# *AI Applications through the Whole Life Cycle of Material Discovery*

**Citation:** Li, Jiali, Lim, Kaizhuo, Yang, Haitao, Ren, Zekun, Raghavan, Shreyaa et al. 2020. "AI Applications through the Whole Life Cycle of Material Discovery." Matter, 3 (2).

**As Published:** 10.1016/J.MATT.2020.06.011

**Publisher:** Elsevier BV

**Persistent URL:** https://hdl.handle.net/1721.1/138475

**Version:** Final published version: final published article, as it appeared in a journal, conference proceedings, or other formally published context

Matter

CellPress
OPEN ACCESS

Review

# AI Applications through the Whole Life Cycle of Material Discovery

Jiali Li,[1] Kaizhuo Lim,[1] Haitao Yang,[1] Zekun Ren,[2] Shreyaa Raghavan,[3] Po-Yen Chen,[1] Tonio Buonassisi,[2,3,*] and Xiaonan Wang[1,*]

**We provide a review of machine learning (ML) tools for material discovery and sophisticated applications of different ML strategies. Although there have been a few published reviews on artificial intelligence (AI) for materials with an emphasis on a single material system or individual methods, this paper focuses on an application-based perspective in AI-enhanced material discovery. It shows how AI strategies are applied through material discovery stages (including characterization, property prediction, synthesis, and theory paradigm discovery). Also, by referring to the ML tutorial, readers can acquire a better understanding of the exact functions of ML methods in each application and how these methods work to realize the targets. We are aiming to enable a better integration of AI methods with the material discovery process. The keys to successful applications of AI in material discovery and challenges to be addressed are also highlighted.**

## INTRODUCTION

New materials define the development of cultures, from the Stone Age to the present day.[1] We interact daily with many thousands of specialized materials as key parts of advanced technology and infrastructure. It is challenging to predict an accurate property-process-structure relationship to design new materials with specific property requirements quickly and precisely. The reasons can be attributed to three main aspects: (1) high dimensionality of features in material design including materials' intrinsic information (e.g., material crystallinity, size, energy level, hydrophobicity) and extrinsic synthesis processes' information (e.g., pH, reaction temperature, concentration); (2) the huge material design space containing a vast amount of possible materials that are difficult to select from (e.g., nanomedicine carriers selection,[2] catalyst ligands selection,[3,4] hybrid solar cells, wearable sensors[5,6]); and (3) the absence of completely known underlying physics and chemistry of complex material systems (e.g., active sites of catalyst,[7,8] molecular targets of drugs,[9] accurate control of high entropy alloy,[10] structure prediction of materials[11–13]). All the challenges are related to the complex material data management in analyzing, understanding, and predicting, which exceed human capability.

Different from experimental measurements of physical and chemical properties, computational material simulation methods (e.g., density functional theory [DFT], molecular dynamics [MD], coarse-grained MD) can calculate various materials' properties by performing simulations as opposed to actual material synthesis. Such methods are relatively faster at predicting materials' properties, yet are not always perfectly accurate. Before combining with machine learning (ML) methods, there are two generations of methods according to the historical development of research in computational materials science: (1) property-structure relationship calculation

### Progress and Potential

Advances in artificial intelligence (AI), especially machine learning (ML), provide enormous tools for processing complex data generated from experimental and computational materials research. With the rapid development of AI methods and the complex nature of interdisciplinary research, a challenge is posed as for which methods to choose for different material systems or context and which steps of the material discovery process would stand to benefit. This paper answers these questions by first introducing ML methods from a material study perspective in a tutorial section. We then discuss how AI can assist in each step through the whole life cycle of material discovery (including characterization, property prediction, synthesis, and theory paradigm discovery) by conducting a thorough literature review in the material application section. Finally, future research efforts should focus on in-depth understandings of descriptors, materials' ML methods, data-driven application strategies, and integration of studies.

and (2) property-structure-composition relationship calculation.[14] These two approaches have enabled the development of extensive databases with properties of organic and inorganic crystals, single molecules, and metal alloys.[15–17] However, the material synthesis conditions are not taken into consideration and the analytical and predictive capabilities of simulation methods depend on clear understanding of a material's underlying physics and chemistry. Besides, even the well-constructed simulation databases can include hypothetical structures that are not thermodynamically stable, and the methods used fail in many cases such as for random or disordered structures or non-zero-K temperatures. The limitations of computational material simulation methods have become apparent for newer material systems, which are significant for material discovery progress but are normally lacking first-principles understanding. Moreover, these simulation methods can be computationally expensive and take infeasibly long to screen vast material design spaces.

Apart from traditional laboratory experiments and computational material simulation approaches, artificial Intelligence (AI) could be an alternative approach that is able to address the material design challenges mentioned above. For example, ML methods have already managed to (1) automate materials' characterization processes and effectively analyze the characterization dataset,[18–21] (2) quickly screen the vast material design space (e.g., reducing the prediction time of DFT from $10^3$ s to $10^{-2}$ s),[22–25] (3) realize property prediction in complex material systems with limited first-principles understanding,[26] (4) directly map high-dimensional synthesis recipes to materials with desired properties,[27,28] and (5) extract generalizable scientific principles from various material systems.[27,29,30] The reason why AI is particularly apt in material design is due to its inherently strong capabilities in handling huge amounts of data as well as high-dimensional analysis. A single material type within a synthesis protocol could contain enormous intrinsic information, such as various physiochemical properties, chemical structure, and composition information. Moreover, with extrinsic synthesis condition information such as temperature, pH, and reaction time, the dimensionality of data can be even higher. All of such information may contribute to the final properties of materials. When combining this rich information of one data point with the high volume of past experimental data (owing to technological innovation in automation, robotics, and computer science), valuable information can be analyzed and discovered within a short period of time by applying suitable ML methods. Here, ML's strengths in analyzing large volumes of high-dimensional data are the key.

As the AI field advances at a rapid pace it poses a challenge regarding which methods to choose and apply for different material systems or context and which steps of the material discovery process would stand to benefit. There have been a few published reviews on AI for materials that provide an excellent overview of the state of the art.[31–34] However, different from those single material system or single method oriented reviews, this paper focuses on an application-based perspective on the latest developments in AI-enhanced material discovery. This approach, accompanied by a thorough ML tutorial, could inspire versatile ideas for materials and AI scientists to contribute to this highly interdisciplinary field.

In this review, by showing how AI can assist in each step through the whole life cycle of material discovery, we aim to provide useful information and links to further resources, enabling better integration of AI methods with the material discovery process. A tutorial of ML tools and their applications in material discovery is first introduced. Different applications of AI in various material discovery stages (including characterization, property prediction, synthesis, and theory paradigm discovery)

[1]Department of Chemical and Biomolecular Engineering, National University of Singapore, 4 Engineering Drive 4, Singapore 117585, Singapore

[2]Singapore-MIT Alliance for Research and Technology SMART, Singapore 138602, Singapore

[3]Massachusetts Institute of Technology, Cambridge, MA 02139, USA

*Correspondence: buonassi@mit.edu (T.B.), chewxia@nus.edu.sg (X.W.)

https://doi.org/10.1016/j.matt.2020.06.011

are then discussed in detail. Finally, a perspective of future directions within this emerging field is provided.

## MACHINE LEARNING TUTORIAL

### Introduction to ML

In this section, we review the most popular and commonly used ML algorithms in the fields of material science. Mitchell provides a definition of ML as "a computer program is said to learn from experience *E* with respect to some class of tasks *T* and performance measure *P*, if its performance at tasks in *T*, as measured by *P*, improves with experience *E*,"[35] which is used as the underlying framework of ML as follows.

#### *The Experience*

Most ML algorithms, based on the type of examples they are given during the training phase, can be classified into either unsupervised or supervised learning.[36] Unsupervised learning involves a dataset of examples with features (measurable properties of the object under analysis) $\boldsymbol{x}$ only and learns some meaningful relationship among the examples such as the probability distribution $p(\boldsymbol{x})$. Supervised learning involves a dataset of examples with features $\boldsymbol{x}$ and corresponding labels $\boldsymbol{y}$, which is the "correct" set of values that are associated with the features. Here, the learning algorithm learns the probability distribution $p(\boldsymbol{y}|\boldsymbol{x})$ or the expected value of a regression $\mathrm{E}(\boldsymbol{y}|\boldsymbol{x})$. There is another main category called reinforcement learning (RL) whereby the agent learns by interacting with an environment to obtain reward feedback. However, RL faces many challenges especially the difficulty of constructing interactive environments with fast feedback. Therefore, it is not as yet widely adopted in material discovery, and this review limits its scope to supervised and unsupervised learning.

#### *The Task*

ML can be applied to various tasks.[36] The task usually involves the processing of examples (or data points) given to the algorithm. The examples contain features/descriptors (e.g., chemical structures, chemical compositions, pH, reaction temperature), which are usually arranged as a vector $\boldsymbol{x}$ to describe a material's physiochemical properties, structural properties, composition properties, or synthesis processes' conditions. The various tasks that can be commonly solved using ML are as follows.

*Classification.* The algorithm is tasked to determine the category or class to which a particular example belongs. This is done by learning a function $f:\mathbb{R}^n \rightarrow \{1, \ldots, m\}$, which maps the feature vector $\boldsymbol{x}$ to one particular class out of $m$ distinct classes. Instead of deciding on one class only, the function can also give a probability distribution over all classes, where each entry in the output vector $\boldsymbol{y} = f(\boldsymbol{x})$ is the probability that the example belongs to a certain class. ML algorithms have been successfully applied to solve material classification tasks. For example, when given a set of synthesis conditions, a classification model can predict whether synthesized materials will be successfully formed[37,38] or which parts of synthesized materials will contain flow defects.[39]

*Regression.* Another common task is regression, whereby the algorithm aims to learn a function $f:\mathbb{R}^n \rightarrow \mathbb{R}$, which determines a continuous value $y$ or a set of continuous values expressed as a vector $\boldsymbol{y}$. One common use of regression models is the prediction of materials' properties such as heat capacity of inorganic solids,[40] Debye temperature,[41] and band gaps of materials.[42]

*Clustering.* This is a form of unsupervised learning task that is useful when there is a large amount of unlabeled data.[43–45] The goal is to organize data points into clusters where items within the same cluster are more "similar" to each other as compared with another cluster. What defines "similar" is dependent on the context and requirements. By organizing items into clusters, one can gain meaningful insight into the data even when the dataset is unlabeled. The clustering methods have been applied to material discovery in material text data mining[46] and analysis of microstructural images.[47]

*Dimension Reduction and Visualization.* When training a model, sometimes we might have too many features. In such scenarios, it would be beneficial to reduce the dimension of the features by mapping them into a lower dimension while preserving as much information as possible.[48,49] This will aid in improving computational efficiency, possibly improving models' performance, preventing overfitting, and helping discover insights for the specific tasks. Dimension reduction has been used in material discovery to improve the results of prediction, such as summarizing full long-time dynamic information into lower-dimension information to achieve better performance.[50,51] In addition, by reducing less relevant features, dimension reduction can be used to discern the underlying physics/chemistry of a material model.[52] Moreover, mapping of high-dimensional data to two-dimensional (2D) or three-dimensional (3D) plots for visualization purposes is also an important function of dimension reduction, as it allows useful insights to be derived from a comprehensible plot. It has been utilized in material discovery to visualize the high-dimensional material design space.[53]

*Efficient Searching.* Carrying out simulations and laboratory experiments to obtain extra data is time-consuming and costly. Efficient searching methods can help identify the most informative extra data points to label and thus minimize the overall efforts in data acquisition. Efficient searching has been incorporated in the material design of ferroelectric perovskites, layered materials, and the MAX ternary carbide/nitride (layered, hexagonal carbides and nitrides that have the general formula $M_{n+1}AX_n$), among others.[26,54,55]

Overall, the identification of which task to be solved is the first step in successfully applying ML in material discovery. However, these five categories of tasks do not always stand independently of each other. For example, efficient searching can be combined with classification or regression to form a closed design loop to search for high-temperature ferroelectric perovskites,[26] and dimension reduction can be used to construct a better feature set to train a classification model of membrane activity.[52]

### The Performance Measure

The performance measure is used to assess an algorithm's performance on a specific task.[36] For classification, the performance of the algorithm can be measured based on the accuracy (proportion of correct output), the error rate (proportion of incorrect output), or a more complex metric based on the confusion matrix such as the Matthews correlation coefficient.[56] For algorithms that output a probability distribution, the log probability can be calculated. For regression, it is common to use mean square error or some forms of norm error. The dataset for a specific task is normally separated into three parts, i.e., training dataset, validation dataset, and test dataset. The algorithm will be trained on the training dataset and further optimized (by tuning hyperparameters such as learning rate and structure of the algorithm model) according to the performance on the validation dataset. The performance of the optimized

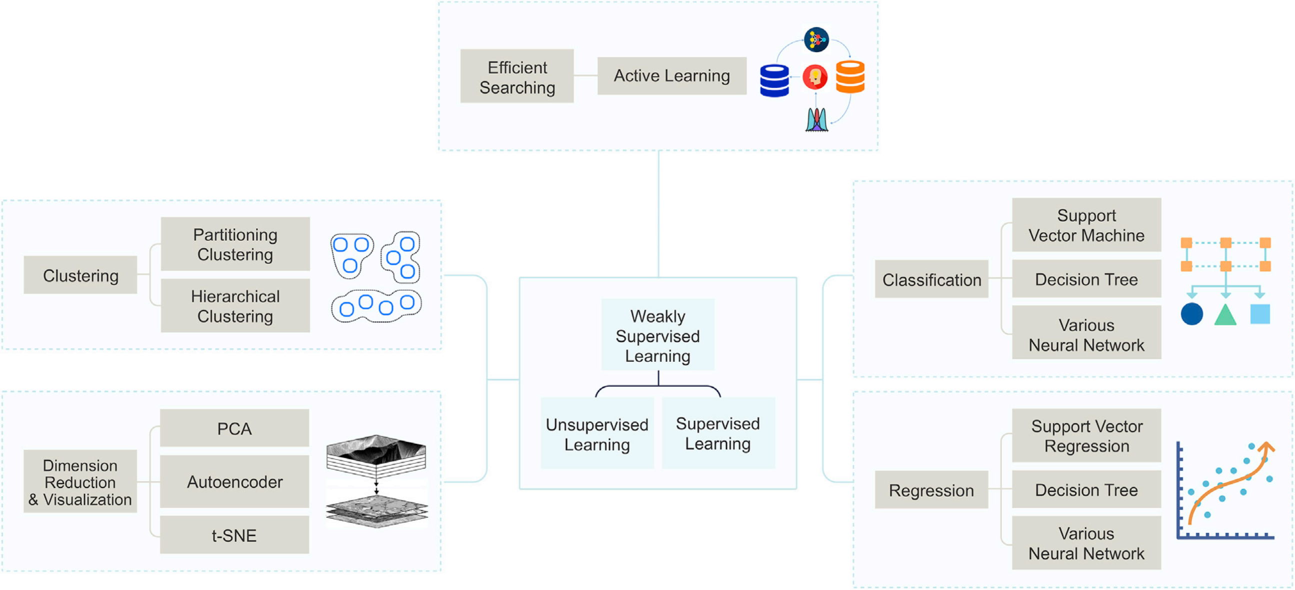

**Figure 1. The Structure of Machine Learning Methods Reviewed in The Tutorial**

supervised learning algorithm is usually measured based on its ability to perform on the test dataset, which is called the test performance and acts as a representation of how well the model can generalize. With respect to unsupervised learning, the clustering procedure can be measured by external and internal indices. External indices need a priori known clustering structure, i.e., they need to have the true label for data in the dataset.[57] For internal indices, the result is evaluated using quantities and features inherent to the dataset.[58] An example is the Calinski-Harabasz index that measures within-cluster coherence and between-cluster isolation.[57] An overview of the metrics and an efficient tool is provided by Wang et al.[59] For dimension reduction, the performance is usually evaluated via loss of quality from the original data (such as cumulative percent variance and the variance of reconstruction error).[60]

In the next section, several models useful for the materials field are introduced. While there have been extensive efforts made in the perpetual search for a "better" model, there is no universally superior algorithm for all problems.[36] Hence, the goal is not to find a universally superior algorithm but to find an algorithm that is best suited for a particular problem. For example, by using a suitable model or imposing different suitable priors (e.g., feature engineering, task splitting) on an ML algorithm, we can develop a model that performs significantly better on a particular problem.[36]

## Model Details

The type of ML to use is context-dependent, and it is important that the correct model is chosen for the appropriate problem to achieve desirable performance without under- or overfitting. Here, several ML models that are frequently used in the field of AI-assisted material discovery are introduced and described. The introduction of these models is sequenced within a supervised learning-unsupervised learning-weakly supervised learning framework. The whole framework is illustrated in Figure 1.

### *Supervised Model*

In a supervised learning model, the model is first trained with a labeled dataset that contains $N$ training examples. For example, the $i$th training example is a pair of two

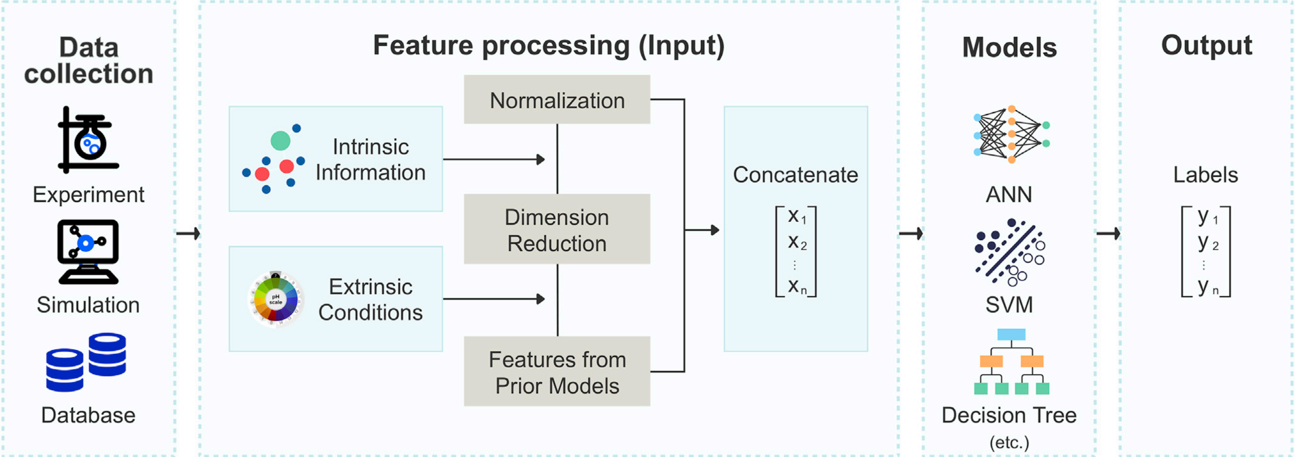

Figure 2. The Framework of Supervised Learning with Two Main Categories of Features

vectors for the features and labels $(\mathbf{x}_i, \mathbf{y}_i)$. The model's goal is to learn the function $f : \mathbf{x} \rightarrow \mathbf{y}$. Which features to feed into the model is extremely crucial and dependent on the context of the problem. Figure 2 shows a general workflow of supervised learning in material discovery. The dataset can be constructed from laboratory experiments, simulations, or existing databases. There are two categories of features commonly used in a material discovery context, namely the intrinsic and extrinsic information of the material system. The intrinsic information is based on the properties of the materials/chemicals used, while the extrinsic information concerns the environment in which the materials exist. The two types of features can be directly used as raw features. However, some further transformations will normally be applied to the raw features before feeding them into the predictive model. The first possible pre-processing transformation is normalization to change the values of features to a common scale. Another possible pre-processing transformation is dimension reduction to reduce the number of features, which is used when there are too many different features, especially compared with the number of labeled training examples. In such a case, it might be desirable to reduce the dimensionality of the features to prevent overfitting. Here, techniques such as principal component analysis (PCA) and variational autoencoders (VAEs) can be applied to reduce the dimensionality of the feature vector, possibly allowing improved model performance.[61,62] Those tools aim to provide the function $f(\mathbf{x}_i) = \tilde{x}_i$, where $\tilde{x}_i$ with reduced dimension is used as an information-rich feature vector for the predictive model. Apart from these general pre-processing techniques, domain-specific transformation can be carried out with a domain-specific prior model. The prior model processes some of the raw features, and its output is then fed to the final predictive model. For example, the transformation of various chemical structures into fixed-length fingerprint vectors can help achieve better predictive accuracy in molecular properties prediction tasks.[63] After these processes, different processed and unprocessed features can be concatenated to form the final feature input vector. This step of deciding which features to use might be the most critical step in ensuring good model performance. If the features provided to the model are inappropriate, no matter how good a model may be, the performance will be poor. After feature processing, different supervised models can be used for the prediction of the final design tasks. The commonly used supervised learning models in material discovery are introduced, and the outputs are normally a series of interested properties (label $y$) that can be represented in a vector form.

*Support Vector Machines.* Support vector machines (SVMs) are efficient classifiers based on Vapnik-Chervonenkis theory or statistical learning theory.[64] An SVM has two main principles:[65,66] First, the pre-processing step maps the input space to a high-dimensional feature space $\Phi : \mathcal{X} \rightarrow \mathcal{F}$, where $\Phi$, $\mathcal{X}$, and $\mathcal{F}$ are the map, input space, and feature space, respectively. Second, the best separating linear hyperplane is found in the feature space. This is possible since the map $\Phi$ can be nonlinear and high-dimensional, making linear separation possible. Also, the hyperplane in $\mathcal{F}$ can be nonlinear in $\mathcal{X}$. The training of an SVM can be computationally efficient, as it involves solving a convex minimization problem. Moreover, the use of the kernel trick where $K(x, x') = \langle \Phi(x), \Phi(x') \rangle$ ($x$ and $x'$ are two different input from $\mathcal{X}$) gives the dot product of the transformed inputs without requiring the explicit map $\Phi$, and thus the convex problem can be efficiently solved. A detailed tutorial on SVM has been given by Smola and Schölkopf.[65]

SVM has been successfully applied in much materials science research. It has been used for classifying whether relaxation in glassy liquids will or will not happen,[67] the emitted fluorescence color bands of silver clusters,[68] and the outcomes of materials' reactions for high-throughput screening purposes.[69] Other than classification tasks, SVM can successfully perform regression tasks and be renamed as support vector regression (SVR). This is used in tasks such as predicting the Debye temperature of an inorganic phosphor host[41] or gas adsorption of metal-organic frameworks (MOFs).[23]

*Decision Trees.* Decision trees (DTs) are sequential logic-based models. A comprehensive overview of DTs is provided by Murthy,[70] and more recent advances with different algorithms can also be found.[71] DTs provide a sequence of logical rules that test an object's features. In a DT, each node represents the feature that is being classified, and the branches that extend downwards from the parent node represent the values that the parent node's feature can take. The parent node is then split recursively into child nodes, with each of their features being tested until a stopping criterion is met, e.g., the maximum depth of the tree is reached. However, DT is prone to overfitting. When a tree is deep and each child node only has a small amount of data, the generalizability of the DT model will be poor. Much work in DT has been conducted in finding efficient algorithms that also minimize the size of the tree to prevent overfitting.[71] Another way to prevent overfitting is to use an ensemble method, e.g., random forest (RF)[72] and gradient boosting (GB).[73] These two methods make predictions based on a combination of outputs of separately trained DTs and they only differ in the way that the collection of DTs is constructed. RF trains each DT independently with a random sample of the original data. The random samples can be obtained by bagging, whereby a random number of samples is drawn with replacement. To reduce the computational difficulty, each DT can only choose a random sample of the whole feature set. A prediction is made from an RF model on an input vector $\boldsymbol{x}$ by taking the average over the individual models (regression) or the class with the most votes (classification). On the other hand, GB builds base learners (the individual model, such as DT or SVM, which makes up the ensemble) sequentially.[74] At each iteration, the next model's parameter is chosen to be most correlated with the negative of the gradient of the loss function. Hence, each additional base learner improves the ensemble's performance. It has been shown that GB will always have better performance than RF if the parameters are finely tuned.[74] However, GB is generally much slower than RF and more sensitive to noisy data.

One of the core strengths of DT is its interpretability. DT has been applied to discovery rule-like information in material discovery processes. Success and failure synthesis experiments of templated vanadium selenites[27] and gold nanoclusters[37] were used to train predictive models as the first step. DTs are built on top of the predictive models to provide interpretability for the domain experts and to allow the understanding of which reaction conditions are important and how they affect the synthesis results.

*Artificial Neural Networks.* The development of the artificial neural network (ANN) has been loosely inspired by neurons in biological neural networks.[75,76] The ANN tries to approximate an unknown function $f^*(\mathbf{x})$ by learning $y = f(\mathbf{x};\theta)$. In a single-layer feedforward ANN, the input $\mathbf{x}\in\mathbb{R}^{n\times1}$ is fed to a layer of hidden units with $m$ nodes. The single layer of nodes has a weight matrix, $W\in\mathbb{R}^{n\times m}$, and a bias vector, $b\in\mathbb{R}^{m\times1}$. The output of the hidden unit is $W^{\mathrm{T}}\mathbf{x} + b$. To be able to approximate nonlinear functions, a nonlinear activation function is applied to the output of the hidden unit, resulting in the final output of the first hidden layer being

$$h^{(1)}(\mathbf{x}) = g\left(W^{(1)^{\mathrm{T}}}\mathbf{x} + b^{(1)}\right). \qquad \text{(Equation 1)}$$

There are many different activation functions, the most commonly used being the rectified linear units (ReLU),[76] where $g(z) = \max\{0, z\}$. A single-layer feedforward ANN has been shown to be a universal approximator, which is able to approximate any function given enough hidden units.[77] However, this theorem does not state how many hidden nodes are required, and in the worst case it will require a huge number of nodes. Hence, deep neural networks (DNNs) with more than one hidden layer are often employed. DNNs or multilayer perceptrons (MLPs) are able to represent the same function as a single-layer ANN but with much fewer nodes, and hence are much more computationally efficient to train and deploy. The output of an MLP is

$$y = h^{(d)}\left(h^{(d-1)}\left(\ldots\left(h^{(1)}(\mathbf{x})\right)\right)\ldots\right). \qquad \text{(Equation 2)}$$

An MLP expresses the prior belief that the function we are trying to learn, $f^*$, can be expressed as a composition of several simpler functions.[78] Informally, each layer of the MLP learns an intermediate representation that is "easier" for the next layer to make use of, allowing the learning of more complicated functions. A possible drawback of MLP is that if the layers are too numerous, it might be difficult to train the deep network.

MLP would be a suitable choice for constructing a predictive model where there is sufficient labeled data normally much larger than simpler ML approaches (e.g., SVMs), depending on how "difficult" the task is. For example, MLP has been used to accurately predict X-ray pulse properties from a free-electron laser[79] and predict the glass-transition temperature of multicomponent oxide glasses.[80] Also, MLPs are often the building blocks of other more complicated models, such as being stacked after a preceding convolutional neural network (CNN) model.

*Convolutional Neural Networks.* CNNs are a specialized form of neural network that are especially suited for processing grid-like data such as 2D or 3D images and audio waveforms.[81,82] CNNs are adopted in processing data while maintaining local spatial relationships that are inherent in data with grid-like topology. Hence, a CNN should be used when the data can be represented as a grid-like structure and the local information inside the data is important.

The architecture of a CNN generally consists of convolutional layers and pooling layers, which are grouped in modules. Normally, grid-like data are input directly to the network and will be processed by several stages of convolution and pooling.

Thereafter, representations from these operations can be used for further steps such as input into an ANN for a classification task.[81] The convolutional layers work as a feature extractor. The feature representations are learned by applying filters and activation functions to the input grid-like data. The $k$th output feature map $Y_k$ can be calculated by the formula

$$Y_k = g\left(W_k * \mathbf{x}_{\text{grid}}\right), \qquad \text{(Equation 3)}$$

where $\mathbf{x}_{\text{grid}}$ is input grid-like data; $W_k$ is the convolutional filter related to the $k$th feature map; the multiplication sign is the 2D convolutional operator, which is used for calculation of inner product of the filter model at each location of the grid-like input; and $g(\bullet)$ is the nonlinear activation function.[83] Besides, the pooling layer is aiming to reduce the spatial size of the representation to reduce the number of parameters and computation in the full network. It can have various forms, the most common of which is max pooling, which is done by applying a max filter to subregions of the initial representation.

Some interesting applications of CNN in material discovery include the use of CNN to robustly classify crystal structures despite the presence of defects,[84] fast extraction of atomic species and type of defects from atomically resolved images,[85] and prediction of the molecular structures directly from atomic force microscopy (AFM) images.[86]

*Graph Neural Networks.* Data might be of a form that cannot be represented in an ordered grid-like fashion. Often graphs, either cyclic or acyclic, directed or undirected, are the most natural representation of certain data. Sequential data through time can be regarded as an acyclic directed graph,[87] and molecules are often best represented as an undirected graph with atoms as the nodes and bonds between atoms as the edges.[22,63] Working directly with graphs as the input allows preservation of the graph topology information, which might convey important dependencies between the various nodes and edges. Graph neural networks (GNNs) have been especially successful in processing molecular structures using a special family of neural networks called the message-passing neural networks (MPNN).[22] MPNNs work with two phases: first, the message-passing phase that acts on each node and runs for $T$ time steps, followed by a readout phase, which converts the final state of all the nodes into a feature vector that summarizes the final state of the entire graph.

GNNs and their variants have been successfully applied to problems where the features or part of the features are graph-like inputs (molecules such as drug molecules[88] or reactant molecules[37]). The GNN is usually placed before the predictive model and acts as a feature pre-processing step to map the variably sized graph into a fixed vector representation. The fixed vector representation is then passed on to the predictive model, usually a DNN of a certain form, for further processing.

*Recurrent Neural Networks.* Recurrent neural networks (RNNs) are useful for analyzing sequential data.[89,90] Sequential data points (e.g., video, audio, DNA sequence) are related to each other across time or space and are not independent. Hence, standard DNNs are not suitable for such tasks.[89] RNNs are able to selectively store information and pass them forward across the sequence. The RNN has hidden layers that store a hidden state for each node at every time step, which provides a "lousy" summary of the past sequence that has occurred.[90] It is "lousy" in the sense that it creates a map from an arbitrarily long sequence of past information to a fixed-length vector representing the hidden states. However, this still allows the RNN to possibly learn a way to encode past information into a rich enough hidden state vector, making it perform well for sequential data. Despite its benefits, the RNN has some problems such as difficulties in optimizing and learning long sequences of data.[90] Thus, many variants of the basic RNN have been developed, such as long-short-term-memory (LSTM)[89,90] cells and gated recurrent units.[90,91] All these variants aim to perform the same task of learning, representing, and detecting key features in the sequence of data that is useful for performing the task at hand, e.g., supervised learning to learn the distribution $p(y|x)$.

RNNs have been successfully applied in analyzing DNA sequences, taking advantage of high-throughput genomic sequencing data to train a hybrid CNN and bidirectional LSTM model (an RNN that can process data in both forward and backward direction which is useful for sequential data that varies across space rather than time) framework that outperforms traditional benchmarks.[92]

### *Unsupervised Model*

Unsupervised models are trained on an unlabeled dataset with the goal of gaining useful insights from the data. Unsupervised learning tasks involve the learning of the underlying probability distribution that generates the dataset, which mainly includes clustering or dimension reduction. In this section, we focus on these two tasks.

*Clustering Algorithms.* There are many different clustering methods, each with their own intended purpose. The two most common forms of clustering are partitioning clustering and hierarchical clustering.

Partitioning clustering aims to partition the items into $k$ clusters.[43,44] Crisp clustering ensures that each item belongs to only one class with an output of {0,1} and fuzzy clustering allows items to belong to clusters in various degrees with values [0,1]. Common methods of partitioning clustering include k-means clustering and k-medians clustering.

Hierarchical clustering aims to show how clusters relate to each other via tree-like structures.[43,45] The tree can be built downward, splitting the top cluster into smaller and smaller clusters (divisive algorithms), or upward by starting with many small clusters and combining them (agglomerative algorithms). Hierarchical clustering provides clusters for a range of values of $k$ (number of clusters) while partitioning clustering only gives one set of clusters of a single value of $k$. Thus, hierarchical clustering can help us understand how smaller clusters are related to larger ones.

k-means has been used to reduce many solar spectral sets into a few characteristic sets to efficiently design solar cells.[93] Moreover, the unsupervised word2vec clustering model is used to capture latent knowledge from materials science literature.[46]

*Principal Component Analysis.* PCA is a common unsupervised learning method used for dimension reduction.[48] It finds a set of linear combinations of the original variables (features), which are called the principal components (PCs). The PCs are orthogonal to one another with the first PC having the highest variance (largest eigenvalue) across the data, followed by the second and so forth. Usually, if the data are $N$ dimensional, $M$ number of PCs are selected where $M < N$. Thus, when the original $N$ dimensional data are projected onto the orthogonal $M$ dimensional basis, the dimension is reduced while maintaining most of the informative variation in the data.[94]

PCA has been used to extract statistically significant information from the observed polar vortices of $PbTiO_3/SrTiO_3$ superlattices[95] and to select useful features in the prediction of higher-selectivity catalysts.[96]

*Autoencoders.* Autoencoders are a special case of feedforward neural networks.[97] Unlike supervised learning where the neural network aims to predict $y$ given $x$, autoencoders aim to learn the identity function. It has two parts, the encoder function $e = f^e(\mathbf{x})$ and the decoder function $r = f^r(e)$. An undercomplete autoencoder is where the dimension of $e$ is less than $\mathbf{x}$ and the latent space acts as an information bottleneck. Hence, the encoder must learn to capture the most important aspects of the features $x$ and encode it in a lower-dimension latent space. The autoencoder is trained by minimizing the following loss function:

$$L(\mathbf{x}, f^r(f^e(\mathbf{x}))). \qquad \text{(Equation 4)}$$

After training, the encoder can be used as pre-processing on the input to give $e = f^e(x)$, after which $e$ is fed into the actual classification or regression model, such as a standard ANN. By reducing the dimensionality of the inputs, the performance tends to improve because the salient aspects of the inputs are captured in the latent space after being encoded.

The autoencoder can be made to have some properties of interest such as sparsity, in which a sparse regularization term is added.[98,99] Also, the autoencoder can be made to be more robust against noise by adding a noise or corruption term, $C(\widehat{\mathbf{x}}|\mathbf{x})$. This is a denoising autoencoder,[100] whereby the autoencoder aims to reconstruct $x$ from a noisy $\widehat{x}$. Another interesting variant is the variational autoencoder (VAE),[101] which aims to learn the true underlying probability distribution of $P^*(X)$. The VAE is trained by drawing some examples $X$ from $P^*$ and trying to learn a distribution $\widehat{P}$ that is as close to $P^*$ as possible. By doing so, we can obtain a generative model that can generate new samples similar to true examples.

Autoencoders have been successfully applied in tasks such as material property prediction[102] and screening of inorganic material synthesis parameters,[103] where the information-rich latent space enables enhanced prediction and generalization capabilities. Moreover, using the generative models that VAEs produce,[102] it is possible to generate novel chemical molecules that have some desired properties by sampling from the latent space and decoding back into a simplified molecular input line entry system (SMILES) representation of the molecule.

*t-Distributed Stochastic Neighbor Embedding.* t-Distributed stochastic neighbor embedding (t-SNE) is a popular unsupervised learning method for dimension reduction and data visualization, which has two stages.[104] First, in the original high-dimensional space it constructs a probability distribution over how likely two pairs of points would be picked. Points that are closer together (e.g., with smaller Euclidean norm) have higher probabilities and vice versa. Second, in the lower-dimensional space, a probability distribution is specified over all points. The algorithm then aims to minimize the Kullback-Leibler divergence, ensuring that the probability distributions in the high- and low-dimensional space are similar to one another. Hence, this allows the mapping of points in high-dimensional space to a lower one. t-SNE is the current state-of-the-art method for data visualization, replacing the traditional PCA. For example, it has been useful in visualizing all adsorption sites of electrocatalysts simulated with DFT.[53]

### Weakly Supervised Model

In supervised learning, a dataset $D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$ is available to train a model with. However, in a weakly supervised setting,[105] there may only exist a small subset of labeled data, with the vast majority being unlabeled, where $D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_l, y_l), x_{l+1}, x_{l+2}, \ldots, x_n\}$. Here, there is only $l$ labeled data and $u = n - l$unlabeled data usually with $u \gg l$. For material discovery, the unlabeled data points here correspond to the search space in an experiment. Since training a good predictive model, such as a neural network, would require sufficient and usually a large amount of data, extensive experiments need to be performed to label these data points. Weakly supervised learning methods can be used to tackle this challenge by training the model using a small set of labeled data.[105]

*Active Learning.* Active learning is the most frequently used weakly supervised learning method in the materials field. The main hypothesis in active learning is that the algorithm can choose the data it wants to learn from and thus requires substantially fewer data for training. Instead of performing a large number of experiments (label all data points), active learning allows a more efficient search by finding a balance between exploration and exploitation, thus reducing the number of expensive and time-consuming experiments that need to be performed. This is done by either selecting examples that are informative and/or representative.[106] Informative examples are those examples which, when labeled, are predicted to best reduce the model's uncertainty, whereas representative examples are those that can well represent the overall input distribution of unlabeled data.

Active learning methods tackle the challenge of efficient searching by mainly two steps: (1) constructing a surrogate model by using the available $l$ labeled data—this surrogate model can be any supervised model mentioned before; (2) constructing a reward function to choose the next data points from which to learn. The reward function can also be called an acquisition function, which can be customized according to specific purposes. Usually, the reward function is constructed to trade off exploration and exploitation based on the predicted value and its uncertainty of the surrogate model. Detailed acquisition functions that may be useful for material discovery have been reviewed by Lookman et al.[31]

There are two main categories of active learning in the materials field. The first category is to find the maximum of an unknown complex objective function (optimum properties of materials) with minimum trials. There are several well-known methods for this category (e.g., Bayesian global optimization, efficient global optimization). Usually the predictive model created by this optimization strategy is more focused on performing well in the region near the global optimum point.[107] For material discovery, instead of mapping the whole material design space, the objective here is to find the global optimum material properties with a minimum number of experimental trials. The second category is more focused on improving the overall performance of a predictive model for investigating not only the best property but also the whole material design space (e.g., query by committee method). By choosing examples that are the most informative/representative for annotation, it provides a more efficient route to labeling unlabeled data points as compared with passive learning (random sampling of unlabeled points). A tutorial of active learning from the computer science perspective has been given by Settles,[106] and a material community-focused review has been published by Lookman et al.[31]

Single- and multiple-objective Bayesian optimization strategies are used to find the composition and structure with optimized materials' properties of layered materials[54] and precipitation-strengthened NiTi shape memory alloys[108] or find the suitable compositions for producing ferroelectric perovskites with highest Curie temperatures where the initial number of compounds with data collected was small.[26] The optimum conditions that produced the desired material could be found with a minimal number of newly labeled data. Also, the query by committee ranking strategy is used to develop accurate and transferrable potentials for predicting molecular energetics.[109]

Overall, each model has its strengths and weaknesses. The key is to choose a suitable model for a specific problem. The applications of these methods in different material discovery stages are discussed in the section Advancement of Material Discovery Assisted by AI.

## ADVANCEMENT OF MATERIAL DISCOVERY ASSISTED BY AI

There are four major parts in the material discovery process: characterization, property prediction, synthesis, and theory paradigm discovery. The first stage, characterization, focuses on ''seeing'' and collecting information about a material and forms the base from which the subsequent three stages of the material discovery arise.

During characterization (especially for imaging and mainly related to instruments), there is an enormous amount of grid-like high-dimensional data (e.g., spectroscopic data and microscopic data), the analysis of which is beyond the capability of human conceivability. AI can aid this scope by automating and enhancing the characterization process, leading to a reduction in manual work, improvement of data quality, and discovery of useful hidden information from the high-dimensional data.

The property prediction aspect focuses on finding a final material with desired or set of desired functionalities. Normally, the information used as input (features) for the predictive model is intrinsic property information of material components, such as the composition, structure, and physiochemical properties, to predict material functionalities. AI-aided property prediction work can be categorized into two main parts. On one hand, for material systems with known first-principle theories, AI can be used to substitute the computational expensive simulation process, actively guide the whole simulation process, or improve the accuracy of the first-principle simulations. On the other hand, for systems with unknown first principles, AI can be used to realize property prediction by efficiently learning from past experience, especially laboratory data, to construct a predictive model.

The synthesis studies aim to search for one or more successful synthesis pathways to form a final material with the desired properties. Here, both intrinsic property information and extrinsic information about the materials' synthesis environment (i.e., synthesis conditions such as pH, the ratio of different reactants, temperature, and time) are used as features in the prediction of the desired synthesis recipes. The synthesis studies to directly connect the synthesis recipe to the desired material functionalities represent one step further than the property prediction.

Finally, the theory discovery aims to discover new phenomena and extract generalizable scientific principles. These new mechanisms could be valuable rules in guiding the design of new materials of a broader range. AI-aided theory paradigm discovery is potentially the most powerful aspect, whereby contributing new mechanisms or efficient design rules to an advanced material system sometimes can trigger a revolution in the understanding of the field.

Although property prediction and synthesis studies are usually carried out before the characterization process in a real experimental procedure, the availability of large volumes of high-dimensional image data, with standard data-augmentation methods (e.g., by rotation and random horizontal/vertical shifts of images) and the mature processing pipeline for image data, make the field of AI in characterization (especially for imaging) more mature. In addition, the accurate and large

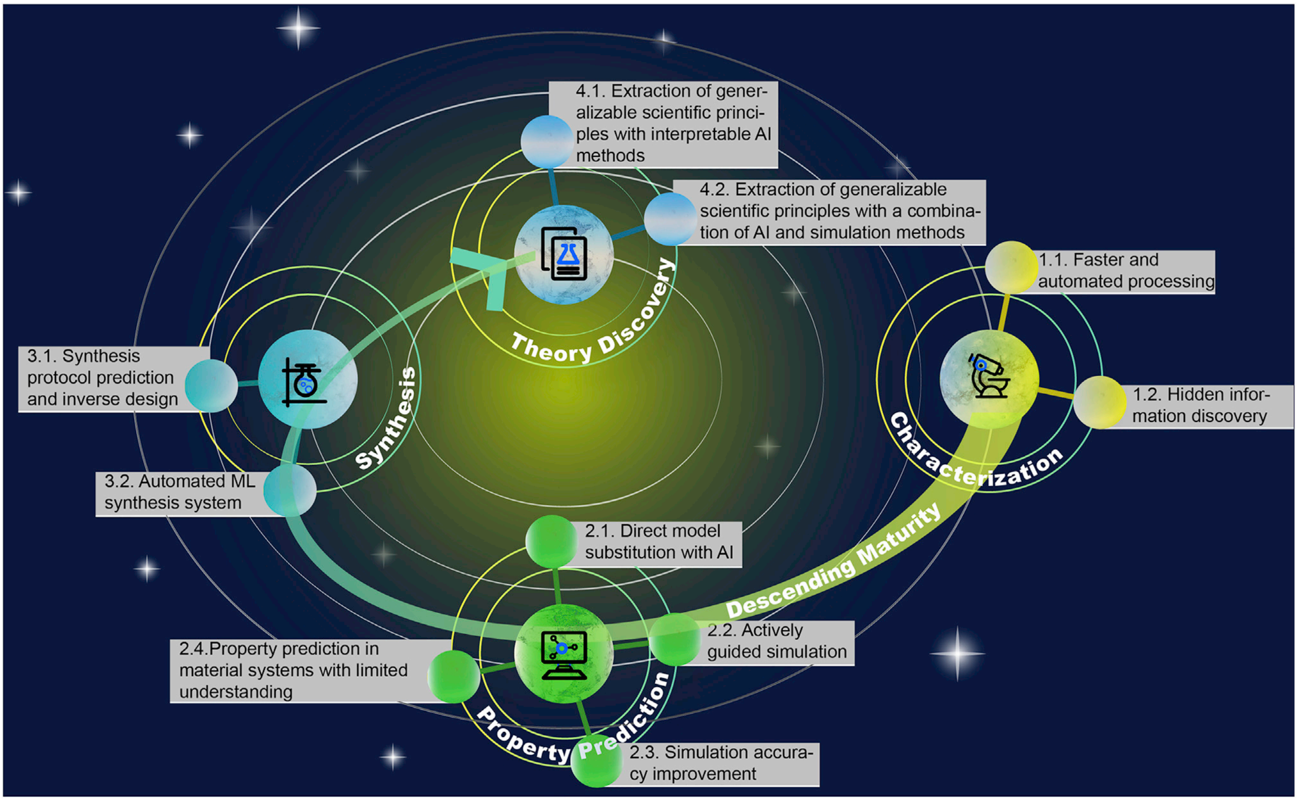
4.1. Extraction of generalizable scientific principles with interpretable AI methods
4.2. Extraction of generalizable scientific principles with a combination of AI and simulation methods
Theory Discovery
1.1. Faster and automated processing
1.2. Hidden information discovery
Characterization
3.1. Synthesis protocol prediction and inverse design
3.2. Automated ML synthesis system
Synthesis
2.1. Direct model substitution with AI
2.2. Actively guided simulation
2.3. Simulation accuracy improvement
2.4.Property prediction in material systems with limited understanding
Property Prediction
Descending Maturity

Figure 3. The Overall Structure of AI Applications through the Whole Life Cycle of Material Discovery

amount of information output from the characterization stage could be the essential first step in applying AI in property prediction and synthesis studies. As a result, characterization will be first introduced. For property prediction, the large amount of simulation databases makes ML applications possible, while there are still challenges in processing the material-specific data efficiently (e.g., chemical molecular structures, crystal structures). For synthesis studies usually laboratory data, which are quite expensive and time-consuming to collect, are required. Moreover, incorporating the complex extrinsic information makes the problem even more challenging. The theory discovery finally tackles the most difficult part in material discovery, which is also the least mature application.

Here, we introduce sections in the sequence of characterization, property prediction, synthesis, and theory paradigm discovery, following a descending maturity trend. The detailed review structure that guides us into the AI-aided material discovery universe is shown in Figure 3.

## AI-Aided Material Characterization

Nowadays, material characterization methods such as super-resolution optical microscopy, free-electron laser, nuclear magnetic resonance (NMR) spectroscopy, and scanning transmission electron microscopy (STEM) can generate a large amount of data. Traditionally, the processing of such information is done manually or by some computationally intensive processing methods. The lack of automation results in the characterization process being time-consuming and accompanied by a potential loss in accuracy due to human errors. For example, finding well-performing parameterizations of complex imaging systems relies on extensive manual work, and

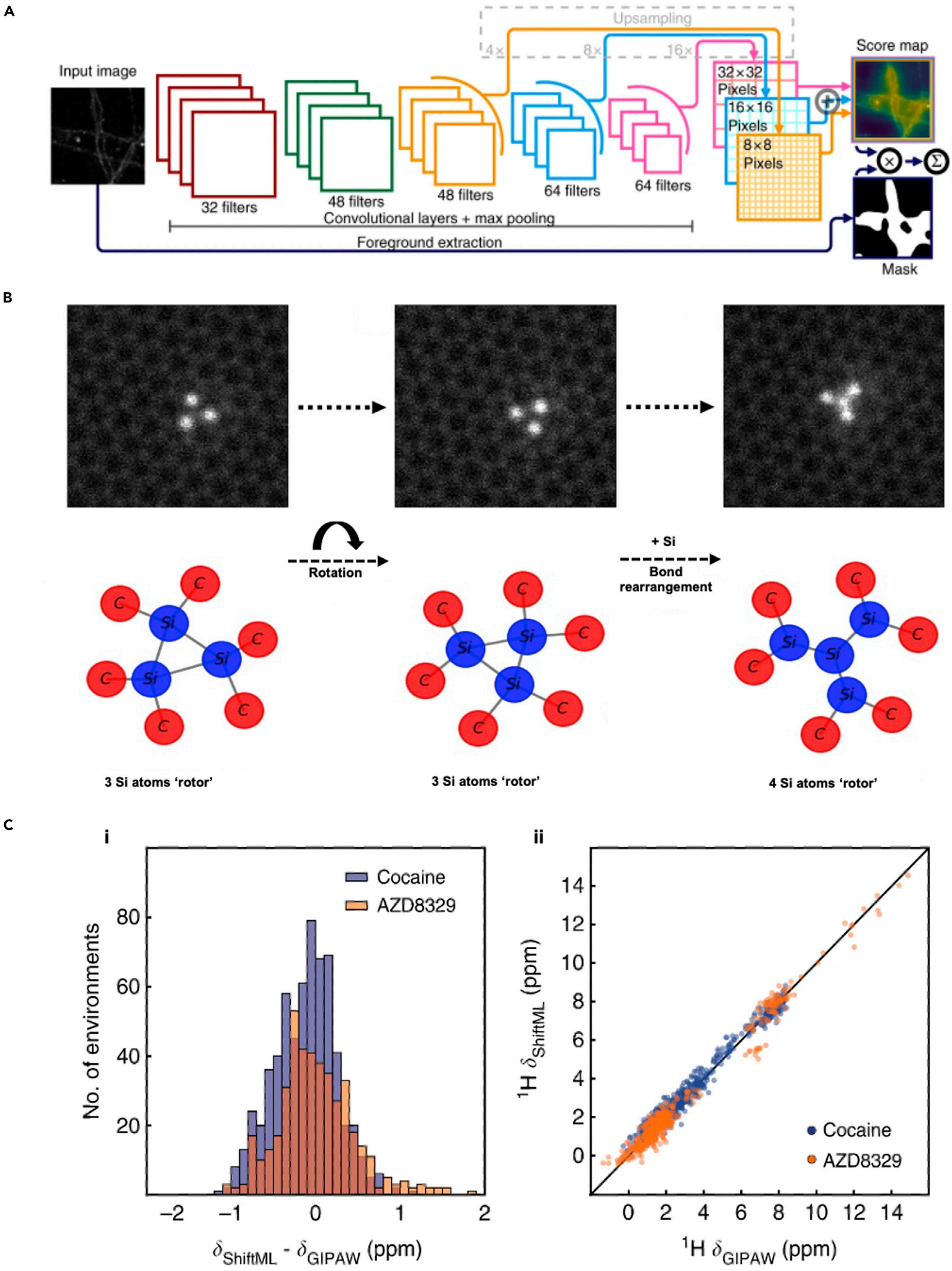
A
Input image
32 filters
48 filters
48 filters
64 filters
64 filters
Convolutional layers + max pooling
Foreground extraction
Upsampling
4×
8×
16×
32×32 Pixels
16×16 Pixels
8×8 Pixels
Score map
Global quality score [0, 1]
Mask
B
Rotation
+ Si
Bond rearrangement
3 Si atoms 'rotor'
3 Si atoms 'rotor'
4 Si atoms 'rotor'
Si
C
C
i
ii
Cocaine
AZD8329
No. of environments
$\delta_{ShiftML} - \delta_{GIPAW}$ (ppm)
−2
−1
0
1
2
20
40
60
80
$^1H\ \delta_{ShiftML}$ (ppm)
$^1H\ \delta_{GIPAW}$ (ppm)
0
2
4
6
8
10
12
14

**Figure 4. Improving Characterization Quality, Speed, and Insights Using ML**

(A) An example of a fully automated characterization process. Proposed fully convolutional network architecture for quality rating. Each convolutional layer is followed by spatial batch normalization and an exponential linear unit activation. Reproduced with permission from Durand et al.[110]

(B) An example of enabling a new kind of characterization. Tracking complex defect transformations on the surface of silicon-doped graphene by using a fully convolutional network framework to atomically resolve scanning transmission electron microscopy. Reproduced with permission from Ziatdinov et al.[85] Copyright 2018, American Chemical Society.

(C) An example of using ML to substitute computational expensive simulation methods to interpret experimental results faster. Comparison of predictions from ML-based ShiftML method and DFT-based GIPAW DFT method for polymorphs of cocaine and AZD8329. Reproduced with permission from Paruzzo et al.[117]

the processing of noisy multidimensional characterization datasets is challenging.[110],[111] In addition, the analysis of pixel-scale imaging data of output information from various characterization methods is difficult to "eyeball." However, with ML advancements the data-driven methods can provide a faster way to process this information with fewer systematic errors by automating the whole process. Moreover, due to the high-dimensional and complex nature of material characterization data, it is difficult for humans to identify the hidden information among them (such as complicated time-dependent crystallization information, collective dynamic data of piezoelectric relaxation studies, and condensed phase molecular properties hidden in a simpler characterization method).[112–114] AI methods are proficient at analyzing such information by design and, as a result, the hidden information inside complex output can be analyzed more efficiently.

### *Faster and Automated Processing*

Automation via AI seems to be a promising direction that has been successfully applied in several published studies. The online single- and multiobjective optimization of imaging parameters, such as the illumination and acquisition settings prior to the different scientific imaging tasks (e.g., living cell and biomaterial imaging), is fully automated by applying kernel regression, full CNN, and Thompson sampling to super-resolution optical microscopy.[110] The online training is achieved by including an expert in the optimization routine. The expert can provide feedback (quality rating) on the trade-offs that can be made among the different objectives, and the kernel regression models' capability to identify the optimum imaging parameters that can be improved by learning this feedback. Next, a deep CNN is proposed to remove the expert and fully automate the overall image quality rating process. With this deep CNN model, the whole optimization process is fully automated. The architecture of this automated image quality rating model is shown in Figure 4A. A CNN-based approach has been developed to automatically detect and recondition the quality of the probe of a scanning tunneling microscope (STM).[115] Also, an automated multistage pattern recognition approach has been developed in the detailed characterization of surface molecular architectures of STM. This is realized by constructing training data via simulation, identifying molecular structure states via Markov random field and classifying the accurate rotational class via a CNN.[116] The automation of the characterization process could not only effectively save money and time, but also be a strong tool to discover and track time-related information. A fully convolutional network can be applied to atomically resolved images produced by STEM to realize the fast extraction of atomic species and the type of defects. Complex atomic and defect transformations identified by this approach, including time-related information of switching between different coordination of silicon dopants in graphene, show the potential of this approach to learn reaction time-series information effectively (Figure 4B). Over 2,000 training images are generated for this work, with a data-augmentation procedure to the original synthetic images by using random horizontal/vertical shifts, rotations, zooming-in/-out, and shear transformations.[85]

Apart from the automatic and fast extraction of information, ML technologies such as ANN and DNN have been applied to improve the accuracy of advanced characterization methods. These are applied to analyze complex patterns from the X-ray pulse of a free-electron laser[79] and noisy data from scanning probe microscopy,[111] respectively. The neural networks can tackle the complex output information to obtain full X-ray pulse information on every shot with high fidelity[79] and map hyperspectral data to lower-dimension material-specific parameters with an order of magnitude higher signal-to-noise ratio than the traditional least-squares approach.[111] In addition, a CNN is implemented in classifying crystal structure from the X-ray diffraction image with no handcrafted engineering involved. This approach could be easily implemented and realize accurate and standardized crystal systems classification.[118]

Aside from applying supervised learning methods, unsupervised learning methods could also help in automation and accuracy improvement of the characterization stage. Unsupervised cluster analysis methods show potential in classifying optical microscope images of exfoliated graphene flakes and can identify automatically the position, shape, and thickness of graphene flakes with high accuracy.[119]

AI methods can not only assist or substitute manual work but also accelerate the computational processing schedule of characterization data. This can achieve high-throughput characterization, which is beneficial to time-consuming NMR and 3D simulation work. A Gaussian process regression model is used to substitute expensive computational DFT to calculate NMR chemical shifts for the structure determination of molecular solids. The ML approach based on local atonic environments accurately predicts chemical shifts of molecular solids and their polymorphs with different structures. The results are shown in Figure 4C, where Figure 4Ci demonstrates the distribution of the difference in calculated chemical shifts between the ML-based ShiftML method and DFT-based GIPAW DFT method, and Figure 4Cii shows the correlation between the results of the two different methods. The results demonstrate that the proposed ML-based method is within DFT accuracy.[117] Also, a conditional generative adversarial network is applied to reconstructing 3D porous media from a single 2D image or limited morphological information. The proposed method reconstructs 3D porous media layer by layer with lower computational cost compared with the traditional pixel-by-pixel multipoint statistical reconstruction method.[120]

Finally, there is a special valuable application aspect of AI in reaching faster characterization, which is called the "inverse imaging problem." This aspect concerns imaging problems when forward models are present and the difficulty is in mapping the image back to the specific material information (e.g., molecular structures and grain orientations) used in simulating the image. The probe particle model has been used to construct a synthesized 3D AFM dataset from 134,000 isolated molecules. A deep CNN is trained to solve the inverse imaging problem by predicting the molecular structures directly from AFM images.[86] In addition, a deep CNN is trained on a simulated four-dimensional STEM dataset to predict structural descriptors of complex oxides.[121] Also, a CNN-based deep learning framework with a customized loss function is built to extract grain orientations from electron backscatter diffraction (EBSD) patterns accurately and rapidly. This approach can replace low-accuracy, dictionary-based indexing methods.[122]

### *Hidden Information Discovery*

Apart from accelerating data processing of the characterization stage, ML can help find hidden information that was previously undiscoverable due to the

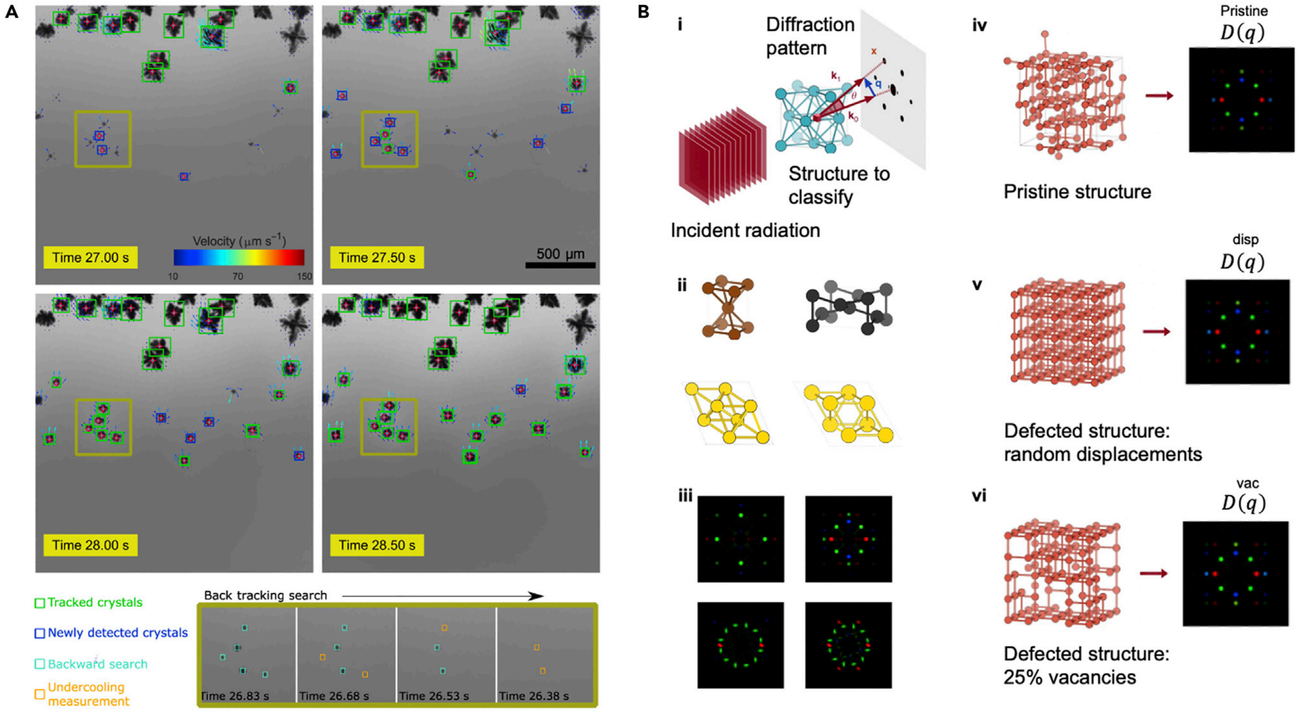

**Figure 5. ML-Assisted Discovery of Hidden Characterization Information**

(A) Hidden information revealed by automated ML tracking technology. Nucleation undercooling measurement method using X-ray radiography and ML. The ML model can automatically track crystals' locations and sizes versus time and quantify multiple processing information. Reproduced with permission from Liotti et al.[112]

(B) (i) Schematic representation of the two-dimensional diffraction fingerprint calculation. (ii and iii) Examples of crystal classes and their 2D diffraction patterns. (iv–vi) Even structures with defects can be classified into the corresponding class, showing that the diffraction fingerprint is robust to defects. Reproduced with permission from Ziletti et al.[84]

complex inner linkage within a dataset. One study used various computer vision algorithms, which were trained to automatically track crystals' locations and sizes in order to learn the effect of cooling rate and solute concentration on nucleation undercooling, crystal formation rate, and crystal growth rate for thousands of separate crystals. The image information is derived from *in situ* synchrotron X-ray radiography (Figure 5A). This approach allows the crystal formation rate, crystal growth rate, and crystal movement to be quantified for each crystal as a function of time, which was not feasible before. This achievement of "seeing" clearer for an ongoing material reaction system could enable a better understanding of a field.[112] In addition, PCA and k-means clustering are applied to robustly identify the onset of a structural phase transition in nanometer-scale volumes by providing collective dynamic data from piezoelectric relaxation studies for ferroelectric relaxors.[113] Moreover, by finding hidden information via a simpler or faster characterization method, one can bypass experimental difficulties in measuring certain properties with simpler methods. Through the application of supervised ML (e.g., RF) and consideration of physiochemical properties such as molecular size, the model can correlate condensed phase physiochemical properties with dynamic gas-phase clustering behavior measured by the differential mobility spectrometry (DMS). This approach combining ML and DMS can be used to quantitatively access a variety of molecular properties relating to an analyte's interaction potential. Such an approach can obtain the properties information faster and with fewer resources than the traditional approaches of directly measuring these properties.[114]

Besides finding hidden information from experimental characterization data alone, ML methods can also be combined with various simulation methods to discover insightful information more effectively. A robust method to classify crystal structures despite the presence of defects is provided by training a CNN based on a 2D image descriptor called diffraction fingerprint. It can dig out the space group information even when the provided 3D materials science structural data are noisy and incomplete. The useful fingerprint is provided by simulating how the crystal structure diffracts incoming incident plane of rays with varying incident angles in 3D space. The learning process of the neural network is uncovered to show the landmarks the model used in achieving good classification performance. The 2D fingerprint calculation process and the robustness of this fingerprint to defects are shown in Figure 5B.[84] In addition, the presence of flexoelectricity in the polar vortices in $PbTiO_3/SrTiO_3$ superlattices is discovered. This is realized by a method based on a combination of unsupervised ML analysis (PCA) of the atomic-scale electron microscopy imaging data and phenomenological phase-field modeling.[95]

Overall, ML methods are successfully applied to material characterization tasks in impressive studies achieving faster speed, obtaining more accurate results, and discovering valuable hidden information. However, the applicability of these studies is not restricted to the characterization stage. The processed characterization data may be used as the input feature to efficiently describe a material system. For example, the 2D diffraction fingerprint from the crystal structure classification task can be used as a promising structural descriptor in properties or synthesis recipe prediction tasks of new crystals, due to its robustness to crystal defects in the crystal's space group classification tasks.[84] In addition, the *in situ* synchrotron X-ray setup could be used in more material systems. The time-series output generated from it can be further analyzed for scientists to learn a dynamic advanced material synthesis system that is far from equilibrium.[112] For a detailed review of ML methods for image processing related to material design, the work of Kalinin et al.[33] is recommended.

## AI-Aided Property Prediction

Property prediction aims to find a suitable function $f : \mathcal{F}_i \rightarrow \mathcal{P}$, where the intrinsic materials' information ($\mathcal{F}_i \in \mathbb{R}^n$) is mapped to the desired functional properties ($\mathcal{P} \in \mathbb{R}^m$). The property prediction work is normally done for material systems where the synthesis protocols are standard and known, i.e., once an optimum final material is found the synthesis of that material is not difficult, or the major interest is only in analyzing the property of the final material.

With the boom of data-driven ML in the materials field, there is a strong need for more data. There have been great advances made for computational material data due to previous efforts from the Material Genome Initiative (lists of the databases can be found in reviews by Butler et al. and Lauri et al.[14,123]). Almost always, it is simpler, cheaper, and faster to generate those data via computational methods as opposed to carrying out actual experiments.

Traditionally, property prediction was carried out by various simulation methods using up considerable computational time. By applying AI to well-understood material systems with known first principles, ML methods are able to partially or completely substitute the computationally expensive simulation processes, to actively guide the whole material simulation workflow, or to improve the accuracy of some simulation methods by utilizing real experimental data or more accurate but even more computationally expensive simulation data. For complex material systems with limited understanding, AI methods can be used to learn the function of property prediction directly from past experimental data and provide a data-driven predictive model. This enables previously impossible property prediction for these systems and hence heralds significant progress in this endeavor.

### AI-Aided Property Prediction in Well-Understood Material Systems

This section first focuses on material systems that are well understood. This does not imply that the system is simple or completely known with full certainty but instead that there is sufficiently known information of the material systems that allows first-principle calculations or simulations. Such well-understood material systems are discussed with respect to the AI-aided property prediction from three major aspects: (1) direct model substitution with AI; (2) actively guided AI in-the-loop simulation; and (3) simulation accuracy improved via AI.

*Direct Model Substitution with AI.* First-principle simulation methods have been used to calculate various properties of different material systems such as intrinsic breakdown strength of perovskites,[124] CO adsorption of gold nanoclusters,[125] and gas adsorption of MOFs.[23] However, the computational cost is very high for these simulation methods (e.g., 60 s to $10^3$ s).[22,23] ML methods can be used to partially or completely substitute the computationally expensive simulation process in order to achieve high-throughput material screening. This allows the fast searching of the vast material design space to quickly find areas that contain the desired functionalities.

A ML model can be trained by existing training datasets to form a data-driven predictive model, which provides the function $f$ that can be used to predict functional properties of new materials with different intrinsic information. However, for such ML models to be effective, good descriptors must be developed and used as feature input. Therefore, we first introduce some important work on chemical feature engineering and then discuss the direct model substitution.

Many past studies have focused on demonstrating the substitution process by applying a novel effectively tailored descriptor set to capture essential information of the specific material system. For example, when predicting the properties of molecules, the molecular structure itself is a discrete form of information. Since gradient-based optimization methods are much faster than discrete optimization methods, efforts have been made to develop a variational autoencoder that transforms discrete molecular representations (SMILES) into continuous latent vector representations. The continuous representations can then be used as input into the predictive ML model. After prediction tasks, the continuous latent vector representations with desired properties can be transformed back to SMILES by a decoder.[102] In addition, the bag-of-bonds method was developed to quickly estimate atomization energies and different electronic properties for a representative set of organic molecules from the GDB-7 simulated database.[126] Moreover, an MPNN framework is constructed to take a molecular graph as input and map it to a continuous vector. This approach is able to achieve state-of-the-art results on an important molecular property prediction benchmark (QM9 Dataset), with a 5-orders-of-magnitude reduction in computational time compared with the conventional DFT approach.[22] A continuous representation of the material structure information is desired, as it contains much more information than discrete representation, enabling ML to learn structure-property relationships more efficiently. Besides the structure features, other descriptors that can describe material intrinsic information such as composition descriptors and physiochemical descriptors are frequently used in ML enhanced simulation work. An extensive review of material descriptors categorized the descriptors into three different levels, from the gross-level property-based descriptors to molecular fragment-level descriptors and further to sub-angstrom-level descriptors.[127]

By utilizing efficient descriptors consisting of both structural and compositional information, DFT calculations of CO adsorption energies of a Au-based nanocluster can be successfully replaced by an RF-based ML framework. The training dataset of this study contains over 2,000 DFT simulated data points. This model is an excellent filtering tool to select first-round candidates for further simulation and analysis. Moreover, the novel descriptors (mainly based on the structure of nanoclusters) developed in this work for nanocluster material can be applied efficiently to other ML nanocluster work (Figure 6A).[125] Also, separate SVMs are trained on data of various properties calculated by first-principle methods to directly discover desired material from the periodic table, and the predicted material properties show good agreement with experimental data.[128] In addition, an ANN is implemented to map directly the conformationally dependent electronic structure of a molecule to coarse-grained (CG) pseudo-atom configurations. This accelerates the simulation by eliminating both back-mapping of CG configuration to atomistic representation and repeated quantum-chemical calculations.[129] Moreover, ANN, SVR, RF, bidirectional neural network, and other regression methods were applied to replace different time-consuming simulation steps in the prediction of intrinsic dielectric breakdown strength of ABX3 perovskites,[124] gas adsorption in MOFs,[23] stability of ternary intermetallic compounds,[130] and chiroptical responses of chiral metamaterials.[131] The models are built with a large variety of physiochemical, structure, and composition descriptors. These methods show promising accuracy and a much faster prediction, which can be utilized in screening a significant number of potential material candidates. Also, an SVM-based ML approach is constructed to classify and predict the crystal structures of ternary equiatomic compositions ABC. The features used are based only on the constituent elements.[132] Finally, several ML models have been trained on high-quality heat capacity data of solid inorganics from NIST-JANAF tables.[40] The trained ML models can be used to rapidly predict heat capacity, which can replace traditional methods such as DFT-based calculation, Cation/anion contribution methods, and Neumann-Kopp estimations, which are limited by either speed or accuracy. This method can be used as a high-throughput screening tool for identifying useful new materials.

Apart from substituting the whole simulation model with ML methods, an ML model can replace some prediction steps of a simulation process and then be combined with the simulation method to achieve both accuracy of *ab initio* methods and high speed for screening. For example, feature selection and gradient boost methods are used in finding the stable lead-free hybrid organic-inorganic perovskites (HOIPs) with desired band-gap properties using a training set composed of 212 simulated data.[133] The ML model can reach high accuracy, as shown in Figure 6B. The ML model is used as a pre-screening tool to select HOIPs with proper band gaps and the DFT runs on the selected candidates to further evaluate their electronic properties and stability. Also, the DFT simulation of Debye temperature of inorganic phosphor host can be replaced by an SVR; this SVR model is used to rapidly predict the Debye temperature of 2,071 potential phosphor hosts.[41] The Debye temperature is then plotted against the DFT calculated band gap to find the compounds with the highest Debye temperature and the largest band gap for synthesis.

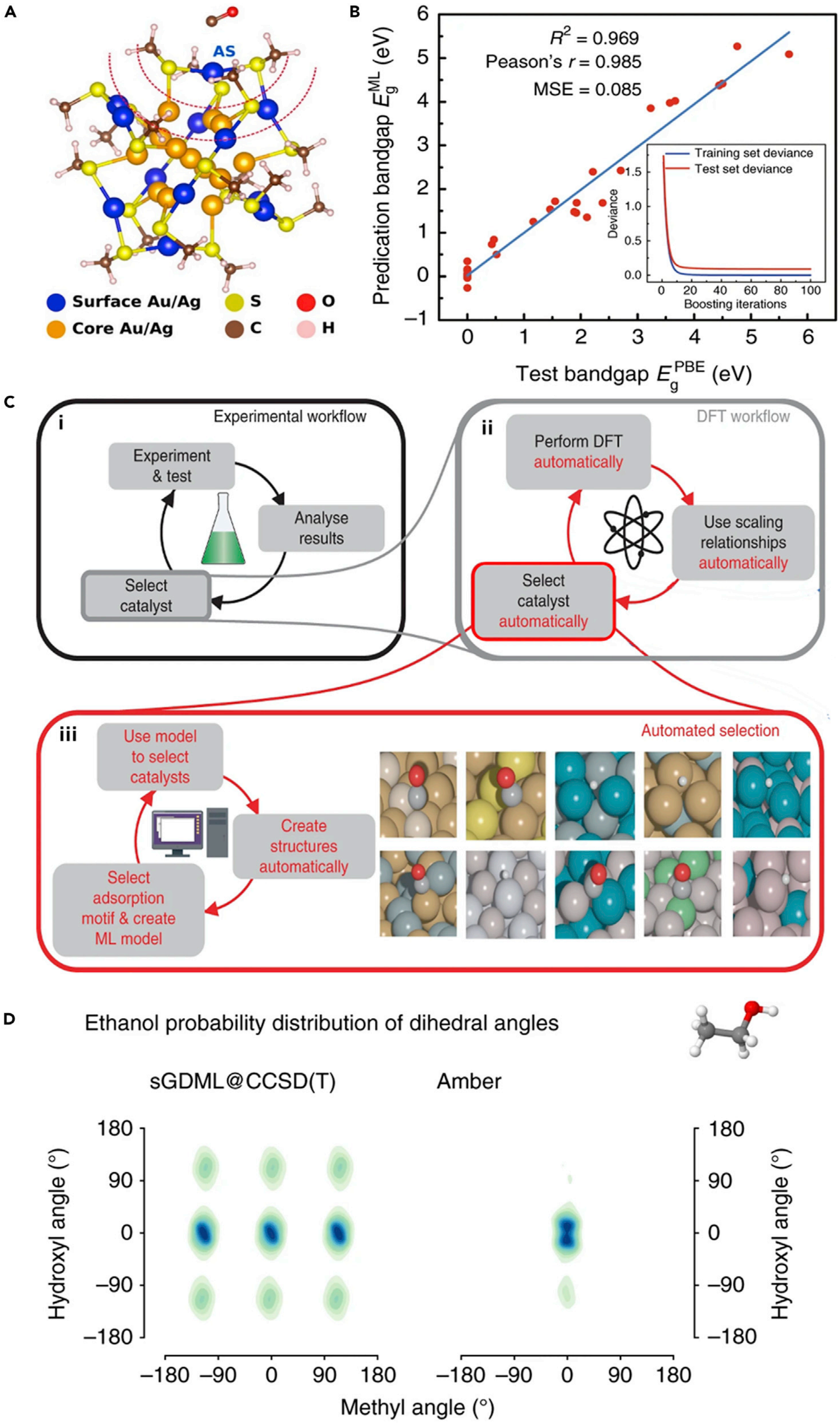
A
AS
Surface Au/Ag
Core Au/Ag
S
C
O
H
B
$R^2$ = 0.969
Peason's $r$ = 0.985
MSE = 0.085
Predication bandgap $E_g^{ML}$ (eV)
Test bandgap $E_g^{PBE}$ (eV)
-1
0
1
2
3
4
5
6
Training set deviance
Test set deviance
Deviance
0.0
0.5
1.0
1.5
Boosting iterations
0
20
40
60
80
100
C
i
Experimental workflow
Experiment & test
Analyse results
Select catalyst
ii
DFT workflow
Perform DFT automatically
Use scaling relationships automatically
Select catalyst automatically
iii
Automated selection
Use model to select catalysts
Create structures automatically
Select adsorption motif & create ML model
D
Ethanol probability distribution of dihedral angles
sGDML@CCSD(T)
Amber
Hydroxyl angle (°)
180
90
0
–90
–180
Methyl angle (°)
–180
–90
0
90
180

**Figure 6. AI-Aided Property Prediction in Well-Understood Material Systems**

(A) Creating new descriptors for nanocluster structures, and using ML as a surrogate model to predict CO adsorption on nanoclusters. Reproduced with permission from Panapitiya et al.[125] Copyright 2018, American Chemical Society.

(B) A partial simulation substitution study with an ML-based band-gap prediction model as the pre-screening tool. This prediction model can reach high accuracy on the test set. Reproduced with permission from Lu et al.[133]

(C) An actively guided DFT workflow to screen catalysts quickly and with *ab initio* accuracy. Panel (iii) shows the proposed ML-based workflow to select candidates systematically and automatically. Reproduced with permission from Tran and Ulissi.[53] Copyright 2018, Springer Nature.

(D) An example of AI-aided simulation accuracy improvement study. This shows that the proposed ML combined MD approach (sGDML@CCSD(T)) can reproduce the dihedral angle probability distributions of ethanol with higher accuracy than the traditional force-field method (Amber force field). Reproduced with permission from Chmiela et al.[134]

*Actively Guided AI-in-the-Loop Simulation.* Different active learning frameworks can be combined with a simulation model to effectively navigate the search space iteratively and identify promising candidates for improving the performance of the surrogate ML model or guiding the simulation work in order to search materials with optimum desired properties. This replaces traditional methods such as grid search or change-one-variable-at-a-time methodology. As discussed in the ML tutorial section, active learning aims to search for the next best unlabeled data point to label. In other words, it aims to find the next best material to simulate based on the greatest expected improvement of the model performance (overall performance or performance on the optimum prediction) or the greatest improvement of the sample input representativeness when the model is updated. This allows a more efficient search of the materials' design space as compared with traditional methods. A Gaussian process regression (GPR) is used as a surrogate model to substitute computationally expensive Poisson-Schrödinger simulations of GaN-based light-emitting diode efficiency, and an efficient global optimization (EGO) strategy is implemented to select the next sample point that maximizes the expected improvement in efficiency while simultaneously accounting for the GP model's uncertainty.[135] A similar work used Bayesian optimization with a surrogate GPR to find layered materials with desired properties.[54] In addition, an ML model is trained on-the-fly as a computationally inexpensive energy predictive surrogate model to accelerate the traditional genetic algorithm (GA) optimization workflow of energy calculations of nanoalloy catalysts. When compared with the traditional "brute force" GA, this approach yields a 50-fold reduction in the number of required energy calculations in this case.[136] In addition, a Bayesian optimization-based crystal structure prediction method is shown to outperform the random search-based method on known systems such as NaCl and $Y_2Co_{17}$. The proposed method can reduce by 30%–40% of the number of searching trials required.[137] Query by committee strategy is also used to develop accurate and transferrable potentials for predicting molecular energetics with a fraction of data required when using naive random sampling techniques. This strategy can improve the overall fitness of the potential by sampling the chemical space when the potential fails to accurately predict the potential energy.[109] Moreover, an active learning framework is used to automatically guide the full-accuracy DFT screening instead of simply substituting the DFT calculation. This is used for the discovery of electrocatalysts for $CO_2$ reduction and $H_2$ evolution. Conventional catalyst simulation-based design work involves the manual analysis of DFT data to gain specific intuition to set the candidates for the next round of calculation. In this work, by creating a surrogate ML predictive model in each round to predict the adsorption sites with near optimum adsorption energies (and add Gaussian noise) for next-round DFT calculation, the whole simulation workflow can be more efficient and automated, as shown in Figure 6C.[53] Finally, a benchmark study of different multiobjective optimization-based adaptive design methods has been conducted on both material simulation data and laboratory data over different material systems. The proposed Maximin method can well balance exploration and exploitation in the data-sparse scenario, which commonly exists in materials' datasets.[138] Overall, the active learning design framework is suitable for simulation-based material discovery, which, however, can also be effectively used in complex material systems with laboratory data. More details will be discussed later. The general framework of active learning is similar for different applications, while the details (e.g., dataset size and feature choice) will vary with different applications case by case.

*Simulation Accuracy Improved via AI.* Apart from the acceleration of computational material design, studies are aiming to improve the accuracy of simulation methods. This target can be realized by several different approaches.

First, an ML model can be trained to correct the difference between more accurate data sources (e.g., lab results or more computationally expensive but more accurate simulation results) and an existing model trained on less accurate but larger amounts of data (usually produced by approximate but faster simulation methods). An ANN is used to correct systematic errors between quantum mechanical calculated heat of formation and experimentally measured data for small molecules. This is realized by using the ANN as an additional layer after the simulation to calibrate it with ground-true laboratory data.[139] In addition, ML corrections have been added to computationally inexpensive approximate legacy quantum methods to more accurately predict various chemical properties.[140]

Second, transfer learning can also help achieve higher simulation and prediction accuracy. In transfer learning for material property prediction, one begins with a model trained on proxy properties with sufficient data and then retrains the pre-trained model on the related (or the same but more accurate) target properties with limited data supply. Normally, when retraining the model, some of the model's parameters are fixed. CCSD(T)/CBS (coupled cluster considering single, double, and perturbative triple excitation calculations, combined with an extrapolation to the complete basis set limit) level accuracy can be achieved on various properties by first training a neural network model on DFT data and then retraining it on a much smaller dataset of gold-standard quantum mechanics calculations (CCSD(T)/CBS) that optimally spans chemical space.[141] In addition, a pre-trained library of material property for transfer learning called XenonPy.MDL has been created. The effectiveness of using transfer learning with this library for different material systems (small molecules, polymers, and inorganic compounds) has been demonstrated.[142] The transfer learning strategy aims not only to improve simulation accuracy but also to bear the potential of improving the prediction accuracy of any material discovery problem where large datasets of properties related to target properties are available.

Finally, ML can be used to calculate more accurate prior information (e.g., molecular force field) that can be further used in the simulation. A gradient-domain ML model trained on a high-level *ab initio* calculated dataset (coupled cluster single-double and perturbative triple [CCSD(T)] level) is combined with MD simulation. The ML model is used to produce a flexible molecular force field with high accuracy within short computational time. This approach allows converged MD simulations with fully quantized electrons and nuclei. The MD simulation can work for flexible molecules with up to a few dozen atoms. This level of exact MD simulation was previously unreachable due to infeasible computational time to produce the global force field at CCSD(T) level. The developed approach could achieve spectroscopic accuracy and rigorous dynamical insights in molecular simulations that improve on previous MD approaches implementing traditional force-field methods, as shown in Figure 6D.[134]

### AI-Aided Property Prediction Study in Material Systems with Limited Understanding

For complex material systems such as the ones with both organic and inorganic parts, complex solid-solutions with fractional site occupancies, large supercells, transition metal oxides with strong electron correlation, and prediction of functional material property not in an ideal situation,[26,143] first-principles simulation might be infeasible due to the limited understanding of such systems. This then leads to the necessity of ground-true laboratory data and a statistical predictive model to analyze them effectively. Such a well-constructed statistical predictive model based on laboratory results can be used to screen material candidates over vast design space.

Active learning approaches have been implemented in the barium titanate material system in order to find high dielectric permittivity material[144] and piezoelectrics with large electrostrains that are not limited to the zero-K condition.[29] Only composition-based descriptors are used in both studies, since the synthesis protocol is kept constant within each study. The first study collects three data points initially and uses GP to construct a surrogate function.[144] With the EGO strategy, the next data point is experimentally conducted. The new data are added and the workflow repeated. The experimental search process is indeed improved as only 6 compositions over 16 possibilities are analyzed to find the optimal composition. For the second study, 61 experimental data points are collected as the database.[29] A number of ML models (e.g., SVR with radial-based kernel, polynomial fits, and gradient tree boosting) and nonparametric bootstrap sampling are used to generate an ensemble of 1,000 models to predict the electrostrain of piezoelectric material. An expected improvement that balances exploration and exploitation is calculated from the ensembled model to suggest the next points to carry out experiments. A design space of 605,000 compositions is screened and leads to the synthesis of a piezoelectric material with the largest electrostrain of 0.23% in the $BaTiO_3$ family.[29] A poorly understood new field of spin-driven thermoelectric effect devices is explored by applying DT methodology with mainly physiochemical and composition descriptors on 112 experimental data. Under the circumstance of lacking understanding of the fundamental physics and material properties responsible for the effect, the ML can exhibit its full potential.[145] In addition, the design of higher-selectivity chiral catalysts is achieved by implementing SVMs and deep feedforward neural networks with 2,150 experimental data.[96] To effectively analyze the selectivity property, a robust structure descriptor called average steric occupancy is invented. This descriptor has the potential to be applied in material design where the property can be affected by 3D steric occupancy. This AI-aided selectivity chiral catalyst design work has the potential to change the way chemists select and optimize chiral catalysts from an empirically guided to a mathematically guided approach. Moreover, a two-step active learning framework is applied in searching for high Curie temperature ferroelectric perovskites in a vast design space that traditional experimental approaches find difficult to achieve. An SVM classifier is used to screen the compositions that can synthesize materials in the perovskite phase, and an SVR model is used to predict the Curie temperature with composition descriptors. This process is completed by incorporating an EGO strategy to choose promising high-temperature candidates for experiments to improve the performance of classification and regression models. This kind of multistep active learning approach could improve the robustness of the ML model and make the whole search process well guided.[26] Furthermore, the glass-transition temperature of multicomponent oxide glasses is predicted by an ANN with a dataset of 55,150 examples and only composition descriptors.[80] Finally, the fluorescence intensity and color are studied for genomic silver nanoclusters by

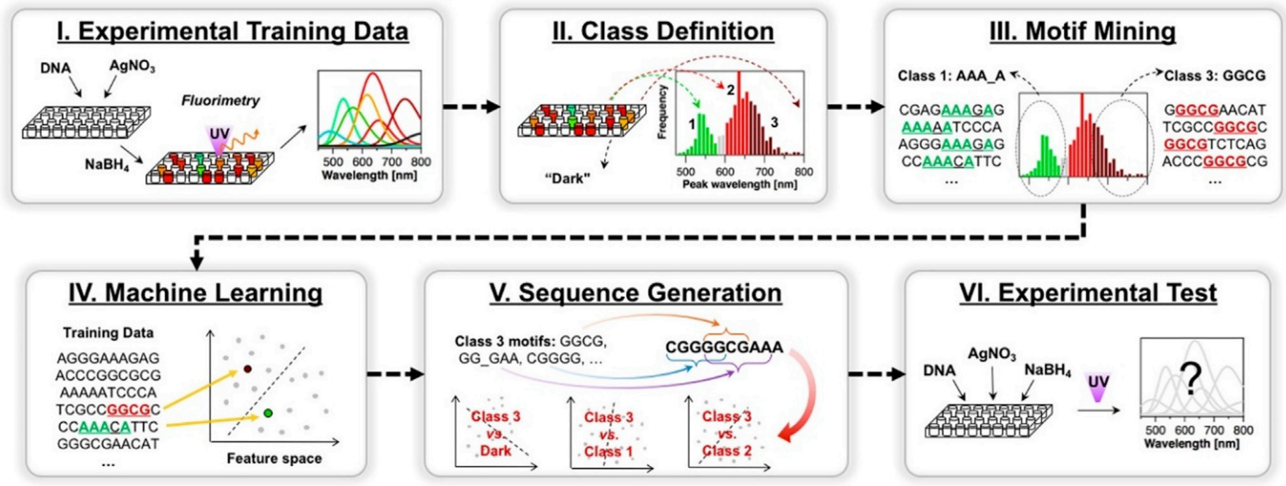

**Figure 7. AI-Aided Property Prediction in Material Systems with Limited Understanding**
Schematic of the sequence design method for DNA templates that stabilize $Ag_N$-DNA within specific color bands. Reproduced with permission from Copp et al.[68] Copyright 2018, American Chemical Society.

SVMs. Since there are four color classes, one-versus-one classification is employed. This shows that instead of the whole DNA sequence, the use of certain DNA base subsequences, or "motifs," can be more effective as the features. The workflow of this fluorescence color study is well illustrated and can be treated as a typical workflow of this section. It includes data generation, task identification, feature construction, model training, inverse design, and experimental validation, as shown in Figure 7.[68]

In summary, for property prediction tasks in both well-understood material systems and the systems with limited understanding, ML could be an efficient tool to overcome several challenges. ML could enhance the simulation process for well-understood systems from three aspects. First, it could speed up the whole simulation process by replacing a part of it or the whole. In addition, by incorporating ML in the loop, the searching and optimization of the whole simulation process could be automated and accelerated. Finally, the simulation accuracy could also be improved by accompanying ML methods. For systems with limited understanding, ML could be an efficient fast screening strategy by learning extensively from both success and failure experimental data.

## AI-Aided Material Synthesis

For material synthesis studies, the goal is to find the map $f : \mathcal{F}_i \times \mathcal{F}_e \rightarrow \mathcal{P}$, where the intrinsic materials' information ($\mathcal{F}_i \in \mathbb{R}^n$) and the extrinsic materials' information ($\mathcal{F}_e \in \mathbb{R}^w$) are mapped to the desired functional properties ($\mathcal{P} \in \mathbb{R}^m$). Here, the interest is in finding a suitable synthesis pathway, given a set of starting materials with their intrinsic material information, to form the final product with desired functional properties. This work is normally done for material systems with nonstandard synthesis protocols using laboratory data. Traditionally, material synthesis is done by material experts with an Edison approach, which is time-consuming and labor-intensive. This approach is usually guided by qualitative information and does not fully utilize previously collected quantitative information. Through AI and ML approaches, extrinsic information, intrinsic information, and past experimental

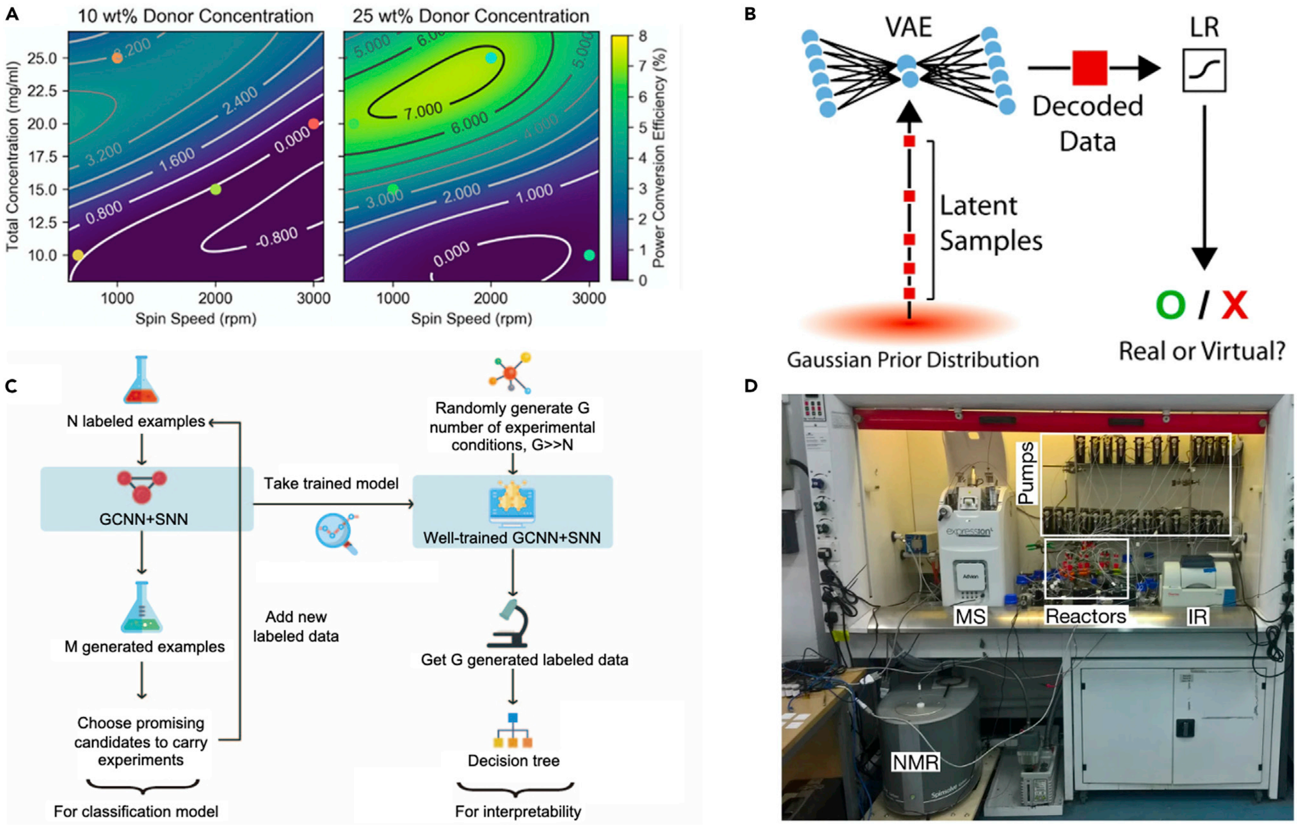

**Figure 8. AI-Aided Material Synthesis**

(A) The power conversion efficiency of organic photovoltaic devices against several synthesis parameters (spin speed, total concentration, donor concentration). Reproduced with permission from Cao et al.[147] Copyright 2018, American Chemical Society.

(B) The general framework for generation of realistic synthesis recipes. Reproduced with permission from Kim et al.[103]

(C) A deep learning framework with a closed learning loop based on SNN. This framework is divided into classification prediction part and interpretation part.[37]

(D) Photograph of the setup of the chemical robot, showing the pumps, reactors, and real-time analytics, including mass spectrometer (MS), NMR device, and infrared spectrometer (IR). Reproduced with permission from Granda et al.[69] Copyright 2018, Springer Nature.

information can be effectively utilized to predict the synthesized material's properties of interest.

Although this field is still at its early stage, some signature work has shown the great potential of applying ML in material synthesis studies. Various ML methods are combined with high-throughput experiments to rapidly discover three new glass-forming systems for metallic glasses. Since diffusion kinetics play a critical role in stabilizing glasses, the different synthesis methods noticeably affect glass-forming ability. To cope with this, various synthesis-method-dependent ML approaches for this system are constructed.[146] In addition, a radial base SVM is used to optimize the power conversion efficiencies (PCE) of bulk heterojunction of organic photovoltaic devices. This improves the PCE through optimizing multiple synthesis variables such as donor concentration, spin speed, and additive concentration. The visualization of PCE versus these synthesis variables is shown in Figure 8A.[147] Potential $SrTiO_3$ synthesis parameters can be suggested by a combination of VAE and logistic regression (LR) binary classifier. The LR is trained to differentiate between text-mined ("real") and VAE-generated ("virtual") synthesis descriptors. The VAE model can learn compressed synthesis representations from sparse descriptors effectively and output novel synthesis recipes through the decoding process automatically. If the output synthesis recipe can pass the LR, it is treated as the potential new synthesis recipe. The framework is shown in Figure 8B.[103] A graph convolutional neural network plus Siamese neural network (SNN) classification model is trained on a small dataset of 54 success and failure experimental data to predict the synthesis recipe of atomically precise gold nanoclusters. This framework provides a possible solution for low data scenarios faced by AI in the materials field. In combination with a DT trained on synthetic data from the well-trained classification model, useful chemical intuition is also derived (Figure 8C).[37] Moreover, the dissolution kinetics, defined as the $SiO_2$ leaching rate, is studied for silicate glasses. Four different ML models are used, namely linear regression, SVR, RF, and ANN. A noticeable point is that the design parameters for the material itself only include composition descriptors, but the target property is affected largely by the external environment such as pH value. This shows that even a material synthesis process can be highly controllable; the extrinsic descriptors that describe how the final material interacts with the external environment can be essential if environment-dependent materials are desired.[148]

Finally, an automated synthesis platform that can predict the reactivity of possible reagent combinations is constructed. This platform consists of an automated reaction system controlled by the AI mind. The AI prediction model for material synthesis is improved after each new round of experiments and can reach the desired accuracy rapidly. This platform is showcased by exploring the Suzuki-Miyaura reaction, which sets a good example of how a future lab could appear. The setup of this chemical robot is shown in Figure 8D.[69]

Overall, ML methods are suitable for application in advanced material synthesis. For material systems where the experimental part still consists of large uncertainty, ML combined with both intrinsic and extrinsic descriptors could assist in finding the complex prediction function. An efficient combination of suitable ML methods and experimental datasets can convert an Edison approach to a quantitatively guided approach that could largely accelerate the material discovery workflow.

### AI-Aided Theory Paradigm Discovery

Rather than being limited to an effective tool to guide the design process of advanced materials, it is always of interest and priority to gain generalizable scientific principles of advanced materials' systems. As already discussed, for material systems with insufficient or limited understanding, normally a prediction model based on laboratory data can be created. By analyzing these prediction models with feature importance and correlations, or model-agnostic methods, more understanding of the system can be derived. Also, a detailed first-principles simulation can be used to analyze the novel materials suggested by the prediction model to find the mechanism behind the desired property gained.

Several successful studies have applied ML to guide the discovery of the theory paradigm for a specific material system. One interesting study is the discovery of a new hidden structural property of glassy liquids called softness.[67] This property was constructed by training an SVM with several structure descriptors on a training dataset containing 6,000 particles at each density. This new softness property of each particle in the system can be treated as a property that is constructed by forming certain complex relationships among the structure descriptors used. By utilizing this new hidden property, one can better predict the relaxation dynamics of glassy liquids, which suggests that the challenge of understanding glass-transition dynamics could be converted into the challenge of understanding the evolution of softness. This sheds light on understanding this complex material system by introducing a new property that is easier to investigate. A CNN model is built to predict the photoelectrochemical power generation of a solar fuel photoanode based on composition and Raman signal descriptors.[149] By analyzing gradients in the trained model, key data relationships that are not readily identified by human inspection or traditional statistical analysis are discovered. This work shows that a gradient analysis approach of material property prediction can be useful to gain understanding of a mechanism. In addition, a layer-by-layer physics domain knowledge informed Bayesian network is developed for optimizing photovoltaic devices' efficiency. Instead of using traditional optimization (process variable-device performance), this two-layer optimization approach (process variable-material properties-device performance) shows better interpretability and reveals new aspects about the devices' design processes. The two-step highly interpretable Bayesian inference framework is shown in Figure 9A.[30] New physical interpretations are also generated for complex, high-dimensional grain-boundary systems.[150] Several ML methods are combined with a high-quality structure descriptor called smooth overlap of atomic positions (SOAP) to capture local atomic environment (LAE) in predicting several crystalline properties. This work uncovers the relationship between the LAEs with particular properties. Once these local structures are understood, they can be optimized for desired behaviors. Moreover, insights of structure-phenomena relations can be extracted by training DTs on data obtained from EBSD scans. The importance of different structure attributes in causing deformation twinning in Mg AZ31 is ranked by the DTs. This result can help the community understand the physical processes associated with tensile twinning in Mg AZ31.[151]

With the assistance of various ML methods, new unwritten guidelines that can be used by synthetic chemists to find the correct synthesis conditions are discovered. An SVM-derived DT model is produced by training on a dataset of 3,955 unique, complete reactions (both success and failure experiments).[27] This model generates three new hypotheses for the crystallization of templated vanadium selenites as shown in Figure 9B. In addition, a GA is combined with a high-throughput robot to search for the optimal synthesis conditions for a prototypical MOF.[152] By analyzing the generated data with an RF model, the chemical intuition that researchers develop in their search for the optimal conditions is captured and quantified. This intuition can be transferred while the detailed chemistry is different for the synthesis of other MOFs.

Finally, with the combination of simulation methods, the work introduced in the previous section to find piezoelectrics with large electrostrains can be further explored. By implementing Landau theory and DFT, it reveals that the large electrostrain is due to the presence of Sn, which allows for the ease of switching of tetragonal domains under an electric field.[29] Also, graph dynamical networks have been implemented to extract statistically relevant dynamics from MD simulations.[51] This approach has revealed previously undiscoverable important dynamical information for various multicomponent amorphous material systems.

To dig the scientific understanding out of high-dimensional complex material data, three key points must be understood. First, a good choice of informatics descriptors for a specific material system and a specific task is necessary. Second, suitable ML models must be chosen for the dataset, and good interpretation methods for these ML models are available, taking advantage of the existing abundant interpretation methods for different ML models.[153] Lastly, a combination of simulation methods and ML models can also reveal important theory paradigms. ML methods can assist

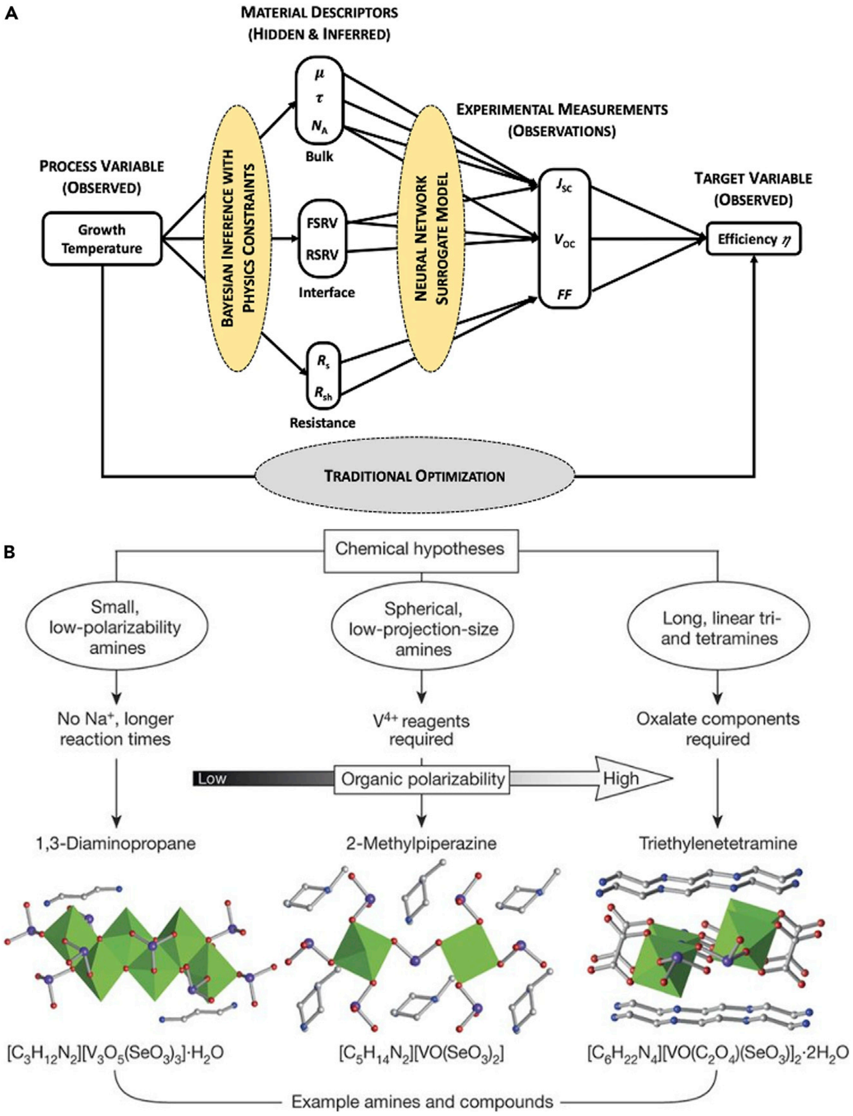
A
MATERIAL DESCRIPTORS (HIDDEN & INFERRED)
μ
τ
N_A
Bulk
FSRV
RSRV
Interface
R_s
R_sh
Resistance
PROCESS VARIABLE (OBSERVED)
Growth Temperature
BAYESIAN INFERENCE WITH PHYSICS CONSTRAINTS
NEURAL NETWORK SURROGATE MODEL
EXPERIMENTAL MEASUREMENTS (OBSERVATIONS)
J_SC
V_OC
FF
TARGET VARIABLE (OBSERVED)
Efficiency η
TRADITIONAL OPTIMIZATION
B
Chemical hypotheses
Small, low-polarizability amines
Spherical, low-projection-size amines
Long, linear tri- and tetramines
No Na+, longer reaction times
V4+ reagents required
Oxalate components required
Low
Organic polarizability
High
1,3-Diaminopropane
2-Methylpiperazine
Triethylenetetramine
[C3H12N2][V3O5(SeO3)3]·H2O
[C5H14N2][VO(SeO3)2]
[C6H22N4][VO(C2O4)(SeO3)]2·2H2O
Example amines and compounds

**Figure 9. AI-Aided Theory Paradigm Discovery**

(A) Schematic of an interpretable ML framework as a physically informed two-step Bayesian network-based process-optimization model. It first links process conditions to material descriptors, then the latter to device performance.[30]

(B) Graphical representation of the three hypotheses generated from the model, and representative structures for each hypothesis. Here three new hypotheses are made about the formation of templated vanadium selenites, categorized by the molecular polarizability of the amine. Reproduced with permission from Raccuglia et al.[27] Copyright 2016, Springer Nature.

in analyzing high-dimensional simulation data to obtain more understanding, and once an optimum material or synthesis recipe is discovered from a large searching space by ML models, detailed simulation can be carried out to analyze the rationale behind the optimum performance.

## FUTURE STEPS

Although an enormous number of articles have been reviewed for different applications of AI in material discovery in this paper, this highly interdisciplinary field is still in the embryonic stage. All of the studies have shown the great potential of applying ML to automate, accelerate, and improve the materials' discovery processes. From what has been reviewed, it can be concluded that the key to achieving good results for AI in materials discovery work includes five major parts: (1) insightful choice of informatics descriptors; (2) suitable models for dealing with specific materials' data structure; (3) efficient labeling by implementing ML in the loop; (4) high-level design to simplify the complex research problem; and (5) the integration of data and methods from different research groups.

Apart from continuing to apply ML in different material systems to achieve different goals, these five major key areas can be treated as the future directions for the community to advance into. By solving the challenges in these directions, our community can discover materials knowledge that was previously not possible to find. For example, solving the challenges can enable us to study complex material synthesis systems that are far from equilibrium with both spatial and temporal complexities (Figure 10).

### Advancement in Descriptors

As introduced in this review, with effective use of suitable descriptors for the material system under study, both good prediction results and in-depth understanding can be achieved. There have been extensive studies in this field; however, a large space still exists for developing and improving useful descriptors such as SOAP[155] and Bag of Bonds[126] that could be used for a group of materials. In addition, suitable features to effectively account for information from more complex materials such as organic and inorganic hybrid systems of ligand-protected nanoclusters and MOFs still need to be discovered. Also, studies that implement the most recent ML methods in creating new material descriptors will be appreciated. New informatics descriptors are created by combining only stoichiometry or even one-hot-encoding form information with attention mechanisms. These new descriptors show promising results in predicting various material properties.[156,157] Moreover, with maturity in high-throughput experimental systems, more complex material synthesis systems can be studied. To acquire the finer information needed for these systems, more sophisticated descriptors such as a descriptor that can capture information for a whole reaction system instead of a single component (e.g., a reaction-system descriptor) should be implemented.[158] Finally, more materials-oriented feature selection methods such as SISSO (sure independence screening and sparsifying operator) are pressing for this emerging field.[159]

### ML Algorithms' Challenges in Complex Material Synthesis Systems

For complex material systems with little first-principle understanding, a statistical predictive model might be trained to predict material properties or synthesis results. Currently, most material systems are limited by data availability, with most ML models being trained on the magnitude of hundreds or a smaller amount of data. This is due to the expensive, labor-intensive, or time-consuming nature of laboratory work. However, the recent development of high-throughput experiments would likely alleviate this problem to a certain extent. We can take a step forward and imagine ways to further utilize the high volume of data of complex material systems generated by high-throughput experiments in the future. Currently, most complex material systems-based ML models use both the intrinsic materials' information ($\mathcal{F}_i \in \mathbb{R}^n$) and the extrinsic materials' information ($\mathcal{F}_e \in \mathbb{R}^w$). Such information sets are both commonly time-independent information. However, most complex reactions of interest are dynamic, far from equilibrium, and spatially heterogeneous. Thus, with better data collection and characterization approaches, we might be

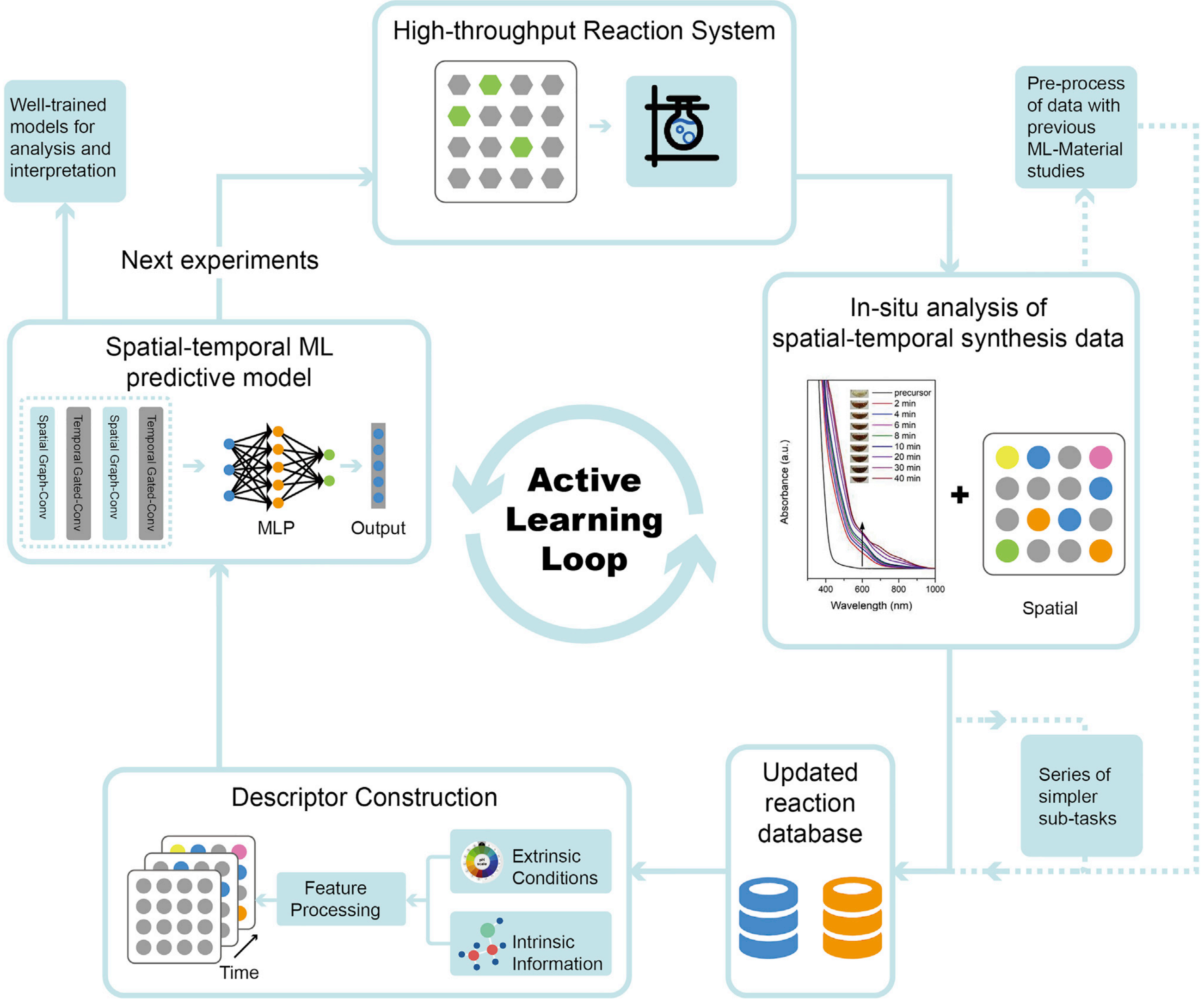

**Figure 10. An Illustration of AI Study in a High-Throughput Material Synthesis System with Spatial-Temporal Complexity**

Achievements in each of the five future directions can push our community one step closer to realizing next-generation material research as shown in the figure. The advancements in each direction for the future directions mentioned can help in this material discovery workflow from different aspects. Within the five future directions, the urging of material-based ML algorithms to analyze spatial-temporal material synthesis data is especially important. The graph indicates UV-visible analysis with time information of gold nanocluster synthesis process. Reproduced with permission from Chen et al.[154] Copyright © 2018, American Chemical Society.

able to incorporate the time dimension into the models. Hence, the intrinsic information ($\mathcal{F}_i \in \mathbb{R}^{n \times t}$) and extrinsic information ($\mathcal{F}_e \in \mathbb{R}^{w \times t}$) can be time-varying and fed into our ML model at each time step to achieve better performance. Certain neural networks such as recurrent or attention-based neural networks could be used to process such time-series features.[89,90] Meanwhile, how to better incorporate chemical information into these ML methods in the materials field is still a topic worth exploring. For example, there can be discrete information of molecular structures as well as continuous information such as temperature, pH, and so forth at each time point. Various methods have been implemented to convert discrete information into continuous data and then combine this with other continuous information to realize the prediction and inverse design tasks of materials or material devices.[62,160,161] However, analyzing these mixed information forms together with a temporal element can be a challenging problem.

Moreover, characterization methods such as attenuated total reflection Fourier transform infrared spectroscopy with focal plane array detectors can capture information for a two-dimensional space.[162] As a result, not only could the bulk properties during the synthesis process of the material be considered, but more detailed spatial information could also be taken into consideration. For example, the spatial information of a material synthesis process as well as materials with heterogeneous structures can be studied. There exists some good work analyzing spatiotemporal chemical information, although there is still much room for improvement.[112,163] Finally, when temporal and spatial chemical information are analyzed together, we should expect to understand the complex system much better. Methods such as graph dynamical networks and spatiotemporal graph convolutional networks can be possible solutions to this challenge,[51,164] while multiple efforts and advances are still needed in methodologies, especially in the fitting materials research domain.

### Machine Learning in the Loop

Since material synthesis experiments are often costly and time-consuming, the data amount is somewhat limited. Although high-throughput material synthesis/process can solve this problem to a certain extent, it is expensive to conduct high-throughput experiments and there are currently great difficulties in realizing this for some complex material systems. Therefore, it remains crucial to develop ML strategies that are effective with sparse data. As a suitable solution, active learning has been well researched theoretically and experimentally. There are many different active learning methodologies such as Bayesian optimization,[31] query-based methods,[106] and closed form variance reduction.[106] However, most active learning research and benchmark tests have only been performed on traditional computer science datasets such as visual or text-based tasks.[165,166] The results of different active learning methods vary greatly depending on the context and task at hand. Thus, it is important to investigate active learning in the context of the materials field. While there have been various studies applying active learning in this area, there is still a strong need to investigate and understand which active learning methods are better for specific classes of material systems or tasks. In particular, the question of how uncertainty should be quantified and different ways to maintain balance between exploitation and exploration in different material systems should be investigated further. There are several pioneering benchmark studies of active learning in materials discovery,[138,167] but more work assessing different material systems with different final targets needs to be done. In addition, studies to converge these benchmarks and create guidelines for the applications are urged. Efforts to make active learning online in the loop of high-throughput experiments will be an important component of next-generation lab work.

### Decomposition of Complex Tasks in Material Discovery

To perform good predictions as well as gain understanding of mechanisms of complex systems, decomposition of a complex task is usually needed. Some material properties or synthesis prediction tasks, as well as the data preparation process, are complex in nature. For example, the prediction of perovskite with high Curie temperature and the multiclass classification of genomic silver nanocluster with specific fluorescence color are difficult tasks to achieve with a single prediction model.[26,68] A clever approach is to decompose this final task into simpler tasks, which normally need fewer efforts in collecting data, whereby the underlying relationships are clearer. As a result, these tasks are easier to train and could reach a higher prediction accuracy than the straightforward approach to analyze the complex task. For example, to synthesize perovskite with high Curie temperature, the formation of the perovskite structure needs to be ensured. The composition suggested by one-step regression Curie temperature prediction failed due to the presence of secondary phases (not pure perovskite structures). To fix this problem, a two-step approach is proposed to decompose the high Curie temperature prediction problem into firstly constructing an SVM classifier to predict whether a perovskite can be synthesized or not to limit composition space, followed by regression for the Curie temperature. This two-step approach realizes the final target to synthesize high-temperature ferroelectric perovskites.[26] In addition, if the final task is not achieved easily, we can try to first complete studies for simpler tasks. For example, the brightness prediction for DNA-templated Ag nanocluster can be studied first and then used as a hint to the design of DNA-templated Ag nanocluster within specific color bands. The motif mining method, high-throughput experimental procedures, and ML implementation experience developed in brightness prediction studies have become the base bricks in realizing the more complicated multiclass classification of fluorescence color of the same material system.[68,168] Although ML methods can find complex information directly from data, the decomposition of tasks can improve the prediction ability of the whole system as well as adding more transparency to the whole process. Instead of mapping the relations between process variables and device performance of photovoltaics directly, a two-step ML framework is proposed to connect target variables to material descriptors first, then to process conditions.[30] Such transparency can enable a better understanding of the whole prediction workflow, which in turn gives more interpretability. For different material systems, the detailed decompositions of tasks are different; however, there will be several general frameworks that could guide different material systems. This is potentially a valuable research direction.

### Integration of Different Applications of AI in Material Discovery

There are many different applications of ML in materials discovery. However, there is a gap between studies of applications during different stages, since each subcommunity normally specializes in their own topic. This gap should be bridged in the long term. Promising methods used in characterization should be effectively combined with ML framework in the property prediction and synthesis component to accelerate the whole process. Several ML frameworks in property prediction should be compiled together to substitute different computationally expensive simulations and be integrated into a single platform. The learned weights of well-trained property prediction models can be transferred to more complex synthesis study models with limited data to achieve better performance. The standardization of different work is preferred to ensure this level of integration. Finally, an integration of efforts through characterization, property prediction, synthesis, and theory should be compiled in high-throughput experimental systems to realize the next generation of materials laboratories.

In conclusion, the keys to successfully applying AI in different material discovery stages are a detailed analysis of the desired tasks and material data structure, a careful choice of informative descriptors, and suitable selections of models and the whole ML strategy. Through such approaches, the application of AI in material discovery will be able to solve the challenges in designing new materials with specific property requirements. This has the potential to accelerate the whole material discovery workflow by speeding up the understanding of experimentation and navigating to optima in multidimensional parameter space for 10- to 100-times acceleration.[169]

## Notations Used in This Review

$\boldsymbol{x}$, feature vector
$p()$, probability distribution
$\boldsymbol{y}$, output vector
$E()$, expected value
$f(\bullet)$, function
$\mathbb{R}$, set of real numbers
$\tilde{\boldsymbol{x}}$, reduced dimension feature vector
$\mathcal{X}$, input space
$\mathcal{F}$, feature space
$\Phi$, map used to transform the input space to a high-dimensional feature space
K, similarity function used in an SVM
$\theta$, parameters of a model
$W$, weight matrix
$b$, bias in neurons
$h$, hidden units in a neural network
$g(\bullet)$, activation function in a neural network
$z$, output from a layer of neurons before transforming by the activation function
$Y_k$, $k$th output feature map
$\boldsymbol{x}_{\text{grid}}$, input grid-like data
$W_k$, convolutional filter related to $k$th feature map
$k$, number of clusters from clustering methods
e, encoder function
r, decoder function
$L$, loss function
$C(\bullet)$, corruption term
$\widehat{\boldsymbol{x}}$, noisy feature vector
$X$, examples from true probability distribution
$P^*$, true probability distribution of data
$\widehat{P}$, learned probability distribution of data
$D$, set of tuples that forms the dataset available
$\mathcal{F}_i$, material's intrinsic information
$\mathcal{P}$, desired functional properties
$\mathcal{F}_e$, material's extrinsic information
$t$, time dimension of experiments

## ACKNOWLEDGMENTS

We acknowledge the Singapore RIE2020 Advanced Manufacturing and Engineering Programmatic grant ''Accelerated Materials Development for Manufacturing'' by the Agency for Science, Technology and Research under grant no. A1898b0043, the Ministry of Education Academic Research Fund R-279-000-532-114, and National Research Foundation, Prime Minister's Office, Singapore under its Campus for Research Excellence and Technological Enterprise program through the Singapore Massachusetts Institute of Technology Alliance for Research and Technology's Low Energy Electronic Systems research program. The authors thank X. Xue for designing figures, LT. Chen for discussing decision tree part, and H.M. Liang for proofreading.

## AUTHOR CONTRIBUTIONS

Conceptualization, X.W. and T.B.; Visualization, J.L., K.L., H.Y., P.-Y.C., and X.W.; Writing – Original Draft, J.L., K.L., H.Y., and X.W.; Writing – Review & Editing, J.L., K.L., H.Y., D.R., S.T., P.-Y.C., T.B., and X.W.; Funding Acquisition, P.-Y.C., T.B., and X.W.; Supervision, X.W.

## REFERENCES

1. Sass, S.L. (1998). The Substance of Civilization: Materials and Human History from the Stone Age to the Age of Silicon (Arcade Publishing).

2. Millard, M., Yakavets, I., Zorin, V., Kulmukhamedova, A., Marchal, S., and Bezdetnaya, L. (2017). Drug delivery to solid tumors: the predictive value of the multicellular tumor spheroid model for nanomedicine screening. Int. J. Nanomedicine *12*, 7993.

3. Berrisford, D.J., Bolm, C., and Sharpless, K.B. (1995). Ligand-accelerated catalysis. Angew. Chem. Int. Ed. *34*, 1059–1070.

4. Chirik, P.J., and Wieghardt, K. (2010). Radical ligands confer nobility on base-metal catalysts. Science *327*, 794–795.

5. Wright, M., and Uddin, A. (2012). Organic-inorganic hybrid solar cells: a comparative review. Sol. Energy Mater. Sol. Cells *107*, 87–111.

6. Amjadi, M., Kyung, K.U., Park, I., and Sitti, M. (2016). Stretchable, skin-mountable, and wearable strain sensors and their potential applications: a review. Adv. Funct. Mater. *26*, 1678–1698.

7. Chen, S., Chen, Z., Siahrostami, S., Higgins, D., Nordlund, D., Sokaras, D., Kim, T.R., Liu, Y., Yan, X., and Nilsson, E. (2018). Designing boron nitride islands in carbon materials for efficient electrochemical synthesis of hydrogen peroxide. J. Am. Chem. Soc. *140*, 7851–7859.

8. Guo, D., Shibuya, R., Akiba, C., Saji, S., Kondo, T., and Nakamura, J. (2016). Active sites of nitrogen-doped carbon materials for oxygen reduction reaction clarified using model catalysts. Science *351*, 361–365.

9. Overington, J.P., Al-Lazikani, B., and Hopkins, A.L. (2006). How many drug targets are there? Nat. Rev. Drug Discov. *5*, 993.

10. Yao, Y., Huang, Z., Xie, P., Lacey, S.D., Jacob, R.J., Xie, H., Chen, F., Nie, A., Pu, T., and Rehwoldt, M. (2018). Carbothermal shock synthesis of high-entropy-alloy nanoparticles. Science *359*, 1489–1494.

11. Graser, J., Kauwe, S.K., and Sparks, T.D. (2018). Machine learning and energy minimization approaches for crystal structure predictions: a review and new horizons. Chem. Mater. *30*, 3601–3612.

12. Oliynyk, A.O., Antono, E., Sparks, T.D., Ghadbeigi, L., Gaultois, M.W., Meredig, B., and Mar, A. (2016). High-throughput machine-learning-driven synthesis of full-heusler compounds. Chem. Mater. *28*, 7324–7331.

13. Ryan, K., Lengyel, J., and Shatruk, M. (2018). Crystal structure prediction via deep learning. J. Am. Chem. Soc. *140*, 10158–10168.

14. Butler, K.T., Davies, D.W., Cartwright, H., Isayev, O., and Walsh, A. (2018). Machine learning for molecular and materials science. Nature *559*, 547–555.

15. Hachmann, J., Olivares-Amaya, R., Atahan-Evrenk, S., Amador-Bedolla, C., Sánchez-Carrera, R.S., Gold-Parker, A., Vogt, L., Brockway, A.M., and Aspuru-Guzik, A. (2011). The harvard clean energy project: large-scale computational screening and design of organic photovoltaics on the world community grid. J. Phys. Chem. Lett. *2*, 2241–2251.

16. Jain, A., Ong, S.P., Hautier, G., Chen, W., Richards, W.D., Dacek, S., Cholia, S., Gunter, D., Skinner, D., Ceder, G., and Persson, K.A. (2013). Commentary: the materials project: a materials Genome approach to accelerating materials innovation. APL Mater. *1*, 011002.

17. Calderon, C.E., Plata, J.J., Toher, C., Oses, C., Levy, O., Fornari, M., Natan, A., Mehl, M.J., Hart, G., Buongiorno Nardelli, M., and Curtarolo, S. (2015). The AFLOW standard for high-throughput materials science calculations. Comput. Mater. Sci. *108*, 233–238.

18. Oviedo, F., Ren, Z., Sun, S., Settens, C., Liu, Z., Hartono, N.T.P., Ramasamy, S., DeCost, B.L., Tian, S.I.P., and Romano, G. (2019). Fast and interpretable classification of small X-ray diffraction datasets using data augmentation and deep neural networks. NPJ Comput. Mater. *5*, https://doi.org/10.1038/s41524-019-0196-x.

19. Wang, H., Xie, Y., Li, D., Deng, H., Zhao, Y., Xin, M., and Lin, J. (2020). Rapid identification of X-ray diffraction patterns based on very limited data by interpretable convolutional neural networks. J. Chem. Inf. Model. *60*, 2004–2011.

20. Viswanathan, G., Oliynyk, A.O., Antono, E., Ling, J., Meredig, B., and Brgoch, J. (2019). Single-crystal automated refinement (SCAR): a data-driven method for determining inorganic structures. Inorg. Chem. *58*, 9004–9015.

21. Ly, C., Olsen, A.M., Schwerdt, I.J., Porter, R., Sentz, K., McDonald, L.W., and Tasdizen, T. (2019). A new approach for quantifying morphological features of U3O8 for nuclear forensics using a deep learning model. J. Nucl. Mater. *517*, 128–137.

22. Gilmer, J., Schoenholz, S.S., Riley, P.F., Vinyals, O., and Dahl, G.E. (2017). Neural message passing for quantum chemistry. Proc. 34th Int. Conf. Mach. Learn. *70*, 1263–1272.

23. Borboudakis, G., Stergiannakos, T., Frysali, M., Klontzas, E., Tsamardinos, I., and Froudakis, G.E. (2017). Chemically intuited, large-scale screening of MOFs by machine learning techniques. NPJ Comput. Mater. *3*, https://doi.org/10.1038/s41524-017-0045-8.

24. Meredig, B., Agrawal, A., Kirklin, S., Saal, J.E., Doak, J.W., Thompson, A., Zhang, K., Choudhary, A., and Wolverton, C. (2014). Combinatorial screening for new materials in unconstrained composition space with machine learning. Phys. Rev. B *89*, 94104.

25. Gaultois, M.W., Oliynyk, A.O., Mar, A., Sparks, T.D., Mulholland, G.J., and Meredig, B. (2016). Perspective: web-based machine learning models for real-time screening of thermoelectric materials properties. Apl Mater. *4*, 53213.

26. Balachandran, P.V., Kowalski, B., Sehirlioglu, A., and Lookman, T. (2018). Experimental search for high-temperature ferroelectric perovskites guided by two-step machine learning. Nat. Commun. *9*, https://doi.org/10.1038/s41467-018-03821-9.

27. Raccuglia, P., Elbert, K.C., Adler, P.D.F., Falk, C., Wenny, M.B., Mollo, A., Zeller, M., Friedler, S.A., Schrier, J., and Norquist, A.J. (2016). Machine-learning-assisted materials discovery using failed experiments. Nature *533*, 73–76.

28. Kim, E., Huang, K., Saunders, A., McCallum, A., Ceder, G., and Olivetti, E. (2017). Materials synthesis insights from scientific literature via text extraction and machine learning. Chem. Mater. *29*, 9436–9444.

29. Yuan, R., Liu, Z., Balachandran, P.V., Xue, D., Zhou, Y., Ding, X., Sun, J., Xue, D., and Lookman, T. (2018). Accelerated discovery of large electrostrains in $BaTiO_3$-based piezoelectrics using active learning. Adv. Mater. *30*, https://doi.org/10.1002/adma.201702884.

30. Ren, Z., Oviedo, F., Thway, M., Tian, S.I.P., Wang, Y., Xue, H., Perea, J.D., Layurova, M., Heumueller, T., and Birgersson, E. (2020). Embedding physics domain knowledge into a bayesian network enables layer-by-layer process innovation for photovoltaics. NPJ Comput. Mater. *6*, https://doi.org/10.1038/s41524-020-0277-x.

31. 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*, https://doi.org/10.1038/s41524-019-0153-8.

32. Tabor, D.P., Roch, L.M., Saikin, S.K., Kreisbeck, C., Sheberla, D., Montoya, J.H., Dwaraknath, S., Aykol, M., Ortiz, C., Tribukait, H., et al. (2018). Accelerating the discovery of

materials for clean energy in the era of smart automation. Nat. Rev. Mater. *3*, 5–20.

33. Kalinin, S.V., Sumpter, B.G., and Archibald, R.K. (2015). Big-deep-smart data in imaging for guiding materials design. Nat. Mater. *14*, 973.

34. Li, H., Zhang, Z., and Liu, Z. (2017). Application of artificial neural networks for catalysis: a review. Catalysts *7*, 306.

35. Mitchell, T.M. (1997). Machine Learning (McGraw-Hill International Editions; McGraw-Hill).

36. Goodfellow, I., Bengio, Y., and Courville, A. (2016). Chapter 5. Machine learning basics. Deep Learning (MIT Press), pp. 96–161.

37. Li, J., Chen, T., Lim, K., Chen, L., Khan, S.A., Xie, J., and Wang, X. (2018). Deep learning accelerated gold nanocluster synthesis. Adv. Intell. Syst. *1*, https://doi.org/10.1002/aisy.201900029.

38. Coley, C.W., Barzilay, R., Jaakkola, T.S., Green, W.H., and Jensen, K.F. (2017). Prediction of organic reaction outcomes using machine learning. ACS Cent. Sci. *3*, 434–443.

39. Cubuk, E.D., Schoenholz, S.S., Rieser, J.M., Malone, B.D., Rottler, J., Durian, D.J., Kaxiras, E., and Liu, A.J. (2015). Identifying structural flow defects in disordered solids using machine-learning methods. Phys. Rev. Lett. *114*, 108001.

40. Kauwe, S.K., Graser, J., Vazquez, A., and Sparks, T.D. (2018). Machine learning prediction of heat capacity for solid inorganics. Integr. Mater. Manuf. Innov. *7*, 43–51.

41. Zhuo, Y., Mansouri Tehrani, A., Oliynyk, A.O., Duke, A.C., and Brgoch, J. (2018). Identifying an efficient, thermally robust inorganic phosphor host via machine learning. Nat. Commun. *9*, https://doi.org/10.1038/s41467-018-06625-z.

42. Dong, Y., Wu, C., Zhang, C., Liu, Y., Cheng, J., and Lin, J. (2019). Bandgap prediction by deep learning in configurationally hybridized graphene and boron nitride. NPJ Comput. Mater. *5*, https://doi.org/10.1038/s41524-019-0165-4.

43. Grira, N., Crucianu, M., and Boujemaa, N. (2004). Unsupervised and semi-supervised clustering: a brief survey. In A Review of Machine Learning Techniques for Processing Multimedia Content. Report of the MUSCLE European Network of Excellence (FP6), pp. 9–16.

44. Celebi, M.E. (2014). Partitional Clustering Algorithms (Springer).

45. Kaufman, L., and Rousseeuw, P.J. (2009). Finding Groups in Data: An Introduction to Cluster Analysis (John Wiley & Sons).

46. Tshitoyan, V., Dagdelen, J., Weston, L., Dunn, A., Rong, Z., Kononova, O., Persson, K.A., Ceder, G., and Jain, A. (2019). Unsupervised word embeddings capture latent knowledge from materials science literature. Nature *571*, 95.

47. Kitahara, A.R., and Holm, E.A. (2018). Microstructure cluster analysis with transfer learning and unsupervised learning. Integr. Mater. Manuf. Innov. *7*, 148–156.

48. Cunningham, P. (2008). Dimension reduction. In Machine Learning Techniques for Multimedia, M. Cord and P. Cunningham, eds. (Springer), pp. 91–112.

49. Kasun, L.L.C., Yang, Y., Huang, G.-B., and Zhang, Z. (2016). Dimension reduction with extreme learning machine. IEEE Trans. Image Process. *25*, 3906–3918.

50. Mardt, A., Pasquali, L., Wu, H., and Noé, F. (2018). VAMPnets for deep learning of molecular kinetics. Nat. Commun. *9*, https://doi.org/10.1038/s41467-017-02388-1.

51. Xie, T., France-Lanord, A., Wang, Y., Shao-Horn, Y., and Grossman, J.C. (2019). Graph dynamical networks for unsupervised learning of atomic scale dynamics in materials. Nat. Commun. *10*, 2667.

52. Lee, E.Y., Fulan, B.M., Wong, G.C.L., and Ferguson, A.L. (2016). Mapping membrane activity in undiscovered peptide sequence space using machine learning. Proc. Natl. Acad. Sci. U S A *113*, 13588–13593.

53. Tran, K., and Ulissi, Z.W. (2018). Active learning across intermetallics to guide discovery of electrocatalysts for $CO_2$ reduction and $H_2$ evolution. Nat. Catal. *1*, 696–703.

54. Bassman, L., Rajak, P., Kalia, R.K., Nakano, A., Sha, F., Sun, J., Singh, D.J., Aykol, M., Huck, P., Persson, K., and Vashishta, P. (2018). Active learning for accelerated design of layered materials. NPJ Comput. Mater. *4*, https://doi.org/10.1038/s41524-018-0129-0.

55. Talapatra, A., Boluki, S., Duong, T., Qian, X., Dougherty, E., and Arróyave, R. (2018). Autonomous efficient experiment design for materials discovery with bayesian model averaging. Phys. Rev. Mater. *2*, 113803.

56. Matthews, B.W. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochim. Biophys. Acta *405*, 442–451.

57. Dudoit, S., and Fridlyand, J. (2002). A prediction-based resampling method for estimating the number of clusters in a dataset. Genome Biol. *3*, https://doi.org/10.1186/gb-2002-3-7-research0036.

58. Thalamuthu, A., Mukhopadhyay, I., Zheng, X., and Tseng, G.C. (2006). Evaluation and comparison of gene clustering methods in microarray analysis. Bioinformatics *22*, 2405–2412.

59. Wang, K., Wang, B., and Peng, L. (2009). CVAP: validation for cluster analyses. Data Sci. J. *8*, 88–93.

60. Valle, S., Li, W., and Qin, S.J. (1999). Selection of the number of principal components: the variance of the reconstruction error criterion with a comparison to other methods. Ind. Eng. Chem. Res. *38*, 4389–4401.

61. Shenai, P.M., Xu, Z., and Zhao, Y. (2012). Applications of principal component analysis (PCA) in materials science. In Principal Component Analysis—Engineering Applications, P. Sanguanasat, ed. (IntechOpen), pp. 25–40.

62. Noh, J., Kim, J., Stein, H.S., Sanchez-Lengeling, B., Gregoire, J.M., Aspuru-Guzik, A., and Jung, Y. (2019). Inverse design of solid-state materials via a continuous representation. Matter *1*, 1370–1384.

63. Duvenaud, D., Maclaurin, D., Aguilera-Iparraguirre, J., Gómez-Bombarelli, R., Hirzel, T., Aspuru-Guzik, A., and Adams, R.P. (2015). Convolutional networks on graphs for learning molecular fingerprints. In Advances in Neural Information Processing Systems 28, C. Cortes, N.D. Lawrence, D.D. Lee, M. Sugiyama, and R. Garnett, eds. (NIPS Foundation), pp. 2215–2223.

64. Cortes, C., and Vapnik, V. (1995). Support-vector networks. Mach. Learn. *20*, 273–297.

65. Smola, A.J., and Schölkopf, B. (2004). A tutorial on support vector regression. Stat. Comput. *14*, 199–222.

66. Chorowski, J., Wang, J., and Zurada, J.M. (2014). Review and performance comparison of SVM-and ELM-based classifiers. Neurocomputing *128*, 507–516.

67. Schoenholz, S.S., Cubuk, E.D., Sussman, D.M., Kaxiras, E., and Liu, A.J. (2016). A structural approach to relaxation in glassy liquids. Nat. Phys. *12*, 469.

68. Copp, S.M., Gorovits, A., Swasey, S.M., Gudibandi, S., Bogdanov, P., and Gwinn, E.G., (2018). Fluorescence color by data-driven design of genomic silver clusters. ACS Nano *12*, https://doi.org/10.1021/acsnano.8b03404.

69. Granda, J.M., Donina, L., Dragone, V., Long, D.L., and Cronin, L. (2018). Controlling an organic synthesis robot with machine learning to search for new reactivity. Nature *559*, 377–381, https://doi.org/10.1038/s41586-018-0307-8.

70. Murthy, S.K. (1998). Automatic construction of decision trees from data: a multi-disciplinary survey. Data Min. Knowl. Discov. *2*, 345–389.

71. Kotsiantis, S.B. (2013). Decision trees: a recent overview. Artif. Intell. Rev. *39*, 261–283.

72. Breiman, L. (2001). Random forests. Mach. Learn. *45*, 5–32.

73. Friedman, J.H. (2002). Stochastic gradient boosting. Comput. Stat. Data Anal. *38*, 367–378.

74. Chen, T.; Guestrin, C. Xgboost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; ACM, 2016; pp 785–794.

75. Jain, A.K., Mao, J., and Mohiuddin, K.M. (1996). Artificial neural networks: a tutorial. Comput. (Long. Beach. Calif). *3*, 31–44.

76. Goodfellow, I., Bengio, Y., and Courville, A. (2016). Chapter 6. Deep feedforward networks. Deep Learning (MIT Press), pp. 161–223.

77. Hornik, K., Stinchcombe, M., and White, H. (1989). Multilayer feedforward networks are universal approximators. Neural Networks *2*, 359–366.

78. Montufar, G.F., Pascanu, R., Cho, K., and Bengio, Y. (2014). On the number of linear regions of deep neural networks. In Advances in Neural Information Processing Systems 27, Z. Ghahramani, M. Welling, C. Cprtes, N.D. Lawrence, and K.Q. Weinberger, eds. (NIPS Foundation), pp. 2924–2932.

79. Sanchez-Gonzalez, A., Micaelli, P., Olivier, C., Barillot, T.R., Ilchen, M., Lutman, A.A., Marinelli, A., Maxwell, T., Achner, A., Agåker, M., et al. (2017). Accurate prediction of X-ray pulse properties from a free-electron laser using machine learning. Nat. Commun. *8*, https://doi.org/10.1038/ncomms15461.

80. Cassar, D.R., de Carvalho, A.C.P.L.F., and Zanotto, E.D. (2018). Predicting glass transition temperatures using neural networks. Acta Mater. *159*, 249–256.

81. Rawat, W., and Wang, Z. (2017). Deep convolutional neural networks for image classification: a comprehensive review. Neural Comput. *29*, 2352–2449.

82. Goodfellow, I., Bengio, Y., and Courville, A. (2016). Chapter 9. Convolutional networks. Deep Learning (MIT Press), pp. 326–366.

83. Yu, D., Wang, H., Chen, P., and Wei, Z. (2014). Mixed pooling for convolutional neural networks. In International Conference on Rough Sets and Knowledge Technology, D. Miao, W. Pedrycz, D. Slezak, G. Peters, Q. Hu, and R. Wang, eds. (Springer), pp. 364–375.

84. Ziletti, A., Kumar, D., Scheffler, M., and Ghiringhelli, L.M. (2018). Insightful classification of crystal structures using deep learning. Nat. Commun. *9*, https://doi.org/10.1038/s41467-018-05169-6.

85. Ziatdinov, M., Dyck, O., Maksov, A., Li, X., Sang, X., Xiao, K., Unocic, R.R., Vasudevan, R., Jesse, S., and Kalinin, S.V. (2017). Deep learning of atomically resolved scanning transmission electron microscopy images: chemical identification and tracking local transformations. ACS Nano *11*, 12742–12752.

86. Alldritt, B., Hapala, P., Oinonen, N., Urtev, F., Krejci, O., Canova, F.F., Kannala, J., Schulz, F., Liljeroth, P., and Foster, A.S. (2020). Automated structure discovery in atomic force microscopy. Sci. Adv. *6*, eaay6913.

87. Scarselli, F., Gori, M., Tsoi, A.C., Hagenbuchner, M., and Monfardini, G. (2008). The graph neural network model. IEEE Trans. Neural Networks *20*, 61–80.

88. Altae-Tran, H., Ramsundar, B., Pappu, A.S., and Pande, V. (2017). Low data drug discovery with one-shot learning. ACS Cent. Sci. *3*, 283–293.

89. Lipton, Z.C., Berkowitz, J., and Elkan, C. (2015). A critical review of recurrent neural networks for sequence learning. arXiv, 1506.00019.

90. Goodfellow, I., Bengio, Y., and Courville, A. (2016). Chapter 10. Sequence modeling: recurrent and recursive nets. In Deep Learning (MIT Press), pp. 367–415.

91. Chung, J., Gulcehre, C., Cho, K., and Bengio, Y. (2015). Gated feedback recurrent neural networks. In Proceedings of the 32nd International Conference on Machine Learning, *vol. 37*, F. Bach and D. Blei, eds (JMLR), pp. 2067–2075.

92. Quang, D., Xie, X., and Dan, Q. (2016). A hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences. Nucleic Acids Res. *44*, e107.

93. Ripalda, J.M., Buencuerpo, J., and García, I. (2018). Solar cell designs by maximizing energy production based on machine learning clustering of spectral variations. Nat. Commun. *9*, https://doi.org/10.1038/s41467-018-07431-3.

94. Bro, R., and Smilde, A.K. (2014). Principal component analysis. Anal. Methods *6*, 2812–2831.

95. Li, Q., Nelson, C.T., Hsu, S.L., Damodaran, A.R., Li, L.L., Yadav, A.K., McCarter, M., Martin, L.W., Ramesh, R., and Kalinin, S.V. (2017). Quantification of flexoelectricity in $PbTiO_3$/$SrTiO_3$ superlattice polar vortices using machine learning and phase-field modeling. Nat. Commun. *8*, https://doi.org/10.1038/s41467-017-01733-8.

96. Zahrt, A.F., Henle, J.J., Rose, B.T., Wang, Y., Darrow, W.T., and Denmark, S.E. (2019). Prediction of higher-selectivity catalysts by computer-driven workflow and machine learning. Science *363*, eaau5631.

97. Goodfellow, I., Bengio, Y., and Courville, A. (2016). Chapter 14. Autoencoders. In Deep Learning (MIT Press), pp. 499–523.

98. Olshausen, B.A., and Field, D.J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. Nature *381*, 607.

99. Lee, H., Battle, A., Raina, R., and Ng, A.Y. (2007). Efficient sparse coding algorithms. In Advances in Neural Information Processing Systems, J.C. Platt, D. Koller, Y. Singer, and S.T. Roweis, eds. (\NIPS Foundation), pp. 801–808.

100. Vincent, P.; Larochelle, H.; Bengio, Y.; Manzagol, P.-A. Extracting and Composing Robust Features with Denoising Autoencoders. In Proceedings of the 25th International Conference on Machine Learning; ACM, 2008; pp 1096–1103.

101. Doersch, C. (2016). Tutorial on variational autoencoders. arXiv, 1606.05908.

102. Gómez-Bombarelli, R., Wei, J.N., Duvenaud, D., Hernández-Lobato, J.M., Sánchez-Lengeling, B., Sheberla, D., Aguilera-Iparraguirre, J., Hirzel, T.D., Adams, R.P., and Aspuru-Guzik, A. (2018). Automatic chemical design using a data-driven continuous representation of molecules. ACS Cent. Sci. *4*, 268–276.

103. Kim, E., Huang, K., Jegelka, S., and Olivetti, E. (2017). Virtual screening of inorganic materials synthesis parameters with deep learning. NPJ Comput. Mater. *3*, https://doi.org/10.1038/s41524-017-0055-6.

104. van der Maaten, L., and Hinton, G. (2008). Visualizing data using T-SNE. J. Mach. Learn. Res. *9*, 2579–2605.

105. Zhou, Z.-H. (2017). A brief introduction to weakly supervised learning. Natl. Sci. Rev. *5*, 44–53.

106. Settles, B. (2009). Active Learning Literature Survey (University of Wisconsin-Madison Department of Computer Sciences).

107. Brochu, E., Cora, V.M., and De Freitas, N. (2010). A tutorial on bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning. arXiv, 1012.2599.

108. Solomou, A., Zhao, G., Boluki, S., Joy, J.K., Qian, X., Karaman, I., Arróyave, R., and Lagoudas, D.C. (2018). Multi-objective Bayesian materials discovery: application on the discovery of precipitation strengthened NiTi shape memory alloys through micromechanical modeling. Mater. Des. *160*, 810–827.

109. Smith, J.S., Nebgen, B., Lubbers, N., Isayev, O., and Roitberg, A.E. (2018). Less is more: sampling chemical space with active learning. J. Chem. Phys. *148*, https://doi.org/10.1063/1.5023802.

110. Durand, A., Wiesner, T., Gardner, M.A., Robitaille, L.É., Bilodeau, A., Gagné, C., De Koninck, P., and Lavoie-Cardinal, F. (2018). A machine learning approach for online automated optimization of super-resolution optical microscopy. Nat. Commun. *9*, 5247.

111. Borodinov, N., Neumayer, S., Kalinin, S.V., Ovchinnikova, O.S., Vasudevan, R.K., and Jesse, S. (2019). Deep neural networks for understanding noisy data applied to physical property extraction in scanning probe microscopy. NPJ Comput. Mater. *5*, https://doi.org/10.1038/s41524-019-0148-5.

112. Liotti, E., Arteta, C., Zisserman, A., Lui, A., Lempitsky, V., and Grant, P.S. (2018). Crystal nucleation in metallic alloys using X-ray radiography and machine learning. Sci. Adv. *4*, https://doi.org/10.1126/sciadv.aar4004.

113. Li, L., Yang, Y., Zhang, D., Ye, Z.G., Jesse, S., Kalinin, S.V., and Vasudevan, R.K. (2018). Machine learning-enabled identification of material phase transitions based on experimental data: exploring collective dynamics in ferroelectric relaxors. Sci. Adv. *4*, eaap8672.

114. Walker, S.W.C., Anwar, A., Psutka, J.M., Crouse, J., Liu, C., Le Blanc, J.C.Y., Montgomery, J., Goetz, G.H., Janiszewski, J.S., Campbell, J.L., and Hopkins, W.S. (2018). Determining molecular properties with differential mobility spectrometry and machine learning. Nat. Commun. *9*, 5096.

115. Rashidi, M., and Wolkow, R.A. (2018). Autonomous scanning probe microscopy in situ tip conditioning through machine learning. ACS Nano *12*, 5185–5189.

116. Ziatdinov, M., Maksov, A., and Kalinin, S.V. (2017). Learning surface molecular structures via machine vision. NPJ Comput. Mater. *3*, https://doi.org/10.1038/s41524-017-0038-7.

117. Paruzzo, F.M., Hofstetter, A., Musil, F., De, S., Ceriotti, M., and Emsley, L. (2018). Chemical shifts in molecular solids by machine learning. Nat. Commun. *9*, 4501.

118. Park, W.B., Chung, J., Jung, J., Sohn, K., Singh, S.P., Pyo, M., Shin, N., and Sohn, K.-S. (2017). Classification of crystal structure using a convolutional neural network. IUCrJ *4*, 486–494.

119. Masubuchi, S., and Machida, T. (2019). Classifying optical microscope images of exfoliated graphene flakes by data-driven machine learning. NPJ 2d Mater. Appl. *3*, 4–6.

120. Feng, J., Teng, Q., He, X., and Wu, X. (2018). Accelerating multi-point statistics reconstruction method for porous media via deep learning. Acta Mater. *159*, 296–308.

121. Oxley, M.P., Yin, J., Borodinov, N., Somnath, S., Ziatdinov, M., Lupini, A.R., Jesse, S., Vasudevan, R.K., and Kalinin, S.V. (2020). Deep learning of interface structures from the 4D STEM data: cation intermixing vs. Roughening. arXiv, 2002.09039.

122. Jha, D., Singh, S., Al-Bahrani, R., Liao, W., Choudhary, A., De Graef, M., and Agrawal, A. (2018). Extracting grain orientations from EBSD patterns of polycrystalline materials using convolutional neural networks. Microsc. Microanal. *24*, 497–502.

123. Himanen, L., Geurts, A., Foster, A.S., and Rinke, P. (2019). Data-driven materials science: status, challenges, and perspectives. Adv. Sci. *6*, 1900808.

124. Kim, C., Pilania, G., and Ramprasad, R. (2016). Machine learning assisted predictions of intrinsic dielectric breakdown strength of ABX3 perovskites. J. Phys. Chem. C *120*, 14575–14580.

125. Panapitiya, G., Avendano-Franco, G., Ren, P., Wen, X., Li, Y., and Lewis, J.P. (2018). Machine-learning prediction of CO adsorption in thiolated, Ag-alloyed Au nanoclusters. J. Am. Chem. Soc. *140*, 17508–17514.

126. Hansen, K., Biegler, F., Ramakrishnan, R., Pronobis, W., Von Lilienfeld, O.A., Müller, K.R., and Tkatchenko, A. (2015). Machine learning predictions of molecular properties: accurate many-body potentials and nonlocality in chemical space. J. Phys. Chem. Lett. *6*, 2326–2331.

127. Ramprasad, R., Batra, R., Pilania, G., Mannodi-Kanakkithodi, A., and Kim, C. (2017). Machine learning in materials informatics: recent applications and prospects. NPJ Comput. Mater. *3*, https://doi.org/10.1038/s41524-017-0056-5.

128. Takahashi, K., and Tanaka, Y. (2016). Material synthesis and design from first principle calculations and machine learning. Comput. Mater. Sci. *112*, 364–367.

129. Jackson, N.E., Bowen, A.S., Antony, L.W., Webb, M.A., Vishwanath, V., and de Pablo, J.J. (2019). Electronic structure at coarse-grained resolutions from supervised machine learning. Sci. Adv. *5*, eaav1190.

130. Schmidt, J., Chen, L., Botti, S., and Marques, M.A.L. (2018). Predicting the stability of ternary intermetallics with density functional theory and machine learning. J. Chem. Phys. *148*, 241728.

131. Ma, W., Cheng, F., and Liu, Y. (2018). Deep-learning-enabled on-demand design of chiral metamaterials. ACS Nano *12*, 6326–6334.

132. Oliynyk, A.O., Adutwum, L.A., Rudyk, B.W., Pisavadia, H., Lotfi, S., Hlukhyy, V., Harynuk, J.J., Mar, A., and Brgoch, J. (2017). Disentangling structural confusion through machine learning: structure prediction and polymorphism of equiatomic ternary phases ABC. J. Am. Chem. Soc. *139*, 17870–17881.

133. Lu, S., Zhou, Q., Ouyang, Y., Guo, Y., Li, Q., and Wang, J. (2018). Accelerated discovery of stable lead-free hybrid organic-inorganic perovskites via machine learning. Nat. Commun. *9*, 3405.

134. Chmiela, S., Sauceda, H.E., Müller, K.R., and Tkatchenko, A. (2018). Towards exact molecular dynamics simulations with machine-learned force fields. Nat. Commun. *9*, https://doi.org/10.1038/s41467-018-06169-2.

135. Rouet-Leduc, B., Barros, K., Lookman, T., and Humphreys, C.J. (2016). Optimisation of GaN LEDs and the reduction of efficiency droop using active machine learning. Sci. Rep. *6*, https://doi.org/10.1038/srep24862.

136. Jennings, P.C., Lysgaard, S., Hummelshøj, J.S., Vegge, T., and Bligaard, T. (2019). Genetic algorithms for computational materials discovery accelerated by machine learning. NPJ Comput. Mater. *5*, https://doi.org/10.1038/s41524-019-0181-4.

137. Yamashita, T., Sato, N., Kino, H., Miyake, T., Tsuda, K., and Oguchi, T. (2018). Crystal structure prediction accelerated by Bayesian optimization. Phys. Rev. Mater. *2*, 13803.

138. Gopakumar, A.M., Balachandran, P.V., Xue, D., Gubernatis, J.E., and Lookman, T. (2018). Multi-objective optimization for materials discovery via adaptive design. Sci. Rep. *8*, https://doi.org/10.1038/s41598-018-21936-3.

139. Hu, L., Wang, X., Wong, L., and Chen, G. (2003). Combined first-principles calculation and neural-network correction approach for heat of formation. J. Chem. Phys. *119*, 11501–11507.

140. Ramakrishnan, R., Dral, P.O., Rupp, M., and Von Lilienfeld, O.A. (2015). Big data meets quantum chemistry approximations: the Δ-machine learning approach. J. Chem. Theor. Comput. *11*, 2087–2096.

141. Smith, J.S., Nebgen, B.T., Zubatyuk, R., Lubbers, N., Devereux, C., Barros, K., Tretiak, S., Isayev, O., and Roitberg, A.E. (2019). Approaching coupled cluster accuracy with a general-purpose neural network potential through transfer learning. Nat. Commun. *10*, https://doi.org/10.1038/s41467-019-10827-4.

142. Yamada, H., Liu, C., Wu, S., Koyama, Y., Ju, S., Shiomi, J., Morikawa, J., and Yoshida, R. (2019). Predicting materials properties with little data using shotgun transfer learning. ACS Cent. Sci. *5*, 1717–1730.

143. Jain, A., Shin, Y., and Persson, K.A. (2016). Computational predictions of energy materials using density functional theory. Nat. Rev. Mater. *1*, 15004.

144. Gao, J., Liu, Y., Wang, Y., Hu, X., Yan, W., Ke, X., Zhong, L., He, Y., and Ren, X. (2017). Designing high dielectric permittivity material in barium titanate. J. Phys. Chem. C *121*, 13106–13113.

145. Iwasaki, Y., Takeuchi, I., Stanev, V., Kusne, A.G., Ishida, M., Kirihara, A., Ihara, K., Sawada, R., Terashima, K., Someya, H., et al. (2019). Machine-learning guided discovery of a new thermoelectric material. Sci. Rep. *9*, https://doi.org/10.1038/s41598-019-39278-z.

146. Ren, F., Ward, L., Williams, T., Laws, K.J., Wolverton, C., Hattrick-Simpers, J., and Mehta, A. (2018). Accelerated discovery of metallic glasses through iteration of machine learning and high-throughput experiments. Sci. Adv. *4*, eaaq1566.

147. Cao, B., Adutwum, L.A., Oliynyk, A.O., Luber, E.J., Olsen, B.C., Mar, A., and Buriak, J.M. (2018). How to optimize materials and devices via design of experiments and machine learning: demonstration using organic photovoltaics. ACS Nano *12*, 7434–7444.

148. Anoop Krishnan, N.M., Mangalathu, S., Smedskjaer, M.M., Tandia, A., Burton, H., and Bauchy, M. (2018). Predicting the dissolution kinetics of silicate glasses using machine learning. J. Non. Cryst. Sol. *487*, 37–45.

149. Umehara, M., Stein, H.S., Guevarra, D., Newhouse, P.F., Boyd, D.A., and Gregoire, J.M. (2019). Analyzing machine learning models to accelerate generation of fundamental materials insights. NPJ Comput. Mater. *5*, https://doi.org/10.1038/s41524-019-0172-5.

150. Rosenbrock, C.W., Homer, E.R., Csányi, G., and Hart, G.L.W. (2017). Discovering the building blocks of atomic systems using machine learning: application to grain boundaries. NPJ Comput. Mater. *3*, https://doi.org/10.1038/s41524-017-0027-x.

151. Orme, A.D., Chelladurai, I., Rampton, T.M., Fullwood, D.T., Khosravani, A., Miles, M.P., and Mishra, R.K. (2016). Insights into twinning in Mg AZ31: a combined EBSD and machine learning study. Comput. Mater. Sci. *124*, 353–363.

152. Moosavi, S.M., Chidambaram, A., Talirz, L., Haranczyk, M., Stylianou, K.C., and Smit, B. (2019). Capturing chemical intuition in synthesis of metal-organic frameworks. Nat. Commun. *10*, https://doi.org/10.1038/s41467-019-08483-9.

153. Molnar, C. (2019). Interpretable Machine Learning: A Guide for Making Black Box Models Explainable (Christoph Molnar).

154. Chen, T., Fung, V., Yao, Q., Luo, Z., Jiang, D., and Xie, J. (2018). Synthesis of water-soluble $[Au_{25}(SR)_{18}]^-$ using a stoichiometric amount of $NaBH_4$. J. Am. Chem. Soc. *140*, 11370–11377.

155. Bartók, A.P., Kondor, R., and Csányi, G. (2013). On representing chemical environments. Phys. Rev. B *87*, 184115.

156. Goodall, R.E.A., and Lee, A.A. (2019). Predicting materials properties without crystal structure: deep representation learning from stoichiometry. arXiv, 1910.00617.

157. Wang, A.Y.-T., Kauwe, S.K.K., Murdock, R.J., and Sparks, T.D. (2020). Compositionally-restricted attention-based network for materials property prediction. ChemRxiv. https://chemrxiv.org/articles/Compositionally-Restricted_Attention-Based_Network_for_Materials_Property_Prediction/11869026/1.

158. Jinich, A., Sánchez-Lengeling, B., Ren, H., Harman, R., and Aspuru-Guzik, A. (2018). A mixed quantum chemistry/machine learning approach for the fast and accurate prediction of biochemical redox potentials and its large-scale application to 315 000 redox reactions. ACS Cent. Sci. *5*, 1199–1210.

159. Ouyang, R., Curtarolo, S., Ahmetcik, E., Scheffler, M., and Ghiringhelli, L.M. (2018). SISSO: a compressed-sensing method for identifying the best low-dimensional descriptor in an immensity of offered candidates. Phys. Rev. Mater. *2*, 83802.

160. Kim, B., Lee, S., and Kim, J. (2020). Inverse design of porous materials using artificial neural networks. Sci. Adv. *6*, eaax9324.

161. Ren, Z.; Oviedo, F.; Xue, H.; Thway, M.; Zhang, K.; Li, N.; Perea, J.D.; Layurova, M.; Wang, Y.; Tian, S. Physics-guided characterization and optimization of solar cells using surrogate machine learning model. In 2019 IEEE 46th Photovoltaic Specialists Conference (PVSC); IEEE, 2019; pp 3054–3058.

162. Kazarian, S.G., and Chan, K.L.A. (2006). Applications of ATR-FTIR spectroscopic imaging to biomedical samples. Biochim. Biophys. Acta *1758*, 858–867.

163. Xie, T., and Grossman, J.C. (2018). Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. Phys. Rev. Lett. *120*, 145301.

164. Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., and Yu, P.S. (2019). A comprehensive survey on graph neural networks. arXiv, 1901.00596.

165. Beluch, W.H.; Genewein, T.; Nürnberger, A.; Köhler, J.M. The power of ensembles for active learning in image classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2018; pp 9368–9377.

166. Settles, B. (2012). Active Learning (Morgan & Claypool).

167. del Rosario, Z., Kim, Y., Rupp, M., Antono, E., and Ling, J. (2019). Assessing the frontier: active learning, model accuracy, and multi-objective materials discovery and optimization. arXiv, 1911.03224.

168. Copp, S.M., Bogdanov, P., Debord, M., Singh, A., and Gwinn, E. (2014). Base motif recognition and design of DNA templates for fluorescent silver clusters by machine learning. Adv. Mater. *26*, 5839–5845.

169. Correa-Baena, J.-P., Hippalgaonkar, K., van Duren, J., Jaffer, S., Chandrasekhar, V.R., Stevanovic, V., Wadia, C., Guha, S., and Buonassisi, T. (2018). Accelerating materials development via automation, machine learning, and high-performance computing. Joule *2*, 1410–1420.