Symmetries in Neural Networks

How Geometric Priors Can Path the Way Towards Global Machine Learning Potentials

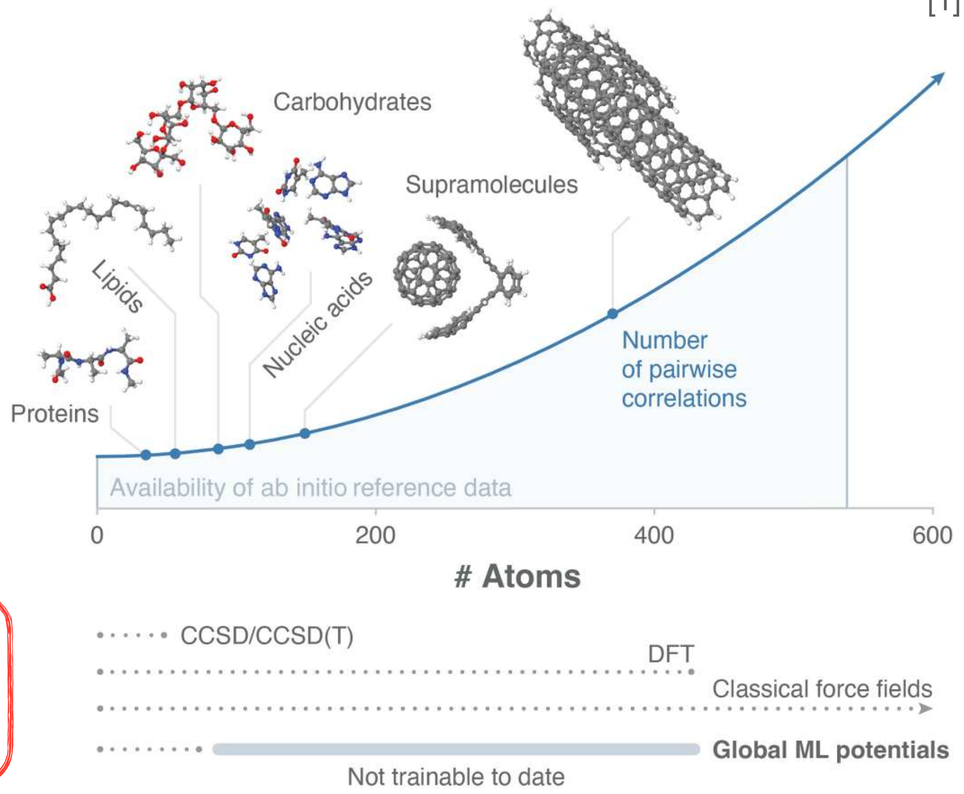
Thorben Frank, TU Berlin

Introduction

- **Goal:** Long time-scale dynamics of atomistic systems for …
  - … protein folding
  - … bio-molecular design
- Traditional *ab-initio* methods only for structures with ~100 atoms
- Machine learning potentials hold the promise to scale beyond 10.000 of atoms

Global machine learning potentials

[1]



[1] Chmiela, Stefan, et al. *Science Advances* 9.2 (2023)

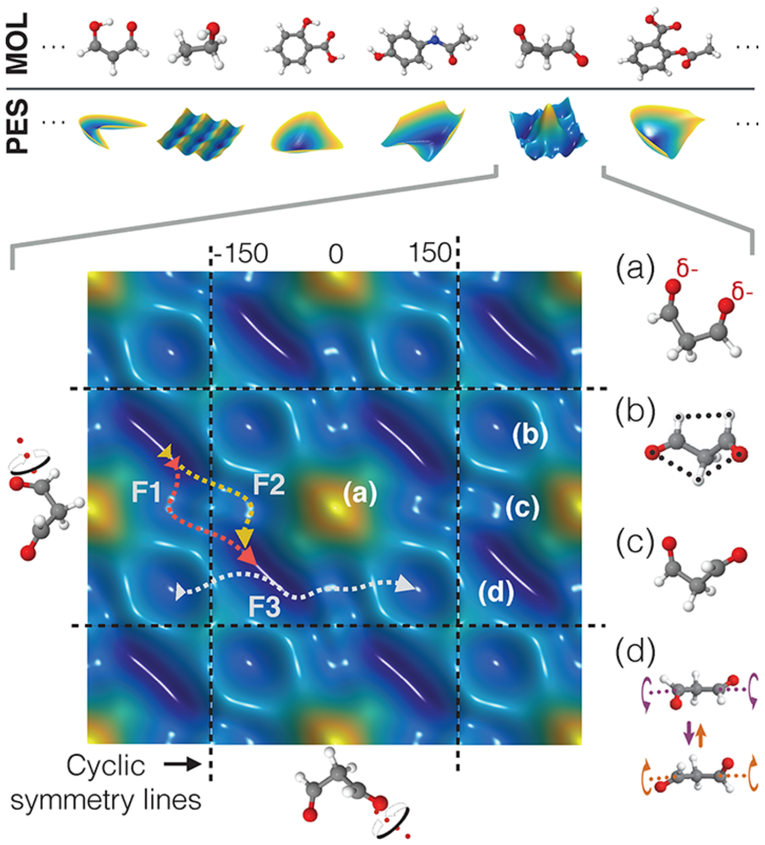
Outline

- Molecular Dynamics for the Impatient
- Machine Learning Potentials
  - Message Passing Neural Networks
- Symmetries in Neural Networks
- Auxiliary Geometric Representations
- (Many) Open Challenges
  - Global Machine Learning Potentials

Potential Energy Surface

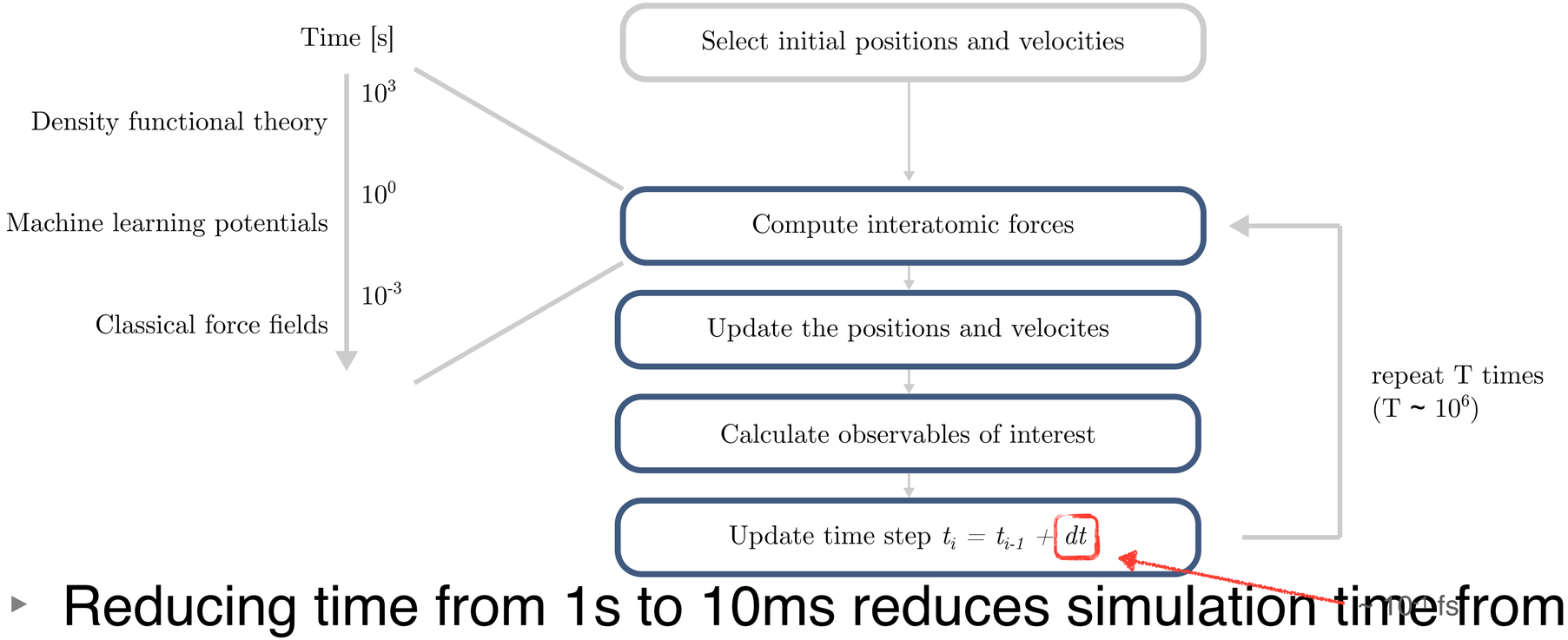
- *Molecular dynamics:*
  Dynamics driven by the interatomic *forces*
- *Atomic positions* $R = \{\vec{r}_1, \ldots, \vec{r}_n | \vec{r}_i \in \mathbb{R}^3\}$
  and *numbers* $Z = \{z_1, \ldots, z_n | z_i \in \mathbb{N}_+\}$
- *Force field:*
  Assigns a force to each atom
  $$\vec{F}_i = \nabla_{\vec{r}_i} f_{\text{pes}}(R, Z)$$
- *Potential energy surface:*
  Born-Oppenheimer
  $$f_{\text{pes}}(Z, R) \mapsto E_{\text{pot}}$$

[2]



[2] Sauceda, Huziel E., et al. Springer, Cham, 2020. 277-307.

Molecular Dynamics Simulation



- Reducing time from 1s to 10ms reduces simulation time from 12d to 3h

Machine Learning Potentials

- Approximate the PES

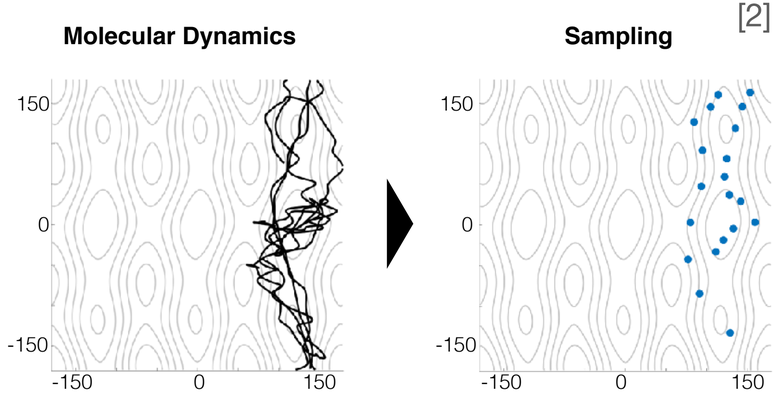
$$E = f_{\text{pes}}(R, Z) \approx f_{\theta^*}(R, Z)$$

- Find the optimal set of parameters θ^* given reference samples $\{(R_k, Z_k, E_k\}_{k=1}^{D}$

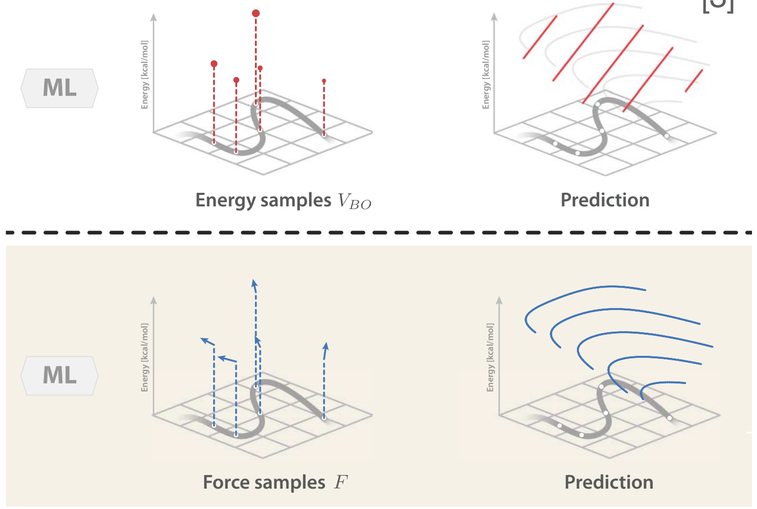
$$\theta^* = \min_\theta \sum_{k=1}^{n_{\text{data}}} \Big(f_\theta(R_k, Z_k) - E_k \Big)^2$$

- Use energy **and** forces

[2]



[3]



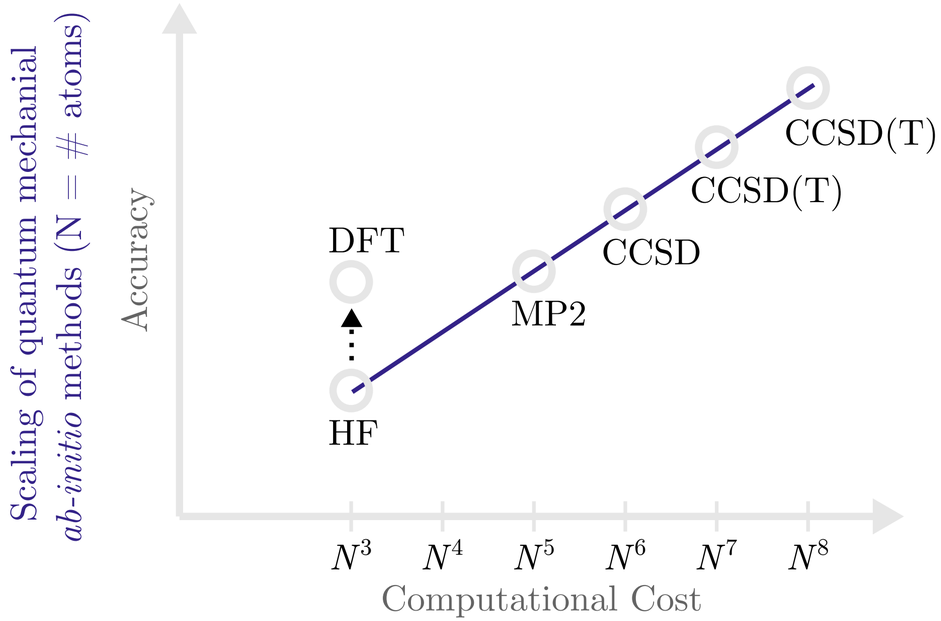
[3] Chmiela, Stefan, et al. *Nature communications* 9.1 (2018): 1-10.

Reference Data

„Garbage In Garbage Out“

- ML model is only as good as the data it is trained on
- High quality reference data becomes more and more expensive
  - Reference data is a precious good
  - *Data efficiency:* Require as little data as possible to yield good results

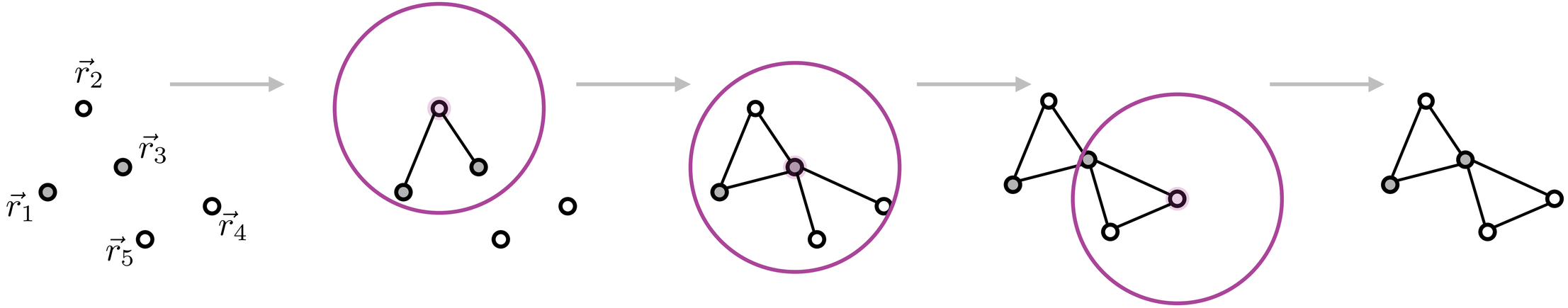
[4]



[4] Frank, Thorben, Oliver Thorsten Unke, and Klaus Robert Muller. *Advances in Neural Information Processing Systems*. 2022.

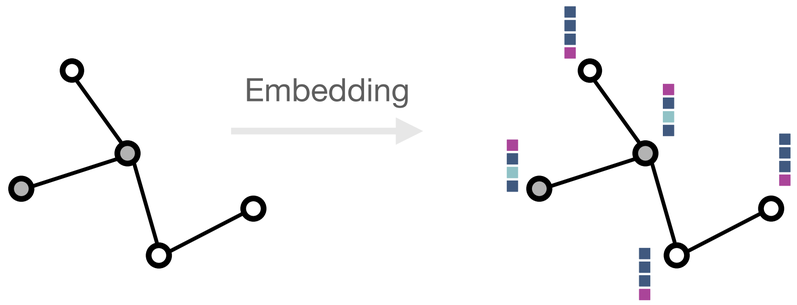
Molecules as Graphs

- Given a data set with molecular geometries
  - Bring it into adequate structure
- Given the atomic positions *R*, construct neighbourhood for each atom as
  - $\mathcal{N}_i = \{j \,|\, r_{ij} \leq r_{\text{cut}}\}$ with $r_{ij} = \|\vec{r}_{ij}\|_2 = \|\vec{r}_j - \vec{r}_i\|_2$

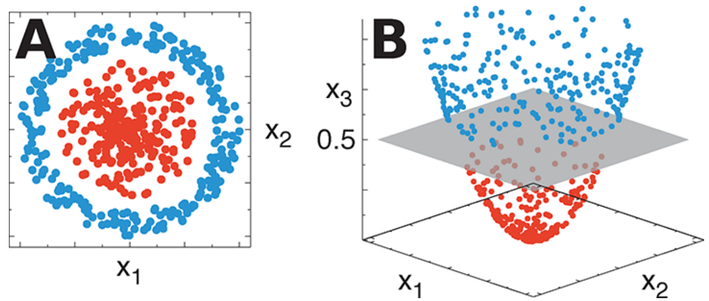


Message Passing Neural Networks

- Neural networks for data that has a graph like structure
- Represent molecule as directed graph with
  - Node values $V = \{ z_1, \ldots, z_n \mid z_i \in \mathbb{N}_+ \}$
  - Edge values $e = \{ \vec{r}_{12}, \ldots, \vec{r}_{kn} | \vec{r}_{ij} \in \mathbb{R}^3 \}$
- Embed the graph:
  - Node features $V_F = \{ f_1, \ldots, f_n \mid f_i \in \mathbb{R}^F \}$ with $f_i^{[t=0]} = f_{\text{emb}}(z_i)$



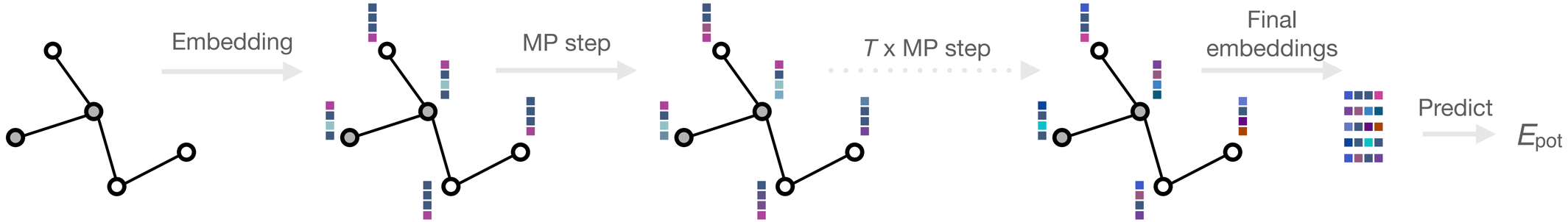
[5]



[5] Unke, Oliver T., et al. *Chemical Reviews* 121.16 (2021): 10142-10186.

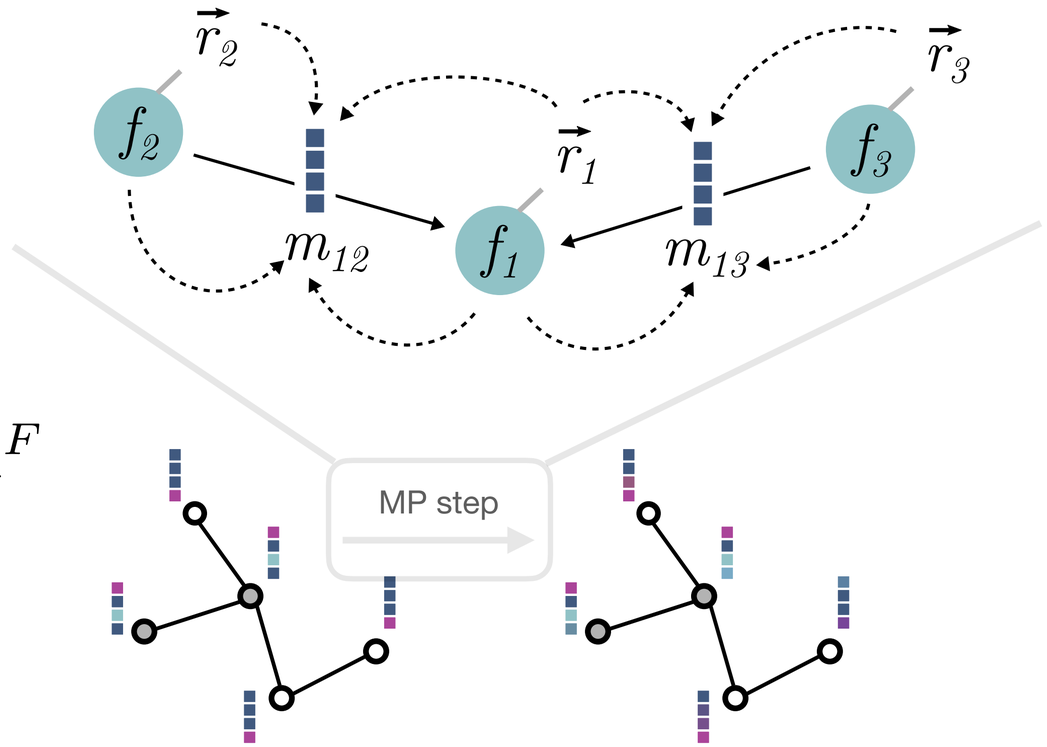
Message Passing Neural Networks

- Iteratively update atomic embeddings given their neighbourhood
  - Message passing step $f_i^{[t+1]} = \mathrm{MPstep}(f_i^{[t]}, R, Z)$
- Energy as sum of atomic contributions
  - $E_{pot} = \sum_{i=1}^{n} E_i$ with $E_i = g_\theta(f_i^{[T]})$ with $g_\theta : \mathbb{R}^F \mapsto \mathbb{R}$



Message Passing Neural Networks

- Message passing step:
  - 1. Message $m_{ij} = m_\theta(f_i, f_j, \vec{r}_i, \vec{r}_j)$
  - 2. Aggregation $m_i = \sum_{j \in \mathcal{N}_i} m_{ij}$
  - 3. Update $f_i^{[t+1]} = u_\theta(f_i^{[t]}, m_i)$
- With $m_\theta : \mathbb{R}^{2F+2d} \mapsto \mathbb{R}^F$ and $u_\theta : \mathbb{R}^{2F} \mapsto \mathbb{R}^F$
- Simultaneously take an update step for all atomic features



SO(3) Equivariance

- Given vector spaces *A*, *B* and group *G,* a function $h : A \mapsto B$ is said to be *G-equivariant* if

$$h\big(\rho(g)a\big) = \sigma(g)h(a)$$

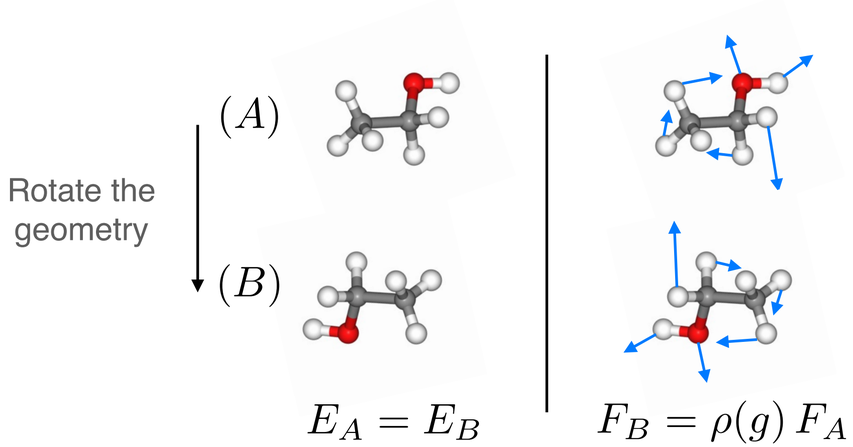
  holds for every $g \in G$, where ρ and σ are group representations in A and B.
- Special case: G-invariant

$$\sigma(g) = \mathrm{Id}$$

- *Example:* Rotations in Euclidean space
  - *G* = SO(3) and $\rho(g)$ = rotation matrices

Invariant Message Passing

- Potential energy is *invariant* w.r.t. rotations
- Forces are *equivariant* w.r.t. rotations
- Invariance can be build by making the message and the features invariant
  - Features: $f_i = f_{\text{emb}}(z_i)$ ✓
  - Message: $m_{ij} = m_\theta(f_i, f_j, ||\vec{r}_i - \vec{r}_j||_2)$ ✓
  - $E_{pot} = \sum_{i=1}^{n} E_i \text{ with } E_i = g_\theta(f_i^{[T]})$ ✓



Message function

$$m_{ij} = m_\theta(f_i, f_j, \vec{r}_i, \vec{r}_j)$$

Aggregation

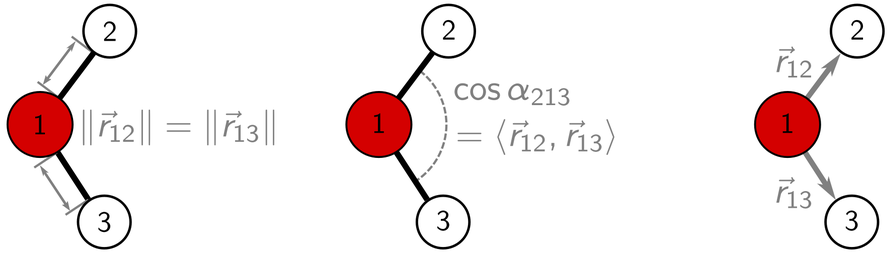
$$m_i = \sum_{j \in \mathcal{N}_i} m_{ij}$$

Update function

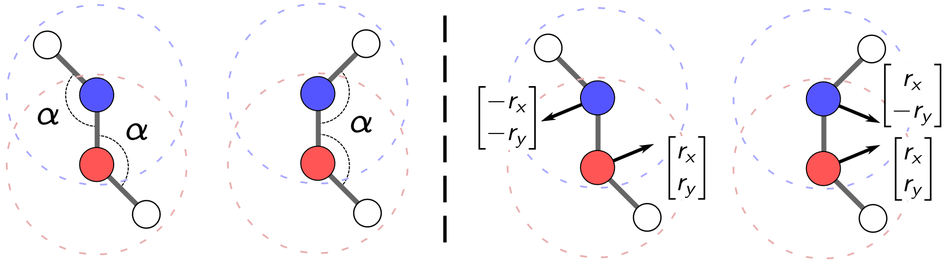
$$f_i^{[t+1]} = u_\theta(f_i^{[t]}, m_i)$$

Is Invariance All You Need?

- Consider message function based on
  - Distances (*invariant*)
  - Angles (*invariant*)
  - Directional vectors (*equivariant*)
- (In-)distinguishable geometries



$$\sum_{j \in \mathcal{N}_i} \|\vec{r}_{ij}\| \qquad \sum_{j \in \mathcal{N}_i} \sum_{k \in \mathcal{N}_i} \alpha_{jik} \qquad \sum_{j \in \mathcal{N}_i} \frac{\vec{r}_{ij}}{\|\vec{r}_{ij}\|}$$



✗ ✗ ✓

[6] Schütt, Kristof, Oliver Unke, and Michael Gastegger. *International Conference on Machine Learning*. PMLR, 2021.

Equivariant Message Passing

- Builds upon equivariant atomic features and messages
- Use *spherical harmonics* as elementary building block
  - For each rotation matrix corresponding *Wigner-D* matrix $D^{(l)} \in \mathbb{R}^{(2l+1)\times(2l+1)}$
- Under rotation of the input, features and messages transform as
  - Equivariant atomic features:

$$f_i\left(M_{\mathrm{rot}}(g)\vec{r}_j\right) = D^{(l)}(g)\, f_i\left(\vec{r}_j\right)$$

  - Equivariant message:

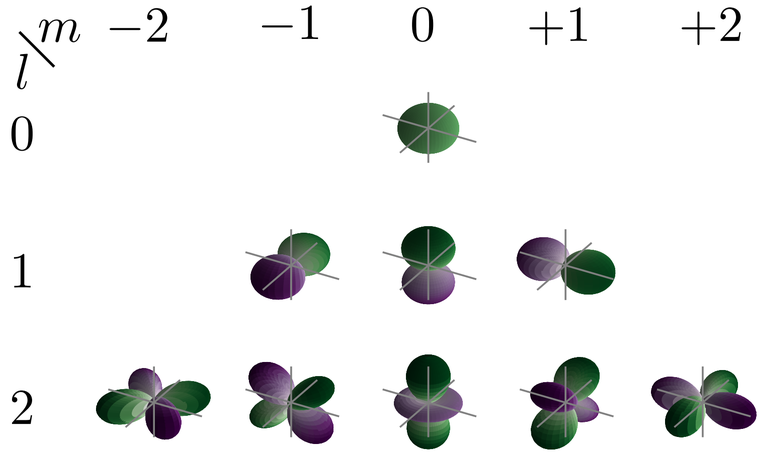
$$m_{ij}\left(f_i, f_j, M_{rot}(g)\vec{r}_{ij}\right) = D^{(l)}(g)\, m_{ij}\left(f_i, f_j, \vec{r}_{ij}\right)$$

Interlude: Real Spherical Harmonics

$$Y_l^m(\hat{r}) : S^2 \mapsto \mathbb{R} \xrightarrow[\text{for a given } l]{\text{collect all } m} \boldsymbol{Y}^{(l)} : S^2 \mapsto \mathbb{R}^{2l+1}$$

$$\text{degree } l \in \{0, \ldots, \infty\}, \text{ order } m \in \{-l, \ldots, +l\} \text{ and unit sphere } S^2$$

- Transformation under rotation
  - $$\boldsymbol{Y}^{(l)}(M_{\text{rot}}(g)\hat{r}) = D^{(l)}(g)\boldsymbol{Y}^{(l)}$$
- Wigner-D matrix $D^{(l)} \in \mathbb{R}^{(2l+1)\times(2l+1)}$
- SO(3) equivariant functions



Equivariant Message Passing

- How to build an equivariant message passing mechanism?
- SO(3) equivariant features

$$\underbrace{\boldsymbol{f}^{(l)}}_{\in\mathbb{R}^{(2l+1)\times F}} \equiv \underbrace{\boldsymbol{Y}^{(l)}(\hat{r})}_{\in\mathbb{R}^{2l+1}} \otimes \underbrace{f}_{\in\mathbb{R}^{F}} = \begin{bmatrix} Y^{(l)}_{-m} f_1 & \dots & Y^{(l)}_{-m} f_F \\ \vdots & \ddots & \vdots \\ Y^{(l)}_{m} f_1 & \dots & Y^{(l)}_{m} f_F \end{bmatrix}$$

$$\implies \boldsymbol{Y}^{(l)}(M_{\text{rot}}\hat{r}) \otimes f = D^{(l)}(M_{\text{rot}})\boldsymbol{Y}^{(l)}(\hat{r}) \otimes f \quad \checkmark$$

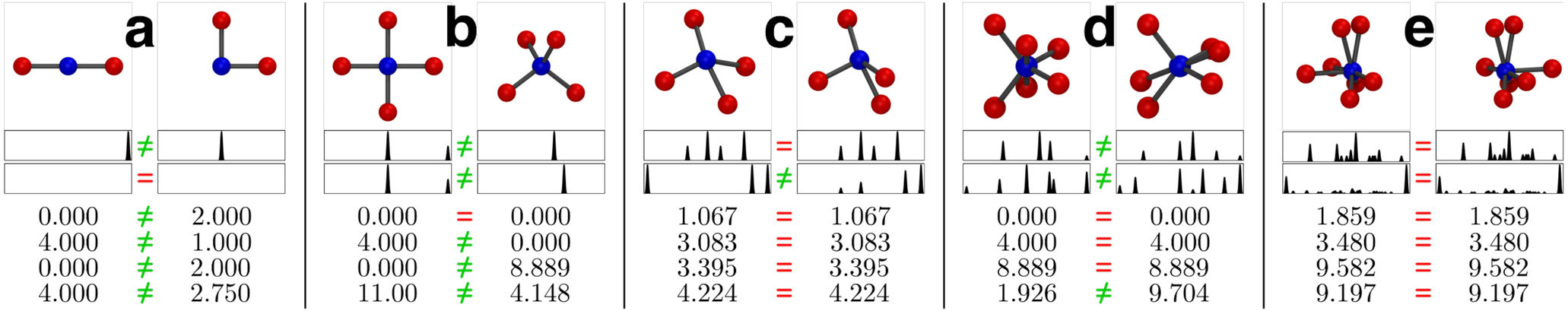
- SO(3) equivariant message

$$\underbrace{\boldsymbol{m}_i}_{\in\mathbb{R}^{(2l+1)\times F}} = \sum_{j\in\mathcal{N}_i} \underbrace{g(r_{ij})}_{\in\mathbb{R}^{F}} \otimes \underbrace{Y^{(l)}(\hat{r}_{ij})}_{\in\mathbb{R}^{2l+1}} \cdot \underbrace{\boldsymbol{f}_i}_{\in\mathbb{R}^{(2l+1)\times F}} \in \mathbb{R}^{(2l+1)\times F}$$

Must preserve equivariance

Local Geometric Expressiveness

- Potential energy is a sum of *local* atomic neighbourhood contributions
  - Capture as much geometric neighbourhood information as possible
- Any function on the sphere: $h(\hat{r}) = \sum_{l=0}^{\infty} c_{ml} Y_m^{(l)}(\hat{r})$



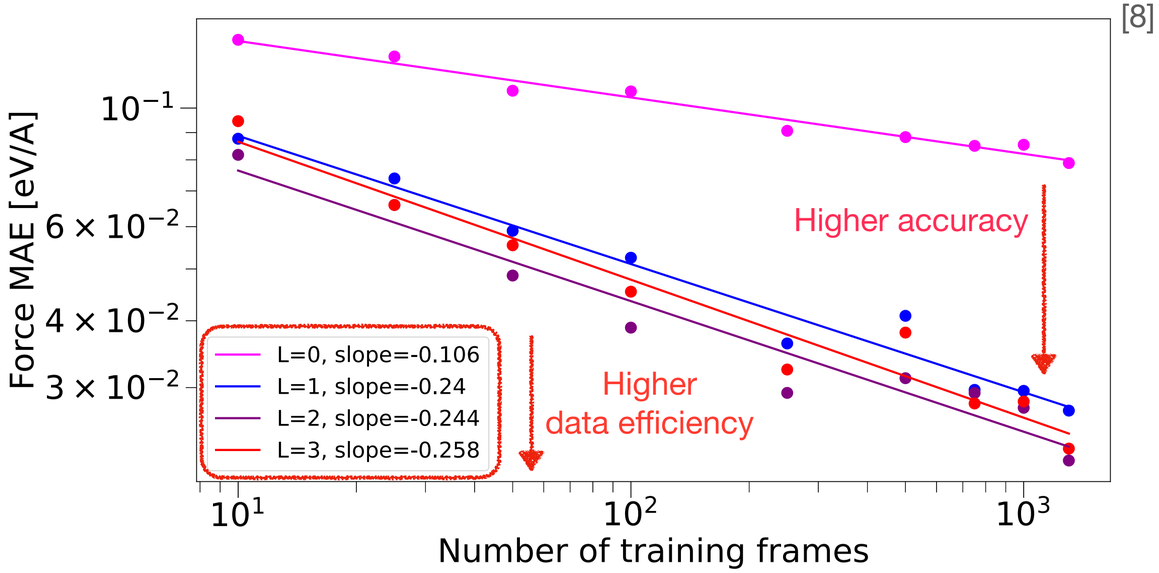
[7]

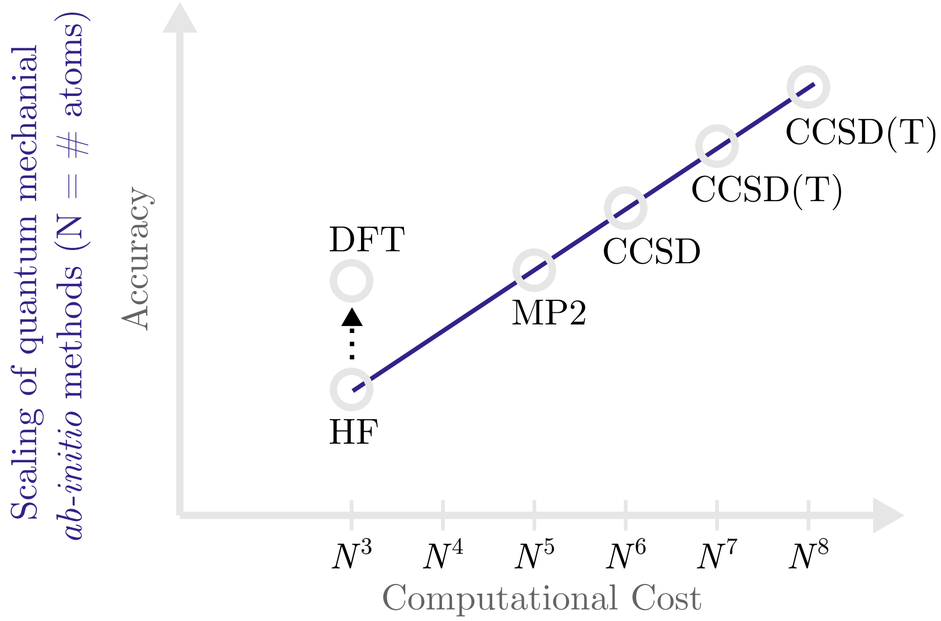
Larger degree *l* necessary to distinguish the neighbourhood of the central blue atom

[7] Unke, Oliver T., et al. *Nature communications* 12.1 (2021): 1-14.

Data Efficiency and Accuracy

- *Reminder:* One need data efficient and accurate models
- Both increase with max. degree *l*

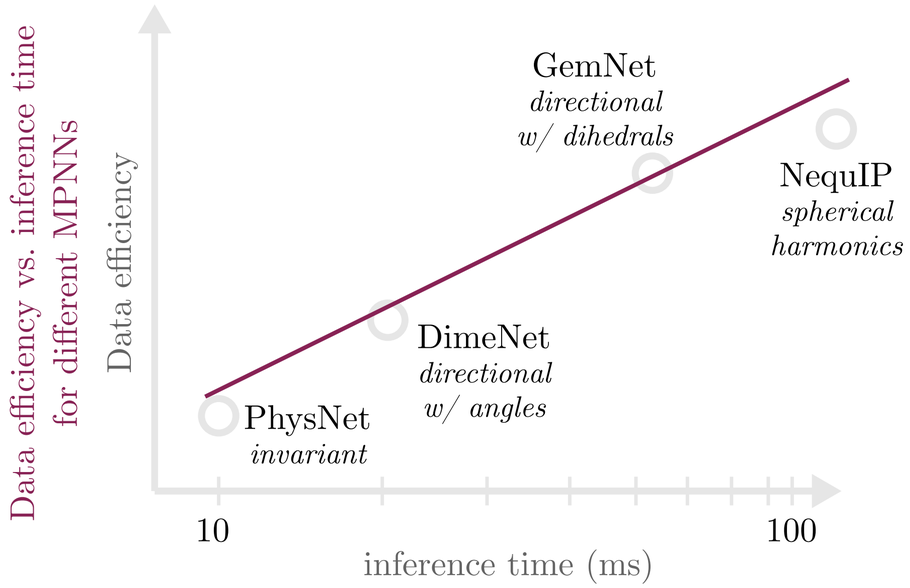


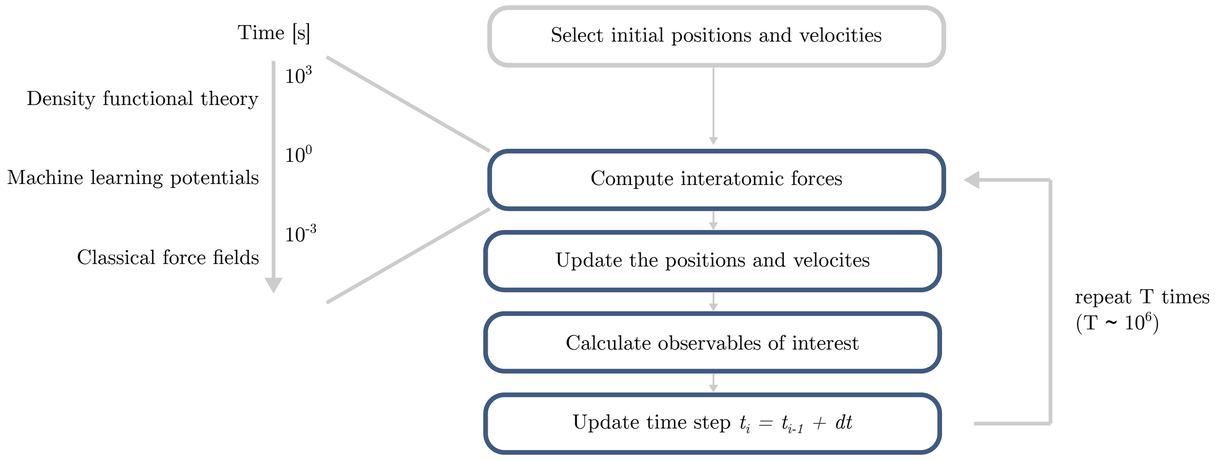


[8] Batzner, Simon et al. *Nature communications* 12.1 (2022): 1-14.

Data Efficiency vs. Complexity

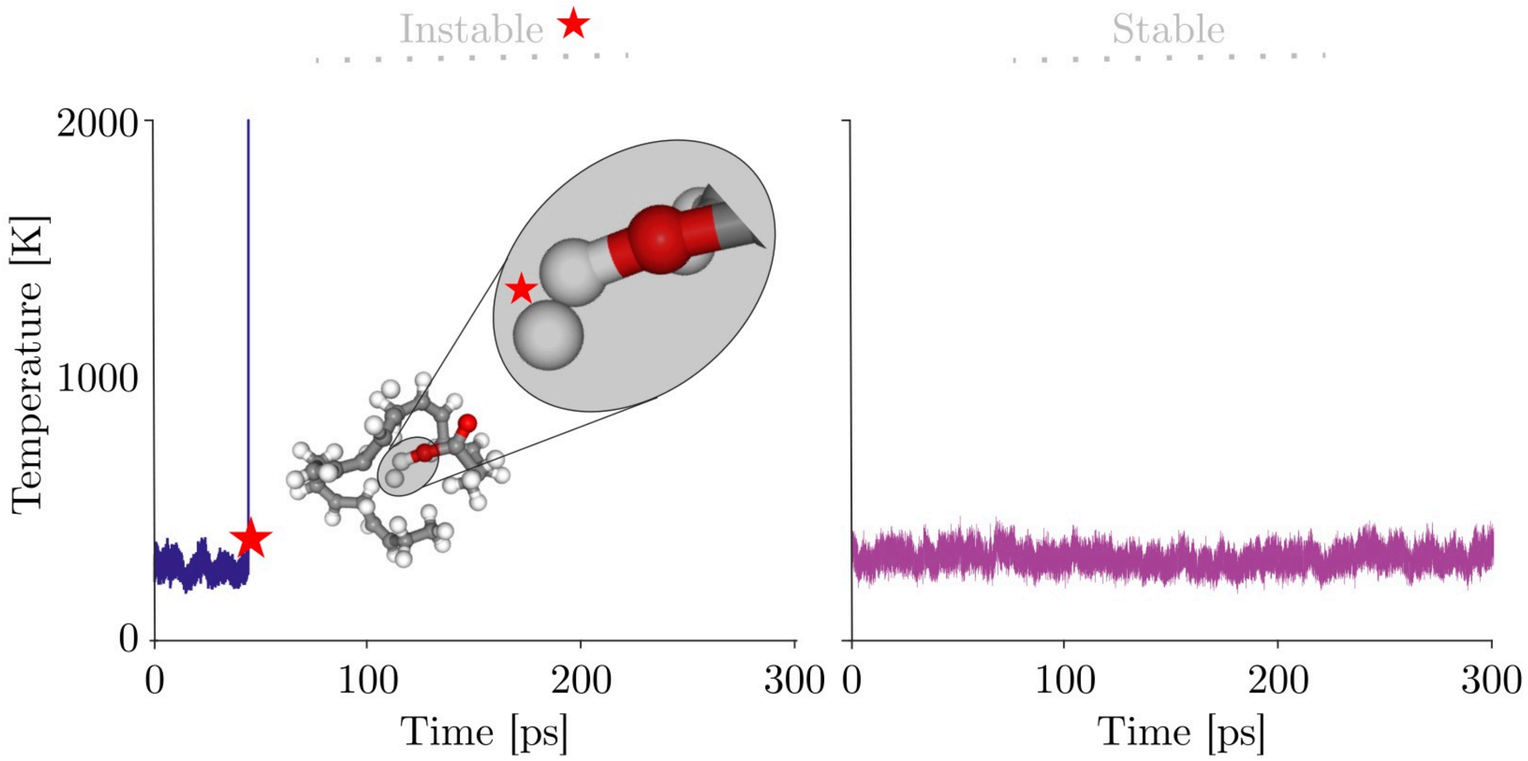
- Current models show a trade-off between data efficiency and time complexity
- **Less** time for data generation but **more** time for MD simulation





Molecular Dynamics Simulation Stability

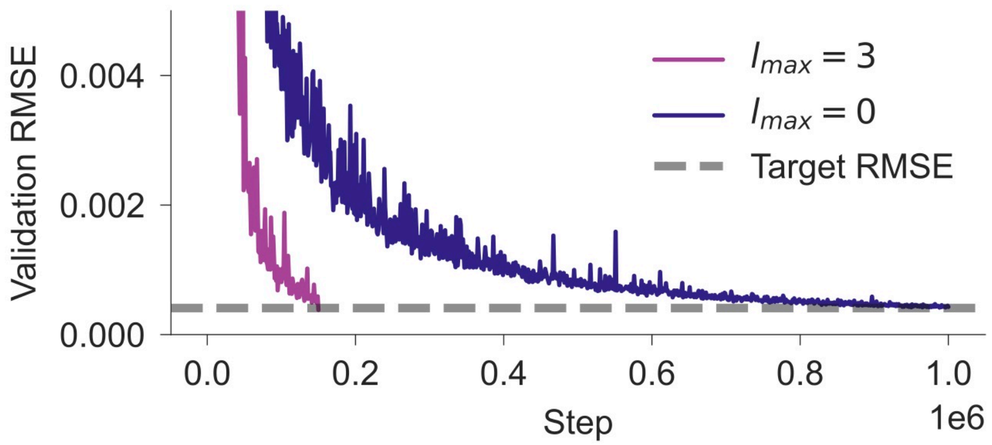
- *MD stability*: Simulation time over which the MD is stable?
- Long MD simulations are required to extract physical observables
- Instability due to non-physical configurations

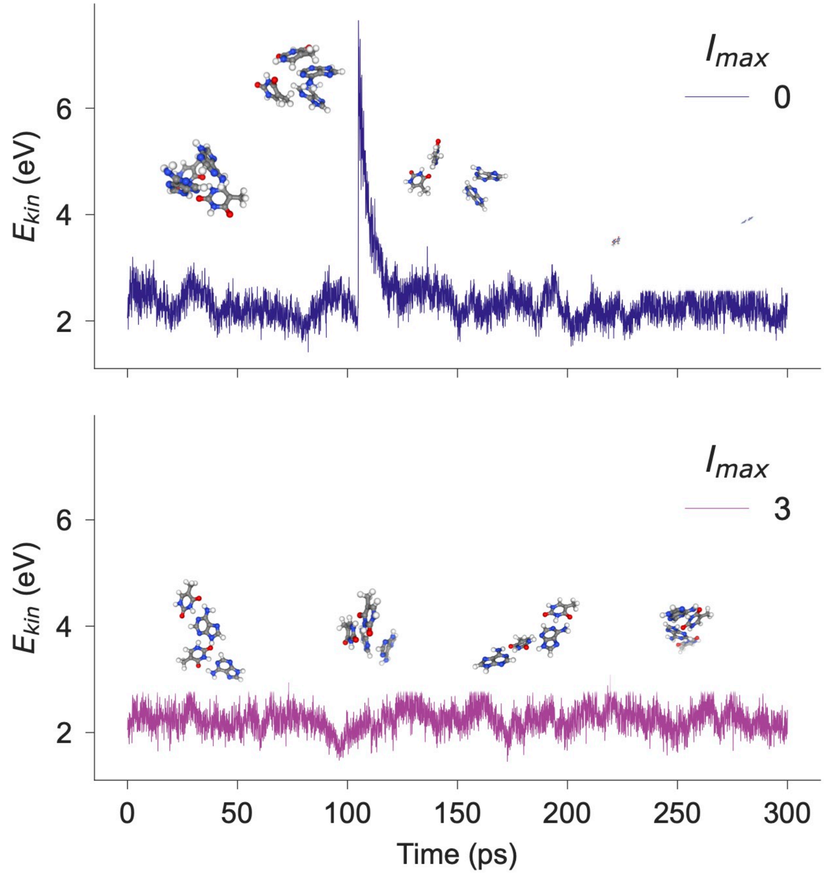


Frank, Thorben, et al. *To appear* (2023)

Equivariance Improves Stability

- Equivariance improves the stability
- Train the **same network**, to the **same error** with and without equivariant information

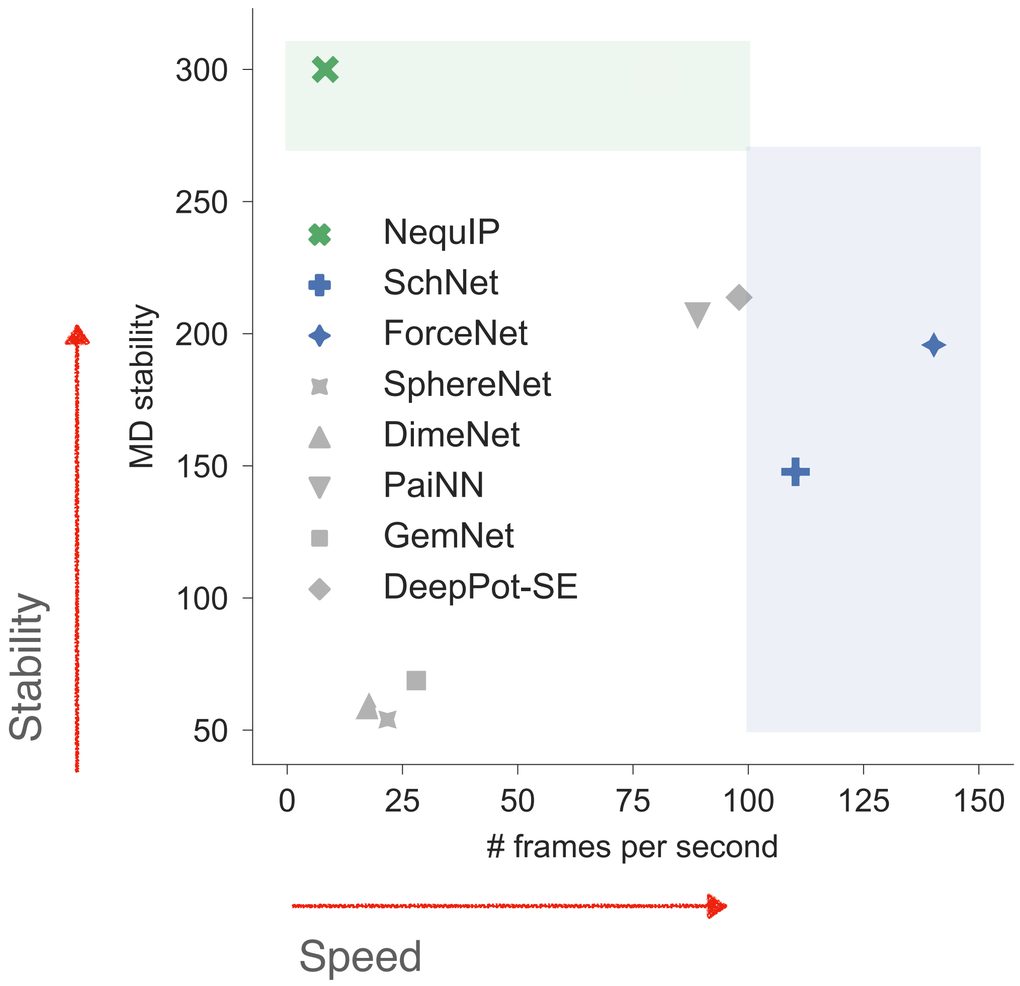




Frank, Thorben, et al. *To appear* (2023)

Molecular Dynamics Simulation Stability

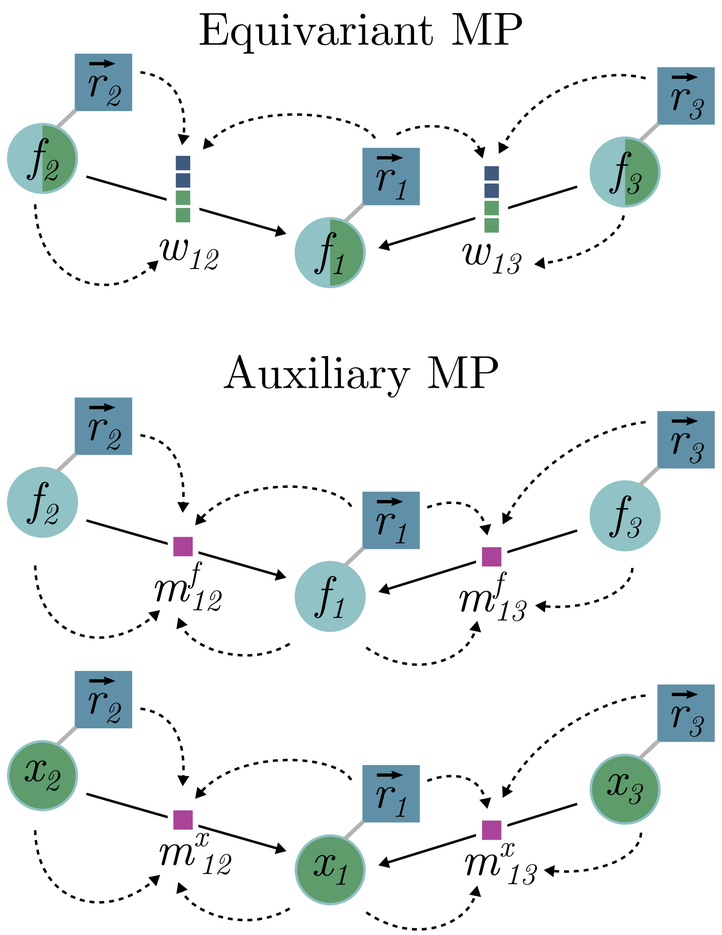
- Only model that is always stable *l = 3*
- Comes at the price of critical slow down of the model
- A similar trade-off as for data efficiency
- Reported times are for systems with 9 to 23 atoms only!
- Scaling to large structures ✗



Auxiliary Equivariant Message Passing

- Auxiliary equivariant message passing
- Introduce auxiliary geometric features $x_i \in \mathbb{R}^{d_G}$ which are updated during message passing

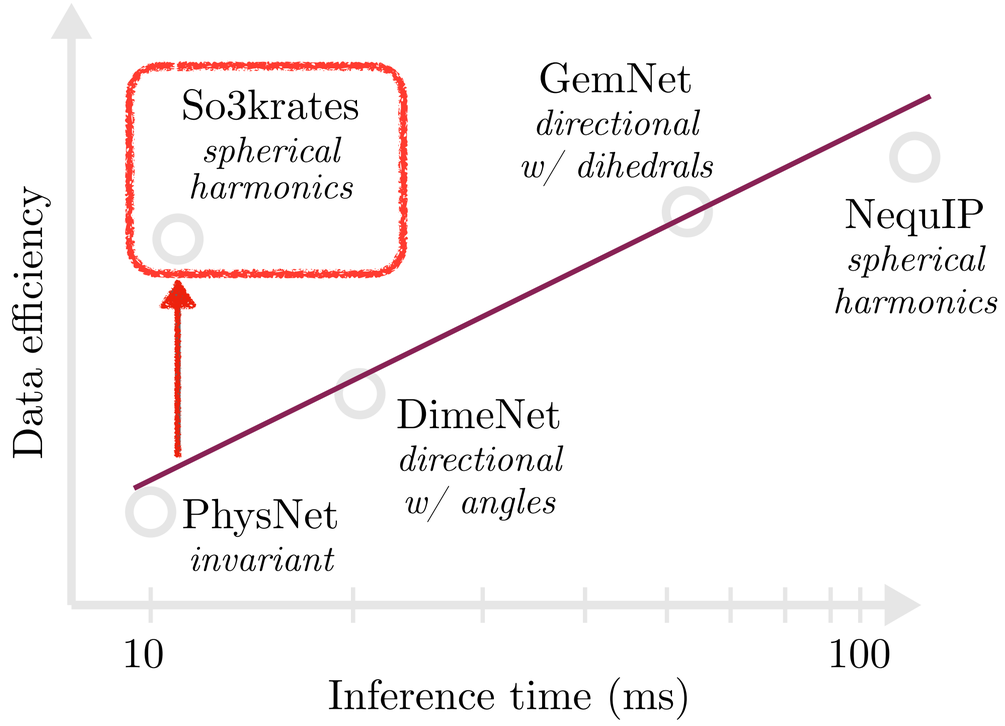
$$
\begin{aligned}
m_{ij}^{f|x} &= m^{f|x}(f_i, f_j, x_i, x_j, a_{ij}) \\
m_i^{f|x} &= \Sigma_{j \in \mathcal{N}_i} m_{ij}^{f|x} \\
f_i^{t+1} &= u(f_i^t, m_i^f) \\
x_i^{t+1} &= u(x_i^t, m_i^x).
\end{aligned}
$$

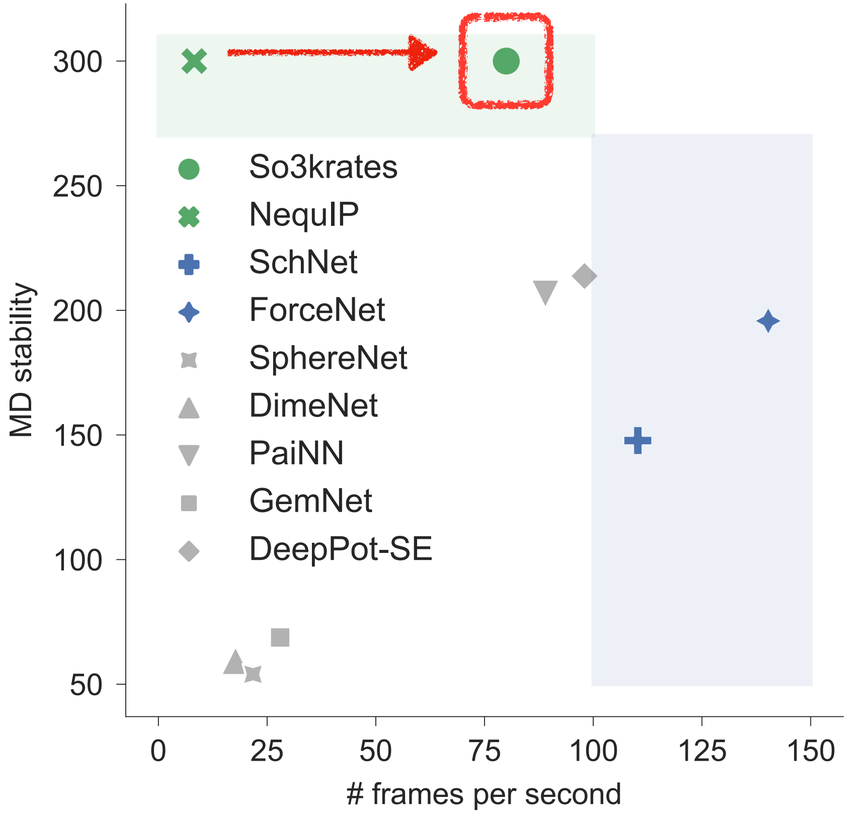


So3krates

Frank, Thorben, Oliver Thorsten Unke, and Klaus Robert Muller. *Advances in Neural Information Processing Systems*. 2022.

Data Efficient, Fast and Stable



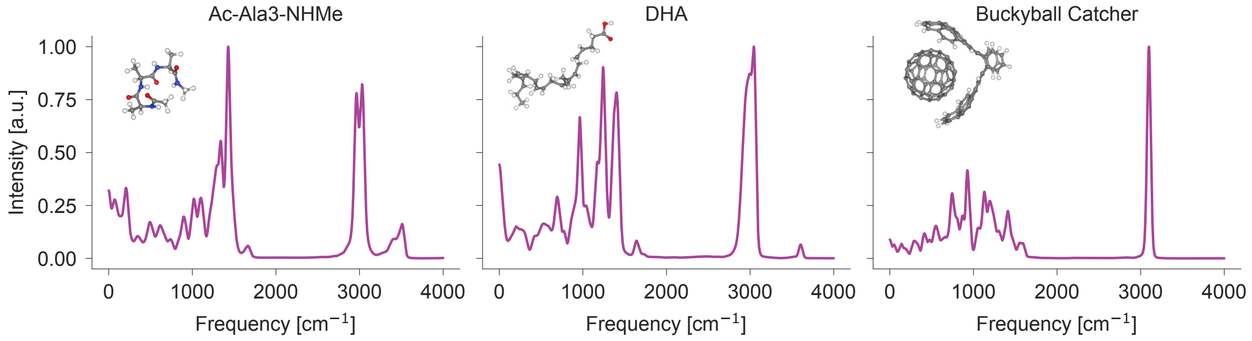


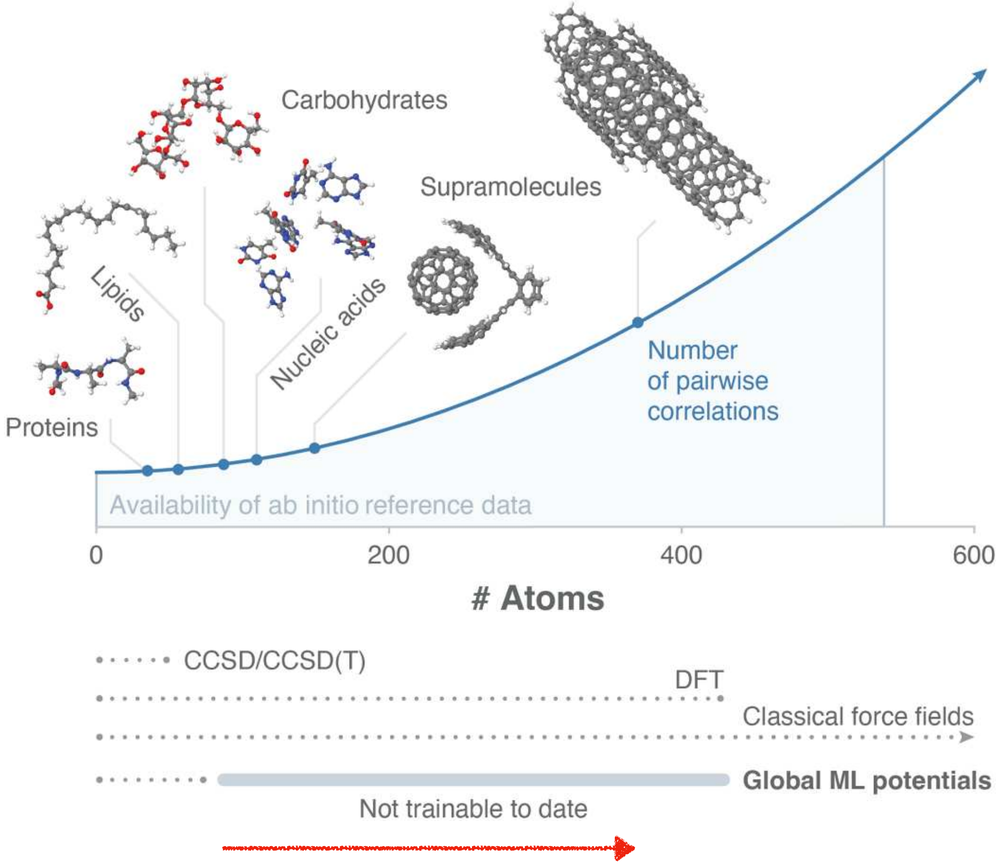
Scale to Large Molecular Structures

- What do we need to tackle large structures?
  - Data efficiency, due to high cost for reference data ✓
  - Stability, to obtain observables from the simulation ✓
  - Efficient models that are feasible to scale to large structures ✓
- Auxiliary message passing fulfils all requirements!

Scaling to Large and Complex Structures

- Calculate the velocity auto-correlation function which requires long and stable MD simulations
- Largest current structure: 370 atoms

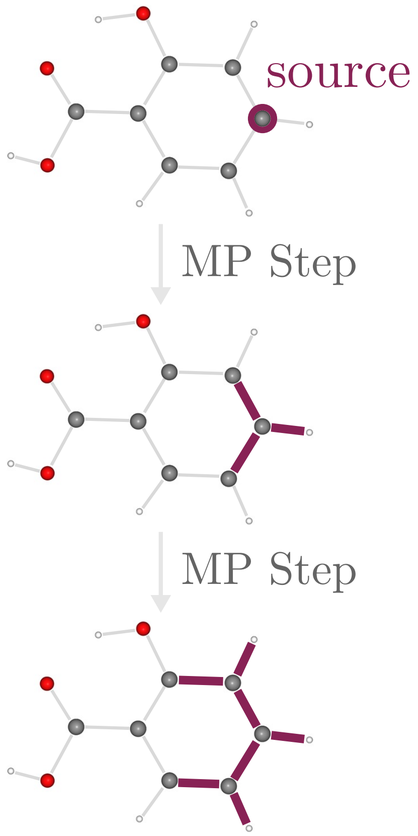




Global Machine Learning Potentials?

- ***Local*** message passing neural networks
- ***Global*** machine learning potentials
- Stacking of multiple message passing layers propagates information (diffusion)
- Effective cutoff

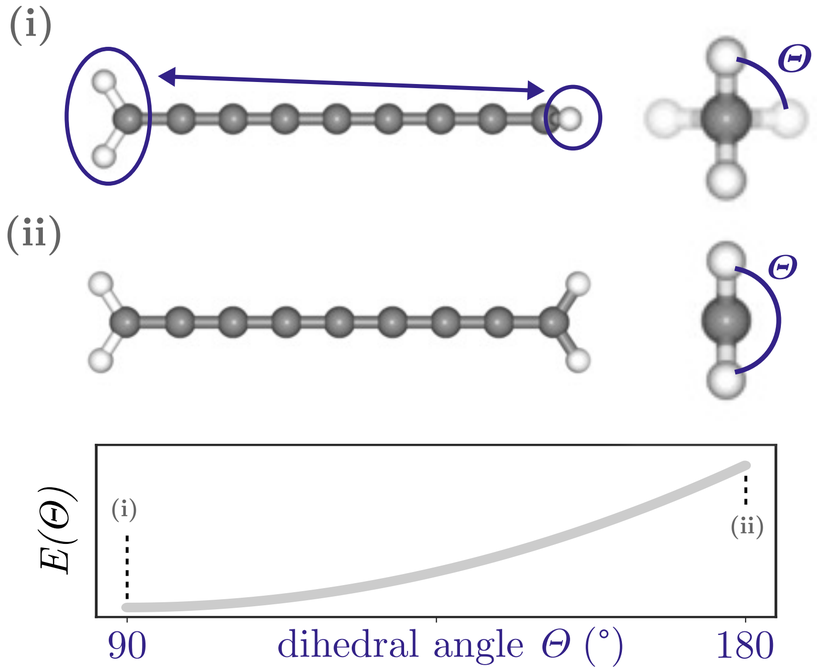
$$r_{\text{cut}}^{\text{eff}} = T \cdot r_{\text{cut}}$$



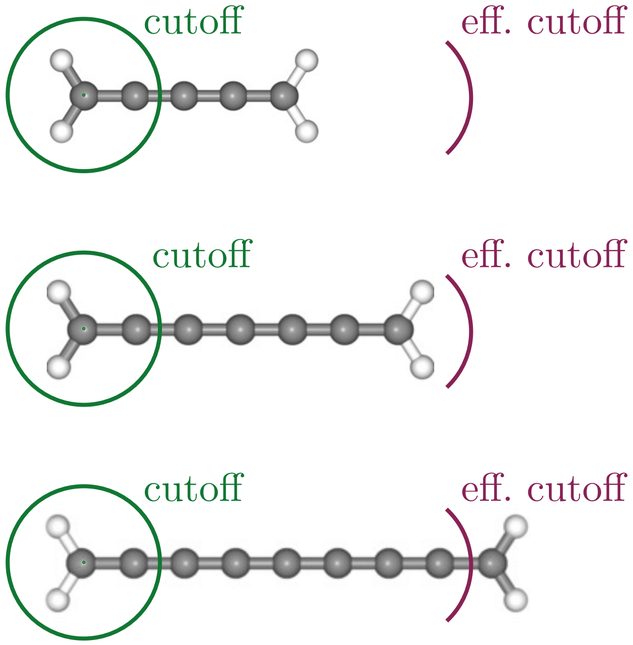
Global Machine Learning Potentials

Electronic Delocalization

- Electronic delocalization: Geometric dependencies on large length scales



$r_{\text{cut}} = 2.5$ and $T = 4$ layers

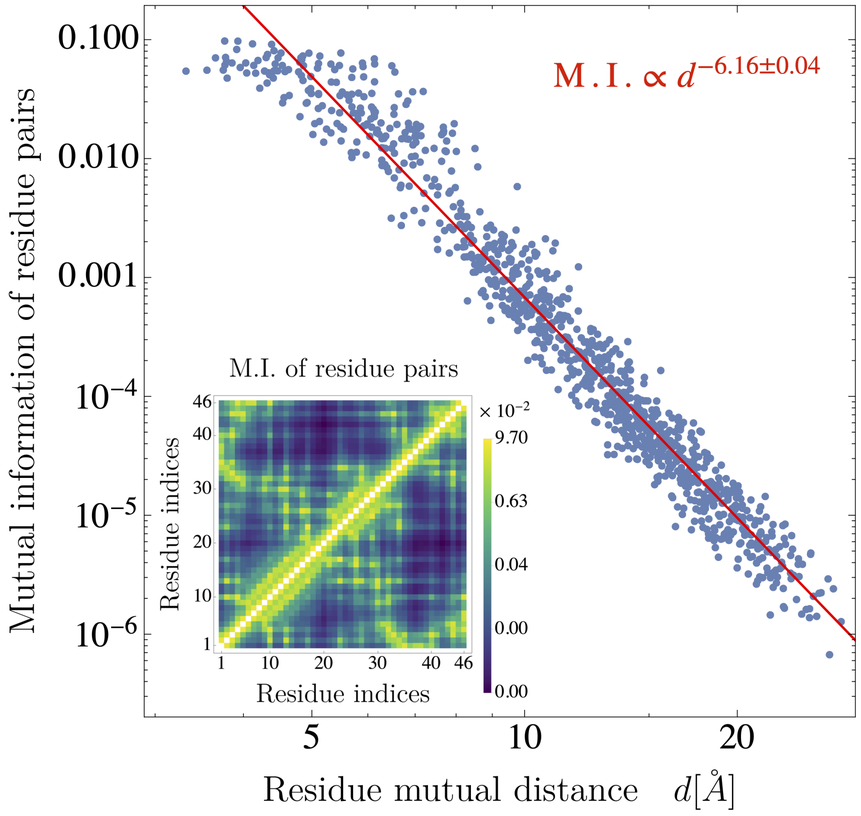


Global Machine Learning Potentials

Long Range Electrostatics, van-der-Waals Interactions, … ???

- Long-range many-body correlations
  - Mutual information extends over large length scales
- *Quick fix:* Get rid of locality assumption!
  - Quadratic scaling in number of atoms
    Unfeasible to train and run!
- Linear scaling without localisation??
  - Open problem!

[9]



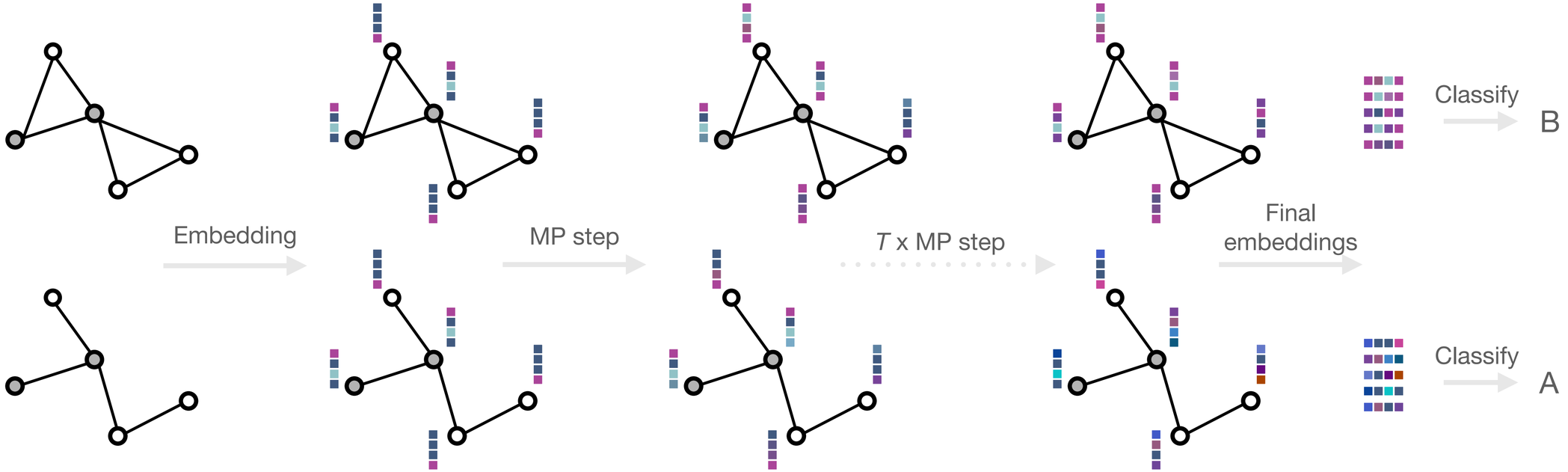
[9] Gori, Matteo, Philip Kurian, and Alexandre Tkatchenko. *arXiv preprint arXiv:2205.11549* (2022).

Thank you for your attention

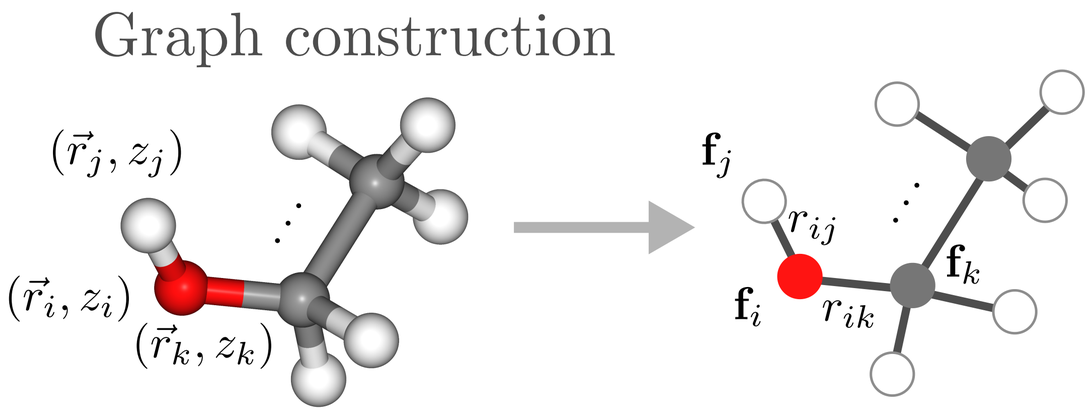
Thank you for your attention

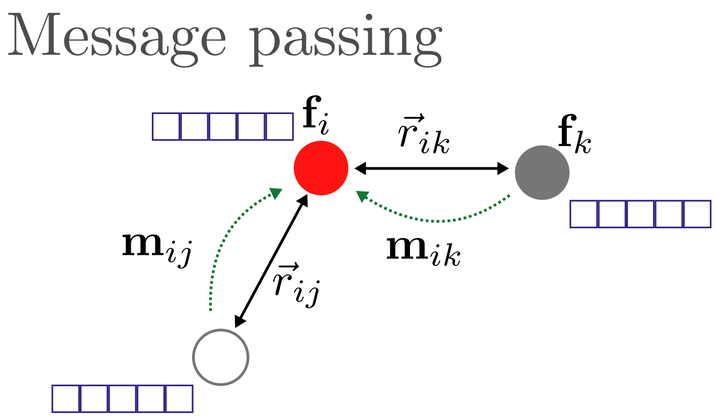
Message Passing Neural Networks

- Iterative updates of the node embeddings given their neighbourhood
- Use the final embeddings to make a prediction



Message Passing Neural Networks





- Construct atomic neighbourhoods $\mathcal{N}_i = \{j \mid r_{ij} \leq r_{\text{cut}}\}$ with $r_{ij} = \|\vec{r}_{ij}\|_2 = \|\vec{r}_j - \vec{r}_i\|_2$
- Potential energy is sum of *atomic energy* contributions

$$E_{pot} = \sum_{i=1}^{n} E_i \text{ with } E_i = g_\theta(f_i^{[T]})$$

Complexity of Equivariant Features

- Source of the complexity of equivariant features?
- Equivariant message passing step

$$m_i = \sum_{j \in \mathcal{N}_i} g(r_{ij}) \otimes Y^{(l)}(\hat{r}_{ij}) \odot f_i$$

Architecture	Operation	Scaling	$l_{\max}$
PhysNet	$\times$	$\mathcal{O}(n \times m \times F)$	0
PaiNN	$\times$	$\mathcal{O}(n \times m \times (l_{\max}+1)^2 \times F)$	1
SpookyNet	$\times$	$\mathcal{O}(n \times m \times (l_{\max}+1)^2 \times F)$	2
NequIP	$\otimes$	$\mathcal{O}(n \times m \times (l_{\max}+1)^4 \times F)$	3