Import and Energy Minimization of Crystallography Open Database Entries using OpenMM

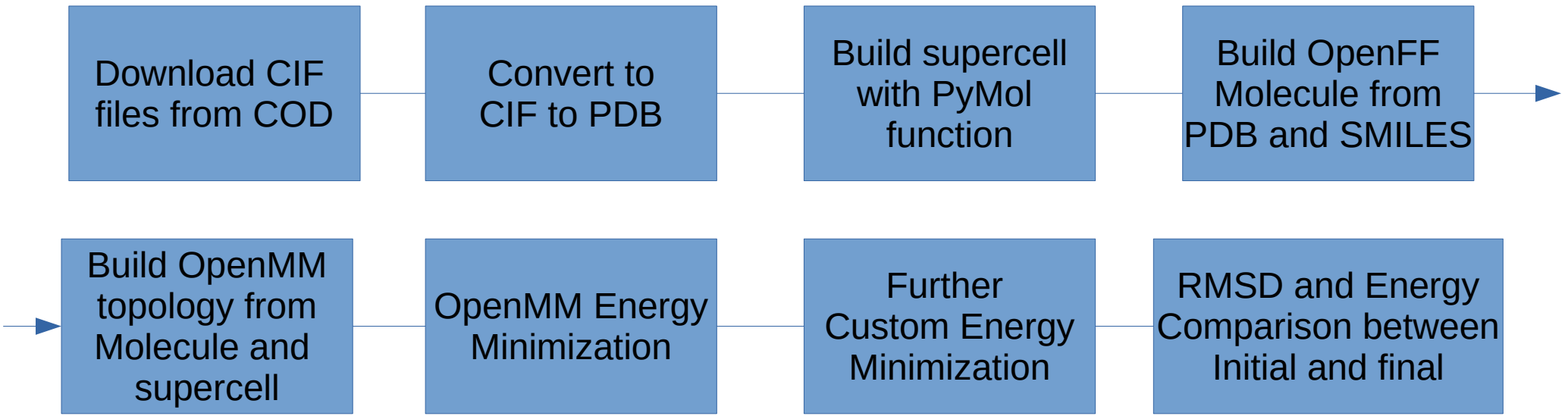
Sam Kennedy
Independent Study Fall 2021 for Dr. Michael Shirts

Import from Crystallography Open Database (COD) – Problem Statement

- Wanted to develop workflow to download and parse a subset of crystal structures from the Crystallography Open Database (COD) and perform energy minimization for various force fields
  - Use OpenMM as MD package
  - Used OpenFF force field (1.0.0 and 2.0.0)

Workflow

- Using Python 3.9 Scripting



Import from Crystallography Open Database (COD) - Issues

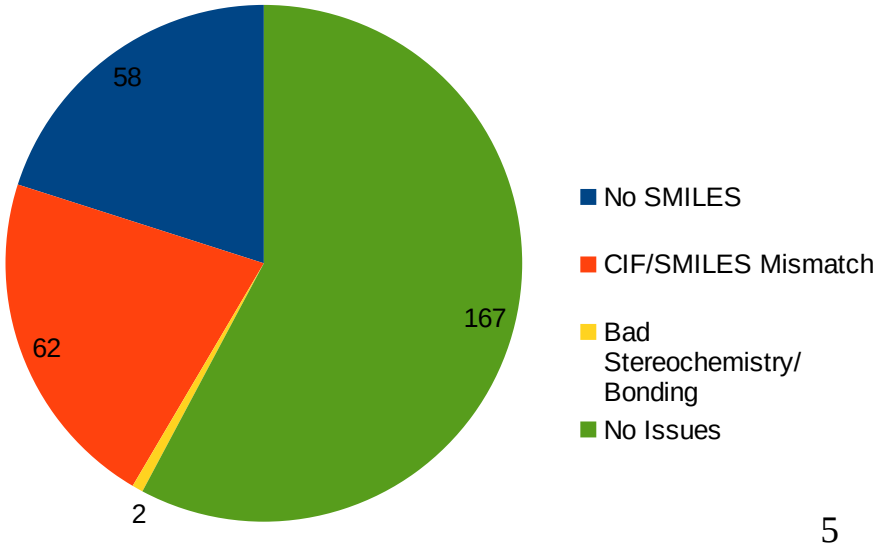
- No easy way to filter organic and non-organic crystals.
- Used DataWarrior copy of COD database which had organic/inorganic filtering capabilities
  - Manually selected a set of ~275 COD IDs for further investigation

Import from Crystallography Open Database (COD) - Issues

I. The SMILES database did not have a string for the COD entry of interest

II. There is a mismatch between the SMILES string and PDB file
   - Most of these errors come from CIF files that are missing atoms in their sets of coordinates

III. The stereochemistry is not fully defined or other faulty bonding information

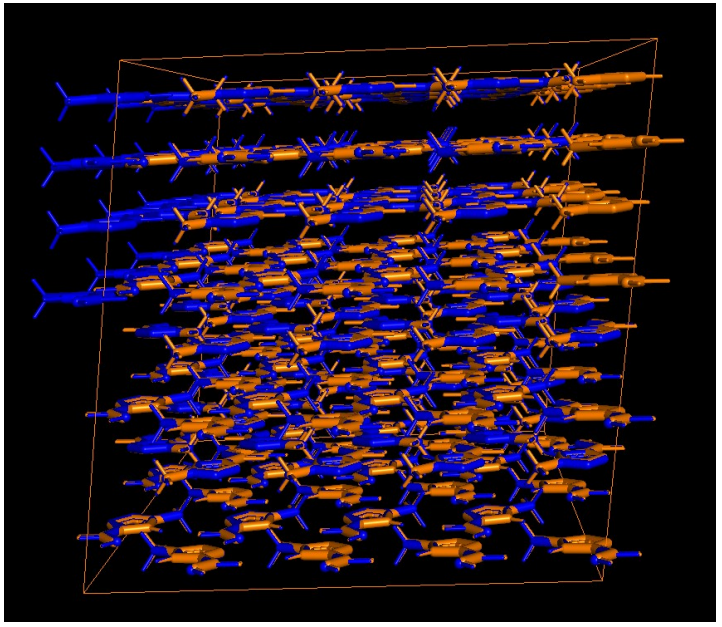


Energy Minimization - Workflow

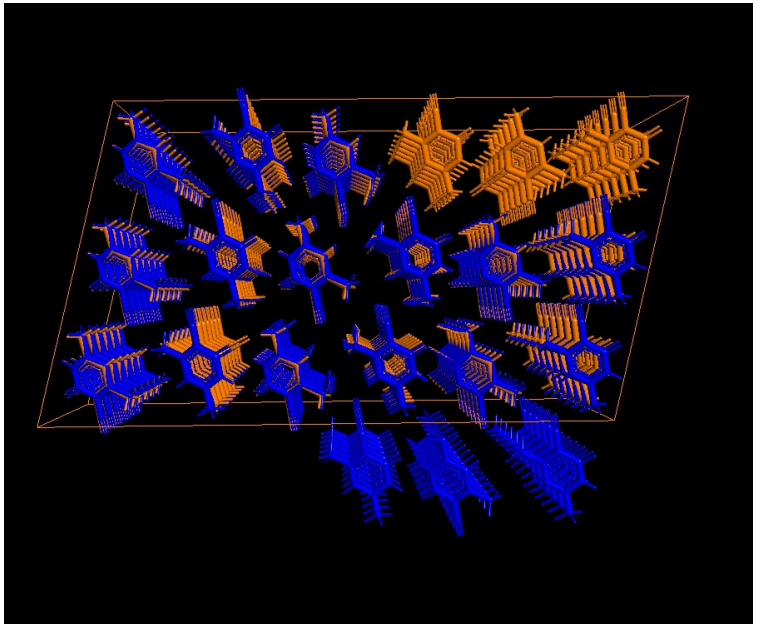
- Minimization in two parts:
  - Initial position minimization with built-in openMM energy minimizer
  - Simultaneous minimization of position and periodic boundary box vectors
    - Jacobian of atom positions taken as forces from simulation. Calculated by finite difference for box vectors.

Energy Minimization - Issues

- Translation of molecules across boundary
  - Orange is initial supercell, blue is minimized cell



COD ID: 1519825



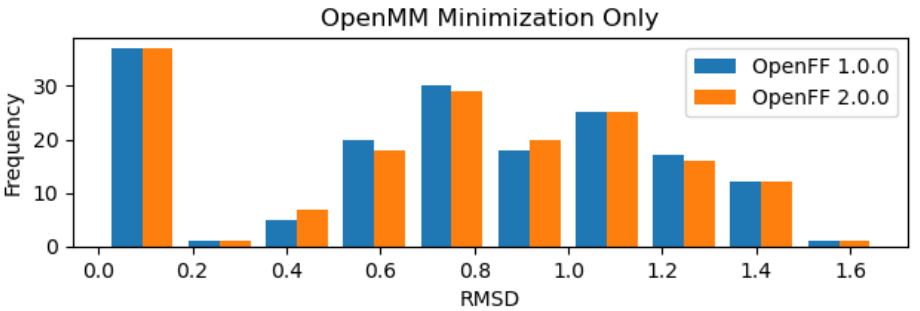
COD ID: 2203241

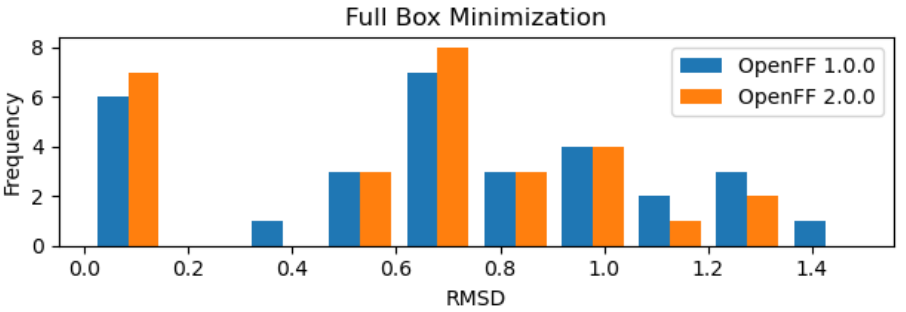
Energy Minimization - Issues

- Energy minimization algorithm fails for many systems due to periodic cell conditions not being in a reduced format during minimizer call.
  - Occasionally periodic boundary is reduced so that it is not larger than 1.8nm also causing an error.
- Modified periodic boundary conditions are not output correctly into the CRYST1 entry of OpenMM generated trajectory.

Energy Minimization – RMSD Results

- No major differences between 1.0 and 2.0
- Both minimization methods show high portion of large RMSD





	Average RMSD (OpenMM Minimization Only)	Average RMSD (Box Minimization)
OpenFF 1.0.0	0.731	0.695
OpenFF 2.0.0	0.726	0.607

Conclusions

- COD is a large database but lack robust search and filtering options
  - SMILES strings are generated for a portion, but not all, of the COD
- Energy minimizer needs additional optimization to avoid “reduced boundary vector” issue and to constrain molecules so they do not migrate across periodic boundary.

References

[1] qual5206, *qual5206/OpenFFCrystalBenchmarking*. 2021. Accessed: Dec. 15, 2021. [Online]. Available: https://github.com/qual5206/OpenFFCrystalBenchmarking

[2] "www.openmolecules.org." https://openmolecules.org/datawarrior/datafiles.html (accessed Dec. 15, 2021).

[3] N. M. O'Boyle, C. Morley, and G. R. Hutchison, "Pybel: a Python wrapper for the OpenBabel cheminformatics toolkit," *Chem. Cent. J.*, vol. 2, no. 1, p. 5, Dec. 2008, doi: 10.1186/1752-153X-2-5.

[4] *Open-Source PyMOL*. Schrodinger, Inc., 2021. Accessed: Dec. 15, 2021. [Online]. Available: https://github.com/schrodinger/pymol-open-source

[5] M. Quirós, S. Gražulis, S. Girdzijauskaitė, A. Merkys, and A. Vaitkus, "Using SMILES strings for the description of chemical connectivity in the Crystallography Open Database," *J. Cheminformatics*, vol. 10, no. 1, p. 23, Dec. 2018, doi: 10.1186/s13321-018-0279-6.