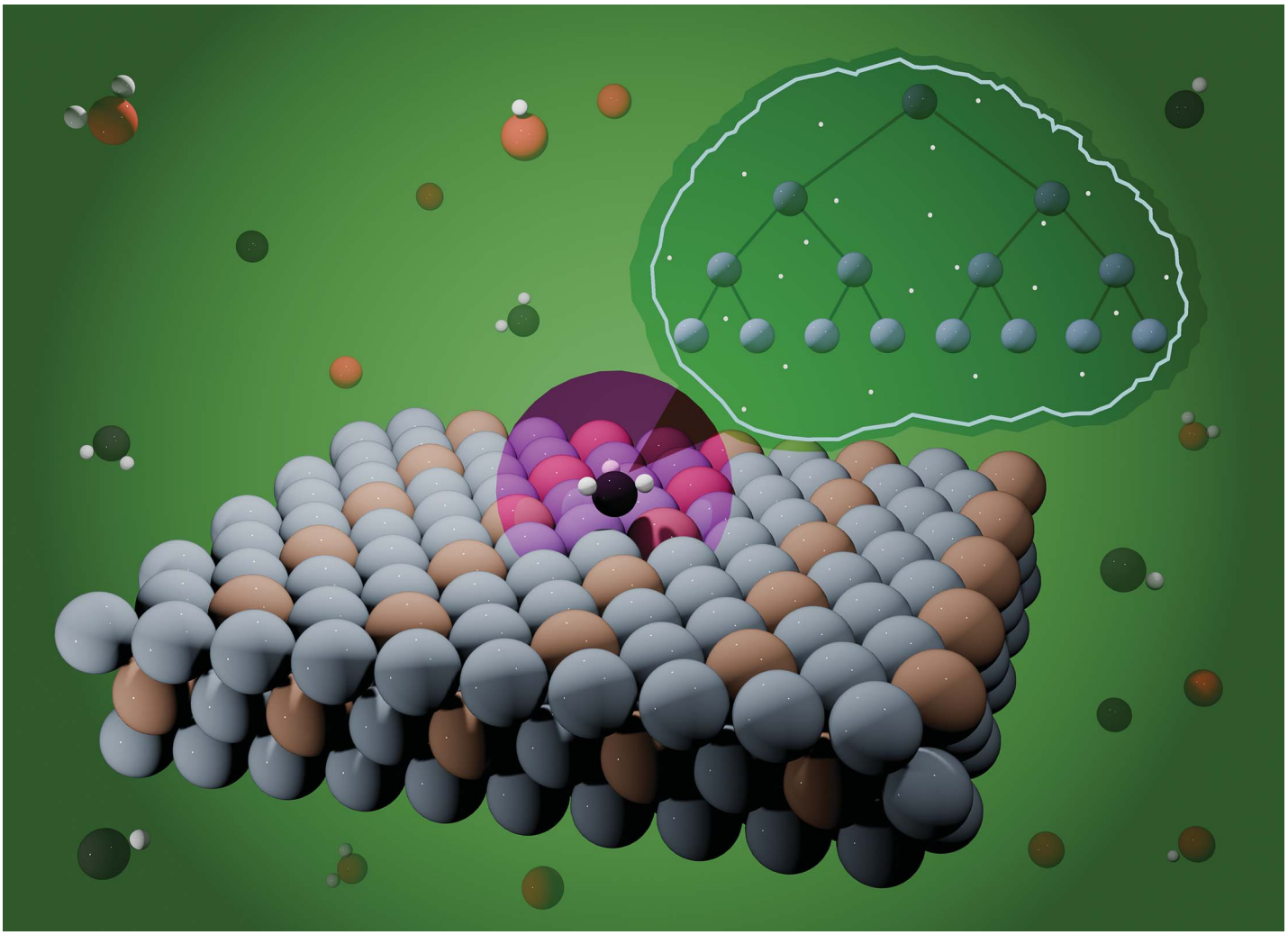
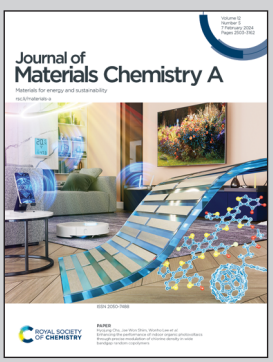


Showcasing the research from Professor A. Comas-Vives from the Institute of Materials Chemistry, TU Wien, Austria, and the Department of Chemistry, Universitat Autònoma de Barcelona, Catalonia, Spain.

Local descriptors-based machine learning model refined by cluster analysis for accurately predicting adsorption energies on bimetallic alloys

A Machine Learning model based on a tree regressor predicts adsorption energies on AB-type bimetallic alloys for species based on C, N, S, O, and atomic H using structural, electronic, and elemental properties. The approach uses local descriptors of the adsorption sites and is refined through cluster analysis. It offers valuable insights into bonding interactions, bringing a valuable tool for screening novel catalytic materials.

As featured in:



See A. Comas-Vives *et al.*, *J. Mater. Chem. A*, 2024, **12**, 2708.

ROYAL SOCIETY OF CHEMISTRY

rsc.li/materials-a

Registered charity number: 207890

Journal of Materials Chemistry A

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal | View Issue

Check for updates

Cite this: J. Mater. Chem. A, 2024, 12, 2708

Local descriptors-based machine learning model refined by cluster analysis for accurately predicting adsorption energies on bimetallic alloys†

A. F. Usuga, [a] C. S. Praveen [bc] and A. Comas-Vives *[ad]

Exploring the vast chemical compound space to provide activity–site relationships on bimetallic catalysts presents significant challenges. It also raises the necessity of developing methodologies capable of overcoming the cost of the computational screening of high-performing heterogeneous catalysts. In the present contribution, we introduce machine learning models enhanced by local descriptors related to the adsorption site for predicting adsorption energies. Additionally, we combined them with cluster analysis to bring valuable tools to detect anomalies in the database, thus enhancing the accuracy and robustness of the predictive models. This approach accurately predicts the adsorption energies of several species containing C, N, S, O, and atomic H adsorbed on AB-type bimetallic alloys with stoichiometric variation of A : B ratios. Among all the evaluated ML-based architectures, the CatBoost model exhibits the best performance with a MAE of 0.019 eV and 0.174 eV for the training and test sets, respectively. The cluster analysis highlights the importance of constructing descriptors containing physicochemical-intuitive insight for describing the bonding interactions. This methodology facilitates the recognition of electronic-structural trends of the surrounding local active site, thereby becoming a potential tool to screen adsorption energies and, ultimately, the catalytic activity.

Received 17th October 2023
Accepted 3rd December 2023

DOI: 10.1039/d3ta06316j

rsc.li/materials-a

1 Introduction

Improving existing catalytic materials for emerging technologies is a resource-intensive and arduous task. Among heterogeneous catalysts, bimetallic alloys have gained significant attention in this context,[1] primarily due to their unique ability to modulate their electronic properties as a function of the composition.[2] These materials are highly promising for catalytic applications due to the versatility of modifying their composition,[3] morphology, and exposed facets.[4] The turning point in utilizing these alloy-based materials is marked by establishing trends linking their activity to various structural and electronic parameters. Density Functional Theory (DFT) calculations are a key asset in rationalizing the catalytic activity at the atomic level and screening novel catalytic materials.[5] However, the computational cost of DFT simulations still limits the exploration of the vast chemical compound space and hinders the high-throughput computational screening of catalytic materials. While DFT calculations can establish correlations, such as scaling relations, between bonding interactions and a simple descriptor on monometallic systems, these correlations with single linear descriptors often fall short in capturing trends for the case of metallic alloys.[6–10] When transition metals-based catalysts are involved in reaction mechanisms, the d-band center is widely used as a descriptor to determine the reactivity, which characterizes the chemisorption based on the electronic structure within the d-band theory.[11,12] However, the accuracy of the d-band center diminishes when dealing with specific transition metals and fails to encompass attributes important in catalysis, such as the adsorption site and type of adsorbate.

Therefore, developing models to explore the vast chemical compound space and establishing connections between bimetallic alloys, their intrinsic characteristics, and their corresponding activity presents a significant challenge. Additionally, the models must account for synergetic effects as the bifunctional activity resulting from the diversity of available active sites.[13] A workaround to alleviate the resource-intensive DFT calculations is to embrace methodologies based on Machine Learning (ML) models.[14–18] ML-based models can potentially yield an accuracy comparable to DFT by effectively capturing complex and non-linear patterns. The progression of ML-based models to incorporate catalysts–adsorbates

[a]Department of Chemistry, Universitat Autònoma de Barcelona, 08193 Cerdanyola del Vallès, Catalonia, Spain. E-mail: aleix.comas@uab.cat

[b]International School of Photonics, Cochin University of Science and Technology, University Road, South Kalamassery, Kalamassery, Ernakulam, Kerala 682022, India

[c]Inter University Centre for Nano Materials and Devices, Cochin University of Science and Technology, University Road, South Kalamassery, Kalamassery, Ernakulam, Kerala 682022, India

[d]Institute of Materials Chemistry, TU Wien, 1060 Vienna, Austria. E-mail: aleix.comas@tuwien.ac.at

† Electronic supplementary information (ESI) available. See DOI: https://doi.org/10.1039/d3ta06316j

interactions has increased the evaluated system's sizes and the number of active sites in the surface models, including materials with strongly localized active sites to extended materials with several and delocalized ones. For strongly localized configurations, ML models have been applied to clusters,[19–22] single-atom catalysts represented by 2D layered materials,[23–26] and monometallic surfaces doped with Transition Metals (TMs).[27–29] The model's accuracy usually relies on the suitability of the features in describing the localized chemical environment of the adsorption site.

In extended systems, it is, however, not obvious to select structural descriptors that can accurately capture the bonding strength between different metallic alloys and adsorbates. The proposed descriptors can be broadly categorized into two main groups: averaged properties and local chemical features. Descriptors based on averaged properties emphasize the significance of incorporating details about each metal composing the alloy.[30–33] Furthermore, including a combination of semi-empirical parameters, such as geometric and tabulated atomic information, capable of differentiating between different surface alloys, enhances accuracy and robustness.[34–36] Similarly, predictive models incorporating electronic properties such as the d-band center,[37–40] Bader charges,[41] and surface energy[42,43] display better accuracy in capturing bonding interactions. These predictive models can effectively screen a broad spectrum of bimetallic systems.

These models may address limitations in describing properties that depend on the specific adsorption site. However, applying these proposed methodologies is constrained in systems with diverse morphologies, potentially losing insight into the factors determining the chemical adsorption strength. In this context, it has been suggested to incorporate descriptors that inherently rely upon the catalyst's active site to improve the poor differentiation in the clustering and prediction of bonding interactions.[44–46] Among the most relevant local chemical descriptors proposed to date, those focused on correlating the nearest neighboring surface atoms to the adsorption site have gained significant attention. These descriptors include the average of elemental properties,[47] atom-specific fingerprints derived from elemental properties,[48,49] and use ML models based on neural networks,[50–52] or graph neural networks.[53–56] However, these descriptors have difficulties adapting to other adsorbates or ML-based model architectures, hindering their transferability to other systems. Alternatively, performing separate correlations for each adsorbate and adsorption site is possible,[57–59] but this approach diminishes the extrapolation ability of the ML-based model.

The current work proposes a machine learning-based methodology to elucidate the catalytic activity of materials built upon bimetallic alloys. We present a strategy that employs ML-based models to establish correlations between the adsorption energy and the adsorption site for various adsorbates. We use predictive model architectures based on Linear Regressors (LR) and Random Forest Regressors (RFR) enhanced by gradient-boosting. Clustering techniques based on dimension reduction from the Uniform Manifold Approximation and Projection (UMAP) methodologies are then used, serving as anomaly detectors in the database. The present work lays the ground for a robust and universal method to predict the chemical bonding strength; it allows an extensive sampling of the chemical compound space defined by the local chemical environment of the adsorption site on bimetallic surfaces and the adsorbates. This study is directed towards comprehensively describing the surrounding local active site. The catalyst database consists of bimetallic alloys with A_nB_m stoichiometry, and the adsorbates are primarily composed of C, O, N, H, and S. The descriptors are formulated using electronic, geometrical, and atomic-elemental properties. These features employ heuristics to avoid human intervention during their construction. This methodology enables the identification of structural-electronic patterns within the local active site environment, serving as a valuable tool in catalyst discovery. Furthermore, the suggested analysis provides a methodology for recognizing accuracy-limit factors that affect high-dimensional models. Proposing an approach where only single-point DFT calculations on the clean surface and gas-phase adsorbate are needed before training the ML-based model.

2 Methodology

2.1 Dataset

The developed framework for predicting the adsorption energy on bimetallic alloys requires the optimized geometries of the clean surface, the gas-phase adsorbate, and the adsorbate over the surface. All the structures were selected from an already reported dataset at the Catalysis Hub repository.[60] The dataset focuses on (111) and (101) facets of FCC bimetallic alloys with A : B stoichiometric ratios of 0%, 25%, 50%, 75%, and 100% between the two elements. The original database was created to predict reaction energies for several elementary reactions. We extracted the relaxed geometries of the adsorbed species on metal surfaces from the database and performed further DFT calculations to set up descriptors for the clean metallic surfaces and the free adsorbates in the gas phase, respectively.

The dataset consists of alloys with A atoms (Ag, Au, Cu, Fe, Pt, or Zn) and B atoms encompassing all metals from groups 3 to 15 and from periods IV to VI, including Al. The chosen adsorbates are C, CH, CH_2, CH_3, H, N, NH, O, OH, H_2O, S, and SH, amounting to a total of 17 343 points in the database. The selected structures include several adsorption sites, namely top, bridge, and hollow. Additionally, the screening process explores several local environments for each adsorption site, obtained by considering a combination of neighboring atoms A and B, as reported in the original dataset. The target property is the adsorption energy of the adsorbates on the metallic surfaces, calculated as follows:

$$E_{\text{bin}} = E_{\text{ads/surf}} - (E_{\text{surf}} + E_{\text{ads}})$$

2.2 Computational details

The parameters for the simulations were adopted from the reported dataset.[60] The calculations were carried out within the

periodic density functional theory (DFT) framework with the plane wave-based Quantum ESPRESSO code.[61] We performed spin-polarized single-point energy calculations using the BEEF-vdW functional on the optimized geometries collected from the database. An optimized collection of ultrasoft pseudopotentials, taken from the open-source GBRV high-throughput pseudopotentials library, was employed for describing the core electrons.[62] The kinetic energy cutoff for wavefunctions and charge density were taken as 35 and 350 Ry, respectively. The Brillouin zone was sampled *via* the Monkhorst–Pack scheme with a $4 \times 4 \times 1$ gamma-centered grid.

2.3 Model architecture and implementation

2.3.1 Supervised machine learning models. We used supervised machine learning models to predict adsorption energies from labeled data. The present approach for predicting the adsorption energy as the target variable builds upon our own prior findings in monometallic systems,[63] where the best-performing models employ analogous descriptors to describe non-linear trends based on Random Forests Regressors (RFR) enhanced by gradient-boosting architectures. The RFR models constitute an ensemble learning method that elaborates on multiple decision trees. The predicted target variable is determined by averaging the predictions from the individual trees. Besides, gradient boosting is employed to augment the performance of decision trees. Its objective is to sequentially train trees that minimize the loss function by considering the gradients derived from it. We have chosen three regressor architectures based on gradient-boosting: XGBoost,[64] CatBoost,[65] and LightGBM.[66] In the first two architectures, the tree expansions follow a depth or level-wise approach, while in the latter, the tree expansions start with a single leaf, resulting in a leaf-wise growth strategy that enhances learning speed.

We also implemented linear regression-based architectures, such as Multiple Linear Regression (MLR) and Kernel Ridge Regression (KRR), to compare their performance with non-linear methods. To assess the performance of the models, we opted for the Mean Squared Error (MSE) and Mean Absolute Error (MAE) metrics as our chosen evaluation criteria. These error metrics are essential for quantifying the deviation of our framework from DFT results. Additionally, gaining insights into the influence of each descriptor on our proposed model is crucial. To address this, we employ the technique proposed by Lundberg and Lee,[67] known as SHapley Additive exPlanations (SHAP). SHAP deconstructs a prediction into a sum of contributions from each descriptor, where each contribution corresponds to a SHAP value. This technique identifies the most relevant correlations between the input features and the predicted outcome, providing valuable insights into the model's functioning.

2.3.2 Supervised clustering model. While our framework primarily focuses on using the proposed descriptors to predict the adsorption energy, we also explore potential trends derived from the same. Our model incorporates an extensive array of descriptors, forming a complex system that poses challenges when dealing with its high dimensionality. To address this limitation, we have used a strategy for reducing the dimensionality of the model *via* the Uniform Manifold Approximation and Projection (UMAP) technique.[68] When comparing various methods to reduce the dimensionality, UMAP stands out for its ability to better preserve the local and global structure of the data. The global structure refers to the degree of closeness among various clusters, while the local structure provides insights into the internal trends within a cluster. This quality proves highly valuable for assessing data similarity within clusters, enabling comparisons between neighboring clusters. Reduced dimensions offer insight into data similarity,[69] while individual values within these reduced dimensions lack inherent physical significance. Thus, we utilize the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) clustering method, implemented within the Scikit-Learn ML library, to comprehend the distribution of reduced dimensions using the labeling provided for the DBSCAN method.[70] The applicability of DBSCAN method is well suited for clusters characterized by compactness and distinct separation, a criterion met in the present case.

2.4 Model training

Given the varying magnitudes of the proposed descriptors, we applied a Robust Scaler transformation using the preprocessing tools of Scikit-learn[70] to normalize the features. Next, we randomly divided the dataset into training and test sets with 70–30 split ratio. Fine-tuning of hyperparameters is carried out based on the loss function metrics and further validated using K-fold cross-validation, employing 10 splits. We considered key parameters such as the number of estimators, maximal depth, and learning rate to enhance the accuracy of the gradient-boosting-based regressor models. The models were trained with simultaneous variations of these key parameters, and the optimal combination was selected based on smaller average errors and deviation in the splits of the K-fold methodology. We adjusted the number of estimators in increments of 200, ranging from 300 to 1500. The learning rate was explored within a range of 0.06 to 0.20, with increments of 0.02, and maximal depth was varied from 3 to 10. Finally, we evaluated the minimum number of training samples and L2 regularization to reduce over-parametrization. The minimum number of samples in the leaf was varied from 0 to 10, while the L2 regularization was evaluated from 0 to 2.0 in increments of 0.02. The tuning of each hyperparameter was carefully balanced to avoid over- and under-parametrized models, aiming at models with optimal predictive capabilities. For the KRR model, the only parameters to tune are the type of kernel and L2 regularization. Detailed information containing the parameters used for each model is provided in Table S1† of the Electronic Supplementary Information (ESI).†

3 Results and discussion

3.1 Feature engineering

The descriptors used for training our machine learning models were exclusively derived from the relaxed configurations of the

clean metal surfaces and the gas-phase adsorbates, employing heuristics to minimize intervention. The modeling of adsorption energy is performed with a set of 34 descriptors. The descriptors are organized separately for the surface and adsorbate and categorized into structural, electronic, and atomic/elemental properties. Our methodology involves constructing descriptors that capture the localized chemical environment of the adsorption site. This methodology for the surface inherently incorporates the geometrical influence associated with the adsorption configuration into most of the proposed features. The construction of surface-related descriptors follows the workflow shown in Fig. 1a. The methodology is based on assembling a sphere on the top of the surface, where the surface atoms inside the sphere are considered responsible for the bonding interactions in the local environment. The sphere's center is 1.5 Å above the top layer (*Z*-axis coordinate), and the *XY*-axis coordinates are determined from the relaxed adsorbate on the surface, as shown in Fig. 1b and c. Although the relaxed coordinates are used to set the sphere, the trained model can be used with arbitrary *XY*-axis coordinates for predicting the adsorption energy at that specific site. The final value for each feature is computed as the average property across all atoms contained within the cutoff sphere. The radius and height of the sphere were varied until suitable values for the general coordination number for every adsorption site were obtained. Still, these heuristics can be modified to compute the features with more or fewer atoms to model complex catalysts or surfaces with higher Miller indices. Electronic and geometric features are obtained from the DFT calculations, while atomic/elemental features are derived from atomic-tabulated properties. In addition, features like the Fermi energy, d-band center, and the work function were also used, although they do not inherently include geometric effects.

Furthermore, we use numerical values to represent different adsorption site types, *i.e.*, 1, 2, 3, and 4 for the top, bridge, fcc-hollow, and hcp-hollow, respectively. The labeling for every site is based on the assigned labeling of the database from which we extract the structures. Finally, the descriptors related to gas-phase adsorbates are developed by considering only the primary element that bonds to the surface, which in the present

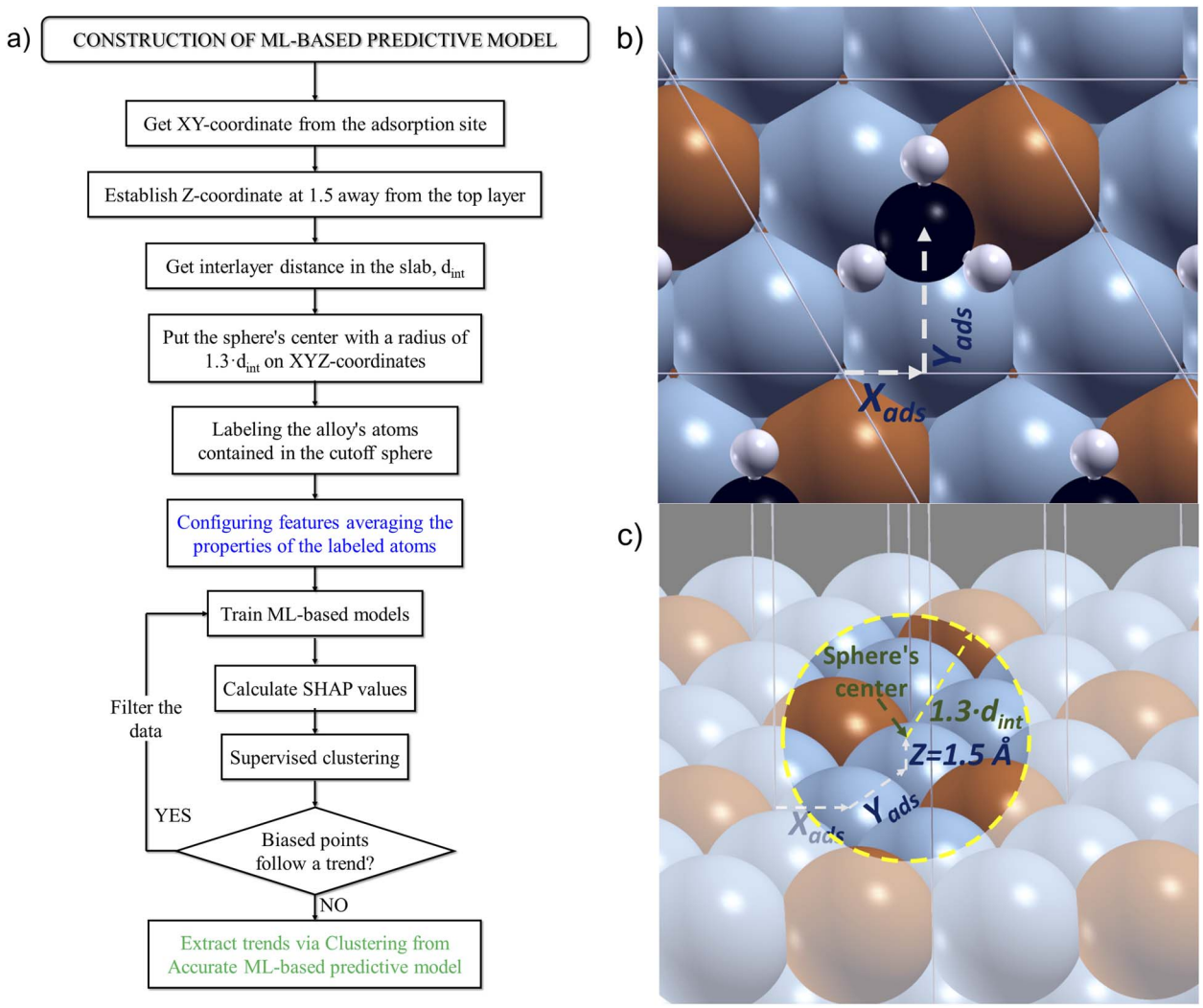


Fig. 1 (a) Schematic workflow of the construction of the surface-related features. (b) Extraction of the *xy*-coordinates for the adsorption site over the surface. (c) Establishing the atomic coordinates for the center of the cutoff sphere of the top surface layer.

case are C, H, N, O, or S. As our database does not include a large number of adsorbates and due to the representation approach, the adsorbate-related features are numerically stratified. The properties used in constructing the descriptors are not only proposed to correlate the adsorption energy but also contribute to providing interpretable physical insights about bonding interactions, as we will show later. A complete list of the used features can be found in the ESI (Table S2†).

3.2 Performance of features

We first explored the correlations between the proposed descriptors and the adsorption energy. Fig. 2 shows Pearson's coefficients, indicating the strength of linear relationships between two attributes, with values closer to ±1.0 signifying stronger correlations. In our case, we observed a weak or near-zero correlation between each descriptor and the adsorption energy. However, it is important to note that linear models frequently fall short in capturing the intricate behavior of adsorption on bimetallic alloys, particularly when considering various types of adsorbates and adsorption sites, as we aim to incorporate in our model. In addition, linear correlations from common descriptors, such as the d-band center, often yield inadequate single-descriptors for adsorption energies. A moderate single-feature linear correlation is evident between the number of surface atoms directly bonded with the adsorbate (denoted as "atoms_surf") and the coordination number and the electronic density at the adsorption site (denoted as "stm_surf"). However, this correlation does not extend to the adsorption energy. It is worth emphasizing that "atoms_surf" directly encodes the adsorption site using a numerically categorical stratification. This implies no discernible trend in adsorption energy based on adsorption sites and type of adsorbates. Thus, it is expected that Machine Learning (ML) models able to capture non-linear patterns will exhibit an improved performance in linking the proposed descriptors with adsorption energies. However, these models must support interpretability to extract physicochemical insights.[71,72]

Subsequently, we evaluated different machine learning regressor architectures to assess if we could enhance the performance of the model to predict adsorption energies *via* local chemical environment descriptors when changing from linear to highly complex non-linear models. The linear predictive model used is based on Multiple Linear Regressors (MLR), while the selected non-linear models correspond to Random Forest Regressors (RFR) enhanced by gradient boosting, in particular, CatBoost, XGBoost, and LightGBM. Additionally, a modification of the linear architecture with a kernel is employed, particularly a Kernel Ridge Regression (KRR). The results of our implemented ML-based models are summarized in Tables 1 and 2. The MLR-based model showcases the weakest overall performance, consistent with the trends observed in the single-feature linear scores from Pearson's coefficients. By employing the KRR-based model, we observed that mapping the descriptors with a Laplacian Kernel leads to a significant improvement. The KRR implementation achieved an average error and R^2 accuracy score in the training set close to that of RFR-based models. The random forest regressor-based models show a similar R^2 accuracy score but slightly higher for the XGBoost model. The CatBoost model displays a higher degree of deviation or bias for the 10 splits of the K-fold methodology. In contrast, the XGBoost outperforms the others concerning the bias in the training set.

To determine the most robust model, comparing the regression metrics for both the training and test sets is needed. While the Mean Squared Error (MSE) penalizes higher-biased points and tends to favor models with smaller robustness. Therefore, to overcome the mentioned MSE limitations, our analysis is based on the Mean Absolute Error (MAE). Using this

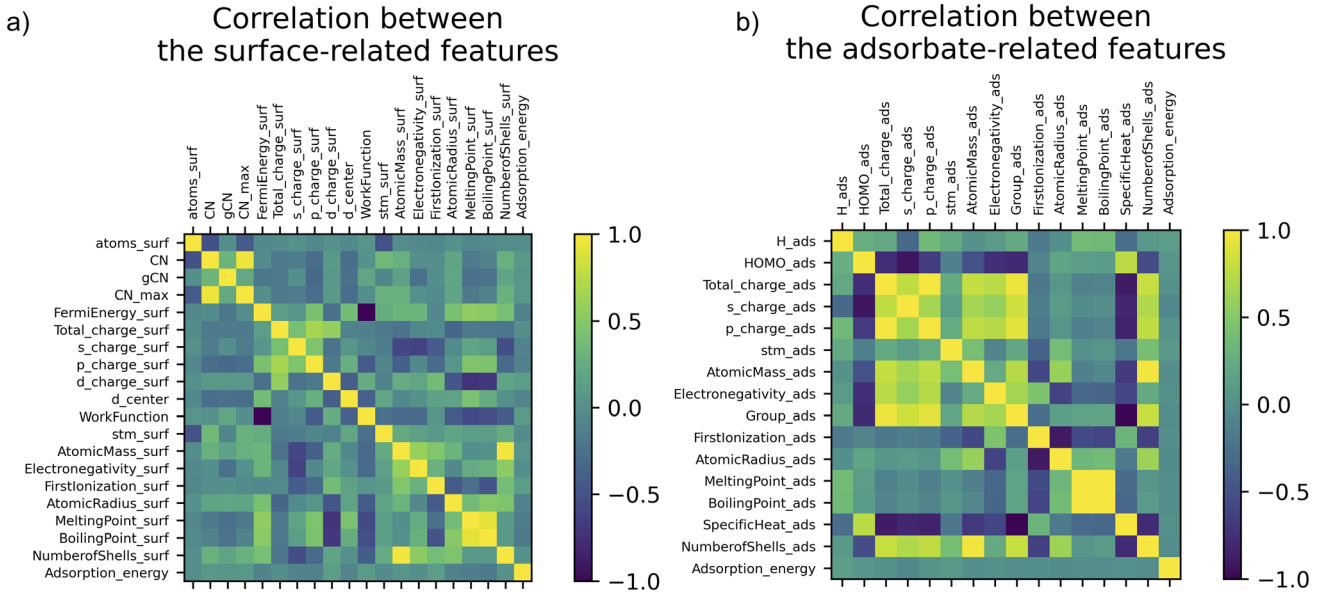


Fig. 2 The linear correlation with Pearson's coefficients between the features and adsorption energy. The data is divided into features related to (a) the clean surface and (b) the gas-phase adsorbate.

Table 1 Summary of the performance of the evaluated ML-based models. Mean accuracy and standard deviation for 10 splits on the K-fold methodology

ML-based model	R^2 (Accuracy score)	Standard deviation %
CatBoost	0.908	1.032
XGBoost	0.910	0.847
LightGBM	0.904	0.883
Kernel ridge	0.900	0.990
Linear regression	0.780	1.452

metric, the CatBoost model exhibits superior performance with a MAE of 0.166 eV for the training set and 0.280 eV for the testing set, followed by the XGBoost model with MAEs of 0.175 and 0.289 eV for the training and test data, respectively.

To thoroughly evaluate the performance of our machine learning-based models, we assess their predictive accuracy for the chemical adsorption strength in both training and testing sets. All RFR architectures perform similarly, with average accuracy of greater than 0.90 (Table 1) and similar MAE and MSE values for the train and test sets (Table 2). Among them, the CatBoost architecture has the smallest MAE value, but presents a higher deviation *via* the K-fold methodology compared to XGBoost and LightGBM algorithms: 1.032% *vs.* 0.847 and 0.833%, respectively. Rationalizing the bias in the CatBoost model requires comparing predicted and reference values of adsorption energies. Fig. 3 illustrates the parity plots, depicting the relationship between DFT-calculated and ML-predicted adsorption energies. At first glance, the CatBoost model may seem to predict energies with a higher bias than the XGBoost model. However, the CatBoost model has fewer scattered data points, suggesting that the CatBoost model performs better overall in MAE, as it has a higher number of points with a smaller MAE than the XGBoost one. The XGBoost and CatBoost models have similar RFR architecture, but with our features and hyperparameters, the CatBoost architecture is more effective in correlating them with the bonding strength. The predictive models generally display non-uniform deviation in specific points, particularly with higher deviations falling within the adsorption energy range of −4.0 to 0.0 eV. This range includes weak adsorption strengths, where the CatBoost model

Table 2 Summary of the performance of the evaluated ML-based architectures, including the metric of the average errors and R^2 score for both the training and testing datasets: Mean Absolute Error (MAE), Mean Squared Error (MSE), and R^2 score

ML-based models	MAE of train [eV]	MSE of train [eV]	R^2 score of train	MAE of test [eV]	MSE of test [eV]	R^2 score of test
CatBoost	0.166	0.116	0.970	0.280	0.279	0.927
XGBoost	0.175	0.105	0.973	0.289	0.262	0.932
LightGBM	0.170	0.104	0.973	0.299	0.279	0.928
Kernel ridge	0.151	0.105	0.973	0.308	0.319	0.917
Linear regression	0.671	0.848	0.781	0.660	0.823	0.786

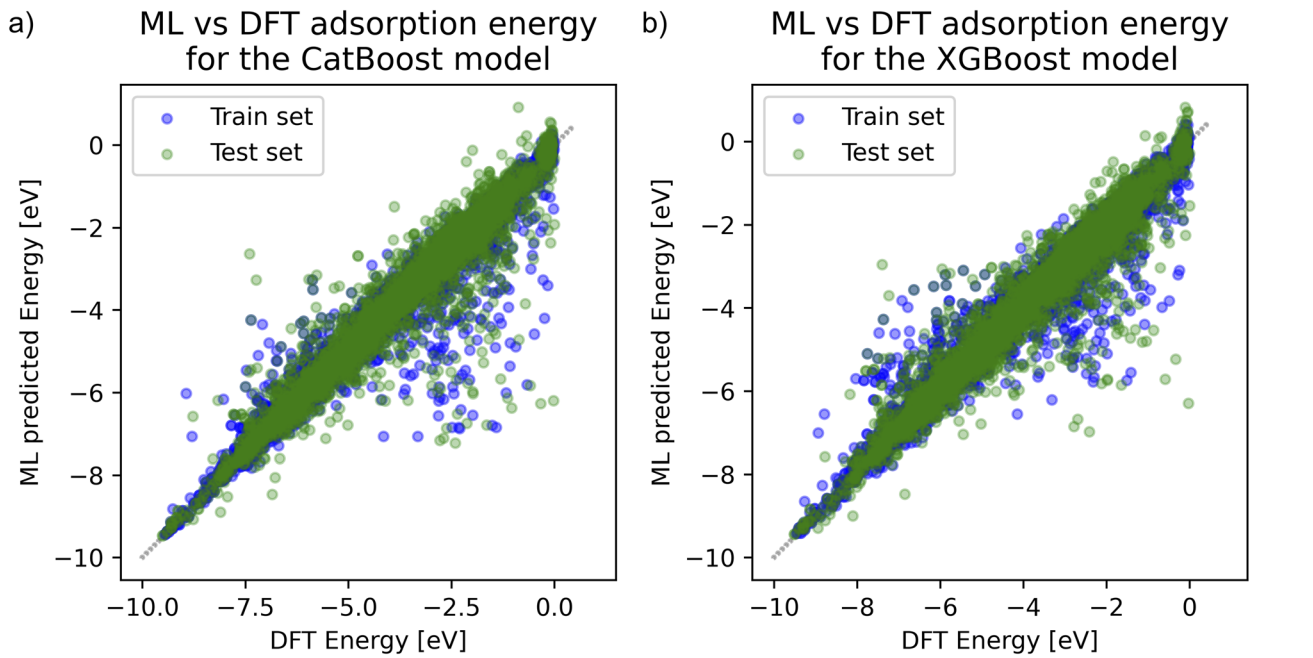


Fig. 3 Parity plot of DFT-calculated *vs.* ML-predicted adsorption energies for (a) the CatBoost and (b) XGBoost models.

tends to overestimate the adsorption energy. The observed behavior suggests that our proposed descriptors may not be adequately configured to capture the bonding interactions associated with these specific data points, an aspect evaluated *via* cluster analysis (*vide infra*).

Then, we use the SHAP methodology to obtain the impact of each suggested descriptor. The SHAP methodology evaluates the influence of individual descriptors, providing graphical insights into the most impactful features on the adsorption energy. In Fig. 4 the SHAP values are summarized, comparing the CatBoost and XGBoost models. When comparing both models, it becomes apparent that the most influential features are generally alike. However, a notable distinction arises in the CatBoost model, where there is a more pronounced reduction in the impact of each feature. This implies that the CatBoost model forces more features to correlate with the adsorption energy. The most influential is the d-partial orbital charge. This outcome aligns with expectations, considering the known dependency of the d-band for bonding interactions in bimetallic catalysts.[12,73]

Similarly, other significant surface-related descriptors following the same trend are the number of surface atoms directly bonded with the adsorbate ("atoms_surf") and the s-partial orbital charges. Regarding descriptors related to the adsorbate, the HOMO energy of the adsorbate ("HOMO_ads") and the number of hydrogen bonds in the adsorbate ("H_ads") significantly influence the predicted adsorption energy. Taking all the observed trends into account, as depicted in Fig. S1 and S2,† it becomes evident that surface-related features exhibit a more extensive influence on predicting the output compared to the relatively limited impact range of adsorbate-related features. Grouping the effect of surface-related features from Fig. S3–S5† in descending order, they are categorized into electronic properties, geometric attributes, and elemental information. Additionally, it can be deduced from the influence of the majority of elemental descriptors that these have a more limited impact on the predictive models. Moreover, it is worth noting that although the electronic properties and geometric attributes include the local description of the adsorption site, they also account for surface structural effects partially extending beyond the area of the surface site.

We further explore the correlations among the most influential features using the SHAP analysis methodology. Our results reveal a categorical tendency between the HOMO of the adsorbate ("HOMO_ads") and the d-partial orbital charges of the surface atom ("d_charge_surf") features, as depicted in Fig. S6.† This categorical trend is expected due to the stratification employed in constructing the adsorbate-related features. The stratification serves as a support for performing clustering within our database. However, clustering the raw database is difficult due to the challenges in analyzing a system with high dimensionality. To decrease the complexity, we opted for a two-dimensional database reduction to explore internal trends. We applied the Uniform Manifold Approximation and Projection (UMAP) methodology to reduce the data dimensionality of the training set. The UMAP reduction is based on the similarity function that preserves local and global information. The initial clustering was performed on the reduced values for the raw normalized database, as depicted in Fig. 5a. The clustering results of the raw data reveal an inadequate clustering, generating indistinct trends in the adsorption energy between each cluster. Cooper *et al.*[74] suggested employing the SHAP values instead of the raw data to improve the poor cluster differentiation and get a better local structure in the data for each cluster. Fig. 5b depicts the clustering using the SHAP values giving the adsorption energy as a reference. It is clear that using the SHAP values generates clusters better classified *via* the 2D reduced features (Fig. 5a). The reduced data from the SHAP values exhibits two discernible trends from the bonding interaction range. Specifically, data points within the middle to high range of adsorption energy show significant similarity, contrasting with the lower values. Continuing extracting information into the data, we have depicted the disparity between DFT-evaluated and ML-predicted values in Fig. 5c. It is noticeable that the points with higher bias are concentrated within specific

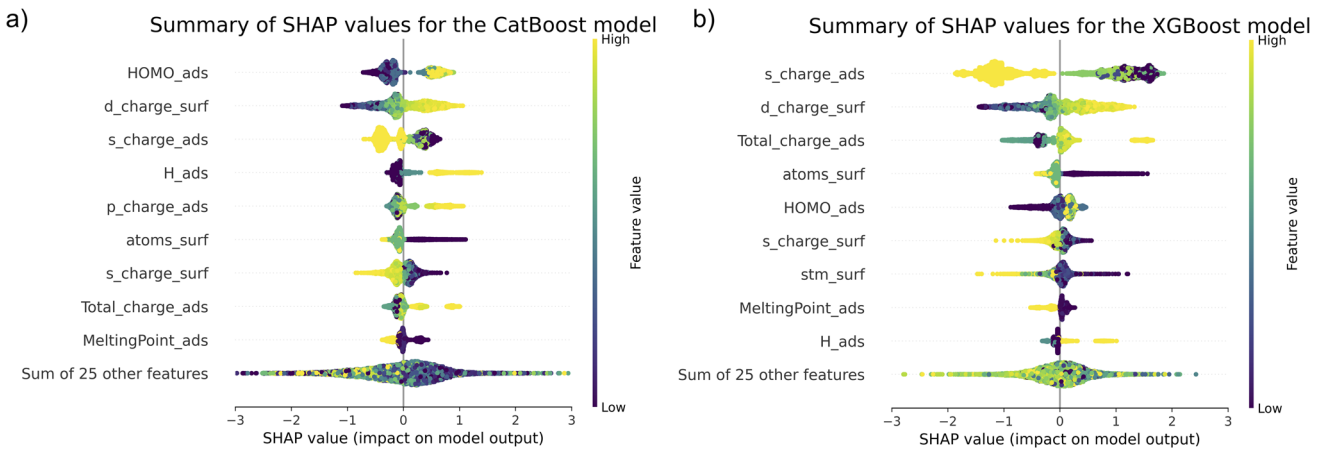


Fig. 4 Summary of the distribution of SHAP values for each descriptor for the architectures: (a) CatBoost and (b) XGBoost. The feature value scale on the right is the same for both ML-based models.

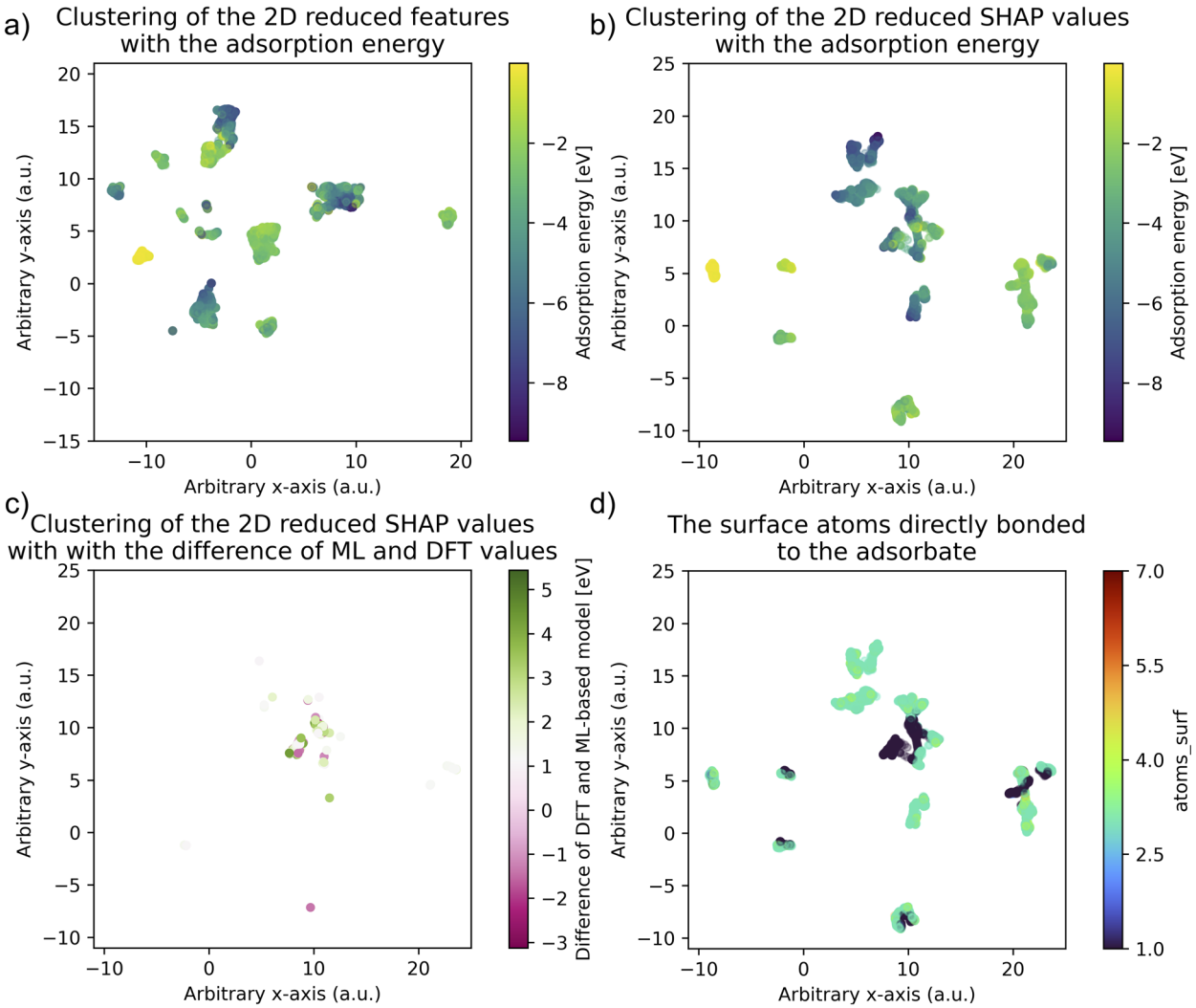


Fig. 5 Dimensionality reduction using the UMAP methodology for (a) applied to all raw normalized data (involving the proposed descriptors) and (b) the SHAP values specifically for the CatBoost model, with the 2D reduced dimension alongside adsorption energy employing the training set. 2D reduced dimension visualization of SHAP values using the CatBoost model for (c) the disparity between DFT-evaluated and ML-predicted adsorption energies values for the points exhibiting higher bias (−1 eV> or 1 eV<), and (d) the surface atoms directly bonded to the adsorbate ("atoms_surf").

clusters, indicating a significant similarity between these points. We labeled each cluster with each feature as a reference to identify the responsible factors for reducing the accuracy of our models.

Our analysis revealed that data points with an "atoms_surf" equal to 1 within the middle to high adsorption energy range introduce higher bias. In other words, some points representing top adsorption sites exhibit lower correlation, as depicted in Fig. 5d. Associating the results in Fig. 3 and 5d, we conclude that our ML-based models tend to overestimate the weakest bonding interactions when fitting the highest adsorption energies at the top sites. It should be noted that the points with a MAE exceeding 1.0 eV account for only 4.3% of the training set, representing a relatively small proportion. Potential solutions to address this issue include increasing the sampling within the range of weak adsorption strengths. Furthermore, it is worth emphasizing the robust architecture of the gradient-boosting methodologies for minimizing the loss function, but when the bias is still high, it does not always imply conserving the physical insight of the model. We will show later this aspect is solved *via* cluster analysis refinement (*vide infra*).

3.3 Database enhanced with anomaly detectors

Supervised clustering serves to identify the internal correlation of points with a higher bias, *i.e.*, as an anomaly detection technique. Employing this methodology, we removed the top sites from the database to retrain the ML-based architectures. Therefore, it allows analyzing the effect on the precision of our proposed surrounding local environment descriptors and to confirm whether the top sites need to be better accounted for in our model and the database. The total number of points in our database decreased from 17 343 to 13 894, representing a reduction of 19.9%. After filtering out this data, the average accuracy significantly increases, with the CatBoost model

Table 3 Summary of the performance of the evaluated ML-based models extracting the top sites. Mean accuracy and standard deviation for 10 splits on the K-fold methodology

ML-based model	R^2 (Accuracy score)	Standard deviation %
CatBoost	0.975	0.388
XGBoost	0.974	0.463
LightGBM	0.975	0.475
Kernel ridge	0.967	0.465
Linear regression	0.855	1.243

having the highest precision. Nevertheless, all RFR architectures yield similar results. Regarding the metrics of averaged errors used, *i.e.*, the MAE, the results are consistent with the predictive model for the entire database. The CatBoost demonstrates the highest robustness, but XGBoost and LightGBM architectures show comparable outcomes, as depicted in Tables 3 and 4.

The decrease in dispersion within the predictive models, excluding the top sites, is depicted in Fig. 6a. Moreover, Table 4 shows a notable reduction of 37.9% for the MAE and 58.1% for the MSE on the test set compared to Table 2 for the CatBoost model. For the training set, achieving a fitting with minimal deviation from the DFT-evaluated adsorption energy using the proposed descriptors based on the surrounding local environment approach is achieved. Simultaneously, the influence of each feature on the predictive model appears to diminish, as shown in Fig. 6b. Notably, the "atoms_surf" descriptor presents a significant reduction in its impact on the model. The electronic properties of both the surface and adsorbate still play a crucial role in correlating the bonding interaction, just in this case, described by the d-partial orbital charge at the surface and the s-partial orbital charge at the adsorbate. Finally, we can

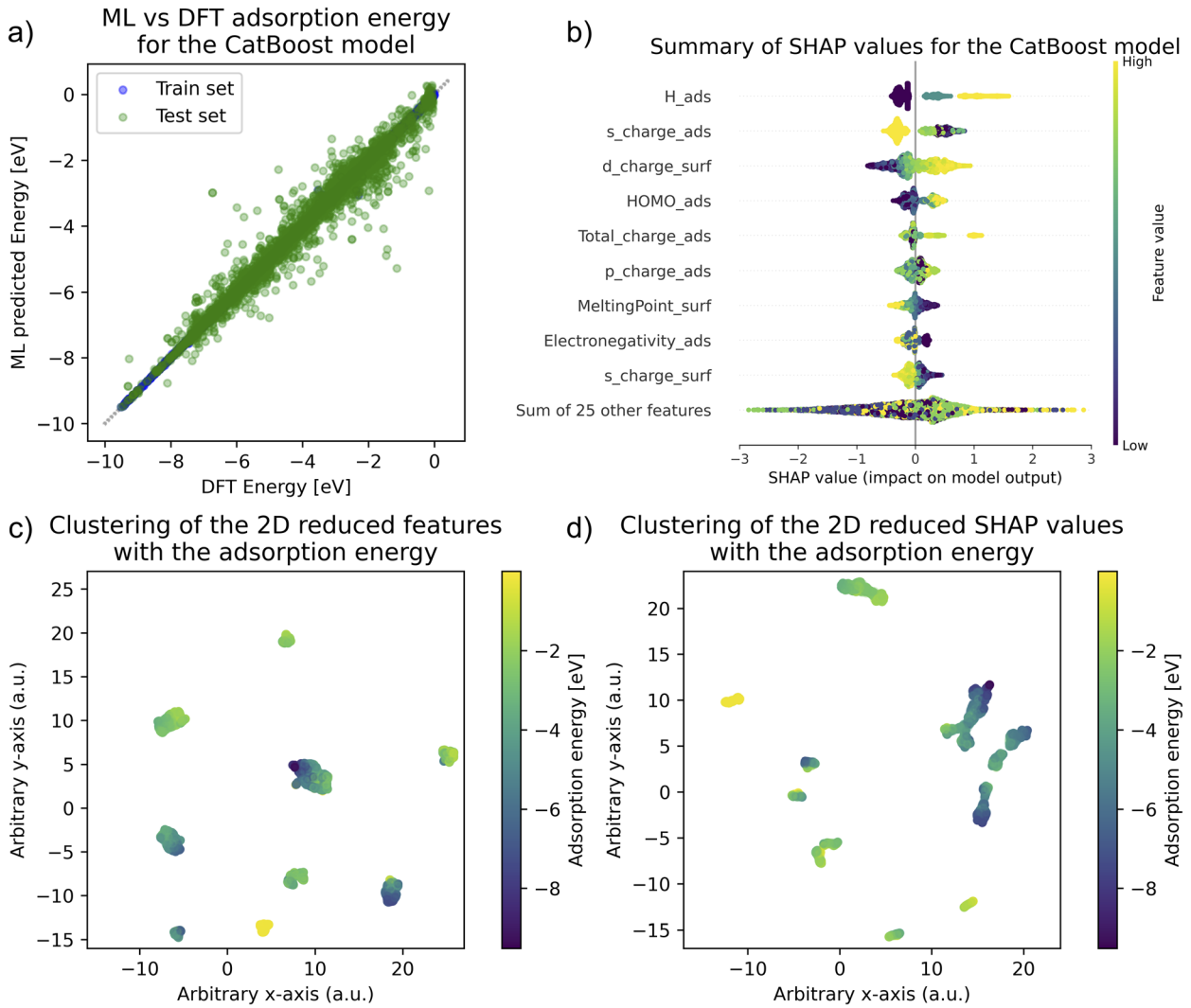


Fig. 6 Performance of ML-based model for the CatBoost methodology extracting top sites from the database. (a) Parity plot of DFT-calculated *vs.* ML-predicted adsorption energies and (b) summary of the distribution of SHAP values for each descriptor. Dimensionality reduction using the UMAP methodology for (c) applied to all raw normalized data (involving the proposed descriptors) and (d) the SHAP values specifically for the CatBoost model, with the 2D reduced dimension alongside adsorption energy employing the training set.

Table 4 Summary of the performance of the evaluated ML-based architectures, including the metric of the average errors and R^2 score for both the training and testing datasets: Mean Absolute Error (MAE), Mean Squared Error (MSE), and R^2 score

ML-based models	MAE of train [eV]	MSE of train [eV]	R^2 score of train	MAE of test [eV]	MSE of test [eV]	R^2 score of test
CatBoost	0.019	0.001	1.000	0.174	0.107	0.973
XGBoost	0.018	0.001	1.000	0.184	0.117	0.970
LightGBM	0.018	0.001	1.000	0.176	0.111	0.972
Kernel ridge	0.061	0.011	0.997	0.212	0.144	0.963
Linear regression	0.561	0.568	0.856	0.559	0.563	0.856

observe a better separation between clusters using the raw normalized data (Fig. 6c), and, therefore, the clustering based on SHAP values (Fig. 6d). This suggests that achieving correlations for the top sites in our dataset can be challenging and may not readily lead to an adequate fit. In addition, removing the top sites from our dataset also decreases the interference among features, helping to mitigate over-parametrization.

By mitigating over-parameterization among features in the predictive model, we can readily assess the impact of the proposed features with each cluster. To identify each cluster, we initially assigned labels to them with the DBSCAN method, as depicted in Fig. S7.† The primary factor influencing the global structure in the clustering is the type of absorbate, in our model, the HOMO energy value of the adsorbate serves to identify each adsorbate, as illustrated in Fig. 7a and shown in Table S3.† Additionally, the global structure is influenced by the number of hydrogen bonds in the adsorbate ("H_ads"), as depicted in Fig. 7b. The adsorbates that do not have hydrogen, such as C, N,

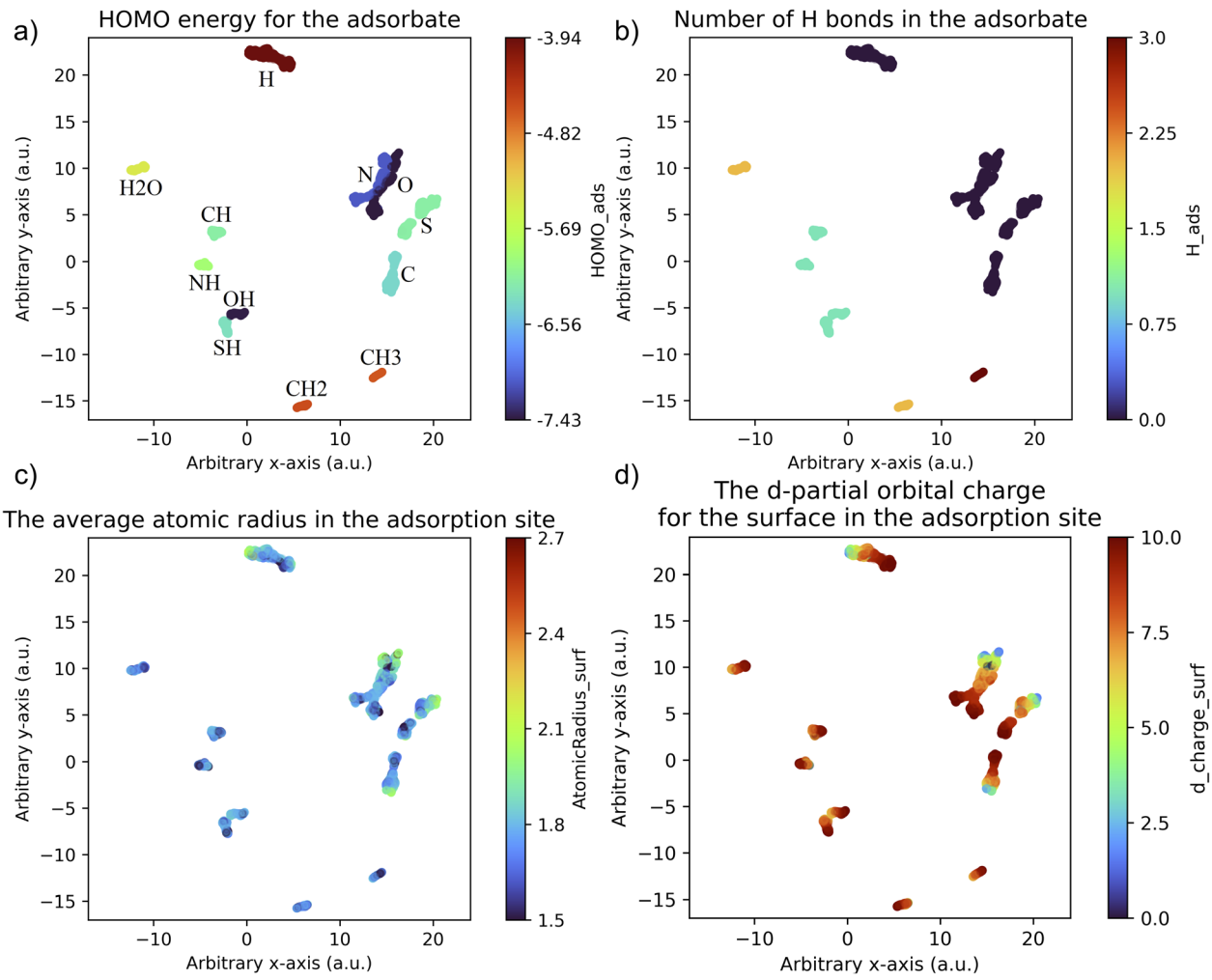


Fig. 7 2D reduced dimension visualization of SHAP values using the CatBoost model for (a) the HOMO values for the adsorbate, (b) the number of hydrogen bonds in the adsorbate, (c) atomic radius in the adsorption site for the surface, and (d) d-partial orbital charge.

S, and O, are close to each other, having global similarity, implying a similar adsorption behavior. A similar trend is found for adsorbates containing a main element and a single H atom bonded to them, *i.e.*, CH, NH, SH, and OH. However, when more than two hydrogen atoms are present on the adsorbate ("H_ads" exceeds 2), the similarity between adsorbates sharing this attribute is lost.

Rationalizing the adsorption energy with the clustering distribution, the electronegativity for the adsorbate helps to explain trends within the bonding strength, as illustrated in Fig. S8.† The smallest bonding interaction corresponds to hydrogen, the element with the lowest electronegativity, while higher adsorption energies are observed for C, N, S, and O. Other observations from "H_ads" are extracted: as expected, when the number of hydrogen atoms increases for C-based adsorbates, the bonding strength decreases, *i.e.*, C is the adsorbate with the higher adsorption energy while the CH_3 presents the lowest values. Furthermore, the bonding interaction of C-based adsorbates depends on the p-partial orbital charge of their C atom (denoted as "p_charge_ads"), when "p_charge_ads" increases the bonding interaction is higher, as observed in Fig. S9.† One specific case worth noting is the water molecule (H_2O), which generally exhibits the smallest values for the bonding interaction. This outcome aligns with expectations, given its stability. Regarding adsorbates with a single hydrogen bond, it becomes evident that SH and OH share a more significant global similarity than CH and NH, which aligns with the p-partial orbital charge of the adsorbate.

Explaining internal trends based on surface-related descriptors can be complex due to the vast amount of information involved. However, specific surface-related trends also arise when examining clusters with adsorbates that do not contain hydrogen. While Fig. 7a for the clusters for C, N, S, and O suggests that the clustering only depends on the kind of adsorbates, these adsorbates are divided into 10 different clusters *via* the DBSCAN method (Fig. S7†). In Fig. S7,† the breaking of the global similarity for C, N, S, and O is explained by the behavior of the d-partial orbital charge ("d_charge_surf") observed in Fig. 7d.

Analyzing the two clusters associated with S in Fig. 7a, each one depends on the "d_charge_surf" values. Furthermore, the d-partial orbital charge is directly correlated with the average atomic radius from the surface atoms at the adsorption site, as depicted in Fig. 7c. The lowest values for the "d_charge_surf" are mainly observed when the average atomic radius in the adsorption site is about 2 Å.

In comparison, the highest values are associated with an average atomic radius below 1.8 Å. Comparing Fig. 7d and 6d shows that the adsorption energy slightly increases when the "d_charge_surf" decreases. Simultaneously, this indicates that the bonding interaction strengthens when the surrounding local environment predominantly comprises atoms with a higher average atomic radius. The clusters associated with C present the same trend as S. Finally, both O and N show relatively close global and local similarities. However, unlike S or C, they display three distinct ranges for the d-partial orbital charge, while there is a correlation between the adsorption energy, atomic radius of the adsorption sites, and the d-partial orbital charge.

4 Conclusions

In this contribution, we propose a methodology that addresses two limitations in machine learning-based models for predicting adsorption energies. First, it addresses the challenge of combining various atom-based adsorbates, and second, it considers the impact of the adsorption site on the chemical binding strength. In this approach, we employ several descriptors, based on the surrounding localized chemical environment of the active site. Within this structure, the ML-based architecture, *via* the CatBoost model, shows the highest performance toward predicting the adsorption energy. The electronic-related features are most impactful in predicting the binding interaction on surface-adsorbate systems, followed by the geometrical features. In contrast, the atomic-related descriptors have the lowest influence on the performance of the model. The CatBoost model has a MAE of 0.166 eV for the training set and 0.280 eV for the testing set.

The anomaly detection technique based on supervised clustering allows us to get trends about the most biased predicted energies. Our model's higher biased points are related to the top adsorption sites. Simultaneously, the top sites are mainly correlated with the range of the weaker bonding interaction, where the correlation may be improved by increasing the sampling at this range. Upon extracting the top adsorption sites, the CatBoost model considerably increases its performance with its MAE, reaching 0.019 eV and 0.174 eV for the training set and test set, respectively. However, it is worth noting that the atomic-elemental-related descriptors play a crucial role in describing physicochemical trends through the supervised clustering methodology. The global and local similarity in the clustering allows us to understand how factors such as the atom-based adsorbates and the alloy's electronic structure correlate with the adsorbate-surface bonding interaction. Our approach underscores features based on chemical intuition in describing the bonding interaction, such as the HOMO energy values for the adsorbate and the d-partial orbital occupancy of the metal surface atoms, although they need to be complemented with several additional features to obtain an accurate prediction of the adsorption energy. Besides, the information extracted *via* supervised learning not only works as an anomaly detection technique, enhancing the mathematical performance of the high-dimensional model but also yields insights with physical coherence and valuable meaning toward the chemical bonding. Further modifications of the approach are envisaged to predict the adsorption energy for more complex adsorbates, such as bidentate binding modes or larger adsorbates, and aspects that will be subject to future work.

Data availability

The ML-based models are available in the GitHub repository: **https://github.com/Anfeus02/Localized-chemical-E_ads-ML.**

Author contributions

CRediT (Contributor Roles Taxonomy) was used for standardized contribution descriptions: A. F. Usuga: data curation, formal analysis, investigation, methodology, software, validation, visualization, writing (original draft), writing (review & editing). C. S. Praveen: data curation, conceptualization, supervision, writing (review & editing). A. Comas-Vives: data curation, conceptualization, project administration, resources, funding acquisition, methodology, supervision, validation, writing (review & editing).

Conflicts of interest

The authors declare that the research was conducted without commercial or financial relationships that could be constructed as a potential conflict of interest.

Acknowledgements

DFT calculations were performed in the HPC "Consorci de Serveis Universitaris de Catalunya (CSUC)". The authors thank the Spanish "Ministerio de Ciencia e Innovación" for funding the "I + D Generación del Conocimiento" project (PID2021-128416NB-I00 and PGC2018-100818-A-I00) and the predoctoral grant (PRE2019-089605). A part of the work has been performed under Project HPC-EUROPA3 (INFRAIA-2016-1-730897), with the support of the EC Research Innovation Action under the H2020 Programme awarded to CSP; in particular, CSP gratefully acknowledges the support of Dr Xavier Solans-Monfort of Computational BioNanoCat in the Department of Chemistry at the Universitat Autònoma de Barcelona (UAB), with HPC resources and support provided by BSC. CSP also acknowledges Cochin University of Science and Technology for the SMNRI project grant and the computing resources provided by Param Sanganak under NSM. CSP also acknowledges DST India for the INSPIRE Faculty Fellowship (IFA18-PH217).

References

1 F. Tao, Synthesis, catalysis, surface chemistry and structure of bimetallic nanocatalysts, *Chem. Soc. Rev.*, 2012, **41**, 7977–7979.
2 J. A. Rodriguez and D. W. Goodman, Surface Science Studies of the Electronic and Chemical Properties of Bimetallic Systems, *J. Phys. Chem.*, 1991, **95**, 4196–4206.
3 M. Sankar, *et al.*, Designing bimetallic catalysts for a green and sustainable future, *Chem. Soc. Rev.*, 2012, **41**, 8099–8139.
4 X. Liu, D. Wang and Y. Li, Synthesis and catalytic properties of bimetallic nanomaterials with various architectures, *Nano Today*, 2012, **7**, 448–466, DOI: 10.1016/j.nantod.2012.08.003.
5 A. Jain, Y. Shin and K. A. Persson, Computational predictions of energy materials using density functional theory, *Nat. Rev. Mater.*, 2016, **1**, 15004.
6 F. Abild-Pedersen, *et al.*, Scaling properties of adsorption energies for hydrogen-containing molecules on transition-metal surfaces, *Phys. Rev. Lett.*, 2007, **99**, 016105.
7 J. Greeley, Theoretical Heterogeneous Catalysis: Scaling Relationships and Computational Catalyst Design, *Annu. Rev. Chem. Biomol. Eng.*, 2016, **7**, 605–635.
8 O. Mamun, K. T. Winther, J. R. Boes and T. Bligaard, A Bayesian framework for adsorption energy prediction on bimetallic alloy catalysts, *npj Comput. Mater.*, 2020, **6**, 177.
9 J. Pérez-Ramírez and N. López, Strategies to break linear scaling relationships, *Nat. Catal.*, 2019, **2**, 971–976, DOI: 10.1038/s41929-019-0376-6.
10 Z. W. Ulissi, A. J. Medford, T. Bligaard and J. K. Nørskov, To address surface reaction network complexity using scaling relations machine learning and DFT calculations, *Nat. Commun.*, 2017, **8**, 14621.
11 B. Hammer and J. Norskov, Why gold is the noblest of all the metals, *Nature*, 1995, **376**, 238–240.
12 B. Hammer and J. B. K. Norskov, Electronic factors determining the reactivity of metal surfaces, *Surf. Sci.*, 1995, **343**, 211–220.
13 J. H. Sinfelt, Catalysis by Alloys and Bimetallic Clusters, *Acc. Chem. Res.*, 1977, **10**, 15–20.
14 W. Yang, T. T. Fidelis and W. H. Sun, Machine Learning in Catalysis, from Proposal to Practicing, *ACS Omega*, 2020, **5**, 83–88, DOI: 10.1021/acsomega.9b03673.
15 R. Ramakrishnan, P. O. Dral, M. Rupp and O. A. Von Lilienfeld, Big data meets quantum chemistry approximations: The Δ-machine learning approach, *J. Chem. Theory Comput.*, 2015, **11**, 2087–2096.
16 J. A. Keith, *et al.*, Combining Machine Learning and Computational Chemistry for Predictive Insights into Chemical Systems, *Chem. Rev.*, 2021, **121**, 9816–9872, DOI: 10.1021/acs.chemrev.1c00107.
17 Z. Zhou, X. Li and R. N. Zare, Optimizing Chemical Reactions with Deep Reinforcement Learning, *ACS Cent. Sci.*, 2017, **3**, 1337–1344.
18 T. Villadsen, N. F. W. Ligterink and M. Andersen, Predicting binding energies of astrochemically relevant molecules via machine learning, *Astron. Astrophys.*, 2022, **666**, A45.
19 M. O. J. Jäger, E. V. Morooka, F. Federici Canova, L. Himanen and A. S. Foster, Machine learning hydrogen adsorption on nanoclusters through structural descriptors, *npj Comput. Mater.*, 2018, **4**, 37.
20 R. Jinnouchi and R. Asahi, Predicting Catalytic Activity of Nanoparticles by a DFT-Aided Machine-Learning Algorithm, *J. Phys. Chem. Lett.*, 2017, **8**, 4279–4283.
21 P. S. Lamoureux, T. S. Choksi, V. Streibel and F. Abild-Pedersen, Combining artificial intelligence and physics-based modeling to directly assess atomic site stabilities: From sub-nanometer clusters to extended surfaces, *Phys. Chem. Chem. Phys.*, 2021, **23**, 22022–22034.
22 R. Chen, *et al.*, Combined first-principles and machine learning study of the initial growth of carbon nanomaterials on metal surfaces, *Appl. Surf. Sci.*, 2022, **586**, 152762.
23 Y. Chen, Y. Zhao, P. Ou and J. Song, Basal plane activation of two-dimensional transition metal dichalcogenides via alloying for the hydrogen evolution reaction: first-

principles calculations and machine learning prediction, *J. Mater. Chem. A*, 2023, **11**, 9964–9975.
24 S. Thomas, F. Mayr, A. Kulangara Madam and A. Gagliardi, Machine learning and DFT investigation of CO, CO2 and CH4 adsorption on pristine and defective two-dimensional magnesene, *Phys. Chem. Chem. Phys.*, 2023, **25**, 13170–13182.
25 C. Gao, *et al.*, Machine learning-enabled band gap prediction of monolayer transition metal chalcogenide alloys, *Phys. Chem. Chem. Phys.*, 2022, **24**, 4653–4665.
26 K. Rossi, *et al.*, Quantitative Description of Metal Center Organization and Interactions in Single-Atom Catalysts, *Adv. Mater.*, 2023, 2307991.
27 D. Wang, *et al.*, Accelerated prediction of Cu-based single-atom alloy catalysts for CO2 reduction by machine learning, *Green Energy Environ.*, 2021, **8**, 820–830.
28 Z. Lu, S. Yadav and C. V. Singh, Predicting aggregation energy for single atom bimetallic catalysts on clean and O* adsorbed surfaces through machine learning models, *Catal. Sci. Technol.*, 2020, **10**, 86–98.
29 Z. K. Han, *et al.*, Single-atom alloy catalysts designed by first-principles calculations and artificial intelligence, *Nat. Commun.*, 2021, **12**, 1833.
30 Z. H. Liu, T. T. Shi and Z. X. Chen, Machine learning prediction of monatomic adsorption energies with non-first-principles calculated quantities, *Chem. Phys. Lett.*, 2020, **755**, 137772.
31 J. Feng, Y. Ji and Y. Li, silico design of copper-based alloys for ammonia synthesis from nitric oxide reduction accelerated by machine learning, *J. Mater. Chem. A*, 2023, **11**, 14195–14203.
32 Z. Gariepy, *et al.*, Machine learning assisted binary alloy catalyst design for the electroreduction of CO2 to C2 products, *Energy Adv.*, 2023, **2**, 410–419.
33 J. Geiger, A. Sabadell-Rendón, N. Daelman and N. López, Data-driven models for ground and excited states for Single Atoms on Ceria, *npj Comput. Mater.*, 2022, **8**, 171.
34 X. Li, *et al.*, A transferable machine-learning scheme from pure metals to alloys for predicting adsorption energies, *J. Mater. Chem. A*, 2022, **10**, 872–880.
35 Z. Li, S. Wang, W. S. Chin, L. E. Achenie and H. Xin, High-throughput screening of bimetallic catalysts enabled by machine learning, *J. Mater. Chem. A*, 2017, **5**, 24131–24138.
36 X. Wan, *et al.*, Machine-learning-assisted discovery of highly efficient high-entropy alloy catalysts for the oxygen reduction reaction, *Patterns*, 2022, **3**, 9.
37 Z. Yang, W. Gao and Q. Jiang, A machine learning scheme for the catalytic activity of alloys with intrinsic descriptors, *J. Mater. Chem. A*, 2020, **8**, 17507–17515.
38 J. Noh, S. Back, J. Kim and Y. Jung, Active learning with non-ab initio input features toward efficient CO2 reduction catalysts, *Chem. Sci.*, 2018, **9**, 5152–5159.
39 X. Ma, Z. Li, L. E. K. Achenie and H. Xin, Machine-Learning-Augmented Chemisorption Model for CO2 Electroreduction Catalyst Screening, *J. Phys. Chem. Lett.*, 2015, **6**, 3528–3533.
40 R. García-Muelas and N. López, Statistical learning goes beyond the d-band model providing the thermochemistry of adsorbates on transition metals, *Nat. Commun.*, 2019, **10**, 4687.
41 M. Rittiruam, *et al.*, First-Principles Density Functional Theory and Machine Learning Technique for the Prediction of Water Adsorption Site on PtPd-Based High-Entropy-Alloy Catalysts, *Adv. Theory Simul.*, 2023, **6**, 2200926.
42 Y. Wang, *et al.*, High-throughput calculations combining machine learning to investigate the corrosion properties of binary Mg alloys, *J. Magnesium Alloys*, 2022, DOI: **10.1016/j.jma.2021.12.007**.
43 S. Saxena, T. S. Khan, F. Jalid, M. Ramteke and M. A. Haider, In silico high throughput screening of bimetallic and single atom alloys using machine learning and ab initio microkinetic modelling, *J. Mater. Chem. A*, 2020, **8**, 107–123.
44 Z. Yang and W. Gao, Applications of Machine Learning in Alloy Catalysts: Rational Selection and Future Development of Descriptors, *Advanced Science*, 2022, **9**, 2106043.
45 T. Mou, *et al.*, Bridging the complexity gap in computational heterogeneous catalysis with machine learning, *Nat. Catal.*, 2023, **6**, 122–136.
46 Z. W. Ulissi, A. R. Singh, C. Tsai and J. K. Nørskov, Automated Discovery and Construction of Surface Phase Diagrams Using Machine Learning, *J. Phys. Chem. Lett.*, 2016, **7**, 3931–3935.
47 J. Zhang, *et al.*, Accurate and efficient machine learning models for predicting hydrogen evolution reaction catalysts based on structural and electronic feature engineering in alloys, *Nanoscale*, 2023, **15**, 11072–11082.
48 K. Tran and Z. W. Ulissi, Active learning across intermetallics to guide discovery of electrocatalysts for CO2 reduction and H2 evolution, *Nat. Catal.*, 2018, **1**, 696–703.
49 F. Liu, P. F. Gao, C. Wu, S. Yang and X. Ding, DFT-based Machine Learning for Ensemble Effect of Pd@Au Electrocatalysts on CO2 Reduction Reaction, *ChemPhysChem*, 2023, **24**, e202200642.
50 J. Lan, *et al.*, AdsorbML: a leap in efficiency for adsorption energy calculations using generalizable machine learning potentials, *npj Comput. Mater.*, 2023, **9**, 172.
51 Z. W. Ulissi, *et al.*, Machine-learning methods enable exhaustive searches for active Bimetallic facets and reveal active site motifs for CO2 reduction, *ACS Catal.*, 2017, **7**, 6600–6608.
52 S. Back, K. Tran and Z. W. Ulissi, Toward a Design of Active Oxygen Evolution Catalysts: Insights from Automated Density Functional Theory Calculations and Machine Learning, *ACS Catal.*, 2019, **9**, 7651–7659.
53 P. G. Ghanekar, S. Deshpande and J. Greeley, Adsorbate chemical environment-based machine learning framework for heterogeneous catalysis, *Nat. Commun.*, 2022, **13**, 5788.
54 P. Reiser, *et al.*, Graph neural networks for materials science and chemistry, *Commun. Mater.*, 2022, **3**, 93.
55 W. Xu, K. Reuter and M. Andersen, Predicting binding motifs of complex adsorbates using machine learning with a physics-inspired graph representation, *Nat. Comput. Sci.*, 2022, **2**, 443–450.

56 S. Pablo-García, *et al.*, Fast evaluation of the adsorption energy of organic molecules on metals via graph neural networks, *Nat. Comput. Sci.*, 2023, **3**, 433–442.
57 W. A. Saidi, Emergence of local scaling relations in adsorption energies on high-entropy alloys, *npj Comput. Mater.*, 2022, **8**, 86.
58 M. Andersen and K. Reuter, Adsorption Enthalpies for Catalysis Modeling through Machine-Learned Descriptors, *Acc. Chem. Res.*, 2021, **54**, 2741–2749.
59 M. Andersen, S. V. Levchenko, M. Scheffler and K. Reuter, Beyond Scaling Relations for the Description of Catalytic Materials, *ACS Catal.*, 2019, **9**, 2752–2759.
60 O. Mamun, K. T. Winther, J. R. Boes and T. Bligaard, High-throughput calculations of catalytic properties of bimetallic alloy surfaces, *Sci. Data*, 2019, **6**, 76.
61 P. Giannozzi, *et al.*, QUANTUM ESPRESSO: A modular and open-source software project for quantum simulations of materials, *J. Phys.: Condens.Matter*, 2009, **21**, 395502.
62 K. F. Garrity, J. W. Bennett, K. M. Rabe and D. Vanderbilt, Pseudopotentials for high-throughput DFT calculations, *Comput. Mater. Sci.*, 2014, **81**, 446–452.
63 C. S. Praveen and A. Comas-Vives, 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.
64 T. Chen and C. Guestrin, XGBoost, in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, 2016, pp. 785–794, DOI: **10.1145/2939672.2939785**.
65 A. V. Dorogush, V. Ershov and A. Gulin, CatBoost: Gradient Boosting with Categorical Features Support, *arXiv*, 2018, preprint, arXiv: 1810.11363, DOI: **10.48550/arXiv.1810.11363**.
66 G. Ke, *et al.*, LightGBM: A Highly Efficient Gradient Boosting Decision Tree, in *Advances in Neural Information Processing Systems*, ed. I. Guyon, Curran Associates, Inc., vol. 30, 2017.
67 S. Lundberg and S.-I. Lee, A Unified Approach to Interpreting Model Predictions, *arXiv*, 2017, preprint, arXiv: 1705.07874, DOI: **10.48550/arXiv.1705.07874**.
68 L. McInnes, J. Healy and J. Melville, UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction, *arXiv*, 2018, preprint, arXiv: 1802.03426, DOI: **10.48550/arXiv.1705.07874**.
69 S. Pablo-García, R. García-Muelas, A. Sabadell-Rendón and N. López, Dimensionality reduction of complex reaction networks in heterogeneous catalysis: From linear-scaling relationships to statistical learning techniques, *Wiley Interdiscip. Rev.: Comput. Mol. Sci.*, 2021, **11**, e1540.
70 L. Buitinck, *et al.*, Scikit-learn: Machine Learning in Python, *J. Mach. Learn. Res.*, 2011, **12**, 2825–2830.
71 J. A. Esterhuizen, B. R. Goldsmith and S. Linic, Interpretable machine learning for knowledge generation in heterogeneous catalysis, *Nat. Catal.*, 2022, **5**, 175–184.
72 H. Wang, Y. Ji and Y. Li, Simulation and design of energy materials accelerated by machine learning, *Wiley Interdiscip. Rev.: Comput. Mol. Sci.*, 2020, **10**, e1421.
73 V. Fung, G. Hu, P. Ganesh and B. G. Sumpter, Machine learned features from density of states for accurate adsorption energy prediction, *Nat. Commun.*, 2021, **12**, 88.
74 A. Cooper, O. Doyle & A. Bourke, Supervised Clustering for Subgroup Discovery: An Application to COVID-19 Symptomatology, in, *Joint European conference on machine learning and knowledge discovery in databases*, Springer, 2021, vol. 1525, pp. 408–422, DOI: **10.1007/978-3-030-93733-1_29**.