InformedXRD:

Reproducible benchmarks and physics-informed evaluation for powder diffraction symmetry classification

Edward G. Friedman[1,2], Elizabeth J. Baggett[3,2], Abhishek Shetty[4], Derrick Chan-Sew[4], Vanellsa Acha[4], Harshitha Dwaracherla[4], Paul Kienzle[2], William Ratcliff[2,5,6]*

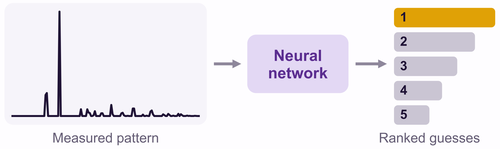
[1] School of Computer Science, Carnegie Mellon University · [2] NIST Center for Neutron Research · [3] Dept. of Physics, Boston College · [4] School of Information, UC Berkeley · [5] Dept. of Materials Science and Engineering and [6]of Physics, University of Maryland | * Corresponding author

Contact: egfriedm@andrew.cmu.edu

Paper website · LinkedIn · Twitter

Background, Problem, & Motivation

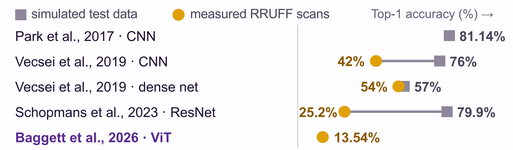
Task: Determine crystal structure symmetry from PXRD pattern using ML models



Why it is hard: 3D information squeezed into 1D pattern, experimental data adds impurity and noise effects



(2D toy; David, *Nature*, 1990)



Top-1 as reported. Test sets differ, so numbers are not comparable.

There have been 4 main evaluation issues/practices that are holding back ML4PXRD

Wrong target (see 1)
Space groups with the same missing peaks look identical.

Non-reproducible benchmarks (see 2)
Each paper picks its own subset of RRUFF / experimental data archives

Flat metrics (see 3 · Tool 2)
A near miss and a wild guess are both just "wrong".

Missing baselines (see 3 · Tool 1)
Nobody checks against guessing the most common labels.

1 Extinction-group mapping

Grade models on what PXRD can resolve

Shared missing peaks in PXRD patterns collapse 230 space groups to 99 extinction groups

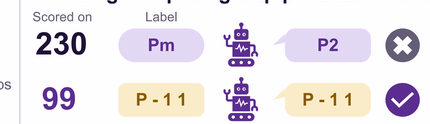
230 space groups (main target of interest): P2, Pm, P2/m

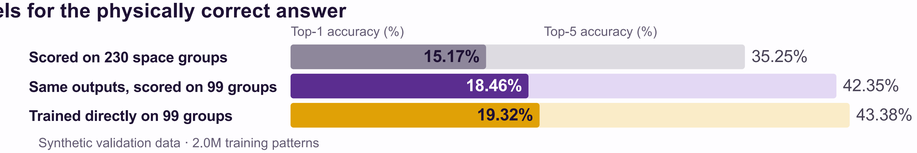


P-11 — one extinction group

99 extinction groups

Grading on space group penalizes models for the physically correct answer





Synthetic validation data · 2.0M training patterns

2 Standardized Benchmark construction method

Build benchmarks from a recipe anyone can rerun

A benchmark should be a recipe, not a hand-picked list

Source archive — raw scans with mixed labels, wavelengths and duplicates

1 **Single label** — one scan, one phase, one label

2 **Wavelength and grid** — same X-ray, same angle grid

3 **Structural grouping** — group scans of the same structure

4 **Family coherence** — keep families that agree

5 **Exact duplicates** — drop identical patterns

6 **Quality strata** — tag each scan by fit quality

7 **Freeze and audit** — lock IDs, date and code

Frozen benchmark — fixed scan IDs, snapshot date and code

Same for every archive: steps 1, 2, 7 · Adapted to each archive: steps 3 to 6

Principle: the rules are code, so anyone can rebuild the set.

Same recipe, two archives: RRUFF and opXRD

1,981 Cu-Kα scans in RRUFF → **RRUFF-473 Benchmark** — 227 minerals · 46 of 99 groups → **RRUFF-325 Benchmark** — 168 minerals · 45 of 99 groups

RRUFF-325 = the RRUFF-473 scans with usable or better fit quality.

92,552 patterns in opXRD → **908** single-phase, one label → **229** measured with Cu-Kα → **opXRD-98** — full 10–80° coverage · 23 extinction groups · a stress test, not a benchmark

3 Evaluation tools

Ask how a model is right, and how it is wrong, via scientific & statistics insight

Demonstrated on 11 models

4 ViT checkpoints
1 residual CNN
6 XRDBench architectures

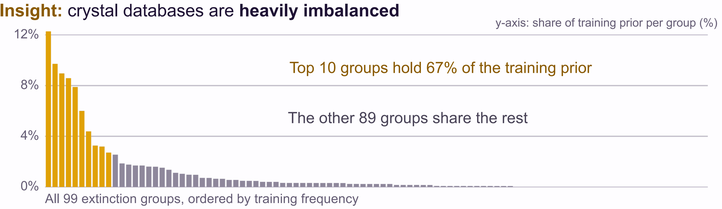
Limits

RRUFF sets cover 46 and 45 of 99 groups. Cu-Kα lab data only. Rescoring needs released checkpoints.

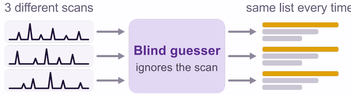
TOOL 1 OF 3 · BASELINES AND SHUFFLE TEST

Does the model beat a guesser that never sees the pattern?

Insight: crystal databases are **heavily imbalanced**

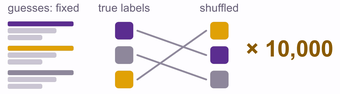


Eval tool: prior-only baseline



Same Top-5 list for every scan.

Eval tool: label-permutation test



× 10,000

No drop after shuffling = not reading the pattern.

Result: a tie at Top-5

Top-1	Blind guesser	12.6%
	Final ViT	10.46%
Top-5	Blind guesser	43.7%
	Final ViT	43.69%

Result: a small real signal

Top-5	Real labels	43.69%
	Labels shuffled	41.48%

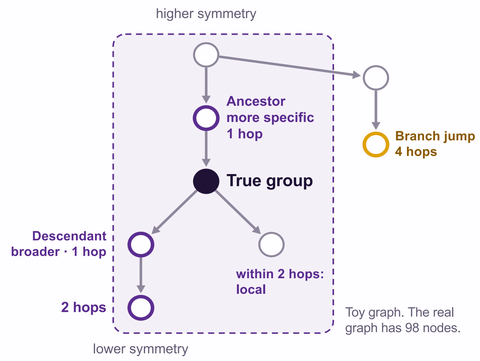
+2.21 points, p = 0.028

Bars: accuracy (%) on RRUFF-325, final ViT. Blind guesser = the test set's label frequencies.

TOOL 2 OF 3 · SYMMETRY-GRAPH METRIC

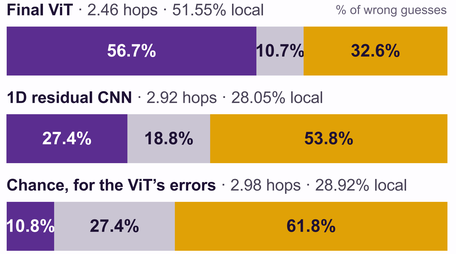
How far off is a wrong guess, and in which direction?

Insight: extinction groups form a natural hierarchy that **separates near misses from wild guesses**



Toy graph. The real graph has 98 nodes.

Eval tool: graph distance and direction of errors



- Descendant: **broader**, lower symmetry
- Ancestor: more **specific**, higher symmetry
- Branch jump: other crystal family (severe)

Result: Top-1 and graph distance crown different models

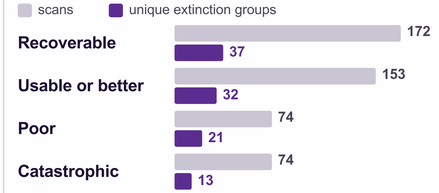


TOOL 3 OF 3 · STRATIFIED REPORTING

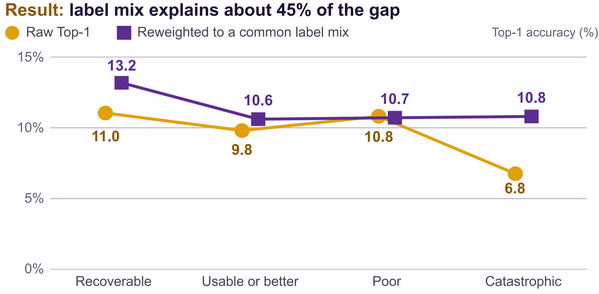
Are scans that are hard to fit also hard for the model?

Insight: the classical Rietveld fit measures pattern quality

Eval tool: RRUFF-473 binned by Rietveld fit quality



Result: label mix explains about 45% of the gap



Catastrophic scans span only 13 groups.

Takeaways

1 **Fix the target to fairly grade models.** 2 **Use standardized and reproducible benchmarks** 3 **Report physics-informed evaluations and dataset baselines instead of Top-k accuracy**

NeurIPS 2026, Evaluations and Datasets Track

Support: Texas Advanced Computing Center (TACC), The University of Texas at Austin; ACCESS allocation PHY250007; Center for High Resolution Neutron Scattering, a partnership between NIST and the National Science Foundation (DMR-2010792). Corresponding author: william.ratcliff@nist.gov

CHRNS — Center for High Resolution Neutron Scattering

NIST — National Institute of Standards and Technology, U.S. Department of Commerce