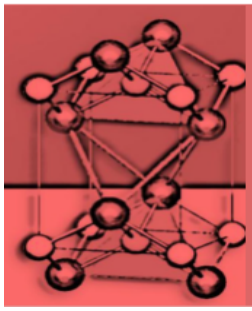

X Workshop on Novel Methods for Electronic Structure Calculations

04th – 06th December 2023
La Plata – Argentina

Molecular identification with high-resolution AFM images, DFT simulations and deep learning

Rubén Pérez [a]

[a] *Departamento de Física Teórica de la Materia Condensada and Condensed Matter Physics Center (IFIMAC), Universidad Autónoma de Madrid, E-28049 Madrid, Spain.*

email: ruben.perez@uam.es

High resolution non-contact atomic force microscopy (HR-AFM) with CO-functionalized metal tips reveals the internal structure of adsorbed organic molecules with unprecedented resolution. resolving intermolecular features, determining bond orders, and characterizing intermediates and final products generated in on-surface reactions [1]. Recent advances in the interpretation using DFT-based methods [2] of the AFM contrast observed in porphycenes [3] and on self-assembled molecular layers driven by either halogen [4] or hydrogen bonds [5], shows that there are clear connections between fundamental chemical properties of the molecules and key features imprinted in HR-AFM images with submolecular resolution.

Inspired by these results, we address the problem of the complete identification (structure and composition) of molecular systems solely based on AFM images exploiting deep learning (DL) techniques. In a first step, we restrict ourselves to a small set of 60 flat molecules and demonstrate the automatic classification of AFM experimental images by a DL model trained essentially with a theoretically generated data set [6]. We analyze the limitations of two standard models for pattern recognition when applied to AFM image classification and develop a model with the optimal depth to provide accurate results and to retain the ability to generalize. We show that a variational autoencoder (VAE) provides a very efficient way to incorporate into the training set, from very few experimental images, characteristic features that assure a high accuracy in the classification of both theoretical and experimental images.

Learning from the successes and the limitations of this proof-of-concept, we have developed QUAM-AFM, the largest data set of simulated AFM images generated from a selection of 685,513 molecules that span the most relevant bonding structures and chemical species in organic chemistry [7]. QUAM-AFM contains, for each molecule, 24 3D image stacks, each consisting of constant-height images simulated for 10 tip–sample distances with a different combination of AFM operational parameters, resulting in a total of 165 million images. The data provided for each molecule includes, besides a set of AFM images, ball-and-stick depictions, IUPAC names, chemical formulas, atomic coordinates, and map of atom heights. In order to simplify the use of the collection as a source of information, we have developed a graphical user interface that allows the search for structures by CID number, IUPAC name, or chemical formula. Using QUAM-AFM, we have designed and trained different deep learning models to go beyond the classification of limited groups of molecules and achieve the complete identification of an arbitrarily complex, unknown molecule, including multimodal recurrent networks (M-RNNs) [8] and Conditional Generative Adversarial Networks (CGANs) [9].

[1]. L. Gross, et al., Angew. Chem.Int. Ed. 57, 3888 (2018).

[2]. M. Ellner, P. Pou, R. Perez, ACS Nano 13, 786 (2019).

[3]. T. K. Shimizu, et al., J. Phys. Chem. C 124, 26759 (2020).

[4]. J. Tschakert, et al., Nat. Commun. 11, 5630 (2020).

[6]. J. Carracedo-Cosme, et al., Nanomaterials 11, 1658 (2021).

[7]. J. Carracedo-Cosme, et al., J. Chem. Inf. Model. 62, 1214 (2022).

[8]. J. Carracedo-Cosme, et al., ACS Appl. Mater. Interfaces 15, 22692 (2023).

[9]. J. Carracedo-Cosme and R. Perez, (2023) submitted (10.48550/arXiv.2205.00447).