Machine Learning-Based Prediction of Activation Energies for Chemical Reactions on Metal Surfaces

Daniel Hutton, Kari Cordes, Carine Michel, Florian Göltl

Version submitted to *Journal of Chemical Information and Modeling*
Published version can be found under the reference *J. Chem. Inf. Model.* 2023, 63, 19, 6006–6013
doi: 10.1021/acs.jcim.3c00740

Machine learning-based prediction of activation energies for chemical reactions on metal surfaces

Daniel J. Hutton[1,a], Kari E. Cordes[1,a], Carine Michel[2], Florian Göltl[1,*]

[1]*The University of Arizona, Department of Biosystems Engineering, 1177 E 4th St., 85719 Tucson, AZ, USA*

[2] *ENSL, CNRS, Laboratoire de Chimie UMR 5182, 46 Allée d'Italie, F69364 Lyon France*

[*]e-mail: fgoeltl@arizona.edu

[a] equal contribution

Abstract:

In computational surface catalysis, the calculation of activation energies of chemical reactions is expensive, which in many cases limits our ability to understand complex reaction networks. Here, we present a universal, machine learning-based approach for the prediction of activation energies for reactions of C, O, and H containing molecules on transition metal surfaces. We rely on generalized Bronsted-Evans-Polanyi relationships in combination with machine learning-based multiparameter regression techniques to train our model for reactions included in the University of Arizona Reaction database. In our best approach, we find a Mean Absolute Error for activation energies within our test set of 0.14 eV if the reaction energy is known and 0.19 eV if the reaction energy is unknown. We expect that this methodology will often replace the explicit calculation of activation energies within surface catalysis when exploring large reaction networks or screening catalysts for desirable properties in the future.

Introduction:

Obtaining accurate activation energies for reactions is the most challenging aspect in electronic structure calculations of molecular interactions with metal surfaces. The difficulties in this task are related to the presence of many local minima for surface adsorbates on metal surfaces and the concept that the activation energy for a reaction is the lowest saddle point on the potential energy surface for the transition between any of the potential local minima of reactants and products. Due to these challenges, obtaining accurate activation energies for even a single surface reaction requires experienced researchers and a computational cost of approximately 10000 CPUh. At the same time, an understanding of the performance or the

Version submitted to *Journal of Chemical Information and Modeling*
Published version can be found under the reference *J. Chem. Inf. Model.* 2023, 63, 19, 6006–6013
doi: 10.1021/acs.jcim.3c00740

computational design of heterogeneous catalysts requires the description of complex reaction networks and the calculation of a large number of activation energies[1].

Due to its excessive computational cost, the explicit calculation of all activation energies for complex reaction networks on multiple metal surfaces is in many cases not feasible and has been identified as limiting factor in applications such as the automated exploration of reaction networks[2–4]. Therefore, methods that rely on descriptors rather than electronic structure calculations for the prediction of activation energies have been developed. These methods are most commonly based on Brønsted-Evans-Polanyi relationships (BEPs)[5–8], which establish a linear correlation between the activation energy of a reaction and its reaction energy[9,10,19–22,11–18]. Additionally, improving BEPs by including additional parameters and utilizing machine learning has been attempted[23–26]. BEP-based methods have become a staple in activation energy predictions and can be highly accurate when studying one specific type of reaction on one specific surface[17]. However, BEPs need to be reparametrized for every reaction studied, which requires a significant number of activation energy calculations using electronic structure theory. To further reduce computational cost, a generally applicable approach to predicting activation energies for any surface reaction is highly desirable[27].

Here, we develop a machine learning-based approach for the prediction of activation energies for reactions of C, O, and H-containing molecules on transition metal surfaces. We first present the University of Arizona Reaction (UAR) database, a database of surface reactions based on literature data. Subsequently, we generalize BEPs, by establishing a set of descriptors for each reaction and use it as input for machine learning based, multi-parameter regression methods to predict activation energies within the UAR database. Finally, we analyze the importance of different descriptor sets on the accuracy of the model.

Version submitted to *Journal of Chemical Information and Modeling*
Published version can be found under the reference *J. Chem. Inf. Model.* 2023, 63, 19, 6006–6013
doi: 10.1021/acs.jcim.3c00740

The University of Arizona Reaction Database

The first step in our work is to develop a database for surface reactions of C, O and H-containing molecules on monometallic late transition metal surfaces. We construct the University of Arizona Reaction database from literature data and include a total of 7161 reactions published in 79 different papers. All energies were converted to eV and forward and backward reactions were added to the database for each datapoint. All datapoints with negative activation energies or activation energies lower than the reaction energies were removed from the database. Furthermore, obvious typos leading to an incorrectly reported reaction stoichiometry were corrected wherever possible.

A statistical overview of the database is given in Figure 1. The datapoints were obtained using seven different electronic structure codes (with localized or plane wave basis sets, different flavors of pseudo-potentials, different computational setups, etc.) and six different density functionals. The included species range from monoatomic molecular fragments to molecules as large as benzalacetone. The reactions range over all possible bond breaking and formation reactions in C, O and H containing molecules, but also include adsorption and functional group transfers. All late transition metals are represented by at least 2% of the set.

Version submitted to *Journal of Chemical Information and Modeling*
Published version can be found under the reference *J. Chem. Inf. Model.* 2023, 63, 19, 6006–6013
doi: 10.1021/acs.jcim.3c00740

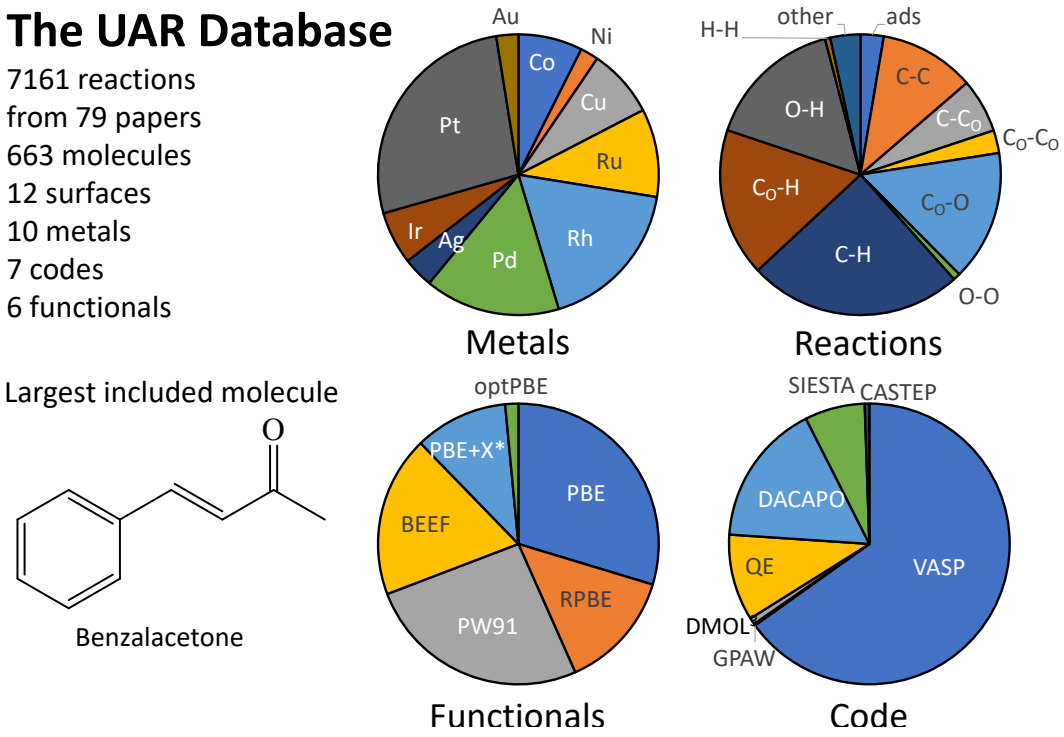


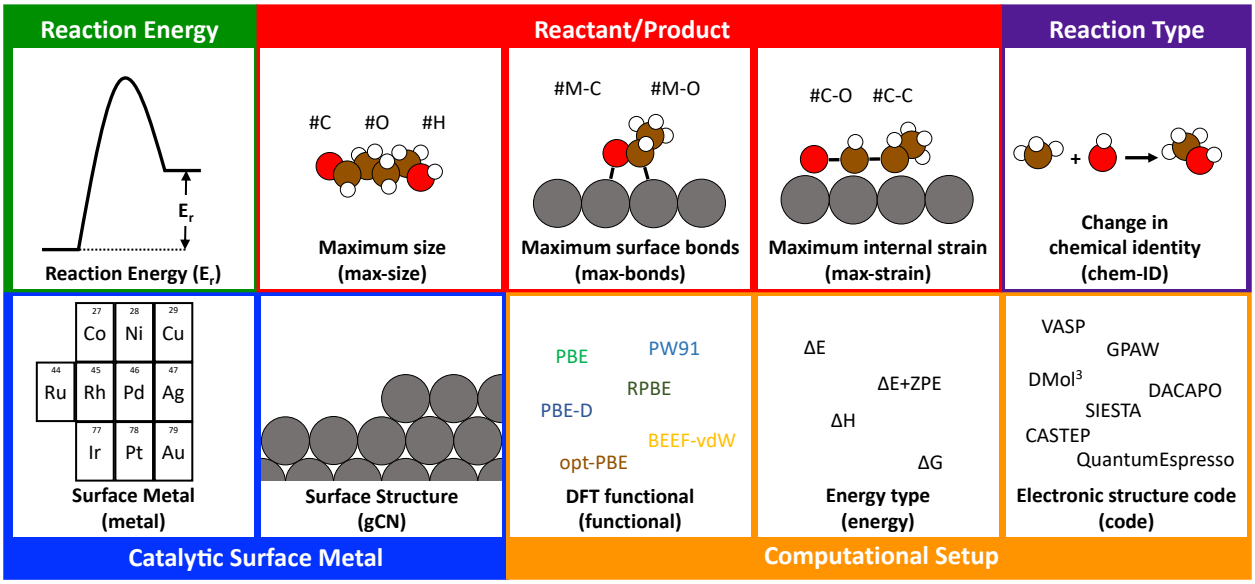
Figure 1: *This work relies on the University of Arizona Reaction (UAR) database. This database contains 7161 datapoints obtained from the computational, heterogeneous catalysis literature. Here, a statistical overview of the UAR database is presented.*

**PBE+X refers to PBE calculations with different dispersion corrections, such as +D2[28], +D3[29], +TS[30], or +dDsC[31].*

The Model:

In this contribution, we rely on generalized BEP relationships[26] to predict activation energies of reactions catalyzed by a metallic surface. In this approach, we correlate the activation energy to the reaction energy as well as descriptor sets related to the reaction type, the reactants and products, the catalytically active metal surface, and the computational setup. The different descriptor sets are schematically represented in Scheme 1.

Version submitted to *Journal of Chemical Information and Modeling*
Published version can be found under the reference *J. Chem. Inf. Model.* 2023, 63, 19, 6006–6013
doi: 10.1021/acs.jcim.3c00740



Scheme 1: *Schematic representation of the different descriptor sets included in the model for the prediction of activation energies. Details considering the mathematical definition of the descriptors used are given in the main text and the Supporting Information, section S2.*

In order to describe the reactants and products, we rely on scaling relationships[32,33]. Scaling relationships are based on bond order parameters, which describe the potential of each atom in a molecule to bind to the surface. Here, the adsorption energy (E_{ads}) of a molecule is expressed as[26]

$$E_{ads} = \sum_{X=C,C_O,O,H} \left(a^X p_X + n^X p_{n^X} + \sum_{Y=C,C_O,O,H} a^{X,Y} p_{X,Y} \right). \qquad (1)$$

In this expression, a^X is the sum over all bond order parameters of atoms of type X (X=C, C_O, O, H ; C stands for C atoms next to C or H, while C_O stands for C atoms adjacent to at least one O atom) in molecule A, n^X counts the number of atoms of type X in molecule A, and $a^{X,Y}$ is obtained as the sum over all products of bond order parameters of adjacent atoms of type X and Y and is related to the strain on the molecule caused by adjacent atoms bonding to the surface. p_X, p_{n^X}, and $p_{X,Y}$ are parameters that describe the contribution of a^X, n^X, and $a^{X,Y}$ to the adsorption strength. It is furthermore expected that the nature of the reactants and products impacts activation energies of surface reactions. The a^X, n^X, and $a^{X,Y}$ describe the chemical identity of

Version submitted to *Journal of Chemical Information and Modeling*
Published version can be found under the reference *J. Chem. Inf. Model.* 2023, 63, 19, 6006–6013
doi: 10.1021/acs.jcim.3c00740

any surface adsorbate, and to describe the reactants and products in our model, we include the maximum value of a^X, n^X, and $a^{X,Y}$ over all reactants/products as descriptors in our model. This choice allows to capture the nature of the reactants and products and keep the number of included descriptors in our model tractable.

Furthermore, inserting equation 1 for each reactant and product into the expression for the reaction energy (E_r) leads to

$$E_r = \sum_{X=C,C_O,O,H} \left(\Delta a^X p_X + \Delta n^X p_{n^X} + \sum_{Y=C,C_O,O,H} \Delta a^{X,Y} p_{X,Y} \right), \qquad (2)$$

where Δa^X, Δn^X, and $\Delta a^{X,Y}$ show the change in surface bonding, the number of atoms adsorbed to the surface and internal strain during a reaction. Since equation 1 describes the chemical identity of each reactant and product, equation 2 describes the change of chemical identity during a reaction. The type of reaction is described by the change of chemical identity of the reactants during a reaction. We therefore include all Δa^X, Δn^X, and $\Delta a^{X,Y}$ as descriptors for the reaction type to our model[26].

Additionally, we rely on the type of metal as given by the column and row in the periodic table and the surface structure as described by the generalized coordination number (gCN)[34,35] for surface atoms involved in the adsorption of reactants and products as descriptors for the catalytically active surface metal.

Furthermore, it is not clear to what degree the computational setup impacts results. Therefore, descriptors related to the computational setup are included, namely, the electronic structure code, the used density functional, and the energy type. Since these three descriptors are poorly described using numerical values, we use one-hot encoding to include them in our model. Overall, this leads to a model with 51 descriptors for each reaction. A summary of all descriptor sets is shown in scheme 1 and their mathematical definitions as well as descriptor values for all molecules studied are given in the Supporting Information, sections S1 and S2. Additionally, a practical example is given in the Supporting Information, section S3.

Subsequently we used Python's scikit-learn package[36] to parameterize three machine learning-based regression methods, namely Random Forest Regression (RFR), Support Vector

Version submitted to *Journal of Chemical Information and Modeling*
Published version can be found under the reference *J. Chem. Inf. Model.* 2023, 63, 19, 6006–6013
doi: 10.1021/acs.jcim.3c00740

Regression (SVR), and Gradient Boosted Regression (GBR). We parameterize these three methods using a training set, test set, and validation set split. Details about the parameterization and the parameters for each method are given in the Methods section.

using a test set size of one datapoint and repeat this procedure for all datapoints. This approach allows us to assign an unambiguous error to each datapoint, and the test set error is reported as an average over all single datapoint test set runs. To avoid overfitting, we parameterize the machine learning-based methods in such a way, that the mean absolute error (MAE) for the test set is minimized.

Results:

During parameterization, we perform a 80%/10%/10% training set/test set/validation set split and find mean absolute errors (MAE) of 0.07 eV/0.19 eV/0.19 eV, 0.11 eV/0.18 eV/ 0.18 eV, and <0.01 eV/0.16 eV/0.16 eV for the training set/test set/validation set for RFR, SVR, and GBR, respectively. However, in a realistic application our model will be trained using the full dataset and will be used to predict activation energies for reactions it was not trained with. If not indicated otherwise, we will focus on our analyses on an approach, where we use a test set size of one datapoint and repeat this procedure for all datapoints. This approach allows us to assign an unambiguous error to each datapoint, and the test set error (MAE_{test}, the MAE for the test) is reported as an average over all single datapoint test set runs. We will use MAE_{test}, the MAE for the test set, as a metric for assessing the quality of each method. Parity plots and error distributions for all methods, as well as the MAEs are shown in Figure 2 and predicted values for each datapoint are given in the Supporting Information. We find that GBR leads to the lowest MAE for the test set (MAE_{test}=0.15 eV), followed by SVR (MAE_{test}=0.17), and RFR (MAE_{test}=0.17 eV). At the same time, linear regression using the same descriptors leads to an MAE_{test} of 0.28 eV for the UAR dataset, which shows the importance of machine learning based regression. Additionally, the methodology described here significantly outperforms previous formulations of gBEPs[26], which lead to a MAE_{test} of 0.23 eV when applied to the UAR database. If not stated otherwise, we will use GBR, the best performing method, in further analysis.

Version submitted to *Journal of Chemical Information and Modeling*
Published version can be found under the reference *J. Chem. Inf. Model.* 2023, 63, 19, 6006–6013
doi: 10.1021/acs.jcim.3c00740

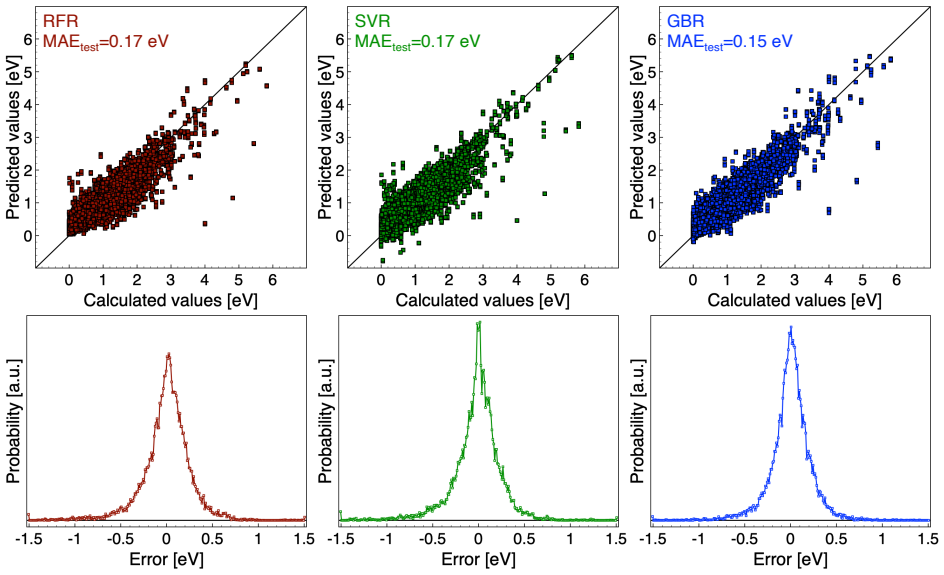


Figure 2: *Parity plots (top), error distributions (bottom), and MAE$_{test}$ values for a test set of one datapoint obtained via Random Forest Regression (RFR, brown), Support Vector Regression (SVR, green), and Gradient Boosted Regression (GBR, blue). Each datapoint is shown as a square, distributions are calculated using a grid of 0.01 eV and a line between datapoints is added to guide the eye.*

To better understand how the inclusion of different descriptor sets affects the performance of our model, we then remove one set of descriptors shown in Scheme 1 at a time and report MAE_{test} values for all methods (see Figure 3 (a)). As expected, removing descriptors for the reaction energy leads to the largest increase in MAE_{test}, even though values still lie below 0.19 eV. Despite the increase in MAE_{test}, it is fascinating to see that our model can predict activation energies for a reaction with an MAE_{test} of 0.19 eV without the prior knowledge of the reaction energy and therefore without performing any electronic structure calculations. Since this is a very important finding to accelerate the screening of large reaction networks, we will also include the model without E_r in our analysis for the remainder of this text (no E_r model). Besides E_r, only the nature of the metal (Pt vs. Rh for instance), the maximum number of surface bonds, and the change in chemical identity leads to an increase in errors significantly larger than 0.01 eV. However, the relatively small increase in MAE_{test} after removing information about the surface metal raises hope that this approach might also be applicable for metal alloy surfaces. All other

Version submitted to *Journal of Chemical Information and Modeling*
Published version can be found under the reference *J. Chem. Inf. Model.* 2023, 63, 19, 6006–6013
doi: 10.1021/acs.jcim.3c00740

descriptor sets on their own only have a small impact on the overall errors. The computational setup barely influences MAE values at all, which is reflected in the minimal impact of the reported energy type and the choice of density functional on overall MAE_{test} values. Furthermore, the lack of impact of electronic structure code on MAE_{test} values agrees with reports in the literature for comparisons between different electronic structure codes[37].

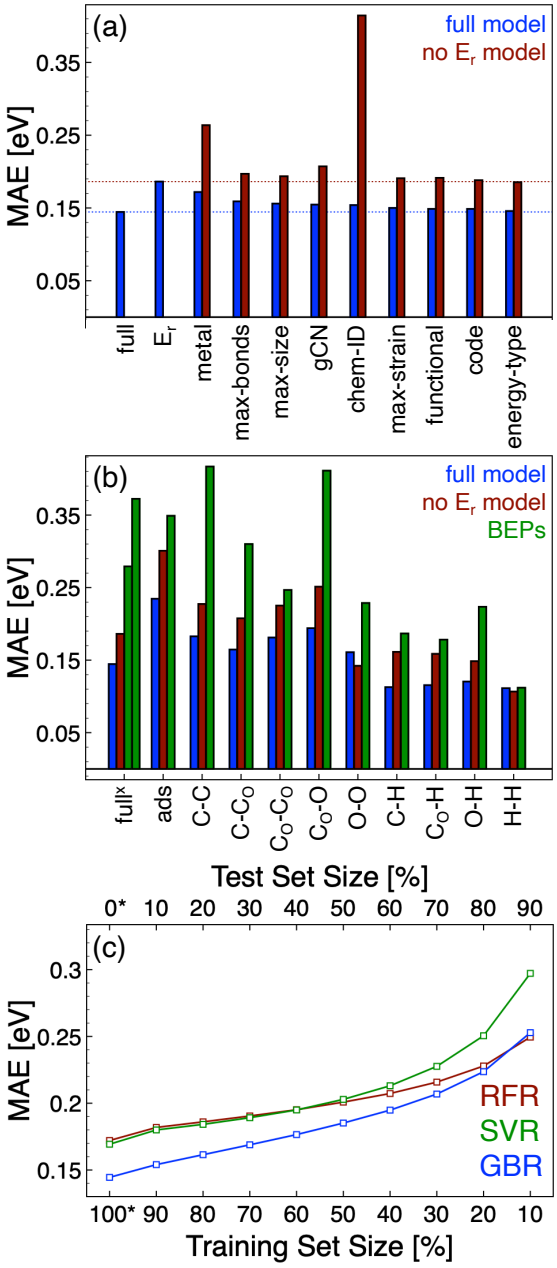


Figure 3: *(a) MAE_{test} values calculated using GBR after the removal of one set of parameters shown in Scheme 1. Blue bars refer to the full model, while brown bars are obtained for the*

Version submitted to *Journal of Chemical Information and Modeling*
Published version can be found under the reference *J. Chem. Inf. Model.* 2023, 63, 19, 6006–6013
doi: 10.1021/acs.jcim.3c00740

model where E_r was removed. Labels on the x-axis are defined in Scheme 1. (b) Comparison between MAE values for adsorption (ads) and bond-breaking reactions, using GBR ("full model", blue), GBR without knowledge of the reaction energy ("no E_r" model, brown) and BEP relationships ("BEPs", green) for different types of reactions. For the full model and the no E_r model, MAE_{test} values are reported, while for BEPs, MAE values for fitting the full database are given. (c) MAE_{test} values for different methods for different training set/test set sizes. Squares were explicitly calculated, and lines were added to guide the eye. Values for RFR are shown in brown, for SVR in green, for GBR in blue.

[x]For the full dataset, two BEP error values are reported, one calculated for the full dataset (right column) and one as weighted average over all reaction types (left column).

[*]The 100%/0% training/test set size corresponds to a test set size of 1 datapoint.

For the no E_r model, MAE_{test} increases to almost 0.40 eV when the descriptor set describing the change in chemical identity is removed (see brown bars in Figure 3 (a)). Otherwise, only information considering the surface metal significantly increases MAE_{test} values in this model.

We compared the performance of our two models (with and without E_r) with the performance of BEP relationships (see Figure 3 (b) and the supporting information Figure S2). It is well known that BEP relationships only work for a set of closely related reactions[15]. We therefore split our dataset into the following subsets for this analysis: C-C, C-C_O, C_O-C_O, C_O-O, O-O, C-H, C_O-H, O-H, and H-H bond-breaking and bond formation reactions, adsorption, and desorption reactions. We then categorize the remaining reactions, which consist mainly of H and O transfer reaction between adsorbates or molecular rearrangement reactions on the surface, as "other" reactions. Bond-breaking and adsorption reactions are shown in Figure 3 (b); bond formation reactions, desorption and other reactions are shown in the supporting information, Figure S2, parameters for BEPs and numerical error values are given in the supporting information, section S5, Figures S3-S6.

We see that our two models (with and without E_r) globally outperform typical BEP relationships with MAE_{test} values for the entire database of 0.15 eV , 0.19 eV and 0.37 eV respectively and only for H-H bond-breaking reactions, a very small subset of reactions, BEPs marginally outperform the no E_r model. Differences are most apparent for C-C and C_O-O bond-breaking and adsorption reactions. Such a behavior is expected, since these reaction categories

Version submitted to *Journal of Chemical Information and Modeling*
Published version can be found under the reference *J. Chem. Inf. Model.* 2023, 63, 19, 6006–6013
doi: 10.1021/acs.jcim.3c00740

include the most diverse sets of reactions[38]. Even though errors increase when the E_r unknown, this approach still outperforms BEP relationships.

Given these encouraging results, we then analyzed the impact of the training set/test set split on obtained results. We choose test set sizes ranging between 10% and 90% and performed 1000 different training set/test set splits and MAE_{test} values are reported as averages over all runs in Figure 3 (c). It is immediately apparent that MAE_{test} values drop with increased training set size. However, already for training set sizes of only 40% of the total data set, MAE_{test} values drop below 0.2 eV using GBR (see Figure 3 (c)). Similar trends are observed for our model without E_r (see supporting information Figure S7). However, overall error values are about 0.04 eV in errors higher, RFR moves closer to GBR, and SVR performs significantly worse than the other methods.

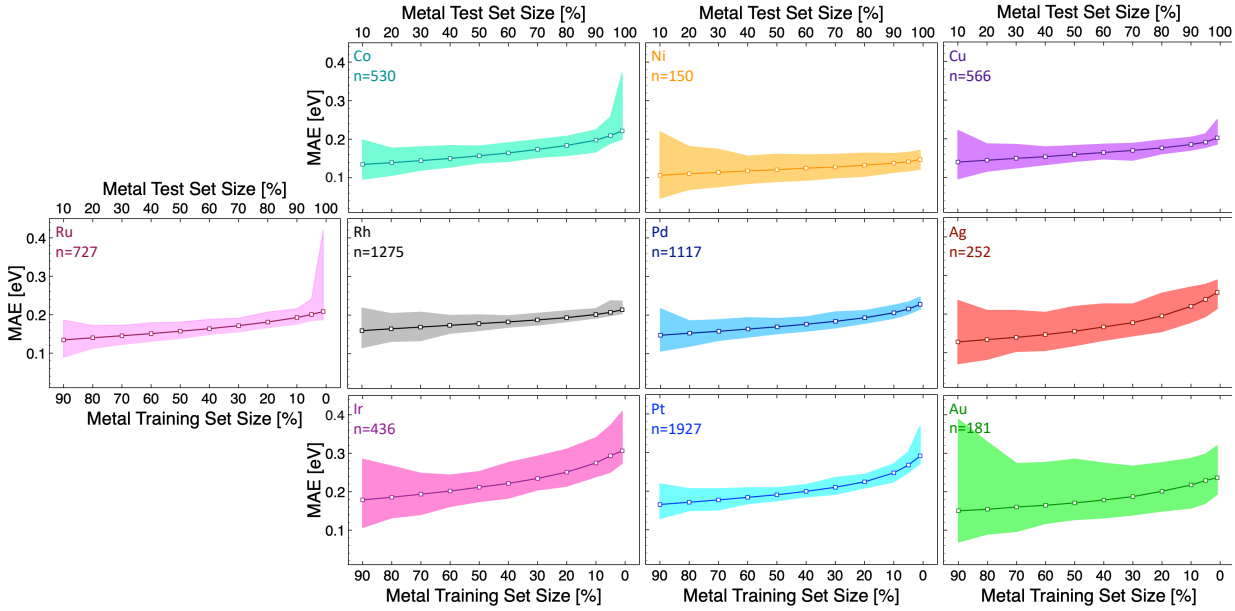


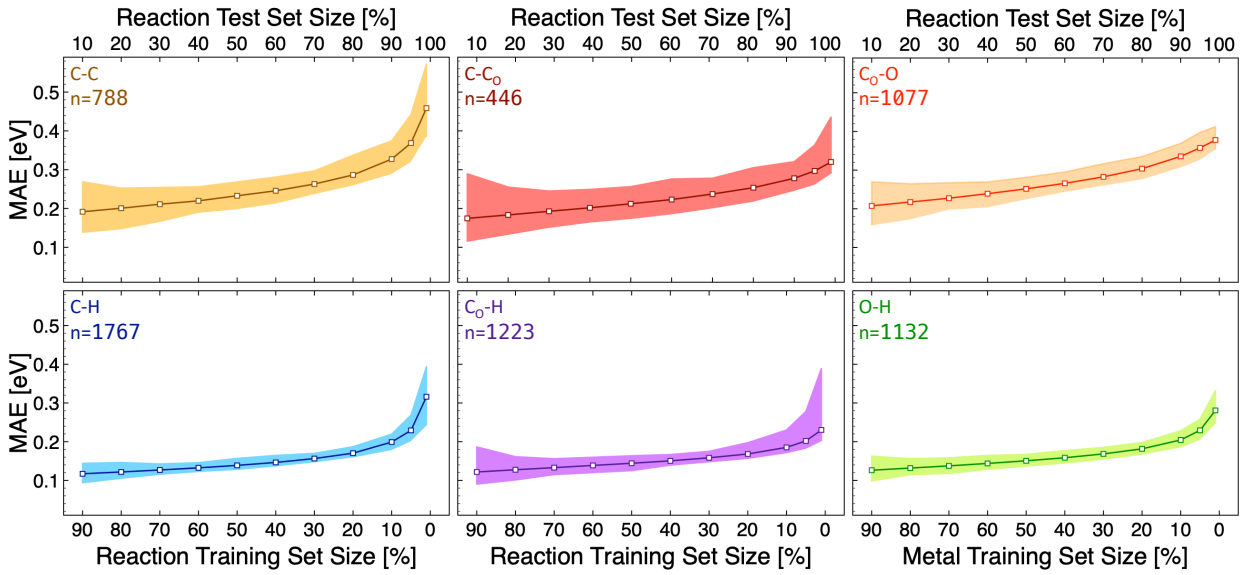
Figure 4: *MAE_{test} values for GBR using a fraction of the metal datapoints as the test set. Squares were explicitly calculated, and the line was added to guide the eye. The shaded area shows the variation in MAE_{test} values for different training- and test set splits. The legend indicates the surface metal and n gives the number of datapoints for each metal.*

So far, our analysis is focused on a limited set of metal surfaces, and it is not clear how easily our model can be extended to other metal surfaces. To better understand this aspect, we remove all reactions over one metal, from the training set and subsequently add in different fractions of the previously removed metal datapoints to the training set and calculate the obtained MAE_{test}

Version submitted to *Journal of Chemical Information and Modeling*
Published version can be found under the reference *J. Chem. Inf. Model.* 2023, 63, 19, 6006–6013
doi: 10.1021/acs.jcim.3c00740

values as an average over 1000 training/test set splits. Results are shown in Fig. 4. We find that for all surface metals, results converge quickly and already after including 20%-30% of the metal datapoints a significant drop in activation energies can be observed. Again, similar trends are observed for our model without knowledge of E_r (see supporting information Figure S8), with overall errors about 0.04 eV higher than for the full model. A thorough analysis for the larger metal datasets (Pt, Pd, Rh) implies that the inclusion of roughly 10%-20% of the datapoints (~150 datapoints) for a new metal might be sufficient to arrive at a reasonably high accuracy for the prediction of activation energies.

As a next step, we performed a similar analysis for different reaction types and focus on reactions where we have more than 300 datapoints available (see Figure 5). As indicated in Figure 3 (b), different reaction types show significant differences in errors. However, for all the reaction types we observe the largest reduction in errors after 10%-20% of the datapoints are included. Even though errors still decrease with increased data, a reasonable accuracy can already be achieved after 20%-30% of the data are included in the training set. These results are encouraging, since it indicates that it might be sufficient to include only a few hundred datapoints for each reaction type when adding a new element such as N or S in our model. Again, errors for a model without knowledge of E_r (see supporting information Figure S9) increase by about 0.04 eV.



Version submitted to *Journal of Chemical Information and Modeling*
Published version can be found under the reference *J. Chem. Inf. Model.* 2023, 63, 19, 6006–6013
doi: 10.1021/acs.jcim.3c00740

Figure 5: *MAE_{test} values for GBR using a fraction of the reaction datapoints as the test set. Squares were explicitly calculated, and the line was added to guide the eye. The shaded area shows the variation in MAE_{test} values for different training- and test set splits. The legend indicates the surface metal and n gives the number of datapoints for each metal.*

Discussion:

So far, we have demonstrated that gBEPs can lead to a prediction of activation energies with an MAE_{test} of 0.14 eV if the reaction energy is known and an MAE_{test} of 0.19 eV if the reaction energy is unknown. Our approach significantly outperforms other approaches reported in the literature[23,25,26], which have focused on significantly smaller databases. However, it is important to note that the reported errors are errors with respect to the DFT data used to train the model, which are typically assumed to show errors of ~0.20 eV[39]. As shown in Figure 3 (c), this increase in accuracy is linked to a significant increase in the number of training data available within the UAR database. The UAR database, however, is obtained from datamining the literature, which leads to a diverse set of data, which has been calculated using different methods and computational setups. While we attempted to capture some of these differences by including descriptors for the functional, electronic structure code, and the reported energy type, other technical aspects, such as the number of layers used in DFT calculations, the energy cut-offs, or convergence criteria were not considered in our analysis. Additionally, we cannot exclude that some of the data included in the UAR database was just misreported in the peer reviewed literature. These differences become apparent when analyzing data for the dehydrogenation of methane over Pt(111) (datapoints 1, 2173, 4203, 5826, and 6311 in the UAR database). Here, the calculated activation energies vary between 0.53 eV and 0.76 eV, while reaction energies vary between -0.30 eV and -0.10 eV. Based on all these factors, we would estimate that data in the UAR database contains noise on the order of 0.05 eV-0.1 eV.

The noise is a measure for errors that can be expected due to differences in the computational strategy in DFT calculations reported in the literature, but also leads to a lower bound for the MAE_{test} of any machine learning based model trained on the UAR dataset. At the same time, the noise in the UAR dataset might be too large to identify small effects on activation

Version submitted to *Journal of Chemical Information and Modeling*
Published version can be found under the reference *J. Chem. Inf. Model.* 2023, 63, 19, 6006–6013
doi: 10.1021/acs.jcim.3c00740

energy predictions, such as the choice of functional, the reported energy type, or the used electronic structure code. However, it might be possible to extract this information training gBEPs on a more uniform dataset, which is expected to be affected with less noise.

Throughout this work, we have chosen a formulation of gBEPs that relies on descriptors for the surface metal and the surface termination. This formulation is ideally suited to describe chemical reactions on monometallic surfaces and can be used in the automated exploration of reaction networks or screening of monometallic surfaces or nanoparticles. However, conventional screening studies of heterogeneous surface catalysts rely on the adsorption energies of molecules or molecular fragments[1,18,40], and can be applied irrespective of the type of surface metal or the surface termination. To explore whether the gBEPs developed in this work are applicable in such a scenario, we replaced the descriptors for the surface metal and the surface termination by the adsorption energies of C, O, H, and CO. All numerical values and details on how adsorption energies were calculated are given in the Supporting Information, section S7. Again, we find similar accuracy in our predictions with MAE_{test}=0.14 eV for the full model and MAE_{test}=0.19 eV for the no E_r model. This demonstrates that gBEPs are flexible in terms of their formulation and can be seamlessly integrated into conventional catalyst screening methodology. Given the accuracy of our approach, we believe that it can directly replace BEPs in catalyst screening studies. Additionally, it might be useful in a pre-screening step, to identify potential rate determining steps. However, it might not be suitable to construct a quantitatively accurate microkinetic model or kinetic Monte-Carlo model for complex surface reactions.

Conclusions:

Overall, we present a machine learning-based approach for the prediction of activation energies for chemical reactions on metal surfaces. The calculation of activation energies using the approach developed in this work is fast, accurate, and easy to apply. Initial tests raise hope that our approach can be extended beyond the single metal surfaces studied here. We expect that this methodology will in many cases, such as the automated exploration of reaction networks or the computational screening of the materials space, replace the explicit calculation

Version submitted to *Journal of Chemical Information and Modeling*
Published version can be found under the reference *J. Chem. Inf. Model.* 2023, 63, 19, 6006–6013
doi: 10.1021/acs.jcim.3c00740

of activation energies in surface catalysis. It will accelerate the understanding of the functioning of surface catalysts and will act as an enabling approach in the design of heterogeneous catalysts.

Methods:

In this work, three different machine-learning based multi-parameter regression techniques were applied, namely Random Forest Regression (RFR), Support Vector Regression (SVR), and Gradient Boosted Regression (GBR). Multiparameter regression analyses were performed using Python's scikit-learn module[36], a part of the Python scikit package. We parameterized these methods by first dividing the dataset into 10 random parts. For each of the part, we split the remaining data into training set (80% of the full data set) and test set (10% of the full data set). To avoid overfitting, we optimized hyperparameters in a way that minimized MAE_{test} over 100 different training/test set splits. The most common parameterization among the ten different datasets was chosen to report the final validation set error. After careful testing, RFR was performed for a total of 1600 trees. SVR was performed using the "rbf" kernel, with the "auto" setting for gamma, an epsilon of 0.01 and a penalty parameter of 160 for C. In GBR, we chose the least squares loss function, a learning rate of 0.1 and a total of 3000 boosting stages and a maximum depth of 7 for each individual boosting estimator. All other parameters were left as the default values.

Data and Software availability: Version 1.0 of the University of Arizona reaction database is accessible at https://github.com/floriangoeltl/The-University-of-Arizona-Reaction-Database . Error values for each datapoint are available in the Supporting Information. All regression analyses were performed using python's scikit-learn package, freely available at https://scikit-learn.org/stable/install.html .

Acknowledgements: Florian Göltl thanks the College of Agriculture and Life Science at the University of Arizona for support. The authors thank Reda Moussawi (University of Arizona) for his help in the construction of the UAR database. We acknowledge computational time at the

National Energy Research Scientific Computing Center (NERSC), a DOE Office of Science User Facility supported by the Office of Science of the U.S. Department of Energy, contract DE-AC02-05CH11231.

Supporting Information: The supporting information contains parameter definitions, a list of included parameters, a practical example, error values for bond-formation, BEP correlations, information on the importance of training/test set sizes for the no-E_r model, a list of all predicted values for the full database, and adsorption energies of C, O, H, and CO and how they were calculated.

References:

(1) Medford, A. J.; Vojvodic, A.; Hummelshøj, J. S.; Voss, J.; Abild-Pedersen, F.; Studt, F.; Bligaard, T.; Nilsson, A.; Nørskov, J. K. From the Sabatier Principle to a Predictive Theory of Transition-Metal Heterogeneous Catalysis. *J. Catal.* **2015**, *328*, 36–42.

(2) Steiner, M.; Reiher, M. Autonomous Reaction Network Exploration in Homogeneous and Heterogeneous Catalysis. *Top. Catal.* **2022**, *65* (1–4), 6–39.

(3) Markgraf, J. T.; Jung, H.; Scheurer, C.; Reuter, K. Exploring Catalytic Reaction Networks with Machine Learning. *Nat. Catal.* **2023**, *6*, 112–121.

(4) Wen, M.; Spotte-Smith, E. W. C.; Blau, S. M.; McDermott, M. J.; Krishnapriyan, A. S.; Persson, K. A. Chemical Reaction Networks and Opportunities for Machine Learning. *Nat. Comput. Sci.* **2023**, *3*, 12–24.

(5) Bell, R. P. The Theory of Reactions Involving Proton Transfers. *Proc. R. Soc. London, Ser. A* **1935**, *154*, 414–429.

(6) Evans, M. G.; Polanyi, M. On the Introduction of Thermodynamic Variables into Reaction Kinetics. *Trans. Faraday Soc.* **1937**, *33*, 448–452.

(7) Evans, M. G.; Polanyi, M. Inertia and Driving Force of Chemical Reactions. *Trans. Faraday Soc.* **1938**, *34*, 11–24.

(8) Bronsted, J. N. Acid and Basic Catalysis. *Chem. Rev.* **1928**, *5* (3), 231–338.

(9) Wang, S.; Temel, B.; Shen, J.; Jones, G.; Grabow, L. C.; Studt, F.; Bligaard, T.; Abild-Pedersen, F.; Christensen, C. H.; Nørskov, J. K. Universal Brønsted-Evans-Polanyi Relations for C-C, C-O, C-N, N-O, N-N, and O-O Dissociation Reactions. *Catal. Letters* **2011**, *141* (3), 370–373.

(10) Wang, S.; Petzold, V.; Tripkovic, V.; Kleis, J.; Howalt, J. G.; Skúlason, E.; Fernández, E. M.; Hvolbæk, B.; Jones, G.; Toftelund, A.; Falsig, H.; Björketun, M.; Studt, F.; Abild-Pedersen, F.; Rossmeisl, J.; Nørskov, J. K.; Bligaard, T. Universal Transition State Scaling Relations for (de)Hydrogenation over Transition Metals. *Phys. Chem. Chem. Phys.* **2011**, *13* (46), 20760–20765.

(11) Pallassana, V.; Neurock, M. Electronic Factors Governing Ethylene Hydrogenation and Dehydrogenation Activity of Pseudomorphic PdML/Re(0001), $Pd_{ML}Ru$(0001), Pd(111), and PdML/Au(111) Surfaces. *J. Catal.* **2000**, *191* (2), 301–317.

(12) Michaelides, A.; Liu, Z. P.; Zhang, C. J.; Alavi, A.; King, D. A.; Hu, P. Identification of General Linear Relationships between Activation Energies and Enthalpy Changes for Dissociation Reactions at Surfaces. *J. Am. Chem. Soc.* **2003**, *125* (13), 3704–3705.

(13) Loffreda, D.; Delbecq, F.; Vigné, F.; Sautet, P. Fast Prediction of Selectivity in Heterogeneous Catalysis from Extended Brønsted-Evans-Polanyi Relations: A Theoretical Insight. *Angew. Chemie* **2009**, *121* (47), 9140–9142.

(14) Sutton, J. E.; Vlachos, D. G. A Theoretical and Computational Analysis of Linear Free Energy Relations for the Estimation of Activation Energies. *ACS Catal.* **2012**, *2* (8), 1624–1634.

(15) van Santen, R. A.; Neurock, M.; Shetty, S. G. Reactivity Theory of Transition-Metal Surfaces: A Br{ø}nsted--Evans--Polanyi Linear Activation Energy: Free-Energy Analysis. *Chem. Rev.* **2010**, *110* (4), 2005–2048.

(16) Yu, L.; Vilella, L.; Abild-Pedersen, F. Generic Approach to Access Barriers in Dehydrogenation Reactions. *Commun. Chem.* **2018**, *1* (1), 2.

(17) Zaffran, J.; Michel, C.; Delbecq, F.; Sautet, P. Trade-Off between Accuracy and Universality in Linear Energy Relations for Alcohol Dehydrogenation on Transition Metals. *J. Phys. Chem. C* **2015**, *119* (23), 12988–12998.

(18) Medford, A. J.; Lausche, A. C.; Abild-Pedersen, F.; Temel, B.; Schjødt, N. C.; Nørskov, J. K.; Studt, F. Activity and Selectivity Trends in Synthesis Gas Conversion to Higher Alcohols. *Top. Catal.* **2014**, *57* (1–4), 135–142.

(19) Filot, I. A. W.; Broos, R. J. P.; Van Rijn, J. P. M.; Van Heugten, G. J. H. A.; Van Santen, R. A.; Hensen, E. J. M. First-Principles-Based Microkinetics Simulations of Synthesis Gas Conversion on a Stepped Rhodium Surface. *ACS Catal.* **2015**, *5* (9), 5453–5467.

(20) Nørskov, J. K.; Bligaard, T.; Logadottir, A.; Bahn, S.; Hansen, L. B.; Bollinger, M.; Bengaard, H.; Hammer, B.; Sljivancanin, Z.; Mavrikakis, M.; Xu, Y.; Dahl, S.; Jacobsen, C. J. H. Universality in Heterogeneous Catalysis. *J. Catal.* **2002**, *209* (2), 275–278.

(21) Vogt, C. Adsorbate Bond Number Dependency for σ- and Π-bonds in Linear Scaling Relationships. *J. Phys. Chem. C* **2023**.

(22) Latimer, A. A.; Kulkarni, A. R.; Aljama, H.; Montoya, J. H.; Yoo, J. S.; Tsai, C.; Abild-Pedersen, F.; Studt, F.; Nørskov, J. K. Understanding Trends in C-H Bond Activation in Heterogeneous Catalysis. *Nat. Mater.* **2017**, *16* (2), 225–229.

(23) Singh, A. R.; Rohr, B. A.; Gauthier, J. A.; Nørskov, J. K. Predicting Chemical Reaction Barriers with a Machine Learning Model. *Catal. Letters* **2019**, *149* (9), 2347–2354.

(24) Takahashi, K.; Miyazato, I. Rapid Estimation of Activation Energy in Heterogeneous Catalytic Reactions via Machine Learning. *J. Comput. Chem.* **2018**, *39* (28), 2405–2408.

(25) Abdelfatah, K.; Yang, W.; Vijay Solomon, R.; Rajbanshi, B.; Chowdhury, A.; Zare, M.; Kundu, S. K.; Yonge, A.; Heyden, A.; Terejanu, G. Prediction of Transition-State Energies of Hydrodeoxygenation Reactions on Transition-Metal Surfaces Based on Machine Learning. *J. Phys. Chem. C* **2019**, *123* (49), 29804–29810.

(26) Göltl, F.; Mavrikakis, M. Generalized Brønsted-Evans-Polanyi Relationships for Reactions on Metal Surfaces from Machine Learning. *ChemCatChem* **2022**, *14*, e202201108.

(27) Göltl, F. Three Grand Challenges for the Computational Design of Heterogeneous Catalysts. *J. Phys. Chem. C* **2022**, *126* (7), 3305–3313.

(28) Grimme, S. Semiempirical GGA-Type Density Functional Constructed with a Long-Range Dispersion Correction. *J. Comput. Chem.* **2006**, *27* (15), 1787–1799.

(29) Grimme, S.; Antony, J.; Ehrlich, S.; Krieg, H. A Consistent and Accurate *Ab Initio*

Parametrization of Density Functional Dispersion Correction (DFT-D) for the 94 Elements H-Pu. *J. Chem. Phys.* **2010**, *132* (15), 154104.

(30) Tkatchenko, A.; Distasio, R. A.; Car, R.; Scheffler, M. Accurate and Efficient Method for Many-Body van Der Waals Interactions. *Phys. Rev. Lett.* **2012**, *108* (23), 1–5.

(31) Steinmann, S. N.; Corminboeuf, C. A Generalized-Gradient Approximation Exchange Hole Model for Dispersion Coefficients. *J. Chem. Phys.* **2011**, *134* (4).

(32) Greeley, J. Theoretical Heterogeneous Catalysis: Scaling Relationships and Computational Catalyst Design. *Annu. Rev. Chem. Biomol. Eng.* **2016**, *7* (1), 605–635.

(33) Abild-Pedersen, F.; Greeley, J.; Studt, F.; Rossmeisl, J.; Munter, T. R.; Moses, P. G.; Skúlason, E.; Bligaard, T.; Nørskov, J. K. Scaling Properties of Adsorption Energies for Hydrogen-Containing Molecules on Transition-Metal Surfaces. *Phys. Rev. Lett.* **2007**, *99* (1), 4–7.

(34) Calle-Vallejo, F.; Martínez, J. I.; García-Lastra, J. M.; Sautet, P.; Loffreda, D. Angewandte Fast Prediction of Adsorption Properties for Platinum Nanocatalysts with Generalized Coordination Numbers. *Angew. Chemie - Int. Ed.* **2014**, *53*, 8316–8319.

(35) Calle-Vallejo, F.; Loffreda, D.; Koper, M. T. M.; Sautet, P. Introducing Structural Sensitivity into Adsorption–Energy Scaling Relations by Means of Coordination Numbers. *Nat. Chem.* **2015**, *7* (5), 403–410.

(36) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, E. Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.

(37) Lejaeghere, K.; Bihlmayer, G.; Björkman, T.; Blaha, P.; Blügel, S.; Blum, V.; Caliste, D.; Castelli, I. E.; Clark, S. J.; Dal Corso, A.; de Gironcoli, S.; Deutsch, T.; Dewhurst, J. K.; Di Marco, I.; Draxl, C.; Dułak, M.; Eriksson, O.; Flores-Livas, J. A.; Garrity, K. F.; Genovese, L.; Giannozzi, P.; Giantomassi, M.; Goedecker, S.; Gonze, X.; Grånäs, O.; Gross, E. K. U.; Gulans, A.; Gygi, F.; Hamann, D. R.; Hasnip, P. J.; Holzwarth, N. A. W.; Iuşan, D.; Jochym, D. B.; Jollet, F.; Jones, D.; Kresse, G.; Koepernik, K.; Küçükbenli, E.; Kvashnin, Y. O.; Locht, I. L. M.; Lubeck, S.; Marsman, M.; Marzari, N.; Nitzsche, U.; Nordström, L.; Ozaki, T.;

Paulatto, L.; Pickard, C. J.; Poelmans, W.; Probert, M. I. J.; Refson, K.; Richter, M.; Rignanese, G.-M.; Saha, S.; Scheffler, M.; Schlipf, M.; Schwarz, K.; Sharma, S.; Tavazza, F.; Thunström, P.; Tkatchenko, A.; Torrent, M.; Vanderbilt, D.; van Setten, M. J.; Van Speybroeck, V.; Wills, J. M.; Yates, J. R.; Zhang, G.-X.; Cottenier, S. Reproducibility in Density Functional Theory Calculations of Solids. *Science (80-.).* **2016**, *351* (6280), aad3000.

(38) Ferrin, P.; Simonetti, D.; Kandoi, S.; Kunkes, E.; Dumesic, J. A.; Norskov, J. K.; Mavrikakis, M. Modeling Ethanol Decomposition on Transition Metals: A Combined Application of Scaling and Bronsted-Evans-Polanyi Relations. *J Am Chem Soc* **2009**, *131* (16), 5809–5815.

(39) Wellendorff, J.; Silbaugh, T. L.; Garcia-Pintos, D.; Nørskov, J. K.; Bligaard, T.; Studt, F.; Campbell, C. T. A Benchmark Database for Adsorption Bond Energies to Transition Metal Surfaces and Comparison to Selected DFT Functionals. *Surf. Sci.* **2015**, *640*, 36–44.

(40) Vojvodic, A.; Nørskov, J. K. New Design Paradigm for Heterogeneous Catalysts. *Natl. Sci. Rev.* **2015**, *2* (2), 140–143.

Version submitted to *Journal of Chemical Information and Modeling*
Published version can be found under the reference *J. Chem. Inf. Model.* 2023, 63, 19, 6006–6013
doi: 10.1021/acs.jcim.3c00740

For Table of Contents Only