SimXRD-4M: Big Simulated X-ray Diffraction Data and Crystal Symmetry Classification Benchmark

Bin Cao [1], Yang Liu[1,2], Zinan Zheng[1], Ruifeng Tan[1], Jia Li[*1,2], Tong-yi Zhang[*1]

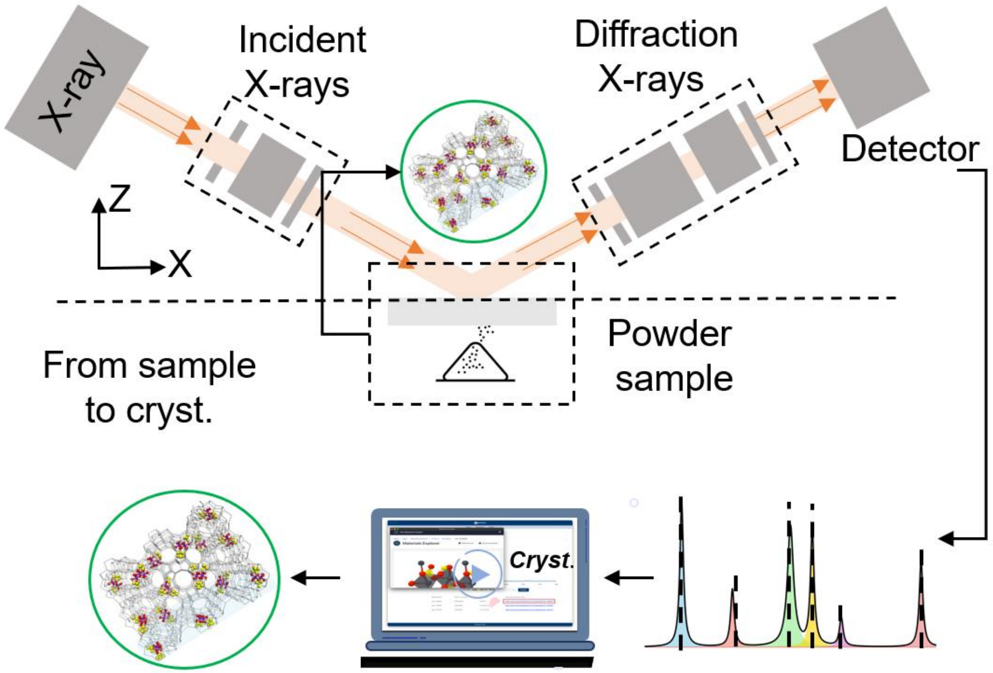
[1] HKUST(GZ) [2] HKUST

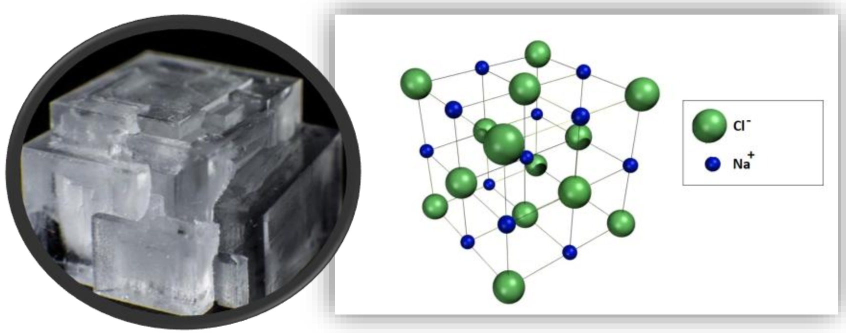
香港科技大学(广州)
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

香港科技大学
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

Motivation

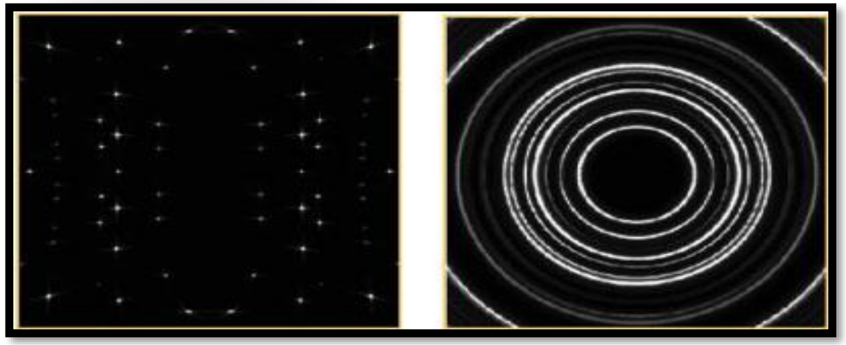
- The AI Lab will pioneer new methods for material synthesis.
- Powder XRD is crucial for detecting crystal structures.
- Intelligent analysis and corresponding software systems must undergo a new revolution.





Matter & Structure

Diffraction pattern

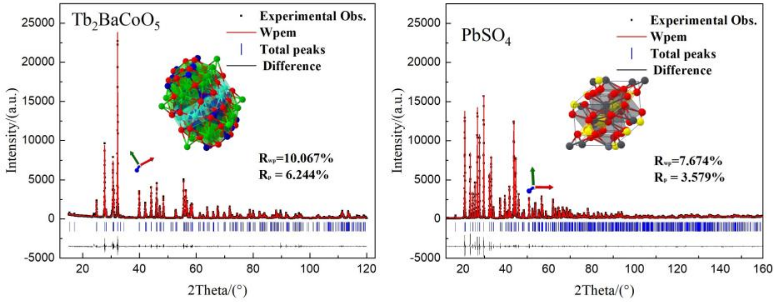


Detection results

Motivation

Table: Summaries of existing powder XRD datasets. ICSD refers to the commercial Inorganic Crystal Structure Database. MP denotes the open-sourced Material Project.

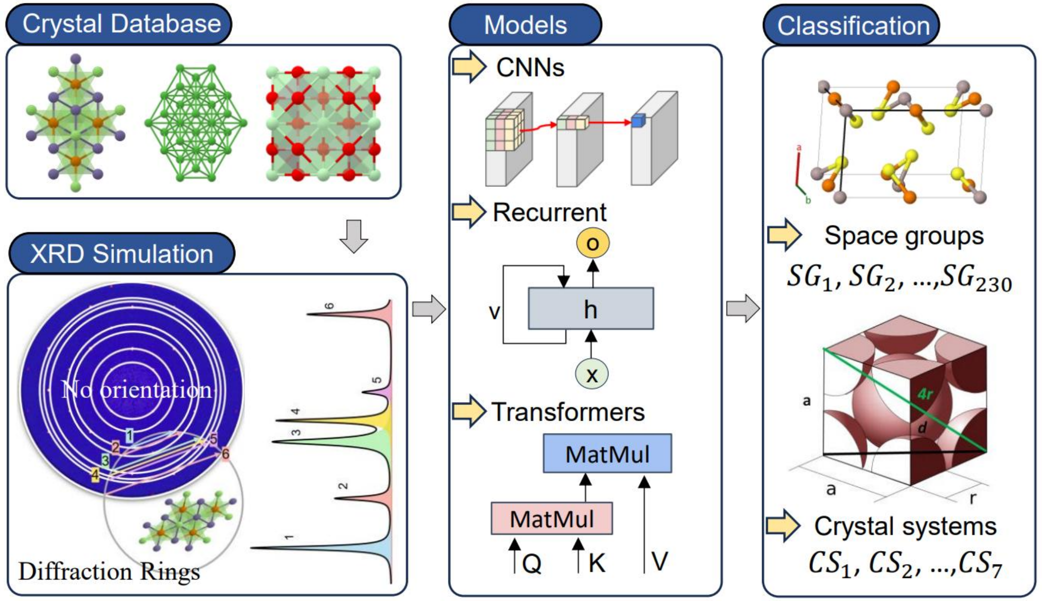
Dataset	#XRD Pattern	#Structure	Open Access	Simulated	Crystal Source	Year
RRUFF (Lafuente et al., 2015)	3,002	3,002	✓	×	-	2015
XRDSP (Suzuki et al., 2020)	169,536	169,536	✓	✓	ICSD	2020
CNN (Lee et al., 2020)	1,785,405	170	×	✓	ICSD	2020
PQNet (Dong et al., 2021)	250,000	1	✓	✓	ICSD	2021
XRDAutoAnalyzer (Szymanski et al., 2021)	38,250	150	✓	✓	ICSD	2021
XRDIsAllYouNeed (Lee et al., 2022)	328,503	189,476&139,027[1]	×	✓	ICSD&MP	2022
AdvancedXRDAnalysis (Lee et al., 2023)	29,569,650	197,131	×	✓	ICSD	2023
CrySTINet (Chen et al., 2024)	100	100	✓	✓	ICSD	2024
CPICANN (Cao, 2024)	692,190	23,073	✓	✓	COD	2024
SimXRD	**4,065,346**	**119,569**	✓	✓	**MP**	2024



Lack of a large-scale and high-quality dataset

Insufficient evaluation on real XRD patterns

Limited exploration on out-library identification



Contributions

Https://github.com/Bin-Cao/SimXRD

- Novel method to generate high-fidelity, million-scale simulated XRD patterns for machine learning.
- Revealing the long-tail distribution of crystals and proposing insights to enhance minority accuracy.
- Providing solutions for studying crystal out-of-library generalization.
- Experimental validation demonstrating robust generalization.

Results | Simulation

$$F = \sum_{j=1}^{N} f_j e^{2\pi i \mathbf{G}^* \cdot \mathbf{R}}$$

grain size

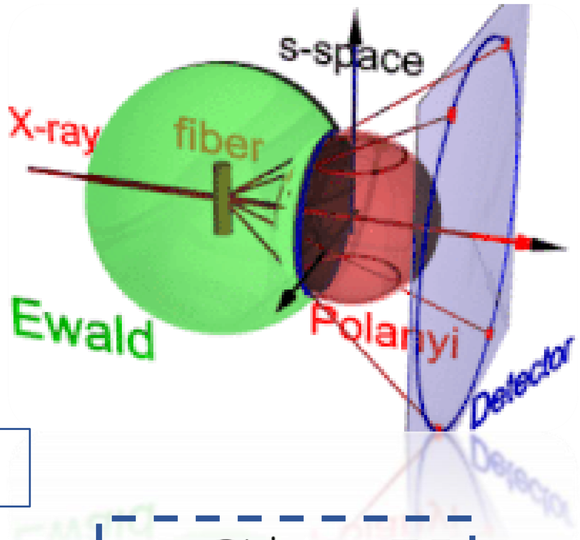
internal stress

$$d = 1/|\mathbf{G}^*|$$

temperature variations

$$\frac{6h^2T}{mk\Theta^2}\left(\phi\left(\frac{\Theta}{T}\right)+\frac{\Theta}{4T}\right)\left(\sin^2\theta\right)/\lambda$$

grain orientation



instrument drift

$$L = \frac{1+\cos^2 2\theta}{\sin^2\theta\cos\theta}$$

instrument noise

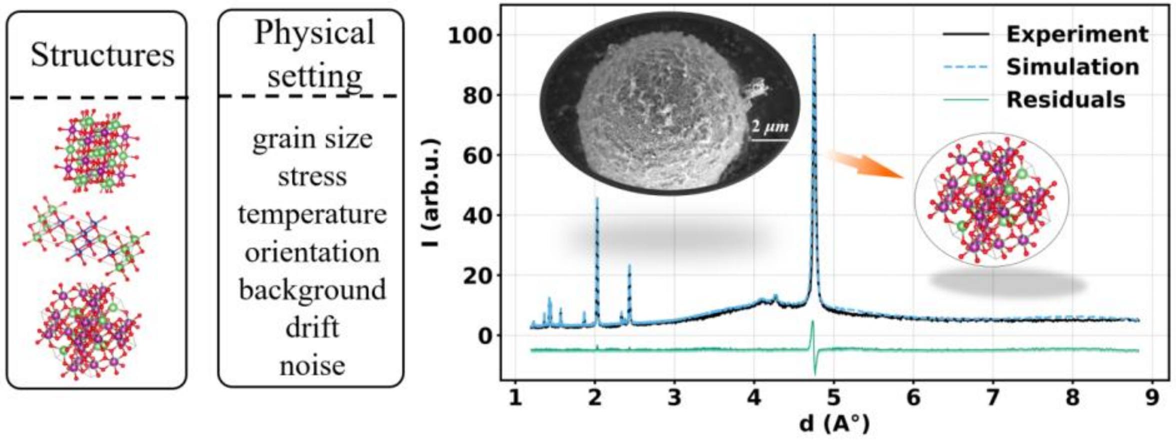
$$y(x) = W * G * S$$

detector geometry

$$\frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{\infty}\left[\frac{\gamma}{(2\theta-t)^2+\gamma^2}\right]\exp\left(-\frac{(2\theta-t)^2}{2\sigma^2}\right)dt$$

scattering induced background

$$\boxed{2\pi\mathbf{G}^* = \mathbf{K}}$$



Contributions

Experimental XRD pattern of a Li-rich layered oxide cathode (Li_2MnO_3) was compared with simulated pattern generated using PysimXRD. The simulation incorporates multiphysical coupling, producing patterns that closely match experimental measurement with minimal residual errors.

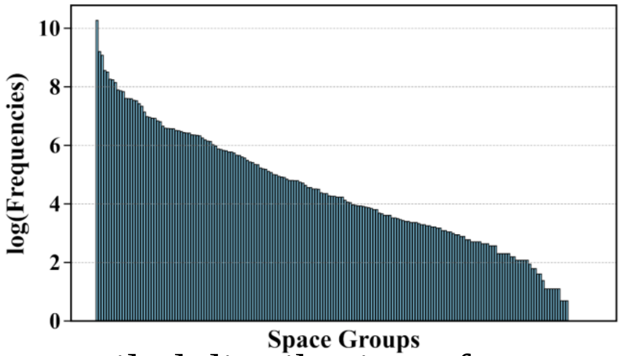
Results | In /out library crystal symmetry classification

Model	# Conv.	# Dropout	# Pooling	Ensemble	Ref.	Crystal System Accuracy	F1	Precision	Recall	Time (ms)	Space Group Accuracy	F1	Precision	Recall	Time (ms)
CNN1	3	✓	AvgPool	×	(Park et al., 2017)	0.559	0.418	0.427	0.431	3.2	0.241	0.002	0.001	0.005	3.4
CNN2	2	×	MaxPool	×	(Lee et al., 2020)	0.466	0.236	0.222	0.283	1.8	0.241	0.002	0.001	0.005	1.9
CNN3	3	×	MaxPool	×	(Lee et al., 2020)	0.531	0.328	0.315	0.369	3.5	0.241	0.002	0.001	0.005	3.6
CNN4	7	✓	MaxPool	×	(Wang et al., 2020)	0.316	0.069	0.045	0.143	1.4	0.241	0.002	0.001	0.005	1.4
CNN5	3	✓	AvgPool	✓	(Maffettone et al., 2021)	0.517	0.394	0.489	0.378	0.6	0.295	0.022	0.025	0.026	0.6
CNN6	7	✓	MaxPool	×	(Dong et al., 2021)	0.316	0.069	0.045	0.143	65.0	0.241	0.002	0.001	0.005	65.0
CNN7	6	✓	MaxPool	✓	(Szymanski et al., 2021)	0.862	0.863	0.887	0.845	6.8	0.588	0.124	0.147	0.130	6.9
CNN8	14	✓	MaxPool	×	(Lee et al., 2022)	0.377	0.162	0.148	0.221	3.1	0.241	0.002	0.001	0.005	3.1
CNN9	3	✓	MaxPool	×	(Le et al., 2023)	0.795	0.817	0.826	0.810	1.9	0.599	0.597	0.681	0.556	1.9
CNN10	4	✓	MaxPool	×	(Le et al., 2023)	0.870	0.888	0.892	0.885	1.8	0.705	0.792	**0.853**	0.759	1.8
CNN11	3	✓	None	×	(Salgado et al., 2023)	**0.902**	**0.922**	**0.932**	**0.914**	4.8	**0.758**	0.750	0.735	**0.828**	4.8
MLP						0.316	0.069	0.045	0.143	1.6	0.241	0.002	0.001	0.005	1.6
RNN						0.381	0.183	0.200	0.202	8.0	0.245	0.003	0.002	0.007	8.1
LSTM						0.728	0.743	0.762	0.728	14.6	0.515	0.156	0.224	0.151	14.6
GRU						0.765	0.788	0.802	0.777	15.1	0.575	0.273	0.400	0.251	15.1
Bidirectional-RNN						0.365	0.155	0.197	0.185	14.7	0.245	0.003	0.002	0.007	14.7
Bidirectional-LSTM						0.791	0.814	0.825	0.805	29.1	0.559	0.384	0.533	0.346	29.1
Bidirectional-GRU						0.800	0.826	0.840	0.816	30.3	0.627	0.451	0.609	0.408	30.3
Transformer						0.338	0.127	0.172	0.155	83.4	0.241	0.002	0.001	0.005	83.5
iTransformer						0.627	0.611	0.652	0.599	1.9	0.388	0.135	0.320	0.118	1.9
PatchTST						0.720	0.752	0.766	0.740	3.3	0.631	**0.811**	0.850	0.784	3.3

In library

Baselines：

- CNN-based Models： CNN-11
- Recurrent Models： RNN,LSTM,GRU. Bidirectional models
- Transformers : Transformer, iTransformer, PatchTST



Long-tailed distribution of space group

Out library

Task	Crystal System Classification				Space Group Classification			
Model	Accuracy	F1	Precision	Recall	Accuracy	F1	Precision	Recall
CNN1	0.627	0.437	0.565	0.472	0.285	0.003	0.002	0.006
CNN2	0.606	0.419	0.428	0.425	0.285	0.003	0.002	0.006
CNN3	0.659	0.501	0.496	0.507	0.285	0.003	0.002	0.006
CNN4	0.378	0.078	0.054	0.142	0.285	0.003	0.002	0.006
CNN5	0.495	0.278	0.285	0.283	0.285	0.003	0.002	0.006
CNN6	0.378	0.078	0.054	0.142	0.285	0.003	0.002	0.006
CNN7	0.673	0.607	0.633	0.608	0.429	0.053	0.050	0.072
CNN8	0.612	0.452	0.448	0.497	0.286	0.003	0.001	0.006
CNN9	0.675	0.632	0.629	0.644	0.439	0.099	0.137	0.113
CNN10	0.692	0.659	0.650	0.674	0.430	0.107	0.168	0.121
CNN11	0.702	0.672	0.659	0.690	0.481	0.136	0.167	**0.150**
MLP	0.378	0.078	0.054	0.142	0.285	0.003	0.002	0.006
RNN	0.409	0.162	0.149	0.178	0.274	0.003	0.002	0.009
LSTM	0.657	0.575	0.583	0.589	0.431	0.070	0.092	0.085
GRU	0.707	0.678	0.656	**0.709**	0.480	0.110	0.143	0.125
Bidirectional-RNN	0.403	0.157	0.146	0.174	0.295	0.004	0.003	0.008
Bidirectional-LSTM	0.704	0.663	0.654	0.678	0.349	0.035	0.044	0.051
Bidirectional-GRU	**0.722**	**0.699**	**0.697**	0.705	**0.498**	**0.138**	**0.192**	0.149
Transformer	0.376	0.138	0.231	0.158	0.285	0.003	0.005	0.006
iTransformer	0.606	0.495	0.519	0.526	0.367	0.053	0.085	0.064
PatchTST	0.656	0.616	0.612	0.627	0.375	0.067	0.093	0.082

Contributions

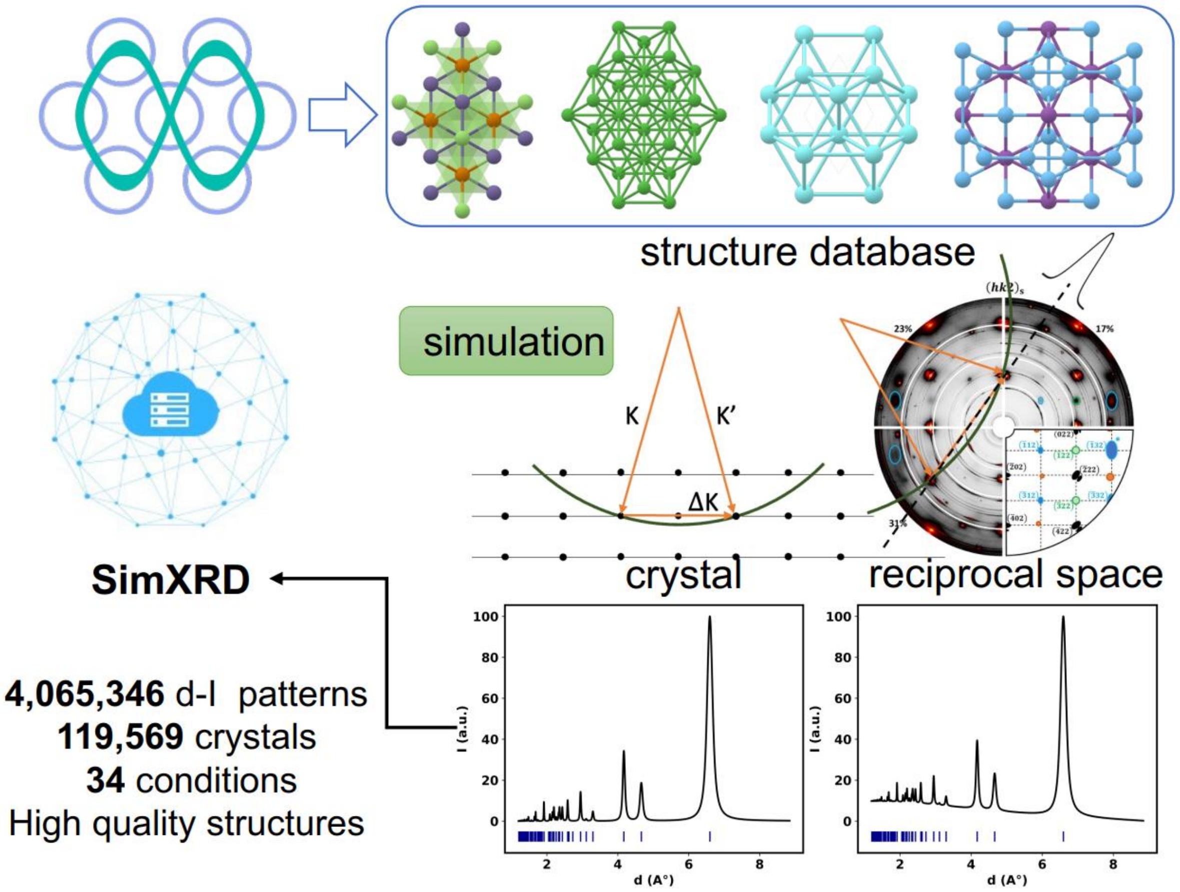
1 Most existing CNN models are unsuitable for symmetry identification within the large scale PXRD database

2 Convolutional neural networks without pooling have achieved the best performance in most tasks

3 Bidirectional recurrent models consistently outperform their unidirectional counterparts

4 PatchTST achieves a significant performance improvement compare to Transformer

Conclusion



SimXRD : Https://openreview.net/pdf?id=mkuB677eMM

1: We introduce **SimXRD**, the **largest open-source** XRD pattern dataset for **symmetry identification**.

2: Data analysis reveals that the symmetry labels follow **a long-tailed distribution**.

3: We evaluate 21 models on two different splitting patterns (**in-library** and **out-of-library**) and find that most existing models struggle to accurately predict the symmetry of low-frequency classes, even when addressing for **class imbalance**. This limitation hinders their real-world applicability.

4: Our results emphasize the importance of modeling long-tailed sequence classification and conducting **comprehensive comparison** to accurately assess the capabilities of various models.