Towards data engineering and model selection in predictive regression of 2D materials properties

Minh-Tuan Dau[1]

[1]Université Côte d'Azur, CNRS, CRHEA, Valbonne, France

mtd@crhea.cnrs.fr

Abstract

Materials science might benefit a shift from the experimental and theoretical works to the paradigm of data exploration promoted by the emergence of AI, machine learning. Many efforts to characterize novel families of 2D materials have been devoted to accelerating the assessment of their electronic-electrochemical-mechanical properties, the nature of defects, *etc*. Data-driven machine learning appears to be a novel and alternative technique that has been shown to be an efficient estimator of the properties of 2D materials.

In the framework of data-driven regression of 2D materials properties, I will address two building blocks: data engineering and model selection (figure 1) which are core process of a ML cycle. For data handling and engineering, an approach in which the descriptor tailoring is based on vectorizing property matrices has been proposed. The generated descriptors result in outstanding metrics for both training and prediction of the electronic properties of 2D materials with respect to the input data extracted from popular 2D materials databases [1, 2, 3]. The model selection will be discussed in the presentation based on classical ensemble models versus neural network-based models. This gives insight into a comprehensive view of data engineering and model selection in the data-assisted prediction of 2D materials properties and beyond.

References

[1] S. Haastrup, M. Strange, M. Pandey, T. Deilmann, P. S. Schmidt *et al.*, 2D Mater., 5 (2018), 042002

[2] J. Davidson, F. Bertoldo, K. S. Thygesen, R. Armiento, npj 2D Mater. Appl., 7 (2023), 26

[3] M.-T. Dau, M. Khalfioui, A. Michon, A. Reserbat-Plantey, S. Vézian, P. Boucaud, Sci. Rep., 13 (2023), 5426

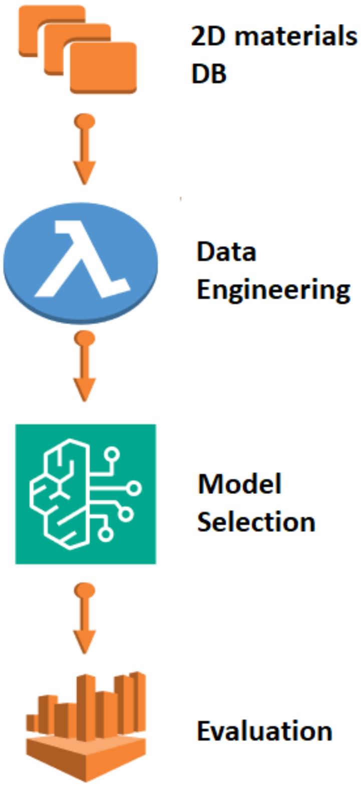

Figure 1. Typical workflow of a ML cycle including data engineering and model selection.