Research Software Engineering to Predict Chemical Properties and Kinetics at High Performance Computing Scale

by

Jackson W. Burns

Submitted to the Department of Chemical Engineering and the Center for Computational Science and Engineering on September 9th in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Chemical Engineering and Computation

Abstract

This thesis presents Research Software Engineering best practices applied to two broad categories: machine learning and physics-based modeling. Many of today's researchers in these and other domains communicate their results through scientific software. For the present work, this spans rule-based, deterministic first-principles systems used for constructing reaction mechanisms to data-driven machine learning models for estimating chemical properties. The former highlights the need for domain-specific skills, in this case with quantum mechanics tools and chemical kinetics theory, whereas the latter emphasizes software package development, and the training and distribution of models.

Research Software Engineering (RSE) is the art and science of developing and distributing such programs in a way that is reproducible, maintainable, and user-friendly. Both domains demonstrate core RSE principles: reproducibility, continuous integration, and validation. Yet despite the central role of RSE, the resulting code is typically not treated as a research product in and of itself.

This thesis will demonstrate that including RSE objectives as primary research goals delivers scientific results that are more correct, trustworthy, and user-friendly. Applying RSE skills even simplifies the process of expanding projects from 'desktop scale' to high performance computing scale, as demonstrated in a subset of the presented applications. Finally, this thesis discusses future directions for scaling and improving these approaches. Recent advancements in scriptable quantum mechanics simulations can take advantage of modern accelerator hardware, offering a future of hyper-scalable dataset generation frameworks. Advances in foundation models for tabular data suggest a new method for pre-training that not only improves performance on small real-world datasets, but also eliminates the need for expensive fine-tuning operations. Both of these promising new directions will be best actualized when also centering RSE best practices.

Thesis Supervisor: William H. Green

Title: Hoyt C. Hottel Professor of Chemical Engineering