Enhanced Descriptor Identification and Mechanism Understanding for Catalytic Activity using Data-Driven Framework: Revealing the Importance of Interactions between Elementary Steps

Journal:	*Catalysis Science & Technology*
Manuscript ID	CY-ART-02-2022-000284.R2
Article Type:	Paper
Date Submitted by the Author:	21-Apr-2022
Complete List of Authors:	Liao, Wenjie; Stony Brook University The State University of New York Liu, Ping; Brookhaven National Lab, Chemistry

SCHOLARONE™
Manuscripts

Enhanced Descriptor Identification and Mechanism Understanding for Catalytic Activity using Data-Driven Framework: Revealing the Importance of Interactions between Elementary Steps

Wenjie Liao[a] and Ping Liu *[a,b]

Accurate identification of descriptors for catalytic activities has long been essential to the in-depth understanding of catalysis and recently to set the basis for catalyst screening. However, commonly used methods suffer from low accuracy in predictability. This study reports an enhanced approach to accurately identify the descriptors from a kinetic dataset using the machine learning (ML) surrogate model. The CO hydrogenation to methanol over Cu-based catalysts was taken as a case study. Our model captures not only the contribution from individual elementary step, but also the interaction between relevant steps within a reaction network, which was found to be essential for high accuracy. As a result, six effective descriptors are identified, which are accurate enough to ensure the trained gradient boosted regression (GBR) model for well prediction of the methanol turn-over-frequency (TOF) over metal (M)-doped Cu(111) model surfaces (M = Au, Cu, Pd, Pt, Ni). More importantly, going beyond the purely mathematic ML model, the catalytic role of each identified descriptor can be revealed by using the model-agnostic interpretation tools, which enhances the insight into the promoting effect of alloying. The trained GBR model outperforms the conventional derivative-based methods in terms of both the predictability and mechanism understanding. It opens alternative possibilities toward accurate descriptor-based rational catalyst optimization.

1. Introduction

Accurate identification of descriptors that can scale well with the catalytic activity and selectivity is of vital importance in catalysis. It can help to develop a quantitative insight into the nature of active site and underlying reaction mechanism[1, 2], to generate explanative and predictive design rules[3], and thus to guide the screening and rational optimization of catalysts[2, 4, 5]. Although the catalytic activity and selectivity can be obtained by kinetic modelling with the chemical master equation, the model usually consists of tens of elementary steps and hundreds of kinetic parameters[6]. Such a high dimensionality is typically solved by the kinetic Monte-Carlo (kMC) methods[7] or by mean-filed approximations[8], making the interpretation very difficult and nonintuitive. Thus, it is necessary to reduce the dimensionality of descriptor space.

There are two commonly used strategies to reduce the descriptor space and extract the effective descriptors. One is to consider the reaction energies or binding energies associated with elementary steps based on their certain correlations with corresponding activation energies, e.g., scaling relations, group additivity-based methods, and Brønsted-Evans-Polanyi relations[9-11]. Only those involved in the highly activated steps are considered as possible effective descriptors. The other is to evaluate the sensitivity of overall rate to the first-order derivative of activation energy for each elementary step. In this way, the rate-limiting steps can be identified, and the activation energies which strongly control the overall rate, are considered as effective descriptors. To do that, the degree of rate control (DRC)[12] or the sensitivity analysis (SA)[1, 13] is typically used. Compared to the former strategy, the latter is more independent that can work without scaling correlations. Besides, it is more straightforward, quantitative, and informative by providing more insight to the reaction mechanisms[14]. These methods have already achieved some success to describe relatively simple reactions over simplified surfaces, where the identified descriptors enable the effective scaling of catalytic activity and selectivity[15-19].

Despite the great advantages, the derivative-based approaches are only informative near the reference points where they are computed, and such localization greatly limits the predictability or accuracy[20]. Furthermore, such one-factor-at-a-time method is not able to capture the complex interaction between two different kinetic parameters such as activation barrier. In the case that the rate of one elementary step strongly depends on the other one, the interaction between the two corresponding activation barriers may need to be accounted, which can affect the overall rate.

To capture such kinetic complexity and enhance the accuracy in model prediction, variance-based global sensitivity analysis methods like Monte-Carlo estimation and Fourier amplitude sensitivity test (FAST) were developed[21-23] and applied to identify the kinetically important elementary steps in theoretical research of catalysis[24-27]. These methods provide a theoretical framework that can promote the full understanding of the model's sensitivity pattern; yet they also suffered from the high computational cost due to the curse of dimensionality and limited to describe relatively simple catalysis, especially when the analytical solution has not been available[28]. Instead, a

[a] Department of Chemistry, State University of New York at Stony Brook, Stony Brook, New York, 11794, United States.

[b] Chemistry Division, Brookhaven National Laboratory, Upton, New York, 11973, United States.

Email: pingliu3@bnl.gov

† Electronic Supplementary Information (ESI) available: Supporting figures and tables containing kMC, ML and DRC results; Reference dataset containing DFT calculated activation energies over M-doped Cu(111) and associated methanol TOF. See DOI: 10.1039/x0xx00000x

relatively cheaper emulator approach was developed which employed a surrogate model, such as high-dimensional model representation (HDMR)[29], to allow the exhaustive calculation of interactive term in global sensitivity analysis. Although the accuracy in prediction can be achieved using various mathematic models, it shows limited capability to provide understanding of kinetics particularly associated with the interactions between elementary steps. Besides, there has no generalized and standardized procedure for global sensitivity analysis[30], and so far the local derivative-based methods have been typically employed to analyse the reaction kinetics[13, 31].

Here, we developed an alternative approach that followed the well-established knowledge-extraction-from-data framework[6] to enhance the accuracy and efficiency of descriptor identification. The commonly used DRC method was also employed for comparison. This new approach takes advantage of both surrogate model in global sensitivity analysis and machine learning (ML). Compared with traditional surrogate model like orthogonal polynomial functions, ML has access to series of flexible non-parametric regression models that provide efficient data-driven function approximation[32, 33]. The ML approach has been already widely applied to the computational catalysis[34], including generating atomistic potentials[35, 36], predicting catalytic properties[15, 37-39], interpreting experimental data[40], and discovering new catalysts[41]. Despite many applications of ML in computational catalysis, the development is still at the early stage, where efforts have been devoted toward establishing explainable and trustworthy ML to promote the current catalytic understanding[34, 42-44]. ML has also been adopted to identify the likely reaction paths within a complex reaction network. Yet the predictability of ML models is limited due to the low accuracy in estimation of activation energy and determination of the rate-limiting steps[45]. More accurate modelling of surface is desirable to provide a strong basis for training.

To enhance the accuracy in modelling the reaction kinetics, the overall rate was considered as a function of the input kinetic parameters in our study, specifically activation energies of all elementary steps involved in a reaction network. Such hypothetic function was then described using a data-driven ML surrogate model together along with the model-agnostic interpretation tools[46], where both the first-order parameter or individual activation energy, and the normalized second-order or product of two activation energies, were considered as descriptors to represent the effect of individual elementary step and the interaction between relevant steps. To demonstrate our data-driven framework, methanol (CH_3OH) production from carbon monoxide (CO) hydrogenation over the doped Cu(111)-based surfaces was used as a case study, which is catalytically interesting due to the advantages of CH_3OH as industrial feedstock for other important chemicals and a renewable energy source[47-49]. The results reveal that our approach is considerably more accurate than the existing based on scaling relations and derivatives, being able to greatly enhance the accuracy in descriptor identification and rate prediction. More importantly, it can also allow the kinetic analysis by evaluating the surrogate model with a negligible addition of computational cost, so that a better mechanism understanding and ultimately design guidelines can be extracted.

2. Theoretical Methods

2.1 KMC Simulations

KMC simulations were conducted based on our previous theoretical study of CO hydrogenation over M-doped Cu(111) single-atom alloy surface[1] or M-Cu(111) in our notation (see SI for detail). They were carried out temperature of 600K to determine the corresponding TOFs of CH_3OH on exposure to 0.1 atm CO and 0.9 atm of H_2 along with a specific combination of activation energies in elementary steps. Each simulation was not considered converged until the statistical noise of is smaller than 0.05 molecule · site^{-1} · s^{-1}.

For all elementary steps involved in the reaction network, their activation barriers (Table S1) and site information (Table S2) were adopted from previous study[1]. In this case, the CO hydrogenation was described by eight elementary reactions (Table S1), including Hydrogen activation (R_0), CO hydrogenation to formyl (*CHO) and its reverse step (R_1 and R_2), *CHO hydrogenation to formaldehyde (*CH_2O, R_3), *CH_2O hydrogenation to methoxy, (*CH_3O, R_4), *CH_3O hydrogenation to *CH_3OH (R_5), *CO desorption (R_6), and *CH_3OH desorption (R_7). The recorded data included the activation energies of each elementary step ($E_n, n = 1 \sim 7$) except hydrogen activation (E_0) as input and the corresponding kMC-simulated CH_3OH TOFs as output. Note that, for simplicity the surface diffusion was not considered, and the reverse reactions were only considered for *CO hydrogenation, as the low stability and likely dehydrogenation of *CHO was reported previously on Cu catalysts[1] and was also observed in the current study as demonstrated below.

To simplify the kMC model and focus on the effect of activation barriers, several assumptions were adopted from previous successful practices in CO and CO_2 hydrogenation[1, 2, 49-51]. Firstly, hydrogen was considered to occur on Cu(111) facilely and the dissociated *H were readily available for reaction. Secondly, the lateral interaction between *CO was ignored due to its low coverage under reaction conditions. Lastly, the desorption of *CH_2O was ignored due to the high pressure of hydrogen (see SI for detail). This simplified kMC model offered an efficient way to establish the accurate and trusted ML framework. While such framework can be easily enriched with more complex kinetics including several competitive pathways running in parallel, distribution of multiple active sites and phases together with lateral interactions, which will be studied in the next step.

The overall reactions were modelled on a 128×128 surface matrix that resembled the Cu(111) and modified Cu(111) surface, where dopant metals account for 1/9 coverage (Figure S1). The rate constants for surface reactions were estimated by transition state theory[52] (eq. 1), where the k_B denotes the Boltzmann constant, T is reaction temperature, h is the Planck constant, E_a is activation barrier for this elementary step, $q_{vib}^{\neq}$ and q_{vib} stands for the vibrational quasi-partition functions in transition state and initial state. For the non-activated exothermic CO adsorption reaction that involves gas phase CO molecule, the rate constant was estimated according to Langmuir theory of adsorption[52, 53] (eq. 2), where p_x is the partial pressure of species of interest in the gas phase and m_x is its molecular weight, and A is the area of one binding site.

$$k_{reac} = \frac{k_B T}{h} \cdot \frac{q_{vib}^{\neq}}{q_{vib}} \cdot exp\left(\frac{-E_a}{k_B T}\right) \qquad (1)$$

$$k_{ads} = \frac{p_x \cdot A}{\sqrt{2\pi \cdot m_x \cdot k_B T}} \qquad (2)$$

2.2 Machine Learning

In the present study, all ML models were implemented with the scikit-learn code package[54]. These models were trained using data in Table S3. The accuracy was measured by root-mean-squared-error (RMSE) between the predicted values and true values.

ML models aim to learn the correlation between TOFs of CH_3OH and twenty-eight engineered descriptors based on activation barriers of elementary steps E_{1-7}. Wherein, seven are first-order descriptors that represent the contribution from each single elementary step to TOF, namely, E_n; while the rest are second-order descriptors were expressed by normalized product, *i.e.*, the harmonic mean (eq. 3), of activation barriers associated with two different elementary steps. Specifically, the harmonic mean of E_n and E_m ($n \neq m$) was employed to capture the non-local behavior[55] originated from interactions between elementary steps. As reported in previous ML studies[56-58], the harmonic mean type second-order descriptors are crucial to increase the predictability and interpretability. Catalytically, $\overline{E}_{n,m}$ can be considered as a weighted E_n by that of an interactive step m in the form of $\frac{2 \cdot E_m}{E_n + E_m}$ to represent the dependence of two relevant elementary steps, which cannot be captured by each individual.

$$\overline{\overline{E}}_{n,m} = \frac{2 \cdot E_n E_m}{E_n + E_m} \qquad (3)$$

Decision tree regression (DTR), Tree-based ensemble models, including random forest (RF) regression, gradient boosted regression (GBR), and the extra tree regression (ETR), and support vector regression (SVR) were employed to learn the correlation between descriptors and CH_3OH TOFs, while the least absolute shrinkage and selection operator (LASSO) regression was also included as a comparison. These models were trained on a spilt training set that contains 80% of data from dataset and were then verified on testing set that contains the rest 20% of data. Another widely applied method, namely the Gaussian process regression (GPR), was not considered since it's sensible to overfitting when dataset was small[59]. Hyperparameters to be defined in these models before training, such as learning rate, minimum loss reduction, maximum depth of a decision tree, and minimum sum of instance weight needed in a child, were optimized by an exhaustive grid search with 5-fold cross-validation (CV) as implemented by scikit-learn's GridSearchCV method[54].

After the model was trained, the effective descriptors were extracted to reduce model's dimensionality. Instead of using typical space projection methods such as principal component analysis (PCA), which suffered from low interpretability of principal components and loss of information[60], permutation feature importance score[61] was employed so that the original chemical meaning of descriptors can be preserved and thus leading to mechanic understanding[62-64]. Permutation feature importance score is defined as the decrease in model's accuracy when a single feature's value is randomly shuffled. This procedure breaks the original relationship between descriptors and the target, thus the drop in model's accuracy is indicative of how much the model depends on the feature. Descriptors with a higher value, in this case greater than 0.05, indicates a higher dependency and is regarded as more important than the others. The value of threshold was set at 0.05 so that most relevant descriptors were preserved in the selected GBR model. As shown in Figure 2, any further exclusion of descriptors would lead to a significant loss of predicting performance on the testing set. Also, it is worth noting that the perturbation-based feature permutation importance ranking only indicates the relevance of feature regarding the model's generalization error[65], thus it's model and feature dependent, and the specific threshold value is only valid in this case.

2.3 Degree of Rate Control

The degree of rate control (DRC) is a widely applied tool to evaluate how overall reaction rate changed with a small perturbation of kinetic parameters and thus determine the effective descriptors. In the present study, the DRC for each elementary step was determined via the procedure reported by Campbell *et al.*[14] (eq. 4). Here, r is TOF of CH_3OH, k_n represents the rate constant of elementary step n, and k_n is the corresponding equilibrium constant. Positive DRC indicates that facilitating step n leads to an increase in overall production rate, and the greater $\chi_n$means the influence of step n is more significant.

$$DRC_n = \chi_n = \frac{k_n}{r} \cdot \left(\frac{\partial r}{\partial k_n}\right)_{k_{m \neq n}, K_n} = \left(\frac{-\partial \ln r}{\partial \frac{E_n}{RT}}\right)_{k_{m \neq n}, K_n} \qquad (4)$$

Following the previous study[5], the DRC was also used to predict the overall rate of product formation on unknow catalyst based on the DRCs for a reference system (eq. 5). Here, r_0 denotes the production rate on reference catalyst 0, χ_n is the DRC for selected key steps on reference catalyst and E_n^0 is the corresponding activation barrier, while E_n^i is the activation barrier for the same elementary step on an unknown catalyst i, and r_i is the estimated production rate.

$$\ln r_i = \ln r_0 + \sum_n \chi_n \cdot \left(\frac{E_n^0 - E_n^i}{RT}\right) \qquad (5)$$

3. Results

3.1 Data-Driven Surrogate Model

The construction of data-driven surrogate model started with data generation and collection, where the kinetic behaviour of CO hydrogenation was sampled across the parameter space of reference system, which was Cu(111) in this case. During sampling, the reference value for each E_n was cited from that of Cu(111), which were reported in our previous study

(Table S1)[1] . Other E_n were generated here by shifting the corresponding reference value randomly and simultaneously in the range of -11.5 ~ 11.5 kCal·mol^{-1} to simulate the tuning of activation barriers induced by doping elements on Cu(111). The range of energy variation was determined in a way that most perturbations induced by metal-doping of Cu(111) considered here could be covered in our dataset. In this way the dataset generated can capture the large number of possibilities including the correlated cases for specific M-Cu(111) systems. If two parameters are identified as correlated thereafter, one of them could be selected as a representative while others could be presented as a function of it. Thus, the extracted relations between key descriptors and target TOF still stand, but only the model will become simpler with less descriptors. Based on each set of E_n ($n = 1{\sim}7$), the kMC simulation was performed to estimate the corresponding TOF of CH_3OH. At the end, it resulted in total 500 samples in the dataset (Table S3), which randomly filled in the parameter space and followed the uniform distribution (Figure S2). The samples with extremely low TOFs (< 0.5 molecule·site^{-1}·s^{-1}) were removed via the under-sampling procedure to avoid the imbalanced dataset problem[66].

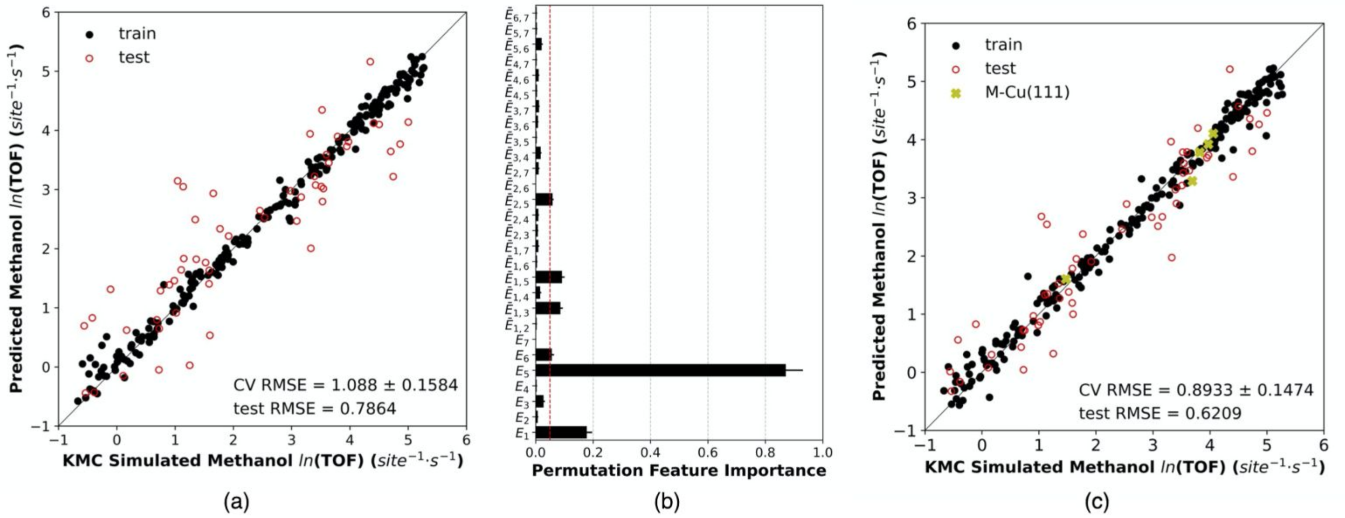

Figure 1. (a) kMC-simulated methanol TOF and on training set (black filled circle) and testing set (red hollow circle) and values predicted by preliminary ML model, (b) permutation feature importance score (black bar) and its standard deviation (solid black line) in the preliminary ML model, (c) kMC simulated methanol TOF on training set (black filled circle), testing set (red hollow circle), and metal-Cu(111) (M = Au, Cu, Pt, Pd, Ni, gold filled cross) and values predicted by refined ML model.

The data collection was followed by construction and training of ML surrogate model based on the twenty-eight descriptors and CH_3OH TOFs via a two-step process. The first step meant to reduce the dimensionality by identifying effective descriptors that had high permutation feature importance scores. Series of existing models, including, LASSO (Figure S3a), SVR (Figure S3b), DTR (Figure S3c), RFR (Figure S3d), ETR (Figure S3e), and GBR (Figure S3f) were trained on the training set with 5-fold cross-validation (CV), where the GBR model scored the best accuracy in terms of RMSE compared with the kMC simulated results (CV RMSE = 1.088±0.1584, test RMSE = 0.7864, Figure 1a). Although the ETR model also had a very competitive performance (CV RMSE = 1.175±0.1544, test RMSE = 0.8865), it suffered from a systematic deviation that tended to overestimate at low TOF and to underestimate at high TOF. Taking both the performance and systematic deviation into account, the GBR model was selected as the most predictive model in our study. Based on the GBR model, the score of permutation feature importance for each descriptor was calculated following the previous study[61] (Figure 1b). Six descriptors with high scores greater than 0.05 in this case, were considered as more significant term to control the TOFs of CH_3OH in the surrogate model than the others and thus the effective descriptors, including three first-order descriptors, *i.e.* activation energies of *CO hydrogenation (R_1) or E_1, *CH_3O hydrogenation (R_5) or E_5, *CO desorption (R_6) or E_6, and three normalized second-order descriptors, harmonic mean in this case, between E_1 and activation energy of *CHO hydrogenation (E_3) or $\overline{E}_{1,3}$, between E_1 and E_5 or $\overline{E}_{1,5}$, and between the activation energy of *CHO dehydrogenation (E_2) and E_5 or $\overline{E}_{2,5}$, were selected. Note that in the current model E_1 and E_2 were treated as independent parameters in the generation of dataset to cover maximum possibilities. However, for a Cu-based system, they are related as the forward and backward activation barriers for *CO hydrogenation, where the difference is determined by the corresponding reaction energy. As will be seen in the following, such dependence between E_1 and E_2 can be well described by the trained GBR model based on these effective descriptors.

In the second step, the GBR model was retrained using only the six effective descriptors (Figure 1c). Although the number of descriptors was greatly reduced from twenty-eight to six, the model's accuracy measured by RMSE was even slightly enhanced than the preliminary model in both training set (0.8933±0.1474) and testing set (0.6209), indicating that the removed descriptors were mostly non-informative. To confirm the effectiveness of the selected descriptors, we removed each of them one at a time from the retrained model separately. The results showed that the removal of each effective descriptor led to a substantial decrease in prediction accuracy in testing set (Figure 2). Interestingly, the accuracy losses after removing normalized second-order descriptors are even larger than removing the first-order descriptors, indicating that the inclusion of all six descriptors are necessary and the normalized second-order descriptors were more essential to achieve high predictability than the first-order descriptors.

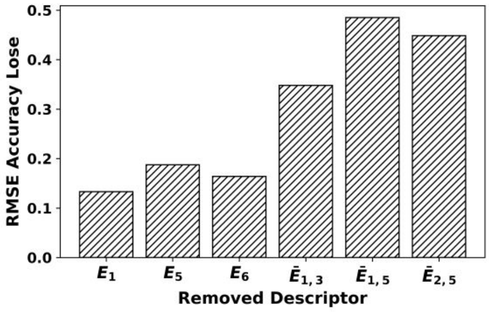

Figure 2. Accuracy loses measured by RMSE in testing set after the indicated descriptor was removed.

3.2 Role of Effective Descriptor

To gain understanding of the contributions from the six effective descriptors, the dependence of TOF of CH_3OH on each descriptor was plotted (Figure 3). The roles of three first-order descriptors are straightforward. The TOF increases with the decreasing activation energies for *CO hydrogenation (E_1, Figure 3a) and *CH_3O hydrogenation (E_5, Figure 3b) as well as the increasing activation energy for *CO desorption barrier (E_6, Figure 3c). It indicates that the CH_3OH production can be facilitated by accelerating the hydrogenation of *CO to *CHO and *CH_3O to *CH_3OH or hindering the *CO desorption, so that *CO can stay on the surface readily for hydrogenation. Here, as the highest energy among E_n for Cu(111) (Table S1), the first-order E_5 is likely the most determinative descriptor for the TOF and two normalized second-order descriptors out of three are also associated with E_5 (Figure 1b). Interestingly, while the TOF increases rapidly with the decreasing E_5, it levels off below 27 kCal·mol^{-1} (Figure 3b). That is, although R_5 is the most activated step on the reference system, Cu(111), and thus has a significant control on the TOF, the corresponding effect as descriptors can vary with the value changed. When E_5 is low enough, it becomes less effective to the TOF; instead, the associated second-order descriptors, $\overline{E}_{1,5}$, and $\overline{E}_{2,5}$ dominate as shown below.

Compared to the first-order descriptors, the situation for the normalized second-order descriptors is more complex. $\overline{E}_{1,5}$ starts to show the effectiveness as descriptor to TOF of CH_3OH at E_5 < 27 kCal·mol^{-1}; while at E_5 > 27 kCal·mol^{-1}, the corresponding effects are much less (Figure 3d). That is, when E_5 is low enough, the corresponding rate for *CH_3O hydrogenation (R_5) alone no longer affects the TOF significantly (Figure 3b); instead, it likely depends on the interaction with prior elementary steps. The *CO hydrogenation to *CHO (R_1), the second highest activated step in the reference Cu(111) system, is one of them. In this case, the occurrence of R_5 depends on that of R_1 and thus the amount of *CHO on the surface. Specifically, the high coverage of *CHO greatly facilitates the *CH_3O hydrogenation. In our model, such interaction between R_1 and R_5 was captured by the harmonic mean, $\overline{E}_{1,5}$. The high TOF of CH_3OH can only be achieved by reducing E_1 (< 18 kCal·mol^{-1}), while keeping E_5 low (< 27 kCal·mol^{-1}, Figure 3d). This transition from first-order descriptors E_5 to second-order descriptor $\overline{E}_{1,5}$ were clearly shown in the bivariant partial dependence plot between E_5 and $\overline{E}_{1,5}$ (Figure 3d). For E_5 greater than 27 kCal·mol^{-1}, the TOF of CH_3OH is almost independent from $\overline{E}_{1,5}$, whereas for E_5 smaller than 27 kCal·mol^{-1} TOF increased as the decrease of $\overline{E}_{1,5}$ and E_5 is almost ineffective. Although R_1 and R_5 are not in direct sequence, the interaction is built via several fast intermediate steps, R_{2-4} (Table S1). A variation in the reaction rate of R_1 can affect the amount of produced intermediates, which will be eventually passed to R_5 via the intermediate steps by varying the amount of reactant intermediates and thus the corresponding rate.

As the reverse step of R_1, the *CHO dehydrogenation to *CO (R_2) also interacts with R_5 at E_5 < 27 kCal·mol^{-1} and shows the significant effect on the TOF for CH3OH production (Figures 3e). A clear dependence between $\overline{E}_{2,5}$ and $\overline{E}_{1,5}$ was observed. Decrease in E_1 and thus $\overline{E}_{1,5}$ corresponds to the increase in E_2 (preferably > 14 kCal·mol^{-1}) and thus $\overline{E}_{2,5}$, which eventually promotes the TOF of CH_3OH production.

Similar situation was also observed between the activation energy of *CO hydrogenation (E_1) and *CHO hydrogenation (E_3). When E_1 is below the critical point (29 kCal·mol^{-1}), the corresponding step R_1 starts to interact with the sequential *CHO hydrogenation (R_3) via $\overline{E}_{1,3}$ (Figure S4). In this case, R_3 helps to remove the unstable *CHO produced facilely from R_1 and prevent it from decomposition. The TOF for CH_3OH production is promoted only when both E_1 (< 29 kCal·mol^{-1}) and E_3 (< 9 kCal·mol^{-1}) were kept low (Figure S4).

The analysis of the effective descriptors allows us to extract the general optimization guidelines that help us to build in-depth understanding of CO hydrogenation over Cu-based catalysts. The catalytic activity of catalysts can be evaluated based on the effective descriptors in a sequence with decreasing control on the TOF of CH_3OH (Figure 3). E_5, which corresponds to the highest importance score (Figure 1b), controls the TOF the most significantly among the six effective descriptors. This is the case for high E_5 (> 27 kCal·mol^{-1} in this case), which introduces the most rapid change in TOF (Figure 3b). Wherein, E_5 corresponds to the rate-limiting step for CO hydrogenation over the Cu-based catalysts, e.g., *CH_3O hydrogenation. The higher E_5 can lead to the lower TOF of CH_3OH. When E_5 decreases (< 27 kCal·mol^{-1}), the variation in TOF with E_5 alone is rather small (Figure 3b). Instead, it is likely replaced by the weighted E_5 or the normalized second-order descriptors of E_5, $\overline{E}_{1,5}$ or $\overline{E}_{2,5}$ (Figure 3d, e). In this case, facilitating the *CO hydrogenation by decreasing $\overline{E}_{1,5}$ or hindering the reverse, *CHO decomposition, via increasing $\overline{E}_{2,5}$

can effectively promote CH3OH production. Here, the variation in $\overline{E}_{2,5}$ (Figure 3e) clearly introduces more significant changes in TOF than $\overline{E}_{1,5}$ (Figure 3d). That is, well control of the *CHO decomposition to *CO can be more effective than that of *CO hydrogenation to *CHO in tuning the TOF.

In comparison with E_5, the control by the other first-order descriptors (E_1 or E_6) is less significant (Figure 3a, c), which is also observed by the importance scores (Figure 1b). E_1 and E_6 correspond to the two steps, which compete for the *CO species on the surface (Table S1). To achieve high TOF, E_1 is desirable to be lower than the critical point, 29 kCal·mol^{-1} in this case, and E_6 is preferred to be higher than E_1 to enable the CO hydrogenation. As seen for that for E_5, E_1 alone works well as descriptor only when the value is high (> 29 kCal·mol^{-1}, Figure 3a). Otherwise, the weighted E_1, $\overline{E}_{1,3}$, is more effective, where both E_1 and E_3 should be kept low to achieve high TOF for CH_3OH (Figure S4). Note that the analysis of both permutation feature importance scores and partial dependence can provide evaluation on the capability of each descriptor to control the TOF in this specific machine-learned surrogate model. Wherein, the partial dependence plot also indicates the effective range and marginal effect of each descriptor. It allows us to roughly estimate the TOF trend of M-Cu(111) based on the value of a descriptor (Figure 3). In this case, the two methods agree on the decrease in control capability for the first-order descriptors going from E_5 to E_1 and E_6, while the partial dependence enables the identification of variation in control capability of the first-order descriptors. Specifically, when a first-order descriptor cannot control well the TOF by itself the corresponding normalized second-order descriptors or the weighted term by the activation barriers of another relevant step should be considered instead.

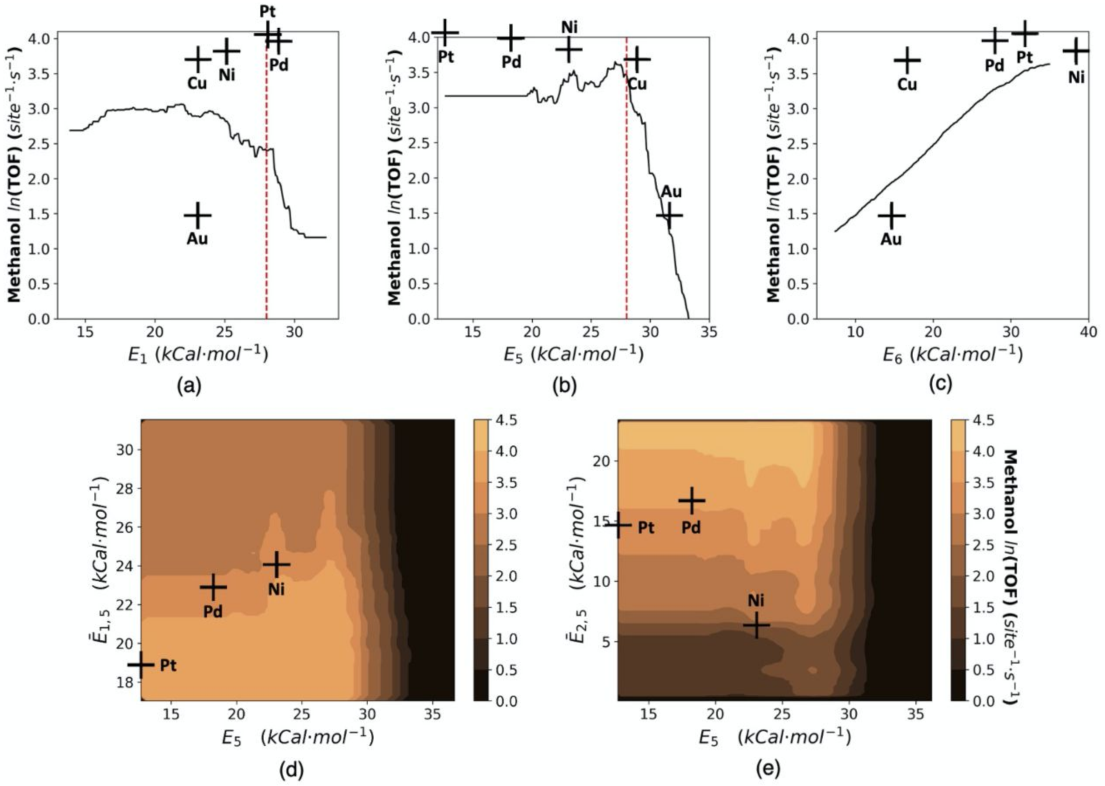

Figure 3. Partial dependence of methanol TOF on (a) E_1, (b) E_5 , (c) E_6, (d) $\overline{E}_{1,5}$, and (e) $\overline{E}_{2,5}$ based on the retrained GBR model. The label "Cu", "Au", "Ni", "Pd" and "Pt" represent Cu(111), Au-Cu(111), Ni-Cu(111), Pd-Cu(111) and Pt-Cu(111), respectively.

While the conventional DRC method only recognized two effective first-order descriptors (Table S4). Like the evaluation according to the permutation feature importance score (Figure 1b), R_5 corresponds to the highest DRC of 0.88 (Table S4) with R_1 as the secondary (DRC = 0.30). Accordingly, the corresponding activation energies, E_5 and E_1, are likely to act as effective descriptors for TOF. Differently, the impact from the other steps and thus the corresponding E_n is rather small and the possible interactions between the elementary steps are completely missed. Given that, the GBR model identifies the effective descriptors more accurately than the DRC method by including the normalized second-order descriptors and provide more in-depth understanding of reaction kinetics

3.3 Model Validation

Although this retrained GBR model has a good predictability on our synthetic dataset, realistic catalytic systems are likely to have different descriptor values and could potentially harm model's performance. To validate the predictability and mechanism understanding on realistic systems, we used the metal (M = Au, Cu, Pd, Pt, Ni, Table S1)-Cu(111) alloy surfaces as testing systems. Although the activation barriers for these systems are within the range of the training set and testing set, the specific values were unknown to the retrained GBR model and reported previously as catalysts for hydrogenation of CO/CO_2 into CH_3OH[67-73]. Here, the kMC simulations were carried out to estimate the TOF of CH_3OH at the same conditions as the data generation. The kMC simulation was based on the DFT-calculated E_n on M-Cu(111) (Table S1), which were cited from our previous study[1]. TOFs were also estimated using the retrained GBR model, where the six descriptors, E_1, E_5 , E_6, $\overline{E}_{1,3}$, $\overline{E}_{1,5}$ and $\overline{E}_{2,5}$ were calculated based on the DFT results (Table S5). The TOFs estimated by DRC model using the identified first-order descriptors E_1 and E_5 were used for comparison. The kMC results were considered as the criteria to evaluate the predictability of the GBR model and DRC method. Indeed, the GBR model displays a decent performance with respect to the kMC-predicted TOF of CH_3OH (Figure 4a and Table S6), showing a decreasing TOF via a sequence: Pt-Cu(111) > Pd-Cu(111) > Ni-Cu(111) > Cu(111) > Au-Cu(111). By comparison, the DRC-predicted TOFs are several orders of magnitude different from the kMC results (Figure 4b and Table S6). More importantly, there is a clear difference in trend of TOF from one system to the next, which is essential to the catalyst screening and is the interest of current study.

The lower predictability of DRC method as compared to the GBR model is mostly associated with the missing normalized second-order descriptors or lack of capture for the non-local behaviours. For instance, E_5 decreases drastically from 28.83 $kCal{\cdot}mol^{-1}$ on Cu(111) to 14.68 $kCal{\cdot}mol^{-1}$ on Pt-Cu(111) (Table S1), which likely indicates a change in mechanism. That is, on Pt-Cu(111) E_5 is not necessarily taken as the effective descriptor anymore (Figure 3b); instead, E_1 and E_6 are likely to have more effect on TOF (Figure 3a,c and Table S1). But the DRC method assumes the same mechanism and keeps the same high weight of E_5 for both Cu(111) and Pt-Cu(111), which results in such a large error. In the current ML model, however, the data-driven framework allows us to vary the E_n simultaneously, and the identified normalized second-order descriptors adaptively change the weight of E_n. In this way the GBR model maintains a good predictability on the TOF even for such non-local case. Wherein, the value of descriptors, *e.g.*, for Pt-Cu(111), are very different from those of reference system, Cu(111) in this case (Figure 4).

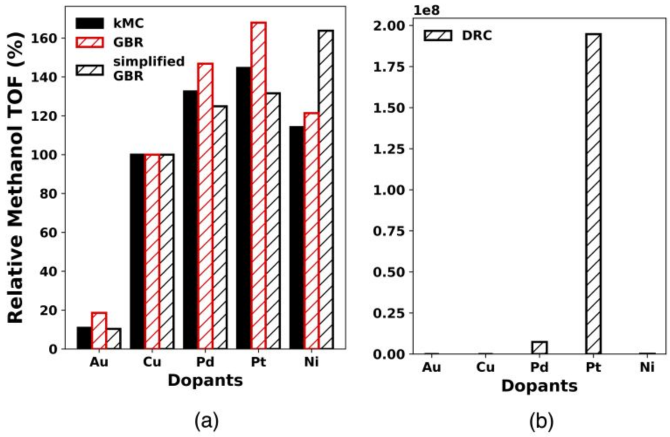

Figure 4. (a) Comparison of the predicted methanol TOF from CO hydrogenation on M-doped Cu(111) between GBR model (shadowed red bar) and simplified GBR model (shadowed black bar) with the kMC simulated values (solid black bar) as criteria.(b) Predicted methanol TOF from CO hydrogenation on M- Cu(111) by DRC method. All TOFs are calibrated to the Cu(111) surface respectively.

The importance of the normalized second-order descriptors is clearly demonstrated by comparing between the GBR model including both first-order and normalized second-order descriptors (GBR, shadowed red bars in figure 4a) and the simplified GBR model only including the first-order descriptors (simplified GBR, shadowed black bars in figure 4a). Although the simplified GBR model retained a good performance on Pd-Cu(111) and Pt-Cu(111), it is not capable to reproduce the trend in TOF of CH_3OH, which is our interest here, by greatly over-estimating the TOF for Ni-Cu(111). According to the kMC simulation, the CO hydrogenation is still hindered since the reverse reaction of *CO hydrogenation (R_2) is almost barrierless (E_2 = 3.69 $kCal{\cdot}mol^{-1}$), and the formed *CHO easily decomposes back to *CO. Such kinetic complexity can be well captured by the GBR model with the identified normalized-second order descriptor $\overline{E}_{2,5}$. Specifically, the TOF of CH_3OH is also sensitive to E_2 when E_5 is low. For the case of Ni-Cu(111), low $\overline{E}_{2,5}$(6.36 $kCal{\cdot}mol^{-1}$) become the major reason why the GBR model recognized it as less active than Pd-Cu(111) and Pt-Cu(111).

3.4 Mechanism Understanding

The retrained GBR model can not only provide the quantitative description of TOF for CO hydrogenation to CH_3OH on M-Cu(111) systems, but also enable the in-depth understanding of the promoting effects by alloying based on the DFT-calculated effective descriptors for M-Cu(111) systems (Table S5). Following the general optimization guidelines extracted from analysis of partial dependence of TOF, we start with the most determinative descriptor E_5 (Figure 3b). Both Au-Cu(111) and Cu(111) (Table S5) correspond to a highly activated R_5 with the corresponding E_5 higher than the critical point (27 $kCal{\cdot}mol^{-1}$). Accordingly, the *CH_3O hydrogenation is likely the rate-limiting step to slow down the overall conversion to CH_3OH (Figure 3b) and thus makes Au-Cu(111) and Cu(111) less active than the other catalysts studied. With the higher E_5 (31.59 $kCal{\cdot}mol^{-1}$), Au-Cu(111) is less active than Cu(111) (E_5 = 28.83 $kCal{\cdot}mol^{-1}$)

With E_5 less than 27 $kCal{\cdot}mol^{-1}$, Pd, Pt, Ni-Cu(111) likely display higher TOF than Cu(111) and Au-Cu(111) (Figure 3b); however, E_5 alone cannot differentiate the sequence of TOF among Pd, Pt, Ni-Cu(111). To do that, the normalized second-order descriptors $\overline{E}_{1,5}$ and $\overline{E}_{2,5}$ should be considered instead. According to the effective $\overline{E}_{2,5}$, an obvious limitation on the TOF over Ni-Cu(111) is observed (Figure 3e). The decomposition of *CHO into *CO (R_2) on Ni-Cu(111) is very facile (E_2 = 3.69 $kCal{\cdot}mol^{-1}$, Table S1), and results in a lower $\overline{E}_{2,5}$ value (6.36 $kCal{\cdot}mol^{-1}$) compared to that over Pt-Cu(111) (14.72 $kCal{\cdot}mol^{-1}$) and Pd-Cu(111) (16.71 $kCal{\cdot}mol^{-1}$, Table S5). As a result, Ni-

This journal is © The Royal Society of Chemistry 20xx

Cu(111) is located in a region that TOF is greatly suppressed (Figure 3e) and is less active than Pd-Cu(111) and Pt-Cu(111).

Following that, the other two first order descriptors, E_1 and E_6, were evaluated to determine the activity of Pd-Cu(111) and Pt-Cu(111). Similarity in partial dependence of TOF between the two surfaces is clearly demonstrated in Figure 3, and the similar TOF is expected. The difference is likely associated with E_6. The *CO desorption from Pd-Cu(111) (E_6 = 27.90 kCal·mol^{-1}) is more facile than *CO hydrogenation (E_1 = 28.83 kCal·mol^{-1}); while in the case of Pt-Cu(111), the trend is opposite (E_6 = 31.82 kCal·mol^{-1}, E_1 = 28.13 kCal·mol^{-1}). That is, *CO prefers the desorption rather than the hydrogenation on Pd-Cu(111), which hinders the CH_3OH production. Although, the difference between E_1 and E_6 is as low as 0.93 kCal·mol^{-1}, the entropic contribution under reaction conditions greatly favours *CO desorption. As a result, Pt-Cu(111) outperforms Pd- Cu(111) toward CO hydrogenation to CH_3OH.

Our results clearly show that using the five effective descriptors, E_1, $\overline{E}_{1,5}$ or $\overline{E}_{2,5}$ and E_1 or E_6 sequencially, the increased TOF of CH_3OH going from Au-Cu(111), Cu(111), Ni-Cu(111), Pd-Cu(111) to Pt-Cu(111) can be well described. By comparison, the contribution from $\overline{E}_{1,3}$ is less significant. The is particular the case for Pd, Pt, Ni-Cu(111). Wherein, E_1 is similarly high for all three systems (28.83 kCal·mol^{-1} for Pd-Cu(111), 28.13 kCal·mol^{-1} for Pt-Cu(111), and 25.14 kCal·mol^{-1} for Ni-Cu(111)). In this case, although E_1 is below the critical level (29 kCal·mol^{-1}), $\overline{E}_{1,3}$ does not act as an alternative descriptor for E_1, and the trend in $\overline{E}_{1,3}$-dependent TOF, Ni-Cu(111) > Pt, Pd-Cu(111), does not follow that in kMC-simulated TOF (Figure S4). Yet, for other systems with lower E_1, $\overline{E}_{1,3}$ it can be more effective to determine TOF.

Given that, the mechanism understanding of effective descriptors can well rationalize the detailed sequence for TOF of CH_3OH among Cu(111) and M-Cu(111) surfaces. More importantly, such understanding also provides the guidance on how to optimize each catalyst with improved TOF: specifically facilitating *CH_3O hydrogenation over Au-Cu(111) and Cu(111) to prevent high E_5, hindering *CHO decomposition over Ni-Cu(111) to prevent low $\overline{E}_{2,5}$, suppressing *CO desorption and/or facilitating the *CO hydrogenation on Pd-Cu(111) and Pt-Cu(111) to prevent higher E_6 than E_1.

Overall, the ability of the current data-driven framework is to enhance the understanding of reaction network on a reference catalytic system obtained from DFT and kMC simulation, where all possible elementary steps are investigated in detail. Based on that, the framework can identify the effective descriptors beyond the single rate-determining step, capture the interaction between relevant elementary steps, and enable the accurate prediction of trend in catalytic activity and provide general principles for further catalyst optimization.

4. Conclusions

The ML-based data-driven surrogate model on the DFT-calculated activation barriers and kMC-simulated TOFs was demonstrated to enhance the accuracy and efficiency in extracting effective descriptors that can control catalytic activity and thus predicting the catalytic activity. Unlike traditional derivative-based methods which only perturb one descriptor at a time, the current ML-based method allows descriptors to vary simultaneously and randomly, so that the ML model can learn non-local behaviours introduced by the interaction between different elementary steps. As a result, a set of effective descriptors, including three first-order descriptors and three normalized second-order descriptors, is identified efficiently.

The trained GBR model based on the effective descriptors predicts the TOF for CH_3OH synthesis from CO hydrogenation over metal-doped Cu(111) alloy surfaces more accurately than that using the DRC model and the simplified GBR model which only includes the three first-order descriptors. More importantly, such model goes beyond the typical mathematic character of ML, being able to greatly enhance the mechanism understanding of promoting effect by alloying as compared to that of DRC and enable the extraction of principles for catalyst optimization toward high CH_3OH production. This approach can be applied to other reactions and catalysts, which opens alternative possibilities to describe the surface reaction kinetics and guide the subsequent optimization of catalysts accurately and effectively.

Conflicts of Interest

Authors claim no conflicts of interest.

Data Availability Statement

Code and original data for reproduction will be available upon request.

Acknowledgements

The research was supported by the Division of Chemical Sciences, Geosciences, & Biosciences, Office of Basic Energy Sciences and carried out at Brookhaven National Laboratory, operated under contract DE-SC0012704 with the US Department of Energy. The DFT calculations were performed using computational resources at the Center for Functional Nanomaterials (CFN), which is a U.S. Department of Energy Office of Science User Facility, at Brookhaven National Laboratory, and at the National Energy Research Scientific Computing Center (NERSC), a U.S. Department of Energy Office of Science User Facility located at Lawrence Berkeley National Laboratory, operated under Contract No. DE-AC02-05CH11231.

References

1. Y. Yang, M. G. White and P. Liu, *J. Phys.l Chem. C*, 2012, **116**, 248-256.
2. S. Kattel, B. Yan, Y. Yang, J. G. Chen and P. Liu, *J. Am. Chem. Soc.*, 2016, **138**, 12440-12450.

3. X. Liu, J. Xiao, H. Peng, X. Hong, K. Chan and J. K. Nørskov, *Nat. Commun.*, 2017, **8**, 15438.
4. M. Zhong, K. Tran, Y. Min, C. Wang, Z. Wang, C.-T. Dinh, P. De Luna, Z. Yu, A. S. Rasouli, P. Brodersen, S. Sun, O. Voznyy, C.-S. Tan, M. Askerka, F. Che, M. Liu, A. Seifitokaldani, Y. Pang, S.-C. Lo, A. Ip, Z. Ulissi and E. H. Sargent, *Nature*, 2020, **581**, 178-183.
5. C. A. Wolcott, A. J. Medford, F. Studt and C. T. Campbell, *J. Catal.*, 2015, **330**, 197-207.
6. A. J. Medford, M. R. Kunz, S. M. Ewing, T. Borders and R. Fushimi, *ACS Catal*, 2018, **8**, 7403-7429.
7. A. P. J. Jansen, *An Introduction To Monte Carlo Simulations of Surface Reactions*, [cond-mat.stat-mech], https://arxiv.org/abs/cond-mat/0303028v1, 2008.
8. C. Karakaya, J. Huang, C. Cadigan, A. Welch, J. Kintner, J. Beach, H. Y. Zhu, R. O'Hayre and R. J. Kee, *Chem. Eng. Sci.*, 2022, **247**, 13.
9. J. K. Nørskov, T. Bligaard and J. Kleis, *Science*, 2009, **324**, 1655.
10. A. J. Medford, A. Vojvodic, J. S. Hummelshøj, J. Voss, F. Abild-Pedersen, F. Studt, T. Bligaard, A. Nilsson and J. K. Nørskov, *J. Catal.*, 2015, **328**, 36-42.
11. M. Salciccioli, M. Stamatakis, S. Caratzoulas and D. G. Vlachos, *Chem. Eng. Sci.*, 2011, **66**, 4319-4355.
12. C. Stegelmann, A. Andreasen and C. T. Campbell, *J. Am. Chem. Soc.*, 2009, **131**, 13563-13563.
13. A. Saltelli, M. Ratto, S. Tarantola and F. Campolongo, *Chem. Rev.*, 2005, **105**, 2811-2828.
14. C. T. Campbell, *ACS Catal.*, 2017, **7**, 2770-2779.
15. K. Tran and Z. W. Ulissi, *Nat. Catal.*, 2018, **1**, 696-703.
16. A. Vojvodic and J. K. Nørskov, *Natl. Sci. Rev.*, 2015, **2**, 140-143.
17. M. M. Montemore and J. W. Medlin, *Catal. Sci. Technol.*, 2014, **4**, 3748-3761.
18. H. Meskine, S. Matera, M. Scheffler, K. Reuter and H. Metiu, *Surf. Sci.*, 2009, **603**, 1724-1730.
19. T. Avanesian and P. Christopher, *ACS Catal.*, 2016, **6**, 5268-5272.
20. J. E. Sutton, W. Guo, M. A. Katsoulakis and D. G. Vlachos, *Nat. Chem.*, 2016, **8**, 331-337.
21. W. Becker and A. Saltelli, *Design for Sensitivity Analysis*, Chapman and Hall/CRC, New York, 1st Edition edn., 2015.
22. A. Saltelli, M. Ratto, T. Andres, F. Campolongo, J. Cariboni, D. Gatelli, M. Saisana and S. Tarantola, in *Global Sensitivity Analysis.*, John Wiley & Sons Ltd, West Sussex, England, 2007, DOI: https://doi.org/10.1002/9780470725184.ch1, pp. 1-51.
23. E. Borgonovo and E. Plischke, *Eur. J. Oper. Res.*, 2016, **248**, 869-887.
24. M. R. Andalibi, P. Bowen, A. Carino and A. Testino, *Comput. Chem. Eng.*, 2020, **140**.
25. S. Dopking, C. P. Plaisance, D. Strobusch, K. Reuter, C. Scheurer and S. Matera, *J. Chem. Phys.*, 2018, **148**.
26. S. Döpking and S. Matera, *Chem. Phys. Lett.*, 2017, **674**, 28-32.
27. H. J. Tian and S. Rangarajan, *ACS Catal.*, 2020, **10**, 13535-13542.
28. S. Lo Piano, F. Ferretti, A. Puy, D. Albrecht and A. Saltelli, *Reliab. Eng. Syst. Saf.*, 2021, **206**, 107300.
29. B. Iooss and P. Lemaître, in *Uncertainty Management in Simulation-Optimization of Complex Systems: Algorithms and Applications*, eds. G. Dellino and C. Meloni, Springer US, Boston, MA, 2015, DOI: 10.1007/978-1-4899-7547-8_5, pp. 101-122.
30. S. Razavi, A. Jakeman, A. Saltelli, C. Prieur, B. Iooss, E. Borgonovo, E. Plischke, S. Lo Piano, T. Iwanaga, W. Becker, S. Tarantola, J. H. A. Guillaume, J. Jakeman, H. Gupta, N. Melillo, G. Rabitti, V. Chabridon, Q. Duan, X. Sun, S. Smith, R. Sheikholeslami, N. Hosseini, M. Asadzadeh, A. Puy, S. Kucherenko and H. R. Maier, *Environ.l Model. Softw.*, 2021, **137**, 104954.
31. T. Turányi, *J. Math. Chem.*, 1990, **5**, 203-248.
32. J. Yang, A. Jakeman, G. Fang and X. Chen, *Environ. Model. Softw.*, 2018, **101**, 289-300.
33. E. H. Y. Beh, F. Zheng, G. C. Dandy, H. R. Maier and Z. Kapelan, *Environ. Model. Softw.*, 2017, **93**, 92-105.
34. J. R. Kitchin, *Nat. Catal.*, 2018, **1**, 230-232.
35. A. Khorshidi and A. A. Peterson, *Comput. Phys. Commun.*, 2016, **207**, 310-324.
36. N. Artrith, A. Urban and G. Ceder, *Phys. Rev. B*, 2017, **96**, 014112.
37. A. R. Singh, B. A. Rohr, J. A. Gauthier and J. K. Nørskov, *Catal. Lett.*, 2019, **149**, 2347-2354.
38. A. Palizhati, W. Zhong, K. Tran, S. Back and Z. W. Ulissi, *J. Chem. Inf. Model.*, 2019, **59**, 4742-4749.
39. O. Mamun, K. T. Winther, J. R. Boes and T. Bligaard, *Sci. Data*, 2019, **6**, 76.
40. N. Artrith, Z. Lin and J. G. Chen, *ACS Catal.*, 2020, **10**, 9438-9444.
41. T. Toyao, Z. Maeno, S. Takakusagi, T. Kamachi, I. Takigawa and K.-i. Shimizu, *ACS Catal.*, 2020, **10**, 2260-2297.
42. Y.-Q. Su, L. Zhang, Y. Wang, J.-X. Liu, V. Muravev, K. Alexopoulos, I. A. W. Filot, D. G. Vlachos and E. J. M. Hensen, *Npj Comp. Mater.*, 2020, **6**, 144.
43. E. O. Ebikade, Y. Wang, N. Samulewicz, B. Hasa and D. Vlachos, *React. Chem. Eng.*, 2020, **5**, 2134-2147.
44. J. Feng, J. L. Lansford, M. A. Katsoulakis and D. G. Vlachos, *Sci. Adv.*, 2020, **6**, eabc3204.
45. Z. W. Ulissi, A. J. Medford, T. Bligaard and J. K. Nørskov, *Nat. Commun.*, 2017, **8**, 14621.
46. W. J. Murdoch, C. Singh, K. Kumbier, R. Abbasi-Asl and B. Yu, *Proc. Natl. Acad. Sci.*, 2019, **116**, 22071.
47. K. C. Waugh, *Catal. Today*, 1992, **15**, 51-75.
48. M. Behrens, F. Studt, I. Kasatkin, S. Kühl, M. Hävecker, F. Abild-Pedersen, S. Zander, F. Girgsdies, P. Kurr, B.-L. Kniep, M. Tovar, R. W. Fischer, J. K. Nørskov and R. Schlögl, *Science*, 2012, **336**, 5.
49. S. Kattel, P. J. Ramírez, J. G. Chen, J. A. Rodriguez and P. Liu, *Science*, 2017, **355**, 1296-1299.
50. W. Liao and P. Liu, *ACS Catal.*, 2020, **10**, 5723-5733.
51. H. Zhang, X. Wang, A. I. Frenkel and P. Liu, *J. Chem. Phys.*, 2021, **154**, 014702.
52. A. P. J. Jansen, in *An Introduction to Kinetic Monte Carlo Simulations of Surface Reactions*, ed. A. P. J. Jansen, Springer Berlin Heidelberg, Berlin, Heidelberg, 2012, DOI: 10.1007/978-3-642-29488-4_4, pp. 73-119.
53. I. Langmuir, *J. Am. Chem. Soc.*, 1916, **38**, 2221-2295.
54. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and E. Duchesnay, *J. Mach. Learn. Res.*, 2011, **12**, 2825-2830.
55. M. Zięba, S. K. Tomczak and J. M. Tomczak, *Expert Syst. Appl.*, 2016, **58**, 93-101.
56. M. A. Newton and A. E. Raftery, *J. R. Stat. Soc. B*, 1994, **56**, 3-26.
57. H. Jung and K.-Y. Chung, *Cluster Computing*, 2014, **17**, 767-774.
58. S. Zheng, C. Ding, F. Nie and H. Huang, *IEEE Trans. Knowl. Data Eng.*, 2019, **31**, 1520-1531.
59. C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*, Mit Press, Cambridge, 2005.
60. J. Lever, M. Krzywinski and N. Atman, *Nat. Methods*, 2017, **14**, 641-642.
61. L. Breiman, *Mach. Learn.*, 2001, **45**, 5-32.

62. G. Chandrashekar and F. Sahin, *Computers & Electrical Engineering*, 2014, **40**, 16-28.
63. J. Cai, J. Luo, S. Wang and S. Yang, *Neurocomputing*, 2018, **300**, 70-79.
64. D. Houtao and G. Runger, arXiv, 2012, preprint, arXiv:1201.1587v3, https://doi.org/10.48550/arXiv.1201.1587.
65. B. Gregorutti, B. Michel and P. Saint-Pierre, *Stat. Comput.*, 2017, **27**, 659-678.
66. M. Galar, A. Fernandez, E. Barrenechea, H. Bustince and F. Herrera, *IEEE Transactio on Systems Man and Cybernetics Part C-Applications and Reviews*, 2012, **42**, 463-484.
67. F. Studt, F. Abild-Pedersen, Q. Wu, A. D. Jensen, B. Temel, J.-D. Grunwaldt and J. K. Nørskov, *J. Catal.*, 2012, **293**, 51-60.
68. Q. Wu, L. D. L. Duchstein, G. L. Chiarello, J. M. Christensen, C. D. Damsgaard, C. F. Elkjær, J. B. Wagner, B. Temel, J.-D. Grunwaldt and A. D. Jensen, *ChemCatChem*, 2014, **6**, 301-310.
69. L. Liu, F. Fan, Z. Jiang, X. Gao, J. Wei and T. Fang, *J. Phys. Chem. C*, 2017, **121**, 26287-26299.
70. D. G. Araiza, A. Gómez-Cortés and G. Díaz, *Catal. Today*, 2020, **356**, 440-455.
71. M. D. Esrafili and B. Nejadebrahimi, *Appl. Surf. Sci.*, 2019, **475**, 363-371.
72. X. Nie, X. Jiang, H. Wang, W. Luo, M. J. Janik, Y. Chen, X. Guo and C. Song, *ACS Catal.*, 2018, **8**, 4873-4892.
73. S. Bai, Q. Shao, P. Wang, Q. Dai, X. Wang and X. Huang, *J. Am. Chem. Soc.*, 2017, **139**, 6827-6830.