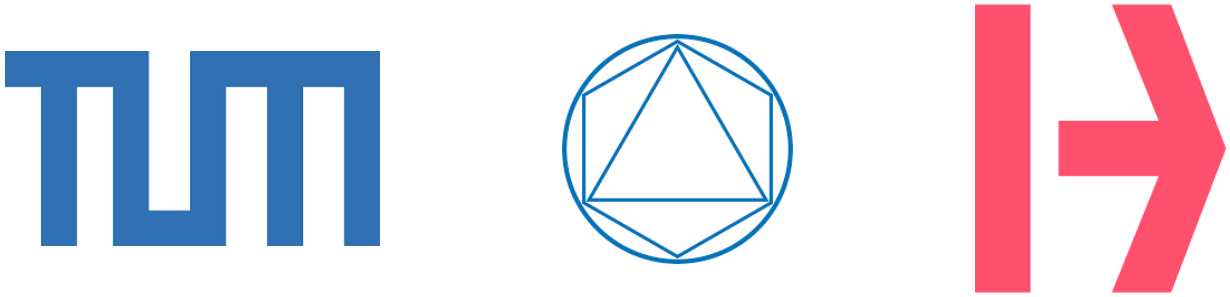

Final Report

Modeling patient variability in scRNA-seq data

M. Moghareh Dehkordi, Zahra Ebrahimi, Sandra Rjeschni and Almut Voigts

Computational Methods for Single-Cell Biology [MA5617]
Department of Mathematics, Technische Universität München, Germany

Supervisor
Vladimir Shitov
Till Richter
Malte Lücken

1 Abstract

This report investigates the variability in single-cell transcriptomics data among COVID-19 patients using patient representation methods like deep learning and distribution-based techniques. Our goal is to dissect the diversity in cellular profiles to inform disease trajectory and patient stratification. We employ a self-supervised, specifically contrastive learning framework, to generate detailed patient representations. This method highlights intricate data patterns, offering insights into cellular behaviors and supporting more personalized treatment strategies. Our findings aim to refine pseudobulk analysis practices and improve disease management by embracing these advanced analytical approaches.

2 Introduction

Single-cell transcriptomics has revolutionized our understanding of cellular heterogeneity within human tissues. However, individual differences in genetics, lifestyle, and acquired characteristics contribute to variability in cellular profiles, impacting disease progression and treatment outcomes. Despite the growing size of single-cell datasets, investigations into inter-individual variability remain limited. This project addresses this gap by evaluating patient variation in COVID-19 single-cell data.

Patient representation task. The collection of single-cell data has significantly advanced in recent decades. As the volume of data generated continues to increase, leveraging this information to accurately represent a patient's health status becomes crucial for advancing research and facilitating the early detection of diseases. {Baysoy et al., 2023}

For this, several patient representation methods have been developed considering single-cell data. Two examples of these approaches are PILOT, which uses optimal transport to facilitate comparison of cell populations across distinct individuals and PhEMD which relies on Earth mover's distance for calculating the distance of two distinct single-cell data samples. {Chen et al., 2020; Joodaki et al., 2023}

Pseudobulk baseline The simplest patient representation method and the common baseline for other approaches in the literature is "pseudobulking". In this process, certain cells, f.e. all cells from one patient, are aggregated together and analyses are run on these freshly defined "bulks". {Squair et al., 2021} While the method is far from new, it is not clear which input, aggregation method and distance metric in pseudobulk performs the best. In this paper, different options are presented and compared against each other. Examples of aggregation methods used to create pseudobulk are summing up the UMI counts of all cells of a cell type or sample or using the mean or median. {Murphy and Skene, n.d.; Ramirez Flores et al., 2023}

To quantify differences between patients' cellular information, a metric for comparison is required. An example of such a metric is the Euclidean distance calculated over pseudobulks. However, Euclidean distance may not be optimal for high-dimensional data, such as single-cell transcriptomics data. Other distance metrics are known to perform better in capturing the variability within such datasets. Choosing the *best* metric is not a simple task. There is no one-size-fits-all solution, as the quality of the metric very much depends on the structure of the data used. {Watson et al., 2022}

The lack of consistency in the choice of distance metrics and aggregation methods in pseudobulking is not the only challenge the method faces. The results are strongly dependent on the input. Using raw counts or PCAs can contain pronounced batch effects and therefore skew the resulting data.

Self-Supervised Learning (SSL) has rapidly emerged as a powerful paradigm in machine learning, enabling models to learn rich representations from unlabeled data. By exploiting the inherent structure of the data, SSL methods facilitate the learning of useful features without the need for explicit annotations, thus overcoming the limitations posed by the scarcity of labeled datasets. Among the various strategies employed within the SSL framework, contrastive learning has proven particularly effective. This approach focuses on learning embeddings by maximizing the agreement between differently augmented views of the same data point while minimizing it between unrelated points, thereby capturing the essential un-

derlying patterns of the dataset. {Balestriero and LeCun, 2022} In the context of single-cell analysis for COVID-19 patients, the application of SSL and, more specifically, contrastive learning techniques, holds significant promise. By treating each patient's cellular profile as a unique data point, we can leverage contrastive SSL to represent each patient based on their single-cell RNA sequencing data.
This study evaluates cell representations, distance metrics, and aggregation techniques for optimal pseudobulk analysis in single-cell transcriptomics, incorporating contrastive learning to enhance patient representation methods. Our objective is to improve transcriptomic analysis accuracy, facilitating a deeper understanding of cellular dynamics and aiding in the development of personalized therapeutic strategies.

3 Datasets

In this study, our methodology primarily incorporated two distinct datasets.
The initial dataset, designated as the Covid-19 Multi-omics Blood Atlas (COMBAT) dataset, included records for 140 patients with 783,704 cells per individual. The variable of interest in this dataset is called 'Outcome', which classifies the clinical status of a patient into one of six ordinal categories: 1 signifies mortality; 2 indicates intubation and mechanical ventilation; 3 corresponds to non-invasive ventilation; 4 is for patients requiring hospitalization with supplemental oxygen; 5 denotes hospitalization without the need for supplemental oxygen, and 6 represents individuals who are not hospitalized. {Wang et al., 2023}
The second dataset was published in the article "*Single-cell multi-omics analysis of the immune response in COVID-19*" by Stephenson et al. It consists of single-cell transcriptomics data from 143 patients across the UK, either healthy or with asymptomatic, mild, moderate, severe and critical COVID-19. 1,141,860 cells were registered and 781,123 passed quality control. Additional factors which influence the outcome include age and severity of infection. {Stephenson et al., 2021}

4 Methods

4.1 Bioinformatics part

To assess patient variability in single-cell data, we employ established patient representation methods for which the paper by {De Donno et al., 2023} is an example of linking heterogeneous patient metadata to the single-cell information.
To determine the optimal pseudobulking strategy, possible combinations of the Euclidean, Cosine and Manhattan metric and the aggregation methods Mean, Median and Sum were analyzed.
The data analysis was performed in scanpy v1.9.8. PCA, Harmony, scVI and scANVI dimensionality reduction methods were applied. For integration methods "Institute" was set as batch covariate in COMBAT dataset, and "Site" in Stephenson et. al. dataset. {Stephenson et al., 2021; Wang et al., 2023}

4.2 Contrastive learning for patient representation

To implement this approach, we adopted the InfoNCE {van den Oord et al., 2019} loss function, a cornerstone of contrastive learning methods within the SSL framework. The InfoNCE loss optimizes the similarity between positive pairs of samples (e.g., different transformations of the same patient's data) relative to negative pairs (e.g., samples from different patients), as given by:

$$L = -\log \frac{\exp(\text{sim}(z_i, z_j)/\tau)}{\sum_{k=1}^{N} \exp(\text{sim}(z_i, z_k)/\tau)}, \qquad (1)$$

where $\text{sim}(z_i, z_j)$ denotes the similarity metric, τ is a temperature scaling parameter, and N is the total number of samples
In our study, we further enhance the robustness of this framework by employing a Cauchy-based similarity metric. This choice is motivated by its effectiveness in handling outliers and its potential to provide a more nuanced differentiation among the samples, crucial for analyzing the complex and heterogeneous nature of single-cell data from COVID-19 patients
We utilized Optuna {Akiba et al., 2019}, which

is a hyperparameter optimization framework that automates the search for the best hyperparameters in machine learning models.

4.2.1 Network Architecture

Our network architecture, designed to process transformed pairs of single-cell data, consists of a fully connected neural network with a variable number of layers and hidden units. Specifically, the network is constructed with:

- An input layer matching the dimensionality of the processed single-cell data.
- Several hidden layers, each followed by a ReLU activation function, with the number of layers and the size of each layer being subject to optimization.
- A final linear output layer producing embeddings of a fixed size.

This design allows for the flexible adaptation of the network's complexity based on the hyperparameters selected during the optimization process. The hyperparameters subject to optimization through Optuna {Akiba et al., 2019}included:

- *Learning rate*, ranging from 1×10^{-4} to 1×10^{-1}, controlling the step size at each iteration while moving toward a minimum of the loss function.
- *Weight decay*, from 1×10^{-7} to 1×10^{-4}, adding a regularization term to the loss to prevent overfitting.
- *Temperature*, between 0.01 and 1, affecting the scaling of similarities in the contrastive loss function.
- *Batch size*, from 4 to 64, determining the number of samples processed before the model is updated.
- *Number of layers* and *hidden size*, from 1 to 7 and 30 to 400 respectively, defining the network's depth and the capacity of each layer.
- *Cell subset size*, from 100 to 6000, specifying the number of cells to sample from each patient to create pseudobulk representations.

4.2.2 Input of the Network

In preparing the input for our neural network, a pseudobulk strategy was adopted to generate input data from three different modes (PCA, scVI, scANVI), which were introduced earlier. This approach involves aggregating cellular data by randomly sampling a predefined number of cells from each patient's dataset, as specified by the *cell subset size* hyperparameter. The gene expressions of these sampled cells are then averaged to create two distinct aggregated profiles for each patient. These aggregated profiles serve as dual views of each patient's cellular data, which are inputted into the network. By analyzing these averaged cellular profiles, the network is trained to discern patterns that are more reflective of patient outcomes, effectively minimizing the influence of noise inherent in single-cell data and emphasizing biologically relevant signals.

4.2.3 Evaluation of the Network's Performance

We proceeded to evaluate the performance of our network, now fine-tuned with the best hyperparameter settings. This evaluation entailed generating latent representations for each patient's data, where the data input was constructed using a pseudobulk approach—averaging gene expressions across a subset of cells from each patient's dataset to form aggregated profiles. The latent representations were obtained by processing these aggregated cellular profiles through the trained network, capturing the essential features of the data.

The generated latent space was evaluated through a k-nearest neighbors (k-NN) test and leveraging the Spearman correlation coefficient (SCC) to measure the model's performance. This method was applied to a test dataset containing previously unseen data, aimed at providing an impartial evaluation of the model's capacity for generalization. The SCC , evaluates the strength of a monotonic relationship between predicted and actual outcomes.

To visualize the relationships among patient outcomes based on these latent representations, we employed Uniform Manifold Approximation and Projection (UMAP) {McInnes et al., 2020}, a technique for dimensionality reduction that facilitates the visualization of high-dimensional data in a two-dimensional space.

5 Results

For choosing the best combination of input type, aggregation method and distance metric, we used the one with the highest f1-score for "Status" in the Stephanson dataset and the highest Spearman-score for "Outcome" in the COMBAT dataset.

Stephenson et al. dataset

The dimensionality of the data set was first reduced using the method PCA.{Korsunsky et al., 2019} It could be observed that the variable "Site", which indicted in which city the sample were taken, was the most influential variable for constructing the principle components and had the highest f1-score for all metrics and all aggregation method used with the method PCA. This strong effect on the clustering can also be observed in Figure 1. To get ride of this

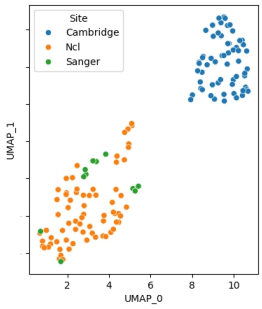

Figure 1: UMAP of the pseudobulk representation of patients from Stephenson et. al. dataset. Each point represents a patient and is obtained as calculating Manhattan distance between vectors of average principal components across the cells of each patient. For visualization purposes, UMAP dimensionality reduction is applied to the matrix of pairwise patients distances.

high batch effect we tried other methods than PCA. Firstly, Harmony was used which does not conclude in satisfactory results, since the batch effect did not really vanish and overall the scores were very low. Therefore, scVI and scANVI were used as methods and a clear reduction of the batch effect was achieved. See therefore also Figure 2. The most effective input type for pseudobulk analysis is *scVI*, with *scANVI* closely following it. They have overall the highest scores for "Status", in combination with a small f1-score for "Site" (not higher than 0.51, in comparison:

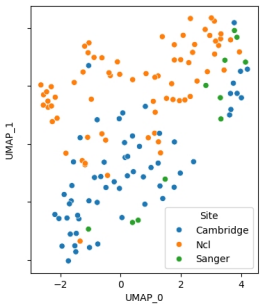

Figure 2: Analogue to Figure 1, however scVI was used as input type.

for PCA the highest score for the batch was 0.91). As a distance metric, the *Manhattan* has the highest score for "Status" in combination with all three aggregation methods and all four input types. For scVI, PCA and Harmony input, *mean* as aggregation method yields optimal results for pseudobulk analysis, whereas scANVI performs better with *Sum* as aggregation method (the f1-score is around 0.08 higher for sum than for mean). So in total we would choose a combination of *scVI*, *Manhattan* and *mean* for the pseudobulk analysis of the Stephanson dataset. An overview of the f1-scores for "Status" with the different input types, aggregation methods and distance metrics is given in Table 1.

COMBAT dataset

Here we compared as imput type PCA, scVI and scanVI. We can observe that the batch effect (noted as "Institute" in the COMBAT dataset) is less of a problem for the input type PCA in general, more for the aggregation method *Sum* (f1-score up to 0.66). Whereby the best Spearman-scores for "Outcome" are achieved with Sum but in combination for PCA with Euclidean distance, scVI with Cosine distance and for scanVI with Manhattan distance. Overall, we would choose as combination scVI with mean and cosine distance as the best combination for the pseudobulk analysis, as it achieves the highest Spearman-score for "Outcome", having an f1-score of zero for the batch "Institute". An overview of the Spearman-scores for "Outcome" with the different input types, aggregation methods and distance metrics is given in Table 2.

Table 1: F1-scores for the "Status" in Stephenson dataset

	Euclidean	Cosine	Manhattan
PCA			
Mean	0.541	0.632	**0.686**
Median	0.405	0.357	0.434
Sum	0.433	0.632	0.469
Harmony			
Mean	0.224	0.132	**0.305**
Median	0.188	0.174	0.192
Sum	0.151	0.132	0.204
scVI			
Mean	0.707	0.712	**0.760**
Median	0.617	0.630	0.689
Sum	0.472	0.712	0.512
scANVI			
Mean	0.603	0.637	0.647
Median	0.355	0.349	0.521
Sum	0.551	0.638	**0.732**

Table 2: Spearman-scores for the "Outcome" in COMBAT dataset using pseudobulk representation

	Euclidean	Cosine	Manhattan
PCA			
Mean	0.516	0.510	0.531
Median	0.441	0.474	0.470
Sum	**0.573**	0.510	0.556
scVI			
Mean	0.601	**0.644**	0.636
Median	0.520	0.599	0.549
Sum	0.564	**0.644**	0.576
scANVI			
Mean	0.582	0.603	0.600
Median	0.565	0.555	0.542
Sum	0.563	0.603	**0.625**

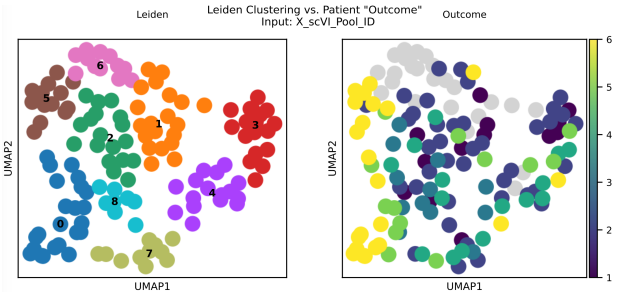

Figure 3: UMAP visualization of latent representations colored by patient outcomes, ranging from 1 (deceased) to 6 (not hospitalized). Healthy patients are observed to cluster together. The input data was processed using the scVI layer, with the optimized hyperparameters: learning rate = 0.0432, weight decay = 4.48×10^{-5}, temperature = 0.904, batch size = 4, number of layers = 1, hidden size = 192, and cell subset size = 5486.

Upon examining the performance of the contrastive model applied to the COMBAT dataset, it was determined that scVI is the most effective input. The SCC of the k-NN evaluation on the test dataset containing unseen data had a value of **0.67**[Figure 3].

The Figure 3 illustrates a UMAP analysis of latent representations derived from patient data. Notably, patients with similar outcomes are clustered together, suggesting that the latent space captures meaningful clinical variations. There is a discernible gradient, with clusters of healthier patients (higher outcome numbers) generally separated from those with more severe conditions (lower outcome numbers). Moreover, a distinct clustering of patients with conditions unrelated to the primary study condition (indicated by gray dots) suggests that the model is capable of differentiating between the studied condition and other health issues.

Table 3: Spearman correlation coefficients and F1 Macro for Severity (Outcome) in COMBAT dataset using Contrastive Learning

	SCC	F1 Macro
scVI	0.602	0.208
scANVI	0.555	0.205
PCA	0.372	0.081

Table 4: Spearman correlation coefficients and F1 Macro for Severity in Stephenson dataset using Contrastive Learning

	SCC	F1 Macro
scVI	0.718	0.389
scANVI	0.754	0.579
PCA	0.584	0.466

In the comparative analysis across both datasets, scVI and scANVI demonstrated superior performance over PCA. The 'Spearman R' score indicates the rank correlation between predicted and true labels, while 'F1 macro' represents the average F1 score across classes, balancing precision and recall.

In comparing the pseudo-bulk and contrastive learning methods, scVI outperforms scANVI with pseudo-bulk on both Stephenson and Combat datasets. However, when employing contrastive learning, scANVI excels on the Stephenson dataset but not on Combat. These outcomes suggest a potential preference for scVI with pseudo-bulk and a dataset-specific advantage for scANVI with contrastive learning.

6 Discussion

We evaluated strategies to perform pseudobulking and developed SSL-based contrastive learning approach for patient representation. The SSL model's performance was directly compared with that of the best pseudobulk method for the COMBAT dataset and assessed for its efficacy on the Stephenson dataset. Contrary to our expectations, the contrastive learning model did not outperform the leading pseudobulk method in the COMBAT dataset analysis. In both datasets, scVI and scANVI latent feautures were clearly the best input types as using them, clinical information about health status was represented better than a technical batch effect of site where the samples were taken. This is inline with single-cell integration benchmark highlighting that these methods are the best for biology preservation while removing technical variability. {Luecken et al., 2022}

Another category worth exploring would be "Days_from_onset" in the Stephenson dataset and "TimeSinceOnset" in the COMBAT dataset. This category is among the highest scoring categories in most cases and a biological argument can be made for the time-span of disease clearly influencing the type of cells present at that stage.

The findings from the evaluation of the SSL-based approach suggest a robust positive correlation, indicating that the latent features effectively encapsulate the biological variations tied to patient prognoses. This assessment, together with the model's ability to cluster healthy subjects within the UMAP visualization, highlights the network's proficiency in deriving representations that not only distinguish among various patient outcomes but also mirror the biological diversity present within single-cell data from COVID-19 patients. Importantly, this correct clustering of patients occurred even though the network was trained in an unsupervised manner and never exposed to the outcome labels.

From a technical perspective, the advancement of this research could be directed towards exploring a number of promising paths:

- *Network Architectures:* What improvements can be achieved by experimenting with varied network architectures, for instance by integrating batch normalization layers?
- *Similarity Functions in InfoNCE:* What are the effects of employing alternative similarity functions, like Gaussian similarity, on the performance of the InfoNCE-based models?
- *Loss Functions:* Could alternative loss functions, such as Margin loss, enhance the contrastive learning models' effectiveness compared to the InfoNCE?
- *Larger Datasets:* Can a model, once trained on an extensive atlas-sized dataset, be effectively utilized and transferred across other same-tissue datasets without retraining?

Acknowledgements

We would like to extend our heartfelt thanks to Vladimir Shitov and Till Richter for their guidance, patience, and support throughout this project. Their weekly meetings and dedication of time have been immensely helpful in our learning journey. We are grateful for all the knowledge and insights they have shared with us, allowing us to explore new methods. Furthermore, we want to express our appreciation for the valuable feedback given by Leon Hetzel, improving the presentations every step of the way. We thank Malte Lücken for setting the topic, playing an integral part in the development of the foundational methodology, and taking part in the seminar.

References

Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). Optuna: A next-generation hyperparameter optimization framework.

Balestriero, R., & LeCun, Y. (2022). Contrastive and non-contrastive self-supervised learning recover global and local spectral embedding methods.

Baysoy, A., Bai, Z., Satija, R., & Fan, R. (2023). The technological landscape and applications of single-cell multi-omics. *Nature Reviews Molecular Cell Biology*, 1–19.

Chen, W. S., Zivanovic, N., Van Dijk, D., Wolf, G., Bodenmiller, B., & Krishnaswamy, S. (2020). Uncovering axes of variation among single-cell cancer specimens. *Nature methods*, *17*(3), 302–310.

De Donno, C., Hediyeh-Zadeh, S., Moinfar, A. A., Wagenstetter, M., Zappia, L., Lotfollahi, M., & Theis, F. J. (2023). Population-level integration of single-cell datasets enables multi-scale analysis across samples. *Nature Methods*, *20*(11), 1683–1692.

Joodaki, M., Shaigan, M., Parra, V., Bülow, R. D., Kuppe, C., Hölscher, D. L., Cheng, M., Nagai, J. S., Goedertier, M., Bouteldja, N., et al. (2023). Detection of patient-level distances from single cell genomics and pathomics data with optimal transport (pilot). *Molecular Systems Biology*, 1–18.

Korsunsky, I., Millard, N., Fan, J., Slowikowski, K., Zhang, F., Wei, K., Baglaenko, Y., Brenner, M., Loh, P.-r., & Raychaudhuri, S. (2019). Fast, sensitive and accurate integration of single-cell data with harmony. *Nature methods*, *16*(12), 1289–1296.

Luecken, M. D., Büttner, M., Chaichoompu, K., Danese, A., Interlandi, M., Müller, M. F., Strobl, D. C., Zappia, L., Dugas, M., Colomé-Tatché, M., et al. (2022). Benchmarking atlas-level data integration in single-cell genomics. *Nature methods*, *19*(1), 41–50.

McInnes, L., Healy, J., & Melville, J. (2020). Umap: Uniform manifold approximation and projection for dimension reduction.

Murphy, A., & Skene, N. (n.d.). A balanced measure shows superior performance of pseudobulk methods in single-cell rna-sequencing analysis. nat commun. 2022; 13: 7851.

Ramirez Flores, R. O., Lanzer, J. D., Dimitrov, D., Velten, B., & Saez-Rodriguez, J. (2023). Multicellular factor analysis of single-cell data for a tissue-centric understanding of disease (J. Park & A. M. Walczak, Eds.). *eLife*, *12*, e93161. https://doi.org/10.7554/eLife.93161

Squair, J. W., Gautier, M., Kathe, C., Anderson, M. A., James, N. D., Hutson, T. H., Hudelle, R., Qaiser, T., Matson, K. J., Barraud, Q., et al. (2021). Confronting false discoveries in single-cell differential expression. *Nature communications*, *12*(1), 5692.

Stephenson, E., Reynolds, G., Botting, R. A., Calero-Nieto, F. J., Morgan, M. D., Tuong, Z. K., Bach, K., Sungnak, W., Worlock, K. B., Yoshida, M., et al. (2021). Single-cell multi-omics analysis of the immune response in covid-19. *Nature medicine*, *27*(5), 904–916.

van den Oord, A., Li, Y., & Vinyals, O. (2019). Representation learning with contrastive predictive coding.

Wang, D., Kumar, V., Burnham, K. L., Mentzer, A. J., Marsden, B. D., & Knight, J. C. (2023). Combatdb: A database for the covid-19 multi-

omics blood atlas. *Nucleic Acids Research*, *51*(D1), D896–D905.

Watson, E. R., Mora, A., Taherian Fard, A., & Mar, J. C. (2022). How does the structure of data impact cell–cell similarity? Evaluating how structural properties influence the performance of proximity metrics in single cell RNA-seq data. *Briefings in Bioinformatics*, *23*(6), bbac387. https://doi.org/10.1093/bib/bbac387