nature catalysis

Review Article

https://doi.org/10.1038/s41929-023-00911-w

Bridging the complexity gap in computational heterogeneous catalysis with machine learning

Received: 12 October 2022

Accepted: 4 January 2023

Published online: 23 February 2023

Check for updates

Tianyou Mou[1,4], Hemanth Somarajan Pillai[1,4], Siwen Wang[1,4], Mingyu Wan[2], Xue Han[1], Neil M. Schweitzer[3], Fanglin Che[2] & Hongliang Xin[1] ✉

Heterogeneous catalysis underpins a wide variety of industrial processes including energy conversion, chemical manufacturing and environmental remediation. Significant advances in computational modelling towards understanding the nature of active sites and elementary reaction steps have occurred over the past few decades. The complexity gap between theory and experiment, however, remains overwhelming largely due to the limiting length and timescales of ab initio simulations, which severely impede the discovery of high-performance catalytic materials. This Review summarizes recent developments and applications of machine learning to narrow and, optimistically, bridge the gap created by the dynamic, mechanistic and chemostructural complexities inherent to the reactive interfaces of practical relevance. We foresee the prospects and challenges of machine learning for the automated design of sustainable catalytic technologies within a data-centric ecosystem that coevolves with computational and data sciences.

Catalysis is a highly complex, multiscale phenomenon of chemical and energy transformations at active sites. Probing underpinning processes to infer design knowledge and strategies for high-performance catalytic materials has been a long-standing goal in catalysis, the realization of which is essential for the transition of our society to a sustainable future. Owing to the intricacies of phase boundaries between a bulk-like substrate and the continuum environment, heterogeneous catalysis as a subdiscipline, on a par with its homogeneous and enzymic counterparts, poses unique challenges, for example, site ambiguity and pathway diversity[1]. With the advent of quantum chemistry and ever-growing computing power, ab initio methods, for example, density functional theory (DFT)[2], have been increasingly used to model catalytic interfaces represented by dozens to hundreds of atoms with open or periodic boundary conditions. The energetics of elementary reaction steps that occur therein can be passed on to multiscale modelling techniques, for example, microkinetics and molecular dynamics, to extend the length and timescales of atomistic simulations towards practical relevance, linking microscopic events to macroscopic observables[3,4]. Within this computational framework, a tremendous number of fundamental insights into how a catalyst possibly functions can be obtained. Indeed, many catalysts were theoretically predicted and further validated by experiments, albeit for relatively simple systems[5]. Nevertheless, computational modelling has arguably pushed the frontier of heterogeneous catalysis to the degree of sophistication today.

Despite advances in depicting active sites and their interactions with environmental factors, there has always been an apparent gap[6,7] between the often idealized model systems amenable to computational modelling and the underlying complexities of operando experiments, which renders the design of industrial catalysts still a largely trial-and-error practice driven by chemical intuition. In retrospect, it has been long recognized that active sites are dynamic on exposure to reactive species, and evolve into site ensembles distinct from

[1]Department of Chemical Engineering, Virginia Polytechnic Institute and State University, Blacksburg, VA, USA. [2]Department of Chemical Engineering, University of Massachusetts Lowell, Lowell, MA, USA. [3]Department of Chemical and Biological Engineering, Northwestern University, Evanston, IL, USA. [4]These authors contributed equally: Tianyou Mou, Hemanth Somarajan Pillai, Siwen Wang. ✉e-mail: hxin@vt.edu

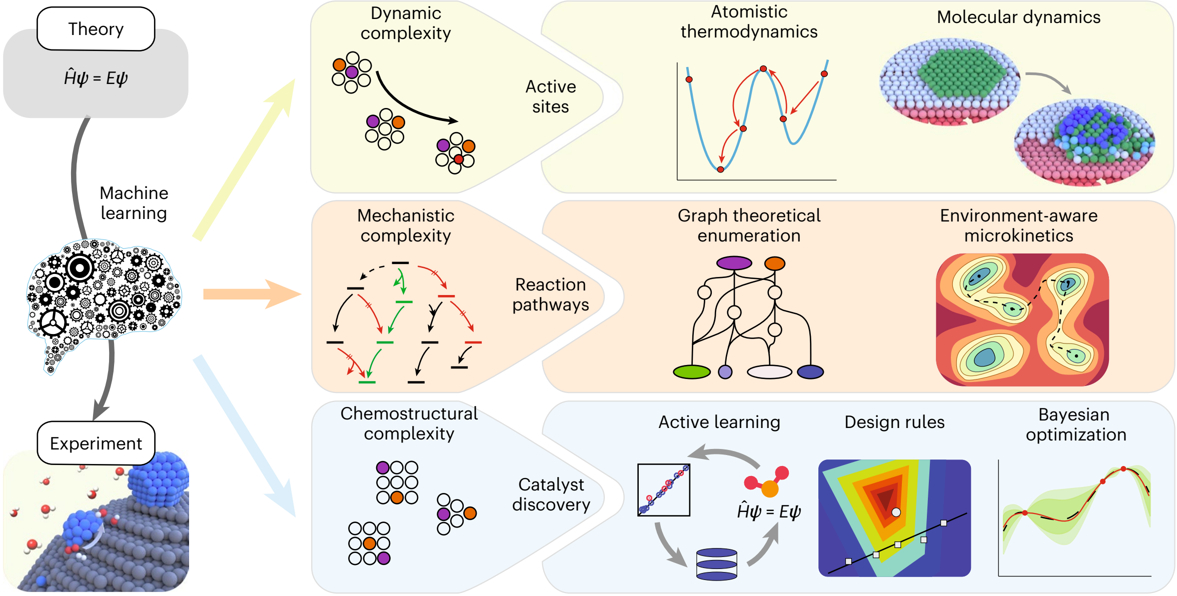

Fig. 1 | Bridging the theory–experiment gap in computational heterogeneous catalysis with machine learning. The dynamic, mechanistic and chemostructural complexities of operando catalytic systems pose grand challenges in revealing the nature of active sites, unravelling reaction pathways and ultimately accelerating catalyst discovery.

as-prepared samples[8]. The dynamic nature of catalysts was brought to attention by Boudart[9] in 1952, stating that "A consequence of the dynamic picture of a catalytic surface presented here is the necessity of devising methods for characterizing the surface during the global catalytic reaction." That is very much true for the need of computational methods to capture the dynamic complexity of active sites under experimentally relevant conditions. Moreover, catalytic reaction networks that involve molecules of immediate interest typically consist of several bond-breaking and formation steps that may be further exacerbated by site coordination. The mechanistic complexity becomes intractable for reactions with chemical species of multiple atoms that can form multidentate adsorption configurations at active sites. Furthermore, the chemostructural complexity of practical catalysts with various chemical and structural promoters embraces the opportunity to design site motifs with the desired properties; however, the number of possible combinations in a hypothesized materials space can be prohibitively large even for high-throughput experimentation and/or computation. Taken together, the referred complexity gap between theory and experiment is too large to be bridged by computational techniques traditionally employed because of the limitations of underlying the ab initio simulations in length and timescales. With more accessible supercomputing and characterization facilities, we are often overwhelmed by huge amounts of data that encode rich information about catalysts and catalytic processes. Attributed to the unique capability of recognizing hidden patterns or correlations in high-dimensional data, artificial intelligence (AI) and machine learning offer exciting new directions and have demonstrated a great potential towards bridging the theory–experiment gap in computational heterogeneous catalysis by learning from data[10–14].

In this Review, we discuss recent developments and applications of machine learning to tackle the aforementioned complexities of heterogeneous catalysis, specifically from the aspects of revealing the nature of active sites, unravelling reaction pathways and ultimately accelerating catalyst discovery (Fig. 1). Although highly promising to bridge the theory–experiment gap in complexity with rapidly evolving AI technologies, the prospects and challenges of machine learning for the automated design of sustainable catalytic technologies are introduced. Implementing a data-centric ecosystem that coevolves with computational and data sciences is essential, and needs cooperative community efforts to build on best practices[15] and lay a solid foundation for future growth.

Revealing the nature of active sites

Computational modelling of active sites at highly complex, heterogeneous catalytic interfaces is currently hindered by a few limiting factors, which include the accuracy–efficiency trade-off in describing the exchange–correlation effects of many-electron systems, the near energetic degeneracy of structurally distinct ensembles and the accessible length and timescales of ab initio simulations. We discuss machine learning algorithms in alleviating some of these issues to capture the dynamic evolution of active sites under experimentally relevant conditions.

Atomistic thermodynamics

An active site can span across a broad configurational and compositional space, which is fundamentally governed by experimental conditions, such as temperature, pressure and adsorbate coverage or collectively the chemical potential(s) of interacting species. Ab initio atomistic thermodynamics is widely used to study the equilibrium behaviour of materials, for example, solid surfaces, under certain conditions with quantum chemistry and statistical mechanics[16]. It is a computationally expensive task to find energetically stable structures, that is, global and local minima of multidimensional potential energy surfaces, given the near energetic degeneracy of structurally distinct ensembles. To tackle these issues, machine learning interatomic potentials (MLIPs)[17–22] were actively developed to predict energies (and forces) from atomistic structures using highly non-linear regression algorithms, for example, high-dimensional neural networks. Broadly categorized as supervised machine learning that learns a target function by training on data with ground truth labels, the mathematical mapping allows for the generation of highly accurate and scalable energy landscapes that can be explored via enhanced sampling techniques[4,7], for example, genetic algorithms, basin hopping and Monte Carlo (Fig. 2a).

Large-scale Monte Carlo simulations enabled by MLIPs showed how the surface structure and composition of CuAu nanoparticles change as a function of size[23], and depict the complexity of active site

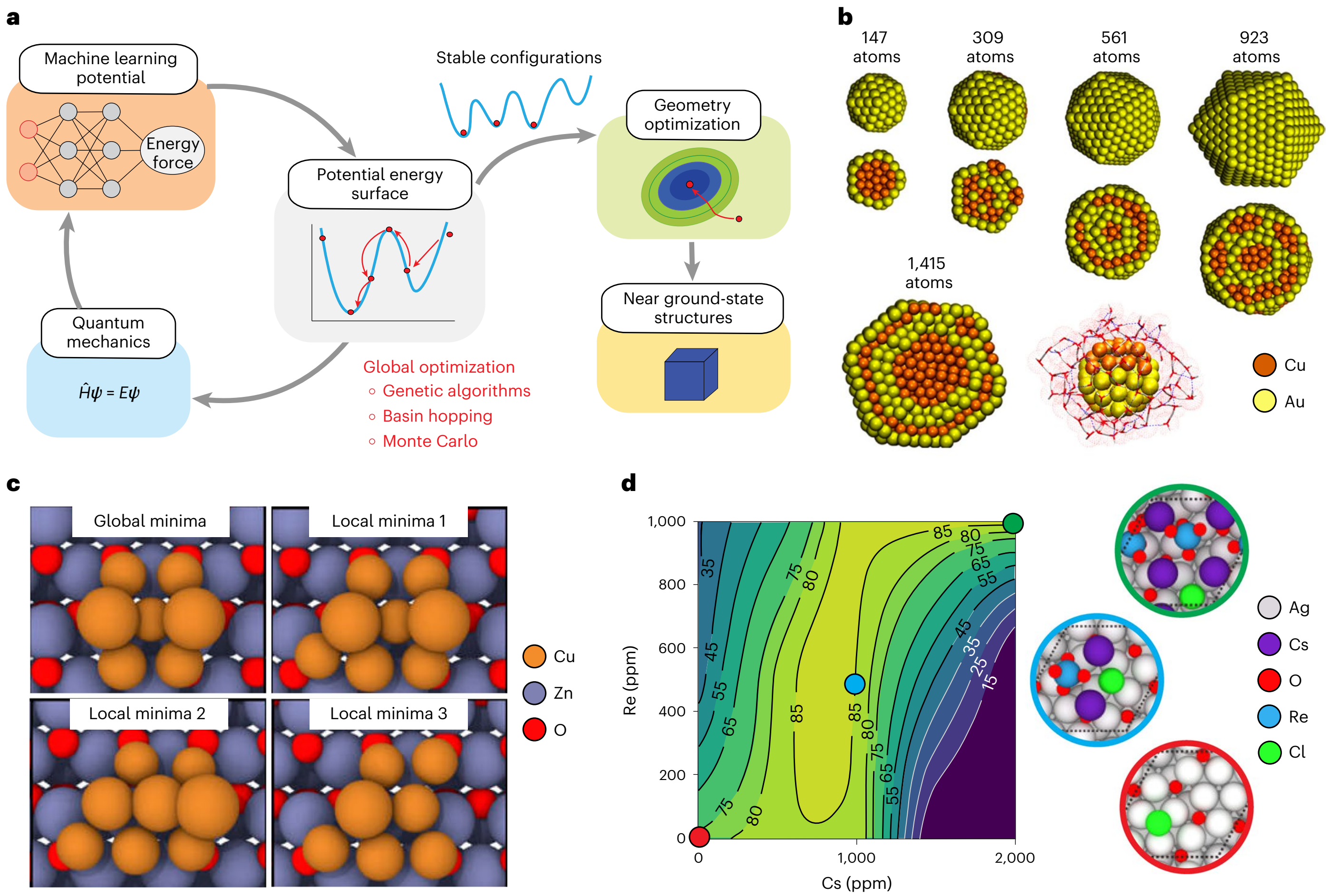

Fig. 2 | Data-enhanced atomistic thermodynamics for exploring configurational spaces. a, Workflow showing how MLIPs are iteratively trained from DFT-calculated structures and utilized in advanced sampling of the surrogate potential energy surface for stable structures. **b**, Optimized CuAu nanoparticles of different sizes via MLIP-enabled Monte Carlo simulations. **c**, Global and local minima structures of Cu_{10} clusters on ZnO(10$\bar{1}$0) optimized via genetic algorithms with machine-learning potentials. **d**, Experimental heat map showing the ethylene epoxide selectivity as a function of Re and Cs concentrations. Circle markers on the heat map indicate various combinations of Re and Cs concentrations for which the global minima structure of Re and Cs on Ag was optimized via simulated annealing with machine-learning potentials. The corresponding global minima structure for each coloured marker is illustrated on the right, where the coloured outline around the structure indicates the corresponding marker on the heat map. Panels **a** and **d**, adapted with permission from ref. [35], American Chemical Society. Panels reproduced with permission from: **b**, ref. [23], American Chemical Society; **c**, ref. [31], AIP.

ensembles on nanoscale systems under aqueous solvation (Fig. 2b). The approach was also used for the CuPdAg system to generate segregation profiles and surface phase diagrams at elevated temperatures[24]. Another important factor is the presence of adsorbed species that can lead to adsorbate-induced surface segregation and reconstruction, by compensating for unfavourable surface configurations via chemisorption. This was demonstrated using Monte Carlo simulations with MLIPs for acrolein adsorption on AgPd alloys, in which acrolein induces the formation of Pd dimers within a Ag host[25]. In the case of acetylene semihydrogenation on PdAg, hydrogen segregates Pd atoms to the surface to form various Pd ensembles, which include dimers, lines and layers with unique reactivity properties[26]. Global optimization with a stochastic surface walking (SSW) algorithm and neural network potentials was rigorously performed to construct a phase diagram of Zn–Cr–O systems, which revealed a stable composition island with a four-coordinated planar Cr^{2+} cation site responsible for the activity and selectivity of syngas (CO/H_2) conversion to give methanol[27]. Reinforcement learning, as another category of machine learning in which a computer agent learns to perform a task through rewarded trial-and-error interactions with an environment, was used to probe surface segregation and its kinetic pathways in NiPdAu alloys powered by MLIPs trained with energetics from effective medium theory[28].

In search of reactive site motifs of atomically dispersed metal catalysts[29], it becomes important to find metastable structures due to their near energetic degeneracy. For a freestanding Pt_{13} cluster, the exploration of the surrogate potential energy surface represented by MLIPs via genetic algorithms came across an ensemble of low-energy metastable structures of hydrogen-covered clusters active for hydrogen evolution and methane activation[30]. The incorporation of supports can be realized via expanding the training data for MLIPs, which thus allows for the modelling of active sites at complex interfaces. Paleico and Behler[31] studied 4–10 atom copper clusters on a ZnO support via Cu–Zn–O MLIPs and genetic algorithms. They were able to find global and local minima structures, with an example shown in Fig. 2c for Cu_{10} clusters. Interestingly, adsorbates such as $*CO$ and $*C_2H_4$ (the asterisk indicates the species is adsorbed on the surface) can drastically change the surface structure of supported metal nanoclusters, flattening them as shown for Pt_{13} on MgO (ref. [32]). A similar effect was observed for $*CO$ on CeO_2-supported Pd_n ($n = 1$–55) (ref. [33]), in which Pd atoms prefer flat structures for small clusters and layered pyramids for large ones.

Revealing adsorbate structures in response to a dynamic change of environmental conditions is computationally challenging due to the large number of possible configurations, especially for molecular adsorbates that can take multidentate adsorption geometries.

Similar to supported nanoclusters, minute energy differences between many adsorbate configurations require sampling of the local and global minima. Basin-hopping Monte Carlo simulations driven by MLIPs showed that at a reasonable coverage of *CO on Pt(553) there can be many kinetically relevant *CO ensembles[34]. Additional complexities arise when promoters and spectator species are considered in catalyst formulations. For example, vinyl chloride and alkali promoters (Cs and Re) are commonly used on Ag catalysts for ethylene epoxidation. Global optimization with machine-learned potentials[35] of O–Cs–Re–Cl systems showed the likely active sites at different Re and Cs concentrations (Fig. 2d), and highlighted the unique roles of both promoters, particularly the formation of ReO_4 clusters, in modulating surface sites for an enhanced selectivity. Beyond transition metal catalysis, machine-learned energetics of the structural configurations of metal oxide surfaces are used for surface Pourbaix diagrams, and provide insights into the nature of active sites towards oxygen evolution[36].

Molecular dynamics

The section above highlights the challenges in modelling active sites under reaction conditions and approaches the problem by finding thermodynamically stable structures. Another important aspect of the dynamic complexity is the real-time evolution of active sites or site ensembles. Ab initio molecular dynamics can be an enabling technique for this purpose. However, it is constrained to pico- to nanosecond timescales and small system sizes due to the formidable computational cost. These restrictions make it difficult to observe surface evolution processes that require long timescales, such as segregation, aggregation and dissolution. In addition, the limited system size prevents the direct modelling of important dynamic scenarios, such as long-range solvation interactions, grain boundaries and complex interfaces. By traversing surrogate potential energy surfaces, extended time and length scales of atomistic simulations can be reached within the framework of machine learning molecular dynamics (MLMD) (Fig. 3a).

Surface structures under dynamic conditions can undergo reconstruction through various elementary steps. This is illustrated in Fig. 3a–c for a Pd monolayer on Ag(111) (ref. [37]). The observation of such phenomena requires high-fidelity atomistic simulations beyond nanoseconds, which necessitates the use of machine-learning potentials. On annealing in a vacuum, the Pd layer is encapsulated by Ag atoms to form isolated Pd atoms. Trajectory analysis classifies several events of dynamic evolution, for example, direct exchange, pop-out and hopping ascent. A similar in situ restructuring was observed for Pd/Au(111) (ref. [38]), on which a subsequent exposure of 0.1 mbar CO enables the Pd monomers to repopulate the surface up to 373 K. Of great interest is the atomistic mechanisms of alloy formation and evolution under dynamic conditions. For example, MLMD simulations of CuZn systems showed that Zn initially alloys near step edges via vacancy generation and direct exchange, but it takes extended timescales (>6 µs) to propagate to terrace regions[39]. Another study focused on the identification of active sites of CuZn alloy nanoparticles for CO_2 electroreduction with SSW powered by MLIPs. Both Cu-heavy CuZn sites and Zn-heavy CuZn sites were found to be stable in dynamic simulations up to 1 ns and could facilitate C–C coupling towards C_{2+} products[40]. In the case of oxide-derived Cu, MLMD simulations generate surface configurations consistent with in situ X-ray absorption spectroscopy experiments when the Cu_2O undergoes a reduction process[41] to form the (100)-like surfaces active towards C_2 product formation (Fig. 3d). Further analysis showed that these sites can be classified into those that favour alcohol (step square sites (s-sq)) and ethylene (planar square (p-sq) and convex square (c-sq)) sites, as depicted in Fig. 3e. Such insights are used to tune the CO_2 electroreduction selectivity by designing specific site motifs (Fig. 3f).

Beyond site structure and composition, another knob of tuning surface reactivity is through the environment that surrounds an active site. Specifically, in electrocatalytic and photoelectrocatalytic applications, solvation can play an important role in dictating the catalytic outcome of active sites. Using Ab initio molecular dynamic simulations, it can be challenging to properly sample all the relevant solvation configurations. Thus, the development of accurate MLIPs was pursued as a way to properly study these complex interfaces[42,43]. For example, the water structure over Pt(111) was interrogated with MLMD simulations[44], which showed a bilayer structure with strongly bound water molecules that form hydrogen bonds with a layer of weakly bound water molecules. This is in contrast with the hexagonal ice structure often used in DFT calculations. Such a disparate water solvation environment can lead to different adsorption energies of key adsorbates. OH adsorption becomes more exothermic in such a water bilayer, especially at higher *OH coverages, as shown from the average adsorption energy profiles[45]. As another example, MLMD simulations of *H on Pt(111) at the aqueous interface showed the formation of *H and *H_2O patches, which resulted in different active sites of hydrogen evolution at high and low *H coverage regions[46].

Compared with ab initio methods, MLIP-powered sampling techniques speed up the exploration of active sites by a few orders of magnitude, which enables the identification of thermodynamically relevant structures. Similarly, MLMD is poised to reveal the nature of active sites under operating conditions by visualizing dynamic processes of site evolution in real time. However, there are still challenges that need to be resolved, which include the number of elements that can be trained at once, accessible timescales beyond microseconds and inaccurate force predictions of out-of-sample configurations that cause unstable dynamic trajectories. Accelerating atomistic simulations towards further extended length and timescales without distorting the intrinsic dynamics holds the key to unlocking its full potential in computational heterogeneous catalysis.

Unravelling reaction pathways

To understand the kinetics of heterogeneous catalytic reactions at given active sites, it is important to unravel reaction pathways from which the rate-limiting factors can be extracted to guide catalyst discovery. For simple gas phase reactions, it is possible to write down elementary steps and investigate their energetics from quantum mechanics. However, this becomes rather cumbersome for complex surface reactions, particularly when the continuum environment (for example, solvation) is included. In this section, we discuss data-enhanced computational methods to automatically generate reaction networks and narrow down possible reaction pathways while we perform rigorous microkinetic simulations[47].

Graph theoretical enumeration

The widely used computational methodology in computational heterogeneous catalysis is to first study possible intermediates and elementary steps via DFT calculations, and then linear scaling relationships[48] can be employed to reduce the dimensions of reactivity descriptors in the kinetics. As the molecules and surface sites involved become more complex, the number of intermediates and elementary steps increase drastically, which makes explicit DFT calculations of energetics prohibitively expensive. To tackle this challenge, many methods were developed to explore reaction pathways[49–51]. For example, an efficient and flexible representation of chemical species is used in chemical graph theory (Fig. 4a), which defines the atoms in the species as nodes and chemical bonds as edges. Each adsorbate–site complex can be largely described using an adjacency list. This representation keeps the information about element symbol, unpaired electron, formal charge and bond connectivity. With this scheme, Gao et al.[52] developed a reaction mechanism generator using a rate-based algorithm[53]. Goldsmith and West extended this gas-phase reaction mechanism generator for catalysis[54]. The framework includes a database of thermodynamic properties and rate coefficients for known species and reaction steps.

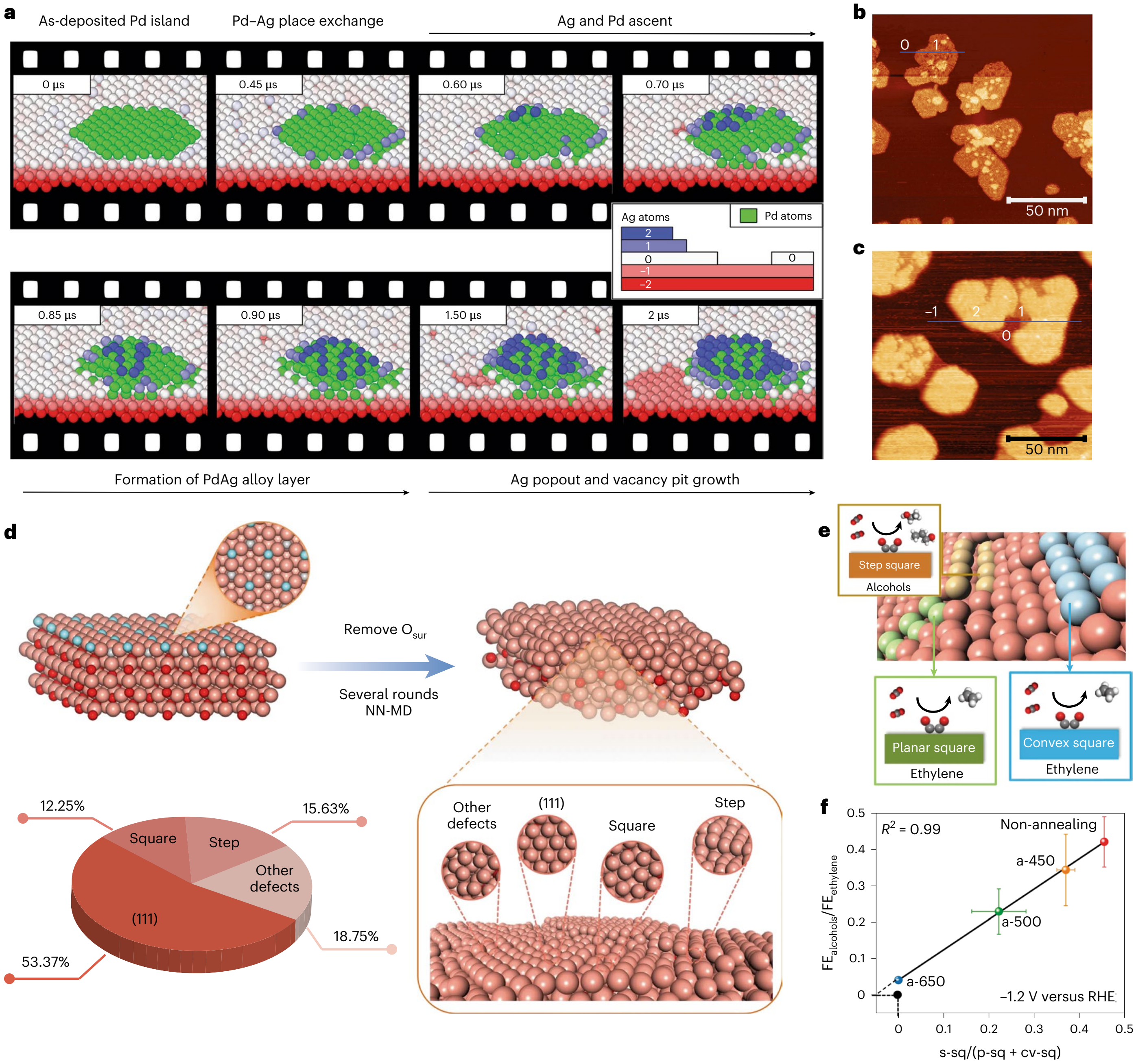

Fig. 3 | MLMD at extended length and timescales. a, MLMD simulations showing the temporal evolution of a Pd_{19} layer on Ag(111). The encapsulation of Pd atoms and vacancy formation can be seen. **b**, A scanning transmission spectroscopy image of Pd islands deposited on Ag(111). **c**, A scanning transmission spectroscopy image of the sample after annealing at 400–450 K showing the encapsulation process. Horizontal blue lines and numbers indicate the layer height profiles: level −1 (subsurface); level 0 (surface); level 1 (deposit); and level 2 (capping). **d**, Simulations of the reduction process on $Cu_2O(111)$ to form ODCu by removing surface oxygen atoms (O_{sur}) and running molecular dynamic simulations with global neural network potential (NN-MD). The various potentially active ensembles on ODCu are depicted. The pie chart shows the fractional distribution of active sites on ODCu. **e**, Site motifs and their preferred products under CO_2 reduction conditions. **f**, The relationship between the ratio of the Faradaic efficiencies (FEs) of alcohols/ethylene and the ratio of s-sq to a sum of p-sq and cv-sq sites. An increase in the s-sq site population is associated with an enhanced alcohol production from CO_2 reduction. a-650, a-500 and a-450 refer to annealing at 650 K, 500 K and 450 K, respectively. RHE, reversed hydrogen electrode. Panel **a** adapted with permission from ref. [37], American Chemical Society. Panels reproduced with permission from: **b**,**c**, ref. [37], American Chemical Society; **d**–**f**, ref. [41] under a Creative Commons license https://creativecommons.org/licenses/by/4.0.

It also includes tools to predict these microkinetic parameters if they are not available. Thus, it is equipped to iteratively generate new reaction pathways and only keep the important ones.

The approach has been used to determine the reaction pathways for catalytic partial oxidation of methane and the concentration profiles of chemical species in a plug flow reactor[54,55] (Fig. 4b). The process fully explores the effects of reaction temperatures, pressures, CH_4/CO_2 ratios and catalysts on kinetic parameters, which identify the optimum variables with the trade-off between H_2 yield and CO_2 reduction[56]. With the graph representation, the most likely reaction pathways for syngas reactions on Rh(111) were theoretically explored by replacing computation-demanding DFT for the reaction energies and barriers

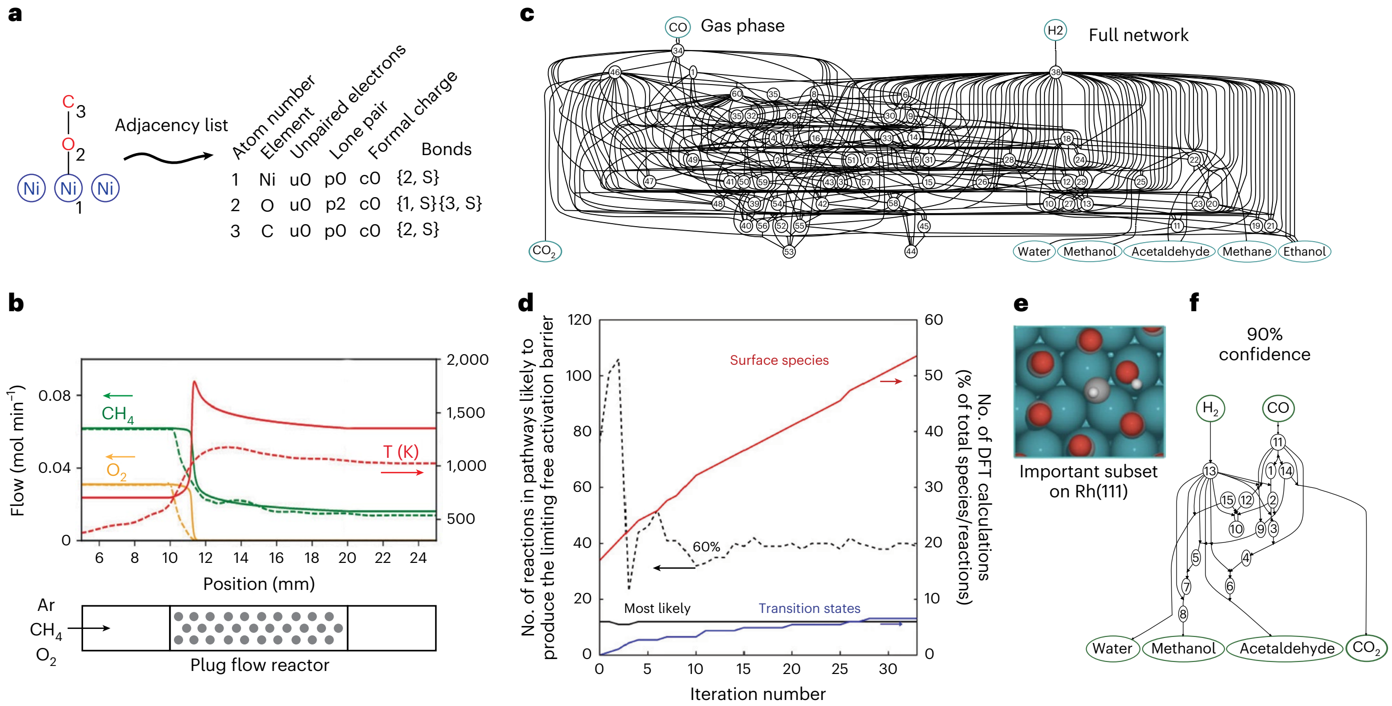

Fig. 4 | Graph theoretical enumeration of reaction pathways with machine learning. a, A graph representation of chemical species[54], in which an adjacency list for CO adsorbed on Ni is shown. Each atom has a list that includes the atom number, element, unpaired electron, lone pair, formal charge and bonds. In this example, it is a single (S) bond between atoms 1 and 2, and atoms 2 and 3. **b**, Automatic mechanism generation of methane-selective oxidation on metal surfaces along with microkinetic models (solid lines) of a plug flow reactor compared with the experimental data (dashed lines) of chemical species and temperature profiles[55]. **c**, A reaction network of syngas to CO_2, water, methane, methanol, acetaldehyde and ethanol from graph theory[57]. Circled numbers are nodes that represent reaction intermediates, whereas lines that connect two nodes are elementary steps from one intermediate to another. **d**, Convergence of the reaction network at each iteration of the process shown in **c**[57]. DFT calculations are performed for important intermediates and transition states at each iteration to improve the performance. **e**, An important subset on Rh(111) showing the scission of CHOH to CH and OH intermediates as the rate-limiting step of ethanol production[57]. **f**, Methanol is still a final product at 90% confidence level given the DFT-level uncertainty[57]. Nodes and lines are working in the same way as those in **c**, but with different adsorbates. Panels adapted with permission from: **a**, ref. [54], American Chemical Society; **b**, ref. [55], American Chemical Society. Panels **c**–**f** reproduced with permission from ref. [57] under a Creative Commons license CC BY 4.0.

with inexpensive machine learning and linear scaling relationships (Fig. 4c). Only after determining the most important steps via classification were self-consistent DFT calculations carried out for energetics, which in turn refine the machine learning models. In this way, one can iteratively reduce the uncertainty and increase the accuracy of the reaction network towards convergence (Fig. 4d–f)[57]. Margraf and Reuter devised a hierarchy algorithm for all the elementary reactions within a chemical subspace[51]. This graph-based method enumerated all the possible reactions by considering bond-breaking reactions as well as methyl and hydroxyl group additions, amenable to constructing reaction networks for a given reaction system that contains C, H and O atoms with up to four non-hydrogen atoms. The ab initio data of such systems should be valuable to train and validate machine learning models of reactivity properties for complex adsorbates, a necessity towards catalysis of high-value chemicals, which include sizeable biomass derivatives.

Environment-aware microkinetics

To unravel kinetically significant reaction pathways, reaction network exploration can be integrated with environment-aware microkinetics (Fig. 5a). In this aspect, microkinetic approaches, which include mean-field microkinetic modelling and kinetic Monte Carlo (KMC) simulations are commonly used. Generally, the active site in a microkinetic model is assumed to be the ground-state structure. Machine learning with compressed sensing algorithms allows for an increased complexity by considering all the high-symmetry sites of a stepped metal surface in CO methanation[58]. An important aspect of microkinetic modelling is the consideration of adsorbate–adsorbate interactions. For mean-field microkinetic modelling, lateral interactions can be considered by modifying the energetics of the intermediates and transition states to be coverage dependent with simple linear relationships or highly non-linear machine learning models[59]. A kinetics-guided pathway search with machine learning was developed to resolve a complex reaction network while considering the coverage of surface intermediates[60]. The key feature of the approach is the automatic identification of kinetically favourable pathways via on-the-fly microkinetic modelling. The reaction sampling is performed using a SSW algorithm enabled by neural network potentials. Low-energy pathways of CO and CO_2 hydrogenation on Cu(211) from SSW simulations were projected onto the collective variables in reduced dimensions, and showed formate and formyl pathways for $^*CO_2$ and *CO hydrogenation, respectively (Fig. 5b)[61]. Capturing bond breaking and/or making while considering the environmental conditions, such as solvation, is challenging even with advanced molecular dynamics sampling techniques. In this aspect, MLIPs were used to study surface reactions, such as CO_2 dissociation on Pt(111) (ref. [62]), N_2 dissociation on Ru(0001) (ref. [63]) and water adsorption and/or dissociation on Pt(110) (ref. [64]). Particularly, Rice et al.[46] used MLMD with enhanced sampling techniques to study the hydrogen evolution reaction at the water–Pt(111) interface. Besides the surface configurations of adsorbates that are part of the active site ensembles, important mechanistic pathways can be unravelled. Specifically, at high coverages of *H the Volmer–Tafel mechanism is favoured, whereas the Volmer–Heyrovsky mechanism is dominant at lower *H coverages. Calegari Andrade et al.[65] studied water dissociation on TiO_2(101) and

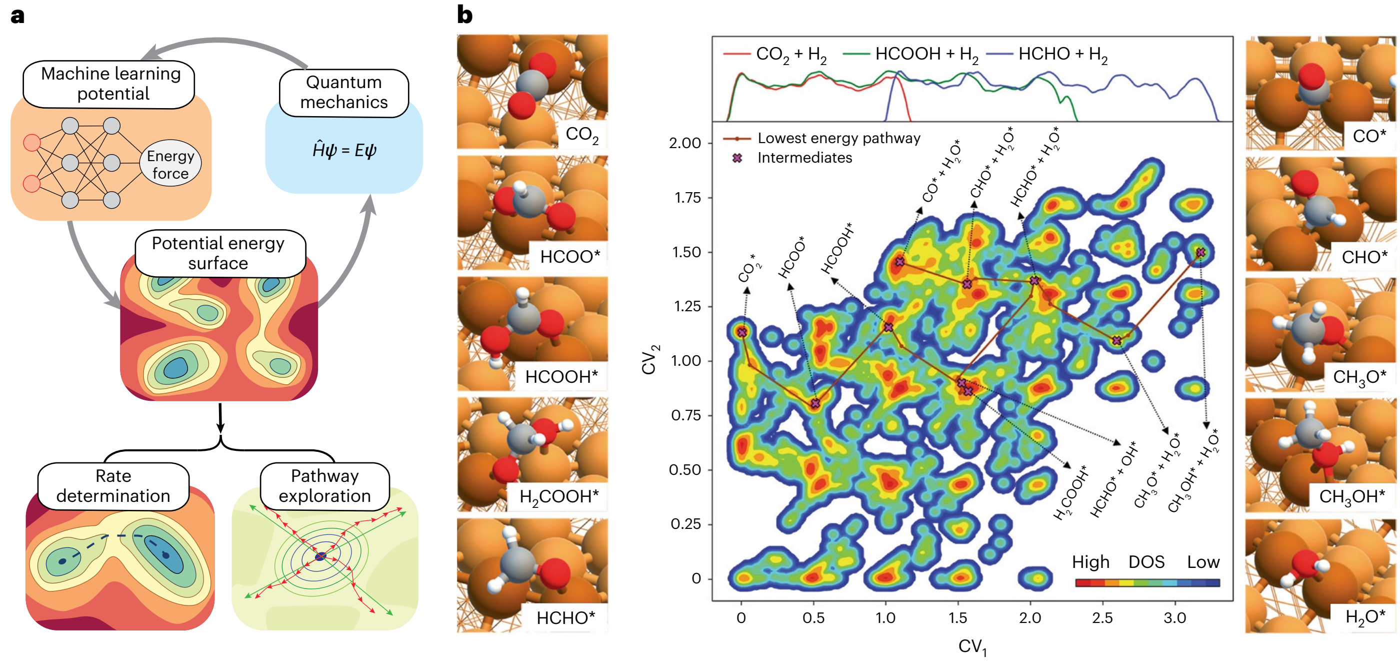

Fig. 5 | Environment-aware microkinetics enabled by machine learning. **a**, Reaction pathway prediction involves the reaction space exploration and rate determination between minima, both of which can be accelerated by machine learning. **b**, Contour plot of 14,958 reaction pairs in methanol synthesis from a CO_2–CO mixture from the low-energy pathways obtained by a microkinetics-guided machine learning pathway search on Cu(211). The intermediates and their optimized geometries are shown. Structures of the coordination function-based collective variable (CV_1 and CV_2) are used in the reduced space to distinguish the states[61]. C, grey; H, white; O, red; terrace Cu, yellow; step-edge Cu, brown. Panel **b** reproduced with permission from ref. [61], American Chemical Society.

calculated the kinetic and thermodynamic parameters via MLMD simulations. Detailed trajectory analysis showed that the mechanism for water dissociation proceeded through a Grotthuss-like mechanism.

KMC simulations would be ideal for capturing complex reaction processes at dynamic surface sites while considering their local environment. However, it is very expensive to enumerate the energetics of all the elementary steps in real-time dynamics for complex systems. With artificial neural networks trained on barrier data, KMC simulations were performed in modelling diffusion processes on the low-index surfaces of copper, and predicted thermodynamically stable surfaces[66]. Within this framework, environment-aware lateral interactions can be integrated explicitly through cluster expansion Hamiltonians or machine learning on a lattice model[59,67]. For example, recent studies using graph neural networks[68] showed that the complex adsorbate–adsorbate interactions can be predicted with a high accuracy by learning from ab initio data. The integration of deep neural networks with KMC and stochastic sampling algorithms for the autonomous exploration of elementary reaction steps is necessary to reveal the full complexity of surface reactions with data-enhanced microkinetic modelling.

Accelerating catalyst discovery

Historical attempts to design improved catalysts largely rely on the Edisonian trial-and-error approach. This strategy might result in suboptimal findings for simple catalytic systems, but is inefficient when searching for high-performance materials of multiple elements and hierarchical structures. In the past few decades, catalyst design has progressively advanced with computation, albeit eclipsed by the chemostructural complexity of practical catalysts. We summarize recent efforts in bridging this theory–experiment gap using machine learning[69] by discussing emergent strategies that actively explore the catalytic materials space with machine learning, formulate design rules by learning from data and optimize catalyst selection within an automated Bayesian framework.

Active machine learning

Although machine learning models efficiently predict catalytic properties once trained, the generation of valuable data for training hinders the design process, largely due to the chemostructural complexity of heterogeneous catalysts. So, it is common that there is limited training data for machine learning algorithms to build on at the beginning. Consequently, the initial model is inadequate to describe the entire design space on an equal footing. In such instances, active learning is useful to iteratively sample the design space, collect additional training data and refine model predictions. This active learning workflow allows the algorithm to identify the most beneficial data to collect and learn from, which leads to a reduced need for data to achieve the same or even improved accuracy compared with that of passive learning.

Active learning works in a closed loop (Fig. 6a), which can be fully automated. Data labelling is a precondition based on the underlying domain problem that ranks candidates with metrics. The often-used metrics of catalytic performance can be described by volcano plots, which use low-dimension reactivity descriptors, for example, the adsorption energies of the reaction intermediates[70,71]. High activity regions indicate the desired descriptor values of active sites. For instance, CO_2 electroreduction to C_2 products (for example, ethylene) on metal alloys is predicted to have the highest activity when CO adsorption is near −0.67 eV relative to its gas phase (exothermic sign convention)[72] (Fig. 6b). An automated machine learning pipeline was used to predict CO adsorption properties on a diverse space of intermetallics with a set of handcrafted physical features, for example, site coordination (Fig. 6c,d). Dimensionality reduction with a *t*-SNE (*t*-distributed stochastic neighbour embedding) representation demonstrates that CuAl site motifs stand out as promising candidates (Fig. 6e)[73]. Electrochemical experiments verified the dealloyed nanoporous CuAl catalysts with an enhanced activity towards CO_2 reduction (Fig. 6f). Gaussian process regression[71,74] is widely applied in active learning because of its inherent uncertainty measures, and was used to discover optimal IrO_3 polymorphs with fewer DFT calculations than

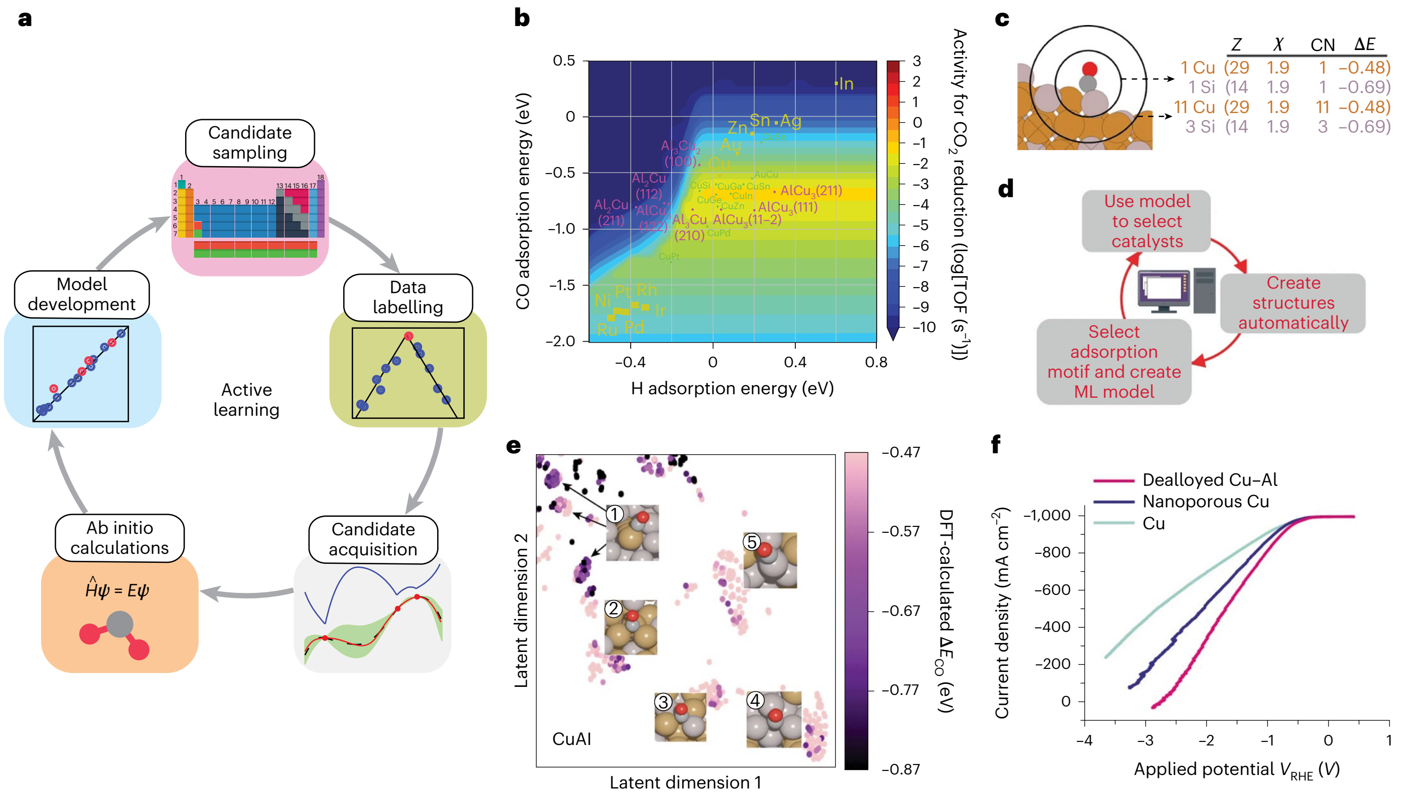

Fig. 6 | Active machine learning for accelerating catalytic materials discovery. a, An active learning workflow, which is composed of candidate space, data labelling, candidate acquisition, ab initio calculations and features and models. **b**, A two-dimensional activity volcano plot for CO_2 electroreduction[72]. TOF, turnover frequency. **c**, Featurization of active sites[70]. Numerical representations include the atomic number (Z), Pauling electronegativity (χ), coordination number (CN) and the adsorption energy at the pure metal of an active site (ΔE). **d**, An automated model selection framework of active learning and surrogate-based optimization[70]. **e**, t-SNE representation of approximately 4,000 adsorption sites on Cu alloys[72]; different point colours represent different DFT-calculated CO adsorption energies (ΔE_{CO}). Representative site motifs are labelled from 1 to 5 in the t-SNE diagram. **f**, C_2H_4 production current density versus potential with dealloyed Cu–Al, nanoporous Cu and evaporated Cu catalysts, which suggests the predicted CuAl alloy has an improved activity[72]. Panels reproduced with permission from: **b**,**f**, ref. [72], Springer Nature Limited; **c**,**d**, ref. [70], Springer Nature Limited. Panel **e** adapted with permission from: ref. [72], Springer Nature Limited.

random selection[75]. A similar approach was applied to the screening of other catalytic systems, which included complex metal oxides[71,74], with easily accessible features to reduce the computational demand[76,77]. As data acquisition is an important part of active learning that makes suggestions of next-round selections by balancing the exploitation and exploration, uncertainty quantification of machine learning models is crucial and remains a fundamental challenge. Recent benchmarks and calibrations are important efforts along this direction to ensure the meaningful convergence of active learning[78,79].

Design rules from machine learning

Machine learning paves the way for the fast screening of high-performance catalysts in a large, hypothesized design space. However, the black-box nature of data-driven machine learning models, for example, deep learning, provides little to no insight. Interpretable machine learning[80–83] offers a path towards opening these black boxes by formulating design rules that circumvent the chemostructural complexity and shed light on the direction of catalyst exploration.

Descriptor-based catalyst design provides physical insights, usually by finding important features. Feature engineering is a process that selects and transforms the most relevant variables from the raw data with the domain knowledge. Current workflows of catalyst design take feature engineering as a tool for optimization and combine it with experimentation for candidate validation[84,85]. For instance, the area-specific resistance (ARS) of perovskite oxides as a reactivity descriptor for oxygen reduction can be predicted by machine learning models with nine readily accessible descriptors[86]. Post hoc analysis of trained deep neural networks provides insights by ranking the importance of each descriptor, which showed the polarization of ionic Lewis acid strengths (ISAs) across metal cations as a key factor (Fig. 7a). This machine-learned design rule sheds light on the fundamental O_2 activation mechanism on perovskites, and leads to an accelerated discovery of improved electrocatalysts with decreased A-site and increased B-site ISAs (Fig. 7b,c). Design rules for stable single atom catalysts on oxide supports were also formulated by learning from data[87,88]. Meanwhile, when there are an enormous number of features to choose from, the compressed sensing method SISSO[89] (sure independence screening and sparsifying operator) provides a suitable solution, which constructs composite descriptors by applying algebraic and/or functional operators to primary features. New descriptors identified by SISSO are more generalizable, and can be applicable to a huge number of systems, which include single atom catalysts[90], perovskite oxides and halides[91], and doped transition metal oxides[92,93]. However, there are drawbacks that limit its applications, which include nearly degenerate models, stability issues on data perturbations and obscure physical interpretations.

In terms of formulating design rules, breaking the adsorption-energy scaling relationships has been a long-lasting effort in catalysis[94,95]. Tuning electronic and geometric descriptors was shown to be effective in tailoring surface reactivity, but remains limited due to the ubiquitous energy-scaling relations[48]. By employing interpretable machine learning, non-scaling behaviour was realized on (100)-type

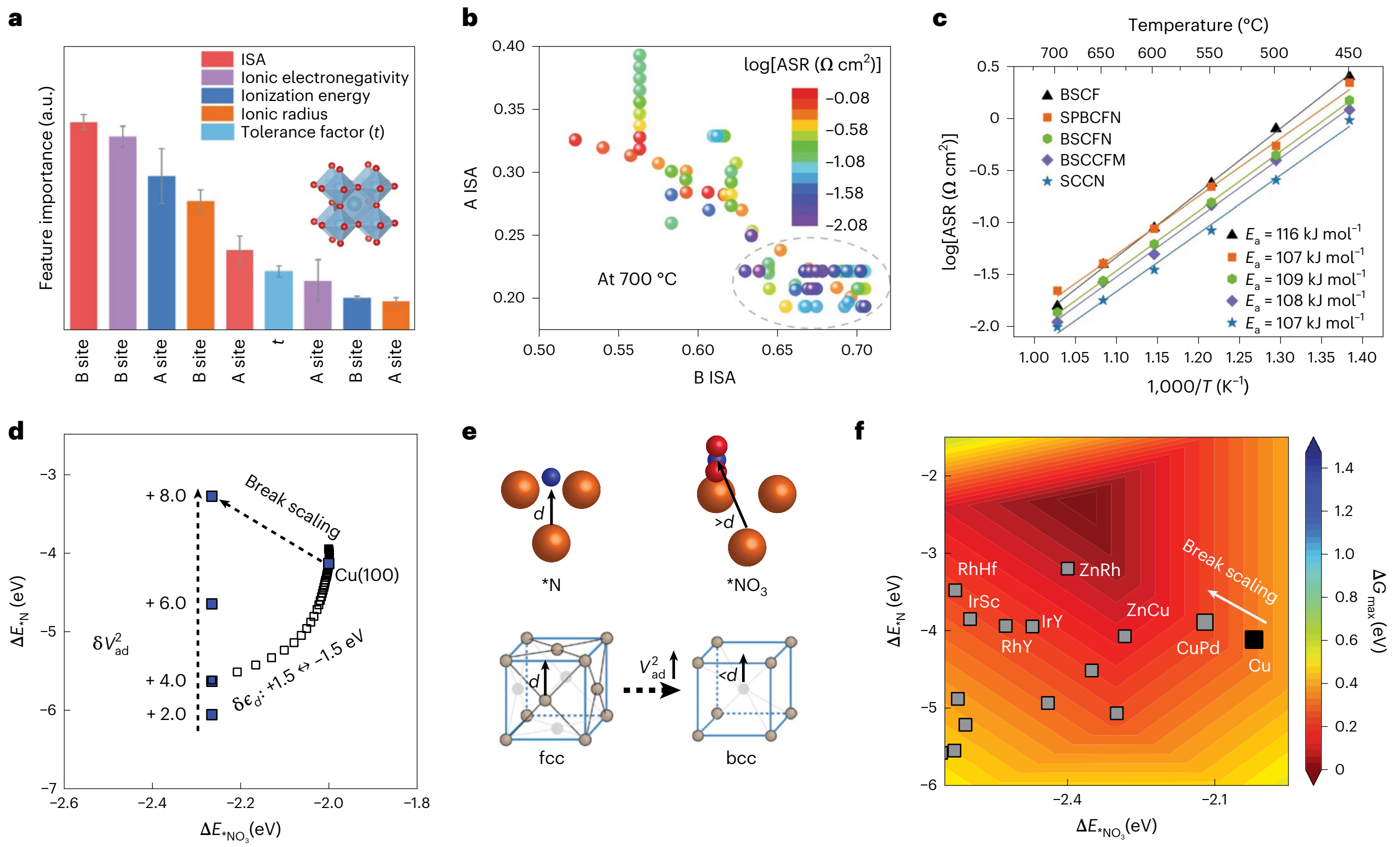

Fig. 7 | Design rules from interpretable machine learning. **a**, Feature importance of a set of ionic descriptors from post hoc analysis of the machine learning models[86]. The ISA is shown as the most important factor. The crystal structure of a perovskite is shown. **b**, Intrinsic oxygen reduction reaction activity characterized by the area specific resistance (ASR) as a function of the ISA values of the A and B sites[86]. **c**, An Arrhenius-type plot of the oxygen reduction activity of selected perovskite oxides[86]. The solid lines represent the least-squares fitting. E_a represents the activation energy. The five selected perovskites are: $Ba_{0.5}Sr_{0.5}Co_{0.8}Fe_{0.2}O_{3-\delta}$ (BSCF, where δ is the oxygen non-stoichiometry), $Sr_{0.6}Ba_{0.2}Pr_{0.2}Co_{0.6}Fe_{0.3}Nb_{0.1}O_3$ (SPBCFN), $Ba_{0.8}Sr_{0.2}Co_{0.6}Fe_{0.2}Nb_{0.2}O_3$ (BSCFN), $Ba_{0.4}Sr_{0.4}Cs_{0.2}Co_{0.6}Fe_{0.3}Mo_{0.1}O_3$ (BSCCFM) and $Sr_{0.9}Cs_{0.1}Co_{0.9}Nb_{0.1}O_3$ (SCCN). **d**, BayesChem-predicted adsorption energies of $*NO_3$ (ΔE_{*NO_3}) and *N (ΔE_{*N}) on Cu(100) by perturbing the electronic structure of adsorption sites via the *d*-band centre ($\delta\varepsilon_d$) and the coupling between the subsurface atom *d* states and adsorbate states (δV^2_{ad}) (ref. [96]). **e**, Schematic illustration of the phase-induced reduction of surface layer separations for breaking linear adsorption-energy scaling relations. For a face-centred cubic (fcc) crystal, the distance between *N in a fourfold hollow site and the subsurface atom is measured by a distance *d*, whereas a body-centred cubic (bcc or B2) lattice gives a smaller distance (*d*), which thus increases the coupling (V^2_{ad}) between the subsurface metal *d* states and the *N *p* states and enables the breaking of linear scaling[96]. **f**, DFT-calculated adsorption energies of $*NO_3$ and *N on (100)-terminated B2 intermetallics close to the activity volcano top[96]. The activity metric was chosen to be the largest Gibbs free energy change (ΔG_{max}) among all the elementary steps at a given electrode potential. The adsorption-energy scaling is broken from Cu(100) to B2 CuPd(100). Panels reproduced with permission from: **a**–**c**, ref. [86], Springer Nature Limited; **d**,**e**, ref. [96] under a Creative Commons License CC BY 4.0. Panel **f** adapted with permission from ref. [96] under a Creative Commons License CC BY 4.0.

sites of ordered B2 intermetallics, attributed to the phase-induced reduction of surface layer separations that leads to a strong Pauli repulsion with the hollow site *N intermediate (Fig. 7d,e)[96]. The physical insights, for example, that govern factors leading to beyond-scaling relationships were provided by the Bayesian model of chemisorption (BayesChem), which was built on the *d*-band theory of chemisorption and Bayesian optimization (BO) by learning from the adsorption properties of metal surfaces[97]. DFT calculations and the activity volcano plot suggested that B2 CuPd nanocubes exhibit a higher activity than Cu for nitrate reduction to ammonia (Fig. 7f), which was validated by electrochemical measurements.

Unsupervised machine learning, a type of algorithm that learns patterns from unlabelled data, can also address the question of what constitutes an adequate explanation. Esterhuizen et al. utilized principal component analysis to provide low-dimensional and interpretable electronic-structure descriptors of near-surface alloys and their reactivity origin[98]. Subgroup discovery algorithms, which overcome the limitations of global models and can identify subgroup correlations based on local effects, were employed in heterogeneous catalysis for the physical understanding of surface reactivity and adsorption-energy scaling relations[88,99]. The approach demonstrated the breaking of the usual scaling relationships between reaction intermediates for nitrogen reduction on single-atom catalysts[99]. Unsupervised machine learning of the literature data was used for knowledge extraction and to elucidate design rules[100,101], for example, in finding catalysts for the oxidative coupling of methane[102] and NO_x reduction[101].

Pure data-driven models have strong predictive capacity but come with the loss of physical intuition due to the black-box nature of complicated formulations. Extracting meaningful knowledge from black-box machine learning models has proved challenging, as the internal logic is not designed for interpretability. The integration of the predictive capacity of black-box models with the intrinsic explainability of physical-based models in interpretable machine learning offers an alternative to open the black boxes and gain physical understanding[69].

Bayesian optimization

In catalysis, the highly complex relationships between material features and catalytic outcomes often render optimization a major challenge. BO plays an indispensable role when functions are obscure, and the generation and evaluation of data are expensive. The approach starts

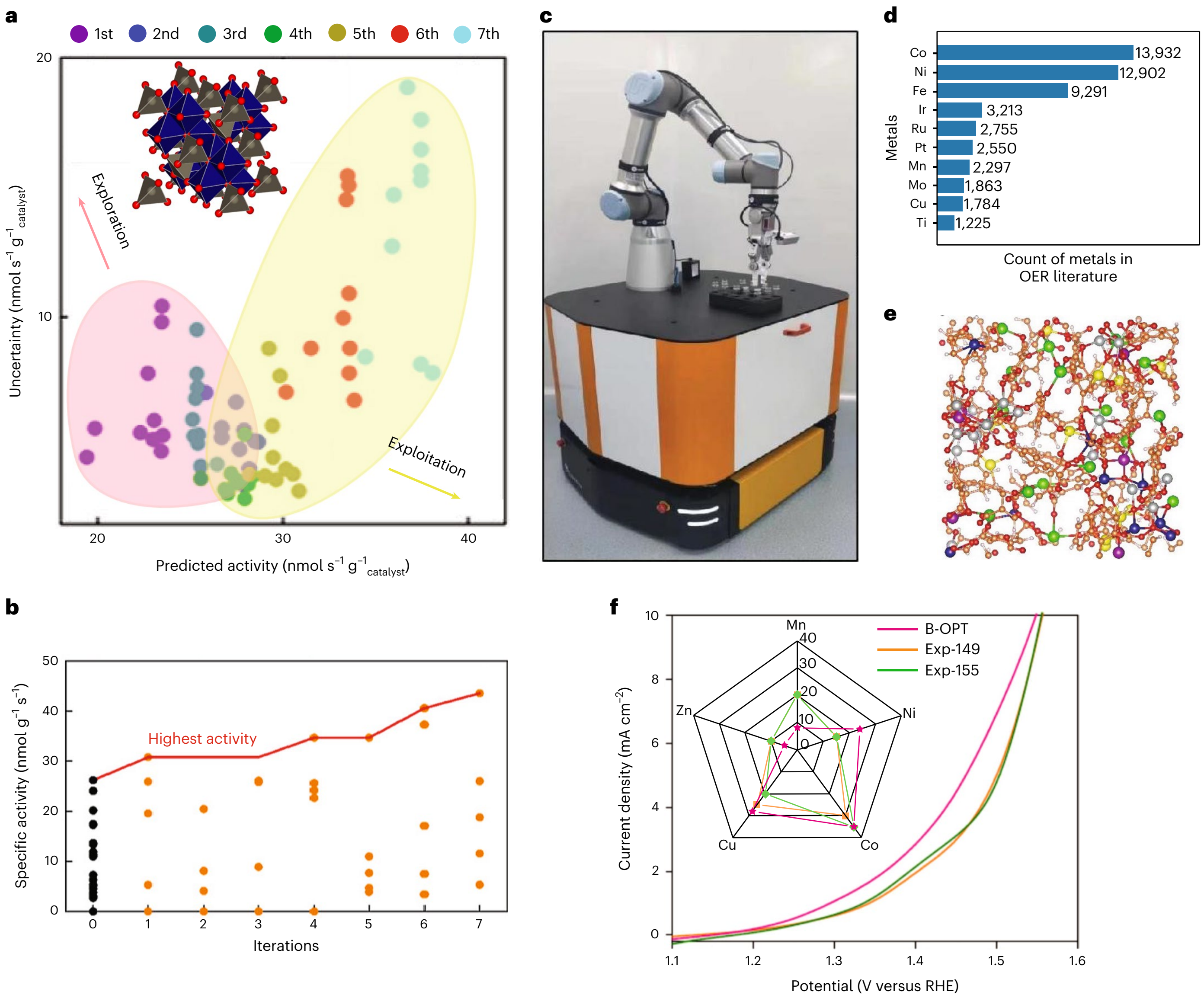

Fig. 8 | BO strategies to find optimal catalysts. a, Predicted specific activity and uncertainty of prediction for each sample over the seven experiment-optimization iterations[103]. The first three iterations in the light pink background focus on exploration, whereas the last four iterations in the light yellow background focus on exploitation. The crystal structure of a spinel oxide is shown. **b**, Iterations of the BO of spinel oxides[103]. Iteration 0 indicates a manual sampling, and iterations 1–7 show the optimization routine. The red line shows the maximum specific activity found over the iterations. **c**, Photograph of the AI chemist[108]. **d**, Frequency of metal co-occurrence provided by machine reading of literature reports[108]. Metals with more than 1,000 counts are shown. OER, oxygen evolution reaction. **e**, An example of simulated structures generated by molecular dynamics[108]. **f**, Current density of the electrochemical oxygen evolution of the BO-suggested sample (B-OPT) and the two best samples from trial-and-error experiments (Exp-149 and Exp-155)[108]. Inset: Kiviat diagram showing the composition ratios of samples. Panel **a** adapted with permission from ref. [103], American Chemical Society. Panels reproduced with permission from: **b**, ref. [103], American Chemical Society; **c**–**f**, ref. [108] under a Creative Commons License CC BY 4.0.

with machine learning surrogate models, for example, gaussian process regression, from the available data, and makes predictions of new data points that can be added into the training data to refine the models. Although bearing similarity to active learning discussed above, the key difference is that the uncertainty captured by the probabilistic model is used to generate an acquisition function, where the optimal point of this function is the set of parameters most likely in increasing the efficiency of optimization. Various acquisition functions are being used to balance the exploitation and exploration of the parameter space, resulting in an iterative improvement to the greatest extent. The Bayesian approach is not focused on the optimization of machine learning surrogate models as in active learning, but on the ability to sample the design space based on available data, aiming for the best achievable rate of convergence in finding optimal solutions.

Nowadays, BO has been routinely used in materials design[103–105]. It has enormous potential to accelerate the predictive discovery of catalyst formulations[104,106], and offers the opportunity for making predictions without the need of physical knowledge[103]. Acquisition functions play a decision-making role for next-round sampling. With limited initial data, the workflow starts from the low-activity region, and the first several iterations focus on exploration to search for high-activity regions, whereas later the optimization shifts towards exploitation to search near the optimal region, which enabled the identification of $(Co_xCu_yM_{1-x-y})_3O_4$ for direct decomposition of nitric oxide into nitrogen[103] (Fig. 8a,b). For propane dehydrogenation to give propylene, a simple acquisition function of decision making was used to find NiMo as a more promising non-precious metal catalyst than Pt, more efficiently than the traditional descriptor-based screening approach[107]. BO was integrated into autonomous catalytic materials discovery by AI chemists (Fig. 8c), who made predictions of the most active electrocatalysts and photocatalysts for oxygen evolution (Fig. 8d–f)[108]. Those AI chemists were designed to read literature to generate a statistical hypothesis and experimental plan, perform continuous experimentations and make selections and predictions with little or no human intervention.

Prospects and challenges

Looking forward, data-enhanced multiscale modelling will be further developed to be an integral tool in computational heterogeneous catalysis to probe active site ensembles and elementary reaction processes under the relevant conditions. Rigorously considering the dynamic and mechanistic complexities in model systems is challenging. In both aspects, developing fast, accurate and self-improving machine learning potentials is the key to enabling high-fidelity atomistic simulations of practical relevance and further deepening our fundamental understanding of heterogeneous catalytic systems. Algorithm advances in terms of the feature representation of the atomic environment beyond the locality approximation[109,110], physics-inspired deep learning architectures[82,111] and advanced data-sampling strategies[112] show promise along this direction.

An improved ab initio description of catalytic processes that explicitly includes solvation, impurities and external energy stimuli at reactive interfaces is needed. Considerable attention to the solvation dynamics at solid–liquid interfaces has been paid in recent years. For example, it was shown that methanol as the solvent can form a reactive hydroxymethyl intermediate at Pd nanoparticles that mediates the proton transfer in the thermocatalytic oxygen reduction towards peroxide formation[113]. In electrocatalytic reactions, the connectivity of hydrogen-bond networks in electric double layers was found to be crucial to understand kinetic pH effects in hydrogen electrocatalysis[114]. As another example, adsorbed alkali metal cations with partial dehydration promote CO_2 electroreduction at metal electrodes[115]. Besides heterogeneous catalysis with thermal, photonic or electrical energy as stimuli, other energy carriers also received intense attention, for example, plasma, surface plasmon and microwave radiation. Recently, it was shown that deep learning with a graph neural network architecture can efficiently predict the electric-field-dependent adsorption energies of surface intermediates in catalytic ammonia synthesis[116].

As machine learning becomes increasingly used in heterogeneous catalytic materials discovery, the demand for interpretability is being considered to ensure that optimal catalysts can be found for the right reason[80,81]. Future applications of machine learning in catalysis have to be aware of the distinction between correlation and causation. Two physical variables with a statistical correlation does not mean one causes the other, and more causal models are needed to ensure the gaining of meaningful knowledge for catalyst design. Accurate machine learning models with physics-informed feature representations or machine-learned high-level features should be used whenever possible to accelerate the catalytic materials discovery, and at the same time to enrich the theory of heterogeneous catalysis by learning insights from data[82–84,117]. One challenge in this endeavour is to incorporate physically tuning parameters in well-established theories, for example, *d*-band theory[97,118], as latent variables in deep learning architectures without sacrificing the accuracy of the model predictions[82].

This Review mainly focuses on bridging the complexity gap in heterogeneous catalysis from a computational perspective. Equally important is applying machine learning tools to enhance experimental data analysis, which includes the microscopic, spectroscopic and kinetic data of catalysts. As demonstrated recently[119–122], machine-learning-assisted X-ray absorption spectroscopy can be used to decipher the active site structures under operando conditions for a broad range of catalytic systems, which include metals[120], alloys[121], metal compounds[122] and zeolites[123]. With ever-growing kinetic data in catalysis, there is an untapped potential to learn directly from those data[124,125]. With enumerated elementary steps in microkinetic modelling, to infer kinetically relevant steps and energetics can be challenging due to data scarcity, and thus requires a probabilistic approach, for example, Bayesian inference. This approach was employed to identify active sites and discern reaction mechanisms[126–129]. Owing to the complex interplay of kinetics at operating conditions, the coverage of intermediates can vary drastically, and thus requires online learning in the iterative Bayesian inference of model parameters.

The field of computational heterogeneous catalysis with machine learning is still in its infancy. There are numerous opportunities to leverage ever-evolving computational and data sciences to advance catalytic materials discovery. An AI and machine learning framework that streamlines data collection with the natural language processing[130,131] of enormous amounts of literature data is highly promising, given the emergent abilities of large language models[132]. Compared with relatively abundant computational data[133–135], the availability of high-quality experimental data in a learnable format is still limited. Recent efforts on rigour and reproducibility in heterogeneous catalysis[136] aim to centralize knowledge and standardized protocols for catalyst synthesis, testing and characterization. Initiatives of this kind might provide a path forward to archive benchmarked, reproducible and abundant experimental data for machine learning. Close collaborations of experimentalists, computational chemists and data scientists hold the key to the sustainable development of a data-centric ecosystem for materials and knowledge discovery in heterogeneous catalysis.

References

1. Chorkendorff, I. & Niemantsverdriet, J. W. *Concepts of Modern Catalysis and Kinetics* (Wiley-VCH, 2007).
2. Nørskov, J. K., Abild-Pedersen, F., Studt, F. & Bligaard, T. Density functional theory in surface chemistry and catalysis. *Proc. Natl Acad. Sci. USA* **108**, 937–943 (2011).
3. Chen, B. W. J., Xu, L. & Mavrikakis, M. Computational methods in heterogeneous catalysis. *Chem. Rev.* **121**, 1007–1048 (2021).
4. Grajciar, L. et al. Towards operando computational modeling in heterogeneous catalysis. *Chem. Soc. Rev.* **47**, 8307–8348 (2018).
5. Nørskov, J. K., Bligaard, T., Rossmeisl, J. & Christensen, C. H. Towards the computational design of solid catalysts. *Nat. Chem.* **1**, 37–46 (2009).
6. Beck, A. et al. Following the structure of copper–zinc–alumina across the pressure gap in carbon dioxide hydrogenation. *Nat. Catal.* **4**, 488–497 (2021).
7. Shi, X. et al. Dynamics of heterogeneous catalytic processes at operando conditions. *J. Am. Chem. Soc. Au* **1**, 2100–2120 (2021).
8. Vogt, C. & Weckhuysen, B. M. The concept of active site in heterogeneous catalysis. *Nat. Rev. Chem.* **6**, 89–111 (2022).
9. Boudart, M. Electronic chemical potential in chemisorption and catalysis. *J. Am. Chem. Soc.* **74**, 1531–1535 (1952).
10. Kitchin, J. R. Machine learning in catalysis. *Nat. Catal.* **1**, 230–232 (2018).
11. Goldsmith, B. R., Esterhuizen, J., Liu, J.-X., Bartel, C. J. & Sutton, C. Machine learning for heterogeneous catalyst design and discovery. *AIChE J.* **64**, 2311–2323 (2018).
12. Schlexer Lamoureux, P. et al. Machine learning for computational heterogeneous catalysis. *ChemCatChem* **11**, 3581–3601 (2019).
13. Medford, A. J., Kunz, M. R., Ewing, S. M., Borders, T. & Fushimi, R. R. Extracting knowledge from data through catalysis informatics. *ACS Catal.* https://doi.org/10.1021/acscatal.8b01708 (2018).
14. Li, H., Jiao, Y., Davey, K. & Qiao, S. Data-driven machine learning for understanding surface structures of heterogeneous catalysts. *Angew. Chem. Int. Ed.* https://doi.org/10.1002/anie.202216383 (2022).
15. Artrith, N. et al. Best practices in machine learning for chemistry. *Nat. Chem.* **13**, 505–508 (2021).

16. Reuter, K. & Scheffler, M. Composition, structure, and stability of RuO_2(110) as a function of oxygen pressure. *Phys. Rev. B* **65**, 035406 (2001).
17. Behler, J. & Parrinello, M. Generalized neural-network representation of high-dimensional potential-energy surfaces. *Phys. Rev. Lett.* **98**, 146401 (2007).
18. Bartók, A. P., Payne, M. C., Kondor, R. & Csányi, G. Gaussian approximation potentials: the accuracy of quantum mechanics, without the electrons. *Phys. Rev. Lett.* **104**, 136403 (2010).
19. 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.* **285**, 316–330 (2015).
20. Shapeev, A. V. Moment tensor potentials: a class of systematically improvable interatomic potentials. *Multiscale Model. Simul.* **14**, 1153–1173 (2016).
21. 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.* **8**, 3192–3203 (2017).
22. Schütt, K. T. et al. SchNet: a continuous-filter convolutional neural network for modeling quantum interactions. *Adv. Neural Inf. Process Syst.* **30**, 992–1002 (2017).
23. Artrith, N. & Kolpak, A. M. Understanding the composition and activity of electrocatalytic nanoalloys in aqueous solvents: a combination of DFT and accurate neural network potentials. *Nano Lett.* **14**, 2670–2676 (2014).
24. Yang, Y., Guo, Z., Gellman, A. J. & Kitchin, J. R. Simulating segregation in a ternary Cu–Pd–Au alloy with density functional theory, machine learning, and Monte Carlo simulations. *J. Phys. Chem. C* **126**, 1800–1808 (2022).
25. Liu, M., Yang, Y. & Kitchin, J. R. Semi-grand canonical Monte Carlo simulation of the acrolein induced surface segregation and aggregation of AgPd with machine learning surrogate models. *J. Chem. Phys.* **154**, 134701 (2021).
26. Li, X.-T., Chen, L., Shang, C. & Liu, Z.-P. In situ surface structures of PdAg catalyst and their influence on acetylene semihydrogenation revealed by machine learning and experiment. *J. Am. Chem. Soc.* **143**, 6281–6292 (2021).
27. 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.* **2**, 671–677 (2019).
28. Yoon, J. et al. Deep reinforcement learning for predicting kinetic pathways to surface reconstruction in a ternary alloy. *Machine Learn. Sci. Technol.* https://doi.org/10.1088/2632-2153 (2021).
29. Chen, Y. et al. A theory-guided X-ray absorption spectroscopy approach for identifying active sites in atomically dispersed transition-metal catalysts. *J. Am. Chem. Soc.* **143**, 20144–20156 (2021).
30. Sun, G. & Sautet, P. Metastable structures in cluster catalysis from first-principles: structural ensemble in reaction conditions and metastability triggered reactivity. *J. Am. Chem. Soc.* **140**, 2812–2820 (2018).
Highlights the importance of low-energy metastable structures of a Pt13 cluster with a modified genetic algorithm driven by machine learning potentials.
31. Paleico, M. L. & Behler, J. Global optimization of copper clusters at the ZnO $(11\bar{1}0)$ surface using a DFT-based neural network potential and genetic algorithms. *J. Chem. Phys.* **153**, 054704 (2020).
32. Kolsbjerg, E. L., Peterson, A. A. & Hammer, B. Neural-network-enhanced evolutionary algorithm applied to supported metal nanoparticles. *Phys. Rev. B* **97**, 195424 (2018).
33. Wang, Y., Su, Y.-Q., Hensen, E. J. M. & Vlachos, D. G. Insights into supported subnanometer catalysts Exposed to CO via Machine-Learning-Enabled Multiscale Modeling. *Chem. Mater.* **34**, 1611–1619 (2022).
34. Sumaria, V. & Sautet, P. CO organization at ambient pressure on stepped Pt surfaces: first principles modeling accelerated by neural networks. *Chem. Sci.* **12**, 15543–15555 (2021).
35. Chen, B. W. J., Wang, B., Sullivan, M. B., Borgna, A. & Zhang, J. Unraveling the synergistic effect of Re and Cs promoters on ethylene epoxidation over silver catalysts with machine learning-accelerated first-principles simulations. *ACS Catal.* **12**, 2540–2551 (2022).
Employs simulated annealing with machine learning potentials to study site structures of promoted Ag(111) with Re, Cs and Cl for ethylene epoxidation.
36. Ulissi, Z. W., Singh, A. R., Tsai, C. & Nørskov, J. K. Automated discovery and construction of surface phase diagrams using machine learning. *J. Phys. Chem. Lett.* **7**, 3931–3935 (2016).
37. Lim, J. S. et al. Evolution of metastable structures at bimetallic surfaces from microscopy and machine-learning molecular dynamics. *J. Am. Chem. Soc.* **142**, 15907–15916 (2020).
Uncovers atomistic processes of the evolution of a Pd layer on Ag(111) with machine learning molecular dynamics.
38. Zhou, C. et al. Dynamical study of adsorbate-induced restructuring kinetics in bimetallic catalysts using the PdAu(111) model system. *J. Am. Chem. Soc.* **144**, 15132–15142 (2022).
39. Halim, H. H. & Morikawa, Y. Elucidation of Cu–Zn surface alloying on Cu(997) by machine-learning molecular dynamics. *ACS Phys. Chem. Au* **2**, 430–447 (2022).
40. Zhen, S. et al. Nature of the active sites of copper zinc catalysts for carbon dioxide electroreduction. *Angew. Chem. Int. Ed.* **61**, e202201913 (2022).
41. Cheng, D. et al. The nature of active sites for carbon dioxide electroreduction over oxide-derived copper catalysts. *Nat. Commun.* **12**, 395 (2021).
Identifies active sites for C–C coupling by machine learning molecular dynamics simulations of reduction processes of oxide-derived copper catalysts.
42. Natarajan, S. K. & Behler, J. Neural network molecular dynamics simulations of solid–liquid interfaces: water at low-index copper surfaces. *Phys. Chem. Chem. Phys.* **18**, 28704–28725 (2016).
43. Quaranta, V., Behler, J. & Hellström, M. Structure and dynamics of the liquid-water/zinc-oxide interface from machine learning potential simulations. *J. Phys. Chem. C* **123**, 1293–1304 (2019).
44. Mikkelsen, A. E. G., Schiøtz, J., Vegge, T. & Jacobsen, K. W. Is the water/Pt(111) interface ordered at room temperature? *J. Chem. Phys.* **155**, 224701 (2021).
45. Mikkelsen, A. E. G. et al. Structure and energetics of liquid water-hydroxyl layers on Pt(111). *Phys. Chem. Chem. Phys.* https://doi.org/10.1039/D2CP00190J (2022).
46. Rice, P. S., Liu, Z.-P. & Hu, P. Hydrogen coupling on platinum using artificial neural network potentials and DFT. *J. Phys. Chem. Lett.* **12**, 10637–10645 (2021).
Employs machine learning molecular dynamics to study active sites of hydrogen evolution at the water/Pt(111) interface.
47. Pablo-García, S., García-Muela, R., Sabadell-Rendó, A. & López, N. Dimensionality reduction of complex reaction networks in heterogeneous catalysis: from linear-scaling relationships to statistical learning technique. *WIREs Comput. Mol. Sci.* **11**, e1540 (2021).
48. Abild-Pedersen, F. et al. Scaling properties of adsorption energies for hydrogen-containing molecules on transition-metal surfaces. *Phys. Rev. Lett.* **99**, 016105 (2007).
49. Rangarajan, S., Bhan, A. & Daoutidis, P. Language-oriented rule-based reaction network generation and analysis: description of RING. *Comput. Chem. Eng.* **45**, 114–123 (2012).
50. Gupta, U. & Vlachos, D. G. Learning chemistry of complex reaction systems via a Python first-principles reaction rule stencil (pReSt) generator. *J. Chem. Inf. Model.* **61**, 3431–3441 (2021).

51. Margraf, J. T. & Reuter, K. Systematic enumeration of elementary reaction steps in surface catalysis. *ACS Omega* **4**, 3370–3379 (2019).

52. Gao, C. W., Allen, J. W., Green, W. H. & West, R. H. Reaction mechanism generator: automatic construction of chemical kinetic mechanisms. *Comput. Phys. Commun.* **203**, 212–225 (2016).

53. Susnow, R. G., Dean, A. M., Green, W. H., Peczak, P. & Broadbelt, L. J. Rate-based construction of kinetic models for complex systems. *J. Phys. Chem. A* **101**, 3731–3740 (1997).

54. Goldsmith, C. F. & West, R. H. Automatic generation of microkinetic mechanisms for heterogeneous catalysis. *J. Phys. Chem. C* **121**, 9970–9981 (2017).

55. Mazeau, E. J., Satpute, P., Blöndal, K., Goldsmith, C. F. & West, R. H. Automated mechanism generation using linear scaling relationships and sensitivity analyses applied to catalytic partial oxidation of methane. *ACS Catal.* **11**, 7114–7125 (2021).

56. Lim, J. Y. et al. Machine learning-assisted CO_2 utilization in the catalytic dry reforming of hydrocarbons: reaction pathways and multicriteria optimization analyses. *Int. J. Energy Res.* **46**, 6277–6291 (2022).

57. Ulissi, Z. W., Medford, A. J., Bligaard, T. & Nørskov, J. K. To address surface reaction network complexity using scaling relations machine learning and DFT calculations. *Nat. Commun.* **8**, 14621 (2017).
Identifies the most likely reaction pathway for syngas conversion on Rh(111) from a multitude of pathways with machine learning.

58. Deimel, M., Reuter, K. & Andersen, M. Active site representation in first-principles microkinetic models: data-enhanced computational screening for improved methanation catalysts. *ACS Catal.* **10**, 13729–13736 (2020).

59. Mou, T., Han, X., Zhu, H. & Xin, H. Machine learning of lateral adsorbate interactions in surface reaction kinetics. *Curr. Opin. Chem. Eng.* **36**, 100825 (2022).

60. Zhang, X.-J. & Liu, Z.-P. Reaction sampling and reactivity prediction using the stochastic surface walking method. *Phys. Chem. Chem. Phys.* **17**, 2757–2769 (2015).

61. Shi, Y.-F., Kang, P.-L., Shang, C. & Liu, Z.-P. Methanol synthesis from CO_2/CO mixture on Cu–Zn catalysts from microkinetics-guided machine learning pathway search. *J. Am. Chem. Soc.* **144**, 13401–13414 (2022).
Uses machine learning interatomic potentials for a surrogate potential energy surface of CO_2 hydrogenation on Cu and finds the most favourable reaction pathways via stochastic surface walking.

62. del Cueto, M. et al. New perspectives on CO_2–Pt(111) interaction with a high-dimensional neural network potential energy surface. *J. Phys. Chem. C* **124**, 5174–5181 (2020).

63. Lee, E. M. Y. et al. Neural network sampling of the free energy landscape for nitrogen dissociation on ruthenium. *J. Phys. Chem. Lett.* **12**, 2954–2962 (2021).

64. Hu, C., Zhang, Y. & Jiang, B. Dynamics of H_2O Adsorption on Pt(110)-(1×2) based on a neural network potential energy surface. *J. Phys. Chem. C* **124**, 23190–23199 (2020).

65. Calegari Andrade, M. F., Ko, H.-Y., Zhang, L., Car, R. & Selloni, A. Free energy of proton transfer at the water–TiO_2 interface from ab initio deep potential molecular dynamics. *Chem. Sci.* **11**, 2335–2341 (2020).

66. Kimari, J. et al. Application of artificial neural networks for rigid lattice kinetic Monte Carlo studies of Cu surface diffusion. *Comput. Mater. Sci.* **183**, 109789 (2020).

67. Vignola, E. et al. A machine learning approach to graph-theoretical cluster expansions of the energy of adsorbate layers. *J. Chem. Phys.* **147**, 054106 (2017).

68. Ghanekar, P. G., Deshpande, S. & Greeley, J. Adsorbate chemical environment-based machine learning framework for heterogeneous catalysis. *Nat. Commun.* **13**, 5788 (2022).

69. Peng, J. et al. Human- and machine-centred designs of molecules and materials for sustainability and decarbonization. *Nat. Rev. Mater.* **7**, 991–1009 (2022).

70. Tran, K. & Ulissi, Z. W. Active learning across intermetallics to guide discovery of electrocatalysts for CO_2 reduction and H_2 evolution. *Nat. Catal.* **1**, 696–703 (2018).

71. Li, Z., Achenie, L. E. K. & Xin, H. An adaptive machine learning strategy for accelerating discovery of perovskite electrocatalysts. *ACS Catal.* **10**, 4377–4384 (2020).

72. Zhong, M. et al. Accelerated discovery of CO_2 electrocatalysts using active machine learning. *Nature* **581**, 178–183 (2020).
Employs an automated machine learning pipeline with active learning to find active catalysts for CO_2 reduction to C_2 products.

73. Van Der, M. Accelerating t-SNE using tree-based algorithm. *J. Mach. Learn. Res* **15**, 3221–3245 (2014).

74. Liu, F., Yang, S. & Medford, A. J. Scalable approach to high coverages on oxides via iterative training of a machine-learning algorithm. *ChemCatChem* **12**, 4317–4330 (2020).

75. Flores, R. A. et al. Active learning accelerated discovery of stable iridium oxide polymorphs for the oxygen evolution reaction. *Chem. Mater.* **32**, 5854–5863 (2020).

76. Noh, J., Back, S., Kim, J. & Jung, Y. Active learning with non-ab initio input features toward efficient CO_2 reduction catalysts. *Chem. Sci.* **9**, 5152–5159 (2018).

77. Liu, X., Cai, C., Zhao, W., Peng, H.-J. & Wang, T. Machine learning-assisted screening of stepped alloy surfaces for C1 catalysis. *ACS Catal.* **12**, 4252–4260 (2022).

78. Hu, Y., Musielewicz, J., Ulissi, Z. & Medford, A. J. Robust and scalable uncertainty estimation with conformal prediction for machine-learned interatomic potentials. *Mach. Learn. Sci. Technol.* **3**, 045028 (2022).

79. Tran, K. et al. Methods for comparing uncertainty quantifications for material property predictions. *Mach. Learn. Sci. Technol.* **1**, 025006 (2020).

80. Esterhuizen, J. A., Goldsmith, B. R. & Linic, S. Interpretable machine learning for knowledge generation in heterogeneous catalysis. *Nat. Catal.* **5**, 175–184 (2022).

81. Omidvar, N. et al. Interpretable machine learning of chemical bonding at solid surfaces. *J. Phys. Chem. Lett.* **12**, 11476–11487 (2021).

82. Wang, S.-H., Pillai, H. S., Wang, S., Achenie, L. E. K. & Xin, H. Infusing theory into deep learning for interpretable reactivity prediction. *Nat. Commun.* **12**, 5288 (2021).

83. Montemore, M. M., Nwaokorie, C. F. & Kayode, G. O. General screening of surface alloys for catalysis. *Catal. Sci. Technol.* **10**, 4467–4476 (2020).

84. Ma, X., Li, Z., Achenie, L. E. K. & Xin, H. Machine-learning-augmented chemisorption model for CO_2 electroreduction catalyst screening. *J. Phys. Chem. Lett.* **6**, 3528–3533 (2015).

85. Sun, Y. et al. Covalency competition dominates the water oxidation structure–activity relationship on spinel oxides. *Nat. Catal.* **3**, 554–563 (2020).

86. Zhai, S. et al. A combined ionic Lewis acid descriptor and machine-learning approach to prediction of efficient oxygen reduction electrodes for ceramic fuel cells. *Nat. Energy* **7**, 866–875 (2022).
Uses machine learning to extract design principles of perovskites for oxygen evolution by learning from experimentally measured area-specific resistances in literature.

87. O'Connor, N. J., Jonayat, A. S. M., Janik, M. J. & Senftle, T. P. Interaction trends between single metal atoms and oxide supports identified with density functional theory and statistical learning. *Nat. Catal.* **1**, 531–539 (2018).
88. Han, Z.-K. et al. Single-atom alloy catalysts designed by first-principles calculations and artificial intelligence. *Nat. Commun.* **12**, 1833 (2021).
89. Ouyang, R., Curtarolo, S., Ahmetcik, E., Scheffler, M. & Ghiringhelli, L. M. SISSO: a compressed-sensing method for identifying the best low-dimensional descriptor in an immensity of offered candidates. *Phys. Rev. Mater.* **2**, 083802 (2018).
90. Andersen, M., Levchenko, S. V., Scheffler, M. & Reuter, K. Beyond scaling relations for the description of catalytic materials. *ACS Catal.* **9**, 2752–2759 (2019).
91. Bartel, C. J. et al. New tolerance factor to predict the stability of perovskite oxides and halides. *Sci. Adv.* **5**, eaav0693 (2019).
92. Xu, W., Andersen, M. & Reuter, K. Data-driven descriptor engineering and refined scaling relations for predicting transition metal oxide reactivity. *ACS Catal.* **11**, 734–742 (2021).
93. Jiang, C. et al. Data-driven interpretable descriptors for the structure–activity relationship of surface lattice oxygen on doped vanadium oxides. *Angew. Chem. Int. Ed.* **61**, e202206758 (2022).
94. Vojvodic, A. & Nørskov, J. K. New design paradigm for heterogeneous catalysts. *Natl Sci. Rev.* **2**, 140–143 (2015).
95. Pérez-Ramírez, J. & López, N. Strategies to break linear scaling relationships. *Nat. Catal.* **2**, 971–976 (2019).
96. Gao, Q. et al. Breaking adsorption-energy scaling limitations of electrocatalytic nitrate reduction on intermetallic CuPd nanocubes by machine-learned insights. *Nat. Commun.* **13**, 2338 (2022).
97. Wang, S., Pillai, H. S. & Xin, H. Bayesian learning of chemisorption for bridging the complexity of electronic descriptors. *Nat. Commun.* **11**, 6132 (2020).
98. Esterhuizen, J. A., Goldsmith, B. R. & Linic, S. Uncovering electronic and geometric descriptors of chemical activity for metal alloys and oxides using unsupervised machine learning. *Chem. Catal.* **1**, 923–940 (2021).
99. Li, H. et al. Subgroup discovery points to the prominent role of charge transfer in breaking nitrogen scaling relations at single-atom catalysts on VS_2. *ACS Catal.* **11**, 7906–7914 (2021).
100. Smith, A., Keane, A., Dumesic, J. A., Huber, G. W. & Zavala, V. M. A machine learning framework for the analysis and prediction of catalytic activity from experimental data. *Appl. Catal. B* **263**, 118257 (2020).
101. Chen, Y., Li, R., Suo, H. & Liu, C. Evaluation of a data-driven, machine learning approach for identifying potential candidates for environmental catalysts: from database development to prediction. *ACS EST Eng.* **1**, 1246–1257 (2021).
102. Mine, S. et al. Analysis of updated literature data up to 2019 on the oxidative coupling of methane using an extrapolative machine-learning method to identify novel catalysts. *ChemCatChem* https://doi.org/10.1002/cctc.202100495 (2021).
103. Zhang, Y. et al. Descriptor-free design of multicomponent catalysts. *ACS Catal.* **12**, 10562–10571 (2022).
104. Nguyen, T. N. et al. High-throughput experimentation and catalyst informatics for oxidative coupling of methane. *ACS Catal.* **10**, 921–932 (2020).
105. Batchelor, T. A. A. et al. Complex-solid-solution electrocatalyst discovery by computational prediction and high-throughput experimentation. *Angew. Chem. Int. Ed.* **60**, 6932–6937 (2021).
106. Williams, T., McCullough, K. & Lauterbach, J. A. Enabling catalyst discovery through machine learning and high-throughput experimentation. *Chem. Mater.* **32**, 157–165 (2020).
107. Wang, T. et al. Theory-aided discovery of metallic catalysts for selective propane dehydrogenation to propylene. *ACS Catal.* **11**, 6290–6297 (2021).
108. An all-round AI-chemist with scientific mind. *Natl Sci. Rev.* https://doi.org/10.1093/nsr/nwac190 (2022).
Bayesian optimization aided by AI chemists for catalyst discovery.
109. Artrith, N., Urban, A. & Ceder, G. Efficient and accurate machine-learning interpolation of atomic energies in compositions with many species. *Phys. Rev. B* **96**, 014112 (2017).
110. Behler, J. & Csányi, G. Machine learning potentials for extended systems: a perspective. *Eur. Phys. J. B* **94**, 142 (2021).
111. Pun, G. P. P., Batra, R., Ramprasad, R. & Mishin, Y. Physically informed artificial neural networks for atomistic modeling of materials. *Nat. Commun.* **10**, 2339 (2019).
112. Bussi, G. & Laio, A. Using metadynamics to explore complex free-energy landscapes. *Nat. Rev. Phys.* **2**, 200–212 (2020).
113. Adams, J. S. et al. Solvent molecules form surface redox mediators in situ and cocatalyze O_2 reduction on Pd. *Science* **371**, 626–632 (2021).
114. Li, P. et al. Hydrogen bond network connectivity in the electric double layer dominates the kinetic pH effect in hydrogen electrocatalysis on Pt. *Nat. Catal.* **5**, 900–911 (2022).
115. Ovalle, V. J., Hsu, Y.-S., Agrawal, N., Janik, M. J. & Waegele, M. M. Correlating hydration free energy and specific adsorption of alkali metal cations during CO_2 electroreduction on Au. *Nat. Catal.* **5**, 624–632 (2022).
116. Wan, M., Yue, H., Notarangelo, J., Liu, H. & Che, F. Deep learning-assisted investigation of electric field–dipole effects on catalytic ammonia synthesis. *J. Am. Chem. Soc. Au* **2**, 1338–1349 (2022).
117. Esterhuizen, J. A., Goldsmith, B. R. & Linic, S. Theory-guided machine learning finds geometric structure–property relationships for chemisorption on subsurface alloys. *Chem* **6**, 3100–3117 (2020).
118. Hammer, B. & Nørsko, J. K. Theoretical surface science and catalysis—calculations and concept. *Adv. Catal.* **45**, 71–129 (2000).
119. Sainju, R. et al. DefectTrack: a deep learning-based multi-object tracking algorithm for quantitative defect analysis of in-situ TEM videos in real-time. *Sci. Rep.* **12**, 15705 (2022).
120. Routh, P. K., Liu, Y., Marcella, N., Kozinsky, B. & Frenkel, A. I. Latent representation learning for structural characterization of catalysts. *J. Phys. Chem. Lett.* **12**, 2086–2094 (2021).
121. Marcella, N. et al. Decoding reactive structures in dilute alloy catalysts. *Nat. Commun.* **13**, 832 (2022).
122. Trummer, D. et al. Deciphering the Phillips catalyst by orbital analysis and supervised machine learning from Cr pre-edge XANES of molecular libraries. *J. Am. Chem. Soc.* **143**, 7326–7341 (2021).
123. Martini, A. et al. Assessing the influence of zeolite composition on oxygen-bridged diamino dicopper(II) complexes in Cu–CHA $DeNO_x$ catalysts by machine learning-assisted X-ray absorption spectroscopy. *J. Phys. Chem. Lett.* **13**, 6164–6170 (2022).
124. Kunz, M. R. et al. Data driven reaction mechanism estimation via transient kinetics and machine learning. *Chem. Eng. J.* **420**, 129610 (2021).
125. Chen, K., Tian, H., Li, B. & Rangarajan, S. A chemistry-inspired neural network kinetic model for oxidative coupling of methane from high-throughput data. *AIChE J.* **68**, e17584 (2022).
126. Savara, A. & Walker, E. A. CheKiPEUQ intro 1: Bayesian parameter estimation considering uncertainty or error from both experiments and theory. *ChemCatChem* **12**, 5385–5400 (2020).
127. Cohen, M. & Vlachos, D. G. Chemical kinetics Bayesian inference toolbox (CKBIT). *Comput. Phys. Commun.* **265**, 107989 (2021).

128. Fricke, C., Rajbanshi, B., Walker, E. A., Terejan, G. & Heyden, A. Propane dehydrogenation on platinum catalysts: identifying the active sites through Bayesian analysis. *ACS Catal.* **12**, 2487–2498 (2022).
129. Walker, E. A., Mitchell, D., Terejanu, G. A. & Heyden, A. Identifying active sites of the water–gas shift reaction over titania supported platinum catalysts under uncertainty. *ACS Catal.* **8**, 3990–3998 (2018).
130. Nandy, A., Duan, C. & Kulik, H. J. Using machine learning and data mining to leverage community knowledge for the engineering of stable metal–organic frameworks. *J. Am. Chem. Soc.* **143**, 17535–17547 (2021).
131. Kim, E. et al. Materials synthesis insights from scientific literature via text extraction and machine learning. *Chem. Mater.* **29**, 9436–9444 (2017).
132. Brown, T. B. et al. Language models are few-shot learners. *Adv. Neural Inf. Process Syst.* **33**, 1877–1901 (2020).
133. Winther, K. T. et al. Catalysis-hub.org, an open electronic structure database for surface reactions. *Sci. Data* **6**, 75 (2019).
134. Chanussot, L. et al. Open Catalyst 2020 (OC20) dataset and community challenges. *ACS Catal.* **11**, 6059–6072 (2021).
135. Álvarez-Moreno, M. et al. Managing the computational chemistry big data problem: the ioChem-BD platform. *J. Chem. Inf. Model.* **55**, 95–103 (2015).
136. Esposito, D. Induce to reproduce. *Nat. Catal.* **5**, 658–661 (2022).

Acknowledgements

H.S.P., S.W. and H.X. acknowledge funding support from the NSF CAREER programme (CBET-1845531). T.M. and X.H. acknowledge funding support from the NSF Chemical Catalysis programme (CHE-2102363). H.X. acknowledges the partial support from the US Department of Energy, Office of Basic Energy Sciences under contract no. DE-SC0023323. T.M. especially thanks the NSF Non-Academic Research Internships for Graduate Students (INTERN) programme for supporting his work in the Brookhaven National Lab chemistry division under the guidance of P. Liu. M.W. acknowledges funding support from the NSF EAGER programme (CBET-2103478). F.C. acknowledges the partial sponsorship by Department of Navy award N00014-22-1-2001 issued by the Office of Naval Research. The United States Government has a royalty-free license throughout the world in all copyrightable material contained herein.

Author contributions

H.X., F.C. and N.S. conceived the idea of the Review. T.M., H.S.P. and S.W. led the manuscript writing. All the authors contributed to the revision of this manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence should be addressed to Hongliang Xin.

Peer review information *Nature Catalysis* thanks Bryan Goldsmith and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.