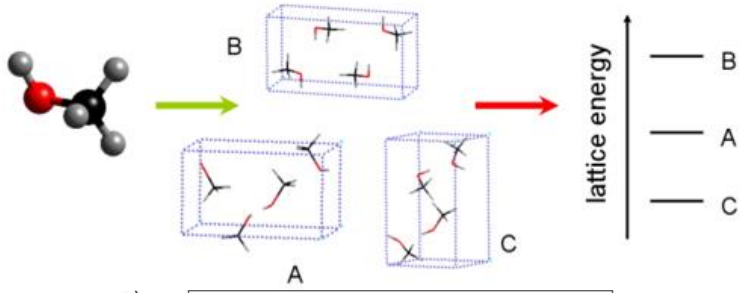
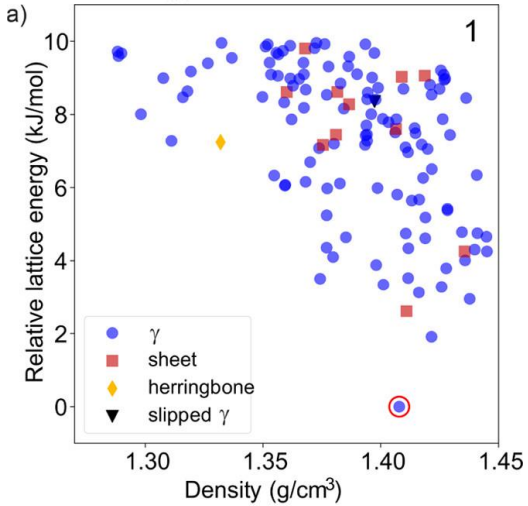
FONS Datathon

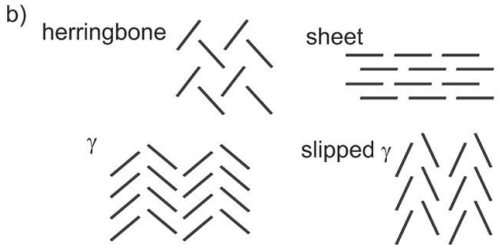
AI to speed up crystal structure prediction

Motivation

- Crystal structure prediction (CSP) is the first principles calculation of the packing of molecules in the solid-state
- From materials to drug discovery, performance is dependent on crystal packing(1)
- A recent venture: Digital navigation of energy–structure–function maps for hydrogen-bonded porous molecular crystals
- CSP is computationally expensive
- The Cambridge structural database (CSD) contains experimentally obtained crystal structures of many types of molecules





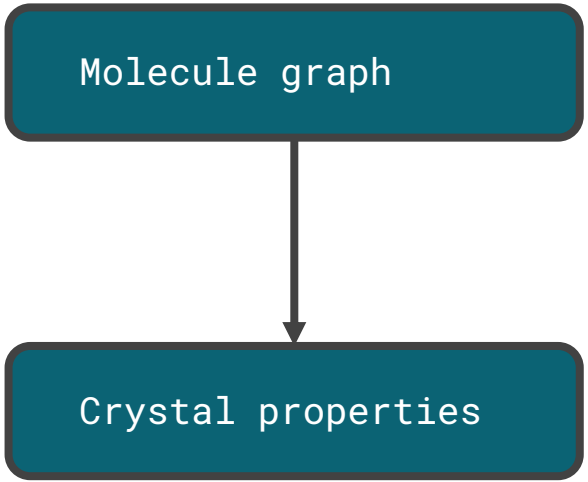


Aim to use this data to improve CSP methods

Goals of the Datathon

From descriptors of molecules to properties of crystal structures

- Given the data provided:
  - Crystal structures packing a single molecule from the CSD
  - Descriptors of the single molecule
  - Properties of the crystal structure and contacts within it
    - Density/packing efficiency
    - What contacts a molecule forms in the solid-state
    - Packing shell
- **Learn from 2D molecular graph to crystal structure properties**



Dataset size: 26,911 entries
80:20 split into training and test set

Pre-reading

Tools:

- What is Z'?
- RDKit: Getting Started In Python
- https://scikit-learn.org/stable/
- Github guides
- Blind tests: CSP Blind Tests

Research articles:

- Space group selection for crystal structure prediction of solvates
- Which conformations make stable crystal structures?
- Data-efficient machine learning for molecular crystal structure prediction
- Large-Scale Computational Screening of Molecular Organic Semiconductors Using Crystal Structure Prediction

Ideas for Prediction Targets

1. Density and/or packing coefficient as a regression problem
2. Classify if a molecule will pack in a centrosymmetric symmetry
3. Predicting Z' (What is Z'?) as a regression problem
4. Predicting the space group of a molecule
5. Predict a molecule's likely contact types from the CSD contact types as a classification problem
6. Model the distribution of close contacts (cluster- and atom-based) in a packing shell as regression problem

Any and all represent a way to speed up CSP by narrowing the search space

We will clarify the weighting of these targets on the day

Quantifying Model Accuracy

Regression Problems

Mean absolute error: $\mathbf{MAE} = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n}$

```
from sklearn.metrics import mean_absolute_error
```

Classification Problems

F1 Score: `F1 = 2 * (precision * recall) / (precision + recall)` [A blog post]

```
from sklearn.metrics import f1_score
```