

Explainable Data-driven Modeling of Adsorption Energy in Heterogeneous Catalysis

Tirtha Vinchurkar,[†] Janghoon Ock,[†] and Amir Barati Farimani^{*,‡}

[†]*Department of Chemical Engineering, Carnegie Mellon University, 5000 Forbes Street, Pittsburgh, PA 15213, USA*

[‡]*Department of Mechanical Engineering, Carnegie Mellon University, 5000 Forbes Street, Pittsburgh, PA 15213, USA*

E-mail: barati@cmu.edu

Abstract

The increasing popularity of machine learning (ML) in catalysis has spurred interest in leveraging data-driven approaches to enhance catalyst design. Despite the promise of machine learning approaches, they often function as black boxes, leaving a gap between physics-based studies and data-driven methodologies. Our study aims to bridge this gap by integrating ML techniques with eXplainable AI (XAI), specifically employing two techniques: post-hoc XAI analysis and Symbolic Regression, to unravel the correlation between adsorption energy and the properties of the adsorbate-catalyst system. Leveraging a large dataset such as the Open Catalyst 2020 (OC20) Dataset, we employ a combination of shallow ML techniques and XAI methodologies. Our investigation involves utilizing multiple shallow machine learning techniques to predict adsorption energy followed by post-hoc analysis for feature importance, inter-feature correlations, and the influence of various feature values on prediction of adsorption energy. Features such as adsorbate electronegativity, coordination number of the adsorbing atom, electronegativity of the adsorbing atom, catalyst electronegativity, and the

number of atoms in the adsorbate stand out in our post-hoc analysis within our feature set for predicting adsorption energy. There is a positive correlation between catalyst and adsorbate electronegativity with the prediction of adsorption energy. Additionally, Symbolic Regression, which helps elucidate mathematical expressions describing the relationship between these important features and adsorption energy, yields results consistent with physics-based equations in previous research. It deduces a mathematical relationship indicating that coordination number of the adsorbing atom is directly proportional to the adsorption energy. Our work establishes a robust framework that integrates ML techniques with XAI, leveraging large datasets like OC20 to enhance catalyst design through model explainability.

Keywords: Catalysis, Computational Catalysis, Adsorption Energy, Explainable AI, Symbolic Regression, High-throughput Screening

Introduction

Rational design of optimal catalysts remains one of the critical challenges in heterogeneous catalysis.^[1] While the primary goal is to enhance catalytic performance, the underlying physics spans across nanoscale phenomena, interfacial chemistry, and reaction engineering.^[2] The multifaceted nature of catalyst surfaces and interfaces makes it particularly difficult to establish universal theoretical frameworks that can reliably predict catalyst reactivity through simple descriptors. Despite these challenges, significant progress has been made in identifying key parameters that govern catalytic behavior.

In catalyst design and development, a critical focus lies on understanding adsorption energy due to its strong correlation with catalytic reactivity, governing the interaction between reactants and catalysts.^[3] The Brønsted-Evans-Polanyi (BEP) relations illustrate volcano-shaped linear relationships between catalytic activity and adsorption energies, highlighting it as a pivotal descriptor.^{[4][7]} The BEP relation demonstrates a linear relationship between activation barrier, E_a , and adsorption energy, E_{ads} , in heterogeneous catalysis. This linear

relationship implies that the catalyst’s reactivity can be primarily characterized by a descriptor, E_{ads} . Modeling adsorption energy provides insights into reactivity and aids in catalyst screening, thus understanding the factors affecting adsorption energy is essential to identify the optimal catalyst for a target reaction.

Density Functional Theory (DFT) is a quantum chemistry-based method widely used to calculate the properties of atomic systems, including adsorption energy in catalysis. Despite its high accuracy, DFT requires substantial computational resources. Consequently, evaluating a large number of materials for catalyst design using expensive quantum chemistry-based approaches like DFT for high-throughput screening becomes infeasible. As a solution, machine learning approaches are gaining much attention due to their potential to accurately predict properties, such as energy and forces, at a reduced computational cost. Notably, extensive databases such as the OC20 dataset^[8] comprise over 1.2 million DFT relaxations of adsorbate-catalyst systems (approximately 250 million single-point calculations) across a substantially broader structure and chemistry space than previously realized. These datasets serve as the foundation for training and deploying machine learning models, aimed at predicting the energy and interatomic forces of the adsorbate-catalyst systems. The extensive availability of material databases like OC20 for adsorbate-catalyst systems has significantly accelerated the adoption of ML approaches in catalysis, particularly Graph Neural Networks (GNNs).^[9-15]

Despite the rise of ML approaches in atomic modeling, their black-box nature obscures the underlying physical insights related to energy and force predictions. This lack of interpretability limits the further application of ML methods, as the contributing factors to predictions remain unclear. Additionally, since ML predictions inherently come with a certain level of uncertainty,^[16] it is more beneficial to gain insights about catalyst systems from ML modeling rather than solely relying on prediction values. To address this challenge, this paper introduces the use of XAI, which uncovers the underlying correlations behind predictions and provides human-understandable explanations, thereby enhancing domain-specific

knowledge.^[17] To achieve explainability, we implement two strategies: firstly, training shallow machine learning models such as Adaboost Regression, XGBoost Regression, Support Vector Regression (SVR), Kernel Ridge Regression (KRR), and Least Absolute Shrinkage and Selection Operator (LASSO) Regression accompanied by post-hoc XAI analysis; secondly, applying Symbolic Regression to gain better understanding of relation between top 5 input features and adsorption energy. The former strategy highlights the importance and correlations of features, while the latter generates mathematical equations that directly expose the potential relationships between input features and target label.

Results and Discussion

Framework

Our study leverages XAI to extract meaningful insights about the relationship between the characteristics of adsorbate-catalyst systems and the adsorption energy, which is the label value for ML predictions. We utilize XAI in two distinct ways as seen in Figure [1](#). First, through shallow machine learning techniques combined with post-hoc XAI analysis. The post-hoc XAI analysis is conducted using the SHAP (SHapley Additive exPlanations) library,^[18] which provides feature importance based on Shapley values. This feature importance is crucial for understanding which characteristics most strongly influence adsorption energy, helping to guide catalyst design by highlighting key factors that impact system performance. Second, we employ Symbolic Regression to derive mathematical equations to compute the adsorption energy based on the top 5 input features with higher Shapely values. Obtaining mathematical equations to calculate adsorption energy is crucial as they enable a clear, quantitative understanding of complex relationships in data, allowing for efficient predictions, easier integration into simulations, and insights into the governing principles of adsorption phenomena. These equations are compared to equations obtained from traditional theory or experimental-based methods, highlighting their potential complementarity.

The feature importance and correlation among features from SHAP analysis, along with the mathematical equations obtained from the Symbolic regression technique, provide better insights into the relation of adsorption energy with structural properties in a data-driven manner. The adsorption energies and feature values like electronegativity, coordination number, and so forth, for both methods are extracted from the adsorbate-catalyst systems in the OC20 dataset. The details of the feature extraction are elaborated in the following section. Integrating ML with XAI on large datasets, such as the OC20 dataset, yields valuable modeling insights that enhance our understanding of catalyst behavior. These insights can be validated through physics-based studies, providing a robust framework for informed catalyst design and accelerating advancements in the field.

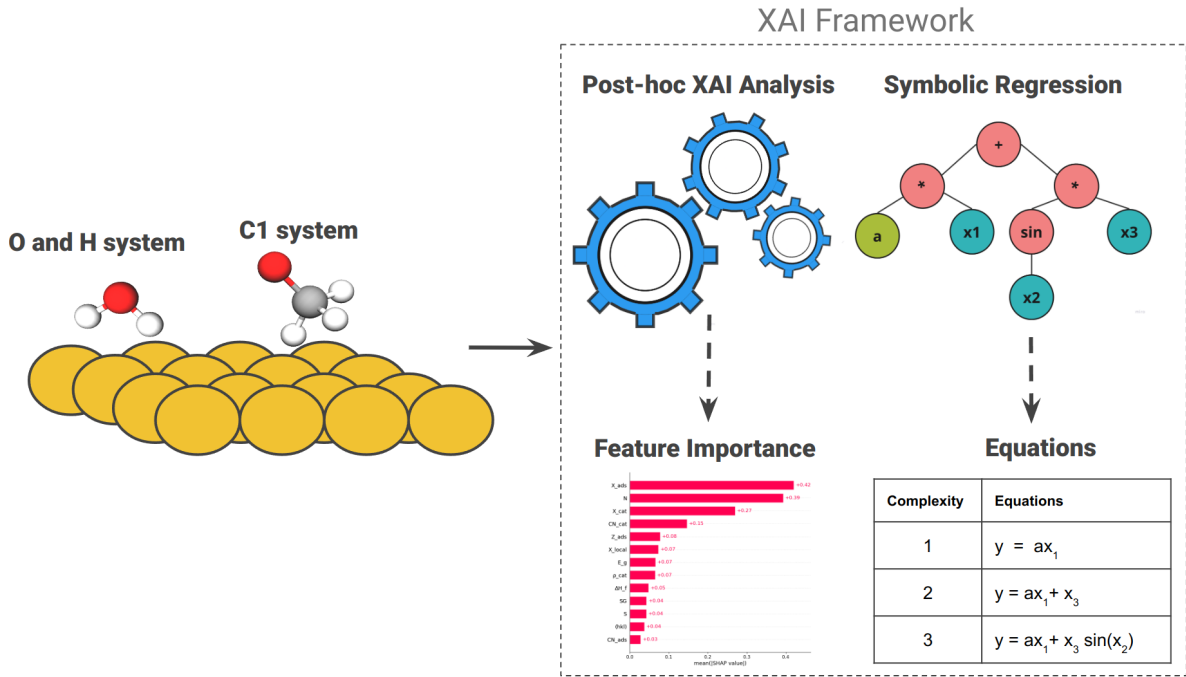


Figure 1: Overview of XAI Methods: Predicting adsorption energy with shallow machine learning models and symbolic regression. Feature importance is derived from shallow machine learning predictions through post-hoc SHAP analysis. Symbolic regression provides mathematical equations alongside its predictions.

Selection of Features

The features are carefully chosen from previous physics-based modeling research regarding adsorption energy computation. The features are categorized into three groups: adsorbate, catalyst-bulk, and surface features, as seen in Table 1. Each feature provides insights into electronic structure, bonding characteristics, and energetic properties of adsorption sites and catalysts. For instance, local electronegativity is determined as the geometric mean of the Pauling electronegativity of metal atoms within the first neighboring shell, offering insight into the local reactivity of catalyst surfaces¹⁹ as follows:

$$\chi_{local} = \prod_{j=1}^{1st\ nn} \chi_j^{01/N} \quad (1)$$

where χ_j^0 is the electronegativity of atom j and N is the total number of atoms within the first neighboring shell including the adsorption site i . While, site type represented numerical classifications of different adsorption site configurations such as bridge sites, ontop sites and hollow-HPC sites.

The features are primarily selected based on previous adsorption modeling studies, as substantiated by the references listed in Table 1, justifying their inclusion in ML-based adsorption energy modeling. Ishioka et al. demonstrated their machine learning models consisting of a Random Forest Classifier and Support Vector Classifier, utilized electronegativity, atomic number, and density as catalytic descriptors and could accurately predict ethylene/ethane selectivity (C2s) in the oxidative coupling of methane reaction.²⁰ The cross-validation score reached a high value of 0.844 for both models, successfully predicting three catalysts with high C2s values, which were subsequently confirmed by experimental validation. Additionally, Roy et al. showed that their model can predict adsorption energies for various intermediates accurately, leading to the identification of seven active catalysts, including CuCoNiZn-based tetrametallic, CuNiZn-based trimetallic, and CuCoZn-based trimetallic alloys. The model construction considered the impact of adsorption sites, encompassing tops,

Table 1: Input features with corresponding notations, references, and descriptions

Category	Features	Notation	Reference	Description
Adsorbate	Adsorbate electronegativity	χ_{ads}	Ishioka et al. ^[20]	Average electronegativity of all atoms in adsorbate molecule
	Sum of adsorbate atomic numbers	ΣZ_{ads}	Tran et al. ^[21]	Sum of atomic numbers of all atoms in adsorbate molecule
	Center atom coordination number	CN_{center}	Calle-Vallejo et al. ^[22]	Coordination number of the central atom in adsorbate molecule
	Adsorbing atom atomic number	Z_{adatom}	Roy et al. ^[23]	Atomic number of adsorbing atom present in adsorbate
	Adsorbing atom electronegativity	χ_{adatom}	Amos et al. ^[24]	Pauling’s electronegativity of adsorbing atom present in adsorbate
	Adsorbing atom coordination number	CN_{adatom}	Calle-Vallejo et al. ^[22]	Coordination number of adsorbing atom present in the adsorbate
Catalyst-bulk	Number of adsorbate atoms	N_{ads}	Amos et al. ^[24]	Total number of atoms present in adsorbate molecule
	Density	ρ_{cat}	Li et al. ^[19]	Density of the overall catalyst structure
	Band gap	E_g	Ma et al. ^[25]	Band gap of catalyst structure
	Space group	SG	Jain et al. ^[26]	Space group number of catalyst
	Formation energy	ΔH_f	Ha et al. ^[27]	Formation energy of catalyst
	Mean atomic number of bulk/catalyst	\bar{z}_{bulk}	Tran et al. ^[21]	Mean atomic number of the catalyst atoms
	Fermi Energy	E_F	Ishioka et al. ^[20]	Fermi Energy of catalyst
Surface	Catalyst electronegativity	χ_{cat}	Ishioka et al. ^[20]	Average electronegativity of the interacting surface atoms of catalyst
	Effective coordination number	CN_{cat}	Calle-Vallejo et al. ^[22]	Average coordination number of the interacting surface atoms of catalyst
	Local electronegativity	χ_{local}	Li et al. ^[19]	Geometric mean of the Pauling electronegativity of catalyst atoms within the first neighboring shell at the adsorption site
	Site type	S	Roy et al. ^[23]	Type of adsorption site based on interaction
	Miller index	hkl	Chen et al. ^[28]	Miller index of catalyst
	Adsorbate-catalyst bond length	d_{ac}	Ha et al. ^[27]	Bond length of adsorbing atom and slab atom

bridges, and hollows, along with their neighboring atoms.^[23] Calle-Vallejo et al. showcased how the utilization of coordination number, incorporating second-nearest neighbors, enabled the identification of three strategies for introducing cavity sites onto the platinum (111) surface. This enhancement proved crucial in improving its efficacy in the oxygen reduction reaction, vital for fuel cells.^[22]

Moreover, Ha et al. employed formation energy as a machine learning descriptor to facilitate the discovery of high-performance single-atom catalysts for electrochemical reactions such as the hydrogen evolution reaction and oxygen evolution/reduction reactions.^[27] The values of area under the Receiver Operating Characteristic (ROC) curve ranged from 0.79 to 0.91. Furthermore, Tran et al. illustrated how atomic number serves as a key descriptor in identifying 131 CO₂ reduction candidate surfaces across 54 alloys and 258 H₂ evolution surfaces across 102 alloys. The Root-Mean-Squared Error (RMSE), Mean Absolute Error (MAE), and median absolute deviation for predictions were 0.46, 0.29, and 0.17 eV, respectively.^[21] The incorporation of structural features, notably the number of atoms alongside elemental properties, contributed to the predictive accuracy in modeling band gaps, resulting in a model capable of predicting band gaps for 2254 light-harvesting materials with a RMSE of 0.23 eV and MAE of 0.14 eV.^[24] Overall, the careful selection of features, informed by both theoretical considerations and empirical evidence from the literature, provides a comprehensive framework for analyzing the complex interplay between catalyst structure, adsorption behavior, and catalytic activity.

Post-hoc XAI analysis

In our study, we employ SHAP analysis on shallow machine learning models to gain insights into feature importance, inter-feature correlations, and the influence of various features on predicting adsorption energy. The chosen machine learning models include Adaboost Regression with a base estimator as Random Forest Regressor, XGBoost Regression, Support Vector Regression (SVR), Kernel Ridge Regression (KRR), and Least Absolute Shrinkage

and Selection Operator (LASSO) Regression. We chose the above ML models based on insights gleaned from comprehensive reviews and studies on machine learning applications for property prediction of materials to ensure the relevance of our analytical framework.^{29,30}

The dataset, comprising 24053 adsorbate-catalyst systems, is divided into an 85:15 ratio for training and testing purposes, with input features and the adsorption energy label provided to the machine learning models. Details about the dataset and data selection are elaborated in the Methods section. The performance outcomes of the models are presented in Table 2. The outcomes demonstrated by all models exhibit comparable performance with MAE values ranging from 0.398 to 0.617. Notably, Adaboost regression stands out as the best performer, demonstrating the lowest MAE value. Therefore, it is used as the reference model for SHAP analysis in the post-hoc analysis. This highlights an important point: post-hoc analysis heavily depends on model performance. In other words, the analysis reflects how the model interprets the features, so advancing the model is essential to gain deeper insights.

Table 2: Model Performance Comparison Based on MAE (Mean Absolute Error) in eV

Model	MAE [eV]
Adaboost Regression	0.398 \pm 0.024
XGBoost Regression	0.461 \pm 0.049
Support Vector Regression	0.477 \pm 0.068
Kernel Ridge Regression	0.509 \pm 0.089
LASSO Regression	0.617 \pm 0.065

In the post-hoc analysis, the primary aspect of model explainability is carried out through SHAP analysis.³¹ Before conducting the SHAP analysis, a correlation matrix (Figure 2(b)) was generated to assess the relationships between input features. The matrix reveals several key relationships, such as a strong positive correlation between X_{ads} and X_{cat} , indicating potential redundancy. Conversely, features like band gap E_g and effective coordination number CN_{cat} exhibit weak correlations, suggesting independent contributions to the model. Additionally, properties like ΔH_f (enthalpy of formation) and ρ_{cat} (density of the catalyst) show moderate correlations with other features, implying they may play a multifaceted role

in adsorption energy predictions. This analysis helps in eliminating highly correlated features to reduce redundancy and improve model efficiency.

The SHAP library has been widely used in various domains, including materials science, to interpret black box machine learning models.¹⁸ This technique computes Shapley values for all features, thereby elucidating each feature’s contribution to the prediction of adsorption energy by assessing the impact of each feature across different combinations of feature values.

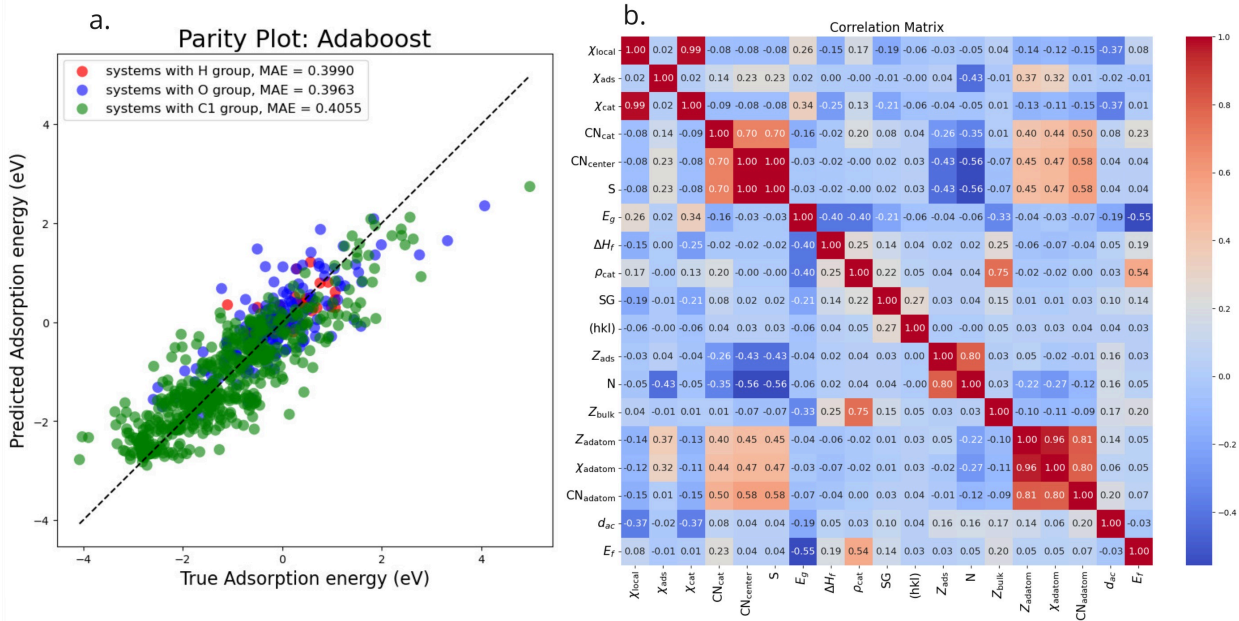


Figure 2: Performance Evaluation and Feature Correlation: (a) parity plot illustrating the performance of Adaboost Regression with Base estimator as Random Forest Regressor (Best model). MAE values are calculated for systems with H, O and C1 group. (b) Correlation matrix depicting the relationships between input features. High values of correlation coefficient have been found amongst a few features such as between Local electronegativity and Catalyst electronegativity, Site type and coordination number of adsorbate molecule.

Figure 3(a) and Figure 3(b) display the feature importance using a radar chart and a summary plot, respectively. From the feature importance, the top five features (higher Shapley values) are adsorbate electronegativity (χ_{ads}), coordination number of the adsorbing atom (CN_{adatom}), electronegativity of the adsorbing atom (χ_{adatom}), catalyst electronegativity (χ_{cat}), and the number of atoms in the adsorbate (N_{ads}), with Shapley values of 0.46, 0.43, 0.25, 0.19, and 0.19, respectively. Electronegativity stands out as the most significant

feature among the catalyst-related factors, which is consistent with findings reported in the literature (as referenced in Table 1). Three of the top five features—electronegativity of the adsorbate, adsorbing atom, and catalyst surface atoms—underscore the pivotal role of electronegativity in determining adsorption behavior on catalytic surfaces.

Conversely, a few features exhibit lower importance. For instance, Formation energy ΔH_f , which plays a crucial role in single-atom catalysts as seen in Ha et al.,²⁷ demonstrates limited relevance in our case, which includes heterogeneous catalysts. This may be due to its nature as a bulk property, making it less effective in predicting the interactions between adsorbates and catalytic surfaces in heterogeneous systems, as reflected in its low Shapely value (+0.06). Similarly, the band gap E_g , a bulk electronic property, holds minimal predictive power for surface interactions, with a Shapely value of +0.06.

Moreover, there appears to be a stronger influence of adsorbate properties compared to catalyst properties in determining adsorption behavior. This may suggest that adsorption energy is more dependent on the type of adsorbate, possibly due to the limited variety of adsorbates, with only three categories (H, O & C1) represented in the dataset, compared to the broader range of catalysts. This imbalance in the number of adsorbates versus catalysts could influence the observed feature importance, highlighting the significant role of data distribution in shaping the outcomes of the feature importance analysis.

Previous studies have shown varying relationships between catalyst electronegativity and adsorption energy. Gao, Wang, et al. proposed an inverse relationship specifically for transition metal catalysts, where higher electronegativity leads to lower adsorption energy.³² Conversely, Trasatti’s research suggests a direct proportionality between metal catalyst electronegativity and adsorption energy.³³

Our analysis reveals a positive correlation between catalyst electronegativity (χ_{cat}) and adsorption energy. This result is evident in the beeswarm plot analysis shown in Figure 3(c), where significant clustering of data points and feature values for catalyst electronegativity can be observed for positive Shapely values. Unlike the findings by Gao and Wang, which

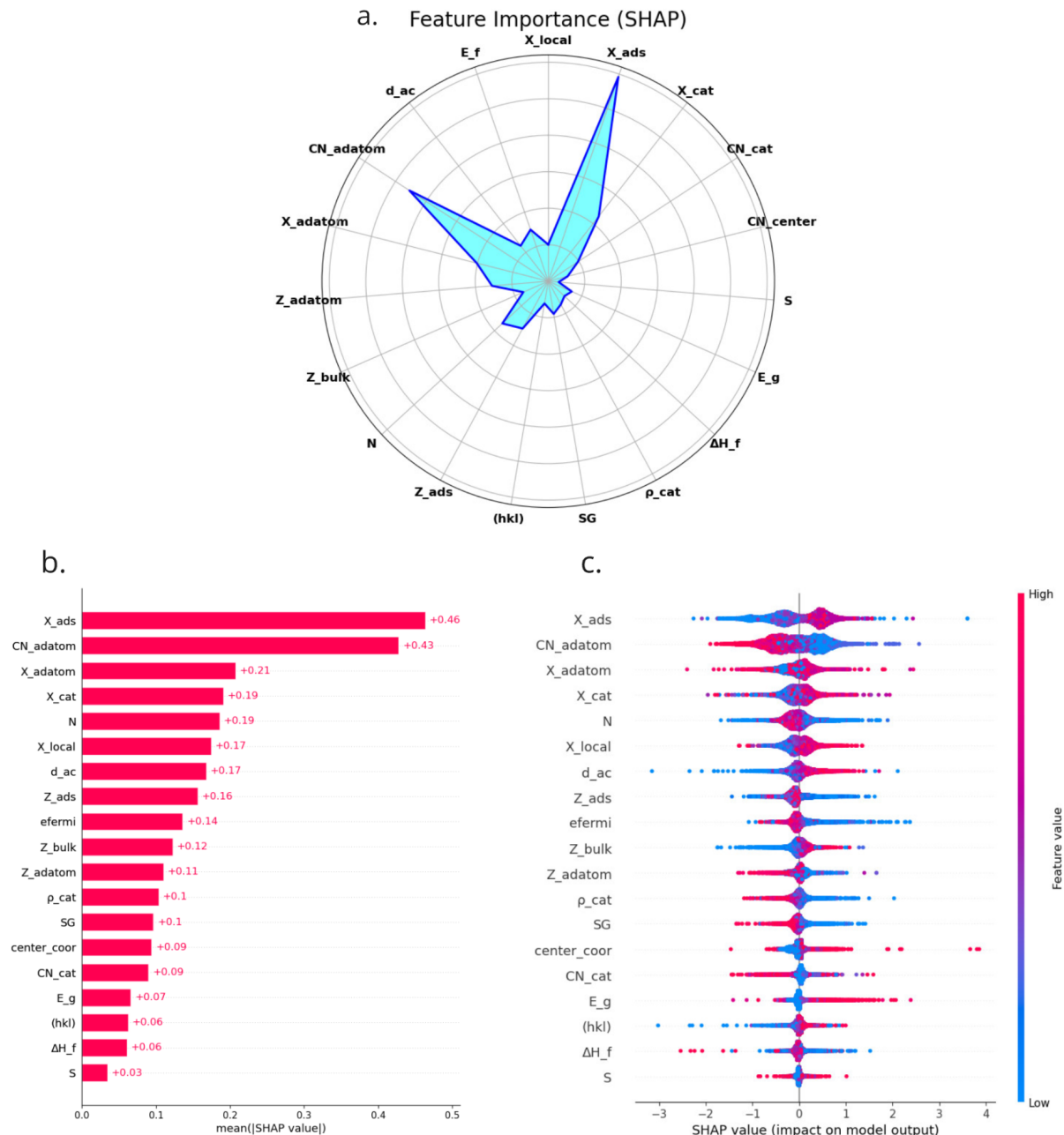


Figure 3: SHAP Analysis: (a) Radar plot illustrating feature importance based on Shapley values. (b) Summary Bar Plot presenting all features. (c) Beeswarm plot depicting the relationship between feature values and Shapley values.

show an inverse relationship between catalyst electronegativity and adsorption energy for transition metal catalysts, our dataset includes a broader range of adsorbates—hydrogen, oxygen, and carbon-based—paired with metal as well as non-metal catalysts. This diversity may account for the differences observed. This underscores the intricate nature of catalyst-

adsorbate interactions, where various factors such as the composition of the dataset and the choice of adsorbates and catalysts play crucial roles in determining the observed outcomes. Additionally, a similar trend is observed for adsorbate electronegativity (χ_{ads}) and local electronegativity (χ_{local}), supporting our findings and emphasizing the complexity of catalyst-adsorbate interactions.

We note a denser clustering of data points with higher feature values corresponding to negative Shapley values for the number of adsorbate atoms in Figure 3(c). This trend suggests a diminished prediction of adsorption energy with an increased number of adsorbate atoms. This relationship may be attributed to the limited number of adsorbates used in the dataset, with a greater number of adsorbate atoms typically observed in the C1 category of adsorbates, which predominantly contributes to this relation.

In adsorption studies, surface characteristics are typically the primary focus, rather than bulk properties. However, for this dataset, the electronegativity of the catalyst bulk may still provide valuable insights into surface properties. This inference is supported by Figure 4, which explores feature-feature correlations and reveals two notable findings. Firstly, Figure 4(a) demonstrates a positive correlation between mean catalyst electronegativity (χ_{cat}) and local electronegativity (χ_{local}). This relationship reflects the influence of catalyst atoms within the first neighboring shell at the adsorption site. It logically follows that the electronegativity of a subset of surface atoms contributes to the overall mean electronegativity of the catalyst lattice. Additionally, a similar correlation is observed in the correlation matrix depicted in Figure 2(b), where local electronegativity and catalyst electronegativity exhibit a high correlation coefficient. This suggests a notable association between surface and bulk electronegativity in this dataset. If a strong correlation between bulk and surface characteristics is established for the target dataset, bulk characteristics could serve as proxies for surface properties, potentially offering greater flexibility in data acquisition, particularly for experimental studies.

The increase in the number of adsorbate atoms correlates with a decrease in overall

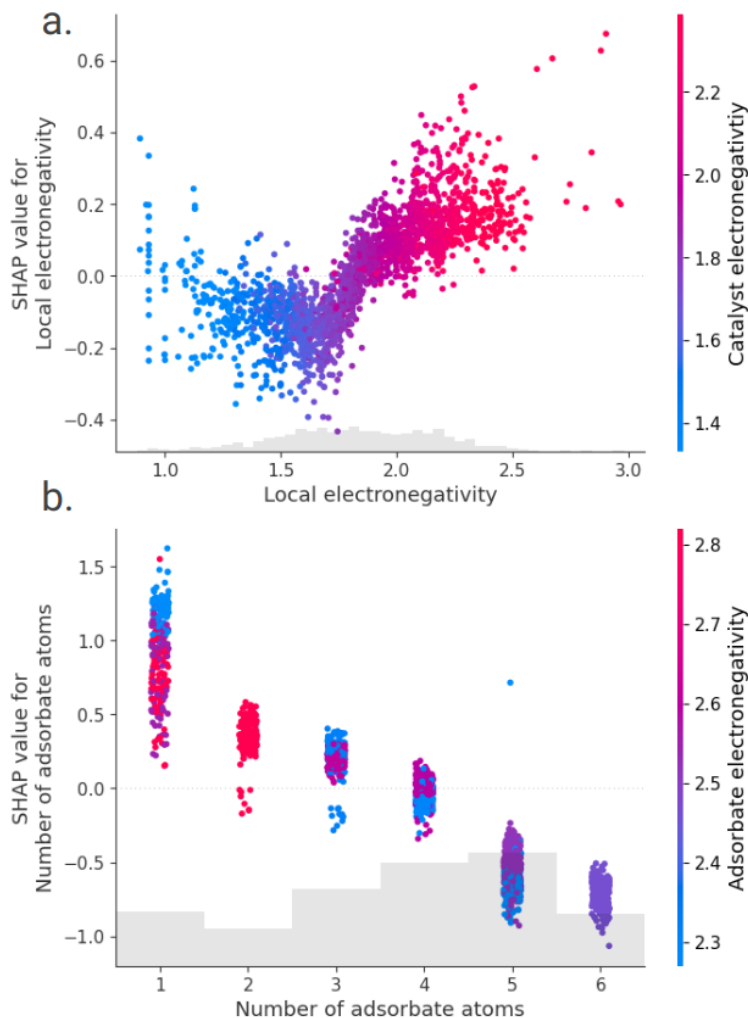


Figure 4: Feature-Feature Correlation: (a) Scatter plot illustrating the relationship between catalyst electronegativity and local electronegativity, indicating higher values of catalyst electronegativity with increasing local electronegativity. (b) Scatter plot demonstrating the correlation between the number of adsorbate atoms and adsorbate electronegativity, revealing a tendency for adsorbate electronegativity to decrease with an increase in the number of adsorbate atoms.

adsorbate electronegativity, as depicted in Figure 4(b). This trend is evident from the transition in color scheme from orange (indicating high adsorbate electronegativity) to purple (indicating lower adsorbate electronegativity). The limited variety of adsorbate types in our dataset likely amplifies the role of adsorbates as key determinants in the adsorption process. The decrease in electronegativity could be attributed to the formation of chemical bonds between the adsorbate atoms. As more adsorbate atoms come into proximity, bonding

interactions occur, causing a redistribution of electron density within the adsorbate system and thereby lowering the overall electronegativity.^[34] Adsorbate bonding and composition are key to catalyst interactions, directly affecting adsorption efficiency. This insight can guide the construction of more robust and targeted datasets for adsorption studies.

Correlation from Symbolic Regression

Symbolic regression is a machine learning-based regression method aimed at identifying an underlying mathematical expression that best describes the relationship between input features and output. In our study, we employed the PySR and SISSO++ libraries, which are open-source tools designed to find an interpretable symbolic expression that optimizes a specified objective.^{[35][36]} In our study, we selected the top five important features with the highest Shapely values from the dataset and aimed to derive meaningful mathematical expressions related to adsorption energy. After removing highly correlated features and conducting post-hoc XAI analysis, the chosen five features were: the number of adsorbate atoms (N), adsorbate electronegativity (χ_{ads}), coordination number of the adsorbing atom (CN_{adatom}), local electronegativity (χ_{local}), and catalyst electronegativity (χ_{cat}). These features serve as the input for deriving interpretable, physically meaningful equations using PySR and SISSO++.

PySR generates candidate equations using genetic programming and evaluates them with a user-defined loss function (e.g., RMSE, MAE).^[37] SISSO++ employs a compressed-sensing-based approach to generate candidate equations by minimizing a user-defined loss function, allowing for interpretable and accurate models.^[36] Both models, PySR and SISSO++, were configured with similar settings to generate equations at different complexity levels. SISSO++ outperformed PySR on both metrics. For more details on the algorithms, refer to the Methods section.

In this framework, the inputs consist of the dataset containing selected important input features and corresponding output label i.e. adsorption energy, while the output of the PySR

and SISSO++ models is a set of mathematical equations that represent the relationships between the input features and output label. These equations serve as interpretable models that capture the underlying patterns and dependencies within the data.

Table 3: Comparison of PySR equations and SISSO++ equations

PySR equations
$E_{\text{ads}} = (-0.26338 \cdot N)$
$E_{\text{ads}} = \frac{\chi_{\text{local}} - N}{\chi_{\text{ads}}}$
$E_{\text{ads}} = (\chi_{\text{local}} + (-0.49684 \cdot N)) - 0.78192$
$E_{\text{ads}} = \frac{\chi_{\text{ads}} - (N + \frac{CN_{\text{adatom}}}{\chi_{\text{local}}})}{\chi_{\text{ads}}}$
$E_{\text{ads}} = ((\chi_{\text{ads}} - (N + (CN_{\text{adatom}}/\chi_{\text{local}}))))/(\chi_{\text{ads}} - 0.18890163))$
SISSO++ equations
$E_{\text{ads}} = c_0 + a_0 \cdot \left(\frac{\chi_{\text{local}} - N}{\chi_{\text{ads}}^6} \right)$
$E_{\text{ads}} = c_0 + a_0 \cdot (CN_{\text{adatom}} - \chi_{\text{ads}} + CN_{\text{cat}} - N) + a_1 \cdot \left(\frac{\chi_{\text{local}} - N}{\chi_{\text{ads}}^3} \right)$
$E_{\text{ads}} = c_0 + a_0 \cdot \cos(\exp(\chi_{\text{ads}})) + a_1 \cdot (CN_{\text{cat}} - N + CN_{\text{adatom}}) + a_2 \cdot \left(\frac{\chi_{\text{local}} - N}{\chi_{\text{ads}}^3} \right)$
$E_{\text{ads}} = c_0 + a_0 \cdot \cos(\chi_{\text{ads}}^2) + a_1 \cdot (\chi_{\text{ads}} + \chi_{\text{local}}) - (\chi_{\text{ads}} \cdot \chi_{\text{local}}) + a_2 \cdot (CN_{\text{adatom}} - \chi_{\text{ads}} + CN_{\text{cat}} - N) + a_3 \cdot \left(\frac{\chi_{\text{local}} - N}{\chi_{\text{ads}}^6} \right)$
$E_{\text{ads}} = c_0 + a_0 \cdot \chi_{\text{ads}} - \chi_{\text{local}} - \exp(-CN_{\text{adatom}}) + a_1 \cdot \sin(\chi_{\text{local}})^6 + a_2 \cdot \cos(\chi_{\text{ads}}^2) + a_3 \cdot (CN_{\text{cat}} - N + CN_{\text{adatom}}) + a_4 \cdot \left(\frac{N}{\chi_{\text{ads}} \cdot \chi_{\text{ads}}^3} \right)$

The equations generated by PySR and SISSO++, as shown in Table 3, consist of five equations for each method, presented in increasing levels of mathematical complexity. PySR equations are simpler, utilizing basic arithmetic operations, while SISSO++ produces more complex equations with higher-order terms that capture subtle, non-linear relationships in the data. PySR begins with a single input feature at complexity 1 and gradually includes more features as complexity increases, integrating all key features by complexity 5 (Table 3). In contrast, SISSO++ uses a broader range of parameters from the outset, incorporating three features at complexity 1 and utilizing all features by complexity 3. This leads to faster convergence rates and enhances its applicability in computational modeling. Another notable advantage of SISSO++ is its ability to consider feature units, ensuring physical relevance and interpretability of the resulting equations. For this study, feature units were specified during the SISSO++ setup, providing context that PySR lacks.

Although SISSO++ and PySR exhibit relatively high root mean square error (RMSE) values of 0.8 eV and 1.2 eV, respectively (details in the SI), SISSO++ successfully captures

the correlation between input features and adsorption energy, aligning with the findings of Gao et al.^[32] Gao’s work shows a direct proportionality between adsorption energy and the effective coordination number, as seen in Equation [2](#). The first term contains ψ which is electronic descriptor, CN_{cat} in second term is effective coordination number and α , λ and θ are system-specific parameters. Similarly, SISSO++ identifies this proportionality from the second dimension onward, where E_{ads} is directly proportional to $|CN_{cat}|$, the effective coordination number (as seen in table [3](#)). These results suggest that at appropriate complexity levels, SISSO++ effectively captures critical feature interactions and underlying relationships, improving predictive accuracy while maintaining alignment with established literature.

$$E_{ads} = 0.1 \cdot \alpha \cdot \psi + \lambda \cdot CN_{cat} + \theta \tag{2}$$

In addition to capturing important feature expressions, the SISSO++ library offers a deterministic approach to model discovery, ensuring reproducible results—an advantage over the genetic programming methods like PySR which are stochastic in nature. SISSO++ also supports multi-task learning, allowing it to solve multiple related tasks simultaneously and leverage shared information to improve performance across all tasks.

Conclusion

Our study successfully integrates explainability into machine learning models through post-hoc SHAP analysis and symbolic regression, highlighting the importance and correlations of key features. By merging machine learning techniques with XAI, we bridge the gap between data-driven approaches and domain-specific knowledge, offering insights into the intricate relationships between input structural features and adsorption energy. One key finding is the direct proportionality between effective coordination number and adsorption energy in catalytic systems. This is corroborated by multiple analytical approaches: SHAP analysis

identified effective coordination number as one of the top five influential features, while symbolic regression method, particularly SISSO++, expressed this direct relationship in its generated equations. Notably, these findings align with and further support the research of Gao et al., underscoring the robustness of this relationship across methodologies.

Additional insights from feature-feature correlations via SHAP analysis revealed a strong correlation between catalyst and local electronegativity, and how adsorbate electronegativity decreases with increase in number of adsorbate atoms for the dataset. The XAI techniques provide crucial insights into ML model predictions by highlighting feature importance and the impact of feature values on adsorption energy. This facilitates focused attention on relevant features, enabling symbolic regression to derive better relationships that can be leveraged for future catalyst optimization. In summary, our study establishes a robust framework that integrates machine learning techniques with XAI, using extensive datasets such as OC20 to improve catalyst design through enhanced model explainability.

Methods

Dataset and Feature selection

The dataset utilized in this study is sourced from the OC20 dataset, which comprises 872,000 trajectories obtained from DFT calculation results, making it suitable for training ML models. From these trajectories, we extracted final frames to obtain relaxed structures of adsorbate-catalyst systems. The OC20 dataset encompasses catalyst systems featuring a diverse array of reactive elements, spanning nonmetals, alkali metals, alkaline earth metals, metalloids, transition metals, and post-transition metals. In contrast, the adsorbate systems consist of 82 molecules categorized into oxygen or hydrogen, C1 molecules, C2 molecules, and nitrogen-containing molecules.

For this study, we curated a dataset for post-hoc analysis by extracting a subset of the OC20 dataset, specifically focusing on systems containing oxygen (O), hydrogen (H), and

C1 category adsorbates. This focus on small molecules minimizes the impact of interatomic interactions within the adsorbate molecules, simplifying comparisons with previous theory-based studies. Following this selection process, the dataset was narrowed down to a subset consisting of 24503 systems. To extract features from this structure data, ASE and Pymatgen packages are utilized, enabling the extraction of relevant information from the relaxed structures of each trajectory. The adsorption energy is retrieved from the OC20 dataset. Ultimately, the structured data, comprising 19 input features and adsorption energy as the output label, is fed into the ML models.

SHAP approach

We use SHAP library to perform post-hoc XAI analysis. SHAP is a technique based on game theory that helps explain how machine learning models make decisions. It connects fair credit distribution with simple explanations by using classic Shapley values. Shapely values are calculated using the Shapley regression values, a concept from cooperative game theory. Shapley regression values are feature importances for linear models in the presence of multicollinearity. This method requires retraining the model on all feature subsets $S \subseteq F$. Mathematically, the Shapley value for a feature i is calculated as follows:

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(F - |S| - 1)!}{|F|!} [f_{S \cup \{i\}}(x_{S \cup \{i\}}) - f_S(x_S)] \quad (3)$$

where F is the total number of features, S is a subset of features excluding feature i , x_S represents the input data with features in subset S , and $x_{S \cup \{i\}}$ represents the input data with features in subset S along with feature i . $f_{S \cup \{i\}}$ is the model prediction using features in subset S along with feature i where as f_S is the model prediction with feature i withheld. The term $|S|!(M - |S| - 1)!/M!$ represents the probability of including feature i in a subset S of size $|S|$, which is adjusted by the number of permutations of subsets containing i .

In essence, the Shapley value for a feature represents the average change in the model

prediction when including that feature compared to all possible combinations of features, weighted by their probabilities. Positive Shapley values signify a positive correlation between features, indicating that as one feature increases, the other feature tends to increase as well. Conversely, negative Shapley values denote a negative correlation, implying that as one feature increases, the other tends to decrease. Our study visualizes Shapley values of all features in various charts such as summary plot, beeswarm plot and scatter plots for a comprehensive understanding of feature importances and correlations.

PySR

PySR (Symbolic Regression using Python) is an open-source library to perform symbolic regression on the dataset.^[37] The core of PySR model’s algorithm involves iteratively evolving populations of mathematical expressions through tournament selection,^[38] mutation, crossover, simplification, and optimization.^{[39][40]} These populations evolve independently, with occasional migration of individuals between them. The evolve-simplify-optimize loop refines the discovered equations by simplifying them to equivalent forms and optimizing constants. This iterative process continues until satisfactory equations representing the relationships between input features and output labels are discovered.

The mathematical equations are produced through a process of evolutionary optimization, where populations of mathematical expressions are iteratively evolved and refined to best fit the dataset. This involves the generation of candidate expressions, their evaluation based on predefined fitness criteria (e.g., goodness of fit to the data), and the application of genetic operators (mutation, crossover) to create new expressions. Additionally, simplification techniques are employed to reduce the complexity of the expressions while preserving their accuracy, and optimization methods are applied to fine-tune the constants within the equations. Through this iterative process, the framework systematically explores the space of mathematical expressions to discover equations that effectively capture the relationships between the input features and output label.

To further refine these equations and mitigate overfitting, we imposed constraints during symbolic regression. We adopted RMSE as the loss function to enable direct comparison with the SISSO++ library at different complexity levels. Given the large dataset, we optimized the model with 50 iterations and enabled Turbo mode, which uses the Loop Vectorization method to accelerate the evaluation among a pool of equations. The `max_size` parameter was set to 7, limiting equation complexity to 7, while the `max_depth` was set to 5, restricting the maximum depth of the mathematical expressions or trees. We set lower values for `max_depth` and `max_size` to prioritize simplicity and interpretability, ensuring that the generated equations remain easy to understand and less prone to overfitting. Additionally, we enabled the `denoise` option, allowing the model to filter out noise in the input data and focus on the underlying relation rather than fitting to noise.

SISSO++

SISSO++ is an advanced feature selection and model construction algorithm that combines two key methods: Sure Independence Screening (SIS) and Subspace Optimization (SO). The SIS method, crucial for handling high-dimensional problems, begins with a large pool of potential features or descriptors. It calculates the correlation between each feature and the target variable (in this case, adsorption energy), then ranks and screens features based on these correlations. A subset of the most promising candidates is selected, and an independence check is performed to avoid redundancy and multicollinearity. This approach effectively reduces the dimensionality of the problem by retaining only the most relevant and independent features. Following the SIS step, the SO method comes into play. It performs a combinatorial optimization within the subspace of selected features to identify the best descriptor (or combination of descriptors) for the target property. SISSO++ applies this SIS-SO process iteratively, refining the selected features through multiple rungs.

The SISSO++ model settings used in this work are optimized to match the PySR model configuration for direct comparison. A descriptor dimensionality of 5 is chosen to capture

sufficient information while controlling model complexity. A descriptor dimensionality of 5 (`desc_dim = 5`) is chosen to capture sufficient information while controlling model complexity. In each iteration, 3 features are selected (`n_sis_select = 3`) to ensure the most relevant variables are prioritized. The model progresses through 2 rungs (`max_rung = 2`), refining the selected features iteratively. To account for residual errors, 1 additional descriptor is added (`n_residual = 1`). The dataset used is specified by `data.csv`, and a 20% fraction of the data is reserved for testing purposes (`leave_out_frac = 0.2`) to ensure robust model validation similar to PySR model. Only the top-performing model is stored (`n_models_store = 1`), and the intercept is not fixed (`fix_intercept = false`) to allow for more flexible linear fitting. These settings are selected to achieve a balance between model performance and computational efficiency.

Data and Software Availability statement

Both the Python code for data retrieval and the data employed in this study are available on GitHub at the following link: https://github.com/tirtha-v/eXplainable_AI

References

- (1) Waclawek, S.; Padil, V. V.; Černík, M. Major advances and challenges in heterogeneous catalysis for environmental applications: a review. *Ecological Chemistry and Engineering S* **2018**, *25*, 9–34.
- (2) Liu, W. Multi-scale catalyst design. *Chemical engineering science* **2007**, *62*, 3502–3512.
- (3) Nørskov, J. K.; Bligaard, T.; Logadottir, A.; Bahn, S.; Hansen, L. B.; Bollinger, M.; Bengaard, H.; Hammer, B.; Sljivancanin, Z.; Mavrikakis, M.; others Universality in heterogeneous catalysis. *Journal of catalysis* **2002**, *209*, 275–278.

- (4) Bligaard, T.; Nørskov, J. K. *Chemical bonding at surfaces and interfaces*; Elsevier, 2008; pp 255–321.
- (5) Cheng, J.; Hu, P.; Ellis, P.; French, S.; Kelly, G.; Lok, C. M. Brønsted- Evans- Polanyi relation of multistep reactions and volcano curve in heterogeneous catalysis. *The Journal of Physical Chemistry C* **2008**, *112*, 1308–1311.
- (6) Bligaard, T.; Nørskov, J. K.; Dahl, S.; Matthiesen, J.; Christensen, C. H.; Sehested, J. The Brønsted–Evans–Polanyi relation and the volcano curve in heterogeneous catalysis. *Journal of catalysis* **2004**, *224*, 206–217.
- (7) Gerashev, A. Relationship between the Brønsted Coefficients of the Steps of a Heterogeneous Catalytic Reaction. *Kinetics and Catalysis* **2022**, *63*, 645–652.
- (8) Chanussot, L.; Das, A.; Goyal, S.; Lavril, T.; Shuaibi, M.; Riviere, M.; Tran, K.; Heras-Domingo, J.; Ho, C.; Hu, W.; others Open catalyst 2020 (OC20) dataset and community challenges. *Acs Catalysis* **2021**, *11*, 6059–6072.
- (9) Ghanekar, P. G.; Deshpande, S.; Greeley, J. Adsorbate chemical environment-based machine learning framework for heterogeneous catalysis. *Nature Communications* **2022**, *13*, 5788.
- (10) Liao, Y.-L.; Smidt, T. Equiformer: Equivariant graph attention transformer for 3d atomistic graphs. *arXiv preprint arXiv:2206.11990* **2022**,
- (11) Gasteiger, J.; Becker, F.; Günnemann, S. Gemnet: Universal directional graph neural networks for molecules. *Advances in Neural Information Processing Systems* **2021**, *34*, 6790–6802.
- (12) Xie, T.; Grossman, J. C. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Physical review letters* **2018**, *120*, 145301.

- (13) Schütt, K.; Kindermans, P.-J.; Saucedo Felix, H. E.; Chmiela, S.; Tkatchenko, A.; Müller, K.-R. Schnet: A continuous-filter convolutional neural network for modeling quantum interactions. *Advances in neural information processing systems* **2017**, 30.
- (14) Gasteiger, J.; Groß, J.; Günnemann, S. Directional message passing for molecular graphs. *arXiv preprint arXiv:2003.03123* **2020**,
- (15) Ock, J.; Magar, R.; Antony, A.; Farimani, A. B. Multimodal Language and Graph Learning of Adsorption Configuration in Catalysis. 2024; <https://arxiv.org/abs/2402.17177>
- (16) Ock, J.; Tian, T.; Kitchin, J.; Ulissi, Z. Beyond independent error assumptions in large GNN atomistic models. *The Journal of Chemical Physics* **2023**, 158, 214702.
- (17) Esterhuizen, J. A.; Goldsmith, B. R.; Linic, S. Interpretable machine learning for knowledge generation in heterogeneous catalysis. *Nature catalysis* **2022**, 5, 175–184.
- (18) Lundberg, S. M.; Lee, S.-I. A unified approach to interpreting model predictions. *Advances in neural information processing systems* **2017**, 30.
- (19) Li, Z.; Ma, X.; Xin, H. Feature engineering of machine-learning chemisorption models for catalyst design. *Catalysis today* **2017**, 280, 232–238.
- (20) Ishioka, S.; Fujiwara, A.; Nakanowatari, S.; Takahashi, L.; Taniike, T.; Takahashi, K. Designing catalyst descriptors for machine learning in oxidative coupling of methane. *ACS Catalysis* **2022**, 12, 11541–11546.
- (21) Tran, K.; Ulissi, Z. W. Active learning across intermetallics to guide discovery of electrocatalysts for CO₂ reduction and H₂ evolution. *Nature Catalysis* **2018**, 1, 696–703.
- (22) 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.
- (23) Roy, D.; Mandal, S. C.; Pathak, B. Machine learning-driven high-throughput screening of alloy-based catalysts for selective CO₂ hydrogenation to methanol. *ACS Applied Materials & Interfaces* **2021**, *13*, 56151–56163.
- (24) Amos, R. D.; Kobayashi, R. Feature Engineering for Materials Chemistry—Does Size Matter? *Journal of Chemical Information and Modeling* **2019**, *59*, 1873–1881.
- (25) Ma, S.; Liu, Z.-P. Machine learning for atomic simulation and activity prediction in heterogeneous catalysis: current status and future. *ACS Catalysis* **2020**, *10*, 13213–13226.
- (26) Jain, A.; Bligaard, T. Atomic-position independent descriptor for machine learning of material properties. *Physical Review B* **2018**, *98*, 214112.
- (27) Ha, M.; Kim, D. Y.; Umer, M.; Gladkikh, V.; Myung, C. W.; Kim, K. S. Tuning metal single atoms embedded in N x C y moieties toward high-performance electrocatalysis. *Energy & Environmental Science* **2021**, *14*, 3455–3468.
- (28) Chen, D.; Shang, C.; Liu, Z.-P. Machine-learning atomic simulation for heterogeneous catalysis. *npj Computational Materials* **2023**, *9*, 2.
- (29) Butler, K. T.; Davies, D. W.; Cartwright, H.; Isayev, O.; Walsh, A. Machine learning for molecular and materials science. *Nature* **2018**, *559*, 547–555.
- (30) Chen, C.; Zuo, Y.; Ye, W.; Li, X.; Deng, Z.; Ong, S. P. A critical review of machine learning of energy materials. *Advanced Energy Materials* **2020**, *10*, 1903242.
- (31) Lundberg, S. M.; Lee, S.-I. In *Advances in Neural Information Processing Systems 30*; Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R., Eds.; Curran Associates, Inc., 2017; pp 4765–4774.

- (32) Gao, W.; Chen, Y.; Li, B.; Liu, S.-P.; Liu, X.; Jiang, Q. Determining the adsorption energies of small molecules with the intrinsic properties of adsorbates and substrates. *Nature communications* **2020**, *11*, 1196.
- (33) Trasatti, S. Electronegativity, work function, and heat of adsorption of hydrogen on metals. *Journal of the Chemical Society, Faraday Transactions 1: Physical Chemistry in Condensed Phases* **1972**, *68*, 229–236.
- (34) Cioslowski, J.; Stefanov, B. B. Electron flow and electronegativity equalization in the process of bond formation. *The Journal of chemical physics* **1993**, *99*, 5151–5162.
- (35) Cranmer, M. Interpretable machine learning for science with PySR and SymbolicRegression.jl. *arXiv preprint arXiv:2305.01582* **2023**,
- (36) Scheffler, M.; Carbogno, C.; Ghiringhelli, L. M.; others SISSO++: A C++ implementation of the sure-independence screening and sparsifying operator approach. *Journal of Open Source Software* **2022**, *7*, 3960.
- (37) Cranmer, M. Interpretable Machine Learning for Science with PySR and SymbolicRegression.jl,” 2023. *arXiv preprint arXiv:2305.01582*
- (38) Goldberg, D. E.; Deb, K. *Foundations of genetic algorithms*; Elsevier, 1991; Vol. 1; pp 69–93.
- (39) Broyden, C. G. The convergence of a class of double-rank minimization algorithms 1. general considerations. *IMA Journal of Applied Mathematics* **1970**, *6*, 76–90.
- (40) Mogensen, P.; Riseth, A. Optim: A mathematical optimization package for Julia. *Journal of Open Source Software* **2018**, *3*.

TOC Graphic

