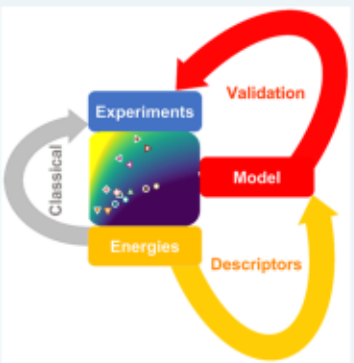
This document is the Accepted Manuscript version of a Published Work that appeared in final form in ACS Catalysis, copyright © American Chemical Society after peer review and technical editing by the publisher. To access the final edited and published work see https://pubs.acs.org/doi/-abs/10.1021/acscatal.1c04345

Generalizing Performance Equations in Heterogeneous Catalysis from Hybrid Data and Statistical Learning

Sergio Pablo-García,[▽] Albert Sabadell-Rendón,[▽] Ali J. Saadun, Santiago Morandi, Javier Pérez-Ramírez,* and Núria López*

ABSTRACT: Activity equations trying to mimic experimental catalytic performance derived from reaction profiles and microkinetic models have been the state of the art in modeling in the last decades. This approach has been able to reproduce semiquantitatively activity volcano plots leading to successful catalyst optimization through the use of descriptors. As systems become more complex (both catalysts and reactants), these methods face increasing limitations. Statistical Learning (SL) techniques can overcome these limitations and improve the search for descriptor-based performance equations. However, the black-box nature of SL techniques makes physical interpretation of the so-obtained models difficult. To advance in the integration of these methodologies to real problems, we have merged experimental activity and selectivity presented as a function of chemical descriptors from Density Functional Theory for the catalyzed hydrodehalogenation of CH_2X_2 (for X = Br, Cl) leading to three main products. The employed Bayesian procedure is able to identify robust equations for activity and selectivity as a function of only two descriptors. This work provides a starting point to solve complex reaction networks using a set of statistical learning tools and hybrid data.



KEYWORDS: *Kinetics, statistical learning, microkinetics, Density Functional Theory, activity, selectivity, descriptors*

■ INTRODUCTION

Equations describing catalytic performance are at the core of industrial heterogeneous catalysis.[1] Stock and Bodenstein found the first kinetic power law rate by investigating reaction rates at variable pressure.[1] This heuristic approach was soon complemented by the mechanistic insight developed by Langmuir, suggesting the existence of equivalent noninteracting active sites on the surface, which evolved into the so-called Langmuir−Hinshelwood−Hougen−Watson (LHHW) kinetics. In this approach, a reaction was described by a set of elementary steps, the rate of which was derived from a LHHW equation. The LHHW rate predictions were compared to the experimentally obtained kinetics data and if the LHHW model was able to fit that data, the mechanism was considered meaningful. As such, the equation was deemed useful to design a reactor for the application at the specified operating conditions; yet, the model is too ideal. Alternatives considering anisotropy were introduced by Temkin,[2] with limited success.

Microkinetic modeling (MK) emerged in the second half of the 20th century.[3,4] In the MK procedure, a mechanism was proposed, and each step had an associated kinetic coefficient k, defining a system of ordinary differential equations (ODEs) with initial conditions. The one-to-one mapping to experiments provided estimates for the kinetic coefficients and equilibrium constants (K) for the steps in the mechanism and the time evolution of intermediates, either as concentration or coverage. Several alternative mechanisms could be tested following this procedure and the quality of the fitting and the reliability of the K and k parameters served to retain or discard a proposed mechanism. The emergence of massive atomistic simulations based on Density Functional Theory (DFT) allowed to investigate reaction profiles in model systems (particularly metals), and changed the traditional use of MK procedures. DFT finds the minimum energy configuration for intermediates on surfaces and the transition states linking the elementary steps. Thus, the corresponding reaction energies, ΔE and activation barriers, E_a could be employed coupled to statistical thermodynamics to obtain estimates for the kinetic coefficients of each elementary step[5] through the Arrhenius and Eyring equations and even for adsorption/desorption processes, where pressure (Hertz−Knudsen) is the relevant variable. Microkinetic models assumptions imply that the number of sites is preserved all through the reaction and the reactivity in all these centers is identical. However, since MK is a mean-field approach, it fails to fully describe highly anisotropic systems in which directionality is crucial. Consequently, even higher-resolution methods, such as kinetic

Monte Carlo (kMC), are required instead. In the kMC approach, all possible steps are simulated on a surface (lattice) explicitly using a stochastic procedure, in which the probability of each process to occur is estimated according to the corresponding k.[6,7] Of course, the more elementary steps are introduced in the kMC simulations, the larger the number of explicit DFT evaluations is needed; therefore, in nonobvious cases, kMC becomes impractical, although recent efforts in code parallelization could improve the predictions.[8] MK and kMC are used to predict macroscopic parameters such as turnover frequency (TOF),[9−18] selectivity,[9,10,16−20] apparent activation energies,[18,19,21−26] and reaction orders.[10,12,19,23,26,27] In addition, from the MK and kMC results, analysis of the most relevant steps can be performed through the degree of rate control (DRC),[14,19,28−31] and the degree of selectivity control (DSC). From DRC and DSC, simplified model equations, which are known as surrogate models, can be inferred as employing exponential dependences to key steps and then expanding the first term of the Taylor expansion of Arrhenius equation to make them linear.[32] In practice, MK and kMC require intensive path investigation to identify the mechanism and some fine-tuning of some of the DFT computed parameters in order to be directly mapped to experimental results for complex systems.[33−36] The consequence is that, in reactor design, power-law equations have been the functions of choice to represent experimental rates. Under this approach, similar functional forms suggest identical mechanisms,[37] whereas the reaction orders are often linked to the participation of a given species in the reaction network, particularly before the rate-determining step. These equations have been employed in the chemical industry to find suitable operation conditions for over a century.

Catalyst development has been linked to the Sabatier principle,[38,39] which is the "not-too-strong-not-too-weak" rule for the optimization within a family of catalytic materials. This rule was formulated deriving in the volcano-shaped functions of a single energy value acting as a catalytic activity descriptor that populate the research works in heterogeneous catalysis. Actually, this generalization of the catalytic activity within a family of catalysts or a family of reactants is rooted on the phenomenological observations by Hammett[40] and Hammond postulates,[41] and Brønsted−Evans−Polanyi (BEP)[42,43] equations. Again, atomistic simulations based on DFT unraveled the nature of these dependencies through the so-called linear-scaling relationships (LSRs), linking thermodynamics to thermodynamics (adsorption of an intermediate to adsorption of the central heteroatom) and thermodynamics to kinetics.[23,36,44−51] In summary, on a metal surface (the most commonly investigated catalyst family), the binding energy of a simple intermediate can be traced back to another with the same heteroatom. The transition state for an elementary step can be linked to initial and final intermediates in a similar manner. Thus, the reaction energies and activation barriers can be estimated only by knowing the slopes for these linear dependencies and the energy parameter of the descriptor.[44,52,53] The dependences can be directly plugged as equations in the MK or kMC modeling, reducing the dimensionality of the catalytic performance problem and leading to computed activity volcanos. In summary, activity volcano plots are obtained from ab initio principles with DFT[37,52] data of a single catalysts, from which the full reaction network is computed. Typically, the energies of the intermediates and transition states are condensed through LSR and the transition state theory to provide the kinetic coefficients, after which they are embedded in the microkinetic models.

This approach has been the gold standard in heterogeneous catalysis[37,54] since the beginning of the 21st century. Catalytic phase changes, dynamics, and deposits of poisons are limiting the progress in the field, because they are not properly taken into account in microkinetic models. Thus, accumulated uncertainties make predictions less accurate when increasing in complexity,[55] which is mainly caused by (i) coverage effects with concomitant adsorbate reorientation or adsorbate energy changes; (ii) catalyst phase or surface transformations; (iii) large reaction networks, where elementary steps grow exponentially;[23] and (iv) highly dynamic materials that have various possible configurations.[56] Consequently, descriptors are chosen based on, to some extent, arbitrary heuristics,[57−59] and the computed MK(DFT) (or kMC(DFT)) rates are still far from experimental values and fail in describing key experimental observables such as selectivity.[59] Recent studies show that the energy profiles require corrections as large as 0.5 eV, with respect to the computed DFT values for some intermediates (the asymmetry leads to further difficulties in assigning meanings to these corrections) to explain experimental observations, particularly in hydrodechlorination processes.[60,61]

In summary, the inference of robust and accurate mathematical expressions to predict activity and product selectivity and finding generally applicable rates within a given family of compounds, are far from obvious. Pioneering approaches employing hybrid data (i.e., from experiments and DFT) have been applied to hydrogen evolution reaction (HER) on transition-metal surfaces,[62] or the methanation reaction on alloys,[63] thus rather simple model catalysts. In these cases, LSRs were employed to correlate experimental activity to descriptors based on physical intuition rather than in statistics and are accordingly difficult to generalize.[64] Since large experimental and computational datasets[65] are being made available through high-throughput techniques, the systematically generated data is amenable for statistical learning (SL) treatments.[64,66−74] Generally, the introduction of data approaches has been steered from the DFT community. Thus, approaches combining DFT and SL have been recently introduced in (i) the derivation of approaches to simplify the DFT computational burden of calculating the energies of many configurations;[75−78] (ii) applying these energies to traditionally MK(DFT) derived volcanoes, providing candidates for electrochemical processes; and (iii) searching for optimized descriptors employed in theoretical studies.[78,79] In parallel, attempts to link experimental catalytic performance can be found in recent literature (for example, in the work of Foppa et al., ∼40 tabulated experimental observables were used to predict the annotated consistent conversion and selectivity of nine vanadium-based catalysts using a symbolic regression protocol (SISSO) to derive a nontrivial ensemble of models[80]).

Among the most relevant procedure for catalytic optimization, descriptors identification, equivalent to dimensionality reduction in SL, has been the most investigated via t-distributed stochastic neighbor embedding (t-SNE),[81] least absolute shrinkage and selection operator (LASSO),[82] and principal component analysis (PCA).[79,83] However, many of these methods are hard to interpret and the optimization parameters are difficult to map to macroscopic variables.

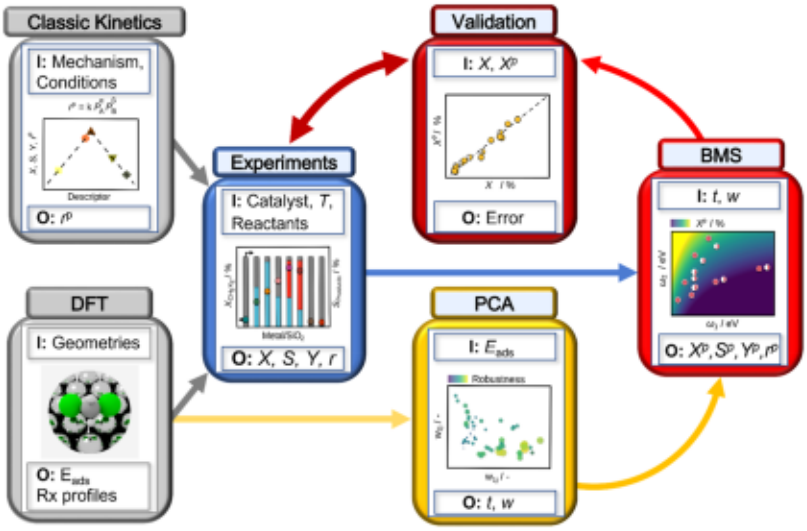


Figure 1. Workflow of the present study, where gray arrows represent classical approaches to catalysis modeling. Colored pathways indicate the route and methods applied for analyzing conversion (X), selectivity (S), yield (Y), and rate (r) of CH_2X_2 hydrodehalogenation. Predicted values of the experimental observables are denoted with the superindex p (as example: observable X, prediction X^p).

Bayesian techniques hold the key to solving many of the interpretability and optimization issues of experimental variables and uncertainties quantification.[84−86] Attempts to generate modeling equations for spectroscopic, spectrometric, and chromatographic purposes,[87] using experimental-only datasets has been performed using Bayesian Statistics. In electrochemical modeling, Bayesian techniques and Gaussian regressor have been employed to predict best compositions for the oxygen reduction reaction in binary and ternary alloys from hybrid data.[88] Functional exploration has been recently been proposed for some properties linking to material degradation.[72,89] In our case, the search for symbolic equations to build parsimonious models, we have employed the Bayesian Machine Scientist (BMS),[90] which allows functional exploration using Markov Chain Monte Carlo.

Product selectivity in homogeneous catalysis has been analyzed through various SL techniques, where, in some studies, computed parameters were taken as variables,[83,91−93] even with small set of experimental data.[94] This type of heuristics with hybrid data has not been attempted in heterogeneous catalysis, which have a "space of ligands", defined as a set of atoms or molecules able to be attached to the surrounding of the active metal centers, less continuous than homogeneous systems. Because of this, the experimental data is sparser, observations are fewer, and the characterization must account for larger space and time scales. In our first attempt to leverage hybrid data approaches to heterogeneous catalysis, we reproduced the conversion of CH_2Br_2 in metal-catalyzed (Fe, Co, Ni, Cu, Ru, Rh, Ag, Ir, and Pt) hydrodebromination using SL techniques, and demonstrated the stability and robustness of the employed algorithms.[95] The hydrodehalogenation reactions constitute a benchmark for SL techniques as the classical MK(DFT) modeling fails due to phase and surface changes (see below), as carbides form for Co (halides for Cu or Ag). In addition, the experimental observation of three different products, some of them ill-defined composition as coke makes its estimation even more difficult. Therefore, setting a robust framework for the use of SL techniques on a technologically challenging reaction for a family of compounds can help us set the key questions, like equation transferability, generalization to a family of reactants, and robustness of the functional forms.

In order to develop a robust framework for the deployment of SL with the aim to obtain reactivity equations of heterogeneous catalysts, we have used the hydrodehalogenation of CH_2X_2 (X = Br, Cl) and compared them to MK(DFT) models. This transformation is a key step in halogen-mediated methane upgrading processes and clearly displays the selectivity issues of the metal catalysts.[95−98] By combining the descriptor identification from PCA and BMS, we present a unique equation search for the reaction rate, CH_2X_2 conversion, and selectivity to different products, thus compiling the performance of the metal catalyst in a generalized form (procedure depicted in Figure 1). Our approach links experimental and theoretical data and sets a robust methodology that could be extrapolated to other heterogeneously catalyzed reactions.

MATERIALS AND METHODS

Density Functional Theory and Microkinetic Model Setup. DFT implemented in the Vienna Ab Initio Simulation Package (VASP 5.4.4)[99] was employed to describe the chemical systems involved in this work. The exchange-correlation energies were obtained using the Generalized Gradient Approximation with the Perdew−Burke−Ernzerhof (GGA PBE-D2) functional,[100,101] reparametrizing the C_6 values for the metals.[102] The inner electrons were represented using the projected-augmented wave (PAW) and the valence electrons by plane waves with a cutoff energy of 450 eV.[103] The Monkhorst−Pack method was used to create the Γ-centered k-point mesh.[104] A four-layer $p(3 \times 3)$-(111) face-centered-cubic slab (fcc) was used to model the metallic surfaces, except for Co and Ru, for which a $p(3 \times 3)$-(0001) body centered cubic slab (hcp) was used. A box of 15 Å × 15 Å × 15 Å was used to model the molecules in gas phase. The vacuum of the slab was implemented with a space in the cell of 15 Å in the z-direction, including the dipole correction due to the asymmetry of the system.[105] For the electronic and ionic relaxations, the threshold criteria were set as 10^{-5} eV and 0.03 eV Å^{-1}, respectively. During optimization, the two upper metal

layers and the adsorbates were allowed to relax, while the rest of the atoms were fixed. The stability of the Cl, F, and I halogenated intermediates was checked via a frequency analysis, while Br and nonhalogenated structures and energies were retrieved from ioChem-BD (10.19061/iochem-bd-1-150 and 10.19061/iochem-bd-1-152).[65,95] Further technical details on DFT simulations can be found in the Supporting Information (SI) (Note S1 and eqs S1 and S2).

In the microkinetic model (mechanism in Figure S1 and Table S1 in the SI), first we assessed the thermodynamic consistency (Table S4 in the SI) and analyzed the reversibility for all metals (Table S5 in the SI). The results of the thermodynamic consistency analysis are in a good agreement with those reported from the third millennium database.[106] The microkinetic runs for all metals were performed using a differential reactor model under the experimental reaction conditions: T = 523 K, P = 1 atm (70% inert gases), initial CH_2Br_2:H_2 ratio = 1:4. Further details on the microkinetic model can be found in Note S2, eqs S3−S7, and Figures S2−S7 in the SI.

Catalyst Preparation. SiO_2-supported metal catalysts were synthesized following the protocol reported by some of us.[95] Commercial SiO_2 (Evonik, AEROPERL 300/30, S_{BET} = 257 $m^2\ g^{-1}$, V_{pore} = 0.95 $cm^3\ g^{-1}$, >99.0%) was calcined at 973 K for 5 h in static air (heating rate = 5 K min^{-1}) prior to its use as support in the catalyst preparation (see Note S3 in the SI for further information). The resulting solids were dried at 373 K for 12 h and optionally calcined in static air at 623 K (heating rate = 5 K min^{-1}) to obtain the SiO_2-supported metal oxides. Subsequently, all samples underwent a reductive treatment in 20 vol % H_2/He (PanGas, purity 5.0) flow for 3 h (heating rate of 10 K min^{-1}) at 573−968 K prior to their use in catalytic tests. The catalysts were referenced as M/SiO_2, where M denotes the metal (i.e., Co, Ni, Cu, Ru, Rh, Ag, Ir, or Pt).

Catalyst Testing. The hydrodehalogenation of the dihalomethanes (CH_2X_2, X = Cl, Br) was performed at ambient pressure in a homemade continuous-flow fixed-bed reactor set up. H_2 (PanGas, purity 5.0), He (Carrier gas, PanGas, purity 5.0), Ar (internal standard, PanGas, purity 5.0) were supplied by a set of digital mass flow controllers (Bronkhorst) and liquid CH_2Br_2 (Acros Organics, 99%) or CH_2Cl_2 (Sigma−Aldrich, >99.9%) was dosed by a syringe pump (Fusion 100, Chemyx) equipped with a water-cooled syringe to a vaporizer unit operated at 393 K. The quartz reactor (internal diameter, d_i = 12 mm) containing the reduced catalyst (catalyst weight, W_{cat} = 0.1−1 g, particle size, d_p = 0.4−0.6 mm) was heated to the desired temperature (T = 423−623 K) in an electrical oven under He flow. The catalyst bed was allowed to stabilize for at least 10 min at desired temperature before the reaction mixture was fed at a total volumetric flow (F_T) of 20−350 cm^3 STP min^{-1} and desired feed composition of CH_2X_2:H_2:Ar:He = 6:24:4.5:65.5 (vol %, X = Cl, Br). Downstream linings were heated at 393 K to prevent the condensation of unconverted reactants and/or products. The content of carbon containing compounds (CH_2X_2, CH_3X, CH_4, C_2H_4, C_2H_6, C_3H_6, and C_3H_8) and of Ar in the reactor outlet gas stream was quantified online via a gas chromatograph equipped with a GS Carbon PLOT column coupled to a mass spectrometer (GC MS, Agilent GC 6890, Agilent MSD 5973N). After the GC MS analysis, the gas stream was passed through two impinging bottles in series containing an aqueous solution of NaOH (1 M) for neutralization prior to its release in the ventilation system.

The conversion of the dihalomethane, $X_{CH_2X_2}$, was calculated using eq 1:

$$X_{CH_2X_2}\ (\%) = \frac{n(CH_2X_2)_{in} - n(CH_2X_2)_{out}}{n(CH_2X_2)_{in}} \times 100 \quad (1)$$

where $n(CH_2X_2)_{in}$ and $n(CH_2X_2)_{out}$ are the molar flows of the reactant at the reactor inlet and outlet, respectively. The selectivity (S_i) to product i (where i = CH_3X, CH_4, C_2H_4, C_2H_6, C_3H_6, and C_3H_8) was calculated according to eq 2:

$$S_j\ (\%) = \frac{\sigma \cdot n(j)_{out}}{n(CH_2X_2)_{in} - n(CH_2X_2)_{out}} \times 100 \quad (2)$$

where $n(j)_{out}$ is the molar flow of product j at the reactor outlet. To take the number of carbon atoms in the products into account, σ = 1, 2, and 3, for products C_1, C_2, and C_3, respectively. The error of the carbon balance, ε_C, used to specify the selectivity to coke, was determined using eq 3:

$$\varepsilon_c\ (\%) = \frac{n(CH_2X_2)_{in} - n(CH_2X_2)_{out} - \sigma \cdot n(j)_{out}}{n(CH_2X_2)_{in}} \times 100 \quad (3)$$

Evaluation of the dimensionless moduli based on the criteria of Carberry, Mears, and Weisz−Prater[107,108] indicated that the catalytic tests were performed in the absence of mass- and heat-transfer limitations.

Statistical Analysis Details. The Random Forest Regressor (RF) and Gaussian Process Regressor (GR) were applied using the Scikit-Learn (0.23.1) package.[109] PCA and BMS were implemented according to our previous work.[79,90,95] The RF was applied with 150 estimator trees with a maximum depth of 3 and optimized using the mean squared error and by bootstrapping the samples. For the GR, the radial-basis function kernel was chosen with a noise parameter (α) of 10^{-5} (see Note S4, eqs S8−S12, and Figure S8 in the SI). The reasons to apply the RF are (i) RF divides the variable space in regions according to the response magnitude (how large is the response according to the variables), and (ii) it presents reasonably accurate predictions on the explored zones. The choice of the GR is based on the following: (i) GR can separate the variable space in regions and the predictions are more accurate than those from the Random Forest, (ii) volcano plots have a similar shape as the Gaussian regressor output (a joint-Gaussian distribution), and (iii) the Gaussian regressor, as the BMS, is based on a Bayesian process.

For each run of the BMS for the two-variables dataset, we choose a priori corresponding to two variables and two parameters (two fitting constants), and a maximum of 10 000 steps (the first 1000 steps were removed), with a thinning of 100 was set. For each nonburnt step, the BMS exported: (i) the complexity of the current function, (ii) the Bayesian Information Criterion (BIC) of the current function, (iii) the Sum of Squared Errors (SSE) of the current function, (iv) the values of the fitting constants, and the function itself. Each run required 1−2 days. The simplest equations then were refitted using the same fitting constants obtained by the BMS to check the output accuracy. Finally, the result was plotted and visually examined. The number of replicates needed for finding appropriate functional forms is dependent on the observable: for conversion and selectivity to CH_4, two replicates were enough, whereas for the selectivity to coke, up to 20 runs were required. The accuracy of the BMS, RF, and GR techniques is

dependent on the observable, but as a general rule, their SSE values follows the following trend: SSE(GR) < SSE(BMS) < SSE(RF). However, the SSE values for selectivity to coke and CH_3X of BMS (510 and 856, respectively) are much better than those obtained with GR (2925 and 3745, respectively).

RESULTS AND DISCUSSION

The approach followed in the present work is illustrated in Figure 1; traditional methods (marked in gray) take advantage of the use of reaction profiles and microkinetic modeling to try to obtain rates and other parameters comparable to experiments. The alternative path (indicated by colors) shows the potential of SL techniques containing hybrid data for the analysis of the performance descriptors of complex catalytic systems.

Catalytic gas-phase hydrodehalogenations is a family of reactions in which halogen elimination from an organic compound is followed by hydrogen addition. The selective transformation of CH_2X_2 into CH_3X is of particular interest, since it is an important step in halogen-mediated natural gas upgrading processes,[95−98] although it presents selectivity issues. In the reaction mechanism, as shown in Figure S1 in the SI, the target CH_3X is formed by the removal of a single halogen and the addition of an H atom. Alternatively, both halides are eliminated and followed by the sequential addition of H and halogen atoms. Coke is formed as a side product from the CH_2 intermediate, whereas CH_4 is generated after hydrogenation of the CH_3 species.

Density Functional Theory Data. The DFT (PBE-D2, see the "Materials and Methods" section) dataset containing the geometries and adsorption energies of 74 intermediates derived from CH_2X_2 (for X = F, Cl, Br, and I) on 8 metallic surfaces (total of 592) is compiled as a comma separated values (csv) file (see the subsection entitled "Density Functional Theory" in the "Materials and Methods" section, eqs S1 and S2 in Note S1 in the SI, and Tables S2 and S3 in the SI). The geometries and raw adsorption energies are available in the ioChem-BD, in 10.19061/iochem-bd-1-152 and 10.19061./iochem-bd-1-228).

Experimental Data. The next step, as shown in Figure 1, is the acquisition of experimental data. A full dataset implies the evaluation of reactivity while taking the entire CH_2X_2 family into account. However, from a practical point of view, dihalomethanes with X = F or I were not included because of handling issues: HF formation or the high boiling point of CH_2I_2, respectively. The experimental dataset contains the reactivity data of SiO_2-supported Fe-, Co-, Ni-, Cu-, Ru-, Rh-, Ag-, Ir-, and Pt-based catalysts (1.0 wt % metal content) in the hydrodehalogenation of CH_2X_2 (X = Cl, Br), using identical reaction conditions (T = 523 K, P = 1 bar, and time-on-stream (tos) = 15 min) to generate a comprehensive and consistent dataset. The systematic evaluation ensures that (i) only single metal component forms are considered; (ii) the nanoparticle size distribution is comparable, as shown in the characterization of the fresh catalyst;[95] (iii) all generated products (CH_3X, coke, CH_4, and other gas phase side products) are measured with comparable accuracy. Briefly, the utilized experimental dataset consists of (i) CH_2X_2 conversion (where $X_{CH_2X_2}$, X = Cl, Br; see Figure 2a), (ii) product selectivity (Figure 2b), (iii) yield to the halomethane, coke, and methane (denoted as S_i and Y_i, respectively (where i = CH_3X, coke, and CH_4), and (iv) the reaction rate ($r_{CH_2X_2}$).

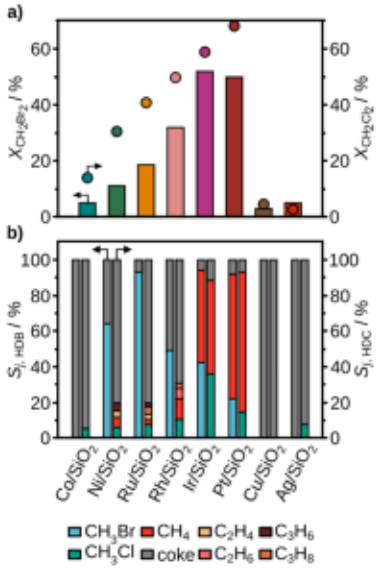


Figure 2. (a) Conversion of CH_2Br_2 and CH_2Cl_2, and (b) product selectivity of the catalysts in CH_2X_2 hydrodehalogenation. In panel (a), the conversion was assessed at a constant space velocity of F_T/W_{cat} = 40 $cm^3\ min^{-1}\ g_{cat}^{-1}$ in hydrodebromination (HDB) and F_T/W_{cat} = 100 $cm^3\ min^{-1}\ g_{cat}^{-1}$ in hydrodechlorination (HDC), while product selectivities in panel (b) were determined at ca. 20% CH_2X_2 conversion achieved by adjusting the space velocity in the range of F_T/W_{cat} = 20−150 $cm^3\ min^{-1}\ g_{cat}^{-1}$ in HDB and 70−350 $cm^3\ min^{-1}\ g_{cat}^{-1}$ in HDC. Other reaction conditions: CH_2Br_2:H_2:Ar:He = 6:24:4.5:65.5 (X = Br, Cl), T = 523 K, P = 1 bar, and tos = 15 min.

Microkinetic Modeling. The state-of-the-art methodologies to determine rates and other parameters take into account the information condensed in reaction energy profiles. In our case, the full DFT networks toward all the products were obtained only for the hydrodebromination and are summarized in Table S1 in the SI.[95] The paths to CH_3Br, CH_4, and coke were considered for all the metals in our study. The reversibility analysis shows that metals cluster in three different groups: Co and Ni; Ru and Rh; and Ir, Pt, Cu, and Ag (see Note S2 in the SI) according to the most relevant elementary steps. Therefore, differently from traditional microkinetic models, where one metal system bears the information for the mechanistic preferences of the full metal data set, we observe that at least three different mechanisms and energy profiles are needed, leading to the selectivity maps that are dependent on the chosen cluster of metals (see Figures S3−S7 in the SI). This points toward the fact that the small errors in the LSR get amplified when trying to address the small energy variations involved in selectivity through the path selection toward different products.

Even if applying the three different microkinetic models, the selectivity patterns depart significantly from experimental observations (see Figure S2, panels (a), (b), and (c)) likely due to the high coverages of poisons as coke and Br. Coke will deposit on the catalyst surface, leading to a decrease over time of the total available active sites, thus limiting the predictability of MK models. Therefore, only systems lean on adsorbed carbons like Ru, Ir, and Pt could be qualitatively represented, consistent with previous observations.[61] If Br is the poison, strong lateral interactions due to the amount of charge dragged from the surface make adsorption energies highly dependent on the particular coverage, thus breaking the mean-field restriction of identical sites. The hydrodehalogenation system

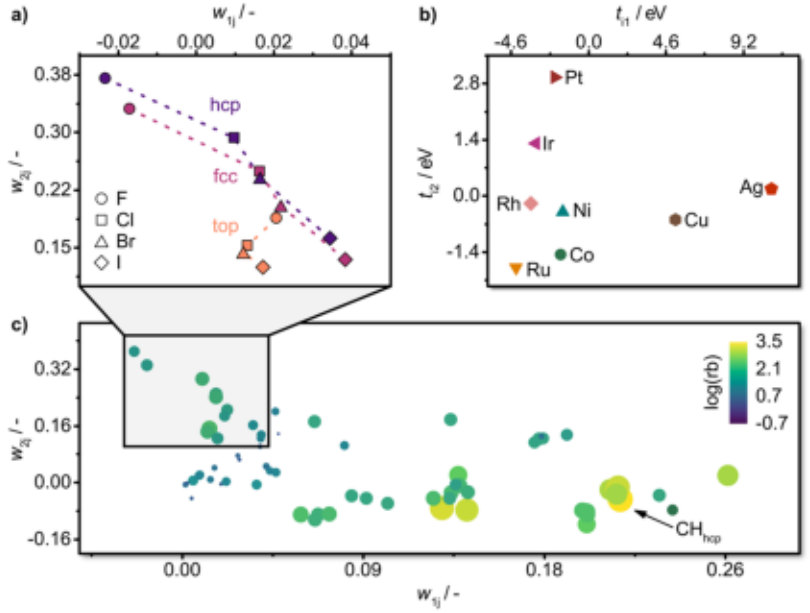


Figure 3. Principal component analysis (PCA) contributions of (a) the halogens to the two principal components, depending on the adsorption position, (b) the metallic surfaces to the principal components, and (c) the absorbed intermediates to the two principal components. The color code and the size represent the robustness term associated with the accuracy of the intermediate as a descriptor.

is thus a very good example to test machine learning approaches.

Principal Component Analysis. PCA was applied to the DFT energy dataset, presented in Figure 3, to reduce the dimensionality by finding the main contributors to the adsorption of the intermediates. Figure 3a shows the resulting dimensionless eigenvectors ($\boldsymbol{w}$) that store the information related to the intermediates (*loading* vectors). The metal characteristics $\boldsymbol{t}$ (*score* vectors) in Figure 3b) represent the energy of the metals (in eV) and are obtained as the product between the centered input matrix and the loading vectors. The three significant principal components cover 89.9%, 7.6%, and 1.6% of the total variance, while weights were 92.8%, 5.2%, and 1.2% (considering only the X = Br dataset).[95] Therefore, when expanding the size of the sampling space by a factor of four, we observe that the contributions are robust but the relative weight between the first and the second component rebalance slightly (89.9% and 7.6%, with respect to 92.8 and 5.2% for Br only and for all the halogens datasets, respectively). Moreover, the results are halogen-independent, allowing generalization. The model can thus estimate the adsorption energy of any intermediate in the network as $E_{ads} = t_1w_1+t_2w_2$, accounting for 97.5% of the values. Selecting the minimum optimal set of intermediate energies that can represent the entire database implies that the selected intermediate should be described exclusively by one of the two principal components, and should provide a higher accuracy when predicting the energies.[79,110] CH_{hcp} and the single halogen atoms X_{hcp} were used as the intermediate descriptors of the first and second component, respectively. The first component is associated with the covalency and can be represented by the E_{ads} of CH, while the second component represents the redox terms, more univocally described by the adsorption of the atomic halogen species.[79] Thus, the PCA term for metals are transferable, while the weights for the adsorbates ($\boldsymbol{w}$) follow the trends according to their position in the periodic table, see Figure 3a). The principal components condense all the information stored in the full DFT intermediate energy dataset and act as descriptors when finding catalyst performance equations for activity and product selectivity.

Bayesian Learning. We have applied the Bayesian Machine Scientist (BMS) to identify the functional forms for the experimental macroscopic observables as a function of the DFT-derived atomistic principal components. BMS was benchmarked against Random Forest and Gaussian Regressors (RF and GR, respectively). RF constitutes the state of the art, while GR follows Bayesian inference. Both reference methods provide reasonable accurate predictions within explored zones, separate space regions according to the response magnitude, and the GR presents functional shapes that are similar to volcano plots, though both techniques belong to the nonexplainable class of SL. The experimental observables ($X_{CH_2X_2}$, S_i, Y_i, and $r_{CH_2X_2}$) are all fitted with BMS and the other methods to the simplest possible meaningful equation taking the two principal components as only variables, as is illustrated in Figure 1.

The BMS methodology was first employed in the bromine reaction subset including the conversion of CH_2Br_2 ($X_{CH_2Br_2}$) and selectivity to CH_3Br (S_{CH_3Br}). While the functional forms of conversion and yield are very simple, showing a direct dependence on the Br adsorption energy and an inverse one with CH adsorption energy (see Note S5, eqs S13 and S14, and Figures S9a and S9c in the SI), the equation for selectivity contained inverse and polynomial terms (up to three degrees) and four constants, which is an indication of potential overfitting (see eq S15 and Figure S9b in the SI). Despite the complexity, the selectivity equation was able to split the $\{E_{ads}(Br), E_{ads}(CH)\}$ space in two areas, separating the catalysts, depending on coke generation.

A natural extension of this approach is to assess whether the functional forms identified for bromine can be extended to the chlorine-containing reactant. Mathematically, this corresponds to the search of an isomorphism between the conversion, selectivity, and yield surfaces for the halides. In doing so, the

Table 1. BMS Equations for the Conversion of CH_2X_2 ($X^p_{CH_2X_2}$, X = Cl, Br), Selectivity to CH_3X, coke, and CH_4 (S_i^p, i = CH_3X, coke, CH_4), Corresponding Yields (Y_i^p), and Reaction Rate, $r^p_{CH_2X_2}$. $\omega_2' = E_{ads}$ (X_{hcp}) + c_3, $\omega_1 = \omega_1' + c_4$, and $\omega_1' = E_{ads}$ (CH_{hcp}) + WF − E_{ea}, and Statistical Error (SSE)[a]

observable	equation	eq number	SSE
$X_{CH_2X_2}$	$X^p_{CH_2X_2} = -\omega_2 c_1 + \frac{{c_2}^2}{\omega_1}$	(4)	260
S_{CH_4}	$S^p_{CH_4} = \left[\omega_2 + \omega_1 + \omega_2\left(\omega_2 - \frac{1}{c_1}\right)(\omega_1 + c_2)^2\right]^2$	(5)	284
S_{coke}	$S^p_{coke} = \frac{\omega_2 c_1}{\omega_2 c_2 - \omega_1}(\omega_2 + c_2)\left[\omega_2 + c_2 + \frac{c_2}{\omega_1 \cos\left(\frac{{\omega_2}^2}{c_2}\right)}\right]$	(6)	510
S_{CH_3X}	$S^p_{CH_3X} = 100 - S^p_{coke} - S^p_{CH_4}$	(7)	856
$Y_{i\in\{CH_3X,coke,CH_4\}}$	$Y^p_{i\in\{CH_3X,coke,CH_4\}} = \frac{X^p_{CH_2X_2} S_i^p}{100}$	(8)	{95, 193, 119}
$r_{CH_2X_2}$	$r^p_{CH_2X_2} = \left(\frac{\omega_1 + c_2\omega_2}{c_1}\right)^2 + \omega_1$	(9)	1687

[a]The c_i coefficients and individual SE are presented in Tables S4 and S5 in the SI. The comparison plots between experimental and predicted data are shown in Figure S9 in the SI.

metal descriptors are fixed t_i, while the halide and fragment contributions w_i are updated (particularly, C-only containing fragments are identical, as the interaction between C-only fragments and halogens is not considered now, see Table S2 in the SI). The predictions for the observables are denoted with a super index p (as for example: $X^p_{CH_2Cl_2}$). These follow the experimental points once the accompanying constants are refitted (see Figure S9d−i, Table S6, and Note S5 in the SI). Thus, the conversion (and, to some extent, the yield) maintain the functional form, recalling the similarities in the volcano shapes found by traditional methods but with two descriptors, as found for the Deacon process,[54] and pointing toward the existence of a generalized equation that can express the performance set for the full halogen family. In turn, the analysis of selectivity, particularly to $S^p_{CH_3Cl}$, shows that, although we do estimate these values correctly, the differences in the fitting constants (proportion of 1 to −20 for Br and Cl, respectively, in c_2) changes the shape of the surface, visible when comparing Figures S9b and S9e in the SI. In the latter, an artificial compression along the x-axis, corresponding to the CH adsorption, is shown.

Intrinsically, the different w values obtained for the different halogens are not enough for the Br equations to be representative of Cl properties as the halogen presence modifies CH adsorption. The adsorbed halogens drag density form the surface (Figure S10 in the SI) and modify the work functions (WF) of metals, which severely affects covalent contributions to the adsorption energy of coadsorbates[111,112] (see Figures S11 and S12 in the SI). The energies of the organic moieties would be affected most, from 0.02 to 0.12 eV per halogen atom. Therefore, rescaling the w parameters corresponding to the CH_n fragments is required. The difference between the metal-only WF and the electron affinity (E_{ea}) of the isolated halogen atom is taken as a proxy for the penalty of the CH_n fragment adsorption energy $\omega_1 = E_{ads}$ (CH_{hcp}) + $WF - E_{ea}$; this allows alignment of the energies of different halogens (bromine and chlorine values are reported in Table S7 in the SI).

Analysis of the Activity Equations. To extend the validation of the approach, the BMS was further applied to the full experimental dataset containing $X_{CH_2X_2}$, S_i, Y_i, and $r_{CH_2X_2}$ (16 experiments and 4 sets of parameters (ω_1 and ω_2); halide adsorption and CH corrected energy adsorptions; see Note S6 in the SI for further information on benchmark datasets). RF and GR were also performed for comparison (Note S7f and Figures S13−S17 in the SI). Each of these searches for performance observable equations was run independently and analyzed separately. The criteria for the best function are (i) simplicity, (ii) interpretability from a physical point of view, and (iii) performance in terms of the lowest sum of square errors (SSE), the acceptance range of which is dependent on the magnitude of the observables: for conversion and the selectivity, the SSE range is between 100 and 1000 (observables between 0 and 100% of product), while for the reaction rate, it is between 100 and 2000 (observables between 0 and 220 s^{-1}). The BMS equations and their associated SSE values are listed in Table 1, and their associated fitting constant are reported in Table S8 in the SI. The selected ensemble of equations, together with the corresponding SSE values and fitting constants for each observable, can be found in Notes S6 and S8, Tables S9−S24, and eqs S16−S70 in the SI. The validation of the BMS predictions was performed via a leave-one-out test (for the results and brief discussion, see Figures S18 and S19, and Note S9, in the SI).

The BMS result for conversion is expressed as eq 4 in Table 1, $X^p_{CH_2X_2}$, showing a simple function (SSE = 260) with a marked volcano character formed by a direct dependence with the adsorption energy of the halogen and an inverse dependence with carbon fragments ω_1. Note that the BMS predictions domain is enclosed in the near range of the input variables (as any other SL technique); thus, the risk of any discontinuity by means of the inverse dependence with ω_1 is prevented. The $X^p_{CH_2X_2}$ dependencies point to a possible poisoning of the metal surfaces with the most exothermic halogen-metal bonds, while conversion improves at lower CH

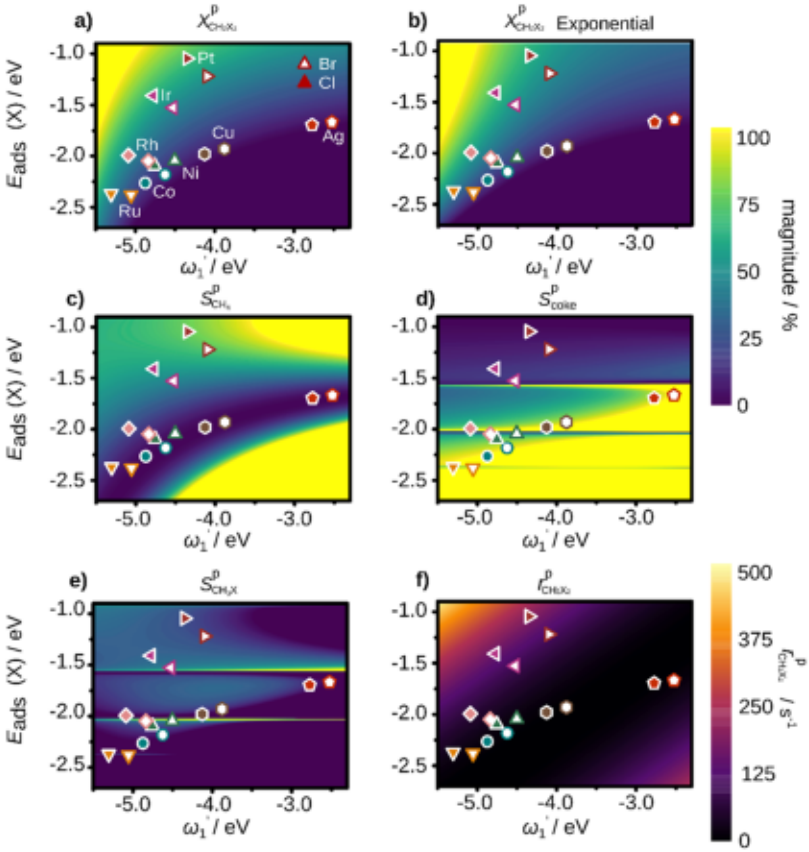


Figure 4. Predicted surfaces using the Bayesian Machine Scientist (Bayesian, eqs 1−11) of (a) CH_2X_2 conversion, ($X^p_{CH_2X_2}$; (b) CH_2X_2 conversion using the exponential model, ($X^p_{CH_2X_2}$ exponential; (c) selectivity to CH_3X, coke, and CH_4 ($S^p_{CH_4}$; (d) S^p_{coke}; (e) $S^p_{CH_3X}$; and (f) CH_2X_2 rate ($r^p_{CH_2X_2}$. All the predictions are presented as a function of the adsorption energy of Br, Cl (denoted as X), and CH. E_{ads}(CH) are corrected with the corresponding metal Work Function (*WF*) and halogen Electron Affinity (E_{ea}) (ω_1') on the hcp sites ($E_{ads}(X_{hcp})$ and $E_{ads}(CH_{hcp})$, respectively) of Co, Ni, Ru, Rh, Ir, Pt, Cu, and Ag.

adsorption values, as illustrated in Figure 4a.[113] According to this interpretation, $X^p_{CH_2X_2}$ is higher for Cl than for Br, because of the higher electronegativity. eq 4 in Table 1 can be interpreted as the surrogate models presented by Campbell et al.[32] To test the hypothesis, we have formulated an exponential model (eq 10) for conversion ($X^p_{CH_2X_2}$).

$$X^p_{CH_2X_2} = c_1 + c_2 e^{-\omega_2} + c_3 e^{-\omega_1} \quad (10)$$

The exponential expression assumes that conversion is related to the desorption rates of poisoning species, that is, CH and Br, and are in agreement with the BMS model, as illustrated in Figure 4b. Indeed, the inverse dependence of conversion with ω_1 found for BMS can be thought in terms of $e^{-x} = (e^x)^{-1}$, thus mapping the equations under the same variable space span. Finally, the SSE values of both models are very similar (SSE(BMS) = 276, SSE(Exponential) = 331), see Table S9 in the SI). These results support the interpretability of the BMS function obtained for conversion (eq 4 in Table 1) as the surrogate model of the reaction rates, taking into account that the BMS-derived expressions are only valid within the descriptors span and cannot be employed for extrapolations. On the other hand, selectivity patterns are too complex to apply to the exponential model satisfactorily.

To test the performance of the BMS algorithm with alternative datasets of the DFT-generated reaction profile (see Note S6 in the SI), we have used two different DFT variables and corresponding experiments: (i) a set of nine adsorption energies (X, H, and the C fragments: CX, CH_nX, and CH_n, where X = Cl, Br, and $n = \{1, 2, 3\}$ on the hcp sites), and (ii) the nonzero barriers of dehydrobromination to fit $X^p_{CH_2Br_2}$ and $S^p_{CH_3Br}$, corresponding to reactions R3 to R5, R7 and R8 in Table S1 in the SI. In all cases, we have used the same number of variables and fitting constants. For the nine-variables case, the BMS algorithm successfully discards four variables (and parameters) for the simplest models. However, the resulting five-variable function (eq 11) is less compact and elegant than the two-variable equation, has a slightly better predictive power (SSE of 223 vs 260), requires 2−3 times higher computational cost (1−2 vs 2−3 days), and more importantly, it is difficult to interpret from a physical point of view, even if it has a negative exponential term, depending on the hydrogen adsorption energy, which partially resembles the Arrhenius equation (indeed, eq S33 in Table S16 in the SI from the two-variables $X^p_{CH_2X_2}$ ensemble also presents exponential dependences with ω_1 and ω_2). Particularly, the rest of the ensemble of the nine-variables functions seem to entangle the different variables. The nine-variables functions ensemble are reported in Table S10 (eqs S16−S20) in the SI, their fitting constants, and SSE values are reported in Table S11 in the SI.

$$X^{p}_{CH_2X_2} = \frac{E_{ads}(CH_{2hcp}) + {c_1}^2}{E_{ads}(CH_{hcp})c_2 \frac{E_{ads}(CHX_{hcp})}{c_3} e^{E_{ads}(H_{hcp})}} \quad (11)$$

For the $X^{p}_{CH_2Br_2}$ when the input space are the reaction barriers of Br-only dataset (only those that are nonzero), after some iterations, the BMS algorithm converges toward an ensemble of functions (Table S12 (eqs S21−S25) in the SI, depending on all five variables, thus, overfitting. Even worse is the performance for $S^{p}_{CH_3Br}$ as the algorithm diverges (Table S13 (eqs S26−S30) in the SI). Therefore, the reduction of the dimensionality of DFT dataset is crucial to achieving meaningful and consistent functional forms, thus reinforcing the role of descriptors in heterogeneous catalysis.

In summary, employing a full energy dataset (including transition states) generates unnecessary issues because the algorithm could generate long sets of coupled descriptors, which increases the optimization time, and is prone to overfitting. In addition, by using PCA, the rationalization and interpretation of the leading terms (descriptors) becomes more transparent.

Analysis of the Selectivity Equations. Next, we addressed selectivity patterns. The selectivity is governed by competing steps in very small adsorption energy ranges at product distribution switches (the so-called cliffs).[91] Three main potential products appear in the hydrodehalogenation process: when hydrogenation dominates, the major product is CH_4 (with a small minority of volatile C2 coupling products), if C−C bond formation prevails, the catalyst cokes. However, the desired product is the semihydrogenated CH_3X that mechanistically shares intermediates with both hydrogenation and coking routes. Therefore, the selective regime to CH_3X corresponds to a narrow area and its selectivity equation is the steepest, hence most difficult to fit. As a rule, and for any complex reaction with a different product distribution, a semireaction path would be the most difficult to approach mathematically. Therefore, the extremes of the catalytic products (coke and methane) are taken as more robust functions than the CH_3X equation. Consequently, the expressions for the selectivity to coke and methane are analyzed first, and the halomethane selectivity is obtained by subtraction from the total. Notice the differences in the predicted selectivities from BMS and the different microkinetic models only for the Br system (see Figure S13 in the SI).

The prediction of selectivity to CH_4, $S^{p}_{CH_4}$ in eq 5 in Table 1 presents a quadratic polynomial dependence with ω_1 and ω_2. The interpretation of these results points to a tradeoff between both variables. Relatively strong metal−CH bonds promote CH_4, while a large halogen adsorption energy inhibits CH_4 production. Only Pt, Ir, and Rh (for Cl) present the right combination between relatively low exothermic halogen adsorption and relatively strong CH bonding (see Figure 4c). This model, as in the conversion case, is simple and the SSE for eq 5 in Table 1 is 284. It is significant that both selectivity terms are polynomials of second degree; this closely follows the traditional modeling of atomistic potential energy surfaces as combinations of parabolas, even in Marcus developments.[114] Remarkably, CH_4 selectivity arises from a balance between both main descriptors.

The S^{p}_{coke} expression (eq 6 in Table 1) shows a direct dependence with the exothermicity of ω_2 and an inverse dependence with the stability of ω_1. These results points to the following: a strong halogen-metal bond results in the complete dehalogenation of CH_2X_2, whereas a strong CH−metal bond leads to lower probabilities to generate coke by C−C coupling. In addition, the S^{p}_{coke} expression clearly divides the $\{\omega_1, \omega_2\}$-space in four different clusters, based on the principal product obtained, recover the main CH_3Br result reported in our previous work,[95] and extends it to CH_3Cl, as illustrated in Figure 4d). Inside the regions, it is possible to see a volcano shape in the ω_2 direction, which limits the maximal coke production subregions (maximal ω_2 and minimal ω_1 modulus). The cosine term modulates the frontier between several Br and Cl points without altering the general dependencies.[115] This is explained by the fact that the computed $\omega_{1,2}$ variables for Br−Ru and Cl−Br are very similar, but the selectivity to coke differ by 1 order of magnitude (79% and 7% for Cl and Br, respectively). Because of the complexity of predicting the S_{coke} behavior, the SSE value is 510.

A direct attempt to predict selectivity to CH_3X, $S^{p}_{CH_3X}$, leads to complex and difficult to interpret functional forms (see Table S19 in the SI). However, it can be obtained from the carbon balance, including the use of selectivity of the other two main compounds from $S^{p}_{CH_4}$ and S^{p}_{coke} (eq 7 in Table 1). The main contribution to the SSE value (856) comes from the propagation of the S^{p}_{coke} errors. The $S^{p}_{CH_3X}$ obtained in this way is illustrated in Figure 4e). From the found dependencies, the more exothermic the halogen adsorption, the narrower the selective range or productive subregions in the $\{\omega_1, \omega_2\}$-space for CH_3Br or CH_3Cl. The area of those productive subregions is dependent on ω_2, as shown in Figure 4e). Thus, larger productive subregions areas imply higher probabilities of finding points with higher associated selectivity values. Physically, this is interpreted as follows: the stronger the halogen is adsorbed to the surface, the lower the selectivity to CH_3X. However, this trend does not apply to CH_3Br selectivity over Ru and Ni, which explains why the BMS divides the different regions using maximum selectivity discontinuities (yellow stripes in Figure 4e).

Analysis of the Yield and Rate Equations. The yields to each of the three main products have been estimated using eq 8 from eq 4 in Table 1 and their respective selectivity expression (eqs 5−7 in Table 1). If we examine the ensembles of yields (Tables S20−S22 and eqs S51−S65 in the SI), the SSE obtained with eq 8 in Table 1 is equal or even lower, compared to the other candidate equations (95, 193, and 119 for CH_3X, coke, and CH_4, respectively). Generally, the separation between regions as shown for the selectivity surfaces are smoothed for yields. The surface for $Y^{p}_{CH_3X}$ shows a mixed dependence with ω_2 and ω_1 between $S^{p}_{CH_4}$ and $X^{p}_{CH_2X_2}$, as illustrated in Figure S13m in the SI). Y^{p}_{coke} presents a remarkable similarity to the surface obtained for $Y^{p}_{CH_3Br}$ (Figure S13p in the SI), which points to a direct squared dependence with ω_2 and an inverse squared dependence with ω_1. The $Y^{p}_{CH_4}$ surface allows a more direct interpretation, similar to $X^{p}_{CH_2X_2}$, as reported in Figure S13s in the SI.

Finally, we have extrapolated our methodology and results to predict the reaction rate of CH_2X_2, $r^{p}_{CH_2X_2}$. The resulting function is reported in eq 9 in Table 1, which is a second-degree polynomial with two terms: (i) the squared difference between ω_1 and ω_2, which is always positive and (ii) ω_1, which is negative. Thus, favored CH adsorption and unfavored X

Model feature	Microkinetic Modeling	Bayesian Machine Scientist
Reactivity complexity	Best in single product reactions	Multiproduct reaction network
Interpretability	Physically interpretable Direct mechanistic insights	Physically interpretable Not direct mechanistic insights
Dataset content and size	Full reaction profile One experimental measure	Only key adsorption energies All experimental data available
Generalizability	Only for simple systems No issues with variables span	For an entire family of reactants Values near to initial variables
Accuracy	Reasonable for activity Poor for selectivity Might need DFT-profiles fitting	Excellent for activity Excellent for selectivity Might need energy rescaling

Figure 5. Comparison between the key features of microkinetic models and the BMS approach. In green if the feature is better or equal for compared to the alternative model, in red if the feature is worst.

adsorption or vice versa (i.e., the difference between ω_1 and ω_2 is high) leads to high rates. The obtained shape is a cliff (Figure 4f). The Cl−Ru point is the main responsible of the error (60% of the 1687 error), similar to the S^P_{coke} case.

Benchmark of BMS versus Common Statistical Learning Techniques. The RF and GR divide the CH_2X_2 conversion, product selectivity and yield, and rate spaces qualitatively in the same manner as BMS (Figures S13 and S14 in the SI for $r^P_{CH_2X_2}$, and Figures S15−S17 in the SI), particularly for conversion (Figures S13a−S13c). The robustness of conversion (eq 4 in Table 1) further reinforces the heuristic knowledge of the strength of volcano plots in catalysis. For product selectivity and yield, and rate, the RF was able to place the frontiers of the regions on the $\{\omega_1, \omega_2\}$ space with similar ω_2 values. The RF is difficult to interpret as no equation is derived (within the given space); GR is even less interpretable, as discerning selectivity with Gaussian functions is difficult, because of the sharp nature of selectivity cliffs. Furthermore, the accuracy of BMS prediction is much better than Random Forest or Gaussian regressors for the selectivity to CH_3X and coke.

Comparison of BMS with Classical Methodologies. In this final section, we would like to discuss the advantages, limitations, and challenges for future hybrid computational models to represent experimental observables, summarized in Figure 5. Microkinetic models have found extreme success in addressing the reactivity of metals for simple reaction networks leading to one product. Hybrid data, in this case, further reinforces the soundness of these models and thus, a complete understanding of the reactivity can be rooted in first-principles data.

With regard to the interpretability of the models, the equations derived from BMS are only qualititavely interpretable, given indications of the most demanding process or the most likely poisons. For instance, we can extrapolate some qualitative insight from the analysis of the conversion surface. For the regions in which we have a maximum poisoning (from CH or the halogen), the most abundant species at the steady state would be carbon fragments or halogens. Under these conditions, we can assume that the rate-determining step is the adsorption of the reactants.[37] In this regard, the insight provided is of less quality that the microkinetic model, in part because the complex DFT analysis profile is summarized into only the key steps. The BMS equations are only useful in the ranges expanded by the variables, and, thus, extrapolations can lead to nonphysically meaningful results. However, there are many aspects where SL methodologies provide less insightful predictions at the atomistic scale. Nevertheless, SL predictions are more robust and useful than standard MK results. Particularly as reactions become more complex and reactivity involves more elementary steps, the relevance of MK becomes more limited.

MK models are no longer able to handle the issues associated with different phase (in some cases, even mixed ones), dynamic rearrangements, site blocking due to poisoning and generally related to the material gap problem. Although we have presented our results for metal-only systems, the extension to structural differences should be the next step in generalization. Another limitation of MK involves the generalization to a family of reactants, as shown here. This is possible by a rescaling in the variable for BMS, but it was less evident for MK models, even if evidence within a reactant family are as old as Brønsted studies.

Regarding the type of data employed by both methods, hybrid sets are very valuable, because they constitute the true benchmark. From the MK standpoint, a single rate or yield and the full reaction profile for one single metal system has been the state of the art. ML techniques can better benefit from the extended systematization of the experimental data, as automatization provides accessibility to larger amounts of identically generated kinetic data, while diminishing the burden of the DFT part, since they are only dependent on the adsorption energies. Thus, automatically generated data are less affected by the intrinsic DFT errors. Besides, algorithms to identify transition states allow the unbiased identification of the descriptors, because they can be rooted in robust SL techniques and not in heuristic priors.

When considering accuracy, MK(DFT) models are excellent in presenting the activity qualitatively, but have severe issues in reproducing selectivity patterns. This is due to the fact that the linear scaling relationships employed to simplify MK equations typically present small errors that do not provide enough accuracy to address the small energy differences involved in the prediction of selectivity. A clustering technique applied to our PCA descriptors for instance would have a tendency to group elements right where the cliff changing selectivity appears (Ni and Co, for instance). Thus, even if very accurate DFT values could be obtained, it is unlikely that sharp selectivity patterns for multiproduct reactions can be obtained using MK models.

In contrast, BMS clearly identifies the patterns when built on robust variables as those derived by PCA. Finally, BMS derived functions could be used as surrogate equations in multiscale approaches and even as a systematic unbiased way to approach standard microkinetic models to experimental data.

CONCLUSIONS

Modeling in catalysis has its early origin on phenomenological observations, giving empirical equations that were simple, such as the power laws. With the introduction of microkinetics, such models became more complex and could only be solved numerically. The gold standard in reaction modeling over the last decades has been coupling density functional theory reaction profiles and mechanisms to microkinetic modeling if needed, the energy profiles are massaged to account for errors due to set up or intrinsic to DFT. However, going further from metal catalytic systems as the material complexity, dynamicity and the number of elementary steps in the reactions increase in an exponential way and, thus, these systems might be not fully addressable by traditional DFT-based methods making classical microkinetic methods fail in their predictions. The availability of consistent, systematic, annotated experimental and computational datasets provides the needed clean raw data to apply SL to apprehend this complexity. The present work shows the possibility to go beyond traditional schemes based on DFT energy profiles for a reaction for which the standard microkinetic modeling fails. To this end, we have qualitatively addressed rates and other experimental parameters using multiscale approaches, in a case that implies a different operando phase for some types of metals (carbides for Co or bromides for Ag and Cu): hydrodehalogenation (X = Cl, Br). The methodology presented is robust enough to generalize the equations to an entire family of reactants.

We have taken a hybrid dataset composed of the activity, selectivity, and rate of hydrodehalogenation reactions on identically prepared metal nanoparticles with similar sizes. The energies of the reaction intermediates were evaluated by DFT and the overall hybrid dataset was employed to feed the SL models (Bayesian Machine Scientist) and derive performance equations that can be physically interpreted. A first step requires identifying the descriptors through the dimensionality reduction of the problem, which is mandatory to derive robust functions. The traditionally identified volcanos are retrieved with data-driven methodologies indicating the robustness of this functional shape to describe heterogeneous catalyst problems. Generalized performance data can be found for a family of hydrodehalogenation reactions. The procedure can be expanded to other sets of catalysts and can be particularly successful when investigating families of similar reactants or larger number of atoms where standard DFT approaches fail, such as (de)polymerizations or biomass conversion. Selectivity involves a more elusive set of parameters, and the extremes of the phase reaction space are easier to describe (semireactions are more ill-defined due to the intricacy of the reaction networks). This problem should be assessed in detail by future studies as SL techniques have been focused on clustering, whereas selectivity is about defining differences in very narrow energy spans. The equations derived from BMS can be used in coarse-grained models and reactor design to avoid the instabilities and error propagation observed when employing microkinetic modeling on the DFT results. Overall, we have shown how the results from BMS can be mapped to previous surrogate models in the literature and outperform MK(DFT), thus having the potential to fill the gap where microkinetic modeling is not possible, because of phase transitions, exceedingly complex reaction networks, multiple products, and selectivity issues. Finally, this hybrid data approach can be used to identify and explain the descriptors in future catalyst design.

ASSOCIATED CONTENT

Supporting Information

Details regarding the computational methods and complementary results reported (Figures S1−S19, Tables S1−S24, eqs S1−S70, and Notes S1−S9 (PDF)

AUTHOR INFORMATION

Corresponding Authors

Javier Pérez-Ramírez − *Institute for Chemical and Bioengineering, Department of Chemistry and Applied Biosciences, ETH Zurich, 8093 Zürich, Switzerland;* orcid.org/0000-0002-5805-7355; Email: jpr@chem.ethz.ch

Núria López − *Institute of Chemical Research of Catalonia, The Barcelona Institute of Science and Technology ICIQ, 43007 Tarragona, Spain;* orcid.org/0000-0001-9150-5941; Email: nlopez@iciq.es

Authors

Sergio Pablo-García − *Institute of Chemical Research of Catalonia, The Barcelona Institute of Science and Technology ICIQ, 43007 Tarragona, Spain*

Albert Sabadell-Rendón − *Institute of Chemical Research of Catalonia, The Barcelona Institute of Science and Technology ICIQ, 43007 Tarragona, Spain*

Ali J. Saadun − *Institute for Chemical and Bioengineering, Department of Chemistry and Applied Biosciences, ETH Zurich, 8093 Zürich, Switzerland*

Santiago Morandi − *Institute of Chemical Research of Catalonia, The Barcelona Institute of Science and Technology ICIQ, 43007 Tarragona, Spain*

Author Contributions

[V]These authors made equal contributions.

Notes

Details regarding all DFT structures and adsorption energies can be found on the ioChem-BD platform in the following links: (i) the complete hydrodehalogenation set can be found in 10.19061/iochem-db-1-228; (ii) the hydrodebromination set can be found in 10.19061/iochem-bd-1-150 and 10.19061/iochem-bd-1-152.

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This work was supported by the ETH Research Grant (No. ETH-43181) and Ministerio de Ciencia e Innovación (Ref. No. RTI2018-101394-BI00). This publication was in part created in NCCR Catalysis, a National Centre of Competence in Research funded by the Swiss National Science Foundation. The authors thank BSC-RES for generously providing

computational resources. The authors also would like to explicitly thank Reviewer 3 for their valuable suggestions.

REFERENCES

(1) Chorkendorff, I.; Niemantsverdriet, J. W. *Concepts of Modern Catalysis and Kinetics*; Wiley, 2003.
(2) Temkin, S. I.; Yakobson, B. I. Diffusion-Controlled Reactions of Chemically Anisotropic Molecules. *J. Phys. Chem.* **1984**, *88*, 2679−2682.
(3) Boudart, M. Kinetics on Ideal and Real Surfaces. *AIChE J.* **1956**, *2*, 62−64.
(4) Aris, R.; Mah, R. H. S. Independence of Chemical Reactions. *Ind. Eng. Chem. Fundam.* **1963**, *2*, 90−94.
(5) Dumesic, J. A.; Aparicio, L. M.; Rekoske, J. E.; Treviño, A. A.; Rudd, D. F. *The Microkinetics of Heterogeneous Catalysis*; American Chemical Society: Washington, DC, 1993.
(6) Chatterjee, A.; Vlachos, D. G. An Overview of Spatial Microscopic and Accelerated Kinetic Monte Carlo Methods. *J. Comput. Mater. Des.* **2007**, *14*, 253−308.
(7) Andersen, M.; Panosetti, C.; Reuter, K. A Practical Guide to Surface Kinetic Monte Carlo Simulations. *Front. Chem.* **2019**, *7*, 202.
(8) Ravipati, S.; Savva, G. D.; Christidi, I.-A.; Guichard, R.; Nielsen, J.; Réocreux, R.; Stamatakis, M. Coupling the Time-Warp Algorithm with the Graph-Theoretical Kinetic Monte Carlo Framework for Distributed Simulations of Heterogeneous Catalysts. *Comput. Phys. Commun.* **2022**, *270*, 108148.
(9) Hansen, M. H.; Nørskov, J. K.; Bligaard, T. First Principles Micro-Kinetic Model of Catalytic Non-Oxidative Dehydrogenation of Ethane over Close-Packed Metallic Facets. *J. Catal.* **2019**, *374*, 161−170.
(10) Stegelmann, C.; Schiødt, N. C.; Campbell, C. T.; Stoltze, P. Microkinetic Modeling of Ethylene Oxidation over Silver. *J. Catal.* **2004**, *221*, 630−649.
(11) Salciccioli, M.; Chen, Y.; Vlachos, D. G. Microkinetic Modeling and Reduced Rate Expressions of Ethylene Hydrogenation and Ethane Hydrogenolysis on Platinum. *Ind. Eng. Chem. Res.* **2011**, *50*, 28−40.
(12) Alexopoulos, K.; Vlachos, D. G. Surface Chemistry Dictates Stability and Oxidation State of Supported Single Metal Catalyst Atoms. *Chem. Sci.* **2020**, *11*, 1469−1477.
(13) Chutia, A.; Thetford, A.; Stamatakis, M.; Catlow, C. R. A. A DFT and KMC Based Study on the Mechanism of the Water Gas Shift Reaction on the Pd(100) Surface. *Phys. Chem. Chem. Phys.* **2020**, *22*, 3620−3632.
(14) Andersen, M.; Plaisance, C. P.; Reuter, K. Assessment of Mean-Field Microkinetic Models for CO Methanation on Stepped Metal Surfaces Using Accelerated Kinetic Monte Carlo. *J. Chem. Phys.* **2017**, *147*, 152705.
(15) Pineda, M.; Stamatakis, M. Beyond Mean-Field Approximations for Accurate and Computationally Efficient Models of on-Lattice Chemical Kinetics. *J. Chem. Phys.* **2017**, *147*, 024105.
(16) Lian, Z.; Ali, S.; Liu, T.; Si, C.; Li, B.; Su, D. S. Revealing the Janus Character of the Coke Precursor in the Propane Direct Dehydrogenation on Pt Catalysts from a KMC Simulation. *ACS Catal.* **2018**, *8*, 4694−4704.
(17) Jørgensen, M.; Grönbeck, H. Selective Acetylene Hydrogenation over Single-Atom Alloy Nanoparticles by Kinetic Monte Carlo. *J. Am. Chem. Soc.* **2019**, *141*, 8541−8549.
(18) Huš, M.; Grilc, M.; Pavlišič, A.; Likozar, B.; Hellman, A. Multiscale Modelling from Quantum Level to Reactor Scale: An Example of Ethylene Epoxidation on Silver Catalysts. *Catal. Today* **2019**, *338*, 128−140.
(19) Singh, S.; Li, S.; Carrasquillo-Flores, R.; Alba-Rubio, A. C.; Dumesic, J. A.; Mavrikakis, M. Formic Acid Decomposition on Au Catalysts: DFT, Microkinetic Modeling, and Reaction Kinetics Experiments. *AIChE J.* **2014**, *60*, 1303−1319.
(20) Kopač, D.; Huš, M.; Ogrizek, M.; Likozar, B. Kinetic Monte Carlo Simulations of Methanol Synthesis from Carbon Dioxide and Hydrogen on Cu(111) Catalysts: Statistical Uncertainty Study. *J. Phys. Chem. C* **2017**, *121*, 17941−17949.
(21) Teschner, D.; Novell-Leruth, G.; Farra, R.; Knop-Gericke, A.; Schlögl, R.; Szentmiklósi, L.; Hevia, M. G.; Soerijanto, H.; Schomäcker, R.; Pérez-Ramírez, J.; López, N. In Situ Surface Coverage Analysis of RuO_2-Catalysed HCl Oxidation Reveals the Entropic Origin of Compensation in Heterogeneous Catalysis. *Nat. Chem.* **2012**, *4*, 739−745.
(22) Nikbin, N.; Caratzoulas, S.; Vlachos, D. G. A First Principles-Based Microkinetic Model for the Conversion of Fructose to 5-Hydroxymethylfurfural. *ChemCatChem.* **2012**, *4*, 504−511.
(23) Li, Q.; García-Muelas, R.; López, N. Microkinetics of Alcohol Reforming for H_2 Production from a FAIR Density Functional Theory Database. *Nat. Commun.* **2018**, *9*, 526.
(24) Frei, M. S.; Mondelli, C.; García-Muelas, R.; Kley, K. S.; Puértolas, B.; López, N.; Safonova, O. V.; Stewart, J. A.; Curulla Ferré, D.; Pérez-Ramírez, J. Atomic-Scale Engineering of Indium Oxide Promotion by Palladium for Methanol Production via CO_2 Hydrogenation. *Nat. Commun.* **2019**, *10*, 3377.
(25) Piccinin, S.; Stamatakis, M. Steady-State CO Oxidation on Pd(111): First-Principles Kinetic Monte Carlo Simulations and Microkinetic Analysis. *Top. Catal.* **2017**, *60*, 141−151.
(26) Huš, M.; Hellman, A. Ethylene Epoxidation on Ag(100), Ag(110), and Ag(111): A Joint Ab Initio and Kinetic Monte Carlo Study and Comparison with Experiments. *ACS Catal.* **2019**, *9*, 1183−1196.
(27) Ovesen, C. V.; Clausen, B. S.; Hammershøi, B. S.; Steffensen, G.; Askgaard, T.; Chorkendorff, I.; Nørskov, J. K.; Rasmussen, P. B.; Stoltze, P.; Taylor, P. A Microkinetic Analysis of the Water−Gas Shift Reaction under Industrial Conditions. *J. Catal.* **1996**, *158*, 170−180.
(28) Campbell, C. T. Future Directions and Industrial Perspectives Micro- and Macro-Kinetics: Their Relationship in Heterogeneous Catalysis. *Top. Catal.* **1994**, *1*, 353−366.
(29) Campbell, C. T. The Degree of Rate Control: A Powerful Tool for Catalysis Research. *ACS Catal.* **2017**, *7*, 2770−2779.
(30) Cortright, R. D.; Dumesic, J. A. Kinetics of Heterogeneous Catalytic Reactions: Analysis of Reaction Schemes. *Adv. Catal.* **2001**, *46*, 161−264.
(31) Kozuch, S.; Shaik, S. A Combined Kinetic-Quantum Mechanical Model for Assessment of Catalytic Cycles: Application to Cross-Coupling and Heck Reactions. *J. Am. Chem. Soc.* **2006**, *128*, 3355−3365.
(32) Wolcott, C. A.; Medford, A. J.; Studt, F.; Campbell, C. T. Degree of Rate Control Approach to Computational Catalyst Screening. *J. Catal.* **2015**, *330*, 197−207.
(33) Rangarajan, S.; Maravelias, C. T.; Mavrikakis, M. Sequential-Optimization-Based Framework for Robust Modeling and Design of Heterogeneous Catalytic Systems. *J. Phys. Chem. C* **2017**, *121*, 25847−25863.
(34) Brezny, A. C.; Landis, C. R. Development of a Comprehensive Microkinetic Model for Rh(Bis(Diazaphospholane))-Catalyzed Hydroformylation. *ACS Catal.* **2019**, *9*, 2501−2513.
(35) 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 Catal.* **2019**, *9*, 4804−4809.
(36) Sutton, J. E.; Vlachos, D. G. Building Large Microkinetic Models with First-Principles' Accuracy at Reduced Computational Cost. *Chem. Eng. Sci.* **2015**, *121*, 190−199.
(37) Nørskov, J. K.; Bligaard, T.; Rossmeisl, J.; Christensen, C. H. Towards the Computational Design of Solid Catalysts. *Nat. Chem.* **2009**, *1*, 37−46.
(38) Sabatier, P. The Method of Direct Hydrogenation by Catalysis. *Nobel Lect.* **1912**.
(39) Balandin, A. A. Modern State of the Multiplet Theor of Heterogeneous Catalysis1. *Adv. Catal.* **1969**, *19*, 1−210.

(40) Hammett, L. P. The Effect of Structure upon the Reactions of Organic Compounds. Benzene Derivatives. *J. Am. Chem. Soc.* **1937**, *59*, 96−103.
(41) Hammond, G. S. A Correlation of Reaction Rates. *J. Am. Chem. Soc.* **1955**, *77*, 334−338.
(42) Brönsted, J. N.; Pedersen, K. Die Katalytische Zersetzung Des Nitramids Und Ihre Physikalisch-Chemische Bedeutung. *Z. Phys. Chem.* **1924**, *108U*, 185−235.
(43) Evans, M. G.; Polanyi, M. Further Considerations on the Thermodynamics of Chemical Equilibria and Reaction Rates. *Trans. Faraday Soc.* **1936**, *32*, 1333−1360.
(44) Nørskov, J. K.; Bligaard, T.; Logadottir, A.; Bahn, S.; Hansen, L. B.; Bollinger, M.; Bengaard, H.; Hammer, B.; Sljivancanin, Z.; Mavrikakis, M.; Xu, Y.; Dahl, S.; Jacobsen, C. J. H. Universality in Heterogeneous Catalysis. *J. Catal.* **2002**, *209*, 275−278.
(45) Mazeau, E. J.; Satpute, P.; Blöndal, K.; Goldsmith, C. F.; West, R. H. Automated Mechanism Generation Using Linear Scaling Relationships and Sensitivity Analyses Applied to Catalytic Partial Oxidation of Methane. *ACS Catal.* **2021**, *11*, 7114−7125.
(46) Majumdar, P.; Greeley, J. Generalized Scaling Relationships on Transition Metals: Influence of Adsorbate-Coadsorbate Interactions. *Phys. Rev. Mater.* **2018**, *2*, 045801.
(47) Medford, A. J.; Vojvodic, A.; Hummelshøj, J. S.; Voss, J.; Abild-Pedersen, F.; Studt, F.; Bligaard, T.; Nilsson, A.; Nørskov, J. K. From the Sabatier Principle to a Predictive Theory of Transition-Metal Heterogeneous Catalysis. *J. Catal.* **2015**, *328*, 36−42.
(48) Valter, M.; dos Santos, E. C.; Pettersson, L. G. M.; Hellman, A. Selectivity of the First Two Glycerol Dehydrogenation Steps Determined Using Scaling Relationships. *ACS Catal.* **2021**, *11*, 3487−3497.
(49) Jørgensen, M.; Grönbeck, H. Scaling Relations and Kinetic Monte Carlo Simulations To Bridge the Materials Gap in Heterogeneous Catalysis. *ACS Catal.* **2017**, *7*, 5054−5061.
(50) Rankin, R. B.; Greeley, J. Trends in Selective Hydrogen Peroxide Production on Transition Metal Surfaces from First Principles. *ACS Catal.* **2012**, *2*, 2664−2672.
(51) Wu, H.; Sutton, J. E.; Guo, W.; Vlachos, D. G. Volcano Curves for in Silico Prediction of Mono- and Bifunctional Catalysts: Application to Ammonia Decomposition. *J. Phys. Chem. C* **2019**, *123*, 27097−27104.
(52) Pérez-Ramírez, J.; López, N. Strategies to Break Linear Scaling Relationships. *Nat. Catal.* **2019**, *2*, 971−976.
(53) Gu, G. H.; Mullen, C. A.; Boateng, A. A.; Vlachos, D. G. Mechanism of Dehydration of Phenols on Noble Metals via First-Principles Microkinetic Modeling. *ACS Catal.* **2016**, *6*, 3047−3055.
(54) Toftelund, A.; Man, I. C.; Hansen, H. A.; Abild-Pedersen, F.; Bligaard, T.; Rossmeisl, J.; Studt, F. Volcano Relations for Oxidation of Hydrogen Halides over Rutile Oxide Surfaces. *ChemCatChem.* **2012**, *4*, 1856−1861.
(55) Bruix, A.; Margraf, J. T.; Andersen, M.; Reuter, K. First-Principles-Based Multiscale Modelling of Heterogeneous Catalysis. *Nat. Catal.* **2019**, *2*, 659−670.
(56) Zhang, Z.; Zandkarimi, B.; Alexandrova, A. N. Ensembles of Metastable States Govern Heterogeneous Catalysis on Dynamic Interfaces. *Acc. Chem. Res.* **2020**, *53*, 447−458.
(57) Falsig, H.; Hvolbæk, B.; Kristensen, I. S.; Jiang, T.; Bligaard, T.; Christensen, C. H.; Nørskov, J. K. Trends in the Catalytic CO Oxidation Activity of Nanoparticles. *Angew. Chemie - Int. Ed.* **2008**, *47*, 4835−4839.
(58) Andersen, M.; Levchenko, S. V.; Scheffler, M.; Reuter, K. Beyond Scaling Relations for the Description of Catalytic Materials. *ACS Catal.* **2019**, *9*, 2752−2759.
(59) Matera, S.; Schneider, W. F.; Heyden, A.; Savara, A. Progress in Accurate Chemical Kinetic Modeling, Simulations, and Parameter Estimation for Heterogeneous Catalysis. *ACS Catal.* **2019**, *9*, 6624−6647.
(60) Bhandari, S.; Rangarajan, S.; Mavrikakis, M. Combining Computational Modeling with Reaction Kinetics Experiments for Elucidating the in Situ Nature of the Active Site in Catalysis. *Acc. Chem. Res.* **2020**, *53*, 1893−1904.
(61) Xu, L.; Stangland, E. E.; Dumesic, J. A.; Mavrikakis, M. Hydrodechlorination of 1,2-Dichloroethane on Platinum Catalysts: Insights from Reaction Kinetics Experiments, Density Functional Theory, and Microkinetic Modeling. *ACS Catal.* **2021**, *11*, 7890−7905.
(62) Nørskov, J. K.; Bligaard, T.; Logadottir, A.; Kitchin, J. R.; Chen, J. G.; Pandelov, S.; Stimming, U. Trends in the Exchange Current for Hydrogen Evolution. *J. Electrochem. Soc.* **2005**, *152*, J23.
(63) Andersson, M. P.; Bligaard, T.; Kustov, A.; Larsen, K. E.; Greeley, J.; Johannessen, T.; Christensen, C. H.; Nørskov, J. K. Toward Computational Screening in Heterogeneous Catalysis: Pareto-Optimal Methanation Catalysts. *J. Catal.* **2006**, *239*, 501−506.
(64) Artrith, N. Learning What Makes Catalysts Good. *Matter* **2020**, *3*, 985−986.
(65) Álvarez-Moreno, M.; De Graaf, C.; López, N.; Maseras, F.; Poblet, J. M.; Bo, C. Managing the Computational Chemistry Big Data Problem: The IoChem-BD Platform. *J. Chem. Inf. Model.* **2015**, *55*, 95−103.
(66) Bo, C.; Maseras, F.; López, N. The Role of Computational Results Databases in Accelerating the Discovery of Catalysts. *Nat. Catal.* **2018**, *1*, 809−810.
(67) Computation and Machine Learning for Chemistry; available via the Internet at: https://www.nature.com/collections/gcijejjahe (accessed June 22, 2021).
(68) Toyao, T.; Maeno, Z.; Takakusagi, S.; Kamachi, T.; Takigawa, I.; Shimizu, K. I. Machine Learning for Catalysis Informatics: Recent Applications and Prospects. *ACS Catal.* **2020**, *10*, 2260−2297.
(69) Sanchez-Lengeling, B.; Aspuru-Guzik, A. Inverse Molecular Design Using Machine Learning:Generative Models for Matter Engineering. *Science* **2018**, *361*, 360−365.
(70) Butler, K. T.; Davies, D. W.; Cartwright, H.; Isayev, O.; Walsh, A. Machine Learning for Molecular and Materials Science. *Nature* **2018**, *559*, 547−555.
(71) Bartók, A. P.; De, S.; Poelking, C.; Bernstein, N.; Kermode, J. R.; Csányi, G.; Ceriotti, M. Machine Learning Unifies the Modeling of Materials and Molecules. *Sci. Adv.* **2017**, *3*, No. e1701816.
(72) Naik, R. R.; Tiihonen, A.; Thapa, J.; Batali, C.; Liu, Z.; Sun, S.; Buonassisi, T. Discovering Equations That Govern Experimental Materials Stability under Environmental Stress Using Scientific Machine Learning. *arXiv* **2021**, https://arxiv.org/abs/2106.10951v1.
(73) Ramprasad, R.; Batra, R.; Pilania, G.; Mannodi-Kanakkithodi, A.; Kim, C. Machine Learning in Materials Informatics: Recent Applications and Prospects. *npj Comput. Mater.* **2017**, *3*, 54.
(74) Jennings, P. C.; Lysgaard, S.; Hummelshøj, J. S.; Vegge, T.; Bligaard, T. Genetic Algorithms for Computational Materials Discovery Accelerated by Machine Learning. *npj Comput. Mater.* **2019**, *5*, 46.
(75) Liu, X.; Xiao, J.; Peng, H.; Hong, X.; Chan, K.; Nørskov, J. K. Understanding Trends in Electrochemical Carbon Dioxide Reduction Rates. *Nat. Commun.* **2017**, *8*, 15438.
(76) Gu, G. H.; Noh, J.; Kim, S.; Back, S.; Ulissi, Z.; Jung, Y. Practical Deep-Learning Representation for Fast Heterogeneous Catalyst Screening. *J. Phys. Chem. Lett.* **2020**, *11*, 3185−3191.
(77) Tran, K.; Neiswanger, W.; Broderick, K.; Xing, E.; Schneider, J.; Ulissi, Z. W. Computational Catalyst Discovery: Active Classification through Myopic Multiscale Sampling. *J. Chem. Phys.* **2021**, *154*, 124118.
(78) Esterhuizen, J. A.; Goldsmith, B. R.; Linic, S. Theory-Guided Machine Learning Finds Geometric Structure-Property Relationships for Chemisorption on Subsurface Alloys. *Chem.* **2020**, *6*, 3100−3117.
(79) García-Muelas, R.; López, N. Statistical Learning Goes beyond the D-Band Model Providing the Thermochemistry of Adsorbates on Transition Metals. *Nat. Commun.* **2019**, *10*, 4687.
(80) Foppa, L.; Ghiringhelli, L. M.; Girgsdies, F.; Hashagen, M.; Kube, P.; Hävecker, M.; Carey, S. J.; Tarasov, A.; Kraus, P.; Rosowski, F.; Schlögl, R.; Trunschke, A.; Scheffler, M. Materials Genes of

Heterogeneous Catalysis from Clean Experiments and Artificial Intelligence. *MRS Bull.* **2021**, DOI: 10.1557/s43577-021-00165-6.
(81) Meyer, B.; Sawatlon, B.; Heinen, S.; Von Lilienfeld, O. A.; Corminboeuf, C. Machine Learning Meets Volcano Plots: Computational Discovery of Cross-Coupling Catalysts. *Chem. Sci.* **2018**, *9*, 7069−7077.
(82) O'Connor, N. J.; Jonayat, A. S. M.; Janik, M. J.; Senftle, T. P. Interaction Trends between Single Metal Atoms and Oxide Supports Identified with Density Functional Theory and Statistical Learning. *Nat. Catal.* **2018**, *1*, 531−539.
(83) Zahrt, A. F.; Henle, J. J.; Rose, B. T.; Wang, Y.; Darrow, W. T.; Denmark, S. E. Prediction of Higher-Selectivity Catalysts by Computer-Driven Workflow and Machine Learning. *Science* **2019**, *363*, No. eaau5631.
(84) Felton, K. C.; Rittig, J. G.; Lapkin, A. A. Summit: Benchmarking Machine Learning Methods for Reaction Optimisation. *Chemistry−Methods* **2021**, *1*, 116−122.
(85) Xiong, J.; Shi, S. Q.; Zhang, T. Y. A Machine-Learning Approach to Predicting and Understanding the Properties of Amorphous Metallic Alloys. *Mater. Des.* **2020**, *187*, 108378.
(86) Sutton, J. E.; Guo, W.; Katsoulakis, M. A.; Vlachos, D. G. Effects of Correlated Parameters and Uncertainty in Electronic-Structure-Based Chemical Kinetic Modelling. *Nat. Chem.* **2016**, *8*, 331−337.
(87) Hibbert, D. B.; Armstrong, N. An Introduction to Bayesian Methods for Analyzing Chemistry Data: Part II: A Review of Applications of Bayesian Methods in Chemistry. *Chemom. Intell. Lab. Syst.* **2009**, *97*, 211−220.
(88) Pedersen, J. K.; Clausen, C. M.; Krysiak, O. A.; Xiao, B.; Batchelor, T. A. A.; Löffler, T.; Mints, V. A.; Banko, L.; Arenz, M.; Savan, A.; Schuhmann, W.; Ludwig, A.; Rossmeisl, J. Bayesian Optimization of High-Entropy Alloy Compositions for Electrocatalytic Oxygen Reduction**. *Angew. Chemie Int. Ed.* **2021**, *60*, 24144−24152.
(89) Rudy, S. H.; Brunton, S. L.; Proctor, J. L.; Kutz, J. N. Data-Driven Discovery of Partial Differential Equations. *Sci. Adv.* **2017**, *3*, No. e1602614.
(90) Guimerà, R.; Reichardt, I.; Aguilar-Mogas, A.; Massucci, F. A.; Miranda, M.; Pallarès, J.; Sales-Pardo, M. A Bayesian Machine Scientist to Aid in the Solution of Challenging Scientific Problems. *Sci. Adv.* **2020**, *6*, No. eaav6971.
(91) Gallarati, S.; Fabregat, R.; Laplaza, R.; Bhattacharjee, S.; Wodrich, M. D.; Corminboeuf, C. Reaction-Based Machine Learning Representations for Predicting the Enantioselectivity of Organocatalysts. *Chem. Sci.* **2021**, *12*, 6879−6889.
(92) Coley, C. W.; Jin, W.; Rogers, L.; Jamison, T. F.; Jaakkola, T. S.; Green, W. H.; Barzilay, R.; Jensen, K. F. A Graph-Convolutional Neural Network Model for the Prediction of Chemical Reactivity. *Chem. Sci.* **2019**, *10*, 370−377.
(93) Schwaller, P.; Vaucher, A. C.; Laino, T.; Reymond, J.-L. Prediction of Chemical Reaction Yields Using Deep Learning. *Mach. Learn. Sci. Technol.* **2021**, *2*, 015016.
(94) Hueffel, J. A.; Sperger, T.; Funes-Ardoiz, I.; Ward, J. S.; Rissanen, K.; Schoenebeck, F. Accelerated Dinuclear Palladium Catalyst Identification through Unsupervised Machine Learning. *Science* **2021**, *374*, 1134−1140.
(95) Saadun, A. J.; Pablo-García, S.; Paunović, V.; Li, Q.; Sabadell-Rendón, A.; Kleemann, K.; Krumeich, F.; López, N.; Pérez-Ramírez, J. Performance of Metal-Catalyzed Hydrodebromination of Dibromomethane Analyzed by Descriptors Derived from Statistical Learning. *ACS Catal.* **2020**, *10*, 6129−6143.
(96) Saadun, A. J.; Zichittella, G.; Paunović, V.; Markaide-Aiastui, B. A.; Mitchell, S.; Pérez-Ramírez, J. Epitaxially Directed Iridium Nanostructures on Titanium Dioxide for the Selective Hydrodechlorination of Dichloromethane. *ACS Catal.* **2020**, *10*, 528−542.
(97) Saadun, A. J.; Kaiser, S. K.; Ruiz-Ferrando, A.; Pablo-García, S.; Büchele, S.; Fako, E.; López, N.; Pérez-Ramírez, J. Nuclearity and Host Effects of Carbon-Supported Platinum Catalysts for Dibromomethane Hydrodebromination. *Small* **2021**, *17*, 2005234.
(98) Zichittella, G.; Pérez-Ramírez, J. Status and Prospects of the Decentralised Valorisation of Natural Gas into Energy and Energy Carriers. *Chem. Soc. Rev.* **2021**, *50*, 2984−3012.
(99) Kresse, G.; Furthmüller, J. Efficiency of Ab-Initio Total Energy Calculations for Metals and Semiconductors Using a Plane-Wave Basis Set. *Comput. Mater. Sci.* **1996**, *6*, 15−50.
(100) Perdew, J. P.; Burke, K.; Ernzerhof, M. Generalized Gradient Approximation Made Simple. *Phys. Rev. Lett.* **1996**, *77*, 3865−3868.
(101) Grimme, S.; Ehrlich, S.; Goerigk, L. Effect of the Damping Function in Dispersion Corrected Density Functional Theory. *J. Comput. Chem.* **2011**, *32*, 1456−1465.
(102) Almora-Barrios, N.; Carchini, G.; Błoński, P.; López, N. Costless Derivation of Dispersion Coefficients for Metal Surfaces. *J. Chem. Theory Comput.* **2014**, *10*, 5002−5009.
(103) Blöchl, P. E. Projector Augmented-Wave Method. *Phys. Rev. B* **1994**, *50*, 17953−17979.
(104) Monkhorst, H. J.; Pack, J. D. Special Points for Brillouin-Zone Integrations. *Phys. Rev. B* **1976**, *13*, 5188−5192.
(105) Neugebauer, J.; Scheffler, M. Adsorbate-Substrate and Adsorbate-Adsorbate Interactions of Na and K Adlayers on Al(111). *Phys. Rev. B* **1992**, *46*, 16067.
(106) Burcat, A.; Ruscic, B.; Chemistry. Third Millenium Ideal Gas and Condensed Phase Thermochemical Database for Combustion (with Update from Active Thermochemical Tables). Technical Report ANL-05/20, Argonne National Laboratory, Argonne, IL, **2005**.
(107) Anderson, J. R.; Boudart, M. Catalysis. In *CATALYSIS—Science and Technology*; Anderson, J. R., Boudart, M., Eds.; Springer: Berlin, Heidelberg, 1996; Vol. *10*.
(108) Mears, D. E. Diagnostic Criteria for Heat Transport Limitations in Fixed Bed Reactors. *J. Catal.* **1971**, *20*, 127−131.
(109) Pedregosa, F.; Michel, V.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Vanderplas, J.; Cournapeau, D.; Varoquaux, G.; Gramfort, A.; Thirion, B.; Grisel, O.; Dubourg, V.; Passos, A.; Brucher, M.; Perrot, M.; Duchesnay, É. Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825−2830.
(110) The robustnes measures the sensitivity and the precision of the descriptor. The robustness parameter is calculated the standard error of the descriptor prediction, and gives information on its variability.
(111) Roman, T.; Groß, A. Periodic Density-Functional Calculations on Work-Function Change Induced by Adsorption of Halogens on Cu(111). *Phys. Rev. Lett.* **2013**, *110* (15), 156804.
(112) Roman, T.; Gossenberger, F.; Forster-Tonigold, K.; Groß, A. Halide Adsorption on Close-Packed Metal Electrodes. *Phys. Chem. Chem. Phys.* **2014**, *16*, 13630−13634.
(113) Regions over 100% in Figure 4 are unexplored zones far from the input points. Furthermore, there is no halogen-metal in our dataset (including F and I) contained in the >100% domain.
(114) Guthrie, J. P. Multidimensional Marcus Theory: An Analysis of Concerted Reactions. *J. Am. Chem. Soc.* **1996**, *118*, 12878−12885.
(115) The cosine term indicates that our predictions for S_{coke} is overfitted due to the difference between Ru-Br and Ru-Cl values, as the RF and GR predictions. The only way to avoid overfitting is increasing the sample size, which is not possible due to the limited points in the Periodic Table.