### Title

The Role of Machine Learning in the Understanding and Design of Materials




### Journal

Journal of the American Chemical Society, 142(48)

### ISSN

0002-7863



### Authors

Moosavi, Seyed Mohamad
Jablonka, Kevin Maik
Smit, Berend



### Publication Date

2020-12-02

### DOI

10.1021/jacs.0c09105

Peer reviewed

JACS
JOURNAL OF THE AMERICAN CHEMICAL SOCIETY

pubs.acs.org/JACS

Perspective

# The Role of Machine Learning in the Understanding and Design of Materials

Seyed Mohamad Moosavi, Kevin Maik Jablonka, and Berend Smit*

Cite This: J. Am. Chem. Soc. 2020, 142, 20273−20287

Read Online

ACCESS | Metrics & More | Article Recommendations

**ABSTRACT:** Developing algorithmic approaches for the rational design and discovery of materials can enable us to systematically find novel materials, which can have huge technological and social impact. However, such rational design requires a holistic perspective over the full multistage design process, which involves exploring immense materials spaces, their properties, and process design and engineering as well as a techno-economic assessment. The complexity of exploring all of these options using conventional scientific approaches seems intractable. Instead, novel tools from the field of machine learning can potentially solve some of our challenges on the way to rational materials design. Here we review some of the chief advancements of these methods and their applications in rational materials design, followed by a discussion on some of the main challenges and opportunities we currently face together with our perspective on the future of rational materials design and discovery.

## INTRODUCTION

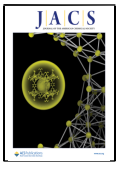
Over the last few decades, materials chemistry research has shifted toward more rational design. There are now many examples, such as metal−organic frameworks (MOFs),[1] polymers,[2] and DNA nanostructures,[3] in which we have such control over the chemistry that we can move away from the traditional trial and error. If we think about rational design, we quickly realize that we are dealing with large numbers: a large number of materials, a large number of applications, and a large number of options. Here we argue that the conventional scientific approach for materials design based on fundamental laws, computational modeling, and experimentation is challenged when encountering these large numbers. Therefore, we are now developing new tools that work on the basis of large data, which might allow us to overcome some of these challenges. The development of modern big-data science methodologies, often called machine learning, will allow us to pursue our aim of understanding and designing of materials in a new way.

Machine learning models try to use the underlying patterns and relationships in data to make new predictions. A classical example is image recognition. We know that there exists a complex relationship between the pixels of an image and their labels (e.g., dog or cat). However, trying to find this complex relationship by writing an equation that takes the images as an input and outputs whether it is an image of a cat or dog is not only extremely challenging but also does not lead to a scalable solution, as we need to develop a new equation for every label. Instead, machine learning methods try to infer this mapping between pixels and labels from observing many examples. The underlying idea of machine learning is to use a training set of many images of all different types of cats and dogs to infer the underlying patterns in order to predict whether an unseen image shows a dog or cat. Similarly, if presented with enough examples, machine learning methods can extract relationships between chemical systems and their properties and performances that would otherwise require solving equations that are too complex.

In this Perspective, we do not deal with the question of "how" to do machine learning. We refer the readers to the comprehensive review articles and books on the topic of how to implement a machine learning project. In specific, excellent resources are available for overviews of the fundamentals of machine learning[4−6] and deep learning[7−10] and their applications in materials design,[11] chemical synthesis,[12−14] and molecular simulation[15] for different classes of materials or applications, e.g., battery materials[16] and nanoporous materials.[17]

Instead, we focus on challenges in the rational design and discovery of materials and how we can use machine learning to address them. Here we use as an example a topic of our research: materials for energy-related applications. We aim to systematically design or discover materials that lead to the optimal energy efficiency for any given application. Typically, we are given some external parameters or constraints that define the problem, for example, the source and sink for carbon capture, the operational pressure for methane and hydrogen storage, or the light spectral distribution and irradiance for solar cells. To systematically find the optimal solution, we typically follow a multistage design process (Figure 1) that starts with identifying the materials search space, followed by predicting or measuring materials properties and evaluating or

Received: August 24, 2020
Published: November 10, 2020









https://dx.doi.org/10.1021/jacs.0c09105
J. Am. Chem. Soc. 2020, 142, 20273−20287

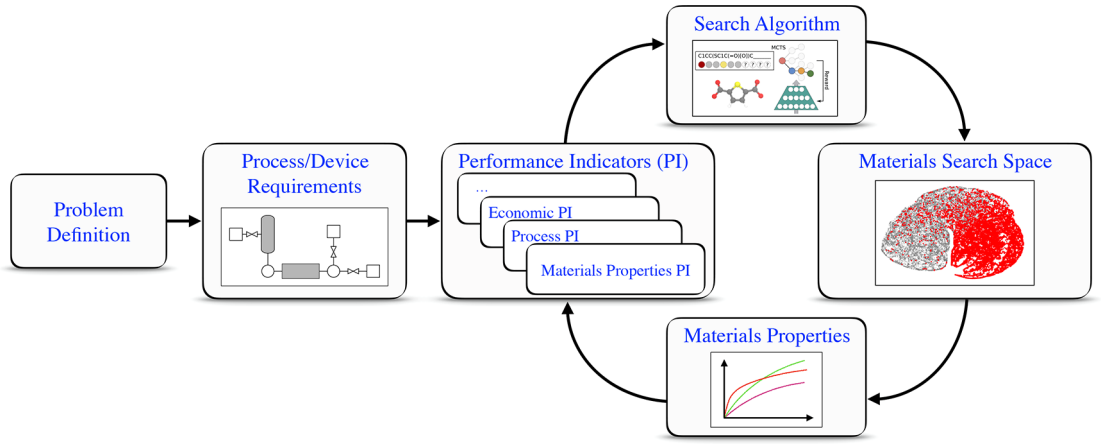



**Figure 1.** Algorithmic approach for holistic rational material design. We start with a problem for which we have conceptualized a solution (e.g., an adsorption process or device) with some requirements. For this concept, we try to find the best materials in the materials search space to maximize the performance indicators. Because of the complexity of performance evaluation, we usually select surrogate parameters (e.g., material properties) that we hope are reasonable surrogates for the performance in the real world.

testing their performance for the target application. We then aim to search the design space using the knowledge we obtain from our observations to find the best-performing setup, considering materials and process. Practically, solving "exactly" the governing equations of the physical laws for all of these stages is far too complex; therefore, we need fundamentally different approaches to become able to deliver solutions for the technological challenges of our time.

The use of machine learning in materials design and discovery is a natural consequence of the problem we try to solve: finding needles in a haystack of materials for any given application. The complexity of this search requires extracting patterns in the form of design rules and/or fast methods for exploring these vast spaces for optimization of processes and materials properties. One can perceive that here machine learning is not the object of the research but the tool for doing better science. Doubtless, our research on materials has progressed so much that we can produce large amounts of high-quality data, which make the field of materials science ready for abrupt growth if correct tools are used.

This new approach in the design and discovery of materials brings us to a new era of materials design and discovery that intrinsically has new opportunities and challenges. We can now effortlessly solve some of the classic problems, and as a consequence, we can focus on more challenging and sometimes new problems. In this Perspective, we review some of the major advancements in materials understanding and design that have been made possible by machine learning. We argue that current progress in machine learning for materials science has been stunning but that breakthroughs are still to come. Hence, we discuss some of the challenges to overcome and our vision for the future direction of research on this topic.

## MATERIALS DESIGN AND DISCOVERY

**Chemical Space and Databases.** Often, finding an optimal material for a given application is presented as inverse design. One starts with the application and systematically narrows down the options to find the holy grail of materials design, the optimal material to synthesize. In the Introduction, we argued that such an approach will be fundamentally limited if we always rely on solving equations, either because of their complexity or our limited understanding of how to simulate certain phenomena. Therefore, we explore here the other extreme, which is using machine learning to infer solutions on the basis of many observations and guide us through the design process. Since this approach is based on having many observations, our starting point is collecting a large amount of data, which can be observations from our materials search space, training data for our models, or simply the data with which we compare our findings. Typically, such large amounts of data are accessible from large databases of materials and properties.

In the past decades, crystal structures of synthesized compounds have been collected in several databases, including the Inorganic Crystal Structure Database (ICSD),[18] the Crystallographic Open Database (COD),[19] the International Centre for Diffraction Data (ICDD),[20] and the Cambridge Structural Database (CSD).[21] In parallel, significant progress has been made in the development of databases of hypothetical materials, i.e., structures generated in silico, making it possible to study materials even before they have ever been synthesized. For instance, in the field of nanoporous materials, this effort has led to the development several hypothetical databases[22] of metal−organic frameworks (MOFs),[23−27] covalent organic frameworks (COFs),[28,29] and zeolites.[30,31] These databases altogether contain millions of chemical compounds.

These databases have constituted the starting point for computational high-throughput screening studies. The computational predictions of these studies have been compiled in repositories and databases that are mainly focused on managing materials properties data. For instance, the Materials Project,[32] the Pauling File,[33] Novel Materials Discovery (NOMAD),[34] and Materials Cloud[35] contain computational materials data. Additionally, databases like the ones from the National Institute of Standards and Technology (NIST) store experimental properties of materials, including adsorption properties of porous materials[36] and thermophysical properties of alloy materials.[37,38]

The size of these databases sounds enormous, yet they represent only a fraction of all possible chemical structures. Since we rely on observations to infer solutions, a primary step is to make sure that we have sufficient diversity in our observations in a database. The underlying fundamental


https://dx.doi.org/10.1021/jacs.0c09105
J. Am. Chem. Soc. 2020, 142, 20273−20287

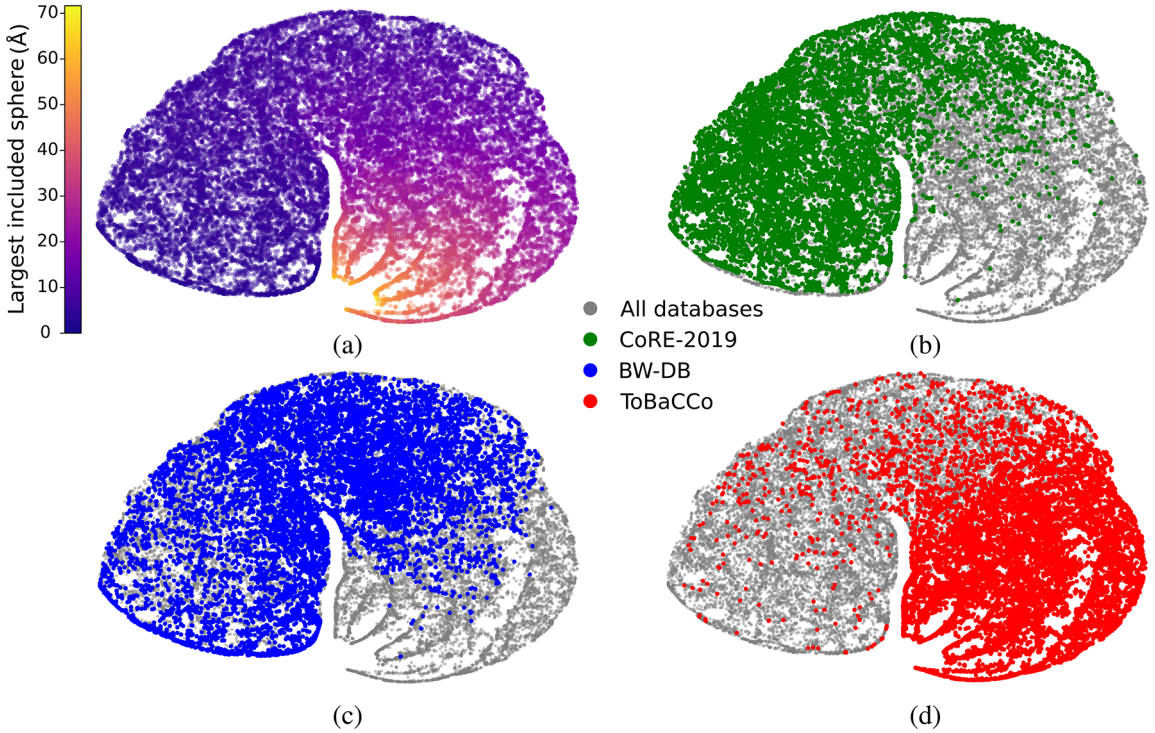



**Figure 2.** Maps of the pore geometry of MOFs. The descriptors of pore geometry of MOFs were mapped to two dimensions using the *t*-distributed stochastic neighbor embedding (t-SNE) method. The *t*-SNE method preserves local similarities such that materials similar to each other are located close to each other in the 2D map. Each dot shows one material, and the structures from different databases are overlaid on top of the collective map structures from all databases. The experimental structures are from the CoRE-2019 database,[44] and the hypothetical structures are from the ToBaCCo[25] and BW-DB[24] databases. From ref 41. CC BY 4.0.

question is how to make sense of the structure of a database without knowing the relation between the material and the performance. In machine learning this question is studied using unsupervised learning, which deals with unlabeled data to infer patterns, for example, classifying materials into different clusters or detecting outliers in a database. We can use unsupervised learning to visualize the current chemical space and analyze the underlying distribution of databases.[39,40]

To illustrate this, we can look at an example of unsupervised learning on MOF databases, several of which are often used as the starting point for materials discovery via high-throughput computational screenings or machine learning. For this reason, it is important to understand how well those databases cover the chemical space of MOFs and how redundant they are.[41] The first step is to map MOF structures onto descriptors to quantify their similarities.[17,42,43] Moosavi et al.[41] used an approach that closely follows the MOF chemistry, in which a MOF is described by four sets of descriptors, for the metal nodes, linkers, functional groups, and pore geometry. The idea is that chemically similar MOFs have similar descriptors, which allows us to quantify this similarity as a distance in descriptor space. One has to note that similarity depends on the application: If we are interested in gas separation, all nonporous MOFs are useless, and hence, for that application they are all the same. However, if we look at optical properties, pore shape is most likely not very relevant.

For an application for which porosity is important, Figure 2 makes a comparison of how the characteristics of pore size and shape are covered by the different databases. While one would like to have a database that has representative samples of all possible geometries, we note a clear difference in the distributions of the geometric properties of the databases. For example, experimental MOFs (CoRE-2019 database[44]) are mainly in the small-pore region, and in silico ones (ToBaCCo database[25]) are mainly in the large-pore region. Indeed, if for a particular application a specific type of pores is desired, the chance of discovering such a material is different in each database. Similar analysis of the chemistry of materials showed a lack of diversity in the metal chemistry of the hypothetical databases. Notably, a lack of diversity can lead to wrong conclusions. For example, Moosavi et al. showed that the importance of metals for carbon capture was underrated in the past studies because of the lack of metal diversity in the analyzed databases. In addition, once we have carried out such an analysis, we can also see whether a new material has a well-known chemistry or pore geometry or whether this material is completely new. As we are dealing with over 90 000 experimental structures, answering this question without such a big-data approach to quantify the similarities of materials is difficult.[41,45−47]

At this point, it is clear that a systematic data management plan is inevitable. Such a plan must cover the full spectrum of data management and curation from discovery to integration and cleaning of data. Interestingly, we can use machine learning methods for data management and curation. For example, we lack tabulated data for many interesting applications and properties. While large amounts of data and scientific knowledge are available through the literature, the challenge is to discover and transform such raw, unstructured data embedded in text into contextualized and structured data. In the context of data discovery and mining, machine learning methods from natural language processing (NLP) can help us to extract data from the literature. For example, to address the lack of data for material synthesis, Kim et al.[48,49] performed


https://dx.doi.org/10.1021/jacs.0c09105
J. Am. Chem. Soc. 2020, 142, 20273−20287

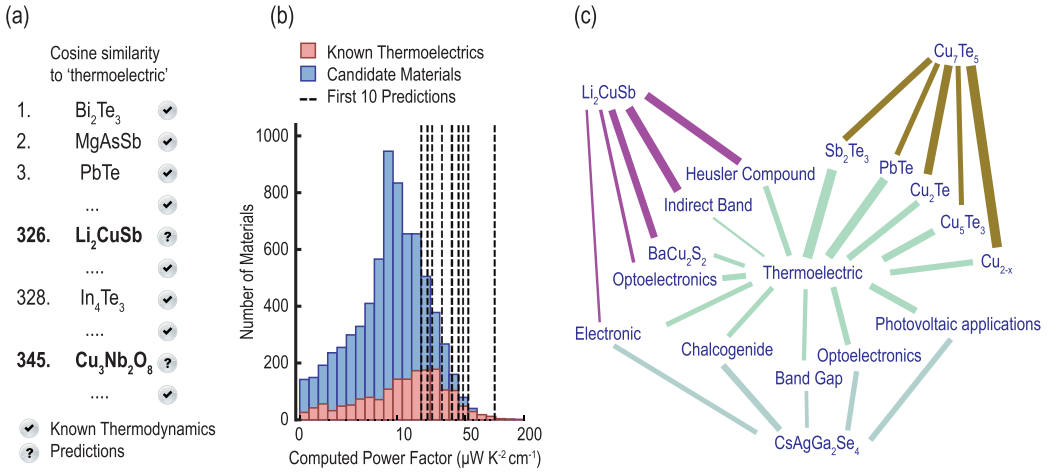



**Figure 3.** Prediction of new materials for thermoelectric applications using data mining of the literature. (a) Materials that are found close to the word "thermoelectric" in the word-embedding space. (b) The power factors of the materials were computed using density functional theory, resulting in the discovery of many new potential materials for the thermoelectric applications. (c) Connecting words between the newly discovered materials and the word "thermoelectric". The figure was redrawn based on ref 50.

text mining of more than 640 000 articles and provided a data set of synthesis parameters for 30 different oxide materials in a structured data format.

An outstanding example is a recent work showing that a method called word embedding can encode knowledge from past publications.[50] Word embeddings are representations of words as high-dimensional vectors such that they preserve relationships between words. For example, the distance between similar words (e.g., "cathode" and "battery") will be smaller in the word embedding space than the distance between dissimilar words (e.g., "ascorbic acid" and "battery"). Tshitoyan et al.[50] analyzed 3.3 million abstracts of materials-science-related articles, containing around 500 000 words, to develop a word embedding that preserves word appearance in the context proximity of the words. Remarkably, this word embedding can capture materials science concepts such as the periodic table and structure−property relationships. For example, they used the word embedding to make predictions of new thermoelectric materials (Figure 3). To find potential materials for thermoelectric applications, they investigated the proximity of the word "thermoelectric" in the word embedding space. A density functional theory prediction of the properties of the materials that were found in this area is shown in Figure 3. The word embedding not only recovered known thermoelectric materials but also discovered several new promising candidates. Interestingly, similar to chemists, the model used common chemical knowledge and intuition, such as similarities in crystal structure or applications, or phrases that describe materials properties for the predictions (see Figure 3c for a depiction of how three of the new potential thermoelectric materials are connected to the word "thermoelectric"). Indeed, dealing with millions of articles to develop such a comprehensive view over the chemical literature is a difficult task to address without machine learning.

Besides data discovery, data curation and cleaning can potentially benefit from machine learning. In specific, we can exploit the statistical nature of machine learning methods to clean the input data itself. Since machine learning models infer the underlying pattern and relationships from many examples, we can use them to identify anomalous cases, i.e., suspicious data points that are different from the majority of similar data points. In structural and materials properties databases, various kinds of errors might occur, such as wrong units for properties, spelling mistakes, data transfer and storage issues, or duplicate structures. An illustrative example is a recent work on the oxidation states of MOFs. The oxidation states of metal centers are determined and reported by chemists for the materials in the CSD. Jablonka et al.[51] developed machine learning models trained on this collection of knowledge that can predict the oxidation states with high accuracy. Coupling uncertainty metrics with these predictions, they were able to identify many incorrect assignments in the CSD. Therefore, their model could be used to flag potential mistakes in a large database such as the CSD with more than a million entries.

A complementary and effective approach is to perform quality control of data at the early stage of data generation. Often the production of big data involves large-scale execution of computational or experimental workflows, for which we would like to have careful control over the quality of data generation as well as resource management to avoid spending valuable resources on fallacious results. Indeed, manual inspection is intractable in these cases, and automation is needed. Using machine learning to control the data generation process is a promising choice in this area of research. An excellent example is the control of time-consuming first-principles calculations on open-shell transition metal complexes.[52] These calculations can frequently fail; for instance, the structure might fall apart during the geometry optimization. To aid automatic detection of these cases, a machine learning classifier model was used to predict simulation outcomes on the sole basis of the chemical composition. Moreover, a complementary classifier model was used to monitor the trajectory/convergence of the calculations, aborting those that had a high chance of failure. Using such models for autonomous job control can avoid generating data that later might be hard to classify into valid and invalid results, enhancing the quality of data generation.

**From Structure to Properties.** The next step is to be able to predict structural properties reliably and with sufficient accuracy. In principle, we can use molecular simulation and quantum mechanics to predict material properties. However, these techniques are limited to an accuracy−efficiency trade-off


https://dx.doi.org/10.1021/jacs.0c09105
J. Am. Chem. Soc. 2020, 142, 20273−20287

and might become computationally prohibitive depending on the system size, the time scales of the physical phenomena of interest, and the number of systems to be investigated. Moreover, some properties like synthesizability are so complex that we still do not have methods to predict them using computer simulations. Machine learning methods hold the promise to shift the paradigm of accuracy and efficiency, enabling exploration of large databases with high accuracy. Broadly speaking, machine learning is used in two main ways in this context: to directly map structures to their properties or to facilitate the development of new modeling methodologies.

Indeed, in recent years several materials properties have been predicted by machine learning methods. Examples include gas adsorption,[41,53−56] catalytic,[42,57−59] thermal,[60,61] thermoelectric,[62,63] bulk mechanical,[64−68] and optical and electronic[67,69,70] properties.

In principle, not solving the complex equations and inferring solutions only from observing many examples allows us to tackle problems that even state-of-the-art theory is limited to answer. In particular, finding solutions for fuzzy problems such as materials synthesis, synthesizability, and oxidation state, for which we do not have a reliable theory, are the areas of research in which data-driven methods can play a significant role. Here, machine learning gives us extremely flexible and elaborate empirical models that can fit the knowledge of individuals or experimental observations and turn them into powerful tools. Interestingly, this flexibility does not necessarily mean that we cannot extract physical insights from these models; it is used only to circumvent the limitations of conventional analytical equations that sometimes are not complex enough to fully capture the behavior of chemical systems. For example, in the case of oxidation states, empirical models that use pairwise distances between atoms to describe local geometries (e.g., the bond valence sum) are not sufficiently elaborate to capture subtle geometric dissimilarities.[51] Moreover, using machine learning can even help us to develop new theories and extract physical insights from the model.[71] For example, Cranmer et al.[72] proposed an approach with which symbolic equations can be derived from a neural network. They used this technique to find a new equation that describes the concentration of dark matter, but one can envision that a similar approach could reveal design rules for materials.

One important area of research for machine learning is to formulate new modeling methods for quantum and statistical mechanics problems. Machine learning approaches for molecular simulation are emerging to solve complex and time-consuming calculations that we typically encounter in modeling of chemical systems. These methods have already had a significant impact on the way that we compute configuration energies and forces and simulate thermodynamic,[73,74] kinetic,[75] electronic,[76] and excited-state[77,78] properties and phenomena.

One of the most significant and earliest applications of machine learning in this area is the development of high-dimensional neural network potentials to extract the potential energy surface of chemical systems from quantum mechanical calculations.[79−81] The underlying assumption here is that the potential energy can be decomposed into a sum of contributions of local environments. Hence, a machine learning model that is trained to map these local atom-centered environments to an energy can be used as a "force field" for simulating chemical systems. Behler and Parrinello[79] introduced a symmetry function formalism that is by design differentiable and encodes the required physical invariances, i.e., the energy of a system is invariant with respect to translation, rotation, and permutation of atoms. Another seminal approach is the Gaussian Approximation Potential (GAP) formalism based on the smooth overlap of atomic positions (SOAP) representation of an atomic environment.[82,83] These potentials provide quantum-mechanical accuracy with the cost of analytical force fields, allowing accurate simulation of large systems on long time scales. Recently, attempts to extend them to different elements of the periodic table have been carried out,[84] and several classes of materials have been successfully modeled using these frameworks.[61,85−87]

One of the main bottlenecks in statistical mechanics is the simulation of rare events: events that take place on a time scale that is short for experiments but is extremely long for a simulation. To simulate rare events in complex systems, which possess potential energy surfaces with multiple minima separated by large energy barriers, it is a challenge to adequately sample the configuration space to reach good statistics. This is the case because the simulation might get trapped in metastable states. For this reason, simulation of these systems requires advanced sampling techniques such as umbrella sampling or replica exchange,[88−91] which try to push the system to move from one minimum configuration to another. In a remarkable recent development,[92] a machine learning model, i.e., an invertible neural network model, was used to map the complex and hard-to-sample configurational space of a chemical system to a distribution that is easy to sample (Figure 4). Such a machine learning model can generate unbiased equilibrium samples, following the Boltzmann distribution, in one shot. These machine learning models, which were named Boltzmann generators, are illustrative examples of the kind of *new* science that we can do using machine learning that we could not do otherwise. They are conceptually different from other established enhanced sampling techniques in that they do not use any collective variable.

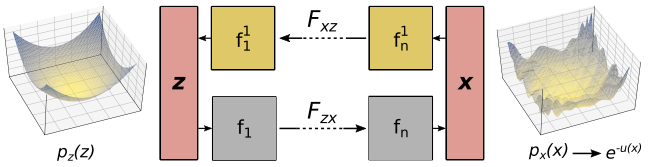



**Figure 4.** Boltzmann generators. An invertible neural network is used to generate independent samples that follow the desired Boltzmann distribution of a molecular system. First, a sample point is chosen from a simple distribution $p_z(\mathbf{z})$, e.g., a Gaussian distribution. Then the neural network transforms this sample to a configuration $\mathbf{x}$ that follows $p_x(\mathbf{x})$, which is a Boltzmann distribution similar to the one of the system. Lastly, to compute the thermodynamic properties, the samples are reweighted to their Boltzmann weight. The figure was redrawn based on refs 92 and 93.

**From Materials Properties to Performance and Application.** Even if we know all of the thermodynamic and transport properties of all of our materials, we still need to understand the techno-economic and engineering requirements of the application in order to develop performance metrics to objectively rank materials.[94] While this step crucially impacts our materials design strategy, it is so challenging that


https://dx.doi.org/10.1021/jacs.0c09105
J. Am. Chem. Soc. 2020, 142, 20273−20287

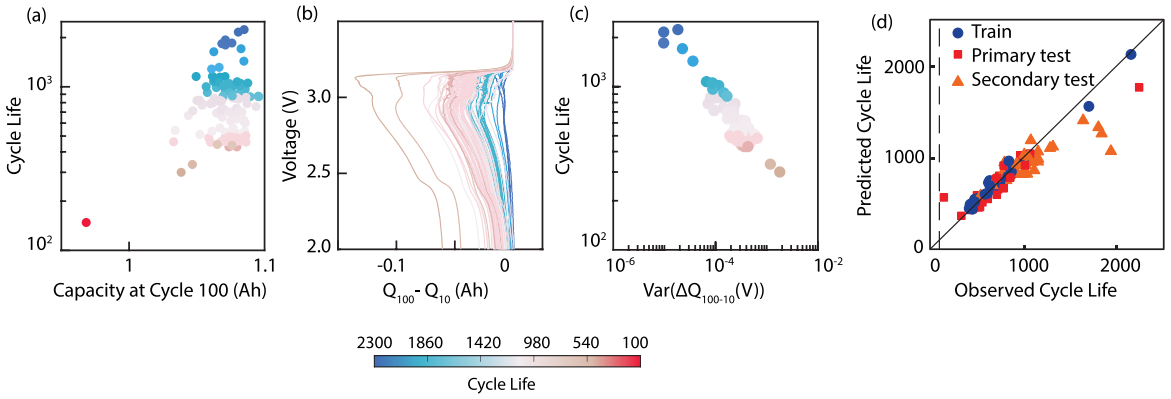



**Figure 5.** Prediction of battery life cycle from early stages. (a) The cycle life is shown with respect to cell capacity at cycle 100. (b, c) Characteristics of the voltage curves of the first cycles were used as features to develop the machine learning models. $Q_{100} - Q_{10}$ is change in discharge capacity between cycle 10 and 100. (d) Predictions of the machine learning model for two test sets. The secondary set was generated after model development. The vertical dashed line shows the 100th cycle, where the predictions were made. The figure was redrawn based on data from ref 102.

we often avoid confronting it. In particular, if we are carrying out research on novel materials, a complete techno-economic metric will be nearly impossible. For example, in many applications the costs will be an important factor.[95] However, how can we estimate the cost of a material that has not yet been synthesized? In the case of MOFs, the abundance of the metal and the complexity of the ligand can be good indications. However, one also has to factor in whether the synthesis can be scaled up easily.[96] Moreover, the engineering design might be totally constrained by nonscientific factors. For example, the adsorption pressure for the vehicular natural gas storage application was set to 65 bar by the Advanced Research Project Agency—Energy (ARPA-E), such that the process could be executed at home, while the minimum discharge pressure was set to 5.8 bar.[97] If one would select a lower minimum discharge pressure, materials with stronger adsorption sites for methane would become more favorable. As a consequence, if these metrics are not well-defined by external agencies, the metrics often become subjective and controversial; each material can be shown to be exceptionally good for one particular property. Therefore, only if we have an understanding of the relative importance of all properties in the context of the full engineering design of an application can we realistically evaluate whether a material will make a real impact. We also need to keep in mind that such metrics might give us the illusion that optimization of only one property will lead us to breakthrough materials. However, because of the complexity of the real-world application and the multistage design process, this is usually not the case.

One step toward unraveling this complexity is to establish an understanding of how materials properties influence the performance in an industrial process. For example, the overwhelming complexity of the evolution of the coupled ordinary/partial differential equations (ODEs/PDEs) underpinning mass and energy balance[98,99] often makes process modeling and optimization be seen as a black box. Using machine learning, we might be able to shine some light on how systems operate. Despite its significance, this topic has not been widely explored to date.[100] In one recent exceptional example, the effect of adsorbent properties on the carbon capture performance was analyzed by Burns et al.[101] Interestingly, they found that the common shortcut metrics for evaluation of materials are insufficient to predict the process-level performance evaluation of materials.

Besides, measuring or computing the performance metric can become a bottleneck in the case of complex processes and applications. An illustrative example is the lifetime estimation of battery cells.[102] The typical lifetime of lithium iron phosphate/graphite cell batteries varies over the range of 150 to 2300 cycles (Figure 5). However, since the battery capacity degradation undergoes a nonlinear process, it is challenging to predict the cycle life from early cycles. For instance, the capacity increases after 100 cycles for more than 81% of cells (see Figure 5a). Therefore, one needs to perform long cycle experiments, which often take months to years to execute. Previously, voltage curves were used for degradation diagnosis.[103] Hence, a machine learning model that monitors the voltage curves from early cycles was developed that can accurately (<4.9% test error) classify cells into long and short cycle life using only the first five cycles (Figure 5b–d). By aborting the long experiments of often hundreds or thousands of cycles for batteries that are not promising, the authors could save huge experimentation costs and time, allowing screening of a large number of candidates.

**Exploring the Design Space.** The final step is to explore the chemical space to find the best-performing candidates. We know that it is not feasible to exhaustively search the chemical space simply because of the exploding number of possible structures. For example, the number of theoretically feasible small drug molecules was estimated to exceed $10^{60}$.[104] Ultimately, screening only the known materials or hypothetical structures is not a solution, as these approaches cover only a limited part of the chemical space and specifically can be biased because of human choices or algorithmic limitations in structure generation.[14,41] Therefore, other search methods are desired to efficiently explore the enormous chemical space.[11,105−107] Crucial in these algorithmic searches is the need to balance between exploration, the process of probing the unseen regions of search space, and exploitation, the process of probing the promising regions.

A very popular class of discrete optimization methods is that of evolutionary algorithms, in particular genetic algorithms (GAs). These methods explore the space by evolving a population of structures through a set of iterative nature-inspired operations to optimize an objective (fitness) function


https://dx.doi.org/10.1021/jacs.0c09105
J. Am. Chem. Soc. 2020, 142, 20273−20287

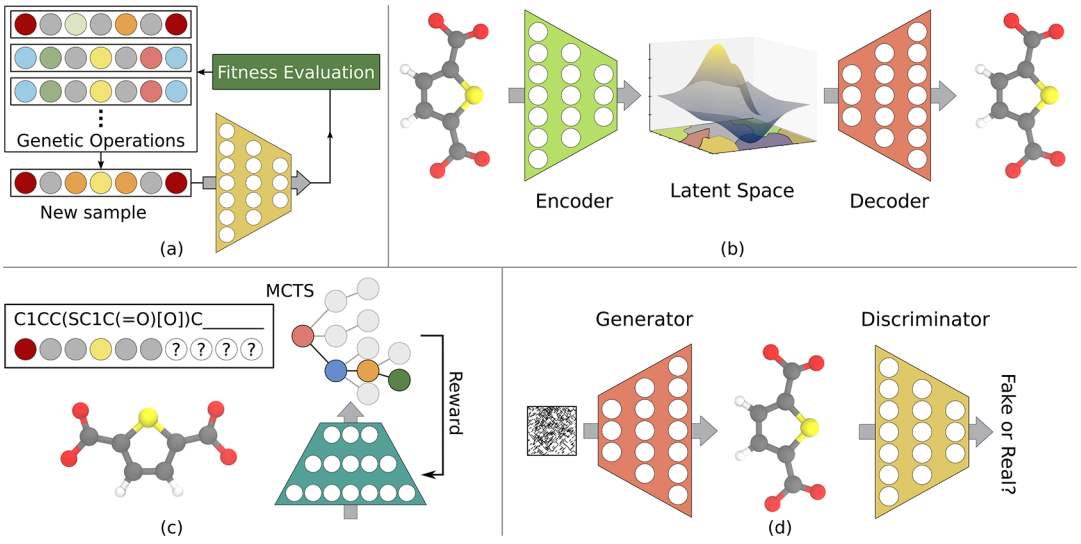



**Figure 6.** Methods for exploring chemical space. (a) Genetic algorithms use genetic operations to generate new samples that can quickly be evaluated by a machine learning model to maximize the fitness score. (b) Variational autoencoders (VAEs) learn a continuous lower-dimensional representation (the latent space) that can be used for gradient-based optimization of properties and recover the optimal chemicals by decoder. (c) Reinforcement-learning-based approach that incorporates Monte Carlo tree search (MCTS) to complete SMILES strings to generate new molecules, maximizing a reward function. (d) In a generative adversarial model, the generator and discriminator compete until the discriminator cannot distinguish generated samples from real empirical samples. By generating new samples, one can explore chemical space to maximize the properties of interest. The figure was redrawn based on ref 11.

(Figure 6a). Since the operations can be tailored and guided by chemical rules, it is a popular choice for chemical design.[108] The idea is that the samples with higher fitness scores have a better chance of survival and are selected more often to pass their genes to new samples. The mutation and permutation of genes, which could be functional groups of a ligand, control the ratio of exploration and exploitation in search. High mutation allows for searching of unexplored regions, while higher permutation ensures local searching. Machine learning can be used to quickly evaluate the fitness of generated samples, accelerating the search for materials discovery. Coupling GA with machine learning has been successfully used for materials synthesis,[109] discovery of transition metal complexes,[105] and organic molecules.[110] In addition, active learning approaches, which use uncertainty estimation in machine learning predictions, allow exploration of regions of space that were not in the training set by adding new data points to the training set on the fly when the model is uncertain.

Alternatively, one can use machine learning methods for the generation of structures.[111] In particular, in the area of organic molecules that follow basic valence rules of Simplified Molecular Input Line Entry System (SMILES) strings, recurrent neural networks (RNNs) or transformers are powerful in completing or generating new sequences of strings. RNNs and transformers have been developed to treat data sequences such as data in natural language processing or voice recognition. To guide and control the generation toward the properties of interest, one powerful approach is Monte Carlo tree search (MCTS). MCTS is used in reinforcement tasks, which involve real-time decision-making for the next moves, e.g., in playing games[112] or control, with a large, complex, and open-ended solution space. In analogy, we can think of the completion of a SMILES string as an open-ended process with a target: we win if the properties of interest improve (see Figure 6c). This approach has been found to be effective for exploring chemical space for different applications, such as MOFs for gas adsorption,[113] synthesis planning,[114,115] and the design of drug molecules.[116]

A desired property to circumvent the expensive optimization in discrete, enormous chemical space is to develop continuous and differentiable representations of chemical structures. If we couple these continuous representations with a generative model that converts a point in the continuous space to a chemical structure, we can perform direct gradient-based optimization of properties. Variational autoencoders (VAEs) are machine learning models that try to do this by learning a lower-dimensional representation of the data that is sufficient to regenerate the original data (Figure 6b). The chief component of a VAE is the lower-dimensional representation, which is called the latent space. By mapping data points to their probability distribution functions in the latent space, we can reach the continuous representation of chemical structures. This approach has recently been applied to organic molecules,[117,118] small molecular graphs,[119] solid-state materials,[120] and porous materials.[121]

An alternative method is generative adversarial networks (GANs). One neural network (the generator) generates new samples, and another one (the discriminator) tries to classify the generated data and some training data as fake or real (Figure 6d). The generator and the discriminator compete until the generator is so good that the discriminator does not have a better chance than 50% in distinguishing fake from real. GANs are finding their position in molecular and materials design,[111] as exemplified by the generation of molecular graphs in MolGAN[106] and an energy grid of guest molecules and zeolite structures in ZeoGAN.[122]

**Synthesis and Autonomous Experimentation.** To practically realize the aim of materials development, we need to be able to synthesize the promising candidates discovered in the previous steps. However, chemical synthesis is a complex, fuzzy process, and our theories are still too limited to guide us through it. Therefore, chemical synthesis mainly rests on


https://dx.doi.org/10.1021/jacs.0c09105
J. Am. Chem. Soc. 2020, 142, 20273−20287

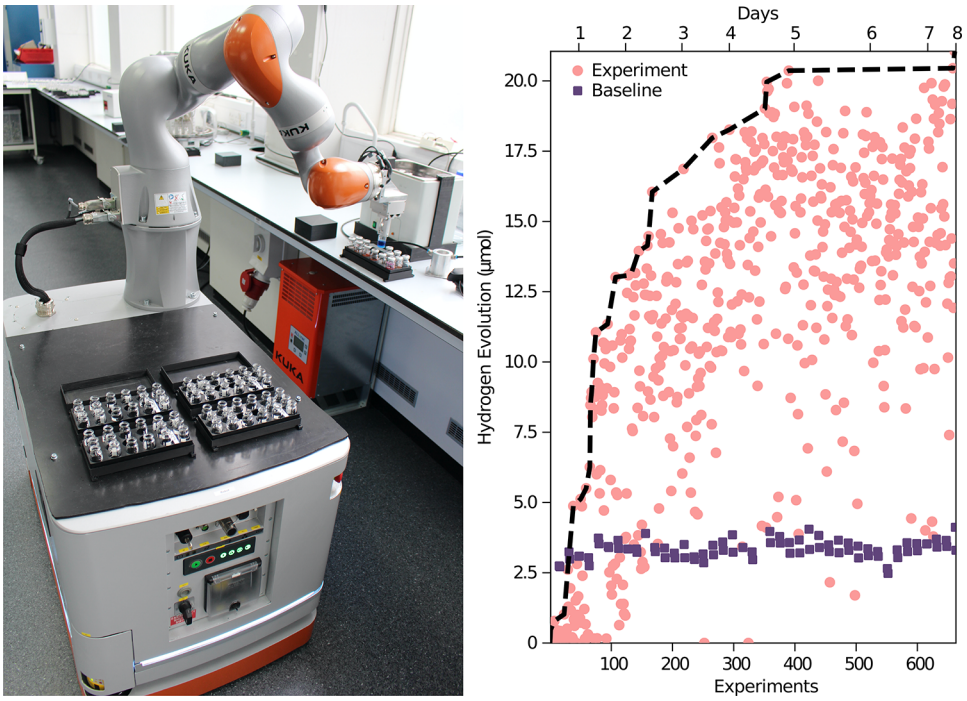



**Figure 7.** A mobile robotic chemist. The robot was used to perform an autonomous search to find a photocatalyst for hydrogen production from water. The robot improved the photocatalytic activity of the initial formulations (indicated by the baseline) by a factor of 6 over 8 days of searching the experimental space, performing 688 experiments. The photograph of the robot was provided by Andrew I. Cooper and Benjamin Burger (University of Liverpool). The figure was redrawn based on ref 137.

unwritten heuristic rules that experienced chemists gain over the course of many experiments. Data-driven approaches are promising alternatives for inferring such chemical intuition if they are presented with a sufficient number of failed and successful experiments. This concept was demonstrated for the synthesis of organic,[123,124] inorganic,[125−128] and MOF[109] materials. Notably, these data-driven methods have the privilege that their predictive performance improves upon provision of more data.

Coupling artificial intelligence with robotic synthesis platforms has taken the idea of autonomous laboratories and experimentation far.[14,129−131] The process of finding, optimizing, and executing synthesis is not only tedious and resource-intensive but also prone to bias and error. One can instead use robotic platforms to reduce the synthesis costs and errors simultaneously, while using artificial intelligence to control the robots. This approach has attracted tremendous attention recently and has led to the development of software like ChemOS[132,133] and hardware like the Chemputer[134] to perform experiments. Various methods have been used to guide these robots, from conventional farthest-point sampling and genetic algorithms[109] to Bayesian optimization,[135−137] again trying to balance exploration and exploitation of chemical synthesis space. The recent work of Burger et al.[137] introducing a mobile robotic chemist (Figure 7) demonstrates how fast this topic of research is growing, and it will be interesting to see whether it leads to the discovery of novel chemistries.

Another promising application of machine learning is in synthesis planning. The challenge here is to identify a feasible route (i.e., the reaction steps, conditions, and reactants) for the synthesis of chemical compounds starting from available chemicals. Ideally, we want a program that takes a target structure as input and provides a list of detailed feasible reaction steps, simultaneously minimizing the number of steps and complexity of the process. Data-driven approaches have recently been explored and shown to be promising for finding the synthesis steps in organic retrosynthesis,[12,114,138,139] suggesting organic directing agents for the synthesis of zeolites,[140] and identifying new phases of inorganic compounds.[127]

## CHALLENGES AND OPPORTUNITIES

**1. Data.** *An Ecosystem with Standards.* The most elementary part of machine learning studies is *data*. The basic standards for data management have been explained in terms of the findable, accessible, interoperable, and reusable (FAIR) guiding principles.[141] In essence, the FAIR principles require (meta)data to be openly retrievable by a unique global and persistent identifier as well as provided with the usage license and detailed provenance. To fully meet the FAIR principles, we must not only develop and use standardized ways of reporting data but also provide ways to access the tools, protocols, codes, and input parameters so that the data can be reproduced. Consequently, developing user-friendly, encouraging ecosystems for sharing and programmatically accessing FAIR data is a fundamental step and a key challenge to unlock the true power of data-driven approaches in the chemical sciences. In addition, it is now common that funding agencies ask for data management plans and require that all data be made publicly available. However, systematically doing these tasks requires a complete rethinking of the way we do research, in which reproducibility and data sharing are the starting point rather than an afterthought to meet the requirements of a journal or funding agency.

*Collection of Experimental Data.* While machine learning using failed experiments can be expected to be one of the most important areas in chemistry,[109,128,142] a large body of these


https://dx.doi.org/10.1021/jacs.0c09105
J. Am. Chem. Soc. 2020, 142, 20273−20287

failed experiments remains unreported. It is too demanding to expect researchers to spend a considerable fraction of their time on documenting failed experiments. Instead, since data are routinely generated over the course of a research project, solutions that are fully integrated with experimental instruments are needed in order to collect data while the user is performing the experiments. Such platforms have remained underdeveloped in chemical sciences. For example, electronic lab notebooks (ELNs),[143−145] allow sharing of protocols, postprocessing scripts, and measurement techniques in a collaborative fashion as well as real-time data acquisition. More importantly, ELNs can allow all of the data (failed and successful) to be published in standard formats with little or no additional effort on the part of the researcher. However, it is essential for the chemistry community to embrace the development of such an open science infrastructure.

*Reproducibility.* Anyone who has tried to reproduce results from the literature can testify that in many cases the articles do not provide all of the information needed to reproduce the results. In the case of computational results for example, often there are unreported parameters (e.g., default parameters in a code), and the reader of the article may be unaware of their importance. However, if these parameters change over time or different ones are used in different groups, it becomes impossible to reproduce the results. The most simple solution is to publish all input files and all scripts along with the article. However, managing this for large-scale calculations using multiple codes becomes intractable, and therefore, one needs a special infrastructure to be able to do this systematically. Recent development of infrastructures in this area, such as Materials Cloud and AiiDA,[35,146−148] and FireWorks[149] are opening promising paths toward addressing these issues. Automation and workflow development and execution tools for machine learning in materials science are also under development, e.g., ChemML[150] and Automatminer.[151] Creating, maintaining, and encouraging the use of these open science infrastructures require the support of the computational chemistry community.

*Data Curation.* As important as they are, data management and curation are among the least enjoyable, time-consuming, tedious, and error-prone tasks. Specifically, since we deal with a large number of data points, e.g., a large number of structures in databases like the CSD, manual inspection is out of question, and the development of new methods is inevitable. Exploring machine learning methods for improving or even building new innovative ways of data curation is an opportunity for future research in chemistry and materials science. Such methods for automatic data curation have recently received attention in many disciplines, including the chemical sciences.[152−155] For instance, by coupling uncertainty estimation methods exploiting the statistical nature of machine learning methods, one can identify mistakes and anomalies in big data. For example, in cases where the machine learning model is confident in its predictions but large discrepancies are observed with the reported data, the user can be warned to double-check the entry to avoid mistakes in databases.

*Data in the Literature.* There is a large body of data stored in the literature. Natural language processing for extracting data from text and image and sequence processing techniques for analyzing spectra would be potentially interesting. Unfortunately, a major obstacle to overcome here is to convert Portable Document Format (PDF) to compatible formats (e.g., plain text). In the future, it might be beneficial for the scientific community to consider reporting in other formats that are better suitable for machine interpretation.

**2. Bias and Uncertainty in the Design Process.** *Novelty, Bias, and Diversity.* Most scientific efforts have been focused on incremental improvements of some shortcut performance indicators, for example, the adsorption capacity and selectivity of MOFs for carbon capture. However, if we consider the full scope of the design process, we realize that such materials properties are only inputs for the next stage of the design (see Figure 1). Therefore, the approach based on incremental improvement of properties is not only limited in finding the true optimal solutions for the full design process but also introduces a strong bias by providing only limited options for the next stage of the design process. For example, for most real-world applications we need a trade-off between multiple properties, and the optimization of only one objective will exclude many solutions that might perform much better in the real problem. If we now also consider that the properties we optimize are not necessarily good surrogates for the practical application, we realize that focusing on the optimization of these metrics will limit our ability to discover novel materials, for which the application might be based on a mechanism of which we are not yet even aware. For this reason, we argue that for a broader perspective over the materials design process, enhancing novelty in each stage will be a better path to success than the optimization of single metrics. Essential here is the development of metrics that allow measurement of such novelty in the evaluation of scientific discoveries. Careful quantification of diversity by extending and developing new metrics in all stages of materials design can help us to reduce such bias.[41,156]

*Uncertainty Quantification and Error Propagation.* Since we are not using physical laws in machine learning models, it is crucially important to be able to identify the domain of applicability of the models for predictions of new systems. However, quantifying uncertainty can be challenging and costly, and this topic has only recently received some attention in the chemical sciences. Several methods have been proposed for quantifying uncertainty,[51,157,158] such as measuring the distance of a new sample to the training data or using ensemble models. Further studies are needed to provide an understanding of the limitation of current methods, to develop more reliable and cheap methods, and to provide guidelines on choosing the method for quantification of uncertainty.

In addition, the error we make in a design process is not limited only to machine learning predictions, as any simulation or experiment also has a level of accuracy. It is therefore important to know how errors propagate through the entire design process. Statistical methods can be used to analyze the sensitivity of the outcomes to the inputs, providing insight on the reliability and relevance of the entire process.

**3. Structure−Property−Performance.** *Featurization.* Further developments are required to apply machine learning for those materials properties that require a tensorial representation, are highly dependent on long-range interactions, or involve dynamics. For example, we are still limited in featurization of materials for those material properties that require tensorial representation, including the stiffness tensor for mechanical properties, the heat conduction tensor for heat transport, and the susceptibility tensor for magnetic/electronic properties. In addition, current representations are limited for properties that rely on structural dynamics. For example, we are aware of the role of flexibility on adsorption properties of


https://dx.doi.org/10.1021/jacs.0c09105
J. Am. Chem. Soc. 2020, 142, 20273−20287

soft porous materials (e.g., in MOFs), yet the commonly used representations do not capture these subtleties.

Additional developments are needed for generative models. For example, sequence-based generative models based on SMILES strings, which have been the main method for generative design of chemicals, cannot generalize to chemistries that do not follow valence-based rules.[11] Also, using generative models with SMILES strings can generate problems since many SMILES strings do not correspond to valid molecules. For this reason, novel representations that are based on a formal grammar have been developed.[159] We note that graph-theoretical descriptors ignore any information related to geometry. Therefore, for any materials and molecular properties that are sensitive to the details of atomic coordinates and geometry, current generative models are limited.

*Molecular Simulation.* The different angles of machine learning techniques for molecular simulation have advanced independently. Examples of these techniques include "machine learned" potentials,[79,83] enhanced sampling methods such as Boltzmann generators,[92] and methods for analysis of molecular dynamics trajectories.[160] The next step is to merge these methods into a toolbox that can be used for different systems at scale. Since these methods work hand in hand with conventional quantum and classical molecular simulation methods, it is of great value to implement and couple them in the existing simulation packages.[161,162]

The development of new modeling techniques will remain fundamentally important for the future of the application of machine learning methods. One of the main pillars of the fast development of data-driven methods in recent years has been the abundance of data, mainly simulated big data due to growing computational power. Hence, it will be continuously important to improve the simulation methods and their accuracy, especially for challenging problems such as nonlinear and noncontinuous phenomena (e.g., instability and regime change), where we still rely heavily on simulation.

*Modeling Complex and Dynamic Processes.* An interesting field of research that has barely been explored for process modeling is the use of machine learning to efficiently solve (nonlinear) partial differential equations.[163,164] These methods have shown great performance for solving complex Navier–Stokes equations in fluid mechanics, e.g., for turbulence applications.[165] Adapting these for process modeling will not only drop the computational costs but also allow the addition of more levels of complexity in modeling by including nonideal effects that are often ignored in such modeling.

*Making Machine Learning Comparable.* The field of machine learning for the design of materials would strongly benefit from establishing reporting standards and using benchmark sets for model comparison and evaluation. Tracking the successful path paved by the researcher in the field of small organic molecules teaches us that using benchmark data sets of molecules and their corresponding properties (labels) allowed them to move fast by enabling them to build on top of previous studies. Without a reference benchmark set of materials and labels, one cannot compare the performance of different featurizations and model architectures, as differences might originate in inhomogeneity in data, such as differences in computational methodologies, or the underlying distribution of structural databases and lack of diversity. Hence, benchmark materials sets with consistent properties need to be developed. Furthermore, since it is not trivial to compare models, agreement on standard reporting methods is needed. A valuable step forward was taken in this direction by Wang et al.[166] who provided guidelines for best practices of machine learning for materials scientists.

*New Learning Algorithms.* Future research on exploring state-of-the-art machine learning methods and expanding them for the chemical sciences is a significant opportunity. In particular, methods like transfer learning, multitask learning, and one-shot learning, which try to facilitate the learning process by transferring parameters or features and/or sharing contextual information, are attractive for cases in which we have little data for one materials class.

**4. Causal and Interpretable Machine Learning.** Explainable machine learning is opening a path toward obtaining fresh insights and developing novel theories. In contrast to the general perception of machine learning models as black boxes, interpreting explainable models can shine some light on the connections between the underlying structure and the corresponding property and performance. For example, machine learning models can be seen as extremely flexible empirical models that, similar to conventional empirical models, can uncover profound novel understanding and knowledge and inspire new theories if interpreted correctly. However, one needs to be cautious to not fall into the trap of correlation versus causation. For example, the number of sunburn cases is correlated with the amount of ice cream sold in a city, which happens obviously because of the dry, hot, sunny days in summer. However, not all cases are that obvious, and further fundamental research is required to find methods to measure the trustworthiness of explanations.[167,168]

In particular, explainable machine learning methods can potentially change the way we study phenomena for which we still have limited theories. In a typical physical system, one can assume that there are a few important terms, such as dimensionless numbers in fluid mechanics, that govern the behavior of the system. Therefore, using machine learning and symbolic equations, one can try to extract the governing equations from large data.[71,72]

**5. Synthesizability.** Perhaps the greatest challenge for computational and data-driven material design and discovery is the synthesizability of discovered structures. Because of the great progress in methods for inverse design, we can maneuver chemical space to find the optimal materials and molecules. However, the full use of this approach is hampered by our ignorance of the synthesizability of the discovered structures. Even if we restrict our search to theoretically valid structures, the discovered structures would often seem impossible to synthesize. Therefore, developing universal solutions for biasing the search for chemicals toward the synthetically accessible parts of chemical space will be an important research direction in the future. An interesting solution is to design structural motifs—that can be incorporated into chemically synthesizable structures—instead of the full structure. For example, instead of discovering a MOF that is optimal for $CO_2$ capture in the presence of water, Boyd et al.[24] discovered a set of adsorption sites, named adsorbaphores. In the next step, they generated a new set of hypothetical structures that contain those adsorbaphores and are also water-repellent. However, in this step, they restricted their search to a set that was guided by experimentalists to be synthesizable. This supervised search relied on the intuition of expert chemists. Indeed, machine learning can help us here in inferring and capturing this intuition. Further research in this direction needs to explore


https://dx.doi.org/10.1021/jacs.0c09105
J. Am. Chem. Soc. 2020, 142, 20273−20287

the extent to which we can encode synthesizability into computational and data-driven materials design.

## OUTLOOK

Machine learning is transforming the way that we approach rational materials design. The inherent complexity of searching the vast spaces of options we face in the process of material design, from materials to processes and applications, requires the development of methods that work best in the limit of large numbers. Machine learning methods provide us this toolbox. Using these methods, we can conceptualize a new way of approaching materials design. The remarkable advancements that we have reviewed in this Perspective are shown as proofs of principle for the components of such an approach. By advancing and merging these components, we can fully exploit all of these advances and realize the power of data-assisted materials design. Indeed, there are still significant challenges on the way, some of which have been mentioned here. However, considering the fast progress in recent years, we can envision that machine learning will be integrated into almost all components of materials design and discovery in the near future.

A pillar of success for the future of this approach is data. When we rely on data to infer the solutions for our problems, the generation of large-scale accurate and reproducible data is vital. Nevertheless, we admit that it is one of the grand challenges for the future of the field. In particular, overcoming some of the challenges on this topic requires introducing new research cultures and collaboration among multiple disciplines from sciences and engineering, including both theoreticians and experimentalists. Therefore, only through an open, disciplined, and collaborative environment based on agreements on data reporting and protocols we will be able to move fast and use the real power of data-driven methods for materials design.

Most of the discoveries in the history of science were not purely rational but relied on the intuition of scientists. Here one can see the scientists as black boxes who have bright intuitions in decision-making. The interesting fact about machine learning models is that once we have a discovery or prediction, we can trace back the paths of decision-making to uncover new insights. Therefore, we can now focus on tailoring the future of materials design using the opportunities that machine learning can bring to us for doing better science.

## AUTHOR INFORMATION


### Corresponding Author

**Berend Smit** − *Laboratory of Molecular Simulation, Institut des Sciences et Ingénierie Chimiques, École Polytechnique Fédérale de Lausanne (EPFL), CH-1951 Sion, Valais, Switzerland;* orcid.org/0000-0003-4653-8562; Email: berend.smit@epfl.ch

### Authors

**Seyed Mohamad Moosavi** − *Laboratory of Molecular Simulation, Institut des Sciences et Ingénierie Chimiques, École Polytechnique Fédérale de Lausanne (EPFL), CH-1951 Sion, Valais, Switzerland;* orcid.org/0000-0002-0357-5729

**Kevin Maik Jablonka** − *Laboratory of Molecular Simulation, Institut des Sciences et Ingénierie Chimiques, École Polytechnique Fédérale de Lausanne (EPFL), CH-1951 Sion, Valais, Switzerland;* orcid.org/0000-0003-4894-4660

Complete contact information is available at: https://pubs.acs.org/10.1021/jacs.0c09105


### Notes

The authors declare no competing financial interest.

## ACKNOWLEDGMENTS

This work was supported by the European Research Council (ERC) under the European Union's Horizon 2020 Research and Innovation Programme (Grant Agreement 666983, MaGic); by NCCR-MARVEL, funded by the Swiss National Science Foundation; in part by the PrISMa Project (299659), funded through the ACT Programme (Accelerating CCS Technologies, Horizon 2020 Project 294766). Financial contributions from the Department for Business, Energy & Industrial Strategy (BEIS) together with extra funding from the NERC and EPSRC Research Councils, United Kingdom, the Research Council of Norway (RCN), the Swiss Federal Office of Energy (SFOE), and the U.S. Department of Energy are gratefully acknowledged. Additional financial support from TOTAL and Equinor is also gratefully acknowledged. K.M.J. acknowledges support from the Swiss National Science Foundation (SNSF) under Grant 200021_172759. We thank Andrew I. Cooper and Benjamin Burger (University of Liverpool) for providing photos of the mobile robot.

## REFERENCES


(1) Yaghi, O. M.; Kalmutzki, M. J.; Diercks, C. S. *Introduction to Reticular Chemistry: Metal−Organic Frameworks and Covalent Organic Frameworks*; John Wiley & Sons, 2019.

(2) Allcock, H. R. Rational design and synthesis of new polymeric material. *Science* **1992**, *255*, 1106−1112.

(3) Jones, M. R.; Seeman, N. C.; Mirkin, C. A. Programmable Materials and the Nature of the DNA Bond. *Science* **2015**, *347*, 1260901.

(4) Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*; Springer Series in Statistics; Springer: New York, 2001; Vol. *1*.

(5) *Machine Learning Meets Quantum Physics*; Schütt, K. T., Chmiela, S., von Lilienfeld, O. A., Tkatchenko, A., Tsuda, K., Müller, K.-R., Eds.; Springer, 2020.

(6) Mehta, P.; Bukov, M.; Wang, C.-H.; Day, A. G.; Richardson, C.; Fisher, C. K.; Schwab, D. J. A high-bias, low-variance introduction to Machine Learning for physicists. *Phys. Rep.* **2019**, *810*, 1−124.

(7) Goodfellow, I.; Bengio, Y.; Courville, A. *Deep Learning*; MIT Press, 2016; http://www.deeplearningbook.org.

(8) LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436−444.

(9) Raghu, M.; Schmidt, E. A Survey of Deep Learning for Scientific Discovery. *arXiv (Computer Science.Machine Learning)*, March 26, 2020, 2003.11755, ver. 1. https://arxiv.org/abs/2003.11755 (accessed 2020-08-24).

(10) Zhang, A.; Lipton, Z. C.; Li, M.; Smola, A. J. *Dive into Deep Learning*. https://d2l.ai/ (accessed 2020-08-24).

(11) Sanchez-Lengeling, B.; Aspuru-Guzik, A. Inverse molecular design using machine learning: Generative models for matter engineering. *Science* **2018**, *361*, 360−365.

(12) Coley, C. W.; Green, W. H.; Jensen, K. F. Machine learning in computer-aided synthesis planning. *Acc. Chem. Res.* **2018**, *51*, 1281−1289.

(13) Häse, F.; Roch, L. M.; Aspuru-Guzik, A. Next-generation experimentation with self-driving laboratories. *Trends Chem.* **2019**, *1*, 282−291.

(14) Gromski, P. S.; Henson, A. B.; Granda, J. M.; Cronin, L. How to explore chemical space using algorithms and automation. *Nat. Rev. Chem.* **2019**, *3*, 119−128.
(15) Noé, F.; Tkatchenko, A.; Müller, K.-R.; Clementi, C. Machine learning for molecular simulation. *Annu. Rev. Phys. Chem.* **2020**, *71*, 361−390.
(16) Aykol, M.; Herring, P.; Anapolsky, A. Machine learning for continuous innovation in battery technologies. *Nat. Rev. Mater.* **2020**, *5*, 725−727.
(17) Jablonka, K. M.; Ongari, D.; Moosavi, S. M.; Smit, B. Big-Data Science in Porous Materials: Materials Genomics and Machine Learning. *Chem. Rev.* **2020**, *120*, 8066−8129.
(18) Hellenbrandt, M. The Inorganic Crystal Structure Database (ICSD)—present and future. *Crystallogr. Rev.* **2004**, *10*, 17−22.
(19) Gražulis, S.; Chateigner, D.; Downs, R. T.; Yokochi, A.; Quirós, M.; Lutterotti, L.; Manakova, E.; Butkus, J.; Moeck, P.; Le Bail, A. Crystallography Open Database—an open-access collection of crystal structures. *J. Appl. Crystallogr.* **2009**, *42*, 726−729.
(20) Gates-Rector, S.; Blanton, T. The Powder Diffraction File: a quality materials characterization database. *Powder Diffr.* **2019**, *34*, 352−360.
(21) Groom, C. R.; Bruno, I. J.; Lightfoot, M. P.; Ward, S. C. The Cambridge Structural Database. *Acta Crystallogr., Sect. B: Struct. Sci., Cryst. Eng. Mater.* **2016**, *72*, 171−179.
(22) Boyd, P. G.; Lee, Y.; Smit, B. Computational development of the nanoporous materials genome. *Nat. Rev. Mater.* **2017**, *2*, 17037.
(23) Wilmer, C. E.; Leaf, M.; Lee, C. Y.; Farha, O. K.; Hauser, B. G.; Hupp, J. T.; Snurr, R. Q. Large-scale screening of hypothetical metal−organic frameworks. *Nat. Chem.* **2012**, *4*, 83.
(24) Boyd, P. G.; et al. Data-driven design of metal−organic frameworks for wet flue gas $CO_2$ capture. *Nature* **2019**, *576*, 253−256.
(25) Gómez-Gualdrón, D. A.; Colón, Y. J.; Zhang, X.; Wang, T. C.; Chen, Y.-S.; Hupp, J. T.; Yildirim, T.; Farha, O. K.; Zhang, J.; Snurr, R. Q. Evaluating topologically diverse metal−organic frameworks for cryo-adsorbed hydrogen storage. *Energy Environ. Sci.* **2016**, *9*, 3279−3289.
(26) Moosavi, S. M.; Boyd, P. G.; Sarkisov, L.; Smit, B. Improving the mechanical stability of metal−organic frameworks using chemical caryatids. *ACS Cent. Sci.* **2018**, *4*, 832−839.
(27) Witman, M.; Ling, S.; Anderson, S.; Tong, L.; Stylianou, K. C.; Slater, B.; Smit, B.; Haranczyk, M. In silico design and screening of hypothetical MOF-74 analogs and their experimental synthesis. *Chem. Sci.* **2016**, *7*, 6263−6272.
(28) Mercado, R.; Fu, R.-S.; Yakutovich, A. V.; Talirz, L.; Haranczyk, M.; Smit, B. In silico design of 2D and 3D covalent organic frameworks for methane storage applications. *Chem. Mater.* **2018**, *30*, 5069−5086.
(29) Martin, R. L.; Simon, C. M.; Smit, B.; Haranczyk, M. In silico design of porous polymer networks: high-throughput screening for methane storage materials. *J. Am. Chem. Soc.* **2014**, *136*, 5006−5022.
(30) Pophale, R.; Cheeseman, P. A.; Deem, M. W. A database of new zeolite-like materials. *Phys. Chem. Chem. Phys.* **2011**, *13*, 12407−12412.
(31) Deem, M. W.; Pophale, R.; Cheeseman, P. A.; Earl, D. J. Computational discovery of new zeolite-like materials. *J. Phys. Chem. C* **2009**, *113*, 21353−21360.
(32) Jain, A.; Ong, S. P.; Hautier, G.; Chen, W.; Richards, W. D.; Dacek, S.; Cholia, S.; Gunter, D.; Skinner, D.; Ceder, G.; Persson, K. A. Commentary: The Materials Project: A materials genome approach to accelerating materials innovation. *APL Mater.* **2013**, *1*, 011002.
(33) Villars, P.; Cenzual, K.; Gladyshevskii, R.; Iwata, S. PAULING FILE - towards a holistic view. *Chem. Met. Alloys* **2018**, *11*, 43−76.
(34) Draxl, C.; Scheffler, M. The NOMAD laboratory: from data sharing to artificial intelligence. *J. Phys.: Mater.* **2019**, *2*, 036001.
(35) Talirz, L.; et al. Materials Cloud, a platform for open computational science. *Sci. Data* **2020**, *7*, 299.
(36) Siderius, D.; Shen, V.; Johnson, R., III; van Zee, R. *NIST/ARPA-E Database of Novel and Emerging Adsorbent Materials*; National Institute of Standards and Technology: Gaithersburg, MD, 2014; 10, T43882.
(37) Pfeif, E.; Kroenlein, K. Perspective: Data infrastructure for high throughput materials discovery. *APL Mater.* **2016**, *4*, 053203.
(38) Wilthan, B.; Pfeif, E. A.; Diky, V. V.; Chirico, R. D.; Kattner, U. R.; Kroenlein, K. Data resources for thermophysical properties of metals and alloys, part 1: structured data capture from the archival literature. *CALPHAD: Comput. Coupling Phase Diagrams Thermochem.* **2017**, *56*, 126−138.
(39) De, S.; Bartók, A. P.; Csányi, G.; Ceriotti, M. Comparing molecules and solids across structural and alchemical space. *Phys. Chem. Chem. Phys.* **2016**, *18*, 13754−13769.
(40) Nicholas, T. C.; Goodwin, A. L.; Deringer, V. L. Understanding the Geometric Diversity of Inorganic and Hybrid Frameworks through Structural Coarse-Graining. *arXiv (Condensed Matter.Materials Science)*, May 20, 2020, 2005.09939, ver. 1. https://arxiv.org/abs/2005.09939 (accessed 2020-08-24).
(41) Moosavi, S. M.; Nandy, A.; Jablonka, K. M.; Ongari, D.; Janet, J. P.; Boyd, P. G.; Lee, Y.; Smit, B.; Kulik, H. J. Understanding the diversity of the metal−organic framework ecosystem. *Nat. Commun.* **2020**, *11*, 4068.
(42) Hansen, K.; Montavon, G.; Biegler, F.; Fazli, S.; Rupp, M.; Scheffler, M.; Von Lilienfeld, O. A.; Tkatchenko, A.; Muller, K.-R. Assessment and validation of machine learning methods for predicting molecular atomization energies. *J. Chem. Theory Comput.* **2013**, *9*, 3404−3419.
(43) Ghiringhelli, L. M.; Vybiral, J.; Levchenko, S. V.; Draxl, C.; Scheffler, M. Big data of materials science: critical role of the descriptor. *Phys. Rev. Lett.* **2015**, *114*, 105503.
(44) Chung, Y. G.; Haldoupis, E.; Bucior, B. J.; Haranczyk, M.; Lee, S.; Zhang, H.; Vogiatzis, K. D.; Milisavljevic, M.; Ling, S.; Camp, J. S.; Slater, B.; Siepmann, J. I.; Sholl, D. S.; Snurr, R. Q. Advances, Updates, and Analytics for the Computation-Ready, Experimental Metal−Organic Framework Database: CoRE MOF 2019. *J. Chem. Eng. Data* **2019**, *64*, 5985−5998.
(45) Lee, Y.; Barthel, S. D.; Dłotko, P.; Moosavi, S. M.; Hess, K.; Smit, B. Quantifying similarity of pore-geometry in nanoporous materials. *Nat. Commun.* **2017**, *8*, 15396.
(46) Moosavi, S. M.; Xu, H.; Chen, L.; Cooper, A. I.; Smit, B. Geometric landscapes for material discovery within energy−structure−function maps. *Chem. Sci.* **2020**, *11*, 5423−5433.
(47) Lee, Y.; Barthel, S. D.; Dłotko, P.; Moosavi, S. M.; Hess, K.; Smit, B. High-throughput screening approach for nanoporous materials genome using topological data analysis: application to zeolites. *J. Chem. Theory Comput.* **2018**, *14*, 4427−4437.
(48) Kim, E.; Huang, K.; Tomala, A.; Matthews, S.; Strubell, E.; Saunders, A.; McCallum, A.; Olivetti, E. Machine-learned and codified synthesis parameters of oxide materials. *Sci. Data* **2017**, *4*, 170127.
(49) Kim, E.; Huang, K.; Saunders, A.; McCallum, A.; Ceder, G.; Olivetti, E. Materials synthesis insights from scientific literature via text extraction and machine learning. *Chem. Mater.* **2017**, *29*, 9436−9444.
(50) Tshitoyan, V.; Dagdelen, J.; Weston, L.; Dunn, A.; Rong, Z.; Kononova, O.; Persson, K. A.; Ceder, G.; Jain, A. Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature* **2019**, *571*, 95−98.
(51) Jablonka, K. M.; Ongari, D.; Moosavi, S. M.; Smit, B. Using collective knowledge to assign oxidation states. *ChemRxiv* **2020**, DOI: 10.26434/chemrxiv.11604129.v1.
(52) Duan, C.; Janet, J. P.; Liu, F.; Nandy, A.; Kulik, H. J. Learning from failure: predicting electronic structure calculation outcomes with machine learning models. *J. Chem. Theory Comput.* **2019**, *15*, 2331−2345.
(53) Fernandez, M.; Boyd, P. G.; Daff, T. D.; Aghaji, M. Z.; Woo, T. K. Rapid and accurate machine learning recognition of high performing metal organic frameworks for $CO_2$ capture. *J. Phys. Chem. Lett.* **2014**, *5*, 3056−3060.

(54) Sun, Y.; DeJaco, R. F.; Siepmann, J. I. Deep neural network learning of complex binary sorption equilibria from molecular simulation data. *Chem. Sci.* **2019**, *10*, 4377−4388.
(55) Anderson, R.; Biong, A.; Gómez-Gualdrón, D. A. Adsorption Isotherm Predictions for Multiple Molecules in MOFs Using the Same Deep Learning Model. *J. Chem. Theory Comput.* **2020**, *16*, 1271−1283.
(56) Ma, R.; Colon, Y. J.; Luo, T. Transfer Learning Study of Gas Adsorption in Metal−Organic Frameworks. *ACS Appl. Mater. Interfaces* **2020**, *12*, 34041−34048.
(57) Kitchin, J. R. Machine learning in catalysis. *Nat. Catal.* **2018**, *1*, 230−232.
(58) Goldsmith, B. R.; Esterhuizen, J.; Liu, J.-X.; Bartel, C. J.; Sutton, C. Machine learning for heterogeneous catalyst design and discovery. *AIChE J.* **2018**, *64*, 2311−2323.
(59) Nandy, A.; Zhu, J.; Janet, J. P.; Duan, C.; Getman, R. B.; Kulik, H. J. Machine Learning Accelerates the Discovery of Design Rules and Exceptions in Stable Metal−Oxo Intermediate Formation. *ACS Catal.* **2019**, *9*, 8243−8255.
(60) Isayev, O.; Oses, C.; Toher, C.; Gossett, E.; Curtarolo, S.; Tropsha, A. Universal fragment descriptors for predicting properties of inorganic crystals. *Nat. Commun.* **2017**, *8*, 15679.
(61) Eckhoff, M.; Behler, J. From molecular fragments to the bulk: development of a neural network potential for MOF-5. *J. Chem. Theory Comput.* **2019**, *15*, 3793−3809.
(62) Gaultois, M. W.; Oliynyk, A. O.; Mar, A.; Sparks, T. D.; Mulholland, G. J.; Meredig, B. Perspective: Web-based machine learning models for real-time screening of thermoelectric materials properties. *APL Mater.* **2016**, *4*, 053213.
(63) Furmanchuk, A.; Saal, J. E.; Doak, J. W.; Olson, G. B.; Choudhary, A.; Agrawal, A. Prediction of seebeck coefficient for compounds without restriction to fixed stoichiometry: A machine learning approach. *J. Comput. Chem.* **2018**, *39*, 191−202.
(64) Agrawal, A.; Deshpande, P. D.; Cecen, A.; Basavarsu, G. P.; Choudhary, A. N.; Kalidindi, S. R. Exploration of data science techniques to predict fatigue strength of steel from composition and processing parameters. *Integr. Mater. Manuf. Innovation* **2014**, *3*, 90−108.
(65) Evans, J. D.; Coudert, F.-X. Predicting the mechanical properties of zeolite frameworks by machine learning. *Chem. Mater.* **2017**, *29*, 7833−7839.
(66) Moghadam, P. Z.; Rogge, S. M.; Li, A.; Chow, C.-M.; Wieme, J.; Moharrami, N.; Aragones-Anglada, M.; Conduit, G.; Gomez-Gualdron, D. A.; Van Speybroeck, V.; Fairen-Jimenez, D. Structure-Mechanical Stability Relations of Metal−Organic Frameworks via Machine Learning. *Matter* **2019**, *1*, 219−234.
(67) Xie, T.; Grossman, J. C. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Phys. Rev. Lett.* **2018**, *120*, 145301.
(68) Mansouri Tehrani, A.; Oliynyk, A. O.; Parry, M.; Rizvi, Z.; Couper, S.; Lin, F.; Miyagi, L.; Sparks, T. D.; Brgoch, J. Machine learning directed search for ultraincompressible, superhard materials. *J. Am. Chem. Soc.* **2018**, *140*, 9844−9853.
(69) Schütt, K. T.; Glawe, H.; Brockherde, F.; Sanna, A.; Müller, K. R.; Gross, E. K. U. How to Represent Crystal Structures for Machine Learning: Towards Fast Prediction of Electronic Properties. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2014**, *89*, 205118.
(70) Häse, F.; Kreisbeck, C.; Aspuru-Guzik, A. Machine Learning for Quantum Dynamics: Deep Learning of Excitation Energy Transfer Properties. *Chem. Sci.* **2017**, *8*, 8419−8426.
(71) Brunton, S. L.; Proctor, J. L.; Kutz, J. N. Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proc. Natl. Acad. Sci. U. S. A.* **2016**, *113*, 3932−3937.
(72) Cranmer, M.; Sanchez-Gonzalez, A.; Battaglia, P.; Xu, R.; Cranmer, K.; Spergel, D.; Ho, S. Discovering Symbolic Models from Deep Learning with Inductive Biases. *arXiv (Computer Science.Machine Learning)*, June 19, 2020, 2006.11287, ver. 1. https://arxiv.org/abs/2006.11287 (accessed 2020-08-24).
(73) Behler, J.; Martoňák, R.; Donadio, D.; Parrinello, M. Metadynamics Simulations of the High-Pressure Phases of Silicon Employing a High-Dimensional Neural Network Potential. *Phys. Rev. Lett.* **2008**, *100*, 185501.
(74) Maresca, F.; Dragoni, D.; Csányi, G.; Marzari, N.; Curtin, W. A. Screw Dislocation Structure and Mobility in Body Centered Cubic Fe Predicted by a Gaussian Approximation Potential. *npj Comput. Mater.* **2018**, *4*, 69.
(75) Mardt, A.; Pasquali, L.; Wu, H.; Noé, F. VAMPnets for Deep Learning of Molecular Kinetics. *Nat. Commun.* **2018**, *9*, 5.
(76) Schütt, K. T.; Arbabzadah, F.; Chmiela, S.; Müller, K. R.; Tkatchenko, A. Quantum-Chemical Insights from Deep Tensor Neural Networks. *Nat. Commun.* **2017**, *8*, 13890.
(77) Dral, P. O.; Barbatti, M.; Thiel, W. Nonadiabatic Excited-State Dynamics with Machine Learning. *J. Phys. Chem. Lett.* **2018**, *9*, 5660−5663.
(78) Westermayr, J.; Gastegger, M.; Marquetand, P. Combining SchNet and SHARC: The SchNarc Machine Learning Approach for Excited-State Dynamics. *J. Phys. Chem. Lett.* **2020**, *11*, 3828−3834.
(79) Behler, J.; Parrinello, M. Generalized neural-network representation of high-dimensional potential-energy surfaces. *Phys. Rev. Lett.* **2007**, *98*, 146401.
(80) Behler, J. Atom-centered symmetry functions for constructing high-dimensional neural network potentials. *J. Chem. Phys.* **2011**, *134*, 074106.
(81) Behler, J. Perspective: Machine learning potentials for atomistic simulations. *J. Chem. Phys.* **2016**, *145*, 170901.
(82) Bartók, A. P.; Payne, M. C.; Kondor, R.; Csányi, G. Gaussian approximation potentials: The accuracy of quantum mechanics, without the electrons. *Phys. Rev. Lett.* **2010**, *104*, 136403.
(83) Bartók, A. P.; Kondor, R.; Csányi, G. On representing chemical environments. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2013**, *87*, 184115.
(84) Smith, J. S.; Isayev, O.; Roitberg, A. E. ANI-1: an extensible neural network potential with DFT accuracy at force field computational cost. *Chem. Sci.* **2017**, *8*, 3192−3203.
(85) Rowe, P.; Deringer, V. L.; Gasparotto, P.; Csányi, G.; Michaelides, A. An accurate and transferable machine learning potential for carbon. *J. Chem. Phys.* **2020**, *153*, 034702.
(86) Bernstein, N.; Bhattarai, B.; Csányi, G.; Drabold, D. A.; Elliott, S. R.; Deringer, V. L. Quantifying Chemical Structure and Machine-Learned Atomic Energies in Amorphous and Liquid Silicon. *Angew. Chem.* **2019**, *131*, 7131−7135.
(87) Jinnouchi, R.; Lahnsteiner, J.; Karsai, F.; Kresse, G.; Bokdam, M. Phase Transitions of Hybrid Perovskites Simulated by Machine-Learning Force Fields Trained on the Fly with Bayesian Inference. *Phys. Rev. Lett.* **2019**, *122*, 225701.
(88) Torrie, G. M.; Valleau, J. P. Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling. *J. Comput. Phys.* **1977**, *23*, 187−199.
(89) Hamelberg, D.; Mongan, J.; McCammon, J. A. Accelerated molecular dynamics: a promising and efficient simulation method for biomolecules. *J. Chem. Phys.* **2004**, *120*, 11919−11929.
(90) Laio, A.; Parrinello, M. Escaping free-energy minima. *Proc. Natl. Acad. Sci. U. S. A.* **2002**, *99*, 12562−12566.
(91) Frenkel, D.; Smit, B. *Understanding Molecular Simulation: From Algorithms to Applications*; Elsevier, 2001; Vol. *1*.
(92) Noé, F.; Olsson, S.; Köhler, J.; Wu, H. Boltzmann generators: Sampling equilibrium states of many-body systems with deep learning. *Science* **2019**, *365*, No. eaaw1147.
(93) Tuckerman, M. E. Machine learning transforms how microstates are sampled. *Science* **2019**, *365*, 982−983.
(94) Brandani, S.; Farmahini, A. H.; Friedrich, D.; Krishnamurthy, S.; Sarkisov, L. Performance-based screening of porous materials for carbon capture. *arXiv (Condensed Matter.Materials Science)*, September 25, 2020, 2009.12289, ver. 1. https://arxiv.org/abs/2009.12289 (accessed 2020-09-25).
(95) Subraveti, S. G.; Roussanaly, S.; Anantharaman, R.; Riboldi, L.; Rajendran, A. Techno-economic assessment of optimized vacuum

swing adsorption for post-combustion $CO_2$ capture from steam-methane reformer flue gas. *Sep. Purif. Technol.* **2021**, *256*, 117832.
(96) DeSantis, D.; Mason, J. A.; James, B. D.; Houchins, C.; Long, J. R.; Veenstra, M. Techno-Economic Analysis of Metal−Organic Frameworks for Hydrogen and Natural Gas Storage. *Energy Fuels* **2017**, *31*, 2024−2032.
(97) Peng, Y.; Krungleviciute, V.; Eryazici, I.; Hupp, J. T.; Farha, O. K.; Yildirim, T. Methane storage in metal−organic frameworks: current records, surprise findings, and challenges. *J. Am. Chem. Soc.* **2013**, *135*, 11887−11894.
(98) Farmahini, A. H.; Krishnamurthy, S.; Friedrich, D.; Brandani, S.; Sarkisov, L. From crystal to adsorption column: challenges in multiscale computational screening of materials for adsorption separation processes. *Ind. Eng. Chem. Res.* **2018**, *57*, 15491−15511.
(99) Farmahini, A. H.; Friedrich, D.; Brandani, S.; Sarkisov, L. Exploring new sources of efficiency in process-driven materials screening for post-combustion carbon capture. *Energy Environ. Sci.* **2020**, *13*, 1018−1037.
(100) Subraveti, S. G.; Li, Z.; Prasad, V.; Rajendran, A. Machine Learning-Based Multiobjective Optimization of Pressure Swing Adsorption. *Ind. Eng. Chem. Res.* **2019**, *58*, 20412−20422.
(101) Burns, T. D.; Pai, K. N.; Subraveti, S. G.; Collins, S. P.; Krykunov, M.; Rajendran, A.; Woo, T. K. Prediction of MOF Performance in Vacuum Swing Adsorption Systems for Postcombustion $CO_2$ Capture Based on Integrated Molecular Simulations, Process Optimizations, and Machine Learning Models. *Environ. Sci. Technol.* **2020**, *54*, 4536−4544.
(102) Severson, K. A.; Attia, P. M.; Jin, N.; Perkins, N.; Jiang, B.; Yang, Z.; Chen, M. H.; Aykol, M.; Herring, P. K.; Fraggedakis, D.; Bazant, M. Z.; Harris, S. J.; Chueh, W. C.; Braatz, R. D. Data-driven prediction of battery cycle life before capacity degradation. *Nat. Energy* **2019**, *4*, 383−391.
(103) Anseán, D.; Dubarry, M.; Devie, A.; Liaw, B.; García, V.; Viera, J.; González, M. Operando lithium plating quantification and early detection of a commercial LiFePO4 cell cycled under dynamic driving schedule. *J. Power Sources* **2017**, *356*, 36−46.
(104) Virshup, A. M.; Contreras-García, J.; Wipf, P.; Yang, W.; Beratan, D. N. Stochastic voyages into uncharted chemical space produce a representative library of all possible drug-like compounds. *J. Am. Chem. Soc.* **2013**, *135*, 7296−7303.
(105) Janet, J. P.; Chan, L.; Kulik, H. J. Accelerating chemical discovery with machine learning: simulated evolution of spin crossover complexes with an artificial neural network. *J. Phys. Chem. Lett.* **2018**, *9*, 1064−1071.
(106) De Cao, N.; Kipf, T. MolGAN: An implicit generative model for small molecular graphs. *arXiv (Statistics.Machine Learning)*, May 30, 2018, 1805.11973, ver. 1. https://arxiv.org/abs/1805.11973 (accessed 2020-08-24).
(107) Mercado, R.; Rastemo, T.; Lindelöf, E.; Klambauer, G.; Engkvist, O.; Chen, H.; Bjerrum, E. J. Graph Networks for Molecular Design. *ChemRxiv* **2020**, DOI: 10.26434/chemrxiv.12843137.v1.
(108) Bao, Y.; Martin, R. L.; Simon, C. M.; Haranczyk, M.; Smit, B.; Deem, M. W. In silico discovery of high deliverable capacity metal−organic frameworks. *J. Phys. Chem. C* **2015**, *119*, 186−195.
(109) Moosavi, S. M.; Chidambaram, A.; Talirz, L.; Haranczyk, M.; Stylianou, K. C.; Smit, B. Capturing chemical intuition in synthesis of metal-organic frameworks. *Nat. Commun.* **2019**, *10*, 539.
(110) Nigam, A.; Friederich, P.; Krenn, M.; Aspuru-Guzik, A. Augmenting genetic algorithms with deep neural networks for exploring the chemical space. *arXiv (Computer Science.Neural and Evolutionary Computing)*, January 15, 2020, 1909.11655, ver. 4. https://arxiv.org/abs/1909.11655 (accessed 2020-08-24).
(111) Elton, D. C.; Boukouvalas, Z.; Fuge, M. D.; Chung, P. W. Deep Learning for Molecular Generation and Optimization - a Review of the State of the Art. *arXiv (Computer Science.Machine Learning)*, May 22, 2019, 1903.04388, ver. 3. https://arxiv.org/abs/1903.04388 (accessed 2020-08-24).
(112) Silver, D.; et al. Mastering the game of Go with deep neural networks and tree search. *Nature* **2016**, *529*, 484−489.
(113) Zhang, X.; Zhang, K.; Lee, Y. Machine Learning Enabled Tailor-Made Design of Application-Specific Metal−Organic Frameworks. *ACS Appl. Mater. Interfaces* **2020**, *12*, 734−743.
(114) Schwaller, P.; Laino, T.; Gaudin, T.; Bolgar, P.; Hunter, C. A.; Bekas, C.; Lee, A. A. Molecular transformer: A model for uncertainty-calibrated chemical reaction prediction. *ACS Cent. Sci.* **2019**, *5*, 1572−1583.
(115) Segler, M. H.; Preuss, M.; Waller, M. P. Planning chemical syntheses with deep neural networks and symbolic AI. *Nature* **2018**, *555*, 604−610.
(116) Yang, X.; Zhang, J.; Yoshizoe, K.; Terayama, K.; Tsuda, K. ChemTS: an efficient python library for de novo molecular generation. *Sci. Technol. Adv. Mater.* **2017**, *18*, 972−976.
(117) Gómez-Bombarelli, R.; Wei, J. N.; Duvenaud, D.; Hernández-Lobato, J. M.; Sánchez-Lengeling, B.; Sheberla, D.; Aguilera-Iparraguirre, J.; Hirzel, T. D.; Adams, R. P.; Aspuru-Guzik, A. Automatic chemical design using a data-driven continuous representation of molecules. *ACS Cent. Sci.* **2018**, *4*, 268−276.
(118) Kusner, M. J.; Paige, B.; Hernández-Lobato, J. M. Grammar variational autoencoder. *arXiv (Statistics.Machine Learning)*, March 6, 2017, 1703.01925, ver. 1. https://arxiv.org/abs/1703.01925 (accessed 2020-08-24).
(119) Simonovsky, M.; Komodakis, N. GraphVAE: Towards generation of small graphs using variational autoencoders. *Lect. Notes Comput. Sci.* **2018**, *11139*, 412−422.
(120) Noh, J.; Kim, J.; Stein, H. S.; Sanchez-Lengeling, B.; Gregoire, J. M.; Aspuru-Guzik, A.; Jung, Y. Inverse design of solid-state materials via a continuous representation. *Matter* **2019**, *1*, 1370−1384.
(121) Yao, Z.; Sanchez-Lengeling, B.; Bobbitt, N. S.; Bucior, B. J.; Kumar, S. G. H.; Collins, S. P.; Burns, T.; Woo, T. K.; Farha, O.; Snurr, R. Q.; Aspuru-Guzik, A. Inverse Design of Nanoporous Crystalline Reticular Materials with Deep Generative Models. *ChemRxiv* **2020**, DOI: 10.26434/chemrxiv.12186681.v1.
(122) Kim, B.; Lee, S.; Kim, J. Inverse design of porous materials using artificial neural networks. *Sci. Adv.* **2020**, *6*, No. eaax9324.
(123) Ley, S. V.; Fitzpatrick, D. E.; Ingham, R. J.; Myers, R. M. Organic synthesis: march of the machines. *Angew. Chem., Int. Ed.* **2015**, *54*, 3449−3464.
(124) de Almeida, A. F.; Moreira, R.; Rodrigues, T. Synthetic organic chemistry driven by artificial intelligence. *Nat. Rev. Chem.* **2019**, *3*, 589−604.
(125) Muraoka, K.; Sada, Y.; Miyazaki, D.; Chaikittisilp, W.; Okubo, T. Linking synthesis and structure descriptors from a large collection of synthetic records of zeolite materials. *Nat. Commun.* **2019**, *10*, 4459.
(126) Corma, A.; Moliner, M.; Serra, J. M.; Serna, P.; Díaz-Cabañas, M. J.; Baumes, L. A. A new mapping/exploration approach for HT synthesis of zeolites. *Chem. Mater.* **2006**, *18*, 3287−3296.
(127) Collins, C.; Dyer, M.; Pitcher, M.; Whitehead, G.; Zanella, M.; Mandal, P.; Claridge, J.; Darling, G.; Rosseinsky, M. Accelerated discovery of two crystal structure types in a complex inorganic phase field. *Nature* **2017**, *546*, 280−284.
(128) Raccuglia, P.; Elbert, K. C.; Adler, P. D.; Falk, C.; Wenny, M. B.; Mollo, A.; Zeller, M.; Friedler, S. A.; Schrier, J.; Norquist, A. J. Machine-learning-assisted materials discovery using failed experiments. *Nature* **2016**, *533*, 73−76.
(129) Kitson, P. J.; Marie, G.; Francoia, J.-P.; Zalesskiy, S. S.; Sigerson, R. C.; Mathieson, J. S.; Cronin, L. Digitization of multistep organic synthesis in reactionware for on-demand pharmaceuticals. *Science* **2018**, *359*, 314−319.
(130) Gromski, P. S.; Granda, J. M.; Cronin, L. Universal Chemical Synthesis and Discovery with 'The Chemputer'. *Trends Chem.* **2020**, *2*, 4−12.
(131) Sanderson, K. Automation: Chemistry shoots for the Moon. *Nature* **2019**, *568*, 577−580.
(132) Roch, L. M.; Häse, F.; Kreisbeck, C.; Tamayo-Mendoza, T.; Yunker, L. P.; Hein, J. E.; Aspuru-Guzik, A. ChemOS: An

orchestration software to democratize autonomous discovery. *PLoS One* **2020**, *15*, No. e0229862.
(133) Roch, L. M.; Häse, F.; Kreisbeck, C.; Tamayo-Mendoza, T.; Yunker, L. P. E.; Hein, J. E.; Aspuru-Guzik, A. ChemOS: Orchestrating autonomous experimentation. *Sci. Rob.* **2018**, *3*, eaat5559.
(134) Steiner, S.; Wolf, J.; Glatzel, S.; Andreou, A.; Granda, J. M.; Keenan, G.; Hinkley, T.; Aragon-Camarasa, G.; Kitson, P. J.; Angelone, D.; Cronin, L. Organic synthesis in a modular robotic system driven by a chemical programming language. *Science* **2019**, *363*, eaav2211.
(135) Epps, R. W.; Bowen, M. S.; Volk, A. A.; Abdel-Latif, K.; Han, S.; Reyes, K. G.; Amassian, A.; Abolhasani, M. Artificial Chemist: An Autonomous Quantum Dot Synthesis Bot. *Adv. Mater.* **2020**, *32*, 2001626.
(136) Häse, F.; Roch, L. M.; Kreisbeck, C.; Aspuru-Guzik, A. Phoenics: A Bayesian optimizer for chemistry. *ACS Cent. Sci.* **2018**, *4*, 1134−1145.
(137) Burger, B.; Maffettone, P. M.; Gusev, V. V.; Aitchison, C. M.; Bai, Y.; Wang, X.; Li, X.; Alston, B. M.; Li, B.; Clowes, R.; Rankin, N.; Harris, B.; Sprick, R. S.; Cooper, A. I. A mobile robotic chemist. *Nature* **2020**, *583*, 237−241.
(138) Coley, C. W.; Barzilay, R.; Jaakkola, T. S.; Green, W. H.; Jensen, K. F. Prediction of organic reaction outcomes using machine learning. *ACS Cent. Sci.* **2017**, *3*, 434−443.
(139) Schwaller, P.; Gaudin, T.; Lanyi, D.; Bekas, C.; Laino, T. Found in Translation": predicting outcomes of complex organic chemistry reactions using neural sequence-to-sequence models. *Chem. Sci.* **2018**, *9*, 6091−6098.
(140) Daeyaert, F.; Ye, F.; Deem, M. W. Machine-learning approach to the design of OSDAs for zeolite beta. *Proc. Natl. Acad. Sci. U. S. A.* **2019**, *116*, 3413−3418.
(141) Wilkinson, M. D.; et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* **2016**, *3*, 160018.
(142) Johansson, S.; Thakkar, A.; Kogej, T.; Bjerrum, E.; Genheden, S.; Bastys, T.; Kannas, C.; Schliep, A.; Chen, H.; Engkvist, O. AI-assisted synthesis prediction. *Drug Discovery Today: Technol.* **2020**, DOI: 10.1016/j.ddtec.2020.06.002.
(143) Patiny, L.; Zasso, M.; Kostro, D.; Bernal, A.; Castillo, A. M.; Bolaños, A.; Asencio, M. A.; Pellet, N.; Todd, M.; Schloerer, N.; Kuhn, S.; Holmes, E.; Javor, S.; Wist, J. The C6H6 NMR Repository: An Integral Solution to Control the Flow of Your Data from the Magnet to the Public. *Magn. Reson. Chem.* **2018**, *56*, 520−528.
(144) Tremouilhac, P.; Nguyen, A.; Huang, Y.-C.; Kotov, S.; Lütjohann, D. S.; Hübsch, F.; Jung, N.; Bräse, S. Chemotion ELN: an Open Source electronic lab notebook for chemists in academia. *J. Cheminf.* **2017**, *9*, 54.
(145) Jablonka, K. M.; Moosavi, S. M.; Asgari, M.; Ireland, C.; Patiny, L.; Smit, B. A Data-Driven Perspective on the Colours of Metal-Organic Frameworks. *ChemRxiv* **2020**, DOI: 10.26434/chemrxiv.13033217.v1.
(146) Huber, S. P. AiiDA 1.0, a scalable computational infrastructure for automated reproducible workflows and data provenance. *Sci. Data* **2020**, *7*, 300.
(147) Pizzi, G.; Cepellotti, A.; Sabatini, R.; Marzari, N.; Kozinsky, B. AiiDA: automated interactive infrastructure and database for computational science. *Comput. Mater. Sci.* **2016**, *111*, 218−230.
(148) Ongari, D.; Yakutovich, A. V.; Talirz, L.; Smit, B. Building a Consistent and Reproducible Database for Adsorption Evaluation in Covalent−Organic Frameworks. *ACS Cent. Sci.* **2019**, *5*, 1663−1675.
(149) Jain, A.; Ong, S. P.; Chen, W.; Medasani, B.; Qu, X.; Kocher, M.; Brafman, M.; Petretto, G.; Rignanese, G.-M.; Hautier, G.; Gunter, D.; Persson, K. A. FireWorks: a dynamic workflow system designed for high-throughput applications. *Concurrency Comput.: Pract. Exper.* **2015**, *27*, 5037−5059.
(150) Haghighatlari, M.; Vishwakarma, G.; Altarawy, D.; Subramanian, R.; Kota, B. U.; Sonpal, A.; Setlur, S.; Hachmann, J. ChemML: A machine learning and informatics program package for the analysis, mining, and modeling of chemical and materials data. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2020**, *10*, No. e1458.
(151) Dunn, A.; Wang, Q.; Ganose, A.; Dopp, D.; Jain, A. Benchmarking Materials Property Prediction Methods: The Matbench Test Set and Automatminer Reference Algorithm. *npj Comput. Mater.* **2020**, *6*, 138.
(152) Lou, P.; Jimeno Yepes, A.; Zhang, Z.; Zheng, Q.; Zhang, X.; Li, C. BioNorm: deep learning-based event normalization for the curation of reaction databases. *Bioinformatics* **2020**, *36*, 611−620.
(153) Rehm, G. et al. QURATOR: Innovative Technologies for Content and Data Curation. *arXiv (Computer Science.Digital Libraries)*, April 25, 2020, 2004.12195, ver. 1. https://arxiv.org/abs/2004.12195 (accessed 2020-08-24).
(154) Lee, K.; Famiglietti, M. L.; McMahon, A.; Wei, C.-H.; MacArthur, J. A. L.; Poux, S.; Breuza, L.; Bridge, A.; Cunningham, F.; Xenarios, I.; Lu, Z. Scaling up data curation using deep learning: An application to literature triage in genomic variation resources. *PLoS Comput. Biol.* **2018**, *14*, e1006390.
(155) Toniato, A.; Schwaller, P.; Cardinale, A.; Geluykens, J.; Laino, T. Unassisted Noise-Reduction of Chemical Reactions Data Sets. *ChemRxiv* **2020**, DOI: 10.26434/chemrxiv.12395120.v1.
(156) Jia, X.; Lynch, A.; Huang, Y.; Danielson, M.; Lang'at, I.; Milder, A.; Ruby, A. E.; Wang, H.; Friedler, S. A.; Norquist, A. J.; Schrier, J. Anthropogenic biases in chemical reaction data hinder exploratory inorganic synthesis. *Nature* **2019**, *573*, 251−255.
(157) Zhang, D.; Lu, L.; Guo, L.; Karniadakis, G. E. Quantifying total uncertainty in physics-informed neural networks for solving forward and inverse stochastic problems. *J. Comput. Phys.* **2019**, *397*, 108850.
(158) Janet, J. P.; Liu, F.; Nandy, A.; Duan, C.; Yang, T.; Lin, S.; Kulik, H. J. Designing in the face of uncertainty: Exploiting electronic structure and machine learning models for discovery in inorganic chemistry. *Inorg. Chem.* **2019**, *58*, 10592−10606.
(159) Krenn, M.; Häse, F.; Nigam, A.; Friederich, P.; Aspuru-Guzik, A. SELFIES: a robust representation of semantically constrained graphs with an example application in chemistry. *arXiv (Computer Science.Machine Learning)*, March 5, 2020, 1905.13741, ver. 2. https://arxiv.org/abs/1905.13741 (accessed 2020-08-24).
(160) Mardt, A.; Pasquali, L.; Wu, H.; Noé, F. VAMPnets for deep learning of molecular kinetics. *Nat. Commun.* **2018**, *9*, 5.
(161) Jinnouchi, R.; Lahnsteiner, J.; Karsai, F.; Kresse, G.; Bokdam, M. Phase transitions of hybrid perovskites simulated by machine-learning force fields trained on the fly with Bayesian inference. *Phys. Rev. Lett.* **2019**, *122*, 225701.
(162) Singraber, A.; Behler, J.; Dellago, C. Library-based LAMMPS implementation of high-dimensional neural network potentials. *J. Chem. Theory Comput.* **2019**, *15*, 1827−1840.
(163) Raissi, M.; Perdikaris, P.; Karniadakis, G. E. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *J. Comput. Phys.* **2019**, *378*, 686−707.
(164) Meng, X.; Li, Z.; Zhang, D.; Karniadakis, G. E. PPINN: Parareal physics-informed neural network for time-dependent PDEs. *Comput. Methods Appl. Mech. Eng.* **2020**, *370*, 113250.
(165) Kutz, J. N. Deep learning in fluid dynamics. *J. Fluid Mech.* **2017**, *814*, 1−4.
(166) Wang, A. Y.-T.; Murdock, R. J.; Kauwe, S. K.; Oliynyk, A. O.; Gurlo, A.; Brgoch, J.; Persson, K. A.; Sparks, T. D. Machine Learning for Materials Scientists: An Introductory Guide Toward Best Practices. *Chem. Mater.* **2020**, *32*, 4954−4965.
(167) Rudin, C. Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. *Nat. Mach Intell.* **2019**, *1*, 206−215.
(168) Gunning, D.; Stefik, M.; Choi, J.; Miller, T.; Stumpf, S.; Yang, G.-Z. XAI—Explainable Artificial Intelligence. *Sci. Rob.* **2019**, *4*, eaay7120.