Title: **ML-enabled assimilation of community geochemical datasets into reactive transport models**

Haruko Wainwright,[1,2] Elliot Chang[3]*, Sergi Molins[1], Wenming Dong[1], Linda Beverly[4], Mavrik Zavarin[3]

[1]Lawrence Berkeley National Laboratory, Berkeley, CA;
[2]Massachusetts Institute of Technology, Cambridge, MA;
[3]Lawrence Livermore National Laboratory, Livermore, CA
[4]California State University, East Bay, Hayward, CA

Contact: (hmwainwright@lbl.gov; hmwainw@mit.edu)

Project Lead Principal Investigator (PI): Haruko Wainwright

BER Program: ESS

Project: DOE Lab-led project

Project Abstract:

Laboratory experiments are critical to interrogate the impact of hydrological and climate perturbations on biogeochemical (BGC) processes in a controlled environment. Hydrologically driven biogeochemical reactions are a key aspect of the Earth system predictability, particularly at dynamic interfaces (e.g., terrestrial-aquatic interfaces, hot spots), governing the cycling of nutrients, metals, and organic matter. Recently, there has been a significant effort to collect and compile relevant experimental data across the community, and to develop machine-readable experimental databases for various elements and species in different conditions. Developing such large database has created a unique opportunity for machine learning (ML) to gain new scientific insights as well as to improve the parameterization and uncertainty quantification within BGC and reactive transport models (RTMs).

In this project, we develop an ML-enabled paradigm shift to integrate laboratory experiments and their data into a framework for subsequent incorporation into Earth Systems models with the FAIR (findability, accessibility, interoperability, and reusability) principle. In particular, we aim to a) establish the experiment-to-simulation pipeline through an open-source python/R suite of codes, (b) develop various unsupervised and supervised learning capabilities coupled with the database, and (c) characterize the parameter and model uncertainties across the global datasets in the Bayesian method and transfer the uncertainty to simulation results in a seamless manner.

During the course of our project, the team has developed a new workflow to govern experiment-to-simulation pipelines initially focused on surface complexation reactions, and has also released the accompanying python/R scripts to the public under LLNL distribution.[1] Additionally, we are developing a high-performing, hybrid ML model informed by the chemical thermodynamic

[1] https://ipo.llnl.gov/technologies/software/llnl-surface-complexation-database-converter-scdc

principles.[2] This new approach exploits the previous FAIR-based database development work and paves the way for a more nuanced perspective between traditional sorption modeling routines and pure ML methods. In parallel, a new Python-based workflow has been developed to quantify parameter and model uncertainties associated with PHREEQC-based geochemical simulations using Bayesian methods. Lastly, we are using this pipeline in the development of a watershed reactive transport model for simulating weathering and ion exchange processes. The work is intended to be general and extensible to other BGC data, and to provide a framework for their implementation in discovery science and Earth Systems modeling. In addition, we envision that this framework will stimulate the developments of other community-wide experimental database in BGC and beyond.

References

1. Zavarin, M., E. Chang, H. Wainwright, N. Parham, R. Kaukuntla, J. Zouabe, A. Deinhart, V. Genetti, S. Shipman, F. Bok and V. Brendler (2022). "Community Data Mining Approach for Surface Complexation Database Development." Environmental Science & Technology **56**(4): 2827-2838.

2. Chang, E., M. Zavarin, L. Beverly, H. Zeng, H. Wanwright (2022). "A Chemistry-Informed Hybrid Machine Learning Approach to Predict Metal Adsorption onto Mineral Surfaces." Journal of Computers and Geosciences [To be submitted]