ChemCatChem

The European Society Journal for Catalysis

Chemistry Europe

European Chemical Societies Publishing

Supported by GECATS GERMAN CATALYSIS SOCIETY

Accepted Article

Title: Design of an Accurate Machine Learning Algorithm to Predict the Binding Energies of Several Adsorbates on Multiple Sites of Metal Surfaces

Authors: CS Praveen and Aleix Comas-Vives

To be cited as: *ChemCatChem* 10.1002/cctc.202000517

Link to VoR: https://doi.org/10.1002/cctc.202000517

WILEY-VCH

10.1002/cctc.202000517

WILEY-VCH

FULL PAPER

Design of an Accurate Machine Learning Algorithm to Predict the Binding Energies of Several Adsorbates on Multiple Sites of Metal Surfaces

CS Praveen[a,b] and Aleix Comas-Vives*[c]

[a] Dr. Praveen CS
International School of Photonics
Cochin University of Science and Technology
University Road, South Kalamassery,
Kalamassery, Ernakulam, Kerala 682022, India

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

[c] Dr. A. Comas-Vives
Department of Chemistry
Universitat Autònoma de Barcelona
08193 Cerdanyola del Vallès, Catalonia, Spain
E-mail: Aleix.Comas@uab.cat

Supporting information for this article is given via a link at the end of the document.

Abstract: In the current work, we design a single unique machine learning (ML) algorithm capable of predicting the binding energies of several C, N and O-based adsorbates and atomic hydrogen on different facets (100, 111, 211) of eleven transition-metals considering an FCC bulk structure (Co, Rh, Ir, Ni, Pd, Pt, Ru, Os, Cu, Ag, Au) with high accuracy with respect to the reference DFT calculations. The selected properties/features are based on already available data or electronic properties easily obtained from DFT calculations of the free adsorbates and on the clean metal surfaces. The mean average error (MAE) for the training set is equal to 0.074 eV, while for the test set (data not used in training) is equal to 0.174 eV, thus having a very small error with respect to the reference DFT calculations. Further modifications of the present algorithm, which is based on an Extragradient boost (XGBoost) regressor in combination with a tree booster, with the addition of few more features might set the ground to develop ML algorithms able to predict the binding energies of adsorbates on more complex catalytic surfaces.

1. Introduction

Machine learning (ML) combined with data availability is revolutionizing our life via the so-called fourth industrial revolution.[1] Matter simulation[2] and the field of chemistry and computational catalysis is not exempt from this paradigm shift. The use of ML algorithms[3] is rapidly increasing in the field of chemistry[4] and catalysis.[5] ML methods can be used to develop atomistic potentials,[6] to search for active motifs in intermetallic structures[7] or to address reaction network complexity,[8] holding great potential in combination with data repositories of electronic structure calculations.[9]

ML algorithms have been used to predict binding energies of adsorbates on surfaces but have been limited to predict the binding energy of one or just a few similar adsorbates of the same family, i. e. oxygen-based for instance, while modifying the surfaces (mostly on different facets of metals).[5d, 10] A key quantity to obtain structure-activity relationships in heterogeneous catalysis is the adsorption energy of the reaction intermediates of interest on the surface of the catalysts and in computational heterogeneous catalysis this is usually evaluated via DFT calculations.[11] In this framework, the first step is to calculate the chemisorption of the different reaction intermediates along the proposed pathways for the reaction of interest. In the screening of catalytic materials adsorption energies can then be combined with Brønsted-Evans-Polanyi (BEP) relationships in order to screen energy barriers based on the calculated thermodynamics.[12] Nevertheless, for real catalysts, such as metal particles, several potential catalytic sites are present and thus in principle the evaluation of the adsorption energies of each adsorbate/intermediate on each different adsorption sites is essential. This, however, increases the computational cost significantly. The usual strategy to reduce computational cost to perform extensive catalytic screening, especially for metal surfaces, consists on evaluating just a few facets and using scaling techniques for describing the adsorption energies of the same family species (for instance, CH_x (x=1,3), OH_x (x=0,1,2) and NH_x (x=0,1,2,3).[13] These scaling relationships are highly relevant since breaking this relationships is actually a strategy to find more active catalytic materials.[14] Another option is the development of descriptors to predict binding energies and here is worth mentioning the success of the *d*-band model to understand the binding of adsorbates on the surface of transition metals.[15] In addition, geometrical parameters are also important to predict binding energies for different facets of the same metal. Calle-Vallejo and Loffreda *et. al.* proposed generalized coordination numbers as a structural descriptor for the adsorption energy of small oxygen- and hydrogen containing adsorbates on Pt nanoparticles of various sizes.[12a, 16] Later, Calle-Vallejo et. al. also introduced structural sensitivity into the scaling relationships among adsorption energies of O-based adsorbates (O, OCH_3 and OH) by using coordination numbers. In addition, former work of Rossmeisl and Wei-Lu and Calle-Vallejo, already showed that

generalized scaling relationships indeed exist also between different family adsorbates.[17] Thus, further generality should be possible in order to account for the binding of adsorbates of different nature simultaneously on multiple adsorption sites of metal surfaces. The goal of the current paper is to design a highly accurate machine learning algorithm, able to predict the adsorption energies of several adsorbates binding either via C, N, O and H (for atomic hydrogen) on the surface sites on many facets of transition-metals simultaneously. The selected adsorbates are some of the key intermediates of the reaction networks of relevant industrial catalytic transformations, *i. e.* the Water-Gas Shift Reaction (WGSR),[18] Steam and Dry Reforming of Methane,[18] the Methanol Synthesis,[19] methanation,[20] Fischer-Tropsch synthesis[21] and ammonia synthesis.[22] A summary of the selected adsorbates, transition metals and the facets considered as well as the their properties/features used to train the machine learning (ML) algorithms to predict the binding energies are shown in Figure 1.

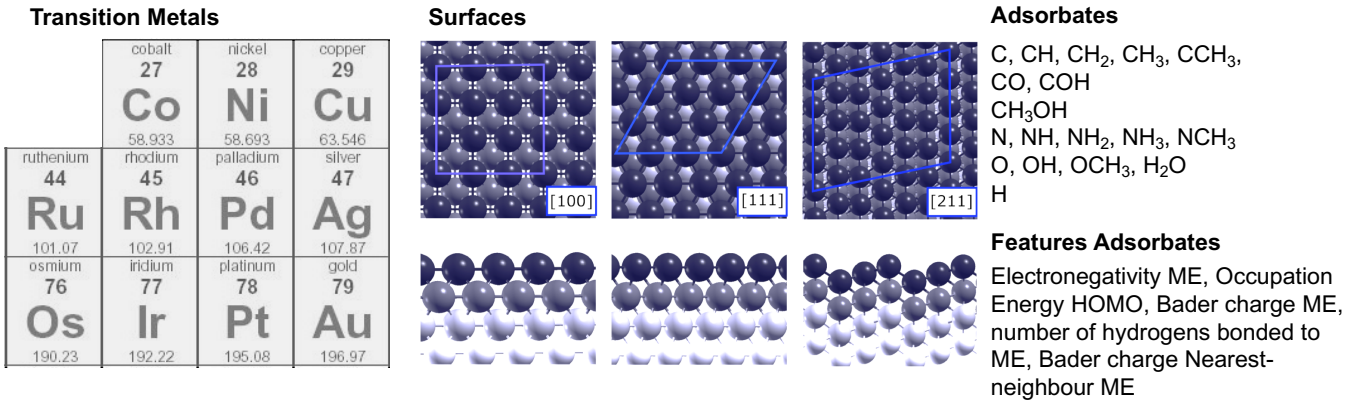


Figure 1. Overview of the metals, facets adsorbates and all related features/properties considered in this study. The goal is predicting the binding energy of the adsorbates on all metals and facets. Acronyms: GCN (Generalized Coordination number), Coordination number (CN), ME (Main element directly bonded to the surface), Occupation (p and s occupation of the main element directly bonded to the surface).

The selected features to train the ML algorithms are based on electronic and geometrical properties of the free adsorbates in the gas phase and the clean metal surfaces. Combining these features and without the need of significant hyperparameter tuning, our model predicts with really high performance the binding energies calculated at DFT level for all evaluated test cases with high accuracy using a single ML algorithm including all the adsorbates at once, in contrast to previous approaches. The present work lays the ground for a universal prediction of binding energies on metal surfaces and further modifications of it could make it even more general; for instance, in order to predict the binding energies of the evaluated adsorbates on the surfaces of related but more complex materials such as alloys or intermetallics.

2. Results and Discussion

2.1 Description of Systems Evaluated

The adsorbates selected are grouped according to the main element directly bonded to the surface, i.e. carbon-based: C, CH, CH_2, CCH_3, CH_3OH, CO, COH; nitrogen-based: N, NH, NH_2, NH_3, NCH_3 and O-based: O, OCH_3, OH and H_2O. Finally, we also included monoatomic H. The selected metals are Co, Rh, Ir, Ni, Pd, Pt, Ru, Os, Cu, Ag and Au, while the selected facets are the 100, 111, and the 211 surfaces. For most of the adsorbates, one adsorption site per adsorbate on each surface was considered. For the adsorption of C, CH and CO adsorbates on the 211 surfaces, two different sites (three-fold and four-fold) were considered for all the metal surfaces. For these adsorbates, four-fold sites are relevant since they have been proposed to be involved in coking processes as well as in key steps of the Fischer-Tropsch Synthesis and methanation reactions.[20-21, 23] Finally, the adsorption of NCH_3 on the Au100 surface was not considered, since the surface reconstructs significantly upon adsorption. Overall, this makes a total number of 623 calculations to perform the subsequent analysis via Machine Learning (ML) algorithms.

2.2 Description of Features used in the ML algorithm

We now briefly explain the features selected for the prediction of the binding energy of different adsorbates into several sites of transition-metal sites. We used a total of 13 features/properties, which are divided into features from the metal surface and those of the adsorbate (see Figure 1). All features are based on properties already of the free adsorbate and the surface sites of the bare metal surfaces. The properties/features to train and test the designed ML algorithms are related to the geometry of the activity sites (generalized coordination numbers)[12a] and coordination numbers), properties of the elements involved in direct bonding (electronegativity), and electronic properties obtained from DFT calculations. For the free adsorbate; energy of the HOMO level, *p* occupation of the main group element directly bonded to the surface (for C, N, and O-based adsorbates), *s*-occupation (for atomic H only), and Bader charges, while for the clean metal surfaces the *d*-occupation and Bader charges of the metal atoms in the top-most layer are obtained from DFT calculations. It is to be noted, however, that the features estimated from DFT corresponds to the free adsorbates in the gas phase and the clean metal surfaces without any adsorbate on them, i.e., no feature corresponding to the optimized adsorbate on the metal surface is taken. The generalized coordination number (GCN) of a surface site is estimated by considering not only the nearest neighbours of the adsorption site, but also the coordination number of those nearest neighbours as well by giving a weight normalized with respect to the maximum coordination number. A detailed description of GCN is provided in the ESI.

2.2.1 Features from the adsorbate in the gas-phase

Energy of the highest occupied band of the free adsorbate (HOMO), Electronegativity, total *p*-occupation (*s*- for atomic hydrogen). Bader charge of the atom directly bonded to the metal surface upon adsorption, number of hydrogens bonded to the atom directly bonded to the surface upon adsorption. If the atom directly bonded to the surface has a nearest neighbour which is a main element, for instance the case of CO, then we also include as feature of the Bader charge of that atom as a feature. If that nearest neighbour has hydrogens bonded to it, i. e. COH, then we also add the number of hydrogens bonded to that atom.

2.2.2 Features from the metal surfaces

Electronegativity (Pauling scale) of the metal, number of the period in the periodic table of elements of the corresponding metal, generalized coordination numbers, as well as coordination number of the atoms composing the adsorption site, type of site: squared (4), tri-fold (3), bridge (2) or top (1), *d*-occupation of the atoms of the metal surface and Bader charges of the top-most layer of the metal surfaces.

2.3 Definition of the Predicted Quantity, *i. e.* Binding Energy.

The quantity we are interested in to predict by means of a ML algorithm is the binding energy of chemisorbed adsorbates on metal surfaces ($E_{adsorption}$)] obtained via DFT calculations. $E_{adsorption}$ is calculated with respect to the energy of the free adsorbates in the gas-phase ($E_{adsorbate}$) and the energy of the specific clean metal surface ($E_{surface}$) using Equation 1 given below.

$$E_{adsorption} = E_{adsorbate/surface} - \left(E_{surface} + E_{adsorbate}\right) \quad (1)$$

All the binding energies calculated at DFT level (see computational details) are reported in the last column of the following supplementary files: final-paper-dataset.csv and final-paper-dataset-hcp.csv (see also ML details).

2.4 Performance of different ML algorithms

We tested different state-of-the-art ML algorithms in combination with the features/properties previously described. In particular, we tested; multivariable linear regression (MLR), Kernel ridge regression (KRR), Gaussian Progressive Regression (GPR) and Extragradient Boost (XGBoost) regression using a Tree based model (*gbtree*) as a booster.[24] Table 1 reports the summary of the performance of all the evaluated ML algorithms, while Table 2 reports their respective statistical metrics.

Table 1. Summary of the performance of the evaluated machine learning algorithms using all the features described along the text.

Algorithm	Training Performance[a]	Test Performance[b]	Accuracy[c]	Standard deviation[d]
MLR	0.748	0.819	0.752	0.061
KRR	0.986	0.970	0.957	0.017
SVR	0.965	0.969	0.958	0.008
GPR	0.972	0.971	0.960	0.011
XGBoost	0.998	0.987	0.987	0.005

[a] Performance of the training set. [b] Performance of the test set. [c] Accuracy based on k-fold (K = 10) cross validation and [d] related final standard deviation.

Table 2. Mean Squared Error (MSE) and Mean Absolute Error (MAE) for all the evaluated machine learning algorithms for both the training and the test sets, respectively.

Algorithm	MSE Training Set (eV)	MAE Training Set (eV)	MSE Test Set (eV)	MAE Test Set (eV)
MLR	1.187	0.884	0.981	0.810
KRR	0.066	0.144	0.162	0.277
SVR	0.163	0.884	0.166	0.301
GPR	0.133	0.255	0.158	0.294
XGBoost	0.008	0.069	0.068	0.198

Overall, the XGBoost regression algorithm performs best in comparison to all tested ML algorithms, in the training and test sets scores and in accuracy, with respective quantities equal to 0.998, 0.987 and 0.987 (Table 1). The multivariate linear regression has a high bias, *i. e.* a low training performance (0.748). Even though all other algorithms (KRR, SVR and GPR) display a rather good performance, the values are lower than for the XGBoost regression algorithm (Table 1). In addition, the mean absolute error (MAE) for the train and the test sets for the XGBoost algorithm are equal to 0.069 and 0.198 eV respectively, which are significantly lower than the MAEs of the other evaluated algorithms (see Table 2). A direct visual representation of the performance of the XGBoost regression algorithm is shown in Figure 2, depicting ML predicted energy (in eV) vs. the binding energy calculated at DFT level (in eV). The training and test sets are represented by blue and orange dots, respectively. The mean square error (MSE) of the training set is equal to 0.008 eV, while the MSE of the test set is equal to 0.069 eV. The analogous graphs for each of the other evaluated algorithms are provided in the ESI (see Figure S1).

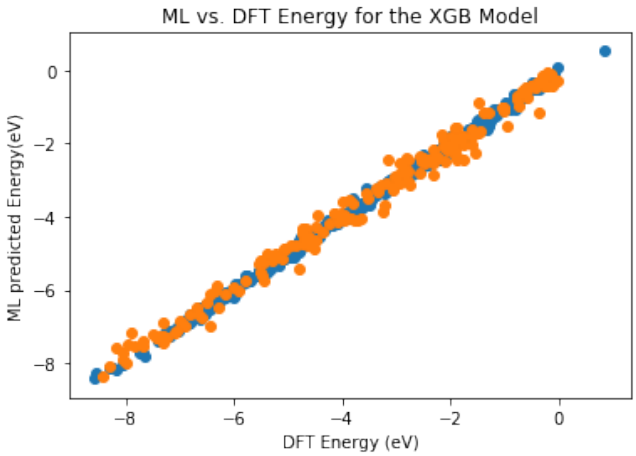


Figure 2. ML predicted vs. DFT calculated binding energy in eV for the training set (blue points, representing 70 % of the data) and the test set (orange points on top of the blue points, representing 30 % of the data).

The learning curve display the performance of a ML algorithm as a function of the training examples and it is used in data science

to evaluate whether additional data is needed to increase the performance of a given ML algorithm. The corresponding plot for the XGBoost algorithm is shown in Figure 3. This Figure indicates that the performance of the algorithm is not expected to improve significantly when increasing the amount of data used for training since the score is already fairly converged when the number of training examples are close to 400. This means that the inclusion of facets with adsorption sites of similar nature than those used, i. e. including for instance other bridge, three or four-fold sites are not likely to be needed to increase significantly the overall performance of the XGBoost algorithm. The learning curves for all the other evaluated ML algorithms can be found in the ESI (Figure S2). It is true, however, that adding just a few additional data from adsorptions on the Ru(0001) hcp-based facet (see Machine Learning details and ESI), does not improve accuracy but further reduces the MAE of the test set to only 0.174 eV (see Table S1 of the ESI). Our work resembles the through study from Andersen et. al. In their work, they used sensing methods (SISSO) to identify descriptors to predict the binding energies of few adsorbates on potential surface sites of transition metals and bimetallic alloys. [25] They combined 18 primary features of the metal surfaces via 8 algebraic and functional operations, obtaining a 8D-ϕ_3, descriptor (trained on the pooled metals and alloyed data set). They use this descriptor and combine it with a non-linear SISSO prediction and obtained a RMSE for the training of pure metals equal to 0.09 eV and for the test data set (alloys) equal to 0.15 eV. In our case, we have only pure metals but a significantly higher number of adsorbates and energy ranges for the binding energy. Our RMSE for the training and test set are equal to 0.10 eV and 0.24 eV respectively. Their better performance in comparison to ours might be due to the following aspects: they did a different fitting for each adsorbate and evaluated more sites/coordination environments for each of the metal surfaces.

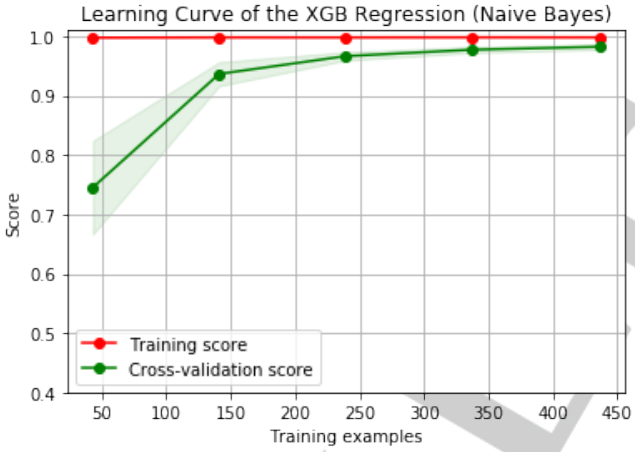


Figure 3. Learning curve corresponding to the Extragradient Boost algorithm with a tree-based booster, which is the ML algorithm that performs best among all tested ML algorithms after hyperparameter tuning.

After having the developed the ML model, the most important step is to understand which of the features considered are most important to determine the final binding energy. We checked different feature importance evaluation methods, based on gain, weight and coverage, but the methods are not consistent with each other (see Figure S3 of the ESI), *i. e.* the order of importance of the evaluated features depends on the evaluation metrics either based on gain, weight or coverage. Further explanation on these three feature importance metrics is given on the ESI. Thus, we decided to use a unified framework to interpret model predictions employing Shapley Additive Explanations (SHAP), which is a game theoretic approach to explain the output of any machine learning model.[26] We then used a recently developed exact algorithm to evaluate tree ensemble methods based on SHAP values to evaluate feature importance in our developed ML algorithm.[27] The summary plot shows both feature importance and the effect of the value of each feature for each data-point for the predicted property of the ML algorithm. Each point on the summary plot is a Shapley value for a feature and a dataset. The position on the *y*-axis is determined by the feature and the position of the *x*-axis by the Shapley (SHAP) value. Overall, this means that features with large absolute SHAP values are the most important ones. Hence, the features are ordered according to their importance from top to bottom, i. e. the *p*-occupation (*s*-occupation for H) of the main element bonded to the surface is the most important one. See also the ESI for the mean SHAP value for each feature (Figure S4), which is another way to visualize feature importance. In addition, Figure 4, shows the effect of the feature value (either low; blue or high; red, as represented by the colorbar of Figure 4) has on our predicted property (SHAP score), i.e. the binding energy. For instance, this means that for low values of orbital occupation of the main element (blue points) the binding energy usually becomes more negative since they have mostly negative SHAP values. Conversely, for high orbital occupation numbers (red points) the binding energy becomes more positive since SHAP values are positive, and thus bonding becomes less favorable. It can be easily inferred that the three most important features correspond to electronic properties of the adsorbate. The most important metal features are also related to its electronic properties: the group the metal occupies in the periodic table of the elements, the Bader charge of the metal as well as its *d*-occupation. Thus, the most important features are electronic parameters as expected for a bonding process. The first geometrical features in decreasing order of importance are the number of metal atoms involved in bonding and the generalized coordination number. It is also worth mentioning that the coordination number of the metal surface is the least important feature.

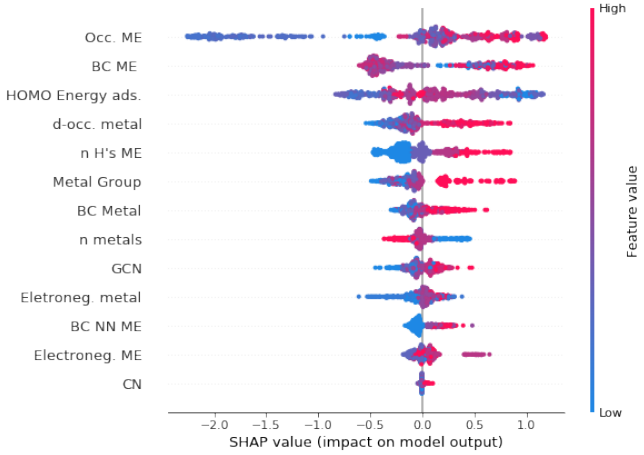


Figure 4. Overview of the most important features for a model, i.e. SHAP values of every feature for every data point. The plot sorts each feature by the sum of SHAP magnitudes over all samples and uses SHAP values to show the distribution of the impacts each feature has on the model output. The color represents the feature value (red, high; blue low).

The SHAP analysis is also useful for feature reduction, since removing up to first three least important features from the SHAP analysis does not alter the MSE, MAE or the accuracy of the algorithm. Thus, the model can be simplified by using only 10 features and leads to the same accuracy. However, removing also the fourth least important feature (i.e., GCN together with the other three least important features of the SHAP analysis) already increases the MAE up to 0.253 eV, and also increases the MSE and decreases the accuracy significantly (see Table S2 of the

ESI). Finally, we also analysed the correlation between all features as well as between each feature and the binding energy of the adsorbate. The resulting heatmap ranges from -1.0 for inverse correlation and 1.0 for direct correlation, while 0.0 means no correlation. The resulting analysis is shown in Figure 5. The generalized coordination number and the coordination number are positively correlated between them. The electronic features of the adsorbate are highly positively correlated between them and analogously, the same happens for the electronic properties of the metal surfaces. Another variable showing a high degree of correlation is the number of metal atoms, which correlates with many electronic features of the adsorbates in an inverse way, especially with the HOMO energy of the adsorbate. Notably, the number of metals involved in bonding with the adsorbate is highly inversely correlated with the binding energy, being the feature that correlates most intensively with the latter quantity. Then, it follows the HOMO energy of the adsorbate, which correlates positively with the binding energy. Finally, as expected and as it should be, the features of the adsorbate and the ones of the metal surfaces do not show any correlation between them.

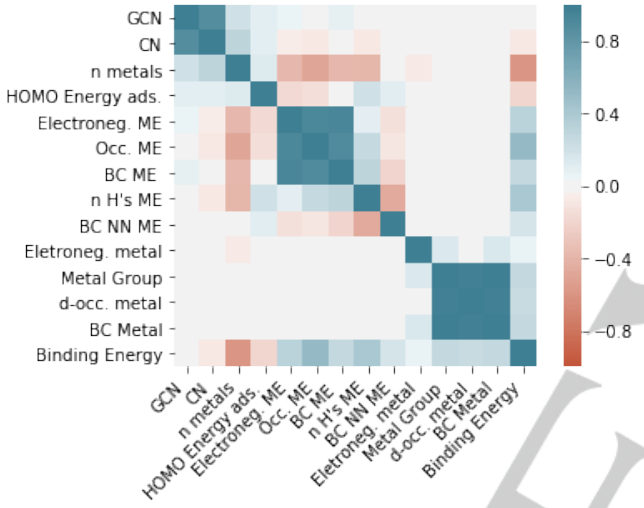


Figure 5. Correlation between all features with each other as well as with the predicted binding energy.

Conclusions

We designed a general ML algorithm that is able to predict the chemisorption energy of several gas-phase adsorbates on several facets of metals, simultaneously. Various ML algorithms were tested, of which Extragradient Boost (XGBoost) regression in combination with a tree booster displays the best performance. The XGBoost regression reproduces with high accuracy the binding energy of chemisorbed adsorbates upon optimization on specific sites of metal surface obtained via DFT calculations. A total number of 13 features based on available or easily obtained properties of the free adsorbate in the gas-phase and on the clean metal surfaces were tested to train the accuracy of the ML algorithm, while the practically same accuracy is obtained with only 10 features. The evaluation of the importance of features bring us back to chemistry and concepts of bonding of adsorbates on surfaces since the most important features determining the binding energy are electronic properties, primarily from the adsorbate and then from the metal. For the adsorbate the most important properties determining the binding energy are the *p*-occupation value of the main element directly bonded to the surface (*s*-occupation value for atomic H) and its Bader charge followed by the HOMO energy of the adsorbate. For the metal, the most important properties are the group the metal occupies in the periodic table of the elements as well as the *d*-occupation value of the metal atoms forming the adsorption site. The most important geometrical parameters correspond to the number of metals directly bonded to the adsorbate as well as to the generalized coordination number of the metal adsorption site.

ML algorithms still have limitations when being used from scratch for atomistic discovery with superior accuracy for any material.[28] Training the algorithms in similar systems than those in the test set is still needed, so to blindly predict the binding energies in unexplored systems still remains challenge. One needs to get the predictions right for a significantly large amount of available data as well as to perform an uncertainty analysis of systems not included in the training. This, however, needs enormous amount of computational time and human effort. On the other hand, a possible strategy towards accurate prediction is identifying features that have a significant effect in yielding high accuracy on wide variety of systems. Therefore, alternative strategies must also need to be identified to reduce the width of the confidence interval of the data predicted using ML for many systems

Thus, our future directions of research in this field will involve the modification of the current ML algorithm to extend it to predict the binding energy of more complex adsorbates, for instance multidentate ones, as well as their adsorption on the surfaces of more complex materials, such as alloys, intermetallics, carbides, nitrides or oxides as well as interfaces such as metal-oxide interfaces.

Computational Details

All periodic DFT calculations were carried out with Vienna Ab Initio Simulation Package (VASP) code[29] and the projector augmented wave (PAW) method,[30] for which interactions between valence electrons and ion cores are described by pseudopotentials and the electronic wavefunctions are expanded in terms of a discrete plane-wave basis set. A plane-wave energy cutoff of 400 eV was used for all calculations. Electron exchange and correlation were treated with the generalized gradient approximation (GGA) within the Perdew-Burke-Ernzerhof (PBE) functional.[31] Nevertheless, we expect the approach also work for other functionals as long as the features and energies are calculated using the same level of theory. Spin-polarized effects were considered for the case of magnetic metals. Brillouin zone sampling was performed using Monkhorst–Pack grids.[32] For the 100 and 111 surface models consist of periodic a 3x3 four-layer slabs and a total of 36 atoms. For the latter two surfaces the metal atoms from the bottom layer were kept fixed in their crystal lattice positions during all optimizations. The surface model of the 211 surfaces corresponds to a 2x4 twelve-layered cell, with a total of 96 metal atoms. For the 211 surfaces, we fixed the three last layers (24 atoms). For the Ru, Co and Os surfaces we also considered the facets from an fcc-phase although their most stable phase in the bulk is the hcp one. We calculated the adsorption energies of the following adsorbates (C, CH, CH_2, CH_3, N. NH, NH_2, NH_3, O, OH, H_2O and CO) on the Ru(0001)-hcp facet. Then, we did two tests to check how this limitation affects our designed ML algorithm. In the first one, the prediction of these binding energies using our previously designed ML algorithm for this completely new data unseen by our algorithm is still good, with MSE and MAE values equal to 0.137 and 0.300 eV, respectively. In the second one, we included this data in our algorithm, formed a new training and test sets and re-evaluated the accuracy, MSE and MAE of the ML algorithm. The accuracy and the MAE of the training set of the algorithm is practically the same, but the MAE value of the test set improves, taking a value of 0.174 eV. (see Table S1 of the ESI). A vacuum separation of 15 Å was in the direction perpendicular to the surface for all evaluated surfaces. The binding energies of all evaluated adsorbates are calculated with respect to the energy of the free

adsorbate in the gas-phase. Spin polarization effects were included in the calculations when needed. In particular, for adsorbates with unpaired electrons as well as for magnetic metal surfaces (Ni, Co and Os).

Machine Learning Details

The dataset is provided as supplementary file (full-paper-dataset.csv). The extended dataset including a few adsorption energies on the Ru0001 surface is also included (full-paper-dataset-hcp.csv). The scikit-learn program[33] was used to train, test and cross-validate all the evaluated machine learning algorithms. We used 70 % of randomly selected data for the training set while the remaining data (30 % of the total) was used for our test set. K-fold cross validation (k = 10) was used to calculate the accuracy of all ML algorithms as reported in Table 1. It is expressed from 0 to 1, being 1 the maximal possible accuracy (100 % in percentage). XGBoost regressor[34] using a tree-based model was imported to scikit-learn. The model used in gradient boosted trees are tree ensembles. A tree ensemble model consists of a set of classification and regression trees (CART). This is the same model than the one used for random forest (RF) and their difference with boosted trees is the training process. When using a tree in boosting methods we build trees one at a time, each new tree helping to correct errors from the previously trained tree. The maximal depth and number of estimators are important to obtain an excellent performance. Thus, we first evaluated the performance as function of the maximal depth values (between 3 and 6) and the number of estimators between 100 and 5000. Accordingly, a value of 4 was selected for the maximal depth value whereas the number of estimators was set to 2000, since this combination of parameters gave the best performance. Further hyperparameter tuning of the XGBoost algorithm was performed for the next following parameters: learning rate and gamma. Learning rate and gamma were screened between 0.01 and 1 for each subsequent order of magnitude, *i. e.* every factor of 10. The best performing values of the learning rate and the gamma were equal to 0.1 and 0.01, respectively. Then, we further screened both the parameters, finally selecting a value for the learning rate and the gamma equal to 0.125 and 0.0075, respectively. Finally, the minimal child weight, alpha and lambda regularization were adjusted, taking respective final values equal to 4, 0.1 and 0.1. The summary of the parameters finally used for the XGBoost algorithm as well as for the other ML algorithms are reported in the ESI. The analysis of the feature importance was performed via the tree explainer method developed by Su-In Lee and co-workers.[27] The correlation heatmap was constructed by means of the seaborn package.[35]

Acknowledgements

CSP acknowledges DST-India for INSPIRE Faculty Fellowship with award number IFA-18 PH217. A.C-V. thanks the Spanish MEC and the European Social Fund (Ramon y Cajal Fellowship: RyC-2016-19930) and the Spanish "Ministerio de Ciencia, Innovación y Universidades" (PGC2018-100818-A-I00) for financial support.

Keywords: Machine Learning • Adsorption Energies • Heterogeneous Catalysis • Metals • Surfaces

[1] K. Schwab in *The Fourth Industrial Revolution*, *Vol.* Encyclopaedia Britannica, inc., **2018**.
[2] A. Aspuru-Guzik, R. Lindh and M. Reiher, *ACS Cent. Sci.* **2018**, *4*, 144-152.
[3] C. Robert, *Chance* **2014**, *27*, 62-63.
[4] a) B. Huang, N. O. Symonds and O. A. v. Lilienfeld, *Handbook of Materials Modeling* **2018**, 1-27; b) A. Jinich, B. Sanchez-Lengeling, H. Ren, R. Harman and A. Aspuru-Guzik, *ACS Cent. Sci.* **2019**, *5*, 1199-1210.
[5] a) P. Schlexer Lamoureux, K. T. Winther, J. A. Garrido Torres, V. Streibel, M. Zhao, M. Bajdich, F. Abild-Pedersen and T. Bligaard, *ChemCatChem* **2019**, *11*, 3581-3601; b) T. Toyao, Z. Maeno, S. Takakusagi, T. Kamachi, I. Takigawa and K.-I. Shimizu, *ACS Catal.* **2019**; c) B. Meyer, B. Sawatlon, S. Heinen, O. A. von Lilienfeld and C. Corminboeuf, *Chem. Sci.* **2018**, *9*, 7069-7077; d) K. Tran and Z. W. Ulissi, *Nat. Catal.* **2018**, *1*, 696-703; e) B. Sawatlon, M. D. Wodrich, B. Meyer, A. Fabrizio and C. Corminboeuf, *ChemCatChem* **2019**, *11*, 4096-4107.
[6] a) V. L. Deringer, M. A. Caro and G. Csányi, *Adv. Mat.* **2019**, *31*, 1902765; b) A. P. Bartók, S. De, C. Poelking, N. Bernstein, J. R. Kermode, G. Csányi and M. Ceriotti, *Sci. Adv.* **2017**, *3*, e1701816; c) S. Chmiela, H. E. Sauceda, K.-R. Müller and A. Tkatchenko, *Nat. Commun.* **2018**, *9*, 3887.
[7] Z. W. Ulissi, M. T. Tang, J. Xiao, X. Liu, D. A. Torelli, M. Karamad, K. Cummins, C. Hahn, N. S. Lewis, T. F. Jaramillo, K. Chan and J. K. Nørskov, *ACS Catal.* **2017**, *7*, 6600-6608.
[8] Z. W. Ulissi, A. J. Medford, T. Bligaard and J. K. Nørskov, *Nat. Commun.* **2017**, *8*, 14621.
[9] a) O. Mamun, K. T. Winther, J. R. Boes and T. Bligaard, *Scientific Data* **2019**, *6*, 76; b) M. Álvarez-Moreno, C. de Graaf, N. López, F. Maseras, J. M. Poblet and C. Bo, *J. Chem. Inf. Model.* **2015**, *55*, 95-103; c) L. M. Ghiringhelli, C. Carbogno, S. Levchenko, F. Mohamed, G. Huhs, M. Lüders, M. Oliveira and M. Scheffler, *Npj Comput. Mater.* **2017**, *3*, 46; d) C. Bo, F. Maseras and N. López, *Nat. Catal.* **2018**, *1*, 809-810.
[10] R. García-Muelas and N. López, *Nat Commun* **2019**, *10*, 4687.
[11] J. K. Nørskov, F. Abild-Pedersen, F. Studt and T. Bligaard, *Proc. Nat. Acad. Sci.* **2011**, *108*, 937-943.
[12] a) F. Calle-Vallejo, J. I. Martínez, J. M. García-Lastra, P. Sautet and D. Loffreda, *Angew. Chem. Int. Ed.* **2014**, *53*, 8316-8319; b) F. Calle-Vallejo, D. Loffreda, M. T. M. Koper and P. Sautet, *Nat. Chem.* **2015**, *7*, 403-410.
[13] Z.-J. Zhao, S. Liu, S. Zha, D. Cheng, F. Studt, G. Henkelman and J. Gong, *Nat. Rev. Mater.* **2019**, *4*, 792-804.
[14] J. Pérez-Ramírez and N. López, *Nat. Catal.* **2019**, *2*, 971-976.
[15] B. Hammer and J. K. Norskov, *Nature* **1995**, *376*, 238-240.
[16] F. Calle-Vallejo, J. Tymoczko, V. Colic, Q. H. Vu, M. D. Pohl, K. Morgenstern, D. Loffreda, P. Sautet, W. Schuhmann and A. S. Bandarenka, *Science* **2015**, *350*, 185-189.
[17] a) F. Calle-Vallejo, J. I. Martínez, J. M. García-Lastra, J. Rossmeisl and M. T. M. Koper, *Physical Review Letters* **2012**, *108*, 116103; b) H.-Y. Su, K. Sun, W.-Q. Wang, Z. Zeng, F. Calle-Vallejo and W.-X. Li, *J. Phys. Chem. Lett.* **2016**, *7*, 5302-5306.
[18] a) L. Foppa, T. Margossian, S. M. Kim, C. Müller, C. Copéret, K. Larmier and A. Comas-Vives, *J. Am. Chem. Soc.* **2017**, *139*, 17128-17139; b) L. Foppa, M.-C. Silaghi, K. Larmier and A. Comas-Vives, *J. Catal.* **2016**, *343*, 196-207; c) L. Foppa, K. Larmier and A. Comas-Vives, *Chimia* **2019**, *73*, 239-244.
[19] a) K. Larmier, W.-C. Liao, S. Tada, E. Lam, R. Verel, A. Bansode, A. Urakawa, A. Comas-Vives and C. Copéret, *Angew. Chem. Int. Ed.* **2017**, *56*, 2318-2323; b) E. Lam, J. J. Corral-Pérez, K. Larmier, G. Noh, P. Wolf, A. Comas-Vives, A. Urakawa and C. Copéret, *Angew. Chem. Int. Ed.* **2019**, *58*, 13989-13996.
[20] L. Foppa, M. Iannuzzi, C. Copéret and A. Comas-Vives, *J. Catal.* **2019**, *371*, 270-275.
[21] a) L. Foppa, M. Iannuzzi, C. Copéret and A. Comas-Vives, *ACS Catal.* **2018**, *8*, 6983-6992; b) L. Foppa, M. Iannuzzi, C. Copéret and A. Comas-Vives, *ACS Catal.* **2019**, *9*, 6571-6582; c) L. Foppa, C. Copéret and A. Comas-Vives, *J. Am. Chem. Soc.* **2016**, *138*, 16655-16668.
[22] a) G. Ertl, *Catal. Rev.* **1980**, *21*, 201-223; b) K. Honkala, A. Hellman, I. N. Remediakis, A. Logadottir, A. Carlsson, S. Dahl, C. H. Christensen and J. K. Nørskov, *Science* **2005**, *307*, 555-558.
[23] R. A. van Santen, A. J. Markvoort, I. A. W. Filot, M. M. Ghouri and E. J. M. Hensen, *Physical Chemistry Chemical Physics* **2013**, *15*, 17038-17063.
[24] T. Chen and C. Guestrin in *XGBoost: A Scalable Tree Boosting System*, *Vol.* Association for Computing Machinery, San Francisco, California, USA, **2016**, pp. 785–794.

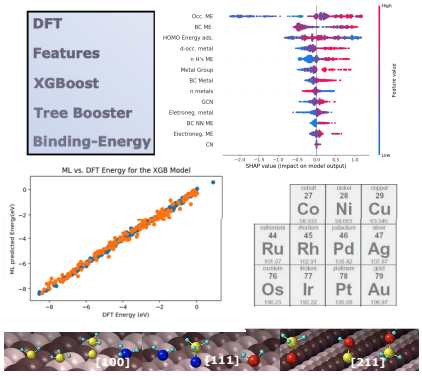
[25] M. Andersen, S. V. Levchenko, M. Scheffler and K. Reuter, *ACS Catal.* **2019**, *9*, 2752-2759.
[26] S. M. Lundberg and S.-I. Lee in *A Unified Approach to Interpreting Model Predictions*, *Vol.* Eds.: I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan and R. Garnett), Curran Associates, Inc., pp. 4765-4774.
[27] S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal and S.-I. Lee, *Nature Machine Intelligence* **2020**, *2*, 56-67.
[28] A. A. Peterson, R. Christensen and A. Khorshidi, *Physical Chemistry Chemical Physics* **2017**, *19*, 10978-10985.
[29] a) G. Kresse and J. Hafner, *Phys. Rev. B* **1993**, *47*, 558-561; b) G. Kresse and J. Furthmüller, *Comput. Mater. Sci.* **1996**, *6*, 15-50; c) G. Kresse and J. Furthmüller, *Phys. Rev. B* **1996**, *54*, 11169-11186.
[30] P. E. Blochl, *Phys. Rev. B* **1994**, *50*, 17953-17979.
[31] J. P. Perdew, K. Burke and M. Ernzerhof, *Phys. Rev. Lett.* **1996**, *77*, 3865.
[32] H. J. Monkhorst and J. D. Pack, *Phys. Rev. B* **1976**, *13*, 5188-5192.
[33] 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. Res.* **2011**, *12*, 2825-2830.
[34] a) J. H. Friedman, *Ann. Statist.* **2001**, *29*, 1189-1232; b) J. H. Friedman, *Comput. Stat. Data Anal.* **2002**, *38*, 367-378.
[35] Seaborn, 10.5281/zenodo.3629446.

10.1002/cctc.202000517

WILEY-VCH

FULL PAPER

Entry for the Table of Contents



We designed a Machine Learning (ML) algorithm able to reproduce with high accuracy and simultaneously the binding energy obtained at DFT level for several C, N, O-based adsorbates and atomic H on several sites of metal surfaces.