Leveraging Supervised Machine Learning to Predict Band Gaps of Modular Materials from Their Molecular Building-Blocks

Malcolm J. A. Jardine[1], Sergi Vela[2], Maria Fumanal[1]
[1]Universitat de Barcelona, Carrer de Martí i Franquès 1-11 Barcelona, Spain
[2]Institut de Química Avançada de Catalunya (IQAC-CSIC), Carrer de Jordi Girona 18, Barcelona, Spain.

mjajardine@ub.edu

Machine learning (ML) offers a powerful yet cost-effective approach for predicting the electronic properties of (a) large systems that are computationally intractable or (b) a large number of systems that cannot be evaluated individually [1]. Both these challenges apply to covalent organic frameworks (COFs), which are extended two- and three-dimensional networks assembled from molecular building blocks [2]. COFs are modular materials that may show distinct physical and electronic properties depending on their constituent building blocks, which makes them an ideal platform for fragment-based computational design. However, the complexity and size of the chemical space resulting from a modular combinatorial approach is prohibitively large, and, therefore, ML offers a practical solution enabling an efficient exploration, prediction, and screening of their material properties.

Recently, Wang *et al.* reported an ML model able to predict band edge energies of 2D-COFs from their molecular precursors [3]. Their study exploits the CoRE-COFs dataset [4], based on experimentally synthesized 2D-COFs (381 structures). Following the same philosophy, we aim to build an ML model able to predict the electronic properties of modular materials, such as COFs, from their constituent monomers. To encompass the widest possible chemical space, we perform supervised learning on the experimentally reported COFs precursors (190 monomers), and leverage the FORMED subset of the CSD molecular database (116k molecules) [5], to include a more diverse, but synthetically available, set of molecular building blocks. We train supervised ML models on these molecular building blocks to predict the band gap of constructed dimers in a computationally efficient manner. A key challenge lies in properly encoding connectivity information between monomers, while retaining model transferability. We assess the impact of using different structural (SOAP, SLATM) and quantum-based (MODA) [6] descriptors on the ML models' performance. Collectively, these efforts will enable an efficient exploration of COF chemical space and offer insight into scalable, data-driven design strategies for molecular-dimers and COFs.

References

[1] Wei, Chu, Sun, *et al.*, InfoMat, 1-3 (2019) 338–358.
[2] Côté, Benin, Ockwig *et al.*, Science 310 (2005) 1166-1170
[3] Wang, Lv, Wan, Wu, Yang, Journal of Physical Chemistry Letters 14-30 (2023), 6757
[4] Tong, Lan, Yang, Zhong, Chemical Engineering Science 168 (2017) 456-464
[5] Blaskovits, Laplaza, Vela, Corminboeuf, Advanced Materials, 36-2 (2024) 2305602
[6] Santiago, Vela, Deumal, Ribas-Arino, Digital Discovery 3-1 (2024) 99-112

Figures

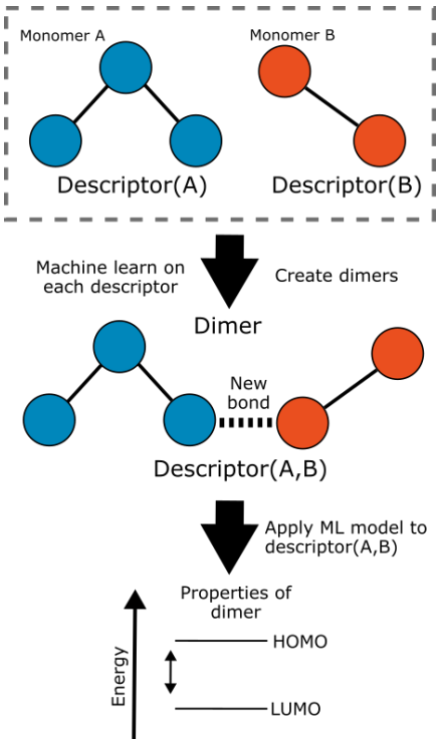

Figure 1. Workflow for implementation of Machine learning (ML) to describe properties of molecular dimers. An ML model is trained on constituent monomers (described with SLATM) to predict properties of constructed molecular dimers.