Data-Efficient Training of Machine Learning Potentials for Aqueous-Phase Heterogeneous Catalysis

Maciej Baradyn[1], Karlo Sovic[1], Johannes T. Margraf[1]

[1]University of Bayreuth, Bavarian Center for Battery Technology (BayBatt), Weiherstraße 26, 95448 Bayreuth, Germany

Machine-learning interatomic potentials (MLIPs) are emerging as powerful tools for extending atomistic materials simulations beyond the length and time scales accessible to first-principles methods. However, their application to chemically and structurally complex materials environments remains limited by the amount and diversity of reference data required for reliable training. This challenge is particularly pronounced at solid–liquid interfaces, where explicit solvent substantially expands the configurational space that must be represented.

Here, we investigate a data-efficient approach for adapting the MACE-MH-1 foundation model to molecular dynamics simulations of aqueous heterogeneous catalytic interfaces, using glycerol adsorbed on Cu(111) in explicit water as a representative system. Rather than generating reference data directly for the full extended interface, we decompose the target chemical environment into smaller, physically motivated subsystems that capture the relevant adsorbate–surface, solvent–surface, and adsorbate–solvent interactions. Using only 25–100 configurations per training dataset, we compare naive fine-tuning with multihead replay fine-tuning and evaluate model accuracy on independent test sets.

Beyond conventional energy and force errors, we examine the ability of the resulting models to transfer to an extended Cu–glycerol–water interface that lies outside the fine-tuning distribution. Molecular dynamics simulations are further validated through liquid-water properties and structural and dynamical observables of the solid–liquid interface. The fine-tuned models reproduce physically meaningful behavior consistent with the underlying DFT reference while requiring only a limited amount of system-specific training data.

Our results highlight how pretrained interatomic foundation models, combined with chemically informed data selection, can reduce the cost of developing accurate machine-learning potentials for complex materials environments. More broadly, this work illustrates the potential of data-efficient AI approaches to enable transferable atomistic simulations of multicomponent interfaces relevant to materials science and heterogeneous catalysis.