

# Exploring Structure-Sensitive Relations for Small Species Adsorption Using Machine Learning

Xue Zong and Dionisios G. Vlachos\*



Cite This: *J. Chem. Inf. Model.* 2022, 62, 4361–4368



Read Online

ACCESS |



Metrics & More

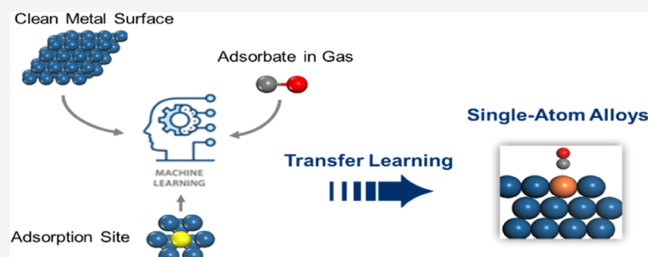


Article Recommendations



Supporting Information

**ABSTRACT:** Accurate prediction of adsorption energies on heterogeneous catalyst surfaces is crucial to predicting reactivity and screening materials. Adsorption linear scaling relations have been developed extensively but often lack accuracy and apply to one adsorbate and a single binding site type at a time. These facts undermine their ability to predict structure sensitivity and optimal catalyst structure. Using machine learning on nearly 300 density functional theory calculations, we demonstrate that generalized coordination number scaling relations hold well for oxygen- and high-valency carbon-binding species but fail for others. We reveal that the valency and the electronic coupling of a species with the surface, along with the site type and its coordination environment, are critical for small species adsorption. The model simultaneously predicts the adsorption energy and preferred site and significantly outperforms linear scalings in accuracy. It can expose the structure sensitivity of chemical reactions and enable enhanced catalyst activity via engineering particle shape and facet defects. The generality of our methodology is validated by training the model with transition metal data and transferring it to predict adsorption energies on single-atom alloys.



## INTRODUCTION

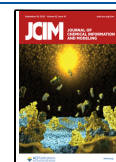
Adsorption is fundamental to the performance and reaction pathways of surface-catalyzed reactions. Despite the rapid development of surface science techniques and theoretical methods, such as density functional theory (DFT), screening the adsorption energies on many materials remains a formidable task. Streamlining this search can be achieved using descriptor-based energy scaling relations.<sup>1–3</sup> Specifically, the pioneering work of Nørskov and co-workers describes the adsorption energies of partially hydrogenated species  $AH_x$  using the atomic adsorption energy of A, e.g., the  $C^*$  adsorption energy as a descriptor for  $CH_x^*$ .<sup>4</sup> Electronic descriptors,<sup>5</sup> such as the d-band center<sup>6–8</sup> and the upper edge of the d-band,<sup>9</sup> have successfully elucidated the trends in adsorption across transition metals. Such linear scalings have transformed our ability to develop microkinetic models and predict better materials for given chemistry. However, these descriptors lack a direct link to the structural features of the active site. This fact severely limits their applicability in identifying the optimal active site and predicting the structure sensitivity of chemical reactions. As a result, engineering the catalyst structure and the active site has remained elusive. Furthermore, these scalings are accurate only when the adsorbates are placed on the same site type; yet, the most stable site changes among surfaces, making the accuracy of the scalings poor. Calle-Vallejo et al. proposed the generalized coordination number (GCN) descriptor model to predict adsorbate binding energies of small oxygen- and hydrogen-containing species on different facets of Pt relevant to the oxygen reduction reaction.<sup>10–12</sup> This vital development allows us to

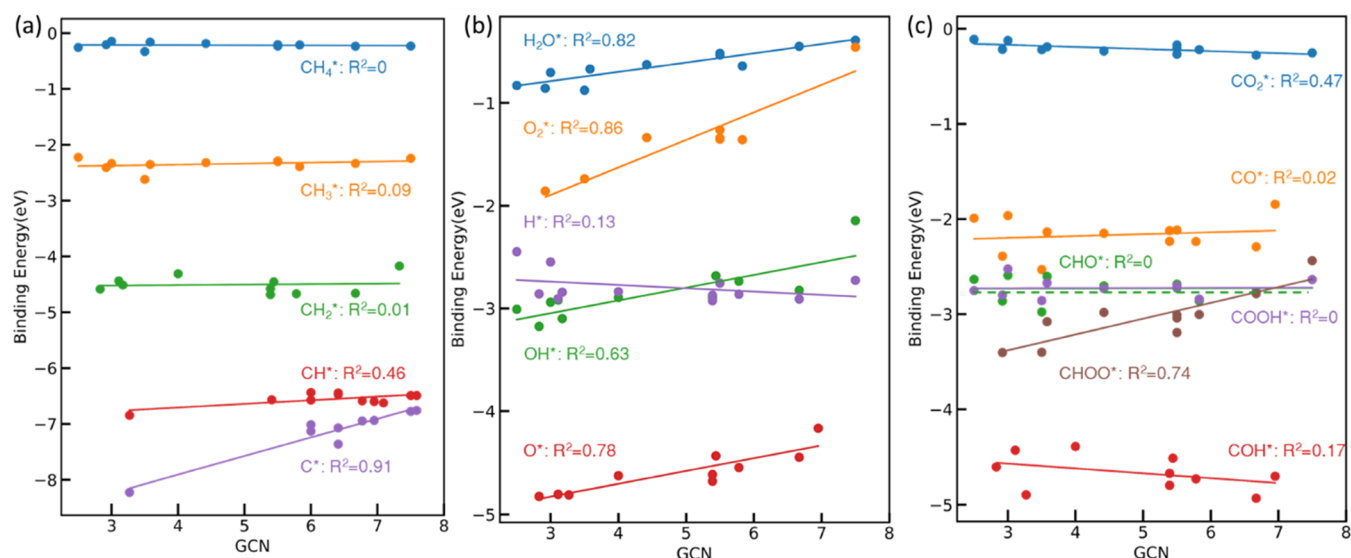
describe structure sensitivity and identify optimal facets, defects, and nanoparticle shapes.<sup>13</sup> However, the extendibility of the GCN model to other species of interest in heterogeneous catalysis has remained elusive. The physical understanding of the scalings is also limited, preventing the transferability to other species and catalysts. Based on a simple physical model with two descriptors,<sup>14</sup> the orbital-wise coordination number has also been proposed to predict the CO and O adsorption energies on Au surfaces. Despite advancing scaling relations, their accuracy and applicability are limited.<sup>15</sup> More accurate approaches beyond a single descriptor are required.

Machine learning (ML) promises to rapidly predict adsorbate binding on multiple adsorption sites of metal surfaces.<sup>16–25</sup> It could capture the adsorbate/surface nonlinear interactions and serve as a high-precision alternative to first-principles modeling.<sup>26</sup> Its usefulness requires easily computable electronic and geometric descriptors.<sup>15,26–29</sup> Takigawa et al.<sup>26</sup> developed an ML model with 12 readily available descriptors to predict adsorption energies of  $CH_4$ -related species on Cu alloys, with a root-mean-squared error (RMSE) <0.3 eV. Praveen and Comas-Vives<sup>28</sup> demonstrated a simple ML algorithm predicting the

Received: July 11, 2022

Published: September 12, 2022





**Figure 1.** Binding energies vs GCN (generalized coordination number). (a)  $\text{CH}_x^*$  species; (b)  $\text{OxHy}^*$  species; and (c)  $\text{COxHy}^*$  species. Regression coefficients  $R^2$  are given in Table S3.

binding energies with a mean absolute error (MAE) of 0.17 eV. Despite promising results, small DFT data and a lack of physical interpretations undermine several ML models.

Here, we combine DFT and ML to establish two models for predicting the adsorption energies of various species on Pt extended surfaces. First, we build a linear scaling GCN model for several species binding through C or O. We demonstrate that unlike the O-binding species in prior work,<sup>10,11</sup> the GCN model is generally inaccurate in describing the structure sensitivity of C-binding species. We accurately predict the oxygen- and carbon-binding species using a minimalistic ML model with descriptors free of DFT calculations involving the valency, molecular weight, number of bonds formed (features of the adsorbate), and the site type and GCN (features of the catalyst structure/active site). By accurately predicting adsorption energies of single-atom alloys using literature transition metal data, we demonstrate the transferability of our model even with such simple descriptors. We discuss the unifying aspects of this model for predicting many small adsorbates and including several site types, a challenge in conventional scaling relations, and its potential prediction of structure sensitivity of surface reactions and enhanced catalytic performance via particle engineering.

## METHODS

**DFT Settings.** DFT calculations were carried out using the Vienna ab initio Simulation Package<sup>30,31</sup> (VASP) and the Atomic Simulation Environment (ASE).<sup>32</sup> We used the projector augmented-wave (PAW) method<sup>33</sup> to describe the electron-ion interactions and the Revised Perdew–Burke–Erzerhof (RPBE) generalized gradient approximation (GGA) exchange–correlation functional.<sup>34</sup> The Kohn–Sham one-electron valence states were expanded in plane-wave basis sets with cutoff energy up to 400 eV. The Fermi population of the Kohn–Sham state was calculated with a Gaussian width of 0.1 eV, and all of the potential energies were extrapolated to 0 K. The convergence criterion for the self-consistent electronic minimization was set to  $10^{-6}$  eV, and the k-point sampling was  $3 \times 3 \times 1$ . At least 20 Å of vacuum was added vertically between

repeated images, and dipole moment corrections were applied to avoid unphysical periodic interactions.

All metal atoms in the two topmost layers of the slabs, the metal adatoms, and the adsorbates were allowed to relax in all directions until the maximum force on any atom was below 0.02 eV/Å. On each facet, we allowed surface species to fully relax on all possible adsorption sites (top, bridge, and hollow) and binding modes (monodentate, bidentate). The binding energy of species A was calculated as  $E_{\text{BE}} = E_{\text{A}^*} - E_{\text{A(g)}} - E_*$ , where \* is an adsorption site and A and A\* are the gas-phase and adsorbed states of a given species. The gas-phase species energies were calculated in cubic boxes of  $20 \text{ Å} \times 20 \text{ Å} \times 20 \text{ Å}$ , and no entropy or zero-point energy corrections were added to those values. Spin polarization was considered when calculating gas-phase energies. Only electronic energies were considered in this work. Zero-point energy corrections could be included easily in the future using frequency correlations with the coordination numbers.<sup>35</sup>

**Adsorbates and Metal Surfaces.** The adsorbates studied in this work included  $\text{CH}_x$  ( $x = 0-4$ ),  $\text{OH}_x$  ( $x = 0-2$ ),  $\text{O}_2$ , H, CO,  $\text{CO}_2$ , COH, CHO, COOH, and CHOO. These are critical intermediates in many important catalytic reactions, such as the water–gas shift reaction (WGS),<sup>36</sup> steam and dry reforming of methane,<sup>36</sup> methane total and partial oxidation reactions, Fischer–Tropsch synthesis,<sup>37,38</sup> and methanation reaction.<sup>39</sup> We chose the low-index surfaces (111), (100), and (110), the (211), (221), and (322) stepped surfaces, and surfaces with n metal adatoms, denoted as nAd@mkl, where m, k, and l are the Miller indices of a plane.

For a given surface atom, the GCN from Calle-Vallejo et al.<sup>10,11</sup> is the sum of weights of its nearest neighbors obtained by dividing their own usual coordination number CN with that of the bulk. We summarized all facets and the corresponding GCNs for all possible adsorption sites in Table S1.

**Machine Learning (ML) Algorithms.** In developing the machine learning algorithms, we included all possible adsorption sites (top, bridge, three-fold, and four-fold hollow) on each surface for adsorbates and stable configurations. Overall, the data set included a total of 295 DFT calculations.

The complete data set is provided as a [Supplementary File](#). We used the scikit-learn program<sup>40</sup> to train, cross-validate, and test all of the machine learning algorithms. Then, 80% of the data was randomly selected and used to train the model, while the remaining 20% was used to test the model performance. K-fold cross-validation ( $k = 10$ ) was used to compare different model performances. The XGBoost regressor<sup>41,42</sup> based on tree ensembles was used, and the hyperparameters were tuned using the GridSearchCV. The hyperparameters in all of the ML algorithms and the tuning process are described in [Table S6](#). Feature importance analysis was performed using the TreeExplainer method and SHAP values developed by Lee et al.<sup>43</sup> for interpreting the impact of having a certain value for a given feature. The correlation heat map was constructed using the Seaborn package.

## RESULTS AND DISCUSSION

**Structure Scaling Relations.** Selected Pt surfaces are shown in [Figure S1](#). Adsorption sites include various terraces, edges, and metal adatoms, spanning a broad range of GCN between 2.5 and 7.5 (see [Table S1](#)), sufficient to typify the heterogeneity of a real catalyst. 295 DFT calculations are done, but only the binding energies on the most favorable adsorption site are plotted against the GCN in [Figure 1](#). The difference between DFT calculations and experimental values<sup>44</sup> is much smaller than the energy variation across surfaces and materials, indicating the accuracy of our DFT results. The preferred adsorption sites of all species are listed in [Table S2](#). The maximum difference in the binding energy of an adsorbate among surfaces is an excellent proxy of structure sensitivity. The binding energies of  $\text{CH}_4^*$  and  $\text{CO}_2^*$  are smaller than 0.25 eV on all surfaces, indicating physisorption. The binding energies of  $\text{CH}_3^*$ ,  $\text{CH}^*$ ,  $\text{CHO}^*$ , and  $\text{COOH}^*$  vary less than 0.4 eV, implying low to modest structure sensitivity.

The statistics of linear regressions of GCN vs binding energies is summarized in [Table S3](#). The binding energies of  $\text{C}^*$ ,  $\text{O}_2^*$ ,  $\text{O}^*$ ,  $\text{H}_2\text{O}^*$ , and  $\text{CHOO}^*$  correlate well with the GCN ( $R^2 > 0.7$ );  $\text{CH}^*$  and  $\text{OH}^*$  correlate with the GCN to an extent. No good correlations exist for the  $\text{CH}_3^*$ ,  $\text{CH}_2^*$ ,  $\text{H}^*$ ,  $\text{CO}^*$ ,  $\text{COH}^*$ ,  $\text{CHO}^*$ , and  $\text{COOH}^*$  binding energies. Overall, GCN correlates well with the binding energies of species binding on the surface via an O atom ( $\text{O}_2^*$ ,  $\text{O}^*$ ,  $\text{H}_2\text{O}^*$ ,  $\text{OH}^*$ ,  $\text{CHOO}^*$ ). Our findings are consistent with those of Calle-Vallejo et al.,<sup>11</sup> who reported GCN as a good descriptor of oxygen-containing species ( $\text{O}_2^*$ ,  $\text{O}^*$ ,  $\text{H}_2\text{O}^*$ ,  $\text{OH}^*$ ,  $\text{OOH}^*$ ,  $\text{H}_2\text{O}_2^*$ ), with the  $\text{OH}^*$  correlation having the lowest  $R^2$ . The results expose that, unlike the O-binding species studied before, the GCN model does not apply to some C-binding species.

The intercept and the slope indicate the adsorption strength and structure sensitivity, respectively.  $\text{C}^*$  and  $\text{O}^*$  have the largest intercepts and the strongest binding energies, consistent with their large valency. The slopes and bond order conservation arguments indicate that  $\text{C}^*$  and  $\text{O}_2^*$  are the most structure-sensitive (slopes of 0.33 and 0.27, respectively), followed by  $\text{CHOO}^*$ ,  $\text{O}^*$ , and  $\text{OH}^*$  (slopes larger than 0.1). The results indicate that oxygen-binding species, e.g., alcohols, ethers, carbonyl-containing compounds, acids, etc., are structure-sensitive. In stark contrast, C-binding species, such as those found in methane activation and hydrogenolysis of larger hydrocarbons, are relatively structure-insensitive. Among the latter species,  $\text{C}^*$  is the only strongly structure-sensitive species. Since  $\text{C}^*$  can lead to coke, we may expect coking to be structure-sensitive.

We rationalize the regression results through the adsorption-site-type sensitivity and the electronic interaction of the metal atom(s) with the central atom. We take  $\text{CH}^*$  as an illustrative example of the former.  $\text{CH}^*$ 's most favorable adsorption site is the hollow site (the linear regression results are plotted in [Figure S2](#)). An  $R^2$  of 0.46 shows a weak correlation between the binding energy and GCN, possibly due to the freedom of the adsorption configuration on hollow sites. If  $\text{CH}^*$  were on a top site, it would have weaker binding but an excellent  $R^2$  of 0.95. On-top binding gives excellent correlations but the other-site-type binding does not; mixed site types break these correlations. When all of the Pt surface adsorption sites are considered, the GCN model does not hold even for  $\text{O}^*$  (see [Figure S3](#)), consistent with the findings of the Koyama group.<sup>45</sup> The ability to account for the site type is essential for accurate scalings. For the latter, we chose the  $\text{CH}_2^*$  and  $\text{O}^*$  adsorption on Pt (100) for illustration.  $\text{CH}_2^*$  and  $\text{O}^*$  both adsorb on a bridge site and have a valency of two. Bader charges ([Table S4](#)) indicate that the interacting Pt surface atoms transfer 0.08 electrons to  $\text{CH}_2^*$  but 0.28  $e^-$  to  $\text{O}^*$  upon adsorption. The more extensive charge transfer to  $\text{O}^*$  manifests a much stronger surface–adsorbate interaction. As the adsorbate varies, the electronic communication with the metal atom(s) is altered, rendering a pure GCN scaling relation inaccurate.

These findings shed light on the limitations of the GCN geometric descriptor model. GCN works well when the species adsorbs on the top site and/or strongly interacts with the metal surface. Using the most stable binding site leads to omitting a large set of valuable data and an inability to describe the actual site of adsorbates. To overcome these limitations, we turn to ML.

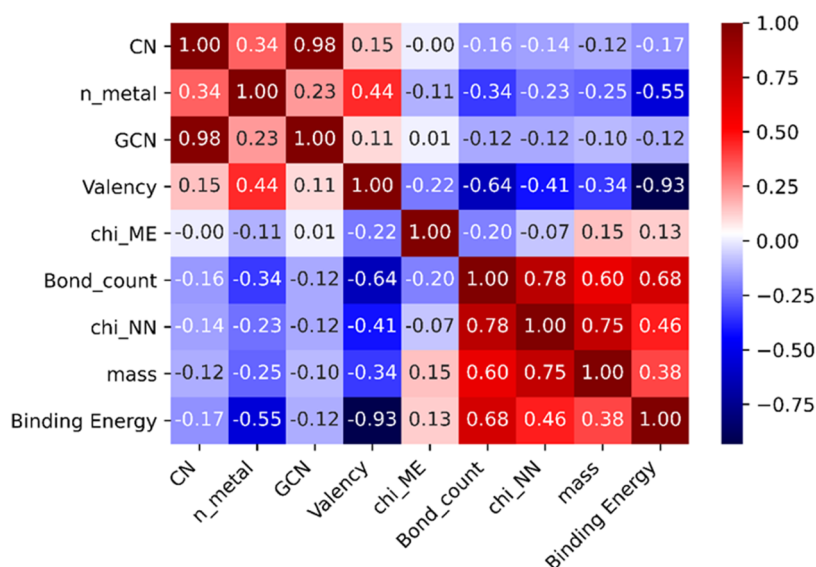
**Featurization and Correlations.** The choice of features is crucial for the predictive performance of the ML model. Inspired by previous studies,<sup>46,47</sup> eight geometric and electronic features of free adsorbates and bare metal surfaces (see [Table 1](#)) were

**Table 1. Features Used in Machine Learning (ML) Models<sup>a</sup>**

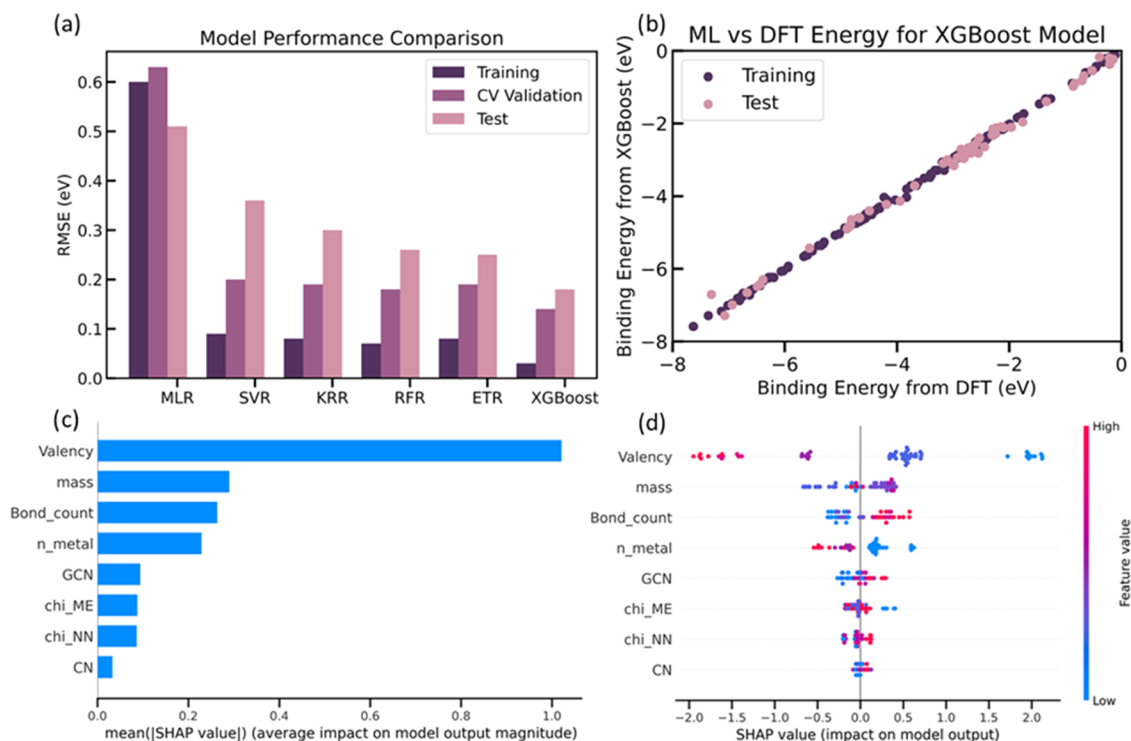
| category        | abbreviation | feature name                                             |
|-----------------|--------------|----------------------------------------------------------|
| free adsorbate  | valency      | adsorbate valency                                        |
|                 | bond_count   | number of bonds of the ME in the molecule                |
|                 | Chi_ME       | electronegativity of ME                                  |
|                 | Chi_NN       | electronegativity of the nearest neighbor atom of the ME |
|                 | mass         | molecular weight of the adsorbate                        |
| adsorption site | n_metal      | number of metal atoms composing the site                 |
|                 | CN           | coordination number of the site                          |
|                 | GCN          | generalized coordination number of the site              |

<sup>a</sup>ME Represents the Main Element Directly Bonded to the Metal Surface.

selected. No features related to metal properties were included in this model since all binding energies were calculated on Pt (for an extension, see below). All descriptor values can be obtained without any DFT calculations (see [Table S5](#)). One set of features includes the adsorbate valency of the main element (ME)<sup>46</sup> directly bonded to the metal surface, the number of bonds of the ME in the molecule (single, double, or triple bonds are differentiated), the electronegativity of the ME and its nearest neighbor in the molecule, and the molecular weight of the adsorbate. A second set includes the GCN, the coordination



**Figure 2.** Correlation between features and the binding energy. n\_metal represents the number of metal atoms involved in bonding, chi\_ME and chi\_NN represent the electronegativity of the main element (ME) and its nearest neighbor. See Table 1 for details of features. Deep red (blue) indicates a strong correlation (anticorrelation).

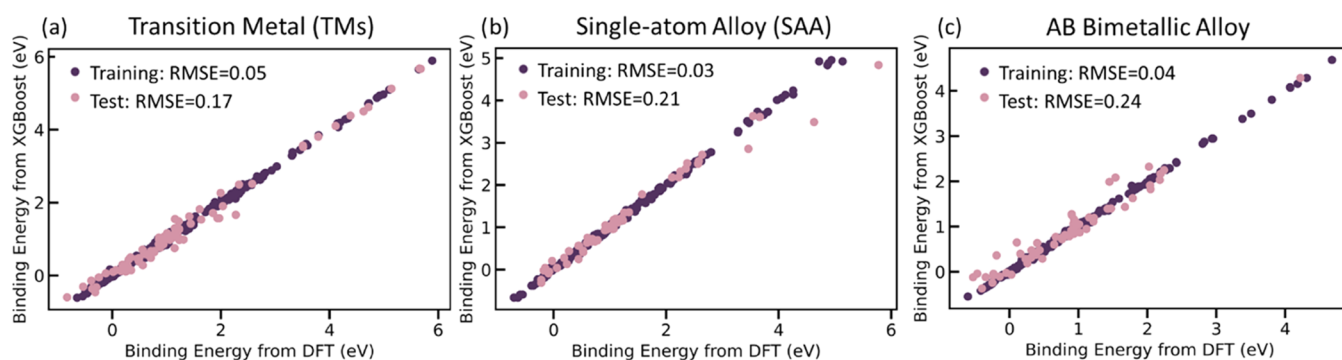


**Figure 3.** ML model performance and selection. (a) RMSE of various ML methods. (b) Binding energies from the XGBoost model vs DFT-calculated values. Color code: purple points are the training set (80% of data); pink points are the test set (20% of data). (c) SHAP values of every feature for all data points show the impact of each feature on the model output. (d) Feature importance based on mean SHAP values. See Table 1 for feature descriptions.

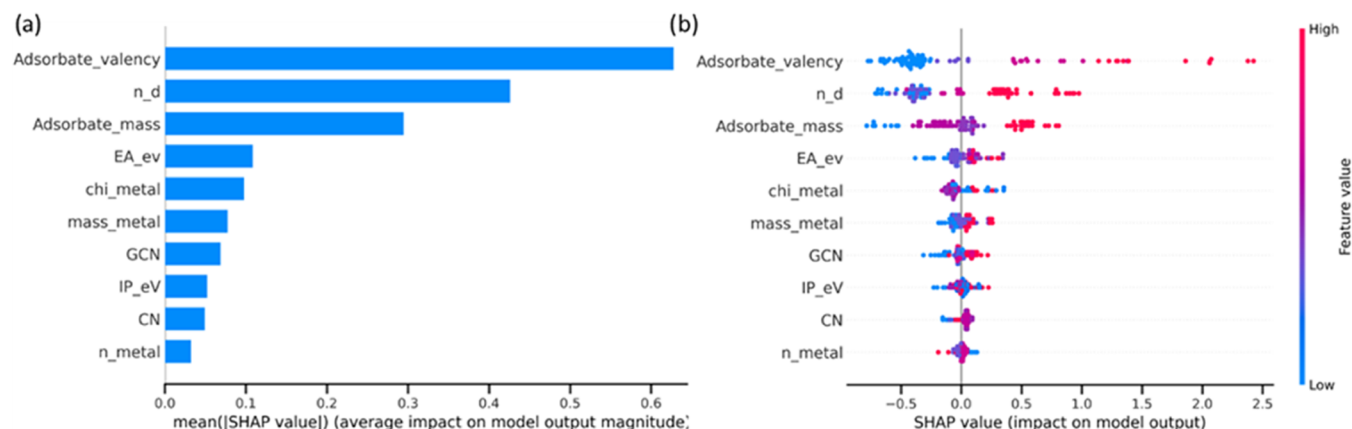
numbers averaged over the site ensemble,<sup>47</sup> and the number of metal atoms composing the adsorption site.

We analyze the correlation among all input features and also with the binding energy of the adsorbates. The resulting heat map is shown in Figure 2. Red (blue) represents direct (inverse) correlation. The geometric descriptors GCN and CN are highly correlated, as expected from the literature,<sup>48</sup> and the electronic descriptors of the adsorbate such as the valency, bond count and chi\_NN are also correlated with each other. Notably, the

adsorbate valency is inversely and most intensively correlated with the binding energy, followed by the bond count and the number of metal atoms of the adsorption site that characterize the site type. The discovery of the correlation of the atom valency with the binding strength follows the bond order conservation principles and is a fundamental aspect of the Norskov linear scaling relations but has not been discussed in the context of geometric descriptor-based models. The type of binding site is also fundamental to adsorption but has not been



**Figure 4.** Results of XGBoost model performance for (a) transition metals; (b) single-atom alloys; and (c) AB bimetallic alloys. Color code: purple points are the training set (80% of data); pink points are the test set (20% of data).



**Figure 5.** SHAP analysis for XGBoost feature importance of transition metals. (a) SHAP values of every feature for every data point show the impact of each feature on the model output. (b) Feature importance based on mean SHAP values. *n<sub>d</sub>* represents the number of d electrons of the metal, EA the electron affinity, *chi<sub>metal</sub>* the electronegativity, and IP the ionization potential. See Table S8 for more details.

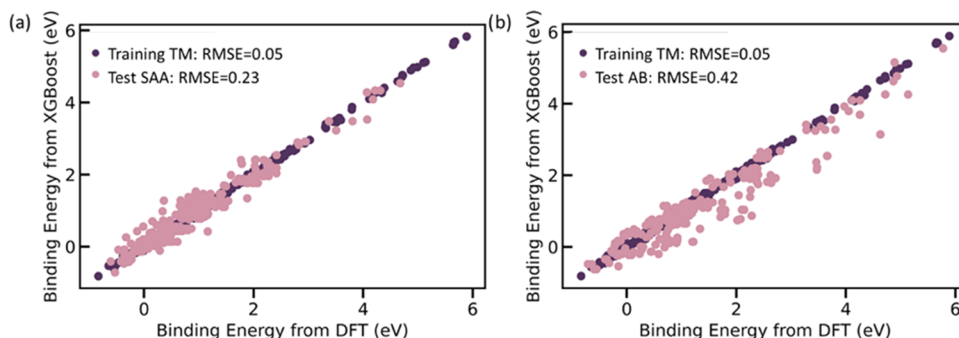
brought into the linear scaling relations, as far as we know, and can be rationalized by bond order conservation principles.

To predict the binding energies accurately, we exploited different machine learning algorithms<sup>49,50</sup> and their optimal model parameters (see Table S6). We tested linear, kernel, and tree ensemble models, including multivariate linear regression (MLR), support vector regression (SVR),<sup>51</sup> kernel ridge regression (KRR),<sup>41,52</sup> random forest regression (RFR),<sup>53</sup> extra tree regression (ETR),<sup>52</sup> and extreme gradient boosting regression (XGBoost). Only simple regression models were selected to ascertain interpretability. The root-mean-squared error (RMSE) was chosen as a performance metric (Figure 3a and Table S7). Unlike the GCN model applicable to a single adsorbate on a single (preferably the top) site type, the ML model can capture all adsorbates and types of adsorption sites at once and is interpretable. It is trained and can provide the most stable site of an adsorbate by comparing the energies for various types; the site with the lowest adsorption energy is the most stable.

The multivariate linear regression (MLR) model is the simplest one, with the test set's highest RMSE value of 0.51 eV. The XGBoost regression algorithm performs the best, with an RMSE of 0.18 eV for the test set. The performance of the SVR, KRR, RFR, and ETR models is good, with the algorithms in the same category, such as SVR and KRR, performing similarly. The XGBoost model performance is depicted in a parity plot in Figure 3b. Plots for the other models are shown in Figure S4. Compared to a previous ML model that used DFT-calculated

properties as descriptors with an RMSE of 0.4 eV,<sup>24</sup> our model shows a much lower error with simple descriptors. The accuracy of our model is similar to a model proposed previously<sup>28</sup> with an RMSE of 0.24 eV, in which the features contained orbital occupation energies that need expensive DFT calculations. Our proposed descriptors are intrinsic properties of the free adsorbates and can be easily and conveniently obtained for applications.

**Feature Importance Analysis.** Exposing the critical features controlling the binding energy is crucial. The XGBoost model's built-in feature analysis is known to be metrics-dependent, and the results are shown in Figure S5. To evaluate the feature importance of the XGBoost model in a unified way, the SHAP (SHapley Additive exPlanations) method proposed by Lee et al.<sup>43,54</sup> is applied. SHAP is a powerful AI technique commonly used to interpret machine learning results based on game theoretically optimal Shapley values. Figure 3c,d shows the feature importance and the effect of each feature on predicting binding energies. The top three most important features all correspond to the electronic properties of the adsorbate, with the adsorbate valency being the dominant feature. Smaller values of valency (well-coordinated) correspond to less negative binding energies. Adsorbate molecular weight and the number of formed bonds also affect the binding energies. The most important features for predicting binding energies are electronic properties, and the key geometric features are the type of adsorption site and the GCN. Feature importance analysis for other tested models is also performed, and the results are



**Figure 6.** Model transferability in predicting adsorption energies of (a) single-atom alloys and (b) AB bimetallic alloys by training the transition metal data.

consistent with the XGBoost model (see Figure S6). Adsorbate valency is the key feature in all cases, followed by bond count, molecular weight, and the type of the adsorption site.

**Transferability of ML Models.** To test the transferability of our selected descriptors and ML model, we extracted a published data set of adsorption energies from Reuter et al.<sup>47</sup> Their data set contains the adsorption energies of six adsorbates (C, CH, CO, H, O, OH) on fcc(111), (110), (100), and (211) facets of nine transition metals (TMs), single-atom alloys (SAA), and AB bimetallic alloys. This data set is internally consistent, obtained using the same DFT setup, and thus it is an ideal one to assess transferability. More details about the data set are given in [Supplementary Note S1](#). Because of varying the metal in this data set, we augmented the previous features with metal-related properties. Unlike prior work using the d-band properties, our selected features do not need DFT calculations. Using the filter-based feature selection method,<sup>55</sup> the final descriptor set includes five features for the metal (electronegativity ( $\chi_{\text{metal}}$ ), first ionization potential (IP), the number of d electrons ( $n_{\text{d}}$ ), electron affinity (EA), and the atomic mass of the metal atom ( $\text{mass}_{\text{metal}}$ )), three for the adsorption site ( $n_{\text{metal}}$ , GCN, CN) and two for the adsorbate (valency and  $\text{adsorbate}_{\text{mass}}$ ). Detailed feature description can be found in [Table S8](#), and features for each metal are summarized in [Table S9](#). With the proposed ten descriptors, we applied the XGBoost model to each data set. Each model's performance is shown in [Figure 4](#).

The TMs model has the smallest test error with an RMSE of 0.17 eV. For SAA and AB bimetallics, the errors are slightly larger than 0.20 eV. The TM model serves as a basic case, and its corresponding feature importance analysis using SHAP is shown in [Figure 5](#). The most important feature is still the adsorbate valency, consistent with our data set, indicating that it is a good universal descriptor. The number of the metal d electrons and the molecular weight of the adsorbate also play a vital role in predicting the adsorption energies. Even with multiple metals in the data set, two of the top three significant features are related to the adsorbate properties, and the features related to the geometric effects contribute less to predicting the adsorption energies.

Unlike typical ML models that are trained and tested on the same kind of data, a transferable model could predict a new system. To achieve this, the descriptors need to be general. First, we test the transferability of our methodology in predicting the adsorption energies of SAAs by training a model on the TM data, as shown in [Figure 6a](#). All metal-related descriptors are calculated as site-arithmetic mean values. The RMSE of the TM-based model on the SAA set is 0.23 eV, close to a ML model

trained on the SAA data set (0.21 eV in [Figure 4b](#)). Predicting adsorption energies using the TM model for this unseen data set is still good. The order of feature importance results (shown in [Figure S7](#)) is the same as the TMs model, suggesting that for SAAs, the dopant only has a subtle effect on the host, and the electronic effect is highly localized on the adsorption site. [Figure 6b](#) shows the model transferability of the TM model to the AB bimetallic alloys. The RMSE of the test AB set is 0.42 eV, comparable to some previous ML models<sup>26</sup> but larger than the model built on the data set itself (0.24 eV in [Figure 4c](#)). The results indicate that bimetallic alloys have relatively nonlocal effects in adsorption compared to SAAs, and the adsorption sites alone are insufficient to correlate very accurately the properties of TMs with AB alloys. The model could be used more for screening. Adding new descriptors or a few AB data points in the training set would further improve the model performance.

## CONCLUSIONS

Adsorption energies are fundamental to all surface processes, and their prediction has been central to modeling efforts and impactful. We presented two adsorption energy models of small species on metal surfaces based on nearly 300 DFT calculations. The first is the traditional linear scaling relation, invoking the generalized coordination number (GCN). It extends the original model from oxygen-binding to carbon-binding species and other selected adsorbates, such as CO, comprising the bulk of chemicals and fuels. Surprisingly, while the model holds well for the oxygen-binding species, it is generally inadequate for the carbon-binding species.  $\text{C}^*$  and  $\text{O}_2^*$  are the most structure-sensitive species. Small species of hydrocarbons appear less structure-sensitive than oxygenates. The correlations are accurate when adsorption entails a single metal atom but not as precise as the species become more dehydrogenated, leading to bridge or hollow binding. We overcome the limitations of the GCN model by building machine learning (ML) models with readily available properties of the gas-phase species and adsorption sites. The XGBoost model gives the lowest root-mean-squared error (RMSE) of 0.18 eV. Analysis reveals the essential features of adsorbates are the valency, the molecular weight, and the number of formed bonds, and of the catalyst active-center are the type and the generalized coordination number. While the valency and GCN have been essential in the published energy and structure scaling relations, respectively, the new features discovered herein unify adsorption of all C and O species, account for various and mixed site types rather than a single site, and improve accuracy. The transferability of the model is assessed, demonstrating its potential for predicting the

adsorption on more complex materials. We anticipate these models to advance catalytic science further.

## ■ ASSOCIATED CONTENT

### SI Supporting Information

The Supporting Information is available free of charge at <https://pubs.acs.org/doi/10.1021/acs.jcim.2c00872>.

(1) Details of the transition metal data set from the literature; (2) performance comparison and feature importance analysis of various tested machine learning (ML) models; (3) statistical analysis of GCN linear regression; (4) summary of implemented features in ML models; and (5) summary of optimized hyperparameters for each ML model (PDF)

Adsorption\_energy\_dataset: DFT-calculated data in this work and extracted data from literature (XLSX)

DFT\_sample\_inputs: sample input files for DFT calculations (ZIP)

python\_scripts: python scripts used for training machine learning models (ZIP)

## ■ AUTHOR INFORMATION

### Corresponding Author

Dionisios G. Vlachos — Department of Chemical and Biomolecular Engineering, University of Delaware, Newark, Delaware 19716, United States; Catalysis Center for Energy Innovation, RAPID Manufacturing Institute, and Delaware Energy Institute (DEI), University of Delaware, Newark, Delaware 19716, United States; [orcid.org/0000-0002-6795-8403](https://orcid.org/0000-0002-6795-8403); Email: [vlachos@udel.edu](mailto:vlachos@udel.edu)

### Author

Xue Zong — Department of Chemical and Biomolecular Engineering, University of Delaware, Newark, Delaware 19716, United States; Catalysis Center for Energy Innovation, RAPID Manufacturing Institute, and Delaware Energy Institute (DEI), University of Delaware, Newark, Delaware 19716, United States

Complete contact information is available at:

<https://pubs.acs.org/doi/10.1021/acs.jcim.2c00872>

### Author Contributions

X.Z. performed all DFT calculations and data analysis and developed the machine learning models. D.G.V. contributed to the idea and provided supervision. Both authors contributed to analyzing the results and writing.

### Notes

The authors declare no competing financial interest.

The data supporting this study's findings are available from the corresponding authors upon reasonable request. The Python code written to perform the machine learning analysis and the complete data set supporting the plots in this work are available on GitHub ([https://github.com/xzong0619/Adsorption\\_Machine\\_Learning](https://github.com/xzong0619/Adsorption_Machine_Learning)).

## ■ ACKNOWLEDGMENTS

Financial support from the RAPID manufacturing institute, supported by the Department of Energy (DOE) Advanced Manufacturing Office (AMO), Award Number DE-EE0007888-9.5, is gratefully acknowledged. RAPID projects at the University of Delaware are also made possible in part by funding provided by the State of Delaware. The Delaware

Energy Institute gratefully acknowledges the support and partnership of the State of Delaware in furthering the essential scientific research being conducted through the RAPID projects.

## ■ REFERENCES

- (1) Nørskov, J. K.; Bligaard, T.; Rossmeisl, J.; Christensen, C. H. Towards the Computational Design of Solid Catalysts. *Nat. Chem.* **2009**, *1*, 37–46.
- (2) Mpourmpakis, G.; Andriotis, A. N.; Vlachos, D. G. Identification of Descriptors for the CO Interaction with Metal Nanoparticles. *Nano Lett.* **2010**, *10*, 1041–1045.
- (3) Montemore, M. M.; Medlin, J. W. Scaling Relations between Adsorption Energies for Computational Screening and Design of Catalysts. *Catal. Sci. Technol.* **2014**, *4*, 3748–3761.
- (4) 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*, 016105–016108.
- (5) 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. *Chem. Sci.* **2013**, *4*, 1245–1249.
- (6) Hammer, B.; Nørskov, J. K. Why Gold Is the Noblest of All the Metals. *Nature* **1995**, *376*, 238–240.
- (7) Hammer, B.; Nørskov, J. K. Theoretical Surface Science and Catalysis—Calculations and Concepts. In *Advances in Catalysis*, Academic Press, 2000; Vol. 45, pp 71–129.
- (8) Hammer, B.; Nørskov, J. K. Electronic Factors Determining the Reactivity of Metal Surfaces. *Surf. Sci.* **1995**, *343*, 211–220.
- (9) Vojvodic, A.; Nørskov, J. K.; Abild-Pedersen, F. Electronic Structure Effects in Transition Metal Surface Chemistry. *Top. Catal.* **2014**, *57*, 25–32.
- (10) 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.
- (11) 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.
- (12) Wang, Y.; Su, Y.-Q.; Hensen, E. J. M.; Vlachos, D. G. Finite-Temperature Structures of Supported Subnanometer Catalysts Inferred via Statistical Learning and Genetic Algorithm-Based Optimization. *ACS Nano* **2020**, *14*, 13995–14007.
- (13) Núñez, M.; Lansford, J. L.; Vlachos, D. G. Optimization of the Facet Structure of Transition-Metal Catalysts Applied to the Oxygen Reduction Reaction. *Nat. Chem.* **2019**, *11*, 449–456.
- (14) Ma, X.; Xin, H. Orbitalwise Coordination Number for Predicting Adsorption Properties of Metal Nanocatalysts. *Phys. Rev. Lett.* **2017**, *118*, 036101–036105.
- (15) Jinnouchi, R.; Asahi, R. Predicting Catalytic Activity of Nanoparticles by a DFT-Aided Machine-Learning Algorithm. *J. Phys. Chem. Lett.* **2017**, *8*, 4279–4283.
- (16) Hansen, K.; Montavon, G.; Biegler, F.; Fazli, S.; Rupp, M.; Scheffler, M.; Von Lilienfeld, O. A.; Tkatchenko, A.; Müller, K. R. Assessment and Validation of Machine Learning Methods for Predicting Molecular Atomization Energies. *J. Chem. Theory Comput.* **2013**, *9*, 3404–3419.
- (17) Ward, L.; Agrawal, A.; Choudhary, A.; Wolverton, C. A General-Purpose Machine Learning Framework for Predicting Properties of Inorganic Materials. *npj Comput. Mater.* **2016**, *2*, 1–7.
- (18) Ulissi, Z. W.; Singh, A. R.; Tsai, C.; Nørskov, J. K. Automated Discovery and Construction of Surface Phase Diagrams Using Machine Learning. *J. Phys. Chem. Lett.* **2016**, *7*, 3931–3935.
- (19) Emery, A. A.; Saal, J. E.; Kirklin, S.; Hegde, V. I.; Wolverton, C. High-Throughput Computational Screening of Perovskites for

- Thermochemical Water Splitting Applications. *Chem. Mater.* **2016**, *28*, 5621–5634.
- (20) Peterson, A. A.; Christensen, R.; Khorshidi, A. Addressing Uncertainty in Atomistic Machine Learning. *Phys. Chem. Chem. Phys.* **2017**, *19*, 10978–10985.
- (21) 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.; Chan, K.; Nørskov, J. K. Machine-Learning Methods Enable Exhaustive Searches for Active Bimetallic Facets and Reveal Active Site Motifs for CO<sub>2</sub> Reduction. *ACS Catal.* **2017**, *7*, 6600–6608.
- (22) Toyao, T.; Maeno, Z.; Takakusagi, S.; Kamachi, T.; Takigawa, I.; Shimizu, K. I. Machine Learning for Catalysis Informatics: Recent Applications and Prospects. *ACS Catal.* **2020**, *10*, 2260–2297.
- (23) Schlexer Lamoureux, P.; Winther, K. T.; Garrido Torres, J. A.; Streibel, V.; Zhao, M.; Bajdich, M.; Abild-Pedersen, F.; Bligaard, T. Machine Learning for Computational Heterogeneous Catalysis. *ChemCatChem* **2019**, *11*, 3581–3601.
- (24) Nayak, S.; Bhattacharjee, S.; Choi, J.-H.; Lee, S. C. Machine Learning and Scaling Laws for Prediction of Accurate Adsorption Energy. *J. Phys. Chem. A* **2020**, *124*, 247–254.
- (25) Kitchin, J. R. Machine Learning in Catalysis. *Nat Catal* **2018**, *1*, 230–232.
- (26) Toyao, T.; Suzuki, K.; Kikuchi, S.; Takakusagi, S.; Shimizu, K. I.; Takigawa, I. Toward Effective Utilization of Methane: Machine Learning Prediction of Adsorption Energies on Metal Alloys. *J. Phys. Chem. C* **2018**, *122*, 8315–8326.
- (27) Pankajakshan, P.; Sanyal, S.; de Noord, O. E.; Bhattacharya, I.; Bhattacharyya, A.; Waghmare, U. Machine Learning and Statistical Analysis for Materials Science: Stability and Transferability of Fingerprint Descriptors and Chemical Insights. *Chem. Mater.* **2017**, *29*, 4190–4201.
- (28) Praveen, C. S.; Comas-Vives, A. Design of an Accurate Machine Learning Algorithm to Predict the Binding Energies of Several Adsorbates on Multiple Sites of Metal Surfaces. *ChemCatChem* **2020**, *12*, 4611–4617.
- (29) 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.
- (30) Kresse, G.; Furthmüller, J. Efficient Iterative Schemes for Ab Initio Total-Energy Calculations Using a Plane-Wave Basis Set. *Phys. Rev. B* **1996**, *54*, 11169–11186.
- (31) Kresse, G.; Furthmüller, J. Efficiency of Ab-Initio Total Energy Calculations for Metals and Semiconductors Using a Plane-Wave Basis Set. *Comput. Mater. Sci.* **1996**, *6*, 15–50.
- (32) Larsen, A. H.; Mortensen, J. J.; Blomqvist, J.; Castelli, I. E.; Christensen, R.; Dulak, M.; Friis, J.; Groves, M. N.; Hammer, B.; Hargus, C.; Hermes, E. D. The Atomic Simulation Environment—a Python Library for Working with Atoms. *J. Phys.: Condens. Matter* **2017**, *29*, 273002–273031.
- (33) Kresse, G.; Joubert, D. From Ultrasoft Pseudopotentials to the Projector Augmented-Wave Method. *Phys. Rev. B* **1999**, *59*, 1758–1775.
- (34) Hammer, B.; Hansen, L. B.; Nørskov, J. K. Improved Adsorption Energetics within Density-Functional Theory Using Revised Perdew-Burke-Ernzerhof Functionals. *Phys. Rev. B* **1999**, *59*, 7413–7421.
- (35) Lansford, J. L.; Mironenko, A. V.; Vlachos, D. G. Scaling Relationships and Theory for Vibrational Frequencies of Adsorbates on Transition Metal Surfaces. *Nat. Commun.* **2017**, *8*, No. 1842.
- (36) Foppa, L.; Margossian, T.; Kim, S. M.; Müller, C.; Copéret, C.; Larmier, K.; Comas-Vives, A. Contrasting the Role of Ni/Al<sub>2</sub>O<sub>3</sub> Interfaces in Water–Gas Shift and Dry Reforming of Methane. *J. Am. Chem. Soc.* **2017**, *139*, 17128–17139.
- (37) Foppa, L.; Iannuzzi, M.; Copéret, C.; Comas-Vives, A. Adlayer Dynamics Drives CO Activation in Ru-Catalyzed Fischer–Tropsch Synthesis. *ACS Catal.* **2018**, *8*, 6983–6992.
- (38) Foppa, L.; Iannuzzi, M.; Copéret, C.; Comas-Vives, A. Facile Fischer–Tropsch Chain Growth from CH<sub>2</sub> Monomers Enabled by the Dynamic CO Adlayer. *ACS Catal.* **2019**, *9*, 6571–6582.
- (39) Foppa, L.; Iannuzzi, M.; Copéret, C.; Comas-Vives, A. CO Methanation on Ruthenium Flat and Stepped Surfaces: Key Role of H-Transfers and Entropy Revealed by Ab Initio Molecular Dynamics. *J. Catal.* **2019**, *371*, 270–275.
- (40) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
- (41) Friedman, J. H. Greedy Function Approximation: A Gradient Boosting Machine. *Ann. Stat.* **2001**, *29*, 1189–1232.
- (42) Friedman, J. H. Stochastic Gradient Boosting. *Comput. Stat. Data. Anal.* **2002**, *38*, 367–378.
- (43) Lundberg, S.; Lee, S.-I. A Unified Approach to Interpreting Model Predictions. 2017, arXiv:1705.07874. arXiv.org e-Print archive. <https://arxiv.org/abs/1705.07874>.
- (44) Wellendorff, J.; Silbaugh, T. L.; Garcia-Pintos, D.; Nørskov, J. K.; Bligaard, T.; Studt, F.; Campbell, C. T. A Benchmark Database for Adsorption Bond Energies to Transition Metal Surfaces and Comparison to Selected DFT Functionals. *Surf. Sci.* **2015**, *640*, 36–44.
- (45) Rivera Rocabado, D. S.; Nanba, Y.; Koyama, M. Density Functional Theory and Machine Learning Description and Prediction of Oxygen Atom Chemisorption on Platinum Surfaces and Nanoparticles. *ACS Omega* **2021**, *6*, 17424–17432.
- (46) 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. *Nat. Commun.* **2020**, *11*, No. 1196.
- (47) Andersen, M.; Levchenko, S. V.; Scheffler, M.; Reuter, K. Beyond Scaling Relations for the Description of Catalytic Materials. *ACS Catal.* **2019**, *9*, 2752–2759.
- (48) Lansford, J. L.; Vlachos, D. G. Spectroscopic Probe Molecule Selection Using Quantum Theory, First-Principles Calculations, and Machine Learning. *ACS Nano* **2020**, *14*, 17295–17307.
- (49) Japkowicz, N.; Shah, M. *Evaluating Learning Algorithms: A Classification Perspective*; Cambridge University Press, 2011.
- (50) Murphy, K. P. *Machine Learning: A Probabilistic Perspective*; MIT Press, 2012.
- (51) Mozer, M. C.; Jordan, M. I.; Petsche, T. *Advances in Neural Information Processing Systems 9: Proceedings of the 1996 Conference*, MIT Press, 1997.
- (52) Geurts, P.; Ernst, D.; Wehenkel, L. Extremely Randomized Trees. *Mach Learn* **2006**, *63*, 3–42.
- (53) Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32.
- (54) Lundberg, S. M.; Erion, G.; Chen, H.; DeGrave, A.; Prutkin, J. M.; Nair, B.; Katz, R.; Himmelfarb, J.; Bansal, N.; Lee, S. I. From Local Explanations to Global Understanding with Explainable AI for Trees. *Nat. Mach. Intell.* **2020**, *2*, 56–67.
- (55) Hira, Z. M.; Gillies, D. F. A Review of Feature Selection and Feature Extraction Methods Applied on Microarray Data. *Adv. Bioinform.* **2015**, *2015*, 1–13.