Predictive Methods and Datasets for Thermodynamic and Kinetic Modeling of Ionic Solutes

by

Jonathan Zheng

Submitted to the
Department of Chemical Engineering
on December 12, 2025 in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY IN CHEMICAL ENGINEERING

ABSTRACT

The ability to quantitatively describe ionization phenomena is essential to designing medicines, developing novel materials, and modeling the time-evolution of many relevant liquid-phase chemical systems. In liquid phase chemistry, acid-base phenomena result in the formation of solvated ions. Biochemical reactions, synthesis steps, and pharmacological mechanisms-of-action also often involve charged reactants and products.

A key property is the "elusive" solvation free energy of the ion, which itself is not directly measurable but for many relevant ions is tied to other observable properties such as the acid dissociation constant (pK_a). However, modeling approaches tend to perform poorly for predicting solvation free energies of ionic solutes. Existing solvation models often have average errors exceeding 3 kcal mol^{-1}. Literature data are also scarce, precluding the parameterization of physics-based solvation models, development of data-driven methods, and benchmarking of such techniques for ionic solutes.

The work in this thesis seeks to address these issues of data scarcity and modeling. The work can be divided into three components:

- **Acid-base phenomena.** First, the data curation aspects of pK_a are discussed. Existing inconsistencies in data usage and terminology have confused the literature, and are clarified herein. Curated datasets for aqueous and non-aqueous pK_a values, based on IUPAC collections of data, are also presented here. However, even with these datasets, many solvent systems still do not have much data (several, for instance, have less than 100 datapoints). To address this data scarcity issue, a method for generating high-quality pK_a predictions in non-aqueous solvents is presented and benchmarked, and then utilized to create "synthetic" data for roughly 3,000 acids in 29 solvents (a total of nearly 80,000 data points).

- **Thermodynamics of ionic solutes.** As aforementioned, pK_a, along with other measurable thermodynamic properties, can be linked back to the ions' solvation energies. This section describes how this thermodynamic relationship is used to generate a new database of hydration free energies for ions, which can then be used to develop simple corrections to existing solvation models to reduce error by roughly 60%. This method is expanded to a large scale, used to generate nearly 6,000 values across 8 solvents.

For comparison, prior to these efforts, the largest such database included only 300 datapoints. Finally, a machine learning model was trained to predict solvation free energies (as well as the substituent properties), the first machine learning model to our knowledge that can predict anionic solvation energies and gas-phase acidities.

- **Kinetics.** The methods of the previous chapters are expanded to reactions involving zwitterions, radicals, and singly-charged anions (S_N2 reactions). The H-abstraction reaction is examined through a dataset of approximately 100 million solvation free energies previously computed in our group. The dataset consists mostly of uncharged closed-shell and open-shell solutes, but contains some zwitterions as well. The error of the calculation method is examined, with a focus on the barrier heights of the reactions. The conformational effects of the zwitterions are examined, showing the surprisingly high sensitivity of the solvation energies to the optimized geometries of such solutes. Next, S_N2 reactions are examined. A dataset of S_N2 rate data was digitized and provided, and used to benchmark a quantum-chemical approach to prediction relative rate coefficients. The method shows good predictive quality, demonstrating the usefulness of solvation models for *relative* properties despite their very high errors for *absolute* energies.

In sum, these efforts combine cheminformatics, quantum chemistry, data science, and machine learning to enable quicker and more accurate prediction of properties related to acid-base and ionization phenomena.

Thesis supervisor: William H. Green, Ph.D.
Title: Hoyt C. Hottel Professor of Chemical Engineering