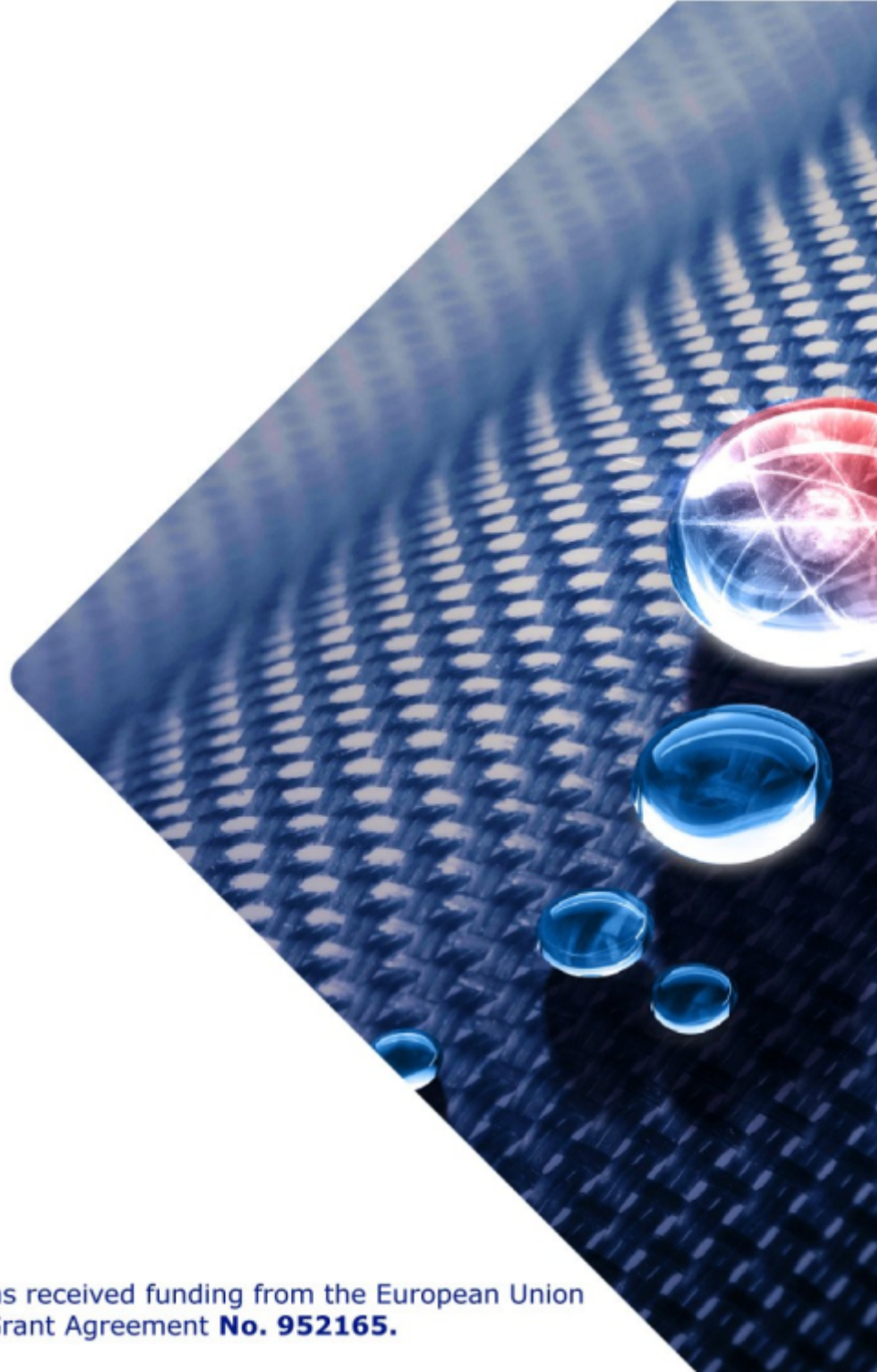
Targeting Real chemical accuracy at the EXascale

TREX Hackathon II

Matthias Rupp, Thomas Bischoff

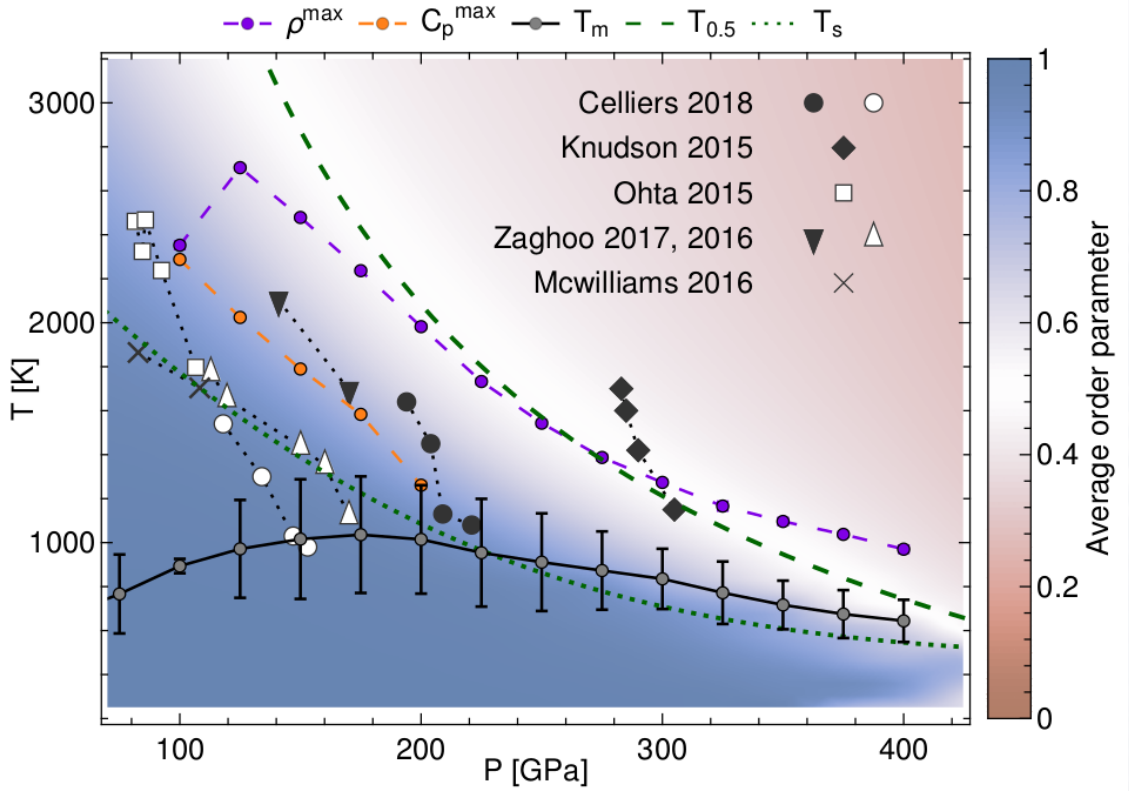
03/03/2022



Targeting Real Chemical Accuracy at the Exascale project has received funding from the European Union Horizoon 2020 research and innovation programme under Grant Agreement **No. 952165.**

Motivation

Phase diagram of hydrogen under pressure :



No consensus in the literature yet :

Cheng, Mazzola, Pickard, Ceriotti, Nature 585, 2020
Tirelli, Tenti, Nakano, Sorella, arXiv 2021
...

→ **Solve this question using QMC and ML !**

ML techniques

Direct learning:

$$\underbrace{\text{QMC}}_{\text{very accurate/heavy}} \longrightarrow \underbrace{f_{\text{ML}}(\text{QMC})}_{\text{very cheap \& not data efficient}}$$

Δ-learning using a DFT baseline:

$$\underbrace{\text{QMC}}_{\text{very accurate/heavy}} \longrightarrow \underbrace{\text{DFT}}_{\text{less accurate/heavy}} + \underbrace{\Delta_{\text{ML}}(\text{QMC - DFT})}_{\text{very cheap \& data efficient}}$$

Δ-learning using a UFP baseline:

$$\underbrace{\text{QMC}}_{\text{very accurate/heavy}} \longrightarrow \underbrace{\text{UFP}}_{\substack{\text{very cheap} \\ \textcircled{1}}} + \underbrace{\Delta_{\text{ML}}(\text{QMC - UFP})}_{\substack{\text{very cheap \& data efficient} \\ \textcircled{2}}}$$

Ultrafast potentials (UFPs) as baseline

Many-body expansion of the energy:

$$E = \underbrace{\sum_{i,j} V_{2b}(r_{ij})}_{\text{two-body terms}} + \underbrace{\sum_{i,j,k} V_{3b}(r_{ij}, r_{ik}, r_{jk})}_{\text{three-body terms}}$$

B-Spline basis for potential terms:

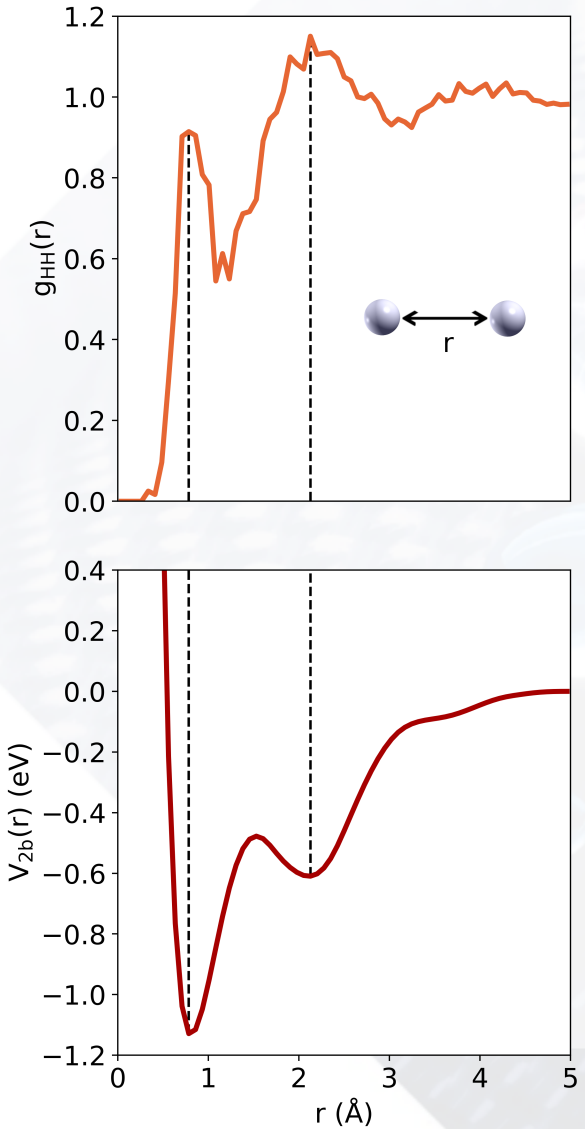
$$V(r) = \sum_n c_n B_n(r)$$

Advantages:

→ fast to evaluate

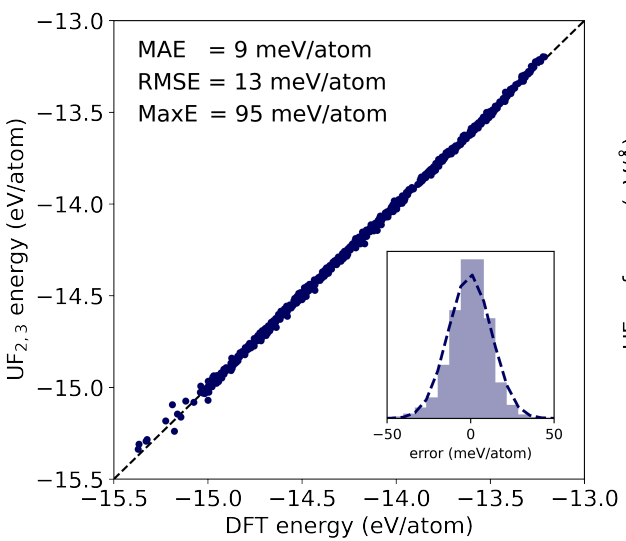
→ intuitive physical interpretation

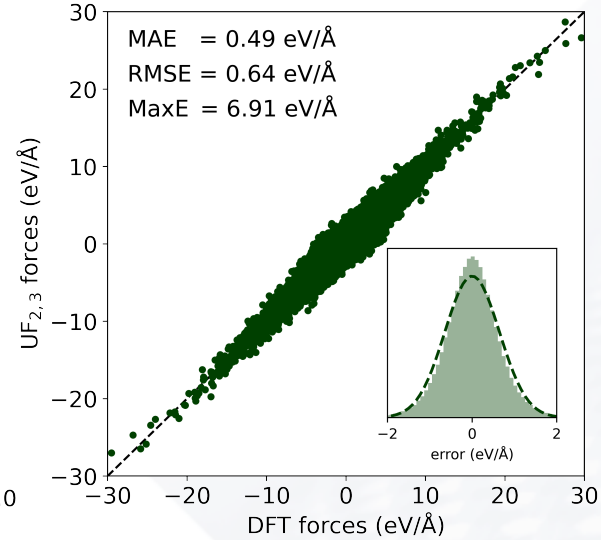
→ robust against holes in the dataset



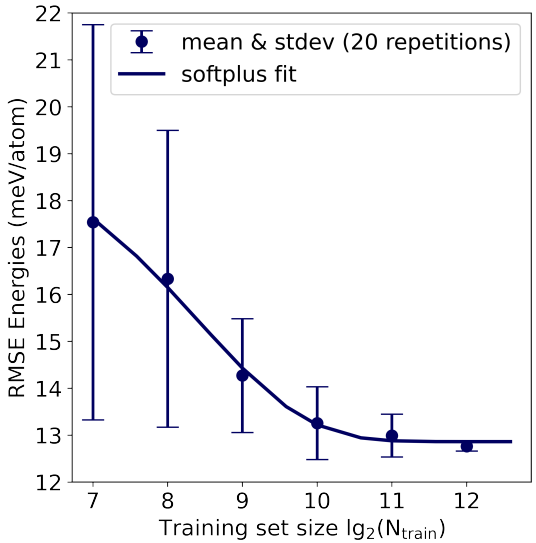
Ultrafast potentials (UFPs) as baseline

Prediction of energies and forces

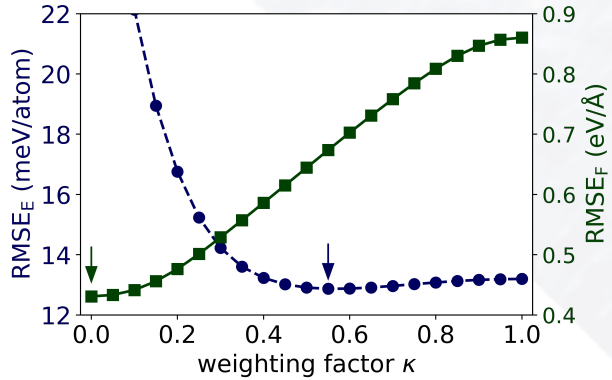




Learning curve



Weighting parameter



Δ - learning

Options for Δ - learning :

	linear model	kernel model	neural networks
interpretability	$++$	$+$	$-$
data efficiency	$++$	$+$	$-$
cost (of training)	$++$	$+$	$--$
achievable accuracy	$-$	$++$	$++$