nature catalysis

PERSPECTIVE

https://doi.org/10.1038/s41929-022-00744-z

Interpretable machine learning for knowledge generation in heterogeneous catalysis

Jacques A. Esterhuizen[1,2], Bryan R. Goldsmith[1,2] ✉ and Suljo Linic[1,2] ✉

Most applications of machine learning in heterogeneous catalysis thus far have used black-box models to predict computable physical properties (descriptors), such as adsorption or formation energies, that can be related to catalytic performance (that is, activity or stability). Extracting meaningful physical insights from these black-box models has proved challenging, as the internal logic of these black-box models is not readily interpretable due to their high degree of complexity. Interpretable machine learning methods that merge the predictive capacity of black-box models with the physical interpretability of physics-based models offer an alternative to black-box models. In this Perspective, we discuss the various interpretable machine learning methods available to catalysis researchers, highlight the potential of interpretable machine learning to accelerate hypothesis formation and knowledge generation, and outline critical challenges and opportunities for interpretable machine learning in heterogeneous catalysis.

Despite their enormous importance, most commercial heterogeneous catalysts have been discovered using trial-and-error experimental approaches that rely on the chemical intuition of catalysis practitioners. The difficulties associated with moving away from empirical experimental approaches towards catalyst design using predictive models are multifaceted, including the fact that heterogeneous catalysis spans time and length scales of more than nine orders of magnitude[1], and that the catalyst performance depends on many variables, such as the catalyst composition, morphology, support material and reaction environment (for example, temperature, solvent and external potential). This large parameter space makes the design and optimization of heterogeneous catalysts challenging.

Researchers have recently turned to machine learning (ML) to accelerate the study and discovery of heterogeneous catalysts, using these tools to navigate the parameter space more efficiently[2–5]. ML is a subfield of artificial intelligence that encompasses methods that self-infer patterns from data. Catalysis researchers leverage these learned patterns to streamline their work in many areas, including the atomistic simulation of reaction conditions[6,7], catalyst surface phase diagram construction[8], reaction mechanism prediction[9,10] and catalyst structure elucidation[11,12]. Most applications of ML in catalysis thus far have used black-box models (see Table 1) to make predictions of computable physical properties (descriptors), such as adsorption or formation energies, that can be related to the catalytic performance (that is, activity or stability)[2–5]. Extracting meaningful physical insights from black-box models has proved challenging, as the internal logic of black-box models is not readily interpretable due to the high degree of complexity of these models.

Interpretable ML methods that merge the predictive capacity of black-box models with the physical interpretability of physics-based models offer an alternative to black-box models. Herein, we refer to interpretable ML as models that extract relevant knowledge about relationships between catalytic variables in the form of succinct data formats such as visualizations, rule sets, or mathematical equations[13]. For example, an interesting fundamental question that interpretable ML can help to address, which we elaborate on below, is determining which physical properties of a catalyst surface govern the chemisorption strength of different adsorbates. In our view, interpretable ML methods present a complementary approach to black-box methods (Fig. 1). Translating the hidden patterns identified by ML models into interpretable information formats can lead to testable theories and hypotheses, further advancing scientific understanding. Knowledge gained from interpretation can help to explain why a model fails to make some predictions accurately and thus guide model improvement. The development and application of interpretable ML algorithms is an active area of research across law, healthcare, business, engineering and science[14–18].

In this Perspective, we discuss the interpretable ML methods that are available to catalysis researchers and the potential of interpretable ML to accelerate hypothesis formation and knowledge generation in the field of heterogeneous catalysis. We frame our discussion by briefly describing black-box models, whose generally opaque internal logic makes extracting physical insights challenging (Fig. 2a). We then introduce two general categories of interpretable ML: grey-box ML methods, which rely on model-agnostic post-hoc analyses to interpret black-box models (Fig. 2b), and glass-box methods in which outputting an interpretation is an inherent feature of the method (Fig. 2c). We highlight studies in heterogeneous catalysis that use interpretable methods (Table 1) and studies from chemistry and materials science research that use methods that have yet to see use in catalysis but may be of interest to the catalysis community. We note that interpretable ML is also helping to improve the design and study of homogeneous catalysts[19–24], which in many respects is a more mature field due to its substantial crossover with molecular design, although discussion of these applications is beyond the scope of this Perspective. Finally, we outline critical challenges for interpretable ML in heterogeneous catalysis.

Black-box methods

Black-box models, such as Gaussian process models or neural networks, are widely used in catalysis. One area that has benefitted from black-box models is computational high-throughput catalyst screening[25–28]. In most cases, these screening studies search

[1]Department of Chemical Engineering, University of Michigan, Ann Arbor, MI, USA. [2]Catalysis Science and Technology Institute, University of Michigan, Ann Arbor, MI, USA. ✉e-mail: bgoldsm@umich.edu; linic@umich.edu

Table 1 | Examples of black-box, grey-box and glass-box ML methods used in catalysis applications

Method	Description	Example application
Black box		
Neural networks[77]	Highly tunable and empirically state-of-the-art predictive models.	High-throughput prediction of CO and H adsorption energies on diverse intermetallic alloys[38].
Gaussian process models[78]	Bayesian models that quantify the prediction uncertainty.	Accelerating the construction of catalyst surface phase diagrams[8].
AdaBoost regressor[79]	Models that refine their focus during training to better fit difficult examples.	Discovery of stable materials such as oxides, phosphides, sulfides and alloys[80].
Grey box		
Global feature-importance scores[81]	Describes a feature's contribution to predictions at the dataset level.	Determining the physical properties that govern CO adsorption during CO_2 electroreduction catalyst screening[25].
Partial dependence plots[31]	Visual explanations of how each feature affects the model output.	Visualizing the impact of small-molecule and oxide-surface properties on chemisorption[34].
Shapley additive explanations[35]	Game-theoretic metric for describing a feature's contribution to an individual prediction.	Quantifying the influence of the catalyst composition and experimental conditions on the selectivity to C_2 products during the oxidative coupling of methane[36].
Glass box		
Symbolic regression[82]	Identifies simple closed-form models for predicting target properties.	Identifying an easily calculable descriptor that predicts chemisorption for various adsorbates on alloys with different compositions and surface facets[40].
Subgroup discovery[83]	Identifies and characterizes subgroups that share common traits in data.	Identifying single-atom catalysts that can break scaling relations for the nitrogen reduction reaction[51].
Generalized additive models[55]	Predictive models that are interpretable because the independent segments of the model decision-making process can be interpreted independently.	Quantifying and understanding chemisorption on alloys[57].
Principal component analysis[84]	An unsupervised ML algorithm that projects the data onto a reduced basis while describing the maximum dataset variance.	Finding electronic-structure descriptors for metal alloys and oxides[49].
Probabilistic graphical models[85]	Predictive models that can be used to enforce causal structure and quantify error.	Quantitatively attributing errors in an activity volcano plot for the oxygen reduction reaction[63].

for model surface sites that bind relevant adsorbates with desired adsorption energies. The motivation for using these approaches is that, in many cases, the design space of possible catalysts is too large to be studied using quantum chemical methods alone. ML models serve as computationally efficient surrogates to minimize expensive quantum chemical calculations, enabling an accelerated screening of the catalyst design space. For example, an ML-accelerated screening of electrochemical carbon dioxide (CO_2) reduction catalysts identified copper–aluminium alloys as active and selective materials based on the computed binding energy of carbon monoxide (CO), which has been proposed to be a descriptor of CO_2 reduction activity[27].

Black-box ML models are referred to as such because the parameters (weights, rules or connections) they learn are so overwhelming in number that directly extracting meaningful insight regarding the different physical behaviours captured by these parameters is unfeasible. For example, in the case of the aforementioned copper–aluminium alloy, the model is too complex to interpret (that is, shed light on the features of aluminium atoms that electronically change the copper atoms to modulate their interaction with the CO adsorbate) and is therefore too complex to explain (that is, contextualize the model's behaviour within the framework of existing CO_2 reduction catalysis knowledge). Nonetheless, their large number of parameters enables black-box models to typically outperform glass-box models in terms of computational accuracy for large and complex datasets. This benefit is sufficient for specific ML applications in catalysis, such as high-throughput active-site screening or creating machine-learned potentials for molecular dynamics or Monte Carlo simulations because accuracy is more desired than interpretation.

Grey-box methods

While a critical problem with black-box models is that it is challenging to interpret the internal logic that led to the conclusions of a model, there exists a class of methods for indirectly extracting interpretable information from black-box ML models after training. These approaches are called post-hoc analysis methods, referred to herein as grey-box methods. Many grey-box methods are model-agnostic and therefore usable with any class of ML model. The information from grey-box methods can take many forms but is usually a set of visualizations or sensitivity measures called feature-importance scores. Grey-box methods can generate explanations that are either global or local. Global explanations allow interpretation of the dataset-level relationships and patterns learned by black-box models, whereas local explanations allow practitioners to understand why black-box models make a specific prediction for a single data point.

The primary grey-box interpretation methods used in catalysis applications so far are global feature-importance scores[25,26,29,30]. A global feature-importance score is a sensitivity measure that describes how an individual feature or combination of features contributes to a model's predictions at the dataset level. These scores yield insight into which features the model generally finds important and allow practitioners to quantify the relative importance of different features for describing a specific type of behaviour. For example, normalized sensitivity coefficients (a measure of feature

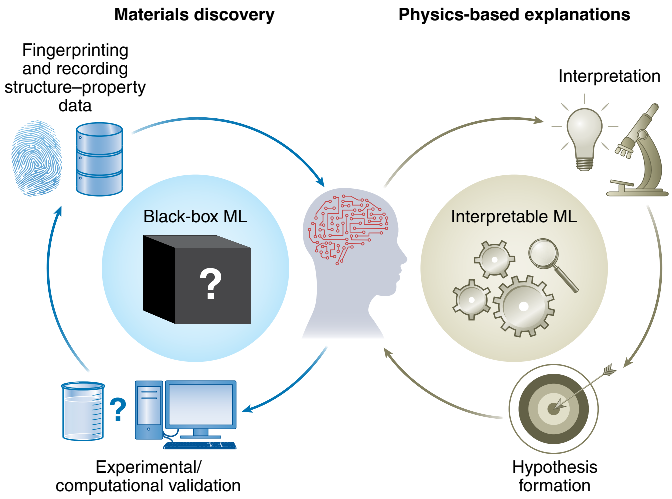
Materials discovery
Physics-based explanations
Fingerprinting and recording structure–property data
Interpretation
Black-box ML
?
Interpretable ML
?
Experimental/ computational validation
Hypothesis formation

Fig. 1 | Synergistic relationship between black-box and interpretable ML approaches. Both interpretable and black-box ML models can be used in materials discovery applications to identify promising catalysts. Validating ML-predicted catalysts via experimental synthesis and characterization and computational validation can lead to materials that exhibit desirable properties such as low cost, high stability, high activity and high selectivity. In addition, fingerprinting (that is, uniquely labelling) and recording the structure–property data from detailed characterization studies can be used to iteratively improve ML models and accelerate catalyst discovery. Interpretable ML algorithms have the advantage of outputting human-interpretable information regarding the patterns or dependencies learned by the ML model. These interpretations enable hypothesis formation about what underlying physical mechanisms might play a role in the task. This knowledge can guide feature selection such that features are linked to the target property and inform additional experiments, thereby improving the ML models and accelerating catalyst discovery further.

importance) were generated for neural networks used to perform high-throughput screening of core–shell alloy catalysts for CO_2 electroreduction using the CO adsorption energy as an activity descriptor (Fig. 3a)[25]. The sensitivity coefficients showed that the local Pauli electronegativity at a catalyst surface site plays a crucial role in predicting the CO adsorption energies, particularly for alloys that contain surface sites with fully occupied *d* bands such as copper (Cu), silver (Ag) and gold (Au). In addition, feature-importance scores of a random forest model have been used to determine the relative impact of the calculated geometric and electronic features of doped nickel phosphide (Ni_2P) catalysts on their hydrogen evolution reaction activity, ultimately identifying the Ni_β–Ni_γ bond length (with the α, β and γ notation used by the authors shown in Fig. 3b) as the most important descriptor of hydrogen evolution reaction activity[29]. These examples highlight the utility and relative ease with which feature-importance scores can extract insight from black-box models.

In addition to global feature-importance scores, there have also been limited efforts to bring post-hoc global visualizations to catalysis. Global visualization methods, such as partial dependence plots[31], accumulated local effects plots[32] and transparent model distillation[16,33], provide visual explanations of the change in the model behaviour subject to a change in only one or two feature values, which allows these effects to be plotted in a line chart or heat map, respectively. One recent application used partial dependence plots to visualize the marginal change in the predicted adsorption energies of various adsorbates such as alkanes, aromatics and amines on group 13 metal oxide surfaces (for example, aluminium oxide and gallium oxide) from a regression model subject to changing the physical properties of the adsorbate and the oxide surface[34]. The partial dependence plots indicated that the adsorbate's highest occupied molecular orbital (HOMO) energy and the oxide's surface energy play crucial roles in determining the adsorption energy (Fig. 3c), with the adsorption strength increasing for a higher-in-energy HOMO and a higher surface energy. Although limited in their applications thus far, we believe that global visualizations will be a valuable tool for researchers to interpret their black-box catalysis models in the future.

Local explanations are an alternative approach for interpreting black-box models. The most common form of local explanation is local feature-importance scores. In contrast to global feature-importance scores, which describe a feature's general contribution across many different predictions, local feature-importance scores describe a feature's contribution to an individual prediction (for example, giving insight into the contributions of electronic or geometric features to the performance of a specific material). One important method that assigns local feature-importance scores is Shapley additive explanations (SHAP)[35], which have been used in several recent catalysis studies[36,37]. For example, a recent study used SHAP in a literature meta-analysis of around 2,000 catalysts for the oxidative coupling of methane (OCM)[36]. SHAP elucidated the relative influence of different catalysts and experimental conditions on the selectivity to desired C_2 products in OCM, suggesting that a high operating temperature, a higher partial pressure of methane relative to oxygen and the presence of lanthanum and sodium in the catalyst were critical parameters for steering OCM towards the desired C_2 products. There have also been efforts in catalysis at providing local explanations, such as generating post-hoc visualizations, for black-box adsorption energy prediction models that show the contributions of individual atoms to the predicted binding energies (Fig. 3d)[38].

Grey-box methods are a promising approach for interpreting black-box models that can often yield plausible explanations of the black-box model's behaviour. Nevertheless, we add a warning that grey-box explanations can also be misleading and misrepresent black-box model behaviour[13,39]. This is because, in most cases, there exists a gap between the simple explanations offered by a grey-box method and the complex behaviour learned by the black-box model. Nonetheless, developing higher resolution grey-box methods remains an open field of research in ML that will undoubtedly benefit advances in catalysis research and in science and engineering in general.

Glass-box methods

Not all ML methods require a grey-box method to interpret the relationships they have uncovered. Some ML methods yield such insights directly, referred to herein as glass-box methods. Generally, these glass-box methods have constraints, such as enforced simplicity, that make direct interpretation of glass-box ML results possible. Glass-box methods are used to find simple analytical expressions that relate input variables to target properties, to identify hidden or underlying structures in the data, to make predictions under enforced modularity or causal structure, or to suggest causal structure directly. It is our view that glass-box methods are preferred if extracting scientific insight is the central objective.

The most applied glass-box method in catalysis applications thus far is symbolic regression[40–44]. Symbolic regression methods (for example, SISSO[45] and genetic programming[46]) algorithmically combine input features using mathematical operators (for example, +, −, ×, ÷ and log) to find functionally simple mathematical expressions that can predict target properties as a function of those features. Models from symbolic regression methods are interpretable because the simplicity of their closed-form analytical expressions allows researchers to step through the models and understand the numerical relationships between the various input features and the

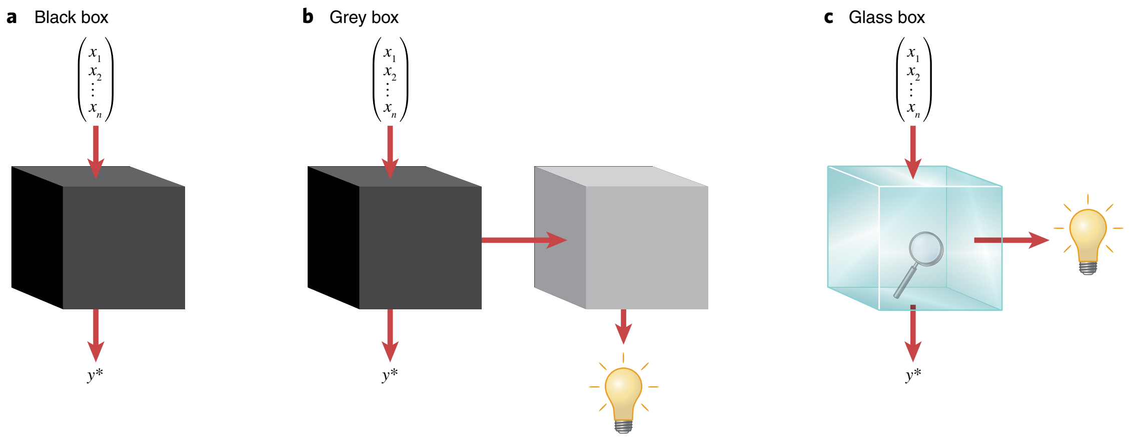

Fig. 2 | Schematic depiction of black-box, grey-box and glass-box ML methods. The objective of supervised learning applications is to predict a target property, y^*. The predictions are made using a set of variables, called features (x_1,x_2,...x_n), that characterize the catalyst. Typically, these features are compositional, electronic or geometric fingerprints of the catalytic system of interest. **a**, Many highly predictive ML approaches fall under the umbrella description of black-box methods because their models contain numerous interdependent weights, rules and connections. Although the many learned patterns can lead to a high predictive accuracy, interpreting them without additional processing is a task that is beyond the means of human cognition. **b**, Post-processing to extract insight from black-box models can be done using grey-box methods, which use a separate model or technique to derive explanations of the black-box model behaviour. **c**, Glass-box methods can output explanations directly. These methods typically have constraints on their functional behaviour that make their interpretation straightforward.

corresponding outputs. For example, researchers used symbolic regression to identify an activity descriptor for the oxygen evolution reaction (OER) on perovskite oxides (oxides with an ABO_3 structure)[43]. From a feature set containing numerous electronic (for example, valence electron structure and electronegativity) and structural features (for example, atomic radii and other structural parameters), symbolic regression identified an activity descriptor (μ/t) that combined two well-known structural parameters of perovskites, the Goldschmidt tolerance factor (t) and the octahedral factor (μ). The identified μ/t descriptor (Fig. 4a) suggested that increasing t, by incorporating large cations at the A site, and decreasing μ, by incorporating small cations at the B site, should lead to higher OER activity. Ultimately, four perovskites with experimental activities higher than previously reported perovskite oxide catalysts were identified for the OER based on these criteria. This example highlights the potential of symbolic regression for building accurate and interpretable models that relate geometric or electronic features of catalysts to their performance.

Another strategy for finding interpretable scientific insights is through unsupervised ML. Unsupervised ML algorithms are pattern-recognition algorithms that probe for underlying structure in the data[47], which can help to accelerate an exploration of the dataset. The two primary methods of unsupervised learning used in catalysis thus far are clustering algorithms and dimensionality reduction algorithms. In clustering, the goal is to identify and segment similar subpopulations, or clusters, in the data. Interpreting the clusters can facilitate an understanding of the similarities and differences between data (for example, spectra, mechanisms and structures). For example, clustering has been used in homogeneous catalysis to help identify the optimal phosphine ligands for a stereoselective Suzuki–Miyaura cross-coupling reaction catalysed by phosphine-ligand-mediated palladium catalysis[48]. To address this challenge, researchers used clustering for the rapid identification of representative phosphine ligand motifs so that the choice of phosphine ligands could be optimized systematically. *K*-means clustering grouped 365 commercially available phosphines into 24 chemically distinct clusters based on their molecular properties such as the HOMO–LUMO gap, Fukui index and volume. One compound from each cluster was experimentally evaluated based on availability, price and anticipated stability, and one of the ligands, 2-(diphenylphosphino)-2′,6′-dimethoxy-1,1′-biphenyl, was identified as having superior performance. Further investigation revealed that structurally similar triarylphosphine ligands were also effective. We note that, in general, clusters are most likely to be interpretable when clustering is performed on low-dimensional data of two or three features, as the clusters and decision boundaries used for cluster assignment can then be easily visualized.

In dimensionality reduction, high-dimensional data (for example, signals, images or large feature sets) are transformed to a lower-dimensional subspace that still captures the essence of the data. In the lower-dimensional space, the data can be visualized and may be interpretable. Nonetheless, the transformation from high to low dimensions can be complex, non-linear and highly parameterized, making it difficult to understand how the lower-dimensional data relate to the original high-dimensional data. One method of developing this understanding is through explanatory visualizations. For example, we recently used a dimensionality reduction technique called principal component analysis (PCA) to establish relationships between the geometric structures of subsurface alloys based on rhodium (Rh), palladium (Pd), iridium (Ir) and platinum (Pt) (in which a ligand metal composes the layer immediately beneath the surface) and the chemisorption strengths of different adsorbates on the alloys[49]. We used electronic-structure descriptors as a bridge to relate the chemisorption strength to the geometric structure and composition of the alloys. To accomplish this goal, we employed PCA on the *d*-projected density of states at the adsorption site to derive principal component (PC) descriptors of the sites' electronic structure. Interpretation of the PC descriptors using signal reconstruction showed that they mapped to the geometric structure of the sites. For example, one of the machine-learned PC descriptors (Fig. 4c), found to be negatively correlated with the chemisorption energy, was associated with a broadening and downshift in the surface electronic *d* band. Relating this machine-learned descriptor to geometric structure revealed that this broadening and downshift is caused by increasing the surface metal size, increasing the ligand metal size and decreasing the number of ligand *d* electrons. Ultimately, the PCA-based approach led us to conclude that selecting larger surface metal atoms, larger ligand metal atoms or ligand metals with fewer *d* electrons lowers the chemisorption energies of most adsorbates. Such insights are critical to designing surface sites with a specific chemical activity.

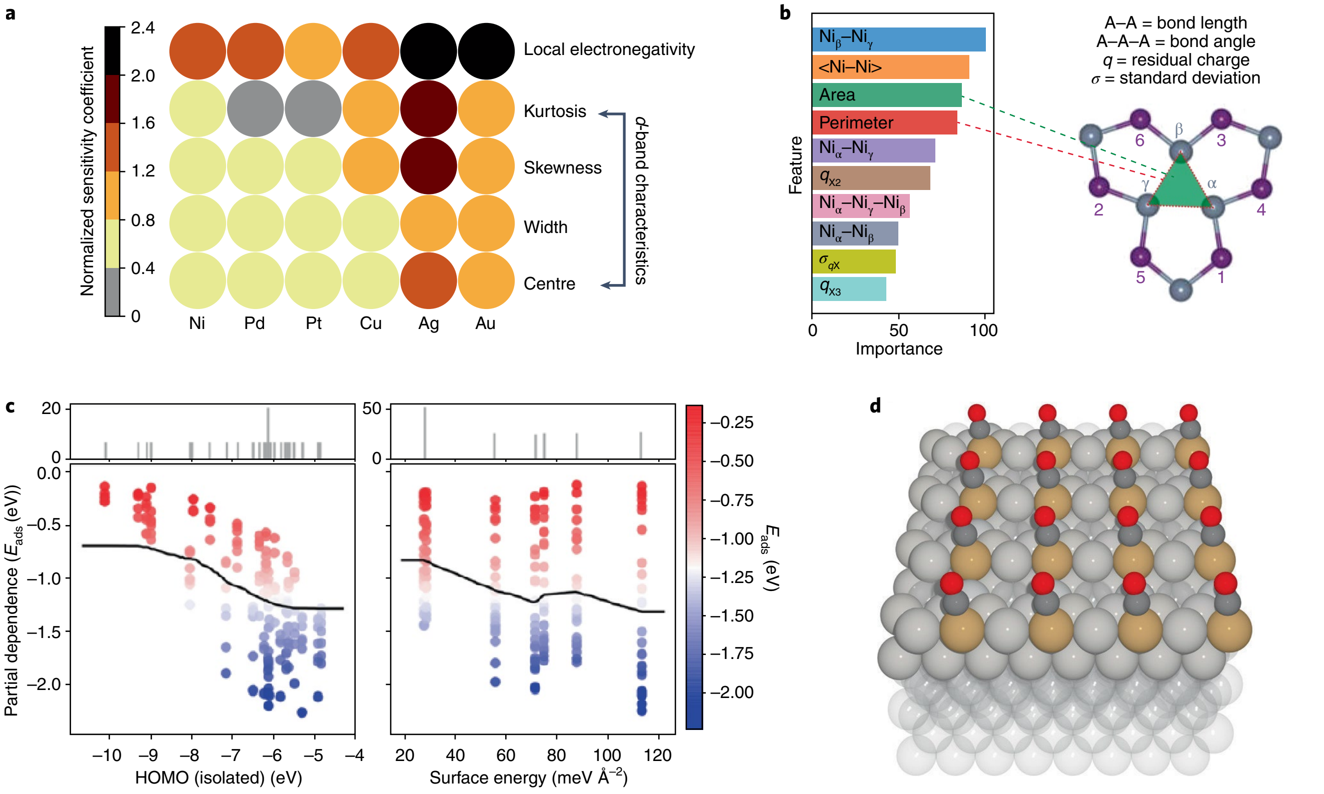

Fig. 3 | Grey-box methods in catalysis applications. a, Normalized sensitivity coefficients are used to assess the feature importance in a neural network for predicting CO adsorption on Ni, Pd, Pt, Cu, Ag and Au core–shell alloys, where the listed metal is the shell. **b**, Relative feature importance on the hydrogen evolution reaction activity of Ni_2P catalysts obtained using a random forest model. Features are defined based on the displayed Ni_2P model system, in which nickel is coloured grey and phosphorus doping sites are coloured purple. The angled brackets on <Ni–Ni> indicate an average bond length. The X notation loosely relates to the dopants. For example, q_{X2} denotes the residual charge on the dopants when there are two dopants, and q_{X3} denotes the residual charge on the dopants when there are three. The standard deviation of the dopant charges is given by σ_{qX}. **c**, Partial dependence plots showing the marginal effect of an adsorbate's HOMO energy and an oxide's surface energy on the adsorption energy (E_{ads}) of the corresponding system. The black lines indicate the partial dependence plot, and the points, coloured according to the adsorption strength, show the actual values. Histograms showing the distribution of the actual values are displayed above. **d**, A saliency map for a crystal graph convolutional neural network model shows which individual atoms in a Cu(211) model contribute the most to the predicted CO binding energy (where O is coloured red, C is coloured dark grey and Cu atoms are coloured light grey and light brown), with the light brown Cu atomic colouring indicating a larger contribution. Panel **a** reproduced with permission from ref. [25], American Chemical Society. Panels adapted with permission from: **b**, ref. [29], American Chemical Society; **c**, ref. [34], American Chemical Society; **d**, ref. [38], American Chemical Society.

Another class of models that probe for latent or underlying structure in a dataset are subgroup discovery (SGD) algorithms[50]. Applications of ML in heterogeneous catalysis often focus on learning a single global model to predict a target property for many different catalysts. However, global models can fail to account for the possibility that the physical mechanism governing a target property may differ across different subgroups of catalysts, with the global model obfuscating or misrepresenting the physics affecting the target property. Therefore, it may sometimes be beneficial to identify and segment physically similar subgroups and learn local models for each subgroup. The interpretation of these subgroups can enable an understanding of the geometric and chemical similarities between the catalysts[51–53]. A recent application of SGD identified single-atom catalysts capable of breaking scaling relations between reaction intermediates for the nitrogen reduction reaction (Fig. 4d)[51]. This analysis demonstrated that scaling relations were only broken by early transition metal atoms, with additional electronic-structure analysis showing that early transition metals offer this advantage because of charge transfer to the support, limiting the amount of charge available for bonding electronegative adsorbates like N* and NH*, where the asterisk indicates an adsorbed species. SGD can also identify regions of feature space where an ML model's predictions are trustworthy, called the domain of applicability. Recent work used such an approach to identify the domain of applicability for ML models predicting the stability of transparent conducting oxides[54]. These examples highlight the potential of SGD for identifying geometrically or chemically similar groupings of catalysts and improving predictive models.

There have also been efforts to leverage glass-box models with enforced modularity for extracting catalysis insights. A model is modular if each feature's contribution to the model decision-making process can be independently interpreted. For example, generalized additive models (GAMs) make predictions by summing independent, potentially non-linear, functions of each input variable[55,56]. Because the overall model is a linear combination of functions that are dependent on only one or two variables of interest, each independent function can be visualized easily, thus shedding light on the model behaviour. We recently used GAMs to develop chemisorption models for OH, Cl, O and S adsorbates on Rh-, Pd-, Ag-, Ir-, Pt- and Au-based subsurface alloys (Fig. 4e)[57]. The GAMs identified a few critical material properties that control the chemisorption strength on the alloys, showing that for a fixed surface atom, the number of *d* electrons in the ligand subsurface metal and the size of the ligand atom are two critical parameters that

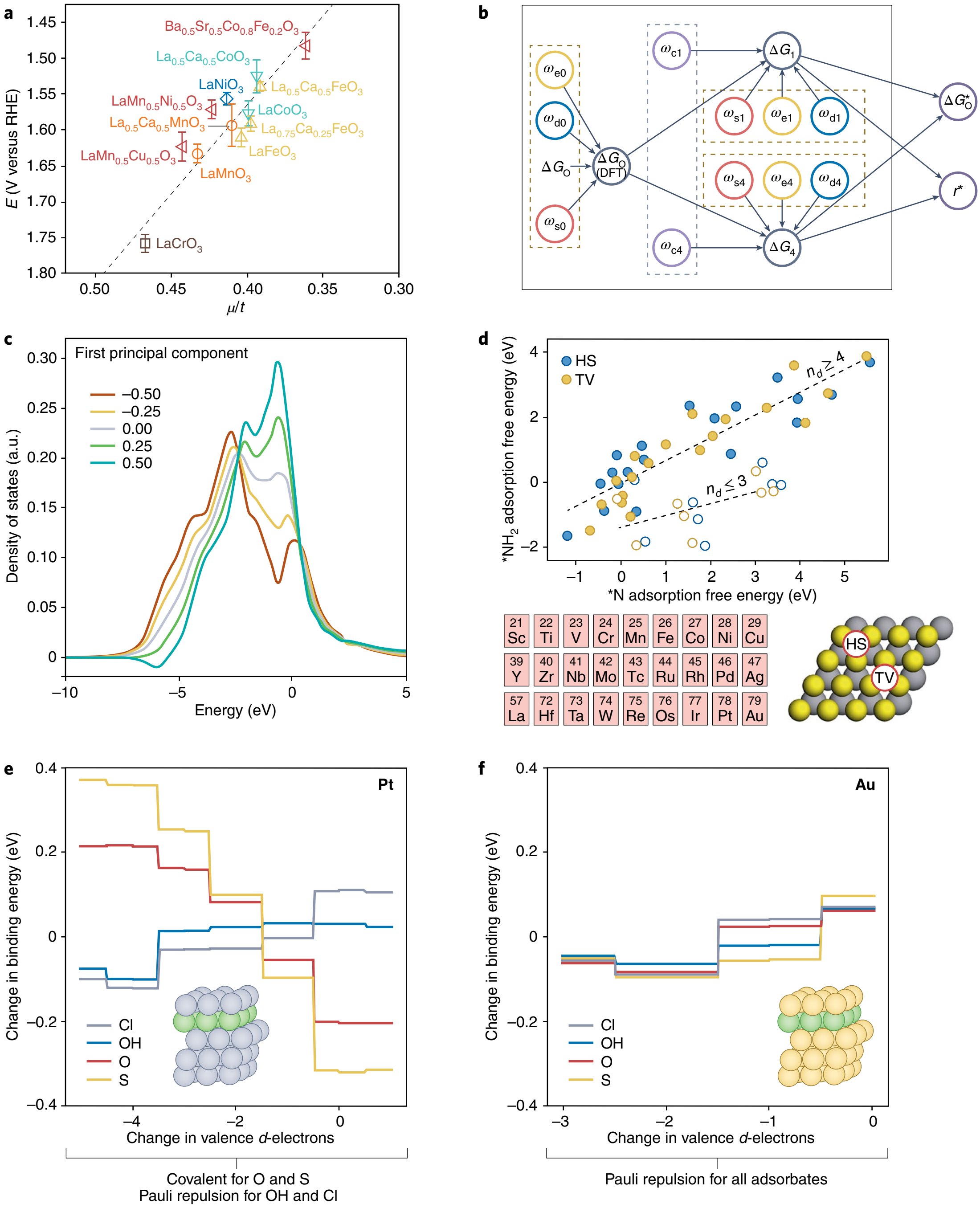

Fig. 4 | Interpreting glass-box ML results. a, The limiting potential (E) at 50 mA cm$^{-2}_{ox}$ of the OER current plotted as a function of the geometric μ/t descriptor identified using symbolic regression. The dashed grey line shows the predicted limiting potential based on the identified μ/t descriptor, given by $E = 2.52\mu/t + 0.55$. The error bars indicate the standard deviation of at least three independent experimental measurements. **b**, A directed acyclic graph was used to predict the optimal oxygen binding energy (ΔG^*_O) and optimal reaction rate (r^*) for the ORR. The uncertainty from sources like solvation treatment (ω_s), density functional theory error (ω_d), experimental error (ω_e) and imperfect empirical correlations (ω_c) were quantified and interpreted using directed acyclic graphs, which increases model credibility and offers insight into how a model's predictive accuracy can be improved. **c**, Signal reconstruction shows the electronic-structure effects captured using a PC descriptor learned on the surface d-projected density of states of subsurface alloys based on Rh, Pd, Ir and Pt. **d**, Calculated $^*NH_2$ adsorption free energy as a function of the calculated *N adsorption free energy on single transition metal atoms on vanadium disulfide (VS_2) supports. All single-atom systems containing four or more d electrons ($n_d \geq 4$; filled points) fulfil the scaling relation, whereas early transition metals with three or fewer d electrons ($n_d \leq 3$; open data points) do not. The overview of the single metal atoms studied (left) and a top view of the VS_2 substrate in which V atoms are coloured grey and S atoms are coloured yellow (right) is shown under the plot. The two single-atom adsorption sites considered are shown: one hexagonal close-packed hollow site atop a V atom (TV) and one face-centred cubic hollow site atop a subsurface hollow site (HS). **e**,**f**, GAMs shed light on the effect of the number of ligand valence d electrons relative to the host metal on the chemisorption energies of O, OH, S, and Cl on Pt (**e**) and Au (**f**) subsurface alloys. The subsurface alloy model systems are shown in the insets, with Pt coloured grey, Au coloured gold and the generic ligand metal atom coloured green. Panel **a** reproduced with permission from ref. [43], Springer Nature Ltd under a Creative Commons license CC BY 4.0. Panels adapted with permission from: **b**, ref. [63], AAAS; **c**, ref. [49], Elsevier; **d**, ref. [51], American Chemical Society; **e**,**f**, ref. [57], Elsevier.

control the chemisorption behaviour, which corroborated earlier reports[58,59]. Notably, the GAM models suggested that the number of *d* electrons in the ligand metal describes the degree of Pauli repulsion between the alloy surface and the adsorbate. Specifically, the GAMs connected the opposite adsorption trends of electron-rich adsorbates, such as OH and Cl, compared with adsorbates that are relatively electron-poor, such as O and S, on platinum alloys to the number of ligand *d* electrons (Fig. 4e). A previous chemisorption study, which used a physics-based chemisorption model[60], attributed similar opposite trends in the chemisorption strength of OH and Cl compared with O and S on platinum alloys to Pauli repulsion[61]. Additional analysis of adsorption trends on metals with completely full *d* bands, such as gold (Fig. 4f), whose chemisorption trends should depend primarily on Pauli repulsion, corroborated this hypothesis as they mirrored the behaviour of the electron-rich adsorbates on platinum. We expect that modular ML models will be a valuable tool for catalysis researchers in the future, especially with the expanding availability of open-source software packages that contain these model classes, such as InterpretML[62].

Generally, interpretations can shed light on the physical correlations between a catalyst's performance and its geometric and chemical properties. Nonetheless, many of these methods require the catalysis researcher to interpret and validate the results to ensure that their ML models do not learn illogical causal connections. However, it is also possible to build models with an enforced causal structure to ensure that ML models will behave in a manner that is consistent with previous physical knowledge[63–65]. Recent work used such an approach to quantify and attribute the errors introduced during the construction of an activity volcano plot for the oxygen reduction reaction (ORR) by platinum-group-metal-based catalysts[63]. Researchers used a probabilistic graphical model, represented by the directed graph in Fig. 4b, to enforce a causal structure between the primary microkinetic input (the oxygen binding energy) and potential error sources when predicting the optimal oxygen binding energy for the ORR. The model accounted for errors in solvation treatment, density functional theory (DFT), experiment and the empirical scaling correlations used to predict the activation and intermediate chemisorption energies. They found that the primary sources of error were errors in the DFT calculations and imperfect correlations between OOH* and OH* binding energies and O* binding energies. There also exist methods, called causal inference methods, to determine causal structure from raw data alone[66]. Although there have only been limited applications of causal inference thus far in the physical sciences[67], the diffusion of advances in causal inference to catalysis research is likely to have a notable impact.

In contrast to the information gap present when interpreting black-box models with grey-box methods, glass-box methods provide a full-resolution explanation of their behaviour. As a result, glass-box methods are preferable for applications where developing scientific insight is the central objective. Although glass-box models may have a worse predictive performance compared with black-box models in practice due to their functional forms that are constrained to enforce modularity, causal structure or simplicity, there is, in theory, no reason that glass-box models cannot be competitive in terms of predictive accuracy. For problems where the data are well structured and the features are physically linked to the target property, there are often little to no performance differences between black-box models and simpler glass-box models[39].

Challenges and opportunities for progress

Despite the successes of interpretable ML in catalysis thus far, the field is still nascent, with substantial challenges to be overcome for it to reach its full potential. Below we highlight the critical challenges that are related to integration with experiment, dataset size, model reusability, and explaining interpretations. Addressing these challenges will require close collaboration between experimentalists, theoreticians and computer scientists.

Integrating interpretable ML with experimental data. Most interpretable ML studies in heterogeneous catalysis thus far have used computational datasets due to a lack of suitably large experimental datasets. Typical ML studies include 100–10,000 training data points[68], which can be prohibitively time-consuming to obtain experimentally. One possible avenue for constructing large experimental datasets for interpretable ML studies is to data-mine the wealth of catalyst data already available in the literature. Despite this approach being the most straightforward because the data already exist, and such processes can be done by applying natural language processing (that is, text extraction[69]), it is questionable if such an approach would yield a coherent dataset amenable to interpretable ML analysis. Data reporting in the field of heterogeneous catalysis often lacks common standards in both experimental operating conditions and catalyst characterization that may limit the interpretability of insights gained from such a dataset. Another approach would be proliferating new combinatoric high-throughput catalyst synthesis and characterization techniques aided using flow reactors, robotics or computer vision[70,71]. Advances in such approaches will facilitate the construction of large experimental datasets with common operating conditions, which may be more amenable to interpretable ML.

Capturing the experimental complexity of catalysts with ML-derived descriptors. An additional challenge facing the integration of interpretable ML with experiments arises from the reliance of most existing interpretable ML studies on computational descriptors, such as the adsorption energies of surface reactive intermediates[2–5]. Despite the successes of descriptor-based screening approaches in heterogeneous catalysis, we emphasize that they have their limitations and can often be inadequate for predicting the behaviour of real catalysts under reaction conditions. Descriptor-based approaches often neglect the complexities present in real catalytic materials, such as environment-induced changes to surface sites, an abundance of diverse surface sites and an inability to synthesize the desired surface sites, among many others. Consequently, generating new hypotheses and descriptor sets that can better predict the behaviour of real catalytic materials under reaction conditions is a critical challenge, but one that we are optimistic that future applications of interpretable ML can help to address.

Dataset size. Even though most ML applications in catalysis thus far have relied on computational datasets generated using quantum chemical calculations, these datasets are still expensive and time-consuming to generate, as many of the calculations rely on high-performance supercomputing resources. Merging black-box ML with interpretable ML may aid future efforts. For example, approaches that leverage black-box surrogate models (for example, active learning[28], Bayesian global optimization[72] and myopic multi-scale sampling[73]) show promise for efficiently constructing large computational datasets for high-throughput catalyst screening. Researchers could use such datasets not only for screening but also to uncover scientific insights using interpretable ML.

Model accessibility and reusability. Most published ML models are seldom reused. One reason for this is that models are usually learned on datasets generated ad hoc; the development of standardized datasets would certainly improve reusability[74]. An additional barrier to reusability is the considerable amount of domain expertise in mathematics and computer science that is required to leverage ML models beyond the typical skill set acquired by chemists, materials scientists and chemical engineers. Furthermore, the features used in the models can provide an additional barrier to use if they are computational or experimental values that are difficult to obtain, as opposed to easily accessible tabulated properties. One route to expanding the accessibility of machine-learned results would be the development of user-friendly web applications that interface with ML models, similar to those provided by the Materials Project[75]. Another route is the proliferation of interpretable ML models. We believe that interpretable models using easily obtainable tabulated properties will help to broaden the community use and impact of machine-learned results. For example, a machine-learned stability descriptor of perovskites identified using symbolic regression has begun to see widespread use, probably due to its closed-form expression and easy computability[76].

Explaining interpretations. As helpful as interpretation tools might be, ML cannot eliminate the role of catalysis scientists in advancing scientific theories and hypotheses. While the ML model and the interpretability method are critical in yielding insightful interpretations, the features used as inputs to the model also play a pivotal role in enabling interpretation. Features should be selected using domain expertise. We believe that, if possible, the best practice is to use features that align with earlier physical explanations, as the interpretation is likely to be more insightful if it reinforces or connects to pre-existing domain knowledge. Such features could include electronic and geometric descriptions of surface sites, spectroscopic data and experimental operating conditions (temperature, pressure and reactant concentrations). Nonetheless, we emphasize that there are no universal criteria for choosing features for interpretable ML applications. The features selected depend heavily on the system being studied, and therefore the need for domain expertise cannot be eliminated.

Even more importantly, it is up to researchers to contextualize, audit and frame interpretable ML models using existing catalysis knowledge. It is unlikely that an ML model in and of itself will discover new or unexpected physics. Quite the opposite is true. If an ML model predicts unexpected physics, the model is more likely to be incorrect (for example, learning correlation rather than causation). The power of interpretable ML lies in its ability to enable researchers to generate hypotheses more easily from data, which can inform additional characterization and experiments. An excellent example of using interpretable ML for hypothesis formation was the previously mentioned work that used SGD to identify single-atom catalysts capable of breaking scaling relations between reaction intermediates for the nitrogen reduction reaction[51]. The researchers used SGD to identify that only early transition metals could break scaling the relations between N* and NH*. They then performed additional electronic-structure analysis to elucidate that early transition metals are superior due to charge transfer to the support, limiting the amount of charge available for bonding electronegative adsorbates. In the same way that the SGD result helped guide the subsequent electronic-structure analysis, future applications of interpretable ML in catalysis should use ML to guide additional computational or experimental corroboration of their hypotheses. Ultimately, the interpretations derived from ML models are meaningless unless catalysis researchers can explain them.

Final remarks

It is near certain that applications of ML in heterogeneous catalysis will increase in the future. The field of heterogeneous catalysis has many complex problems that are ripe for exploration with ML, such as uncovering structure–mechanism–reactivity relationships for multi-component catalysts, elucidating the physicochemical properties that govern photocatalytic, plasma-catalytic and electrocatalytic reactions, and analysing microscopic and spectroscopic data. The early success of ML for accelerating research progress inspires confidence in its ability to help tackle these challenges. Nevertheless, applications of ML to generate new knowledge and hypotheses remain few and far between, mainly because most applications of ML in catalysis thus far have used black-box models. Whereas black-box models can be advantageous from the pragmatic standpoint of high predictive accuracy, directly interpreting their behaviour is intractable. Several interpretable ML methods have been employed in recent heterogeneous catalysis studies, which we broadly group into grey-box and glass-box methods. Grey-box methods enable the interpretation of black-box models, but relying solely on these methods is risky due to a potential information gap between the black-box model and the explanation. Glass-box models, in which interpretation is an inherent feature of the model, are superior for applications where developing scientific insight is the primary objective. Ultimately, it is our view that further applications of interpretable ML methods in heterogeneous catalysis will accelerate knowledge generation in the field.

Data availability

The panels of Fig. 3 and Fig. 4 were adapted from refs. [25,29,34,38,43,49,51,57,63] and have associated raw data.

Received: 12 October 2021; Accepted: 20 December 2021;
Published online: 17 March 2022

References

1. Vlachos, D. G. in *Advances in Chemical Engineering* Vol. 30 (ed. Marin, G. B.) 1–61 (Academic, 2005).
2. Goldsmith, B. R., Esterhuizen, J., Liu, J.-X., Bartel, C. J. & Sutton, C. Machine learning for heterogeneous catalyst design and discovery. *AlChE J.* **64**, 2311–2323 (2018).
3. Schlexer Lamoureux, P. et al. Machine learning for computational heterogeneous catalysis. *ChemCatChem* **11**, 3581–3601 (2019).
4. Kitchin, J. R. Machine learning in catalysis. *Nat. Catal.* **1**, 230–232 (2018).
5. Toyao, T. et al. Machine learning for catalysis informatics: recent applications and prospects. *ACS Catal.* **10**, 2260–2297 (2020).
6. 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).
7. Boes, J. R. & Kitchin, J. R. Modeling segregation on AuPd(111) surfaces with density functional theory and Monte Carlo simulations. *J. Phys. Chem. C* **121**, 3479–3487 (2017).
8. 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).
9. Peterson, A. A. Acceleration of saddle-point searches with machine learning. *J. Chem. Phys.* **145**, 074106 (2016).
10. 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).
11. Kolsbjerg, E. L., Peterson, A. A. & Hammer, B. Neural-network-enhanced evolutionary algorithm applied to supported metal nanoparticles. *Phys. Rev. B* **97**, 195424 (2018).
12. Jennings, P. C., Lysgaard, S., Hummelshøj, J. S., Vegge, T. & Bligaard, T. Genetic algorithms for computational materials discovery accelerated by machine learning. *NPJ Comput. Mater.* **5**, 46 (2019).
13. Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R. & Yu, B. Definitions, methods, and applications in interpretable machine learning. *Proc. Natl Acad. Sci. USA* **116**, 22071–22080 (2019).
14. Caruana, R. et al. Intelligible models for healthcare: predicting pneumonia risk and hospital 30-day readmission. In *Proc. 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 1721–1730 (ACM, 2015).
15. Unceta, I., Nin, J. & Pujol, O. Towards global explanations for credit risk scoring. Preprint at https://arxiv.org/abs/1811.07698 (2018).
16. Tan, S., Caruana, R., Hooker, G. & Lou, Y. Distill-and-compare: auditing black-box models using transparent model distillation. *Proc. 2018 AAAI/ACM Conference on AI, Ethics, and Society* 303–310 (ACM, 2018)
17. Azodi, C. B., Tang, J. & Shiu, S.-H. Opening the black box: interpretable machine learning for geneticists. *Trends Genet.* **36**, 442–455 (2020).
18. Dybowski, R. Interpretable machine learning as a tool for scientific discovery in chemistry. *New J. Chem.* **44**, 20914–20920 (2020).
19. Rothenberg, G. Data mining in catalysis: separating knowledge from garbage. *Catal. Today* **137**, 2–10 (2008).

20. Janet, J. P. & Kulik, H. J. Resolving transition metal chemical space: feature selection for machine learning and structure–property relationships. *J. Phys. Chem. A* **121**, 8939–8954 (2017).
21. Ahneman, D. T., Estrada, J. G., Lin, S., Dreher, S. D. & Doyle, A. G. Predicting reaction performance in C–N cross-coupling using machine learning. *Science* **360**, 186–190 (2018).
22. Maley, S. M. et al. Quantum-mechanical transition-state model combined with machine learning provides catalyst design features for selective Cr olefin oligomerization. *Chem. Sci.* **11**, 9665–9674 (2020).
23. Reid, J. P. & Sigman, M. S. Holistic prediction of enantioselectivity in asymmetric catalysis. *Nature* **571**, 343–348 (2019).
24. Gallarati, S. et al. Reaction-based machine learning representations for predicting the enantioselectivity of organocatalysts. *Chem. Sci.* **12**, 6879–6889 (2021).
25. 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).
26. Li, Z., Wang, S., Chin, W. S., Achenie, L. E. & Xin, H. High-throughput screening of bimetallic catalysts enabled by machine learning. *J. Mater. Chem. A* **5**, 24131–24138 (2017).
27. Zhong, M. et al. Accelerated discovery of CO_2 electrocatalysts using active machine learning. *Nature* **581**, 178–183 (2020).
28. 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).
29. Wexler, R. B., Martirez, J. M. P. & Rappe, A. M. Chemical pressure-driven enhancement of the hydrogen evolving activity of Ni_2P from nonmetal surface doping interpreted via machine learning. *J. Am. Chem. Soc.* **140**, 4678–4683 (2018).
30. Wexler, R. B., Qiu, T. & Rappe, A. M. Automatic prediction of surface phase diagrams using ab initio grand canonical Monte Carlo. *J. Phys. Chem. C* **123**, 2321–2328 (2019).
31. Friedman, J. H. Greedy function approximation: a gradient boosting machine. *Ann. Stat.* **29**, 1189–1232 (2001).
32. Apley, D. W. & Zhu, J. Visualizing the effects of predictor variables in black box supervised learning models. Journal of the Royal Statistical Society: Series B (Statistical Methodology) **82**, 1059–1086 (2020).
33. Tan, S., Caruana, R., Hooker, G., Koch, P. & Gordo, A. Learning global additive explanations for neural nets using model distillation. Preprint at https://arxiv.org/abs/1801.08640 (2018).
34. Liu, C. et al. Frontier molecular orbital based analysis of solid–adsorbate interactions over group 13 metal oxide surfaces. *J. Phys. Chem. C* **124**, 15355–15365 (2020).
35. Lundberg, S. M. & Lee, S.-I. A unified approach to interpreting model predictions. In *Proc. 31st International Conference on Neural Information Processing Systems* (eds Guyon, I. et al.) 4768–4777 (Curran Associates, 2017).
36. 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* **13**, 3636–3655 (2021).
37. Ding, R. et al. Machine learning-guided discovery of underlying decisive factors and new mechanisms for the design of nonprecious metal electrocatalysts. *ACS Catal.* **11**, 9798–9808 (2021).
38. Back, S. et al. Convolutional neural network of atomic surface structures to predict binding energies for high-throughput screening of catalysts. *J. Phys. Chem. Lett.* **10**, 4401–4408 (2019).
39. Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.* **1**, 206–215 (2019).
40. Andersen, M., Levchenko, S., Scheffler, M. & Reuter, K. Beyond scaling relations for the description of catalytic materials. *ACS Catal.* **9**, 2752–2759 (2019).
41. Jonayat, A. S. M., van Duin, A. C. T. & Janik, M. J. Discovery of descriptors for stable monolayer oxide coatings through machine learning. *ACS Appl. Energy Mater.* **1**, 6217–6226 (2018).
42. 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).
43. Weng, B. et al. Simple descriptor derived from symbolic regression accelerating the discovery of new perovskite catalysts. *Nat. Commun.* **11**, 3513 (2020).
44. Liu, C.-Y., Zhang, S., Martinez, D., Li, M. & Senftle, T. P. Using statistical learning to predict interactions between single metal atoms and modified MgO(100) supports. *NPJ Comput. Mater.* **6**, 102 (2020).
45. 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).
46. Wang, Y., Wagner, N. & Rondinelli, J. M. Symbolic regression in materials science. *MRS Commun.* **9**, 793–805 (2019).
47. Murphy, K. P. *Machine Learning: A Probabilistic Perspective* (MIT Press, 2012).
48. Christensen, M. et al. Data-science driven autonomous process optimization. *Commun. Chem.* **4**, 112 (2021).
49. 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).
50. Atzmueller, M. Subgroup discovery. *WIREs Data Min. Knowl. Discov.* **5**, 35–49 (2015).
51. 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).
52. Goldsmith, B. R., Boley, M., Vreeken, J., Scheffler, M. & Ghiringhelli, L. M. Uncovering structure-property relationships of materials by subgroup discovery. *New J. Phys.* **19**, 013031 (2017).
53. Foppa, L. & Ghiringhelli, L. M. Identifying outstanding transition-metal-alloy heterogeneous catalysts for the oxygen reduction and evolution reactions via subgroup discovery. *Top. Catal.* https://doi.org/10.1007/s11244-021-01502-4 (2021).
54. Sutton, C. et al. Identifying domains of applicability of machine learning models for materials science. *Nat. Commun.* **11**, 4428 (2020).
55. Hastie, T. J. & Tibshirani, R. J. *Generalized Additive Models* (Chapman and Hall, 1990).
56. Lou, Y., Caruana, R. & Gehrke, J. Intelligible models for classification and regression. In *Proc. 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 150–158 (ACM, 2012).
57. 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).
58. Mavrikakis, M., Hammer, B. & Nørskov, J. K. Effect of strain on the reactivity of metal surfaces. *Phys. Rev. Lett.* **81**, 2819–2822 (1998).
59. Kitchin, J. R., Nørskov, J. K., Barteau, M. A. & Chen, J. G. Role of strain and ligand effects in the modification of the electronic and chemical properties of bimetallic surfaces. *Phys. Rev. Lett.* **93**, 156801 (2004).
60. Hammer, B., Morikawa, Y. & Nørskov, J. K. CO chemisorption at metal surfaces and overlayers. *Phys. Rev. Lett.* **76**, 2141–2144 (1996).
61. Xin, H. & Linic, S. Communications: exceptions to the d-band model of chemisorption on metal surfaces: the dominant role of repulsion between adsorbate states and metal d-states. *J. Chem. Phys.* **132**, 221101 (2010).
62. Nori, H., Jenkins, S., Koch, P. & Caruana, R. InterpretML: a unified framework for machine learning interpretability. Preprint at https://arxiv.org/abs/1909.09223 (2019).
63. Feng, J., Lansford, J. L., Katsoulakis, M. A. & Vlachos, D. G. Explainable and trustworthy artificial intelligence for correctable modeling in chemical sciences. *Sci. Adv.* **6**, eabc3204 (2020).
64. Wang, S., Pillai, H. S. & Xin, H. Bayesian learning of chemisorption for bridging the complexity of electronic descriptors. *Nat. Commun.* **11**, 6132 (2020).
65. 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).
66. Pearl, J. Causal inference in statistics: an overview. *Stat. Surv.* **3**, 96–146 (2009).
67. Schölkopf, B. et al. Modeling confounding by half-sibling regression. *Proc. Natl Acad. Sci. USA* **113**, 7391–7398 (2016).
68. Andersen, M. & Reuter, K. Adsorption enthalpies for catalysis modeling through machine-learned descriptors. *Acc. Chem. Res.* **54**, 2741–2749 (2021).
69. Kim, E. et al. Materials synthesis insights from scientific literature via text extraction and machine learning. *Chem. Mater.* **29**, 9436–9444 (2017).
70. Tabor, D. P. et al. Accelerating the discovery of materials for clean energy in the era of smart automation. *Nat. Rev. Chem.* **3**, 5–20 (2018).
71. Yang, L. et al. Discovery of complex oxides via automated experiments and data science. *Proc. Natl Acad. Sci. USA* **118**, e2106042118 (2021).
72. 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).
73. Tran, K. et al. Computational catalyst discovery: Active classification through myopic multiscale sampling. *J. Chem. Phys.* **154**, 124118 (2021).
74. Chanussot, L. et al. Open Catalyst 2020 (OC20) dataset and community challenges. *ACS Catal.* **11**, 6059–6072 (2021).
75. Jain, A. et al. Commentary: The Materials Project: a materials genome approach to accelerating materials innovation. *APL Mater.* **1**, 011002 (2013).
76. Bartel, C. J. et al. New tolerance factor to predict the stability of perovskite oxides and halides. *Sci. Adv.* **5**, eaav0693 (2019).
77. Goodfellow, I., Bengio, Y. & Courville, A. *Deep Learning* (MIT Press, 2016).
78. Rasmussen, C. E. in *Advanced Lectures on Machine Learning* (eds Bousquet, O. et al.) 63–71 (Springer, 2004).
79. Freund, Y. & Schapire, R. E. A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Syst. Sci.* **55**, 119–139 (1997).

80. Montoya, J. H. et al. Autonomous intelligent agents for accelerated materials discovery. *Chem. Sci.* **11**, 8517–8532 (2020).
81. Morris, M. D. Factorial sampling plans for preliminary computational experiments. *Technometrics* **33**, 161–174 (1991).
82. Augusto, D. A. & Barbosa, H. J. C. Symbolic regression via genetic programming. In *Proc. Vol.1. Sixth Brazilian Symposium on Neural Networks* 173–178 (IEEE, 2000).
83. Herrera, F., Carmona, C. J., González, P. & del Jesus, M. J. An overview on subgroup discovery: foundations and applications. *Knowl. Inf. Syst.* **29**, 495–525 (2011).
84. Hastie, T., Friedman, J. & Tibshirani, R. *The Elements of Statistical Learning* (Springer, 2001).
85. Koller, D. & Friedman, N. *Probabilistic Graphical Models: Principles and Techniques* (MIT Press, 2009).

Acknowledgements

This work was supported by the US DOE Office of Basic Energy Sciences, Division of Chemical Sciences (DE-SC0021008) (analysis of alloy chemisorption) and the CBET-National Science Foundation under DMREF grant no. 2116646. We acknowledge support from the Michigan Institute for Data Science (MIDAS) PODS Grant. J.A.E. acknowledges support from the University of Michigan J. Robert Beyster Computational Innovation Graduate Fellows Program.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence should be addressed to Bryan R. Goldsmith or Suljo Linic.

Peer review information *Nature Catalysis* thanks Johannes Margraf 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.