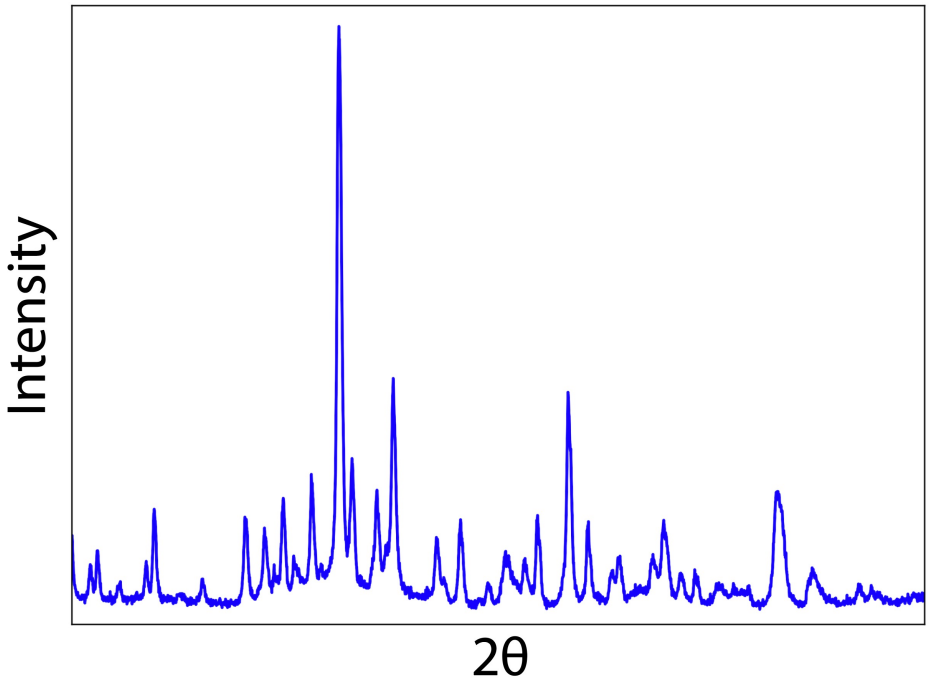
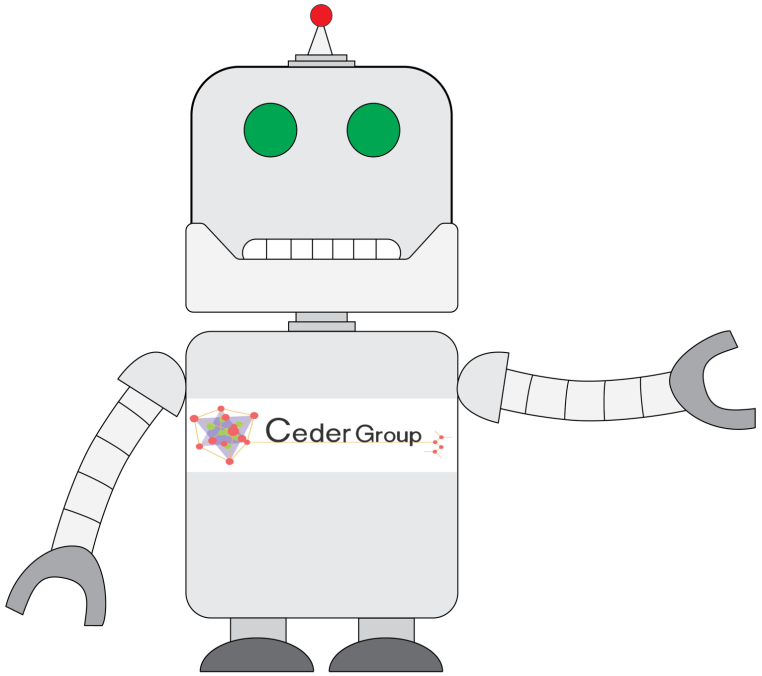
Deep learning to automate phase identification from multi-phase XRD spectra





Phases	Probabilities
$LiMnO_2$	92%
Li_2TiO_3	88%
LiF	64%

Nathan Szymanski, Ceder group at UC Berkeley

Preprint: https://arxiv.org/abs/2103.16664

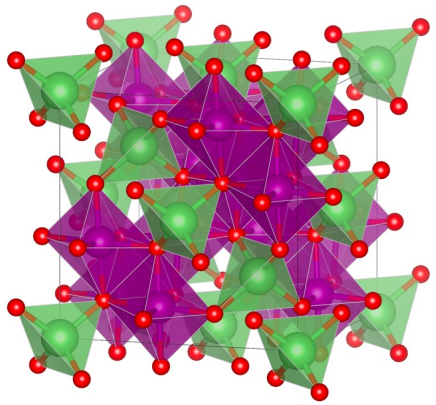
Github: https://github.com/njszym/XRD-AutoAnalyzer

Email: nathan_szymanski@berkeley.edu

Berkeley UNIVERSITY OF CALIFORNIA

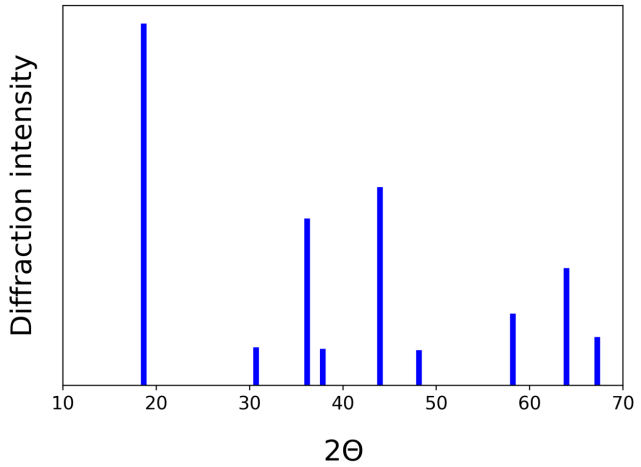
BERKELEY LAB Lawrence Berkeley National Laboratory

XRD: a cornerstone of inorganic materials research

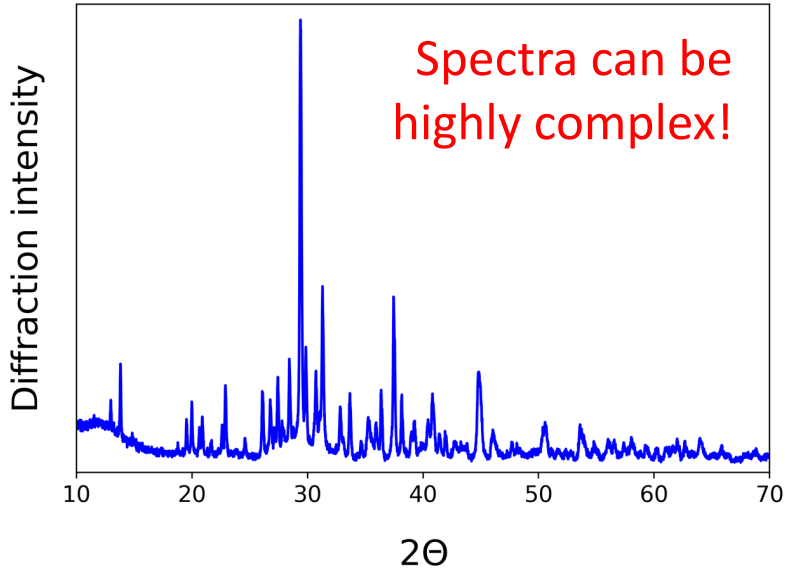


Crystal structure ↔ X-ray diffraction pattern

XRD: a *chemical fingerprint* for phase ID



Characterization for exploratory syntheses:



What's in the sample? Solving this problem widely remains a manual task

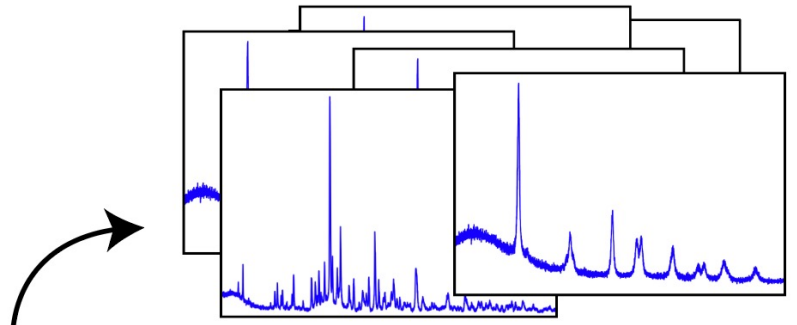
A growing need to reliably automate XRD analysis

Robotic hardware is accelerating experimentation – can software keep up?

1) **High-throughput** and **parallelized experiments**

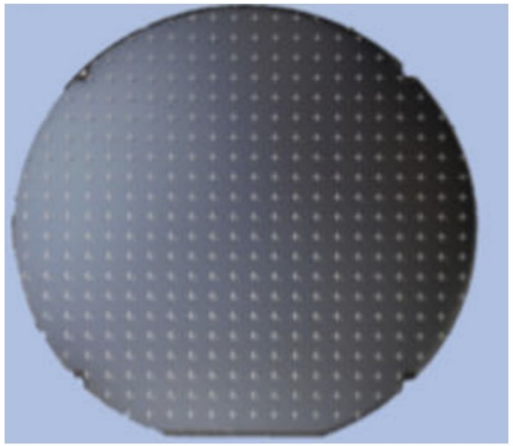
- Generate large amounts of data requiring analysis
- EX: combinatorial thin film synthesis

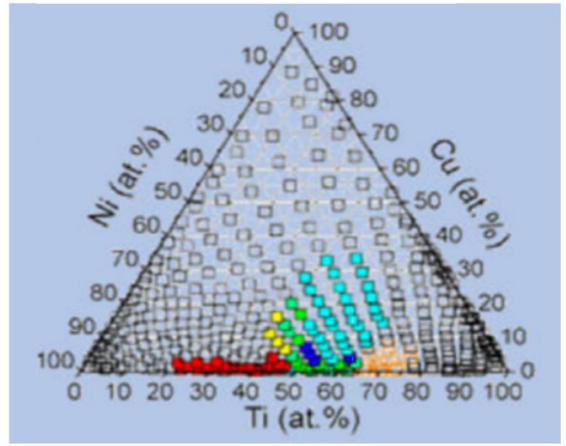
XRD dataset



Phase diagram

Grid of samples with distinct compositions





Ludwig, npj Comp. Mat. (2019).

A growing need to reliably automate XRD analysis

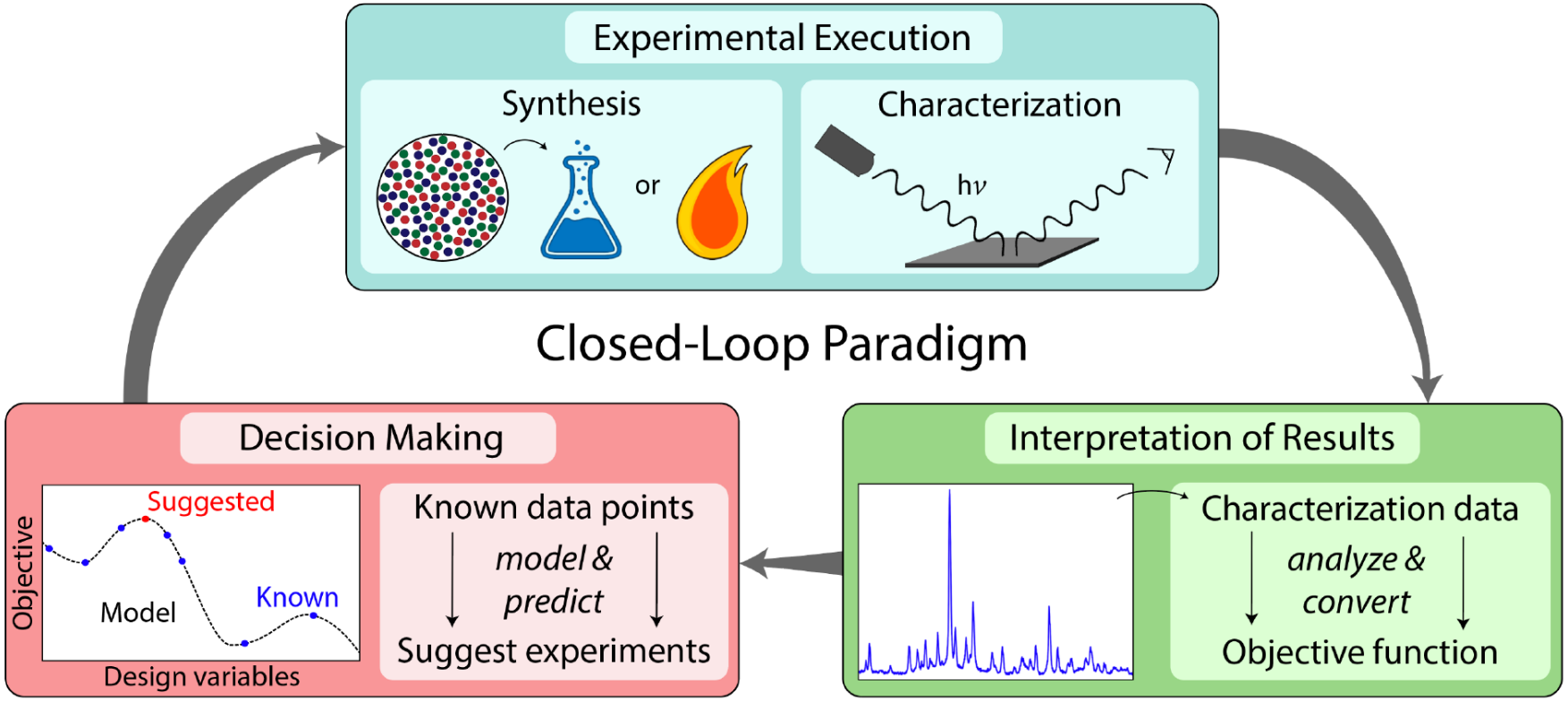
Robotic hardware is accelerating experimentation – can software keep up?

2) Self-driving laboratories for **closed-loop experiments**

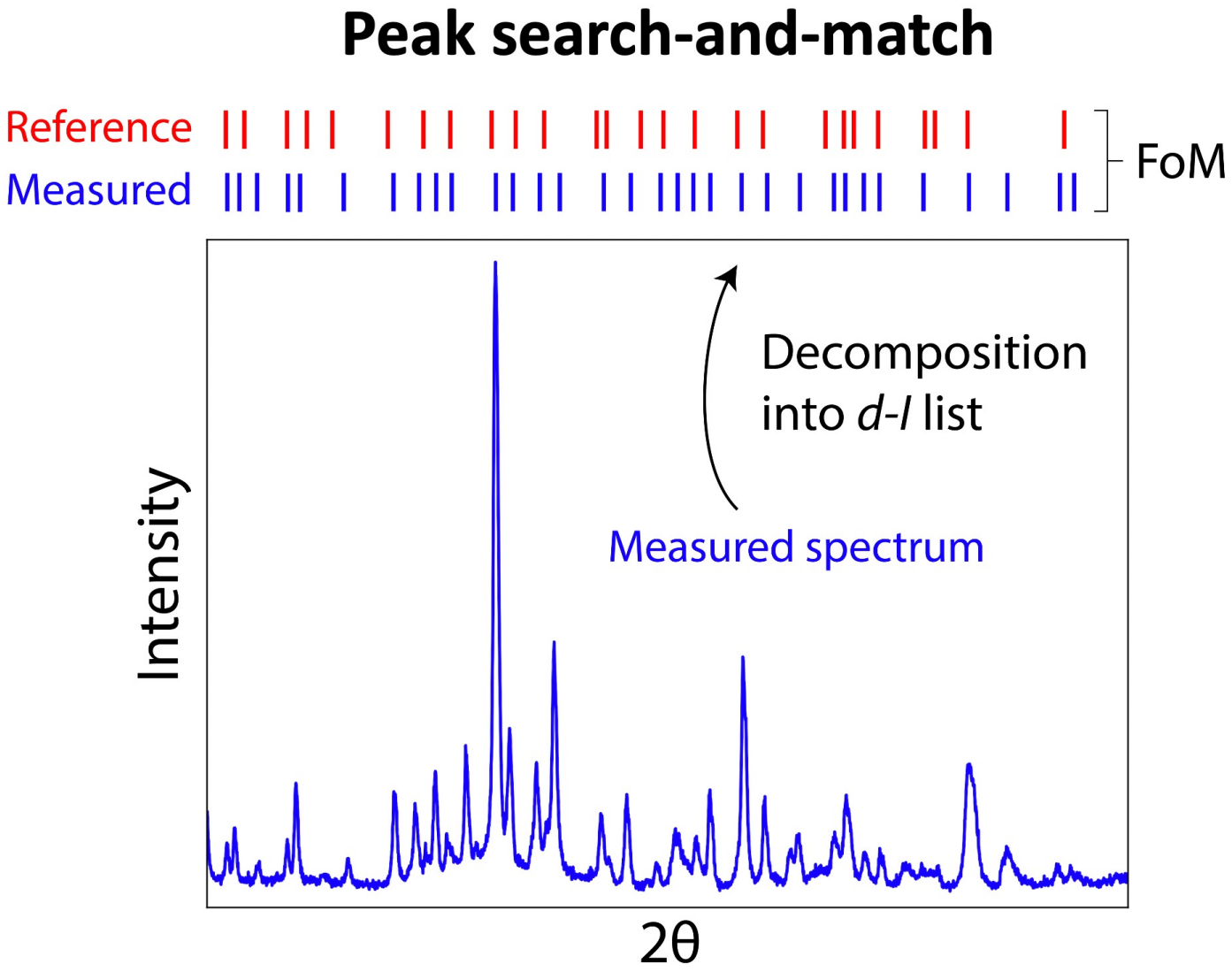
- Experiments are guided by previous results
- Goal is to remove human from the loop

Can we make inorganic materials synthesis a rapid and autonomous process?

Reliable phase identification is critical to learn from each synthesis attempt



Traditional methods to automate phase ID



Approach:

- Extract peak positions and intensities
- Compare with known materials (e.g., from the ICSD or ICDD)

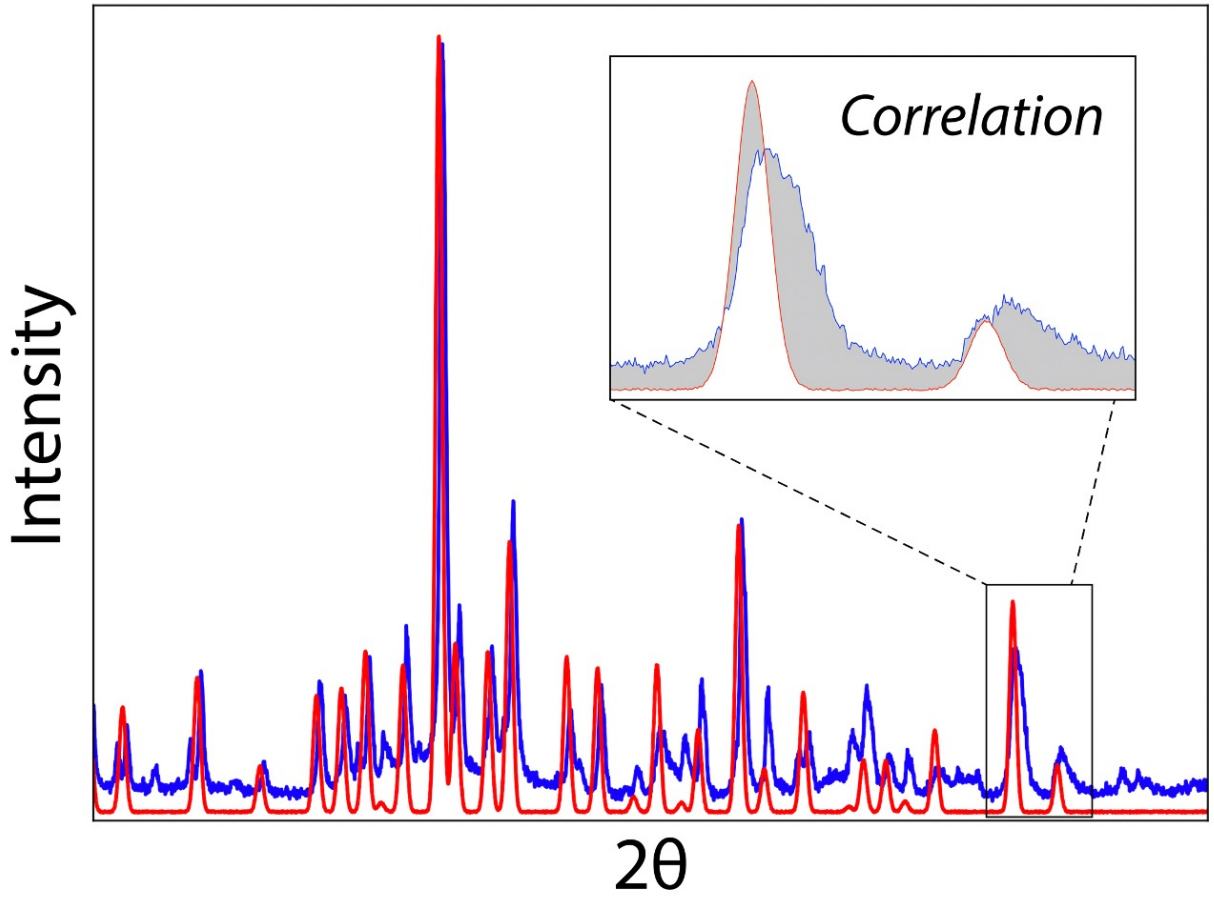
Limitations:

- Peak extraction is unreliable when:
  - Peaks overlap
  - Peaks “blend in” with noise
  - Impurities are present

Traditional methods to automate phase ID

Full-profile analysis

Measured spectrum & Simulated spectrum



Approach:

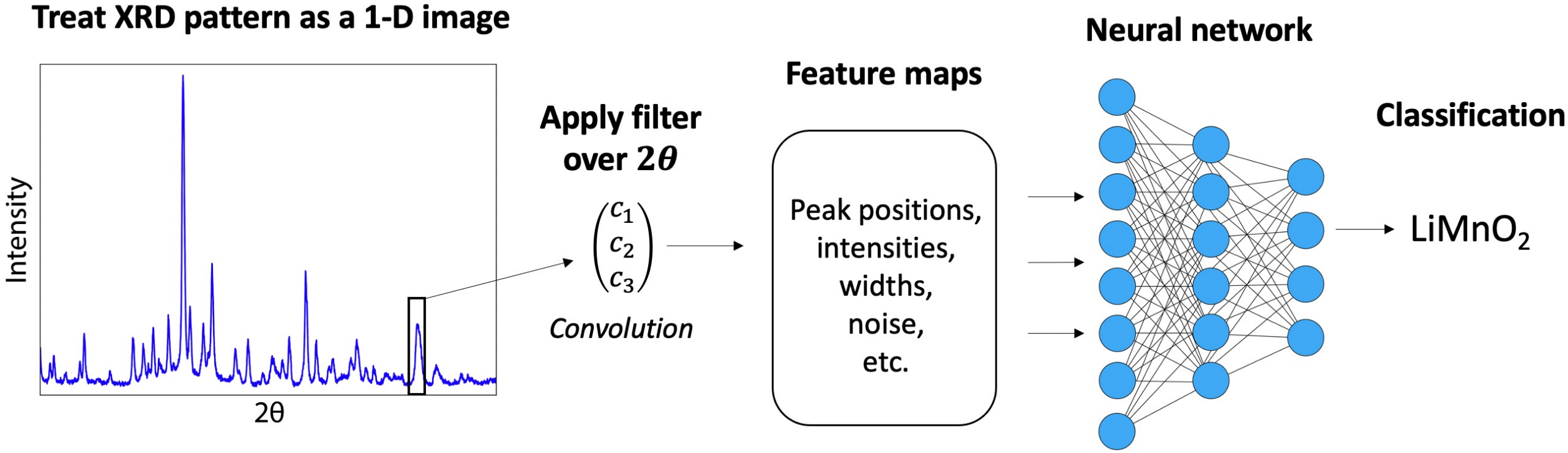
- Simulate spectra of known phases
- Quantify overlap between measured and simulated spectra

Limitations:

- Patterns become dissimilar when experimental artifacts are present, common artifacts include:
  - Strain, texture, small particle size, poor crystallinity

An improved approach: convolutional neural networks

Convolutional neural networks (CNNs) are widely used for image classification



Recent work suggests that CNNs outperform traditional methods for:

- Symmetry classification
- Single-phase identification

Park *et al.*, IUCrJ (2017).
Oviedo *et al.*, npj Comp. Mat. (2019).
Lee *et al.*, Nat. Commun. (2020).
Maffettone *et al.*, Nat. Comp. Sci. (2021).

New developments in our work

1) Physics-informed data augmentation

2) Accounting for non-stoichiometry

3) A probabilistic treatment of multi-phase mixtures

Training the CNN using simulated XRD spectra

What data is used to train the CNN?

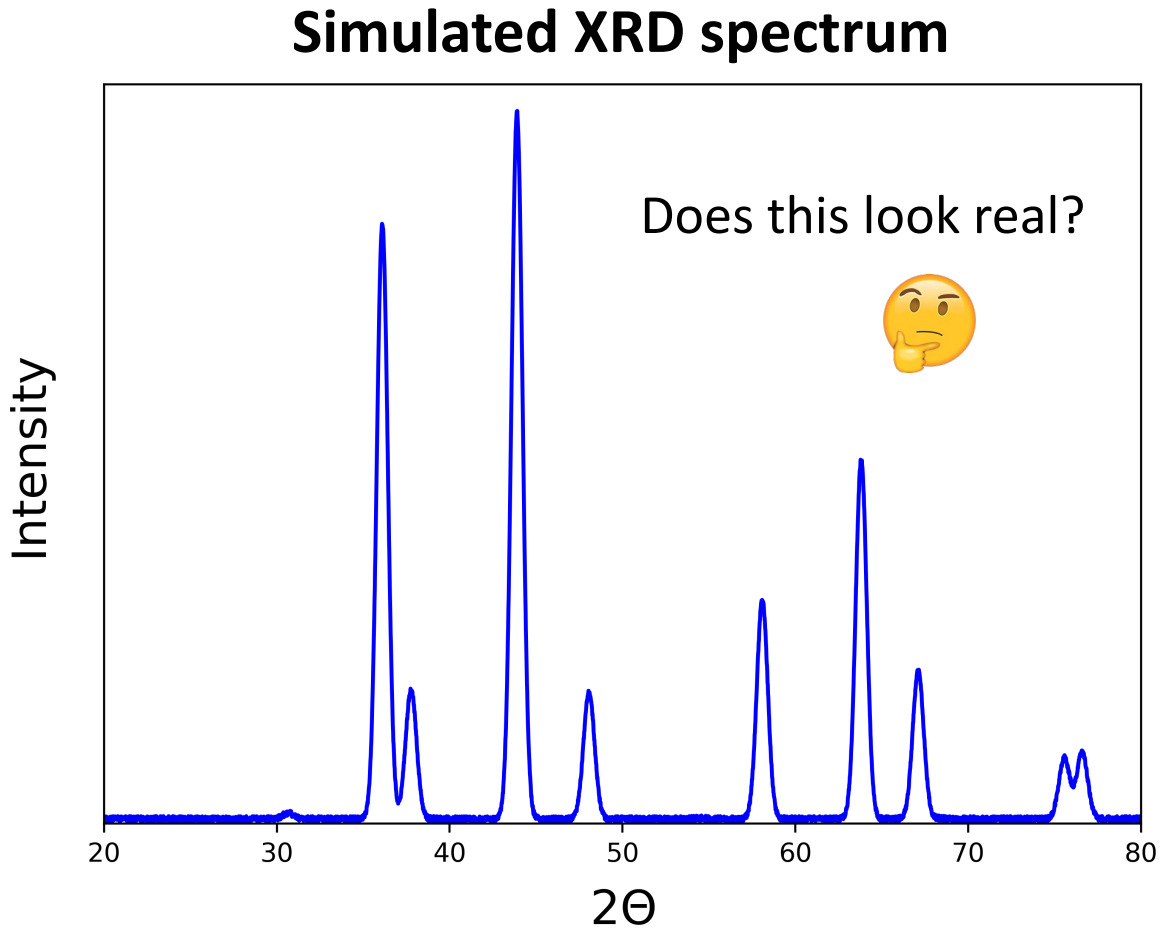
- Limited number of experimental XRD spectra
- Luckily, XRD spectra are easily simulated

Issues to consider:

- One spectrum per phase is not enough
- Ideal spectrum may not reflect experiment

Solution: perform *data augmentation*

Key question:
How to augment simulated spectra?



Physics-informed data augmentation

Derive perturbations to spectra:

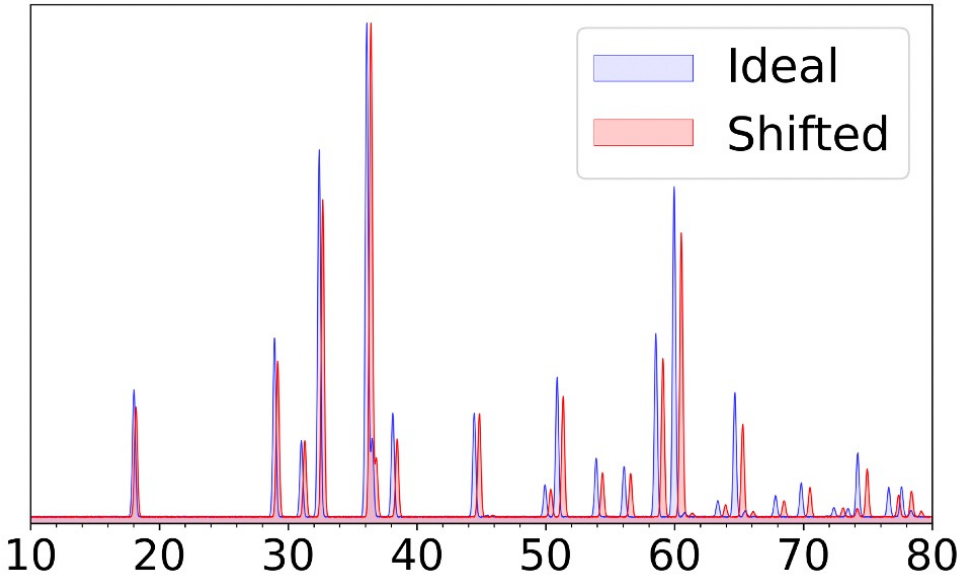
1) Shifts in peak positions ↔ strain in the unit cell:

$$\sigma_{ij} = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} \\ \sigma_{31} & \sigma_{32} & \sigma_{33} \end{pmatrix}$$

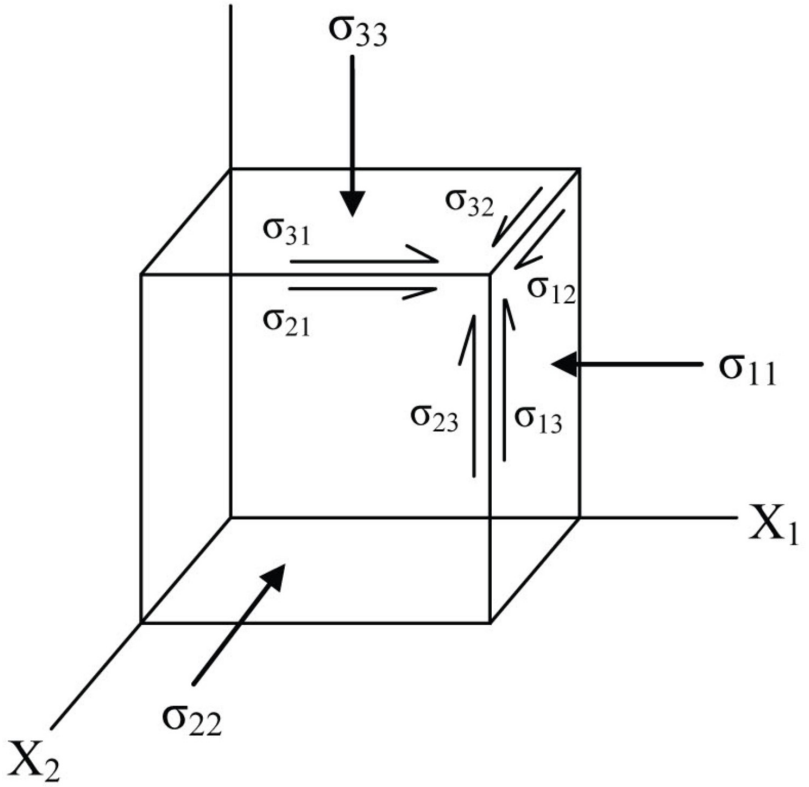
Must preserve symmetry!

2) Variations in peak intensities ↔ texture in the powder:

3) Broadening of peak widths ↔ small domain size:



Bounds: $\pm 4\%$ strain



Physics-informed data augmentation

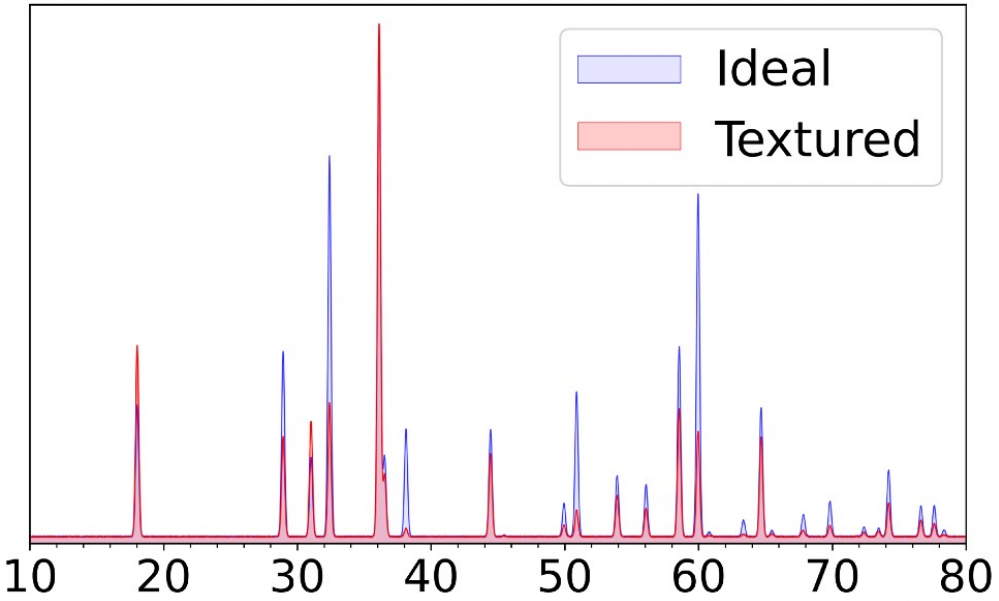
Derive perturbations to spectra:

1) Shifts in peak positions ↔ strain in the unit cell:

2) Variations in peak intensities ↔ texture in the powder:

$$\text{Intensity scaling} \propto \langle hkl \rangle_{\text{preferred}} \cdot [h'k'l']_{\text{peak}} \qquad \text{Bounds: } \pm 50\% \text{ intensity}$$

3) Broadening of peak widths ↔ small domain size:



Physics-informed data augmentation

Derive perturbations to spectra:

1) Shifts in peak positions ↔ strain in the unit cell:

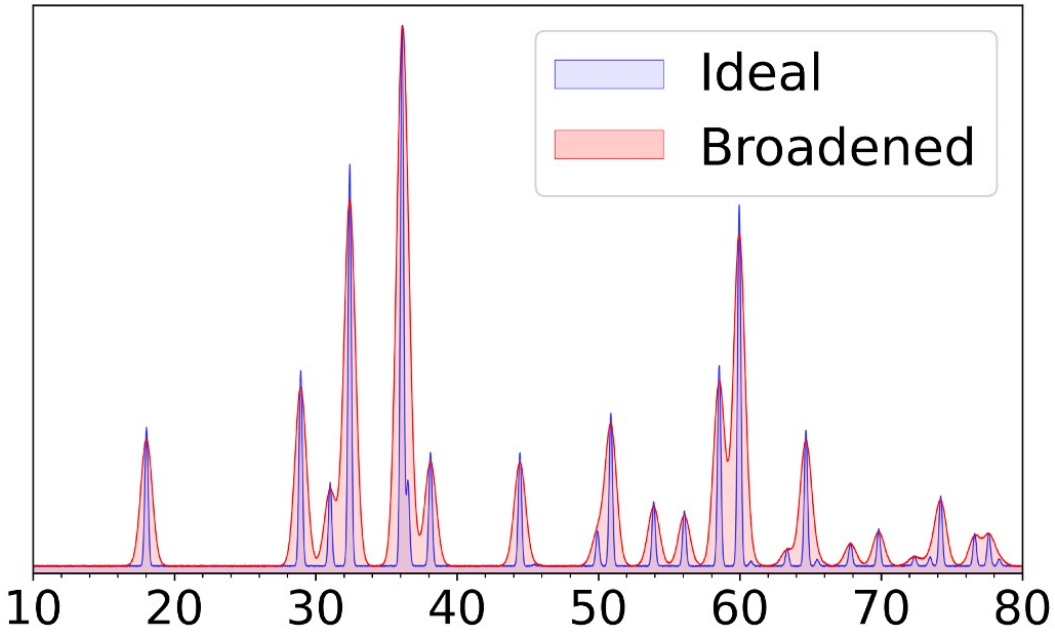
2) Variations in peak intensities ↔ texture in the powder:

3) Broadening of peak widths ↔ small domain size:

Scherrer equation: $\tau = \frac{K\lambda}{\beta cos\theta}$

$\text{FWHM } (\beta) \propto^{-1} \text{ domain size } (\tau)$

$\text{Bounds: 1 nm (broad) to 100 nm (narrow)}$

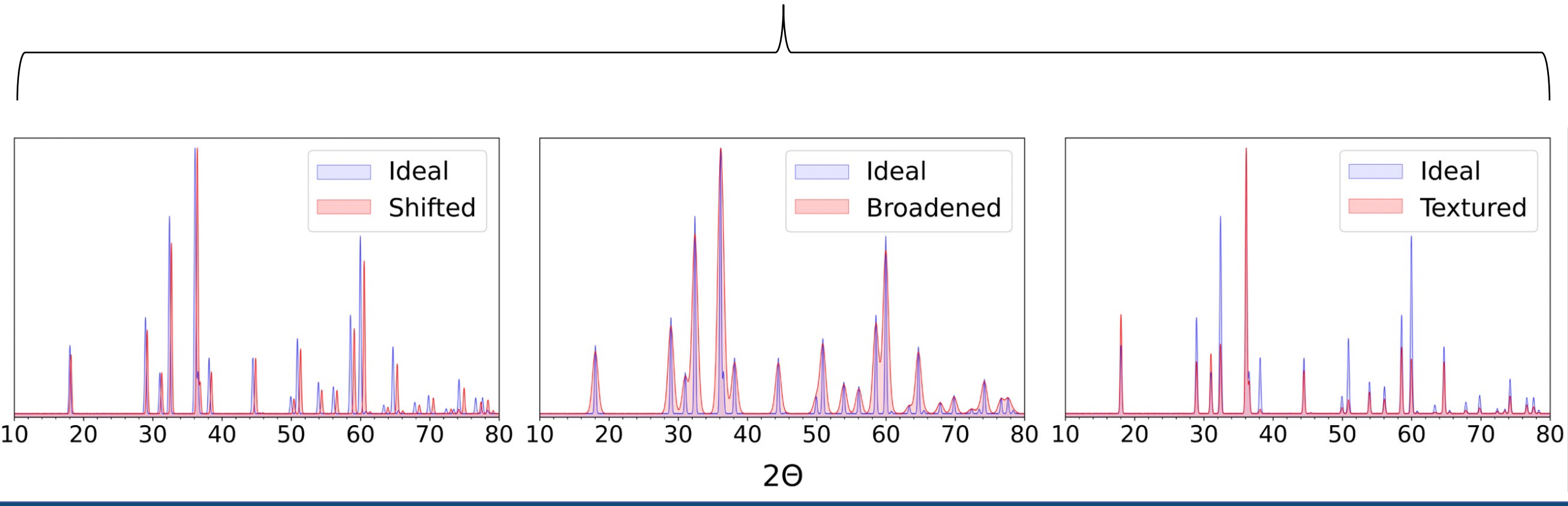


Physics-informed data augmentation

For a given reference phase:

50 spectra simulated from each artifact; changes sampled from normal distribution

→ 150 augmented spectra (per phase) to train the CNN

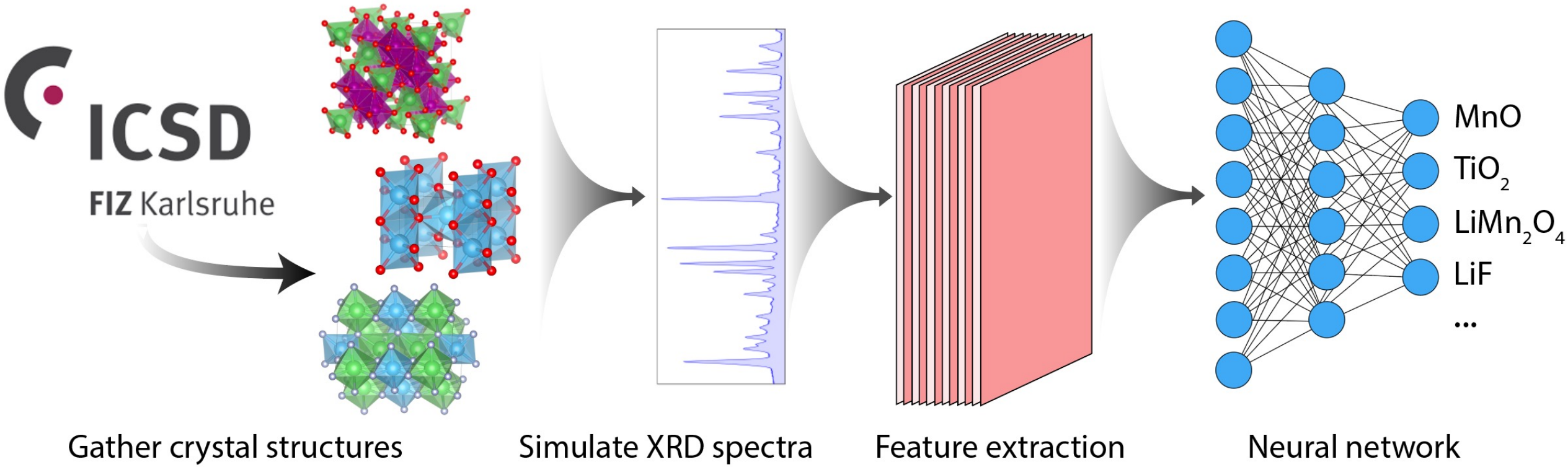


Training a model for the Li-Mn-Ti-O-F chemical space

Li-Mn-Ti-O-F:

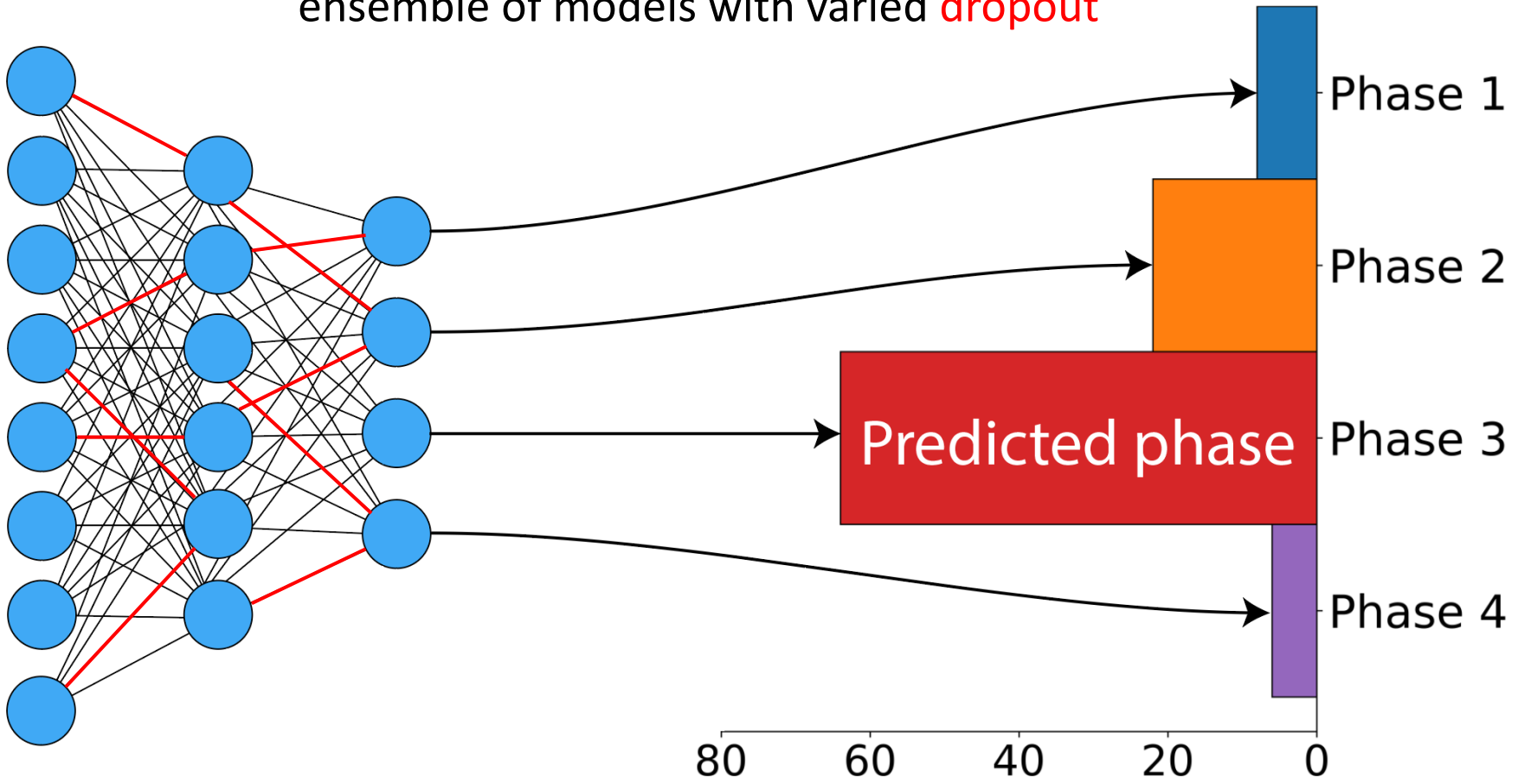
- Useful system for many battery materials
- Challenging test case for diffraction due to similarity in Mn/Ti and O/F scattering factors

Training data: 140 phases extracted from the ICSD → **21,000 simulated spectra**



Ensemble approach to yield probabilistic predictions

Make many predictions using an
ensemble of models with varied dropout



Ensemble NN

Probability distribution (%)

Test results on single-phase spectra

From the simulated spectra:

- 80/20 split for training/testing

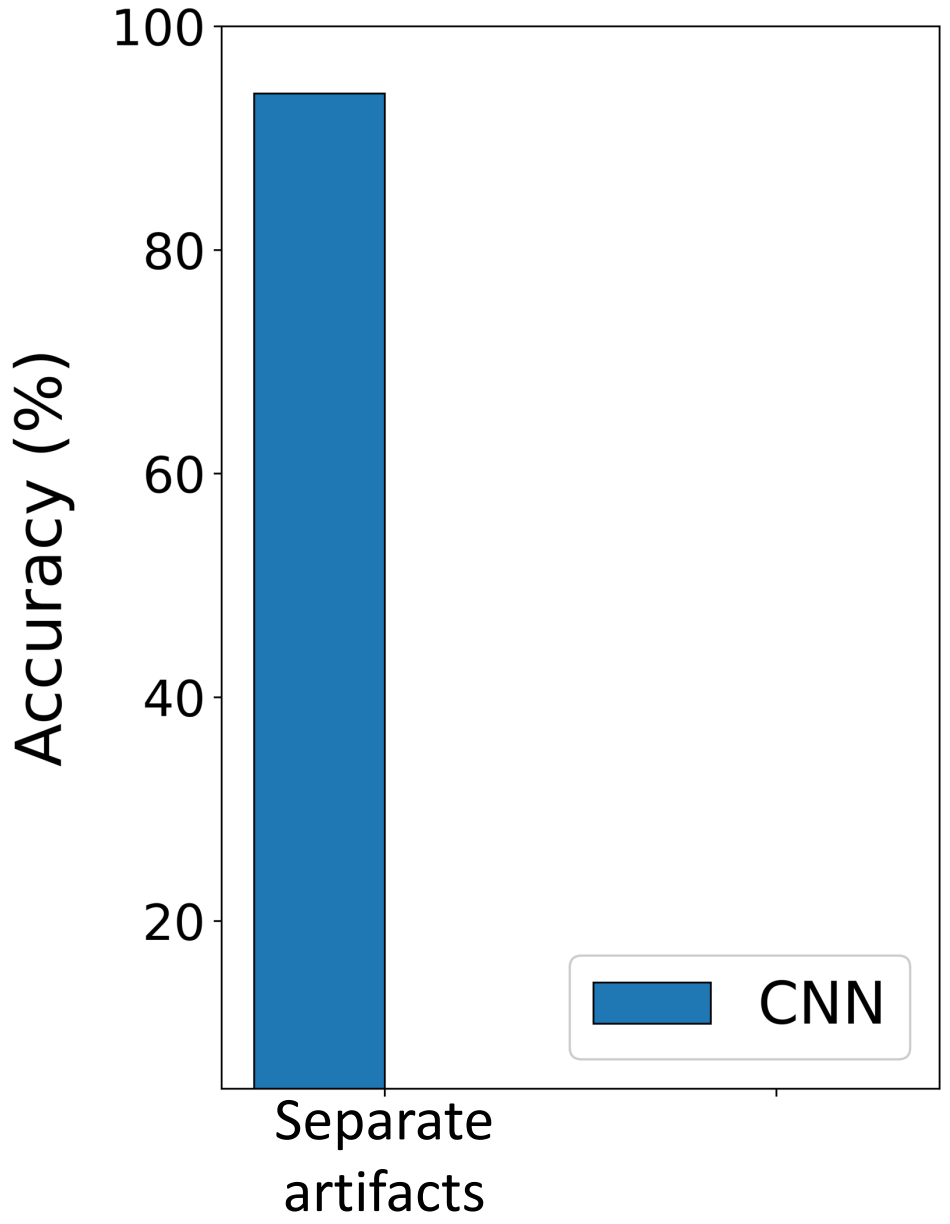
Test results from the CNN:

- **94% of all test spectra are correctly classified**

Accuracy is promising, but what is our baseline?

Training set: 16,800 spectra

- ✓ Single-phase
- ✓ Stoichiometric
- ✓ Separate artifacts
- ✓ Peak perturbations
- ✗ Background effects
- ✗ Mixed artifacts



Test results on single-phase spectra

JADE from MDI:

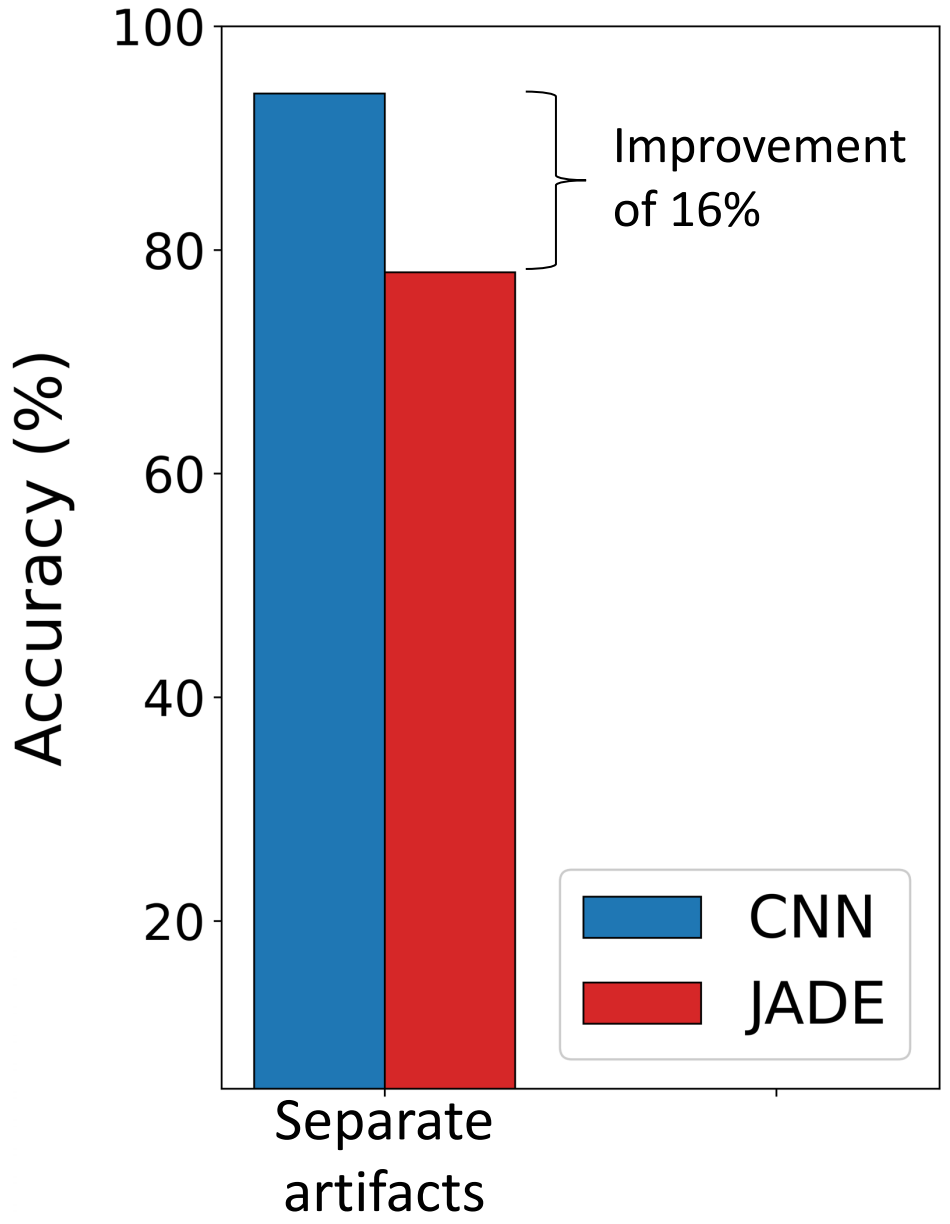
- Peak search-match + full-profile comparison

Results from JADE:

- **Only 78% of all spectra are correctly classified**

Training set: 16,800 spectra

✓ Single-phase	✓ Peak perturbations
✓ Stoichiometric	✗ Background effects
✓ Separate artifacts	✗ Mixed artifacts



Test results: our CNN outperforms traditional methods

Tests on more complicated spectra:

- *All artifacts + diffuse & noisy backgrounds*

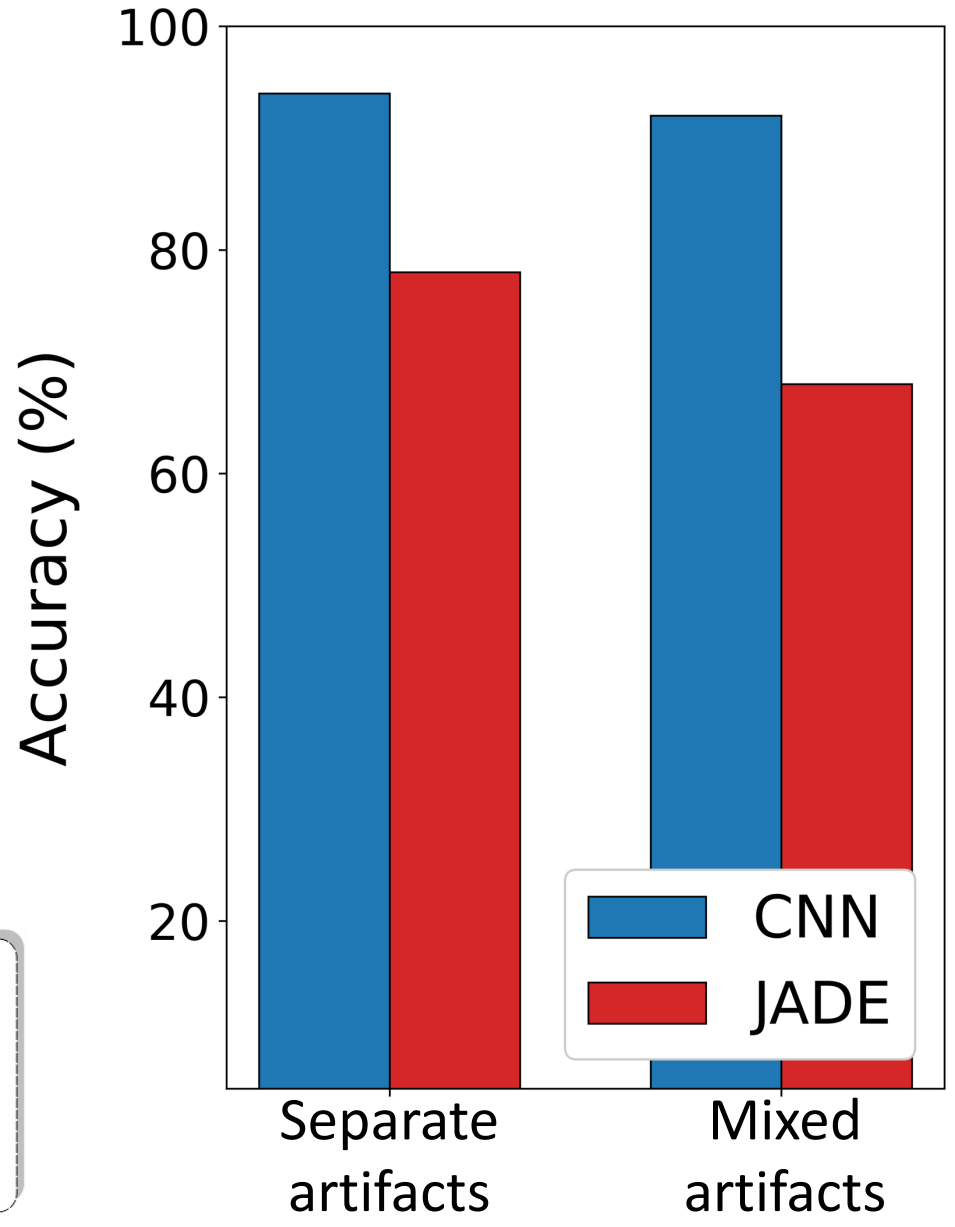
Results from the CNN vs. JADE:

- **92% accuracy with the CNN**
- **68% accuracy with JADE**

Improvement of 24%

Test set: 4,200 spectra

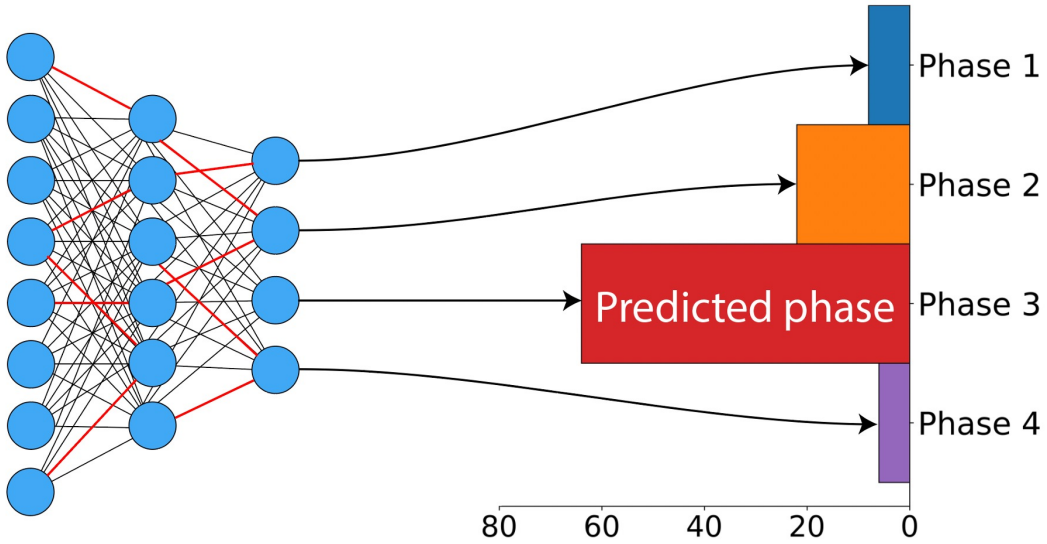
- ✓ Single-phase
- ✓ Stoichiometric
- ✓ Separate artifacts
- ✓ Peak perturbations
- ✓ Background effects
- ✓ Mixed artifacts

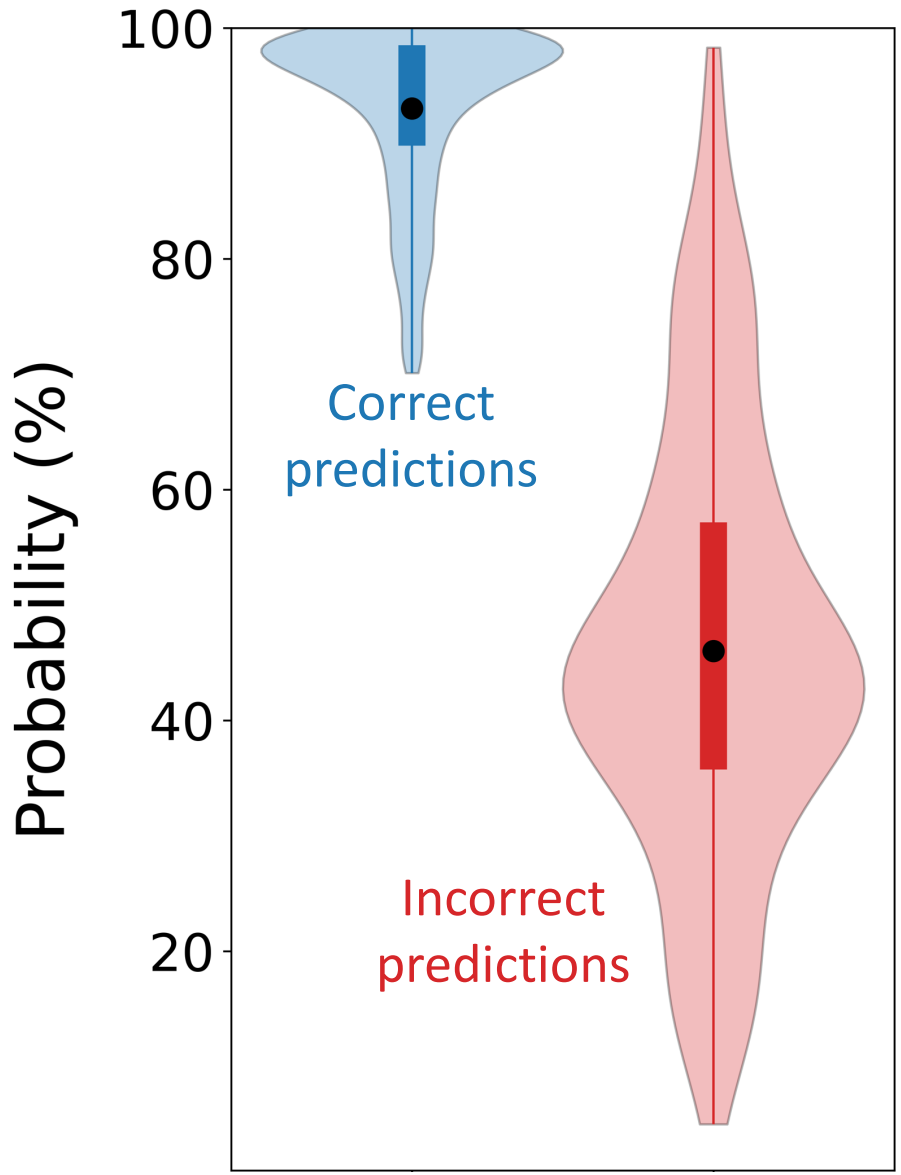


Test results: probabilities provide insight into accuracy

Probabilities provide a measure of confidence:

- Correct/incorrect predictions are clearly distinguished by their probabilities
- High probability → prediction is reliable
- Low probability → use caution





Limitations of current models

1) Physics-informed data augmentation

2) Accounting for non-stoichiometry

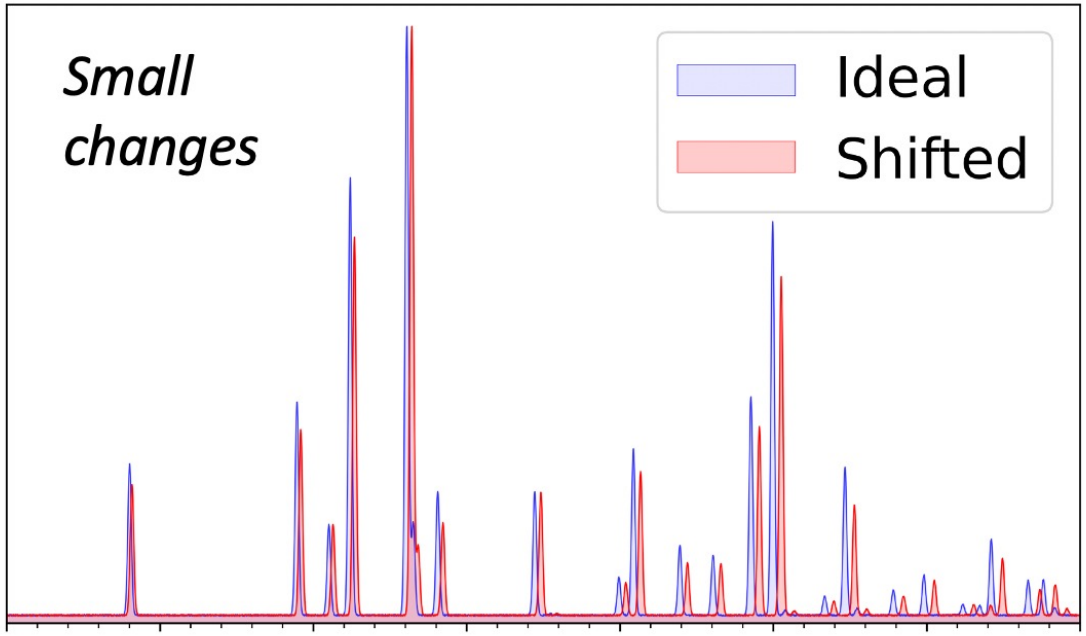
3) A probabilistic treatment of multi-phase mixtures

Non-stoichiometry: beyond perturbative treatment

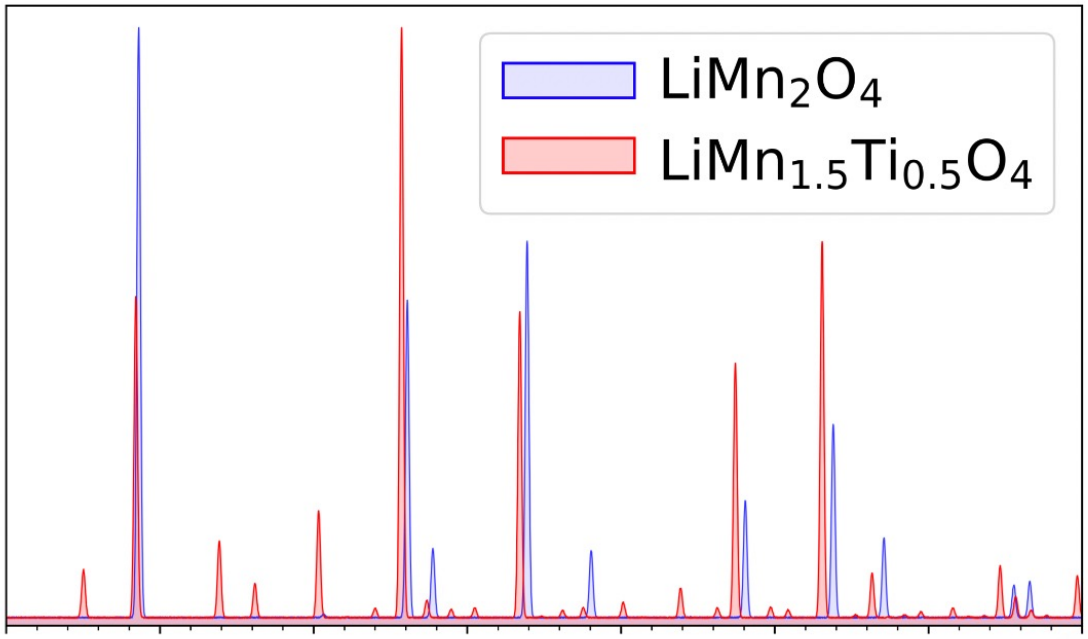
Data augmentation accounts for minor perturbations from idealized reference patterns

However, larger changes occur when composition deviates from expected stoichiometry

Perturbative treatment of artifacts



Beyond perturbations

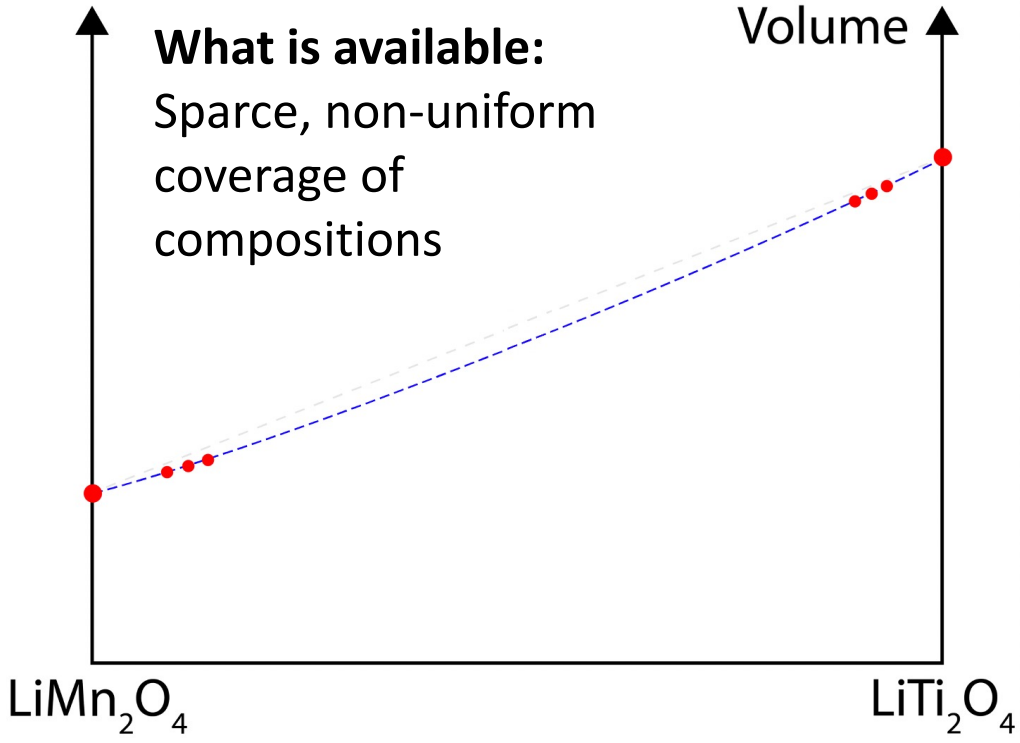


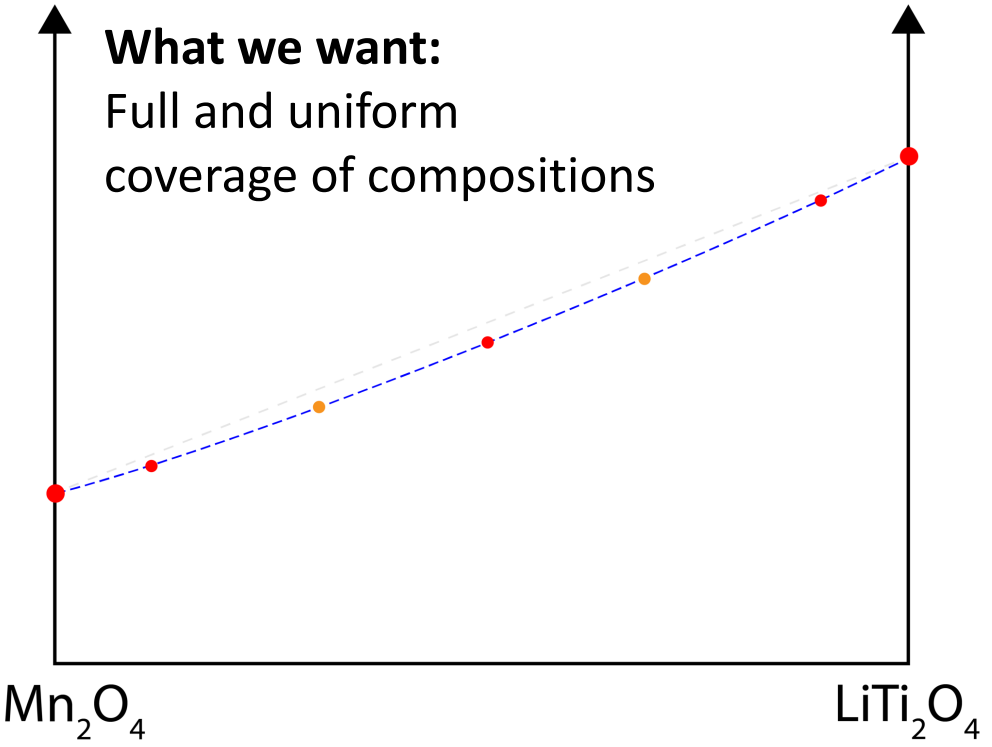
Experimentally reported solid solutions are insufficient

Can we just use whatever solid solutions are available on the ICSD?

- No: the ICSD covers narrow regions of chemical space while leaving others sparse – bias toward highly-studied materials

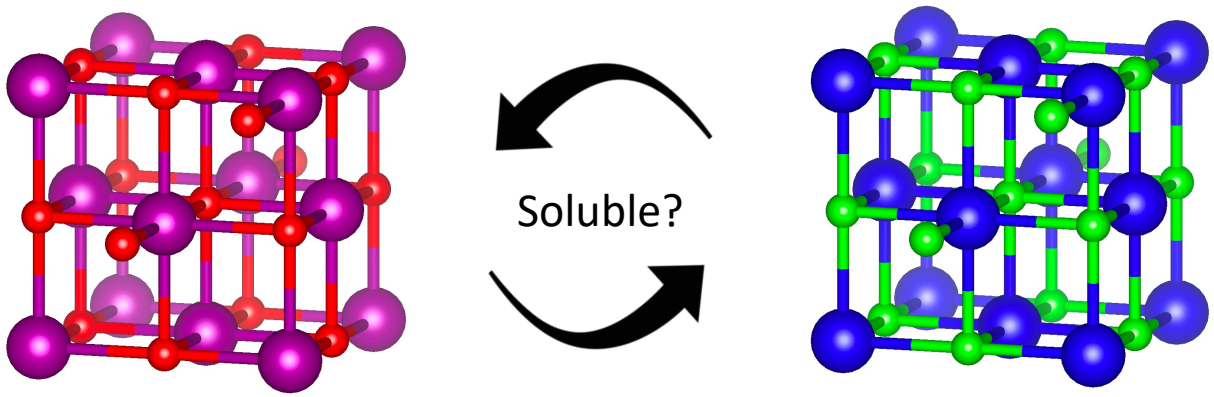
ICSD
FIZ Karlsruhe





Creating hypothetical solid solutions

1) Enumerate pairs of stoichiometric phases
2) For each pair, check **whether the structures are isomorphic**
3) Check whether equivalent sites contain **ions comparable in size** ($\leq 15\%$)



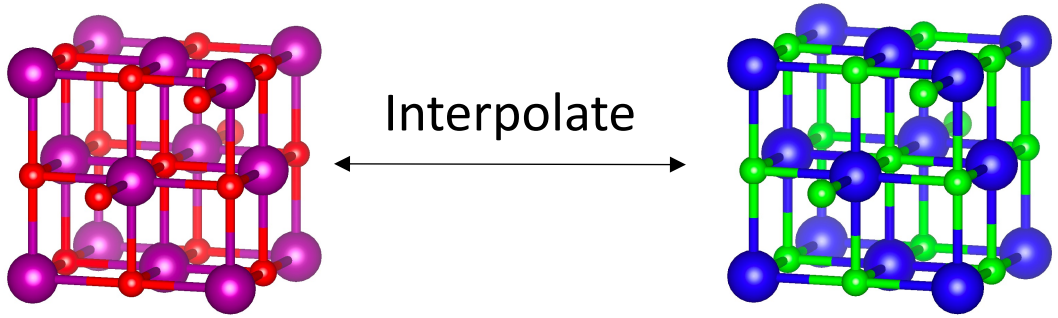
4) If both criteria are satisfied, assume solubility is possible
5) Interpolate a grid of hypothetical solid solutions by assuming **Vegard's law** holds for:
   - Lattice parameters
   - Atomic positions
   - Site occupancies

e.g., Mn – $Mn_{0.75}Ti_{0.25}$ – $Mn_{0.5}Ti_{0.5}$ – $Mn_{0.25}Ti_{0.75}$ – Ti

Testing on non-stoichiometric materials: Li-Mn-Ti-O-F

Training set:

- Hypothetical solid solutions



155 new reference phases in the Li-Mn-Ti-O-F space

Test set:

- Real non-stoichiometric phases from the ICSD

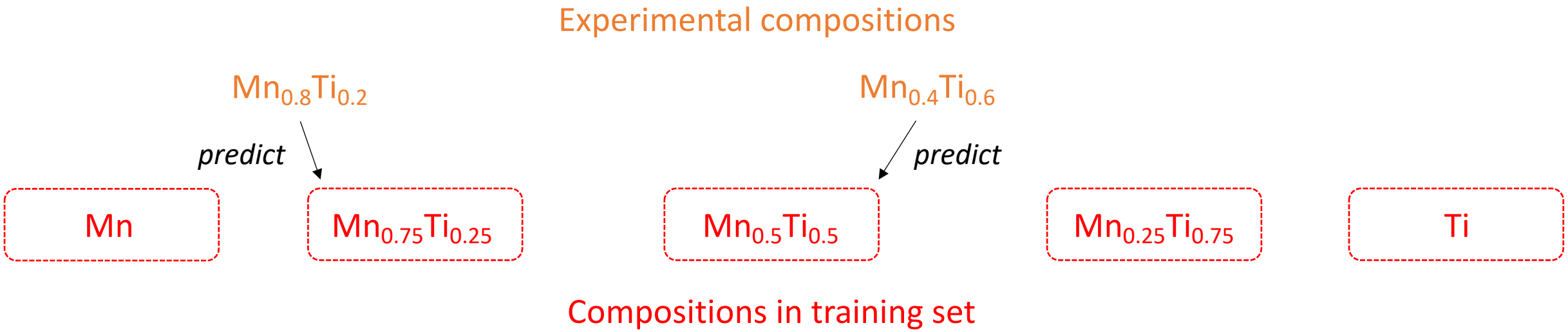
ICSD
FIZ Karlsruhe

20 experimental structures with unique compositions

Testing on non-stoichiometric materials

Classifications:

- **Classifications are discrete** – can only predict what we've trained on



Quantifying performance based on two metrics:

- **Structure:** is the predicted structure isomorphic to the true structure?
- **Composition:** mole fraction error between true and predicted compositions

Testing on non-stoichiometric materials: structure

Before including hypothetical solid solutions:

- **11/20 (55%)** of structures correctly identified

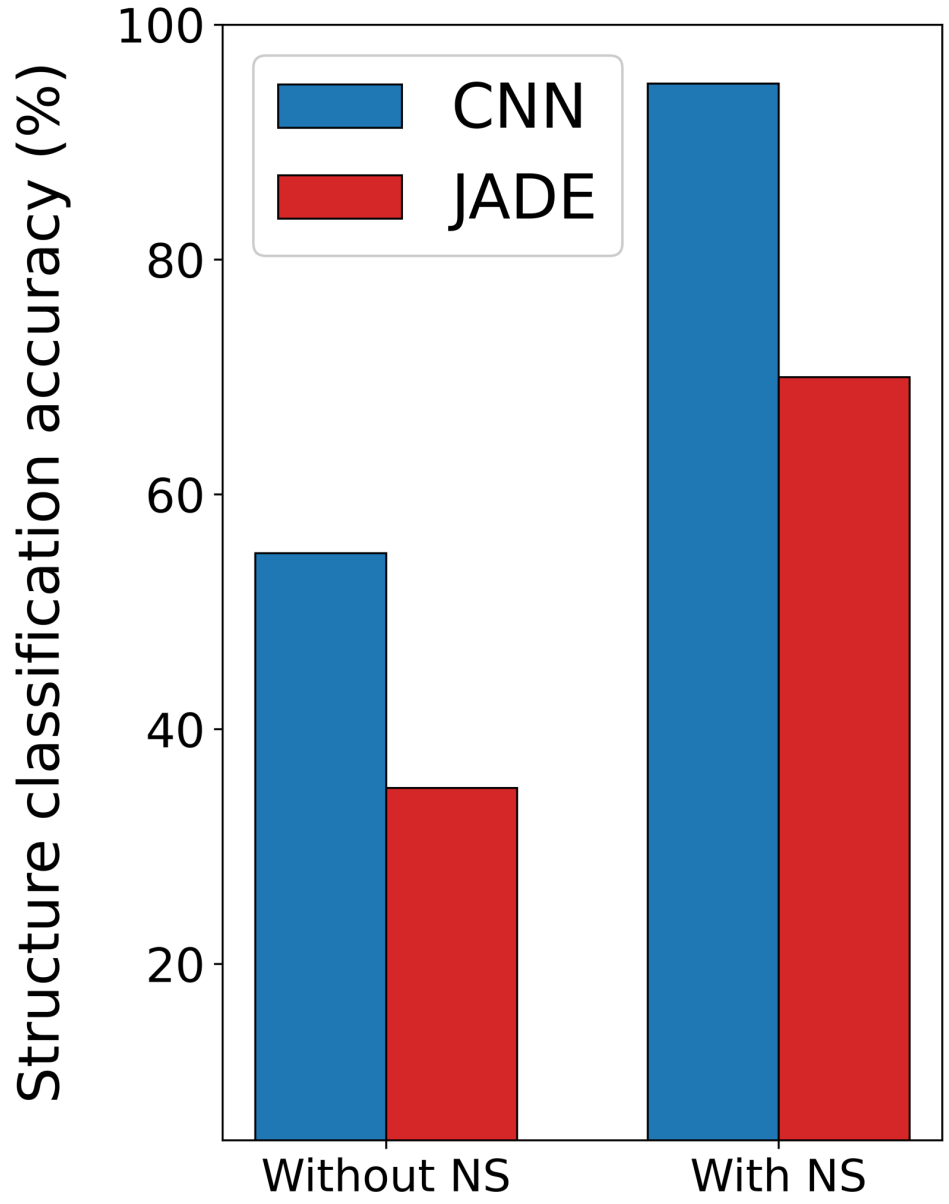
After introducing hypothetical solid solutions:

- **19/20 (95%)** of structures correctly identified

Including non-stoichiometry → 40% improvement

Training set:

Without NS: 16,800 spectra
✗ Non-stoichiometric (hypothetical)
With NS: 13,800 spectra
✓ Non-stoichiometric (hypothetical)



Testing on non-stoichiometric materials: compositions

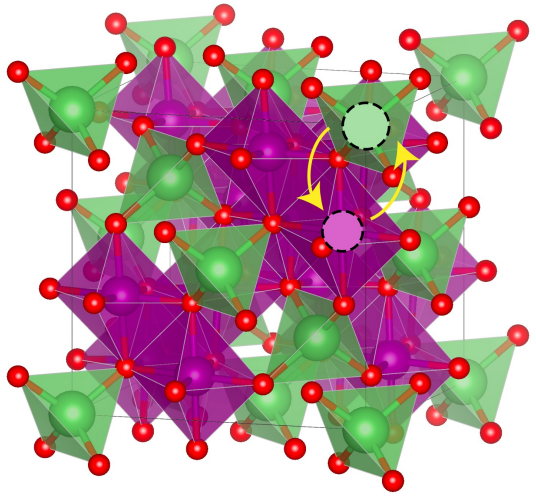
Composition is more difficult to predict:

- Lattice parameters often follow Vegard's
  → **structure classification is reliable**

  Peak positions

- The basis may not follow Vegard's law; some ions can swap sites in the structure
  → **limited accuracy for composition**

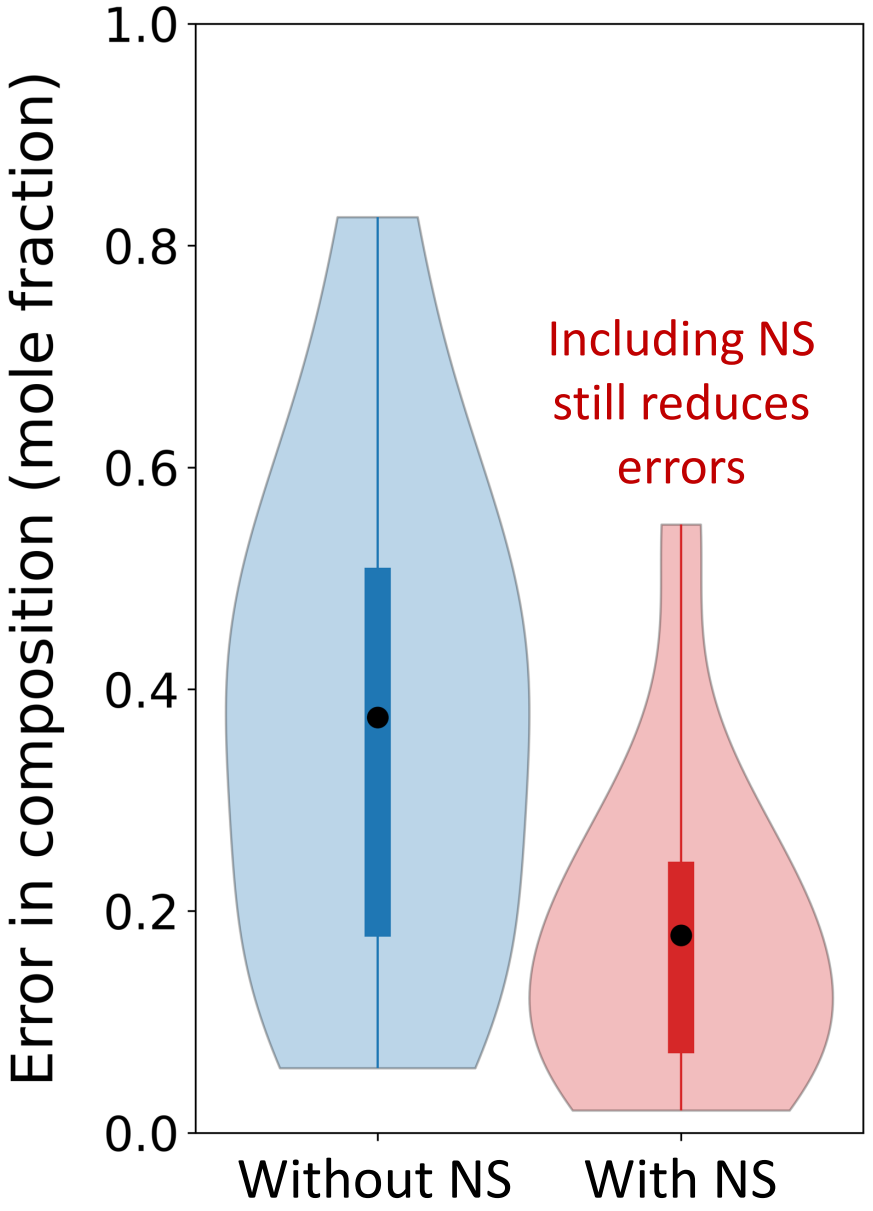
  Peak intensities



Example:

Li/TM ion swap in spinel $LiMn_{1-x}Ti_xO_4$

This algorithm is **not for refinement**



Limitations of current models

1) Physics-informed data augmentation

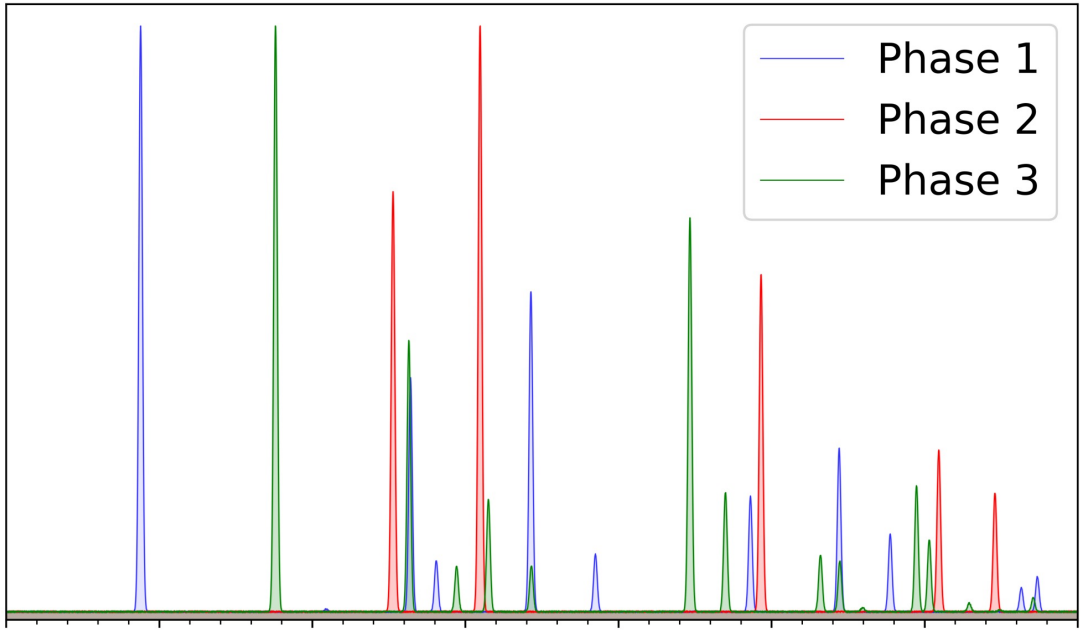
2) Accounting for non-stoichiometry

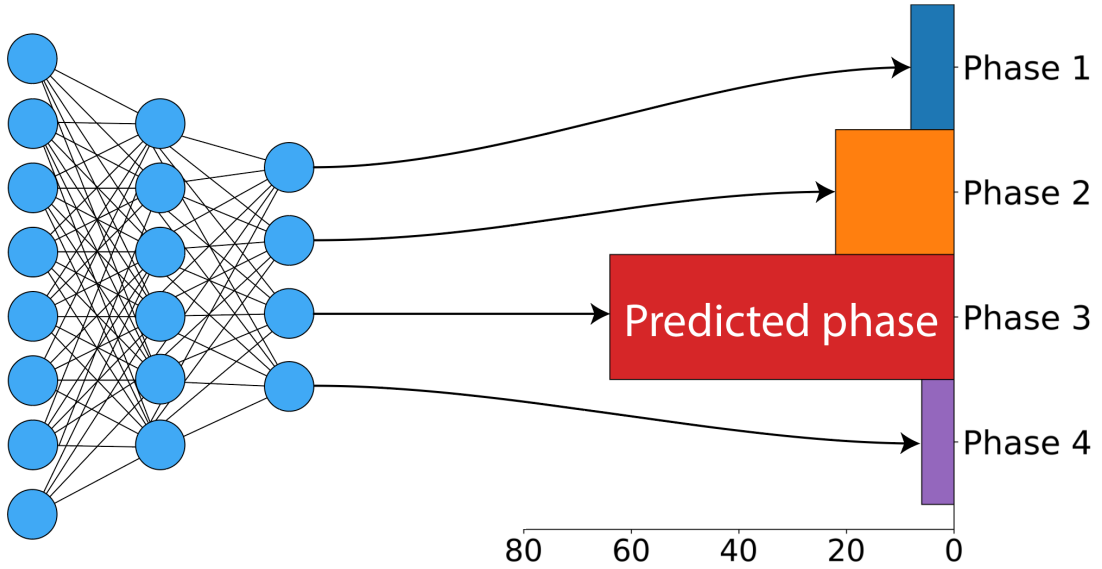
3) A probabilistic treatment of multi-phase mixtures

Possible methods for multi-phase identification

Approach #1 [a]:

- Choose phases with high probabilities
- **Problem: model may confuse similar reference phases**





Approach #2 [b]:

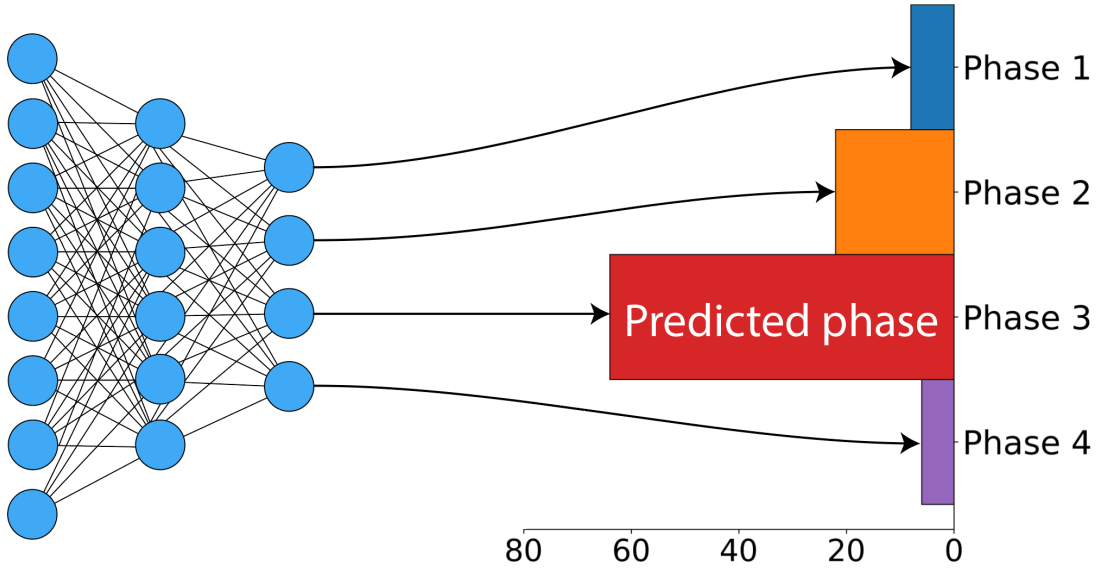
- Train on simulated multi-phase spectra
- **Problem: linear combinations of spectra → combinatorial data explosion**

a) Lee *et al.*, Nat. Commun. (2020).
b) Maffettone *et al.*, Nat. Comp. Sci. (2021).

Possible methods for multi-phase identification

Approach #1:

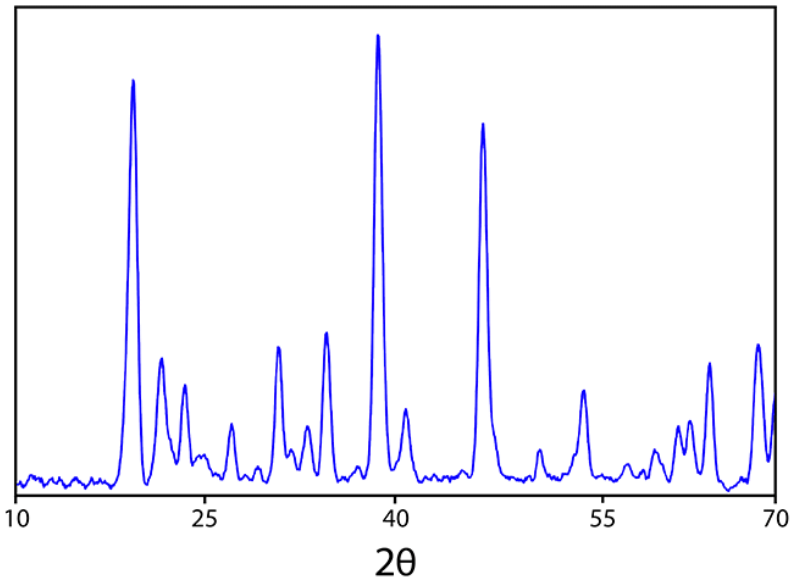
- Choose phases with high probabilities
- **Problem: model may confuse similar reference phases**

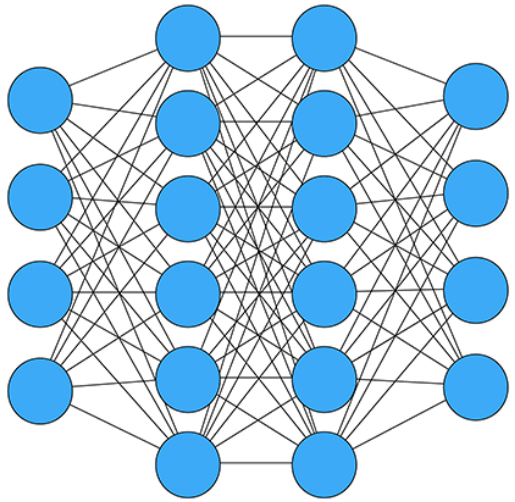


Our solution:

Use probability to guide an iterative approach of phase identification and profile subtraction

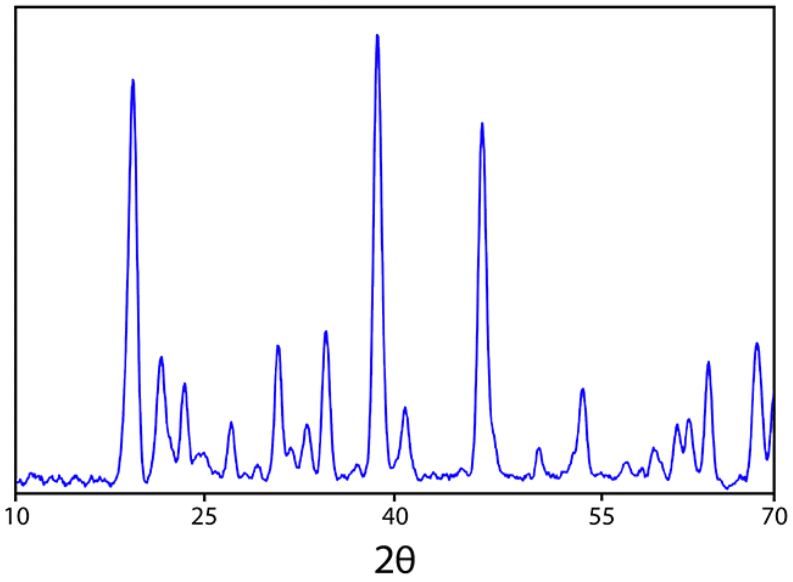
Our approach: phase ID + profile subtraction

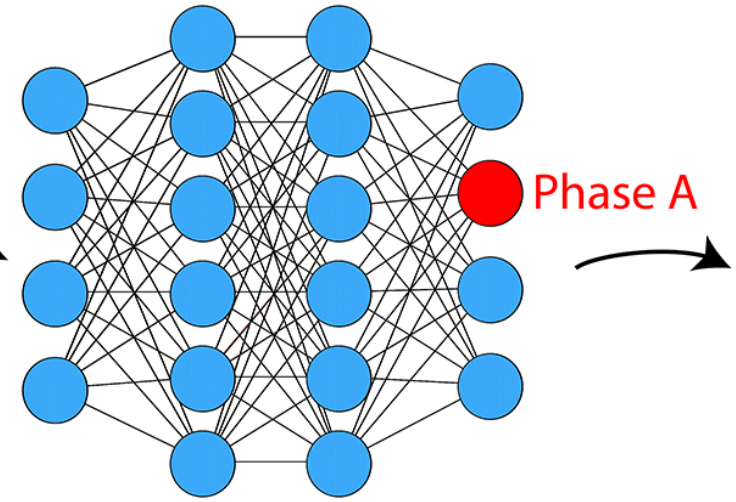


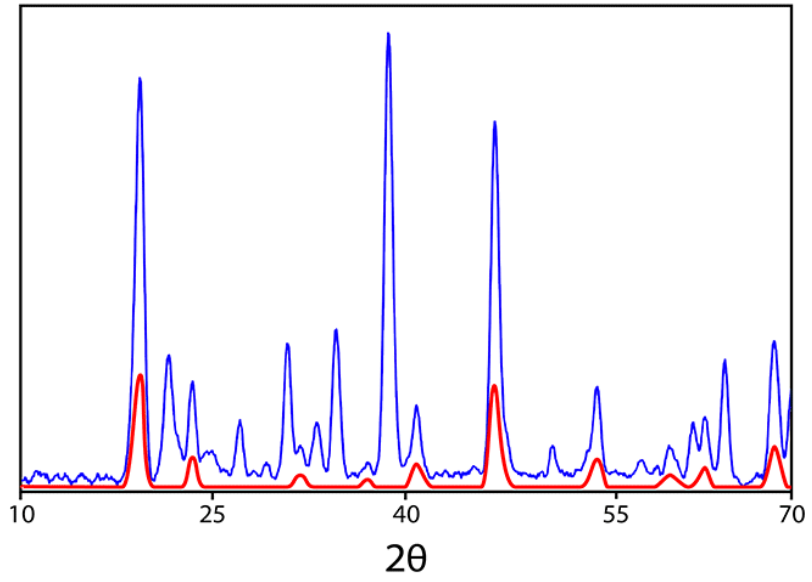


Plug measured spectrum into CNN to identify the first phase

Our approach: phase ID + profile subtraction

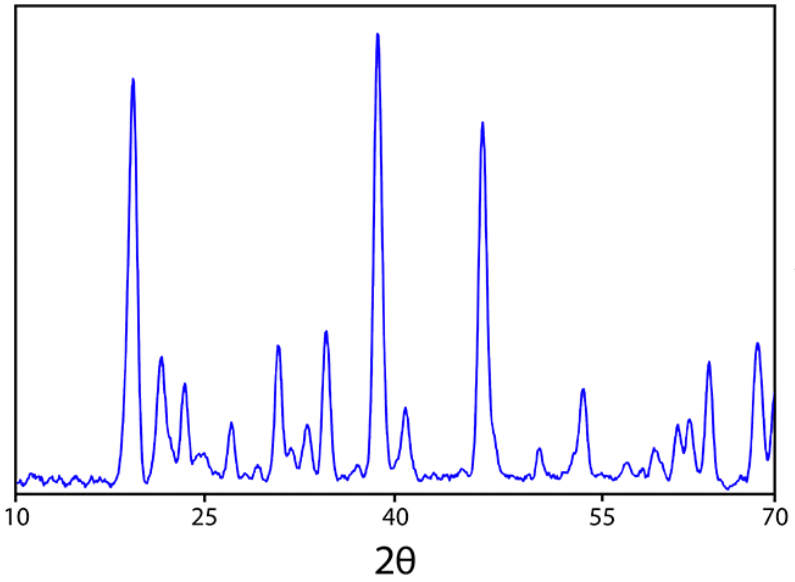


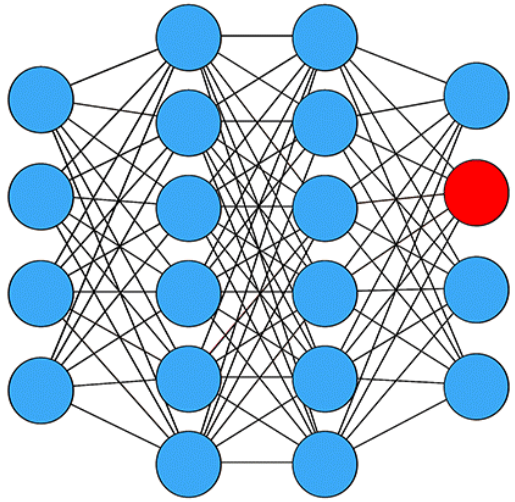


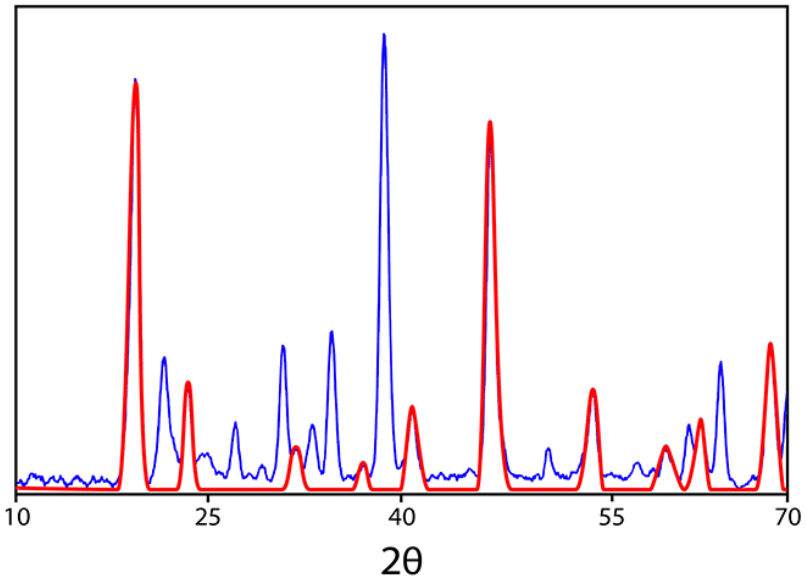


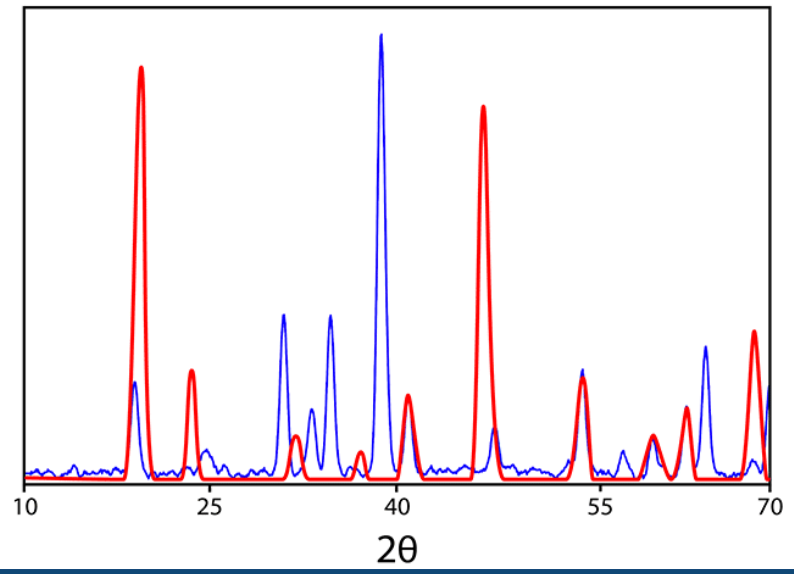
Fit the identified peaks to the measured spectrum

Our approach: phase ID + profile subtraction



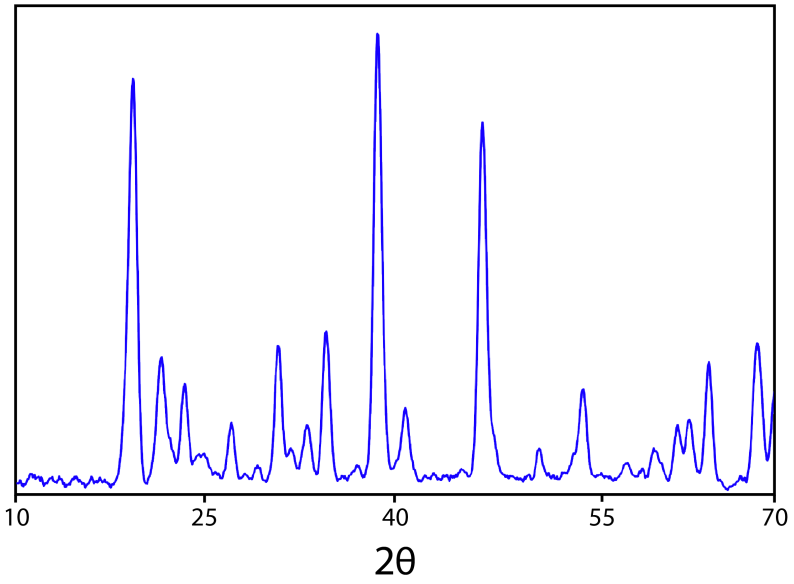


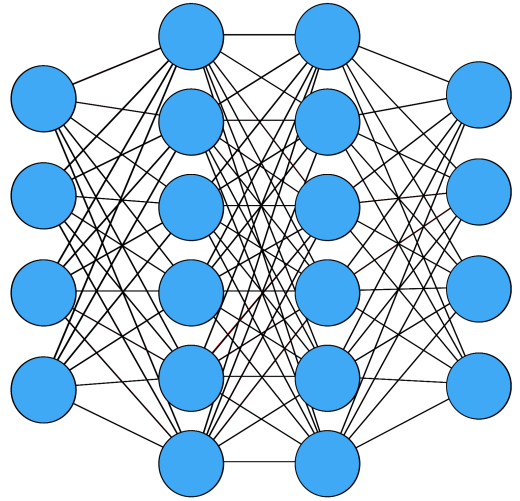


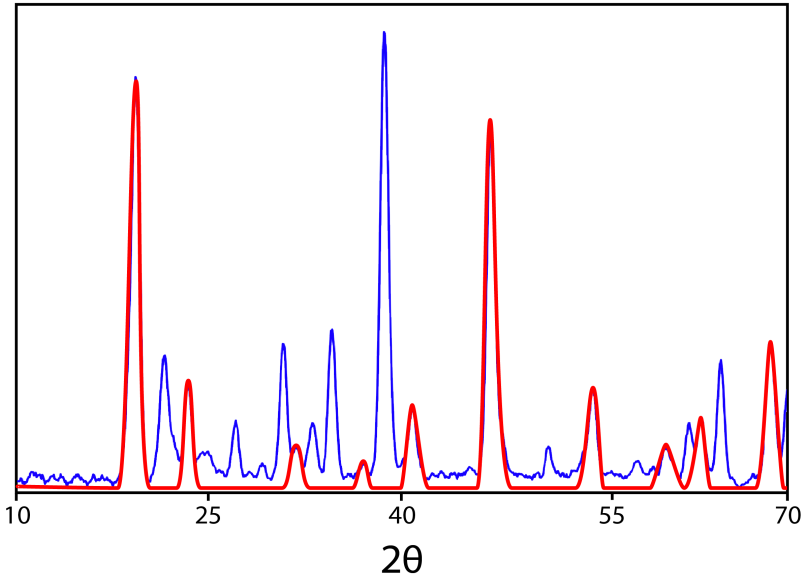


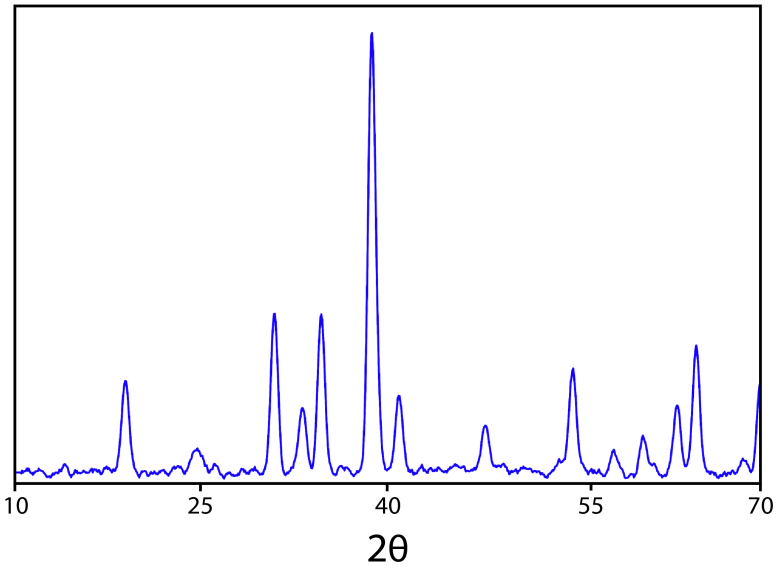
Subtract those peaks to obtain a new spectrum, which now excludes the known phase

Our approach: phase ID + profile subtraction





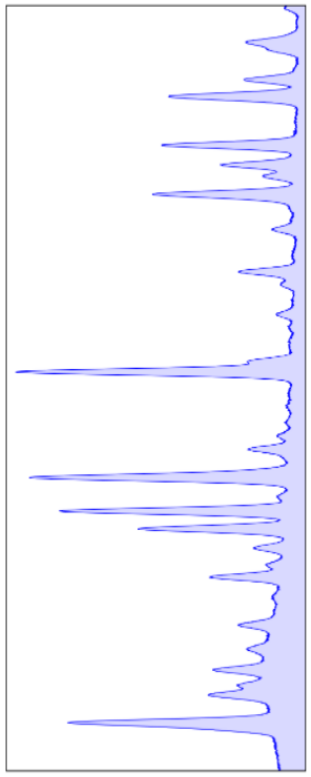




Repeat until all major peaks have been identified

How to use probability as a guiding metric?

Multi-phase spectrum



Li_2MnO_3
Confidence: 55%

MnO_2
Confidence: 52%

LiF
Confidence: 36%

48% ✗

Low *average* confidence

What if we just assume the most probable phase is correct at each step?

- First phase affects spectrum after profile subtraction
- Spectrum affects identification of remaining phases

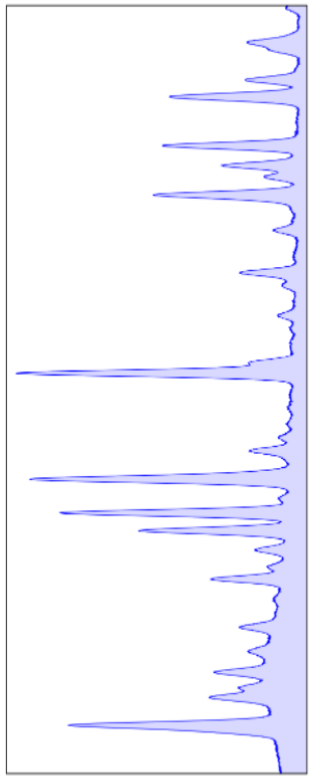
→ **First step is over-prioritized**

True phases: Li_2TiO_3 + Mn_3O_4 + Li_2O

A branching algorithm to maximize confidence

Solution: maximize the average confidence over *all* predicted phases

Multi-phase spectrum



Branching algorithm

- Li_2MnO_3 Confidence: 55%
  - MnO_2 Confidence: 52% → LiF Confidence: 36% — 48% ✗
  - TiO_2 Confidence: 48% → MnO Confidence: 24% — 42% ✗
- Li_2TiO_3 Confidence: 45%
  - Mn_3O_4 Confidence: 88% → Li_2O Confidence: 92% — 75% ✓
  - $LiMn_2O_4$ Confidence: 12% → MnO Confidence: 31% — 29% ✗

Higher *average* confidence

True phases: Li_2TiO_3 + Mn_3O_4 + Li_2O

Test results for multi-phase classification

Methods tested:

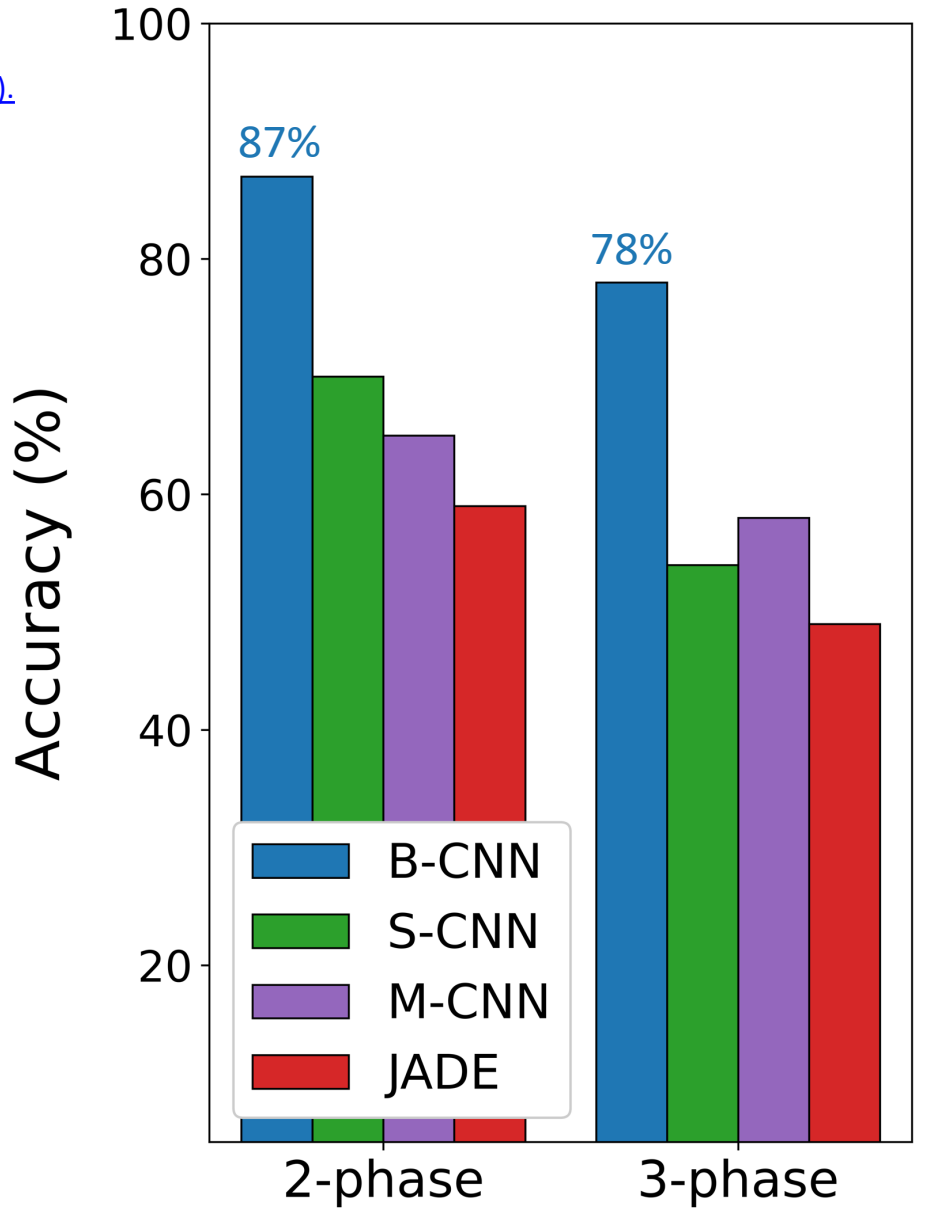
a) Lee *et al.*, Nat. Commun. (2020).
b) Maffettone *et al.*, Nat. Comp. Sci. (2021).

- B-CNN: branching algorithm
- S-CNN[a]: choose the 2 or 3 most probable phases
- M-CNN[b]: train on idealized multi-phase spectra
- JADE: traditional profile-matching

The B-CNN is shown to outperform all other methods

Training set (B-CNN):
16,800 spectra
- ✓ Single-phase
- ✗ Multi-phase
- ✓ Peak perturbations
- ✓ Stoichiometric
- ✗ Background effects

Test set: 3,000 spectra
- ✗ Single-phase
- ✓ Multi-phase
- ✓ Peak perturbations
- ✓ Stoichiometric
- ✓ Background effects



Validation on experimental spectra

Experimental procedure	Anticipated artifact	CNN	JADE
Single-phase			
Pristine samples	None	10/10	9/10
Kapton tape overlaid	Diffuse baseline	9/10	8/10
Rapid XRD scan	Noisy baseline	10/10	7/10
Thick samples	Shifts in 2θ	5/6	2/6
Ball milled	Broadening	5/5	4/5
Partially disordered	Intensity variation	5/6	4/6
Solid solutions	Non-stoichiometry	4/4	3/4
Multi-phase			
Two-phase mixtures	None	10/10	7/10
Three-phase mixtures	None	13/15	9/15
Overall accuracy:		**71/76 (93.4%)**	**53/76 (71.4%)**

Good match with simulated tests

Code availability

- Our approach can be **generalized to any arbitrary composition space**
- Code to automate data augmentation and model training:

https://github.com/njszym/XRD-AutoAnalyzer

- Only requires a set of reference phases

Preprint: https://arxiv.org/abs/2103.16664

Email: nathan_szymanski@berkeley.edu

Acknowledgements

Professor Gerbrand Ceder

Chris Bartel

Yan Zeng

Ceder Group

Howard Tu

Preprint: https://arxiv.org/abs/2103.16664
Github: https://github.com/njszym/XRD-AutoAnalyzer
Email: nathan_szymanski@berkeley.edu
Twitter: @NJSzymanski

NSF

JCESR | JOINT CENTER FOR ENERGY STORAGE RESEARCH