Trends in Chemistry

CellPress

Spotlight

Algorithm-derived feature representations for explainable AI in catalysis

Noushin Omidvar[1] and Hongliang Xin[1,*]

Machine learning (ML) has emerged as a critical tool in catalysis, attributed to its capability of finding complex patterns in high dimensional and heterogeneous data. A recently published article in *Chem Catalysis* (Esterhuizen *et al*.) used unsupervised ML for uncovering electronic and geometric descriptors of the surface reactivity of metal alloys and oxides.

With recent advances in computing infrastructures and quantum-chemical modeling tools, a large amount of data on the energetic properties of catalytic materials are becoming available through open-access repositories (e.g., Catalysis Hub [1], ioChem-BD [2], Computational Materials Repository [3], and Open Catalyst Project [4]). These massive data resources, if fully harnessed, can potentially bring viable routes toward autonomous catalyst design strategies with artificial intelligence (AI). To date, researchers have developed a variety of machine learning (ML) models for predicting the catalytic properties of materials (e.g., adsorption energies as reactivity descriptors in Sabatier volcano plots [5]). However, as the algorithms become increasingly complicated, humans are challenged to explain how the black-box models work and how they come to a specific prediction. Recently, significant works have been undertaken in catalysis to address this challenge. Explainable artificial intelligence (XAI) concepts, algorithms, and tools were used to develop interpretable ML models for the reactivity properties prediction of solid surfaces, with the goal of attaining new scientific insights.

In a recent paper in *Chem Catalysis*, Esterhuizen and colleagues [6] used a combination of XAI tools to automate the discovery of electronic and geometric descriptors for predicting the surface reactivity of metal alloys and oxides. They exploited the merits of unsupervised ML algorithms [e.g., principal component analysis (PCA)] to derive dimensionality-reduced feature representations of the atom projected *d*-electronic density of states (DOS) (Figure 1). These PCA-derived principal components (PCs) lead to accurate regression models of O, C, N, and H adsorption energies on layered metal alloys, outperforming the models that directly use traditional *d*-band descriptors (filling, center, width, upper edge, skewness, and kurtosis) and on par with the models using the complete DOS. More importantly, partial dependence plots offer physical explainability of PCs, with the observation that the chemisorption of those four adsorbates becomes stronger as the scores of the first PC increase, while their bindings become weaker with the increase of the second PC descriptor. Notably, how the PCs capture the electronic structure effect was also elaborated by explanations from the reconstruction of DOS as a function of each PC descriptor score via a signal reconstruction technique (Figure 1). With the decrease of the first PC, the reconstructed DOS is shown to have a wider *d*-band and lower *d*-band center due to a higher degree of orbital overlap. The reconstruction of DOS based on the second PC score, however, explains changes in the higher-order *d*-band moment predominantly. With higher second PC scores, the number of *d*-electrons increases. The PCA-derived feature representations can also connect the composition and geometry of surface alloys with adsorption energies. Varying surface and subsurface ligand metals, they revealed that the ligand metal's size is the main geometric factor that influences the first PC score. Increasing the *d*-orbital radius of ligand metals is associated with a lower first PC and higher degree of orbital overlap and thus results in a broader and lower-energy *d*-band. This ultimately causes a weaker adsorption. In addition, the number of valence *d*-electrons in the surface metals mainly determines the second PC and its increase leads to a higher second PC and lower upper band edge relative to the Fermi level and weaker chemisorption. Exploring the strain effect on the descriptors and the chemisorption strength also made it clear that the strain has a substantial impact on the first PC, implying that the first PC contains information about the degree of orbital overlap between metal atoms. However, the geometric strain has little effect on the second PC, which confirms that the second PC mainly reflects the valence *d*-electron number of the surface metal atoms. Interestingly, this explainable ML framework is not limited to layered alloys and can be extended to other catalytic systems. It was shown that it could derive feature representations of surface oxygen reactivity of metals, rutile metal oxides, and perovskite metal oxides, which capture the relative occupations of the O–M bonding and antibonding orbitals. This work showcased that unsupervised ML of informative features as an XAI approach can simplify the relationship to be learned from data and draw physical insights for guiding catalyst design.

The importance of this PCA approach particularly comes into the light as building intrinsically interpretable models, models that can talk for themselves. Regarding all forms of linear regression as being under the ML umbrella, there is a long history in heterogeneous catalysis in developing descriptor-based reactivity models, such as the *d*-band theory of chemisorption [7], scaling relations [8], and group additivity [9]. Montemore and colleagues [16] used a linear combination of deliberately selected electronic structure descriptors of metal

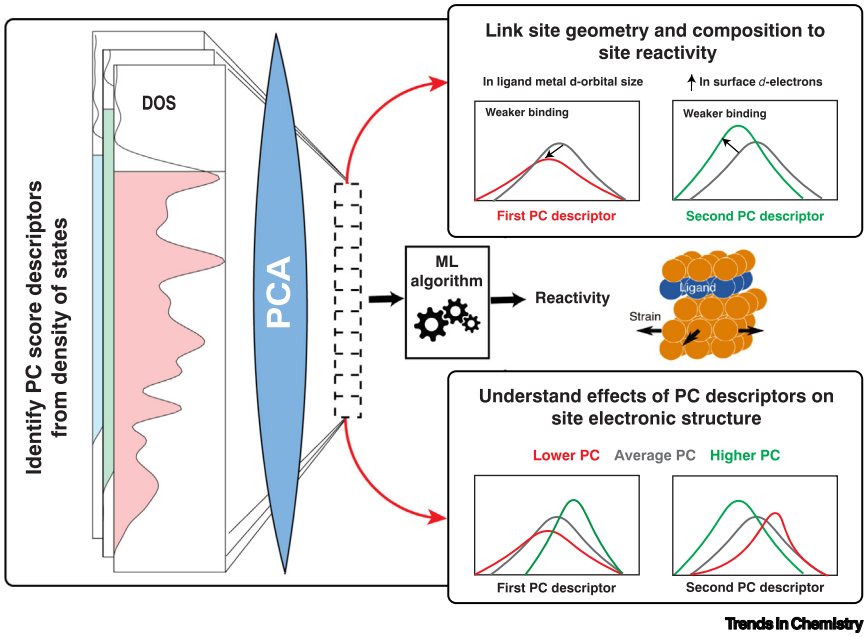

Figure 1. Automated identification of explainable electronic-structure descriptors using principal component analysis (PCA). PCA reduces the high-dimension of density of states (DOS) features to low-dimension principal component (PC) score descriptors. PC descriptors link electrical structure, geometry, and composition and catalytic properties of materials. Abbreviations: ML, machine learning.

surfaces to build a multivariate model and predict adsorption energies of different adsorbates for a wide range of chemistries. Their model is descriptively accurate because of its transparency, while being too restrictive to describe the nonlinearity of chemical bonding processes and thus has limited prediction accuracy, as shown in Figure 2. In another effort, Esterhuizen and colleagues [10] used a decision tree-based generalized additive model (iGAM), which is an extension of the linear regression model, for quantifying the relationship between the geometric structure of an adsorption site and site reactivity. It can provide insights into the contribution of each feature to the adsorption properties by nonlinear mapping (shape functions) included in the algorithm formulation. Inheriting the mathematical transparency of linear regression models, symbolic regression with compressed sensing algorithms (e.g., SISSO [11]) were used to find sparse feature representations for predicting the adsorption energies of descriptor species relevant to CO methanation and oxygen evolution on transition metals and their oxides. This type of symbolic regression can make them reasonably accurate while also providing descriptive accuracy. Going forward, the model explainability can be realized by training the ML models to produce not just the predictions but also the explanation. Using theory-based models is a straightforward approach for training explanation. The BayesChem [12] model, which is built upon the *d*-band reactivity theory, could offer quantitative explanations of the underlying mechanism of chemical bonding, which is difficult to get using purely data-driven regression models. Nonetheless, this XAI approach of developing transparent inherently typically restricts the complexity of model algorithms and comes with performance costs (i.e., as their explainability improves, the performance deteriorates). Therefore, there appears a tradeoff between model explainability and performance, as illustrated in Figure 2. In another work, where Fung and colleagues [13] used convolutional neural networks [14] to automatically extract high-level feature representations of the electronic DOS, a perturbative XAI approach was exploited to gain insights from the model after training. For that, they performed transformations on the input DOS and monitored their impact on their DOSnet ML prediction. For instance, a band-shift transformation corresponding to electron/hole doping showed that a downshift of DOS leads to weaker bindings, consistent with the *d*-band theory of chemisorption. They also benefited from the concept of feature importance to give explanations on which parts of the DOS are accountable for the chemisorption energy prediction. Applying these types of explanatory XAI approaches on black-box models can explore the reactivity changes from hypothetical perturbations like doping, alloying, and strain. Developing hybrid ML frameworks that build explainable models on top of complex ones has been an attempt to break this tradeoff, shown in Figure 2. In a hybrid model, a complicated network can be used to do what it is good at (i.e., learning representations of data) and a transparent model often rooted on theories can be applied to infer insights and explanation. Wang and colleagues [15] have recently proposed the theory-infused neural network (TinNet) approach that integrates deep learning algorithms with the *d*-band theory of chemisorption for reactivity prediction of transition-metal surfaces models. TinNet passes the output activations from the graph-based neural network in the regression module into the theory module where the explainability is encoded. In addition to highly accurate predictions, the trained parameters of the TinNet can provide utmost explanations within the theoretical framework. With TinNet-derived feature representations in terms of interaction parameters, a deconvolution of the *d*-contributed adsorption energy into Pauli repulsion and orbital hybridization becomes readily available and can shed further light

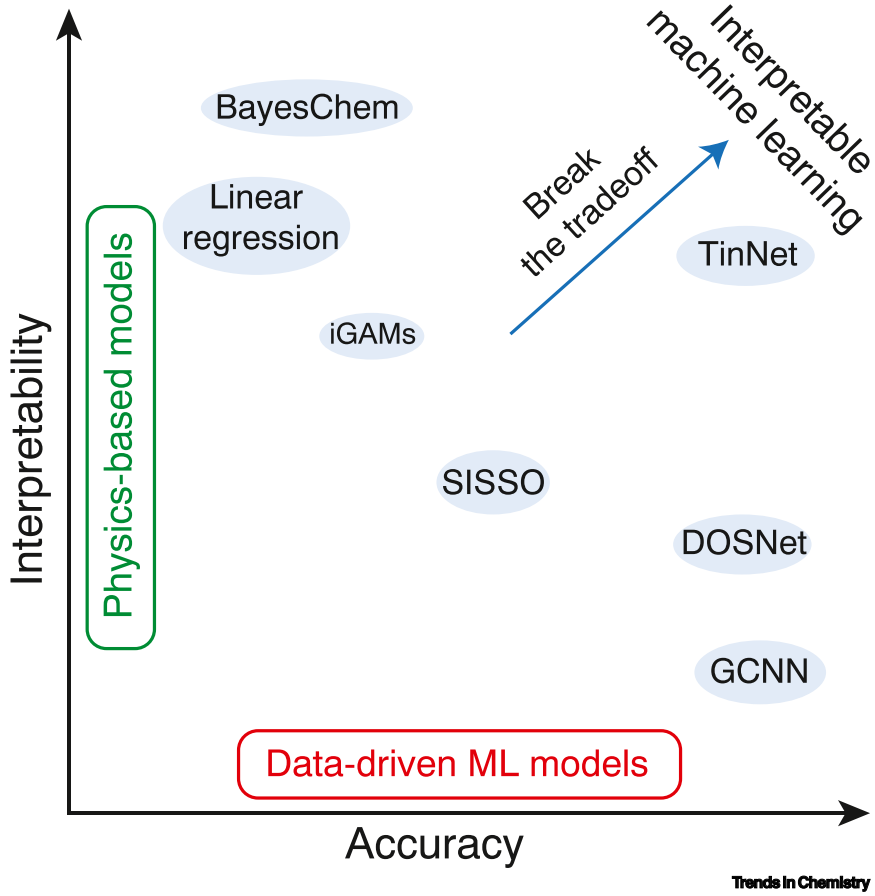

Figure 2. The tradeoff between interpretability and accuracy of interpretable machine learning models. Abbreviations: ML, machine learning.

on the nature of chemical bonding at metal surfaces.

From our view, algorithm-derived feature representations play a key role in developing interpretable ML models for revealing novel scientific knowledge. Integration of theories into ML algorithms shows promise to break the tradeoff of model predictive accuracy and interpretation integrity and presents new opportunities for ML discovery of novel materials with desirable catalytic properties.

Declaration of interests

No interests are declared.

[1]Department of Chemical Engineering, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061, USA

*Correspondence:
hxin@vt.edu (H. Xin).

https://doi.org/10.1016/j.trechm.2021.10.001

References

1. Winther, K.T. *et al.* (2019) Catalysis-Hub.org, an open electronic structure database for surface reactions. *Sci. Data* 6, 75
2. Álvarez-Moreno, M. *et al.* (2015) Managing the computational chemistry big data problem: the ioChem-BD platform. *J. Chem. Inf. Model.* 55, 95–103
3. Landis, D.D. *et al.* (2012) The computational materials repository. *Comput. Sci. Eng.* 14, 51–57
4. Chanussot, L. *et al.* (2021) Open Catalyst 2020 (OC20) dataset and community challenges. *ACS Catal.* 11, 6059–6072
5. Medford, A.J. *et al.* (2015) From the Sabatier principle to a predictive theory of transition-metal heterogeneous catalysis. *J. Catal.* 328, 36–42
6. Esterhuizen, J.A. *et al.* (2021) Uncovering electronic and geometric descriptors of chemical activity for metal alloys and oxides using unsupervised machine learning. *Chem Catal.* 1, 923–940
7. Hammer, B. and Nørskov, J.K. (1995) Electronic factors determining the reactivity of metal surfaces. *Surf. Sci.* 343, 211–220
8. Abild-Pedersen, F. *et al.* (2007) Scaling properties of adsorption energies for hydrogen-containing molecules on transition-metal surfaces. *Phys. Rev. Lett.* 99, 016105
9. Salciccioli, M. *et al.* (2010) 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* 114, 20155–20166
10. Esterhuizen, J.A. *et al.* (2020) Theory-guided machine learning finds geometric structure-property relationships for chemisorption on subsurface alloys. *Chem* 6, 3100–3117
11. Andersen, M. and Reuter, K. (2021) Adsorption enthalpies for catalysis modeling through machine-learned descriptors. *Acc. Chem. Res.* 54, 2741–2749
12. Wang, S. *et al.* (2020) Bayesian learning of chemisorption for bridging the complexity of electronic descriptors. *Nat. Commun.* 11, 6132
13. Fung, V. *et al.* (2021) Machine learned features from density of states for accurate adsorption energy prediction. *Nat. Commun.* 12, 88
14. Back, S. *et al.* (2019) Convolutional neural network of atomic surface structures to predict binding energies for high-throughput screening of catalysts. *J. Phys. Chem. Lett.* 10, 4401–4408
15. Wang, S.-H. *et al.* (2021) Infusing theory into deep learning for interpretable reactivity prediction. *Nat. Commun.* 12, 5288
16. Montemore, M.M. *et al.* (2020) General screening of surface alloys for catalysis. *Catal. Sci. Technol.* 10, 4467–4476. https://doi.org/10.1039/D0CY00682C