Title: Towards a foundation model for flow cytometry

Project Duration: 3 months (full-time) / 6 months (part-time)

Keywords: self-supervised, pretraining, transformers

Project Background

In this project we will be developing a prototype for a foundation model that encodes data obtained from flow-cytometric measurements. Medical applications often suffer from a large heterogeneity in the data between data sources, e.g., due to differing analysis protocols, demographic composition or measurement devices. This makes pooling data across data sources difficult. Integrating such variable sources into a single foundation model (FM) could be a first step towards further improvements in patient outcomes, especially, e.g., for the detection of rare diseases. Small or medium-sized FMs have the potential to integrate such diverse sources into a single model by learning unified representations from a large source of heterogeneous data. This has been done widely across domains, e.g., for vision (see [2]) and has also been tried for flow cytometry [1].

In this project we want to take important first steps towards such a foundation model by training an initial prototype starting from similar attempts in literature [2]. The goal is to evaluate a self-supervision objective that yields robust representations which are helpful for down stream tasks, e.g., single cell classification.

Data Data for flow cytometry is obtained by measuring single cells in a patient sample in a flow cytometer. This results in ~3-5 .fcs files per patient, each measuring ~8-12 different variables. FCS measurements are measured in sets of 1e4 - 1e6 cells, making it suitable for NN architectures that operate on sets (e.g., transformers). In this project, we'll be using a large dataset, collected from several centers, that covers a variety of biological and technical sources of variation. We will evaluate the learned representations on a single-cell classification dataset with available labels.

The core deliverables of this project are:

- Test and adapt transformer architectures for flow cytometry (FCS) data for self-supervised pretraining.
- Design and evaluate different novel (self-supervised) pre-training objectives.
- Evaluate the learned representation on an interesting downstream task: Single-cell classification.

This project is for you if

- You are highly motivated and self-sufficient
- You are a good python programmer
- You have a strong background in ML
- You had some prior exposure to self-supervised learning approaches, e.g. for images, text or other modalities

- You are interested in self-supervised learning and foundation models for domain specific data - here data from flow cytometers.
- You have a basic interest in medicine/biology and ML in the medical domain.

About us Blood is a unique window into your immune system, but analysis of these millions of tests today is done manually and subjectively. We at hema.to build a commercially available AI platform for clinical blood cancer testing. Our goal is to improve testing by providing faster, better and more reliable testing of immune disease with the mission of digital precision medicine for millions of patients, plug&play to the existing global cytometry infrastructure.

Literature

1. https://openreview.net/pdf?id=2mq6uezuGj
2. https://arxiv.org/abs/2304.07193
3. https://arxiv.org/abs/2310.07338
4. https://arxiv.org/abs/2207.14255

Contact

Dr. Florian Pfisterer

florian@hema.to

Hemato - www.hema.to

If you are interested, please send an email to florian@hema.to, preferably with a short CV and some info on you background and motivation.