Machine learning accelerated first principles study of the hydrodeoxygenation of propanoic acid

Wenqiang Yang[1, ‖], Kareem E. Abdelfatah[2, ‖], Subrata Kumar Kundu[1], Biplab Rajbanshi[1, 4], Gabriel A. Terejanu[3, *], Andreas Heyden[1, *]

[1] *Department of Chemical Engineering, University of South Carolina, 301 S. Main Street, Columbia, SC 29208, USA.*

[2] *Department of Computer Science and Engineering, University of South Carolina, 301 S. Main Street, Columbia, SC 29208, USA.*

[3] *Department of Computer Science, University of North Carolina at Charlotte, Charlotte, North Carolina 28262, United States*

[4] *Department of Chemistry, Visva-Bharati University, Santiniketan 731235, India*

[‖] These authors contributed equally to this work.
Corresponding author email: heyden@cec.sc.edu; gabriel.terejanu@uncc.edu
Phone: +1-803-777-5025

ABSTRACT

The complex reaction network of catalytic biomass conversions often involves hundreds of surface intermediates and thousands of reaction steps, greatly hindering the rational design of metal catalysts for these conversions. Here, we present a framework of machine learning (ML) accelerated first-principles studies for the hydrodeoxygenation (HDO) of propanoic acid over transition metal surfaces. The microkinetic model (MKM) is initially parameterized by ML-predicted energies and iteratively improved by identifying the rate-determining species and steps (RDS), computing their energies by DFT, and re-parameterizing the MKM until all the RDS are computed by DFT. The Gaussian process (GP) model performs significantly better than the linear ridge regression model for predicting both the adsorption free energies and transition state free energies. Parameterized with energies from the GP model, only 5-20% of the full reaction network has to be computed by DFT for the MKM to possess DFT-level accuracy for the TOF and dominant reaction pathway. While the linear ridge regression model performs worse than the GP model, its performance is greatly improved when only transition states are predicted by the regression model and adsorption energies are computed by DFT. Overall, we find a high accuracy in adsorption free energies is more important for a reliable MKM than a high accuracy in TS free energies. Finally, based on the GP model with G_{OH} and G_{CHCHCO} as catalyst descriptors, we build two-dimensional volcano plots in activity and selectivity that can help design promising alloy catalysts for HDO reactions of organic acids.

Keywords: Biomass conversion, Large reaction network, Density functional theory, Machine learning, Iterative microkinetic model, Volcano plot

1. Introduction

The study of the usage of biomass as an alternative to fossil fuels has greatly increased due to environmental and climate issues brought about by the combustion of fossil fuels and their derivatives[1]. However, biomass-derived fuels have their drawbacks, such as high viscosity, poor oxidation stability, low energy density, high cloud point temperature, etc., due to their high oxygen content[2-5]. Deoxygenation therefore becomes very important in the chemical utilization of biomass. Currently, the heterogeneous catalyzed hydrodeoxygenation (HDO) process, which is conducted in both vapor and liquid phases, is widely used for the conversion of biomass to valuable products[6, 7]. Conventional petrochemical-based HDO catalysts, such as sulfided $NiMo/Al_2O_3$ and $CoMo/Al_2O_3$ cannot be used for the conversion of biomass due to the low sulfur content in biomass, high sulfur content in the final product, and short catalyst lifetimes[6]. To overcome these limitations of conventional sulfided catalysts, new metal catalysts must be developed for the HDO of biomass, which requires a comprehensive understanding of the HDO mechanism on the catalyst surface. Unfortunately, the surface catalyzed reaction network of biomass conversion often involves hundreds of intermediates and thousands of elementary reaction steps which greatly limits the computational and experimental study of this complex reaction network. Furthermore, supported metal particles display multiple surface facets and active sites and in principle, all of these sites need to be studied to identify the most relevant active sites.

Microkinetic modeling based on parameters obtained from density functional theory (DFT) calculations has been proven to be a powerful tool for understanding reaction mechanisms in heterogenous catalysis[8, 9], however, most models only consider small molecules with typically only one or two carbon atoms to reduce the size of the reaction network and the corresponding computational effort for computing the free energy of the various species and transition states (TS). Fortunately, most surface species and TS have only a negligible impact on the various quantities of interest (such as turnover frequency and selectivity), and in fact only a few key species and TS control the rate of the overall reaction network such that most computational resources should be focused on these highly important surface species while the majority of the reaction network can be described with lower accuracy and less costly approaches[8]. In other words, identifying a meaningful (but

cheap) model to estimate the properties of most species and TS of the reaction network while locating the most important species is of high importance for any computational (and experimental) study of complex reaction networks.

Wolcott et al.[10] recently used the idea that only a few species and TS determine the overall performance of a catalyst and introduced a degree of rate control approach to catalyst design that aims at identifying the most relevant species and TS for a reference system. It then assumes that the same reaction mechanism, rate-controlling species, and degrees of rate control also hold for other catalyst surfaces. Under these assumptions, to predict the rate of reaction for a new catalyst system relative to the rate of a reference system, it is necessary to only determine the free energy for the most rate-controlling species (identified for the reference system) on the new catalyst surface (by DFT, experiment, or semiempirical methods). Unfortunately, the reliability of this approach strongly depends on the degrees of rate control not changing when changing the catalyst surface, which is unlikely for large reaction networks of biomass catalysis that typically contain many partial rate controlling steps or for catalyst surfaces that are very different to the reference system.

Alternatively, the idea of a few important species is not utilized, and it is instead attempted to correlate the energies of the various species and TS to descriptor values. Thus, the microkinetic models are only a function of descriptor values, and volcano curves in activity and selectivity can be generated that are only functions of these descriptors. Specifically, one of the most popular, and also simplest, approaches proposed by Nørskov, Bligaard, and coworkers[11-15] is to use linear scaling relations to relate the adsorption energy of all reaction intermediates on the catalyst surface to one or two adsorption energies of small molecules or atomic species such as C, O, OH and CO. Instead of adsorption energies, intrinsic properties of the catalyst such as coordination number, d-band moments, the density of states near the Fermi level, etc. can also be used as descriptors for predicting adsorption energies[16-18]. Here, nonlinear functions of the descriptor values have sometimes been found to be superior for multimetallic transition metal surfaces.[16] However, only very few studies considered molecules with more than two carbon atoms or attempted to model larger species on multiple catalyst surfaces such that it is currently still unknown how practically useful these linear and nonlinear approaches are for predicting the adsorption

energy of larger biomass molecules. In this context, more complicated machine learning models based on group additivity[14, 19-23], 'Coulomb' matrix[24-26], bag-of-bonds[27], etc. have recently been proposed to predict the adsorption energy of larger adsorbates. These methods sometimes require the geometry and coordinates of the adsorbed species for reliable predictions which is however often not available and which we found to be very difficult to predict with lower-level methods such as molecular mechanics or DFTB.

Next, TS (free) energies (or forward barriers) have to be predicted for a family of reactions, or the same reaction over multiple active sites, or a combination of both, and again linear correlations such as transition-state scaling[28] (TSS) or Brønsted–Evans–Polanyi[29] (BEP) correlations are commonly used for this task that relate the TS energy to adsorption energy information of the (dissociated) product state or the reactant and product states. Often, a reaction family is defined based on the bond type that is undergoing bond cleavage (e.g., C-C, C-O, C-H, O-H scission), and correlations are built for each family. Given that this linear relation has been shown to roughly hold over various catalyst facets, it has been widely used in the computational catalyst screening community[15, 16, 28-41]. For example, catalyst screening based on linear scaling has been used for the design of heterogeneous catalysts for C-H and N-H bond cleavage[10, 39, 42-50].

However, for large reaction networks typical of biomass conversion, it is more challenging to identify reliable linear scaling relationships as the surface and TS energies are not only dependent on the state of the reactive moiety but also the conformation and strain in the non-reactive part of the surface species that is often quite different for different species on different surfaces. Thus, more descriptors would have to be used in a linear model (which again requires more data during parameterization which might not be practical), or more complex nonlinear machine learning models are tried. Ulissi and coworkers[49] proposed a framework that is based on a Gaussian process model using group additivity for the adsorption energy predictions. In combination with transition-state scaling relations (TSS), the key species and the most likely reaction mechanism have been identified for the Rh(111) catalyzed alcohol production from syngas. Similarly, Sutton et al.[50] presented a systematic hierarchical multiscale framework based on group additivity and Brønsted–Evans–Polanyi (BEP) relations for parameterizing the microkinetic model of ethanol steam reforming over Pt. The reaction system contained 67 surface species and

160 reaction steps, and they found that their method delivers first-principles accuracy at a significantly reduced computational cost. Despite the success of these models in determining the reaction mechanism, they still suffer from some intrinsic drawbacks of the models being used to estimate the surface intermediate adsorption and TS energy. Specifically, the usefulness of the linear relations for the transition state (BEP and TSS) is not clear for larger molecules and more complicated chemistries[33, 38, 51]. For instance, Sutton et al.[52] reported that while the classic BEP relation is sufficient for C-H cleavage, it is inadequate for O-H and C-OH bond scission in ethanol on the Pt(111) and Rh(111) surfaces. Similar observations have been made by Lee et al. for the C-O scission of methoxy and hydroxyl groups in guaiacol on Pt(111)[51].

In this study, we investigate the practical usefulness of machine learning models for parameterizing microkinetic models of reaction networks of biomass conversion on metal surfaces when no DFT data are available for a new surface (but only for other surfaces). Next, we test if the microkinetic models can be improved iteratively by identifying rate and selectivity determining surface species and TS and computing these species by DFT, i.e., can we utilize the idea that only a few species and TS are kinetically important for each surface but that these important species are not necessarily known or are the same for each surface. Essentially, we are testing whether the machine learning models are accurate enough that the microkinetic model solution (with all DFT data) is within the convergence radius of the iterative loop. Here, we compare the performance of stacked-Gaussian Process models for adsorption and TS (free) energies against linear models that use TSS and BEP relations for predicting the TS free energies. While the results are necessarily system, data, and machine learning model dependent, we still hope to provide guidelines when machine learning tools are likely useful and can accelerate computational catalysis studies. As a model system, we studied the HDO of propanoic acid. The HDO of carboxylic acids is relevant for green diesel production from esters[53, 54] and the stabilization of pyrolysis oils[55, 56]. In addition, short chain (C2-C5) volatile fatty acids (VFAs) can readily be produced from various organic wastes by methane-arrested anaerobic digestion, and utilization of these VFAs often requires HDO to, e.g., alcohols and aldehydes[57]. Thus, the identification of novel catalysts for the HDO of propionic acid is important for various biomass conversion processes.

2. Methods

2.1. Computational methods

All computational data utilized in this study are based on DFT calculations with the projector augmented wave (PAW) method[58] as implemented in the Vienna Ab Initio Simulation Package (VASP).[59, 60] The generalized gradient approximation (GGA) with the Perdew and Wang 1991 functional (PW91) was used to treat the exchange-correlation effects[61, 62]. While PW91 lacks van der Waals interactions that are often considered important for the adsorption of larger molecules, its effect on surface reaction energies is generally small[63, 64] and given that PW91 already tends to overestimate adsorption energies, the lack of van der Waals interactions in this study should only have a small impact on the conclusions[65]. An energy cutoff of 400 eV is used and the energy convergence criterion was set to 10^{-7} eV. All structures were relaxed until the Hellmann-Feynman force on each atom was smaller than 0.01 eV $Å^{-1}$. To simulate the surface, a $3 \times 2\sqrt{3}$ (111) or (0001) surface model was constructed with four metal atom layers separated by a 15 Å vacuum gap. The dipole correction was applied to the direction perpendicular to the surface. For all surface calculations, the bottom two layers were fixed to their bulk positions while the top two layers were fully relaxed in all directions. Figure S11 of the Supporting Information illustrates convergence of the adsorption energies with slab thickness to within less than 0.1 eV. In the vibrational frequency calculations, all metal atoms were fixed at their optimized positions. For free energy calculations, frequencies below 100 cm^{-1} were shifted to 100 cm^{-1} for the calculation of partition functions so that errors associated with the harmonic approximation are minimized for the small frequencies[66]. The Brillouin zone integration was sampled by a 4×4×1 k-point mesh based on the Monkhorst-Pack scheme[67]. To locate the TS of the elementary reaction steps, a combination of the climbing image nudged elastic band (CI-NEB) method and the dimer method was used[68-71]. In this approach, NEB calculations are conducted for 20-30 steps without full convergence. These NEB results serve as an initial guess for the structure and reaction coordinate of subsequent dimer optimizations of the TS structure. For dimer-optimized TS structures, we compute all vibrational frequencies and confirm that the TS structure is a first-order saddle point and the eigenvector corresponding to the imaginary frequency describes the movement

from the reactant to the product state. All dimer-optimized TS structures are provided in the Supplementary Information. Similarly, for all ground states DFT energies and structures are provided in the Supporting Information of this paper.

2.2. Machine learning descriptors and models

We have previously studied the HDO of propionic acid over five transition metal surfaces, including Pt(111)[72], Pd(111)[73], Cu(111)[74], Rh(111)[75], and Ru(0001)[76]. New data for Ag(111), Re(0001), Pd(100), and Ni(111) have been generated for this study. For Ag(111) and Re(0001), we only computed adsorption energies, i.e., no transition states. As machine learning models for parameterizing the microkinetic models, we utilize our prior research on modeling both surface species and TS for the HDO of propionic acid.

The adsorption energies of the surface intermediates were predicted using both linear and nonlinear models with molecular fingerprints as species descriptors as illustrated in Fig. S1 of the SI and adsorption energies as material descriptors. The molecular fingerprints can be generated from either the chemical formula or the SMILES representation of the molecules[77]. In the molecular fingerprints, the atoms are divided into subclasses by the element number and the number of free valences of the atom (the number of unpaired electrons; we assume O and C can have a maximum of 1 and 3 unpaired electrons – isolated adsorbed C and O atoms are not part of this study, explaining why 2 and 4 unpaired electrons are not possible in this study), which can identify the type and number of unsaturated atoms in the molecules and capture the likely bonding information between the molecules and the catalyst surface. For instance, the carbon atom is divided into C^0, C^1, C^2, and C^3, which correspond to a carbon atom that has 0, 1, 2, and 3 free electrons that can form bonds to the metal surface, respectively. Besides, the number of various bonds was also identified based on the atom types (we assume single bonds for all the atom-atom connections except for the C=O double bond), such as C^0-O^0 and C^2-O^1, etc. For instance, one C^0-O^0 and C^2-O^1 bond will be in $CH_3CH_2CH_2OH$ and CH_3CH_2CO, respectively. Only C-H, C-C, C-O and C=O bonds are considered in this work since only C, H, and O atoms are found in all the surface intermediates, and C-C double and triple bonds typically strongly interact with surface metal atoms leading to an effective single C-C bond and various bonds with the metal surface atoms. Besides the molecular fingerprints,

descriptors to represent the catalysts (metal surfaces) are also needed. From principal component analysis (PCA) and varimax rotation, we found that about 98% of the variance of the adsorption data used in this work can be explained by the adsorption energy of CHCHCO and OH[78]. Therefore, the adsorption energy of CHCHCO and OH were used as descriptors for the catalysts (metal surface). To compare the performance of the nonlinear and linear models, Gaussian process regression and ridge regression were used since they are among the best models in predicting the adsorption (free) energies of surface intermediates according to our previous studies[78, 79].

For the TS free energy predictions, various categories of descriptors and ML models were studied[79]. We use metal descriptors (such as the adsorption energies of H, C, OH, CH_3CH and CHCHCO), molecular fingerprints of the reactants and products and reaction descriptors (such as the adsorption energy of the reactant, the sum of adsorption energies of the products and the reaction energies), to describe the variabilities of the TS free energies. The performance of the various combinations of these descriptors and ML models in predicting the transition state energies was systematically investigated in our previous study[79]. The best descriptors for the ridge regression and the Gaussian process are $E_R\_E_{p1}\_E_{p2}\_FP_R\_FP_{p1}\_E_{CH3CH}$ and $E_R\_E_{p1}\_E_{p2}\_FP_R\_FP_{p2}\_E_{CH3CH}$, respectively, where E_R, E_{p1}, E_{p2}, and E_{CH3CH} are the adsorption free energy of the reactant of the (dissociation) reaction, the adsorption energy of the smaller product and the bigger product of the reaction and adsorbed CH_3CH, respectively, while FP_R, FP_{p1} and FP_{p2} are the fingerprints of the reactant, smaller product and bigger product of the reaction, respectively. Therefore, these two sets of descriptors were used for the prediction of the TS free energies for the linear model (ridge regression) and nonlinear model (Gaussian process) in this work, respectively. The well-known transition state scaling relationship (TSS)[28] which linearly correlates the transition state energy to the product energy was also studied. Except when otherwise specified, all the energies used in this work are free energies calculated at a temperature of 498 K such that the results from ML predictions can directly be used in a microkinetic model.

2.3. The iterative descriptor-based approach

As shown in Fig. 1, our approach consists of 4 steps after building the database: (1) predict the surface intermediate adsorption free energies at reaction conditions with the optimized machine learning model and proper descriptors; (2) predict the transition state free energies with the optimized models based on the predicted adsorption free energies from step 1; (3) parameterize the MKM with the predicted adsorption and transition state free energies, run the microkinetic model and sensitivity analysis to determine the rate-controlling species and transition states, followed by DFT calculations for these rate-controlling species and TS if they are only predicted from ML models. After this step, there are two different self-consistent loops, the outer loop (retraining) and the inner loop (non-retraining). In the self-consistent retraining loop, the DFT calculations of the rate-controlling species will be added to the database and the ML model used in step 1 and 2 will be retrained and re-optimized. New predictions from the reoptimized model and the rate-control species will then parameterize the MKM again to obtain new rate-controlling species. This process will continue until all the rate-controlling species identified by the MKM are calculated via DFT. In the self-consistent non-retraining loop, the DFT calculated rate-controlling species and TS will not be used for the ML training and will instead only be used to refine the MKM parameters. New rate-controlling species will then be identified by the refined MKM. For both approaches, the loop will stop when all the rate-controlling species are calculated by DFT; (4) Finally, based on the optimized model and catalyst descriptors, volcano plots on catalyst activity and selectivity will be built, which can provide a guide to catalyst design and synthesis. We reiterate that no coordinate information is used for the ML models predicting the adsorption/TS energies of surface species. Only the DFT calculated values of the OH and CHCHO binding energies are used as ML input feature identifying the metal surface. All other energy-related input features are from ML-predicted adsorption energies, for instance, the reaction energies, reactant energies, and product energies etc. The non-energy related input features are the fingerprints of the adsorbates that are based on a 1D description of the adsorbed species without any surface information.

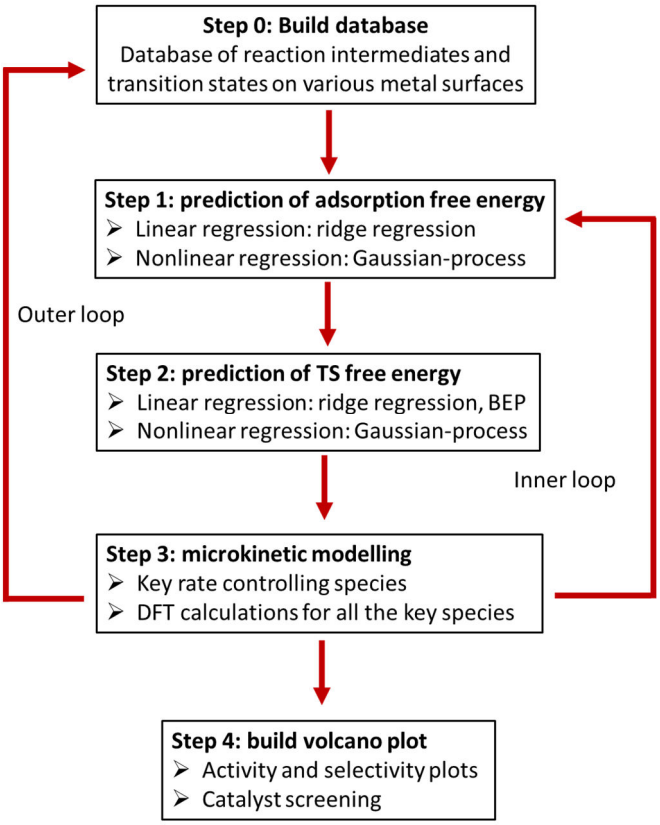


Figure 1. Schematic representation of the iterative descriptor-based approach that is used to identify the reaction mechanisms and activity of the HDO of propionic acid on transition metal surfaces.

3. Results and Discussion

3.1 Approach validation: HDO of propanoic acid through decarbonylation and decarboxylation

3.1.1 Adsorption energy and transition state energy predictions

To benchmark the accuracy of the various approaches for predicting adsorption and transition state free energies, we considered the reaction network shown in the supplementary information Fig. S2 that only contains decarbonylation (DCN) and decarboxylation (DCX) pathways to alkane products. The reason we did not consider pathways to produce alcohols and aldehydes in this validation study is related to us having only a complete database, i.e., all adsorbed species and TS, for only the DCN and DCX pathways. Alcohol and aldehyde production can occur on some metals such as Rh and Cu, however, these pathways are neglected in this validation study (but they are considered in

the catalyst design study discussed in section 3.3). Specifically, the reaction network consists of 29 surface intermediates and 41 surface reaction steps per surface. There are an overall of 232 surface intermediates adsorption data points (8 metal surfaces) and 246 transition states (only 6 metal surfaces having transition states available – we did not compute any TS for Re and Ag as these metals are very oxophillic and very noble, respectively, and hardly display any catalytic activity).

For the training and testing process of the machine learning models, data points for all metals have been used for training while leaving out one metal for testing, and the data points from Re(0001) and Ag(111) surfaces are always used in the ML model training for adsorption energy prediction since we do not have transition state energies on these two metal surfaces. In other words, we are assuming a somewhat realistic scenario that the (complete) reaction network has been studied for all known metal catalysts in the database and we attempt to use this information to predict the performance for an unknown metal surface. To make a fair comparison, we tried all combinations of known and unknown surfaces. The MAEs of the adsorption and TS free energy predictions for the six metal surfaces are shown in Table 1. For the prediction of the adsorption free energy of the surface intermediates, the molecular fingerprints, G_{CHCHCO} and G_{OH} are used as the input features/descriptors. Generally, the predictions of the Gaussian process can achieve an MAE that is significantly smaller than for the ridge regression predictions, and except for the Cu(111) surface, the MAEs of the Gaussian process predictions are below 0.2 eV for all the other metal surfaces. The ridge regression prediction errors are usually more than 0.2 eV for all the studied six metal surfaces. Based on the predicted adsorption energies, linear (ridge regression) and nonlinear (Gaussian process) models were used to predict the transition state energies. Again, the Gaussian process regression outperformed the linear ridge regression model. The widely used TSS relationship which uses the product energy as the input descriptor was also used to predict the transition state energy. Not surprisingly, a very large MAE of the predictions is found for all the six metal surfaces, which agrees very well with results from previous studies demonstrating that BEP and TSS are inadequate for O-H, C-O and C-OH bond scission in biomass conversion reactions[33, 38, 51, 52]. We note that large prediction errors, e.g., larger 0.2 eV, can occur, especially for TS predictions, on some metal surfaces, which could lead (at 498 K) to a two-order magnitude

shift in the compute rate constant. However, we identify all RDS with a degree of rate control, χ_i, larger than 0.01 in each MKM and calculate the energy of these states by DFT and then re-parameterize the MKM. Thus, the only states with large errors are those with a degree of rate-control of less than 0.01, i.e., they are not important. At 498 K, an energy error of $G_{error} = 0.2\ eV$ (or 0.4 eV) for, e.g., a transition state with a degree of rate control of 0.01 leads to a change in rate of less than 5 % (or 10%); $\frac{r_{err}}{r_{true}} = exp\left(\chi_i \frac{G_{error}}{RT}\right) < 1.05\ (or\ 1.1)$. Thus, even though we may have two or even four orders of magnitude errors in calculated reaction rate constants for some steps, their effect on the TOF is small.

Table 1. MAEs in eV of the adsorption free energy and transition state free energy prediction at 498 K with different ML models. RR represents ridge regression, GP represents Gaussian process regression, and TSS is transition state scaling.

	Adsorption free energy		**TS free energies**		
	RR	**GP**	**RR**	**TSS**	**GP**
Cu(111)	0.399	0.232	0.319	0.568	0.290
Ni(111)	0.217	0.158	0.232	0.338	0.223
Pd(111)	0.231	0.124	0.212	0.520	0.149
Pt(111)	0.300	0.194	0.301	0.455	0.278
Rh(111)	0.214	0.122	0.242	0.396	0.193
Ru(0001)	0.294	0.145	0.269	0.675	0.222

3.1.2 Reaction kinetics: linear vs nonlinear

Next, we tried the *non-retraining approach* (inner loop in Fig. 1) and started with the nonlinear stackedGP-based model for predicting the free energies of the "new" catalyst surface. The details of the microkinetic model and the sensitivity analysis[80, 81] to identify the rate-controlling species are included in section S3 and S4 in the supplementary information. It is to be noted that we only consider a species to be rate-controlling when its corresponding degree of rate control is greater than 0.01. The predicted reaction rates of the DCN and DCX path on the six metal surfaces at each iteration loop from the Gaussian process based MKM are shown in Fig. 2. For comparison, the corresponding rates from the MKM with parameters obtained by DFT are also shown. We note that no lateral interaction model is used here and thus, the rates are artificially low due to high CO and H coverage; however, the model should still be meaningful to judge the practical usefulness of an iterative cycle for improving the reliability of a machine learning model and the

consequences on the predicted catalyst activity, etc. Generally, the Gaussian process performs very well in predicting the reaction rate of the DCN path on all six studied metal surfaces, and the predicted rate converges to the DFT calculated values very fast within 2-3 cycles. Next, the predicted DCN rates are always larger than those of the DCX path, demonstrating that the DCN path is preferred over the DCX path on all of the six surfaces which agrees with the DFT calculations. For the DCX rate prediction, only on the Pd(111) and Cu(111) surfaces can the GP-based approach get a similar DCX rate as that computed by an MKM parameterized by DFT. Considering that the DCX mechanism is not favored over the studied metal surfaces and its contribution to the TOF is negligible relative to the DCN mechanism, it is not surprising that the GP-based approach cannot correctly capture the rate-controlling species for the DCX path. We anticipate that determining the DCX path correctly would require us to also compute the degree of rate control for the DCX rate; however, this is likely not practical as more DFT calculations would be needed.

The predicted rate-controlling species from the final cycle of the six studied metal surfaces are shown in Table 2. In the table, we also report the total number of DFT calculated surface intermediates and TS to converge our iterative scheme. The stackedGP-based approach can identify almost all rate-determining species and TS as computed by DFT. Only for the Ni(111), Pt(111), and Rh(111) surfaces do we observe that the GP-based approach failed to identify a few minor rate-controlling species. However, since the importance of these species is minor, the predicted TOFs are still almost the same as those predicted in the DFT parameterized model. For instance, for Ni(111), the TOF from the GP-based approach and DFT parameterized models are 1.14×10^{-8} s^{-1} and 1.18×10^{-8} s^{-1}, respectively, while for Pt(111) they are 3.93×10^{-8} s^{-1} and 3.25×10^{-8} s^{-1}, respectively. Usually, the number of required TS computed by DFT for the new surface is 2 to 10 for the GP-based approach (out of a total of 41 in the reaction network), and the number of required surface intermediate adsorption energies computed by DFT is 2-6 (out of a total of 29 in the reaction network). In other words, only 5% to 20% of the full reaction network has to be computed for the GP-based approach to get a DFT level of accuracy in TOF predictions for the HDO of propionic acid, demonstrating a greatly reduced computational cost. Next, in Table 2, the rate-controlling species from the full DFT calculations parameterized MKM also show a significant variation across metals, demonstrating that the approach proposed

by Wolcott et al.[10] for catalyst design, which assumes the same rate-controlling species and degrees of rate control for all catalyst surfaces, is not appropriate for a complex reaction network like biomass conversion.

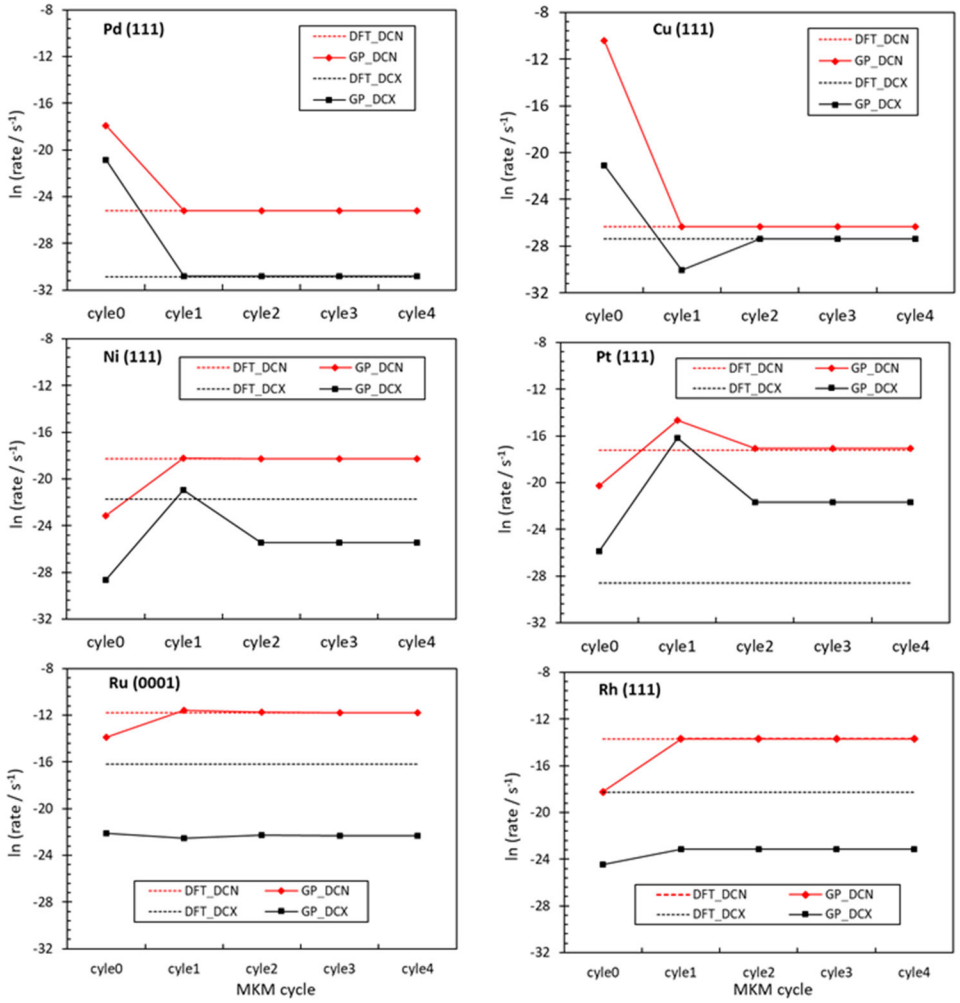


Figure 2. The predicted reaction rate of DCN and DCX paths at each MKM cycle for the six studied transition metal surfaces based on stacked Gaussian process predictions. MKM results based on pure DFT calculations are shown as DFT_DCN and DFT_DCX, respectively. Rates are very low due to the neglect of any lateral interactions. The reaction conditions are T = 473 K and partial pressures of 1 bar for propanoic acid, 0.1 bar for H_2 and 0.001 bar for CO.

Table 2. Predicted degrees of rate control of the key species in the final cycle for the HDO of propionic acid from various models. Stacked ridge and ridge+TSS denote a ridge regression model was used for both the adsorption free energy and transition state free energy prediction except that only the product free energy is used as input for the TS prediction in the TSS, while stacked GP means a Gaussian process regression model was

used for both the adsorption and transition state free energy predictions. TRC and KRC denote the total number of DFT calculations needed for the surface intermediates and transition states to reach MKM convergence on the specific surface, respectively. The two columns under "org_ads" correspond to results from models parameterized with DFT calculated intermediate adsorption energies but with ridge regression and TSS predicted TS energies, respectively. The "org_ts' column shows the results from the MKM model parameterized with DFT calculated TS energies but with ridge regression predicted intermediate adsorption energies. The final converged TOFs from the various models are also shown. "--" means the degree of rate control is less than 0.01. The columns with results from full DFT-based MKM are also shown as reference.

	DFT	stacked ridge	stacked GP	ridge+ TSS	org_ads		org_ts
					ridge_ts	TSS	ridge_ads
Ni(111)							
CH_2CH	-0.13	--	--	--	--	--	--
CO	-0.57	--	-0.57	--	-0.57	-0.57	-0.56
H	-2.10	-3.70	-2.10	-3.70	-2.10	-2.10	-2.13
TS1	0.93	0.94	0.96	0.95	0.96	0.96	0.90
TS8	-0.10	--	--	--	--	--	-0.03
TS18	-0.06	--	--	--	--	--	-0.02
KRC	--	17	4	9	3	1	--
TRC	--	7	3	6	--	--	25
TOF	1.18×10^{-8}	2.97×10^{-8}	1.14×10^{-8}	2.94×10^{-8}	1.14×10^{-8}	1.14×10^{-8}	1.20×10^{-8}
Pt(111)							
CO	-0.34	--	-0.34	--	-0.34	-0.44	--
H	-1.86	-2.86	-1.86	-2.90	-1.86	-2.40	-2.86
TS1	0.06	0.08	0.05	0.07	0.06	--	0.08
TS2	0.16	0.16	N/A	0.17	0.16	0.05	0.16
TS5	0.74	0.73	0.90	0.68	0.74	--	0.73
TS6	--	--	--	--	--	--	-0.39
TS7	--	--	--	--	--	0.91	--
TS8	--	--	--	--	--	--	-0.35
KRC	--	9	7	20	6	5	--
TRC	--	3	3	13	--	--	5
TOF	3.25×10^{-8}	5.33×10^{-8}	3.93×10^{-8}	5.59×10^{-8}	3.25×10^{-8}	9.47×10^{-9}	5.33×10^{-8}
Rh(111)							
CH_2CH COOH	0.05	--	--	--	--	--	--
COOH	-0.07	--	--	--	--	--	0.01
CO	-0.98	--	-0.98	--	-0.97	-0.75	-0.98
H	-0.87	-3.71	-0.87	-3.70	-0.87	-0.67	-0.87
TS1	0.89	0.93	0.89	0.96	0.89	--	0.90
TS5	0.06	0.03	0.07	--	0.07	0.96	0.07
KRC	--	8	2	16	4	9	--
TRC	--	2	2	8	--	--	18
TOF	1.12×10^{-6}	2.27×10^{-5}	1.11×10^{-6}	2.15×10^{-5}	1.10×10^{-6}	7.73×10^{-8}	1.12×10^{-6}
Ru(0001)							

CH_3CH_2COO	-1.34	-1.33	-1.33	-1.23	-1.24	-1.24	-1.34
H	-0.49	-0.49	-0.48	-0.45	-0.45	-045	-0.49
TS2	0.22	0.24	0.23	0.48	0.45	0.45	0.22
TS5	0.24	0.24	0.25	0.48	0.48	0.48	0.24
TS9	0.26	0.25	0.24	--	--	--	0.26
TS41	0.24	0.23	0.24	--	--	--	0.24
TS31	--	--	--	--	0.02	0.02	--
TS32	--	--	--	--	0.08	--	--
KRC	--	12	8	17	5	7	--
TRC	--	6	5	11	--	--	5
TOF	7.67×10^{-6}	7.50×10^{-6}	7.68×10^{-6}	1.15×10^{-5}	1.16×10^{-5}	1.18×10^{-5}	7.67×10^{-6}
			Pd(111)				
CO	-0.61	--	-0.61	--	-0.61	-0.60	--
H	-1.12	-2.92	-1.12	--	-1.12	-1.10	-2.93
TS1	0.05	0.08	0.05	--	0.05	--	0.08
TS5	0.91	0.88	0.91	--	0.91	--	0.88
TS29	--	--	--	0.96	--	0.96	--
KRC	--	20	4	1	2	1	--
TRC	--	11	2	0	--	--	11
TOF	1.14×10^{-11}	4.43×10^{-11}	1.14×10^{-11}	3.97×10^{-8}	1.13×10^{-11}	3.95×10^{-14}	4.44×10^{-11}
			Cu(111)				
CH_3CH_2COO	-0.32	--	-0.32	--	-0.32	-0.32	--
H	-2.23	--	-2.23	--	-2.23	-2.22	-2.88
TS1	--	0.96	--	0.95	--	--	--
TS5	0.70	--	0.70	--	0.70	0.71	0.70
TS29	0.26	--	0.26	--	0.26	0.26	0.26
KRC	--	4	10	3	5	33	--
TRC	--	1	6	4	--	--	12
TOF	4.92×10^{-12}	9.33×10^{-6}	4.92×10^{-12}	9.23×10^{-6}	4.92×10^{-12}	4.93×10^{-12}	7.27×10^{-12}

A *non-retraining linear modeling approach*, using the same descriptors as in the stackedGP models, was also tested. Two different linear approaches for the TS are used which are Ridge regression and TSS. To clarify, in both approaches, the linear models are identical except that the product free energy is only used as the input feature (descriptor) for the TS prediction in the TSS approach which is widely used in heterogenous catalysis, especially in the small molecules activation. Generally, as shown in Fig. 3, both approaches converge very fast in predicting the rate of the DCN and DCX pathways. However, there is a large discrepancy between the predicted rates and the values predicted by the MKM from DFT calculations due to the failure to correctly identify the rate-controlling species as shown in Table 2. It appears that the DFT solution has been outside the convergence radius of the linear models. For the Pt(111) and Ru(0001) surfaces, the Ridge regression

can capture the most important rate-controlling species and predict a TOF very close to the DFT calculated model. However, it needs more DFT calculations than the GP-based approach. For instance, as shown in Table 2, for Ru(0001), the required total DFT calculations (surface intermediates and TS) are 18 and 13, respectively, for the Ridge regression and the GP-based approach. The same trend was found for the TSS approach. For instance, for the Pt(111) surface, even though the predicted TOF from the TSS approach is close to the value from the DFT model, the required DFT calculations (total of 33) are significantly higher than those needed for the GP-based approach (total of 10), as shown in Table 2. The critical failure of the linear models is their inability to obtain a good estimation of the TOF due to them identifying wrong rate-controlling species. For instance, on Rh(111), shown in Table 2, even with a total number of 24 new DFT calculations, the TSS approach failed to identify CO as the key species and predicted a TOF of 2.15×10^{-5} s^{-1} which is almost one order of magnitude larger than the MKM parameterized by DFT (1.12×10^{-6} s^{-1}). In contrast, the GP-based approach captured all the most important rate-controlling species and predicted a TOF of 1.11×10^{-6} s^{-1} with only 4 DFT calculations. We note that we do not believe that the failure necessarily originates from the use of linear models but from the higher prediction error of critical species and TS.

Interestingly, even the dominant pathway cannot always be identified with some of the linear models. Figure 4 illustrates the dominant reaction pathways of the HDO of propanoic acid for the Rh(111) surface as predicted by the various models. The dominant pathways from the GP-based approach agrees very well with the MKM parameterized by DFT, which follows a DCN reaction path, $CH_3CH_2COOH \rightarrow CH_3CH_2CO \rightarrow CH_3CHCO \rightarrow CH_3CCO \rightarrow CH_3C \rightarrow CH_2C \rightarrow CH_2CH \rightarrow CH_2CH_2$. In contrast, for the linear ridge regression model, the α-carbon needs to be less dehydrogenated prior to decarbonylation. In other words, the linear ridge regression model cannot correctly capture the dominant pathway. Figure S3 to S7 in the Supporting Information illustrates the dominant pathways for the five other metal surfaces.

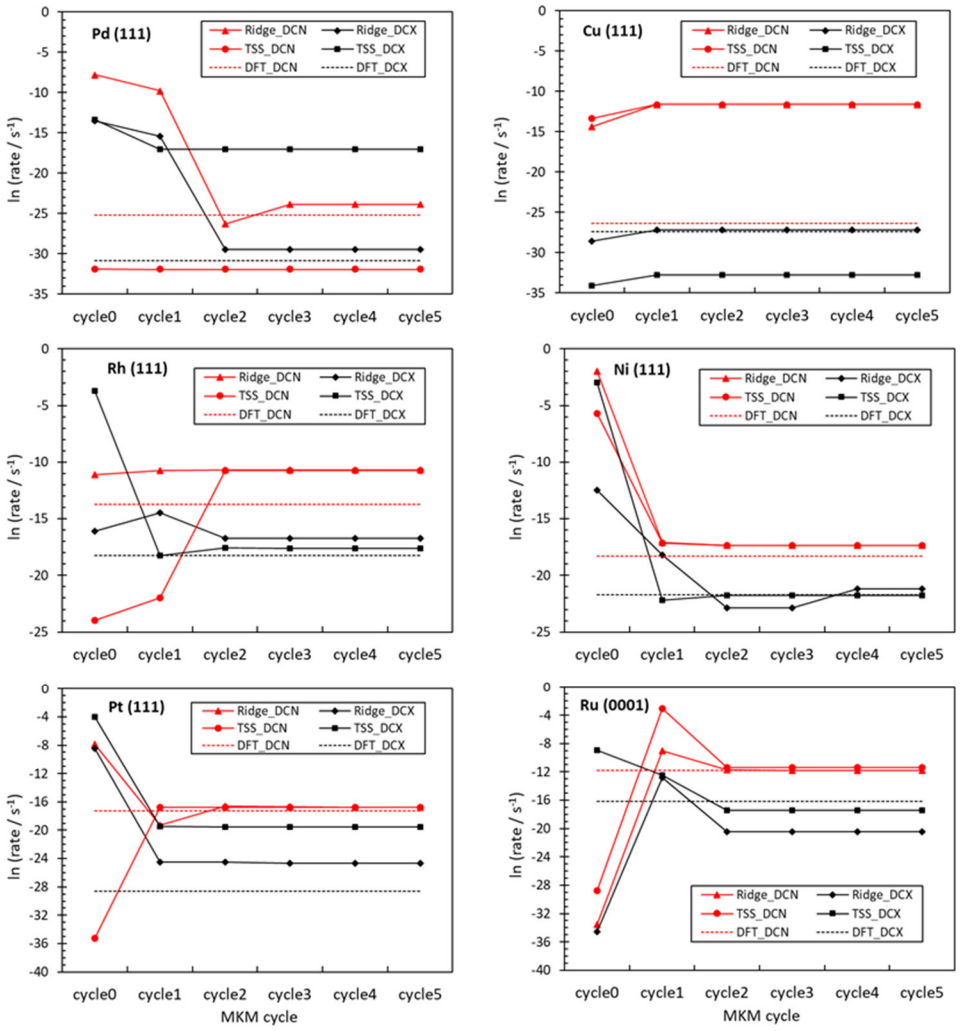


Figure 3. The predicted reaction rate of DCN and DCX pathways of each MKM cycle for the six studied transition metal surfaces from linear regression models and all DFT calculations. Results from the model using TSS for the TS predictions are shown as TSS_DCN and TSS_DCX, and the corresponding results from the ridge regression are shown as Ridge_DCN and Ridge_DCX. The reaction conditions are T = 473 K and partial pressures of 1 bar for propanoic acid, 0.1 bar for H_2 and 0.001 bar for CO.

In summary, the nonlinear Gaussian process models outperform the linear regression models in both the identification of the rate-controlling species and the dominant reaction mechanism. Most importantly, the errors appear to be too large in the linear models to accurately converge to the accurate reaction mechanism and rate-controlling species which often leads to large errors in the activity (TOF) of the transition metal surfaces. In other words, the errors of these models are too large for the models to be

practically useful for a computational catalysis study of the conversion of biomass molecules.

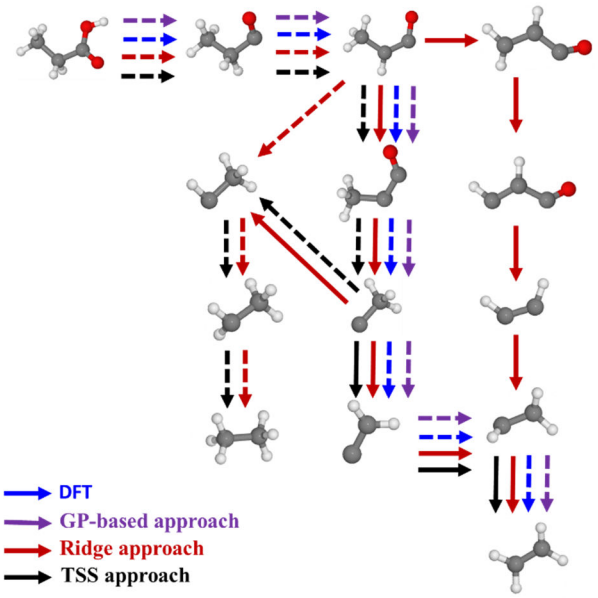


Figure 4. Dominant pathways for the Rh(111) surface determined by various models. Dashed arrows are steps with larger reaction rates while the solid arrows are steps with competitive but reduced reaction rates to the dashed arrow steps.

3.1.3 Importance of the accuracy of the adsorption free energy predictions

Since two sets of predicted free energies are used in the microkinetic model: the free energy of the surface intermediates and the TS, it is instructive to explore which set of free energies is of greater importance in the determination of the catalyst activity (TOF). Given that the linear ridge regression model performs poorly in predicting the key surface and TS species and the TOF, we tested whether the performance of the linear regression model approach could be improved if it is provided with either all surface intermediate free energies or all the TS free energies from DFT calculations. In other words, we are testing for which set of free energies, it is more important to have reliable data and we use the non-retraining approach for this test. In the following, we call the model with all surface intermediates computed by DFT while all TS predicted by ridge regression and TSS, org_ads_ridge_ts and org_ads_TSS_ts, respectively. The model with all TS computed by DFT and all surface intermediates computed by ridge regression, ridge_ads_org_ts,

The predicted TOFs from the three models are shown in Fig. 5. Across all six studied metal surfaces, the org_ads_ridge_ts model displayed a very satisfactory performance despite the relatively poor prediction performance of the TS model. For instance, for the Rh(111) surface, the org_ads_ridge_ts model predicts a TOF of 1.10×10^{-6} s^{-1} which is practically identical to the MKM parameterized by DFT. We note that when using ridge regression also for the surface intermediates, a TOF of $2.15\text{x}10^{-5}$ s^{-1} was predicted which is more than one order of magnitude too large. Generally, the iterative non-retraining model converges to the correct results after computation of only about 10% of the TS in the org_ads_ridge_ts model. In contrast, convergence is slow in the ridge_ads_org_ts model (see Figure 5) although at least for the Ni(111), Rh(111), and Ru(0001) surfaces the model converges to the correct TOF and identifies the most relevant rate controlling species. For instance, for the Ni(111) surface, the convergence of the iterative loop is only obtained after the DFT computation of 25 out of 29 surface intermediates. Furthermore, for the other three metal surfaces (Pt, Pd, and Cu), there are significant discrepancies in the converged TOFs relative to the MKM parameterized by DFT due to the failure of capturing some key surface intermediates as rate-controlling species. Reliable free energies for surface intermediates appear to be significantly more important than TS which is quite noteworthy given that most microkinetic models require the computation of significantly more TS than surface intermediates and given that the computation of TS is about one order of magnitude more expensive than the computation of ground states. This observation also agrees with findings from Sutton et al. for the ethanol steam reforming on Pt[50].

As discussed previously in section 3.1.2, the TSS approach which uses predicted adsorption energies from ridge regression and TS energies from TSS relationship estimations, performs extremely poor in capturing the rate-controlling species and predicting the TOF. We, therefore, investigated if the performance of the TSS model could be improved by using all DFT adsorption energies while the TS is still predicted from the TSS relationships that often display large errors (Table 1). As shown in Fig. 5, an improvement in the prediction of the TOF is observed with the org_ads_TSS_ts model for the Ni(111), Cu(111) and Ru(0001) surfaces, and the predicted TOF agrees quite well with the MKM parameterized by DFT which again demonstrates that as long as the surface

intermediates are computed by DFT, the TS model is of lower importance. However, great discrepancies are found for the Rh(111), Pd(111), and Pt(111) surfaces. Also, to reach convergence, the org_ads_TSS_ts model requires a larger number of DFT calculations than the org_ads_ridge_ts model. For instance, as shown in Table 2, for the Cu(111) surface, the org_ads_TSS_ts model needs 33 TS DFT calculations while the model with a better TS model, org_ads_ridge_ts, requires only 5 TS DFT calculations. We conclude that while an accurate model for the surface intermediates is more important than an accurate model for the TS, if the TS model is not appropriate for the chemistry, e.g., the TSS model is not appropriate for larger molecules, then no reliable predictions can be made.

3.1.4 Retraining versus non-retraining of the machine learning models

To investigate the importance of retraining during the iterative loop, we also studied the performance of the retraining approach (outer loop in Fig. 1). We limit the discussion here to the stackedGP model as predictions are most reliable with this model and we already observed a satisfactory performance for predicting the reaction kinetics in the non-retraining approach. During the retraining process, the rate-controlling species identified at each cycle will be added to the database and the GP model will be reoptimized for the next cycle. New predictions of the adsorption energies and TS from the reoptimized GP will be used for the next MKM cycle. The prediction MAEs from the reoptimized stackedGP model at each retraining cycle, as shown in Table S1, S3, S5, S7, S9 and S11, respectively, for the six studied metal surfaces, suggest that the retraining does not decrease the prediction errors. For some metal surfaces, the MAEs even increase, especially for the Ru(0001) surface where relative to the starting cycle, the corresponding MAEs for the surface intermediate and TS free energy predictions from the final reoptimized GP model are increased by 0.06 eV and 0.02 eV, respectively. For most of the studied metal surfaces, the retraining approach can predict almost the same TOFs and rate-controlling species as that from the full DFT calculations parameterized model. However, the retraining approach requires more DFT calculations to be included in the model than the non-retraining approach. Furthermore, for the Pt(111) and Cu(111) surfaces, there exists a great discrepancy between the retraining approach predicted TOF and the MKM parameterized by DFT, which originates from the retraining approach not correctly capturing all rate-controlling species on these two surfaces, as shown in Table S10 and Table S11. In contrast,

as discussed in section 3.1.1, the non-retraining approach converges to the correct TOF and identifies all critical rate-controlling species for all the studied six metal surfaces with significantly fewer DFT calculations. Therefore, the non-retraining approach is better than the retraining approach in studying the DCN and DCX-based reaction network of the HDO of propanoic acid. Updating the database with DFT-calculated rate-controlling species and retraining the machine learning models does not improve the performance of our iterative scheme. We suspect that this behavior can be explained by the bias that may be introduced into the ML model during the retraining process when adding only a few data points of the metal of interest. This is due to the small reaction network (consisting of only 29 intermediates and 41 transition states) that we studied which has only a few rate-controlling species that need to be calculated by DFT. By adding a few species to the database, a bias may be introduced that causes an increased error in the predictions. Especially, if the rate-controlling species cannot be found on the first retraining cycle, an increased error may occur on its prediction and may lead to a wrong dominant pathway which makes the true rate-controlling species never be found by the model. Therefore, a very large discrepancy occurs between the TOFs from the DFT model and ML predictions. For instance, the adsorption energy of H has a very strong effect in determining the rate of reaction on the Pt(111) surface, which indicates the accuracy of the H adsorption free energy prediction will largely determine the accuracy of the predicted TOF. As shown in Fig. S8 and Table S12, respectively, the absolute prediction error for the H adsorption free energy on the Pt(111) surface is increased by 0.1 eV after retraining and H adsorption has not been identified by the retraining model as a rate-controlling species, which lead to a predicted TOF that is one order of magnitude larger than the DFT value. However, if the retraining approach was used on a more complicated reaction network that needs more DFT calculations being included in the database, the bias would be reduced and the retraining approach could perform better. In the supporting information, we studied the retraining approach on an extended reaction network that contains the alcohol and aldehyde pathways (see section S7). Here, we found an improved performance of the retraining approach; however, overall, more DFT calculations are needed in the retraining approach, and we still conclude that retraining does not benefit the overall efficiency.

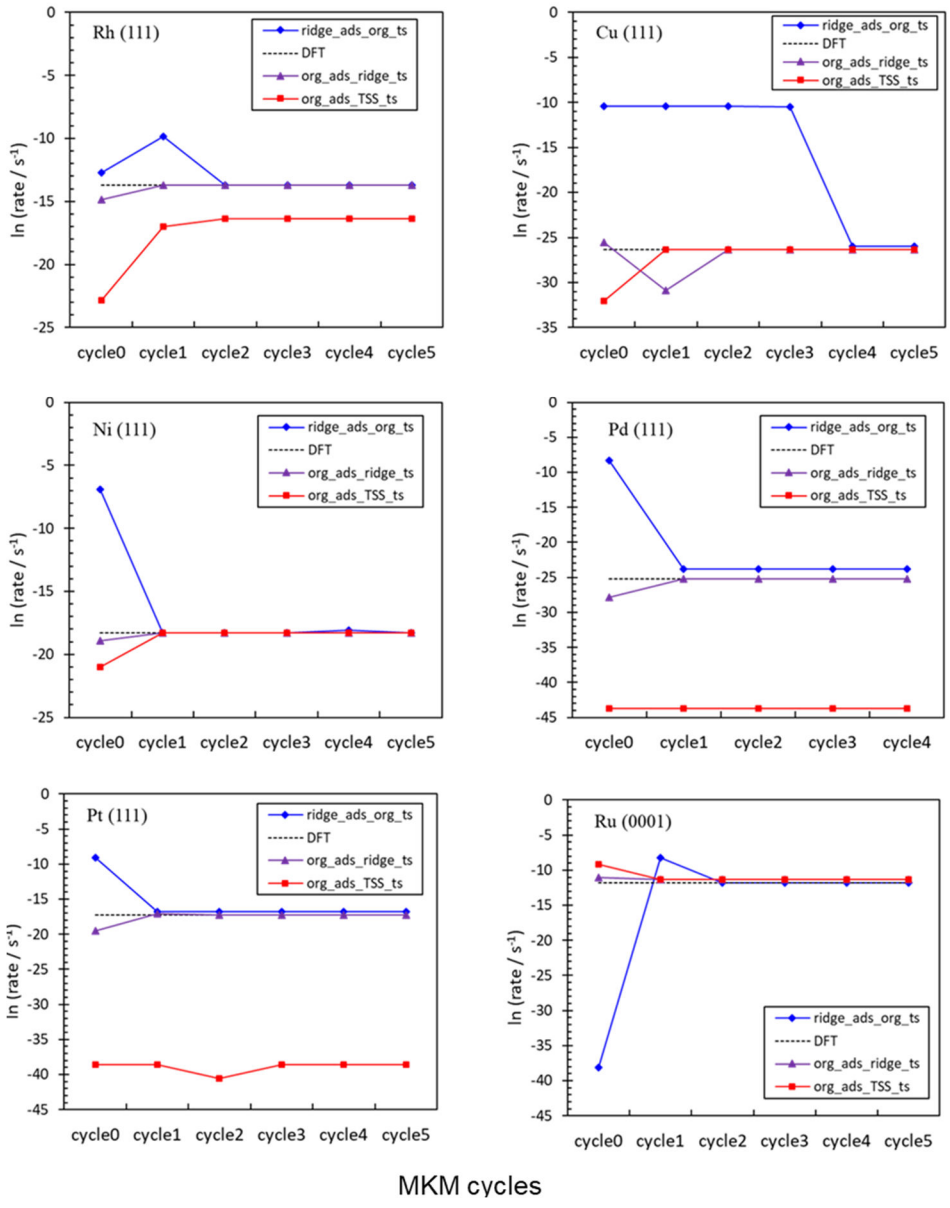


Figure 5. Predicted reaction rate at each MKM cycle for the six studied transition metal surfaces parameterized by the various ML models and DFT calculations. ridge_ads_org_ts indicates that all surface intermediate free energies are predicted by machine learning models while the TS are computed by DFT, org_ads_ridge_ts indicates that all surface intermediates are computed by DFT and only the TS are predicted by machine learning models, org_ads_TSS_ts corresponds to a model with all surface intermediates computed by DFT and all TS predicted by the quite unreliable TSS model. The reaction conditions are T = 473 K and partial pressures of 1 bar for propanoic acid, 0.1 bar for H_2 and 0.001 bar for CO.

3.2 Catalyst screening: activity and selectivity

Finally, the catalyst activity and primary product formation rates (selectivity) of the HDO of propanoic acid are mapped as a function of the adsorption free energies of CHCHCO (G_CHCHCO) and OH (G_OH) for our microkinetic model utilizing the GP model. Here, we considered the extended reaction network for the HDO of propanoic acid, which also includes alcohol and aldehyde formation in addition to decarboxylation and decarbonylation pathways. Section S7 in the supporting information shows that the GP-based model is also validated on this extended reaction network. G_CHCHCO and G_OH are used as the metal descriptors since the GP model for adsorption energy predictions can predict adsorption energies on unknown metal catalysts with only these two DFT calculations, and we wanted to limit the number of metal descriptors to two. We note that in the ML model validation process, the adsorption energies of CH_3CH are used in the TS predictions and, therefore, should also be considered as metal descriptors. However, here its value is from adsorption GP model predictions that use G_CHCHCO and G_OH as input, we thus did not consider it as a necessary metal descriptor since it correlates very well to the two descriptors through the GP model.

Figure 6(a) illustrates the activity as a function of G_CHCHCO and G_OH in the limit of zero conversion. Generally, most studied metals in this work are not located in the most active region. However, a pronounced activity maximum for the HDO of propanoic acid is found, and the Rh(111) surface appears to be the most active (111) surface among the studied metals. Figure 6(a) furthermore suggests that bimetallic alloying could be a possible approach for increasing the activity of a metal catalyst. For instance, Re or Ru have a relatively strong OH and CHCHO binding energy compared to Rh, adding Ru or Re, i.e., forming a Ru-Rh/Re-Rh bimetallic alloy or single metal alloy, the OH and CHCHO binding will most likely be stabilized on a Rh surface and shift Rh-based catalysts to a high activity region on the volcano plot. Thus, a Ru-Rh or Re-Rh catalyst could be an optimal catalyst that binds both CHCHCO and OH optimally. Interestingly, even though we only studied (111) surfaces in this work, the descriptor values of Pd(100) place this surface close to the top of the activity curve. This is interesting as it agrees very well with our previous DFT study that showed that the Pd(100) surface is much more active than the Pd(111) surface for the HDO of propanoic acid.[82] It appears that our model can, to some extent, extrapolate to other surface facets even though our database and ML models are based on (111) surface data only. Fig. 7

illustrates the formation rates towards the various primary products, propionaldehyde, propanol, decarbonylation and decarboxylation products in the limit of low conversion. Propionaldehyde has the highest formation rate (highest selectivity) on all metal catalysts in the limit of low conversion, which agrees well with our recent experimental study on the HDO of propanoic acid over Pt catalysts[83]. Some DCN products and propanol can also form, but they are negligible compared to propionaldehyde formation. The DCX path is never favored on any metal catalyst.

Next, we analyzed how our results change when increasing the conversion to 5-10%. Specifically, aldehyde production is thermodynamically less favorable than alcohol or alkane production in the presence of hydrogen. Thus, increasing the conversion even by a small amount can substantially change our observations. We developed a CSTR reactor model for this purpose. Section S8 in the SI describes the model equations. Figure 6(b) illustrates the activity plot at practically meaningful conversion, and Figure 8 displays corresponding rates to aldehyde, alcohol, decarbonylation, and decarboxylation products. As expected, rates are lower at higher conversion rates except for alcohol production and decarbonylation. Propanol and decarbonylation pathways can follow from propanal. Thus, once propanal is produced, it can readsorb and be converted to alkanes and alcohols. We also observe that the peak propanoic acid conversion shifts to higher adsorption energies for CHCHCO. Overall, the Rh(111) surface and the Pd(100) surface remain active for converting propanoic acid. At 5-10% conversion, propanal, propanol, and, to a lesser degree, DCN products are produced. While this does not 100% agree with experimental and computational observations that the Pd(100) surface primarily produces decarbonylation products[82], some level of deviation is expected given that the heat map is generated by a reactor model parameterized by GP predicted energies and not direct DFT calculations and that the lateral interaction model is developed as an average lateral interaction model of 6 metal surfaces. Nevertheless, the rate difference between decarbonylation and alcohol production is not large for Pd(100), and qualitatively, the trend that metals such as Ru, which have a stronger binding energy for OH, favor alcohol production while Pt and Pd favor more alkane production is reproduced by our model.

Considering that Rh and Pd are active in dehydroxylation (high primary propionaldehyde selectivity) and that Re and Ru are active in propionaldehyde

hydrogenation, we can conclude from our model that a Rh (Pd)-Ru (Re) alloy system could have a very high activity and selectivity towards propanol production. Next, decarboxylation is practically not observed by our model. Still, alkane production through DCN can be optimized when starting with Pd and Pt catalysts and doping them with an element that does not strongly bind oxygen species such as OH but increases the binding energy for carbon species such as CHCHCO. For instance, Ir could be a promising dopant as it binds O moderately, but binds C much stronger than O[84]. While not a doping strategy, Pt(100) is likely also more active than Pt(111) and should display a good selectivity to alkanes. The overall reaction trend of carboxylic acid deoxygenation to an aldehyde followed by alcohol or alkane production agrees very well with our recent experimental study of the HDO of propanoic acid over a Pt/SiO_2 catalyst[83].

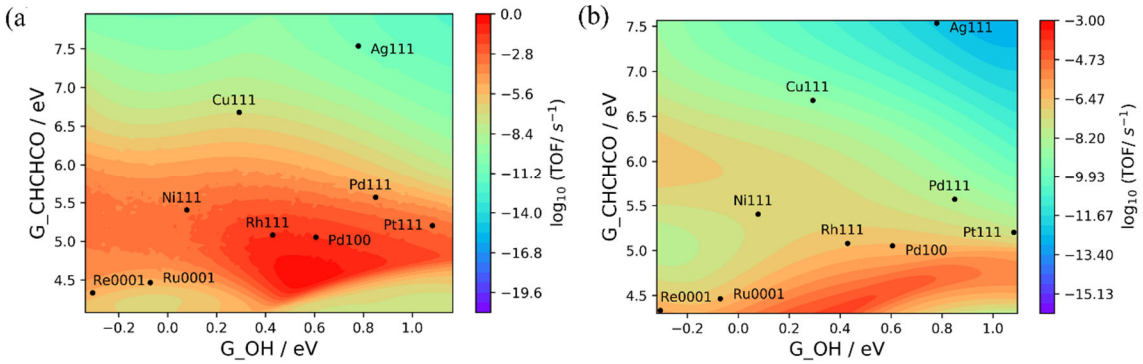


Figure 6. Theoretical activity maps for the HDO of propanoic acid over transition metal surfaces with the TOFs from (a) differential reactor model and (b) CSTR model at a conversion of the reactant of 5-10%. TOFs are determined as a function of the adsorption energies of CHCHCO and OH, respectively. The reaction conditions are T = 473 K and partial pressures of 1 bar for propanoic acid, 0.1 bar for H_2 and 0.001 bar for CO.

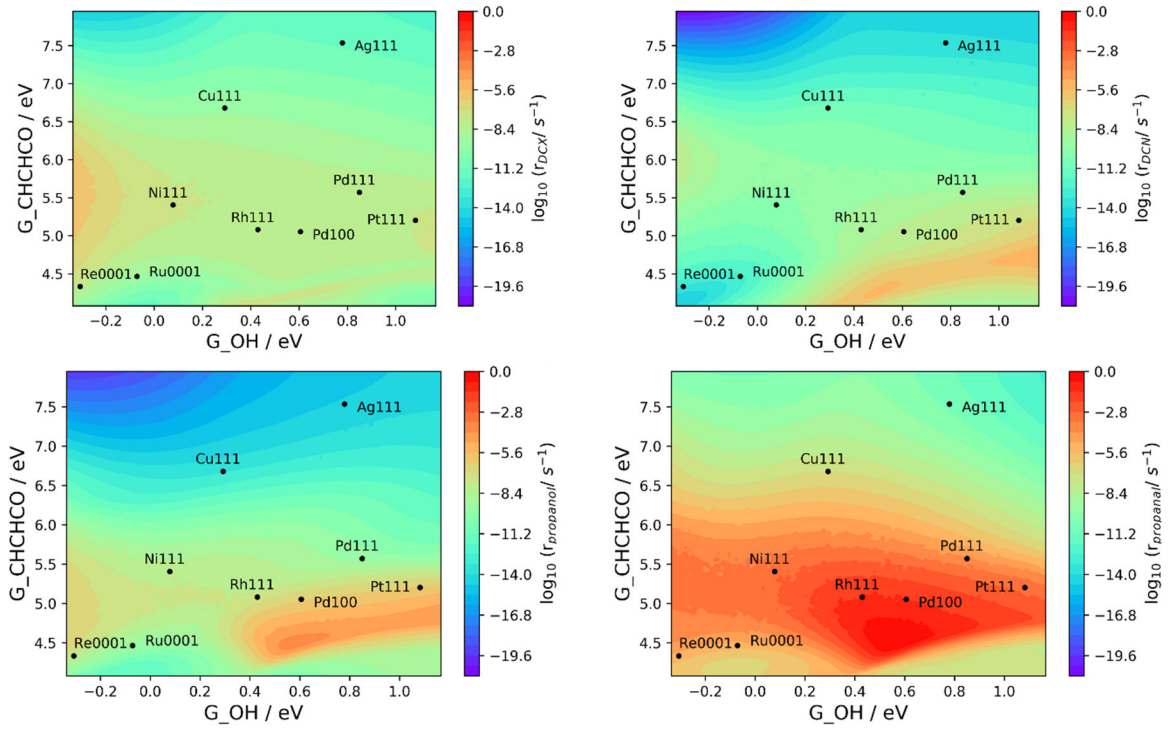


Figure 7. Theoretical reaction rate maps for the DCX path, DCN path, propanol and propionaldehyde production during the HDO of propanoic acid over transition metal surfaces for a differential reactor model. The reaction rates are determined as a function of the adsorption energies of CHCHCO and OH, respectively. The reaction conditions are T = 473 K and partial pressures of 1 bar for propanoic acid, 0.1 bar for H_2, and 0.001 bar for CO.

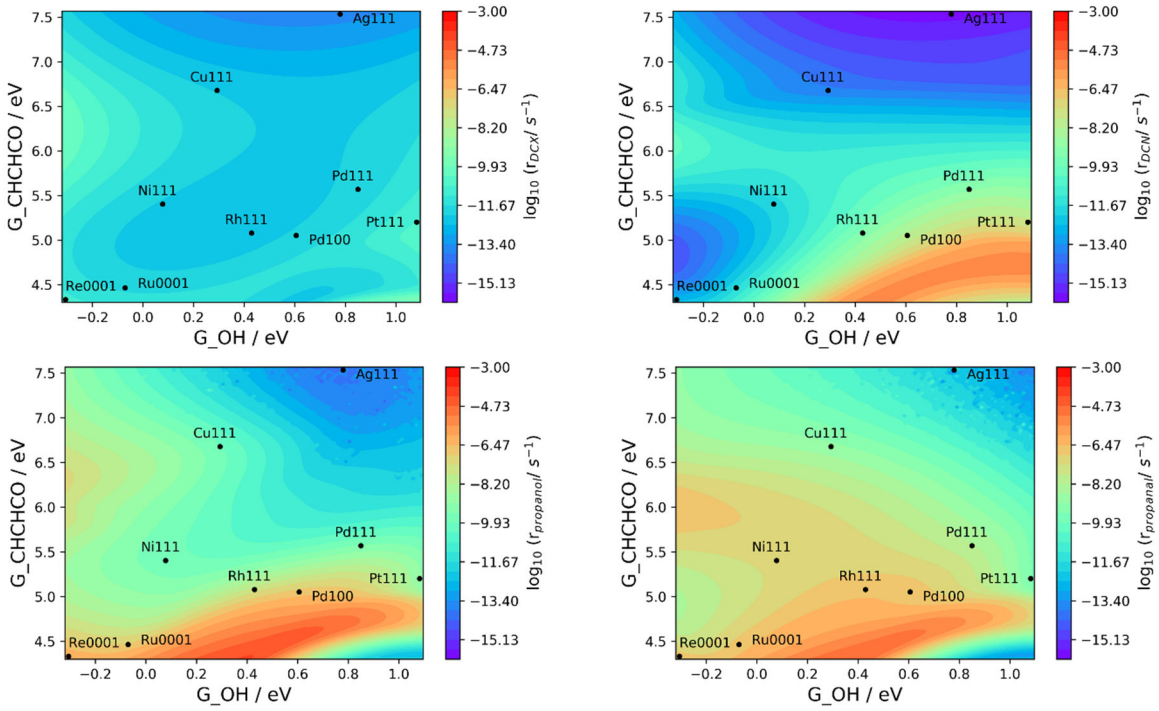


Figure 8. Theoretical reaction rate maps for the DCX path, DCN path, propanol and propionaldehyde production during the HDO of propanoic acid over transition metal surfaces for a CSTR model at a conversion of the reactant of 5-10%. The reaction rates are determined as a function of CHCHCO and OH adsorption energies, respectively. The reaction conditions are T = 473 K and partial pressures of 1 bar for propanoic acid, 0.1 bar for H_2, and 0.001 bar for CO.

4. Summary

We proposed a framework of machine learning (ML) accelerated first-principles calculations to reduce the computational effort for studying the complex reaction network of biomass molecule conversions. As a model system, we studied the HDO of propanoic acid over transition metal surfaces. The MKM uses ML predictions for both the adsorption free energies and transition state free energies of the various species. The MKM is iteratively improved by identifying the RDS, computing the rate-controlling species and transition states from DFT, and re-parameterizing the MKM until all the RDS are calculated by DFT. To predict the adsorption energies of reaction intermediates, the GP and ridge regression models are used since we previously found these two models are among the best non-linear and linear ML models in predicting adsorption energies. GP is found to perform better than ridge regression with smaller prediction MAEs, and based on

the predicted adsorption energies, ML models are built and optimized to predict the TS energies. Besides the GP and ridge regression models, the TSS, which works very well in small molecule reactions, for instance, CO, CH_4, H_2, etc., is also used in predicting the TS energies. GP is still found to be the best model, while TSS predictions have the largest MAEs, ranging from 0.338 eV to 0.675 eV, demonstrating that TSS is not an appropriate model for TS in biomass conversion reactions. With the ML-predicted energies, two different approaches are used to parameterize the MKM, namely the retraining and non-retraining approach. For the retraining approach, the identified RDS from each MKM cycle are calculated by DFT and then added to the database to retrain the ML model giving new energy predictions for the various states in the MKM. In contrast, for the non-retraining approach, the ML model is not retrained and only the free energies of the RDS is updated in the MKM by the DFT calculated free energies. The non-retraining approach based on GP predictions is found to be far superior to the retraining approach for the small reaction network of the DCN and DCX network. Here, it only needs to calculate 5-20% of the states in the full reaction network to correctly predict the activity (TOF) of the metal surface and identify the dominant reaction mechanisms and RDS at DFT-level accuracy. For a larger reaction network which includes alcohol and aldehyde production, the prediction quality of the retraining approach improved; however, more DFT calculations were needed when retraining the ML models. We therefore conclude that retraining does not benefit the overall efficiency for most computational catalysis problems. Since the adsorption free energies and the TS free energies are essential for MMK and they are all from ML, we explored the accuracy of which set of free energies is of greater importance in the determination of a catalyst's activity (TOF). Specifically, we provided the MKM with either full set of DFT-calculated adsorption free energies or TS free energies and ML-predicted energies of the other. The model with full DFT-calculated adsorption free energies showed a much better performance, predicting more accurate TOFs and RDS while requiring fewer DFT calculations, demonstrating the greater importance of accurate adsorption free energies. Finally, we built activity and selectivity volcano plots for the HDO of propanoic acid over metal catalysts from GP model predictions using G_{OH} and G_{CHCHCO} as the metal descriptors. Based on the volcano plot, we provide suggestions for improving the activity and selectivity to specific products through alloying.

Supporting Information

The fingerprint of CH_3CCOO, the full reaction network of the HDO of propanoic acid, detailed information about the MKM, dominant pathways for the DCN&DCX reaction network from both ML- and DFT-based MKM, the performance of the retraining approach on DCN&DCX reaction network and the extended reaction network, detailed information about the CSTR model and the adsorption energy of propanoic acid on various slab with different atom layers.

The optimized adsorption structures of the adsorbates and transition states.

Conflicts of interest

The authors declare no competing financial interests.

ACKNOWLEDGMENT

We gratefully acknowledge financial support from the U.S. Department of Energy, Office of Basic Energy Science, Catalysis Science program under Award DE-SC0007167 (approximately half of the gas phase data and the overall model that contains all machine learning models) and the National Science Foundation under Grant No. DMREF-1534260 (approximately half of the initial gas phase data and individual surface models). Computational resources have been provided by the National Energy Research Scientific Computing Center (NERSC) which is supported by the Office of Science of the U.S. Department of Energy and in part by XSEDE under grant number TG-CTS090100. Computational resources from CASCADE cluster from Environmental Molecular Sciences Laboratory (EMSL) under Pacific Northwest National Laboratory (PNNL) are also used for the DFT calculations. Finally, computing resources from the USC High Performance Computing Group are gratefully acknowledged.

Reference

(1) Alonso, D. M.; Bond, J. Q.; Dumesic, J. A. Catalytic conversion of biomass to biofuels. *Green Chemistry* **2010**, *12* (9), 1493. DOI: 10.1039/c004654j.
(2) Bozell, J. J.; Petersen, G. R. Technology development for the production of biobased products from biorefinery carbohydrates—the US Department of Energy's "Top 10" revisited. *Green Chemistry* **2010**, *12* (4), 539. DOI: 10.1039/b922014c.
(3) Braden, D. J.; Henao, C. A.; Heltzel, J.; Maravelias, C. C.; Dumesic, J. A. Production of liquid hydrocarbon fuels by catalytic conversion of biomass-derived levulinic acid. *Green Chemistry* **2011**, *13* (7), 1755. DOI: 10.1039/c1gc15047b.
(4) Chheda, J. N.; Huber, G. W.; Dumesic, J. A. Liquid-phase catalytic processing of biomass-derived oxygenated hydrocarbons to fuels and chemicals. *Angew Chem Int Ed Engl* **2007**, *46* (38), 7164-7183. DOI: 10.1002/anie.200604274.
(5) Gallezot, P. Conversion of biomass to selected chemical products. *Chem Soc Rev* **2012**, *41* (4), 1538-1558. DOI: 10.1039/c1cs15147a.
(6) Gallezot, P. Catalytic conversion of biomass: challenges and issues. *ChemSusChem* **2008**, *1* (8-9), 734-737. DOI: 10.1002/cssc.200800091.
(7) Wang, H.; Male, J.; Wang, Y. Recent Advances in Hydrotreating of Pyrolysis Bio-Oil and Its Oxygen-Containing Model Compounds. *ACS Catalysis* **2013**, *3* (5), 1047-1070. DOI: 10.1021/cs400069z.
(8) Chorkendorff, I.; Niemantsverdriet, J. W. *Concepts of modern catalysis and kinetics*; 2017.
(9) Olcay, H.; Xu, L.; Xu, Y.; Huber, G. W. Aqueous-Phase Hydrogenation of Acetic Acid over Transition Metal Catalysts. *ChemCatChem* **2010**, *2* (11), 1420-1424. DOI: 10.1002/cctc.201000134.
(10) Wolcott, C. A.; Medford, A. J.; Studt, F.; Campbell, C. T. Degree of rate control approach to computational catalyst screening. *Journal of Catalysis* **2015**, *330*, 197-207. DOI: 10.1016/j.jcat.2015.07.015.
(11) Abild-Pedersen, F.; Greeley, J.; Studt, F.; Rossmeisl, J.; Munter, T. R.; Moses, P. G.; Skulason, E.; Bligaard, T.; Norskov, J. K. Scaling properties of adsorption energies for hydrogen-containing molecules on transition-metal surfaces. *Phys Rev Lett* **2007**, *99* (1), 016105. DOI: 10.1103/PhysRevLett.99.016105.
(12) Fernandez, E. M.; Moses, P. G.; Toftelund, A.; Hansen, H. A.; Martinez, J. I.; Abild-Pedersen, F.; Kleis, J.; Hinnemann, B.; Rossmeisl, J.; Bligaard, T.; Norskov, J. K. Scaling relationships for adsorption energies on transition metal oxide, sulfide, and nitride surfaces. *Angew Chem Int Ed Engl* **2008**, *47* (25), 4683-4686. DOI: 10.1002/anie.200705739.
(13) Jones, G.; Studt, F.; Abild-Pedersen, F.; Nørskov, J. K.; Bligaard, T. Scaling relationships for adsorption energies of C2 hydrocarbons on transition metal surfaces. *Chemical Engineering Science* **2011**, *66* (24), 6318-6323. DOI: 10.1016/j.ces.2011.02.050.
(14) Vorotnikov, V.; Vlachos, D. G. Group Additivity and Modified Linear Scaling Relations for Estimating Surface Thermochemistry on Transition Metal Surfaces: Application to Furanics. *The Journal of Physical Chemistry C* **2015**, *119* (19), 10417-10426. DOI: 10.1021/acs.jpcc.5b01696.

(15) Greeley, J. Theoretical Heterogeneous Catalysis: Scaling Relationships and Computational Catalyst Design. *Annu Rev Chem Biomol Eng* **2016**, *7*, 605-635. DOI: 10.1146/annurev-chembioeng-080615-034413.
(16) Andersen, M.; Levchenko, S. V.; Scheffler, M.; Reuter, K. Beyond Scaling Relations for the Description of Catalytic Materials. *ACS Catalysis* **2019**, *9* (4), 2752-2759. DOI: 10.1021/acscatal.8b04478.
(17) Calle-Vallejo, F.; Loffreda, D.; Koper, M. T.; Sautet, P. Introducing structural sensitivity into adsorption-energy scaling relations by means of coordination numbers. *Nat Chem* **2015**, *7* (5), 403-410. DOI: 10.1038/nchem.2226.
(18) Li, Z.; Ma, X.; Xin, H. Feature engineering of machine-learning chemisorption models for catalyst design. *Catalysis Today* **2017**, *280*, 232-238. DOI: 10.1016/j.cattod.2016.04.013.
(19) Benson, S. W.; Buss, J. H. Additivity Rules for the Estimation of Molecular Properties. Thermodynamic Properties. *The Journal of Chemical Physics* **1958**, *29* (3), 546-572. DOI: 10.1063/1.1744539.
(20) Eigenmann, H. K.; Golden, D. M.; Benson, S. W. Revised group additivity parameters for the enthalpies of formation of oxygen-containing organic compounds. *The Journal of Physical Chemistry* **1973**, *77* (13), 1687-1691. DOI: 10.1021/j100632a019.
(21) Gu, G. H.; Vlachos, D. G. Group Additivity for Thermochemical Property Estimation of Lignin Monomers on Pt(111). *The Journal of Physical Chemistry C* **2016**, *120* (34), 19234-19241. DOI: 10.1021/acs.jpcc.6b06430.
(22) Guo, N.; Caratzoulas, S.; Doren, D. J.; Sandler, S. I.; Vlachos, D. G. A perspective on the modeling of biomass processing. *Energy & Environmental Science* **2012**, *5* (5), 6703. DOI: 10.1039/c2ee02663e.
(23) Salciccioli, M.; Chen, Y.; Vlachos, D. G. Density Functional Theory-Derived Group Additivity and Linear Scaling Methods for Prediction of Oxygenate Stability on Metal Catalysts: Adsorption of Open-Ring Alcohol and Polyol Dehydrogenation Intermediates on Pt-Based Metals. *The Journal of Physical Chemistry C* **2010**, *114* (47), 20155-20166. DOI: 10.1021/jp107836a.
(24) Rupp, M.; Tkatchenko, A.; Muller, K. R.; von Lilienfeld, O. A. Fast and accurate modeling of molecular atomization energies with machine learning. *Phys Rev Lett* **2012**, *108* (5), 058301. DOI: 10.1103/PhysRevLett.108.058301.
(25) Rupp, M.; Ramakrishnan, R.; von Lilienfeld, O. A. Machine Learning for Quantum Mechanical Properties of Atoms in Molecules. *The Journal of Physical Chemistry Letters* **2015**, *6* (16), 3309-3313. DOI: 10.1021/acs.jpclett.5b01456.
(26) Huang, B.; von Lilienfeld, O. A. Communication: Understanding molecular representations in machine learning: The role of uniqueness and target similarity. *J Chem Phys* **2016**, *145* (16), 161102. DOI: 10.1063/1.4964627.
(27) Hansen, K.; Biegler, F.; Ramakrishnan, R.; Pronobis, W.; von Lilienfeld, O. A.; Muller, K. R.; Tkatchenko, A. Machine Learning Predictions of Molecular Properties: Accurate Many-Body Potentials and Nonlocality in Chemical Space. *J Phys Chem Lett* **2015**, *6* (12), 2326-2331. DOI: 10.1021/acs.jpclett.5b00831.
(28) Wang, S.; Petzold, V.; Tripkovic, V.; Kleis, J.; Howalt, J. G.; Skulason, E.; Fernandez, E. M.; Hvolbaek, B.; Jones, G.; Toftelund, A.; et al. Universal transition state scaling relations for (de)hydrogenation over transition metals. *Phys Chem Chem Phys* **2011**, *13* (46), 20760-20765. DOI: 10.1039/c1cp20547a.

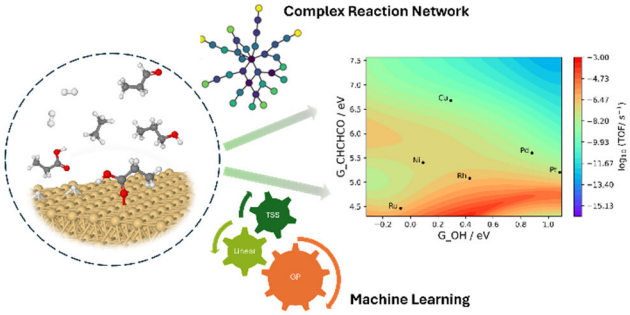
(29) van Santen, R. A.; Neurock, M.; Shetty, S. G. Reactivity theory of transition-metal surfaces: a Bronsted-Evans-Polanyi linear activation energy-free-energy analysis. *Chem Rev* **2010**, *110* (4), 2005-2048. DOI: 10.1021/cr9001808.
(30) Bligaard, T.; Nørskov, J. K.; Dahl, S.; Matthiesen, J.; Christensen, C. H.; Sehested, J. The Brønsted–Evans–Polanyi relation and the volcano curve in heterogeneous catalysis. *Journal of Catalysis* **2004**, *224* (1), 206-217. DOI: 10.1016/j.jcat.2004.02.034.
(31) Cheng, J.; Hu, P.; Ellis, P.; French, S.; Kelly, G.; Lok, C. M. Brønsted−Evans−Polanyi Relation of Multistep Reactions and Volcano Curve in Heterogeneous Catalysis. *The Journal of Physical Chemistry C* **2008**, *112* (5), 1308-1311. DOI: 10.1021/jp711191j.
(32) 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. *Catalysis Letters* **2010**, *141* (3), 370-373. DOI: 10.1007/s10562-010-0477-y.
(33) Wang, S.; Vorotnikov, V.; Sutton, J. E.; Vlachos, D. G. Brønsted–Evans–Polanyi and Transition State Scaling Relations of Furan Derivatives on Pd(111) and Their Relation to Those of Small Molecules. *ACS Catalysis* **2014**, *4* (2), 604-612. DOI: 10.1021/cs400942u.
(34) 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. *Journal of Catalysis* **2015**, *328*, 36-42. DOI: 10.1016/j.jcat.2014.12.033.
(35) Zaffran, J.; Michel, C.; Delbecq, F.; Sautet, P. Towards more accurate prediction of activation energies for polyalcohol dehydrogenation on transition metal catalysts in water. *Catalysis Science & Technology* **2016**, *6* (17), 6615-6624. DOI: 10.1039/c6cy00865h.
(36) Plessow, P. N.; Abild-Pedersen, F. Examining the Linearity of Transition State Scaling Relations. *The Journal of Physical Chemistry C* **2015**, *119* (19), 10448-10453. DOI: 10.1021/acs.jpcc.5b02055.
(37) 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. DOI: 10.1021/ja027366r.
(38) Zaffran, J.; Michel, C.; Auneau, F.; Delbecq, F.; Sautet, P. Linear Energy Relations As Predictive Tools for Polyalcohol Catalytic Reactivity. *ACS Catalysis* **2014**, *4* (2), 464-468. DOI: 10.1021/cs4010503.
(39) Motagamwala, A. H.; Ball, M. R.; Dumesic, J. A. Microkinetic Analysis and Scaling Relations for Catalyst Design. *Annu Rev Chem Biomol Eng* **2018**, *9*, 413-450. DOI: 10.1146/annurev-chembioeng-060817-084103.
(40) Sandberg, R. B.; Hansen, M. H.; Nørskov, J. K.; Abild-Pedersen, F.; Bajdich, M. Strongly Modified Scaling of CO Hydrogenation in Metal Supported TiO Nanostripes. *ACS Catalysis* **2018**, *8* (11), 10555-10563. DOI: 10.1021/acscatal.8b03327.
(41) Nørskov, J. K.; Bligaard, T.; Logadottir, A.; Bahn, S.; Hansen, L. B.; Bollinger, M.; Bengaard, H.; Hammer, B.; Sljivancanin, Z.; Mavrikakis, M.; et al. Universality in Heterogeneous Catalysis. *Journal of Catalysis* **2002**, *209* (2), 275-278. DOI: 10.1006/jcat.2002.3615.

(42) Loffreda, D.; Delbecq, F.; Vigne, F.; Sautet, P. Fast prediction of selectivity in heterogeneous catalysis from extended Bronsted-Evans-Polanyi relations: a theoretical insight. *Angew Chem Int Ed Engl* **2009**, *48* (47), 8978-8980. DOI: 10.1002/anie.200902800.
(43) Grabow, L. C.; Studt, F.; Abild-Pedersen, F.; Petzold, V.; Kleis, J.; Bligaard, T.; Norskov, J. K. Descriptor-based analysis applied to HCN synthesis from NH3 and CH4. *Angew Chem Int Ed Engl* **2011**, *50* (20), 4601-4605. DOI: 10.1002/anie.201100353.
(44) Nandy, A.; Zhu, J.; Janet, J. P.; Duan, C.; Getman, R. B.; Kulik, H. J. Machine Learning Accelerates the Discovery of Design Rules and Exceptions in Stable Metal–Oxo Intermediate Formation. *ACS Catalysis* **2019**, *9* (9), 8243-8255. DOI: 10.1021/acscatal.9b02165.
(45) Wang, Y.; Xiao, L.; Qi, Y.; Mahmoodinia, M.; Feng, X.; Yang, J.; Zhu, Y. A.; Chen. Towards rational catalyst design: boosting the rapid prediction of transition-metal activity by improved scaling relations. *Phys Chem Chem Phys* **2019**, *21* (35), 19269-19280. DOI: 10.1039/c9cp04286e.
(46) Guo, W.; Stamatakis, M.; Vlachos, D. G. Design Principles of Heteroepitaxial Bimetallic Catalysts. *ACS Catalysis* **2013**, *3* (10), 2248-2255. DOI: 10.1021/cs4005166.
(47) Deshlahra, P.; Iglesia, E. Reactivity and Selectivity Descriptors for the Activation of C–H Bonds in Hydrocarbons and Oxygenates on Metal Oxides. *The Journal of Physical Chemistry C* **2016**, *120* (30), 16741-16760. DOI: 10.1021/acs.jpcc.6b04604.
(48) Jaraíz, M.; Rubio, J. E.; Enríquez, L.; Pinacho, R.; López-Pérez, J. L.; Lesarri, A. An Efficient Microkinetic Modeling Protocol: Start with Only the Dominant Mechanisms, Adjust All Parameters, and Build the Complete Model Incrementally. *ACS Catalysis* **2019**, *9* (6), 4804-4809. DOI: 10.1021/acscatal.9b00522.
(49) Ulissi, Z. W.; Medford, A. J.; Bligaard, T.; Norskov, J. K. To address surface reaction network complexity using scaling relations machine learning and DFT calculations. *Nat Commun* **2017**, *8*, 14621. DOI: 10.1038/ncomms14621.
(50) Sutton, J. E.; Vlachos, D. G. Building large microkinetic models with first-principles′ accuracy at reduced computational cost. *Chemical Engineering Science* **2015**, *121*, 190-199. DOI: 10.1016/j.ces.2014.09.011.
(51) Lee, K.; Gu, G. H.; Mullen, C. A.; Boateng, A. A.; Vlachos, D. G. Guaiacol hydrodeoxygenation mechanism on Pt(111): insights from density functional theory and linear free energy relations. *ChemSusChem* **2015**, *8* (2), 315-322. DOI: 10.1002/cssc.201402940.
(52) Sutton, J. E.; Vlachos, D. G. A Theoretical and Computational Analysis of Linear Free Energy Relations for the Estimation of Activation Energies. *ACS Catalysis* **2012**, *2* (8), 1624-1634. DOI: 10.1021/cs3003269.
(53) Donnis, B.; Egeberg, R. G.; Blom, P.; Knudsen, K. G. Hydroprocessing of Bio-Oils and Oxygenates to Hydrocarbons. Understanding the Reaction Routes. *Topics in Catalysis* **2009**, *52* (3), 229-240. DOI: 10.1007/s11244-008-9159-z.
(54) Huber, G. W.; O'Connor, P.; Corma, A. Processing biomass in conventional oil refineries: Production of high quality diesel by hydrotreating vegetable oils in heavy vacuum oil mixtures. *Applied Catalysis A: General* **2007**, *329*, 120-129. DOI: 10.1016/j.apcata.2007.07.002.

(55) Wildschut, J.; Mahfud, F. H.; Venderbosch, R. H.; Heeres, H. J. Hydrotreatment of Fast Pyrolysis Oil Using Heterogeneous Noble-Metal Catalysts. *Industrial & Engineering Chemistry Research* **2009**, *48* (23), 10324-10334. DOI: 10.1021/ie9006003.
(56) Elliott, D. C. Historical Developments in Hydroprocessing Bio-oils. *Energy & Fuels* **2007**, *21* (3), 1792-1815. DOI: 10.1021/ef070044u.
(57) Holtzapple, M. T.; Wu, H.; Weimer, P. J.; Dalke, R.; Granda, C. B.; Mai, J.; Urgun-Demirtas, M. Microbial communities for valorizing biomass using the carboxylate platform to produce volatile fatty acids: A review. *Bioresour Technol* **2022**, *344* (Pt B), 126253. DOI: 10.1016/j.biortech.2021.126253.
(58) Kresse, G.; Joubert, D. From ultrasoft pseudopotentials to the projector augmented-wave method. *Physical Review B* **1999**, *59* (3), 1758-1775. DOI: 10.1103/PhysRevB.59.1758.
(59) Kresse, G.; Hafner, J. Ab initio molecular dynamics for liquid metals. *Phys Rev B Condens Matter* **1993**, *47* (1), 558-561. DOI: 10.1103/physrevb.47.558.
(60) Kresse, G.; Furthmüller, J. Efficiency of ab-initio total energy calculations for metals and semiconductors using a plane-wave basis set. *Computational Materials Science* **1996**, *6* (1), 15-50. DOI: 10.1016/0927-0256(96)00008-0.
(61) Perdew, J. P.; Wang, Y. Accurate and simple analytic representation of the electron-gas correlation energy. *Phys Rev B Condens Matter* **1992**, *45* (23), 13244-13249. DOI: 10.1103/physrevb.45.13244.
(62) Perdew, J. P.; Yue, W. Accurate and simple density functional for the electronic exchange energy: Generalized gradient approximation. *Phys Rev B Condens Matter* **1986**, *33* (12), 8800-8802. DOI: 10.1103/physrevb.33.8800.
(63) Lu, J.; Faheem, M.; Behtash, S.; Heyden, A. Theoretical investigation of the decarboxylation and decarbonylation mechanism of propanoic acid over a Ru(0 0 0 1) model surface. *Journal of Catalysis* **2015**, *324*, 14-24. DOI: 10.1016/j.jcat.2015.01.005.
(64) Behtash, S.; Lu, J.; Williams, C. T.; Monnier, J. R.; Heyden, A. Effect of Palladium Surface Structure on the Hydrodeoxygenation of Propanoic Acid: Identification of Active Sites. *The Journal of Physical Chemistry C* **2015**, *119* (4), 1928-1942. DOI: 10.1021/jp511618u.
(65) Faheem, M.; Saleheen, M.; Lu, J.; Heyden, A. Ethylene glycol reforming on Pt(111): first-principles microkinetic modeling in vapor and aqueous phases. *Catalysis Science & Technology* **2016**, *6* (23), 8242-8256. DOI: 10.1039/c6cy02111e.
(66) Haworth, N. L.; Wang, Q.; Coote, M. L. Modeling Flexible Molecules in Solution: A pKa Case Study. *J Phys Chem A* **2017**, *121* (27), 5217-5225. DOI: 10.1021/acs.jpca.7b04133.
(67) Monkhorst, H. J.; Pack, J. D. Special points for Brillouin-zone integrations. *Physical Review B* **1976**, *13* (12), 5188-5192. DOI: 10.1103/PhysRevB.13.5188.
(68) Henkelman, G.; Uberuaga, B. P.; Jonsson, H. A climbing image nudged elastic band method for finding saddle points and minimum energy paths. *Journal of Chemical Physics* **2000**, *113* (22), 9901-9904. DOI: Pii [S0021-9606(00)71246-3] Doi 10.1063/1.1329672.
(69) Henkelman, G.; Jonsson, H. A dimer method for finding saddle points on high dimensional potential surfaces using only first derivatives. *Journal of Chemical Physics* **1999**, *111* (15), 7010-7022. DOI: Doi 10.1063/1.480097.

(70) Olsen, R. A.; Kroes, G. J.; Henkelman, G.; Arnaldsson, A.; Jonsson, H. Comparison of methods for finding saddle points without knowledge of the final states. *J Chem Phys* **2004**, *121* (20), 9776-9792. DOI: 10.1063/1.1809574.
(71) Heyden, A.; Bell, A. T.; Keil, F. J. Efficient methods for finding transition states in chemical reactions: comparison of improved dimer method and partitioned rational function optimization method. *J Chem Phys* **2005**, *123* (22), 224101. DOI: 10.1063/1.2104507.
(72) Yang, W.; Solomon, R. V.; Lu, J.; Mamun, O.; Bond, J. Q.; Heyden, A. Unraveling the mechanism of the hydrodeoxygenation of propionic acid over a Pt (1 1 1) surface in vapor and liquid phases. *Journal of Catalysis* **2020**, *381*, 547-560. DOI: 10.1016/j.jcat.2019.11.036.
(73) Lu, J.; Behtash, S.; Faheem, M.; Heyden, A. Microkinetic modeling of the decarboxylation and decarbonylation of propanoic acid over Pd(111) model surfaces based on parameters obtained from first principles. *Journal of Catalysis* **2013**, *305*, 56-66. DOI: 10.1016/j.jcat.2013.04.026.
(74) Rajbanshi, B.; Yang, W.; Yonge, A.; Kundu, S. K.; Fricke, C.; Heyden, A. Computational Investigation of the Catalytic Hydrodeoxygenation of Propanoic Acid over a Cu(111) Surface. *The Journal of Physical Chemistry C* **2021**, *125* (35), 19276-19293. DOI: 10.1021/acs.jpcc.1c05240.
(75) Yang, W.; Solomon, R. V.; Mamun, O.; Bond, J. Q.; Heyden, A. Investigation of the reaction mechanism of the hydrodeoxygenation of propionic acid over a Rh(1 1 1) surface: A first principles study. *Journal of Catalysis* **2020**, *391*, 98-110. DOI: 10.1016/j.jcat.2020.08.015.
(76) Anderson, J. S.; Cutsail, G. E., 3rd; Rittle, J.; Connor, B. A.; Gunderson, W. A.; Zhang, L.; Hoffman, B. M.; Peters, J. C. Characterization of an Fe identical withN-NH2 Intermediate Relevant to Catalytic N2 Reduction to NH3. *J Am Chem Soc* **2015**, *137* (24), 7803-7809. DOI: 10.1021/jacs.5b03432.
(77) Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences* **2002**, *28* (1), 31-36. DOI: 10.1021/ci00057a005.
(78) Chowdhury, A. J.; Yang, W.; Walker, E.; Mamun, O.; Heyden, A.; Terejanu, G. A. Prediction of Adsorption Energies for Chemical Species on Metal Catalyst Surfaces Using Machine Learning. *The Journal of Physical Chemistry C* **2018**, *122* (49), 28142-28150. DOI: 10.1021/acs.jpcc.8b09284.
(79) 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. *The Journal of Physical Chemistry C* **2019**, *123* (49), 29804-29810. DOI: 10.1021/acs.jpcc.9b10507.
(80) Campbell, C. T. The Degree of Rate Control: A Powerful Tool for Catalysis Research. *ACS Catalysis* **2017**, *7* (4), 2770-2779. DOI: 10.1021/acscatal.7b00115.
(81) Stegelmann, C.; Andreasen, A.; Campbell, C. T. Degree of rate control: how much the energies of intermediates and transition states control rates. *J Am Chem Soc* **2009**, *131* (23), 8077-8082. DOI: 10.1021/ja9000097.
(82) Kundu, S. K.; Vijay Solomon, R.; Yang, W.; Walker, E.; Mamun, O.; Bond, J. Q.; Heyden, A. Surface structure sensitivity of hydrodeoxygenation of biomass-derived

organic acids over palladium catalysts: a microkinetic modeling approach. *Catalysis Science & Technology* **2021**, *11* (18), 6163-6181. DOI: 10.1039/d1cy01029h.
(83) Gopeesingh, J.; Zhu, R.; Schuarca, R.; Yang, W.; Heyden, A.; Bond, J. Q. Kinetic and Mechanistic Analysis of the Hydrodeoxygenation of Propanoic Acid on Pt/SiO2. *Industrial & Engineering Chemistry Research* **2021**, *60* (45), 16171-16187. DOI: 10.1021/acs.iecr.1c03032.
(84) Cao, A.; Nørskov, J. K. Spin Effects in Chemisorption and Catalysis. *ACS Catalysis* **2023**, *13* (6), 3456-3462. DOI: 10.1021/acscatal.2c06319.


Complex Reaction Network
G_CHCHCO / eV
G_OH / eV
Cu
Pd
Ni
Pt
Rh
Ru
TSS
Linear
GP
Machine Learning

TOC