Training arbitrary models

Joe, Davide & Philip— Metatensor workshop June 2025

What we are going to do today?

1. Using *metatrain* to train a *GAP* and *PET* model on ethanol

2. Run MD with LAMMPS

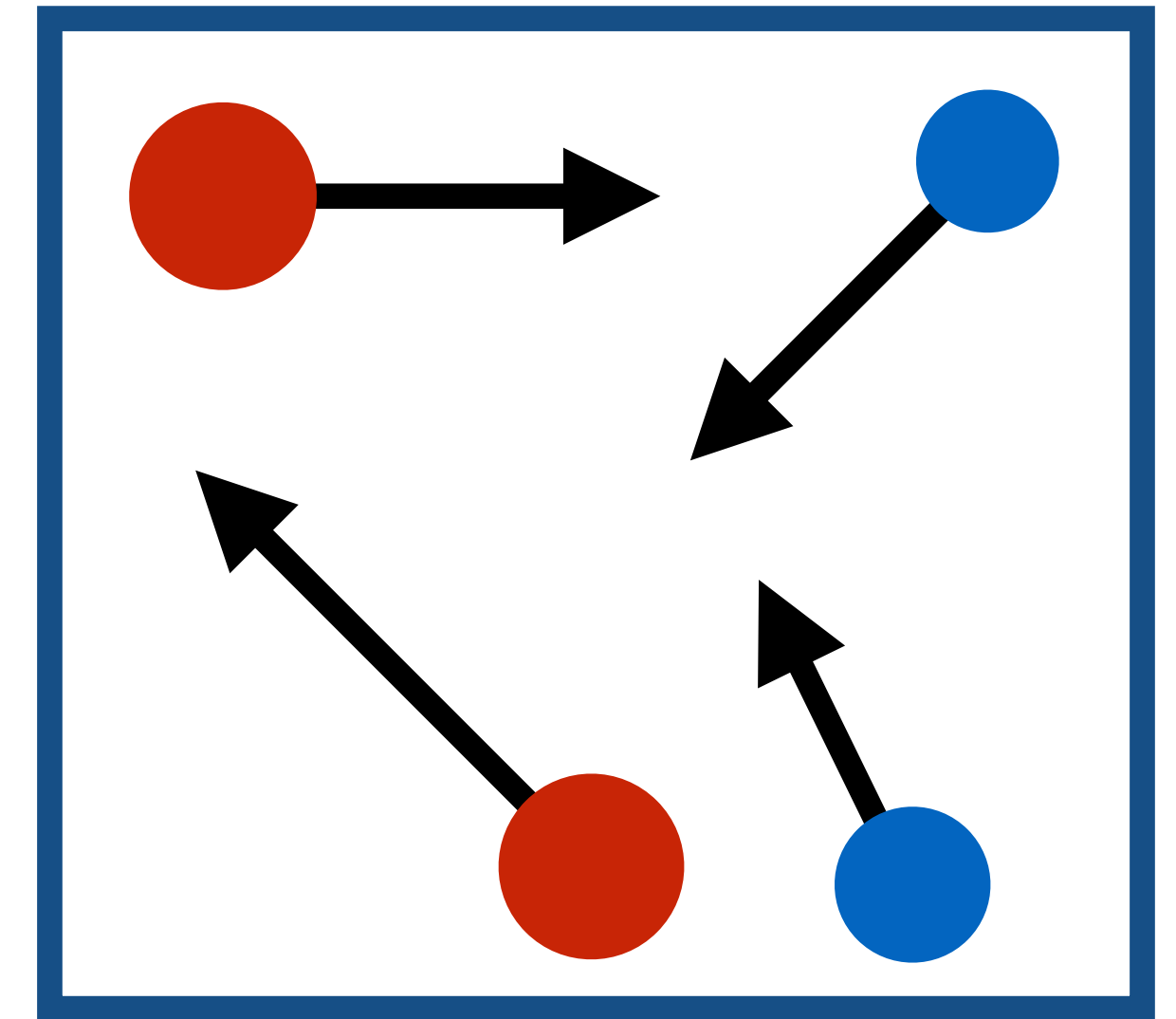3. Quantify the uncertainty of the trained models

Ethanol dataset

- 1000 ethanol structures taken from the rmd17 dataset

- rmd17: revsied molecular dynamics 2017 introduced in Chimela et. al. Sci. Adv. 2017

- Energies and forces for molecular dynamics trajectories of eight organic molecules. Level of theory DFT: PBE+vdW-TS.

- Structure sampled from ab initio molecular dynamics (AIMD)

In short: DO NOT train a model on more than 1000 samples from this dataset. Data already published with 50K samples on the original MD17 dataset should be considered meaningless due to this fact and due to the noise in the original data.

Reminder: Molecular Dynamics

```
call init
t = 0
do while t < tmax
    call force()
    call integrate()
    t=t+delta
enddo
```

from ML

integrator

$$\boldsymbol{F}_i = -\frac{\partial}{\partial \boldsymbol{r}_i} U(\boldsymbol{r}_1, \dots, \boldsymbol{r}_i, \boldsymbol{r}_N)$$

$$\boldsymbol{r}_i(t + \Delta t) \approx 2\boldsymbol{r}_i(t) - \boldsymbol{r}_i(t - \Delta t)\frac{1}{m}\boldsymbol{F}_i(t)\Delta t^2$$

Reminder: Atomistic ML Architectures

Model	Representation / Fingerprint	Training	Body Order	Elements Supported	Equivariance	Scalability	Summary
Gaussian Approximation Potential (GAP)	Fixed mathematical descriptors (e.g., SOAP)	Linear algebra (sparse regression)	Fixed	Few (< 6 re-commended)	✔ Equivariant	⚠️ Limited (small datasets)	**Good for small datasets, precise with limited data**
Point Edge Transformer (PET)	Learned representations (on-the-fly)	Stochastic gradient descent	Unlimited	Many (∞)	❌ Non-equivariant	✅ Scales to large datasets	**Good for large datasets, flexible learning**

Let's get started

https://github.com/metatensor/
Workshop-spring-2025