Efficient Workflow Automation for Materials Modeling: Towards Predictive AI Systems using Synthetic High Throughput Dataset Generation

Mario Vozza[1,2], Tommaso Forni[1,2], Fabio Le Piane[1,3] and Francesco Mercuri[1]

In the field of materials science, there has been a significant increase in the need for streamlined computational techniques aimed at generating predictive datasets tailored for artificial intelligence (AI) applications. Density Functional Tight Binding (DFTB) simulations serve as a powerful tool for elucidating atomic-scale interactions and material properties. However, the manual preparation of DFTB simulations can be time-consuming, hindering the rapid generation of large-scale datasets necessary for training AI models.

This study presents a comprehensive approach to automating the generation and analysis of materials science datasets, specifically focusing on defected graphene structures. This workflow efficiently handles computations for various material properties, including energy and charge transport. By automating these procedures, we can efficiently generate extensive datasets wherein each structure is correlated with its corresponding properties. This tight coupling between structures and properties provides a robust foundation for training predictive models. Additionally, beyond the output properties from DFTB, we have augmented the dataset with synthetic Scanning Tunnelling Microscopy (STM) images generated using the Local Density of States (LDOS). This expansion opens ways for correlating experimental measurements directly with the examined structure in future analyses, enhancing the dataset extension.

With the dataset we created, we employed object detection techniques to identify defects within the graphene flakes. Subsequently, we extracted these defects from the structure's image and utilised classical computer vision techniques to derive features from these defects. The aim was to predict material properties based on the defect geometry using eXtreme Gradient Boosting (XGBoost). Moreover, having access to STM images allows us to correlate images and material properties using convolutional neural networks (CNNs). A pivotal element of this study was the robust integration of data, whereby all outputs from the simulations, including the generated STM images, were stored within a NoSQL database like MongoDB. This integrated approach enhances data management capabilities, allowing for easier scalability and ensuring the consistency and reliability of the dataset. By centralising the storage of simulation outputs and images, researchers could seamlessly access and analyse the data, fostering collaboration and accelerating scientific discoveries in materials science.

[1] CNR - ISMN, Via P. Gobetti 101, 40129 Bologna, Italy; mario.vozza@polito.it
[2] Politecnico di Torino, Corso Castelfidardo 39, 10138 Torino, Italy
[3] University of Bologna, via Zamboni 33, Bologna, 40126, Italy