

Structured Representation Learning for Cytometry: Cell Annotation and Population Discovery

Auteur : Bodart, Fanny

Promoteur(s) : Louppe, Gilles

Faculté : Faculté des Sciences appliquées

Diplôme : Master en ingénieur civil biomédical, à finalité spécialisée

Année académique : 2024-2025

URI/URL : <http://hdl.handle.net/2268.2/23237>

Avertissement à l'attention des usagers :

Tous les documents placés en accès ouvert sur le site le site MatheO sont protégés par le droit d'auteur. Conformément aux principes énoncés par la "Budapest Open Access Initiative"(BOAI, 2002), l'utilisateur du site peut lire, télécharger, copier, transmettre, imprimer, chercher ou faire un lien vers le texte intégral de ces documents, les disséquer pour les indexer, s'en servir de données pour un logiciel, ou s'en servir à toute autre fin légale (ou prévue par la réglementation relative au droit d'auteur). Toute utilisation du document à des fins commerciales est strictement interdite.

Par ailleurs, l'utilisateur s'engage à respecter les droits moraux de l'auteur, principalement le droit à l'intégrité de l'oeuvre et le droit de paternité et ce dans toute utilisation que l'utilisateur entreprend. Ainsi, à titre d'exemple, lorsqu'il reproduira un document par extrait ou dans son intégralité, l'utilisateur citera de manière complète les sources telles que mentionnées ci-dessus. Toute utilisation non explicitement autorisée ci-avant (telle que par exemple, la modification du document ou son résumé) nécessite l'autorisation préalable et expresse des auteurs ou de leurs ayants droit.

Abstract

Flow cytometry enables the characterization of cell types based on the expression of specific surface and intracellular markers. It is widely used in both research and clinical settings to analyze cell populations. Recent advances in the field now allow the simultaneous measurement of numerous markers, resulting in high-dimensional datasets. Thus, the conventional manual gating approach is no longer suitable for analyzing such complex data. While several machine learning methods have been proposed for automated cell classification, most focus solely on known populations. Conversely, unsupervised methods can discover novel subpopulations but lack interpretability and do not support direct annotation.

In this work, we propose a model capable of addressing these complementary goals within a unified semi-supervised framework. Our approach leverages structured representation learning through a deep generative model to achieve (1) accurate classification of known immune cell populations, (2) discovery of novel subpopulations, and (3) characterization of immune population dynamics across experimental conditions.

We introduce MARVIN: *Structured Representation Learning for Cytometry: Cell Annotation and Population Discovery*, a mixture-based variational autoencoder with a latent space explicitly structured by cell type. By modeling the latent space as a Gaussian mixture, MARVIN enables both annotation and subpopulation discovery within a unified framework.

To evaluate its performance, we benchmark MARVIN on public cytometry datasets and compare it to Scyan (Blampey et al., 2023), a recent generative model designed for cytometry data. We assess MARVIN's ability to recover masked subpopulations specific to peanut allergy and analyze immune response dynamics before and after allergen exposure. MARVIN reliably identifies relevant novel (unseen) subpopulations and captures their shifts across different experimental conditions.

This dual functionality makes MARVIN a powerful tool for both exploratory research and routine clinical analysis. We plan to apply this framework to investigate immune activation patterns in an ongoing clinical trial focused on vaccine response in immunocompromised patients.