

RESEARCH ARTICLE | SEPTEMBER 30 2025

## A research database for experimental electrocatalysis: Advancing data sharing and reusability

Special Collection: [Data Science for Catalysis](#)

Ruchika Mahajan ; Ashton M. Aleman ; Colin F. Crago ; Suman Bhasker-Ranganath ;  
Melissa E. Kreider ; Jose A. Zamora Zeledon ; Johanna Schröder ; Gaurav A. Kamat;  
McKenzie A. Hubert ; Adam C. Nielander ; Thomas F. Jaramillo; Michaela Burke Stevens  ;  
Johannes Voss  ; Kirsten T. Winther  



*J. Chem. Phys.* 163, 124704 (2025)

<https://doi.org/10.1063/5.0280821>



### Articles You May Be Interested In

How to extract kinetic information from Tafel analysis in electrocatalysis

*J. Chem. Phys.* (December 2023)

Epitaxial oxide thin films for oxygen electrocatalysis: A tutorial review

*J. Vac. Sci. Technol. A* (November 2021)

The surface states of transition metal X-ides under electrocatalytic conditions

*J. Chem. Phys.* (March 2023)

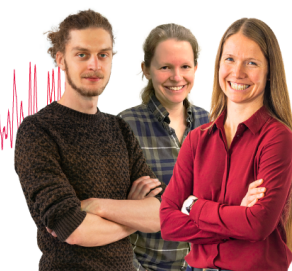
### Webinar From Noise to Knowledge

May 13th – Register now



Zurich  
Instruments

Universität  
Konstanz










# A research database for experimental electrocatalysis: Advancing data sharing and reusability

Cite as: J. Chem. Phys. 163, 124704 (2025); doi: 10.1063/5.0280821

Submitted: 14 May 2025 • Accepted: 20 August 2025 •

Published Online: 30 September 2025



Ruchika Mahajan,<sup>1,2</sup>  Ashton M. Aleman,<sup>1,2</sup>  Colin F. Crago,<sup>1,2</sup>  Suman Bhasker-Ranganath,<sup>1,2</sup>   
Melissa E. Kreider,<sup>1,2</sup>  Jose A. Zamora Zeledon,<sup>1</sup>  Johanna Schröder,<sup>1,2</sup>  Gaurav A. Kamat,<sup>1,2</sup>  
McKenzie A. Hubert,<sup>1,2</sup>  Adam C. Nielander,<sup>2</sup>  Thomas F. Jaramillo,<sup>1,2</sup>  Michaela Burke Stevens,<sup>2,a)</sup>   
Johannes Voss,<sup>2,a)</sup>  and Kirsten T. Winther<sup>2,a)</sup> 

## AFFILIATIONS

<sup>1</sup> SUNCAT Center for Interface Science and Catalysis, Department of Chemical Engineering, Stanford University, Stanford, California 94305, USA

<sup>2</sup> SUNCAT Center for Interface Science and Catalysis, SLAC National Accelerator Laboratory, Menlo Park, California 94025, USA

**Note:** This paper is part of the JCP Special Topic, Data Science for Catalysis.

**a)** Authors to whom correspondence should be addressed: [mburkes@slac.stanford.edu](mailto:mburkes@slac.stanford.edu); [vossj@slac.stanford.edu](mailto:vossj@slac.stanford.edu); and [winther@slac.stanford.edu](mailto:winther@slac.stanford.edu)

## ABSTRACT

The availability of high-fidelity catalysis data is essential for training machine learning models to advance catalyst discovery. Furthermore, the sharing of data is crucial to ensure the comparability of scientific results. In electrocatalysis, where complex experimental conditions and measurement uncertainties pose unique challenges, structured data collection and sharing are critical to improving reproducibility and enabling robust model development. Addressing these challenges requires standardized approaches to data collection, metadata inclusion, and accessibility. To support this effort, we have developed an extensive data infrastructure that curates and organizes multimodal data from electrocatalysis experiments, making them openly available through the catalysis-hub.org platform. Our datasets, comprising 241 experimental entries, provide detailed information on reaction conditions, material properties, and performance metrics, ensuring transparency and interoperability. By structuring electrocatalysis data in web-based as well as machine-readable formats, we aim to bridge the gap between experimental and computational research, allowing for improved benchmarking and predictive modeling. This work highlights the importance of well-structured, accessible data in overcoming reproducibility challenges and advancing machine learning applications in catalysis. The framework we present lays the foundation for future data-driven research in electrocatalysis and offers a scalable model for other experimental disciplines.

Published under an exclusive license by AIP Publishing. <https://doi.org/10.1063/5.0280821>

## I. INTRODUCTION

The discovery of high-performing catalysts is needed to improve the energy efficiency and price competitiveness of electrochemical energy technologies, including water splitting for clean H<sub>2</sub> production, electrochemical CO<sub>2</sub> reduction toward carbon-based and liquid fuels, electrochemical N<sub>2</sub> reduction to ammonia, and power generation in electrochemical fuel cells. To advance these technologies, stable, active, and selective catalyst materials must be discovered and tested in conjunction with the electrochemical cell, which includes the catalyst material, catalyst matrix and

conductive support, electrolyte, polymeric binder/ionomer, promoters, and proton/anion exchange membranes. Therefore, optimizing electrocatalytic performance is a highly complex task with a vast parameter space. In this regard, there is an opportunity for data-driven methodologies to accelerate the discovery of not only optimized catalyst material composition but also electrocatalytic cell configuration and operation parameters. However, artificial intelligence (AI) methodologies are relatively under-utilized in experimental heterogeneous catalysis,<sup>1</sup> where a significant limitation for the successful application of AI is the availability of consistent, well-structured, and reliable experimental datasets.

Over the past decades, the field of materials science has seen a tremendous advance in the generation, curation, storage, and distribution of open materials data. Important repositories include the Inorganic Crystal Structure Database (ICSD),<sup>2</sup> the Cambridge Structural Database (CSD),<sup>3</sup> and the International Center for Diffraction Data (ICDD),<sup>4</sup> providing crystallographic refinement data for inorganic and organic compounds, which serve as valuable reference data for material characterization. In parallel, computational materials repositories have been developed, with an overall focus on crystallographic, i.e., periodically representable, materials, including the Materials Project,<sup>5</sup> the Open Quantum Materials Database (OQMD),<sup>6</sup> Materials Cloud,<sup>7</sup> and more.<sup>8</sup> The increasing availability of open and structured data has also aided the development of ML-based methodology for material property prediction<sup>9</sup> and discovery,<sup>10</sup> enabling the application of deep learning that requires a larger amount of training (1000+ data points as a guideline).

While these databases have steadily grown to support complex and large-scale datasets, they often focus on domain-specific or well-structured data types, particularly computational data for crystalline materials. At the same time, broader efforts have been made to integrate diverse materials data into unified platforms. One such example is the Materials Data Facility (MDF), which was introduced as a flexible and scalable infrastructure to support the publication, discovery, and reuse of diverse materials data.<sup>11</sup> However, MDF has faced challenges in achieving accelerated growth in data submissions over time. This may be due to its broader scope, the complexity of its metadata requirements, and limited engagement from the user community. These limitations suggest that offering a generalized, all-purpose platform may not be sufficient on its own. Considering these challenges, we believe successful data structures are more likely to gain wider adoption and long-term utility when they are tailored to the specific needs of their intended user communities. Learning from these experiences is essential in designing future platforms that can achieve lasting impact and broader adoption.

The field of catalysis has also seen significant advances in the development of computational catalysis data repositories, with a focus on storing surface slab geometries and calculated properties.<sup>12–15</sup> For example, adsorption and reaction energy datasets on catalysis-hub.org have aided in the development of ML models for surface stability and reactivity, including adsorption energy predictions<sup>16,17</sup> serving as surrogate models for DFT computations. More recently, the large collection of training data on the Open Catalyst Project database has enabled the training of ML interatomic potentials to accelerate surface slab simulations, as well as the training of deep-learning models for adsorption energy surrogates.<sup>14,15</sup> In terms of experimental catalysis, open datasets have been limited to smaller collections produced by individual groups or larger, consistent datasets produced through a simplified protocol in a high-throughput manner. For example, datasets such as CatTestHub<sup>18</sup> for heterogeneous catalysis serve as valuable open-access repositories for benchmarking gas-phase thermal catalytic reactions such as methanol dehydrogenation, formic acid decomposition, and Hofmann elimination. Another significant effort by Senocrate *et al.*<sup>19</sup> involved developing a comprehensive measurement system for electrochemical CO<sub>2</sub> reduction by integrating commercial analytical instruments and sensors with open

source software, supporting standardized, high throughput data collection. More recently, Open Catalyst Experiments 2024 (OCx24) by Abed *et al.*<sup>20</sup> released one of the largest experimental collections, comprising 572 synthesized samples and 441 gas diffusion electrodes tested for the CO<sub>2</sub> reduction and hydrogen evolution reactions. There are also several efforts focused on experimental data in materials science. For example, the Materials Provenance Store (MPS)<sup>21</sup> supports heterogeneous and distributed data from experimental workflows; MaterialsAtlas.org<sup>22</sup> integrates both experimental and computational materials data; and the High-Throughput Experimental Materials Database (HTEM-DB)<sup>23</sup> collects data from large-scale combinatorial materials experiments. While the database platforms discussed above mark a significant step forward, the data are primarily shared in CSV-based formats, without a programmatic interface and with limited access to raw measurements, which could restrict broader reuse and integration.

In parallel, recent advances in self-driven laboratories have begun to address challenges in data availability and reproducibility by enabling the generation of large, systematic, and high-quality datasets in chemistry and materials science.<sup>24</sup> Combined high-throughput synthesis and characterization work has even been used to screen over 300 000 material combinations in a single study.<sup>25</sup> Mining data from published peer-reviewed papers is another route taken by researchers seeking to acquire sufficient data to train machine learning models, with the difference between experimental setups and procedures as well as a lack of experimental details being a large concern for data consistency. Deep language models have furthermore been applied to extract insights from research papers;<sup>26–29</sup> however, the ability of such models to accurately compare information across datasets produced at different locations and with different setups is unclear. Recent platforms such as the Digital Catalysis Platform (DigCat)<sup>30</sup> collect data from published papers and combine it with models and AI tools. This helps with visualization, comparison, and interactive questions to support reliable and faster catalyst discovery. However, because DigCat uses data from different studies, it can have inconsistencies, differences in experiments, and missing details that may affect model training and reproducibility.

To advance the development of AI methodology in the field of catalysis, we need to increase the availability of high-quality datasets, preferably through data repositories that adhere to best practices for data curation and sharing. Such best practices have been provided by the GO FAIR initiative, with the recommendation that data should be Findable, Accessible, Interoperable, and Reusable (FAIR).<sup>31</sup> Findable data are generally registered or indexed in a searchable resource, are assigned unique identifiers such as permanent URLs, and contain additional metadata about the context and quality of the data. Accessibility can be ensured by making data retrievable under a standardized protocol, such as an HTTP or FTP request, for example, through a website. Interoperability means that a formal and broadly applicable language is used for knowledge representation, such as having meaningful keywords to represent data as well as making sure that the data are machine readable. Reusability of data largely comes down to ensuring sufficient metadata and provenance for users to understand how the data were produced and the history of the data, as well as the ability to tell if the data will be useful in another context.

All these practices must be improved for experimental catalysis data sharing. Reusability is particularly important in the context of ML model training and for ensuring the reproducibility of scientific research. Here, a major concern is to provide enough metadata and granularity to ensure that results can be understood and reproduced. In this work, we present an open research database for electrocatalysis that allows for the comparison of a diverse set of experimental setups and testing protocols covering key reactions such as the oxygen reduction reaction (ORR), oxygen evolution reaction (OER), and hydrogen evolution reaction (HER). Our platform is designed to follow the FAIR principles, with a strong focus on standardizing experimental electrocatalysis data formats and making them accessible through an intuitive, web-based interface. Users can explore and compare data directly on the website or access it programmatically via our publicly available GraphQL API (<https://api.catalysis-hub.org>) as well as the CatHub Python API (<https://github.com/SUNCAT-Center/CatHub>), supporting both manual exploration and automated workflows. It supports scalable, standardized data access through interactive visualizations, cross-study comparisons, and programmatic APIs—enabling deeper analysis and integration with machine learning workflows. A detailed description of the data structure, API features, and visualization tools follows in Sec. II.

## II. RESULTS

In this work, we report an open research database for electrocatalysis, made publicly available under <https://experimental.catalysis-hub.org/>. An additional outcome of this

work is the development of a data collection and storage framework that can capture the complexity of electrocatalytic materials and testing procedures. In this framework, we collect metrics related to the composition, morphology, testing, and catalytic performance of electrocatalysts, considering all steps of the experimental workflow as described in Sec. II A. Importantly, we are collecting and storing a wide selection of data and metadata important for data reproducibility and reusability. This includes characterization spectra (XRD, XPS, and XAS) as well as electrochemical testing (I–V curves), used to assess catalytic performance and stability.

### A. Experimental workflow

The three major thrusts of an experimental workflow in electrocatalysis research are synthesis, characterization, and catalyst testing (illustrated in Fig. 1). Synthesis involves designing or following a procedure to create a catalyst material of interest. Characterization involves using techniques (often spectroscopy, microscopy, and/or crystallography) to understand the composition, structure, and morphology of the synthesized material. Catalyst testing involves applying a protocol relevant to catalytic operating conditions to evaluate the performance of a catalyst. Finally, post-characterization is done after catalytic testing to identify changes to the material that happened during electrocatalysis, as catalyst materials are often chemically and structurally dynamic.<sup>32</sup> Each step in the experimental workflow generates unique data metrics that contribute to a holistic understanding of a catalyst material.

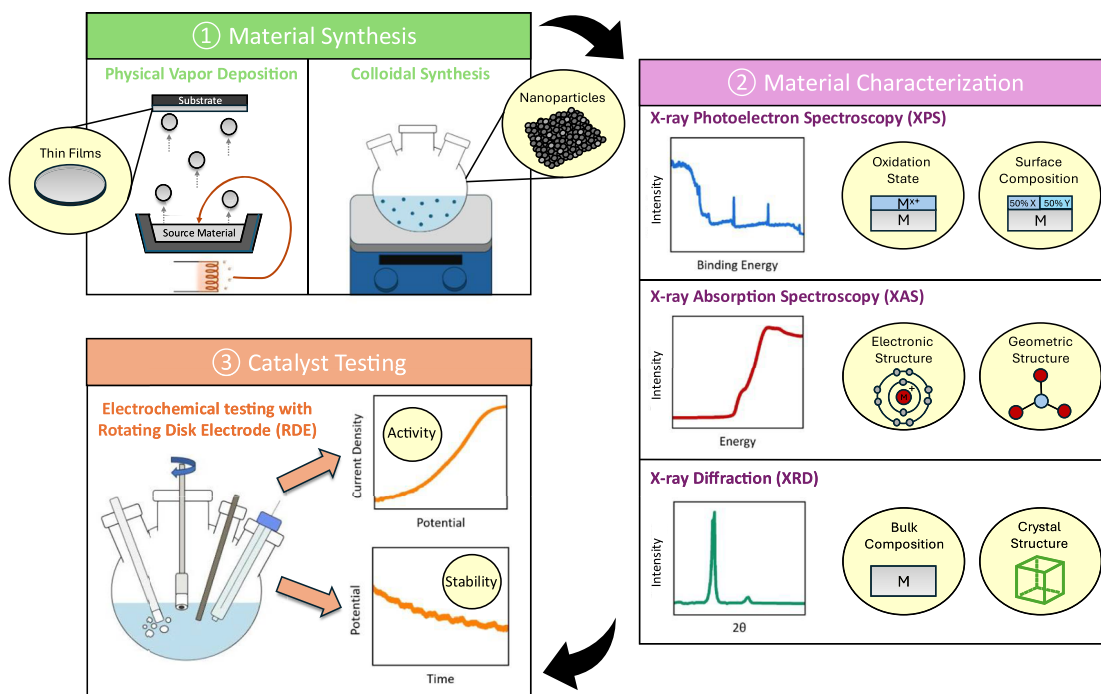


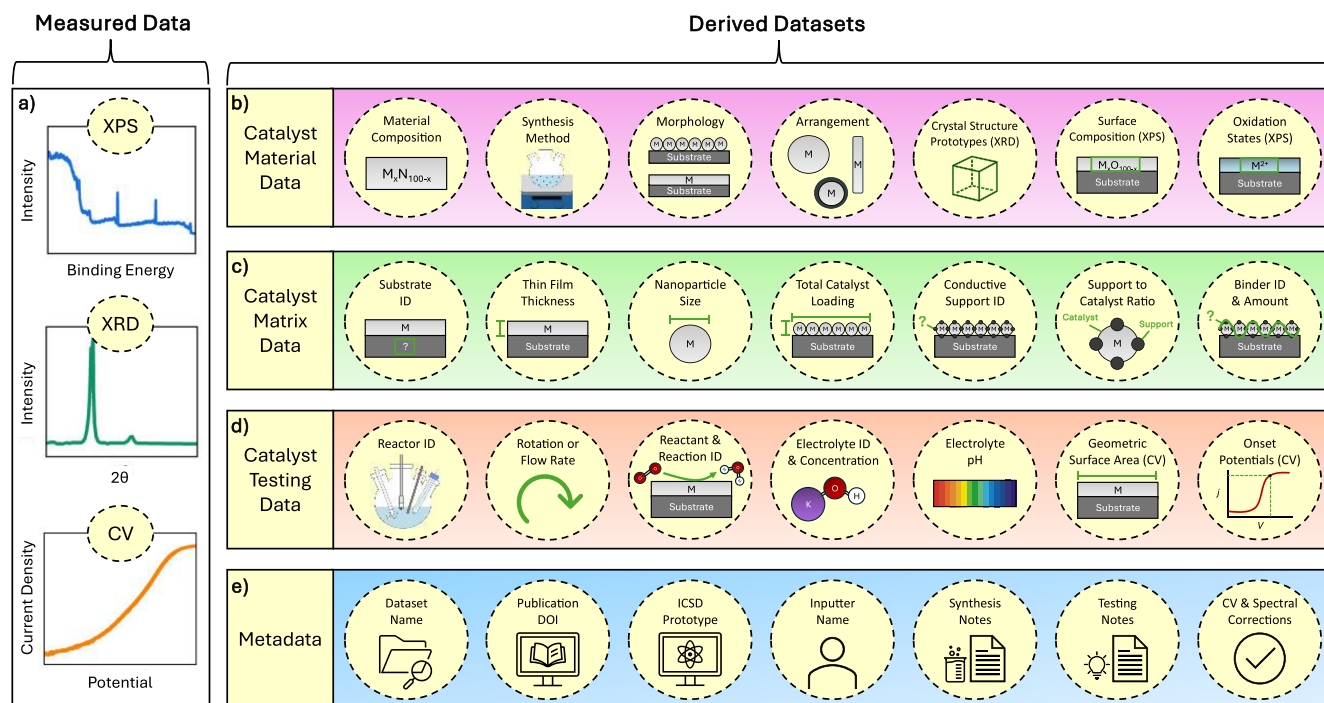
FIG. 1. Three major thrusts of an experimental workflow for electrocatalysis, including (1) material synthesis, (2) material characterization, and (3) catalyst testing.

Synthesis methods are chosen depending on the aimed type and morphology of the catalyst. For more fundamental studies, thin film catalysts are desirable that intend to minimize the chemical and structural complexity of the synthesized materials, and thus they are more analogous to conventional DFT slab models. Thin films can be fabricated, e.g., with vacuum deposition techniques such as thermal or electron beam evaporation, physical vapor deposition,<sup>33</sup> or reactive sputtering.<sup>34</sup> Our key database input parameters for thin films are the substrate, the targeted deposition rate, film thickness, and geometric surface area of the sample. Powder or nanoparticle catalysts, on the other hand, are more often used in commercial energy conversion devices and are desirable for more applied studies, as they have a higher surface area and, thus, an increased number of available active sites. Powder catalysts can be synthesized from a wide array of wet chemical and thermal processes, e.g., our group often uses a colloidal approach. Important database input parameters for powders regarding the catalyst matrix are the substrate, the type of conductive support (called henceforth conductive support ID), the catalyst-support ratio, the catalyst loading, the binder ID and amount, and the geometric surface area of the sample.

The three main characterization methods in our workflow are x-ray photoelectron spectroscopy (XPS), x-ray absorption spectroscopy (XAS), and x-ray diffraction (XRD). XPS survey spectra provide insights into surface composition, and high-resolution spectra can be fitted to resolve surface chemical states and oxidation states.<sup>35</sup> For XPS, we collect intensity and binding energy arrays

for both survey and element-specific high-resolution scans, along with other necessary parameters and outputs such as the C 1s peak shift for calibration. Derived data include the fitted atomic surface composition and oxidation states. XAS spectra provide insights into bulk chemical composition and structure.<sup>36</sup> For XAS, we collect normalized intensity vs energy data into our database, along with the specified transition (i.e., K-edge or L-edge). Diffractograms from XRD can identify the bulk crystal structure of a catalyst material. For XRD, we store the raw intensity vs 2 theta (the angle between the incident and diffracted beam) in our database, as well as results derived from our analysis, such as the corresponding ICSD ID,<sup>2</sup> space group, and lattice parameter. As mentioned previously, catalyst characterization performed after electrochemical testing can contribute valuable insight to catalyst restructuring. Although we are not currently including post-testing characterization data in our uploaded data, the functionality is accommodated in our data structure and will be included in future data uploads.

For electrochemical characterization, we use a standard rotating disk electrode (RDE) system to screen the catalytic activity with cyclic voltammetry (CV), i.e., electrode potential is swept in a catalytically relevant window, and the current density response is measured. We also evaluate electrochemical stability by applying a constant current density/potential and measuring the resulting potential/current density over a longer period of time,<sup>37</sup> followed by another CV to reevaluate activity after the stability test. For each of these electrochemical experiments (pre-CV, stability test, and post-CV), time, potential, and current density columns are used as input



**FIG. 2.** (a) Data structure for experimental catalysis data, including measured data in the form of XPS, XRD, and CV curves. (b)–(d) Derived data and tabulated metrics for the catalyst material, matrix, and testing, respectively. (e) Metadata that provide additional context to the data.



into our database. Furthermore, we extract and tabulate the onset potential at a range of current densities from the CV curves, together with the maximum mass-transfer limited current density. We also track important experimental parameters such as electrolyte ID and concentration, sparged gas, rotation rate, and reaction of interest (e.g., oxygen reduction or oxygen evolution).

## B. Data structure

In correspondence with the collection workflow described above, we have developed a data structure specific to electrocatalysis, including tabulated properties of the *catalyst material*, *catalyst matrix*, and *catalyst testing*, as well as *metadata*, as illustrated in the schematic in Fig. 2. Higher modality data are collected for XRD and XPS characterization spectra as well as CV curves in the form of arrays. The choice of catalyst material serves as an entry point for our data structure, where the chemical composition, synthesis method, and morphology are the main entries. We store the assumed chemical composition as targeted by the synthesis together with metrics extracted and interpreted from characterization spectra. From the XRD, we derive the crystal space group obtained from a comparison to ICSD crystal structure prototypes. In the case of signatures of more than one crystal symmetry in the XRD diffractogram, we provide a list of space groups and prototype IDs. From the XPS, we extract the measured oxidation state and surface composition for transition metal elements, which will generally differ from the synthesis target composition. To facilitate data reusability, we are storing post-processed spectra in separate XRD and XPS tables with links to the catalyst material ID. For the XPS, we store the survey spectra (giving an overview of peaks over a larger energy range), as well as high resolution XPS peaks in a narrower energy range. We also provide the C 1s shift that was used to correct the XPS peak positions, which are particularly relevant for reproducibility purposes.

Moving beyond the impact of the unique catalyst material, there is significant complexity when considering the full catalyst matrix, as illustrated in Fig. 2(c). Important catalyst matrix parameters include catalyst thin film thickness, substrate, conductive support, binder/ionomers, and the total catalyst loading, as well as the ratio between the different components. These parameters can all have a significant effect on catalyst performance and should be tabulated to foster data reusability. Furthermore, the prospect of optimizing catalyst matrix parameters with ML methods further motivates data collection in this space. For example, our recent work on developing interpretable ML models for the performance of binary transition metal antimonates found the catalyst loading and the conductive support ratio to be key parameters for catalyst performance.<sup>38</sup> For the catalysis testing data, collected parameters include properties of the reactor, the electrolyte, and the reactants, as shown in Fig. 2(d). Furthermore, catalytic activity metrics extracted from the CV curves are tabulated, including the mass-transfer-limited current density and the onset potential for a range of current densities.

To add context to the collected datasets, we also assign additional metadata. Each data point is assigned to a dataset name (generally chosen based on the chemical composition of the catalyst studied), as well as the name of the person responsible for adding the data point to the table. If a data entry is part of a publication, we link

it to the publication DOI. In addition, we link the ICSD crystal structure prototype IDs where applicable (where the prototype ID refers to the crystal structure, while the composition of the ICSD reference will generally differ). The data collection was carried out by manually collecting, curating, and labeling entries that were then added to a shared spreadsheet. We used one main table for tabulated metrics and separate tables for spectra and CV curves, carefully linking the data in different tables through unique identifiers.

## C. Datasets and availability

To facilitate the access and visualization of our catalysis data collection, we have developed the Catalysis-hub Experimental (CatHubExp) database, a web-based tool available at <https://experimental.catalysis-hub.org>. The goal is to provide a comprehensive repository that researchers can explore, analyze, and use to further their research. Currently, 13 datasets are made publicly available and can be accessed on the *Publications* page. Here, data are browsed on a dataset basis, with each dataset corresponding to a distinct journal publication with an assigned DOI and corresponding link. Data were collected with a focus on catalyst discovery through the study of high-performing catalyst compositions and morphologies. Therefore, the catalytic reaction, synthesis method, material class, and catalyst matrix generally vary across datasets. Chemical reactions currently include oxygen reduction reaction (ORR), oxygen evolution reaction (OER), and hydrogen evolution reaction (HER), with 195, 43, and 03 entries, respectively. A diverse collection of catalyst materials is sampled, including thin film metals and alloys of Ir–Pt,<sup>39</sup> Ir–X (X = Sn, Cr, Ti, Ni),<sup>40</sup> Ag–Pd,<sup>33</sup> and Ag–Cu systems;<sup>41</sup> transition metal antimonates;<sup>42</sup> carbon nanotubes;<sup>43</sup> metal–organic frameworks;<sup>44</sup> and Ru-based pyrochlores,<sup>45</sup> phosphides,<sup>46</sup> and nitrides.<sup>34</sup>

Consistent with the Catalysis-hub computational data collection, each dataset is assigned a permanent identifier constructed from aggregating the name of the first author, the title of the paper, and the publication year. Clicking a selected publication ID links to a separate page with detailed dataset information about synthesized samples, their properties (e.g., chemical, structural, and physical), experimental methods, different reaction types, material morphologies, their electrocatalytic performance, and related publications, among other relevant data, as discussed in Sec. II B.

In addition to the *Publications* page entry point, data can be browsed through the *Catalyst Material* page (both linked via buttons on the main page), which enables users to filter data based on catalyst composition and visualize data entries across publications. Researchers can use an interactive periodic table for element-based and formula-based searches to explore materials. After a search, a material table appears as shown in Fig. 3, displaying key properties such as composition, DOI, morphology, reaction, pH, and onset potential as default visible columns. Users can customize the table view by selecting additional columns through the “Manage columns” option directly within the table. This feature also provides advanced filtering options to refine data by functionalities, e.g., search for any keywords across the table, direct DOI links connected to source publications, and a CSV “Export Table” that allows users to download filtered datasets. In addition, users can sort the material table columns or search the data using various filters, such as material properties, reaction types, onset potential,

## Material/Sample Search

## Explore Catalyst Materials

Search Material Data by Chemical Composition or Properties.

Materials -Mn

All Reactions

73 materials match your search (3 selected)

Search across all colur

Export Table

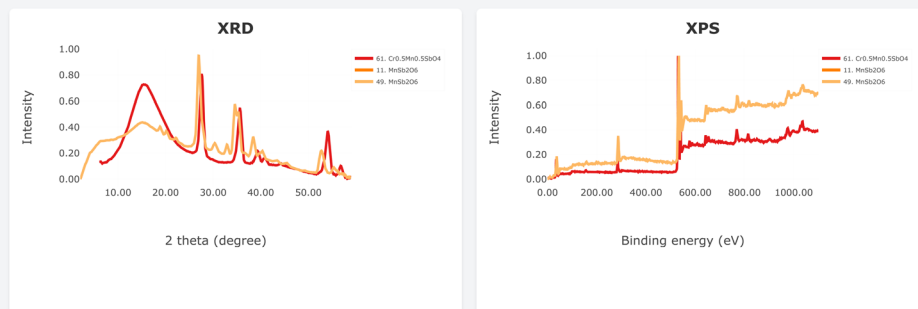
SN	Composition	DOI	Morphology	Reaction	Onset poten (V vs. RHE at +/-100 $\mu\text{A}/\text{cm}^2$ )	Onset potential (V vs. RHE at +/-1000 $\mu\text{A}/\text{cm}^2$ )	pH	Replicate	Onset potential (V vs. RHE at +/-0.1 $\text{mA}/\text{cm}^2$ )
<input type="checkbox"/> 31	Cr <sub>0.5</sub> Mn <sub>0.5</sub> SbO <sub>4</sub>	<a href="#">DOI: 10.1002/cphc.202400010</a>	nanoparticle	ORR	0.86	-	13	3	0.857 +/- 0.001
<input type="checkbox"/> 18	MnSb <sub>2</sub> O <sub>6</sub>	<a href="#">DOI: 10.1002/cphc.202400010</a>	nanoparticle	ORR	0.85	-	13	3	0.851 +/- 0.006
<input checked="" type="checkbox"/> 49	MnSb <sub>2</sub> O <sub>6</sub>	<a href="#">DOI: 10.1002/cphc.202400010</a>	nanoparticle	ORR	0.85	-	13	-	-
<input type="checkbox"/> 16	MnSb <sub>2</sub> O <sub>6</sub>	<a href="#">DOI: 10.1002/cphc.202400010</a>	nanoparticle	ORR	0.84	-	13	1	0.851 +/- 0.006
<input type="checkbox"/> 20	MnSb <sub>2</sub> O <sub>6</sub>	<a href="#">DOI: 10.1002/cphc.202400010</a>	nanoparticle	ORR	0.84	-	13	2	0.863 +/- 0.013
<input type="checkbox"/> 39	MnSb <sub>2</sub> O <sub>6</sub>	<a href="#">DOI: 10.1021/acsnano.3c0042</a>	nanoparticle	ORR	0.84	-	13	4	0.823 +/- 0.01
<input checked="" type="checkbox"/> 61	Cr <sub>0.5</sub> Mn <sub>0.5</sub> SbO <sub>4</sub>	<a href="#">DOI: 10.1002/cphc.202400010</a>	nanoparticle	ORR	0.84	-	13	-	-
<input type="checkbox"/> 35	Fe <sub>0.5</sub> Mn <sub>0.5</sub> Sb <sub>2</sub> O <sub>6</sub>	<a href="#">DOI: 10.1002/cphc.202400010</a>	nanoparticle	ORR	0.83	-	13	1	0.823 +/- 0.005
<input type="checkbox"/> 47	MnSb <sub>2</sub> O <sub>6</sub>	<a href="#">DOI: 10.1002/cphc.202400010</a>	nanoparticle	ORR	0.83	-	13	-	-
<input type="checkbox"/> 3	MnSb <sub>2</sub> O <sub>6</sub>	<a href="#">DOI: 10.1002/cphc.202400010</a>	nanoparticle	ORR	0.82	-	13	3	0.776 +/- 0.034
<input checked="" type="checkbox"/> 11	MnSb <sub>2</sub> O <sub>6</sub>	<a href="#">DOI: 10.1021/acsnano.3c0042</a>	nanoparticle	ORR	0.82	-	13	2	0.823 +/- 0.01

3 rows selected

Rows per page: 501-50 of 69

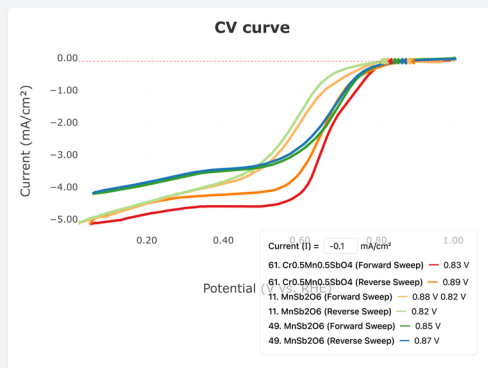
## Characterization spectra

X-ray diffraction (XRD) and photoelectron spectroscopy (XPS) results are shown below.



## Electrochemical testing

Cyclic voltammetry (CV) curves are shown below. Select the current density value in the legend below to re-calculate the onset potentials.



**FIG. 3.** Advanced catalyst materials search: Researchers can select materials with checkboxes to analyze their electrochemical properties. Here, all Mn-based materials were selected, filtered by ORR reaction type, and further refined by the highest onset potential. Based on the selection, only the corresponding XRD, XPS, and CV graphs are displayed, facilitating detailed data analysis.

etc. A subset of the data entries corresponds to repeated electrochemical measurements for already existing catalyst morphologies and testing parameters. This is indicated by the “Replicate” column, where entries are assigned an integer number if one or more repeated measurements exist. For these entries, we report the averaged onset potentials as well as error bars (standard deviations) in a separate column, shown as MEAN+/-STD. Including such statistics is valuable for assessing the data uncertainty and quality. They improve transparency, provide a precise measure of variability, and help assess the reliability of the data.

Finally, researchers are invited to contribute data to foster database scalability and collaboration. As will be mentioned in the following, our developed data structure and storage framework is flexible enough to accommodate additional columns for derived tabulated data. In addition, the data structure allows for empty/null values of columns, since some metrics might not be relevant for all catalyst systems. Moreover, to promote transparency and long-term data reliability, if a data entry is later found to be flawed either by contributors or users, it will not be removed from the database. Instead, such entries will be retained and clearly marked with a label and clarifying note in the metadata to indicate the issue. This policy ensures reproducibility, allows users to make informed decisions about data quality, and helps in building trust in the database over the long term. In Sec. II D, we will briefly describe the system architecture and implementation details of the underlying developed system and demonstrate the key features of CatHubExp.

#### D. System architecture

The CatHubExp database platform consists of several entities that handle different stages of the data flow, as illustrated in Fig. 4. First, data are manually collected, curated, and organized in a spreadsheet-based tabular workspace, using standardized tables with well-defined column names corresponding to our developed data structure shown in Fig. 2. Different data modalities, such as key-value pairs, characterization spectra, and CV curves, are stored in separate tables and linked through unique material and sample identifiers.

Data are subsequently transferred to a cloud database, using a customized Python workflow to parse data from regular tables to a structured query language (SQL) format. The structured data format efficiently manages relationships between catalyst material,

testing samples, CV, and characterization tables and has enhanced capabilities for data retrieval through tailored queries. With a PostgreSQL database backend, we obtain a good compromise between structure and flexibility, utilizing the JSONB format to accommodate an unlimited number of key value pairs that enable additional table columns in the future.

Next, a Python web API is used to create the GraphQL endpoint (<https://api.catalysis-hub.org>), now providing access to experimental as well as computational Catalysis-hub data. The application enables flexible querying of complex data structures, efficient data fetching by selecting specific fields, and simplified frontend integration. It also provides a machine-readable data interface, making it accessible in a JSON-like format through an HTTP request. Experimental data are found in *publicationExp*, *materials*, *samples*, *echemical*, *xps*, and *xrd* tables. For instance, materials can be queried by publication (pubId), as illustrated in Fig. 5.

Finally, the front-end application (<https://experimental.catalysis-hub.org>) serves as the main user interface. It is architected using modern web technologies with React.js, a JavaScript framework. It is a component-based architecture that allows the application to efficiently manage dynamic user interfaces (UI) by reusing components and maintaining a virtual document object model (DOM) for fast rendering. This modular approach enhances scalability and simplifies development, making it easier to extend functionality in the future. The frontend system uses the Material-UI library for data display and management and the Apollo GraphQL client for data fetching from the backend services. Important features include interactive data visualization, search by materials and publications, and filtering capabilities. The application also displays experimental data through customizable graphs and tables and provides data export functionality for enhanced usability.

#### E. Data analysis

In this section, we will briefly discuss the interactive tools, allowing researchers to select data entries for further analysis, perform real time calculations, and visualize results through dynamic graphs and filtering options.

The platform provides high-quality plots for XRD and XPS material characterization spectra, which offer deeper insights when multiple materials are selected simultaneously, as shown in Fig. 3.

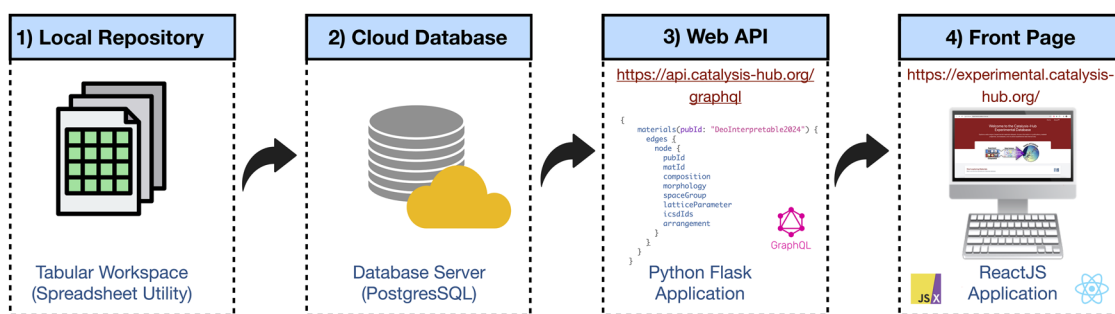
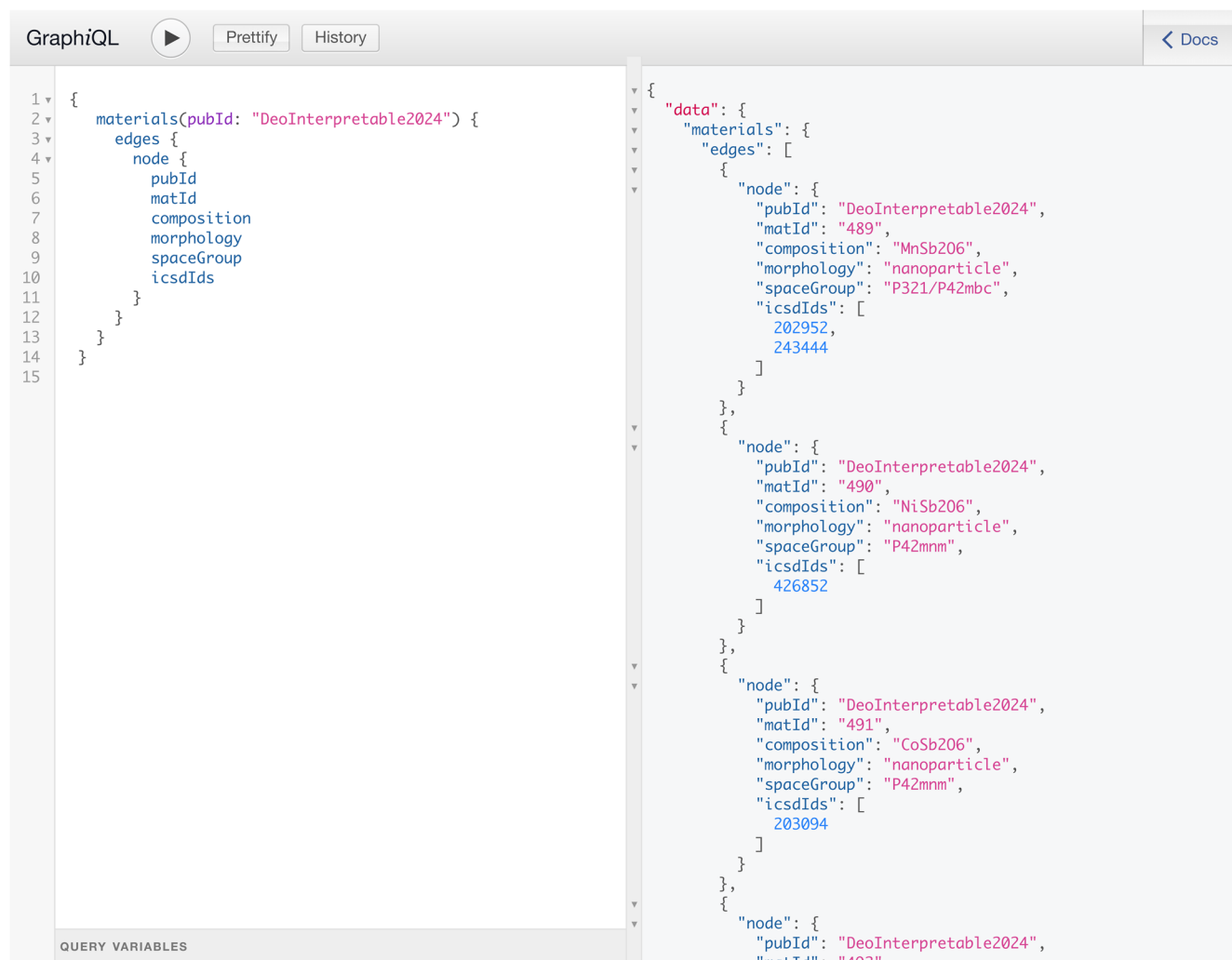


FIG. 4. Schematic representation of the flow of data, illustrating the connections between the database server, backend, and frontend applications.





The screenshot displays the GraphQL web interface. On the left, a query is written in a code editor with line numbers 1 through 15. The query is a nested structure: a root object with a 'materials' field, which is a function call 'materials(pubId: "DeoInterpretable2024")' containing an 'edges' field, which is an array of 'node' objects. Each 'node' object contains fields: 'pubId', 'matId', 'composition', 'morphology', 'spaceGroup', and 'icsdIds'. The right side of the interface shows the JSON response, which is a 'data' object containing a 'materials' field. This field is an array of 'edges', each containing a 'node' object with the same fields as specified in the query. The response shows three nodes with different 'matId' values (489, 490, 491) and different 'icsdIds' arrays. The interface also includes buttons for 'Prettify', 'History', and 'Docs'.

```
1 {
2   materials(pubId: "DeoInterpretable2024") {
3     edges {
4       node {
5         pubId
6         matId
7         composition
8         morphology
9         spaceGroup
10        icsdIds
11      }
12    }
13  }
14 }
15
```

```
{
  "data": {
    "materials": {
      "edges": [
        {
          "node": {
            "pubId": "DeoInterpretable2024",
            "matId": "489",
            "composition": "MnSb206",
            "morphology": "nanoparticle",
            "spaceGroup": "P321/P42mbc",
            "icsdIds": [
              202952,
              243444
            ]
          }
        },
        {
          "node": {
            "pubId": "DeoInterpretable2024",
            "matId": "490",
            "composition": "NiSb206",
            "morphology": "nanoparticle",
            "spaceGroup": "P42mm",
            "icsdIds": [
              426852
            ]
          }
        },
        {
          "node": {
            "pubId": "DeoInterpretable2024",
            "matId": "491",
            "composition": "CoSb206",
            "morphology": "nanoparticle",
            "spaceGroup": "P42mm",
            "icsdIds": [
              203094
            ]
          }
        }
      ]
    }
  }
}
```

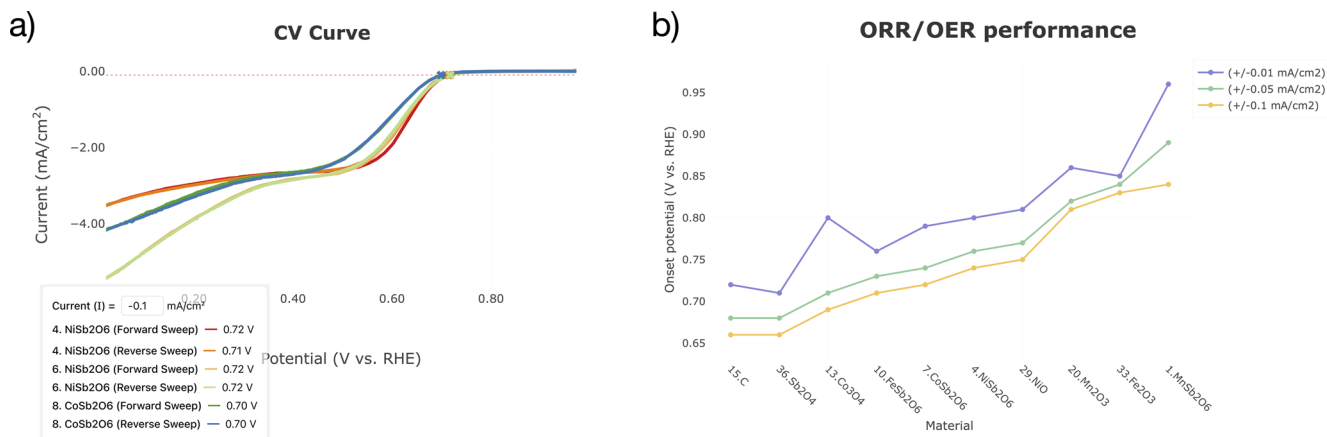
**FIG. 5.** Example of a GraphQL query for a materials search by publication ID executed in the API web interface. The web API can be accessed at <https://api.catalysis-hub.org/graphql>.

We note that spectral intensities are normalized to allow for easier comparison. In addition, the platform supports the visualization and straightforward analysis of cyclic voltammetry (CV) curves, including both forward and reverse sweeps. Researchers can select the current density value from the CV graph's legend to recalculate the onset potentials. As shown in Fig. 6(a), a current density of  $-0.1 \text{ mA/cm}^2$  is selected by default, and the graph provides a real time updated onset potential. Furthermore, comparing multiple graphs across different materials facilitates in-depth data analysis.

Another key feature allows users to visualize ORR/OER performance for a selected publication across different onset potentials. As shown in Fig. 6(b), the graph displays onset potentials

at three different current densities ( $+/-0.01$ ,  $+/-0.05$ , and  $+/-0.1 \text{ mA/cm}^2$ ) for various materials within the selected publication, enabling a comprehensive comparison of electrochemical performance.

Moreover, one can see the x-axis and y-axis data points by hovering over any graph, which makes the visualizations easier to understand and interact with. As an added capability, all graphs can be downloaded as high-quality PNG images, and the platform includes features such as zoom in, zoom out, autoscaling, and advanced filtering options to refine the visual analysis. Overall, the platform provides a range of tools for analyzing and comparing data, making it easier for researchers to explore and interpret their results.



**FIG. 6.** (a) Cyclic Voltammetry (CV) curve with forward and reverse sweeps. Users can select a current density value from the legend to recalculate the onset potential. Here, an example is shown for  $-0.1 \text{ mA/cm}^2$ . (b) Catalyst performance visualization for electrochemical analysis, shown here for the ORR onset potential.

### III. DISCUSSION

One of the overarching goals of this work is to improve data availability in electrocatalysis to leverage ML-guided catalyst design, where surrogate models are applied to identify and explore catalyst-reaction spaces of high interest. Generating and curating training data from experiments is resource-intensive and remains the major bottleneck for applying ML and deep learning to experimental catalysis. Although literature mining and combining data from different experimental setups can help expand datasets, ensuring their consistency remains a significant hurdle. Consequently, models trained on limited and incompatible data are prone to making inaccurate predictions, particularly when extrapolating to unknown spaces. For example, ML models trained on features based on catalyst composition are unlikely to perform well across datasets, particularly when the catalyst matrix structure and testing conditions impact performance. Moreover, uncertainty measures such as error bars can be used in machine learning models to help them learn better and make more reliable predictions.

Thus, a main deliverable of this work is to improve reproducibility and comparability in electrocatalysis research, creating a platform for publishing datasets with sufficient (meta)data to compare results from different sources, including high-performing and low-performing systems. The inclusion of catalyst characterization spectra and CV curves serves as an important means to improve the reproducibility and comparability of data and the ability to build more generalizable models.

Through the inclusion of catalyst characterization information in our database, we also enable a close connection to computational modeling, for example, by integrating experimental features with DFT-based features in ML models (e.g., from high-throughput computations). While atomic features capturing periodic trends, thermodynamics, and physical properties are often inexplicit, descriptors from theory, such as structural, energetic, vibrational, and electronic properties derived from DFT, provide higher fidelity in describing catalytic reactivity. Here, important computational metrics include surface adsorption and reaction

energies as well as atomic structures, such as those stored on our computational database counterpart at [www.catalysis-hub.org](http://www.catalysis-hub.org). In addition, phase diagrams and Pourbaix stability predictions are highly relevant for electrocatalysis, where the pH- and voltage-dependent restructuring and phase changes can be predicted with computation.

In the absence of experimental data, ML models trained on DFT-derived descriptors enable high-throughput screening of chemical spaces to guide experiments and further refine predictions based on feedback. However, integration of experimental descriptors into the ML model training can significantly enhance the power for the prediction of realistic catalytic performance.

Our study on binary transition metal antimony oxides highlights the value of incorporating experimental descriptors into ML model training.<sup>38</sup> By integrating atomic, theoretical, and experimental descriptors with human-interpretable models to predict the ORR onset potential, catalyst loading and conductive support-to-catalyst ratio emerged as key experimental features. Electronegativity and metal-oxygen bond length were identified among atomic and theoretical descriptors. Despite being trained on limited experimental data combined with bulk descriptors, ML models effectively capture trends in electrochemical performance.

Descriptors computed from surface slab models representing thin-film morphologies and nanoparticle facets will offer a more realistic depiction of active sites. Incorporating surface descriptors is expected to improve predictive accuracy. In this regard, the experimental database can further enhance ML models by

1. Using crystal structure(s) and high-symmetry facets derived from XRD to inform the model surface structures used for computed surface descriptors.
2. Use XPS and XAS insights to further refine the computational structural models, considering oxidation states, binding modes and adsorption sites of chemical intermediates, and catalyst-support interactions.
3. Integrating data from non-static models generated through molecular dynamics (MD) simulations with experimental

post-characterization data to capture catalyst structure evolution under reaction conditions.

ML models can also be trained to identify correlations between spectral features and catalyst structure-activity relationships, using the data of experimental spectra alone or in combination with the data of simulated spectra from high throughput computations.<sup>47</sup> These models enable automated spectral interpretation, streamlining the reverse engineering of promising catalysts through spectral inversion techniques.

While our data collection is relatively small (considering the volume of data needed to train ML models), we hope that our data collection framework, data structures, and ontologies can serve as inspiration for other researchers. Furthermore, we are continuously expanding our database with new datasets generated in the SUNCAT group, and datasets will generally be available after submission or acceptance of the corresponding publications. While our current data collection is highly focused on oxygen-based reactions (OER/ORR), a future goal is to expand the structure to carbon and nitrogen chemistries, such as CO<sub>2</sub> reduction to C1 and C2 products (CO<sub>2</sub>R) as well as nitrogen and nitrate reduction (N<sub>2</sub>R, NO<sub>3</sub>R). For these reactions, an additional complexity lies in optimizing the selectivity of the catalyst with respect to the desired products.<sup>48</sup> Therefore, a future goal is to expand our data structure to include tabulated selectivity metrics. Moreover, establishing a centralized repository to systematically organize all relevant experimental data and metrics for a specific reaction ensures adherence to standardized protocols and minimizes the risk of false positives, such as those highlighted in Ref. 49 for the electrochemical reduction of nitrogen to ammonia. We encourage the broader experimental community to share their data, even if it follows different formats. The idea is that, as long as metadata are available, it is feasible to build a centralized database either via APIs or using AI tools to integrate and analyze such diverse datasets. A successful example is our computational counterpart of Catalysis-hub (CatHub), which focuses on computational data and has grown steadily over time. We hope to enable a similar trajectory for experimental data in catalysis. An additional goal for our future data collection is the inclusion of post-testing characterization spectra. While this functionality is already accommodated in the data structure, it has not been prioritized in the data collection effort so far.

#### IV. CONCLUSION

In our database, we are advancing the standards of data collection, storage, and sharing for experimental electrocatalysis through the open web-based platform at <https://experimental.catalysis-hub.org>. Recognizing the importance of making experimental datasets accessible to the broader scientific community for reproducibility and machine learning model development, we systematically identify and collect data on key features relevant to catalyst synthesis, characterization, and testing. Currently, our database hosts electrocatalysis data for widely studied reactions such as the oxygen evolution reaction and oxygen reduction reaction. The platform catalogs unique entries from individual experiments, including details on the catalyst material, matrix, testing conditions, and metadata. In addition, a key feature of the platform is its ability to store multimodal information from various measurement

techniques, including XRD, XPS, and CV curves. With a simple query, users can seamlessly access and import datasets spanning different material classes and publications directly from the web interface. We hope that our open database and ontology for experimental electrocatalysis will serve as inspiration for other researchers and facilitate a broader sharing of data in the community.

#### ACKNOWLEDGMENTS

This research was supported by the U.S. Department of Energy, Office of Science, Office of Basic Energy Sciences, Chemical Sciences, Geosciences, and Biosciences Division, Catalysis Science Program to the SUNCAT Center for Interface Science and Catalysis. This research used resources of the National Energy Research Scientific Computing Center (NERSC), a Department of Energy User Facility using NERSC award BES-ERCAP0032527.

#### AUTHOR DECLARATIONS

##### Conflict of Interest

The authors have no conflicts to disclose.

##### Author Contributions

**Ruchika Mahajan:** Methodology (equal); Software (equal); Visualization (equal); Writing – original draft (equal). **Ashton M. Aleman:** Data curation (equal); Visualization (equal); Writing – review & editing (equal). **Colin F. Crago:** Data curation (equal); Writing – original draft (equal). **Suman Bhasker-Ranganath:** Writing – original draft (equal); Writing – review & editing (equal). **Melissa E. Kreider:** Data curation (equal); Methodology (equal). **Jose A. Zamora Zeledon:** Data curation (equal). **Johanna Schröder:** Data curation (equal); Writing – review & editing (equal). **Gaurav A. Kamat:** Data curation (equal). **McKenzie A. Hubert:** Data curation (equal). **Adam C. Nielander:** Data curation (supporting); Writing – review & editing (supporting). **Thomas F. Jaramillo:** Conceptualization (equal); Funding acquisition (equal). **Michaela Burke Stevens:** Conceptualization (equal); Data curation (equal); Methodology (equal); Supervision (equal). **Johannes Voss:** Conceptualization (equal); Supervision (equal). **Kirsten T. Winther:** Conceptualization (equal); Methodology (equal); Supervision (equal); Writing – original draft (equal).

#### DATA AVAILABILITY

Collected data can be browsed and downloaded from our web platform at <https://experimental.catalysis-hub.org> and the GraphQL API at <https://api.catalysis-hub.org>, as well as programmatically via the CatHub Python API at <https://github.com/SUNCAT-Center/CatHub>.

#### REFERENCES

- 1 M. Suvarna and J. Pérez-Ramírez, “Embracing data science in catalysis research,” *Nat. Catal.* 7(6), 624–635 (2024).
- 2 D. Zagorac, H. Müller, S. Ruehl, J. Zagorac, and S. Rehme, “Recent developments in the inorganic crystal structure database: Theoretical crystal structure data and related features,” *J. Appl. Crystallogr.* 52(5), 918–925 (2019).

- <sup>3</sup>C. R. Groom, I. J. Bruno, M. P. Lightfoot, and S. C. Ward, "The Cambridge structural database," *Acta Crystallogr., Sect. B: Struct. Sci., Cryst. Eng. Mater.* **72**(2), 171–179 (2016).
- <sup>4</sup>S. N. Kabekkodu, A. Dosen, and T. N. Blanton, "PDF-5+ : A comprehensive Powder Diffraction File™ for materials characterization," *Powder Diff.* **39**(2), 47–59 (2024).
- <sup>5</sup>A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, and K. A. Persson, "Commentary: The materials project: A materials genome approach to accelerating materials innovation," *APL Mater.* **1**(1), 011002 (2013).
- <sup>6</sup>J. E. Saal, S. Kirklin, M. Aykol, B. Meredig, and C. Wolverton, "Materials design and discovery with high-throughput density functional theory: The open quantum materials database (OQMD)," *JOM* **65**(11), 1501–1509 (2013).
- <sup>7</sup>L. Talirz, S. Kumbhar, E. Passaro, A. V. Yakutovich, V. Granata, F. Gargiulo, M. Borelli, M. Uhrin, S. P. Huber, S. Zoupanos, C. S. Adorf, C. W. Andersen, O. Schütt, C. A. Pignedoli, D. Passerone, J. VandeVondele, T. C. Schulthess, B. Smit, G. Pizzi, and N. Marzari, "Materials cloud, a platform for open computational science," *Sci. Data* **7**(1), 299 (2020).
- <sup>8</sup>M. Esters, C. Oses, S. Divilov, H. Eckert, R. Friedrich, D. Hicks, M. J. Mehl, F. Rose, A. Smolyanyuk, A. Calzolari, X. Campilongo, C. Toher, and S. Curtarolo, "aflow.org: A web ecosystem of databases, software and tools," *Comput. Mater. Sci.* **216**, 111808 (2023).
- <sup>9</sup>K. Choudhary, B. DeCost, C. Chen, A. Jain, F. Tavazza, R. Cohn, C. W. Park, A. Choudhary, A. Agrawal, S. J. L. Billinge, E. Holm, S. P. Ong, and C. Wolverton, "Recent advances and applications of deep learning methods in materials science," *npj Comput. Mater.* **8**(1), 59 (2022).
- <sup>10</sup>C. Zeni, R. Pinsler, D. Zügner, A. Fowler, M. Horton, X. Fu, S. Shysheya, J. Crabbé, L. Sun, J. Smith, B. Nguyen, H. Schulz, S. Lewis, C.-W. Huang, Z. Lu, Y. Zhou, H. Yang, H. Hao, J. Li, R. Tomioka, and T. Xie, "MatterGen: A generative model for inorganic materials design," *arXiv:2312.03687* (2023).
- <sup>11</sup>B. Blaiszik, K. Chard, J. Pruney, R. Ananthakrishnan, S. Tuecke, and I. Foster, "The materials data facility: Data services to advance materials science research," *JOM* **68**(8), 2045–2052 (2016).
- <sup>12</sup>M. Álvarez-Moreno, C. De Graaf, N. López, F. Maseras, J. M. Poblet, and C. Bo, "Managing the computational chemistry big data problem: The ioChem-BD platform," *J. Chem. Inf. Model.* **55**(1), 95–103 (2015).
- <sup>13</sup>K. T. Winther, M. J. Hoffmann, J. R. Boes, O. Mamun, M. Bajdich, and T. Bligaard, "Catalysis-Hub.org, an open electronic structure database for surface reactions," *Sci. Data* **6**(1), 75 (2019).
- <sup>14</sup>L. Chanussot, A. Das, S. Goyal, T. Lavril, M. Shuaibi, M. Riviere, K. Tran, J. Heras-Domingo, C. Ho, W. Hu, A. Palizhati, A. Sriram, B. Wood, J. Yoon, D. Parikh, C. L. Zitnick, and Z. Ulissi, "Open catalyst 2020 (OC20) dataset and community challenges," *ACS Catal.* **11**(10), 6059–6072 (2021).
- <sup>15</sup>R. Tran, J. Lan, M. Shuaibi, B. M. Wood, S. Goyal, A. Das, J. Heras-Domingo, A. Kolluru, A. Rizvi, N. Shoghi, A. Sriram, F. Therrien, J. Abed, O. Voznyy, E. H. Sargent, Z. Ulissi, and C. L. Zitnick, "The open catalyst 2022 (OC22) dataset and challenges for oxide electrocatalysts," *ACS Catal.* **13**(5), 3066–3084 (2023).
- <sup>16</sup>O. Mamun, K. T. Winther, J. R. Boes, and T. Bligaard, "High-throughput calculations of catalytic properties of bimetallic alloy surfaces," *Sci. Data* **6**(1), 76 (2019).
- <sup>17</sup>B. M. Comer, J. Li, F. Abild-Pedersen, M. Bajdich, and K. T. Winther, "Unraveling electronic trends in O\* and OH\* surface adsorption in the MO<sub>2</sub> transition-metal oxide series," *J. Phys. Chem. C* **126**(18), 7903–7909 (2022).
- <sup>18</sup>A. S. Burté, A. Nair, L. C. Grabow, P. J. Dauenhauer, S. L. Scott, and O. A. Abdelrahman, "CatTestHub: A benchmarking database of experimental heterogeneous catalysis for evaluating advanced materials," *J. Catal.* **442**, 115902 (2025).
- <sup>19</sup>A. Senocrate, F. Bernasconi, P. Kraus, N. Plainpan, J. Trafkowski, F. Tolle, T. Weber, U. Sauter, and C. Battaglia, "Parallel experiments in electrochemical CO<sub>2</sub> reduction enabled by standardized analytics," *Nat. Catal.* **7**(6), 742–752 (2024).
- <sup>20</sup>J. Abed, J. Kim, M. Shuaibi, B. Wander, B. Duijff, S. Mahesh, H. Lee, V. Gharakhanyan, S. Hoogland, E. Irtem, J. Lan, N. Schouten, A. U. Vijayakumar, J. Hatrick-Simpers, J. R. Kitchin, Z. W. Ulissi, A. van Vugt, E. H. Sargent, D. Sinton, and C. L. Zitnick, "Open catalyst experiments 2024 (OCx24): Bridging experiments and computational models," *arXiv:2411.11783* (2024).
- <sup>21</sup>M. J. Statt, B. A. Rohr, D. Guevarra, S. K. Suram, T. E. Morrell, and J. M. Gregoire, "The materials provenance Store," *Sci. Data* **10**(1), 184 (2023).
- <sup>22</sup>J. Hu, S. Stefanov, Y. Song, S. S. Omeel, S.-Y. Louis, E. M. D. Siriwardane, Y. Zhao, and L. Wei, "MaterialsAtlas.org: A materials informatics web app platform for materials discovery and survey of state-of-the-art," *npj Comput. Mater.* **8**(1), 65 (2022).
- <sup>23</sup>A. Zakutayev, J. Perkins, M. Schwarding, R. White, K. Munch, W. Tumas, N. Wunder, and C. Phillips (2017). "High throughput experimental materials database," NREL Data Catalog. Golden, CO: National Renewable Energy Laboratory. <https://doi.org/10.7799/1407128>
- <sup>24</sup>J. Gregoire (2021). "Event-sourced database containing the results of high-throughput electrochemistry experiments and the associated analyses," Caltech-DATA. <https://doi.org/10.22002/9ME7V-D5J04>
- <sup>25</sup>L. Yang, J. A. Haber, Z. Armstrong, S. J. Yang, K. Kan, L. Zhou, M. H. Richter, C. Roat, N. Wagner, M. Coram, M. Berndt, P. Riley, and J. M. Gregoire, "Discovery of complex oxides via automated experiments and data science," *Proc. Natl. Acad. Sci. U. S. A.* **118**(37), e2106042118 (2021).
- <sup>26</sup>X. Chen, Y. Gao, L. Wang, W. Cui, J. Huang, Y. Du, and B. Wang, "Large language model enhanced corpus of CO<sub>2</sub> reduction electrocatalysts and synthesis procedures," *Sci. Data* **11**(1), 347 (2024).
- <sup>27</sup>Y. Gao, L. Wang, X. Chen, Y. Du, and B. Wang, "Revisiting electrocatalyst design by a knowledge graph of Cu-based catalysts for CO<sub>2</sub> reduction," *ACS Catal.* **13**(13), 8525–8534 (2023).
- <sup>28</sup>Y. Su, X. Wang, Y. Ye, Y. Xie, Y. Xu, Y. Jiang, and C. Wang, "Automation and machine learning augmented by large language models in a catalysis study," *Chem. Sci.* **15**(31), 12200–12233 (2024).
- <sup>29</sup>Z. Zheng, F. Florit, B. Jin, H. Wu, S. Li, K. Y. Nandiwale, C. A. Salazar, J. G. Mustakis, W. H. Green, and K. F. Jensen, "Integrating machine learning and large language models to advance exploration of electrochemical reactions," *Angew. Chem., Int. Ed.* **64**(6), e202418074 (2025).
- <sup>30</sup>D. Zhang and H. Li, "Digital catalysis platform (DigCat): A gateway to big data and AI-powered innovations in catalysis," *ChemRxiv:10.26434/chemrxiv-2024-9lpb9* (2024).
- <sup>31</sup>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, and B. Mons, "The FAIR guiding principles for scientific data management and stewardship," *Sci. Data* **3**(1), 160018 (2016).
- <sup>32</sup>B. R. Wygant, K. Kawashima, and C. B. Mullins, "Catalyst or precatalyst? The effect of oxidation on transition metal carbide, pnictide, and chalcogenide oxygen evolution catalysts," *ACS Energy Lett.* **3**(12), 2956–2966 (2018).
- <sup>33</sup>J. A. Zamora Zeledón, M. B. Stevens, G. T. K. K. Gunasooriya, A. Gallo, A. T. Landers, M. E. Kreider, C. Hahn, J. K. Nørskov, and T. F. Jaramillo, "Tuning the electronic structure of Ag-Pd alloys to enhance performance for alkaline oxygen reduction," *Nat. Commun.* **12**(1), 620 (2021).
- <sup>34</sup>M. E. Kreider, M. B. Stevens, Y. Liu, A. M. Patel, M. J. Statt, B. M. Gibbons, A. Gallo, M. Ben-Naim, A. Mehta, R. C. Davis, A. V. Ievlev, J. K. Nørskov, R. Sinclair, L. A. King, and T. F. Jaramillo, "Nitride or oxynitride? Elucidating the composition–activity relationships in molybdenum nitride electrocatalysts for the oxygen reduction reaction," *Chem. Mater.* **32**(7), 2946–2960 (2020).
- <sup>35</sup>M. C. Biesinger, B. P. Payne, A. P. Grosvenor, L. W. M. Lau, A. R. Gerson, and R. S. C. Smart, "Resolving surface chemical states in XPS analysis of first row transition metals, oxides and hydroxides: Cr, Mn, Fe, Co and Ni," *Appl. Surf. Sci.* **257**(7), 2717–2730 (2011).
- <sup>36</sup>I. Timoshenko and B. Roldan Cuenya, "In situ/operando electrocatalyst characterization by X-ray absorption spectroscopy," *Chem. Rev.* **121**(2), 882–961 (2021).
- <sup>37</sup>C. C. L. McCrory, S. Jung, J. C. Peters, and T. F. Jaramillo, "Benchmarking heterogeneous electrocatalysts for the oxygen evolution reaction," *J. Am. Chem. Soc.* **135**(45), 16977–16987 (2013).
- <sup>38</sup>S. Deo, M. E. Kreider, G. Kamat, M. Hubert, J. A. Zamora Zeledón, L. Wei, J. Matthews, N. Keyes, I. Singh, T. F. Jaramillo, F. Abild-Pedersen, M. Burke Stevens, K. Winther, and J. Voss, "Interpretable machine learning models for practical

antimonate electrocatalyst performance,” *ChemPhysChem* **25**(13), e202400010 (2024).

<sup>39</sup>A. L. Strickler, A. Jackson, and T. F. Jaramillo, “Active and stable Ir@Pt core-shell catalysts for electrochemical oxygen reduction,” *ACS Energy Lett.* **2**(1), 244–249 (2017).

<sup>40</sup>A. L. Strickler, R. A. Flores, L. A. King, J. K. Nørskov, M. Bajdich, and T. F. Jaramillo, “Systematic investigation of iridium-based bimetallic thin film catalysts for the oxygen evolution reaction in acidic media,” *ACS Appl. Mater. Interfaces* **11**(37), 34059–34066 (2019).

<sup>41</sup>D. Higgins, M. Wette, B. M. Gibbons, S. Siahrostami, C. Hahn, M. Escudero-Escribano, M. García-Melchor, Z. Ulissi, R. C. Davis, A. Mehta, B. M. Clemens, J. K. Nørskov, and T. F. Jaramillo, “Copper silver thin films with metastable miscibility for oxygen reduction electrocatalysis in alkaline electrolytes,” *ACS Appl. Energy Mater.* **1**(5), 1990–1999 (2018).

<sup>42</sup>G. T. K. K. Gunasooriya, M. E. Kreider, Y. Liu, J. A. Zamora Zeledón, Z. Wang, E. Valle, A.-C. Yang, A. Gallo, R. Sinclair, M. B. Stevens, T. F. Jaramillo, and J. K. Nørskov, “First-row transition metal antimonates for the oxygen reduction reaction,” *ACS Nano* **16**(4), 6334–6348 (2022).

<sup>43</sup>Z. Lu, G. Chen, S. Siahrostami, Z. Chen, K. Liu, J. Xie, L. Liao, T. Wu, D. Lin, Y. Liu, T. F. Jaramillo, J. K. Nørskov, and Y. Cui, “High-efficiency oxygen reduction to hydrogen peroxide catalysed by oxidized carbon materials,” *Nat. Catal.* **1**(2), 156–162 (2018).

<sup>44</sup>G. Chen, M. B. Stevens, Y. Liu, L. A. King, J. Park, T. R. Kim, R. Sinclair, T. F. Jaramillo, and Z. Bao, “Nanosized zirconium porphyrinic metal-organic

frameworks that catalyze the oxygen reduction reaction in acid,” *Small Methods* **4**(10), 2000085 (2020).

<sup>45</sup>M. A. Hubert, A. M. Patel, A. Gallo, Y. Liu, E. Valle, M. Ben-Naim, J. Sanchez, D. Sokaras, R. Sinclair, J. K. Nørskov, L. A. King, M. Bajdich, and T. F. Jaramillo, “Acidic oxygen evolution reaction activity–stability relationships in Ru-based pyrochlores,” *ACS Catal.* **10**(20), 12182–12196 (2020).

<sup>46</sup>L. A. King, M. A. Hubert, C. Capuano, J. Manco, N. Danilovic, E. Valle, T. R. Hellstern, K. Ayers, and T. F. Jaramillo, “A non-precious metal hydrogen catalyst in a commercial polymer electrolyte membrane electrolyser,” *Nat. Nanotechnol.* **14**(11), 1071–1074 (2019).

<sup>47</sup>J. Benavides-Hernández and F. Dumeignil, “From characterization to discovery: Artificial intelligence, machine learning and high-throughput experiments for heterogeneous catalyst design,” *ACS Catal.* **14**(15), 11749–11779 (2024).

<sup>48</sup>H. Wu, A. Singh-Morgan, K. Qi, Z. Zeng, V. Mougél, and D. Voiry, “Electrocatalyst microenvironment engineering for enhanced product selectivity in carbon dioxide and nitrogen reduction reactions,” *ACS Catal.* **13**(8), 5375–5396 (2023).

<sup>49</sup>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, and I. Chorkendorff, “A rigorous electrochemical ammonia synthesis protocol with quantitative isotope measurements,” *Nature* **570**(7762), 504–508 (2019).