PhD Position F/M Statistical Learning on Flow Cytometry Data for the early characterization of Acute Myeloid Leukemia

- Contact: Christèle ETCHEGARAY (christele.etchegaray@inria.fr), Baudouin DENIS DE SENNEVILLE (baudouin.denisdesenneville@math.u-bordeaux.fr),

- Title: **Statistical learning on Flow Cytometry data for the early characterization of Acute Myeloid Leukemia**

- Keywords:

Acute Myeloid Leukemia, Flow Cytometry, Machine Learning, Deep Learning, Artificial Intelligence

- Scientific Research context:

Acute Myeloid Leukemia (AML) is an aggressive form of bone marrow cancer characterized by the proliferation of immature blood cells. The typical treatment is intensive chemotherapy that starts as early as possible. For some patients, this treatment turns out to be ineffective. Alternative treatment and/or inclusion in a clinical trial could be proposed if only these patients could be identified from the diagnosis. A recent study (Itzykson et al., 2021) proposed a therapeutic decision tool based on cytogenetic and molecular biomarkers (chromosomal abnormalities, mutations). It is able to classify patients in three groups based on the adequacy of intensive chemotherapy (favorable, adverse or intermediate). Unfortunately, these biomarkers are obtained too late to inform the initial therapeutic decision.
In this PhD thesis, the goal is to develop **statistical learning approaches** for **flow cytometry data** obtained at diagnosis, in order to **predict the cytogenetic and molecular prognosis markers** for each patient.

- Work description:

The first goal is to go beyond the manual treatment of flow cytometry data performed by the clinicians by establishing a **data preprocessing algorithm**. Flow cytometry data appear as large dimensional tables where, for each patient, tens of thousands of cells are individually characterized by two markers of size and granularity, and 10 markers for expression in surface proteins. A first task will focus on cell outliers filtering using a strategy based on unsupervised clustering techniques such as Self-Organizing Maps (Van Gassen et al., 2015). This work will lead to the development of a R library.

The second goal is to develop **deep-learning models for the prediction of the presence of mutations**. Convolutive Neural Networks will be adapted to the specificities of flow cytometry data (e.g robustness with respect to markers permutation), extending the previous work from (Hu et al., 2020). The effect of some settings in the data preprocessing or cell subsampling will be investigated. Interpretability of the predictions will be assessed by permutation methods. Possible further development will aim at predicting a mutation rate (regression) rather than a binary mutation status (classification).

The third goal is to supplement the previous approach to build a **model for the prediction of the chemotherapy-adequacy group** from flow cytometry data. This stratification arise from the combination of a 3-class cytogenetic risk group with some mutation landscapes. First, there will be exploration of strategies for combining mutation models. A second task will focus on the prediction of the cytogenetics risk group (classification). A third task will

consist in building a decision tree approach to combine these models. The resulting model will then be validated on an independent dataset.

- Required Knowledge and background:

- Background in applied mathematics, ideally with specialized courses on programming, machine or/and deep learning.
- Appetite for interdisciplinary work with clinicians.

- References:

- Didi, I. et al. Artificial intelligence-based prediction models for acute myeloid leukemia using real-life data: A DATAML registry study. Leuk Res. 2024 Jan;136:107437
- Hu et al., (2020). A robust and interpretable end-to-end deep learning model for cytometry data. Proceedings of the National Academy of Sciences, 117(35), 21373-21380.
- Itzykson et al., 2021. Genetic identification of patients with AML older than 60 years achieving longterm survival with intensive chemotherapy. Blood 138, 507–519.
- Van Gassen, S., Callebaut, B., Van Helden, M. J., Lambrecht, B. N., Demeester, P., Dhaene, T., & Saeys, Y. (2015). FlowSOM: Using self-organizing maps for visualization and interpretation of cytometry data. Cytometry Part A, 87(7), 636-645.

- Duration & start date: PhD 36 months from October 1st