Genarris 2.0: A Random Structure Generator for Molecular Crystals

Noa Marom

Materials Science & Engineering

Carnegie Mellon University
Molecular Crystals

Used for *e.g.*, pharmaceuticals, organic electronics

Weak dispersion (van der Waals) interactions produce potential energy landscapes with many local minima close in energy

Molecular crystals often exhibit **polymorphism**, the ability of the same molecule to crystallize in several structures

Polymorphs may have different physical/chemical properties!

The challenge: given a 2D stick diagram of a molecule, predict all of its possible polymorphs

Requires searching a high-dimensional space with a high accuracy
Two Layers of Code

<table>
<thead>
<tr>
<th>Structure Prediction</th>
<th>Electronic Structure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Composition Structure</td>
<td>Energy Electronic properties</td>
</tr>
</tbody>
</table>

**Numeric atom-centered orbital basis sets**

**All flavors of semi-local and hybrid DFT**

**Pairwise (TS) and many-body dispersion (MBD) methods**

*Genarris* generates random structures of crystals of (semi-) rigid molecules


*GAtor* genetic algorithm (GA) for structure prediction of crystals of (semi-) rigid molecules


Genarris: Workflow

Genarris generates curated sets of random structures with physical constraints for:

- Initial populations for CSP algorithms
- Training sets for machine learning

**New features in Genarris 2.0:**

- Updated to Python 3 with MPI4Py parallelization
- Machine learned model for estimating the solid state molecular volume from the single molecule structure
- Generation in all compatible space groups with $Z=1$ including in special Wyckoff positions
- Fast hierarchical structure checks to prevent unphysically close contacts between molecules
- Special close contact settings for strong H-bonds
- Seamless execution of user-defined workflows
- New “Robust” workflow for clustering and down-selection
Genarris: Volume Estimation

Goal: estimate solid state molecular volume from single molecule structure

For 2,173 polymorph pairs extracted from CSD, the percent volume difference between polymorphs has a standard deviation of 2.95%.

Simple model: Monte Carlo estimation of volume occupied by vdw spheres + linear fit.
To improve the model add features: topological molecular fragments

Linear model coefficients found by Bayesian regression

Fragments that appear less than 30 times in the dataset are eliminated

64 unique fragments left

Top 5 fragments: HC, HO, OC, ON
The machine learned model based on MC+ molecular topological fragments gives a standard deviation of 2.93%.

Genarris: Structure Generation

A list of compatible space groups and Wyckoff positions is built

Compatible space groups for benzene with 6/mmm symmetry and Z=4:

Unlike general positions, only specific orientations are allowed on special Wyckoff positions
By default, atoms belonging to different molecules cannot be closer than \( Sr = 0.85 \) of the sum of their vDW radii.

Based on statistical analysis of the CSD, new defaults have been implemented for strong hydrogen bonds:

- \( OH\cdots O \):
  - \( Sr = 0.55 \) (25,294 interactions)

- \( NH\cdots O \):
  - \( Sr = 0.59 \) (17,189 interactions)

- \( OH\cdots F \):
  - \( Sr = 0.59 \) (4,801 interactions)

- \( OH\cdots N \):
  - \( Sr = 0.58 \) (20,874 interactions)

- \( NH\cdots N \):
  - \( Sr = 0.62 \) (15,247 interactions)

- \( NH\cdots F \):
  - \( Sr = 0.58 \) (14,627 interactions)
Genarris: Clustering and Down-Selection

Benzene with 4 molecules/unit cell

-Compute vector representation of molecular packing
-AP clustering and selection of 10% of structures
-Single point energy evaluation with PBE+TS
-AP clustering and selection of 10% with the lowest energy
-Geometry optimization using PBE+TS

Pool of diverse and low energy structures

The predicted volume agrees with experiment
The mean of the volume histogram is close to the target
Structures are generated in most of the allowed space groups with general and special positions

Diversity-based selection
Genarris: Clustering and Down-Selection

Benzene with 4 molecules/unit cell

The final pool provides a balance of energy and diversity.

The experimental structure is found in the final pool.
Generating Molecular Crystal Surfaces with Ogre

Ogre generates slab models of molecular crystal surfaces

The bulk crystal structure is cleaved along a user-defined Miller plane

A molecular graph representation is used to identify and repair cleaved molecules

User-defined number of layers and vacuum space are added

Ogre streamlines surface energy calculations and Wulff shape construction.

Strong H-bonds play an important role in the surface energy of aspirin.

The aspirin crystal shape is in good agreement with experiment.

Acknowledgements

Fall 2019

TBD: Extend GAtor and Genarris to more complex systems

Download GAtor, Genarris, and Ogre: www.noamarom.com