A Software Ecosystem for the Data-Driven Design of Chemical Systems and the Exploration of Chemical Space

Mojtaba Haghighatlari[1], William Evangelista[1], Mohammad Atif Faiz Afzal[1], Ching-Yen Shih[1], Bryan A. Moore[1], Mikhail Pechagin[1], Yujie Tian[1], Johannes Hachmann[1,2,3]

(1) Department of Chemical and Biological Engineering,
University at Buffalo, The State University of New York, Buffalo, New York, United States.
(2) Computational and Data-Enabled Science and Engineering Graduate Program,
University at Buffalo, The State University of New York, Buffalo, New York, United States.
(3) New York State Center of Excellence in Materials Informatics, Buffalo, New York, United States.

Trial-and-error research approaches are increasingly ill equipped to meeting the complex challenges involved in the discovery and design of next-generation chemistry and materials. Our work recognizes the great opportunities that are arising with the shift towards data-driven *in silico* research and a rational design paradigm. These approaches are poised to mitigate the inefficiencies, shortcomings, and limitations of traditional trial-and-error research. However, the notion to utilize modern data science in the chemistry context is so recent that much of the basic infrastructure has not yet been developed, or is still in its infancy. The existing tools and expertise tend to be in-house, specialized, or otherwise unavailable to the community at large. Data science is thus in practice beyond the scope and reach of most researchers in the field. Our work aims to chart new paths in this area by creating an open, general-purpose software ecosystem designed to overcoming this situation, filling the prevalent infrastructure gap, and thus making data-driven research a viable and widely accessible proposition. Our software ecosystem fuses *in silico* modeling (in particular computational quantum chemistry), high-throughput screening techniques, and Big Data analytics into an integrated research infrastructure. We have been developing the necessary methods, algorithms, protocols, and codes, and assembled them in three loosely connected program suites: ***ChemHTPS*** provides an automated platform for the virtual high-throughput screening of compound and material candidate libraries as well as reaction networks; ***ChemBDDB*** offers a database and data model template for the massive information volumes created by data-intensive projects; and ***ChemML*** is a machine learning and informatics toolbox for the validation, analysis, mining, and modeling of such data sets.

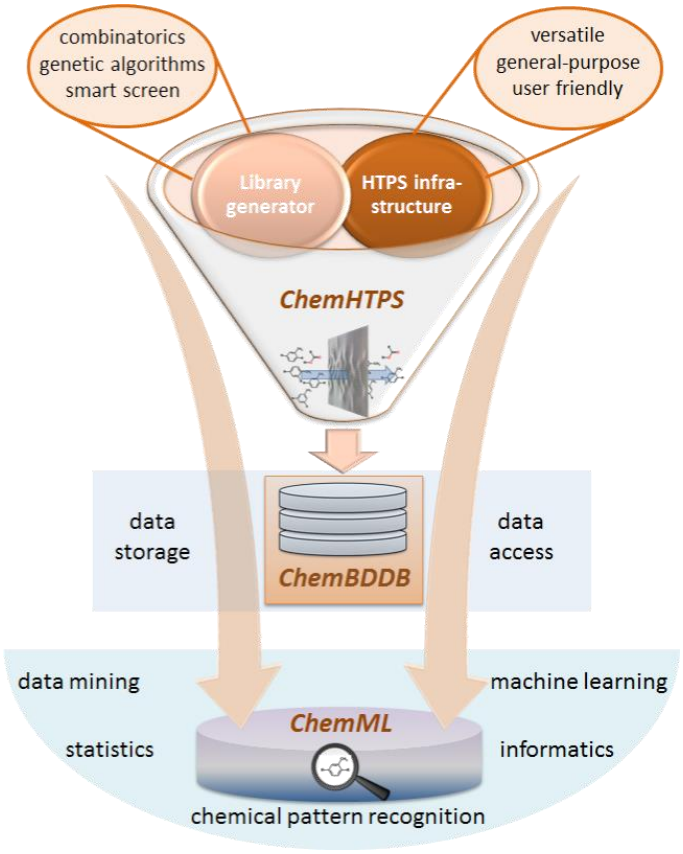

Fig. 1: Schematic of our software ecosystem.