Attention Is Not All You Need for Diffraction

Physics-informed transformers for crystal symmetry classification

William Ratcliff
NIST Center for Neutron Research
Dept. of Materials Science & Engineering, U. Maryland

National Academies Workshop: Frontiers of AI-Realized Materials

arXiv:2604.23811

Collaborators

- Elizabeth Baggett (Boston College / NIST)
- Edward Friedman (Carnegie Mellon / NIST)
- Karen Cao (Caltech / NIST)
- Derrick Chan-Sew (UC Berkeley)
- Harshita Dwarcherla (UC Berkeley)
- Paul Kienzle (NIST)
- Satvik Lolla (Berkeley / NIST)
- Dan Luu Nguyen (UC Berkeley)
- Abhishek Shetty (UC Berkeley)
- Vanellsa Acha (UC Berkeley)
- Ichiro Takeuchi (U Maryland)

Computational resources: Texas Advanced Computing Center (Stampede3, Vista)

Outline

1. Powder diffraction, systematic absences, and why symmetry is hard
2. CNNs: first attempts, scaling limits, and prior work
3. Transformers: a better architecture for this physics
4. Extinction groups: the information-theoretically correct target
5. Physics-informed ViT: coordinate channels, dual heads, and fusion
6. Training curriculum and calibration
7. Topological error analysis: errors on the subgroup DAG
8. Preferred orientation, the catastrophic paradox, and NN→Pawley pipeline

The Diffraction Problem

What can a powder pattern tell us about crystal symmetry?

Powder Diffraction in 60 Seconds

The experiment:

- Shine X-rays or neutrons at a polycrystalline powder
- Measure scattered intensity as a function of angle (2θ)
- Result: a 1D pattern of peaks

What the peaks encode:

- Peak positions → lattice geometry (unit cell)
- Peak intensities → atomic arrangement
- Systematic absences → symmetry elements (screw axes, glide planes)

The inverse problem:

- Given a pattern, determine the crystal symmetry
- Traditionally requires expert crystallographer + manual iteration
- Poorly suited to high-throughput or autonomous experiments

Systematic Absences: The Symmetry Fingerprint

What are systematic absences?

Certain symmetry elements (screw axes, glide planes, lattice centering) cause specific reflections to vanish — their structure factor is exactly zero.

Example: 2_1 screw axis along b

A 2-fold rotation + half-lattice translation

Causes: 0k0 reflections with odd k are absent

If you see intensity at (030) but not (010), that's a diagnostic clue

Example: Body-centered (I) lattice

Extra lattice point at the cell center

Causes: hkl reflections with h+k+l = odd are absent

This eliminates roughly half of all possible peaks

Why this matters for classification:

- The pattern of which peaks are present vs. absent is the primary clue for determining symmetry
- But in real data, noise adds intensity where absences should be, and preferred orientation can suppress peaks that should be present
- This is why the problem is hard — and why the model needs to see long-range patterns, not just local peaks

Why Is This Hard?

The loss landscape is non-convex

- Segal et al. showed that gradient-based refinement can converge to incorrect local minima even from good starting points

Multiple space groups produce identical patterns

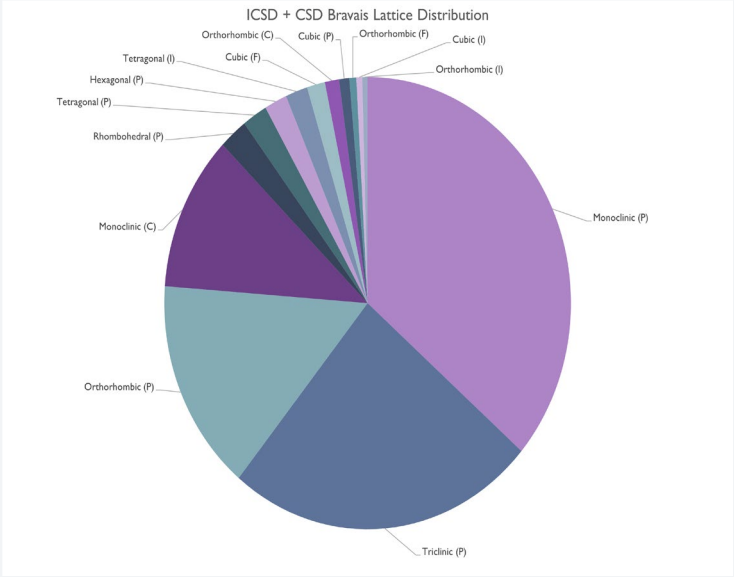
- Friedel's law, shared reflection conditions → fundamental ambiguity

Real data is messy

- Preferred orientation, impurity phases, background noise, sample displacement
- These can add or remove intensity where symmetry demands specific values

Extreme class imbalance in databases

- A few space groups dominate; most have very few examples

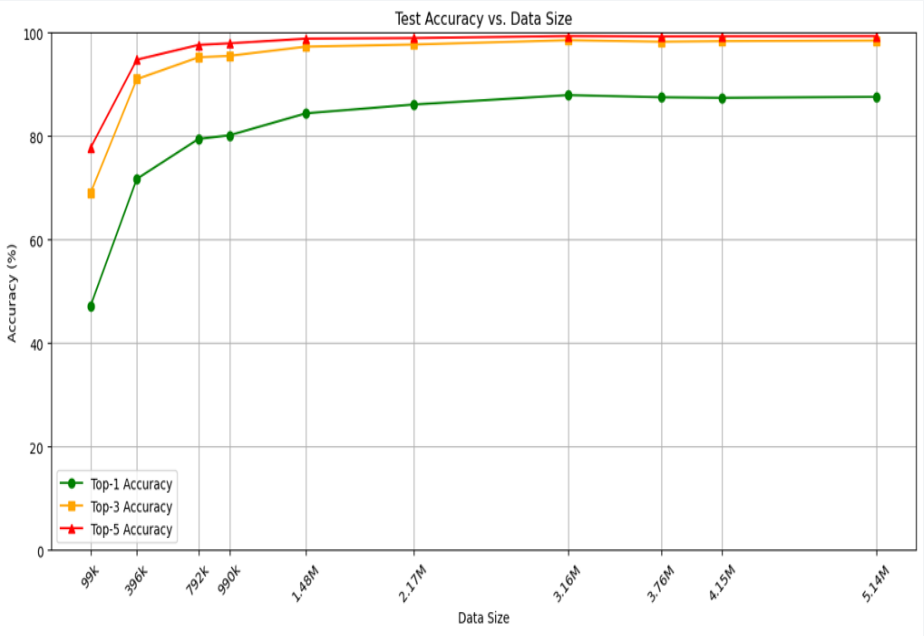


ICSD + CSD Bravais lattice distribution

CNNs for Diffraction

The first generation of ML approaches

CNNs: Strong Start, Hard Ceiling



ResNet-18 Top-1 saturates at ~88% regardless of data size

What is a CNN?

- Slides small filters across the input to detect local features
- Our baseline: 1D ResNet-18 for diffraction patterns

The problem: architectural ceiling

- Local receptive field misses long-range peak relationships
- Translation invariance is wrong: a peak at 20° means something fundamentally different from 60°
- Performance plateaus even with more data

Need an architecture that sees the whole pattern at once

Prior Work: Scaling CNNs to 261 Million Patterns

Schopmans et al. (2023) pushed CNN scaling to the limit:

- Trained ResNet models on up to 261 million on-the-fly synthetic diffractograms
- Used ICSD-informed synthesis with strong chemical priors
- Reported ~25% Top-1 on 942 hand-filtered RRUFF patterns (145 space groups)
- **Observed performance plateau beyond ~10^8 structures**

The lesson:

- Even 261M patterns could not overcome the CNN's architectural ceiling
- The local receptive field is a fundamental bottleneck for this task
- **More data is not enough — we need a different architecture**

This motivates our move to transformers, which can see the entire pattern at once

Transformers

Self-attention: seeing the whole pattern at once

What Is a Transformer?

Self-Attention

- Every position attends to every other position simultaneously

Three learned projections:

- Query: "what am I looking for?"
- Key: "what do I contain?"
- Value: "what information do I carry?"

- Attention score = how relevant each position is to each other
- Multi-head: learn multiple relationship types in parallel

Why This Matters for Diffraction

- Peak ratios encode lattice geometry
  - Comparing low-angle and high-angle peaks simultaneously

- Systematic absences are global
  - Missing peaks at specific positions reveal symmetry

- Not translation invariant
  - A peak at 20° and 60° mean different things

Transformers can model all of this natively

Extinction Groups

Asking the right question

The Information-Theoretic Target

230 space groups → 99 extinction groups

- Multiple space groups share identical reflection conditions
- They are indistinguishable by powder diffraction
- Training on space groups penalizes models for failing to distinguish the indistinguishable

First hint (different dataset sizes — suggestive, not controlled):

Target	Data	Top-1	Top-3	Top-5
Space Groups (230)	2.3M	37%	74%	88%
Extinction Groups (99)	990k	**80%**	**95%**	**97%**

Controlled comparison (matched 2.0M corpus, same architecture):

Target	Top-1	Top-3	Top-5
SG → post-hoc EG collapse	8.61%	19.37%	26.76%
Direct EG training	**19.32%**	**34.28%**	**43.38%**

Even after collapsing SG predictions into EG space, direct EG training wins decisively

Extinction Groups: Examples

Each extinction group bundles space groups that produce identical systematic absences:

EG	Crystal System	Extinction Symbol	Bundled Space Groups
1	Monoclinic	P – 1 1	$P2_{11}$ (3), Pm11 (6), P2/m11 (10)
2	Monoclinic	P 2_1 1 1	$P2_{111}$ (4), $P2_1$/m11 (11)
4	Monoclinic	P 2_1/c 1 1	$P2_1$/c11 (14)
7	Orthorhombic	P – – –	P222 (16), Pmm2 (25), Pmmm (47), ...
10	Orthorhombic	P 2_1 2_1 2_1	$P2_12_12_1$ (19)
81	Cubic	P – –	P23 (195), Pm-3 (200), P432 (207), P-43m (215), Pm-3m (221)
99	Triclinic	P –	P1 (1), P-1 (2)

Key observations:

- EG 81 bundles 5 cubic space groups that all have no systematic absences (P lattice, no screws/glides)
- EG 4 contains only $P2_1/c$ — one of the most common mineral space groups
- EG 99 (triclinic P1/P-1) is the lowest-symmetry catch-all: a common "sink" for noisy predictions

Physics-Informed Architecture

Encoding crystallographic knowledge into the network

Physics-Informed Vision Transformer

Coordinate Channel

Two-channel input:
intensity + $\sin^2(\theta)$
Physical ruler for
reciprocal space

Physics-Aware PE

Patch-center $\sin^2(\theta)$
passed through an MLP
and added to token
positional embeddings

Split Head

37-bit structured target:
7 crystal system
5 lattice centering
25 operator bits

Auxiliary Head

99-way direct EG
classifier: robust to
noise, learns joint
distribution

Fusion Decoder: $p_{fused} = \alpha \cdot p_{split} + (1 - \alpha) \cdot p_{aux}$

- Split head: structural regularizer — forces attention to weak systematic-absence cues
- Auxiliary head: robust probabilistic classifier — dominates deployment accuracy
- Fusion: complementary failure modes → improved Top-1 over either alone

Positional Encoding: Physics-Aware Patch Embeddings

A standard transformer has no notion of order — it sees a bag of tokens

We need to tell it where each piece of the pattern lives

Standard positional encoding

Adds a fixed index: "I am token #1, #2, #3..."

- Encodes absolute position only
- Position is measured in token index, not diffraction geometry
- But diffraction lives in reciprocal-space coordinates, not patch count

Our physics-aware PE

Each patch's center $\sin^2(\theta)$ → small MLP → embedding vector

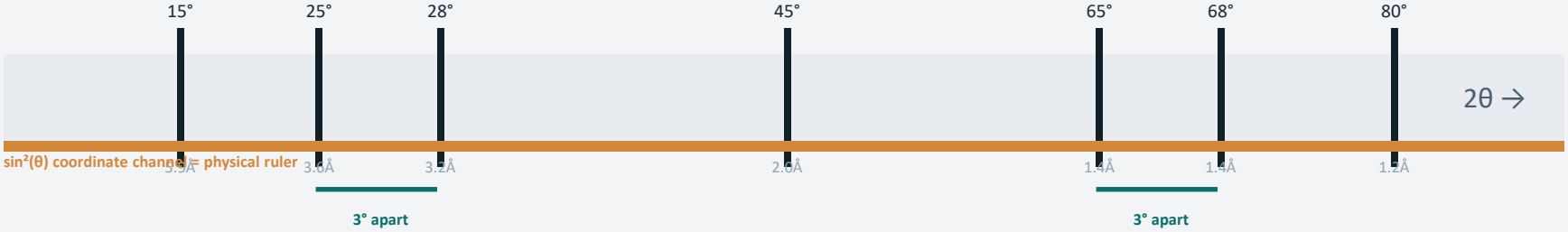
- **This is added to the standard learned absolute PE**
- The MLP learns a smooth reciprocal-space ruler (nearly 1D, monotone in Q^2)

Not RoPE (relative), not sinusoidal — it's a learned absolute map from diffraction geometry

Ablation: removing the physics-aware PE collapses split-head validity to 0% — it carries the critical positional prior for structured rule decoding

But token-level physics PE does not expose the full pointwise geometry inside each patch — that's why we also need the coordinate channel...

Two Encodings, Two Jobs



sin²(θ) Coordinate Channel

"Where am I?"

- Each patch knows its absolute position in reciprocal space
- A peak at d=3.6Å is a fundamentally different reflection than a peak at d=1.4Å
- Without this, the model can't tell low-angle from high-angle features

Answers: what d-spacing is this?

Physics-Aware Positional Embedding

"What reciprocal-space region does this token live in?"

- Each patch's center sin²(θ) → MLP → added to the learned PE
- Tokens from 20° and 70° start with different physical position codes
- The MLP learns a Q²-like ruler (PC1 explains 99.98% of variance)

Biases attention: which reciprocal-space neighborhood am I in?

Neither alone is sufficient. Together they give the model a complete spatial understanding of the 1D reciprocal-space signal.

How Fusion Works

Two heads, two failure modes, one prediction

Split Head (rule path)

Predicts 37 crystallographic bits → deterministic lookup to EG

- Strong when diagnostic absences are clear
- Brittle on noisy real data (logits → diffuse → low confidence)

Auxiliary Head (statistical path)

Direct 99-way softmax over extinction groups

- Robust to noise — learns continuous joint distribution
- Can be overconfident without calibration

$$\text{Fusion:} \quad p_{fused} = \alpha \cdot p_{split} + (1 - \alpha) \cdot p_{aux}$$

Why fusion helps:

- When rule evidence is weak, split logits are near zero → induced EG distribution is diffuse → aux dominates automatically
- When absences are clear, split path sharpens the prediction beyond what aux alone provides
- **If split head contributed nothing new, fusion could not improve Top-1 — but it does**

Training Curriculum

Architecture alone is not enough

Three-Stage Training Curriculum

Phase 1

Uniform Synthetic Pretraining

1.38M patterns balanced across all 99 EGs
Prevents collapse to common low-symmetry classes
Builds symmetry-balanced geometric engine

Phase 2

RRUFF-Style Fine-Tuning

2.35M realistic synthetic patterns
Noise, backgrounds, impurities, broadening
Single largest driver of real-data improvement

Phase 3

Bayesian Inference + Calibration

Empirical Bayes prior from mineral frequencies
Temperature scaling (T=5) resolves overconfidence
Decouples structure from geological bias

Key insight: architecture and data design must be co-optimized. The best transformer fails on real data without the right curriculum.

Calibration & Real-Data Results

Post-hoc calibration changes everything

Temperature Scaling: The Missing Ingredient

The problem: the model is frozen at T ≈ 0

- Fine-tuning absorbs geological frequency into the weights (logits already encode the prior)
- Adding an external Bayesian prior then double-counts geological bias
- Result: the system is "too cold" — locked into the ground state (most common class)

The fix: warm it up. Divide logits by T=5 before adding the log-prior

- This reduces effective energy differences, letting other states (rarer EGs) compete again

Uncalibrated (T=1)

Top-1: 2.15% | Top-5: 14.46%

ECE: 0.457 | NLL: 6.58

Calibrated (T=5)

Top-1: 9.54% | Top-5: 43.08%

ECE: 0.049 | NLL: 3.70

Same model weights. No retraining. Just calibrated inference.

Topological Error Analysis

Not all errors are created equal

Errors on the Subgroup Graph

Standard Top-1 treats every misclassification as equally wrong

But crystallographically, some errors are much more reasonable than others

We map predictions onto the DAG of maximal translationengleiche subgroups

What is this graph?

- Nodes = 99 extinction groups
- Edges = loss of a single symmetry operation (without changing the translation lattice)
- Moving down = losing symmetry (descendant)
- Moving up = gaining symmetry (ancestor)

This lets us ask:

- How far away (in graph hops) are the model's errors?
- Do errors tend to go up (hallucinating symmetry) or down (conservative)?

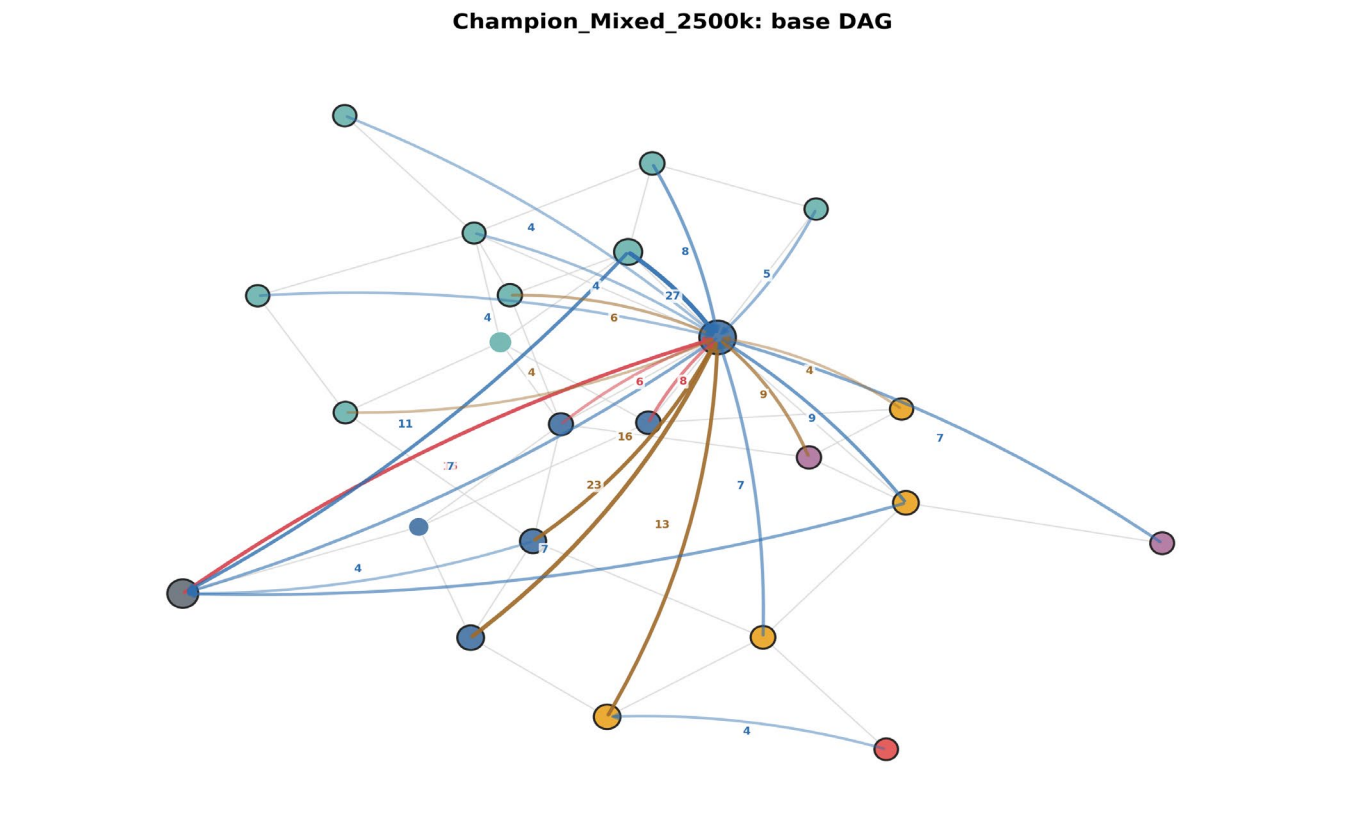
Topological Error Structure on RRUFF-325

Model	Top-1	Mean dist.	≤2 hops	Desc./Anc.
Legacy aux (uncalib.)	4.92%	3.50	17.8%	64/158
Stage-2c (calibrated)	9.54%	2.72	**38.4%**	**191/27**
PO-only, 1 epoch	**13.54%**	2.51	**53.0%**	157/28
Final large mixed	10.46%	**2.46**	51.6%	165/31

Key findings:

- Calibration makes errors more local: mean distance drops from 3.50 to 2.72
- Calibrated model is conservative: 191 descendant vs. 27 ancestor errors
- *Like a cautious crystallographer: when unsure, default to lower symmetry*
- Legacy model hallucinated symmetry: 158 ancestor errors

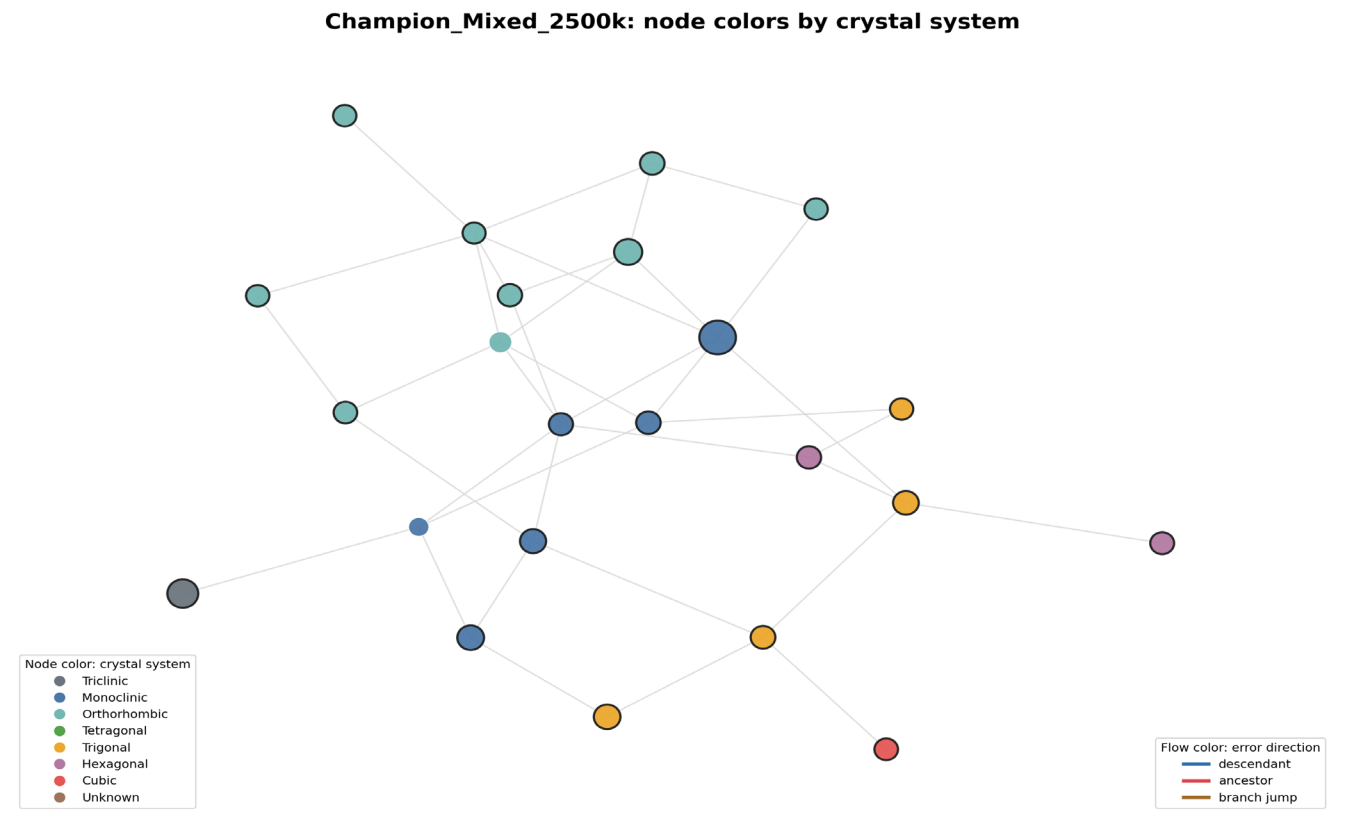
Visualizing Errors on the Subgroup DAG



Base DAG

- Nodes = extinction groups
- Edges = subgroup relations
- Node size = error count

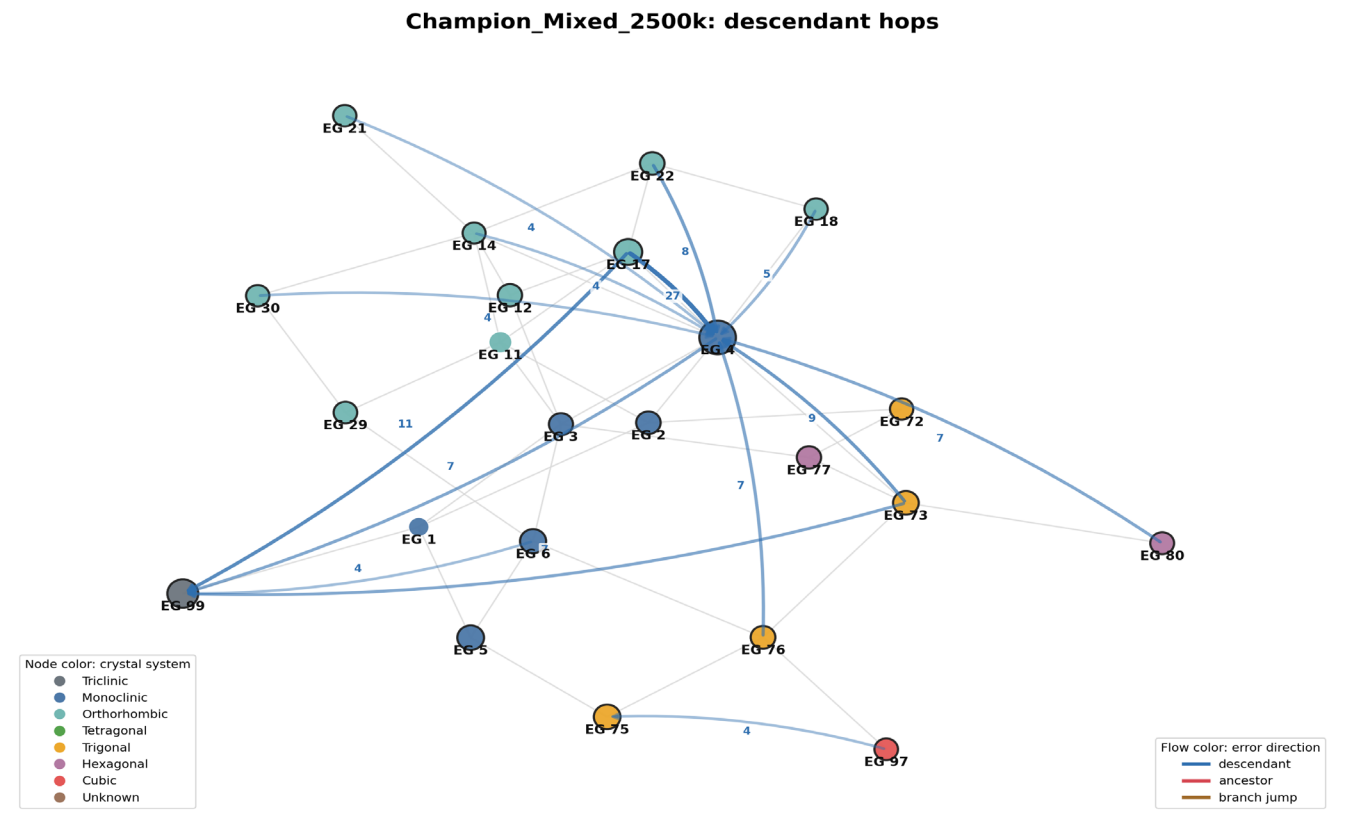
Crystal System Neighborhoods



Crystal systems

- Node color = crystal system
- Errors cluster within crystallographic neighborhoods

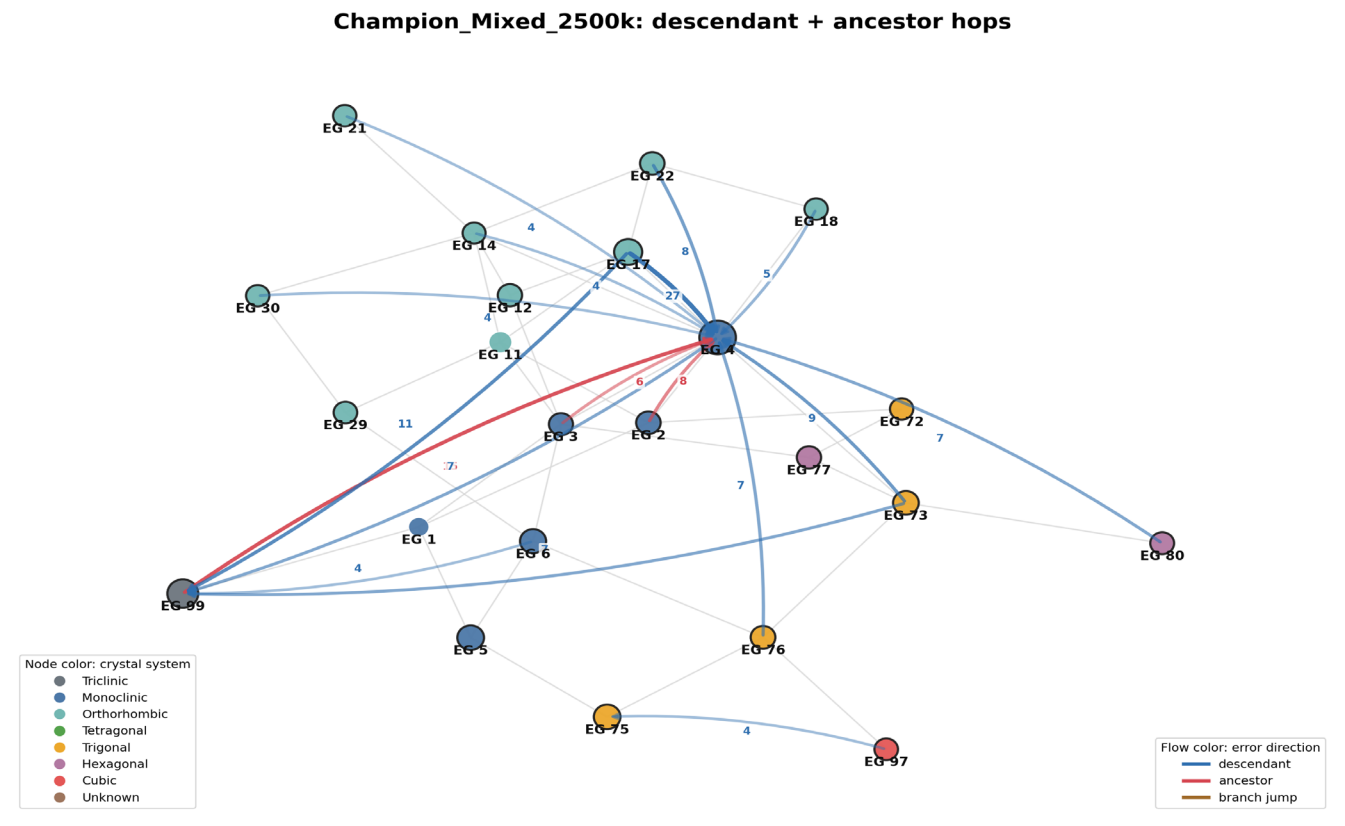
Descendant Errors: Conservative Defaults



Descendant flows

- 165 errors flow downward to lower symmetry
- *Model acts like a cautious crystallographer*
- Noise adds intensity where absences should be → reject higher symmetry

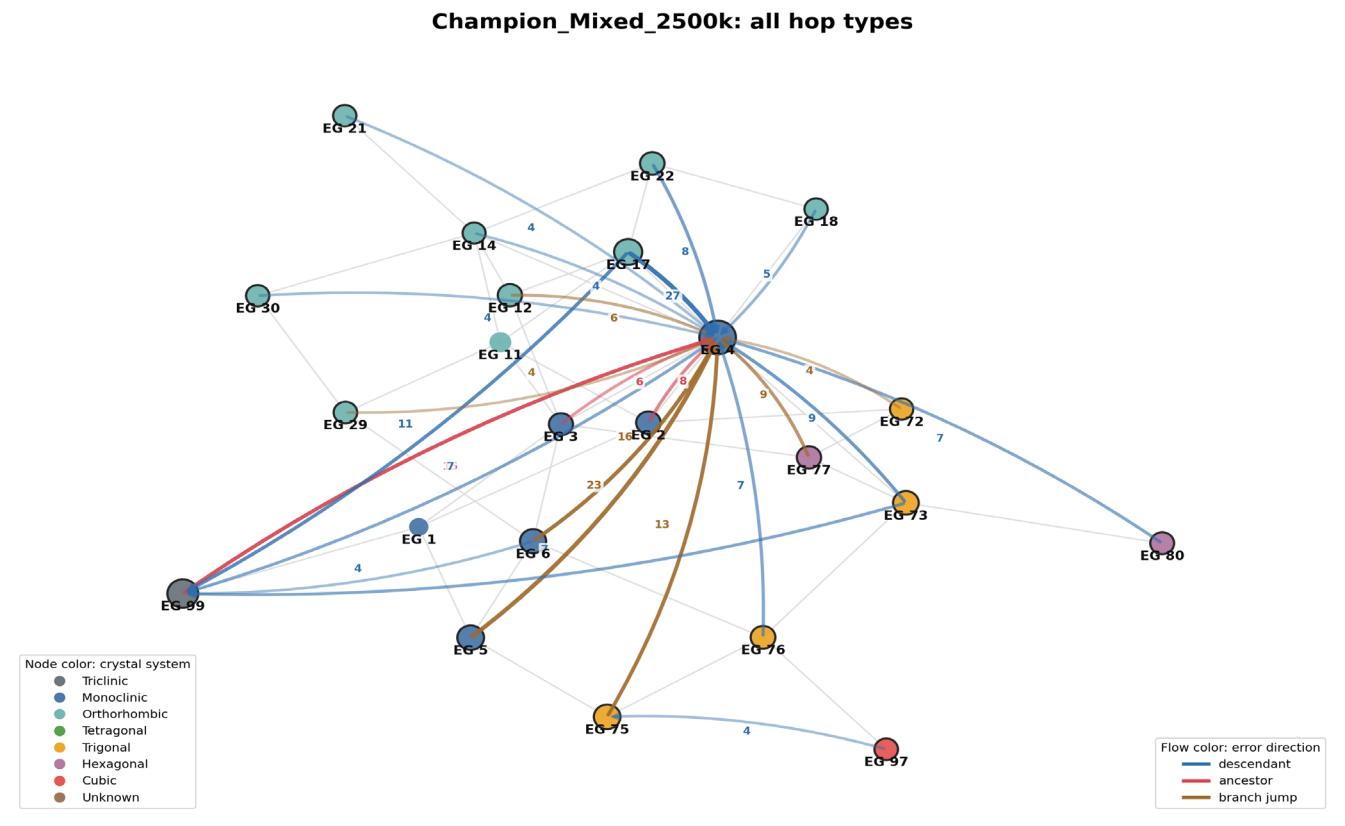
Ancestor Errors: Symmetry Hallucinations



Ancestor flows

- Only 31 errors hallucinate higher symmetry
- Strongly suppressed vs. legacy model (158 ancestors)
- Calibration is key to eliminating false high-symmetry guesses

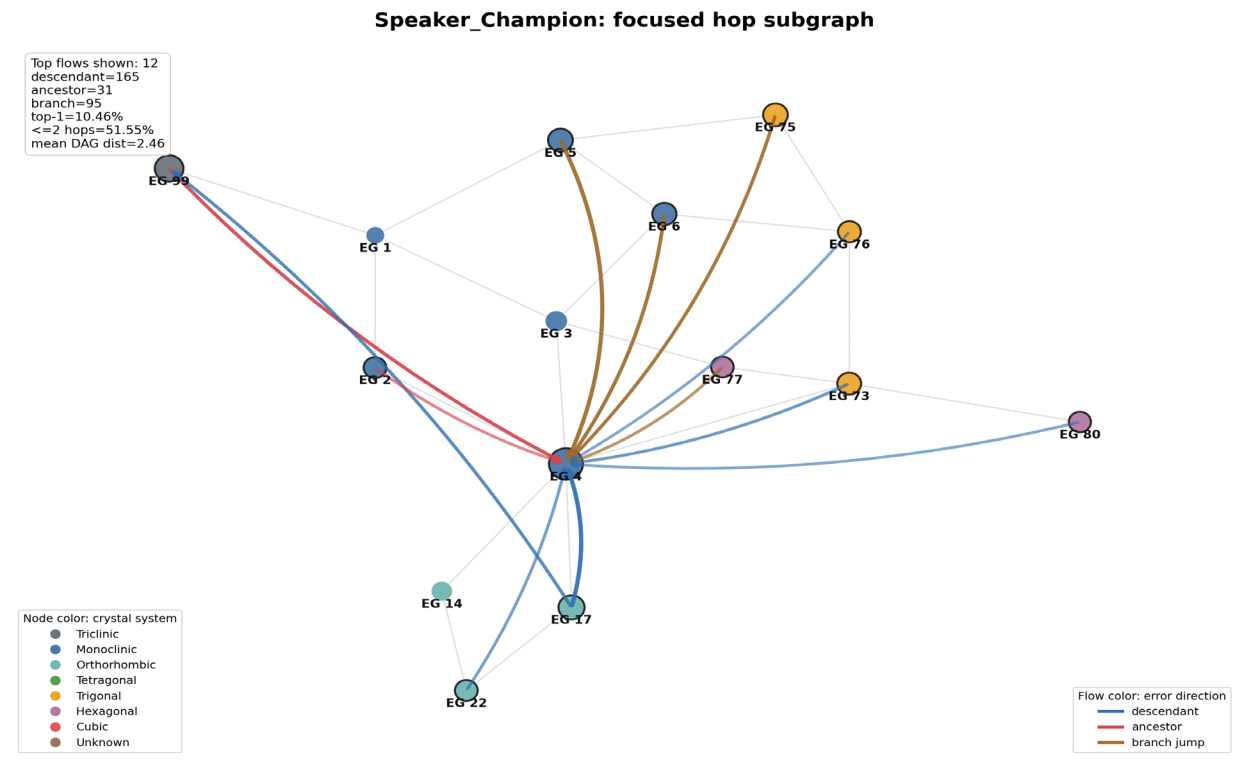
Branch Jumps: Texture Aliasing



Branch jumps

- 95 errors hop to crystallographic cousins
- PO erases diagnostic reflections → nearby cousins become ambiguous
- *"Texture aliasing" limit of 1D powder data*

Complete Error Flow: Final Mixed Model



Champion model summary

- Top-1: 10.46%
- **Mean DAG dist: 2.46**
- ≤2 hops: 51.55%

- 165 desc / 31 anc / 95 branch

Errors are local, physically structured, and emerge without topological supervision

Preferred Orientation Changes the Error Landscape

What is preferred orientation? In an ideal powder, crystallites point in all directions equally and every reflection gets its fair share of intensity. In real samples, plate-like or needle-like crystals tend to align (e.g. flat on the sample holder). This systematically strengthens some reflections and suppresses others — distorting the pattern away from the ideal powder average.

The physics of descendant bias:

- Higher-symmetry EGs mandate strict absences (zero intensity)
- Real noise and PO add or remove intensity where symmetry demands specific values
- Model rejects unsupported higher symmetry → conservative descendant default

What PO-aware training does:

- We add March-Dollase PO to synthetic training → model learns that some missing reflections are texture artifacts, not true absences
- This weakens descendant bias and tightens the local error neighborhood

- *Texture aliasing: once PO erases diagnostic reflections, nearby crystallographic cousins become ambiguous → lateral branch hops*

Graceful Topological Degradation

The model's errors are not random — they are physically structured

Uncalibrated legacy model:

- Mean graph distance 3.50, mostly ancestor hallucinations (guessing higher symmetry)

Calibrated non-PO model:

- Mean distance 2.72, strongly descendant-biased (conservative)
- *Like a cautious crystallographer: rejecting unsupported symmetry*

PO-aware mixed model:

- Mean distance 2.46, tightest locality of all
- Learns that missing reflections can be texture artifacts, weakening descendant bias
- Branch hops increase → "texture aliasing" limit of 1D powder data

This graceful degradation emerges without any explicit topological supervision

What Is Pawley Fitting?

A classical (non-ML) method for analyzing diffraction patterns:

Given a candidate space group and approximate lattice parameters:

- 1. Calculate where peaks should appear (from Bragg's law + symmetry)
- 2. Fit the peak intensities freely (no atomic model needed)
- 3. Measure how well the calculated peak positions match the data

Unlike Rietveld refinement (which needs a full atomic model), Pawley fitting only needs the lattice and symmetry — making it ideal for symmetry ranking.

The problem with using it alone:

- A global sweep over all candidate space groups takes hours per pattern
- Lower-symmetry models can overfit noise, and accidental absences mimic true absences
- Without a strong prior on where to look, the search is unstable and brittle

Solution: use the neural network to narrow the search, then Pawley to verify locally

From Neural Proposer to Classical Verifier

The topology enables a practical hybrid pipeline:

1. Neural model proposes

Calibrated ViT localizes the search to a small DAG neighborhood

- Typically ≤2 hops from ground truth
- Seconds, not hours

2. Pawley verifies locally

Classical sparse Pawley fitting ranks only candidates within bounded branch

- Completes in seconds to minutes
- vs. hours for global sweep

3. Bounded result

Space-group ranking within the local branch

- Physically local even when not exact

Results on RRUFF-325:

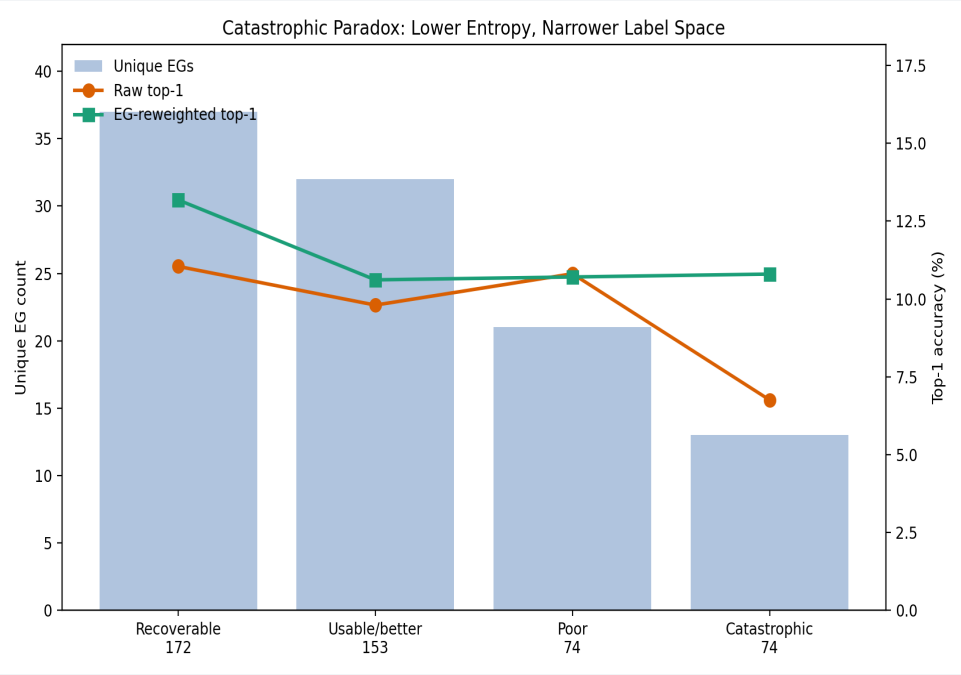
- Rhombohedral branch: 34 cases, all completed — 8/25 hR cases exact top-ranked space group
- Cubic face-centered: 9 cases, all completed — locally ambiguous but stays within cF branch
- Monoclinic: hardest regime — 61 cases attempted, 51 completed after stabilization (low-angle truncation, reflection clustering, NNLS solver)

Key insight: the NN converts an intractable global classical search into a bounded local verification problem

The Catastrophic Paradox

When worse data is easier to classify

The Catastrophic Paradox



Surprising finding:

Patterns with the worst Rietveld fits are among the easiest to classify!

Explanation:

- Catastrophic patterns: only 13 EGs, 38 minerals (narrow target)
- Recoverable patterns: 37 EGs, 111 minerals (broad target)
- *Lower label-space entropy makes classification easier even when data quality is worse*
- Many catastrophic minerals are highly textured/cleavable — the neural network recognizes the Bragg-angle topology even when intensities are badly distorted

Context: Comparison with Prior Work

How does this compare?

- Schopmans et al.: ~25% Top-1 on 942 hand-filtered RRUFF patterns (145 SGs, 261M training patterns)
- Our best Top-1: 16.70% on RRUFF-473 (99 EGs, no exclusions, algorithmically curated)
- Our best Top-5: 52.22% on RRUFF-473

Key differences that prevent direct comparison:

- Classification target: space groups vs. extinction groups
- Evaluation set: hand-filtered vs. algorithmically curated (retains hard patterns)
- Data scale: 261M vs. 5.14M training patterns
- Data generation: ICSD-informed vs. physics-first curriculum

Prior-only baseline (no diffraction pattern): ~9.7% Top-1

- Our Top-5 of 43-52% far exceeds this → model uses real diffraction evidence to re-rank

Summary

Key Results

- Extinction groups are the right target (8.6% → 19.3% in matched control)
- Transformers outscale CNNs on diffraction
- Physics-informed architecture matters
- Three-stage curriculum bridges sim-to-real gap
- Post-hoc calibration is essential, not cosmetic
- Errors are topologically local and physically interpretable
- Graceful degradation without topological supervision

Design Pattern for Scientific ML

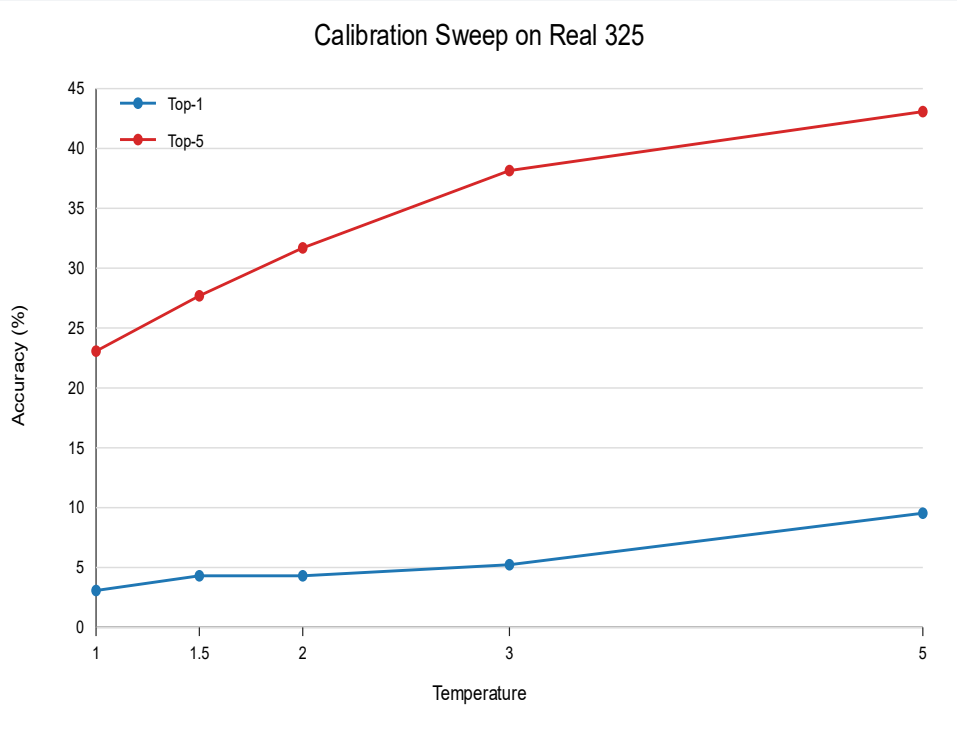
- Encode measurement physics in the architecture
- Let curriculum manage domain shift
- Treat calibrated inference as a first-class design parameter

the physics must be in the architecture,
the care must be in the data,
and inference must be calibrated.

arXiv:2604.28164 | william.ratcliff@nist.gov

Backup Slides

Calibration Sweep on RRUFF-325



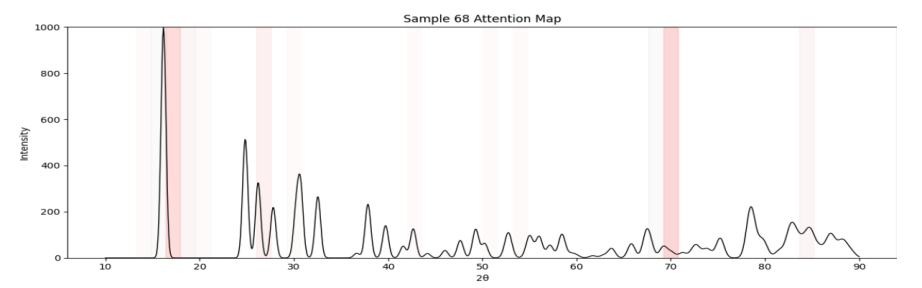
Softening the auxiliary logits (T=5) before adding the geological prior:

- Decouples structural evidence from geological overconfidence
- Temperature scaling restored usable uncertainty

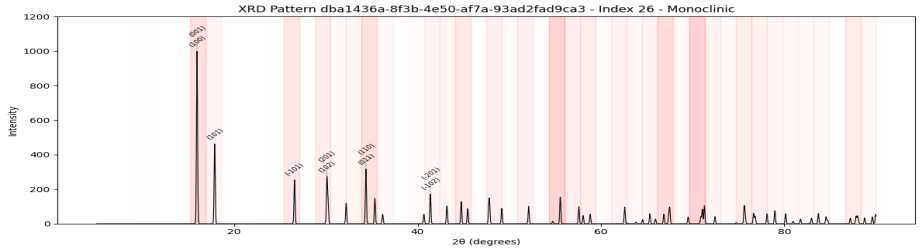
What Does the Model See?

Attention maps and saliency analysis

Attention Maps: Transformer vs. CNN



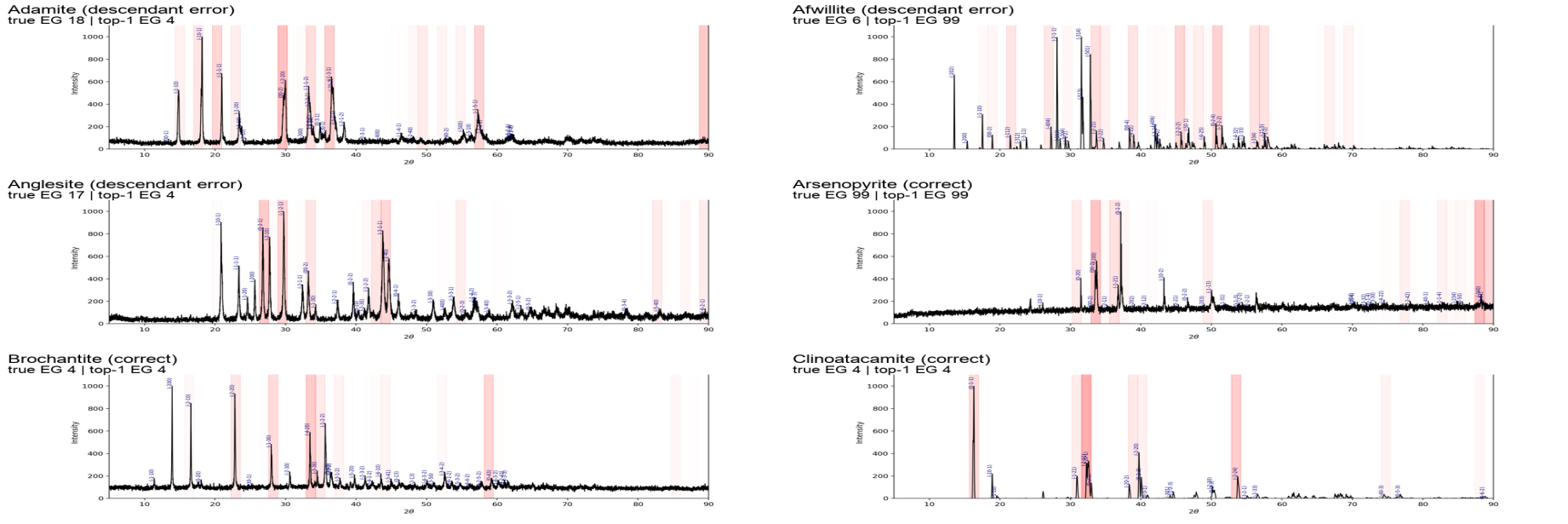
Regular Transformer



Vision Transformer

- **Transformers attend to both primary and minor peaks**
- CNN saliency concentrates on tallest peaks only
- *Caveat: qualitative diagnostics, not proof of mechanism*

Calibrated Attention on Real Minerals



Six Stage-2c calibrated attention overlays on RRUFF-325