Autonomous High-Throughput Computations in Catalysis

Stephan N. Steinmann, Angga Hermawan, Mohammed Bin Jassar, Zhi Wei Seh

Autonomous High-Throughput Computations in Catalysis

Stephan N. Steinmann,[1,*] Angga Hermawan,[2] Mohammed Bin Jassar[1,3] and Zhi Wei Seh[4,5*]

[1]Univ Lyon, ENS de Lyon, CNRS, Laboratoire de Chimie UMR 5182, Lyon, France
[2]Faculty of Textile Science and Technology, Shinshu University, 3-15-1 Tokida, Ueda City, Nagano, 386-8567, Japan
[3]Stellantis Centre Technique Vélizy A, 78140 VELIZY VILLACOUBLAY, France
[4]Institute of Materials Research and Engineering, Agency for Science, Technology and Research (A*STAR), 2 Fusionopolis Way, Innovis, 138634, Singapore
[5]Lead contact
*E-mail: stephan.steinmann@ens-lyon.fr; sehzw@imre.a-star.edu.sg

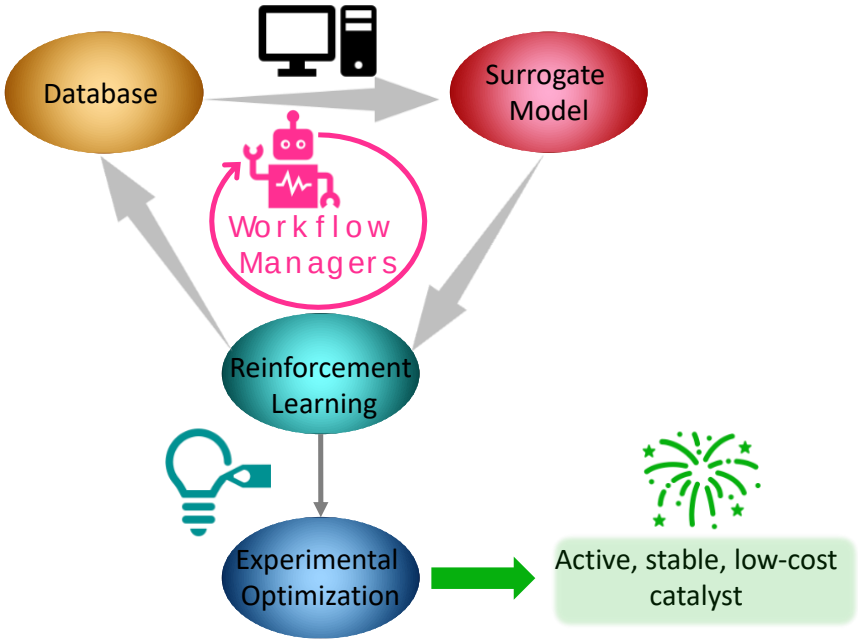
Bigger Picture

Autonomous atomic-scale computations can significantly accelerate catalyst development, but the necessary (software) infrastructure is not yet widely adopted or even known in the community. Particularly, reinforcement learning is well adapted to the needs of targeted catalyst development with reduced human cost. Accounting for catalyst stability under working conditions is challenging, but achievable by a combination of high-throughput computations and machine-learning. Autonomous computations should be complemented by autonomous laboratories, which allow high-throughput experimentation and catalyst optimization through machine-learning, but require large investments from research facilities.

Summary

Autonomous atomistic computations are excellent tools to accelerate the development of heterogeneous (electro-)catalysts. In this perspective, we critically review the achieved progress to accelerate high-throughput screening aimed at identifying promising catalyst materials, via databases, workflow managers and machine-learning techniques. Outstanding challenges are also discussed extensively: the modification and stability of catalyst surfaces under realistic reaction conditions is key for meaningful predictions. Furthermore, adequately accounting for solvent effects remains a topic of active research particularly relevant for biomass transformations and electrocatalysis. Finally, efficient, autonomous workflows for investigating active sites of amorphous catalysts remain underdeveloped. The computations can also be supplemented with autonomous laboratories, which allow to perform sophisticated experiments driven by artificial intelligence-augmented design of experiments, reducing human-time investment for optimizing synthesis and reaction conditions as well as catalyst characterizations. The combination of autonomous computations and laboratories promise to power the dearly needed transition to a sustainable chemical industry.

Graphical Abstract



Keywords

Heterogeneous catalysis; autonomous computations; autonomous laboratories; high-throughput screening; interface; electrocatalysis; machine-learning

UN Sustainable Goals

SDG9: Industry, Innovation and Infrastructure

Introduction

Heterogeneous (electro-)catalysis is at the heart of the production of chemicals, from refineries and petrochemicals to the manufacture of ammonia and sulfuric acid. While these processes have been studied and optimized over decades, the very recent political action, promising to drastically reduce CO_2 emissions, calls upon chemists to devise novel catalysts to achieve more efficient catalytic conversion. These catalysts not only need to be highly active, but also stable under long-term operating conditions.[1]

Two prototypical examples are water-splitting and biomass conversion.[2] Stable, energy-efficient water-splitting catalysts based on Earth-abundant materials (e.g., transition-metal dichalcogenides), in combination with renewable electricity, typically from wind or solar-power, would allow us to temporarily store energy from intermittent sources on a grid level. Such “green” hydrogen would also be a decarbonized chemical that can be used throughout the industry. In particular, the Haber-Bosch process, responsible for ammonia production and thus essentially all nitrogen atoms in chemical compounds from polymers to colorants, fertilizers and drugs, currently relies on hydrogen produced via methane reforming, releasing large amounts of CO_2.[3] Biomass conversion based on ligno-cellulosic feedstocks is also a

promising avenue for transitioning from the petrochemical industry to a renewable, carbon-neutral industry producing all the commodities (plastics, detergents, etc) via new pathways. In biomass conversion, the major challenge compared with the current petrochemical processes is that the bio-sourced molecules are highly oxygenated and much less volatile than the petrochemicals. Hence, catalysts working efficiently in water have to be developed.[4] Ideally, the two aspects (hydrogen production and biomass conversion) could be coupled for a sustainable energy future.[5]

Accelerating catalyst development with workflow managers and databases

In this section we discuss the various tools that have become available to accelerate atomistic computations relevant to the understanding and design of heterogeneous catalysis. The main focus is dedicated to workflow managers, but data repositories and automated transition state searches are also succinctly reviewed and summarized in Table 1.

Mass- (and charge-) transport have a significant impact on the overall kinetics and depend on the meso-scale properties of the catalytic layer as illustrated for CO_2 reduction[6] and solar hydrogen production.[7] However, the intrinsic activity of a catalyst is determined by its atomic-scale structure and sets an upper maximum to the performance. In heterogeneous (electro-)catalysis, this intrinsic activity, which is the focus of the current perspective, is intimately connected to the complex interface between the solid catalyst and either gas-phase (typical for petrochemistry) or the (aqueous) liquid-phase, typical for electrocatalysis and biomass conversion. Note, that our perspective mostly covers metallic and transition-metal oxide/sulphide catalysts, but does not address porous catalysts such as zeolites and metal-organic frameworks.

The kinetics over heterogeneous catalyst has been extensively studied by computations over the last decades, most of the time under idealized conditions.[8,9] Typically, the catalyst model is a single-crystal surface on which a single adsorbate is placed. Furthermore, the vast majority of studies focuses on small (1-2 carbon atoms) reactants and intermediates. These simplified surface models lend themselves for high-throughput in silico screening studies,[10,11] which generally require a workflow manager. The main purpose of a workflow manager in computational chemistry is to orchestrate all the computations to run on supercomputer facilities, tracking failures, correcting the most basic errors and dynamically adapting settings when necessary. In general, the workflow manager can also be used to construct the starting geometries and to chain computations, e.g., screening adsorption modes and only performing frequency computations for the lowest energy ones. Furthermore, the workflow manager can also be responsible to store the results in a database and thus to avoid redundancies.

In practice, several workflow managers have either been adapted from material screening to screening of adsorption modes, or specifically developed for this purpose. One of the most abstract and complete workflow managers is AiiDA,[12,14] which not only automates the execution of workflows and the storage of the generated data, but also scrupulously keeps track of the provenance of the data along a workflow and allows integrated storage of the obtained results on the MaterialsCloud repository.[16] An alternative is provided by AFLOW,[18,20] which is a high-throughput computational framework with an associated database,[22] geared towards bulk properties of materials, but is also capable of generating and analyzing surface

free energies for various high-index surfaces in complex multi-component compounds and nanoparticles. AFLOW is also directly applicable to estimate stabilities of explored structures, e.g., by automatically determining formation energies and constructing binary convex hulls for alloys. Similarly, the MaterialsProject[24] has developed a suite of programs to manage workflows (called FireWorks[13]), generate input files (Pymatgen[15]) and correct common failures (Custodian). Failure management is one of the difficult to efficiently automate aspects of computational catalysis: On the one hand there are technical failures (errors in input files, failures of hardware) and on the other hand there are unphysical results due to convergence failures (electronic structure or geometry) or initial geometries that are too far from the intended optimized geometries. Finally, the ASE toolkit[28] also provides the possibility to create and manage complex workflows if the user creates the super-structure: ASE provides all the necessary elementary bricks, but is not a workflow manager in itself. For example, the Atomic Simulation Recipes rely on ASE to carry out "routine" workflows, called "recipes".[30] The interface between the recipe and the high-performance computing facility that carries out the individual tasks is provided by MyQueue.[32] Atomate[34], which is a bundled version of the MaterialsProject tools, has, for instance, been specifically extended in order to allow high-throughput screening of CO_2 reduction reaction intermediates (i.e., CO) on semi-conductor surfaces, automatically identifying surface terminations, their relative surface energies and optimizing the adsorbates on automatically identified adsorption sites.[17] Similarly, the Generalized Adsorption Simulator for Python, GASpy, has been developed to rapidly screen surface free energies and adsorption energies of small molecules on (inter-)metallic surfaces.[37] The exploitation of graph-theory allows CatKit[39] to feature the unique capability of explicitly handling bi-dentate adsorption modes of small molecules, significantly expanding the applicability and efficiency in heterogeneous catalysis applications. Bayesian Optimisation Structure Search, BOSS, accelerates the orientational exploration of adsorption modes via a Bayesian surrogate model, thus reducing the number of DFT computations required to identify low-energy adsorption modes.[86]

In contrast to rigid, small adsorbates that can be handled by several adsorption generation tools, flexible, polyfunctional molecules, typical for biomass, pose the challenge of dealing with the vast phase-space of adsorption configurations. Screening this configurational space has very recently been proposed based on a reductionist's approach to the problem, i.e., converting it into a series of chemically intuitive "decisions". The following efficient, though transparent, approach has been implemented in DockOnSurf:[42] First, the conformational space is sampled in the absence of the surface. Then, the conformations are adsorbed on the surface by combining possible anchoring points (functional groups on the molecule) with adsorption sites on the surface, together with a rotational sampling to explore the relative orientations of the molecule with respect to the surface normal. Finally, non-physical configurations (where the adsorbate collides with the surface) are identified, adjusted or removed. DockOnSurf has, for instance, automatically generated low-energy configurations of sorbitol on hematite. In this example, the lowest energy conformation in the gas-phase leads to an adsorption mode that is 0.9 eV above the most stable adsorbate conformation. Inspection of the different adsorption modes reveals that internal hydrogen-bonding inherited from the gas-phase conformation makes several functional unavailable for strong interactions with the surface. In contrast, a conformer with a higher moment of inertia (2500 vs 2000 amu $Å^2$), which exposes more of its functional groups, leads to the lowest adsorption

configuration. This demonstrates that adsorbing gas-phase configurations other than the lowest-energy conformer is crucial for polyfunctional molecules.

Workflow managers have been instrumental for producing large databases, mostly dedicated to materials properties. For an extensive review on materials databases the reader is referred to ref [19]. For example, the database of Materials Project contains elastic tensors for more than 1200 materials, which has been driven by a Fireworks workflow and covers about five times more inorganic materials than all experimentally known databases on elastic tensors[13]. Similarly, ASR has been used to construct a dedicated database of 2D materials (C2DB).[21] C2DB contains 4000 entries, covering a large variety of materials, their energies and a large range of properties, from atomic charges and vibrational spectra up to photophysical properties relevant to photo-electrocatalysis such as effective masses, exciton binding energies, etc. Analogously, the exfoliation energies of more than 5000 2D materials have been computed by an AiiDA driven workflow and deposited on the MaterialsCloud repository.[23] Another typical example is the Open Quantum Materials Database (OQMD),[47] which is an extensive database of crystal structures and provides computational estimates of their formation energies, which can be conveniently compared to experiment. To build this database, the authors have developed an on-purpose workflow manager (qmpy). The most relevant purposedly-built database for heterogeneous catalysis comes from the Open Catalyst Project and is currently called OC20.[25] This database combines surfaces from more than 5000 unique unary, binary and ternary bulk materials retrieved from Materials Project with 86 adsorbates (ranging from H, N, O to functionalized C_2 and N_2 species). OC20 has been constructed based on a combination of established libraries such as pymatgen and CatKit[39], the latter for managing the adsorption process. Finally, we also mention NOMAD-lab,[82] which is a user-driven database, with associated tools for analysis and machine-learning. The central idea behind NOMAD-lab is that the larger the database, the more powerful it is and that computational chemists around the world are constantly generating data that could be integrated in such an overarching database, fully embracing the “big-data” paradigm. In other words, in contrast to the “purpose-built” databases mentioned above, NOMAD-lab is a repository for long-term storage and sharing of computational data. While this means that data is heterogeneous (setups, codes, etc), the initiative augments individual efforts of high-throughput computations by collecting and sharing the various datasets. This diversity of origins of data allows comparisons and cross-validations between various methods and computational setups.

Beyond the determination of the adsorption modes and corresponding thermodynamics of a reaction scheme, determining reaction kinetics requires the determination of activation energies. Automating, or at least accelerating, the location of transition states is an active field of research. For example, PathFinder has been specifically developed to speed up the location of transition states in ionic materials, exploiting charge density profiles to improve the initial guess of the minimum energy pathway.[26] In a related spirit, Opt'n Path[27] performs the interpolation not in the commonly used Cartesian coordinates but in chemically meaningful internal coordinates, providing sensible starting points for initializing nudged-elastic band computations[29] or any of the accelerated equivalents, such as the approxNEB[26] reflective-middle-image NEB[31], the Bayesian-augmented NEB[33] or the Gaussian Process Regression augmented NEB.[35] In a comparison for CO_2 activation by platinum, the later method has been found to be particularly robust among the accelerated NEB methods.[36] Last

but not least, AFIR,[58] a computationally very intensive, but fully automated procedure to locate unknown transition states, has also been transposed to heterogeneous catalysis. Applied to the oxidation of CO on Pt(111), the method was capable of automatically identifying relevant diffusion pathways and various C-O bond formation mechanisms, leading to CO_2 or CO_3, spanning 133 local minima, 272 approximate and 26 true transition states. The latter have been selected based on a microkinetic model identifying the kinetic bottlenecks in the reaction network.[38]

Merging machine-learning and workflow managers: autonomous computations

The following section discusses how machine learning and in particular reinforcement learning can be combined with workflow managers in order to reach autonomous computations, i.e., efficient, self-regulating workflows, limiting "useless" computations.

High-throughput computations, as discussed above, are very powerful. Nevertheless, in order to efficiently search through the vast chemical space of possible catalysts, the screening can, and should be, accelerated by surrogate models (see Figure 1 for a schematic overview). Surrogate models are mathematical functions that try to predict the outcome of a costly computation based on much simpler to obtain descriptors, such as geometric fingerprints, graphs or intrinsic material properties such as the d-band center. These surrogate models nowadays mostly rely on machine learning (ML),[40] gradually replacing their ancestor which is a linear scaling relationship.[41] Despite this trend towards sophisticated ML, the prediction of transition state energies for hydrodeoxygenation (particularly relevant for biomass conversion) demonstrated that linear models lead to essentially the same quality as more advanced, non-linear models.[43] Nevertheless, in general, machine-learning holds high promises for accelerating molecule, material,[44,45] drug[46] and catalyst[48,49] development based on a data-driven paradigm. Note that surrogate models in the context of catalysis design can be classified into two broad categories: (i) interatomic potentials (also called machine learning potentals, MLP) that are used as an alternative to DFT and (ii) effective models that are used to circumvent energy evaluations of atomistic models altogether. In the remainder of the text we are concerned with the latter approaches unless stated otherwise. Combining surrogate-models with more expensive DFT computations allows to arrive at autonomous computational workflows, where the surrogate models are used to decide which additional computations need to be carried out to either (in)validate the prediction of the model or to make the model more robust (see Figure 1). Such a surrogate-model driven workflow has, for instance, been applied to discover novel (nearly stable) binary materials,[50] and to the screening of hydrogen evolution/oxidation catalysts,[51] where the compromise between activity and stability is key (see Figure 2a and 2b). The same basic principles can be applied to reaction networks, where the surrogate model is used to predict adsorption and activation energies, while microkinetic and DFT simulations are used to ascertain rate-limiting steps and their energetics. This scheme is illustrated by Figure 2c, where the reaction network of syngas to C_2 species over Rh(111) is extensively studied. Even though the full reaction network is considered, only about 50% of the intermediates and even only 10% of all transition states have been explicitly been determined, while the energetics of the others are estimated via surrogate models.[52]

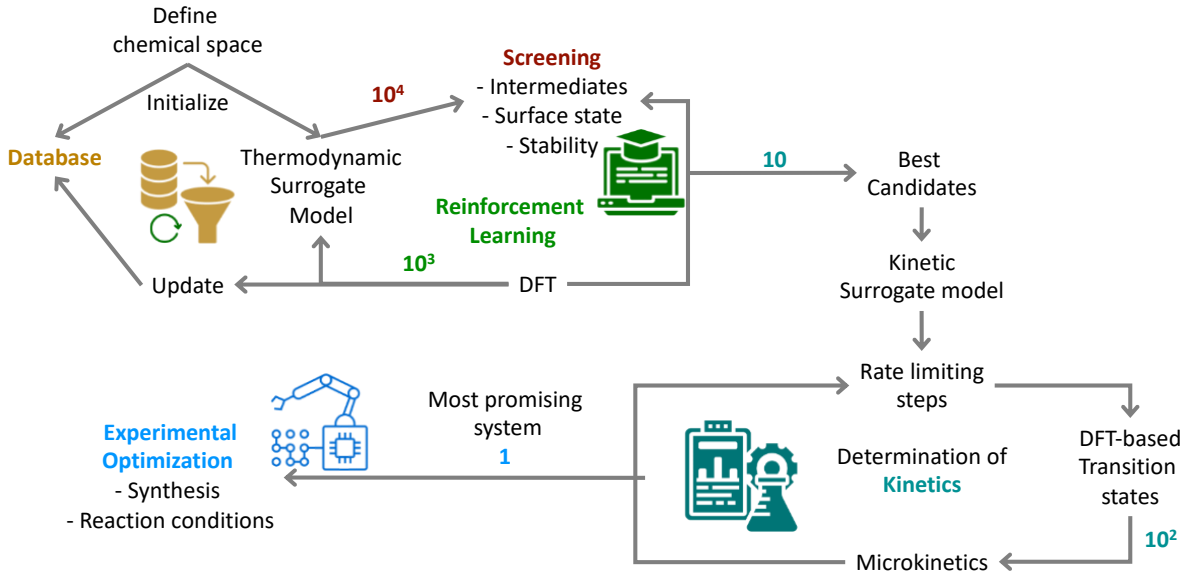


Figure 1. Overall schematic for autonomous catalyst discovery up to experimental testing. Numbers give rough estimates of considered systems and/or computations. Key aspects are highlighted in color, starting from a database (gold) that is continuously updated by a workflow manager, surrogate-model based screening (dark red) of active sites and the optimization of promising leads via reinforcement learning (dark green). The kinetics of the most promising candidates have to be assessed (teal), before experiments optimize (typically relying on Bayesian optimization) the synthesis and reaction conditions (light blue).

In order to deliver on the promise of accelerating catalyst design, we now have to transition from "proof-of-principle" demonstrations, where the goal was to demonstrate the power of ML (like the usage of OC20 for training graph-convolutional neural networks for replacing DFT energy evaluations by MLPs[25]), to problem-driven applications, i.e., actually developing novel catalysts, such as demonstrated for CO_2 electroreduction.[53] In this context, generative models become a corner stone. A generative model is a function proposing structures outside of the training set. In the simplest case a generative model is able to combine a given number of building blocks. For molecules, it would, for example, be able to suggest adding functional groups. For a catalyst, it could propose surface modifications (adsorbates, defects) and substitutional doping. Of course one can load a database with potential active sites and screen their activity for a given reaction. However, first, the particular reaction conditions should be automatically taken into account so that only stable catalysts and their relevant surface state are considered; and second, it would be much more powerful to organize the chemical universe of catalysts in an explorable manner. While this has been successfully achieved for organic molecules[54,55] and, as an extension, partially to porous solids with organic building blocks,[56] to the best of our knowledge no equivalent algorithm exists for metals, sulphides, oxides, MXenes, etc. The reason for this lack of generative models is enrooted in chemistry: for organic chemistry, the algorithm can rely on well-established valence-based rules, while such a simple classification does not apply to solids and their surfaces. Similarly, the large diversity of surfaces, and thus of potential active sites, limits "universal" generative models that would allow prediction of the adsorption site and geometry of adsorbates on the catalyst, in analogy to what has been achieved for conformations of organic molecules.[76] Once reasonable structures are available, running the corresponding DFT computations in a high-throughput workflow is feasible, limiting useless computations. Then, the DFT results can be

conveniently analyzed by ab initio thermodynamics[78] to identify the resting state of the catalyst and the energy requirements to create vacancies in the adsorbate layer etc.

In the absence of these "universal" generative models, and in combination with the "problem-driven" ML studies, we are advocating to increasingly rely on reinforcement learning (RL) techniques. The core of reinforcement learning is the policy that is being learned which maximizes a defined reward. This technique has been very successful for learning strategy games: for example AlphaGo is based on RL.[57] In chemistry, RL is not yet widely adopted by the community, but we believe that this will change soon: RL has already proven its efficiency in optimizing various chemistry-relevant functions. For instance, RL has been trained not only for efficiently optimizing geometries of organic molecules,[59] but also for determining the lowest energy pathway in the complex Haber-Bosch reaction over Fe(111), effectively learning chemical kinetics.[60] Furthermore, RL has been applied to construct efficient training sets for MLPs[61,62] and even to identify efficient, highly accurate, compressions of wave functions.[63] The power of RL relies on two complementary aspects: On the one hand, RL is typically applied as an active learning framework, i.e. the training set is constructed on the fly, according to the needs of the model being trained and the promising regions being explored. Active learning limits the number of "useless" computations, so that only comparably small training sets are required. Being able to train surrogate models with small training sets is all the more important when applying ML to solve novel problems for which the required training set is not established beforehand. RL is also ideally suited for transfer learning: Having learned an optimal policy for one problem might be an excellent starting point for learning an optimal policy on a related problem. This has, for instance, been applied to the wave function compression, where the policy for optimal selection of Slater determinants around equilibrium distance has been shown to significantly accelerate the identification of the optimal combination of Slater determinants for stretched distances.[63] Transfer learning is also closely linked to combining information from different sources. Take, for instance, surrogate models, DFT computations and experiment. Combining these three sources of information in an optimal way is not obvious, but under the condition that they show reasonably similar trends, the algorithms on how to minimize the overall cost for solving a given optimization problem have now been developed.[87] Moreover, given that RL is, by construction, problem and system oriented, devising generative models is more manageable: Within a class of systems it is rather natural to select the sites of interest, both for surface modifications (typically substitutional doping) and adsorption sites.

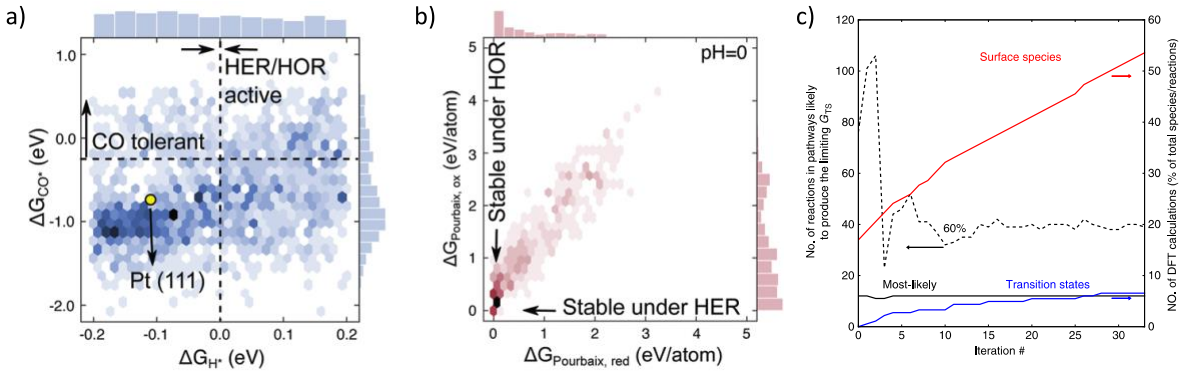


Figure 2. Surrogate-model based estimates of a) activity and b) stability, of hydrogen evolution and oxidation catalysts. Taken from ref [51]. c) Representation of the convergence of the syngas to C2 species over Rh(111)

network. At each iteration, DFT computations are added for key intermediates and transition states. Less than 10% of transition states and 60% of the intermediates are required. Taken from ref [52].

Challenges to be addressed

In the following section, we outline four major challenges for autonomous workflows applied to catalysis: (i) solvation effects, (ii) maintaining a curated database and making it efficiently searchable via well-crafted descriptors, (iii) efficient workflows for complex interfaces, including surface states reflecting the reaction conditions, (iv) workflows for exploring amorphous catalysts.

In the next 5 years or so, the community will have to embrace the challenge of dealing with the solvent: While implicit solvents have been available in the molecular community for more than thirty years,[64] only the last 5-10 years have seen their implementation and application in heterogeneous catalysis.[65] While these implicit solvent models are ideally suited for high-throughput computations and very valuable for electrocatalysis[66], they tend to significantly underestimate solvation effects, especially on metals, where the electrostatic interactions (captured by implicit solvents) is negligible compared to the near-chemisorption of solvent and complex many-body effects are non-negligible even for small adsorbates.[67–69] Therefore, competitive adsorption between general adsorbates and solvent molecules has to be considered.[70] The development of generally applicable hybrid molecular mechanics/quantum mechanics solvation models[70–72] or approximations thereof[73,74] is likely to become the most adapted strategy, provided transferable force fields with an acceptable accuracy can be devised.[75] Alternatively, purposedly built MLPs (as very recently illustrated in the context of electrocatalysts[77]) could become standard practice once a robust autonomous workflow for their fitting has been developed. Ultimately, these computationally rather heavy solvation models might be replaced by some surrogate ML models in the spirit of implicit solvent models, but with improved accuracy compared to the electrostatics-based[101,103] current state of the art. However, given the scarcity of benchmark data, we believe that this will only come in a second stage, estimated to be 5-10 years ahead of us, when atomistic models have been made available to generate the necessary benchmark data.

Given the data-driven nature of ML, one might argue that establishing and maintaining large database is key. This not only allows to avoid duplicating computations, but allows to train models with existing data, for instance to initialize surrogate models.[79] In the context of catalysis, the databases are most useful for defining the search space: Which types of surfaces exist? – What types of active sites could be envisioned? Currently, the largest such database is OC20.[25] To unleash the power of such databases, the user should also be able to judge the stability and feasibility of a given structure. Hence, a "feasible surface" database with experimentally observed surface structures would be particularly valuable, together with the corresponding conditions. Depending on the in situ/operando experimental characterization technique, the surface state can now be precisely determined, especially with scanning-tunneling microscopy or transmission electron microscopy.[106] Establishing the corresponding database could be seen as an analogue to the protein data bank[80] or the open reaction database[81] but for surfaces. The user might then investigate the evolution of the structure under his particular conditions or trying to adapt the reaction conditions towards the ones under which the promising surface has already been observed. In line with the move towards open-science and big-data, the community as a whole would also benefit from one, unified

database which will be continuously fed by the computations performed and published, which is the goal of NOMAD-lab[82]. Having one, huge, database would also enable automatic, self-driven identification of outliers (i.e., wrong/unconverged computations), so that extracting curated datasets would become easy. Even beyond dataset curation, robust data consistency checks, in order to filter out erroneous computations, will need to be developed in order to fully benefit from autonomous computations. On a related topic, the development of suitable descriptors (graphs, fingerprints) for surfaces would make these databases more searchable and would allow to classify the entries: well-crafted descriptors make surrogate models more robust and thus more useful. Typically used descriptors are simple cut-off distances,[83] Behler-Parinello's symmetry functions[84] or the smooth overlap of atomic positions.[85] Improved descriptors (see ref [113] for a comparison) might ultimately lead to powerful generative models that allow to efficiently explore chemical space.

Active sites not resembling a single-crystal surface are still very challenging to address. First of all, the active site might be at a triple-phase boundary (catalyst, support and reaction medium). While such systems have been regularly described in the literature, high-throughput construction and evaluation of their activity has, to the best of our knowledge, not been reported. Nevertheless, the BOSS scheme provides a promising avenue towards an efficient construction of complex interfaces as illustrated by the C_{60}-TiO_2 interface,[86] or the tetracyanoethylene/Cu(111) interface.[88] Similarly, MPInterfaces allows one to efficiently construct solid/solid or solid/ligand interfaces.[116] Second, the surface state might evolve under the reaction conditions:[89] ranging from coverage with solvent or reaction intermediates to surface reconstruction or even catalyst decomposition/dissolution. These modifications of the surface state by the reaction medium are still inadequately tackled, even though they lend themselves for autonomous computations, as adding/removing adsorbates can be automatized efficiently. An instructive example on how the reaction conditions can influence the reactivity is provided by the evolution of Pt_7 clusters supported on γ-alumina under C-H activation conditions. Extensive computations predict that the cluster dynamically reconstructs as a function of the reaction progress, changing shape and hydrogen coverage.[90]

Finally, amorphous catalysts can, from an experimental point of view, be as convenient as crystalline ones, or even more promising due to potentially higher specific surface area. It is only very recently that grain-boundaries can be conveniently constructed automatically.[120] From an atomistic modelling point of view, amorphous catalysts are very delicate to address (see ref [91] for an remarkable study of Pt nanowires as hydrogen evolution catalysts and ref [92] for the extensive modelling of high-entropy alloys), as not even the starting position of the atoms is well known and a validation of a given model compared to experiment is necessarily very indirect. For example, MoS_3 is discussed as a potential hydrogen evolution catalyst, but its atomic structure remains debated with various experiments leading to characteristics that are difficult to reconcile with a single atomistic model.[93]

Towards self-driven experimental laboratories

In this final section we review the progress in autonomous laboratories, i.e., hardware that is managed by artificial intelligence, enabling high-throughput, reproducible catalysis synthesis, characterizations and performance tests.

The impressive progress in terms of conceptualization and building both the software[125] and hardware solutions for autonomous experimentations offers a bright perspective to the field of heterogeneous (electro-)catalysis. In particular, autonomous laboratories have been shown to be well suited to optimize a given chemical system. In other words, the boundaries of the problem need to be defined by the user, while the "recipe" of the synthesis is optimized via a surrogate model with experimental factors as inputs. This has been, for example, demonstrated for the synthesis of nanoparticles with a specific shape.[94] Note, that even the shape analysis can be fully automatized by coupling automated TEM with machine-learning image processing techniques.[95] A similar approach has also been demonstrated for optimizing the synthesis protocol of few-layer WTe_2,[96] or for determining optimal organic additives for Cu-based CO_2 reduction electrocatalysts.[97] These studies also perfectly illustrate that the experimental search space is much more restricted (only one material) compared to computational screenings (see also Figure 1). However, the feasibility of given active sites can only be estimated on a very rough level based on computations, so that dedicated experiments are required to tune synthesis conditions that might be able to lead to the sought-after active sites and validate the results experimentally.[98,99] Autonomous high-throughput experimentation has been developed several years ago to explore a vast chemical compounds space, e.g., metals, organics, organometallics, inorganic solids, etc.[100] As a result, researchers were able to discover new chemical compounds with more exotic properties than that of conventional synthesis and have obtained big data on catalysts at an unprecedented scale and speed.[102] However, since high-throughput experimentation is more cost intensive that the conventional approach, design of experiments involving catalyst synthesis, characterization and testing should be carefully planned to maximize information output with minimum number of experiments.

With rapid progress in automation control, combinatorial high-throughput synthesis of catalysts can be made fully autonomous with the aid of a guided-robotics arm connected to the control system. Thus, optimized parameter conditions, e.g., catalyst composition, mixing sequence, reaction treatments, and so on can be obtained with minimal human intervention.[104,105,107] For instance, sputtering is a versatile process for catalyst synthesis in both laboratory and industrial scale, enabling deposition of a thin film of catalyst with controllable thickness. High-throughput depositions allow creating a library of catalyst materials with a controllable composition gradient and a large range of film thickness.[108] This is usually done using robotic arm-assisted sputtering in which the precursors are sputtered through a series of masks consisting of some overlaying masks. The masking combination is dependent on the pre-designed catalyst composition. For instance, combinatorial synthesis of binary catalysts by sputtering deposition uses binary masks made of primary and secondary masks. For ternary catalyst composition, a ternary mask is used with the composition of primary, secondary, and tertiary masks. The outcomes of thin-film library materials entirely depend on the composition of the catalysts and the reaction condition under which the experiments are conducted. The approach is suitable for synthesizing ternary, quaternary, or even higher-order mixtures of elements to produce a thousand catalysts in a parallel reactor in a concise amount of time.[99]

On the other hand, pulsed laser deposition offers rapid and homogeneous deposition of many materials using an ablation from a high-energy UV laser. This method has also been adopted for the combinatorial catalyst library, which uses a typical series of quaternary masks in a so-

called multi-plume pulsed-laser deposition system.[109] The autonomous robotic arm helps to rotate the samples holders that usually house pellet precursors and subsequent transfer for post-treatment or characterization. The use of robotic arms is also employed in the sol-gel synthesis of catalysts. A library of catalysts is typically prepared autonomously by robotic arm and pipette to take precursor solutions and transfer them to small vials (2-5 ml capacity) as microreactors in which the sol-gel reaction occurs.[110] More recently, jet dispensing equipped in the automatic printing technology (see Figure 3a) was utilized for high-throughput synthesis of a library of cocrystals.[111] Precursor ink was formulated with a predetermined concentration to prepare the gradient library in parallel. The method guarantees a faster speed of fluid dispending with a highly accurate compositional gradient, thus reducing the amount of experimentation and saving production time. The record has achieved 1,000,000 formulations within one operating hour.[111]

While combinatorial synthesis involves the preparation of vast arrays of the gradient materials, high-throughput characterization accelerates the discovery of structure-property relationships.[112] High-throughput characterization can be made fully autonomous by the robotic arm. An example is the development of automated rotating sample changer for X-ray diffraction to identify crystallographic features of catalysts. In D8 ADVANCE, developed by Bruker, which can measure up to 90 samples in parallel, the robot arm transfers the sample to the rotation sample stage, allowing permanent rotation and automatic positioning adjacent to the X-ray beam.[114] It can be equipped with AUTO-CHANGER consisting of a loading station for up to 6 sample magazine towers, a robotic sample handler with integrated gripper, and a rotation sample stage mounted to the goniometer. When a magazine tower is loaded or removed for refilling the new set of samples, the machine automatically detects it. Moreover, the handling gripper and transfer robot ensures safe transportation of sample to and from the measurement setup. For crystallography under cryogenic conditions, RoboDiff has been developed and has processed more than 20,000 molecules, including catalysts (see Figure 3b).[115] A robotic sample changer has also been installed in a small-angle X-ray scattering setup that can perform hundreds of measurement automatically with small sample consumption (± 5 μL).[117]

Raman spectroscopy is a powerful and non-destructive tool to obtain surface properties of catalysts and elucidate reaction mechanisms.[118] In a modern high-throughput Raman technology setup, a robotic system is employed to move samples and acquire data. This is typically done by the deposition of molecular or solid catalysts onto the multi-well plate attached to an automated sample stage.[119] Achieving laser beam focus is one of biggest challenges in measuring high-throughput Raman spectra for non-experts, and thus autofocus technology is developed to allow laser beam refocusing during sample holder rotation.[121] Several advanced Raman technologies such as UV resonance Raman spectroscopy, surface-enhanced Raman spectroscopy, time- and spatially resolved Raman spectroscopy can also gather information on how catalytic mechanisms occur by probing the solutions or reactions intermediates for catalytic CO_2 reduction, water splitting, water purification, etc.[122]

High-throughput catalyst testing is of great significance in accelerating catalyst development.[123] For instance, automated analysis of catalytic products can be performed by gas chromatography-mass spectrometry (GC-MS) and high-performance liquid chromatography, using a robotic handling pipette that is designed to provide reliable and

accurate liquid injection, sample preparation and pretreatment. Having a miniaturized electrochemical workstation is convenient for conducting parallel catalyst testing, which aims to shrink chemical laboratories to lab-on-a-chip system. Microfluidic reactors are sophisticated setups used to test catalyst activity. Their advantages are their versatility, small volumes, fast operation speeds, capability of parallelization, as well as well-controlled parameters (e.g., temperature, pressure, etc).[124] For example, researchers have studied a gradient catalyst consisting of Cu, Pd and Au ($Cu_xPd_yAu_{(1-x-y)}$ alloy) connected to individual microfluidic channels, where each end of channel is accessible by a programmable and movable liquid handling robot-equipped GC-MS nozzle which rapidly screens 100 H_2/D_2 exchange products within 10 minutes.[126]

For larger catalytic systems, a lab-made high-throughput catalyst testing setup has been established for oxidative methane coupling which screens 20 catalysts under 216 different conditions.[123] This can be achieved by the use of pneumatically actuated diaphragm valves in autosampler which is connected to a diaphragm pump and the inlet of a quadrupole mass spectrometer (see Figure 3c). Developing a parallel module that evaluates the activity of solid-state catalysts simultaneously is one of the future interests in the field. The 16-parallel high-throughput reactor systems, developed by hte GmbH, are specially designed to screen multi-catalytic reactions in a wide range of process parameters, to operate in gas and liquid feed streams, to perform in plug-flow, fix-bed, and trickle-bed reactors, and to couple with online GC-MS or offline analysis. It covers a broad range of catalyst volumes from small quantities of solid powders to massive quantities of shaped materials.[127] All of these operations are autonomous to reduce human error during preparation, handling, and testing of catalysts.

Finally, all the experimental results, including the raw data and metadata containing the experimental parameters, should be transferred automatically to cloud-based servers, which can then be analyzed by automated-data analysis and visualization tools.[128] The most critical role of high-throughput experimentation is to find the structure-activity relationship of the catalyst. Thus, the algorithm developed for automated data analysis should estimate or predict the optimum synthesis condition as feedback to the high-throughput experimentation. This is an ideal concept of data-guided combinatorial synthesis and data-driven catalysts discovery. One can extend this framework to additional parameters to make further predictions, enabling faster and more efficient catalyst development.

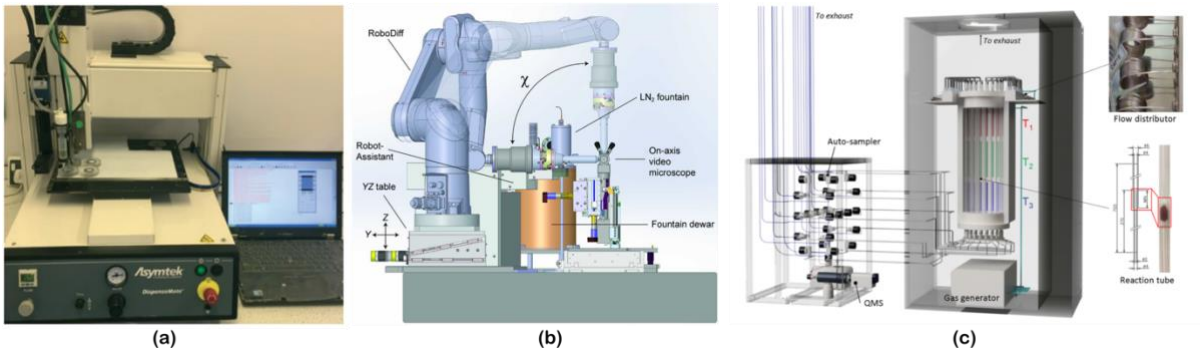


Figure 3. High-throughput experimentation for catalyst development. a) Jet dispensing printing technology. Taken from ref [111]. b) RoboDiff for cryogenic diffraction measurement. Taken from ref [115]. c) High-throughput catalyst testing system for oxidative coupling of methane. Taken from ref [123].

Table 1 Synoptic table of workflow managers and related tools.

Tool	Brief description of main characteristics	Ref.
AiiDA	High-level workflow manager, keeping track of data provenance. Can be coupled to the MaterialsCloud repository.[16]	[12,14]
AFLOW	Workflow manager for high-throughput in silico screenings. Results can be easily stored in AFLOWlib.[22]	[18,20]
Atomate	Workflow manager, input file generation and failure management developed by the MaterialsProject.[24] The workflow can move from bulk structures to surfaces, enumeration of adsorption sites, adsorbing small adsorbates on them and computing a wide range of properties.	[34]
ASE	Toolkit to manipulate and carry out computations, ranging from bulks to molecules, driven by DFT or empirical potentials.	[28]
ASR	Collection of "recipes" for the computation of almost any property derived from electronic structure computations. Heavily depends on ASE.	[30]
MyQueue	Manager for submitting (interdependent) individual computations, useful for high-throughput screenings.	[32]
GASpy	Dynamic database (unknown entries will be computed upon request) and workflow manger, including predefined workflows for adsorption energy computations from bulks and molecules to surface and adsorption site enumeration.	[37]
DockOnSurf	Workflow for screening adsorption modes of flexible, polyfunctional molecules on user-defined adsorption sites.	[42]
Qmpy	Workflow and database manager for screening bulk materials and their properties, including formation energies, with VASP. Built for OQMD.	[47]
CatKit	Tool to generate initial structures of adsorbates on surfaces. It exploits graph-theory and symmetry to enumerate all unique adsorption sites and to adsorb mono or bi-dentate adsorbates along high-symmetry directions.	[39]
BOSS	Workflow for screening adsorption modes while reducing costly DFT computations by accelerating geometry optimizations via Bayesian optimization.	[86]
NOMAD-lab	Repository for results from most electronic structure codes. Raw-output files are stored, minimizing information loss.	[82]

AFIR	Algorithm to automatically determine transition states for arbitrary reactions.	[58]
Dscribe	Library implementing various atomic structure-based descriptors.	[113]
ChemOS	High-level operating system for coupling automated laboratory equipment with machine-learning.	[125]
MISO	Algorithm optimally combining information from hierarchical levels with increasing cost and decreasing uncertainty.	[87]
Atomsk	Toolkit to create grain-boundaries and bulk dislocations.	[120]
MPInterfaces	Toolkit to generate complex interfaces, heavily depends on the tools underlying Atomate.	[116]

Conclusions

In summary, we firmly believe that catalyst development can be accelerated by high-throughput, autonomous computations that identify promising (active) and realistic (under given reaction conditions) catalyst surfaces. Given the excessive complexity to predict the feasibility and (long-term) stability of a given catalyst from first principles atomistic computations, our vision for optimal theory-guided catalyst design consists of in silico screening of the chemical space to identify promising compositions and active site motives by high-throughput, autonomous computations. This screening of the chemical space is followed by experimental ML-enhanced optimization of the synthesis protocol and the reaction conditions to achieve active and stable catalysts within the computationally identified family. This experimental optimization can integrate any user-defined cost function, for example a trade-off between price, activity and stability. With time, the autonomous laboratories might become as available as supercomputing facilities, opening a new branch of catalysis research, requiring skills somewhere between experimental and computational sciences.

Will these autonomous workflows and laboratories replace trained, highly-skilled researchers? – This seems rather unlikely to us. Rather, we believe that automating the tedious parts of catalyst development should be seen as a liberating action, enabling chemists and chemical engineers to focus on developing promising hypotheses and efficient ways of testing them, rather than spending their time on the Edisonian trial and error approach. Similarly, computational chemists could focus on making models more relevant, coming up with good descriptors and generating insight, rather than having to spend time on repetitive and error-prone human-based, construct and collect, actions. Furthermore, recent advances in extracting knowledge from data sets in terms of human interpretable hypothesis are an exciting step towards the generation of general rules and insights[129] and thus making the intellectual journey through chemical space even more enjoyable.

Acknowledgements

Z.W.S. acknowledges the support of the Singapore National Research Foundation (NRF-NRFF2017-04). S.N.S. acknowledges the support by Région Auvergne Rhône-Alpes through

the project Pack Ambition Recherche 2018 MoSHi. M.B.J. gratefully acknowledges Stellantis and the ANRT for his PhD fellowship.

Author contributions

All authors contributed to writing and editing the manuscript.

Conflicts of interest

There are no conflicts to declare.

References

1. Hu, X., and Yip, A.C.K. (2021). Heterogeneous Catalysis: Enabling a Sustainable Future. Front. Catal. *1*, 3.

2. Reina, T.R., and Odriozola, J.A. (2020). Heterogeneous Catalysis for Energy Applications (The Royal Society of Chemistry).

3. Marakatti, V.S., and Gaigneaux, E.M. (2020). Recent Advances in Heterogeneous Catalysis for Ammonia Synthesis. ChemCatChem *12*, 5838–5857.

4. Huo, J., Tessonnier, J.-P., and Shanks, B.H. (2021). Improving Hydrothermal Stability of Supported Metal Catalysts for Biomass Conversions: A Review. ACS Catal., 5248–5270.

5. Simoes, M., Baranton, S., and Coutanceau, C. (2012). Electrochemical Valorisation of Glycerol. ChemSusChem *5*, 2106–2124.

6. Kas, R., Yang, K., Bohra, D., Kortlever, R., Burdyny, T., and Smith, W.A. (2020). Electrochemical CO2 reduction on nanostructured metal electrodes: fact or defect? Chem. Sci. *11*, 1738–1749.

7. Modestino, M.A., Hashemi, S.M.H., and Haussener, S. (2016). Mass transport aspects of electrochemical solar-hydrogen generation. Energy Environ. Sci. *9*, 1533–1551.

8. Norskov, J.K., Bligaard, T., Hvolbaek, B., Abild-Pedersen, F., Chorkendorff, I., and Christensen, C.H. (2008). The nature of the active site in heterogeneous metal catalysis. Chem Soc Rev *37*, 2163–2171.

9. van Santen, R.A. (2017). Modern heterogeneous catalysis: an introduction (Wiley-VCH Verlag).

10. Greeley, J., Jaramillo, T.F., Bonde, J., Chorkendorff, I., and Norskov, J.K. (2006). Computational high-throughput screening of electrocatalytic materials for hydrogen evolution. Nat Mater *5*, 909–913.

11. Norskov, J.K., Bligaard, T., Rossmeisl, J., and Christensen, C.H. (2009). Towards the computational design of solid catalysts. Nat Chem *1*, 37–46.

12. Huber, S.P., Zoupanos, S., Uhrin, M., Talirz, L., Kahle, L., Häuselmann, R., Gresch, D., Müller, T., Yakutovich, A.V., Andersen, C.W., et al. (2020). AiiDA 1.0, a scalable computational infrastructure for automated reproducible workflows and data provenance. Sci. Data *7*, 300.

13. Jain, A., Ong, S.P., Chen, W., Medasani, B., Qu, X., Kocher, M., Brafman, M., Petretto, G., Rignanese, G., Hautier, G., et al. (2015). FireWorks: a dynamic workflow system designed for high-throughput applications. Concurr. Comput. Pract. Exp. *27*, 5037–5059.

14. Uhrin, M., Huber, S.P., Yu, J., Marzari, N., and Pizzi, G. (2021). Workflows in AiiDA: Engineering a high-throughput, event-based engine for robust and modular computational workflows. Comput. Mater. Sci. *187*, 110086.

15. Ong, S.P., Richards, W.D., Jain, A., Hautier, G., Kocher, M., Cholia, S., Gunter, D., Chevrier, V.L., Persson, K.A., and Ceder, G. (2013). Python Materials Genomics (pymatgen): A robust, open-source python library for materials analysis. Comput. Mater. Sci. *68*, 314–319.

16. Talirz, L., Kumbhar, S., Passaro, E., Yakutovich, A.V., Granata, V., Gargiulo, F., Borelli, M., Uhrin, M., Huber, S.P., Zoupanos, S., et al. (2020). Materials Cloud, a platform for open computational science. Sci. Data *7*, 299.

17. Andriuc, O., Siron, M., Montoya, J.H., Horton, M., and Persson, K.A. (2021). Automated Adsorption Workflow for Semiconductor Surfaces and the Application to Zinc Telluride. J. Chem. Inf. Model. *61*, 3908–3916.

18. Curtarolo, S., Setyawan, W., Hart, G.L.W., Jahnatek, M., Chepulskii, R.V., Taylor, R.H., Wang, S., Xue, J., Yang, K., Levy, O., et al. (2012). AFLOW: An automatic framework for high-throughput materials discovery. Comput. Mater. Sci. *58*, 218–226.

19. Himanen, L., Geurts, A., Foster, A.S., and Rinke, P. (2019). Data-Driven Materials Science: Status, Challenges, and Perspectives. Adv. Sci. *6*, 1900808.

20. Calderon, C.E., Plata, J.J., Toher, C., Oses, C., Levy, O., Fornari, M., Natan, A., Mehl, M.J., Hart, G., Buongiorno Nardelli, M., et al. (2015). The AFLOW standard for high-throughput materials science calculations. Comput. Mater. Sci. *108*, 233–238.

21. Haastrup, S., Strange, M., Pandey, M., Deilmann, T., Schmidt, P.S., Hinsche, N.F., Gjerding, M.N., Torelli, D., Larsen, P.M., Riis-Jensen, A.C., et al. (2018). The Computational 2D Materials Database: high-throughput modeling and discovery of atomically thin crystals. 2D Mater. *5*, 042002.

22. Curtarolo, S., Setyawan, W., Wang, S., Xue, J., Yang, K., Taylor, R.H., Nelson, L.J., Hart, G.L.W., Sanvito, S., Buongiorno-Nardelli, M., et al. (2012). AFLOWLIB.ORG: A distributed materials properties repository from high-throughput ab initio calculations. Comput. Mater. Sci. *58*, 227–235.

23. Mounet, N., Gibertini, M., Schwaller, P., Campi, D., Merkys, A., Marrazzo, A., Sohier, T., Castelli, I.E., Cepellotti, A., Pizzi, G., et al. (2018). Two-dimensional materials from high-throughput computational exfoliation of experimentally known compounds. Nat. Nanotechnol. *13*, 246–252.

24. Jain, A., Ong, S.P., Hautier, G., Chen, W., Richards, W.D., Dacek, S., Cholia, S., Gunter, D., Skinner, D., Ceder, G., et al. (2013). Commentary: The Materials Project: A materials genome approach to accelerating materials innovation. APL Mater. *1*, 011002.

25. Chanussot, L., Das, A., Goyal, S., Lavril, T., Shuaibi, M., Riviere, M., Tran, K., Heras-Domingo, J., Ho, C., Hu, W., et al. (2021). Open Catalyst 2020 (OC20) Dataset and Community Challenges. ACS Catal. *11*, 6059–6072.

26. Rong, Z., Kitchaev, D., Canepa, P., Huang, W., and Ceder, G. (2016). An efficient algorithm for finding the minimum energy path for cation migration in ionic materials. J. Chem. Phys. *145*, 074112.

27. Fleurat-Lessard, P. Reaction Path. http://pfleurat.free.fr/ReactionPath.php.

28. Larsen, A.H., Mortensen, J.J., Blomqvist, J., Castelli, I.E., Christensen, R., Dulak, M., Friis, J., Groves, M.N., Hammer, B., Hargus, C., et al. (2017). The atomic simulation environment—a Python library for working with atoms. J. Phys. Condens. Matter *29*, 273002.

29. Henkelman, G., Uberuaga, B.P., and Jónsson, H. (2000). A climbing image nudged elastic band method for finding saddle points and minimum energy paths. J. Chem. Phys. *113*, 9901–9904.

30. Gjerding, M., Skovhus, T., Rasmussen, A., Bertoldo, F., Larsen, A.H., Mortensen, J.J., and Thygesen, K.S. (2021). Atomic Simulation Recipes: A Python framework and library for automated workflows. Comput. Mater. Sci. *199*, 110731.

31. Mathiesen, N.R., Jónsson, H., Vegge, T., and García Lastra, J.M. (2019). R-NEB: Accelerated Nudged Elastic Band Calculations by Use of Reflection Symmetry. J. Chem. Theory Comput. *15*, 3215–3222.

32. Mortensen, J., Gjerding, M., and Thygesen, K. (2020). MyQueue: Task and workflow scheduling system. J. Open Source Softw. *5*, 1844.

33. Koistinen, O.-P., Dagbjartsdóttir, F.B., Ásgeirsson, V., Vehtari, A., and Jónsson, H. (2017). Nudged elastic band calculations accelerated with Gaussian process regression. J. Chem. Phys. *147*, 152720.

34. Mathew, K., Montoya, J.H., Faghaninia, A., Dwarakanath, S., Aykol, M., Tang, H., Chu, I., Smidt, T., Bocklund, B., Horton, M., et al. (2017). Atomate: A high-level interface to generate, execute, and analyze computational materials science workflows. Comput. Mater. Sci. *139*, 140–152.

35. Garrido Torres, J.A., Jennings, P.C., Hansen, M.H., Boes, J.R., and Bligaard, T. (2019). Low-Scaling Algorithm for Nudged Elastic Band Calculations Using a Surrogate Machine Learning Model. Phys. Rev. Lett. *122*, 156001.

36. Meyer, R., Schmuck, K.S., and Hauser, A.W. (2019). Machine Learning in Computational Chemistry: An Evaluation of Method Performance for Nudged Elastic Band Calculations. J. Chem. Theory Comput. *15*, 6513–6523.

37. Tran, K., Palizhati, A., Back, S., and Ulissi, Z.W. (2018). Dynamic Workflows for Routine Materials Discovery in Surface Science. J. Chem. Inf. Model. *58*, 2392–2400.

38. Sugiyama, K., Sumiya, Y., Takagi, M., Saita, K., and Maeda, S. (2019). Understanding CO oxidation on the Pt(111) surface based on a reaction route network. Phys. Chem. Chem. Phys. *21*, 14366–14375.

39. Boes, J.R., Mamun, O., Winther, K., and Bligaard, T. (2019). Graph Theory Approach to High-Throughput Surface Adsorption Structure Generation. J. Phys. Chem. A *123*, 2281–2285.

40. Chen, C., Zuo, Y., Ye, W., Li, X., Deng, Z., and Ong, S.P. (2020). A Critical Review of Machine Learning of Energy Materials. Adv. Energy Mater. *10*, 1903242.

41. Bligaard, T., Norskov, J.K., Dahl, S., Matthiesen, J., Christensen, C.H., and Sehested, J. (2004). The Bronsted-Evans-Polanyi relation and the volcano curve in heterogeneous catalysis. J Catal *224*, 206–217.

42. Martí, C., Blanck, S., Staub, R., Loehlé, S., Michel, C., and Steinmann, S.N. (2021). DockOnSurf: A Python Code for the High-Throughput Screening of Flexible Molecules Adsorbed on Surfaces. J. Chem. Inf. Model. *61*, 3386–3396.

43. Abdelfatah, K., Yang, W., Vijay Solomon, R., Rajbanshi, B., Chowdhury, A., Zare, M., Kundu, S.K., Yonge, A., Heyden, A., and Terejanu, G. (2019). Prediction of Transition-State Energies of Hydrodeoxygenation Reactions on Transition-Metal Surfaces Based on Machine Learning. J. Phys. Chem. C *123*, 29804–29810.

44. Moosavi, S.M., Jablonka, K.M., and Smit, B. (2020). The Role of Machine Learning in the Understanding and Design of Materials. J. Am. Chem. Soc. *142*, 20273–20287.

45. Pollice, R., dos Passos Gomes, G., Aldeghi, M., Hickman, R.J., Krenn, M., Lavigne, C., Lindner-D'Addario, M., Nigam, A., Ser, C.T., Yao, Z., et al. (2021). Data-Driven Strategies for Accelerated Materials Design. Acc. Chem. Res. *54*, 849–860.

46. Bannigan, P., Aldeghi, M., Bao, Z., Häse, F., Aspuru-Guzik, A., and Allen, C. (2021). Machine learning directed drug formulation development. Adv. Drug Deliv. Rev. *175*, 113806.

47. Kirklin, S., Saal, J.E., Meredig, B., Thompson, A., Doak, J.W., Aykol, M., Rühl, S., and Wolverton, C. (2015). The Open Quantum Materials Database (OQMD): assessing the accuracy of DFT formation energies. Npj Comput. Mater. *1*, 15010.

48. Schlexer Lamoureux, P., Winther, K.T., Garrido Torres, J.A., Streibel, V., Zhao, M., Bajdich, M., Abild-Pedersen, F., and Bligaard, T. (2019). Machine Learning for Computational Heterogeneous Catalysis. ChemCatChem *11*, 3581–3601.

49. Toyao, T., Maeno, Z., Takakusagi, S., Kamachi, T., Takigawa, I., and Shimizu, K. (2020). Machine Learning for Catalysis Informatics: Recent Applications and Prospects. ACS Catal. *10*, 2260–2297.

50. Montoya, J.H., Winther, K.T., Flores, R.A., Bligaard, T., Hummelshøj, J.S., and Aykol, M. (2020). Autonomous intelligent agents for accelerated materials discovery. Chem. Sci. *11*, 8517–8532.

51. Back, S., Na, J., Tran, K., and Ulissi, Z.W. (2020). In silico discovery of active, stable, CO-tolerant and cost-effective electrocatalysts for hydrogen evolution and oxidation. Phys. Chem. Chem. Phys. *22*, 19454–19458.

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

53. Zhong, M., Tran, K., Min, Y., Wang, C., Wang, Z., Dinh, C.-T., De Luna, P., Yu, Z., Rasouli, A.S., Brodersen, P., et al. (2020). Accelerated discovery of CO2 electrocatalysts using active machine learning. Nature *581*, 178–183.

54. Zhou, Z., Kearnes, S., Li, L., Zare, R.N., and Riley, P. (2019). Optimization of Molecules via Deep Reinforcement Learning. Sci. Rep. *9*, 1–10.

55. Krenn, M., Häse, F., Nigam, A., Friederich, P., and Aspuru-Guzik, A. (2020). Self-referencing embedded strings (SELFIES): A 100% robust molecular string representation. Mach. Learn. Sci. Technol. *1*, 045024.

56. Yao, Z., Sánchez-Lengeling, B., Bobbitt, N.S., Bucior, B.J., Kumar, S.G.H., Collins, S.P., Burns, T., Woo, T.K., Farha, O.K., Snurr, R.Q., et al. (2021). Inverse design of nanoporous crystalline reticular materials with deep generative models. Nat. Mach. Intell. *3*, 76–86.

57. Silver, D., Huang, A., Maddison, C.J., Guez, A., Sifre, L., van den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., et al. (2016). Mastering the game of Go with deep neural networks and tree search. Nature *529*, 484–489.

58. Maeda, S., and Morokuma, K. (2011). Finding Reaction Pathways of Type A + B → X: Toward Systematic Prediction of Reaction Mechanisms. J. Chem. Theory Comput. *7*, 2335–2345.

59. Ahuja, K., Green, W.H., and Li, Y.-P. (2021). Learning to Optimize Molecular Geometries Using Reinforcement Learning. J. Chem. Theory Comput. *17*, 818–825.

60. Lan, T., and An, Q. (2021). Discovering Catalytic Reaction Networks Using Deep Reinforcement Learning from First-Principles. J. Am. Chem. Soc. *143*, 16804–16812.

61. Zhang, L., Lin, D.-Y., Wang, H., Car, R., and E, W. (2019). Active learning of uniformly accurate interatomic potentials for materials simulation. Phys. Rev. Mater. *3*, 023804.

62. Staub, R., and Steinmann, S. (2021). Replacing Chemical Intuition by Machine Learning: a Mixed Design of Experiments - Reinforcement Learning Approach to the Construction of Training Sets for Model Hamiltonians. 10.26434/chemrxiv-2021–6v4n0.

63. Goings, J.J., Hu, H., Yang, C., and Li, X. (2021). Reinforcement Learning Configuration Interaction. J. Chem. Theory Comput. *17*, 5482–5491.

64. Tomasi, J., and Persico, M. (1994). Molecular Interactions in Solution: An Overview of Methods Based on Continuous Distributions of the Solvent. Chem. Rev. *94*, 2027–2094.

65. Mathew, K., Sundararaman, R., Letchworth-Weaver, K., Arias, T.A., and Hennig, R.G. (2014). Implicit solvation model for density-functional study of nanocrystal surfaces and reaction pathways. J Chem Phys *140*, 084106.

66. Abidi, N., Lim, K.R.G., Seh, Z.W., and Steinmann, S.N. (2021). Atomistic modeling of electrocatalysis: Are we there yet? WIREs Comput. Mol. Sci. *11*, e1499.

67. Zhang, Q., and Asthagiri, A. (2019). Solvation effects on DFT predictions of ORR activity on metal surfaces. Catal. Today *323*, 35–43.

68. Heenen, H.H., Gauthier, J.A., Kristoffersen, H.H., Ludwig, T., and Chan, K. (2020). Solvation at metal/water interfaces: An ab initio molecular dynamics benchmark of common computational approaches. J. Chem. Phys. *152*, 144703.

69. Rendón-Calle, A., Builes, S., and Calle-Vallejo, F. (2020). Substantial improvement of electrocatalytic predictions by systematic assessment of solvent effects on adsorption energies. Appl. Catal. B Environ. *276*, 119147.

70. Clabaut, P., Schweitzer, B., Götz, A.W., Michel, C., and Steinmann, S.N. (2020). Solvation Free Energies and Adsorption Energies at the Metal/Water Interface from Hybrid Quantum-Mechanical/Molecular Mechanics Simulations. J. Chem. Theory Comput. *16*, 6539–6549.

71. Saleheen, M., and Heyden, A. (2018). Liquid-Phase Modeling in Heterogeneous Catalysis. ACS Catal. *8*, 2188–2194.

72. Zhang, X., DeFever, R.S., Sarupria, S., and Getman, R.B. (2019). Free Energies of Catalytic Species Adsorbed to Pt(111) Surfaces under Liquid Solvent Calculated Using Classical and Quantum Approaches. J. Chem. Inf. Model. *59*, 2190–2198.

73. Weitzner, S.E., Akhade, S.A., Varley, J.B., Wood, B.C., Otani, M., Baker, S.E., and Duoss, E.B. (2020). Toward Engineering of Solution Microenvironments for the CO2 Reduction

Reaction: Unraveling pH and Voltage Effects from a Combined Density-Functional–Continuum Theory. J. Phys. Chem. Lett. *11*, 4113–4118.

74. Jeanmairet, G., Levesque, M., and Borgis, D. (2020). Tackling Solvent Effects by Coupling Electronic and Molecular Density Functional Theory. J. Chem. Theory Comput.

75. Clabaut, P., Fleurat-Lessard, P., Michel, C., and Steinmann, S.N. (2020). Ten Facets, One Force Field: The GAL19 Force Field for Water–Noble Metal Interfaces. J. Chem. Theory Comput. *16*, 4565–4578.

76. Wang, S., Witek, J., Landrum, G.A., and Riniker, S. (2020). Improving Conformer Generation for Small Rings and Macrocycles Based on Distance Geometry and Experimental Torsional-Angle Preferences. J. Chem. Inf. Model. *60*, 2044–2058.

77. Naserifar, S., Chen, Y., Kwon, S., Xiao, H., and Goddard, W.A. (2021). Artificial Intelligence and QM/MM with a Polarizable Reactive Force Field for Next-Generation Electrocatalysts. Matter *4*, 195–216.

78. Reuter, K., and Scheffler, M. (2001). Composition, structure, and stability of ${\mathrm{RuO}}_{2}(110)$ as a function of oxygen pressure. Phys. Rev. B *65*, 035406.

79. Yamada, H., Liu, C., Wu, S., Koyama, Y., Ju, S., Shiomi, J., Morikawa, J., and Yoshida, R. (2019). Predicting Materials Properties with Little Data Using Shotgun Transfer Learning. ACS Cent. Sci. *5*, 1717–1730.

80. wwPDB consortium, Burley, S.K., Berman, H.M., Bhikadiya, C., Bi, C., Chen, L., Costanzo, L.D., Christie, C., Duarte, J.M., Dutta, S., et al. (2019). Protein Data Bank: the single global archive for 3D macromolecular structure data. Nucleic Acids Res. *47*, D520–D528.

81. Kearnes, S.M., Maser, M.R., Wleklinski, M., Kast, A., Doyle, A.G., Dreher, S.D., Hawkins, J.M., Jensen, K.F., and Coley, C.W. (2021). The Open Reaction Database. J. Am. Chem. Soc.

82. Draxl, C., and Scheffler, M. (2019). The NOMAD laboratory: from data sharing to artificial intelligence. J. Phys. Mater. *2*, 036001.

83. Gu, G.H., Noh, J., Kim, S., Back, S., Ulissi, Z., and Jung, Y. (2020). Practical Deep-Learning Representation for Fast Heterogeneous Catalyst Screening. J. Phys. Chem. Lett. *11*, 3185–3191.

84. Behler, J., and Parrinello, M. (2007). Generalized Neural-Network Representation of High-Dimensional Potential-Energy Surfaces. Phys. Rev. Lett. *98*, 146401.

85. Bartók, A.P., Kondor, R., and Csányi, G. (2013). On representing chemical environments. Phys. Rev. B *87*, 184115.

86. Todorović, M., Gutmann, M.U., Corander, J., and Rinke, P. (2019). Bayesian inference of atomistic structure in functional materials. Npj Comput. Mater. *5*, 1–7.

87. Herbol, H.C., Poloczek, M., and Clancy, P. (2020). Cost-effective materials discovery: Bayesian optimization across multiple information sources. Mater. Horiz. *7*, 2113–2123.

88. Egger, A.T., Hörmann, L., Jeindl, A., Scherbela, M., Obersteiner, V., Todorović, M., Rinke, P., and Hofmann, O.T. (2020). Charge Transfer into Organic Thin Films: A Deeper Insight through Machine-Learning-Assisted Structure Search. Adv. Sci. *7*, 2000992.

89. Yoon, J., Cao, Z., Raju, R.K., Wang, Y., Burnley, R., Gellman, A.J., Farimani, A.B., and Ulissi, Z.W. (2021). Deep reinforcement learning for predicting kinetic pathways to surface reconstruction in a ternary alloy. Mach. Learn. Sci. Technol. *2*, 045018.

90. Sun, G., Fuller, J.T., Alexandrova, A.N., and Sautet, P. (2021). Global Activity Search Uncovers Reaction Induced Concomitant Catalyst Restructuring for Alkane Dissociation on Model Pt Catalysts. ACS Catal. *11*, 1877–1885.

91. Gu, G.H., Lim, J., Wan, C., Cheng, T., Pu, H., Kim, S., Noh, J., Choi, C., Kim, J., Goddard, W.A., et al. (2021). Autobifunctional Mechanism of Jagged Pt Nanowires for Hydrogen Evolution Kinetics via End-to-End Simulation. J. Am. Chem. Soc. *143*, 5355–5363.

92. Batchelor, T.A.A., Pedersen, J.K., Winther, S.H., Castelli, I.E., Jacobsen, K.W., and Rossmeisl, J. (2019). High-Entropy Alloys as a Discovery Platform for Electrocatalysis. Joule *3*, 834–845.

93. Sahu, A., Steinmann, S.N., and Raybaud, P. (2020). Size-Dependent Structural, Energetic, and Spectroscopic Properties of MoS3 Polymorphs. Cryst. Growth Des. *20*, 7750–7760.

94. Tao, H., Wu, T., Aldeghi, M., Wu, T.C., Aspuru-Guzik, A., and Kumacheva, E. (2021). Nanoparticle synthesis assisted by machine learning. Nat. Rev. Mater. *6*, 701–716.

95. Yao, L., Ou, Z., Luo, B., Xu, C., and Chen, Q. (2020). Machine Learning to Reveal Nanoparticle Dynamics from Liquid-Phase TEM Videos. ACS Cent. Sci. *6*, 1421–1430.

96. Xu, M., Tang, B., Lu, Y., Zhu, C., Lu, Q., Zhu, C., Zheng, L., Zhang, J., Han, N., Fang, W., et al. (2021). Machine Learning Driven Synthesis of Few-Layered WTe_2 with Geometrical Control. J. Am. Chem. Soc. *143*, 18103–18113.

97. Guo, Y., He, X., Su, Y., Dai, Y., Xie, M., Yang, S., Chen, J., Wang, K., Zhou, D., and Wang, C. (2021). Machine-Learning-Guided Discovery and Optimization of Additives in Preparing Cu Catalysts for CO2 Reduction. J. Am. Chem. Soc. *143*, 5755–5762.

98. McCullough, K., Williams, T., Mingle, K., Jamshidi, P., and Lauterbach, J. (2020). High-throughput experimentation meets artificial intelligence: a new pathway to catalyst discovery. Phys. Chem. Chem. Phys. *22*, 11174–11196.

99. Ludwig, A. (2019). Discovery of new materials using combinatorial synthesis and high-throughput characterization of thin-film materials libraries combined with computational methods. Npj Comput. Mater. *5*, 1–7.

100. Turner, H.W., Volpe, A.F., and Weinberg, W.H. (2009). High-throughput heterogeneous catalyst research. Surf. Sci. *603*, 1763–1769.

101. Marenich, A.V., Cramer, C.J., and Truhlar, D.G. (2009). Universal Solvation Model Based on Solute Electron Density and on a Continuum Model of the Solvent Defined by the Bulk Dielectric Constant and Atomic Surface Tensions. J. Phys. Chem. B *113*, 6378–6396.

102. Nursam, N.M., Wang, X., and Caruso, R.A. (2015). High-Throughput Synthesis and Screening of Titania-Based Photocatalysts. ACS Comb. Sci. *17*, 548–569.

103. Sundararaman, R., and Goddard, W.A. (2015). The charge-asymmetric nonlocally determined local-electric (CANDLE) solvation model. J. Chem. Phys. *142*, 064107.

104. Potgieter, K., Aimon, A., Smit, E., von Delft, F., and Meijboom, R. (2021). Robotic Catalysis: A High-Throughput Method for Miniature Screening of Mesoporous Metal Oxides. Chemistry–Methods *1*, 192–200.

105. Stroyuk, O., Raievska, O., Langner, S., Kupfer, C., Barabash, A., Solonenko, D., Azhniuk, Y., Hauch, J., Osvet, A., Batentschuk, M., et al. (2021). High-Throughput Robotic Synthesis and Photoluminescence Characterization of Aqueous Multinary Copper–Silver Indium Chalcogenide Quantum Dots. Part. Part. Syst. Charact. *38*, 2100169.

106. Tao, F. (Feng), and Crozier, P.A. (2016). Atomic-Scale Observations of Catalyst Structures under Reaction Conditions and during Catalysis. Chem. Rev. *116*, 3487–3539.

107. Krska, S.W., DiRocco, D.A., Dreher, S.D., and Shevlin, M. (2017). The Evolution of Chemical High-Throughput Experimentation To Address Challenging Problems in Pharmaceutical Synthesis. Acc. Chem. Res. *50*, 2976–2985.

108. Shi, Y., Yang, B., Rack, P.D., Guo, S., Liaw, P.K., and Zhao, Y. (2020). High-throughput synthesis and corrosion behavior of sputter-deposited nanocrystalline Alx(CoCrFeNi)100-x combinatorial high-entropy alloys. Mater. Des. *195*, 109018.

109. Mao, S.S., and Zhang, X. (2015). High-Throughput Multi-Plume Pulsed-Laser Deposition for Materials Exploration and Optimization. Engineering *1*, 367–371.

110. Cong, P., Doolen, R.D., Fan, Q., Giaquinta, D.M., Guan, S., McFarland, E.W., Poojary, D.M., Self, K., Turner, H.W., and Weinberg, W.H. (1999). High-Throughput Synthesis and Screening of Combinatorial Heterogeneous Catalyst Libraries. Angew. Chem. Int. Ed. *38*, 483–488.

111. Scoutaris, N., Nion, A., Hurt, A., and Douroumis, D. (2016). Jet dispensing as a high throughput method for rapid screening and manufacturing of cocrystals. CrystEngComm *18*, 5079–5082.

112. Ortega, C., Otyuskaya, D., Ras, E.-J., Virla, L.D., Patience, G.S., and Dathe, H. (2021). Experimental methods in chemical engineering: High throughput catalyst testing — HTCT. Can. J. Chem. Eng. *99*, 1288–1306.

113. Himanen, L., Jäger, M.O.J., Morooka, E.V., Federici Canova, F., Ranawat, Y.S., Gao, D.Z., Rinke, P., and Foster, A.S. (2020). DScribe: Library of descriptors for machine learning in materials science. Comput. Phys. Commun. *247*, 106949.

114. Bruker, D8 ADVANCE. https://www.bruker.com/en/products-and-solutions/diffractometers-and-scattering-systems/x-ray-diffractometers/d8-advance-family/d8-advance.html.

115. Nurizzo, D., Bowler, M.W., Caserotto, H., Dobias, F., Giraud, T., Surr, J., Guichard, N., Papp, G., Guijarro, M., Mueller-Dieckmann, C., et al. (2016). RoboDiff: combining a sample changer and goniometer for highly automated macromolecular crystallography experiments. Acta Crystallogr. Sect. Struct. Biol. *72*, 966–975.

116. Mathew, K., Singh, A.K., Gabriel, J.J., Choudhary, K., Sinnott, S.B., Davydov, A.V., Tavazza, F., and Hennig, R.G. (2016). MPInterfaces: A Materials Project based Python tool for high-throughput computational screening of interfacial systems. Comput. Mater. Sci. *122*, 183–190.

117. Round, A., Felisaz, F., Fodinger, L., Gobbo, A., Huet, J., Villard, C., Blanchet, C.E., Pernot, P., McSweeney, S., Roessle, M., et al. (2015). BioSAXS Sample Changer: a robotic sample changer for rapid and reliable high-throughput X-ray solution scattering experiments. Acta Crystallogr. D Biol. Crystallogr. *71*, 67–75.

118. Goldrick, S., Umprecht, A., Tang, A., Zakrzewski, R., Cheeks, M., Turner, R., Charles, A., Les, K., Hulley, M., Spencer, C., et al. (2020). High-Throughput Raman Spectroscopy Combined with Innovate Data Analysis Workflow to Enhance Biopharmaceutical Process Development. Processes *8*, 1179.

119. Mondol, A.S., Patel, M.D., Rüger, J., Stiebing, C., Kleiber, A., Henkel, T., Popp, J., and Schie, I.W. (2019). Application of High-Throughput Screening Raman Spectroscopy (HTS-RS) for Label-Free Identification and Molecular Characterization of Pollen. Sensors *19*, 4428.

120. Hirel, P. (2015). Atomsk: A tool for manipulating and converting atomic data files. Comput. Phys. Commun. *197*, 212–219.

121. Coffey, P., Smith, N., Lennox, B., Kijne, G., Bowen, B., Davis-Johnston, A., and Martin, P.A. (2021). Robotic arm material characterisation using LIBS and Raman in a nuclear hot cell decommissioning environment. J. Hazard. Mater. *412*, 125193.

122. Westley, C., Xu, Y., Carnell, A.J., Turner, N.J., and Goodacre, R. (2016). Label-Free Surface Enhanced Raman Scattering Approach for High-Throughput Screening of Biocatalysts. Anal. Chem. *88*, 5898–5903.

123. Nguyen, T.N., Nhat, T.T.P., Takimoto, K., Thakur, A., Nishimura, S., Ohyama, J., Miyazato, I., Takahashi, L., Fujima, J., Takahashi, K., et al. (2020). High-Throughput Experimentation and Catalyst Informatics for Oxidative Coupling of Methane. ACS Catal. *10*, 921–932.

124. Roberts, E.J., Habas, S.E., Wang, L., Ruddy, D.A., White, E.A., Baddour, F.G., Griffin, M.B., Schaidle, J.A., Malmstadt, N., and Brutchey, R.L. (2017). High-Throughput Continuous Flow Synthesis of Nickel Nanoparticles for the Catalytic Hydrodeoxygenation of Guaiacol. ACS Sustain. Chem. Eng. *5*, 632–639.

125. Roch, L.M., Häse, F., Kreisbeck, C., Tamayo-Mendoza, T., Yunker, L.P.E., Hein, J.E., and Aspuru-Guzik, A. (2020). ChemOS: An orchestration software to democratize autonomous discovery. PLOS ONE *15*, e0229862.

126. Kondratyuk, P., Gumuslu, G., Shukla, S., Miller, J.B., Morreale, B.D., and Gellman, A.J. (2013). A microreactor array for spatially resolved measurement of catalytic activity for high-throughput catalysis science. J. Catal. *300*, 55–62.

127. Sundermann, A., and Gerlach, O. (2016). High-Throughput Screening as a Supplemental Tool for the Development of Advanced Emission Control Catalysts: Methodological Approaches and Data Processing. Catalysts *6*, 23.

128. Moses, O.A., Chen, W., Adam, M.L., Wang, Z., Liu, K., Shao, J., Li, Z., Li, W., Wang, C., Zhao, H., et al. (2021). Integration of data-intensive, machine learning and robotic experimental approaches for accelerated discovery of catalysts in renewable energy-related reactions. Mater. Rep. Energy *1*, 100049.

129. Friederich, P., Krenn, M., Tamblyn, I., and Aspuru-Guzik, A. (2021). Scientific intuition inspired by machine learning-generated hypotheses. Mach. Learn. Sci. Technol. *2*, 025027.