ANNUAL REVIEWS

Annual Review of Materials Research

Evolving the Materials Genome: How Machine Learning Is Fueling the Next Generation of Materials Discovery

Changwon Suh,[1] Clyde Fare,[2] James A. Warren,[3] and Edward O. Pyzer-Knapp[2]

[1]Nexight Group, Silver Spring, Maryland 20910, USA

[2]IBM Research, Daresbury WA4 4AD, United Kingdom; email: epyzerk3@uk.ibm.com

[3]Material Measurement Laboratory, National Institute of Standards and Technology, Gaithersburg, Maryland 20899, USA

Annu. Rev. Mater. Res. 2020. 50:1–25

First published as a Review in Advance on April 21, 2020

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

https://doi.org/10.1146/annurev-matsci-082019-105100

Keywords

materials genome, materials discovery, artificial intelligence, machine learning, data

Abstract

Machine learning, applied to chemical and materials data, is transforming the field of materials discovery and design, yet significant work is still required to fully take advantage of machine learning algorithms, tools, and methods. Here, we review the accomplishments to date of the community and assess the maturity of state-of-the-art, data-intensive research activities that combine perspectives from materials science and chemistry. We focus on three major themes—learning to see, learning to estimate, and learning to search materials—to show how advanced computational learning technologies are rapidly and successfully used to solve materials and chemistry problems. Additionally, we discuss a clear path toward a future where data-driven approaches to materials discovery and design are standard practice.

1. INTRODUCTION: MACHINE LEARNING AND THE MATERIALS DISCOVERY WORKFLOW

The discovery—or, preferably, the design—of new materials lies at the heart of technological progress. Finding faster, more reliable methods for obtaining materials with the properties and performance needed to realize manufactured products ranging from airplane wings, to cardiac stents, to the next generation of computer chips remains one of the great challenges of modern science and engineering. The Materials Genome Initiative (MGI) (1), a US government effort to accelerate materials discovery, design, development, and deployment into manufactured products, and numerous analogous efforts around the world are all seeking to find these faster, more reliable methods and to lower the barriers to accessing such methods for manufacturers.

Of course, until very recently, the development of new materials was essentially relegated to the domain of artisans. Many materials innovations were the product of serendipity. Yet beyond these happy accidents that provided a leap forward, a more systematic approach to developing new materials relied on an iterative, labor-intensive process of trial and error, where most of the knowledge informing the development came from information handed down from master to apprentice. Such processes are reliable, expensive, and slow. Improvements often occur over decades (or generations).

With the advent of modern materials science and engineering, the situation was substantially improved; the knowledge was often encoded in predictive physical models. Yet even with these advances, the complexity of materials systems has made broad use of these models, and of simulations based on these models, challenging. In the past 20 years, however, the situation has significantly improved. This can be attributed to several factors, including the advent of ever faster computational power, yielding predictive power with direct engineering relevance, as well as substantial efforts to promote this paradigm shift, such as integrated computational materials engineering (ICME) (2) and the MGI.

The rise of the ICME paradigm reflects the availability of models that span the length scales and timescales relevant to understanding the processing-structure-properties (PSP) linkages that underpin materials design, as well as the intellectual and technical capacity to integrate these models into a predictive engine. With such tools one can, in principle, attack the inverse design problem, wherein the properties of the system are set by a designer and data and software tools can then be used to determine the processing steps needed to create the material. ICME has become a dominant approach in metallurgical materials design and has been making inroads in the design of both ceramics and polymers. Indeed, the success of the ICME approach provided a substantial impetus for the founding of the MGI. However, to date, one of the primary challenges with ICME is the cost, in both human expertise and computational power, required for the significant return on investment that can be realized from such techniques. It is in this context that the recent understanding of the potential of artificial intelligence (AI) to influence materials design has generated substantial excitement.

In the traditional view of ICME, the computational models that undergird the approach are, at their core, physics-based relationships that model specific phenomena. However, AI approaches such as machine learning are, in essence, nothing more than computational algorithms to construct a model relating inputs to outputs. Thus, a machine learning approach can replace a traditional physical model—albeit with specific risks, as its domain of applicability is often poorly characterized and thus its failure is impossible to predict in advance. Nonetheless, such machine-learning-based approaches come with several advantages: First, the algorithms can often provide a model where none exists, and second, AI algorithms are often extraordinarily faster than traditional models. For these reasons, AI—and specifically machine learning—is emerging as a potent tool for materials design (3) (**Figure 1**).

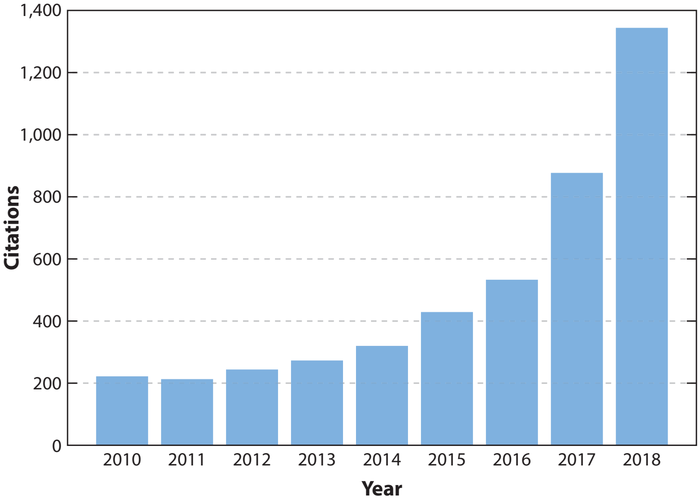

Figure 1

Number of research articles written annually on machine learning in materials discovery from 2010 to 2018. There has been a near-exponential growth in such publications since the start of the decade. Data were retrieved from Web of Science in August 2019.

The use of the word genome in the MGI is by analogy with the Human Genome Initiative, evoking the notion that knowledge of basic building blocks allows for scientific discovery. Indeed, there have been quite a few genome efforts within materials science—for example, the ceramic genome (4), glass genome (5), nanomaterials genome (6), nanoporous materials genome (7), polymer genome (8), high-pressure materials genome (9), and structure genome (10). MGI and machine learning for materials are intimately connected: One can discover and make materials more quickly and inexpensively by determining quantitative structure-property relationships (QSPRs)—materials genes, to abuse the analogy—through the use of machine learning.

Additionally, when combined with robotic laboratory systems, machine learning systems can enable autonomous research. In this case, the desired properties of the materials system are inputs, and the machine-learning-controlled setup can autonomously search processing and composition space in a quest to discover the next material with the desired properties. These types of systems promise to completely disrupt existing approaches to materials design and to more fully realize the goals of truly intelligent, and automated, materials design (11).

Much of the potential upside to the application of machine learning to materials design seems self-evident, but of course, challenges remain. There are significant impediments to implementation, beyond the need to address concerns that naturally arise when a new, unfamiliar technique becomes prominent. Perhaps the most pressing need is the requirement for a robust data infrastructure. The MGI is striving to make the maximum amount of research data widely available, with sufficient metadata to make the data maximally useful (12). This requires many additional considerations regarding metadata and the standards required for interoperability of disparate materials data providers. While significant progress has been made, there is still much to be done.

According to Tom M. Mitchell (13), a pioneer of machine learning, learning is an approach to improving problem-solving through prior knowledge. He defined machine learning as follows: “A computer program is said to learn from experience *E* with respect to some class of tasks *T* and performance measure *P*, if its performance at tasks in *T*, as measured by *P*, improves with experience *E*” (13, p. 2). Aligning with Mitchell’s treatment, we focus on three major themes:

1. learning to see (how machine learning can represent materials data for use in the production of AI models);
2. learning to estimate (how, once materials data have been correctly represented, they can be used to build powerful surrogate models that are cheap to execute yet accurate); and
3. learning to search (how, once we are able to estimate the properties of materials on the basis of data, we can exploit this ability to build powerful and creative search strategies).

2. LEARNING TO SEE

2.1. Representations

Applications of machine learning to molecular species require as a first step some means to represent the molecules of interest in a manner appropriate for algorithmic ingestion or production. The question of how best to construct such representations, which are typically either tensors of numbers or strings of characters, has been a focus of research within the chemometrics community (14), where there is a long history of investigation into how best to efficiently search chemical databases for molecules matching specific substructures and how to construct QSPRs. In the latter case, it has been recognized that the choice of representation has a greater impact than the choice of which learning algorithm to apply (15, 16).

Representations typically subdivide into either carefully collated collections of handcrafted one-dimensional descriptors, each of which is a single summary statistic (e.g., atom counts, bond counts, molecular weight, experimental or computational properties), or more systematic representations of molecular structure. The latter are a focus here.

Within other fields, notably image and natural language processing, the deep learning revolution has led to a move away from investigating different representations in favor of generating simple, complete, basic representations that are then transformed during the training process into a representation learned by the network that expresses features useful for the inference problem at hand (see the sidebar titled Requirements of a Materials Descriptor). Thus, choice of architecture and choice of data set subsume choice of representation. Within materials and chemical research, deep learning has shown considerable promise. Unlike for image processing, where the objects of interest are intrinsically digital and an obvious simple, complete, basic representation is readily available, in the case of molecules, even when deep networks are being used, questions remain: how to construct the initial representation, whether to use the molecular graph or full three-dimensional coordinates, and how to represent the atoms and/or bonds themselves. Further nonneural methods making use of carefully constructed three-dimensional representations still offer performance competitive with that of deep neural networks.

REQUIREMENTS OF A MATERIALS DESCRIPTOR

For a representation to efficiently capture materials information, a key requirement is obeying the physical principles that are known to govern molecular species. The following principles are essential to correctly representing molecules and materials:

1. The representation must be invariant to the order in which atoms are provided in the input.
2. The representation must be invariant to the order in which bonds are provided in the input.
3. The representation must be invariant to the order in which atomic pairs are labeled in the input.

In general, any learning algorithm using a representation that is not blind to these principles will possess the variational flexibility capable of distinguishing between physically meaningless differences.

2.2. Nonneural Fingerprints

Two-dimensional representations encode a molecular graph specified by the identities of the atoms along with the bonding that links those atoms together. As there is no encoding of coordinates, two-dimensional representations are implicitly invariant to translation and rotation, and the remaining symmetry consideration is thus the order in which the atoms and bonds are provided.

A litany of such representations are possible (17); some of the most frequently used are the simplified molecular input line entry system (SMILES) (18), the International Chemical Identifier (InChI) (19), the Molecular Access System (MAACS) (20), and extended connectivity fingerprints (ECFP) (21).

Both SMILES and InChI are character-string-based representations. The SMILES string is built up by traversing and recording the heavy atoms present in a modified version of the molecular graph where rings have been broken and the atoms with broken bonds annotated to indicate their presence. InChI, by contrast, is simply a layer-wise specification of the molecular formula, connectivity, bound hydrogens, charges, stereochemistry, and isotopes and is not invariant to permutation of atomic index. SMILES is invariant to these mutations, and due to its character-based nature, it invites the use of modeling techniques appropriate for natural language processing. Correspondingly, it has found popular usage within deep learning and in particular for the generation of novel molecular species.

An alternative representation method is the use of binary vectors also known as chemical fingerprints. Many such encodings have been developed as a means to search databases of molecules according to the presence or absence of particular subgroups, and they have since been taken up by the chemical machine learning community. MAACS keys are a 166-dimensional binary vector, where each element, or key, corresponds to some molecular feature such as the presence of a particular functional group. ECFP is a member of a family of circular fingerprints that constitute one of most widely used tensor-based topological representations. When building an ECFP representation, the number of paths that pass through each atom within a particular cutoff radius is encoded for each atom, and the result is compressed via hashing to fit a predefined length. This systematically characterizes each atom, in contrast to the predetermined molecular features specified by the MAACS keys.

Due to the hashing present in ECFP and the fixed number of molecular features captured by MAACS keys, neither constitutes a complete representation that allows full reconstruction of the molecular graph it is encoding. This is in contrast to SMILES and InChI, and they are thus better suited to regression and classification problems than to the generation of novel molecular species.

The state of the art for representation of a molecular graph has moved to the use of deep learning, though there is still some research into nonneural methods such as the N-gram graph representation (22). Within neural methods, the aforementioned representations—in particular SMILES—are still in use as basic starting points from which deep networks can extract patterns. They are complemented in this use by more basic representations that correspond to a matrix specifying the pairwise connectivity of atoms within the molecule and some set of vectors specifying the atomic identities. This basic representation violates the invariance to permutation symmetry, so enforcing this within the final learned representation falls to the neural architecture.

While encodings of the molecular graph are implicitly invariant to rotation and translation (as these degrees of freedom are not present), the same is not true when one is attempting to capture the full atomic coordinates. Thus, consideration of how to achieve symmetry invariance is more challenging and is a key issue in the search for more expressive three-dimensional molecular representations.

An early three-dimensional representation constructed by Behler & Parrinello (23) encodes atoms within the molecule via some number of radial and angular symmetry functions. A later development following similar principles, designed by the Csányi group (24), is the smooth overlap of atomic positions.

Another early attempt at three-dimensional molecular representation, this time inspired directly by the molecular Hamiltonian, was developed by the von Lilienfeld group (25). Known as the Coulomb matrix, this is a matrix representation where the off-diagonal elements correspond to the Coulomb nuclear repulsion terms between atom pairs while diagonal elements encode the atomic charge. One downside of this design is a lack of permutational invariance. Further work by the von Lilienfeld group that sought to overcome this shortcoming, and to increase the accuracy of machine learning methods applied to ab initio computational data, led to the bag of bonds representation (26); the bonds, angles, machine learning representation (27); the FFLA crystal representation (28); the spectrum of London and Axilrod-Teller-Muto potentials (29); and the molecular alchemical radial angular distribution along with histograms of distances, angles, and dihedral angles (30). Overviews of the performance of the three-dimensional fingerprints listed above for various different learning architectures can be found in Faber et al.'s (30) work.

Just as string-based representations can allow for the application of language processing machinery to molecular inference, representations that are based on the voxel—the three-dimensional analog of the pixel—can allow for the application of three-dimensional scene processing machinery to molecular inference. Along these lines, some voxel-based three-dimensional molecular structure representations have started to appear (31, 32).

While deep neural representations for encoding full atomic structure are gaining popularity, research into nonneural three-dimensional representations continues to thrive and indeed offers comparable performance. More recent developments include the many-body tensor representation (33) and wavelet-transform-based methods (34), which, together with histograms of distances, angles, and dihedral angles, offer competitive performance to neural methods.

2.3. Neural Fingerprints

The progress described above in the ability of machines to see molecules has required the cumulative efforts of human researchers investigating and discovering useful representations. A complementary approach is the use of deep neural networks to learn an effective representation.

In some sense, this passes the job of discovery to the algorithms themselves, albeit at the cost of pushing the problem for researchers into the search for effective network architectures. The promise of progress in automatic discovery of appropriate architectures (35) seems poised to reduce this barrier even further.

Neural fingerprints that encode the molecular graph were introduced by Duvenaud et al. (36). Inspired by how ECFP worked, Duvenaud et al. adapted the convolution neural network architectures previously used to learn representations of images by progressively extracting higher-order structure from the raw image, finally learning representations of molecules by progressively extracting higher-order structure from the raw graph structure. This initial work was complemented by Kearnes et al. (37), who added more sophisticated ways of dealing with permutational invariance, and was further generalized within a message-passing framework (38).

The Behler & Parrinello (23) symmetry functions have been extended by Smith et al. (39) in their neural network potential ANI to build single-atom atomic environment vectors as the molecular representation for a richer embedding of local atomic information. An alternative approach, known as tensor field networks, operates on point cloud representations (40), and a close analog—deep tensor networks—operates on featurized interatomic distance matrices combined with a vector of atomic identities (41, 42).

Neural fingerprints are generated by attempting to solve a particular inference task. In doing so, the final fingerprint expresses molecular information suitable for solving the inference problem it was trained on. Multitask learning, where multiple chemical properties are learned at once, allows for the generation of more general transferable fingerprints. Perhaps counterintuitively, forcing a fingerprint to express information suitable for multiple chemical tasks can often improve the performance of the individual subtasks (43), a phenomenon well known in the machine learning community (44). Any neural fingerprint architecture can be combined with multitask learning, so the question of how to choose complementary tasks to control the information expressed within the fingerprint has been investigated (45).

3. LEARNING TO ESTIMATE: BUILDING QUANTITATIVE STRUCTURE-PROPERTY RELATIONSHIP MODELS FOR MATERIALS DISCOVERY WITH MACHINE LEARNING

The crux of materials design and discovery is to identify mutuality of the four materials elements—structure/composition, properties, synthesis/processing, and performance—which is often depicted by Flemings's (46) tetrahedron. The four elements have been at the core of materials design in the name of QSPRs or PSP linkages in the materials community (47). As described in Section 2, QSPRs are considered to be some of the fundamental building blocks of the MGI. The high-dimensional, complex nature of QSPRs creates a vast search space where the use of machine learning makes it effective to enhance navigation and linear/nonlinear mapping methods to find or design new materials at much lower computational cost. Machine learning has a wide range of applications to the development of QSPR models, and it is particularly useful when physics-based models do not exist or when we need to link the different timescales or length scales in materials modeling.

The differences between materials properties observed in the laboratory and those observed in the simulated world can be addressed only through the close interaction between theory, computation, and experiments. Materials discovery is not only about the new materials but also about enhancement of existing materials. Thus, machine learning can provide capabilities that will increase the speed of the discovery and development of materials through the QSPR estimation tasks. For the sake of argument, in this section we briefly discuss only a few tasks to learn crystal structure, microstructure, and materials characteristics. Other major machine learning tasks, such as materials search via high-throughput screening or Bayesian optimization by navigating QSPR space, property prediction, and synthesis optimization, are discussed in Section 4.

Materials property prediction starts with identifying the chemical composition and crystal structure. In the early 1980s, there was a great deal of discussion about the use of materials descriptors. For example, Zunger's (48) early work using pseudopotential orbital radii to create structural separation plots (i.e., stability maps that separate different structural types) helped to predict the stable crystal structure of known AB-type binary compounds. This brute force type of classification approach for structure recognition has extended to data-mining-based QSPR mapping (49), and more recently, similar concepts were applied to deep neural network approaches. Thus, two classical descriptors related to crystal stability and synthesizability—the Pauling electronegativity and Shannon ionic radii—were used to construct deep neural networks that predict the density functional theory (DFT) formation energies of $C_3A_2D_3O_{12}$ garnets and ABO_3 perovskites from the identified relationship among composition, structure ordering, and energy (50). To avoid crude and fuzzy structure classifications, more advanced learning techniques such as general recursive schemes are now used to automatically classify structures, making it possible to curate the materials structure genealogy (51). In parallel, machine learning has become a more attractive option by offering tools such as the random forest algorithms to overcome the high cost of theoretical calculations such as DFT or force field modeling for each crystal structure (52). For further information, we refer readers to the recent review about structure prediction methods by Needs and colleagues (53).

Materials observables such as the microstructure are related not only to materials properties but also to the postsynthesis processing of a material, and mapping of a huge microstructure-property space is thus a complex task. Therefore, identification of tailored or optimal microstructures for desired materials properties remains a major challenge for inverse design. Recently, McCue et al. (54) used data-mining-assisted automated image analysis of published nanoporous gold electron micrograph images to explore PSP relationships. To identify the key microstructure representations, Wodo and colleagues (55) developed a surrogate method for compact microstructure-property mapping by treating microstructures as graphs. According to recent work by Agrawal and colleagues (56), deep neural networks such as generative adversarial networks (GANs) can learn the mapping between latent variables and microstructures, leading to an optimal microstructure with targeted material properties via optimization of the latent variables.

Huge amounts of complex microstructure-related data about materials are being generated by recent advancements in characterization techniques, including X-ray photoelectron spectroscopy for mapping chemical images, electron backscatter diffraction for identifying grain boundary types, and cathodoluminescence for studying the distribution of recombination centers. Moreover, the complexity of the image data becomes more significant when the microstructural responses are connected in terms of external environments such as stress. To handle this complexity of morphological and crystallographic data (such as grain volume, grain shape, and orientation distribution regarding the morphology and spatial arrangement of grains), there are various efforts across the materials community to identify or develop QSPRs via machine learning computations. Among these, it is notable that a method based on deep neural networks finds its own applications in the area of high-resolution electron microscopy. With the aid of deep neural networks, for example, it is possible to identify the positions of atoms in a lattice and the types of atomic species in real time and to detect and classify defects from scanning transmission electron microscopy data (57). Convolutional neural networks (CNNs) were trained to classify atomic structures in high-resolution transmission electron microscopy (58). Kalinin and colleagues (59) showed the effective use of deep CNNs for real-time phase analysis (i.e., phase formation and evolution mapping) to automatically identify symmetry classes in atomically resolved images from electron and scanning probe microscopies. Apart from their usefulness as a tool to detect structural features, such deep learning algorithms can also be applied to automated high-resolution electron microscopy calibration that aims at enhancing the robustness of measurement. For example, Xu & LeBeau (60) showed how to extract electron microscopy parameters from position-averaged convergent-beam electron diffraction patterns analyzed automatically by a CNN.

When we apply machine learning to materials discovery and design, the types of data from which we can learn range from characterized results such as images, to modeling results, to processing histories. In other words, a goal should be to collect data or develop databases to train from various sources of information. The materials discovery and design community continues to develop methodologies for collecting scientific information—in particular, to automatically collect complex processing data—for insight into synthesis recipes for materials.

Natural language processing (NLP) is getting more attention from the literature as an automated data population tool. Olivetti and colleagues (61–63) successfully used NLP to understand zeolite synthesis-structure relationships and created an autonomously compiled synthesis planning resource. Using NLP and semisupervised relationship extractions from the literature, Court & Cole (64) developed an autogenerated database of chemical compounds including Curie and Néel magnetic phase transition temperatures. Semiautomatic mining was used to tag processing conditions and physical property information in the literature to establish the property-processing relationship of oxide films in the pulsed laser deposition process (65). Similar approaches include automated image collection from the published literature and machine learning training on a text corpus in scientific articles (54, 66). The key challenge in NLP or text mining for automatic data collection is to accurately codify literature information, such as entire synthesis routes, on the basis of synthesis ontology (67, 68).

While many literature information attempts report successful results, it is also just as important to learn from failed attempts. As Norquist and colleagues (69) pointed out, failed experiments are a great source of information for the next set of successful experiments. They trained a support vector machine on a failed data set to predict reaction successes or failures, and the results were remarkable.

Automated and/or autonomous materials discovery and design processes are essential to speed up results and minimize human error, and there are many recent activities in this area. Good examples include automated molecular and alloy design (70, 71), automated theoretical calculations such as ab initio calculations (72), and fully automated machine-learned potentials with active learning (73). Robots are now used for autonomous assembly of materials such as van der Waals superlattices (74). Autonomous manufacturing at the atomic scale (i.e., atomic fabrication processes) is realized by using deep learning such as a CNN. Here, deep learning is used to identify surface features to precisely pattern atomic structures (75). Aspuru-Guzik and colleagues (76) pointed out that the immediate task in enabling smart automation in the materials discovery process is to integrate autonomous synthesis planning, automated chemical synthesis, and autonomous experimentation as a closed-loop workflow.

3.1. The Importance of Data

It is commonly understood that the power of machine learning is inherently tied to the quality of the data sets. Unfortunately, the materials and chemistry communities often work with a limited number of databases or small data sets. Compared with materials databases created by theoretical calculations or simulations, experimental databases containing synthesis procedures/history or processing conditions are still too rare to allow understanding of QSPRs even in the community (77). One way to overcome limited-data problems is to employ transfer learning by using a model trained with extensive data for a new task with limited data. For example, transfer learning is becoming more popular in the microstructure reconstruction area. This is mainly because the area aims at constructing statistically equivalent microstructures with very limited information about the original structure (78). Zhang & Ling (79) proposed to approximately predict targeted properties by using nonexpensive modeling and then to use the estimation to establish machine learning models that enhance prediction accuracy. A limitation of transfer learning is that it requires a very well-trained model to apply to specific target tasks (80). Other approaches to overcome limited-data problems include meta-learning, neural Turing machines, Bayesian frameworks, fast surrogate machine learning models, and machine learning models constrained with dimensional analysis and scaling laws (81–83).

Additionally, heterogeneous data sets or scattered data from various sources are common in the materials and chemistry areas. While integration of databases or scattered data sets is beneficial for better understanding of QSPRs, it is possible to miss significant features. A recent paper suggested approaches that individually model data from distinct sources first and then employ a stacked ensembling that combines simpler models by using two layers—the multiple-model layer and the prediction layer—of machine learning (84). Similarly, the combination of different learning algorithms (for example, the combination of machine learning with evolutionary algorithms) is possible. Transfer learning using ensemble neural networks is useful for addressing the issues of heterogeneous data and quality (85, 86). Another approach to deal with a heterogeneous database or scarce data is simultaneous multitask learning—such as the sure-independence screening and sparsifying operator—that identifies key descriptors representing multiple target properties simultaneously (87).

3.2. Current Applications of Machine Learning Algorithms to Materials Design

Materials designers widely use machine learning for key tasks such as correlation, prediction, and optimization of design parameters to precisely control processes and achieve target properties. A more novel application, however, is the use of machine learning to integrate individual models describing chemical and physical phenomena occurring at different timescales and length scales to comprehensively understand QSPRs. The algorithms described in this section are examples of the wide variety of machine learning algorithms used for materials discovery and design today.

While increasing the number of parameters in materials modeling, simulation, synthesis, and characterization provides materials designers with additional design options (e.g., designers can select or combine parameters to target certain properties), it also gives rise to the curse of dimensionality. To address this problem effectively, materials designers often use high-dimensional visualization techniques. These techniques not only provide low-dimensional visual representations but also illuminate relationships between variables. Suh et al. (88) demonstrated the role of various visualization techniques for *N*-dimensional data generated by high-throughput experimentation. They introduced several techniques to the high-throughput experimentation community, including parallel coordinates, radial visualization mapping, heat maps, and glyph plots. Recently, Rickman (89) used parallel coordinates to create materials property charts showing the property correlations between different materials classes.

Dimensionality reduction algorithms are used to identify and visualize the structure of high-dimensional data sets in low-dimensional space. One example of a traditional least-squares-loss-based spectral decomposition algorithm is principal component analysis. Principal component analysis is a linear dimensionality reduction method that finds directions of maximal variance in data to preserve as much of the original high-dimensional data structure as possible. This method serves as a major dimensionality reduction technique for dealing with various types of materials data, such as powder X-ray diffraction patterns (90). Although most QSPRs are expressed nonlinearly, attempts to nonlinearly capture high-dimensional data structure—such as manifold learning, which focuses on local distances between features—have been quite rare for materials problems. A good example of manifold learning for QSPRs is a diffusion map approach combined with hierarchical clustering to learn optimal thin-film process conditions to develop an Al-doped ZnO layer in copper indium gallium diselenide solar cells (91). Recently, similar types of manifold learning, such as the t-distributed stochastic neighbor embedding (t-SNE) algorithm, are becoming more pervasive for obtaining low-dimensional representations in a nonlinear way. The t-SNE algorithm focuses on preserving the local distances of high-dimensional data while identifying the data's global structure. For example, Zakutayev et al. (77) demonstrated the power of the t-SNE algorithm for visualizing the most common compositions in the High Throughput Experimental Materials Database.

Ever since Li (92) showed the applicability and effectiveness of classification and regression trees—or simply decision trees—for understanding materials behavior by using creep rupture data on austenitic stainless steels, the materials community has used this approach in various applications, such as microstructure optimization, due to its simple interpretability (93). However, to avoid the overfitting problem in decision tree algorithms, a random forest algorithm can be employed that simultaneously uses multiple decision trees. The random forest algorithm has numerous applications, including capturing nonlinear dependencies in data and better predicting properties such as the critical temperature of superconductors (94). Decision tree algorithms have evolved over the years, leading to approaches such as the Monte Carlo tree search, which approximates optimal decisions in QSPR space (95). One of the most popular classification and regression algorithms in materials design procedures, on the other hand, is a support vector machine (SVM). SVMs find the best line or hyperplane that separates the data into classes. They use a type of mathematical transformation called the kernel trick to find high-dimensional space where data classes are linearly separable and then project the high-dimensional information to the original low-dimensional space. SVM has been used for various types of classification and regression problems, such as predicting the voltage of electrode materials (96).

The goal of deep learning in materials design is to automatically identify complex relationships between input and output in terms of QSPRs, making it more computationally efficient than traditional learning algorithms. For example, CrystalGAN—a GAN-based architecture—automatically generates chemically stable crystallographic structures, such as new ternary crystallographic structures, by using existing binary information (97). Agrawal and colleagues (98) presented a deep neural network model called ElemNet. The model uses the elemental composition of a compound as the input, captures the physical and chemical interactions between elements, and predicts materials properties such as the formation enthalpies of compounds. Similarly, the deep learning architecture SchNet allows one to perform various quantum chemistry tasks, such as property prediction of molecules or materials, potential energy surfaces, and force fields (42).

4. LEARNING TO SEARCH

The ability to accurately model processes and properties by using data-driven techniques is certainly powerful and scalable, but it does fall foul of one of the tenets of successful research—never stop learning. When we build a model and never update it, that model is frozen with the knowledge—and associated biases—of a particular point in time, and from that moment it begins to become stale.

Increasingly, materials discovery is being treated as an active learning problem, in which models are continually updated as new information flows in. This, in itself, poses an interesting challenge: How do we ensure that models built in this way do not themselves end up biased by the methodologies that were used to determine which data were collected? In the world of machine learning, this is known as the exploration-exploitation trade-off. A purely exploratory approach to collecting data (sometimes known as a diversity-driven approach) will try to build the most generally applicable models by considering as wide a range of data points as possible. While this process is much less prone to bias, the resulting models are often less predictive in any one specific case. A purely exploitative approach to model building will result in a model that is fantastic for a specific use case but does not transfer well to data beyond those used to train it. Clearly, the key is to balance these two extremes to build an efficient materials search pipeline.

There are two major approaches taken when searching materials space: the active ranking and reprioritization of an existing library, undertaken with the understanding that only a small percentage of the enumerated library will be tested, and the generative approach, where completely new materials are dreamed up, often through decoding of a constructed latent space or through an evolutionary approach. Both of these methods have their pros and cons; the construction of a library, for example, affords the user the close control over the kinds of materials suggested for testing that is by definition lacking in the purely generative approach. This can, however, lead to bias in library construction and to a subtle favoring of particular materials subclasses. Here, we look at methodologies that fall into both camps, with particular attention to deep reinforcement learning and Bayesian optimization.

4.1. Bayesian Optimization

Bayesian optimization is an efficient black-box optimization technique often used for the optimization of expensive functions. In contrast to many other optimization techniques, Bayesian optimization does not require the formulation of gradients but instead formulates a response surface, which itself is used as a surrogate function. Typically, this response surface will be generated using a Bayesian model known as a Gaussian process, although other models, such as Bayesian neural networks, have also been used, with notable successes.

Essential to the success of Bayesian optimization is the fact that the optimization algorithm takes into account both the predictions of the model and the uncertainties in these predictions. This combination of exploration and exploitation is encoded through an acquisition function, which typically relies on the concept of improvement, usually formulated as

$$\gamma(x) = \frac{\mu(x) - f^*}{\sigma^2(x)}, \qquad 1.$$

where f^* is the best target value observed so far, $\mu(x)$ is the predicted mean, and $\sigma^2(x)$ is the corresponding variance.

Given some set of data already observed, the next data point to be acquired is found through maximization of the acquisition function. The most commonly used acquisition function is known as expected improvement (EI):

$$\mathrm{EI}(x) = \mu(x) - f^*\Phi(\gamma) + \sigma(x)\phi(\gamma), \qquad 2.$$

where Φ denotes the cumulative distribution function of the standard normal distribution, ϕ denotes the probability density function of the standard normal distribution, and γ denotes the improvement described in Equation 1.

After this data point is acquired, the Bayesian model is refitted to include the new data, and the cycle repeats until either the user runs out of budget or the acquisition function becomes zero (indicating no advantage to sampling any more data) (**Figure 2**).

Bayesian optimization is particularly conceptually suited to chemical search problems, as it effectively mimics the scientific discovery process of hypothesize, test, observe, (re)evaluate (99, 100). Perhaps because of this, Bayesian optimization has been applied in a wide variety of areas, facilitated by its black-box nature and its conceptual similarity to the scientific method, which aid in its uptake in the scientific community.

Due to the similarities between Bayesian optimization and classic design-of-experiments (DoE) methodologies, it is unsurprising that researchers have begun to find success in using this technique to tune experimental conditions for materials synthesis and processing. Venkatesh and colleagues (101) used Bayesian optimization to accelerate the synthesis of polymer fibers—a complex challenge that had defied classical DoE approaches due to the high dimensionality and complex interdependencies of the parameters. Through the use of automated synthesis techniques, the researchers were able to automatically screen conditions, and they followed a multiobjective design protocol with experiments judged on deviations from ideal fiber length and diameter. The

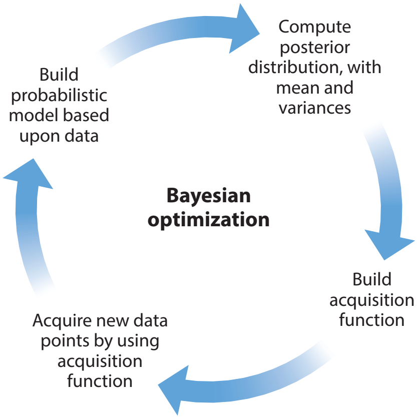
Compute posterior distribution, with mean and variances
Build acquisition function
Acquire new data points by using acquisition function
Build probabilistic model based upon data
Bayesian optimization

Figure 2

The cycle of Bayesian optimization: Models are continuously updated as new data are acquired, a process determined by the acquisition function.

authors state that their procedure will never self-terminate—practically true, although Bayesian optimization terminates when there is a global zero score for the acquisition function, which is rare in high-dimensional search—and they recommend tracking the best found value for a termination heuristic. When there has been a long plateau in the best found value, the search is terminated.

Meredig and colleagues (102) have also used Bayesian optimization for automating the materials discovery process. Unlike in many applications of Bayesian optimization, they used random forests for their underlying model. While these models are not inherently Bayesian, as are the more traditionally used Gaussian processes, they are more scalable. In this study, the uncertainty estimates necessary for a Bayesian optimization approach were calculated by using bias-corrected infinitesimal jackknife and jackknife-after-bootstrap estimates, which were proposed as a high-quality alternative methodology. The researchers tested their methodology on a wide range of materials discovery and processing problems, including maximizing the magnetic deformation of magnetocaloric materials, the critical temperature of superconductors, the figure of merit of thermoelectrics, and the fatigue strength of steel. The methodology was shown to outperform random search, an achievement that, while not often cited as a challenging benchmark, was deemed to be significant because the materials and process parameters searched over had already been deemed worthy of investigation, so random search over this biased set should have produced reasonable results.

The similarity of Bayesian optimization to DoE methodologies has also been exploited in the area of combinatorial library screening, an increasingly important tool in materials discovery. Common Bayesian Optimization Library (COMBO), a toolkit developed by Tsuda and colleagues (103), has been applied to the acceleration of materials discovery workflows. One metric for success is the *k* recall, which measures how many of the top *k* candidates are found at a fixed budget. With a budget of 300 iterations, and setting $k = 30$, for instance, a random search has a score of 40%. When COMBO is used with the same configuration, this score is enhanced to 83%, which is indicative of the power of these kinds of optimization techniques to drive efficiency.

In the real world, optimization problems are rarely sequential, instead requiring multiple evaluations of an objective to occur simultaneously. One clear example of this is the connection between Bayesian optimization and robotic screening. In a robotic screen or assay, there are a fixed number of slots that can be used to obtain data, and there is no significant benefit to not using all available slots. Groves & Pyzer-Knapp (104) developed *k*-means batch Bayesian optimization (KMBBO) and its high-dimensional analogy, KMBBO with compressed sensing (CS-KMBBO), to deal efficiently with this situation. In KMBBO, the acquisition function is split into sections by using a *k*-means algorithm, with *k* determined by the number of slots available. The authors credit the efficiency of this approach to its combination of peak picking via *k* means, which exploits the current acquisition function, with the spherical repulsion term in the *k*-means clustering algorithm, which ensures that any slots that do not correspond to peaks are instead filled with a diversity-driven subset. On a drug discovery challenge, KMBBO achieved very low regret after only 10 epochs of sampling (for a total of 100 molecules), when the library being sampled consisted of around 20,000 molecules.

Given the proven ability of Bayesian optimization techniques to rapidly winnow large chemical libraries, it is unsurprising that there has been some significant success in their application to high-throughput virtual screening. Pyzer-Knapp and colleagues (105) have demonstrated the use of parallel and distributed Thompson sampling to rapidly screen quantum mechanical simulations of the highest occupied molecular orbital, the lowest unoccupied molecular orbital, and the power conversion efficiency to efficiently discover high-potential candidates for organic photovoltaics. In this study, the researchers were able to discover candidates with significantly high power conversion efficiency around 30 times quicker than other search techniques, by exploiting a new parallel Bayesian methodology to evaluate multiple candidates simultaneously.

Arróyave and colleagues (106) used Bayesian optimization to accelerate a high-throughput search of the MAX ternary carbide/nitride space (where M is an early transition metal, A is an A-group element, and X is C or N) through DFT calculations. In this study, bulk and shear modulus were optimized separately as single-objective optimizations, and the results were then compared with a multiobjective optimization process in which both were optimized simultaneously (see the sidebar titled Single-Objective or Multiobjective Search?). In addition, in this study, a Bayesian model averaging approach was used to automatically select features for inclusion in the model driving the optimization, which was shown to be robust to the dimensionality reduction performed in this way.

4.2. Reinforcement Learning

Reinforcement learning is a subfield of AI that is targeted at learning how to make decisions that maximize a reward over a time horizon. In the reinforcement learning paradigm, agents learn how to maximize this reward through interactions with their environment. In contrast to Bayesian optimization, reinforcement learning is better considered as learning to control a process rather than strictly optimize it. Reinforcement learning has shown great potential in learning to play games such as the Atari suite and Go (107), as well as in simulation challenges such as protein folding (108).

SINGLE-OBJECTIVE OR MULTIOBJECTIVE SEARCH?

While many search problems may initially look to be single objective (i.e., maximizing some desired property), in an industrial context, this is rarely the case. A material that is viewed as the best material from a single-property perspective could be rejected due to processing concerns, toxicity concerns, or even financial concerns. Machine learning techniques, through their general applicability, can offer a single framework that, given data, can model all of these objectives, thereby allowing a simpler route to multiobjective optimization.

DEFINITIONS FOR REINFORCEMENT LEARNING

Reinforcement learning is a terminology-driven literature, so it is worth defining some key terms.

- **Agent:** Agent is the name for an entity that takes actions.
- **Action (A):** The set {**A**} contains all possible actions that the agent can take.
- **Discount factor:** The predicted future rewards are often adjusted by the discount factor to dampen their effect on the agent's choice of action.
- **Environment:** The environment is the world through which the agent moves; it takes the agent's current state and action as input and returns as output the agent's reward and next state.
- **State (S):** A state is a concrete and immediate situation—that is, a specific place and moment—in which the agent finds itself.
- **Reward (R):** A reward is the feedback by which we measure the success or failure of an agent's actions.
- **Policy (π):** The policy is the strategy the agent employs to determine the next action from the current state.
- **Value (V):** The value is the expected long-term return with discount, as opposed to the short-term reward R.
- ***Q* value:** The Q value, also known as the action value, is conceptually similar to value except that it also takes into account the current action a.

The goal of reinforcement learning is to pick the best known action for any given state, and it is therefore necessary to rank the actions and assigned values relative to one another (see the sidebar titled Definitions for Reinforcement Learning). To achieve this, state-action pairs are mapped to the values they are expected to produce with the Q function. The Q function takes as its input an agent's state (S) and action (A) and maps them to probable rewards. This can be calculated by the Bellman equation:

$$Q_{\text{new}}(S,A) = Q(S,A) + \alpha R(S,A) + \gamma\ \{\max[(S',A') - Q(S,A)]\}, \qquad 3.$$

where γ is the discount factor and α is a learning rate.

Reinforcement learning is the process of running the agent through sequences of state-action pairs, observing the rewards that result, and adapting the predictions of the Q function to those rewards until it accurately predicts the best path for the agent to take.

If one is able to calculate the reward for any given action in any given state, then optimization is simply implemented as a lookup table; this is known as a Q table. Unfortunately, the calculation of the Q function, and construction of a Q table, can be formidably expensive (**Figure 3**). Deep Q learning (sometimes referred to by the more general term deep reinforcement learning) replaces the construction of a lookup table with the construction of a model that can relate states and actions to estimated rewards. This model is often a deep neural network—hence the term deep reinforcement learning—and can allow users to access a much larger and more complicated space of potential actions.

Deep reinforcement learning lends itself to the challenge of materials design and process optimization. For example, Popova et al. (109) have developed a computational design strategy based upon deep reinforcement learning, which they term reinforcement learning for structural evolution (ReLeaSE). In their strategy, two deep neural networks are trained—one for generative tasks and one for predictive tasks. These two networks are trained separately but are then combined and used jointly to generate the desired chemical libraries. After training the separate networks by using traditional supervised techniques, ReLeaSE uses reinforcement learning to bias the generative task so that molecules with desired functions and properties are generated. The authors found that

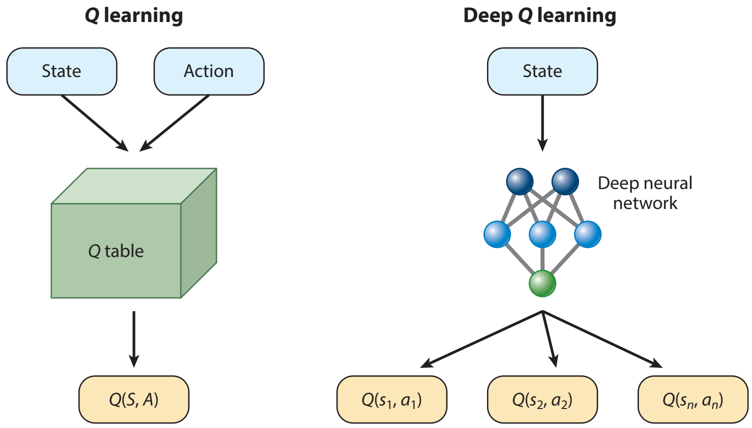

Figure 3

The process of *Q* learning in its most vanilla form (*left*) involves the determination of *Q* values given specific states and actions through what is essentially a lookup table. Deep *Q* learning (*right*) replaces this expensive computation by replacing the lookup table with a deep neural network that maps states and actions to a predicted *Q* value.

the majority of compounds generated by the ReLeaSE method were novel structures, displaying significant differences from the molecules used to train the model; this creativity demonstrates the potential for reinforcement learning to generate desirable targets outside of existing data sets. In a similar manner, Olivecrona et al. (110) used deep reinforcement learning to design SMILES strings that were predicted to have desirable materials properties. The authors demonstrated how this approach generates strong results across a wide range of tasks that are relevant to materials design, such as generating as yet unseen analogs to a given molecular structure and generating novel structures predicted to have highly desirable properties.

Since reinforcement learning can easily be understood as learning to control, it is understandable that its use for controlling and optimizing reaction conditions has been investigated. One particularly strong example is presented by Zare and colleagues (111). These authors used deep reinforcement learning to discover optimal reaction conditions in a flow-chemistry setup. They found that this methodology outperformed a state-of-the-art black-box optimization algorithm, stable noisy optimization by branch and fit (SNOBFIT), by using 71% fewer steps on both simulations and real reactions. As experiments are naturally expensive to evaluate in both time and cost, they used simulations to bootstrap the learning, greatly decreasing the cost of the optimization.

4.3. Generative Models

While methods such as Bayesian optimization and reinforcement learning are extremely powerful on their own, it is arguable that the real evolution of the materials genome comes when these techniques are combined with an element of creativity, generating new materials that have not been conceived of by a scientist at all (112). These creative algorithms are referred to as generative models, as instead of predicting the behavior of a known input (material), they generate new inputs (materials) that are predicted to have powerful properties. A more detailed review of the application of generative modeling to molecular species can be found in Reference 113.

The most popular generative methods are variational autoencoders (VAEs) (114) and GANs (115); the former have seen more use within molecular and materials science (113, 116).

VAEs are neural networks trained to reproduce their input but structured in two pieces: The first is an encoder that takes a base representation and learns an effective neural fingerprint, and the second is a decoder that learns how to reconstruct the molecular species from the neural fingerprint. The space spanned by the neural fingerprint is known as the latent space. After training, novel molecules can be generated by using the decoder and randomly sampling (or indeed searching) the latent space. In this field, VAEs were first applied by Gómez-Bombarelli et al. (117) by using SMILES as the base representation. Implicitly, what is desired is for the VAE to map molecules to a latent space such that the latent space is densely populated (i.e., there are no gaps where regions of the latent space do not correspond to meaningful chemical species) and such that chemically similar molecules are close to one another. There are indications that autoencoders do indeed achieve these aims (118). Notions of what precisely is meant by chemically similar can be altered by cotraining the autoencoder alongside property predictions (119). This approach allows for specialized use to generate molecular species with particular properties of interest, for example, for drug discovery (120). One issue with the use of SMILES as the base representation is the challenge of achieving a densely populated latent space; this is due to the ease with which one can construct invalid SMILES strings that do not correspond to valid molecular structures. Attempts to solve this have included introducing a grammar (116) to further constrain the generated strings or switching to a different base representation. The main appeal of SMILES is its status as a complete encoding of the molecular graph that does not violate any symmetry considerations. Both NeVAE (121) and the junction tree encoder (122) are non-SMILES-based autoencoder architectures. Like the two-dimensional neural fingerprints (36–38), NeVAE uses a base representation composed of a matrix of bond orders and a vector of atom identities. This representation is not implicitly invariant to permutation; however, the NeVAE architecture enforces this symmetry. The junction tree encoder operates on a similar base representation but maps it into two pieces: a traditional encoding of the graph connectivity alone, as in the two-dimensional fingerprints and NeVAE, and an additional latent space representation known as the junction tree that encodes the identities of clusters of atoms. These are then combined during reconstruction.

Winter et al. (123) created an autoencoder-like architecture; their innovation was to focus on translation between different base descriptors rather than simply reconstructing a single base descriptor in an attempt to extract more chemically meaningful content that is less biased.

An alternative to VAEs, GANs operate on the basis of networks that again divide into two components. Instead of an encoder and decoder trained to reproduce molecules, GANs operate via a generator and discriminator. The latter receives outputs from the generator as well as real examples of training molecules and is trained to distinguish between the two, while the former receives an initial vector that plays the role of the representation/latent space and is trained to produce outputs that fool the discriminator. Once trained, the generator can be used in place of the VAE decoder to generate new molecular species, though unlike in a VAE, the inverse transformation, which maps a molecule to its representation, is not possible. While pure GANs have not seen much application to materials generation, some examples exist (124), and combinations of VAEs and GANs have seen some usage (31, 120, 125).

A different approach to the generation of new materials is to use an evolutionary algorithm. Evolutionary algorithms have been inspired by mechanisms found in biological evolution, such as reproduction, mutation, recombination, and selection. Berardo et al. (126) applied evolutionary algorithms to the construction of tetrahedral cages with applications as porous materials. They used crossover and mutation operations to successfully build libraries of cage molecules that were optimized for porosity, window size, and symmetry—a fitness function that was determined through a survey of the existing literature on successful porous organic cages.

Evolutionary algorithms are typically bottlenecked by the computational expense of evaluating the fitness function (127), so in the materials domain, it is more usual to see common fitness functions, such as relative energy, evaluated by using cheaper semiempirical potentials than to see them evaluated by more accurate methods such as DFT. Jennings et al. (86), however, used a machine-learning-based surrogate model to improve the description of the fitness surface at a low computational cost, reducing the total number of energy minimizations required to fully search the convex hull of local minima from 16,000 to 300. Another example of the fusion of genetic algorithms with machine learning models was undertaken by Simmons and colleagues (128). These authors employed a deep learning approach to build a model that biases the evolution of a genetic algorithm, with fitness evaluations performed using both simulation and experimental approaches. This is a particularly powerful approach, as it gives the evolutionary algorithm the ability to continuously learn and draw inferences from its experience to accelerate the creative and generative processes.

5. OUTLOOK

As the field of machine-learning-aided materials research and development quickly evolves, the materials and chemistry community recognizes that more integrated efforts will play a significant role in advancing materials discovery and design procedures. There are various technical reviews—for example, reviews of materials informatics (129), the history of high-throughput experimentation and combinatorial approaches (130), machine learning applications in continuum mechanics of materials (131), deep learning for molecular design (132), deep materials informatics (133), and deep learning in chemistry (134), as well as a community-wide road map (135)—that describe the community's accomplishments and future directions. As outlined in the Future Directions, the community focuses on understanding fundamental behaviors of materials by using machine learning, establishing data sharing and integration, and increasing our experience via a convergence approach. For example, pioneering machine learning work toward fundamental materials problems includes automatically identifying material phase transitions (136), understanding microstructurally small fatigue-crack-driving forces (137), and predicting fracture propagation and failing modes (138). To accomplish these tasks, it would be powerful to perform machine-learning-based inverse design instead of forward modeling (112, 139). The lack of a robust method or standard protocol for exchanging data between data sets or databases using complex schemas makes data integration and sharing difficult. The use of more standardized methods—such as categorical query language, an XML-based data schema such as the NanoMine polymer nanocomposite schema, or machine-readable formats and open collaborative frameworks—would be useful to reduce data-sharing problems and facilitate identifying QSARs (140–142). Convergence approaches to realize the materials genome include the National Institute of Standards and Technology's High-Throughput Experimental Materials Collaboratory (HTE-MC). In collaboration with the National Renewable Energy Laboratory, HTE-MC is focusing on combining high-throughput materials synthesis, processing, and characterization (i.e., physical systems) with data management (i.e., virtual processes) (143). Similar concepts exist, such as the concepts of integrated collaborative environments, online high-throughput computational infrastructure, or high-performance cloud computing, to effectively explore quantitative structure-property relationships (144, 145).

SUMMARY POINTS

1. Artificial intelligence is revolutionizing materials discovery.
2. We have moved beyond use cases of the quantitative structure-property relationship type that simply map inputs to predictions, and we can see impacts in many stages of the discovery process.

3. Linking different types of data will enable a significant shift in the paradigm of materials discovery.
4. It is important to control, or at least understand, bias in the methods and indeed within the data themselves.
5. Through learning to see, estimate, and search with artificial intelligence, we can begin to emulate the scientific process, enabling a smart, automated materials discovery workflow.

FUTURE DIRECTIONS

1. Insights into the fundamental behaviors of materials should be sought. While most machine learning tasks are focused on the prediction of materials properties, it is equally important to use learning techniques to focus on understanding materials behaviors and performance as a result of collected properties.
2. We should increase our experience via data integration. Integration of scattered data from different sources of information is important for deeper understanding of quantitative structure-property relationships.
3. We should increase our experience via convergence approaches. Integration of virtual and physical systems will be powerful in accelerating the discovery of new materials.

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. The mention of commercial products, their source, or their use in connection with the material reported herein is not to be construed as either an actual or an implied endorsement by the National Institute of Standards and Technology, International Business Machines, or Nexight Group.

ACKNOWLEDGMENTS

C.S. thanks the Nexight Group team—Ross Brindle, Jacqueline Bowen, and Jack Holmes—for lending their support in the development of this article. Their time, effort, and contributions made this critical work possible. E.O.P.-K. and C.F. were supported by the Science and Technology Facilities Council Hartree Centre's Innovation Return on Research program, funded by the UK Department for Business, Energy and Industrial Strategy.

LITERATURE CITED

1. Natl. Sci. Technol. Counc. 2011. *Materials Genome Initiative for Global Competitiveness*. White Pap., Comm. Technol., Natl. Sci. Technol. Counc., Washington, DC. **https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/materials_genome_initiative-final.pdf**
2. Natl. Res. Counc. 2008. *Integrated Computational Materials Engineering: A Transformational Discipline for Improved Competitiveness and National Security*. Washington, DC: Natl. Acad. Press
3. Holm EA. 2019. In defense of the black box. *Science* 364:26–27
4. Zhang W, Du Y, Chen L, Peng Y, Zhou P, et al. 2014. Design of new gradient cemented carbides and hard coatings through ceramic genome. In *High Temperature Ceramic Matrix Composites 8*, ed. L Zhang, D Jiang, pp. 1–13. Hoboken, NJ: Wiley

5. Mauro J, Tandia A, Vargheese K, Mauro Y, Smedskjaer M. 2016. Accelerating the design of functional glasses through modeling. *Chem. Mater.* 28:4267–77
6. Qian C, Siler T, Ozin G. 2015. Exploring the possibilities and limitations of a nanomaterials genome. *Small* 11:64–69
7. Boyd P, Lee Y, Smit B. 2017. Computational development of the nanoporous materials genome. *Nat. Rev. Mater.* 2:17037
8. Mannodi-Kanakkithodi A, Chandrasekaran A, Kim C, Huan T, Pilania G, et al. 2018. Scoping the polymer genome: a roadmap for rational polymer dielectrics design and beyond. *Mater. Today* 21:785–96
9. Amsler M, Hegde V, Jacobsen S, Wolverton C. 2018. Exploring the high-pressure materials genome. *Phys. Rev. X* 8:041021
10. Peng B, Goodsell J, Pipes R, Yu W. 2016. Generalized free-edge stress analysis using mechanics of structure genome. *J. Appl. Mech.* 83:101013
11. Ren F, Ward L, Williams T, Laws KJ, Wolverton C, et al. 2018. Accelerated discovery of metallic glasses through iteration of machine learning and high-throughput experiments. *Sci. Adv.* 4:eaaq1566
12. Warren JA. 2018. The Materials Genome Initiative and artificial intelligence. *MRS Bull.* 43:452–57
13. Mitchell TM. 1997. *Machine Learning*. Boston: WCB McGraw-Hill
14. Todeschini R, Consonni V. 2009. *Molecular Descriptors for Chemoinformatics*, Vol. 2: *Appendices, References*. Weinheim, Ger.: Wiley-VCH. 2nd ed.
15. Young SS, Yuan F, Zhu M. 2012. Chemical descriptors are more important than learning algorithms for modelling. *Mol. Inform.* 31:707–10
16. Polishchuk PG, Kuźmin VE, Artemenko AG, Muratov EN. 2013. Universal approach for structural interpretation of QSAR/QSPR models. *Mol. Inform.* 32:843–53
17. Todeschini R, Consonni V. 2009. *Molecular Descriptors for Chemoinformatics*, Vol. 1: *Alphabetical Listing*. Weinheim, Ger.: Wiley-VCH. 2nd ed.
18. Weininger D. 1988. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* 28:31–36
19. Heller S, McNaught A, Stein S, Tchekhovskoi D, Pletnev I. 2013. InChI—the worldwide chemical structure identifier standard. *J. Cheminform.* 5:7
20. Durant JL, Leland BA, Henry DR, Nourse JG. 2002. Reoptimization of MDL keys for use in drug discovery. *J. Chem. Inf. Comput. Sci.* 42:1273–80
21. Rogers D, Hahn M. 2010. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* 50:742–54
22. Liu S, Chandereng T, Liang Y. 2018. N-gram graph, a novel molecule representation. arXiv:1906:09206 [cs.LG]
23. Behler J, Parrinello M. 2007. Generalized neural-network representation of high-dimensional potential-energy surfaces. *Phys. Rev. Lett.* 98:146401
24. Bartók AP, Kondor R, Csányi G. 2013. On representing chemical environments. *Phys. Rev. B* 87:184115
25. Rupp M, Tkatchenko A, Müller KR, von Lilienfeld OA. 2012. Fast and accurate modeling of molecular atomization energies with machine learning. *Phys. Rev. Lett.* 108:058301
26. Hansen K, Biegler F, Ramakrishnan R, Pronobis W, von Lilienfeld OA, et al. 2015. Machine learning predictions of molecular properties: accurate many-body potentials and nonlocality in chemical space. *J. Phys. Chem. Lett.* 6:2326–31
27. Huang B, von Lilienfeld OA. 2016. Communication. Understanding molecular representations in machine learning: the role of uniqueness and target similarity. *J. Chem. Phys.* 145:161102
28. Faber FA, Lindmaa A, von Lilienfeld OA, Armiento R. 2016. Machine learning energies of 2 million elpasolite (ABC_2D_6) crystals. *Phys. Rev. Lett.* 117:135502
29. Huang B, von Lilienfeld OA. 2017. The "DNA" of chemistry: scalable quantum machine learning with "amons." arXiv:1707.04146 [physics.chem-ph]
30. Faber FA, Hutchison L, Huang B, Gilmer J, Schoenholz SS, et al. 2017. Prediction errors of molecular machine learning models lower than hybrid DFT error. *J. Chem. Theory Comput.* 13:5255–64
31. Kuzminykh D, Polykovskiy D, Kadurin A, Zhebrak A, Baskov I, et al. 2018. 3D molecular representations based on the wave transform for convolutional neural networks. *Mol. Pharm.* 15:4378–85

32. Skalic M, Jiménez J, Sabbadin D, De Fabritiis G. 2019. Shape-based generative modeling for de novo drug design. *J. Chem. Inf. Model.* 59:1205–14

33. Huo H, Rupp M. 2017. Unified representation of molecules and crystals for machine learning. arXiv:1704.06439 [physics.chem-ph]

34. Eickenberg M, Exarchakis G, Hirn M, Mallat S. 2017. Solid harmonic wavelet scattering: predicting quantum molecular energy from invariant descriptors of 3D electronic densities. *Adv. Neural Inf. Process. Syst.* 30:6522–31

35. Elsken T, Metzen JH, Hutter F. 2019. Neural architecture search: a survey. *J. Mach. Learn. Res.* 20:1–21

36. Duvenaud D, Maclaurin D, Aguilera-Iparraguirre J, Gómez-Bombarelli R, Hirzel T, et al. 2015. Convolutional networks on graphs for learning molecular fingerprints. *Adv. Neural Inf. Process. Syst.* 28:2224–32

37. Kearnes S, McCloskey K, Berndl M, Pande V, Riley P. 2016. Molecular graph convolutions: moving beyond fingerprints. *J. Comput.-Aided Mol. Des.* 30:595–608

38. Gilmer J, Schoenholz SS, Riley PF, Vinyals O, Dahl GE. 2017. Neural message passing for quantum chemistry. *Proc. Mach. Learn. Res.* 70:1263–72

39. Smith JS, Isayev O, Roitberg AE. 2017. ANI-1: an extensible neural network potential with DFT accuracy at force field computational cost. *Chem. Sci.* 8:3192–203

40. 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]

41. Schütt KT, Arbabzadah F, Chmiela S, Müller KR, Tkatchenko A. 2017. Quantum-chemical insights from deep tensor neural networks. *Nat. Commun.* 8:13890

42. Schütt KT, Sauceda HE, Kindermans PJ, Tkatchenko A, Müller KR. 2018. SchNet—a deep learning architecture for molecules and materials. *J. Chem. Phys.* 148:241722

43. Ramsundar B, Kearnes S, Riley P, Webster D, Konerding D, Pande V. 2015. Massively multitask networks for drug discovery. arXiv:1502.02072 [stat.ML]

44. Ruder S. 2017. An overview of multi-task learning in deep neural networks. arXiv:1706.05098 [cs.LG]

45. Fare C, Turcani L, Pyzer-Knapp EO. 2018. Powerful, transferable representations for molecules through intelligent task selection in deep multitask networks. arXiv:1809.06334 [physics.chem-ph]

46. Flemings M. 1999. What next for departments of materials science and engineering? *Annu. Rev. Mater. Sci.* 29:1–23

47. Pollock TM, Van der Ven A. 2019. The evolving landscape for alloy design. *MRS Bull.* 44:238–46

48. Zunger A. 1980. Systematization of the stable crystal structure of all AB-type binary compounds: a pseudopotential orbital-radii approach. *Phys. Rev. B* 22:5839–72

49. Suh C, Rajan K. 2009. Data mining and informatics for crystal chemistry: establishing measurement techniques for mapping structure-property relationships. *Mater. Sci. Technol.* 25:466–71

50. Ye W, Chen C, Wang Z, Chu IH, Ong S. 2018. Deep neural networks for accurate predictions of crystal stability. *Nat. Commun.* 9:3800

51. Himanen L, Rinke P, Foster A. 2018. Materials structure genealogy and high-throughput topological classification of surfaces and 2D materials. *NPJ Comput. Mater.* 4:52

52. Graser J, Kauwe S, Sparks T. 2018. Machine learning and energy minimization approaches for crystal structure predictions: a review and new horizons. *Chem. Mater.* 30:3601–12

53. Oganov A, Pickard C, Zhu Q, Needs R. 2019. Structure prediction drives materials discovery. *Nat. Rev. Mater.* 4:331–48

54. McCue I, Stuckner J, Murayama M, Demkowicz M. 2018. Gaining new insights into nanoporous gold by mining and analysis of published images. *Sci. Rep.* 8:6761

55. Du P, Zebrowski A, Zola J, Ganapathysubramanian B, Wodo O. 2018. Microstructure design using graphs. *NPJ Comput. Mater.* 4:50

56. Yang Z, Li X, Brinson L, Choudhary A, Chen W, Agrawal A. 2018. Microstructural materials design via deep adversarial learning methodology. *J. Mech. Des.* 140:111416

57. Ziatdinov M, Dyck O, Maksov A, Li X, Sang X, et al. 2017. Deep learning of atomically resolved scanning transmission electron microscopy images: chemical identification and tracking local transformations. *ACS Nano* 11:12742–52

58. Madsen J, Liu P, Kling J, Wagner J, Hansen T, et al. 2018. A deep learning approach to identify local structures in atomic-resolution transmission electron microscopy images. *Adv. Theory Simul.* 1:1800037

59. Vasudevan R, Laanait N, Ferragut E, Wang K, Geohegan D, et al. 2018. Mapping mesoscopic phase evolution during E-beam induced transformations via deep learning of atomically resolved images. *NPJ Comput. Mater.* 4:30

60. Xu W, LeBeau J. 2018. A deep convolutional neural network to analyze position averaged convergent beam electron diffraction patterns. *Ultramicroscopy* 188:59–69

61. Jensen Z, Kim E, Kwon S, Gani T, Román-Leshkov Y, et al. 2019. A machine learning approach to zeolite synthesis enabled by automatic literature data extraction. *ACS Cent. Sci.* 5:892–99

62. Kim E, Huang K, Tomala A, Matthews S, Strubell E, et al. 2017. Machine-learned and codified synthesis parameters of oxide materials. *Sci. Data* 4:170127

63. Kim E, Jensen Z, van Grootel A, Huang K, Staib M, et al. 2019. Inorganic materials synthesis planning with literature-trained neural networks. arXiv:1901.00032 [cond-mat.mtrl.sci]

64. Court C, Cole J. 2018. Auto-generated materials database of Curie and Néel temperatures via semi-supervised relationship extraction. *Sci. Data* 5:180111

65. Young S, Maksov A, Ziatdinov M, Cao Y, Burch M, et al. 2018. Data mining for better material synthesis: the case of pulsed laser deposition of complex oxides. *J. Appl. Phys.* 123:115303

66. Onishi T, Kadohira T, Watanage I. 2018. Relation extraction with weakly supervised learning based on process-structure-property-performance reciprocity. *J. Sci. Technol. Adv. Mater.* 19:649–59

67. Kim E, Huang K, Kononova O, Ceder G, Olivetti E. 2019. Distilling a materials synthesis ontology. *Matter Opin.* 1:8–12

68. Schwaller P, Gaudin T, Lányi D, Bekas C, Laino T. 2018. "Found in translation": predicting outcomes of complex organic chemistry reactions using neural sequence-to-sequence models. *Chem. Sci.* 9:6091–98

69. Raccuglia P, Elbert K, Adler P, Falk C, Wenny M, et al. 2016. Machine-learning-assisted materials discovery using failed experiments. *Nature* 533:73–76

70. Dimitrov T, Kreisbeck C, Becker J, Aspuru-Guzik A, Saikin S. 2019. Autonomous molecular design: then and now. *ACS Appl. Mater. Interfaces* 11:24825–36

71. Boyce B, Uchic M. 2019. Progress toward autonomous experimental systems for alloy development. *MRS Bull.* 44:273–80

72. Toher C, Oses C, Curtarolo S. 2019. Automated computation of materials properties. In *Materials Informatics: Methods, Tools and Applications*, ed. O Isayev, A Tropsha, S Curtarolo, pp. 181–222. Weinheim, Ger.: Wiley-VCH

73. Smith J, Nebgen B, Lubbers N, Isayev O, Roitberg A. 2018. Less is more: sampling chemical space with active learning. *J. Chem. Phys.* 148:241733

74. Masubuchi S, Morimoto M, Morikawa S, Onodera M, Asakawa Y, et al. 2018. Autonomous robotic searching and assembly of two-dimensional crystals to build van der Waals superlattices. *Nat. Commun.* 9:1413

75. Rashidi M, Croshaw J, Mastel K, Tamura M, Hosseinzadeh H, Wolkow RA. 2019. Autonomous atomic scale manufacturing through machine learning. arXiv:1902.08818 [cond-mat.mtrl-sci]

76. Tabor D, Roch L, Saikin S, Kreisbeck C, Sheberla D, et al. 2018. Accelerating the discovery of materials for clean energy in the era of smart automation. *Nat. Rev. Mater.* 3:5–20

77. Zakutayev A, Wunder N, Schwarting M, Perkins J, White R, et al. 2018. An open experimental database for exploring inorganic materials. *Sci. Data* 5:180053

78. Li X, Zhang Y, Zhao H, Burkhart C, Brinson L, Chen W. 2018. A transfer learning approach for microstructure reconstruction and structure-property predictions. *Sci. Rep.* 8:13461

79. Zhang Y, Ling C. 2018. A strategy to apply machine learning to small datasets in materials science. *NPJ Comput. Mater.* 4:25

80. Lubbers N, Lookman T, Barros K. 2017. Inferring low-dimensional microstructure representations using convolutional neural networks. *Phys. Rev.* 96:052111

81. Butler K, Davies D, Cartwright H, Isayev O, Walsh A. 2018. Machine learning for molecular and materials science. *Nature* 559:547–55

82. Nyshadham C, Rupp M, Bekker B, Shapeev A, Mueller T, et al. 2019. Machine-learned multi-system surrogate models for materials prediction. *NPJ Comput. Mater.* 5:51

83. Kumar N, Rajagopalan P, Pankajakshan P, Bhattacharyya A, Sanyal S, et al. 2019. Machine learning constrained with dimensional analysis and scaling laws: simple, transferable, and interpretable models of materials from small datasets. *Chem. Mater.* 31:314–21

84. Sparks T, Kauwe S, Welker T. 2018. Extracting knowledge from DFT: experimental band gap predictions through ensemble learning. ChemRxiv 7236029. **https://doi.org/10.26434/chemrxiv.7236029.v1**

85. Paul A, Jha D, Al-Bahrani R, Liao W, Choudhary A, Agrawal A. 2019. Transfer learning using ensemble neural networks for organic solar cell screening. arXiv:1903.03178 [cs.LG]

86. Jennings P, Lysgaard S, Hummelshøj J, Vegge T, Bligaard T. 2019. Genetic algorithms for computational materials discovery accelerated by machine learning. *NPJ Comput. Mater.* 5:46

87. Ouyang R, Ahmetcik E, Carbogno C, Scheffler M, Ghiringhell L. 2019. Simultaneous learning of several materials properties from incomplete databases with multi-task SISSO. *J. Phys. Mater.* 2:024002

88. Suh C, Sieg S, Heying M, Oliver J, Maier W, Rajan K. 2009. Visualization of high-dimensional combinatorial catalysis data. *J. Comb. Chem.* 11:385–92

89. Rickman J. 2018. Data analytics and parallel-coordinate materials property chart. *NPJ Comput. Mater.* 4:5

90. Mueller T, Kusne A, Ramprasad R. 2016. Machine learning in materials science: recent progress and emerging applications. *Rev. Comput. Chem.* 29:186–73

91. Suh C, Biagioni D, Glynn S, Scharf J, Contreras M, et al. 2011. Exploring high-dimensional data space: identifying optimal process conditions in photovoltaics. In *2011 37th IEEE Photovoltaic Specialists Conference*, pp. 762–67. New York: IEEE

92. Li Y. 2006. Predicting materials properties and behavior using classification and regression trees. *Mat. Sci. Eng. A* 433:261–68

93. Liu R, Kumar A, Chen Z, Agrawal A, Sundararaghavan V, Choudhary A. 2015. A predictive machine learning approach for microstructure optimization and materials design. *Sci. Rep.* 5:11551

94. Stanev V, Oses C, Kusne AG, Rodriguez E, Paglione J, et al. 2018. Machine learning modeling of superconducting critical temperature. *NPJ Comput. Mater.* 4:29

95. Dieb T, Ju S, Shiomi J, Tsuda K. 2019. Monte Carlo tree search for materials design and discovery. *MRS Commun.* 9:532–36

96. Joshi R, Eickholt J, Li L, Fornari M, Barone V, Peralta J. 2019. Machine learning the voltage of electrode materials in metal-ion batteries. *ACS Appl. Mater. Interfaces* 11:18494–503

97. Nouira A, Sokolovska N, Crivello J. 2018. CrystalGAN: learning to discover crystallographic structures with generative adversarial networks. arXiv:1810.11203v3 [cs.LG]

98. Jha D, Ward L, Paul A, Liao W, Choudhary A, et al. 2018. ElemNet: deep learning the chemistry of materials from only elemental composition. *Sci. Rep.* 8:17593

99. Frazier PI, Wang J. 2016. Bayesian optimization for materials design. In *Information Science for Materials Discovery and Design*, ed. T Lookman, F Alexander, K Rajan, pp. 45–75. Cham, Switz.: Springer

100. Gubernatis J, Lookman T. 2018. Machine learning in materials design and discovery: examples from the present and suggestions for the future. *Phys. Rev. Mater.* 2:120301

101. Li C, de Celis Leal DR, Rana S, Gupta S, Sutti A, et al. 2017. Rapid Bayesian optimisation for synthesis of short polymer fiber materials. *Sci. Rep.* 7:5683

102. Ling J, Hutchinson M, Antono E, Paradiso S, Meredig B. 2017. High-dimensional materials and process optimization using data-driven experimental design with well-calibrated uncertainty estimates. *Integr. Mater. Manuf. Innov.* 6:207–17

103. Ueno T, Rhone TD, Hou Z, Mizoguchi T, Tsuda K. 2016. COMBO: an efficient Bayesian optimization library for materials science. *Mater. Discov.* 4:18–21

104. Groves M, Pyzer-Knapp EO. 2018. Efficient and scalable batch Bayesian optimization using K-means. arXiv:1806.01159 [stat.ML]

105. Hernández-Lobato JM, Requeima J, Pyzer-Knapp EO, Aspuru-Guzik A. 2017. Parallel and distributed Thompson sampling for large-scale accelerated exploration of chemical space. *Proc. Mach. Learn. Res.* 70:1470–79

106. Talapatra A, Boluki S, Duong T, Qian X, Dougherty E, Arróyave R. 2018. Autonomous efficient experiment design for materials discovery with Bayesian model averaging. *Phys. Rev. Mater.* 2:113803

107. Silver D, Hubert T, Schrittwieser J, Antonoglou I, Lai M, et al. 2018. A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. *Science* 362:1140–44

108. Senior AW, Evans R, Jumper J, Kirkpatrick J, Sifre L, et al. 2020. Improved protein structure prediction using potentials from deep learning. *Nature* 577:706–10

109. Popova M, Isayev O, Tropsha A. 2018. Deep reinforcement learning for de novo drug design. *Sci. Adv.* 4:eaap7885

110. Olivecrona M, Blaschke T, Engkvist O, Chen H. 2017. Molecular de-novo design through deep reinforcement learning. *J. Cheminform.* 9:48

111. Zhou Z, Li X, Zare RN. 2017. Optimizing chemical reactions with deep reinforcement learning. *ACS Cent. Sci.* 3:1337–44

112. Sanchez-Lengeling B, Aspuru-Guzik A. 2018. Inverse molecular design using machine learning: generative models for matter engineering. *Science* 361:360–65

113. Xu Y, Lin K, Wang S, Wang L, Cai C, et al. 2019. Deep learning for molecular generation. *Future Med. Chem.* 11:567–97

114. Doersch C. 2016. Tutorial on variational autoencoders. arXiv:1606.05908 [stat.ML]

115. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, et al. 2014. Generative adversarial nets. *Adv. Neural Inf. Process. Syst.* 27:2672–80

116. Jørgensen PB, Schmidt MN, Winther O. 2018. Deep generative models for molecular science. *Mol. Inform.* 37:1700133

117. 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.* 4:268–76

118. Blaschke T, Olivecrona M, Engkvist O, Bajorath J, Chen H. 2018. Application of generative autoencoder in de novo molecular design. *Mol. Inform.* 37:1700123

119. Kang S, Cho K. 2019. Conditional molecular design with deep generative models. *J. Chem. Inf. Model.* 59:43–52

120. Kadurin A, Aliper A, Kazennov A, Mamoshina P, Vanhaelen Q, et al. 2016. The cornucopia of meaningful leads: applying deep adversarial autoencoders for new molecule development in oncology. *Oncotarget* 8:10883–90

121. Samanta B, De A, Jana G, Chattaraj PK, Ganguly N, Gomez-Rodriguez M. 2018. NeVAE: a deep generative model for molecular graphs. arXiv:1802.05283 [cs.LG]

122. Jin W, Barzilay R, Jaakkola T. 2018. Junction tree variational autoencoder for molecular graph generation. arXiv:1802.04364 [cs.LG]

123. Winter R, Montanari F, Noé F, Clevert DA. 2019. Learning continuous and data-driven molecular descriptors by translating equivalent chemical representations. *Chem. Sci.* 10:1692–701

124. De Cao N, Kipf T. 2018. MolGAN: an implicit generative model for small molecular graphs. arXiv:1805.11973 [stat.ML]

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

126. Berardo E, Turcani L, Miklitz M, Jelfs KE. 2018. An evolutionary algorithm for the discovery of porous organic cages. *Chem. Sci.* 9:8513–27

127. Lysgaard S, Landis DD, Bligaard T, Vegge T. 2014. Genetic algorithm procreation operators for alloy nanoparticle catalysts. *Top. Catal.* 57:33–39

128. Patra TK, Meenakshisundaram V, Hung JH, Simmons DS. 2017. Neural-network-biased genetic algorithms for materials design: evolutionary algorithms that learn. *ACS Comb. Sci.* 19:96–107

129. Rickman J, Lookman T, Kalinin S. 2019. Materials informatics: from the atomic-level to the continuum. *Acta Mater.* 168:473–510

130. Maier W. 2019. Early years of high-throughput experimentation and combinatorial approaches in catalysis and materials science. *ACS Comb. Sci.* 21:437–44

131. Bock F, Aydin R, Cyron C, Huber N, Kalidindi S, Klusemann B. 2019. A review of the application of machine learning and data mining approaches in continuum materials mechanics. *Front. Mater.* 6:110

132. Elton DC, Boukouvala Z, Fuge MD, Chung PW. 2019. Deep learning for molecular design—a review of the state of the art. *Mol. Syst. Des. Eng.* 4:828–49

133. Agrawal A, Choudhary A. 2019. Deep materials informatics: applications of deep learning in materials science. *MRS Commun.* 9:779–92

134. Mater A, Coote M. 2019. Deep learning in chemistry. *J. Chem. Inf. Model.* 59:2545–59

135. Alberi K, Nardelli M, Zakutayev A, Mitas L, Curtarolo S, et al. 2018. The 2019 materials by design roadmap. *J. Phys. D* 52:013001

136. Li L, Yang Y, Zhang D, Ye ZG, Jesse S, et al. 2018. Machine learning-enabled identification of material phase transitions based on experimental data: exploring collective dynamics in ferroelectric relaxors. *Sci. Adv.* 4:8672

137. Rovinelli A, Sangid M, Proudhon H, Ludwig W. 2018. Using machine learning and a data-driven approach to identify the small fatigue crack driving force in polycrystalline materials. *NPJ Comput. Mater.* 4:35

138. Schwarzer M, Rogan B, Ruan Y, Song Z, Lee D, et al. 2019. Learning to fail: predicting fracture evolution in brittle materials using recurrent graph convolutional neural networks. *Comput. Mater. Sci.* 162:322–32

139. Zunger A. 2018. Inverse design in search of materials with target functionalities. *Nat. Rev. Chem.* 2:0121

140. Karcher S, Willighagen E, Rumble J, Ehrhart F, Evelo C, et al. 2018. Integration among databases and data sets to support productive nanotechnology: challenges and recommendations. *Nanoimpact* 9:85–101

141. Brown K, Spivak D, Wisnesky R. 2019. Categorical data integration for computational science. *Comput. Mater. Sci.* 164:127–32

142. Zhao H, Wang Y, Lin A, Hu B, Yan R, et al. 2018. NanoMine schema: an extensible data representation for polymer nanocomposites. *APL Mater.* 6:111108

143. Natl. Inst. Stand. Techol. 2017. High-Throughput Experimental Materials Collaboratory. *National Institute of Standards and Technology*. **https://www.nist.gov/programs-projects/high-throughput-experimental-materials-collaboratory**

144. Jacobsen M, Fourman J, Porter K, Wirrig E, Benedict M, Ward C. 2016. Creating an integrated collaborative environment for materials research. *Int. Mater. Manuf. Innov.* 5:232–44

145. Yang X, Wang Z, Zhao X, Song J, Yu C, et al. 2018. MatCloud, a high-throughput computational materials infrastructure: present, future visions, and challenges. *Chin. Phys. B* 27:110301