CHEMICAL ENGINEERING TRANSACTIONS

VOL. 94, 2022

A publication of
AIDIC
The Italian Association
of Chemical Engineering
Online at www.cetjournal.it

Guest Editors: Petar S. Varbanov, Yee Van Fan, Jiří J. Klemeš, Sandro Nižetić
Copyright © AIDIC Servizi S.r.l.
ISBN 978-88-95608-93-8 **ISSN** 2283-9216

DOI: 10.3303/CET2294122

A Machine Learning-Accelerated Density Functional Theory (ML-DFT) Approach for Predicting Atomic Adsorption Energies on Monometallic Transition Metal Surfaces for Electrocatalyst Screening

Jan Goran T. Tomacruz[a], Karl Ezra S. Pilario[b], Miguel Francisco M. Remolona[c], Allan Abraham B. Padama[d], Joey D. Ocon[a,*]

[a]Laboratory of Electrochemical Engineering, Department of Chemical Engineering, University of the Philippines Diliman, Quezon City 1101, Philippines
[b]Process Systems Engineering Laboratory, Department of Chemical Engineering, University of the Philippines Diliman, Quezon City 1101, Philippines
[c]Chemical Engineering Intelligence Learning Laboratory, Department of Chemical Engineering, University of the Philippines Diliman, Quezon City 1101, Philippines
[d]Materials Computations Group, Institute of Mathematical Sciences and Physics, University of the Philippines Los Baños, Los Baños 4031, Laguna, Philippines
jdocon@up.edu.ph

The global mission to reduce fossil fuel consumption has led to the escalating demand for electrochemical energy storage (EES) devices such as fuel cells and batteries. Computational techniques like Density Functional Theory (DFT) have recently been coupled with Machine Learning (ML) for high-throughput material screening and discovery. Transition metal surfaces are popular electrocatalyst candidates, but predictive ML regression models have only been applied to select metals such as Pt and Cu. Additionally, characterizing the contributions of each feature is challenging, especially on black-box models. In this work, regression models were trained to predict the adsorption energies of carbon, hydrogen, and oxygen on 27 fcc (111) monometallic surfaces and applied model-agnostic interpretation methods to evaluate feature importance. Over 200 adsorption energies on transition metal surfaces were collected from Catalysis-hub.org, a surface reaction database. A dataset was constructed for each adsorbate, and was composed of ten surface atomic, surface electronic, and adsorbate properties collected from online databases and DFT calculations on adsorbate-free surfaces. Then, the fine-tuned random forest regression, Gaussian process regression, and artificial neural network models predicted atomic adsorption energies while permutation feature importance calculated feature contributions. All ML model accuracies were found to be competitive with those from literature, with Gaussian process regression reporting the lowest errors of the three models. Coordination number was also found to have the largest contributions for all models. ML-DFT methodologies such as this can be expanded to accommodate alloys and more adsorbates for a wider screening of potential EES materials.

1. Introduction

The undeniable threat of climate change continues to be present despite the ongoing COVID-19 pandemic. A major contributor in anthropogenic pollution is the CO_2 emissions from the energy industry, which was documented to be 32 x 10^9 t of CO_2 in 2020. This is due to the world's high reliance on fossil fuels and non-renewable resources to provide the increasing energy demand (BP, 2020). A growing field of research is focused on utilizing electrochemical reactions to decarbonize the industry, especially in the areas of carbon capture and utilization, fuel cell reactions, and water electrolysis (She et al., 2017).

Paper Received: 15 April 2022; Revised: 15 June 2022; Accepted: 28 June 2022
Please cite this article as: Tomacruz J.G.T., Pilario K.E.S., Remolona M.F.M., Padama A.A.B., Ocon J.D., 2022, A Machine Learning-Accelerated Density Functional Theory (ML-DFT) Approach for Predicting Atomic Adsorption Energies on Monometallic Transition Metal Surfaces for Electrocatalyst Screening, Chemical Engineering Transactions, 94, 733-738 DOI:10.3303/CET2294122

Efficient discovery of electrocatalytic materials with high activities is essential to provide competitive technologies in these applications. Quantum molecular modelling techniques such as Density Functional Theory (DFT) are typically used to calculate adsorption energies, which act as proxies for electrocatalytic activities (Greeley, 2016). However, using DFT alone to identify promising electrocatalysts from a large set of candidate materials is a time-consuming endeavour (Schleder et al., 2019). With the boom of the Information Age, Machine Learning (ML) methods have evolved to incorporate big data into material science. By augmenting DFT calculations with ML regression algorithms (ML-DFT), adsorption energies on multiple adsorbates and adsorption sites can be predicted using only DFT calculations on a smaller subset of candidate materials, as opposed to individually calculating each of their adsorption energies through DFT. Researchers such as Nayak et al. (2020) and Wang et al. (2020) used this ML-DFT approach to train regression models using material property databases on transition metal surfaces to predict their adsorption energies.

However, the use of ML has limitations in terms of interpretability. Although there are existing regression models that have attributes that can provide property insights, deep machine learning techniques such as neural networks tend to be treated as a black box. This leads to difficulty in the assessment of regression models aside from model accuracy and observing structure-property relationships from their results (Murdoch et al., 2019). It is believed that this research gap can be addressed by applying post-hoc interpretation methods such as Permutation Feature Importance (PFI) to determine the feature contributions in the predictive model. The application of these techniques on transition metal surfaces for atomic adsorption is a timely novelty, as these surfaces have extensive material data (Winther et al., 2019), and existing DFT-ML studies have yet to apply these methods on these surfaces. In this work, ML models were trained to predict the adsorption energies of single-atom adsorbates on 27 monometallic transition metal surfaces with minimal DFT calculations. This study also used PFI to identify the features with the highest contributions in each ML model. With this approach, candidate electrocatalyst materials can be screened more efficiently and with less computational expenses.

2. Methodology

The methodology of this work is summarized in Figure 1 below. Three surface reaction datasets were built for this study, with one for each adsorbate. Data were collected through online databases and DFT calculations. Afterward, ML regression models were trained to predict the adsorption energies of hydrogen, carbon, and oxygen. Finally, their model accuracies and feature importances were calculated.

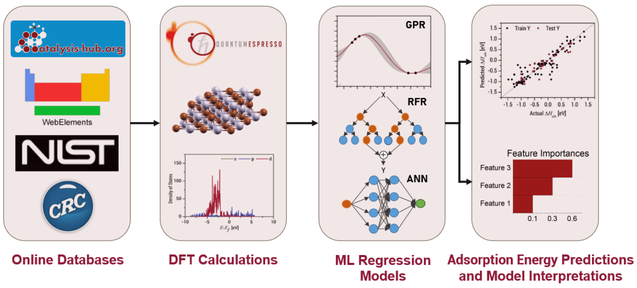

Figure 1. Schematic of methodology flow in this work

2.1 Surface reaction datasets

The monometallic transition metal surfaces were sourced from Catalysis-hub.org (Winther et al., 2019) and listed as the following elements in Figure 2a below.

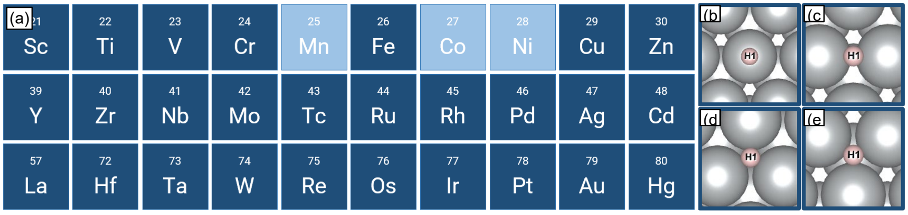

Figure 2. a) The d-block transition metals accounted for in this work are in dark blue; The different adsorption sites of hydrogen (H1) on the metallic surface: (b) top, (c) bridge, (d) hollow-fcc, and (e) hollow-hcp

The team of Mamun et al. (2019) conducted DFT calculations on these three-layered metallic slabs at the face-centred cubic (111) plane. One surface reaction dataset for each adsorbate was constructed for the study, where each data point represents each adsorption site on a monometallic surface, as seen in Figures 2b-e. From Catalysis-hub.org, there are 93 recorded energies for hydrogen adsorption, 74 for carbon adsorption, and 76 for oxygen adsorption. The features of the surface reaction dataset are listed in Table 1. These features were classified into three sets of properties: surface atomic, adsorbate, and surface electronic. The surface atomic properties were defined here as quantities that are constant given an atomic identity. On the other hand, the coordinate number acts as the sole adsorbate property, as described by its adsorption site.

Table 1. Proposed features for the surface reaction dataset

Property Set	Feature	Definition	Source of data
Surface Atomic	Pauling Electronegativity	Ability of atoms to attract electrons to itself	CRC [a]
	Ionization energy	Required energy to remove an electron	NIST [b]
	Atomic radius	Distance from nucleus to outermost orbital	WebElements [c]
	Sublimation energy	Transition energy from solid to gas phase	WebElements [c]
	Molar volume	Volume of one mole at standard conditions	WebElements [c]
	Lattice parameter	One dimension in a crystal cell structure	WebElements [c]
Adsorbate	Coordination number	Number of neighboring atoms to the adsorbate	Catalysis-hub [d]
Surface Electronic	d-Band center (ϵ_c)	Center of d-band energy states at the top layer	DFT calculations
	d-Band width (ϵ_w)	Width of d-band energy states at the top layer	DFT calculations
	d-Band filling (ϵ_f)	Filling of d-band energy states at the top layer	DFT calculations

[a] (Rumble, 2021), [b] (NIST, 2022), [c] (WebElements, 2022), [d] (Winther et al., 2019)

Table 2. Comparison of the three regression models and their hyperparameter tunings in this study

Model Name	RFR	GPR	ANN
ML algorithm family	Ensemble methods	Kernel methods	Neural network methods
Working principle [a] [b]	Hierarchical splitting of dataset for relationship mapping	Non-parametric prediction from probabilistic model	Construction of a multi-layer network of calculation nodes
Advantages [a] [b]	Averaging of trees reduce overfitting Robust to outliers	Can interpolate from small datasets Calculates uncertainty	Highly customizable architecture Can build complex models
Disadvantages [a] [b]	Black box behavior	Inefficient with high-dimensional datasets	Computationally expensive and black box behavior
Hyperparameter tuning technique	Grid search CV (10-folds)	Grid search CV (10-folds)	Randomized search CV (10-folds)
Number of combinations in hyperparameter tuning	768	56	60
Optimized hyperparameters	max_depth	alpha	optimizer
	max_features	kernel	activation function
	max_leaf_nodes	-	first hidden layer neurons
	min_weight_fraction_leaf	-	second hidden layer neurons
	n_estimators	-	

[a] (Theobald, 2017), [b] (Scikit-learn.org, 2022a)

2.2 DFT calculations

Electronic properties were the third property set, which described the d-band behaviour of the top layer of the surface. It is noteworthy that DFT calculations were conducted on only adsorbate-free surfaces, reducing the number of calculations by a factor of nine. The computational software QUANTUM ESPRESSO version 6.6 (Giannozzi et al., 2009) was used for this purpose and the Bayesian Error Estimation Functional with van der Waals correlation (BEEF-vdW) was the assigned DFT functional (Wellendorf et al., 2012). The cell was set to a k-point mesh of 10×10×1 and a vacuum space of 20 Å. All cells were also set to a 500-eV plane-wave cutoff, and a 5,000-eV density cutoff, except for the cell, which was set to a 1000-eV plane-wave cutoff, and a 10,000-eV density cutoff. Spin-unpolarized calculations were performed on the transition metals except for Fe and Co, which had spin-polarized calculations with starting magnetizations of 1.25 and 1. The partial density of states of d-states was obtained to calculate the surface electronic properties, as seen in Eqs (1-3) (Nørskov et al., 2014).

$$\epsilon_c = \frac{\int_{-\infty}^{\infty} n(E) \cdot n(E' - \epsilon_c)}{\int_{-\infty}^{\infty} n(E)} \qquad (1)$$

$$\epsilon_{\mathrm{w}} = \left(\frac{\int_{-\infty}^{\infty} \mathrm{n(E)} \cdot \mathrm{n}(\mathrm{E}' - \epsilon_{\mathrm{c}})^2}{\int_{-\infty}^{\infty} \mathrm{n(E)}} \right)^{\frac{1}{2}} \tag{2}$$

$$\epsilon_{\mathrm{f}} = \frac{\int_{-\infty}^{\mathrm{E_f}} \mathrm{n(E')}}{\int_{-\infty}^{\infty} \mathrm{n(E)}} \tag{3}$$

2.3 ML prediction models

The completed datasets were then simplified using principal component analysis (PCA), which reduces the number of input parameters of the regression models with minimal information loss (i.e., a cumulative variance of 95 %) (Jollife and Cadima, 2016). Three regression models were then used to predict adsorption energies: random forest regression (RFR), Gaussian process regression (GPR), and artificial neural networks (ANN). Their details as well as their strengths and weaknesses are briefly discussed in Table 2. The dataset was then divided into a ratio of 80 % training set and 20 % testing set, where the hyperparameters of the training set were optimized via grid search cross-validation or randomized search cross-validation with 10 folds, as listed also in Table 2.

2.4 Adsorption energy predictions and model interpretations

After model training, the predicted adsorption energies were plotted against those from the testing set. The model accuracies were represented through the root mean square error (RMSE). Using PFI, feature importances were calculated by observing the R^2 drop whenever a feature is shuffled. A higher R^2 indicates a higher importance (Scikit-learn.org, 2022b), as seen in Eq (4):

$$i_j = \mathrm{s} - \frac{1}{10} \sum_{k=1}^{10} s_{k,j} \tag{4}$$

The process of hyperparameter tuning, model fitting, and accuracy and feature importance evaluations are repeated 50 times to account for the randomized train-test splits in the dataset.

3. Results and Discussions

The dimensionality of each adsorbate dataset was reduced to five principal components using PCA. With an 80 % : 20 % train-test split and 50 trials, all regression models were able to accurately predict the adsorption energies of hydrogen, carbon, and oxygen, as seen in Figure 3. Out of the three models, GPR demonstrates the lowest RMSE for all adsorbates due to its probabilistic approach, which is advantageous in small-sized datasets. Regression analyses show that the average RMSEs of these models are also comparable to those from models constructed by Nayak, et al. (2020) – adsorption energy and Wang et al., (2020) – catalyst screening.

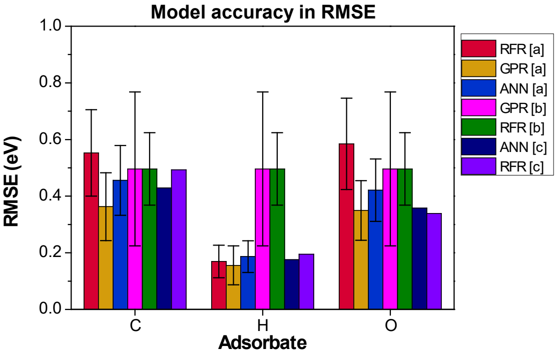

Figure 3. Box plots comparing the RMSE of the trained adsorption energy models of [a] this work to [b] Nayak et al., 2020, and [c] Wang et al., 2020. Note that there was no indicated standard deviation in [c].

Additionally, the calculated feature importance levels from PFI reveal that the coordination number of the adsorbate provides the highest contributions to the regression models in GPR, as illustrated in Figure 4. The features with the next highest average contributions, such as the d-band center and the lattice parameter, vary based on the adsorbate dataset. Although these findings can be interpreted as the confirmation of how the d-band model is an effective predictor of adsorbate energy (Nørskov et al., 2014), the vast difference in feature importance can be attributed to the lower correlation of the coordination number compared to the other two property sets. Because most of the properties in the surface reaction datasets are dependent on atomic identity and adsorbate-free DFT calculations, the values of all the features aside from coordination number are the same when comparing two surface-adsorbate systems with the same composition but different adsorption sites. A limitation of PFI is that it assumes that features are independent of each other (Scikit-learn, 2022a). The PCA results confirm this, as the number of input parameters was effectively halved with 95 % of the dataset information retained. As a result, PFI interprets surface atomic and surface electronic properties as of similar importance when evaluated. These findings were also consistent with RFR and ANN models.

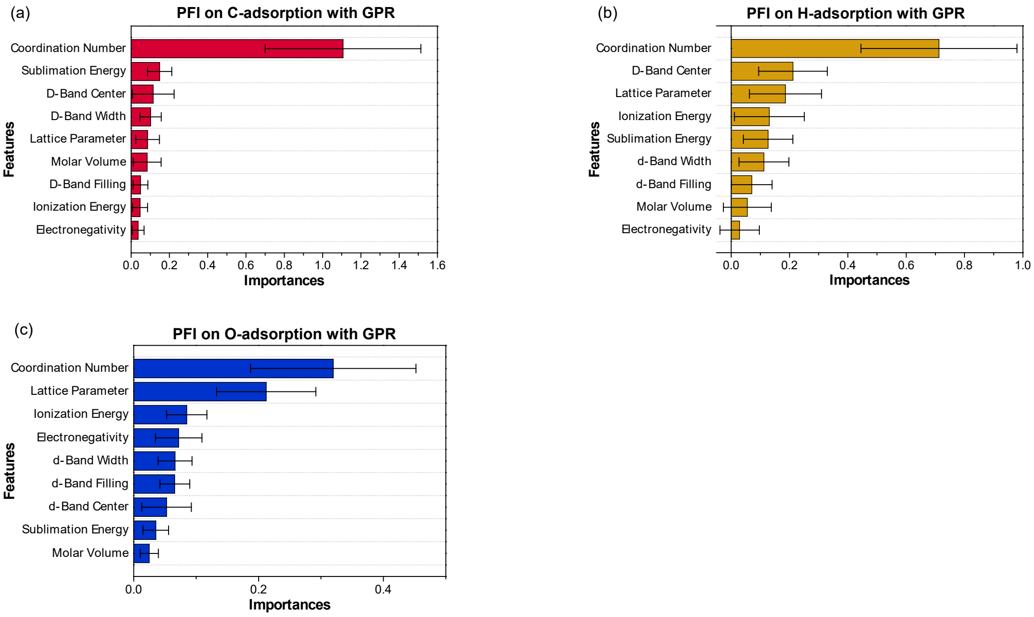

Figure 4. The feature importance levels for a) carbon, b) hydrogen, and c) oxygen adsorption using GPR

4. Conclusion

In this work, the adsorption energies of carbon, hydrogen, and oxygen were predicted using surface reaction datasets composed of surface atomic, surface electronic, and adsorbate properties. These datasets were constructed through data collection from online databases and DFT calculations on adsorbate-free monometallic transition metal surfaces. Then the RFR, GPR, and ANN models were tuned and trained 50 times, and PFI evaluated the contributions of each feature per model.

The data shows that all three regression models were as accurate as similarly constructed ML models in the literature. The small size of the datasets allowed Gaussian process regression to exercise the highest accuracy with RMSEs of 0.36±0.12 eV for carbon adsorption, 0.16±0.07 eV for hydrogen adsorption, and 0.35±0.11 eV for oxygen adsorption. Finally, PFI shows that coordination number is a key feature as it provides the highest contributions in model training. These findings serve as a successful proof of concept of ML-DFT methodologies for adsorption energy prediction, which can be expanded to alloys and mixtures such as High Entropy Alloys (HEAs) and MXenes. Future work can be focused on ML-DFT studies on these materials, provided that their surface reaction databases have enough DFT calculations. The material screening search space can also include more facets such as bcc (111) and hcp (0001), as some metals are naturally present in this form. Finally, grouped interpretation methods such as the calculation of Shapley values and Grouped Permutation Feature Importances can be applied to account for correlations inside property sets.

Nomenclature

ϵ_{c} – d-Band center
ϵ_{w} – d-Band width
ϵ_{f} – d-Band filling
E – energy, eV
$n(E)$ – density of states at E
E_{f} – fermi energy, eV
E' – energy subtracted by fermi energy, eV

i_j – feature importance of feature j, -
s – coefficient of determination (R^2), -
k – Iteration number in PFI, -
j – feature counter, -
$s_{k,j}$ – coefficient of determination (R^2) of kth shuffling of variable j
eV – electron-Volts, 1.60 x 10^{-19} Joules

Acknowledgements

J.G.T.T. would like to acknowledge the Engineering Research and Development for Technology (ERDT) Scholarship for their financial support. The authors would like to acknowledge The Commission on Higher Education – Philippine California Advanced Research Institutes (CHED-PCARI) through the CIPHER Project (IIID 2018-008). The authors would also like to thank the Computing and Archiving Research Environment (COARE) of the Department of Science and Technology – Advanced Science and Technology Institute (DOST-ASTI) for facilitating the computations required for this study.

References

BP, 2020. Statistical Review of World Energy Globally Consistent Data on World Energy Markets and Authoritative Publications in the Field of Energy, 66, <https://www.bp.com/content/dam/bp/business-sites/en/global/corporate/pdfs/energy-economics/statistical-review/bp-stats-review-2020-full-report.pdf>, accessed 26/06/2022.

Giannozzi P., Baroni S., Bonini N., Calandra M., Car R., Cavazzoni C., Ceresoli D., Chiarotti G.L., Cococcioni M., Dabo I., et al., 2009. QUANTUM ESPRESSO: A Modular and Open-Source Software Project for Quantum Simulations of Materials. Journal of Physics Condensed Matter, 21(39), 395502.

Greeley J., 2016. Theoretical Heterogeneous Catalysis: Scaling Relationships and Computational Catalyst Design. Annual Review of Chemical and Biomolecular Engineering, 7, 605–635.

Jollife I.T., Cadima J., 2016. Principal Component Analysis: A Review and Recent Developments. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 374, DOI: 10.1098/rsta.2015.0202.

Mamun O., Winther K.T., Boes J.R., Bligaard T., 2019. High-Throughput Calculations of Catalytic Properties of Bimetallic Alloy Surfaces. Scientific Data, 6, 1–9.

Murdoch W.J., Singh C., Kumbier K., Abbasi-Asl R., Yu B., 2019. Definitions, Methods, and Applications in Interpretable Machine Learning. Proceedings of the National Academy of Sciences of the United States of America, 116, 22071–22080.

Nayak S., Bhattacharjee S., Choi J.H., Lee S.C., 2020. Machine Learning and Scaling Laws for Prediction of Accurate Adsorption Energy. Journal of Physical Chemistry A, 124, 247–254.

NIST, 2022, NIST: Atomic Spectra Database - Ionization Energies <physics.nist.gov/PhysRefData/ASD/ionEnergy.html>, accessed 31/03/2022.

Nørskov J.K., Studt F., Abild-Pedersen F., Bligaard T., 2014. Fundamental Concepts in Heterogeneous Catalysis. John Wiley & Sons, Inc, Hoboken, NJ, USA.

Rumble J.M. (Ed.), 2021. CRC - Handbook of Chemistry and Physics 102nd Edition, <hbcp.chemnetbase.com/faces/contents/ContentsResults.xhtml>, accessed 31/03/2022.

Schleder G.R., Padilha A.C.M., Acosta C.M., Costa M., Fazzio A., 2019. From DFT to Machine Learning: Recent Approaches to Materials Science–a Review. Journal of Physics: Materials, 2, 032001.

Scikit-learn, 2022a. 1.7. Gaussian Processes. <https://scikit-learn.org/stable/modules/gaussian_process.html>, accessed 31/03/2022.

Scikit-learn, 2022b. 4.2. Permutation Feature Importance. <scikit-learn.org/stable/modules/permutation_importance.html>, accessed 31/03/2022.

She Z.W., Kibsgaard J., Dickens C.F., Chorkendorff I., Nørskov J.K., Jaramillo T.F., 2017. Combining Theory and Experiment in Electrocatalysis: Insights into Materials Design. Science, 355(6321), DOI: 10.1126/science.aad4998.

Theobald E., 2017. Machine Learning for Absolute Beginners: A Plain English Introduction 2nd Edition; Scatterplot Press, ISBN: 978-1549617218.

Wang T.R., Li J.C., Shu W., Hu S.L., Ouyang R.H., Li W.X., 2020. Machine-Learning Adsorption on Binary Alloy Surfaces for Catalyst Screening. Chinese Journal of Chemical Physics, 33, 703–711.

WebElements, 2022. The Periodic Table of the Elements by WebElements. <www.webelements.com/>, accessed 31/03/2022.

Wellendorff J., Lundgaard K.T., Møgelhøj A., Petzold V., Landis D.D., Nørskov J.K., Bligaard T., Jacobsen K.W., 2012. Density Functionals for Surface Science: Exchange-Correlation Model Development with Bayesian Error Estimation. Physical Review B - Condensed Matter and Materials Physics, 85, 235149.

Winther K.T., Hoffmann M.J., Boes J.R., Mamun O., Bajdich M., Bligaard T., 2019. Catalysis-Hub.Org, an Open Electronic Structure Database for Surface Reactions. Scientific Data, 6, 1–10.