Dimensionality reduction of complex reaction networks in heterogeneous catalysis: From linear-scaling relationships to statistical learning techniques

Sergio Pablo-García | **Rodrigo García-Muelas** | **Albert Sabadell-Rendón** | **Núria López**

Institute of Chemical Research of Catalonia, The Barcelona Institute of Science and Technology, Tarragona, Spain

Correspondence
Núria López, Institute of Chemical Research of Catalonia, The Barcelona Institute of Science and Technology, Av. Països Catalans 16, 43007 Tarragona, Spain.
Email: nlopez@iciq.es

Funding information
Ministerio de Ciencia e Innovación, Grant/Award Number: RTI2018-101394-B-I00

Abstract

The mechanistic analysis in heterogeneous catalysis is based on listing all elementary steps and evaluating explicitly their energies. To this end, computational models based on Density Functional Theory have become a standard to estimate the information needed in mechanistic studies. Typically, either the minimum energy paths or those with the smaller span are summarized in reaction profiles. Such simplifications gather a lot of information, although further dimensionality reduction is required to obtain the most relevant descriptors of catalytic activity and generate the so-called volcano plots. The selection of descriptors has been traditionally based on simple intermediates, such as central atoms in small molecules (as C in CH_4), which have good thermodynamic correlations to other fragments containing them. Yet, in emerging processes (recent studies), the number of intermediates involved increase, configurational effects and lateral interactions become significant, and complex materials with low symmetry are employed, thus the simple rules encapsulated in linear scaling relationships lose their predictive power due to error accumulation. At the same time, large datasets generated for the intermediates call for statistical analysis and thus these techniques are being leveraged to chemical systems, particularly to reduce their dimensionality.

This article is categorized under:
Structure and Mechanism > Reaction Mechanisms and Catalysis
Structure and Mechanism > Computational Materials Science
Electronic Structure Theory > Ab Initio Electronic Structure Methods

KEYWORDS

Density Functional Theory, descriptors, dimensionality reduction, heterogeneous catalysis, linear-scaling relationships

1 | INTRODUCTION

Many industrially relevant chemical processes rely on the activation of a few atoms with a set of common intermediates characterized by a few main atoms, like the Haber–Bosch process for generating ammonia.[1,2] As raw materials taken as reactants become more complex,[3] selectivity issues start to be central to attain performance[4–6] since separations are ecologically and economically costly. In addition, stability is the ultimate challenge that limits catalytic implementation and, although widely academically neglected, it constitutes a must in the design of new processes.[7–9] Finally, catalytic architectures at the industrial level are characterized by complexity including the active material, carrier,[10] molecular modifiers[11] or dopants[12] plus the binders[13,14] while computational models barely go beyond pure crystals and simple orientations.[15] Therefore, complexity is an intrinsic parameter to catalysis and arises from three sources: the catalyst's structure, the reaction network, and environmental factors: (i) The catalyst's complexity appears through in homogeneous structures with multiple potential active sites; (ii) The reaction network leading from reactants to desired and unwanted products may have hundreds or thousands of intermediates and transition states; (iii) Environmental effects can be caused by pressure, local concentrations being different from the bulk of the fluid phase, external forces.[16] Their combination masks our understanding of catalytic processes.

Linear-scaling relationships, LSR, generate constraints in the available chemical space that define volcanoes for activity and selectivity. Thus, catalyst optimization relies in our ability to break such limitations.[17,18] Although this can be done effectively in labs, the search for a wider phase-space and the relatively small knowledge of complex materials has severely limited our ability to predict their catalytic properties. Heuristics have been at the core of catalytic development and large experimental databases were even generated in the early days of catalysis such as the Haber–Bosch process. This heuristic knowledge has been mainly stored in companies while academia has discouraged the publication of negative results. In the later years, there has been a change in paradigm.[19,20] Besides, computational results can now be obtained cheaply and systematically for simple but widely used families of materials, including metals and alloys. DFT can be employed to extract energies of phases that might not exist under reaction conditions due to phase changes or poisoning, and this widens our ability to interrogate the phase space more comprehensively. Spanning a larger range of, for instance, adsorption energies, simplifies the evaluation of the properties and descriptor identification. Finally, negative results can be generated with minor cost. This has paved the way for integrating statistical learning (machine learning, ML) techniques into computational heterogeneous catalysis. ML can simplify complexity through different perspectives.[21] The leverage of these techniques to chemical systems has only been possible after many years due to the following: (i) ability to generate wide set of data with significant similar error independent of the code[22]; (ii) emergence of databases with computational data (or where to introduce this computational data)[19,23–27]; (iii) the awareness to move towards open science models.[20] In this context, it is interesting that such approaches have been already proved successful in biology of proteins where large databases for crystal structures have been available for the last 50 years allowing the very recent successful implementation of artificial intelligence (AI) algorithms.[28] In heterogeneous catalysis, like in many other communities, there is a need for AI algorithms that ensure their interpretability. This would allow to provide new synthetic routes while retaining the scientific understanding on the process. In the following we present the most acute issues to address complexity and how dimensionality reduction tools coming from statistical learning techniques can allow a more robust approach.

2 | LINEAR SCALING RELATIONSHIPS

Sabatier identified that the rate of some reactions depends on a single parameter, namely oxide formation, which is the first descriptor for activity.[29] These so-called volcano plots were introduced by Balandin[30] However, descriptors have been rather elusive and they have mostly been employed after extensive Density Functional Theory simulations. The reason for that is that in many cases phases change for the terminal parts of a given descriptor (i.e., for too exothermic O adsorption energies the system is likely to evolve to the oxide phase). Even DFT systems can provide data for ideal scenarios where a phase that does not exist under reaction conditions but its properties can be computed, thus decoupling phase stability and reactivity in an effective manner.

2.1 | Descriptors à-la-antique

There are two main families of linear scaling relationships as described below. In the first one, structural or topological features define the thermochemistry (e.g., describing the adsorption energy of intermediates). In the second one, the thermochemistry defines the kinetics.

Early thermochemical models were based on group additivity rules and applied widely on molecular systems.[31] On heterogeneous catalysis, their application started on reactions involving hydrogenation of rather inert species, like the Haber–Bosch process and methanation. These could be extensively computed by the 1990 and ended up with simplified descriptors, the corresponding central atoms (as N in NH_3 or C in CH_x).[32–37] Thus, for the activity in NH_3 formation, N_2 is physisorbed and the NH_x are found to depend on the energy of the Nitrogen atom. Although the adsorption energy of H is sometimes used as descriptor,[38] it also correlates to that of N or C on transition metals[35,39] thus leaving N as the unique descriptor. The origin for the dependence of NH_x, $x = 1–3$ with N can be related to valence considerations and therefore the relative energies ended up wrapping up to a single term.[32] These dependencies on the properties of metals where further extended to relationships between central atoms[34] and to structural effects.[40,41] The electronic structure of molecular systems can be much more convoluted and thus LSR have not been derived till very recently.[42] Extensions to oxides, nitrides and other compounds were also proposed[43] but, due to the semiconductor nature of some of these materials further refinements are needed.[44,45]

In oxidation processes the equations turned out not to be so straightforward and thus multivariable approaches were proposed.[46] For CO oxidation, it was found that O binding energy as a single descriptor was insufficient, and thus heuristically at least a second contribution was needed, namely CO. More recently, it was found that the two descriptors come from the decomposition of the metal-adsorbate chemical bond into covalent and electrostatic terms, and the chemical space of adsorbates on pure metals is 2D.[39,47] Two-dimensional plots have also been employed in metal oxides, particularly in the oxygen evolution reaction.[48,49]

2.2 | Kinetics as a function of thermodynamic parameters

Besides purely structural descriptors defining thermochemistry, kinetic parameters can also be approximated from the former ones following the so-called Bell-, or Brønsted–Evans–Polanyi relations.[50–54] There, the activation energy of a given family of elementary step (e.g., hydrogenations), E_a is put as a linear function of the reaction energy, ΔE. The factor multiplying ΔE is exactly 0.5 for symmetric reactions, such in S_N2, and approaches either 0.0 or 1.0 if the transition state structure resembles more the initial or final states respectively. These scaling relationships were later on extended to put the energy of the transition state E_{TS} as a function of either the initial or final states, E_{IS} or E_{FS}.[4,55–57] The factors multiplying E_{IS} or E_{FS} were shown to be necessarily 1.0 in order to be universal; this is, independent on the chosen energy references.[57] Older qualitative approaches, nonlinear methods such as Shustorovich Bond Order Conservation theory[58] and the UBI-QEP method[59] have been trying to rationalize the origin of such structure, thermochemistry, and kinetics correlations in terms of bonds.

3 | MULTI-SCALE MODELING

To get observables directly comparable to experiments, the energy profiles obtained via DFT have often been used as input to multi-scale models.[6,60–63] In those models, the LSR devised in the previous section have been employed to approximate the input parameters of the ordinary differential equations (ODE) that define the chemical kinetics of a particular mechanism, and thus obtain the volcano plots. Here we describe the most popular among these methods: Microkinetics (MK), Kinetic Monte Carlo (KMC), and computational fluid dynamics (CFD) simulations. MK is a mean-field method that solves the entire coupled ODE system defined by each one of the different balances present in a reaction network,[63] and the boundary conditions of the experimental settings. MK can provide the composition as a function of time and can take the kinetic parameters from LSR. Thus, volcano plots can be generated by coupling simple DFT and MK via LSR. MK models can provide valuable insights in both homogeneous and heterogeneous catalytic systems. For instance, precise reproduction of experimental kinetic data has been achieved for the condensation of n-butylamine and benzaldehyde[64] and Rh-based hydroformylation.[65] In single atom catalysis, MK has been applied to investigate the oxygen reduction reaction (ORR) with metal doped graphene.[66] In heterogeneous catalysis, several examples exists, such as alcohol reforming[57] or ethylene epoxidation on Ag.[67] The limitations in the use of DFT or LSR-derived energies became also

evident in MK modeling.[5] For instance, in formic acid decomposition on Au/SiC and Pt/C, Mavrikakis et al.[68] tuned iteratively the DFT parameters by introducing coverage effects and improving the active site model until they reach the experimental values. Kinetic parameters can also be estimated employing Bayesian Statistics.[62] Those errors can be even bigger if employing LSR to account for the dependences between different adsorption energies or barriers. This points out towards the needs of much more accurate LSR and descriptors through ML techniques.

For highly anisotropic systems the spatial information is crucial and thus Kinetic Monte Carlo is needed.[69,70] Instead of solving the ODE system defined in pure MK, KMC calculates the probabilities to transform the lattice current state to all possible future states. These are related to the kinetic constant, generally via Eyring equation, where the thermodynamic parameters are obtained by DFT. Next, the following state is selected, the lattice is updated, and the simulation time is advanced.[69–71] KMC has been implemented in several different codes, such as SPARKS,[72] ZACROS,[73] and kmos.[74] In transition metals, for example, KMC has been used to study the water shift reaction on Pd(100).[75] In alloys, the mobility of atoms on metal surfaces[76] for CO oxidation on $RuO_2(110)$[77,78] or to find Cu and Fe percolation with amine arylation activity on graphitic carbon nitride.[79] The main drawback of this approach is the number of configurations for which DFT simulations are needed. The number of barriers to be evaluated is explosive, limiting the use of these codes. The energies for intermediates can be found from cluster expansions, while LSR can be employed to estimate the barrier without performing the DFT transition state search itself. This explains why descriptors of simple systems are found even if DFT evaluations are done in the low coverage regime.

MK,[80] or, in some cases, KMC solvers[81] can be coupled to CFD codes that solve numerically the possible balances present on the system, for example, mass, momentum, or energy. The governing equations are a set of partial differential equations (PDE) and ODE.[82–84] Applications include CH_4 partial oxidation on Rh,[85] CO oxidation on $RuO_2(110)$ and Pd(100),[81] and ethylene oxidation on silver.[67] In summary, as the number of intermediates and reactions rises, the system becomes more difficult (or virtually impossible) to tackle with DFT coupled to MK, KMC or CFD alone. Likely ML techniques with adequate uncertainty estimation can ensure a seamless input to multi-scale modeling.[86,87]

4 | AUTOMATED NETWORK GENERATION AND ANALYSIS

First, reaction networks present strong dependencies in the nature of the intermediates adsorbed on the surface.[5,6] This can be seen as an intrinsically symmetry due to the locality of the chemical bond[4,32,54,88] and can be employed in the dimensionality reduction. As this holds for many intermediates, the structures linking them are also subjected to these dependencies and thus linear scaling relationships, LSR, appear both for the energies of intermediates and transition states (Figure 1). However, as reaction processes are larger the number of related structures increases and this coupled to the intrinsic error associated to the linear fittings in LSR make the predictions, particularly of selectivity, more uncertain. The uncertainties in the activation energies get magnified when producing activity and selectivity, as they are introduced into an exponential term (Arrhenius equation); as a result, typical errors are approximately of 2–3 orders of magnitude.[5,6]

4.1 | Labeling and generation

The simplest set of reactions correspond to the addition or removal of an atom or moiety. Given the active role of surfaces participating in these events they are the easiest. In comparison, concerted steps are much less common catalyzed by surfaces. Ideally, more advanced reaction search needs to implement a labeling technique to easily classify the intermediates and the reactions involved in a reaction network. For example, SMILES labeling allows the codification of an entire molecule using simple text string.[89] Another important step during the generation of a network is the definition of the connectivity between the different elements of the network. Graph theory becomes a powerful tool when connectivity plays an important role in the system, allowing to model the moieties that compose the network as nodes that are connected between edges. However, the use of graph theory is not restricted to the network, molecules can also be converted to graphs, defining their atoms as nodes connected between their chemical bonds (edges). This approach simplifies representation of molecules, easing their generation and analysis.[90] Much progress has been made in this direction, materializing into tools to generate and model complex reaction networks, as for example NetGen,[91] RING,[92] RMGcat,[93] among others.[3,94] The use of these methods can be extended to other branches of chemistry and provide a complete set of analysis tools to extract information from a reaction network. For a discussion on the scope and limitations of all these methods the readers are directed to the recent review by Vernuccio and Broadbelt.[95] However, these

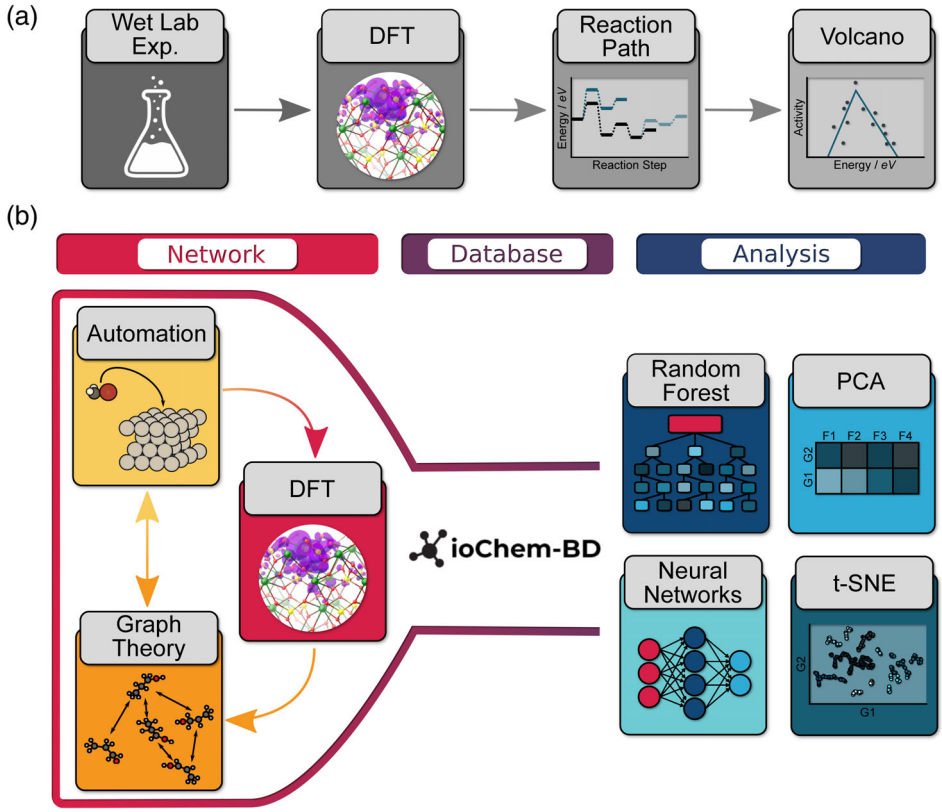

FIGURE 1 (a) Representation of the old workflows versus (b) the automated generation of reaction networks coupled with statistical post-processing. In the first approach the experimental data was complemented by density functional energy calculations to obtain the reaction paths, and the linear-scaling relationships between them. The result is the volcano plot. Alternatively, the new procedures automatically set up the calculations that are then stored in a database that are analyzed with the statistical approaches

graphs are codified to be optimal in machine language and their exploration is rather difficult, thus, requiring image techniques to be understandable by humans. This is where graphical programming languages such as DOT[96] and graph serialization tools[97] excel, generating illustrations of the network that are highly interpretable for the naked eye. Graph networks are incredibly flexible, they allow not only to focus on specific parts of the network via the generation of subgraphs but also to manipulate the size of the network and coupled to automation techniques.

4.1.1 | Automation frameworks

As the reaction network grows in complexity, the amount of DFT calculations to fully describe an entire system becomes a challenge. During these years, the need to create frameworks simplifying input generation process for massive DFT studies has emerged. Initiatives such ASE[98] or Open Babel Project[99] simplify the generation of input files, drastically reducing the time needed to prepare these files via scripting. The combination of these automation frameworks with graph theory becomes an extremely powerful tool. Graph-modeled networks can generate the connectivity of the molecules inside the network and link them via elementary steps, while automation frameworks can be used to calculate these intermediates. The generated DFT data can be then integrated inside the network and used to further expand the network, generating a positive feedback loop. As the preparation and classification of DFT data is not straightforward, workflows become imperative during the automatic data generation for large systems (more than 102–103 DFT points).[100–103] Workflows allow the application of recipes that effectively automate the energy calculations via taking the control of the decision making and error handling processes. Fireworks[104] and AiiDA[105,106] are only two examples of the available frameworks that implement these workflows in production environments. Due to the smart assemble of graphs, the information contained in those graph-modeled networks is already sorted and classified, being prone to be stored into databases.

4.1.2 | Feature extraction and databasing

As the flow of data increases, it also does the need to sort, classify and store this data to be available worldwide and understandable for everyone. Online databases and chemical repositories provide an essential service to supply this need.[107] An increasing number of initiatives started emerging last decade, some of them aiming to store general DFT data such as Nomad[24] and ioChem-BD[23] while others aiming to be more specific such as Materials Project Initiative,[108] Materials Cloud,[25] Computational Materials Repository[26] and Catalysis Hub.[109] Some of these databases, such as ioChem-BD,[23] provide an automatic refinement process, where the data provided is analyzed and preprocessed before becoming public, extracting the most valuable features and explicitly exposing them to the final users. However, the true potential of these databases lies in their role as nexus between the data generation and the data exploration/analysis. Statistical learning techniques provide a powerful tool to predict and inspect the behavior of chemical systems.[19,27] Nonetheless, the accuracy of these techniques relies on the amount and the quality of the data available to describe the system. Thus, the classification and storage provided by online databases plays a center role during the discovery process.

4.1.3 | Missing pieces in network coding

Networks in organic modeling like in prebiotic chemistry are more advanced[110] and such approaches need to be introduced for the reactivity on surfaces. The complexity of the graph-modeled networks resides in the connectivity between the fragments that compose the network. When the complexity of the intermediates that belong to the network increases, the entire process falls apart. Many efforts have been made to globally describe convoluted reaction networks[111] and the interactions between all the components of a chemical system.[112] Although a full description of a chemical system is virtually the most accurate approach to solve these problems, complex molecules tend to be computationally expensive and performing a full DFT analysis for a network involving a nonnegligible number of complex molecules and their interactions with the rest of the system is impractical. Smart chemical space sampling[113] and thermodynamic prediction via machine learning[114] are some of the solutions proposed to reduce the DFT weight of these systems. However, the biggest challenge of analyzing complex reaction networks is to predict of experimental rates and selectivities. Here, microkinetics is the most suitable tool to predict activities, selectivities, reaction orders, preferred reaction paths, and most-abundant reaction intermediates. Graph-modeled networks excel generating the full set of microkinetic equations.[93] However, the usage of automation lead to a combinatory explosion of intermediate and transition state energies that need to be considered in the microkinetic model. Graph analysis can be used to prune the microkinetic model through the extraction of specific subgraphs of the network and/or applying a kinetic criterion to the transition states. However, these models are limited, and cannot accurately predict reaction rates by themselves.[5,6] To overcome these issues, many efforts have been made to apply statistical learning techniques and to include other phenomena traditionally neglected in DFT and microkinetic models, as described next.

5 | SIMPLIFICATIONS IMPLICIT TO AUTOMATIC MODELING

In the automated multiscale modeling of heterogeneous catalytic processes complexity arises from three sources, namely (i) the catalyst's structure, (ii) the molecular complexity of the reactant, and (iii) environmental effects, Figure 2.

5.1 | Catalyst complexity

Computational models generated automatically normally assume simple crystalline systems. In contrast, real catalysts may have nonregular shapes and have different ensembles,[115] dopants can change reaction paths[12] and the intrinsic nature of active sites with clear speciation can be elusive.[116,117] More generally, a catalytic material is composed by the active phase and a support intended to be an inactive carrier and meant for the dilution of the expensive catalytic metal phase. Depending on the carrier, the “inert” simplification can be disputed. The carrier may either affect the electronic

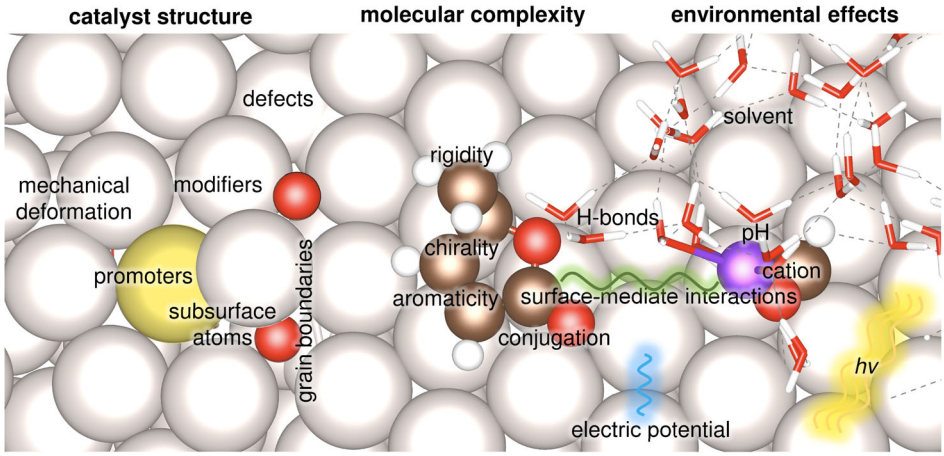

FIGURE 2 Sources of complexity in heterogeneous catalysis at the levels of the material, the molecule, and the external factors

and geometric structure of the active phase in the so-called strong metal-support interaction, or actively interact with the reactant as a co-catalyst. Depending on the particular effect, their activity would need to be taken into account in a separately from the main active site.

In addition, LSR are very well-developed for metals and alloys where the deviations from the d-band model are relatively small and the surface orientation and coordination can be approximated via metal coordination scaling.[32,41] Oxides and many other materials have complex electronic structures that need to be considered accurately.[118] However, the LSR relations have lower accuracy when applied in oxides, sulfides, nitrides and in general multicomponent phases where two of the elements have marked differences in electronegativity.[43,48,119] Seven pillars controlling the reactivity of oxides were identified by Grasselli in 2002, namely the host structure, the strength of metal-oxygen bonds, lattice oxygens, redox properties, multifunctionality, active site isolations, and phase cooperation.[119] As these effects are overlapping, the minimum descriptors needed were found to be the vacancy formation energies as well as the basicity and acidity of the reaction centers but still much research is needed in this area.[120]

5.2 | Reactant complexity

The reaction network leading from reactants to desired and unwanted products may have hundreds or thousands of intermediates and transition states, depending on the size of reactants and products. LSR in heterogeneous catalysis were developed for very simplified molecular systems, rarely containing more than two central atoms.[32,33,88] However, deviations occur when increasing the size and chemical complexity of the molecule.[31,121] First, when several alcohols and amines functional groups are present, the number of conformations grows exponentially[122] and the number of hydrogen bonds need to be maximized. Besides, functional groups may repeal each other mediated by the surface, thus breaking the additivity of thermochemical rules. Long hydrocarbon chains tend to maximize their interaction with the surface[122] and within themselves.[121] Molecules containing rings also have rigidity,[31] which constrains which parts of them may effectively interact with the catalyst.[123] The deviations are exacerbated for conjugated and aromatic molecules,[31,124] as the energies to form intramolecular double bonds shall be compared with the molecule-surface bonds.[39] Besides, such molecules may exist in an intermediate between two states depending on the metal[39] or have stable counterintuitive radical forms.[125] Moreover, chiral centers can be formed or controlled either in multifunctional catalysts or through bifunctional strategies.[126,127] Finally, future algorithms would need to detect if well-known organic-chemistry-textbook reactions could take place independently on the catalyst-mediated reactions. For instance, in aqueous phase, a large part of the adsorbates do interact with the solvent and their acid/base properties, thus the steps containing the formation of tautomers and zwitterions should be included in the reaction network. Also, some intermediates may fully desorb and react in the solvent without the mediation of the catalysts.[128] These complementary reaction sets are also needed with a balanced estimation error.

5.3 | External factors

Besides the catalyst and the reactant, the reaction environment is another source of complexity. The more pervasive of them are the solvent interactions, which can affect the potential energy of all intermediates in a given reaction network, and may fully switch the selectivity.[57,129,130] The presence of solvent may also modify diffusion of key reactants and products, thus affecting their local concentration around the surface and the local pH. Diffusion may in turn affect the coverage. Reaction profiles are typically described in the low-coverage regime for all the intermediates, considering a mean-field approximation where all lateral interactions are neglected. This greatly simplifies the definition of linear-scaling relationships although it might introduce severe deviations, particularly for large fragments. On large molecules, steric effects may also govern selectivity. For instance, for acrolein hydrogenation on Pd, it was found that more space is required to hydrogenate the C═O functional group over the C═C one, and a different final product would be found in low and high coverage regimes.[131] In materials with small pore sizes, confinement dramatically affects the properties of the fluid phase. Finally, for photo- and electrochemical applications, external potentials can excite the electronic structure of both the catalyst and the adsorbate. The combination of all these effects altogether obscures the understanding of catalytic processes and may limit the accuracy of the models unless all relevant effects are considered.

6 | DESCRIPTORS FROM MACHINE LEARNING TECHNIQUES

Finding appropriate descriptors for complex systems is not evident. To solve this issue, machine learning techniques are being introduced. This has the following benefits: (i) the descriptors are statistically robust and (ii) the descriptor acquisition is easily automated. The statistical techniques for extracting descriptors can be clustered in two main groups: feature selection and classification, and dimensionality reduction, as illustrated in Figure 3. The first group selects the most representative features without changing the final dimension, while the second one transforms those features into a lower dimension.

6.1 | Feature selection and classification

Two of the most popular feature selection techniques are the Least Absolute Shrinkage and Selection Operator (LASSO, Figure 3 panel b), and the Elastic Net Regularization (ENR). LASSO and ENR are supervised ML techniques, in which a Y response variable is needed to perform the feature selection. Those methods are also used for regression. The LASSO method is a feature selection method based in a regression (linear, logistic, or others), in which data is shrunk towards a central point, that is, the mean. LASSO adds a penalty equals to the absolute value of the coefficient (L1-regularization). Thus, regression coefficients that corresponds to correlated variables are near zero. LASSO returns simple models avoiding overfitting but if the number of points is much larger than the number of features, LASSO tends to select all the number of features in an arbitrary way. LASSO has been applied to homogeneous catalysis for predicting regioselectivities of alkenes,[132] and the electronic structure of transition metal complexes with different organic ligands,[133] and with a L2-regularization method (Kernel Ridge Regression), to select the best catalysts in cross-coupling reactions.[134] In heterogeneous catalysis, LASSO has been used for generating a method for screening new possible catalysts[135] and to investigate properties of single atom catalysis.[136,137]

The ENR method is a variation of the LASSO method with quadratic penalty term (L2-regularization). ENR shares the strong points of LASSO, with the extra benefit that solves the issue of the number of features extracted but still has the issue of the correlated variables and with a high computational cost. ENR has been applied in homogeneous catalysis for quantifying steric effects on organic molecules[138] and as benchmark regressor in the screening of (111)-bimetallic alloys in heterogeneous catalysis.[139]

Feature selection methods can also be coupled to classification methods to make predictions. Two of the most preferred ML classifiers in catalysis are Artificial Neural Networks (ANN), and Random Forest Classifier (RFC, Figure 3 panel c), which are supervised ML techniques. The ANN method consists in a set of connected input and output units (neurons), where each connection has an associated weight. During the training process, the network adjusts the weights to obtain the feature classification. ANN performance is particularly good for nonlinear data with large number of features and once trained, the ANN is a very fast method. However, ANN outputs are not interpretable (black box), and strongly depends on the training data (more than another classification or regression ML methods). Thus,

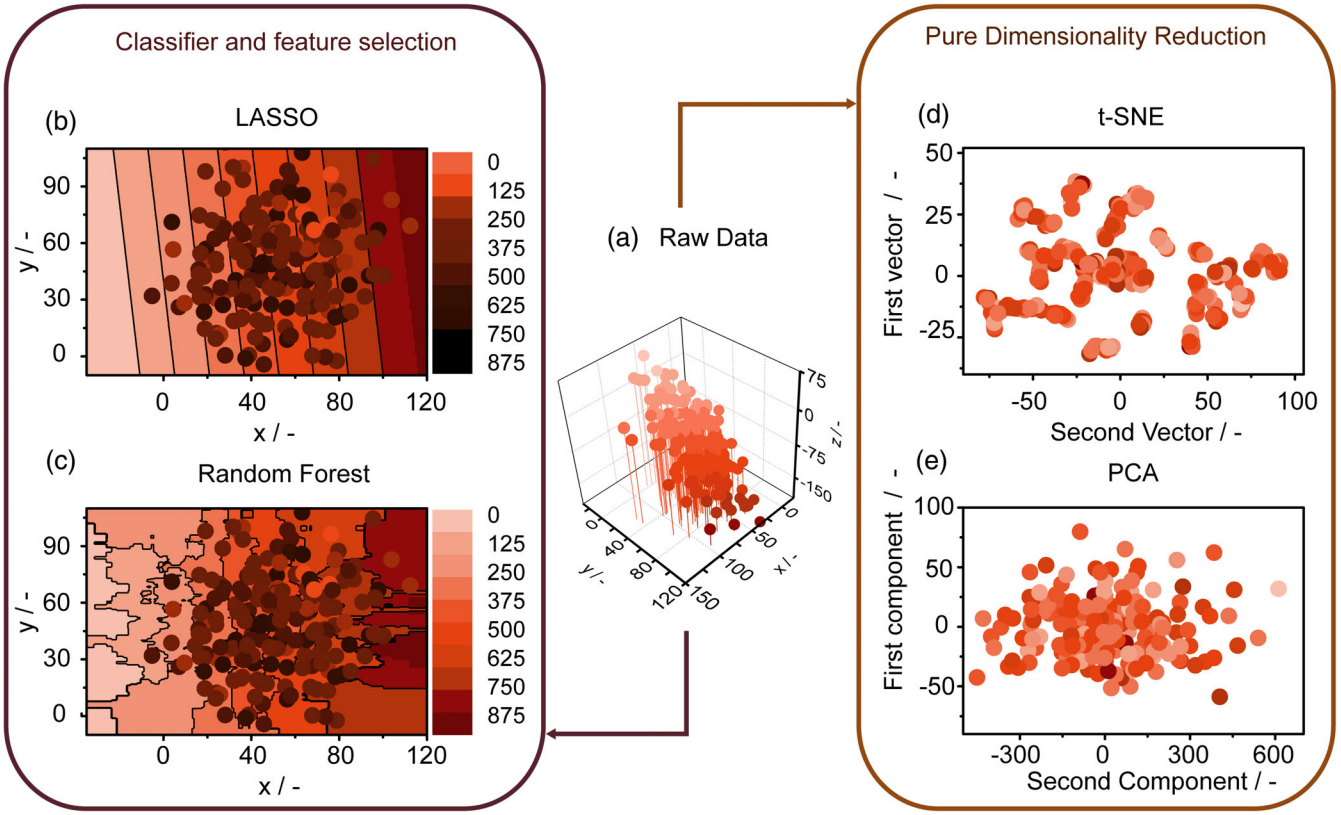

FIGURE 3 Schematic representation of dimensionality reduction techniques grouped as classifiers, feature selectors and pure dimensionality reductors: (a) 3 dimensional plot of $f(x,y,z)$, where x and y are two normal randomly distributed variables, and z is a linear combination of x and y; (b) LASSO (c) random forest classifier, (d) t-SNE, and (e) PCA applied on data illustrated in panel (a)

generalization to other datasets is more difficult and overfitting can appear. As example, in homogeneous catalysis ANN has been applied in predicting and analyzing 60×10^3 cross-coupling reactions[140] and the prediction of formation enthalpies of hydrocarbons,[141] while in heterogeneous catalysis, surface properties,[142] and CO_2 reduction[3] on bimetallic surfaces have been explored.

The RFC is based in generating a set of decision trees, extract a prediction for every tree and select the best solution by scoring all the solutions in the ensemble. The features are selected as a function of the weight that they have in the final decision. The robustness of the method depends on the number of trees used but in general RFC is highly accurate and robust and avoids overfitting (via forest averaging). Again, has the drawbacks of a black box model, limiting interpretability and the cost raises with the number of trees used. RFC has been applied olefin oligomerization using Cr as homogeneous catalyst[143] and in heterogeneous catalysis to screen C_2 transformation catalysts,[144] or the HER evolution on NiP_2 systems.[145] RFC has been also used together with multi-scale modeling to MK and KMC coupled to CFD for simulating ethylene oxidation on Ag[86] and in CO oxidation on $RuO_2(110)$,[87] respectively.

Other supervised nonlinear techniques for classification and regression are convolutional graph and graph embedding neural networks applied on organic molecules with biological interest,[146,147] and diffusion maps, applied on proteins.[148] Even if these last methods are not very popular in our field, they show a huge potential for large systems like big organic molecules synthesis in homogenous catalysis, or mapping all the possible interactions of complex surfaces, such as oxides in heterogenous catalysis.

6.2 | Dimensionality reduction

Common dimensionality reduction techniques are t-Stochastic Neighbor Embedding (t-SNE, Figure 3 panel d) and Principal Component Analysis (PCA, Figure 3 panel e). t-SNE transforms the distances between observations into conditional probabilities. Then, instead of comparing distances between the point x_i and its neighbor x_j, the method measures the

probability for x_j to be selected assuming a Gaussian Probability Density Distribution (PDD) centered in x_i. Nearly points have a high probability, while further observations have almost 0 probabilities to be selected. Later, the algorithm generates two analog observations y_i and y_j in a lower dimension space, and again calculates the conditional probability of the point y_j to be selected. The final output is a visually attractive a 2 or 3D map, in which the input data have been grouped in such way that dense clusters are expanded, and sparse set of points are condensed. t-SNE reduces dimensionality in systems in which other techniques fail at providing a very visual understanding of the data. However, 3D is the maximum dimension for the new reduced space and extrapolation to new datasets is limited. In homogeneous catalysis these techniques have allowed the construction of data-driven volcanoes,[114] while in the heterogeneous catalysis context, it has been employed as visualization technique in water oxidation,[149] or to derive new CO_2 electrocatalytic Cu-based bimetallic materials.[113]

The PCA reduces the dimensionality by projecting all data points into the directions that capture most of the variability, called principal components. This projection is done via diagonalization of the covariance matrix, thus ensuring that the corresponding vectors are orthogonal. The eigenvalues contain the explained variance of each corresponding eigenvector and are taken as criterion for selecting those that explains more. The PCA presents the following advantages: (i) the precision of dimensionality reduction can be controlled by including more principal components, (ii) the results and predictions are generalizable among similar data-sets, (iii) as the master equation is a bilinear model, it is can be tuned to be interpretable, thus describing additive phenomena such as thermochemistry, and (iv) the equation is highly modular and, provided that the physical interpretation is known, it can be extended to include other terms, such as solvation and coverage. The main drawback of PCA is that it cannot capture nonlinear correlations, although they are easier to spot in spaces with lower dimensionality. In homogeneous catalysis PCA has been also applied in asymmetric catalysis.[150] An iterative supervised variant, where PCA was recursively applied after fitting the obtained descriptors to the response variable, was applied for spotting selective ligands for the pyrrole synthesis.[151] The interpretable flavor of PCA has been successful applied in the decomposition of alcohols[39] and hydrodebromination reactions[47] on metals. There it was found that the largest source of variability almost coincides with the covalent part of the metal-adsorbate bond, commonly mapped to the d-band center on transition metals.[32,152] The second largest source of variability was associated with the relative redox character of the bonds, in line with the classical view of Pauling,[153] as well as the recent interpretation of differences on the coupling matrices.[154] Thus, PCA is a highly versatile method, able to reduce the dimensionality while preserving most of the system information. PCA paves the way for a universal methodology to extract robust and physically meaningful descriptors that can be related to experimental observables.[39,47] In addition, it is possible to overcome the linearity issues of PCA by using the kernel-PCA (kPCA) method, which uses different shaped kernels (as example, the corresponding kernel (k) for linear PCA is $k(\mathbf{x}, \mathbf{y}) = \mathbf{x}\cdot\mathbf{y}$). In kPCA, the covariance matrix is never explicitly diagonalized; the kernel function, which is defined as the dot product of the mapping on the new space of the nonlinear combinations (k is a matrix), plays the role of the covariance matrix instead. Then, the eigenvalues and the eigenvectors of the kernel are the principal components of the new space. The kPCA has been used to study the enzyme lactate hydrogenase.[155] All these dimensionality reduction techniques have a huge potential in shortening the estimations through Density Functional Theory based on data already available.

7 | CONCLUSION

Heterogeneous Catalysis has been based on the use of descriptors derived heuristically. However, as complexity increases simplifications based on simple arguments and chemical intuition fade. During the last years, several approaches to identify descriptors through statistical techniques have been put forward. In the present work we have revised the key implementation aspects regarding automation, structure generation and data extraction and storage, the first bottleneck when working with data-driven approaches. We have identified several areas that require further attention, particularly when increasing the multicomponent phase of materials that serve as catalysts, when larger molecules cannot be represented by small surrogates, and when external forces such as solvation or electric potentials are imposed to the system. Finally, we present pure dimensionality reduction techniques, like t-SNE or PCA in heterogeneous catalysis, that constitute promising and robust tools for descriptors identification thus paving the way to more advanced models that can account for activity and selectivity in an interpretable manner.

ACKNOWLEDGMENTS

This work was supported by the Spanish Ministries of Science and Innovation, and Universities (RTI2018-101394-B-I00). The authors thank BSC-RES for generously providing computational resources.

CONFLICT OF INTEREST

The author declares no conflict of interest.

AUTHOR CONTRIBUTIONS

Sergio Pablo-García: Conceptualization; data curation; formal analysis; investigation; methodology; software; supervision; validation; visualization; writing-original draft; writing-review & editing. **Rodrigo García-Muelas:** Conceptualization; data curation; formal analysis; investigation; methodology; project administration; software; supervision; validation; visualization; writing-original draft; writing-review & editing. **Albert Sabadell-Rendón:** Conceptualization; data curation; formal analysis; investigation; methodology; resources; software; supervision; validation; visualization; writing-original draft; writing-review & editing. **Núria López:** Conceptualization; funding acquisition; resources; writing-review & editing.

DATA AVAILABILITY STATEMENT

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

ORCID

Rodrigo García-Muelas https://orcid.org/0000-0002-2219-5027
Albert Sabadell-Rendón https://orcid.org/0000-0003-2905-1541
Núria López https://orcid.org/0000-0001-9150-5941

REFERENCES

1. Chorkendorff I, Niemantsverdriet JW. *Concepts of modern catalysis and kinetics*. Weinheim, Germany: John Wiley & Sons; 2017. https://doi.org/10.1002/3527602658.
2. Cui X, Tang C, Zhang Q. A review of Electrocatalytic reduction of dinitrogen to ammonia under ambient conditions. *Adv Energy Mater*. 2018;*8*(22):1800369. https://doi.org/10.1002/aenm.201800369.
3. Ulissi ZW, Medford AJ, Bligaard T, Nørskov JK. To address surface reaction network complexity using scaling relations machine learning and DFT calculations. *Nat Commun*. 2017;*8*(1):1–7. https://doi.org/10.1038/ncomms14621.
4. Sutton JE, Vlachos DG. A theoretical and computational analysis of linear free energy relations for the estimation of activation energies. *ACS Catal*. 2012;*2*(8):1624–34. https://doi.org/10.1021/cs3003269.
5. Sutton JE, Guo W, Katsoulakis MA, Vlachos DG. Effects of correlated parameters and uncertainty in electronic-structure-based chemical kinetic Modelling. *Nat Chem*. 2016;*8*(4):331–7. https://doi.org/10.1038/nchem.2454.
6. Bruix A, Margraf JT, Andersen M, Reuter K. First-principles-based multiscale Modelling of heterogeneous catalysis. *Nat Catal*. 2019;*2*(8):659–70. https://doi.org/10.1038/s41929-019-0298-3.
7. Greeley J, Jaramillo TF, Bonde J, Chorkendorff IB, Nørskov JK. Computational high-throughput screening of electrocatalytic materials for hydrogen evolution. *Nat Mater*. 2006;*5*(11):909–13. https://doi.org/10.1038/nmat1752.
8. Saal JE, Kirklin S, Aykol M, Meredig B, Wolverton C. Materials design and discovery with high-throughput Density Functional Theory: the open quantum materials database (OQMD). *JOM*. 2013;*65*(11):1501–9. https://doi.org/10.1007/s11837-013-0755-4.
9. Singh AK, Mathew K, Zhuang HL, Hennig RG. Computational screening of 2D materials for photocatalysis. *J Phys Chem Lett*. 2015;*6*:1087–98. https://doi.org/10.1021/jz502646d.
10. Ahmadi M, Mistry H, Roldan Cuenya B. Tailoring the catalytic properties of metal nanoparticles via support interactions. *J Phys Chem Lett*. 2016;*7*(17):3519–33. https://doi.org/10.1021/acs.jpclett.6b01198.
11. Almora-Barrios N, Novell-Leruth G, Whiting P, Liz-Marzan LM, López N. Theoretical description of the role of halides, silver, and surfactants on the structure of gold nanorods. *Nano Lett*. 2014;*14*(2):871–5. https://doi.org/10.1021/nl404661u.
12. García-Muelas R, Dattila F, Shinagawa T, Martín AJ, Pérez-Ramírez J, López N. Origin of the selective electroreduction of carbon dioxide to formate by chalcogen modified copper. *J Phys Chem Lett*. 2018;9(24):7153–9. https://doi.org/10.1021/acs.jpclett.8b03212.
13. Mitchell S, Michels N-L, Pérez-Ramírez J. From powder to technical body: the undervalued science of catalyst scale up. *Chem Soc Rev*. 2013;*42*(14):6094–112. https://doi.org/10.1039/c3cs60076a.
14. Hargreaves JSJ, Munnoch AL. A survey of the influence of binders in zeolite catalysis. *Cat Sci Technol*. 2013;*3*(5):1165. https://doi.org/10.1039/c3cy20866d.

15. Sholl D, Steckel JA. *Density Functional Theory: a practical introduction*. Hoboken, United States of America: John Wiley & Sons; 2011.
16. Ardagh MA, Birol T, Zhang Q, Abdelrahman OA, Dauenhauer PJ. Catalytic resonance theory: SuperVolcanoes, catalytic molecular pumps, and oscillatory steady state. *Cat Sci Technol*. 2019;*9*(18):5058–76. https://doi.org/10.1039/c9cy01543d.
17. Vojvodic A, Nørskov JK. New design paradigm for heterogeneous catalysts. *Natl Sci Rev USA*. 2015;*2*(2):140–3. https://doi.org/10.1093/nsr/nwv023.
18. Pérez-Ramírez J, López N. Strategies to break linear scaling relationships. *Nature Catalysis*. 2019;*2*:971–6. https://doi.org/10.1038/s41929-019-0376-6.
19. Butler KT, Davies DW, Cartwright H, Isayev O, Walsh A. Machine learning for molecular and materials science. *Nature*. 2018;*559*(7715):547–55. https://doi.org/10.1038/s41586-018-0337-2.
20. Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, et al. Comment: the FAIR guiding principles for scientific data management and stewardship. *Sci Data*. 2016;*3*(1):160018. https://doi.org/10.1038/sdata.2016.18.
21. Hastie T, Tibshirani R, Friedman J. The elements of statistical learning. *Springer series in statistics*. Volume 99. 2nd ed. New York, NY: Springer New York; 2009. https://doi.org/10.1007/978-0-387-84858-7.
22. Lejaeghere K, Bihlmayer G, Bjorkman T, Blaha P, Blugel S, Blum V, et al. Reproducibility in Density Functional Theory calculations of solids. *Science*. 2016;*351*(6280):aad3000–0. https://doi.org/10.1126/science.aad3000.
23. Álvarez-Moreno M, De Graaf C, López N, Maseras F, Poblet JM, Bo C. Managing the computational chemistry big data problem: the IoChem-BD platform. *J Chem Inf Model*. 2015;*55*(1):95–103. https://doi.org/10.1021/ci500593j.
24. NoMaD Repository [Internet]. Https://Nomad-Coe.Eu. Accessed 16 Dec 2020.
25. Materials Cloud [Internet]. Https://Www.Materialscloud.Org/Home. Accessed 16 Dec 2020.
26. CatApp Database—Computational materials repository [Internet]. Https://Cmr.Fysik.Dtu.Dk/. Accessed 16 Dec 2020.
27. Tabor DP, Roch LM, Saikin SK, Kreisbeck C, Sheberla D, Montoya JH, et al. Accelerating the discovery of materials for clean energy in the era of smart automation. *Nat Rev Mater*. 2018;*3*(5):5–20. https://doi.org/10.1038/s41578-018-0005-z.
28. Callaway E. It will change everything: DeepMind's AI makes gigantic leap in solving protein structures. *Nature*. 2020;*588*:203–4. https://doi.org/10.1038/d41586-020-03348-4.
29. Sabatier P. La Catalyse En Chimie Organique; Encyclopédie de science chimique appliquée. *Ch Béranger*. 1913. https://doi.org/10.14375/NP.9782369430186.
30. Balandin AA. Modern state of the Multiplet theory of heterogeneous catalysis. *Adv Catal*. 1969;*19*:1–210. https://doi.org/10.1016/S0360-0564(08)60029-2.
31. Benson SW. Thermochemical kinetics: *methods for the estimation of thermochemical data and rate parameters*. Wiley: New York, United States of America 1976.
32. Abild-Pedersen F, Greeley J, Studt F, Rossmeisl J, Munter TR, Moses PG, et al. Scaling properties of adsorption energies for hydrogen-containing molecules on transition-metal surfaces. *Phys Rev Lett*. 2007;*99*(1):016105. https://doi.org/10.1103/PhysRevLett.99.016105.
33. Salciccioli M, Edie SM, Vlachos DG. Adsorption of acid, ester, and ether functional groups on Pt: fast prediction of thermochemical properties of adsorbed oxygenates via DFT-based group additivity methods. *J Phys Chem C*. 2012;*116*(2):1873–86. https://doi.org/10.1021/jp2091413.
34. Montemore MM, Medlin JW. A unified picture of adsorption on transition metals through different atoms. *J Am Chem Soc*. 2014;*136*(26):9272–5. https://doi.org/10.1021/ja504193w.
35. Skúlason E, Bligaard T, Gudmundsdóttir S, Studt F, Rossmeisl J, Abild-Pedersen F, et al. A theoretical evaluation of possible transition metal electro-catalysts for N2 reduction. *Phys Chem Chem Phys*. 2012;*14*(3):1235–45. https://doi.org/10.1039/c1cp22271f.
36. Medford AJ, Vojvodic A, Hummelshøj JS, Voss J, Abild-Pedersen F, Studt F, et al. From the Sabatier principle to a predictive theory of transition-metal heterogeneous catalysis. *J Catal*. 2015;*328*:36–42. https://doi.org/10.1016/j.jcat.2014.12.033.
37. Montemore MM, Medlin JW. Site-specific scaling relations for hydrocarbon adsorption on hexagonal transition metal surfaces. *J Phys Chem C*. 2013;*117*(39):20078–88. https://doi.org/10.1021/jp4076405.
38. Bagger A, Ju W, Varela AS, Strasser P, Rossmeisl J. Electrochemical CO_2 Reduction: a classification problem. *Chem Phys Chem*. 2017;*18*(22):3266–73. http://dx.doi.org/10.1002/cphc.201700736.
39. 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*(1):4687. https://doi.org/10.1038/s41467-019-12709-1.
40. Calle-Vallejo F, Tymoczko J, Colic V, Vu QH, Pohl MD, Morgenstern K, et al. Finding optimal surface sites on heterogeneous catalysts by counting nearest neighbors. *Science*. 2015;*350*(6257):185–9. https://doi.org/10.1126/science.aab3501.
41. Batchelor TAA, Pedersen JK, Winther SH, Castelli IE, Jacobsen KW, Rossmeisl J. High-entropy alloys as a discovery platform for electrocatalysis. *Joule*. 2019;*3*(3):834–45. https://doi.org/10.1016/j.joule.2018.12.015.
42. Busch M, Fabrizio A, Luber S, Hutter J, Corminboeuf C. Exploring the limitation of molecular water oxidation catalysts. *J Phys Chem C*. 2018;*122*(23):12404–12. https://doi.org/10.1021/acs.jpcc.8b03935.
43. Fernández EM, Moses PG, Toftelund A, Hansen HA, Martínez JI, Abild-Pedersen F, et al. Scaling relationships for adsorption energies on transition metal oxide, sulfide, and nitride surfaces. *Angew Chem Int Ed*. 2008;*47*(25):4683–6. https://doi.org/10.1002/anie.200705739.
44. Moser M, Czekaj I, López N, Pérez-Ramírez J. The virtue of defects: stable bromine production by catalytic oxidation of hydrogen bromide on titanium oxide. *Angew Chem Int Ed*. 2014;*126*(33):8772–7. https://doi.org/10.1002/anie.201483371.

45. Divanis S, Kutlusoy T, Boye IMI, Man IC, Rossmeisl J. Oxygen evolution reaction: a perspective on a decade of atomic scale simulations. *Chem Sci*. 2020;*11*:2943–50. https://doi.org/10.1039/C9SC05897D.
46. Falsig H, Hvolbæk B, Kristensen IS, Jiang T, Bligaard T, Christensen CH, et al. Trends in the catalytic CO oxidation activity of nanoparticles. *Angew Chem Int Ed*. 2008;*47*:4835–9. https://doi.org/10.1002/anie.200801479.
47. Saadun AJ, Pablo-Garcia S, Paunovic V, Li Q, Sabadell-Rendón A, Kleemann K, et al. Performance of metal-catalyzed hydrodebromination of dibromomethane analyzed by descriptors derived from statistical learning. *ACS Catal*. 2020;*10*:6129–43. https://doi.org/10.1021/acscatal.0c00679.
48. Vojvodic A, Calle-Vallejo F, Guo W, Wang S, Toftelund A, Studt F, et al. On the behavior of Brønsted–Evans–Polanyi relations for transition metal oxides. *J Chem Phys*. 2011;*134*(24):244509. https://doi.org/10.1063/1.3602323.
49. Bajdich M, García-Mota M, Vojvodic A, Nørskov JK, Bell AT. Theoretical investigation of the activity of cobalt oxides for the electrochemical oxidation of water. *J Am Chem Soc*. 2013;*135*(36):13521–30. https://doi.org/10.1021/ja405997s.
50. Bell RP. The theory of reactions involving proton transfers. *Proc R Soc Lond A Math Phys Eng Sci*. 1936;*154*(882):414–29. https://doi.org/10.1098/rspa.1936.0060.
51. Evans MG, Polanyi M. Inertia and driving force of chemical reactions. *Trans Faraday Soc*. 1938;*34*:11–24. https://doi.org/10.1039/tf9383400011.
52. Nørskov JK, Bligaard T, Logadottir A, Bahn S, Hansen LB, Bollinger M, et al. Universality in heterogeneous catalysis. *J Catal*. 2002;*209*(2):275–8. https://doi.org/10.1006/jcat.2002.3615.
53. Bligaard T, Nørskov JK, Dahl S, Matthiesen J, Christensen CH, Sehested J. The Brønsted–Evans–Polanyi relation and the volcano curve in heterogeneous catalysis. *J Catal*. 2004;*224*(1):206–17. https://doi.org/10.1016/j.jcat.2004.02.034.
54. Zaffran J, Michel C, Delbecq F, Sautet P. Trade-off between accuracy and universality in linear energy relations for alcohol dehydrogenation on transition metals. *J Phys Chem C*. 2015;*119*(23):12988–98. https://doi.org/10.1021/acs.jpcc.5b01703.
55. Loffreda D, Delbecq F, Vigné F, Sautet P. Fast prediction of selectivity in heterogeneous catalysis from extended Brønsted–Evans–Polanyi relations: a theoretical insight. *Angew Chem Int Ed*. 2009;*48*(47):8978–80. https://doi.org/10.1002/anie.200902800.
56. García-Muelas R, Li Q, López N. Density Functional Theory comparison of methanol decomposition and reverse reactions on metal surfaces. *ACS Catal*. 2015;*5*(2):1027–36. https://doi.org/10.1021/cs501698w.
57. Li Q, García-Muelas R, López N. Microkinetics of alcohol reforming for H2 production from a FAIR Density Functional Theory database. *Nat Commun*. 2018;*9*(1):526. https://doi.org/10.1038/s41467-018-02884-y.
58. Shustorovich E. The bond-order conservation approach to chemisorption and heterogeneous catalysis: applications and implications. *Adv Catal*. 1990;*37*:101–63. https://doi.org/10.1016/S0360-0564(08)60364-8.
59. Shustorovich E. The UBI-QEP method: a practical theoretical approach to understanding chemistry on transition metal surfaces. *Surf Sci Rep*. 1998;*31*(1–3):1–119. https://doi.org/10.1016/S0167-5729(97)00016-2.
60. López N, Almora-Barrios N, Carchini G, Błoński P, Bellarosa L, García-Muelas R, et al. State-of-the-art and challenges in theoretical simulations of heterogeneous catalysis at the microscopic level. *Cat Sci Technol*. 2012;*2*(12):2405. https://doi.org/10.1039/c2cy20384g.
61. Vlachos DG. Multiscale modeling for emergent behavior, complexity, and combinatorial explosion. *AIChE J*. 2012;*58*(5):1314–25. https://doi.org/10.1002/aic.13803.
62. Matera S, Schneider WF, Heyden A, Savara A. Progress in accurate chemical kinetic modeling, simulations, and parameter estimation for heterogeneous catalysis. *ACS Catal*. 2019;*9*(8):6624–47. https://doi.org/10.1021/acscatal.9b01234.
63. Motagamwala AH, Dumesic JA. Microkinetic modeling: a tool for rational catalyst design. *Chem Rev*. 2021;*121*(2):1049–76. https://doi.org/10.1021/acs.chemrev.0c00394.
64. Pérez-Soto R, Besora M, Maseras F. The challenge of reproducing with calculations raw experimental kinetic data for an organic reaction. *Org Lett*. 2020;*22*(8):2873–7. https://doi.org/10.1021/acs.orglett.0c00367.
65. Brezny AC, Landis CR. Development of a comprehensive microkinetic model for Rh(Bis[Diazaphospholane])-catalyzed hydroformylation. *ACS Catal*. 2019;*9*(3):2501–13. https://doi.org/10.1021/acscatal.9b00173.
66. Rebarchik M, Bhandari S, Kropp T, Mavrikakis M. How noninnocent spectator species improve the oxygen reduction activity of single-atom catalysts: microkinetic models from first-principles calculations. *ACS Catal*. 2020;*10*(16):9129–35. https://doi.org/10.1021/acscatal.0c01642.
67. Linic S, Barteau MA. Construction of a reaction coordinate and a microkinetic model for ethylene epoxidation on silver from DFT calculations and surface science experiments. *J Catal*. 2003;*214*(2):200–12. https://doi.org/10.1016/S0021-9517(02)00156-2.
68. 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*(9):1893–904. https://doi.org/10.1021/acs.accounts.0c00340.
69. Chatterjee A, Vlachos DG. An overview of spatial microscopic and accelerated kinetic Monte Carlo methods. *J Comput Mater Des*. 2007;*14*(2):253–308. https://doi.org/10.1007/s10820-006-9042-9.
70. Stamatakis M, Vlachos DG. Unraveling the complexity of catalytic reactions via kinetic Monte Carlo simulation: current status and Frontiers. *ACS Catal*. 2012;*7*:2648–63. https://doi.org/10.1021/cs3005709.
71. Andersen M, Panosetti C, Reuter K. A practical guide to surface kinetic Monte Carlo simulations. *Front Chem*. 2019;*7*:202. https://doi.org/10.3389/fchem.2019.00202.
72. Slepoy A, Thompson AP, Plimpton SJ. A constant-time kinetic Monte Carlo algorithm for simulation of large biochemical reaction networks. *J Chem Phys*. 2008;*128*(20):205101. https://doi.org/10.1063/1.2919546.

73. Nielsen J, D'Avezac M, Hetherington J, Stamatakis M. Parallel kinetic Monte Carlo simulation framework incorporating accurate models of adsorbate lateral interactions. *J Chem Phys*. 2013;*139*(22):224706. https://doi.org/10.1063/1.4840395.
74. Hoffmann MJ, Matera S, Reuter K. Kmos: a lattice kinetic Monte Carlo framework. *Comput Phys Commun*. 2014;*185*(7):2138–50. https://doi.org/10.1016/j.cpc.2014.04.003.
75. Chutia A, Thetford A, Stamatakis M, Catlow CRA. 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*(6):3620–32. https://doi.org/10.1039/c9cp05476f.
76. Mahlberg D, Groß A. Vacancy assisted diffusion on single-atom surface alloys. *ChemPhysChem*. 2021;*22*(1):29–39. https://doi.org/10.1002/cphc.202000838.
77. Reuter K, Scheffler M. First-principles kinetic Monte Carlo simulations for heterogeneous catalysis: application to the CO oxidation at Ru O2 (110). *Phys Rev B Condens Matter Mater Phys*. 2006;*73*(4):045433. https://doi.org/10.1103/PhysRevB.73.045433.
78. Pogodin S, López N. A more accurate kinetic Monte Carlo approach to a monodimensional surface reaction: the interaction of oxygen with the RuO2(110) surface. *ACS Catal*. 2014;*4*(7):2328–32. https://doi.org/10.1021/cs500414p.
79. Vorobyeva E, Gerken VC, Mitchell S, Sabadell-Rendón A, Hauert R, Xi S, et al. Activation of copper species on carbon nitride for enhanced activity in the arylation of amines. *ACS Catal*. 2020;*10*(19):11069–80. https://doi.org/10.1021/acscatal.0c03164.
80. Vandewalle LA, Marin GB, Van Geem KM. CatchyFOAM: Euler–Euler CFD simulations of fluidized bed reactors with microkinetic modeling of gas-phase and catalytic surface chemistry. *Energy Fuel*. 2020;*35*:2545–61. https://doi.org/10.1021/acs.energyfuels.0c02824.
81. Matera S, Maestri M, Cuoci A, Reuter K. Predictive-quality surface reaction chemistry in real reactor models: integrating first-principles kinetic Monte Carlo simulations into computational fluid dynamics. *ACS Catal*. 2014;*4*(11):4081–92. https://doi.org/10.1021/cs501154e.
82. Maestri M, Cuoci A. Coupling CFD with detailed microkinetic modeling in heterogeneous catalysis. *Chem Eng Sci*. 2013;*96*:106–17. https://doi.org/10.1016/j.ces.2013.03.048.
83. Maffei T, Gentile G, Rebughini S, Bracconi M, Manelli F, Lipp S, et al. A multiregion operator-splitting CFD approach for coupling microkinetic modeling with internal porous transport in heterogeneous catalytic reactors. *Chem Eng J*. 2016;*283*:1392–404. https://doi.org/10.1016/j.cej.2015.08.080.
84. Cuoci A, Frassoldati A, Faravelli T, Ranzi E. A computational tool for the detailed kinetic modeling of laminar flames: application to C2H4/CH4 Coflow flames. *Combust Flame*. 2013;*160*(5):870–86. https://doi.org/10.1016/j.combustflame.2013.01.011.
85. Donazzi A, Maestri M, Michael BC, Beretta A, Forzatti P, Groppi G, et al. Microkinetic modeling of spatially resolved autothermal CH4 catalytic partial oxidation experiments over Rh-coated foams. *J Catal*. 2010;*275*(2):270–9. https://doi.org/10.1016/j.jcat.2010.08.007.
86. Partopour B, Paffenroth RC, Dixon AG. Random forests for mapping and analysis of microkinetics models. *Comput Chem Eng*. 2018;*115*:286–94. https://doi.org/10.1016/j.compchemeng.2018.04.019.
87. Bracconi M, Maestri M. Training set design for machine learning techniques applied to the approximation of computationally intensive first-principles kinetic models. *Chem Eng J*. 2020;*400*:125469. https://doi.org/10.1016/j.cej.2020.125469.
88. Salciccioli M, Chen Y, Vlachos DG. 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. *J Phys Chem C*. 2010;*114*(47):20155–66. https://doi.org/10.1021/jp107836a.
89. Weininger D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J Chem Inf Comput Sci*. 1988;*28*(1):31–6. https://doi.org/10.1021/ci00057a005.
90. Wen M, Blau SM, Spotte-Smith EWC, Dwaraknath S, Persson KA. BonDNet: a graph neural network for the prediction of bond dissociation energies for charged molecules. *Chem Sci*. 2021;*12*(5):1858–68. https://doi.org/10.1039/D0SC05251E.
91. Pfaendtner J, Broadbelt LJ. Mechanistic modeling of lubricant degradation. 2. The autoxidation of decane and octane. *Ind Eng Chem Res*. 2008;*47*(9):2897–904. https://doi.org/10.1021/ie071481z.
92. Rangarajan S, Bhan A, Daoutidis P. Language-oriented rule-based reaction network generation and analysis: description of RING. *Comput Chem Eng*. 2012;*45*:114–23. https://doi.org/10.1016/j.compchemeng.2012.06.008.
93. Goldsmith CF, West RH. Automatic generation of microkinetic mechanisms for heterogeneous catalysis. *J Phys Chem C*. 2017;*121*(18):9970–81. https://doi.org/10.1021/acs.jpcc.7b02133.
94. Kim Y, Kim JW, Kim Z, Kim WY. Efficient prediction of reaction paths through molecular graph and reaction network analysis. *Chem Sci*. 2018;*9*(4):825–35. https://doi.org/10.1039/C7SC03628K.
95. Vernuccio S, Broadbelt LJ. Discerning complex reaction networks using automated generators. *AIChE J*. 2019;*65*(8):e16663. https://doi.org/10.1002/aic.16663.
96. Ellson J, Gansner ER, Koutsofios E, North SC, Woodhull G. Graphviz and dynagraph—static and dynamic graph drawing tools. *Graph drawing software*. Berlin: Springer-Verlag; 2003. p. 127–48.
97. Hagberg A, Swart P, Chult S. *Exploring network structure, dynamics, and function using networkX*. Los Alamos, NM: Los Alamos National Lab. (LANL); 2008.
98. Larsen HA, Mortensen JJ, Blomqvist J, Castelli IE, Christensen R, Dułak M, et al. The atomic simulation environment—a python library for working with atoms. *J Phys Condens Matter*. 2017;*29*(27):273002. https://doi.org/10.1088/1361-648X/aa680e.
99. Open Babel: Te Open Source Chemistry Toolbox. Http://Openbabel.Org/Wiki/Main_Page. Accessed 16 Dec 2020.
100. Montoya JH, Persson KA. A high-throughput framework for determining adsorption energies on solid surfaces. *NPJ Comput Mater*. 2017;*3*(1):1–4. https://doi.org/10.1038/s41524-017-0017-z.
101. Tran K, Palizhati A, Back S, Ulissi ZW. Dynamic workflows for routine materials discovery in surface science. *J Chem Inf Model*. 2018;*58*(12):2392–400. https://doi.org/10.1021/acs.jcim.8b00386.

102. Kahle L, Marcolongo A, Marzari N. High-throughput computational screening for solid-state Li-ion conductors. *Energ Environ Sci*. 2020;*13*(3):928–48. https://doi.org/10.1039/C9EE02457C.
103. Pablo-García S, Álvarez-Moreno M, López N. Turning chemistry into information for heterogeneous catalysis. *Int J Quantum Chem*. 2021;*121*(1):e26382. https://doi.org/10.1002/qua.26382.
104. Jain A, Ong SP, Chen W, Medasani B, Qu X, Kocher M, et al. FireWorks: a dynamic workflow system designed for high-throughput applications. *Concurr Comput Pract Exp*. 2015;*27*(17):5037–59. https://doi.org/10.1002/cpe.3505.
105. Pizzi G, Cepellotti A, Sabatini R, Marzari N, Kozinsky B. AiiDA: automated interactive infrastructure and database for computational science. *Comput Mater Sci*. 2016;*111*:218–30. https://doi.org/10.1016/j.commatsci.2015.09.013.
106. AiiDA. Http://Www.Aiida.Net. Accessed 16 Dec 2020.
107. Bo C, Maseras F, López N. The role of computational results databases in accelerating the discovery of catalysts. *Nat Catal*. 2018;*1*(11):809–10. https://doi.org/10.1038/s41929-018-0176-4.
108. Jain A, Ong SP, Hautier G, Chen W, Richards WD, Dacek S, et al. Commentary: the materials project: a materials genome approach to accelerating materials innovation. *APL Mater*. 2013;*1*(1):011002. https://doi.org/10.1063/1.4812323.
109. Catalysis Hub [Internet]. Https://Www.Catalysis-Hub.Org. Accessed 8 Mar 2021.
110. Wołos A, Roszak R, Żądło-Dobrowolska A, Beker W, Mikulak-Klucznik B, Spólnik G, et al. Synthetic connectivity, emergence, and self-regeneration in the network of prebiotic chemistry. *Science*. 2020;*369*(6511):eaaw1955. https://doi.org/10.1126/science.aaw1955.
111. Eslamibidgoli MJ, Eikerling MH. Approaching the self-consistency challenge of electrocatalysis with theory and computation. *Curr Opin Electrochem*. 2018;*9*:189–97. https://doi.org/10.1016/j.coelec.2018.03.038.
112. Vandichel M, Busch M, Laasonen K. Oxygen evolution on metal-oxy-hydroxides: beneficial role of mixing Fe, Co, Ni explained via bifunctional edge/acceptor route. *ChemCatChem*. 2020;*12*(5):1436–42. https://doi.org/10.1002/cctc.201901951.
113. Zhong M, Tran K, Min Y, Wang C, Wang Z, Dinh C-T, et al. Accelerated discovery of CO_2 electrocatalysts using active machine learning. *Nature*. 2020;*581*(7807):178–83. https://doi.org/10.1038/s41586-020-2242-8.
114. Wodrich MD, Fabrizio A, Meyer B, Corminboeuf C. Data-powered augmented volcano plots for homogeneous catalysis. *Chem Sci*. 2020;*11*(44):12070–80. https://doi.org/10.1039/D0SC04289G.
115. Dattila F, García-Muelas R, López N. Active and selective ensembles in oxide-derived copper catalysts for CO_2 reduction. *ACS Energy Lett*. 2020;5(10):3176–84. https://doi.org/10.1021/acsenergylett.0c01777.
116. Frei MS, Mondelli C, García-Muelas R, Kley KS, Puértolas B, López N, et al. Atomic-scale engineering of indium oxide promotion by palladium for methanol production via CO_2 hydrogenation. *Nat Commun*. 2019;*10*(1):3377. https://doi.org/10.1038/s41467-019-11349-9.
117. Frei MS, Mondelli C, García-Muelas R, Morales-Vidal J, Philipp M, Safonova OV, et al. Nanostructure of nickel-promoted indium oxide catalysts drives selectivity in CO_2 hydrogenation. *Nat Commun*. 2021;*12*:1960.
118. Kauppinen MM, Korpelin V, Verma AM, Melander MM, Honkala K. Escaping scaling relationships for water dissociation at interfacial sites of zirconia-supported Rh and Pt clusters. *J Chem Phys*. 2019;*151*(16):164302. https://doi.org/10.1063/1.5126261.
119. Grasselli RK. Fundamental principles of selective heterogeneous oxidation catalysis. *Top Catal*. 2002;*21*(1–3):79–88. https://doi.org/10.1023/A:1020556131984.
120. Capdevila-Cortada M, Vilé G, Teschner D, Pérez-Ramírez J, López N. Reactivity descriptors for ceria in catalysis. *Appl Catal Environ*. 2016;*197*:299–312. https://doi.org/10.1016/j.apcatb.2016.02.035.
121. Wodrich MD, Corminboeuf C, von Schleyer P. Systematic errors in computed alkane energies using B3LYP and other popular DFT functionals. *Org Lett*. 2006;*8*(17):3631–4. https://doi.org/10.1021/ol061016i.
122. García-Muelas R, López N. Collective descriptors for the adsorption of sugar alcohols on Pt and Pd(111). *J Phys Chem C*. 2014;*118*(31):17531–7. https://doi.org/10.1021/jp502819s.
123. Li Q, López N. Chirality, rigidity, and conjugation: a first-principles study of the key molecular aspects of lignin depolymerization on Ni-based catalysts. *ACS Catal*. 2018;*8*(5):4230–40. https://doi.org/10.1021/acscatal.8b00067.
124. Gonthier JF, Steinmann SN, Wodrich MD, Corminboeuf C. Quantification of "fuzzy" chemical concepts: a computational perspective. *Chem Soc Rev*. 2012;*41*(13):4671–87. https://doi.org/10.1039/c2cs35037h.
125. Gu J, Wu W, Danovich D, Hoffmann R, Tsuji Y, Shaik S. Valence bond theory reveals hidden delocalized diradical character of polyenes. *J Am Chem Soc*. 2017;*139*(27):9302–16. https://doi.org/10.1021/jacs.7b04410.
126. Han X, Xia Q, Huang J, Liu Y, Tan C, Cui Y. Chiral covalent organic frameworks with high chemical stability for heterogeneous asymmetric catalysis. *J Am Chem Soc*. 2017;*139*(25):8693–7. https://doi.org/10.1021/jacs.7b04008.
127. Szőllősi G. Asymmetric one-pot reactions using heterogeneous chemical catalysis: recent steps towards sustainable processes. *Cat Sci Technol*. 2018;*8*(2):389–422. https://doi.org/10.1039/C7CY01671A.
128. Ting LRL, García-Muelas R, Martín AJ, Veenstra FLP, Chen ST, Peng Y, et al. Electrochemical reduction of carbon dioxide to 1-butanol on oxide-derived copper. *Angew Chem Int Ed*. 2020;*59*(47):21072–9. https://doi.org/10.1002/anie.202008289.
129. Garcia-Ratés M, López N. Multigrid-based methodology for implicit solvation models in periodic DFT. *J Chem Theory Comput*. 2016;*12*(3):1331–41. https://doi.org/10.1021/acs.jctc.5b00949.
130. Garcia-Ratés M, García-Muelas R, López N. Solvation effects on methanol decomposition on Pd (111), Pt (111), and Ru (0001). *J Phys Chem C*. 2017;*121*(25):13803–9. https://doi.org/10.1021/acs.jpcc.7b05545.
131. Tuokko S, Pihko PM, Honkala K. First principles calculations for hydrogenation of acrolein on Pd and Pt: Chemoselectivity depends on steric effects on the surface. *Angew Chem Int Ed*. 2016;*55*(5):1670–4. https://doi.org/10.1002/anie.201507631.

132. Banerjee S, Sreenithya A, Sunoj RB. Machine learning for predicting product distributions in catalytic regioselective reactions. *Phys Chem Chem Phys*. 2018;*20*(27):18311–8. https://doi.org/10.1039/C8CP03141J.
133. Janet JP, Kulik HJ. Predicting electronic structure properties of transition metal complexes with neural networks. *Chem Sci*. 2017;*8*(7): 5137–52. https://doi.org/10.1039/C7SC01247K.
134. Meyer B, Sawatlon B, Heinen S, von Lilienfeld OA, Corminboeuf C. Machine learning meets volcano plots: computational discovery of cross-coupling catalysts. *Chem Sci*. 2018;*9*(35):7069–77. https://doi.org/10.1039/C8SC01949E.
135. Suzuki K, Toyao T, Maeno Z, Takakusagi S, Shimizu K, Takigawa I. Statistical analysis and discovery of heterogeneous catalysts based on machine learning from diverse published data. *ChemCatChem*. 2019;*11*(18):4537–47. https://doi.org/10.1002/cctc.201900971.
136. O'Connor NJ, Jonayat ASM, Janik MJ, Senftle TP. Interaction trends between single metal atoms and oxide supports identified with Density Functional Theory and statistical learning. *Nat Catal*. 2018;*1*(7):531–9. https://doi.org/10.1038/s41929-018-0094-5.
137. Su Y-Q, Zhang L, Wang Y, Liu J-X, Muravev V, Alexopoulos K, et al. Stability of heterogeneous single-atom catalysts: a scaling law mapping thermodynamics to kinetics. *NPJ Comput Mater*. 2020;*6*(1):144. https://doi.org/10.1038/s41524-020-00411-6.
138. Yamaguchi S, Nishimura T, Hibe Y, Nagai M, Sato H, Johnston I. Regularized regression analysis of digitized molecular structures in organic reactions for quantification of steric effects. *J Comput Chem*. 2017;*38*(21):1825–33. https://doi.org/10.1002/jcc.24791.
139. Li Z, Wang S, Chin WS, Achenie LE, Xin H. High-throughput screening of bimetallic catalysts enabled by machine learning. *J Mater Chem A*. 2017;*5*(46):24131–8. https://doi.org/10.1039/C7TA01812F.
140. Burello E, Farrusseng D, Rothenberg G. Combinatorial explosion in homogeneous catalysis: screening 60,000 cross-coupling reactions. *Adv Synth Catal*. 2004;*346*(13–15):1844–53. https://doi.org/10.1002/adsc.200404170.
141. Wodrich MD, Corminboeuf C. Reaction enthalpies using the neural-network-based X1 approach: the important choice of input descriptors. *J Phys Chem A*. 2009;*113*(13):3285–90. https://doi.org/10.1021/jp9002005.
142. Palizhati A, Zhong W, Tran K, Back S, Ulissi ZW. Toward predicting Intermetallics surface properties with high-throughput DFT and convolutional neural networks. *J Chem Inf Model*. 2019;*59*(11):4742–9. https://doi.org/10.1021/acs.jcim.9b00550.
143. Maley SM, Kwon D-H, Rollins N, Stanley JC, Sydora OL, Bischof SM, et al. Quantum-mechanical transition-state model combined with machine learning provides catalyst design features for selective Cr olefin oligomerization. *Chem Sci*. 2020;*11*(35):9665–74. https://doi.org/10.1039/D0SC03552A.
144. Takahashi K, Miyazato I, Nishimura S, Ohyama J. Unveiling hidden catalysts for the oxidative coupling of methane based on combining machine learning with literature data. *ChemCatChem*. 2018;*10*(15):3223–8. https://doi.org/10.1002/cctc.201800310.
145. Wexler RB, Martirez JMP, Rappe AM. Chemical pressure-driven enhancement of the hydrogen evolving activity of Ni_2P from nonmetal surface doping interpreted via machine learning. *J Am Chem Soc*. 2018;*140*(13):4678–83. https://doi.org/10.1021/jacs.8b00947.
146. Wang X, Li Z, Jiang M, Wang S, Zhang S, Wei Z. Molecule property prediction based on spatial graph embedding. *J Chem Inf Model*. 2019;*59*(9):3817–28. https://doi.org/10.1021/acs.jcim.9b00410.
147. Harada S, Akita H, Tsubaki M, Baba Y, Takigawa I, Yamanishi Y, et al. Dual graph convolutional neural network for predicting chemical networks. *BMC Bioinform*. 2020;*21*(S3):94. https://doi.org/10.1186/s12859-020-3378-0.
148. Boninsegna L, Gobbo G, Noé F, Clementi C. Investigating molecular kinetics by variationally optimized diffusion maps. *J Chem Theory Comput*. 2015;*11*(12):5947–60. https://doi.org/10.1021/acs.jctc.5b00749.
149. Palkovits R, Palkovits S. Using artificial intelligence to forecast water oxidation catalysts. *ACS Catal*. 2019;*9*(9):8383–7. https://doi.org/10.1021/acscatal.9b01985.
150. Amar Y, Schweidtmann AM, Deutsch P, Cao L, Lapkin A. Machine learning and molecular descriptors enable rational solvent selection in asymmetric catalysis. *Chem Sci*. 2019;*10*(27):6697–706. https://doi.org/10.1039/C9SC01844A.
151. See XY, Wen X, Wheeler TA, Klein CK, Goodpaster JD, Reiner BR, et al. Iterative supervised principal component analysis driven ligand design for regioselective Ti-catalyzed pyrrole synthesis. *ACS Catal*. 2020;*10*(22):13504–17. https://doi.org/10.1021/acscatal.0c03939.
152. Hammer B, Nørskov JK. Electronic factors determining the reactivity of metal surfaces. *Surf Sci*. 1995;*343*(3):211–20. https://doi.org/10.1016/0039-6028(96)80007-0.
153. Pauling L. The nature of the chemical bond. IV. The energy of single bonds and the relative electronegativity of atoms. *J Am Chem Soc*. 1932;*54*(9):3570–82. https://doi.org/10.1021/ja01348a011.
154. Montemore MM, Medlin JW. Predicting and comparing C—M and O—M bond strengths for adsorption on transition metal surfaces. *J Phys Chem C*. 2014;*118*(5):2666–72. https://doi.org/10.1021/jp5001418.
155. Antoniou D, Schwartz SD. Toward identification of the reaction coordinate directly from the transition state ensemble using the kernel PCA method. *J Phys Chem B*. 2011;*115*(10):2465–9. https://doi.org/10.1021/jp111682x.