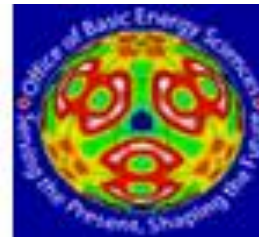


Machine learning materials science from experimental data

S.J.L. Billinge

Columbia University,

Brookhaven National Laboratory



A NEXT **GEN**ERATION
SYNTH**ESIS** CENTER



George Box: All models are wrong, some models are useful

Andy Millis: I want to be able to work on some classical problems

Columbia University in the City of New York

Brookhaven National Laboratory

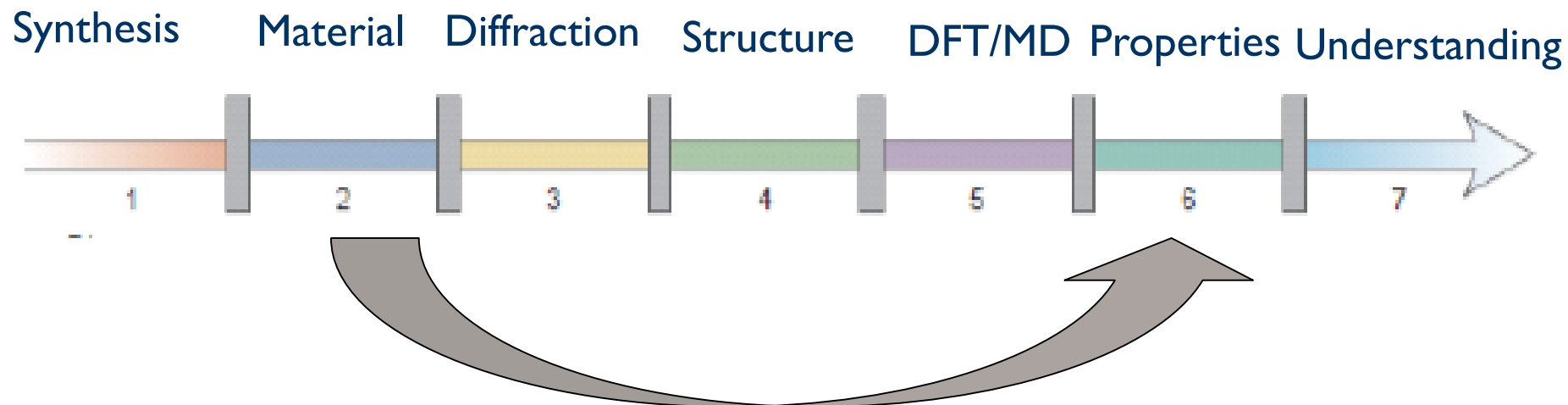


National Synchrotron Light Source-II
(NSLS-II) =>

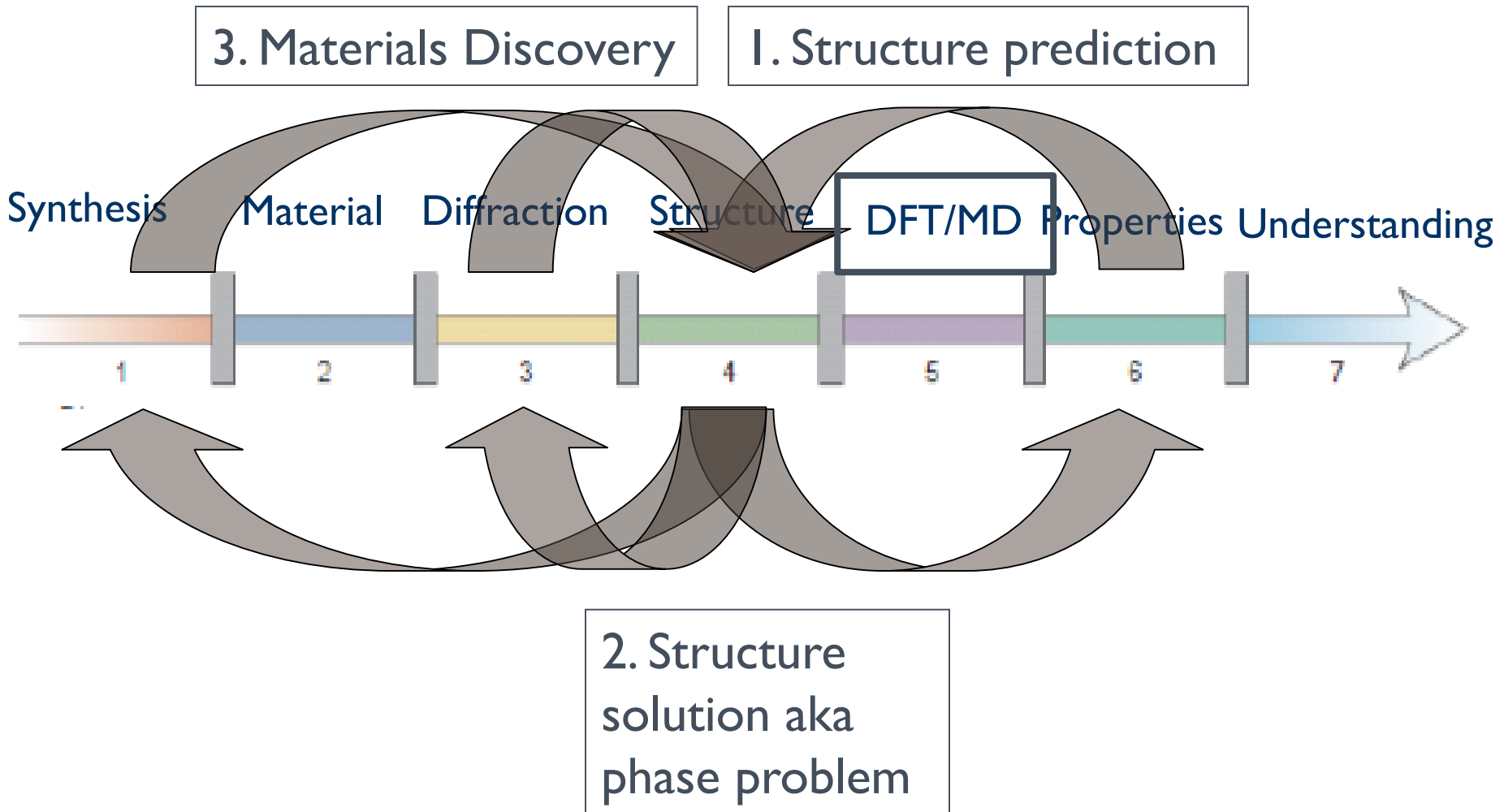
- XPD beamline
- Coherence
- Small beams
- High energy resolution
- Resonant scattering



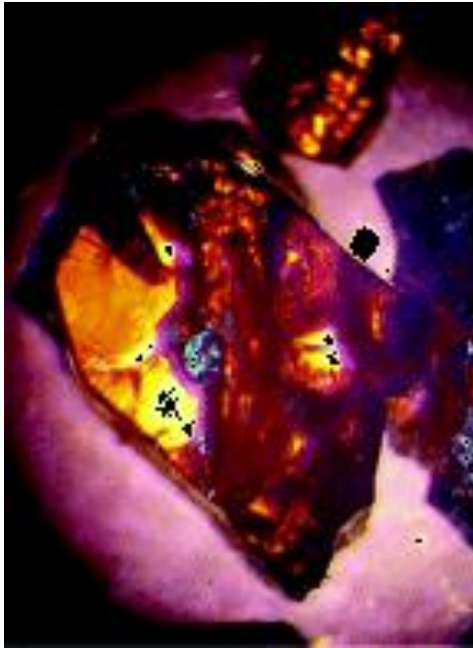
Materials Discovery



Materials Genomics



The Crystal Structure Problem



- Problem:
 - Here is a crystal, what is its structure?
- Solution:
 1. Give it to your grad student
 2. She puts it on the x-ray machine
 3. ...Pushes the button
 1. Machine tells you the structure
 2. Or Machine gets stuck
 1. Throw away the crystal
 2. Make it the subject of her thesis

Crystallography is largely a solved problem

From LiGaTe₂: A New Highly Nonlinear
Chalcopyrite Optical Crystal for the Mid-IR
L. Isaenko, et al., J. Crystal Growth, 5, 1325
– 1329 (2005)

Quantum Dot solar cells

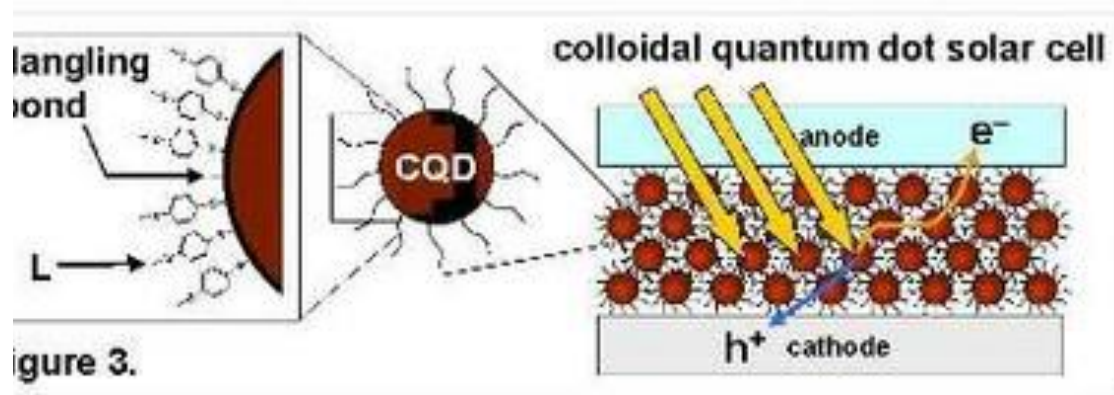
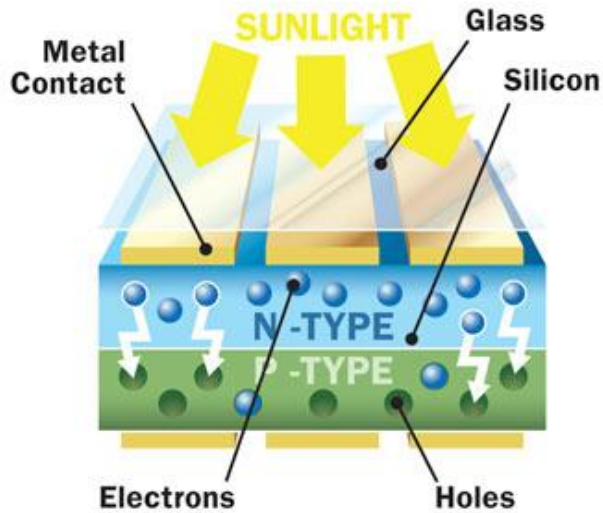
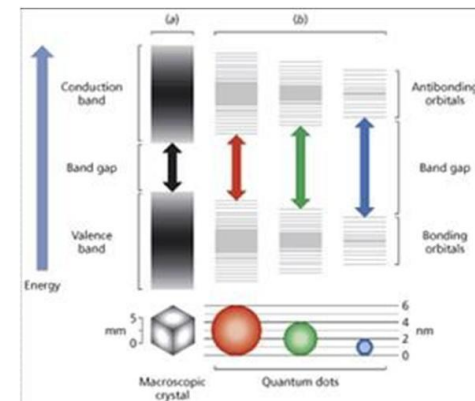
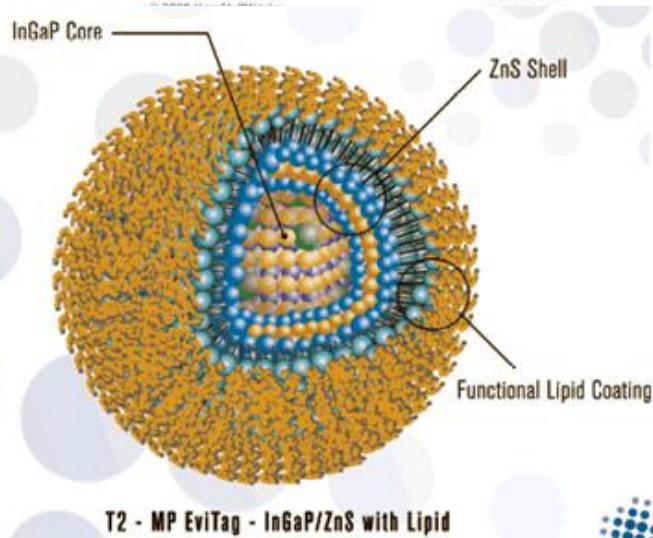
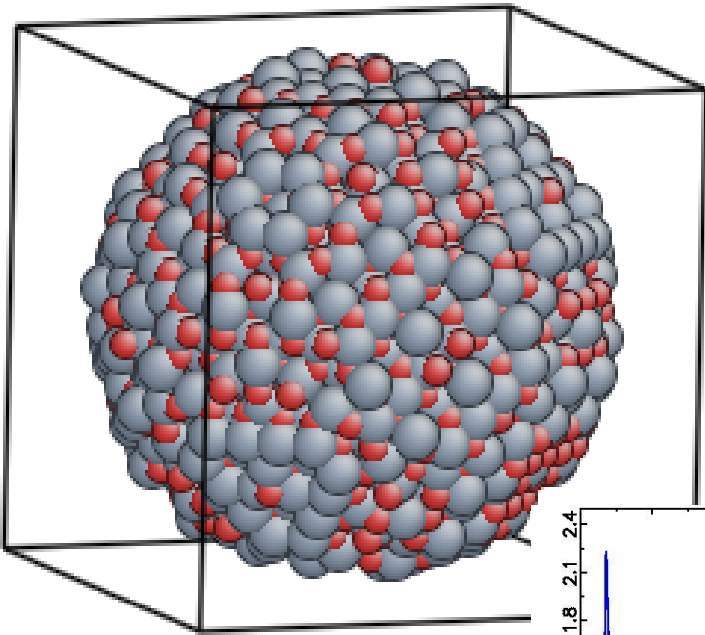


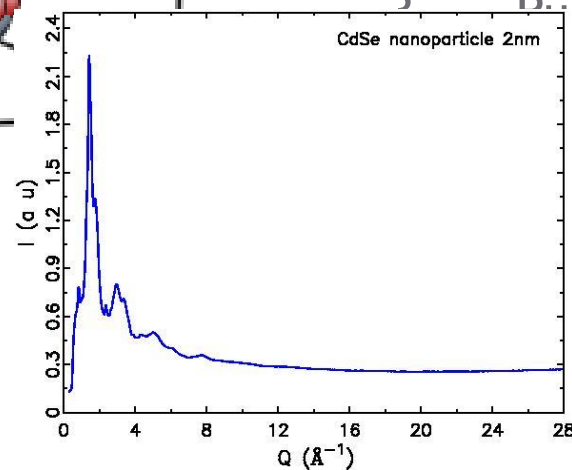
Figure 3.



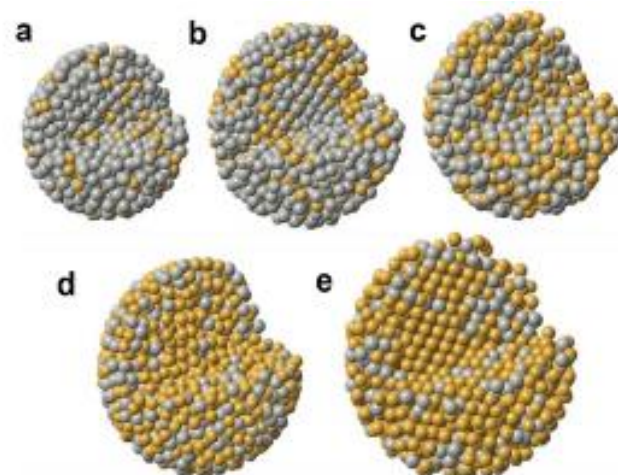
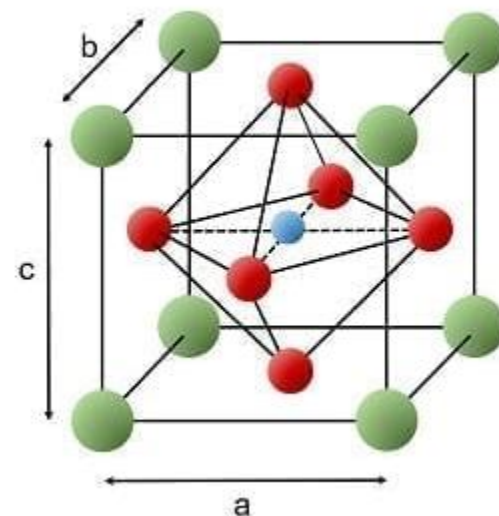
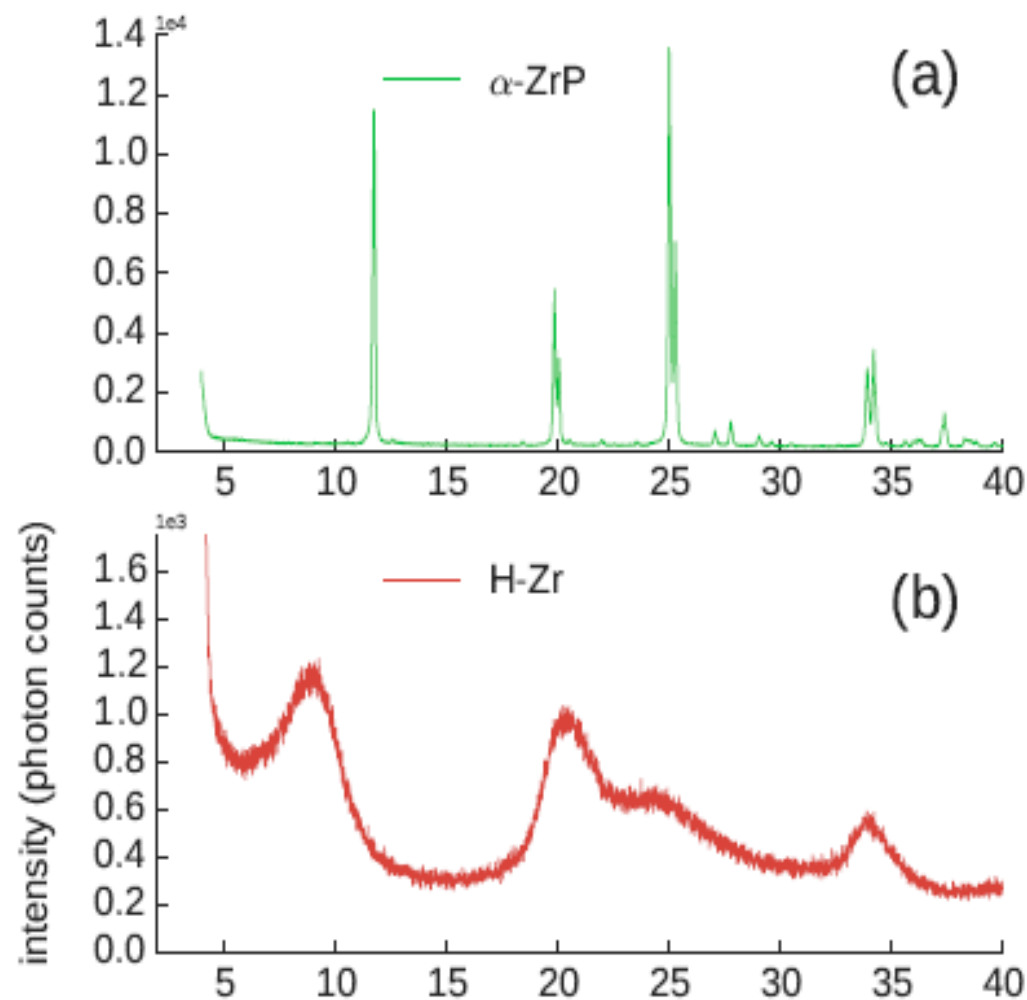
The Nanostructure Problem



- Problem:
 - Here is a nanoparticle, what is its structure?
- Solution:
 1. Give it to your grad student
 2. She puts it on the x-ray machine
 3. She pushes the button



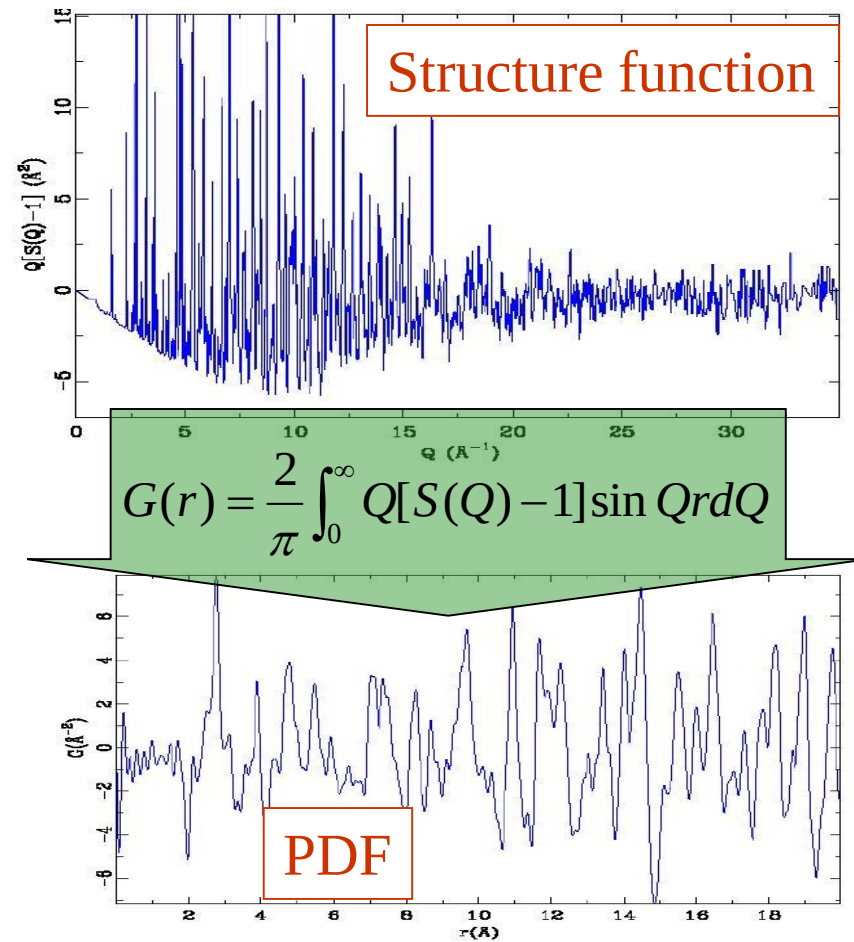
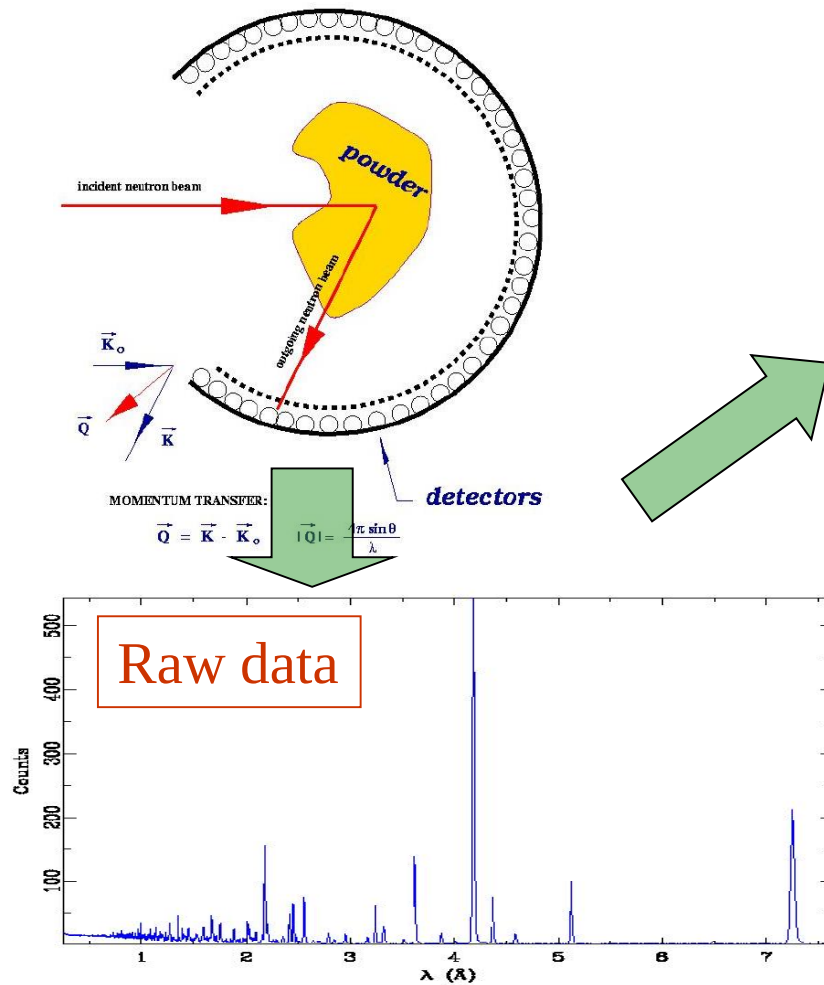
The Challenge with Real Material Structure Determination

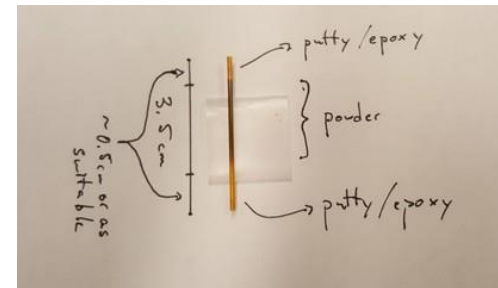
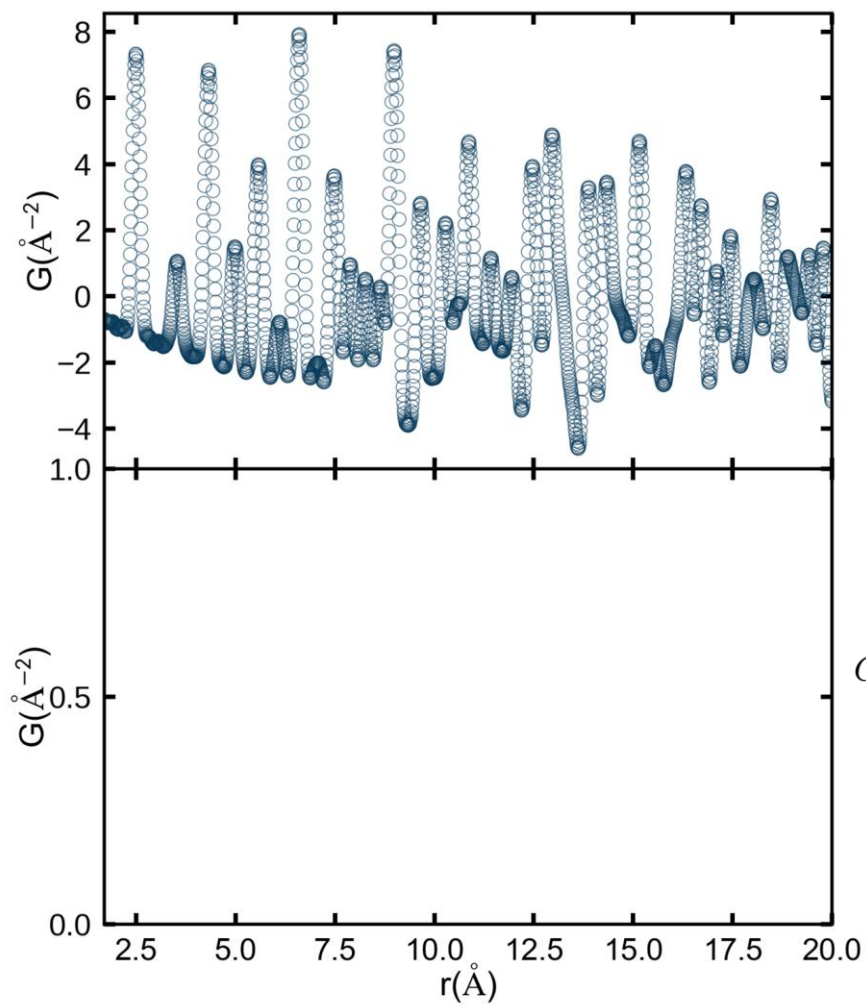


Nanostructure inverse problems

... are often ill-posed

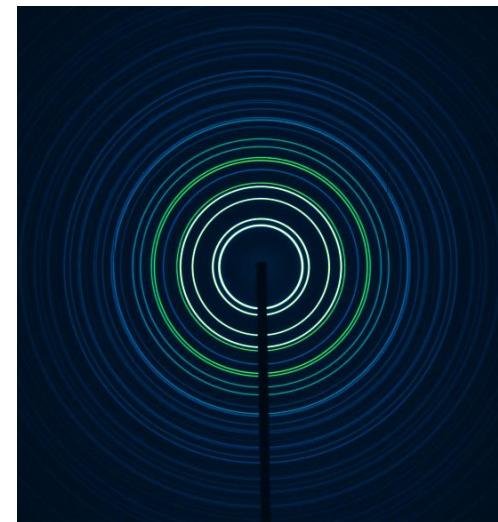
The atomic Pair Distribution Function

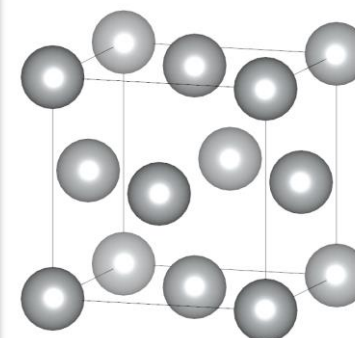
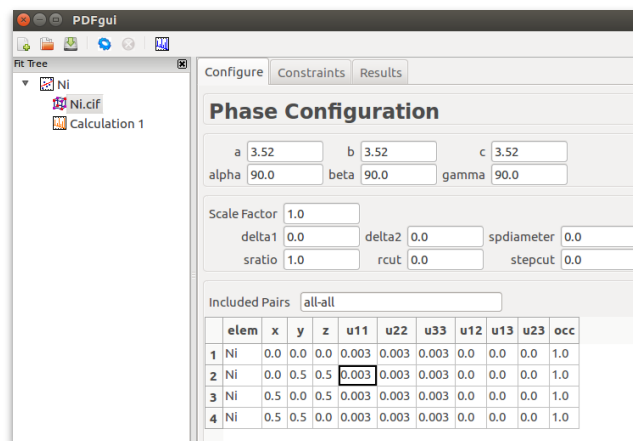
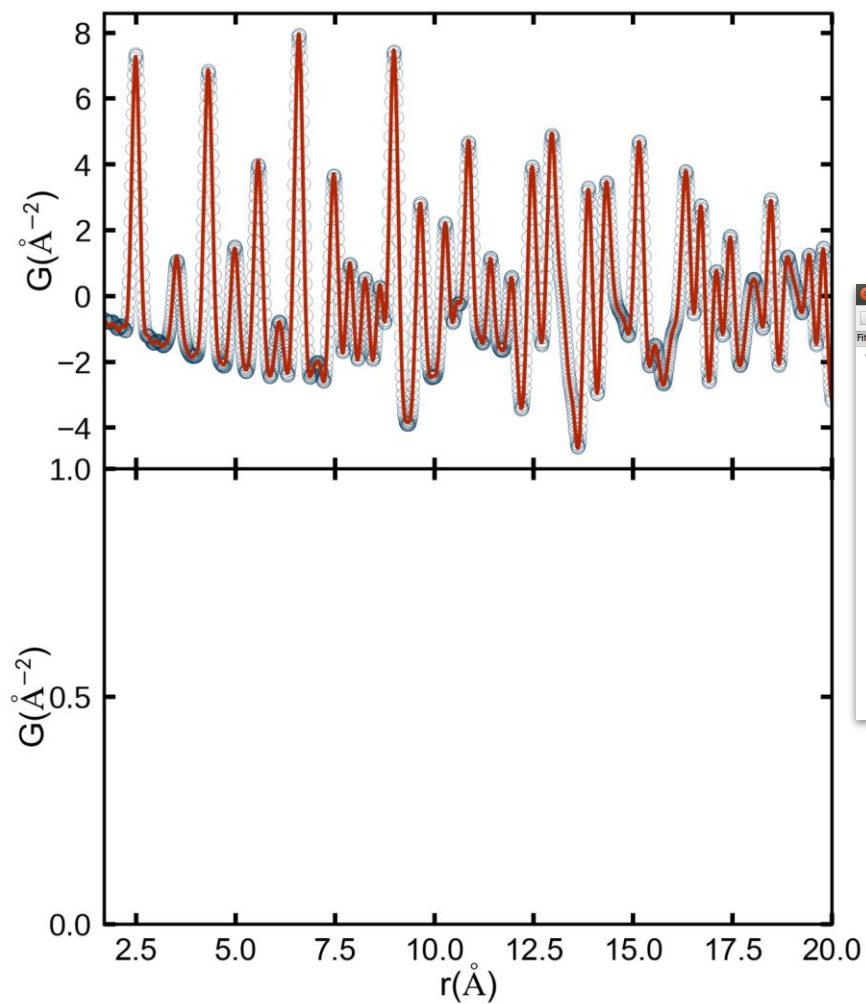




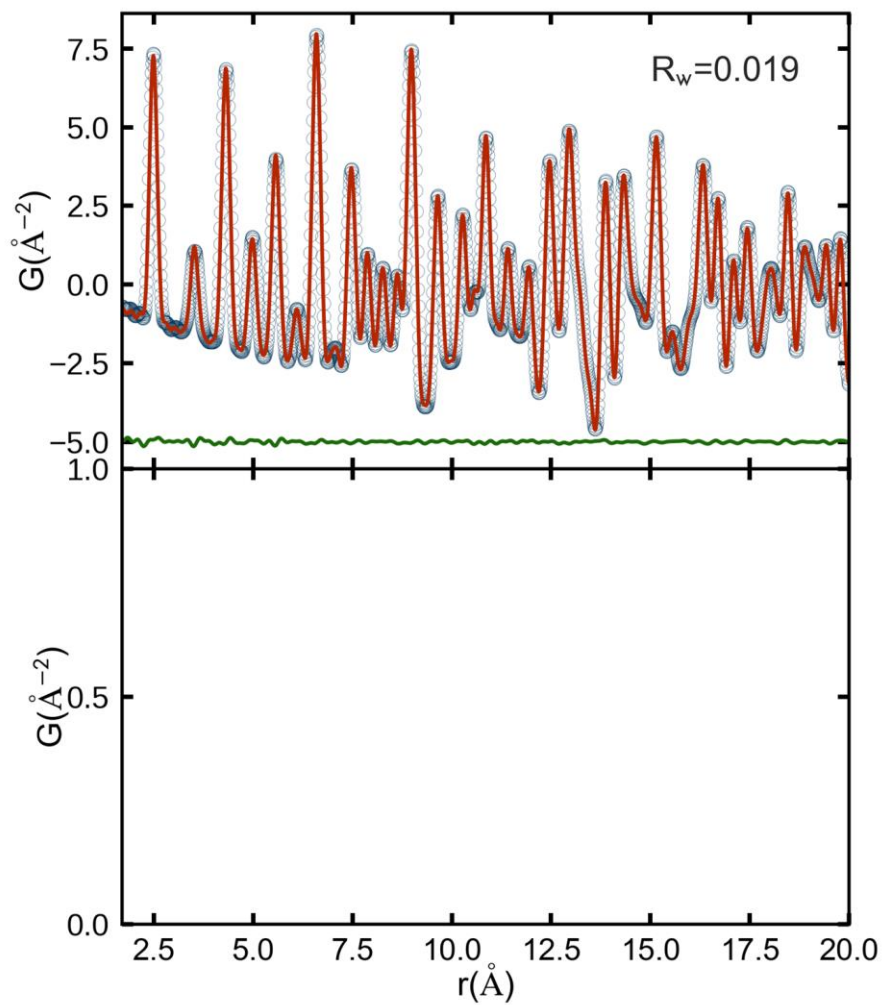
...

$$G(r) = \frac{2}{\pi} \int_{Q_{\min}}^{Q_{\max}} F(Q) \sin(Qr) dQ$$





$$G_{\text{calc}}(r) = \frac{1}{Nr} \sum_i \sum_{j \neq i} \left[\frac{b_i b_j}{\langle b \rangle^2} \delta(r - r_{ij}) \right] - 4\pi r \rho_0$$



$$R_w = \sqrt{\frac{\sum_{i=1}^n [G_{obs}(r_i) - G_{calc}(r_i, P)]^2}{\sum_{i=1}^n G_{obs}(r_i)^2}}$$

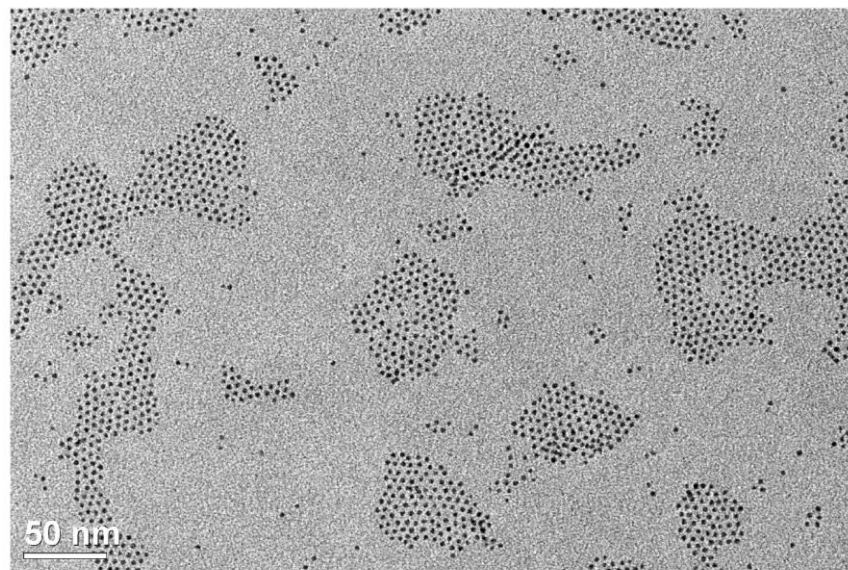
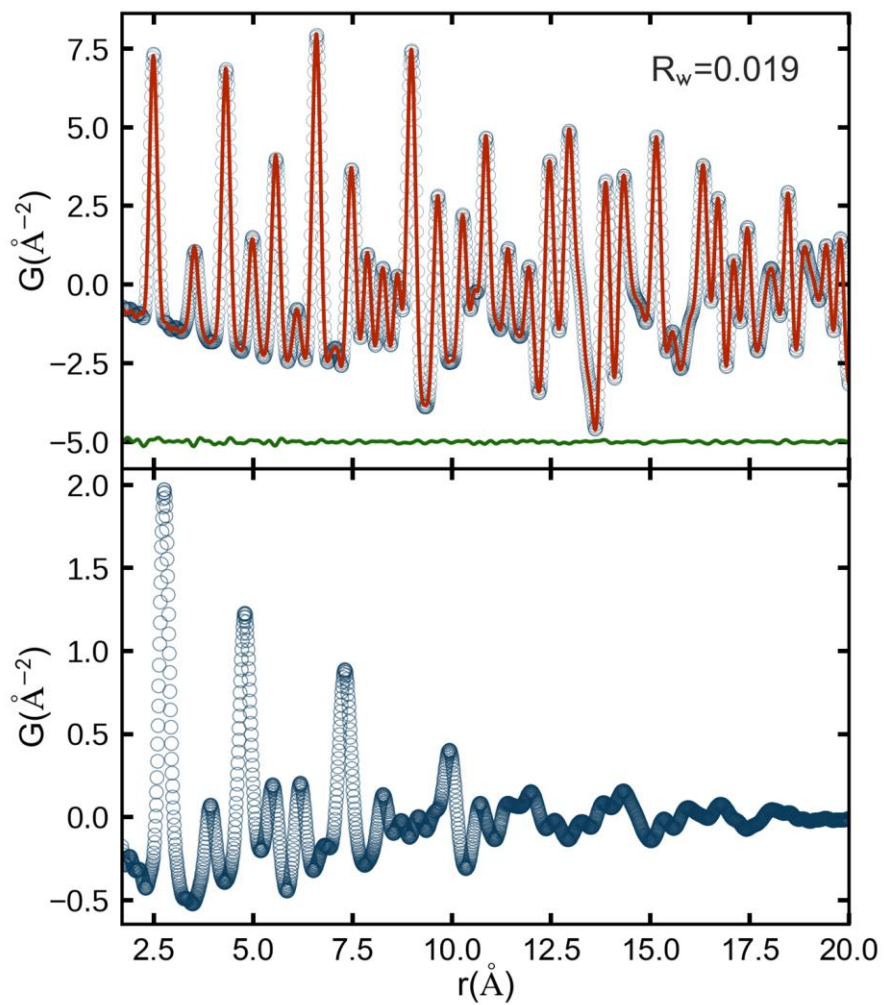
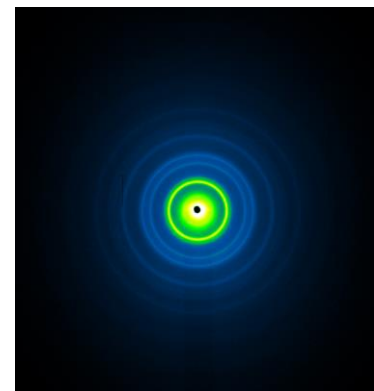
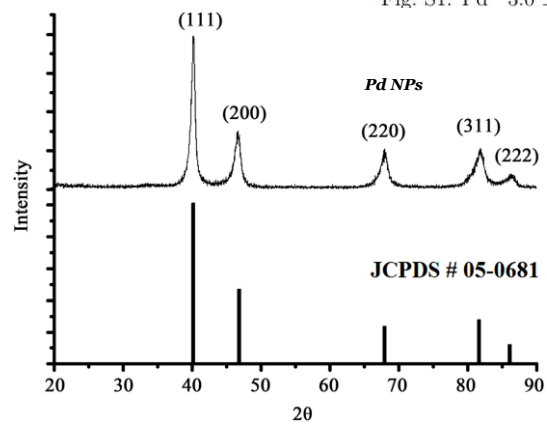
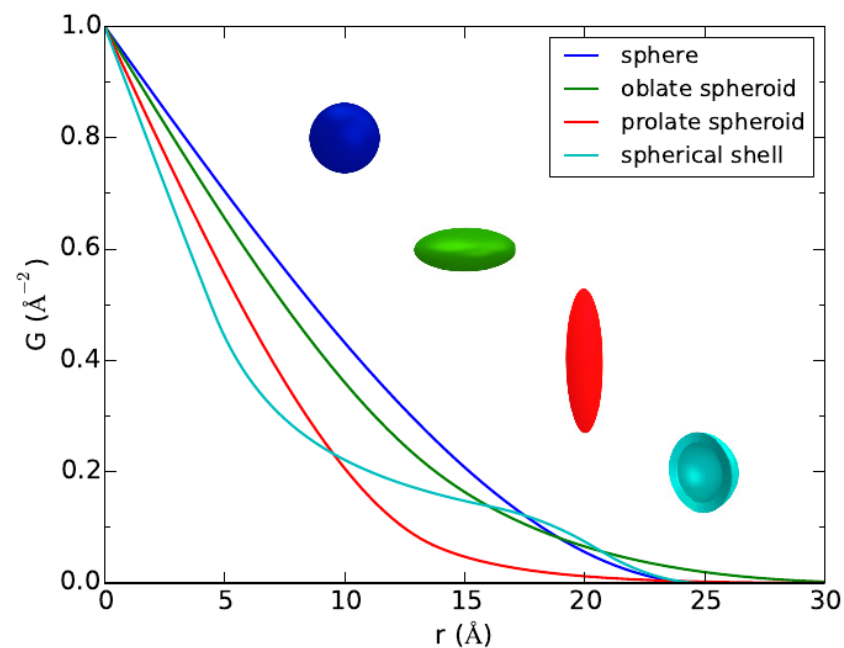
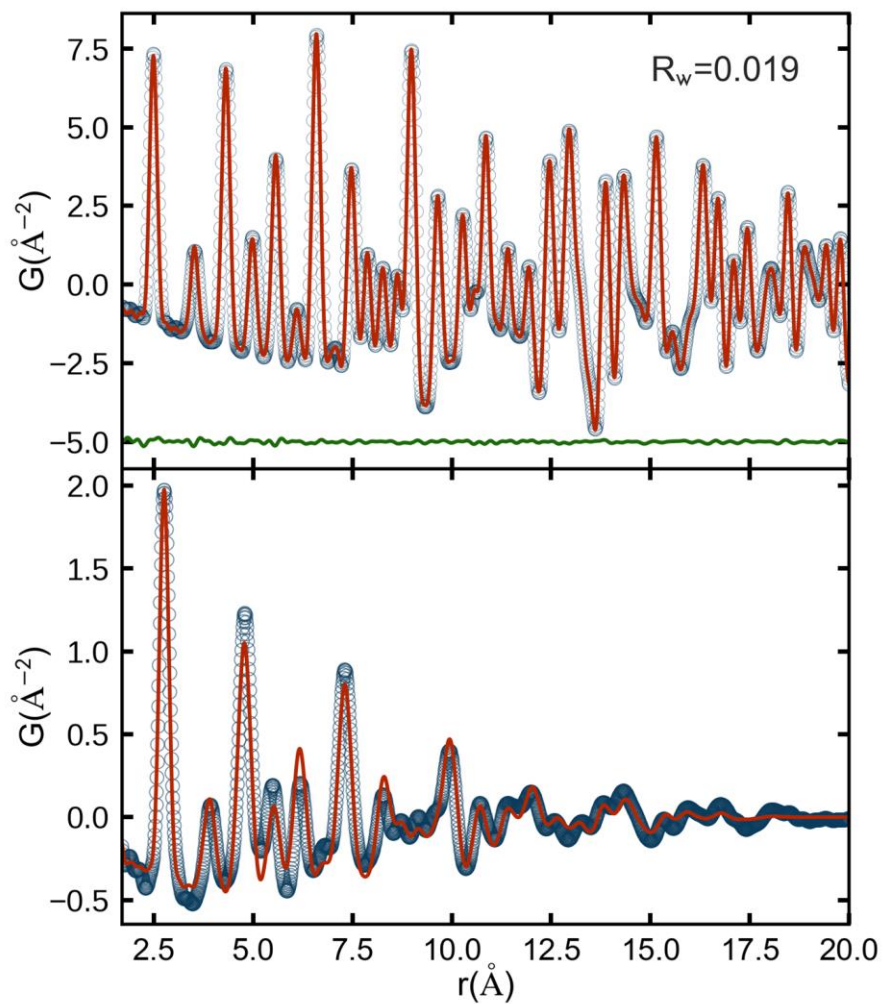


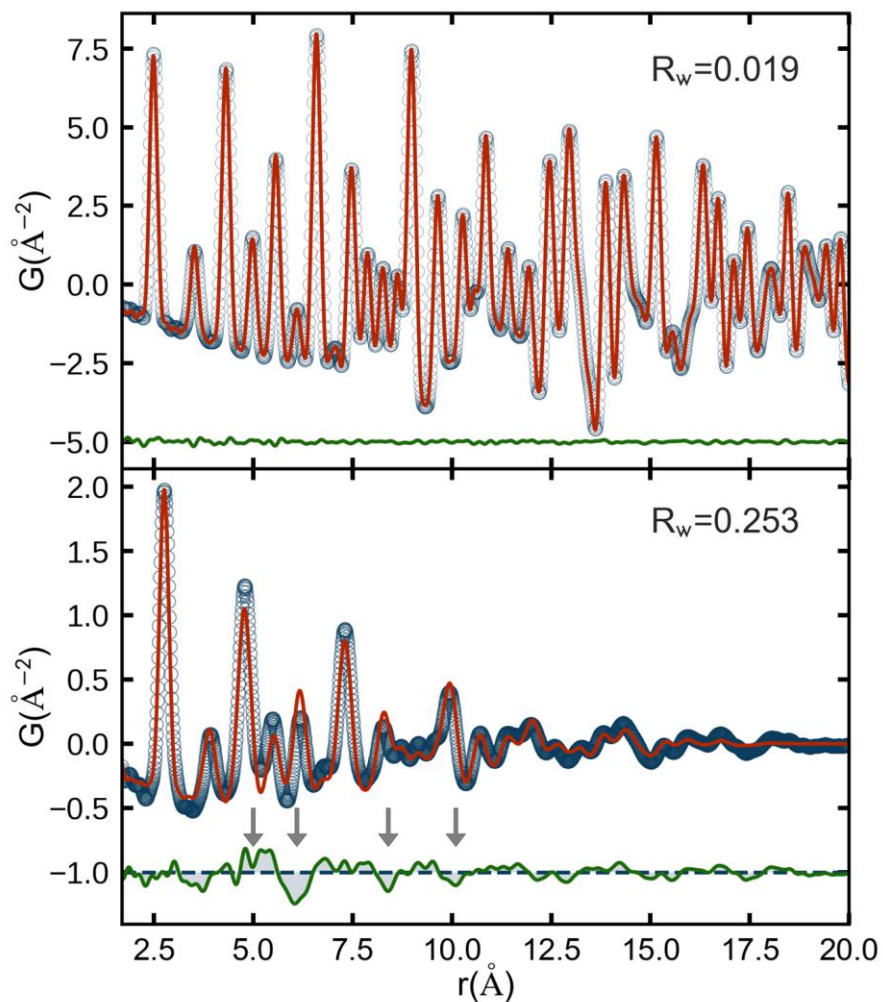
Fig. S1: Pd^P 3.0 ± 0.3 nm





$$G(r, d)_s = G(r) f(r, d)$$

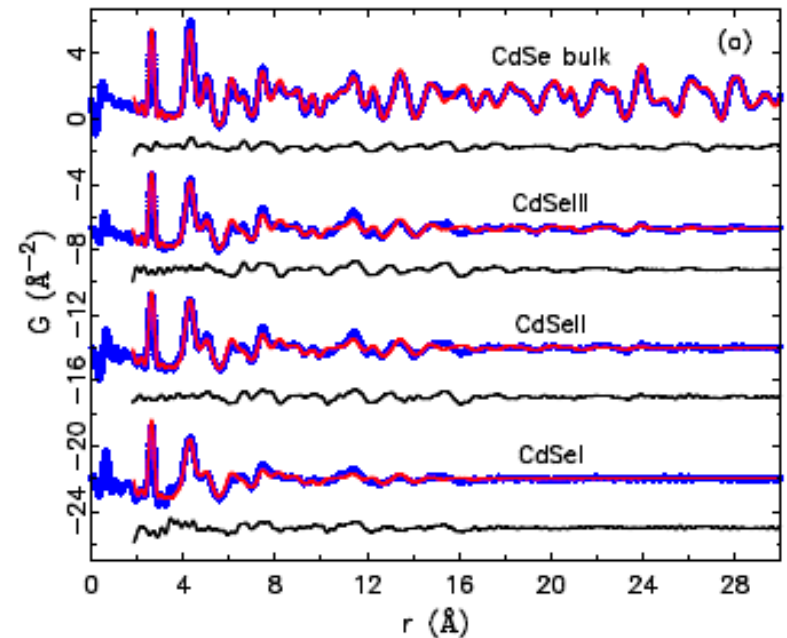
$$f(r, d) = \left[1 - \frac{3r}{2d} + \frac{1}{2} \left(\frac{r}{d} \right)^3 \right] \Theta(d - r)$$



Banerjee, SJLB, et al. “Improved models for metallic nanoparticle cores from atomic pair distribution function (PDF) analysis”. In: J. Phys. Chem. C 122.51 (2018), pp. 29498–29506. doi: 10.1021/acs.jpcc.8b05897.

AC fits to nanoparticle data are goodbut not that good!

Here it may be noted that the agreement factors achieved with the PDF fits based on lattice-type models are usually in the range of 15–20% as is the case with the present fits. Such R_w values appear somewhat high when compared to goodness-of-fit factors resulted from Rietveld refinement of XRD data in reciprocal space. This mostly reflects the fact that the atomic PDFs being fit take both the sharp, Bragg like features and the diffuse component of the XRD data into account while Rietveld fits consider only the former. The inherently higher absolute value of the goodness-of-fit factors in PDFs fits, however, does not affect their functional purpose as a residuals quantity that guides the refinement of the structural parameters of the tested Bravais lattices based models.



- Excuses excuses excuses!
- Can we do better?

Name	Composition	D (nm)	Surfact.	Beamline	Morph.
Pd ^P	Pd	3.0 ^a	OAm	XPD	P
CoPd ^P	Co _{0.2} Pd _{0.8}	8.6 ^a	OAm, TOP	XPD	P
PdNi ^W	Pd _{0.83} Ni _{0.17}	2.3 ^a	ODA	X17A	W
Ag ^P	Ag	38.0 ^b	PVP	XPD	P
AgAu ^P	Ag _{0.5} Au _{0.5}	5.0 ^b	PVP	XPD	P
PdAu ^W	Pd _{0.9} Au _{0.1}	2.1 ^a	ODA	X17A	W
Pd ^W	Pd	2.0 ^a	ODA	X17A	W
Au ^P	Au	9.0 ^b	PVP	XPD	P
PtRu ^{P,C}	Pt _{0.5} Ru _{0.5}	3.5 ^a	PFSA	XPD	P,C
Pt ^P	Pt	3.0 ^a	OAm, TOP	XPD	P
Au ^{Cl1}	Au	1-2	p-MBA	ID11	Cl1
Au ^{Cl2}	Au	1-2	SC6	11IDB	Cl2
Ni ^B	Ni	Bulk	None	XPD	B

^a Average size by transmission electron microscopy (TEM)

^b Average size by differential centrifugal sedimentation (DCS)

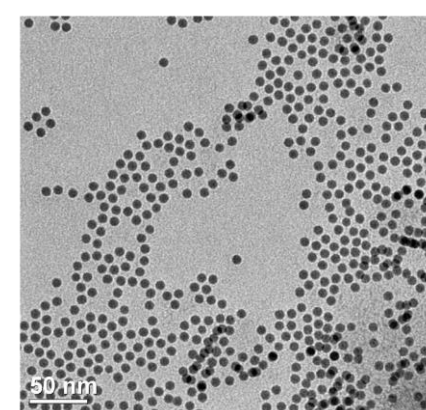
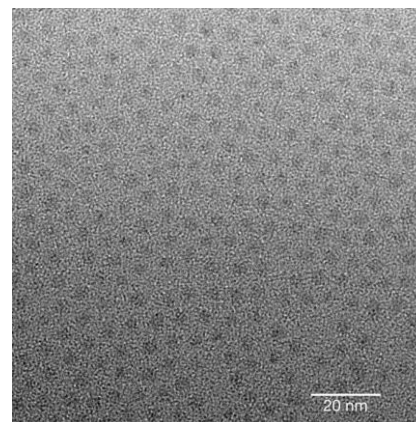
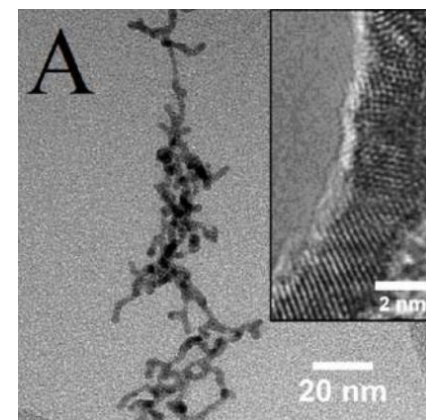
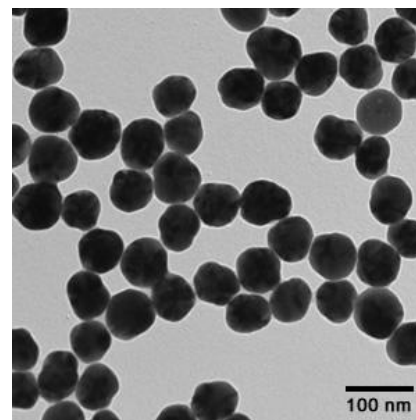
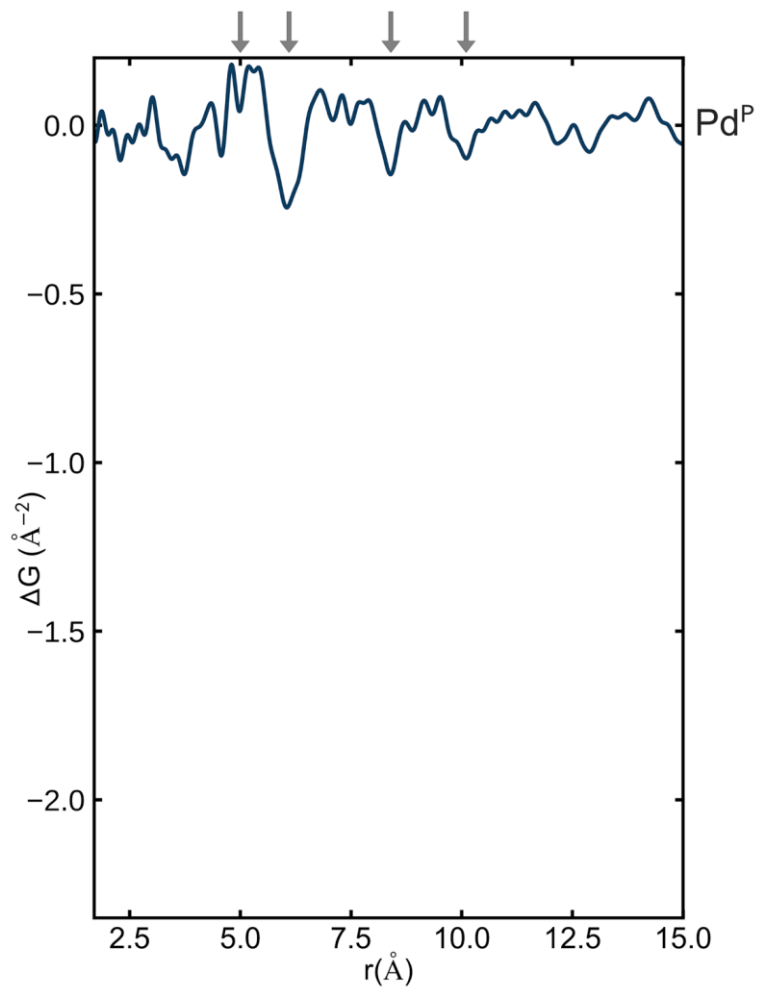
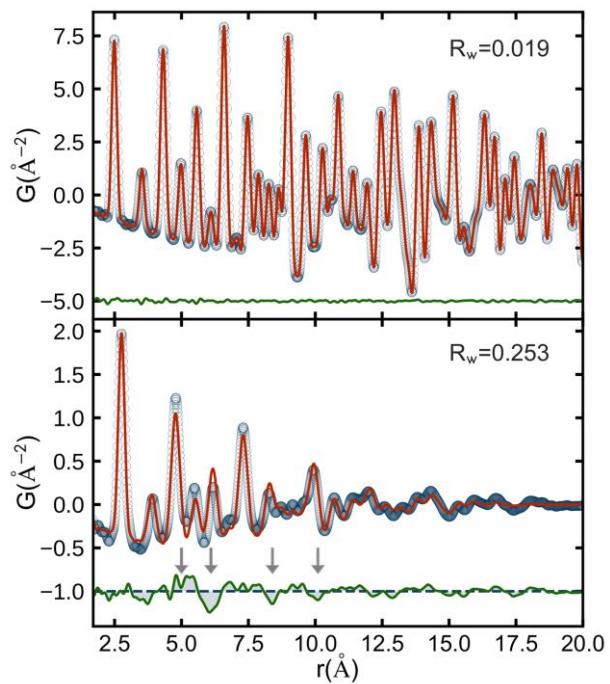
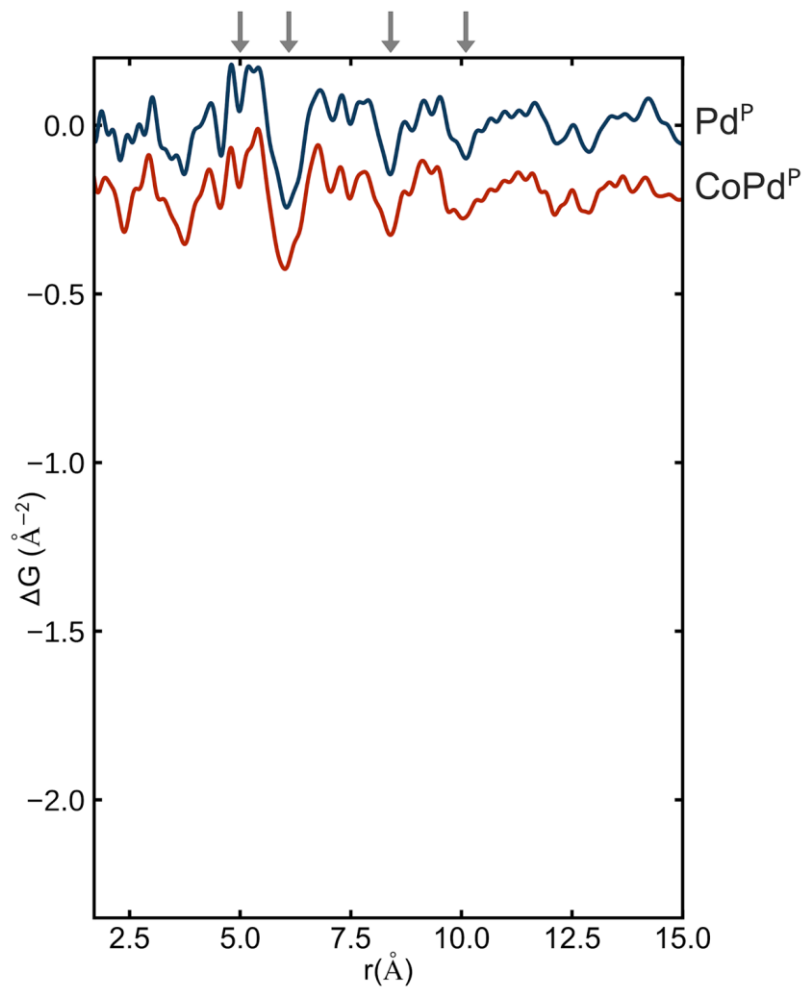
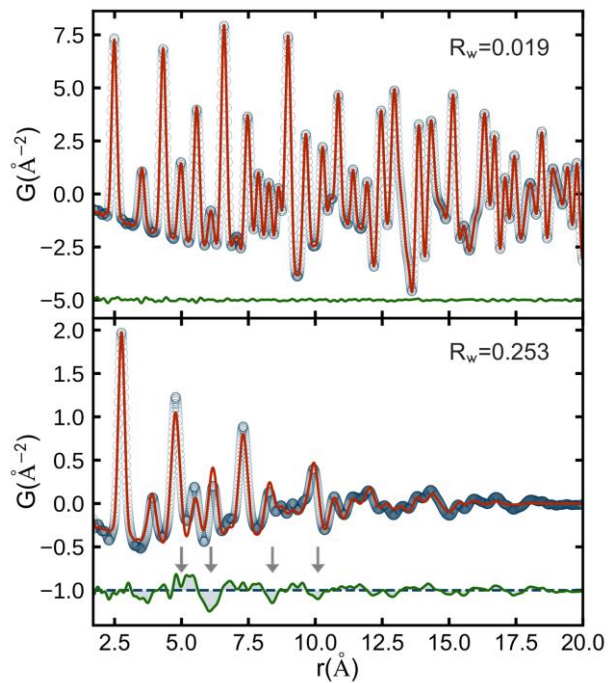


Fig. S2: CoPd^P 8.6 ± 1.7 nm

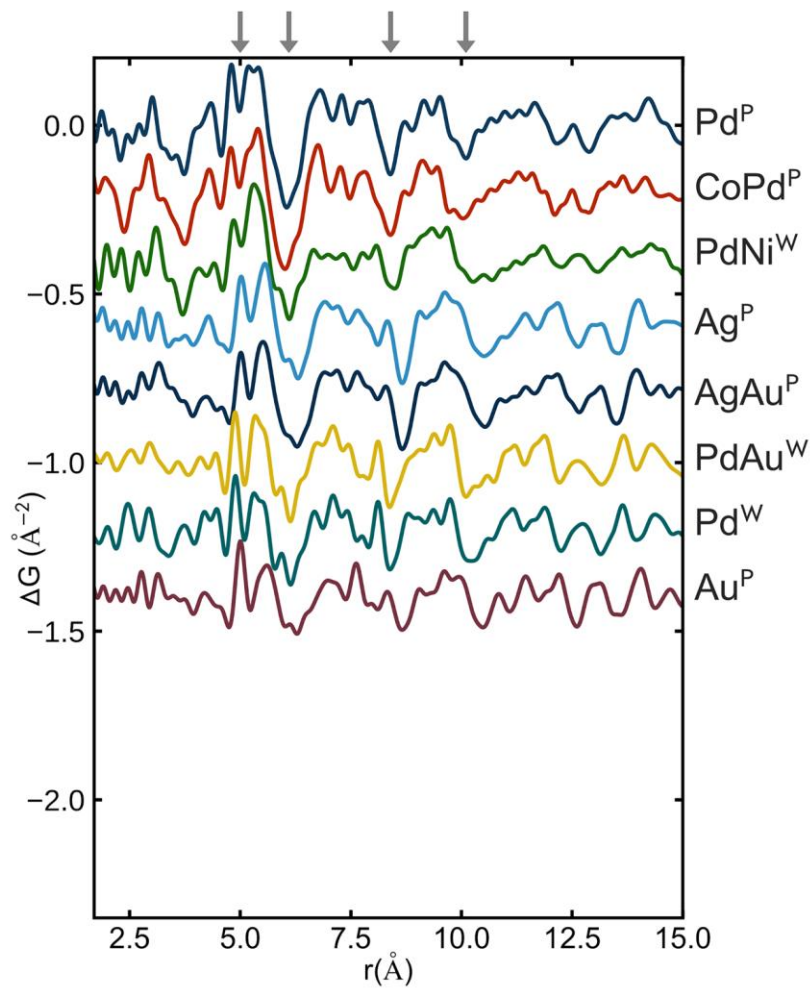
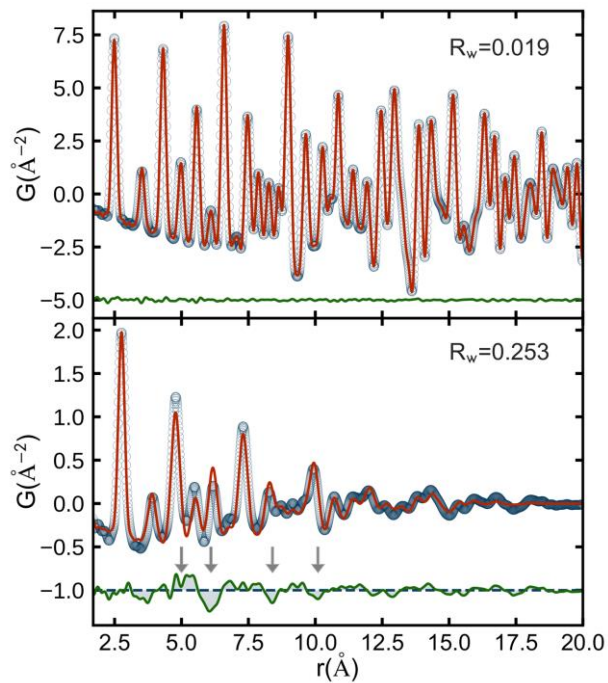
Residuals



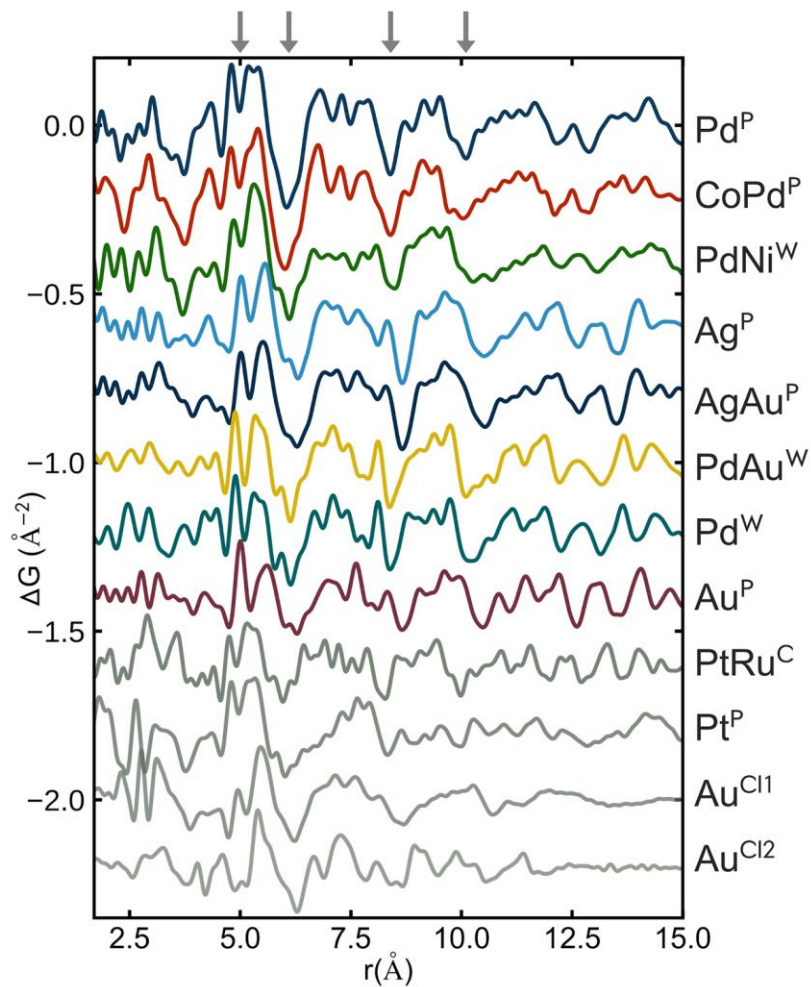
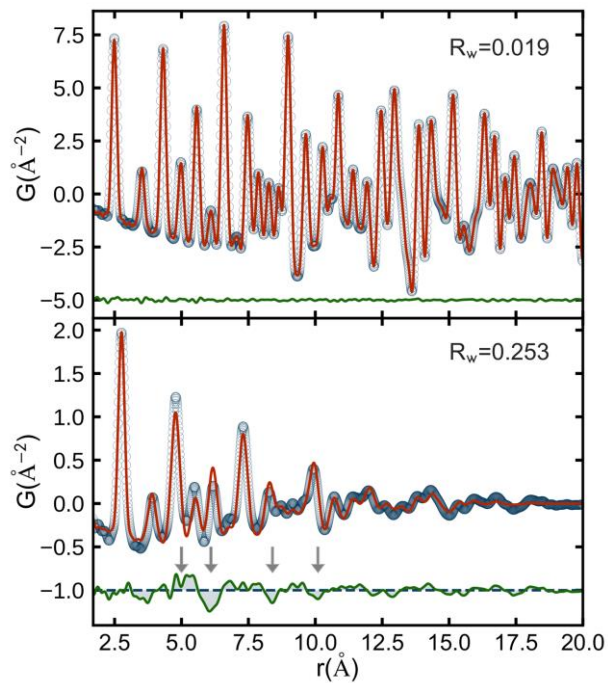
Residuals



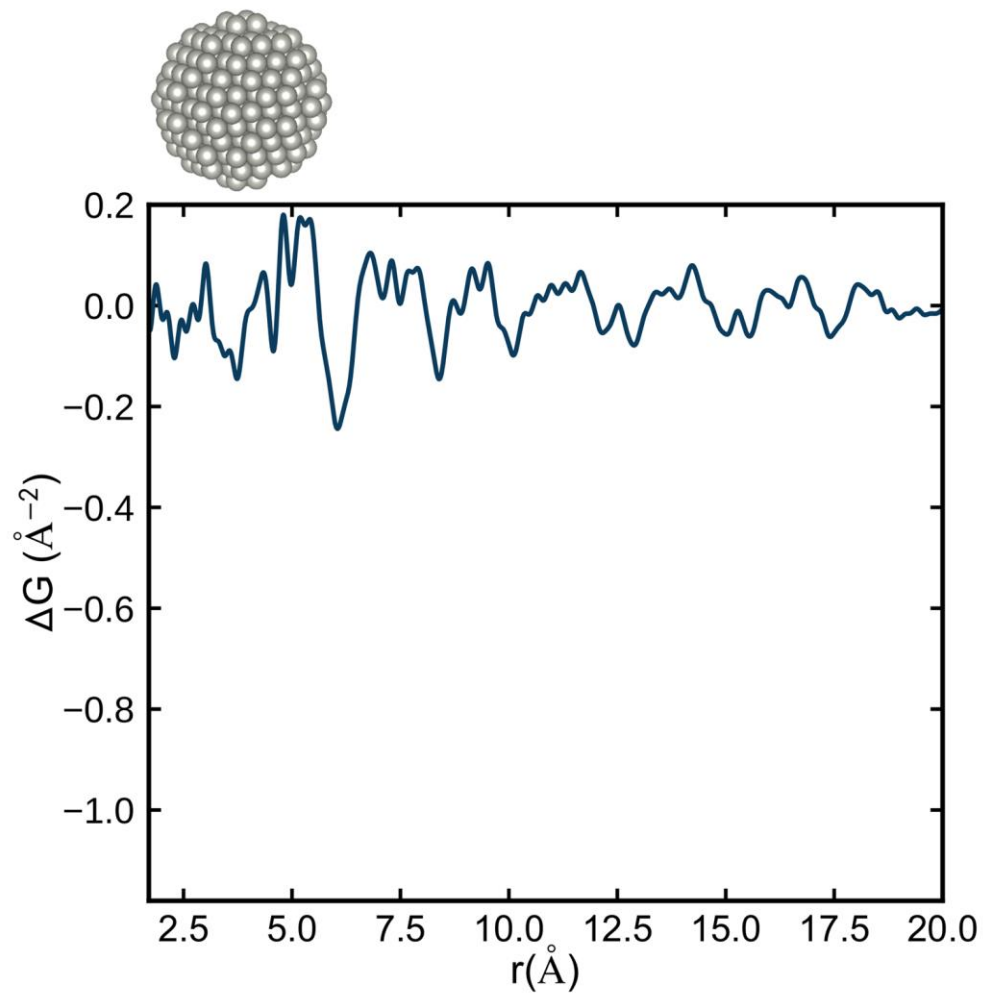
Residuals



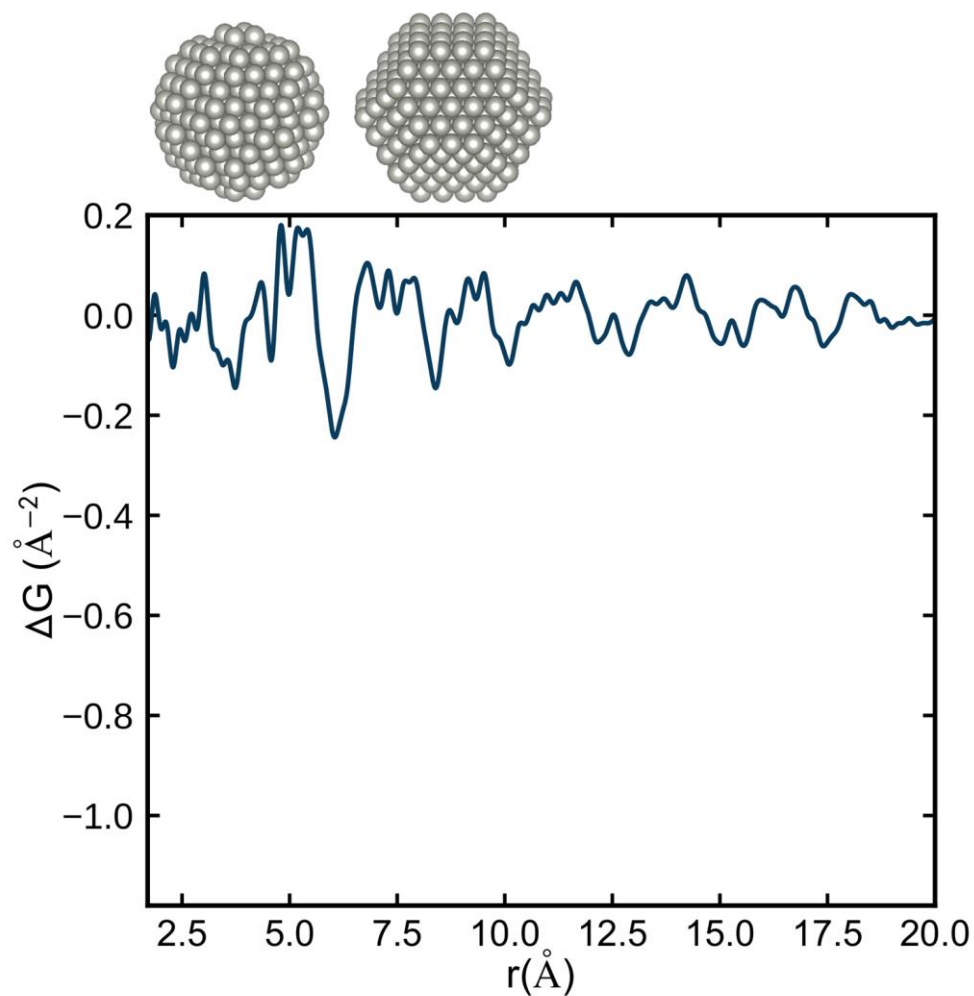
Residuals



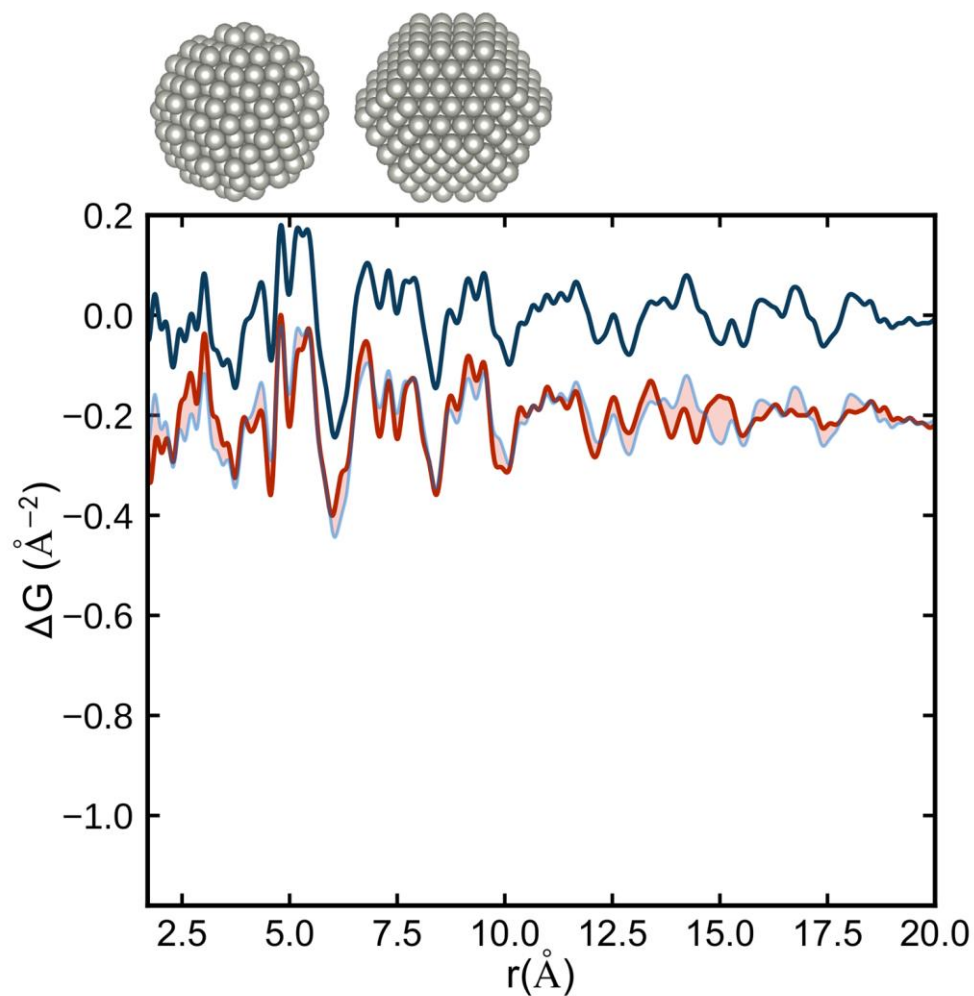
Testing discrete models



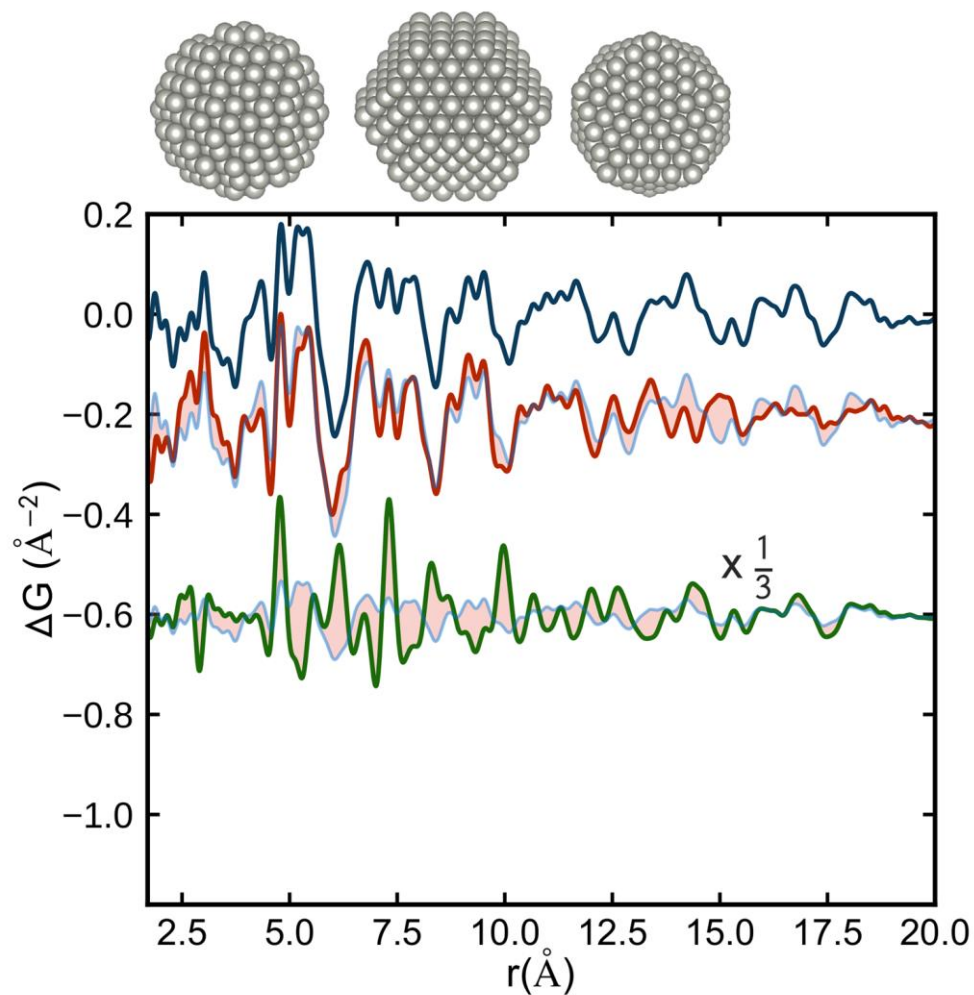
Testing discrete models



Testing discrete models

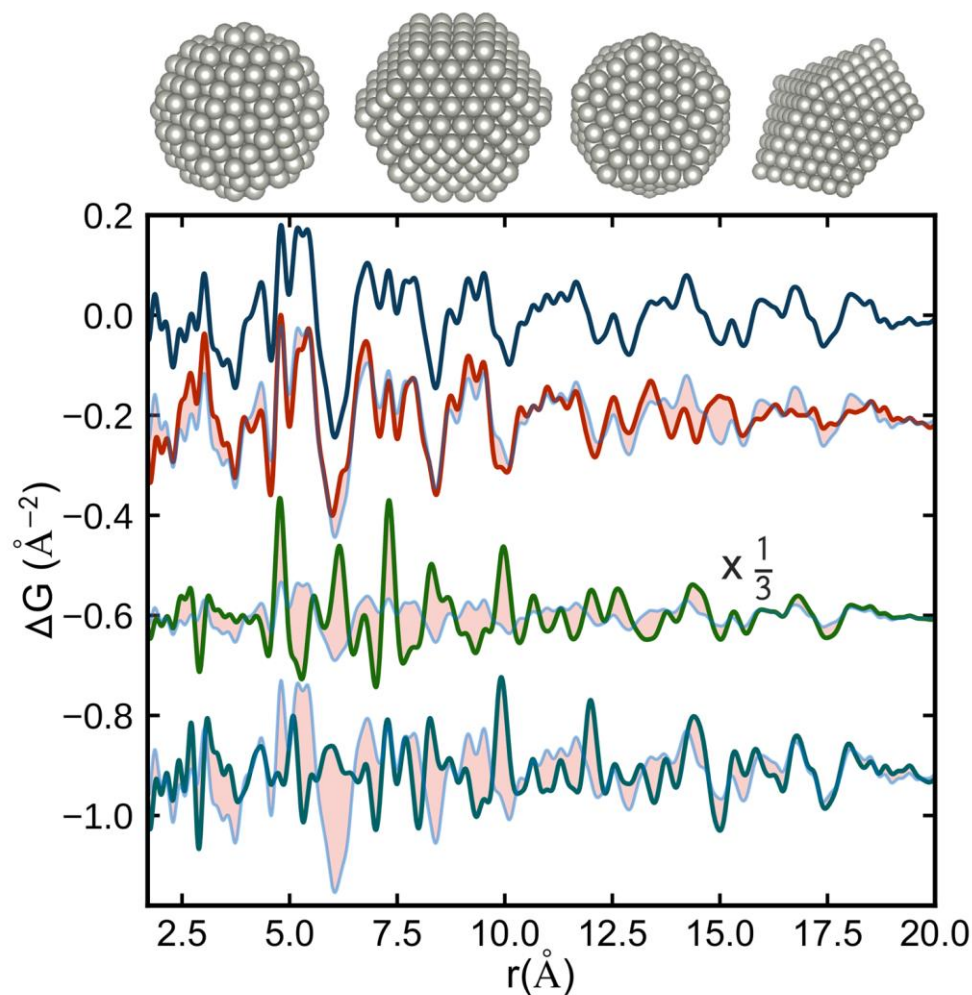


Testing discrete models



Testing discrete models

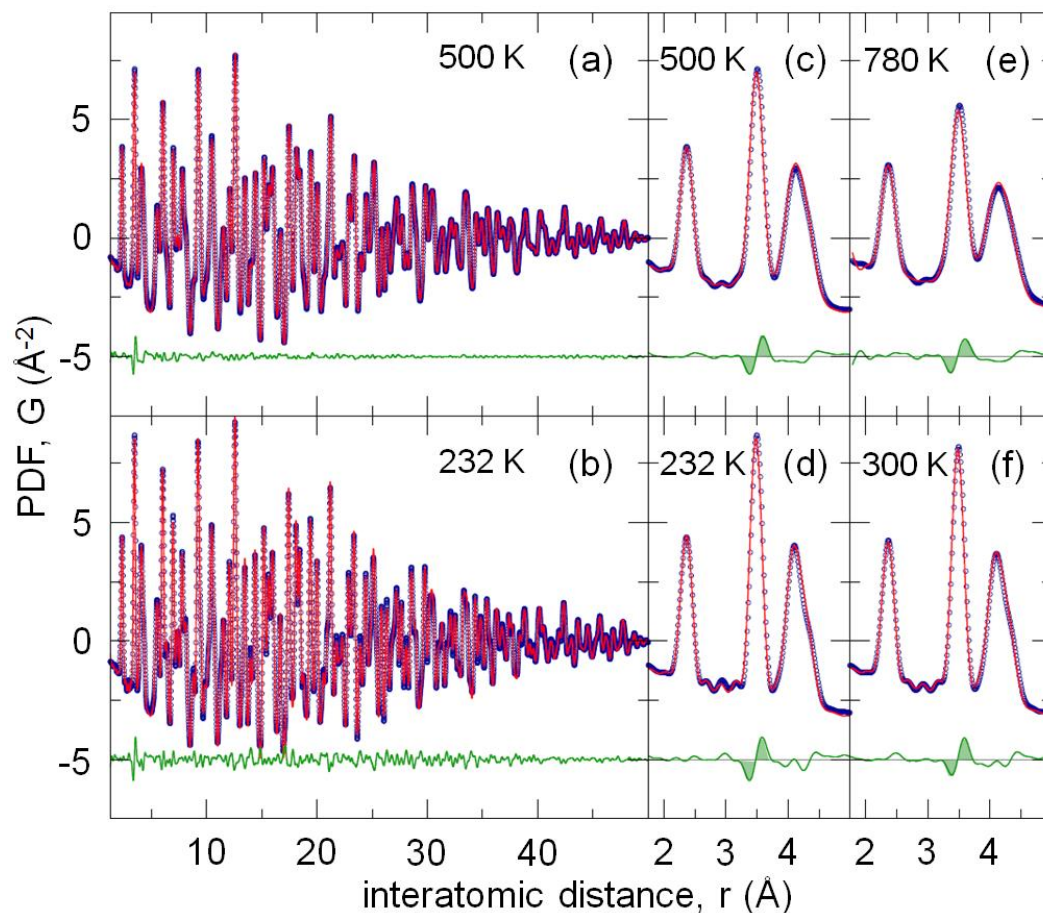
Banerjee, SJLB, et al. “Improved models for metallic nanoparticle cores from atomic pair distribution function (PDF) analysis”. In: J. Phys. Chem. C 122.51 (2018), pp. 29498–29506. doi: 10.1021/acs.jpcc.8b05897.



Automating cluster finding

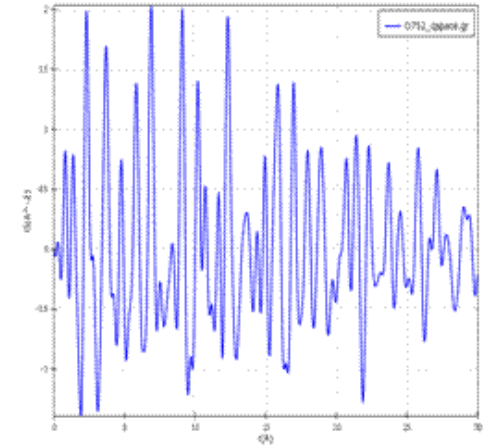
Standard approach

- I PDF
- I model
- 20 parameters
- Vary the parameters until the model agrees as well as possible with the data.
- Emphasis on parameter estimation, not on model selection
- Challenge is finding the right model

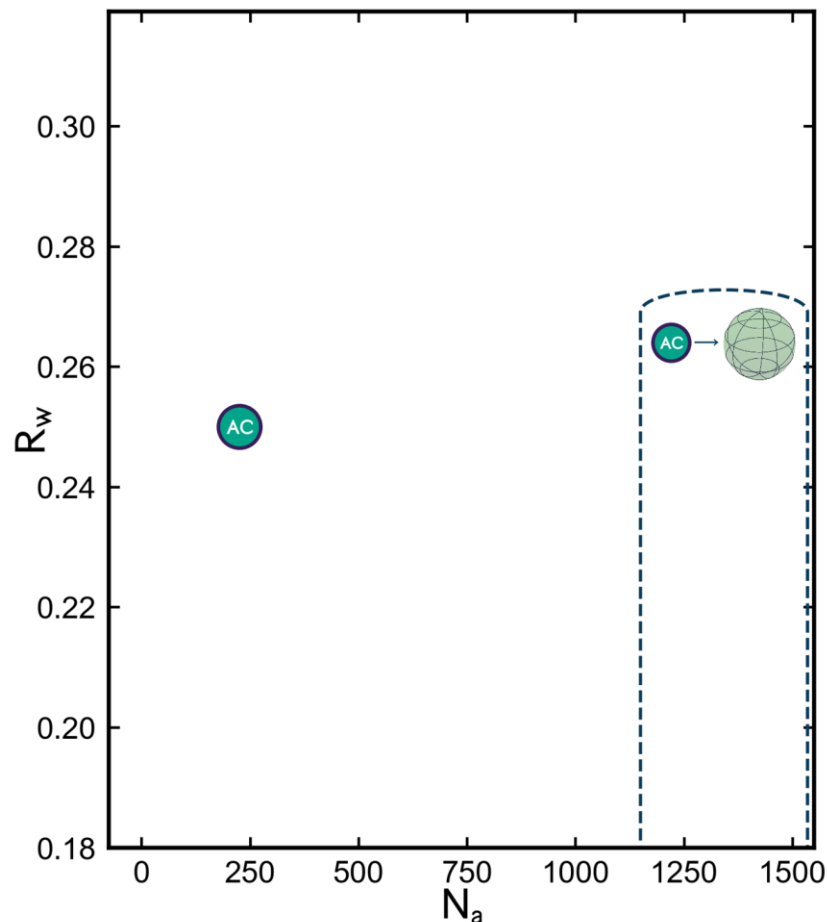


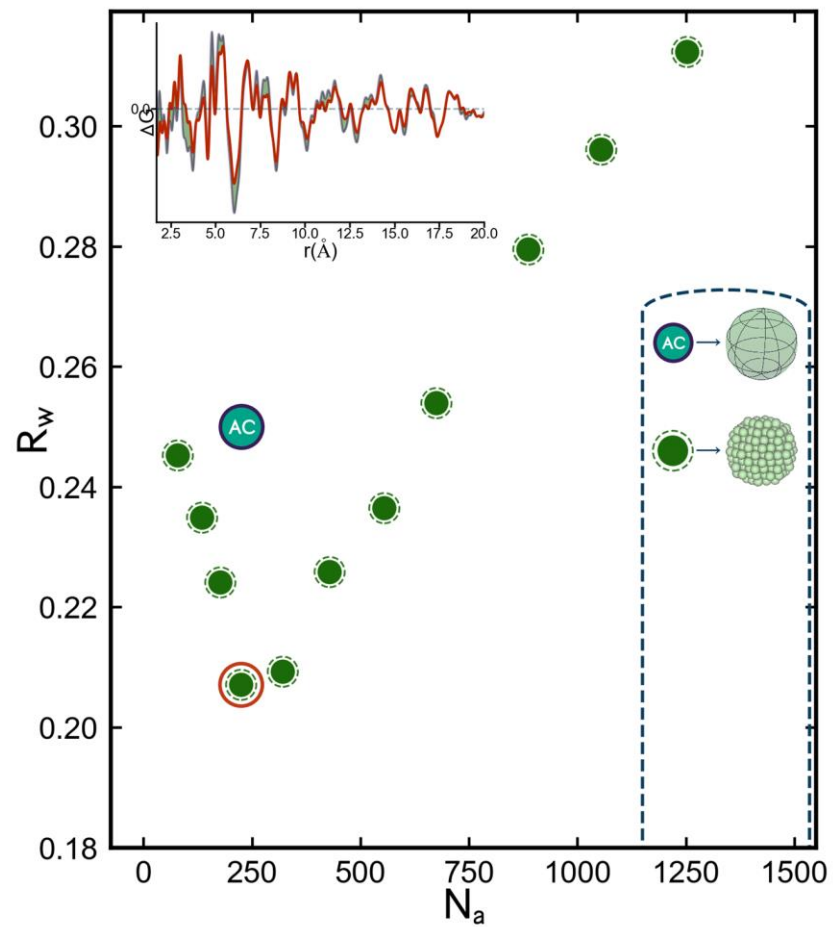
New approach: Cluster-Mining

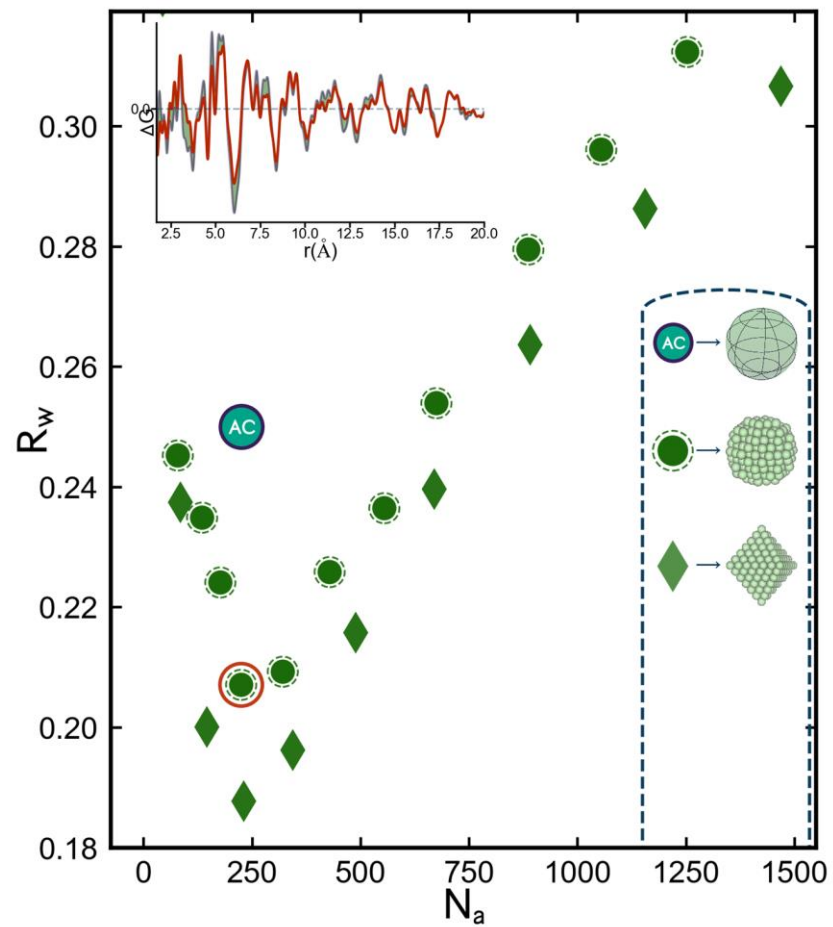
- 1 PDF
- many models
- few parameters
- Iterate over large numbers of models
- Emphasis on model selection
- Advantage: find multiple nearby models!
- Challenge: structure must be in the structure-mine

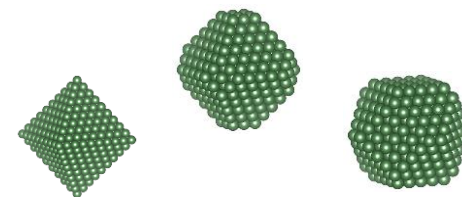
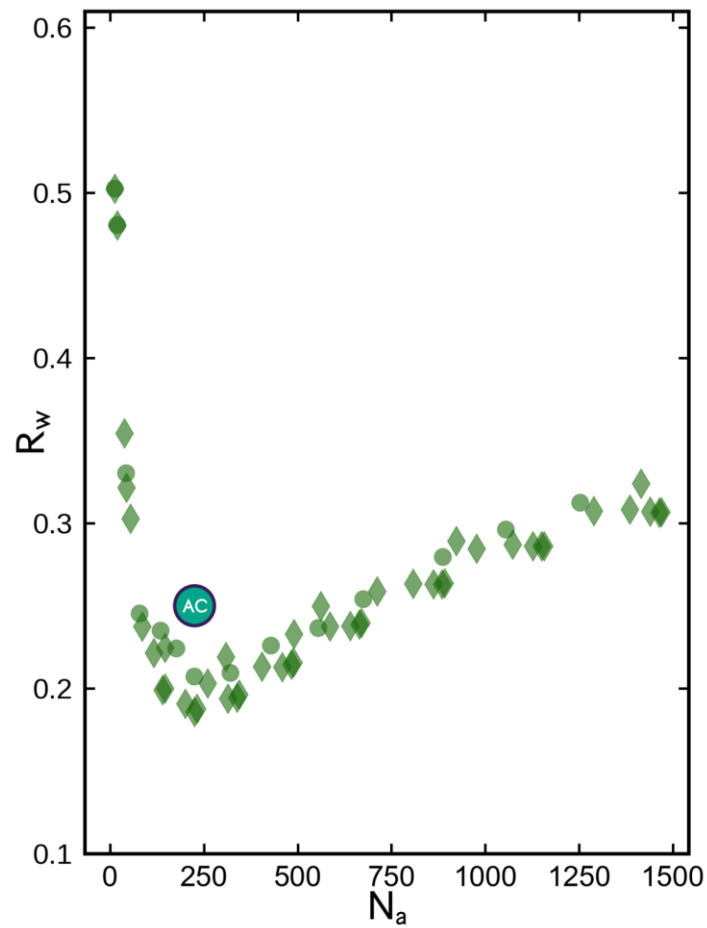


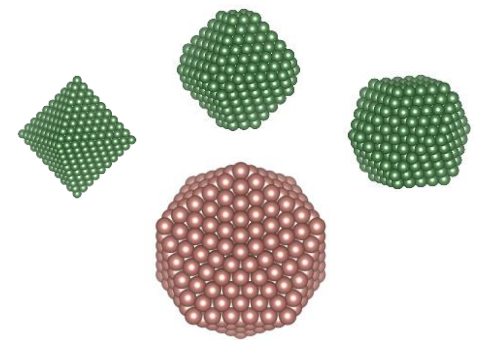
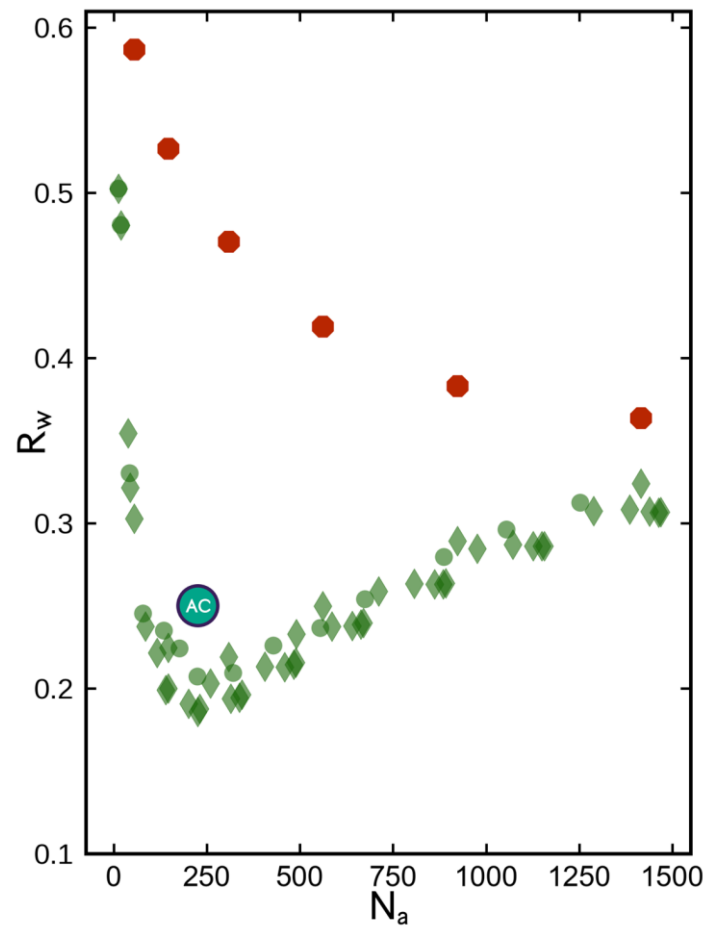
Banerjee, SJLB, et al. “Cluster-mining: An approach for determining core structures of metallic nanoparticles from atomic pair distribution function data”. In: *Acta Crystallogr. A* 76.1 (2020), pp. 24–31. doi: 10.1107/S2053273319013214.



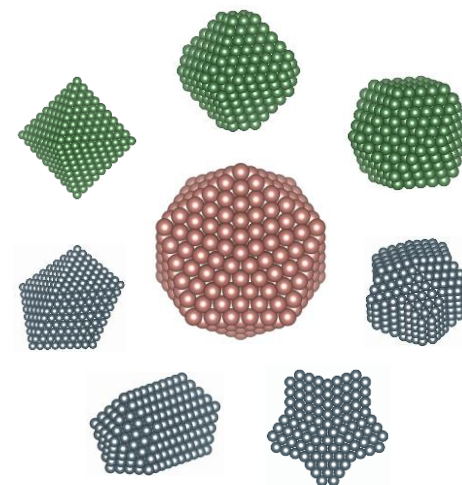
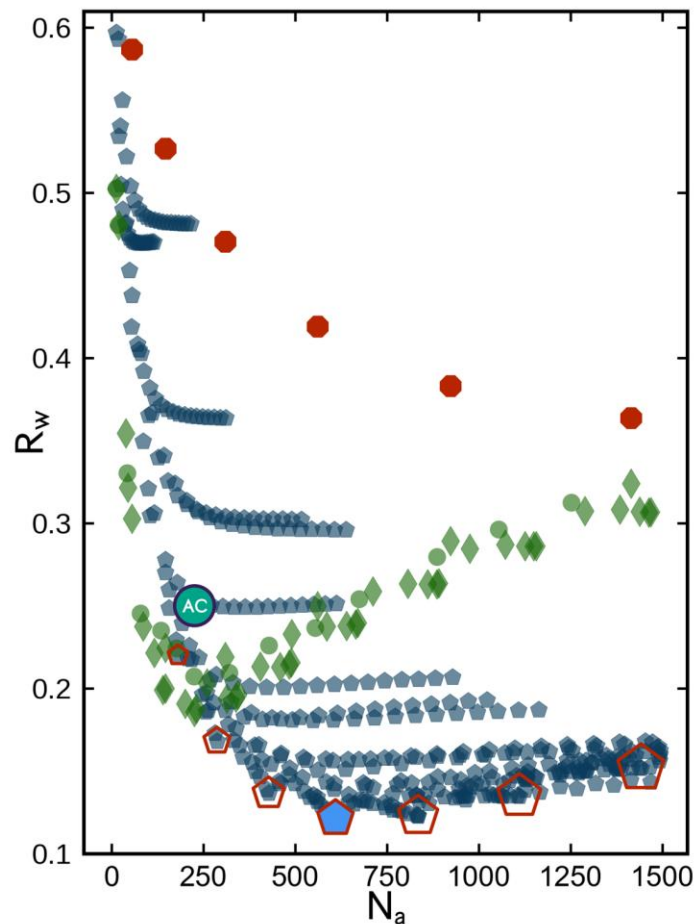








Banerjee, SJLB, et al. “Cluster-mining: An approach for determining core structures of metallic nanoparticles from atomic pair distribution function data”. In: *Acta Crystallogr. A* 76.1 (2020), pp. 24–31. doi: 10.1107/S2053273319013214.



Structure-mining: An automated screening of large numbers of candidate structures to the atomic pair distribution function (PDF)

Long Yang, Pavol Juhas, Maxwell W. Terban, Simon J. L. Billinge
Acta Crystallographica Section A, in press



Introduction

Crystal structural database



Materials Project Database
Crystallography Open Database

PDF auto-refinement



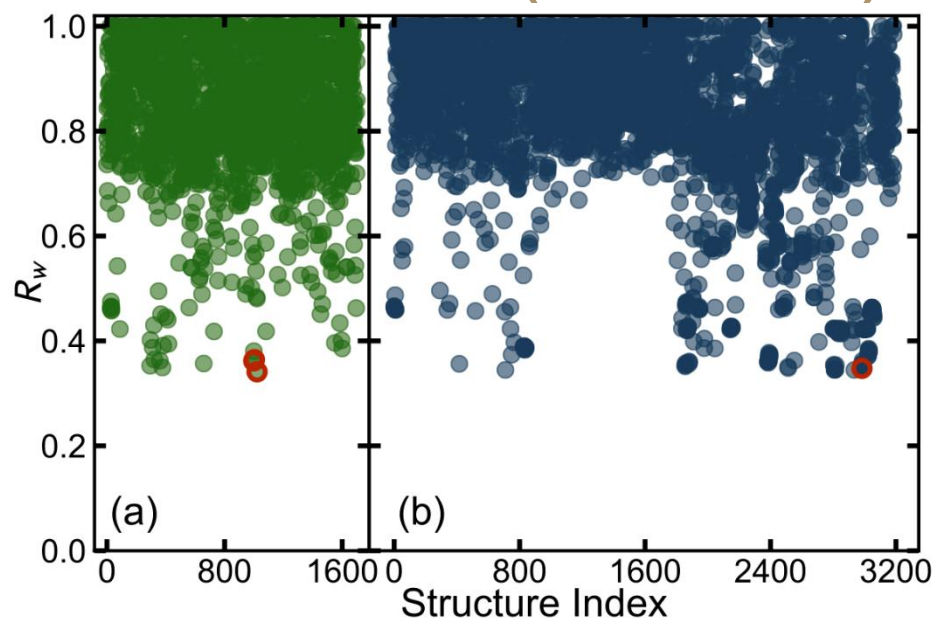
DiffPy-CMI

R_w sorted structure-mining result

No.	R_w	formula	space_group	Ba- U_{iso-r}	O- U_{iso-r}	Ti- U_{iso-r}
20	0.143308	Ba2 O6 Ti2	Amm2	0.002085	0.018062	0.008006
21	0.144425	Ba2 O6 Ti2	Amm2	0.002027	0.017803	0.008684
22	0.145316	Ba2 O6 Ti2	Amm2	0.001999	0.017465	0.009089
26	0.146433	Ba2 O6 Ti2	Amm2	0.001973	0.016756	0.009420
30	0.148121	Ba2 O6 Ti2	Amm2	0.001957	0.017870	0.009978
33	0.151343	Ba2 O6 Ti2	Amm2	0.002535	0.006989	0.008263
2	0.155847	Ba1 O3 Ti1	P4mm	0.002735	0.011571	0.006421
8	0.161942	Ba1 O3 Ti1	P4mm	0.002634	0.016226	0.008579
0	0.161942	Ba1 O3 Ti1	P4mm	0.002634	0.016226	0.008579
31	0.163028	Ba1 O3 Ti1	R3m	0.002704	0.016278	0.007618
29	0.163086	Ba2 O6 Ti2	Amm2	0.002827	0.002921	0.005437
28	0.163120	Ba1 O3 Ti1	R3m	0.002699	0.016083	0.007675
27	0.163405	Ba1 O3 Ti1	R3m	0.002681	0.015975	0.007887
25	0.163844	Ba1 O3 Ti1	R3m	0.002663	0.015968	0.008149
17	0.163934	Ba1 O3 Ti1	R3m	0.002649	0.015838	0.008251
24	0.164318	Ba1 O3 Ti1	R3m	0.002630	0.015253	0.008431
16	0.164382	Ba1 O3 Ti1	R3m	0.002639	0.015667	0.008449
18	0.164820	Ba1 O3 Ti1	R3m	0.002614	0.015019	0.008652
32	0.165866	Ba1 O3 Ti1	R3m	0.002599	0.014932	0.009055
3	0.165928	Ba1 O3 Ti1	P4mm	0.002610	0.015074	0.009552
9	0.166332	Ba1 O3 Ti1	P4mm	0.002616	0.015758	0.009675
4	0.167933	Ba1 O3 Ti1	Pmm2	0.002485	0.015098	0.009475
23	0.169423	Ba1 O3 Ti1	P4mm	0.002625	0.016343	0.010339
11	0.170190	Ba1 O3 Ti1	P4/mmm	0.002643	0.017426	0.010542
10	0.170190	Ba1 O3 Ti1	P4/mmm	0.002643	0.017426	0.010542
15	0.210176	Ba1 O3 Ti1	Pm-3m	0.004582	0.017209	0.013186
1	0.210176	Ba1 O3 Ti1	Pm-3m	0.004582	0.017209	0.013186
13	0.210176	Ba1 O3 Ti1	Pm-3m	0.004582	0.017209	0.013186
12	0.210176	Ba1 O3 Ti1	Pm-3m	0.004582	0.017209	0.013186
5	0.210176	Ba1 O3 Ti1	Pm-3m	0.004582	0.017209	0.013186
14	0.210176	Ba1 O3 Ti1	Pm-3m	0.004582	0.017209	0.013186
6	0.366657	Ba1 O3 Ti1	Pm-3m	0.005790	0.079941	0.012610
7	0.573221	Ba6 O18 Ti6	P6 ₃ /mmc	0.006970	0.046913	0.004137
19	0.707676	Ba1 O3 Ti1	P4mm	0.004189	0.047858	0.248950

Example result of BaTiO₃ nanoparticle heuristic- I

Heuristic-4 (*-*Si-O)



- Structure-mining found the same model as in prior work, MPD No. 1003 ($\text{NaFeSi}_2\text{O}_6$) and COD No. 2983 ($\text{NaFeSi}_2\text{O}_6$), s.g.: C 2/c.
- It also returns some structures with space group C 2, such as MPD No. 998 ($\text{Na}_{0.83}\text{FeSi}_2\text{O}_6$), which may be viewed as a very similar structure but with a lowered symmetry and deficient atoms at some sites
- It also returns some structures substituting at Na or Fe sites by other elements. For example, MPD No. 1021 ($\text{NaGaSi}_2\text{O}_6$).

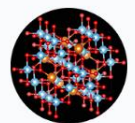
To the cloud!!!!



Logout from Simon 

 PDF IN THE CLOUD

structureMining



Auto search for the best structures from an experimental PDF.

start

The Billinge Group, Columbia University



Log



PDF IN THE CLOUD

structureMining

Given a measured (or calculated) PDF, structureMining will search databases to find the best structures to fit it.

Please see the strucureMining paper for more information. Please cite the structureMining paper if this helps you get a publishable result.

Upload Data

PDF file: No file selected.

☒ X-ray ☐ Neutron

Composition:

Optional Parameter:

[The Billinge Group](#), Columbia University



PDF IN THE CLOUD

structureMining

Results

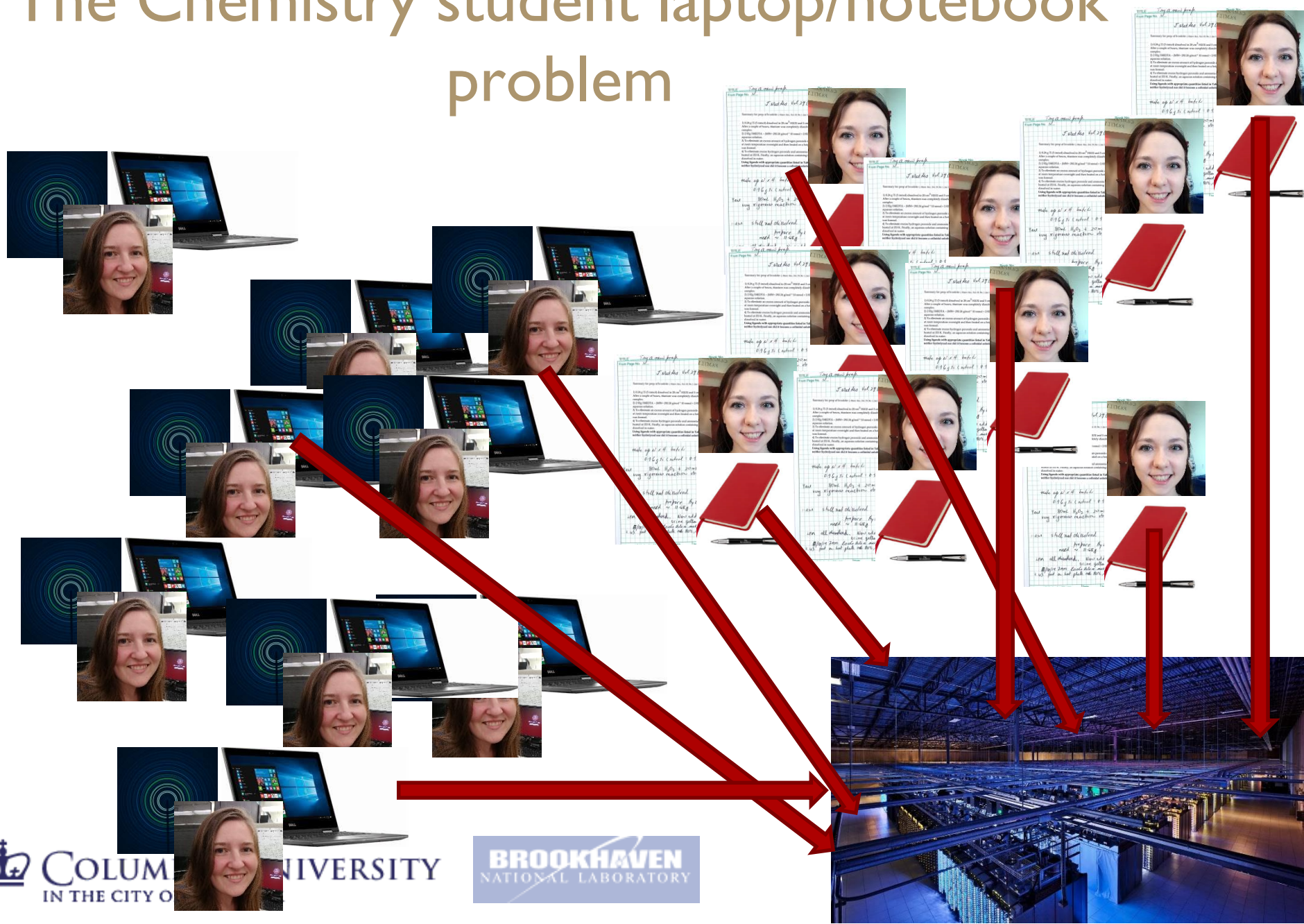
SM found total 3 structures and 3 structures with weighted profile agreement factor, $R_w < 0.5$.

[Download Full Table](#)

Expand Table

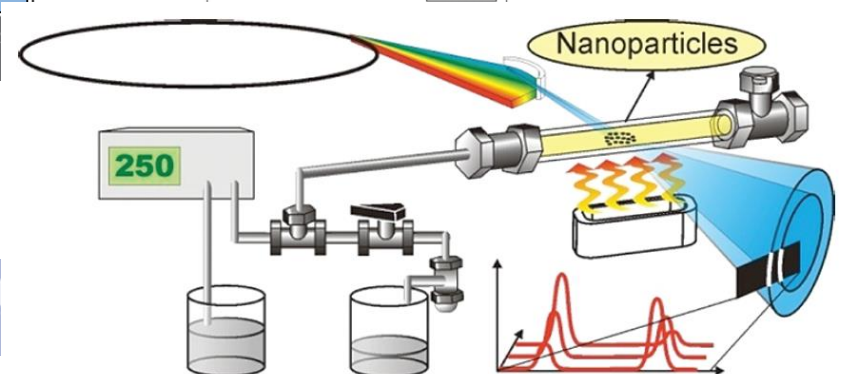
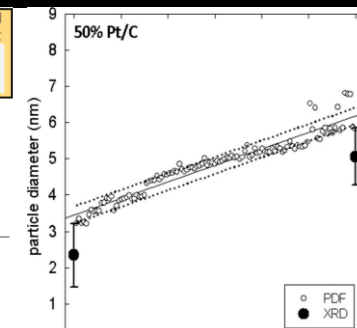
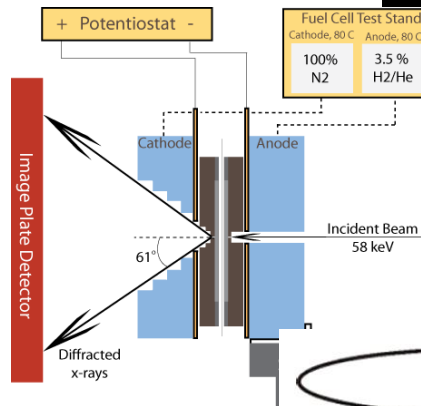
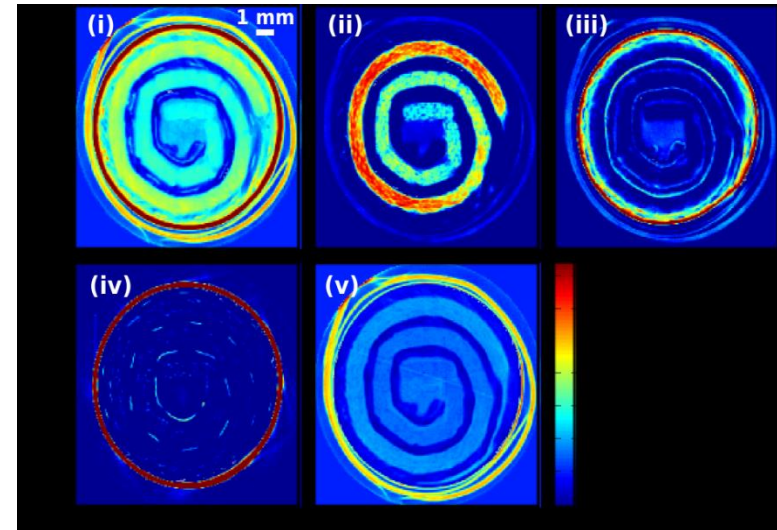
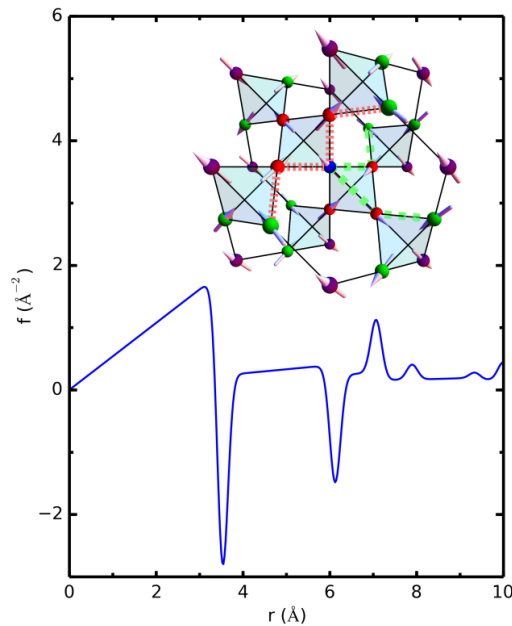
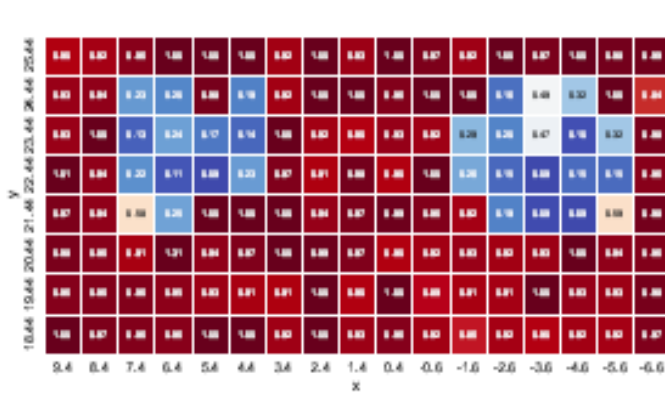
	rw	formula	space_group	db	db_id	ref	download
0	0.140570	As ₄ Fe ₄ Sr ₂	I4/mmm	COD	9016374	Tegel. Journal of Physics: Condensed Matter. 20 (2008) 452201_1--452201_5	 
1	0.140673	As ₈ Fe ₈ Sr ₄	Fmmm	COD	9016523	Tegel. Journal of Physics: Condensed Matter. 20 (2008) 452201_1--452201_5	 
2	0.325028	As ₂ Fe ₂ Sr ₁	I4/mmm	MPD	mp-4488	Saha et al. Journal of Physics: Conference Series. 273 (2011) 012104/1-012104/4	 

The Chemistry student laptop/notebook problem



The analyzed structured labelled data from everyone in a database problem

and all this with spatial and time dependence



This new approach to modeling opens the door for Machine Learning/Deep learning

- I don't know an person who can look at a PDF and tell the s.g. (except perhaps if it is an fcc material or something)

Question:

- Can we train a Machine Learning model to recognize it?

Answer:

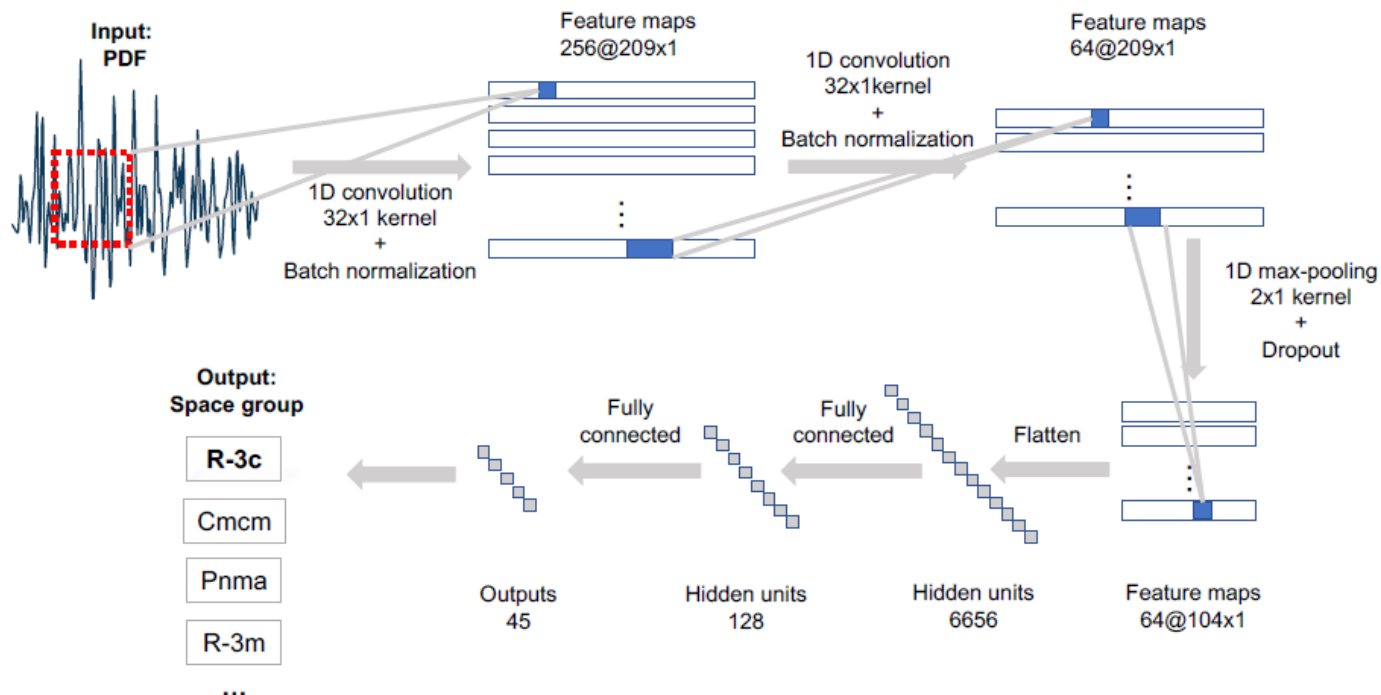
- Yes!

Chia-Hao Liu, Yunzhe Tao, Daniel J. Hsu, Qiang Du, and Billinge, Simon J. L., *Acta Crystallogr. A* 75 (2019), pp. 633–643.
doi: 10.1107/S2053273319005606.



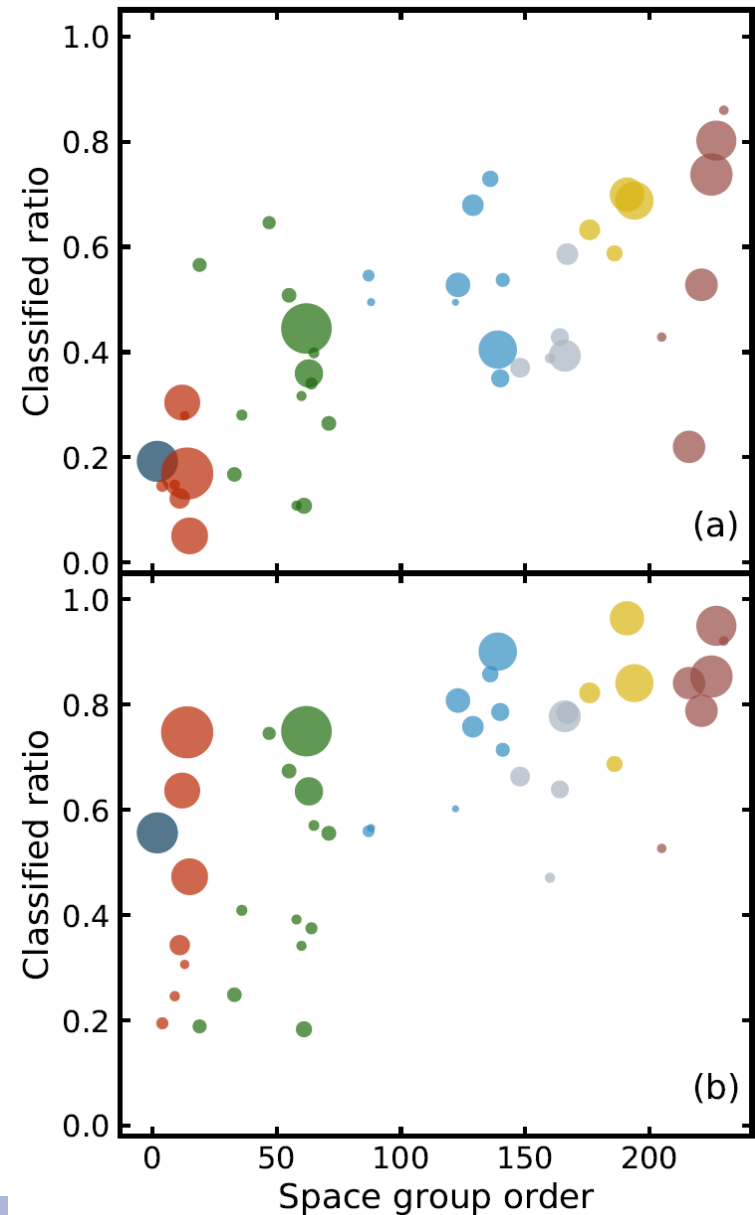
Details

- Work of Yunzhe Tao and Chia-Hao Liu
- PDF itself is the input feature vector
- 101,802 structures in 45 space-groups
- Train on 80% of data, 20% retained for testing

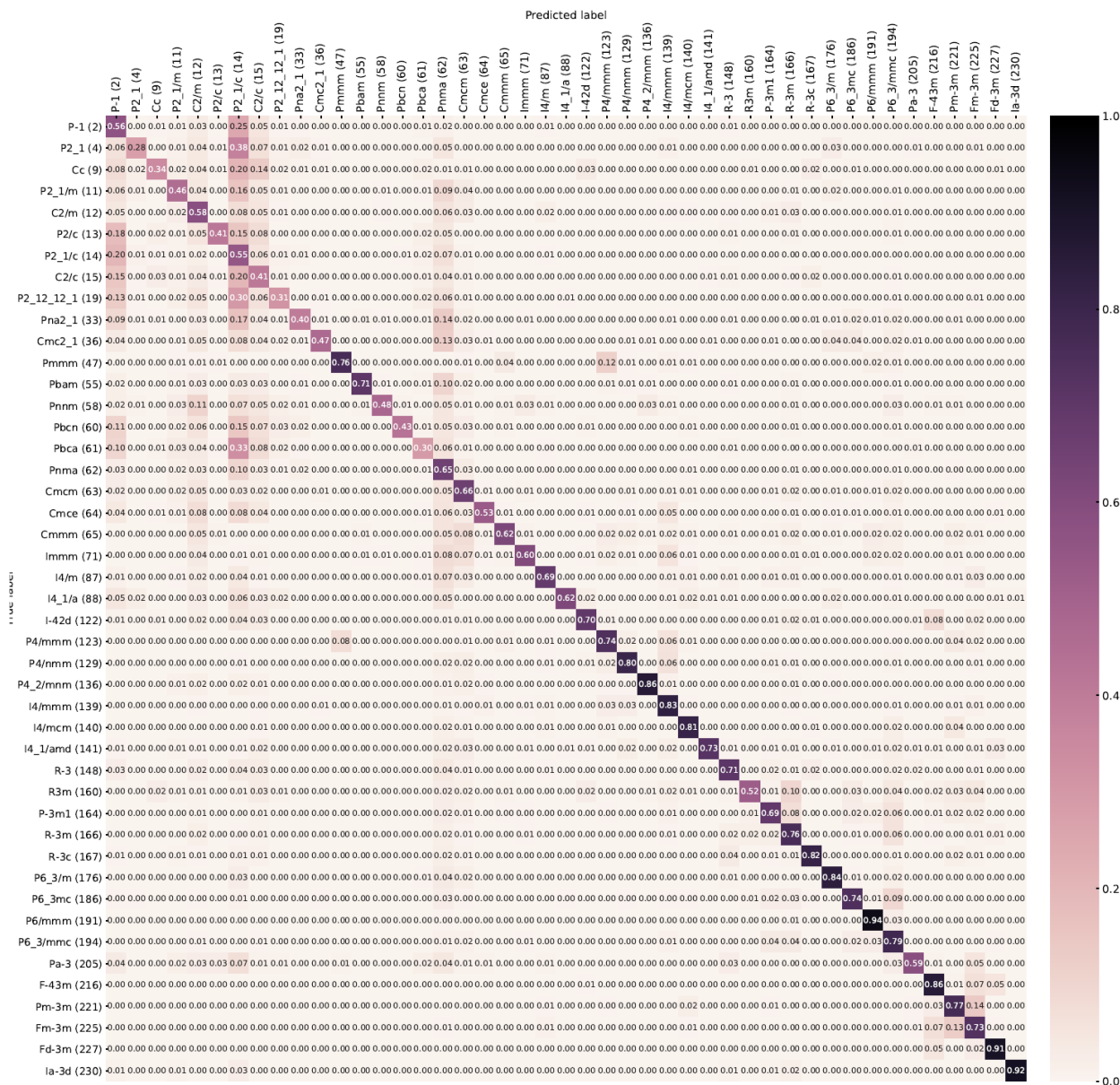
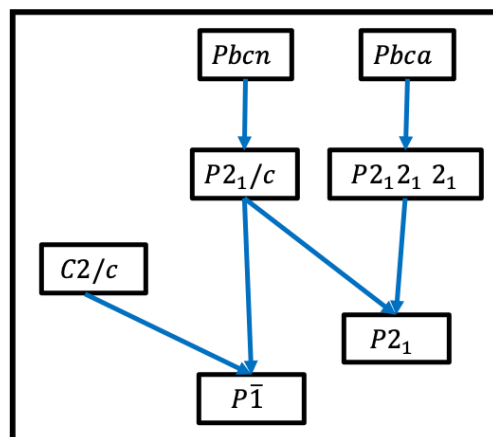


Space Groups from PDFs

- First try, Logistic Regression
- Second try, Convolution Neural Net
 - Classification ratio's are for finding the right s.g. in the top 6



- Rows: true label
- Columns: predicted label



Acknowledgements

- A special thank you to all my current and former students and post-docs
- Also my many wonderful collaborators, mentioned during the talk
- Facilities:
 - APS, CHESS, NSLS, NSLS-II (and people therein)
 - MLNSC, ISIS, SNS (and people therein)
- Funding: DOE-BES and NSF-DMR

