Exploring Material Similarity using Graph-Based Crystal Structure Analysis and Machine Learning

Karen Cao

Background

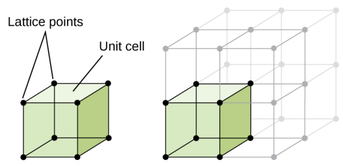
https://icsd.products.fiz-karlsruhe.de/

Inorganic Crystal Structure Database (ICSD)

- World's largest database for inorganic crystal structures
- 200,000+ compounds
- Provides Crystallographic Information File (CIF) for each crystal structure with unit cell, atomic coordinates, space group, and etc. information

Crystal Structures

- 3D arrangement of atoms, molecules, or ions in a crystalline solid
- Spatial arrangement determined by interactions between the atoms such as bonds
- Unit Cell - 3D structure in the form of a parallelepiped that is repeated
- Lattice Point - reference point that is repeated within the crystal structure with identical surroundings



https://courses.lumenlearning.com/suny-chem-atoms-first/chapter/lattice-structures-in-crystalline-solids-2/

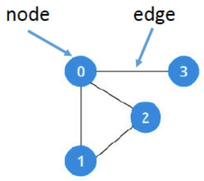
Space Groups

- Composed of Bravais type and symmetry operations that describe the arrangement of atoms in unit cell of a crystal structure
- Hermann-Mauguin (HM) symbols
- Bravais types
  - P – Primitive
  - I – Body Centered
  - F – Face Centered
- Symmetry Operations
  - Rotations
  - Reflections
  - Inversions
  - Glide Plane/Screw Axis
- Each space group is associated with a unique number (230 space groups)

http://www.chem.latech.edu/~upali/chem281/281lab/X-ray/XRD.htm

https://youtu.be/NaRjCfqT9GE

Space Group - I 4 m m
I – Body Centered
4 - 4-fold rotational symmetry axis, crystal looks same after 90 degree rotation
m – vertical mirror plane
m – horizontal mirror plane



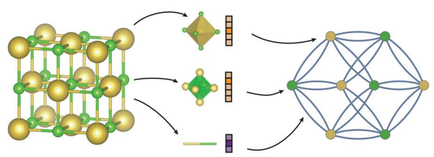
https://andysbrainbook.readthedocs.io/en/latest/FunctionalConnectivity/CONN_ShortCourse/CONN_AppendixA_GraphTheory.html

Graphs

- Mathematical structures used to model relationships
- Composed of nodes and edges
- Nodes represent entities or objects
- Edges represent interactions or connections between nodes

Objective

- Investigate the types of crystal structures already discovered and how they are distributed
- Represent crystal structures as graphs in order to create comparisons and place them in clusters or communities



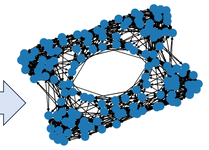
https://doi.org/10.5220/0005209202710278

Methods

Transforming crystal structures into graphs

- Takes in a CIF file as input and returns a graph constructed with the JARVIS and DGL python libraries
- Main functions
  - Nearest_neighbor_edges - k-NN (k-nearest neighbor) edge list
  - Radius_graph - edge list based on a specified cutoff radius after several conversions and calculations
- Created by NIST researchers Dr. Kamal Choudhary and Brian DeCost

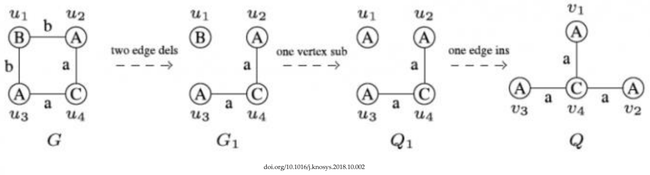
Top of CIF File



Graph visualization with NetworkX Python Library

Measuring Similarity between Graphs

- Graph Edit Distance (GED) - minimum cost to transform one graph into another through a sequence of edit operations
- Edit operations - node and edge insertion, deletion, substitution
- The idea of using graph edit distance was inspired by NIST researcher Dr. Debra Audus



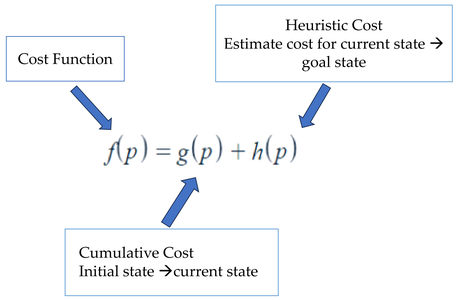
doi.org/10.1016/j.knosys.2018.10.002

Edit Operations:
1. Two edge deletions
2. One vertex substitution
3. One edge insertion

Each operation costs 1 → Total cost of 4

Calculating Graph Edit Distance (GED)

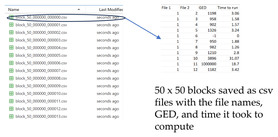
- Based on the A* Search Algorithm which finds shortest paths very efficiently
- Priority queue to store nodes and edges
- Elements are retrieved based on their "priority value" or order in which they are added (First in first out – FIFO)
- Apply each possible operation to create a new graph state and calculate the cost function f(p)
- Remove node with smallest f(p) from priority queue
- Used the NetworkX Python Library to calculated GED
- Returns the best GED calculation within a maximum number of seconds to execute
- Calculating GED is computationally expensive and time consuming so limiting run time was necessary
- Only able to calculate GEDs between 750 crystal structures



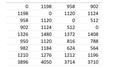
Creating a Distance Matrix

Ran the processes parallel on multiple nodes of a clustered server to increase efficiency of GED calculations

Submit batch scripts in the terminal to create 50 x 50 blocks of the distance matrix



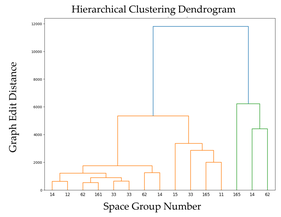
50 x 50 blocks saved as csv files with the file names, GED, and time it took to compute



Concatenated the block csv files into one dataframe and iterated through to create a 750 x 750 distance matrix

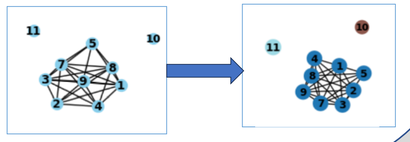
Hierarchical Clustering

- Groups similar crystal structures into groups/clusters
- How it works:
  - Treats each graph as a separate cluster
  - Repeatedly merges two clusters that are closest
  - Iterates until all clusters merged together

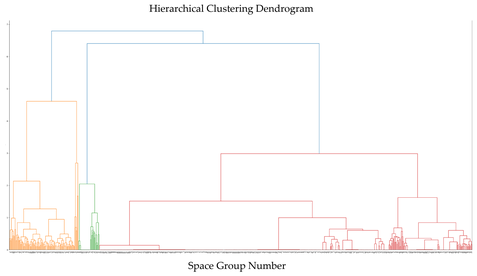


Community Detection

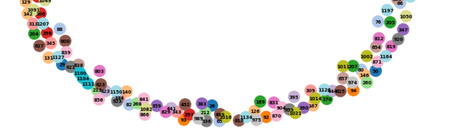
- Locates tightly connected nodes in a graph
- Graph with nodes for each crystal and edges constructed based on cutoff GED
- How it works:
  - Assigns different community to nodes
  - Considers each neighboring community for placing nodes
  - Node placed in neighbor community based on modularity
  - Modularity -



Results



CIF File Name	Space Group	Cluster
Cif Name 11	Space Group 14	1
Cif Name 279	Space Group 62	1
Cif Name 306	Space Group 14	1
Cif Name 369	Space Group 58	1
Cif Name 383	Space Group 176	1
Cif Name 392	Space Group 13	1
Cif Name 842	Space Group 58	1
Cif Name 904	Space Group 224	1
Cif Name 908	Space Group 224	1
Cif Name 923	Space Group 14	1
Cif Name 969	Space Group 157	1
Cif Name 981	Space Group 33	1
Cif Name 1021	Space Group 173	1
Cif Name 1049	Space Group 229	1
Cif Name 1050	Space Group 229	1
Cif Name 1051	Space Group 229	1
Cif Name 1083	Space Group 36	1
Cif Name 1091	Space Group 96	1
Cif Name 638	Space Group 7	2
Cif Name 651	Space Group 92	2
Cif Name 654	Space Group 2	2
Cif Name 800	Space Group 216	2
Cif Name 801	Space Group 216	2
Cif Name 812	Space Group 152	2
Cif Name 813	Space Group 61	2
Cif Name 376	Space Group 33	3
Cif Name 390	Space Group 62	3



Community	Nodes with CIF Name Labels
Community 1	10
Community 2	11
Community 3	20
Community 4	25
Community 5	26
Community 6	29
Community 7	32
Community 8	35
Community 9	941, 942, 948, 17, 966, 33, 41, 43, 96
Community 10	50
Community 11	52
Community 12	60
Community 13	65
Community 14	66
Community 15	76
Community 16	78
Community 17	79
Community 18	82
Community 19	84
Community 20	88
Community 21	91
Community 22	92
Community 23	93
Community 24	94
Community 25	95
Community 26	102
Community 27	104
Community 28	110
Community 29	123
Community 30	131
Community 31	140
Community 32	165
Community 33	167
Community 34	169
Community 35	170
Community 36	172
Community 37	176
Community 38	204
Community 39	205
Community 40	207
Community 41	209
Community 42	210
Community 43	211
Community 44	212

Future Work

- Closely investigate the clusters and communities to understand the types of structures which have not been investigated as much, helping facilitate the generation of new materials
- How has this distribution has changed over time?
- Continue calculating the graph edit distance for a complete 200,000 x 200,000 distance matrix

References

Abu-Aisheh, Z., Raveaux, R., Ramel, J., & Martineau, P. (2015). An Exact Graph Edit Distance Algorithm for Solving Pattern Recognition Problems. *International Conference on Pattern Recognition Applications and Methods*. https://doi.org/10.5220/0005209202710278

Jayawickrama, T. D. (2021, December 29). Community Detection Algorithms. Towards Data Science. *Medium*. _https://towardsdatascience.com/community-detection-algorithms-9bd8951e7dae

Patlolla, C. R. (2022, August 15). Understanding the concept of Hierarchical clustering Technique. Towards Data Science. *Medium*. https://towardsdatascience.com/understanding-the-concept-of-hierarchical-clustering-technique-c6e8243758ec

Sands, D. E. (2012). *Introduction to Crystallography*. Courier Corporation.

Xie, T., & Grossman, J. C. (2018). Crystal Graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Physical Review Letters*, *120*(14). https://doi.org/10.1103/physrevlett.120.145301

2.3. Clustering. (n.d.). Scikit-learn. https://scikit-learn.org/stable/modules/clustering.html#hierarchical-clustering

Communities — NetworkX 3.1 documentation. (n.d.). https://networkx.org/documentation/stable/reference/algorithms/community.html

graph_edit_distance — NetworkX 3.1 documentation. (n.d.). https://networkx.org/documentation/stable/reference/algorithms/generated/networkx.algorithms.similarity.graph_edi t_distance.html

Acknowledgements

- Mentor: Dr. William Ratcliff
- Special thanks to Mr. Paul Kienzle
- NCNR Mentors: Dr. Paul Butler, Mr. Jeff Krzywon, Dr. Yun Liu
- CHRNS Sponsors: Julie Borchers, Leland Harriger