Determination of Adsorption Energies from DFT databases using Machine Learning techniques

Jose´ I. Arsuaga[a] and Ana I.Torres[a]

[a]*Instituto de Ingeniería Química, Facultad de Ingeniería, Universidad de la República, J. Herrera y Reissig 565, Montevideo 11300,Uruguay*
aitorres@fing.edu.uy,+59827142714 ext. 18102

Abstract

This paper discusses the estimation of adsorption energies for reaction intermediates for a given metallic surface and molecule. Regression models are learned from DFT data available in the literature in a two step approach. First, metallic surfaces are characterized by a principal component analysis (PCA) followed by a suitable orthonormal rotation to find a set of species that can be used as descriptors for the metallic surface. Then, different machine learning techniques are considered for the regression using the previous descriptors for the metallic surface and molecular descriptors such as the number and type of bonds for the adsorbate. With the available data, CH_3, CO_2 and CH_2 were found to explain 93% of the total variance, thus were used as surface descriptors. Threeof the tested models were found to adjust similarly well to validation data.

Keywords: Adsorption energies, Machine Learning, Electrocatalysis

1. Introduction

Electrocatalytic reactions have recently gained a lot of attention as they can be powered by electric energy from renewable non-programmable sources to achieve zero or even negative carbon processes. In particular, reduction of CO_2 on metallic surfaces is a promising process to transform industrial CO_2 emissions into valuable fuels and products. However, electrocatalytic processes are characterized by their very low selectivity; for example in the case of CO_2 many C1 and C2 products including acids (e.g formic acid,acetic acid), alcohols (e.g methanol and ethanol), and light hydrocarbons (for example ethylene) have been experimentally reported.

It is clear that the properties of the metallic surface play a role on the selectivity. As examples, copper and copper alloys or copper oxides seem to favor the production of ethylene and methanol from CO_2 (Dinh et al. (2018), Wang et al. (2018)), and platinum the production of CH_4 (Umeda et al. (2020)). Yet, the reaction mechanisms are not completely understood, and the lack of understanding hinders reactor and process design. Reaction network generators can be used to build the possible reaction pathways; still, thermodynamical properties of the proposed adsorbates need to be known to infer which of the pathways are feasible. One option to compute the required thermodynamical properties is the use of density functional theory (DFT). However, this approach requires very specialized knowledge, is time consuming and resource intensive in terms of computational power. Thus, it may not be suitable if a large amount of DFT-derived data is needed for calculations in a particular application.

In this work, we have taken another approach to estimate the required thermodynamic properties (adsorption energies) which is based on machine learning using data from DFT that is already available in the literature. Figure 1 schematizes the procedure. The rationale is that adsorption

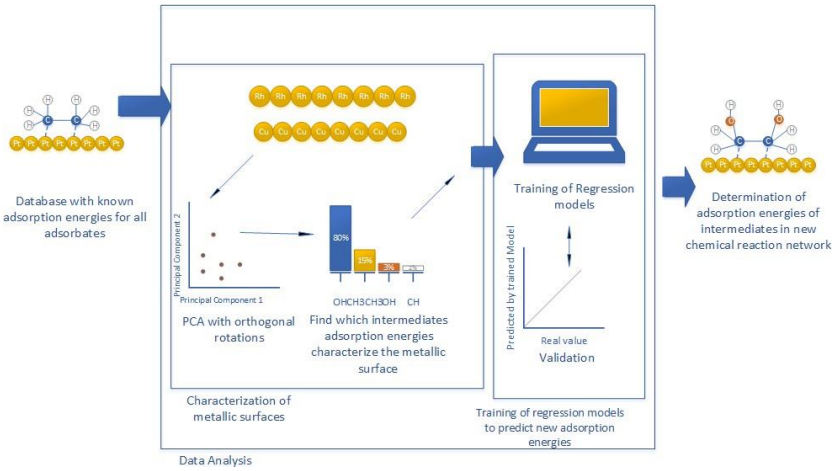

Figure 1: Schematic of the machine learning approach to find a model for estimating adsorption energies of adsorbates in metallic surfaces.

energies depend on properties of the adsorbates (molecules) and properties of the metallic surfaces. While it is fairly clear which properties should be considered as descriptors of the molecules (e.g. atoms, type and amount of bonds, etc.), it is not that clear which ones should be considered for the metallic surfaces. Therefore, the main hypothesis is that given a dataset containing the adsorption energies for several adsorbates in different metallic surfaces, it is possible to find a set of descriptors for the surfaces that are based on a subset of these energies. This idea is not new, Chowdhury et al. (2018) has proposed a similar approach using a dataset of 29 molecules and 8 metallic surfaces. In here, we have largely expanded the dataset including new surfaces and adsorbates, as a result, a different set of descriptors is obtained. On the basis of these new descriptors for the surfaces, and those of the molecules, a regression model was trained to predict adsorption energies of other molecules on the previous surfaces. The longer term objective is to use this regression model together with network generators to predict preferred reaction pathways for electrocatalytic reactions of different species on different surfaces.

2. Selection of descriptors of the metallic surface

As mentioned in the introduction, in order to find a correlation able to estimate the adsorption energies of any pair of adsorbate/metallic surface, a way of characterizing the surfaces needs to be found. A traditional DFT approach to describe a surface would need for example information on the atoms that compose the surface and their geometrical arrangement. These characteristics will affect the adsorption energies of all the adsorbates, although the effect on the possible adsorbates is different for each one, and depends mainly on the adsorbate itself. The idea then is to find those adsorbates whose energy of adsorption changes the most when the metallic surface changes. If this set of adsorbates is small, then using them as descriptors of the surface is a very practical and efficient way to characterize the surface, as instead of performing DFT calculations for all the components, we need DFT calculations for just a few.

2.1. PCA analysis of the energy of adsorption data

Principal component analysis (PCA) is a technique that given a data matrix $M_{N\times P}$, finds a new space of reduced dimensions which conserves a maximum amount of the variance of the original

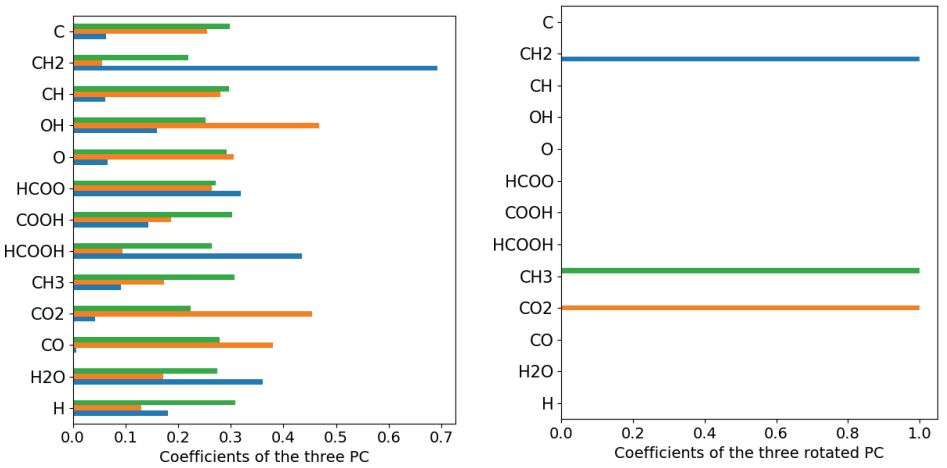

Figure 2: PCA analysis: left plain PCA; right PCA+varimax rotation. In both, green, blue and red are first, second and third component respectively.

data. This new space can be built by finding the eigenvectors of the covariance matrix (Σ_M) of the data in M. To establish which are the principal components of the data in M we look for the eigenvectors of Σ_M associated with the largest eigenvalues. In practice, those that added together make up for a certain threshold of the variance (usually 90-95%).

In our case, M is the matrix of adsorption energies (taken from DFT databases); each M_{np} entry is the adsorption energy for the *p-th* adsorbate on the *n th* metallic surface. Figure 2 showsthe results of applying PCA to a dataset built using adsorption energies reported in Plauck et al. (2016); Herron et al. (2012, 2013, 2014); Xu et al. (2018a); Bai et al. (2019); Ford et al. (2010); Mavrikakis et al. (2002); Ojeda et al. (2010); Scaranto and Mavrikakis (2016a); Singh et al. (2014);Scaranto and Mavrikakis (2016b); Ferrin et al. (2012); Greeley and Mavrikakis (2002); Ford et al.(2005); Chen et al. (2019); Xu et al. (2018b); Krekelberg et al. (2004); Hahn and Mavrikakis (2014); Grabow and Mavrikakis (2011); Gokhale et al. (2008); Li et al. (2016); Herron et al. (2014); Salciccioli et al. (2010, 2012); Lu et al. (2015, 2012); Schmidt and Thygesen (2018); Wellendorff et al. (2015). In here, it is important to mention that to apply PCA-techniques, M has to be complete, which means that only those adsorbates for which we found DFT data for *all* the surfaces of interest (Cu,Pt,Pd,Rh,Re,Ru,Ag,Au,Fe,Ir,Os,Co,Ni) were included.

Figure 2-left shows the three principal components as green (first component), orange (second component) and blue (third component) bars. Together these three are able to explain 93% of the variance of the original dataset (results obtained using the `scikit-learn` package in Python Pedregosa et al. (2011)). The adsorption energy for each adsorbate can then be expressed in terms of these three principal components; the absolute value of the weights that need to be applied are represented in the figure by the length of each bar.

2.2. PCA with varimax rotation

Unfortunately, the results in Fig.2-left are of little use as they are, as the principal components lack of physical interpretation. It would be desirable to have results where each component is clearly dominated by one or a few adsorbates. This is accomplished by finding a new set of orthogonal axis that represent a basis of the same space as the principal components, but in which the axis align better with some of the adsorbates. In this way, the coefficients of many of the adsorbates

Table 1: Summary of the performance of the regression

Method	KRR-poly	KRR-rbf	KSVR-poly	KSVR-rbf	CART	RF
RMSE	0.22	0.25	0.23	0.27	0.2	0.15
RMSE wo/H	0.12	0.13	0.12	0.13	0.18	0.11
sign mismatch	CO_2	-	CO_2	-	O	-

become zero, and those that are non-zero can be interpreted as the descriptors.

What was discussed was solved by Kaiser (1958) who proposed to find the new axes by solving the optimization problem in Eq. 1.

$$\max_{R} \sum_{j=1}^{j=K} \left(\frac{1}{N} \sum_{i=1}^{i=N} ((A R_{ij})^2)^2 - \alpha \frac{1}{N} (A' R_{ij})^2)^2 \right) \quad (1)$$

$$s.t : R^T R = I$$

In here, A is the original $P \times K$ eigenvector matrix ($K = 3$ as there are three principal componentsin our case study) and R the $K \times K$ rotation matrix. α is a parameter of the problem, if $\alpha = 1$ Eq.1 is the Varimax rotation.

Fig.2-right shows the results when applying the varimax rotation. These results indicates that CH_3, CO_2 and CH_2, as first (79% of the variance), second (9% of the variance) and third component (5% of the variance) respectively, can be used as descriptors of the metallic surface (results also obtained with `scikit-learn`). Notice that this is different from the results in Chowdhury et al. (2018) who obtained OH and $CHCHCO$ as descriptors. The difference lies in the expansion of the dataset to include data from different sources, as the exactly same results as in the Chowdhuryet al. (2018) are obtained when considering their database.

3. Learning a model to predict adsorption energies

After a suitable set of descriptors for the metallic surfaces is found, a regression problem that uses them and those of the adsorbates, can be formulated to learn a model for the energies of adsorption from data. As descriptors of the molecules we have considered the number and type of bonds in the adsorbate; we have also added facet and coverage as additional descriptors for the metallic surface when available. Notice now that completeness of the data is not required for this step, thus all available data can be used.

The following techniques were considered for learning the model: Kernel Ridge Regression (KRR, with polynomial and radial basis functions as kernels), Kernel Support Vector Regression (KSVR, with polynomial and radial basis functions as kernels), Classification and regression trees (CART) and Random Forest (RF). For the sake of space, we will not describe these methods, they are well explained in several references including `scikit-learn` documentation Pedregosa et al. (2011). In all cases, an 8-fold cross validation scheme was performed to define the set of possible values for the hyperparameters for each technique. Learning/ validation division of the dataset was 85/15% respectively. A summary of the results is in Table 1. As seen KRR-rbf KSVR-rbf and RF provided similar results in terms of RMSE over the validation set. Most importantly, they did not predict positive energies as negative nor the other way around. This was a problem that we observed when using polynomial kernels (wrong sign prediction for CO_2) and CART (wrong sign prediction for O). From a physical viewpoint, this is troublesome because it would imply that certain adsorbates cannot be adsorbed when in reality they can. In here, it has to be commented that we identified

some outliers for the energy of adsorption of *H* (2 out of 90 datapoints for H) from the reported DFT data; the table includes the RMSE results with and without considering these outliers

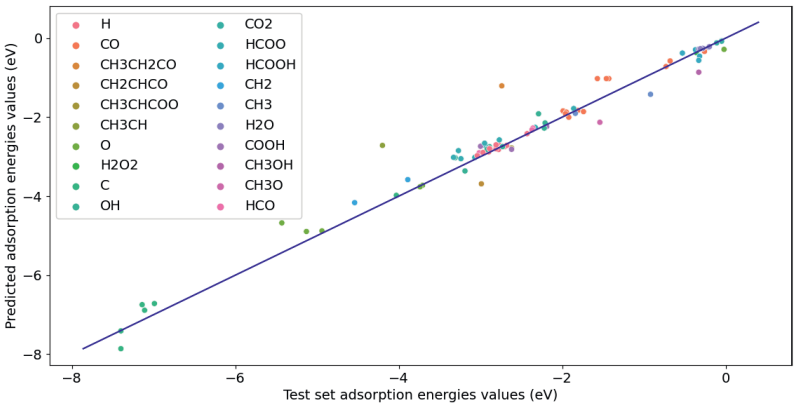

Figure 3: Results of the regression using Random Forest. Adjustment for $y = x$: $R^2 = 0.96$

Finally, Fig.3 presents the predicted vs DFT adsorption energies for data in the validation set. This type of plot showing a good regression is typical for all the techniques we have tested. This stresses the need of verifying that the chosen model correctly assigns the sign of the energy of adsorption for all data in the validation set, a point that may be overseen by just looking for the best RMSE and predicted vs real data fitting.

4. Conclusion

A large data set of DFT-based adsorption energies for different metallic surfaces and adsorbates was used to train a regression model. In a first step, PCA followed by Varimax rotation was found to be able to characterize the metallic surfaces using CH_3, CO_2 and CH_2 as principal components, with a loss of information less than 10% in terms of variance of the data. Six regression models based on either Kernel Ridge, Support Vector, CART or Random Forest were considered. All models were found to provide good estimations in terms of RMSE, but some had trouble in assigning a correct sign to those adsorbates whose energy of adsorption was close to zero. Those that can correctly estimate the sign could be used together with reaction network generators to predict thermodynamically feasible pathways.

References

Y. Bai, D. Kirvassilis, L. Xu, M. Mavrikakis, 2019. Atomic and molecular adsorption on ni(111). Surface Science 679, 240–253.

B. W. J. Chen, D. Kirvassilis, Y. Bai, M. Mavrikakis, Apr 2019. Atomic and molecular adsorption on ag(111). The Journal of Physical Chemistry C 123 (13), 7551–7566.

A. J. Chowdhury, W. Yang, E. Walker, O. Mamun, A. Heyden, G. A. Terejanu, 2018. Prediction of adsorption energies for chemical species on metal catalyst surfaces using machine learning. The Journal of Physical Chemistry C 122 (49), 28142–28150.

C. T. Dinh, T. Burdyny, M. Kibria, A. Seifitokaldani, C. Gabardo, F. P. Garc´ıa de Arquer, A. Kiani, J. Edwards, P. Luna, O. Bushuyev, C. Zou, R. Quintero-Bermudez, Y. Pang, D. Sinton, E. Sargent, 05 2018. Co 2 electroreduction to ethylene via hydroxide-mediated copper catalysis at an abrupt interface. Science 360, 783–787.

P. Ferrin, S. Kandoi, A. U. Nilekar, M. Mavrikakis, 2012. Hydrogen adsorption, absorption and diffusion on and in transition metal surfaces: A dft study. Surface Science 606 (7), 679–689.

D. C. Ford, A. U. Nilekar, Y. Xu, M. Mavrikakis, 2010. Partial and complete reduction of o2 by hydrogen on transition metal surfaces. Surface Science 604 (19), 1565–1575.

D. C. Ford, Y. Xu, M. Mavrikakis, 2005. Atomic and molecular adsorption on pt(111). Surface Science 587 (3), 159–174.

A. A. Gokhale, J. A. Dumesic, M. Mavrikakis, Jan 2008. On the mechanism of low-temperature water gas shift reaction on copper. Journal of the American Chemical Society 130 (4), 1402–1414.

L. C. Grabow, M. Mavrikakis, Apr 2011. Mechanism of methanol synthesis on cu through co2 and co hydrogenation. ACS Catalysis 1 (4), 365–384.

J. Greeley, M. Mavrikakis, Jun 2002. A first-principles study of methanol decomposition on pt(111). Journal of the American Chemical Society 124 (24), 7193–7201.

K. Hahn, M. Mavrikakis, Feb 2014. Atomic and molecular adsorption on re(0001). Topics in Catalysis 57 (1), 54–68.

J. A. Herron, J. Scaranto, P. Ferrin, S. Li, M. Mavrikakis, Dec 2014. Trends in formic acid decomposition on model transition metal surfaces: A density functional theory study. ACS Catalysis 4 (12), 4434–4445.

J. A. Herron, S. Tonelli, M. Mavrikakis, 2012. Atomic and molecular adsorption on pd(111). Surface Science 606 (21), 1670–1679.

J. A. Herron, S. Tonelli, M. Mavrikakis, 2013. Atomic and molecular adsorption on ru(0001). Surface Science 614, 64–74.

H. F. Kaiser, 1958. The varimax criterion for analytic rotation in factor analysis. Psychometrika 23 (3), 187–200.

W. P. Krekelberg, J. Greeley, M. Mavrikakis, Jan 2004. Atomic and molecular adsorption on ir(111). The Journal of Physical Chemistry B 108 (3), 987–994.

S. Li, J. Scaranto, M. Mavrikakis, Oct 2016. On the structure sensitivity of formic acid decomposition on cu catalysts. Topics in Catalysis 59 (17), 1580–1588.

J. Lu, S. Behtash, A. Heyden, 2012. Theoretical investigation of the reaction mechanism of the decarboxylation and decarbonylation of propanoic acid on pd(111) model surfaces. The Journal of Physical Chemistry C 116 (27), 14328–14341.

J. Lu, M. Faheem, S. Behtash, A. Heyden, 2015. Theoretical investigation of the decarboxylation and decarbonylation mechanism of propanoic acid over a ru(0001) model surface. Journal of Catalysis 324, 14–24.

M. Mavrikakis, J. Rempel, J. Greeley, L. B. Hansen, J. K. Nørskov, 2002. Atomic and molecular adsorption on rh(111). The Journal of Chemical Physics 117 (14), 6737–6744.

M. Ojeda, R. Nabar, A. U. Nilekar, A. Ishikawa, M. Mavrikakis, E. Iglesia, 2010. Co activation pathways and the mechanism of fischer–tropsch synthesis. Journal of Catalysis 272 (2), 287–297.

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, E. Duchesnay, 2011. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research 12, 2825–2830.

A. Plauck, E. E. Stangland, J. A. Dumesic, M. Mavrikakis, 2016. Active sites and mechanisms for h2o2 decomposition over pd catalysts. Proceedings of the National Academy of Sciences 113 (14), E1973–E1982.

M. Salciccioli, Y. Chen, D. G. Vlachos, 2010. Density functional theory-derived group additivity and linear scaling methods for prediction of oxygenate stability on metal catalysts: Adsorption of open-ring alcohol and polyol dehydrogenation intermediates on pt-based metals. The Journal of Physical Chemistry C 114 (47), 20155–20166.

M. Salciccioli, S. M. Edie, D. G. Vlachos, 2012. Adsorption of acid, ester, and ether functional groups on pt: Fast prediction of thermochemical properties of adsorbed oxygenates via dft-based group additivity methods. The Journal of Physical Chemistry C 116 (2), 1873–1886.

J. Scaranto, M. Mavrikakis, 2016a. Density functional theory studies of hcooh decomposition on pd(111). Surface Science 650, 111–120, the surface science of heterogeneous catalysis: In honor of Robert J. Madix.

J. Scaranto, M. Mavrikakis, 2016b. Hcooh decomposition on pt(111): A dft study. Surface Science 648, 201–211, special issue dedicated to Gabor Somorjai's 80th birthday.

P. S. Schmidt, K. S. Thygesen, 2018. Benchmark database of transition metal surface and adsorption energies from many-body perturbation theory. The Journal of Physical Chemistry C 122 (8), 4381–4390.

S. Singh, S. Li, R. Carrasquillo-Flores, A. C. Alba-Rubio, J. A. Dumesic, M. Mavrikakis, 2014. Formic acid decomposition on au catalysts: Dft, microkinetic modeling, and reaction kinetics experiments. AIChE Journal 60 (4), 1303–1319.

M. Umeda, Y. Niitsuma, T. Horikawa, S. Matsuda, M. Osawa, 2020. Electrochemical reduction of co2 to methane on platinum catalysts without overpotentials: Strategies for improving conversion efficiency. ACS Applied Energy Materials 3 (1), 1119–1127.

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, T. F. Jaramillo, jul 2018. Electrochemical carbon monoxide reduction on polycrystalline copper: Effects of potential, pressure, and pH on selectivity toward multicarbon and oxygenated products. ACS Catalysis 8 (8), 7445–7454.

J. Wellendorff, T. L. Silbaugh, D. Garcia-Pintos, J. K. Nørskov, T. Bligaard, F. Studt, C. T. Campbell, 2015. A benchmark database for adsorption bond energies to transition metal surfaces and comparison to selected dft functionals. Surface Science 640, 36–44, reactivity Concepts at Surfaces: Coupling Theory with Experiment.

L. Xu, D. Kirvassilis, Y. Bai, M. Mavrikakis, 2018a. Atomic and molecular adsorption on fe(110). Surface Science 667, 54–65.

L. Xu, J. Lin, Y. Bai, M. Mavrikakis, Jun 2018b. Atomic and molecular adsorption on cu(111). Topics in Catalysis 61 (9), 736–750.