SAND2022-2440C

Sandia National Laboratories

SAND2022-2110 C

Scalable Solutions for Training Machine Learned Interatomic Potentials

Mitchell Wood[1], Charles Sievers[1,2], Danny Perez[3], Nicholas Lubbers[3] and Aidan Thompson[1]

[1]Sandia National Labs, Albuquerque, New Mexico, USA

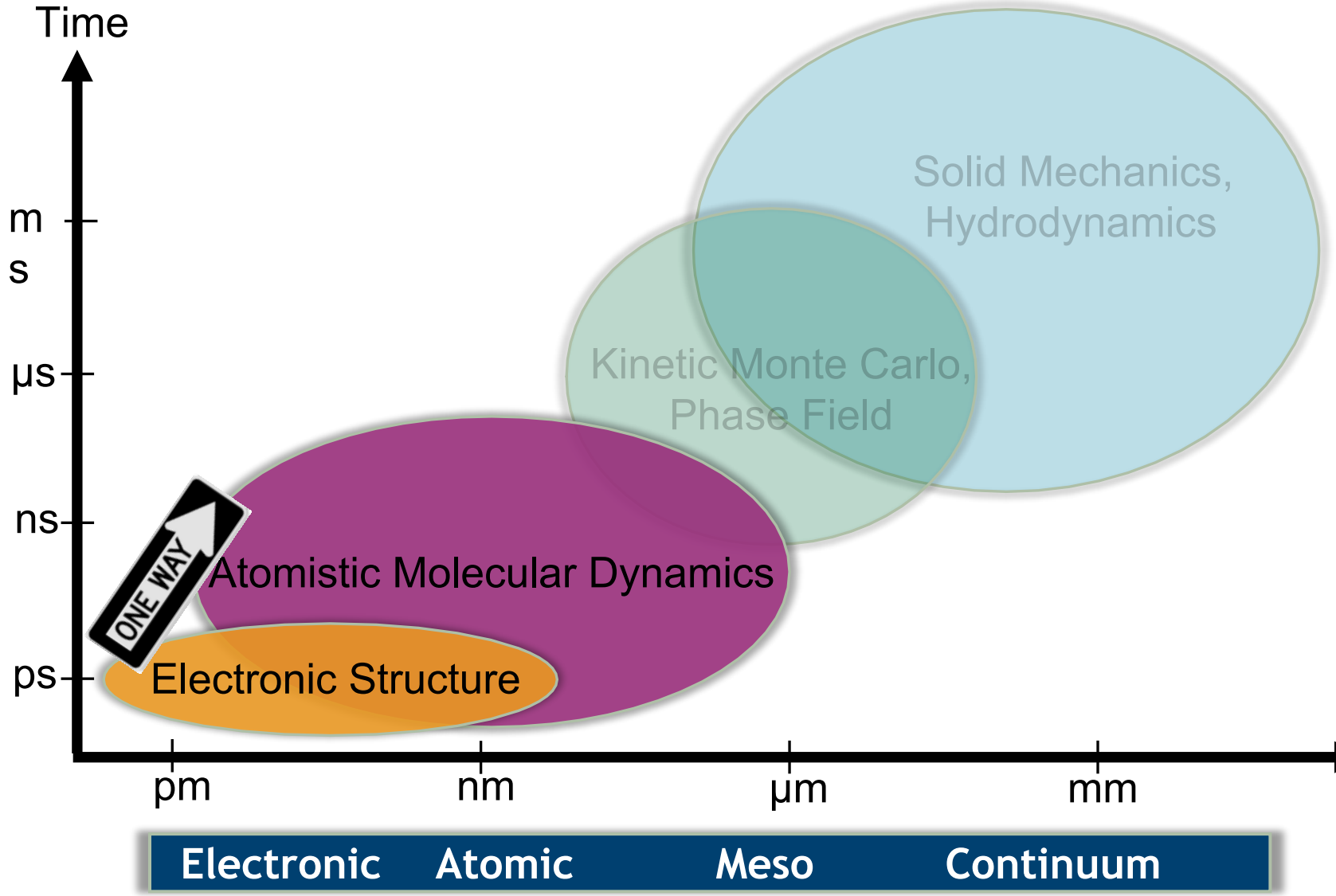
[2]University of California Davis, Davis, California, USA

[3]Los Alamos National Labs, Los Alamos, New Mexico, USA

Sandia National Laboratories is a multi-mission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC., a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

Interatomic Potentials as Multi-Scaling



- IAP can be useful without being physically motivated
- Preserving accuracy through scales while becoming computationally efficient
- Need to be cautious of what is promised with machine learning, most of MD will be extrapolation

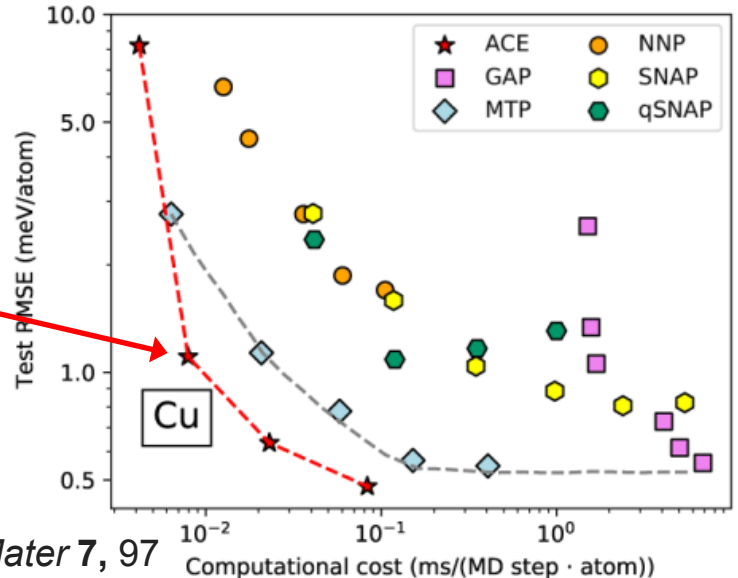
Accuracy Demands

Training Data Sets for Material Science

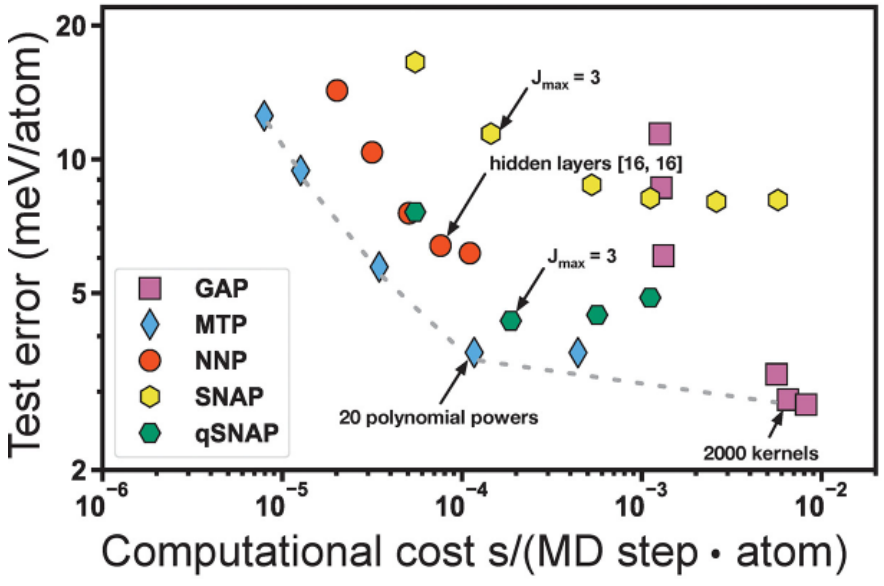
- S.P. Ong (UCSD) and group members generated ‘standardized’ training data for Cu, Ni, Mo, Li, Si, Ge
- Comparing ML-IAP implemented in LAMMPS, cost assessed on CPU only
- Go test your method against ours!

https://github.com/materialsvirtuallab/mlearn

*Uses a modified training set



Lysogorskiy *et. al. npj Comput Mater* **7,** 97



Zuo *et. al. J. Phys. Chem. A* (2020), 124, 4,

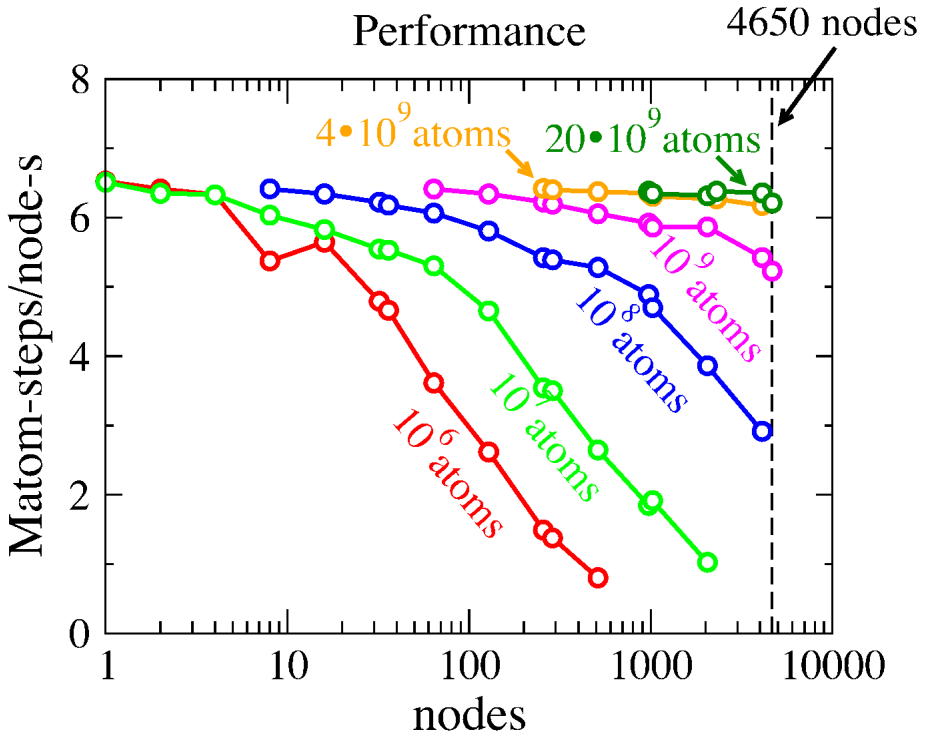
Efficiency Demands

2021 Gordon Bell Finalist

- Uses SNAP ML-IAP for high pressure Carbon
- Team from USF, Sandia, NERSC, NVIDIA, KTH : doi.org/10.1145/3458817.3487400

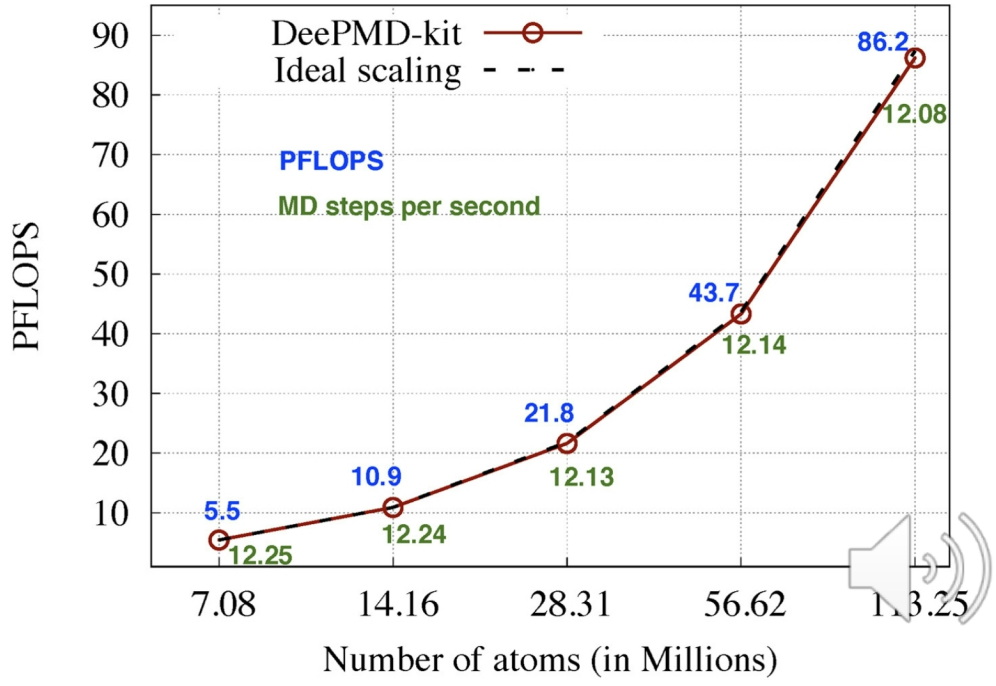
2020 Gordon Bell Winner

- Neural Network based ML-IAP for Water, Cu
- Team from Princeton, Berkeley, LBNL, IAPCM(Beijing), Peking Univ.: doi.org/10.1016/j.cpc.2020.107624



*Both heavily rely on GPU parallelization!

*Both interface with and run in LAMMPS!



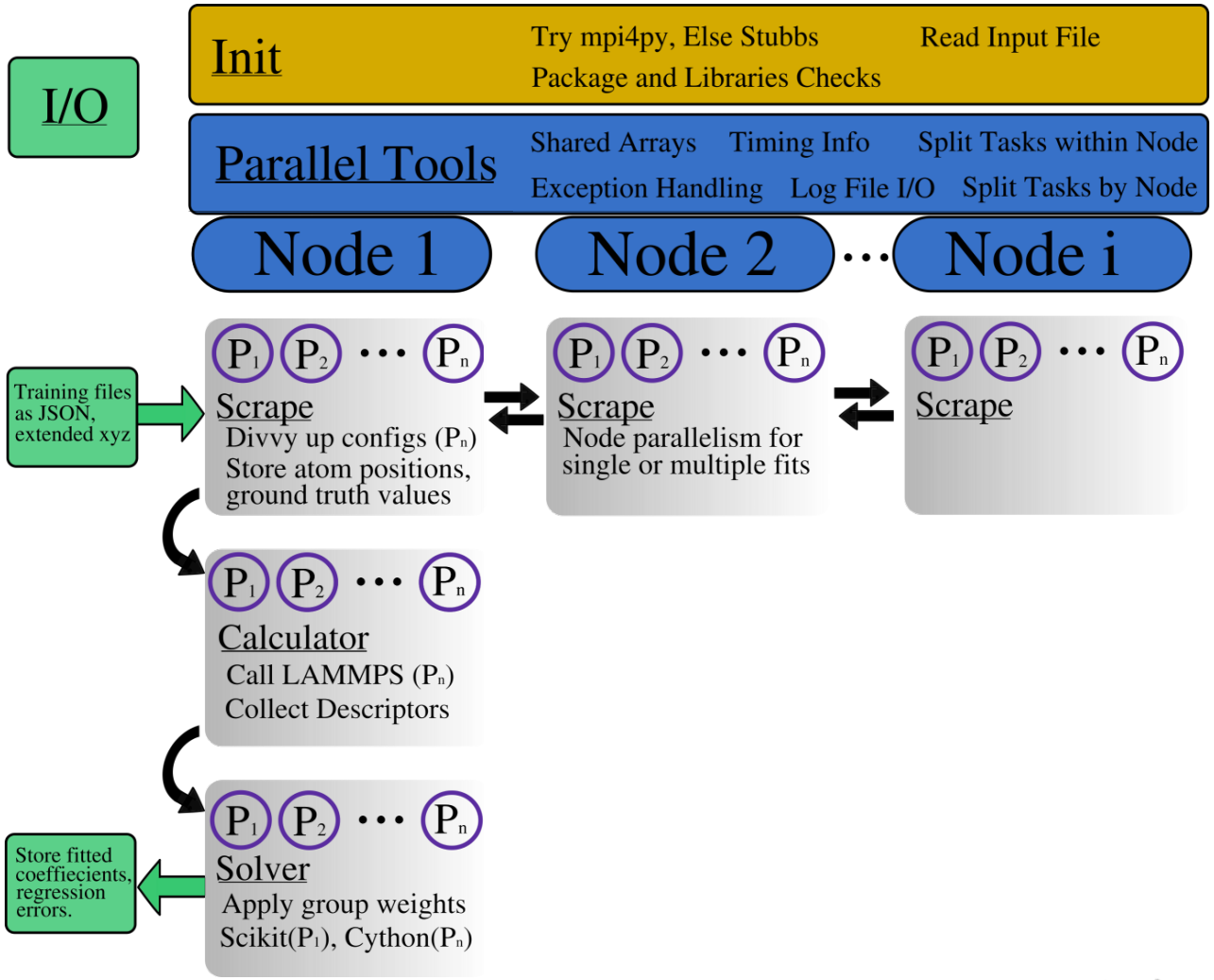
Accelerating Model Development

FitSNAP.py Breakdown

- Software tools needed to realize the accuracy-cost tradeoff continuum
- Python interface for user ease of use Object oriented framework for developer ease of use

- ML-IAP can be ‘overlapped’ with other physical models (coulombic, magnetic spins, ion core repulsion)
- Training data storage as dictionaries:

```
Dataset =
{"Label":
  "Example containing 1 configurations, each with 2 atoms",
  "LatticeStyle": "angstrom","EnergyStyle": "electronvolt","StressStyle": "kB",
  "Data":
    [{"NumAtoms": 2,
      "Lattice": [[ 2.4223151206970215, 0.0, 0.0],[ 0.0, 2.139922618865967, 0.0
      "Energy": -14.30234864,
      "Stress": [[ -37.90633, 6.66761, 0.00398],[ 6.66761, 752.92696, 0.00498],
      "AtomTypes": [ "Fe", "Fe"],
      "Positions": [[ 0.0, 0.0, 0.0],[ 2.26141, 1.33745, 1.97781]],
      "Forces": [[ 0.809971, 0.872872, 0.000495],[ -0.809971, -0.872872, -0.000
      "Spins": [[ 0.10080178569846865, 0.03968183670838251, 0.9126822442927977,
      }]}}
```

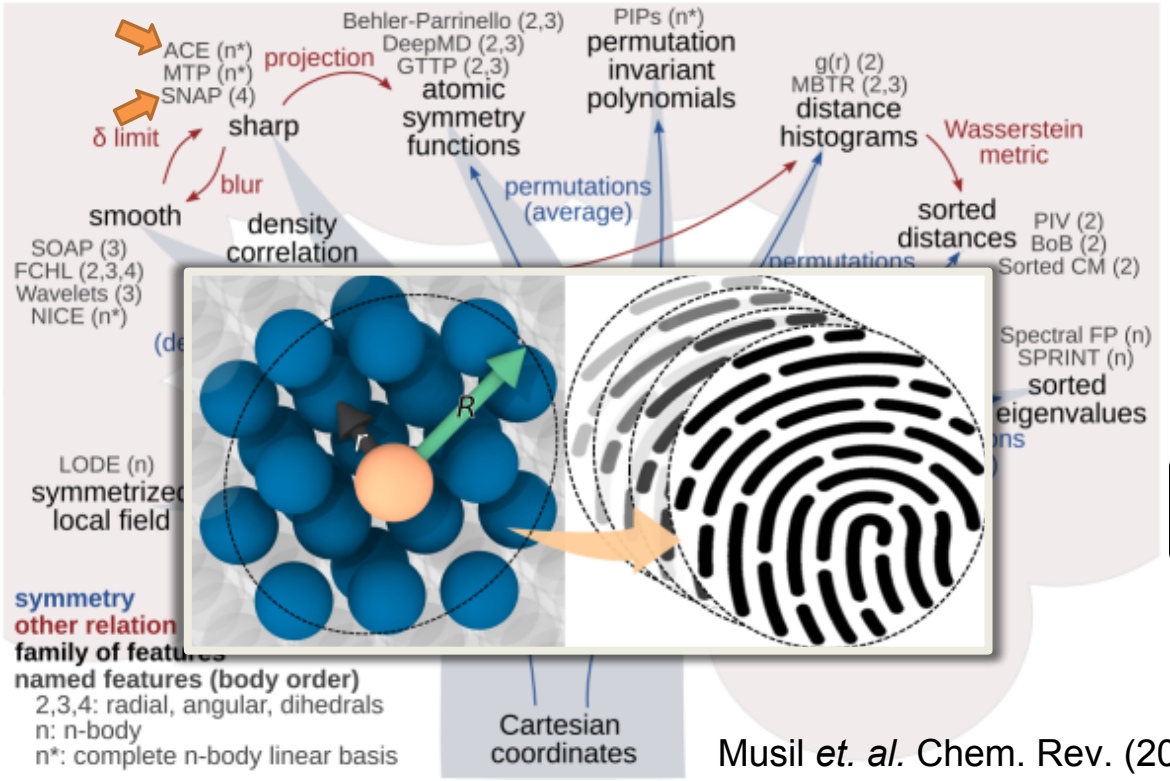


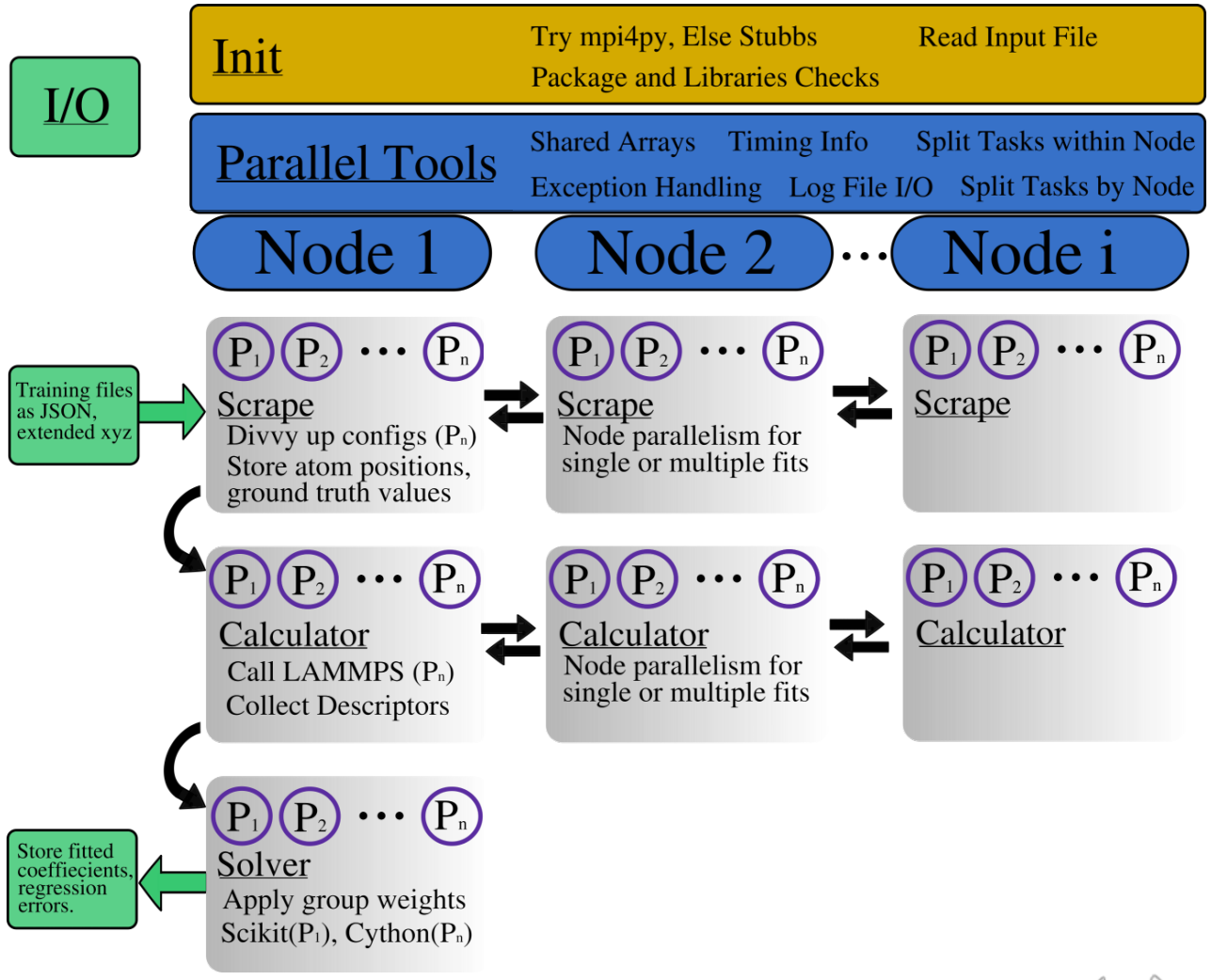
*Auto conversion from VASP OUTCAR coming soon!

Descriptors

LAMMPS Breakdown

- Calculator class calls LAMMPS to convert atomic coordinates into descriptors.
- Thread parallel implementation via Mpi4Py and LAMMPS python library interface.





Musil *et. al.* Chem. Rev. (2021) 121,

Regression

SNAP Model Form

$$E^i_{SNAP} = \alpha_0 + \sum_k \left[\alpha^{(1)}_k \left(B^i_k - B_{k_0}\right) \quad + \alpha^{(n)}_k (\ldots)^n\right]$$

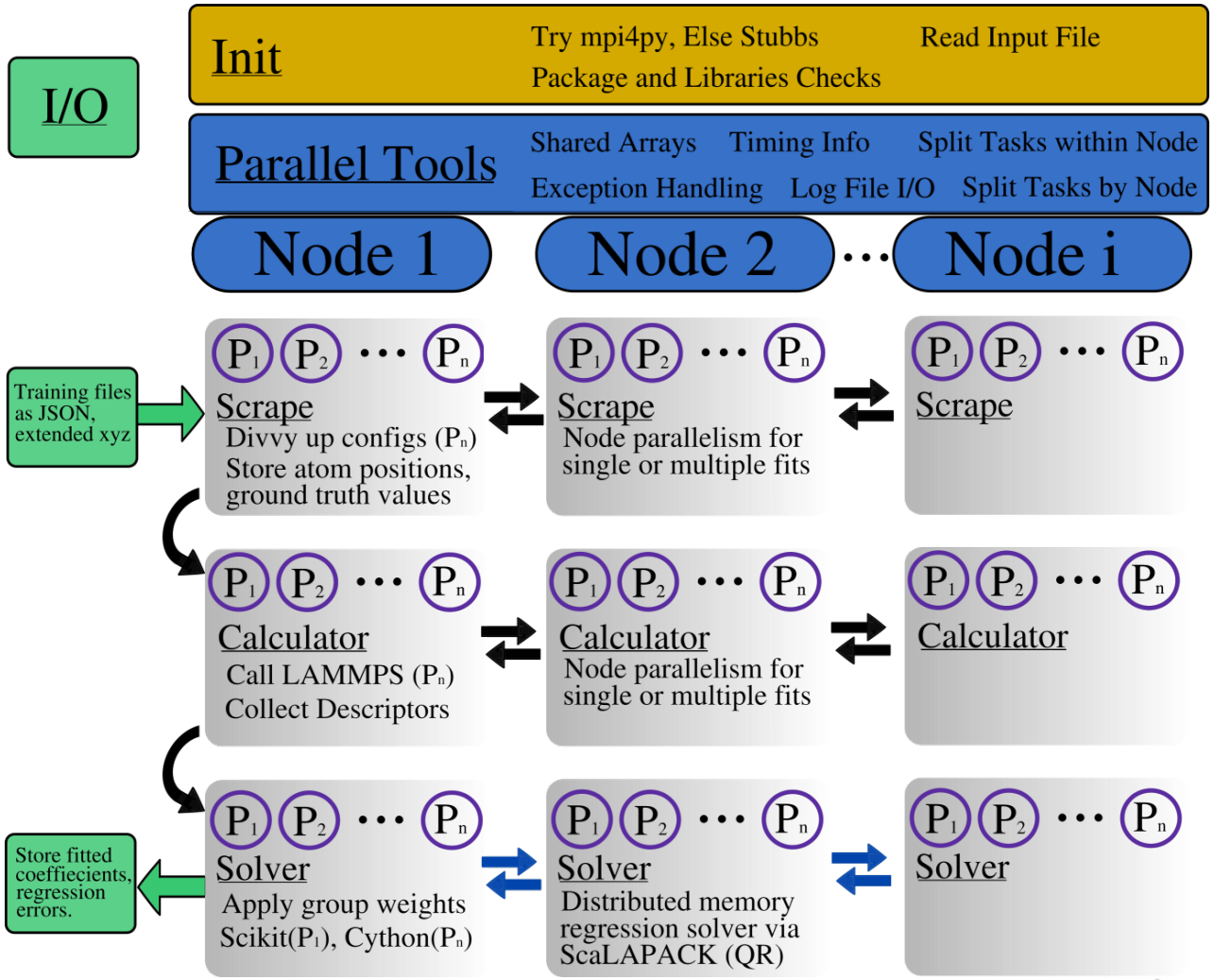
$$min\left(\|w \cdot D\alpha - T\|^2 - \gamma_l \|\alpha\|^l\right)$$

w : Weights

D : Set of descriptors

T : DFT training

- D is a N*M matrix
  - N Scales with number of training points
  - M Scales with the descriptor expansion
  - N>>M, can exceed local memory



*Coupling to AutoDiff and Pytorch coming soon!

Parallelization

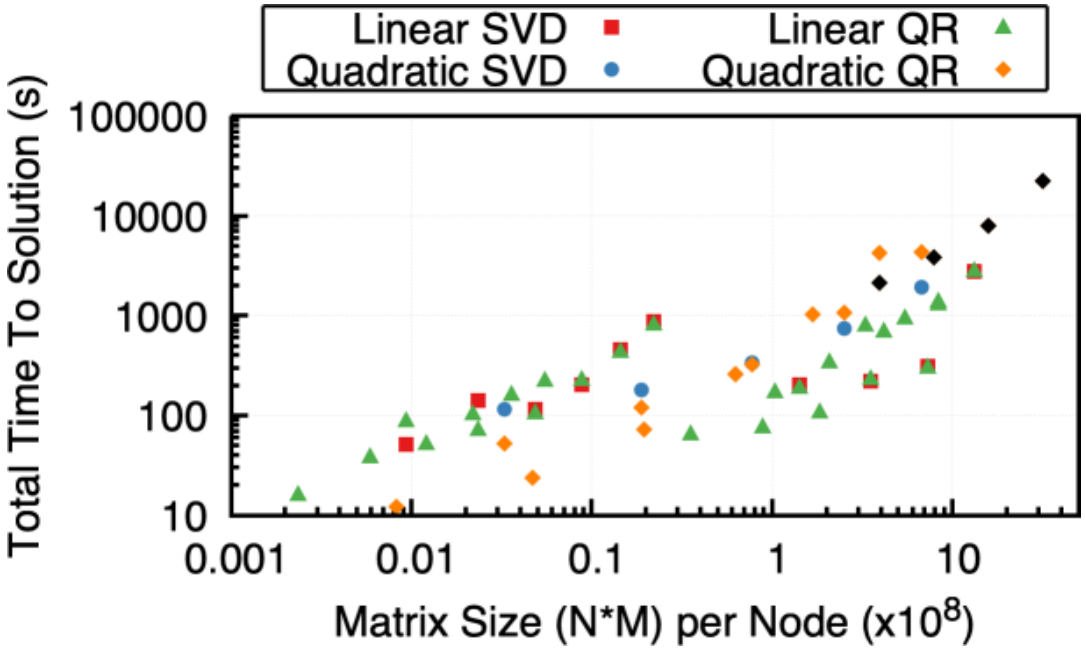
Timing Breakdown

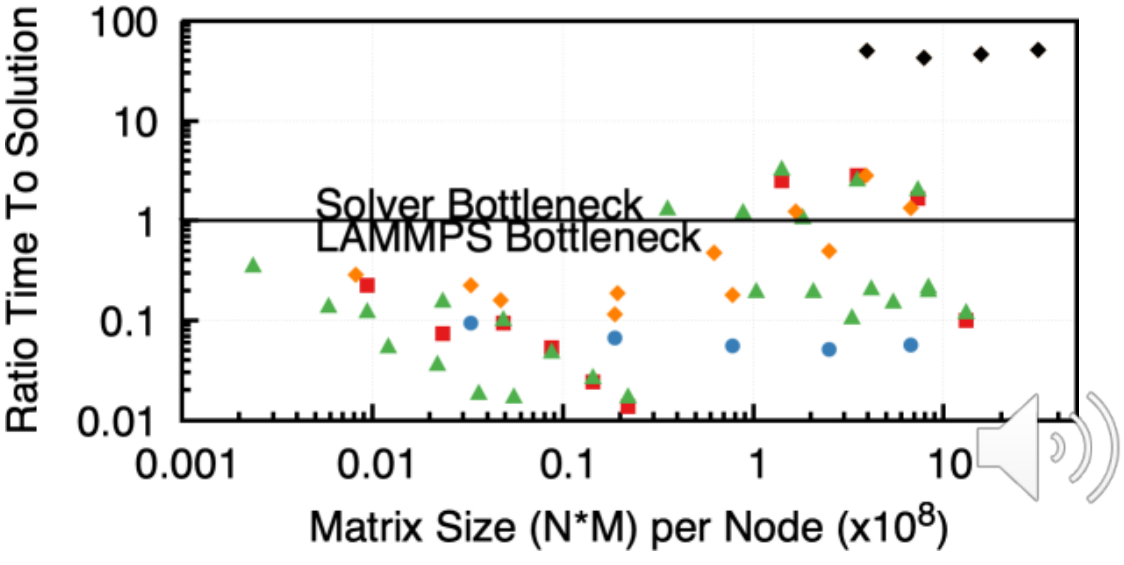
- Cython backend to Solver class allow for distributed memory regression → QR Decomposition via ScaLAPACK
- Each nodes' object handles its own set of training Data → ML-IAP fitting only limited by resource availability

- Points are increasing descriptor basis, Quadratic results in order of magnitude larger M sizes

◆ N*M matrix exceeds the 128Gb of local shared memory, only possible with a distributed solver.

- Gradient descent solvers are implemented, but are slow and add hyperparameters to fits



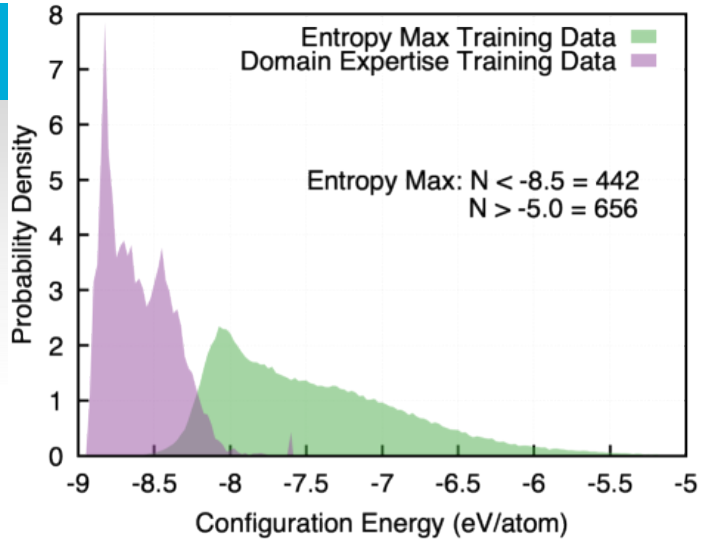


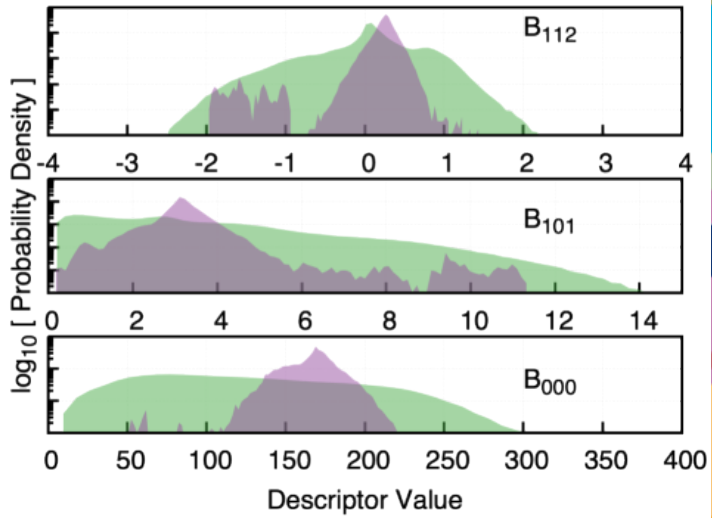
Potpourri – Training Sets

+Montes, Pereyra arXiv:2201.09829

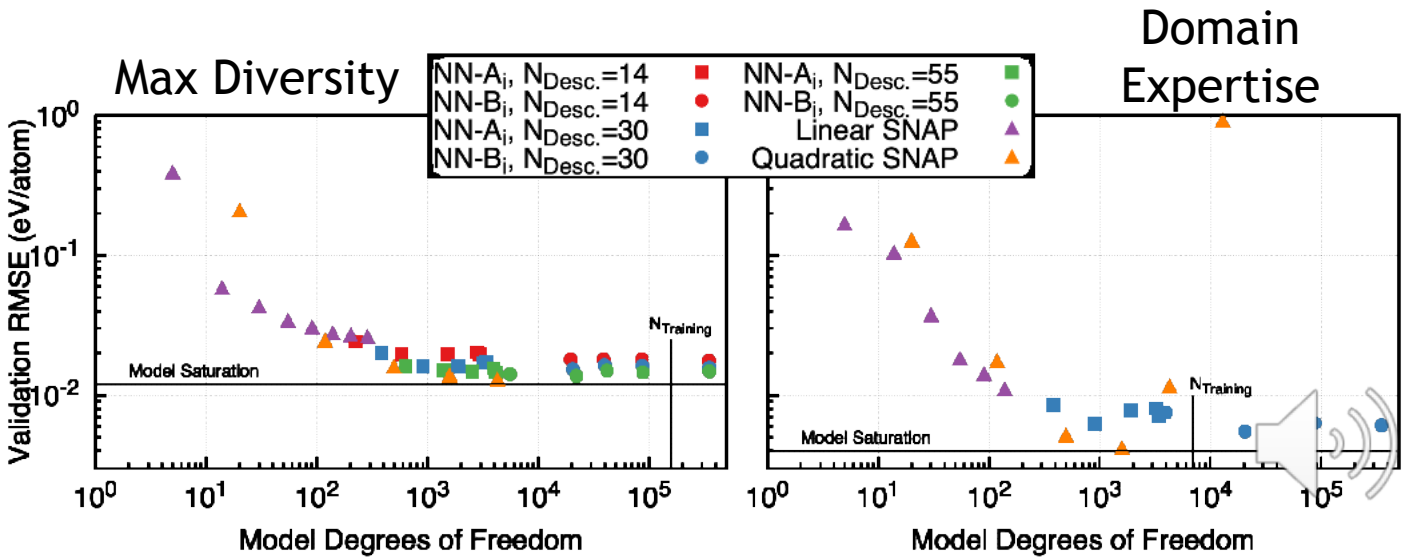
Model form - Training Pairing

- SNAP models are really only tied to bispectrum components as descriptors, model form is flexible
- How complex of a model is needed to capture the training set? Linear? Deep Neural Network?





- Accuracy of ***all*** model forms saturates, true of simple linear and NN models!
- Observed for user constructed and automated training set generation!



Software Development

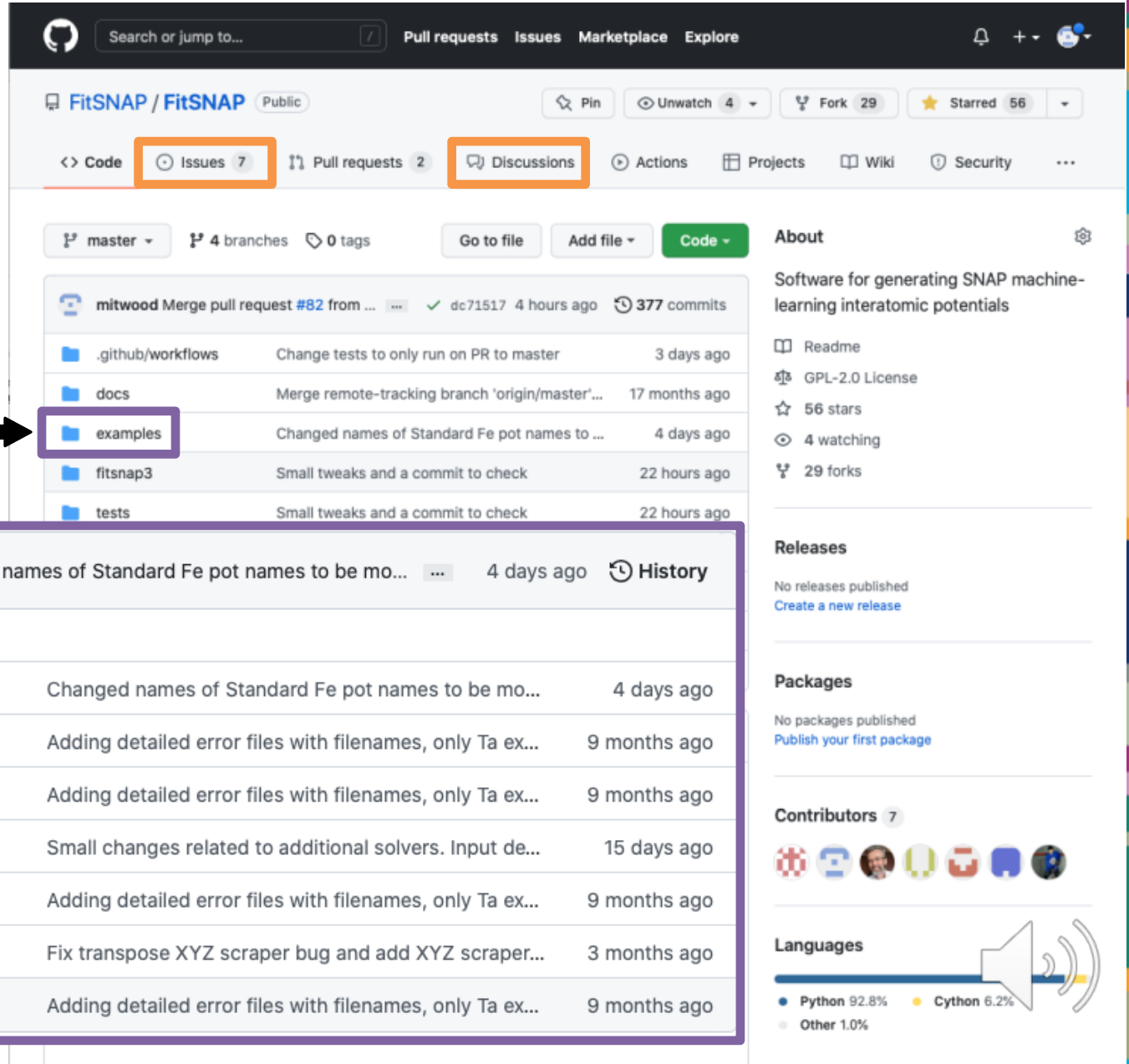
https://github.com/FitSNAP/FitSNAP

- pip install fitsnap
  or
  git clone https://…

- Open issues, request features, and discuss with developers/users via GitHub

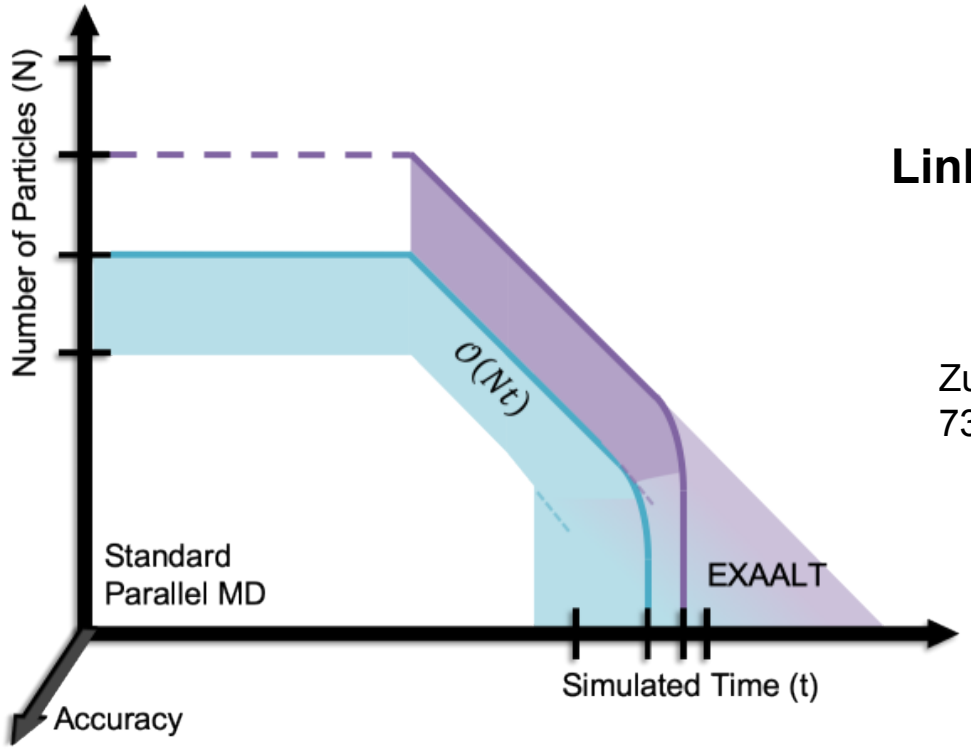
- Example cases roughly follow our publications;

Fe: SNAP + Magnetic Spin
InP: Explicit Multi-Element Descriptors
Ta_PACE: Atomic Cluster Expansion Model
Etc.



Conclusions and Path Forward

- The EXAALT project is ensuring Exascale-ready MD software beyond the length, time-scales of standard MD
- While harder to quantify, the fidelity of our MD simulations needs to be a key consideration at the Exascale



Links and References:

https://github.com/FitSNAP/FitSNAP

https://github.com/materialsvirtuallab/mlearn

Zuo *et. al. J. Phys. Chem. A* (2020), 124, 4, 731–745

Lysogorskiy *et. al. npj Comput Mater* **7,** 97 (2021) doi.org/10.1145/3458817.3487400

doi.org/10.1016/j.cpc.2020.107624

Musil *et. al.* Chem. Rev. (2021) 121, 9759−9815

Contact Information:

mitwood@sandia.gov

EXASCALE COMPUTING PROJECT

Los Alamos NATIONAL LABORATORY EST. 1943

Sandia National Laboratories

NVIDIA

U.S. DEPARTMENT OF ENERGY | Office of Science

OAK RIDGE National Laboratory

NERSC