THE JOURNAL OF PHYSICAL CHEMISTRY LETTERS
A JOURNAL OF THE AMERICAN CHEMICAL SOCIETY

pubs.acs.org/JPCL

Perspective

Interpretable Machine Learning of Chemical Bonding at Solid Surfaces

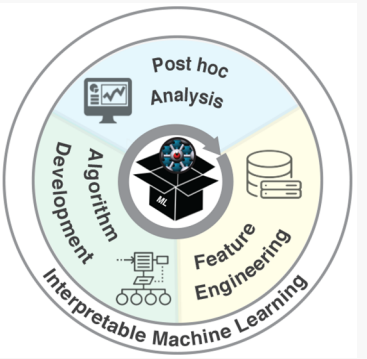
Noushin Omidvar, Hemanth S. Pillai, Shih-Han Wang, Tianyou Mou, Siwen Wang, Andy Athawale, Luke E. K. Achenie, and Hongliang Xin*

Cite This: *J. Phys. Chem. Lett.* 2021, 12, 11476–11487

Read Online

ACCESS | Metrics & More | Article Recommendations

ABSTRACT: Understanding the nature of chemical bonding and its variation in strength across physically tunable factors is important for the development of novel catalytic materials. One way to speed up this process is to employ machine learning (ML) algorithms with online data repositories curated from high-throughput experiments or quantum-chemical simulations. Despite the reasonable predictive performance of ML models for predicting reactivity properties of solid surfaces, the ever-growing complexity of modern algorithms, e.g., deep learning, makes them black boxes with little to no explanation. In this Perspective, we discuss recent advances of interpretable ML for opening up these black boxes from the standpoints of feature engineering, algorithm development, and post hoc analysis. We underline the pivotal role of interpretability as the foundation of next-generation ML algorithms and emerging AI platforms for driving discoveries across scientific disciplines.



Chemical bonding at solid surfaces underpins many technological processes, including industrial separations, pollution remediation, and interconversion of energies mediated by molecular carriers.[1] For catalytic reactions occurring at gas−solid or liquid−solid interfaces, adsorption of reactive species onto surface atoms is a prerequisite for bond breaking and formation, thus playing a pivotal role in kinetics.[2] Attributed to linear scaling relationships,[3−6] adsorption energies of one or two simple intermediates at site ensembles largely dictate the activity and selectivity of catalytic materials. As one of the oldest rules in catalysis, the Sabatier principle highlights the importance of such reactivity descriptors in the volcano-shaped plots of catalytic performance, in which optimal sites should have "just right" binding affinities toward descriptor species, neither too strong to get poisoned nor too weak to be limited by activating stiff chemical bonds.[7] With recent advances in computing infrastructures and quantum-chemical modeling tools, e.g., density functional theory (DFT), it has become a routine practice to unravel the functional mechanisms of existing catalysts and computationally design improved ones followed by experimental validation.[8] However, the high computational cost of DFT simulations in a combinatorial and high-throughput workflow restricts the size of the chemical space and the structural complexity of active sites that can be explored, prompting the development of a new paradigm for catalytic materials discovery.

In recent years, machine learning (ML) algorithms have been increasingly used to predict energetic properties of catalytic materials, particularly for metals and metal compounds where ever-growing data sets exist in open-access repositories, e.g., Catalysis Hub, Computational Materials Repository, ioChem-BD, and Open Catalyst Project.[9−18] However, the predictive performance of purely data-driven ML models comes with the loss of physical intuition as they typically have complicated mathematical formulations which make them black boxes. As autonomous materials discovery platforms with AI agents are actively being developed in the catalysis realm, there is an immediate demand for revealing the rationale of the decision-making. In this regard, the field of interpretable ML has attracted attention over the past few years, and various strategies are implemented to improve the physical understanding arising from ML models.[19] Nevertheless, interpretability is a notoriously controversial concept, and there is no formal definition agreed upon among domain experts, ML practitioners, and algorithm developers. Broadly speaking, interpretability can be considered as the extraction of knowledge from data, while the knowledge relevancy is implied by the attained insights.[20]

In this Perspective, we critically review recent advances of interpretable ML for opening up black boxes and specifically for predicting reactivity properties of solid surfaces from the

Received: October 6, 2021
Accepted: November 15, 2021

https://doi.org/10.1021/acs.jpclett.1c03291
J. Phys. Chem. Lett. 2021, 12, 11476−11487

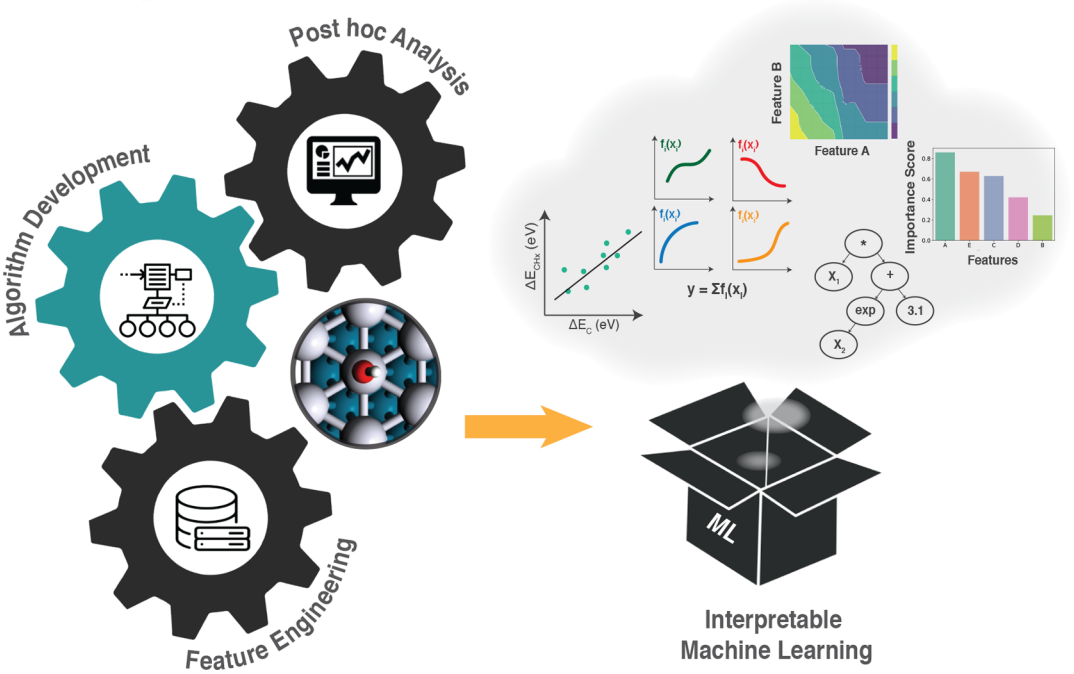


Figure 1. Interpretable machine learning of chemical bonding at solid surfaces can be achieved by feature engineering, algorithm development, and post hoc analysis.

aspects of feature engineering, algorithm development, and post hoc analysis (see Figure 1). We review the long-lasting effort of the catalysis community to find electronic, geometric, and energetic descriptors as physically intuitive feature representations of catalytic sites. We elaborate ML algorithms that are interpretable to some extent because of the constraint of model complexity or the integration of domain knowledge. We also look into post hoc analysis tools, e.g., permutation feature importance, that can provide model-agnostic interpretations. Finally, promising prospects and existing challenges regarding interpretable ML to accelerate catalytic materials discovery are discussed.

Feature Engineering. The very first step in developing ML models is to construct features that are representative of data samples. Employing informative features may simplify the relationships that need to be learned from data and allow physical interpretation of model predictions. Identifying physically transparent and relevant features using domain knowledge has been an integral part of fundamental catalysis, and many physics-inspired features were established based on the electronic, geometric, and energetic information on adsorption sites.

Electronic descriptors of site atoms include electron configurations from the periodic table of elements and more complex electronic structures that can be obtained from quantum-chemical calculations. One of the simplest descriptors of this type for transition metals and their compounds, e.g., metal oxides, is the number of valence electrons of *d*-metal atoms.[21] It is designed from the intuition that surface metal atoms interact with adsorbing species to maximize their overall stability by satisfying electron-counting rules. The *d*-band characteristics, i.e., the moments of the electronic density of *d*-states distribution projected onto site atoms, have also been widely used to capture the general trends of adsorption energies. They are considered physics-inspired and highly informative because their development is rooted in the theory of chemisorption. According to the *d*-band theory widely applied to transition-metal systems, the adsorption process can be conceptually separated into two interaction steps of adsorbate frontier orbitals with the metal *sp*-states and then *d*-states, sequentially (see Figure 2a). Because the contribution from the *sp*-states (ΔE_{sp}) is approximated as a constant for transition-metal surfaces of a given site type, the variation of binding energies is solely governed by the *d*-states (ΔE_d). Hammer and Nørskov in the 1990s introduced the *d*-band center, i.e., the first moment of the electronic density of *d*-states distribution relative to the Fermi level (E_F), as a key descriptor for understanding reactivity trends of many catalytic systems (see Figure 2b), including pristine transition-metal surfaces,[22] metal alloys,[23,24] and surfaces with structural defects (strains and steps)[23,25] or poisons/promoters.[25]

Vojvodic et al.[27] and Xin et al.[28] identified the *d*-band upper edge, defined as $\epsilon_d + W_d/2$ (W_d: *d*-bandwidth) or the maximum peak position of the Hilbert transform of the electronic density of *d*-states distribution, as an improved reactivity descriptor that explicitly considers higher-order characteristics of the *d*-band electronic structure. None of those descriptors, however, can truly capture the reactivity trends of chemical bonding involving the adsorbate of almost completely filled valence shells and site atoms of nearly fully occupied *d*-states (e.g., hydroxyl adsorption on late transition metals and their alloys). To resolve this puzzle within the *d*-band theoretical framework, interatomic coupling strength was recognized as a crucial factor governing Pauli repulsion interactions, which often dominate the *d*-band contribution to adsorption energies.[29]

While most of the electronic descriptors require self-consistent quantum-chemical calculations, geometric descriptors of surface reactivity are defined as the structural properties of an adsorption site under the influence of surroundings. One of the most intuitive geometric descriptors is the regular coordination number (CN), which is defined as the number of

https://doi.org/10.1021/acs.jpclett.1c03291
J. Phys. Chem. Lett. 2021, 12, 11476−11487

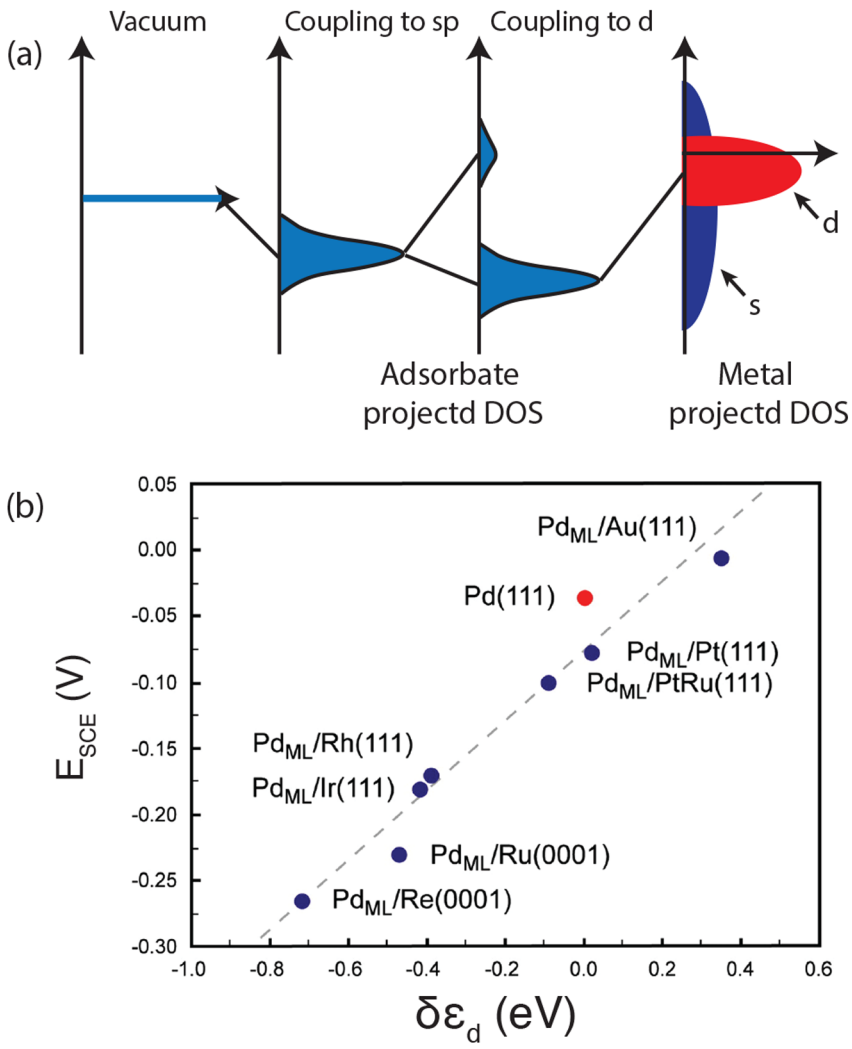


Figure 2. (a) Schematic illustration of the formation of a chemical bond between an adsorbate valence level and the *sp*- and *d*-states of a transition metal surface. (b) The changes in the hydrogen-desorption potentials for Pd overlayers on a variety of metals scale with the shift of the *d*-band center ($\delta\epsilon_d$). Adapted from ref 26. Copyright 2005, John Wiley & Sons.

atoms the site ensemble is directly bonded to. The maximum coordination for an atom in an fcc/hcp metal crystal is 12. Because of bond breaking at interfaces, surface atoms have reduced coordination numbers, e.g., 9 for the fcc {111}-facet. To make up for the lack of coordination, surface sites show the tendency toward the formation of new chemical bonds. This results in negative correlations between the CN of a site ensemble and the adsorption strength, consistent with the bond-order conservation principle.[30] Using the CN of surface metal atoms along with their curvature angles as supplemental features, Mpourmpakis et al.[31] elucidated how the size, shape, and symmetry of Au nanoparticles impact site reactivity in CO oxidation. In a similar vein, the generalized coordination number ($\overline{CN}$) of a surface site was engineered to consider the local environment of each coordinating atom.[32] The ($\overline{CN}$) of an atom i with n_i nearest neighbors is defined as $\overline{CN}_i = \sum_{j=1}^{n_i} CN_j/CN_{max}$, in which CN_{max} is the maximal coordination of a bulk atom. This descriptor provides fundamental insights into the structure−reactivity relationships that successfully guide the design of Pt cavity sites with a slightly weaker (∼0.1 eV) *OH binding than Pt(111), resulting in improved catalytic performance for oxygen reduction in fuel cells.[33] The ($\overline{CN}$) can reflect the general trends of adsorption energies on pure metal surfaces and shape-specific nanoparticles, although it is not directly applicable for describing complex systems with lattice strains and metal ligands. Ma and Xin[34] proposed the orbitalwise coordination number CN^{α} (α = s or d) as a reactivity descriptor for metal nanocatalysts, which explicitly takes into account interatomic interactions within the tight-binding theory. It is defined as $CN_i^{\alpha} = M_{2,i}^{\alpha}/(t_{nn}^{\alpha,\infty})^2$, in which α represents *s*- or *d*-orbitals, $M_{2,i}^{\alpha}$ is the second moment of the projected density of states onto the α-orbital at the site i, and $(t_{nn}^{\alpha,\infty})^2$ is the sum of the square of the α-electron hopping integrals to relevant valence orbitals of a neighboring atom in the reference bulk. Hopping integrals depend on the orbital size, shape (symmetry), and internuclear distance as approximated on the solid-state table.[35] This descriptor outperforms semiempirical bond-counting descriptors for characterizing the surface reactivity of metal nanoparticles of varying size, shape, and composition, which is attributed to its consideration of lattice strains and metal ligands (Figure 3). For transition-metal oxides, the adjusted coordination number

https://doi.org/10.1021/acs.jpclett.1c03291
J. Phys. Chem. Lett. 2021, 12, 11476−11487

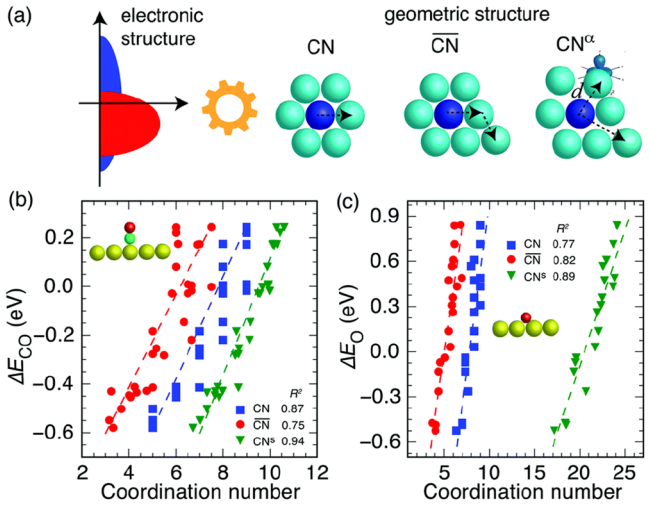


Figure 3. (a) The electronic structure of a surface metal site is linked to its coordination number descriptors. The coordination numbers of metal sites correlate with the adsorption energies of *CO (b) and *O (c) at an on-top and a hollow site of Au nanoparticles, respectively. Reprinted from ref 38. Copyright 2018, The Royal Society of Chemistry.

(ACN)[36] and bond-energy-integrated orbitalwise coordination number ($\overline{CN}^{sd}$)[37] were developed as engineered features to understand site reactivity trends, showing promise for generalizing the coordination concept to complex catalytic systems.

Another type of physics-inspired feature that can provide fundamental insights into chemisorption processes are energetic descriptors. With the adsorbate valency concept, Abild-Pedersen and colleagues showed that the binding energies of a hydrogenated species at metal surfaces can be linearly described by that of the center heavy atoms, the slope of which is the ratio of the adsorbate valencies, e.g., $(4 - x)/4$ for $*CH_x$ scaling with *C (Figure 4a).[3] This series of linear scaling relationships (LSRs) have been generalized to hydrocarbons with multiple carbon atoms and transition-metal oxides, sulfides, nitrides, and carbides.[4,39] The dependency of LSRs on site types such as terraces, steps, kinks, or metal adatoms has also been included to improve the predictive accuracy and relevancy for structure-sensitive systems.[40] For complex adsorbates with multiple functional groups, the group additivity concept has been shown to be efficient for predicting binding energies as the sum of individual molecular-fragment contributions, in analogy to the Benson group-increment theory (BGIT) originally developed for gas-phase molecules.[41] By leveraging both LSRs and group additivity formulas, energetic descriptors have been employed to predict adsorption properties of hydrocarbons, oxygenates, furanics, and aromatics on various metal surfaces (Figure 4b).[42−46]

Having highly informative features simplifies the relationship that the model must learn and allows practitioners to apply model-based interpretability approaches. Despite great contributions on physics-informed features, there are still bottlenecks for many trained ML models. Feature engineering aids the construction of such new informed feature sets by drawing on the practitioner's current domain experience as well as insights obtained from the data through exploratory data analysis. By extracting higher-order features, feature engineering may uncover potential physical information present in the raw input data and make ML models more practical.

Algorithm Development. Another strategy of achieving model interpretability is to utilize or develop algorithms that are intrinsically interpretable, i.e., algorithms being readily descriptive of the relationship between input features and the target. Regarding all forms of linear regression as being under the ML umbrella, intrinsically interpretable ML has a long history in heterogeneous catalysis. Descriptor-based models for understanding reactivity trends of solid catalysts, such as the linearized *d*-band model of chemisorption, scaling relations, group additivity, and the Brønsted−Evans−Polanyi (BEP) relationship, all fall under this type of model by learning from a typically small data set of adsorption properties.

Linear regression models can be generally written as $y = \beta_0 + \sum_{i=1}\beta_i x_i$. The betas ($\beta_i$) represent feature weights or coefficients, and β_0 is the intercept or bias. These unknown parameters can be optimized when met with data as part of the learning process. For instance, a linearized *d*-band model predicts the change in chemisorption energy $\delta\Delta E$ from one metal surface to another with two linear terms,[47] $\delta\Delta E = k_1\delta\epsilon_d + k_2\delta V_{ad}^2$. The first term denotes the covalent contribution due to the hybridization of metal *d*-states with the adsorbate resonance states that are formed after being embedded into the delocalized metal *sp*-states. The second term denotes the Pauli repulsion contribution caused by the orthogonalization of metal *d*-states and the adsorbate resonance states prior to orbital hybridization. An optimized model with learned parameters (k_1 and k_2) predicts that a positive shift in the *d*-band center leads to a more exothermic covalent interaction, while a larger coupling matrix element squared V_{ad}^2 is associated with weaker chemisorption if and only if Pauli repulsion dominates reactivity trends. The model rationalizes the observed exceptions to the traditional *d*-band model of chemisorption involving late transition metals and electron-rich adsorbates and proves to be useful in guiding catalytic materials design.[47] Montemore et al.[48] also used a linear combination of deliberately selected electronic descriptors of surface sites to build multivariate reactivtiy models of different adsorbates relevant for a wide range of chemistries (Figure 5). These descriptors include the *d*-band center (ϵ_d), the number of *p*-electrons (n_p), the coupling matrix element squared between the adsorbate state(s) and metal *d*-states (V_{ad}^2), and the filling of the *d*-band (f_d). All these descriptors can be obtained from look-up tables or estimated by submodels built upon neighboring atoms' features. The model is descriptively accurate because of its transparency on how each of these features affects the electronic descriptors and ultimately the adsorption energies.

Despite being highly transparent, linear regression models are too restrictive to accurately describe the nonlinearity in chemical bonding. To include nonlinear correlations while retaining the interpretability of linear regression models, generalized additive models (GAMs) have attracted attention recently. These models are linear combinations of nonlinear single-feature components, namely, shape functions, $y = \sum_{i=1} f_i(x_i)$.[49] A shape function can be any nonlinear relationship of choice, and the contribution from each feature can be directly quantified. Thus, it is capable of providing a descriptive understanding of model predictions. Esterhuizen et al.[50] used a decision-tree-based generalized additive model (iGAM) for unraveling factors that influence the chemisorption strength of different electron-rich and electron-poor adsorbates on model

https://doi.org/10.1021/acs.jpclett.1c03291
J. Phys. Chem. Lett. 2021, 12, 11476−11487

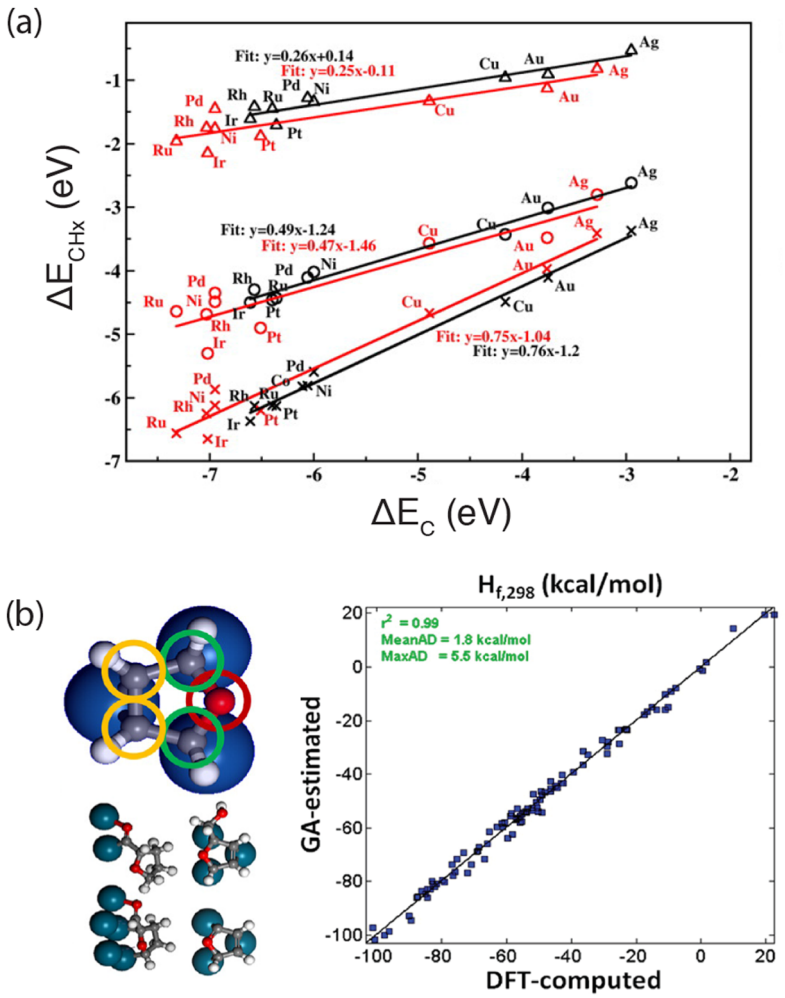


Figure 4. Energetic descriptors are physics-inspired features that are being used widely to predict chemisorption energies of adsorbate species. (a) Adsorption energies of CH_x intermediates scale with adsorption energies of the C atom. Adapted from ref 3. Copyright 2007, American Physical Society. (b) Group additivity has been used to predict adsorption properties of furanic derivatives on Pd(111). Adapted from ref 45. Copyright 2014, American Chemical Society.

alloys with subsurface ligands (Figure 6). The model supports that the strain in the surface metal layer, the number of *d*-electrons in the ligand metal, and the *d*-orbital size of the ligand metal are the essential characteristics of an adsorption site. Energy deconvolution through shape functions consolidates the knowledge that OH adsorption is less influenced by strain than other adsorbates (O, Cl, and N) because of the varying degree of site coordination and the distinct nature of chemical bonding.[47] The feature shape for the number of *d*-electrons in the ligand metal, on the other hand, sheds light on a new aspect that has not been directly explored before. It showed that the binding strength of the electron-rich adsorbates (OH, Cl, and F) becomes weaker as the filling of the *d*-states in the ligand metal increases, whereas an opposite behavior was observed for relatively electron-poor adsorbates (O, S, and CH_x), in agreement with previous studies.[29,47] The predictive accuracy of GAMs can be improved by including pairwise interactions resulting in a model known as generalized additive models plus interactions (termed GA^2 Ms).[51] Nevertheless, pairwise interactions of primary features make the algorithm convoluted and not easily explainable.

Similar to linear regression, symbolic regression is an algorithm that offers structural interpretability with analytic equations. Unlike traditional regression techniques with predefined model structures, symbolic regression optimizes both mathematical formulas and parameters while learning from data. Symbolic regression comprises enormously vast combinations of mathematical operations (+, sqrt, exp, sin, cos, log, etc.) on features. Finding the best formula becomes an optimization challenge that has been commonly tackled with genetic programming and Bayesian optimization.[52] The attained formulation can offer descriptive understanding of the underlying correlations. However, similar to any other interpretable models, a compromise between accuracy and interpretability needs to be considered. As the formula becomes more complicated (potentially more accurate), the model gets less interpretable. An example of employing symbolic regression with attention to this trade-off is the work of Weng et al.,[53] in which a simple descriptor was created to describe the OER activity of perovskite oxides (ABO_3). They found nine mathematical formulas that satisfy the requirements of being descriptive and reasonably predictive. Among those nine, μ/t showed the best compromise, where $t = (r_A + r_O)/\sqrt{2}(r_B + r_O)$ and $\mu = r_B/r_O$ are the tolerance and octahedral factors, respectively. This model provides physical insights for a rational design strategy, i.e., incorporation of

https://doi.org/10.1021/acs.jpclett.1c03291
J. Phys. Chem. Lett. 2021, 12, 11476−11487

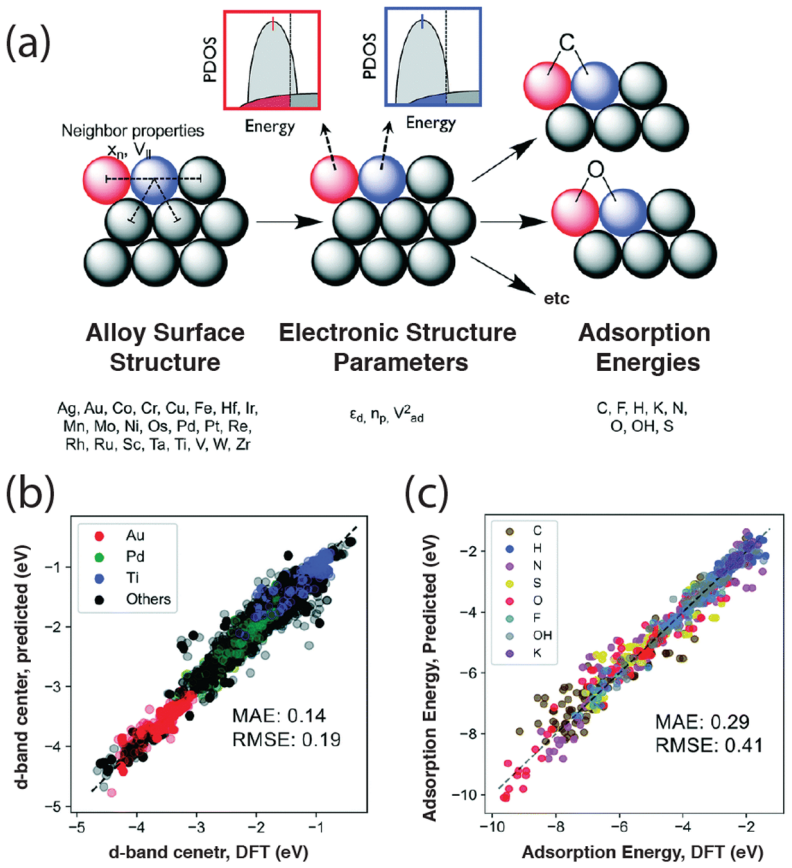


Figure 5. (a) With physical understanding of chemical bonding at surfaces, basic structural features can be chosen to predict electronic structure parameters and then adsorption energies. The general linear models based on the electronic structure properties gave good accuracy for predicting the electronic structure parameter, *d*-band center (b), and adsorption energies of various adsorbate species (c). Adapted from ref 48. Copyright 2020, The Royal Society of Chemistry.

large group IA and IIA cations (like K^+, Rb^+, Cs^{2+}, and Ca^{2+}) on the A site (increasing t) and small 3*d* TM cations on the B site (decreasing μ) resulting in candidate catalysts with improved OER activity.

Symbolic regression has shown a promising aptitude for developing interpretable ML models. However, having to consider a vast dimension of formulas and features is a weakness of this algorithm. Ideally, only a few features can be highly relevant in the sense of defining the functionality of the system,[54] not to mention incorporating sparsity into symbolic regression can make ML models relevant while being descriptive. Compressed sensing (CS), as a sparse feature selection algorithm, provides low-dimensional models that identify highly relevant descriptors and predict output targets simultaneously.[55] Two popular CS algorithms in the field of materials discovery are the least absolute shrinkage and selection operator (LASSO) and the sure-independence screening and sparsifying operator (SISSO). A specific target, chemisorption energy for instance, can be predicted by a linear combination of a collection of features (Θ), $y = \beta\Theta$. In contrast to the standard regression, LASSO applies the desire for sparsity and enforces zero contributions for some of the features by penalizing the magnitude of β coefficients.[10,12,56−58] Nonetheless, LASSO has stability issues for correlated features and the Sure SISSO algorithm has been introduced to alleviate this problem.[55] SISSO also assumes that the property of interest is a linear function of candidate descriptors that are nonlinear functions of primary features. Those high-level descriptors are derived in iterative steps by applying mathematical operators on primary features and then ranked with the degree of correlation with the target. By keeping the top ranks, SISSO can identify optimal *n*-dimensional descriptors out of the immense feature spaces. The SISSO approach has given researchers physical insights about the underlying relationship for the prediction of material properties, by providing sparse analytical models.[59,59−63] Andersen et al.[64] used this algorithm to find sparse feature representations for predicting the adsorption enthalpies of key reaction intermediates relevant to CO methanation and oxygen evolution on transition metals and their oxides. The compressed sensing technique derives algebraic models as a combination of primary features, including atomic, bulk, surface, and site properties (Figure 7). This model unbiasedly uncovered the physical factors of surface reactivity that were acquired by the *d*-band model, including the radius of the *d*-orbitals in the bulk TM (r_d), the filling and width of the *d*-band, coupling matrix element squared (V^2_{ad}), Pauling electronegativity (PE), and the density of states at the Fermi level of the *sp*- and *d*-bands. For metal oxides, the least complicated 1D SISSO model identified the *d*-bandwidth (W_d) and charge transfer energy (CTE), the energy difference between the unoccupied metal *d*-band and filled oxygen 2*p*-states, as the most important descriptors. Moreover, more complex descriptors revealed the importance of additional primary features, e.g., angular-resolved local order parameters. These features are correlated weakly with the target

https://doi.org/10.1021/acs.jpclett.1c03291
J. Phys. Chem. Lett. 2021, 12, 11476−11487

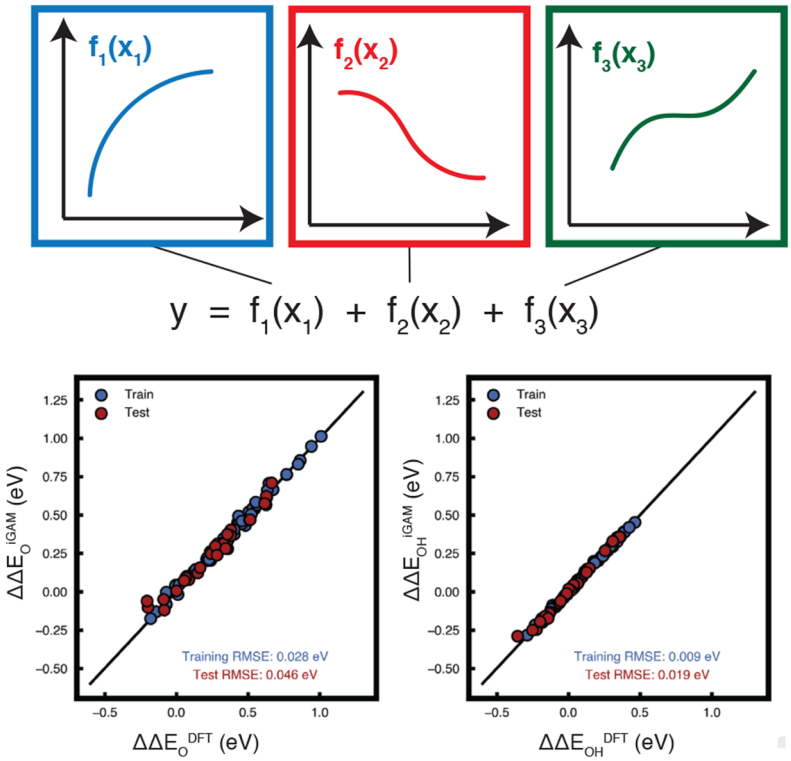


Figure 6. Additive structure of iGAMs allows for display of each feature x_i's contribution to the adsorption energies. Parity plots for DFT-calculated and iGAM-predicted adsorption energies on Pt alloys show reasonable predictive accuracy. Adapted from ref 50. Copyright 2020, Elsevier.

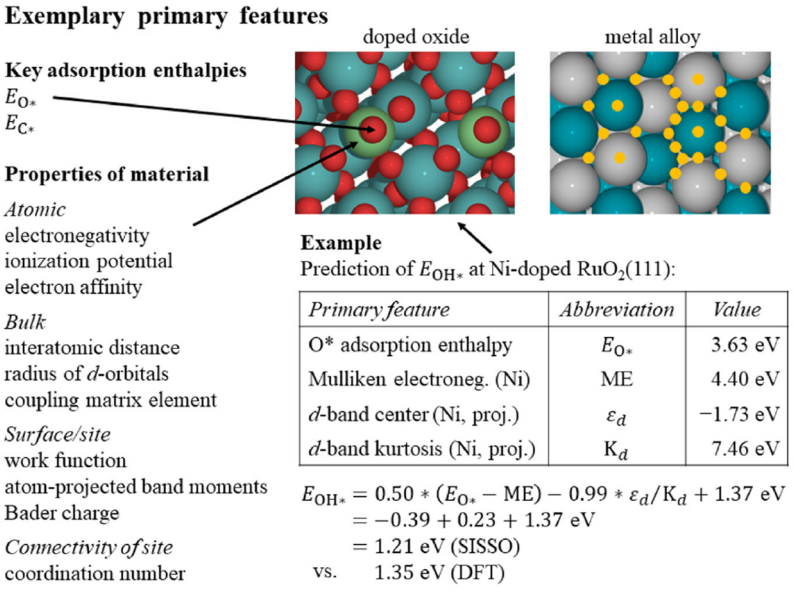


Figure 7. The SISSO algorithm uses primary features to identify descriptors for prediction of a material's properties. The adsorption energy of OH is predicted by one of the simplest recognized models for TMO surfaces. Yellow circles mark symmetry-inequivalent adsorption sites on the stepped metal alloy (top, bridge, and 3-fold- and 4-fold-coordinated sites). The primary features are averaged over these atoms if the adsorption site is made up of numerous atoms (all but top sites). Reprinted from ref 64. Copyright 2021, American Chemical Society.

individually, but they appear to be capable of capturing part of the target that electronic primary features do not if combined with other features, which emphasizes the relevance of feature interactions.

Clustering is also one type of learning algorithms with interpretability potential resulting from identifying local patterns and relationships. These unsupervised algorithms divide the data set into related subgroups. The subgroup-specific local models of adsorption energies can be more accurately descriptive compared to global models as the underlying mechanism may differ for different groups of materials (such as *d*-block metals vs *p*-block metals).[65] Identification of subgroups makes models more comprehensible by understanding similarities and differences between

https://doi.org/10.1021/acs.jpclett.1c03291
J. Phys. Chem. Lett. 2021, 12, 11476−11487

subgroups and providing a simpler pattern of the data locally. Ghiringhelli et al. were able to offer physical knowledge of the local behavior of O and OH adsorption properties of transition-metal alloys using the subgroup-discovery (SGD) local technique. Their model's boolean SG rules, also known as selectors, could reveal alloy surface sites that deviate from linear scaling relationships.[66] This approach also aided the physical interpretability of SISSO models for H binding on single-atom alloy catalysts (SAACs) by providing local trends of primary features appearing in identified descriptors of the model.[67] For instance, multiple subgroups of surfaces that bind strongly with the H atom were distinguished by common selectors, including the condition that *d*-band center of the top surface layer host metals > −0.17. Their analysis also showed that a subgroup of SAACs with strong and intermediate binding energies of H atoms is mainly regulated by the properties of the host metal rather than those of the guest metals.

Most of the interpretable ML models we discussed so far rely on the traditionally hand-crafted features identified by human experts. Depending solely on the highly specialized domain knowledge to develop features, however, may restrict the generalizability of ML models. The pursuit of AI has been pushing the exploitation of alternatives to manual feature extraction. In this aspect, representation-learning algorithms are broadly used for identifying relevant features. In a recent attempt, Esterhuizen et al.[68] applied principal component analysis (PCA), an unsupervised dimension reduction technique, to the atom-projected density of *d*-states with the goal of identifying electronic and geometric factors as algorithm-derived feature representations. This approach reduces the high-dimensional *d*-DOS into a small set of principal components (PCs) that are readily applicable in the development of more accurate ML models for predicting chemisorption energies in comparison to hand-crafted descriptors. More importantly, reconstruction of density of states signals from PC descriptors provides insights into how the materials' electronic structure, surface geometry, and composition are linked and, ultimately, affect the bonding strength. Specifically, the first PC descriptor was shown to mainly capture the effects from the size of surface and ligand metals, while the second PC descriptor is correlated with the number of valence *d*-electrons in surface metals. The convolutional neural network, or CNN for short, as a deep learning framework is also well-established for feature learning in image recognition. Over the past few years, there has been huge attention paid to representation learning to extract the high-level features from graphs. Xie et al.[69] leveraged this tool for predicting material properties using the crystal graph convolutional neural network (CGCNN), which is initiated by converting crystal structures into graphs with atoms as nodes and their connections as edges. The crystal graph encodes the atomic information and bond interactions between atoms. Convolutional layers are then built on top to get new graphs in which each node represents the local environment of the atom. Using a pooling layer at this stage, a feature vector representing the whole crystal is generated. The global vector from linear pooling becomes the key piece in the model interpretability by providing contributions from local chemical environments to the target property. This algorithm has been employed to predict surface reactivity by learning from a large data set. For instance, the original CGCNN code has been modified to collect neighbor information using Voronoi polyhedra to predict CO and H binding energies on a variety of metal alloy surfaces.[70] Fung et al.[71] developed the DOSnet models for a wide range of metal alloys and adsorbates, using the CNN algorithm to automatically featurize the electronic density of states and provide physically meaningful interpretations of reactivity trends. Although the interpretability may be compromised, representation learning is a promising ML approach to be explored for automatically gaining physically intuitive information. Integrating scientific understanding of physical interactions into the algorithms is an emerging field to generalize the models and make them interpretable. A Bayesian learning approach was developed by Wang et al.[72] based on the *d*-band reactivity theory and Newns−Anderson-type model Hamiltonians. They used a Bayesian inference algorithm (Bayeschem) to learn the model parameters on a small ab initio data set. Bayschem enables quantitative investigation of chemical bonding of simple adsorbates on metal surfaces by providing orbitalwise insights. Bayeschem prediction performance is, however, compromised by its interpretability. The theory-infused neural network (TinNet) was recently developed[73] that infuses the *d*-band theory of chemisorption into deep learning networks to predict reactivity properties of transition-metal surfaces. As shown in Figure 8, the TinNet framework contains two sequential components: a regression

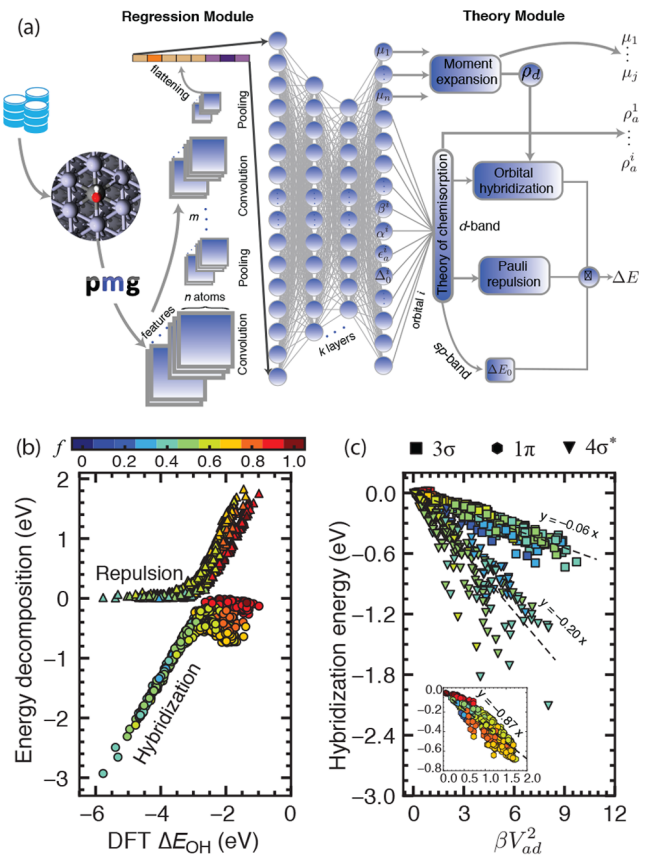


Figure 8. (a) The TinNet framework with a regression module and a theory module sequentially integrates scientific understanding of physical interactions into the learning algorithm. (b) Contributions of orbital hybridization and Pauli repulsion from metal *d*-states to *OH adsorption energies deconvoluted using TinNet models. (c) TinNet shows coupling integral squared (V_{ad}^2) for 3σ, 1π, and 4σ* orbitals linearly correlates with the corresponding orbital hybridization energy. Adapted from ref 73. Copyright 2021, Wang et al.

https://doi.org/10.1021/acs.jpclett.1c03291
J. Phys. Chem. Lett. 2021, 12, 11476−11487

module and a theory module. The input into the regression module built with convolutional neural networks is the feature representation of the adsorbate−substrate system that encodes the atomic information and bonding interactions of each atom with its neighboring environment. The output units from the regression module then serve as unknown parameters in the theory module that is built upon the *d*-band theory of chemisorption for predicting adsorption properties of a *d*-metal site. For a given adsorbate, the predicted adsorption energy from TinNet can be deconvoluted into the orbital hybridization and Pauli repulsion contributions, which can give a detailed physical interpretation not attainable from a purely regression-based model. This model also provides the projected density of states onto the adsorbate frontier orbital(s) and *d*-band moments of the adsorption site for interpretation.

Post Hoc Analysis. To this point we have discussed the interpretability that can be achieved before and during the process of learning through feature engineering and algorithm development, respectively. Statistical analysis after training may also offer model interpretation. Post hoc analysis typically separates the interpretation from the model development and has been shown to be generic in generating explanations for black box models, irrespective of learning algorithms. These approaches include visualizing the relationship between the features and output target, measuring the contribution of individual features to the model prediction, and approximating models with interpretable surrogate models. By leveraging the visualization ability of humans as our core cognitive skill, we can greatly improve the transparency of ML models.[74] Partial dependence plots (PDPs), individual conditional expectation (ICE) plots,[75] and accumulated local effects (ALEs)[76,77] are examples of visualization techniques that have been used to explain ML models. For instance, Liu et al.[78] used PDPs to interpret tree ensemble regression models of the interactions between small molecules and group 13 metal-oxide surfaces, revealing that higher HOMO energies, surface energies, and dipole moments statistically lead to stronger molecular bindings.

Feature importance analysis determines which features have the most impact on model predictions. Permutation feature importance (PFI) originally introduced for random-forest models measures the change in prediction error by permuting feature values and is the most prevalent technique for computing and displaying feature contributions.[80] Li et al.[79] utilized sensitivity analysis based on feature permutation to physically comprehend and explain the underlying variables that influence adsorbate−metal interactions in feed-forward neural network models. This sensitivity analysis highlighted the significant role of the lower moments of the metal *d*-band in site reactivity in comparison to higher moments, a higher dependency of CO adsorption on *d*-band features than OH adsorption, and a notable effect of *sp*-band properties on OH adsorption (Figure 9). Fung et al.[71] also used a similar approach to obtain insights into which parts of the DOS are accountable for the prediction from the deep neural network model (DOSnet) with the full DOS as inputs. Looking into hydrogen adsorption on Pt, they observed a decrease in the bonding strength if low-energy states are masked and an increase when masking the high-energy states. Other local and global feature analysis metrics have also been introduced as post hoc techniques to explain ML models, e.g., leave-one-covariate-out (LOCO)[81] and Shapley value.[82] Surrogate

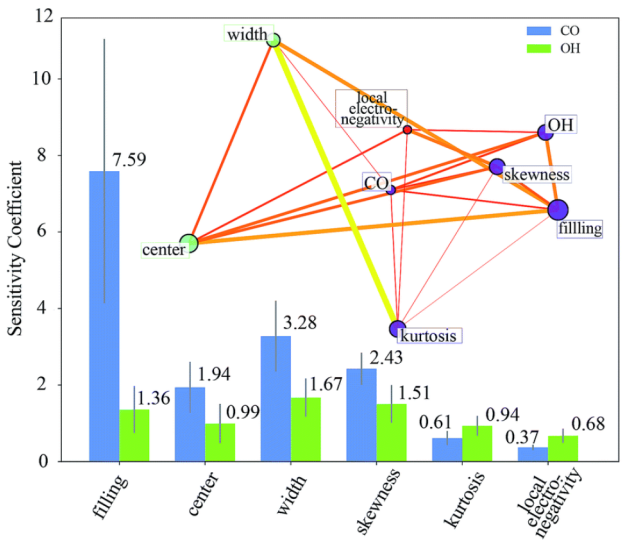


Figure 9. Feature importance scores for *CO and *OH adsorption models based on the sensitivity analysis of neural network models imply an important role for the lower moments of the metal *d*-band in site reactivity in comparison to higher moments and the distinct bonding character of two adsorbates. Reprinted from ref 79. Copyright 2017, The Royal Society of Chemistry.

models, generated by training a simpler model with the original model's input data and predictions, can give insights into complicated black box models.[83] Local surrogate models like LIME (local interpretable model-agnostic explanations) focus on explaining why specific predictions were produced and identifying which parts of a given input contribute the most to a prediction.[84] SHAP (SHapley Additive exPlanations) is also a surrogate model inspired by LIME to explain the prediction of an instance by computing the importance rating for each feature for a specific prediction based on the Shapley values.[85,86]

In summary, feature engineering can help with the relevance of models by introducing informative, physics-based features. Algorithm-based interpretability may give undistorted explanations of the chemisorption process by introducing constraints into the structural form of ML models. A trade-off between model predictive accuracy and interpretation integrity is often manifested in algorithm development. Integration of domain knowledge into ML algorithms can possibly break this trade-off and offer further insights into the nature of chemical bonding. Post hoc interpretability allows maintaining the high predictive accuracy of ML models while providing explanations on how the predictions are made. Despite significant progress in all three aspects, many challenges remain to be addressed on the path toward truly interpretable models for driving catalytic materials discovery. Feature engineering is still limited by the lack of understanding of the electronic structure and its relationship with chemical bonding, particularly for complex materials. One promising direction in engineering of new informative features is addressing this shortfall by looking into high-order features, e.g., convolutions of the electron density.[87] Algorithm development with integrated domain theories heavily relies on the improvement of network architectures and physical models. It is critical to apply approaches that transform the network so that its essential building blocks, i.e., the architecture and underlying calculations, become more understandable by humans. Network compression methods

https://doi.org/10.1021/acs.jpclett.1c03291
J. Phys. Chem. Lett. 2021, 12, 11476−11487

such as pruning and distillation are examples of approaches that can be employed to gain smaller and more interpretable networks.[88−90] TinNet, integrating a Newns−Anderson physical model, is demonstrated to work very well for transition metals; however, its generalization to *sp*-metals and metal compounds remains challenging. Thus, it is vital to focus on the generalizability of the model while developing future theory-integrated models. Despite the availability of many powerful post hoc tools for the interpretation of black box models, their use in the field has been limited. It is also beneficial to explore various post hoc approaches to get more information out of the black box models that have previously been trained. Nonetheless, these approaches should be used cautiously as they may raise the risks and concerns of generating explanations that are products of artifacts learned by the model rather than genuine information from the data. Even so, interpretability should be an integral part of ML that can push forward catalysis science. Given the current challenges and technological hurdles in data availability, standardization, and sharing mechanisms, the future looks extremely bright indeed for deploying an explainable ML framework that is directly interpretable, tractable, and trustworthy. If successful, an AI agent can be built to discover plausible models from data and automatically present its findings as user-centered, human-level explanations. With all that excitement, developing quantifiable metrics is the key toward interpretable ML for catalytic materials discovery. Establishing such metrics should be considered an essential part of future efforts in interpretable ML frameworks.

AUTHOR INFORMATION

Corresponding Author

Hongliang Xin − *Department of Chemical Engineering, Virginia Polytechnic Institute and State University, Blacksburg, Virginia 24061, United States;* orcid.org/0000-0001-9344-1697; Email: hxin@vt.edu

Authors

Noushin Omidvar − *Department of Chemical Engineering, Virginia Polytechnic Institute and State University, Blacksburg, Virginia 24061, United States;* orcid.org/0000-0001-6766-8548

Hemanth S. Pillai − *Department of Chemical Engineering, Virginia Polytechnic Institute and State University, Blacksburg, Virginia 24061, United States;* orcid.org/0000-0003-3131-7396

Shih-Han Wang − *Department of Chemical Engineering, Virginia Polytechnic Institute and State University, Blacksburg, Virginia 24061, United States;* orcid.org/0000-0003-4418-2080

Tianyou Mou − *Department of Chemical Engineering, Virginia Polytechnic Institute and State University, Blacksburg, Virginia 24061, United States;* orcid.org/0000-0002-7389-6712

Siwen Wang − *Department of Chemical Engineering, Virginia Polytechnic Institute and State University, Blacksburg, Virginia 24061, United States;* orcid.org/0000-0003-3582-5398

Andy Athawale − *Department of Chemical Engineering, Virginia Polytechnic Institute and State University, Blacksburg, Virginia 24061, United States;* orcid.org/0000-0003-3685-5898

Luke E. K. Achenie − *Department of Chemical Engineering, Virginia Polytechnic Institute and State University, Blacksburg, Virginia 24061, United States;* orcid.org/0000-0001-9850-5346

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jpclett.1c03291

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

N.O., H.S.P., S.-H.W., T.M., S.W., A.A., L.E.K.A., and H.X. acknowledge the partial financial support from the NSF CAREER program (CBET-1845531). The computational resource used in this work is provided by Advanced Research Computing at Virginia Polytechnic Institute and State University.

REFERENCES

(1) Nilsson, A.; Pettersson, L. G.; Norskov, J. *Chemical bonding at surfaces and interfaces*; Elsevier, 2011.

(2) Chorkendorff, I.; Niemantsverdriet, J. W. *Concepts of modern catalysis and kinetics*; John Wiley & Sons, 2017.

(3) Abild-Pedersen, F.; Greeley, J.; Studt, F.; Rossmeisl, J.; Munter, T. R.; Moses, P. G.; Skulason, E.; Bligaard, T.; Nørskov, J. K. Scaling properties of adsorption energies for hydrogen-containing molecules on transition-metal surfaces. *Phys. Rev. Lett.* **2007**, *99*, 016105.

(4) Fernández, E. M.; Moses, P. G.; Toftelund, A.; Hansen, H. A.; Martínez, J. I.; Abild-Pedersen, F.; Kleis, J.; Hinnemann, B.; Rossmeisl, J.; Bligaard, T.; et al. Scaling relationships for adsorption energies on transition metal oxide, sulfide, and nitride surfaces. *Angew. Chem., Int. Ed.* **2008**, *47*, 4683−4686.

(5) Wang, S.; Temel, B.; Shen, J.; Jones, G.; Grabow, L. C.; Studt, F.; Bligaard, T.; Abild-Pedersen, F.; Christensen, C. H.; Nørskov, J. K. Universal brønsted-evans-polanyi relations for c−c, c−o, c−n, n−o, n−n, and o−o dissociation reactions. *Catal. Lett.* **2011**, *141*, 370−373.

(6) Jones, G.; Studt, F.; Abild-Pedersen, F.; Nørskov, J. K.; Bligaard, T. Scaling relationships for adsorption energies of C2 hydrocarbons on transition metal surfaces. *Chem. Eng. Sci.* **2011**, *66*, 6318−6323.

(7) Medford, A. J.; Vojvodic, A.; Hummelshøj, J. S.; Voss, J.; Abild-Pedersen, F.; Studt, F.; Bligaard, T.; Nilsson, A.; Nørskov, J. K. From the Sabatier principle to a predictive theory of transition-metal heterogeneous catalysis. *J. Catal.* **2015**, *328*, 36−42.

(8) Zhao, Z.-J.; Liu, S.; Zha, S.; Cheng, D.; Studt, F.; Henkelman, G.; Gong, J. Theory-guided design of catalytic materials using scaling relationships and reactivity descriptors. *Nature Reviews Materials* **2019**, *4*, 792−804.

(9) Tian, H.; Rangarajan, S. Predicting adsorption energies using multifidelity data. *J. Chem. Theory Comput.* **2019**, *15*, 5588−5600.

(10) Chowdhury, A. J.; Yang, W.; Walker, E.; Mamun, O.; Heyden, A.; Terejanu, G. A. Prediction of adsorption energies for chemical species on metal catalyst surfaces using machine learning. *J. Phys. Chem. C* **2018**, *122*, 28142−28150.

(11) Noh, J.; Back, S.; Kim, J.; Jung, Y. Active learning with non-ab initio input features toward efficient CO 2 reduction catalysts. *Chemical science* **2018**, *9*, 5152−5159.

(12) Hoyt, R. A.; Montemore, M. M.; Fampiou, I.; Chen, W.; Tritsaris, G.; Kaxiras, E. Machine learning prediction of H adsorption energies on Ag alloys. *J. Chem. Inf. Model.* **2019**, *59*, 1357−1365.

(13) Ulissi, Z. W.; Tang, M. T.; Xiao, J.; Liu, X.; Torelli, D. A.; Karamad, M.; Cummins, K.; Hahn, C.; Lewis, N. S.; Jaramillo, T. F.; et al. Machine-learning methods enable exhaustive searches for active bimetallic facets and reveal active site motifs for CO2 reduction. *ACS Catal.* **2017**, *7*, 6600−6608.

(14) Winther, K. T.; Hoffmann, M. J.; Boes, J. R.; Mamun, O.; Bajdich, M.; Bligaard, T. Catalysis-Hub. org, an open electronic structure database for surface reactions. *Sci. Data* **2019**, *6*, 1−10.

(15) Landis, D. D.; Hummelshøj, J. S.; Nestorov, S.; Greeley, J.; Dułak, M.; Bligaard, T.; Nørskov, J. K.; Jacobsen, K. W. The computational materials repository. *Comput. Sci. Eng.* **2012**, *14*, 51−57.
(16) Saal, J. E.; Kirklin, S.; Aykol, M.; Meredig, B.; Wolverton, C. Materials design and discovery with high-throughput density functional theory: the open quantum materials database (OQMD). *JOM* **2013**, *65*, 1501−1509.
(17) Álvarez-Moreno, M.; de Graaf, C.; Lopez, N.; Maseras, F.; Poblet, J. M.; Bo, C. Managing the computational chemistry big data problem: the ioChem-BD platform. *J. Chem. Inf. Model.* **2015**, *55*, 95−103.
(18) Chanussot, L.; Das, A.; Goyal, S.; Lavril, T.; Shuaibi, M.; Riviere, M.; Tran, K.; Heras-Domingo, J.; Ho, C.; Hu, W.; et al. Open Catalyst 2020 (OC20) Dataset and Community Challenges. *ACS Catal.* **2021**, *11*, 6059−6072.
(19) Schmidt, J.; Marques, M. R.; Botti, S.; Marques, M. A. Recent advances and applications of machine learning in solid-state materials science. *npj Computational Materials* **2019**, *5*, 83.
(20) Murdoch, W. J.; Singh, C.; Kumbier, K.; Abbasi-Asl, R.; Yu, B. Definitions, methods, and applications in interpretable machine learning. *Proc. Natl. Acad. Sci. U. S. A.* **2019**, *116*, 22071−22080.
(21) Calle-Vallejo, F.; Inoglu, N. G.; Su, H.-Y.; Martinez, J. I.; Man, I. C.; Koper, M. T.; Kitchin, J. R.; Rossmeisl, J. Number of outer electrons as descriptor for adsorption processes on transition metals and their oxides. *Chemical Science* **2013**, *4*, 1245−1249.
(22) Hammer, B.; Nørskov, J. K. Theoretical surface science and catalysis—calculations and concepts. *Adv. Catal.* **2000**, *45*, 71−129.
(23) Kitchin, J. R.; Nørskov, J. K.; Barteau, M. A.; Chen, J. Role of strain and ligand effects in the modification of the electronic and chemical properties of bimetallic surfaces. *Phys. Rev. Lett.* **2004**, *93*, 156801.
(24) Kitchin, J.; Nørskov, J. K.; Barteau, M.; Chen, J. Modification of the surface electronic and chemical properties of Pt (111) by subsurface 3d transition metals. *J. Chem. Phys.* **2004**, *120*, 10240−10246.
(25) Abild-Pedersen, F.; Greeley, J.; Nørskov, J. K. Understanding the effect of steps, strain, poisons, and alloying: methane activation on Ni surfaces. *Catal. Lett.* **2005**, *105*, 9−13.
(26) Kibler, L. A.; El-Aziz, A. M.; Hoyer, R.; Kolb, D. M. Tuning reaction rates by lateral strain in a palladium monolayer. *Angew. Chem., Int. Ed.* **2005**, *44*, 2080−2084.
(27) Vojvodic, A.; Nørskov, J.; Abild-Pedersen, F. Electronic structure effects in transition metal surface chemistry. *Top. Catal.* **2014**, *57*, 25−32.
(28) Xin, H.; Vojvodic, A.; Voss, J.; Nørskov, J. K.; Abild-Pedersen, F. Effects of d-band shape on the surface reactivity of transition-metal alloys. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2014**, *89*, 115114.
(29) Xin, H.; Linic, S. Communications: Exceptions to the d-band model of chemisorption on metal surfaces: The dominant role of repulsion between adsorbate states and metal d-states. *J. Chem. Phys.* **2010**, *132*, 221101.
(30) Shustorovich, E. The bond-order conservation approach to chemisorption and heterogeneous catalysis: applications and implications. *Adv. Catal.* **1990**, *37*, 101.
(31) Mpourmpakis, G.; Andriotis, A. N.; Vlachos, D. G. Identification of descriptors for the CO interaction with metal nanoparticles. *Nano Lett.* **2010**, *10*, 1041−1045.
(32) Calle-Vallejo, F.; Martínez, J. I.; García-Lastra, J. M.; Sautet, P.; Loffreda, D. Fast prediction of adsorption properties for platinum nanocatalysts with generalized coordination numbers. *Angew. Chem., Int. Ed.* **2014**, *53*, 8316−8319.
(33) Calle-Vallejo, F.; Tymoczko, J.; Colic, V.; Vu, Q. H.; Pohl, M. D.; Morgenstern, K.; Loffreda, D.; Sautet, P.; Schuhmann, W.; Bandarenka, A. S. Finding optimal surface sites on heterogeneous catalysts by counting nearest neighbors. *Science* **2015**, *350*, 185−189.
(34) Ma, X.; Xin, H. Orbitalwise coordination number for predicting adsorption properties of metal nanocatalysts. *Phys. Rev. Lett.* **2017**, *118*, 036101.
(35) Harrison, W. A. *Electronic structure and the properties of solids: the physics of the chemical bond*; Courier Corporation, 2012.
(36) Fung, V.; Tao, F. F.; Jiang, D.-e. General structure−reactivity relationship for oxygen on transition-metal oxides. *J. Phys. Chem. Lett.* **2017**, *8*, 2206−2211.
(37) Wu, D.; Dong, C.; Zhan, H.; Du, X.-W. Bond-energy-integrated descriptor for oxygen electrocatalysis of transition metal oxides. *J. Phys. Chem. Lett.* **2018**, *9*, 3387−3391.
(38) Wang, S.; Omidvar, N.; Marx, E.; Xin, H. Coordination numbers for unraveling intrinsic size effects in gold-catalyzed CO oxidation. *Phys. Chem. Chem. Phys.* **2018**, *20*, 6055−6059.
(39) Vojvodic, A.; Hellman, A.; Ruberto, C.; Lundqvist, B. I. From electronic structure to catalytic activity: A single descriptor for adsorption and reactivity on transition-metal carbides. *Phys. Rev. Lett.* **2009**, *103*, 146103.
(40) Calle-Vallejo, F.; Loffreda, D.; Koper, M. T.; Sautet, P. Introducing structural sensitivity into adsorption−energy scaling relations by means of coordination numbers. *Nat. Chem.* **2015**, *7*, 403−410.
(41) Benson, S. W.; Cruickshank, F.; Golden, D.; Haugen, G. R.; O'neal, H.; Rodgers, A.; Shaw, R.; Walsh, R. Additivity rules for the estimation of thermochemical properties. *Chem. Rev.* **1969**, *69*, 279−324.
(42) Kua, J.; Faglioni, F.; Goddard, W. A. Thermochemistry for Hydrocarbon Intermediates Chemisorbed on Metal Surfaces: CH n-m (CH3) m with n= 1, 2, 3 and m ≤ n on Pt, Ir, Os, Pd, Rh, and Ru. *J. Am. Chem. Soc.* **2000**, *122*, 2309−2321.
(43) Salciccioli, M.; Chen, Y.; Vlachos, D. G. Density functional theory-derived group additivity and linear scaling methods for prediction of oxygenate stability on metal catalysts: adsorption of open-ring alcohol and polyol dehydrogenation intermediates on Pt-based metals. *J. Phys. Chem. C* **2010**, *114*, 20155−20166.
(44) Salciccioli, M.; Edie, S.; Vlachos, D. Adsorption of acid, ester, and ether functional groups on Pt: fast prediction of thermochemical properties of adsorbed oxygenates via DFT-based group additivity methods. *J. Phys. Chem. C* **2012**, *116*, 1873−1886.
(45) Vorotnikov, V.; Wang, S.; Vlachos, D. G. Group additivity for estimating thermochemical properties of furanic compounds on Pd (111). *Ind. Eng. Chem. Res.* **2014**, *53*, 11929−11938.
(46) Vorotnikov, V.; Vlachos, D. G. Group additivity and modified linear scaling relations for estimating surface thermochemistry on transition metal surfaces: Application to furanics. *J. Phys. Chem. C* **2015**, *119*, 10417−10426.
(47) Xin, H.; Holewinski, A.; Linic, S. Predictive structure−reactivity models for rapid screening of Pt-based multimetallic electrocatalysts for the oxygen reduction reaction. *ACS Catal.* **2012**, *2*, 12−16.
(48) Montemore, M. M.; Nwaokorie, C. F.; Kayode, G. O. General screening of surface alloys for catalysis. *Catal. Sci. Technol.* **2020**, *10*, 4467−4476.
(49) Hastie, T. J.; Tibshirani, R. J. *Generalized additive models*; Routledge, 2017.
(50) Esterhuizen, J. A.; Goldsmith, B. R.; Linic, S. Theory-Guided Machine Learning Finds Geometric Structure-Property Relationships for Chemisorption on Subsurface Alloys. *Chem.* **2020**, *6*, 3100−3117.
(51) Lou, Y.; Caruana, R.; Gehrke, J.; Hooker, G. Accurate intelligible models with pairwise interactions. *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*; 2013; pp 623−631.
(52) Yin, W.-J. Density functional theory-free descriptor for the practical discovery of perovskite catalysts. *Comput. Mater. Sci.* **2021**, *193*, 110342.
(53) Weng, B.; Song, Z.; Zhu, R.; Yan, Q.; Sun, Q.; Grice, C. G.; Yan, Y.; Yin, W.-J. Simple descriptor derived from symbolic regression accelerating the discovery of new perovskite catalysts. *Nat. Commun.* **2020**, *11*, 3513.
(54) Brunton, S. L.; Proctor, J. L.; Kutz, J. N. Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proc. Natl. Acad. Sci. U. S. A.* **2016**, *113*, 3932−3937.

https://doi.org/10.1021/acs.jpclett.1c03291
J. Phys. Chem. Lett. 2021, 12, 11476−11487

(55) 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. *Physical Review Materials* **2018**, *2*, 083802.
(56) Dean, J.; Taylor, M. G.; Mpourmpakis, G. Unfolding adsorption on metal nanoparticles: Connecting stability with catalysis. *Science advances* **2019**, *5*, No. eaax5101.
(57) Bucior, B. J.; Bobbitt, N. S.; Islamoglu, T.; Goswami, S.; Gopalan, A.; Yildirim, T.; Farha, O. K.; Bagheri, N.; Snurr, R. Q. Energy-based descriptors to rapidly predict hydrogen storage in metal−organic frameworks. *Molecular Systems Design & Engineering* **2019**, *4*, 162−174.
(58) 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 Computational Materials* **2020**, *6*, 102.
(59) Fung, V.; Hu, G.; Sumpter, B. Electronic band contraction induced low temperature methane activation on metal alloys. *J. Mater. Chem. A* **2020**, *8*, 6057−6066.
(60) Xu, W.; Andersen, M.; Reuter, K. Data-Driven Descriptor Engineering and Refined Scaling Relations for Predicting Transition Metal Oxide Reactivity. *ACS Catal.* **2021**, *11*, 734−742.
(61) Zhao, W.; Chen, L.; Zhang, W.; Yang, J. Single Mo 1 (W 1, Re 1) atoms anchored in pyrrolic-N 3 doped graphene as efficient electrocatalysts for the nitrogen reduction reaction. *J. Mater. Chem. A* **2021**, *9*, 6547−6554.
(62) Bartel, C. J.; Sutton, C.; Goldsmith, B. R.; Ouyang, R.; Musgrave, C. B.; Ghiringhelli, L. M.; Scheffler, M. New tolerance factor to predict the stability of perovskite oxides and halides. *Science advances* **2019**, *5*, No. eaav0693.
(63) Deimel, M.; Reuter, K.; Andersen, M. Active Site Representation in First-Principles Microkinetic Models: Data-Enhanced Computational Screening for Improved Methanation Catalysts. *ACS Catal.* **2020**, *10*, 13729−13736.
(64) Andersen, M.; Reuter, K. Adsorption Enthalpies for Catalysis Modeling through Machine-Learned Descriptors. *Acc. Chem. Res.* **2021**, *54*, 2741.
(65) Goldsmith, B. R.; Boley, M.; Vreeken, J.; Scheffler, M.; Ghiringhelli, L. M. Uncovering structure-property relationships of materials by subgroup discovery. *New J. Phys.* **2017**, *19*, 013031.
(66) Foppa, L.; Ghiringhelli, L. M. Identifying outstanding transition-metal-alloy heterogeneous catalysts for the oxygen reduction and evolution reactions via subgroup discovery. *Top. Catal.* **2021**, DOI: 10.1007/s11244-021-01502-4.
(67) Han, Z.-K.; Sarker, D.; Ouyang, R.; Mazheika, A.; Gao, Y.; Levchenko, S. V. Single-atom alloy catalysts designed by first-principles calculations and artificial intelligence. *Nat. Commun.* **2021**, *12*, 1833.
(68) 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. Catalysis* **2021**, *1*, 923−940.
(69) Xie, T.; Grossman, J. C. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Phys. Rev. Lett.* **2018**, *120*, 145301.
(70) Back, S.; Yoon, J.; Tian, N.; Zhong, W.; Tran, K.; Ulissi, Z. W. Convolutional neural network of atomic surface structures to predict binding energies for high-throughput screening of catalysts. *J. Phys. Chem. Lett.* **2019**, *10*, 4401−4408.
(71) Fung, V.; Hu, G.; Ganesh, P.; Sumpter, B. G. Machine learned features from density of states for accurate adsorption energy prediction. *Nat. Commun.* **2021**, *12*, 88.
(72) Wang, S.; Pillai, H. S.; Xin, H. Bayesian learning of chemisorption for bridging the complexity of electronic descriptors. *Nat. Commun.* **2020**, *11*, 6132.
(73) Wang, S.-H.; Pillai, H. S.; Wang, S.; Achenie, L. E.; Xin, H. Infusing theory into deep learning for interpretable reactivity prediction. *Nat. Commun.* **2021**, *12*, 5288.
(74) Vellido, A. The importance of interpretability and visualization in machine learning for applications in medicine and health care. *Neural computing and applications* **2020**, *32*, 18069−18083.
(75) Goldstein, A.; Kapelner, A.; Bleich, J.; Pitkin, E. Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *journal of Computational and Graphical Statistics* **2015**, *24*, 44−65.
(76) 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)* **2020**, *82*, 1059−1086.
(77) Ivonina, M. V.; Orimoto, Y.; Aoki, Y. Quantum chemistry−machine learning approach for predicting and elucidating molecular hyperpolarizability: Application to [2.2] paracyclophane-containing push−pull polymers. *J. Chem. Phys.* **2021**, *154*, 124107.
(78) Liu, C.; Li, Y.; Takao, M.; Toyao, T.; Maeno, Z.; Kamachi, T.; Hinuma, Y.; Takigawa, I.; Shimizu, K.-i. Frontier molecular orbital based analysis of solid−adsorbate interactions over Group 13 metal oxide surfaces. *J. Phys. Chem. C* **2020**, *124*, 15355−15365.
(79) 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* **2017**, *5*, 24131−24138.
(80) Breiman, L. Random forests. *Machine learning* **2001**, *45*, 5−32.
(81) Lei, J.; G'Sell, M.; Rinaldo, A.; Tibshirani, R. J.; Wasserman, L. Distribution-free predictive inference for regression. *J. Am. Stat. Assoc.* **2018**, *113*, 1094−1111.
(82) Shapley, L. S. In *Contributions to the Theory of Games (AM-28), Vol. II*; Kuhn, H. W., Tucker, A. W., Eds.; Princeton University Press: Princeton, 1953; pp 307−318.
(83) Molnar, C.; Casalicchio, G.; Bischl, B. Interpretable machine learning−a brief history, state-of-the-art and challenges. *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*; 2020; pp 417−431.
(84) Ribeiro, M. T.; Singh, S.; Guestrin, C. "Why should i trust you?" Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*; 2016; pp 1135−1144.
(85) Lundberg, S. M.; Lee, S.-I. A unified approach to interpreting model predictions. *Proceedings of the 31st international conference on neural information processing systems*; 2017; pp 4768−4777.
(86) Zhang, S.; Lu, T.; Xu, P.; Tao, Q.; Li, M.; Lu, W. Predicting the Formability of Hybrid Organic−Inorganic Perovskites via an Interpretable Machine Learning Strategy. *J. Phys. Chem. Lett.* **2021**, *12*, 7423−7430.
(87) Lei, X.; Medford, A. J. Design and analysis of machine learning exchange-correlation functionals via rotationally invariant convolutional descriptors. *Physical Review Materials* **2019**, *3*, 063801.
(88) Han, S.; Mao, H.; Dally, W. J. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. *arXiv* **2015**, 1510.00149.
(89) Hinton, G.; Vinyals, O.; Dean, J. Distilling the knowledge in a neural network. *arXiv* **2015**, 1503.02531.
(90) Abbasi-Asl, R.; Yu, B. Structural Compression of Convolutional Neural Networks with Applications in Interpretability. *Frontiers in Big Data* **2021**, *4*. DOI: 10.3389/fdata.2021.704182