Efficient mapping of CO adsorption on Cu1xMx bimetallic alloys via machine learning

Journal of
Materials Chemistry A

PAPER

View Article Online
View Journal

Cite this: DOI: 10.1039/d3ta06915j

Efficient mapping of CO adsorption on $Cu_{1-x}M_x$ bimetallic alloys *via* machine learning†

Mattia Salomone, *[a] Michele Re Fiorentin, [a] Francesca Risplendi, [a] Federico Raffone, [a] Timo Sommer, [b] Max García-Melchor *[b] and Giancarlo Cicero*[a]

The electrochemical reduction of CO_2 (CO_2RR) has the potential to allay the greenhouse gas effect while also addressing global energy challenges by producing value-added fuels and chemicals (mostly C_2 molecules such as ethylene and ethanol). However, due to the complicated chemical pathways involved, achieving high selectivity and efficiency towards specific reduction products remains challenging. In fact, the design of more selective and efficient catalysts often relies on trial-and-error approaches, which are very time consuming and resource intensive. In response, driven by the inherent importance of CO adsorption energy in the conversion of CO_2 into C_{2+} hydrocarbons and alcohols, we propose a two-step approach employing machine learning classification and regression algorithms to predict CO binding energies on CuM(111)/(100) (M = Al, Ti, V, Fe, Co, Ni, Zn, Nb, Mo, Ru, Pd, Ag, Cd, Sn, Sb, Hf, W, Ir, Pt, Au) bimetallic surfaces. Firstly, we assess the stability of each adsorption site by utilizing classification algorithms. Subsequently, focusing exclusively on the stable sites, we employ regression models to predict the adsorption energies of CO. Remarkably, by employing a Gradient Boosting Classifier for classification, together with a Gradient Boosting Regressor for regression, we predict CO binding energies with a high level of robustness and accuracy for Cu bimetallic alloys with up to 17% surface impurity concentrations. The accuracy of our models is demonstrated by F1 scores exceeding 96% and a mean square error below 0.05 eV^2 for the classification and regression parts, respectively. These remarkable results highlight the adaptability of our approach and its capability for efficiently screening Cu-based CO_2RR electrocatalysts, enabling rapid evaluation of promising candidates for future in-depth explorations.

Received 10th November 2023
Accepted 3rd May 2024

DOI: 10.1039/d3ta06915j

rsc.li/materials-a

1 Introduction

The development of technologies for the capture and conversion of CO_2 to value-added compounds has been driven by the drastic environmental consequences of anthropogenic climate change.[1] The electrochemical reduction of CO_2 (CO_2RR) to industrially relevant chemicals using reasonably priced renewable energy sources is one of the most promising strategies.[2] Currently, this process has reached a high degree of technical readiness in the production of single-carbon molecules like CO and formic acid.[3,4] However, due to undesirable side reactions and relatively limited selectivity, the CO_2RR to C_{2+} products (such as ethylene and ethanol) with higher market potential faces many challenges.[5] To date, this process has mainly been reported for copper-based catalysts in alkaline media.[6] However, while alkaline conditions promote the synthesis of desirable C_{2+} products, high pH values dramatically reduce the CO_2 utilization efficiency due to hydroxide ion reactivity and the consequent precipitation of (bi)carbonate.[7] Shifting to more acidic conditions would lower (bi)carbonate formation, but also suppress the already low faradaic efficiencies towards C_{2+} products, as it would favour the competing hydrogen evolution reaction.[8] Several studies have demonstrated that the reduction of CO on Cu leads to a product distribution comparable to that of the CO_2RR on Cu, revealing that CO is an important reaction intermediate.[9–14] Consequently, the binding energy of adsorbed CO (*CO) is considered as a key reaction descriptor for the production of C_{2+} chemicals.[15–17] Metals that bind CO too strongly will be poisoned by this intermediate, while metals that bind CO too weakly will release it before it can react further, following the Sabatier principle.[18] The awareness of CO being a key intermediate in the CO_2RR to hydrocarbons and alcohols has led to increasing interest in studying CO electroreduction (CORR) as a proxy for understanding CO_2RR trends. This approach is advantageous because it considers fewer reaction steps/intermediates. In addition, a better understanding of

[a]Dipartimento di Scienza Applicata e Tecnologia, Politecnico di Torino, 10129 Torino, Italy. E-mail: mattia.salomone@gmail.com; giancarlo.cicero@polito.it

[b]CRANN and AMBER Research Centres, School of Chemistry, Trinity College Dublin, College Green, Dublin 2, Ireland. E-mail: garciamm@tcd.ie

† Electronic supplementary information (ESI) available. See DOI: https://doi.org/10.1039/d3ta06915j

CORR catalysis is important because it could provide an alternate route for the CO_2RR that can potentially overcome selectivity problems if the process is split into two separate steps: (1) CO_2 reduction to CO and (2) CO reduction to the desired products (tandem approach).[19,20] To avoid (bi)carbonate precipitation and efficiency losses, CO_2 is first reduced to CO under mildly acidic conditions, followed by the reduction of CO to the required C_{2+} product(s) in alkaline media. However, despite this tandem process being very promising, a deeper understanding of the second step is required since many C_{2+} reaction intermediates exhibit similar binding energies, compromising CORR selectivity. Altering the local environment surrounding the active sites can improve the catalytic activity, selectivity, and stability of electrocatalysts. For example, alloying,[21,22] surface doping,[23] ligand modification,[24] and interface engineering[25] have been adopted as catalyst engineering strategies. Among these methods, copper alloying has proven particularly useful in improving catalytic performance while retaining long-term stability.

In this context, density functional theory (DFT) is a powerful and predictive tool for guiding catalyst design and synthesis.[26–28] However, DFT simulations require a significant amount of time to be performed and considering that the designing of novel catalysts generally depends on a trial-and-error approach,[29] this aspect represents a considerable bottleneck when investigating a large number of metal alloys with different concentrations and types of impurities.[30–33] Recent advancements in the integration of machine learning (ML) with *ab initio*-obtained data have provided an opportunity to accelerate the high-throughput screening of electrocatalysts. In supervised learning, the algorithm learns a predictive function from a dataset of labeled data, enabling it to make predictions on new, unlabeled data – the term label denotes the target property of interest, *e.g.* the binding energy of a reaction intermediate. The goal is to identify complex patterns and correlations within the known data, which can then be extrapolated to accurately forecast unexplored data. Once the training set is sufficiently large and representative of the catalytic property of interest, the trained model can be used to predict this quantity also for systems not considered during the training process,[29,34–37] without the need for further DFT simulations. This approach has seen applications in various chemical fields,[38–40] including catalysis.[15,41,42] Different models exist depending on the structure of the learned function and the associated learning algorithm, offering a range of possible approaches to address different learning tasks. ML models can be used both for classification and regression tasks. In classification, the learner is required to map the input space into predefined classes, *i.e.* the label (target property) is discrete. In contrast, within regression, the model maps the input space into a real-value domain, thus leading to a continuous label.[43,44]

In this work we present an ML-based model to predict CO binding energies on various Cu bimetallic surfaces $Cu_{1-x}M_x$(111)/(100), with $0.028 \leq x \leq 0.168$ (*ca.* 3–17%) and M = Al, Ti, V, Fe, Co, Ni, Zn, Nb, Mo, Ru, Pd, Ag, Cd, Sn, Sb, Hf, W, Ir, Pt, Au. Our investigations focus on metallic Cu surfaces as host matrices, owing to the favorable reduction of copper oxides to metallic Cu under applied potentials and pH values pertinent to the CO_2RR.[45–48] The novelty of our approach lies in dividing the surface analysis into two separate steps. First, with classification algorithms we evaluate the stability of a given binding site, then, focusing solely on the stable sites, we employ regression models to predict CO adsorption energies. To obtain realistic generalization errors,[49] we compare the performance of multiple ML algorithms, developed using readily available chemical and geometrical properties as features. In most cases, the ML models show excellent performances for both classification and regression tasks. In particular, we find the Gradient Boosting[50,51] Classifier to be the best model to predict the stability of CO binding sites, while the Gradient Boosting Regressor performs best in the prediction of CO adsorption energies. Adopting the same two-step approach, we employ these algorithms to predict CO binding energies on $Cu_{1-x}Ag_x$ and $Cu_{1-x}Au_x$ bimetallic alloys with relatively high surface impurity concentrations ($x \leq 0.168$). Despite training our algorithms at low concentrations, the overall performance of the ML models remains comparable to those obtained with the original test set, highlighting the versatility and transferability of our ML models. According to our findings, our two-step ML approach can accurately capture the CO behavior on a wide range of bimetallic alloys, considering the different impurity concentrations as well as different surface facets of various guest species. Hence, this method is envisioned to enable the high throughput screening of CORR electrocatalysts for the sustainable and selective production of chemicals and fuels.

2 Methods

2.1 DFT computational approach

ML models were trained on DFT-calculated CO adsorption energies (ΔE_{CO}) on Cu-based bimetallic alloys, namely $Cu_{0.972}M_{0.028}$(111)/(100) surfaces. The ΔE_{CO} values (considering C as the atom binding to the surface) were calculated with the Quantum Espresso package[52,53] using ultrasoft pseudopotentials[54] and the Perdew–Burke–Ernzerhof (PBE) exchange–correlation functional.[55,56] Although this exchange–correlation functional tends to overestimate the binding energies of adsorbates on metal surfaces, it consistently captures the adsorption trends of molecules. This ensures a faithful depiction of the overall behavior of the CO molecule. Valence electrons were described using plane waves with a cutoff energy of 40 Ry, while a cutoff of 400 Ry was adopted for the charge density and pseudopotentials. Four-layered Cu(111)/(100) surface slabs were constructed with a vacuum layer of 12 Å. Structures were subsequently relaxed by minimizing the atomic forces with a convergence threshold of 10^{-5} Ry per Bohr and sampling the Brillouin zone using a $10 \times 10 \times 1$ Monkhorst–Pack (MP) *k*-point grid.[57] Using these models, we built $Cu_{0.972}M_{0.028}$(111)/(100) surface slabs with a 6×6 periodicity by substitutional alloying 1 surface Cu atom with 20 different M species (*i.e.* Al, Ti, V, Fe, Co, Ni, Zn, Nb, Mo, Ru, Pd, Ag, Cd, Sn, Sb, Hf, W, Ir, Pt and Au). For these surfaces, the *k*-point grid was reduced to $2 \times 2 \times 1$ to keep the same *k*-point density. In all the calculations, the bottom two layers were fixed, while the topmost layers were allowed to relax.

2.2 Structure of the dataset: labels and features

To generate the dataset, we adsorbed a CO molecule onto all the unique binding sites within the first and second nearest neighbor positions adjacent to the guest species M in the $Cu_{1-x}M_x$(111)/(100) surfaces, shown in Fig. 1c and d. Subsequently, for each binding site we calculated the corresponding CO adsorption energy, defined as

$$\Delta E_{CO} = E_{slab+CO} - E_{slab} - E_{CO},$$

where $E_{slab+CO}$, E_{slab} and E_{CO} are the energies of the Cu slab with the adsorbed CO, the pristine surface, and an isolated CO molecule, respectively. In the case of the pristine surfaces, the possible binding sites are shown in Fig. 1a and b. In particular, for the Cu(111) surface, the CO molecule can adsorb on top (**T**) and bridge (**B**) sites, as well as two different hollow sites, namely fcc (**H1**) and hcp (**H2**) (see Fig. 1a). However, the **B** site was found to be unstable, and the CO molecule moved into an **H1** site, which is the most stable adsorption on Cu(111).[58] This behavior was also observed on all (111) surfaces, and therefore the **B** sites were not included in the (111) surface dataset. Similarly, for the Cu(100) surface, CO can bind on **T** and **B** sites, as well as a hollow site (**H**), as depicted in Fig. 1b. Upon introduction of the guest M species, the number of unique adsorption sites significantly increases, giving rise to the two patterns shown in Fig. 1c and d for Cu(111) and Cu(100), respectively. We note that the presence of the guest atom in Cu(100) gives rise to two distinct bridge sites per atom (**B1** and **B2** in light/dark blue) because of their varying distances from the impurity. The rest of the sites for both (111) and (100) surfaces remain either unchanged (**T** in yellow and **H** in light/dark red) or are equivalent (white dots in Fig. 1c and d) due to surface symmetry. Overall, this leads to 25 and 21 unique sites for the $Cu_{0.972}M_{0.028}$(111)/(100) surfaces, respectively. These, combined with the 20 guest atoms considered in this work, result in 920 entries in our dataset ($20 \times 21 + 20 \times 25 = 920$).

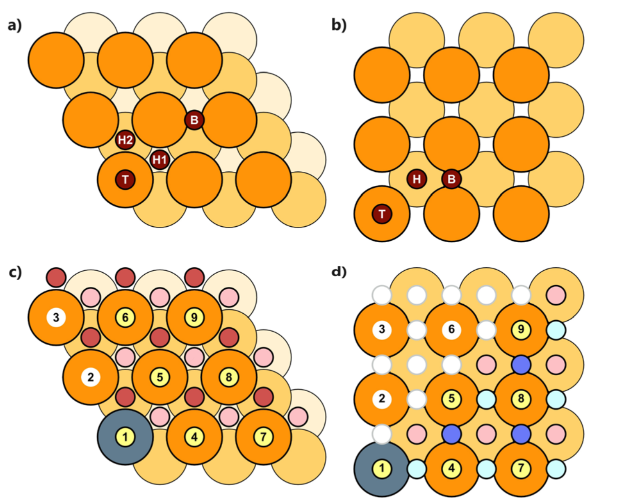

Fig. 1 Labeling of adsorption sites for the pristine (a) Cu(111) and (b) Cu(100) surfaces. Binding sites for the defective (c) $Cu_{0.972}M_{0.028}$(111) and (d) $Cu_{0.972}M_{0.028}$(100) surfaces. Atom numbers from 1 (impurity) to 9 are used to distinguish atoms. Atoms 2, 4, and 5 are nearest neighbors to guest species, while 3, 6, 7, 8, and 9 are next-nearest neighbors. The presence of the impurity introduces new non-equivalent adsorption sites, namely 3 per atom in Cu(111) and 4 in Cu(100). Cu(111): light red and red dots represent hollow fcc (H1) and hcp (H2) binding sites. Cu(100): dots in light blue/blue and light red denote bridge (B1/B2, resulting from their different distance from the guest atom) and hollow sites, respectively. Yellow dots represent top sites and white dots symmetrically equivalent sites not included in the dataset.

Next, each adsorption energy was represented by a vector of 15 features, of which 13 are related to the chemical properties of the M atom,[59,60] and 2 are associated with the geometrical aspects of the CO binding site. To create a highly interpretable, general, accessible and computationally efficient ML model, the 13 chemical features were chosen to be properties readily available in the literature which do not require any further DFT simulations. These features include the atomic number (AtNumb), periodic table group and period (AtGroup and AtPer), atomic radius (AtRad), atomic mass (AtMass), electron affinity (ElecAff), Pauli electronegativity (PaElec), and ionization energy (IonEn), density (Density), melting point (MeltPoint), boiling point (BoilPoint), surface energy (SurfEn), and work function (WorkFun) of the M species. The last two features refer to the hexagonal surface of the guest atom in its crystalline phase. The remaining 2 geometrical properties are the distance between the CO adsorption site and the substitutional impurity, and the generalized coordination number (GCN)[61] of the adsorption site proposed by Calle-Vallejo *et al.* To make the model more general, these features were calculated on the ideal surface geometries of the pristine Cu(111)/(100) surfaces. This assumes that the distance between a binding site and the M atom is not influenced by the impurity type. In addition, when calculating the GCN for a specific adsorption site, the local strain was not included. Finally, we note that Cu features were not included in the models as they would be constant for all the elements in the dataset.

2.3 Implementation of the ML models

The CO adsorption energies on the given $Cu_{1-x}M_x$(111)/(100) surfaces were obtained in two steps: (1) determining if a binding site is stable or unstable, and (2) calculating the CO adsorption energy on stable sites. A binding site was deemed stable if the CO molecule remains adsorbed on the surface and does not migrate to a different site during the DFT relaxation, otherwise the site was labelled as unstable. To address this issue, we sampled a wide range of ML algorithms belonging to several different categories, namely linear and distance-based models, support vector machines, kernel methods, decision tree, tree ensemble methods and feedforward NNs. In particular, we trained eight different supervised models (most of them used both for classification and regression), namely Gradient Boosting[50,51] Classifier and Regressor (GBC and GBR), Support Vector[62] Classifier and Regressor (SVC and SVR, with

Table 1 Hyperparameters used for the classification (upper table) and regression (lower table) algorithms, optimized considering a 4-fold cross validation

Classification algorithm	Hyperparameters
Gradient boosting	n_estimators = 300, max_depth = 5, learning_rate = 0.7
Support vector	kernel = linear, class_weight = {0 : 1.1, 1 : 0.4}, kernel = poly, degree = 25, coef0 = 5.0
Random forest	n_estimators = 100, max_depth = 10
Decision tree	max_depth = 10
Neural networks	number of layers = 4, neurons = 50/50/50/1, activation = sigmoid, loss = binary_crossentropy, optimizer = Adam
Logistic regression	solver = liblinear, C = 2, class_weight = {0 : 1.1, 1 : 0.4}
K-nearest neighbors	n_neighbors = 2, algorithm = auto, p = 2
Regression algorithm	**Hyperparameters**
Gradient boosting	n_estimators = 314, max_depth = 4, learning_rate = 0.15
Support vector	kernel = rbf, C = 4, gamma = 0.06, epsilon = 0.001
Random forest	n_estimators = 6
Decision tree	max_depth = 67
Neural networks	number of layers = 4, neurons = 80/80/80/1, activation = relu, loss = mae, optimizer = rmsprop
Gaussian process	kernel = RBF()·RQ (length_scale = 1), alpha = 1, n_restarts_optimizer = 15, length_scale_bounds = $(1 \times 10^{-5}, 100)$

linear, *polynomial* and *rbf* kernels), Random Forest[63] Classifier and Regressor (RFC and RFR), Decision Tree[64] Classifier and Regressor (DTC and DTR), Neural Networks[42] (NNs) for both tasks, Logistic Regression[65] (LR) for classification, K-Nearest Neighbor Classifier[66] (KNN) for classification and Gaussian Process Regressor[67] (GPR) for regression. With the exception of NNs, which were implemented using Keras[68] with a TensorFlow[69] back-end, all ML algorithms were developed using the open-source library Scikit-Learn.[70] In training our models and optimizing hyperparameters, we employed the function RandomizedSearchCV with a 4-fold cross-validation, as implemented in Scikit-Learn, utilizing 75% of our dataset. While our original plan designed the remaining 25% for testing, we noted that the train/test split significantly influenced the resulting metrics, likely due to the constrained size of our dataset. To mitigate this variance and bolster the robustness of our results, we opted to employ the complete dataset for metric computation. This was accomplished through 100 randomized train/test splits (75 : 25), applying the previously determined hyperparameters. The predictive accuracy was then ascertained by averaging across various metrics from these 100 trials, a method adopted from Saxena *et al.*[41] In Table 1 we report the hyperparameters considered for the classification and regression algorithms. To implement the feature scaling, features are centered by removing the mean and then scaled to unit variance.

3 Results and discussion

3.1 Dataset analysis

ML models were trained and implemented using the dataset of 920 entries as described in Sections 2.2 and 2.3. In Fig. 2, we

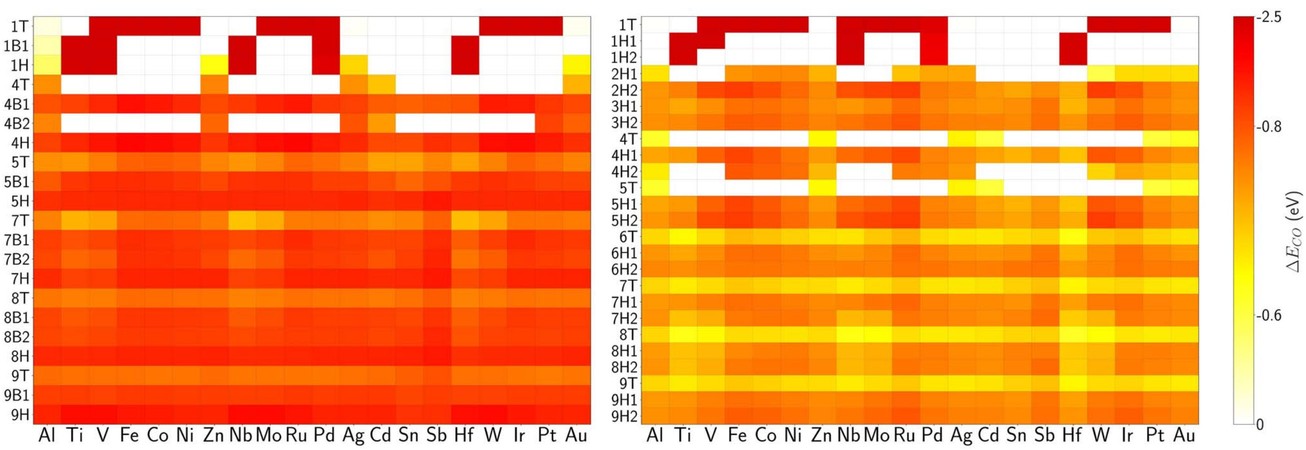

Fig. 2 Matrix representation of the adsorption energy dataset. The left panel shows the CO adsorption energies on Cu(100), while the right panel reports their values on Cu(111). The adsorption energy is 0 eV for non-stable sites.

Table 2 F1 (upper table), MSE and R^2 values (lower table) obtained in the assessment of the performance of the different ML algorithms on the majority (Maj) and minority (Min) classes. These results correspond to averages over 100 random training/test data splits

Classification	F1 training set		F1 test set	
Algorithm	Maj	Min	Maj	Min
GBC	1.000	1.000	0.988	0.934
SVC_{linear}	0.952	0.791	0.949	0.781
SVC_{poly}	1.000	1.000	0.967	0.827
RFC	1.000	1.000	0.966	0.812
DTC	1.000	0.998	0.985	0.915
NNs	0.976	0.870	0.965	0.815
LR	0.953	0.783	0.948	0.760
KNN	0.969	0.856	0.926	0.672

Regression	MSE (eV^2)		R^2	
Algorithm	Training	Test	Training	Test
GBR	0.000	0.003	1.000	0.970
SVC_{rbf}	0.005	0.017	0.949	0.827
RFR	0.001	0.007	0.985	0.933
DTR	0.000	0.006	1.000	0.938
NNs	0.004	0.010	0.942	0.848
GPR	0.000	0.049	1.000	0.533

show the CO adsorption energies (ΔE_{CO}) obtained on all the unique sites of the substitutionally doped $Cu_{1-x}M_x(111)/(100)$ surfaces. From these data, we observe how sites located far from the impurity (these are positions 3, 6, 7, 8 and 9 in Cu(111), and 5, 7, 8 and 9 in Cu(100)) are only slightly affected by the guest atom. Consequently, their adsorption energies behave like the pristine surfaces, *i.e.* $\Delta E_{CO}^{T} < \Delta E_{CO}^{H1} < \Delta E_{CO}^{H2}$ for Cu(111) and $\Delta E_{CO}^{T} < \Delta E_{CO}^{B1} \approx \Delta E_{CO}^{B2} < \Delta E_{CO}^{H}$ for Cu(100). In addition, we note that CO binds stronger on Cu(100) compared to Cu(111), in agreement with previous theoretical studies.[71,72] In general, adsorption energies far from the impurity were found to be within the range of −0.60 to −0.90 eV.

On the other hand, binding energies in the vicinity of the guest atom are strongly influenced to the extent that some adsorption sites become unstable. For the stable sites, binding energies span from thermoneutral (**T** sites with M = Al, Zn, Ag and Au) to *ca.* −2.4 eV (**T** sites with M = Fe, Co, Ru, W and Ir). Altogether, it is evident that predicting site stability and CO binding energies is a complex non-linear challenge to address where ML algorithms can have a major impact.

3.2 Performance of ML models

It has been demonstrated that there is no single optimal ML model that works for any problem.[41,73] Hence it is always advisable to test different ML algorithms to assess which one performs best for a specific application. In this work, we trained the ML models summarized in Table 1 and evaluated their performance following the procedure described in Section 2.3. The estimation of their accuracy is based on two indices for the classification part (*i.e.* F1 scores on the stable/majority and unstable/minority classes), and two different indices for the regression part (*i.e.* mean squared error, MSE, and coefficient of determination, R^2). The results of this analysis are summarized in Table 2 and Fig. 3. The closer the F1 and R^2 scores are to 1, the better the performance of the classification and regression, respectively.

For the classification part, it is important to highlight that the dataset is strongly asymmetric since most of the adsorption sites are stable (780/920 ≈ 85% of the dataset). This however does not influence tree-based algorithms and NNs. In fact, GBC, RFC, DTC, and NNs perform extremely well both on the majority and minority classes (Fig. 3a). SVC with the polynomial kernel and KNN also exhibit high F1 scores for both classes, while the imbalance of the dataset compromises the accuracy of simpler models, such as LR and SVC with a linear kernel. To overcome this issue, we applied a penalty which weighs differently the two classes, *i.e.* 0 and 1 for minority and majority classes, respectively. With this correction, the performance of the LR and SVC algorithms improves significantly on the minority class. Overall, we find the GBC model to be the best performing algorithm with an average F1 score of 0.96.

Regarding the regression part, based on the MSE and the R^2 values computed for the training (left columns) and test (right columns) datasets (see Table 2 and Fig. 3b) we conclude that all methods perform remarkably well, except GPR. This is due to the strong dependence of GPR on the random split used for training. Indeed, from the 100 different random splits used for this model, we observe very good performances ($R^2 = 0.89$, MSE $= 0.005$ eV^2), as well as some poor results ($R^2 = 0.25$, MSE $= 1.98$ eV^2). Analyzing the parity plots for the four best algorithms (GBR, RFR, NNs, and DTR), shown in Fig. 4, we conclude that these models exhibit a high level of accuracy in the prediction of CO binding energies on both Cu(111) and Cu(100) surfaces. Furthermore, these models have the capability of making forecasts across a wide range of values, which is crucial to accurately describe CO adsorption in the sites near the impurity. These sites are particularly important for fine tuning catalytic performance, and therefore their behavior is the most important aspect to be captured in the ML models. Despite the excellent performance displayed by DTR, RFR, and NNs, the GBR algorithm shows overall the most promising predictive power for regression.

Based on all the results discussed above, we propose the GBC model for the classification part, and the GBR algorithm for the regression task.

3.3 Feature sensitivity analysis

After training the ML models, it is possible to obtain the relative feature importance. As an example, we have examined the feature importance in the GBC/GBR algorithms based on two metrics, namely the frequency of its usage for node splitting in the decision tree, and the performance improvement resulting from that split. These metrics are then averaged over all the trees within the model to assess the final feature importance,[41] as depicted in Fig. 5. As far as the stability of a specific adsorption site is concerned, it can be observed that the distance from the impurity D_{Im} (the feature 'Distance') plays

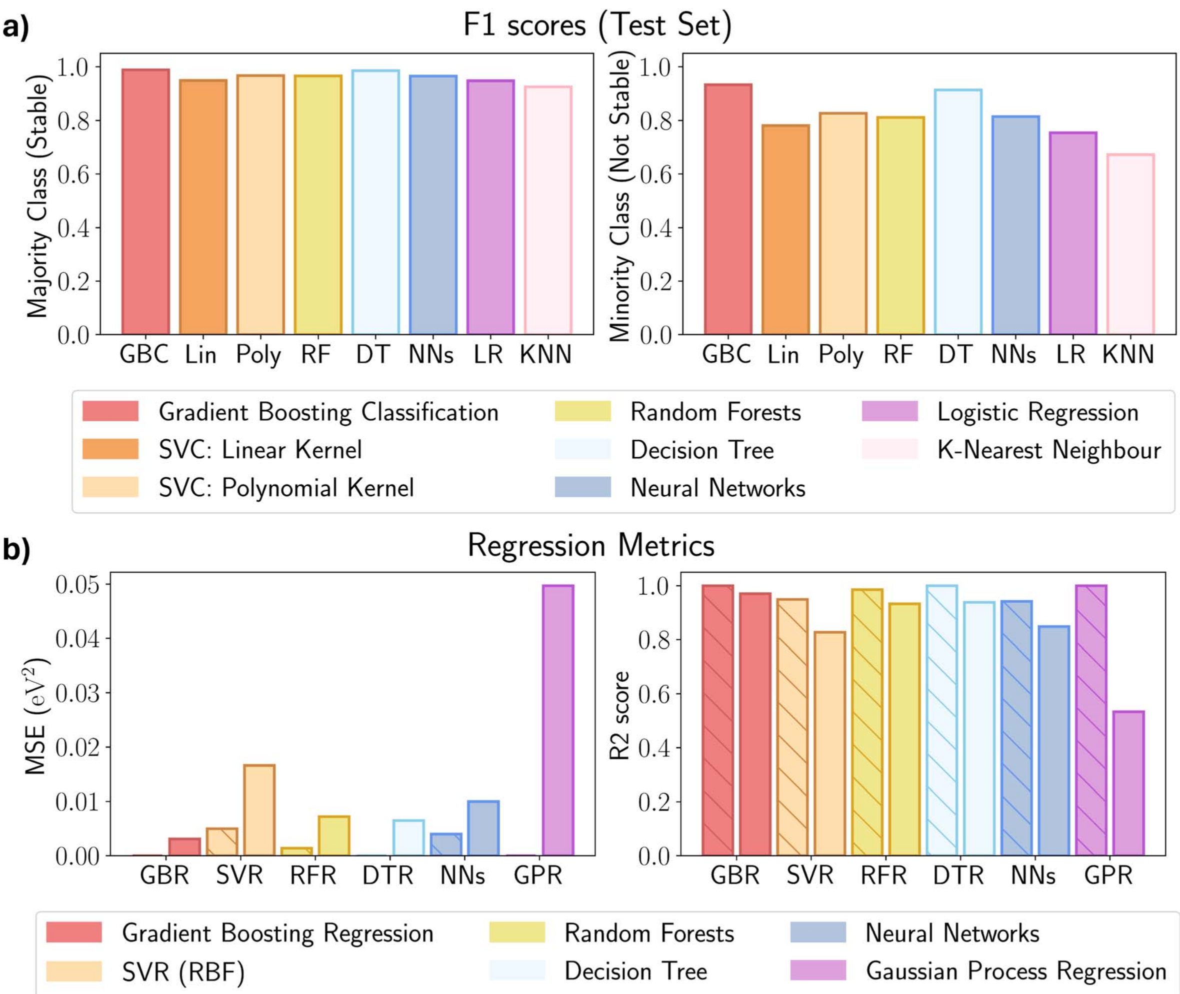

Fig. 3 (a) F1 scores obtained on the test set using different ML classification algorithms. The F1 scores on the majority/stable (minority/not stable) class are reported on the left (right) panel. (b) MSE (left) and R^2 scores (right) obtained with the different ML regression models. Bars with diagonal lines indicate the values in the training set, while solid bars identify the values in the test set.

a fundamental role (Fig. 5, left panel), as beyond a certain value (*ca.* 2.7 Å) all the sites considered in the dataset are predicted to be stable. Hence this feature holds substantial significance since it seems to correlate positively with stable adsorptions. However, from Fig. 2 we can also see that the stability of adsorption sites also depends on other factors, since some sites behave differently even when they are at the same distance from the impurity (different ΔE_{CO} within a given row, Fig. 2). The relative feature importance obtained with GBC also reveals that properties such as the group, surface energy, work function, ionization energy, and electron affinity of the guest species contribute to the predictive power of the ML algorithm. Interestingly, none of these features emerge as being more relevant than the others, indicating a complex relationship among them in determining whether a site is stable for CO adsorption or not. The GCN does not significantly enhance the description of site stability as this is well described by the chemical properties previously discussed when $D_{Im} \leq 2.7$ Å. This outcome is accentuated by the observation that identical site types can exhibit different stabilities contingent upon the impurity

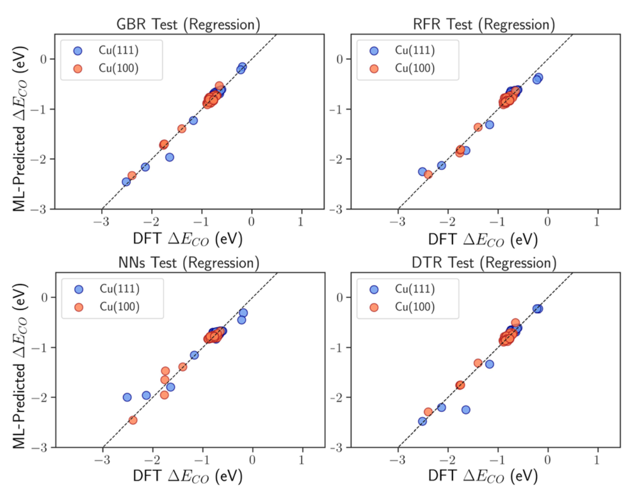

Fig. 4 Parity plots with the four best performing algorithms. Blue dots indicate CO adsorption energies on the Cu(111) surface, while orange dots denote the binding energies on Cu(100). The *y* axis represents the adsorption energies predicted by the ML algorithms, while the *x* axis the DFT-calculated values. The closer the dots are to the black dashed line ($y = x$), the better the model performs.

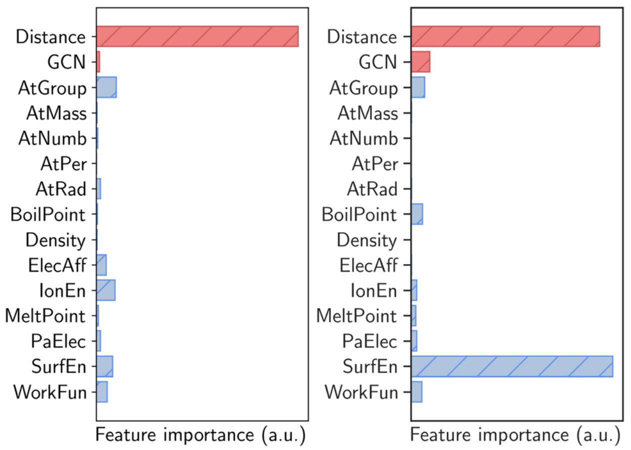

Fig. 5 Relative feature importance of the GBC (left) and GBR (right) models. The 13 chemical features are represented by blue bars, while the 2 geometrical properties are in orange.

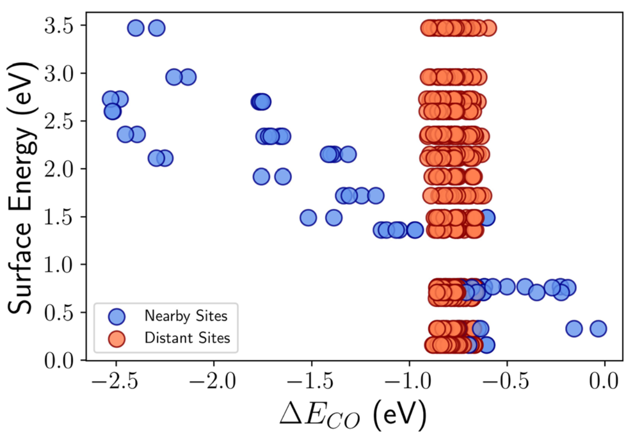

Fig. 6 Relationship between CO adsorption energies, guest atom surface energies, and distance (D_{Im}) of the adsorption site from the guest atom. Blue points (orange points) are for $D_{\mathrm{Im}} \leq 2.7$ Å ($D_{\mathrm{Im}} > 2.7$ Å).

involved. In contrast, for $D_{\mathrm{Im}} > 2.7$ Å, all sites are found to be stable, reinforcing the premise that proximity to the impurity is the dominant factor in describing this behavior, irrespective of the type of binding site.

Regarding the regression part, we find that the distance from the impurity is also important for the prediction of CO binding energies. In fact, we observe that when D_{Im} is *ca.* ≤ 2.7 Å the adsorption energies exhibit substantial variability, whereas when $D_{\mathrm{Im}} > 2.7$ Å, there are less fluctuations. However, we note that the labels also depend on the guest atom's surface energy, as shown in Fig. 6, consistently with other studies.[74,75] In particular, an interesting linear relationship between surface and binding energies is observed for adsorption sites around the M species (blue dots), with higher surface energies resulting in increased CO adsorption strengths, while the two quantities become uncorrelated for $D_{\mathrm{Im}} \geq 2.7$ Å (orange dots). Other important chemical characteristics including group, ionization energy, and boiling temperature exhibit comparable behaviors (see Fig. S1†), supporting the idea that the effect of the substitutional impurities is very local. Consequently, the GCN feature becomes more important in the regression part, because this is the only feature which provides relevant information for predicting CO adsorption energies far from the impurity.

Overall, we conclude that the most important chemical features to predict stability and CO binding energies on the surfaces investigated in this work are the surface energy, ionization energy, electron affinity, group, and boiling and melting points of the guest species M, which we rationalize as follows.

• Surface energies describe surface reactivity; the higher the surface energy, the higher the reactivity.

• The ionization energy and electron affinity dictate how easily an electron can be transferred between the surface metal atoms and the adsorbate.

• The group is related to the number of electrons in the outermost electron shell, and hence determines the chemical behaviour of an element.

• Boiling and melting points, on the other hand, are indicative of the strength of chemical bonds between atoms.

Based on these observations, we can conclude that GBC and GBR models successfully capture the underlying physics and intricate interactions within our systems. Therefore, the exceptional predictive performance of the GBC and GBR models, coupled with the understanding of the key chemical features, allows us to gain valuable insights into the CO adsorption trends on different Cu surfaces.

With this knowledge, we next evaluated the impact of feature selection on the model performance. For this, we removed the least important features from the dataset to create a reduced feature vector consisting of 7 elements, namely 5 chemical properties of the alloying materials (*i.e.* IonEn, AtGroup, BoilPoint and MeltPoint, and SurfEn) and 2 geometrical properties (*i.e.* Distance and GCN). The results from this analysis are summarized in Table S1.† GBC shows slightly improved F1 scores, increasing from 0.934 to 0.941 and from 0.988 to 0.989 for the minority and majority classes, respectively. Similarly, the GBR model displays an enhanced performance with the R^2 score increasing from 0.970 to 0.978 and the MSE decreasing from 0.003 to 0.002 eV^2. This suggests that the selected features contain the most relevant information needed for the prediction of CO binding energies, and therefore removing the least important features does not significantly affect the models' accuracy. Importantly, this feature selection process can potentially reduce the risk of overfitting, leading to more robust and generalizable predictions across different catalyst surfaces, while enhancing the efficiency of the ML models in terms of computational cost and interpretability.

3.4 Data scalability

We next evaluated the impact of the number of available data points on the performance of the GBC and GBR models. While maintaining the test set fixed at 25% of the entire dataset, we varied the number of points in the training set by selecting different percentages from the remaining 75% of the dataset, specifically 50%, 60%, 70%, 80%, and 90%. Data points in these subsets were randomly sampled. Results are shown in Fig. 7.

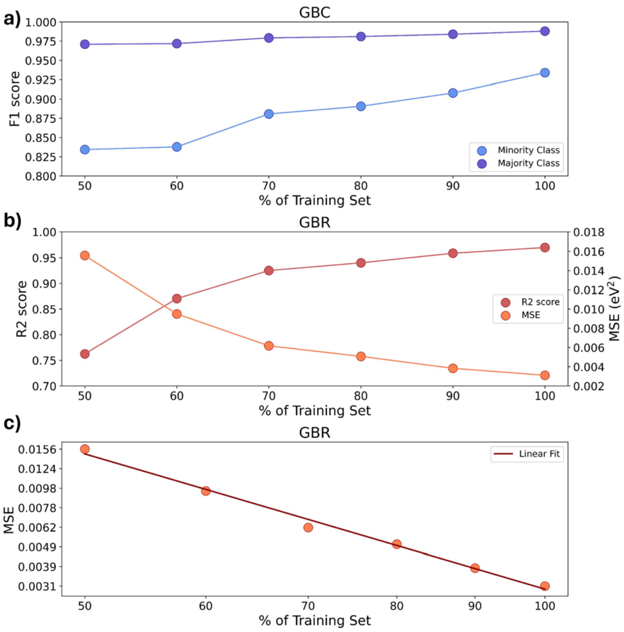

Fig. 7 Evaluation of the (a) GBC and (b) GBR models performance across varying percentages of training set points. (c) Logarithmic relationship between MSE and the percentage of training set points.

In the classification part, the F1 score for the majority class (stable sites) sees marginal improvements in comparison to the minority class, as depicted in Fig. 7a. This observation likely stems from the relatively large number of data points for stable sites, even when only a small portion of the training set is utilized. In contrast, the dataset for unstable sites is much smaller, which introduces more variability in the model's predictions as different percentages of the training set are sampled.

Regarding the regression part, as shown in Fig. 7b, the performance metrics significantly improve with the addition of more data, aligning with our expectations. Furthermore, the log–log plot presented in Fig. 7c details the relationship between the MSE and the number of training data points. Typically, these plots reveal a linear trend that can be extrapolated to estimate the data volume required to achieve a certain performance level. The linear arrangement of the data points on our plot suggests a consistent training process. By extrapolation, we estimate that to reduce the MSE by a factor of e (approximately 2.718), the volume of training data would need to increase by approximately 50%. For our dataset, this equates to an additional 300 data points (calculated as $0.5 \times 0.75 \times 780$). However, given the high accuracy already demonstrated by our models, and the marginal improvement anticipated from expanding the dataset, we have opted to maintain the current dataset size.

3.5 CO binding energy predictions on higher stoichiometry $Cu_{1-x}M_x$ alloys

Given the strong influence of D_{Im}, both in the classification and the regression parts, we next explored the possibility of exploiting our models to describe $Cu_{1-x}M_x$(111)/(100) surfaces with a higher impurity content compared to the dataset (*i.e.* $0.028 \leq x \leq 0.168$). For this, we consider Au and Ag as guest species since these alloys have been widely studied for CO_2RR applications.[76–80] Firstly, we predicted the ground state structures of $Cu_{1-x}Ag_x$ and $Cu_{1-x}Au_x$ alloys in the range $0 \leq x \leq 0.3$ by means of a cluster expansion method[81–86] (see the ESI†). Then, we selected three of the ground state structures with intermediate impurity concentrations and predicted the CO adsorption energies using the two-step ML method described above. Specifically, we focused on Cu(111) surfaces with 11.1% Au ($Cu_{0.889}Au_{0.111}$) and 16.7% Ag ($Cu_{0.833}Ag_{0.167}$), and a Cu(100) surface with 12.5% Au ($Cu_{0.875}Au_{0.125}$), shown in Fig. S2 and S3, and S9a and b,† respectively.

Notably, despite training our ML models on structures with very low surface impurity concentrations (*ca.* 3%), we were pleased to see that they also perform remarkably well for higher concentrations. In particular, for the CuM(111) surfaces, GBC predicts the stability of Ag and Au binding sites with 100% accuracy, while for the CuM(100) surface, the average F1 score is 0.90. We attribute this difference to the fact that the (100) surface is inherently less stable than the (111), and therefore, the introduction of a larger number of substitutional impurities has a larger negative impact on the former, thus reducing the predictive power of the ML algorithms. Nevertheless, we note that the majority of the trained models exhibit average F1 scores in the original test set that are comparable to the obtained Cu bimetallic surfaces with high impurity concentrations. This highlights the potential of these methods for the accelerated prediction of CO binding energies for CuM alloys.

In addition, we our GBR model can predict ΔE_{CO} values on CuAu(111), CuAg(111) and CuAu(100) with remarkable accuracy, *i.e.* R^2/MSE (eV^2) = 0.789/0.004, R^2/MSE (eV^2) = 0.872/0.004 and R^2/MSE (eV^2) = 0.733/0.006, respectively. From the parity plots in Fig. 8 we also observe that our models can accurately predict ΔE_{CO} both in the vicinity and far away from the impurity. While GBR seems to slightly underestimate ΔE_{CO}, overall the results can be deemed very satisfactory considering that the impurity concentration in these systems is *ca.* 4–6 times larger than the one used for training. The adsorption energy maps

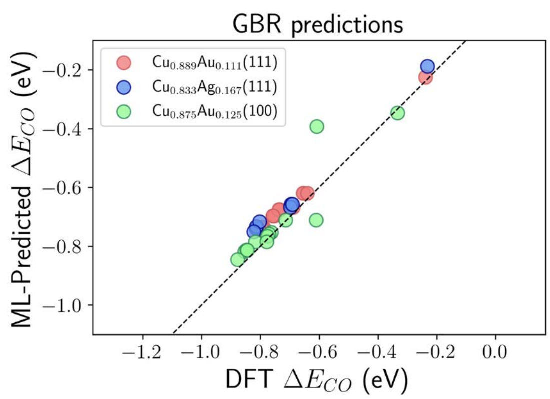

Fig. 8 Parity plot obtained with GBR at varying impurity concentrations in the Cu(111) and Cu(100) surfaces. Different colors correspond to the different structures considered.

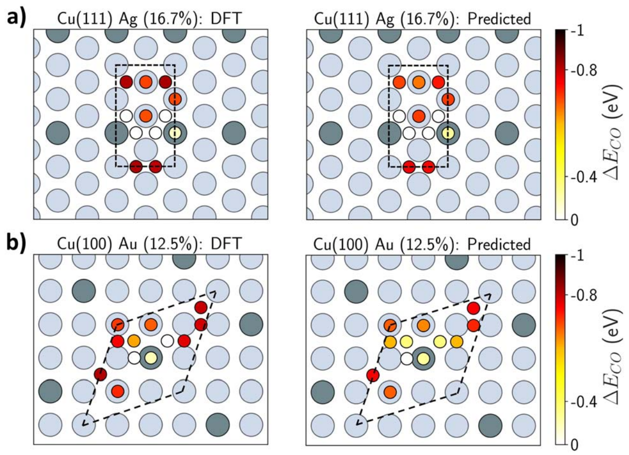

Fig. 9 DFT-calculated (left) and ML-predicted (right) CO adsorption energy maps over the (a) $Cu_{0.833}Ag_{0.167}(111)$ and (b) $Cu_{0.875}Au_{0.125}(100)$ surfaces. The dashed black lines highlight the unit cells. The smaller dots correspond to different binding sites for CO, with the colors representing their adsorption energies. White dots denote unstable sites. Color code: Cu (light blue), Au/Ag (dark gray).

calculated by DFT and predicted by our two-step ML approach (GBC + GBR) on the $Cu_{0.833}Ag_{0.167}(111)$ and $Cu_{0.875}Au_{0.125}(100)$ surfaces are shown in Fig. 9. The small dots in the maps denote the different binding sites for CO with their colors expressing their associated ΔE_{CO} values, while white dots indicate unstable sites. We note that for the $Cu_{0.833}Ag_{0.167}(111)$ surface, the **B** site between the two impurities is stable. However, since for the (111) surface we did not consider **B** as a possible adsorption site, our model is unable to predict this configuration. This site is stable only due to the particular geometrical configuration of the substitutional impurities, hence it represents a pathological case that could be addressed by expanding the dataset with systems showing this kind of structure.

In summary, observing the maps reported in Fig. 9, we conclude that our two-step ML approach is capable of accurately predicting CO adsorption energies on Cu bimetallic (111) and (100) surfaces with high impurity concentrations up to *ca.* 17%. This remarkable capability of these models is envisioned to enable the high-throughput screening of bimetallic alloys to accelerate the discovery of CO_2RR electrocatalysts.

4 Conclusions

In this work we report a novel two-step ML approach based on classification and regression algorithms to predict CO adsorption energies on Cu-based bimetallic alloys. Among the ML models developed herein, GBC exhibits the best performance in classifying the type of binding sites (as stable or unstable), while GBR outperforms all the models in the prediction of CO binding energies on the identified stable sites. The features used to describe the Cu bimetallic (111) and (100) surfaces include the readily available chemical and geometrical properties of the guest species and binding sites, respectively, which prove very advantageous. Firstly, they make our ML models highly interpretable as they contain much descriptive information about adsorbate–surface interactions. For example, the chemical features of the impurity, namely the ionization energy, atomic group, surface energy and electron affinity, have a strong influence on the surrounding Cu atoms within a radius of 2.7 Å, while the GCN does not have a significant effect on the classification of stable/unstable binding sites. In addition, we find an interesting linear correlation between the surface energy of the guest species and the CO binding energies in the vicinity of the impurity. Notably the boiling and melting points, as well as the GCN of the binding site, become more important in the prediction of CO adsorption energies than in the classification of the binding site. Secondly, our approach is accessible and computationally efficient, since it is built on readily available features which do not require any further calculations.

Our machine learning models demonstrate exceptional performance, accurately predicting CO binding energies on a wide array of Cu-based bimetallic alloys. This accuracy is maintained even when the models are tested on systems with substitutional impurity concentrations *ca.* 6 times higher than those in the training set, showcasing the models' versatility and broad applicability. Leveraging the CO binding energy as a reaction descriptor for C_2/C_{2+} product formation, our study highlights Cu surfaces doped with Ni and V as promising candidate materials. These metals exhibit a stronger CO binding compared to Cu, favoring C–C coupling by increasing the likelihood of CO molecules being in close proximity. Significantly, the CO binding strength on these metals is optimal, reducing the risk of CO surface poisoning. Nonetheless, a comprehensive assessment of catalytic performance must also account for the H binding energy[87] and the surface coverage of the electrocatalyst under relevant applied potentials and pH.[88] The universal and adsorbate-independent features incorporated into our models guarantee their applicability to a range of CO_2RR intermediates and electrocatalytic processes. In addition, this versatility holds great promise for applying our methodology to investigate various adsorbates across different Cu : M ratios, thereby expediting the design of efficient, selective electrocatalysts for the CO_2RR. Such an approach is poised to transform the production of chemical fuels and value-added compounds, enabling the rapid screening and experimental synthesis of novel materials with superior catalytic performance.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

The CINECA award under the ISCRA initiative and HPC@POLITO are acknowledged for the availability of high-performance computing resources and support. MRF, FR and GC acknowledge the High-Performance Computing, Big Data, and Quantum Computing Research Centre, established under the Italian National Recovery and Resilience Plan (PNRR).

Notes and references

1 S. J. Davis, N. S. Lewis, M. Shaner, S. Aggarwal, D. Arent, I. L. Azevedo, S. M. Benson, T. Bradley, J. Brouwer, Y.-M. Chiang, C. T. M. Clack, A. Cohen, S. Doig, J. Edmonds, P. Fennell, C. B. Field, B. Hannegan, B.-M. Hodge, M. I. Hoffert, E. Ingersoll, P. Jaramillo, K. S. Lackner, K. J. Mach, M. Mastrandrea, J. Ogden, P. F. Peterson, D. L. Sanchez, D. Sperling, J. Stagner, J. E. Trancik, C.-J. Yang and K. Caldeira, *Science*, 2018, **360**, eaas9793.

2 N. Mac Dowell, P. S. Fennell, N. Shah and G. C. Maitland, *Nat. Clim. Change*, 2017, **7**, 243–249.

3 C.-T. Dinh, F. P. García de Arquer, D. Sinton and E. H. Sargent, *ACS Energy Lett.*, 2018, **3**, 2835–2840.

4 S. Verma, Y. Hamasaki, C. Kim, W. Huang, S. Lu, H.-R. M. Jhong, A. A. Gewirth, T. Fujigaya, N. Nakashima and P. J. A. Kenis, *ACS Energy Lett.*, 2018, **3**, 193–198.

5 K. Jiang, S. Siahrostami, T. Zheng, Y. Hu, S. Hwang, E. Stavitski, Y. Peng, J. Dynes, M. Gangisetty, D. Su, K. Attenkofer and H. Wang, *Energy Environ. Sci.*, 2018, **11**, 893–903.

6 Y. Zheng, A. Vasileff, X. Zhou, Y. Jiao, M. Jaroniec and S.-Z. Qiao, *J. Am. Chem. Soc.*, 2019, **141**, 7646–7659.

7 E. Jeng and F. Jiao, *React. Chem. Eng.*, 2020, **5**, 1768–1775.

8 H. Ooka, M. C. Figueiredo and M. T. M. Koper, *Langmuir*, 2017, **33**, 9307–9313.

9 S. Nitopi, E. Bertheussen, S. B. Scott, X. Liu, A. K. Engstfeld, S. Horch, B. Seger, I. E. L. Stephens, K. Chan, C. Hahn, J. K. Nørskov, T. F. Jaramillo and I. Chorkendorff, *Chem. Rev.*, 2019, **119**, 7610–7672.

10 Y. Hori, A. Murata and R. Takahashi, *J. Chem. Soc., Faraday Trans.*, 1989, **1**, 2309–2326.

11 D. W. DeWulf, T. Jin and A. J. Bard, *J. Electrochem. Soc.*, 1989, **136**, 1686.

12 E. Bertheussen, T. V. Hogg, Y. Abghoui, A. K. Engstfeld, I. Chorkendorff and I. E. L. Stephens, *ACS Energy Lett.*, 2018, **3**, 634–640.

13 Y. Hori, R. Takahashi, Y. Yoshinami and A. Murata, *J. Phys. Chem. B*, 1997, **101**, 7075–7081.

14 L. Wang, S. A. Nitopi, E. Bertheussen, M. Orazov, C. G. Morales-Guio, X. Liu, D. C. Higgins, K. Chan, J. K. Nørskov, C. Hahn and T. F. Jaramillo, *ACS Catal.*, 2018, **8**, 7445–7454.

15 N. Zhang, B. Yang, K. Liu, H. Li, G. Chen, X. Qiu, W. Li, J. Hu, J. Fu, Y. Jiang, M. Liu and J. Ye, *Small Methods*, 2021, **5**, 2100987.

16 A. A. Peterson, F. Abild-Pedersen, F. Studt, J. Rossmeisl and J. K. Nørskov, *Energy Environ. Sci.*, 2010, **3**, 1311–1315.

17 K. P. Kuhl, T. Hatsukade, E. R. Cave, D. N. Abram, J. Kibsgaard and T. F. Jaramillo, *J. Am. Chem. Soc.*, 2014, **136**, 14107–14113.

18 P. Sabatier, *Ber. Dtsch. Chem. Ges.*, 1911, **44**, 1984–2001.

19 M. Jouny, W. Luc and F. Jiao, *Nat. Catal.*, 2018, **1**, 748–755.

20 W. Luc, X. Fu, J. Shi, J.-J. Lv, M. Jouny, B. H. Ko, Y. Xu, Q. Tu, X. Hu, J. Wu, Q. Yue, Y. Liu, F. Jiao and Y. Kang, *Nat. Catal.*, 2019, **2**, 423–430.

21 Y. C. Li, Z. Wang, T. Yuan, D.-H. Nam, M. Luo, J. Wicks, B. Chen, J. Li, F. Li, F. P. G. de Arquer, Y. Wang, C.-T. Dinh, O. Voznyy, D. Sinton and E. H. Sargent, *J. Am. Chem. Soc.*, 2019, **141**, 8584–8591.

22 C. Chen, Y. Li, S. Yu, S. Louisia, J. Jin, M. Li, M. B. Ross and P. Yang, *Joule*, 2020, **4**, 1688–1699.

23 E. L. Clark, C. Hahn, T. F. Jaramillo and A. T. Bell, *J. Am. Chem. Soc.*, 2017, **139**, 15848–15857.

24 A. K. Buckley, M. Lee, T. Cheng, R. V. Kazantsev, D. M. Larson, W. A. Goddard III, F. D. Toste and F. M. Toma, *J. Am. Chem. Soc.*, 2019, **141**, 7355–7364.

25 S. B. Varandili, J. Huang, E. Oveisi, G. L. De Gregorio, M. Mensi, M. Strach, J. Vavra, C. Gadiyar, A. Bhowmik and R. Buonsanti, *ACS Catal.*, 2019, **9**, 5035–5046.

26 K. Jiang, H.-X. Zhang, S. Zou and W.-B. Cai, *Phys. Chem. Chem. Phys.*, 2014, **16**, 20360–20376.

27 X. Zhang, A. Chen, L. Chen and Z. Zhou, *Adv. Energy Mater.*, 2022, **12**, 2003841.

28 C. Salvini, M. Re Fiorentin, F. Risplendi, F. Raffone and G. Cicero, *J. Phys. Chem. C*, 2022, **126**, 14441–14447.

29 A. Chen, X. Zhang and Z. Zhou, *InfoMat*, 2020, **2**, 553–576.

30 J. E. Sutton and D. G. Vlachos, *Chem. Eng. Sci.*, 2015, **121**, 190–199.

31 Z. Yang, W. Gao and Q. Jiang, *J. Mater. Chem. A*, 2020, **8**, 17507–17515.

32 D. Cheng, F. R. Negreiros, E. Aprà and A. Fortunelli, *ChemSusChem*, 2013, **6**, 944–965.

33 A. Jain and T. Bligaard, *Phys. Rev. B*, 2018, **98**, 214112.

34 K. T. Butler, D. W. Davies, H. Cartwright, O. Isayev and A. Walsh, *Nature*, 2018, **559**, 547–555.

35 G. H. Gu, C. Choi, Y. Lee, A. B. Situmorang, J. Noh, Y.-H. Kim and Y. Jung, *Adv. Mater.*, 2020, **32**, 1907865.

36 T. Zhou, Z. Song and K. Sundmacher, *Engineering*, 2019, **5**, 1017–1026.

37 X. Wang, S. Ye, W. Hu, E. Sharman, R. Liu, Y. Liu, Y. Luo and J. Jiang, *J. Am. Chem. Soc.*, 2020, **142**, 7737–7743.

38 K. Rao, Q. K. Do, K. Pham, *et al.*, *Top. Catal.*, 2020, **63**, 728–741.

39 T.-H. Hung, Z.-X. Xu, D.-Y. Kang and L.-C. Lin, *J. Phys. Chem. C*, 2022, **126**, 2813–2822.

40 D. Ologunagba and S. Kattel, *Energies*, 2020, **13**(9), 2182.

41 S. Saxena, T. S. Khan, F. Jalid, M. Ramteke and M. A. Haider, *J. Mater. Chem. A*, 2020, **8**, 107–123.

42 B. R. Goldsmith, J. Esterhuizen, J.-X. Liu, C. J. Bartel and C. Sutton, *AIChE J.*, 2018, **64**, 2311–2323.

43 S. B. Kotsiantis, *Informatica*, 2007, **31**, 249–268.

44 V. Nasteski, *Horizons*, 2017, **4**, 51–62.

45 L. Mandal, K. R. Yang, M. R. Motapothula, D. Ren, P. Lobaccaro, A. Patra, M. Sherburne, V. S. Batista, B. S. Yeo, J. W. Ager, J. Martin and T. Venkatesan, *ACS Appl. Mater. Interfaces*, 2018, **10**, 8574–8584.

46 Z. Ni, H. Liang, Z. Yi, R. Guo, C. Liu, Y. Liu, H. Sun and X. Liu, *Coord. Chem. Rev.*, 2021, **441**, 213983.

47 T. Li, A. Ciotti, M. Rahaman, C. W. S. Yeung, M. García-Melchor and E. Reisner, *ChemRxiv*, 2023, DOI: 10.26434/chemrxiv-2023-146f9.
48 J.-J. Velasco-Velez, R. V. Mom, L.-E. Sandoval-Diaz, L. J. Falling, C.-H. Chuang, D. Gao, T. E. Jones, Q. Zhu, R. Arrigo, B. Roldan Cuenya, A. Knop-Gericke, T. Lunkenbein and R. Schlögl, *ACS Energy Lett.*, 2020, **5**, 2106–2111.
49 K. Hansen, G. Montavon, F. Biegler, S. Fazli, M. Rupp, M. Scheffler, O. A. von Lilienfeld, A. Tkatchenko and K.-R. Müller, *J. Chem. Theory Comput.*, 2013, **9**, 3404–3419.
50 J. H. Friedman, *Ann. Stat.*, 2001, 1189–1232.
51 J. H. Friedman, *Comput. Stat. Data Anal.*, 2002, **38**, 367–378.
52 P. Giannozzi, S. Baronni and N. Bonini, *J. Phys.: Condens.Matter*, 2009, **21**, 395502.
53 P. Giannozzi, O. Andreussi and T. N. Brumme, *J. Phys.: Condens.Matter*, 2017, **29**, 465901.
54 M. Schlipf and F. Gygi, *Comput. Phys. Commun.*, 2015, **196**, 36–44.
55 J. Perdew, K. Burke and Y. Wang, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 1996, **54**(23), 16533–16539.
56 M. Salomone, M. Re Fiorentin, G. Cicero and F. Risplendi, *J. Phys. Chem. Lett.*, 2021, **12**, 10947–10952.
57 H. J. Monkhorst and J. D. Pack, *Phys. Rev. B*, 1976, **13**, 5188.
58 L. Xu, J. Lin, Y. Bai and M. Mavrikakis, *Top. Catal.*, 2018, **61**, 736–750.
59 R. Tran, Z. Xu, B. Radhakrishnan, D. Winston, W. Sun, K. A. Persson and S. P. Ong, *Sci. Data*, 2016, **3**, 160080.
60 R. Tran, X.-G. Li, J. H. Montoya, D. Winston, K. A. Persson and A. P. Ong, *Surf. Sci.*, 2019, **687**, 48–55.
61 F. Calle-Vallejo and A. S. Bandarenka, *ChemSusChem*, 2018, **11**, 1824–1828.
62 C. Cortes and V. Vapnik, *Mach. Learn.*, 1995, **20**, 273–297.
63 L. Breiman, *Mach. Learn.*, 2001, **45**, 5–32.
64 X. Wu, V. Kumar, J. R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. J. McLachlan, A. Ng, B. Liu, S. Y. Philip, *et al.*, *Knowl. Inf. Syst.*, 2008, **14**, 1–37.
65 E. Bisong, in *Logistic Regression*, Apress, Berkeley, CA, 2019, pp. 243–250.
66 A. Mucherino, P. J. Papajorgji and P. M. Pardalos, in *k-Nearest Neighbor Classification*, Springer New York, New York, NY, 2009, pp. 83–106.
67 C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*, The MIT Press, 2005.
68 F. Chollet, *keras*, 2015, GitHub, https://github.com/fchollet/keras.
69 M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving and M. Isard *et al.*, *12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16)*, 2016, pp. 265–283.
70 F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and E. Duchesnay, *J. Mach. Learn. Technol.*, 2011, **12**, 2825–2830.
71 S. Vollmer, G. Witte and C. Wöll, *Catal. Lett.*, 2001, **77**, 1–3.
72 J. H. Montoya, C. Shi and J. K. Nørskov, *J. Phys. Chem. Lett.*, 2015, **6**(11), 2032–2037.
73 T. F. Sterkenburg and P. D. Grünwald, *Synthese*, 2021, **199**, 9979–10015.
74 T. Toyao, K. Suzuki, S. Kikuchi, S. Takakusagi, K.-i. Shimizu and I. Takigawa, *J. Phys. Chem. C*, 2018, **122**, 8315–8326.
75 H. Zhuang, A. J. Tkalych and E. A. Carter, *J. Phys. Chem. C*, 2016, **120**, 23698–23706.
76 T. Ahmad, S. Liu, M. Sajid, K. Li, M. Ali, L. Liu and W. Chen, *Nano Res. Energy*, 2022, **1**, e9120021.
77 J. Monzó, Y. Malewski, R. Kortlever, F. J. Vidal-Iglesias, J. Solla-Gullón, M. T. M. Koper and P. Rodriguez, *J. Mater. Chem. A*, 2015, **3**, 23690–23698.
78 Z. Chang, S. Huo, W. Zhang, J. Fang and H. Wang, *J. Phys. Chem. C*, 2017, **121**, 11368–11379.
79 S. Zhang, S. Zhao, D. Qu, X. Liu, Y. Wu, Y. Chen and W. Huang, *Small*, 2021, **17**, 2102293.
80 S. Dai, T.-H. Huang, W.-I. Liu, C.-W. Hsu, S.-W. Lee, T.-Y. Chen, Y.-C. Wang, J.-H. Wang and K.-W. Wang, *Nano Lett.*, 2021, **21**, 9293–9300.
81 A. van de Walle and G. Ceder, *J. Phase Equilib.*, 2002, **23**, 348.
82 F. Raffone, C. Ataca, J. C. Grossman and G. Cicero, *J. Phys. Chem. Lett.*, 2016, **7**, 2304–2309.
83 F. Raffone, F. Savazzi and G. Cicero, *Phys. Chem. Chem. Phys.*, 2021, **23**, 11831–11836.
84 M. Salomone, F. Raffone, M. Re Fiorentin, F. Risplendi and G. Cicero, *Nanomaterials*, 2022, **12**, 2079–4991.
85 F. Raffone, F. Savazzi and G. Cicero, *J. Phys. Chem. Lett.*, 2019, **10**, 7492–7497.
86 V. Ankit Kumar, R. Federico and C. Giancarlo, *Nanomater. Nanotechnol.*, 2020, **10**, 1847980420955093.
87 J. Hussain, H. Jónsson and E. Skúlason, *ACS Catal.*, 2018, **8**, 5240–5249.
88 A. Ciotti and M. García-Melchor, *Curr. Opin. Electrochem.*, 2023, **42**, 101402.