AARHUS UNIVERSITY

midt
regionmidtjylland

MASTER'S THESIS IN BIOINFORMATICS

Interpretable Deep Learning for Single-Cell transcriptomics

A Linearly Decoded Variational Autoencoder

Author:
David Martin-Pestana
Student Number: 202303407

Supervisor:
Nicolai J. Birkbak
Professor

SOLIDUM PETIT IN PROFUNDIS · UNIVERSITAS ARHUSIENSIS ·

Department of Molecular Medicine
Cancer Evolution & Immunology Group

June, 2025

Acknowledgements

There are many people I'd like to thank for their support during the last 22 months of my Master's, whether nearby or halfway across Europe, but I will try to keep it brief.

I guess it is never easy to find your place at a new country, but the support of my parents, family, friends and colleagues at the Bioinformatics Research Centre and Department of Molecular Medicine made it much easier, so my first thanks go to them.

On a more academic note, I am deeply thankful to the Cancer Evolution and Immunology group for having me over the last 14 months since I started my first project with them. I also extend this thanks to every colleague in the Department of Molecular Medicine, working along such competent and bright people really elevates being here to a whole new level.

I am specially grateful to Mikkel H. Schierup, Moi Coll-Maciá and Iker Rivas-González for giving me the opportunity to work with them during my Master's. I couldn't have asked for a better first "real-life" work experience and the skills I gained under their supervision extend well over hte realm of bioinformatics.

Saving this for last because it concerns my next chapter. I would like to thank Nicolai Birkbak for giving me the opportunity to continue in the group as a PhD candidate over the next three years. I truly look forward to working full time with everyone in the group.

Abstract

The immune system undergoes profound changes with aging and in response to diseases such as COVID-19, resulting in significant inter-individual variability. Traditional bulk analysis methods obscure crucial intra-individual heterogeneity, limiting the identification of precise immune health markers. Single-cell RNA sequencing (scRNA-seq) overcomes these limitations, providing high-resolution insights into cellular diversity, yet introduces challenges such as data sparsity, batch effects, and interpretability.

In this thesis, I developed and validated an interpretable computational pipeline utilizing a Linearly Decoded Variational Autoencoder (LDVAE) to harmonize heterogeneous scRNA-seq datasets, enabling the identification of biologically meaningful gene signatures correlated with clinical outcomes. The LDVAE effectively integrated data from three independent COVID-19 Peripheral Blood Mononuclear Cells (PBMCs) datasets, significantly reducing batch-related technical noise while preserving biologically relevant information. Predictive modeling using derived gene signatures achieved stratification of patients by disease severity (macro-average AUC up to 0.74), highlighting classical monocytes as critical predictors of clinical outcomes.

On the other hand, application of this pipeline to healthy aging datasets revealed limitations associated with incomplete batch correction due to uneven age distributions across datasets, highlighting the sensitivity of batch integration methods to dataset composition. Despite these challenges, the results demonstrate the potential of LDVAE-based approaches for developing interpretable immune health markers, facilitating personalized medicine strategies.

Abbreviations

- AIDA: Asian Immune Diversity Index
- AUC: Area Under the Curve
- CCA: Canonical Correlation Analysis
- COVID-19: Coronavirus Disease 2019
- DLGM: Deep Latent Gaussian Model
- DLVM: Deep Latent Variable Model
- ELBO: Evidence Lower BOund
- FDR: False Discovery Rate
- GO: Gene Ontology
- HSC/MPP: Hematopoietic Stem Cells and Multipotent Progenitors
- KL: Kullback–Leibler divergence
- LDVAE: Linearly Decoded Variational Autoencoder
- MAIT: Mucosal-associated invariant T cells
- MNN: Mutual Nearest Neighbors
- NK: Natural Killer (cells)
- PBMC: Peripheral Blood Mononuclear Cells
- PCA: Principal Component Analysis
- QC: Quality Control
- ROC: Receiver Operating Characteristic
- scRNA-seq: Single-cell RNA sequencing
- scVI: Single-cell Variational Inference
- STRING: Search Tool for the Retrieval of Interacting Genes/Proteins
- t-SNE: t-distributed Stochastic Neighbor Embedding
- TCR: T Cell Receptor
- UMAP: Uniform Manifold Approximation and Projection
- UMI: Unique Molecular Identifier
- VAE: Variational Autoencoder
- XGBoost: eXtreme Gradient Boosting
- ZINB: Zero-Inflated Negative Binomial

Contents