Exploring foundation model approaches for flow cytometry

Supervisor: Paul Martini

Note. Teams of up to two people are encouraged.

Research question. Recently proposed *deep learning–based* methods for analyzing *flow cytometry* data are predominantly supervised. In practice, labeled data is typically available for only a small number of samples, whereas large amounts of unlabeled data are readily accessible. Can this unlabeled data be leveraged to train a *foundation model* that improves performance on downstream tasks?

Background. Flow cytometry (FC) is a laboratory technique used to detect and measure physical and chemical characteristics of population of cells or particles in a solution. In medicine, particularly in hematology and immunology, FC is mainly applied to characterize and count types of white blood cells in the evaluation of infectious diseases, autoimmune disorders, immunodeficiencies, or in the diagnosis of blood cancers such as leukemias or lymphomas. A flow cytometer has a fixed number of channels m (typically $m \in \{10, 11, \ldots, 30\}$), corresponding to a marker protein panel, across which it captures one measurement for each cell in a sample (sample $\approx$ patient). Typically, there are $10000 \leq n \leq 1000000$ cells per sample leading to a data matrix $X \in \mathbb{R}^{n \times m}$ where $n >> m$. The data is unlabeled and labels (e.g. cell type, disease) are only available through a manual gating process.

Here, "foundation model" refers to pretraining on unlabeled flow cytometry data to obtain reusable embeddings that transfer across tasks or datasets.

Project outline. The project would proceed in three stages:

(0) Extensive literature search. (Investigate whether embedding methods for single-cell RNA sequencing data can be adopted for FC data.)

(1) Benchmark self-supervised representation learning techniques (e.g. AE, VAE, MAE, contrastive learning, other existing methods). Evaluation strategy:

- Annotated dataset available [Bini et al., 2024].
- Clustering quality and neighborhood preservation
- Downstream classification task

(2) Marker panels are not standardized across FC datasets. Thus, for foundation model training an alignment step is necessary to integrate embeddings from different datasets. Explore the feasibility of different alignment approaches (e.g. adversarial domain adaptation, cycle-consistent mappings, own ideas). Evaluate whether integrating multiple datasets improves performance compared to training on each dataset individually.

Requirements.

- Basic knowledge in deep learning
- Python programming, experience with PyTorch
- Independent and rigorous work-style

References

L. Bini, F. Nassajian Mojarrad, M. Liarou, T. Matthes, and S. Marchand-Maillet. Flowcyt: A comparative study of deep learning approaches for multi-class classification in flow cytometry. In *Conference on Health, Inference, and Learning (CHIL)*, 2024.