Revisiting material structure in the time of AI

Simon J. L. Billinge

[1]Columbia University

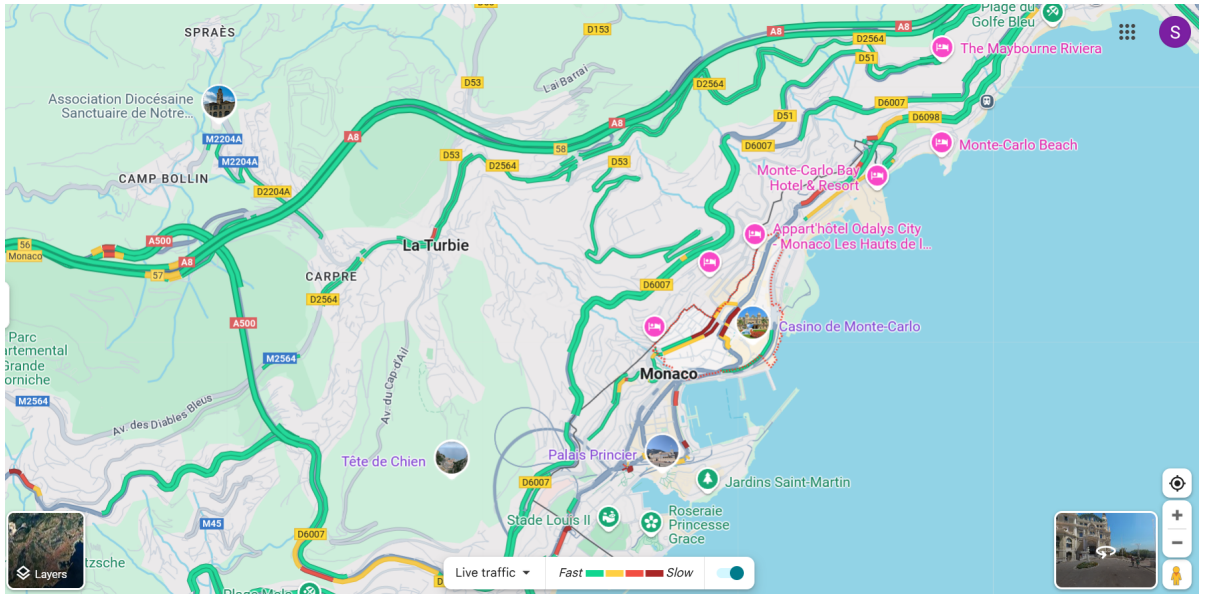
July 9, 2025

What I learned about materials from AI

What I learned about materials from AI

What I learned about materials from AI

Let me recap a bit of my journey

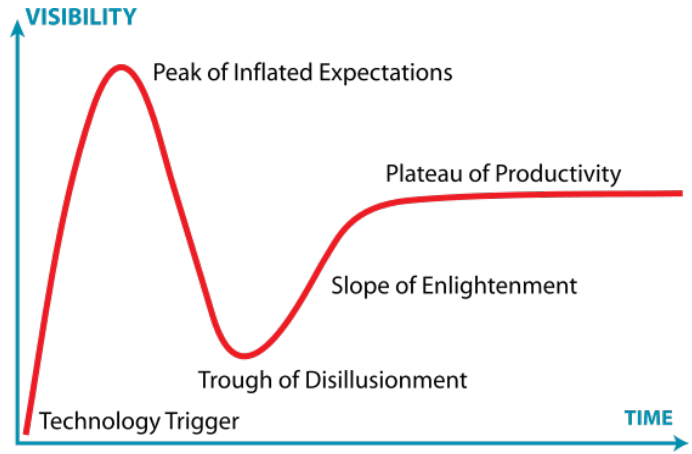


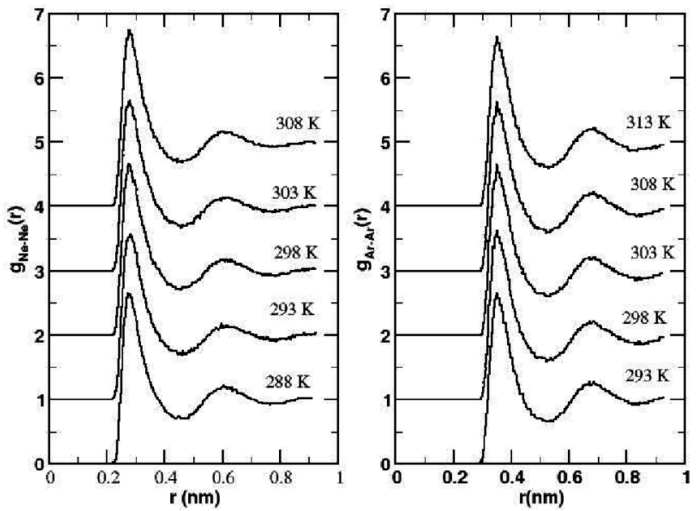
What I learned about materials from AI

It started with a classification problem

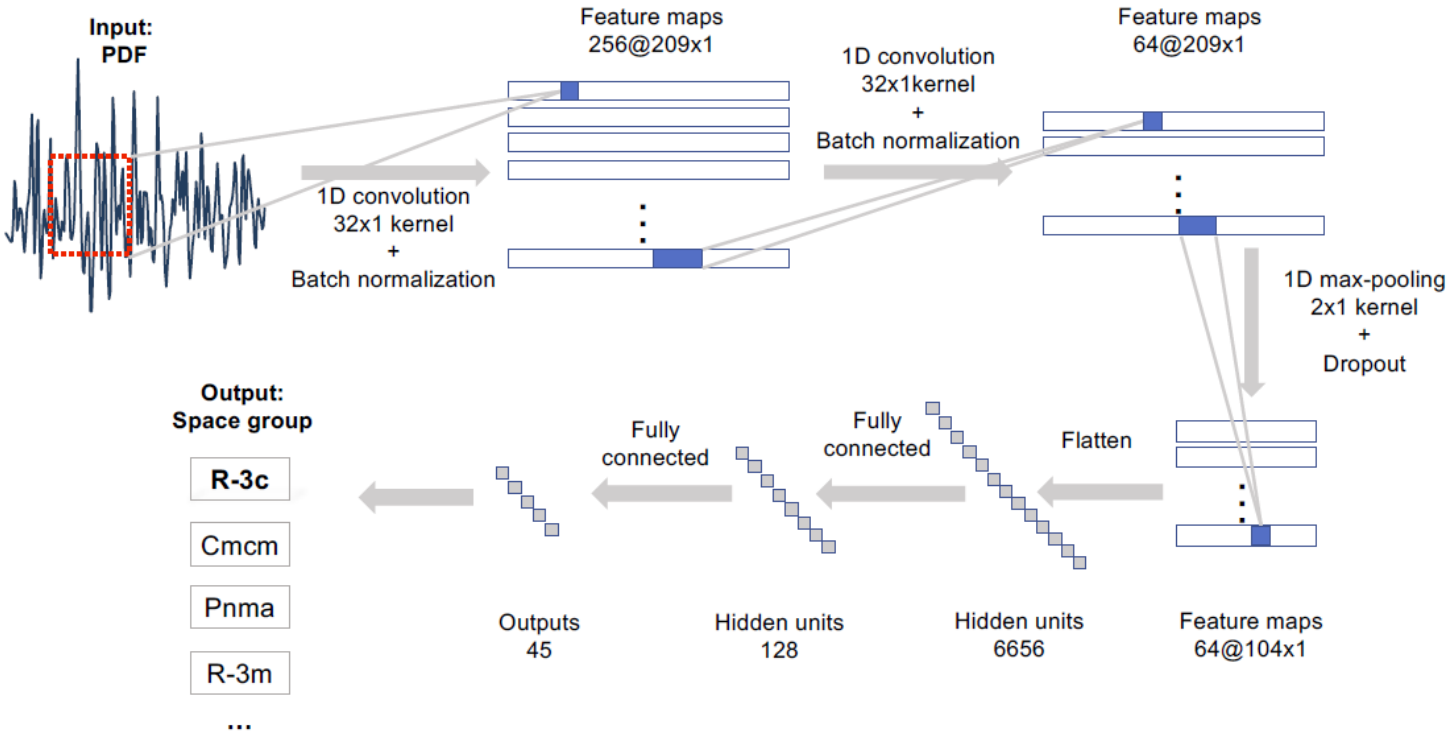
- We can't get space-group from PDF
- But PDF is just the FT of diffraction data
- We can get SG from diffraction data

So the information is there, can a ML model figure out a surrogate forward problem?





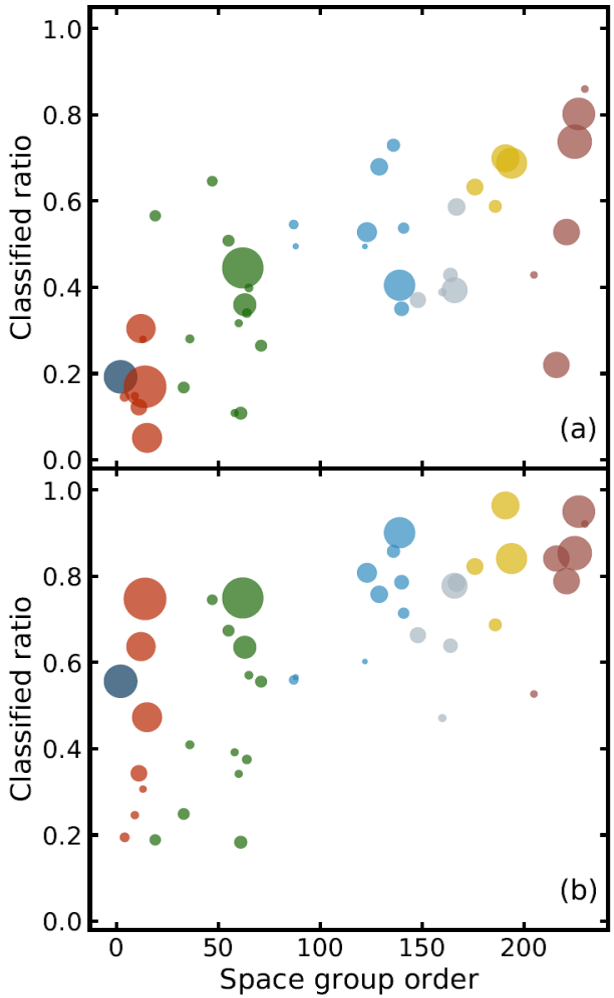
Given a PDF, what was the space-group of the material?



collaboration with group of Qiang Du (Columbia U)

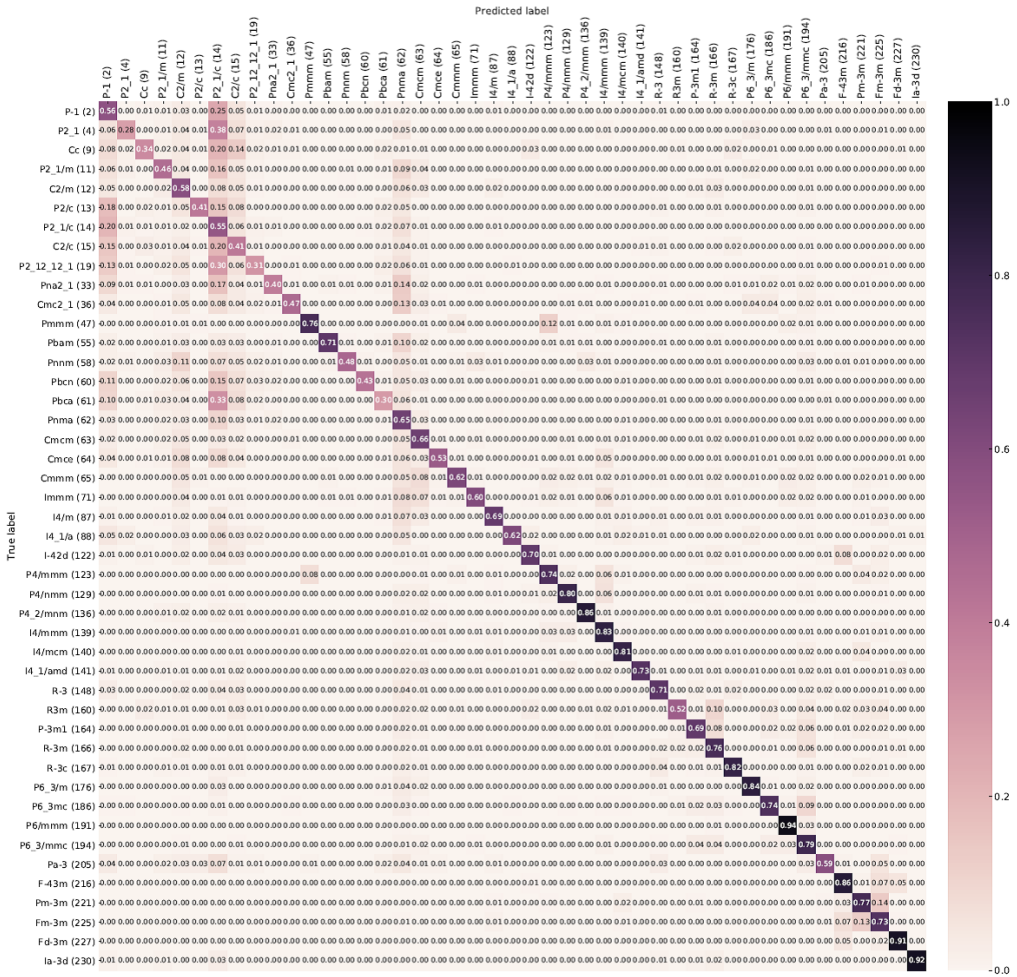
Chia-Hao Liu, SJLB et al., Acta Cryst. (2019), 10.1107/S2053273319005606

How well did it perform?



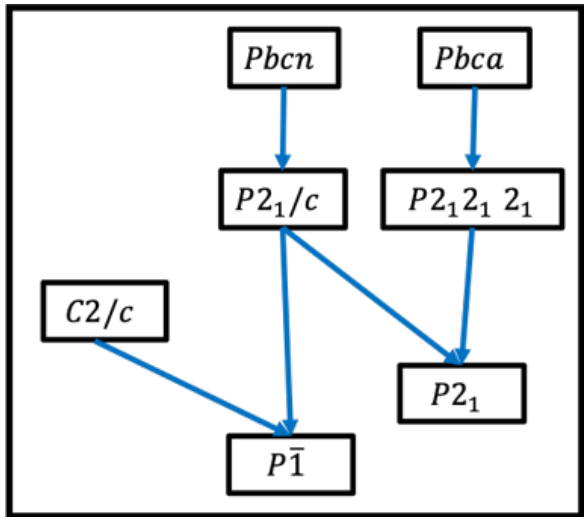
- Top six classification ratio
- (top) Logistic Regression baseline
- (bottom) CNN 91% overall
- Color indicates crystal system
- Size of spot indicates sample-size

We learned crystallography from the model



Matrix of confusion

- tear drops on space-groups C2/1, Pnma
- connected by group-sub-group relationships



Can we do structure solution with Generative AI?

The goal

- Input: the PDF or a powder diffraction pattern
- Output: the structure of the material

That would be cool!

What I learned about AI by applying it to structure solution

What I thought I knew

- Gen-AI is doing interpolation

What I learned about AI by applying it to structure solution

What I thought I knew

- Gen-AI is doing interpolation
- But how do we interpolate structures into each other?

What I learned about AI by applying it to structure solution

What I thought I knew

- Gen-AI is doing interpolation
- But how do we interpolate structures into each other?
- For close-packed structures, I can interpolate from fcc to bcc by introducing stacking-fault defects.

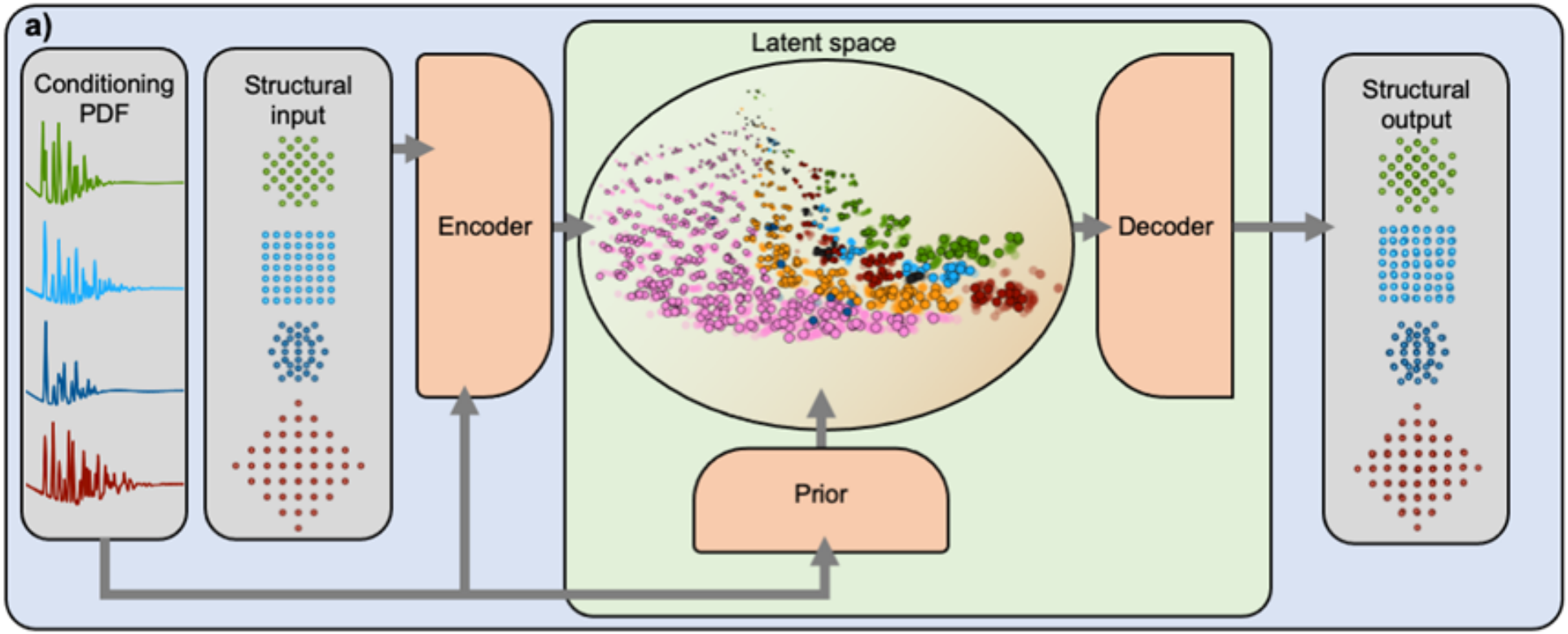
Let's do close-packed structures that interpolate

Case 1: close packed structures

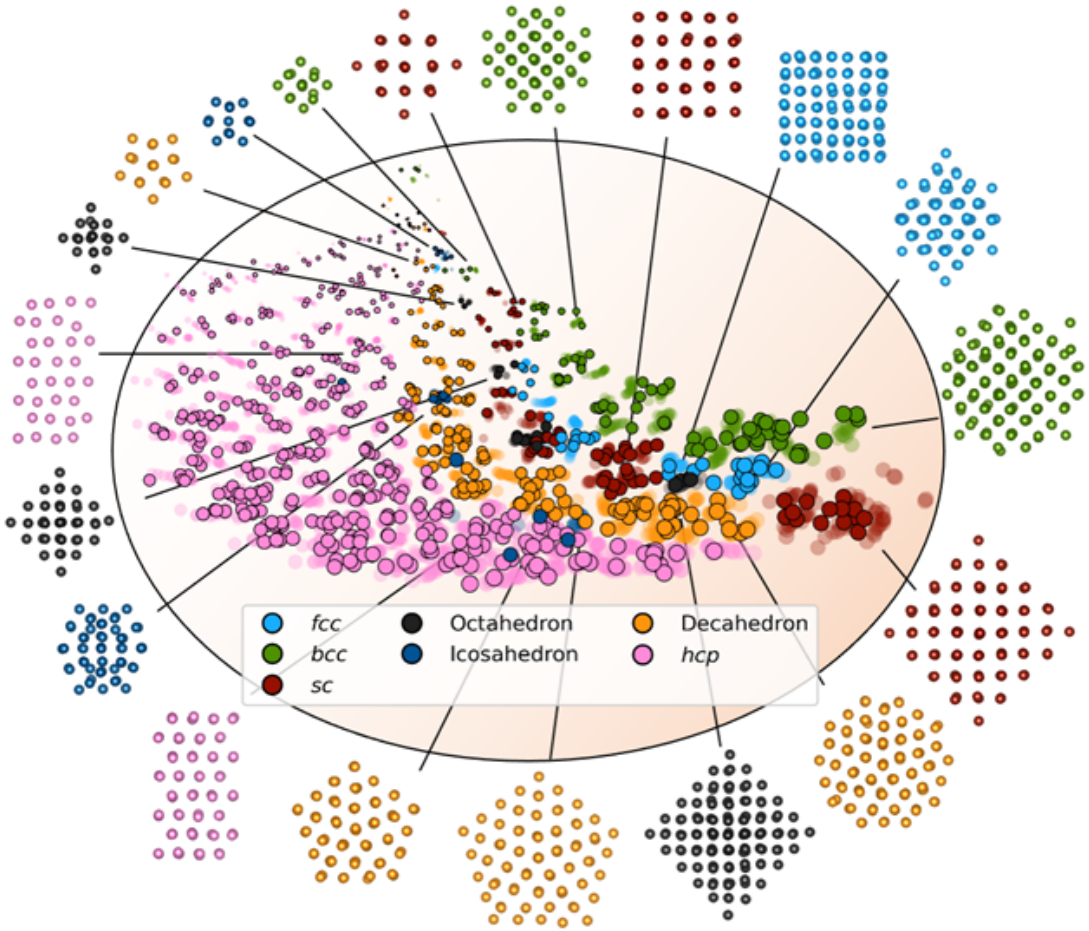
- Graph Convolutional Variational Autoencoder (CVAE)
- Collaboration with Kirsten Jensen, Raghav Selvan, U. Copenhagen
- Work of students Emil Kjaer and Andy Anker
- Kjaer, Anker, et al., Digital Discovery (2023), 10.1039/D2DD00086E

Structure solution using Neural Nets

Deepstruc: close packed nanoparticles with a CVAE



Deepstruc: close packed nanoparticles with a CVAE



What I learned about AI by applying it to structure solution

What I thought I knew

- Gen-AI is doing interpolation
- But how do we interpolate structures into each other?
- For close-packed structures, I can interpolate from fcc to bcc by introducing stacking-fault defects.
- It worked!

What I learned about AI by applying it to structure solution

What I thought I knew

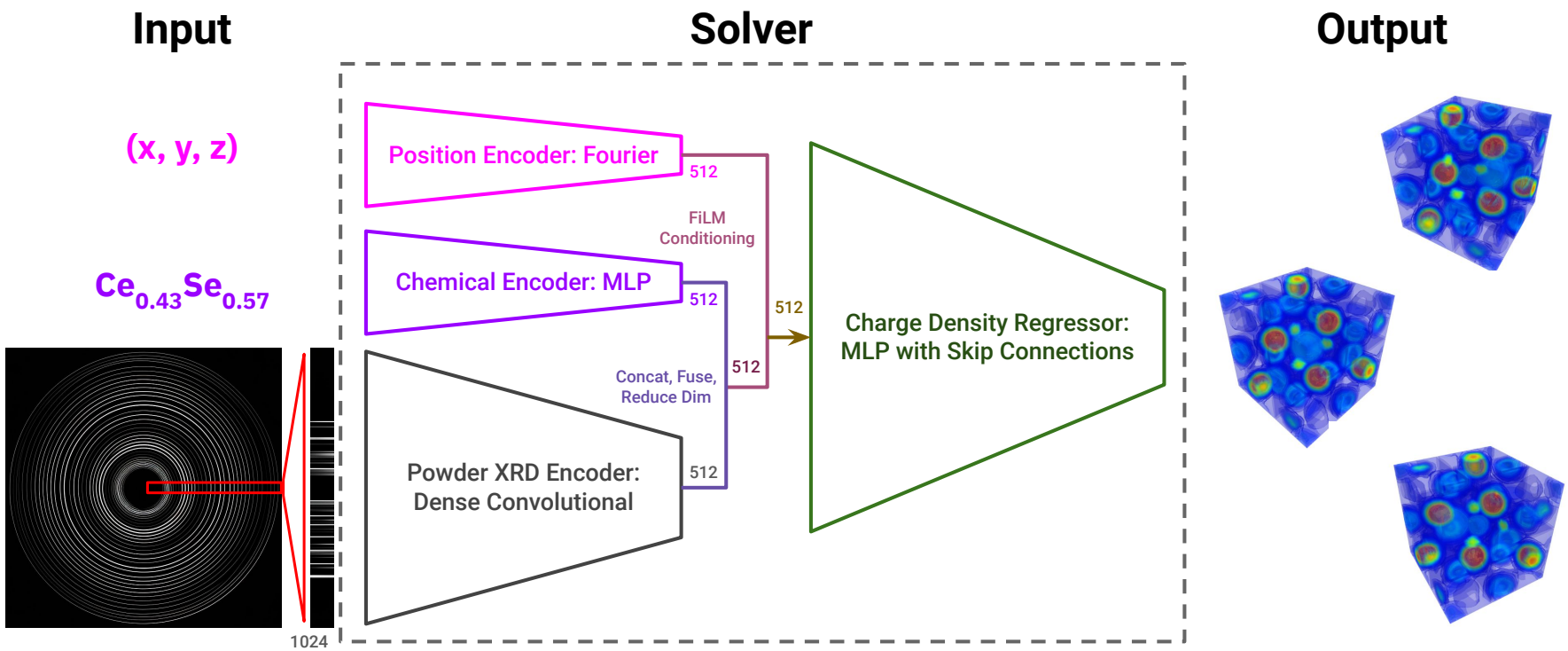
- Gen-AI is doing interpolation
- But how do we interpolate structures into each other?
- For close-packed structures, I can interpolate from fcc to bcc by introducing stacking-fault defects.
- It worked!
- But in general? If I draw a line from NaCl to nickel, I don't find $CaCO_4$ in between

Let's change the problem to something continuous....electron-density!

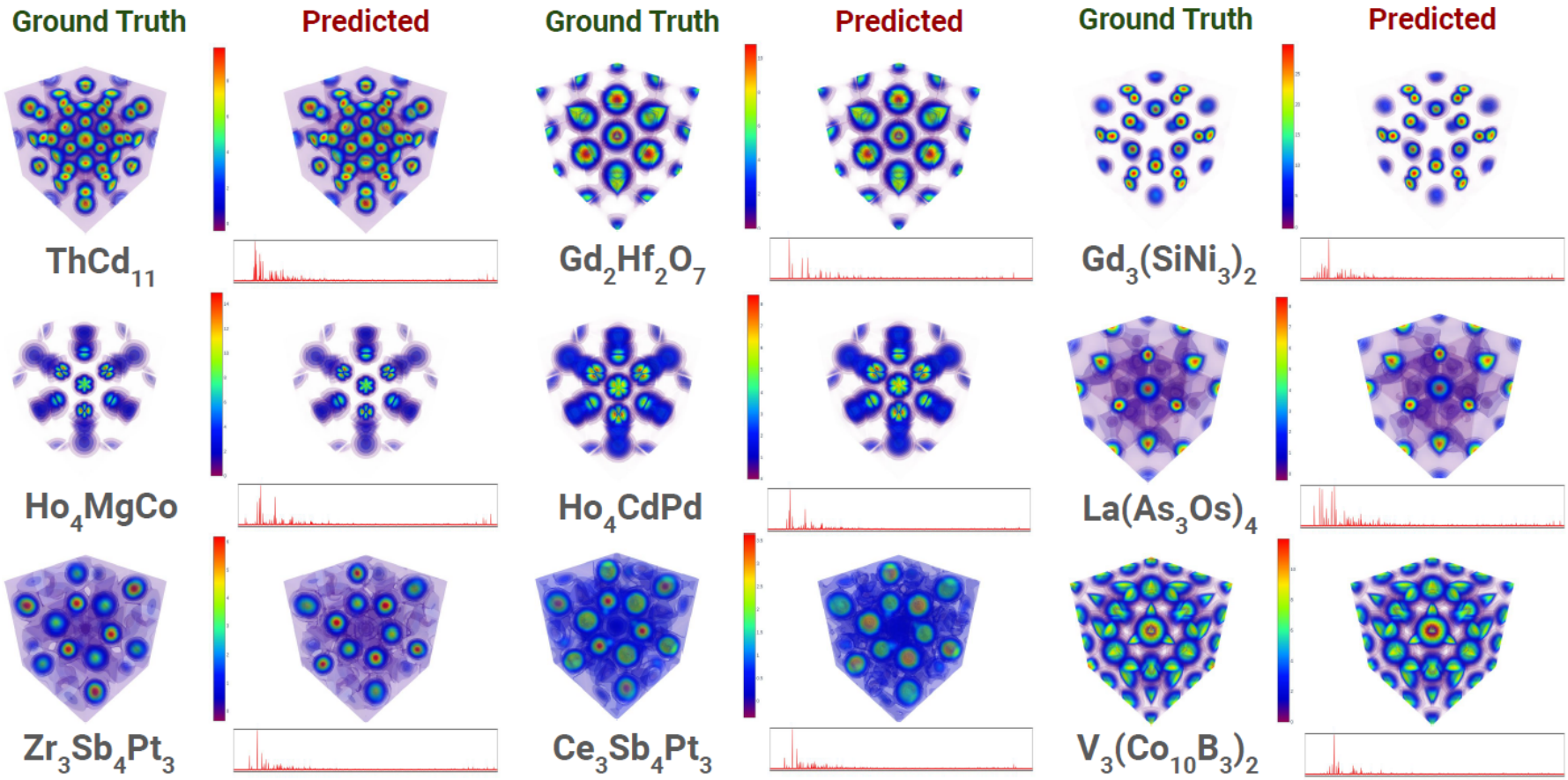
Case 2: electron density determination

- Collaboration with Hod Lipson (Columbia University)
- Work of students Gabriel Guo and Ling Lan
- G. Guo et al., npj-Computational Materials, to be published
- Graph Convolutional Variational Autoencoder (CVAE)
- input: powder diffraction pattern and composition
- output: electron density distribution
- Trained on cubic and trigonal (non-orthogonal but high symmetry) crystal systems

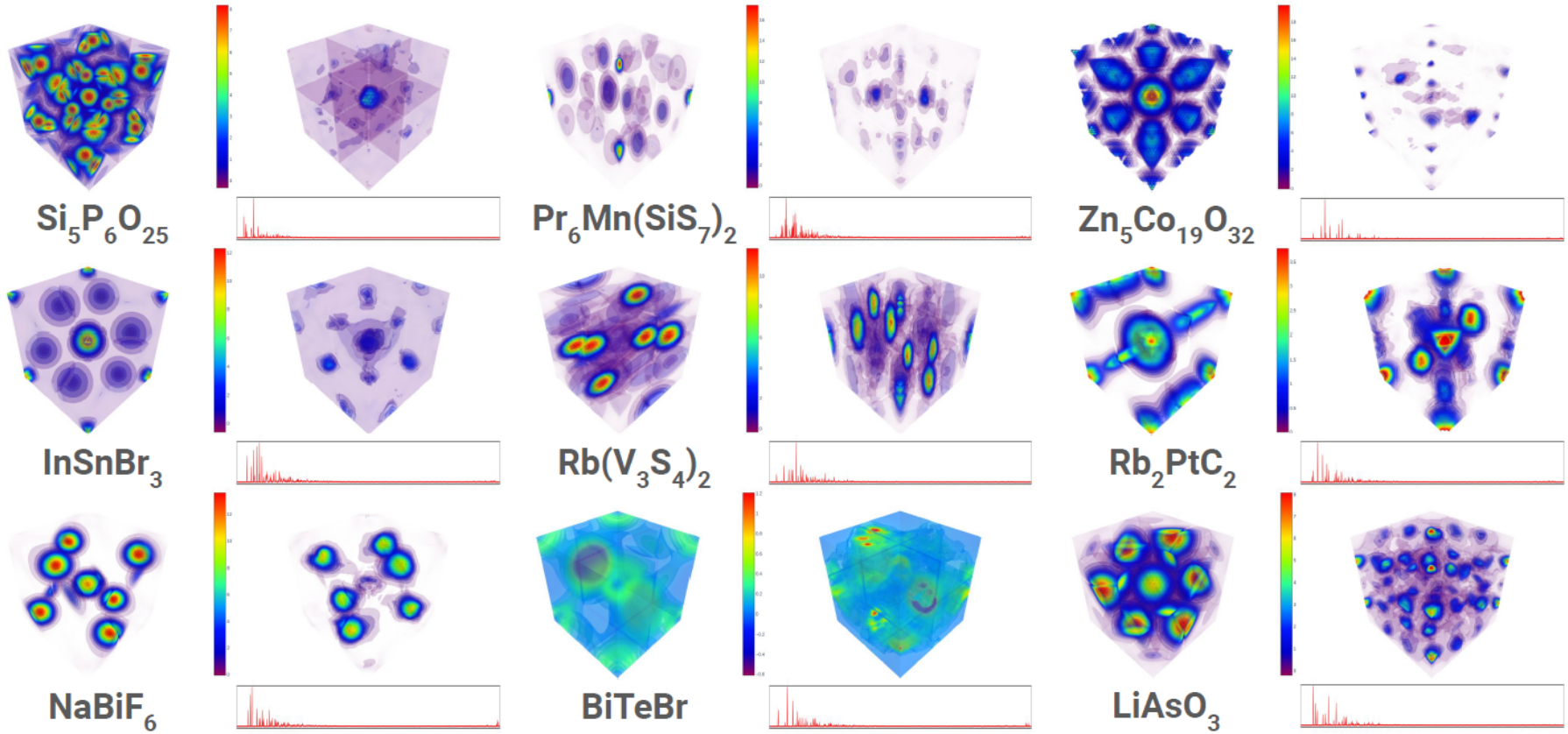
CrystalNet: Electron density reconstruction with CVAE



CrystalNet: Electron density reconstruction with CVAE



CrystalNet: Electron density reconstruction with CVAE



What I learned about AI by applying it to structure solution

What I learned

- Gen-AI is doing interpolation
- But how do we interpolate structures into each other?
- In general? If I draw a line from NaCl to nickel, I don't find $CaCO_4$ in between

What I learned about AI by applying it to structure solution

What I learned

- Gen-AI is doing interpolation
- But how do we interpolate structures into each other?
- In general? If I draw a line from NaCl to nickel, I don't find $CaCO_4$ in between
- But the interpolation is happening in the latent space of the auto-encoder, not in our familiar "structure-space"

What I learned about AI by applying it to structure solution

What I learned

- Gen-AI is doing interpolation
- But how do we interpolate structures into each other?
- In general? If I draw a line from NaCl to nickel, I don't find $CaCO_4$ in between
- But the interpolation is happening in the latent space of the auto-encoder, not in our familiar "structure-space"

The magic of diffusion models

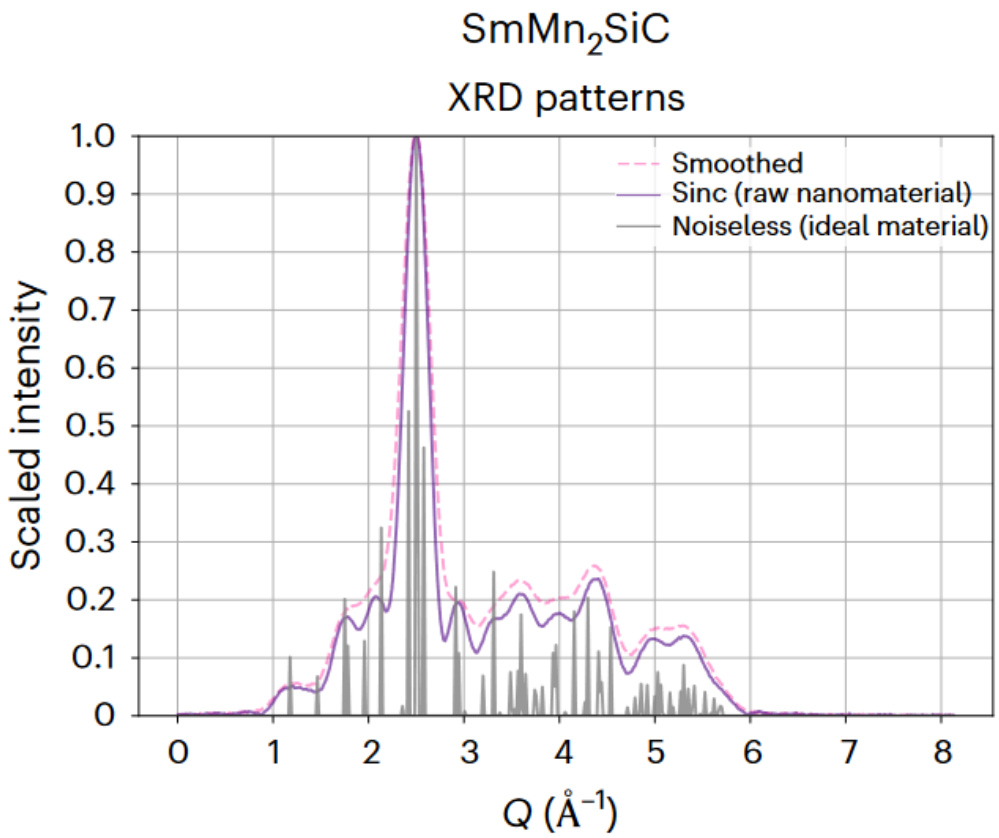
Case 3: structure solution of nanoparticles

- Collaboration with Hod Lipson (Columbia University)
- Work of students Gabriel Guo and Tristan Saidi
- G. Guo et al., Nature Materials, (2025) https://doi.org/10.1038/s41563-025-02220-y
- input: nanostructure powder diffraction pattern and composition
- output: "structure": unit cell, coordinates and coloring

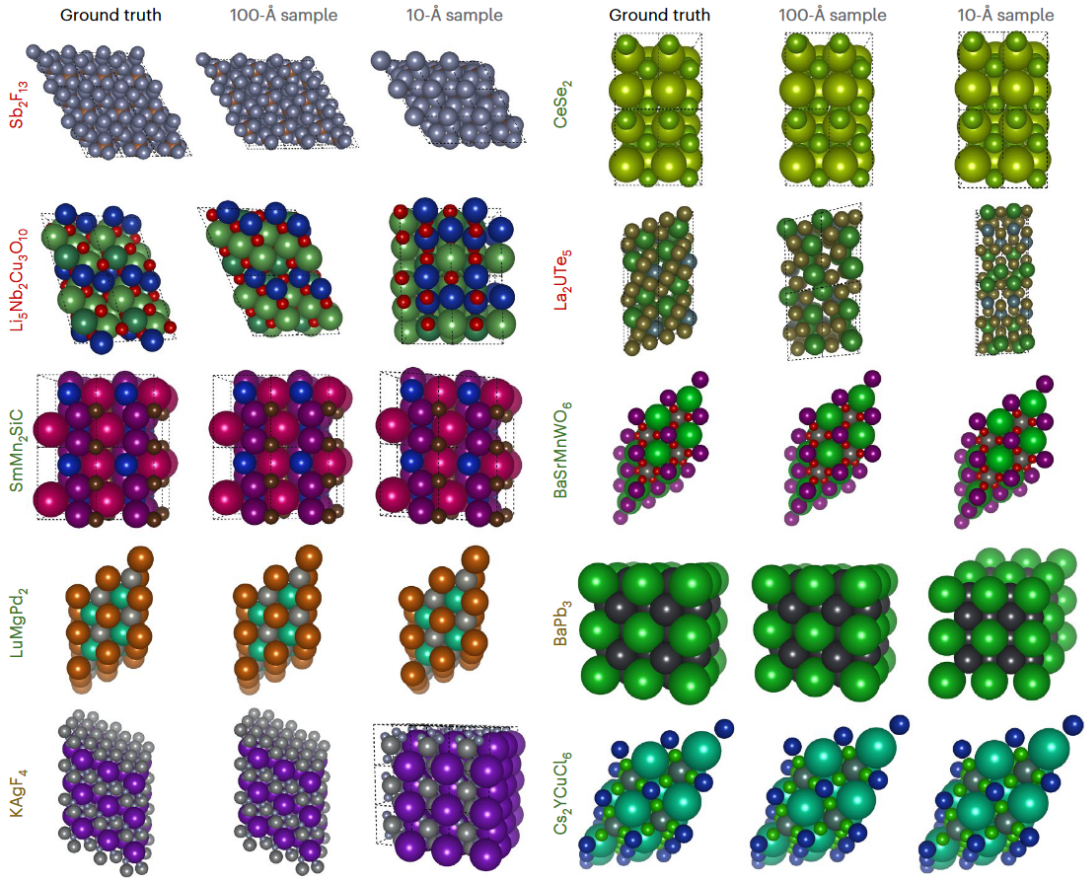
PXRDNet: structure solution with diffusion model

PXRDNet

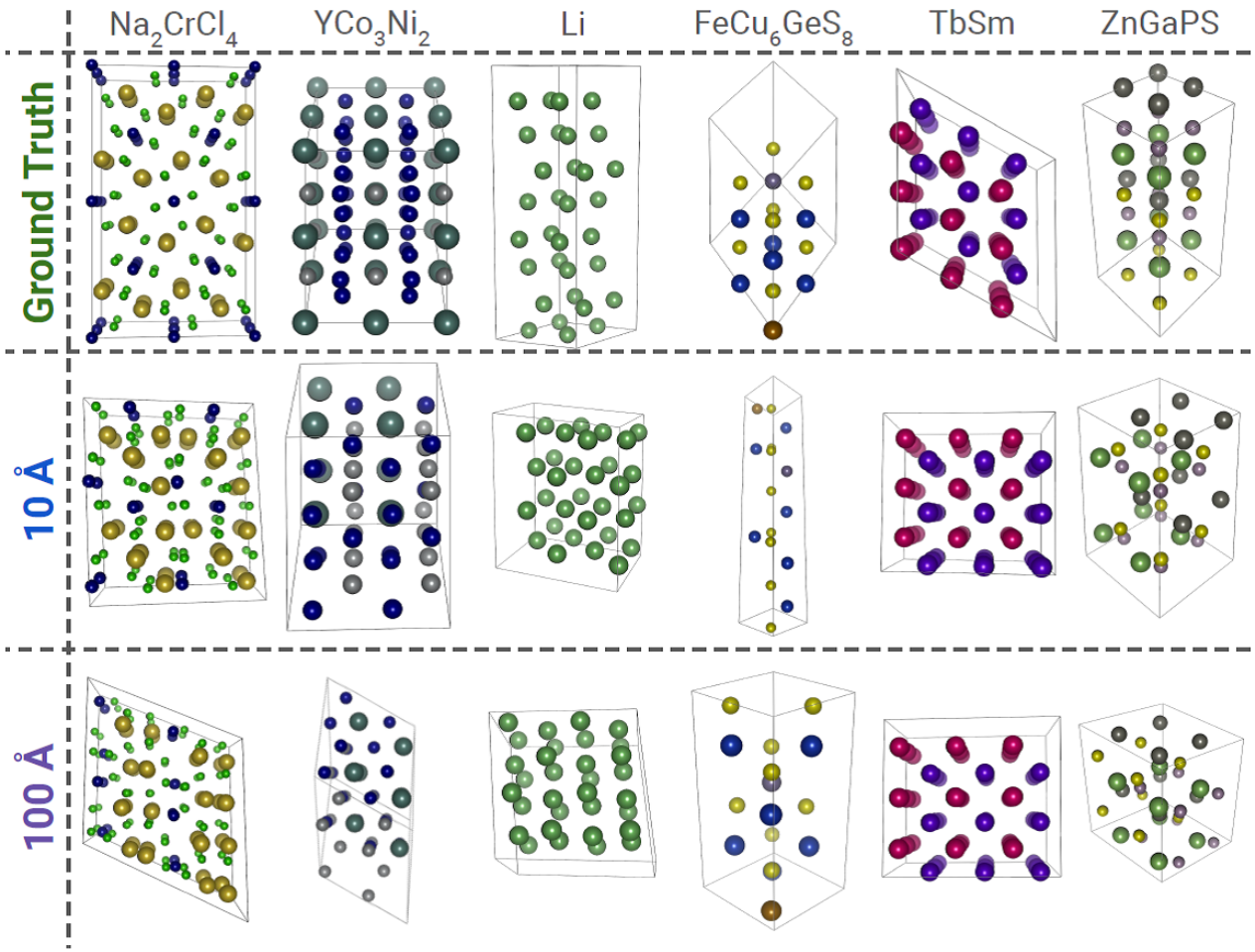
- 100 Å or 10 Å NP powder data. Low information content!
- Based on CDVAE (T. Xie, Proc. International Conference on Learning Representations (2022).
- SE(3) equivariant graph NN autoencoder for composition, lattice parameters and number of atoms.
- Denoising diffusion via noise-conditioned score networks for the atomic coordinates



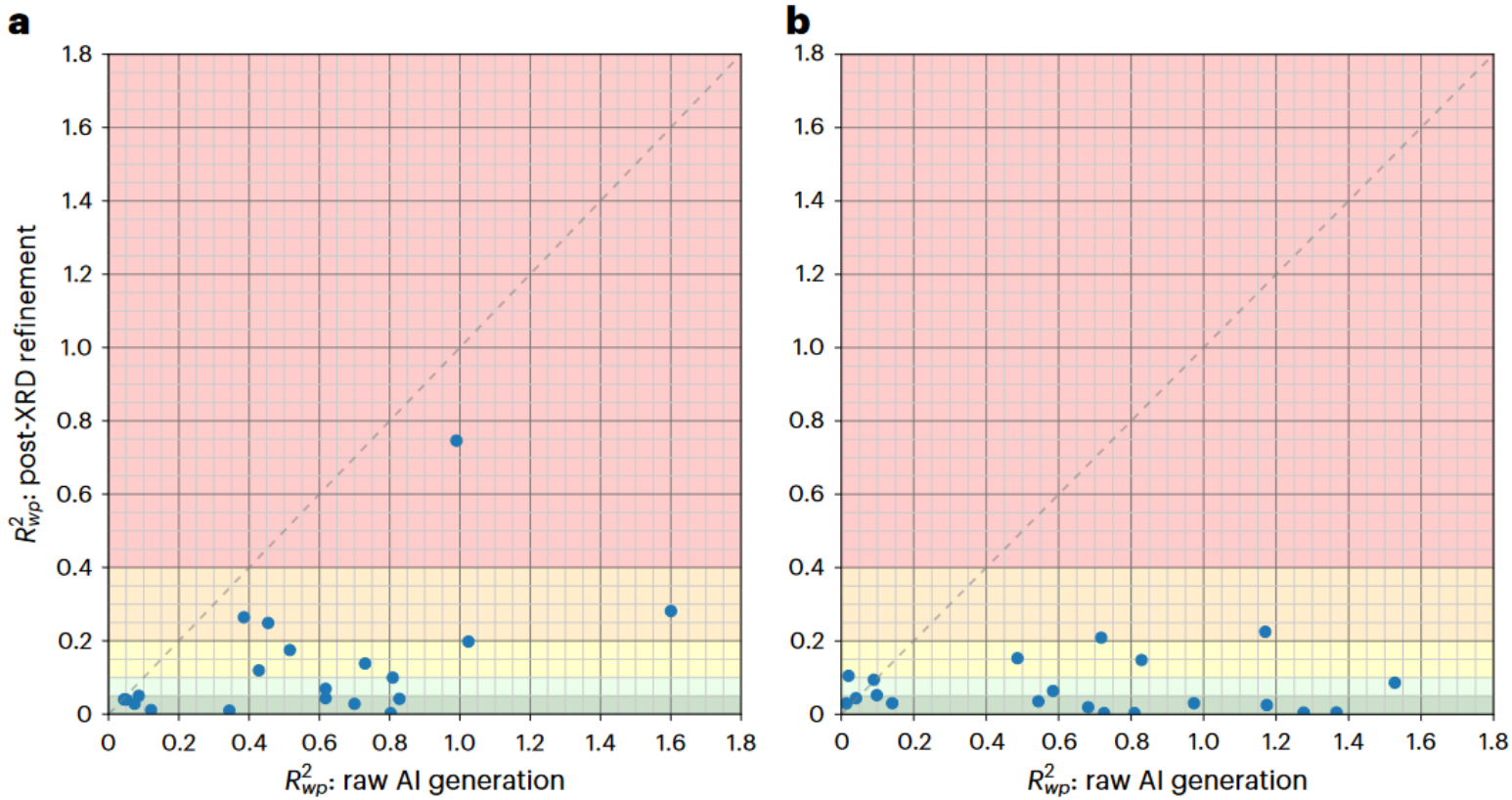
PXRDNet: structure solution with diffusion model



PXRDNet: structure solution with diffusion model



PXRDNet: structure solution with diffusion model



Diffusion models are the s$%t

Diffusion models are the s$%t

- They build latent spaces/embeddings that can interpolate non-interpolatable things
- They do it by starting with a large population of "good" things
- They then randomize them in a well defined way
- Any point in this space can be denoised
- A point in the space can be picked by a prompt, which could be a desired target

Diffusion models are the s$%t

Diffusion models are the s$%t

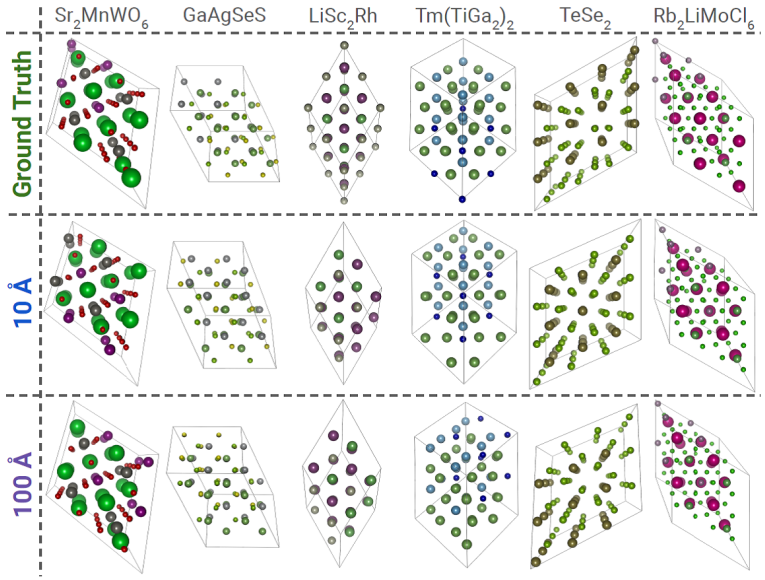
- They build latent spaces/embeddings that can interpolate non-interpolatable things
- They do it by starting with a large population of "good" things
- They then randomize them in a well defined way
- Any point in this space can be denoised
- A point in the space can be picked by a prompt, which could be a desired target

They work!

Example: PXRDnet

It works!

- After training the latent space and conditioning on (rather small amounts of) diffraction data, it makes a rather good guess of the structure solution
- Enforces SO(3) equivariance so only interested in rigid bodies

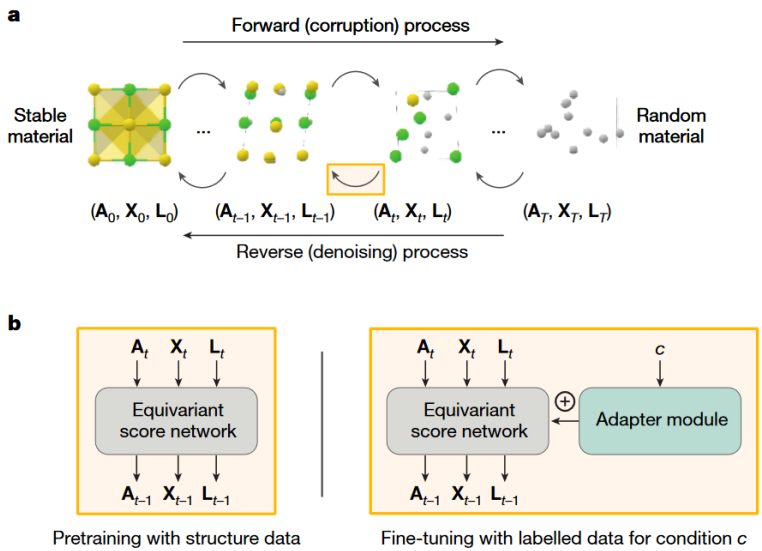


Example: matterGen

Let's throw some serious Microsoft money at the same problem

- Trained on a dataset that is $> 10\times$ larger than PXRDnet using serious Microsoft compute
- In most respects, in common with PXRDnet
- Can be conditioned on many different labels. Has to be trained, but the foundation model means that it can be trained for the new task on a small amount of data
- At the time of writing of the MS could only be conditioned on scalars. I'm sure that has changed.

It works!



What I learned about AI - Summary

- Diffusion models are the sh$%t
- Emerging foundation models built on large, diverse, datasets are adaptable for many task with modest retraining

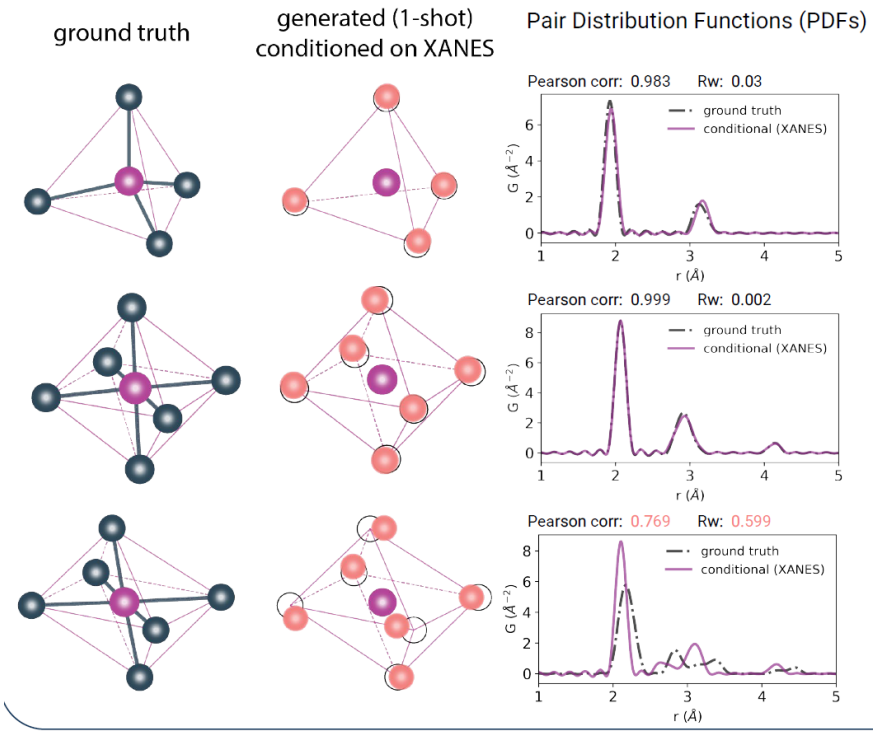
They work

Local structure from XANES with a diffusion model

Solving local structure from XANES (not XAFS!)

- Work of Aniv Ray and Tina Na Narong
- Collaboration with Steven Torrisi and Hod Lipson
- Use a diffusion model to extract local structure (with some surprises)

Please see Tina's poster



What I learned about materials by applying AI to structure solution

What I learned about materials by applying AI to structure solution

Structure solution

Structure of the problem

- Input is data, a diffraction pattern or PDF
- Output is a structure
- To train the model we need to tell it when it got the structure right vs wrong
- We know the structure, but we need a categorical success/failure measure
- How do we do that?

Structure solution

How do we assess if a structure solution is right?

- Need a measure of similarity between structures
- Need a threshold when structures are sufficiently similar to determine them as being "the same" and so "the structure is solved"

Structure solution

How do we assess if a structure solution is right?

- Need a measure of similarity between structures
- Need a threshold when structures are sufficiently similar to determine them as being "the same" and so "the structure is solved"
- btw, this problem affects Materials Discovery. MatterGen talks about SUN materials, "stable", "unique", "new" (i.e., not the same as a known structure).

When are structures the same problem (WASTS problem)

But wait, there are many papers in the literature that do this without much discussion. How do they do it?

- They use a structure similarity metric in pymatgen "structureMatcher"
- But this appears not to have been heavily validated. We are taking on blind faith that is working as hoped
- Some chemists seem to disagree

When are structures the same problem (WASTS problem)

But wait, there are many papers in the literature that do this without much discussion. How do they do it?

- They use a structure similarity metric in pymatgen "structureMatcher"
- But this appears not to have been heavily validated. We are taking on blind faith that is working as hoped
- Some chemists seem to disagree

The structure definition problem

What is a structure?

The structure definition problem

Online Dictionary of Crystallography Navigation Search Online Dictionary of Search

Crystal structure

Actions

From Online Dictionary of Crystallography

Structure cristalline (*Fr*). Kristallstruktur (*Ge*). Struttura cristallina (*It*). 結晶構造 (*Ja*). Estructura cristalina (*Sp*).

A **crystal structure** is a crystal pattern consisting of atoms.

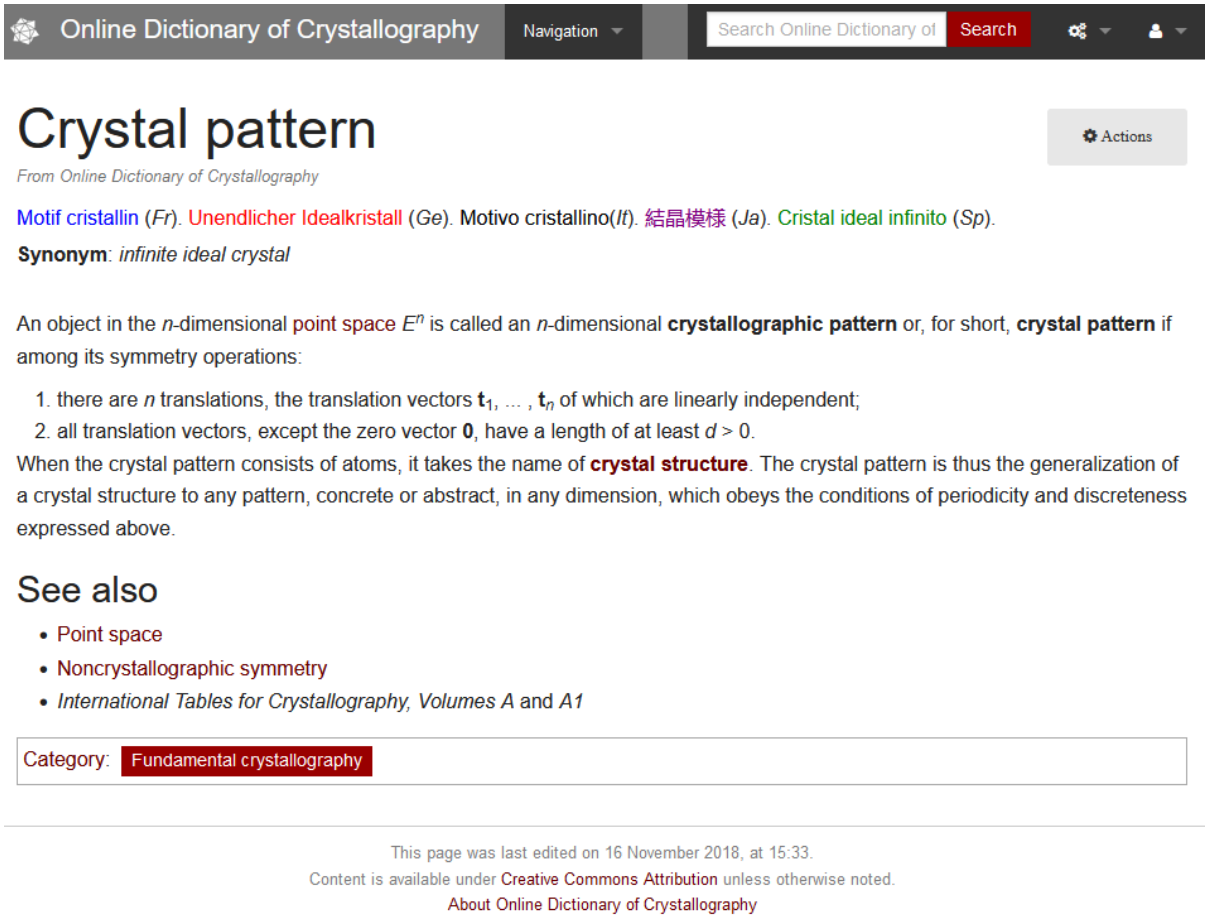
See also

- Point space
- Noncrystallographic symmetry
- *International Tables of Crystallography, Volumes A and A1*

Category: Fundamental crystallography

This page was last edited on 9 November 2017, at 17:29.
Content is available under Creative Commons Attribution unless otherwise noted.
About Online Dictionary of Crystallography

The structure definition problem



Online Dictionary of Crystallography | Navigation | Search Online Dictionary of | Search

Crystal pattern

Actions

From Online Dictionary of Crystallography

Motif cristallin (*Fr*). Unendlicher Idealkristall (*Ge*). Motivo cristallino(*It*). 結晶模様 (*Ja*). Cristal ideal infinito (*Sp*).

Synonym: *infinite ideal crystal*

An object in the *n*-dimensional point space E^n is called an *n*-dimensional **crystallographic pattern** or, for short, **crystal pattern** if among its symmetry operations:

1. there are *n* translations, the translation vectors $\mathbf{t}_1, \ldots, \mathbf{t}_n$ of which are linearly independent;
2. all translation vectors, except the zero vector **0**, have a length of at least $d > 0$.

When the crystal pattern consists of atoms, it takes the name of **crystal structure**. The crystal pattern is thus the generalization of a crystal structure to any pattern, concrete or abstract, in any dimension, which obeys the conditions of periodicity and discreteness expressed above.

See also

- Point space
- Noncrystallographic symmetry
- *International Tables for Crystallography, Volumes A* and *A1*

Category: Fundamental crystallography

This page was last edited on 16 November 2018, at 15:33.
Content is available under Creative Commons Attribution unless otherwise noted.
About Online Dictionary of Crystallography

The structure definition problem

What is a structure, chemist's definition?

- Definition based on space-group classification
- Structures that change without breaking symmetry are the same structure

The structure definition problem

What is a structure, chemist's definition?

- Definition based on space-group classification
- Structures that change without breaking symmetry are the same structure
- In this world it is easy to say that two structures are the same (continuous crystallography)
- much harder to say whether two unequal structures are the same in the chemist's sense

The structure definition problem

Case 1: thermal expansion

- Take material
- Warm it up
- The unit cell expands (without breaking (point) symmetry)
- The structure hasn't changed (according to Chemists)
- The structure has changed (according to continuous crystallography measures like PDD)
- Actually, it has changed, all the bonds are different lengths which affects properties

The structure definition problem

Case 2: miniscule phase transition

- Take material
- Move one atom off a special position by a tiny amount
- The structure has changed (according to Chemists), the space group has changed
- The structure has changed (according to continuous crystallography measures like TPP), but by a miniscule amount

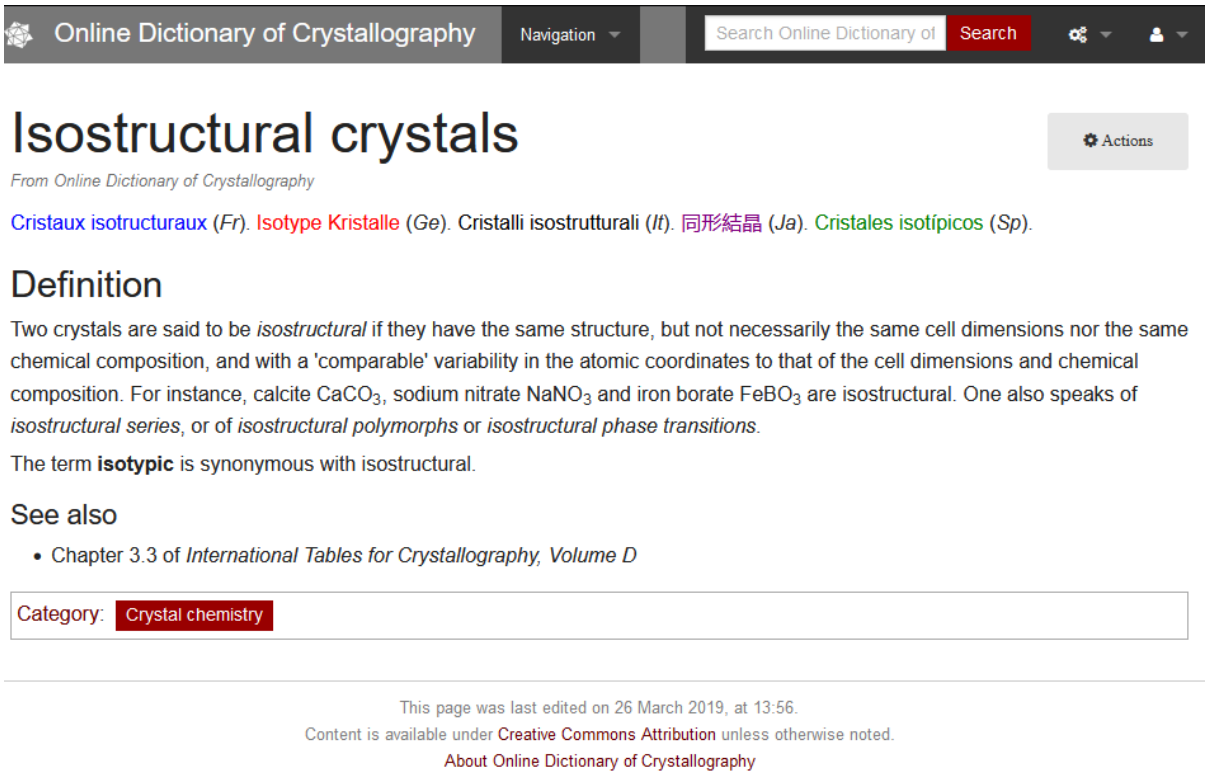
The structure definition problem

Case 2: miniscule phase transition

- Take material
- Move one atom off a special position by a tiny amount
- The structure has changed (according to Chemists), the space group has changed
- The structure has changed (according to continuous crystallography measures like TPP), but by a miniscule amount

Case 2 the continuous crystallography thinks the materials are closer than Case 1, but chemists think otherwise

The structure definition problem



Online Dictionary of Crystallography | Navigation | Search Online Dictionary of | Search

Isostructural crystals

Actions

From Online Dictionary of Crystallography

Cristaux isotructuraux (*Fr*). Isotype Kristalle (*Ge*). Cristalli isostrutturali (*It*). 同形結晶 (*Ja*). Cristales isotípicos (*Sp*).

Definition

Two crystals are said to be *isostructural* if they have the same structure, but not necessarily the same cell dimensions nor the same chemical composition, and with a 'comparable' variability in the atomic coordinates to that of the cell dimensions and chemical composition. For instance, calcite $CaCO_3$, sodium nitrate $NaNO_3$ and iron borate $FeBO_3$ are isostructural. One also speaks of *isostructural series*, or of *isostructural polymorphs* or *isostructural phase transitions*.

The term **isotypic** is synonymous with isostructural.

See also

- Chapter 3.3 of *International Tables for Crystallography, Volume D*

Category: Crystal chemistry

This page was last edited on 26 March 2019, at 13:56.
Content is available under Creative Commons Attribution unless otherwise noted.
About Online Dictionary of Crystallography

The structure definition problem

Summary

As a community we need more practical definitions of what it means for two structures to be the same or similar

Ways to compare structures: structure metrics

How do crystallographers compare structures?

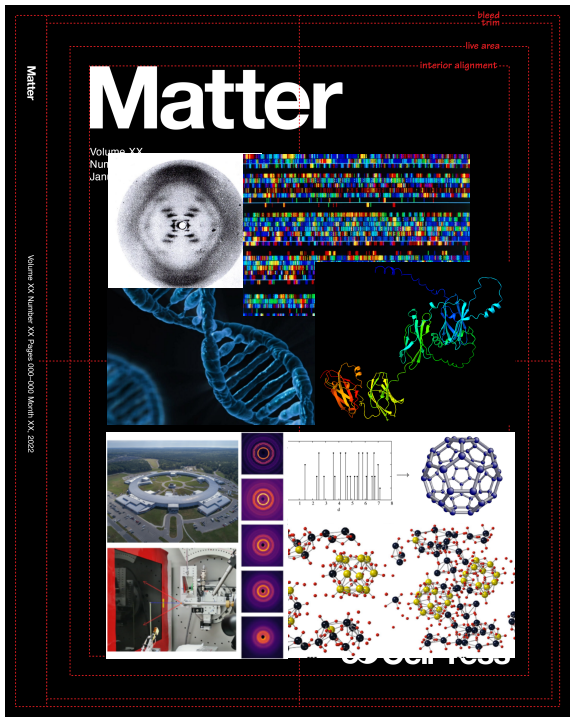
- Gold standard is Rw between measured and computed structure-factors or diffraction data coupled with crystallographic expertise
- CIFcheck (does some symmetry and chemistry common sense)

Ways to compare structures: continuous structure metrics

continuous structure representations

- PDF as a "materials genome", a 1D function that "codes for" 3D structure
- AMD (reduce the PDF to a scalar) - metric of Vitaliy Kurlin (U Liverpool)
- PDD (matrix of partial PDFs)
- smooth overlap of atomic positions (SOAP)
- atom-centered symmetry functions (ACSF)

AMD and PDD are not experimentally accessible but are valid metrics, and PDD is complete



Comparing StructureMatcher and PDF

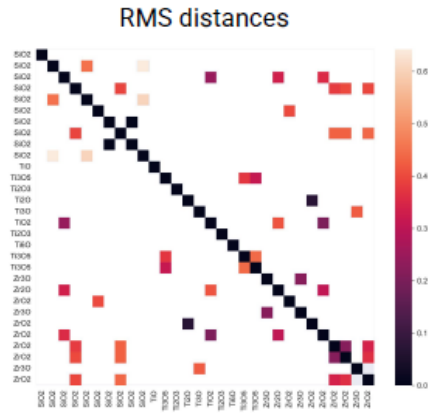
Baseline: pymatgen.StructureMatcher module

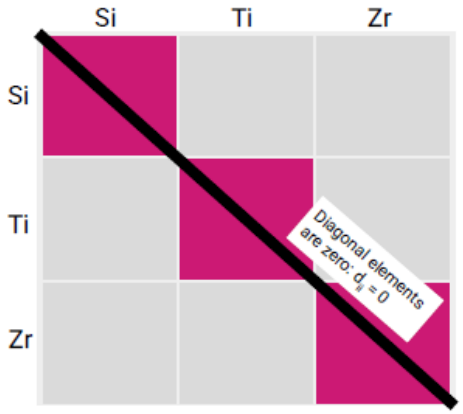
- Pymatgen: open-source Python library for materials analysis
- StructureMatcher module compares structures based off their atomic coordinates.

$$distance(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

Distance measure:

- Root-mean-square distance between the atomic positions of two structures
- Ignores atomic species, only accounts for atomic sites and lattice parameters
- Permutes atomic sites to minimize RMS, account for rotations & symmetries





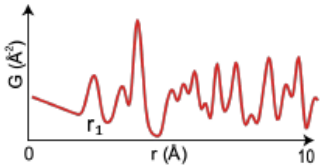
Pair Distribution Function (PDF):
"histogram" of bond lengths in the structure, experimentally measurable from diffraction data.

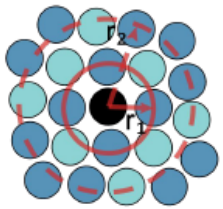
Distance measures:

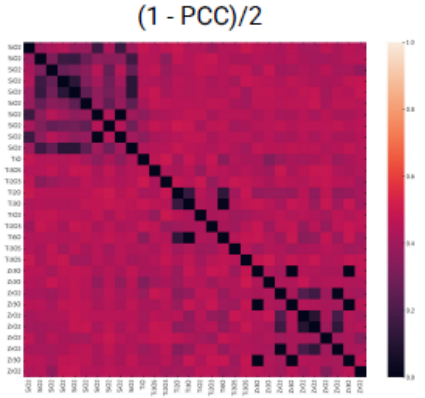
- Pearson correlation coefficient (PCC)
- Symmetrized R_w

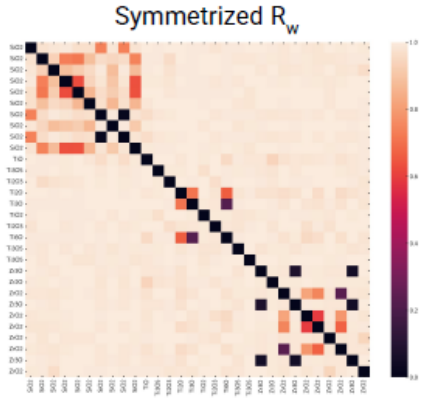
$$R_w = \frac{\sum_i (x_i - y_i)^2}{\sum_i (\vec{x}_i \vec{y}_i)}$$

**Pairs of PDFs are "morphed" before computing PCC and R_w. This accounts for differences in scale between structures, which affects R_w, PCC greatly.*









Work of Tina Na Narong, Zoe Zachko and Andrew Yang, Collaboration with Steven Torrisi

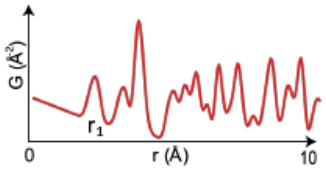
Comparing PDF and PDD/AMD

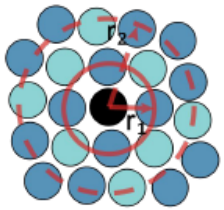
Pair Distribution Function (PDF):
"histogram" of bond lengths in the structure, experimentally measurable from diffraction data.

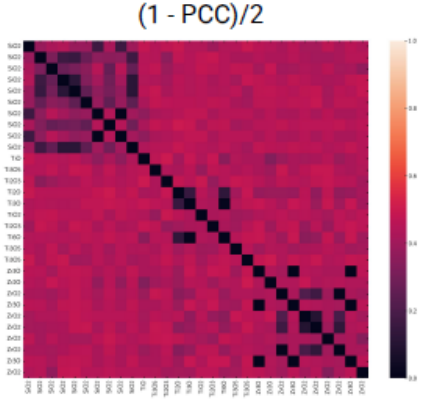
Distance measures:
- Pearson correlation coefficient (PCC)
- Symmetrized R_w

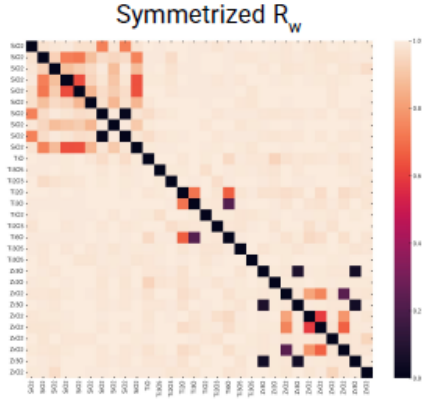
$$R = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}}$$

**Pairs of PDFs are "morphed" before computing PCC and R_w. This accounts for differences in scale between structures, which affects R_w PCC greatly.*



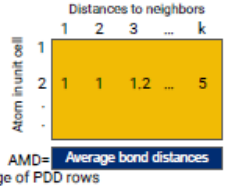






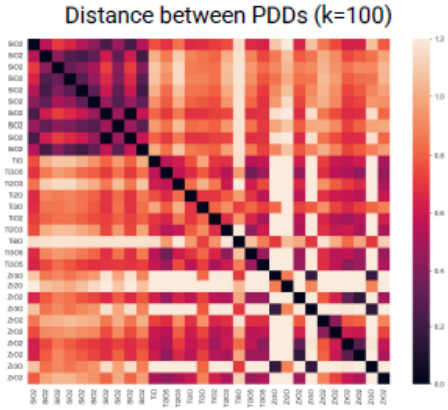
Pointwise Distance Distribution (PDD)
- Matrix of interatomic distances in the structure (isometry invariant)
- Each row of a PDD matrix lists *k* shortest distances (ascending order) from an atom in the unit cell to other atoms

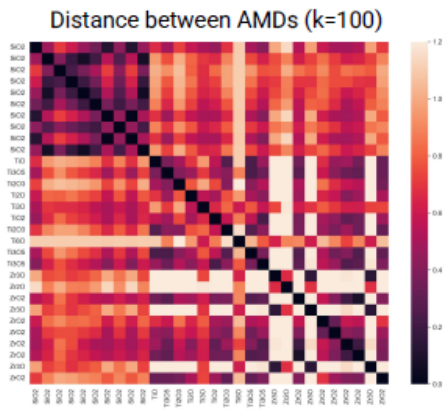
Distance measure: Earth's Mover Distance (EMD)
→ treat PDD matrix elements as "piles" of Earth
→ solve a linear programming problem to quantify the minimum "mass transport work" needed to make the two PDDs identical



Average Minimum Distance (AMD)
- AMD is an average of PDD: vector of *k* average shortest distances

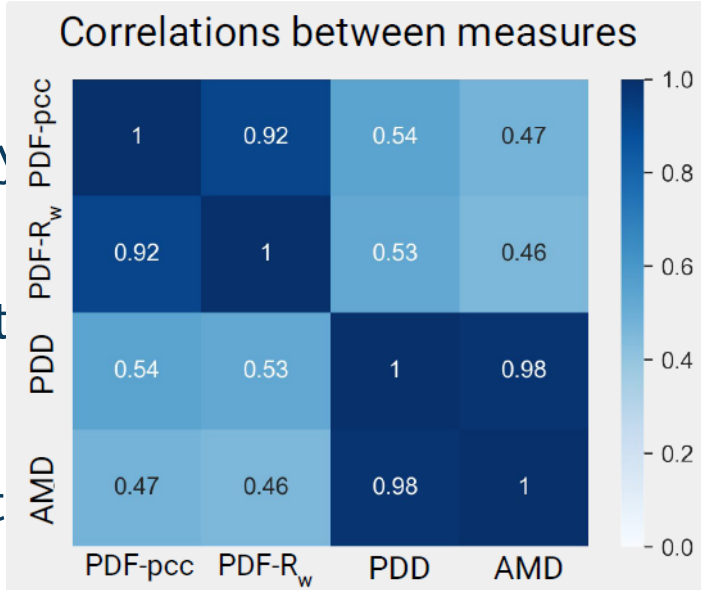
Distance measure: Chebyshev distance $D(x,y) = \max_i |x_i - y_i|$





Comparing structure metrics

- different similarity metrics differ in which structures they think are similar
- Which is the rightest
- We don't know
- How similar of result do they provide?



Correlations with StructureMatcher
(only distances below threshold included)

PDF-pcc	0.72
PDF-R$_w$	0.83
PDD	0.63
AMD	0.55

Summary

- Diffusion models are the s$%t
- They can be used to make really great foundation models
- As with LLMs, we will likely all be using these in the future
- But:

Summary

- Diffusion models are the s$%t
- They can be used to make really great foundation models
- As with LLMs, we will likely all be using these in the future
- But: be careful what you wish for...the answers are only as good as the question that was asked
- Bias in the data, bias in the network/noising protocol, choice of success metric
- These things are still being figured out

With GenAI, the problem is the problem, not the solution

Acknowledgements

- My current and former students and post-docs
- Beamline and software teams
- Collaborators

- Funding (DOE-BES, NSF-DMR, TRI, Columbia-DSI)
- The Facilities!