Software development and its application for predicting optical properties in molecular space: ChemML program suite

Mojtaba Haghighatlari,1,* Johannes Hachmann,1, 2, 3

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

*Email: mojtabah@buffalo.edu

Abstract: The idea of utilizing modern data science in chemical and materials research has recently gained considerable attention. However, tools and techniques that could facilitate this work have oftentimes not yet been developed or are still in their infancy. Existing expertise tends 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 address this situation by creating a program suite and software toolbox that is designed to overcoming this situation, filling the prevalent infrastructure gap, and thus making the application of big data analytics in the chemical and materials context – e.g., via machine learning and informatics – a viable and widely accessible proposition. Over the past two years we have been developing ChemML1, our ML and informatics program package for the validation, analysis, mining, and modeling of large-scale chemical data sets. The overall design of ChemML consists of two main frameworks called ChemML Wrapper and ChemML Library. In total ChemML Wrapper offers more than 100 functions and classes from its own library or from external packages. A key consideration of our work is to make ChemML as comprehensive and user-friendly as possible, so that it can be readily employed by machine learning researchers without the need for excessive expert knowledge. This presentation will detail the code design and modular structure of first release of ChemML, including its capabilities, and methodological advances for accurately predicting properties of organic molecules in different applications. In this perspective, we explore deep learning architectures for predicting optical properties of organic molecules and polymers. These architectures merge different feature representation methods and design methodologies to achieve efficient and generalizable predictive models. All these models can further be used in virtual high-throughput screening for accelerating materials discovery and as a guide for rational design.

References: