Perspective

Machine Learning for Heterogeneous Catalyst Design and Discovery

Bryan R. Goldsmith, Jacques Esterhuizen, and Jin-Xun Liu
Dept. of Chemical Engineering, University of Michigan, Ann Arbor, MI 48109-2136

Christopher J. Bartel
Dept. of Chemical and Biological Engineering, University of Colorado Boulder, Boulder, CO 80309

Christopher Sutton
Fritz-Haber-Institut der Max-Planck-Gesellschaft, Theory Dept., Faradayweg 4-6, Berlin D-14195, Germany

DOI 10.1002/aic.16198
Published online May 25, 2018 in Wiley Online Library (wileyonlinelibrary.com)

Keywords: heterogeneous catalysis, machine learning, data mining, compressed sensing, computational catalysis

Introduction

Advances in machine learning (ML) are making a large impact in many fields, including: artificial intelligence,[1] materials science,[2,3] and chemical engineering.[4] Generally, ML tools learn from data to find insights or make fast predictions of target properties.[5] Recently, ML is also greatly influencing heterogeneous catalysis research[6] due to the availability of ML (e.g., Python Scikit-learn[7], TensorFlow[8]) and workflow management tools (e.g., ASE,[9] Atomate[10]), the growing amount of data in materials databases (e.g., Novel Materials Discovery Laboratory,[11] Citrination,[12] Materials Project,[13] CatApp[14]), and algorithmic improvements.

New catalysts are needed for sustainable chemical production, alternative energy, and pollution mitigation applications to meet the demands of our world's rising population. It is a challenging endeavor, however, to make novel heterogeneous catalysts with good performance (i.e., stable, active, selective) because their performance depends on many properties: composition, support, surface termination, particle size, particle morphology, and atomic coordination environment.[15] Additionally, the properties of heterogeneous catalysts can change under reaction conditions through various phenomena such as Ostwald ripening, particle disintegration, surface oxidation, and surface reconstruction.[16] Many heterogeneous catalyst structures are disordered or amorphous in their active state, which further complicates their atomic-level characterization by modeling and experiment.[17]

Computational modeling using quantum mechanical (QM) methods such as density functional theory (DFT)[18,19] can accelerate catalyst screening by enabling rapid prototyping and revealing active sites and structure-activity relations. The high computational cost of QM methods, however, limits the range of catalyst spaces that can be examined. Recent progress in merging ML with QM modeling and experiments promises to drive forward rational catalyst design.[20] Therefore, it is timely to highlight the ability of ML tools to accelerate heterogeneous catalyst research. A key question we aim to address in this perspective is how machine learning can aid heterogeneous catalyst design and discovery.

ML has been used in catalysis research since at least the 1990s. Early studies used neural networks to correlate catalyst physicochemical properties and reaction conditions with measured catalytic performance,[21,22] but these studies were limited in the number of systems considered. Recently, ML has been applied to the high-throughput screening of heterogeneous catalysts and found to be predictive and applicable across a broad space of catalysts. ML algorithms such as decision trees, kernel ridge regression, neural networks, support vector machines, principal component analysis, and compressed sensing can help create predictive models of catalyst target properties, which are typically figures of merit corresponding to stability, activity, selectivity.[23–25]

In this perspective, we discuss various areas where ML is making an impact on heterogeneous catalysis research. ML is also aiding homogeneous catalysis research and shares many similarities (and differences) with ML for heterogeneous catalysis, but this discussion is beyond the perspective's scope (for interested readers, see Ref. 26–28). Here, we emphasize the ability of ML combined with QM calculations to speed-up the search for optimal catalysts in combinatorial large spaces, such as alloys. ML-derived interatomic potentials for accurate and fast catalyst simulations will also be assessed, as well as the opportunity for ML to help find descriptors of catalyst performance in large datasets. The use of ML to aid transition state search algorithms (to compute reaction mechanisms) will

Correspondence concerning this article should be addressed to B. R. Goldsmith at bgoldsm@umich.edu.

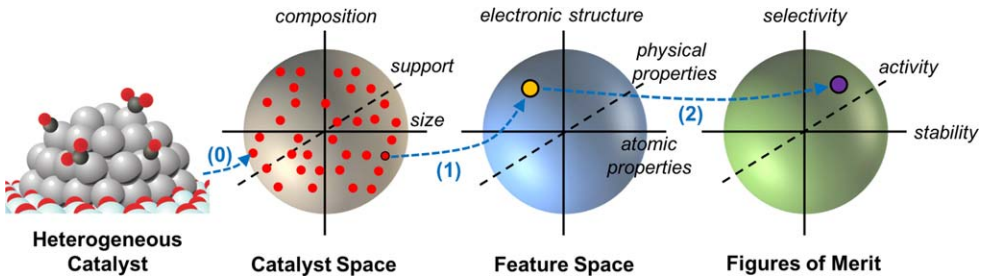

Figure 1. (0) A heterogeneous catalyst sample within some larger dataset (catalyst space)—containing catalysts with different composition, support type, and particle size—can be described by its (1) features within some feature space, which is made up of electronic-structure properties, physical properties, and atomic properties. Machine learning algorithms can (2) build models or find descriptors that map the features describing the catalysts to their figures of merit. Figure adapted from Ref. 24 with permission from Elsevier.

also be discussed. Last, an outlook on future opportunities for ML to assist catalyst discovery will be given.

Impact of Machine Learning on Heterogeneous Catalysis

We first note a few general details about machine learning. For supervised learning of a dataset, a matrix of input features (i.e., properties from which the machine can learn) is constructed and a learning algorithm identifies an analytical or numerical relationship between this matrix and the target property of interest. Typically, in physical sciences, it is desirable that this model has an interpretable form. Caution must be taken to avoid generating flawed models because of poor input feature construction or overfitting the model to the training data. In contrast to supervised learning, unsupervised learning algorithms (such as *k*-means clustering or principal component analysis) find patterns and regularities in data without a target property.

A general workflow for building ML models of catalysts is shown in Figure 1. First a dataset containing various catalysts must be created. Next, each catalyst is described by its features (often called fingerprints or representations), which can consist of electronic-structure properties, physical properties, and atomic properties. Importantly, the features should capture the important physicochemical properties of the materials, should be much easier to compute than the target property, and uniquely define each material. Then machine learning tools can be used to find patterns, build models, or discover descriptors that map the features describing the catalyst to their figures of merit.

We will discuss both supervised and unsupervised learning algorithms applied to heterogeneous catalysis problems in this perspective. Several approaches are described that include a structural representation (e.g., SOAP[29,30]) to produce an accurate model of catalyst properties, whereas other data analytics methods such as SISSO aim to search over a vast space of possible features to find the most accurate and meaningful descriptor.[31] Subgroup discovery extends this feature selection process to identify the ideal features or descriptors for subpopulations of catalyst data. Such ML tools (among many others discussed in the following sections) are poised to become routine methods in the physical sciences for building predictive models and understanding data.

Active site determination and catalyst screening

The conventional route to discover and develop catalysts with desired properties has been through experimental testing and involves candidate materials being synthesized and tested a few samples at a time, which is costly and time consuming. High-throughput screening of combinatorial catalyst libraries can aid catalyst discovery by helping to search through vast design spaces.[32] Machine learning can assist screening efforts by helping to navigate the catalyst search space by finding correlations or by speeding up calculations of the target property.

Researchers have applied ML on experimental data to train models that predict catalytic performance of materials based on their synthesis conditions and composition as model input features.[33,34] Such ML approaches can guide the synthesis of better catalysts, but experimental catalysis data is often limited and hard to obtain, which can lead to models that are not generalizable across diverse chemical spaces. QM modeling can more easily generate larger datasets than experiments or fill in gaps in experimental data, from which ML models can then be trained.

One widely studied class of catalysts that present a combinatorial challenge is alloy nanoparticles, which are used in applications such as fuel cells,[35] biomass conversion,[36] and natural gas conversion[37] due to their compositional tunability and potential multifunctionality.[38] It is challenging to identify optimal catalyst compositions and active sites on alloy catalysts because of the many possible unique structures (e.g., surface facets and adsorbate configurations) due to their compositional diversity and reduction in symmetry (relative to monometallic nanoparticles). Despite the many possible surface facets on alloy catalysts and their potential contributions to catalyst performance, researchers typically model only a few stable facets, usually the (111), (100), or (110) because of the computational expense of modeling every surface. Yet, the active sites contributing the most to the observed rate are often not sites on the most stable surface,[17,39] so modeling only a few stable facets could misrepresent the catalytically active surface.

Recent works show ML can be integrated with QM methods to overcome the computational bottleneck of pure QM

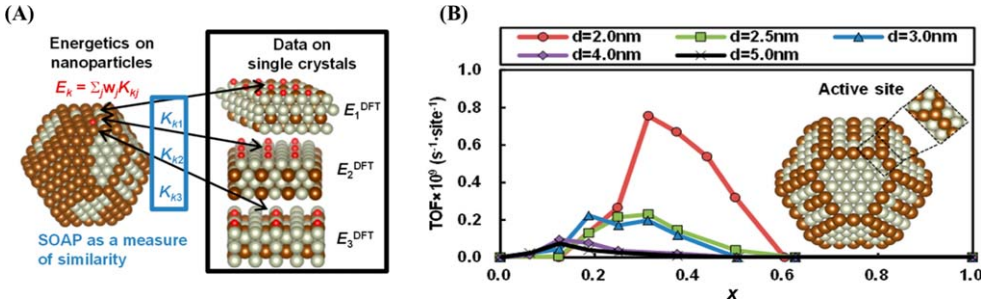

Figure 2. (A) Bayesian linear regression scheme, using SOAP as the kernel, to predict energetics of reaction intermediates on truncated octahedral $Rh_{(1-x)}Au_x$ nanoparticle catalysts. The nanoparticle and reaction intermediate energetics are estimated based on training data of adsorbate binding energies on single crystal surfaces obtained using density functional theory (DFT) calculations. E_k is the energy of the kth reaction intermediate on the nanoparticle, K_{kj} is the SOAP kernel, and w_j are the regression coefficients. (B) Predicted turnover frequencies (TOF) per surface site at 500 K for the direct decomposition of NO on $Rh_{(1-x)}Au_x$ nanoparticles with diameters between 2 and 5 nm, computed from the energetics of the Bayesian linear regression, Brønsted−Evans−Polanyi relations, and microkinetic modeling. The active site structure, which are the corners of the $Rh_{(1-x)}Au_x$ alloy nanoparticle, is shown inset. Oxygen atom = Red sphere; Rhodium atom = Silver sphere; Gold atom = Brown sphere. Nitrogen and NO are not shown. Adapted with permission from Ref. 40. Copyright 2017 American Chemical Society.

modeling strategies and enable accurate screening of large alloy catalyst spaces.[40–42] For example, using Bayesian linear regression (trained on DFT-computed adsorption energies) and Brønsted−Evans−Polanyi relations (which relates the enthalpy of reaction to the activation energy),[43] the effects of alloy composition, nanoparticle size, and surface segregation on NO decomposition turnover frequency (TOF) by $Rh_{(1-x)}Au_x$ nanoparticles were explored, Figure 2.[40] SOAP (smooth overlap atomic position) was used as the kernel in their Bayesian linear regression scheme to approximate the similarity between two local atomic environments based on overlap integrals of three-dimensional atomic distributions.[29,30] After the SOAP-based model is trained, it enables quick estimates of reaction energetics on alloy nanoparticles using only energetic data of single crystal surfaces, Figure 2A. This analysis suggests 2 nm $Rh_{(1-x)}Au_x$ particles with $x \approx 0.33$ have a high TOF, with the most active sites being at the nanoparticle corners, Figure 2B, whereas larger nanoparticles are less active. This work shows kinetic analysis using energetics estimated by ML can be useful to predict size-dependent activity of alloy nanoparticles with reduced computational expense.

Neural networks (NNs) and linear scaling relations[44] (relating adsorption energies of similar species) were used to screen >1000 bimetallic alloys as methanol electrooxidation catalysts for direct methanol fuel cells.[41] The NNs were trained on ~1000 DFT-computed CO and OH adsorption energies on (111)-terminated alloy surfaces using the electronic properties of the metal surface site (e.g., d-band center[45]) and the physical properties of the substrate (e.g., atomic radius) as NN input features. The NNs identified several compositions of transition metal alloys (e.g., Pt/Ru, Pt/Co, Pt/Fe) and structural motifs that exhibit lower theoretical limiting potentials (defined as the minimal potential where all reaction steps are downhill in free energy) than Pt, which agrees with experiments.

A combined DFT and NN iterative approach was used to exhaustively screen Ni_xGa_y bimetallic surfaces for CO_2 reduction activity.[46] CO binding energy was chosen as the target property for screening active facets because surfaces that weakly adsorb CO are linked to greater activity for CO_2 reduction.[47] The Ni_xGa_y system is difficult to model using DFT alone because each composition can exhibit several stable structures at reducing potentials, with each structure having dozens of possible exposed surface facets. The use of a NN to accelerate the search process reduced the number of DFT calculations by an order of magnitude and enabled the study of four bulk compositions (Ni, NiGa, Ni_3Ga, and Ni_5Ga_3), 40 surface facets, and 583 unique adsorption sites for CO_2 reduction activity.

Ultimately, NiGa(210), NiGa(110), and Ni_5Ga_3(021) were predicted to be among the most active surface facets for CO_2 reduction. These active facets all display active Ni atoms surrounded by surface Ga atoms, which rationalizes experimental reports of Ni_xGa_y activity.[48] Some of these active facets could have been missed using conventional, nonexhaustive, search strategies.

Surface phase diagrams help to determine catalyst active sites and reaction mechanisms because they reveal the expected composition and surface phase as a function of temperature, pressure, potential, or dopant concentration.[49] Surface phase diagrams are difficult to obtain by experiment, thus QM modeling is advantageous to predict stable surface structures under reaction conditions. A DFT-trained Gaussian process regression (GPR) model was shown to more quickly and comprehensively predict catalyst surface phase diagrams than conventional intuition-based approaches.[42] Specifically, rapid construction of Pourbaix diagrams, which map surface phases as a function of applied potential and pH, was shown for IrO_2 and MoS_2 surfaces under conditions relevant to the electrocatalytic reduction of N_2 to NH_3.[42] The GPR model, trained on 20–30 adsorbate configurations computed using DFT, estimates the probability that a given set of surface coverages contains configurations relevant to the Pourbaix-stable phase.[42] The computational cost to obtain Pourbaix diagrams of IrO_2 and MoS_2 was reduced by three times using the GPR model compared with manually trying adsorbate configurations informed by physical intuition. Unintuitive and stable surface coverages were identified using GPR that were missed using approaches based on physical intuition.

These studies show ML combined with QM modeling can enable the systematic screening of large catalyst spaces and give unexpected solutions to complex catalysis problems. ML permits exhaustive searches of a given design space with dramatically reduced computational expense compared with QM calculations, revealing both intuitive and unintuitive information. Such ML approaches are expected to be adopted by the community to help identify active catalyst facets and alloy compositions.

Finding descriptors and patterns in catalysis data

A descriptor is a computationally inexpensive surrogate model for some more complicated figure of merit,[50] such as stability, activity, and selectivity in heterogeneous catalysis. The most prevalent descriptor in heterogeneous catalysis is the energy of the d-band center with respect to the Fermi level,[45] which is connected to the interaction between adsorbate valence states and the d-states of a transition metal surface. Consequently, molecule adsorption energies on transition metal surfaces linearly correlate with the d-band center, which can then be related to catalyst activity through linear scaling relations.[45] Other catalyst descriptors[51] derived by intuition exist such as the "generalized" coordination number[52] or "orbital-wise" coordination number,[53] which can estimate the chemical reactivity of nanoparticle catalysts by rationally counting the atoms (or their orbital overlap) that influence the electronic structure of each catalyst site. Such descriptors are powerful but have limitations in accuracy and generalizability. For example, very electronegative adsorbates on substrates with a nearly filled d-band (e.g., OH adsorption on platinum alloys) are a family of common adsorbate-substrate systems that are not well described by the d-band model.[54]

More accurate and generalizable descriptors to predict catalyst figures of merit may exist but remain undiscovered. ML tools for descriptor identification could surpass human intuition to find new, potentially superior, descriptors. It is also possible ML tools could combine known descriptors in unintuitive ways to produce a single more accurate descriptor. To find catalyst descriptors using ML, the set of potential features from which the descriptor is learned must contain the chemistry and physics relevant to the target property of interest. Thus, generating or constructing relevant catalyst features for a given problem is critical.

Using catalyst features that do not require QM calculations can accelerate catalyst prediction and screening. For example, although the d-band center predicts adsorption energies on metal surfaces, its computation requires QM (typically, DFT) calculations. A kernel ridge regression* (KRR)[55] model was trained to predict CO adsorption energy on 263 alloy surfaces using the d-band width of the muffin-tin orbital and the geometric mean of electronegativity as features, which both can be obtained without QM calculations.[56] After training, this KRR model was used to screen CO_2 reduction reaction core-shell catalysts, with Cu_3Zr@Cu and Cu_3Y@Cu predicted to be more active than Au-based catalysts. Another study used gradient boosting regression to quickly estimate the d-band center for 11 monometallic and 110 bimetallic surfaces based on tabulated features such as the density and the enthalpy of fusion of each metal.[57] Because adsorption energies are related to catalyst activity through linear scaling relations, rapidly predicting adsorption energies can yield catalyst activity trends on metal and alloy surfaces.

Although nonlinear regression models are predictive and can consist of physically motivated features,[58] a common criticism is their relative lack of physical interpretability due to their high dimensionality and nonlinearity. Yet, sensitivity analysis can be applied to random forests or neural networks to estimate the relative importance of features in the model.[41,59] Nonetheless, if the goal is to understand the chemical mechanism of catalysts instead of simply fitting data, then low dimensional models are desirable.[60]

Compressed sensing based feature selection methods can give linear, low-dimensional models (i.e., the number of dimensions is just the number of terms in a linear expansion), which offers a robust and fast approach to find simple descriptors of materials to predict target properties.[50,61] In particular, a recently created algorithm called Sure Independence Screening and Sparsifying Operator (SISSO) finds low-dimensional descriptors out of a huge feature space (billions of features) within the framework of compressed-sensing based dimensionality reduction.[31] SISSO has been used by some of the authors to find an improved descriptor to predict the stability of perovskite oxide and halide materials using an experimental dataset.[62] The linearity and simplicity of the descriptors found by SISSO can make them more transferable to materials outside of the training set than nonlinear models, which are prone to overfitting. Although not currently applied to an example relevant for catalysis, SISSO is expected to aid the discovery of descriptors that map catalyst features to their figures of merit.

Data mining methods are powerful ML tools to find nontrivial insights in big data and to help build predictive models. Efforts have been made to integrate data mining methods with heterogeneous or homogeneous catalysis data to promote catalyst characterization and to build quantitative structure-property relationship models.[63–67] An early study used data mining to help make predictive models of cyclohexene epoxidation yield by mesoporous titanium-silicate catalysts.[63] In this study, principal component analysis† (PCA) was used to extract spectra features from X-ray diffraction (XRD) characterization data of 63 catalysts. The composition of the starting synthesis gel and XRD spectra features were used as NN inputs to classify the catalyst epoxide yield. XRD spectra features markedly improved catalyst performance predictions compared with using only synthesis parameters.

Besides helping to extract predictive features, data mining can find trends in catalytic reactions.[64,68] For example, selective hydrogenation of 5-ethoxymethylfurfural was examined over 96 bimetallic catalysts and 16 metal catalysts supported on either SiO_2 or Al_2O_3.[64] Each catalyst was tested in two solvents (diethyl carbonate, 1,4-dioxane) and three temperatures. Using PCA, major trends in the dataset regarding the impact of the support, temperature, solvent, and metal for the hydrogenation of 5-ethoxymethylfurfural were found; for example, SiO_2-supported catalysts typically have much lower activity than Al_2O_3-supported catalysts and higher conversions are

*KRR is a nonlinear version of ridge regression similar to the least squares procedure, except it penalizes the sizes of the regression coefficients. The type of nonlinearity in KRR is determined by the choice of kernel.

†PCA is a method that transforms a number of features into a smaller number of uncorrelated features called principal components, which best separate the data points.

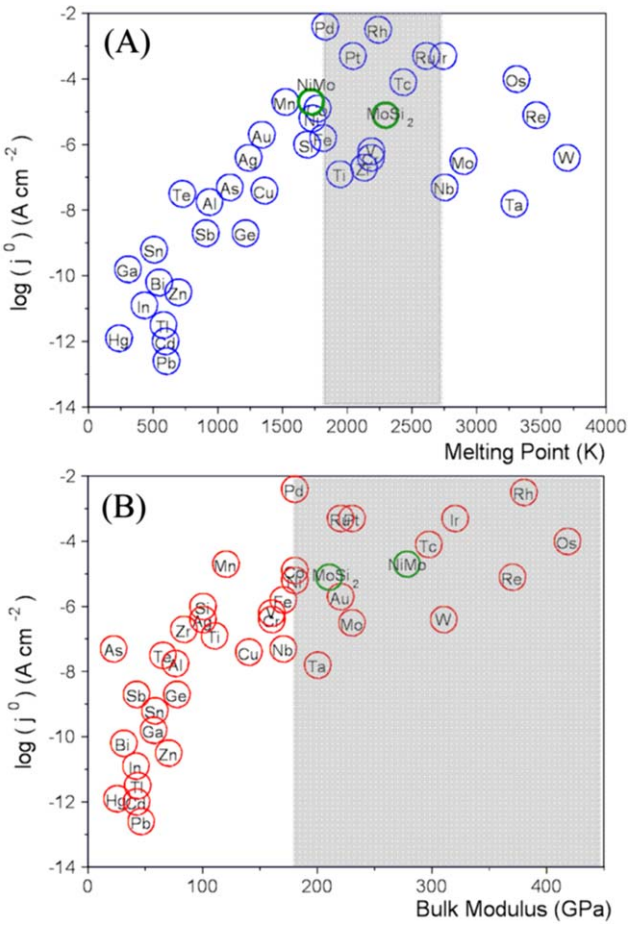

Figure 3. Log(j^o) for the hydrogen evolution reaction in acid vs. (A) melting point and (B) bulk modulus for the elemental metals. Gray regions indicate optimum ranges of the melting point and bulk modulus. NiMo and $MoSi_2$ (green circles) follow the melting point and bulk modulus correlations of the elemental metals. Adapted from Ref. 66 with permission. Copyright 2013 American Chemical Society.

obtained using diethyl carbonate as a solvent compared to 1,4-dioxane.

Data mining found strong correlations between bulk material properties of elemental metals and their experimental hydrogen evolution reaction (HER) kinetics in acid. A dataset containing 38 elemental metals and 50 bulk materials properties were mined for correlations with HER exchange current densities (j^0) using the Reshef algorithm.[66] Interestingly, the melting point and bulk modulus of the metals gave correlations slightly stronger than those of the d-band center for HER activity, and these correlations remained true for the promising NiMo HER electrocatalyst and a previously untested $MoSi_2$ catalyst, Figure 3. These case studies show that data mining tools can find hidden patterns in experimental catalysis data and suggest regions in "catalyst space" where improved catalysts are found.

Most ML applications in catalysis infer a global prediction model for some property of interest, but the underlying mechanism for a desired catalyst property could differ for different catalysts within a large amount of data. Consequently, a global model fitted to the entire dataset may be difficult to interpret and incorrectly describe the physical mechanisms. One could instead partition the dataset into chemically similar catalyst subgroups via clustering algorithms and train a separate model on each subgroup, which can increase prediction accuracy by reducing the different physicochemical effects that each ML model must describe. As an alternative, local pattern search algorithms such as subgroup discovery (SGD) could be used to automatically find and describe subgroups.[69]

SGD aims to find and describe local subpopulations in which the target property takes on a useful distribution.[70] The SGD algorithm consists of three main parts: (1) the use of a description language for finding subgroups within a given pool of data; (2) the definition of utility functions that formalize the interestingness of subgroups; and (3) the use of a search algorithm to find selectors that describe interesting subgroups. One of the authors has shown SGD can be used to find descriptors that predict the stable crystal structure for the 82 octet AB binary materials, as well as find patterns and correlations between structural and electronic properties of gold clusters (Au_5–Au_{14}).[71] Unlike global modeling algorithms, SGD could identify potentially unintuitive groupings of catalysts, which (a) enables understanding of physicochemical similarity between systems, and (b) can be used to improve predictive models.

Machine-learned interatomic potentials for catalyst simulation

Modeling catalysts under reaction conditions using QM is computationally expensive because the cost of these approaches scales unfavorably with system size, thus QM applications remain limited to small catalytic systems (hundreds of atoms). To overcome this size constraint, ML is being used to develop interatomic potentials (mathematical functions for computing the potential energy of a system of atoms) trained with data generated by QM, which estimate interaction energies with increased numerical efficiency compared with QM methods.[72] Therefore, these machine-learned interatomic potentials (MLPs) can speed-up simulations by several orders of magnitude while keeping comparable accuracy to QM methods.[73] The small computational cost of MLPs compared with QM methods promises to make them useful to catalytic systems at extended length and time scales, and aid near-exhaustive catalyst structure searches, see Figure 4.

After catalyst structures under operating conditions are determined, mechanistic modeling and microkinetic simulations can be performed to obtain insights and make catalyst predictions, which can next be confirmed by catalyst synthesis, characterization, and testing. Further advances in MLPs are needed, however, to fulfill the vision outlined in Figure 4. In the following section, we will discuss some progress, challenges, and opportunities for MLPs to model catalysis, as well as some ambitious MLPs, which may one day circumvent the need for traditional QM modeling of catalysts.

MLPs have undergone great advances in recent years, which is laying the foundation for MLP applications to catalysis studies. For example, the first molecular dynamics simulation with a machine-learned density functional (trained on DFT

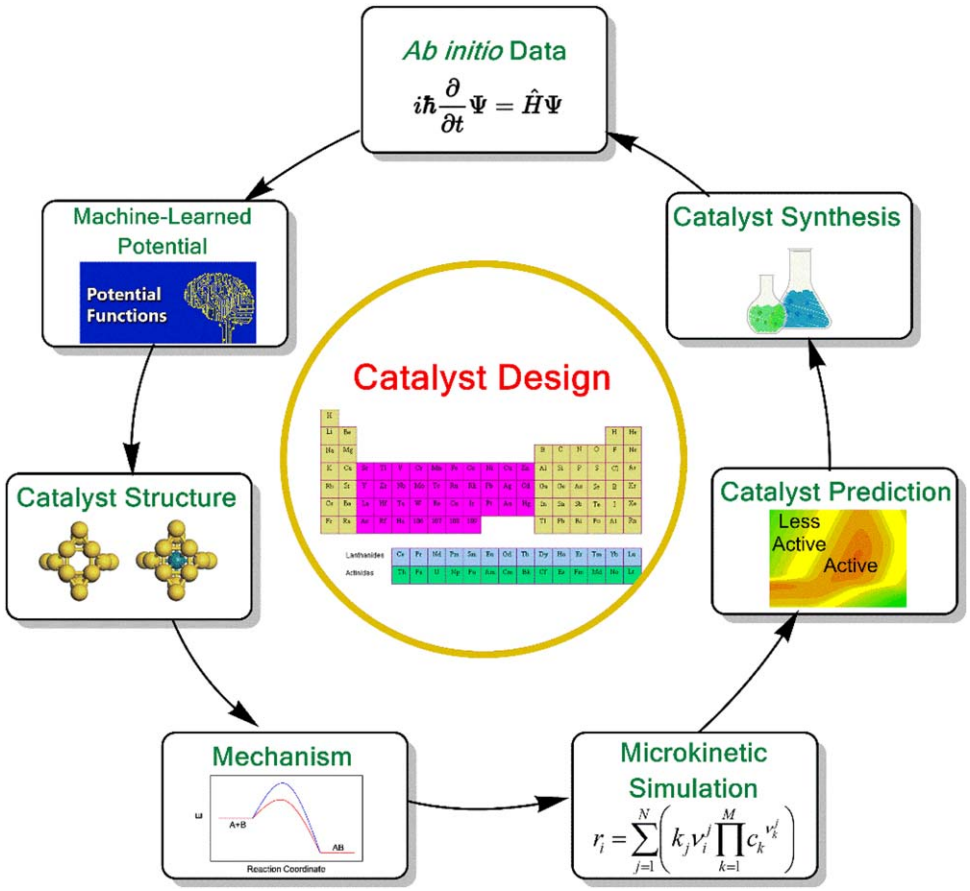

Figure 4. Machine-learned interatomic potentials, trained on high-quality data generated by quantum mechanical (ab initio) methods, can accelerate catalyst structure searches and simulate greater time and length scales. After stable catalyst structures under operating conditions are determined, mechanistic analysis and microkinetic simulations can be performed to extract catalyst design insights and make catalyst predictions, which can next be verified by catalyst synthesis, characterization, and testing. Data of the synthesized catalyst can be obtained by ab initio calculations to close the workflow cycle.

reference data) was used to simulate intramolecular proton transfer within malonaldehyde.[74] MLPs made of deep tensor neural networks can perform highly accurate molecular dynamics simulations of small molecules, classify the relative stability of aromatic rings, as well as give insights on local molecular chemical potentials.[75]

The accuracy of NN interatomic potentials are competitive against popular force fields such as ReaxFF.[76,77] ReaxFF is a bond order-based force field that can predict bond formation/breaking reactions. The Behler–Parrinello neural network (BPNN) potential, which uses symmetry functions to represent the chemical environment of each atom in the system, was benchmarked against ReaxFF for predicting the equation of state, vacancy formation and diffusion barriers for bulk gold, surface diffusion and slipping barriers for gold surfaces, and the most stable gold nanocluster structures for Au_6 and Au_{38}.[76] BPNN was fitted to 9734 DFT calculations (using PBE) and gave an RMSE of 0.021 eV/atom on the validation set, whereas ReaxFF had an RMSE of 0.136 eV/atom over the entire dataset.[76] Although able to achieve high accuracy, one drawback of NN-based MLPs is their computational expense among potentials, which is 1–2 orders of magnitude higher than ReaxFF and classical interatomic potentials because of the more complex representation of the system that is used in combination with the NN.[76,78]

MLPs are being increasingly used to model catalyst dynamics and predict stable surfaces and structures under reaction conditions. Dynamics in catalysis are so ubiquitous that catalysts have been referred to as "living" systems. For example, the distribution and concentration of vacancy sites in catalyst supports can change under reaction conditions and impact catalytic performance.[79,80] Ostwald ripening (the growth of larger nanoparticles from smaller nanoparticles), or nanoparticle disintegration into single atoms are also common dynamic phenomena that can change nanoparticle activity and selectivity.[81,82] A NN interatomic potential combined with grand canonical Monte Carlo (GCMC) predicted the surface coverage of oxygen atoms on a Pd(111) surface as a function of temperature and pressure.[83] Additionally, the NN potential was used with nudged elastic band calculations to predict the

minimum energy pathway for oxygen adatom diffusion on Pd(111) in the dilute limit.

One major challenge is to determine stable catalyst structures under reaction conditions, for example, small nanoclusters can adopt a diverse array of unintuitive structures at elevated temperatures.[84] Supported nanoclusters covered with reactants could adopt a stable geometry or an ensemble of geometries different than those covered with reaction intermediates or products.[84] MLPs could help determine supported nanocluster geometries in the presence of adsorbates through combination of structure-searching methods such as genetic algorithms, basin-hopping and GCMC.[85–90]

Fast and predictive reactive MLPs would be indispensable for simulating challenging systems such as catalysis at liquid/solid interfaces, for which a detailed solvent description is required (e.g., solvent can participate directly in reactions and modify the surface coverage of intermediates) but difficult to achieve in practice.[91] MLPs have been used to study structural and dynamical properties of interfacial water at low-index copper surfaces, including water probability densities, molecular orientations, and hydrogen-bond lifetimes.[92] Combining a MLP with Monte Carlo enabled the characterization of the equilibrium surface structure and composition of bimetallic Au/Cu nanoparticles in aqueous solution, which are relevant CO_2 reduction catalysts.[93,94] Future work involving QM/MLP methods to simulate the active site with high fidelity (using QM) and the rest of environment (using a MLP) would be valuable to model larger catalytic systems and reactions in solution.

One drawback of MLPs is the large amount of data typically needed to achieve predictive accuracy, which often requires many thousands of geometry configurations for training. Recently it was shown, however, that gradient-domain machine learning, which uses exclusively atomic gradient information instead of atomic energies, can construct accurate MLPs from only 1000 geometries obtained from molecular dynamics trajectories (e.g., for benzene, toluene, ethanol, and aspirin).[95] This approach enables molecular dynamics simulations with DFT accuracy for small molecules three orders of magnitude faster than simulations using explicit DFT calculations. Another strategy is to directly machine learn energy functionals (within the framework of Kohn-Sham DFT), which should yield large savings in computer time and allow larger catalytic systems to be studied.[74,96]

Many thousands of scientific articles published each year use QM methods, so these types of machine learning works are exciting because they promise to allow the construction of fast potentials with QM accuracy to simulate catalyst systems. MLPs have shown success to examine molecules, metal surfaces containing adsorbates, and nanoparticles. Yet progress is needed to increase the transferability and generalizability of MLPs, especially for modeling bond-breaking reactions across full catalytic cycles. Developing MLPs to model reactions across full catalytic cycles is challenging because: (1) it is hard to obtain sufficient training data of relevant bond breaking reactions and (2) it is more difficult for MLPs to interpolate bond breaking events than nonbond-breaking events due to the greater change in the chemical properties of a given system. Another challenge to overcome is the difficulty in training accurate MLPs for condensed-phase systems containing above four different elements (because of the exponentially growing size of configuration space with the number of elements). Some of the challenges regarding training MLPs will be alleviated with larger training datasets of accurate QM data becoming more available in data repositories, and from improvements in approaches to understand uncertainty in model predictions.[97] Progress in data sharing and data reuse techniques (e.g., transfer learning)[98] would also promote usage of MLPs to study catalysts via easier access to training data. With the growing availability of software for machine learning potentials such as AMP,[99] PROPhet,[100] and TensorMol[101] it is evident that MLPs will keep being extended.

Accelerating the discovery of catalytic mechanisms

Designing heterogeneous catalysts for a specific reaction requires knowledge of the rate-controlling transition states and intermediates.[102] To understand the key elementary steps and surface abundance intermediates with atomistic detail, the stable structures and the corresponding transition states (TS) that connect them must be known. On the potential energy surface (PES), stable reactant molecules, product molecules, and reaction intermediates are in local or global minima. Catalyst geometry optimization methods to find minima usually involve Conjugate Gradient or Quasi-Newton Raphson methods. A more difficult problem than finding minima is to locate TS structures on heterogeneous catalysts (e.g., bond breaking reactions of adsorbates), which correspond to first-order saddle points on the PES.

TS searching algorithms have aided many computational mechanistic analyses of heterogeneous catalysts. Some of these algorithms are: the Cerjan-Miller algorithm, Climbing-Image Nudged Elastic Band, Dimer method, Force Reversed method, Growing String, and the Single-Ended Growing String.[103–108] Once the transition states for elementary steps are known, catalyst activation free energy barriers and rate constants can be computed.[109] Thus, creating more efficient algorithms to navigate the PES and locate transition states is important to help understand catalytic reactions.

ML can accelerate TS searches and minimum energy path (MEP) finding algorithms. The MEP is the lowest-energy path connecting two minima on the PES (i.e., the path of maximum statistical weight in a system at thermal equilibrium), thus it is kinetically relevant. To accelerate MEP and TS search calculations, a DFT-trained NN was used to estimate the PES for which nudged elastic band (NEB) computations were carried out.[110] Another study used Gaussian process regression (GPR) to speed-up NEB searches to find MEPs for a benchmark system involving 13 rearrangement transitions of a heptamer island on a model solid surface.[111] These ML approaches are surely going to accelerate calculations of MEPs for heterogeneous catalytic processes involving small adsorbates. However, better computational scaling of the GPR calculations will be needed to accelerate MEP calculations of larger systems. Looking ahead, we believe the future of TS and MEP path searching lies in combining ML with automated reaction path search methods.[112,113] Such approaches would create the possibility of exhaustively searching heterogeneous catalyst reaction pathways in an automated fashion to find the relevant thermodynamic and kinetic information of the full catalytic cycle.

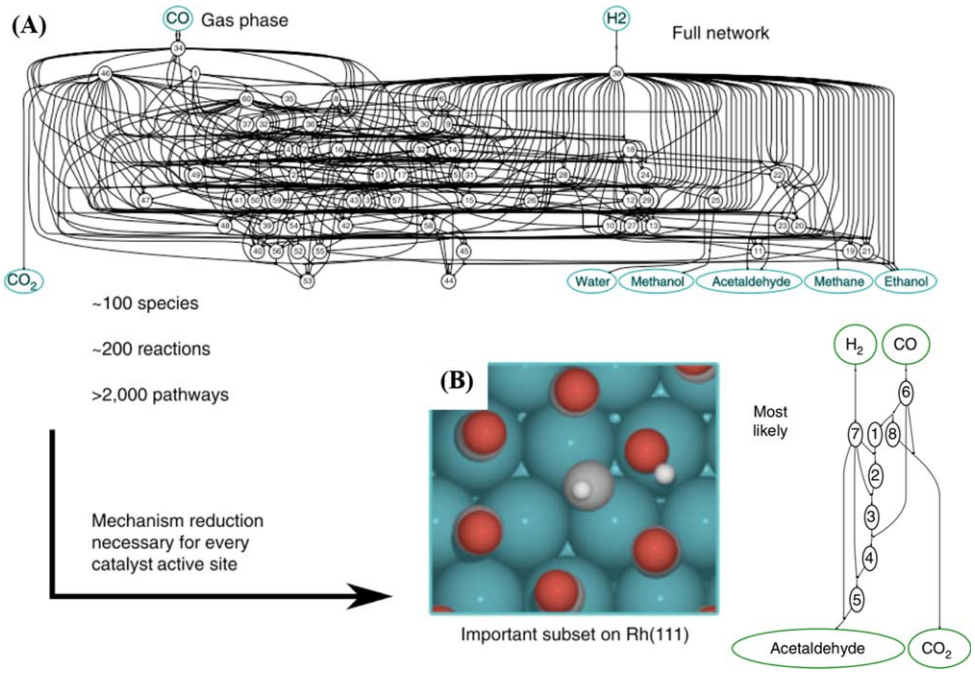

Figure 5. (A) Reaction network for the reaction of $CO + H_2$ (syngas) to CO_2, water, methanol, acetaldehyde, methane, and ethanol, including surface intermediates (containing up to two carbon and two oxygen atoms). (B) The reduced reaction network for $CO + H_2$ reactivity on Rh(111) indicates acetaldehyde and CO_2 are the major products, which is confirmed by experiment. The reduction of the reaction network (A) to the reduced reaction network (B) is achieved using a machine learning aided reaction network optimization framework. Oxygen atom = Red sphere; Rhodium atom = green sphere; Carbon atom = Grey sphere; Hydrogen atom = white sphere. Figure adapted from Ref. 116.

ML approaches also show promise to aid mechanistic studies by helping to address reaction network complexity in a systematic fashion.[114,115] QM modeling can yield insights into reaction mechanisms and improved catalysts for reactions of small molecules, but it is typically computationally prohibitive for complex reaction networks involving large molecules. As a step toward enabling accurate and fast computational predictions of reaction networks, an optimization framework using GPR was applied to study the reaction of syngas ($CO + H_2$) over Rh(111) catalysts under experimentally relevant operating conditions (573 K and 1 atm of gas phase reactants), Figure 5.[114] A reaction network for syngas conversion over Rh(111) is shown in Figure 5A, which has hundreds of species, hundreds of possible reactions, and more than two thousand possible reaction pathways to consider. Starting from a few DFT energies of the intermediates in the reaction network, a computationally inexpensive GPR scheme was used to predict the free energy for all intermediates in the reaction network. TS linear scaling relations were exploited to estimate the activation energies for all reactions in the network, and a simple classifier was used to select the potential rate-limiting steps. Through an iterative GPR model refinement process, where only potential rate-limiting steps were further analyzed using the climbing-image nudged elastic band algorithm, a probable reaction network was identified, Figure 5B.

The most probable reaction mechanism was found using DFT to calculate only 5% of transition state energies and 40% of intermediate species energies, and the mechanism matches the experimentally observed selectivity of Rh(111) toward making acetaldehyde. For analyzing more complex reaction pathways, advances in graph theory-based regression approaches can be used to quickly estimate needed thermochemistry and activation energies.[115] This example once again shows that ML can make more efficient use of CPU time by leveraging catalyst data already obtained by QM methods.

Opportunities and Prospects

Machine learning is a valuable addition to a researcher's toolkit for generating knowledge about heterogeneous catalysts. ML combined with computational modeling or experiments is creating avenues for rapidly screening heterogeneous catalysts, finding descriptors of catalyst performance, and aiding catalyst synthesis. A major application of ML in catalysis is to train predictive models based on quantum mechanical data to enable the systematic screening of large catalyst spaces for adsorbate binding strength and activity. ML approaches can help identify active catalyst facets and alloy compositions. Additionally, applications of machine-learned interatomic potentials promise to allow the simulation of catalytic systems at larger length scales or longer time scales with high accuracy, albeit further methodological development is needed. Other cutting-edge methods for descriptor identification such as SISSO and subgroup discovery can search over a huge space of possible features to find descriptors of catalyst stability, activity, and selectivity.

Literature on heterogeneous catalysis is mounting with numerous catalysts being synthesized, characterized, and tested for catalytic performance. Organizing all the generated catalyst information in databases for storage, query, and sharing is key to fully exploit the power of ML to construct predictive models and to find patterns in catalysis data. However, manually extracting catalyst knowledge from published literature is tedious, time consuming, and can be error prone. Natural language processing and ML would allow automated text and data extraction to uncover scientific insights from this large body of catalysis information. This area is ripe to develop for the catalysis community. Some advances on the text-mining front have already been made in the chemistry[116] and materials science communities.[117,118] Tools are needed to extract catalysis information such as kinetics, thermodynamics, particle size, operating temperature, and synthesis conditions.[68,119] Being able to extract large amounts of catalyst information to fill databases would create routes for innovation through data mining studies.

Another area ready for further innovation is machine learning for catalyst imaging (e.g., scanning transmission electron microscopy, scanning tunneling microscopy, and atomic force microscopy) and spectroscopic (e.g., infrared, X-ray absorption near edge structure) analysis. For example, ML could help generate higher quality images or improved spectra with decreased sampling time, or help interpret experimental spectra.[120,121] Importantly, imaging and spectroscopic data contains quantitative structural and functional information, albeit with high complexity. ML models that map imaging and spectroscopic data to structure-property information would be valuable for catalyst understanding and help link models and experiments.[122,123] Recently, a neural network converted XANES spectra of Pt nanoparticles into information about their atomic-coordination environment to assist with their structural characterization.[123] The neural network was trained on Pt nanoparticle XANES simulations and validated against experiment. This result suggests rapid spectroscopic determination of catalyst morphology is becoming closer to reality through the aid of ML.

From accelerating catalyst active site determination to finding descriptors and patterns in catalysis data, in recent years machine learning has proven to be versatile and useful for aiding heterogeneous catalyst understanding, design, and discovery. The power of machine learning has just begun to be exploited in heterogeneous catalysis research, with much room remaining for advancement (e.g., text mining, image analysis, machine-learned interatomic potentials, and reaction path search algorithms). Further development of machine learning software, algorithms, and techniques promises to aid heterogeneous catalysis design and discovery in the years to come.

Acknowledgments

The authors thank Saswata Bhattacharya, Sergey Levchenko, Suljo Linic, Runhai Ouyang, and Matthias Scheffler for helpful discussions about machine learning for catalysis. B.R.G acknowledges start-up funding from University of Michigan, Ann Arbor. C.S. gratefully acknowledges funding through a postdoctoral fellowship by the Alexander von Humboldt Foundation.

Literature Cited

1. Silver D, Schrittwieser J, Simonyan K, Antonoglou I, Huang A, Guez A, Hubert T, Baker L, Lai M, Bolton A, Chen Y, Lillicrap T, Hui F, Sifre L, van den Driessche G, Graepel T, Hassabis D. Mastering the game of go without human knowledge. *Nature*. 2017; 550 (7676):354.
2. Ramprasad R, Batra R, Pilania G, Mannodi-Kanakkithodi A, Kim C. Machine learning in materials informatics: recent applications and prospects. *Npj Comput Mater*. 2017;3 (1):54.
3. Tabor DP, Roch LM, Saikin SK, Kreisbeck C, Sheberla D, Montoya JH, Dwaraknath S, Aykol M, Ortiz C, Tribukait H, Amador-Bedolla C, Brabec CJ, Maruyama B, Persson KA, Aspuru-Guzik A. Accelerating the discovery of materials for clean energy in the era of smart automation. *Nat Rev Mater*. 3;5:2018.
4. Beck DA, Carothers JM, Subramanian VR, Pfaendtner J. Data science: accelerating innovation and discovery in chemical engineering. *AIChE J*. 2016;62 (5):1402.
5. Friedman J, Hastie T, Tibshirani R. *The Elements of Statistical Learning*, *Vol*. 1. Springer series in statistics, New York, 2001.
6. Kitchin JR. Machine learning in catalysis. *Nat Catal*. 2018;1(4):230.
7. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V. Scikit-learn: machine learning in python. *J Mach Learn Res*. 2011;12:2825.
8. Abadi M, Barham P, Chen J, Chen Z, Davis A, Dean J, Devin M, Ghemawat S, Irving G, Isard M, Kudlur M, Levenberg J, Monga R, Moore S, Murray DG, Steiner B, Tucker P, Vasudevan V, Warden P, Wicke M, Yu Y, Zheng X. *TensorFlow: A System for Large-Scale Machine Learning*, OSDI, 2016:265.
9. Hjorth Larsen A, Jørgen Mortensen J, Blomqvist J, Castelli IE, Christensen R, Dułak M, Friis J, Groves MN, Hammer B, Hargus C, Hermes ED, Jennings PC, Bjerre Jensen P, Kermode J, Kitchin JR, Leonhard Kolsbjerg E, Kubal J, Kaasbjerg K, Lysgaard S, Bergmann Maronsson J, Maxson T, Olsen T, Pastewka L, Peterson A, Rostgaard C, Schiøtz J, Schütt O, Strange M, Thygesen KS, Vegge T, Vilhelmsen L, Walter M, Zeng Z, Jacobsen KW. The Atomic Simulation Environment—A Python library for working with atoms. *J Phys Condens Matter*. 2017;29(27):273002.
10. Mathew K, Montoya JH, Faghaninia A, Dwarakanath S, Aykol M, Tang H, Chu I-h, Smidt T, Bocklund B, Horton M, Dagdelen J, Wood B, Liu Z-K, Neaton J, Ong SP, Persson K, Jain A. Atomate: a high-level interface to generate, execute, and analyze computational materials science workflows. *Comput Mater Sci*. 2017; 139:140.
11. Ghiringhelli LM, Carbogno C, Levchenko S, Mohamed F, Huhs G, Lüders M, Oliveira M, Scheffler M. Towards efficient data exchange and sharing for big-data driven materials science: metadata and data formats. *Npj Comput Mater*. 2017;3(1):46.
12. O'Mara J, Meredig B, Michel K. Materials data infrastructure: a case study of the citrination platform to

examine data import, storage, and access. *JOM*. 2016; 68(8):2031.
13. Jain A, Ong SP, Hautier G, Chen W, Richards WD, Dacek S, Cholia S, Gunter D, Skinner D, Ceder G, Persson KA. Commentary: the Materials Project: a materials genome approach to accelerating materials innovation. *APL Mater*. 2013;1(1):011002.
14. Hummelshøj JS, Abild-Pedersen F, Studt F, Bligaard T, Nørskov JK. CatApp: a web application for surface chemistry and heterogeneous catalysis. *Angew Chem Int Ed*. 2012;124(1):278.
15. van Santen RA. *Modern Heterogeneous Catalysis: An Introduction*. Weinheim, Germany: John Wiley & Sons, 2017:592.
16. Kalz KF, Kraehnert R, Dvoyashkin M, Dittmeyer R, Gläser R, Krewer U, Reuter K, Grunwaldt JD. Future challenges in heterogeneous catalysis: understanding catalysts under dynamic reaction conditions. *ChemCatChem*. 2017;9(1):17.
17. Goldsmith BR, Peters B, Johnson JK, Gates BC, Scott SL. Beyond ordered materials: understanding catalytic sites on amorphous solids. *ACS Catal*. 2017;7(11):7543.
18. Gross EK, Dreizler RM. *Density Functional Theory*, *Vol*. 337. Berlin/Heidelberg, Germany: Springer Science & Business Media, 2013.
19. Carter EA. Challenges in modeling materials properties without experimental input. *Science*. 2008;321(5890):800.
20. Ras E-J, Rothenberg G. Heterogeneous catalyst discovery using 21st century tools: a tutorial. *RSC Adv*. 2014; 4(12):5963.
21. Hattori T, Kito S. Neural network as a tool for catalyst development. *Catal Today*. 1995;23(4):347.
22. Sasaki M, Hamada H, Kintaichi Y, Ito T. Application of a neural network to the analysis of catalytic reactions analysis of NO decomposition over Cu/ZSM-5 zeolite. *Appl Catal A*. 1995;132(2):261.
23. Mueller T, Kusne AG, Ramprasad R. Machine learning in materials science: recent progress and emerging applications. *Rev Comput Chem*. 2016;29:186.
24. Rothenberg G. Data mining in catalysis: separating knowledge from garbage. *Catal Today*. 2008;137(1):2.
25. Fernandez M, Barron H, Barnard AS. Artificial neural network analysis of the catalytic efficiency of platinum nanoparticles. *RSC Adv*. 2017;7(77):48962.
26. Maldonado AG, Rothenberg G. Predictive modeling in homogeneous catalysis: a tutorial. *Chem Soc Rev*. 2010; 39(6):1891.
27. Janet JP, Kulik HJ. Resolving transition metal chemical space: feature selection for machine learning and structure–property relationships. *J Phys Chem A*. 2017; 121(46):8939.
28. Janet JP, Chan L, Kulik HJ. Accelerating chemical discovery with machine learning: simulated evolution of spin crossover complexes with an artificial neural network. *J Phys Chem Lett*. 2018; 9 (5):1064.
29. Bartók AP, Kondor R, Csányi G. On representing chemical environments. *Phys Rev B*. 2013;87(21):184115.
30. Bartók AP, Kondor R, Csányi G. Erratum: on representing chemical environments [Phys. Rev. B 87, 184115 (2013)]. *Phys Rev B*. 2017;96(1):019902.
31. Ouyang R, Curtarolo S, Ahmetcik E, Scheffler M, Ghiringhelli LM. SISSO: a compressed-sensing method for systematically identifying efficient physical models of materials properties. *arXiv preprint arXiv:1710.03319*, 2017.
32. Senkan SM. High-throughput screening of solid-state catalyst libraries. *Nature*. 1998;394(6691):350.
33. Baumes L, Farrusseng D, Lengliz M, Mirodatos C. Using artificial neural networks to boost high-throughput discovery in heterogeneous catalysis. *Mol Inform*. 2004; 23(9):767.
34. Baumes L, Serra J, Serna P, Corma A. Support vector machines for predictive modeling in heterogeneous catalysis: a comprehensive introduction and overfitting investigation based on two real applications. *J Comb Chem*. 2006;8(4):583.
35. Cleve TV, Moniri S, Belok G, More KL, Linic S. Nanoscale engineering of efficient oxygen reduction electrocatalysts by tailoring the local chemical environment of Pt surface Sites. *ACS Catal*. 2017;7(1):17.
36. Alonso DM, Wettstein SG, Dumesic JA. Bimetallic catalysts for upgrading of biomass to fuels and chemicals. *Chem Soc Rev*. 2012;41(24):8075.
37. Yu W, Porosoff MD, Chen JG. Review of Pt-based bimetallic catalysis: from model surfaces to supported catalysts. *Chem Rev*. 2012;112(11):5780.
38. Andersen M, Medford AJ, Nørskov JK, Reuter K. Scaling-relation-based analysis of bifunctional catalysis: the case for homogeneous bimetallic alloys. *ACS Catal*. 2017;7(6):3960.
39. Peters B, Scott SL. Single atom catalysts on amorphous supports: a quenched disorder perspective. *J Chem Phys*. 2015;142(10):104708.
40. Jinnouchi R, Asahi R. Predicting catalytic activity of nanoparticles by a DFT-aided machine-learning algorithm. *J Phys Chem Lett*. 2017;8(17):4279.
41. Li Z, Wang S, Chin WS, Achenie LE, Xin H. High-throughput screening of bimetallic catalysts enabled by machine learning. *J Mater Chem A*. 2017;5(46):24131.
42. Ulissi ZW, Singh AR, Tsai C, Nørskov JK. Automated discovery and construction of surface phase diagrams using machine learning. *J Phys Chem Lett*. 2016;7(19): 3931.
43. van Santen RA. *Molecular Catalytic Kinetics Concepts*. Weinheim: WILEY-VCH Verlag GmbH, 2010.
44. Greeley J. Theoretical heterogeneous catalysis: scaling relationships and computational catalyst design. *Annu Rev Chem Biomol Eng*. 2016;7(1):605.
45. Nørskov JK, Bligaard T, Rossmeisl J, Christensen CH. Towards the computational design of solid catalysts. *Nat Chem*. 2009;1(1):37.
46. Ulissi ZW, Tang MT, Xiao J, Liu X, Torelli DA, Karamad M, Cummins K, Hahn C, Lewis NS, Jaramillo TF, Chan K, Nørskov JK. Machine-learning methods enable exhaustive searches for active bimetallic facets and reveal active site motifs for CO_2 reduction. *ACS Catal*. 2017;7(10):6600.
47. Peterson AA, Nørskov JK. Activity descriptors for CO_2 electroreduction to methane on transition-metal catalysts. *J Phys Chem Lett*. 2012;3(2):251.

48. Torelli DA, Francis SA, Crompton JC, Javier A, Thompson JR, Brunschwig BS, Soriaga MP, Lewis NS. Nickel–gallium-catalyzed electrochemical reduction of CO_2 to highly reduced products at low overpotentials. *ACS Catal*. 2016;6(3):2100.
49. Reuter K, Stampf C, Scheffler M. Ab initio atomistic thermodynamics and statistical mechanics of surface properties and functions. In: Yip S, editor. *Handbook of Materials Modeling*, Dordrecht: Springer, 2005:149.
50. Ghiringhelli LM, Vybiral J, Levchenko SV, Draxl C, Scheffler M. Big data of materials science: critical role of the descriptor. *Phys Rev Lett*. 2015;114(10): 105503.
51. Sinthika S, Waghmare UV, Thapa R. Structural and electronic descriptors of catalytic activity of graphene-based materials: first-principles theoretical analysis. *Small*. 2018;14(10):1703609.
52. Calle-Vallejo F, Tymoczko J, Colic V, Vu QH, Pohl MD, Morgenstern K, Loffreda D, Sautet P, Schuhmann W, Bandarenka AS. Finding optimal surface sites on heterogeneous catalysts by counting nearest neighbors. *Science*. 2015;350(6257):185.
53. Ma X, Xin H. Orbitalwise coordination number for predicting adsorption properties of metal nanocatalysts. *Phys Rev Lett*. 2017;118(3):036101.
54. Xin H, Linic S. Communications: exceptions to the d-band model of chemisorption on metal surfaces: the dominant role of repulsion between adsorbate states and metal d-states. *J Chem Phys*. 2010;132 (22):221101.
55. Rupp M. Machine learning for quantum mechanics in a nutshell. *Int J Quantum Chem*. 2015;115(16):1058.
56. Noh J, Kim J, Back S, Jung Y. Catalyst design using actively learned machine with non-ab initio input features towards CO_2 reduction reactions. *arXiv preprint arXiv:1709.04576*, 2017.
57. Takigawa I, Shimizu K-I, Tsuda K, Takakusagi S. Machine-learning prediction of the d-band center for metals and bimetals. *RSC Adv*. 2016;6 (58):52587.
58. Li Z, Ma X, Xin H. Feature engineering of machine-learning chemisorption models for catalyst design. *Catal Today*. 2017;280 (Part 2):232.
59. Wexler RB, Martirez JMP, Rappe AM. Chemical pressure-driven enhancement of the hydrogen evolving activity of Ni2P from nonmetal surface doping interpreted via machine learning. *J Am Chem Soc*. 2018; 140(13):4678.
60. Pankajakshan P, Sanyal S, de Noord OE, Bhattacharya I, Bhattacharyya A, Waghmare U. Machine learning and statistical analysis for materials science: stability and transferability of fingerprint descriptors and chemical insights. *Chem Mater*. 2017;29(10):4190.
61. Ghiringhelli LM, Vybiral J, Ahmetcik E, Ouyang R, Levchenko SV, Draxl C, Scheffler M. Learning physical descriptors for materials science by compressed sensing. *New J Phys*. 2017;19(2):023017.
62. Bartel CJ, Sutton C, Goldsmith BR, Ouyang R, Musgrave CB, Ghiringhelli LM, Scheffler M. New tolerance factor to predict the stability of perovskite oxides and halides. *arXiv preprint arXiv:1801.07700*, 2018.
63. Corma A, Serra JM, Serna P, Moliner M. Integrating high-throughput characterization into combinatorial heterogeneous catalysis: unsupervised construction of quantitative structure/property relationship models. *J Catal*. 2005;232(2):335.
64. Ras E-J, McKay B, Rothenberg G. Understanding catalytic biomass conversion through data mining. *Top Catal*. 2010;53(15–18):1202.
65. Madaan N, Shiju NR, Rothenberg G. Predicting the performance of oxidation catalysts using descriptor models. *Catal Sci Technol*. 2016;6(1):125.
66. Leonard KC, Bard AJ. Pattern recognition correlating materials properties of the elements to their kinetics for the hydrogen evolution reaction. *J Am Chem Soc*. 2013; 135(42):15885.
67. Ras E-J, Louwerse MJ, Rothenberg G. New tricks by very old dogs: predicting the catalytic hydrogenation of HMF derivatives using Slater-type orbitals. *Catal Sci Technol*. 2012;2(12):2456.
68. Odabaşı Ç, Günay ME, Yıldırım R. Knowledge extraction for water gas shift reaction over noble metal catalysts from publications in the literature between 2002 and 2012. *Int J Hydrogen Energy*. 2014;39(11):5733.
69. Boley M, Goldsmith BR, Ghiringhelli LM, Vreeken J. Identifying consistent statements about numerical data with dispersion-corrected subgroup discovery. *Data Min Knowl Discov*. 2017;31(5):1391.
70. Herrera F, Carmona CJ, González P, Del Jesus MJ. An overview on subgroup discovery: foundations and applications. *Knowl Inf Syst*. 2011;29(3):495.
71. Goldsmith BR, Boley M, Vreeken J, Scheffler M, Ghiringhelli LM. Uncovering structure-property relationships of materials by subgroup discovery. *New J Phys*. 2017;19(1):013031.
72. Shapeev AV. Moment tensor potentials: a class of systematically improvable interatomic potentials. *Multiscale Model Sim*. 2016;14(3):1153.
73. Botu V, Batra R, Chapman J, Ramprasad R. Machine learning force fields: construction, validation, and outlook. *J Phys Chem C*. 2017;121(1):511.
74. Brockherde F, Vogt L, Li L, Tuckerman ME, Burke K, Müller K-R. Bypassing the Kohn-Sham equations with machine learning. *Nat Commun*. 2017;8(1):872.
75. Schütt KT, Arbabzadah F, Chmiela S, Müller KR, Tkatchenko A. Quantum-chemical insights from deep tensor neural networks. *Nat Commun*. 2017;8:13890.
76. Boes JR, Groenenboom MC, Keith JA, Kitchin JR. Neural network and ReaxFF comparison for Au properties. *Int J Quantum Chem*. 2016;116(13):979.
77. Dolgirev PE, Kruglov IA, Oganov AR. Machine learning scheme for fast extraction of chemically interpretable interatomic potentials. *AIP Adv*. 2016;6(8):085318.
78. Behler J. First principles neural network potentials for reactive simulations of large molecular and condensed systems. *Angew Chem Int Ed*. 2017;56(42):12828.
79. Campbell CT, Peden CH. Oxygen vacancies and catalysis on ceria surfaces. *Science*. 2005;309(5735):713.
80. Su Y-Q, Filot IAW, Liu J-X, Tranca I, Hensen EJM. Charge transport over the defective $CeO_2(111)$ surface. *Chem Mater*. 2016;28(16):5652.

81. Goldsmith BR, Sanderson ED, Ouyang R, Li W-X. CO- and NO-Induced disintegration and redispersion of three-way catalysts rhodium, palladium, and platinum: an ab initio thermodynamics study. *J Phys Chem C*. 2014; 118(18):9588.
82. Su Y-Q, Liu J-X, Filot IAW, Hensen EJM. Theoretical study of ripening mechanisms of Pd clusters on ceria. *Chem Mater*. 2017;29(21):9456.
83. Boes JR, Kitchin JR. Neural network predictions of oxygen interactions on a dynamic Pd surface. *Mol Simul*. 2017;43(5–6):346.
84. Zhai H, Alexandrova AN. Fluxionality of catalytic clusters: when it matters and how to address it. *ACS Catal*. 2017;7(3):1905.
85. Ouyang R, Xie Y, Jiang D-e. Global minimization of gold clusters by combining neural network potentials and the basin-hopping method. *Nanoscale*. 2015;7(36):14817.
86. Senftle TP, van Duin AC, Janik MJ. Methane activation at the Pd/CeO_2 interface. *ACS Catal*. 2017;7(1):327.
87. Boes JR, Kitchin JR. Modeling segregation on AuPd(111) surfaces with density functional theory and Monte Carlo simulations. *J Phys Chem C*. 2017;121(6):3479.
88. Zhai H, Alexandrova AN. Ensemble-average representation of Pt clusters in conditions of catalysis accessed through GPU accelerated deep neural network fitting global optimization. *J Chem Theory Comput*. 2016; 12(12):6213.
89. Sun G, Sautet P. Metastable structures in cluster catalysis from first-principles: structural ensemble in reaction conditions and metastability triggered reactivity. *J Am Chem Soc*. 2018;140(8):2812.
90. Liu J-X, Su Y, Filot IA, Hensen EJ. A linear scaling relation for CO oxidation on CeO_2-supported Pd. *J Am Chem Soc*. 2018;140(13):4580.
91. Sievers C, Noda Y, Qi L, Albuquerque EM, Rioux RM, Scott SL. Phenomena affecting catalytic reactions at solid–liquid interfaces. *ACS Catal*. 2016;6(12):8286.
92. Natarajan SK, Behler J. Neural network molecular dynamics simulations of solid–liquid interfaces: water at low-index copper surfaces. *Phys Chem Chem Phys*. 2016;18(41):28704.
93. Artrith N, Kolpak AM. Understanding the composition and activity of electrocatalytic nanoalloys in aqueous solvents: a combination of DFT and accurate neural network potentials. *Nano Lett*. 2014;14(5):2670.
94. Artrith N, Kolpak AM. Grand canonical molecular dynamics simulations of Cu–Au nanoalloys in thermal equilibrium using reactive ANN potentials. *Comput Mater Sci*. 2015;110:20.
95. Chmiela S, Tkatchenko A, Sauceda HE, Poltavsky I, Schütt KT, Müller K-R. Machine learning of accurate energy-conserving molecular force fields. *Sci Adv*. 2017; 3(5):e1603015.
96. Li L, Snyder JC, Pelaschier IM, Huang J, Niranjan UN, Duncan P, Rupp M, Müller KR, Burke K. Understanding machine-learned density functionals. *Int J Quantum Chem*. 2016;116(11):819.
97. Peterson AA, Christensen R, Khorshidi A. Addressing uncertainty in atomistic machine learning. *Phys Chem Chem Phys*. 2017;19(18):10978.
98. Hutchinson ML, Antono E, Gibbons BM, Paradiso S, Ling J, Meredig B. Overcoming data scarcity with transfer learning. *arXiv preprint arXiv:1711.05099*, 2017.
99. Khorshidi A, Peterson AA. Amp: a modular approach to machine learning in atomistic simulations. *Comput Phys Commun*. 2016;207:310.
100. Kolb B, Lentz LC, Kolpak AM. Discovering charge density functionals and structure-property relationships with PROPhet: a general framework for coupling machine learning and first-principles methods. *Sci Rep*. 2017;7(1):1192.
101. Yao K, Herr JE, Toth DW, Mckintyre R, Parkhill J. The TensorMol-0.1 model chemistry: a neural network augmented with long-range physics. *Chem Sci*. 2018; 9(8):2261.
102. Campbell CT. The degree of rate control: a powerful tool for catalysis research. *ACS Catal*. 2017;7(4):2770.
103. Hratchian HP, Schlegel HB. Finding minima, transition states, and following reaction pathways on ab initio potential energy surfaces. In Dykstra C, Frenking G, Kim K, Scuseria G (editors.), *Theory and Applications of Computational Chemistry: The first forty years*. Amsterdam: Elsevier, 2005:195.
104. Heyden A, Bell AT, Keil FJ. Efficient methods for finding transition states in chemical reactions: comparison of improved dimer method and partitioned rational function optimization method. *J Chem Phys*. 2005; 123(22):224101.
105. Schlegel HB. Exploring potential energy surfaces for chemical reactions: an overview of some practical methods. *J Comput Chem*. 2003;24(12):1514.
106. Zimmerman PM. Single-ended transition state finding with the growing string method. *J Comput Chem*. 2015; 36(9):601.
107. Jafari M, Zimmerman PM. Reliable and efficient reaction path and transition state finding for surface reactions with the growing string method. *J Comput Chem*. 2017;38(10):645.
108. Sun K, Zhao Y, Su H-Y, Li W-X. Force reversed method for locating transition states. *Theor Chem Acc*. 2012;131(2):1118.
109. Peters B. *Reaction Rate Theory and Rare Events*, 1 ed. Amsterdam, Netherlands: Elsevier Science, 2017.
110. Peterson AA. Acceleration of saddle-point searches with machine learning. *J Chem Phys*. 2016;145(7): 074106.
111. Koistinen O-P, Dagbjartsdóttir FB, Ásgeirsson V, Vehtari A, Jónsson H. Nudged elastic band calculations accelerated with Gaussian process regression. *J Chem Phys*. 2017;147(15):152720.
112. Martínez-Núñez E. An automated method to find transition states using chemical dynamics simulations. *J Comput Chem*. 2015;36(4):222.
113. Zimmerman PM. Navigating molecular space for reaction mechanisms: an efficient, automated procedure. *Mol Simul*. 2015; 41 (1–3):43.
114. Ulissi ZW, Medford AJ, Bligaard T, Nørskov JK. To address surface reaction network complexity using scaling relations machine learning and DFT calculations. *Nat Commun*. 2017;8:14621.

115. Gu GH, Plechac P, Vlachos DG. Thermochemistry of gas-phase and surface species via LASSO-assisted subgraph selection. *React Chem Eng*. 2018.
116. Krallinger M, Rabal O, Lourenço A, Oyarzabal J, Valencia A. Information retrieval and text mining technologies for chemistry. *Chem Rev*. 2017;117(12): 7673.
117. 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.
118. 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(21):9436.
119. Swain MC, Cole JM. ChemDataExtractor: A toolkit for automated extraction of chemical information from the scientific literature. *J Chem Inf Model*. 2016;56(10): 1894.
120. Report of the basic research needs workshop for catalysis science. *Basic Research Needs for Catalysis Science to Transform Energy Technologies*; US DOE Office of Science (United States), 2018:57.
121. Timoshenko J, Keller KR, Frenkel AI. Determination of bimetallic architectures in nanometer-scale catalysts by combining molecular dynamics simulations with x-ray absorption spectroscopy. *J Chem Phys*. 2017;146(11):114201.
122. Kalinin SV, Sumpter BG, Archibald RK. Big-deep-smart data in imaging for guiding materials design. *Nat Mater*. 2015;14(10):973.
123. Timoshenko J, Lu D, Lin Y, Frenkel AI. Supervised machine-learning-based determination of three-dimensional structure of metallic nanoparticles. *J Phys Chem Lett*. 2017;8(20):5091.