

Representations of Materials for Machine Learning

James Damewood,¹ Jessica Karaguesian,^{1,2}
Jaclyn R. Lunger,¹ Aik Rui Tan,¹ Mingrou Xie,^{1,3}
Jiayu Peng,¹ and Rafael Gómez-Bombarelli¹

¹Department of Materials Science and Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA; email: jkdamewo@mit.edu, lungerja@mit.edu, atan14@mit.edu, jypeng@mit.edu, rafagh@mit.edu

²Center for Computational Science and Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA; email: jkarag@mit.edu

³Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA; email: mrx@mit.edu

ANNUAL
REVIEWS **CONNECT**

www.annualreviews.org

- Download figures
- Navigate cited references
- Keyword search
- Explore related articles
- Share via email or social media

Annu. Rev. Mater. Res. 2023. 53:399–426

First published as a Review in Advance on
April 18, 2023

The *Annual Review of Materials Research* is online at
matsci.annualreviews.org

<https://doi.org/10.1146/annurev-matsci-080921-085947>

Copyright © 2023 by the author(s). This work is licensed under a Creative Commons Attribution 4.0 International License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. See credit lines of images or other third-party material in this article for license information.



Keywords

representation, feature engineering, machine learning, materials science, crystal structure, generative model

Abstract

High-throughput data generation methods and machine learning (ML) algorithms have given rise to a new era of computational materials science by learning the relations between composition, structure, and properties and by exploiting such relations for design. However, to build these connections, materials data must be translated into a numerical form, called a representation, that can be processed by an ML model. Data sets in materials science vary in format (ranging from images to spectra), size, and fidelity. Predictive models vary in scope and properties of interest. Here, we review context-dependent strategies for constructing representations that enable the use of materials as inputs or outputs for ML models. Furthermore, we discuss how modern ML techniques can learn representations from data and transfer chemical and physical information between tasks. Finally, we outline high-impact questions that have not been fully resolved and thus require further investigation.

1. INTRODUCTION

Energy and sustainability applications demand the rapid development of scalable new materials technologies. Big data and machine learning (ML) have been proposed as strategies to rapidly identify needle-in-the-haystack materials that have the potential for revolutionary impact.

High-throughput experimentation platforms based on robotized laboratories can increase the efficiency and speed of synthesis and characterization. However, in many practical open problems, the number of possible design parameters is too large to be analyzed exhaustively. Virtual screening somewhat mitigates this challenge by using physics-based simulations to suggest the most promising candidates, reducing the cost but also the fidelity of the screens (1–3).

Over the past decade, hardware improvements, new algorithms, and the development of large-scale repositories of materials data (4–9) have enabled a new era of ML methods. In principle, predictive ML models can identify and exploit nontrivial trends in high-dimensional data to achieve accuracy comparable with or superior to first-principles calculations but with an orders-of-magnitude reduction in cost. In practice, while a judicious model choice is helpful in moving toward this ideal, such ML methods are also highly dependent on the numerical inputs used to describe systems of interest—the so-called representations. Only when the representation is composed of a set of features and descriptors from which the desired physics and chemistry are emergent can the promise of ML be achieved.

Thus, the problem that materials informatics researchers must answer is, How can we best construct this representation? Previous work has provided practical advice for constructing materials representations (10–14), namely that (*a*) the similarity or difference between two data points should match the similarity or difference between representations of those two data points, (*b*) the representation should be applicable to the entire materials domain of interest, and (*c*) the representation should be easier to calculate than the target property.

Representations should reflect the degree of similarity between data points such that similar data have similar representations and as data points become more different their representations diverge. Indeed, the definition of similarity will depend on the application. Consider, as an example, a hypothetical model predicting the electronegativity of an element, excluding noble gases. One could attempt to train the model using atomic number as input, but this representation violates the above principle, as atoms with a similar atomic number can have significantly different electronegativities (e.g., fluorine and sodium), forcing the model to learn a sharply varying function whose changes appear at irregular intervals. Alternatively, a representation using period and group numbers would closely group elements with similar atomic radii and electron configurations. Over this new domain, the optimal prediction will result in a smoother function that is easier to learn.

The approach used to extract representation features from raw inputs should be feasible over the entire domain of interest—all data points used in training and deployment. If data required to construct the representation are not available for a particular material, ML screening predictions cannot be made.

Finally, for the ML approach to remain a worthwhile investment, the computational cost of obtaining representation features and descriptors for new data should be smaller than that of obtaining the property itself through traditional means, either experimentally or with first-principles calculations. If, for instance, accurately predicting a property calculated by density-functional theory (DFT) with ML requires input descriptors obtained from DFT on the same structure and at the same level of theory, the ML model does not offer any benefit.

A practicing materials scientist will notice a number of key barriers to forming property-informative representations that satisfy these criteria. First, describing behavior often involves quantifying structure-to-property relationships across length scales. The diversity of possible

atomistic structure types considered can vary over space groups, supercell size, and disorder parameters. This challenge motivates researchers to develop flexible representations capable of capturing local and global information based on atomic positions. Beyond this idealized picture, predicting material performance relies upon understanding the presence of defects, the characteristics of the microstructure, and reactions at interfaces. Addressing these concerns requires extending previous notions of structural similarity or developing new specialized tools. Furthermore, atomistic structural information is not available without experimental validation or extensive computational effort (15, 16). Therefore, when predictions are required for previously unexplored materials, models must rely on more readily available descriptors such as those based on elemental composition and stoichiometry. Lastly, due to experimental constraints, data sets in materials science can often be scarce, sparse, and restricted to relatively few and self-similar examples. The difficulty in constructing a robust representation in these scenarios has inspired strategies to leverage information from high-quality representations built for closely related tasks through transfer learning.

In this review, we analyze how representations of solid-state materials (**Figure 1**) can be developed given constraints on the format, quantity, and quality of available data. We discuss the justifications, benefits, and trade-offs of different approaches. This discussion is meant to highlight methods of particular interest rather than provide exhaustive coverage of the literature.

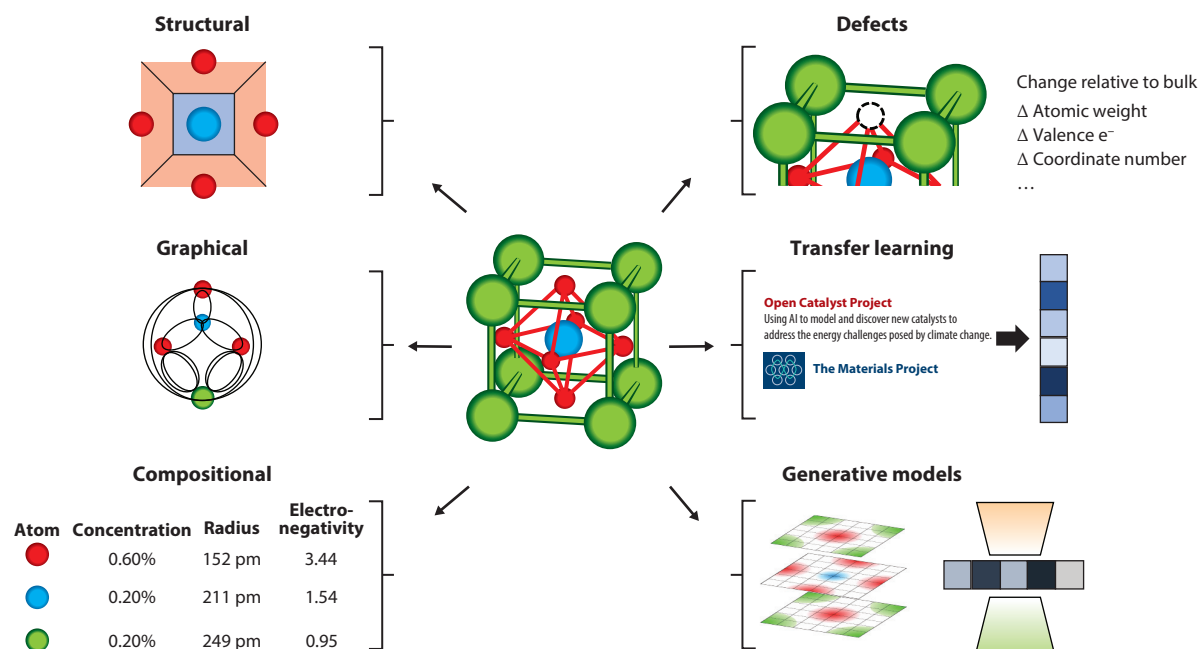


Figure 1

Summary of representations for perovskite SrTiO_3 . (*Top left*) 2D cross section of Voronoi decomposition. Predictive features can be constructed from neighbors and the geometric shape of cells (17). (*Middle left*) Crystal graph of SrTiO_3 constructed assuming periodic boundary conditions and used as an input to graph neural networks (18). (*Bottom left*) Compositional data, including concentrations, and easily accessible atomic features, such as electronegativities and atomic radii (19). Data taken from Reference 20. (*Top right*) Deviations from a pristine bulk structure induced by an oxygen vacancy to predict formation energy (21). (*Middle right*) Representations can be learned from large repositories using deep neural networks. The latent physical and chemical information can be leveraged in related but data-scarce tasks. (*Bottom right*) Training of generative models capable of proposing new crystal structures by placing atoms in discretized volume elements (22–25).

We comment on current limitations and open problems whose solutions would have high impact. In summary, we provide readers with an introduction to the current state of the field and exciting directions for future research.

2. STRUCTURAL FEATURES FOR ATOMISTIC GEOMETRIES

Simple observations in material systems (e.g., higher ductility of face-centered cubic metals compared with body-centered cubic metals) have made it evident that material properties are highly dependent on crystal structure, from coordination and atomic ordering to broken symmetries and porosity. For a computational materials scientist, this presents the question of how to algorithmically encode information from a set of atom types ($a_1, a_2, a_3, \dots$), positions ($x_1, x_2, x_3, \dots$), and primitive cell parameters into a feature set that can be effectively utilized in ML.

For ML methods to be effective, it is necessary that the machine-readable representation of a material's structure fulfills the criteria outlined in Section 1 (10–14). Notably, scalar properties (such as heat capacity or reactivity) do not change when translations, rotations, or permutations of atom indexing are applied to the atomic coordinates. Therefore, to ensure representations reflect the similarities between atomic structures, the representations should also be invariant to those symmetry operations.

2.1. Local Descriptors

One strategy to form a representation of a crystal structure is to characterize the local environment of each atom and consider the full structure as a combination of local representations. This concept was applied by Behler & Parrinello (26), who proposed the atom-centered symmetry functions (ACSFs). ACSF descriptors (**Figure 2a**) can be constructed using radial, G_i^1 , and angular, G_i^2 , symmetry functions centered on atom i ,

$$G_i^1 = \sum_{j \neq i}^{\text{neighbors}} e^{-\eta(R_{ij}-R_s)^2} f_c(R_{ij}), \quad 1.$$

$$G_i^2 = 2^{1-\zeta} \sum_{j,k \neq i}^{\text{neighbors}} (1 + \lambda \cos \theta_{ijk})^\zeta e^{-\eta(R_{ij}^2 + R_{ik}^2 + R_{jk}^2)} f_c(R_{ij}) f_c(R_{ik}) f_c(R_{jk}), \quad 2.$$

with the tunable parameters λ , R_s , η , and ζ . R_{ij} is the distance between the central atom i and atom j , and θ_{ijk} corresponds to the angle between the vector from the central atom to atom j and the vector from the central atom to atom k . The cutoff function f_c screens out atomic interactions beyond a specified cutoff radius and ensures the locality of the atomic interactions. Because symmetry functions rely on relative distances and angles, they are rotationally and translationally invariant. Local representations can be constructed from many symmetry functions of the type G_i^1 and G_i^2 with multiple settings of tunable parameters to probe the environment at varying distances and angular regions. With the set of localized symmetry functions, neural networks can then predict local contributions to a particular property and approximate global properties as the sum of local contributions. The flexibility of this approach allows for modification of the G_i^1 and G_i^2 functions (27, 28) or higher-capacity neural networks for element-wise prediction (28).

In search of a representation with fewer hand-tuned parameters and a more rigorous definition of similarity, Bartók et al. (12) proposed a rotationally invariant kernel for comparing environments based on the local atomic density. Given a central atom, the Smooth Overlap of Atomic Positions (SOAP) defines the atomic density function $\rho(\mathbf{r})$ as a sum of Gaussian functions centered at each neighboring atom within a cutoff radius (**Figure 2b**). The choice of Gaussian function is motivated

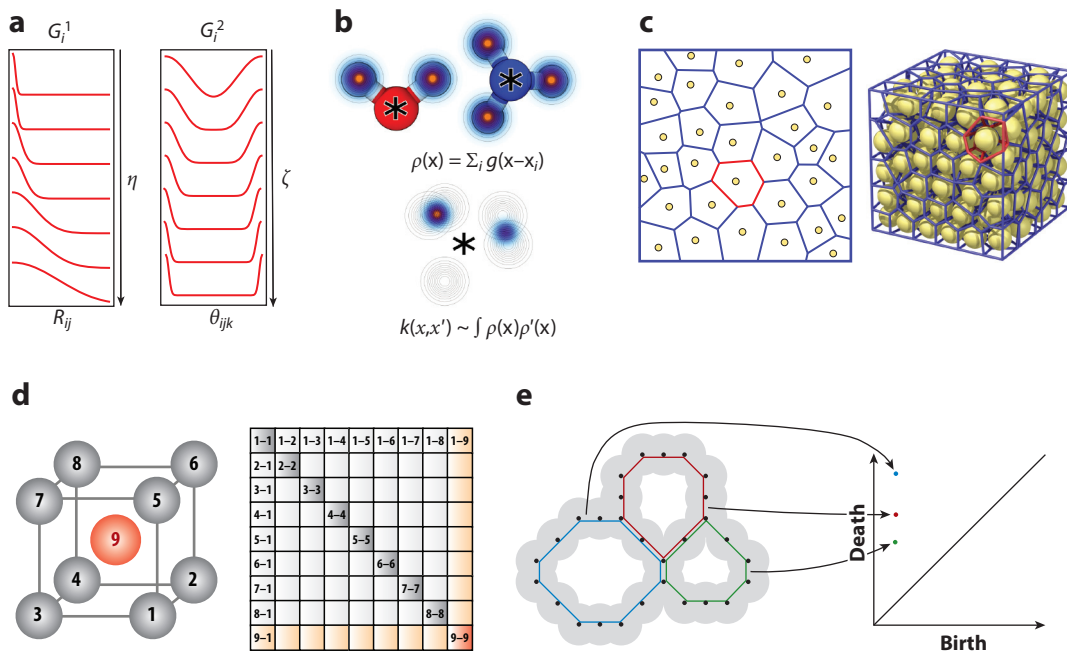


Figure 2

(a) Examples of radial, G_i^1 , and angular, G_i^2 , symmetry functions from the local atom-centered symmetry function descriptor proposed by Behler & Parrinello (26). (b) In the Smooth Overlap of Atomic Positions (SOAP) descriptor construction, the atomic neighborhood density of a central atom is defined by the sum of Gaussian functions around each neighboring atom. A kernel function can then be built to compare the different environments by computing the density overlap of the atomic neighborhood functions. Panel adapted with permission from Reference 36 (CC BY-NC 4.0). (c) Voronoi tessellation in two and three dimensions. Yellow circles and spheres show particles while the blue lines equidistantly divide the space between two neighboring particles. Polygonal spaces encompassed by the blue lines are the Voronoi cells. Panel adapted from Reference 37 (CC BY 4.0). (d) Illustration of a Coulomb matrix where each element in the matrix shows the Coulombic interaction between the labeled particles in the system on the left. Diagonal elements show self-interactions. (e) The births and deaths of topological holes in (left) a point cloud are recorded on (right) a persistence diagram. Persistent features lie far from the parity line and indicate more significant topological features. Panel adapted from Reference 38 (CC BY 4.0).

by the intuition that representations should be continuous such that small changes in atomic positions should result in correspondingly small changes in the metric between two configurations. With a basis of radial functions $g_n(\mathbf{r})$ and spherical harmonics $Y_{lm}(\theta, \phi)$, $\rho(\mathbf{r})$ for central atom i can be expressed as

$$\rho_i(\mathbf{r}) = \sum_j \exp\left(-\frac{|\mathbf{r} - \mathbf{r}_{ij}|^2}{2\sigma^2}\right) = \sum_{nlm} c_{nlm} g_n(\mathbf{r}) Y_{lm}(\hat{\mathbf{r}}), \quad 3.$$

and the kernel can be computed (12, 29) as

$$K(\rho, \rho') = \mathbf{p}(\mathbf{r}) \cdot \mathbf{p}'(\mathbf{r}), \quad 4.$$

$$\mathbf{p}(\mathbf{r}) \equiv \sum_m c_{nlm} (c_{n'lm})^*, \quad 5.$$

where c_{nlm} are the expansion coefficients in Equation 3. In practice, $\mathbf{p}(\mathbf{r})$ can be used as a vector descriptor of the local environment and is also referred to as a power spectrum (12). SOAP has

demonstrated extraordinary versatility for materials applications both as a tool for measuring similarity (30) and as a descriptor for ML algorithms (31). Furthermore, the SOAP kernel can be used to compare densities of different elements by adding an additional factor that provides a definition for similarity between atoms where, for instance, atoms in the same group could have higher similarity (29). The mathematical connections between different local atomic density representations including ACSFs and SOAP are elucidated by a generalized formalism introduced by Willat et al. (32), offering a methodology through which the definition of new variants can be clarified.

Instead of relying on the density of nearby atoms, local representations can be derived from a Voronoi tessellation of a crystal structure. The Voronoi tessellation segments space into cells such that each cell contains one atom and the surrounding spatial region for which the contained atom is closer than any other atom (**Figure 2c**). From these cells, Ward et al. (17) identified a set of descriptive features including an effective coordination number computed using the area of the faces, lengths and volumes of nearby cells, ordering of the cells based on elements, and atomic properties of nearest neighbors weighted by the area of the intersecting face. When combined with compositional features (19), their representation results in better performance on predictions of formation enthalpy for the Inorganic Crystal Structure Database than partial radial distribution functions (PRDFs) (33) (see also figure 1 in Reference 17). In subsequent work, these descriptors have facilitated the prediction of experimental heat capacities in MOFs (34). Similarly, Isayev et al. (35) replaced faces of the Voronoi tessellation with virtual bonds and separated the resulting framework into sets of linear (up to four atoms) and shell-based (up to nearest neighbors) fragments. Additional features related to the atomic properties of constituent elements were associated with each fragment, and the resulting vectors were concatenated with attributes of the supercell. In addition to demonstrating accurate predictive capabilities, models could be interpreted through the properties of the various fragments. For instance, predictions of the band gap could be correlated with the difference in ionization potential in two-atom linear fragments, a trend that could be exploited to design the properties of materials through tuning of the composition (35).

2.2. Global Descriptors

Alternatively, to more explicitly account for interactions beyond a fixed cutoff, atom types and positions can be encoded into a global representation that reflects geometric and physical insight. Inspired by the importance of electrostatic interactions in chemical stability, Rupp et al. (39) proposed the Coulomb matrix (**Figure 2d**), which models the potential between electron clouds:

$$M_{i,j} = \begin{cases} Z_i^{2.4} & \text{for } i = j \\ \frac{Z_i Z_j}{|r_i - r_j|} & \text{for } i \neq j \end{cases} \quad 6.$$

Due to the fact that off-diagonal elements are dependent on only relative distances, Coulomb matrices are rotation and translation invariant. However, the representation is not permutation invariant since changing the labels of the atoms will rearrange the elements of the matrix. While originally developed for molecules, the periodicity of crystal structures can be added to the representation by considering images of atoms in adjacent cells, replacing the $\frac{1}{|r_i - r_j|}$ dependence with another function with the same small distance limit and periodicity that matches the parent lattice, or using an Ewald sum to account for long-range interactions (11). BIGDML (40) further improved results by restricting predictions from the representation to be invariant to all symmetry operations within the space group of the parent lattice and demonstrated effective implementations on tasks ranging from H interstitial diffusion in Pt to phonon density of states (DOS). While this approach has been able to effectively model long-range physics, these representations rely on a

fixed supercell and may not be able to achieve the same chemical generality as local environments (40).

Global representations have also been implemented with higher-order tensors. PRDFs are 3D non-permutation-invariant matrices $g_{\alpha\beta r}$ whose elements correspond to the density of element β in the environments of element α at radius r (33). The many-body tensor representation (MBTR) provides a more general framework (10) that can quantify k -body interactions and account for chemical similarity between elements. The MBTR is translationally, rotationally, and permutationally invariant and can be applied to crystal structures by summing over atoms in the primitive cell. While MBTR exhibited better performance than SOAP or Coulomb matrices for small molecules, its accuracy may not extend to larger systems (10).

Another well-established method for representing crystal structures in materials science is the cluster expansion. Given a parent lattice and a decoration σ defining the element that occupies each site, Sanchez et al. (41) sought to map this atomic ordering to material properties and proposed evaluating the correlations between sites through a set of cluster functions. Each cluster function Φ is constructed from a product of basis functions ϕ over a subset of sites. To ensure the representation is appropriately invariant, symmetrically equivalent clusters are grouped into classes denoted by α . The characteristics of the atomic ordering can be quantified by averaging cluster functions over the decoration $\langle \Phi_\alpha \rangle_\sigma$, and properties q of the configuration can be predicted as

$$q(\sigma) = \sum_{\alpha} J_{\alpha} m_{\alpha} \langle \Phi_{\alpha} \rangle_{\sigma},$$

where m_{α} is a multiplicity factor that accounts for the rate of appearance of different cluster types, and J_{α} are parameters referred to as effective cluster interactions that must be determined from fits to data (42). While cluster expansions have been constructed for decades and provided useful models for configurational disorder and alloy thermodynamics (42), cluster expansions assume the structure of the parent lattice, and models cannot generally be applied across different crystal structures (43, 44). Furthermore, due to the increasing complexity of selecting cluster functions, implementations are restricted to binary and ternary systems without special development (45). Additional research has extended the formalism to continuous environments (atomic cluster expansion) by treating σ as pairwise distances instead of site occupancies and constructing ϕ from radial functions and spherical harmonics (46). The atomic cluster expansion framework has provided a basis for more sophisticated deep learning approaches (47).

2.3. Topological Descriptors

Topological data analysis has found favor over the past decade in characterizing structure in complex, high-dimensional data sets. When applied to the positions of atoms in amorphous or crystalline structures, topological methods reveal underlying geometric features that inform behavior in downstream predictions, such as phase changes, reactivity, and separations. In particular, persistent homology (PH) is able to identify significant structural descriptors that are both machine readable and physically interpretable. The data can be probed at different length scales (formally called filtrations) by computing a series of complexes that each include all sets of points where all pairwise distances are less than the corresponding length (48). Analysis of complexes by homology in different dimensions reveals holes or voids in the data manifold, which can be described by the range of length scales they are observed at (persistences) as well as when they are produced (births) and disappear (deaths). Emergent features with significant persistence values are less likely to be caused by noise in the data or as an artifact of the chosen length scales. In practice, multiple persistences, births, and deaths produced from a single material

can be represented together by persistence diagrams (**Figure 2e**) or undergo additional feature engineering to generate a variety of descriptors as ML inputs (49).

While PH has been applied to crystal structures in the Open Quantum Materials Database (50), the method is particularly useful in the analysis of porous materials. The identified features (births, deaths, persistences) hold direct physical relevance to traditional structural features used to describe pore geometries. For instance, persistent 2D deaths represent the largest sphere that can be included inside the pores of the materials. Krishnapriyan et al. (38) have shown that these topological descriptors outperform traditional structural descriptors when predicting carbon dioxide adsorption under varying conditions for metal-organic frameworks (MOFs), as did Lee et al. (51) for zeolites for methane storage capacities. Representative cycles can trace the topological features back to the atoms that are responsible for the hole or the void, creating a direct relationship between structure and predicted performance (see figure 4 in Reference 38). Similarity methods for comparing barcodes can then be used to identify promising novel materials with similar pore geometries for targeted applications.

A caveat is that PH does not inherently account for system size and is thus size dependent. The radius cutoff, or the supercell size, needs to be carefully considered to encompass all significant topological features and allow comparison across systems of interest. In the worst case scenario, the computation cost per filtration for a structure is $O(N^3)$, where N is the number of sets of points defining a complex. Although the cost is alleviated by the sparsity of the boundary matrix (52), the scaling is poor for structures whose geometric features exceed unit cell lengths. The benefit of using PH features to capture more complex structural information has to be carefully balanced with the cost of generating these features.

3. LEARNING ON PERIODIC CRYSTAL GRAPHS

In the previous section, we describe many physically inspired descriptors that characterize materials and can be used to efficiently predict properties. The use of differentiable graph-based representations in convolutional neural networks, however, mitigates the need for manual engineering of descriptors (53, 54). Indeed, advances in deep learning and the construction of large-scale materials databases (4–9) have made it possible to learn representations directly from structural data. From a set of atoms $a_1, a_2, a_3, \dots$ located at positions $x_1, x_2, x_3, \dots$, materials can be converted to a graph $G(V, E)$ defined as the set of atomic nodes V and the set of edges E connecting neighboring atoms. Many graph-based neural network architectures were originally developed for molecular systems, with edges representing bonds. By considering periodic boundary conditions and defining edges as connections between neighbors within a cutoff radius, graphical representations can be leveraged for crystalline systems. The connectivity of the crystal graph thus naturally encodes local atomic environments (18).

When used as inputs to ML algorithms, the graph nodes and edges are initialized with an associated set of features. Nodal features can be as simple as a one-hot vector of the atomic number or can explicitly include other properties of the atomic species (e.g., electronegativity, group, period). Edge features are typically constructed from the distance between the corresponding atoms. Subsequently, a series of convolutions parameterized by neural networks modifies node and/or edge features based on the current state of their neighborhood (**Figure 3a**). As the number of convolutions increases, interactions from further away in the structure can propagate, and graph features become tuned to reflect the local chemical environment. Finally, node and edge features can be pooled to form a single vector representation for the material (53, 55).

Crystal Graph Convolution Neural Networks (CGCNN) (18) and Materials Graph Networks (MEGNet) (56) have become benchmark algorithms capable of predicting properties across solid-state materials domains including bulk, surfaces, disordered systems, and 2D materials

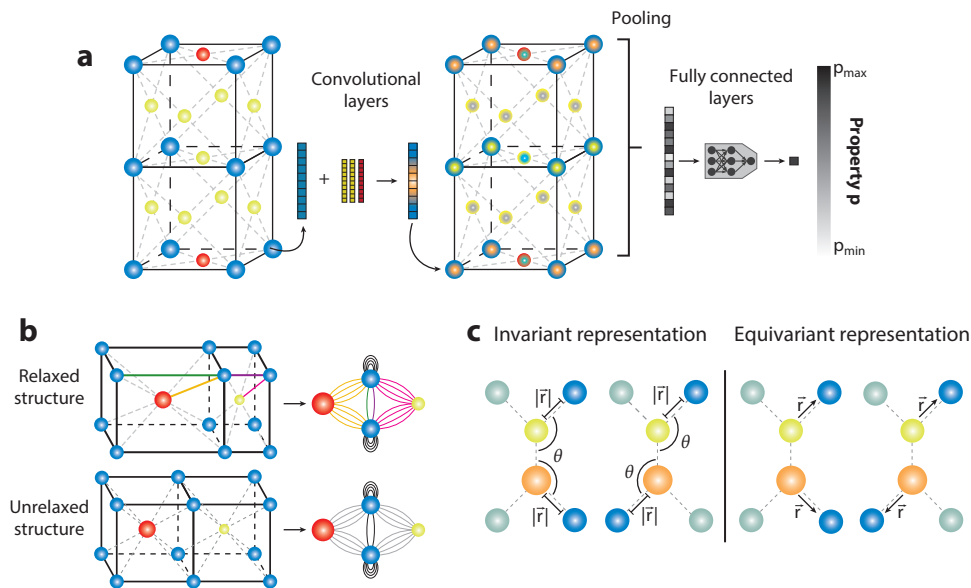


Figure 3

(a) General architecture of graph convolutional neural networks for property prediction in crystalline systems. Three-dimensional crystal structure is represented as a graph with nodes representing atoms and edges representing connections between nearby atoms. Features (e.g., nodal, edge, angular) within local neighborhoods are convolved, pooled into a crystal-wide vector, and mapped to the target property. Panel adapted from Reference 62. (b) Information loss in graphs built from pristine structures. Geometric distortions of ground-state crystal structures are captured as differing edge features in graphical representations. This information is lost in graphs constructed from corresponding unrelaxed structures. (c) Ability of graphical representations to distinguish toy structures. Assuming a sufficiently small cutoff radius, the invariant representation—using edge lengths and/or angles—cannot distinguish the two toy arrangements, while the equivariant representation with directional features can. Panel adapted from Reference 64. Abbreviation: CGCNN, Crystal Graph Convolution Neural Network.

(57, 58). Atomistic Line Graph Neural Network (ALIGNN) extended these approaches by including triplet three-body features in addition to nodes and edges and exhibited superior performance to CGCNN over a broad range of regression tasks including formation energy, band gap, and shear modulus (59). Other variants have used information from Voronoi polyhedra to construct graphical neighborhoods and augment edge features (60) or initialized node features based on the geometry and electron configuration of nearest-neighbor atoms (61).

While these methods have become widespread for property prediction, graph convolution updates based on only the local neighborhood may limit the sharing of information related to long-range interactions or extensive properties. Gong et al. (63) demonstrated that these models can struggle to learn materials properties reliant on periodicity, including characteristics as simple as primitive cell lattice parameters. As a result, while graph-based learning is a high-capacity approach, performance can vary substantially by the target use case. In some scenarios, methods developed primarily for molecules can be effectively implemented out of the box with the addition of periodic boundary conditions, but especially in the case of long-range physical phenomena, optimal results can require specialized modeling.

Various strategies to account for this limitation have been proposed. Gong et al. (63) found that if the pooled representation after convolutions was concatenated with human-tuned descriptors,

errors could be reduced by 90% for related predictions, including phonon internal energy and heat capacity. Algorithms have attempted to more explicitly account for long-range interactions by modulating convolutions with a mask defined by a local basis of Gaussians and a periodic basis of plane waves (65), employing a unique global pooling scheme that could include additional context such as stoichiometry (66), or constructing additional features from the reciprocal representation of the crystal (67). Other strategies have leveraged assumptions about the relationships among predicted variables, such as representing phonon spectra by using a Gaussian mixture model (68).

Given the promise and flexibility of graphical models, improving the data efficiency, accuracy, generalizability, and scalability of these representations is an active area of research. While our discussion above of structure-based material representations relied on the translational and rotational invariance of scalar properties, this characteristic does not continue to hold for higher-order tensors. Consider a material with a net magnetic moment. If the material is rotated 180° around an axis perpendicular to the magnetization, the net moment then points in the opposite direction. The moment was not invariant to the rotation but instead transformed alongside the operation in an equivariant manner (69). For a set of transformations described by the group G , equivariant functions f satisfy $g * f(x) = f(g * x)$ for every input x and every group element g (69, 70). Recent efforts have shown that by introducing higher-order tensors to node and edge features (**Figure 3c**) and restricting the update functions such that intermediate representations are equivariant to the group $E3$ (encompassing translations, rotations, and reflections in R^3), models can achieve state-of-the-art accuracy on benchmark data sets and even exhibit comparable performance to structural descriptors in low-data [(O)100 data points] regimes (64, 69, 71). Further accuracy improvements can be made by explicitly considering many-body interactions beyond edges (47, 72, 73). Such models, developed for molecular systems, have since been extended to solid-state materials and shown exceptional performance. Indeed, Chen et al. (74) trained an equivariant model to predict phonon DOS and was able to screen for high-heat capacity targets, tasks identified to be particularly challenging for baseline CGCNN and MEGNet models (63). Therefore, equivariant representations may offer a more general alternative to the specialized architectures described above.

A major restriction of these graph-based approaches is the requirement for the positions of atomic species to be known. In general, ground-state crystal structures exhibit distortions that allow atoms to break symmetries, which are computationally modeled with expensive DFT calculations. Graphs generated from pristine structures lack representation of relaxed atomic coordinates (**Figure 3b**), and resulting model accuracy can degrade substantially (75, 76). These graph-based models are therefore often most effective at predicting properties of systems for which significant computational resources have already been invested, thus breaking the advice from Section 1. As a result, their practical usage often remains limited when searching broad regions of crystal space for an optimal material satisfying a particular design challenge.

Strategies have therefore been developed to bypass the need for expensive quantum calculations by using unrelaxed crystal prototypes as inputs. Gibson et al. (76) trained CGCNN models on data sets composed of both relaxed structures and a set of perturbed structures that map to the same property value as the fully relaxed structure. The data augmentation incentivizes the CGCNN model to predict similar properties within some basin of the fully relaxed structure and was demonstrated to improve prediction accuracy on an unrelaxed test set. Alternatively, graph-based energy models can be used to modify unrelaxed prototypes by searching through a fixed set of possibilities (77) or using Bayesian optimization (78) to find structures with lower energy. Lastly, structures can be relaxed using a cheap surrogate model (e.g., a force field) before a final prediction is made. The accuracy and efficiency of such a procedure will fundamentally rely on the validity and compositional generalizability of the surrogate relaxation approach (75).

4. CONSTRUCTING REPRESENTATIONS FROM STOICHIOMETRY

The phase, crystal system, and atomic position of materials are not always available when modeling materials systems, rendering structural and graphical representations impossible to construct. In the absence of these data, material representations can also be built purely from stoichiometry (the concentration of the constituent elements) and without knowledge of the geometry of the local atomistic environments. Despite their lack of structural information and apparent simplicity, these methods provide unique benefits for materials science researchers. First, descriptors used to form compositional representations such as common atomic properties (e.g., atomic radii, electronegativity) do not require computational overhead and can be readily found in existing databases (19). In addition, effective models can often be built using standard algorithms for feature selection and prediction that are implemented in freely available libraries (79), increasing accessibility to non-experts when compared with structural models. Lastly, when used as tools for high-throughput screening, compositional models identify a set of promising elemental concentrations. Compared with the suggestion of particular atomistic geometries, stoichiometric approaches may be more robust as they make weaker assumptions about the outcomes of attempted syntheses.

Composition-based rules have long contributed to efficient materials design. William Hume-Rothery and Linus Pauling designed rules for determining the formation of solid solutions and crystal structures that include predictions based on atomic radii and electronic valence states (80, 81). However, many exceptions to their predictions can be found (82).

ML techniques offer the ability to discover and model relationships between properties and physical descriptors through statistical means. Meredig et al. (83) demonstrated that a decision tree ensemble trained using a feature set of atomic masses, positions in the periodic table, atomic numbers, atomic radii, electronegativities, and valence electrons could outperform a conventional heuristic on predicting whether ternary compositions would have formation energies <100 meV/atom. Ward et al. (19) significantly expanded this set to 145 input properties, including features related to the distribution and compatibility of the oxidation states of constituent atoms. Their released implementation, MagPie, can be a useful benchmark or starting point for the development of further research methods (79, 84, 85). Furthermore, if a fixed structural prototype (e.g., elpasolite) is assumed, these stoichiometric models can be used to analyze compositionally driven variation in properties (86, 87).

Even more subtle yet extremely expressive low-dimensional descriptors can be obtained by initializing a set with standard atomic properties and computing successive algebraic combinations of features, with each calculation being added to the set and used to compute higher-order combinations in the next round. While the resulting set will grow exponentially, compressive sensing can then be used to identify the most promising descriptors from sets that can exceed 10^9 possibilities (88, 89). Ghiringhelli et al. (90) found descriptors that could accurately predict the relative stability of zinc blende and rock salt phases for binary compounds, and Bartel et al. (91) identified an improved tolerance factor τ for the formation of perovskite systems (Table 1). While these approaches do not derive their results from a known mechanism, they do provide

Table 1 Example descriptors determined through compressive sensing

Descriptor	Prediction	Variables
$\frac{IP(B)-EA(B)}{r_p(A)^2}$	Energy difference between AB structure prototypes	IP, ionization potential EA, electron affinity r_p , density of valence of p orbital
$\frac{r_X}{r_B} - n_A(n_A - \frac{r_A/r_B}{\ln(r_A/r_B)})$	Stability of ABX_3 perovskite	n_Y , oxidation state of Y r_Y , ionic radius of Y

enough interpretability to enable the extraction of physical insights for the screening and design of materials.

When large data sets are available, deep neural networks tend to outperform traditional approaches, and that is also the case for compositional representations. The size of modern materials science databases has enabled the development of information-rich embeddings that map elements or compositions to vectors as well as the testing and validation of deep learning models. Chemically meaningful embeddings can be constructed by counting all compositions in which that element appeared in the Materials Project (92) or learned through the application of natural language processing to previously reported results in the scientific literature (93). These data-hungry methods were able to demonstrate that their representations could be clustered based on atomic group (92) and could be used to suggest new promising compositions based on similarity with the best known materials (93). The advantages of training deep learning algorithms with large data sets are exemplified by ElemNet, which uses only a vector of fractional stoichiometry as input. Despite its apparent simplicity, when >3,000 training points were available, ElemNet performed better than a MagPie-based model at predicting formation enthalpies (94).

While the applicability of ElemNet is limited to problem domains with $O(10^3)$ data points, more recent methods have significantly reduced this threshold. ROOST (95) represented each composition as a fully connected graph with nodes as elements, and properties were predicted using a message-passing scheme with an attention mechanism that relied on the stoichiometric fraction of each element. ROOST substantially improved on ElemNet, achieving better performance than MagPie in cases with only hundreds of training examples. Meanwhile, CrabNet (96) forms element-derived matrices as a sum of embeddings of each element's identity and stoichiometric fraction. This approach achieves similar performance to ROOST by updating the representation using self-attention blocks. The fractional embedding can take log-scale data as input such that even dopants in small concentrations can have a significant effect on predictions. Despite the inherent challenges of predicting properties purely from composition, these recent and significant modeling improvements suggest that continued algorithmic development could be an attractive and impactful direction for future research projects.

Compositional models have the advantage that they can suggest new systems to experimentalists without requiring a specific atomic geometry and, likewise, can learn from experimental data without necessitating an exact crystal structure (97). Owing to their ability to incorporate experimental findings into ML pipelines and provide suggestions with fewer experimental requirements (e.g., synthesis of a particular phase), compositional models have become attractive methods for materials design. Zhang et al. (97) trained a compositional model using atomic descriptors on previous experimental data to predict Vicker's hardness and validated their model by synthesizing and testing eight metal disilicides. Oliynyk et al. (87) identified new Heusler compounds while also verifying their approach on negative cases where they predicted a synthesis would fail. Another application of their approach enabled the prediction of the crystal structure prototype of ternary compounds with greater than 96% accuracy. By training their model to predict the probability associated with each structure, they were able to experimentally engineer a system (TiFeP) with multiple competing phases (98).

While researchers have effectively implemented compositional models as methods for materials design, their limitations should be considered when selecting a representation for ML studies. Fundamentally, compositional models will only provide a single prediction for each stoichiometry regardless of the number of synthesizable polymorphs. While training models to only predict properties of the lowest-energy structure is physically justifiable (99), extrapolation to technologically relevant metastable systems may still be limited. Additionally, graph-based structural models such as CGCNN (18) or MEGNet (56) generally outperform compositional

models (84). Therefore, composition models are most practically applicable when atomistic resolution of materials is unavailable and thus structural representations cannot be effectively constructed.

5. DEFECTS, SURFACES, AND GRAIN BOUNDARIES

Mapping the structure of small molecules and unit cells to materials properties has been a reasonable starting point for many applications of materials science modeling. However, materials design often requires understanding larger length scales beyond the small unit cell, such as in defect and grain boundary engineering and in surface science (100). In catalysis, for example, surface activity is highly facet dependent and cannot be modeled using the bulk unit cell alone. It has been shown that the (100) facet of RuO_2 , a state-of-the-art catalyst for the oxygen evolution reaction (OER), has an order-of-magnitude-higher current for OER than the active site on the thermodynamically stable (110) facet (101). Similarly, small unit cells are not sufficient for modeling transport properties, where size, orientation, and characteristics of grain boundaries play a large role. In order to apply ML to practical materials design, it is therefore imperative to construct representations that can characterize environments at the relevant length scales.

Defect engineering offers a common and significant degree of freedom through which materials can be tuned. Data science can contribute to the design of these systems as fundamental mechanisms are often not completely understood even in longstanding cases such as carbon in steels (103). Dragoni et al. (31) developed a Gaussian approximate potential (104) using SOAP descriptors for face-centered cubic iron that could probe vacancies, interstitials, and dislocations, but their model was confined to a single phase of one element and required DFT calculations incorporating $O(10^5)$ unique environments to build the interpolation.

Considering that even a small number of possible defects significantly increases combinatorial complexity, a general approach for predicting properties of defects from pristine bulk structure representations could accelerate computation by orders of magnitude (**Figure 4a**). For example, Varley et al. (102) observed simple and effective linear relationships between vacancy formation energy and descriptors derived from the band structure of the bulk solid. While their model only considered one type of defect, their implementation limits computational expense by demonstrating that only DFT calculations on the pristine bulk were required (102). Structure- and composition-aware descriptors of the pristine bulk have additionally been shown to be predictive

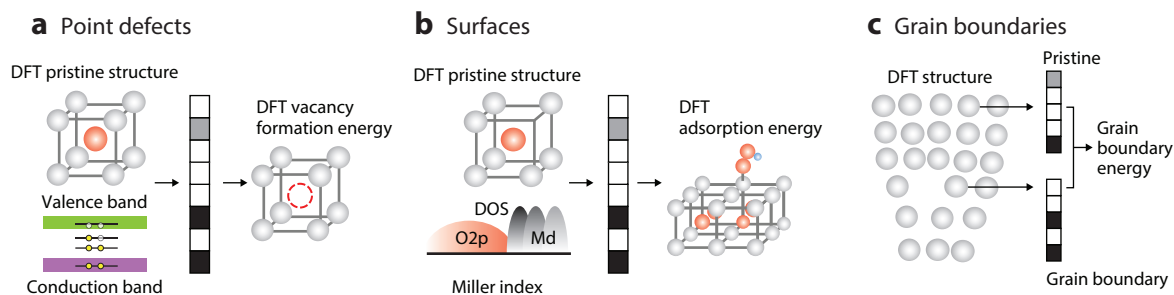


Figure 4

(a) Point defect properties are learned from a representation of the pristine bulk structure and additional relevant information on conduction and valence band levels. Panel adapted with permission from Reference 102; copyright 2017 American Chemical Society. (b) Surface properties are learned from a combination of pristine bulk structure representation, Miller index, and DOS information. (c) Local environments of atoms near a grain boundary and atoms in the pristine bulk are compared to learn grain boundary properties. Abbreviations: DFT, density-functional theory; DOS, density of states.

of vacancy formation in metal oxides (105, 106) and site and antisite defects in AB intermetallics (107). To develop an approach that can be used over a broad range of chemistries and defect types, Frey et al. (21) formed a representation by considering relative differences in characteristics (atomic radii, electronegativity, etc.) of the defect structure compared with the pristine parent. Furthermore, because reference bulk properties could be estimated using surrogate ML models, no DFT calculations were required for prediction of either formation energy or changes in electronic structure (21). We also note that in some cases it may be judicious to design a model that does not change significantly in the presence of defects. For these cases, representations based on simulated diffraction patterns are resilient to site-based vacancies or displacements (108).

Like in defect engineering, ML for practical design of catalyst materials requires representations beyond the single unit cell. Design of catalysts with high activity crucially depends on interactions of reaction intermediates with materials surfaces based on the Sabatier principle, which argues that activity is greatest when intermediates are bound neither too weakly nor too strongly (109). From a computational perspective, determining absorption energies involves searches over possible adsorption active sites, surface facets, and surface rearrangements, leading to a combinatorial space that can be infeasible to exhaustively cover with DFT. Single-dimension descriptors based on electronic structure have been established that can predict binding strengths and provide insight on tuning catalyst compositions, such as the metal d-band center for metals (110) and oxygen 2p-band center for metal oxides (111). Additional geometric approaches include describing the coordination of the active site (generalized coordination number in metals, adjusted generalized coordination number in metal oxides) (112). Based on the success of these simple descriptors, ML models have been developed to learn binding energy using the DOS and geometric descriptors of the pristine bulk structure as features (**Figure 4b**) (113).

However, these structural and electronic descriptors are often not generalizable across chemistries (110, 114), limiting the systems over which they can be applied and motivating the development of more sophisticated ML techniques. To reduce the burden on high-throughput DFT calculations, active learning with surrogate models using information from pure metals and active-site coordination has been used to identify alloy and adsorbate pairs that have the highest likelihood of producing near-optimal binding energies (115). Furthermore, when sufficient data (>10,000 examples) are available, modifications of graph-convolutional models have also predicted binding energies with high accuracy even in data sets with up to 37 elements, enabling discovery without detailed mechanistic knowledge (114). To generalize these results, the release of Open Catalyst 2020 and its related competition (6, 9) have provided over one million DFT energies for training new models and a benchmark through which new approaches can be evaluated (75). While significant advancements have been made, state-of-the-art models still exhibit high errors for particular adsorbates and nonmetallic surface elements, constraining the chemistry over which effective screening can be conducted (75). Furthermore, the complexity of the design space relevant for ML models grows considerably when accounting for interactions between adsorbates and different surface facets (116).

Beyond atomistic interactions, a material's mechanical and thermal behavior can be significantly modulated by processing conditions and the resulting microstructure. Greater knowledge of local distortions introduced at varying grain boundary incident angles would give computational materials scientists a more complete understanding of how experimentally chosen chemistries and synthesis parameters will translate into device performance. Strategies to quantify characteristics of grain boundary geometry have included reducing computational requirements by identifying the most promising configurations with virtual screening (117), estimating the grain boundary free volume as a function of temperature and bulk composition (118), treating the microstructure as a graph of nodes connected across a grain boundary (54, 119), and predicting the energetics,

and hence feasibility, of solute segregation (120). While the previous approaches did not include features based on the constituent atoms and were only benchmarked on systems with up to three elements, recent work has demonstrated that the excess energy of the grain boundary relative to the bulk can be approximated across compositions with five variables defining its orientation and the bond lengths within the grain boundary (**Figure 4c**) (121).

Further research has tried to map the local grain boundary structure to function. Algorithmic approaches to grain boundary structure classification have been developed [see, for example, VoroTop (122)], but such approaches typically rely on expert users and do not provide a continuous representation that can smoothly interpolate between structures (123). To eliminate these challenges, Rosenbrock et al. (124) proposed computing SOAP descriptors for all atoms in the grain boundary, clustering vectors into classes, and identifying grain boundaries through their local environment classes. The representation was not only predictive of grain boundary energy, temperature-dependent mobility, and shear coupling but also provided interpretable effects of particular structures within the grain boundary. A related approach computed SOAP vectors relative to the bulk structure when analyzing thermal conductivity (125). Representations based on radial and angular structure functions can also quantify the mobility of atoms within a grain boundary (126). When combined, advancing models for grain boundary stability as well as structure-to-property relationships opens the door for the functional design of grain boundaries.

6. TRANSFERABLE INFORMATION BETWEEN REPRESENTATIONS

Applications of ML to materials science are limited by the scope of compositions and structures over which algorithms can maintain sufficient accuracy. Thus, building large-scale, diverse data sets is the most robust strategy to ensure trained models can capture the relevant phenomena. However, in most contexts, materials scientists are confronted with sparsely distributed examples. Ideally, models can be trained to be generalizable and exhibit strong performance across chemistries and configurations even with few to no data points in a given domain. In order to achieve this, representations and architectures must be chosen such that models can learn to extrapolate beyond the space observed in the training set. Effective choices often rely on inherent natural laws or chemical features that are shared between the training set and extrapolated domain such as physics constraints (127, 128), the geometric (129, 130) and electronic (131, 132) structure of local environments, and positions of elements in the periodic table (133, 134). For example, Li et al. (129) were able to predict absorption energies on high-entropy-alloy surfaces after training on transition metal data by using the coordination number and electronic properties of neighbors at the active site. While significant advancements have been made in the field, extrapolation of ML models across materials spaces typically requires specialized research methods and is not always feasible.

Likewise, it is not always practical for a materials scientist to improve model generality by just collecting more data. In computational settings, some properties can be reliably estimated only with more expensive, higher levels of theory, and for experimentalists, synthetic and characterization challenges can restrict throughput. The deep learning approaches that have demonstrated exceptional performance over a wide range of tests cases discussed in this review can require at least $\sim 10^3$ training points, putting them seemingly out for the realm of possibility for many research projects. Instead, predictive modeling may fall back on identifying relationships between a set of human-engineered descriptors and target properties.

Alternatively, the hidden, intermediate layers of deep neural networks can be conceptualized as a learned vector representation of the input data. While this representation is not directly

interpretable, it must still contain physical and chemical information related to the prediction task, which downstream layers for the network utilize to generate model outputs. Transfer learning leverages these learned representations from task A and uses them in the modeling of task B. Critically, task A can be chosen to be one for which a large number of data points are accessible (e.g., prediction of all DFT formation energies in the Materials Project), and task B can be of limited size (e.g., predicting experimental heats of formation of a narrow class of materials). In principle, if task A and task B share an underlying physical basis (the stability of the material), the features learned when modeling task A may be more informationally rich than a human-designed representation (135). With this more effective starting point, subsequent models for task B can reach high accuracy with relatively few new examples.

The most straightforward methods to implement transfer learning in the materials science community follow a common procedure: (a) train a neural network model to predict a related property (task A) for which $>O(1,000)$ data points are available (pretraining), (b) fix parameters of the network up to a chosen depth d (freezing), and (c) given the new data set for task B, either retrain the remaining layers, where parameters can be initialized randomly or from the task A model (fine-tuning), or treat the output of the model at depth d as an input representation for another ML algorithm (feature extraction) (136, 137). The robustness of this approach has been demonstrated across model classes including those using only composition [ElemNet (135, 137), ROOST (95)], crystal graphs [CGCNN (138)], and equivariant convolutions [GemNet (139)]. Furthermore, applications of task B range from experimental data (95, 135) to DFT-calculated surface absorption energies (139).

The sizes of the data sets for task A and task B will determine the effectiveness of a transfer learning approach in two ways. First, the quality and robustness of the representation learned for task A will increase as the number of observed examples (the size of data set A) increases. Secondly, as the size of data set B decreases, data become too sparse for an ML model to learn a reliable representation alone, and prior information from the solution to task A can provide an increasingly useful method to interpolate between the few known points. Therefore, transfer learning typically exhibits the greatest boosts in performance when task A has orders-of-magnitude more data than task B (135, 138).

In addition, the quality of information sharing through transfer learning depends on the physical relationship between task A and task B. Intuitively, the representation from task A provides a better guide for task B if the tasks are closely related. For example, Kolluru et al. (139) demonstrated that transfer learning from models trained on the Open Catalyst Dataset (6) exhibited significantly better performance when applied to the absorption of new species than energies of less-related small molecules. While it is difficult to choose the optimal task A for a given task B a priori, shotgun transfer learning (136) has demonstrated that the best pairing can be chosen experimentally by empirically validating a large pool of possible candidates and selecting top performers.

The depth d from which features should be extracted from task A to form a representation can also be task dependent. Kolluru et al. (139) provided evidence that to achieve optimal performance, more layers of the network should be allowed to be retrained in step (c) as the connection between task A and task B becomes more distant. Gupta et al. (137) arrived at a similar conclusion that the early layers of deep neural networks learned more general representations and performed better in cross-property transfer learning. Inspired by this observation that representations at different neural network layers contain information with varying specificity to a particular prediction task, representations for transfer learning that combine activations from multiple depths have been proposed (139, 140).

When tasks are sufficiently different, freezing neural network weights may not be the optimal strategy, and instead representations for task B can include predictions for task A as descriptors. For instance, Cubuk et al. (141) observed that structural information was critical to predict Li conductivity but was only available for a small set of compositions for which crystal structures were determined. By training a separate surrogate model to predict structural descriptors from the composition and using those approximations in subsequent Li conductivity models, the feasible screening domain was expanded by orders of magnitude. Similarly, Greenman et al. (142) used $O(10,000)$ time-dependent DFT calculations to train a graph neural network, the estimates of which could be used as an additional descriptor for a model predicting experimental peaks in absorption spectra. Representations have also been sourced from the output of generative models. Kong et al. (143) trained a generative adversarial network (GAN) to sample electronic DOS given a particular material composition. Predictions of absorption spectra were improved by concatenating stoichiometric data with the average DOS sampled from the generative model.

7. GENERATIVE MODELS FOR INVERSE DESIGN

While, in principle, ML methods can significantly reduce the time required to compute material properties, and material scientists can employ these models to screen for a set of target systems by rapidly estimating the stability and performance, the space of feasible materials precludes a naive global optimization strategy in most cases. Generative models including variational autoencoders (VAEs) (1, 144), GANs (145, 146), and diffusion models (147, 148) can be trained to sample from a target distribution and have proven to be capable strategies for optimization in high-dimensional molecular spaces (1, 149). While some lessons can be drawn from the efforts of researchers in the computational chemistry community, generative models face unique challenges for proposing crystals (150, 151). First, the diversity of atomic species increases substantially when compared with small organic molecules. In addition, given a composition, a properly defined crystal structure requires both the positions of the atoms within the unit cell and the lattice vectors and angles that determine the system's periodicity. This definition is not unique, and the same material can be described after rotations or translations of atomic coordinates as well as integer scaling of the original unit cell. Lastly, many state-of-the-art materials for catalysis (e.g., zeolites, MOFs) can have unit cells including >100 of atoms, increasing the dimensionality of the optimization problem (150, 151).

One attempt to partially address the challenges of generative modeling for solid materials design is a voxel representation (150), in which unit cells are divided into volume elements and models are built using techniques from computer vision. Hoffmann et al. (22) represented unit cells using a density field that could be further segmented into atomic species and were able to generate crystals with realistic atomic spacings. However, atoms could be mistakenly decoded into other species with a nearby atomic number and most of the generated structures could not be stably optimized with a DFT calculation. Alternate approaches could obtain more convincing results, but over a confined region of material space (154). iMatgen (**Figure 5a**) invertibly mapped all unit cells into a cube with Gaussian-smeared atomic density and trained a VAE coupled with a surrogate energy prediction. The model was able to rediscover stable structures but was constrained over the space of vanadium oxides (23). A similar approach constructed a separate voxel representation for each element and employed a GAN trained alongside an energy constraint to explore the phases of Bi-Se (155). In order to resolve some of the limitations of Hoffmann et al. (22), Court et al. (24) reduced segmentation errors by augmenting the representation with a matrix describing the occupation (0,1) of each voxel and a matrix recording the atomic number of occupied voxels. Their model was able to propose new materials that exhibited

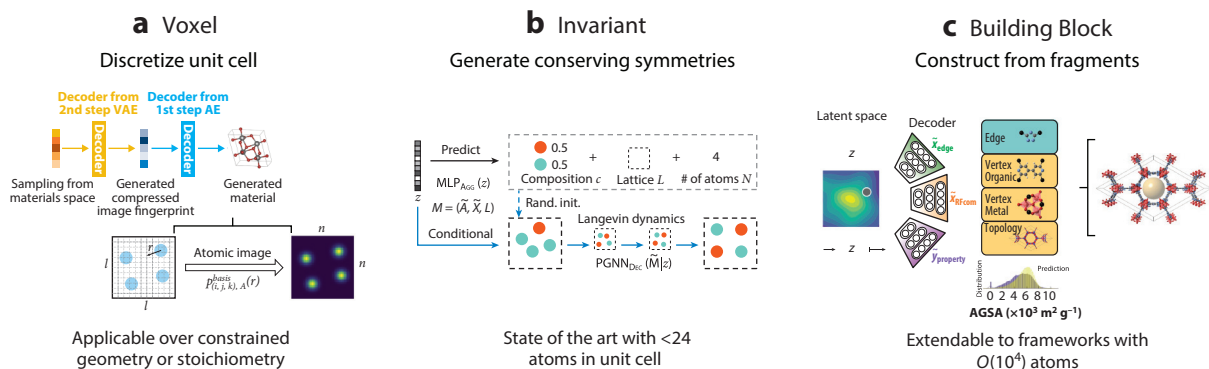


Figure 5

Approaches for crystal structure generative models. (a) Models based on voxel representations defined positions of atoms by discretizing space into finite volume elements but were not applied generally over the space of crystal structures (22–25, 154). Panel adapted with permission from Reference 23. (b) Restricting the generation process to be invariant to permutations, translations, and rotations through an appropriately constrained periodic decoder (PGNN_{DIC}) results in sampling structures exhibiting more diversity and stability. Panel adapted with permission from Reference 152. (c) When features of the material can be assumed, such as a finite number of possible topologies connecting substructures, the dimensionality of the problem can be substantially reduced and samples over larger unit cell materials can be generated. Panel adapted with permission from Reference 153. Abbreviations: AE, autoencoder; VAE, variational autoencoder.

chemical diversity and could be further optimized with DFT, but its analysis was restricted to cubic systems. Likewise, compositions of halide perovskites with optimized band gaps can be proposed using a voxelized representation of a fixed perovskite prototype (25).

Voxel representations can be relaxed to continuous coordinates in order to develop methods that are more comprehensively applicable over crystal space. Kim et al. (156) represented materials using a record of the unit cell as well as a point cloud of fractional coordinates of each element. The approach proposed lower-energy structures than iMatgen for V–O binaries and was also applicable over more diverse chemical spaces (Mg–Mn–O ternaries). Another representation including atomic positions along with elemental properties can be leveraged for inverse design over spaces that vary in both composition and lattice structure. In a test case, the model successfully generated new materials with negative formation energy and promising thermoelectric power factor (154). While these models have demonstrated improvements in performance, they lack the translational, rotational, and scale invariances of real materials and are restricted to sampling particular materials classes (152, 156).

Recently, alternatives that account for these symmetries have been proposed. Fung et al. (157) proposed a generative model for rotationally and translationally invariant atom-centered symmetry functions from which target structures could be reconstructed. Crystal diffusion VAEs (Figure 5b) leveraged periodic graphs and SE(3) equivariant message-passing layers to encode and decode their representation in a translationally and rotationally invariant way (152). They also proposed a two-step generation process during which they first predicted the crystal lattice from a latent vector and subsequently sampled the composition and atomic positions through Langevin dynamics. Furthermore, they established well-defined benchmark tasks and demonstrated that for inverse design their method was more flexible than voxel models with respect to a crystal system and more accurate than point cloud representations at identifying crystals with low formation energy.

Scaling solid-state generative modeling techniques to unit cells with 10⁴ atoms would enable inverse design of porous materials that are impossible to explore exhaustively but demonstrate

exceptional technological relevance. Currently, due to the high number of degrees of freedom, sampling from these spaces requires imposing physical constraints in the modeling process. Such restrictions can be implemented as postprocessing steps or integrated into the model representation. ZeoGAN (158) generated positions of oxygen and silicon atoms in a $32 \times 32 \times 32$ grid to propose new zeolites. While some of the atomic positions proposed directly from their model violated conventional geometric rules, they could obtain feasible structures by filtering out divergent compositions and repairing bond connectivity through the insertion or deletion of atoms. Alternatively, Yao et al. (153) designed geometric constraints directly into the generative model by representing MOFs by their edges, metal/organic vertices, and distinct topologies (RFcodes) (Figure 5c). Because this representation is invertible, all RFcodes correspond to a structurally possible MOF. By training a VAE to encode and decode this RFcode representation, they demonstrated the ability to interpolate between structures and optimize properties. In general, future research should balance more stable structure generation against the possible discovery of new motifs and topologies.

8. DISCUSSION

In this review, we introduce strategies for designing representations for ML in the context of challenges encountered by materials scientists. We discuss local and global structural features as well as representations learned from atomic-scale data in large repositories. We note additional research that extends beyond idealized crystals to include the effects of defects, surfaces, and microstructure. Furthermore, we acknowledge that in practice the availability of data can be limited both in quality and in quantity. We describe methods to mitigate this including developing models based on compositional descriptors alone or leveraging information from representations built for related tasks through transfer learning. Finally, we analyze how generative models have improved by incorporating symmetries and domain knowledge. As data-based methods have become increasingly essential for materials design, optimal ML techniques will play a crucial role in the success of research programs. The previous sections demonstrate that the choice of representation will be among these pivotal factors and that novel approaches can open the door to new modes of discovery. Motivated by these observations, we conclude by summarizing open problems with the potential to have a large impact on the field of materials design.

8.1. Trade-Offs of Local and Global Structural Descriptors

Local structural descriptors including SOAP (12) have become reliable metrics to compare environments with a specific cutoff radius and, when properties can be defined through short-range interactions, have demonstrated strong predictive performance. Characterizing systems based off of local environments allows models to extrapolate to cases where global representations may vary substantially (e.g., an extended supercell of a crystal structure) (14) and enables highly scalable methods of computation that can extend the practical limit of simulations to much larger systems (159). However, Unke et al. (160) notes that the required complexity of the representation can grow quickly when modeling systems with many distinct elements, and the quality of ML predictions will be sensitive to the selected hyperparameters, such as the characteristic distances and angles in atom-centered symmetry functions. Furthermore, it is unclear if these high-quality results extend to materials characteristics that strongly depend on long-range physics or the periodicity of the crystal. On the other hand, recent global descriptors (40) can more explicitly model these phenomena but have not exhibited the same generality across space groups and system sizes. Strategies exploring appropriate combinations of local and long-range features (161) have the potential to break through these trade-offs to provide more universal models for material property prediction.

8.2. Prediction from Unrelaxed Crystal Prototypes

If relaxed structures are required to form representations, the space over which candidates can be screened is limited to those materials for which optimized geometries are known. Impressively, recent work (162, 163) has shown that ML force fields, even simple models with relatively high errors, can be used to optimize structures and obtain converged results that are lower in energy than those obtained using Vienna Ab initio Simulation Package (164). Their benchmarking on the OC20 (6) data set and lower accuracy requirements suggest that the approach could be generalizable across a wide class of material systems and thus significantly expand the availability of structural descriptors. Similarly, Chen & Ong (165) demonstrated that a variant of MEGNet could perform high-fidelity relaxations of unseen materials with diverse chemistries and that leveraging the resulting structures could improve downstream ML predictions of energy when compared with unrelaxed inputs. The strong performance of these approaches and their potential to significantly increase the scale and effectiveness of computational screening motivates high-value research questions concerning the scale of data sets required for training, the generalizability over material classes, and the applicability to prediction tasks beyond stability.

8.3. Applicability of Compositional Descriptors

Compositional descriptors are typically readily available as tabulated values, but even state-of-the-art models do not perform as well as the best structural approaches. However, there is some evidence that the scale of improvement when including structural information is property dependent. System energies can be conceptualized as a sum of site energies that are highly dependent on the local environment, and graph neural networks provide significantly more robust predictions of materials stability (84). On the other hand, for properties dependent on global features such as phonons (vibrations) or electronic band structures (band gap), the relative improvement may not be as large (99, 166, 167). Identifying common trends connecting tasks for which this difference is the least significant would provide more intuition about which scenarios compositional models are most appropriate for. Furthermore, in some modeling situations, structural information is available but over only a small fraction of the data set. To maximize the value of these data, more general strategies involving transfer learning (141) or combining separate composition and structural models (85) should be developed.

8.4. Extensions of Generative Models

Additional symmetry considerations and the implementation of diffusion-based architectures led to generative models that improved significantly over previous voxel approaches. While this strategy is a promising direction for small unit cells, efforts pertaining to other parameters critical to material performance including microstructure (168), dimensionality (169), and surfaces (170) should also be pursued. In addition, research groups have sidestepped some of the challenges of materials generation by designing approaches that only sample material stoichiometry (171). While these methods limit the full characterization of new materials through a purely computational pipeline, there may be cases where they are sufficient to propose promising regions for experimental analysis.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

AUTHOR CONTRIBUTIONS

J.D. was involved in the writing of all sections. A.R.T. and M.X. collaborated on the writing for Section 2 and designed **Figure 2**. J.K. collaborated on the writing of Section 3 and designed **Figure 3**. J.R.L. collaborated on the writing for Section 5 and designed **Figure 4**. J.P. provided valuable insights for the organization and content of the article. R.G.-B. selected the topic and focus of the review, contributed to the central themes and context, and supervised the project. All authors participated in discussions and the reviewing of the final article.

ACKNOWLEDGMENTS

The authors acknowledge financial support from the Advanced Research Projects Agency-Energy (ARPA-E), US Department of Energy, under award DE-AR0001220, the National Defense Science and Engineering Graduate Fellowship (to J.D.), the National Science Scholarship from Agency for Science, Technology and Research (to M.X.), and the Asahi Glass Company (to A.R.T.). R.G.-B. thanks the Jeffrey Cheah Chair in Engineering. The authors would like to thank Anna Bloom for editorial contributions.

LITERATURE CITED

1. Gómez-Bombarelli R, Wei JN, Duvenaud D, Hernández-Lobato JM, Sánchez-Lengeling B, et al. 2018. Automatic chemical design using a data-driven continuous representation of molecules. *ACS Cent. Sci.* 4(2):268–76
2. Peng J, Schwalbe-Koda D, Akkiraju K, Xie T, Giordano L, et al. 2022. Human- and machine-centred designs of molecules and materials for sustainability and decarbonization. *Nat. Rev. Mater.* 7:991–1009
3. Pyzer-Knapp EO, Suh C, Gómez-Bombarelli R, Aguilera-Iparraguirre J, Aspuru-Guzik A. 2015. What is high-throughput virtual screening? A perspective from organic materials discovery. *Annu. Rev. Mater. Res.* 45:195–216
4. Jain A, Ong SP, Hautier G, Chen W, Richards WD, et al. 2013. Commentary: the Materials Project: a materials genome approach to accelerating materials innovation. *APL Mater.* 1(1):011002
5. Kirklin S, Saal JE, Meredig B, Thompson A, Doak JW, et al. 2015. The Open Quantum Materials Database (OQMD): assessing the accuracy of DFT formation energies. *npj Comput. Mater.* 1(1):15010
6. Chanussot L, Das A, Goyal S, Lavril T, Shuaibi M, et al. 2021. Open Catalyst 2020 (OC20) dataset and community challenges. *ACS Catal.* 11(10):6059–72. Correction. 2021. *ACS Catal.* 11(21):13062–65
7. Curtarolo S, Setyawan W, Wang S, Xue J, Yang K, et al. 2012. Aflowlib.org: a distributed materials properties repository from high-throughput ab initio calculations. *Comput. Mater. Sci.* 58:227–35
8. Ward L, Dunn A, Faghaninia A, Zimmermann NE, Bajaj S, et al. 2018. Matminer: an open source toolkit for materials data mining. *Comput. Mater. Sci.* 152:60–69
9. Tran R, Lan J, Shuaibi M, Goyal S, Wood BM, et al. 2022. The Open Catalyst 2022 (OC22) dataset and challenges for oxide electrocatalysis. arXiv:2206.08917 [cond-mat.mtrl-sci]
10. Huo H, Rupp M. 2022. Unified representation of molecules and crystals for machine learning. *Mach. Learn. Sci. Technol.* 3:045017
11. Faber F, Lindmaa A, von Lilienfeld OA, Armiento R. 2015. Crystal structure representations for machine learning models of formation energies. *Int. J. Quantum Chem.* 115(16):1094–101
12. Bartók AP, Kondor R, Csányi G. 2013. On representing chemical environments. *Phys. Rev. B* 87(18):184115
13. von Lilienfeld OA, Ramakrishnan R, Rupp M, Knoll A. 2015. Fourier series of atomic radial distribution functions: a molecular fingerprint for machine learning models of quantum chemical properties. *Int. J. Quantum Chem.* 115(16):1084–93
14. Musil F, Grisafi A, Bartók AP, Ortner C, Csányi G, Ceriotti M. 2021. Physics-inspired structural representations for molecules and materials. *Chem. Rev.* 121(16):9759–815

15. Glass CW, Oganov AR, Hansen N. 2006. USPEX—evolutionary crystal structure prediction. *Comput. Phys. Commun.* 175(11–12):713–20
16. Wang Y, Lv J, Zhu L, Ma Y. 2010. Crystal structure prediction via particle-swarm optimization. *Phys. Rev. B* 82(9):094116
17. Ward L, Liu R, Krishna A, Hegde VI, Agrawal A, et al. 2017. Including crystal structure attributes in machine learning models of formation energies via Voronoi tessellations. *Phys. Rev. B* 96(2):024104
18. Xie T, Grossman JC. 2018. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Phys. Rev. Lett.* 120(14):145301
19. Ward L, Agrawal A, Choudhary A, Wolverton C. 2016. A general-purpose machine learning framework for predicting properties of inorganic materials. *npj Comput. Mater.* 2(1):16028
20. Haynes WM. 2016. *CRC Handbook of Chemistry and Physics*. Boca Raton, FL: CRC Press
21. Frey NC, Akinwande D, Jariwala D, Shenoy VB. 2020. Machine learning-enabled design of point defects in 2D materials for quantum and neuromorphic information processing. *ACS Nano* 14(10):13406–17
22. Hoffmann J, Maestrati L, Sawada Y, Tang J, Sellier JM, Bengio Y. 2019. Data-driven approach to encoding and decoding 3-D crystal structures. arXiv:1909.00949 [cs.LG]
23. Noh J, Kim J, Stein HS, Sanchez-Lengeling B, Gregoire JM, et al. 2019. Inverse design of solid-state materials via a continuous representation. *Matter* 1(5):1370–84
24. Court CJ, Yildirim B, Jain A, Cole JM. 2020. 3-D inorganic crystal structure generation and property prediction via representation learning. *J. Chem. Inform. Model.* 60(10):4518–35
25. Choubisa H, Askerka M, Ryczko K, Voznyy O, Mills K, et al. 2020. Crystal site feature embedding enables exploration of large chemical spaces. *Matter* 3(2):433–48
26. Behler J, Parrinello M. 2007. Generalized neural-network representation of high-dimensional potential-energy surfaces. *Phys. Rev. Lett.* 98(14):146401
27. Behler J. 2011. Atom-centered symmetry functions for constructing high-dimensional neural network potentials. *J. Chem. Phys.* 134(7):074106
28. Smith JS, Isayev O, Roitberg AE. 2017. ANI-1: an extensible neural network potential with DFT accuracy at force field computational cost. *Chem. Sci.* 8(4):3192–203
29. De S, Bartók AP, Csányi G, Ceriotti M. 2016. Comparing molecules and solids across structural and alchemical space. *Phys. Chem. Chem. Phys.* 18(20):13754–69
30. Schwalbe-Koda D, Jensen Z, Olivetti E, Gómez-Bombarelli R. 2019. Graph similarity drives zeolite diffusionless transformations and intergrowth. *Nat. Mater.* 18(11):1177–81
31. Dragoni D, Daff TD, Csányi G, Marzari N. 2018. Achieving DFT accuracy with a machine-learning interatomic potential: thermomechanics and defects in bcc ferromagnetic iron. *Phys. Rev. Mater.* 2(1):013808
32. Willatt MJ, Musil F, Ceriotti M. 2019. Atom-density representations for machine learning. *J. Chem. Phys.* 150(15):154110
33. Schütt KT, Glawe H, Brockherde F, Sanna A, Müller KR, Gross EK. 2014. How to represent crystal structures for machine learning: towards fast prediction of electronic properties. *Phys. Rev. B* 89(20):205118
34. Moosavi SM, Novotny B, Ongari D, Moubarak E, Asgari M, et al. 2022. A data-science approach to predict the heat capacity of nanoporous materials. *Nat. Mater.* 21(12):1419–25
35. Isayev O, Oses C, Toher C, Gossett E, Curtarolo S, Tropsha A. 2017. Universal fragment descriptors for predicting properties of inorganic crystals. *Nat. Commun.* 8(1):15679
36. Bartók AP, De S, Poelking C, Bernstein N, Kermode JR, et al. 2017. Machine learning unifies the modeling of materials and molecules. *Sci. Adv.* 3(12):e1701816
37. Lazar EA, Lu J, Rycroft CH. 2022. Voronoi cell analysis: the shapes of particle systems. *Am. J. Phys.* 90(6):469
38. Krishnapriyan AS, Montoya J, Haranczyk M, Hummelshøj J, Morozov D. 2021. Machine learning with persistent homology and chemical word embeddings improves prediction accuracy and interpretability in metal-organic frameworks. *Sci. Rep.* 11(1):8888
39. Rupp M, Tkatchenko A, Müller KR, von Lilienfeld OA. 2012. Fast and accurate modeling of molecular atomization energies with machine learning. *Phys. Rev. Lett.* 108(5):058301

40. Saucedo HE, Gálvez-González LE, Chmiela S, Paz-Borbón LO, Müller KR, Tkatchenko A. 2022. BIGDML—towards accurate quantum machine learning force fields for materials. *Nat. Commun.* 13(1):3733
41. Sanchez JM, Ducastelle F, Gratias D. 1984. Generalized cluster description of multicomponent systems. *Phys. A Stat. Mech. Appl.* 128(1–2):334–50
42. Chang JH, Kleiven D, Melander M, Akola J, Garcia-Lastra JM, Vegge T. 2019. CLEASE: a versatile and user-friendly implementation of cluster expansion method. *J. Phys. Condens. Matter* 31(32):325901
43. Hart GL, Mueller T, Toher C, Curtarolo S. 2021. Machine learning for alloys. *Nat. Rev. Mater.* 6(8):730–55
44. Nyshadham C, Rupp M, Bekker B, Shapeev AV, Mueller T, et al. 2019. Machine-learned multi-system surrogate models for materials prediction. *npj Comput. Mater.* 5(1):51
45. Yang JH, Chen T, Barroso-Luque L, Jadidi Z, Ceder G. 2022. Approaches for handling high-dimensional cluster expansions of ionic systems. *npj Comput. Mater.* 8(1):133
46. Drautz R. 2019. Atomic cluster expansion for accurate and transferable interatomic potentials. *Phys. Rev. B* 99(1):014104
47. Batatia I, Kovács DP, Simm GNC, Ortner C, Csányi G. 2022. Mace: higher order equivariant message passing neural networks for fast and accurate force fields. arXiv:2206.07697 [stat.ML]
48. Carlsson G. 2020. Topological methods for data modelling. *Nat. Rev. Phys.* 2(12):697–708
49. Pun CS, Lee SX, Xia K. 2022. Persistent-homology-based machine learning: a survey and a comparative study. *Artif. Intell. Rev.* 55(7):5169–213
50. Jiang Y, Chen D, Chen X, Li T, Wei GW, Pan F. 2021. Topological representations of crystalline compounds for the machine-learning prediction of materials properties. *npj Comput. Mater.* 7(1):28
51. Lee Y, Barthel SD, Dłotko P, Moosavi SM, Hess K, Smit B. 2018. High-throughput screening approach for nanoporous materials genome using topological data analysis: application to zeolites. *J. Chem. Theory Comput.* 14(8):4427–37
52. Buchet M, Hiraoka Y, Obayashi I. 2018. Persistent homology and materials informatics. In *Nanoinformatics*, ed. I Tanaka, pp. 75–95. Singapore: Springer Nat.
53. Duvenaud D, Maclaurin D, Aguilera-Iparraguirre J, Gómez-Bombarelli R, Hirzel T, et al. 2015. Convolutional networks on graphs for learning molecular fingerprints. arXiv:1509.09292 [cs.LG]
54. Reiser P, Neubert M, Eberhard A, Torresi L, Zhou C, et al. 2022. Graph neural networks for materials science and chemistry. *Commun. Mater.* 3(1):93
55. Schütt KT, Saucedo HE, Kindermans PJ, Tkatchenko A, Müller KR. 2018. SchNet – a deep learning architecture for molecules and materials. *J. Chem. Phys.* 148(24):241722
56. Chen C, Ye W, Zuo Y, Zheng C, Ong SP. 2019. Graph networks as a universal machine learning framework for molecules and crystals. *Chem. Mater.* 31(9):3564–72
57. Fung V, Zhang J, Juarez E, Sumpter BG. 2021. Benchmarking graph neural networks for materials chemistry. *npj Comput. Mater.* 7(1):84
58. Chen C, Zuo Y, Ye W, Li X, Ong SP. 2021. Learning properties of ordered and disordered materials from multi-fidelity data. *Nat. Comput. Sci.* 1(1):46–53
59. Choudhary K, DeCost B. 2021. Atomistic line graph neural network for improved materials property predictions. *npj Comput. Mater.* 7(1):185
60. Park CW, Wolverton C. 2020. Developing an improved crystal graph convolutional neural network framework for accelerated materials discovery. *Phys. Rev. Mater.* 4(6):063801
61. Karamad M, Magar R, Shi Y, Siahrostami S, Gates ID, Farimani AB. 2020. Orbital graph convolutional neural network for material property prediction. *Phys. Rev. Mater.* 4(9):093801
62. Karagunesian J, Lunger JR, Shao-Horn Y, Gomez-Bombarelli R. 2021. *Crystal graph convolutional neural networks for per-site property prediction*. Paper presented at the 35th Conference on Neural Information Processing Systems, online, Dec. 13
63. Gong S, Xie T, Shao-Horn Y, Gomez-Bombarelli R, Grossman JC. 2022. Examining graph neural networks for crystal structures: limitations and opportunities for capturing periodicity. arXiv:2208.05039 [cond-mat.mtrl-sci]
64. Schütt KT, Unke OT, Gastegger M. 2021. Equivariant message passing for the prediction of tensorial properties and molecular spectra. arXiv:2102.03150 [cs.LG]

65. Cheng J, Zhang C, Dong L. 2021. A geometric-information-enhanced crystal graph network for predicting properties of materials. *Commun. Mater.* 2(1):92
66. Louis SY, Zhao Y, Nasiri A, Wang X, Song Y, et al. 2020. Graph convolutional neural networks with global attention for improved materials property prediction. *Phys. Chem. Chem. Phys.* 22(32):18141–48
67. Yu H, Hong L, Chen S, Gong X, Xiang H. 2023. Capturing long-range interaction with reciprocal space neural network. arXiv:2211.16684 [cond-mat.mtrl-sci]
68. Kong S, Ricci F, Guevarra D, Neaton JB, Gomes CP, Gregoire JM. 2022. Density of states prediction for materials discovery via contrastive learning from probabilistic embeddings. *Nat. Commun.* 13(1):949
69. Batzner S, Musaelian A, Sun L, Geiger M, Mailoa JP, et al. 2021. E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. *Nat. Commun.* 13(1):2453
70. Geiger M, Smidt T. 2022. e3nn: Euclidean neural networks. arXiv:2207.09453 [cs.LG]
71. Thomas N, Smidt T, Kearnes S, Yang L, Li L, et al. 2018. Tensor field networks: rotation- and translation-equivariant neural networks for 3d point clouds. arXiv:1802.08219 [cs.LG]
72. Gastegger J, Becker F, Günnemann S. 2021. GemNet: universal directional graph neural networks for molecules. *Adv. Neural Inform. Proc. Syst.* 34:6790–802
73. Gastegger J, Shuaibi M, Sriram A, Günnemann S, Ulissi Z, et al. 2022. GemNet-OC: developing graph neural networks for large and diverse molecular simulation datasets. arXiv:2204.02782 [cs.LG]
74. Chen Z, Andrejevic N, Smidt T, Ding Z, Xu Q, et al. 2021. Direct prediction of phonon density of states with Euclidean neural networks. *Adv. Sci.* 8(12):2004214
75. Kolluru A, Shuaibi M, Palizhati A, Shoghi N, Das A, et al. 2022. Open challenges in developing generalizable large-scale machine-learning models for catalyst discovery. *ACS Catal.* 12(14):8572–81
76. Gibson J, Hire A, Hennig RG. 2022. Data-augmentation for graph neural network learning of the relaxed energies of unrelaxed structures. *npj Comput. Mater.* 8(1):211
77. Schmidt J, Pettersson L, Verdozzi C, Botti S, Marques MA. 2021. Crystal graph attention networks for the prediction of stable materials. *Sci. Adv.* 7(49):7948
78. Zuo Y, Qin M, Chen C, Ye W, Li X, et al. 2021. Accelerating materials discovery with Bayesian optimization and graph deep learning. *Mater. Today* 51:126–35
79. Dunn A, Wang Q, Ganose A, Dopp D, Jain A. 2020. Benchmarking materials property prediction methods: the Matbench test set and Automatminer reference algorithm. *npj Comput. Mater.* 6(1):138
80. Callister WD, Rethwisch DG. 2018. Materials science and engineering: an introduction. Hoboken, NJ: Wiley. 10th ed.
81. Pauling L. 1929. The principles determining the structure of complex ionic crystals. *J. Am. Chem. Soc.* 51(4):1010–26
82. George J, Waroquiers D, Stefano DD, Petretto G, Rignanese GM, Hautier G. 2020. The limited predictive power of the Pauling rules. *Angew. Chem. Int. Ed.* 132(19):7639–45
83. Meredig B, Agrawal A, Kirklin S, Saal JE, Doak JW, et al. 2014. Combinatorial screening for new materials in unconstrained composition space with machine learning. *Phys. Rev. B* 89(9):094104
84. Bartel CJ, Trewartha A, Wang Q, Dunn A, Jain A, Ceder G. 2020. A critical examination of compound stability predictions from machine-learned formation energies. *npj Comput. Mater.* 6(1):97
85. Stanev V, Oses C, Kusne AG, Rodriguez E, Paglione J, et al. 2018. Machine learning modeling of superconducting critical temperature. *npj Comput. Mater.* 4(1):29
86. Faber FA, Lindmaa A, Lilienfeld OAV, Armiento R. 2016. Machine learning energies of 2 million elpasolite (ABC_2D_6) crystals. *Phys. Rev. Lett.* 117(13):135502
87. Oliynyk AO, Antono E, Sparks TD, Ghadbeigi L, Gaultois MW, et al. 2016. High-throughput machine-learning-driven synthesis of full-heusler compounds. *Chem. Mater.* 28(20):7324–31
88. Ouyang R, Curtarolo S, Ahmetcik E, Scheffler M, Ghiringhelli LM. 2018. SISSO: a compressed-sensing method for identifying the best low-dimensional descriptor in an immensity of offered candidates. *Phys. Rev. Mater.* 2(8):083802
89. Ghiringhelli LM, Vybiral J, Ahmetcik E, Ouyang R, Levchenko SV, et al. 2017. Learning physical descriptors for materials science by compressed sensing. *N. J. Phys.* 19(2):023017
90. Ghiringhelli LM, Vybiral J, Levchenko SV, Draxl C, Scheffler M. 2015. Big data of materials science: critical role of the descriptor. *Phys. Rev. Lett.* 114(10):105503

91. Bartel CJ, Sutton C, Goldsmith BR, Ouyang R, Musgrave CB, et al. 2019. New tolerance factor to predict the stability of perovskite oxides and halides. *Sci. Adv.* 5(2):eaav0693
92. Zhou Q, Tang P, Liu S, Pan J, Yan Q, Zhang SC. 2018. Learning atoms for materials discovery. *PNAS* 115(28):E6411–17
93. Tshitoyan V, Dagdelen J, Weston L, Dunn A, Rong Z, et al. 2019. Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature* 571(7763):95–98
94. Jha D, Ward L, Paul A, Liao W-k, Choudhary A, et al. 2018. *ElemNet*: deep learning the chemistry of materials from only elemental composition. *Sci. Rep.* 8(1):17593
95. Goodall RE, Lee AA. 2020. Predicting materials properties without crystal structure: deep representation learning from stoichiometry. *Nat. Commun.* 11(1):6280
96. Wang AYT, Kauwe SK, Murdock RJ, Sparks TD. 2021. Compositionally restricted attention-based network for materials property predictions. *npj Comput. Mater.* 7(1):77
97. Zhang Z, Tehrani AM, Oliynyk AO, Day B, Brgoch J. 2021. Finding the next superhard material through ensemble learning. *Adv. Mater.* 33(5):2005112
98. Oliynyk AO, Adutwum LA, Rudyk BW, Pisavadia H, Lotfi S, et al. 2017. Disentangling structural confusion through machine learning: structure prediction and polymorphism of equiatomic ternary phases *ABC*. *J. Am. Chem. Soc.* 139(49):17870–81
99. Tian SIP, Walsh A, Ren Z, Li Q, Buonassisi T. 2022. What information is necessary and sufficient to predict materials properties using machine learning? arXiv:2206.04968 [cond-mat.mtrl-sci]
100. Artrith N. 2019. Machine learning for the modeling of interfaces in energy storage and conversion materials. *J. Phys. Energy* 1(3):032002
101. Rao RR, Kolb MJ, Giordano L, Pedersen AF, Katayama Y, et al. 2020. Operando identification of site-dependent water oxidation activity on ruthenium dioxide single-crystal surfaces. *Nat. Catal.* 3:516–25
102. Varley JB, Samanta A, Lordi V. 2017. Descriptor-based approach for the prediction of cation vacancy formation energies and transition levels. *J. Phys. Chem. Lett.* 8(20):5059–63
103. Zhang X, Wang H, Hickel T, Rogal J, Li Y, Neugebauer J. 2020. Mechanism of collective interstitial ordering in Fe–C alloys. *Nat. Mater.* 19(8):849–54
104. Deringer VL, Bartók AP, Bernstein N, Wilkins DM, Ceriotti M, Csányi G. 2021. Gaussian process regression for materials and molecules. *Chem. Rev.* 121(16):10073–141
105. Wan Z, Wang QD, Liu D, Liang J. 2021. Data-driven machine learning model for the prediction of oxygen vacancy formation energy of metal oxide materials. *Phys. Chem. Chem. Phys.* 23(29):15675–84
106. Witman M, Goyal A, Ogitsu T, McDaniel A, Lany S. 2022. Graph neural network modeling of vacancy formation enthalpy for materials discovery and its application in solar thermochemical water splitting. ChemRxiv. <https://doi.org/10.26434/chemrxiv-2022-frcns>
107. Medasani B, Gamst A, Ding H, Chen W, Persson KA, et al. 2016. Predicting defect behavior in b2 intermetallics by merging ab initio modeling and machine learning. *npj Comput. Mater.* 2(1):1
108. Ziletti A, Kumar D, Scheffler M, Ghiringhelli LM. 2018. Insightful classification of crystal structures using deep learning. *Nat. Commun.* 9(1):2775
109. Medford AJ, Vojvodic A, Hummelshøj JS, Voss J, Abild-Pedersen F, et al. 2015. From the Sabatier principle to a predictive theory of transition-metal heterogeneous catalysis. *J. Catal.* 328:36–42
110. Zhao ZJ, Liu S, Zha S, Cheng D, Studt F, et al. 2019. Theory-guided design of catalytic materials using scaling relationships and reactivity descriptors. *Nat. Rev. Mater.* 4(12):792–804
111. Hwang J, Rao RR, Giordano L, Katayama Y, Yu Y, Shao-Horn Y. 2017. Perovskites in catalysis and electrocatalysis. *Science* 358(6364):751–56
112. Calle-Vallejo F, Tymoczko J, Colic V, Vu QH, Pohl MD, et al. 2015. Finding optimal surface sites on heterogeneous catalysts by counting nearest neighbors. *Science* 350(6257):185–89
113. Fung V, Hu G, Ganesh P, Sumpter BG. 2021. Machine learned features from density of states for accurate adsorption energy prediction. *Nat. Commun.* 12(1):88
114. Back S, Yoon J, Tian N, Zhong W, Tran K, Ulissi ZW. 2019. Convolutional neural network of atomic surface structures to predict binding energies for high-throughput screening of catalysts. *J. Phys. Chem. Lett.* 10(15):4401–8
115. Tran K, Ulissi ZW. 2018. Active learning across intermetallics to guide discovery of electrocatalysts for CO₂ reduction and H₂ evolution. *Nat. Catal.* 1(9):696–703

116. Ghanekar PG, Deshpande S, Greeley J. 2022. Adsorbate chemical environment-based machine learning framework for heterogeneous catalysis. *Nat. Commun.* 13(1):5788
117. Kiyohara S, Oda H, Miyata T, Mizoguchi T. 2016. Prediction of interface structures and energies via virtual screening. *Sci. Adv.* 2(11):e1600746
118. Hu C, Zuo Y, Chen C, Ong SP, Luo J. 2020. Genetic algorithm-guided deep learning of grain boundary diagrams: addressing the challenge of five degrees of freedom. *Mater. Today* 38:49–57
119. Dai M, Demirel MF, Liang Y, Hu JM. 2021. Graph neural networks for an accurate and interpretable prediction of the properties of polycrystalline materials. *npj Comput. Mater.* 7:103
120. Huber L, Hadian R, Grabowski B, Neugebauer J. 2018. A machine learning approach to model solute grain boundary segregation. *npj Comput. Mater.* 4(1):64
121. Ye W, Zheng H, Chen C, Ong SP. 2022. A universal machine learning model for elemental grain boundary energies. *Scr. Mater.* 218:114803
122. Lazar EA. 2017. Vorotop: Voronoi cell topology visualization and analysis toolkit. *Model. Simul. Mater. Sci. Eng.* 26(1):015011
123. Priedeman JL, Rosenbrock CW, Johnson OK, Homer ER. 2018. Quantifying and connecting atomic and crystallographic grain boundary structure using local environment representation and dimensionality reduction techniques. *Acta Mater.* 161:431–43
124. Rosenbrock CW, Homer ER, Csányi G, Hart GL. 2017. Discovering the building blocks of atomic systems using machine learning: application to grain boundaries. *npj Comput. Mater.* 3(1):29
125. Fujii S, Yokoi T, Fisher CA, Moriwake H, Yoshiya M. 2020. Quantitative prediction of grain boundary thermal conductivities from local atomic environments. *Nat. Commun.* 11(1):1854
126. Sharp TA, Thomas SL, Cubuk ED, Schoenholz SS, Srolovitz DJ, Liu AJ. 2018. Machine learning determination of atomic dynamics at grain boundaries. *PNAS* 115(43):10943–47
127. Batatia I, Batzner S, Kovács DP, Musaelian A, Simm GNC, et al. 2022. The design space of E(3)-equivariant atom-centered interatomic potentials. arXiv:2205.06643 [stat.ML]
128. Axelrod S, Shakhnovich E, Gómez-Bombarelli R. 2022. Excited state non-adiabatic dynamics of large photoswitchable molecules using a chemically transferable machine learning potential. *Nat. Commun.* 13(1):3440
129. Li X, Li B, Yang Z, Chen Z, Gao W, Jiang Q. 2022. A transferable machine-learning scheme from pure metals to alloys for predicting adsorption energies. *J. Mater. Chem. A* 10(2):872–80
130. Harper DR, Nandy A, Arunachalam N, Duan C, Janet JP, Kulik HJ. 2022. Representations and strategies for transferable machine learning improve model performance in chemical discovery. *J. Chem. Phys.* 156(7):074101
131. Unke OT, Chmiela S, Gastegger M, Schütt KT, Sauceda HE, Müller KR. 2021. SpookyNet: learning force fields with electronic degrees of freedom and nonlocal effects. *Nat. Commun.* 12(1):7273
132. Husch T, Sun J, Cheng L, Lee SJ, Miller TF. 2021. Improved accuracy and transferability of molecular-orbital-based machine learning: organics, transition-metal complexes, non-covalent interactions, and transition states. *J. Chem. Phys.* 154(6):064108
133. Zheng X, Zheng P, Zhang RZ. 2018. Machine learning material properties from the periodic table using convolutional neural networks. *Chem. Sci.* 9(44):8426–32
134. Feng S, Fu H, Zhou H, Wu Y, Lu Z, Dong H. 2021. A general and transferable deep learning framework for predicting phase formation in materials. *npj Comput. Mater.* 7(1):10
135. Jha D, Choudhary K, Tavazza F, Liao W-k, Choudhary A, et al. 2019. Enhancing materials property prediction by leveraging computational and experimental data using deep transfer learning. *Nat. Commun.* 10(1):10
136. Yamada H, Liu C, Wu S, Koyama Y, Ju S, et al. 2019. Predicting materials properties with little data using shotgun transfer learning. *ACS Central Sci.* 5(10):1717–30
137. Gupta V, Choudhary K, Tavazza F, Campbell C, Liao W-k, et al. 2021. Cross-property deep transfer learning framework for enhanced predictive analytics on small materials data. *Nat. Commun.* 12(1):6595
138. Lee J, Asahi R. 2021. Transfer learning for materials informatics using crystal graph convolutional neural network. *Comput. Mater. Sci.* 190:110314
139. Kolluru A, Shoghi N, Shuaibi M, Goyal S, Das A, et al. 2022. Transfer learning using attentions across atomic systems with graph neural networks (TAAG). *J. Chem. Phys.* 156(18):184702

140. Chen C, Ong SP. 2021. AtomSets as a hierarchical transfer learning framework for small and large materials datasets. *npj Comput. Mater.* 7(1):173
141. Cubuk ED, Sendek AD, Reed EJ. 2019. Screening billions of candidates for solid lithium-ion conductors: a transfer learning approach for small data. *J. Chem. Phys.* 150(21):214701
142. Greenman KP, Green WH, Gómez-Bombarelli R. 2022. Multi-fidelity prediction of molecular optical peaks with deep learning. *Chem. Sci.* 13(4):1152–62
143. Kong S, Guevarra D, Gomes CP, Gregoir JM. 2021. Materials representation and transfer learning for multi-property prediction. *Appl. Phys. Rev.* 8(2):021409
144. Kingma DP, Welling M. 2013. Auto-encoding variational Bayes. arXiv:1312.6114 [stat.ML]
145. Goodfellow IJ, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, et al. 2014. Generative adversarial networks. *Adv. Neural Inf. Proc. Syst.* 27:2672–80
146. Nouiara A, Sokolovska N, Crivello JC. 2018. CrystalGAN: learning to discover crystallographic structures with generative adversarial networks. In *Proceedings of the AAAI 2019 Spring Symposium on Combining Machine Learning with Knowledge Engineering*, ed. A Martin, K Hinkelmann, A Gerber, D Lenat, F van Harmelen, P Clark. <https://ceur-ws.org/Vol-2350/paper18.pdf>
147. Shi C, Luo S, Xu M, Tang J. 2021. Learning gradient fields for molecular conformation generation. *Proc. Mach. Learn. Res.* 139:9558–68
148. Ho J, Jain A, Abbeel P. 2020. Denoising diffusion probabilistic models. *Adv. Neural Inf. Proc. Syst.* 33:6840–51
149. Schwalbe-Koda D, Gómez-Bombarelli R. 2020. Generative models for automatic chemical design. *Lect. Notes Phys.* 968:445–67
150. Noh J, Gu GH, Kim S, Jung Y. 2020. Machine-enabled inverse design of inorganic solid materials: promises and challenges. *Chem. Sci.* 11(19):4871–81
151. Fuhr AS, Sumpter BG. 2022. Deep generative models for materials discovery and machine learning-accelerated innovation. *Front. Mater.* 9:182
152. Xie T, Fu X, Ganea OE, Barzilay R, Jaakkola T. 2021. Crystal diffusion variational autoencoder for periodic material generation. arXiv:2110.06197 [cs.LG]
153. Yao Z, Sánchez-Lengeling B, Bobbitt NS, Bucior BJ, Kumar SGH, et al. 2021. Inverse design of nanoporous crystalline reticular materials with deep generative models. *Nat. Mach. Intell.* 3(1):76–86
154. Ren Z, Tian SIP, Noh J, Oviedo F, Xing G, et al. 2022. An invertible crystallographic representation for general inverse design of inorganic crystals with targeted properties. *Matter* 5(1):314–35
155. Long T, Fortunato NM, Opahle I, Zhang Y, Samathrakris I, et al. 2021. Constrained crystals deep convolutional generative adversarial network for the inverse design of crystal structures. *npj Comput. Mater.* 7(1):66
156. Kim S, Noh J, Gu GH, Aspuru-Guzik A, Jung Y. 2020. Generative adversarial networks for crystal structure prediction. *ACS Central Sci.* 6(8):1412–20
157. Fung V, Jia S, Zhang J, Bi S, Yin J, Ganesh P. 2022. Atomic structure generation from reconstructing structural fingerprints. *Mach. Learn. Sci. Technol.* 3:045018
158. Kim B, Lee S, Kim J. 2020. Inverse design of porous materials using artificial neural networks. *Sci. Adv.* 6(1):eaax9324
159. Musaelian A, Batzner S, Johansson A, Sun L, Owen CJ, et al. 2023. *Nat. Commun.* 14(1):579
160. Unke OT, Chmiela S, Sauceda HE, Gastegger M, Poltavsky I, et al. 2021. Machine learning force fields. *Chem. Rev.* 121(16):10142–86
161. Grisafi A, Nigam J, Ceriotti M. 2021. Multi-scale approach for the prediction of atomic scale properties. *Chem. Sci.* 12:2078
162. Schaarschmidt M, Riviere M, Ganose AM, Spencer JS, Gaunt AL, et al. 2022. Learned force fields are ready for ground state catalyst discovery. arXiv:2209.12466 [cond-mat.mtrl-sci]
163. Lan J, Palizhati A, Shuaibi M, Wood BM, Wander B, et al. 2023. AdsorbML: accelerating adsorption energy calculations with machine learning. arXiv:2211.16486 [cond-mat.mtrl-sci]
164. Kresse G, Hafner J. 1993. Ab initio molecular dynamics for liquid metals. *Phys. Rev. B* 47(1):558–61
165. Chen C, Ong SP. 2022. A universal graph deep learning interatomic potential for the periodic table. *Nat. Comput. Sci.* 2(11):718–28

166. Legrain F, Carrete J, Roekeghem AV, Curtarolo S, Mingo N. 2017. How chemical composition alone can predict vibrational free energies and entropies of solids. *Chem. Mater.* 29(15):6220–27
167. Zhuo Y, Tehrani AM, Brgoch J. 2018. Predicting the band gaps of inorganic solids by machine learning. *J. Phys. Chem. Lett.* 9(7):1668–73
168. Hsu T, Epting WK, Kim H, Abernathy HW, Hackett GA, et al. 2020. Microstructure generation via generative adversarial network for heterogeneous, topologically complex 3D materials. *JOM* 73(1):90–102
169. Song Y, Siriwardane EM, Zhao Y, Hu J. 2020. Computational discovery of new 2D materials using deep learning generative models. *ACS Appl. Mater. Interfaces* 13(45):53303–13
170. Liu Z, Zhu D, Rodrigues SP, Lee KT, Cai W. 2018. Generative model for the inverse design of metasurfaces. *Nano Lett.* 18(10):6570–76
171. Dan Y, Zhao Y, Li X, Li S, Hu M, Hu J. 2020. Generative adversarial networks (GAN) based efficient sampling of chemical composition space for inverse design of inorganic materials. *npj Comput. Mater.* 6(1):84