The Open Catalyst 2025 (OC25) Dataset and Models for Solid-Liquid Interfaces

Sushree Jagriti Sahoo [1] **Mikael Marschin** [2] **Daniel S. Levine** [1] **Zachary W. Ulissi** [1] **C. Lawrence Zitnick** [1] **Joel B. Varley** [3] **Joseph A. Gauthier** [2] **Nitish Govindarajan** [4] **Muhammed Shuaibi** [1]

[1]*FAIR at Meta* [2]*Department of Chemical Engineering, Texas Tech University, Lubbock, TX 79409, USA* [3]*Materials Science Division, Lawrence Livermore National Laboratory, Livermore, CA 94550, USA* [4]*School of Chemistry, Chemical Engineering and Biotechnology, Nanyang Technological University, 21 Nanyang Link, Singapore 637371, Singapore*. Correspondence to: Nitish Govindarajan nitish.govindarajan@ntu.edu.sg.

1. Introduction

Over the past few years, the development of MLIPs has revolutionized the field of atomistic simulations. What used to be an onerous task that individual research groups would tackle for each system of interest, creating significant redundancy of effort across the community, is increasingly becoming a matter of downloading freely available foundational MLIPs and fine-tuning them for specific system(s) of interest. This has been enabled by the availability of large, open datasets spanning a variety of systems and applications in materials science: bulk materials, crystalline organic frameworks, molecules, and catalysts to train generalized MLIPs. However, the solid-liquid interface, crucial to understanding electrochemical transformation technologies, is largely missing from these open datasets. It should come as no surprise then, that a MLIP that has never been explicitly shown how long-range charges interact, or how a metal surface might behave in the presence of explicit water/solvent molecules, may struggle to adequately model the structure and properties of the electrified double layer crucial to the understanding of electrocatalytic processes. While recent dataset efforts have tried to address this, they are rather limited in both size and chemical diversity, often exploring very specific chemistries and applications.

In this work, we introduce the Open Catalyst 2025 (OC25) dataset [1] and models for solid-liquid interfaces from Meta FAIR to broaden the capabilities of generalized foundational MLIPs and to shed light on fundamental aspects of the (electro)chemical transformation technologies that are likely to be critical in our transition to a low carbon economy. The OC25 dataset contains over 7 million single-point DFT calculations sampled from a wide range of transition metal and metal oxide surfaces in contact with several relevant electrolytes, including different cations and anions that have been used in experimental studies. We train several baseline models on the OC25 dataset, demonstrating their utility over existing state-of-the-art models. The dataset, models, and code are all open source and freely available to allow the community to expand upon the work we demonstrate here.

2. Substantial section

The OC25 dataset is constructed from a combination of DFT relaxations and molecular dynamics simulations of catalyst/solvent/ion structures. Structures consist of a catalyst surface, a solvent, and at least 1 adsorbate molecule. Structures may also contain multiple adsorbates and an ion present to better capture reactive environments and electrocatalytic systems. Surfaces are sampled from unique bulk structures in the Materials Project, including oxides. Eight commonly used solvents are sampled across varying solvent depths, with system sizes of 144 atoms on average. Adsorbates are randomly sampled from 98 unique molecules, including OC20 species and additional reactive intermediates. Overall, the OC25 dataset consists of 7,801,261 single-point calculations, spanning 1,511,270 unique systems and 88 unique elements. OC25 structures correspond to highly off-equilibrium configurations, a property that we have seen to aid in training ML models. To accomplish this, short time-scale MD simulations at high temperature (1000K) were run, minimizing the redundancy in structures that can come from fully relaxed configurations, i.e. OC20 and OC22. The inclusion of off-equilibrium configurations is reflected in the force distributions of OC25 (Figure 2), with forces much higher than those of OC20 and OC22.

OC25 constitutes the largest and most diverse solid-liquid interface dataset available, encompassing a broad spectrum of unique surfaces, solvents, ions, and adsorbate combinations. Despite its extensive coverage of solid-liquid interfaces, OC25 has inherent limitations. Although the systems included are approximately twice as large as those in OC20/OC22, they still fall short of the very large cell sizes needed to rigorously capture bulk-like solvent effects and lower ion concentrations that are often utilized in experiments. The dataset features average ion concentrations of ~2M, with a minimum of 0.38M. This is substantially more concentrated than many experiments in electrocatalysis; however, we also note that under reaction conditions and applied potentials, ion concentrations will generally be higher near the (charged) interface compared to the bulk electrolyte. Solvent depths average around 5.6, with a maximum of 10, corresponding to only a limited number of solvent layers. Although the models in this work may still be able to accurately predict interfacial properties and reactivity, these aspects remain to be tested in future studies. During the generation of this dataset, spin polarization was disabled. As a result, materials for which magnetism is an important consideration may not be accurately represented, and may require

additional fine-tuning.

An important quantity that is currently not explored in this work is the interface workfunction, which can be related to the electrochemical potential measured in experiments. Estimating the workfunction for surface/electrolyte/vacuum interfaces involves evaluating two quantities: the Fermi level and the vacuum potential above the electrolyte. Although interface workfunctions can be estimated by performing single-point DFT evaluations on representative configurations from molecular dynamics simulations using the OC25 models, it would be highly desirable to have ML models that can predict the interface workfunctions directly. Access to the interface workfunction during the simulation can also enable constant potential (grand-canonical) simulations to estimate the thermodynamics and kinetics of electrochemical reactions. Developing models that can accurately predict both the Fermi level and the vacuum potential, and in turn the interfacial workfunction, is an exciting direction for future research in the study of solid-liquid interfaces.

Models trained on OC25 achieve strong predictive accuracy, with energy and force errors reaching as low as 0.105 eV and 0.015 eV/, respectively for the eSEN-S-cons model. Out-of-distribution errors are larger for unseen solvents and ions, providing opportunities for improving model generalizability. Novel modeling approaches to better account for long-range interactions and charge can be especially useful for this dataset due to the presence of electrolytes.

The curation of OC25 provided valuable insights into the convergence and consistency of DFT-calculated force labels and its impacts on MLIP training. In practice, ensuring that the net force across the system is zero is particularly important in training models directly on AIMD data, as some DFT codes (e.g., VASP) do not automatically remove this drift. Additionally, we observe that filtering out calculations where the total force drift exceeds acceptable thresholds can improve data quality. For OC25, this threshold was chosen to be 1 eV/, but we encourage future efforts to perform their own analysis. Perhaps most surprisingly, models trained on the unfiltered, less-converged data were able to predict the more tightly converged set of forces very accurately, suggesting that models possess a notable resilience to moderate label noise.

Building on the success of OC20, OC25 introduces explicit modeling of the solid-liquid interface in catalytic systems. Covering an even broader range of chemical diversity, the OC25 dataset, along with its baseline models, hopes to serve as a valuable resource for the broader scientific community.

2.1 Related work

The Open Catalyst 2020 and 2022 datasets (OC20/OC22), Catalysis-Hub, GASpy, and AQCat represent the largest catalyst datasets for training MLIPs. The diversity and scale of these datasets have led to the development of several state-of-the-art ML models for the community. Although models trained on these datasets have demonstrated successful applications for gas-phase heterogeneous catalysis, little has been done for solid-liquid and electrified solid-liquid interfaces. The Open Catalyst 2025 (OC25) dataset hopes to bridge this gap by presenting the largest catalyst solid-liquid interface dataset.

2.2 Figures and tables

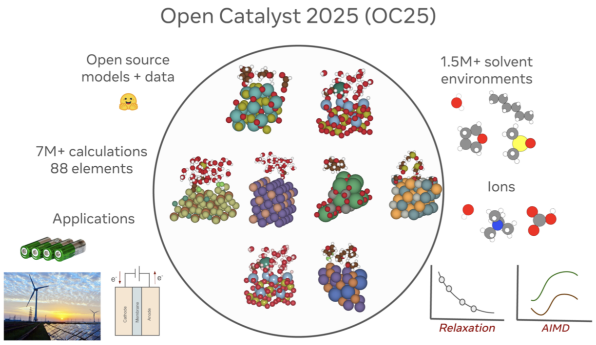

Fig. 1: Overview of OC25, including dataset statistics, sampling strategies, relevant applications, and sample snapshots of the dataset.

Table 1: Size of the train, validation, and test splits.

	Split	Size	Description
Train	All	7,395,512	Training set
Val	Val	203,630	OOD combos
5*Test	Test	202,119	OOD combos
	Solvent	11,111	OOD solvents
	Ion	7,176	OOD ions
	Both	6,989	OOD solvents+ions
	Solvation	5,713	$\Delta\tilde{E}_{solv}$

Acknowledgments

N. G. acknowledges support from a startup grant at NTU (award number 024462-00001). M.M. and J.A.G. gratefully acknowledge support from The Welch Foundation under Grant Number D-2188-20240404. J.B.V. acknowledges support by the U.S. Department of Energy, Lawrence Livermore National Laboratory (LLNL) under Contract No. DE-AC52-07NA27344.

References

[1] Sushree Jagriti Sahoo, Mikael Maraschin, Daniel S. Levine, Zachary Ulissi, C. Lawrence Zitnick, Joel B Varley, Joseph A. Gauthier, Nitish Govindarajan, and Muhammed Shuaibi. The open catalyst 2025 (oc25) dataset and models for solid-liquid interfaces, 2025.