Read Me File for the Data Set: "Predicting Crystal Structures using Digital Alchemy Inverse Materials Design and the Random Forest Technique of Machine Learning"

Yina Geng, Greg van Anders, and Sharon C. Glotzer

Contact: Yina Geng yinageng@umich.edu

Description of the Data Set:

The data are the 13 target structures used in developing our model for predicting colloidal crystal structures from the geometries of particular shapes. The target structures are:

- simple cubic (SC),
- body-centered cubic (BCC),
- face-centered cubic (FCC),
- simple chiral cubic (SCC),
- hexagonal (HEX-1-0.6),
- diamond (D),
- graphite (G),
- honeycomb (H),
- body-centered tetragonal (BCT-1-1-2.4),
- high-pressure Lithium (Li),
- Manganese (beta-Mn)
- Uranium (beta-U),
- Tungsten (beta-W).

At least nine simulations were run on each of the target structures. All of the data are formatted as .pos files.

Research Background:

The data were generated as a part of a model in predicting crystal structures solely from knowledge about the colloidal particles without the need for simulations or experiments. We used the Digital Alchemy inverse materials design approach to find optimal and near-optimal hard, convex, colloidal, polyhedral shapes for 13 target structures. We then used the Random Forest technique from machine learning to classify these shapes. Of the 10 measures of shape, two – the dihedral angle (cos(d)) and the trace of the moment of inertia tensor (Tr(I)) – are sufficient to predict the self-assembly behavior of a shape. Our model can be used to inform experiments and select building blocks for self-assembling nanoparticle superlattices and colloidal crystals.

Methodology:

To construct the predictive model we performed Alchemical Monte Carlo (Alch-MC) simulations on the target structures. We placed a minimum of N = 100 particles in a periodic simulation box, with the exact number chosen to be a multiple of the number of particles in the unit cell of one of the 13 target structures. Particle shapes were initialized with as many as 64 vertices randomly generated to create a convex shape. Monte Carlo (MC) sweeps were performed to allow particle translations, rotations, and shape moves via vertex re-location.

What's Needed to Use the Data Set:

Each folder contains the Alch-MC result for one target structure. For example, the folder name "Li_V32_pf0.65" means that the target structure is Li, number of vertices is 32, and packing fraction is 0.65. We save one simulation result in a .pos file, which is dumped every 5e3 Monte Carlo sweeps. For each dump frame, it has simulation box information (learn more at http://hoomd-blue.readthedocs.io/en/stable/box.html), shape definition, and particles position and orientation information. We are interested in the shape definition in this project. The shape definition is in the format "def A "poly3d number_of_vertices x_y_z_coordinates_of_each_vertex color_of_shape"".

We provide a python script to extract all the particle vertices information. You can run it using python version > 3 with "python getShapeVerticesFromPos.py"

Note: In the machine learning model, we use the optimal particle shapes at the equilibrium state. We only use shapes from last 100 frames in each .pos file. We randomly select 1000 shapes for one target structure and this is the input to the machine learning model.

Suggested Citation to the Data:

Geng, Yina, van Andres, Greg, and Glotzer, Sharon C. (2018) "Predicting Crystal Structures using Digital Alchemy Inverse Materials Design and the Random Forest Technique of Machine Learning [dataset]" University of Michigan. https://deepblue.lib.umich.edu/data/concern/generic_works/6q182k84r?locale=en.