Plan Overview

A Data Management Plan created using DMPonline

Title: Automated Crystal Phase Identification using Physics- Informed Transformers

Creator: Beulah Saka

Contributor: Dr Alex Ganose, Sandip De

Affiliation: Imperial College London

Template: MSc in Digital Chemistry: Research Data Management Statement (DMS)

Project abstract:

Crystalline materials diffract X-rays, and the resulting pattern can act almost like a fingerprint of the structure that caused it. Machine learning models have been trained extensively to read these patterns and identify a space group that matches the pattern, with accuracies above 80%. Those figures are obtained using datasets skewed in space-group distribution or with chemical contrast that can carry some of the signal. Improving these models by supervising them with differentiable physical laws and auxiliary targets has been widely proposed. It is not known whether that guidance improves symmetry assignment, or whether the limit lies in what a powder pattern records. Here we show that three forms of crystallographic guidance leave space-group accuracy unchanged, and that accuracy instead depends on a single step in the augmentation process. On 69,000 single-element structures balanced across all 230 space groups, four model variants span 0.54%. Of four distortions applied to the simulated patterns, three reduce accuracy by 2.7% between them while additive noise reduces it by 20.9%, because it alone fills the zeros in which systematic absences are recorded. Removing the noise doubles accuracy without changing the ranking. Target accuracies degrade in proportion to how much correct predictions depend on absent reflections; hence, the binding constraint is the input pattern itself. Applied to phase identification, predicted symmetry retrieves the correct structure from 69,000 candidates 0.1% of the time alone, but raises classical pattern matching from 76.9% to 80.2% as a prior on it. Symmetry prediction is therefore an addition to search-match rather than a replacement. We anticipate that noise levels will need to be reported and justified in work of this kind, since the choice affects accuracy more than any architectural decision tested here.

ID: 209123

Start date: 13-04-2026

End date: 20-08-2026

Last modified: 18-08-2026

Copyright information:

Automated Crystal Phase Identification using Physics- Informed Transformers

Data Collection

What data did you create or collect?

I generated 69,000 synthetic crystal structures (dummy-Cu, all 230 space groups) and simulated their powder XRD patterns myself, then built five augmented versions of that pattern set. I also recorded model-training output checkpoints, per-epoch logs, predictions, and metrics for every sweep and experiment run.

Did you reuse data (e.g. data sourced from a 3rd party data provider, data repository, database…)?

Yes, some machine learning training was done on the ICSD data; however, none of those results made it into the final report because they were exploratory.

What file formats did you use?

CSV and JSON for metadata and run results, HDF5 (.h5) and NumPy (.npz) for patterns and model arrays, PyTorch (.pt) for model checkpoints, Excel (.xlsx) for the final tables, and PDF/PNG for figures.

Did you create any software or write any code to create, process or analyse data? This includes any code used to train/test machine learning methods.

Yes. Sandip De supplied most of the structure generator, pattern simulator, and augmentation pipeline code; however, this changed over time as the research progressed. I created the transformer architecture code, and I wrote the training code for all four model variants for this project.

Ethical and Legal Responsibilities

Did your research involve human participants?

- No

Did you process or collect personal data?

- No

Are there any IP or copyright restrictions which might influence your use or sharing of the data?

Yes, I am not able to share the ICSD data i inherited as it is owned by other individuals.

Data Documentation and Quality Assurance

How has the data been documented to ensure it can be understood?

Every top-level results directory has a short README describing what was computed, which file holds the results, and the conclusion drawn from it.

What quality assurance processes did you adopt?

Every generated structure is verified with spglib against its intended space group before being added to the dataset. Model parameters were selected based on grid search results to ensure the model performs at its maximum capacity. Data was kept in the same directory throughout the project as to avoid corruption.

Data Storage and Security

How much data did you generate?

Roughly 8 GB stored locally (6.3 GB of structures and patterns, 1.9 GB of model results), plus a further set of large augmented pattern datasets (~1.1 GB each) generated and kept only on the remote GPU machine used for training, not mirrored locally.

How did you store and back-up your data during the project?

Data and code were kept on a local machine and a remote GPU server (TitanX) used for training,

How did you manage access and security?

Data is was not moved to any other directory other than my user accounts on the TitanX GPU or imperials HPC for model training. No one has had access to the main machine used for the project since the project began. The machine is password locked.

Data Preservation and Sharing

What are the plans for long-term preservation and data sharing of the data you have generated?

The model's training code is shared via a GitHub repository for anyone to access. Access to results and data is limited, subject to supervisor approval.

Will there be any restrictions on accessing the data?

The scripts on Github can be run to generate data so there should be no restrictions to the model's scripts. Access to the generated dataset is subject to supervisor approval.

Responsibilities and Resources

What was your responsibility in implementing this plan?

As the student, I was responsible for generating, verifying and documenting the data myself, and for keeping the record (READMEs) up to date as the project progressed.

If you would have to restart the same project, what would you change in the way you managed your data?

I would write the per-directory READMEs and standardise run-naming conventions from the start, not at the end. Several early runs had to be identified after the fact by cross-referencing scripts, and two results were briefly lost earlier in the project to a naming clash before run names were made to carry their full configuration.

Planned Research Outputs

Report - "Automated Crystal Phase Identification using Physics- Informed Transformers"

The output of this project is an Encoder only Transformer with a convolutional neural network backbone trained to classify synthetic structures.

Planned research output details

Title	DOI	Type	Release date	Access level	Repository(ies)	File size	License	Metadata standard(s)	May contain sensitive data?	May contain PII?
Automated Crystal Phase Identification using Physi ...		Report	2026-08-20	Open	None specified		None specified	None specified	No	No