ACCOUNTS of chemical research

pubs.acs.org/accounts Article

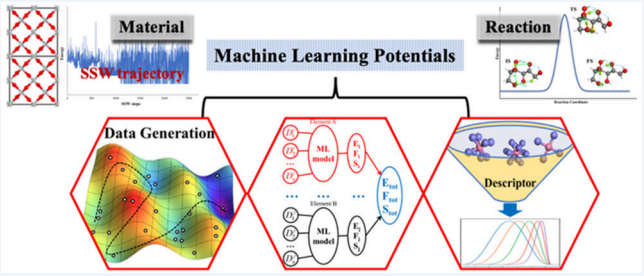
Large-Scale Atomic Simulation via Machine Learning Potentials Constructed by Global Potential Energy Surface Exploration

Published as part of the Accounts of Chemical Research special issue "Data Science Meets Chemistry".

Pei-Lin Kang, Cheng Shang,* and Zhi-Pan Liu*

Cite This: *Acc. Chem. Res.* 2020, 53, 2119–2129 Read Online

CONSPECTUS: Atomic simulations based on quantum mechanics (QM) calculations have entered into the tool box of chemists over the past few decades, facilitating an understanding of a wide range of chemistry problems, from structure characterization to reactivity determination. Due to the poor scaling and high computational cost intrinsic to QM calculations, one has to either sacrifice accuracy or time when performing large-scale atomic simulations. The battle to find a better compromise between accuracy and speed has been central to the development of new theoretical methods.



The recent advances of machine-learning (ML)-based large-scale atomic simulations has shown great promise to the benefit of many branches of chemistry. Instead of solving the Schrödinger equation directly, ML-based simulations rely on a large data set of accurate potential energy surfaces (PESs) and complex numerical models to predict the total energy. These simulations feature both a high speed and a high accuracy for computing large systems. Due to the lack of a physical foundation in numerical models, ML models are often frustrated in their predictivity and robustness, which are key to applications. Focusing on these concerns, here we overview the recent advances in ML methodologies for atomic simulations on three key aspects. Namely, the generation of a representative data set, the extensity of ML models, and the continuity of data representation. While global optimization methods are the natural choice for building a representative data set, the stochastic surface walking method is shown to provide the desired PES sampling for both minima and transition regions on the PES. The current ML models generally utilize local geometrical descriptors as an input and consider the total energy as the sum of atomic energies. There are many flavors of data descriptors and ML models, but the applications for material and reaction predictions are still limited, not least because of the difficulty to train the associated vast global data sets. We show that our recently designed power-type structure descriptors together with a feed-forward neural network (NN) model are compatible with highly complex global PES data, which has led to a large family of global NN (G-NN) potentials. Two recent applications of G-NN potentials in material and reaction simulations are selected to illustrate how ML-based atomic simulations can help the discovery of new materials and reactions.

KEY REFERENCES

- Huang, S.-D.; Shang, C.; Zhang, X.-J.; Liu, Z.-P. Material discovery by combining stochastic surface walking global optimization with a neural network. *Chem. Sci.* **2017**, *8*, 6327−6337.[1] *A "global-to-global" approach is proposed to systematically construct the machine learning global potential for atomic simulations, which combines the stochastic surface walking (SSW) method and neural network (NN) techniques.*
- Ma, S.; Huang, S.-D.; Liu, Z.-P. Dynamic coordination of cations and catalytic selectivity on zinc−chromium oxide alloys during syngas conversion. *Nature Catal.* **2019**, *2*, 671−677.[2] *The SSW-NN method is applied to construct the thermodynamics phase diagram of the Zn−Cr−O ternary system, which leads finally to the clarification of the active site in a ZnCrO catalyst for the syngas conversion.*
- Kang, P.-L.; Shang, C.; Liu, Z.-P. Glucose to 5-Hydroxymethylfurfural: Origin of Site-Selectivity Resolved by Machine Learning Based Reaction Sampling. *J. Am. Chem. Soc.* **2019**, *141*, 20525−20536.[3] *The SSW-NN method is applied to explore the reaction network of glucose pyrolysis, which identifies the lowest energy pathway and the critical role of the retro-Michael-addition reaction in the site-selectivity.*

Received: July 22, 2020
Published: September 17, 2020

https://dx.doi.org/10.1021/acs.accounts.0c00472
Acc. Chem. Res. 2020, 53, 2119–2129

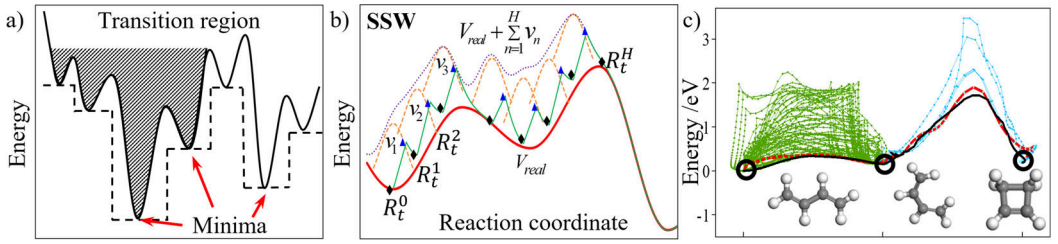


Figure 1. (a) A schematic of the 1D PES. The dotted lines indicate the transformed PES, as treated by basin hopping and genetic algorithm (Lamarckian-type) methods; the crossed area indicates the filling of the energy wells by added potentials, as utilized in metadynamics. (b) An illustration of the SSW method in a 1D PES (also see eq 1). The red, orange, purple, and green curves represent the real PES (V_{real}), the Gaussian functions (v_n, $n = 1, 2, ...H$), the modified PES, and the searching trajectory, respectively. (c) SSW trajectories for the PES global exploration of C_4H_6 molecules. The lowest energy trajectory is highlighted by the red color. For comparison, the lowest energy pathway is shown as a black curve that is extrapolated using the intrinsic reaction coordinate. Reproduced with permission from ref 50. Copyright 2013 American Chemical Society.

1. CHALLENGES TO ATOMIC SIMULATION IN CHEMISTRY

Material and reaction simulations are at the frontline of chemistry research. These simulations typically deal with slow chemical bond evolutions in complex environments, such as on surfaces, interfaces, and heterojunctions of materials. Unsurprisingly, they often resort to quantum mechanics (QM) calculations,[4−7] i.e., by solving the Schrödinger equation, to obtain accurate energetics and other thermodynamics quantities. Even with the power of modern supercomputing facilities, these QM calculations are still severely limited by the spatial and temporal scale of the target system, typically within hundreds of atoms and a few picoseconds.[8] Considering the time-scale of chemical syntheses (e.g., from seconds to hours) and the typical microstructure of synthesized materials (e.g., from nano to micrometers), QM calculations are apparently doomed in the battle toward "simulating experiments on a chip".

With the recent breakthroughs in deep learning methods and the advancement of graphics-processing-unit (GPU) computing, the 21st century has witnessed a boom in artificial intelligence applications.[9] The marriage between artificial intelligence and simulation techniques has brought new hopes for large-scale atomic simulation. Historically, the introduction of advanced machine learning (ML) techniques into atomic simulations can date back to 1990s for the potential energy surface (PES) construction of small systems (e.g., molecules interacting with a frozen surface, considering several degrees of freedom[10,11]). The idea is to utilize a ML model, such as neural networks (NNs), to fit the QM PES data set, which produces a numerical function, i.e., the ML potential, for fast PES evaluation. Equipped with highly complex numerical functions, ML potentials can reach the required level of accuracy to properly describe chemical reactions beyond those of traditional force field methods. The progress toward large-scale atomic simulations for materials and chemical reactions in general has been unfortunately slow, not least because of the high cost in collecting QM PES data set. For a long period, the methodology development has been center stage, focusing on inventing new PES sampling methods and designing new ML models to train large data sets.[1,12−30]

As an alternative to QM calculations, any qualified ML models for atomic simulations require, to warrant their usage, high computational speeds using low-scaling calculations, with an accuracy comparable to QM calculations. While this can indeed be achieved with the increasing complexity of numerical functions, it was soon realized that the predictivity and the transferability of ML models are equally, if not more, important in applications. Since there are innumerable materials and reactions, and as a training data set will never be complete, one has to set up some basic guidelines for ML models that can maximally enhance the predictivity and the transferability. Three aspects, (1) the representativity of the data set, (2) the extensity of ML models, and (3) the continuity of data representation, must rank top in the list of guidelines.

1.1. Representativity

Common to all ML applications, the quality of a data set used for training is always of most concern, which determines largely the predictivity of ML models. The data set should ideally cover as many as possible different structures, which acts as a representative sample to all likely structures on a global PES.

1.2. Extensity

As energy is an extensive quantity and is the output of ML models, it is essential that ML models are able to describe different sizes of systems, from atoms to bulky materials, on an equal footing.

1.3. Continuity

To obtain the first derivative of energy, i.e., force, ML models should be able to produce a continuous derivative for energy with both a high numerical accuracy and a high efficiency.

This Account serves to overview the important methodology progress for ML-based atomic simulations on the aforementioned three aspects. We will highlight in particular the stochastic surface walking global optimization with the neural network potential (SSW-NN) method[1,24] developed by our group that satisfies all basic guidelines and already proves its power in large-scale atomic simulations. The SSW-NN method is implemented in LASP[30] software (Large-scale Atomic Simulation with Neural Network Potential; accessible from www.lasphub.com), which implements the PES data generation, NN potential training, and atomic simulation in one platform. On the basis of LASP, a large set of global NN (G-NN) potentials is now established via the automated global-to-global approach,[1] which supports a wide range of elements across the periodic table.

2. GLOBAL DATA GENERATION

Unlike QM calculations, where the high-quality PES is solved from the Schrödinger equation, ML-based atomic simulation cannot be performed until a high-quality PES data set for training ML potential is available. This very first step turns out to be the most computationally extensive, which has largely limited the wide applications of ML-based atomic simulation for years. The key solution for this is to design efficient PES sampling

https://dx.doi.org/10.1021/acs.accounts.0c00472
Acc. Chem. Res. 2020, 53, 2119−2129

methods for data set generation. The data set should be representative and compact.

Naturally, a simple way to obtain a representative data set is to exploit the existing database. For organic molecules, for example, GDB-X[31] and QM-X[32,33] (X = the number of non-H atoms) are well-designed chemical data sets, containing a vast chemical space with virtually all possible molecules satisfying the chemical bonding octet rule. Smith et al. utilized the normal mode sampling (NMS) method to displace the database molecules and obtain a huge data set of 17.2 million structures.[22] These structures are, however, limited to the PES regions near minima, as indicated in Figure 1a, which cannot be utilized for reaction exploration.

In principle, the global PES sampling methods, as schematically illustrated in Figure 1a, would be the best choice for PES data generation. Among various methods applied to date, the most popular one remains to be molecular dynamics (MD)-based approaches, as represented by simulated annealing.[12,14,26,27] Simulated annealing explores PES via repeated heating and cooling cycles and have been utilized in many publications.[35−37] The enhanced MD method such as metadynamics[38,39] and iMD-VR[40] can be the valuable supplement to add the reaction data. The active learning approach developed by Smith et al. was also utilized to select distinct data from MD trajectories.[41,42] Nevertheless, MD sampling for data generation suffers from the "short-sighted" problem due to the exponentially low probability to overcome high reaction barriers 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 input phases (also discussed below). This will inevitably lead to the inadequacy of thus-obtained ML potentials for predicting unknown materials and reactions. On the other hand, other global optimization methods, such as basin-hopping,[43] the evolutionary algorithm (EA),[28] the genetic algorithm (GA),[44] and the particle swarm optimization method (PSO),[29] 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 a fast global minimum (GM) search. They generally obtain the PES data from the structural relaxation trajectories and thus may well miss key reaction channels at the transition regions (see Figure 1a).

In 2017, our group proposed to utilize the SSW global optimization trajectories for generating a PES data set, which turns out to be successful for a wide range of materials and reactions.[1−3,24,45−49] The SSW method was initially developed for global optimization and pathway searching of aperiodic systems, such as molecules and clusters, and was then extended to periodic crystals.[50−52] Compared to other global optimization methods with aggressive structure perturbation, the SSW method visits PES with a small step-size by exploiting the second-derivative (vibrational mode) information. SSW is able to sample the structural patterns at the transition region, and this allows finding unknown chemical reactions.

$$V_{\mathrm{mod}} = V_{\mathrm{real}} + \sum_{n=1}^{H} v_n = V_{\mathrm{real}} + \sum_{n=1}^{H} w_n \times \exp\left[-\frac{((\mathbf{R_t} - \mathbf{R_t}'')\cdot\mathbf{N_t}'')^2}{2 \times ds^2}\right] \quad (1)$$

$$P = \begin{cases} \exp\left[\dfrac{E_{\mathrm{new}} - E_{\mathrm{old}}}{RT}\right], & \text{when } E_{\mathrm{new}} > E_{\mathrm{old}} \\ 1, & \text{otherwise} \end{cases} \quad (2)$$

The SSW method combines bias-potential-driven dynamics[53] and the Metropolis Monte Carlo (MC) method.[54] The former is a standard technique utilized to overcome high barriers between minima on a PES, as shown in eq 1, where the consecutive bias potentials (Gaussian function, v_n, n = 1 to H) are added to modify the PES (V_{mod}), moving a structural configuration from minima to a high energy position on the PES (see Figure 1b), and the latter is a common method in PES sampling to select states according to the Boltzmann distribution, as shown in eq 2. The SSW method adopts a random mode generation and a constrained softening technique to obtain an optimal mode ($\mathbf{N_t}''$ in eq 1), along which the bias-potentials are added. Figure 1c shows the SSW trajectories for C_4H_6 molecules,[50] where SSW can overcome the high barriers on PES and identifies both minima and the low energy pathways between them.

The differences between MD and SSW in PES sampling can be clearly illustrated by Figure 2, where we performed both the MD simulation and a variable-cell SSW search on a TiO_2 crystal containing 48 atoms. The MD simulation is carried out at 1500 K and 1 atm, with an isothermal−isobaric (NPT) ensemble for 0.5 ns. Using the SSW and MD trajectories with similar total energy/force evaluation steps (5×10^5), we randomly collected 5000 configurations for each to view the 2D PES contour plots, as shown in Figure 2a and b, where the energy is plotted against the distance-weighted Steinhart order parameter (OP_2).[34,55] We can see immediately that SSW visits a large area of the PES, covering tens of different crystal phases, while MD only explores the local PES around the initial structure, i.e., the rutile phase,

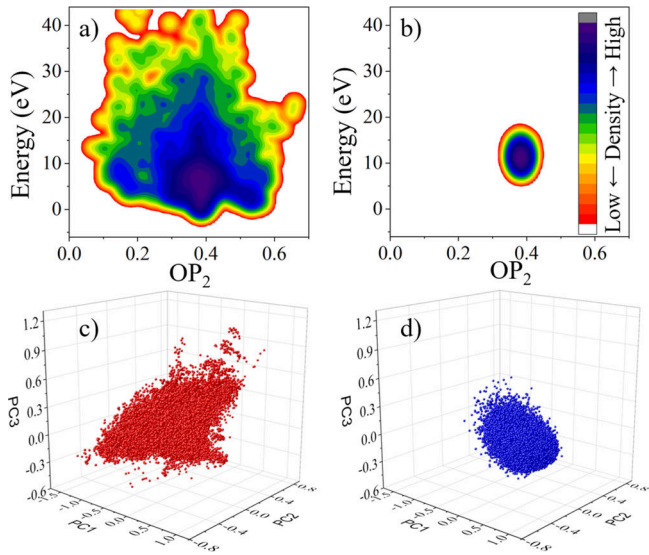


Figure 2. Comparison of data generated by SSW (a and c) and MD (b and d). Panels a and b show the PES contour plot of SSW and MD trajectories for the TiO_2 system. The x-axis is a distance-weighted Steinhart order parameter (OP_2),[34] and the y-axis is the total energy of the system corresponding to the GM. The color indicates the density of states of structures. Panels c and d show the three most distinct features of the Ti atoms in SSW and MD trajectories from the principle component analysis (PCA) based on the structure descriptors of the TiO G-NN potential.[30]

https://dx.doi.org/10.1021/acs.accounts.0c00472
Acc. Chem. Res. 2020, 53, 2119−2129

even at 1500 K (TiO_2 melts above 2000 K).[56] In addition, we also analyzed all the data by using principle component analysis (PCA), which is a technique to classify data sets according to their features. On the basis of a set of atomic structure descriptors of the Ti element (described in detail section 3 below) used in the TiO G-NN potential (accessible from the LASP Web site), PCA identifies the three most distinguishable feature components, namely PC1 to PC3. It can be found that the data from MD trajectories are limited in the atomic environment, as reflected by the small space. For comparison, the SSW data are more diverse, which covers the majority of the MD data and extends to the atomic environment belonging to many other solid phases.

Provided with the global data set, the thus-generated ML potential (also see section 3) can have both good predictivity and transferability. In Figure 3a, we have compared the energies of three TiO_2 minima (48 atoms) using DFT and the TiO G-NN potential.[1] Our G-NN potential is trained from SSW global data sets containing no more than 36 atoms, and thus, the 48-atom structures can be a good test for the extensity and transferability of the potential. The three minima are rutile, and the two other high energy configurations have P1 symmetry that cannot be reduced to any smaller unit cell. For the rutile phase, the energy difference between DFT and G-NN methods is only 0.46 meV/atom. And the difference increases to 1.9 and 4.2 meV/atom for structure **B** (a defective anatase phase) and **C** (a porous structure), respectively. It suggests that the global data set with only systems below 36 atoms does already contain the key information required to compute the energy of 48-atom systems. Figure 3b shows the typical computational time for NN and DFT benchmarked on the TiO_2 system, where NN as a linearly scaling approach can be significantly faster than DFT in large-scale atomic simulations.

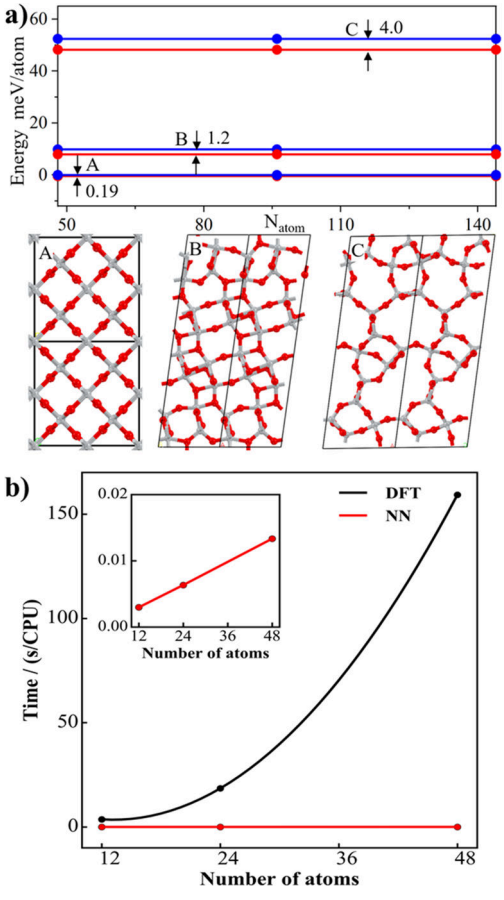


Figure 3. (a) The energy versus the system size plot in the TiO_2 system. Three structures are named as **A** (rutile), **B**, and **C** (low symmetry structures). The energy difference (meV/atom) between DFT (red) and G-NN (blue) is indicated, which is invariant with the change of the system size. (b) Computational time (s/CPU) of a single point calculation for different size TiO_2 systems (12, 24, and 48 atoms in a unit cell) by DFT (black) and NN (red) methods. Panel b is reproduced with permission from ref 1. Copyright 2017 Royal Society of Chemistry.

3. MACHINE LEARNING MODELS AND DATA DESCRIPTOR

Once a data set is established, one can exploit the PES quantities of the data set, including the structure coordinates, the total energy, and the atomic forces, to train a ML model. The procedure belongs to a standard application of supervised machine learning, and thus, there are many off-the-shelf mathematical functions and numerical fitting methods that can be selected for immediate usage. However, in order to produce a good ML potential, two basic rules must be obeyed, which guide the recent design of ML models and the selection of their input, namely, the data descriptors.

Rule 1 involves the invariance to the system size. ML models need to be flexible to describe systems with different sizes accurately, from atoms to solids. The supercelling of the structure (the increase of atom numbers) should not change the energy per atom.

Rule 2 involves continuity in data descriptors. For yielding atomic forces with high numerical accuracy, the input of the ML model should be continuous and derivable with respect to the atomic coordinate.

Restricted by these two rules, ML models adopt the local properties as inputs that are invariant to the system size; their data descriptors preserve the translational, rotational, and permutational invariance of system in order to yield a continuous and derivable PES.

In fact, the empirical force field methods[57] developed since the 1960s that decompose total energy into two-body (bond), three-body (angle), and four-body (dihedral angle) terms obey the above two rules. However, these internal coordinates, e.g., bonds and angles, are often too short-ranged to describe solid materials, and the function forms are also not complex enough to simulate chemical reactions. In 2007, Behler and Parrinello proposed a ML model,[12] named as a high-dimensional neural network (HDNN), where the total energy of system is expanded as the sum of individual atomic energies, as shown in eq 3. The atomic energy can be trained with a feed-forward neural network (NN) by linking the local chemical environment as represented by a set of structure descriptors $\mathbf{D}_i$ with the total energy, as shown in eq 4. Each $\mathbf{D}_i$ is an atom-centered symmetry function (ACSF), which are constructed by summing a series of two-body radial and/or three-body angular functions.

$$E = \sum_{i=1}^{N_{atoms}} E_i \tag{3}$$

$$E_i = f(D_i) \tag{4}$$

Since then, many ML models were proposed and, almost without exception, utilize the basic principles of HDNN, i.e., eqs 3 and 4 (also see Figure 4). The local chemical environment, by

https://dx.doi.org/10.1021/acs.accounts.0c00472
Acc. Chem. Res. 2020, 53, 2119−2129

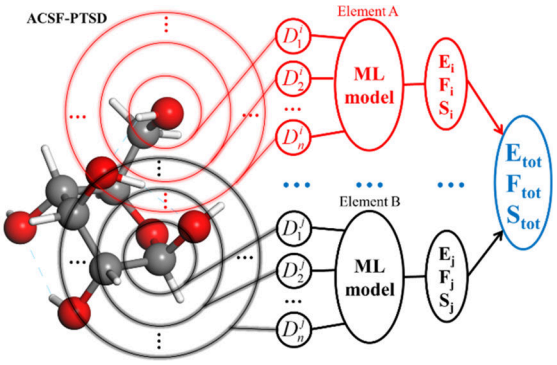


Figure 4. Illustration of ML models. Left: The atomic vector $\mathbf{D}_i$ is constructed from a series of element-discriminated functions based on coordinates (see eqs 5−12 for power-type structural descriptor (PTSD) functions[24]). Right: ML models that decompose the system properties, including energy, force, and stress (E_{tot}, F_{tot}, and S_{tot}) as the sum of atomic contributions (E_i, F_i, and S_i, where i and j are the atom indices), where the atomic vector $\mathbf{D}_i$ is utilized as the input for the models.

contrast, has many alternative representations, such as the Gaussian-type ACSF proposed by Behler and Parrinello,[12,13] the smooth overlap of atomic positions (SOAP),[16,21,25] graph convolutional neural networks (GCN),[23,26,58] the power-type structural descriptor (PTSD) developed by us,[1,24,30] and so on.[19,27] Figure 3a also shows the G-NN potential that follows eqs 3 and 4 and utilizes PTSDs as inputs for the potential, which indeed satisfies quite nicely *rule 1*, where the same structure with different supercell sizes produces exactly the same energy per atom.

The PTSD proposed by our group in 2018,[24] as shown in eqs 5−12, belongs to ACSF for representing the atomic environment (Figure 4). A set of structure descriptors labelled as S1−S6 can be obtained by combining these atom-centered power-type structure descriptors S1−S6. It is a set of highly sophisticated descriptors, which are developed to be compatible with the SSW global optimization data set. In PTSD, the traditional two-body and three-body terms are included in addition to the inclusion of four-body terms and the introduction of spherical functions, which enhance the structure discrimination. In particular, the combination of the power function and spherical harmonic function in S2 and S5 mimics the atomic wave functions, which provides a convenient way to couple the radial and angular information of an atom. The four-body term in S6 can describe the dihedral angle and is thus critical for describing organic molecular configurations and reactions.

$$f_c(r_{ij}) = \begin{cases} 0.5 \times \tanh^3\left[1 - \frac{r_{ij}}{r_c}\right], & \text{for } r_{ij} \le r_c \\ 0, & \text{for } r_{ij} > r_c \end{cases} \tag{5}$$

$$R^n(r_{ij}) = r_{ij}^n \cdot f_c(r_{ij}) \tag{6}$$

$$S_i^1 = \sum_{j \neq i} R^n(r_{ij}) \tag{7}$$

$$S_i^2 = \left[\sum_{m=-L}^{L} \left| \sum_{j \neq i} R^n(r_{ij}) Y_{Lm}(\mathbf{r}_{ij}) \right|^2 \right]^{1/2} \tag{8}$$

$$S_i^3 = 2^{1-\zeta} \sum_{j,k \neq i} (1 + \lambda \cos\theta_{ijk})^{\zeta} \cdot R^n(r_{ij}) \cdot R^m(r_{ik}) \cdot R^p(r_{jk}) \tag{9}$$

$$S_i^4 = 2^{1-\zeta} \sum_{j,k \neq i} (1 + \lambda \cos\theta_{ijk})^{\zeta} \cdot R^n(r_{ij}) \cdot R^m(r_{ik}) \tag{10}$$

$$S_i^5 = \left[\sum_{m=-L}^{L} \left| \sum_{j,k \neq i} R^n(r_{ij}) \cdot R^m(r_{ik}) \cdot R^p(r_{jk}) \cdot (Y_{Lm}(\mathbf{r}_{ij}) + Y_{Lm}(\mathbf{r}_{ik})) \right|^2 \right]^{1/2} \tag{11}$$

$$S_i^6 = 2^{1-\zeta} \sum_{j,k,l \neq i} (1 + \lambda \cos\delta_{ijkl})^{\zeta} \cdot R^n(r_{ij}) R^m(r_{ik}) R^p(r_{il}) \tag{12}$$

In the eqs 5−12, r_{ij} is the internuclear distance between atom i and j and θ_{ijk} is the angle centered at i atom with j and k being neighbors (i, j, and k are atom indices). The key ingredients in the PTSD are the cutoff function f_c that decays to zero beyond the r_c (eq 5), the power-type radial function, the trigonometric angular functions, and the spherical harmonic function. S^1 and S^2 are two-body functions; S^3, S^4, and S^5 are three-body functions; and S^6 is a four-body function. Obviously, the PTSD functions

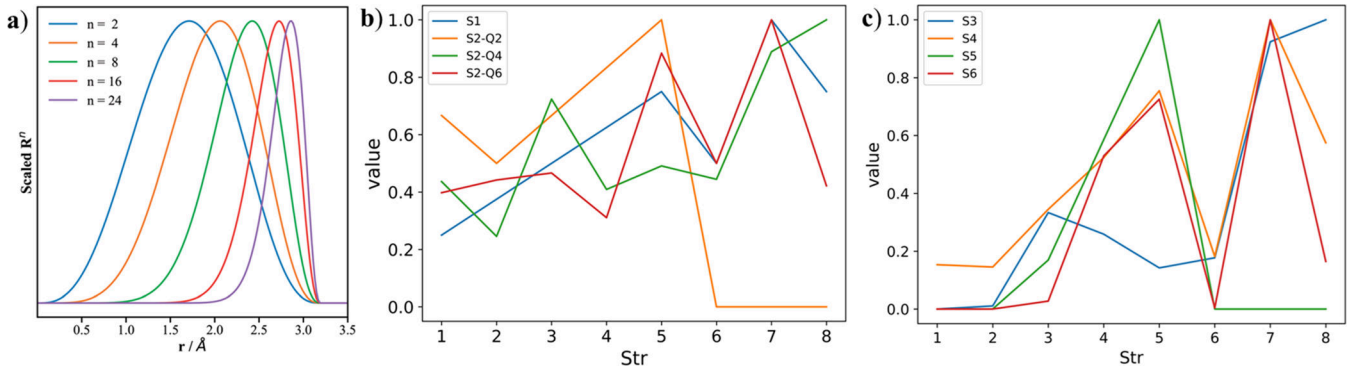


Figure 5. (a) Plots of the radial part of the PTSDs, R^n (see eq 6), for the same cutoff radius of 3.2 Å but with different powers n. The x-axis is the distance r, while the y-axis is the function value scaled to (0, 1). (b and c) Plots of the two-body functions of PTSDs (S1 and S2) and three-body and four-body functions of the PTSDs (S3−S6) for a center atom in different coordination environments corresponding to structures 1 to 8 labeled on the x-axis. Legend: structure 1, line; structure 2, triangle; structure 3, square; structure 4, pentagon; structure 5, hexagon; structure 6, tetrahedron; structure 7, cube; structure 8, octahedron. Panel a is reproduced with permission from ref 24. Copyright 2019 Royal Society of Chemistry.

https://dx.doi.org/10.1021/acs.accounts.0c00472
Acc. Chem. Res. 2020, 53, 2119−2129

also satisfy *rule 2*, where the derivatives of the atomic coordinate can be derived analytically.

Figure 5a shows that the power function with only one parameter (n) when combined with the decaying cutoff function can create radial distributions with flexible peaks and shapes, which effectively extract the structure information in that particular radial window. The combination of different powers (n, m, and p) in three-body functions can couple conveniently different radial distributions. Figure 5b and c illustrate a set of S1 to S6 values for describing the atoms with different coordination shells, i.e., two- (line), three- (triangle), four- (square), five- (pentagon), six- (hexagon), four- (tetrahedron), eight- (cube), and six-coordination (octahedron) shells. All bond distances are the same (2 Å) in these structures, and the power functions in all PTSDs have the same power of $n = 2$. As shown in Figure 5b, the spherical harmonic function with a different angular moment of $L = 2$, 4, or 6 (see eq 8) can extract angular information of the local environments even in two-body functions. Figure 5c shows that the S3−S6 three-body and four-body terms can further yield different values for the local environment, particularly for high-coordination numbers (>3) and thus are useful complements to the two-body terms. It is the large variation of the PTSD values that achieve the ability to sensitively discriminate different structures.

Similar to many possible flavors of data descriptors, there are a zoo of numerical forms for ML models, which differ in the number of fitting parameters and the complexity of the function form. While feed-forward NNs are often the choice, the function forms in other ML techniques have also been used, such as Gaussian process regression (GPR),[17,18,61,62] kernel ridge regression (KRR),[28] support vector machines (SVMs),[63,64] and spectral neighbor analysis potentials (SNAPs).[65] It is not straightforward to compare different ML models directly due to different target systems involved in the literature. Nevertheless, we have compiled the reported accuracy of different ML models in Table 1 and Table 2, where the data sets are all constructed by

Table 1. ML Applications in Literatures, Where the Global Dataset from Global Optimization Methods or from the Large Existing Database Are Utilized for Training Potentials[a]

system	description	model	DatGen	RMSE, meV/atom
Si[59]	ACSF[b]	NN	MD/MC	5
GeTe[14]	ACSF[b]	NN	MD	5.6
CHON[22]	ANI	NN	NMS	3
LiSi[44]	ADF/RDF[c]	NN	GA	6.3
B[29]	ACSF[b]	GPR	PSO	53

[a]The listed data include the data descriptors, ML models, data generation schemes, and root mean squared errors (RMSEs) in the energy. [b]ACSF in the Behler scheme.[13] [c]Descriptor based on the expansion of the radial and angular distribution functions.[60]

global optimization techniques or from an existing large database and not from a single MD trajectory or a single target reaction. These examples thus represent the state-of-the-art for constructing PES with ML potentials.

Specifically, Table 1 lists the ML potentials by using non-SSW data collection methods, including MD, NMS, GA, and PSO. They can be compared with Table 2, which lists the G-NN potentials reported so far. Both Table 1 and Table 2 contain CHON and B potentials. The CHON potential (ANI) in Table 1 utilizes the NMS for data collection and can be used only for predicting the structures nearby minima, while the CHON G-NN can be used for reaction prediction, as shown below in section 4.2. For a B potential, a single-element system with diverse phases, a high-accuracy ML potential is technically difficult to construct: the reported accuracy from SSW-NN data sets is 12 meV/atom, and that from the PSO data sets is above 50 meV/atom.

Table 2. G-NN Potentials Generated by the SSW-NN Method[a]

system	NN arch.	data set size	RMSE, meV/atom
TiOH[66]	201−50−50−1	143786	9.8
B[24]	173−110−110−1	165423	12.4
H_2O[45]	156−50−5−1	58825	2.1
ZnCrO[2]	324−80−60−60−1	38285	4.3
CoO[46]	148−80−50−50−1	42246	12.1
AuCeO[47]	183−60−50−50−1	33654	6.1
CHON[3]	407−120−80−80−1	94854	10.1
YZrO[49]	188−60−50−50−1	28803	7.7

[a]Listed data includes the feed-forward NN architecture (nodes and layers), the global dataset size and the RMSE in energy.

It should be emphasized that the complexity of functional forms and the number of structure descriptors required for ML are relevant to the structure diversity of the data set. This can be seen from the network size for different G-NN potentials listed in Table 2, where two to three hidden layers are generally required in NN architecture for achieving the low RMSE in energy (e.g., below 10 meV/atom).

4. APPLICATIONS

Due to the local representation of data descriptors and the numerical functions in ML models, ML potential simulations are, in principle, linear scaling with the system size,[67] and the speed can be more than 4 orders of magnitude faster than DFT, even for medium size systems (e.g., 48 atoms, also see Figure 3b).[1] Thus, a major field of ML potential applications is structure prediction, e.g., searching for new crystal phases,[1,59] and recently, ML potentials are also utilized to expedite the reaction space exploration for finding reaction mechanisms. Recent years have seen interesting ML applications in material science, cluster science, and biology.[44,68,69] In the following section, we present two SSW-NN applications to illustrate ML applications on structure prediction and reaction network exploration.

4.1. Material Simulation for Predicting Phase Diagrams

Multielement oxide represents a class of solid materials with complex structures. Force field methods were often utilized in the atomic simulation of oxides, especially when the atomic force can be approximated by fixed charge ionic interactions, e.g., in ZrO_2. For many other oxides, such as $Zn_xCr_yO_z$, where Cr has variable oxidation states (e.g., Cr^{3+} and Cr^{4+}), it has long been a problem to characterize the atomic structure of oxides with varying Zn:Cr ratios. With the advent of the SSW-NN method, it is now possible to establish the global PES for these complex oxide systems, and thus, large-scale ML-based atomic simulations can be utilized to understand these materials' physicochemical properties.

ZnCr oxide is one of the first generation of industry catalysts for syngas-to-methanol conversion,[70] being extensively studied in experiments since the 1930s. While the Zn:Cr ratios can significantly influence the syngas-to-methanol catalytic activity and selectivity, the structure of the catalyst has been debated for

https://dx.doi.org/10.1021/acs.accounts.0c00472
Acc. Chem. Res. 2020, 53, 2119−2129

decades.[71−74] By using the SSW-NN method, we recently established the ternary ZnCrO global data set (see Table 2).[2] It contains structures from 10 to 84 atoms per cell and cover different Zn:Cr:O ratios, i.e., ZnO, CrO_x, and $ZnCr_xO_y$, with different morphology forms, e.g., bulks, layers, and clusters. The large varieties of Zn:Cr ratios for structures in data sets provide the basis for predicting different oxidation states for Cr in different chemical environments.

The extensive SSW global search for different ZnCrO compositions was then performed by using the ZnCrO G-NN potential, which leads to the ternary phase diagram for ZnCrO, as shown in Figure 6. The spinel-type crystalline structures

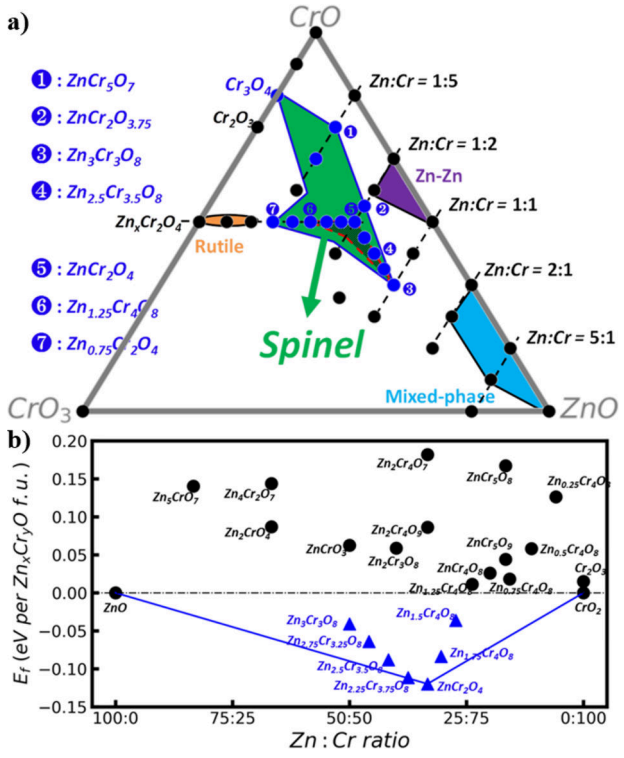


Figure 6. (a) Ternary Zn−Cr−O phase diagram. The green region maps out the compositions, with the spinel-type skeleton structure as the global minimum, and the blue circles labeled by numbers indicate the composition. Only the spinel ZnCrO phases in the red dashed triangle are thermodynamically allowed. (b) Convex hulls for all the ZnCrO structures are indicated by the blue line. The blue triangles and black circles represent the negative and positive formation energies compared to the ZnO and CrO_2 phases, respectively. f.u. = formula unit. Reproduced with permission from ref 2. Copyright 2019 Springer Nature.

appear as a major motif for the ZnCrO materials, where the Zn composition ranges from Zn:Cr ratios of 0:1 to 1:1. The thermodynamics phase diagram of Zn−Cr−O further reveals the presence of a small stable composition island, i.e., Zn:Cr:O = 6:6:16−3:8:16, where the oxide tends to crystallize into a spinel crystal phase (Figure 6b). While $ZnCr_2O_4$ is the most stable phase in thermodynamics, the change of the Zn:Cr ratio from 1:2 to 1:1 ($Zn_3Cr_3O_8$) leads to the generation of a series of metastable crystal phases, which contain the unusual $[ZnO_6]$ octahedra (O_h) in bulk. By further investigating the surface phase diagram, we found that, owing to the presence of $[ZnO_6]_{Oh}$, the oxygen vacancy (O_v) formation ability increases appreciably and extends from the surface to subsurface with the increase of the Zn:Cr ratio. Under syngas reaction conditions, the catalyst surfaces of $Zn_3Cr_3O_8$ expose unprecedented four-coordinated planar Cr^{2+} cations, while only five-coordinated Cr cations with a pyramid geometry can be exposed on the catalyst surfaces of $ZnCr_2O_4$. This subtle structural difference turns out to be critical in how it profoundly affects syngas conversion activity and selectivity.[2]

4.2. Reaction Simulation for Finding Reaction Pathways and Mechanisms

Organic reactions lie at the heart of chemistry. Recent years have witnessed the encouraging progress of ML techniques in the reverse design of synthetic routes based on the existing experimental database of organic synthesis.[76−79] These knowledge-based approaches, however, lack the insight into the mechanisms of chemical reactions and by no means can replace atomic simulations. To apply ML techniques to simulate organic reactions, there are two outstanding difficulties in building the global PES of organic reactions. First, the configurational and reactive space of organic molecules is astronomically huge, involving many elements (at least C, H, O, or N) and many different bonds with different bond orders (C−C, C═C, C−H, or C−O···). Second, the transition region of organic reactions is often narrow in space and has a high energy with respect to the initial state due to covalent bond making/breaking (e.g., >0.7 eV).

We recently applied the SSW-NN method to resolve the glucose pyrolysis reaction network.[3] As glucose with the $C_6H_{12}O_6$ formula virtually has enumerable likely pathways involving C−O−H elements in organic chemistry. The construction of a general-purpose C−H−O G-NN potential is necessary to describe the large number of possible reaction pathways. Our SSW-NN simulation for data set generation focused on collecting basic local reaction patterns by considering that the global pathways can be seen as combinations of simpler local patterns. First, we have utilized the short-time SSW sampling of molecular crystal systems (in variable periodic cells) containing C−H−O−N elements via DFT calculations. The starting molecular structures for these simulations are randomly selected from QM9.[32,33] On the basis of this small data set from DFT sampling, the first NN potential is trained and then utilized for global sampling of organic molecules and molecular crystals that are performed iteratively to expand the reactive PES data set. The starting structures are also randomly selected from the QM9 database. Finally, the SSW-NN global sampling for the glucose pyrolysis reaction network is used to further improve the transferability of the G-NN potential for glucose chemistry. The starting structures are updated and randomly selected from glucose global optimization SSW trajectories. The final data set has 94854 structures, containing nearly all (78) bonding patterns with C−H−O−N elements (Table 2).

Benefiting from the low cost of the G-NN PES, we can achieve a deep exploration of the reaction tree starting from D-glucose. In total, we managed to sample 1200000 minima and collected more than 150000 reaction pairs. After removing duplicate reactions and recording only the lowest barrier connection between pairs, the final reaction database, as shown in Figure 7, contains 4455 unique molecules and 6407 different reactions with 3488 reaction patterns. We have carefully analyzed the pathways in our reaction database to identify the pathways to 5-hydroxymethylfurfural (HMF), a major and valuable product observed in experiments.[80−82]

https://dx.doi.org/10.1021/acs.accounts.0c00472
Acc. Chem. Res. 2020, 53, 2119−2129

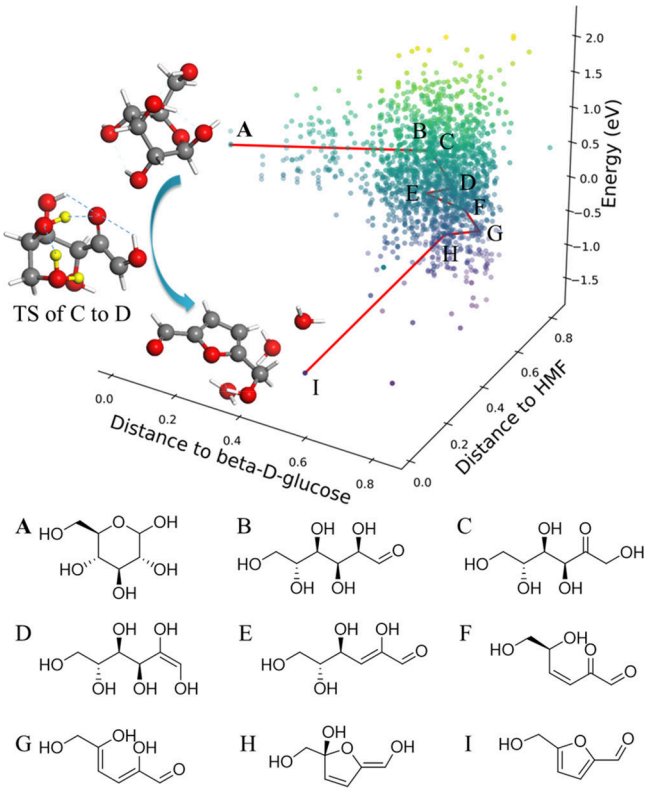


Figure 7. Reaction database and the lowest energy pathway sampled by the SSW-NN method for glucose pyrolysis. The *x*- and *y*-axes are the similarity distances of intermediates with respect to *β*-D-glucose and HMF, respectively. The similarity distances are calculated using the fingerprint algorithm in RDkit using the Tanimoto similarity.[75] The lowest energy pathway from beta-D-glucose to HMF is marked by a red line, and the structures along the lowest energy pathway, **A** to **I**, are shown below.

In the lowest energy pathway, *β*-D-glucopyranose undergoes ring-opening, isomerization, tautomerization, dehydration, and cyclization to HMF (see Figure 7). Overall, the rate-determining step belongs to the enol−keto tautomerization reaction (**C** → **D**), with a barrier of 1.91 eV (with respect to the most stable configuration of *β*-D-glucopyranose hereafter), which is 0.19 eV lower than the previous pathways (2.10 eV in *β*-H elimination).[3] The enol−keto tautomerization reaction overall benefits the mechanism due to the opening of the retro-Michael-addition route in the subsequent dehydration reactions and the avoidance of direct *β*-H elimination.

5. CONCLUDING REMARKS

Current ML techniques greatly expand the time-scale and the scope of atomic simulations. This perspective overviews three key aspects of ML potentials that are critical to the performance, i.e., the quality of the data set, the ML models, and the structure descriptors. We show that the major drawbacks of ML potentials, e.g., the lack of transferability and robustness that are intrinsic to the numerical function with a huge number of fitting parameters, can be largely circumvented with the help of efficient global PES sampling techniques. In the future, we believe that there is still ample room to improve ML potential techniques, which could further expedite material and reaction simulations and produce new research directions. To be specific, we elaborate two ongoing research directions in our group.

5.1. ML Potentials Integrated with Electronic Structure Information

All current ML potentials only take into account the geometrical information on structures and output the total energy of the system. This is certainly associated with the high cost in computing electronic structures (e.g., atomic charge) for new structures based on quantum mechanics. However, considering that the electronic structure characteristics are intrinsic to many important applications, the low-cost methods to correlate the electronic structure with the geometrical structure are certainly much more desirable and deserve future research efforts.

5.2. Integrated ML Models for Material and Reaction Prediction

Apart from atomic simulations, ML techniques have been applied to many other fields in chemical property prediction, such as reaction mechanisms,[76] band gaps,[58] structure of proteins,[83,84] and so on. In fact, property prediction is a coarse-grained predictor compared to atomic simulations but can provide quick ideas on possible reaction patterns and desirable catalyst compositions, to name a few. It is therefore desirable to couple them with atomic simulations, particularly for global material and reaction searches, to accelerate the simulation. On the other hand, the atomic simulation can validate these predicted properties and provide new data for the construction of such a model.

AUTHOR INFORMATION

Corresponding Authors

Cheng Shang − *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/0000-0001-7486-1514; Email: cshang@fudan.edu.cn

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

Author

Pei-Lin Kang − *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*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.accounts.0c00472

Notes

The authors declare no competing financial interest.

Biographies

Pei-Lin Kang received a BS degree from Fudan University (Shanghai, China) in 2017 and currently studies as a PhD student at Fudan university under the supervision of Prof. Zhi-Pan Liu. He is working on machine learning potential construction and reaction exploration methods.

Cheng Shang graduated in Chemistry at Fudan University where he also received his PhD degree in 2013 under the supervision of Prof. Zhi-

https://dx.doi.org/10.1021/acs.accounts.0c00472
Acc. Chem. Res. 2020, 53, 2119−2129

Pan Liu. He then moved to the University of Cambridge, where he joined the theory group supervised by Prof. David Wales. In 2015 he joined Fudan University, where he currently holds a position as an associate professor in Physical Chemistry. His current research interests are mainly focused on methodology development for potential energy surface exploration and applications on crystal structure predication.

Zhi-Pan Liu graduated in Chemistry at Shanghai JiaoTong University (Shanghai, China). He received his PhD degree in 2003 at Queens University of Belfast under the supervision of Prof. Peijun Hu. He then moved to the University of Cambridge, where he joined the surface science group supervised by Prof. Sir David King. In 2005 he joined Fudan University, where he has since stayed, with a position as a Changjiang professor in Physical Chemistry. His current research interests include methodology development for potential energy surface exploration, theory on heterogeneous catalysis, and machine-learning-based atomic simulation.

■ ACKNOWLEDGMENTS

This work was supported by the National Key Research and Development Program of China (2018YFA0208600) and the National Science Foundation of China (91945301, 91745201, 21533001, and 91645201).

■ REFERENCES

(1) Huang, S.-D.; Shang, C.; Zhang, X.-J.; Liu, Z.-P. Material discovery by combining stochastic surface walking global optimization with a neural network. *Chem. Sci.* **2017**, *8*, 6327−6337.
(2) Ma, S.; Huang, S.-D.; Liu, Z.-P. Dynamic coordination of cations and catalytic selectivity on zinc−chromium oxide alloys during syngas conversion. *Nat. Catal.* **2019**, *2*, 671−677.
(3) Kang, P.-L.; Shang, C.; Liu, Z.-P. Glucose to 5-Hydroxymethylfurfural: Origin of Site-Selectivity Resolved by Machine Learning Based Reaction Sampling. *J. Am. Chem. Soc.* **2019**, *141*, 20525−20536.
(4) Hohenberg, P.; Kohn, W. J. P. r. Inhomogeneous electron gas. *Phys. Rev.* **1964**, *136*, B864−871.
(5) Parr, R. G.; Yang, W. *Density Functional Theory of Atoms and Molecules*; Springer Netherlands: Dordrecht, The Netherlands, 1980.
(6) Car, R.; Parrinello, M. Unified Approach for Molecular Dynamics and Density-Functional Theory. *Phys. Rev. Lett.* **1985**, *55*, 2471−2474.
(7) Zhao, Y.; Truhlar, D. G. Density Functionals with Broad Applicability in Chemistry. *Acc. Chem. Res.* **2008**, *41*, 157−167.
(8) Rousseau, R.; Glezakou, V.-A.; Selloni, A. Theoretical insights into the surface physics and chemistry of redox-active oxides. *Nat. Rev. Mater.* **2020**, *5*, 460−475.
(9) Jordan, M. I.; Mitchell, T. M. Machine learning: Trends, perspectives, and prospects. *Science* **2015**, *349*, 255−260.
(10) Blank, T. B.; Brown, S. D.; Calhoun, A. W.; Doren, D. J. Neural-Network Models of Potential-Energy Surfaces. *J. Chem. Phys.* **1995**, *103*, 4129−4137.
(11) Liu, T.; Fu, B.; Zhang, D. H. Six-dimensional quantum dynamics study for the dissociative adsorption of HCl on Au(111) surface. *J. Chem. Phys.* **2013**, *139*, 184705.
(12) Behler, J.; Parrinello, M. Generalized Neural-Network Representation of High-Dimensional Potential-Energy Surfaces. *Phys. Rev. Lett.* **2007**, *98*, 146401.
(13) Behler, J. Atom-centered symmetry functions for constructing high-dimensional neural network potentials. *J. Chem. Phys.* **2011**, *134*, 074106.
(14) Sosso, G. C.; Miceli, G.; Caravati, S.; Behler, J.; Bernasconi, M. Neural network interatomic potential for the phase change material GeTe. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2012**, *85* (17), 174103.
(15) Rupp, M.; Tkatchenko, A.; Muller, K. R.; von Lilienfeld, O. A. Fast and accurate modeling of molecular atomization energies with machine learning. *Phys. Rev. Lett.* **2012**, *108*, 058301.
(16) Bartók, A. P.; Kondor, R.; Csányi, G. On representing chemical environments. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2013**, *87* (18), 184115.
(17) Li, Z.; Kermode, J. R.; De Vita, A. Molecular dynamics with on-the-fly machine learning of quantum-mechanical forces. *Phys. Rev. Lett.* **2015**, *114*, 096405.
(18) Bartók, A. P.; Csányi, G. Gaussian approximation potentials: A brief tutorial introduction. *Int. J. Quantum Chem.* **2015**, *115*, 1051−1057.
(19) Behler, J. Perspective: Machine learning potentials for atomistic simulations. *J. Chem. Phys.* **2016**, *145*, 170901.
(20) Kolb, B.; Zhao, B.; Li, J.; Jiang, B.; Guo, H. Permutation invariant potential energy surfaces for polyatomic reactions using atomistic neural networks. *J. Chem. Phys.* **2016**, *144*, 224103.
(21) De, S.; Bartok, A. P.; Csanyi, G.; Ceriotti, M. Comparing molecules and solids across structural and alchemical space. *Phys. Chem. Chem. Phys.* **2016**, *18*, 13754−13769.
(22) Smith, J. S.; Isayev, O.; Roitberg, A. E. ANI-1: an extensible neural network potential with DFT accuracy at force field computational cost. *Chem. Sci.* **2017**, *8*, 3192−3203.
(23) Schutt, K. T.; Arbabzadah, F.; Chmiela, S.; Muller, K. R.; Tkatchenko, A. Quantum-chemical insights from deep tensor neural networks. *Nat. Commun.* **2017**, *8*, 13890.
(24) Huang, S.-D.; Shang, C.; Kang, P.-L.; Liu, Z.-P. Atomic structure of boron resolved using machine learning and global sampling. *Chem. Sci.* **2018**, *9*, 8644−8655.
(25) Jäger, M. O. J.; Morooka, E. V.; Federici Canova, F.; Himanen, L.; Foster, A. S. Machine learning hydrogen adsorption on nanoclusters through structural descriptors. *NPJ. Comput. Mater.* **2018**, *4*, 37.
(26) Schutt, K. T.; Sauceda, H. E.; Kindermans, P. J.; Tkatchenko, A.; Muller, K. R. SchNet - A deep learning architecture for molecules and materials. *J. Chem. Phys.* **2018**, *148*, 241722.
(27) Zhang, L.; Han, J.; Wang, H.; Car, R.; E, W. Deep Potential Molecular Dynamics: A Scalable Model with the Accuracy of Quantum Mechanics. *Phys. Rev. Lett.* **2018**, *120*, 143001.
(28) Jacobsen, T. L.; Jorgensen, M. S.; Hammer, B. On-the-Fly Machine Learning of Atomic Potential in Density Functional Theory Structure Optimization. *Phys. Rev. Lett.* **2018**, *120*, 026102.
(29) Tong, Q.; Xue, L.; Lv, J.; Wang, Y.; Ma, Y. Accelerating CALYPSO structure prediction by data-driven learning of a potential energy surface. *Faraday Discuss.* **2018**, *211*, 31−43.
(30) Huang, S. D.; Shang, C.; Kang, P. L.; Zhang, X. J.; Liu, Z. P. LASP: Fast global potential energy surface exploration. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2019**, *9* (6), e1415.
(31) Fink, T.; Bruggesser, H.; Reymond, J. L. Virtual exploration of the small-molecule chemical universe below 160 Da. *Angew. Chem., Int. Ed.* **2005**, *44*, 1504−1508.
(32) Ruddigkeit, L.; van Deursen, R.; Blum, L. C.; Reymond, J. L. Enumeration of 166 billion organic small molecules in the chemical universe database GDB-17. *J. Chem. Inf. Model.* **2012**, *52*, 2864−2875.
(33) Ramakrishnan, R.; Dral, P. O.; Rupp, M.; von Lilienfeld, O. A. Quantum chemistry structures and properties of 134 kilo molecules. *Sci. Data* **2014**, *1* (1), 140022.
(34) Zhang, X.-J.; Shang, C.; Liu, Z.-P. Pressure-induced silica quartz amorphization studied by iterative stochastic surface walking reaction sampling. *Phys. Chem. Chem. Phys.* **2017**, *19*, 4725−4733.
(35) Kirkpatrick, S.; Gelatt, C. D.; Vecchi, M. P. Optimization by Simulated Annealing. *Science* **1983**, *220*, 671−680.
(36) Pannetier, J.; Bassas-Alsina, J.; Rodriguez-Carvajal, J.; Caignaert, V. Prediction of crystal structures from crystal chemistry rules by simulated annealing. *Nature* **1990**, *346*, 343−345.
(37) Schön, J. C.; Jansen, M. First Step Towards Planning of Syntheses in Solid-State Chemistry: Determination of Promising Structure Candidates by Global Optimization. *Angew. Chem., Int. Ed. Engl.* **1996**, *35*, 1286−1304.
(38) Gastegger, M.; Marquetand, P. High-Dimensional Neural Network Potentials for Organic Reactions and an Improved Training Algorithm. *J. Chem. Theory Comput.* **2015**, *11*, 2187−2198.

https://dx.doi.org/10.1021/acs.accounts.0c00472
Acc. Chem. Res. 2020, 53, 2119−2129

(39) Herr, J. E.; Yao, K.; McIntyre, R.; Toth, D. W.; Parkhill, J. Metadynamics for training neural network model chemistries: A competitive assessment. *J. Chem. Phys.* **2018**, *148*, 241710.
(40) Amabilino, S.; Bratholm, L. A.; Bennie, S. J.; Vaucher, A. C.; Reiher, M.; Glowacki, D. R. Training Neural Nets To Learn Reactive Potential Energy Surfaces Using Interactive Quantum Chemistry in Virtual Reality. *J. Phys. Chem. A* **2019**, *123*, 4486−4499.
(41) Smith, J. S.; Nebgen, B.; Lubbers, N.; Isayev, O.; Roitberg, A. E. Less is more: Sampling chemical space with active learning. *J. Chem. Phys.* **2018**, *148*, 241733.
(42) Smith, J. S.; Zubatyuk, R.; Nebgen, B.; Lubbers, N.; Barros, K.; Roitberg, A. E.; Isayev, O.; Tretiak, S. The ANI-1ccx and ANI-1x data sets, coupled-cluster and density functional theory properties for molecules. *Sci. Data* **2020**, *7* (1), 134.
(43) Wales, D. J.; Doye, J. P. K. Global Optimization by Basin-Hopping and the Lowest Energy Structures of Lennard-Jones Clusters Containing up to 110 Atoms. *J. Phys. Chem. A* **1997**, *101*, 5111−5116.
(44) Artrith, N.; Urban, A.; Ceder, G. Constructing first-principles phase diagrams of amorphous LixSi using machine-learning-assisted sampling with an evolutionary algorithm. *J. Chem. Phys.* **2018**, *148*, 241711.
(45) Guan, S.-H.; Shang, C.; Huang, S.-D.; Liu, Z.-P. Two-Stage Solid-Phase Transition of Cubic Ice to Hexagonal Ice: Structural Origin and Kinetics. *J. Phys. Chem. C* **2018**, *122*, 29009−29016.
(46) Kong, F.-C.; Li, Y.-F.; Shang, C.; Liu, Z.-P. Stability and Phase Transition of Cobalt Oxide Phases by Machine Learning Global Potential Energy Surface. *J. Phys. Chem. C* **2019**, *123*, 17539−17547.
(47) Huang, S.-D.; Shang, C.; Liu, Z.-P. Ultrasmall Au clusters supported on pristine and defected CeO_2: Structure and stability. *J. Chem. Phys.* **2019**, *151*, 174702.
(48) Li, Y.-F.; Liu, Z.-P. Active Site Revealed for Water Oxidation on Electrochemically Induced δ-MnO2: Role of Spinel-to-Layer Phase Transition. *J. Am. Chem. Soc.* **2018**, *140*, 1783−1792.
(49) Guan, S.-H.; Zhang, K.-X.; Shang, C.; Liu, Z.-P. Stability and anion diffusion kinetics of Yttria-stabilized zirconia resolved from machine learning global potential energy surface exploration. *J. Chem. Phys.* **2020**, *152*, 094703.
(50) Shang, C.; Liu, Z.-P. Stochastic Surface Walking Method for Structure Prediction and Pathway Searching. *J. Chem. Theory Comput.* **2013**, *9*, 1838−1845.
(51) Zhang, X.-J.; Shang, C.; Liu, Z.-P. From Atoms to Fullerene: Stochastic Surface Walking Solution for Automated Structure Prediction of Complex Material. *J. Chem. Theory Comput.* **2013**, *9*, 3252−3260.
(52) Shang, C.; Zhang, X.-J.; Liu, Z.-P. Stochastic surface walking method for crystal structure and phase transition pathway prediction. *Phys. Chem. Chem. Phys.* **2014**, *16*, 17845−17856.
(53) Laio, A.; Parrinello, M. Escaping free-energy minima. *Proc. Natl. Acad. Sci. U. S. A.* **2002**, *99*, 12562−12566.
(54) Metropolis, N.; Rosenbluth, A. W.; Rosenbluth, M. N.; Teller, A. H.; Teller, E. Equation of State Calculations by Fast Computing Machines. *J. Chem. Phys.* **1953**, *21*, 1087−1092.
(55) Steinhardt, P. J.; Nelson, D. R.; Ronchetti, M. Bond-orientational order in liquids and glasses. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1983**, *28*, 784−805.
(56) O'Neil, M. J. *The Merck Index: An Encyclopedia of Chemicals, Drugs, and Biologicals*, 15th ed.; The Royal Society of Chemistry: Cambridge, UK, 2013; pp 1755.
(57) Truhlar, D. G.; Gordon, M. S. From force fields to dynamics: Classical and quantal paths. *Science* **1990**, *249*, 491−498.
(58) Xie, T.; Grossman, J. C. Crystal Graph Convolutional Neural Networks for an Accurate and Interpretable Prediction of Material Properties. *Phys. Rev. Lett.* **2018**, *120*, 145301.
(59) Behler, J.; Martonak, R.; Donadio, D.; Parrinello, M. Metadynamics simulations of the high-pressure phases of silicon employing a high-dimensional neural network potential. *Phys. Rev. Lett.* **2008**, *100*, 185501.
(60) Artrith, N.; Urban, A.; Ceder, G. Efficient and accurate machine-learning interpolation of atomic energies in compositions with many species. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2017**, *96* (1), 014112.
(61) Caro, M. A.; Deringer, V. L.; Koskinen, J.; Laurila, T.; Csanyi, G. Growth Mechanism and Origin of High sp∧{3} Content in Tetrahedral Amorphous Carbon. *Phys. Rev. Lett.* **2018**, *120*, 166101.
(62) Mocanu, F. C.; Konstantinou, K.; Lee, T. H.; Bernstein, N.; Deringer, V. L.; Csanyi, G.; Elliott, S. R. Modeling the Phase-Change Memory Material, Ge2Sb2Te5, with a Machine-Learned Interatomic Potential. *J. Phys. Chem. B* **2018**, *122*, 8998−9006.
(63) Balabin, R. M.; Lomakina, E. I. Support vector machine regression (LS-SVM)−an alternative to artificial neural networks (ANNs) for the analysis of quantum chemistry data? *Phys. Chem. Chem. Phys.* **2011**, *13*, 11710−11718.
(64) Vitek, A.; Stachon, M.; Kromer, P.; Snael, V. Towards the Modeling of Atomic and Molecular Clusters Energy by Support Vector Regression. In *2013 5th International Conference on Intelligent Networking and Collaborative Systems*, Xi'an, China, September 9−11, 2013; pp 121−126.
(65) Thompson, A. P.; Swiler, L. P.; Trott, C. R.; Foiles, S. M.; Tucker, G. J. Spectral neighbor analysis method for automated generation of quantum-accurate interatomic potentials. *J. Comput. Phys.* **2015**, *285*, 316−330.
(66) Ma, S.; Huang, S.-D.; Fang, Y.-H.; Liu, Z.-P. TiH Hydride Formed on Amorphous Black Titania: Unprecedented Active Species for Photocatalytic Hydrogen Evolution. *ACS Catal.* **2018**, *8*, 9711−9721.
(67) Shang, C.; Huang, S.-D.; Liu, Z.-P. Massively parallelization strategy for material simulation using high-dimensional neural network potential. *J. Comput. Chem.* **2019**, *40*, 1091−1096.
(68) Sun, G.; Sautet, P. Toward Fast and Reliable Potential Energy Surfaces for Metallic Pt Clusters by Hierarchical Delta Neural Networks. *J. Chem. Theory Comput.* **2019**, *15*, 5614−5627.
(69) Yao, K.; Herr, J. E.; Toth, D. W.; McKintyre, R.; Parkhill, J. The TensorMol-0.1 model chemistry: a neural network augmented with long-range physics. *Chem. Sci.* **2018**, *9*, 2261−2269.
(70) Waugh, K. Methanol synthesis. *Catal. Today* **1992**, *15*, 51−75.
(71) Molstad, M. C.; Dodge, B. F. Zinc Oxide−Chromium Oxide Catalysts for Methanol Synthesis. *Ind. Eng. Chem.* **1935**, *27*, 134−140.
(72) Errani, E.; Trifiro, F.; Vaccari, A.; Richter, M.; Del Piero, G. Structure and reactivity of Zn-Cr mixed oxides. Role of non-stoichiometry in the catalytic synthesis of methanol. *Catal. Lett.* **1989**, *3*, 65−72.
(73) Bradford, M. C.; Konduru, M. V.; Fuentes, D. X. Preparation, characterization and application of Cr2O3/ZnO catalysts for methanol synthesis. *Fuel Process. Technol.* **2003**, *83*, 11−25.
(74) Song, H.; Laudenschleger, D.; Carey, J. J.; Ruland, H.; Nolan, M.; Muhler, M. Spinel-Structured ZnCr2O4 with Excess Zn Is the Active ZnO/Cr2O3 Catalyst for High-Temperature Methanol Synthesis. *ACS Catal.* **2017**, *7*, 7610−7622.
(75) *RDKit: Open-source cheminformatics*. https://www.rdkit.org/.
(76) Segler, M. H. S.; Preuss, M.; Waller, M. P. Planning chemical syntheses with deep neural networks and symbolic AI. *Nature* **2018**, *555*, 604−610.
(77) Coley, C. W.; Green, W. H.; Jensen, K. F. Machine Learning in Computer-Aided Synthesis Planning. *Acc. Chem. Res.* **2018**, *51*, 1281−1289.
(78) Coley, C. W.; Jin, W.; Rogers, L.; Jamison, T. F.; Jaakkola, T. S.; Green, W. H.; Barzilay, R.; Jensen, K. F. A graph-convolutional neural network model for the prediction of chemical reactivity. *Chem. Sci.* **2019**, *10*, 370−377.
(79) Schwaller, P.; Laino, T.; Gaudin, T.; Bolgar, P.; Hunter, C. A.; Bekas, C.; Lee, A. A. Molecular transformer: A model for uncertainty-calibrated chemical reaction prediction. *ACS Cent. Sci.* **2019**, *5*, 1572−1583.
(80) Fang, Y.; Li, J.; Chen, Y.; Lu, Q.; Yang, H.; Wang, X.; Chen, H. Experiment and Modeling Study of Glucose Pyrolysis: Formation of 3-Hydroxy-γ-butyrolactone and 3-(2H)-Furanone. *Energy Fuels* **2018**, *32*, 9519−9529.

https://dx.doi.org/10.1021/acs.accounts.0c00472
Acc. Chem. Res. 2020, 53, 2119−2129

(81) Patwardhan, P. R.; Satrio, J. A.; Brown, R. C.; Shanks, B. H. Product distribution from fast pyrolysis of glucose-based carbohydrates. *J. Anal. Appl. Pyrolysis* **2009**, *86*, 323−330.
(82) Mayes, H. B.; Nolte, M. W.; Beckham, G. T.; Shanks, B. H.; Broadbelt, L. J. The Alpha−Bet(a) of Glucose Pyrolysis: Computational and Experimental Investigations of 5-Hydroxymethylfurfural and Levoglucosan Formation Reveal Implications for Cellulose Pyrolysis. *ACS Sustainable Chem. Eng.* **2014**, *2*, 1461−1473.
(83) Jianlin Cheng; Tegge, A.N.; Baldi, P. Machine Learning Methods for Protein Structure Prediction. *IEEE Rev. Biomed. Eng.* **2008**, *1*, 41−49.
(84) Wang, J.; Cao, H.; Zhang, J. Z. H.; Qi, Y. Computational Protein Design with Deep Learning Neural Networks. *Sci. Rep.* **2018**, *8*, 6349.

https://dx.doi.org/10.1021/acs.accounts.0c00472
Acc. Chem. Res. 2020, 53, 2119−2129