Using an Open Source Python Toolbox (Signac) to Manage High Dimensional Research Data

Chrisy Xiyu Du

Mechanical Engineering

My background

Assistant Professor in Mechanical Engineering (Jan. 2023)

PhD in the Glotzer Group at University of Michigan

Highly computation driven group with many open-source softwares

https://github.com/glotzerlab

Signac

Python based data-management software

Completely Open source

Can be installed with conda and pip

Slack community

Documentation: https://docs.signac.io/en/latest/

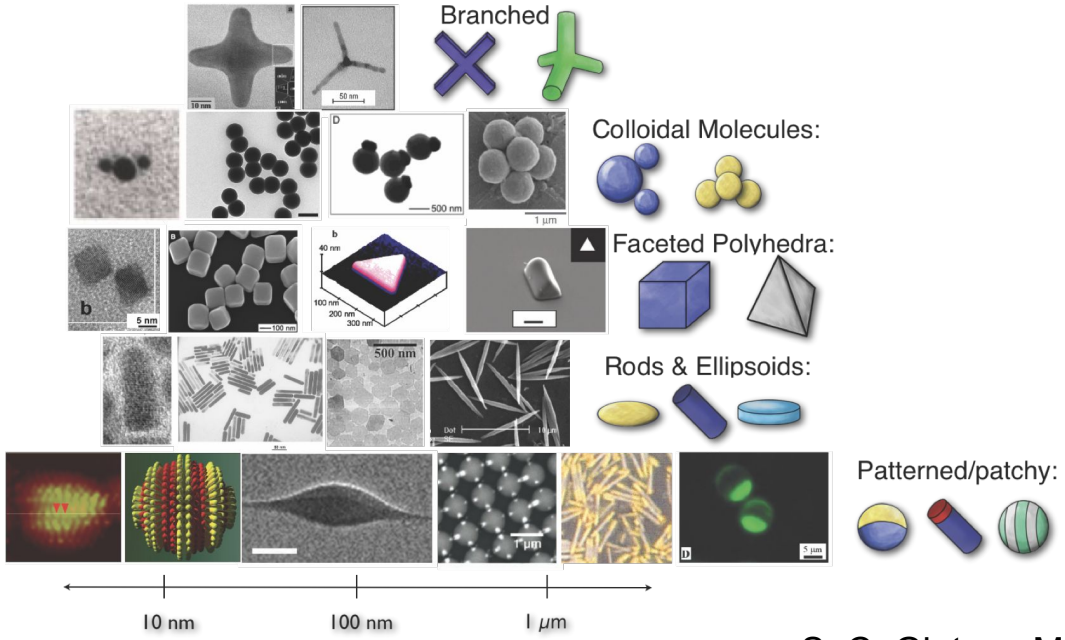
Why do I need it?

My research: use theory and computer simulation to design the next generation of soft materials.

Why do I need it?

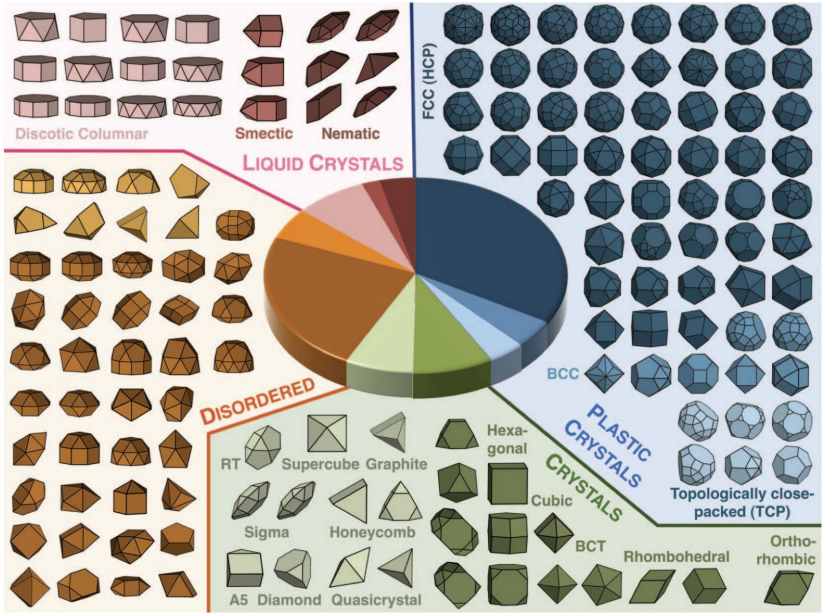
Movie: Big Hero 6 (gif by aviscranio.tumblr.com)

Large parameter space

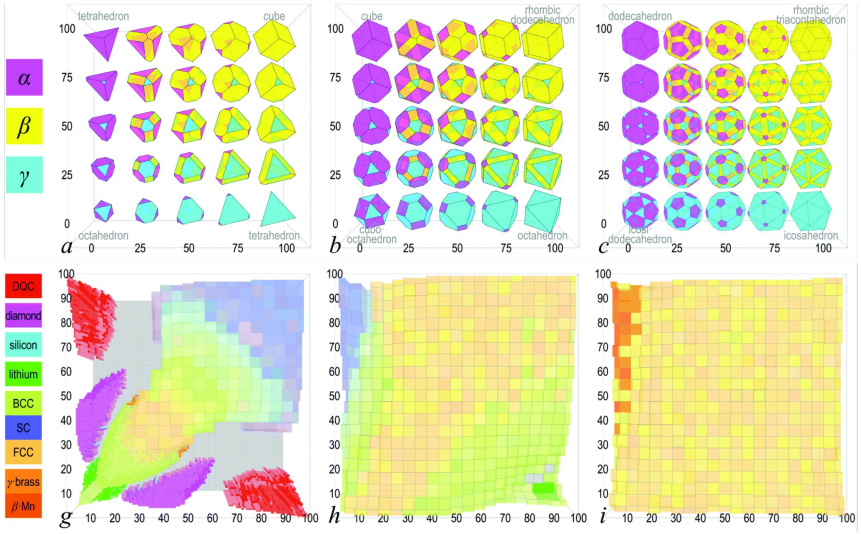


S. C. Glotzer, M. J. Solomon (2007)

Large parameter space



P. F. Damasceno, M. Engel, S. C. Glotzer, Science (2012)



D. Klotsa, E. R. Chen, M. Engel, and S. C. Glotzer, Soft Matter (2018)

What kind of data/metadata?

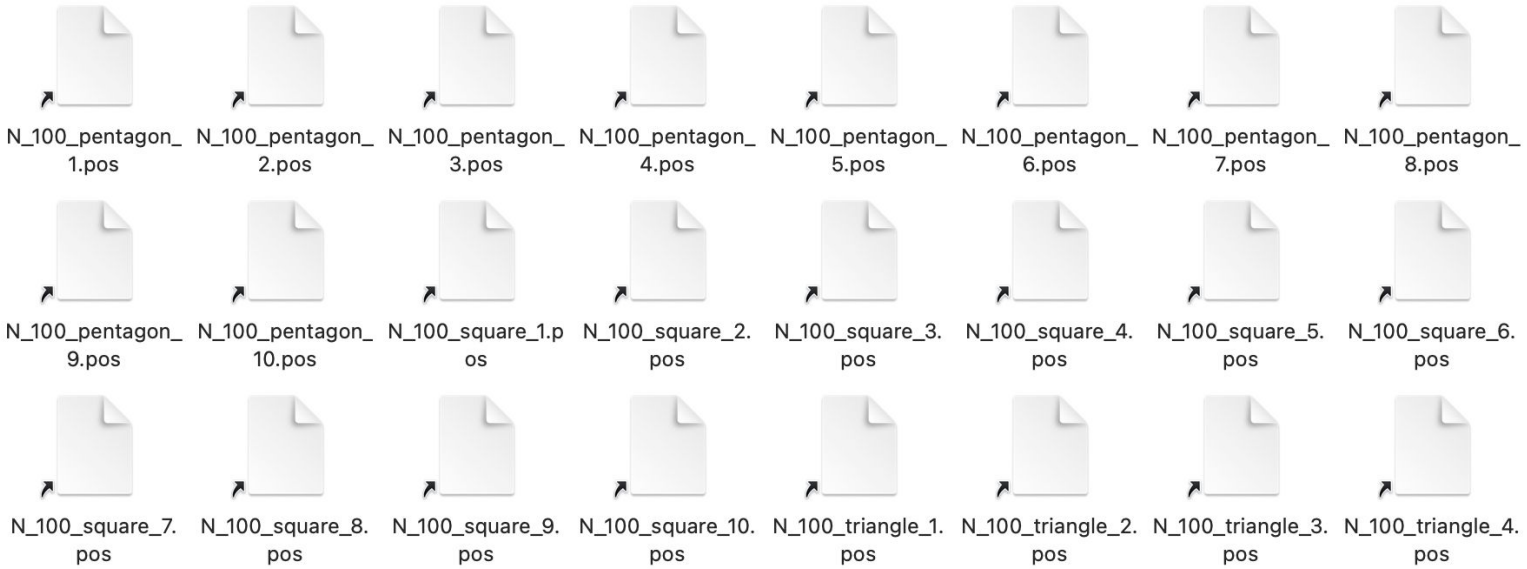
Metadata:

Number of particles, system density, temperature, how many steps, interaction type, etc

Data:

Positions of every particle, measured system information (energy, pressure, etc)

Before signac…



Metadata is a part of the file name for the data file.

Not sustainable

Hard to parse when doing batch analysis

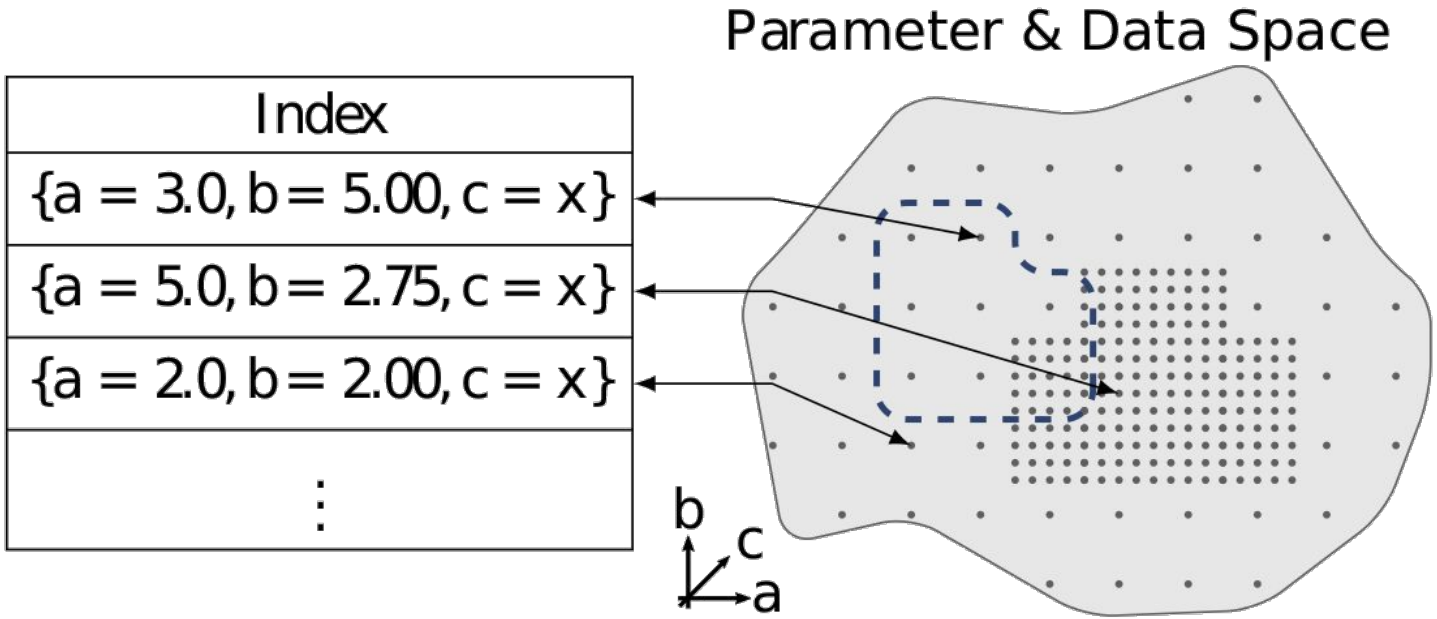
Not robust to parameter expansion

Naming convention needs human knowledge

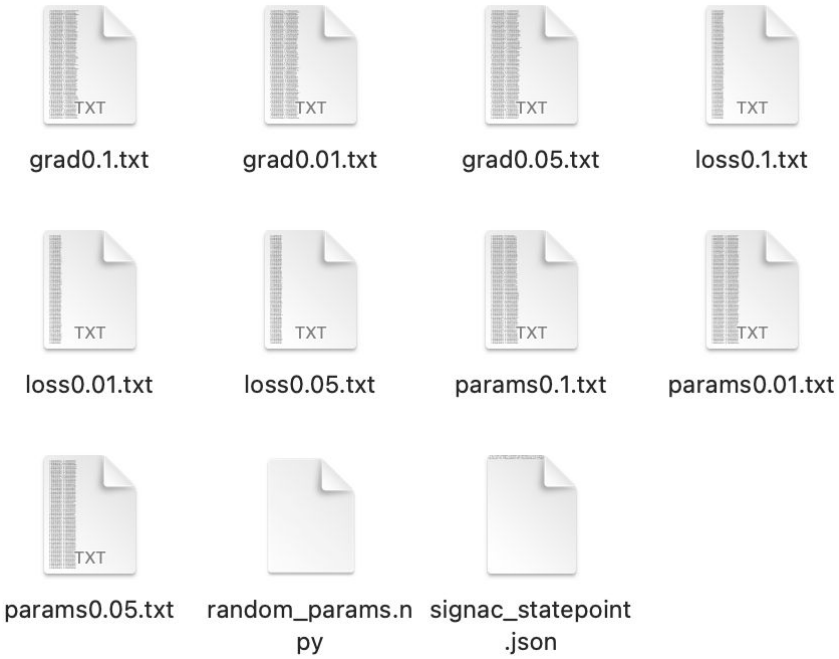
Hard to share data with others

Can we do better?

Introducing Signac



Signac workspace

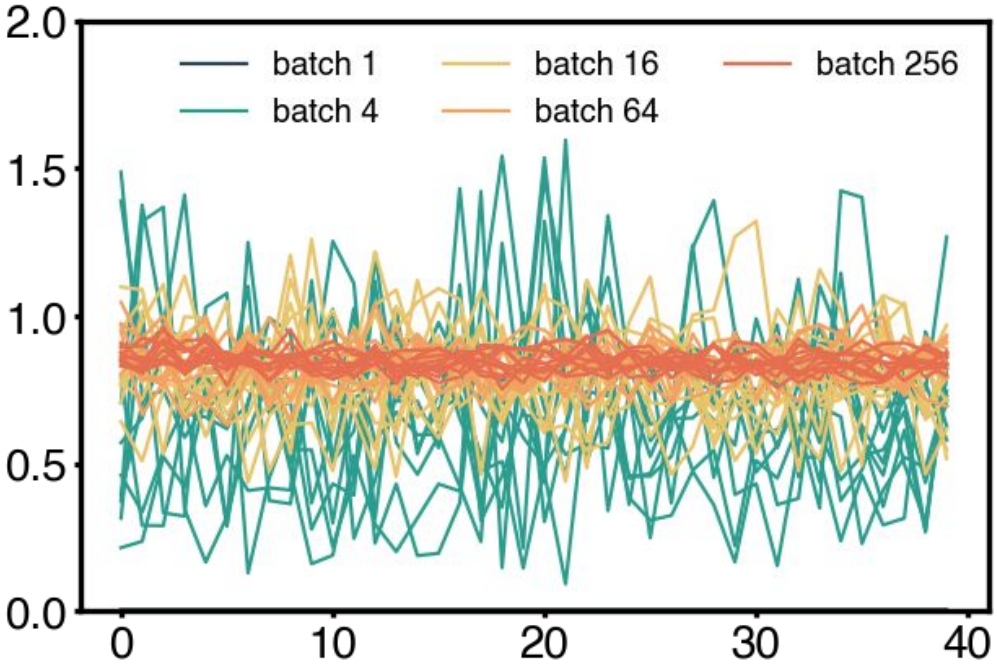


Signac statepoint


```
{"N": 6, "s_type": "octahedra", "num_patch": 10, "type_patch": 2,
"r": 2.5, "phi": 0.05, "num_steps": 40000, "scale": 1.3, "kT": 0.8,
"D0": 4.0, "dt": 0.0001, "rand_search_count": 50, "batch_size": 16,
"loop_batch": 4, "num_steps_opt": 1000, "opt_steps": 100,
"learning_rate": 0.1, "closeness_penalty": 0.0,
"closeness_penalty_nbrs": 1, "seed": 86711402, "replica": 3}
```


These metadata are being encoded into the hex string of the statepoint folders for quick search.

Post processing



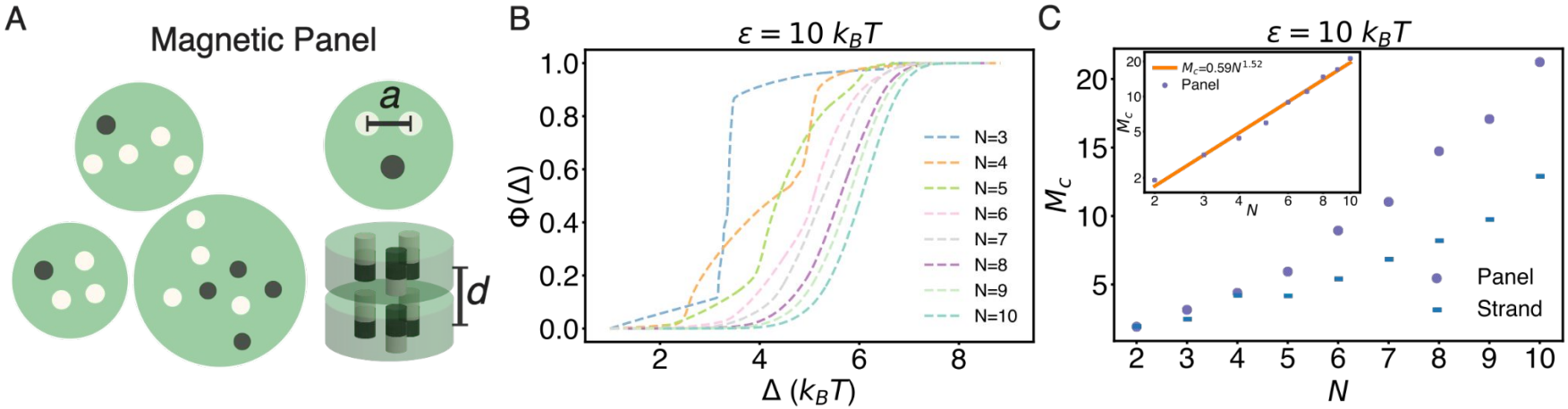
Post processing

```
colors = ['#264653', '#2a9d8f', '#e9c46a', '#f4a261', '#e76f51']
batch_size = [1, 4, 16, 64, 256]

fig = plt.figure(num = 1, figsize = (9, 6), dpi = 80, facecolor = None, edgecolor = 'k')


for i in range(len(batch_size)):
    for job in project.find_jobs({'batch_size': batch_size[i]}):
        ave_loss = np.load(job.fn("std_loss.npy"))
        if job.sp.replica == 1:
            plt.plot(ave_loss, color = colors[i], linewidth = 2, label = 'batch '+str(batch_size[i]))
        else:
            plt.plot(ave_loss, color = colors[i], linewidth = 2)
plt.legend(frameon = False, fontsize = 18, ncol=3)
plt.ylim([0, 2.0])
```

Integration with HPC

```
~/ideal_gas_project $ python project.py status
Using environment configuration: StandardEnvironment
Fetching status: 100%|
Fetching labels: 100%|

Overview: 9 jobs/aggregates, 0 jobs/aggregates with eligible operations.

label            ratio
---------------  ------------------------------------------------------------
volume_computed  |                                                | 9/9 (100.00%)

operation/group
-----------------


[U]:unknown [R]:registered [I]:inactive [S]:submitted [H]:held [Q]:queued [A]:active [E]:error
rror
```

My publication using Signac



Over 1500 statepoint folders, other project has over 10k statepoints

CX. Du, HA. Zhang, T. Pearson, J. Ng, P. McEuen, I. Cohen, and M. P. Brenner, Soft Matter (2022)

Signac

Python based data-management software

Completely Open source

Can be installed with conda and pip

Slack community

Documentation: https://docs.signac.io/en/latest/