GDCh Viewpoint Article Angewandte Chemie International Edition www.angewandte.org

How to cite: *Angew. Chem. Int. Ed.* **2023**, *62*, e202302971
doi.org/10.1002/anie.202302971

FAIR Data

Achieving Digital Catalysis: Strategies for Data Acquisition, Storage and Use

*Clara Patricia Marshall, Julia Schumann, and Annette Trunschke**

Abstract: Heterogeneous catalysis is an important area of research that generates data as intricate as the phenomenon itself. Complexity is inherently coupled to the function of the catalyst and advance in knowledge can only be achieved if this complexity is adequately captured and accounted for. This requires integration of experiment and theory, high data quality and quality control, close interdisciplinary collaboration, and sharing of data and metadata, which is facilitated by the application of joint data management strategies. This Viewpoint Article first discusses the potential of a digital transition in catalysis research. Then, a summary of the current status in terms of data infrastructure in heterogeneous catalysis is presented, defining the various types of (meta-) data, from catalyst synthesis to functional analysis. Finally, an already implemented working concept for local data acquisition and storage is introduced and the benefits and further development directions for catalysis data use and sharing are discussed.

1. Introduction

In the paradigm of the information revolution, data is in focus.[1] Reliable experimental and calculated data are extremely valuable resources for research[2] and key to answering important scientific questions related to both the sustainable manufacture of chemicals and materials and the transformation of the global energy system, which are tasks that cannot be solved without catalysis.[3] Disruptive innovations[4] in catalyst technology require a deeper fundamental understanding of the physical principles underlying catalytic processes. This can be achieved by a closer integration of experiment and theory, which is facilitated by the progressively widespread use of artificial intelligence methods (Figure 1).[5] For this purpose, the data must be reliable and accessible, which places high demands on their acquisition and storage.

As with functional materials in general, catalyst data are particularly complex. The complexity is due in part to the wide range of length and time scales involved in the multitude of different coupled processes, such as diffusion, adsorption, surface reaction, or heat transport, that affect the relationships between structure and function of a catalyst. In addition, the chemical feedback between catalyst and reaction medium (i.e., the mutual influence of temporally fluctuating composition of the reacting phase and the interface) must be taken into account, which emphasizes the importance of metadata. Finally, catalysis research is an interdisciplinary field that requires the incorporation of a wide variety of methods from inorganic, organic, analytical, physical, and computational chemistry, engineering and chemical physics, all of which generate data in a great variety of structures (Figure 1).

The aim of this article is to analyze the current state of the data infrastructure in catalysis research and to propose and discuss solutions and future directions for data management, focusing on the experimental practice in thermal heterogeneous catalysis. For this, our own approaches and experiences are also introduced. We anticipate that the lessons learned from this analysis will be applicable to other areas of catalysis and, even more broadly, to other disciplines in chemistry and physics, and benefits can be derived from mutual inspiration.[6] In redesigning the existing data infrastructure in catalysis research, it seems useful to consider the FAIR data principles, which are first briefly explained.

2. The FAIR concept

2.1. Definition of FAIR data

The “FAIR Data Principles” are guidelines proposed on the initiative of a worldwide consortium of stakeholders from academia, industry, funding agencies, and scientific publishers.[7] The idea was to improve the ability of machines to automatically search and exploit data and to facilitate the free exchange of data between users. The abbreviation FAIR stands for the terms **F**indable, **A**ccessible, **I**nteroperable, and **R**eusable. ***Findable*** means that data and metadata are given a globally unique and persistent identifier so that they are registered or indexed

[*] Dr. C. P. Marshall, Dr. J. Schumann, Dr. A. Trunschke
Department of Inorganic Chemistry, Fritz-Haber-Institut der Max-Planck-Gesellschaft
Faradayweg 4–6, 14195 Berlin (Germany)
E-mail: trunschke@fhi-berlin.mpg.de

Dr. J. Schumann
Consortium FAIRmat, c/o Humboldt-Universität zu Berlin
Zum Großen Windkanal 2, 12489 Berlin (Germany)

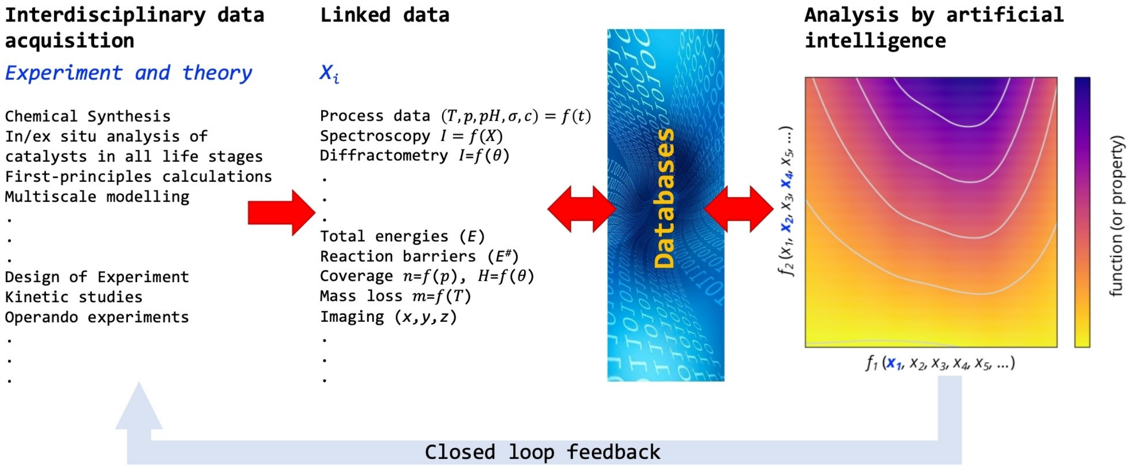

Interdisciplinary data acquisition
Experiment and theory
Chemical Synthesis
In/ex situ analysis of catalysts in all life stages
First-principles calculations
Multiscale modelling
.
.
.
Design of Experiment
Kinetic studies
Operando experiments
.
.
.
Linked data
X_i
Process data $(T,p,pH,\sigma,c)=f(t)$
Spectroscopy $I=f(X)$
Diffractometry $I=f(\theta)$
.
.
.
Total energies (E)
Reaction barriers $(E^{\#})$
Coverage $n=f(p)$, $H=f(\theta)$
Mass loss $m=f(T)$
Imaging (x,y,z)
.
.
.
Databases
Analysis by artificial intelligence
$f_2\,(x_1, x_2, x_3, x_4, x_5, \ldots)$
$f_1\,(x_1, x_2, x_3, x_4, x_5, \ldots)$
function (or property)
Closed loop feedback

Figure 1. Data acquisition in catalysis research provides the basis for understanding fundamental relationships between physical material properties and the performance of a catalytically active material to discover new catalysts based on this knowledge. Data analysis in catalysis requires coping with the complex entanglement of different data types resulting from the integration of experiment and theory, as well as the interdisciplinary nature of experimental catalysis research. Artificial intelligence and accelerated development (e.g., by design of experiment (DoE) methods) are playing an increasingly important role, which imposes high demands on the quality of the data and the documentation of metadata.

in searchable resources. ***Accessible*** implies that (meta-) data can be retrieved with standardized, open, and universally implementable communication protocols, if necessary, after authentication procedures. The term ***Interoperable*** denotes that (meta-) data use a formal, widely applicable language for knowledge representation. Finally, (meta−) data should be ***Reusable***, meaning that it must be released with a clear and accessible data usage license and adequately described with accurate and relevant attributes.[6,8] The FAIR principles thus essentially relate to the accessibility and reuse of data in digital repositories.

2.2. Development of legal regulations

To take advantage of the new opportunities arising from the synergies between data value chains and data transparency, reliable regulations must be created. It has been proposed that mandatory "research data management plans" (RDMPs) should be established for publicly funded research projects to align the interests of data producers in terms of control over and responsibility for their data with FAIR principles.[8b]

Various licenses have been developed in the past for the free use of software, known under the term "General Public License" (GPL). The users of such free software are allowed to utilize, analyze, distribute and also further develop the software, which has greatly advanced community-driven developments in catalysis informatics[5i] and computational chemistry. A network-based version management forum for software development projects is GitHub.[9]

"Creative Commons" are used in publishing data. To allow datasets to be cited and referenced, automatic generation of a Digital Object Identifier (DOI) should be associated with publication. In a repository for research data from chemistry ("Chemotion") developed as an alternative to commercial databases, for example, the data are made available and licensed as Open Data, including generation of DOIs with publication, automatic plausibility checks and verification of formal aspects of the uploaded content, and workflows for peer review of the submitted data.[8c]

2.3. The benefits of FAIR data in catalysis

Not only in the catalysis community, but more generally in chemistry,[6,10] in the field of energy transformation and storage, such as for the synthesis[11] and application[12] of metal–organic frameworks (MOFs) and battery research,[13] in bioscience,[14] or material science,[15] archiving and making data available according to FAIR standards bring the advantage that knowledge once acquired is made readily available to the entire research field and beyond, and is preserved over the long term. In this way, research resources can be saved by avoiding duplication of effort. On the other hand, reproduction experiments are absolutely necessary to build future studies reliably on the state of knowledge. However, because of missing metadata, reproduction of literature results in catalysis research is difficult at present, but becomes possible through generation of interoperable (meta-) data.

Another major advantage is that FAIR data enables cross- disciplinary linking and combination of data from different areas. In this way, data that have been studied in a specific context can be evaluated in the light of a completely new research question. For example, materials studied in semiconductor science[18] may be equally interesting for catalysis.[19]

2.4. About the need to change the paradigm

The original idea of facilitated machine data-driven approaches will advance catalysis research immensely, so it becomes critical to integrate the FAIR concept into the transformation of the catalysis data infrastructure. However, easy access to data in repositories is an important component,[8a,d] but not the only requirement, to be considered when building a future-oriented catalysis data infrastructure. Equally important, a paradigm shift is needed to drive research forward. This concerns in particular the integration of experiment and theory, which does not yet sufficiently account for the complex and metastable nature of functional interfaces. In other words, this means that in both fundamental studies of catalyst function by experiment and theory and technical catalyst development, too little attention is paid to the formation of the active phase in interaction with the reacting molecules.

The atomic structure of the catalyst under catalytic conditions is generally not known and therefore the models on which theory is based are often highly simplified,[20] and thus, not necessarily representative. Methodological developments in theoretical chemistry aim to account for the complexity and dynamics of catalysts by combining machine learning (ML) techniques that narrow down the number of possible active surface structures of interest with precise DFT calculations of these structures so identified,[21] and the application of first-principles-based multiscale models.[22] Experiments provide important reference data in this regard. The experimental information that enters the theory must be reliable and therefore based on rigorously conducted experiments on realistic (i.e., catalytically active) systems.

Complex chemical reactions important for the transformation of the current materials, and energy economy that embraces a closed carbon cycle and a hydrogen-based energy sector (e.g., electrocatalytic water splitting or thermocatalytic hydrogenation of carbon oxides), require complex catalyst systems. A paragon example is the interface between the catalytically active metal Cu and the matrix ZnO, which is crucial for the performance of the large-scale Cu-ZnO/Al_2O_3 catalyst applied for methanol synthesis from carbon oxides.[23] During pretreatment and operation, local coverage of the Cu surface with partially reduced ZnO_x modifications occurs, which depends on the energy input and the chemical potential of the gas phase (Figure 2).[17,24] At reduction temperatures higher than 250 °C, bulk Cu–Zn alloys are formed, which lead to a decrease of the methanol formation rate. Alloy formation is favored by higher hydrogen concentrations in the reducing gas.[16] But if the highly sensitive formation of CuZn alloys is confined to the surface, the methanol synthesis may be significantly promoted.[25] Therefore, the formation procedure, especially the reduction conditions, the operation temperature and the feed composition, have a delicate influence on the performance of Cu–ZnO-based methanol synthesis catalysts.[16,25,26]

Generally, high activity and selectivity of catalysts in complex reactions require also chemical or structural complexity of the catalyst material, and the catalyst system usually turns into a difficult-to-characterize active phase under working conditions that is sensitive to the process parameters. It is not always clear a priori which metadata have to be recorded, and a complete acquisition of all data is technically difficult and requires development time to automate experiments. Therefore, in order to master the complexity and to be able to obtain reproducible data, new experimental working methods must be developed in which the determination of important functional properties follows standardized operating procedures (SOPs).[27] This includes the definition of inter-laboratory benchmarks,[27a,28] essential for quality control, and the documentation of the workflow of catalyst tests.[27g,28c] Automation of experiments supports this approach and will relieve the scientist from routine tasks.[28c]

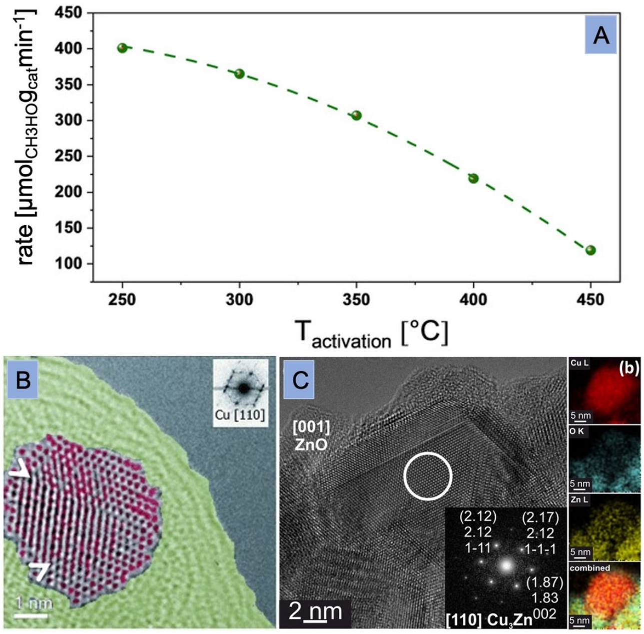


Figure 2. A) Influence of the reduction temperature ($T_{activation}$) on the formation rate of methanol at 30 bar and 230 °C in synthesis gas (8 CO_2/6 CO/59 H_2/27 He).[16] B) TEM images of a Cu–ZnO/Al_2O_3 catalyst after reduction at 250 °C exhibiting metallic Cu embedded in ZnO_x.[17] (C) TEM images of a Cu–ZnO/Al_2O_3 catalyst after reduction at 350 °C showing Cu–Zn alloy formation.[16] In (C) elemental EELS mapping of Cu L-, O K-, and Zn L-edges after reductive activation at 400 °C is also presented. Reproduced with permission from John Wiley and Sons.

Thus, a paradigm shift in catalysis means:

- Generate and use FAIR data in theory and experiment.
- Consider the formation of the catalyst under reaction conditions in theory and experiment and link all metadata including the workflow (detailed procedure) of the experiment to the catalytic data.
- Work according to SOPs and validate catalyst data through interlaboratory testing using benchmark catalysts.

3. The diverse character of catalysis data

The development of a data infrastructure in catalysis research is a particular challenge due to the interdisciplinary nature of the research field (Figure 1). In addition, the chemistry is extremely diverse in terms of both the types of reactions and the materials chemistry of the catalyst: 1) the catalyst and the reaction medium may be in liquid, solid, or a gaseous state; 2) the reaction can occur in homogeneous phase or at various interfaces of any combination of solid, liquid, or gaseous matter; 3) the catalyst may be a molecular species or an extended solid; 4) the reaction temperatures range from below 0 °C to above 1000 °C; 5) likewise, the pressure can vary from negative pressure to high pressure, such as 150–300 bar in ammonia synthesis; 6) catalyzed reactions can be conducted in batch or continuously… just to name a few cases to illustrate the diversity. In other words, the chemistry, process parameters, and physical methods used to determine physical properties of the catalyst and the reaction medium are widely varying.

The common element is that the target value in catalysis research is always a kinetic parameter, generally the rate of the formation of a certain product. The actual measured integral reaction rate in turn depends on many factors, such as process conditions, reactor geometry, energy and mass transport effects, solvents, or impurities. Intrinsically, however, the rate is related to physical and chemical characteristics of the catalyst. As a result, in catalysis research, unlike in related scientific disciplines such as materials science,[30] it seems to be useful to structure the data in a material-centric (catalyst-centric) way (Figure 3), not in an experiment-centric way.

However, since the physicochemical properties of the catalyst, and thus its performance, as already described in section 2.4, strongly depend on the process parameters and the previous history (catalyst synthesis, pretreatment and formation under reaction conditions), not only the rate and all measured physico-chemical properties of the catalyst should be directly linked, but also all data and metadata from the different life stages of the catalyst, including time. Notably, metadata include additional information such as a timestamp, device used, units, or operator. The workflow of the experiment from catalyst synthesis to the acquisition of the kinetic value is also part of the metadata. This multilayered entanglement should be considered when setting up electronic laboratory notebooks (ELNs) and repositories. It means that experiment-centered data from synthesis and characterization (fields bordered with dashed lines in Figure 3) must be linked to the catalyst, resulting in an overall catalyst-centered structure (Figure 3). In the following, the specifics of the data from the different unit operations in catalysis research will be examined in more detail.

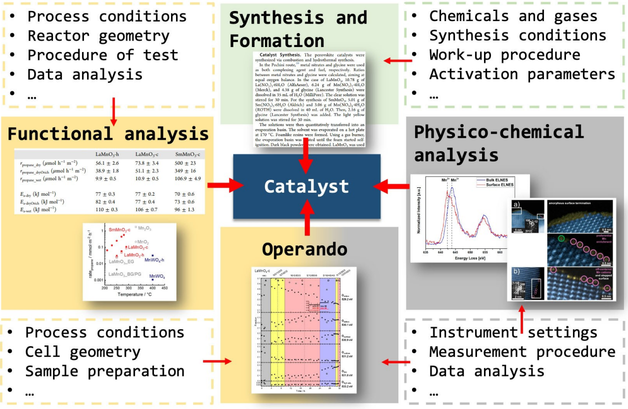


Figure 3. Components of catalyst-centric acquisition of experimental data. Examples of data and metadata from experiment-centric studies (functional analysis, operando analysis, etc.) are given in the boxes bordered with dashed lines. Screenshots of published data in the boxes for the various types of investigations are shown to illustrate the general very diverse form (text, tables, xy plots, images) in which data are currently commonly presented. Reused under Creative Commons CC-BY4.0.[29]

3.1. Synthesis data

Synthesis is probably the most sensitive and least discussed aspect in catalysis when addressing data management. The preparation of a catalyst is the first milestone in an investigation, and most reproducibility issues in later steps can be ascribed to synthesis protocols that lack detail. Currently, synthesis in scientific journals is generally described in text form (Figure 3). Details such as glassware used, lot number of chemicals used, and order of addition of reactants, belong to the synthesis data, in addition to the concentration of solutions, aging time, or temperature. Automated synthesis reactors and digitalized sensors facilitate this work even further and improve reproducibility by recording time-resolved data such as pH, temperature, or spectra.[31] It also enables mathematical methods, such as the design of experiment (DoE), to be used to accelerate studies.[31a,32] In heterogeneous catalysis, however, the often diverse, sequential unit operations required to prepare a catalyst, ranging, for example, from liquid-phase condensation chemistry to annealing at high temperatures, represent a technical challenge for automation.[28c,33] The pragmatic quick solution here is stepwise modular automation and manual transfer between unit operations.[28c] It is necessary to prepare a sufficiently large batch of a catalyst (in gram scale) to be able to perform all characterization and the physicochemical and functional analysis. Data sourced from the same batch contributes significantly to improving the reliability of catalysis data.[27g,34]

3.2. Data of physicochemical analysis

Here the type of data is just as varied because numerous methods fall into this category and it could, for example, consist of an xy file (e.g., an infrared spectrum) or a matrix in the form of an image (e.g., from transmission electron microscopy (TEM)) (Figure 3). The device used and the measurement parameters (metadata) afford very important information for third-party users of the data. For example, the xy file of an X-ray diffraction pattern measured with a Cu X-ray source, represented as intensity versus 2θ, is different from the xy file measured with a Mo X-ray source.

The characterization of the catalyst should be done at all stages and should include not only that of the freshly prepared catalyst precursor, but especially an analysis of the spent catalyst.

Furthermore, it is very important to define the conditions of the sample transfer. Catalysts generally do not need to be studied ex situ, as most characterization techniques today have capabilities for in situ pretreatment under defined conditions. Operando measurements[35] are preferred since kinetic quantities can be directly assigned to intrinsic physical properties of the catalyst (e.g., determined spectroscopically). However, operando studies are not always applicable for physical and methodological reasons. In these experiments, two investigations that are already complex in themselves must be linked to the catalyst. Particularly in synchrotron-based experiments,[36] immense sets of metadata are associated with both the physical quantity, such as a binding energy or an X-ray absorption spectrum, which on their own require complex data analysis,[37] and the reaction rate (e.g., measured by mass spectrometry).[38] For technical reasons, the reactor geometry of an operando cell is often different from that of a laboratory reactor, which can result in different kinetic operando data compared to the data of the same catalyst measured under the same reaction conditions in a fixed bed reactor.[39] Common sources of errors in operando measurements are gas bypass, inadequate temperature control, and transport limitations.[40]

3.3. Data of functional analysis

This data is the core data in the field,[41] and consists primarily of chemical analytics (e.g., concentration of compounds in the reacting phase determined by techniques such as different types of chromatography, thermal conductivity, mass spectrometry, or infrared detectors) coupled to time, which is then transformed to Figures of merits such as selectivity to a product as a function of the conversion of a substrate or a reaction rate as a function of the reaction time or the temperature. In high-throughput experiments, which are often applied in catalysis, these data are generated in parallel for several or a large number of catalysts partly at the expense of accuracy, but reliable enough to achieve a catalyst ranking.[5f,h,8d]

The rates are used to estimate further kinetic parameters such as apparent activation energies and reaction orders, and they are employed in modeling, resulting in mathematical rate expressions. To avoid pitfalls in the acquisition of kinetic data, recommendations are given in the literature for the selection of the experimental reactors, reaction conditions, best evaluation practices, and kinetic modeling methods.[27a,42] Here, collaboration between research groups can contribute to determine realistic error bars by measuring reaction rates on catalytic reference materials and reactions.[28a–c] Such measurements seem to be rather routine, but are absolutely necessary to determine valid data, which is crucial for analyses using artificial intelligence. The scientific benefits of such laborious approaches have already been demonstrated.[34]

3.4. Computational data

Today, modeling and data science play an essential role in advancing science by contributing to the understanding of catalytic processes and significantly driving the development of new catalytically active materials.[22,43] Sharing of data in computational catalysis is already more established compared to experimental research. Especially density functional theory (DFT) data is widely released on open repositories,[44] and exchange of analysis codes and software is widespread on GitHub.[9] However, the rigorous determination of descriptors in computational chemistry requires compliance with best practices[45] and benchmarking.[5g,43] Since the data generated in first-principles-based calculations and data science originate from many different computer codes, they are rather heterogeneous. In order to be able to include such data in freely accessible repositories, they must be reprocessed, normalized and converted into a common, code-independent format using parsers, as has been implemented, for example, in the Novel Materials Discovery (NOMAD) database, a platform for the exchange and use of materials science data.[2,44a] Other platforms[44b–c] directly provide adsorption energies for adsorbates on surfaces, in some cases even reaction barriers for surface reactions, making it easier to combine data from different data sets for catalysis applications (e.g. for the development of microkinetic models or for machine learning).

Incorporating data-driven approaches, such as ML, necessitates the awareness, that the accuracy of the developed model is inherently limited by the quality of the input data used to train the models, adhering to the phrase "garbage in, garbage out".[5e] Therefore, it is imperative for either computational or experimental data utilized for model training, to be recorded as accurately, reliably, and reproducibly as possible.[27g] Furthermore, we have to be aware that, any data that is used for ML applications might have served different objectives in their original studies, which will impact the types and accuracy of data. Large-scale screening or high-throughput studies might have cruder settings, with lower accuracy compared to a detailed study done for a single surface/catalyst with the aim of developing a microkinetic model.

Many different algorithms are available that vary in the size of the dataset required, in the precision of predictions, in complexity and in whether there is a gain in understanding of the underlying principles associated with the prediction.[5a,46] Neural networks and ensemble-based methods, for example, offer little insight,[46,47] but can help bridge the gap between different time and length scales of typical DFT calculations and reactor studies and enable fast screening in catalysts (materials) informatics.[5h] Interpretable ML algorithms, on the other hand, can reveal the physical laws underlying catalyst's properties and enable further development of catalyst systems based on hypotheses that follow from this analysis, which can accelerate catalyst discovery in a more sustainable manner.[5a,34b] These transparent methods in catalysis informatics[5i] include, for example, tree-based classification and regression methods.[5a,46] Active learning algorithms using Bayesian optimization techniques have been developed to rapidly identify candidates with tailored properties for further validation either through experimental synthesis or more elaborate computations in research campaigns.[5h,46]

When done right, ML can help address current problems in first-principles studies such as studying the realistic complexity and dynamics of the active sites, including coverage and solvation effects, and overcoming differences in temporal and spatial dimensions of catalytic processes.[5]

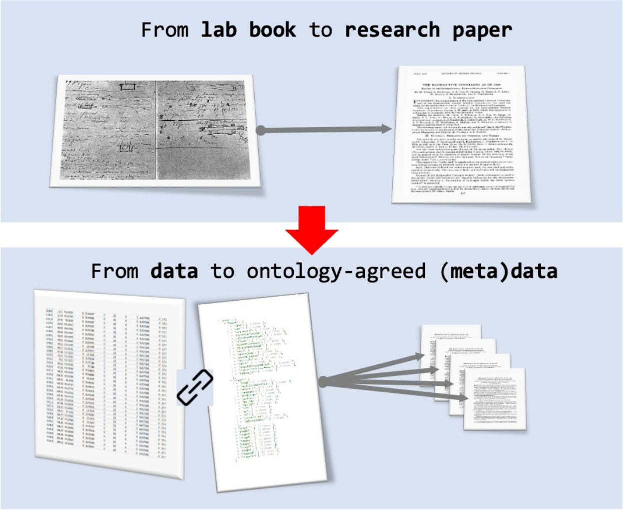

From lab book to research paper
From data to ontology-agreed (meta)data

Figure 4. Present (top) and future (bottom) of data and metadata use. Machine-readable open data can be analyzed under different perspectives in the future and thus be included in several publications.

3.5. Individual and research community responsibilities

From a top-down approach, the definition of catalysis data implies the establishment of an ontology, and thus, requires wide community acceptance. This can only be achieved by active involvement of all stakeholders and needs to be led by initiatives with a wide target, such as the US Materials Genome Initiative (MGI)[48] or the NFDI initiative in Germany.[49] From a bottom-up approach, scientists at the bench need to grasp the complexity and intricacy of catalysis data and their collection, moving from note-taking as a memory aid towards comprehensive reusable documentation (Figure 4). Such awareness should be encouraged by appropriate mentoring from the early stages in the scientific career, setting the example of rigor over novelty in research. This is not always the case, and quality suffers because in today's "publish or perish" culture, quantity often takes priority over quality, leading to a lack of disruptive insights.[4]

4. The current catalysis data infrastructure

4.1. Catalysis data are difficult to find

Catalysis data are currently essentially only findable through literature searches. Finding datasets in catalysis publications is complicated because papers typically do not contain standardized data objects and policies for making data available vary across journals. Publishers are already contributing by requiring that data be uploaded to repositories before publication. However, the data that support the findings are most frequently only available from the corresponding author upon request or uploaded to the Supporting Information. Increasingly, databases and community-wide repositories are also being used (see Table 1 for examples). An overview of possible databases for different research areas is provided, for instance, on the website of Open Research Europe, an open access publishing platform for the publication of results obtained in various European funding programs.[56] However, even if

Table 1: Examples of databases relevant for catalysis and data-driven infrastructure services.

Name	Host	Link	Content, remarks
NOMAD[44a]	Humboldt University Berlin	https://nomad-lab.eu/	Computational and experimental catalysis
Catalysis Hub[44b]	Stanford University	https://www.catalysis-hub.org/	Computational catalysis
CADS[50]	Hokkaido University	https://cads.eng.hokudai.ac.jp/	Catalysis data, 12 data sets (>12000 entries)
Kinetics Database[51]	NIST	https://kinetics.nist.gov/kinetics/index.jsp	Chemical kinetics for gas-phase reactions
SCIENCE DATA BANK[52]	Chinese Academy of Sciences	https://www.scidb.cn/en	8 886 results for "catalyst"
Cambridge Structural Database (CSD)[53]	CCDC and FIZ Karlsruhe	https://www.ccdc.cam.ac.uk/structures/	Structural data
SwissCAT+[54]	ETH Zürich, EPFL	https://swisscatplus.ch/	Data-driven infrastructure
zenodo[55]	CERN	https://zenodo.org/	2645 results for "catalyst" in January 2023, but mainly pdf's of articles
Materials project[44c]	Lawrence Berkeley National Laboratory	https://materialsproject.org/catalysis	Open Catalyst 2020 Dataset (>500 000 entries of adsorption energies)
Materials cloud[44d]	Swiss National Supercomputing Centre	https://www.materialscloud.org/home	27 publication results for "catalyst"

some catalyst databases have already been set up,[50b,57] their use is not very common.

4.2. Most of the catalysis data are not accessible

Much of the data are not retrievable with universal communication protocols, if available in archives at all. This applies not only to the valuable data collected in industry when optimizing commercial catalysts, which are protected for intellectual property reasons, but also to many results, especially negative findings,[58] in academic research. Therefore, catalyst data accessible today, are generally not diverse enough. Particularly poor catalysts, which are often not mentioned in publications are, however, of great importance for data sets that are evaluated using machine learning.[58b,59] Often as well, for a given reaction, the entire community focuses on only one or a few different catalyst compositions. So, for example, the majority of catalysts studied so far in ammonia decomposition contain Ru as the active metal.[60]

4.3. Catalysis data are not interoperable and not reusable

A standardized data structure for catalysis data is completely lacking, and the archived data are most frequently not released, for example, with a Digital Object Identifier or any other unique identifier (ID).[8b] For kinetic data, it should be possible to come to a standard data structure as is common, for example, for structural data in the Inorganic Crystal Structure Database[61] or for the evaluation of surface area and pore size distribution.[62] Standards for data models in materials science are, for example, Crystallographic Information Files (CIF) for crystal structures,[63] or the NeXus format for Neutron, X-ray, and muon science.[64] Recommendations for performing kinetic studies have existed for a long time,[27a,b,42a–c,65] but a standardized data output that is universally applicable to various catalyzed reactions has not been developed or has not become commonly accepted. To accelerate progress in this area, the following information could be requested when publishing:

- Complete, standardized sets of data and metadata of kinetic measurements, as is already common practice in the publication of crystallographic data of new structures, including the workflow of pretreatment and test.
- Provision of the catalytic data always also numerically and not only in the form of a graphical representation.
- Indication of the formulas used to calculate the evaluated data.
- Details on criteria used to test for mass transport phenomena, such as the Carberry number and the Wheeler-Weisz criterion.[41b]

However, the development of these requirements is the responsibility of the community and cannot be accomplished by the editors alone.

4.4. About the necessity of developing an ontology

If a minimum set of data is determined according to prescribed methods that are generally accepted, this enables rapid cross-laboratory comparison of catalysts. To facilitate knowledge sharing, agreements are needed on the development of a generally accepted language for representing knowledge (i.e., an ontology) that specifies not only the vocabulary used for the designation of experimental parameters or particular catalyst properties, but also the links between the data. Generally speaking, an ontology is a specification of a conceptualization.[66] The conceptualization summarizes the objects and concepts that are assumed to be of interest in a particular domain and describes the relationships that exist between them (e.g., that the composition of a material is related to its electronic or magnetic properties, or the crystal structure to structural porosity).[58a,66] No ontology of catalysis science exists, but there is an awareness that development is necessary.[8b,48,58a,67] Efforts to develop common ontologies are currently led, for example, by the NFDI4Cat consortium in Germany,[8d,68] but the catalysis community must increase their involvement and get into action to make the proposal a reality and develop community standards.

4.5. Catalysis data are not always reproducible

The frequently observed insufficient reproducibility of published data in heterogeneous catalysis is primarily not due to inadequate experimental practice,[27b,28d,42c,69] but to the inherent metastable nature of any active catalyst.[17,70] As a consequence, catalysis data have no value without complete specification of metadata. Emphasis must be placed on comprehensive documentation, but it is sometimes difficult to see in advance which parameters belong to a complete metadata set. Catalyst synthesis can be difficult to reproduce even when all reaction parameters are described very precisely, for example, when trace impurities can be introduced by the starting compounds[71] or the synthesis conditions have a sensitive effect on the metal particle size.[72] Lot numbers of the chemicals used, the gas purity, and the time sequence of the unit operations must therefore be part of the metadata of catalyst synthesis. Poor reproducibility sometimes also results from catalyst synthesis in batches that are too small and, in particular, if the batch size is not specified. However, miniaturization is not an option in view of the complex problem of upscaling.[73] Also for the purposes of fundamental research, it is important to synthesize a catalyst on a gram scale in order to be able to perform all investigations on only one batch and thus arrive at reliable structure–function relationships.[27g,34]

Furthermore, the preparation of the catalyst is not completed with chemical synthesis and thermal pre-treatment of a material. In catalysis research, it is well known that the active catalyst is formed only under the reaction conditions of the specific catalytic reaction from the fresh precursor introduced into a test reactor.[74] The conditioning in the so-called formation phase may involve either activation or deactivation. The result of the transformation of the precursor into a specific modification of an active catalyst (i.e., the resulting steady-state performance) also depends on the workflow applied to achieve the steady state (e.g., whether a temperature variation was started at low or high temperature).[75] Therefore, the details of pressing and sieving, and all formation metadata, including the sequence and duration of explicit treatments, must be linked with the specific steady-state performance. As long as full automation of experiments and/or automatic recording of all process parameters is technically not yet possible in individual laboratories, standard operating procedures for catalyst testing are useful and are urgently needed.[27a,c,e–g,28b,65,76] However, despite this awareness, SOPs are not widely adopted in the general laboratory practice in catalysis. In automation solutions, criteria must be inserted to verify that steady state has been reached,[27d] and to check whether the data are falsified due to mass and energy transport limitations.

4.6. When is a catalysis data set complete?

The catalytic properties are determined by physical parameters of the catalyst material and evolve in interaction with the reacting molecules. The underlying relationships are complex functions (Figure 1, right). Interpretable machine learning can help to uncover the hidden relationships, paving the way for rational design of new, improved catalysts.[5a,34] In order to gather the necessary data for this, it is important to comprehensively characterize catalysts at all stages of their life (freshly synthesized, formed (activated), spent). However, the question arises as to which characterization methods are important and whether the analysis may only be carried out under working conditions (in operando),[39,77] or whether certain important properties of catalysts are already reflected in the ex situ analysis of the freshly prepared precursor. For this purpose, it is necessary to conduct systematic studies by using reliable experimental data. Interpretable machine learning can likewise contribute to identify relevant characterization methods.[34b]

4.7. Published catalysis data are not directly comparable

Published data pose the challenge of non-uniformity. Databases are still being created using literature data.[78] Natural language processing (NLP) and information retrieval (IR) methods were applied,[79] such as those already successfully developed for materials science.[80] However, the analysis of published data by data science is risky because even the classic comparison in review articles shows that catalysts with seemingly the same composition tested under apparently identical reaction conditions have different catalytic performances, such as Ru or Ni catalysts in the decomposition of ammonia,[60,81] or vanadium-oxide-based catalysts in the oxidation of alkanes.[82] The reason may be experimental errors, but more likely it is the metastable nature of the catalyst and the resulting influence of the workflow of the investigation on the outcome of the experiment, as discussed above. Such inconsistencies in the literature data due to missing metadata but also lack of diversity, leading to the absence of entire potential phase spaces, could be the explanation for why the analysis of published catalytic data (e.g. for the oxidative coupling of methane) using machine learning methods generally did not yield any insights beyond the empirical knowledge already available.[78,83]

4.8. Catalysis data are not Big Data

Rigorous catalysis experiments are time-consuming and material-intensive. Therefore, catalyst data are generally not Big Data. However, they can become Big Data if catalyst data are rigorously measured, linked to all metadata, and become FAIR, making individual measurements available for analysis to the entire community and beyond, according to the FAIR principles. Thus, very large datasets can be generated if the entire catalyst community works to create shared datasets and different aspects of the data can be analyzed by different research groups and thus can be included several times in published results (Figure 4, bottom right). As a consequence, more machine learning methods can then be applied that absolutely require Big Data. To achieve this goal, it is necessary to change paradigms in catalysis research and adapt methods for measuring catalysis data to meet the needs of the coming digital age. Furthermore, the results of first-principles computations accelerated using machine learning can significantly increase the available data sets.[5c]

5. Data acquisition and storage concepts

As outlined above, classical approaches to data acquisition and storage in catalysis need to be revised.[5h–i,8a,d,84] Therefore, in this section, we discuss practical aspects and already existing solutions for important components of a local research data infrastructure in catalysis. All developed tools described below have been released for open access and feedback is welcome. References to the documentation are given in the corresponding chapters.

5.1. The central role of the scientist

One of the main components in a research data infrastructure, besides data itself, is the scientist (Figure 5). This might seem contradictory, particularly if one thinks that we will move from human-designed and -controlled experiments performed to test a hypothesis based on domain knowledge towards bias-free data acquisition and autonomous catalyst discovery.

The latter is a complex task,[28c] but even if it can be technically mastered in the future, the principal question

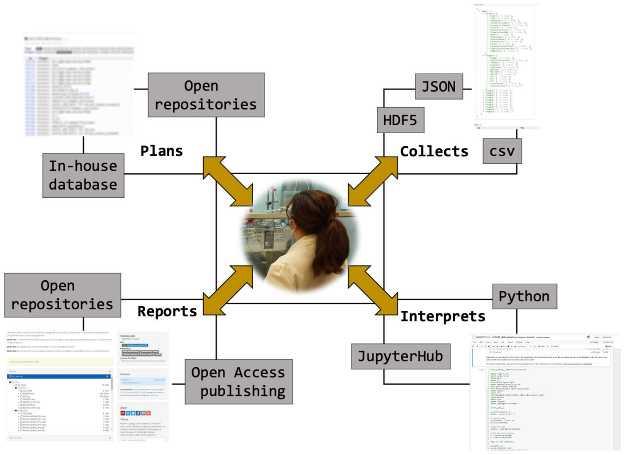


Figure 5. Pragmatic data acquisition and storage concept on the way to autonomous catalyst discovery.

remains as to whether the intervention of a scientist will always be required to propose, adjust, and technically approve experimental procedures. A new sense of responsibility should be fostered in young generations so that they understand the importance of consistency, optimization experiments, and negative results.[58b–d] Systematic completeness instead of the "one-gold" experiment should become the rule. In particular, as more data science concepts permeate the catalysis community, it is both an opportunity and a challenge for scientists to change their mindset and increase their information technology (IT) literacy, by doing hands-on practice with all the open-source available tools, as some tutorials encourage.[85]

Advances in catalysis science can only be made if both experimentalists and theoreticians can effectively communicate with each other, and we should do our part to bridge that gap by increasing the affinity to each other's fields. Data stewards have a particular facilitative and supportive function in this process. The implementation of modern data infrastructures in scientific institutions cannot be achieved without data stewards, which should also be taken into account in strategic planning.

5.2. Standard operating procedures

Any research project begins with a hypothesis based on a research question, either derived from prior knowledge or coming from a Design of Experiment (DoE) approach to understand a novel subject (i.e., it begins with a plan; Figure 5). In the field of catalysis, this can be generalized for most experimental studies to (i) synthesis experiments, (ii) specific characterization methods, and (iii) functional analysis in a reaction of interest under defined reaction conditions. The research question quite often refers to the performance of a catalyst in relation to its intrinsic physicochemical properties. Materials chemistry is frequently varied (e.g., the active metal content in a mixed metal oxide). These materials are synthesized, characterized in all phases of the life of a catalyst including the freshly prepared precursor, the activated, and the spent catalyst (by default with the determination of elemental composition, surface area and texture, analysis by diffraction, various spectroscopic techniques, temperature-programmed methods, and microscopy), and studied in a specific reaction while performing kinetic experiments. As an example, Figure 6 shows the planned workflow for a project in which catalysts for the oxidation of short-chain alkanes were investigated.[27g,34] For each of the unit operations in the study, a standard operation procedure (handbook) was pre-determined already at the planning stage of the project, with the help of both previous knowledge and some optimization experiments, which were also properly documented.[27g]

Generally, the instructions are method dependent. The SOP for a particular method (e.g., in situ Raman spectroscopy), should be optimized for each project, as the samples can vary (e.g., different wavelength of the laser excitation, or different measurement parameters are necessary for different samples and research questions). In this way, categories and instances are defined for each SOP (i.e., for each type of experiment). A collection of these SOPs can then be organized into a handbook for a specific project,[27g,34b] which does not necessarily have to be in text format as in the cited examples. It can be organized in a machine-readable file such as in the JavaScript Object Notation (JSON) format, which can be linked to the data generated with the corresponding method. Actually, this is preferable, as it allows documentation of the handbook according to which measurements were performed, and makes it easier to modify SOPs if required in a project due to newly gained scientific knowledge.

SOPs should preferably include benchmark catalysts and models on which experimental or calculated data, respectively, can be validated. In some industrial laboratories it is common practice to verify the performance of analytical techniques and staff through national accreditation rounds in which participating laboratories analyze standard samples provided by an auditing body. Applying this concept to academia would immensely increase the reproducibility, reliability, and quality of catalyst data, although this would be a major challenge in terms of the scale and diversity of topics in global catalysis research. However, this should ultimately be possible as SOPs are well established for other disciplines—in particular in the biosciences[86]—and actively applied and discussed amidst Good Practices (GxP), also as part of codes of conduct.[87] Concepts such as data integrity and storage,[88] or recommendations on how to write SOPs,[89] have already been discussed in these fields and can be directly applied to other subjects. As experiments following SOPs can be demanding for a researcher, a major challenge of this approach is how to manage the motivation of the scientists involved and enhance creative work. From our own experience, we can remark that providing automation options and insisting that SOPs are a minimum requirement (any additional experiment is not discouraged) offer a good counterbalance. Additional advantages from SOPs,

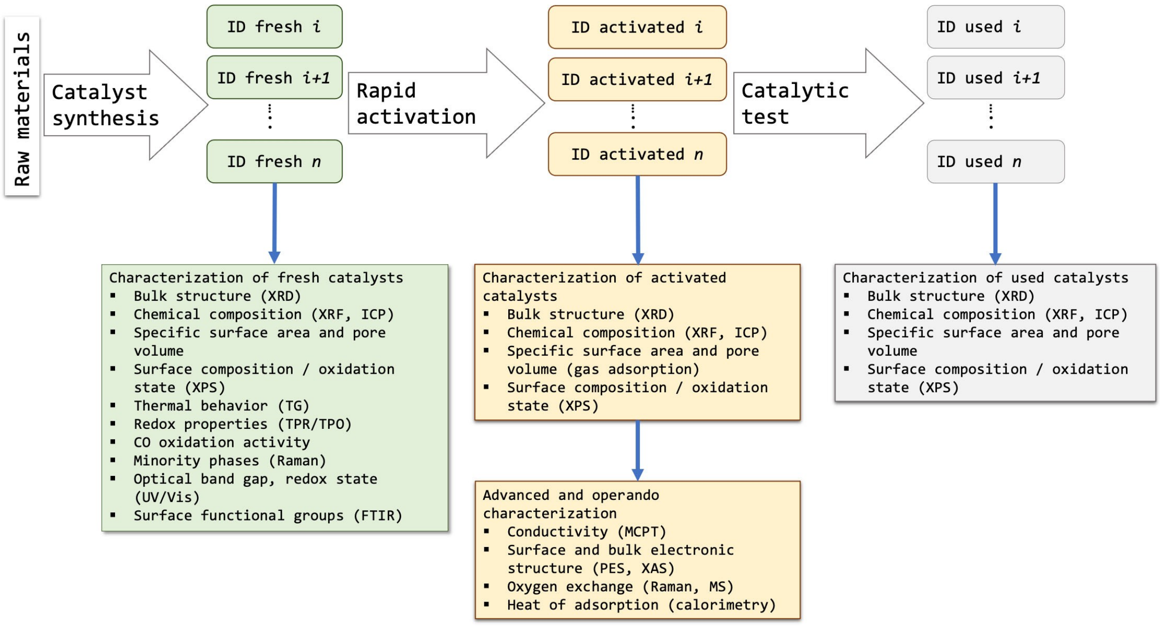

Raw materials
Catalyst synthesis
ID fresh *i*
ID fresh *i+1*
ID fresh *n*
Rapid activation
ID activated *i*
ID activated *i+1*
ID activated *n*
Catalytic test
ID used *i*
ID used *i+1*
ID used *n*
Characterization of fresh catalysts
Bulk structure (XRD)
Chemical composition (XRF, ICP)
Specific surface area and pore volume
Surface composition / oxidation state (XPS)
Thermal behavior (TG)
Redox properties (TPR/TPO)
CO oxidation activity
Minority phases (Raman)
Optical band gap, redox state (UV/Vis)
Surface functional groups (FTIR)
Characterization of activated catalysts
Bulk structure (XRD)
Chemical composition (XRF, ICP)
Specific surface area and pore volume (gas adsorption)
Surface composition / oxidation state (XPS)
Advanced and operando characterization
Conductivity (MCPT)
Surface and bulk electronic structure (PES, XAS)
Oxygen exchange (Raman, MS)
Heat of adsorption (calorimetry)
Characterization of used catalysts
Bulk structure (XRD)
Chemical composition (XRF, ICP)
Specific surface area and pore volume
Surface composition / oxidation state (XPS)

Figure 6. Workflow and planned experiments in a project on selective oxidation in which work was carried out according to a handbook.[27g,34] In the project, twelve catalysts were prepared, activated, analyzed in all stages and tested in ethane, propane, and *n*-butane oxidation according to standard operating procedures.

such as improved time planning, rigorous and comparable in-house and inter-laboratory results, are only some of the obvious benefits. The initial effort needed for developing the SOP pays off by easing the path in the measurement protocols, and can also be reused for other projects. Our vision is that, in the near future, scientists at a global level working on similar topics can use the same SOP. In this way, data produced anywhere in the world, once shared, can be directly compared as long as it follows the same minimum requirement.

Once an agreement is formed, a platform is needed, where the SOPs and the data collected over a project are available for those involved. We propose that this is first done locally for each laboratory, and after publication, data can be shared (see Section 6) either in overarching repositories (Table 1) or by conceding open access to the data entries in the local data structures.[27g,34b]

5.3. Electronic laboratory notebooks and databases

The digital format to record observations and results has significant advantages over the paper alternative. Clear data records, enhanced quality and ethical control, overview of samples and experiments, and easier data transfer, are among the many assets to be considered, which are also in line with several transparency initiatives, such as the Lindau Guidelines.[91]

The advantages are already recognized by researchers at various levels, and versions of Electronic Laboratory Notebooks and Laboratory Information Management Systems (LIMS) have flourished in past years.[92] ELNs for general use can be very suitable but have two main disadvantages: 1) that the transition for the scientist is not always straightforward, and 2) that ELNs usually need to be complemented with a database/data archive, used as a repository. Solutions adapted to catalysis are under development.[5h,93] The concept of research data management (RDM) at the Department of Inorganic Chemistry, Fritz-Haber-Institut der Max-Planck-Gesellschaft, provides an enhanced option, as the focus is on automating processes to ease the transition and on integrating both the ELN functionalities with a data archive, which has been implemented in the AC/CatLab data archive (Figures 7 and 8).[90]

The AC/CatLab data archive (Figure 7), which was launched in 2003, and is currently used in the department, has recently been further developed into an ELN tailored to the needs of research in heterogeneous catalysis.[90] It has a flexible structure and can be expanded and adapted upon demand. It works upon classifying data into document types, and assigning each entry to a specific project

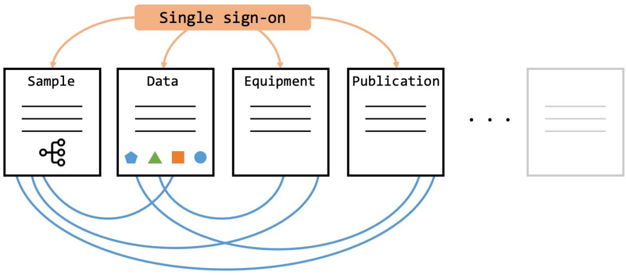
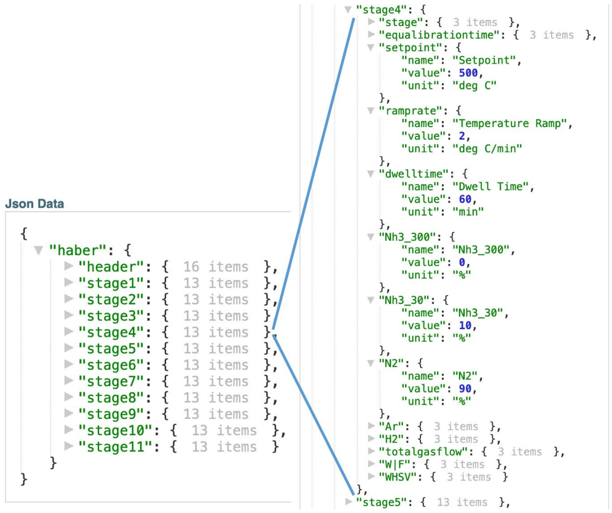


Figure 7. AC/CatLab data archive: flexible architecture of an ELN for catalysis research.[90]



```
{
  ▼ "haber": {
     ▶ "header": {  16 items  },
     ▶ "stage1": {  13 items  },
     ▶ "stage2": {  13 items  },
     ▶ "stage3": {  13 items  },
     ▶ "stage4": {  13 items  },
     ▶ "stage5": {  13 items  },
     ▶ "stage6": {  13 items  },
     ▶ "stage7": {  13 items  },
     ▶ "stage8": {  13 items  },
     ▶ "stage9": {  13 items  },
     ▶ "stage10": {  13 items  },
     ▶ "stage11": {  13 items  }
    }
}
```

```
▼ "stage4": {
   ▶ "stage": {  3 items  },
   ▶ "equalibrationtime": {  3 items  },
   ▼ "setpoint": {
         "name": "Setpoint",
         "value": 500,
         "unit": "deg C"
     },
   ▼ "ramprate": {
         "name": "Temperature Ramp",
         "value": 2,
         "unit": "deg C/min"
     },
   ▼ "dwelltime": {
         "name": "Dwell Time",
         "value": 60,
         "unit": "min"
     },
   ▼ "Nh3_300": {
         "name": "Nh3_300",
         "value": 0,
         "unit": "%"
     },
   ▼ "Nh3_30": {
         "name": "Nh3_30",
         "value": 10,
         "unit": "%"
     },
   ▼ "N2": {
         "name": "N2",
         "value": 90,
         "unit": "%"
     },
   ▶ "Ar": {  3 items  },
   ▶ "H2": {  3 items  },
   ▶ "totalgasflow": {  3 items  },
   ▶ "W|F": {  3 items  },
   ▶ "WHSV": {  3 items  }
  },
▶ "stage5": {  13 items  },
```

Figure 8. Metadata is not entered into predefined fields, but is contained in a single field in the form of JSON scripts, which can also be searched. These scripts are both machine- and human-readable. Here, metadata of an automated test reactor is shown and exemplarily unfolded, revealing the setpoint temperature and the heating rate in test segment 4 of an experiment and what gas composition is used.[90]

(single sign-on, Figure 7). The access authorization is governed by the involvement of the respective user in the corresponding projects.

The document types are called *Sample*, *Data*, *Publication*, *Chemical*, *Gas*, *Equipment*, and *Instrument.* All the document types can be filtered, so that every item is included, and making an inventory is fairly easy.

Each item/entry receives a unique identifier (entry ID), including the samples, chemicals, and hardware. The ID together with a Quick Response (QR) code and a description can be printed on labels by pressing on a printer button. When these chemically resistant labels are scanned on a sample, chemical packaging, gas bottle, or on a device in the laboratory, for example with a cell phone, the entry in the archive is opened and the corresponding information can be obtained.

Mandatory fields in each data type have been kept to an absolute minimum, as most of the information can be stored in a searchable JSON field in a flexible and future-proof way (Figure 8).[90]

Both the document types and the fields in each document type can be extended as desired (Figure 7). Any amount of data can be uploaded to an entry and there is a comment field that can contain information about the data. In addition, *each entry* can be linked to *any other entry* (Figure 7). A rapid visualization of the links is in development. The different data types are explained in more detail in the following.

5.3.1. Document type Sample

In this entry, information of the sample (in house prepared, from collaboration partners, industry standards) is summarized. Normally, this is a text, but data of automated catalyst synthesis or different types of metadata can be added in the various subfields (preparator, date of synthesis, or receipt) and a JSON file can then be developed upon the requests of each project SOP/Handbook. Data related to the synthesis (e.g., a pH-time curve), pictures taken during synthesis, or a set-up flow chart can be uploaded here.

When the sample is analyzed, a link is made from the *Sample* entry to the *Data* entry with the results. Thus, it can be seen at a glance which examinations were performed on the sample with a particular ID. Any treatment of a sample (e.g., calcination, pressing and sieving, catalytic test) results in the creation of a new sample (with a new unique sample ID) for which a new entry is created and which is linked to the parent sample. In this way, the life history of a catalyst can be followed precisely and characterization data obtained at each stage, such as the XRD of a freshly prepared, calcined or spent sample, can be clearly assigned.

5.3.2. Document type Data

Any type of file can be uploaded here (XRD diffractogram, TEM images, calibration graphs, etc.). In most of the cases, these entries contain data of characterization experiments or catalyst tests and are, therefore, linked to the corresponding sample entry as explained above. The metadata of the analysis is covered in a JSON file (Figure 8). JSON files are automatically generated in data-driven projects for experimental plans and in automated measurements for the metadata. For manual experiments, web-based Graphical User Interfaces (GUIs) are provided for the generation of JSON scripts using the metadata entered manually via the GUI.

The JSON files, including the metadata in all document types, are machine readable and searchable and, most importantly, can also be adjusted retroactively by using scripts (e.g., in an automated way) if community ontologies change or a different SOP needs to be followed in another project. It is valuable to upload, besides the proprietary software file, the data in an accessible format (e.g., comma-separated values (csv) or Hierarchical Data Format 5 (HDF5)). Direct visualization of data is of great advantage and work is in progress to implement this for all data types.

5.3.3. Document types Chemical/Gas/Equipment/Instrument

Each chemical or gas that is purchased is included in the data type *Chemical* or *Gas*, respectively, and information reported comprises opening date, current user, location, purity, safety data sheets (SDS), certificate of analysis, etc. Chemicals can then be labeled and traceable upon performance of various experiments. The entries are linked to *Samples* or *Data.* So it is clear which chemical was used in a catalyst synthesis or which gas was used in catalyst testing.

Every commercial device or in-house developed set-up is encompassed as *Equipment*; for instance, a complete test reactor set-up, or a spectrometer with gas delivery infrastructure and analytics. If the *Equipment* consists of exchangeable parts, such as mass flow controllers, oven, thermocouple, etc., these are denoted *Instruments* and consist of an entry type of their own linked to the corresponding *Equipment* entry. Only inventoried parts are included. Data such as calibration curves, date, and reason of repairs, are uploaded here.

If an *Instrument* is exchanged in an *Equipment* (e.g., because a mass flow controller or a valve is defective), this results in the creation of a new *Equipment* entry that is a descendant of the original *Equipment* (the two entries are linked). In this way, different versions of setups can be clearly tracked and assigned to the experiments that were performed with it. The *Data* entry that contains analysis results measured using a particular setup is linked to the current version of the *Equipment* used for the measurement. The same applies to *Sample* entries and the *Equipment* used for the synthesis.

5.3.4. Document type Publication

Here reports, posters, conference presentations, internal meetings, or research articles are included. The main goal is to improve the collaboration and information flow in the projects.

5.3.5. Interface and access

What is particularly important, the AC/CatLab data archive incorporates an Application Programming Interface (API) that can be used to remotely request, upload, and access data from automated and manual experiments or theory, and link the local repository to overarching repositories (Figure 9). Furthermore, specific entries (*Samples*, *Data*, *Chemicals*, *Publication*, etc.) can also become public after publication, upon approval from the project-assigned Principal Investigator (PI) (see for example references [95] and [34b]).

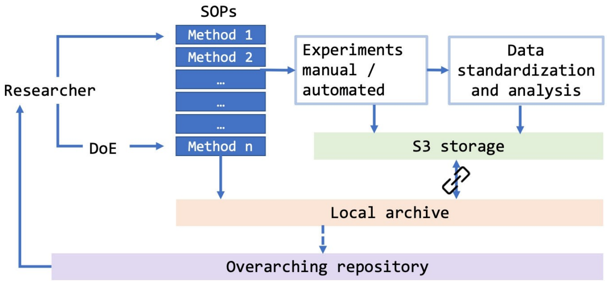


Figure 9. The example of the local data infrastructure at the Department of Inorganic Chemistry, Fritz-Haber-Institut der Max-Planck-Gesellschaft implemented in the CatLab project.[90,94]

5.4. Experiment automation and data upload

Automation is one of the key aspects in our strategy.[94,96] Automated experiments provide comparable data, especially when performed repeatedly on many samples. Human errors in the experimental procedure are thus minimized. In addition, once those tasks are automated, they can relieve the scientist from the time consumed not only in the experiment (automated experiments), but also in the data storage (automated upload), and the first preliminary data analysis and handling (automated standard data analysis) (Figure 9).

There are two main approaches for automated experiments, (i) purchasing a commercial device, or (ii) developing an internal experimental set-up. The former has the advantage that it is an easy solution, usually established for a specific task, but lacks the flexibility of the latter.

In an internal set-up, an automation concept can be elaborated upon specific requests, non-proprietary file extensions can be used, and it is easier to adapt to future experiments, albeit time- and resource-consuming at the beginning. Control software such as LabVIEW or EPICS[94a] can be used to integrate the different instruments (e.g., ovens, detectors, mass flow controllers, etc.) and also create a Graphical User Interface for the users. Data collection in fully automated experiments[94b–c] consist of an automatic processing of a recipe, creating the database entry for the data and the method, recording and uploading the measurement files to the data entry, generating and uploading a JSON file that contains all the metadata of the measurement to the data entry and, finally, establishing all links (e.g., to *Sample*, *Equipment*, *Chemicals*, *Gases*) and, if appropriate, also generating new sample entries and further links (such as the sample entry of a spent catalyst if the data are catalyst test data, which will be linked to the sample entry of the pressed and sieved sample and to the data entry containing the test data).

In experiments not automated or automated with proprietary software, the scientist is also assisted with automatic uploading of metadata and data to the archive.[96] The generation of a JSON file that contains the metadata is automated with the help of a JSON-file generator, where the scientist interacts with a graphical user interface (a form on a web page) to provide the specific information for each field. As different types of data exist in catalysis, as described in Section 3, a dedicated JSON generator is developed for each case. The uploading of the measurement files is carried out by an archiver appliance program, which is connected to a dedicated S3 bucket (a cloud online storage) for each equipment. The software ensures that a data entry is generated and that all data and metadata are uploaded correctly to that entry and that the entry is appropriately linked to other related entries.

Automation is valuable, though expensive both in terms of human and financial resources. Therefore, it can only be implemented in steps and also not for all experiments for scientific and methodological reasons. Thus, manual experiments continue to play an important role in our data strategy. Any manual experiment can still be documented in our data architecture within a JSON generator. It is our advice to start using these tools in a systematic manner even for manual experiments, before all devices and processes are ultimately automated. This will provide important preliminary work for the establishment of automation concepts and promote close cooperation between the catalysis researcher and the information technology specialist.

6. Data use and sharing

Once data is collected, it can only be useful if knowledge can be gained from it. In particular, raw data is of upmost value for future use (Figure 4) as it is only intrinsically connected to the experiment and its design, and no human interaction, neither interpretation nor analysis has been performed. Raw data can be re-used, re-analyzed, or added to other datasets for comparison. The definition of outliers, models, and statistical noise, can change with the paradigm and tools available at a defined time. From a philosophical view, the possible interpretations and analyses are then as limitless as the number of scientists that work with it, thus highlighting the importance of open licenses for datasets of raw data, careful metadata documentation, and dedicated data curation. For this to take place, we need as a community to undergo a cultural change, and put the value of scientific insight into rigorous systematic experiments instead of gradual novelty.

In the future of data analysis in catalysis, we should learn from other disciplines such as informatics, where clean code is the aim.[97] Version control (e.g., via GitHub) is also a common feature that can become useful,[9] as shown already by other disciplines.[98] Specific data analysis in publications should be reported, as there are several tools available to make data analysis a much more clear procedure. For example, along with the raw data from a measurement, one could provide a Python script describing the data analysis,[99] as our group did in an earlier example (Figure 5).[100] This automation of data analysis will also make the scientist's work easier and the data more reliable, even if some effort is required in the first set-up stage until successful deployment. As stated above, the required IT-literacy from the new generation of scientists will be higher than it is now, but as digital natives join academia in the next years, the progression can be smooth.

Standardized data analysis appears as an alluring possibility, only if handbook protocols (SOPs), as described above, are followed, and easy accessible reference samples (e.g. reference materials from government agencies such as the *Bundesanstalt für Materialforschung und -prüfung* (BAM) in Germany,[101] or benchmark catalysts provided by industry[28b]) are tested in inter-laboratory experiments. Huge efforts are still needed in this regard, as consensus of how data is analyzed in each method vary on the type of sample, device availability in the different laboratories, and technical skills. Historical examples exist of well-established protocols, for instance in benchmarking water-splitting catalysts,[27c] the analysis of data from nitrogen physisorption,[62] or in temperature-programmed experiments.[76a,102]

The subsequent logical step of analyzing data is to share it, both raw and analyzed data, and its interpretation. This is usually done in publications, where descriptive text accompanies graphical representations or tables, simplified for better understanding. We need to shift out of this paradigm, and share data instead (Figure 4). The importance of open science initiatives focusing on open, machine-usable data has recently been highlighted in this context.[10] For this to be possible, journal editors should ask authors to submit their research data in repositories (Table 1), and follow guidelines to ensure compliance with the FAIR principles. The European Commission recommends to use a community-recognized repository, and in case that does not exist, use a certified repository.[56] For catalysis, no such specific repository exists yet, and it should currently be our main task as a community to establish a catalysis database or a catalysis platform for searching in different repositories. We need to organize a data curation system within the community to avoid creating "data dumps" where scientists just upload files without clear naming of metadata, or in a proprietary format. Community-driven initiatives (such as The Turing Way in data science)[103] should be used as inspiration to foster the development of collaborative data sharing. We will all benefit from data sharing in the end, as new knowledge will be gained, unnecessary experiments will not need to be repeated, and the main responsibility relies on all of us taking the step further by actively contributing to the discussion.

7. Conclusions

The deeper reason for the complexity in catalysis lies in the kinetic nature of functional materials as opposed to thermodynamically stable modifications of compounds, which have no function. Each function is inherently associated with desirable (high activity) or undesirable (deactivation) reconstructions of the catalyst, and the complicated underlying processes are not easily understood. To gain understanding, complexity must be captured and accounted for. This can only be achieved by changing the way we work with and handle data.

The transformation to digital working schemes is progressing rapidly in many areas, and heterogeneous catalysis must increase its pace, as otherwise the conventional research approach will fail to deliver disruptive discoveries of new catalyst technologies to solve the key challenges of our time. Redesigning the existing data infrastructure in catalysis research is only possible with a community-wide open discussion—to which this article also aims to contribute—and by taking small but steady and rigorous steps toward data-driven research concepts, defined workflows, comprehensive data and metadata storage, and fair data use and sharing in the local research environment. In this regard, we consider the following points to be particularly important:

- The paradigm in catalysis research has to change, away from superficial efforts aimed at short-term publication success toward systematic and rigorous experimentation in close cooperation with theoreticians and IT specialists to generate data with the required quality for data science.
- To achieve this, working according to Standard Operating Procedures or Handbooks is required at least for a central part of the experiments, where data sets comparable to all other catalyst laboratories are generated. Both the kinetic and characterization data must be validated against benchmarks, and these tests must be published along with the data from the research project to be published. Initially, the initiative to generate such benchmarks may come from individual projects in which different laboratories collaborate and one cooperation partner (e.g., the industry partner) synthesizes the benchmark catalyst. In the future, however, we see the responsibility for defining and providing benchmarks for different catalyzed reactions community-wide resting with national and international research organizations. Alternatively, research communities and appointed laboratories could be available to review results that are to be published. This means that catalysts which show outstanding performance are submitted for independent verification prior to publication.
- Catalysis research is interdisciplinary and diverse. Therefore, different solutions for optimized local data infrastructures will emerge. One concept, which has already been largely implemented in the daily scientific work of the authors, is presented in Section 5 of this article. It is vitally important that the developed tools, such as ELNs and local repositories, are sufficiently flexible to accommodate changes in research directions, applied methods, and ontology.
- Sophisticated ontologies are critical, as are curated repositories that remain useful and searchable even as the amount of archived information grows over time. It is important to develop community-wide accepted methods for quality assessment before data are released.
- Repositories must have powerful application programming interfaces to enable data exchange. The chemist of the future will need to be skilled in using sophisticated tools for data retrieval and processing, and collaboration with mathematicians for documenting, archiving, and extracting value from information will be helpful.
- To increase acceptance, local solutions should be designed in such a way that the researcher does not have to adapt (e.g., by increasing the workload, as would be the case by filling in numerous mandatory fields in a database), but that the solutions are adapted to the needs of the research. This requires close cooperation between the researchers and IT specialists, which should be mediated by a data steward who has insight into both areas and who also fulfils a certain control function.
- Automated solutions and standard operating procedures that include benchmarks support the data management and prepare the way for autonomous catalyst discovery, which is still far from reality.

All these changes and developments can be driven locally in individual laboratories. The establishment of overarching repositories, in compliance with access rights and intellectual property issues, and the development of a catalysis research ontology, however, are the responsibilities of the entire community, led by national and international initiatives for research data infrastructures. But there will not be any progress unless every researcher is engaged actively, from increasing IT literacy as a catalysis researcher, establishing a local initiative, and appointing a data steward/ess, up to mentoring the younger generation to accelerate the change. The paradigm shift we need to cross the bridge towards digital catalysis will only arise from our own involvement as a scientific community, which this article will have hopefully contributed to.

Acknowledgements

The design of the local data infrastructure of the Department of Inorganic Chemistry of the Fritz-Haber-Institut (FHI) der Max-Planck-Gesellschaft (MPG) in Berlin under the direction of Prof. Dr. Robert Schlögl is the result of a joint effort of all co-workers over many years in collaboration with the Computer Support Group of the FHI. Here, we would also like to thank BASF SE for sponsoring tasks within the framework of the BasCat project, through which this work was promoted. BasCat is a collaboration between Technical University Berlin, FHI, and the clusters of excellence "Unified Concepts in Catalysis"/"Unifying Systems in Catalysis" (UniCat https://www.unicat.tu-berlin.de/ UniSysCat, https://www.unisyscat.de). The Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy—EXC 2008-390540038-UniSysCat) is acknowledged. Currently, the automation is funded by the Federal Ministry of Education and Research (BMBF) in the framework of the CatLab project, FKZ 03EW0015B. We thank Abdulrhman Moshantaf from our group, and Michael Wesemann, Patrick Oppermann, and Heinz Junkes from the Computer Support Group of the FHI for the IT developments and continuous support of automation and database projects. The collaboration with the Humboldt University Berlin, funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation), in the framework of the project FAIRmat—FAIR Data Infrastructure for Condensed-Matter Physics and the Chemical Physics of Solids, project number 460197019, is also acknowledged. We thank Dr. Lucas Foppa, the NOMAD Laboratory at the FHI, and Michael Wesemann, Computer Support Group at the FHI, for the provision of images, and Maria Victoria Gatti for her help in designing the TOC image.

Conflict of Interest

The authors declare no conflict of interest.

Data Availability Statement

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

Disclaimer

The opinions expressed in this publication are the view of the authors and do not necessarily reflect the opinions or views of *Angewandte Chemie International Edition/Angewandte Chemie*, the Publisher, the GDCh, or the affiliated editors. The Publisher cannot be held responsible for the contents and contemporaneity of any external websites.

Keywords: Automation · Autonomous Catalyst Discovery · Clean Data · Digital Catalysis · Local Data Infrastructure

[1] N. Seedat, F. Imrie, M. van der Schaar, *arXiv* **2022**, https://doi.org/10.48550/arXiv.2211.05764.
[2] C. Draxl, M. Scheffler, *MRS Bull.* **2018**, *43*, 676.
[3] R. Schlögl, *ChemSusChem* **2010**, *3*, 209.
[4] M. Park, E. Leahey, R. J. Funk, *Nature* **2023**, *613*, 138.
[5] a) J. A. Esterhuizen, B. R. Goldsmith, S. Linic, *Nat. Catal.* **2022**, *5*, 175; b) J. R. Kitchin, *Nat. Catal.* **2018**, *1*, 230; c) H. Li, Y. Jiao, K. Davey, S.-Z. Qiao, *Angew. Chem. Int. Ed.* **2023**, *62*, e202216383; d) J. Peng, D. Schwalbe-Koda, K. Akkiraju, T. Xie, L. Giordano, Y. Yu, C. J. Eom, J. R. Lunger, D. J. Zheng, R. R. Rao, S. Muy, J. C. Grossman, K. Reuter, R. Gómez-Bombarelli, Y. Shao-Horn, *Nat. Rev. Mater.* **2022**, *7*, 991; e) T. Toyao, Z. Maeno, S. Takakusagi, T. Kamachi, I. Takigawa, K. Shimizu, *ACS Catal.* **2020**, *10*, 2260; f) T. Williams, K. McCullough, J. A. Lauterbach, *Chem. Mater.* **2020**, *32*, 157; g) P. Schlexer Lamoureux, K. T. Winther, J. A. Garrido Torres, V. Streibel, M. Zhao, M. Bajdich, F. Abild-Pedersen, T. Bligaard, *ChemCatChem* **2019**, *11*, 3581; h) K. Takahashi, J. Ohyama, S. Nishimura, J. Fujima, L. Takahashi, T. Uno, T. Taniike, *Chem. Commun.* **2023**, *59*, 2222; i) A. J. Medford, M. R. Kunz, S. M. Ewing, T. Borders, R. Fushimi, *ACS Catal.* **2018**, *8*, 7403.
[6] S. Herres-Pawlis, F. Bach, I. J. Bruno, S. J. Chalk, N. Jung, J. C. Liermann, L. R. McEwen, S. Neumann, C. Steinbeck, M. Razum, O. Koepler, *Angew. Chem. Int. Ed.* **2022**, *61*, e202203038.
[7] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, J. Bouwman, A. J. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C. T. Evelo, R. Finkers, A. Gonzalez-Beltran, A. J. G. Gray, P. Groth, C. Goble, J. S. Grethe, J. Heringa, P. A. C. 't Hoen, R. Hooft, T. Kuhn, R. Kok, J. Kok, S. J. Lusher, M. E. Martone, A. Mons, A. L. Packer, B. Persson, P. Rocca-Serra, M. Roos, R. van Schaik, S.-A. Sansone, E. Schultes, T. Sengstag, T. Slater, G. Strawn, M. A. Swertz, M. Thompson, J. van der Lei, E. van Mulligen, J. Velterop, A. Waagmeester, P. Wittenburg, K. Wolstencroft, J. Zhao, B. Mons, *Sci. Data* **2016**, *3*, 160018.
[8] a) P. S. F. Mendes, S. Siradze, L. Pirro, J. W. Thybaut, *ChemCatChem* **2021**, *13*, 836; b) A. Salazar, B. Wentzel, S. Schimmler, R. Gläser, S. Hanf, S. A. Schunk, *Chem. Eur. J.* **2023**, *29*, e202202720; c) P. Tremouilhac, C.-L. Lin, P.-C. Huang, Y.-C. Huang, A. Nguyen, N. Jung, F. Bach, R. Ulrich, B. Neumair, A. Streit, S. Bräse, *Angew. Chem. Int. Ed.* **2020**, *59*, 22771; d) C. Wulf, M. Beller, T. Boenisch, O. Deutschmann, S. Hanf, N. Kockmann, R. Kraehnert, M. Oezaslan, S. Palkovits, S. Schimmler, S. A. Schunk, K. Wagemann, D. Linke, *ChemCatChem* **2021**, *13*, 3223.
[9] "GitHub", can be found under https://github.com/ **2022** (accessed April 09, 2023).
[10] K. M. Jablonka, L. Patiny, B. Smit, *Nat. Chem.* **2022**, *14*, 365.
[11] a) S. M. Moosavi, A. Nandy, K. M. Jablonka, D. Ongari, J. P. Janet, P. G. Boyd, Y. Lee, B. Smit, H. J. Kulik, *Nat. Commun.* **2020**, *11*, 4068; b) Y. Luo, S. Bag, O. Zaremba, A. Cierpka, J. Andreo, S. Wuttke, P. Friederich, M. Tsotsalas, *Angew. Chem. Int. Ed.* **2022**, *61*, e202200242.
[12] M. Fernandez, P. G. Boyd, T. D. Daff, M. Z. Aghaji, T. K. Woo, *J. Phys. Chem. Lett.* **2014**, *5*, 3056.
[13] T. Lombardo, M. Duquesnoy, H. El-Bouysidy, F. Årén, A. Gallo-Bueno, P. B. Jørgensen, A. Bhowmik, A. Demortière, E. Ayerbe, F. Alcaide, M. Reynaud, J. Carrasco, A. Grimaud, C. Zhang, T. Vegge, P. Johansson, A. A. Franco, *Chem. Rev.* **2022**, *122*, 10899.
[14] K. Karako, Y. Chen, W. Tang, *BioSci. Trends* **2018**, *12*, 553.
[15] M. Scheffler, M. Aeschlimann, M. Albrecht, T. Bereau, H.-J. Bungartz, C. Felser, M. Greiner, A. Groß, C. T. Koch, K. Kremer, W. E. Nagel, M. Scheidgen, C. Wöll, C. Draxl, *Nature* **2022**, *604*, 635.
[16] E. Frei, A. Gaur, H. Lichtenberg, L. Zwiener, M. Scherzer, F. Girgsdies, T. Lunkenbein, R. Schlögl, *ChemCatChem* **2020**, *12*, 4029.
[17] T. Lunkenbein, J. Schumann, M. Behrens, R. Schlögl, M. G. Willinger, *Angew. Chem. Int. Ed.* **2015**, *54*, 4544.
[18] H. Tao, T. Wu, M. Aldeghi, T. C. Wu, A. Aspuru-Guzik, E. Kumacheva, *Nat. Rev. Mater.* **2021**, *6*, 701.
[19] T. Hisatomi, K. Domen, *Nat. Catal.* **2019**, *2*, 387.
[20] K. Reuter, C. P. Plaisance, H. Oberhofer, M. Andersen, *J. Chem. Phys.* **2017**, *146*, 040901.
[21] a) Z. W. Ulissi, M. T. Tang, J. Xiao, X. Liu, D. A. Torelli, M. Karamad, K. Cummins, C. Hahn, N. S. Lewis, T. F. Jaramillo, K. Chan, J. K. Nørskov, *ACS Catal.* **2017**, *7*, 6600; b) W. Xu, K. Reuter, M. Andersen, *Nat. Comput. Sci.* **2022**, *2*, 443.
[22] A. Bruix, J. T. Margraf, M. Andersen, K. Reuter, *Nat. Catal.* **2019**, *2*, 659.
[23] a) Y. Kanai, T. Watanabe, T. Fujitani, T. Uchijima, J. Nakamura, *Catal. Lett.* **1996**, *38*, 157; b) M. Behrens, F. Studt, I. Kasatkin, S. Kühl, M. Hävecker, F. Abild-Pedersen, S. Zander, F. Girgsdies, P. Kurr, B.-L. Kniep, M. Tovar, R. W. Fischer, J. K. Nørskov, R. Schlögl, *Science* **2012**, *336*, 893.
[24] J. Schumann, J. Kröhnert, E. Frei, R. Schlögl, A. Trunschke, *Top. Catal.* **2017**, *60*, 1735.
[25] S. Kuld, M. Thorhauge, H. Falsig, C. F. Elkjær, S. Helveg, I. Chorkendorff, J. Sehested, *Science* **2016**, *352*, 969.
[26] a) S. Derrouiche, H. Lauron-Pernot, C. Louis, *Chem. Mater.* **2012**, *24*, 2282; b) M. V. Twigg, M. S. Spencer, *Top. Catal.* **2003**, *22*, 191; c) A. Beck, M. A. Newton, M. Zabilskiy, P. Rzepka, M. G. Willinger, J. A. van Bokhoven, *Angew. Chem. Int. Ed.* **2022**, *61*, e202200301.
[27] a) T. Bligaard, R. M. Bullock, C. T. Campbell, J. G. Chen, B. C. Gates, R. J. Gorte, C. W. Jones, W. D. Jones, J. R. Kitchin, S. L. Scott, *ACS Catal.* **2016**, *6*, 2590; b) J. G. Chen, C. W. Jones, S. Linic, V. R. Stamenkovic, *ACS Catal.* **2017**, *7*, 6392; c) I. Spanos, A. A. Auer, S. Neugebauer, X. Deng, H. Tüysüz, R. Schlögl, *ACS Catal.* **2017**, *7*, 3768; d) S. L. Scott, *ACS Catal.* **2018**, *8*, 8597; e) S. Z. Andersen, V. Čolić, S. Yang, J. A. Schwalbe, A. C. Nielander, J. M. McEnaney, K. Enemark-Rasmussen, J. G. Baker, A. R. Singh, B. A. Rohr, M. J. Statt, S. J. Blair, S. Mezzavilla, J. Kibsgaard, P. C. K. Vesborg, M. Cargnello, S. F. Bent, T. F. Jaramillo, I. E. L. Stephens, J. K. Nørskov, I. Chorkendorff, *Nature* **2019**, *570*, 504; f) B. H. R. Suryanto, H.-L. Du, D. Wang, J. Chen, A. N.

Simonov, D. R. MacFarlane, *Nat. Catal.* **2019**, *2*, 290; g) A. Trunschke, G. Bellini, M. Boniface, S. J. Carey, J. Dong, E. Erdem, L. Foppa, W. Frandsen, M. Geske, L. M. Ghiringhelli, F. Girgsdies, R. Hanna, M. Hashagen, M. Hävecker, G. Huff, A. Knop-Gericke, G. Koch, P. Kraus, J. Kröhnert, P. Kube, S. Lohr, T. Lunkenbein, L. Masliuk, R. Naumann d'Alnoncourt, T. Omojola, C. Pratsch, S. Richter, C. Rohner, F. Rosowski, F. Rüther, M. Scheffler, R. Schlögl, A. Tarasov, D. Teschner, O. Timpe, P. Trunschke, Y. Wang, S. Wrabetz, *Top. Catal.* **2020**, *63*, 1683.

[28] a) S. Weber, R. T. Zimmermann, J. Bremer, K. L. Abel, D. Poppitz, N. Prinz, J. Ilsemann, S. Wendholt, Q. Yang, R. Pashminehazar, F. Monaco, P. Cloetens, X. Huang, C. Kübel, E. Kondratenko, M. Bauer, M. Bäumer, M. Zobel, R. Gläser, K. Sundmacher, T. L. Sheppard, *ChemCatChem* **2022**, *14*, e202101878; b) G. C. Bond, P. B. Wells, *Appl. Catal.* **1985**, *18*, 221; c) A. Trunschke, *Catal. Sci. Technol.* **2022**, *12*, 3650; d) A. Bhan, W. N. Delgass, *J. Catal.* **2022**, *405*, 419.

[29] G. Koch, M. Hävecker, D. Teschner, S. J. Carey, Y. Wang, P. Kube, W. Hetaba, T. Lunkenbein, G. Auffermann, O. Timpe, F. Rosowski, R. Schlögl, A. Trunschke, *ACS Catal.* **2020**, *10*, 7007.

[30] E. Soedarmadji, H. S. Stein, S. K. Suram, D. Guevarra, J. M. Gregoire, *npj Comput. Mater.* **2019**, *5*, 79.

[31] a) C. W. Coley, N. S. Eyke, K. F. Jensen, *Angew. Chem. Int. Ed.* **2020**, *59*, 22858; b) P. S. Gromski, A. B. Henson, J. M. Granda, L. Cronin, *Nat. Chem. Rev.* **2019**, *3*, 119; c) D. P. Tabor, L. M. Roch, S. K. Saikin, C. Kreisbeck, D. Sheberla, J. H. Montoya, S. Dwaraknath, M. Aykol, C. Ortiz, H. Tribukait, C. Amador-Bedolla, C. J. Brabec, B. Maruyama, K. A. Persson, A. Aspuru-Guzik, *Nat. Chem. Rev.* **2018**, *3*, 5; d) I. I. Ivanova, Y. G. Kolyagin, I. A. Kasyanov, A. V. Yakimov, T. O. Bok, D. N. Zarubin, *Angew. Chem. Int. Ed.* **2017**, *56*, 15344; e) S. A. Pelster, R. Kalamajka, W. Schrader, F. Schüth, *Angew. Chem. Int. Ed.* **2007**, *46*, 2299; f) M. Sanchez Sanchez, F. Girgsdies, M. Jastak, P. Kube, R. Schlögl, A. Trunschke, *Angew. Chem. Int. Ed.* **2012**, *51*, 7194; g) M. Behrens, D. Brennecke, F. Girgsdies, S. Kissner, A. Trunschke, N. Nasrudin, S. Zakaria, N. F. Idris, S. B. A. Hamid, B. Kniep, R. Fischer, W. Busser, M. Muhler, R. Schlögl, *Appl. Catal. A* **2011**, *392*, 93.

[32] a) A. Bubliauskas, D. J. Blair, H. Powell-Davies, P. J. Kitson, M. D. Burke, L. Cronin, *Angew. Chem. Int. Ed.* **2022**, *61*, e202116108; b) C. W. Coley, N. S. Eyke, K. F. Jensen, *Angew. Chem. Int. Ed.* **2020**, *59*, 23414.

[33] E. Stach, B. DeCost, A. G. Kusne, J. Hattrick-Simpers, K. A. Brown, K. G. Reyes, J. Schrier, S. Billinge, T. Buonassisi, I. Foster, C. P. Gomes, J. M. Gregoire, A. Mehta, J. Montoya, E. Olivetti, C. Park, E. Rotenberg, S. K. Saikin, S. Smullin, V. Stanev, B. Maruyama, *Matter* **2021**, *4*, 2702.

[34] a) L. Foppa, L. M. Ghiringhelli, F. Girgsdies, M. Hashagen, P. Kube, M. Hävecker, S. J. Carey, A. Tarasov, P. Kraus, F. Rosowski, R. Schlögl, A. Trunschke, M. Scheffler, *MRS Bull.* **2021**, *46*, 1016; b) L. Foppa, F. Rüther, M. Geske, G. Koch, F. Girgsdies, P. Kube, S. J. Carey, M. Hävecker, O. Timpe, A. V. Tarasov, M. Scheffler, F. Rosowski, R. Schlögl, A. Trunschke, *J. Am. Chem. Soc.* **2023**, *145*, 3427.

[35] A. Chakrabarti, M. E. Ford, D. Gregory, R. Hu, C. J. Keturakis, S. Lwin, Y. Tang, Z. Yang, M. Zhu, M. A. Bañares, I. E. Wachs, *Catal. Today* **2017**, *283*, 27.

[36] a) H. Bluhm, M. Hävecker, A. Knop-Gericke, M. Kiskinova, R. Schloegl, M. Salmeron, *MRS Bull.* **2007**, *32*, 1022; b) Y. W. Choi, H. Mistry, B. Roldan Cuenya, *Curr. Opin. Electrochem.* **2017**, *1*, 95.

[37] a) N. Marcella, Y. Liu, J. Timoshenko, E. Guan, M. Luneau, T. Shirman, A. M. Plonka, J. E. S. van der Hoeven, J. Aizenberg, C. M. Friend, A. I. Frenkel, *Phys. Chem. Chem. Phys.* **2020**, *22*, 18902; b) J.-D. Grunwaldt, C. G. Schroer, *Chem. Soc. Rev.* **2010**, *39*, 4741.

[38] A. M. Wernbacher, P. Kube, M. Hävecker, R. Schlögl, A. Trunschke, *J. Phys. Chem. C* **2019**, *123*, 13269.

[39] N. E. Tsakoumis, A. P. E. York, D. Chen, M. Rønning, *Catal. Sci. Technol.* **2015**, *5*, 4859.

[40] F. C. Meunier, *Chem. Soc. Rev.* **2010**, *39*, 4602.

[41] a) R. J. Berger, F. Kapteijn, J. A. Moulijn, G. B. Marin, J. De Wilde, M. Olea, D. Chen, A. Holmen, L. Lietti, E. Tronconi, Y. Schuurman, *Appl. Catal. A* **2008**, *342*, 3; b) F. Kapteijn, J. A. Moulijn, *Handbook of Heterogeneous Catalysis, Vol. 4*, 2nd ed. (Eds.: G. Ertl, H. Knözinger, F. Schüth, J. Weitkamp), Wiley-VCH, Weinheim, **2008**, pp. 2019–2045, https://onlinelibrary.wiley.com/doi/abs/10.1002/9783527610044.hetcat0108.

[42] a) R. J. Berger, E. H. Stitt, G. B. Marin, F. Kapteijn, J. A. Moulijn, *CATTECH* **2001**, *5*, 36; b) J. J. Birtill, *Ind. Eng. Chem. Res.* **2007**, *46*, 2392; c) U. I. Kramm, R. Marschall, M. Rose, *ChemCatChem* **2019**, *11*, 2563; d) G. B. Marin, V. V. Galvita, G. S. Yablonsky, *J. Catal.* **2021**, *404*, 745.

[43] B. W. J. Chen, L. Xu, M. Mavrikakis, *Chem. Rev.* **2021**, *121*, 1007.

[44] a) "NOMAD", can be found under https://nomad-lab.eu/ **2014** (accessed February 27, 2023)b) "Catalysis Hub", can be found under https://www.catalysis-hub.org/ **2019** (accessed February 27, 2023); c) "Materials Project", can be found under https://materialsproject.org/catalysis **2013** (accessed April 18, 2023); d) "Materials Cloud", can be found under www.materialscloud.org **2018** (accessed April 18, 2023).

[45] a) N. Artrith, K. T. Butler, F.-X. Coudert, S. Han, O. Isayev, A. Jain, A. Walsh, *Nat. Chem.* **2021**, *13*, 505; b) A. Y.-T. Wang, R. J. Murdock, S. K. Kauwe, A. O. Oliynyk, A. Gurlo, J. Brgoch, K. A. Persson, T. D. Sparks, *Chem. Mater.* **2020**, *32*, 4954.

[46] G. Pilania, *Comput. Mater. Sci.* **2021**, *193*, 110360.

[47] A. Tompos, M. Sanchez-Sanchez, L. Végvári, G. P. Szijjártó, J. L. Margitfalvi, A. Trunschke, R. Schlögl, K. Wanninger, G. Mestl, *Catal. Today* **2021**, *363*, 45.

[48] H. S. Stein, J. M. Gregoire, *Chem. Sci.* **2019**, *10*, 9640.

[49] "NFDI", can be found under https://www.nfdi.de/?lang=en **2021** (accessed February 27, 2023).

[50] a) "CADS", can be found under https://cads.eng.hokudai.ac.jp/ **2018** (accessed February 27, 2023); b) J. Fujima, Y. Tanaka, I. Miyazato, L. Takahashi, K. Takahashi, *React. Chem. Eng.* **2020**, *5*, 903.

[51] "Kinetics Database", can be found under https://kinetics.nist.gov/kinetics/index.jsp **2021** (accessed February 27, 2023).

[52] "SCIENCE DATA BANK", can be found under https://www.scidb.cn/en **2015** (accessed February 27, 2023).

[53] "Cambridge Structural Database (CSD)", can be found under https://www.ccdc.cam.ac.uk/structures/, (accessed February 27, 2023).

[54] "SwissCAT+", can be found under https://swisscatplus.ch/ **2022** (accessed February 27, 2023).

[55] "zenodo", can be found under https://zenodo.org/ **2013** (accessed April 19, 2023).

[56] "Open Research Europe Data Guidelines", can be found under https://open-research-europe.ec.europa.eu/for-authors/data-guidelines#generaldata **2022** (accessed December 14, 2022).

[57] a) C. Bo, F. Maseras, N. López, *Nat. Catal.* **2018**, *1*, 809; b) K. T. Winther, M. J. Hoffmann, J. R. Boes, O. Mamun, M. Bajdich, T. Bligaard, *Sci. Data* **2019**, *6*, 75.

[58] a) L. Takahashi, I. Miyazato, K. Takahashi, *J. Chem. Inf. Model.* **2018**, *58*, 1742; b) F. Strieth-Kalthoff, F. Sandfort, M. Kühnemund, F. R. Schäfer, H. Kuchen, F. Glorius, *Angew. Chem. Int. Ed.* **2022**, *61*, e202204647; c) T. Taniike, K.

Takahashi, *Nat. Catal.* **2023**, *6*, 108; d) P. Raccuglia, K. C. Elbert, P. D. F. Adler, C. Falk, M. B. Wenny, A. Mollo, M. Zeller, S. A. Friedler, J. Schrier, A. J. Norquist, *Nature* **2016**, *533*, 73.

[59] M. Vennewald, A. Iemhoff, D. Ditz, R. Palkovits, *Catal. Sci. Technol.* **2022**, *12*, 1741.

[60] I. Lucentini, X. Garcia, X. Vendrell, J. Llorca, *Ind. Eng. Chem. Res.* **2021**, *60*, 18560.

[61] "Inorganic Crystal Structure Database ", can be found under https://icsd.fiz-karlsruhe.de, **2023** (accessed January 5, 2023).

[62] M. Thommes, K. Kaneko, A. V. Neimark, J. P. Olivier, F. Rodriguez-Reinoso, J. Rouquerol, K. S. W. Sing, *Pure Appl. Chem.* **2015**, *87*, 1051.

[63] "Crystallographic Information Framework", can be found under https://iucr.org/resources/cif, (accessed April 09, 2023).

[64] "NeXus Format", can be found under https://nexusformat.org/ **2023** (accessed April 09, 2023).

[65] J. Macht, R. T. Carr, E. Iglesia, *J. Catal.* **2009**, *264*, 54.

[66] T. R. Gruber, *Int. J. Man-Mach. Stud.* **1995**, *43*, 907.

[67] L. Wilbraham, S. H. M. Mehr, L. Cronin, *Acc. Chem. Res.* **2021**, *54*, 253.

[68] A. S. Behr, M. Völkenrath, N. Kockmann, *Research Square* **2023**, https://doi.org/10.21203/rs.3.rs-2457909/v1.

[69] a) F. Schüth, M. D. Ward, J. M. Buriak, *Chem. Mater.* **2018**, *30*, 3599; b) S. Moniri, T. Van Cleve, S. Linic, *J. Catal.* **2017**, *345*, 1.

[70] a) Z. Zhang, B. Zandkarimi, A. N. Alexandrova, *Acc. Chem. Res.* **2020**, *53*, 447; b) R. Schlögl, *Angew. Chem. Int. Ed.* **2015**, *54*, 3465.

[71] a) L. E. Y. Nonneman, A. G. T. M. Bastein, V. Ponec, R. Burch, *Appl. Catal.* **1990**, *62*, L23; b) N. Yang, A. J. Medford, X. Liu, F. Studt, T. Bligaard, S. F. Bent, J. K. Nørskov, *J. Am. Chem. Soc.* **2016**, *138*, 3705; c) F. Stavale, X. Shao, N. Nilius, H.-J. Freund, S. Prada, L. Giordano, G. Pacchioni, *J. Am. Chem. Soc.* **2012**, *134*, 11380.

[72] a) A. Sápi, T. Rajkumar, J. Kiss, Á. Kukovecz, Z. Kónya, G. A. Somorjai, *Catal. Lett.* **2021**, *151*, 2153; b) P. Mäki-Arvela, D. Y. Murzin, *Appl. Catal. A* **2013**, *451*, 251.

[73] R. Price, M. Cassidy, J. G. Grolig, G. Longo, U. Weissen, A. Mai, J. T. S. Irvine, *Adv. Energy Mater.* **2021**, *11*, 2003951.

[74] a) S. K. Wilkinson, M. J. H. Simmons, E. H. Stitt, X. Baucherel, M. J. Watson, *J. Catal.* **2013**, *299*, 249; b) M. Armbrüster, M. Behrens, F. Cinquini, K. Föttinger, Y. Grin, A. Haghofer, B. Klötzer, A. Knop-Gericke, H. Lorenz, A. Ota, S. Penner, J. Prinz, C. Rameshan, Z. Révay, D. Rosenthal, G. Rupprechter, P. Sautet, R. Schlögl, L. Shao, L. Szentmiklósi, D. Teschner, D. Torres, R. Wagner, R. Widmer, G. Wowsnick, *ChemCatChem* **2012**, *4*, 1048.

[75] S. J. Tauster, *Acc. Chem. Res.* **1987**, *20*, 389.

[76] a) D. E. De Vos, S. Ernst, C. Perego, C. T. O'Connor, M. Stöcker, *Microporous Mesoporous Mater.* **2002**, *56*, 185; b) C. Wei, R. R. Rao, J. Peng, B. Huang, I. E. L. Stephens, M. Risch, Z. J. Xu, Y. Shao-Horn, *Adv. Mater.* **2019**, *31*, 1806296.

[77] M. A. Bañares, *Catal. Today* **2005**, *100*, 71.

[78] U. Zavyalova, M. Holena, R. Schlögl, M. Baerns, *ChemCatChem* **2011**, *3*, 1935.

[79] Y. Zhang, C. Wang, M. Soukaseum, D. G. Vlachos, H. Fang, *J. Chem. Inf. Model.* **2022**, *62*, 3316.

[80] a) E. Kim, K. Huang, A. Tomala, S. Matthews, E. Strubell, A. Saunders, A. McCallum, E. Olivetti, *Sci. Data* **2017**, *4*, 170127; b) V. Tshitoyan, J. Dagdelen, L. Weston, A. Dunn, Z. Rong, O. Kononova, K. A. Persson, G. Ceder, A. Jain, *Nature* **2019**, *571*, 95.

[81] a) K. E. Lamb, M. D. Dolan, D. F. Kennedy, *Int. J. Hydrogen Energy* **2019**, *44*, 3580; b) S. Ristig, M. Poschmann, J. Folke, O. Gómez-Cápiro, Z. Chen, N. Sanchez-Bastardo, R. Schlögl, S. Heumann, H. Ruland, *Chem. Ing. Tech.* **2022**, *94*, 1413.

[82] Y. Gambo, S. Adamu, A. A. Abdulrasheed, R. A. Lucky, M. S. Ba-Shammakh, M. M. Hossain, *Appl. Catal. A* **2021**, *609*, 117914.

[83] a) K. Takahashi, I. Miyazato, S. Nishimura, J. Ohyama, *ChemCatChem* **2018**, *10*, 3223; b) S. Mine, M. Takao, T. Yamaguchi, T. Toyao, Z. Maeno, S. M. A. Hakim Siddiki, S. Takakusagi, K.-i. Shimizu, I. Takigawa, *ChemCatChem* **2021**, *13*, 3636.

[84] A. Nieva de la Hidalga, J. Goodall, C. Anyika, B. Matthews, C. R. A. Catlow, *Catal. Commun.* **2022**, *162*, 106384.

[85] S. Palkovits, *ChemCatChem* **2020**, *12*, 3995.

[86] a) M. K. Tuck, D. W. Chan, D. Chia, A. K. Godwin, W. E. Grizzle, K. E. Krueger, W. Rom, M. Sanda, L. Sorbara, S. Stass, W. Wang, D. E. Brenner, *J. Proteome Res.* **2009**, *8*, 113; b) S. R. Mager, M. H. A. Oomen, M. M. Morente, C. Ratcliffe, K. Knox, D. J. Kerr, F. Pezzella, P. H. J. Riegman, *Eur. J. Cancer* **2007**, *43*, 828.

[87] "Guidelines for Safeguarding Good Research Practice—Code of Conduct", can be found under https://zenodo.org/record/6472827 **2022** (accessed February 27, 2023).

[88] a) S. Davis, J. Usansky, S. Mitra-Kaushik, J. Kellie, K. Honrine, E. Woolf, J. Adams, R. Kelly, J. Evens, S. Pine, H. Hochreiner, M. Dawes, J. Kentner, S. Crawford, *Bioanalysis* **2021**, *13*, 1313; b) H. Alosert, J. Savery, J. Rheaume, M. Cheeks, R. Turner, C. Spencer, S. S. Farid, S. Goldrick, *Biotechnol. J.* **2022**, *17*, 2100609.

[89] a) P. Mehra, G. Minhas, W. Costa Pereira, *Quality Assurance Implementation in Research Labs* (Eds.: A. Anand), Springer Singapore, Singapore, **2021**, pp. 45–62, https://doi.org/10.1007/978-981-16-3074-3_4; b) S. Hollmann, M. Frohme, C. Endrullat, A. Kremer, D. D'Elia, B. Regierer, A. Nechyporenko, *PLoS Comput. Biol.* **2020**, *16*, e1008095.

[90] "fhimpg/archive ", can be found under https://github.com/fhimpg/archive **2022** (accessed February 27, 2023).

[91] "Lindau Guidelines", can be found under https://lindauguidelines.org/ **2023** (accessed February 27, 2023).

[92] a) H. K. Machina, D. J. Wild, *J. Lab. Autom.* **2013**, *18*, 264; b) R. Kwok, S. Kanza, *Nature* **2018**, *560*, 269; c) S. Guerrero, A. López-Cortés, J. M. García-Cárdenas, P. Saa, A. Indacochea, I. Armendáriz-Castillo, A. K. Zambrano, V. Yumiceba, A. Pérez-Villa, P. Guevara-Ramírez, O. Moscoso-Zea, J. Paredes, P. E. Leone, C. Paz-Y-miño, *PLoS Comput. Biol.* **2019**, *15*, e1006918; d) D. Bromfield Lee, *J. Chem. Educ.* **2018**, *95*, 1102; e) P. Tremouilhac, A. Nguyen, Y. C. Huang, S. Kotov, D. S. Lütjohann, F. Hübsch, N. Jung, S. Bräse, *J. Cheminf.* **2017**, *9*, 54.

[93] H. Gossler, J. Riedel, E. Daymo, R. Chacko, S. Angeli, O. Deutschmann, *Chem. Ing. Tech.* **2022**, *94*, 1798.

[94] a) "EPICS in the Max-Planck-Society", can be found under https://epics.mpg.de/index.php?*n*=Main.HomePage?userlang=en **2019** (accessed February 27, 2023); b) "Haber - Catalytic test reactor for ammonia decomposition", can be found under https://gitlab.fhi.mpg.de/fhi-ac/haber **2022** (accessed February 27, 2023); c) "Ertl - Catalytic test reactor for CO oxidation", can be found under https://gitlab.fhi.mpg.de/fhi-ac/ertl **2022** (accessed February 27, 2023).

[95] P. Kube, J. Dong, N. S. Bastardo, H. Ruland, R. Schlögl, J. T. Margraf, K. Reuter, A. Trunschke, *Nat. Commun.* **2022**, *13*, 7504.

[96] "Automatic file transformation and data upload in electron microscopy sessions", can be found under https://gitlab.fhi.mpg.de/fhi-ac/velox **2021** (accessed February 27, 2023).

[97] R. C. Martin, *Clean code: A handbook of agile software craftsmanship*, Prentice Hall, Upper Saddle River, **2009**.

[98] S. P. Gilroy, B. A. Kaplan, *Perspect. Behav. Sci.* **2019**, *42*, 565.

[99] P. Jupyter, M. Bussonnier, J. Forde, J. Freeman, B. Granger, T. Head, C. Holdgraf, K. Kelley, G. Nalvarte, A. Osheroff, M.

Pacer, Y. Panda, F. Perez, B. Ragan-Kelley, C. Willing, *Proceedings of the 17th Python in Science Conference (SCIPY 2018)* **2018**, pp. 113–120, https://conference.scipy.org/proceedings/scipy2018/project_jupyter.html.

[100] P. Kraus, E. H. Wolf, C. Prinz, G. Bellini, A. Trunschke, R. Schlögl, *Digital Discovery* **2022**, *1*, 241.

[101] "Reference Material", can be found under https://webshop.bam.de/webshop_en/reference-material.html **2022** (accessed December 13, 2022).

[102] D. A. M. Monti, A. Baiker, *J. Catal.* **1983**, *83*, 323.

[103] "The Turing Way: A handbook for reproducible, ethical and collaborative research", can be found under https://zenodo.org/record/6909298 **2022** (accessed February 27, 2023).

Manuscript received: February 27, 2023
Version of record online: May 31, 2023