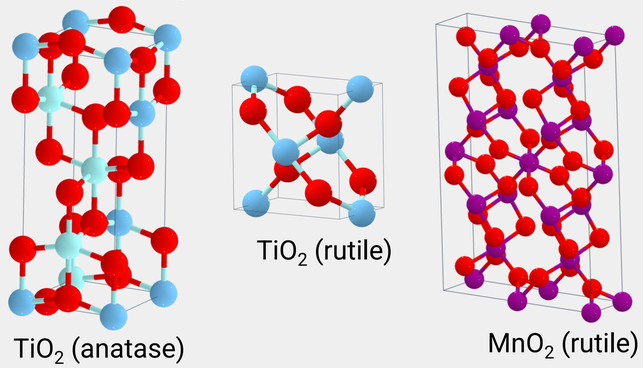
How to Tell Structures Apart: Data-Driven Analysis of Structural Similarity Measures

Tanaporn Na Narong*[1], Zoe N. Zachko*[1], Andrew Yang*[1], Steven B. Torrisi[2] & Simon J. L. Billinge[1]

[1]Department of Applied Physics and Applied Mathematics, Columbia University, NY 10027
[2]Toyota Research Institute, Los Altos, CA 94022

Which structures are the same/different? And how different are they?



It depends... We can look at their compositions, atoms' positions, bond lengths, etc.

What if I give you 2 million structures?

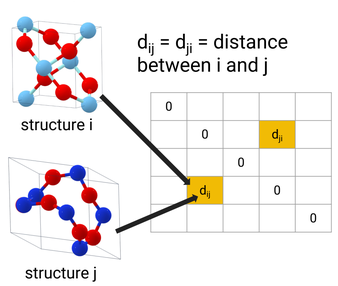
Problems

- Millions of "new" materials have been predicted by AI (e.g. Google DeepMind [1]). How do we know they are truly different from existing structures?
- How do we validate if two structures are the same? If they are different, how different are they?
  - Use case: compare ground truth vs AI/ML predicted structures

Goal: develop a data-driven framework to systematically analyze and compare different structural similarity measures

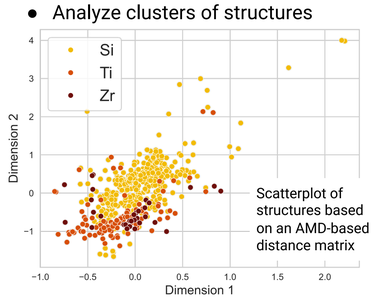
Methods

1. Compute a distance matrix:



2. Do different measures agree?

- Compute correlations between the distance matrices
- Analyze clusters of structures



Datasets (preliminary): oxides with many polymorphs from Materials Project

Si Ti Zr + O

Distances between structures are based on:

- pymatgen.StructureMatcher (baseline)
- Pair distribution function (PDF)
- Pointwise distance distribution (PDD)
- Average minimum distance (AMD)

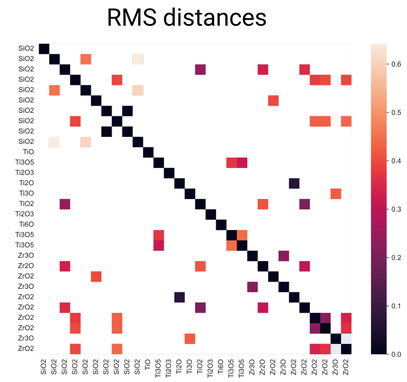
Baseline: pymatgen.StructureMatcher module

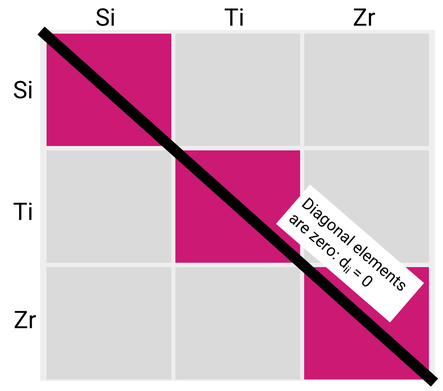
- Pymatgen: open-source Python library for materials analysis [2]
- StructureMatcher module compares structures based off their atomic coordinates.

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

Distance measure:

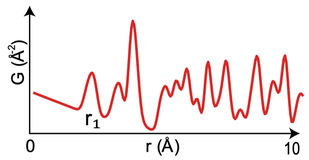
- Root-mean-square distance between the atomic positions of two structures
- Ignores atomic species, only accounts for atomic sites and lattice parameters
- Permutes atomic sites to minimize RMS, account for rotations & symmetries

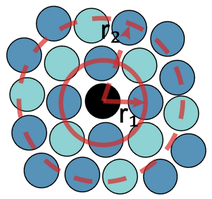




Pair Distribution Function (PDF):

"histogram" of bond lengths in the structure, experimentally measurable from diffraction data.



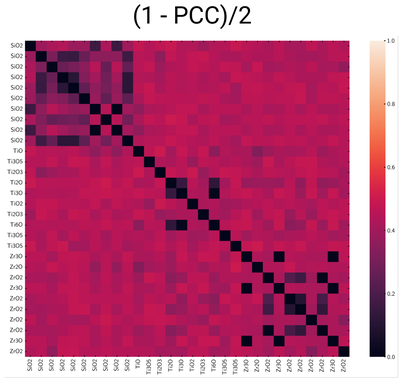


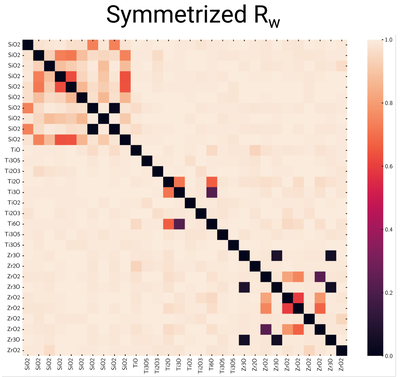
Distance measures:

- Pearson correlation coefficient (PCC)
- Symmetrized R_w

$$R_{w,sym}(\vec{x}, \vec{y}) = \frac{\sum_i (x_i - y_i)^2}{\sqrt{\sum_i x_i^2 \sum_i y_i^2}}$$

**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.*

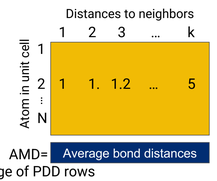




Pointwise Distance Distribution (PDD)

- 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

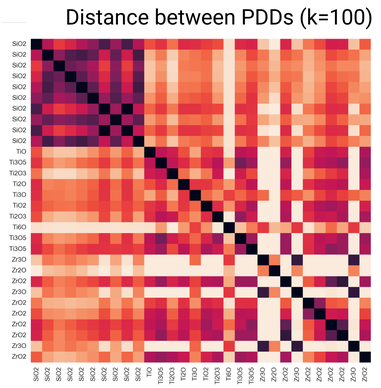
Distance measure: Earth's Mover Distance (EMD)
→ treat PDD matrix elements as "piles" of Earth
→ solve a linear programming problem to quantify the minimum "mass transport work" needed to make the two PDDs identical

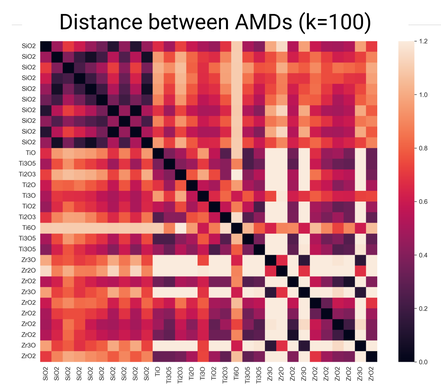


Average Minimum Distance (AMD)

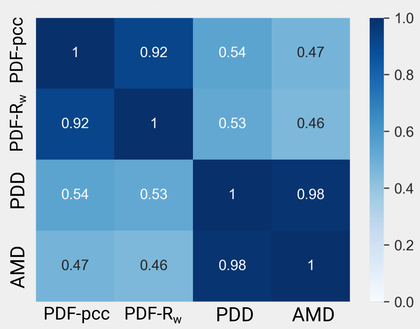
- AMD is an average of PDD: vector of *k* average shortest distances

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





Correlations between measures



Correlations with StructureMatcher (only distances below threshold included)

PDF-pcc	0.72
PDF-R_w	0.83
PDD	0.63
AMD	0.55

Summary & Outlook

- Structures within the same chemical group (Si/Ti/Zr) show higher similarity than across different groups
- Ti-O structures tend to be more similar to Zr-O than Si-O, but we need a larger sample size to statistically verify this trend
- Similarity measures that are based on the same structural descriptors are highly correlated (PDF-PCC and PDF-R_w / PDD and AMD)
- Future work: incorporate larger datasets, cluster analysis, and additional similarity measures

References

1. Merchant, A., Batzner, S., Schoenholz, S. S., Aykol, M., Cheon, G., & Cubuk, E. D. (2023). Scaling deep learning for materials discovery. Nature, 624(7990), 80-85.
2. Ong, S. P., Richards, W. D., Jain, A., Hautier, G., Kocher, M., Cholia, S., ... & Ceder, G. (2013). Python Materials Genomics (pymatgen): A robust, open-source python library for materials analysis. Computational Materials Science, 68, 314-319.
3. Billinge, S. J. (2019). The rise of the X-ray atomic pair distribution function method: a series of fortunate events. Philosophical Transactions of the Royal Society A, 377(2147), 20180413.
4. Widdowson, D., & Kurlin, V. (2021). Pointwise distance distributions of periodic point sets. arXiv preprint arXiv:2108.04798.