Machine Learning for the Discovery, Design, and Engineering of Materials

Chenru Duan,[1,2,#], Aditya Nandy,[1,2,#] and Heather J. Kulik[1,*]

[1]*Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139*

[2]*Department of Chemistry, Massachusetts Institute of Technology, Cambridge, MA 02139*

[#]These authors contributed equally.

*Corresponding author: phone: 617-253-4584, email: hjkulik@mit.edu

email addresses: crduan@mit.edu (C.D.), nandy@mit.edu (A.N.), hjkulik@mit.edu (H.J.K.)

ORCIDs: 0000-0003-2592-4237 (C.D.), 0000-0001-7137-5449 (A.N.), 0000-0001-9342-0191 (H.J.K.)

Keywords: machine learning, artificial intelligence, density functional theory, high-throughput screening, transition metal chemistry, organometallics

ABSTRACT: Machine learning (ML) has become a part of the fabric of high-throughput screening and computational discovery of materials. Despite its increasingly central role, challenges remain in fully realizing the promise of ML. This is especially true for the practical acceleration of the engineering of robust materials and the development of design strategies that surpass trial and error or high-throughput screening alone. This Review covers recent advances in algorithms and in their application that are starting to make inroads toward: i) the discovery of new materials through large-scale enumerative screening, ii) the design of materials through identification of rules and principles that govern materials properties, and iii) the engineering of practical materials by satisfying multiple objectives. We conclude with opportunities for further advancement to realize machine learning as a widespread tool for practical computational materials design.

1. Introduction.

With ever-increasing computing power and developments in algorithms in the past few years, the face of computational materials science has evolved significantly. In the mid-2000s, accurate but low-cost first-principles, density functional theory (DFT) became a widespread tool for the high-throughput screening and discovery of new materials(1-3). In the present era, machine learning (ML) methods have benefitted from growing data sets to overhaul computational materials science again(4). Nowhere is this more true than in efforts to uncover new materials. While many ML models, and even their application to challenging spaces such as transition-metal catalysis, were established in prior decades(5-8), it is the availability of larger data sets(9) and easy to use tools(10, 11) that have particularly transformed this landscape. Nevertheless, the number of ways ML can be applied in computational materials science is diverse. In some cases, the benefit of ML is transparent by reducing computational cost, whereas in other cases, traditional physics-based modeling may be more expedient, predictive, and interpretable. Alternatively, in cases where no physics-based modeling protocol is established, ML can be used to bring computational materials science closer to counterpart experimental efforts, such as by predicting quantities related to synthesis(12-16) or materials stability(17). For each property or material of interest, the optimal approach may vary. In this Review, we will explore representative demonstrations of ML in computational materials science as well as outstanding challenges. A comprehensive discussion of the historical relationship between physics-based modeling and the incorporation of ML in transition-metal chemistry is provided in Ref. (18).

First, we define three key areas of opportunity in computational materials science for the application of data science and ML algorithms: i) to discover, ii) to design, and iii) to engineer new materials. To strictly delineate these three terms is challenging, as they have some overlap that is increased when ML models are brought into consideration (Figure 1). We define *discovery* as the approach to characterize enumerated materials until a “best” material is found, a process which ML regression models can greatly accelerate (Figure 1). The *design* of materials is more narrowly defined than discovery here, as it refers to the determination of a structure–property relationship and its use to tune a target property (Figure 1). ML models and features can accelerate this design process by providing more complex and accurate structure–property mappings than simple heuristics (Figure 1). We distinguish *engineering* of materials as an evolution of the design paradigm, wherein development of practical and robust materials is carried out through automated optimization of one or more properties (Figure 1). Here, ML models are essential to accelerate the optimization of materials in high-dimensional spaces to find “needles” in “haystacks” (Figure 1). We focus in this review on transition-metal-containing complexes and porous metal–organic frameworks (MOFs) due to the combination of challenge and promise that their materials space imparts. For these materials, computational materials science has until recently required accurate first-principles DFT or bespoke force fields for property prediction, but recent demonstrations have shown them to be tractable targets for ML. We also touch upon other materials (e.g., organic molecules) and efforts based on experimental data sets.

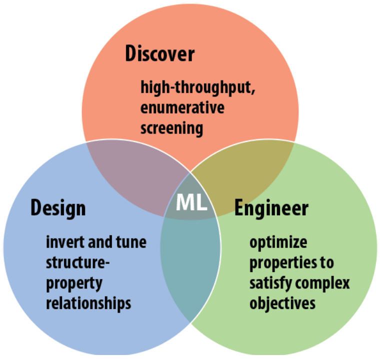
Discover
high-throughput, enumerative screening
Design
invert and tune structure-property relationships
ML
Engineer
optimize properties to satisfy complex objectives

Figure 1. Venn diagram of three approaches to computational materials science that can be aided by machine learning (ML): discovery that is accelerated by high-throughput DFT or ML models; designing by identifying and inverting structure–property relationships; and engineering by using optimization algorithms to satisfy multiple trade-offs, accelerated with ML models.

2. Discovery through Accelerated High-throughput Screening.

The discovery of materials as a fundamental goal of scientists and engineers is not new. One key difference from earlier serendipity-driven discovery efforts is the increasing availability of automation strategies(1-4, 19-24) and related large repositories of experimental data(25) and computation-ready datasets(9, 26-30) (e.g., for materials repurposing). These advances have come in combination with increased computing power for physics-based (e.g., DFT) models as well as the development of machine learning models trained on these large data sets for rapid property prediction. As a result, the scale of new materials that can be screened for discovery has been transformed from hundreds by trial and error to thousands or millions. Here, we outline strategies for enumerating materials to discover lead compounds that satisfy specific or as-yet unknown property targets. We describe how pairing discrete enumerated spaces with ML models can be used to reveal trends that would have otherwise been missed in smaller data sets. We conclude this section by presenting some remaining challenges for ML-accelerated materials discovery.

2.1. Approaches for Screening and Enumeration.

Screening of available experimental or enumerated data sets can be accelerated with ML regression models. In this approach, some traditional calculation or experimentation must be carried out on a sufficiently representative data set to train the ML model so that it can make predictions on the remaining, unseen compounds. For many models, it is expected that the ML model will interpolate between unseen compounds better than it will extrapolate, but a measure of chemical similarity or dissimilarity must be available to make this assessment. Chemical fingerprints (i.e., matching the presence of specific functional groups) or ML representations can be used to define similarity. ML representations obtained from feature selection (see Sec. 3) are particularly useful for this task.

Experimental data sets for the direct screening of transition-metal complexes and closely related metal–organic framework (MOF) materials are relatively small and scarce. The largest data set of experimentally-characterized metal-organic (e.g., MOFs or TMCs) compounds is the set of structures obtained from X-ray diffraction in the Cambridge Structural Database (CSD)(25). Geometric structures of many transition-metal complexes are available (ca. 150k or more) in the CSD(25). Chung and coworkers have refined a subset of the CSD believed to contain porous MOFs in a set known as the computation ready experimental MOF dataset (CoRE MOF), which has recently been updated to contain over 14k MOFs.(29) Such sets are typically augmented with simulation data, e.g., as is the case for the 86k structures in the tmQM dataset of DFT and low-cost, semi-empirical properties(31) or for specific target properties such as gas adsorption in MOFs(29) (see Sec. 2.2).

Because experimental data sets are both small and limited in available properties, enumeration has played an important role in data generation for ML models and ML-based screening (Figure 2). Focusing on transition-metal complex enumeration, symmetry has often been exploited(22, 26, 32, 33) (i.e., by requiring that ligands of a transition-metal complex be identical), in part because higher-symmetry complexes are more frequently synthesized (e.g., as indicated by frequency in the CSD). Even with such constraints, the enumeration of all feasible compounds can rapidly lead to a larger space (ca. billions) than is tractable to study with direct computation or even ML regression model-accelerated screening. As a result, the materials space is typically constrained for the target application and to reduce the combinatorial challenge of enumeration. Ways in which this constraint can be applied include limiting the number of components that are varied, such as the possible elements in binary or ternary solid-state alloys(34) or the ligands in a transition-metal complex(26) (Figure 2). The range of system sizes(35) or types of geometric arrangement(27, 28) can also be constrained to reduce the search space.

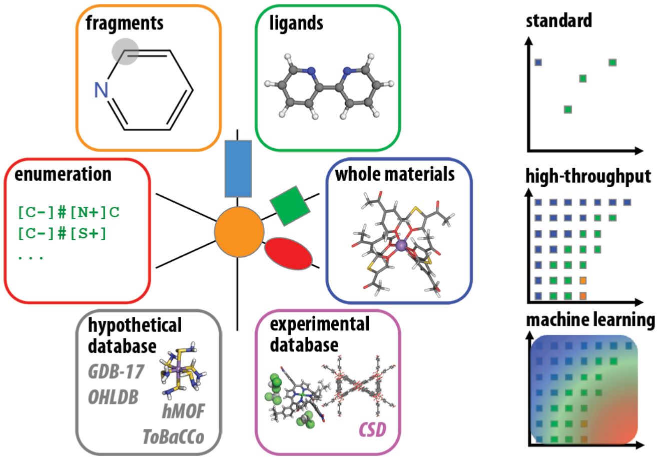

Figure 2. (Left) Types of data set development approaches that can be used in combination for for discovery of transition-metal complexes and metal–organic materials (schematic of a metal and three ligand types in an octahedral metal environment shown in inset). Approaches include: enumeration (e.g., of SMILES strings) with weak constraints on allowed chemistry to generate molecules, fragments of stable motifs for subsequent fusion (e.g., indicated by gray circle in inset), from common ligands, and data sets of entire materials. Databases generated from hypothetical enumeration of molecules (e.g., GDB-17 organic molecules(35) or OHLDB transition-metal complexes(26)) and MOFs (e.g., hMOF(27) or ToBaCCo(28)) can be subsequently screened. Whole materials are often extracted from these databases or from experimental databases such as the CSD(25). (Right) Schematic of how scale of a standard screen (top) compares to high-throughput (middle) and machine learning interpolation (bottom) for screening studies, with an equivalent property color scale from blue to green to red used in all three panes.

Enumerative strategies include ones that define connectivity (e.g., from a simplified molecular input line entry string, SMILES) based on heuristic constraints including satisfaction of the octet rule(35), although some benefit has been found to weakening heuristic constraints(26). The GDB-17(35) set of organic molecules of up to 17 heavy atoms with a limited number (i.e., C, N, O, S and halogen) of elements contains 66 billion compounds, and a subset of it formed the basis of the QM9 dataset widely used for organic molecule ML(9). When aiming to generalize such an approach to small transition-metal complexes, Gugler et al.(26) relaxed constraints on the octet rule to sample a wider array of chemical bonding in ligands while constraining octahedral complexes to no more than 13 heavy (i.e., C, N, O, P, or S) atoms.

Another strategy is to constrain the space of enumerated compounds to ones where fragments are based on experimentally synthesizable and stable materials(20, 36, 37). For example, stable five- and six-membered carbon-containing conjugated rings with a limited number (i.e., 1 or 2) heteroatoms (e.g., N, O, S) are common motifs in many inorganic materials. Through combinatorial fusion and subsequent functionalization, Janet et al.(37) constructed a space of 2.8M homoleptic transition-metal complexes. Yet another strategy is to constrain enumeration to common ligands based on the frequency of their occurrence in inorganic complexes in the CSD and to assemble them into more varied combinations and with distinct metal centers(22, 32, 33) (Figure 2).

For MOF screening, enumeration of large sets of feasible compounds is typically carried out by leveraging the reticular structure of a MOF(27) (Figure 2). A MOF consists of the inorganic secondary building unit (SBU, i.e., a metal node), linkers, and functional groups. An additional consideration for MOFs is the favored topological net (i.e., the crystal structure formed)(28), although early enumeration focused on simple cubic unit cells(27). Combinatorial enumeration with small changes in linker chemistry (e.g., functional groups) and metal SBUs can result in large data sets of 100k or more MOFs(27, 28, 38, 39), although how well such sets represent the diversity of synthetically accessible materials has been questioned(40) (see Sec 2.3). Similar strategies have also been demonstrated in the enumeration of zeolites(41).

2.2. Demonstrations of Enumerative Materials Screening.

Once data sets of hypothetical materials(34), molecules (e.g., QM9(9) or OHLDB(26)) and MOFs (e.g., hMOF(27), BW-DB(38), or ToBaCCo(28)) or their experimentally based (e.g., CoRE-MOF(29) or tmQM(31)) counterparts are assembled, their sizes can easily number in the thousands to hundreds of thousands of unique compounds (Figure 2). As a result, these sets can be employed both for direct high-throughput screening and as training sets for ML models such as artificial neural networks (ANNs) or kernel-based (e.g., Gaussian process regression or kernel ridge regression, KRR) models.

Once an ML model is trained, enumerative screening of a larger, related set of compounds can be beneficial to reveal trends in structure–property relationships. The power of dataset enumeration for virtual screening was realized over 15 years ago by Rothenberg and coworkers(42), but the high computational cost of virtual screening with DFT at the time limited the scope of approaches based on computational chemistry, instead favoring the use of experimental results.(6) For example, after training an ANN on 412 experimental catalysts and reaction conditions, 66,000 new combinations were then predicted with the trained ANN.(6)

More recently, Nandy et al.(43) trained an ANN on an 80/20 train/test split of 712 open-shell transition-metal complexes for the prediction of the reaction energetics for metal-oxo formation to a test set accuracy of 5.5 kcal/mol. To identify metal- and spin-state-dependent trends, they enumerated and predicted properties of a set of nearly 10k catalysts comprised of the original 29 ligands in the dataset (Figure 3). Because only 5% of these compounds were previously seen in model training, the revealed trends provide a more general interpretation, despite being inherently interpolative in nature (Figure 3). Key observations included the strong dependence on metal and spin state of the stability of formed terminal metal-oxos (Figure 3). A similar approach was carried out by Liu et al.(44), in this case to identify metal and spin-state dependence of multireference (MR) character (i.e., as judged through ratios of the non-dynamic and total correlation diagnostics(45), r_{ND}, from fractional occupation DFT). Starting from a data set of 4,865 complexes, an ANN trained to predict the multireference diagnostic r_{ND} to 3% mean absolute percent error (i.e., 0.018) was applied to a theoretical space of 187.2k new complexes, only 1% of which had been seen in training data. Through this analysis, Liu et al. observed(44) that although MR character is expected to be higher in low-spin states, this trend is only seen for some (i.e., Cr or Mn) metals (Figure 3).

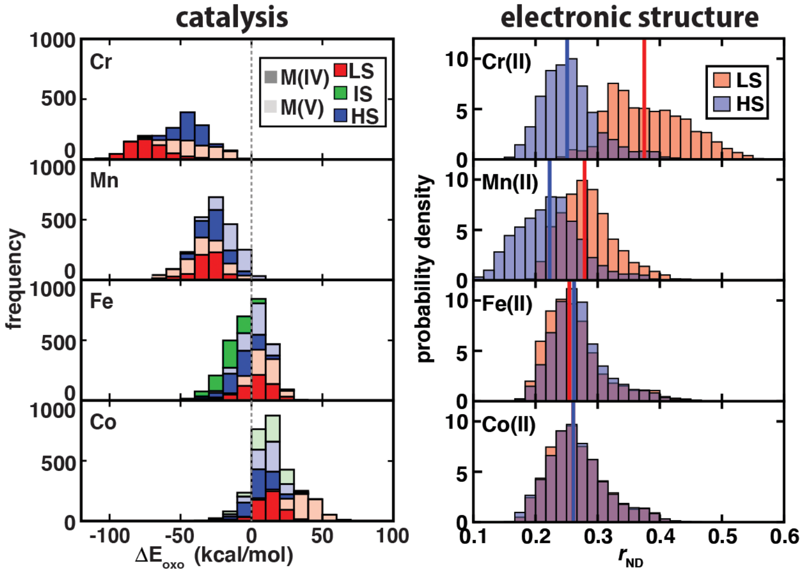

Figure 3. Examples of trends revealed for catalysis (left) and degree of multireference character in electronic structure (right) from enumerative scoring of materials from Refs. (43, 44). (Left) Distribution of oxo formation energies (ΔE$_{oxo}$, in kcal/mol, bin size: 10 kcal/mol) as predicted by an ANN from Ref. (43) for 9,860 transition-metal complexes with four metals in up to three spin and two oxidation states. Unnormalized counts are shown on the y-axis, and histogram is colored by spin (red for low-spin, LS, green for intermediate-spin, IS, and blue for high-spin, HS). The stacked histogram is shaded by oxidation state, with oxidation state +5 complexes represented by translucent coloring, and oxidation state +4 complexes represented by opaque coloring. (Right) Normalized probability density distribution of r_{ND} (unitless, bin size 0.0167) as predicted by the ANN from Ref. (44) for 11,700 M(II) complexes (M = Cr, Mn, Fe, Co) in two spin states. The histograms are colored by spin (red for low-spin, LS, and blue for high-spin, HS). The median of each distribution is indicated by a vertical line with the same corresponding color. Adapted with permission from Refs. (43, 44). Copyright 2019 and 2020 American Chemical Society.

Enumerated sets can improve ML model generalization in comparison to models based on sets of transition-metal complexes formed from commonly-studied ligands. For example, the OHLDB set from Gugler et al.(26) of small transition-metal complexes with diverse metal-coordination environments was used to improve the generalization of earlier ANN models for property prediction. Incorporation of OHLDB compounds reduced errors of ML models (i.e., ANNs) on experimental (i.e., CSD) complexes by around 25%.(26, 46) This improvement was observed despite the larger size of the CSD complexes, because the two sets shared similarities in direct metal-coordination environment.

Enumerative screening has been most widely exploited for screening MOF materials for gas adsorption and separation. While the hypothetical space of MOF materials is large, the cost of classical molecular mechanics (MM) calculations to determine properties such as surface area or gas uptake are relatively low in comparison to DFT calculations required for catalysis and other applications. Demonstrations have included the screening of the over 13k MOFs in the ToBaCCo set(28) for hydrogen storage(47) or other similarly large MOF sets for methane storage(39, 48).

2.3. Outstanding Challenges for Enumeration in Transition-Metal Chemistry.

Despite the power of enumerative screening, caveats should be noted. First, once spaces have been enumerated, high-throughput screening with DFT is most widely used to collect property information. However, each discrete calculation requires considerable computational effort and can lead to failed outcomes if the DFT calculation of the new, hypothetical material does not converge(49) (Figure 2). The ML models that interpolate a space from discrete DFT results may have the least available information where calculations are unsuccessful and thus may be the most unreliable there (Figure 2). It may be necessary to validate the ML-interpolated and DFT-level predictions with experiment to ensure predictions are realistic and materials are synthesizable, as was done for organic light-emitting diodes(50) or MOFs for wet flue gas capture(51). Additionally, the choice of data selected for training the ML model can bias its predictions. Moosavi et al.(40) noted that KRR models trained to predict gas separation characteristics (e.g., CO_2 adsorption) on different data sets (e.g., experimental CoRE-MOF(29) vs hypothetical BW-DB(38) or ToBaCCo(28)) were not transferable. The influence of data set bias on ML-driven design will be discussed next in Sec. 3. Overall, despite some caveats, ML-driven enumerative screening is a powerful approach to reveal trends in large sets that may not have been obvious with the standard approach (Figure 2).

3. Design of Materials via Abstracted Principles.

In contrast with discovery, we distinguish materials design as the process by which structure–property relationships are inferred over a dataset and then often used to iteratively improve upon known designs. ML models, analysis of their key features(32, 52), and dimensionality reduction (i.e., unsupervised learning)(53-56) represent important tools for building structure–property relationships and mapping chemical space(57). Despite their recent prominence, it is worth noting that closely related quantitative structure–property or structure–activity relationships, typically involving linear models, have long been used in transition-metal catalysis(58-60). We describe approaches to building and interpreting primarily ML-based structure–property relationships and then discuss recent demonstrations and remaining challenges.

3.1. Approaches for Inferring and Interpreting Structure–Property Relationships.

A key ingredient in building structure–property relationships is the development of descriptors or representations that encode the key chemical features that dictate materials properties. We use the term descriptors to refer to *ad hoc* features that have been heavily engineered or hand-picked (Figure 4). We use the term representations to indicate higher-dimensional feature vectors that are most beneficial in non-linear models of increasing complexity (Figure 4). The selection of essential descriptors or features is an approach to building structure–property relationships. In multiple linear regression (MLR) models, descriptors may be selected one at a time either by identifying those that individually correlate with the desired property or by adding them only if they reduce the error of the fit (Figure 4). It is recommended that cross-validation error on subsets of training data be used for descriptor selection because it reduces overfitting that might occur if goodness of fit is used as the metric instead (Figure 4). In a similar fashion, one-shot models such as least absolute shrinkage and selection operator (LASSO) or random forests are widely used both as ML models for property prediction(61) and for selection of important features(22, 32, 40, 43) (Figure 4). In random forest models, predictions are made based on a series of binary decision trees. The effect of omitting trees with specific features can provide a measure of the relative importance of features, and the features that have a limited effect on the error of the model can be omitted (Figure 4). LASSO is an MLR model with a modified loss function that sets uninformative features to zero via regularization (Figure 4).

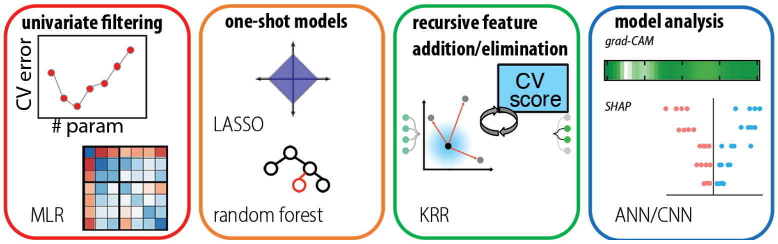

Figure 4. Different approaches to feature selection (top) and the associated model complexity/type (bottom) with increasing complexity from left to right: multiple linear regression selection by cross-validation or comparison of univariate correlations; one-shot feature selection and model fitting in LASSO and random forest; recursive feature addition with selection by CV error in more complex models such as KRRs; and model interpretation of deep neural networks such as gradient-weighted class-activation map (grad-CAM) and Shapley additive explanations (SHAP) scores.

While these models are widely used and fast to train, kernel models and deep learning (e.g., ANN) models can generally achieve lower errors on comparable data sets as long as the data set sizes number in the hundreds of points(6, 32, 33). Kernel ridge regression (KRR) or Gaussian process models inherently encode structure–property relationships by weighting the predictions of new compounds by the most proximal (i.e., similar) molecules in the relevant feature space. Thus, good prediction by a KRR model suggests a good feature set has been chosen. As a result, recursive feature addition (RFA) or elimination (RFE) can be used to identify the best features as judged by KRR test set errors (Figure 4). Because exhaustive enumeration of all possible combinations of initial features for the RFA along with model hyperparameter selection is prohibitive, pre-ordering features for addition(22), e.g., with random forest, is one preferred alternative (Figure 4). When KRR models are applied with tailored feature sets, their prediction accuracy can rival or surpass(22) deep learning (e.g., fully connected ANNs or convolutional neural networks, CNNs) models. Nevertheless, ANNs can be preferable both in the case of larger data sets and in chemical discovery applications either with enumeration of large spaces (see Sec. 2) or in materials optimization (e.g., with active learning, see Sec. 4).

Although typically viewed as ‘black box’, interpretation of ANNs has been achieved in a number of ways in recent years (Figure 4). The response of an ANN model to changes in variables measured with Shapley Additive Explanation (SHAP) scores(62) can guide feature importance. In the case of CNNs, other approaches include identifying model focus (e.g., with gradient weighted class-activation map or grad-CAM(63), Figure 4). It is worth noting that an ANN consists of a series of layers that essentially carry out feature selection, because each layer manipulates the space in which the data is represented and reorients points closer or further to each other. From the last layer of the model known as the final latent space, a linear regression is typically carried out to make the property prediction. As a result, analysis of latent space distances has been useful in materials design(64) and in identifying model uncertainty by determining support of a prediction on a new compound with respect to similar available data(46, 49) or finding the most useful places for data enrichment(65).

Dimensionality reduction is another approach to simplify analysis of compounds either in feature space or model (i.e., latent) space. Formally an unsupervised learning technique because the data distribution is learned on unlabeled data, dimensionality reduction can be beneficial to identifying the most essential descriptors that distinguish compounds. Principal component analysis (PCA) projects a high-dimensional space into a smaller space spanned by the key combinations of features. When most of the variance of the data can be explained by the first few PCs, they can act as a useful reduced feature set(54, 66, 67). Other methods such as t-distributed stochastic neighbor embedding (t-SNE)(68) or uniform manifold approximation and projection (UMAP)(69) aim to preserve apparent pairwise distances approximately in a reduced two-dimensional space so that it can be readily visualized. Like PCA, UMAP dimensions can also be used as a feature set, whereas those from t-SNE cannot. These techniques can be applied to the latent space of models just as readily as the feature space, for example, to highlight regions of promise for enrichment(37, 65) or to reveal governing design principles(64).

3.2. Developing Structure–Property Relationships with Machine Learning.

Identifying key descriptors or features in a structure–property relationship mapping has been demonstrated as an approach to revealing design principles(18, 22, 32, 52) and enabling iterative design(70). Early efforts in this area can be traced back to the use of linear models and descriptors.(71) While descriptors and representations have evolved with the increasing use of more sophisticated ML models in recent years, demonstrations in the early 2000s(5-8, 42) bear striking similarity to many modern applications.(32, 43, 72-74) Nevertheless, the source and scope of data set sizes has evolved. Closely related inverse design strategies are outlined in Ref. (75).

Because structural data of transition-metal complexes has been widely available for some time (e.g., from the CSD(25)), many early descriptors were motivated by analyzing CSD structures(76), even prior to the development of affordable computational modeling (e.g., with DFT). In this case, successful descriptors were generally demonstrated in single or multiple linear regression models(59, 60). These include relationships between catalysis or ligand binding and the steric bulk (e.g., Tolman cone angle(77, 78) or buried volume(79)) of a ligand. Other 3D-focused approaches include the development of maps of electrostatics with probe molecules(71), an approach which has recently gained favor in deep learning models for predicting gas storage characteristics of MOFs(80), or experimental homogeneous catalyst activities(81), where it is sometimes called stereocartography(82). In open-shell transition-metal complexes, the metal–ligand bond length is a sensitive reporter of the molecule's experimental spin state. The likelihood of a transition-metal complex to exhibit spin-crossover behavior has been predicted based on the ligand nitrogen–nitrogen separation(83), and similar features have been invoked for predicting single molecule magnet behavior.(84) An ANN trained on topological descriptors to predict DFT metal–ligand bond lengths(52) was used to classify experimental spin states(85) and even correct DFT-based energetic predictions of spin states.

As quantum chemical modeling became tractable, first-principles descriptors were used, with frontier orbital energies or partial charges playing an important role in building transferable models capable of predicting experimental catalyst activities(59, 61). Fey and coworkers augmented sets of CSD ligands with quantum chemical descriptors to further distinguish their properties(53, 54). In some cases, deep learning (e.g., ANN) models have been developed to predict frontier orbital energies(22, 43, 86), which can be applied to explain experimental activity of catalysts known to be correlated to such energies(86). Alternative first-principles reactivity descriptors(87), such as those measuring interaction between catalysts and reactants(88), have also been fruitfully employed with deep learning models to predict experimental catalytic activities.

Importantly, interrogation of the key descriptors, even in the case of a limited data set, can provide further insight into what maximizes catalytic activity(89).

Beyond trial-and-error, *ad hoc* selection of descriptors, rigorous feature selection with ML models has emerged as a powerful tool for inferring design principles(22, 32, 43, 52, 72, 90, 91). Low-cost, topological(5, 7, 32, 92-94), three-dimensional(95-100), one-hot chemical encoding(92, 101), and geometric descriptors (e.g., for MOF pores)(102-104) are often employed for this purpose in transition-metal chemistry. Applying random-forest-ranked recursive feature addition (RF-RFA) to KRR models, expected heuristics have been reinforced by feature selection approaches(22, 32, 105), such as the spin splitting of a complex being strongly dependent on the metal-local electronic environment (e.g., from the electronegativities of surrounding atoms, Figure 5). Conversely, other properties such as the frontier orbital energies(22), band gap(44, 105), or ionization/redox potential(32, 52) have been found to depend much more on the global structure of the molecule (Figure 5). Such significant differences suggest opportunities for orthogonal design of the two properties(52). This type of analysis has also demonstrated the circumstances in which some heuristics can be expected to fail. The HOMO level has been proposed to be important(106) in predicting metal-oxo formation in homogeneous catalysts, but feature selection has shown that HOMO energetics depend much less on the metal than does strongly metal-dependent oxo formation. The similarity of metal-local features for predicting metal-oxo formation and spin-state ordering also highlight why strong dependence of reaction energetics on spin-state and metal (see Sec. 2) can be expected in this class of catalysts(43). Importantly, in sufficiently balanced data sets(43, 105), these observations are relatively insensitive to the DFT functional on which the data was trained (Figure 5).

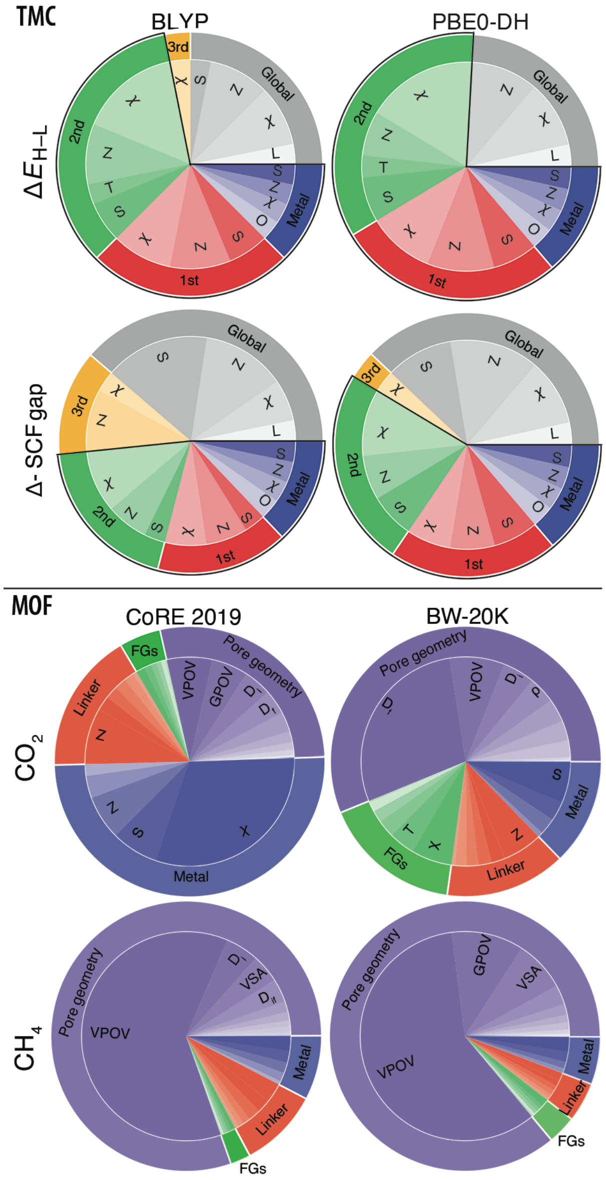

Figure 5. Feature analysis for representative materials and properties. (top) Transition-metal complexes (TMC) adiabatic spin splitting ($\Delta E_{\text{H-L}}$, top row) and HOMO–LUMO gap from Δ-SCF (bottom row) with revised autocorrelation functions (RACs) grouped by property (inset) and most distant atom in the RAC: metal (blue), 1st coordination sphere (red), 2nd coordination sphere (green), 3rd coordination sphere (orange), and more distant (global, gray) for the BLYP GGA (left) and the PBE0-DH double hybrid (right) on around 2000 TMCs from Ref. (105). (bottom) MOF grand canonical Monte-Carlo (GCMC) gas adsorption for CO_2 (top row) and CH_4 (bottom row) applied to the CoRE-MOF 2019 data set (left) and BW-20K hypothetical MOF data set (right). The MOF RACs and results from Ref. (40) show the features grouped by being metal-centered RACs (blue), linker-centered (red), functional group-centered (FGs, green), or derived from pore geometry characteristics (purple).

3.3. Outstanding Challenges in ML-accelerated Design Strategies.

The development of structure–property relationships is not without its challenges, most of which stem from the balance and size of data sets on which the relationships are built. For example, in experimental data sets extracted from the literature(14, 15, 19), there can be positive bias for reporting only successful experiments(16, 107, 108), leading to challenges in classification tasks(17). Additionally, the availability and consistency of reporting of key properties may be limited, with only structural data or related properties widely available except in very recent demonstrations(17). It has also been observed that experimental data sets (e.g., of MOFs in CoRE-MOF(29)) do not necessarily have the same characteristics as hypothetical sets (e.g., hMOF(27) or ToBaCCo(28)).(40) As a result, extracted design principles show a significant difference when obtained from feature importance analysis of ML models trained on these different sets (Figure 5).(40) Interestingly, the hypothetical sets do not necessarily fully include and go beyond the experimental sets into uncharted materials space. Instead, the experimental sets exhibit greater diversity of metal SBUs that are identified only in experimentally derived structure–property relationships to be important for gas adsorption. Thus, challenges for design stem from the challenges of enumerative discovery described in Sec. 2.

4. Engineering Materials to Address Challenges.

We next distinguish engineering of materials with ML by the act of optimization of one or more properties. Along with engineering of the activity, other practical considerations include stability(17), solubility(37), synthesizability(109), cost(74, 110), and optimal conditions (e.g., for reaction or operation)(111-113). Here, it may be necessary to work directly with experimental data from extracted literature reports(14, 15, 19, 114) or high-throughput experimentation(115, 116) because, unlike activity, there may not be a good way to model these factors due to their dependence on complex phenomena and often unknown factors. When multiple properties are being optimized, the search space naturally grows in a way that benefits from ML. We next describe some approaches and demonstrations of engineering materials through ML-accelerated optimization as well as the challenges that remain.

4.1. Single-objective Optimization and Iterative Approaches.

Optimization of a single property is feasible either with direct experimentation(117, 118) or physics-based modeling(119) but can often be accelerated with ML(43, 120, 121). Genetic algorithms are one automated approach to improve materials over successive "generations" (Figure 6).(122, 123) Genes suitable for optimizing transition-metal complexes or MOFs could be the metals or inorganic SBUs(118) as well as ligands/linkers(37, 43), or functional groups(124) (Figure 6). Discrete pools of each of these possible choices gives rise to a large space(125), similar to those discussed for full enumeration (see Sec. 2). At each generation, a series of operations is carried out, first to score the population to identify the fittest compounds and then to carry out mutation or crossover to ensure diversity of the population of optimizing materials (Figure 6). This approach proceeds until terminated, and the local nature of the optimization usually motivates a series of runs(120) to ensure all best leads are found. Jensen and coworkers(36, 126) have devised ring closure operations for transition-metal complexes, as implemented in their DENOPTIM code(126), that can alter ligand denticities to identify the optimal denticity (e.g., for spin-crossover compounds(36)).

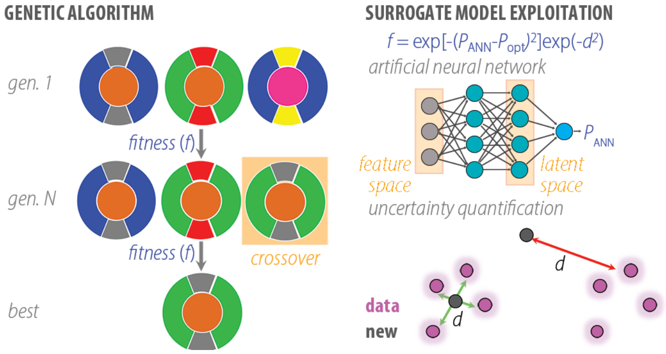
GENETIC ALGORITHM
gen. 1
fitness (f)
gen. N
fitness (f)
crossover
best
SURROGATE MODEL EXPLOITATION
$f = \exp[-(P_{ANN}-P_{opt})^2]\exp(-d^2)$
artificial neural network
feature space
latent space
P_{ANN}
uncertainty quantification
d
data
new
d

Figure 6. (Left) Schematic of a genetic algorithm applied to materials defined by three genes (circle and two surrounding arcs). At each generation, fitness is evaluated, and the fittest survive with crossover (shown) and mutation applied to encourage diversity. (Right) Demonstration of how a surrogate model can be exploited, e.g., in combination with a genetic algorithm. (top) The fitness function in a genetic algorithm uses the ANN for scoring in a composite function that penalizes points at a high distance, *d*, from training data (middle) A schematic of an ANN property prediction is shown with the input feature space and the latent space both highlighted. (bottom) The distance of a new point in either feature or latent space to available training data (pink) can be close (shown schematically at bottom left) or far (shown at bottom right).

Although fitness evaluation for each generation can be parallelized, the computational cost may remain high. As a result, use of genetic algorithms with experimental results typically involves efficient assays, for example using rapid colorimetric outputs(117). Because fitness evaluation dictates the overall time for optimization, multiple linear regression (i.e., quantitative structure–property relationship, or QSPR) models(59, 60), force field or semi-empirical models(36), or ML regression models (e.g., with ANNs)(120) are preferable in comparison to full DFT evaluation. The molSimplify automatic design (mAD)(22) and DENOPTIM codes(126) both flexibly apply any type of fitness evaluation, and mAD has a series of built in ANNs to ensure rapid evaluation(22). In addition to these automatic optimization procedures, iterative exploitation of developed structure–property relationships (see Sec. 3) can be used to identify ways to change one property without altering another(17, 52) (see Sec 4.3). Alternatively, kernel-based Gaussian processes (discussed in Sec 4.2) are also commonly used for optimization, especially in continuous spaces.

4.2. Multi-Objective Optimization Algorithms.

Optimizing multiple objectives for property engineering is often necessary in practice, and this is where acceleration through the use of ML models for search in high-dimensional spaces is most needed. For the optimization of multiple properties, a Pareto front represents a hypersurface that passes through the series of compounds that have the best combination of properties (e.g., price and activity(110), catalyst selectivity and conversion(127), or different measures of MOF surface area(128), Figure 7). To ensure enrichment at the Pareto front with ML-accelerated discovery, a number of active learning strategies have been proposed.(37, 129-132) In all cases, some initial pool of discrete candidates is curated (see Sec. 2), and initial DFT or experimental data is obtained (Figure 7). From this data, an ML model such as a Gaussian process (GP) or ANN may be trained to make predictions on the larger set of compounds in the full discrete space (Figure 7). Most optimization procedures then focus on identifying the new points that in combination provide the most information to the model and have the greatest promise of enriching or superseding the current Pareto set (Figure 7). When these approaches use Bayesian statistics to predict these quantities, they are referred to as Bayesian optimization.

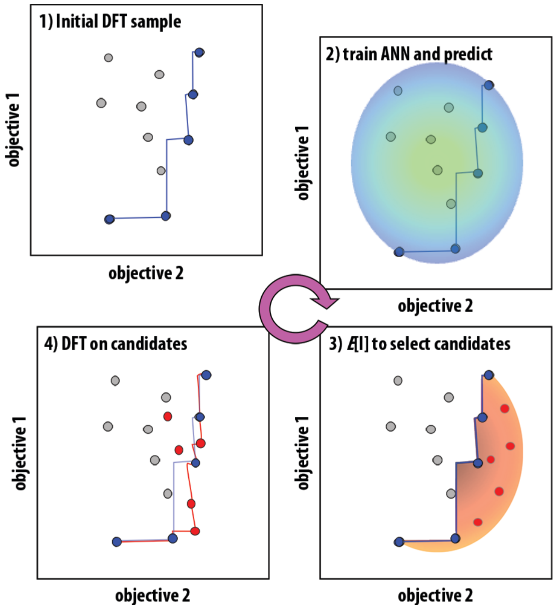

Figure 7. Example of steps in an active learning workflow (clockwise from top left): 1) initial data generation (e.g., with DFT, circle symbols) and determination of a Pareto front (blue lines and dots); 2) training and evaluation of a machine learning model (e.g., an ANN) across a large space of hypothetical compounds (smooth shaded space); 3) scoring and selection of compounds expected to go past the Pareto front (e.g., with expected improvement, *E*[I]); 4) further evaluation of selected candidates with DFT. The cycle, including ANN training and prediction, is repeated, as indicated by inset magenta arrow.

A specific type of efficient global optimization (EGO) affords a score, known as the expected improvement, *E*[I], based on both the predicted property from the ML model as well as the uncertainty of the ML model.(133) Once promising points have been identified by the ML model, they are then evaluated with DFT or experiment and used to retrain the model (Figure 7). Repeating this procedure ensures enrichment of data for the ML model as well as optimal leads close to the Pareto front.

4.3. Demonstrations of Materials Optimization.

Single-objective, ML-driven or QSPR-driven optimization with genetic algorithms has been widely demonstrated for both experimental and computational improvements in heterogeneous(134-136) and homogeneous(43, 137) catalysis, spin-crossover complexes(120), among others(121). The promise of these techniques is even greater when multiple tradeoffs are considered. ML-driven optimization has also been demonstrated in scanning optimal catalyst and reaction space combinations(138), selection of solvents to improve reaction outcomes(139), or in the design of experiments to optimize turnover number in CO_2 reduction catalysts(140).

Janet and coworkers used(37) EGO with the 2D expected improvement criterion for multi-objective optimization of complexes with earth-abundant (i.e., Cr, Mn, Fe, or Co) metals for redox flow battery applications. From a space of 2.8M candidate homoleptic transition-metal complexes that were sufficiently bulky (ca. 100–200 atoms) to be resistant to membrane crossover, they simultaneously improved the predicted solubility in non-polar aqueous electrolytes (i.e., as judged through logP) and the redox potential of the compounds (Figure 8). They showed that lookahead errors (i.e., predictions on future generations in the EGO run) were better with a multi-task ANN than with a GP model, suggesting the ANN generalized better to previously unseen compounds. The best compounds in terms of simultaneous solubility and redox characteristics had a common motif of a fusion of five- and six-membered O-coordinating rings around a high-spin Mn center with small polar groups (Figure 8). The need for small polar groups to increase solubility without reducing redox potential could have been inferred from the strong size dependence of oxidation potential that had been noted in prior design abstractions(32), but the ring structure was new (Figure 8). In five generations, the best-performing materials all significantly exceeded(37) those that could have been observed in diversity-oriented or random search-based screens (Figure 8). Properties were optimized at least 500-fold faster through the EGO algorithm, achieving design leads in weeks instead of decades. These performance improvements are consistent with observations of global optimization for other materials (e.g., transition-metal oxides(141), hybrid perovskites(142)), where the improvement with each property optimization dimension is generally multiplicative.

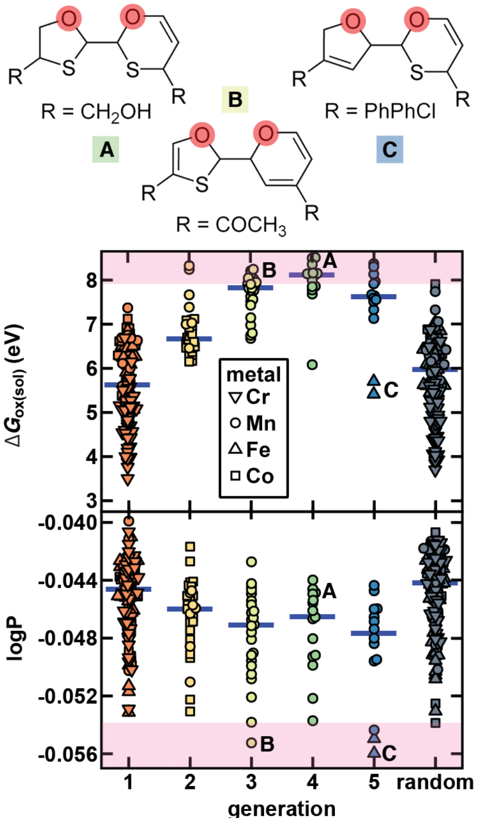

Figure 8. Distribution of $\Delta G_{\mathrm{ox(sol)}}$ (top) and logP (bottom) values for each generation (colors by generation and symbols as in inset) alongside a larger random sample (gray symbols). The mean

value for each generation is indicated with a blue horizontal line. The points exceeding those obtained from the random sample in performance are shown in a magenta shaded region. Three complexes are labeled and shown at top with each ligand shown as a skeleton structure: the highest $\Delta G_{\mathrm{ox(sol)}}$ Mn complex (A), the best trade-off Mn complex (B), and the highest-magnitude logP Fe complex (C). Adapted with permission from Ref. (37). Copyright 2020 American Chemical Society.

In some cases, ML-accelerated DFT-based or MM-based property optimization is insufficient, and researchers have extracted experimental data from the literature to guide materials engineering principles. This effort is most mature in extracting and predicting synthesis recipes(12-15, 19, 114) or other quantities (e.g., MOF surface area(143)) that are reported uniformly. Nandy and coworkers(17) recently demonstrated extraction of properties related to thermal stability and solvent removal stability in MOFs, two key characteristics for the design of MOFs as useful catalysts and in gas storage. They exploited(17) the CoRE-MOF(29) set to ensure that topological(32, 40) representations could be applied to train GP and ANN classifiers (i.e., for solvent removal) and regression models (i.e., for breakdown temperatures) on sets of over 1000 reported experimental results. Through these models, Nandy et al. showed(17) it was possible to redesign linker chemistry (e.g., functional groups) to enhance stability in known materials, suggesting a strategy for re-engineering materials with improved properties. Thus, the multiplicative nature of ML acceleration and the opportunities created by the proliferation of experimental data sets suggest that it may soon be possible to engineer materials for practical considerations in much shorter timescales than traditional experimentation allows.

4.4. Outstanding Challenges for Materials Engineering.

In the case of QSPRs or ML models, a question of domain of applicability arises that can be challenging to answer before embarking on a materials optimization campaign. While this can partly be controlled by the choice of ligands or metals in the pool of candidates, an alternative approach is to modify the fitness function to ensure exploitation of a model only in regions well supported by training data. Similar to early multi-objective estimates of model uncertainty in optimization(144), Janet and coworkers employed(120) a composite fitness function that penalized points distant either in feature space(120) or in the latent space(43, 46) of an ANN model (Figure 6). The fitness is thus highest where both the target property is satisfied and the compound is not too distant from training data, controlling errors on newly discovered compounds to little more than the baseline test error of the models for applications in spin-crossover design(46, 120) or catalysis(43) (Figure 6).

For multi-objective optimization with model exploration, i.e., where promising regions of space with high model uncertainty are identified and then incorporated as new data to an updated model, GPs are a natural choice because they provide inherent uncertainty estimates. However, GP property evaluation scales supralinearly with the data set size, making them less attractive from a time-to-evaluation standpoint when data set sizes are large unless specific approaches are employed that exploit sparsity to reduce scaling. Additionally, ANNs have been shown(37) to have better generalization errors, indicating they have more promise for predicting new points past the Pareto front (Figure 7). As a final caveat, prediction based on literature data can suffer from positive publication bias(107), and it can be challenging to automatically extract large data sets of experimental properties if there is no universal reporting standard. In classification problems, some have proposed to intentionally inject hypothetical, negative results to improve classifier performance(145). Despite these caveats, ML-driven multi-objective optimization and engineering remains a promising way to reduce time to discovery of materials with optimal tradeoffs from decades to weeks and to uncover new ways to repurpose existing materials.

5. Conclusions and Outlook.

ML has become a part of the fabric of high-throughput screening and computational discovery of materials. Despite its increasingly central role, challenges remain in fully realizing the promise of ML, especially for the practical acceleration of the engineering of robust materials and elucidation of design strategies that surpass trial-and-error or high-throughput DFT-based or experimentation-based screening alone. This Review covered recent advances in both the algorithms and in their application that are starting to make inroads toward: i) the discovery of new materials through ML-accelerated large-scale enumerative screening, ii) the design of materials through identification of rules and principles that govern materials properties, and iii) the engineering of practical materials especially by satisfying multiple objectives.

Beyond the approaches discussed here, numerous opportunities remain for further realizing the potential of ML-accelerated computational materials science. While enumeration is one successful strategy for discovering new materials that can be accelerated with ML regression models, the potential bias in either the rules for enumeration or in the training data available to the ML model is still poorly understood. The use of either published experimental or computational results can be expected to lead to a bias towards successes, which can then influence the ML models that lack information about unsuccessful compounds. Another opportunity is for the continued development of generative models, which have begun to be applied in materials science.(146-149) However, generative models can lead to compounds that are not experimentally feasible and can inherit the biases in the underlying data distribution. It remains to be seen how to best enumerate new compounds with finite available resources while maintaining realism, but ML models that predict calculation success(49) could provide one useful way to ensure that calculations are fruitful.

Uncertainty in all aspects of ML-accelerated computational materials science remains a challenge. For simple materials (e.g., organic molecules), transfer learning from a low level of theory to a higher level one has been demonstrated(150), but a hierarchy of tractable methods remains out of reach for most correlated materials such as transition-metal complexes. While training on a single DFT functional is common and can be expected to bias predicted lead compounds(52), consensus approaches that are trained on an ensemble of functionals can provide more robust predictions when ground-truth theory or experiment are unavailable.(105) Another source of uncertainty is in the ML model itself and how well it generalizes to new spaces. Transferable descriptors or representations as well as improved uncertainty quantification metrics(46) along with ever-growing data sets will be useful in addressing these challenges.

In the design and engineering of new materials, DFT-based ML models can only go so far. They are suitable for predicting properties such as activity but they may miss longer-term goals of selectivity or stability essential to real-world materials predictions. While careful extraction of experimental literature data(17) or high-throughput experimentation can address some of these challenges, large data sets are seldom available for new classes of materials, either due to a lack of prior experimentation or inconsistent reporting in the scientific literature. Another challenge in building structure–property relationships is that many of the approaches discussed here required knowledge of the structure, which itself can evolve or be unknown during the operative lifecycle (e.g., in catalysis). Even in atomically precise transition-metal complexes, challenges can arise in understanding when the electronic structure may differ (e.g., for distinct, excited electronic states) while the geometric structure remains largely unchanged. Finally, there can be uncertainty in both the experimental measurement and conditions that may challenge the robustness of any automated extraction protocol that aims to leverage data across numerous experimental setups. Despite these challenges, increased availability and size of datasets of materials properties, both from DFT and from experiment in challenging materials spaces such as transition-metal chemistry, will undoubtedly unlock new opportunities for ML-accelerated computational materials science in the years to come.

ACKNOWLEDGMENT

This work has been generously supported over the years by the Office of Naval Research under grant numbers N00014-17-1-2956, N00014-18-1-2434, and N00014-20-1-2150, DARPA grant D18AP00039, the Department of Energy under grant numbers DE-SC0018096 and DE-SC0012702, the National Science Foundation under grant numbers CBET-1704266 and CBET-1846426, an AAAS Marion Milligan Mason Award, a Career Award at the Scientific Interface from the Burroughs Wellcome Fund, and an Alfred P. Sloan Fellowship in Chemistry. A.N. was partially supported by a National Science Foundation Graduate Research Fellowship under Grant #1122374. The authors thank Adam H. Steeves for providing a critical reading of the manuscript.

LITERATURE CITED

1. Curtarolo S, Hart GLW, Nardelli MB, Mingo N, Sanvito S, Levy O. 2013. The high-throughput highway to computational materials design. *Nat. Mater.* 12: 191-201
2. Jain A, Ong SP, Hautier G, Chen W, Richards WD, et al. 2013. Commentary: The Materials Project: A materials genome approach to accelerating materials innovation. *APL Mater.* 1: 011002
3. Agrawal A, Choudhary A. 2016. Perspective: Materials informatics and big data: Realization of the "fourth paradigm" of science in materials science. *APL Mater.* 4: 053208
4. Dimitrov T, Kreisbeck C, Becker JS, Aspuru-Guzik A, Saikin SK. 2019. Autonomous Molecular Design: Then and Now. *ACS Appl. Mater. Interfaces* 11: 24825-36
5. Burello E, Rothenberg G. 2003. Optimal Heck Cross-Coupling Catalysis: A Pseudo-Pharmaceutical Approach. *Adv. Synth. Catal.* 345: 1334-40
6. Burello E, Farrusseng D, Rothenberg G. 2004. Combinatorial Explosion in Homogeneous Catalysis: Screening 60,000 Cross-Coupling Reactions. *Adv. Synth. Catal.* 346: 1844-53

7. Burello E, Rothenberg G. 2005. Topological Mapping of Bidentate Ligands: A Fast Approach for Screening Homogeneous Catalysts. *Adv. Synth. Catal.* 347: 1969-77
8. Landrum GA, Penzotti JE, Putta S. 2005. Machine-learning models for combinatorial catalyst discovery. *Meas. Sci. Technol.* 16: 270-77
9. Ramakrishnan R, Dral PO, Rupp M, Von Lilienfeld OA. 2014. Quantum chemistry structures and properties of 134 kilo molecules. *Sci. Data* 1: 140022
10. Wu ZQ, Ramsundar B, Feinberg EN, Gomes J, Geniesse C, et al. 2018. MoleculeNet: a benchmark for molecular machine learning. *Chem. Sci.* 9: 513-30
11. Haghighatlari M, Vishwakarma G, Altarawy D, Subramanian R, Kota BU, et al. 2020. 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.* 10: e1458
12. Huo H, Rong Z, Kononova O, Sun W, Botari T, et al. 2019. Semi-supervised machine-learning classification of materials synthesis procedures. *npj Comput. Mater.* 5: 1-7
13. Kim E, Huang K, Jegelka S, Olivetti E. 2017. Virtual screening of inorganic materials synthesis parameters with deep learning. *npj Comput. Mater.* 3: 53
14. Kim E, Huang K, Saunders A, McCallum A, Ceder G, Olivetti E. 2017. Materials synthesis insights from scientific literature via text extraction and machine learning. *Chem. Mater.* 29: 9436-44
15. Kononova O, He T, Huo H, Trewartha A, Olivetti EA, Ceder G. 2021. Opportunities and challenges of text mining in aterials research. *iScience* 24: 102155
16. Moosavi SM, Chidambaram A, Talirz L, Haranczyk M, Stylianou KC, Smit B. 2019. Capturing chemical intuition in synthesis of metal-organic frameworks. *Nat. Commun.* 10: 539
17. Nandy A, Duan C, Kulik HJ. submitted. Using Machine Learning and Data Mining to Leverage Community Knowledge for the Engineering of Stable Metal-Organic Frameworks. arXiv:2106.13327
18. Nandy A, Duan C, Taylor MG, Liu F, Steeves AH, Kulik HJ. 2021. Computational Discovery of Transition-Metal Complexes: From High-throughput Screening to Machine Learning. *Chem. Rev.* 121: 9927-10000
19. Swain MC, Cole JM. 2016. ChemDataExtractor: A Toolkit for Automated Extraction of Chemical Information from the Scientific Literature. *J. Chem. Inf. Model.* 56: 1894-904
20. Foscato M, Venkatraman V, Occhipinti G, Alsberg BK, Jensen VR. 2014. Automated Building of Organometallic Complexes from 3D Fragments. *J. Chem. Inf. Model.* 54: 1919-31
21. Ioannidis EI, Gani TZH, Kulik HJ. 2016. molSimplify: A toolkit for automating discovery in inorganic chemistry. *J. Comput. Chem.* 37: 2106-17
22. Nandy A, Duan C, Janet JP, Gugler S, Kulik HJ. 2018. Strategies and Software for Machine Learning Accelerated Discovery in Transition Metal Chemistry. *Ind. Eng. Chem. Res.* 57: 13973-86
23. Ward L, Dunn A, Faghaninia A, Zimmermann NER, Bajaj S, et al. 2018. Matminer: An open source toolkit for materials data mining. *Comput. Mater. Sci.* 152: 60-69
24. Ward L, Agrawal A, Choudhary A, Wolverton C. 2016. A general-purpose machine learning framework for predicting properties of inorganic materials. *npj Comput. Mater.* 2: 16028

25. Groom CR, Bruno IJ, Lightfoot MP, Ward SC. 2016. The Cambridge structural database. *Acta Crystallogr., Sect. B: Struct. Sci., Cryst. Eng. Mater.* 72: 171-79
26. Gugler S, Janet JP, Kulik HJ. 2020. Enumeration of de novo inorganic complexes for chemical discovery and machine learning. *Mol. Syst. Des. Eng.* 5: 139-52
27. Wilmer CE, Leaf M, Lee CY, Farha OK, Hauser BG, et al. 2012. Large-scale screening of hypothetical metal–organic frameworks. *Nat. Chem.* 4: 83
28. Colón YJ, Gómez-Gualdrón DA, Snurr RQ. 2017. Topologically guided, automated construction of metal–organic frameworks and their evaluation for energy-related applications. *Cryst. Growth Des.* 17: 5801-10
29. Chung YG, Haldoupis E, Bucior BJ, Haranczyk M, Lee S, et al. 2019. Advances, Updates, and Analytics for the Computation-Ready, Experimental Metal–Organic Framework Database: CoRE MOF 2019. *J. Chem. Eng. Data* 64: 5985-98
30. Saal JE, Kirklin S, Aykol M, Meredig B, Wolverton C. 2013. Materials Design and Discovery with High-Throughput Density Functional Theory: The Open Quantum Materials Database (OQMD). *JOM* 65: 1501-09
31. Balcells D, Skjelstad BB. 2020. tmQM Dataset—Quantum Geometries and Properties of 86k Transition Metal Complexes. *J. Chem. Inf. Model.* 60: 6135-46
32. Janet JP, Kulik HJ. 2017. Resolving transition metal chemical space: feature selection for machine learning and structure-property relationships. *J. Phys. Chem. A* 121: 8939-54
33. Janet JP, Kulik HJ. 2017. Predicting Electronic Structure Properties of Transition Metal Complexes with Neural Networks. *Chem. Sci.* 8: 5137-52
34. Meredig B, Agrawal A, Kirklin S, Saal JE, Doak JW, et al. 2014. Combinatorial screening for new materials in unconstrained composition space with machine learning. *Phys. Rev. B* 89: 094104
35. Ruddigkeit L, van Deursen R, Blum LC, Reymond J-L. 2012. Enumeration of 166 Billion Organic Small Molecules in the Chemical Universe Database GDB-17. *J. Chem. Inf. Model.* 52: 2864-75
36. Foscato M, Houghton BJ, Occhipinti G, Deeth RJ, Jensen VR. 2015. Ring Closure to Form Metal Chelates in 3D Fragment-Based De Novo Design. *J. Chem. Inf. Model.* 55: 1844-56
37. Janet JP, Ramesh S, Duan C, Kulik HJ. 2020. Accurate Multiobjective Design in a Space of Millions of Transition Metal Complexes with Neural-Network-Driven Efficient Global Optimization. *ACS Cent. Sci.* 6: 513-24
38. Boyd PG, Woo TK. 2016. A Generalized Method for Constructing Hypothetical Nanoporous Materials of any Net Topology from Graph Theory. *CrystEngComm* 18: 3777-92
39. Martin RL, Haranczyk M. 2014. Construction and Characterization of Structure Models of Crystalline Porous Polymers. *Cryst. Growth Des.* 14: 2431-40
40. Moosavi SM, Nandy A, Jablonka KM, Ongari D, Janet JP, et al. 2020. Understanding the diversity of the metal-organic frameworks ecosystems. *Nat. Commun.* 11: 4068
41. Deem MW, Pophale R, Cheeseman PA, Earl DJ. 2009. Computational Discovery of New Zeolite-Like Materials. *J. Phys. Chem. C* 113: 21353-60
42. Hageman JA, Westerhuis JA, Frühauf H-W, Rothenberg G. 2006. Design and Assembly of Virtual Homogeneous Catalyst Libraries –Towards In Silico Catalyst Optimisation. *Adv. Synth. Catal.* 348: 361-69

43. Nandy A, Zhu J, Janet JP, Duan C, Getman RB, Kulik HJ. 2019. Machine Learning Accelerates the Discovery of Design Rules and Exceptions in Stable Metal- Oxo Intermediate Formation. *ACS Catal*. 9: 8243-55

44. Liu F, Duan C, Kulik HJ. 2020. Rapid Detection of Strong Correlation with Machine Learning for Transition-Metal Complex High-Throughput Screening. *J. Phys. Chem. Lett*. 11: 8067-76

45. Ramos-Cordoba E, Matito E. 2017. Local Descriptors of Dynamic and Nondynamic Correlation. *J. Chem. Theory Comput*. 13: 2705-11

46. Janet JP, Duan C, Yang T, Nandy A, Kulik HJ. 2019. A quantitative uncertainty metric controls error in neural network-driven chemical discovery. *Chem. Sci*. 10: 7913-22

47. Gómez-Gualdrón DA, Colón YJ, Zhang X, Wang TC, Chen Y-S, et al. 2016. Evaluating Topologically Diverse Metal–Organic Frameworks for Cryo-Adsorbed Hydrogen Storage. *Energy Environ. Sci*. 9: 3279-89

48. Martin RL, Simon CM, Smit B, Haranczyk M. 2014. In silico Design of Porous Polymer Networks: High-Throughput Screening for Methane Storage Materials. *J. Am. Chem. Soc*. 136: 5006-22

49. Duan C, Janet JP, Liu F, Nandy A, Kulik HJ. 2019. Learning from Failure: Predicting Electronic Structure Calculation Outcomes with Machine Learning Models. *J. Chem. Theory Comput*. 15: 2331-45

50. Gomez-Bombarelli R, Aguilera-Iparraguirre J, Hirzel TD, Duvenaud D, Maclaurin D, et al. 2016. Design of efficient molecular organic light-emitting diodes by a high-throughput virtual screening and experimental approach. *Nat. Mater*. 15: 1120-27

51. Boyd PG, Chidambaram A, García-Díez E, Ireland CP, Daff TD, et al. 2019. Data-Driven Design of Metal–Organic Frameworks for Wet Flue Gas CO_2 Capture. *Nature* 576: 253-56

52. Janet JP, Liu F, Nandy A, Duan C, Yang T, et al. 2019. Designing in the Face of Uncertainty: Exploiting Electronic Structure and Machine Learning Models for Discovery in Inorganic Chemistry. *Inorg. Chem*. 58: 10592-606

53. Durand DJ, Fey N. 2019. Computational Ligand Descriptors for Catalyst Design. *Chem. Rev*. 119: 6561-94

54. Mansson RA, Welsh AH, Fey N, Orpen AG. 2006. Statistical Modeling of a Ligand Knowledge Base. *J. Chem. Inf. Model*. 46: 2591-600

55. Beyreuther S, Hunger J, Huttner G, Mann S, Zsolnai L. 1996. Conformation of tripod Metal Templates in CH3C(CH2PPh2)3MLn (n = 2, 3): Neural Networks in Conformational Analysis. *Chem. Ber*. 129: 745-57

56. Ceriotti M. 2019. Unsupervised machine learning in atomistic simulations, between predictions and understanding. *J. Chem. Phys*. 150: 150901

57. Virshup AM, Contreras-García J, Wipf P, Yang W, Beratan DN. 2013. Stochastic voyages into uncharted chemical space produce a representative library of all possible drug-like compounds. *J. Am. Chem. Soc*. 135: 7296-303

58. Cruz VL, Martinez S, Ramos J, Martinez-Salazar J. 2014. 3D-QSAR as a Tool for Understanding and Improving Single-Site Polymerization Catalysts. A Review. *Organometallics* 33: 2944-59

59. Sigman MS, Harper KC, Bess EN, Milo A. 2016. The Development of Multidimensional Analysis Tools for Asymmetric Catalysis and Beyond. *Acc. Chem. Res*. 49: 1292-301

60. Zahrt AF, Athavale SV, Denmark SE. 2020. Quantitative Structure–Selectivity Relationships in Enantioselective Catalysis: Past, Present, and Future. *Chem. Rev.* 120: 1620-89

61. Ahneman DT, Estrada JG, Lin S, Dreher SD, Doyle AG. 2018. Predicting Reaction Performance in C–N Cross-Coupling Using Machine Learning. *Science* 360: 186-90

62. Lundberg SM, Lee S-I. 2017. *A unified approach to interpreting model predictions*. Presented at Proceedings of the 31st international conference on neural information processing systems

63. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. 2017. *Grad-cam: Visual explanations from deep networks via gradient-based localization*. Presented at Proceedings of the IEEE international conference on computer vision

64. Gomez-Bombarelli R, Wei JN, Duvenaud D, Hernandez-Lobato JM, Sanchez-Lengeling B, et al. 2018. Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules. *ACS Cent. Sci.* 4: 268-76

65. Iovanac NC, Savoie BM. 2019. Improved Chemical Prediction from Scarce Data Sets via Latent Space Enrichment. *J. Phys. Chem. A* 123: 4295-302

66. Cundari TR, Sârbu C, Pop HF. 2002. Robust Fuzzy Principal Component Analysis (FPCA). A Comparative Study Concerning Interaction of Carbon−Hydrogen Bonds with Molybdenum−Oxo Bonds. *J. Chem. Inf. Comput. Sci.* 42: 1363-69

67. Saadun AJ, Pablo-García S, Paunović V, Li Q, Sabadell-Rendón A, et al. 2020. Performance of Metal-Catalyzed Hydrodebromination of Dibromomethane Analyzed by Descriptors Derived from Statistical Learning. *ACS Catal.* 10: 6129-43

68. van der Maaten L, Hinton G. 2008. Visualizing data using t-SNE. *J. Mach. Learn. Res.* 9: 2579-605

69. McInnes L, Healy J, Melville J. 2018. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv:1802.03426*

70. Maley Steven M, Kwon D-H, Rollins N, Stanley JC, Sydora OL, et al. 2020. Quantum-Mechanical Transition-State Model Combined with Machine Learning Provides Catalyst Design Features for Selective Cr Olefin Oligomerization. *Chem. Sci.* 11: 9665-74

71. Martínez S, Cruz VL, Ramos J, Martínez-Salazar J. 2012. Polymerization Activity Prediction of Zirconocene Single-Site Catalysts Using 3D Quantitative Structure–Activity Relationship Modeling. *Organometallics* 31: 1673-79

72. Friederich P, dos Passos Gomes G, De Bin R, Aspuru-Guzik A, Balcells D. 2020. Machine Learning Dihydrogen Activation in the Chemical Space Surrounding Vaska's Complex. *Chem. Sci.* 11: 4584-601

73. Cordova M, Wodrich MD, Meyer B, Sawatlon B, Corminboeuf C. 2020. Data-Driven Advancement of Homogeneous Nickel Catalyst Activity for Aryl Ether Cleavage. *ACS Catal.* 10: 7021-31

74. Meyer B, Sawatlon B, Heinen S, von Lilienfeld OA, Corminboeuf C. 2018. Machine learning meets volcano plots: computational discovery of cross-coupling catalysts. *Chem. Sci.* 9: 7069-77

75. Freeze JG, Kelly HR, Batista VS. 2019. Search for Catalysts by Inverse Design: Artificial Intelligence, Mountain Climbers, and Alchemists. *Chem. Rev.* 119: 6595-612

76. Orpen A. 2002. Applications of the Cambridge Structural Database to molecular inorganic chemistry. *Acta Crystallogr., Sect. B: Struct. Sci.* 58: 398-406

77. Tolman CA. 1970. Phosphorus Ligand Exchange Equilibriums on Zerovalent Nickel. Dominant Role for Steric Effects. *J. Am. Chem. Soc.* 92: 2956-65
78. Tolman CA. 1977. Steric Effects of Phosphorus Ligands in Organometallic Chemistry and Homogeneous Catalysis. *Chem. Rev.* 77: 313-48
79. Falivene L, Cao Z, Petta A, Serra L, Poater A, et al. 2019. Towards the online computer-aided design of catalytic pockets. *Nat. Chem.* 11: 872-79
80. Bucior BJ, Bobbitt NS, Islamoglu T, Goswami S, Gopalan A, et al. 2019. Energy-Based Descriptors to Rapidly Predict Hydrogen Storage in Metal–Organic Frameworks. *Mol. Syst. Des. Eng.* 4: 162-74
81. Zahrt AF, Henle JJ, Rose BT, Wang Y, Darrow WT, Denmark SE. 2019. Prediction of Higher-Selectivity Catalysts by Computer-Driven Workflow and Machine Learning. *Science* 363: eaau5631
82. Lipkowitz KB, D'Hue CA, Sakamoto T, Stack JN. 2002. Stereocartography: A Computational Mapping Technique That Can Locate Regions of Maximum Stereoinduction around Chiral Catalysts. *J. Am. Chem. Soc.* 124: 14255-67
83. Phan H, Hrudka JJ, Igimbayeva D, Lawson Daku LvM, Shatruk M. 2017. A simple approach for predicting the spin state of homoleptic Fe (II) tris-diimine complexes. *J. Am. Chem. Soc.* 139: 6437-47
84. Holleis L, Shivaram BS, Balachandran PV. 2019. Machine learning guided design of single-molecule magnets for magnetocaloric applications. *Appl. Phys. Lett.* 114: 222404
85. Taylor MG, Yang T, Lin S, Nandy A, Janet JP, et al. 2020. Seeing is Believing: Experimental Spin States from Machine Learning Model Structure Predictions. *J. Phys. Chem. A* 124: 3286-99
86. Chang AM, Freeze JG, Batista VS. 2019. Hammett Neural Networks: Prediction of Frontier Orbital Energies of Tungsten–Benzylidyne Photoredox Complexes. *Chem. Sci.* 10: 6844-54
87. Yada A, Nagata K, Ando Y, Matsumura T, Ichinoseki S, Sato K. 2018. Machine Learning Approach for Prediction of Reaction Yield with Simulated Catalyst Parameters. *Chem. Lett.* 47: 284-87
88. Mikami K. 2020. Interactive-quantum-chemical-descriptors enabling accurate prediction of an activation energy through machine learning. *Polymer* 203: 122738
89. Yang WH, Fidelis TT, Sun WH. 2020. Prediction of catalytic activities of bis(imino)pyridine metal complexes by machine learning. *J. Comput. Chem.* 41: 1064-67
90. Janet JP, Gani TZH, Steeves AH, Ioannidis EI, Kulik HJ. 2017. Leveraging Cheminformatics Strategies for Inorganic Discovery: Application to Redox Potential Design. *Ind. Eng. Chem. Res.* 56: 4898-910
91. Henle JJ, Zahrt AF, Rose BT, Darrow WT, Wang Y, Denmark SE. 2020. Development of a Computer-Guided Workflow for Catalyst Optimization. Descriptor Validation, Subset Selection, and Training Set Analysis. *J. Am. Chem. Soc.* 142: 11578-92
92. Fanourgakis GS, Gkagkas K, Tylianakis E, Froudakis GE. 2020. A Universal Machine Learning Algorithm for Large-Scale Screening of Materials. *J. Am. Chem. Soc.* 142: 3814-22
93. Xie T, Grossman JC. 2018. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Phys. Rev. Lett.* 120: 145301
94. Moreau G, Broto P. 1980. The autocorrelation of a topological structure: A new molecular descriptor. *Nouv. J. Chim.* 4: 359

95. Fernandez M, Trefiak NR, Woo TK. 2013. Atomic Property Weighted Radial Distribution Functions Descriptors of Metal–Organic Frameworks for the Prediction of Gas Uptake Capacity. *J. Phys. Chem. C* 117: 14095-105

96. Eckhoff M, Lausch KN, Blöchl PE, Behler J. 2020. Predicting oxidation and spin states by high-dimensional neural networks: Applications to lithium manganese oxide spinels. *J. Chem. Phys.* 153: 164107

97. Lunghi A, Sanvito S. 2020. Surfing Multiple Conformation-Property Landscapes via Machine Learning: Designing Single-Ion Magnetic Anisotropy. *J. Phys. Chem. C* 124: 5802-06

98. Pardakhti M, Moharreri E, Wanik D, Suib SL, Srivastava R. 2017. Machine Learning Using Combined Structural and Chemical Descriptors for Prediction of Methane Adsorption Performance of Metal Organic Frameworks (MOFs). *ACS Comb. Sci.* 19: 640-45

99. Duan C, Liu F, Nandy A, Kulik HJ. 2020. Data-Driven Approaches Can Overcome the Cost-Accuracy Tradeoff in Multireference Diagnostics. *J. Chem. Theory Comput.* 16: 4373-87

100. Rupp M, Tkatchenko A, Muller KR, von Lilienfeld OA. 2012. Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning. *Phys. Rev. Lett.* 108: 058301

101. Borboudakis G, Stergiannakos T, Frysali M, Klontzas E, Tsamardinos I, Froudakis GE. 2017. Chemically Intuited, Large-Scale Screening of MOFs by Machine Learning Techniques. *npj Comput. Mater.* 3: 40

102. Anderson R, Rodgers J, Argueta E, Biong A, Gómez-Gualdrón DA. 2018. Role of Pore Chemistry and Topology in the CO_2 Capture Capabilities of MOFs: From Molecular Simulation to Machine Learning. *Chem. Mater.* 30: 6325-37

103. Martin RL, Smit B, Haranczyk M. 2011. Addressing Challenges of Identifying Geometrically Diverse Sets of Crystalline Porous Materials. *J. Chem. Inf. Model.* 52: 308-18

104. Willems TF, Rycroft CH, Kazi M, Meza JC, Haranczyk M. 2012. Algorithms and Tools for High-Throughput Geometry-Based Analysis of Crystalline Porous Materials. *Microporous Mesoporous Mater.* 149: 134-41

105. Duan C, Chen S, Taylor MG, Liu F, Kulik HJ. Machine learning to tame divergent density functional approximations: a new path to consensus materials design principles. arXiv:2106.13109

106. Liao P, Getman RB, Snurr RQ. 2017. Optimizing Open Iron Sites in Metal–Organic Frameworks for Ethane Oxidation: A First-Principles Study. *ACS Appl. Mater. Interfaces* 9: 33484-92

107. Jia X, Lynch A, Huang Y, Danielson M, Lang'at I, et al. 2019. Anthropogenic biases in chemical reaction data hinder exploratory inorganic synthesis. *Nature* 573: 251-55

108. Raccuglia P, Elbert KC, Adler PD, Falk C, Wenny MB, et al. 2016. Machine-learning-assisted materials discovery using failed experiments. *Nature* 533: 73-76

109. Umegaki T, Watanabe Y, Nukui N, Omata K, Yamada M. 2003. Optimization of Catalyst for Methanol Synthesis by a Combinatorial Approach Using a Parallel Activity Test and Genetic Algorithm Assisted by a Neural Network. *Energy Fuels* 17: 850-56

110. Andersson M, Bligaard T, Kustov A, Larsen K, Greeley J, et al. 2006. Toward Computational Screening in Heterogeneous Catalysis: Pareto Optimal Methanation Catalysts. *J. Catal.* 239: 501-06

111. Corma A, Serra J, Serna P, Valero S, Argente E, Botti V. 2005. Optimisation of Olefin Epoxidation Catalysts with the Application of High-Throughput and Genetic Algorithms Assisted by Artificial Neural Networks (Softcomputing Techniques). *J. Catal.* 229: 513-24

112. Cuéllar MP, Lapresta-Fernández A, Herrera JM, Salinas-Castillo A, Pegalajar MdC, et al. 2015. Thermochromic sensor design based on Fe(II) spin crossover/polymers hybrid materials and artificial neural networks as a tool in modelling. *Sens. Actuators, B* 208: 180-87

113. Gustafson JA, Wilmer CE. 2019. Intelligent Selection of Metal–Organic Framework Arrays for Methane Sensing via Genetic Algorithms. *ACS Sens.* 4: 1586-93

114. Tshitoyan V, Dagdelen J, Weston L, Dunn A, Rong Z, et al. 2019. Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature* 571: 95-98

115. Schweidtmann AM, Clayton AD, Holmes N, Bradford E, Bourne RA, Lapkin AA. 2018. Machine Learning Meets Continuous Flow Chemistry: Automated Optimization Towards the Pareto Front of Multiple Objectives. *Chem. Eng. J.* 352: 277-82

116. Batra R, Chen C, Evans TG, Walton KS, Ramprasad R. 2020. Prediction of water stability of metal–organic frameworks using machine learning. *Nat. Mach. Intell.* 2: 704-10

117. Kreutz JE, Shukhaev A, Du W, Druskin S, Daugulis O, Ismagilov RF. 2010. Evolution of Catalysts Directed by Genetic Algorithms in a Plug-Based Microfluidic Device Tested with Oxidation of Methane by Oxygen. *J. Am. Chem. Soc.* 132: 3128-32

118. Watanabe Y, Umegaki T, Hashimoto M, Omata K, Yamada M. 2004. Optimization of Cu Oxide Catalysts for Methanol Synthesis by Combinatorial Tools Using 96 Well Microplates, Artificial Neural Network and Genetic Algorithm. *Catal. Today* 89: 455-64

119. Chung YG, Gómez-Gualdrón DA, Li P, Leperi KT, Deria P, et al. 2016. In Silico Discovery of Metal-Organic Frameworks for Precombustion CO_2 Capture using a Genetic Algorithm. *Sci. Adv.* 2: e1600909

120. Janet JP, Chan L, Kulik HJ. 2018. Accelerating Chemical Discovery with Machine Learning: Simulated Evolution of Spin Crossover Complexes with an Artificial Neural Network. *J. Phys. Chem. Lett.* 9: 1064-71

121. Jennings PC, Lysgaard S, Hummelshøj JS, Vegge T, Bligaard T. 2019. Genetic Algorithms for Computational Materials Discovery Accelerated by Machine Learning. *npj Comput. Mater.* 5: 46

122. Leardi R. 2001. Genetic Algorithms in Chemometrics and Chemistry: a Review. *J. Chemom.* 15: 559-69

123. Le TC, Winkler DA. 2016. Discovery and Optimization of Materials Using Evolutionary Approaches. *Chem. Rev.* 116: 6107-32

124. Collins SP, Daff TD, Piotrkowski SS, Woo TK. 2016. Materials Design by Evolutionary Optimization of Functional Groups in Metal-Organic Frameworks. *Sci. Adv.* 2: e1600954

125. Lee S, Kim B, Cho H, Lee H, Lee SY, et al. 2021. Computational Screening of Trillions of Metal–Organic Frameworks for High-Performance Methane Storage. *ACS Appl. Mater. Interfaces* 13: 23647-54

126. Foscato M, Venkatraman V, Jensen VR. 2019. DENOPTIM: Software for Computational de Novo Design of Organic and Inorganic Molecules. *J. Chem. Inf. Model.* 59: 4077-82

127. Llamas-Galilea J, Gobin OC, Schüth F. 2009. Comparison of Single- And Multiobjective Design of Experiment in Combinatorial Chemistry for the Selective Dehydrogenation of Propane. *J. Comb. Chem.* 11: 907-13

128. Martin RL, Haranczyk M. 2013. Insights into Multi-Objective Design of Metal–Organic Frameworks. *Cryst. Growth Des.* 13: 4208-12

129. Gopakumar AM, Balachandran PV, Xue D, Gubernatis JE, Lookman T. 2018. Multi-Objective Optimization for Materials Discovery via Adaptive Design. *Sci. Rep.* 8: 3738

130. Yuan R, Liu Z, Balachandran PV, Xue D, Zhou Y, et al. 2018. Accelerated Discovery of Large Electrostrains in BaTiO3-Based Piezoelectrics Using Active Learning. *Adv. Mater.* 30: 1702884

131. del Rosario Z, Rupp M, Kim Y, Antono E, Ling J. 2020. Assessing the Frontier: Active Learning, Model Accuracy, and Multi-Objective Candidate Discovery and Optimization. *J. Chem. Phys.* 153: 024112

132. Zhang Y, Lee AA. 2019. Bayesian semi-supervised learning for uncertainty-calibrated prediction of molecular properties and active learning. *Chem. Sci.* 10: 8154-63

133. Forrester AIJ, Keane AJ. 2009. Recent advances in surrogate-based optimization. *Prog. Aeronaut. Sci.* 45: 50-79

134. Rodemerck U, Baerns M, Holena M, Wolf D. 2004. Application of a Genetic Algorithm and a Neural Network for the Discovery and Optimization of New Solid Catalytic Materials. *Appl. Surf. Sci.* 223: 168-74

135. Huang K, Chen F-Q, Lü D-W. 2001. Artificial Neural Network-Aided Design of a Multi-Component Catalyst for Methane Oxidative Coupling. *Appl. Catal., A* 219: 61-68

136. Huang K, Zhan X-L, Chen F-Q, Lü D-W. 2003. Catalyst Design for Methane Oxidative Coupling by Using Artificial Neural Network and Hybrid Genetic Algorithm. *Chem. Eng. Sci.* 58: 81-87

137. Chu Y, Heyndrickx W, Occhipinti G, Jensen VR, Alsberg BK. 2012. An Evolutionary Algorithm for de Novo Optimization of Functional Transition Metal Compounds. *J. Am. Chem. Soc.* 134: 8885-95

138. Rizkin BA, Hartman RL. 2019. Supervised machine learning for prediction of zirconocene-catalyzed alpha-olefin polymerization. *Chem. Eng. Sci.* 210: 115224

139. Amar Y, Schweidtmann A, Deutsch P, Cao LW, Lapkin A. 2019. Machine learning and molecular descriptors enable rational solvent selection in asymmetric catalysis. *Chem. Sci.* 10: 6697-706

140. Siebert M, Krennrich G, Seibicke M, Siegle AF, Trapp O. 2019. Identifying high-performance catalytic conditions for carbon dioxide reduction to dimethoxymethane by multivariate modelling. *Chem. Sci.* 10: 10466-74

141. Chen H-Z, Zhang Y-Y, Gong X, Xiang H. 2014. Predicting New TiO_2 Phases with Low Band Gaps by a Multiobjective Global Optimization Approach. *J. Phys. Chem. C* 118: 2333-37

142. Herbol HC, Hu W, Frazier P, Clancy P, Poloczek M. 2018. Efficient search of compositional space for hybrid organic–inorganic perovskites via Bayesian optimization. *npj Comput. Mater.* 4: 51

143. Park S, Kim B, Choi S, Boyd PG, Smit B, Kim J. 2018. Text mining metal–organic framework papers. *J. Chem. Inf. Model.* 58: 244-51

144. Scott DJ, Manos S, Coveney PV. 2008. Design of Electroceramic Materials Using Artificial Neural Networks and Multiobjective Evolutionary Algorithms. *J. Chem. Inf. Model.* 48: 262-73
145. Cáceres EL, Mew NC, Keiser MJ. 2020. Adding stochastic negative examples into machine learning improves molecular bioactivity prediction. *J. Chem. Inf. Model.* 60: 5957-70
146. Kim B, Lee S, Kim J. 2020. Inverse design of porous materials using artificial neural networks. *Sci. Adv.* 6: eaax9324
147. Yao Z, Sánchez-Lengeling B, Bobbitt NS, Bucior BJ, Kumar SGH, et al. 2021. Inverse design of nanoporous crystalline reticular materials with deep generative models. *Nat. Mach. Intell.* 3: 76-86
148. Dan Y, Zhao Y, Li X, Li S, Hu M, Hu J. 2020. Generative Adversarial Networks (GAN) Based Efficient Sampling of Chemical Composition Space for Inverse Design of Inorganic Materials. *npj Comput. Mater.* 6: 84
149. Jensen Z, Kwon S, Schwalbe-Koda D, Paris C, Gómez-Bombarelli R, et al. 2021. Discovering Relationships between OSDAs and Zeolites through Data Mining and Generative Neural Networks. *ACS Cent. Sci.* 7: 858-67
150. Ramakrishnan R, Dral PO, Rupp M, von Lilienfeld OA. 2015. Big data meets quantum chemistry approximations: the Δ-machine learning approach. *J. Chem. Theory Comput.* 11: 2087-96

Terms and Definitions list:

Machine learning (ML): This is the study of computer algorithms that improve automatically based on training/sample data without explicitly being programmed to do so. These methods are closely related to statistical learning.

Artificial neural network (ANN): A powerful class of nonlinear models with easy-to-compute derivatives. They were originally conceptualized as a model of how neurons communicate in the brain because they typically contain activation functions that "turn on" in response to a signal. In this Account, we are generally referring to a class of feed-forward, fully-connected ANNs with a few layers that are sometimes also called deep neural networks (DNNs).

Kernel ridge regression (KRR): A family of models that use a potentially nonlinear kernel to assign similarity between different inputs. KRR models are powerful regression models that are relatively easy to train and understand.

Domain of applicability: The physico-chemical or structural space, knowledge, or information on which the training set of a model has been developed and thus for which the model is suitable to make predictions in new compounds. Typically, some measure of similarity in either the full feature space or a principal component analysis (PCA) is employed to determine the likelihood a model's domain of applicability extends to a new compound.

Genetic (evolutionary) algorithms: This is a class of optimization algorithms inspired by biological evolution, typically involving "generations", reproduction/mutation/recombination, and selection. Solutions in optimization are individuals in a population, and their fitness is evaluated to guide the evolution of the population.

featurization: The process of assigning features to an input. Features are usually numerical constructs based on the chemical structure or constituent atomic properties of a molecule or material.

regularized multiple linear regression (MLR): This is a class of multiple linear regression techniques wherein the fitting objective (typically called the loss function) is modified to penalize model complexity (i.e., regularized) to avoid overfitting. The most well known and widely applied of these techniques is the least absolute shrinkage and selection operator (i.e., LASSO). LASSO behaves like a standard MLR approach but penalizes coefficients associated with uninformative (i.e., uncorrelated) features.

revised autocorrelations (RACs): Extensions of Moreau-Broto autocorrelation (AC) functions, which are products of heuristic properties on the molecular graph distinguished by the number of bond paths the atoms are apart and thus encode no explicit 3D information. RACs were introduced in Ref. (32) to modify the scope of the AC function to focus on ligand-centered and metal-centered products as well as to introduce difference-based evaluations. They also are averaged over equatorial and axial ligands distinctly. The most common application of RACs include several heuristics of oxidation state, HF exchange, spin state, and denticity. For the heuristic properties, electronegativity, covalent radius, identity (i.e., 1), topology (i.e., number of atoms to which an atom is connected), and nuclear charge are used. The conventional application uses a distance cutoff of 3 for the maximum number of bond paths. This typically leads to an approximately 150-dimensional feature vector.

feature selection/feature engineering: Feature selection broadly refers to either by-hand or automated approaches to eliminate uninformative features and identify a sparse set of representative features needed in model training. This is a necessary step for simpler models, whereas deep neural networks essentially act through successive layers to carry out feature selection. Feature engineering is a closely related synonym but more commonly refers to by hand identification of ad hoc features derived from experience with the problem (e.g., inorganic chemistry) at hand.

distances: This generally refers to the difference (i.e., Euclidean distance) of two compounds in their feature space or in a model's space (e.g., the latent space).

similarity: A measure of how comparable two molecules are either in their chemical composition, as judged through feature space distances, or properties. Typically, one aims in simple models for the measure of similarity in the feature space to be comparable to the similarity in property space.

hyperparameter: A parameter whose value controls the model complexity/selection and training process as opposed to parameters that are obtained via training. Examples include the learning rate, mini-batch size, or even number of hidden layers and nodes in an artificial neural network, the kernel width and regularization in a KRR model, the degree of regularization in an MLR model, and so on.

kernel width: This refers to a key hyperparameter in KRR models. A KRR model prediction is influenced by the nearest neighbors to an applied test point, where the influence of points decays with the inverse of the kernel width (typically a Gaussian or exponential decay function, for example). If the kernel width is narrow, then only the most proximal neighbors influence prediction. Conversely, if the kernel width is high, then a larger number of points influence model prediction.

hidden layers: An interior layer of nodes in an ANN, i.e. a layer whose output is not directly observed. Normally, all layers except the input and output layers are hidden layers.

active learning: A special case of machine learning in which new data points are labeled and used in iterative model training, typically at multiple steps, such that the model is iteratively improved, as judged through reduction in MAE rather than trained in one shot. The EGO with 2D-EI approach referenced later in the glossary represents one example of this approach.

uninformative features: Portions of a feature set input vector that do not correlate to output properties and therefore do not improve model performance when they are included in the feature set.