Machine Learning for Atomic Simulation and Activity Prediction in Heterogeneous Catalysis: Current Status and Future

Sicong Ma and Zhi-Pan Liu*

ABSTRACT: Heterogeneous catalysis, for its industrial importance and great complexity in structure, has long been the testing ground of new characterization techniques. Machine learning (ML) as a starring tool in data science brings new opportunities for chemists to interpret, simulate, and predict complex reactions in heterogeneous catalysis. Here we review the current status of ML methods and applications in heterogeneous catalysis by following two main streams: the top-down approach by learning experiment data and the bottom-up approach for making predictions from first-principles, which differ in the data source. We focus more on the latter, where ML interacts intimately with first-principles calculations for predicting the key properties (e.g., molecular adsorption energy) and evaluating potential energy surface (PES) to expedite the atomic simulation. The ML-based PES exploration represents the top gear that can largely replace the traditional roles of first-principles calculations for structure determination and activity evaluation but requires efficient methods for data set generation, sensitive structure descriptors to discriminate structures, and iterative self-learning to refine the ML potential. We illustrate these key ingredients of ML-based atomic simulation using the SSW-NN method developed by our group as the example. Three cases of SSW-NN application are presented to elaborate how ML can expedite the material and reaction simulation and lead to new findings on catalyst structure and reaction channels. The future directions of ML-based applications in heterogeneous catalysis are also discussed.

KEYWORDS: machine learning, heterogeneous catalysis, potential energy surface, density functional theory, global optimization, SSW-NN, LASP

1. CHALLENGES OF THEORETICAL CATALYSIS

Heterogeneous catalysis is where material science meets chemical reactions. It is no wonder that every progress in material characterization technique, from the earliest X-ray diffraction (XRD) to the latest transmission electron microscopy (TEM), brings great forward in catalysis science. Theoretical simulations with density functional theory (DFT) calculations enter into the tool boxes of catalysis community in the past 30 years and have become an essential complement to experiment for understanding catalyst structures and catalytic activity. Not only are the reaction mechanisms clarified for a significant number of heterogeneous catalytic reactions, but also general rules for predicting catalytic sites and reactivity are established. To date, much fundamental knowledge from well-defined, crystalline surfaces have been gleaned and the research focus shifts more toward catalytic systems under realistic conditions. In this regard, DFT-based theoretical simulation faces the ever-increasing challenges to cope with the rapidly increased degree-of-freedoms caused by both the high structure complexity and the huge reaction space in catalysis.

The Haber–Bosch process, as a textbook example for illustrating the complexity of heterogeneous catalysis, utilizes alkali-promoted fused iron as catalyst for converting N₂/H₂ to ammonia. It generally contains multiple elements and multiple components: in addition to iron, the presence of alumina, calcium oxide, and potassium are also important to the performance; the catalysts are prepared from magnetite (Fe₃O₄) followed by high-temperature reduction. Apart from the complexity of catalysts, the reaction occurs at high temperature (500 °C) and high pressure (300 atm) with many intermediates, including adsorbed H and NHₓ species. The past decades have seen many DFT-based calculations on Fe-catalyzed ammonia synthesis, which were however limited to single-crystal Fe surfaces (e.g., Fe(111)).

Received: August 9, 2020
Revised: September 11, 2020
Fe(211)\(^9\) as what did in surface science experiment.\(^9\) Obviously, it would be highly desirable for theoretical simulations to predict the active catalyst structures under reaction conditions and to explore the intriguing questions relevant to catalysis, such as the roles of promoters and the effect of particle size.

In fact, the dilemma of catalytic reaction simulation is rooted in the fundamental kinetics laws governed by the Master Equation (eq 1).

\[
\frac{d\mathbf{P}}{dt} = \mathbf{A}\mathbf{P}
\]

where the rate of \(\mathbf{P}\) (a vector represents the population of different states) depends on the transition rate matrix \(\mathbf{A}\). According to transition state theory (TST),\(^{10,11}\) the matrix element \(A_{ij}\) in transition rate matrix is the rate constant between \(i\) and \(j\) states, as summarized in eq 2.

\[
k_{TST} = \frac{k_BT}{h} e^{-\Delta G_{TS}/k_BT}
\]

It suggests that (i) the building of the matrix needs to run over all (important) states, the number of which explodes with the increase of system size and complexity; (ii) the occurrence of states separated by high barriers are exponentially slower compared to those with low barriers. For a reaction to occur at 500 °C with 1 s\(^{-1}\) rate, the free-energy barrier is about 2 eV, which suggests that if the temperature drops to 400 °C, the rate is reduced to 0.01 s\(^{-1}\). This answers the basic questions in catalysis (e.g., why ammonia synthesis requires high temperatures: the stepwise hydrogenations from N to NH\(_3\) have an overall barrier of 1.8–2 eV).\(^{12,13}\)

In general, both the catalyst structure evolution under reaction conditions and the catalytic reactions of molecule belong to rare events with high reaction barriers, which require a long time-scale (from seconds to hours) in simulation in order to observe these processes. As \textit{ab initio} (e.g., DFT) molecular dynamics (MD)\(^{14,15}\) relies on Born–Oppenheimer approximation to solve Schrodinger equations and Newtonian laws to move atoms, the typical time-scale is often limited to less than a nanosecond, i.e. 10\(^6\) times \textit{ab initio} calculations with a typical time-step of 1 fs in MD. On the other hand, the computation speed of DFT, although is sensitive to the complexity of the employed density functional (e.g., PBE,\(^{15}\) HSE06\(^{16}\)), has a poor scaling of at least \(O(N\ln N)\) and is efficient only for small-sized systems within hundreds of atoms. Therefore, instead of straightforward MD, alternative low-cost approaches have to be taken in order to predict catalytic reactions from theory.

Since the late 1990s, the mainstream to study catalytic reactions converges to locate the transition state (TS) of reactions on potential energy surface (PES), from which the reaction barrier and the activity can be determined from theory.\(^{17}\) Many efficient TS search methods have been developed for heterogeneous catalytic reactions. They can be classified as the single-ended approach, i.e. starting from TS-like structure developed for heterogeneous catalytic reactions. They can be classified as the single-ended approach, i.e. starting from TS-like structure (the dimer method and its improved versions\(^{18–20}\)), and double-ended approach, i.e. starting from initial and final states (nudged elastic band\(^{21,22}\) and double-ended surface walking\(^{23}\)). As the TS search requires the information on the structure configurations, it becomes frustrated for reactions with complex reaction configurations such as those involved in surface structure restructuring and unusual reaction intermediates in reaction network.

Another popular way for activity prediction is based on the Bell-Evans–Polanyi (BEP) principle,\(^{24}\) which avoids the time-consuming TS search by establishing the approximate linear correlation between the reaction barrier \(E_s\) and the reaction enthalpy \(\Delta H\) (the energy difference between final and initial states), see eq 3.

\[
E_s = a\Delta H + b
\]

Between 2001 to 2003, several groups (e.g., Neurock, Hu, and Nørskov groups, etc.)\(^{25–27}\) determined independently the parameters of \(a\) and \(b\) in BEP relation by DFT for a large number of surface reactions, which provides the quantitative basis for plotting the Volcano curve of activity, the Sabatier principle known since 1910s, as shown schematically in Figure 1a. The binding strengths of products and reactants need to be in a fine balance, neither too strong nor too weak, to achieve the highest activity.

The BEP approach has been further extended to predict the activity of complex multistep reactions by using simple descriptors that can be computed facilely. For example, the \(d\)-band center \((\varepsilon_d)\) proposed by Hammer and Nørskov\(^{28}\) was known to be a good indicator for the molecular adsorption...
strength on metal surfaces, which can be computed by averaging the occupied d-states \((n_d)\) in eq 4.

\[
e_{\text{d}} = \frac{\int_{-\infty}^{\infty} n_d(e) e \, de}{\int_{-\infty}^{\infty} n_d(e) \, de}
\]  

(4)

Except for the electronic descriptors, the geometry-based descriptors are of most interests. The generalized coordination number (CN) proposed by Sautet group39,40 connects the atomic coordination environment (eq 5) with the binding energies of reaction intermediates for oxygen reduction reaction.

\[
\text{CN} = \sum_{j=1}^{n} \frac{\text{cn}(j)}{\text{cn}_{\text{max}}}
\]

(5)

The sum includes all of the first-nearest neighbor atoms \((\text{cn}(j))\) and the division by the maximum number of first-nearest neighbors in the bulk \((\text{cn}_{\text{max}})\) ensures that CN spans the range between 0 and 12 in face-centered cubic metals. Based on these simple descriptors, the high-throughput screening techniques were then developed for fast evaluating the activity of different materials across the periodic table.3,41

After obtaining the elementary reaction data from first-principles (e.g., the reaction energy and barrier), the mean-field microkinetics or kinetic Monte Carlo (KMC) can be utilized to simulate chemical reactions on surfaces, which will yield the time-resolved catalytic activity and selectivity and provide the fine details on the surface status during reaction dynamics. Although the first-principles data-based KMC simulations remain largely limited to few well-defined crystalline surfaces and relatively simple reactions, the success of these works by carefully benchmarking with experimental kinetic data do represent a marvelous hallmark in theoretical catalysis.3,35

Apart from the difficulty in considering the structural complexity of the catalysts and the tedious computation of the kinetics data for all elementary steps, a major concern for kinetics simulations originates from the intrinsic error of DFT calculations (~0.1 eV error) in computing adsorption energies, reaction barriers, and vibrational entropies for certain molecules/reactions (e.g., CO adsorption). These DFT errors have to be corrected empirically in kinetics simulations in order to yield a more realistic rate for the experimental conditions.

With the development of method and the massive production of data, the 21st century has witnessed a great advance in data-driven science (e.g., machine learning (ML)) and its applications in catalysis research. Similar to the high-throughput approaches that use geometry descriptors for finding trends, ML methods also utilize the geometry-based information for the energy and the property prediction, which renders the high speed in computation. ML applications on energy materials,35 catalysis informatics,36 and heterogeneous catalyst design and discovery37,38 have been reviewed recently. Different from these previous contributions, here we will focus more on the progress of theoretical catalysis via atomic simulation, which is now benefited significantly by ML techniques. We will review ML-based atomic simulation methods utilized in catalysis, in particular stochastic surface walking (SSW) global optimization based on neural network (NN) potential method (SSW-NN) developed in our group.37,40 These ML methods announce the advent of a post-DFT era in theoretical catalysis, which aim to solve challenging catalysis problems that are traditionally intractable from first-principles calculations, either in the time-scale or in the size and the structural complexity of the system.

2. ML FOR HETEROGENEOUS CATALYSIS

As an interdisciplinary technology, ML was invented in the fields of computer science and statistics science,41 which aims to identify trends, rules, or functions from the existing data, instead of from physical laws. With dozens of ML algorithms developed to date, from kernel ridge regression to support vector regression and to sophisticated deep NN, ML techniques have been successfully applied to many fields of science that routinely have access to big data.

The applications in heterogeneous catalysis started in 1990s,42,43 mainly on learning experimental data. For example, Kito et al.44 used NN to predict the product distribution of ethylene oxidative hydrogenation based on experimental catalyst data. They chose different products, including styrene, benzaldehyde, benzene, toluene, CO, and CO₂ as the outputs of NN, and nine different independent variables as input, including the valence, surface area of the catalyst, amount of catalyst, typical valence, ionic radius, coordination number, electronegativity, partial charge of oxygen ion, and standard heat of formation of oxides. The high cost in obtaining a large, consistent data set is the major bottleneck for ML applications in experiment. Compared with experiment, the data produced from DFT calculations is self-consistent and much more economic to obtain. This is particularly advantageous for materials and reactions that are unlikely to synthesize or to occur under ambient conditions. Furthermore, many material gene databases with DFT results are being established around the world,45 which, as a speed-up, enables the convenient data mining for ML of novel materials.

Broadly speaking, there are two branches for ML applications in catalysis (i.e., the top-down and bottom-up approaches). The former is to directly predict catalytic performance based on existing experimental data as those started from 1990s. The latter learns the results from quantum mechanics (QM) calculations, aiming to predict property and evaluate PES, which can be regarded as a replacement of QM for complex catalysis applications, as shown in Figure 1b.

2.1. Top-Down Approach: Learning from Experimental Data. To build a ML model for catalytic performance prediction, it is essential to first define the data descriptors for catalyst. Traditional concepts for catalyst include, for example, the particle size, surface area, valence state, surface composition and coordination, which can be determined from advanced characterization techniques, such as XRD, TEM, Brunauer—Emmett—Teller—surface area analysis, X-ray photoelectron spectroscopy, and so on. However, because of the high cost of structure characterization, it is a common practice that only the catalysts with good catalytic performance are characterized in detail, causing the lack of data for most catalysts reported in literatures. In practice, the low-cost descriptors are adopted in ML applications, typically from the catalyst synthesis conditions (e.g., feed composition and ratio, synthesis time and temperature, pH value, etc.) and the operation conditions of reaction (e.g., catalyst amount, reaction temperature, time, reactant amount, etc.).

For instance, Sasaki et al.42 showed that the yield and byproducts of NO decomposition over the Cu/ZSM-5 zeolite catalyst can be predicted by a NN model using the experimental conditions as input such as compositional...
quantity, temperature, and pressure. Similarly, Rahman et al.\textsuperscript{46} used the temperature, reaction time, substrate molar ratio, and enzyme amount as the inputs for a NN model, which can predict the yield of lipase-catalyzed dioctyladipate synthesis. These models help to optimize the synthesis conditions and reaction conditions to maximize the reaction activity. However, they are not expected to be transferable (i.e., only work for the experiment data for each research group) because the consistency of data cannot be guaranteed due to some hidden variables (e.g., stirring rate, experimental equipment).

ML-assisted literature analysis (i.e., the data mining technology) is a possible solution to overcome the data limitation and eliminate the inconsistency between data. By using the text mining and information extraction of natural language processing, Kim et al.\textsuperscript{17} established a ML model using natural language processing and decision tree to predict the critical parameters for synthesizing oxide materials from 12 000 scientific articles about metal oxide synthesis (Figure 2). For instance, they found that most hydrothermal reactions are carried out between 150 and 200 °C for 12 or 24 h. The hydrothermal condition used to synthesize both simple and complex oxides occur at the similar and only modestly high hydrothermal condition used to synthesize both simple and pentanary oxides. Figure adapted from ref 47. Copyright 2017 American Chemical Society.

It should be emphasized that the conclusions from the top-down ML approaches rely on the scope of published experimental data; it is virtually impossible to predict the performance of unknown catalysts and find unknown reactions. The emergence of ML robots\textsuperscript{52} might be the solution for collecting a large amount of experimental data without human input. There is much to explore in the future of automated experiments.

2.2. Bottom-Up Approach: Predicting from First-Principles. It is more attractive to make prediction from first-principles, i.e. without input from experiment. Such a standard ML procedure would consist of four steps: (i) building a data set from QM calculations, often DFT, which should contain various structures and their properties (e.g., energy, band gap, d-band center, etc.); (ii) identifying the features that can link the structure with the properties for prediction. These features can be basic constant of atom/elements, the geometrical coordinate or other low-cost computable properties (e.g., atomic charge); (iii) designing ML models and training ML models on the data set; (iv) predicting properties using ML models for unknown structures.

Typically, step (i) is the most time-consuming step, which requires heavy QM computation on many candidate structures. We will show later that the efficient PES sampling techniques are critical for obtaining a desirable data set. The steps of (ii) and (iii) are application-oriented where the chemical intuition, the experience, and also trial-and-error tests are typical recipes to identify the suitable features and to construct appropriate ML models. In the following, we will overview the ML literatures on heterogeneous catalysis from two aspects according to the output of ML models (i.e., adsorption energy prediction and PES prediction).

2.2.1. Adsorption Energy Prediction. Based on the BEP principle, i.e. the linear dependence of the reaction barrier on the adsorption energy, the adsorption energy is regarded as a key parameter in heterogeneous catalysis and is also a low-cost computable property. The successful prediction of adsorption energy can lead to a qualitative evaluation of catalytic activities of system. The numerical accuracy for the adsorption energy prediction should be high (e.g., below 0.1 eV) because the value itself is typically –3 to 1 eV and an error of 0.1 eV may totally change the activity.

For finding the active site of CO reduction on NiGa alloy, Ulissi et al.\textsuperscript{53} developed a NN-aided discovery approach which relates the CO adsorption energies with different NiGa bimetallic surface sites. The surface site is described by seven different Ni structures, i.e. the coordination number (CN) of Ni, CN$_{Ga}$ the average CN for Ni–Ni bonds (CN$_{Ni-Ni}$), the average CN$_{Ni-Ga}$ the average CN$_{Ga-Ni}$ the average CN$_{Ga-Ga}$ and the fraction of the Ni in the alloy. This led to the discovery of a previously unconsidered active site, single Ni atoms surrounded by surface Ga atoms, which is shown to exhibit a step-like kinetic behavior and the best thermodynamics for CO reduction.\textsuperscript{54} Jinnouchi and Asahi\textsuperscript{55,56} predicted the rates of NO decomposition as a function of composition and particle size by using a ML scheme, where the combination of Bayesian linear regression and the local similarity kernel is utilized to connect the st. Ni, and NO adsorption energies. The catalytic activities of NO decomposition on Rh$_{n}$–Au alloy nanoparticles were then estimated by Sabatier analysis based on the predicted adsorption energy and the available BEP relationship.

Apart from the catalytic activity estimation on a single catalyst, the ML-based adsorption energy prediction can be extended for the purpose of the high-throughput catalyst screening and the reaction network optimization. For catalyst screening, Tran and Ulissi\textsuperscript{57} constructed the relationship between structural fingerprints and the CO and H adsorption energies. Guided by the optimal adsorption energy of CO and H ($E_{CO} = \sim 0.67$ eV and $E_{H} = \sim 0.27$ eV) for CO$_2$RR and hydrogen evolution reaction (HER), respectively, they performed a systematic screening of alloys with 31 different elements that encompasses 50% of d-block elements and 33% of p-block elements. Finally, 131 candidate surfaces from 5 bulk alloys for CO$_2$ reduction and 258 surfaces from 102 bulk alloys for H$_2$ evolution were identified. This method successfully accelerated the design of the CuAl alloy catalyst.

---

**Figure 2.** Two-dimensional hexagonally binned normalized histograms of hydrothermal reaction and calcination times and temperatures for binary and pentanary oxides. Figure adapted from ref 47. Copyright 2017 American Chemical Society.
which provides multiple sites and surface orientations with near-optimal CO binding for both efficient and selective CO
reduction.\(^\text{58,59}\) For finding reaction mechanism, Nørskov and
co-workers\(^\text{60}\) proposed a reaction network optimization
framework using Gaussian process regression (GPR) 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), as shown in Figure
3. 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. The BEP relation was 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 (Figure 3a). Through an iterative GPR model refinement
process, where only potential rate-limiting steps were analyzed
via further DFT calculations, a probable reaction network from
syngas to acetaldehyde was finally identified (Figure 3b).

Instead of the quantitative prediction of adsorption energy,
ML using the same data set of adsorption can be utilized to
provide qualitative understanding on what controls the
adsorption strength. Toyao et al.\(^\text{61}\) developed a simple ML
model with 12 descriptors for predicting the adsorption
energies of CH\(_x\) and H species on Cu-based alloys. Among 12
descriptors, the element group, surface energy, and melting
point were found to be the key factors affecting the adsorption
energy (Figure 4a). Rappe and co-workers constructed a
regularized random forest ML model to quantify the
importance of different descriptors of the Ni,P (0001) surface
structure on determining the free energy of H adsorption. The
results showed that the Ni−Ni bond length is the most
important descriptor for HER activity: a shorter Ni−Ni bonds
give higher HER activity (Figure 4b).\(^\text{62}\) Similarly, Xin et al.\(^\text{63,64}\)
built a ML model with a large number of different descriptors
for predicting CO binding on alloys. They identified the
important role from the d-band shape and sp-band filling on
CO binding, even for coinage metals. Indeed, this information
provides the alternative to the adsorption energy for activity
prediction.

Since the adsorption energy can be related to d-band center
for metals, ML is also utilized to predict directly the d-band center
as to provide a more general model for different adsorbates. Takigawa et al.\(^\text{65,66}\) developed a ML-based model
by gradient boosting regression method to predict the d-band
centers for 11 metals (Fe, Co, Ni, Cu, Ru, Rh, Pd, Ag, Ir, Pt,
Au) and their pairwise bimetals for two different structures
(metal impurities and overlayer-covered metal surfaces). It is
shown that the d-band centers are reasonably well predicted
with average root-mean-square error (RMSE) less than 0.5 eV,
even only six descriptors and 25% data given for training. In

---

**Figure 3.** (a) Reaction network for the reaction of syngas (CO+H\(_2\)) to CO\(_2\), water, methane, methanol, acetaldehyde and ethanol, including surface-adsorbed intermediates with up to two carbons and two oxygens (C1/C2 chemistries). (b) The reduced network for syngas reactivity on Rh (1 1 1), producing acetaldehyde selectively as confirmed by the 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. Figure adapted from ref \(^\text{60}\). Copyright 2017 Nature Publishing Group.
Enhances the OER activity. The covalency strength can be constructed to predict the adsorption energy of CH as the catalyst stability, the active site structure, and the exact reaction kinetics, cannot be totally neglected, which indeed become increasingly important for evaluating the peak activity of volcano curve. The atomic simulation remains as the key tool to go beyond BEP relation in order to determine the activity under realistic conditions.

2.2.2. Atomic Simulation for Fast PES Exploration. QM and statistic mechanics are two branches of fundamental sciences that play on the battle ground of PES. The former provides the energy and latter derives thermodynamics from the energy landscape. It is of central significance to develop fast and reliable PES evaluation methods to meet the ever-increasing demand for simulating complex systems, such as those in heterogeneous catalysis. ML-based atomic simulation emerged in recent years is such a promising field, which relies on a ML potential for evaluating PES, an analytic numerical function with numerous adjustable parameters. Compared with empirical force fields, the ML potential is advantageous in studying complex material systems as well as chemical reactions as long as the training data set contains the related PES data.

The energy expression to correlate structure with energy is basic to the ML model for PES evaluation. Since energy is an extensive quantity, the analytic expression of energy can be conventionally expanded as the summation of different functional forms constituted by local geometrical variables, such as bond, angles, and dihedral angles. The many-body expansion (MBE) method is a typical approach, which, being identical to the force field methods, splits the total energy as the interaction terms of one-, two-, and three-body terms ($E_{ij}$, $E_{ijk}$ and $E_{ijk}$, respectively), as shown in eq 6.

$$E = \sum_i E_i + \sum_{ij} E_{ij} + \sum_{ijk} E_{ijk}$$

Instead of using physically meaningful functions, each term utilizes complex numerical functions, the parameters of which can be trained via ML techniques. For example, Bartok et al. used Bayesian inference techniques to correct DFT one-body and two-body energies for water, which significantly improves the description of the structure and dynamics of liquid water. Because of the sharp increase of the many-body terms (four-body and above) with the increase of system size, the MBE method is problematic for describing large size systems and also those with salient many-body bonding characteristics (metallic systems) and thus not suitable for heterogeneous catalysis.

Behler and Parrinello proposed a high-dimensional neural network (HDNN) scheme in 2007, as shown in Figure 5, that splits the total energy as the summation of individual atomic energy, as shown in eq 7.

$$E = \sum_i E_i$$

where $E_i$ for each atom is the output of a standard feed-forward NN. The input of NN is a set of structural descriptors to describe the atom bonding environment. The NN parameters can be trained using the first-principles PES data set. Compared with the MBE method, the atomic energy summation method is linear scaling with respect to system size and thus is suitable for simulating large material systems. In HDNN architecture, the functional form of NN can be replaced by other numerical representations, especially when target systems are relatively simple. For example, GPR proposed by Bartok et al. have been successfully utilized.
for a number crystalline systems, functional materials, and amorphous solids.\textsuperscript{71-78}

To describe the local atomic environment, Behler proposed a series of Gaussian-type symmetry functions as the input of NN. The most used two-body $G^2$ and three-body $G^4$ are shown in eqs 8 and 9,\textsuperscript{34,85} where $r_{ij}$ is the internuclear distance between atoms $i$ and $j$, and $\theta_{ijk}$ is the angle centered at the $i$ atom with neighbors $j$ and $k$. These structural descriptors are invariant with respect to the permutational, rotational, and translational movement of the system and thus can produce a continuous PES for deriving the first energy derivatives (i.e., the atomic force):

$$f_c(r_{ij}) = \begin{cases} 0.5 \times \tanh \left[ \frac{3}{r_{ij}} \right] & \text{for } r_{ij} \leq r_c \\ 0 & \text{for } r_{ij} > r_c \end{cases}$$

(8)

$$G^2 = \sum_{ij} e^{-t(r_{ij} - r^*)} f_c(r_{ij})$$

(9)

$$G^4 = 2^{3-t} \sum_{ijk} (1 + \lambda \cos \theta_{ijk})^t e^{-t(r_{ij}^2 + r_{jk}^2 + r_{ki}^2)} f_c(r_{ijk})$$

(10)

With the similar idea, other local geometry representations (e.g., bispectrum and smooth overlap of atomic positions)\textsuperscript{34,85} were also proposed for PES construction.

While HDNN was initially developed for material systems, it has been utilized for catalysis-related problems. Bose and Kitchin\textsuperscript{86} investigated the oxygen interaction with Pd (111) by constructing a PdO HDNN potential. The grand-canonical MC simulations on a (10 $\times$ 10) Pd slabs at 600 K was performed, which found that the mean oxygen coverage is around 0.50 monolayer. The oxygen diffusion barrier from HDNN potential is 0.40 eV, which are in agreement with experimental results of 0.4$\sim$0.5 eV.\textsuperscript{77} Artrith et al.\textsuperscript{78} utilized HDNN potential for optimizing the structure of Cu/ZnO. The canonical (NVT) MD simulation at 1000 K was carried out for a large Cu cluster (612 atoms) supported on ZnO (10$\bar{1}$0) surface, a system of 7524 atoms. The yielded Cu cluster is highly distorted, particularly at the interface, which provides a structural model for understanding the true catalyst morphology under reaction conditions.

Undoubtedly, a major problem for ML-based atomic simulation lies at the high cost of producing the large PES data set required for training ML potential, which must involve extensive QM calculations for a large number of structural configurations. Ideally, the PES data set for ML should be representative and compact but needs to be as global as possible to cover both minima and transition regions on PES. In 2017, our group proposed to utilize the SSW\textsuperscript{79-80} global optimization trajectories for constructing the representative PES data set, and the combination with NN technique (SSW-NN) for atomic simulation turns out to be successful for simulating a wide range of materials and reactions and also for guiding experiment.\textsuperscript{81,82} The SSW-NN method is now implemented in LASP software (Large-scale Atomic Simulation with neural network Potential, accessible from www.lasphub.com), which incorporates the PES data generation, NN potential training, and the atomic simulation in one platform. By using LASP, a large set of global NN (G-NN) potentials is now established via the automated global-to-global learning procedure. The G-NN potentials cover a wide range of elements across the periodic table. In the following sections, we will introduce the SSW-NN method and its three applications to heterogeneous catalysis.

3. SSW-NN FOR HETEROGENEOUS CATALYSIS

3.1. SSW-NN Methodology. In principle, the global PES sampling methods would be an outstanding choice for PES data generation. Simulated annealing has been the most popular method, which explores PES via repeated heating and cooling cycles.\textsuperscript{83} The enhanced MD method such as metadynamics\textsuperscript{84,85} can be the valuable supplement to add the reaction data. Nevertheless, MD sampling for data generation still suffers from the "short-sighted" problem because of the exponentially low probability to overcome high reaction barrier at low temperatures and the preference of trapping at high-entropy structure regions at high temperatures. As a result, the PES data thus generated are often overwhelmingly redundant, being highly localized to a few initial phases. This will inevitably lead to the inadequacy of thus-obtained ML potentials for predicting unknown materials and reactions. However, the other global optimization methods, such as basin-hopping,\textsuperscript{86} evolutionary algorithm,\textsuperscript{87} genetic algorithm,\textsuperscript{88,89} and the particle swarm optimization method (CALYPSO),\textsuperscript{90} have also been tested for on-the-fly ML model training of materials in recent years. These methods transform the PES by overlooking the transition region between minima to realize fast global minimum (GM) search. Generally, they sample the PES data from the structural relaxation trajectories and thus may well miss key reaction channels.

Figure 5. Schematic structure of HDNN architecture for a ternary system. Figure adapted from ref 72. Copyright 2017 Wiley Publishing Group.
Our group developed SSW method in 2013 for global optimization of aperiodic systems, which utilized a more sophisticated power-type structure descriptors (G-NN potential and performing the global optimization method is an iterative global-to-global approach for generating the new minimum. At the end of the SSW step, a structure selection module (e.g., direction is re-chosen a random direction to perturb the structure after the method, each SSW step (from one minimum to another) will utilize the energy configuration moving smoothly from a local minimum to a high-energy configuration along one random mode direction. The climbing mechanism of SSW method learns from the TS location method, the bias-potential driven constrained-Broyden-dimer (BP-CBD) method, that utilize bias potentials to overcome the barrier between minima. In one particular SSW step, a modified PES, as shown in Figure 6a, is utilized for moving from the current minimum, $R^n_0$, to a high energy configuration, $R^n_i$ (the climbing), in which a series of bias Gaussian potential is added one by one consecutively along the direction. Once the $R^n_i$ is reached, all bias potential are removed and the local optimization is performed to quench the structure to a new minimum. Different from the BP-CBD method, each SSW step (from one minimum to another) will choose a random direction to perturb the structure after the direction is refined (softened) using the biased-CBD method. At the end of the SSW step, a structure selection module (e.g., in Metropolis Monte Carlo scheme) is applied to accept/refuse the new minimum.

Inherited from SSW global optimization, the SSW-NN method is an iterative global-to-global approach for generating G-NN potential and performing the global optimization (PTSD), which is developed to be compatible with the SSW global optimization data set. In PTSD, not only are the traditional two-body and three-body terms included but also the four-body terms are added and the spherical functions are introduced to enhance the structure discrimination. For example, eqs 11–13 show the two-body terms utilized in PTSD, where $R^n$ is the power function to describe the radial distribution and the $Y_{lm}$ is the spherical function to take into account the angular distribution of the chemical environment.

$$R^n(r) = r^n f_j(r) \quad (11)$$

$$S_i^1 = \sum_{j \neq i} R^n(r_{ij}) \quad (12)$$

$$S_i^2 = \left[ \sum_{m=-L}^{L} \left| \sum_{j \neq i} R^n(r_{ij})Y_{lm}(r_{ij}) \right|^{2/3} \right]^{1/2} \quad (13)$$

In the beginning of SSW-NN self-learning procedure (Figure 6b), the NN potential is initially obtained by learning a small data set (typically less than 1000 structures) collected from short-time SSW sampling based on DFT, which are often restricted to small systems (below 20 atoms) of known configurations. The data set needs to be calculated by DFT with a high-accuracy setup. Next, the SSW global optimization based on NN potential will be carried out extensively, starting from a variety of initial structures, mainly randomly constructed, with different morphology, including bulk, surface and clusters, different chemical compositions, and different number of atoms per cell. After each iteration of global optimization, a small data set with diverse structures on PES is screened out by selecting either randomly or from those exhibiting new atomic environment (e.g., out-of-bounds in structural descriptors, unrealistic energy/force/curvature). These additional data will be calculated by DFT with the same high-accuracy setup and then added to the training data set for a new iteration of NN potential update. Typically, more than ~100 iterations are required to finally obtain a transferable G-NN potential. The accuracy for G-NN potential is typically 5~10 meV/atom for RMSE of energy and 0.1~0.2 eV/Å for RMSE of force.

### 3.2. LASP Software

LasP software project initiated in the early of 2018. The initial objective is to merge two major simulation tools, SSW global optimization and G-NN calculations, developed by our group into a single package for the better and simpler usage. LasP is now shaped toward a software platform for many purposes, not only for atomistic simulation but also for PES data building and exchange and even for G-NN potential generation. A large set of powerful simulation techniques has been assembled into LASP program (the current version 2.2) to simplify the usage and to enlarge the scope of the current atomistic simulation. This can be seen in Figure 7, which overviews the current modules in LASP. Apart from the G-NN computation for energy and force evaluation, LASP implements standard data-exchange interfaces to connect with common PES evaluation packages, which allows for the PES data generation using QM calculation. For PES exploration at the finite temperature and pressure conditions, LASP also provides the standard and restrained MD functionalities, in addition to the structure local/global geometry optimization and TS search. All these functionalities can be switched on and off in LASP with short keywords, and
thus a variety of different tasks can be fulfilled facilely within one platform, such as SSW-NP PES global exploration, PES data generation using QM calculations, G-NN potential creation, and MD simulation based on G-NN potential.

### 3.3. SSW-NN Applications

#### 3.3.1. Case 1: Y-Stabilized ZrO$_2$ Structural Evolution

**Y**-stabilized ZrO$_2$ (YSZ) is one of the most important ion conducting solids for solid oxide fuel cells and solid oxide electrolysis cells. The composition and temperature-dependence of the oxygen anion transportation have been the focus in a large volume of literatures. The observations from experiments and simulation, we have computed the anion diode from empirical potentials, where the computed oxygen anion transport have been the focus in a large volume of literatures. The composition and temperature-dependence of the oxygen anion transportation have been the focus in a large volume of literatures. The observations from experiments are, however, at odds with the computational results with respect to the cubic-ZrO$_2$ and cubic-Y$_2$O$_3$. The composition and temperature-dependence of the oxygen anion transport have been the focus in a large volume of literatures. The observations from experiments are, however, at odds with the computational results.

#### 3.3.2. Case 2: Syngas Conversion on ZnCrO Catalysts

Zinc–chromium oxide (ZnCrO) catalyst is the first generation industry catalysts for syngas-to-methanol. Experiments show that the catalytic activity is significantly affected by Zn:Cr ratios. The best activity and selectivity is achieved at Zn:Cr = 1:1, while the Zn:Cr = 2:1 catalysts yield rather poor activity. The atomic structure of ZnCrO catalysts with the Zn:Cr = 1:1 remains however uncertain. To resolve where and how syngas conversion occurs on ZnCrO catalyst, SSW-NN simulations were performed to obtain the Zn$_x$Cr$_{1-x}$O global PES based on the ZnCrO G-NN potential, which was obtained by exploring different compositions of Zn–Cr–O three element systems. The final G-NN has three-hidden layers (324-80-60-61 net), equivalent to 103,743 network parameters in total. The final RMSE of energy and force are 4.3 meV/atom and 0.128 eV/Å respectively.

By using the GMS of different ZnCrO compositions, the thermodynamics phase diagram of Zn–Cr–O is thus constructed, as shown in Figure 9a. It reveals a small stable

---

**Figure 7.** Architecture of LASP code and the modules implemented.

**Figure 8.** (a) Thermodynamic convex hull diagram for Y$_2$Zr$_{1-2x}$O$_{2x}$ with respect to the cubic-ZrO$_2$ and cubic-Y$_2$O$_3$. (b) Oxygen diffusion coefficient D vs the concentration of Y$_2$O$_3$ at different temperatures. Figure adapted from refs 98 (a) and 99 (b). Copyright 2020 American Institute of Physics and Copyright 2020 American Chemical Society. 

---

13221

https://dx.doi.org/10.1021/acscatal.0c03472
ACS Catal. 2020, 10, 13213–13226
syngas conversion reaction on the most stable Zn and thus DFT calculations were applied to investigate the G-NN potential for catalytic reactions are considerably huge describe the syngas reaction. The e due to the lack of C and H elements, it cannot be applied to surface structure determination under reaction conditions and, selectivity profoundly.

The microkinetics simulation further con where the rate-determining step is the hydrogenation of CH sites on the amorphous TiO phase diagram of TiO composition island (i.e., Zn:Cr:O = 6:6:16~3:8:16), where the oxide alloy crystallizes into a spinel phase. At Zn:Cr = 1:1, a Zn,Cr,0, metastable crystal phase is present, also with the spinel crystal structure, but contains the highest concentration of unusual [ZnO] octahedra (O₈) in bulk compared to the other spinel crystals. This subtle structural difference turns out to be critical to affect the syngas conversion activity and selectivity profoundly.

The ZnCrO G-NN potential was utilized for bulk and surface structure determination under reaction conditions and, due to the lack of C and H elements, it cannot be applied to describe the syngas reaction. The efforts to build five-element G-NN potential for catalytic reactions are considerably huge and thus DFT calculations were applied to investigate the syngas conversion reaction on the most stable Zn₆Cr₈O₂₈ and Zn₆Cr₈O₂₈ surfaces determined from G-NN. It was found that the Zn₆Cr₈O₂₈ surface can generate the four-coordinated planar Cr²⁺ cation site on surface that is critical for methanol activity and selectivity. As shown in Figure 9b, the reaction follows CO → CHO → CH₂O → CH₂OH → CH₃OH mechanism, where the rate-determining step is the hydrogenation of CH₂O step, which also determines the selectivity of CH₂OH or CH₃OH. The microkinetics simulation further confirms the sharp difference in activity and selectivity observed in experiment. The origin of unusual[ZnO₈] octahedral (O₈) in bulk as compared to the other spinel crystals. This subtle structural difference turns out to be critical to affect the syngas conversion activity and selectivity profoundly.

3.3.3. Case 3: Origin of Amorphous TiO₂Hₓ for HER. Black titania (TiO₂) exhibit a core–shell structure with the amorphous shell coated on the anatase crystals has been found to exhibit the superior HER activity. The amorphous shell structure which is responsible for the enhanced HER activity is unclear. To resolve the HER active sites on the amorphous TiO₂Hₓ shell, a Ti–O–H three elements NN potential was constructed to describe the PES of TiO₂Hₓ system based on large TiO₂Hₓ global data set with 143 786 structures. The network involves two-hidden layers, each with 50 neurons, equivalent to 38 103 network parameters in total. The final RMSE of energy and force are 9.8 meV/atom and 0.22 eV/Å respectively. This TiOH NN potential can replace theoretically the DFT calculations not only in structural identification of TiO₂Hₓ catalyst but also in catalysis reaction involving water, O₂, H₂, and H₂O.

By performing SSW-NN simulation, the thermodynamics phase diagram of TiO₂ bulk and surface in contact with H₂ at different temperature and pressures can thus be determined quantitatively. Among common anatase surfaces, only the ridged anatase (112) surface was found to reconstruct significantly by surface H atom and a local high H coverage, 0.69 ML, can gradually built up during the surface amorphization with the exposure of 25% TiOH, 50% TiOH, and 25% TiOH, atoms on surface (Figure 10a). The Ti–O bond length has a wide distribution, from 1.8 to 2.2 Å, as compared with 1.9~2.1 Å on perfect TiO₂ surface. This high H coverage not only renders the black color of the amorphous TiO₂ but also provides a low energy reaction channel for HER: a
transient Ti−H hydride becomes likely to form on the exposed Ti atoms of amorphous surface. The nascent TiH hydride can react facilely with the neighboring OH to produce H₂, where the barrier is more than 1 eV lower than the traditional H coupling channel via two surface OH groups (barrier >1.9 eV), as shown in Figure 10b.\textsuperscript{115,116} The TiH/OH coupling pathway for hydrogen production is ~24 orders of magnitude faster than the direct OH/OH coupling, indicating that the TiH/OH mechanism is the only likely reaction for the HER on amorphous TiO₂ material. The experiment also revealed the presence of TiH hydride on amorphous TiO₂, where the characteristic H chemical shift δ value in ²H nuclear magnetic resonance spectroscopy has a negative peak (−0.6 ppm) and this peak of δ grows with the increase of hydrogenation time.\textsuperscript{117}

4. CONCLUDING REMARKS

The ever-increasing ease to access big data has reshaped our society and also the way of chemical research. In the past decade, a variety of ML techniques have been developed, either based on experimental data or using first-principles results, which demonstrate their great power to solve outstanding challenges in heterogeneous catalysis that are otherwise difficult to treat with traditional methods. While here we retrospect briefly the history of theoretical studies on surface reactions and introduce the main branches in ML methods and applications for catalysis, our main focus is put on the ML-based atomic simulation, which has the great potential to replace DFT for large-scale atomic simulation. As a representative in this field, SSW-NN method implemented in LASP software is elaborated, highlighting the key features of algorithm and the main functionalities. Three cases studies of SSW-NN (i.e., YZrO, ZnCrO, and TiOH systems) for material and reaction simulation are presented to demonstrate the ability of ML for large-scale simulation and the finding of new structures and new reactions.

Apart from the successful applications, the current ML methods do exhibit many obvious shortcomings that need to improve in the future. For example, the limited availability and possible inconsistency are generally present for experimental data, which is a key obstacle for ML to guide experiments; the intrinsic approximation in BEP relation casts doubts on ML-based catalyst and activity prediction using simple descriptors (e.g., the adsorption energy). ML-based atomic simulation using G-NN potentials, although has reached to the stage for wide applications, still suffers from a number of scientific and technical difficulties, as listed in the following.

(i) The high-cost to build multielement G-NN potentials for materials and reactions. For example, for CuZnAl catalyst for methanol synthesis, it would require a six-element potential, CuZnAlCH₂O, to explore the phase space of catalyst structures and reactions. The construction of such a G-NN potential is, however, extremely challenging due to the cost to generate the representative PES data;

(ii) The lack of electronic structure information in the current ML potentials. The electronic structure information, such as atomic charge and density of states, if available with a low-cost, would facilitate greatly to interpret different electronic spectra in experiment data and to understand better how catalysts evolve during reaction;

(iii) The questionable accuracy for simulating long-range disordering of complex systems, such as glassy materials. There are general concerns on both the data set generation using small unit cells and the use of local atomic descriptors for fitting the energy of long-range disordered structures and the interaction therein.

(iv) the difficulty to go beyond DFT accuracy. For the great difficulty in generating global data set, DFT calculation remains to be the practical choice for constructing ML potentials. This effectively restricts the accuracy of ML potentials to the level of DFT.

(v) Overall, there are still ample rooms to design new ML models and invent new ML methods to better cope with the above issues. Luckily, heterogeneous catalysis is complex enough to drive the development of these new technologies for exciting future.

AUTHOR INFORMATION

Zhi-Pan Liu — Collaborative Innovation Center of Chemistry for Energy Material, Shanghai Key Laboratory of Molecular Catalysis and Innovative Materials, Key Laboratory of Computational Physical Science, Department of Chemistry, Fudan University, Shanghai 200433, China; orcid.org/0000-0002-2906-5217; Email: zpliu@fudan.edu.cn

Sicong Ma — Collaborative Innovation Center of Chemistry for Energy Material, Shanghai Key Laboratory of Molecular Catalysis and Innovative Materials, Key Laboratory of Computational Physical Science, Department of Chemistry, Fudan University, Shanghai 200433, China; orcid.org/0000-0001-5894-5910

Complete contact information is available at: https://pubs.acs.org/10.1021/acscatal.0c03472

Notes
The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This work was supported by the National Key Research and Development Program of China (2018YFA0208600), National Science Foundation of China (91945301, 91745201 and 21533001), China Post-Doctoral Science Foundation (2019M661340).

REFERENCES


Site Motifs for CO