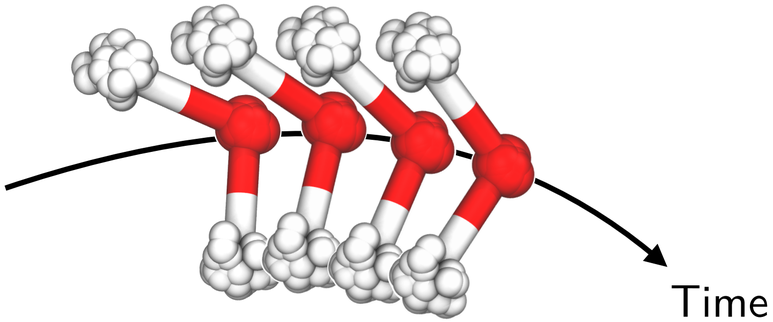
Machine learning interatomic potentials – active learning techniques

Elia Stocco, Shubham Sharma, Krystof Brezina and **Mariana Rossi**
Max Planck Institute for the Structure and Dynamics of Matter
Hamburg, Germany

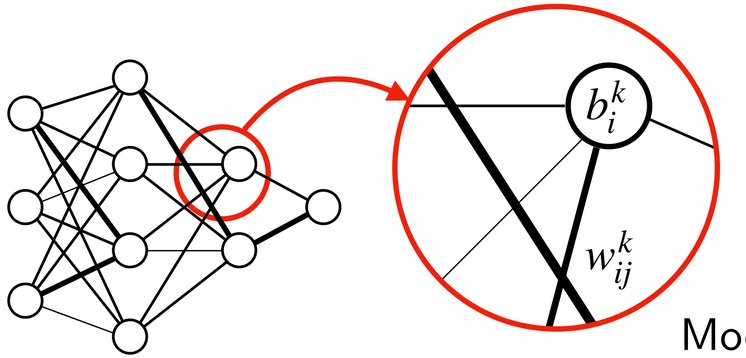
CNPEM/Ilum – Max Planck Meeting on Electronic Structure and Materials Informatics
Campinas, São Paulo, Brazil
July 3, 2025

mpsd
Max-Planck-Institut für
Struktur und Dynamik der Materie

Machine-learning interaction potentials (MLIPs)



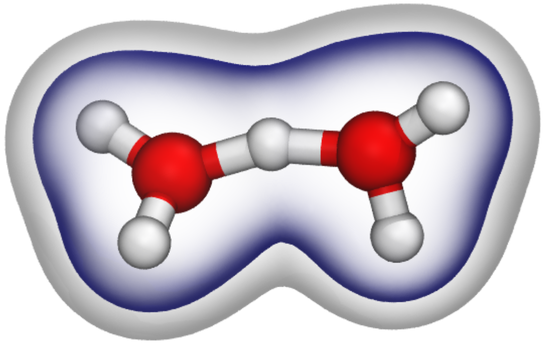
$$\mathbf{M}\ddot{\mathbf{q}} = -\nabla E(\mathbf{q})$$



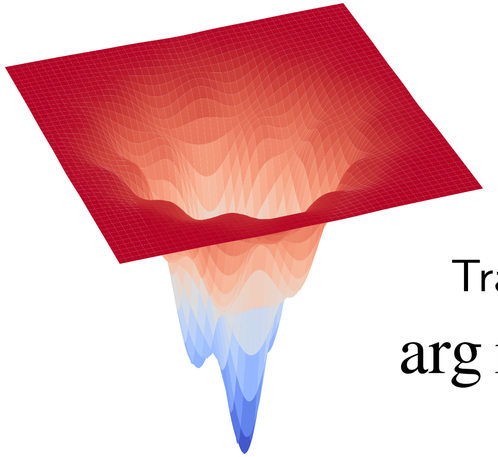
Model parameters

$\theta = (\mathbf{w}, \mathbf{b})$

$$E(\mathbf{q}) = f(\mathbf{q}; \theta)$$



$$E(\mathbf{q}) = \langle \Psi_0 | \hat{H}_{\text{electronic}}(\mathbf{q}) | \Psi_0 \rangle$$



Training

$$\arg\min_{\theta} \mathscr{L}(D, \theta)$$

What do we require from the training data set D?

- **Uniform** and **dense enough** coverage of the thermally available configuration space
- **Interpolation** vs. **extrapolation** prediction regime
- **Compact** and **robust** data sets

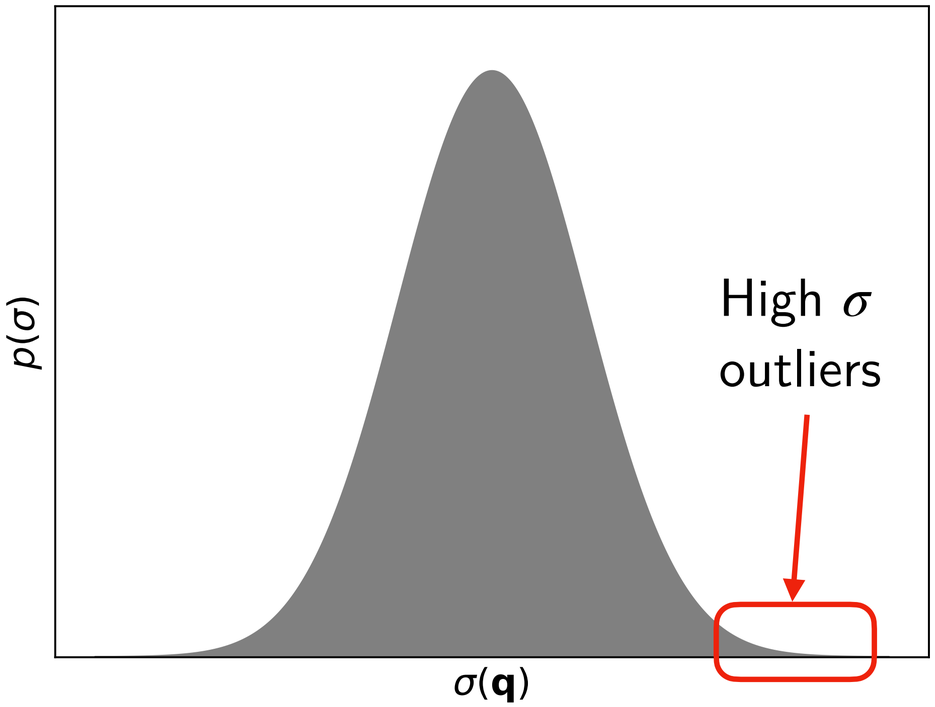
Tent-pole analogy of **good** training sets:

Active learning

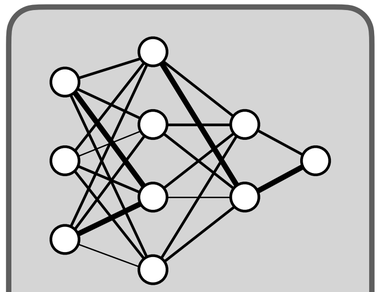
Active learning: selection of new data points based on some *uncertainty* measure σ

Regime 1: filtering a *labeled* candidate set

Regime 2: selecting relevant data from an unseen, exploratory candidate set (*unlabeled*)



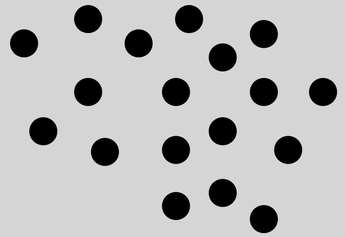
Regime 1: labeled candidates



NNP

$E(\mathbf{q}) = f(\mathbf{q};\theta)$

Prediction



Candidate set

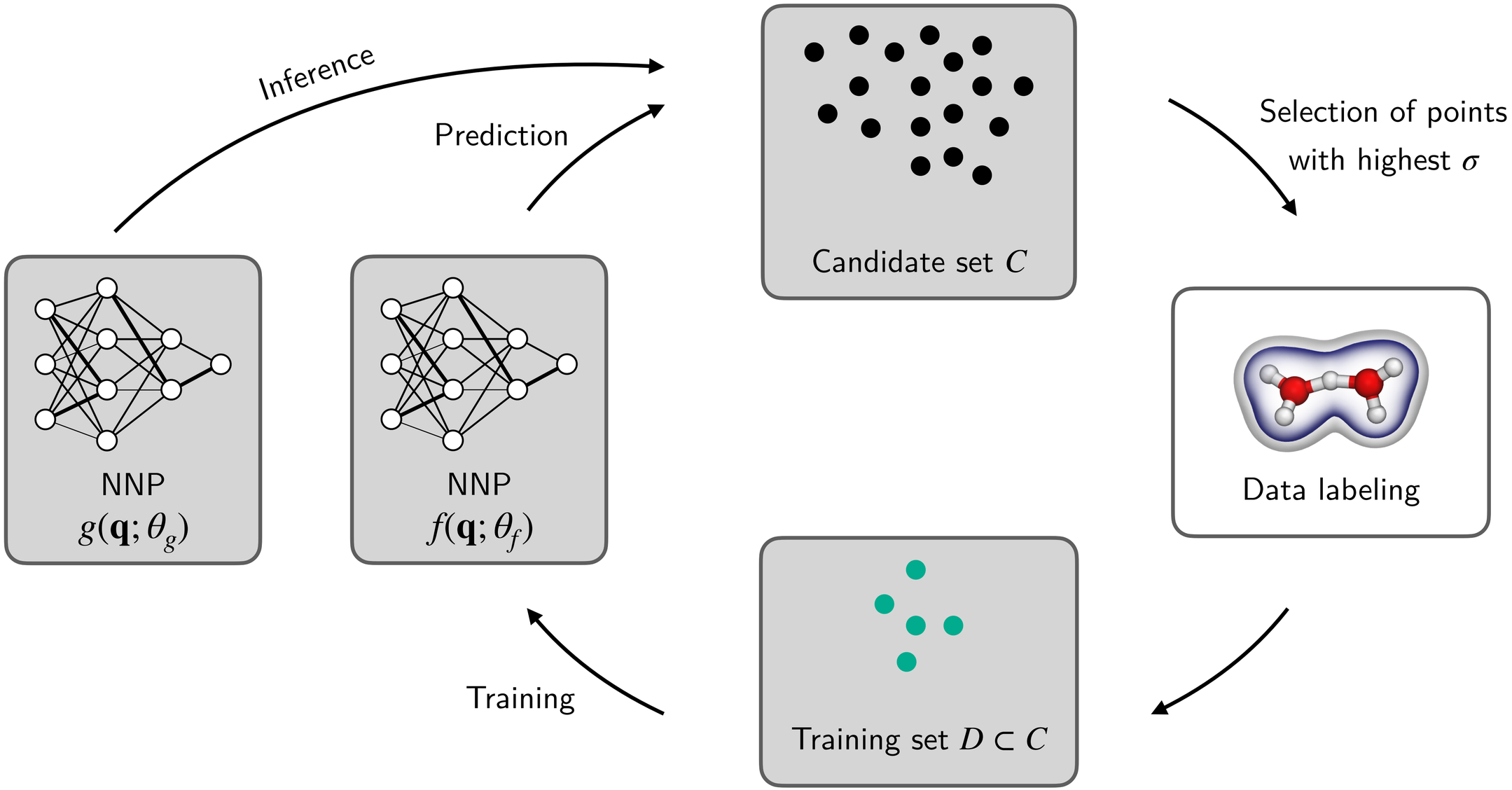
$C = \{\mathbf{q}_j, E(\mathbf{q}_j)\}_{j=1}^{N}$

Selection of points with highest σ

Training set $D \subset C$

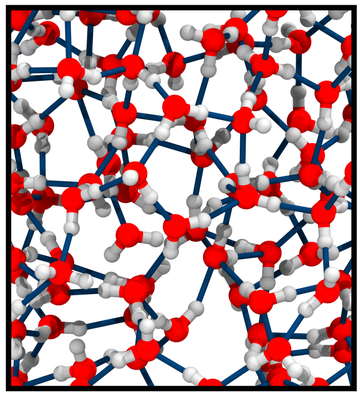
Training

Regime 2: refinement using unlabeled candidates



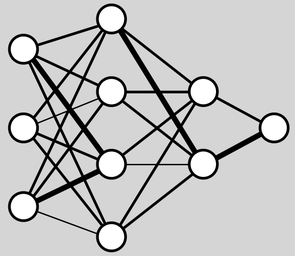
A real-world example: thermodynamic state points of water

Regime 1



Liquid, DFT

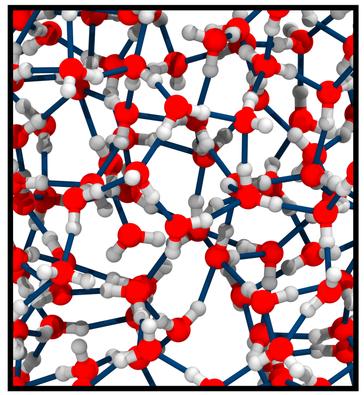
Active learning



NNP

$E(\mathbf{q}) = g(\mathbf{q}; \theta_g)$

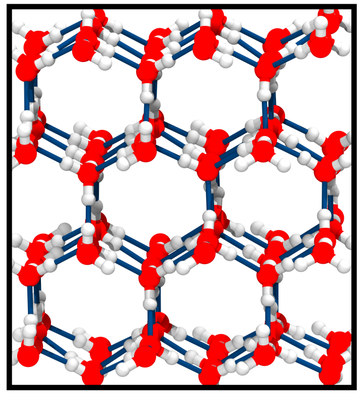
Inference



Liquid, NNP

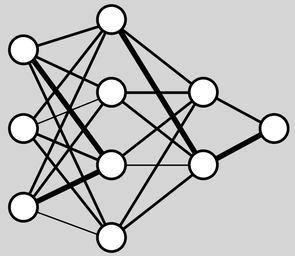
Inference

Regime 2



Ice I_h

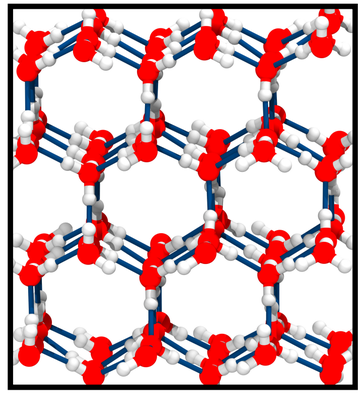
Active learning
+ labeling



NNP

$E(\mathbf{q}) = f(\mathbf{q}; \theta_f)$

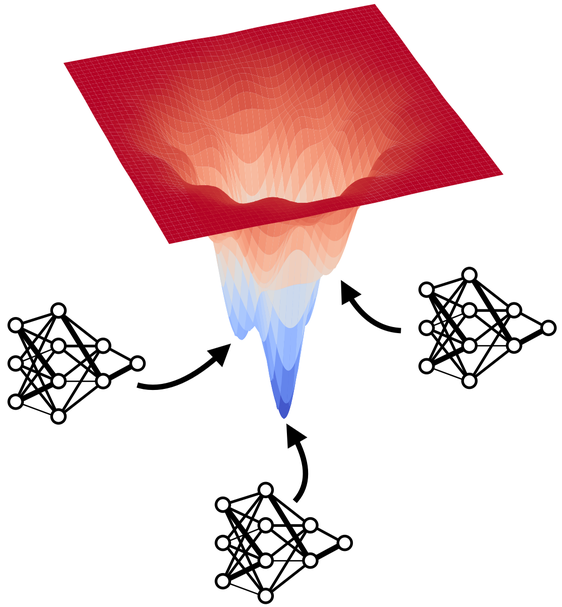
Inference



Ice I_h, NNP

Schran C., Brezina K., Marsalek O. JCP, 153, 104105, **2020**

Committee models & Query by committee

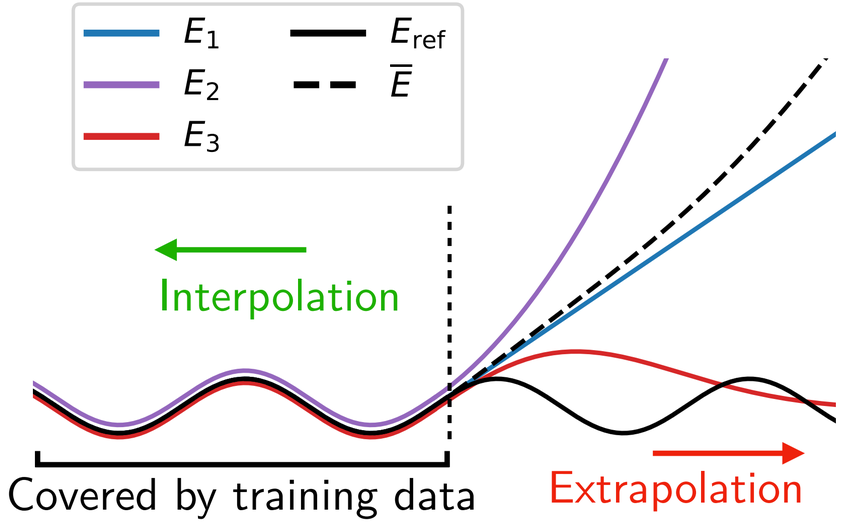


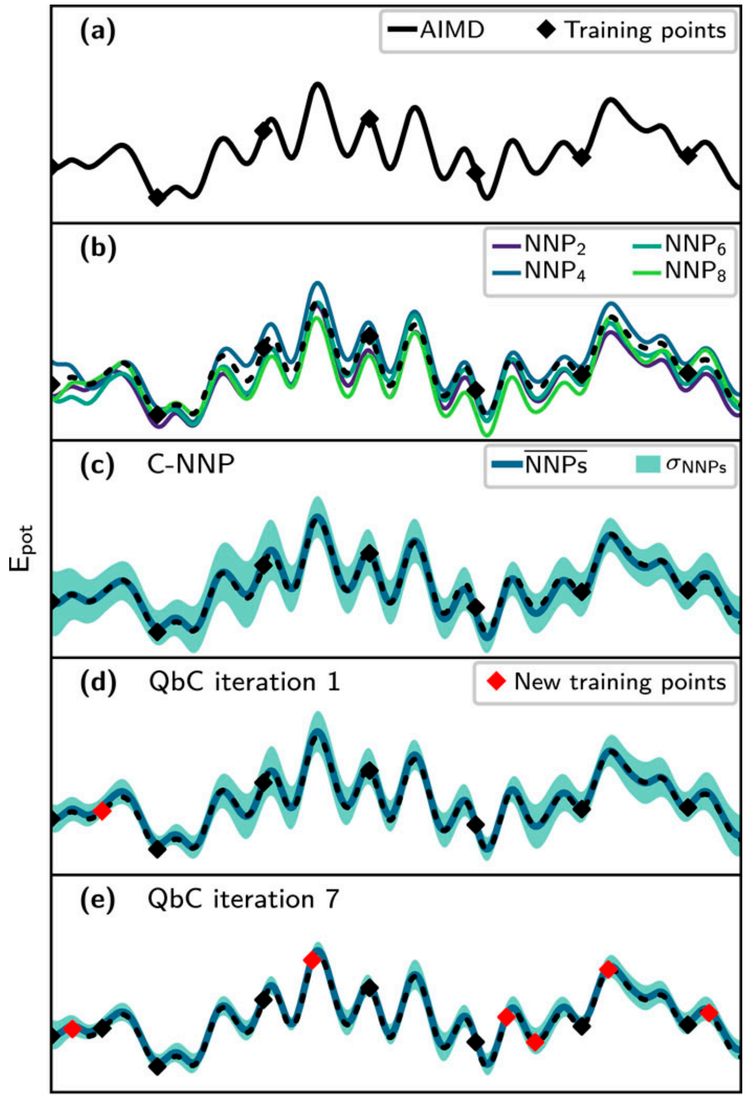
$\mathbf{q}$ → C-NNP → $\{E_i(\mathbf{q})\}_{i=1}^{N_c}$, $\overline{E}(\mathbf{q})$, $\sigma_E(\mathbf{q})$

Committee disagreement

Low disagreement 📉 → low generalization error 🥰

High disagreement 📈 → high generalization error 😭





Schran C., Brezina K., Marsalek O. JCP, 153, 104105, **2020**

Disagreement: energies or forces?

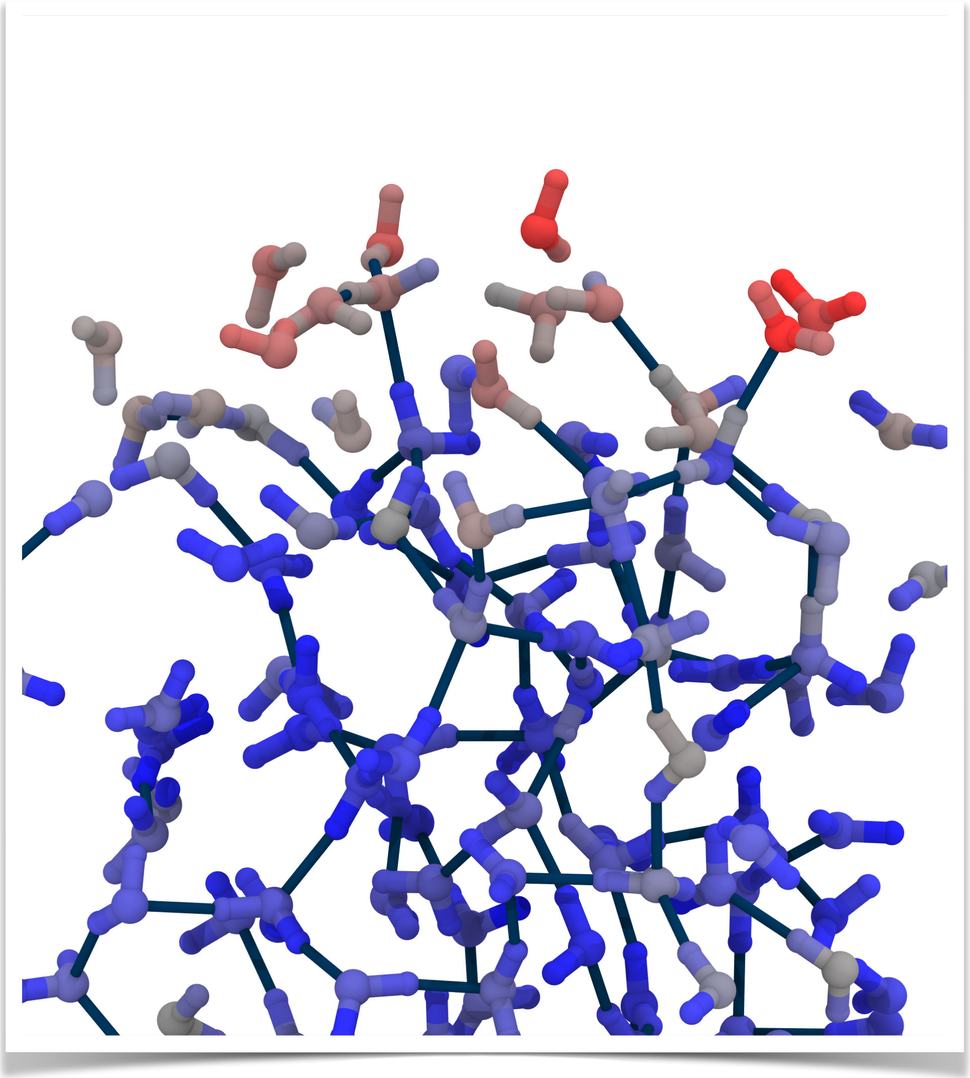
$$\sigma_E(\mathbf{q}) = \sqrt{\frac{1}{N_\mathrm{c}} \sum_{n=1}^{N_\mathrm{c}} \left[E_n(\mathbf{q}) - \overline{E}(\mathbf{q})\right]^2}$$

$$\overline{\sigma}_F(\mathbf{q}) = \frac{1}{N} \sum_{\alpha=1}^{N} \sqrt{\frac{1}{N_\mathrm{c}} \sum_{n=1}^{N_\mathrm{c}} \left[\mathbf{F}_{\alpha n}(\mathbf{q}) - \overline{\mathbf{F}}_\alpha(\mathbf{q})\right]^2}$$

Atoms

Committee members

Force disagreement is typically used in active-learning applications for practical reasons.



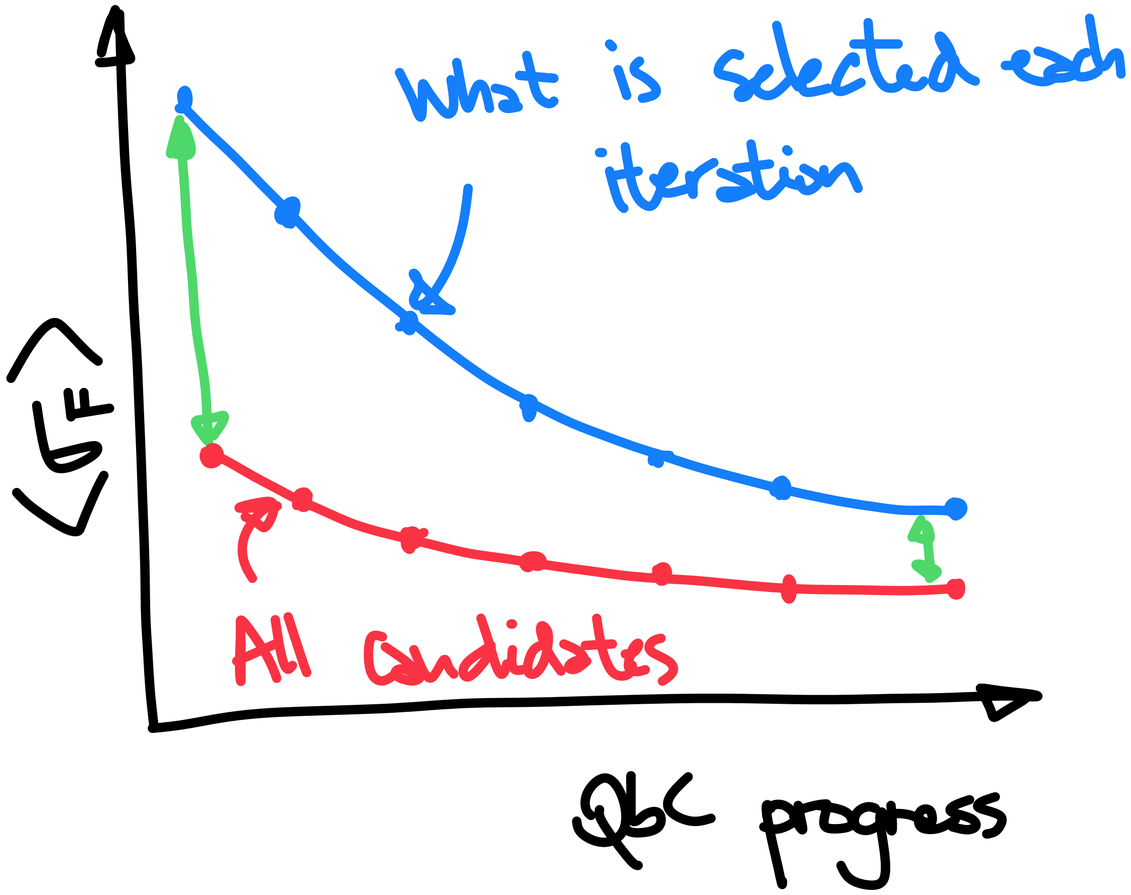
Schran C., Brezina K., Marsalek O. JCP, 153, 104105, **2020**

Takehome messages: controlling QbC convergence

QbC goal: extract as much valuable **unlearned information** from the candidate set in as **few samples** as possible

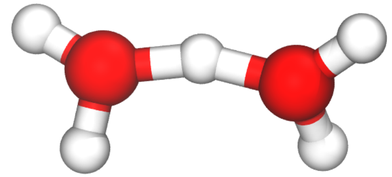
Convergence: this happens when the σ for the newly selected batch of structure **plateaus** and **approaches** the σ for the remaining candidates

Generalization error: acceptable error over an independent **test set**

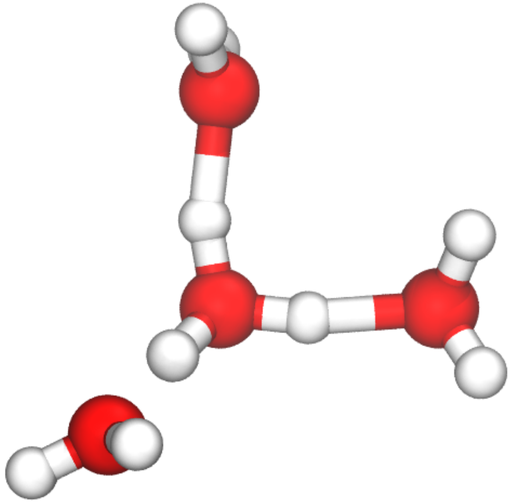


Goals of the this tutorial

- **Actively select labeled data points** for the Zundel cation using **QbC** with a committee of **MACE [1]** models
- Explore the properties and the evolution of **committee disagreement** over the course of the QbC cycle
- **Compare the quality** of active-learned and random-sampled models
- Perform **exploratory** inference MD using the Zundel-trained model on the Eigen cation
- Use QbC with on-the-fly data labeling with **FHI-aims [2]** to **improve** the initial Eigen model



Zundel cation ($H_5O_2^+$)



Eigen cation ($H_9O_3^+$)

[1] Batatia I. et al. NeuIPS, 36, **2023**
[2] Abbot J. W. et al. arXiv:2505.00125, **2025**