SimXRD-4M : Datasheet

June 8, 2024

1 SimXRD-4M

We introduce SimXRD-4M, the largest open-source simulated X-ray diffraction (XRD) pattern dataset, designed to accelerate the development of crystallographic informatics. SimXRD-4M consists of 4,065,346 simulated powder X-ray diffraction patterns, representing 119,569 distinct crystal structures under 33 different simulated conditions that mimic real-world variations. This dataset underscores the academic significance and engineering innovation of simulated XRD patterns within this interdisciplinary field.

2 Datasheet

Motivation

For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

The dataset was created to address a fundamental task in crystallography: precisely identifying crystal types by analyzing XRD patterns across various physical conditions. Traditionally, this involves a search-match process [1] that iterates through numerous known powder diffraction patterns until a satisfactory alignment is achieved with the target XRD pattern. While prevalent, this method is time-consuming [1, 15, 9].

To address these challenges, recent studies [12, 23, 16, 27, 11, 24, 4, 20] have trained neural networks on simulated XRD datasets, treating XRD patterns as sequences and aiming to classify them according to specific symmetries (e.g., crystal systems or space groups). Despite significant progress, previous research has largely focused on specific materials [6, 12, 27]. For instance, Lee et al. [12] concentrated on mixtures of 38 distinct binary and ternary crystals in the Sr-Li-Al-O inorganic compounds.

Furthermore, the geometric similarity of crystals and various experimental physical factors can result in XRD patterns with similar peak distributions. Therefore, to develop models with high generalization ability, it is crucial to have sufficient coverage of structures and close alignment with experimental patterns under different physical conditions.

Creating such a high-fidelity simulated XRD pattern dataset is necessary yet challenging due to the numerous domain-specific simulation processes involved, such as determining crystal stability, instrumental settings, and physical environments.
SimXRD-4M was developed to fill this gap.

Who created this dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

Bin Cao[1,2] Yang Liu[1,3], Zinan Zheng[1], Ruifeng Tan[1,2], Jia Li[1,3], Tong-yi Zhang[1,2]
[1]Hong Kong University of Science and Technology (Guangzhou)
[2]Guangzhou Municipal Key Laboratory of Materials Informatics
[3]Hong Kong University of Science and Technology

Composition

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.
Each row in the SimXRD dataset represents a diffraction pattern, consisting of 3051 pairs of values representing lattice plane distance (d) and diffraction intensity (I). Additionally, each instance includes information such as chemical formula (string), atomic elements (list), space group (integer), and crystal system (integer).

How many instances are there in total (of each type, if appropriate)?
The SimXRD dataset comprises 4,065,346 simulated powder XRD patterns, each associated with a chemical formula, a set of atomic elements, a space group value, and a crystal system value.

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).
SimXRD-4M does not contain all possible instances of XRD patterns. However, it covers all the crystals contained in the latest Material Project [10] (almost all inorganic crystals), denoted as MP-2024.1. The diffraction pattern for a given crystal varies depending on various physical settings. We provided 33 coupling conditions for diffraction. Since many physical settings affecting diffraction patterns are continuous values, it is impossible to provide all real situations. The specific settings are determined based on our domain knowledge and experience in XRD practice.

What data does each instance consist of? "Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description.
Each instance in the dataset consists of features extracted from simulated XRD patterns. Below is an example of the Python code used to access and process the data:
The data includes:

- `element`: A list of chemical symbols representing the elements in the crystal (e.g., `['C', 'H', 'O']`).
- `latt_dis`: A list of lattice plane distances.
- `intensity`: A list of diffraction intensities.
- `spg`: An integer representing the space group number.
- `crysystem`: An integer representing the crystal system number.

```python
from ase.db import connect
databs = connect("./binxrd.db")
for row in databs.select():
    atoms = row.toatoms()
    element = atoms.
    get_chemical_symbols()
    latt_dis = eval(getattr(row, '
    latt_dis'))
    intensity = eval(getattr(row,
    'intensity'))
    spg = eval(getattr(row, 'tager
    '))[0]
    crysystem = eval(getattr(row,
    'tager'))[1]
```

Listing 1: Example Python code

Is there a label or target associated with each instance? If so, please provide a description.

Formally, SimXRD considers the following multi-class sequence classification problem. Given the XRD pattern $X = [x_1, x_2, \cdots, x_n] \in R^n$ where x_i is the i-th intensity value and X is arranged by lattice plane distance, the objective is to predict its symmetry $Y \in R^k$. Here n is the feature dimension, which is 3501 in our dataset. $k = 7$ for crystal system classification and $k = 230$ for space group classification.

SimXRD has clearly categorized labels of crystal systems (integers, 1-7) and space groups (integers, 1-230).

Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.

No

Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.

We recommend two types of data splits: (1) As shown in our benchmark, the dataset is split according to different simulation environments. Both training and testing datasets contain the same structures but under different simulation environments, corresponding to the in-library identification in XRD phase identification. (2) We recommend splitting the dataset according to crystal types. For both training and testing datasets, they contain different structures, corresponding to the out-library identification in XRD phase identification. We provide the code for generating such splitting based on SimXRD OutlibDataProcessor.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide

descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.
SimXRD relies on the crystal database. Currently, it contains all the recorded data in the Material Project (MP) crystal database. The MP database is open-sourced. Further development of SimXRD may require supplementing new crystal structures after a certain period when significant new crystals are added to the MP database.

Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals non-public communications)? If so, please provide a description.
No

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.
No

Does the dataset relate to people? If not, you may skip the remaining questions in this section.
No

Does the dataset identify any subpopulations (e.g., by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.
No

Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? If so, please describe how.
No

Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.
No

Collection Process

How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

Crystal data The crystal structures utilized in this study were sourced from Material Project [10], a comprehensive, searchable database containing information about solid-state materials and molecules. It provides detailed data on these materials' physical properties, such as elastic tensors, band structures, and formation energies, derived from electronic structure calculations, offering a more comprehensive reference for crystal studies. We utilized the latest dataset, denoted as MP-2024.1, encompassing a to-

tal of 154,718 crystallographic structures of January 2024.

Crystal Filtering To ensure consistency between recorded space group numbers (crystal systems) and atomic arrangements, and to enhance the quality of crystal structures, we conducted a thorough examination of each structure in the MP database. We utilized Spglib [25], a library designed for identifying and managing crystal symmetries, to check all structures. Structures displaying broken symmetry, duplication, or discrepancies in space groups were excluded. Additionally, only structures containing up to 500 atoms per lattice cell were retained, effectively encompassing nearly all inorganic materials in the MP dataset. A total of 119,569 crystal structures were screened.

Simulation in SimXRD The XRD patterns are determined by both the crystal's intricate structure and practical factors including the state of the specimen and the instrumental parameters, are simulated by custom software WPEM [2]

What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? How were these mechanisms or procedures validated?

The crystal database MP and screening package Spglib are open-sourced. The simulation software is WPEM, homemade and not completely open-sourced yet, but its effectiveness has been recognized by multiple peer researches in material and physical communities[13, 3, 22, 19, 7]. Additionally, the details of the simulation process are provided in mathematics in our paper.

Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.

Crystal data were retrieved from the MP database in January 2024. The simulated data were generated during the period from January to May 2024.

Were any ethical review processes conducted (e.g., by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.

No ethical problems were involved.

Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?

The crystal database MP is open for acquisition.

Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.

Not applicable

If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).

Not applicable

Preprocessing/cleaning/labeling

Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remainder of the questions in this section.

All spectrum data contained in SimXRD provide full information with XRD features and labels. We did not perform any further preprocessing of the data.

Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.

We did not perform any preprocessing of the data beyond what is contained in SimXRD.

Is the software used to preprocess/clean/label the instances available? If so, please provide a link or other access point.

We provide a data loader associated with our dataset as dataloader.

Uses

Has the dataset been used for any tasks already? If so, please provide a description.

Although the dataset has not been utilized for specific tasks, we offer some benchmarks to illustrate its potential in library space group and crystal system classification tasks.

We evaluate three types of sequence classification models:

- **CNN-based Models**: We assess all existing CNN architectures proposed for symmetry identification. As many models lack specific names, we classify them based on their convolution and pooling layers, and whether they employ ensemble learning and dropout layers.
- **Recurrent Models**: We select three fundamental recurrent neural networks: RNN [17], LSTM [8], and GRU [5]. Additionally, since XRD patterns can be analyzed from both forward and backward perspectives, we evaluate the bidirectional performance of RNNs, LSTMs, and GRUs [21].
- **Transformers**: Transformers have demonstrated effectiveness in various sequence modeling tasks. We evaluate the performance of raw transformers [26] along with two advanced transformer models - iTransformer [14] and PatchTST [18].

In addition to these models, we include MLP as a straightforward baseline. For recurrent models and transformers, we first use them to learn sequence representations and then employ an MLP to predict symmetry.
For further details, please refer to our paper.

Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.

Please refer to SimXRD-4M.

Is there anything about the composition of the dataset or the way it

was collected and preprocessed/-cleaned/labeled that might impact future uses? For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?
Not applicable.

Distribution

Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.
No

How will the dataset be distributed (e.g., tarball on website, API, GitHub) Does the dataset have a digital object identifier (DOI)?
The dataset can be acquired from the website SimXRD-4M.

When will the dataset be distributed?
Starting from June 1, 2024.

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.
No

Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.
No

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.
No, the dataset is open for any application.

Maintenance

Who will be supporting/hosting/-maintaining the dataset?
Mr. Cao Bin, conducting research at Hong Kong University of Science and Technology (Guangzhou).

How can the owner/curator/manager of the dataset be contacted (e.g., email address)?
Email address: bcao686@connect.hkust-gz.edu.cn

Is there an erratum? If so, please provide a link or other access point.
No

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?
If any updates are necessary in future development, we will provide essential

updates and communicate all revisions and changes through the database website SimXRD-4M. We will provide the updated version and maintain a history of versions.

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.

Not applicable

Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to users.

All older versions of the dataset will continue to be supported/hosted/maintained. The database website SimXRD-4M will record all historical datasets and the newest dataset versions along with commit histories.

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to other users? If so, please provide a description.

Yes, we accept user contributions to extend/augment/build on/contribute to the dataset. If any user wishes to contribute, they need to provide crystal data and simulation software for strict screening by us to ensure data quality. The application can be launched by submitting a pull request on our dataset GitHub repository SimXRDGithub or by contacting the author directly for data merging.

References

[1] Angela Altomare, Corrado Cuocci, Carmelo Giacovazzo, Anna Moliterni, and Rosanna Rizzi. Qualx: a computer program for qualitative analysis using powder diffraction data. *Journal of Applied Crystallography*, 41(4):815–817, 2008.

[2] Bin Cao. Whole pattern fitting of powder x-ray diffraction by expectation maximum algorithm, 2024.

[3] Bin Cao, Tianhao Su, Shuting Yu, Tianyuan Li, Taolue Zhang, Jincang Zhang, Ziqiang Dong, and Tong-Yi Zhang. Active learning accelerates the discovery of high strength and high ductility lead-free solder alloys. *Materials & Design*, 241:112921, 2024.

[4] Kamal Choudhary, Brian DeCost, Chi Chen, Anubhav Jain, Francesca Tavazza, Ryan Cohn, Cheol Woo Park, Alok Choudhary, Ankit Agrawal, Simon JL Billinge, et al. Recent advances and applications of deep learning methods in materials science. *npj Computational Materials*, 8(1):59, 2022.

[5] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.

[6] Siti Fatimah, Risti Ragadhita, Dwi Fitria Al Husaeni, and Asep Bayu Dani Nandiyanto. How to calculate crystallite size from x-ray diffraction (xrd) using scherrer method. *ASEAN Journal of Science and Engineering*, 2(1):65–76, 2022.

[7] Amy XY Guo, Bin Cao, Zihan Wang, Xiao Ma, and Shan Cecilia Cao. Fabricated high-strength, low-elastic modulus biomedical ti-24nb-4zr-8sn alloy via powder metallurgy. *Materials*, 16(10):3845, 2023.

[8] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[9] International Centre for Diffraction Data. Mdi jade software. https://www.icdd.com/mdi-jade/. Accessed: 2024-05-18.

[10] Anubhav Jain, Shyue Ping Ong, Geoffroy Hautier, Wei Chen, William Davidson Richards, Stephen Dacek, Shreyas Cholia, Dan Gunter, David Skinner, Gerbrand Ceder, et al. Commentary: The materials project: A materials genome approach to accelerating materials innovation. *APL materials*, 1(1), 2013.

[11] Byung Do Lee, Jin-Woong Lee, Junuk Ahn, Seonghwan Kim, Woon Bae Park, and Kee-Sun Sohn. A deep learning approach to powder x-ray diffraction pattern analysis: Addressing generalizability and perturbation issues simultaneously. *Advanced Intelligent Systems*, 5(9):2300140, 2023.

[12] Jin-Woong Lee, Woon Bae Park, Jin Hee Lee, Satendra Pal Singh, and Kee-Sun Sohn. A deep-learning technique for phase identification in multiphase inorganic compounds using synthetic xrd powder patterns. *Nature communications*, 11(1):86, 2020.

[13] Tongxing Lei, Bin Cao, Wenbo Fu, Xiuling Shi, Zhiyu Ding, Qi Zhang, Junwei Wu, Kaikai Li, and Tong-Yi Zhang. A li-rich layered oxide cathode with remarkable capacity and prolonged cycle life. *Chemical Engineering Journal*, page 151522, 2024.

[14] Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long. itransformer: Inverted transformers are effective for time series forecasting. *CoRR*, abs/2310.06625, 2023.

[15] Binfeng Lv, Zhenjie Feng, Xiaowei Sun, Jian Cao, Yu Gao, Shihui Chang, Cheng Dong, and Jincang Zhang. iqual: a new computer program for qualitative phase analysis with efficient search–match capability. *Journal of Applied Crystallography*, 57(2), 2024.

[16] Phillip M Maffettone, Lars Banko, Peng Cui, Yury Lysogorskiy, Marc A Little, Daniel Olds, Alfred Ludwig, and Andrew I Cooper. Crystallography companion agent for high-throughput materials discovery. *Nature Computational Science*, 1(4):290–297, 2021.

[17] Larry Medsker and Lakhmi C Jain. *Recurrent neural networks: design and applications*. CRC press, 1999.

[18] Yuqi Nie, Nam H. Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023.

[19] Yin Qin, Bin Cao, Xiao-Ye Zhou, Zhuorui Xiao, Hanxiang Zhou, Zhenyi Zhao, Yibo Weng, Jianshuai Lv, Yang Liu, Yan-Bing He, et al. Orthorhombic (ru, mn) 2o3: A superior electrocatalyst for acidic oxygen evolution reaction. *Nano Energy*, 115:108727, 2023.

[20] Jerardo E Salgado, Samuel Lerman, Zhaotong Du, Chenliang Xu, and Niaz Abdolrahim. Automated classification of big x-ray diffraction data using deep learning models. *npj Computational Materials*, 9(1):214, 2023.

[21] Mike Schuster and Kuldip K Paliwal. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, 45(11):2673–2681, 1997.

[22] Linlin Sun, Bin Cao, Qingshuang Ma, Qiuzhi Gao, Jiahao Luo, Minglong Gong, Jing Bai, and Huijun Li. Machine learning-assisted composition design of w-free co-based superalloys with high γ-solvus temperature and low density. *Journal of Materials Research and Technology*, 29:656–667, 2024.

[23] Nathan J Szymanski, Christopher J Bartel, Yan Zeng, Qingsong Tu, and Gerbrand Ceder. Probabilistic deep learning approach to automate the interpretation of multi-phase diffraction spectra. *Chemistry of Materials*, 33(11):4204–4215, 2021.

[24] Leslie Ching Ow Tiong, Jeongrae Kim, Sang Soo Han, and Donghun Kim. Identification of crystal symmetry from noisy diffraction patterns by a shape analysis and deep learning. *npj Computational Materials*, 6(1):196, 2020.

[25] Atsushi Togo, Kohei Shinohara, and Isao Tanaka. `Spglib`: a software library for crystal symmetry search, 2024.

[26] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008, 2017.

[27] Hong Wang, Yunchao Xie, Dawei Li, Heng Deng, Yunxin Zhao, Ming Xin, and Jian Lin. Rapid identification of x-ray diffraction patterns based on very limited data by interpretable convolutional neural networks. *Journal of chemical information and modeling*, 60(4):2004–2011, 2020.