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] Dept. of Physics, University of Maryland · * Corresponding author

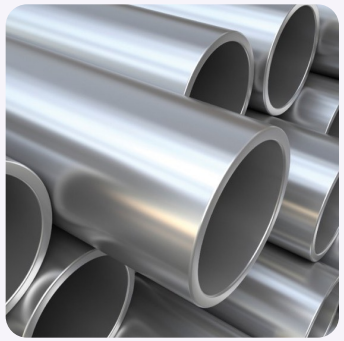


NeurIPS 2026 · Evaluations and Datasets Track

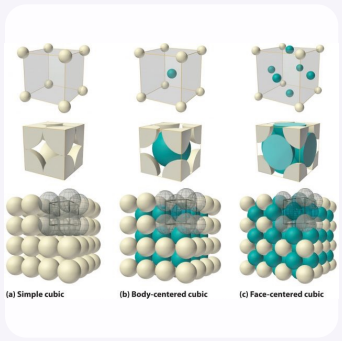
NIST NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY U.S. DEPARTMENT OF COMMERCE

NSF CHRNS CENTER FOR HIGH RESOLUTION NEUTRON SCATTERING NIST

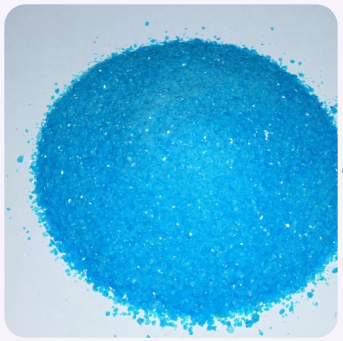
Powder X-Ray Diffraction (PXRD) encodes information about a crystal structure into a pattern



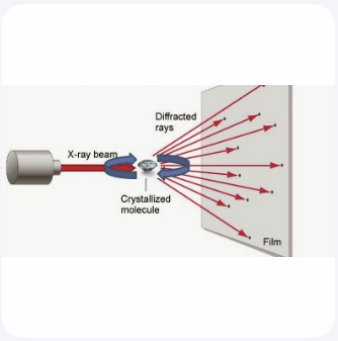
Material



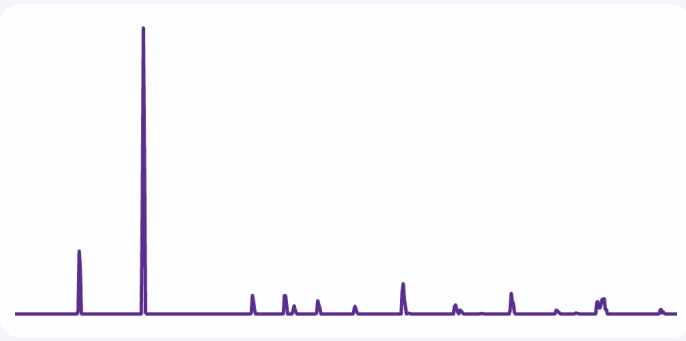
Atoms in a lattice



Ground to powder



X-ray beam in

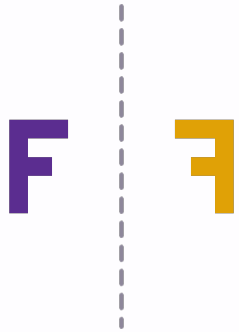


Pattern out: intensity vs. angle

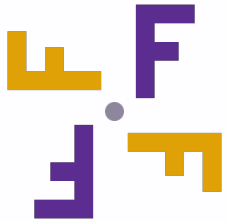
Measured scan: quartz (RRUFF R040031)

Peak positions and heights encode symmetry and atomic arrangement.

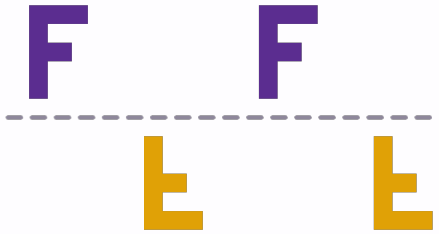
Crystals are sorted by symmetry into 230 space groups



Mirror



Rotation



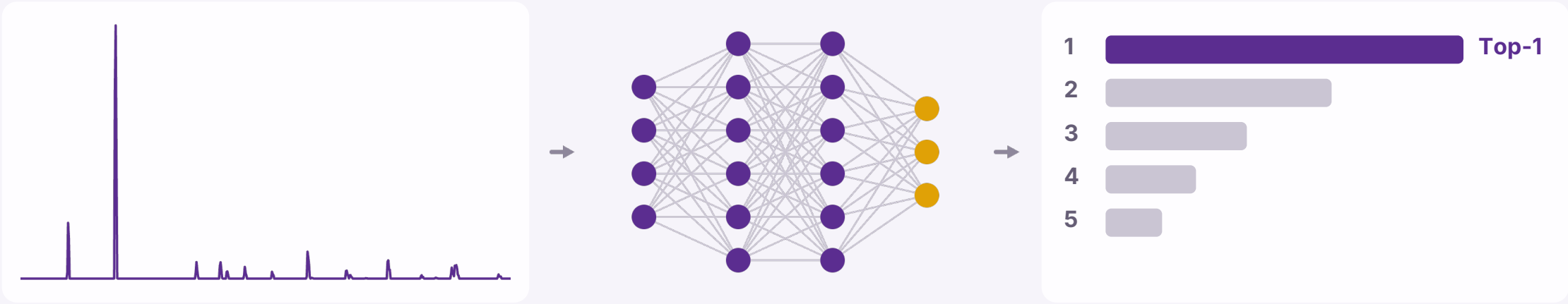
Glide: mirror + half step

230

space groups

every way symmetry operations can combine in a 3D crystal

ML-4-PXRD: Training models to learn crystal structure symmetry from pattern



Measured pattern

Neural network

Ranked symmetry labels

Trained on simulated patterns

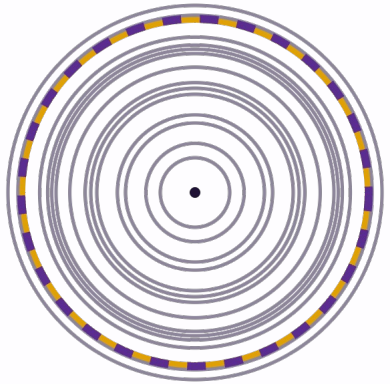
Tested on measured patterns

Why it is hard: PXRD squeezes 3D information into one signal



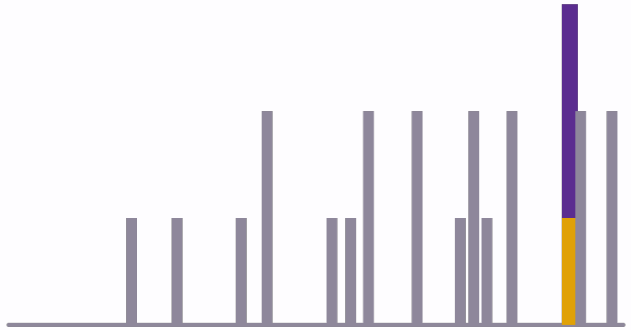
Single crystal

separate spots, each with a direction



Powder

every orientation at once: spots smear into rings



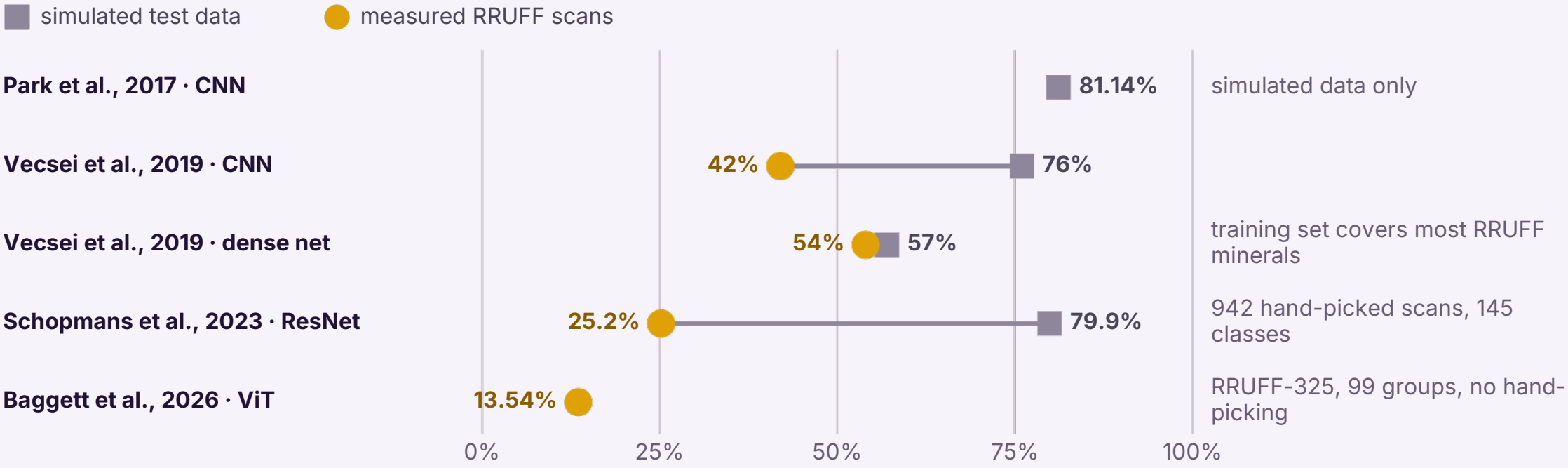
What we measure

one curve: different reflections share a peak

Real scans add: preferred orientation · noise and background · impurity phases

Source: W. I. F. David, Nature, 1990

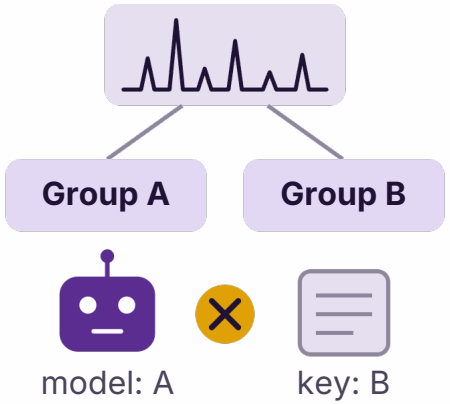
Reported accuracy is high on simulated data and drops on measured scans



Each study uses its own test subset, label set and filters, so the numbers are not comparable.

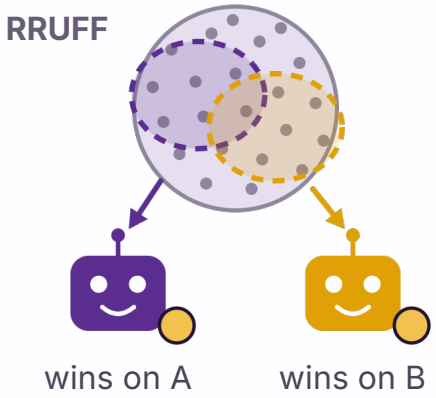
ML-4-PXRD has four main issues when it comes to evaluation

1) Wrong target



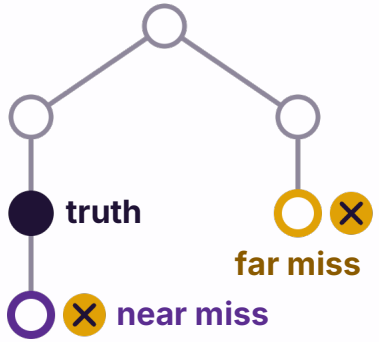
PXRD alone cannot infer a crystals space group

2) Non-reproducible benchmarks



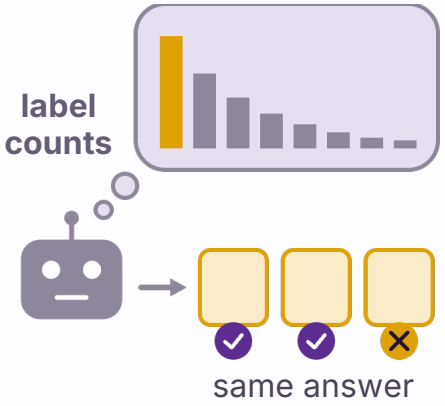
Scores cant be compared since papers use different, non-specified, subset of RRUFF

3) Flat metrics



Target has a inherent hierarchy, yet we give the same penalty far misses and near misses the same.

4) Missing baselines



Model can score by guessing the most common labels. Without a guess-only baseline, nobody can tell.

Three contributions, demonstrated on 11 models

1

Using extinction groups & Extinction-group mapping

230 → 99

space groups → extinction groups

2

Standardized Benchmark construction method

RRUFF-473 | RRUFF-325 | opXRD-98

rules written as code, frozen sets

3

Evaluation tools

- Prior-only baselines
- Label-permutation test
- Symmetry-graph error metric
- Stratified reporting

Demonstrated on 11 models

4 ViT checkpoints | 1 residual CNN | 6 XRDBench architectures

Mapping, RRUFF sets, symmetry graph and ViT checkpoints were first released with Baggett et al., 2026.

Contribution 1: grade on what PXRD can actually resolve

1

Using extinction groups & Extinction-group mapping

230 → 99

space groups → extinction groups

2

Benchmark construction method

RRUFF-473 · RRUFF-325 · opXRD-98

rules written as code, frozen sets

3

Evaluation tools

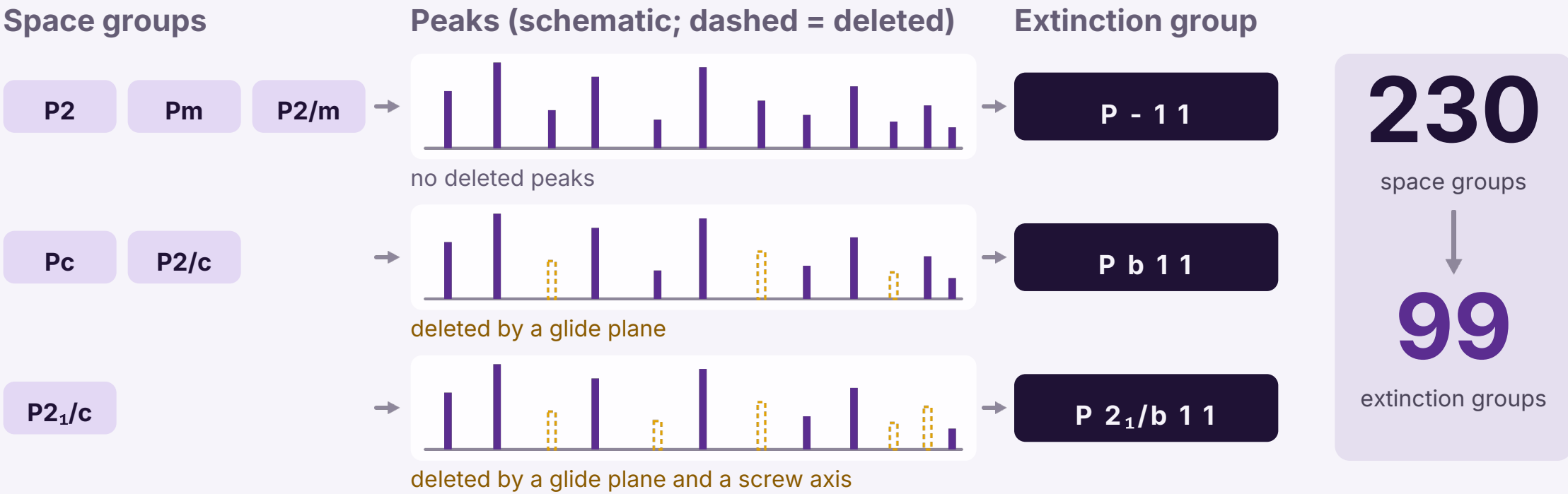
- Prior-only baselines
- Label-permutation test
- Symmetry-graph error metric
- Stratified reporting

Demonstrated on 11 models

4 ViT checkpoints · 1 residual CNN · 6 XRDBench architectures

Mapping, RRUFF sets, symmetry graph and ViT checkpoints were first released with Baggett et al., 2026.

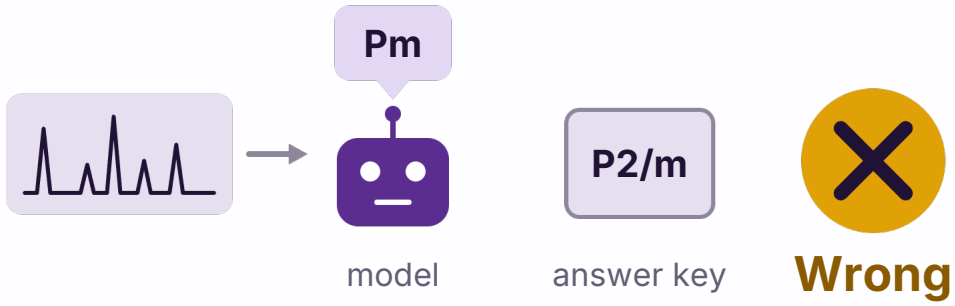
Crystallography Insight: PXRD causes some space groups to have the same missing peaks, collapsing 230 space groups to 99 extinction groups



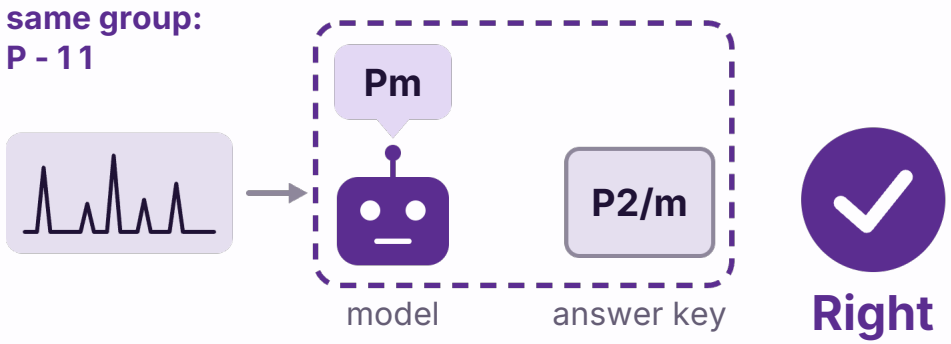
Same deleted peaks, same extinction group. Powder data cannot split them further.

Grading (and training) on space group instead of extinction groups punishes a physically correct answer

Graded on 230 space groups

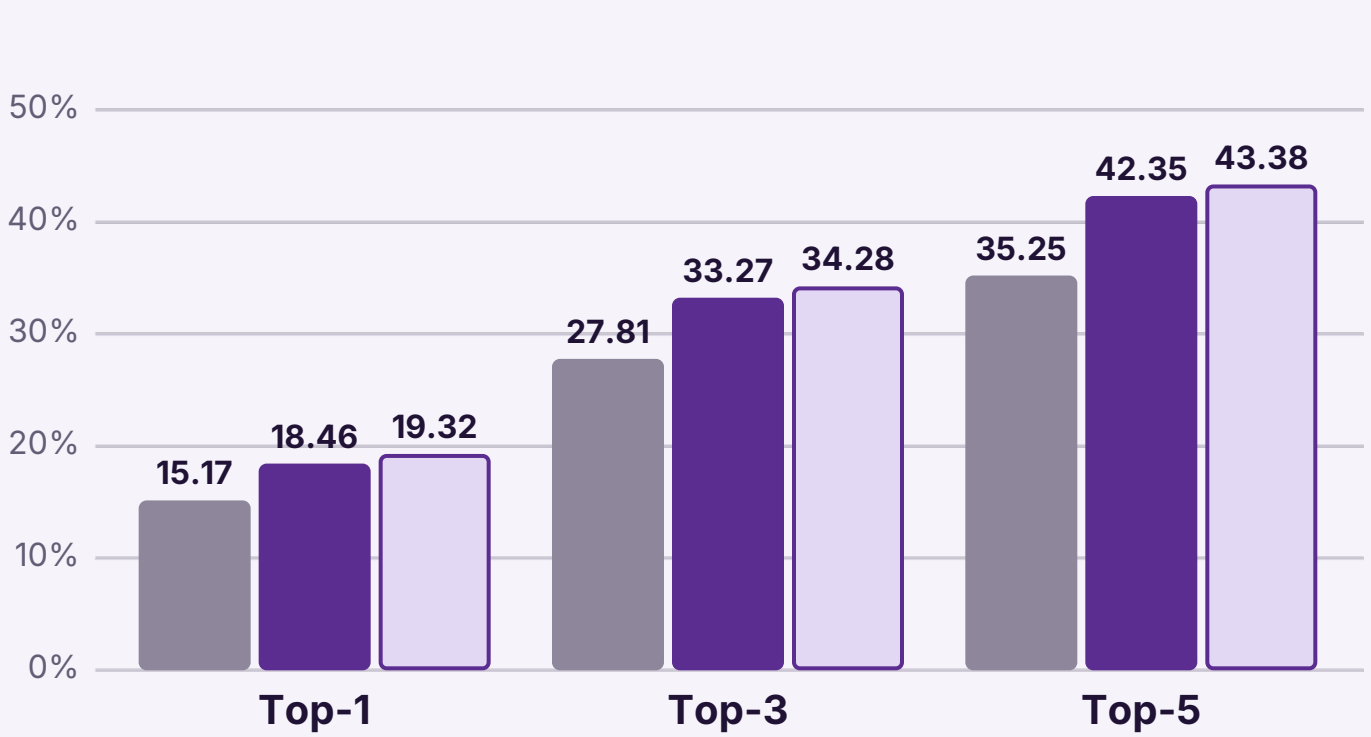


Graded on 99 extinction groups



Pm and P2/m sit in the same extinction group. Powder diffraction cannot tell them apart.

Fair target, same predictions: 15.17% → 18.46% Top-1



Scored on 230 space groups

Same outputs, scored on 99 extinction groups

Trained directly on 99 groups, same amount of per group

What we release

A verified 230 → 99 lookup table, as JSON and CSV. Any 230-way model can be rescored after the fact.

Direct training adds +0.86 Top-1 at full scale, +2.77 at one-eighth of the data.

Synthetic validation data, 2.0M-pattern training set.

Contribution 2: build benchmarks from a recipe anyone can rerun

1

Extinction-group mapping

230 → 99

space groups → extinction groups

2

Benchmark construction method

RRUFF-473 RRUFF-325 opXRD-98

rules written as code, frozen sets

3

Evaluation tools

- Prior-only baselines
- Label-permutation test
- Symmetry-graph error metric
- Stratified reporting

Demonstrated on 11 models

4 ViT checkpoints · 1 residual CNN · 6 XRDBench architectures

Mapping, RRUFF sets, symmetry graph and ViT checkpoints were first released with Baggett et al., 2026.

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

Raw archive in → **Frozen benchmark out**

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

Same for every archive: steps 1, 2, 7

Adapted to each archive: steps 3 to 6

The inclusion rules are code. Anyone can rebuild the set from a frozen snapshot.

RRUFF: 1,981 scans become two frozen benchmarks

1,981 Cu-Kα scans in RRUFF

family-coherence checks

RRUFF-473

473 scans · 227 minerals · 46 of 99 groups

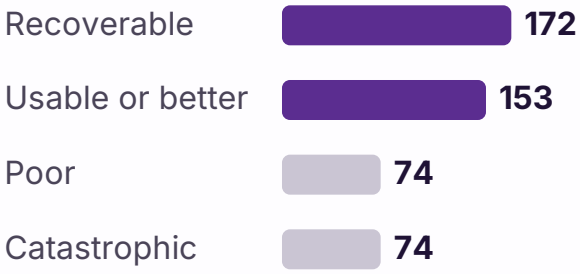
fit-quality cutoffs

RRUFF-325

325 scans · 168 minerals · 45 of 99 groups

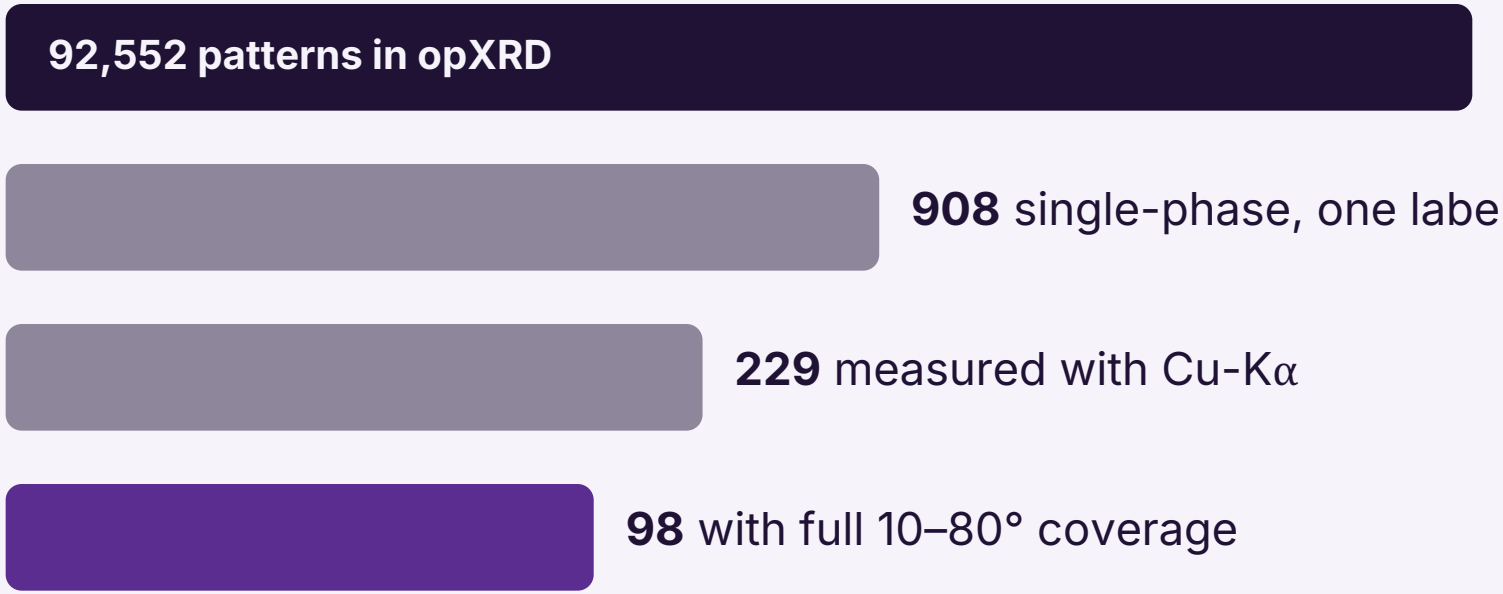
Frozen scan IDs, snapshot date and code reproduce both sets exactly.

RRUFF-473 by fit quality



Purple rows form RRUFF-325

Same recipe, new archive: 92,552 opXRD patterns yield 98



Bar widths on a log scale.

opXRD-98

28 space groups
23 extinction groups

A stress test, not a benchmark

The label space and both scoring readouts transferred without modification.

Contribution 3: ask how a model is right, and how it is wrong

1

Extinction-group mapping

230 → 99

space groups → extinction groups

2

Benchmark construction method

RRUFF-473 RRUFF-325 opXRD-98

rules written as code, frozen sets

3

Evaluation tools

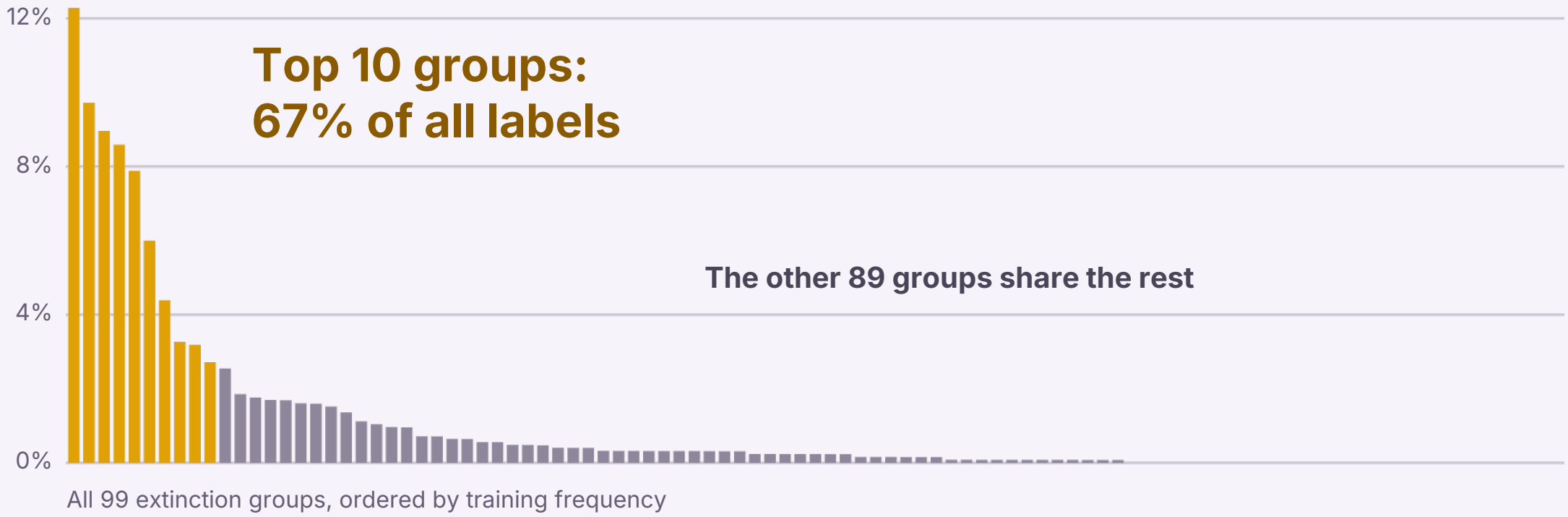
- Tool 1 · Baselines + shuffle test
- Tool 2 · Symmetry-graph metric
- Tool 3 · Stratified reporting

Demonstrated on 11 models

4 ViT checkpoints | 1 residual CNN | 6 XRDBench architectures

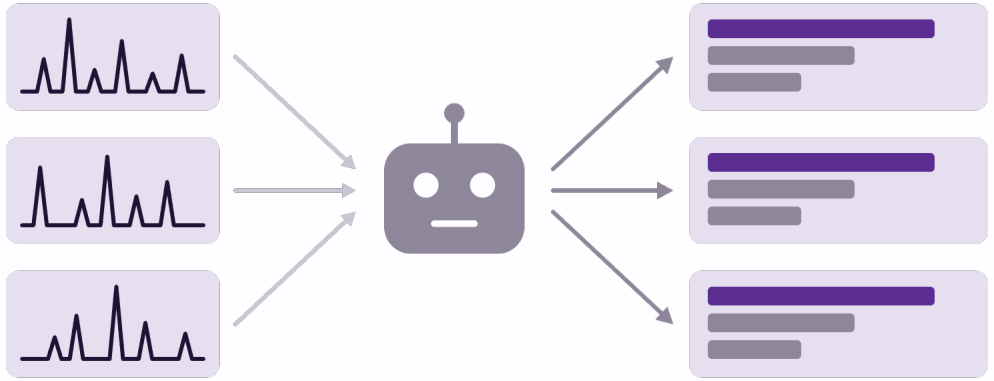
Mapping, RRUFF sets, symmetry graph and ViT checkpoints were first released with Baggett et al., 2026.

Labels are heavily imbalanced: ten groups hold 67% of the training prior



Two controls: guess by frequency, then shuffle the labels

1 · Prior-only baseline

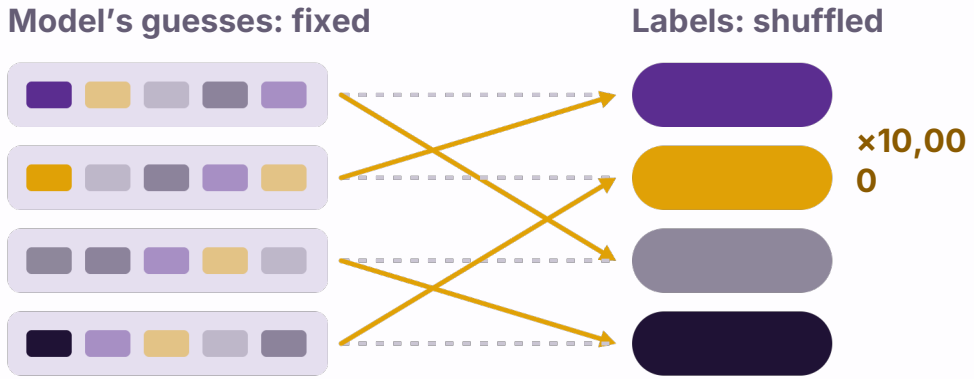


Ignores the pattern. Stamps the same Top-5 list on every scan.

Deployable: uses training frequencies

Oracle: uses the test set's own frequencies

2 · Label-permutation test

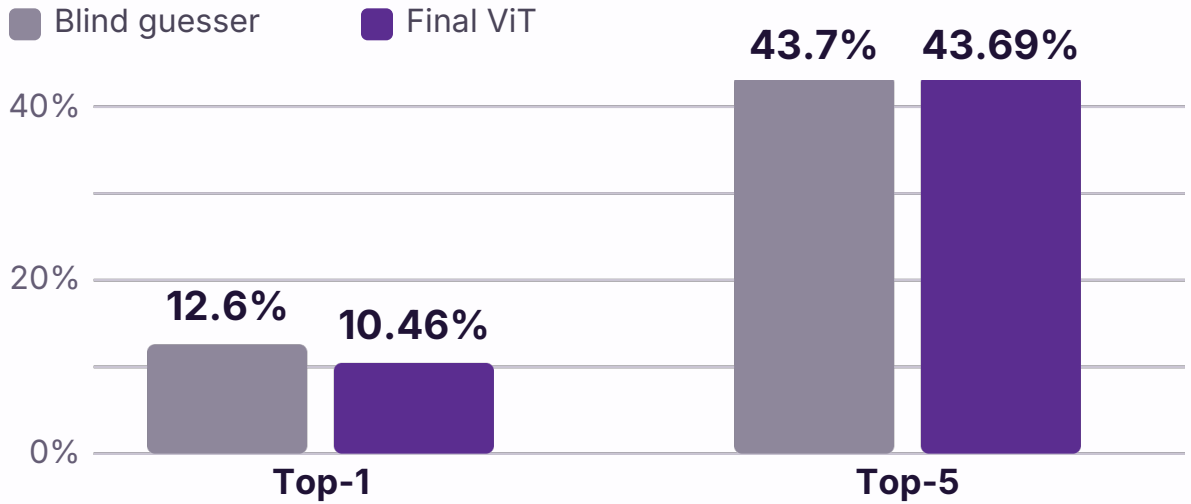


If the score does not drop after shuffling, the model is not reading the pattern.

Two controls, two answers: a tie, and a small real signal

1 Prior-only baseline

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



Answer: no clear win. They tie at Top-5.

2 Label-permutation test

Does the model use anything specific to each pattern?

Top-5 accuracy

43.69% real labels

41.48% labels shuffled

+2.21 points, p = 0.028

Answer: yes, a little.

Both on RRUFF-325, final ViT. Blind guesser = the test set's own label frequencies (oracle prior).

A near miss should not cost the same as a wild guess

Question: what is g? (9.8 m/s²)	Top-1 grading	A physics teacher
Student A 9.7 m/s²	✗ Wrong	Close
Student B 98 m/s²	✗ Wrong	Way off





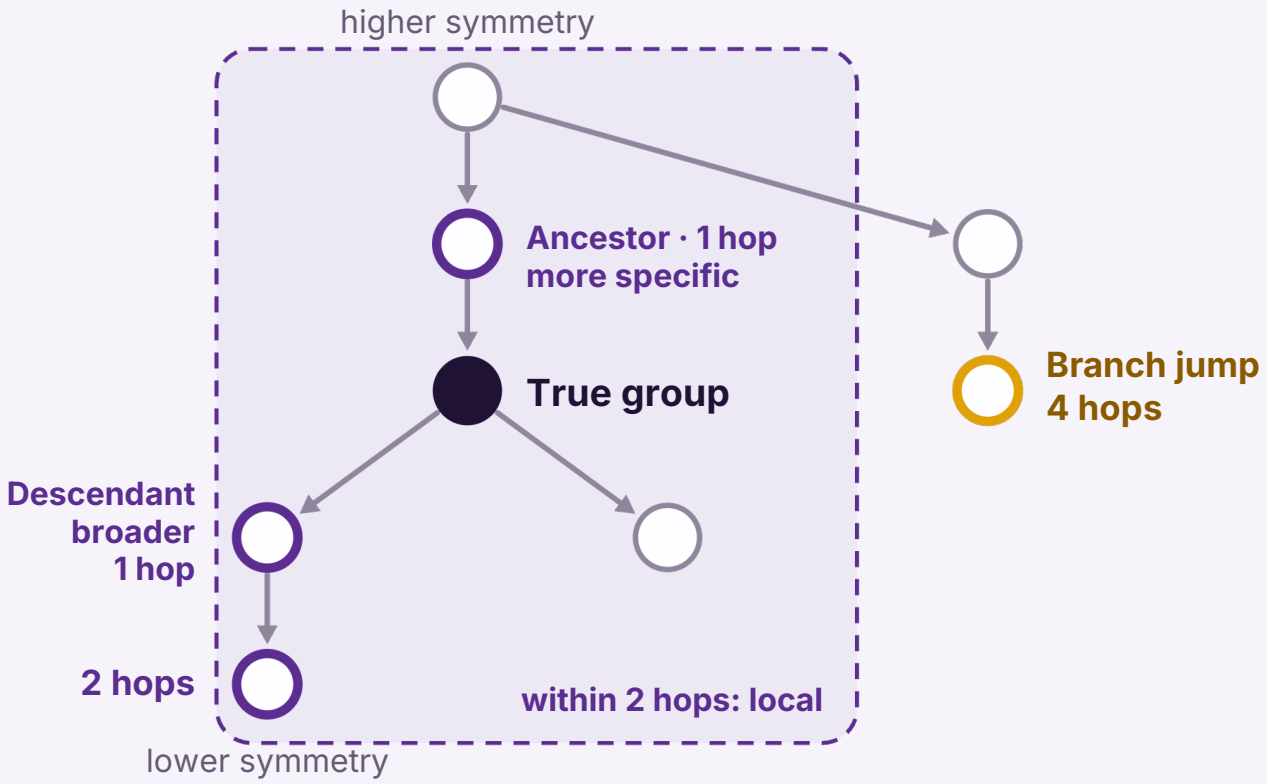
Same idea for symmetry: a neighboring subgroup is a near miss, a different crystal family is a wild guess.

Direction matters too: broad guesses are safer than overly specific ones

	**Truth: tetragonal** like a rectangle		**Measured: one peak family, split in two** dashed outlines below mark these two peaks
	**Orthorhombic** like a parallelogram	 allows 3 peaks; two can merge	**Broader: fits, only less specific** Descendant · lower symmetry · the conservative error
	**Cubic** like a square	 allows 1 peak; no split possible	**More specific: contradicted by the split** Ancestor · higher symmetry than the data support
	**Hexagonal** like a triangle	 peaks at other angles	**Unrelated: a different peak sequence** Branch jump · the severe error

Top-1 marks all three wrong:
Noise erases evidence of high symmetry first, so a nearby lower-symmetry guess is the expected, safer failure.

The symmetry graph measures how far off a guess is



98 nodes

one per extinction group

Edges

subgroup relations, from higher to lower symmetry

Distance

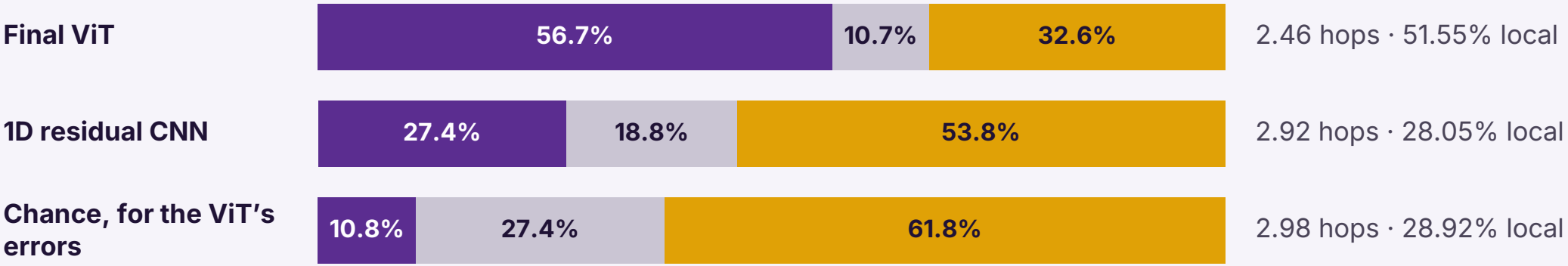
hops between the guess and the truth

Distance, direction and locality separate safe errors from severe ones

Distance
mean hops from the truth

Direction
broader (lower symmetry), more specific (higher), or a branch jump

Locality
share of errors within 2 hops



Descendant (broader): fell to lower symmetry

Branch jump: different crystal family (severe)

Ancestor (more specific): claimed higher symmetry

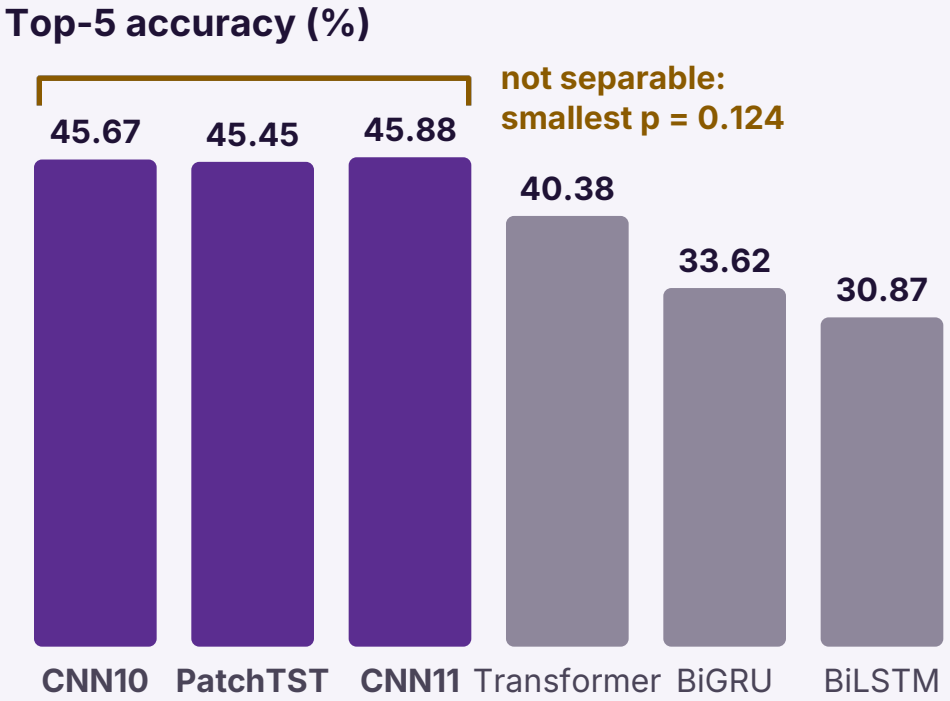
RRUFF-325, wrong Top-1 predictions only.

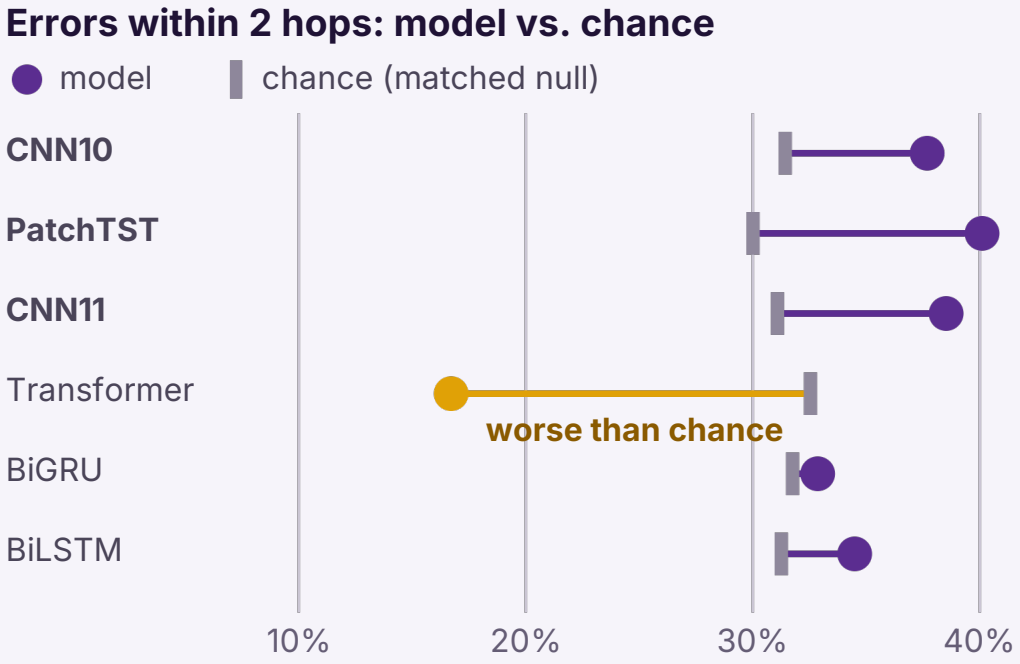
Top-1 and graph distance crown different models



RRUFF-325. Lower distance is better. ViT checkpoints from Baggett et al., 2026.

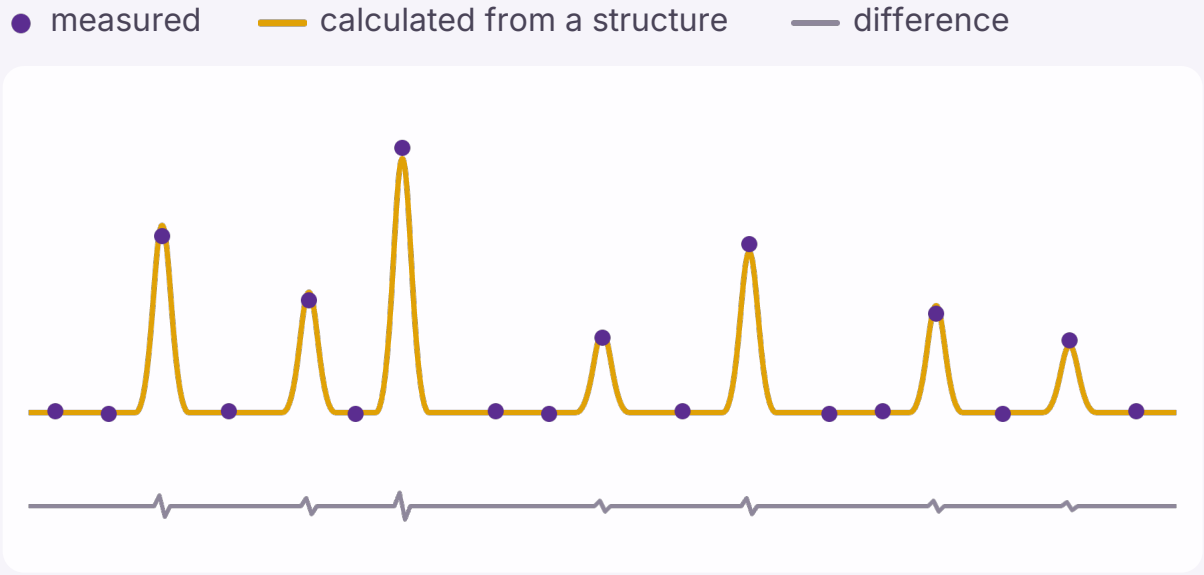
Top-k cannot separate the top three architectures; their error structure differs





RRUFF-473. Six architectures trained under the XRDBench recipe, rescored on 99 extinction groups.

Rietveld fitting: the classical way to check a pattern against a structure



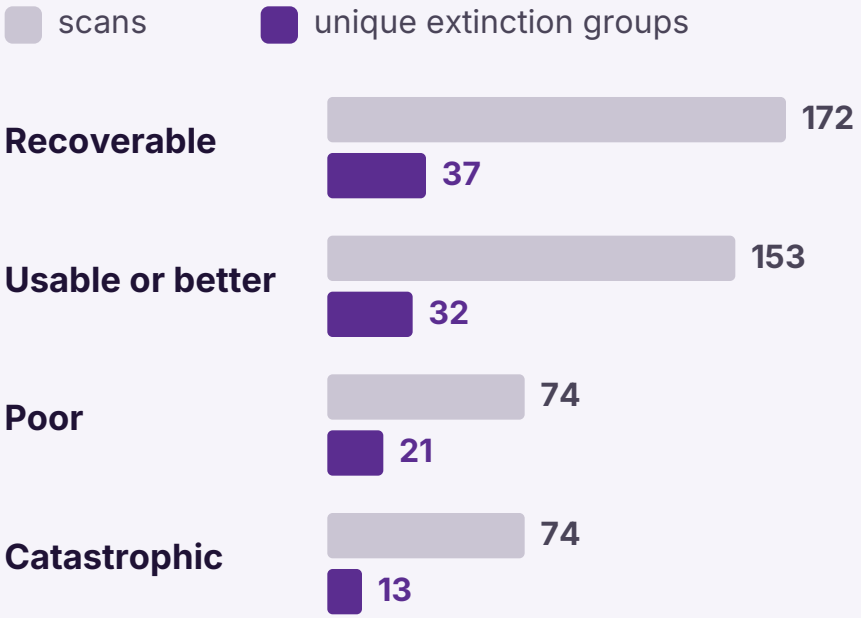
How a fit works

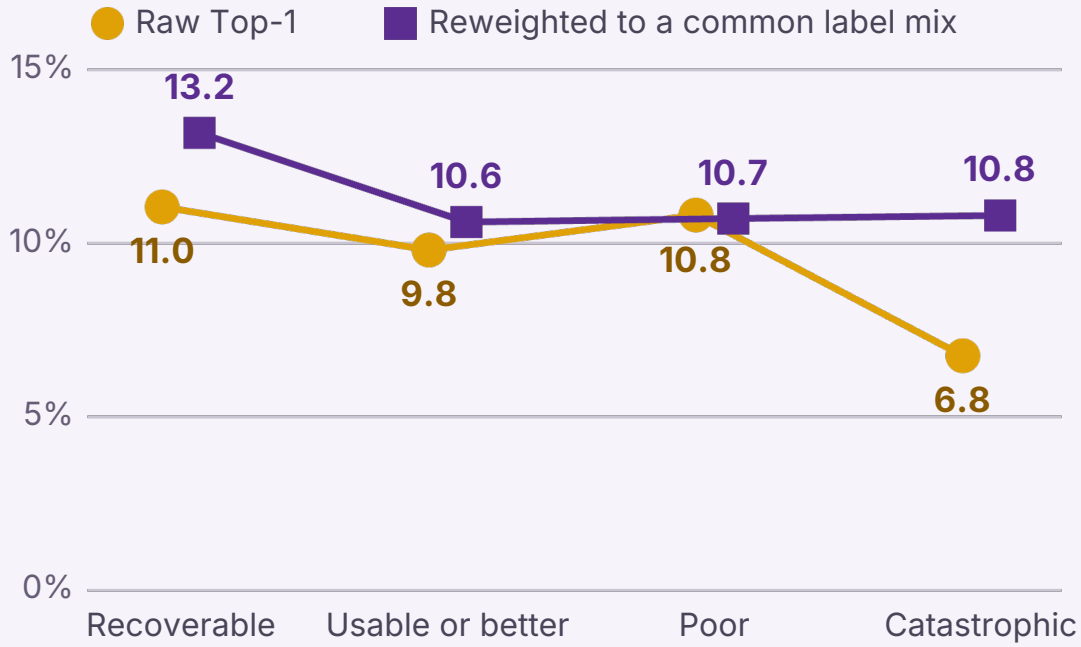
1. Start with known crystal
2. Simulate the whole pattern: peak positions, heights, shapes, background
3. Tune the parameters by least squares until it matches

How well it fits gives each scan a grade: Recoverable · Usable or better · Poor · Catastrophic

The question: are scans that are hard to fit also hard for the model?

Fit quality does not predict difficulty; label mix explains about 45% of the gap





Catastrophic scans span only 13 groups. Controlling for label mix closes about 45% of the gap.

TAKEAWAYS

Evaluation design changes what you can conclude

1. **Grade on what the measurement can resolve: 99 extinction groups**
2. **Build benchmarks from code, then freeze them**
3. **Report baselines, error structure and strata, not only Top-k**

Limits: the RRUFF sets cover 46 and 45 of the 99 groups, Cu-Kα laboratory data only; only models with released checkpoints can be rescored.

Website

edwardf8.github.io
/informedxrd

Twitter

@EdwardFrie41242

ACKNOWLEDGMENTS

Thank you

- Texas Advanced Computing Center (TACC), The University of Texas at Austin
- ACCESS allocation PHY250007, “Applications of AI to Diffraction”
- Center for High Resolution Neutron Scattering, a partnership between NIST and the National Science Foundation (DMR-2010792)

Corresponding author: william.ratcliff@nist.gov

NIST NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY U.S. DEPARTMENT OF COMMERCE

NSF CHRNS CENTER FOR HIGH RESOLUTION NEUTRON SCATTERING NIST