Identifying essential mass cytometry markers in classifying cells belonging or not to Covid patients with neural networks

Gabrielle CHIRON* [1] , Evann DREUMONT* [1], Elie-Julien EL HACHEM [2]
[1] Biology, Modeling and Data Analysis Bachelor at Sorbonne University and Roscoff's Marine Biology Station
[2] Sorbonne University, INSERM, Nutrition and Obesities: Systemic Approaches, NutriOmique, 91 Boulevard de l'hôpital, 75013, Paris, France
*These authors have equally contributed
Corresponding authors' e-mail address: gabrielle.chiron@etu.sorbonne-universite.fr, evann.dreumont@etu.sorbonne-universite.fr

Introduction

Following the hit of Covid-19, numerous data on positive and negative SARS-CoV-2 patients were released, among which mass cytometry files obtained from analyzing their circulating immune cells (PBMCs). Such files were produced within the Covidom project, from which our data proceeds [1].

Mass cytometry is a technique allowing the characterization and classification of cell populations through the marking of cell antigens, thanks to specific antibodies labeled with stable metal isotopes. The latter can be quantified by time of flight (TOF) mass spectrometry.

Our goal was to use neural networks and deep learning to identify which mass cytometry markers are essential to the classification of immune cells in function of their origin (SARS-CoV-2+ or SARS-CoV-2- patient). The principle of this commonly called 'deep cell profiling' is to run a simple two-layered neural network and select the neurons (i.e. markers) with the biggest weight. These markers could be investigated for the development of preventive therapies against viral pathologies.

We also sought to build the most accurate neural network possible for the prediction of the cells' origin. This could be used for diagnosing patients.

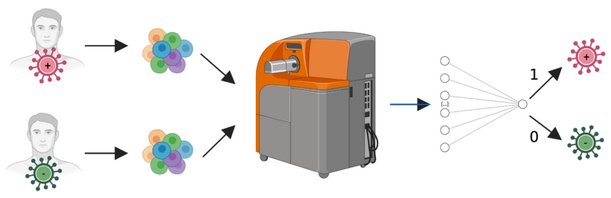


Figure 1: Obtaining mass cytometry data for deep cell profiling.

Methods

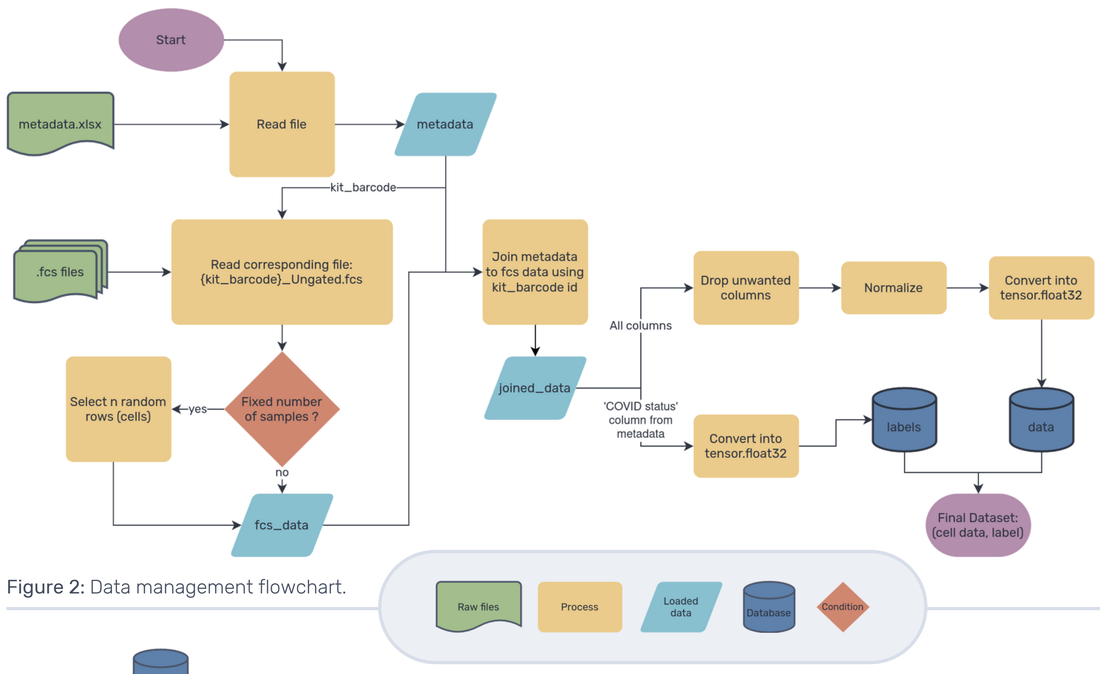


Figure 2: Data management flowchart.

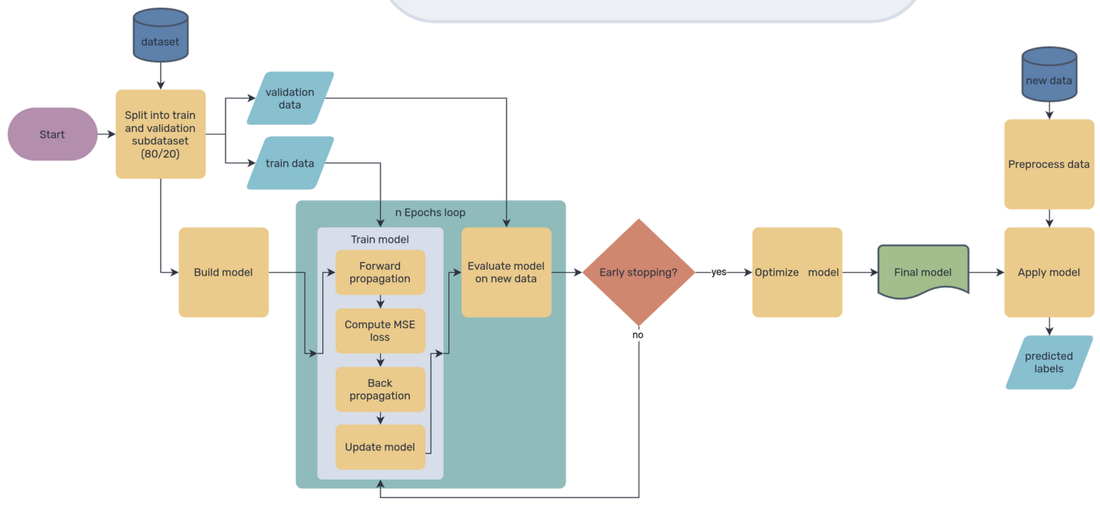


Figure 3: Neural network process flowchart.

A. Data Management

The raw files we used possess data from 99 patients [1]:

- 69 SARS-CoV-2+
- 30 SARS-CoV-2-

For each patient:

- We loaded cytometry data for 10 000 cells.
- We assumed the cell status is the same as the patient's SARS-CoV-2 status.

⇒ Final cohort - 990 000 cells (690 000 SARS-CoV-2+ and 300 000 SARS-CoV-2-).

We transformed this data into our Final dataset containing tuples with cell data (61 cytometry markers) and labels (SARS-CoV-2 status).

B. Neural Network (NN)

Our NN takes in input the cell data (a 61-parameter tensor). Depending on the objective:

- The NN is a perceptron: a linear regression applied on the input.
- The NN is a DNN (Deep NN): a projection of the input on to a 500-neurons hidden layer, followed by a linear regression of this layer.

Hidden layers apply a ReLU activation function. To prevent overfitting, dropouts and L1 regularization could be applied during the training process.

In each case, the output layer (i.e. a single neuron) applies a sigmoid function to provide the probability of SARS-CoV-2+ status for the cell.

Results

Figure 4: Unoptimized perceptron's training results.

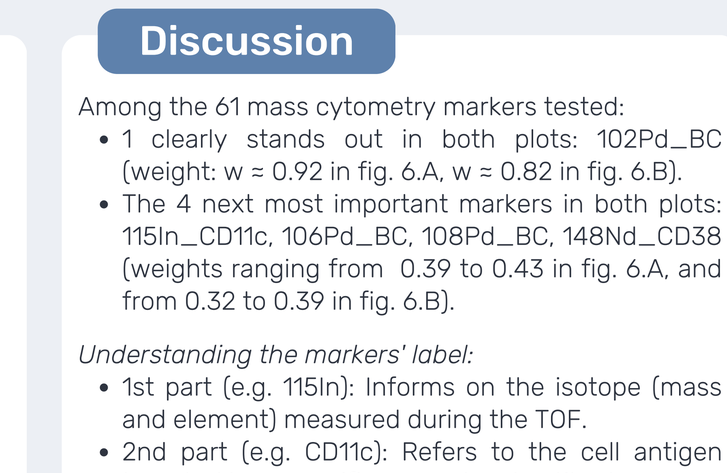
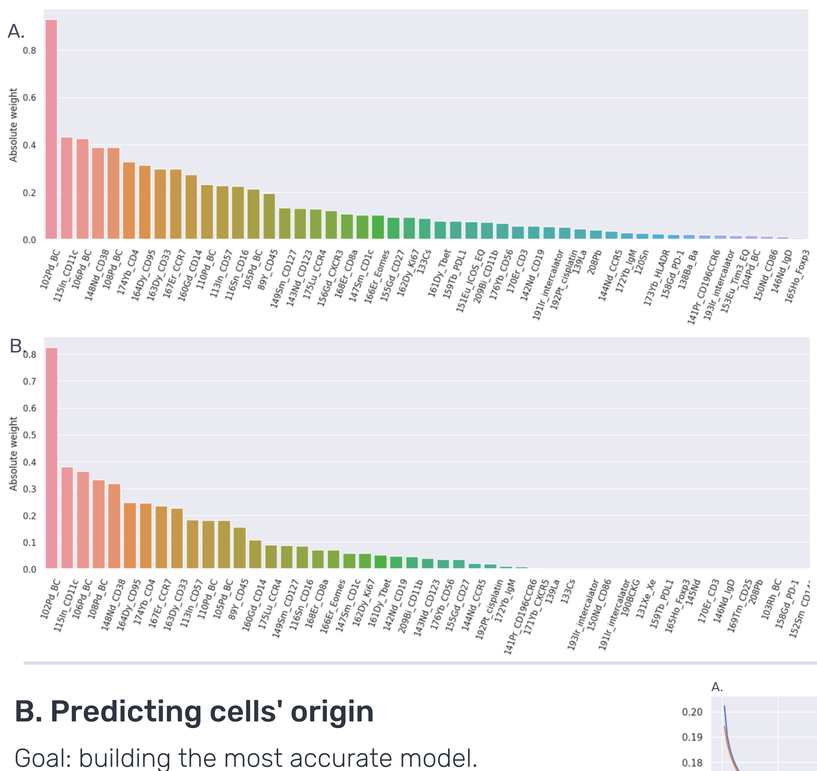


Figure 5: two-layered NN architecture schematics

A. Cell profiling

Goal: identifying essential mass cytometry markers to the classification of cells in function of their origin (SARS-CoV-2+ or SARS-CoV-2- patient).

- Usage of a perceptron (two-layered NN) (fig. 5): allows us to know the importance of cytometry markers in the prediction of cells' origin, as they each correspond to an input neuron's weight in this NN type.
- 30 epochs: the unoptimized model's loss, accuracy and F1 score already stagnate (fig. 4). (Accuracy ≈ 74% at 30 epochs in B.). An L1-regularized model nearly gives identical loss, accuracy and F1 results.
- Plots of cytometry markers' absolute weight, after training the model (fig. 6): The introduction of regularization (B.) reduces the number of important (i.e. heavy-weighted) markers and reduces their weight, hence avoiding overfitting.

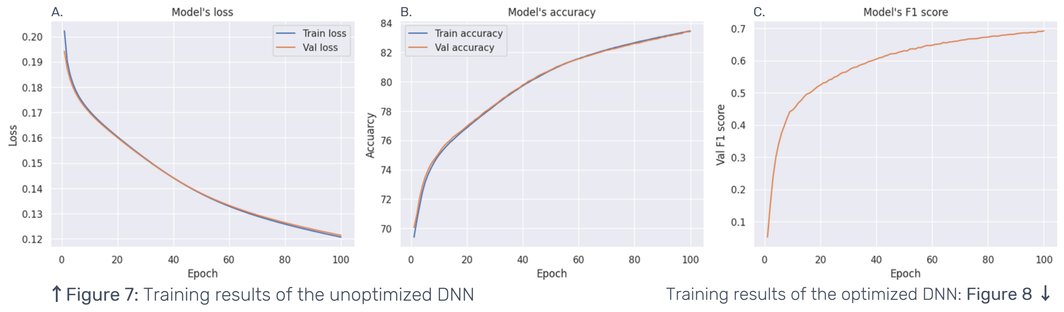


← **Figure 6:** Markers' weight in unoptimized (A.) and L1-regularized (B.) NN.

B. Predicting cells' origin

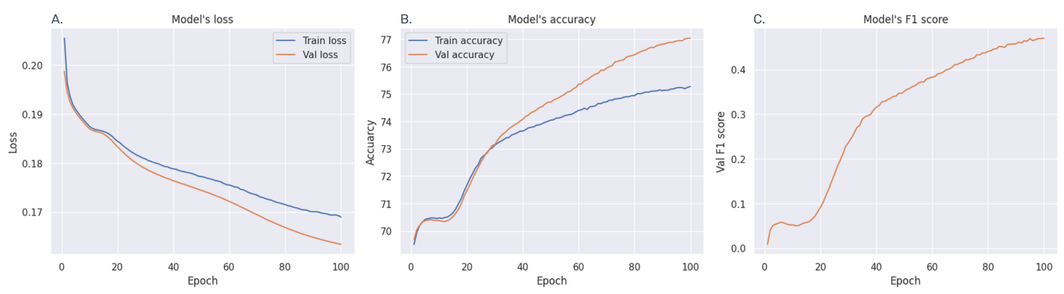
Goal: building the most accurate model.

- Usage of a DNN, with an hidden layer of 500 neurons: allows better accuracy.
- 100 epochs: in this NN type, more epochs are needed before the model's loss, accuracy and F1 score converge - 100 are barely enough, but chosen due to computing resources limitations.
- 1st deep model - without optimization (fig. 7): allows slightly better accuracy (83% in B.), but appears to overfit, meaning it will give poor results on new data.
- 2nd deep model - with optimization (dropouts in the first layer and L1-regularization, fig. 8): allows decent accuracy (77% in B.), without overfitting.



↑ **Figure 7:** Training results of the unoptimized DNN

Training results of the optimized DNN: **Figure 8** ↓



Discussion

Among the 61 mass cytometry markers tested:

- 1 clearly stands out in both plots: 102Pd_BC (weight: w ≈ 0.92 in fig. 6.A, w ≈ 0.82 in fig. 6.B).
- The 4 next most important markers in both plots: 115In_CD11c, 106Pd_BC, 108Pd_BC, 148Nd_CD38 (weights ranging from 0.39 to 0.43 in fig. 6.A, and from 0.32 to 0.39 in fig. 6.B).

Understanding the markers' label:

- 1st part (e.g. 115In): Informs on the isotope (mass and element) measured during the TOF.
- 2nd part (e.g. CD11c): Refers to the cell antigen targeted by a specific monoclonal antibody.

⇒ Essential markers to discriminate cells originating from a SARS-CoV-2 + or - patient:

- BCs (Blood Cells)' antigens
- CD11c antigen
- CD38 antigen

BC:

- Not informative for understanding the cells' classification: does not refer to a specific cellular type (all PBMCs are BCs), nor to specific antigens.

CD11c:

- Mainly known as a dendritic cell marker, but also expressed by some macrophages. [2] Implies that patients infected with SARS-CoV-2 know great changes in their dendritic cell population.
- Coherent with dendritic cells' role: antigen-presenting cells playing a key part in the adaptive answer to diseases, such as SARS-CoV-2.
- Coherent with CD11c's role: A high expression level of CD11c on dendritic cells is correlated with a successful response against SARS-CoV-2. [3]

CD38:

- widely expressed in a variety of human tissues and cells, especially in the immune system. [4]
- Role in the immune response: CD38 is an NADase enhanced by viral-induced metabolic changes and allowing anti-viral and inflammatory responses. [5]
- A high proportion of CD38-expressing cells is correlated with non-survival to SARS-CoV-2, [6] as it may lead to fatal hyperimmune response (notably through a cytokine storm). [7]
- CD38 is highly expressed as a consequence of aging. Contributes to making the elderly and individuals with age-related syndromes more susceptible to SARS-CoV-2. [7]

References

[1] https://flowrepository.org/id/FR-FCM-Z367
[2] Helft, Julie et al. "GM-CSF Mouse Bone Marrow Cultures Comprise a Heterogeneous Population of CD11c(+)MHCII(+) Macrophages and Dendritic Cells." Immunity vol. 42,6 (2015): 1197-211. doi:10.1016/j.immuni.2015.05.018
[3] Hasan, Amal et al. "Fatal COVID-19 is Associated with Reduced HLA-DR, CD123 or CD11c Expression on Circulating Dendritic Cells." Journal of inflammation research vol. 15 5665-5675. 10 Oct. 2022, doi:10.2147/JIR.S360207
[4] Li, Yanli et al. "CD38 as an immunomodulator in cancer." Future oncology (London, England) vol. 16,34 (2020): 2853-2861. doi:10.2217/fon-2020-0401
[5] Horenstein, Alberto L et al. "CD38 in the age of COVID-19: a medical perspective." Physiological reviews vol. 101,4 (2021): 1457-1486. doi:10.1152/physrev.00046.2020
[6] Bobcakova, Anna et al. "Activated CD8+CD38+ Cells Are Associated With Worse Clinical Outcome in Hospitalized COVID-19 Patients." Frontiers in immunology vol. 13 861666. 14 Mar. 2022, doi:10.3389/fimmu.2022.861666
[7] Zeidler, Julianna D et al. "Implications of the NADase CD38 in COVID pathophysiology." Physiological reviews vol. 102,1 (2022): 339-341. doi:10.1152/physrev.00007.2021

Links

Code and poster available at: *github.com/LeGmask/COVID_cyTOF*