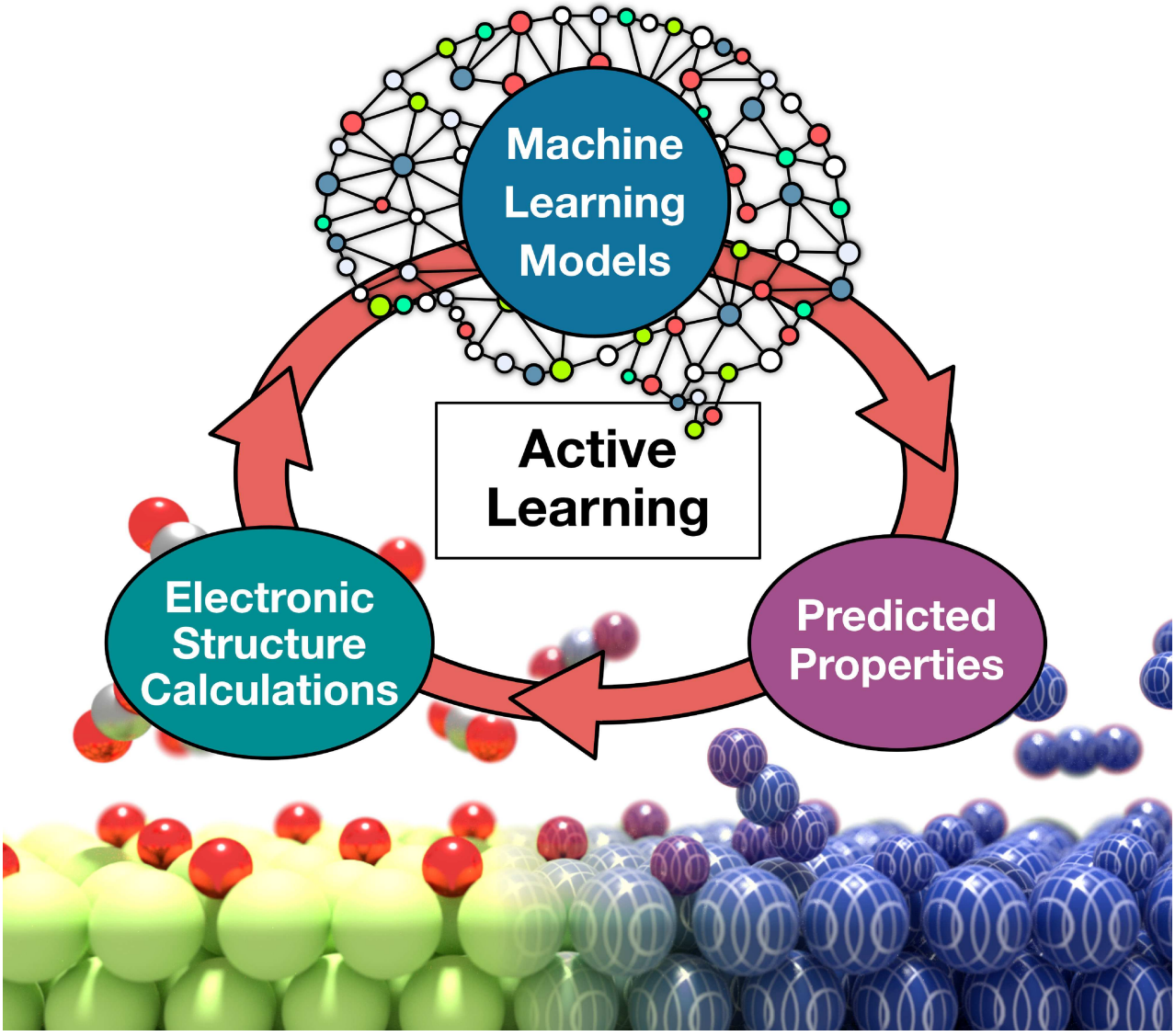
ChemPubSoc Europe
DOI: 10.1002/cctc.201900595
CHEMCATCHEM Reviews

VIP *Very Important Paper*

Machine Learning for Computational Heterogeneous Catalysis

Philomena Schlexer Lamoureux,*[a, b] Kirsten T. Winther,[a, b] Jose Antonio Garrido Torres,[a, b] Verena Streibel,[a, b] Meng Zhao,[a, b] Michal Bajdich,[a, b] Frank Abild-Pedersen,[a, b] and Thomas Bligaard[a, b]



Big data and artificial intelligence has revolutionized science in almost every field – from economics to physics. In the area of materials science and computational heterogeneous catalysis, this revolution has led to the development of scientific data repositories, as well as data mining and machine learning tools to investigate the vast materials space. The goal of using these tools is to establish a deeper understanding of the relations between materials properties and activity, selectivity and stability – the important figures of merit in catalysis. Based on these insights, catalyst design principles can be established, which hopefully lead us to discover highly efficient catalysts to solve pressing issues for a sustainable future and the synthesis of highly functional materials, chemicals and pharmaceuticals. The inherent complexity of catalytic reactions quests for machine learning methods to efficiently navigate through the high-dimensional hyper-surfaces in structure optimization problems to determine relevant chemical structures and transition states. In this review, we show how cutting edge data infrastructures and machine learning methods are being used to address problems in computational heterogeneous catalysis.

1. Introduction

In order to advance in solving the energy and environment crisis, it is necessary to develop highly efficient catalysts for the generation and storage of fuels as well as efficient and sustainable ways of producing fertilizers, base and fine chemicals.[1] With the advent of the digital era, data science and machine learning have largely advanced and shaped the fields of chemistry,[2–4] materials science[4–7] and homogeneous catalysis.[8,9] Data about heterogeneous catalysts (chemical or electro-chemical) can be used to train machine learning models in order to predict quantities of interest, such as measures of the important figures of merit - stability, activity and selectivity. As we will discuss in more detail in this review, machine learning and data science can be used to establish an understanding on underlying materials properties that define the quality of a material as a catalyst for a specific reaction. To clarify, whereas *data science* is the umbrella term for data-related operations used to transform data into insights or predictions, *machine learning*, being a sub-field of data science, deals with algorithms that can be used to categorize data and predict future outcomes. In this context, it is also worth specifying that *deep learning*, on the other hand, denotes a certain type of algorithm class based on artificial neural networks.

As heterogeneous catalysis lies at the intersection between chemistry, solid-state physics and materials science, it gains immediate benefits from data science related developments in those fields. The presence of complex reaction sites and interfaces in heterogeneous catalysis (including heterogeneous electrocatalysts), however, constitutes a major distinction of the field from those mentioned above. It is therefore practical to think of heterogeneous catalysts as "active phases", constituent of an evolving material in the reaction environment. Heterogeneous catalysis is furthermore a multi-scale phenomenon, to which machine learning can be readily applied to improve the various levels of theory used to describe the effects arising at different time and length scales. For instance, machine learning has been exploited for the development of efficient force fields[10,11] to accelerate the description of potential energy surfaces, neural network model chemistries,[12] the initialization of micro-kinetic models[13] and reactor design.[14] However, since the catalyst functionality is often determined by microscopic enviroments such as step sites and defects,[15] atomistic models, in particular in combination with density functional theory (**DFT**), remain the workhorses of computational heterogeneous catalysis research. The combination of DFT-based materials description with cutting-edge data science and machine learning techniques and novel data repositories[16] therefore holds an immense potential for the acceleration of catalyst discovery.

Historically, the design of materials, including heterogeneous catalysts, was based on intuition and experiment. This approach translates into long commercialization time lines of 10–20 years, which not only stifle investment in early stage research but also hinder the solution of important energy and environment challenges.[17,18] An advantage of computational studies is that they can produce more comprehensive materials databases in a shorter time frame than the experimental approaches. Having faster access to a more comprehensive data collection is definitely beneficial for the application of data science related techniques.[19] For this reason, DFT-based materials databases are steadily growing and efficient data structures along with sharing strategies are under development. Data science and machine learning methods can help us to transform the data into actionable insights. However, there are domain-specific challenges to overcome, such as finding unified methods, which are able to accurately describe a larger range of materials types: Even within the DFT framework different materials need different tweaks, such as accurate van-der-Waals description or methods to obtain the correct band gap. Another domain-specific challenge is for instance the usage of simplified atomic models and thereof derived descriptors, which may not capture the relevant catalyst functionality.

In this review, we will focus on the application of machine learning techniques in combination with electronic structure simulations, in particular density functional theory (DFT). We will first outline some important technical aspects, like data structures, featurization and surrogate machine learning mod-

[a] *Dr. P. Schlexer Lamoureux, Dr. K. T. Winther, Dr. J. A. Garrido Torres, Dr. V. Streibel, Dr. M. Zhao, Dr. M. Bajdich, Dr. F. Abild-Pedersen, Dr. T. Bligaard*
SUNCAT Center for Interface Science and Catalysis, SLAC National Accelerator Laboratory, 2575 Sand Hill Road, Menlo Park, California, 94025, United States
E-mail: schlexer@stanford.edu
[b] *Dr. P. Schlexer Lamoureux, Dr. K. T. Winther, Dr. J. A. Garrido Torres, Dr. V. Streibel, Dr. M. Zhao, Dr. M. Bajdich, Dr. F. Abild-Pedersen, Dr. T. Bligaard*
Department of Chemical Engineering, Stanford University, 443 Via Ortega, Stanford, CA 94305, United States

This manuscript is part of the Special Issue dedicated to the Women of Catalysis.

els. Then, we will provide an overview of recent work in the field and discuss future opportunities and perspectives.

2. Machine Learning Concepts

The main goal of catalysis research is to optimize catalysts in terms of activity, selectivity, stability, sustainability and price. Computational heterogeneous catalysis research is well suited

Philomena Schlexer Lamoureux, PhD in Materials Science and Nanotechnology, joined SUNCAT as a Feodor Lynen Research Fellow in 2017. She combines first-principle modeling with machine learning and data science to establish site-specific quantitative structure-property relationships for catalysis. She developed and implemented methods for atomic site-based feature analysis and established models to understand and efficiently calculate electro-catalytic reactions. (ca. 30 papers, 1 chapter, May 2019)

Kirsten Winther obtained her PhD in Physics from the Technical University of Denmark in 2015, where she applied electronic structure theory and many-body approaches to study electronic excitations in atomic scale materials and van der Waals heterostructures. She joined the SUNCAT center at Stanford University as a postdoc in 2017, where she is developing the Catalysis-Hub.org database platform for computational catalysis as well as new methodologies for automated materials discovery.

Jose A. Garrido Torres obtained his PhD in Chemistry in 2017 at the University of St Andrews, UK. His research focused in developing quantum mechanical methods and modeling systems related to the field of surface science, involving molecular electronics and heterogeneous catalysis. He joined the SUNCAT center at Stanford University in 2017. His research focuses on developing and implementing surrogate machine learning algorithms for accelerating and automating the search of novel materials.

Meng Zhao received her PhD in Chemistry in 2017 at Case Western Reserve University, USA. She employed materials modeling to simulate solid-liquid interfaces for unraveling surface reaction mechanisms in electrochemical applications. She joined the SUNCAT center at Stanford University in 2017. She focuses on developing a web application to screen materials electrochemical stabilities to facilitate materials innovation.

Verena Streibel (née Pfeifer) studied Materials Science at the Technical University Darmstadt. She obtained her PhD in Chemistry in 2016 from the Technical University Berlin for her doctoral studies with Prof. Robert Schlögl at the Fritz-Haber-Institut der Max-Planck-Gesellschaft. Therein, she developed in situ electrochemical cells for Ambient Pressure XPS and XAS to observe electronic structure fingerprints of oxygen-evolving iridium surfaces. In 2018, she joined the SUNCAT Center at Stanford University as a postdoctoral Feodor Lynen Research Fellow. Here she uses DFT, site-specific scaling relations, and micro-kinetic modelling to identify the next generation of methanol synthesis catalysts.

Michal Bajdich received his PhD in Physics at North Carolina State University and is a staff scientist at SUNCAT since 2013. His research focuses on understanding and discovering catalysts for electro-catalytic reactions based on oxides and related materials. He combines first-principle methods with catalysis informatics and has developed theoretical models for electrochemistry and XAS spectroscopy, as well as computational DFT methodology.

Frank Abild-Pedersen got his PhD in Theoretical Physics at Technical University of Denmark, and is Senior Staff Scientist at SUNCAT since 2007. His research interests cover the theoretical description of catalysis and materials properties. In particular, he is interested in modeling catalysis relevant phenomena; reactivity, selectivity promotion, poisoning and deactivation and carefully testing devised models against well-defined experiments through collaborations with experts in the field. He has 121 published papers (cited ca. 15,278 times, H-index 56, Google Scholar May 2019), and successfully filed 2 patents.

Thomas Bligaard got his PhD in Theoretical Physics at Technical University of Denmark, and is Co-director of the SUNCAT Center for Interface Science and Catalysis since 2018. His research interest cover electronic structure theory, kinetic modeling, optimization, materials search, Bayesian statistics, machine learning, and heterogeneous catalysis. He is author of more than 80 papers, 1 textbook, 3 book chapters, and 2 patents. Cited more than 14,000 times, H-index 47 (ISI, January 2019).

to understand and enhance the first three aspects, which is reflected in two major approaches:

(i) Understanding and quantifying which catalyst properties are relevant for the catalyst figures of merit, i.e. activity, selectivity and stability.
(ii) Screening the materials space in terms of composition and structure, based on those activity descriptors, to find better catalysts.

In the context of machine learning, the first approach is closely related to identifying significant features (synonyms are descriptors, predictors, independent variables) from raw data. Features can be used to understand and predict the target properties (synonyms are outcome, dependent variable). Whereas the targets are activity, selectivity, and stability, the features can be any physical property or even synthesis methods. Commonly, features are based on the structure and composition of the material.[20]

The ideal machine learning procedure begins with a problem to be solved. Having formulated the question, we proceed with the (desirably bias-free) data acquisition, then process the data to make it accessible to the machine learning model, then train a predictive and/or explanatory model on the data, and finally evaluate the model performance. We will follow this procedural logic in this section, walking through the three major steps in machine learning: (1) data acquisition and storage, (2) featurization and (3) modeling. In the following section, we will address all of these aspects.

2.1. Data Infrastructures and Sharing

The amount, nature and accuracy of the data ultimately determine the set of questions that can be answered via machine learning, as well as the machine learning models that can be used. Although data acquisition via density functional theory (DFT) is much faster than experimental screening, it is still limited by computing speed, access and cost. This limitation becomes important as the atomic model size increases with the complexity of the catalyst. Not only therefore, data sharing and recycling is key for the advancement of the field. In the next sections, we will outline data sharing principles and potentials.

2.1.1. Computational Materials Databases

Given the solid-state nature of heterogeneous catalysts, the field has a major conceptual and technical intersection with materials informatics, where, not surprisingly, computational databases have been flourishing in the past decade. Recently, Lin provided an overview on 13 computational materials databases.[21] Some of these databases have grown extensively. Prominent examples are Materials Project,[22] MaterialsCloud/AIIDA,[23] Novel Materials Discovery (NoMaD),[24] the Open Quantum Materials Database (OQMD),[25] the AFLOWlib repository[26] and the Computational Materials Repository (CMR),[27] which combined have many hundred thousands of entries. Also, specialized databases featuring phonon calculations[28] and 2D materials[29] have recently been presented. Important databases for experimentally observed crystal structures includes The Cambridge Database,[30] The Inorganic Crystal Structure Database,[31] the Crystallographic Open Database (COD)[32] and the Inorganic Material Database (AtomWork).[33] A majority of the entries in the databases mentioned above are solid-state materials. Whenever the catalyst performance can be related to bulk properties, solid-state materials databases constitute a valuable resource for a direct catalyst screening. Because bulk properties cannot capture surface-specific effects on the catalyst activity, a database of surface reaction energetics was realized first by CatApp[34] and recently Catalysis-Hub.org,[16] an open repository for chemical reactions on catalytic surfaces, featuring 100,000+ chemical adsorption and reaction energies from electronic structure calculations. On this web-platform users can search for reactions energies among a large range of surface compositions, including single crystals, metal alloys, transition-metal oxides and 2D-materials, all combined with a graphical user interface (GUI), which allows to access and visualize the atomic geometries and stored properties in the internet browser.

A prerequisite for using DFT calculations for data science and machine learning approaches is that information is structured and made accessible in such a way that smaller, relevant subsets can be retrieved from large sets of data. Retrieval should be efficient such that data can be selected based on properties of interest to the model in question in a fast and automated manner. According to the FAIR guiding principles for data management,[35] data should be *findable*, *accessible*, *interoperable* and *reusable* in order to have maximum impact.

The term findable entails that data and metadata should be easy to discover for humans and be machine-readable. Also, data and metadata should be accessible with standard protocols (such as through an http request), ensuring simple and unified user access. In practice, this accesscan be achieved by providing user-friendly web-interfaces as well as application programming interfaces (APIs). Such an API exists for the Materials Project, the Materials Application Programming Interface (MAPI), which is an interface to programmatically query and interact with the database based on the Representational State Transfer (REST) pattern for the web. Many of the databases mentioned above use similar approaches, with data fetching over http, ensuring transferability of a query to different databases. In addition, Python wrappers such as the pymatgen library of Materials Project, support data access. Data should be interoperational in the sense that it should be integrated with other data when relevant, and coupled to applications or workflows to analyze, process and store data. Here, using the Resource Description Framework (RDF) for data interchange, ensures that links between data and metadata can be made machine readable. Reusable means that data should be described with enough details to reproduce the obtained result, and to be used in another context than initially intended, which especially relevant in a scientific concept. A meaningful metadata labeling also depends on the development of

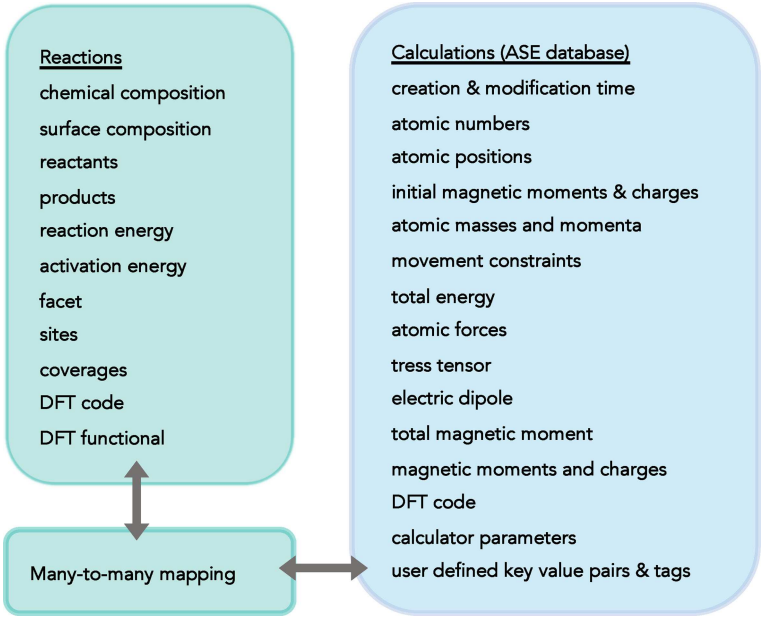


Figure 1. Schematic of the SQL data structure of Catalysis-Hub.org, used to store reaction energies (Reactions table, green) and DFT calculations (ASE database, blue). Since each reaction energy involves several DFT calculations (and single DFT calculations can potentially be used for several reactions), a many-to-many mapping schema is used to preserve connections between the table rows.

appropriate vocabularies, or *ontologies*, for heterogeneous catalysis and electrochemistry. An ontology for the periodic table in relation to catalytic materials was recently proposed.[36,37] Moving forward, an ontology for the structure of the molecule-surface interface could be developed and integrated with a recently proposed ontology for chemical reactions.[38]

As accommodating the principles above relies heavily on finding efficient ways to store and retrieve data, we will discuss some relevant data structures for storing DFT calculations for machine learning purposes.

2.1.2. Computational Materials Data Storage Principles

In computational materials science, data is predominantly generated by DFT calculations, where both the input parameters (DFT code, version, functional, k-points etc.) and output of the calculations should be stored to ensure reproducibility. DFT results of interest include the total potential energy, atomic numbers and positions, forces, charges, magnetic moments and Kohn-Sham/GW band gaps. Depending on the properties of interest, detailed information about the electronic structure, such as the Kohn-Sham eigenvalues and wavefunctions can also be of relevance, although these take up significantly more storage space. Therefore, condensed information about the electronic structure, such as the density of states and the electronic band energies along high-symmetry paths (band structure) is usually a better choice.

Relational databases using structured query language (SQL), such as SQLite, mySQL and PostgreSQL, are generally well suited to store well-defined collections of properties. Data is arranged in ordered tables, with columns and rows, that can be linked to capture connections between different types of entries. The power of relational databases is the ability to select subsets of data with high efficiency, by constraining the value of one or more columns. The Atomic Simulation Environment (ASE), a popular software package for handling atomic structures,[39] provides database functionality that uses SQLite and PostgreSQL backends to store atomic structures, calculation details and results, as well as user-defined meta-data. Properties such as atomic numbers, positions, forces, magnetic moments and charges are stored in arrays in the table, while calculator parameters and user-defined key value pairs are stored in dictionaries. The ASE database is used on the servers of the computational materials repository (CMR) and Catalysis-Hub.org, and can be used locally. On Catalysis-Hub, the data structure is supplemented with additional tables to store pre-parsed reaction energies and barriers, so that these properties are easily available. An illustration of the (simplified) Catalysis-Hub data structure is shown in Figure 1.

For some applications, a more flexible database structure, provided by noSQL/not-only SQL database backends, is desirable. In this formalism, no strict structure is imposed on the data. Instead, data can be viewed as a collection of key-value pairs, which can accommodate changing data formats and structures. For example, the Materials Project utilizes noSQL database formalism of MongoDB to provide a flexible storage of user-defined workflows. PostgreSQL, which is used for Catalysis-Hub as well as MaterialsCloud,[23] also provide noSQL-like features and is therefore a good compromise for accommodating structured as well as non-structured data.

2.2. Materials Featurization

In machine learning, featurization is the process of choosing and engineering relevant input parameters (features) that are characteristic for the properties of interest. As catalysis is a multi-scale phenomenon, features could, in principle, be chosen from any length scale. That is, using information based on continuum models and macroscopic, microscopic, atomistic, and electronic properties. Especially in experimental high-throughput studies, non-atomistic features, such as synthesis conditions and meso-scale structuring, are popular.[40,41] A large part of computational studies is based on electronic structure methods because of their precision in describing chemical bonds. Here, the featurization can be based on the atomic and electronic structure. Given the fact that the activity (rate of reaction) of a catalyst is ultimately defined on an atomic scale, there should in principle be a direct mapping from the structure and composition of the active phase to its activity. This ideal case is far from being computable currently, but approximate models and methods like density functional theory (DFT), together with first-principle thermodynamics and transition state theory, have served as work horses in computational catalysis research.[42]

The representation of a material in terms of features is also called fingerprinting. The subset of relevant features, which can be used themselves as simple surrogate models for the target (figure of merit), are also called descriptors,[43] and we will use both terms interchangeably. Features should preferably have a physical meaning and be easily accessible, i.e. easy to compute or look up. Furthermore, they should be universal, making them valid strong predictors across a variety of different catalyst types. Ideally, it should be possible to infer the catalyst material from a given set of promising features, which can be seen as the reverse process[7] of featurization.

Descriptor-based approaches have a long-standing history in catalysis. They date back to the Sabatier principle[44] and the Brønsted-Evans-Polayi[45,46] relationships developed in the early 20th century, which relate reaction energetics to activity trends. With the advances in DFT and surface science, adsorption energies of molecules on transition metal surfaces evolved as useful descriptors.[47–52] Scaling relations, together with activity maps, and the d-band model constitute a comprehensive set of concepts to understand and predict catalytic activity and selectivity by mapping the rate of a full catalytic reaction onto a few descriptors.[42]

In the context of DFT, we can distinguish between electronic structure descriptors and atomic structure descriptors. The level of theory from which the feature is established is an important parameter affecting its quality, complexity and accessibility. The d-band center or width, for instance, requires the computation of the density of states (DOS). As much of the transition metal chemistry is defined by the alignment of the metal's *d*-band center and the adsorbate frontier orbitals, it is indeed a valuable feature.[50,53,54] Other features that can be derived from the electronic structure are, for example, band gaps, electron mobilities and number of states at the Fermi level. Recently, the average *2p*-state energy of surface oxygen atoms was shown to be a strong descriptor for the oxygen reactivity at metal and metal-oxide surfaces.[55,56] The computation of the electronic structure, however, is computationally expensive, which constitutes a limiting factor for the usage of this feature type.[57] However, features that are usually based on electronic structure methods can also be estimated from computationally inexpensive methods: Gradient-boosted decision trees have been used to predict the metals *d*-band center using six tabulated experimental properties.[58]

DFT calculations require atomic structure models consisting of atomic identities (element ID) and coordinates representing the active phase. It is possible to generate atomic structures without DFT calculations and, especially in the case of bulk materials, machine learning can provide valuable interpolation schemes to predict materials properties solely based on their structure.[59] The featurization process based on atomic structures usually requires manually constructed feature vectors or complex transformation of atom coordinates.[60] The features are then based on geometric and compositional metrics, such as: stoichiometry, coordination numbers,[61,62] distances, angles, bulk symmetry operations,[63] formal oxidation states and formal electron counts. The computation of these metrics is not trivial, as many of them rely on empiric parameters, such as cut-off distances for nearest-neighbor detection. Depending on the reaction, the active structure can be represented by a bulk phase, by a surface or by more complex models as shown in Figure 2. The simplicity of periodic bulk models enables straight-forward high-throughput computations, giving rise to the vast amount of data stored in online materials databases. An example of an advanced bulk materials featurization scheme is based on Voronoi-tessellation.[64,65]

For some reactions, the catalyst activity is significantly defined by surface properties and especially in the case of surface-sensitive reactions, using surface-specific descriptors is of vital importance.[48,66] As complex functionality emerging through surface-specific attributes is not expected to be captured by bulk descriptors, the development of efficient surface descriptors is key. Features based on coordination numbers have the advantage that they do not rely on absolute atomic positions, and can be generated without computing the global structure minimum. Good estimates for empirical cut-off parameters are needed, however, in order to reduce errors. Generalized and orbital-wise coordination numbers have been successfully used as features in linear scaling relations between DFT adsorption energies on mono-metallic surfaces.[67–69] An alternative purely coordination-based approach uses the binding energy of the metal adsorption sites as a feature to predict adsorption energies of thermo-chemical descriptors on mono-metallic surfaces and nanoparticles.[70,71] This approach unveils a new family of linear scaling relations between adsorption site stabilities and adsorption energies. The advantage of this approach is that by including information of the chemical environment, it is readily extendable to bi- and possibly multi-metallic systems.[72,73] To facilitate the feature generation, machine learning models (see next sections) can be trained to automatically extract features from the data, an approach is called model stacking. The prime example of model stacking

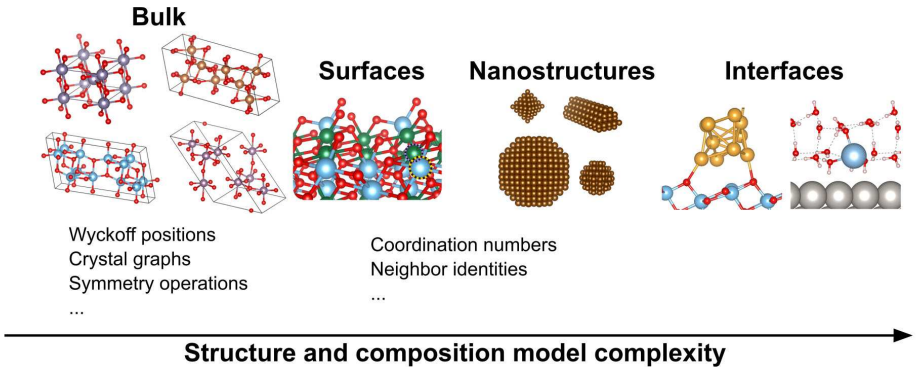


Figure 2. The emergence of structure and composition feature complexity with model accuracy. Highly functionalized catalysts may quest for highly complex features.

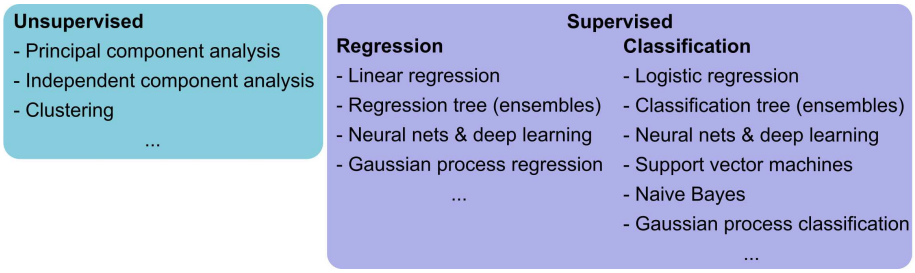


Figure 3. Overview on machine learning models categories with popular examples.

are (convolutional) deep neural networks, often used in computer vision.[60] Another example for the circumvention of a "manual" featurization, is the usage of a direct similarity measure between atomic neighbor environments called smooth overlap of atomic positions (SOAP), which has been used for the development of neural network potentials.[74]

2.3. Machine Learning Models

As outlined before, the machine learning pipeline ideally begins with the selection of a target problem, upon which data are acquired, featurized and then finally used for modeling. The goal of the modeling may be mere prediction, a causal understanding or even a mechanistic understanding. Depending on the goal, some models are more appropriate than others. Machine learning models trained on materials/atomic data sets are typically good at *interpolating* to similar systems, but show less accuracy *extrapolating* to other material types, which is a general phenomenon in machine learning. Therefore the ability to generalize predictions to data dissimilar of those of the training set is a desirable property of a machine learning model.

In the following, we will outline and compare a few popular machine learning model types and discuss their suitability for the different goals. We can generally distinguish between unsupervised and supervised models. The categorization of methods is summarized in Figure 3.

2.3.1. Unsupervised Models

Unsupervised models comprise all methods that find patterns within unlabeled data where no target values are supplied. Such approaches are well suited to find hidden patterns in the data, which may help to establish causal and mechanistic theories and can be used to facilitate later predictive studies. Although target values are not used in the context of the loss function, they can still be added as another feature dimension.

One of the simplest and most popular unsupervised learning algorithms is ***K*-means clustering**, where the purpose of clustering is to assign data points to different groups (clusters) based on similarity. The algorithm assigns each data point to one of k clusters, where k is a hyper-parameter which has to be chosen in advance. *K*-means then uses a similarity function, for instance the L^1 or L^2 distance, where p and q are two different points (vectors) in the space of interest. In the case of k-means clustering p_i are the centroids/seeds and q_i the data points in feature space. When using the L^1 or L^2 distance as a loss function p_i are the targets and q_i the predicitons.:

$$L^1 = \sum_i |p_i - q_i| \tag{1}$$

$$L^2 = \sqrt{\sum_i |p_i - q_i|^2}, \tag{2}$$

Starting from k initial points (seeds), the distance between each data point to these points is calculated. The points closest to the initial seeds are added to the cluster. Then the new cluster centroid is calculated as the average of all points in that cluster. Varying k and initial seed positions, the procedure can then be optimized using for instance pooled variances of the mere sum of distances between final cluster members and their centroids. In the case that target values are known, we can improve clustering by using the target information. This can be done via the adjusted random index,[75] mutual information,[76] homogeneity, completeness and V-measures,[77] just to mention a few. There are various related clustering methods, such as hierarchical clustering, mixture models, and density-based spatial clustering.[78] Clustering has been used to featurize spectroscopic data from experimental catalysts.[8] The resulting data set was then fed into quantitative structure-property relationship (QSPR) models, including logistic regression, neural networks and decision trees, achieving high accuracy in predicting catalytic performance.

The K-means clustering is easy to implement, understand and visualize. Due to it's simplicity, it is easy to adjust, and it segments large datasets while being computationally efficient.[79] There are, however, also a few assumptions and drawbacks: (1) Resulting clusters are spherical and of equal size, which causes problems if features have different scales or variances, whereby the later can be addressed using Gaussian Mixture models[80] (2) the algorithm can get stuck in local minima, due to insufficient sampling of random initial centroids, (3) the algorithm won't differentiate between homogeneous and clustered data.

Another popular unsupervised method is **principal component analysis (PCA)**, where a set of correlated variables is transformed to an orthogonal basis of uncorrelated variables. In machine learning terms, PCA corresponds to a linear transformation of the feature space to a (sub-)space that minimizes linear correlation between the feature dimensions. The linear correlation between the different features is given by the off-diagonal entries of the covariance matrix $\mathbf{C}_X$, equation (3), where $\mathbf{X}$ is the data matrix with features as rows and samples as columns. Note that we only account for linear correlation between features.

$$\mathbf{C}_X = \frac{1}{n}\mathbf{X}\mathbf{X}^T \tag{3}$$

$$\mathbf{D} = \mathbf{P}\mathbf{C}_X\mathbf{P}^T \tag{4}$$

To find the new uncorrelated feature dimensions, we can simply compute the eigenvectors of the covariance matrix, equation (4), by finding the matrix $\mathbf{P}$ that diagonalizes $\mathbf{C}_X$. The eigenvectors in this example are the row vectors of $\mathbf{P}$ and constitute the principal components of $\mathbf{X}$. PCA was successfully used to determine the most predictive activity descriptors of dual-site transition metal oxide oxygen evolution catalysts.[81] Unfortunately, PCA typically lacks intuitive physical interpretability.

PCA can be used to analyze the feature relevance by selecting features according to the magnitude (from largest to smallest in absolute values) of their PC coefficients (loadings). PCA is especially helpful in prediction-oriented applications, as opposed to mechanistic studies. Here, PCA serves as powerful dimensionality reduction, which makes training and using models for prediction faster, i.e. when used as a feature engineering (pre-processing) method. PCA also is based on a few assumptions and and has some drawbacks: (1) It's based on the assumption that variance equals importance and thus needs feature normalization. (2) Components lack physical interpretability. (3) The algorithm does not optimize for feature separability. (4) It accounts only for linear combination of features only, i.e. is an inherently linear method. (5) It is computationally inefficient for large data sets due to the eigenvalue decomposition.

Other interesting, less commonly used unsupervised learning methods are self-organizing maps,[82] and hidden Markov models,[83] a similar supervised method is independent component analysis (ICA) which optimizes the data separability.[84]

2.3.2. Supervised Models

In supervised learning methods, the machine learning model maps a set of features to a setof targets. The model predicts outcomes $\hat{\mathbf{y}}$, which are compared to the true outcome $\mathbf{y}$ by a similarity function. The similarity measure is used in a loss function, which outputs a loss measure. The loss is then used to optimize the model. The goal is to minimize the loss with respect to the model's parameters (weights and hyper-parameters), which generally corresponds to a non-convex optimization problem. As shown in Figure 3, we can distinguish between regression and classification. Regression works with continuous numeric outcomes, as for instance band gaps, adsorption energies, turnover frequencies, selectivities, molecular weight of polymers etc. Classification, on the other hand, can be used to classify materials as "conductor"/semi-conductor"/"insulator". Supervised models are primarily predictive models, although causal and/or mechanistic insights can be gained as well. The model complexity is related to (1) bias and (2) variance. A biased model with too few parameters oversimplifies the relation between variables and is therefore said to under-fit". A model with too many parameters may result in too large high prediction variability and is thus said to over-fit". In the following, we will shortly outline the most popular methods and give some examples on how they can be applied to computational heterogeneous catalysis problems.

2.3.3. Linear Models

A simple and very commonly used method is linear regression. Linear regression has been widely used to establish scaling relations, descriptor analysis and feature selection, as we will discuss in more detail below. The simplest linear regression version is ordinary least squares (OLS), in which the loss function to be

minimized is the sum of squares ($L = \sum_i |y_i - \hat{y}_i|^2$) of the predicted and true outcome differences. This method has an analytical solution which can be readily computed, as shown in equation (5), giving the coefficient matrix $\hat{\beta}$.

$$\hat{\beta} = (X^T X)^{-1} X^T y \quad (5)$$

$$\hat{y} = X\hat{\beta} \quad (6)$$

With increasing feature space, OLS is prone to over-fitting. This is because features may be irrelevant or suffer of multi-collinearity. Therefore, regularization methods have been developed of which the most popular are (1) Ridge regularization[85] (based on L^2 distance) and (2) LASSO[86] (least absolute shrinkage and selection operator), which is based on the L^1 distance and therefore favors sparsity. Finally, the elastic net method[87] combines the two approaches. Linear models assume that the error on the target has the same distributions across all feature values, which corresponds to negligible heteroscedasticity, i.e. where features have different distributional variances.

Linear models are widely applicable to many problems in heterogeneous computational catalysis. The analysis of linear regression coefficients has brought insights about the role of the bond order between adsorbate and surface in scaling relations[47] using the CH_3 to C scaling as an example (Figure 4). This observation has enabled a fundamental understanding of the parameters that define catalytic activity.[47,50,71,88–91] Linear regression has furthermore been successfully used to predict the *d*-band center within a mean absolute error of about 0.26 eV from a set of experimental elementary descriptors.[58]

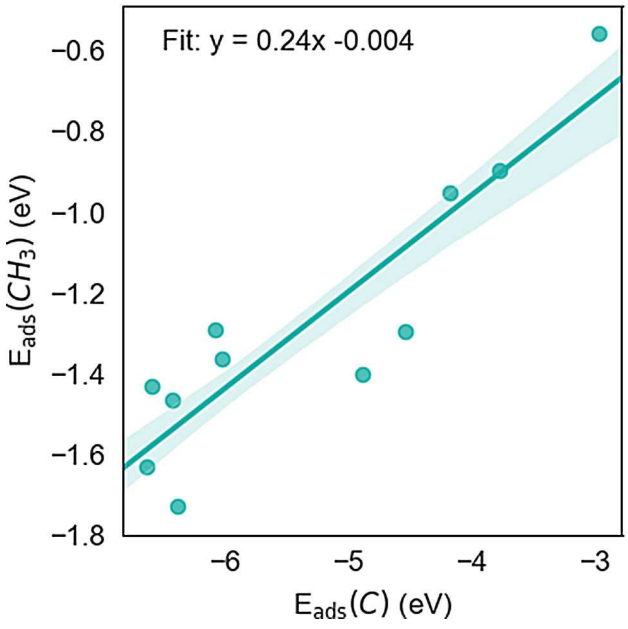


Figure 4. Scaling relations between adsorption energies of CH_3 and C on metal surfaces. The coefficient of the linear regression fit (0.24) is related to the bond order difference between the two adsorbates. The 68% confidence interval on the predicted data is shown. Data re-plotted from reference [47].

Recently, Hong used linear regression and LASSO regularization for automatic feature selection, which revealed the importance of the transition metal e_g/t_{2g} electron occupancy in determining the catalytic activity.[92]

Linear regression is a powerful, yet simple method, which can be used in diverse ways to understand and predict outcomes. It has been widely applied to many problems in computational heterogeneous catalysis, and lies at the core of linear scaling relations. In combination with regularization techniques, or using non-linearly engineered features (e.g. new feature a*b from features a and b), it can deliver valuable insights on feature importance and physical relations. Drawbacks are that in it accounts only for linear feature interactions. Linear regression furthermore assumes that there is no or little multicollinearity between features (otherwise the solutions are not unique anymore) and cannot account for heteroscedasticity.

2.3.4. Kernel Methods

As mentioned above, linear regression can be augmented to include non-linear correlations by manually engineering non-linear features combinations, or by using the kernel trick in combination with a non-linear kernel. The kernel trick enables us to calculate the similarity of two points, given by the dot product $\mathbf{x} \cdot \mathbf{x}$, in a space that is higher-dimensional than the feature space. This trick is done by computing $\mathbf{k}(\mathbf{x} \cdot \mathbf{x})$, where $\mathbf{k}$ is the kernel. Mathematically, the process corresponds to mapping the data points into a higher-dimensional feature space and then computing the dot product there, however, the transformation is never made explicitly. As long as the mapping function exists, it is sufficient to calculate $\mathbf{k}(\mathbf{x} \cdot \mathrm{thbfx})$. We can use this trick and substitute the dot products that arise in the solutions for the linear regression equations, see equation (5).

When the kernel trick is applied in combination with Ridge regularization, we speak of kernel ridge regression (KRR).[93] The form of the model learned by kernel ridge regression is identical to support vector regression (SVR),[94] but their loss functions are different. There are quite a few examples how KRR has been applied to computational heterogeneous catalysis problems. For instance, KRR has been used to predict DFT adsorption energies of hydrogen on a large number of sites on different types of nanoclusters.[95–97] For this, the cluster sites were featurized using the SOAP method, see section 2, and then KRR was used to predict the adsorption energy.[95] In another study, Noh *et al.* presented a machine learning model to predict CO adsorption energies on metal alloy surfaces.[57] They used simple non-first-principle input features, such as the linear muffin-tin orbital theory (LMTO)-based *d*-band width and the geometric mean of electro-negativity. By combining these features with an active learning algorithm, they obtained highly accurate adsorption energies.[57]

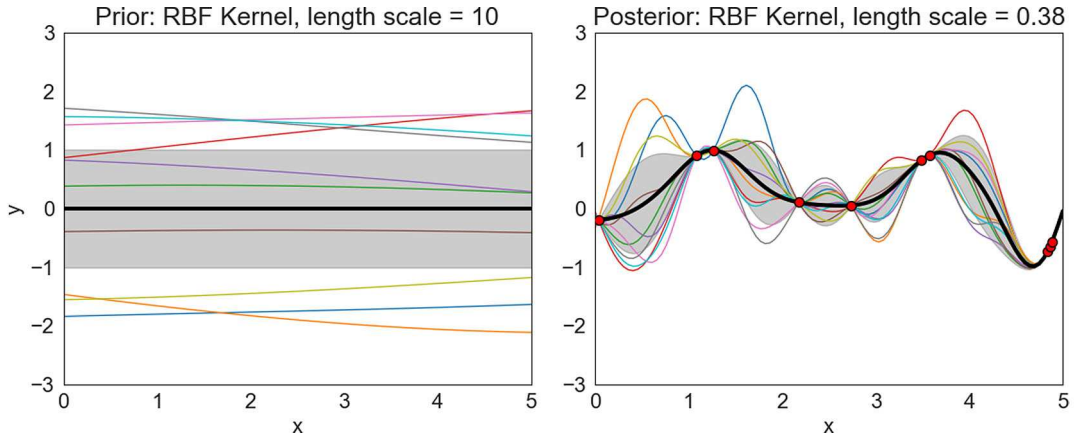


Figure 5. Prior and posterior distribution over functions mean (black), standard deviation (grey) and a few examples from the function space (colors).

2.3.5. Gaussian Processes

The kernel trick is often combined with a Gaussian process, which delivers not only predictions but also uncertainty measures for those predictions. A Gaussian process uses Bayesian probability theory and assumes a multivariate normal distribution over functions with a continuous domain, i.e. the numeric feature space. The form of the functions is given by the kernel. A commonly used kernel is the radial basis function (RBF) kernel, shown in equation (7), where x_i and x_j are two data vectors with the different feature values as entries and l is the length scale parameter, which is learned by fitting the kernel to the data. Shorter length scales across many independent features increase the model complexity, and thus the risk of over-fitting. The complexity of a kernel can be quantified by it's rank.[98]

$$k(x_i, x_j) = \exp\left(-\frac{1}{2l^2}(x_i - x_j)\right) \tag{7}$$

The starting distribution over the functions is also called *prior*. A sample of functions drawn from an RBF prior, together with mean and standard deviation of all function values is shown in Figure 5. After fitting the kernel to the data, we obtain the *posterior* distribution over functions that take the data values of our training set, shown as red points in Figure 5. Note that the method is not sparse and hence stores" the training data in the model, 5 for prediction.

Gaussian processes have been widely used to estimate potential energy surfaces (PES), for instance in the form of interatomic potentials[99] and surrogate models for first-principle calculations, see section 2. Ulissi *et al.* used Gaussian processes to predict adsorption energies, transition state scaling and rate-limiting steps for the complex reaction network of the syngas reaction over Rh(111).[100] In a different study, Ulissi *et al.* used Gaussian processes to generate approximate Pourbaix diagrams.[101] Takigawa *et al.* found that Gaussian process regression showed superior accuracy compared to five linear models and five nonlinear models with a prediction error of the *d*-band center below 0.2 eV compared to the density functional theory benchmark.[54]

Gaussian processes are very powerful, and hyperparameters can be neatly optimized by maximizing the marginal likelihood. Furthermore, Gaussian processes predictions come with estimated errors of that prediction, which are empirical confidence intervals. Gaussian processes are not sparse, that is they use and iterpolate between training observations, which makes them accurate, but also computationally inefficient during prediction. They also loose efficiency in higher-dimensional features spaces.

2.4. Decision Trees & Ensemble Methods

Another popular and powerful supervised learning method are decision trees. Decision trees divide up the feature space with consecutive linear decision boundaries by answering a series of "if-then-else" questions related to the descriptor values at each branch point until a terminal leaf node is hit.[8] The working principle is shown in Figure 6.

Decision trees use a greedy mechanism that divides the feature space at each decision point upon the feature which best predicts the targets without considering later decisions. The tree learns from the training data by going through all possible features to split on, optimize the splitting threshold (parameters a_1, a_2, etc.) with respect to a loss function (e.g. sum of residual squares). The predicted value is obtained by a simple model (e.g. the average target value of all points in the selected feature region or the majority vote of class assignments).

In order to prevent over-fitting, there are various methods available for decision trees. One of the most popular is to create sub-samples of the training data (e.g. via bootstrapping) and fit a tree on each of the subsets using a random sub-choice of the features. The final model can then use the majority vote or any other combined statistic from the ensemble of trees, which is also called *random forests*.[102]

Random forests and decision tree have been successfully used for combinatorial materials screening[103] and are becom-

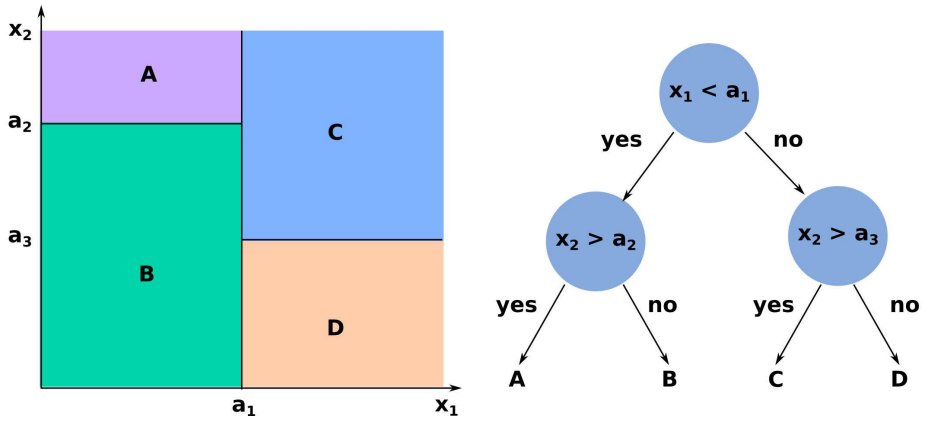


Figure 6. A decision tree: The feature space is split on a feature at a time (either x_1 or x_2) with the parameter (e.g. a_1) that minimizes the loss function at this particular split point.

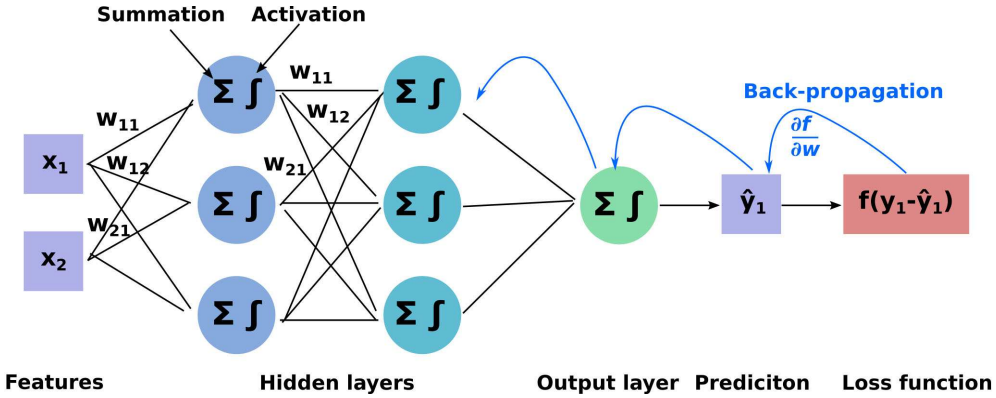


Figure 7. A neural network: The features are weighted by a weight matrix of dimensions $f \times n$ where f = number of features and n = number of nodes, where each node is an artificial neuron. The weighted features are summed and the result is used in the activation function. The output layer is the last layer of nodes and can output one or more predicted values.

ing more popular in computational catalysis: In a recent study, decision tress were applied to determine the empirical rules and conditions that lead to high catalytic performance (high CO conversion) of the water-gas shift reaction.[104] Janet *et al.*[105] used random forests to perform feature selection in order to establish structure-property relationships of molecular transition metal catalysts.

Decision trees are easy to understand and interpret, well-suited for visualization and closely mimics the human decision-making process. Decision trees process numerical and categorical features, which results in reduction of necessary pre-processing. Decision trees are non-parametric models, so no assumptions about the shape of data are made. They are fast for inference and feature selection is accounted for automatically, as unimportant features will not influence the result. Multi-collinearity also does not affect the results. Without proper regularization, decision trees tend to overfit and the greedy mechanism does not try different sequences of the feature. Furthermore, additional data can change the result and the model is not easily adapted to new data.

2.5. Neural Networks

Another very powerful and unique model type are neural networks (NNs), loosely inspired by the structure of biological brain networks. In a NN model, the input is passed to a collection of nodes (artificial neurons, more specifically so-called perceptrons), which are connected through edges with some defined weight. The node layers can be seen as distinct individual layers (hidden layers), which can have specialized functionality. The basic working principle of NNs is illustrated in Figure 7. Each input feature has a set of weights for each node in the first hidden layer. These weight vectors make up the first weight matrix of the neural network. In each node, the weighted features are summed and the sum is inserted to a non-linear activation function, such as a sigmoid function. There are many different activation functions, of which the most used is the sigmoid and the rectified linear unit (ReLU) function as it is less sensitive to bad weight initialization. The process is repeated within the number of hidden layers. The last layer is the output node. It can output one or several values, which can be used for regression and classification problems. The neural

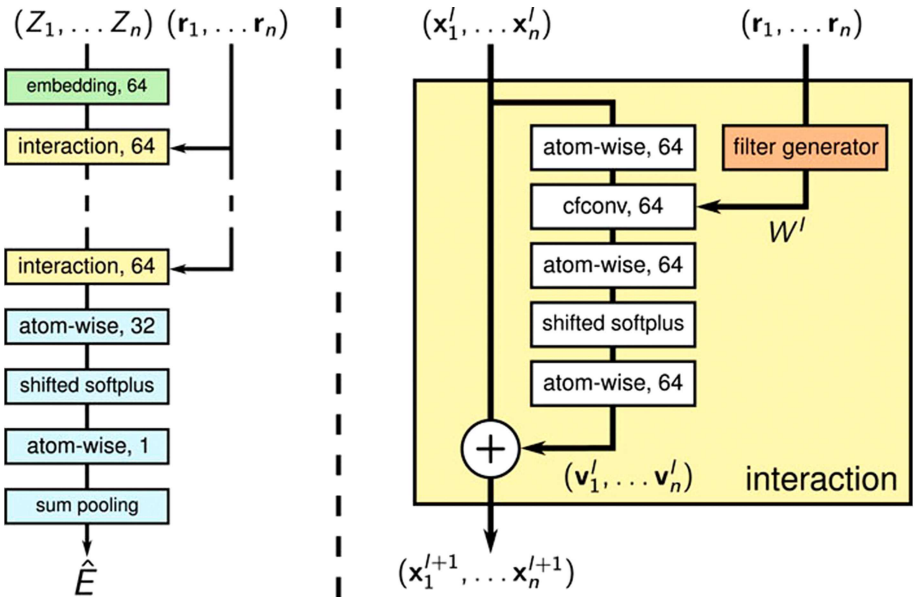


Figure 8. Left: schematic view of the SchNet architecture. Right: Interaction block architecture. Reproduced with permission from Ref. [108]. Copyright 2018 AIP Publishing.

network is trained by a process called back-propagation. In short, the chain rule for deriving the individual outputs with respect to the weights is applied, Figure 7. Neural networks have been used as surrogate models in place of computationally expensive first principle calculations to derive the total energy of the system. We address this special application in more detail in section 3.2.

Unfortunately, the interpretablity of the resulting weight matrices and node outputs is limited. However, this is an important aspect in mechanistic and causal studies, such as they arise in materials science and catalysis as we aim to understand the relation between features (materials properties) and outcomes (other materials properties, turn over frequency etc.). The problem can be addressed by performing a sensitivity analysis. That is, for instance randomizing one of the features and in how far this affects the predictive power. Fernandez *et al.* have used neural networks and decision trees to establish structure-property relations using a set of theoretical Pt nanoparticles.[106] In order to understand the relevance of the different features, they trained the neural networks with two features each and examined those with good performance (measured by coefficient of determination R^2) closer.

2.6. Deep Learning

An important sub-field of machine learning is deep learning. The quintessential deep learning models are deep feedforward neural networks, which employ multiple layers of artificial neurons. These models are called feedforward, because the information flows through the layers of neurons without feedback connections. The inclusion of feedback connections creates recurrent neural networks. Deep learning can be seen as an extreme case of model stacking. In addition to the basic feedforward hidden layer architecture, we can add layers with diverse functionality, such as pooling layers and normalization layers used in convolutional neural networks. Thereby, earlier layers in deep (convolutional) networks can be interpreted as automatic feature extractors. Deep neural networks in particular have been seen as black box models due to the lack of weight interpretability. However, progress has been achieved on analyzing trained models, as well as designing more interpretable architectures: For instance Schütt et al. developed a deep tensor neural network (DTNN) that enables spatially and chemically resolved insights into quantum-mechanical observables of molecular systems.[107]

In the DTNN framework, interactions are modeled by tensor layers, in which atom representations and interatomic distances are combined using a parameter tensor. The DTNN has been further improved to develop the SchNet network,[108] whose functionality is shown in Figure 8. The atoms with atomic number Z_i and positions r_i are attributed feature tuples to x_i. The network is used to compute potential energy surfaces. More examples on how machine learning is used to compute potential energy surfaces are given in section 3.2.

Neural networks and deep networks are very powerful and account for non-linear feature interactions. The performance of NN typically keeps increasing with the amount of training data in situations where other algorithms already reach an asymptotic performance level. Depending on the architecture, neural networks can be computationally quite efficient. Drawbacks are the (possible) complexity of architectures and their implementation. Again depending on the architecture and the problem at hand, neural networks may need a lot of training data to give good results. Furthermore, neural networks lack intuitive interpretability, although analysis methods are improving.

3. Recent Trends

In this section, we summarize recent developments in the field. We can roughly distinguish two main trends how machine learning has been used for computational heterogeneous catalysis research: (1) Fast combinatorial searches in the vast space of solid bulk materials based on descriptors and(2) using machine-learning to accelerate traditional optimization algorithms, e.g. atomic structure relaxation and transition state search.

3.1. Materials Screening

Tools and data from materials informatics can be applied to screen the materials space in terms of structure and composition of bulk (and other) materials to understand and predict active solid-state catalysts, an approach commonly termed catalysis informatics.[36,109] As described before, the important figures of merit are stability, activity and selectivity. In the following, we will outline how machine learning has been applied to predict and understand these entities.

3.1.1. Materials Stability

In general, the stability of a compound is highly dependent on the physical conditions, such as the temperature and pressure. In the case of electro-chemical systems, the applied potential and the electrolyte pH have a significant impact on the stability of the bulk as well as surface structures. Just as phase diagrams maps the thermodynamic equilibrium of temperatures vs. pressures, Pourbaix diagrams map out the stable phases as a function of electro-chemical potentials and electrolyte pH. The conventional approach of generating these diagrams is not trivial as it usually requires meticulous and costly experimentation even for simple binary or ternary systems. For a multi-component system beyond ternary, this approach becomes unfeasible.

To overcome the failure of the conventional strategy, computer-aided approaches have been employed. Phase diagrams construction is on the basis of phenomenological modeling (Calphad methodology[110]) which extrapolates thermo-chemical properties of multi-component systems from those of lower order systems.[111] A Pourbaix diagram, on the other hand, is generated by defining the boundaries of convex hulls via a comparison of the formation energies for all possible phases within the ranges of electrolyte pH and potentials. There are two crucial requirements leading to the success of computing phase diagrams or Pourbaix diagrams. The first is realistic Gibbs energies for all the possible phases and the second is the availability of efficient algorithms to define the equilibrium. While the latter has been addressed by various commercial or in-house software packages, including Thermo-Calc,[112] PandatTM,[113] MTDATA[114] and FactSage[115] for phase diagrams, and Pymatgen[116] and ASE[39] for Pourbaix diagrams, the former remains challenging. The challenge of obtaining Gibbs energies remains in two-fold. The first fold is that exhaustively searching and enumerating all the possible phases in the system and the second fold is data sources of collecting Gibbs energies. We only focus on data sources in this section, and the optimal exhaustive search and enumeration strategy will be discussed under section 3.2.

A conventional way of obtaining Gibbs energies is either from experiments or accurate theoretical calculations. However, Gibbs energies are available only for a limited number of known inorganic compounds, which becomes a bottleneck of the comprehensive material stability screening for a large variety of the systems. Recently, machine learning techniques and neural networks begin to draw materials scientists' attention and have been employed to predict materials chemical properties including Gibbs energies. Some researches focus on featurization: Bartel *et al.*[117] used SISSO (sure independence screening and sparsifying operator)-based fingerprints to predict temperature-dependent Gibbs energies for inorganic crystalline solids with a decent accuracy of 50 meV/atom, along with thousands of temperature-dependent phase diagrams generated. Ward *et al.*[64] combined composition-based descriptors with Voronoi tessellation derived attributes, employed OQMD as training data and decision tree model to map formation enthalpies with mean absolute error of 80 meV/atom. Takahashi *et al.*[118] utilized the Gaussian mixture model and random forest classification to reveal the important descriptors that determine the crystal structure of single and binary compounds in OQMD, based on descriptor importance rankings. Almost at the same time, Seko *et al.*[119] proposed to use descriptive statistics as descriptorsto represent a compound for predicting cohesive energies. The descriptors were generated from atomic and structural information. Some researches focus on machine learning algorithms: Schmidt *et al.*[120] benchmarked the performances of various machine learning methods (including ridge regression, random forests, extremely randomized trees and neural networks) in regards of their abilities of predicting thermodynamic stability of solids. Within the algorithms they tested, extremely randomized trees provided the best performance. Most recent research[121] focuses on graph convolutional neural networks, instead of constructing vector matrix to represent the structures, graphs of atoms connection are used to train the multi-layer neural network to predict the formation energies with a reasonable mean absolute error of 130 meV/atom. Other researches took both featurization and algorithms implementation into consideration: Gossett *et al.*[122] built a machine learning workflow/pipeline (AFLOW-ML) using raw materials data as input and electronic, thermal and mechanical properties as outputs. Ward *et al.*[123] developed a python-based platform (matminer) which streamlines materials data collection and data mining, and Picklum *et al.* have recenlty presented the MatCALO platform,[124] for providing a machine learning predicting for the composition of materials that satisfies additional criteria such as specified ranges of hardness and deformation. All above efforts contribute to more accurate Gibbs energies predictions, which leads to more rational Pourbaix diagrams. One machine learning application on constructing surface Pourbaix diagrams has been demonstrated by Ulissi,[101] in

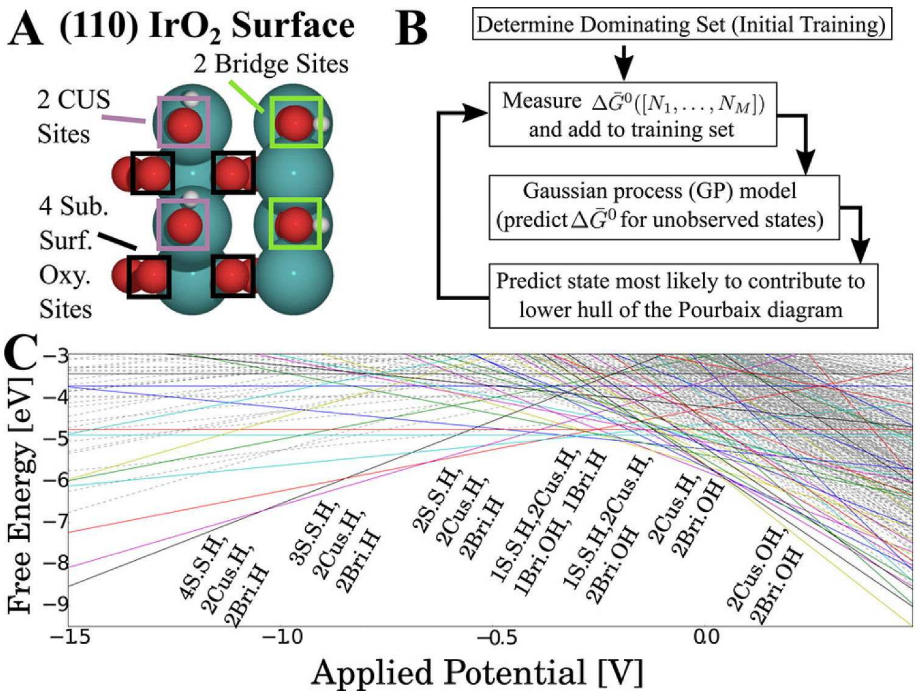


Figure 9. ML aided approach of constructing Pourbaix diagram for an IrO_2 surface. (A) three adsorption sites considered on a 2×2 IrO_2(110) slab: CUS (Coordination unsaturated sites), brige sites and subsurface oxygen sites. (B) A flowchart of adopting supervised Gaussian process model for Pourbaix diagram construction. (C) A generated Pourbaix diagram with the lower hulls labeled. Reprinted with permission from Ref. [101]. Copyright 2016 American Chemical Society.

which a Gaussian process regression model was trained to predict coverage-dependent free energies of surfaces and Pourbaix diagrams for the IrO_2(110) and MoS_2 surfaces are generated with a significantly reduced complexity. Their machine learning aided approach of producing Pourbaix diagram for an IrO_2 surface is shown in Figure 9.

3.1.2. Catalytic Activity and Selectivity

The recent transition from Edisonian to knowledge-based design of catalysts entails identifying correlations between descriptors and catalytic performance metrics like activity and selectivity. These descriptors can be integrated as features in the machine-learning context. This integration helps to speed up the selection process of catalyst candidates warranting experimental study.

Since obtaining rates and selectivities for every possible catalyst is both experimentally and computationally prohibitively expensive, the energy space of intermediates is usually parametrized using descriptors. A prominent example for such activity- and selectivity-determining descriptors are adsorption energies of catalytic intermediates that can be rather accurately determined using density functional theory.[48] Volcano plots based on scaling[47] and BEP[45,46] relations of these descriptors have helped to shed light on favorable adsorption properties for a given reaction.[42] Nevertheless, screening the vast materials space of multimetallic alloys and oxides even only for these selected properties still remains a formidable task. In recent years, the machine-learning community has tackled this challenge by incorporating their tools into workflows aiming to reduce the number of expensive electronic structure calculations and by developing predictive adsorption energy models. Since these predicted adsorption energy values are directly mapped onto rates and selectivities, it is crucial to accurately predict these descriptors.

Automated structure generation presents a first step toward accelerated screening of catalyst candidates. Montoya and Persson[125] set out along this path by automatically constructing distinct adsorbate geometries enabling high-throughput workflows to perform DFT calculations of adsorption energies. Recently, Boes *et al.*[126] further elaborated in this direction and reported on a graph-theory-based approach to enumerate surfaces and unique chemical adsorption structures. Boes *et al.* created unique, systematic surface representations for any possible adsorption structure and generated three dimensional initial guesses for these structures.

Predictive models of adsorption energies of common catalytic descriptors like OH*, CO*, or NO* on multimetallic surfaces have flourished in the past years. These models can be roughly assigned to physics-based[72,73,127–129] and data-driven[57,59,69,130–138] approaches. While the former approaches start their predictions from pre-determined physical models, like the *d*-band theory,[53] bond order conservation,[139] or electrostatic interactions,[140] the latter use large sets of materials properties and try to identify and use trends that might not have a known analytical expression.[24] In the following, we will provide an overview of the data-driven machine-learning efforts

in this field with quantitative evaluations of how accurately adsorption energies are being predicted.

Several regression schemes have been used to predict adsorption energies on alloys. Jinnouchi *et al.*,[134,135] for example, employed a Bayesian linear regression scheme to predict adsorption energies of the key intermediates of *NO* decomposition of *RhAu* nanoparticles based on DFT data of single crystalline slab surfaces with MAEs of 0.1 eV. Building on the fact that sites with similar atomic configurations are expected to have similar activities, Jinnouchi *et al.* employed the smooth overlap atomic position (SOAP)[141] similarity kernel as a descriptor. They use these predictions to understand surface segregation, size and composition dependences of catalytic rates and obtained detailed structures of active sites on *RhAu* nanoparticles during *NO* decomposition. Toyao *et al.*,[136] in turn, found that extra tree regression combined with twelve database-available descriptors is able to predict adsorption energies of methane related intermediates on Cu-based alloys with a RMSE of 0.26 eV. They use their findings to identify dopants that stabilize CH_3 compared to CH_2 to optimize the efficiency of methane utilization for methanol synthesis and oxidative coupling of methane (OCM). Finally, Noh *et al.*[57] combined kernel ridge regression with an active learning algorithm. The authors chose the linear muffin-tin orbital theory (LMTO)-based *d*-band width and the geometric mean of electronegativity, hence simple non-first-principle input features, as descriptors. Thereby, they obtain CO* adsorption energies with an RMSE of 0.05 eV and identify $Cu_3Y@Cu^*$ as effective electrochemical CO_2 to *CO* reduction catalyst with low overpotential.

Other approaches have focused on employing neural networks. Ma *et al.*[130] and Li *et al.*[131] trained artificial neural networks with sets of DFT-calculated adsorption energies and electronic structure fingerprints like e.g. moments of the *d*-band distribution. They used these neural networks to predict adsorption energies of catalytic intermediates on multimetallic surfaces within RMSEs of 0.1 eV–0.2 eV. Ma *et al.* show how this approach can be combined with scaling relations to enable high-throughput catalyst screening. Using this approach, they were able to suggest multimetallic alloys with improved activity and selectivity towards CO_2 electro-reduction. Ulissi *et al.*[59] and Tran *et al.*[133] used neural network potentials to predict adsorption energies on intermetallic compounds. By combining active machine learning[142] with surrogate-based optimization,[143] their models teach themselves, predict facets containing promising active site motifs and suggest motifs that should be investigated next, i.e. the models guide the DFT calculations and constantly update themselves as new calculated data becomes available. Using this approach, Tran *et al.*[133] screened 1499 intermetallics for selectivity in the CO_2 reduction and hydrogen evolution reactions. Within this space, they identified 131 candidate surfaces of 54 intermetallics to be potentially active in CO_2 reduction. A number of these candidates had already been validated by experiment before,[133] which encourages experimental investigation of the other proposed candidates.

Finally, researchers have sought to expand theories of bonding by extracting analytical models from data-driven frameworks. Andersen *et al.*[138] employed compressed sensing, namely SISSO (sure independence screening and sparsifying operator), to predict adsorption energies of atomic and molecular adsorbates on mono- and multimetallic transition metal surfaces. The descriptors found with this method are expressed as nonlinear functions of clean catalyst surface properties like coordination numbers, *d*-band-moments, and the density of states at the Fermi level, which can be obtained from a single DFT calculation. This method predicts adsorption energies within RMSEs of 0.2 eV. Figure 10 summarizes how machine learning enhanced adsorption energy predictions can speed up the selection of catalyst candidates to be tested in experiment.

Machine learning for computational heterogeneous catalysis has not only helped to develop predictive models for adsorption energies but has also revealed other physical quantities that determine the activity and selectivity of catalysts.

Hong *et al.*,[92] for example, used LASSO regularization for automatic feature selection from aset of 14 descriptors derived from electronic structure calculations. By relating these descriptors to experimental oxygen evolution activities of transition-metal oxide catalysts, the authors unveiled the importance of the transition metal e_g/t_{2g} electron occupancy in determining the catalytic activity. Wexler *et al.*[144] used a regularized random forest machine-learning algorithm to investigate the relative importance that structure and charge descriptors have on the HER activity of $Ni_2P(0001)$. Thereby, the authors identify the Ni–Ni bond length, which can be tuned through dopants, as the most important activity descriptor.

The above mentioned examples show that, while still being in its infancy, machine-learning enhanced activity and selectivity screening based on catalytic descriptors has great potential to efficiently identify suitable catalyst candidates to be tested by experiment within the vast materials space of multimetallic alloys and compounds.

3.2. Structure Optimization and Transition-State Search

In order to obtain an relevant reaction mechanisms and kinetics, two crucial quantities of the potential energy surfaces (PES) have to be determined: (1) minimum energy structures corresponding to global or local minima on the PES, i.e. initial and inal states, and (2) the saddle points connecting these minima, i.e. transition states. Computing energy and forces via first-principles is computationally demanding and shows unfortunate scaling with the system size (number of electrons), as well as the complexity of the system. Finding and utilizing mathematical methods (e.g. computer algorithms) to reduce the number of first-principle calculations involved in the optimization of atomic structures is therefore highly beneficial for computational heterogeneous catalysis research.

Over the past two decades, the development of efficient algorithms to optimize atomic structures has been extensively investigated.[145–148] Traditional optimization methods (e.g. gradient-based, Newtonian dynamics) often require calculating the

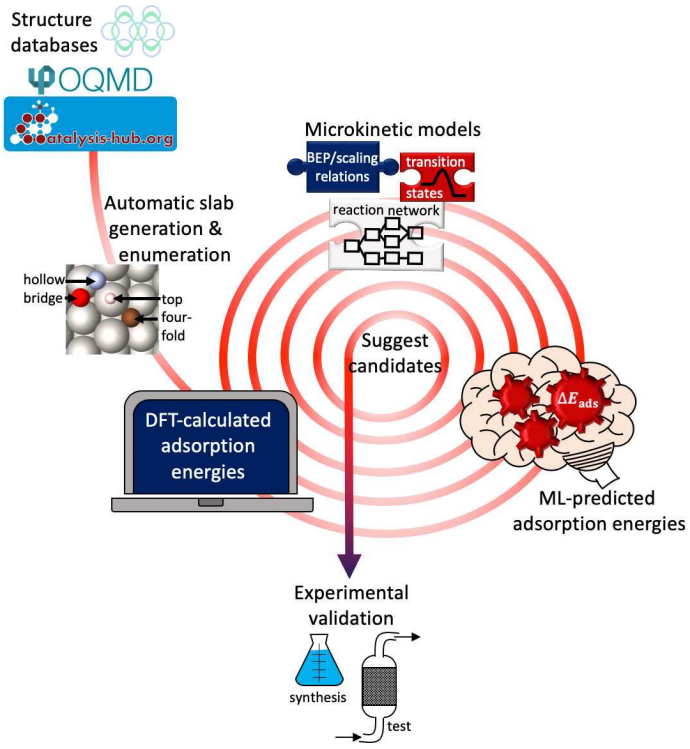


Figure 10. Machine learning-enhanced catalyst candidate prediction: Bulk and surface structures retrieved from structure databases like materialsproject.org, OQMD, catalysis-hub.org, etc. are used for automated slab generation and enumeration of possible adsorption sites. In an iterative process, limited numbers of DFT-calculated adsorption energies and machine-learning-predicted adsorption energies are used to inform microkinetic models to eventually suggest promising catalyst candidates that should be investigated by experiment.

energy and/or forces of the system at each optimization step. Therefore, machine learning (ML) surrogate models have been recently proposed as promising alternatives.[149–153] The surrogate ML model should provide an approximation of the "true" first-principle PES using only a few first principle calculations.

The use of surrogate machine learning models has been successfully used to accelerate the exploration of energy landscapes of atomic structures by means of local and global structure optimization.[149–153] Hammer and coworkers have shown that the use of supervised learning can be exploited to enhance global optimization algorithms of molecular structures,[150] highlighting the importance of selecting an adequate acquisition function to optimally navigate through high-dimensional systems. Exploring the vast energy landscape of complex hyper-surfaces is prohibitively expensive from a computational point of view. Thus, the ideal acquisition function must be formulated to intrinsically contain a balance between exploration (acquiring data of regions with high uncertainty) and exploitation (acquiring data on regions with high-probability of finding low-energy structures). Machine learning models have also been combined with gradient-based optimizers to accelerate energy minimization of atomic structures. Garijo del Río *et al.* have shown how a gradient-based local optimizer, such as the Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm[154–157] can be used to minimize the predicted energy from the ML models.[152] This gradient-based optimizer is ultimately dictating the form of the acquisition function in charge of exploring the potential energy landscape towards the nearest local minimum.

The use of surrogate ML models has served to accelerate the search of transition states by means of decreasing the number of function evaluations required to converge Nudged Elastic Band (NEB) calculations. The original NEB method[158,159] has been extensively used in the field of materials science for calculating reaction barriers when the initial and final states for that given transition are known. One of the major drawbacks of this method is that an accurate description of the path ultimately relies on including an adequate number of images describing the transition from reactants to products. The high computational cost of optimizing those images towards a converged minimum energy path (MEP) has made NEB one of the most controversial methods in computational heterogeneous catalysis, since it is a robust method to find saddle points but it is also computationally very expensive.

Recently, Peterson and coworkers have proposed a ML-based approach using convoluted Neural Networks (NN)[160] to alleviate the expensive cost of optimizing NEB calculations. Figure 11a and b show a comparison between the original NEB method[158] and the NN assisted algorithm, respectively, on the diffusion of a Au atom across an Al surface in which some Al atoms are substituted with Pt atoms. The original NEB method evaluates all the images along the path in each iteration, i.e. when the images of the band are moved towards obtaining a converged MEP (represented by the red line and circles in

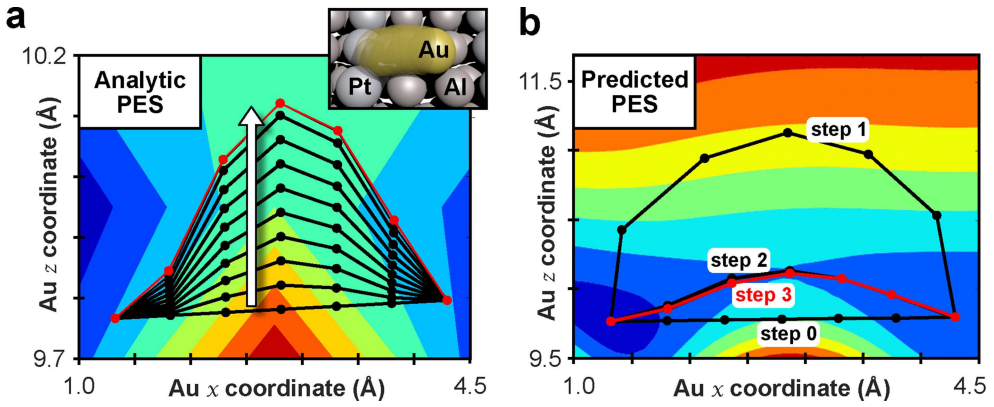


Figure 11. Comparison between the (a) original NEB and (b) Neural Network assisted surrogate NEB[160] methods for the diffusion of a Au atom on a fcc(100) Al surface with Pt alloyed. The black circles represent the first-principle calculations whilst the red circles show the converged Minimum Energy Pathways (MEP). A white arrow is superimposed in (a) to highlight the NEBoptimization sequence towards finding a converged MEP. This sequence is highlighted in (b) by the "step" labels, from the initial interpolation (step 0) to the converged MEP (step 3). The dark blue color on the PES represents low energies whilst the red color represents high energies. Note that the Au "z" coordinate and color-scale selected for the analytic and predicted PES are different, however, the two methods provide the same converged MEP. Reproduced with permission from Ref. [160]. Copyright 2016 AIP Publishing.

Figure 11a). The NN approach relies on training a model with the experiences (first-principle calculations, see black circles in step 0" in Figure 11b) to obtain a predicted PES that can be optimized with almost no extra cost. The result of this optimization is then calculated (see black circles in step 1" in Figure 11b) and the model is retrained with all the experiences collected. The algorithm keeps collecting data, retraining the model, improving the predictions and optimizing NEB calculations on the predicted PES until the first-principle data of the predicted path is analytically converged (see red line and circles in step 3" in Figure 11b). From the comparison between the original and surrogate NN methods in Figure 11 it is evident that the performance of the classical NEB method is surpassed by the NN approach. This improvement can be determined by counting the number of first-principle function evaluations (black circles in Figure 11a,b) to obtain the converged MEP.

ML-assisted NEB algorithms using Gaussian Process Regression (GPR) were recently introduced by Jónsson and coworkers.[161] The main advantage of using GPR relies on the fact that the model can capture the uncertainty estimates of the process. The authors presented a GPR-assisted NEB algorithm that evaluates the geometry of the image presenting the highest uncertainty of the optimized predicted path in each iteration of the surrogate model. The algorithm updates the GP each time that a new structure is evaluated, this procedure is known as one-image-evaluation (OIE) method. As proposed by Garrido Torres *et al.*, the uncertainty estimates can also be combined with the results obtained from the *ab initio* calculations to define convergence criteria.[162] This work also highlights the importance of using an adequate acquisition function in order to achieve a predicted MEP that mimics the analytic MEP. This agreement has been validated by performing a benchmark using a number of acquisition functions with different weights on their exploration and exploitation components, e.g. evaluating the NEB images with high uncertainty but also targeting the images with highest predicted energy along the NEB (with high probability to find a saddle point). This approach has served to reduce the number of function evaluations required to obtain an accurate description of the MEP (see Figure 12). In this case, the authors have used the uncertainty estimates and analytic forces to decouple the computational cost of the NEB calculations (in terms of function evaluations) from the number of images describing the NEB, which, for decades, has been one of the major drawbacks of using the NEB-based approach.

NEB calculations are performed on the known reaction pathways that are assumed to be important. A more systematic way of detecting unknown reaction pathways and determining reaction mechanisms is to automate exhaustive search. Several search algorithms are available, including single-ended growing string method (SSM),[163] growing string method (GSM),[164] freezing string method (FSM),[165] KinBot[166] and single-component artificial force induced reaction method (SC-AFIR).[167] Maeda *et al.*[167] benchmarked the performances of these five search algorithms and they found SC-AFIR can detect more reaction channels due to its high versatility. The automated reaction path search method should be used with a caution as selecting wrong path among tremendous options leads to false conclusions. ML techniques can aid to increase the likelihood of selecting the correct pathways. Selective related work includes Coley *et al.*[168] employing neural networks to recognize the patterns on reaction paths from literatures and to target the best candidates of the major products through candidate ranking, and Ulissi *et al.*[59] in our group presenting a machine learning framework of optimization under uncertainties, where a surrogate model was iterated to predict the most important reaction steps. The combination of DFT calculations/experimental results with machine learning aided search algorithms can significantly accelerate the rate of completing complicated reaction maps of reaction of interests.

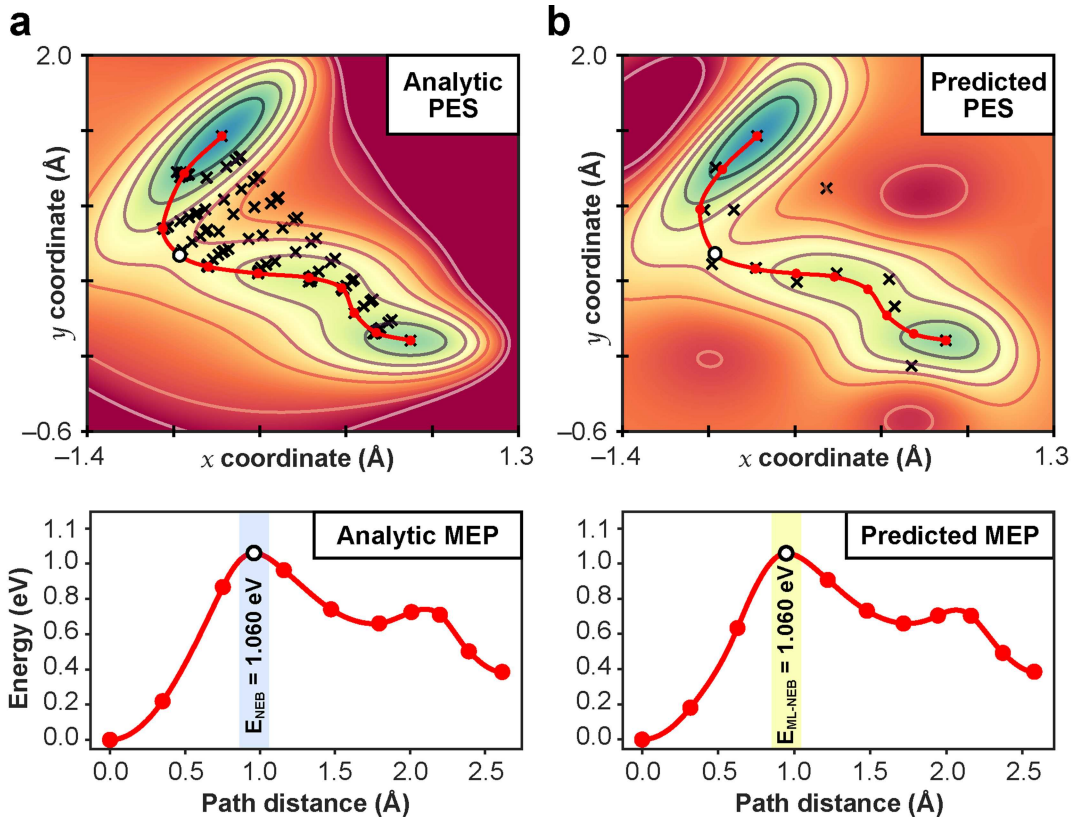


Figure 12. Comparison between the (a) original Nudged Elastic Band (NEB)[158] and (b) Gaussian Process Regression (GPR)-assisted surrogate NEB[161,162] methods of the two-dimensional Müller-Brown PES. The black crosses represent the first-principle calculations whilst the red circles show the converged Minimum Energy Pathways (MEP). The saddle points are highlighted by the white circles. The blue colors on the PES represents low energies whilst the red color represents high energies. Reproduced with permission from Ref. [160]. Copyright 2019 APS.

4. Future Directions

4.1. Software Tools and Workflows

Advances in machine learning approaches for theoretical surface science and catalysis (and material science in general), relies deeply on the development of sophisticated software tools for managing the generation and storage of data as well as the integration with feature design and model generation. In particular in the framework of active learning, atomic structure and feature generation, model prediction and candidate acquisition must be integrated with first-principle job submissions and evaluation in order to obtain an efficient feedback loop for automated predictions.

Recent advances in this field include the CatLearn Python module[169] which allows to generate descriptors from atomic structures of surfaces and nanoparticles, automates descriptor selection, and offers active learning functions. This package also includes some of the active-learning routines introduced above, such as the structural optimization (ML-Min) and transition-state search (ML-NEB[162]) using GPR.

Also, the recently developed CatKit Python module,[126] enables automatic enumeration of all possible adsorption sites on a surface structure, by using a unique graph representation of a provided bulk structure and any slab which can be produced from that bulk. Using such a catalog of possible structures is necessary for an automated informatics approach to catalytic systems. For instance, Noh *et al.* used active learning to accelerate their machine learning workflow and to enhance the predictive accuracy of CO adsorption energy on binary alloy surfaces by more than 0.1 eV.[57]

Development and utilization of database resources for electronic structure calculations is another prerequisite for efficient model generation. In connection to active learning workflows, outputs of calculations must be stored in a structured manner that allows for on-the-fly model generation as well as later retrieval. Also, with the continuous expansion of open electronic structure databases, the amount of data available for model generation will increase significantly, possibly enabling different approaches.

4.2. Advances in Model Complexity

Future generation and material models should aim to bridge the gap between simplified descriptor-based bulk materials screening and detailed mechanistic/kinetic studies by developing meaningful surface descriptors of a variety of materials types beyond transition metals, as for instance metal oxides, functionalized by intrinsic and/or extrinsic defects. While

impressive advances in predicting bulk oxide properties via machine learning techniques have already been realized,[60,170] accurate activity trends are still subject to current research. Because of the inherent complexity of oxide surfaces, screening studies of this type focus on predicting or using bulk properties, like the $e_g/t_{2t\,g}$ occupancy, to understand activity trends.[92,171,172] This approach works especially well if the properties of the surface sites correlate smoothly with the bulk[173] and also if the intermediates adsorb on the same surface site, and lastly if the BEP scaling relations are closely satisfied. However, especially the second condition constitutes a problem, as intermediate scaling relations limit the catalyst activity and product selectivity.[174] A strategy to solve this problem is the diversification of the active centers of different elementary reactions and adsorbate-dependent cooperative effects.[81,175,176] Site diversification can be achieved by introducing dual surface sites. The open question remains if machine learning algorithms can contribute to discovery of functional dual-site catalysts in the future.

5. Summary

In this review article, we introduced common and cutting edge machine learning and data science topics and their application to computational heterogeneous catalysis. We discussed the importance of datasharing and the computational catalysis-related challenges to systematically acquire, store and share data. We outlined the working principle of popular machine learning algorithms and how they are applied to computational heterogeneous catalysis problems. Two important directions of how machine learning is currently utilized have been identified: (1) The combinatorial search for catalyst materials and (2) the acceleration of potential energy exploration.

Acknowledgements

This work was supported by the U.S. Department of Energy, Chemical Sciences, Geosciences, and Biosciences (CSGB) Division of the Office of Basic Energy Sciences, via Grant DE-AC02-76SF00515 to the SUNCAT Center for Interface Science and Catalysis. PS and VS gratefully acknowledge the Alexander von Humboldt Foundation (AvH) for financial support.

Conflict of Interest

The authors declare no conflict of interest.

Keywords: Computational chemistry · density functional theory · heterogeneous catalysis · machine learning · materials science

[1] Z. W. Seh, J. Kibsgaard, C. F. Dickens, I. Chorkendorff, J. K. Nørskov, T. F. Jaramillo, *Science* **2017**, *355*, 4998.
[2] A. J. Medford, M. R. Kunz, S. M. Ewing, T. Borders, R. Fushimi, *ACS Catal.* **2018**, *8*, 7403–7429.
[3] B. R. Kowalski, *J. Chem. Inf. Comput. Sci.* **1975**, *15*, 201–203.
[4] K. T. Butler, D. W. Davies, H. Cartwright, O. Isayev, A. Walsh, *Nature* **2018**, *559*, 547.
[5] K. Rajan, *Mater. Today* **2005**, *8*, 38–45.
[6] S. R. Kalidindi, M. De Graef, *Annu. Rev. Mater. Res.* **2015**, *45*, 171–193.
[7] A. Jain, G. Hautier, S. P. Ong, K. Persson, *J. Mater. Res.* **2016**, *31*, 977–994.
[8] G. A. Landrum, J. E. Penzotti, S. Putta, *Meas. Sci. Technol.* **2004**, *16*, 270.
[9] N. Fey, *Chem. Cent. J.* **2015**, *9*, 38.
[10] J. Behler, *Phys. Chem. Chem. Phys.* **2011**, *13*, 17930–17955.
[11] V. Botu, R. Batra, J. Chapman, R. Ramprasad, *J. Phys. Chem. C* **2016**, *121*, 511–522.
[12] J. E. Herr, K. Yao, R. McIntyre, D. W. Toth, J. Parkhill, *J. Chem. Phys.* **2018**, *148*, 241710.
[13] R. Moros, H. Kalies, H. Rex, S. Schaffarczyk, *Comput. Chem. Eng.* **1996**, *20*, 1257–1270.
[14] S. R. Upreti, K. Deb, *Comput. Chem. Eng.* **1997**, *21*, 87–92.
[15] K. Reuter, H. Metiu, *Handb. Mater. Model.* **2018**, 1–11.
[16] K. Winther, M. J. Hoffmann, O. Mamun, J. R. Boes, J. K. Nørskov, M. Bajdich, T. Bligaard, *Sci. Data,* **2019**, *6,* **75**.
[17] S. R. Kalidindi, A. J. Medford, D. L. McDowell, *JOM* **2016**, *68*, 2126–2137.
[18] A. Jain, K. A. Persson, G. Ceder, *APL Mater.* **2016**, *4*, 053102.
[19] M. De Jong, W. Chen, T. Angsten, A. Jain, R. Notestine, A. Gamst, M. Sluiter, C. K. Ande, S. Van Der Zwaag, J. J. Plata„ C. Toher, S. Curtarolo, G. Ceder, K. A. Persson, M. Asta, *Sci. Data* **2015**, *2*, 150009.
[20] A. Zakutayev, N. Wunder, M. Schwarting, J. D. Perkins, R. White, K. Munch, W. Tumas, C. Phillips, *Sci. Data* **2018**, *5*, 180053.
[21] L. Lin, *Mater. Perform. Charact.* **2015**, *4*, 148–169.
[22] A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, K. A. Persson, *APL Mater.* **2013**, *1*, 011002.
[23] G. Pizzi, A. Cepellotti, R. Sabatini, N. Marzari, B. Kozinsky, *Comput. Mater. Sci.* **2016**, *111*, 218–230.
[24] C. Draxl, M. Scheffler, *MRS Bull.* **2018**, *43*, 676–682.
[25] J. E. Saal, S. Kirklin, M. Aykol, B. Meredig, C. Wolverton, *JOM* **2013**, *65*, 1501–1509.
[26] S. Curtarolo, W. Setyawan, S. Wang, J. Xue, K. Yang, R. H. Taylor, L. J. Nelson, G. L. Hart, S. Sanvito, M. Buongiorno-Nardelli, N. Mingo, O. Levyh, *Comput. Mater. Sci.* **2012**, *58*, 227–235.
[27] D. D. Landis, J. S. Hummelshoj, S. Nestorov, J. Greeley, M. Dulak, T. Bligaard, J. K. Norskov, K. W. Jacobsen, *Comput. Sci. Eng.* **2012**, *14*, 51–57.
[28] A. Togo, *Phonon database at Kyoto university*, http://phonondb.mtl.kyoto-u. ac.jp.
[29] S. Haastrup, M. Strange, M. Pandey, T. Deilmann, P. S. Schmidt, N. F. Hinsche, N. Gjerding, D. Torelli, P. M. Larsen, A. C. Riis-Jensen, J. Gath, K. W. Jacobsen, J. J. Mortensen, T. Olsen, K. S. Thygesen, *2D Mater.* **2018**, *5*, 042002.
[30] F. H. Allen, *Acta Crystallogr. Sect. B* **2002**, *58*, 380–388.
[31] A. Belsky, M. Hellenbrandt, V. L. Karen, P. Luksch, *Acta Crystallogr. Sect. B* **2002**, *58*, 364–369.
[32] S. Gražulis, A. Daškevič, A. Merkys, D. Chateigner, L. Lutterotti, M. Quiros, R. Serebryanaya, P. Moeck, R. T. Downs, A. Le Bail, *Nucleic Acids Res.* **2011**, *40*, D420–D427.
[33] Y. Xu, M. Yamazaki, P. Villars, *Jpn. J. Appl. Phys.* **2011**, *50*, 11RH02.
[34] J. S. Hummelshøj, F. Abild-Pedersen, F. Studt, T. Bligaard, J. K. Nørskov, *Angew. Chem. Int. Ed.* **2012**, *51*, 272–274; *Angew. Chem.* **2012**, *124*, 278–280.
[35] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, J. Bouwman, A. J. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C. T. Evelo, R. Finkers, A. Gonzalez-Beltran, A. J. Gray, P. Groth, C. Goble, J. S. Grethe, J. Heringa, P. A. t. Hoen, R. Hooft, T. Kuhn, R. Kok, J. Kok, S. J. Lusher, M. E. Martone, A. Mons, A. L. Packer, B. Persson, P. Rocca-Serra, M. Roos, R. van Schaik, S.-A. Sansone, E. Schultes, T. Sengstag, T. Slater, G. Strawn, M. A. Swertz, M. Thompson, J. van der Lei, E. van Mulligen, J. Velterop, A. Waagmeester, P. Wittenburg, K. Wolstencroft, J. Zhao, B. Mons, *Sci. Data* **2016**, *3*, 160018.
[36] K. Takahashi, L. Takahashi, I. Miyazato, J. Fujima, Y. Tanaka, T. Uno, H. Satoh, K. Ohno, M. Nishida, K. Hirai, J. Ohyama, T. N. Nguyen, S. Nishimura, T. Taniike, *ChemCatChem* **2019**, *11*, 1146–1152.
[37] L. Takahashi, I. Miyazato, K. Takahashi, *J. Chem. Inf. Model.* **2018**, *58*, 1742–1754.

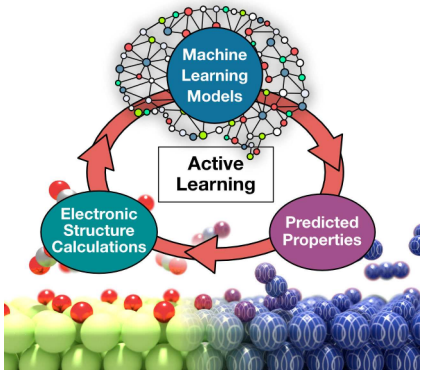
[38] G. Grethe, G. Blanke, H. Kraut, J. M. Goodman, *J. Cheminf.* **2018**, *10*, 22.
[39] A. H. Larsen, J. J. Mortensen, J. Blomqvist, I.E. Castelli, R. Christensen, M. Dułak, J. Friis, M. N. Groves, B. Hammer, C. Hargus, E. D. Hermes, P. C. Jennings, P. B. Jensen, J. Kermode, J. R. Kitchin, E. L. Kolsbjerg, J. Kubal, K. Kaasbjerg, S. Lysgaard, J. B. Maronsson, T. Maxson, T. Olsen, L. Pastewka, A. Peterson, C. Rostgaard, J. Schiøtz, O. Schütt, M. Strange, K. S. Thygesen, T. Vegge, L. Vilhelmsen, M. Walter, Z. Zeng, K. W. Jacobsen, *J. Phys. Condens. Matter* **2017**, *29*, 273002.
[40] A. Corma, J. Serra, P. Serna, M. Moliner, *J. Catal.* **2005**, *232*, 335–341.
[41] C. Klanner, D. Farrusseng, L. Baumes, M. Lengliz, C. Mirodatos, F. Schüth, *Angew. Chem.* **2004**, *116*, 5461–5463; *Angew. Chem. Int. Ed.* **2004**, *43*, 5347–5349.
[42] J. K. Nørskov, F. Studt, F. Abild-Pedersen, T. Bligaard, *Fundamental concepts in heterogeneous catalysis*, John Wiley & Sons, **2014**.
[43] B. R. Goldsmith, J. Esterhuizen, J.-X. Liu, C. J. Bartel, C. Sutton, *AIChE J.* **2018**, *64*, 2311–2323.
[44] P. Sabatier, *La catalyse en chimie organique.*, Libr. Polytech. Paris Liege, **1920**.
[45] J. Brönsted, K. Pedersen, *Z. Phys. Chem.* **1924**, *108*, 185–235.
[46] M. Evans, M. Polanyi, *Trans. Faraday Soc.* **1936**, *32*, 1333–1360.
[47] F. Abild-Pedersen, J. Greeley, F. Studt, J. Rossmeisl, T. Munter, P. G. Moses, E. Skulason, T. Bligaard, J. K. Nørskov, *Phys. Rev. Lett.* **2007**, *99*, 016105.
[48] J. K. Nørskov, T. Bligaard, B. Hvolbæk, F. Abild-Pedersen, I. Chorkendorff, C. H. Christensen, *Chem. Soc. Rev.* **2008**, *37*, 2163.
[49] J. Greeley, J. K. Nørskov, *J. Phys. Chem. C* **2009**, *113*, 4932–4939.
[50] A. J. Medford, A. Vojvodic, J. S. Hummelshøj, J. Voss, F. Abild-Pedersen, F. Studt, T. Bligaard, A. Nilsson, J. K. Nørskov, *J. Catal.* **2015**, *328*, 36–42.
[51] R. Parsons, *Trans. Faraday Soc.* **1958**, *54*, 1053–1063.
[52] F. Besenbacher, I. Chorkendorff, B. S. Clausen, B. Hammer, A. M. Molenbroek, J. K. Nørskov, I. Stensgaard, *Science* **1998**, *279*, 1913–1915.
[53] B. Hammer, J. Nørskov, *Surf. Sci.* **1995**, *343*, 211–220.
[54] I. Takigawa, K. ichi Shimizu, K. Tsuda, S. Takakusagi in *Nanoinformatics*, Springer Singapore, **2018**, pp. 45—64.
[55] C. F. Dickens, J. H. Montoya, A. R. Kulkarni, M. Bajdich, J. K. Nørskov, *Surf. Sci.* **2019**, *681*, 122–129.
[56] C. Dickens, A. Latimer, *CS229 Final Report: Learning Chemistry from Moment to Moment*, **2018**.
[57] J. Noh, S. Back, J. Kim, Y. Jung, *Chem. Sci.* **2018**, *9*, 5152–5159.
[58] I. Takigawa, K. ichi Shimizu, K. Tsuda, S. Takakusagi, *RSC Adv.* **2016**, *6*, 52587–52595.
[59] Z. W. Ulissi, M. T. Tang, J. Xiao, X. Liu, D. A. Torelli, M. Karamad, K. Cummins, C. Hahn, N. S. Lewis, T. F. Jaramillo, K. Chan, J. K. Nørskov, *ACS Catal.* **2017**, *7*, 6600–6608.
[60] T. Xie, J. C. Grossman, *Phys. Rev. Lett.* **2018**, *120*, year.
[61] F. Calle-Vallejo, J. Tymoczko, V. Colic, Q. H. Vu, M. D. Pohl, K. Morgenstern, D. Loffreda, P. Sautet, W. Schuhmann, A. S. Bandarenka, *Science* **2015**, *350*, 185–189.
[62] X. Ma, H. Xin, *Phys. Rev. Lett.* **2017**, *118*, year.
[63] A. Jain, T. Bligaard, *Phys. Rev. B* **2018**, *98*, 214112.
[64] L. Ward, R. Liu, A. Krishna, V. I. Hegde, A. Agrawa, A. Choudhary, W. Chris, *Phys. Rev. B* **2017**, *96*, 024104.
[65] R. Jalem, M. Nakayama, Y. Noda, T. Le, I. Takeuchi, Y. Tateyama, H. Yamazaki, *Science and Technology of Advanced Materials* **2018**, *19*, 231–242.
[66] P. Panagiotopoulou, D. I. Kondarides, X. E. Verykios, *Appl. Catal. B* **2009**, *88*, 470–478.
[67] F. Calle-Vallejo, J. I. Martínez, J. M. García-Lastra, P. Sautet, D. Loffreda, *Angew. Chem. Int. Ed.* **2014**, *53*, 8316–8319; *Angew. Chem.* **2014**, *126*, 8456–8459.
[68] X. Ma, H. Xin, *Phys. Rev. Lett.* **2017**, *118*, 036101.
[69] Z. Li, X. Ma, H. Xin, *Catal. Today* **2017**, *280*, 232–238.
[70] L. T. Roling, L. Li, F. Abild-Pedersen, *J. Phys. Chem. C* **2017**, *121*, 23002–23010.
[71] L. T. Roling, F. Abild-Pedersen, *ChemCatChem* **2018**, *10*, 1643–1650.
[72] L. T. Roling, T. S. Choksi, F. Abild-Pedersen, *Nanoscale* **2019**, 4438–4452.
[73] T. S. Choksi, L. T. Roling, V. Streibel, F. Abild-Pedersen, *J. Phys. Chem. Lett.*, **2019**, 10,1852-1859**2019**.
[74] A. P. Bartók, R. Kondor, G. Csányi, *Phys. Rev. B* **2013**, *87*, 184115.
[75] L. Hubert, P. Arabie, *J. Classif.* **1985**, *2*, 193–218.
[76] A. Strehl, J. Ghosh, *J. Mach. Learn. Res.* **2003**, *3*, 583–617.
[77] J. B. Hirschberg, A. Rosenberg **2007**.
[78] H.-P. Kriegel, P. Kröger, J. Sander, A. Zimek, *Interdiscip. Sci. Rev.* **2011**, *1*, 231–240.
[79] M. K. Pakhira, *2014 International Conference on Computational Intelligence and Communication Networks*, **2014**, pp. 1047–1051.
[80] C. Fraley, A. E. Raftery, *Comput. J.* **1998**, *41*, 578–588.
[81] M. G. Mavros, J. J. Shepherd, T. Tsuchimochi, A. R. Mclsaac, T. V. Voorhis, *J. Phys. Chem. C* **2017**, *121*, 15665–15674.
[82] T. Kohonen, *Biological Cybernetics* **1982**, *43*, 59–69.
[83] L. E. Baum, T. Petrie, *Ann. Math. Stat.* **1966**, *37*, 1554–1563.
[84] W. Deng, Y. Liu, J. Hu, J. Guo, *Pattern Recognit. Lett.* **2012**, *45*, 4438–4450.
[85] A. N. Tikhonov, *Dokl. Akad. Nauk SSSR*, **1943**, 195–198.
[86] F. Santosa, W. W. Symes, *SIAM J. Sci. Comput.* **1986**, *7*, 1307–1330.
[87] H. Zou, T. Hastie, *J. Royal Stat. Soc.* **2005**, *67*, 301–320.
[88] P. N. Plessow, F. Abild-Pedersen, *J. Phys. Chem. C* **2015**, *119*, 10448–10453.
[89] M. Fields, C. Tsai, L. D. Chen, F. Abild-Pedersen, J. K. Nørskov, K. Chan, *ACS Catal.* **2017**, *7*, 2528–2534.
[90] J. Greeley, *Annu. Rev. Chem. Biomol. Eng.* **2016**, *7*, 605–635.
[91] F. Calle-Vallejo, J. I. Martínez, J. M. García-Lastra, P. Sautet, D. Loffreda, *Angew. Chem. Int. Ed.* **2014**, *53*, 8316–8319; *Angew. Chem.* **2014**, *126*, 8456–8459.
[92] W. T. Hong, R. E. Welsch, Y. Shao-Horn, *J. Phys. Chem. C* **2015**, *120*, 78–86.
[93] K. P. Murphy, *Machine learning: a probabilistic perspective*, MIT press, **2012**.
[94] C. M. Bishop, *Pattern recognition and machine learning*, Springer, **2006**.
[95] M. O. Jäger, E. V. Morooka, F. F. Canova, L. Himanen, A. S. Foster, *npj Comput. Mater.* **2018**, *4*, 37.
[96] J. Greeley, T. F. Jaramillo, J. Bonde, I. Chorkendorff, J. K. Nørskov, *Nat. Mater.* **2006**, *5*, 909–913.
[97] E. Skúlason, V. Tripkovic, M. E. Björketun, S. Gudmundsdottir, G. Karlberg, J. Rossmeisl, T. Bligaard, H. Jónsson, J. K. Nørskov, *J. Phys. Chem. C* **2010**, *114*, 18182–18197.
[98] O. Roy, M. Vetterli, *2007 15th European Signal Processing Conference*, **2007**, pp. 606–610.
[99] A. P. Bartók, M. C. Payne, R. Kondor, G. Csányi, *Phys. Rev. Lett.* **2010**, *104*, 136403.
[100] Z. W. Ulissi, A. J. Medford, T. Bligaard, J. K. Nørskov, *Nat. Commun.* **2017**, *8*, 14621.
[101] Z. W. Ulissi, A. R. Singh, C. Tsai, J. K. Nørskov, *J. Phys. Chem. Lett.* **2016**, *7*, 3931–3935.
[102] L. Breiman, *Machine learning* **2001**, *45*, 5–32.
[103] B. Meredig, A. Agrawal, S. Kirklin, J. E. Saal, J. Doak, A. Thompson, K. Zhang, A. Choudhary, C. Wolverton, *Phys. Rev. B* **2014**, *89*, 094104.
[104] Ç. Odabaşı, M. E. Günay, R. Yıldırım, *Int. J. Hydrogen Energy* **2014**, *39*, 5733–5746.
[105] J. P. Janet, H. J. Kulik, *J. Phys. Chem. A* **2017**, *121*, 8939–8954.
[106] M. Fernandez, H. Barron, A. S. Barnard, *RSC Adv.* **2017**, *7*, 48962–48971.
[107] K. T. Schütt, F. Arbabzadah, S. Chmiela, K. R. Müller, A. Tkatchenko, *Nat. Commun.* **2017**, *8*, 13890.
[108] K. T. Schütt, H. E. Sauceda, P.-J. Kindermans, A. Tkatchenko, K.-R. Müller, *J. Chem. Phys.* **2018**, *148*, 241722.
[109] T. Hattori, S. Kito, *Catal. Today* **1995**, *23*, 347–355.
[110] H. Lukas, S. G. Fries, B. Sundman, *Computational Thermodynamics – The Calphad Method*, Cambridge, **2007**.
[111] A. Y. Chang, S. Chen, F. Zhang, X. Yan, F. Xie, R. Schmid-Fetzer, A. W. Oates, *Prog. Mater. Sci.* **2004**, *49*, 313–345.
[112] J.-O. Andersson, T. Helander, L. Höglund, P. Shi, B. Sundman, *Calphad* **2002**, *26*, 273–312.
[113] W. Cao, S.-L. Chen, F. Zhang, K. Wu, Y. Yang, Y. Chang, R. Schmid-Fetzerc, W. Oates, *Calphad* **2009**, *33*, 328–342.
[114] R. H. Davies, A. T. Dinsdale, J. A. Gisby, J. A. J. Robinson, S. M. Martin, *Calphad* **2002**, *36*, 229–271.
[115] C. Bale, E. Bélisle, P. Chartrand, S. Decterov, G. Eriksson, A. Gheribi, K. Hack, I.-H. Jung, Y.-B. Kang, J. Melançon, A. Pelton, S. Petersen, C. Robelin, J. Sangster, P. Spencer, M.-A. Van Ende, *Calphad* **2016**, *54*, 35–53.
[116] K. A. Persson, B. Waldwick, P. Lazic, G. Ceder, *Phys. Rev. B* **2012**, *85*, 235438.
[117] C. J. Bartel, S. L. Millican, A. M. Deml, J. R. Rumptz, W. Tumas, A. W. Weimer, S. Lany, V. Stevanović, C. B. Musgrave, A. M. Holder, *Nat. Commun.* **2018**, *9*, 4168.
[118] K. Takahashi, T. Lauren, *J. Phys. Chem. Lett.* **2019**, *10*, 283.
[119] A. Seko, H. Hayashi, K. Nakayama, A. Takahashi, I. Tanaka, *Phys. Rev. B* **2017**, *95*, 144110.

[120] J. Schmidt, J. Shi, P. Borlido, L. Chen, S. Botti, M. A. L. Marques, *Chem. Mater.* **2017**, *29*, 50905103.
[121] T. Xie, J. C. Grossman, *Phys. Rev. Lett.* **2018**, *120*, 145301.
[122] E. Gossetta, C. Tohera, C. Osesa, O. Isayevc, F. Legraind, F. Rosea, E. Zurekf, J. Carreteg, N. Mingod, A. Tropshac, S. Curtaroloa, *Comput. Mater. Sci.* **2018**, *152*, 134–145.
[123] L. Ward, A. Dunn, A. Faghaninia, N. E. Zimmermannc, S. Bajaj, Q. Wang, J. Montoya, J. Chen, K. Bystromd, M. Dyllag, K. Chard, M. Astad, K. A. Persson, G. J. Snyder, I. Foster, A. Jain, *Comput. Mater. Sci.* **2018**, *152*, 60–69.
[124] M. Picklum, M. Beetz, *Comput. Mater. Sci.* **2019**, *163*, 50–62.
[125] J. H. Montoya, K. A. Persson, *npj Comput. Mater.* **2017**, *3*, 1–3.
[126] J. R. Boes, O. Mamun, K. Winther, T. Bligaard, *J. Phys. Chem. A* **2019**, *0*, 2281–2285.
[127] A. Holewinski, H. Xin, E. Nikolla, S. Linic, *Curr. Opin. Chem. Eng.* **2013**, *2*, 312–319.
[128] H. Xin, A. Holewinski, S. Linic, *ACS Catal.* **2012**, *2*, 12–16.
[129] K. Saravanan, J. R. Kitchin, O. A. von Lilienfeld, J. A. Keith, *J. Phys. Chem. Lett.* **2017**, *8*, 5002–5007.
[130] X. Ma, Z. Li, L. E. K. Achenie, H. Xin, *J. Phys. Chem. Lett.* **2015**, *6*, 3528–3533.
[131] Z. Li, S. Wang, W. S. Chin, L. E. Achenie, H. Xin, *J. Mater. Chem. A* **2017**, 24131–24138.
[132] S. Wang, N. Omidvar, E. Marx, H. Xin, *Phys. Chem. Chem. Phys.* **2018**, *20*, 6055–6059.
[133] K. Tran, Z. W. Ulissi, *Nat. Catal.* **2018**, *1*, 696–703.
[134] R. Jinnouchi, R. Asahi, *J. Phys. Chem. Lett.* **2017**, *8*, 4279–4283.
[135] R. Jinnouchi, H. Hirata, R. Asahi, *J. Phys. Chem. C* **2017**, *121*, 26397–26405.
[136] T. Toyao, K. Suzuki, S. Kikuchi, S. Takakusagi, K. I. Shimizu, I. Takigawa, *J. Phys. Chem. C* **2018**, *122*, 8315–8326.
[137] A. J. Chowdhury, W. Yang, E. Walker, O. Mamun, A. Heyden, G. A. Terejanu, *J. Phys. Chem. C* **2018**, *122*, 28142–28150.
[138] M. Andersen, S. Levchenko, M. Scheffler, K. Reuter, *ACS Catal.* **2019**, *0*, 2752–2759.
[139] E. Shustorovich, *Surf. Sci. Rep.* **1986**, *6*, 1–63.
[140] T. Choksi, P. Majumdar, J. P. Greeley, *Angew. Chem. Int. Ed.* **2018**, *57*, 15410–15414.
[141] A. P. Bartók, R. Kondor, G. Csányi, *Phys. Rev. B* **2013**, *87*, 184115.
[142] B. Settles, *Synthesis Lectures on Artificial Intelligence and Machine Learning* **2012**, *6*, 1–114.
[143] Z.-H. Han, K.-S. Zhang in *Real-World Applications of Genetic Algorithms* (Ed.: O. Roeva), IntechOpen, Rijeka, **2012**, Chapter 17.
[144] R. B. Wexler, J. M. P. Martirez, A. M. Rappe, *J. Am. Chem. Soc.* **2018**, *140*, 4678–4683.
[145] D. M. Deaven, K.-M. Ho, *Phys. Rev. Lett.* **1995**, *75*, 288.
[146] D. Daven, N. Tit, J. Morris, K. Ho, *Chem. Phys. Lett.* **1996**, *256*, 195–200.
[147] D. J. Wales, J. P. Doye, *J. Phys. Chem. A* **1997**, *101*, 5111–5116.
[148] E. Bitzek, P. Koskinen, F. Gähler, M. Moseler, P. Gumbsch, *Phys. Rev. Lett.* **2006**, *97*, 170201.
[149] V. Botu, R. Ramprasad, *Int. J. Quantum Chem.* **2015**, *115*, 1074–1083.
[150] M. S. Jørgensen, U. F. Larsen, K. W. Jacobsen, B. Hammer, *J. Phys. Chem. A* **2018**, *122*, 1504–1509.
[151] T. Jacobsen, M. Jørgensen, B. Hammer, *Phys. Rev. Lett.* **2018**, *120*, 026102.
[152] E. G. del Río, J. J. Mortensen, K. W. Jacobsen, *arXiv preprint arXiv:1808.08588* **2018**.
[153] E. Iype, S. Urolagin, *J. Chem. Phys.* **2019**, *150*, 024307.
[154] C. G. Broyden, *J. Inst. Math. Its Appl.* **1970**, *6*, 76–90.
[155] R. Fletcher, *Comput. J.* **1970**, *13*, 317–322.
[156] D. Goldfarb, *Math. Comput.* **1970**, *24*, 23–26.
[157] D. F. Shanno, *Math. Comput.* **1970**, *24*, 647–656.
[158] H. Jónsson, G. Mills, K. W. Jacobsen **1998**.
[159] G. Henkelman, B. P. Uberuaga, H. Jónsson, *J. Chem. Phys.* **2000**, *113*, 9901–9904.
[160] A. A. Peterson, *J. Chem. Phys.* **2016**, *145*, 074106.
[161] O.-P. Koistinen, F. B. Dagbjartsdóttir, V. Ásgeirsson, A. Vehtari, H. Jónsson, *J. Chem. Phys.* **2017**, *147*, 152720.
[162] J. A. G. Torres, P. C. Jennings, M. H. Hansen, J. R. Boes, T. Bligaard, *Phys. Rev. Lett.* **2019**, *122*, 156001.
[163] P. M. Zimmerman, *J. Comput. Chem.* **2015**, *36*, 601.
[164] S. M. Sharada, P. M. Zimmerman, A. T. Bell, M. Head-Gordon, *J. Comput. Chem.* **2012**, *8*, 5166.
[165] Y. V. Suleimanov, W. H. Green, *J. Chem. Theory Comput.* **2015**, *11*, 4248.
[166] J. Zador, H. N. Najm, *KinBot 1.0: a code for automatic PES exploration*, Sandia National Laboratories Technical Report No. SAND2012-8095, **2013**.
[167] S. Maeda, Y. Harabuch, *J. Chem. Theory Comput.* **2019**, *15*, 2111.
[168] R. a. J. T. S. Coley, C. W. Barzilay, W. H. Green, K. F. Jensen, *ACS Cent. Sci.* **2017**, *3*, 434.
[169] *CatLearn Code*, arXiv:1904.00904, Accessed: 2019–04-01.
[170] G. Pilania, A. Mannodi-Kanakkithodi, B. P. Uberuaga, R. Ramprasad, J. E. Gubernatis, T. Lookman, *Sci. Rep.* **2016**, *6*, 19375.
[171] J. Suntivich, K. J. May, H. A. Gasteiger, J. B. Goodenough, Y. Shao-Horn, *Science* **2011**, *334*, 1383–1385.
[172] A. Vojvodic, J. K. Norskov, *Science* **2011**, *334*, 1355–1356.
[173] F. Calle-Vallejo, O. A. Díaz-Morales, M. J. Kolb, M. T. M. Koper, *ACS Catal.* **2015**, *5*, 869–873.
[174] A. Vojvodic, J. K. Nørskov, *Natl. Sci. Rev.* **2015**, *2*, 140–149.
[175] C. Roy, B. Sebok, S. B. Scott, E. M. Fiordaliso, J. E. Sørensen, A. Bodin, D. B. Trimarco, C. D. Damsgaard, P. C. K. Vesborg, O. Hansen, I. E. L. Stephens, J. Kibsgaard, I. Chorkendorff, *Nat. Catal.* **2018**, *1*, 820–829.
[176] R. B. Sandberg, M. H. Hansen, J. K. Nørskov, F. Abild-Pedersen, M. Bajdich, *ACS Catal.* **2018**, *8*, 10555–10563.

Manuscript received: April 1, 2019
Revised manuscript received: May 14, 2019
Accepted manuscript online: May 15, 2019
Version of record online: ■■, ■■■

REVIEWS

Advances in machine learning and data science hold great potential for the rapid computational screening of solid-state catalyst materials candidates and to accelerate the computation of potential energy landscapes. In this review, we outline important data science and machine learning concepts and show how these are applied in the field of computational heterogeneous catalysis. We cover topics like data storage and sharing, materials featurization and using machine learning models for predictive and mechanistic studies.



Dr. P. Schlexer Lamoureux, Dr. K. T. Winther, Dr. J. A. Garrido Torres, Dr. V. Streibel, Dr. M. Zhao, Dr. M. Bajdich, Dr. F. Abild-Pedersen, Dr. T. Bligaard*