MULTI-FIDELITY DEEP LEARNING AND ACTIVE LEARNING FOR MOLECULAR OPTICAL PROPERTIES

KEVIN P. GREENMAN, WILLIAM H. GREEN, *Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA*; RAFAEL GÓMEZ-BOMBARELLI, *Department of Materials Science and Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA*.

A variety of physics-based and statistical methods have been developed to guide molecular design based on optical properties. Each method has a trade-off between cost, accuracy, and generalizability. While methods such as time-dependent density functional theory (TD-DFT) are often generalizable across chemical space due to their foundations in physics, they are relatively slow and are less suitable for screening large libraries of molecules. Statistical or machine learning methods are fast, but their performance is highly dependent on the choice of training data and representation. This makes them useful for design within chemical families that already have large datasets available, but less useful for de novo design tasks that explore new parts of chemical space. We propose a new deep learning method that leverages a combination of low fidelity (TD-DFT) and high fidelity (experimental) data sets to predict molecular optical properties with improved accuracy and generalizability over existing statistical methods. We also illustrate the importance of non-random data splitting strategies to assess generalizability of predictions for spectra in condensed phase. Finally, we demonstrate the use of active learning for model improvement by gathering new experimental data in regions of high prediction uncertainty.