MACHINE LEARNING FOR MOLECULAR PROPERTY PREDICTIONS AND A SOFTWARE ECOSYSTEM THAT ENABLES IT

Johannes Hachmann*
University at Buffalo, The State University of New York
Buffalo, NY 14260

Abstract Overview

This presentation considers the role of machine learning and modern data science in the chemical and materials domain. We will discuss the development of data-derived prediction models and the elucidation of structure-property relationships that can be used for accelerated discovery, rational design, inverse engineering, and the exploration of chemical space. We will showcase our contributions to this field and highlight our software tools and methodological advances on proof-of-concept case studies.

Keywords

machine learning, data-driven *in silico* research, data mining, structure-property relationships

Introduction

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 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 determine the behavior of molecules, materials, and reactions (see Fig. 1). These insights form our foundation for the rational design and inverse engineering of novel compounds with tailored properties.

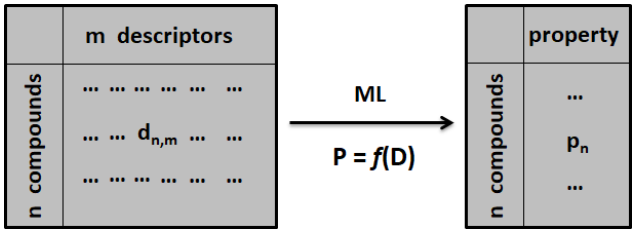


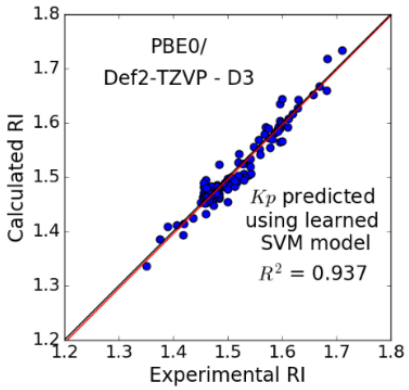
Figure 1. Mathematical setup of the structure-property relationship problem. We employ machine learning to recover the unknown mapping function f to compute the target property p of a given structure, represented in terms of descriptors D.

We will provide specific discovery and design examples of high-refractive index polymers for lens components (see Fig. 2), deep eutectic solvents for supercapacitors, and organic semiconductors. In this context, 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

* To whom all correspondence should be addressed

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.

(a)



(b)

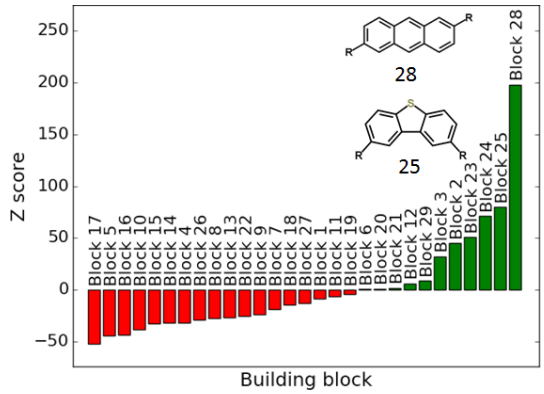


Figure 2. (a) Comparison between experimental refractive index values and those of a hybrid prediction model (partly data-derived, partly physics-based). (b) Z-scores from a hypergeometric distribution analysis identifying prevalent moiety patterns in the most promising candidates from a high-refractive-index polymer screening study. The molecular building blocks with large positive values are significantly overexpressed in the top candidates and thus correlate with desirable performance.

Finally, we will discuss our software ecosystem for data-driven *in silico* research (see Fig. 3) that enables all this work, both on the application as well as on the method development side. It consists of four loosely connected program suites: *ChemLG* is a generator for compound and material candidate libraries that allows us to enumerate chemical space (i.e., performing data definition); *ChemHTPS* provides an automated platform for the virtual high-throughput screening of these libraries (i.e., performing data generation); *ChemBDDB* offers a database and data model template for the massive information volumes created by data-intensive projects (i.e., performing data storage); and *ChemML* is a machine learning and informatics toolbox for the validation, analysis, mining, and modeling of such data sets (i.e., performing data mining).

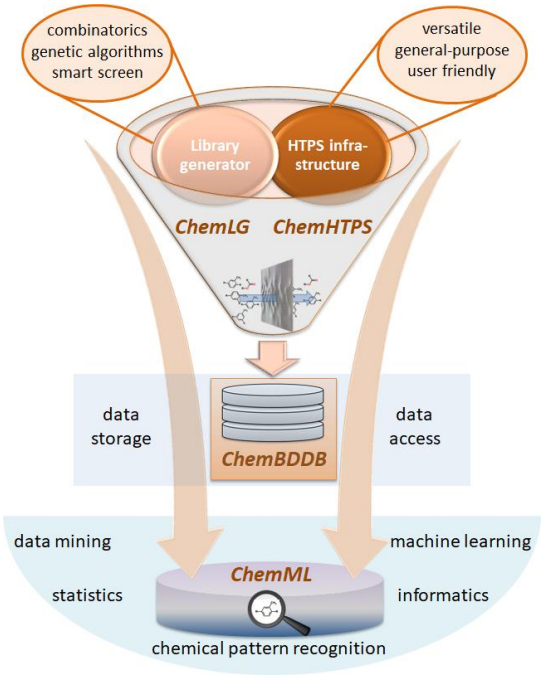


Figure 3. Schematic of the ChemEco software ecosystem for data-driven in silico research comprised of ChemLG, ChemHTPS, ChemBDDB, and ChemML codes.

The notion to utilize modern data science in chemistry 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. By contributing this open, general-purpose, comprehensive, easy-to-use software ecosystem, we aim to chart new paths in this area and help in overcoming this situation, filling the prevalent infrastructure gap, and thus making data-driven research a viable and widely accessible proposition for the community. This includes a development platform and testbed for methods (Fig. 4).

Acknowledgments

This work was supported by start-up funds provided through the University at Buffalo (UB), the National Science Foundation (NSF) CAREER program (grant No. OAC-1751161), and the New York State Center of Excellence in Materials Informatics (grant No. CMI-1148092-8-75163). Computing time on the high performance computing clusters 'Rush', 'Alpha', 'Beta', and 'Gamma' was provided by the UB Center for Computational Research (CCR).

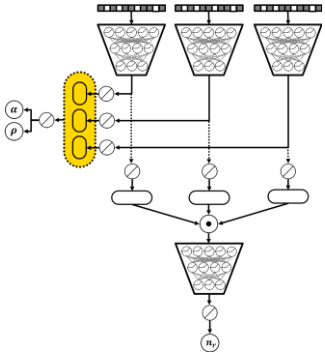


Figure 4. Network architecture that incorporates the non-linearity of the Lorentz-Lorenz equation.

References

Afzal, M.A.F., Haghighatlari, M., Ganesh, S.P., Cheng, C., Hachmann, J. (2019). Accelerated Discovery of High-Refractive-Index Polyimides via First-Principles Molecular Modeling, Virtual High-Throughput Screening, and Data Mining. *J. Phys. Chem. C*, in print.

Afzal, M.A.F., Cheng, C., Hachmann, J. (2018). Combining First-Principles and Data Modeling for the Accurate Prediction of the Refractive Index of Organic Polymers, *J. Chem. Phys.* 148, 241712.

Hachmann, J., Afzal, M.A.F., Haghighatlari, M., Pal, Y. (2018). Building and Deploying a Cyberinfrastructure for the Data-Driven Design of Chemical Systems and the Exploration of Chemical Space. *Mol. Simul.* 44, 921-929.

Haghighatlari, M., Hachmann, J. (2019). Advances of Machine Learning in Molecular Modeling and Simulation. *Curr. Opin. Chem. Eng.* 23, 51-57.