A Machine Learning Shortcut to Physics-Based Modeling and Simulations

Johannes Hachmann[a,b,c]

[a] *Department of Chemical and Biological Engineering,*
University at Buffalo, The State University of New York, Buffalo, New York, USA
[b] *Computational and Data-Enabled Science and Engineering Graduate Program,*
University at Buffalo, The State University of New York, Buffalo, New York, USA
[c] *New York State Center of Excellence in Materials Informatics, Buffalo, New York, USA*
hachmann@buffalo.edu

The process of creating new chemistry and materials is increasingly driven by computational modeling and simulation, which allow us to characterize compounds of interest before pursuing them in the laboratory. However, traditional physics-based approaches (such as first-principles quantum chemistry) tend to be computationally demanding, in which case they may not be a practically viable option for large-scale screening studies that could efficiently explore the vastness of chemical space.

In this presentation, we will show how we employ machine learning to develop data-derived prediction models that are alternatives to physics-based models, and how we utilize them in massive-scale hyperscreening studies at a fraction of the cost. Aside from conducting such data-driven discovery, we also employ data mining techniques to develop an understanding of the hidden structure-property relationships that define the behavior of molecules, materials, and reactions. These insights form our foundation for the rational design and inverse engineering of novel compounds with tailored properties.

We will highlight our work on physics-infused machine learning models that seek to improve the robustness and range of applicability of purely data-derived models; on adapting cutting-edge data science techniques for chemical applications (e.g., transfer learning, active learning, and advanced network architectures for deep learning); and on meta-machine learning, i.e., to (machine) learn how to apply machine learning in the chemical domain. We will also show how we use data science techniques to advance, augment, and correct traditional molecular modeling and simulation methods.

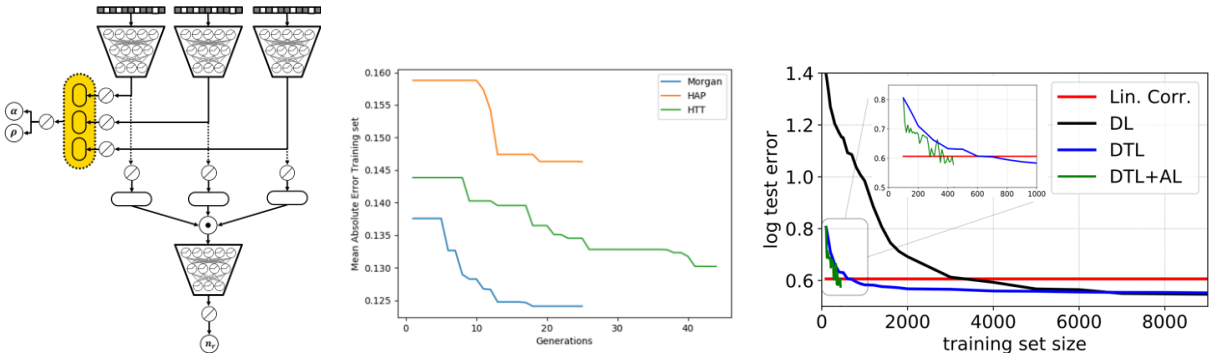

Figure. 2. (a) Network architecture that incorporates the non-linearity of the Lorentz-Lorenz equation; (b) ML model optimization within fixed features spaces for RI predictions via evolutionary learning; (c) Learning curves of traditional deep learning (DL) of static DFT polarizabilities vs DL plus transfer learning (DTL) vs DTL plus active learning (DTL+AL). The results demonstrate a dramatic reduction in the size of the required training set size.

References

1. J. Hachmann, M.A.F. Afzal, M. Haghighatlari, Y. Pal, *Mol. Simul.* **44** (2018), 921-929.
2. M. Haghighatlari, J. Hachmann, *Curr. Opin. Chem. Eng.* **23** (2019), 51-57.