INL-JOU--18-50172-Revision-0

Extracting knowledge from data through catalysis informatics

Andrew J Medford, M. Ross Kunz, Sarah M. Ewing, Tammie Borders, Rebecca Fushimi

The INL is a
U.S. Department of Energy
National Laboratory
operated by
Battelle Energy Alliance

June 2018

Prepared for the U.S. Department of Energy
Office of Energy Efficiency and Renewable Energy
Under DOE Idaho Operations Office
Contract DE-AC07-05ID14517

Extracting knowledge from data through catalysis informatics

Andrew J. Medford,*,† M. Ross Kunz,‡ Sarah M. Ewing,‡ Tammie Borders,‡ and Rebecca Fushimi*,‡,¶

†*School of Chemical & Biomolecular Engineering, Georgia Institute of Technology, Atlanta, GA, USA*

‡*Biological and Chemical Processing Department, Energy and Environmental Science and Technology, Idaho National Laboratory, PO Box 1625, Idaho Falls, ID 83415, USA*

¶*Center for Advanced Energy Studies, 995 University Boulevard, Idaho Falls, ID 83401, USA*

E-mail: andrew.medford@chbe.gatech.edu; rebecca.fushimi@inl.gov

Abstract

Catalysis informatics is a distinct sub-field that lies at the intersection of cheminformatics and materials informatics, but with unique challenges arising from the dynamic, surface-sensitive, and multi-scale nature of heterogeneous catalysis. The ideas behind catalysis informatics can be traced back decades, but the field is only recently emerging due to advances in data infrastructure, statistics, machine learning, and computational methods. In this work we review the field from early works on expert systems and knowledge engines to more recent approaches utilizing machine-learning and uncertainty quantification. The data-information-knowledge hierarchy is introduced and used to classify various developments. The chemical master equation and micro-kinetic models are proposed as a quantitative representation of catalysis knowledge, which can be used to generate explanative and predictive hypotheses for the understanding and discovery of catalytic materials. We discuss future prospects for the field, including improved quantitative coupling of experiment/theory, advanced micro-kinetic models, and the development of open-source software tools. Ultimately, integration of existing chemical and physical models with emerging statistical and computational tools presents a promising route toward the automated design, discovery, and optimization of heterogeneous catalytic processes.

1 Introduction

The term "catalysis informatics" has been used increasingly in the field of heterogeneous catalysis since as early as 2001,[1] yet there is no clear definition of the phrase. The term "informatics" appears relatively few times in the catalysis literature,[1–10] with a few works referring to "catalysis informatics" specifically in several rather different contexts.[1,4,7–10] In this work we define catalysis informatics as the extraction of knowledge from information via the design, representation and organization of data sets and the application of data mining and analysis tools to accelerate the discovery and understanding of heterogeneous catalytic materials. This definition includes many of the machine-learning techniques that have recently been applied to the field of catalysis,[7,10–21] but the focus is on the extraction of actionable and interpretable knowledge, rather than the identification of patterns or acceleration of methods.[22] We limit the scope to hetero-

geneous catalytic materials which have a number of unique challenges. We note that informatics approaches are also relevant to homogeneous[18,23] and biological catalysts;[24] however, these applications are largely covered by the existing fields of cheminformatics[25,26] and bioinformatics[27] respectively. Furthermore, we focus primarily on methods that are unique to the challenges of heterogeneous catalysis. Cheminformatics (or chemometrics) has been established for decades[25] and involves the extraction of knowledge from molecular structures that are relevant to the chemical products, reactants, and intermediates of catalytic reactions.[28] The field of "materials informatics" is also relatively well established,[29,30] and addresses properties of solid materials that may act as catalysts. Similarly, "atomistic informatics" approaches such as machine-learning force fields[31,32] and large databases of atomic structures/properties[33,34] have recently become prevalent. While all of these fields are relevant to heterogeneous catalysis, they are not unique to it and will only be discussed in situations where there is significant overlap.

The field of heterogenous catalysis has several unique attributes that make it particularly challenging from an informatics perspective, and for these reasons we argue that "catalysis informatics" is a necessary addition to the existing zoo of informatics subfields. For one, the phenomenon of heterogeneous catalysis involves the interaction of a molecule with a solid surface. These systems are often treated as distinct in informatics approaches; molecules are discrete entities and many solid surfaces are periodic. Different mathematical descriptions are often employed to describe these fundamentally different types of atomic-scale systems. This issue persists at the macro-scale, where catalytic properties depend on both the material (surface) structure and the chemical environment in which the catalyst is operating. Thus, "catalysis informatics" lies at the intersection of materials informatics and cheminformatics. This fact breaks the traditional "process-structure-property" paradigm of materials informatics, since the relevant surface structure will be linked to both the structure of the solid surface and the chemical potential of gas-phase species in the reaction environment;[35] for this reason we propose process-[structure+environment]-property relationships that are more appropriate for catalysis (see Fig. 1). Furthermore, heterogeneous catalysis is an inherently local surface phenomenon, often controlled by defects, typically with sizes on the order of a few ångström.[36–39] This leads to a situation in which easily measured bulk properties (e.g. composition) are not directly related to catalytic properties (although indirect correlations are possible)[40] since the composition of the surface termination is often different from that of the bulk. These properties lead to a "featurization challenge" in catalysis, referring to the fact that it is difficult to identify the salient features that control the catalytic behavior of a material at both the atomic and macro scales.

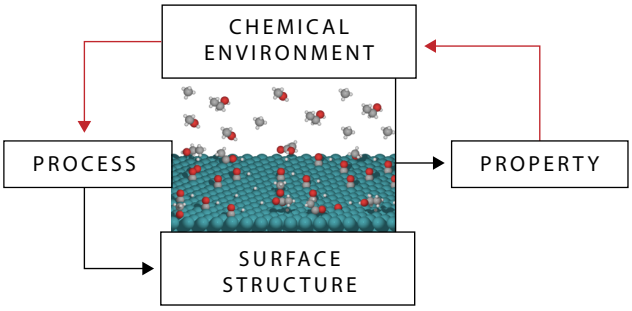

Figure 1: Schematic of process-[structure+environment]-property relationship for catalysis. Red arrows depict dynamic processes since the catalytic property of a material will affect the chemical environment (i.e. gas composition changes as a function of conversion), and the chemical environment can induce in-situ structural changes (i.e. surface reconstruction). At steady-state the red arrows are no longer relevant, although the description of "process" will include processes that happen after the reaction initiates.

An additional challenge arises due to the fact that heterogeneous catalysis is a dynamic multi-scale phenomenon that spans >9 orders of magnitude in time and length scales.[41] The dissociation of molecular bonds is rare on the timescale of atomic motion; atoms move on the

timescale of picoseconds, yet the frequency of bond dissociation occurs on the order of seconds. Furthermore, the surface state of a catalytic material is highly sensitive to the reaction conditions, and active sites may be dynamic, forming in situ during the course of a reaction. This also deviates from the static "process-structure-property" framework, since the catalytic properties affect the reaction environment, creating a feedback loop that effectively acts as a "process" on the surface (see Fig. 1) and makes direct observation of catalytic events or active sites extremely difficult. Instead, catalytic properties are measured indirectly at the macro-scale, and properties of the active site and rate-limiting step must be inferred from these observations. Alternatively, computational techniques such as density functional theory (DFT) can be applied to obtain direct insight into the energetics of active sites and bond-breaking events. However, these techniques require knowledge of the atomic-scale structure of the active site, which must be assumed or inferred from experiment. This makes the prediction of emergent phenomena extremely difficult. Furthermore, even if an active site structure is known, the quantum mechanical techniques that can be applied to catalytic systems are of limited accuracy. For example, DFT with the conventional generalized-gradient approximation (GGA) is known to exhibit errors on the order of 0.2 eV for adsorption energies,[42] and even hybrid techniques do not have systematically improvable accuracy.[43] Wavefunction methods such as coupled cluster are prohibitively expensive for periodic systems needed to model extended surfaces,[44] and approximating surfaces with clusters is also expensive for wavefunction methods and the accuracy converges slowly with cluster size or requires complex QM/(QM/MM) embedding approaches.[45–48] These factors result in an "uncertainty challenge" since there is rarely a reliable ground truth.

Despite these enormous complexities, the field of heterogeneous catalysis has made steady and continuous progress in the development of industrial catalysts. This has occurred primarily though two strategies: trial-and-error discovery[49,50] and systematic reductionist approaches using model catalysts.[51] The history of the ammonia synthesis catalyst, one of the first and most widely-studied catalytic systems, provides a useful illustration of these two principles. The original ammonia synthesis catalyst was reported by Haber in the early 1900's and subsequently optimized through trial-and-error testing led by Mittasch and Bosch.[49] This testing included both single- and multi-component catalysts, and it is estimated that over 10,000 catalytic tests were performed to discover the iron-based catalyst that is still widely used in industry.[49,52,53] More modern work has focused on ammonia synthesis as a prototypical reaction,[51] and a wide range of model catalysts based on iron, ruthenium, and molybdenum have been studied with surface-science and spectroscopic techniques.[53] These studies yielded fundamental insight into the nature of the reaction mechanism and active sites that are active for ammonia synthesis, and the development of novel catalysts such as the ruthenium-based catalysts that maximize the abundance of the active step sites,[54] and Co-Mo catalysts that exhibit optimal nitrogen binding energies.[55] The trial-and-error and reductionist approaches are complimentary. Trial-and-error approaches are "broad" searches in materials space, while reductionist approaches are "deep" searches into the fundamental mechanism of a reaction. The application of data science techniques is a promising route to integrating these two philosophies of catalyst design and can accelerate both discovery and understanding of catalytic systems.

This work reviews the field of "catalysis informatics" including early "expert systems"[56–59] and data-driven empirical correlations[60–65] to more recent "knowledge extraction" engines[66–68] and descriptor-based microkinetic approaches.[40,69,70] We discuss the central role of the chemical "master equation"[71] (Eq. 1) and micro-kinetic models[72] in catalysis and classify informatics approaches as either model-based or model-free. We also identify emerging opportunities in the field, and discuss the prospects for using a data-driven framework for fusing computational and experimental information

to systematically identify critical parameters of a catalytic system and utilize this knowledge toward the prediction of novel catalytic materials. Ultimately, we expect that "catalysis informatics" will continue to grow in importance and accelerate the catalyst development process both in academia and industry.

2 From Data to Knowledge

The paradigm of the data-information-knowledge pyramid (see Fig. 2) is a useful conceptual framework for assessing informatics approaches,[22,73] although the specifics of this mapping are somewhat subjective and vary by field. The philosophy of knowledge extraction and discovery from data is well-studied in the computational and statistical communities, and numerous reviews[22,74–76,76] and texts[77–81] are recommended for a more thorough perspective. In this work we seek to map these concepts to the context of heterogeneous catalysis (Fig. 2), and propose the following definitions:

- data: catalytic reaction (e.g. activity, selectivity, stability), surface characterization (e.g. XPS, DRIFTS), materials characterization (e.g. density, surface area, crystal structure), reaction environment (e.g. temperature, pressure, concentrations)
- information: analysis of data using statistical tools (e.g. regression, classification) and/or underlying physical models to organize, understand, and utilize data.
- knowledge: integration of information derived from statistical/physical models with approximations to the chemical master equation (e.g. micro-kinetic models, kinetic Monte Carlo) to establish quantitative and generalizable insight into the nature of the active site and mechanism of the catalytic process.

Establishing clear conceptual lines between these categories is not straightforward, since "data" can often refer to the result of a model

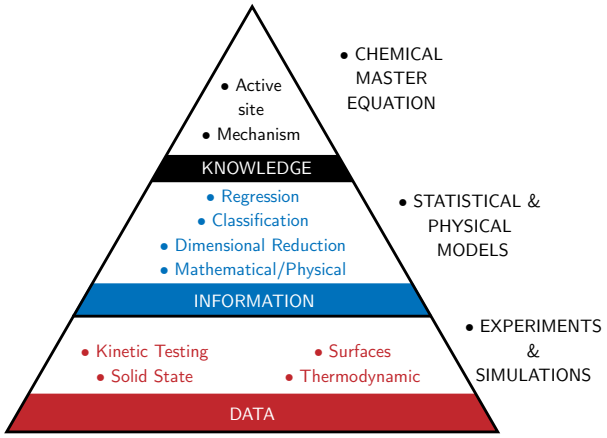

Figure 2: Schematic of data-information-knowledge hierarchy.

applied to raw data, and the line between "information" and "knowledge" is philosophical in nature. Here we propose that "data" refers to data as measured, simulated, or computed using clearly defined and recorded devices and protocols. Thus "derived data" resulting from non-trivial post-analysis falls into the category of "information", since the derived data will depend on underlying assumptions. For example, the adsorption energy and vibrational frequencies of an adsorbate computed using DFT are "data", while the Gibb's free energy of adsorption at a specific temperature computed using the harmonic approximation of statistical mechanics would be "information" derived from this data. The distinction between "information" and "knowledge" presented here relies primarily on the mapping to the chemical "master equation"[71] that ultimately governs surface kinetics. The master equation represents a full state of knowledge about the catalytic system, from which all observable catalytic properties can be derived. However, in practical heterogeneous catalysis applications the equation is typically reduced to micro-kinetic models that employ numerous simplifications including truncated reaction mechanisms and mean-field approximations;[72] these models are not as rigorous as the master equation, but still provide useful organizing principles for knowledge extraction. In this work we use the term "master equation" in the general sense which en-

compasses mean-field or phenomenological approximations[72,82] as well as stochastic[83–85] and deterministic[86,87] solutions to explicit lattice models. Essentially, the micro-kinetic model consists of a reaction mechanism and corresponding thermodynamic equilibrium and kinetic rate constants. This represents a clear definition, and emphasizes the fact that the micro-kinetic model is the mathematical link connecting the state variables of the system (temperature, pressure, etc.), intrinsic parameters of a catalytic surface (e.g. equilibrium and rate constants), the chemical reaction mechanism, and the observed global reaction rate. Given an accurate kinetic model it is possible to generalize the behavior of a catalytic material;[72,88] thus we concur with Caruthers et. al. that "the [kinetic] model is a quantitative representation of knowledge about the catalytic system." [66]

These proposed boundaries between data, information, and knowledge are fuzzy; the fine details that distinguish these categories are less critical than the broad ideas which will serve as a basis for the following sections. A more detailed overview of the data-information-knowledge continuum is illustrated in Fig. 3. The following sections discuss the various components of this continuum to describe how data flows to knowledge through the process of conceptualization and model construction, and how this knowledge can be used to generate new hypotheses. Testing these new hypotheses leads to an expansion of available data, and enables model refinement. This hierarchy shares similarities with multi-scale modeling which seeks to link catalytic reaction data to a micro-kinetic model;[89] however, informatics is broader in that more types of data are considered and models need not have any physical basis. We note that the data-information-knowledge hierarchy is not meant as a system to rank the impact of various approaches; one could argue that innovation at the data level is most impactful, since data is the "raw material" of informatics and has the potential to broadly affect efficient and robust generation of information and knowledge. Furthermore, while knowledge is the ultimate pursuit of academic studies, profit is the driver for most industrial research, and we anticipate that in some scenarios approaches at the "information" level may be more economical. Ultimately, the field of catalysis informatics will require innovation at, and integration across, all levels of the data-information-knowledge hierarchy to become a widely adopted and practically impactful discipline.

2.1 Data

Data is the central currency of quantitative science, and forms the foundation of any informatics approach. More generally, data is the foundation of all empirical science, and hence should be recorded and stored in a systematic and reproducible way. Informatics is an inherently data-driven field, and high-quality, machine-readable data with appropriate contextual meta-data will enhance the efficiency with which catalysis informatics can convert data to knowledge. This section briefly considers issues of (big) data storage for catalysis, and outlines some key types of catalysis data along with their context and purpose.

2.1.1 Data Storage

The general workflow of designing experiments and recording results generates data, but is not sufficient to constitute "informatics". Informatics refers to the systematic and quantitative extraction of information and knowledge from large amounts of data, which requires structured storage of machine-readable data.In the context of catalysis, a wide range of materials and chemical properties are relevant, and numerous databases exist. In this regard, the data of "catalysis informatics" is largely the same as the data of materials and cheminformatics. For example, crystallographic databases, thermodynamic databases, and computational databases are all relevant to catalysis (see Ref. 90 for a list of examples). In addition, the CatApp database[91] contains a large number of computational results specifically for catalytic processes, and numerous databases have been developed for the structures of nanoporous catalytic materials such as zeolites and metal-

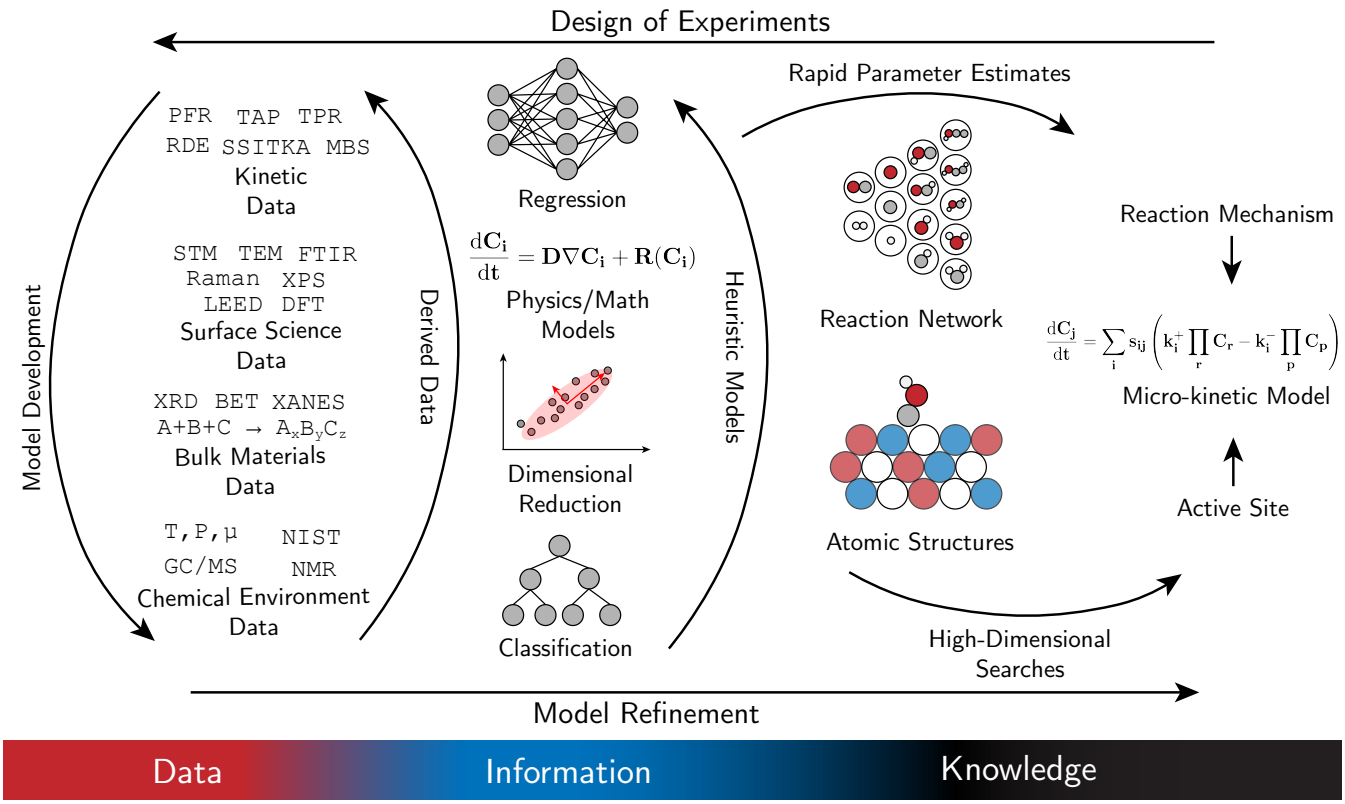

Figure 3: Schematic of relationship between data, information, and knowledge in heterogeneous catalysis. Data abbreviations are listed in Table 2 and are meant to be representative not comprehensive. Models based on statistics and/or physics can be used to create derived data, to establish heuristic relationships between data, or to rapidly estimate parameters for reaction network or atomic-scale surface models. The reaction network and atomic-scale structure are high-dimensional structures that can be distilled to reaction mechanism(s) and active site(s) via high-dimensional searches. Ultimately this leads to design of experiments and refinement of micro-kinetic models based on the chemical master equation.

organic frameworks (MOFs).[92–94] Despite the vast amount of available materials data, it is arguable as to whether or not it constitutes "big data".[90] Typically, big data is considered in terms of the three key metrics, or the "3V's", of volume, velocity, and variety (see Table 1). Data-centric companies like Google, Facebook, and YouTube must deal with volumes of data on the scale of exabytes (10^9 GB), and velocities on the scale of petabytes (10^6 GB) per day.[90,95] In comparison, the relatively mature Materials Project repository[33] contains less than 10^6 total entries accumulated over several years, and this far exceeds the size of any catalysis-specific database. Based on this one can conclude that materials (and catalysis) data does not qualify as "big" based on volume or velocity.[90]

One commonality between materials data and truly "big" data is variety. The tremendous variety of materials and catalysis data exceeds that of the more traditional data science sectors. Data in business or social media are typically generated within a controlled (digital) context, and often managed by a single entity. Conversely, catalysis data spans a broad range of scientific disciplines (e.g. materials science, chemistry, chemical engineering), and are often generated based on experiments involving physical systems that include inherent uncertainty. Furthermore, the materials sector has a variety of stakeholders including academic groups, national labs, and industrial research groups and there is a lack of standardized data formats.[96] This leads to data that is typically siloed into discipline-specific databases like the ones mentioned previously, if it is cap-

tured at all.[90,97] There have been several efforts to amalgamate data from these varied sources using common data formats and application programming interfaces (APIs) such as the NIST Materials Data Curation System and the Citrination database.In particular, the Citrination database includes catalysis-specific data from the CatApp database and the literature. The ability to access data from a variety of discipline-specific databases is crucial for catalysis, where data from a range of fields are relevant. Yet, there is also a large amount of catalysis data that has been generated through high-throughput combinatorial approaches[2,66,98] but is not openly available or stored in any accessible way. The problem is exacerbated by the vast number of possible representations needed for catalysis data. For example, molecules and reaction mechanisms can be represented by graphs, matrices, or text strings, each of which has various advantages and disadvantages.[77] Establishing more open databases and amalgamating this data into common machine-readable formats is a first step toward harnessing catalysis data to enable new discoveries.[68]

In addition to the canonical "3V's" of big data a plethora of other "V's" have emerged.[99,100] There are an additional two that are relevant to catalysis data: veracity and volatility.[100] The accuracy (veracity) of data is always important, but given the difficulty of identifying critical features of catalysts and the inherent uncertainty associated with the data it is especially relevant to the field of catalysis. In the case of experimental data reproducibility is often an issue due to the fact that the catalytic properties can be sensitive to the presence of surface defects that are difficult to measure.[101] For example, Rh/SiO_2 catalysts that are nominally the same exhibit conversions that vary by a factor of 5 based on synthesis details such as Rh precursor, SiO_2 supplier and support washing procedures.[102] This sensitivity is attributed to a combination of defects and impurities, two difficult-to-control factors that have proven critical in numerous other catalytic processes including the ammonia synthesis and oxygen evolution reactions.[103–105] As a result of the Arrhenius dependence, often in catalysis a small amount of active sites may have a marked impact on the rate. The implication is that it is difficult to make an *a priori* assessment of which process details are pertinent or superfluous. For this reason all synthesis process and reaction setup details are crucial contextual meta-data for experimental catalysis. Furthermore, this meta-data should be electronically linked to the corresponding catalytic testing data in order to maintain data veracity; this is far from common practice, as most of these details are buried in supplementary information or not reported at all. The development of benchmark datasets and procedures for surface science data[42] and various catalytic processes[106,107] is also critical to establishing consensus on testing and reporting procedures, and will ultimately improve data veracity in the field of catalysis.[108]

The fifth and final "V" of big data that we discuss is volatility,[100] referring in this context to the volatility of (meta-)data structures. The significant detail needed in the meta-data for processing and reaction conditions, coupled with the wide range of classes of catalyst materials (e.g. metals, oxides, zeolites, supported/unsupported, etc.), reaction types (e.g. thermocatalysis, electrocatalysis, photocatalysis), reactor setups (e.g. plug flow, batch, rotating disk, etc.) and data types (e.g. kinetic, thermodynamic, spectroscopic, computational, etc.) makes development of a universal schema for catalyst data practically intractable. For example, researchers may discover that a previously-unrecorded detail (e.g. support supplier) is actually critical, requiring the meta-data of prior records be updated. Another scenario is a researcher who decides to create composite catalysts from different classes (e.g. metal particles supported on zeolites), where the resulting composite catalyst class will inherit features from both sub-classes, and also have features unique to the composite. This volatility suggests that flexible data structures such as JSON[109] and PIF[96] will be necessary to structure catalyst data, and schema-free database technologies such as MongoDB[110] and ElasticSearch[111] can aid in search and retrieval.

Table 1: Five V's of "Big" Data

V	Description	Unit	Issue in catalysis?
Volume	Size of data	bytes	No
Velocity	Flux of data	bytes/time	No
Variety	Diversity of data	count	Yes
Veracity	Uncertainty of data	probability	Yes
Volatility	Stability of data structure	time	Yes

These approaches aid with integration of heterogeneous data while providing the machine-readability that is a critical enabler of informatics approaches.

2.1.2 Data Types & Purpose

The variety and volatility of catalysis data makes it futile to attempt to comprehensively list all types of relevant data, or propose details of how the data should be stored; however, it is useful to consider some basic data types and the purposes for which the data may be used. We propose that there are four types of data relevant to catalysis: data on the chemical environment, data on the catalyst material, data on the active surface, and data on reaction kinetics. The latter two types are unique to catalysis, and are related to the micro-scale interactions between the catalyst active sites and molecules (surface science), and macro-scale observations of the results of these interactions (catalytic reaction). We briefly describe some representative data of each type (collected in Table 2), and discuss relevant meta-data. In general data should be stored along with meta-data clearly describing the device used to measure it, the protocols employed, and the units in which it is measured. By associating this meta-data with the relevant data, the veracity of catalysis data can be greatly improved.

Data on the chemical environment describe in essence the chemical potentials of product and reactant species that are not bound to the catalyst surface. This includes the reaction conditions (temperature, pressure, concentrations, illumination, applied potential, etc.), as well as the nature and concentrations of the reactants and products. These data also include the thermochemical properties of product and reactant species, solvents, electrolytes, or other spectator species; these data are readily available from sources such as the NIST Webbook[112] or PubChem database[113] for many common chemical species. In recording chemical environment data for specific catalytic reactions it is worth considering the devices that are used to measure the data. For example, product concentrations are often measured with analytical techniques such as gas-chromatography/mass-spectrometry (GC/MS) or nuclear magnetic resonance (NMR). There are often implicit assumptions made in the analysis of this raw data to determine product concentration (e.g. calibration curves). Any such assumptions should be noted as meta-data for concentrations, and where possible the raw data should be stored as well. Another key consideration for chemical environment data is the fact that it is often measured at the macro-scale, while the micro-scale environment is most relevant to the catalytic phenomena. The difference between the two is typically due to thermal and mass transport, and in general the micro-scale chemical environment will not be spatially homogeneous. For this reason it is key that data on temperature, flow rate, and pressure include details of the devices and protocols used to measure them. Ideally the corresponding catalytic reaction data can be utilized to assess the influence of these transport effects.

Data on the material describe the static properties of the catalyst material that are independent of the reaction environment. This primarily corresponds to the bulk properties of the catalyst material, but may also include characterizations such as particle size distribution, surface area, or dopant concentration. We note that in practice these properties may change during the reaction, but such changes would

generally be observable through characterization after the catalytic testing. More specifically, materials data is not related to the intermediate species or materials states that may occur during the course of a reaction. Materials data for catalysis is in effect the same as data for materials informatics, and hence databases such as Materials Project[33] and the Citrination Platform[114] can be leveraged. Further, we consider the synthesis conditions or supplier of catalysts to fall into the category of materials data. Process data is particularly challenging to store and analyze due to the fact that it is time and history dependent - heating two individual solutions and then mixing them is not the same as mixing two individual solutions and then heating them. This time- and history-dependent nature of process data has been identified as a challenge in the materials informatics community,[115] and is beyond the scope of this work. Nonetheless, recording the details of synthesis procedure provides key context for catalyst materials and should be recorded as meta-data. Knowledge of supplier and precursor purity is of particular importance, due to the potentially large influence of impurities.[102,105]

Surface science data is not unique to catalysis, but it is significantly less abundant than chemical or materials data. This data is surface-sensitive, and hence corresponds to the most critical region of the material for a catalytic reaction. Surface science data can be measured outside of the reaction environment (ex situ, often in ultra-high vacuum), in a similar environment to the reaction (in situ), or under functioning catalytic conditions (operando). Surface science measurements are typically complex, and indeed surface science is an entire field of research, hence we do not seek to cover possible types of meta-data. Rather, we briefly discuss the types of techniques that may be used, and the importance of in situ and operando data. In general, surface science involves selectively probing the surface region, including adsorbed species, using (photo)electrons (XPS, LEED, TEM), infrared radiation (FTIR, Raman), or atoms/molecules (ISS, TOF-SIM, TPR, TPD, TAP). This data corresponds to information regarding the adsorption energies and geometric structure of intermediate states on the catalyst surface. We also consider DFT to be primarily a source of surface-science data in catalysis, since it is commonly used to provide adsorption energies. In measuring surface science data it is important to consider the local reaction environment at the surface during the measurement, since catalyst surfaces are often dynamic and surface data in one environment (e.g. single crystal at low pressure) may not be valid at other conditions (e.g. nanoparticle at high pressure). This has been dubbed the "pressure and materials gap" [116,117] and highlights the importance of linking chemical environment data with surface science data. Furthermore, the dynamic nature of catalysis means that the most valuable surface-science data will also be time-dependent, effectively measuring the properties of the surface during a working reaction. The storage and meta-data for time-dependent surface-science data will require additional meta-data linking it to the corresponding catalytic reaction data.

The final, and most directly relevant, category of data is reaction kinetics data. This is effectively time-dependent chemical environment data, from which the key catalytic metrics of such as activity, selectivity, and stability can be derived. Catalytic reaction data needs meta-data corresponding to the relevant details of the reactor setup (PFR, CSTR, RDE, TAP, etc.) and care should be taken to minimize or quantify heat and mass transport effects that may cause discrepancies between the chemical environment at the catalyst surface and the chemical environment measured at the reactor inlet/outlet. Further, colloquial metrics such as "activity" are not well-defined. Where possible the activity should be reported as turnover frequency, which is derived from numerous assumptions and additional surface-science data; these assumptions and supplementary data should be linked to the catalytic reaction data. Even the catalytic rate requires context, since the rate will in general be time-dependent. Recording the time at which the rate was measured along with any assumptions regarding steady-state will provide potentially important context. The metrics of selectivity

and stability are even less well-defined than activity, and it is all too common to find data in the literature that cannot be directly compared due to differing definitions and lack of detail in how these metrics are computed; where possible raw concentration vs. time data should be recorded and stored along with catalytic reaction metrics. Comprehensive guidelines for reporting catalytic reaction data are beyond the scope of this work, but have been discussed elsewhere.[106–108] In the context of informatics we note that as data infrastructure improves it will become increasingly easy to store raw and derived catalytic data along with the appropriate meta-data, and this will facilitate improved performance in informatics approaches.

2.2 Information

Data is transformed to information by analysis to extract patterns and quantifiable relationships using statistical and/or physical models; here we focus primarily on statistical and data-driven models. In the context of catalysis informatics, we consider information to be derived data or patterns that do not rely on the chemical master equation and do not directly address questions of active site or reaction mechanism. This information can be broadly classified into two levels: macro-scale information and micro-scale information. Macro-scale information does not rely on a fundamental or atomic-scale perspective, but rather seeks a direct empirical relation between some macro-scale descriptor of catalyst process/structure and the resulting catalytic performance in terms of activity, selectivity, stability, etc. In contrast, micro-scale information is focused on understanding the fundamental interactions between a catalyst surface and the adsorbed intermediate species that exist during the process of catalysis; however, this information does not directly relate to the practical catalytic performance. The gap between micro- and macro-scale information has been dubbed the "materials and pressure gap" [116–118] due to the fact that it is manifested by a difference between complex materials and conditions relevant to practical catalysis and the far simpler model systems that are often used to extract fundamental information. This gap has been addressed by the development of a host of in-situ/operando spectroscopies and transient kinetic techniques. These approaches seek to extract fundamental information under more relevant conditions by using surface-sensitive techniques and/or transient modifications of the reaction conditions. The in-situ, operando, and transient kinetic techniques provide the most insight, but they also require the most time/resources and the most complex analysis. The field of informatics has the potential to automate and accelerate the analysis of complex data from these "gap bridging" techniques, as well as the extraction of information at the macro and micro-levels.

2.2.1 Macro-scale informatics

Catalyst design takes place in an extremely high-dimensional space, even at the macro-scale. Catalyst materials may contain numerous elements, and additional variables such as support material, particle size, dispersion, and loading – this results in a combinatorial explosion of possibilities. High-throughput testing typically focuses on the composition of catalyst materials, and seeks to test as many catalyst compositions as possible.[119–122] This is a highly efficient version of the Edisonian trial-and-error approach through which catalysts have been developed for over a century,[49,50,52] and requires specialized setups (e.g. parallel reactors or printed catalysts) and rapid approaches for monitoring activity (e.g. thermal or FTIR product detection).[119–121,123,124] Needless to say, high-throughput combinatorial studies generate a tremendous amount of data, which presents an informatics challenge. For this reason, high-throughput studies were early adopters of machine learning and data science techniques, and several informatics frameworks have been proposed to manage high-throughput catalytic data.[2,3,66,125–127] Reviews of high-throughput catalyst testing have been published previously,[128–130] so here we focus primarily on applications of informatics techniques for searching through the high-dimensional space and machine-learning ap-

proaches for identifying patterns in the data. Further, we note that while combinatorial methods provide a diverse search of composition, they typically only link this data to the most basic kinetic information. The search can provide information on what bulk compositions are active, but limited guidance as to why certain combinations perform better than others, unless they are coupled with micro-kinetic models as discussed further Sec. 2.3.

One illustrative example of combinatorial testing is the case of mixed-metal oxides with 5 transition-metal components there are over 150 million possible compositions. Duff et. al. designed a high-throughput scheme for testing mixed-metal oxide ethylene epoxidation catalysts capable of screening 10,000 compositions per day. Even at this extremely high rate of testing it would take over 40 years to explore the entire parameter space, the authors employed a genetic algorithm to guide the search.[98] This "evolutionary" approach to materials screening continues to be commonly employed in high-throughput searches,[131–135] and is an example of the importance of informatics approaches to high-throughput experiments. However, the authors point out that screening at this rate is highly susceptible to "false negatives", due to lack of control over the exact composition of the active sites that are synthesized/tested in this manner. Therefore, combinatorial approaches are effective for identifying promising areas of activity, but are a less effective means to exclude areas of inactivity.

High-throughput experimental data lends itself well to regression models, since the measurements are taken under nearly identical conditions with only controlled variables changing. Typically the variables are related to catalyst composition, but high-throughput testing can also be used to rapidly probe the effect of reaction parameters (e.g. temperature) or materials parameters (e.g. particle size) for a fixed catalyst composition,[139] providing information on both material and environment. Researchers were utilizing artificial neural networks (ANNs) to establish quantitative [process/structure+environment]-property relationships for zeolite and oxide catalysts as early as the 90's. This approach continues to be employed in numerous studies.[12,60–64,135,140–148] These ANN models act as a black box that connects an input space to an output space, essentially an extremely flexible non-linear regression model capable of quantifying complex relationships in high-dimensional spaces; however, they are reliable only for interpolation, and it is often difficult to obtain insight from the resulting model.[149]

In addition to high-throughput data it is possible to extract large amounts of data from the literature to search for patterns. The group of Yıldırım has used decision trees to extract insight from thousands of literature data points for dry reforming of methane,[150] biodiesel production,[151] water-gas shift,[152] and CO oxidation.[153–155] The decision tree approach leads to interpretable models that provide insight into the most important macroscopic variables for catalyst processing and operating conditions. Another statistical method that is common in the analysis of macroscopic catalysis data is partial least squares (PLS) and principal component analysis (PCA).[65,156–161] These approaches seek maximize variance (PCA) or covariance (PLS) while reducing the dimension in order to classify catalysts or establish quantitative relationships between input and output variables.[162] These techniques are slightly more transparent than ANN's since they provide direct insight into the relative importance of various input variables, however they are also linear models and are hence less powerful in describing complex relationships. Ultimately, ANNs, PCA, PLS, and other statistical classification and regression models are valuable tools for extracting information from data, but their ability to generalize beyond the input data is limited unless they are connected to kinetic models (see Sec. 2.3).

2.2.2 Micro-scale informatics

The foundation for the extraction of catalytic knowledge is availability of micro-scale information about the rates of interaction between the catalyst surface and products, reactants, and intermediate species. Surface science experi-

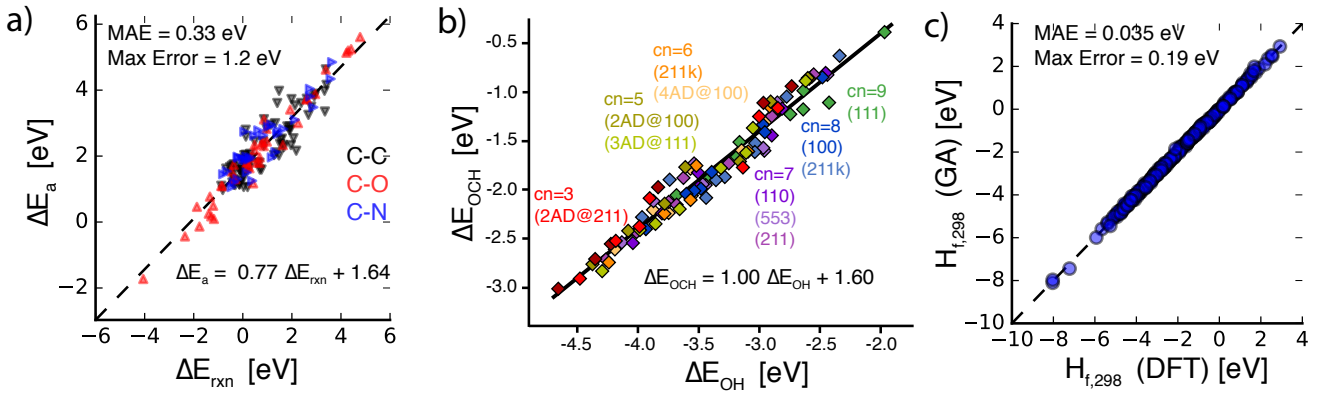

Figure 4: Illustration of Brønsted-Evans-Polanyi relationships for transition-metal (211) surfaces (a) (reproduced from data in Ref. 136), scaling for OCH* species on other transition-metal facets based on generalized coordination number (b) (reproduced from data in Ref. 137), and prediction of complex furanic compounds on Pd(111) using group additivity (c) (reproduced from data in Ref. 138).

ments on well-defined surfaces provide a valuable source of such information. These techniques include ultra-high vacuum (UHV) experiments such as XPS, LEIS, LEED, TPD, TPR, MBS, and others that provide insight into the structure and energetics of a catalyst surface.[101] The use of informatics in the experimental surface science community has been far less prevalent than that observed with macroscopic characterization techniques.

One area similar to combinatorial catalyst testing is high-throughput surface science.[163] These experiments provide surface-specific information as a function of bulk alloy composition, and hence provide more direct insight than macro-scale combinatorial approaches. For example, Boes et. al. utilized high-throughput surface science in conjunction with DFT calculations in order to understand the bulk-composition-dependent hydrogen adsorption energy for Cu/Pd alloys.[164] Collections of benchmarks for adsorption energies and transition-state energies for several adsorbates on a range of different transition-metal surfaces were also recently published.[42,165,166] This experimental information provides important verification of the accuracy of widely-used computational approaches. Another example of informatics in experimental surface science is the use of sophisticated regression models to extract more accurate binding energies from TPR results, where it was shown that results are significantly improved with the proper choice of objective function.[167]

There is also considerable micro-scale information in image data, an area where machine-learning has substantial potential as demonstrated by recent successes in other fields such as disease diagnosis.[168] The microscopy community has begun applying machine-learning techniques to classify and analyze image data with great success.[169–172] In particular, automated analysis and tracking of surface defects and surface structures from atomic-scale microscopy data has the potential to provide accurate statistical information about the dynamic nature of catalyst surfaces.[171] These techniques may also be applicable to other types of catalysis data such as 2D spectra that can be represented as images. Although these developments have not been widely applied to catalytic materials, they have significant potential to provide valuable information such as the nature and prevalence of defect sites on model surfaces and the resulting impact on kinetic function. The multi-scale nature and close coupling between the data and kinetics leads to many experimental approaches that bridge the materials gap and/or are linked to kinetic models, which will be discussed in Sec. 2.2.3 and 2.3 respectively. Though the experimental space for microscopic measurements may be smaller than

that presently achieved with macroscopic combinatorial screening, experimental micro-scale catalysis data can provide significant progress toward knowledge of *why* certain materials perform better.

One major route to obtaining atomic-scale insight in catalysis is electronic structure theory, most commonly based on DFT. Computational models can provide detailed insight into the energetics of molecules binding at surfaces, and these quantities are critical for understanding the [structure-environment]-property linkages in catalysis.[4,69,70,173,174] However, DFT calculations are also computationally expensive, and this cost can become prohibitive in the context of high-throughput studies. Hence, numerous methods based on physical, chemical, and data-driven models have been explored to develop quantitative linkages between the molecular/electronic structure of catalyst surfaces and the associated adsorption energy of intermediate molecules.

A classic example of a physically-derived correlation between catalyst electronic structure is the *d*-band model[175,176] and associated adsorption energy "scaling relations".[177] These linear correlations between the binding energies of various intermediates drastically reduce the number of calculations needed to predict catalyst activity, and have spurred the growth of computational catalyst screening studies that will be discussed later (Sec. 2.3.2). While these original correlations were physically-derived only for transition-metal surfaces, many other linear and non-linear relationships between adsorption energies and electronic structure parameters have been developed for a wide range of materials classes.[178–182] These "descriptor-based" approaches predict adsorption and/or transition-state energies based on a few easily-obtainable inputs such as electronic structure parameters,[175,178,183–189] (generalized) coordination numbers,[137,190–194] or other adsorption energies;[136,177,179,180,195,196] several examples are shown in Fig. 4a-b. They are typically semi-empirical, and increasingly employ machine-learning techniques.[14,17,197] Recent work has also demonstrated that statistical analysis can be used to identify optimal descriptors.[198] These descriptors are based primarily on the properties of the catalyst surface/material, and the parameters of the model vary with as the adsorbate changes. For small molecules the model itself is relatively simple, but for larger molecules and/or more complex active site geometries group-additivity[28,89,138,199–204] (Fig. 4c) and bond-order conservation[205–207] approaches can be used. Other studies have utilized informatics techniques such as compressive sensing,[208,209] graph theory,[16] and regression[210] to accelerate the process of constructing and parameterizing models for treating lateral adsorbate-adsorbate interactions. This combination of techniques provides a route to utilize known data of binding energies of an adsorbate on a number of catalyst surfaces in order to rapidly predict the binding energy of the same (or similar) adsorbate on new materials, different active sites, or different coverage conditions.

An alternative approach to accelerating the calculation of adsorption energies is to use the molecular structures as inputs rather than descriptors. The atomic structure can be "fingerprinted" using numerous techniques that quantify the local environment of individual atoms[11,31,212,213] or properties of the entire system.[214] These fingerprints serve as inputs to regression models such as neural networks, kernel ridge regression, or Gaussian process regression. Neural network models have been shown to yield highly accurate predictions with errors <5 meV when trained to results from thousands of DFT calculations.[31] These machine-learned models can effectively serve as force-fields for molecular dynamics simulations, where the flexibility of the model is able to reproduce the complex quantum-mechanical phenomena that govern the reactions of molecular species at surfaces and have been used successfully to study bond breaking,[215–217] solvent effects,[218–220] support/particle effects,[211] and segregation/reconstruction[220,221] in metallic and oxide systems at time and length scales that are impractical for DFT.[222] An example of the architecture and accuracy of a an atomistic neural network for Cu particles on ZnO is shown in Fig. 5. A comparison of neu-

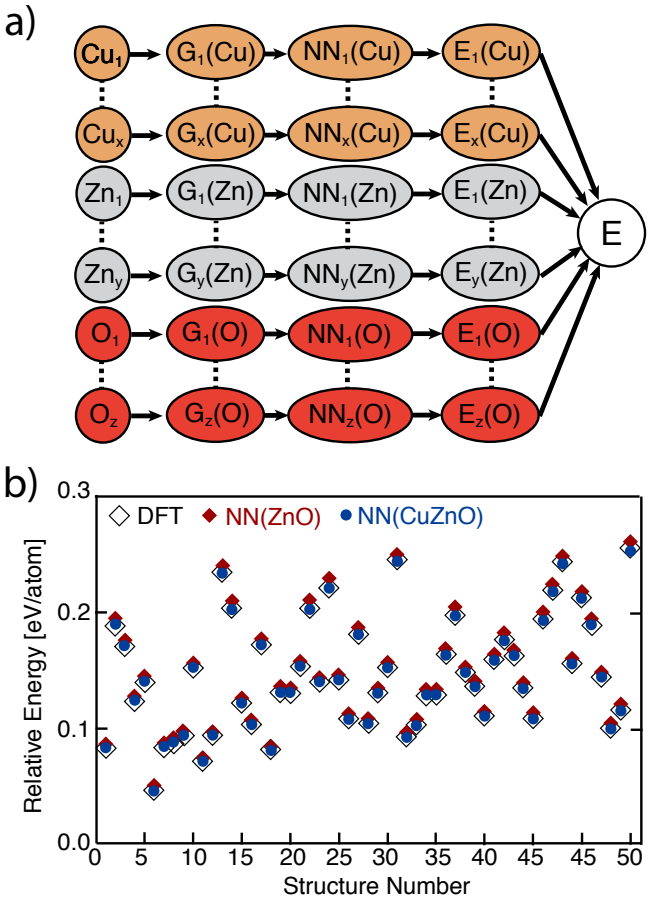

Figure 5: Schematic of atomistic neural network architecture (a) and results for Cu particles on a ZnO support compared to DFT (b). Reproduced from data in Ref. 211

.

ral network force-fields and the ReaxFF[9] force-field for the properties of Au clusters demonstrated that neural networks are more accurate, but also require more computational resources both in training and evaluation.[223] Hence, neural network force-fields provide an intermediate step between DFT and ReaxFF in the accuracy/cost tradeoff. An alternative approach is to use probabilistic models such as Gaussian process regression (GPR) or Bayesian inference to predict adsorption energies.[13,214,224] These approaches are typically somewhat less accurate, but also require significantly less training data (10-100 DFT calculations) and include built-in uncertainty estimate that can be reduced by the addition of targeted training data and used in search procedures to identify stable surfaces, active site motifs, and reaction mechanisms, as discussed further in Sec. 2.3.1.

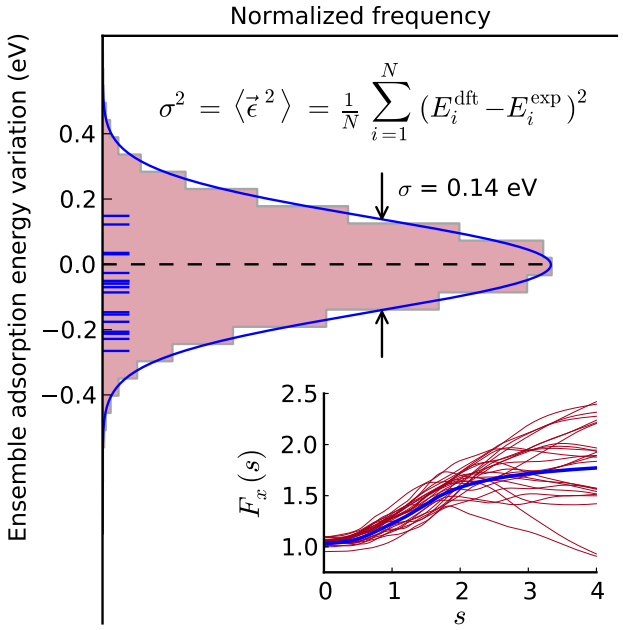

Figure 6: The predicted errors in 17 adsorption energies for BEEF-vdW (blue horizontal lines) and predicted uncertainty (shaded Gaussian). The corresponding ensemble of exchange-enhancement factors is shown in the inset. From Ref. 225. Reprinted with permission from AAAS.

Another issue with micro-scale data obtained from DFT is that the treatment of exchange-correlation requires non-systematic approximations, making it difficult to assess the reliability of DFT models. Empirically the accuracy of adsorption energies is found to be on the order of 0.2 eV when compared to experiment,[42] but these comparisons (and the accuracy of the experimental numbers) are limited by the challenging nature of adsorption energy measurements. A lower bound of the cumulative error of all intermediate steps in a reaction pathway can be obtained by comparison between calculated and experimental gas-phase reaction energies, which are well-known.[112] In several cases this results in rather large errors that must be corrected in order to obtain reliable reaction equilibrium behavior. For example, the oxygen reduction/evolution reaction energy is off by ∼0.7 eV for most GGA functionals due to poor treatment of the triplet state of O_2, so the experimental value is typically used

instead.[226] A similar error is observed for the CO_2 reduction reaction, and a statistical analysis was utilized to identify that this error was consistent for molecules involving the O=C=O backbone.[227] An alternative to correcting these errors in an ad-hoc fashion is to quantify the uncertainty in the adsorption energy. This can be achieved by fitting probability distributions to the results from various functional approximations,[228,229] or in a more systematic way by mapping uncertainty back to the parameters of the functional. The latter approach has become particularly prevalent in the catalysis community due to the development of the "Bayesian Error Estimation Functional" (BEEF).[230–232] This family of functionals is developed specifically for surface science[42,166,230], and utilizes an ensemble-based approach to propagate uncertainty from the parameters of the GGA model to calculated energies (Fig. 6); importantly, all of these approaches capture correlations in error between various binding energies which have been shown to be important for uncertainty propagation.[225,229] Furthermore, the combination of error ensembles from BEEF functionals has been combined with a statistical analysis of reaction energy errors to establish simple but accurate correction schemes for the adsorption energies of molecules containing the O=C=O backbone,[43] and a similar ensemble-based approach has been developed to quantify uncertainty in predictions from neural network potentials.[233] These examples illustrate the utility of statistical and informatics approaches in improving the accuracy and assessing the reliability of information derived from electronic structure theory simulations.

2.2.3 Bridging the gap with informatics

The most powerful type of catalytic information is derived from data obtained from methods that effectively bridge the gap between micro- and macro-scale information. These techniques extract surface-sensitive micro-scale information from complex catalysts (e.g. supported nanoparticles) and conditions (e.g. high temperature and pressure) that approach (i.e. in situ), or are equivalent to (i.e. operando), practical operating conditions.[35] This can be achieved by probing the catalyst surface using photons/electrons/X-rays (spectroscopy/microscopy),[35,234–238] reactant/intermediate molecular species (transient kinetics),[239–242] or both simultaneously (modulation excitation spectroscopy).[243,244] These advanced techniques produce rich information about the nature of the reaction mechanism and active site, but also require significant overhead in the form of complex reactor design and sophisticated mathematical/statistical analysis to convert data to information. The recent developments in data science provide many open opportunities to more efficiently extract and store information from these valuable experimental approaches. There have been several applications of machine-learning approaches to analyze data from several spectroscopic/microscopic techniques including Raman spectra,[245] infrared spectra,[246] X-ray emission,[247] time of flight secondary ion mass spectra,[248] XANES,[249] and microscopy;[169,170,172,250] however, these novel approaches are not widely applied to catalysis studies. One exception is a recent study where in situ XANES measurements were used in conjunction with a neural network model in order to determine the structure of platinum clusters, as illustrated in Fig. 7.[15]

In the case of transient kinetics approaches, the two most common techniques are steady-state isotopic transient kinetic analysis[252–254] (SSITKA) and temporal analysis of products (TAP).[255–257] These approaches have required advanced mathematical, statistical, and/or physical models to extract information from raw data.[242,251,258,259,259,260] For example, a phenomenological analysis of the product distribution was developed to extract a "reactivity fingerprint" that can characterize a catalyst's kinetic response without the assumption of a kinetic model.[239] These "reactivity" quantities contain physicokinetic information and provide a basis of screening the active surfaces of a catalyst material. Furthermore, these fingerprints can be combined with learning models such as decision trees to determine appropriate kinetic models, effectively converting kinetic in-

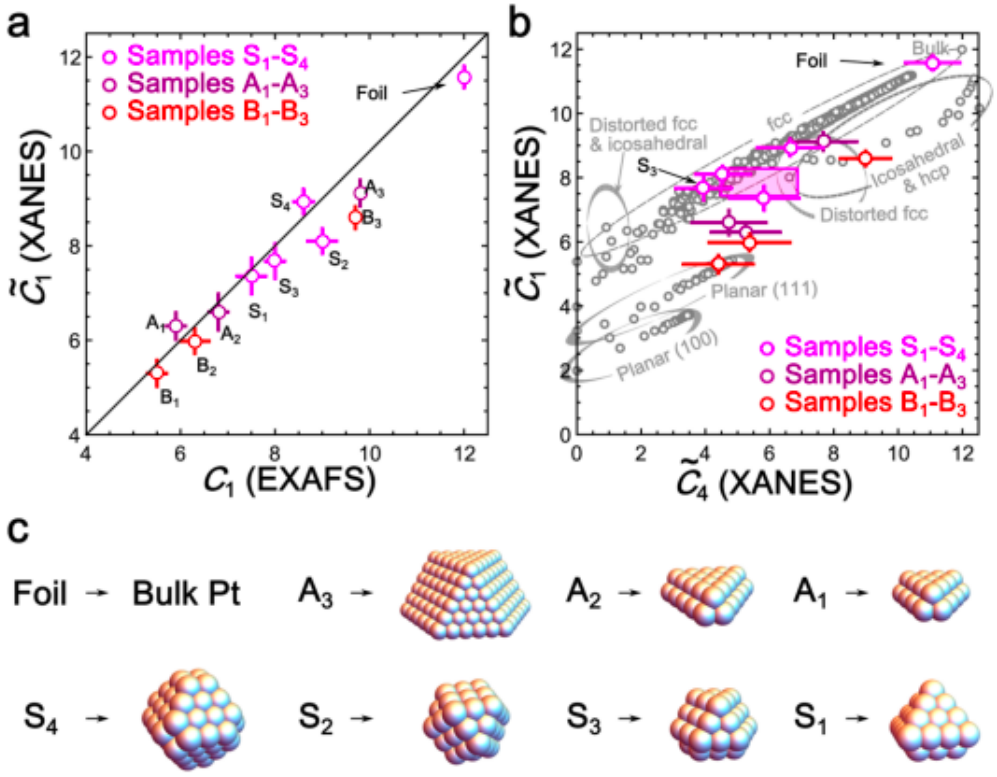

Figure 7: Neural network models were used to determine the structure of Pt clusters based on EXAFS data. Comparison of coordination number for the first shell shows agreement between neural nets and conventional analysis (a), and the neural net model is used to predict coordination numbers for the fourth shell (b) and assign 3D cluster models (c). Reprinted with permission from Ref. 15. Copyright 2017 American Chemical Society.

formation to knowledge in a systematic way.[261] Another example is the "Y-procedure" method that provides a route to convert primary pulse response exit flux data into information on the reaction rate, gas and surface concentration of reactants/products without the assumption of a kinetic model.[241,251,258] This presents an opportunity to exploit the transient experiment as a high-throughput screening tool of surface composition. The pulse response forces a change in the surface coverage of species and with the Y-procedure analysis the intrinsic rate constant can be extracted from any point in the rate, gas concentration, and surface concentration space. Figure 8 demonstrates such data for the oxidation of platinum. A similar exploration of this parameter space would be experimentally intensive and less precise in a conventional steady-state reactor system.

The technique of modulation excitation spectroscopy (MES) effectively combines operando spectroscopy and transient kinetics by measuring the spectroscopic response to a transient in reactant concentration.[243,244] This relatively new class of techniques has proven particularly useful in the field of heterogeneous catalysis because it is able to deconvolute the spectroscopic response of (irrelevant) spectator species and that of relevant reaction intermediates.[244,262] This is achieved by repeatedly pulsing reactant concentration and analyzing the spectral response; MES has been successfully applied to numerous surface-sensitive spectroscopies including XAS,[263] FTIR,[243] DRIFTS,[264,265] and PM-IRRAS,[266] and others[244,262] in order to elucidate active sites and reaction mechanisms. The MES approach combines the strengths of operando spectroscopy and transient kinetics to provide substantial insight into the catalytic process, but it also inherits the complexity of both approaches. The reactor setups for MES are typically more elaborate than operando re-

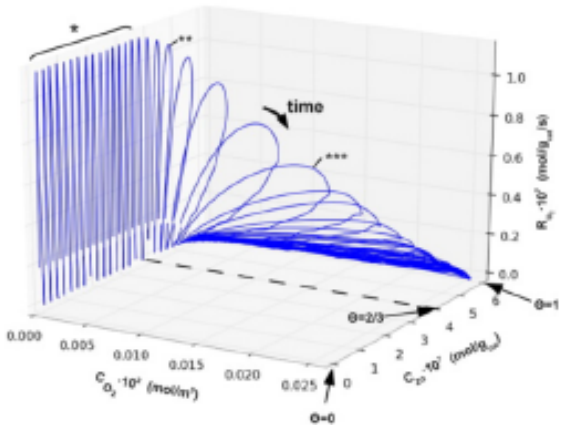

Figure 8: Intrinsic rate/concentration space calculated from transient experiments of Pt oxidation. Reprinted from Ref. 251, Copyright (2011), with permission from Elsevier.

actors, and the resulting datasets are complex time-resolved spectra that require involved mathematical analysis that is often a barrier for implementation.[244] There has been relatively little effort to apply emerging data science techniques to these rich datasets, although enhanced fitting procedures have been used to improve the quality of information extracted from EXAFS MES data.[267] Utilizing data science techniques to systematically manage MES data and extract information is a promising future direction in catalysis informatics.

2.3 Knowledge

The ultimate goal of catalysis informatics is to extract generalizable knowledge from raw data. The chemical master equation and microkinetic models are the mathematical embodiment of the catalytic reaction mechanism and energetics of surface reactions, providing a common framework for understanding a catalytic reaction as a function of reaction conditions and catalyst composition.[66] The Markovian "master equation" of chemical kinetics is given by:

$$\frac{\mathrm{d}P(S_i)}{\mathrm{d}t} = \sum_j \left(A_{ji}P(S_j) - A_{ij}P(S_i)\right) \qquad (1)$$

where S_i is a state of the surface, $P(S_i)$ is the probability of observing state i, and A_{ij} is a transition matrix containing the probability of a transition between state i and j. The simplicity of this equation is deceiving, and the devil lies in the details of the definition of a state S_i, since the number of possible states grows exponentially with the size of the surface being modeled.[86,268] Due to this high dimensionality the master equation is typically solved stochastically through kinetic Monte Carlo algorithms,[268,269] or deterministically through mean-field approximations;[72,89] in this work we use the term "micro-kinetic model" generically to refer to any approximation/algorithm used to solve the master equation. Conceptually, the transition-probability matrix corresponds to the rate constants of various elementary steps, and the probability of observing a state corresponds to the concentration of different chemical species. This correspondence is clearer in the common mean-field approximation where the probabilities are assumed to be related to concentrations:

$$r_i = k_i^+ \prod_r C_r - k_i^- \prod_p C_p \qquad (2)$$

where r_i is the rate of elementary step$i$, $k_i^\pm$ is the forward/reverse rate constant for elementary step i, and $C_{r/p}$ are the concentrations of reactant/product species. In order to obtain the rate of change for a concentration C_j of species j the rates of relevant elementary steps must be summed:

$$\frac{\mathrm{d}C_j}{\mathrm{d}t} = \sum_i s_{ij} r_i \qquad (3)$$

where s_{ij} are stoichiometric coefficients for the number of j molecules in elementary step i. Combining Eqs. 2 and 3 yields:

$$\frac{\mathrm{d}C_j}{\mathrm{d}t} = \underbrace{\sum_i s_{ij}}_{\text{mechanism}} \left(\overbrace{k_i^+}^{\text{active site}} \underbrace{\prod_r C_r}_{\text{mechanism}} - k_i^- \prod_p C_p \right) \qquad (4)$$

From this mean-field version of the master equation it is clear that the micro-kinetic model is an embodiment of the active site and reac-

tion mechanism of a catalytic reaction, with rate/equilibrium constants corresponding to the intrinsic properties of the active site, and the structure of the equation arising from the reaction mechanism. While Eq. 4 is a significant simplification the conceptual correspondences hold regardless of complexity. For example, coverage effects could be accounted for by letting rate constants depend on concentration ($k_i^{\pm}(C_j)$); in this case, the relevant "active site" may include adsorbates other than the product/reactant in addition to the catalyst material itself. This illustrates the key role of micro-kinetic models as quantitative representations of catalysis knowledge.

Micro-kinetic models are central to both computational and experimental approaches, and ultimately have the ability to both explain *why* catalysts function and predict *what* the catalytic function of a new material will be, as shown schematically in Fig. 9. In this section we review the successful use of kinetic models in both explanative and predictive contexts, with a focus on approaches where data science and informatics has been used in conjunction with more traditional kinetic modeling approaches. We conclude by revisiting the concept of a catalysis "knowledge engine" that integrates various types of catalysis data,[66] and discuss how recent advances in machine learning and informatics may help advance this concept.

2.3.1 Explanative Approaches

Explanative statistical models seek to identify a causal explanation for an observed phenomenon.[270] In the context of catalysis, we consider the micro-kinetic model to be the causal explanation, and informatics approaches seek to link this model to micro- and/or macro-scale information in a quantitative way. Once a model and its parameters have been identified, the model may become predictive since the behavior at different reaction conditions can be predicted and tested; however, in this section we focus on approaches that seek a causal explanation for the catalytic behavior of a material with a static chemical composition. More specifically, explanative approaches seek to *identify the relevant reaction mechanism(s) and active site(s)* for a given catalyst material under a range of relevant operating conditions. In quantitative terms this amounts to constructing the mathematical form of the kinetic model and estimating the relevant parameters (see Eq. 4). In practice, two basic strategies are used to achieve this goal: "top down" and "bottom up". Here we define "top down" approaches as those that seek to establish the model and estimate parameters based on macro-scale reaction kinetics data from complex catalysts, while "bottom up" approaches utilize micro-scale surface science information from well-defined systems (we note that this differs from an alternative definition based on the complexity of the kinetic model rather than the catalyst systems[261]). Ultimately most approaches seek to bridge the gap between micro- and macro-scale descriptions of the kinetic process, but this is typically achieved in a qualitative manner. For example, intuition or information about micro-scale

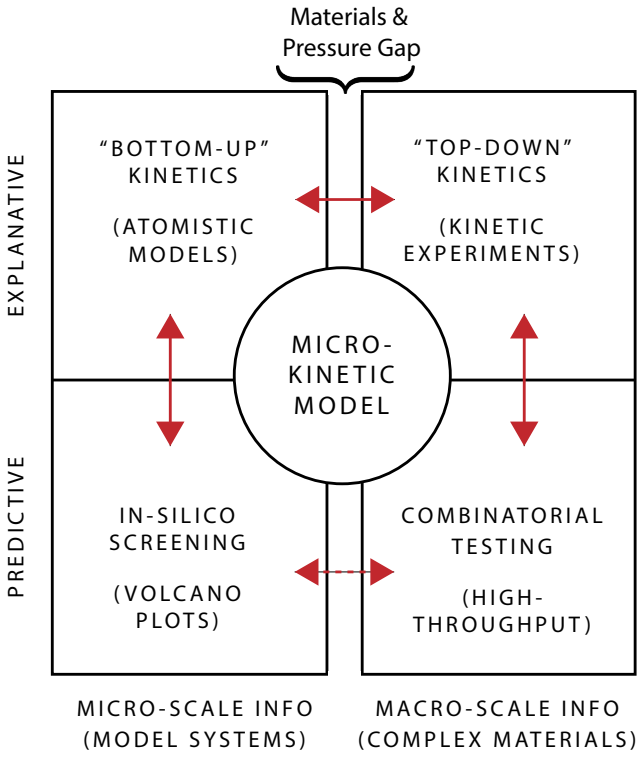

Figure 9: Classification of computational/experimental and predictive/explanative approaches illustrating the central role of the micro-kinetic model.

processes may inform model construction in a top-down approach, and qualitative comparison to macro-scale kinetics is critical to assess and validate bottom-up models. However, some approaches do seek to quantitatively bridge the gap,[167,271–273] often with the use of statistical models; this is an emerging and important area for catalysis informatics.

The idea of constructing and parameterizing kinetic models is far from new. Kinetic models have been central to catalysis from the early days, including power-law kinetics with no molecular insight, Langmuir-Hinshelwood-Hougen-Watson (LHHW) models that require some chemical intuition, and micro-kinetic models that link directly to elementary processes.[72,89,274] The most basic realization of a kinetic model is the Arrhenius equation, which is routinely parameterized using Arrhenius plots. More advanced approaches utilize non-linear regression to establish the parameters of LHHW and/or micro-kinetic models. These parameter estimation approaches could be considered early informatics approaches, but have been covered extensively in previous reviews and texts.[72,101] Similarly, the use of computational methods (primarily DFT) to provide energetics of elementary processes to parameterize kinetic models from the bottom up is well-established.[173,174] Here we focus on extensions of these approaches that include statistical models and/or uncertainty quantification.

Reaction mechanism: The first step to establishing a micro-kinetic model is determination of the set of differential equations from the underlying chemical reaction mechanism. This step is far from trivial, as reaction networks grow exponentially with the number of atoms in the largest molecule.[28,89] This complexity is exacerbated for surface reactions, where the possibility of multi-site mechanisms and coverage effects can lead to even more possibilities.[41,86] In practice, the reaction mechanism is often determined from a combination of qualitative analysis of experimental data and chemical intuition; however, as reaction mechanisms and catalyst surfaces become more complex these intuitive approaches become impractical so automated, systematic techniques must be employed. There are two general approaches to systematic mechanism construction: local and global. Local approaches start from a set of reactants and iteratively construct, parameterize, and solve the differential equations corresponding to all possible next elementary steps. The results are used to identify the most relevant steps, and the algorithm iterates forward through the reaction network, growing a "core" of relevant reactions and discarding all reactions that are deemed improbable. This approach has been used extensively and successfully in the combustion chemistry community[28,88,275] and a version of the Reaction Mechanism Generator (RMG) has recently been adapted to surface reactions suggesting that this is a promising route to model construction.[276] These approaches inherently rely on data science methods since elementary step enumeration and model parameterization happens on the fly. The determination of elementary steps uses graph-theory and the rapid parameterization requires databases of known parameter values as well as models for quickly predicting the values of unknown parameters.[28] Local mechanism construction approaches are efficient because they avoid irrelevant regions of the reaction network, and are advantageous because they deliver a simplified and parameterized model that can be further refined using additional experimental or computational data. The main disadvantage of local model generation is the reliance on accurate on-the-fly parameterization. Inaccurate parameters can lead to incorrect mechanisms, and the unavailability of accurate adsorption and transition-state energies for surface reactions makes this particularly challenging for heterogeneous catalysis.

The global approach to model construction avoids this problem by first considering all elementary steps connecting reactants to products and subsequently relying on parameterization to simplify the model. Global approaches are more comprehensive, but they are also less efficient since they sample all possible elementary steps rather than just the most probable ones, leading to a combinatorial explosion of complexity as the size of the reaction net-

work increases. This can result in networks with $> 10^4$ total reactions, corresponding to a similar number of differential equations and parameters which are expensive or intractable to accurately parameterize and solve. Informatics approaches can aid in this complexity reduction problem; for example, Ulissi et. al. showed that a combination of a simple kinetic model with parameters obtained from machine-learning and uncertainty quantification can significantly reduce the effort needed in model simplification.[13] Another strategy is the use of rule-based approaches that embed chemical intuition into the model generation process. An example is the Rule Input Network Generator (RING) software[277–279] that utilizes graph theory to efficiently generate reaction mechanisms for complex chemistries based on user-defined rules. There are numerous other approaches that have been developed for construction and analysis of reaction networks, commonly based on graph theory and/or analogies to electrical circuits;[280–290] however, these techniques are not widely used in practice, possibly due to their mathematical complexity or practical challenges in implementation. Catalysis informatics can help to improve and unify these different techniques by providing more accurate parameter estimates through databases (see Sec. 2.1), regression models (see Sec. 2.2) and integration of these estimates with existing mathematical techniques for reaction network generation/analysis to establish data-driven rules for automatic mechanism generation.

Kinetic parameters and active sites: Once a reaction mechanism and corresponding micro-kinetic model have been identified the next step is determining the values of the model parameters and simplifying the model as necessary. The main challenge is the large number of parameters that arise from the reaction energy (equilibrium constant) and transition-state energy (rate constant) of each elementary step. This can result in models with $> 10^2$ parameters, but the kinetic behavior is typically controlled by only a few rate-limiting steps.[89,271,291] Thus the goal is not only to find

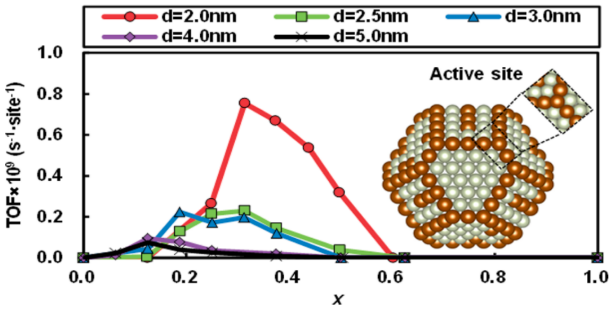

Figure 10: Neural networks combined with a micro-kinetic model are able to predict effects of particle size and composition for RhAu bimetallic particles for the NO decomposition reaction. Reprinted with permission from Ref. 19. Copyright 2017 American Chemical Society.

the parameters of the model, but also identify which ones are truly meaningful and which ones are irrelevant to the catalytic behavior.[89,292] This amounts to determining an optimum-complexity kinetic model, which is analogous to the concept of an optimum-complexity regression model from data science and statistics.[78] The two main strategies for model parameterization are "top down" models that rely on regression to macro-scale experimental data of catalyst performance and "bottom-up" models that utilize results from atomic-scale computational results to determine parameters. The top-down approach is advantageous in its ability to accurately reproduce experimental results measured on real catalytic systems; however, the large number of fitting parameters makes regression highly susceptible to over-fitting, particularly for steady-state kinetics where the behavior will be controlled by only a few parameters.[89,293] Statistical techniques such as latent variable analysis (based on PLS, see Sec. 2.2) have been employed to systematically assess the appropriate number of parameters[294] and reduce model complexity.[295] An alternative approach is to utilize genetic algorithms or Bayesian probability theory to estimate the uncertainty on fitted parameters,[67,296,297] effectively determining which parameters are well-determined based on the experimental data; this has the additional advantage of providing an error estimate on the computed quantities.

The other disadvantage of the top-down approach is the fact that it does not provide any direct insight into the relevant active site(s).

The bottom-up approach overcomes this disadvantage by utilizing atomic-scale models and quantum-mechanical techniques such as DFT in order to parameterize a kinetic model from first-principles.[82,174] However, the bottom-up approach relies on an accurate active-site model, and can also be very expensive since each parameter requires a DFT calculation; for models with $> 10^2$ elementary steps this is effectively intractable. This is exacerbated by the fact that there are a semi-infinite number of possible surface configurations of a given material, leading to a challenge in systematically determining the atomic-scale structure of the relevant active site(s). In general a number of approaches have been explored for sampling active-site configurations including genetic algorithms,[299] (constrained) minima hopping,[300,301] and comparison to experiment.[15,302] Recently, machine-learning techniques have demonstrated the ability to accelerate this process by rapidly estimating DFT energies (see Sec. 2.2), and these techniques have been integrated with various search algorithms to determine the relevant active site and mechanism. Several recent reports have utilized machine-learning models along with thermodynamic stability of nanoparticle surfaces to assess the activity of bimetallic particles for CO_2 reduction[11] and NO decomposition[19] (Fig. 10). Machine-learning potentials have also been used to accelerate genetic algorithm searches for the active site structure of platinum clusters,[303] and coupled with a Bayesian search algorithm to efficiently identify active site structures and adsorbate configurations.[224,304,305] Informatics approaches have also been employed to rapidly identify active sites in nanoporous materials such as zeolites and MOFs. These materials have the additional challenge that the bulk framework structure can take many possible forms. These frameworks are often tabulated in databases (see Sec. 2.1), which can be analyzed to identify promising active sites. For example, Matsuoka et. al. applied a combination of descriptor-based techniques, inexpensive force-fields, and DFT to identify strong Brønsted acid sites in a database of over 500,000 zeolite structures.[306] Evolutionary algorithms have also been applied to accelerate the determination of zeolite structures,[127,307] providing a possible route to determine active site structures without relying on databases. These examples suggest that the numerous advances in machine-learning for adsorption energy prediction (see Sec. 2.2) combined with search and optimization algorithms will enable comprehensive determination of active-site structures for bottom-up models in the future.

A further disadvantage of the bottom-up approach is that the estimates of adsorption and transition-state energies are not perfectly accurate, even when they are computed with the proper active site model, due to deficiencies in DFT. This necessitates the development of strategies to quantify the uncertainty (see Sec. 2.2) and propagate it through kinetic models to assess the reliability of computed rates. For example, Medford et. al. showed that correlations in error between DFT binding energies leads to a cancellation effect such that the error on computed rates is lower than would be naively expected from the error on DFT energies,[225] and similar results were obtained independently by Sutton et. al.[271] Nonetheless, agreement between bottom-up kinetic models and experimental results is rarely quantitative, even for the most advanced models,[308] likely due to the numerous simplifications and assumptions that are required in bottom-up models and the inaccuracy of underlying quantum-mechanical approximations. This necessitates the use of combined bottom-up/top-down approaches that seek to provide atomic-scale models that are quantitatively consistent with experimentally observed results. One approach to achieving this is to use bottom-up parameters as initial estimates for top-down regression models, as illustrated by the results of Grabow et. al. where a bottom-up model for methanol synthesis over a Cu catalyst was refined through top-down regression yielding excellent agreement with experiment.[273] Recently, the groups of Vlachos[271] and Heyden[228] have utilized uncertainty quantification to assess the

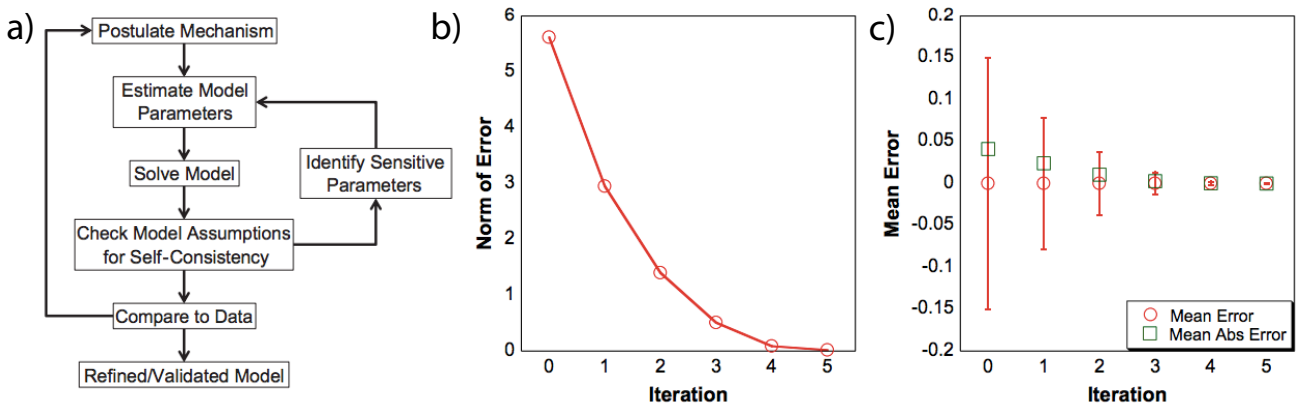

Figure 11: Schematic of iterative model refinement for automated micro-kinetic model construction and convergence of the approach for ethanol steam reforming on Pt. Reprinted from Ref. 298, Copyright (2015), with permission from Elsevier.

probability that bottom-up micro-kinetic models explain observed experimental results. Recent work by Sutton et. al. integrated parameter estimation, first principles modeling, sensitivity analysis, and experimental data in order to systematically refine and parameterize a complex multi-scale micro-kinetic model of ethanol steam reforming on platinum catalysts (Fig. 11).[298] This is a key example of how informatics approaches can distill experimental and computational data into knowledge about a catalytic process.

2.3.2 Predictive Approaches

Predictive statistical models produce testable estimates of new or future observations.[270] In the context of catalysis this could mean the prediction of catalytic behavior under different reaction conditions, or the catalytic performance of a new catalyst material. The former type of prediction arises naturally from the micro-kinetic model, as discussed in the prior section; furthermore, predictive approaches that do not make use of a kinetic model are discussed in Sec. 2.2. In this section we will focus on approaches that utilize a micro-kinetic model to *predict new catalytically active or selective materials*. This predictive power is typically predicated on an accurate explanation of the catalytic mechanism and active site, and the ability to rationally predict/discover catalytic materials provides strong validation of the underlying explanative micro-kinetic model. Hence, the two approaches are intimately related and should be seen as complementary.

The task of predicting catalytically active/selective materials from micro-kinetic models is challenging because it requires the simultaneous prediction of multiple materials properties (adsorption energies) that interact in a complex and active-site-specific way to control catalytic behavior.[309] This is in contrast to many other materials design problems such as thermodynamic stability, solar energy, or topological insulators which depend on relatively few properties that can be derived directly from the electronic structure of the bulk material.[310] This complexity suggests a role for informatics approaches; indeed multiple examples have shown that a combination of computational simulations, experimental measurements, statistical correlations, and physical models can guide the discovery of active and selective catalytic materials.[70,181,309]

The most common approach to prediction/discovery of catalysts is the use of descriptor-based micro-kinetic models which leads to the well-known "volcano plots" developed by the Nørskov group.[82,311] While this approach is not traditionally considered "informatics", the use of (linear) correlations between adsorption energies is at the core of its success. These adsorption-energy scaling relations, and similar approaches as discussed in Sec. 2.2, serve to reduce the dimensional-

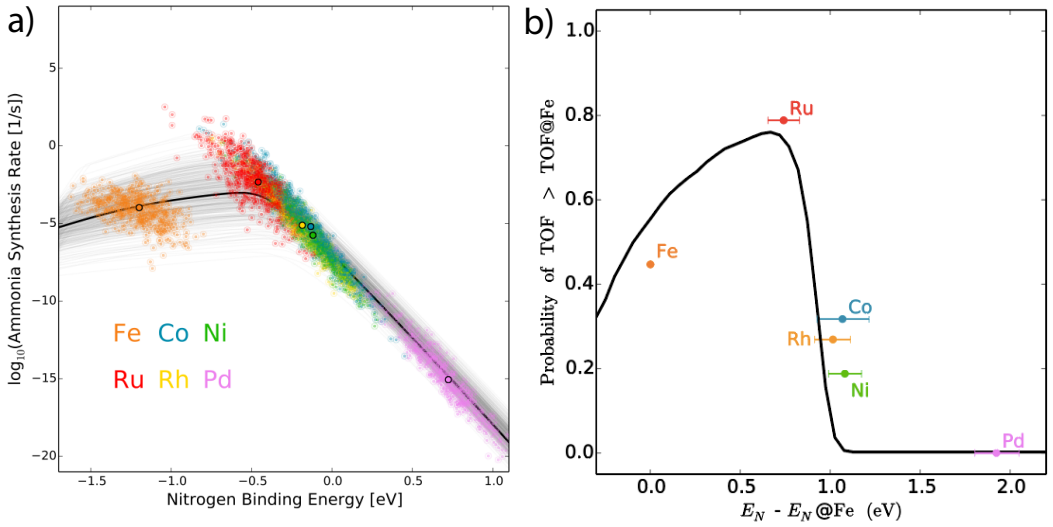

Figure 12: Ensemble of predicted rates and volcano curves for ammonia synthesis over stepped transition-metal surfaces based on propagation of BEEF-vdW energy ensembles (a). Probability that a catalyst will have activity higher than Fe based on statistical analysis of ensembles (b). From Ref. 225. Adapted with permission from AAAS.

.

ity of the large parameter space that underlies micro-kinetic models. This combination of physically-inspired correlations and physical models enables a projection of catalytic activity onto a low-dimensional "descriptor space" consisting of key adsorption energies or electronic-structure properties.[309] The low-dimensional space enables high-throughput screening and optimization of novel catalytic materials, enabling prediction of novel chemical compositions that will be active/selective for a reaction of interest.[309,311–313] The descriptor-based kinetic analysis is effectively a dimensional reduction algorithm designed specifically for the problem of catalysis, and is a prototype of knowledge-driven predictive approaches in catalysis informatics.

The use of volcano plots for catalyst screening is well-developed, and covered in multiple other reviews[70,181,309] and texts.[82,314] Here we review it briefly and discuss opportunities for integration of informatics approaches. An early example of the predictive power of the volcano plot is the discovery of CoMo catalysts for ammonia synthesis. These alloys were predicted by "interpolating" between the binding energies of Mo (too reactive) and Co (too noble) to find an optimum binding energy, and the predictions were experimentally verified.[55] Since then more sophisticated approaches have been developed based on multiple descriptors, and DFT calculations are typically used to screen new catalyst candidates rather than the interpolation principle.[309] While volcano plots were originally developed to predict catalytic activity the same descriptor-based approach has also proven successful for predicting selectivity.[229,315–318] Furthermore, the approach has been implemented in the `CatMAP` software tool.[40] This open-source informatics tool facilitates efficient integration of adsorption energy data, regression models for scaling relations, and micro-kinetic models.

One important role of catalysis informatics in the prediction of new catalyst materials is quantification of uncertainty and analyzing sensitivity of models. The construction of volcano plots requires many assumptions and approximations, and by quantifying and propagating uncertainty it is possible to provide statistical estimates of the confidence on predictions; this improves the robustness and reliability of the predictive models.[225,229,271,319,320] The quantification of uncertainty is closely related to sensitivity analysis, in that both assess the influence of changing parameters on model predictions;[271] however the "degree of

rate control" sensitivity analysis[291] has physical meaning and can be measured in principle. Recently a scheme was proposed for using degree of rate control to make new predictions.[321] Early work on uncertainty quantification showed that the predicted optimum for transition-metal alloy ammonia decomposition catalysts is more sensitive to the inclusion of adsorbate-adsorbate interactions than to perturbations in the adsorption energies.[319] More recently, correlations in underlying DFT errors have been shown to lead to cancellation of error, so that position of the predicted optimum of volcano plots is more precisely determined than the absolute DFT energies. This has been illustrated with probability distributions based on multiple DFT approximations[271] and via propagation of energy ensembles from the BEEF-vdW functional (Fig. 12a). The results can be interpreted in terms of a probability "volcano", illustrating that regardless of underlying error the volcano plots provide a way of establishing the region of descriptor-space where the probability of discovering an active catalyst is highest (Fig. 12b).[225] In addition, the influence of uncertainty in the linear scaling relations on predicted optima for activity and selectivity has been assessed for ethanol hydrodeoxygenation.[229] The results indicate that absolute error on predicted rate is rather large ($\pm$ 2 orders of magnitude), and the position of the optimum can vary substantially based on errors in linear energy scaling relationships, but that the region of maximum probability is consistent with the region of maximum activity (Fig. 13a-b). Furthermore, the selectivity is found to be very poorly determined if carbon binding is greater than zero (see Fig. 13c). This is consistent with other analyses of uncertainty on selectivity,[102,318] and indicates that quantitative selectivity predictions based on DFT calculations or volcano plots should be regarded as qualitative indicators of regions where the probability of high selectivity is maximal. Improving the rigor of uncertainty analysis in catalyst predictions is a key area where informatics will play a role, and the probabilistic interpretation of volcano plots presents an opportunity to integrate these predictions with other statistical methods such as Bayesian optimization.

Finally, we turn to the expert systems and knowledge engines that are arguably the first examples of catalysis informatics. The use of "expert systems" was popular in the 80's, prior to the rise of machine learning, and numerous attempts were made to use this approach to develop artificial intelligence systems capable of integrating information in order to discern catalytic mechanisms and design catalysts.[56–58,60] However, these approaches were ultimately unsuccessful because expert systems rely on a significant amount of human input and were not able to generalize findings to new systems or design catalysts, likely due to the previously mentioned challenges in featurizing catalysts and accounting for uncertainty in catalysis data. More recently, Caruthers et. al. proposed the development of a catalysis "knowledge engine" capable of converting kinetic data from high throughput experiments into a parameterized kinetic model.[66] Similar ideas have been proposed by others,[88,272] with the key idea being coupling of high-throughput experimental testing to micro-kinetic and surface science models. This is a promising approach to systematically converting data into knowledge; however, these systems are not publicly available, and significant advances have been made in the field of computational catalysis and machine learning since the early reports. Nonetheless, the concept of a "knowledge engine" capable of drawing from existing data, suggesting new data to be measured/computed, and ultimately providing a parameterized micro-kinetic model with uncertainty bounds on all parameters is a powerful informatics strategy. Furthermore, coupling such a system to the recent developments in descriptor-based catalyst screening can effectively "close the loop" of explanative, predictive, computational and experimental approaches (Fig. 9); the development of these "knowledge engines" an enticing goal for the field of catalysis informatics.

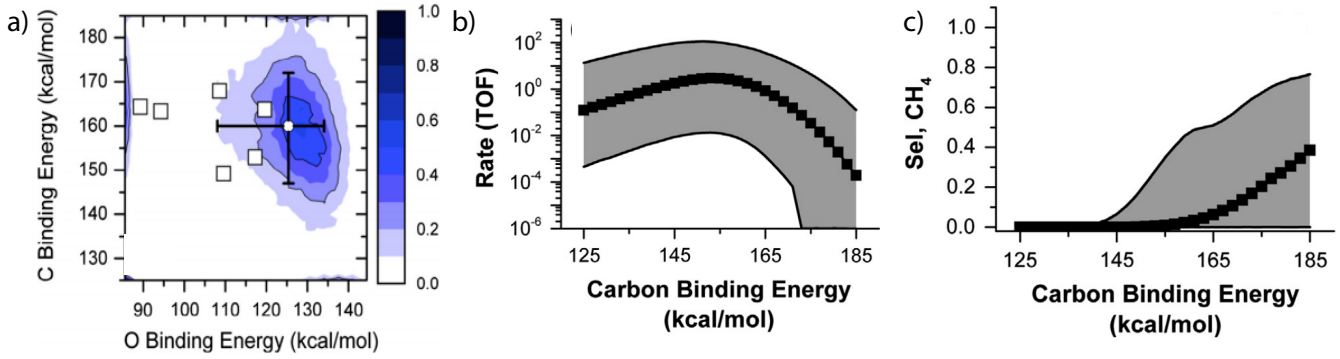

Figure 13: Uncertainty in position carbon and oxygen binding energy for ethanol hydrodeoxygenation maximum due to errors in adsorption energy scaling relations (a). Error in turnover frequency (b) and selectivity toward CH_4 (c). Adapted from Ref. 229

.

3 Future Opportunities

The future of catalysis informatics is bright due to recent advances in computational modeling, widespread availability of machine-learning models, and increasingly sophisticated data infrastructure. In this section we briefly discuss three areas where there is significant potential for development. These areas are: i) closer coupling of experimental data and computational models, ii) advances in micro-kinetic modeling, and iii) the development of open-source software implementations. Research in these areas can enable the development of catalysis "knowledge engines" which integrate all aspects of catalysis informatics. This key goal, first identified over a decade ago,[66] is now within sight, but will require a combination of technical advances and community efforts in order to be achieved.

The interplay between experimental and computational data is strong in the field of catalysis due to the complementary strengths of these techniques. Collaborations between theoretical and experimental research have shown great success at understanding catalytic processes[173,174] and discovering new catalytic materials.[70,181] However, the integration of information is often qualitative or semi-quantitative, and typically happens through an arduous process of in-person meetings, email exchanges, and phone calls. Experimentalists and theoreticians rarely exchange raw data, and any data exchange typically happens well after the experiments/calculations have been completed. Further, there are numerous examples of superficial "agreement" between theory and experiment which do not hold up to scrutiny (we omit citations, but expect that anyone familiar with the literature is aware of examples). Catalysis informatics has the potential to improve these weaknesses through at least two approaches: improved data infrastructure and probabilistic frameworks. The first challenge is primarily practical, since the technology to immediately upload data to cloud-based infrastructure is well-established, and is increasingly used by industrial labs.[322] In addition, "e-collaboration" platforms to facilitate rapid transfer of data and communication are also being developed in the materials informatics community.[30,97] Similar approaches could be adopted by catalysis researchers to enable rapid, or even real-time, integration of data from diverse theoretical and experimental sources. This holds particular promise for techniques that rely on mathematically intensive analysis such as TAP[256,258] or MES.[243,244] Transient kinetic techniques are presently underutilized but these rich data sources can provide direct connection to micro-kinetic model parameters calculated from DFT. This experimental data should be utilized for model reduction, decreasing the parameter set and guiding atomistic calculations making the dialog between researchers more productive.

The second challenge arises in how to inte-

grate or "fuse" data from the variety of techniques used to analyze heterogeneous catalysts (see Sec. 2.1). One promising approach is the use of statistical frameworks capable of assessing the probability that data from different sources agree.[225,228,229,271,319] Probabilities are agnostic to the type of data, and hence provide a natural framework for data integration; however, significant work is needed to improve the statistical rigor and generality of current approaches. The adoption of these probabilistic frameworks along with improved data infrastructure has the possibility of enabling real-time theory/experiment comparisons. This close coupling will facilitate feedback loops for i) automated design of experiment to identify the most relevant experiment to validate or falsify a hypothesis and ii) systematic model refinement by rapidly identifying structural issues in the model arising from an incorrect active site, incomplete mechanism, or inaccurate approximation to the master equation.

Another key area for informatics research is the development of advanced approaches to micro-kinetic modeling. Micro-kinetic models are proposed as the central representation of knowledge for catalysis, yet existing approaches suffer from a number of limitations. Micro-kinetic models are typically based on the deterministic mean-field approximation to the master equation or stochastic kMC solutions of it.[89] The mean-field or "phenomenological" models are more commonly employed because they are intuitive to set up, and yield a set of coupled ordinary differential equations that are relatively easy to solve and analyze. Mean-field models also require relatively few parameters, and yield deterministic solutions that can be analyzed and refined with sensitivity analyses. However, the accuracy of mean-field models is limited in situations of high coverage,[85] and their extension to multiple active sites is not straightforward.[86] These deficiencies can be mitigated by the addition of more complex coverage-dependent rate constants,[102,319,323] but the simplicity and robustness of the model is decreased. In contrast, kMC approaches seek to exactly solve the master equation based on a lattice-based representation of the surface. This makes inclusion of adsorbate-adsorbate interactions and multiple active sites straightforward, but comes at the cost of a combinatorial explosion of parameters and issues with convergence and sensitivity analysis of stochastic solutions.[268,269,324] The application of statistical and machine-learning techniques can accelerate parameter estimation (see Sec. 2.2), and a number of techniques have recently been developed to perform sensitivity analysis in kMC simulations,[325–328] but significantly more computational effort and human expertise is required relative to mean-field models. One promising alternative is the development of deterministic solutions to the master equation. Several techniques have been proposed,[86,87] though they are mathematically complex and have not been widely applied in practice. Additionally, uncertainty propagation through micro-kinetic models is an important and increasingly common practice[198,225,229,271,319] , but typically relies on ensemble-based approaches that significantly increase the computational cost of micro-kinetic models. The ideal micro-kinetic modeling approach for integration with catalysis informatics frameworks would satisfy the following criteria: i) systematically improvable through inclusion of adsorbate-adsorbate interactions and multiple/dynamic active sites, ii) deterministic solution that is tractable for large reaction networks, iii) facile integration with uncertainty quantification, iv) easily differentiable with respect to parameters, v) implemented, tested, and released for public use.

The final, and perhaps most important, strategy for the future success of catalysis informatics is the development of open-source databases and software tools. The first reported software tool for automatic data-driven catalyst discovery was DECADE, which was developed decades ago.[57,58] Since then several additional software tools for catalyst optimization have been reported,[3,66–68] yet to our knowledge these tools are not widely adopted in industry or academia, or even currently available. While there could be many reasons for this, we believe that the centralized development and lack of open-source code is a significant contributor. Since these tools were de-

veloped there have been significant advances in infrastructure for open-source software development through tools like `git` and associated repositories (Github, Gitlab, Bitbucket, etc.)[329] Furthermore, the machine-learning community has seen vibrant growth and rampant adoption of open-source implementations of algorithms in packages like `scikit-learn`, `Tensorflow`, `Torch`, and `Theano`.[330,331] Notably, many of these packages are developed commercially by companies like Google and Facebook, indicating that there are economic incentives to release open-source code in order to drive adoption.[332] Increasingly, developments in computational catalysis and catalysis informatics are similar to those in machine-learning since they are mathematically intensive, and/or do not have closed-form solutions and/or require substantial amounts of data to be reproduced. This leads to a situation where new advances are impractical or impossible to transfer out of the group where they are developed if an open-source software implementation is not made available.[20] While software development is not formally incentivized by the academic funding structure, there are several informal incentives: i) increased adoption of methods by the community, leading to higher citations and scientific impact, ii) improved reproducibility, iii) community-driven maintenance and documentation, and iv) efficient use of resources by reducing duplicated effort or development of redundant techniques. Several open-source tools and databases are already available and widely used by the catalysis community including `ASE`,[333] `CatApp`,[91] `CatMAP`,[40] `RMG-Cat`,[276] `kmos`,[324] and `AMP`.[212] Many of these tools are developed (or have interfaces) in the `python` programming language, which is also widely used in the machine-learning community.[330,331] This indicates that new tools in catalysis informatics can leverage existing software by supporting a `python` interface, and that learning `python` is a worthwhile exercise for students and researchers interested in catalysis informatics. Ultimately, we expect that catalysis knowledge engines will not be monolithic entities, but rather amalgamations of many tools and databases connected through statistical models to solve specific problems in heterogeneous catalysis.

Table 2: Examples of data sources and types in catalysis (not intended to be an exhaustive list).

Device or Technique(s)	Primary Data[a]	Data type(s)[b]	Analyzed Data Example(s)
NIST Database	Thermochemical quantities	Material, Environment	Heat capacities, Reaction enthalpies, free energies, etc.
DFT	Potential energies	Material, Surface	Adsorption energies, reaction barriers
Material Synthesis	Reagents, concentrations, etc.	Material	Procedure, recipe
XRD	Scattering diffractogram	Material	Unit cell dimensions, structure, crystalline phase(s), particle size
EXAFS, XANES	Absorption spectra	Material, surface	Interatomic distances, coordination number, Debye-Waller factors, oxidation state
BET	Pressure, adsorption isotherm	Material	Surface area, porosity
XPS, Auger, UPS	Photoelectron spectra	Surface	Chemical shift, atomic concentration, oxidation state; work function
LEED	Scattering diffractogram	Surface	Bond distances, lattice dimensions
TEM	Topographic image	Surface	Bond distances, symmetry and space group, elemental composition, oxidation states
STM	Topographic image	Surface	Occupied/unoccupied electronic states
FTIR	Interferogram, absorption spectra	Surface, Environment	Surface vibrational frequencies, gas phase rotational frequencies
Raman	Scattering spectra	Surface, Environment	Lattice vibrational frequencies, crystallinity
TPD, TPR	Desorption temperature profile	Kinetic	Adsorbate coverage, adsorption energy/activation energy of desorption; reaction activation energy
TAP	Exit flux	Kinetic	Intrinsic rate constants, surface residence time, numbers of active sites, microporous diffusivity
GC/MS following PFR, CSTR	Concentration	Kinetic, Environment	Apparent rate constants, conversion, selectivity
SSITKA	Concentration	Kinetic	Apparent rate constants, surface coverages, surface residence time, numbers of active sites

[a] Current and voltage are typically the most primary data source for experiments but models for converting to physicochemical quantities at this level are generally straightforward and reliable.

[b] "Material" is short for bulk material data, and "Environment" is short for chemical environment data.

4 Conclusions

The field of catalysis informatics seeks to systematically and quantitatively organize, represent, and convert data to actionable knowledge about heterogeneous catalysis through the application of statistical and physical models. Knowledge about catalytic processes is embodied in the chemical master equation and microkinetic models, and can be used to generate explanative and predictive hypotheses that enable optimization of operating conditions, discovery of new catalyst materials, and refinement of informatics models. Informatics is closely related to efforts in multi-scale modeling, but is distinguished by application of statistical tools and heuristics in addition to physical models, and the integration of data sources beyond catalytic behavior. Catalysis informatics is closely related to cheminformatics and materials informatics; however, it is distinguished by the fact that catalysis is a dynamic process controlled by structure of the materials surface, which is intimately linked to the surrounding chemical environment. The time-dependent nature of catalysis leads to challenges in data representation and analysis, and the explicit coupling to the chemical environment breaks the paradigm of process-structure-property relationships commonly employed in both cheminformatics and materials informatics. For this reason we propose that a dynamic process-[structure+environment]-property framework is more appropriate for catalysis informatics, and expect that further development of this framework may ultimately lead to developments in other informatics fields.

Catalysis informatics as a field is currently still emerging, despite roots that reach back decades. Recent advances in data infrastructure, statistics, machine learning, and computing in general have the potential to broadly impact the field of heterogeneous catalysis. Presently there is significant room for improvement in the systematic storage and access of catalytic reaction data, most of which is disseminated only through the scientific literature. We propose that although catalysis data is not "big" in size, there are big problems to be surmounted; in particular, the variety of relevant data makes systematic organization challenging and leads to volatile data structures, and the complexity of catalytic measurements leads to a challenge in maintaining data veracity. However, the emergence of schema-free databases enables researchers to store data in a way that is organized yet flexible, and can aid in amalgamating data from diverse sources/schema. We propose that the basic categories of catalysis data are: materials data, chemical environment data, surface science data, and catalytic reaction data. While these broad categorizations leave many open problems to be solved, they may also aid in organization and storage of catalysis data in the future.

The recent rise in machine-learning approaches is revolutionizing the extraction of information from data. This includes the analysis of macro-scale high-throughput catalytic testing data to identify heuristic patterns that link composition and synthesis conditions directly to catalyst performance, as well as numerous recent developments in the rapid prediction of atomic-scale adsorption energies. The discovery and impact of adsorption-energy "scaling relations" along with the rampant increase in adsorption energy data computed with DFT has led to the development of a wide range of approaches for rapidly correlating electronic structure, active-site structure, and adsorbate structure to molecular adsorption energies. Related approaches utilize the full atomic structure of surfaces and adsorbates as inputs to machine-learning models to rapidly predict energies. These machine-learning forcefields are sufficiently flexible to describe a range of complex reactive environments at surfaces and have the potential to significantly increase the time and length scales available for atomic-scale simulations in catalysis. There have also been substantial advances in "bridging the gap" between "micro-scale" surface science models and realistic, functioning catalysts that provide "macro-scale" catalytic reaction data. The rise of in-situ, operando, and transient techniques provides an exciting experimental toolbox for generating data related to micro-scale phenomena under macro-scale conditions. Techniques

such as in-situ X-ray techniques, TAP reactors, and modulation-excitation spectroscopy provide complex data that requires intensive data analysis to extract useful information, and application of machine-learning and data science approaches to this rich source of data is a promising direction for catalysis informatics.

The conversion of information to knowledge requires a quantitative definition of knowledge. We propose that for catalysis this is the Markovian chemical master equation, and the micro-kinetic models that are commonly used to represent or approximate it in the field of catalysis. This requires knowledge of the active sites that participate in the catalytic reaction, along with the chemical mechanisms that connect reactants to products. The challenge of determining active sites and mechanisms are similar in that they require a high-dimensional search over atomic structures (active sites) and chemical reaction networks (reaction mechanisms) where the dimensionality of the search space rapidly becomes intractable for comprehensive investigation. The development of informatics approaches to accelerate the prediction of the energies of atomic surface structures and adsorbed intermediate species enables high-dimensional searches, global optimizations, and iterative model refinement schemes to systematically identify the most probable active sites and mechanisms from a semi-infinite number of possibilities. This results in micro-kinetic models that are capable of explaining the behavior of known catalysts and optimizing their performance. The resulting micro-kinetic models can also be used to discover new catalysts by generating predictions of the catalytic behavior of materials that have not yet been studied. The prospect of integrating high-throughput experimentation with explanative and predictive micro-kinetic models has the potential to result in "knowledge engines" to accelerate the understanding and discovery of heterogeneous catalysts. This is a challenging goal, but the catalysis community is expansive and strong, and there is significant momentum in the field of catalysis informatics. Through collaborative efforts and the development of open-source databases and software tools the prospect of catalysis knowledge engines for automated catalyst design and discovery is realistic.

Acknowledgement The authors thank Gregory Yablonsky and Lars Grabow for discussion and suggested references. This work was fully supported by the U.S. Department of Energy (USDOE), Office of Energy Efficiency and Renewable Energy (EERE), Advanced Manufacturing Office Next Generation R&D Projects under contract no. DE-AC07-05ID14517. Accordingly, the U.S. Government retains a nonexclusive, royalty-free license to publish or reproduce the published form of this contribution, or allow others to do so, for U.S. Government purposes.

References

(1) Knapman, K. Development of a useful combinatorial catalysis informatics platform. How did your company handle the thousands of decisions it made today? *Chimica Oggi* **2001**, *19*, 9–12.

(2) Farrusseng, D.; Baumes, L.; Mirodatos, C. Data Management for Combinatorial Heterogeneous Catalysis: Methodology and Development of Advanced Tools. *ChemInform* **2004**, *35*.

(3) Farrusseng, D.; Clerc, F.; Mirodatos, C.; Azam, N.; Gilardoni, F.; Thybaut, J.; Balasubramaniam, P.; Marin, G. Development of an Integrated Informatics Toolbox: HT Kinetic and Virtual Screening. *Combinatorial Chemistry & High Throughput Screening* **2007**, *10*, 85–97.

(4) Fronczek-Munter, T.; Nrskov, J. Towards Catalysis Informatics - Materials design using Density Functional Theory. Ph.D. thesis, 2008.

(5) Lausche, A. C.; Hummelshøj, J. S.; Abild-Pedersen, F.; Studt, F.; Nørskov, J. K. Application of a new informatics tool in heterogeneous catalysis: Analysis of methanol dehydrogenation on transition metal catalysts for the

production of anhydrous formaldehyde. *Journal of Catalysis* **2012**, *291*, 133–137.

(6) Celse, B.; Rebours, S.; Gay, F.; Coste, P.; Bourgeois, L.; Zammit, O.; Lebacque, V. Integration of an Informatics System in a High Throughput Experimentation. Description of a Global Framework Illustrated Through Several Examples. *Oil & Gas Science and Technology – Revue d'IFP Energies nouvelles* **2013**, *68*, 445–468.

(7) Li, Z.; Ma, X.; Xin, H. Feature engineering of machine-learning chemisorption models for catalyst design. *Catalysis Today* **2017**, *280*, 232–238.

(8) Raybaud, P.; Toulhoat, H. Molecular Modeling and High-Throughput Experimentation (HTE): Meeting the Challenges of Catalysts, Chemicals and Materials Design. *Oil and Gas Science and Technology* **2006**, *61*, 579–592.

(9) Nilsson, A.; Pettersson, L.; Nørskov, J. *Chemical Bonding at Surfaces and Interfaces*; Elsevier BV, 2008; p 312.

(10) Yada, A.; Nagata, K.; Ando, Y.; Matsumura, T.; Ichinoseki, S.; Sato, K. Machine Learning Approach for Prediction of Reaction Yield with Simulated Catalyst Parameters. *Chemistry Letters* **2018**, *47*, 284–287.

(11) Ulissi, Z. W.; Tang, M. T.; Xiao, J.; Liu, X.; Torelli, D. A.; Karamad, M.; Cummins, K.; Hahn, C.; Lewis, N. S.; Jaramillo, T. F.; Chan, K.; Nørskov, J. K. Machine-Learning Methods Enable Exhaustive Searches for Active Bimetallic Facets and Reveal Active Site Motifs for CO_2 Reduction. *ACS Catalysis* **2017**, *7*, 6600–6608.

(12) Li, H.; Zhang, Z.; Liu, Z. Application of Artificial Neural Networks for Catalysis: A Review. *Catalysts* **2017**, *7*, 306.

(13) Ulissi, Z. W.; Medford, A. J.; Bligaard, T.; Nørskov, J. K. To address surface reaction network complexity using scaling relations machine learning and DFT calculations. *Nature Communications* **2017**, *8*, 14621.

(14) Li, Z.; Wang, S.; Chin, W. S.; Achenie, L. E.; Xin, H. High-throughput screening of bimetallic catalysts enabled by machine learning. *Journal of Materials Chemistry A* **2017**, *5*, 24131–24138.

(15) Timoshenko, J.; Lu, D.; Lin, Y.; Frenkel, A. I. Supervised Machine-Learning-Based Determination of Three-Dimensional Structure of Metallic Nanoparticles. *The Journal of Physical Chemistry Letters* **2017**, *8*, 5091–5098.

(16) Vignola, E.; Steinmann, S. N.; Vandegehuchte, B. D.; Curulla, D.; Stamatakis, M.; Sautet, P. A machine learning approach to graph-theoretical cluster expansions of the energy of adsorbate layers. *The Journal of Chemical Physics* **2017**, *147*, 054106.

(17) Ma, X.; Li, Z.; Achenie, L. E. K.; Xin, H. Machine-Learning-Augmented Chemisorption Model for CO_2 Electroreduction Catalyst Screening. *The Journal of Physical Chemistry Letters* **2015**, *6*, 3528–3533.

(18) Landrum, G. A.; Penzotti, J.; Putta, S. Machine-Learning Models for Combinatorial Catalyst Discovery. *MRS Proceedings* **2003**, *804*.

(19) Jinnouchi, R.; Asahi, R. Predicting Catalytic Activity of Nanoparticles by a DFT-Aided Machine-Learning Algorithm. *The Journal of Physical Chemistry Letters* **2017**, *8*, 4279–4283.

(20) Kitchin, J. R. Machine learning in catalysis. *Nature Catalysis* **2018**, *1*, 230–232.

(21) Goldsmith, B. R.; Esterhuizen, J.; Liu, J.-X.; Bartel, C. J.; Sutton, C. Machine learning for heterogeneous cata-

lyst design and discovery. *AIChE Journal* **2018**, *64*, 2311–2323.

(22) Fayyad, U.; Piatetsky-Shapiro, G.; Smyth, P. From Data Mining to Knowledge Discovery in Databases. *AI Magazine* **1996**, *17*, 054106.

(23) Fey, N. Lost in chemical space? Maps to support organometallic catalysis. *Chemistry Central Journal* **2015**, *9*.

(24) Alderson, R. G.; Ferrari, L. D.; Mavridis, L.; McDonagh, J. L.; Mitchell, J. B.; Nath, N. Enzyme Informatics. *Current Topics in Medicinal Chemistry* **2012**, *12*, 1911–1923.

(25) Kowalski, B. R. Chemometrics: Views and Propositions. *Journal of Chemical Information and Modeling* **1975**, *15*, 201–203.

(26) Lahana, R. Cheminformatics – decision making in drug discovery. *Drug Discovery Today* **2002**, *7*, 898–900.

(27) Benton, D. Bioinformatics — principles and potential of a new multidisciplinary tool. *Trends in Biotechnology* **1996**, *14*, 261–272.

(28) Sumathi, R.; Jr., W. H. G. A priori rate constants for kinetic modeling. *Theoretical Chemistry Accounts: Theory, Computation, and Modeling (Theoretica Chimica Acta)* **2002**, *108*, 187–213.

(29) Rajan, K. Materials informatics. *Materials Today* **2005**, *8*, 38–45.

(30) Kalidindi, S. R.; Graef, M. D. Materials Data Science: Current Status and Future Outlook. *Annual Review of Materials Research* **2015**, *45*, 171–193.

(31) Behler, J. Neural network potential-energy surfaces in chemistry: a tool for large-scale simulations. *Physical Chemistry Chemical Physics* **2011**, *13*, 17930.

(32) Botu, V.; Batra, R.; Chapman, J.; Ramprasad, R. Machine Learning Force Fields: Construction, Validation, and Outlook. *The Journal of Physical Chemistry C* **2017**, *121*, 511–522.

(33) Jain, A.; Ong, S. P.; Hautier, G.; Chen, W.; Richards, W. D.; Dacek, S.; Cholia, S.; Gunter, D.; Skinner, D.; Ceder, G.; Persson, K. a. The Materials Project: A materials genome approach to accelerating materials innovation. *APL Materials* **2013**, *1*, 011002.

(34) Curtarolo, S.; Setyawan, W.; Wang, S.; Xue, J.; Yang, K.; Taylor, R. H.; Nelson, L. J.; Hart, G. L.; Sanvito, S.; Buongiorno-Nardelli, M.; Mingo, N.; Levy, O. AFLOWLIB.ORG: A distributed materials properties repository from high-throughput ab initio calculations. *Computational Materials Science* **2012**, *58*, 227–235.

(35) Bañares, M. A. Operando Spectroscopy: the Knowledge Bridge to Assessing Structure-Performance Relationships in Catalyst Nanoparticles. *Advanced Materials* **2011**, *23*, 5293–5301.

(36) Boudart, M. *Advances in Catalysis*; Elsevier, 1969; pp 153–166.

(37) Yates, J. T.; Szabó, A.; Henderson, M. A. *Structure-Activity and Selectivity Relationships in Heterogeneous Catalysis, Proceedings of the ACS Symposium on Structure-Activity Relationships in Heterogeneous Catalysis*; Elsevier, 1991; pp 273–290.

(38) Somorjai, G. The surface science of heterogeneous catalysis. *Surface Science* **1994**, *299-300*, 849–866.

(39) Nørskov, J. K.; Bligaard, T.; Hvolbæk, B.; Abild-Pedersen, F.; Chorkendorff, I.; Christensen, C. H. The nature of the active site in heterogeneous metal catalysis. *Chemical Society Reviews* **2008**, *37*, 2163.

(40) Medford, A. J.; Shi, C.; Hoffmann, M. J.; Lausche, A. C.; Fitzgibbon, S. R.; Bligaard, T.; Nørskov, J. K. CatMAP: A Software Package for Descriptor-Based Microkinetic Mapping of Catalytic Trends. *Catalysis Letters* **2015**, *145*, 794–807.

(41) Vlachos, D. G. *Advances in Chemical Engineering - Multiscale Analysis*; Elsevier, 2005; pp 1–61.

(42) Wellendorff, J.; Silbaugh, T. L.; Garcia-Pintos, D.; Nørskov, J. K.; Bligaard, T.; Studt, F.; Campbell, C. T. A benchmark database for adsorption bond energies to transition metal surfaces and comparison to selected DFT functionals. *Surface Science* **2015**, *640*, 36–44.

(43) Christensen, R.; Hansen, H. A.; Vegge, T. Identifying systematic DFT errors in catalytic reactions. *Catal. Sci. Technol.* **2015**, *5*, 4946–4949.

(44) Booth, G. H.; Grüneis, A.; Kresse, G.; Alavi, A. Towards an exact description of electronic wavefunctions in real solids. *Nature* **2012**, *493*, 365–370.

(45) Govind, N.; Wang, Y.; da Silva, A.; Carter, E. Accurate ab initio energetics of extended systems via explicit correlation embedded in a density functional environment. *Chemical Physics Letters* **1998**, *295*, 129–134.

(46) Solans-Monfort, X.; Sodupe, M.; Branchadell, V.; Sauer, J.; Orlando, R.; Ugliengo, P. Adsorption of NH3and H2O in Acidic Chabazite. Comparison of ONIOM Approach with Periodic Calculations. *The Journal of Physical Chemistry B* **2005**, *109*, 3539–3545.

(47) Zimmerman, P. M.; Head-Gordon, M.; Bell, A. T. Selection and Validation of Charge and Lennard-Jones Parameters for QM/MM Simulations of Hydrocarbon Interactions with Zeolites. *Journal of Chemical Theory and Computation* **2011**, *7*, 1695–1703.

(48) Sharifzadeh, S.; Huang, P.; Carter, E. A. All-electron embedded correlated wavefunction theory for condensed matter electronic structure. *Chemical Physics Letters* **2009**, *470*, 347–352.

(49) Mittasch, A.; Frankenburger, W. The historical development and theory of ammonia synthesis. *Journal of Chemical Education* **1929**, *6*, 2097.

(50) Mittasch, A.; Frankenburg, W. *Advances in Catalysis*; Elsevier, 1950; pp 81–104.

(51) Boudart, M. Model catalysts: reductionism for understanding. *Topics in Catalysis* **2000**, *13*, 147–149.

(52) A., D. J.; W., H. G.; Michel, B. *Handbook of Heterogeneous Catalysis*; American Cancer Society, 2008; Chapter 1.1.

(53) Schloegl, R. Catalytic Synthesis of Ammonia — A "Never-Ending Story"? *ChemInform* **2003**, *34*.

(54) Hubert, B.; Olaf, H.; Alexander, B.; Martin, M. The AmmoniaSynthesis Catalyst of the Next Generation: BariumPromoted OxideSupported Ruthenium. *Angewandte Chemie International Edition* **2001**, *40*, 1061–1063.

(55) Jacobsen, C. J. H.; Dahl, S.; Clausen, B. S.; Bahn, S.; Logadottir, A.; Nørskov, J. K. Catalyst Design by Interpolation in the Periodic Table: Bimetallic Ammonia Synthesis Catalysts. *Journal of the American Chemical Society* **2001**, *123*, 8404–8405.

(56) Bañares-Alcántara, R.; Westerberg, A.; Ko, E.; Rychener, M. Decade—A hybrid expert system for catalyst selection—I. Expert system consideration. *Computers & Chemical Engineering* **1987**, *11*, 265–277.

(57) Banãres-Alcántara, R.; Ko, E.; Westerberg, A.; Rychener, M. DECADE—a

hybrid expert system for catalyst selection—II. Final architecture and results. *Computers & Chemical Engineering* **1988**, *12*, 923–938.

(58) Seshan, K. DECADE - A hybrid expert system for catalyst selection. *Applied Catalysis* **1989**, *55*, N4–N5.

(59) Hu, X. D.; Foley, H. C.; Stiles, A. B. Design of alcohol synthesis catalysts assisted by a knowledge-based expert system. *Industrial & Engineering Chemistry Research* **1991**, *30*, 1419–1427.

(60) Sasaki, M.; Hamada, H.; Kintaichi, Y.; Ito, T. Application of a neural network to the analysis of catalytic reactions Analysis of NO decomposition over Cu/ZSM-5 zeolite. *Applied Catalysis A: General* **1995**, *132*, 261 – 270.

(61) Hattori, T.; Kito, S. Neural network as a tool for catalyst development. *Catalysis Today* **1995**, *23*, 347 – 355, The Impact of Computers on Catalyst Research and Development.

(62) Serra, J. M.; Corma, A.; Chica, A.; Argente, E.; Botti, V. Can artificial neural networks help the experimentation in catalysis? *Catalysis Today* **2003**, *81*, 393 – 403, European Workshop on Combinatorial Catalysis.

(63) Cundari, T. R.; Deng, J.; Pop, H. F.; Srbu, C. Structural Analysis of Transition Metal B-X Substituent Interactions. Toward the Use of Soft Computing Methods for Catalyst Modeling. *Journal of Chemical Information and Modeling* **2000**, *40*, 1052 – 1061.

(64) Cundari, T. R.; Russo, M. Database Mining Using Soft Computing Techniques. An Integrated Neural Network-Fuzzy Logic-Genetic Algorithm Approach. *Journal of Chemical Information and Modeling* **2001**, *41*, 281 – 287.

(65) Pasadakis, N.; Yiokari, C.; Varotsis, N.; Vayenas, C. Characterization of hydrotreating catalysts using the principal component analysis. *Applied Catalysis A: General* **2001**, *207*, 333 – 341.

(66) Caruthers, J. Catalyst design: knowledge extraction from high-throughput experimentation. *Journal of Catalysis* **2003**, *216*, 98–109.

(67) Katare, S.; Bhan, A.; Caruthers, J. M.; Delgass, W. N.; Venkatasubramanian, V. A hybrid genetic algorithm for efficient parameter estimation of large kinetic models. *Computers & Chemical Engineering* **2004**, *28*, 2569–2581.

(68) Metaxas, K.; Thybaut, J. W.; Morra, G.; Farrusseng, D.; Mirodatos, C.; Marin, G. B. A Microkinetic Vision on High-Throughput Catalyst Formulation and Optimization: Development of an Appropriate Software Tool. *Topics in Catalysis* **2009**, *53*, 64–76.

(69) Hammer, B.; Nørskov, J. *Advances in Catalysis*; Elsevier, 2000; pp 71–129.

(70) Nørskov, J. K.; Bligaard, T.; Rossmeisl, J.; Christensen, C. H. Towards the computational design of solid catalysts. *Nature Chemistry* **2009**, *1*, 37–46.

(71) McQuarrie, D. A. Stochastic approach to chemical kinetics. *Journal of Applied Probability* **1967**, *4*, 413–478.

(72) Dumesic, J. A.; Rudd, D. F.; Aparicio, L. M.; Rekoske, J. E.; Trevino, A. A. *The Microkinetics of Heterogeneous Catalysis*; American Chemical Society, 1993.

(73) Rowley, J. The wisdom hierarchy: representations of the DIKW hierarchy. *Journal of Information Science* **2007**, *33*, 163–180.

(74) Yang, Q.; Wu, X. 10 Challenging Problems in Data Mining Research. *International Journal of Information Technology & Decision Making* **2006**, *05*, 597–604.

(75) Wu, X.; Kumar, V.; Quinlan, J. R.; Ghosh, J.; Yang, Q.; Motoda, H.; McLachlan, G. J.; Ng, A.; Liu, B.; Yu, P. S.; Zhou, Z.-H.; Steinbach, M.; Hand, D. J.; Steinberg, D. Top 10 algorithms in data mining. *Knowledge and Information Systems* **2007**, *14*, 1–37.

(76) Gorban, A. N.; Zinovyev, A. Principal manifolds and graphs in practice: from molecular biology to dynamical systems. *International Journal of Neural Systems* **2010**, *20*, 219–232.

(77) Constales, D.; Yablonsky, G. S.; D'hooge, D. R.; Thybaut, J. W.; Marin, G. B. *Advanced Data Analysis and Modelling in Chemical Engineering*; Elsevier, 2016.

(78) Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition (Springer Series in Statistics)*; Springer, 2016.

(79) Rokach, L.; Maimon, O. Z. *Data Mining and Knowledge Discovery Handbook*; Springer, 2010.

(80) Vapnik, V. *The Nature of Statistical Learning Theory (Information Science and Statistics)*; Springer, 2013.

(81) Aggarwal, C. C. *Data Mining: The Textbook*; Springer, 2015.

(82) Nørskov, J. K.; Studt, F.; Abild-Pedersen, F.; Bligaard, T. *Fundamental Concepts in Heterogeneous Catalysis*; Wiley, 2014.

(83) Gillespie, D. T. Exact stochastic simulation of coupled chemical reactions. *The Journal of Physical Chemistry* **1977**, *81*, 2340–2361.

(84) Tovbin, Y. Lattice-gas model in kinetic theory of gas-solid interface processes. *Progress in Surface Science* **1990**, *34*, 1–235.

(85) Temel, B.; Meskine, H.; Reuter, K.; Scheffler, M.; Metiu, H. Does phenomenological kinetics provide an adequate description of heterogeneous catalytic reactions? *The Journal of Chemical Physics* **2007**, *126*, 204711.

(86) Herschlag, G. J.; Mitran, S.; Lin, G. A consistent hierarchy of generalized kinetic equation approximations to the master equation applied to surface catalysis. *The Journal of Chemical Physics* **2015**, *142*, 234703.

(87) Gelß, P.; Matera, S.; Schütte, C. Solving the master equation without kinetic Monte Carlo: Tensor train approximations for a CO oxidation model. *Journal of Computational Physics* **2016**, *314*, 489–502.

(88) de Vijver, R. V.; Devocht, B. R.; Geem, K. M. V.; Thybaut, J. W.; Marin, G. B. Challenges and opportunities for molecule-based management of chemical processes. *Current Opinion in Chemical Engineering* **2016**, *13*, 142–149.

(89) Salciccioli, M.; Stamatakis, M.; Caratzoulas, S.; Vlachos, D. A review of multiscale modeling of metal-catalyzed reactions: Mechanism development for complexity and emergent behavior. *Chemical Engineering Science* **2011**, *66*, 4319–4355.

(90) Hill, J.; Mulholland, G.; Persson, K.; Seshadri, R.; Wolverton, C.; Meredig, B. Materials science with large-scale data and informatics: Unlocking new opportunities. *MRS Bulletin* **2016**, *41*, 399–409.

(91) Hummelshj, J. S.; Abild-Pedersen, F.; Studt, F.; Bligaard, T.; Nrskov, J. K. CatApp: A Web Application for Surface Chemistry and Heterogeneous Catalysis. *Angewandte Chemie International Edition* **2012**, *51*, 272–274.

(92) Earl, D. J.; Deem, M. W. Toward a Database of Hypothetical Zeolite Structures. *Industrial & Engineering Chemistry Research* **2006**, *45*, 5449–5454.

(93) Pophale, R.; Cheeseman, P. A.; Deem, M. W. A database of new zeolite-like materials. *Physical Chemistry Chemical Physics* **2011**, *13*, 12407.

(94) Chung, Y. G.; Camp, J.; Haranczyk, M.; Sikora, B. J.; Bury, W.; Krungleviciute, V.; Yildirim, T.; Farha, O. K.; Sholl, D. S.; Snurr, R. Q. Computation-Ready, Experimental Metal–Organic Frameworks: A Tool To Enable High-Throughput Screening of Nanoporous Crystals. *Chemistry of Materials* **2014**, *26*, 6185–6192.

(95) Kaisler, S.; Armour, F.; Espinosa, J. A.; Money, W. Big Data: Issues and Challenges Moving Forward. 2013 46th Hawaii International Conference on System Sciences. 2013.

(96) Michel, K.; Meredig, B. Beyond bulk single crystals: A data format for all materials structure–property–processing relationships. *MRS Bulletin* **2016**, *41*, 617–623.

(97) Kalidindi, S. R.; Medford, A. J.; McDowell, D. L. Vision for Data and Informatics in the Future Materials Innovation Ecosystem. *JOM* **2016**, *68*, 2126–2137.

(98) Duff, D. G.; Ohrenberg, A.; Voelkening, S.; Boll, M. A Screening Workflow for Synthesis and Testing of 10,000 Heterogeneous Catalysts per Day– Lessons Learned. *Macromolecular Rapid Communications* **2004**, *25*, 169–177.

(99) Saggi, M. K.; Jain, S. A survey towards an integration of big data analytics to big insights for value-creation. *Information Processing & Management* **2018**,

(100) *NIST Big Data Interoperability Framework: Volume 7, Standards Roadmap*; 2015.

(101) Chorkendorff, I.; Niemantsverdriet, J. W. *Concepts of Modern Catalysis and Kinetics*; Wiley-VCH Verlag GmbH & Co. KGaA, 2003.

(102) Yang, N.; Medford, A. J.; Liu, X.; Studt, F.; Bligaard, T.; Bent, S. F.; Nørskov, J. K. Intrinsic Selectivity and Structure Sensitivity of Rhodium Catalysts for C_{2+} Oxygenate Production. *Journal of the American Chemical Society* **2016**, *138*, 3705–3714.

(103) Dahl, S.; Logadottir, A.; Egeberg, R. C.; Larsen, J. H.; Chorkendorff, I.; Törnqvist, E.; Nørskov, J. K. Role of Steps inN2Activation on Ru(0001). *Physical Review Letters* **1999**, *83*, 1814–1817.

(104) Egeberg, R.; Dahl, S.; Logadottir, A.; Larsen, J.; Nørskov, J.; Chorkendorff, I. N2 dissociation on Fe(110) and Fe/Ru(0001): what is the role of steps? *Surface Science* **2001**, *491*, 183–194.

(105) Trotochaud, L.; Young, S. L.; Ranney, J. K.; Boettcher, S. W. Nickel–Iron Oxyhydroxide Oxygen-Evolution Electrocatalysts: The Role of Intentional and Incidental Iron Incorporation. *Journal of the American Chemical Society* **2014**, *136*, 6744–6753.

(106) McCrory, C. C. L.; Jung, S.; Peters, J. C.; Jaramillo, T. F. Benchmarking Heterogeneous Electrocatalysts for the Oxygen Evolution Reaction. *Journal of the American Chemical Society* **2013**, *135*, 16977–16987.

(107) McCrory, C. C. L.; Jung, S.; Ferrer, I. M.; Chatman, S. M.; Peters, J. C.; Jaramillo, T. F. Benchmarking Hydrogen Evolving Reaction and Oxygen Evolving Reaction Electrocatalysts for Solar Water Splitting Devices. *Journal of the American Chemical Society* **2015**, *137*, 4347–4357.

(108) Bligaard, T.; Bullock, R. M.; Campbell, C. T.; Chen, J. G.; Gates, B. C.; Gorte, R. J.; Jones, C. W.; Jones, W. D.; Kitchin, J. R.; Scott, S. L. Toward Benchmarking in Catalysis Science: Best Practices, Challenges, and Opportunities. *ACS Catalysis* **2016**, *6*, 2590–2602.

(109) Jackson, W. *JSON Quick Syntax Reference*; Apress, 2016; pp 21–29.

(110) Chodorow, K.; Dirolf, M. *MongoDB: The Definitive Guide*, 1st ed.; O'Reilly Media, Inc., 2010.

(111) Gormley, C.; Tong, Z. *Elasticsearch: The Definitive Guide*, 1st ed.; O'Reilly Media, Inc., 2015.

(112) Burgess, D. In *NIST Chemistry WebBook, NIST Standard Reference Database Number 69*; Linstrom, P., Mallard, W., Eds.; National Institute of Standards and Technology: Gaithersburg MD, 20899.

(113) Kim, S.; Thiessen, P. A.; Bolton, E. E.; Chen, J.; Fu, G.; Gindulyte, A.; Han, L.; He, J.; He, S.; Shoemaker, B. A.; Wang, J.; Yu, B.; Zhang, J.; Bryant, S. H. PubChem Substance and Compound databases. *Nucleic Acids Research* **2015**, *44*, D1202–D1213.

(114) O'Mara, J.; Meredig, B.; Michel, K. Materials Data Infrastructure: A Case Study of the Citrination Platform to Examine Data Import, Storage, and Access. *JOM* **2016**, *68*, 2031–2034.

(115) Brough, D. B.; Wheeler, D.; Warren, J. A.; Kalidindi, S. R. Microstructure-based knowledge systems for capturing process-structure evolution linkages. *Current Opinion in Solid State and Materials Science* **2017**, *21*, 129–140.

(116) Goodman, D. Catalysis: from single crystals to the "real world". *Surface Science* **1994**, *299-300*, 837–848.

(117) Park, J. Y.; Somorjai, G. A. *Current Trends of Surface Science and Catalysis*; Springer New York, 2013; pp 3–17.

(118) Freund, H.-J.; Kuhlenbeck, H.; Libuda, J.; Rupprechter, G.; Bumer, M.; Hamann, H. Bridging the pressure and materials gaps between catalysis and surface science: clean and modified oxide surfaces. *Topics in Catalysis* **2001**, *15*, 201–209.

(119) Hendershot, R. J.; Snively, C. M.; Lauterbach, J. High-Throughput Heterogeneous Catalytic Science. *ChemInform* **2006**, *37*.

(120) Senkan, S. Combinatorial Heterogeneous Catalysis—A New Path in an Old Field. *Angewandte Chemie International Edition* **2001**, *40*, 312–329.

(121) Turner, H. W.; Volpe, A. F.; Weinberg, W. High-throughput heterogeneous catalyst research. *Surface Science* **2009**, *603*, 1763–1769.

(122) Green, M. L.; Choi, C. L.; Hattrick-Simpers, J. R.; Joshi, A. M.; Takeuchi, I.; Barron, S. C.; Campo, E.; Chiang, T.; Empedocles, S.; Gregoire, J. M.; Kusne, A. G.; Martin, J.; Mehta, A.; Persson, K.; Trautt, Z.; Duren, J. V.; Zakutayev, A. Fulfilling the promise of the materials genome initiative with high-throughput experimental methodologies. *Applied Physics Reviews* **2017**, *4*, 011105.

(123) Yang, K.; Bedenbaugh, J.; Li, H.; Peralta, M.; Bunn, J. K.; Lauterbach, J.; Hattrick-Simpers, J. Development of a High-Throughput Methodology for Screening Coking Resistance of Modified Thin-Film Catalysts. *ACS Combinatorial Science* **2012**, *14*, 372–377.

(124) Sasmaz, E.; Mingle, K.; Lauterbach, J. High-Throughput Screening Using Fourier-Transform Infrared Imaging. *Engineering* **2015**, *1*, 234–242.

(125) Klanner, C.; Farrusseng, D.; Baumes, L.; Mirodatos, C.; Schth, F. How to Design Diverse Libraries of Solid Catalysts? *QSAR & Combinatorial Science* **2003**, *22*, 729–736.

(126) Schuth, F.; Baumes, L.; Clerc, F.; Demuth, D.; Farrusseng, D.; Llamasgalilea, J.; Klanner, C.; Klein, J.; Martinezjoaristi, A.; Procelewska, J. High throughput experimentation in oxidation catalysis: Higher integration and "intelligent" software. *Catalysis Today* **2006**, *117*, 284–290.

(127) Baumes, L. A.; Kruger, F.; Jimenez, S.; Collet, P.; Corma, A. Boosting theoretical zeolitic framework generation for the determination of new materials structures using GPU programming. *Physical Chemistry Chemical Physics* **2011**, *13*, 4674.

(128) Jandeleit, B.; Schaefer, D. J.; Powers, T. S.; Turner, H. W.; Weinberg, H. W. Combinatorial Materials Science and Catalysis. *Angewandte Chemie International Edition* **1999**, *38*, 2494–2532.

(129) Derouane, E. G., Lemos, F., Corma, A., Ribeiro, F. R., Eds. *Combinatorial Catalysis and High Throughput Catalyst Design and Testing*; Springer Netherlands, 2000.

(130) Naccache, C. *Principles and Methods for Accelerated Catalyst Design and Testing*; Springer Netherlands, 2002; pp 245–256.

(131) Baumes, L. A.; Collet, P. Examination of genetic programming paradigm for high-throughput experimentation and heterogeneous catalysis. *Computational Materials Science* **2009**, *45*, 27–40.

(132) Baumes, L. A.; Blansché, A.; Serna, P.; Tchougang, A.; Lachiche, N.; Collet, P.; Corma, A. Using Genetic Programming for an Advanced Performance Assessment of Industrially Relevant Heterogeneous Catalysts. *Materials and Manufacturing Processes* **2009**, *24*, 282–292.

(133) Gusel, L.; Brezocnik, M. Application of genetic programming for modelling of material characteristics. *Expert Systems with Applications* **2011**, *38*, 15014–15019.

(134) Le, T. C.; Winkler, D. A. Discovery and Optimization of Materials Using Evolutionary Approaches. *Chemical Reviews* **2016**, *116*, 6107–6132.

(135) Patra, T. K.; Meenakshisundaram, V.; Hung, J.-H.; Simmons, D. S. Neural-Network-Biased Genetic Algorithms for Materials Design: Evolutionary Algorithms That Learn. *ACS Combinatorial Science* **2017**, *19*, 96–107.

(136) Wang, S.; Temel, B.; Shen, J.; Jones, G.; Grabow, L. C.; Studt, F.; Bligaard, T.; Abild-Pedersen, F.; Christensen, C. H.; Nørskov, J. K. Universal Brønsted-Evans-Polanyi Relations for C–C, C–O, C–N, N–O, N–N, and O–O Dissociation Reactions. *Catalysis Letters* **2010**, *141*, 370–373.

(137) Calle-Vallejo, F.; Loffreda, D.; Koper, M. T. M.; Sautet, P. Introducing structural sensitivity into adsorption–energy scaling relations by means of coordination numbers. *Nature Chemistry* **2015**, *7*, 403–410.

(138) Vorotnikov, V.; Wang, S.; Vlachos, D. G. Group Additivity for Estimating Thermochemical Properties of Furanic Compounds on Pd(111). *Industrial & Engineering Chemistry Research* **2014**, *53*, 11929–11938.

(139) Duan, S.; Kahn, M.; Senkan, S. High-Throughput Nanoparticle Catalysis: Partial Oxidation of Propylene. *Combinatorial Chemistry & High Throughput Screening* **2007**, *10*, 111–119.

(140) Kite, S.; Hattori, T.; Murakami, Y. Estimation of catalytic performance by neural network — product distribution in oxidative dehydrogenation of ethylbenzene. *Applied Catalysis A: General* **1994**, *114*, L173–L178.

(141) Hou, Z.-Y.; Dai, Q.; Wu, X.-Q.; Chen, G.-T. Artificial neural network aided design of catalyst for propane ammoxidation. *Applied Catalysis A: General* **1997**, *161*, 183 – 190.

(142) Huang, K.; Chen, F.-Q.; Lü, D.-W. Artificial neural network-aided design of a multi-component catalyst for methane oxidative coupling. *Applied Catalysis A: General* **2001**, *219*, 61 – 68.

(143) Végvári, L.; Tompos, A.; Gőbls, S.; Margitfalvi, J. Holographic research strategy for catalyst library design. *Catalysis Today* **2003**, *81*, 517–527.

(144) Holeňa, M.; Baerns, M. Feedforward neural networks in catalysis. *Catalysis Today* **2003**, *81*, 485–494.

(145) Corma, A.; Serra, J.; Serna, P.; Valero, S.; Argente, E.; Botti, V. Optimisation of olefin epoxidation catalysts with the application of high-throughput and genetic algorithms assisted by artificial neural networks (softcomputing techniques). *Journal of Catalysis* **2005**, *229*, 513–524.

(146) Holena, M.; Baerns, M. *Handbook of Heterogeneous Catalysis*; American Chemical Society, 2008; Chapter 2.2, pp 66–81.

(147) Cukic, T.; Kraehnert, R.; Holena, M.; Herein, D.; Linke, D.; Dingerdissen, U. The influence of preparation variables on the performance of Pd/Al2O3 catalyst in the hydrogenation of 1,3-butadiene: Building a basis for reproducible catalyst synthesis. *Applied Catalysis A: General* **2007**, *323*, 25–37.

(148) Panahi, P. N.; Niaei, A.; Tseng, H.-H.; Salari, D.; Mousavi, S. M. Modeling of catalyst composition–activity relationship of supported catalysts in NH3–NO-SCR process using artificial neural network. *Neural Computing and Applications* **2015**, *26*, 1515–1523.

(149) Rothenberg, G. Data mining in catalysis: Separating knowledge from garbage. *Catalysis Today* **2008**, *137*, 2 – 10, Recent Developments in Combinatorial Catalysis Research and High-Throughput Technologies.

(150) Şener, A. N.; Günay, M. E.; Leba, A.; Yıldırım, R. Statistical review of dry reforming of methane literature using decision tree and artificial neural network analysis. *Catalysis Today* **2018**, *299*, 289–302.

(151) Tapan, N. A.; Yıldırım, R.; Günay, M. E. Analysis of past experimental data in literature to determine conditions for high performance in biodiesel production. *Biofuels, Bioproducts and Biorefining* **2016**, *10*, 422–434.

(152) Odabaşı, Ç.; Günay, M. E.; Yıldırım, R. Knowledge extraction for water gas shift reaction over noble metal catalysts from publications in the literature between 2002 and 2012. *International Journal of Hydrogen Energy* **2014**, *39*, 5733–5746.

(153) Gnay, M. E.; Yildirim, R. Neural network Analysis of Selective CO Oxidation over Copper-Based Catalysts for Knowledge Extraction from Published Data in the Literature. *Industrial & Engineering Chemistry Research* **2011**, *50*, 12488–12500.

(154) Günay, M. E.; Yildirim, R. Developing global reaction rate model for CO oxidation over Au catalysts from past data in literature using artificial neural networks. *Applied Catalysis A: General* **2013**, *468*, 395–402.

(155) Günay, M. E.; Yildirim, R. Knowledge Extraction from Catalysis of the Past:

A Case of Selective CO Oxidation over Noble Metal Catalysts between 2000 and 2012. *ChemCatChem* **2013**, *5*, 1395–1406.

(156) Wold, S. Principal Component Analysis. *Chemometrics and Intelligent Laboratory Systems* **1987**, *2*, 37–52.

(157) Rothenberg, G. Principal Component Analysis of Catalytic Functions in the Composition Space of Heterogeneous Catalysts. *QSAR and Combinatorial Science* **2007**, *26*, 528 535.

(158) Teixeira, F.; Mosquera, R. A.; Melo, A.; Freire, C.; Cordeiro, M. N. D. S. Principal component analysis of Mn(salen) catalysts. *Phys. Chem. Chem. Phys.* **2014**, *16*, 25364–25376.

(159) Pattiya, A.; Titiloye, J. O.; Bridgwater, A. Evaluation of catalytic pyrolysis of cassava rhizome by principal component analysis. *Fuel* **2010**, *89*, 244 – 253.

(160) Race, A. M.; Steven, R. T.; Palmer, A. D.; Styles, I. B.; Bunch, J. Memory Efficient Principal Component Analysis for the Dimensionality Reduction of Large Mass Spectrometry Imaging Data Sets. *Analytical Chemistry* **2013**, *85*, 3071–3078, PMID: 23394348.

(161) Andriotis, A. N.; Mpourmpakis, G.; Broderick, S.; Rajan, K.; Datta, S.; Sunkara, M.; Menon, M. Informatics guided discovery of surface structure-chemistry relationships in catalytic nanoparticles. *The Journal of Chemical Physics* **2014**, *140*, 094705–0947058.

(162) Izenman, A. J. *Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning*, 1st ed.; Springer Publishing Company, Incorporated, 2008.

(163) Kitchin, J. R.; Gellman, A. J. High-throughput methods using composition and structure spread libraries. *AIChE Journal* **2016**, *62*, 3826–3835.

(164) Boes, J. R.; Gumuslu, G.; Miller, J. B.; Gellman, A. J.; Kitchin, J. R. Estimating Bulk-Composition-Dependent H2 Adsorption Energies on CuxPd1–x Alloy (111) Surfaces. *ACS Catalysis* **2015**, *5*, 1020–1026.

(165) Gautier, S.; Steinmann, S. N.; Michel, C.; Fleurat-Lessard, P.; Sautet, P. Molecular adsorption at Pt(111). How accurate are DFT functionals? *Physical Chemistry Chemical Physics* **2015**, *17*, 28921–28930.

(166) Sharada, S. M.; Bligaard, T.; Luntz, A. C.; Kroes, G.-J.; Nørskov, J. K. SBH10: A Benchmark Database of Barrier Heights on Transition Metal Surfaces. *The Journal of Physical Chemistry C* **2017**, *121*, 19807–19815.

(167) Savara, A. Simulation and fitting of complex reaction network TPR: The key is the objective function. *Surface Science* **2016**, *653*, 169–180.

(168) de Bruijne, M. Machine learning approaches in medical image analysis: From detection to diagnosis. *Medical Image Analysis* **2016**, *33*, 94–97.

(169) Kalinin, S. V.; Sumpter, B. G.; Archibald, R. K. Big-deep-smart data in imaging for guiding materials design. *Nature Materials* **2015**, *14*, 973–980.

(170) Ziatdinov, M.; Dyck, O.; Maksov, A.; Li, X.; Sang, X.; Xiao, K.; Unocic, R. R.; Vasudevan, R.; Jesse, S.; Kalinin, S. V. Deep Learning of Atomically Resolved Scanning Transmission Electron Microscopy Images: Chemical Identification and Tracking Local Transformations. *ACS Nano* **2017**, *11*, 12742–12752.

(171) Ziatdinov, M.; Maksov, A.; Kalinin, S. V. Learning surface molecular structures via machine vision. *npj Computational Materials* **2017**, *3*, 31.

(172) Vlcek, L.; Maksov, A.; Pan, M.; Vasudevan, R. K.; Kalinin, S. V. Knowledge Extraction from Atomically Resolved Images. *ACS Nano* **2017**, *11*, 10313–10320.

(173) Neurock, M. Perspectives on the First Principles Elucidation and the Design of Active Sites. *ChemInform* **2003**, *34*.

(174) van Santen, R. A.; Neurock, M. *Molecular Heterogeneous Catalysis: A Conceptual and Computational Approach*; Wiley-VCH, 2006.

(175) Hammer, B.; Nørskov, J. K. Why gold is the noblest of all the metals. *Nature* **1995**, *376*, 238–240.

(176) Hammer, B.; Nørskov, J. Electronic factors determining the reactivity of metal surfaces. *Surface Science* **1995**, *343*, 211–220.

(177) Abild-Pedersen, F.; Greeley, J.; Studt, F.; Rossmeisl, J.; Munter, T. R.; Moses, P. G.; Skúlason, E.; Bligaard, T.; Nørskov, J. K. Scaling Properties of Adsorption Energies for Hydrogen-Containing Molecules on Transition-Metal Surfaces. *Physical Review Letters* **2007**, *99*.

(178) Suntivich, J.; Gasteiger, H. A.; Yabuuchi, N.; Nakanishi, H.; Goodenough, J. B.; Shao-Horn, Y. Design principles for oxygen-reduction activity on perovskite oxide catalysts for fuel cells and metal–air batteries. *Nature Chemistry* **2011**, *3*, 546–550.

(179) Man, I. C.; Su, H.-Y.; Calle-Vallejo, F.; Hansen, H. A.; Martínez, J. I.; Inoglu, N. G.; Kitchin, J.; Jaramillo, T. F.; Nørskov, J. K.; Rossmeisl, J. Universality in Oxygen Evolution Electrocatalysis on Oxide Surfaces. *ChemCatChem* **2011**, *3*, 1159–1165.

(180) Fernández, E.; Moses, P.; Toftelund, A.; Hansen, H.; Martínez, J.; Abild-Pedersen, F.; Kleis, J.; Hinnemann, B.; Rossmeisl, J.; Bligaard, T.; Nørskov, J. Scaling Relationships for Adsorption Energies on Transition Metal Oxide, Sulfide, and Nitride Surfaces. *Angewandte Chemie* **2008**, *120*, 4761–4764.

(181) Greeley, J. Theoretical Heterogeneous Catalysis: Scaling Relationships and Computational Catalyst Design. *Annual Review of Chemical and Biomolecular Engineering* **2016**, *7*, 605–635.

(182) Montemore, M. M.; Medlin, J. W. Scaling relations between adsorption energies for computational screening and design of catalysts. *Catal. Sci. Technol.* **2014**, *4*, 3748–3761.

(183) Vojvodic, A.; Ruberto, C.; Lundqvist, B. I. Atomic and molecular adsorption on transition-metal carbide (111) surfaces from density-functional theory: a trend study of surface electronic factors. *Journal of Physics: Condensed Matter* **2010**, *22*, 375504.

(184) İnoğlu, N.; Kitchin, J. R. New solid-state table: estimating d-band characteristics for transition metal atoms. *Molecular Simulation* **2010**, *36*, 633–638.

(185) Schweitzer, N.; Xin, H.; Nikolla, E.; Miller, J. T.; Linic, S. Establishing Relationships Between the Geometric Structure and Chemical Reactivity of Alloy Catalysts Based on Their Measured Electronic Structure. *Topics in Catalysis* **2010**, *53*, 348–356.

(186) Xin, H.; Holewinski, A.; Schweitzer, N.; Nikolla, E.; Linic, S. Electronic Structure Engineering in Heterogeneous Catalysis: Identifying Novel Alloy Catalysts Based on Rapid Screening for Materials with Desired Electronic Properties. *Topics in Catalysis* **2012**, *55*, 376–390.

(187) Xin, H.; Vojvodic, A.; Voss, J.; Nørskov, J. K.; Abild-Pedersen, F. Effects ofd-band shape on the surface reactivity of transition-metal alloys. *Physical Review B* **2014**, *89*.

(188) Holewinski, A.; Xin, H.; Nikolla, E.; Linic, S. Identifying optimal active sites for heterogeneous catalysis by metal alloys based on molecular descriptors and electronic structure engineering. *Current Opinion in Chemical Engineering* **2013**, *2*, 312–319.

(189) Calle-Vallejo, F.; Inoglu, N. G.; Su, H.-Y.; Martínez, J. I.; Man, I. C.; Koper, M. T. M.; Kitchin, J. R.; Rossmeisl, J. Number of outer electrons as descriptor for adsorption processes on transition metals and their oxides. *Chemical Science* **2013**, *4*, 1245.

(190) Ma, X.; Xin, H. Orbitalwise Coordination Number for Predicting Adsorption Properties of Metal Nanocatalysts. *Physical Review Letters* **2017**, *118*.

(191) Calle-Vallejo, F.; Martínez, J. I.; García-Lastra, J. M.; Sautet, P.; Loffreda, D. Fast Prediction of Adsorption Properties for Platinum Nanocatalysts with Generalized Coordination Numbers. *Angewandte Chemie* **2014**, *126*, 8456–8459.

(192) Montemore, M. M.; Medlin, J. W. Site-Specific Scaling Relations for Hydrocarbon Adsorption on Hexagonal Transition Metal Surfaces. *The Journal of Physical Chemistry C* **2013**, *117*, 20078–20088.

(193) Zhao, Z.; Chen, Z.; Zhang, X.; Lu, G. Generalized Surface Coordination Number as an Activity Descriptor for CO_2 Reduction on Cu Surfaces. *The Journal of Physical Chemistry C* **2016**, *120*, 28125–28130.

(194) Guo, W.; Vlachos, D. G. Effect of local metal microstructure on adsorption on bimetallic surfaces: Atomic nitrogen on Ni/Pt(111). *The Journal of Chemical Physics* **2013**, *138*, 174702.

(195) Jones, G.; Studt, F.; Abild-Pedersen, F.; Nørskov, J. K.; Bligaard, T. Scaling relationships for adsorption energies of C2 hydrocarbons on transition metal surfaces. *Chemical Engineering Science* **2011**, *66*, 6318–6323.

(196) Wang, S.; Petzold, V.; Tripkovic, V.; Kleis, J.; Howalt, J. G.; Skúlason, E.; Fernández, E. M.; Hvolbæk, B.; Jones, G.; Toftelund, A.; Falsig, H.; Björketun, M.; Studt, F.; Abild-Pedersen, F.; Rossmeisl, J.; Nørskov, J. K.; Bligaard, T. Universal transition state scaling relations for (de)hydrogenation over transition metals. *Physical Chemistry Chemical Physics* **2011**, *13*, 20760.

(197) Göltl, F.; Müller, P.; Uchupalanun, P.; Sautet, P.; Hermans, I. Developing a Descriptor-Based Approach for CO and NO Adsorption Strength to Transition Metal Sites in Zeolites. *Chemistry of Materials* **2017**, *29*, 6434–6444.

(198) Krishnamurthy, D.; Sumaria, V.; Viswanathan, V. Maximal Predictability Approach for Identifying the Right Descriptors for Electrocatalytic Reactions. *The Journal of Physical Chemistry Letters* **2018**, *9*, 588–595.

(199) Benson, S. W.; Buss, J. H. Additivity Rules for the Estimation of Molecular Properties. Thermodynamic Properties. *The Journal of Chemical Physics* **1958**, *29*, 546–572.

(200) Kua, J.; Goddard, W. A. Chemisorption of Organics on Platinum. 2. Chemisorption of C_2H_x and CH_x on Pt(111). *The Journal of Physical Chemistry B* **1998**, *102*, 9492–9500.

(201) Kua, J.; Faglioni, F.; Goddard, W. A. Thermochemistry for Hydrocarbon Intermediates Chemisorbed on Metal Surfaces: $CH_{n-m}(CH_3)$ with n= 1, 2, 3 and m≤n on Pt, Ir, Os, Pd, Rh, and Ru. *Journal of the American Chemical Society* **2000**, *122*, 2309–2321.

(202) Salciccioli, M.; Chen, Y.; Vlachos, D. G. Density Functional Theory-Derived

Group Additivity and Linear Scaling Methods for Prediction of Oxygenate Stability on Metal Catalysts: Adsorption of Open-Ring Alcohol and Polyol Dehydrogenation Intermediates on Pt-Based Metals. *The Journal of Physical Chemistry C* **2010**, *114*, 20155–20166.

(203) Gu, G. H.; Vlachos, D. G. Group Additivity for Thermochemical Property Estimation of Lignin Monomers on Pt(111). *The Journal of Physical Chemistry C* **2016**, *120*, 19234–19241.

(204) Gu, G. H.; Schweitzer, B.; Michel, C.; Steinmann, S. N.; Sautet, P.; Vlachos, D. G. Group Additivity for Aqueous Phase Thermochemical Properties of Alcohols on Pt(111). *The Journal of Physical Chemistry C* **2017**, *121*, 21510–21519.

(205) Shustorovich, E.; Bell, A. T. The thermochemistry of C2 hydrocarbons on transition metal surfaces: The bond-order-conservation approach. *Surface Science* **1988**, *205*, 492–512.

(206) Shustorovich, E. *Advances in Catalysis*; Elsevier, 1990; pp 101–163.

(207) Shustorovich, E. The UBI-QEP method: A practical theoretical approach to understanding chemistry on transition metal surfaces. *Surface Science Reports* **1998**, *31*, 1–119.

(208) Nelson, L. J.; Ozoliņš, V.; Reese, C. S.; Zhou, F.; Hart, G. L. W. Cluster expansion made easy with Bayesian compressive sensing. *Physical Review B* **2013**, *88*.

(209) Nelson, L. J.; Hart, G. L. W.; Zhou, F.; Ozoliņš, V. Compressive sensing as a paradigm for building physics models. *Physical Review B* **2013**, *87*.

(210) Hoffmann, M. J.; Medford, A. J.; Bligaard, T. Framework for Scalable Adsorbate–Adsorbate Interaction Models. *The Journal of Physical Chemistry C* **2016**, *120*, 13087–13094.

(211) Artrith, N.; Hiller, B.; Behler, J. Neural network potentials for metals and oxides - First applications to copper clusters at zinc oxide. *physica status solidi (b)* **2012**, *250*, 1191–1203.

(212) Khorshidi, A.; Peterson, A. A. Amp : A modular approach to machine learning in atomistic simulations. *Computer Physics Communications* **2016**, *207*, 310–324.

(213) Pilania, G.; Wang, C.; Jiang, X.; Rajasekaran, S.; Ramprasad, R. Accelerating materials property predictions using machine learning. *Scientific Reports* **2013**, *3*.

(214) Ulissi, Z. W.; Singh, A. R.; Tsai, C.; Nørskov, J. K. Automated Discovery and Construction of Surface Phase Diagrams Using Machine Learning. *The Journal of Physical Chemistry Letters* **2016**, *7*, 3931–3935.

(215) Ludwig, J.; Vlachos, D. G. Ab initio molecular dynamics of hydrogen dissociation on metal surfaces using neural networks and novelty sampling. *The Journal of Chemical Physics* **2007**, *127*, 154716.

(216) Artrith, N.; Behler, J. High-dimensional neural network potentials for metal surfaces: A prototype study for copper. *Physical Review B* **2012**, *85*.

(217) Boes, J. R.; Kitchin, J. R. Neural network predictions of oxygen interactions on a dynamic Pd surface. *Molecular Simulation* **2017**, *43*, 346–354.

(218) Natarajan, S. K.; Behler, J. Neural network molecular dynamics simulations of solid–liquid interfaces: water at low-index copper surfaces. *Physical Chemistry Chemical Physics* **2016**, *18*, 28704–28725.

(219) Quaranta, V.; Hellström, M.; Behler, J. Proton-Transfer Mechanisms at the Water–ZnO Interface: The Role of Presolvation. *The Journal of Physical Chemistry Letters* **2017**, *8*, 1476–1483.

(220) Natarajan, S. K.; Behler, J. Self-Diffusion of Surface Defects at Copper–Water Interfaces. *The Journal of Physical Chemistry C* **2017**, *121*, 4368–4383.

(221) Boes, J. R.; Kitchin, J. R. Modeling Segregation on AuPd(111) Surfaces with Density Functional Theory and Monte Carlo Simulations. *The Journal of Physical Chemistry C* **2017**, *121*, 3479–3487.

(222) Behler, J. First Principles Neural Network Potentials for Reactive Simulations of Large Molecular and Condensed Systems. *Angewandte Chemie International Edition* **2017**, *56*, 12828–12840.

(223) Boes, J. R.; Groenenboom, M. C.; Keith, J. A.; Kitchin, J. R. Neural network and ReaxFF comparison for Au properties. *International Journal of Quantum Chemistry* **2016**, *116*, 979–987.

(224) Carr, S. F.; Garnett, R.; Lo, C. S. Accelerating the search for global minima on potential energy surfaces using machine learning. *The Journal of Chemical Physics* **2016**, *145*, 154106.

(225) Medford, A. J.; Wellendorff, J.; Vojvodic, A.; Studt, F.; Abild-Pedersen, F.; Jacobsen, K. W.; Bligaard, T.; Nørskov, J. K. Assessing the reliability of calculated catalytic ammonia synthesis rates. *Science* **2014**, *345*, 197–200.

(226) Nørskov, J. K.; Rossmeisl, J.; Logadottir, A.; Lindqvist, L.; Kitchin, J. R.; Bligaard, T.; Jónsson, H. Origin of the Overpotential for Oxygen Reduction at a Fuel-Cell Cathode. *The Journal of Physical Chemistry B* **2004**, *108*, 17886–17892.

(227) Peterson, A. A.; Nørskov, J. K. Activity Descriptors for CO_2 Electroreduction to Methane on Transition-Metal Catalysts. *The Journal of Physical Chemistry Letters* **2012**, *3*, 251–258.

(228) Walker, E.; Ammal, S. C.; Terejanu, G. A.; Heyden, A. Uncertainty Quantification Framework Applied to the Water–Gas Shift Reaction over Pt-Based Catalysts. *The Journal of Physical Chemistry C* **2016**, *120*, 10328–10339.

(229) Sutton, J. E.; Vlachos, D. G. Effect of errors in linear scaling relations and Brønsted–Evans–Polanyi relations on activity and selectivity maps. *Journal of Catalysis* **2016**, *338*, 273–283.

(230) Wellendorff, J.; Lundgaard, K. T.; Møgelhøj, A.; Petzold, V.; Landis, D. D.; Nørskov, J. K.; Bligaard, T.; Jacobsen, K. W. Density functionals for surface science: Exchange-correlation model development with Bayesian error estimation. *Physical Review B* **2012**, *85*.

(231) Wellendorff, J.; Lundgaard, K. T.; Jacobsen, K. W.; Bligaard, T. mBEEF: An accurate semi-local Bayesian error estimation density functional. *The Journal of Chemical Physics* **2014**, *140*, 144107.

(232) Lundgaard, K. T.; Wellendorff, J.; Voss, J.; Jacobsen, K. W.; Bligaard, T. mBEEF-vdW: Robust fitting of error estimation density functionals. *Physical Review B* **2016**, *93*.

(233) Peterson, A. A.; Christensen, R.; Khorshidi, A. Addressing uncertainty in atomistic machine learning. *Physical Chemistry Chemical Physics* **2017**, *19*, 10978–10985.

(234) Wachs, I. E. Extending surface science studies to industrial reaction conditions: mechanism and kinetics of methanol oxidation over silver surfaces. *Surface Science* **2003**, *544*, 1–4.

(235) Weckhuysen, B. M. Snapshots of a working catalyst: possibilities and limitations of in situ spectroscopy in the field of heterogeneous catalysis. *Chemical Communications* **2002**, 97–110.

(236) Helveg, S.; López-Cartes, C.; Sehested, J.; Hansen, P. L.; Clausen, B. S.; Rostrup-Nielsen, J. R.; Abild-Pedersen, F.; Nørskov, J. K. Atomic-scale imaging of carbon nanofibre growth. *Nature* **2004**, *427*, 426–429.

(237) Hansen, P. L. Atom-Resolved Imaging of Dynamic Shape Changes in Supported Copper Nanocrystals. *Science* **2002**, *295*, 2053–2055.

(238) Creemer, J.; Helveg, S.; Hoveling, G.; Ullmann, S.; Molenbroek, A.; Sarro, P.; Zandbergen, H. Atomic-scale electron microscopy at ambient pressure. *Ultramicroscopy* **2008**, *108*, 993–998.

(239) Shekhtman, S. O.; Yablonsky, G. S.; Gleaves, J. T.; Fushimi, R. "State defining" experiment in chemical kinetics—primary characterization of catalyst activity in a TAP experiment. *Chemical Engineering Science* **2003**, *58*, 4843–4859.

(240) Yablonsky, G.; Constales, D.; Gleaves, J. Multi-Scale Problems in the Quantitative Characterization of Complex Catalytic Materials. *Systems Analysis Modelling Simulation* **2002**, *42*, 1143–1166.

(241) Yablonsky, G. S.; Redekop, E. A.; Constales, D.; Gleaves, J. T.; Marin, G. B. Rate-Reactivity Model: A New Theoretical Basis for Systematic Kinetic Characterization of Heterogeneous Catalysts. *International Journal of Chemical Kinetics* **2016**, *48*, 304–317.

(242) Berger, R. J.; Kapteijn, F.; Moulijn, J. A.; Marin, G. B.; Wilde, J. D.; Olea, M.; Chen, D.; Holmen, A.; Lietti, L.; Tronconi, E.; Schuurman, Y. Dynamic methods for catalytic kinetics. *Applied Catalysis A: General* **2008**, *342*, 3–28.

(243) Baurecht, D.; Fringeli, U. P. Quantitative modulated excitation Fourier transform infrared spectroscopy. *Review of Scientific Instruments* **2001**, *72*, 3782–3792.

(244) Müller, P.; Hermans, I. Applications of Modulation Excitation Spectroscopy in Heterogeneous Catalysis. *Industrial & Engineering Chemistry Research* **2017**, *56*, 1123–1136.

(245) Liu, J.; Osadchy, M.; Ashton, L.; Foster, M.; Solomon, C. J.; Gibson, S. J. Deep convolutional neural networks for Raman spectrum recognition: a unified solution. *The Analyst* **2017**, *142*, 4067–4074.

(246) Gastegger, M.; Behler, J.; Marquetand, P. Machine learning molecular dynamics for the simulation of infrared spectra. *Chem. Sci.* **2017**, *8*, 6924–6935.

(247) Rossouw, D.; Burdet, P.; de la Peña, F.; Ducati, C.; Knappett, B. R.; Wheatley, A. E. H.; Midgley, P. A. Multicomponent Signal Unmixing from Nanoheterostructures: Overcoming the Traditional Challenges of Nanoscale X-ray Analysis via Machine Learning. *Nano Letters* **2015**, *15*, 2716–2720.

(248) Ievlev, A. V.; Belianinov, A.; Jesse, S.; Allison, D. P.; Doktycz, M. J.; Retterer, S. T.; Kalinin, S. V.; Ovchinnikova, O. S. Automated Interpretation and Extraction of Topographic Information from Time of Flight Secondary Ion Mass Spectrometry Data. *Scientific Reports* **2017**, *7*.

(249) Zheng, C.; Mathew, K.; Chen, C.; Chen, Y.; Tang, H.; Dozier, A.; Kas, J. J.; Vila, F. D.; Rehr, J. J.; Piper, L. F. J.; Persson, K. A.; Ong, S. P. Automated generation and ensemble-learned matching of X-ray absorption spectra. *npj Computational Materials* **2018**, *4*.

(250) Belianinov, A.; He, Q.; Kravchenko, M.; Jesse, S.; Borisevich, A.; Kalinin, S. V. Identification of phases, symmetries and defects through local crystallography. *Nature Communications* **2015**, *6*, 7801.

(251) Redekop, E. A.; Yablonsky, G. S.; Constales, D.; Ramanchandran, P. A.; Pherigo, C.; Gleaves, J. T. The Y-procedure methodology for the interpretation of transient kinetic data: analysis of irreversible adsorption. *CHEMICAL ENGINEERING SCIENCE* **2011**, *66*, 6441–6452.

(252) Happel, J. Transient tracing. *Chemical Engineering Science* **1978**, *33*, 1567.

(253) Ledesma, C.; Yang, J.; Chen, D.; Holmen, A. Recent Approaches in Mechanistic and Kinetic Studies of Catalytic Reactions Using SSITKA Technique. *ACS Catalysis* **2014**, *4*, 4527–4547.

(254) Shannon, S. L.; Goodwin, J. G. Characterization of Catalytic Surfaces by Isotopic-Transient Kinetics during Steady-State Reaction. *Chemical Reviews* **1995**, *95*, 677–695.

(255) Gleaves, J. T.; Ebner, J. R.; Kuechler, T. C. Temporal Analysis of Products (TAP)—A Unique Catalyst Evaluation System with Submillisecond Time Resolution. *Catalysis Reviews* **1988**, *30*, 49–116.

(256) Gleaves, J. T.; Yablonsky, G.; Zheng, X.; Fushimi, R.; Mills, P. L. Temporal analysis of products (TAP)—Recent advances in technology for kinetic analysis of multi-component catalysts. *Journal of Molecular Catalysis A: Chemical* **2010**, *315*, 108–134.

(257) Morgan, K.; Maguire, N.; Fushimi, R.; Gleaves, J. T.; Goguet, A.; Harold, M. P.; Kondratenko, E. V.; Menon, U.; Schuurman, Y.; Yablonsky, G. S. Forty years of temporal analysis of products. *Catalysis Science & Technology* **2017**, *7*, 2416–2439.

(258) Yablonsky, G.; Constales, D.; Shekhtman, S.; Gleaves, J. The Y-procedure: How to extract the chemical transformation rate from reaction-diffusion data with no assumption on the kinetic model. *Chemical Engineering Science* **2007**, *62*, 6754–6767.

(259) Hoost, T. E.; Jr., J. G. G. Nonparametric determination of reactivity distributions from isotopic transient kinetic data. *Journal of Catalysis* **1992**, *134*, 678–690.

(260) Shannon, S.; Goodwin, J. Use of linear modeling in steady-state isotopic-transient kinetic analysis of surface-catalyzed reactions: Application to plug-flow reactors. *Applied Catalysis A: General* **1997**, *151*, 3–26.

(261) Redekop, E. A.; Yablonsky, G. S.; Constales, D.; Ramachandran, P. A.; Gleaves, J. T.; Marin, G. B. Elucidating complex catalytic mechanisms based on transient pulse-response kinetic data. *Chemical Engineering Science* **2014**, *110*, 20–30.

(262) Urakawa, A.; Burgi, T.; Baiker, A. Sensitivity enhancement and dynamic behavior analysis by modulation excitation spectroscopy: Principle and application in heterogeneous catalysis. *Chemical Engineering Science* **2008**, *63*, 4902–4909.

(263) Ferri, D.; Newton, M. A.; Nachtegaal, M. Modulation Excitation X-Ray Absorption Spectroscopy to Probe Surface Species on Heterogeneous Catalysts. *Topics in Catalysis* **2011**, *54*, 1070–1078.

(264) Cavers, M.; Davidson, J.; Harkness, I.; Rees, L.; McDougall, G. Spectroscopic Identification of the Active Site for CO Oxidation on Rh/Al2O3 by Concentration Modulation in situ DRIFTS. *Journal of Catalysis* **1999**, *188*, 426–430.

(265) Ortelli, E.; Wokaun, A. Use of periodic variations of reactant concentrations in

time resolved FTIR studies of heterogeneously catalysed reactions. *Vibrational Spectroscopy* **1999**, *19*, 451–459.

(266) Urakawa, A.; Bürgi, T.; Baiker, A. Modulation Excitation PM-IRRAS: A New Possibility for Simultaneous Monitoring of Surface and Gas Species and Surface Properties. *CHIMIA International Journal for Chemistry* **2006**, *60*, 231–233.

(267) König, C. F. J.; van Bokhoven, J. A.; Schildhauer, T. J.; Nachtegaal, M. Quantitative Analysis of Modulated Excitation X-ray Absorption Spectra: Enhanced Precision of EXAFS Fitting. *The Journal of Physical Chemistry C* **2012**, *116*, 19857–19866.

(268) Chatterjee, A.; Vlachos, D. G. An overview of spatial microscopic and accelerated kinetic Monte Carlo methods. *Journal of Computer-Aided Materials Design* **2007**, *14*, 253–308.

(269) Sabbe, M. K.; Reyniers, M.-F.; Reuter, K. First-principles kinetic modeling in heterogeneous catalysis: an industrial perspective on best-practice, gaps and needs. *Catalysis Science & Technology* **2012**, *2*, 2010.

(270) Shmueli, G. To Explain or to Predict? *Statistical Science* **2010**, *25*, 289–310.

(271) Sutton, J. E.; Guo, W.; Katsoulakis, M. A.; Vlachos, D. G. Effects of correlated parameters and uncertainty in electronic-structure-based chemical kinetic modelling. *Nature Chemistry* **2016**, *8*, 331–337.

(272) Dellamorte, J. C.; Barteau, M. A.; Lauterbach, J. Opportunities for catalyst discovery and development: Integrating surface science and theory with high throughput methods. *Surface Science* **2009**, *603*, 1770–1775.

(273) Grabow, L. C.; Mavrikakis, M. Mechanism of Methanol Synthesis on Cu through CO_2 and CO Hydrogenation. *ACS Catalysis* **2011**, *1*, 365–384.

(274) Fogler, H. S. *Elements of Chemical Reaction Engineering (5th Edition) (Prentice Hall International Series in the Physical and Chemical Engineering Sciences)*; Prentice Hall, 2016.

(275) Gao, C. W.; Allen, J. W.; Green, W. H.; West, R. H. Reaction Mechanism Generator: Automatic construction of chemical kinetic mechanisms. *Computer Physics Communications* **2016**, *203*, 212–225.

(276) Goldsmith, C. F.; West, R. H. Automatic Generation of Microkinetic Mechanisms for Heterogeneous Catalysis. *The Journal of Physical Chemistry C* **2017**, *121*, 9970–9981.

(277) Rangarajan, S.; Bhan, A.; Daoutidis, P. Language-oriented rule-based reaction network generation and analysis: Description of RING. *Computers & Chemical Engineering* **2012**, *45*, 114–123.

(278) Rangarajan, S.; Bhan, A.; Daoutidis, P. Language-oriented rule-based reaction network generation and analysis: Applications of RING. *Computers & Chemical Engineering* **2012**, *46*, 141–152.

(279) Rangarajan, S.; Kaminski, T.; Wyk, E. V.; Bhan, A.; Daoutidis, P. Language-oriented rule-based reaction network generation and analysis: Algorithms of RING. *Computers & Chemical Engineering* **2014**, *64*, 124–137.

(280) Fishtik, I.; Callaghan, C. A.; Datta, R. Reaction Route Graphs. I. Theory and Algorithm. *The Journal of Physical Chemistry B* **2004**, *108*, 5671–5682.

(281) Fishtik, I.; Callaghan, C. A.; Datta, R. Reaction Route Graphs. II. Examples of Enzyme- and Surface-Catalyzed Single Overall Reactions. *The Journal of Physical Chemistry B* **2004**, *108*, 5683–5697.

(282) Fishtik, I.; Callaghan, C. A.; Datta, R. Reaction Route Graphs. III. Non-Minimal Kinetic Mechanisms. *The Journal of Physical Chemistry B* **2005**, *109*, 2710–2722.

(283) Rao, S.; van der Schaft, A.; Jayawardhana, B. A graph-theoretical approach for the analysis and model reduction of complex-balanced chemical reaction networks. *Journal of Mathematical Chemistry* **2013**, *51*, 2401–2422.

(284) Sakamoto, A.; Kawakami, H.; Yoshikawa, K. A graph theoretical approach to complex reaction networks. *Chemical Physics Letters* **1988**, *146*, 444–448.

(285) Maio, F. D.; Lignola, P. KING, a KInetic Network Generator. *Chemical Engineering Science* **1992**, *47*, 2713–2718.

(286) Hillewaert, L. P.; Dierickx, J. L.; Froment, G. F. Computer generation of reaction schemes and rate equations for thermal cracking. *AIChE Journal* **1988**, *34*, 17–24.

(287) Temkin, O. N.; Bonchev, D. G. Application of graph theory to chemical kinetics: Part 1. Kinetics of complex reactions. *Journal of Chemical Education* **1992**, *69*, 544.

(288) Temkin, O. N.; Zeigarnik, A. V.; Bonchev, D. *Graph Theoretical Approaches to Chemical Reactivity*; Springer Netherlands, 1994; pp 241–275.

(289) Temkin, O. N.; Zeigarnik, A. V.; Bonchev, D. G. Application of Graph Theory to Chemical Kinetics. Part 2. Topological Specificity of Single-Route Reaction Mechanisms. *Journal of Chemical Information and Modeling* **1995**, *35*, 729–737.

(290) Zeigarnik, A. V.; Temkin, O. N.; Bonchev, D. Application of Graph Theory to Chemical Kinetics. 3. Topological Specificity of Multiroute Reaction Mechanisms. *Journal of Chemical Information and Computer Sciences* **1996**, *36*, 973–981.

(291) Stegelmann, C.; Andreasen, A.; Campbell, C. T. Degree of Rate Control: How Much the Energies of Intermediates and Transition States Control Rates. *Journal of the American Chemical Society* **2009**, *131*, 13563–13563.

(292) Baek, B.; Aboiralor, A.; Wang, S.; Kharidehal, P.; Grabow, L. C.; Massa, J. D. Strategy to improve catalytic trend predictions for methane oxidation and reforming. *AIChE Journal* **2016**, *63*, 66–77.

(293) Aghalayam, P.; Park, Y. K.; Vlachos, D. G. Construction and optimization of complex surface-reaction mechanisms. *AIChE Journal* **2000**, *46*, 2017–2029.

(294) Sjöblom, J.; Creaser, D. New approach for microkinetic mean-field modelling using latent variables. *Computers & Chemical Engineering* **2007**, *31*, 307–317.

(295) Karst, F.; Maestri, M.; Freund, H.; Sundmacher, K. Reduction of microkinetic reaction models for reactor optimization exemplified for hydrogen production from methane. *Chemical Engineering Journal* **2015**, *281*, 981–994.

(296) Blau, G.; Lasinski, M.; Orcun, S.; Hsu, S.-H.; Caruthers, J.; Delgass, N.; Venkatasubramanian, V. High fidelity mathematical model building with experimental data: A Bayesian approach. *Computers & Chemical Engineering* **2008**, *32*, 971–989.

(297) Hsu, S.-H.; Stamatis, S. D.; Caruthers, J. M.; Delgass, W. N.; Venkatasubramanian, V.; Blau, G. E.; Lasinski, M.; Orcun, S. Bayesian Framework for Building Kinetic Models of

Catalytic Systems. *Industrial & Engineering Chemistry Research* **2009**, *48*, 4768–4790.

(298) Sutton, J. E.; Vlachos, D. G. Building large microkinetic models with first-principles accuracy at reduced computational cost. *Chemical Engineering Science* **2015**, *121*, 190–199.

(299) Vilhelmsen, L. B.; Hammer, B. A genetic algorithm for first principles global structure optimization of supported nano structures. *The Journal of Chemical Physics* **2014**, *141*, 044711.

(300) Goedecker, S. Minima hopping: An efficient search method for the global minimum of the potential energy surface of complex molecular systems. *The Journal of Chemical Physics* **2004**, *120*, 9911–9917.

(301) Peterson, A. A. Global Optimization of Adsorbate–Surface Structures While Preserving Molecular Identity. *Topics in Catalysis* **2013**, *57*, 40–53.

(302) Zheng, W.; Zhang, J.; Xu, H.; Li, W. NH3 Decomposition Kinetics on Supported Ru Clusters: Morphology and Particle Size Effect. *Catalysis Letters* **2007**, *119*, 311–318.

(303) Sun, G.; Sautet, P. Metastable Structures in Cluster Catalysis from First-Principles: Structural Ensemble in Reaction Conditions and Metastability Triggered Reactivity. *Journal of the American Chemical Society* **2018**, *140*, 2812–2820.

(304) Packwood, D. M.; Hitosugi, T. Rapid prediction of molecule arrangements on metal surfaces via Bayesian optimization. *Applied Physics Express* **2017**, *10*, 065502.

(305) Todorovic, M.; Gutmann, M. U.; Corander, J.; Rinke, P. Efficient Bayesian Inference of Atomistic Structure in Complex Functional Materials. 2017.

(306) Matsuoka, T.; Baumes, L.; Katada, N.; Chatterjee, A.; Sastre, G. Selecting strong Brønsted acid zeolites through screening from a database of hypothetical frameworks. *Physical Chemistry Chemical Physics* **2017**, *19*, 14702–14707.

(307) Abdelkafi, O.; Idoumghar, L.; Lepagnot, J.; Paillaud, J.-L.; Deroche, I.; Baumes, L.; Collet, P. Using a novel parallel genetic hybrid algorithm to generate and determine new zeolite frameworks. *Computers & Chemical Engineering* **2017**, *98*, 50–60.

(308) Honkala, K. Ammonia Synthesis from First-Principles Calculations. *Science* **2005**, *307*, 555–558.

(309) Medford, A. J.; Vojvodic, A.; Hummelshøj, J. S.; Voss, J.; Abild-Pedersen, F.; Studt, F.; Bligaard, T.; Nilsson, A.; Nørskov, J. K. From the Sabatier principle to a predictive theory of transition-metal heterogeneous catalysis. *Journal of Catalysis* **2015**, *328*, 36–42.

(310) Curtarolo, S.; Hart, G. L. W.; Nardelli, M. B.; Mingo, N.; Sanvito, S.; Levy, O. The high-throughput highway to computational materials design. *Nature Materials* **2013**, *12*, 191–201.

(311) Bligaard, T.; Nørskov, J.; Dahl, S.; Matthiesen, J.; Christensen, C.; Sehested, J. The Brønsted–Evans–Polanyi relation and the volcano curve in heterogeneous catalysis. *Journal of Catalysis* **2004**, *224*, 206–217.

(312) Wang, Z.; Hu, P. Some Attempts in the Rational Design of Heterogeneous Catalysts Using Density Functional Theory Calculations. *Topics in Catalysis* **2015**, *58*, 633–643.

(313) Wang, Z.; Hu, P. Towards rational catalyst design: a general optimization framework. *Philosophical Transactions*

of the Royal Society A: Mathematical, Physical and Engineering Sciences **2016**, *374*, 20150078.

(314) Bligaard, T.; Nørskov, J. *Chemical Bonding at Surfaces and Interfaces*; Elsevier, 2008; pp 255–321.

(315) Studt, F.; Abild-Pedersen, F.; Bligaard, T.; Sorensen, R. Z.; Christensen, C. H.; Nørskov, J. K. Identification of Non-Precious Metal Alloy Catalysts for Selective Hydrogenation of Acetylene. *Science* **2008**, *320*, 1320–1322.

(316) Medford, A. J.; Lausche, A. C.; Abild-Pedersen, F.; Temel, B.; Schjødt, N. C.; Nørskov, J. K.; Studt, F. Activity and Selectivity Trends in Synthesis Gas Conversion to Higher Alcohols. *Topics in Catalysis* **2013**, *57*, 135–142.

(317) Kulkarni, A. R.; Zhao, Z.-J.; Siahrostami, S.; Nørskov, J. K.; Studt, F. Cation-exchanged zeolites for the selective oxidation of methane to methanol. *Catalysis Science & Technology* **2018**, *8*, 114–123.

(318) Schumann, J.; Medford, A. J.; Yoo, J. S.; Zhao, Z.-J.; Bothra, P.; Cao, A.; Studt, F.; Abild-Pedersen, F.; Nørskov, J. K. Selectivity of Synthesis Gas Conversion to $C2_+$ Oxygenates on fcc(111) Transition-Metal Surfaces. *ACS Catalysis* **2018**, 3447–3453.

(319) Ulissi, Z.; Prasad, V.; Vlachos, D. Effect of multiscale model uncertainty on identification of optimal catalyst properties. *Journal of Catalysis* **2011**, *281*, 339–344.

(320) Deshpande, S.; Kitchin, J. R.; Viswanathan, V. Quantifying Uncertainty in Activity Volcano Relationships for Oxygen Reduction Reaction. *ACS Catalysis* **2016**, *6*, 5251–5259.

(321) Wolcott, C. A.; Medford, A. J.; Studt, F.; Campbell, C. T. Degree of rate control approach to computational catalyst screening. *Journal of Catalysis* **2015**, *330*, 197–207.

(322) May, M. Companies in the cloud: Digitizing lab operations. *Science* **2017**, *355*, 532–534.

(323) Lausche, A. C.; Medford, A. J.; Khan, T. S.; Xu, Y.; Bligaard, T.; Abild-Pedersen, F.; Nørskov, J. K.; Studt, F. On the effect of coverage-dependent adsorbate–adsorbate interactions for CO methanation on transition metal surfaces. *Journal of Catalysis* **2013**, *307*, 275–282.

(324) Hoffmann, M. J.; Matera, S.; Reuter, K. kmos: A lattice kinetic Monte Carlo framework. *Computer Physics Communications* **2014**, *185*, 2138–2150.

(325) Meskine, H.; Matera, S.; Scheffler, M.; Reuter, K.; Metiu, H. Examination of the concept of degree of rate control by first-principles kinetic Monte Carlo simulations. *Surface Science* **2009**, *603*, 1724–1730.

(326) Hoffmann, M. J.; Engelmann, F.; Matera, S. A practical approach to the sensitivity analysis for kinetic Monte Carlo simulation of heterogeneous catalysis. *The Journal of Chemical Physics* **2017**, *146*, 044118.

(327) McGill, J. A.; Ogunnaike, B. A.; Vlachos, D. G. Efficient gradient estimation using finite differencing and likelihood ratios for kinetic Monte Carlo simulations. *Journal of Computational Physics* **2012**, *231*, 7170–7186.

(328) Núñez, M.; Robie, T.; Vlachos, D. G. Acceleration and sensitivity analysis of lattice kinetic Monte Carlo simulations using parallel processing and rate constant rescaling. *The Journal of Chemical Physics* **2017**, *147*, 164103.

(329) *Introduction to Scientific and Technical Computing*; CRC Press, 2016; pp 39–53.

(330) Nelli, F. *Python Data Analytics*; Apress, 2015; pp 237–264.

(331) Ketkar, N. *Deep Learning with Python*; Apress, 2017; pp 159–194.

(332) Bitzer, J.; Schröder, P. J. *The Economics of Open Source Software Development*; Emerald Group Publishing Limited, 2006.

(333) Larsen, A. H.; Mortensen, J. J.; Blomqvist, J.; Castelli, I. E.; Christensen, R.; Dułak, M.; Friis, J.; Groves, M. N.; Hammer, B.; Hargus, C.; Hermes, E. D.; Jennings, P. C.; Jensen, P. B.; Kermode, J.; Kitchin, J. R.; Kolsbjerg, E. L.; Kubal, J.; Kaasbjerg, K.; Lysgaard, S.; Maronsson, J. B.; Maxson, T.; Olsen, T.; Pastewka, L.; Peterson, A.; Rostgaard, C.; Schiøtz, J.; Schütt, O.; Strange, M.; Thygesen, K. S.; Vegge, T.; Vilhelmsen, L.; Walter, M.; Zeng, Z.; Jacobsen, K. W. The atomic simulation environment—a Python library for working with atoms. *Journal of Physics: Condensed Matter* **2017**, *29*, 273002.

Graphical TOC Entry

Some journals require a graphical entry for the Table of Contents. This should be laid out "print ready" so that the sizing of the text is correct. Inside the `tocentry` environment, the font used is Helvetica 8 pt, as required by *Journal of the American Chemical Society*.
The surrounding frame is 9 cm by 3.5 cm, which is the maximum permitted for *Journal of the American Chemical Society* graphical table of content entries. The box will not resize if the content is too big: instead it will overflow the edge of the box.
This box and the associated title will always be printed on a separate page at the end of the document.