Towards a foundation model for flow cytometry

Supervisor: Paul Martini

Note. Teams of up to two people are encouraged.

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. [Bini et al., 2024] A flow cytometer has a fixed number of channels m (typically $m \in \{10, 11, \ldots, 30\}$) across which it captures one measurement for each cell in a sample (sample $\approx$ patient). Typically, there 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.

Project outline. Recently published deep learning based methods for the analysis of FC data mostly rely on labeled data. Since such data usually is only available for few samples, the development of a foundation model for FC data which can be trained in a self-supervised fashion is paramount. The project would proceed as follows:

- Extensive literature search. Possible starting points:
  - Hu et al. [2020]
  - Wödlinger et al. [2022]
  - Kowarsch et al. [2022]
  - Weijler et al. [2023]
  - Fisch et al. [2024]
  - Bini et al. [2024]

  Also, investigate whether methods for single-cell RNA sequencing data embedding can be adopted for FC data.
- Try out existing methods on a labeled, public benchmark dataset (e.g. [Bini et al., 2024]) extract the embeddings they produce and assess their goodness
- Assess whether existing methods/architectures can be adopted to be trained in a self-supervised fashion
- From these findings conceptualize and implement a prototype of an own model

Requirements.

- Python programming, experience with PyTorch
- Willingness to dive into the intricacies of DL with FC data
- 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.

L. Fisch, M. Heming, A. Schulte-Mecklenbeck, C. C. Gross, S. Zumdick, C. Barkhau, D. Emden, J. Ernsting, R. Leenings, K. Sarink, N. R. Winter, U. Dannlowski, H. Wiendl, G. Meyer Zu Hörste, and T. Hahn. Gatenet: A novel neural network architecture for automated flow cytometry gating. *Computers in Biology and Medicine*, 179:108820, 2024. doi: 10.1016/j.compbiomed.2024.108820. URL https://www.sciencedirect.com/science/article/pii/S0010482524009053.

Z. Hu, A. Tang, J. Singh, S. Bhattacharya, and A. J. Butte. A robust and interpretable end-to-end deep learning model for cytometry data. *Proceedings of the National Academy of Sciences*, 117(35):21373–21380, 2020. doi: 10.1073/pnas.2003026117. URL https://www.pnas.org/doi/abs/10.1073/pnas.2003026117.

F. Kowarsch, L. Weijler, M. Wödlinger, M. Reiter, M. Maurer-Granofszky, A. Schumich, E. O. Sajaroff, S. Groeneveld-Krentz, J. G. Rossi, L. Karawajew, R. Ratei, and M. N. Dworzak. Towards self-explainable transformers for cell classification in flow cytometry data. In *Interpretability of Machine Intelligence in Medical Image Computing: 5th International Workshop, IMIMIC 2022, Held in Conjunction with MICCAI 2022, Singapore, September 22, 2022, Proceedings*, pages 22–32, Berlin, Heidelberg, 2022. Springer-Verlag. doi: 10.1007/978-3-031-17976-1_3. URL https://doi.org/10.1007/978-3-031-17976-1_3.

L. Weijler, F. Kowarsch, M. Reiter, P. Hermosilla, M. Maurer-Granofszky, and M. Dworzak. Fate: Feature-agnostic transformer-based encoder for learning generalized embedding spaces in flow cytometry data, 2023. URL https://arxiv.org/abs/2311.03314.

M. Wödlinger, M. Reiter, L. Weijler, M. Maurer-Granofszky, A. Schumich, E. O. Sajaroff, S. Groeneveld-Krentz, J. G. Rossi, L. Karawajew, R. Ratei, and M. N. Dworzak. Automated identification of cell populations in flow cytometry data with transformers. *Computers in Biology and Medicine*, 144:105314, 2022. doi: 10.1016/j.compbiomed.2022.105314. URL https://www.sciencedirect.com/science/article/pii/S0010482522001068.