Can accurate machine learning models pinpoint the best materials efficiently?

A.P. Sarikas*, G.S. Fanourgakis, G.E. Froudakis
Department of Chemistry, Voutes Campus, University of Crete, GR-71003 Heraklion Crete, Greece

The intrinsic properties of metal organic frameworks (MOFs), such as their large surface areas and porosity, render them as one of the most promising adsorbents for hydrogen storage systems. Additionally, their modular nature has led to an enormous expansion of the available material pool which in turns raises the question if an efficient identification of the best materials for a specific application, is feasible.
In the recent years, Machine Learning (ML) techniques have been established as the main tool for the exploration of large material databases, since they can greatly accelerate this process. In this work, "traditional" ML models and self-consistent ML-based models [1] are compared with respect to their ability to efficiently pinpoint the best materials of a database. As a case study, we have used hydrogen adsorption in MOFs at different thermodynamic conditions. Despite their accuracy, traditional models struggle to identify the most promising materials without compromising computational cost. On the other hand, self-consistent models can even reduce by two orders of magnitude the amount of reference data required for the identification of the best materials compared to traditional approaches. Nevertheless, the are still factors that complicate the efficient pinpointing of novel materials.

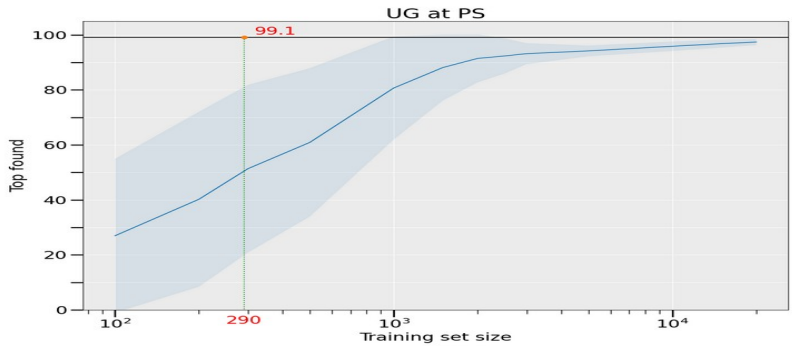

Figure 1: Blue line represents the number of the top 100 materials identified by the traditional model (y-axis) as function of the training set size (x-axis, logarithmic scale), while blue shaded area shows the respective standard deviation. The number of the top 100 materials that the SC model identifies and its training set size are represented by the black and green line, respectively (i.e. 290 materials are needed for training in order to identify 99 of the top 100 performing materials, while with traditional ML models you can find approximately 50 if using the same training set. Moreover, for obtaining 99 with traditional ML models you will need a training set of more than 10.000). UG and PS stand for usable gravimetric (capacity) and pressure-swing (conditions), respectively.

Acknowledgement

[1] This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101058547 (MOST-H2).

References

[1] G. Fanourgakis, K. Gkagkas, E. Tylianakis and G. Froudakis, The Journal of Physical Chemistry C, **124**, 36 (2020).

* antonios.sarikas@gmail.com