Predicting the properties of molecular materials: multiscale simulation workflows meet machine learning

Fabio Le Piane, Matteo Baldoni and Francesco Mercuri

DAIMON Team, CNR-ISMN, Via P. Gobetti 101, 40129 Bologna, Italy.
fabio.lepiane@ismn.cnr.it
www.daimoners.eu

Key Words: *Multiscale Modelling, Workflows, Ontology, Artificial Intelligence, Machine Learning.*

In recent years, machine learning (ML) methods have applied with success to studies of the properties of molecular materials. The vast majority of these studies are focused on the properties of individual molecules, targeting the correlation between molecular structure and resulting properties. The properties of several technological materials constituted by molecular aggregates, however, depend on both molecular structure and on aggregation morphology, as for example the case of nanoscale materials. Computational methods for predicting the properties of molecular materials must therefore integrate the properties of individual molecules with information about aggregation morphology, which, in turn, can be related to materials fabrication and processing. The definition of a modelling paradigm able to simulate and predict the properties of molecular materials as a function of molecular structure and aggregation/fabrication conditions can potentially enable high-throughput development of novel materials for technological applications. In this work, we design and implement a computational workflow for the simulation of the properties of molecular materials integrated with a ML scheme for enhancing the computational workload. The workflow is based on a multi-scale top-down approach, in which target properties are defined from the application to the molecular scale. The workflow is implemented through top-down hierarchical data structures, which connects the properties of molecular materials at the nanoscale to the atomistic/electronic scale. Modelling data are generated by applying domain-specific simulation protocols based on atomistic molecular dynamics and density functional theory calculations. ML approaches are therefore applied to enable the scale reduction, providing a local mapping at a lower scale of the properties of large molecular aggregates, reducing greatly the overall computational load. The proposed approach is applied to the evaluation of intermolecular electronic couplings of aggregates of organic molecular semiconductors, a key quantity for the development of materials for advanced electronics. Preliminary studies suggest the relevance of the specific set of features considered for representing intermolecular properties, which depend on the aggregation morphology. Work is in progress to assess the interplay between the structure of individual molecules and the structure of aggregates in determining the performance of ML predictions of the properties of molecular materials. Moreover, an additional speedup of the whole workflow is obtained by optimizing the implementation of the integration between the multiscale simulation workflow and the ML engine.

In order to facilitate the entire process, we started the development of an ontology materials science, with particular focus on molecular materials. The core of the projects revolves around the idea that a chemico-physical object is represented by its structural features and properties, while computational and experimental workflows are separately represented but results are 100% interoperable. We linked our work to pre-existing ontologies, like MDO and EMMO, trying to keep them as compatible as possible. Succeeding in developing this ontology should

help re-organizing the community working on molecular materials, which is a field historically filled with many different formats, procedures, and conventions. We hope to help researchers to both speed up their personal work while also making communications between different teams easier and quicker. Finally, the development of unifying standards will also serve as an incredibly powerful accelerator for the application of machine learning and all kinds of data intensive applications to materials science.