CDD/Post-doctorate at CEA: Machine/deep learning approaches for the elucidation of small molecule structures

Laboratory/Company: CEA

Duration: 12 mois

Contact: etienne.thevenot@cea.fr

Deadline for application: 2023-11-30

Context: Mass Spectrometry (MS) based metabolomics is a powerful technology for the discovery of biomarkers that can stratify patients. Within the MetaboHUB French infrastructure for metabolomics and fluxomics, a significant effort is dedicated to the development of innovative computational methods, software libraries and workflows for the processing, analysis, and interpretation of metabolomics data.

Determining the 2D structure of a metabolite from MS data is a major challenge. Since spectral libraries of reference compounds are scarce, *in silico* strategies have been developed to match experimental spectra directly to molecules, for which the structure is known, but no spectra is available [1]. The current reference method relies on the prediction of a vector of chemical descriptors using a set of Support Vector Machines; this fingerprint can subsequently be matched to those from known compounds in databases [2]. Performances, however, remain currently limited to 30% of correct structures [3]. Recently, alternatives based on artificial neural networks have been suggested to further take into account the interactions between features [4].

Method: The first task will focus on the benchmark of the recent prediction tools against the consortium's data (peakforest.org), as well as against those from the CASMI challenge data [3]. The model will then be enriched with new input features and output molecular properties, and the architecture will be optimized to improve the performances. Finally, the algorithms will be implemented into FAIR software libraries and computational workflows for high-throughput and reproducible structure recommendation.

Main responsibilities:

- Identify the open source prediction tools
- Implement a pipeline for FAIR comparison of their performances
- Build a training database of all publicly available spectra
- Build a comprehensive list of molecular descriptors
- Propose alternative learning architectures to increase the prediction performances
- Implement the selected solution in FAIR software libraries and computational workflows

Keywords: machine learning, deep learning, cheminformatics

References:

[1] Nguyen et al. (2019) Recent advances and prospects of computational methods for metabolite identification: a review with emphasis on machine learning approaches. *Briefings in Bioinformatics*, 20, 2028–2043.

[2] Dührkop et al. (2015) Searching molecular structure databases with tandem mass spectra using CSI:FingerID. *PNAS*, 112, 12580–12585.

[3] Schymanski et al. (2017) Critical Assessment of Small Molecule Identification 2016: automated methods. *Journal of Cheminformatics*, 9, 22.

[4] Fan et al. (2020) MetFID: artificial neural network-based compound fingerprint prediction for metabolite annotation. *Metabolomics*, 16, 104.

Profile:

Bachelor's degree (Bac +5) or PhD in machine learning, deep learning, cheminformatics or computational mass spectrometry

MetaboHUB and CEA are committed to promoting gender equality, and female candidates are encouraged to apply.

Skills:

- Proficiency in Python and PyTorch
- Familiarity with RDKit
- Familiarity with QSAR approaches (an advantage)
- Familiarity with Singularity containers (an advantage)
- Ability to work independently and collaborate effectively within a multidisciplinary consortium.
- Good communication and documentation skills.

Location: You will join the metabolomics data science team (Odiscé; odisce.github.io) at CEA Saclay and interact with the colleagues from the MetaboHUB consortium.