Predicting Hydrogen Adsorption Energies on Platinum Nanoparticles and Surfaces with Machine Learning

Lea Gašparič
lea.gasparic@ijs.si
Jožef Stefan Institute, Jožef Stefan international postgraduate schoole
Ljubljana, Slovenia

Anton Kokalj
tone.kokalj@ijs.si
Jožef Stefan Institute, Jožef Stefan international postgraduate school
Ljubljana, Slovenia

Sašo Džeroski
saso.dzeroski@ijs.si
Jožef Stefan Institute, Jožef Stefan international postgraduate school
Ljubljana, Slovenia

Abstract

The growing interest in hydrogen gas as a fuel drives research into environmentally friendly hydrogen production methods. One viable approach of obtaining hydrogen is the electrocatalysis of water, which includes the hydrogen evolution reaction (HER) as one of the half-reactions. In the search of highly active catalysts for the HER, machine learning can be effectively utilized to develop models for calculating hydrogen adsorption energy, a key descriptor of catalytic activity. In this study, we learned models for predicting hydrogen adsorption energy on platinum. We used various machine-learning (ML) techniques on two datasets, one for extended surfaces and the other for nanoparticles. The respective results reveal that ML models for extended surfaces are more accurate than those for nanoparticles, and that the features describing the local environment are the most significant for the predictions. For surfaces, the coordination number is the most relevant feature, while the d-band center is the most important for nanoparticles. The ML models developed in this study lack sufficient accuracy to provide reliable results, highlighting the need for further investigation with additional features or larger datasets.

Keywords

platinum, hydrogen, DFT calculations, decision trees, feature ranking

1 Introduction

A lot of scientific and societal interest is devoted to hydrogen fuel, which can generate electrical power by producing water as a byproduct. One environmentally friendly method of producing hydrogen is through the electrocatalysis of water, where hydrogen and oxygen gases are formed. This process involves two reactions: oxygen and hydrogen evolution reactions. Considerable effort is being directed towards improving catalysts for both reactions and understanding the fundamental processes involved [21, 13]. In this contribution, we will focus on the hydrogen evolution reaction (HER), for which platinum is known to be a highly active catalyst due to its near-optimal hydrogen adsorption free energy [15, 21]. However, the high cost of platinum motivates ongoing research of alternative materials.

Information Society 2024, 7–11 October 2024, Ljubljana, Slovenia

https://doi.org/10.70314/is.2024.scai.8689

The mechanism of HER includes adsorbed hydrogen atom (H*) as an intermediate. Consequently, the adsorption energy of hydrogen is often used as a descriptor of the catalytic activity of the material [15, 21]. The most straightforward approach to obtain the adsorption energies is with density-functional theory (DFT) calculations. However, as the size of the system and the number of different adsorption sites increase, a full DFT analysis becomes computationally unfeasible. To address this challenge, machine-learning methods can be employed to predict hydrogen adsorption energies based on DFT results, enabling the investigation of more complex systems [10]. For example, bimetallic nanoparticles were investigated by Jäger et al. [8] and Zhang et al. investigated amorphous systems [20].

This contribution focuses on the use of machine learning for predicting hydrogen adsorption energies on platinum using electronic and geometric descriptors. Two separate datasets were constructed, one for surfaces and the other for nanoparticles. By employing supervised learning and attribute ranking, we built ML models, assessed their accuracy and analyzed whether the two datasets exhibit similar correlations. The idea of the contribution is illustrated in Figure 1.

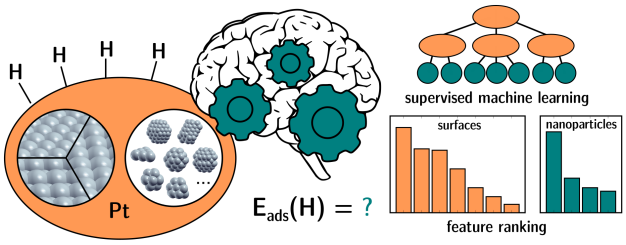

Figure 1: Supervised machine learning and feature ranking was performed for hydrogen adsorption energy on platinum catalysts modeled as surfaces and nanoparticles.

2 Materials and Methods

2.1 DFT Calculations and Datasets

We utilized DFT calculations to calculate hydrogen adsorption energies (a target variable for ML) and electronic descriptors for ML. We also utilized geometric descriptors. Two datasets were constructed, one for platinum nanoparticles and the other for platinum surfaces.

DFT calculations were performed with the Perdew-Burke-Ernzerhof (PBE) approximation [17], a plane-wave basis set, and PAW pseudopotentials [3]. Energy cutoffs were set to 50 and 575 Ry for wavefunctions and electron density,

respectively. Methfessel-Paxton smearing [12] of 0.02 eV was employed.

Pt(111), Pt(100), and Pt(110) surface slab models were constructed with the calculated lattice parameter of bulk Pt (3.97 Å). The models of Pt(111) and Pt(100) surfaces consist of 4 atomic layers, with the bottom layer fixed to bulk positions, while Pt(110) has 6 atomic layers with the bottom two layers fixed. To achieve a greater variety of adsorption sites, Pt(111) and Pt(100) were also modeled with a missing-row defect. All surface models are shown in Figure 2. Calculations accounted for the dipole correction and periodic images of slabs were separated by at least 15 Å of vacuum. Different sizes of surface supercells were used, and the k-point grid for (1×1) surface unit cells of Pt(111), Pt(100), and Pt(110) were 12×12×1, 11×11×1, and 11×8×1, respectively. For larger supercells, the number of k-points was adapted accordingly.

Calculations with nanoparticles were performed with the gamma k-point and Martyna-Tuckerman correction for isolated systems [11]. Nanoparticles were modeled with different shapes and sizes, consisting of 3 and up to 116 atoms. Their periodic images were separated by at least 15 Å of vacuum. All calculations were preformed with the Quantum ESPRESSO package [5].

The hydrogen adsorption energy was calculated as:

$$E_{\mathrm{ads}} = E_{\mathrm{H}*} - E_{*} - \frac{1}{2}E_{\mathrm{H_2}} \tag{1}$$

where $E_{\mathrm{H}*}$ is the calculated energy of optimized adsorption system, E_* is the energy of the standalone platinum system, and $E_{\mathrm{H_2}}$ is the energy of the hydrogen molecule. All performed calculations included only one adsorbed H atom per supercell or nanoparticle.

As an electronic descriptor, we used the d-band center, which is considered to be a good indicator of metal reactivity [6]. It was obtained through DFT calculations using the following equation:

$$\varepsilon_{\mathrm{d}} = \frac{\int_{-\infty}^{\infty} n_{\mathrm{d}}(E)EdE}{\int_{-\infty}^{\infty} n_{\mathrm{d}}(E)dE} \tag{2}$$

where E is the energy and n_{d} is the projected density of states on d-orbitals of the atoms forming the adsorption site.

For the geometric descriptors, we determined the average coordination number of Pt atoms forming the adsorption site, as well as the generalized coordination number (GCN) of the adsorption site [2], calculated as:

$$\mathrm{GCN}(i) = \sum_{j=1}^{N_i} \frac{\mathrm{CN}(j)}{\mathrm{CN}_{\mathrm{max}}} \tag{3}$$

where i denotes an atom or a group of atoms forming the adsorption site, N_i is the number of first nearest neighbors of i, which are denoted with j. $\mathrm{CN}(j)$ is the coordination number of atom j and $\mathrm{CN}_{\mathrm{max}}$ is the maximal coordination of a given site found in the bulk material.

In addition, the type of adsorption site was used as a descriptor. For extended surfaces, the coverage of H atoms, the surface area per H atom and surface type were also used for learning. For nanoparticles, some descriptors relevant to the size of nanoparticles were also utilized, in particular: the number of all atoms (N_{all}) in the nanoparticle, the number of surface atoms (N_{surf}), the maximal (r_{max}) and minimal (r_{min}) distances from the center of the nanoparticle to the surface atoms and the distance from the center of the nanoparticle to the adsorption site (r_{ads}). The datasets for surfaces and nanoparticles contained 46 and 85 data points, respectively.

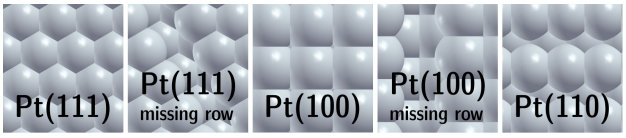

Figure 2: Models of extended surfaces used to calculate hydrogen adsorption energies.

2.2 Machine-Learning Methods

The prepared datasets were analyzed using the Weka software package [4]. The target value in both datasets is the hydrogen adsorption energy, making this a regression task. Supervised machine learning was employed to develop models for predicting the target value, which were evaluated by 10-fold cross-validation.

One of the used methods is linear regression, that computes the linear relationship between the target value and the descriptors. The relevant descriptors included in the equation were selected according to the M5 method [18]. This method iteratively removes descriptors with the smallest effect on the model until the error of the model no longer decreases.

We also used the random forest method [7, 1] with 100 trees of unlimited depth. With this method, multiple decision trees were constructed by selecting relevant features from a random subset of $\mathrm{int}(\log_2(m)+1)$ features, where m is the total number of features. The final values are the averages of the predictions from the individual trees.

To obtain an explainable ML model, we also built regression trees using the M5' method [18, 19]. In this method, trees are built by splitting the training sets according to attributes that maximize the standard deviation reduction. After the trees are constructed, they are pruned to avoid overfitting and smoothed to address discontinuities between the leaves. For our datasets, we used unpruned trees to prevent the formation of trees that are too small and give poor predictions. We also restricted tree branching to a minimum of 6 instances per leaf node for surfaces and 20 for nanoparticles to avoid overfitting the data and to ensure trees of sufficient size.

We also performed variable importance estimation and ranking for our selected descriptors with all data points used as a test set. To evaluate the importance of the descriptors with respect to hydrogen adsorption energy, we employed two methods: ReliefF [9] and correlation [16]. The ReliefF method is more sensitive to feature interactions and works by calculating the distances between training instances and identifying the 'nearest hit' and 'nearest miss'. It then adjusts the weights of the differing descriptors between the target and nearest instances. The correlation method evaluates the Pearson correlation coefficient [16]

between the features and the target variable, without accounting for interactions between features. Both methods provide scores ranging from −1 to 1, with 1 being the highest score. For the ReliefF method, a score of −1 indicates the worst importance score, while for Pearson's correlation, a score of −1 indicates anti-correlation, and 0 indicates no correlation.

3 Results and Discussion

3.1 Machine-Learning Models

Supervised machine learning was performed using linear regression, random forest, and M5' regression tree. The obtained Pearson's correlation coefficients and root mean squared errors (RMSE) between true and predicted values are shown in Table 1.

We can observe that not all ML models provide better RMSE values compared to those calculated with a simple arithmetic average, referred to as the default predictor. For surfaces, linear regression and random forest perform the best and yield similar results. The regression tree model performs the worst and has higher RMSE compared to the default predictor. For nanoparticles, all methods yield errors close to those of the default predictor and correlation coefficients bellow 0.5.

The obtained results indicate that with the selected descriptors, the hydrogen adsorption energies are more accurately predicted on surfaces, which are simpler as compared to nanoparticles. Surfaces have high symmetry and only a handful of different adsorption sites, while nanoparticles have different shapes and sizes, consist of different facets, and each nanoparticle has numerous different adsorption sites. This gives a huge variety of adsorption sites that can make the prediction of adsorption energies harder.

Considering the best models, the obtained adsorption energies have an error of ±0.13 eV for surfaces and ±0.22 eV for nanoparticles. Due to the exponential dependence of reaction rate and adsorption energy, even a small error in adsorption energy hugely affects the reaction rate. Hence, the models, particularly for nanoparticles, do not provide sufficiently accurate results for any practical use.

The selected ML models also provide insights into the relations between the considered features and the target variable. The linear regression model for nanoparticles includes only the d-band center and a factor for the hollow adsorption site, whereas the equation for surfaces is more complex. It includes adsorption site, surface type, and both coordination numbers. This indicates that for nanoparticles, the d-band center is the most relevant factor, while for surfaces, geometric factors exhibit greater predictive value. The regression-tree models shown in Figure 3 have lower accuracy and, consequently, are less reliable.

The ML models could be improved by expanding the dataset or by calculating additional descriptors. For surfaces, more data can be obtained through calculations on a wider variety of surface types and by accounting for different surface defects. However, expanding the dataset for nanoparticles is limited by their size, since DFT calculations for larger particles are computationally too demanding. Therefore, a larger number of different smaller particles can be tested instead. Using more sophisticated descriptors such as atom-centered symmetry functions, smooth overlap of atomic positions and many body tensor representation could also improve the results, but would require different sampling of adsorption structures. The use of transfer learning from pre-trained models based on chemical structures could also lead to significant improvements.

Table 1: Pearson's correlation coefficients (CC) and root mean squared errors (RMSE) in eV units for all three used ML methods. For comparison, RMSE of the default predictor is also given.

	surfaces		Nanoparticles	
	CC	RMSE	CC	RMSE
linear regression	0.71	0.13	0.38	0.22
random forest	0.69	0.13	0.34	0.22
M5' decision tree	0.49	0.19	0.34	0.22
default predictor	/	0.18	/	0.23

3.2 Feature Ranking

Feature ranking was performed for both surfaces and nanoparticles, with the results presented in Figure 4. The ReliefF and correlation importance criteria provide different rankings of features. For surfaces, the coordination number is identified as the most relevant descriptor, followed by the generalized coordination number. In contrast, for nanoparticles, the d-band center is the most important descriptor. Features describing the size of different nanoparticles show lower relevance for predictions. The most relevant features in both data sets describe the local environment of the adsorption site, indicating the local nature of adsorption.

The importance of the d-band center is already well-documented in the literature [14], as it correlates with the reactivity of metals. As seen from the graphs, the d-band center is not so strongly correlated with the hydrogen binding energy on surfaces. This can be attributed to the fact that on a perfectly flat surface, all surface atoms have the same d-band center. In contrast, on nanoparticles, the d-band center varies for each adsorption site because the atoms are not equivalent. Therefore, the d-band center is

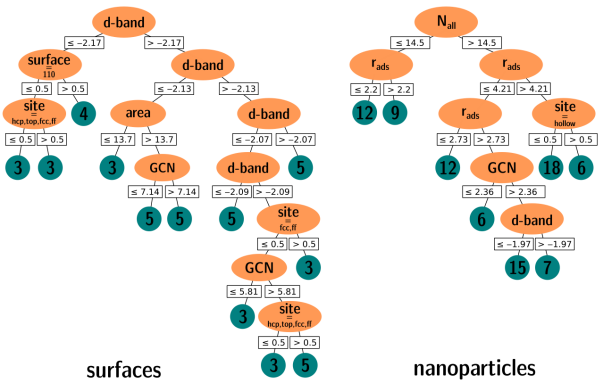

Figure 3: Schematic representation the obtained random-tree models for ideal surfaces and nanoparticles. Nodes are denoted with orange and the resulting classes are represented with turquoise circles and include the number of data points in the class.

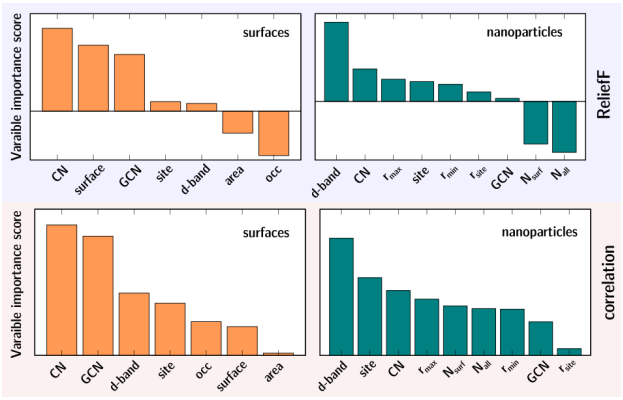

Figure 4: Variable importance scores calculated by the ReliefF and correlation criteria. Importance scores for correlation criteria are given as absolute values.

expected to be more relevant for nanoparticles. For the ranking based on correlation, the calculated factors for the d-band center are negative. This indicates that a lower d-band center corresponds to a higher adsorption energy and consequently a less reactive site, which is physically intuitive.

It is also interesting to note that the surface type descriptor is not very relevant according to correlation, yet it becomes the second most important feature when other descriptor are considered. This can be attributed to the fact that this descriptor has the same value for all adsorption sites on the same surface. However, when combined with other descriptors, it can give additional information, as similar adsorption sites on different surfaces can yield considerably different adsorption energies.

4 Conclusion

We applied different ML techniques to predict the adsorption energy of hydrogen on platinum surfaces and nanoparticles using simple geometric and electronic descriptors. Models for predicting adsorption energy on surfaces performed better, with the linear regression and random forest methods showing the highest correlation coefficient and accuracy. In contrast, predictions for nanoparticles yielded lower correlation coefficients and accuracy similar to the one calculated by a default predictor. Therefore, the models presented in this contribution do not provide accurate estimation of hydrogen adsorption energies. Utilizing more sophisticated descriptors and larger training data sets could enhance the performance of these models.

Differences between datasets are also evident in feature ranking. For surfaces, coordination numbers are the most relevant descriptors, while for nanoparticles, the d-band center shows the highest relevance. All these relevant descriptors are related to the local environment of the adsorption site, indicating that adsorption is a local phenomenon.

References

[1] Leo Breiman. 2001. Random forests. *Machine Learning*, 45, 1, (Oct. 2001), 5–32. DOI: 10.1023/A:1010933404324.

[2] Federico Calle-Vallejo, José I. Martínez, Juan M. García-Lastra, Philippe Sautet, and David Loffreda. 2014. Fast prediction of adsorption properties for platinum nanocatalysts with generalized coordination numbers. *Angew. Chem. Int. Ed.*, 53, 32, (Aug. 2014), 8316–8319. DOI: 10.1002/anie.201402958.

[3] Andrea Dal Corso. 2014. Pseudopotentials periodic table: From H to Pu. *Comput. Mater. Sci.*, 95, (Dec. 2014), 337–350. (files: H.pbe-kjpaw_psl.1.0.0.UPF, Pt.pbe-n-kjpaw_psl.1.0.0.UPF). DOI: 10.1016/j.commatsci.2014.07.043.

[4] Eibe Frank, Mark A. Hall, and Ian H. Witten. 2016. *The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques". Fourth Edition.* Morgan Kaufmann. https://ml.cms.waikato.ac.nz/weka/Witten_et_al_2016_appendix.pdf.

[5] Paolo Giannozzi et al. 2009. QUANTUM ESPRESSO: a modular and open-source software project for quantum simulations of materials. *J. Phys: Condens. Matter*, 21, 39, 395502. Code available from http://www.quantum-espresso.org/. DOI: 10.1088/0953-8984/21/39/395502.

[6] Bjørk Hammer and Jens K. Nørskov. 1995. Electronic factors determining the reactivity of metal surfaces. *Surf. Sci.*, 343, 3, (Dec. 1995), 211–220. DOI: 10.1016/0039-6028(96)80007-0.

[7] Tin Kam Ho. 1995. Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition*. Vol. 1. IEEE, 278–282.

[8] Marc O. J. Jäger, Yashasvi S. Ranawat, Filippo Federici Canova, Eiaki V. Morooka, and Adam S. Foster. 2020. Efficient machine-learning-aided screening of hydrogen adsorption on bimetallic nanoclusters. *ACS Comb. Sci.*, 22, 12, (Dec. 2020), 768–781. DOI: 10.1021/acscombsci.0c00102.

[9] Igor Kononenko, Edvard Šimec, and Marko Robnik-Šikonja. 1997. Overcoming the myopia of inductive learning algorithms with RELIEFF. *Applied Intelligence*, 7, 1, (Jan. 1997), 39–55. DOI: 10.1023/A:1008280620621.

[10] Jin Li et al. 2023. Machine learning-assisted low-dimensional electrocatalysts design for hydrogen evolution reaction. *Nano-Micro Lett.*, 15, 1, (Oct. 2023), 227–27. DOI: 10.1007/s40820-023-01192-5.

[11] Glenn J. Martyna and Mark E. Tuckerman. 1999. A reciprocal space based method for treating long range interactions in ab initio and force-field-based calculations in clusters. *J. Chem. Phys.*, 110, 6, (Feb. 1999), 2810–2821. DOI: 10.1063/1.477923.

[12] Michael Methfessel and Anthony Thomas Paxton. 1989. High-precision sampling for brillouin-zone integration in metals. *Phys. Rev. B*, 40, 6, (Aug. 1989), 3616–3621. DOI: 10.1103/PhysRevB.40.3616.

[13] Bishnupad Mohanty, Piyali Bhanja, and Bikash Kumar Jena. 2022. An overview on advances in design and development of materials for electrochemical generation of hydrogen and oxygen. *Mater. Today Energy*, 23, (Jan. 2022), 100902. DOI: 10.1016/j.mtener.2021.100902.

[14] Anders Nilsson, Lars G. M. Pettersson, Bjørk Hammer, Thomas Bligaard, Claus Hviid Christensen, and Jens K. Nørskov. 2005. The electronic structure effect in heterogeneous catalysis. *Catal. Lett.*, 100, 3, (Apr. 2005), 111–114. DOI: 10.1007/s10562-004-3434-9.

[15] Jens Kehlet Nørskov, Thomas Bligaard, Ashildur Logadottir, John R. Kitchin, Jingguang G. Chen, Stanislav Pandelov, and Ulrich Stimming. 2005. Trends in the exchange current for hydrogen evolution. *J. Electrochem. Soc.*, 152, 3, (Jan. 2005), J23. DOI: 10.1149/1.1856988.

[16] Karl Pearson. 1895. Vii. note on regression and inheritance in the case of two parents. *proceedings of the royal society of London*, 58, 347-352, 240–242.

[17] John P. Perdew, Kieron Burke, and Matthias Ernzerhof. 1996. Generalized gradient approximation made simple. *Phys. Rev. Lett.*, 77, 18, (Oct. 1996), 3865–3868. DOI: 10.1103/PhysRevLett.77.3865.

[18] John R et al. Quinlan. 1992. Learning with continuous classes. In *5th Australian joint conference on artificial intelligence*. Vol. 92. World Scientific, 343–348.

[19] Yong Wang and Ian H Witten. 1997. Inducing model trees for continuous classes. In *Proceedings of the ninth European conference on machine learning* number 1. Vol. 9. Citeseer, 128–137.

[20] Jiawei Zhang, Peijun Hu, and Haifeng Wang. 2020. Amorphous catalysis: machine learning driven high-throughput screening of superior active site for hydrogen evolution reaction. *J. Phys. Chem. C*, 124, 19, (May 2020), 10483–10494. DOI: 10.1021/acs.jpcc.0c00406.

[21] Jing Zhu, Liangsheng Hu, Pengxiang Zhao, Lawrence Yoon Suk Lee, and Kwok-Yin Wong. 2020. Recent advances in electrocatalytic hydrogen evolution using nanoparticles. *Chem. Rev.*, 120, 2, (Jan. 2020), 851–918. DOI: 10.1021/acs.chemrev.9b00248.