


## RESEARCH ARTICLE OPEN ACCESS

# opXRD: Open Experimental Powder X-Ray Diffraction Database

Daniel Hollarek<sup>1,2</sup> | Henrik Schopmans<sup>1,2</sup> | Jona Östreicher<sup>1,2</sup> | Jonas Teufel<sup>1,2</sup> | Bin Cao<sup>3</sup> | Adie Alwen<sup>4</sup> | Simon Schweidler<sup>2</sup> | Mriganka Singh<sup>5</sup> | Tim Kodalle<sup>5,6</sup> | Hanlin Hu<sup>7</sup> | Gregoire Heymans<sup>8</sup> | Maged Abdelsamie<sup>9,10</sup> | Arthur Hardiagon<sup>11</sup> | Alexander Wiczorek<sup>12</sup> | Siarhei Zhuk<sup>12</sup> | Ruth Schwaiger<sup>13</sup> | Sebastian Siol<sup>12</sup> | François-Xavier Coudert<sup>11</sup> | Moritz Wolf<sup>14</sup> | Carolin M. Sutter-Fella<sup>5</sup> | Ben Breitung<sup>2</sup> | Andrea M. Hodge<sup>4</sup> | Tong-yi Zhang<sup>3</sup> | Pascal Friederich<sup>1,2</sup> 

<sup>1</sup>Institute of Theoretical Informatics, Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany | <sup>2</sup>Institute of Nanotechnology, Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany | <sup>3</sup>Guangzhou Municipal Key Laboratory of Materials Informatics, Advanced Materials Thrust, Hong Kong University of Science and Technology (Guangzhou) (HKUST), Guangzhou, China | <sup>4</sup>Department of Chemical Engineering and Materials Science, University of Southern California (USC), Los Angeles, California, USA | <sup>5</sup>Molecular Foundry Division, Lawrence Berkeley National Laboratory (LBNL), Berkeley, California, USA | <sup>6</sup>Advanced Light Source, Lawrence Berkeley National Laboratory, Berkeley, California, USA | <sup>7</sup>Hoffmann Institute of Advanced Materials, Shenzhen Polytechnic, Shenzhen, China | <sup>8</sup>Lawrence Berkeley National Laboratory (LBNL), Chemical Sciences Division, Berkeley, California, USA | <sup>9</sup>Material Science and Engineering Department, King Fahd University of Petroleum and Minerals (KFUPM), Dhahran, Saudi Arabia | <sup>10</sup>Interdisciplinary Research Center for Intelligent Manufacturing and Robotics, King Fahd University of Petroleum and Minerals (KFUPM), Dhahran, Saudi Arabia | <sup>11</sup>Chimie ParisTech, PSL University, CNRS, Institut de Recherche de Chimie Paris, Paris, France | <sup>12</sup>Empa-Swiss Federal Laboratories for Materials Science and Technology (EMPA), Dübendorf, Switzerland | <sup>13</sup>Institute of Energy Materials and Devices, Forschungszentrum Juelich GmbH, Juelich, Germany | <sup>14</sup>Engler-Bunte-Institut & Institute of Catalysis Research and Technology, Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany

**Correspondence:** Pascal Friederich ([pascal.friederich@kit.edu](mailto:pascal.friederich@kit.edu))

**Received:** 7 March 2025 | **Revised:** 29 April 2025 | **Accepted:** 8 May 2025

**Funding:** German Research Foundation (DFG); Federal Ministry of Education and Research (BMBF), Grant/Award Number: 01DM21001B; Helmholtz Foundation Model Initiative; France 2030 framework by Agence Nationale de la Recherche, Grant/Award Number: ANR-22-PEXD-0009 of PEPR DIADEM; Office of Science, Office of Basic Energy Sciences, of the U.S. Department of Energy, Grant/Award Number: DE-AC02-05CH11231; Guangzhou-HKUST (GZ) Joint Funding Program, Grant/Award Number: 2023A03J0003; National Science Foundation (NSF), Grant/Award Number: DMR-2227178 and OISE-2106597; Helmholtz Research Program; Strategic Focus Area-Advanced Manufacturing (SFA-AM); Ministry of Science; Research and Arts Baden-Württemberg

**Keywords:** crystal structure determination | machine learning | open-access data | phase identification | powder x-ray diffraction

## ABSTRACT

Powder X-ray diffraction (pXRD) experiments are a cornerstone for materials structure characterization. Despite their widespread application, analyzing pXRD diffractograms still presents a significant challenge to automation and a bottleneck in high-throughput discovery in self-driving labs. Machine learning promises to resolve this bottleneck by enabling automated powder diffraction analysis. A notable difficulty in applying machine learning to this domain is the lack of sufficiently sized experimental datasets, which has constrained researchers to train primarily on simulated data. However, models trained on simulated pXRD patterns showed limited generalization to experimental patterns, particularly for low-quality experimental patterns with high noise levels and elevated backgrounds. With the Open Experimental Powder X-ray Diffraction Database (opXRD), we provide an openly available and easily accessible dataset of labeled and unlabeled experimental powder diffractograms. Labeled opXRD data can be used to evaluate the performance of models on experimental data and unlabeled opXRD data can help improve the performance of models on experimental data, for example, through transfer learning methods. We collected 92,552 diffractograms, 2179 of them labeled, from a wide spectrum of material classes. We hope this ongoing effort can guide machine learning research toward fully automated analysis of pXRD data and thus enable future self-driving materials labs.

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2025 The Author(s). *Advanced Intelligent Discovery* published by Wiley-VCH GmbH.

## 1 | Introduction

The advent of high-throughput experiments holds the prospect of significantly accelerating the speed of materials discovery [1]. The synthesis and characterization of novel materials are becoming increasingly efficient and automated, increasing the throughput of samples in experimentation pipelines [2–4].

After fabricating a new material, a number of analysis techniques can be used to characterize the sample. One method that can be used for phase identification, phase quantification, grain size characterization, and to determine the crystal structure of a new material is powder X-ray diffraction (pXRD). When using pXRD measurements, crystal structures are typically determined through Rietveld refinement. In Rietveld refinement, an initial crystal structure model is fitted to the observed diffractogram by iteratively updating the structural model. Each update of the structural model seeks to minimize the difference between the observed diffractogram and the diffractogram simulated from the current structural model [5, 6]. As Rietveld refinement is a local optimization method, the result of the refinement procedure is generally only as good as the initial structural model the process started from.

Manually performing Rietveld refinement is time-consuming and often requires expert knowledge. It is not scalable to the degree required to keep up with advances in throughput and efficiency in other steps of the experimentation pipeline. The refinement process requires the operator to determine an initial structural model from which the refinement can start and as well as initial values for parameters that characterize the background [7]. The structural model is usually obtained using search-match software, which identifies crystal structures with similar powder diffraction patterns from a database of crystal structures with accompanying powder diffraction patterns. However, an initial structural model obtained from such a database is not guaranteed to lead to an accurate structure solution through Rietveld refinement, especially not for novel structures. Additionally, attempting to refine all crystal structure parameters at once is known to lead to unphysical results [4]. Hence, parameters are refined iteratively, with each iteration only refining a limited set of parameters. Finding the correct order in which to refine structure parameters and finding the correct values for initial background parameters both present problems that add to the difficulty of the refinement process.

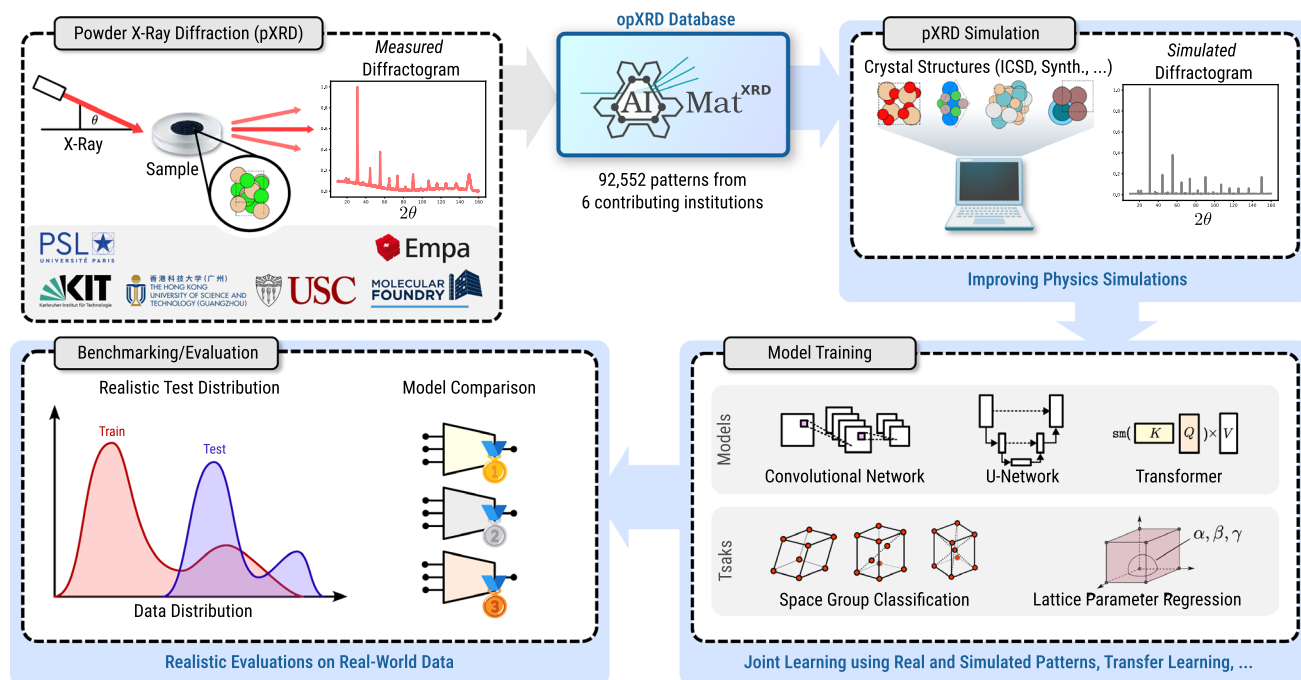
Machine learning has the potential to speed up the manual analysis of powder diffractograms and keep pace with an automated high-throughput experimentation environment [8, 9]. Models can be either trained to predict crystal structure information directly given a diffractogram, or they can be used to automate the conventional refinement workflow. In the latter case, a model would first predict an initial crystal structure [9], which is then refined by a second model trained to perform the refinement process [10]. So far, due to an absence of labeled datasets with experimental diffractograms [11], machine learning in this domain has largely relied on diffractograms simulated from known structures [12, 13] or, most recently, from generated synthetic crystals [14]. Models trained on datasets with simulated diffractograms have already shown strong performance in predicting phases [12, 15, 16], lattice parameters [17–20], space

group [12, 14, 20–26], and crystallite size [17, 26] from simulated diffractograms. However, the performance substantially drops off when these models are applied to data originating from experiments [11, 14, 20, 21, 23]. This discrepancy in performance arises due to imperfections in experimental data, which are not present in diffraction patterns modeled under ideal conditions. This is discussed in more detail below.

Both labeled and unlabeled datasets of experimental powder diffractograms hold significant value for machine learning-based pXRD analysis, particularly with regard to bridging the performance gap between simulated and experimental domains. Labeled experimental data can be used to test and benchmark existing and new automated analysis approaches. This enables researchers to gauge how well a given model would perform under real-world conditions if integrated into an automated experimentation pipeline. Unlabeled experimental data enables machine learning researchers to evaluate how closely their simulations represent experimental data and modify their simulation algorithms accordingly. Unlabeled data can also find applications in transfer learning approaches to transfer model capabilities from the domain of simulated diffractograms to the domain of experimental diffractograms. While some experimental powder databases exist, their utility is limited by the fact that they are either small or not openly accessible.

In this work, we introduce an open powder X-ray diffraction (opXRD) database featuring a broad range of patterns collected from experiments. The objective of this work is to introduce and disseminate a large, open experimental pXRD dataset, paving the way for future studies that will evaluate and benchmark its impact on model performance. With a total of 92,552 patterns collected from 6 contributing institutions, the opXRD database exceeds the size of the previously largest database of openly accessible experimental powder diffraction data by two orders of magnitude. To the best of our knowledge, the largest database of this type is the RRUFF database, containing 1290 experimental powder diffraction patterns [27, 28]. Larger commercial datasets such as the PDF5+ [29] and the Linus Pauling File [30] exist, but their utility is limited by fees and restrictive licenses. License terms of commercial datasets, such as the PDF5+ and the Linus Pauling File, prohibit or restrict the publication of models trained on their data. In contrast, the opXRD database is both free and imposes no restrictions on how its data is used. Figure 1 provides an overview of machine learning workflows enabled and supported by the opXRD database.

Of the 92,552 patterns in the opXRD database, 2179 patterns come with at least partial structural information of the underlying sample. Of these 2179 labeled patterns, 912 have labels of the full crystal structure. While the fraction of samples with structural labels (2.35%) may appear small, this fraction represents the largest openly available collection of experimentally derived, structure-annotated pXRD patterns. As a comparison, the RRUFF database, often used for benchmarking ML models, contains partial labels for 1290 patterns, but no atomic coordinates. The opXRD database is larger in size, richer in labels, and broader in represented experimental setups compared to previous openly available datasets. Given the inherently labor-intensive nature of manual labeling in pXRD analysis, it is impractical to expect a fully labeled dataset at the scale of



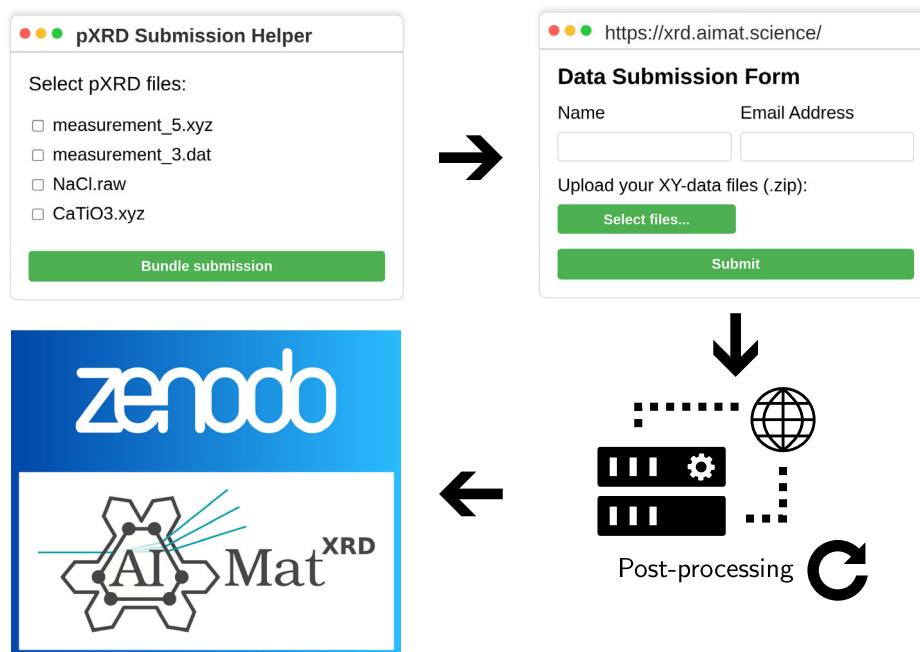
**FIGURE 1** | Experimental powder X-ray diffraction (pXRD) patterns from several contributors are collected in the opXRD database. The proposed open-access database of experimental data aims to support each step in the pXRD-related machine learning workflow by informing better physics simulations, supplying model training data, and providing a foundation for realistic performance evaluations.

simulated training datasets, which commonly exceed  $10^6$  patterns [14, 28, 31]. Therefore, next to the option of benchmarking models and methods on the labeled subset, opXRD is designed to complement vast simulated datasets. This can be achieved through the adjustment of simulation parameters by comparing with the unlabeled subset of the database, and through transfer learning strategies. We now want to discuss these two options of utilizing the unlabeled portion of our database in more detail.

The neglected effects that lead to discrepancies between simulated patterns and patterns stemming from experiments are largely known. Unaccounted effects may include preferred crystallite orientation, variations in grain size, crystal defects, the impact of temperature on the scattering process, internal stress, the non-monochromaticity of the X-ray source, and X-ray-induced fluorescence [21, 32, 33]. Additionally, varying experimental setups produce distinct powder diffraction patterns on the same sample. Features that may vary between experimental setups include the shape of diffraction peaks, the wavelength and polarization of the employed X-ray source, and the detector geometry [21, 32, 33]. The recorded scattering angles may also be slightly falsified if the sample is displaced from its intended position [21, 34]. As these and more neglected effects are integrated into the simulation process, real powder diffraction data can be used to evaluate how closely simulated data matches up with real data. While direct comparisons are only possible on labeled patterns, comparing the strength and prevalence of features between simulated and real data can nevertheless provide information about the fidelity of the simulation. Taking into account all neglected effects without making approximations will incur significant computational costs that will lower the size of the generated training data. A more efficient approach could be to use real experimental data to identify the effects that have the largest impact in practice and model them heuristically.

The second way in which unlabeled experimental data can serve to bridge the performance gap between simulated and experimental domains is through transfer learning. The objective of transfer learning is to transfer the capabilities of a model learned on a source domain in which labeled data is abundant to a target domain in which labeled data is sparse [35]. In this context, the source domain is simulated powder diffraction patterns and the target domain is experimental powder diffraction patterns. Many approaches to transfer learning have been proposed, particularly in the domain of image classification [36, 37]. These existing techniques can be adapted to facilitate transfer learning in the context of pXRD patterns. Seddiki et al. have already successfully applied transfer learning in the domain of mass spectrometry to boost the accuracy of mass spectrum classification models [38]. Since both mass spectrometry data and pXRD data are one-dimensional, this work demonstrates the merit of transfer learning in a setting similar to pXRD.

The opXRD database is intended as a growing, community-driven initiative. The database we present here is the first version, but we hope to further increase the database size through active engagement with the pXRD community. Our primary objective is to minimize the effort and thus the barrier to contributing experimental data to the opXRD database. Thus, we developed a program that helps to find and share data from pXRD lab computers. Users can select their most common pXRD file types, the program lists all files of that type, and users can select or deselect certain folders or files for sharing. Selected contributions will be uploaded to opXRD, processed to a common file format, and—if wanted—published on Zenodo on behalf of the contributors, before becoming part of the opXRD database. If labels are available, they can be shared with opXRD as well. Further details can be found on the opXRD website (<https://xrd.aimat.science/>). An overview of this process is given in Figure 2 below.



**FIGURE 2** | Overview of the data collection pipeline. Datasets are submitted using an online submission form, optionally with the help of our submission helper software. After post-processing and data homogenization, we offer the creation of a Zenodo entry for each user submission and subsequently include the submission in the opXRD database.

As argued by Aranda and Kroon-Batenburg et al. [39, 40], sharing raw powder diffraction data is not only in the interest of furthering machine learning research but is also in line with open science principles. It furthers the ability of other researchers to reproduce published work and in turn adds to the credibility of the publisher of the data. Compared to publishing data individually, publishing data on the opXRD database has the added benefit of contributing to a large, homogeneous dataset with a standardized interface. This makes the data more easily accessible to other researchers and provides more value to researchers seeking large quantities of data. However, further data annotation with metadata is required to fully fulfill the FAIR data principles.

The opXRD database contains pXRD patterns from single- and multiphase materials from a wide variety of material classes, including high-entropy materials, perovskites, and commercial catalysts. Some of the XRD data was collected on thin films rather than on true powder samples, which may influence the quality of the data in regards to full structure resolution. Additionally, some of the data was collected in grazing-angle geometry rather than in the usual Bragg-Brentano geometry employed in powder diffraction. The broad range of available experimental samples contained in the opXRD database makes it possible to apply state-of-the-art ML approaches to the domain of pXRD analysis. We hope that the opXRD database can drive ML research in this field towards more advanced automated analysis workflows that can accelerate materials science research through ready application in high-throughput experimentation pipelines. Details of the experiments of research groups contributing to the opXRD database are discussed in Section 3. A detailed description of how to acquire and use opXRD data is given in Section 4, and Section 5 describes how further data can be contributed.

### 1.1 | Review of Machine Learning-Based pXRD Analysis

To showcase the need for datasets such as the one presented in this publication, we now discuss some recent approaches that apply machine learning methods to classification and regression tasks for powder diffractograms.

In 2020, Lee et al. trained a deep convolutional neural network (CNN) using simulated diffractograms based on structures from the ICSD, which is able to classify occurring phases in diffractograms of a specific compound pool [41]. In 2022, they further developed models based on fully convolutional neural networks and transformer encoders that predict the crystal system, the space group, and other structural properties, such as the band gap [42]. With their best model for the crystal system prediction on ICSD structures, they achieved a test accuracy of 92.2%. In 2017, Park et al. reached a test accuracy of roughly 81% for a CNN, which classifies space groups of simulated single-phase diffractograms [12].

A regression analysis on lattice parameters within a broader framework encompassing all material classes was conducted by Chitturi et al. [18] in 2021. They developed a distinct CNN for each crystal system, utilizing a merged dataset from both the ICSD and the Cambridge Structural Database, and managed to achieve a mean absolute percentage error of about 10% for the lattice lengths, although they encountered difficulties in accurately predicting angles. In 2024, Zhang et al. introduced a convolutional self-attention neural network trained on simulated patterns to classify crystal types [20]. Their model was tested on 23,073 unary, binary, and ternary inorganic crystal structures sourced from the COD. The study observed a noticeable performance drop when the pre-trained model



was applied to real experimental patterns as opposed to simulated data. However, their recent work [21] proposes using convolutional peak descriptors that consider the detector’s geometry, which reduces the performance gap in their benchmark tests.

Neural networks trained purely on experimental diffractograms can perform well when the range of samples is narrow and the data is collected only on a single machine [13, 43]. However, in a more general setting with a wide range of investigated samples and employed diffractometers, training neural networks purely on experimental diffractograms becomes infeasible. This is because of the limited availability of labeled experimental diffractograms relative to the scope of the task. However, in 2023, Salgado et al. [31] showed that adding a fraction of experimental patterns to a simulated training dataset improves the performance on unseen experimental patterns. They used 50% of the experimental patterns contained in the RRUFF database and added those to their large simulated training set. Then they tested their model’s performance on the other half of the RRUFF database and achieved a performance increase in the 230-way space group classification accuracy of 11 percentage points compared to the same model only trained on simulated patterns.

In 2024, Schuetzke et al. trained a classifier to classify if a diffractogram stems from an amorphous, single-phase, or multi-phase sample [44]. Due to the lack of experimental pXRDs, they built a pipeline to augment simulated diffractograms of a reference structure by, among other things, slightly varying the underlying crystal lattice. For spinel structures, they reported an accuracy of 100%, but they also proved that their approach can be transferred to other datasets.

In 2023, Schopmans et al. presented an approach to generate synthetic crystal structures and their corresponding pXRD patterns on the fly during the training process [14]. This approach defeats the issue of a limited dataset size, which limits the depth of neural networks that can be trained. However, the accuracy dropped substantially when we applied our space group classification model to experimental patterns from the RRUFF database. Augmenting our simulated patterns with background, noise, and impurities helps to bring simulated diffractograms closer to experimental ones, making models trained on them more performant on experimental diffractograms. However, this augmentation process could be improved by incorporating background

and noise statistics from a broader experimental pXRD database, such as the one presented in this publication.

It becomes apparent that the more general the task is, the more challenging the transfer to experimental data becomes. For example, the space group classification task across all material systems is very general. Therefore, transferring it to the application on experimental diffraction patterns is difficult [14, 23, 42]. On the contrary, there are some successful approaches that also work well on experimental data, but those are mostly methods that do phase determination in a limited compound space, making the task less complex [41, 44].

The current volume of experimental pXRD patterns is insufficient to effectively train ML models, highlighting an urgent need for a comprehensive experimental pXRD database. The most advanced ML models currently are trained on approximately  $10^5$ – $10^6$  simulated diffractograms [14, 31]. This is, to the best of our knowledge, two orders of magnitude larger than the largest currently curated experimental dataset, the PDF-5+, with approximately  $2 \times 10^4$  experimental patterns. It is even one order of magnitude larger than the approximately  $10^5$  unlabeled diffractograms in the initial version of the opXRD dataset we present here.

To make ML-based pXRD data identification practical for experimental use and automate structure prediction despite lacking experimental training data, two key approaches are essential. First, developing more sophisticated simulation methods to better approximate experimental patterns [21] by using statistics from experimental diffractograms. Second, creating an experimental database that enables transfer learning to bridge the gap between simulated and real-world data. For both of these steps, the development of opXRD is particularly significant, as it will provide a comprehensive experimental benchmark for the community, allowing fair comparison of baseline models and accurate evaluation of their applicability in real experimental situations.

## 2 | Existing Datasets

To contextualize opXRD within the current environment of experimental powder diffraction data, the list below provides an overview of the largest crystal structure databases that offer access to experimental powder diffraction data. For an overview of these databases, refer to Table 1 below.

**TABLE 1** | Overview of experimental powder diffraction databases: The column “O.A.” indicates whether or not the database is open-access. The availability of the chemical composition, space groups, lattice parameters, and the full structure of the underlying samples is indicated by the columns “Comp.,” “Spg.,” “Lattice” and “Full structure”, respectively.

Name	No. Patterns	O.A.	Comp.	Spg.	Lattice	Full Structure	Year est.
Linus Pauling file	21,700	✗	✓	✓	✓	✓	2002
Powder Diffraction File <sup>a</sup>	20,800	✗	✓	✓	✓	52%	1941
RRUFF	1290	✓	✓	✓	✓	✗	2006
Crystallography Open Database	1052	✓	✓	✓	85%	85%	2003
PowBase	169	✓	✓	✗	✗	✗	1999

<sup>a</sup>The PDF lists the Material Platform for Data Science (MPDS) as a database source. Since the MPDS is hosted by the Pauling File project, there is likely significant overlap in the experimental patterns available in the PDF and the Linus Pauling File.

## 2.1 | Linus Pauling File [45]

The Linus Pauling File is a largely commercial crystal structure database published and maintained by the Pauling File project [29]. It is currently distributed as Pearson Crystal data [46] and the Materials Platform for Data Science (MPDS) [47]. The database, first published in 2002, currently contains more than 534,000 crystal structures [47] and 21,700 corresponding experimental powder diffraction patterns [46]. This makes the Pauling file, to the best of our knowledge, the largest collection of experimental powder diffraction data available to researchers. As of November 2024, Pearson's crystal data is available to researchers through a purchase of a 1-year license starting at a price point of 2200€ [48]. The MPDS is partially open, with the open portion of the MPDS data accessible through a web interface [47]. API access to the full MPDS can be purchased through a 1-year license starting at 9500€ [49]. We asked the Pauling File project whether the experimental powder diffraction data is accessible through the MPDS API. The Pauling File project responded that this data is not currently provided through the API, but could be offered in the future at the request of customers.

## 2.2 | Powder Diffraction File [50]

The Powder Diffraction File (PDF), published and maintained by the International Center for Diffraction Data (ICDD), is a large collection of materials with accompanying powder diffraction data first published in 1941 [28]. According to the ICDD, the latest release of the PDF, the PDF5+, contains over a million materials with accompanying powder diffraction data. However, since most of these powder diffraction patterns are simulated, we asked the ICDD about the number of experimental diffraction patterns in the PDF5+. We were told that 20,800 of the patterns in the PDF5+ stemmed from experiments and that 10,954 of these patterns were accompanied by the atomic coordinates of the underlying structures. Since the PDF5+ lists the MPDS as a database source, there is likely a significant overlap in the experimental patterns found in the PDF5+ and those found in the Pauling file. Currently, the PDF5+ is available to researchers through a purchase of a 1-year license starting at a price point of \$6265. However, the ICDD does not allow researchers to train machine learning models on PDF5+ data, regardless of whether the resulting models are published [51].

## 2.3 | RRUFF [52]

The RRUFF Mineral Database, first published in 2006, provides detailed information on minerals, including their chemical compositions, crystallography, and spectroscopic data [27]. Managed by the University of Arizona, it was created to serve as a public repository for mineral identification and research. It contains 1290 powder diffraction patterns stemming from experiments, each labeled with the lattice parameters and composition of the underlying structures. The RRUFF data is openly accessible on its official website [52].

## 2.4 | Crystallography Open Database [53]

The Crystallography Open Database (COD) is an open-access collection of crystal structures founded in 2003 [54]. It currently provides over 500,000 crystal structures. Of these files, 1052 contains the experimental powder diffraction data that was used to determine the underlying crystal structures of the investigated samples. Hence, the experimental powder diffraction data contained in the COD is mostly labeled with the full crystal structure information. The data is openly accessible in the form of .cif files on the official COD website [53].

## 2.5 | PowBase [55]

PowBase is a database of 169 mostly unlabeled experimental powder diffraction patterns collected and maintained by crystallography researcher Armel Le Bail starting in 1999. PowBase is an initiative suggested in the Structure Determination by Powder Diffractometry (SDPD) mailing list, which was co-maintained by Le Bail. The COD is another community initiative that grew out of this mailing list. As of March 2025, all 169 patterns are still freely available for download on the official website [55].

There is also publicly available powder diffraction data uploaded to datasets on Zenodo. However, this data is split into disparate entries that typically only contain the work of a single research project. Additionally, extracting powder diffraction data at scale is hindered by the fact that the data is often given in plain text files in non-standardized formats, which are difficult to parse automatically. We are currently planning a systematic large-scale extraction of powder diffraction data from databases like Zenodo with the help of a large language model. This data will be included in a future release of the opXRD database.

While not strictly speaking a powder diffraction database, the High-Throughput Experimental Materials Database (HTEM) by the National Renewable Energy Laboratory (NREL) is a valuable source of X-ray diffraction data [56]. Currently, the HTEM database contains 65,779 thin-film samples with corresponding X-ray diffraction data [57]. Each database entry includes the elemental composition of the underlying sample but does not provide any information on its structure. HTEM data is open-access and can be downloaded through an API provided by NREL.

Aside from the databases mentioned above, we have also investigated several other crystal structure resources in search of experimental powder diffraction data. Crystal structure resources that were investigated but not found to contain any appreciable amount of publicly available experimental powder diffraction data include the Inorganic Crystal Structure Database [58], the Cambridge Structural Database [59], the Materials Project database [60], the Crystallographic and Crystallochemical Database [61], the Bilbao Incommensurate Crystal Structure Database [62], the Mineralogy Database [63], the IUCr Raw data letters [64], the U.S. Naval Research Laboratory Crystal Lattice-Structures [65], the Athena Mineral database [66] and the Protein data bank [67]. The lack of experimental powder diffraction data in these databases is to be expected, as most structure solutions are achieved through single-crystal diffraction.

### 3 | opXRD Database

In collaboration with several other research institutions, we have collected a database of 92,552 experimental powder diffraction patterns. Of these patterns, 2179 are at least partially labeled with structural information, and 912 are labeled with the full crystal structure of the underlying material. The following research institutions contributed data to the opXRD database: The French National Centre for Scientific Research (CNRS), Hong Kong University of Science and Technology (Guangzhou) (HKUST), University of Southern California (USC), Lawrence Berkeley National Laboratory (LBNL), Empa-Swiss Federal Laboratories for Materials Science and Technology (EMPA), and the Karlsruhe Institute of Technology (KIT). We have taken measures to ensure data validation through both manual and automated processes. Before parsing the data, we established the file formats and data organization of each submission to ensure that the files were compatible with our custom parsing mechanism. As part of the automatic parsing process, we filtered the submitted datasets to exclude patterns with invalid features such as only one unique recorded angle, negative angles, less than 50 recorded angles total, or all intensities being zero. After the parsing, we also manually inspected a random selection of patterns from each submitted dataset for any anomalies that would warrant further investigation.

We standardized the associated structural information according to the standards described by Setyawan et al. [68] using PYMATGEN. In particular, this standardization enforces a single crystal axis convention throughout all opXRD data. Table 2 provides an overview of the contributions of each institution.

The variance of the data was analyzed using principal component analysis (PCA). PCA can be applied to datasets  $X \subset \mathbb{R}^N$  to reduce the number of components needed to describe points  $p \in X$  up to some tolerance in lost accuracy. In the context of PCA, the cumulative explained variance ratio is a measure of how much of the variance in the dataset  $X$  can be explained using a given number of components. For a rigorous definition of PCA and the explained variance ratio, we refer to the literature [69]. Here, PCA was performed on datasets of X-ray diffraction patterns. These datasets  $X$  are subsets of  $\mathbb{R}^N$  with  $N = 512$  since each pattern  $p \in X$  was standardized to have 512 intensity values spread

out evenly from  $0^\circ$  to  $180^\circ$  using zero padding and interpolation with cubic splines. Hence, the maximal components that could be needed to describe a dataset of diffraction data in this context is  $N = 512$ . However, the maximal number of components is even lower for datasets that contain less than 512 patterns. In this case, the maximal number of components is equal to the number of patterns in the dataset since each pattern can add at most one degree of freedom to the dataset  $X \in \mathbb{R}^N$ . Hence, the maximum number of components  $N_{\max}$  of a pattern dataset  $X$  is given as follows:

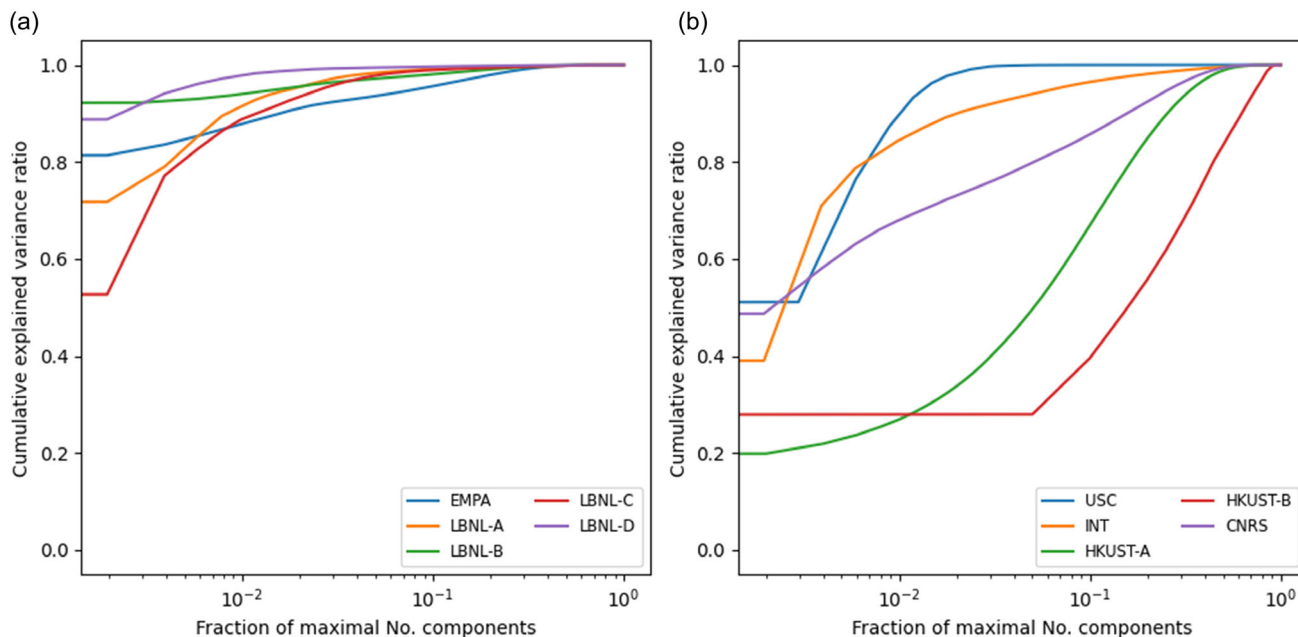
$$N_{\max} = \min(N_{\text{values}}, N_{\text{patterns}}) \quad (1)$$

Here  $N_{\text{values}} = 512$  is the number of recorded intensity values per pattern and  $N_{\text{patterns}}$  is the number of patterns in the dataset  $X$ . Figure 3 below shows the cumulative explained variance ratio over the fraction of maximal No. components  $N_{\max}$  as defined above. In this figure, a faster convergence of the cumulative variance ratio towards one indicates that the patterns in this dataset are largely similar. For example, the contributions by USC and LBNL contain many very similar patterns. The patterns in the USC dataset are similar because the underlying samples are all variations of CuNi and CuAl alloys. The patterns submitted by LBNL are similar because they stem from in-situ recordings where several hundred or several thousand patterns were collected over time per sample. In contrast, the CNRS and the HKUST contributions each are collections that encompass many research projects over a large period of time and thus exhibit a high degree of variability between individual patterns.

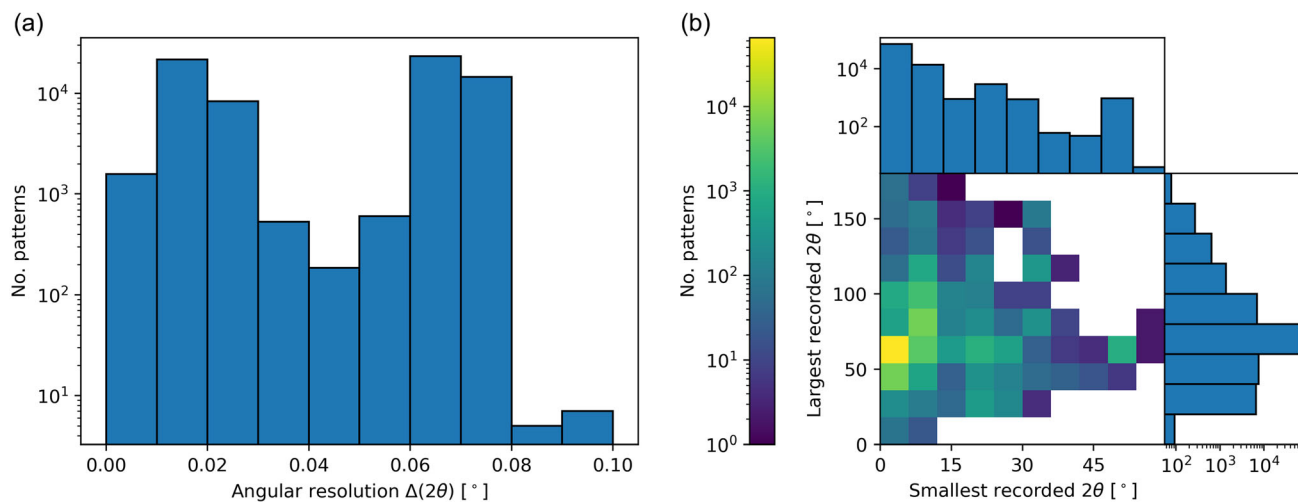
Figure 4 illustrates the distributions of pattern properties in the opXRD database. Nearly all patterns have an angular resolution smaller than  $\Delta(2\theta) = 0.1^\circ$ . Here, the angular resolution is defined as the range of recorded angles divided by the number of recorded intensity values along that range. For most patterns, the lowest recorded angle is smaller than  $30^\circ$  and the highest recorded angle is smaller than  $120^\circ$ . The start-to-end angle distribution reveals that all diffractograms start in a narrow window between 0 and approximately 50, while they end between 50 and 150, with the majority of patterns going from 0 to approximately 70. Unlike most ML approaches using synthetic data over the full angle range with fixed resolution, the opXRD dataset has a strongly varying angle range and resolution. Hence, working

**TABLE 2** | Overview of the contributions to the opXRD database: The availability of the chemical composition, space groups, lattice parameters, and the full structure of the underlying samples is indicated by the columns “Comp.,” “Spg.,” “Lattice” and “Full structure.” respectively.

Institution	No. Patterns	Comp.	Spg.	Lattice	Full Structure	Research Project
CNRS	1052	✓	85%	✓	85%	Diffraction data extracted from the COD
USC	338	✓	✓	90%	✗	Study of CuNi and CuAl alloys
HKUST (GZ)	520	4%	4%	4%	4%	Phase identification dataset
EMPA	770	✓	63%	✗	✗	Metal halide perovskites, Zn-V-N libraries
INT	19,796	✗	✗	✗	✗	Compilation of various projects
IKFT	64	✗	✗	✗	✗	Commercial catalysts, metals, metal oxides
LBNL	70,012	✗	✗	✗	✗	Perovskites precursors, Mn-Sb-O system
$\sum$ Labeled	2,179	79%	66%	63%	42%	Partially and fully labeled opXRD data
$\sum$ Unlabeled	90,373	✗	✗	✗	✗	Unlabeled opXRD data



**FIGURE 3** | Explained variance ratio over the fraction of the maximum number of components for each dataset contributed to the opXRD database: (a) Contributions EMPA, LBNL-A, LBNL-B, LBNL-C, LBNL-D, (b) contributions USC, INT, HKUST-A, HKUST-B, CNRS. Here, the maximum number of components refers to  $N_{\max}$  as defined in Equation (1). Datasets contributed by the same institution are labeled alphabetically in the order in which they are described in the texts towards the end of this section.



**FIGURE 4** | Histograms detailing properties of all diffraction patterns in the opXRD database: (a) distribution of angular resolutions, (b) distribution of smallest and largest recorded  $2\theta$  values.

with this data requires additional pre-processing methods such as padding and interpolation, or more flexible ML models beyond standard CNNs.

Figure 5 illustrates how structural properties are distributed among the structures underlying the labeled subset of the opXRD database. Most structures include either N, C, or O atoms, and have unit cells that contain less than 100 atoms and are smaller than  $10 \text{ \AA}^3$ . The most common space groups include the orthorhombic  $Pnma$ , the monoclinic  $P2_1/c$ , and the cubic  $Fm3m$ .

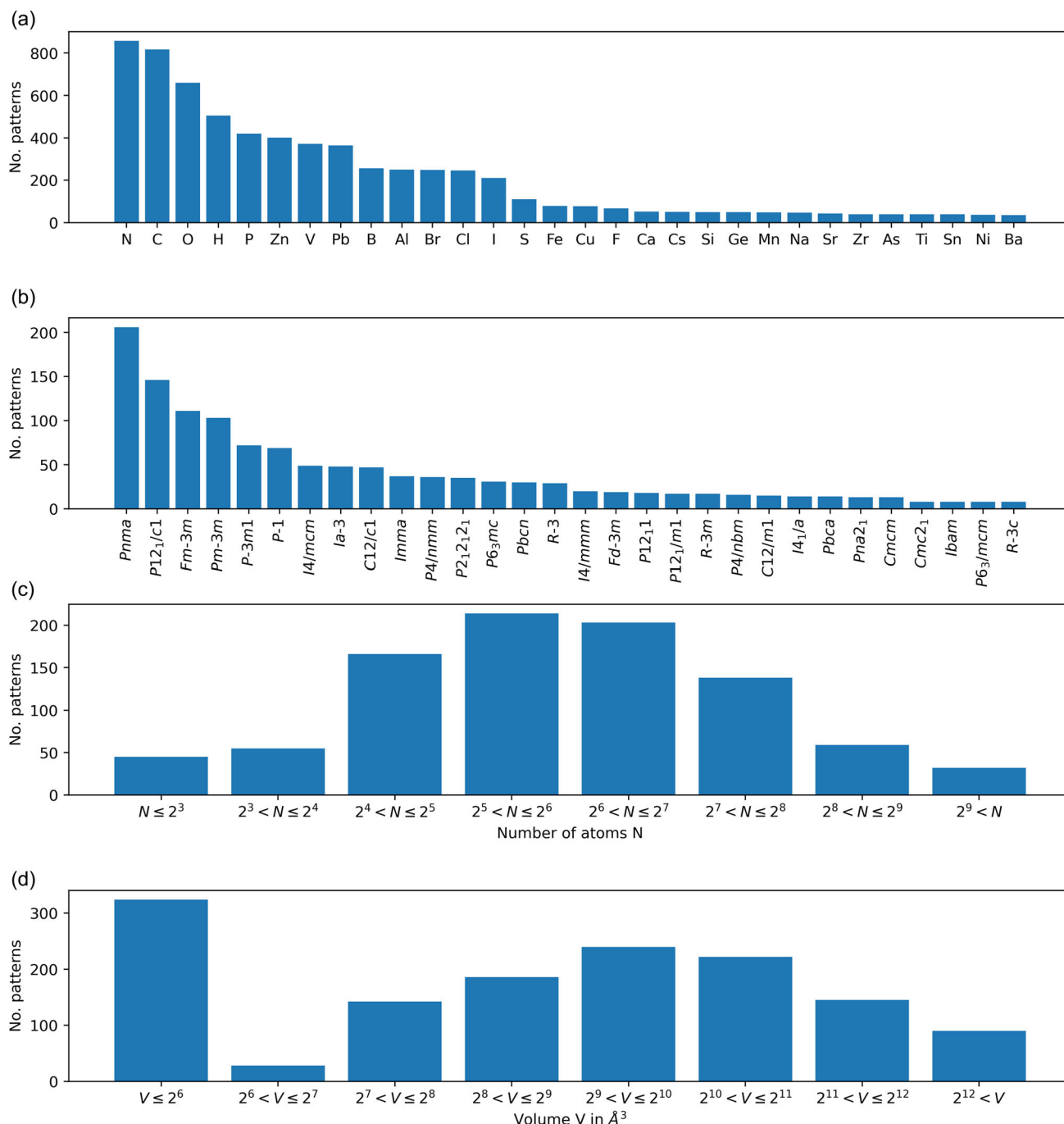
In the following, we will describe the datasets contributed by each of the collaborating research groups and institutions.

Each paragraph includes a description of the investigated materials and how X-ray diffraction data was collected. If applicable, the presence of thin-film samples or atypical diffraction geometries is indicated. Most data was collected using Cu radiation sources, which has a  $K_{\alpha 1}$  wavelength of  $\lambda = 1.54056 \text{ \AA}$  and a  $K_{\alpha 2}$  wavelength of  $\lambda = 1.54439 \text{ \AA}$ .

### 3.1 | Institut De Recherche De Chimie Paris, CNRS

Experimental pXRD data was extracted from the Crystallography Open Database (COD) [70, 71]. The COD is, to our knowledge,





**FIGURE 5** | Histograms detailing properties of the structures underlying labeled diffraction patterns in the opXRD database: (a) distribution of structures containing specified elements, (b) distribution of structures containing space groups, (c) distribution of number of atoms  $N$  contained in unit cell, (d) distribution of unit cell volume  $V$  in  $\text{\AA}^3$ .

the largest open-access collection of experimental crystal structures of organic, inorganic, and metal-organic compounds and minerals, containing more than 500,000 entries. The data in the COD are placed in the public domain and licensed under the CC0 License. Of the entire COD database, 5432 structures contained at least one tag from the CIF\_POW dictionary, that is, a tag relating to powder diffraction studies. These 5432 structures only account for 1% of the total COD database, but this is to be expected since most crystal structures are resolved from single-crystal diffraction. Of these 5432 files, most contained only

metadata related to the powder diffraction experiment, but did not include the raw data of the pattern itself. We could extract raw experimental pXRD patterns from 1052 files in total, after curation of a small number of files with clearly invalid data.

The pXRD data from the COD database are of high quality, with a median resolution of  $\Delta(2\theta) = 0.013^\circ$  and an average number of 9190 points measured per pattern. They span a wide chemical space, including organic, inorganic, and hybrid structures, and 75 different elements of the periodic table.

### 3.2 | Guangzhou Municipal Key Laboratory of Materials Informatics, HKUST (GZ)

Two datasets were contributed to the opXRD database. The first dataset (HKUST-A) is a selected subset of a small-scale experimental powder X-ray database developed over the past 2 years, called the X-Ray Phase Identification Public Experimental Dataset (XRed) (<https://github.com/WPEM/XRED>). The primary goal of XRed is to support the advancement of intelligent phase identification technology by providing a foundation for data collection in future large-scale machine learning applications. XRed primarily focuses on metal and metal-oxide particles, with data collected using diffractometers such as the Empyrean 3.0, Aeris, and Bruker D8 Advance, all employing Cu X-ray sources. The dataset HKUST-A contains 21 pXRD patterns, each labeled with a corresponding CIF file that documents the refined structure. Data are categorized by elemental systems and include original experimental files, spanning single-phase to five-phase mixtures, as well as mixtures designed for various research tasks.

In addition to XRed, the opXRD database integrates an experimental dataset composed of powder diffraction data sourced from open-access publications and collaborating institutions (HKUST-B). These institutions have provided the data with full authorization for research purposes. Compared to XRed, this dataset offers broader chemical element coverage, encompassing ionic, atomic, and metallic crystals. It is also larger, containing 499 entries. However, unlike XRed, these data entries are not accompanied by CIF files.

### 3.3 | Laboratory for Surface Science and Coating Technologies, Empa

Combinatorial Zn–V–N libraries were synthesized using radio-frequency co-sputtering of Zn and V in a mixed Ar and N<sub>2</sub> plasma. An orthogonal deposition temperature and composition gradient was created, resulting in a deposition temperature of 220°C for samples 1–9 and 114°C for samples 37–45. The composition for each sample was determined using X-ray fluorescence (XRF) spectroscopy, which was further calibrated through Rutherford backscattering spectroscopy (RBS) based on selected samples. The newly identified and isolated semiconductor Zn<sub>2</sub>VN<sub>3</sub> was identified to exhibit a cation-disordered wurtzite structure as verified by additional GI-XRD and SAED measurements [72].

Tin halide perovskites were deposited using single-step spin-coating as reported elsewhere [73]. Methylammonium lead iodide libraries with varying degrees of residual PbI<sub>2</sub> were deposited using a two-step procedure involving both thermal evaporation of PbI<sub>2</sub> and subsequent spin-coating of a methylammonium solution. The relative phase fractions were quantified using supplementary azimuthal angle scans coupled with structural factors and geometrical factors as reported elsewhere [74]. Fully inorganic lead perovskite libraries were prepared using thermal co-evaporation of lead and cesium halide salts. All metal halide perovskite libraries were measured within a custom-made X-ray transparent inert-gas dome, resulting in the presence of minor additional features within the  $\theta = 19\text{--}31^\circ$  range. For all

combinatorial libraries where any phases are specified, the complete set of phases is reported in the metadata.

XRD data was measured using a Bruker D8 Discover equipped with a Cu radiation source in a Bragg–Brentano geometry. For the reported datasets, the instrument was equipped with a Goebel mirror effectively removing the Cu K <sub>$\beta$</sub>  radiation. The data set originates from the combinatorial exploration of the Zn–V–N compositional space, as well as data gathered from multiple research activities on more established metal halide perovskite semiconductors. All data was collected from thin films deposited on borosilicate glass. The Zn–V–N films showed some preferential out-of-plane orientation, while for the perovskites the preferential orientation was minimal, resulting in the presence of all reflections.

### 3.4 | Institute of Nanotechnology, KIT

X-ray diffraction data was collected from a wide range of research projects conducted at the Institute of Nanotechnology over the past 10 years. A major part of the research focused on high-entropy materials, which involved incorporating many different elements into single-phase structures, leading to peak shifts or phase separations. Most of those multi-component complex materials appeared in various structures, including rock-salt, spinel, fluorite, perovskite, and delafossite. The samples were prepared either in powder or in bulk form; therefore, powder XRD was performed on samples with adjusted height. The samples were prepared using various synthesis techniques, mostly solid-state or wet chemical syntheses, to obtain the desired structures. Consequently, particle size and crystallinity varied significantly. The sample set also includes samples that were not successfully measured or where phases could not be identified.

The X-ray diffraction data were collected on a Bruker D8 Advance using a Cu radiation source or a STOE Stadi P diffractometer equipped with a Ga-jet X-ray source. The samples were initially recorded for various research projects over the last ten years and were measured with different step sizes, times per step, and over different angle ranges, but all using Cu K <sub>$\alpha$</sub>  or Ga K <sub>$\beta$</sub>  radiation. The samples mostly contained transition metal oxides, sulfides, and fluorides. To improve statistics, the samples were rotated during the entire measurement. Some air-sensitive samples were measured using a transparent polymer dome for protection. This dome led to increased background noise over the first 20° and slightly decreased pattern resolution.

### 3.5 | Institute of Catalysis Research and Technology, KIT

A variety of samples were analyzed, including commercial catalysts, bulk reference materials, porous metal oxide particles, and nanoparticles. The latter were synthesized via the surfactant-free benzyl alcohol route [75, 76]. The cobalt oxide (CoO or Co<sub>3</sub>O<sub>4</sub>) and cerium oxide (CeO<sub>2</sub>) nanoparticles were in the size range of 4–16 nm according to the Scherrer equation. A series of porous Al<sub>2</sub>O<sub>3</sub> materials, which were prepared by calcination of boehmite (AlOOH) at various temperatures, represents crystalline samples

with limited long-range structure and various contributions of  $\text{Al}_2\text{O}_3$  polymorphs.

X-ray diffraction (XRD) was conducted with an X'Pert Pro MPD (Panalytical) in Bragg-Brentano geometry using a Cu X-ray source. The patterns were acquired in the  $2\theta$  range of  $5-80^\circ$  with a step size of  $0.016711^\circ$  or  $0.033420^\circ$  and a total acquisition time of 40–120 min. This study has been carried out with the support of Angelina Barthelmeß, Elisabeth Herzinger, and Henning Hinrichs.

### 3.6 | Molecular Foundry Division & Advanced Light Source & Chemical Sciences Division, LBNL

In total, four different datasets were collected. The first dataset (LBNL-A) was collected from spin-coating and annealing triplecation metal-halide perovskite precursor solutions with the composition  $\text{Cs}_{0.05}(\text{MA}_{0.23}\text{FA}_{0.77})\text{Pb}_{1.1}(\text{I}_{0.77}\text{Br}_{0.23})_3$  onto various substrates. Here, MA stands for Methylammonium and FA stands for Formamidinium. The substrates onto which these solutions were coated include glass, which is amorphous, and GaAs wafers, which are single crystalline. Other substrates were stacks of glass/indium tin oxide, stacks of GaAs/CIGS, and stacks of glass/CIGS. Here, CIGS stands for a stack of Mo, Cu (In, Ga)  $\text{Se}_2$ , Cds, and ZnO. Some of the substrates were additionally covered with a self-assembling monolayer of MeO-2PACz. The GaAs substrates were prepared by Dr. Jiro Nishinaga from the National Institute of Advanced Industrial Science and Technology (AIST) in Japan [77] and the glass/CIGS substrates by Dr. Christian Kaufmann and his team at Helmholtz-Zentrum Berlin (HZB) in Germany [78]. Data collection was performed in situ during thin-film deposition using a custom-made spin-coating and annealing stage [79].

A second dataset (LBNL-B) was collected from spin-coating metal-halide perovskite precursor solutions with varying compositions of  $\text{MAPb}(\text{I}_{1-x}\text{Br}_x)_3$  spin-coated onto glass substrates. Here, MA = Methylammonium and  $x = 0, 0.33, 0.5, 0.67, 1$ . The substrates were preheated to different temperatures, including  $30^\circ\text{C}$ ,  $50^\circ\text{C}$ ,  $70^\circ\text{C}$ , and  $90^\circ\text{C}$ , and the spin-coating process was performed at a constant temperature on the preheated substrates. For both datasets, diffraction data were continuously measured during spin-coating, chemical induction of crystallization, and annealing of the samples, at  $100^\circ\text{C}$  and  $110^\circ\text{C}$  respectively. The diffraction data was recorded with a frequency of about 0.561/s and 0.541/s. Each in situ measurement consisted of about 500–1000 individual diffractograms. Depending on the substrate, each series of diffractograms shows an evolution from substrate only to a combination of polycrystalline perovskite,  $\text{PbI}_2$ , and substrate via several intermediate phases.

For these two datasets, experimental XRD data were collected at beamline 12.3.2 of the Advanced Light Source, the synchrotron at Lawrence Berkeley National Laboratory. The data were collected using a photon energy of 10 keV ( $\lambda = 1.23984 \text{ \AA}$ ), selected using a Si (111) monochromator. Measurements were taken in grazing incidence geometry, that is, using a beam incidence angle of  $1^\circ$ . Two-dimensional diffraction images were recorded using a Dectris Pilatus 1 M area detector at an angle between  $34^\circ$  and

$36^\circ$  with a sample-to-detector distance of roughly 190 mm. The two-dimensional data were calibrated using an  $\text{Al}_2\text{O}_3$  calibration standard and integrated along the azimuthal angle.

A third dataset (LBNL-C) was collected by observing the phase evolution of an Mn-Sb-O system with varying annealing temperatures. The temperatures used to analyze the crystal structure of the Mn-Sb-O system were chosen depending on the number of phase transitions appearing for a certain temperature range. Few changes in the crystal structure appear between room temperature and  $300^\circ\text{C}$  and phase transitions appeared from  $300^\circ\text{C}$  until  $850^\circ\text{C}$ . No phase transition appeared when cooling down. Therefore, the crystal structure was measured every  $100^\circ\text{C}$  between room temperature and  $300^\circ\text{C}$ ; every  $50^\circ\text{C}$  between  $300^\circ\text{C}$  and  $850^\circ\text{C}$ ; and every  $200^\circ\text{C}$  when cooling down. The heating and cooling rates were fixed for all the experiments at  $50^\circ\text{C}/\text{min}$  and the holding time was fixed to 2 min.

This data was collected using the in situ Rigaku-SmartLab3 kW diffractometer. This tool operates with SmartLab Studio II software, which can measure the X-ray diffraction during the annealing process. This enables directly showing all the phase transitions when annealing in various atmospheres such as  $\text{O}_2$ , Ar, and  $\text{NH}_3$ . Phase transitions are analyzed with the in situ XRD tool up to  $850^\circ\text{C}$  in this work. Most of the in situ experiments were performed under an air-like 20%  $\text{O}_2$  and 80% Ar environment (Ar flow: 50 sccm,  $\text{O}_2$  flow: 10 sccm). When a 100% Ar environment is fixed, an Ar flow of 60 sccm is input. The Bragg-Brentano (BB) mode is preferred in terms of geometry because it is more adapted in the analysis of scarce phases such as  $\text{MnSb}_2\text{O}_6$  rutile. The angular step used in the recording was 0.01 and the scanning rate was  $10/\text{min}$ .

A fourth dataset (LBNL-D) was collected from a two-step spin-coating process using metal-organic frameworks (MOFs) in perovskite precursor solutions, deposited onto glass substrates. In the first step, a nanoscale thiol-functionalized UiO-66-type Zr-based MOF ( $\text{UiO}-66-(\text{SH})_2$ ) was added to the  $\text{PbI}_2$  precursor. This was followed by the deposition of an organic mixture solution containing FAI, MACl, and MABr in the second step. The incorporation of MOFs aids in suppressing perovskite vacancy defects, thereby enhancing device stability and efficiency. To further investigate the influence of  $\text{UiO}-66-(\text{SH})_2$  on perovskite thin-film formation during the annealing process, a time-resolved GIWAXS experiment was conducted. The measurements were performed using a setup similar to that of LBNL-A and B.

## 4 | Usage

The opXRD database is hosted on Zenodo (<https://zenodo.org/records/14254270>) and can be downloaded by any user without any barriers or restrictions.

We also provide a Python library 'opxrd' to easily download and interface with the dataset. The opXRD library is designed for easy integration with common machine learning frameworks such as *PyTorch*. This makes it an ideal resource for researchers developing and benchmarking sim-to-real transfer strategies in

pXRD data analysis. The instructions for how to install this library can be found in the repository associated with the library. The repository of this library is located at <https://github.com/aimat-lab/opxrd>. The opxrd library includes options for data-loading, standardization, plotting, and the conversion to PyTorch tensors. We provide a Jupyter Notebook (<https://colab.research.google.com/github/aimat-lab/opXRD/blob/main/opxrd/usage.ipynb>) that showcases these functionalities in more detail. This notebook also illustrates how to interface with the opXRD database through Python.

## 5 | Summary and Outlook

With the opXRD database, a curation of 92,552 unlabeled and 2179 at least partially labeled experimental powder X-ray diffraction patterns from a wide range of different materials systems, we provide the largest currently available source of experimental XRD patterns. With this, we address the need for experimental data that arises when developing algorithms and analysis tools for pXRD data, both based on machine learning and classical approaches. The data can be used for the actual method development and for testing. Our dataset is a valuable and so far missing resource to drive further developments in the automated analysis of XRD data. Looking forward, opXRD is expected to play an important role in the development of advanced transfer learning approaches that integrate large-scale simulated data with real experimental patterns, ultimately driving the automation and accuracy of pXRD analysis. Future work will also include comprehensive benchmark evaluations to quantify the performance improvements achieved by incorporating opXRD into transfer learning pipelines.

Rather than a finished project, the opXRD database is an ongoing effort to collect experimental powder XRD data. We invite everyone working with experimental powder XRD to submit any data they would like to publicly share to the dataset, to further improve its utility and thus aid further developments in this field. Our submission page (<https://xrd.aimat.science/>) will continue to stay available for the submission of data. As new submissions come in, newer versions of the opXRD database incorporating these submitted datasets will be released. As the opXRD database grows further, we look forward to expanding this website to become a comprehensive community resource from which the database can be governed. Planned resources on this site include a contributor list, version list, and changelogs as well as comprehensive versioning, license, citation, and attribution practice statements.

We will keep updating and maintaining the dataset with new incoming submissions.

### Acknowledgments

H.S. acknowledges financial support by the German Research Foundation (DFG) through the Research Training Group 2450 “Tailored Scale-Bridging Approaches to Computational Nanoscience”. P.F. and D.H. acknowledge support by the Federal Ministry of Education and Research (BMBF) under Grant No. 01DM21001B (German-Canadian Materials Acceleration Center). J.Oe. and P.F.

acknowledge financial support from the Helmholtz Foundation Model Initiative within Project “SOL-AI”. Part of this work was funded under the France 2030 framework by Agence Nationale de la Recherche (project ANR-22-PEXD-0009 of PEPR DIADEM). Work at the Molecular Foundry was supported by the Office of Science, Office of Basic Energy Sciences, of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231. Work at the Advanced Light Source (ALS) was done at beamline 12.3.2. The ALS is a DOE Office of Science User Facility under contract no. DE-AC02-05CH11231. The development of the online phase identification platform is supported by the Guangzhou-HKUST (GZ) Joint Funding Program (No. 2023A03J0003). Work by the USC group was supported by the National Science Foundation (NSF) grant numbers DMR-2227178 and OISE-2106597. M.W. acknowledges funding by the Helmholtz Research Program “Materials and Technologies for the Energy Transition (MTET), Topic 3: Chemical Energy Carriers”. Work by the Empa group was supported by the Strategic Focus Area–Advanced Manufacturing (SFA–AM) through the project Advancing manufacturability of hybrid organic–inorganic semiconductors for large area optoelectronics (AMYS) as well as the Empa internal research call 2020. We thank BWCloud, funded by the Ministry of Science, Research and Arts Baden–Württemberg, for providing cloud server infrastructure.

### Conflicts of Interest

The authors declare no conflicts of interest.

### Data Availability Statement

The opXRD database is hosted on Zenodo (<https://zenodo.org/records/14254270>) and can be downloaded by any user without any barriers or restrictions.

### References

1. Y. Liu, Z. Hu, Z. Suo, et al., *Science China Technological Sciences* 62 (2019): 521.
2. B. P. MacLeod, F. G. L. Parlane, T. D. Morrissey, et al., *Science Advances* 6 (2020): 20–eaaz8867.
3. A. Ludwig, *npj Computational Materials* 5 (2019): 1.
4. Y. Ozaki, Y. Suzuki, T. Hawai, K. Saito, M. Onishi, and K. Ono, *npj Computational Materials* 6 (2020): 1.
5. R. E. Dinnebier, A. Leineweber, and J. S. O. Evans, *Rietveld Refinement: Practical Powder Diffraction Pattern Analysis Using TOPAS* (De Gruyter, 2019).
6. D. A. F. Cano, A. R. C. Quispe, R. R. Vellamin, J. A. O. Anticona, J. González, and J. A. R. Guivar, *Revista de Investigación de Física* (2021).
7. L. B. McCusker, R. B. Von Dreele, D. E. Cox, D. Louër, and P. Scardi, *Journal of Applied Crystallography* 32, no. 1 (1999): 36.
8. A. Agrawal and A. Choudhary, *MRS Communications* 9 (2019): 779.
9. V.–A. Surdu and R. György, *Applied Sciences* 13 (2023): 9992.
10. Z. Feng, Q. Hou, Y. Zheng, et al., *Computational Materials Science* 156 (2019): 310.
11. H. Wang, Y. Xie, D. Li, et al., *Journal of Chemical Information and Modeling* 60, no. 4 (2020): 2004.
12. W. Park, J. Chung, J. Jung, et al., *IUCrJ* 4 (2017): 486.
13. B. D. Lee, J.-W. Lee, J. Ahn, S. Kim, W. Park, and K. Sohn, *Advanced Intelligent Systems* 5 (2023): 2300140.
14. H. Schopmans, P. Reiser, and P. Friederich, *Digital Discovery* 2, no. 5 (2023): 1414–1424.
15. D. Chen, Y. Bai, S. Ament, et al., *Nature Machine Intelligence* 3, no. 9 (2021): 812.



16. M.-C. Chang, S. Ament, M. Amsler, et al., arXiv:2308.07897, 2023.
17. H. Dong, K. Butler, D. Matras, et al., *npj Computational Materials* 7 (2021): 1.
18. S. R. Chitturi, D. Ratner, R. C. Walroth, et al., *Journal of Applied Crystallography* 1799 (2021): 54.
19. S. Habershon, E. Cheung, K. Harris, and R. Johnston, *Journal of Physical Chemistry A* 108 (2004): 711.
20. S. Zhang, B. Cao, T. Su, et al., *IUCrJ* 11, no. Pt 4 (2024): 634.
21. B. Cao, Y. Liu, Z. Zheng, R. Tan, J. Li, and T.-y. Zhang, arXiv preprint arXiv:2406.15469, 2024.
22. F. Oviedo, Z. Ren, S. Sun, et al., *npj Computational Materials* 5 (2018): 1.
23. P. M. Vecsei, K. Choo, J. Chang, and T. Neupert, *Physical Review B* 99 (2019): 245120.
24. A. N. Zaloga, V. V. Stanovov, O. E. Bezrukova, P. S. Dubinin, and I. S. Yakimov, *Materials Today Communications* 25 (2020): 101662.
25. Y. Suzuki, H. Hino, T. Hawaii, K. Saito, M. Kotsugi, and K. Ono, *Scientific Reports* 10 (2020): 21790.
26. A. Chakraborty and R. Sharma, *The Visual Computer* 38 (2021): 1275.
27. B. Lafuente, R. T. Downs, H. Yang, and N. Stone, *1. The Power of Databases: The RRUFF Project* (De Gruyter, 2015).
28. T. Armbruster and R. M. Danisi, *Highlights in Mineralogical Crystallography* (De Gruyter, 2015).
29. S. Gates-Rector and T. Blanton, *Powder Diffraction* 34 (2019): 352.
30. P. Villars, K. Cenzual, R. Gladyshevskii, and S. Iwata, *Chemistry of Metals and Alloys* 3, no. 4 (2018): 43.
31. J. E. Salgado, S. Lerman, Z. Du, C. Xu, and N. Abdolrahim, *npj Computational Materials* 9, no. 1 (2023): 214.
32. Y. Waseda, E. Matsubara, and K. Shinoda, *X-Ray Diffraction Crystallography* (Springer, 2011).
33. V. Pecharsky and P. Zavalij, *Fundamentals of Powder Diffraction and Structural Characterization of Materials* (Springer, 2023).
34. B. S. Hulbert and W. M. Kriven, *Journal of Applied Crystallography* 56, no. 1 (2023): 160.
35. F. Zhuang, Z. Qi, K. Duan, et al., *Proceedings of the IEEE* 109 (2021): 43.
36. L. A. Gatys, A. S. Ecker, and M. Bethge, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016): 2414–2423.
37. Y. Ganin and V. Lempitsky, *Proceedings of the 32nd International Conference on Machine Learning* 37 (2015): 1180–1189.
38. K. Seddiki, P. Saudemont, F. Precioso, et al., *Nature Communications* 11 (2020): 5595.
39. M. Aranda, *Journal of Applied Crystallography* 51 (2018): 1739–1744.
40. L. M. J. Kroon-Batenburg, M. P. Lightfoot, N. T. Johnson, and J. R. Helliwell, *Structural Dynamics* 11 (2024): 011301.
41. J.-W. Lee, W. B. Park, J. H. Lee, S. P. Singh, and K.-S. Sohn, *Nature Communications* 11 (2020): 1–86.
42. B. D. Lee, J.-W. Lee, W. B. Park, et al., *Advanced Intelligent Systems* 4 (2022): 7–2200042.
43. J. R. Hattrick-Simpers, B. DeCost, A. G. Kusne, et al., *Integrating Materials and Manufacturing Innovation* 10, no2 (2021): 311.
44. J. Schuetzke, S. Schweidler, F. R. Muenke, et al., *Advanced Intelligent Systems* 6, no. 3 (2024): 2300501.
45. “P. F. project,” *Linus Pauling File Product Descriptions*, accessed November 27, 2024, <https://web.archive.org/web/20240221221553/https://paulingfile.com/index.php?p=products#PAULING%20FILE%20products>.
46. “A. International,” *Pearson’s Crystal Data Product Description*, accessed November 27, 2024, <https://web.archive.org/web/20240617123612/https://www.crystalimpact.com/pcd/>.
47. P. Villars, *Mpds Access Link*, accessed December 04, 2024, <https://mpds.io/#start>.
48. “C. Impact,” *Pearson’s Crystal Data Product Offering*, accessed December 10, 2024, <https://shop-crystalimpact.de/en/p/pearson-s-crystal-data-one-year-single-license>.
49. “P. F. project,” *Mpds Api Product Description*, accessed December 10, 2024, <https://mpds.io/#products>.
50. “ICDD,” *Pdf5 Product Description*, accessed November 27, 2024, <https://www.icdd.com/pdf-5/>.
51. “ICDD,” *Pdf5+ license*, accessed March 07, 2025, <https://www.icdd.com/licensing-process/\#1528471154226-933e5cc6-8da7>.
52. U o A Department of Geosciences, Ruff access link, accessed November 27, 2024, [https://web.archive.org/web/20241007175010/https://rruff.info/about/about\\_general.php](https://web.archive.org/web/20241007175010/https://rruff.info/about/about_general.php).
53. “C. maintainers,” *Crystallography Open Database*, accessed November 27, 2024, <https://www.crystallography.net/cod/>.
54. S. Gražulis, D. Chateigner, R. T. Downs, et al., *Journal of Applied Crystallography* 42, no. 4 (2009): 726–729.
55. A. L. Bail, *Powbase*, accessed March 07, 2025, <http://www.cristal.org/powbase/index.html>.
56. A. Zakutayev, N. Wunder, M. Schwarting, et al., *Scientific Data* 5 (2018): 1.
57. “N. R. E. Laborator,” *High-Throughput Experimental Database Statistics*, accessed March 07, 2025, <https://hitem.nrel.gov/stats>.
58. F. Karlsruhe, *Icsd Access Link*, accessed November 27, 2024, <https://icsd.products.fiz-karlsruhe.de/>.
59. “C. C. D. Centre,” *Cambridge Structural Database Access Link*, accessed November 27, 2024, <https://www.ccdc.cam.ac.uk/structures/>.
60. “M. Project,” *Materials Project Database Website Access Link*, accessed November 27, 2024, <https://next-gen.materialsproject.org/>.
61. “R. A. o. S. Institute of Experimental Mineralogy,” *Crystallographic and Crystallochemical Database Website Access Link*, accessed November 27, 2024, <https://database.iem.ac.ru/mincryst/index.php>.
62. “U of the Basque Country,” *Bilbao Incommensurate Crystal Structure Database Access Link*, accessed November 27, 2024, <https://www.cryst.ehu.es/bincstrdb/search/>.
63. D. Barthelmy, *Mineralogy Database Access Link*, accessed November 27, 2024, <https://webmineral.com/>.
64. “I. U. of Crystallography (IUCr),” *Iucr Raw Data Letters Access Link*, accessed November 27, 2024, <https://iucrdata.iucr.org/x/index.html>.
65. “U. N. R. Laboratory,” *Crystal Lattice-Structures Access Link*, accessed November 27, 2024, <https://www.atomic-scale-physics.de/lattice/>.
66. P. Perroud, *Athena Mineral Database Access Link*, accessed November 27, 2024, <https://athena.unige.ch/athena/mineral/mineral.html>.
67. “R. C. for Structural Bioinformatics,” *Protein Data Bank Access Link*, accessed November 27, 2024, <https://www.rcsb.org/>.
68. W. Setyawan and S. Curtarolo, *Computational Materials Science* 49, no. 2 (2010): 299.
69. I. T. Jolliffe and J. Cadima, “Philosophical Transactions of the Royal Society A: Mathematical,” *Physical and Engineering Sciences* 374 (2016): 20150202.

70. S. Gražulis, D. Chateigner, R. Downs, et al., *Journal of Applied Crystallography* 42 (2009): 726.
71. A. Vaitkus, A. Merkys, T. Sander, et al., *Journal of Cheminformatics* 15, no. 1 (2023): 123.
72. S. Zhuk, A. A. Kistanov, S. C. Boehme, et al., *Chemistry of Materials* 33, no. 23 (2021): 9306.
73. A. Wiczorek, H. Lai, J. Pious, F. Fu, and S. Siol, *Advanced Materials Interfaces* 10, no. 7 (2023): 2201828.
74. A. Wiczorek, A. G. Kuba, J. Sommerhäuser, L. N. Caceres, C. M. Wolff, and S. Siol, *Journal of Materials Chemistry A* 12 (2024): 7025.
75. M. Wolf, S. J. Roberts, W. Marquart, et al., *Dalton Transactions* 48 (2019): 36–13858.
76. M. Wolf, N. Fischer, and M. Claeys, *Materials Chemistry and Physics* 213 (2018): 305.
77. J. Nishinaga, T. Nagai, T. Sugaya, H. Shibata, and S. Niki, *Applied Physics Express* 11, no. 8 (2018): 082302.
78. M. D. Heinemann, R. Mainz, F. Österle, et al., *Scientific Reports* 7 (2017): 1.
79. T. Song, Z. Yuan, M. Mori, et al., *Advanced Functional Materials* 30 (2019): 6.

### Supporting Information

Additional supporting information can be found online in the Supporting Information section.