Emna Harigua
Institut Pasteur de Tunis

Title: Benchmarking machine learning approaches for computer-aided drug discovery

Abstract: Computer-Aided Drug Discovery (CADD) has gained momentum with recent advances in data analysis and artificial intelligence (AI). In this context, reliable datasets available to the community are of utmost importance to be used in the development of new AI approaches for the discovery of new therapeutic molecules. Considering chemical information on active and inactive molecules against a given pathogen or disease, Ligand-based Drug Discovery (LBDD) approaches rely on a set of techniques ranging from encoding chemical structures into molecular fingerprints to Machine Learning (ML) and Deep Learning (DL) algorithms to deliver novel biologically active molecules. As part of our CADD work against SARS-CoV-2, we performed an extensive literature search to integrate information on 2,610 molecules with a validated effect against SARS-CoV and/or SARS-CoV-2. The dataset presented a heterogeneous content with respect to the validation experiments and/or the target pathogen, which was considered as an additional challenge for the training of ML and DL algorithms. First, we encoded the chemical structures of these molecules by multiple systems to be easily usable as input data for conventional ML algorithms or DL architectures. Then, we evaluated the performance of seven ML algorithms and four DL algorithms to classify molecules into two categories: active and inactive. DL algorithms demonstrated lower ROC-AUC scores compared to baseline ML algorithms (RF and SVM). The DL models with the most promising performances were optimized through hyperparameters' tuning. The ROC-AUC scores obtained by cross-validation reached 85%, demonstrating the ability of these algorithms to correctly predict the activities of anti-coronavirus molecules. Additionally, the DL algorithms demonstrated a higher generalization power as compared to RF when tested through a stratified validation against subsets of the heterogeneous dataset. An external validation step on the collection of FDA-approved drugs revealed a superior potential of DL algorithms to perform the conversion of these drugs against SARS-CoV-2 based on our data, with a TPR equal to 40%. The present results bring novel insights on the power of DL algorithms, mainly Graph Convolution-based approaches, in CADD. The dataset herein collected was made publicly available and is of interest to the large scientific community working in the field of Drug Discovery and Design.