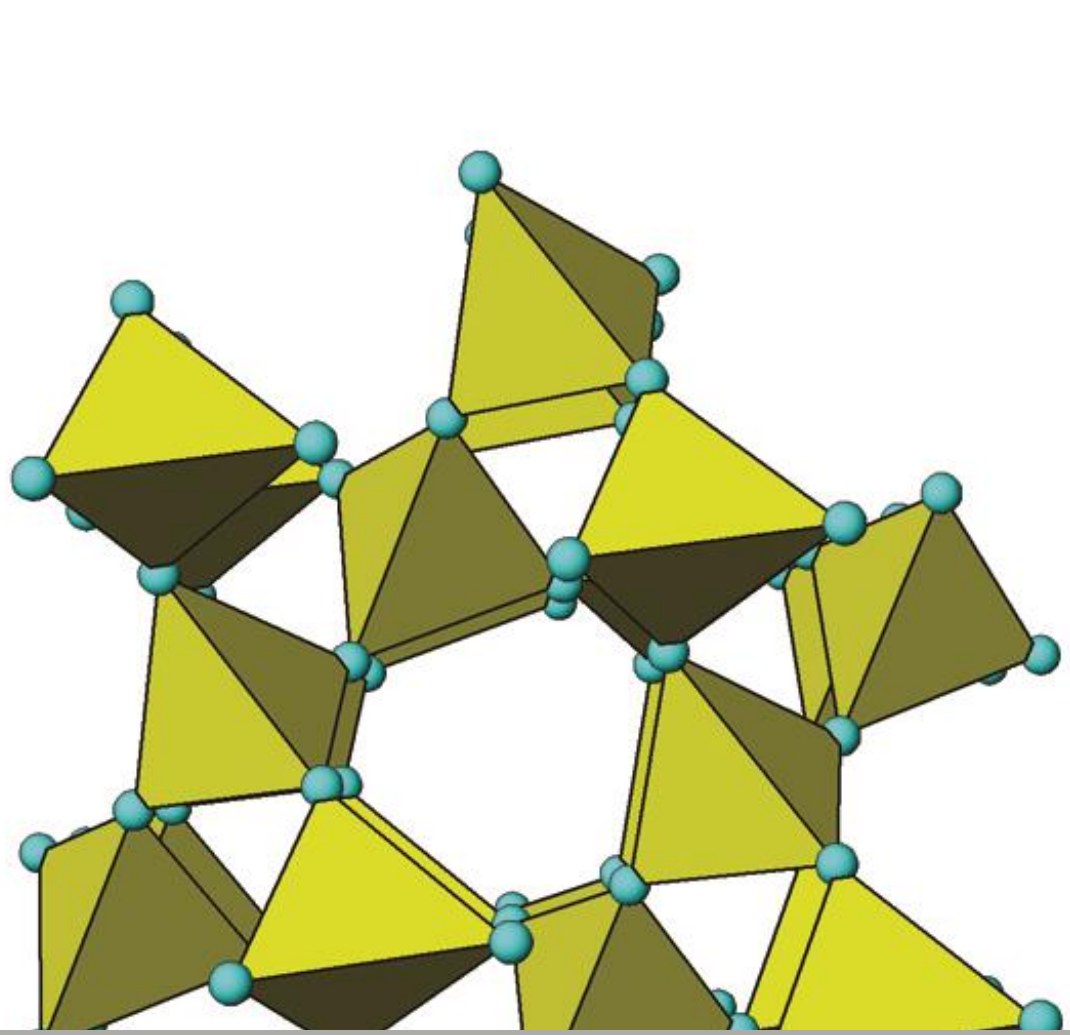


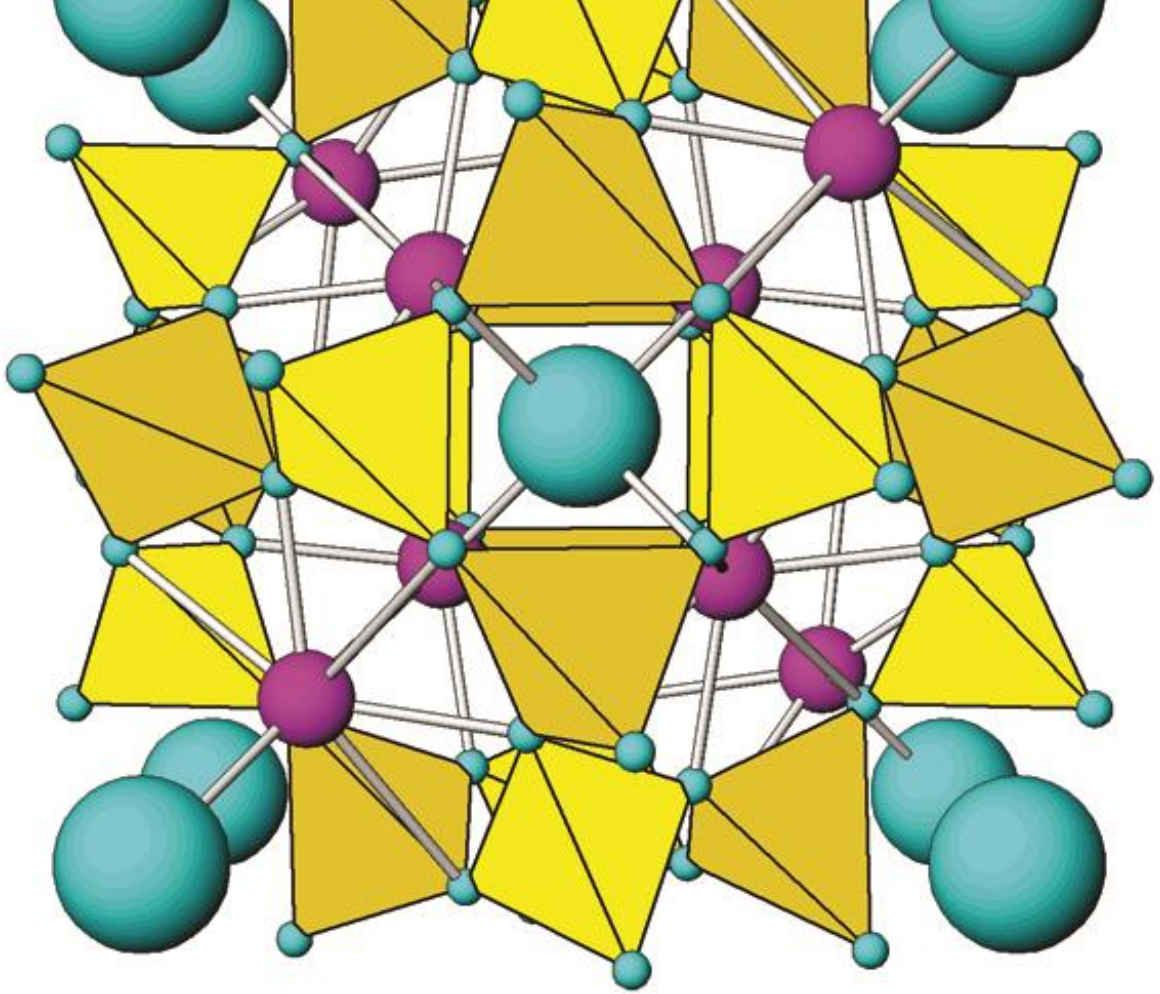


# Classifying Crystal Structure using Deep Learning

EDWARD FRIEDMAN | B.S COMPUTER SCIENCE, CARNEGIE MELLON UNIVERSITY

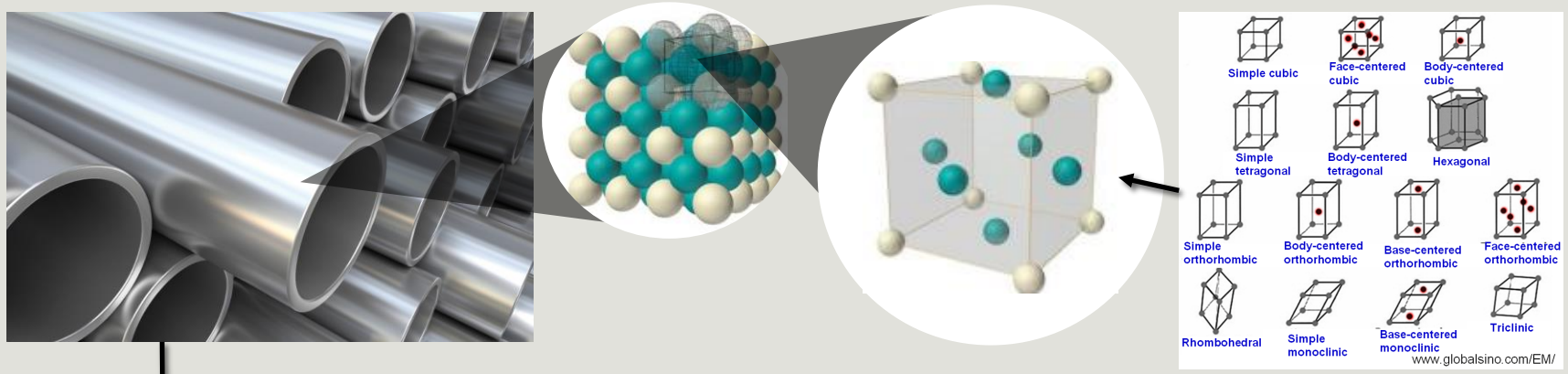
MENTOR: WILLIAM RATCLIFF





# Background, Methodology, & Past Work

# Crystallography & Powder Diffraction

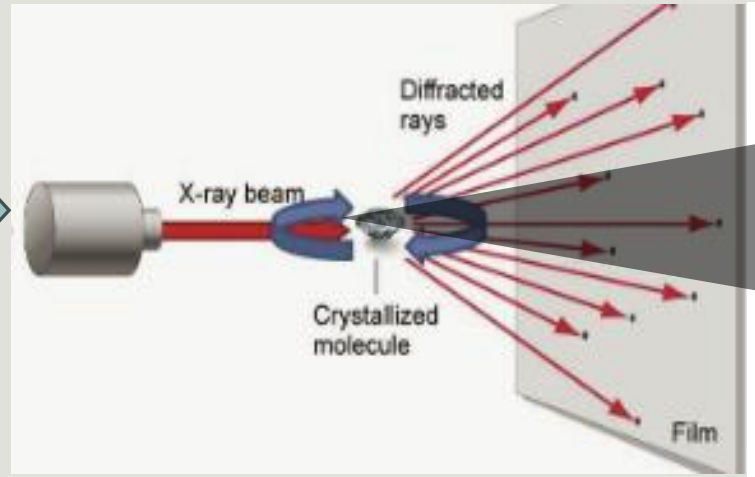




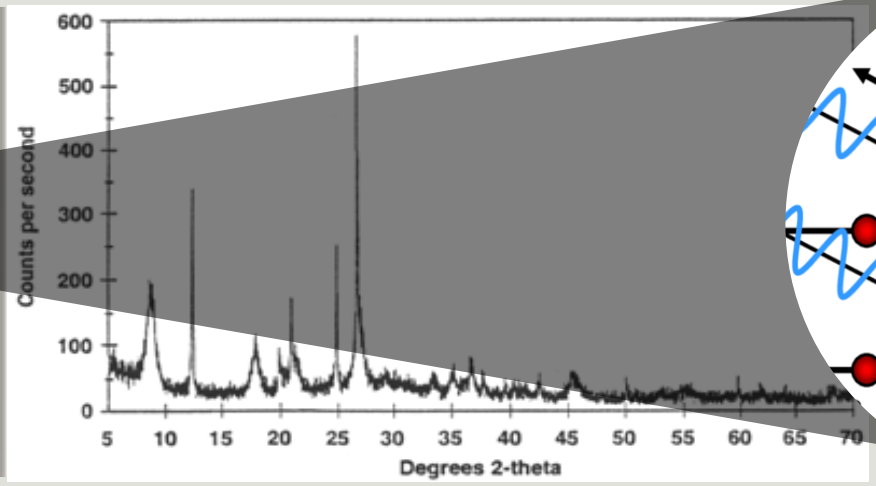

Space Group(230 Possible Groups)
How do we Classify???
Powder Diffraction!!!

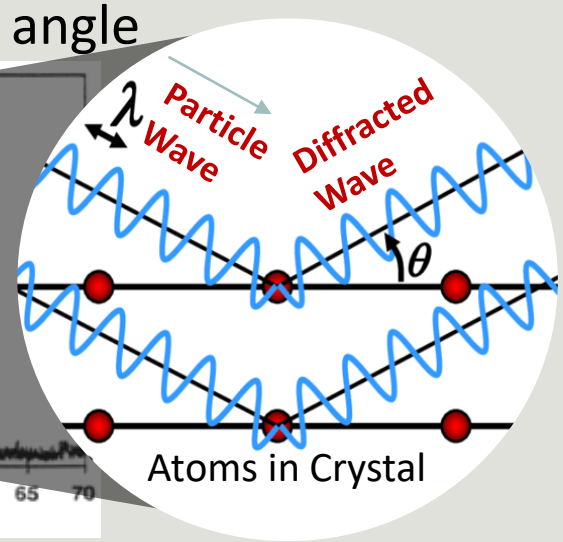
Unknown Powder of Crystals

**IN**: Neutron/X-Ray beams

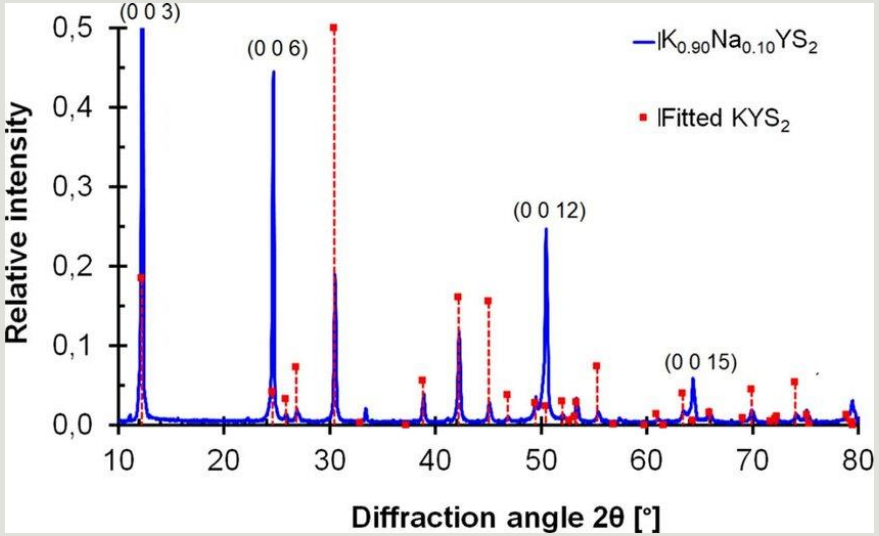



**Out**: Particle intensities @ each angle

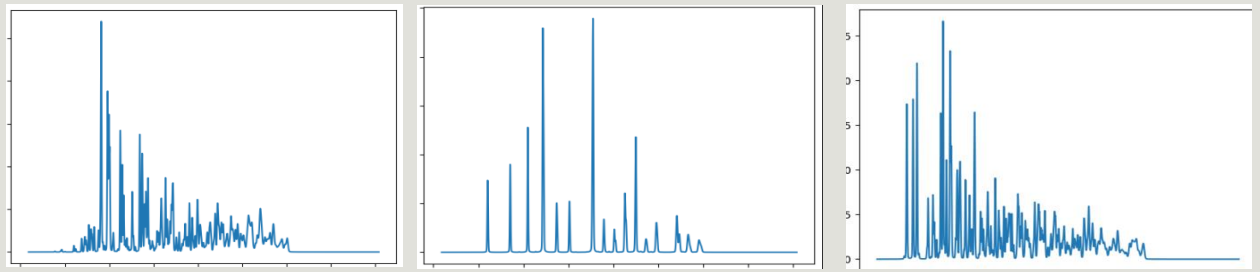


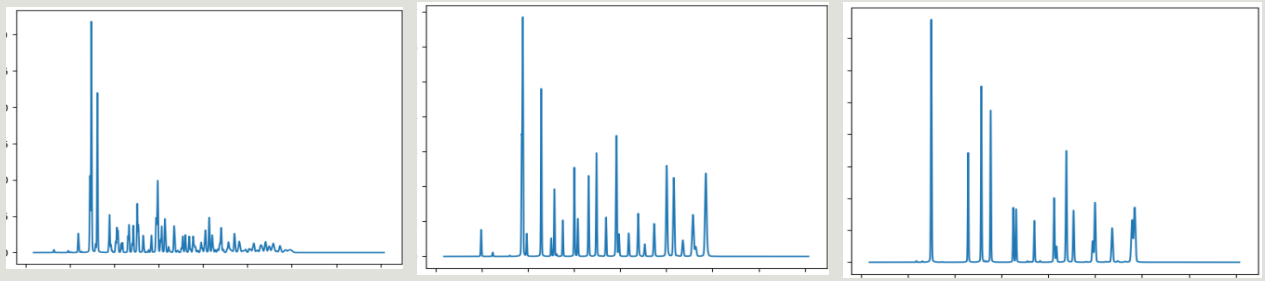
# Classification and Analysis: The Hard Part




3 examples of Crystals from the same SPG



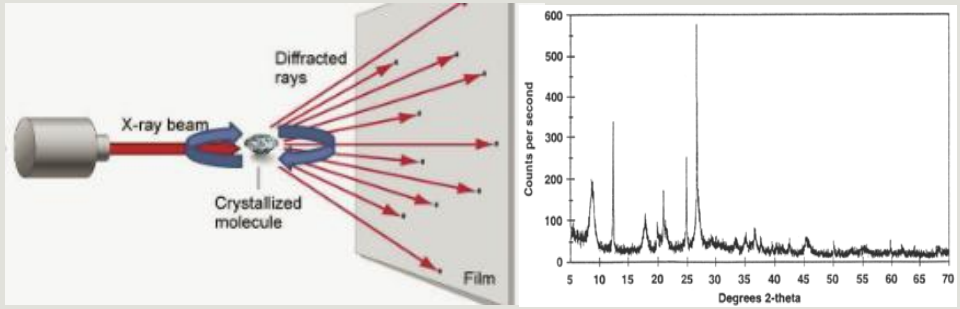
3 examples of Crystals from a diff SPG



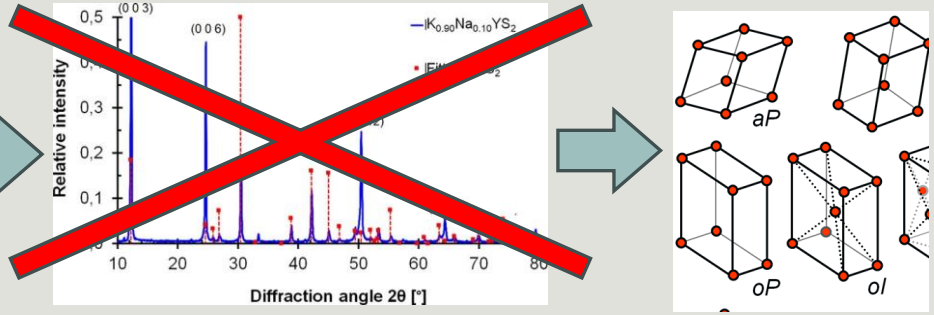
# The Goal

Unknown Crystal Structure

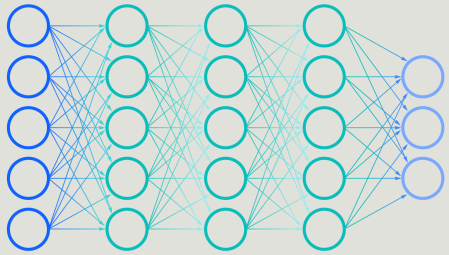

Diffraction Experiments




Indexing/Classification Space Group

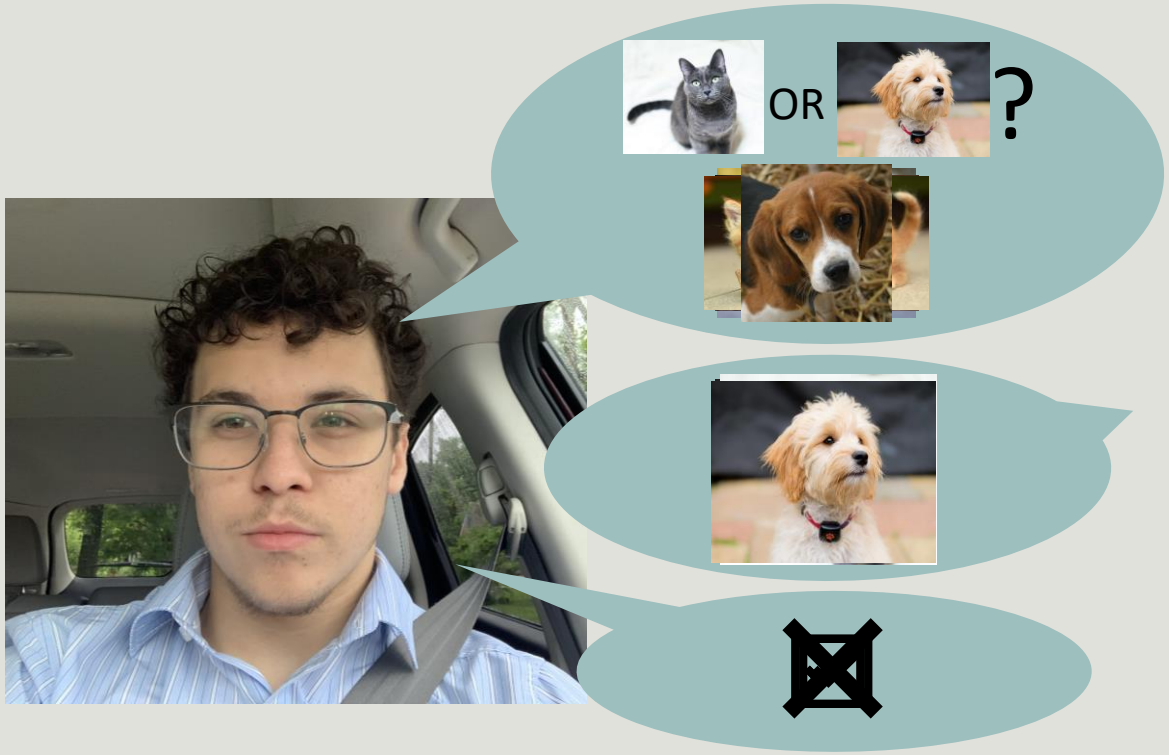


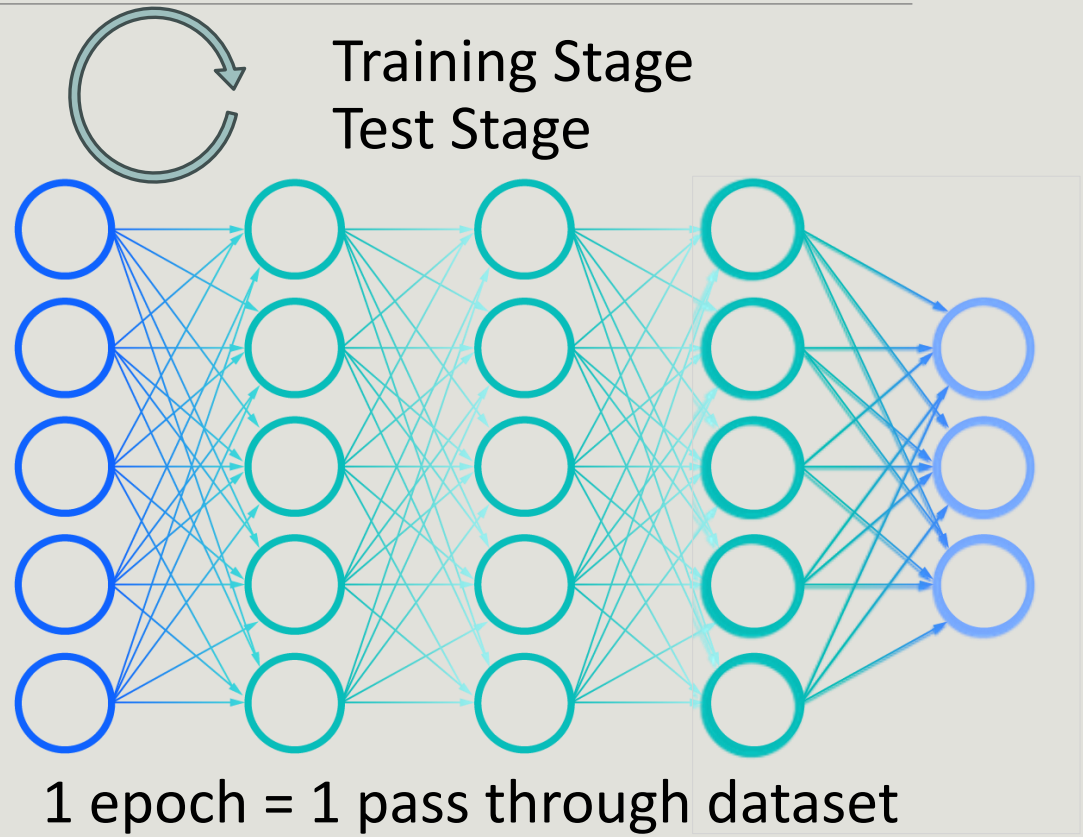

Why Deep Learning??!



Neural Network

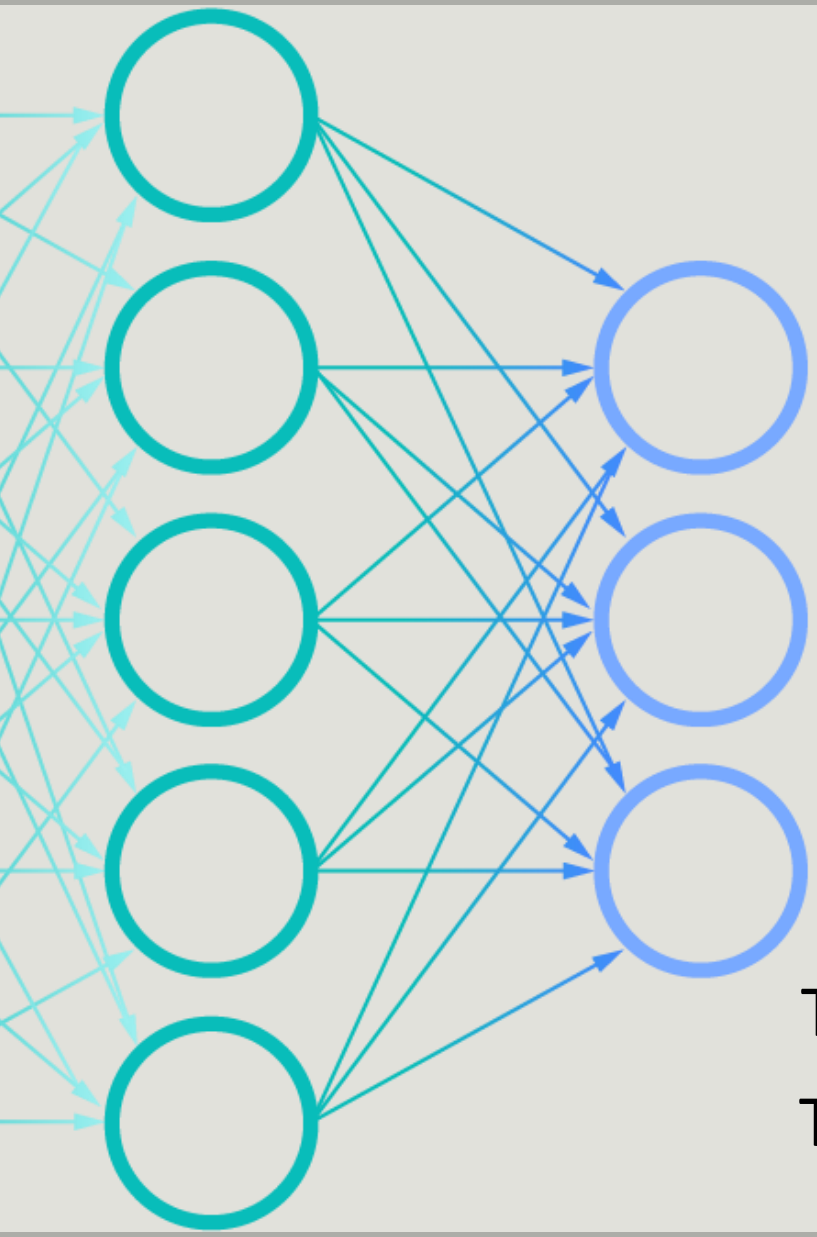
# Training Neural Networks 101 (Without the Math)




Training = Homework,

Testing = Exam



# What do we get?


%,

%] $\left(\sum \%_n\right) = 1$

[0.88,0.12]

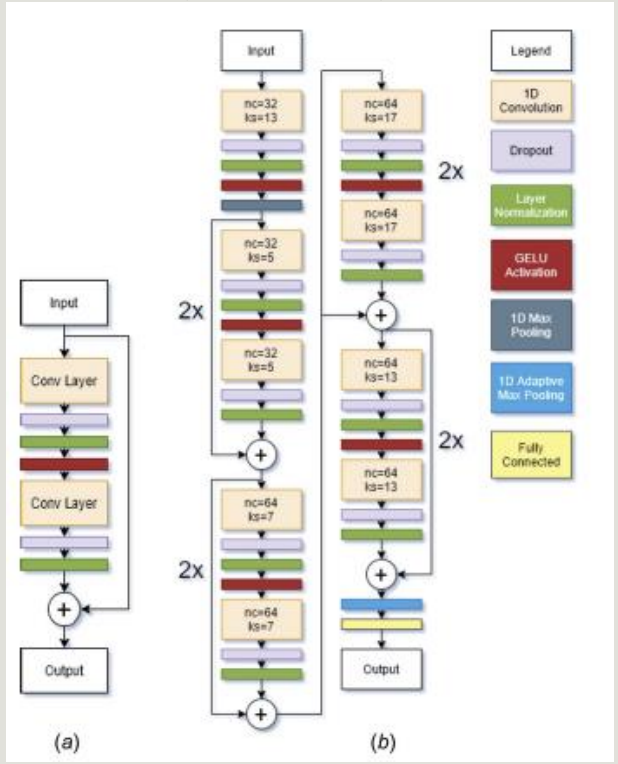
Top 1 Accuracy: If label is highest model prediction

Top 3 Accuracy: If Label among top 3 model predictions

# Methodology

## Architecture: ResNet CNN

(Lolla et al.,2022)




A semi-supervised deep-learning approach for automatic crystal structure classification

S Lolla, H Liang, AG Kusne, I Takeuchi... - Applied ..., 2022 - journals.iucr.org

## Synthetic Data

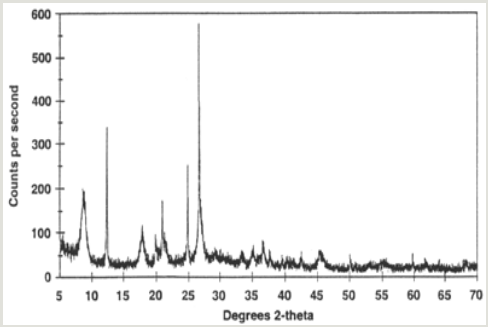
Crystal Structure Information Databases (ICSD, CSD, COD, etc)

ICSD

PyXtal

PyXtal Data Augmentation (Random structures)

FullProf Suite / PyCrysFML




Database of 1D diffraction patterns(used for training)

Insightful classification of crystal structures using deep learning [PD
A Ziletti, D Kumar, M Scheffler, LM Ghiringhelli Ful
Nature communications, 2018 · nature.com

**Abstract**
Computational methods that automatically extract knowledge from data are critical for enabling data-driven materials science. A reliable identification of lattice symmetry is a crucial first step for materials characterization and analytics. Current methods require a user-specified threshold, and are unable to detect average symmetries for defective structures. Here, we propose a machine learning-based approach to automatically classify structures by crystal symmetry. First, we represent crystals by calculating a diffraction

SHOW MORE

☆ Save 99 Cite Cited by 383 Related articles All 14 versions

Neural networks trained on synthetically generated crystals can extract structural information from ICSD powder X-ray diffractograms
H Schopmans, P Reiser, P Friederich
Digital Discovery, 2023 · pubs.rsc.org

Machine learning techniques have successfully been used to extract structural information such as the crystal space group from powder X-ray diffractograms. However, training directly on simulated diffractograms from databases such as the ICSD is challenging due to its limited size, class-inhomogeneity, and bias toward certain structure types. We propose an alternative approach of generating synthetic crystals with random coordinates by using the symmetry operations of each space group. Based on this approach, we

SHOW MORE

☆ Save 99 Cite Cited by 7 Related articles All 4 versions »

Crystal structure prediction via deep learning
K Ryan, J Lengyel, M Shatruk
Journal of the American Chemical Society, 2018 · ACS Publications

We demonstrate the application of deep neural networks as a machine-learning tool for the analysis of a large collection of crystallographic data contained in the crystal structure repositories. Using input data in the form of multiperspective atomic fingerprints, which describe coordination topology around unique crystallographic sites, we show that the neural-network model can be trained to effectively distinguish chemical elements based on the topology of their crystallographic environment. The model also identifies structurally

SHOW MORE

☆ Save 99 Cite Cited by 396 Related articles All 11 versions »

Automated classification of big X-ray diffraction data using deep learning models
JE Salgado, S Lerman, Z Du, C Xu... - npj Computational ..., 2023 - nature.com
In current in situ X-ray diffraction (XRD) techniques, data generation surpasses human analytical capabilities, potentially leading to the loss of insights. Automated techniques require human intervention, and lack the performance and adaptability required for material exploration. Given the critical need for high-throughput automated XRD pattern analysis, we present a generalized deep learning model to classify a diverse set of materials' crystal systems and space groups. In our approach, we generate training data with a holistic ...
☆ Save 99 Cite Cited by 25 Related articles All 7 versions »

A semi-supervised deep-learning approach for automatic crystal structure classification
S Lolla, H Liang, AG Kusne, I Takeuchi... - Applied ..., 2022 - journals.iucr.org
The structural solution problem can be a daunting and time-consuming task. Especially in

Towards end-to-end structure determination from x-ray diffraction data using deep learning
G Guo, J Goldfeder, L Lan, A Ray, AH Yang, B Chen, SJL Billinge, H Lipson

Crystal structure determi
machine learning
EA Riesel, T Mackey, H Nilforosh
Journal of the American Chemic

Powder X-ray diffraction (
However, complete struct
consuming and is often in
learning (ML) approaches to PXRD analysis predict only a subset of the total information that comprises a crystal structure. We developed a pioneering generative ML model designed to solve crystal structures from real-world experimental PXRD data. In addition

SHOW MORE

☆ Save 99 Cite Cited by 17 Related articles All 3 versions »

# This is not a "unique" problem.

l science of determining the structure of
rm, by analyzing their x-ray diffraction (XRD)
available as crystalline powder, powder
many fields. However, powder crystallography
n, and therefore the structural inference
typically involves a laborious process of iterative design, structural refinement, and domain

SHOW MORE

☆ Save 99 Cite Cited by 8 Related articles All 5 versions »

Powder X-ray diffraction pattern is all you need for machine-learning-based symmetry identification and property prediction
BD Lee, JW Lee, WB Park, J Park, MY Cho, S Pal Singh, M Pyo, KS Sohn
Advanced Intelligent Systems, 2022 · Wiley Online Library

Herein, data-driven symmetry identification, property prediction, and low-dimensional embedding from powder X-Ray diffraction (XRD) patterns of inorganic crystal structure database (ICSD) and materials project (MP) entries are reported. For this purpose, a fully convolutional neural network (FCN), transformer encoder (T-encoder), and variational autoencoder (VAE) are used. The results are compared to those obtained from a well-established crystal graph convolutional neural network (CGCNN). A task-specified small

SHOW MORE

☆ Save 99 Cite Cited by 46 Related articles All 4 versions »

Crystal structure assignment for unknown compounds from x-ray diffraction patterns with deep learning [PDF] acs.org
L Chen, B Wang, W Zhang, S Zheng, Z Chen, M Zhang, C Dong, F Pan, S Li Full View
Journal of the American Chemical Society, 2024 · ACS Publications

Determining the structures of previously unseen compounds from experimental characterizations is a crucial part of materials science. It requires a step of searching for the structure type that conforms to the lattice of the unknown compound, which enables the pattern matching process for characterization data, such as X-ray diffraction (XRD) patterns. However, this procedure typically places a high demand on domain expertise, thus creating an obstacle for computer-driven automation. Here, we address this

SHOW MORE

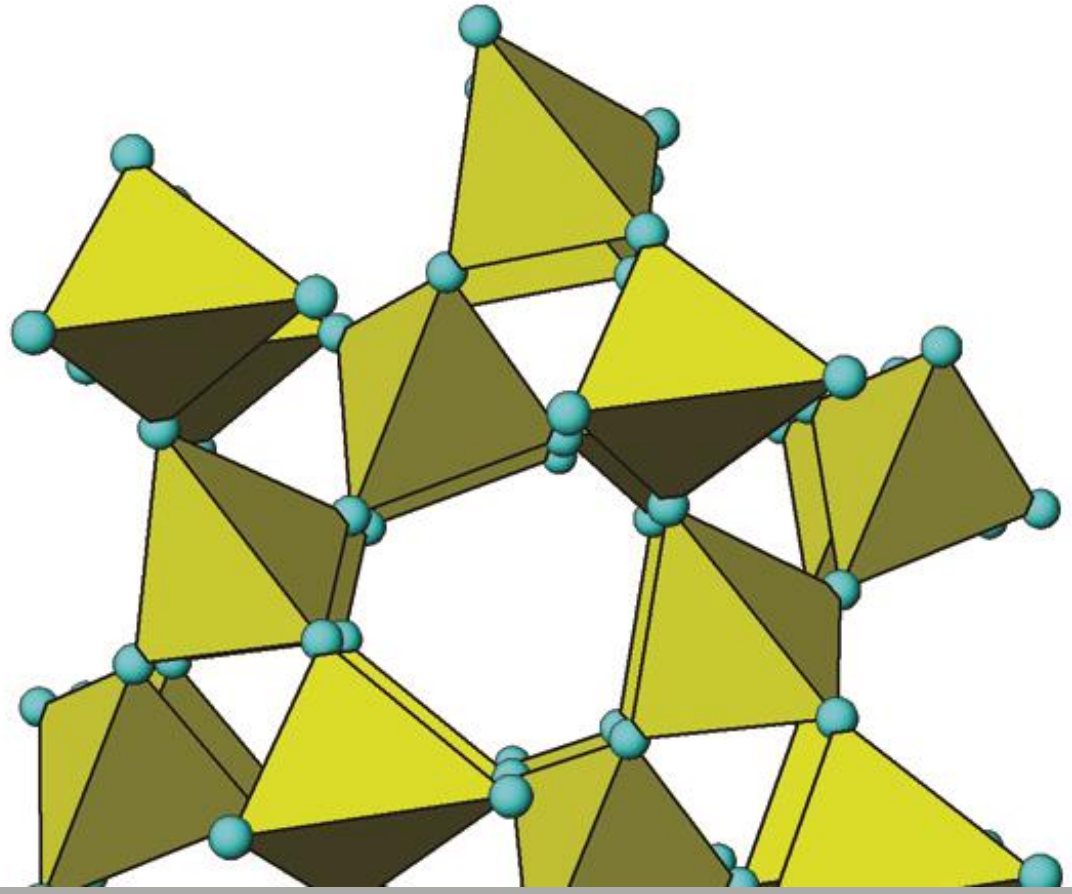
☆ Save 99 Cite Cited by 14 Related articles All 6 versions »

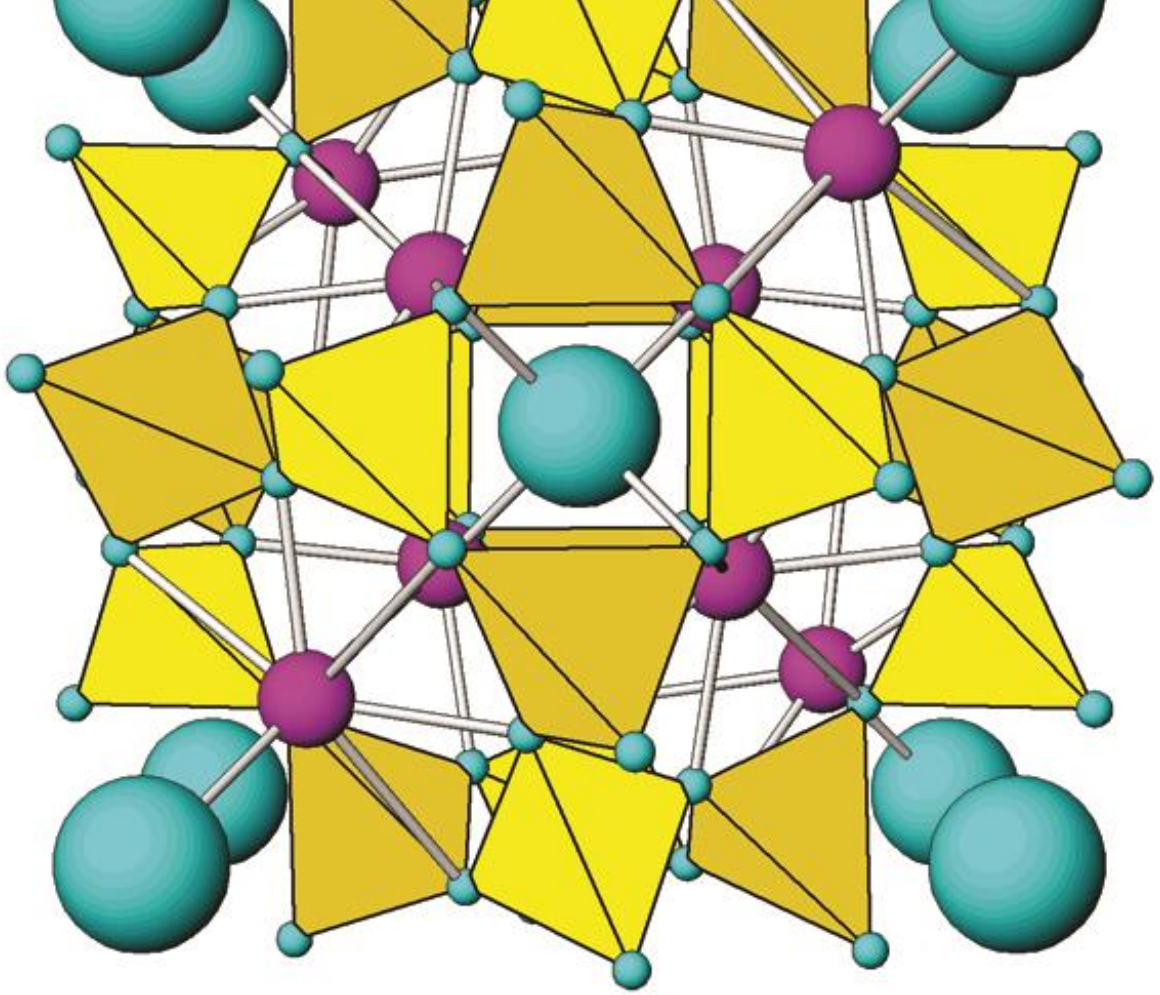
Classification of crystal structure using a convolutional neural network
WB Park, J Chung, J Jung, K Sohn, SP Singh, M Pyo, N Shin, KS Sohn
IUCrJ, 2017 · journals.iucr.org

A deep machine-learning technique based on a convolutional neural network (CNN) is introduced. It has been used for the classification of powder X-ray diffraction (XRD) patterns in terms of crystal system, extinction group and space group. About 150 000 powder XRD patterns were collected and used as input for the CNN with no handcrafted engineering involved, and thereby an appropriate CNN architecture was obtained that allowed determination of the crystal system, extinction group and space group. In sharp

SHOW MORE

☆ Save 99 Cite Cited by 262 Related articles All 12 versions »

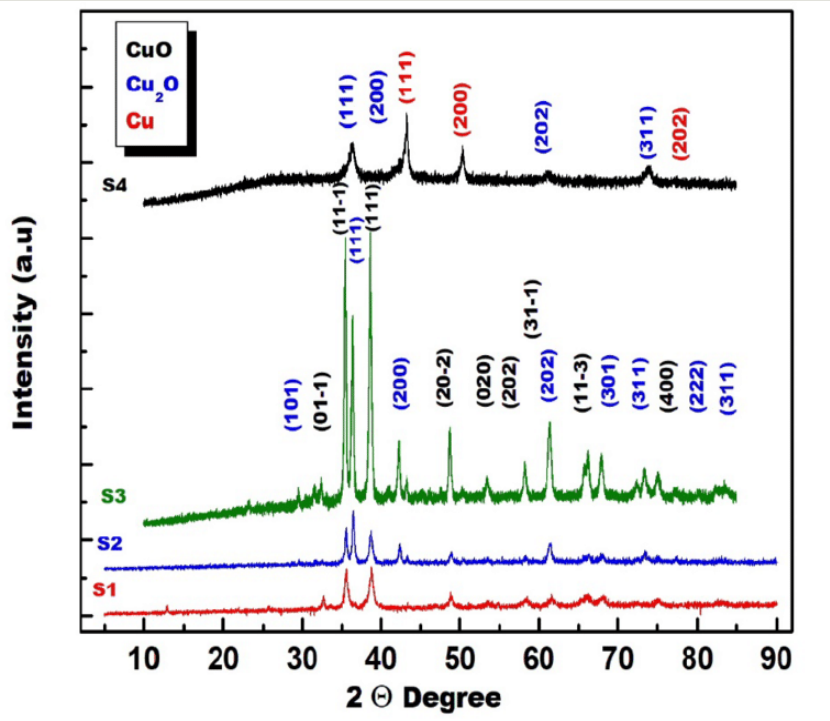




# Addressing Previous and Current Limitations

# Issue 1: Labeling in previous studies

Diffraction Pattern

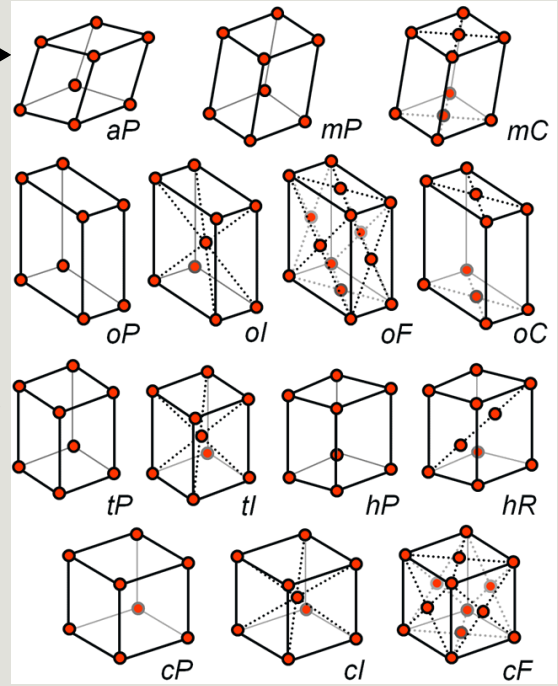



Pattern to Ext Group

Extinction Group, (99 Possible)

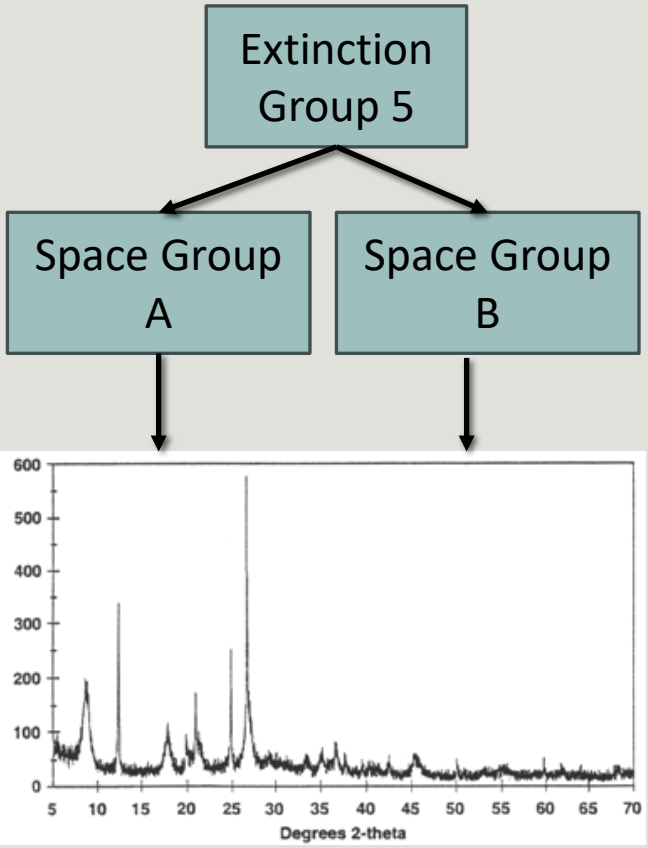
Ext Group to SG (w more experiments)

Space Groups(230 Possible)

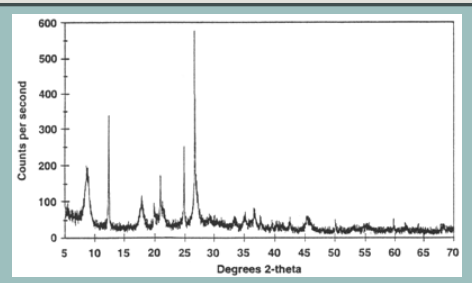



**Space Groups that are apart of the same extinction group have identical patterns**

# Example of why this is a problem




*Space Group A and B have identical diffraction patterns

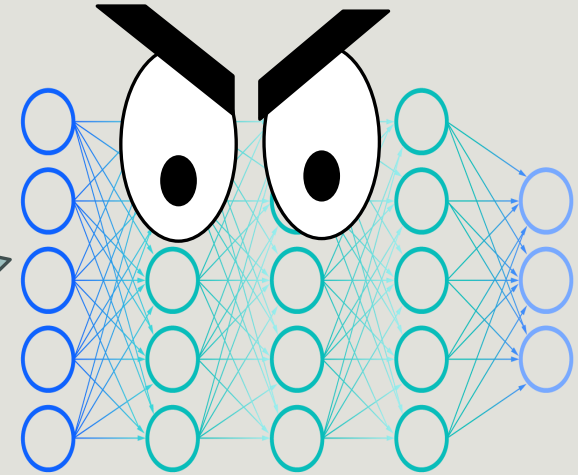



?

Space Group A!

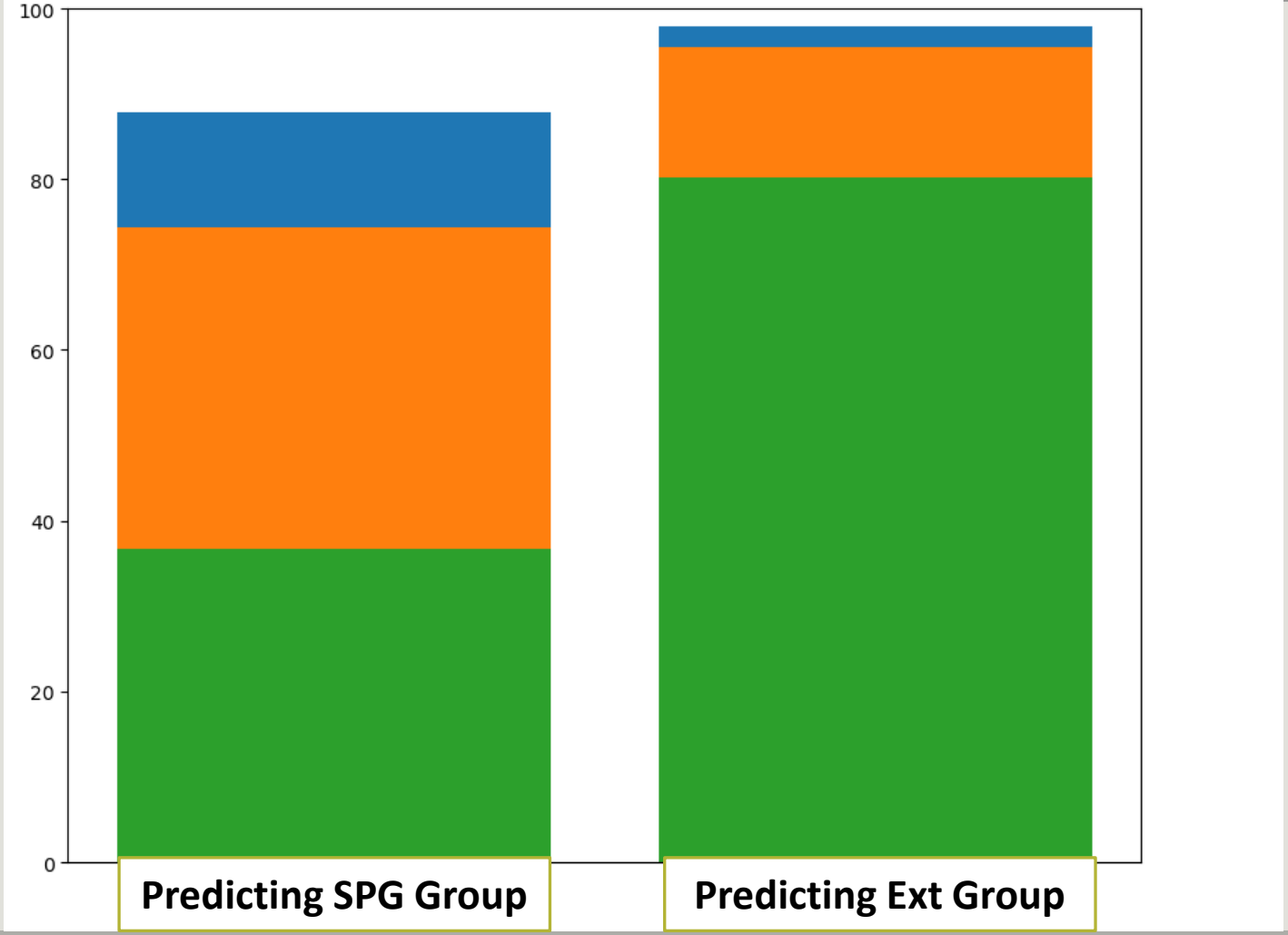
NO. Its Space Group B

What???



When testing, top 3 accuracy might catch this issue, yet model is still being confused during training

# Results: Space Groups vs Extinction Groups





- Both models have 10k per class
- Using Balanced Synthetic Data(We will get there later)
- **SPG Model achieved top 1 85% accuracy when outputs/labels were mapped to EXT**

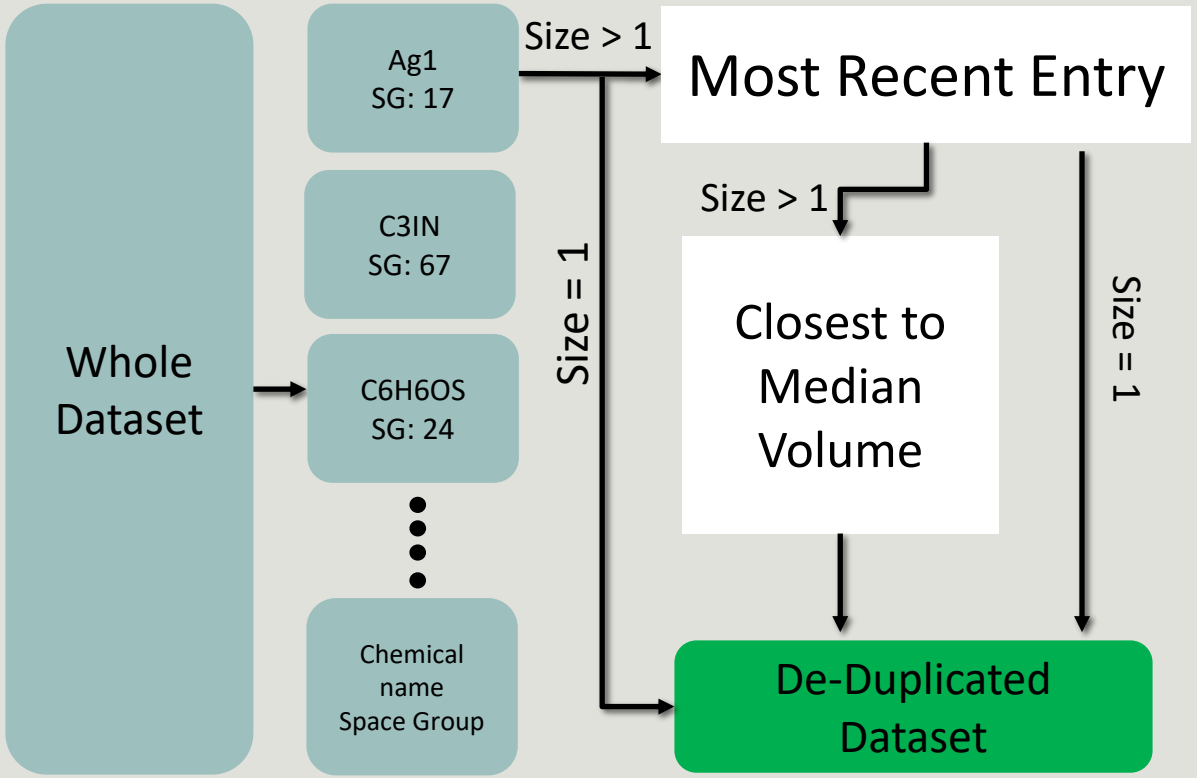
# Issue 2: Duplicates & De-Duplicating

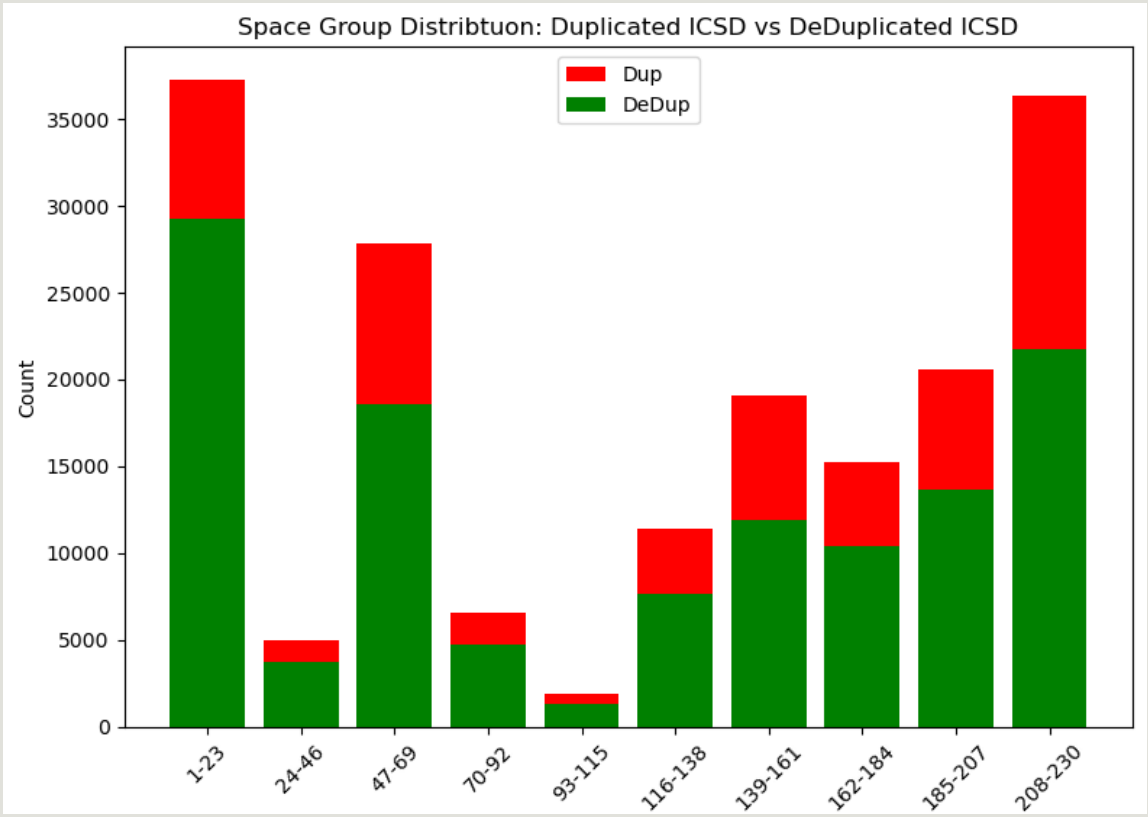
ICSD COD The Cambridge Crystallographic Data Centre

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Ag0.02Cu1.98S3Si1 | 155130 | 2006 | 4/1/2007 | monoclinic | 9 | 6.3405 | 11.257 | 6.2812 | 90 | 107.464 | 90 | 427.66 |
| Ag0.02Li0.98 | 58311 | 1969 | 4/1/2004 | cubic | 229 | 3.4913 | 3.4913 | 3.4913 | 90 | 90 | 90 | 42.56 |
| Ag0.034In0.037Sb0.764Te0.165 | 94288 | 2001 | ######## | hexagonal | 166 | 4.347 | 4.347 | 11.2415 | 90 | 90 | 120 | 183.96 |
| Ag0.034In0.037Sb0.764Te0.165 | 94289 | 2001 | ######## | hexagonal | 166 | 4.3553 | 4.3553 | 11.276 | 90 | 90 | 120 | 185.23 |
| Ag0.034In0.037Sb0.764Te0.165 | 94290 | 2001 | ######## | hexagonal | 166 | 4.3696 | 4.3696 | 11.5759 | 90 | 90 | 120 | 191.41 |
| Ag0.034In0.037Sb0.764Te0.165 | 94291 | 2001 | ######## | hexagonal | 166 | 4.3747 | 4.3747 | 5.8087 | 90 | 90 | 120 | 96.27 |
| Ag0.034In0.037Sb0.764Te0.165 | 426574 | 2013 | 2/1/2014 | hexagonal | 166 | 4.338 | 4.338 | 11.004 | 90 | 90 | 120 | 179.33 |
| Ag0.034In0.037Sb0.764Te0.165 | 426575 | 2013 | 2/1/2014 | hexagonal | 166 | 4.3032 | 4.3032 | 11.2623 | 90 | 90 | 120 | 180.61 |
| Ag0.03Cd0.985O1 | 29297 | 1960 | 1/1/1980 | cubic | 225 | 4.69385 | 4.69385 | 4.69385 | 90 | 90 | 90 | 103.42 |
| Ag0.03Mg0.97 | 58325 | 1950 | 4/1/2004 | hexagonal | 194 | 3.1936 | 3.1936 | 5.1769 | 90 | 90 | 120 | 45.73 |
| Ag0.04Cu3.96 | 604103 | 1993 | 8/1/2008 | cubic | 225 | 3.62 | 3.62 | 3.62 | 90 | 90 | 90 | 47.44 |

SAME CHEMICAL + SPACE GROUP → REALLY SIMILAR CRYSTAL STRUCTURES → REALLY SIMILAR XRD PATTERNS

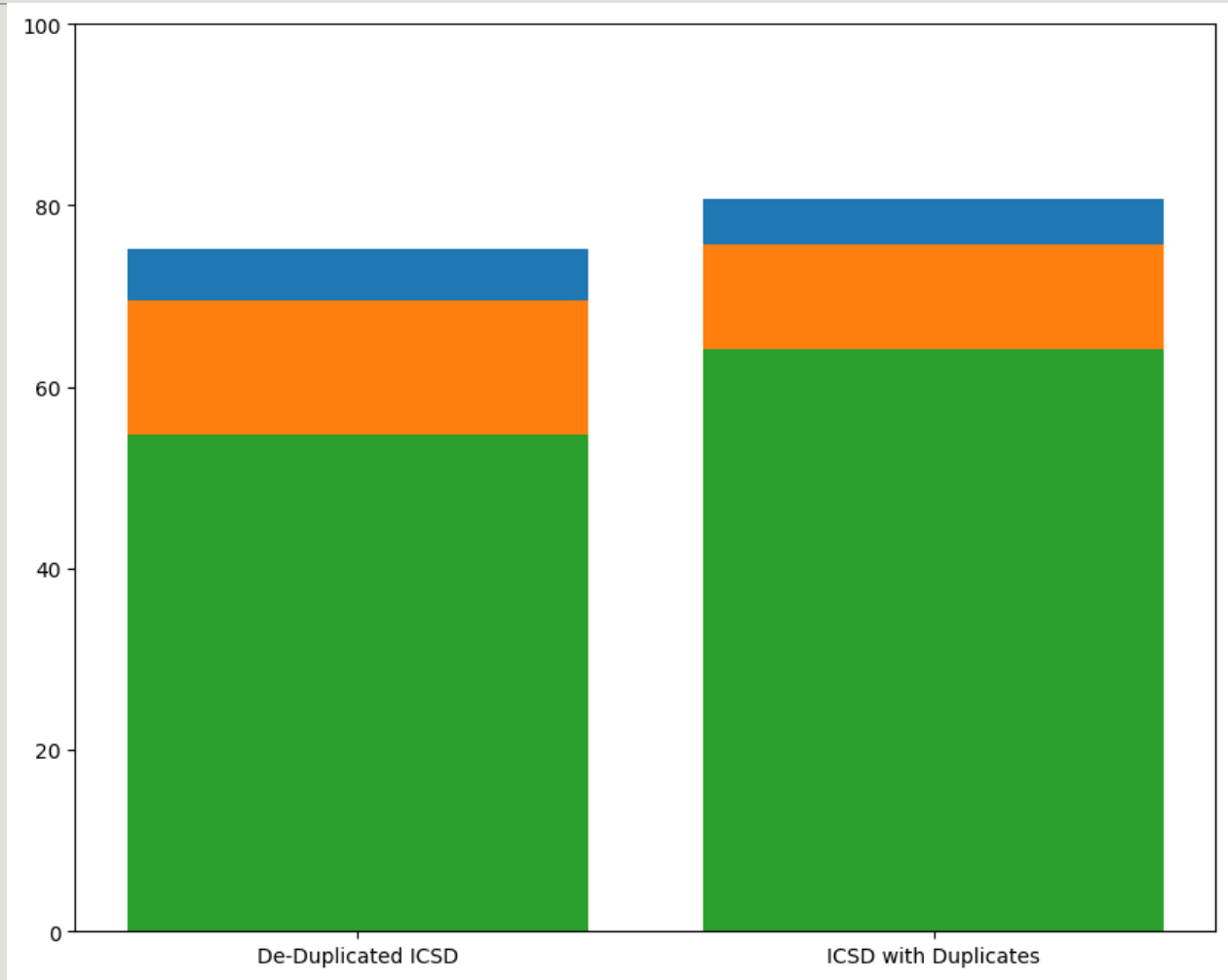
# Solution: De-Duping




ICSD Lost ~33% of data (~180k -> ~120k)!!!

# Result: Duplicated vs De-Duplicated

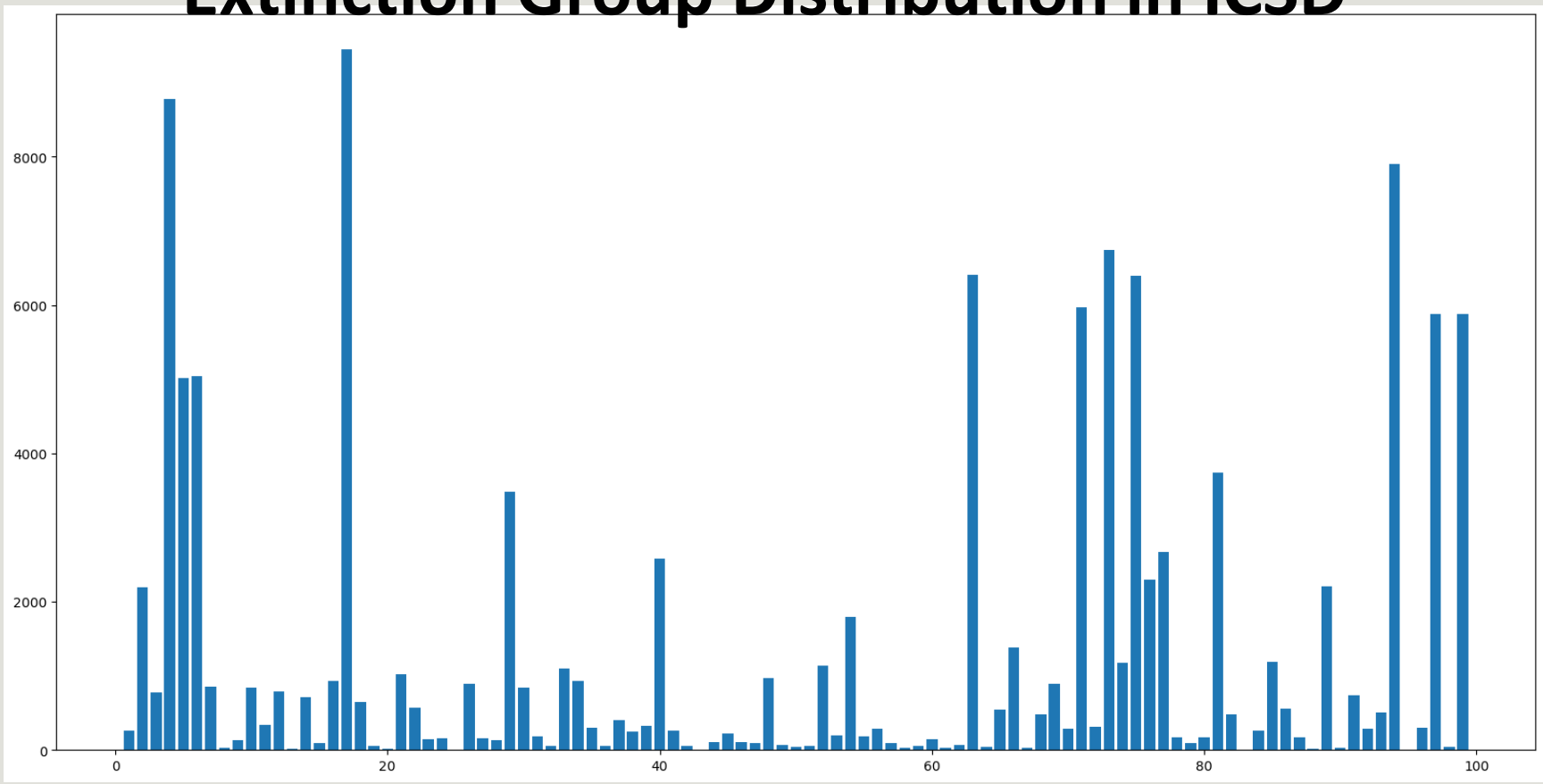




Two Potential Explanations

- Less Data (180k -> 120k), data scaling tests show this is worse

- Removing the “Cheating” in bias datasets

# Issue 3: Label Biases in Databases

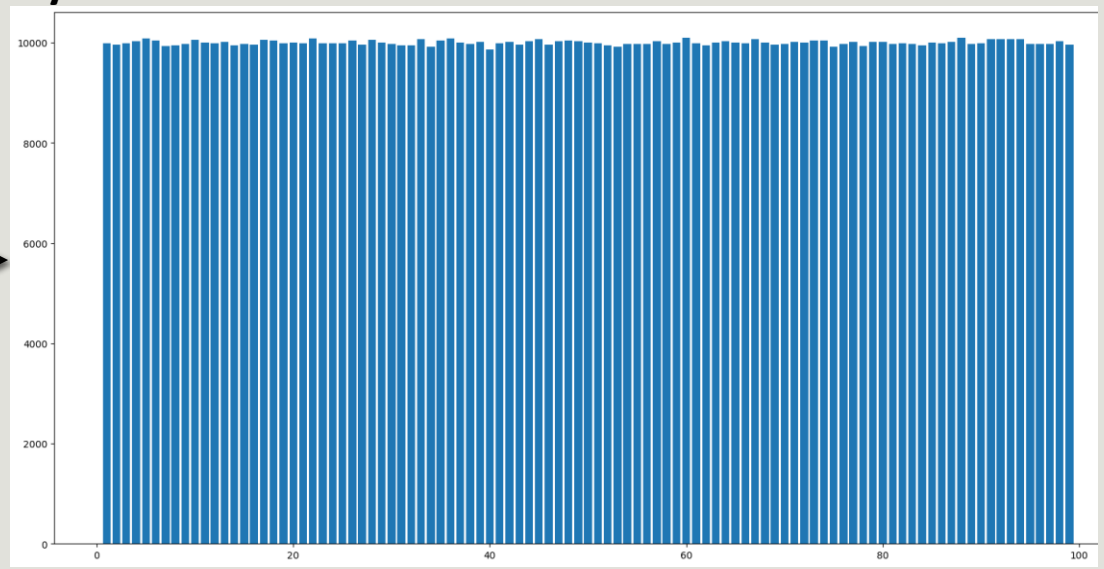


# Solution: Random Crystal Structure Database

Overarching Question: Can Neural Networks find **mapping from diffraction pattern to Crystal Structure**?

**We don't ONLY Need real, physical crystals to do this**

Request: Dataset with {x} Entries → For each Ext Group, generate x/99 crystal structures → RANDOM CRYSTAL GENERATOR MAGIC




Two Methods to generate patterns:
Reflection(Less Realistic)
PyCrysFML(more Realistic)

BALANCED DATASET

*Both methods create crystals that might be unnatural, yet still have "valid" patterns

# Results: Biased vs Balanced Dataset

| Trained Model | Biased Test Set | Balanced Test Set |
|---|---|---|
| **Top-1 Accuracy** | | |
| Biased Trained Model | | |
| Larger Unbiased Trained Model | | |
| **Top-3 Accuracy** | | |
| Biased Trained Model | | |
| Larger Unbiased Trained Model | | |
| **Top-5 Accuracy** | | |
| Biased Trained Model | | |
| Larger Unbiased Trained Model | | |

Biased Set -> ICSD (120k Entries)
Balanced Set -> 1.23 million “Fake” Crystals,

Bias on Unbiased Performance < Unbiased on Bias Performance

# Duplicates vs Bias, and why they're both bad

## Duplicates

- Very (VERY VERY) Similar Entries

- Leads to cheating, model is "memorizing" certain patterns on test

Bias

- Some classes/groups are represented more then others

- Possible Origins:
  - Man-Made:
    - Errors in Classification(Misclassifying complex groups as similar ones)
    - Certain types of materials are of more interest(Semiconductors, Magnetic Materials, etc)
  - Nature Made:
    - More Materials fall under certain groups than others

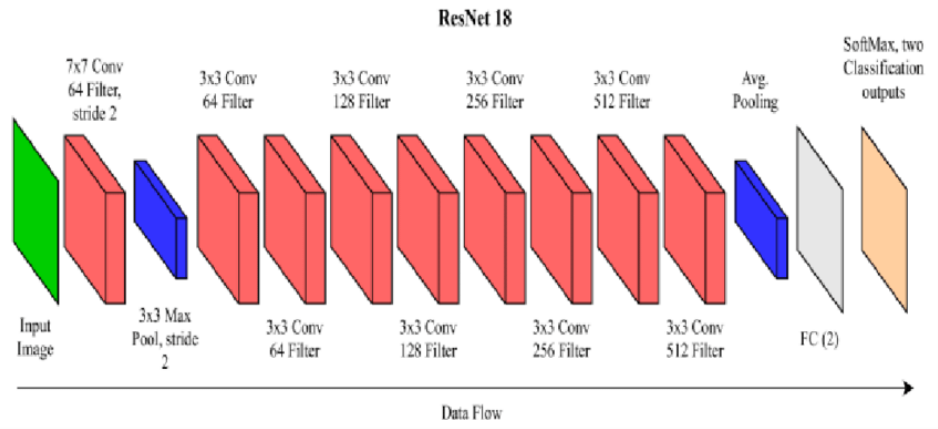
**- More Bias = Model learning distribution rather than mapping**

# Issue 4: ResNet CNN Itself

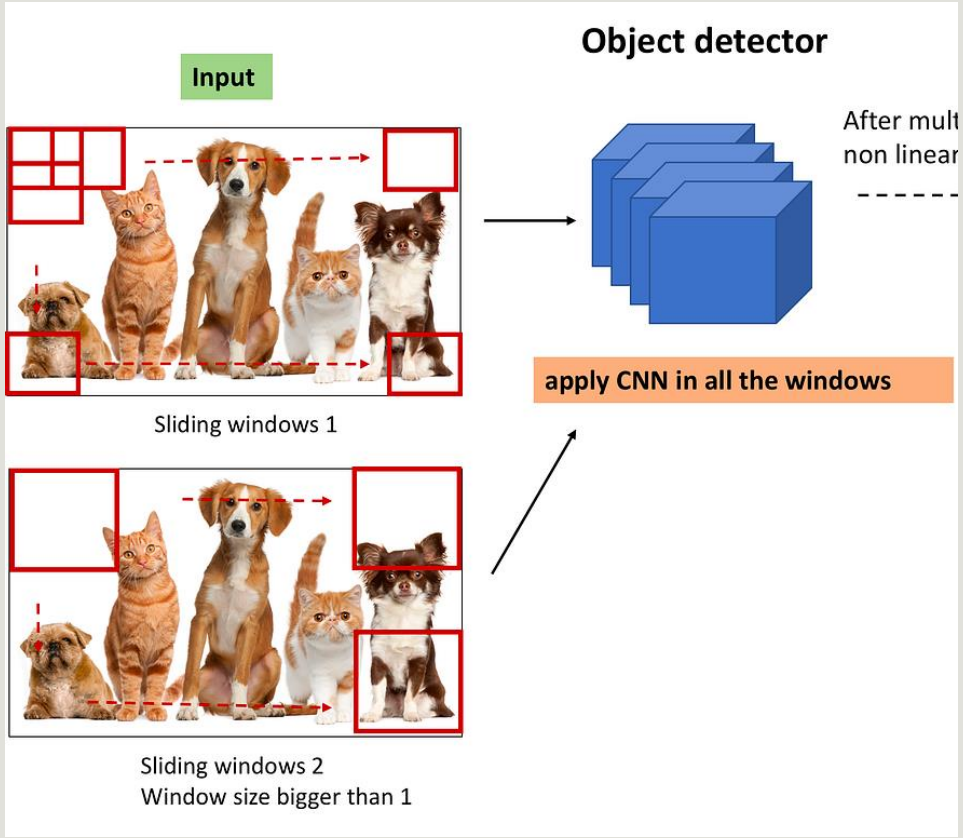
## Potential Issue 1: Not Scaling

"AI Moore's Law": More data + compute = increase performance, **unless there's a bottle neck**

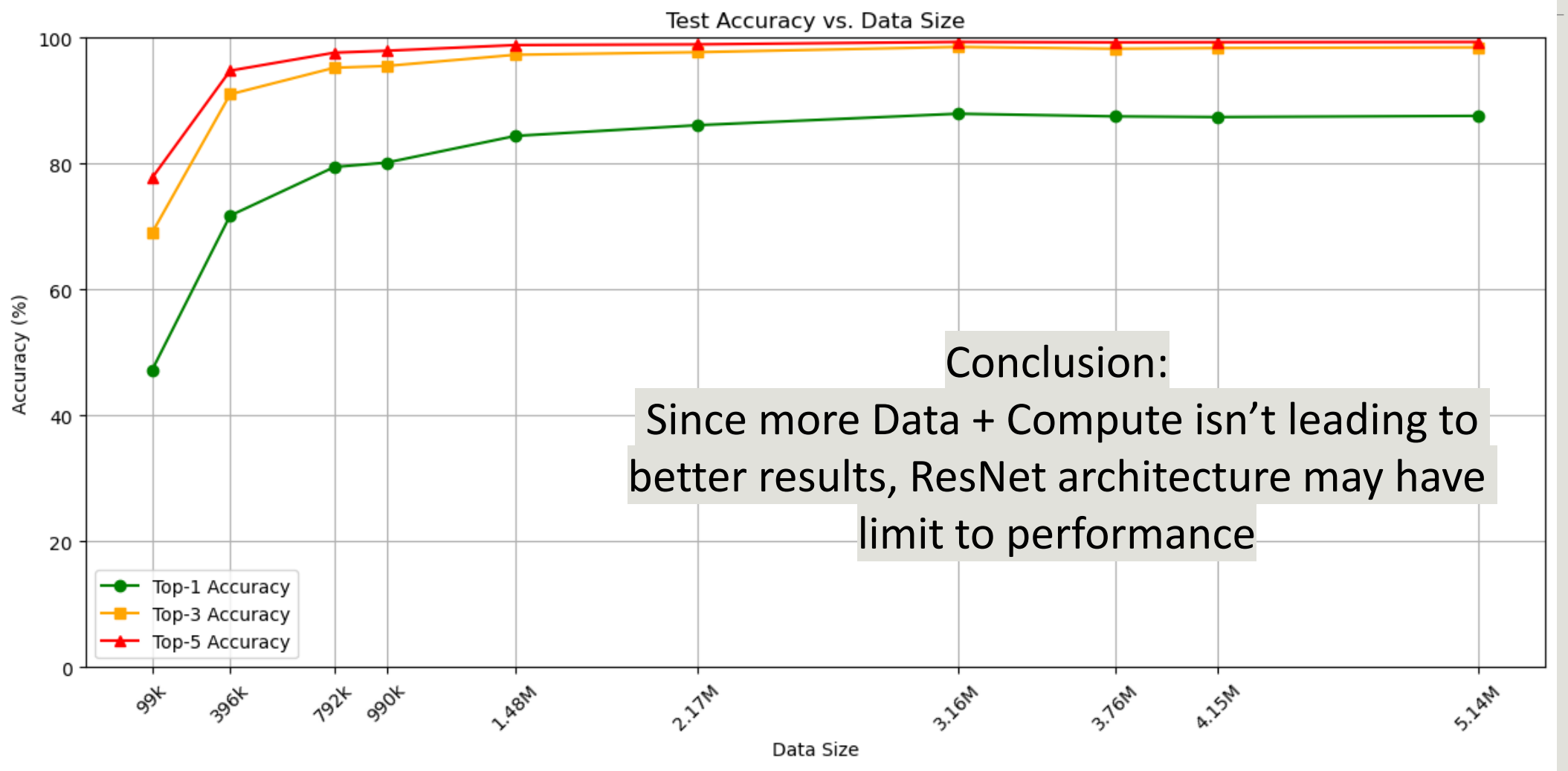
Architecture itself **might** be limited




## Potential Issue 2: CNN Localization



# Scaling Data – Unbiased Data




Conclusion:

Since more Data + Compute isn't leading to better results, ResNet architecture may have limit to performance

# Conclusions and Next Steps

- De-duping matters, Past Studies Results have slight skew
- Bias Datasets don't generalize well, like at all
- Testing on extinction groups not only leads to better performance, but our model just gets less confused
- ResNet is reaching a bottleneck, meaning its time to shift to a new architecture
- We do better then majority of previous studies

**Next Steps**

- Get CNN to Reproduce high results on PyCrysFML data and real data
- (MAYBE) Train  a transformer
- Publish!

# Acknowledgments

Collaborators:

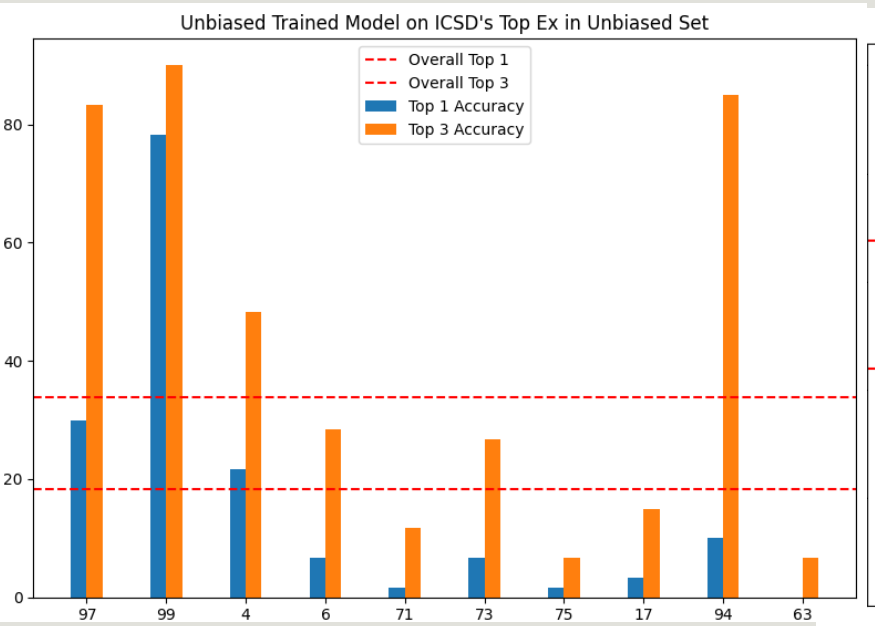
Elizabeth Baggett[1 2], William Ratcliff[1 3], Paul Kienzle[1], Derrick Chan-Sew[4], Abhishek Shetty[4], Vanellsa Acha[4]

[1]National Institute and Standards in Technology, [2]Boston College, [3]University of Maryland, [4]University of California, Berkeley.

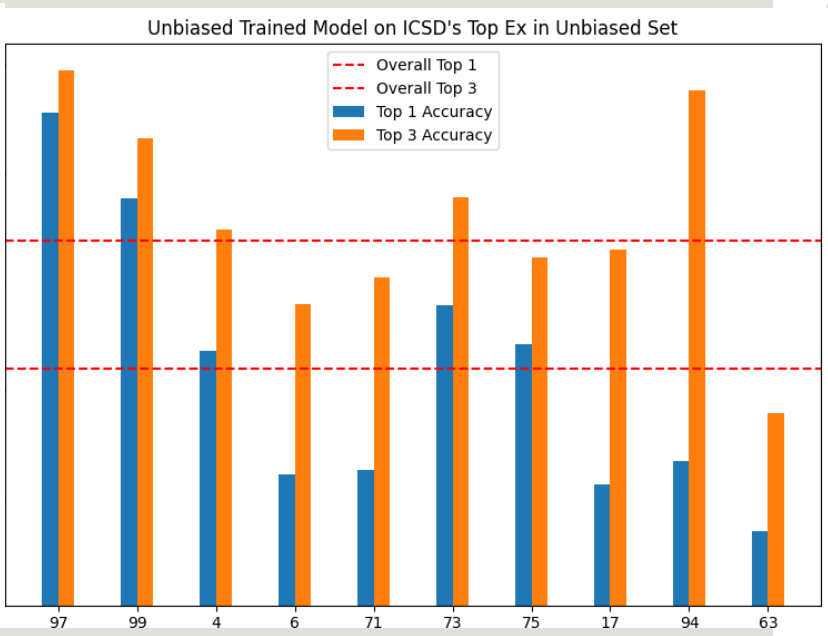
Special thanks to:

- Julie Borchers, Leland Harriger (CORE Program Directors)
- Center of High-Resolution Neutron Scattering (CHRNS)
- UTexas Advanced Computing Center
- NIST HPC Center
- National Institute of Standards and Technology(NIST) and NIST Center of Neutron Research(NCNR)

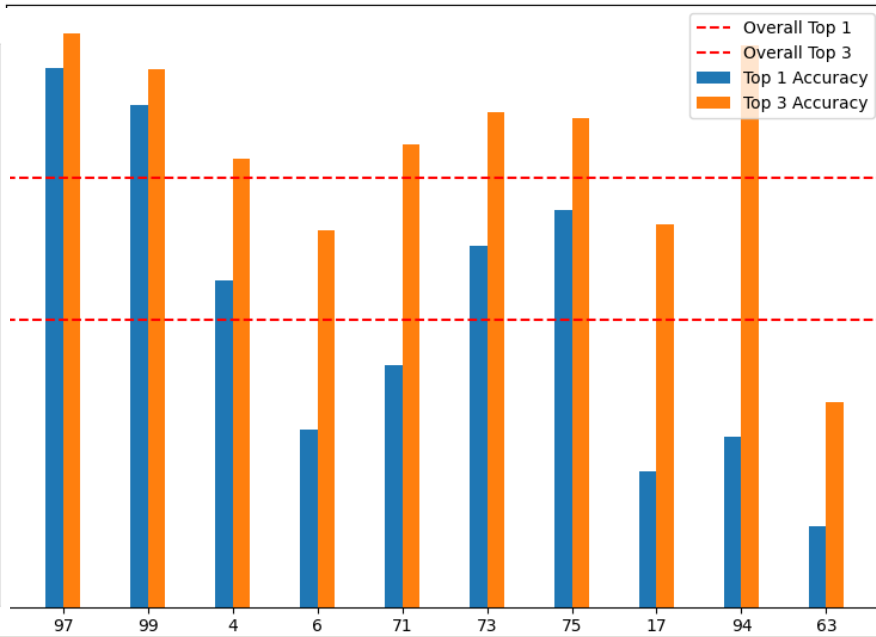
# Why our Balanced Models Underperform?




120k Amount




1.23m Amount

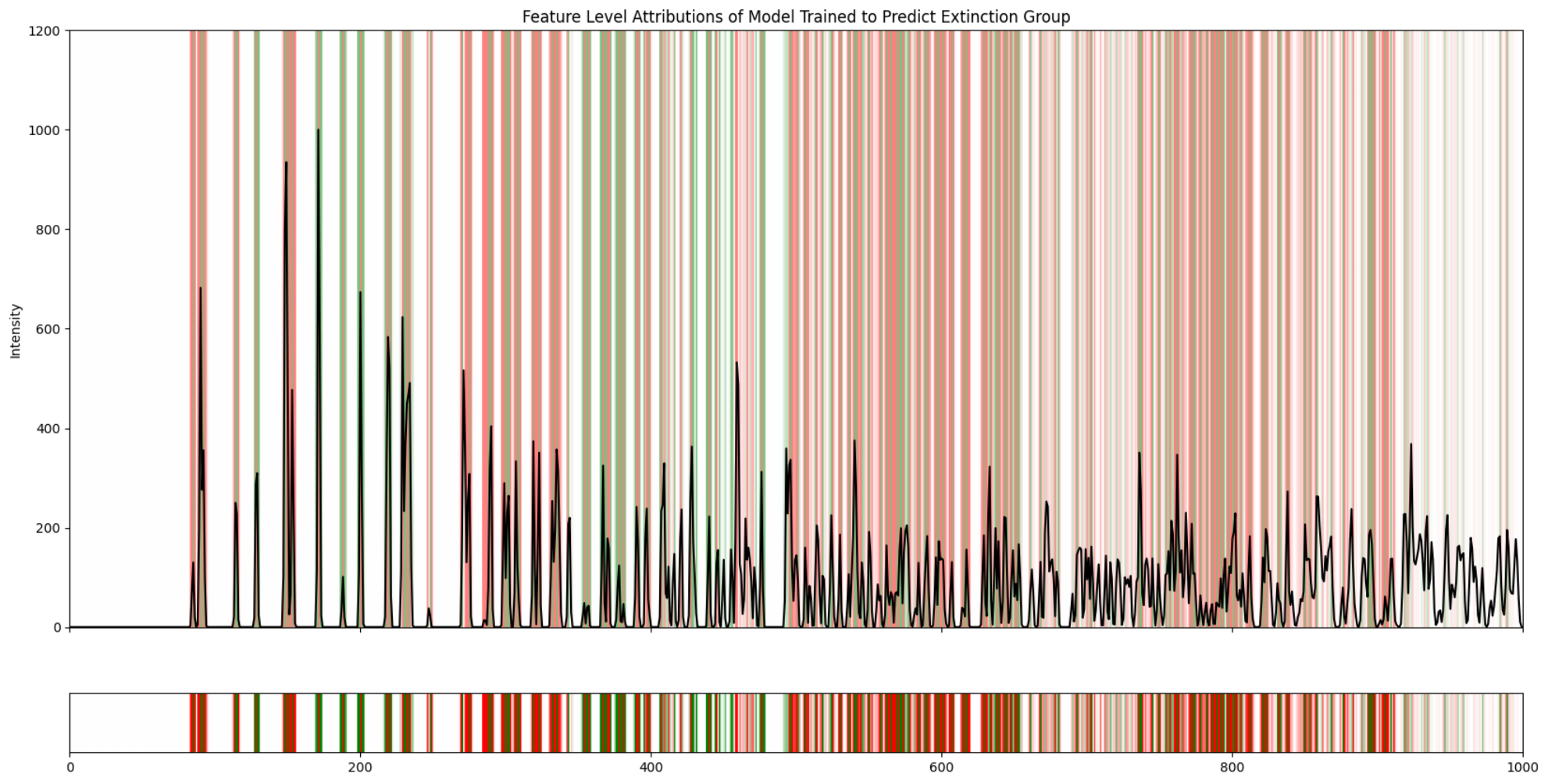



2m Amount

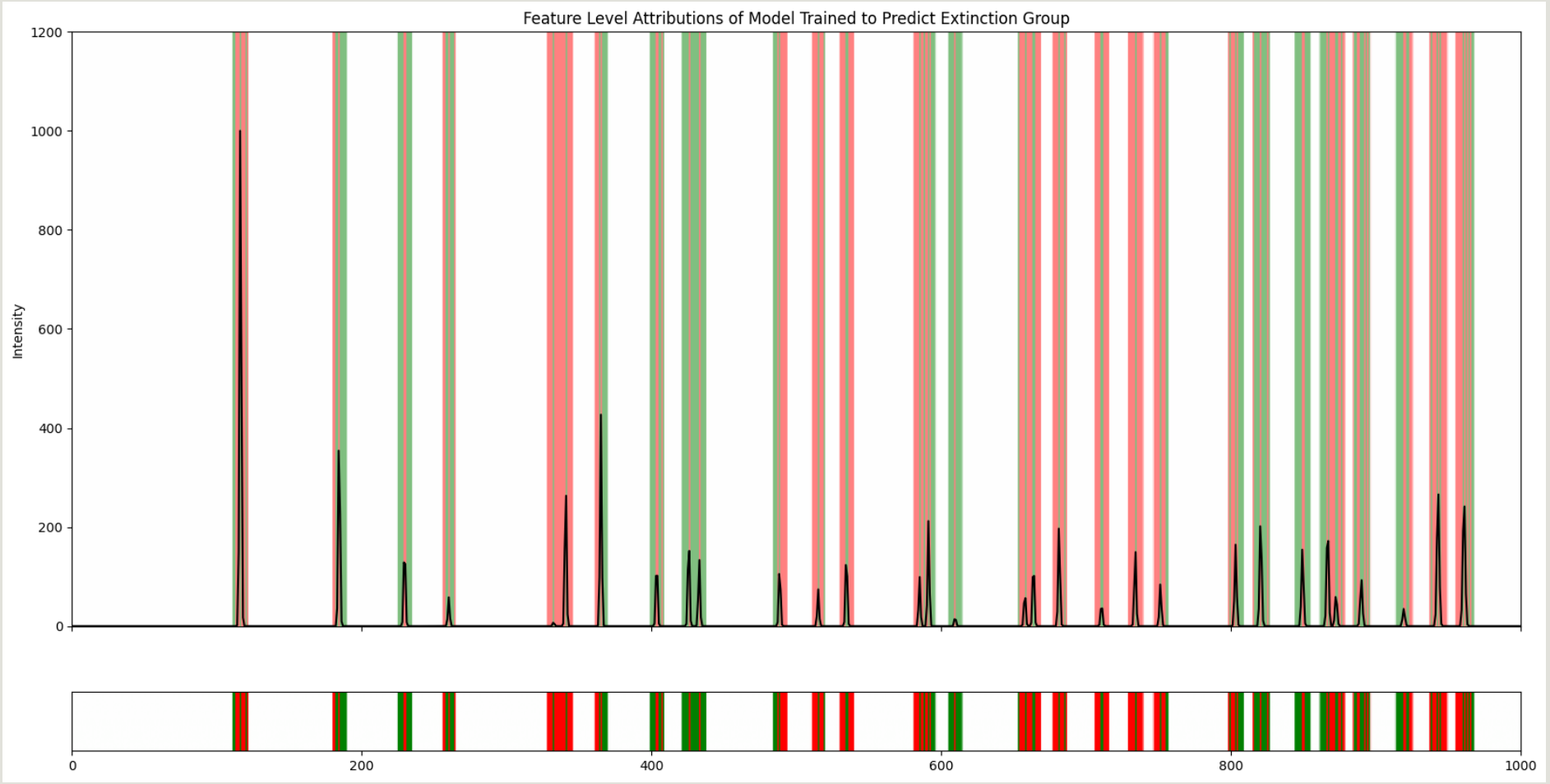
# Why are are against bias: Model doesn’t learn

# Why we are against bias: Its not Natural

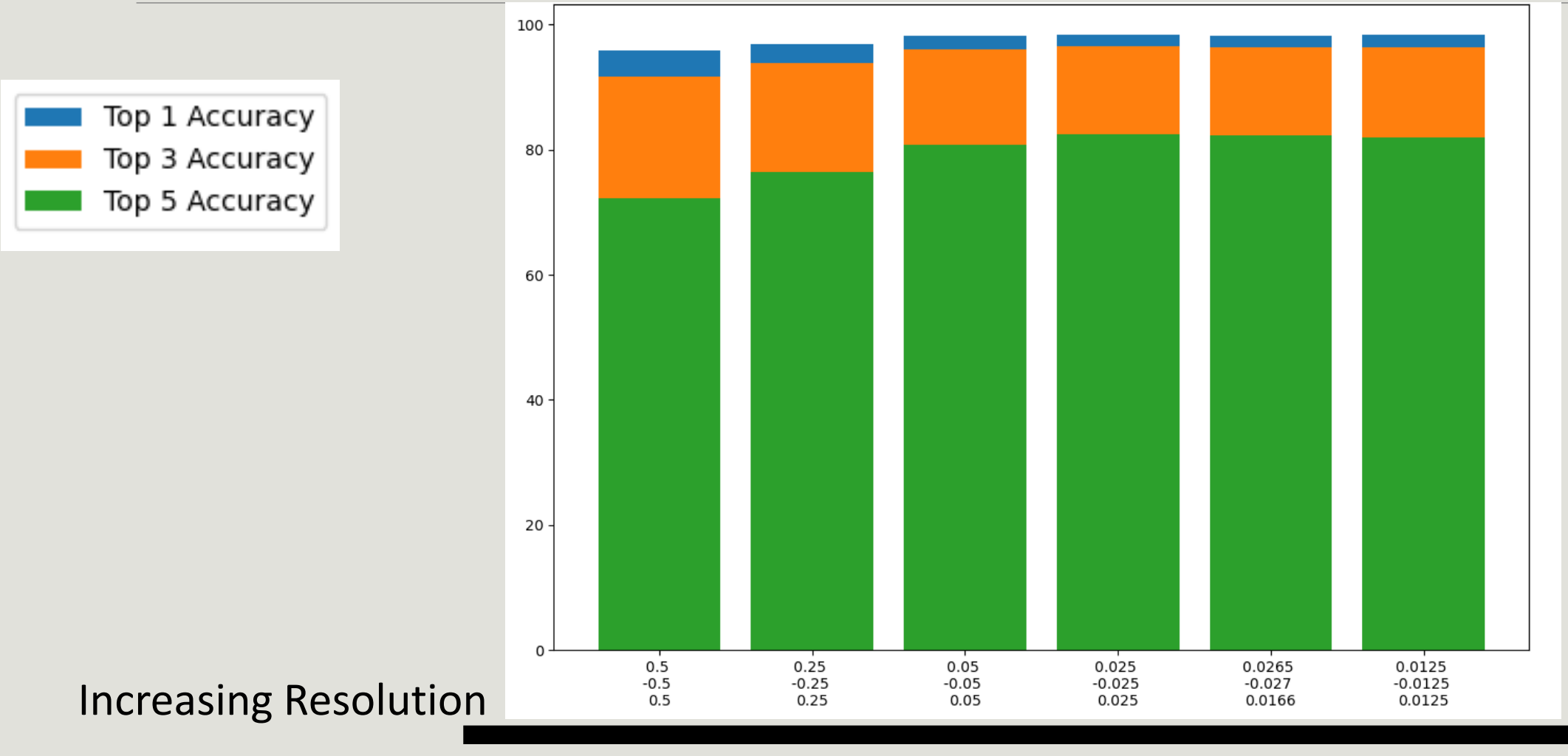
# Interpretability – Ext Group



# Interpretably – Space Group



# Resolution


Top 1 Accuracy
Top 3 Accuracy
Top 5 Accuracy
100
80
60
40
20
0
0.5
-0.5
0.5
0.25
-0.25
0.25
0.05
-0.05
0.05
0.025
-0.025
0.025
0.0265
-0.027
0.0166
0.0125
-0.0125
0.0125
Increasing Resolution