Evaluating the importance of input data representation for single cell deep learning classifier

Bruno Puczko-Szymański

June 25, 2025

Abstract

The SARS-CoV-2 pandemic, along with lung cancer and other respiratory diseases, is responsible for a very high number of deaths around the world. Understanding the cellular diversity of the lungs can contribute to the development of more effective drugs and therapies. Currently, one of the most popular methods used to study the cellular composition of a given tissue is single-cell RNA sequencing (scRNA-seq). Identifying cell types based on sequencing results is a complex task that requires expert knowledge. To simplify and accelerate this process, machine learning methods are proposed. In this work, we propose a deep learning classifier trained on scRNA-seq data obtained from COVID-19 patients. Additionally, we found that training a model in gene expression space suffers from severe overfitting, so we explore different data representations including the usage of external models such as SCVI and SCANVI which seems to deal with this problem and also achieve better results. The whole code for our analysis can be found in Github repository

1 Introduction

Single-cell RNA sequencing (scRNA-seq) has revolutionized transcriptomic analysis by enabling the investigation of gene expression at the resolution of individual cells. This level of granularity allows researchers to characterize cellular heterogeneity, uncover novel cell types, and study dynamic processes such as immune responses and disease progression [1]. Deep learning models have shown great promise in addressing the complexity of high-dimensional biological data, particularly in supervised tasks like cell type prediction [2]. However, these models are also susceptible to overfitting, especially when trained on raw gene expression profiles with tens of thousands of features. Moreover, imbalanced class distributions, typical in scRNA-seq datasets, further complicate the development of robust classifiers. To mitigate these issues, dimensionality reduction and representation learning techniques are increasingly used to transform raw gene expression data into more compact and informative embeddings. Among the most widely used are scVI and scANVI—variational autoencoder (VAE)-based models that learn low-dimensional latent representations of cells. While scVI is an unsupervised method that captures transcriptomic variation, scANVI extends scVI with semi-supervised learning, incorporating available cell-type labels to improve interpretability and classification performance [3]. In this study, we investigate the effectiveness of deep learning classifiers trained on raw gene expression versus embeddings generated by scVI and scANVI. Using a publicly available scRNA-seq dataset of bronchoalveolar lavage fluid from 10 COVID-19 patients [4], which includes annotations for 28 immune and epithelial cell types, we evaluate the impact of different preprocessing pipelines on classification performance. We explore how the number of highly variable genes (HVGs) affects model accuracy, and assess how representation learning methods influence generalization and overfitting. Our findings show that while raw expression models are prone to overfitting—especially as the number of HVGs increases—scVI and scANVI offer significant improvements in both accuracy and robustness. Notably, scANVI embeddings lead to the best overall results, demonstrating the value of

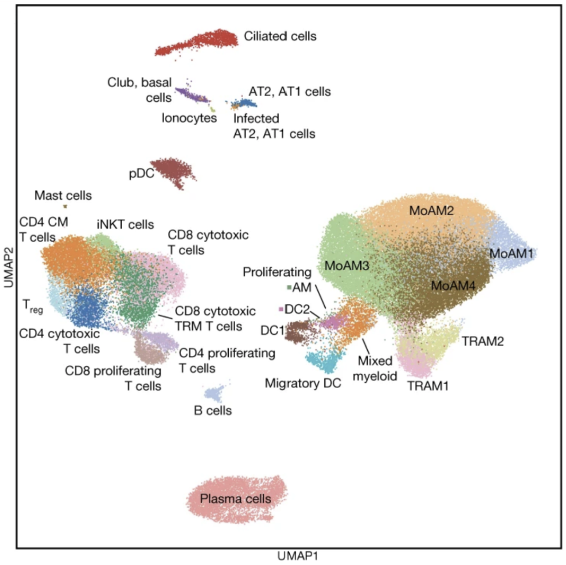

Figure 1: UMAP projection of the cells in our data. [4]

label-aware representation learning in noisy, high-dimensional single-cell data. This work highlights the importance of careful data representation in scRNA-seq classification tasks and supports the use of generative latent models for scalable, interpretable analysis.

2 Methods and Materials

2.1 Dataset Description

In this study, we used the single-cell RNA sequencing (scRNA-seq) dataset published by Rogan A. Grant et al. [4]. The dataset was generated from bronchoalveolar lavage fluid samples collected from 10 patients with COVID-19. It contains gene expression profiles for 77,146 cells across 21,819 genes. The authors provided both raw count data and a preprocessed version of the dataset. According to the original publication, the data were processed using a standard single-cell RNA-seq analysis pipeline, which included filtering low-quality cells, removing mitochondrial genes, normalization, and standardization. Additionally, the authors identified and labeled the 4,488 most highly variable genes, which were used in our analyses. Each cell in the dataset is annotated with a corresponding cell type label, resulting in a total of 28 distinct cell types: AT2/AT1 cells, B cells, CD4 central memory (CM) T cells, CD4 cytotoxic T cells, CD4 proliferating T cells, CD8 cytotoxic T cells, CD8 cytotoxic TRM T cells, CD8 proliferating T cells, Ciliated cells, Club/Basal cells, DC1, DC2, Infected AT2/AT1 cells, Ionocytes, Mast cells, Migratory DC, Mixed myeloid, MoAM1, MoAM2, MoAM3, MoAM4, Plasma cells, Proliferating alveolar macrophages (Prolif. AM), TRAM1, TRAM2, Treg, iNKT cells, pDC. As illustrated in Figure 2, the distribution of cell types in the dataset is highly imbalanced. This heterogeneity in cell type proportions was taken into account during model design and evaluation. The data can be downloaded from Gene Expression Omnibus with accession number GSE155249. It is possible to download both RAW and preprocessed data.

2.2 Data Preparation for Deep Learning Classification

To train and evaluate our deep learning models, we split the dataset into training, validation, and test sets using a 70/20/10 ratio. Given the strong class imbalance across the 28 cell types, we employed stratified sampling to maintain consistent class distributions across all subsets. This ensured that rare

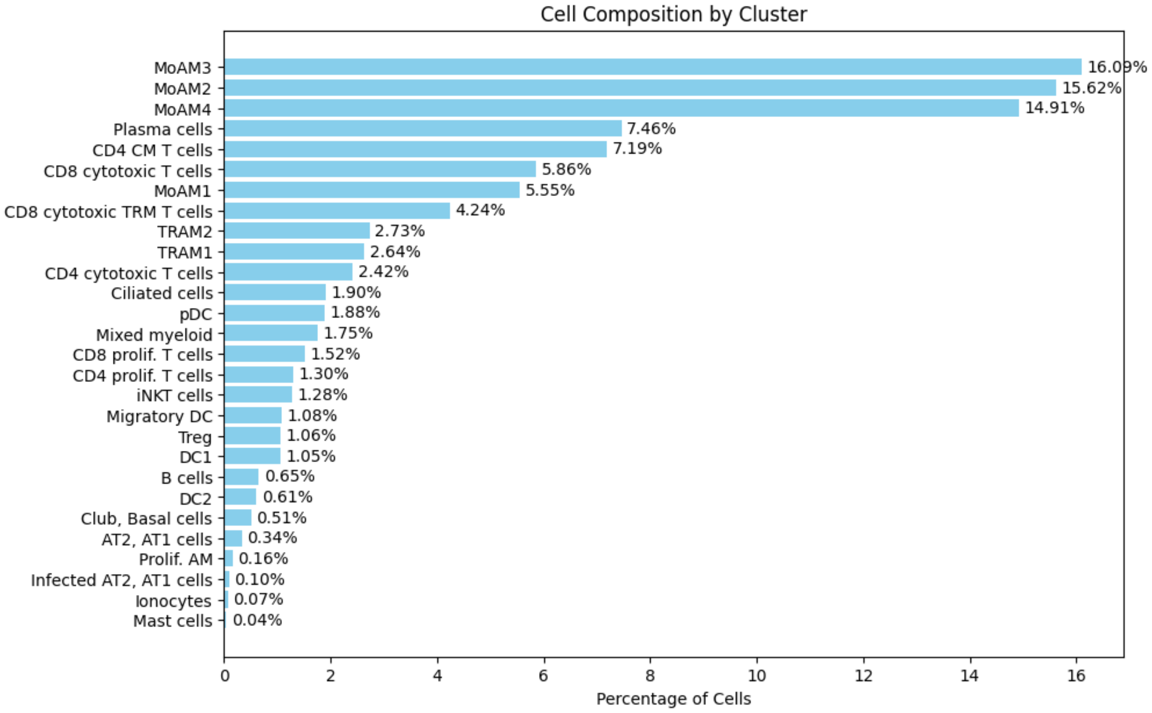

Figure 2: Cell type distribution in the dataset from Rogan A. Grant et al.

cell types were proportionally represented in each split, reducing the risk of biased model performance. Model architecture exploration and hyperparameter tuning were performed using the validation set, while final performance metrics were computed on the held-out test set. This separation of validation and test sets ensured that the reported results reflect the model's generalization ability and are not biased by model selection.

2.3 Base Model Architecture Selection

Before evaluating the impact of different input data representations, we first aimed to identify a reasonably strong and computationally efficient base architecture for our classifier. The goal of this exploration was not to find the best-performing model possible, but rather to establish a solid baseline that balances predictive performance with training efficiency. The first parameter we investigated was the depth of the neural network, i.e., the number of hidden layers. We evaluated architectures with 2 to 4 hidden layers, using the following layer sizes in descending order: 512, 256, 128, and 64 neurons. All configurations showed similar validation accuracy, suggesting that the number of hidden layers had limited impact on performance. However, we observed severe overfitting across all tested depths: models achieved over 96% accuracy on the training set but only around 86% on the validation set. Based on this observation, we selected a network with 3 hidden layers of sizes 512, 256, and 128 neurons as a compromise between performance and simplicity for subsequent experiments. We also tuned the learning rate and found that a value of $1e^{-4}$ yielded stable and efficient convergence. The batch size was set to 128, and training was performed using the Adam optimizer with weight decay to mitigate overfitting. All models were trained for 15 epochs. The cross-entropy loss function was used as the training objective. To address overfitting and further improve generalization, we tested the inclusion of batch normalization and dropout layers. We applied batch normalization and a dropout rate after each hidden layer. These additions had a positive effect on training stability and slightly reduced overfitting, although a performance gap between training and validation remained. A summary of training and validation metrics for the different configurations are shown in Table 1. The final base architecture used in subsequent experiments included three hidden layers (512, 256, 128), ReLU activations, batch normalization and dropout (0.5), a learning rate of $1e^{-4}$, weight decay of $1e^{-5}$, and training for 15 epochs using Adam optimizer and cross-entropy loss.

Model	Accuracy Train	Accuracy Val	Precision Train	Precision Val
2 layers + BN + D(0.2)	0.97	0.88	0.97	0.83
3 layers + BN + D(0.3)	0.97	0.87	0.93	0.87
3 layers + BN + D(0.5)	0.97	0.87	0.96	0.90
3 layers	1.0	0.88	1.0	0.90
3 layers + BN + D(0.5) + WD	0.96	0.87	0.96	0.90
4 layers + BN + D(0.5)	0.97	0.87	0.93	0.82

Table 1: Performance metrics for selected models. The selected architecture is highlighted in red color. BN is Batch Normalization, D is Dropout and WD is Weight Decay.

2.4 scVI and scANVI

The setup and usage of both scVI and scANVI was carried out according to the tutorial from https://docs.scvi-tools.org/en/latest/tutorials/notebooks/scrna/harmonization.html. We found that classifiers trained on the embedding of these models need a learning rate equal to $7e^{-3}$. All other parameters were left as a default as suggested by the tutorial authors.

3 Results

3.1 Impact of the Number of Highly Variable Genes on Classifier Performance

We investigated how the number of highly variable genes (HVGs) influences the performance of the classifier. The results are shown in Table 2. The first observation is that models trained with a smaller number of HVGs, such as 100, 200, and 500, perform significantly worse across all evaluation metrics. This can be attributed to the limited amount of information available to the model, which hinders its ability to accurately classify cell types. Interestingly, overfitting is minimal or entirely absent in models using fewer HVGs. This aligns with the idea that reducing the input dimensionality also reduces noise, which is particularly high in single-cell RNA-seq data. For models trained with 1000, 2000, and 4488 HVGs, we observe a steady improvement in performance metrics. However, these models also exhibit an increasing overfitting, as evidenced by the growing gap between training and validation metrics. When the number of HVGs increases further, to 10,000 and the complete set of 21,819 genes, overfitting becomes a major issue. Notably, this increase in input size does not lead to better performance on the validation set, suggesting diminishing returns and possibly the introduction of additional noise that negatively affects generalization.

3.2 Performance of Models Trained on scVI and scANVI Embeddings

We evaluated the performance of models trained on embeddings generated using the scVI and scANVI frameworks. Both methods produced latent embeddings of dimensionality 10, computed using two sets of HVGs: 1,200 and 4,369 genes, respectively. The results are shown in Table 3. Our results clearly show that models trained on scVI and scANVI embeddings substantially outperform those trained directly on raw gene expression data. This performance gain can be attributed in part to the ability of latent embeddings to capture informative structure in the data while reducing dimensionality and noise. Additionally, models trained on larger HVG sets generally performed better than those trained on the smaller subset, likely due to the richer feature space. Interestingly, overfitting was not

No. HVGs	Accuracy Train	Accuracy Val	Precision Train	Precision Val	F1 Train	F1 Val
100	0.61	0.63	0.53	0.57	0.48	0.52
200	0.69	0.70	0.62	0.62	0.58	0.60
500	0.79	0.78	0.74	0.72	0.70	0.70
1000	0.85	0.81	0.87	0.84	0.80	0.79
2000	0.93	0.84	0.93	0.87	0.89	0.83
4488	0.99	0.87	0.98	0.90	0.97	0.87
10000	0.99	0.84	0.99	0.87	0.99	0.84
21819	0.99	0.85	0.99	0.88	0.99	0.85

Table 2: Performance metrics for models with different number of HVGs.

observed in models trained on scVI and scANVI embeddings, in contrast to models trained directly on gene expression vectors, which suffered from significant overfitting. This indicates that the latent representations produced by these methods are not only compact but also more generalizable. Furthermore, scANVI consistently outperformed scVI across all tested configurations, which aligns with expectations given that scANVI extends scVI by incorporating label supervision during training. The best performance overall was achieved by the model trained on scANVI embeddings with 4,369 HVGs, which reached scores exceeding 0.94 across all evaluation metrics (accuracy, precision, and F1). In the following subsection, we analyze the behavior and performance of this best-performing model in more detail.

Model	Accuracy Train	Accuracy Val	Precision Train	Precision Val	F1 Train	F1 Val
scVI 1200	0.81	0.84	0.83	0.87	0.82	0.84
scVI 4369	0.83	0.86	0.84	0.87	0.83	0.86
scANVI 1200	0.91	0.93	0.93	0.94	0.92	0.94
scANVI 4369	0.96	0.98	0.94	0.96	0.94	0.96

Table 3: Performance metrics for models with different number of HVGs.

3.3 Deeper look into best models

Up to this point, we identified two best-performing models: one trained directly on the gene expression space using 4,488 HVGs, and another trained on the latent representation generated by scANVI using 4,369 HVGs. In this subsection, we analyze their performance in greater detail by evaluating them on a completely unseen test dataset. The corresponding confusion matrices are presented in Figure 3 and 4, which provides information on which cell types are most frequently misclassified and how the performance of the model varies between different parts of the label space. The first observation from Figure 3 is that the model performs particularly poorly on two cell types: Infected AT2, AT1 cells and Proliferating AM. The likely explanation is class imbalance: those cell types are among the least abundant in the dataset, which may have led to insufficient training examples for the model to learn distinguishing features. Additionally, Proliferating AM cells are frequently misclassified as DC2 (16.7%), Mixed myeloid (25%), and MoAM3 (16.7%). As shown in Figure 1, these cell types are spatially clustered together in the UMAP projection. This spatial proximity suggests that their gene expression profiles are similar, which may explain the classifier's difficulty in distinguishing between them.

The classifier trained on the scANVI latent space demonstrates strong performance across nearly all cell types (Figure 4). On the Figure 5 we can see that the latent space from scANVI model in fact does a good job to separated different cell types. As expected, the most challenging classes were Mixed myeloid cells and Proliferating AM. Notably, in this case, the Prolif. AM cells were

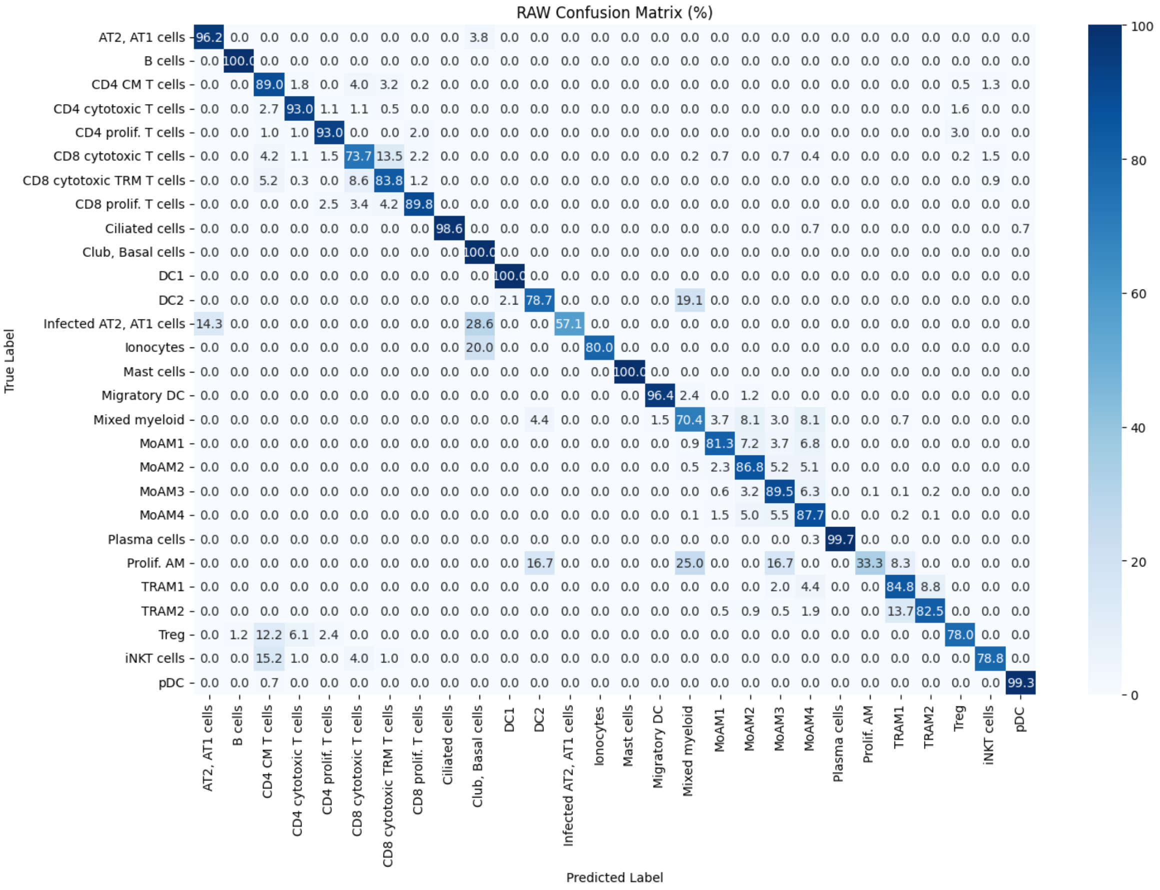

Figure 3: Confusion matrix for classifier trained on the gene space.

most frequently misclassified as B cells. Despite these specific difficulties, all remaining cell types were classified with high accuracy (above 96%), indicating robust generalization. These results suggest that the model successfully captures essential information for cell-type discrimination within a 10-dimensional embedding space - a significant reduction compared to the 4488-dimensional gene expression space used in earlier models. This observation underscores two key points: first, that scANVI produces compact and informative representations; second, that raw expression data can be highly noisy, making models trained directly on gene space more prone to overfitting. The results highlight the importance of using carefully regularized models and informative data representations in single-cell classification tasks.

4 Discussion

In this work, we developed a simple yet effective deep learning architecture for classifying 28 distinct cell types based on single-cell RNA sequencing expression data. We first explored the impact of using different numbers of highly variable genes (HVGs) on classification performance. Our results indicated that while increasing the number of genes provided richer information, it also significantly increased the risk of overfitting, especially in models trained directly on high-dimensional gene expression space. To address this issue, we employed two generative models, scVI and scANVI, to obtain low-dimensional embeddings of size 10 from the same data. Classifiers trained on scVI embeddings achieved performance comparable to those trained on raw gene expression, but with substantially reduced overfitting. More notably, classifiers trained on scANVI embeddings demonstrated

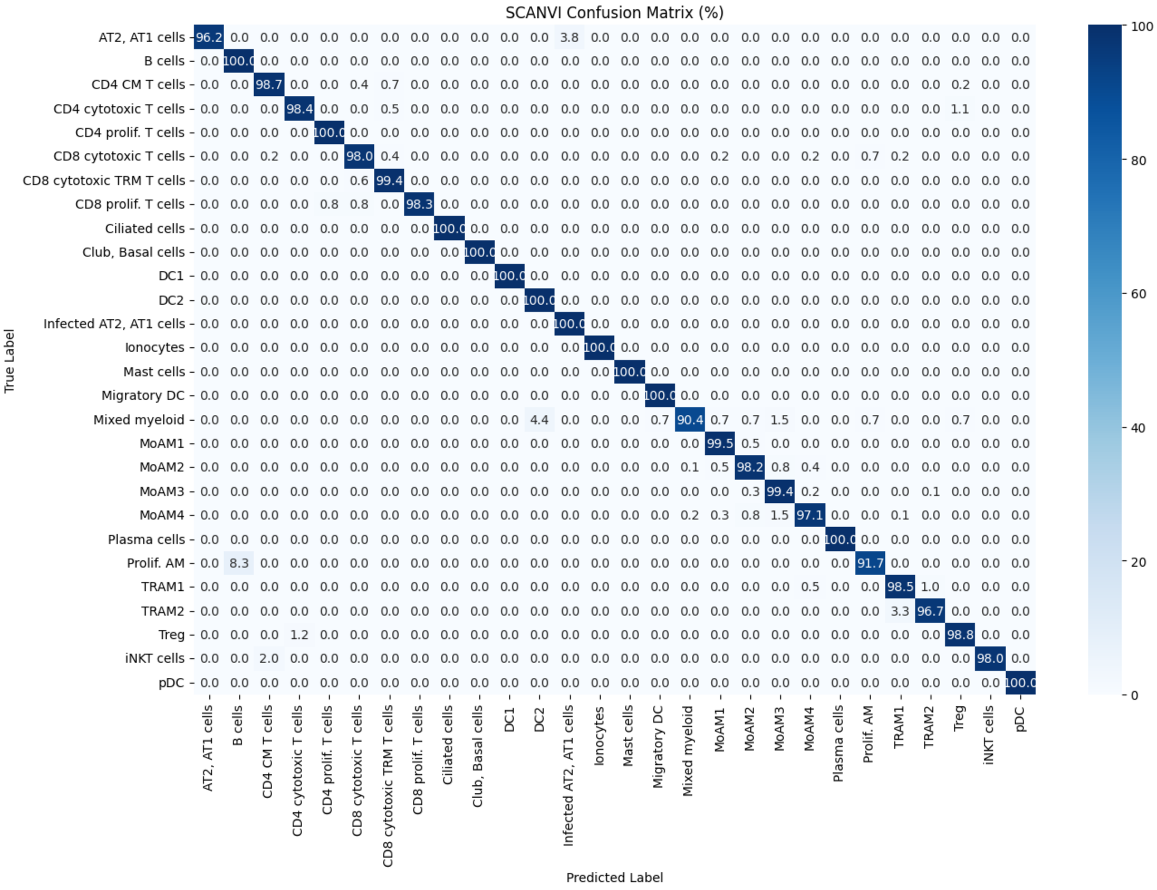

Figure 4: Confusion matrix for classifier trained on the latent space obtained from the scANVI model.

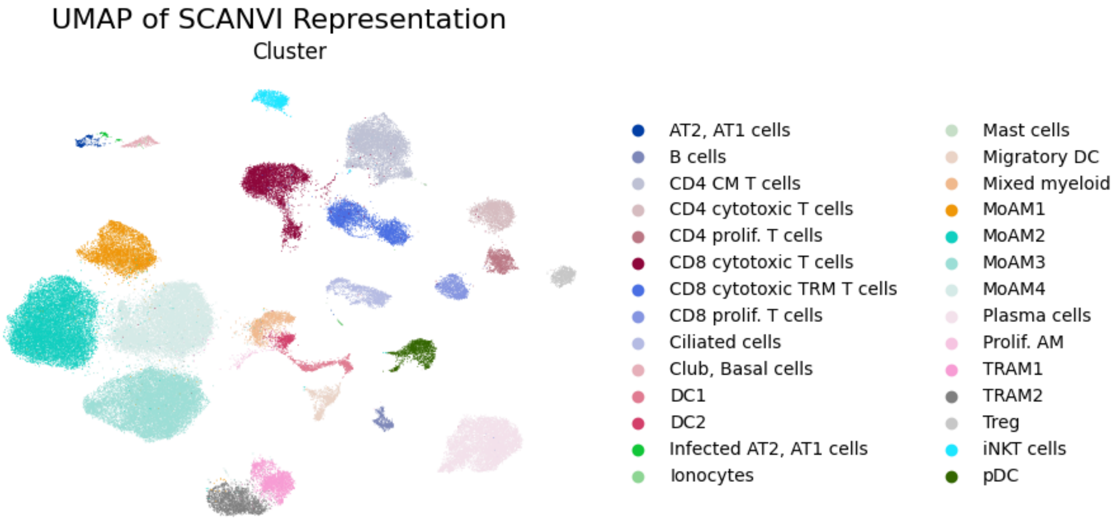

Figure 5: UMAP projection of the embedding obtained using scANVI model.

a significant improvement in accuracy, precision, and F1 score, without clear signs of overfitting. This is consistent with the semi-supervised nature of scANVI, which incorporates label information during embedding learning. Our findings emphasize the importance of careful data preprocessing and representation learning in single-cell analysis. Raw scRNA-seq data is inherently noisy and sparse, which can mislead even well-regularized models. By using structured latent representations, we can achieve both high classification performance and strong generalization to unseen data.

5 Acknowledgment

In this article we used ChatGPT (OpenAI, 2025) to assist with language editing, formatting suggestions, and structure refinement of the manuscript. All final content was reviewed and approved by the authors.

References

1. Haque, A., Engel, J., Teichmann, S. A. & Lönnberg, T. A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. *Genome medicine* **9,** 1–12 (2017).
2. Ma, F. & Pellegrini, M. ACTINN: automated identification of cell types in single cell RNA sequencing. *Bioinformatics* **36,** 533–538 (2020).
3. Xu, C. *et al.* Probabilistic harmonization and annotation of single-cell transcriptomics data with deep generative models. *Molecular systems biology* **17,** e9620 (2021).
4. Grant, R. A. *et al.* Circuits between infected macrophages and T cells in SARS-CoV-2 pneumonia. *Nature* **590,** 635–641 (2021).