org/*
[154] Joshi R P, et al. 2019 Machine Learning the Voltage of Electrode Materials in Metal-Ion Batteries *ACS Appl Mater Interfaces* **11** 18494-18503
[155] Dave A, et al. 2020 Autonomous Discovery of Battery Electrolytes with Robotic Experimentation and Machine Learning *Cell Rep Phys Sci* **1** 100264
[156] Kim S, et al. 2022 Accelerated battery life predictions through synergistic combination of physics-based models and machine learning *Cell Rep Phys Sci* **3** 101023
[157] He K, Zhang X, Ren S, and Sun J 2016 Deep residual learning for image recognition *Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit (CVPR)* 770-778
[158] Chen X, Choi J, and Li X 2022 Machine Learning a Million Cycles as 2D Images from Practical Batteries for Electric Vehicle Applications *ACS Energy Lett* **7** 4362-4367
[159] Christians J A, Habisreutinger S N, Berry J J, and Luther J M 2018 Stability in Perovskite Photovoltaics: A Paradigm for Newfangled Technologies *ACS Energy Lett* **3** 2136-2143
[160] Snaith H J 2018 Present status and future prospects of perovskite photovoltaics *Nat Mater* **17** 372-376
[161] Case C, Beaumont N, and Kirk D 2019 Industrial Insights into Perovskite Photovoltaics *ACS Energy Lett* **4** 2760-2762
[162] Zhu H, et al. 2023 Long-term operating stability in perovskite photovoltaics *Nat Rev Mater* **8** 569-586
[163] Chen J, et al. 2022 Machine learning-driven design of promising perovskites for photovoltaic applications: A review *Surf Interfaces* **35** 102470
[164] Polman A, et al. 2016 Photovoltaic materials: Present efficiencies and future challenges *Science* **352** aad4424
[165] Yu Y, Tan X, Ning S, and Wu Y 2019 Machine Learning for Understanding Compatibility of Organic–Inorganic Hybrid Perovskites with Post-Treatment Amines *ACS Energy Lett* **4** 397-404
[166] Lu S, et al. 2019 Rapid Discovery of Ferroelectric Photovoltaic Perovskites and Material Descriptors via Machine Learning *Small Methods* **3** 1900360
[167] Stanley J C, Mayr F, and Gagliardi A 2019 Machine Learning Stability and Bandgaps of Lead-Free Perovskites for Photovoltaics *Adv Theory Simul* **3** 1900178
[168] She C, et al. 2021 Machine learning-guided search for high-efficiency perovskite solar cells with doped electron transport layers *J Mater Chem* **9** 25168-25177
[169] Yang W, et al. 2022 Exploring the Effects of Ionic Defects on the Stability of CsPbI(3) with a Deep Learning Potential *Chemphyschem* **23** e202100841

[170] Liu M, Seeger A, and Guo R 2023 Cross-Linked Polymer Membranes for Energy-Efficient Gas Separation: Innovations and Perspectives *Macromolecules* **56** 7230-7246
[171] Sidhikku Kandath Valappil R, Ghasem N, and Al-Marzouqi M 2021 Current and future trends in polymer membrane-based gas separation technology: A comprehensive review *J Ind Eng Chem* **98** 103-129
[172] Wessling M, et al. 1994 Modelling the permeability of polymers: a neural network approach *J Membr Sci* **86** 193-198
[173] Zhu G, et al. 2020 Polymer genome–based prediction of gas permeabilities in polymers *J Polym Eng* **40** 451-457
[174] Barnett J W, et al. 2020 Designing exceptional gas-separation polymer membranes using machine learning *Sci Adv* **6** eaaz4301
[175] Wilson A N, et al. 2023 PolyID: Artificial Intelligence for Discovering Performance-Advantaged and Sustainable Polymers *Macromolecules* **56** 8547-8557
[176] Gorai P, Stevanović V, and Toberer E S 2017 Computationally guided discovery of thermoelectric materials *Nat Rev Mater* **2** 17053
[177] Wang X, et al. 2023 A Critical Review of Machine Learning Techniques on Thermoelectric Materials *J Phys Chem Lett* **14** 1808-1822
[178] Snyder G J and Snyder A H 2017 Figure of merit ZT of a thermoelectric device defined from materials properties *Energy Environ Sci* **10** 2280-2283
[179] Zhu T, et al. 2021 Charting lattice thermal conductivity for inorganic crystals and discovering rare earth chalcogenides for thermoelectrics *Energy Environ Sci* **14** 3559-3566
[180] Li Y, et al. 2022 Large Data Set-Driven Machine Learning Models for Accurate Prediction of the Thermoelectric Figure of Merit *ACS Appl Mater Interfaces* **14** 55517-55527
[181] Gaultois M W, et al. 2013 Data-Driven Review of Thermoelectric Materials: Performance and Resource Considerations *Chem Mater* **25** 2911-2920
[182] Zhan T, Fang L, and Xu Y 2017 Prediction of thermal boundary resistance by the machine learning method *Sci Rep* **7** 7109
[183] Ong S P, et al. 2013 Python Materials Genomics (pymatgen): A robust, open-source python library for materials analysis *Comput Mater Sci* **68** 314-319
[184] Kumar N, Kim S B, Lee S Y, and Park S J 2022 Recent Advanced Supercapacitor: A Review of Storage Mechanisms, Electrode Materials, Modification, and Perspectives *Nanomaterials (Basel)* **12**
[185] Yan J, Wang Q, Wei T, and Fan Z 2013 Recent Advances in Design and Fabrication of Electrochemical Supercapacitors with High Energy Densities *Advanced Energy Materials* **4**
[186] Ghosh S, Rao G R, and Thomas T 2021 Machine learning-based prediction of supercapacitor performance for a novel electrode material: Cerium oxynitride *Energy Storage Materials* **40** 426-438

[187] Du Q, et al. 2021 3D hierarchical porous carbon matching ionic liquid with ultrahigh specific surface area and appropriate porous distribution for supercapacitors *Nanoscale* **13** 13285-13293
[188] Vinodh R, et al. 2020 A review on porous carbon electrode material derived from hypercross-linked polymers for supercapacitor applications *Journal of Energy Storage* **32**
[189] Wang T, et al. 2023 Machine-learning-assisted material discovery of oxygen-rich highly porous carbon active materials for aqueous supercapacitors *Nat Commun* **14** 4607
[190] Kranthiraja K and Saeki A 2022 Machine Learning-Assisted Polymer Design for Improving the Performance of Non-Fullerene Organic Solar Cells *ACS Appl Mater Interfaces* **14** 28936-28944
[191] Zhu M-X, et al. 2021 Rational Design of High-Energy-Density Polymer Composites by Machine Learning Approach *ACS Applied Energy Materials* **4** 1449-1458
[192] Lopez J, Mackanic D G, Cui Y, and Bao Z 2019 Designing polymers for advanced battery chemistries *Nature Reviews Materials* **4** 312-330
[193] Wu S, et al. 2019 Machine-learning-assisted discovery of polymers with high thermal conductivity using a molecular design algorithm *npj Comput Mater* **5**
[194] Chen L, et al. 2020 Frequency-dependent dielectric constant prediction of polymers using machine learning *npj Comput Mater* **6**
[195] Xu P, Ji X, Li M, and Lu W 2023 Small data machine learning in materials science *npj Comput Mater* **9** 42
[196] Ding K, Xu Z, Tong H, and Liu H 2022 Data Augmentation for Deep Graph Learning: A Survey *ACM SIGKDD Explor* **24** 61-77
[197] Gao Z, et al. 2021 Training Robust Graph Neural Networks with Topology Adaptive Edge Dropping *arXiv:2106.02892*
[198] Wang Y, et al. 2020 GraphCrop: Subgraph Cropping for Graph Classification *arXiv:2009.10564*
[199] Zheng W, et al. 2022 Machine learning for imbalanced datasets: Application in prediction of 3d-5d double perovskite structures *Comput Mater Sci* **209** 111394
[200] Gupta V, et al. 2021 Cross-property deep transfer learning framework for enhanced predictive analytics on small materials data *Nat Commun* **12** 6595
[201] Jha D, et al. 2019 Enhancing materials property prediction by leveraging computational and experimental data using deep transfer learning *Nat Commun* **10** 5316
[202] Cubuk E D, Sendek A D, and Reed E J 2019 Screening billions of candidates for solid lithium-ion conductors: A transfer learning approach for small data *J Chem Phys* **150** 214701
[203] Chen C, et al. 2021 Learning properties of ordered and disordered materials from multi-fidelity data *Nat Comput Sci* **1** 46-53
[204] Kailkhura B, et al. 2019 Reliable and explainable machine-learning methods for accelerated material discovery *npj Comput Mater* **5** 108

[205] Qiu C, et al. 2021 Nanophotonic inverse design with deep neural networks based on knowledge transfer using imbalanced datasets *Opt Express* **29** 28406-28415
[206] Liu B, et al. 2017 Retrosynthetic Reaction Prediction Using Neural Sequence-to-Sequence Models *ACS Cent Sci* **3** 1103-1113
[207] Coley C W, Green W H, and Jensen K F 2019 RDChiral: An RDKit Wrapper for Handling Stereochemistry in Retrosynthetic Template Extraction and Application *J Chem Inf Model* **59** 2529-2537
[208] Chen L, Kern J, Lightstone J P, and Ramprasad R 2021 Data-assisted polymer retrosynthesis planning *Appl Phys Rev* **8** 031405
[209] Xu C, Wang Y, and Barati Farimani A 2023 TransPolymer: a Transformer-based language model for polymer property predictions *npj Comput Mater* **9** 64
[210] Qiu H, et al. 2024 PolyNC: a natural and chemical language model for the prediction of unified polymer properties *Chem Sci* **15** 534-544
[211] Hase F, Roch L M, and Aspuru-Guzik A 2018 Chimera: enabling hierarchy based multi-objective optimization for self-driving laboratories *Chem Sci* **9** 7642-7655
[212] Messac A, Ismail-Yahaya A, and Mattson C A 2003 The normalized normal constraint method for generating the Pareto frontier *Struct Multidiscipl Optim* **25** 86-98
[213] Mueller-Gritschneder D, Graeb H, and Schlichtmann U 2009 A Successive Approach to Compute the Bounded Pareto Front of Practical Multiobjective Optimization Problems *SIAM J Optim* **20** 915-934
[214] Kim I Y and de Weck O L 2005 Adaptive weighted sum method for multiobjective optimization: a new method for Pareto front generation *Struct Multidiscipl Optim* **31** 105-116
[215] Eisele V and Schmitt-Landsiedel D 1991 Optimization and architectural evaluation of regular combinatoric structures *Microprocessing and Microprogramming* **32** 69-73
[216] Sattari K, Xie Y, and Lin J 2021 Data-driven algorithms for inverse design of polymers *Soft Matter* **17** 7607-7622
[217] Eugene E A, et al. 2023 Learning and optimization under epistemic uncertainty with Bayesian hybrid models *Comput Chem Eng* **179** 108430
[218] Wang K and Dowling A W 2022 Bayesian optimization for chemical products and functional materials *Curr Opin Chem Eng* **36** 100728