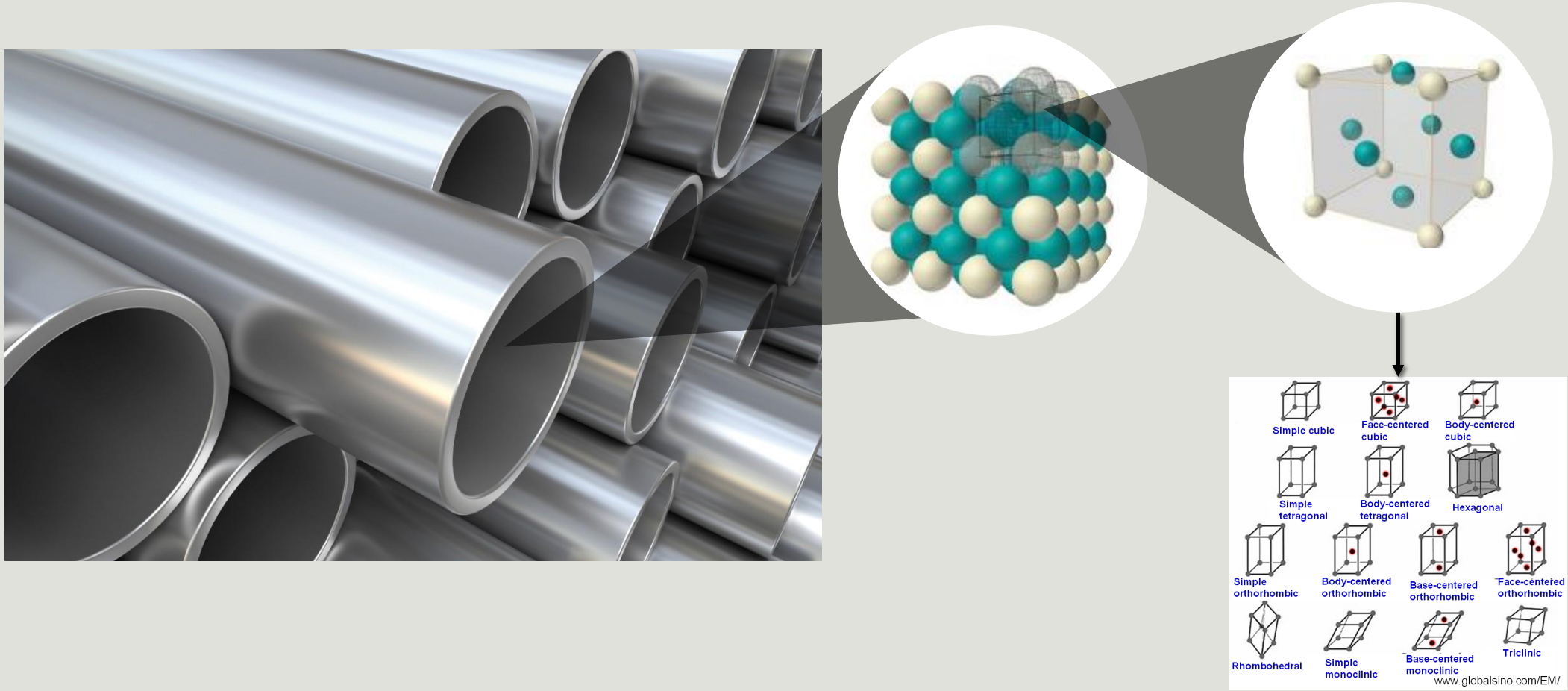
Classifying Crystal Structure using Deep Learning

EDWARD FRIEDMAN, WHEATON HIGH SCHOOL, SILVER SPRING MD

MENTOR: WILLIAM RATCLIFF

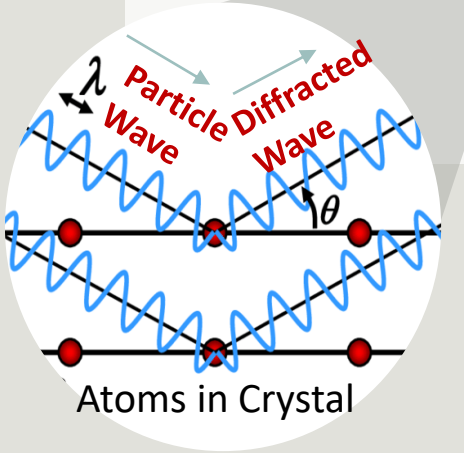
Crystallography



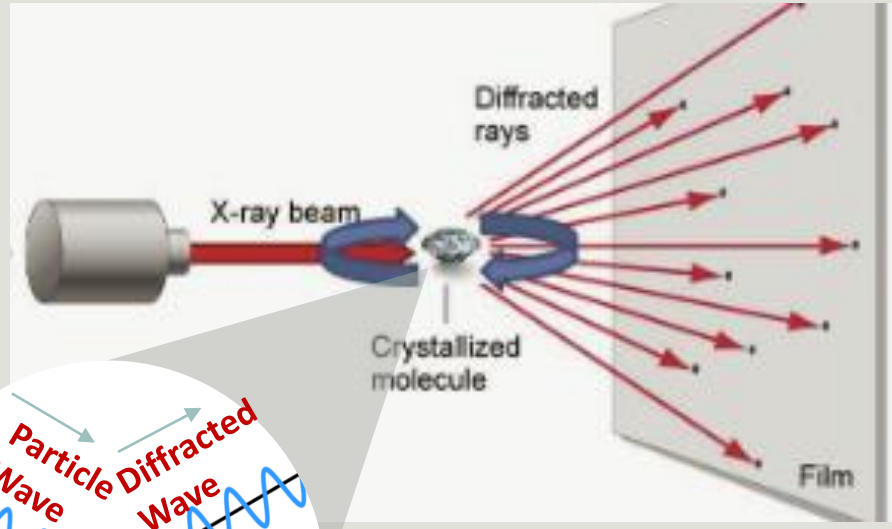
Powder Diffraction

Unknown Power of Crystals

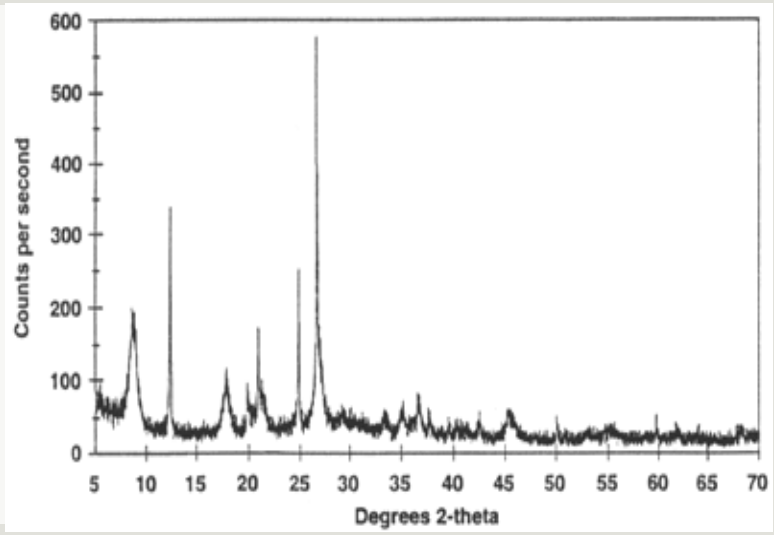
IN: Neutron beams



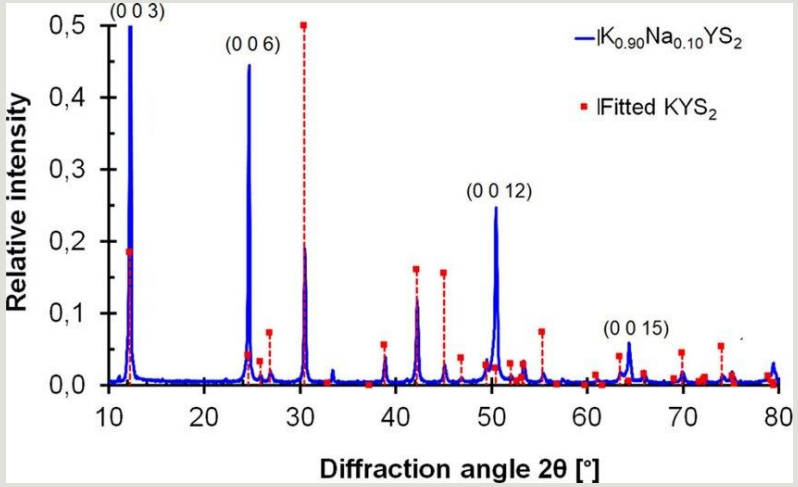
Atoms in Crystal



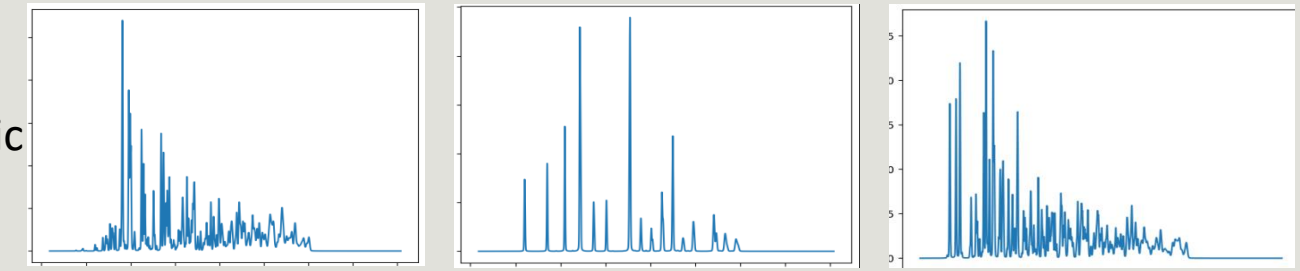
Out: Particle intensities at each angle



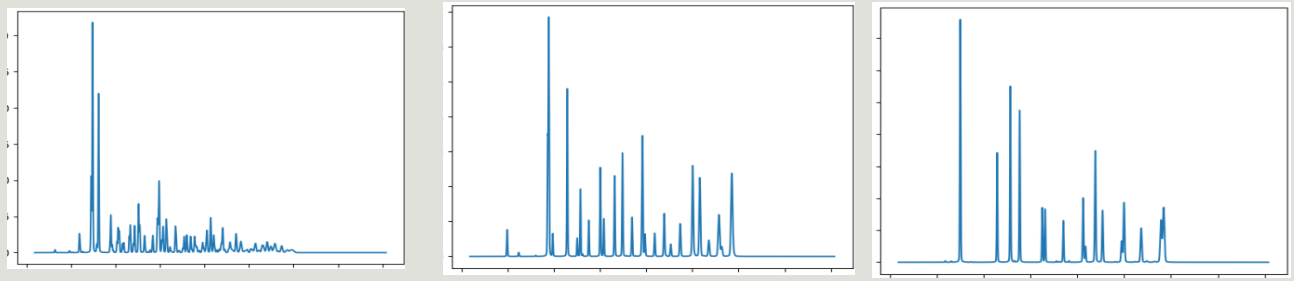
Classification and Analysis: The Hard Part



3 examples of P-Orthorhombic crystals



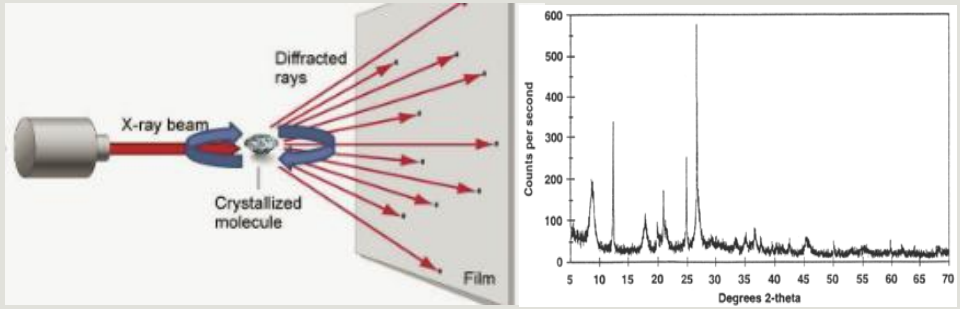
3 examples of I-Tetragonal crystals



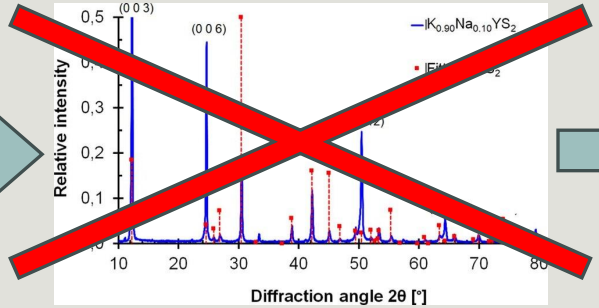
The Goal

Unknown Crystal Structure

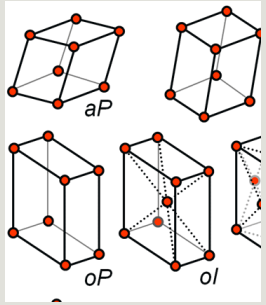
Diffraction Experiments

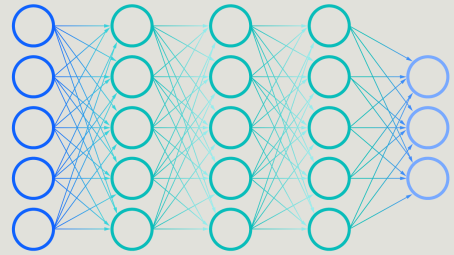


Indexing/Classification



Bravais lattice & Space Group



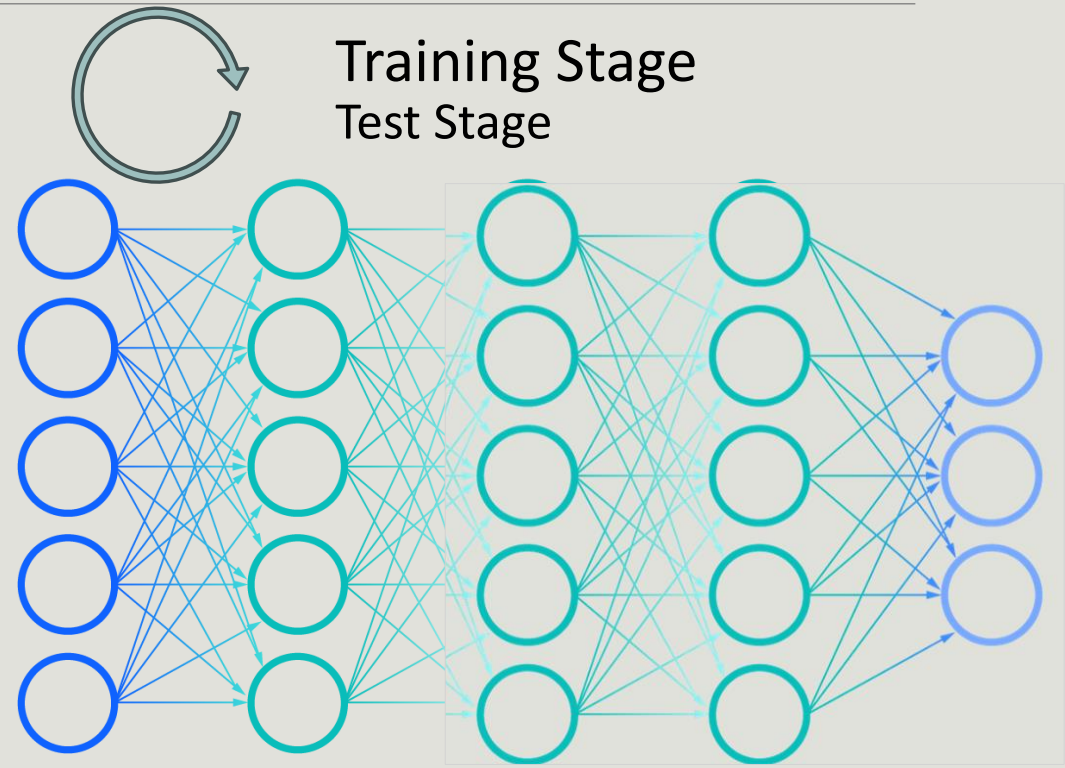


Neural Network

Why Deep Learning??!

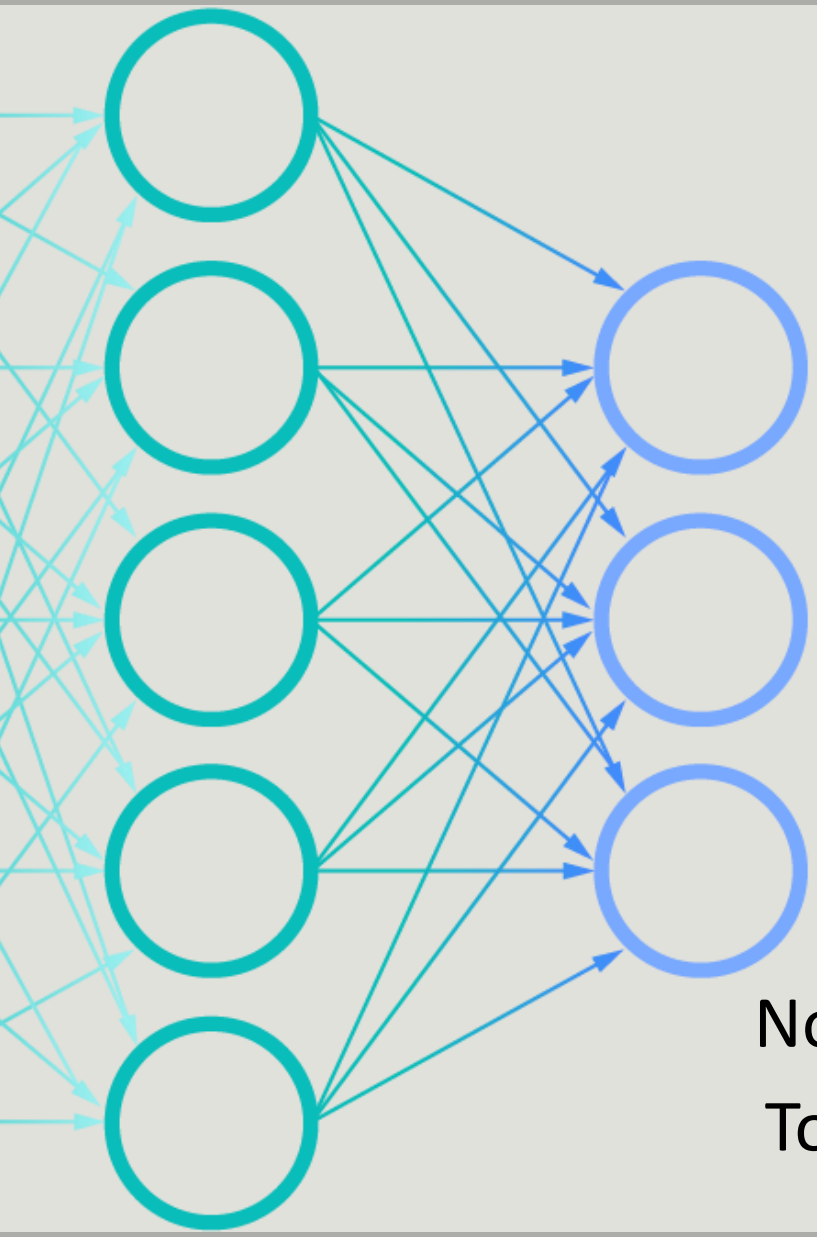
Deep Learning and Neural Networks(NN)

Training = Homework,
Testing = Exam



1 epoch = 1 pass through dataset

How? Really good math



What do we get?

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

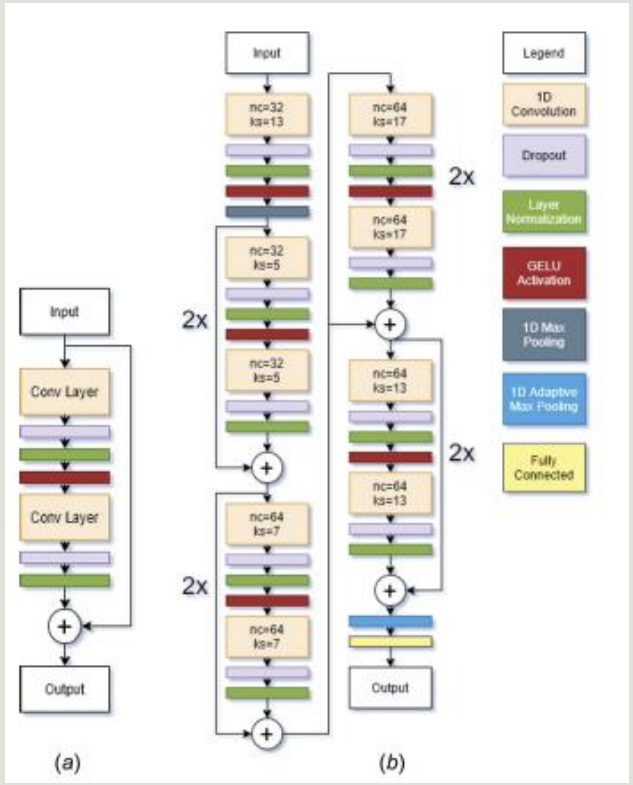
[0.88,0.12] →

Normal/Top 1 Accuracy: If label is highest model prediction

Top 3 Accuracy: If Label among top 3 model predictions

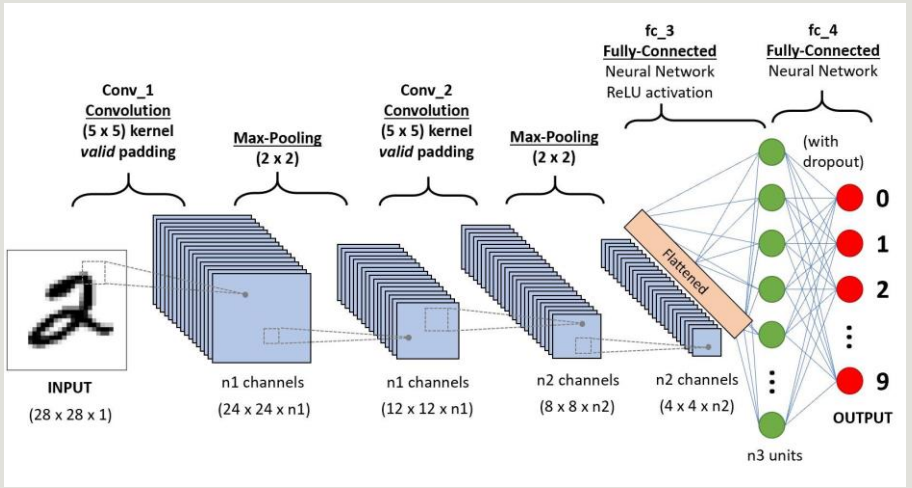
Methodology

ResNet CNN



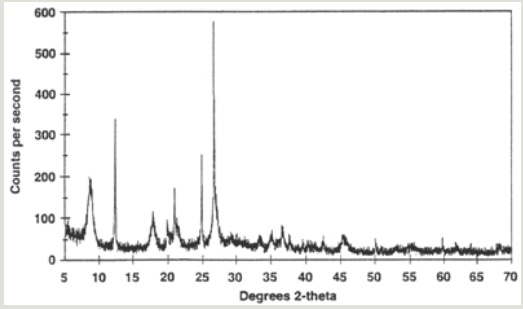
(Lolla et al.,2021)

CNN



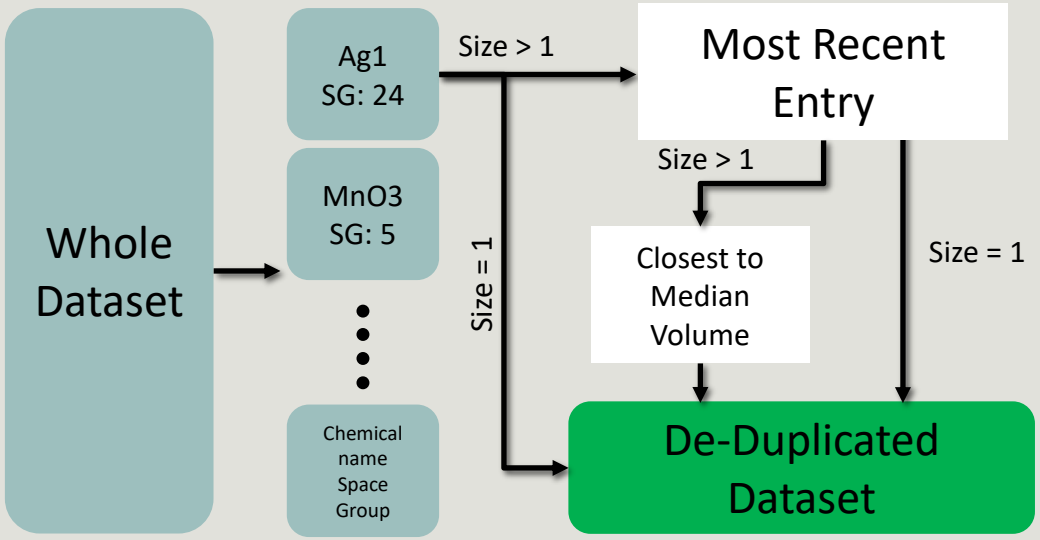
ICSD

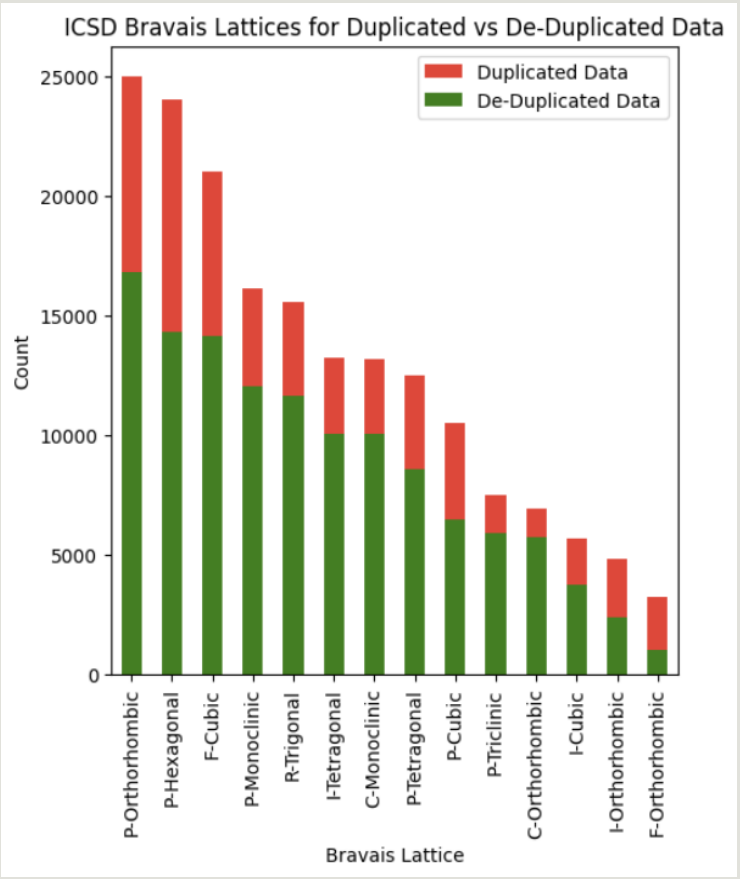
180k Set



De-Duplicating

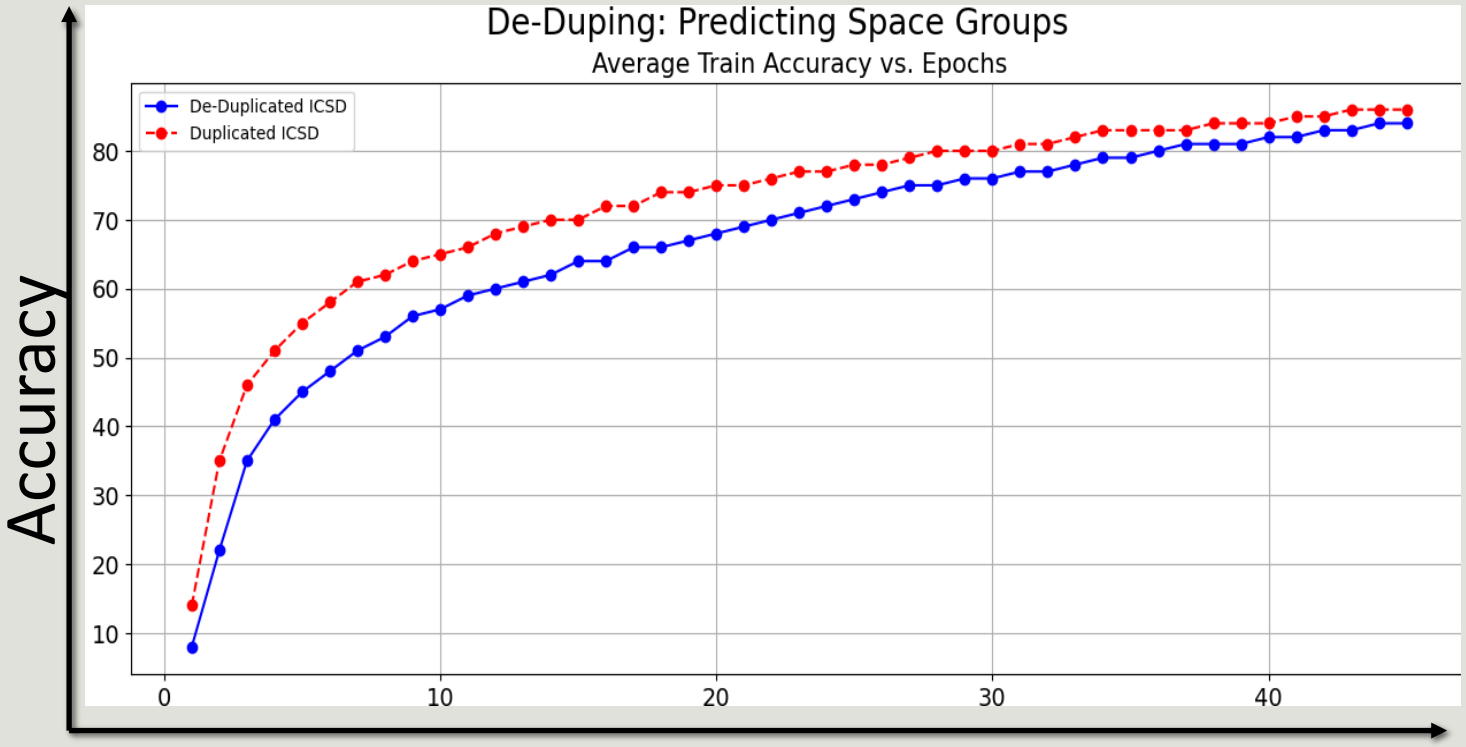
Chemical Name		PubYear			Space Group							
Ag0.02Cu1.98S3Si1	155130	2006	4/1/2007	monoclinic	9	6.3405	11.257	6.2812	90	107.464	90	427.66
Ag0.02Li0.98	58311	1969	4/1/2004	cubic	229	3.4913	3.4913	3.4913	90	90	90	42.56
Ag0.034In0.037Sb0.764Te0.165	94288	2001	########	hexagonal	166	4.347	4.347	11.2415	90	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

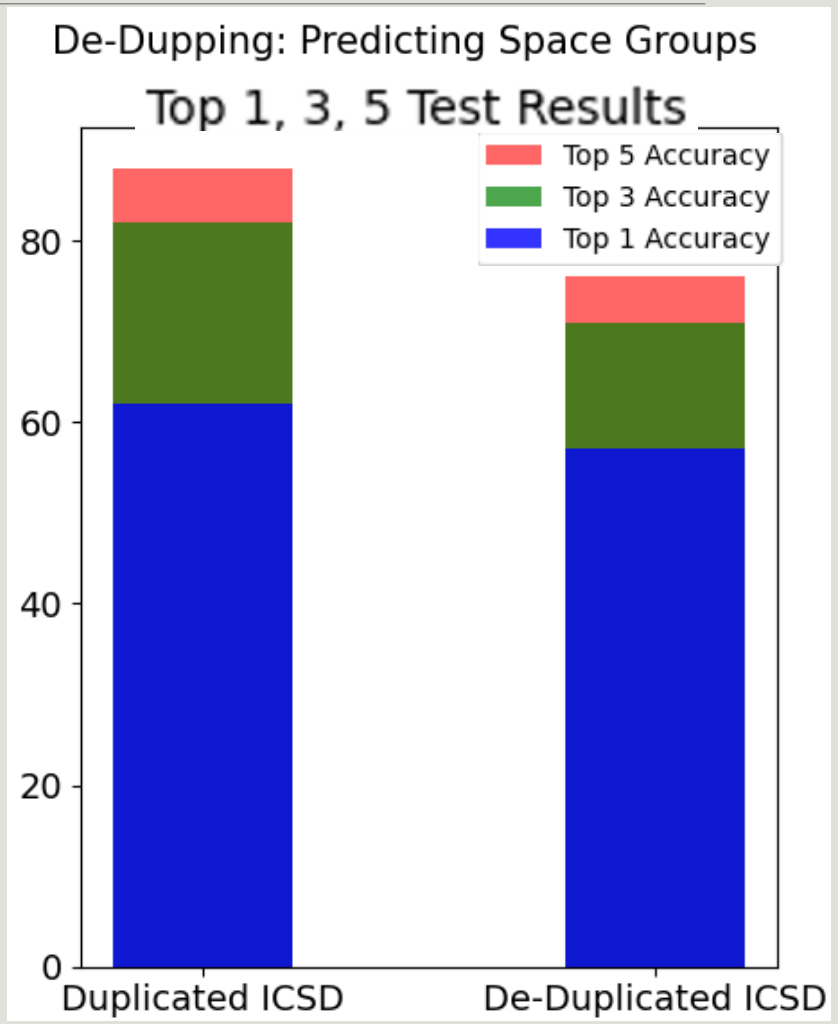




Lost ~33% of data (~180k -> ~120k)

Duplicated Vs De-Duplicated Data Performance

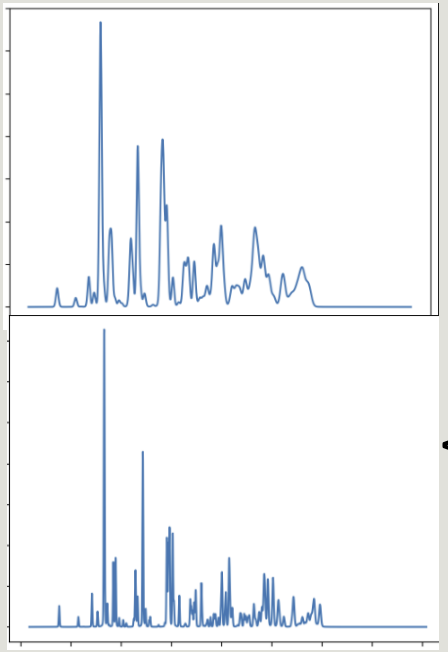




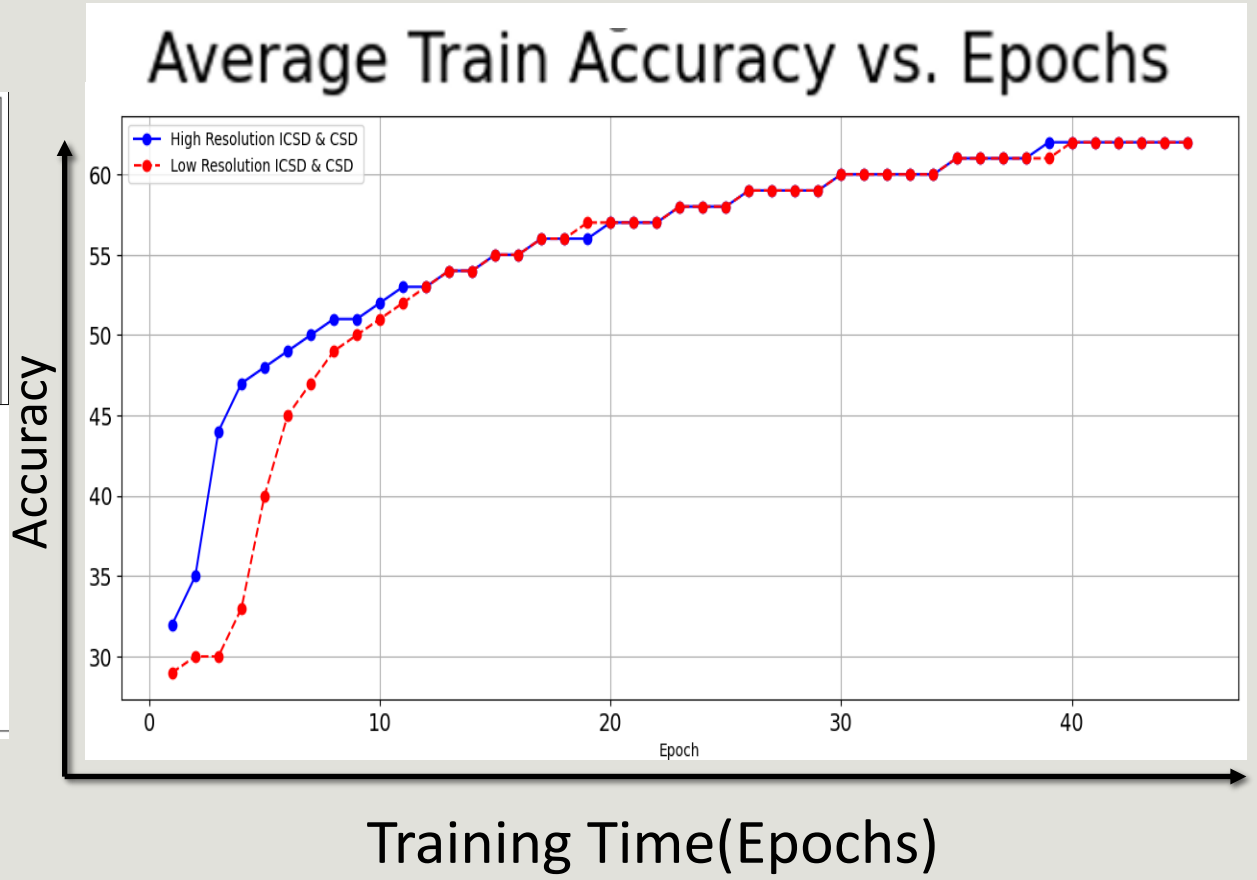
Loosing ~33% of data

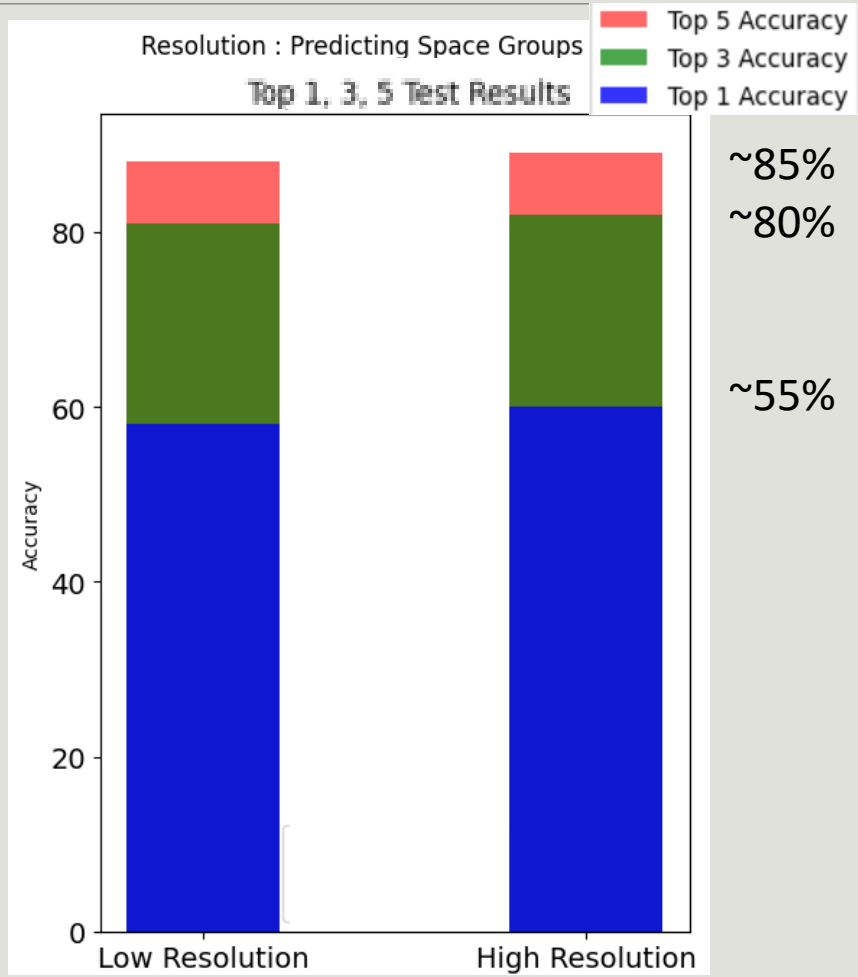
Does the resolution of the experiment help?

Low Res



High Res

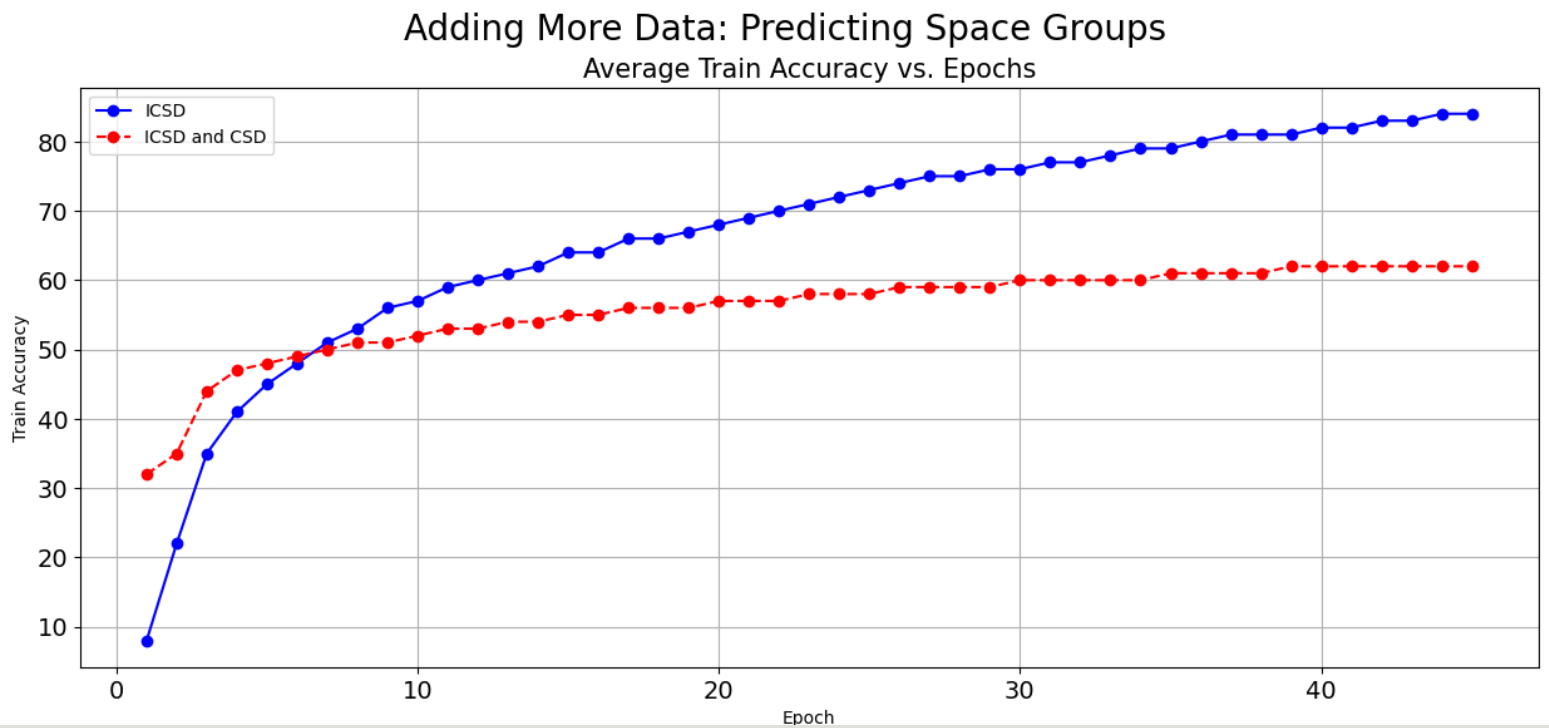


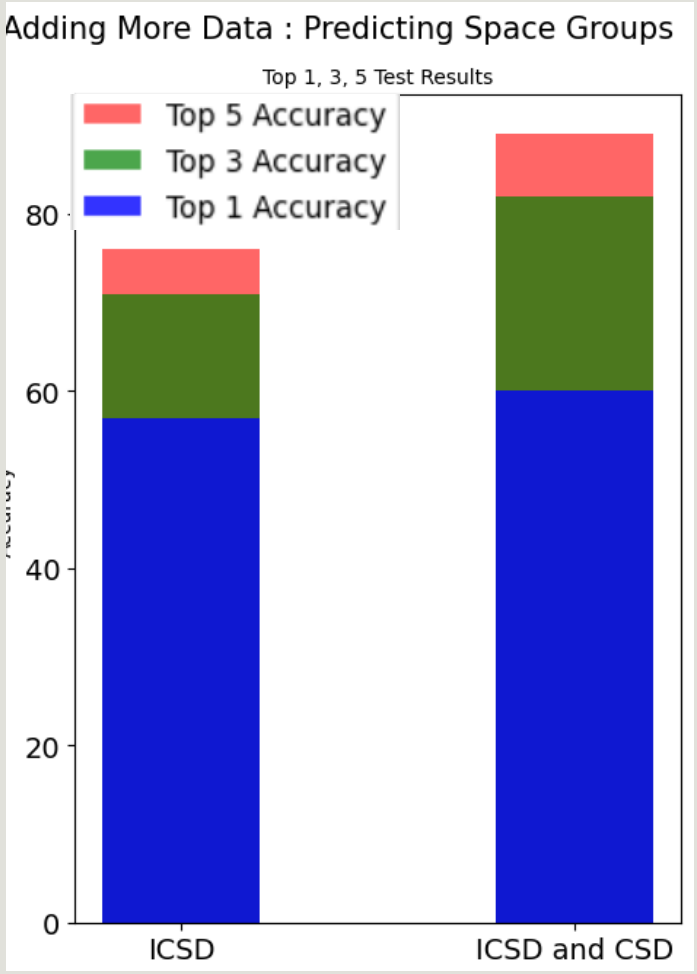


Let's add the more data to try to get better?

ICSD ~120k → CSD + ICSD ~1.1 million

Training



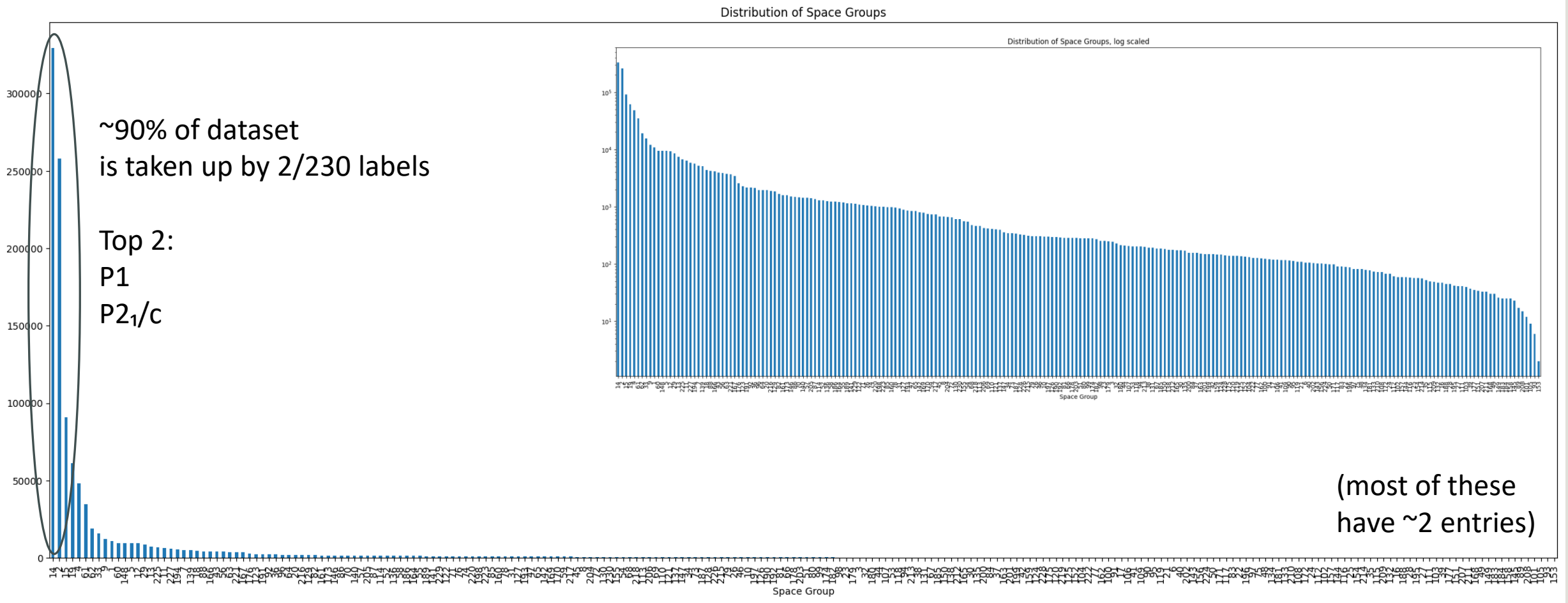


~91%

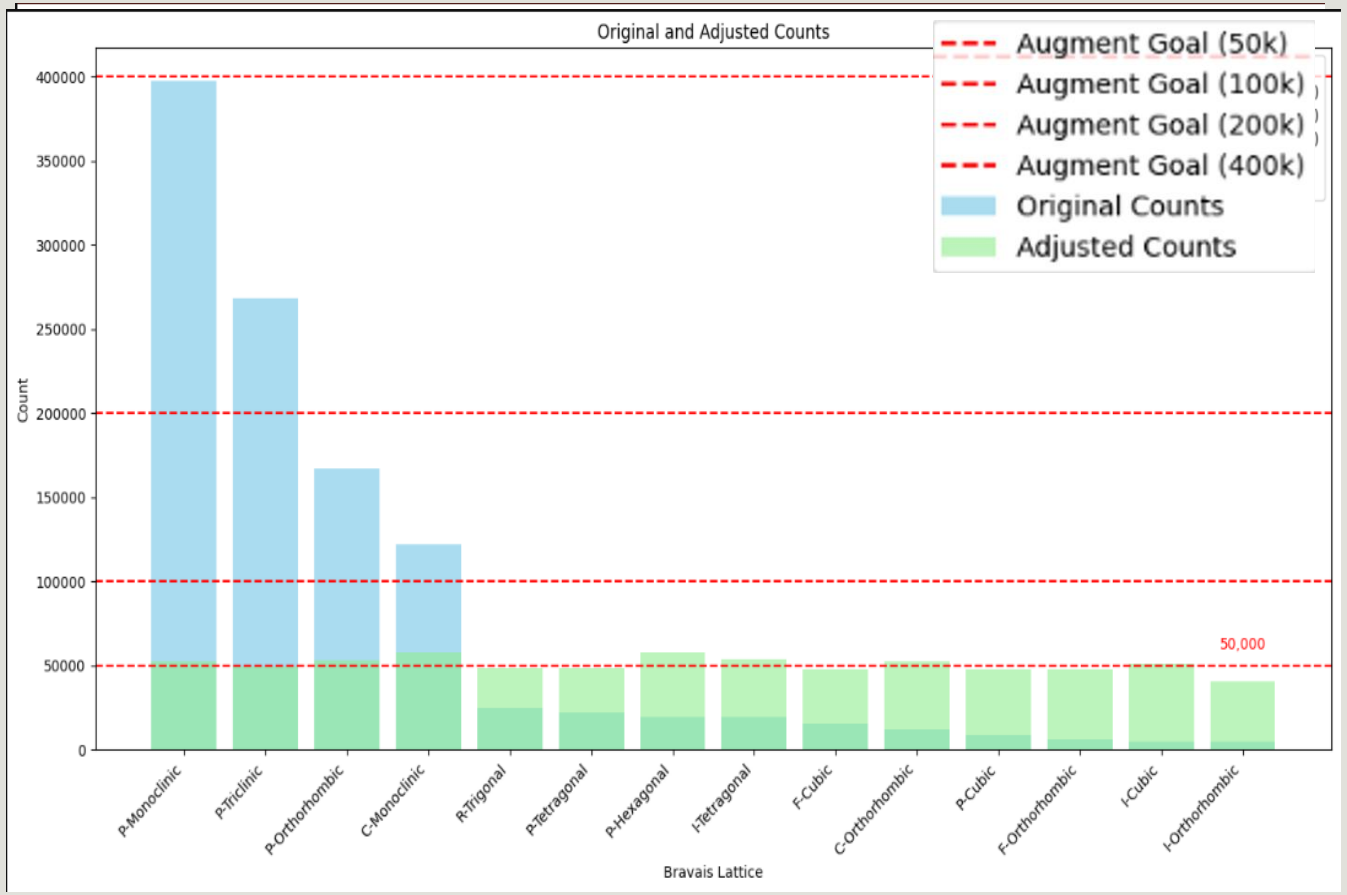
~80%

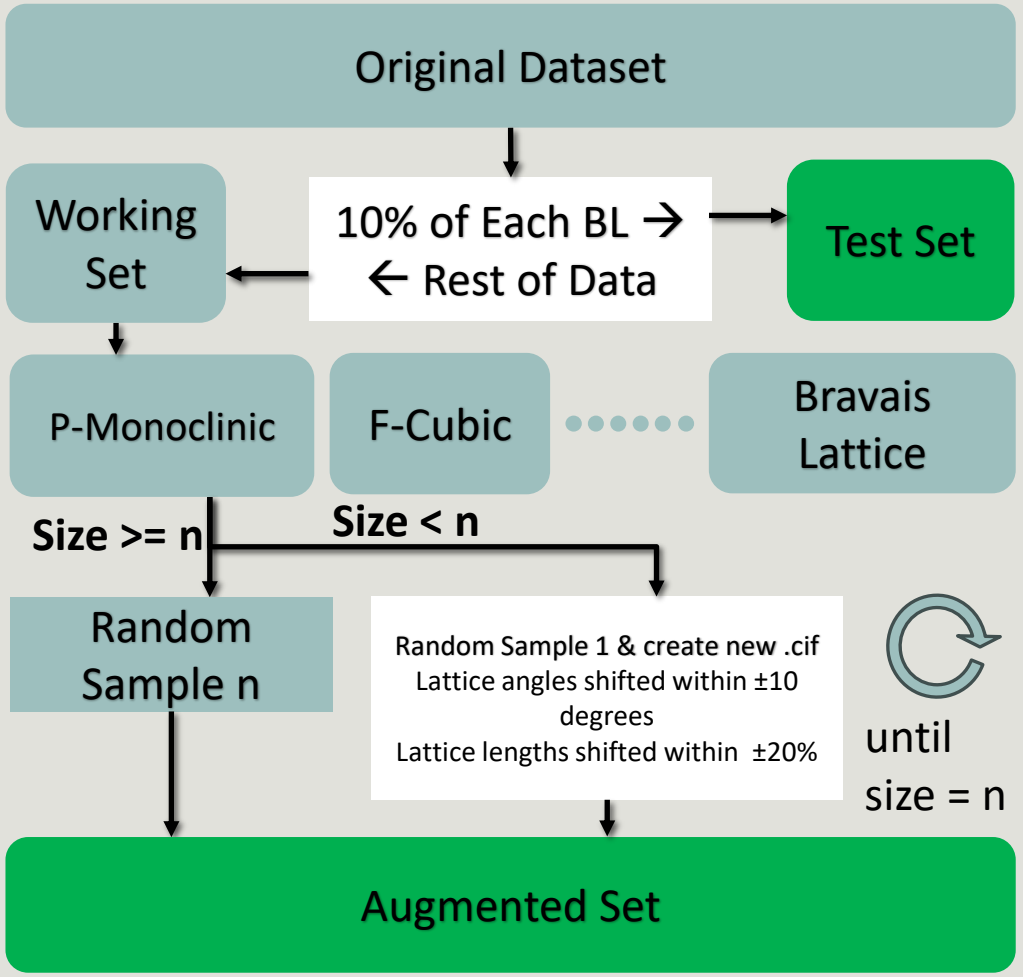
~58%

The real issue: Label Imbalance

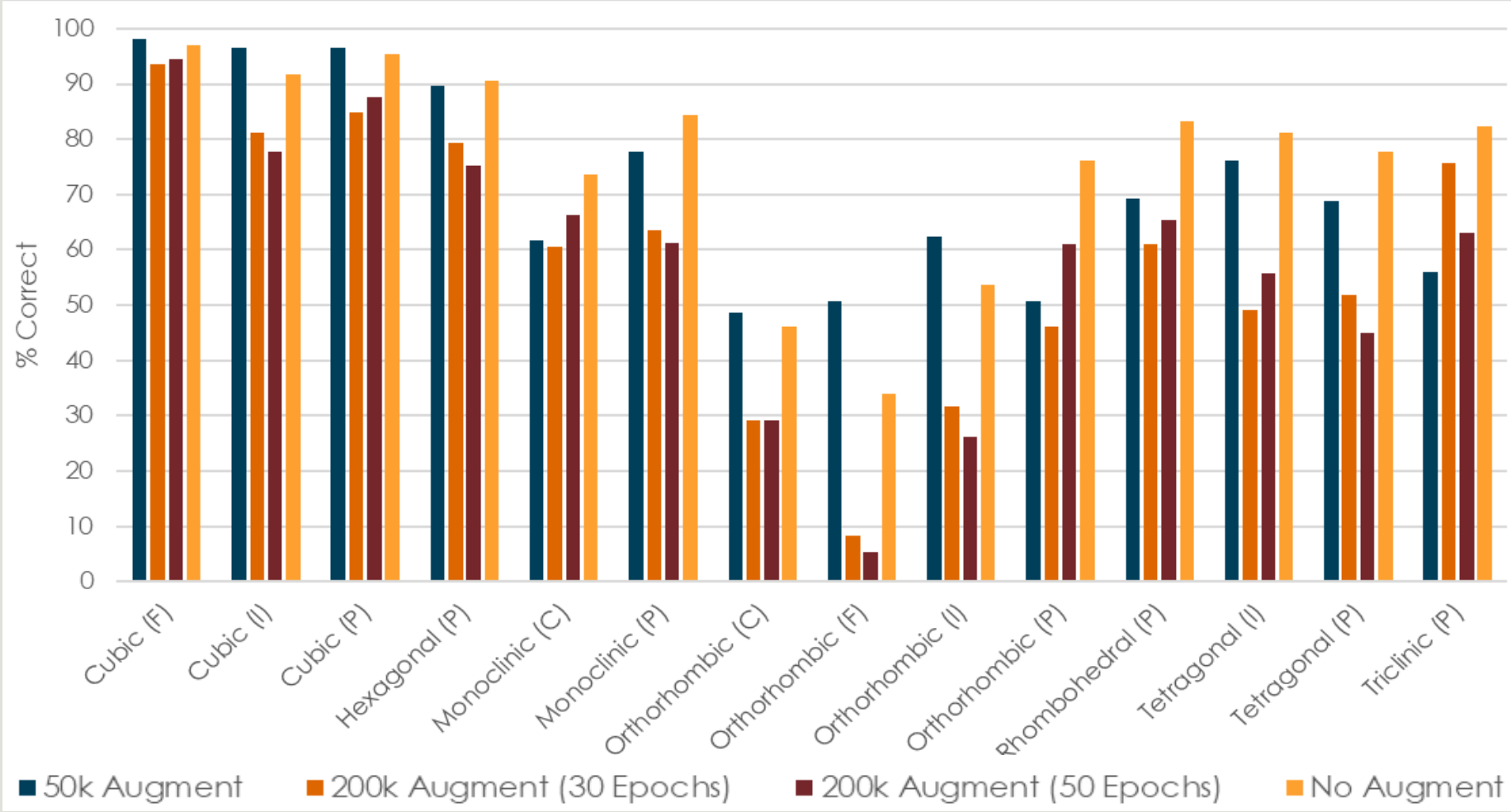


Data Augmentation





Data Augmentation to 50k Balanced



Conclusions and Next Steps

- De-Duping matters
- Our resolution testing didn't change much
- Adding more data helped, but it wasn't enough to overcome label imbalance
- Data augmentation is promising and strongly needed
- Moving to more complex cases

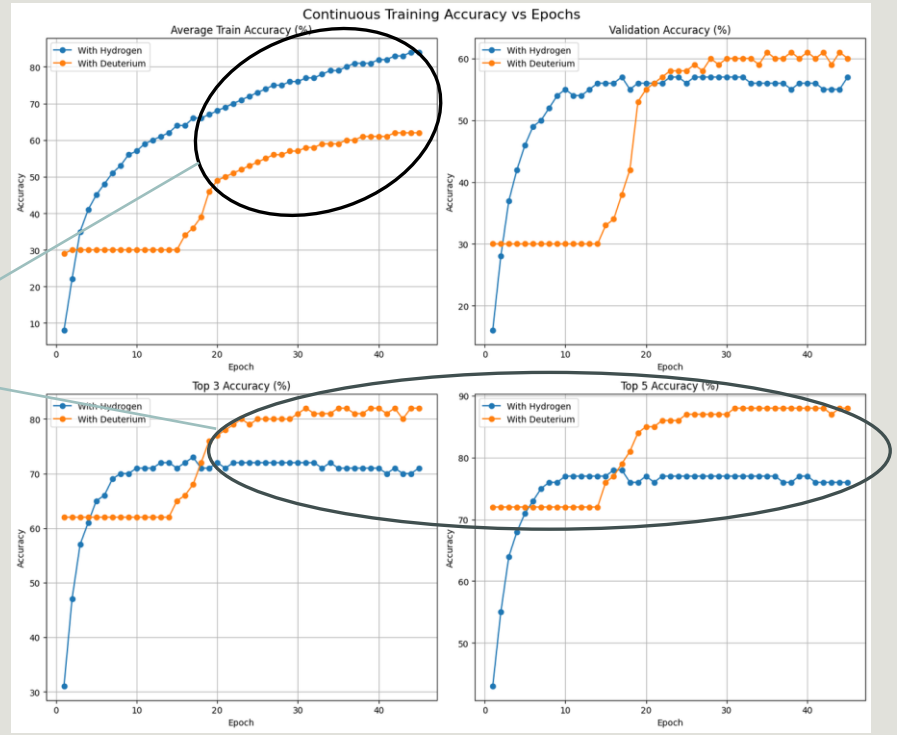
Acknowledgments

Collaborators:

- Elizabeth Baggett (Boston College, National Institute and Standards in Technology)
- William Ratcliff (National Institute of Standards and Technology)
- Paul Kienzle (National Institute of Standards and Technology)

Special thanks to:

- Julie Borchers, Leland Harriger
- National Institute of Standards and Technology
- Center of High-Resolution Neutron Scattering
- UTexas Advanced Computing Center

Swapping Hydrogen with Deuterium

Det. Could get better