Benchmarking Deep Generative Models and Classical Baselines for Single-Cell Analysis of Peripheral Blood and Thymic Tissues in Myasthenia Gravis

María Dolores Navarro Maturana[*], Daria Lykova[**], Yitian Tan[***]

[*] SPS, Columbia University
[**] Columbia College, Columbia University
[***] SEAS, Columbia University

GitHub Repository:
https://github.com/daria-ly/Applied_DataScience-Final-Project-.git

I. INTRODUCTION

Single-cell RNA sequencing (scRNA-seq) has become a central tool for studying immune-mediated diseases by enabling high-resolution characterization of cellular heterogeneity and transcriptional states. In autoimmune disorders such as myasthenia gravis (MG), scRNA-seq offers a powerful means of probing both systemic immune alterations in peripheral blood and localized immune dysregulation within disease-relevant tissues, particularly the thymus. MG is characterized by autoantibody production against neuromuscular junction components and is strongly associated with thymic abnormalities, including hyperplasia, ectopic germinal centers, and altered T cell development. Understanding how immune cell states differ between peripheral blood and thymic tissue is therefore essential for elucidating MG pathophysiology. A central challenge in scRNA-seq analysis lies in learning low-dimensional representations that faithfully capture biologically meaningful structure while accounting for technical noise and dataset-specific variation. A wide range of representation-learning methods has been proposed, ranging from classical linear techniques to probabilistic factor models and deep generative approaches. While increasingly expressive models are often assumed to yield superior biological insight, their performance depends critically on the structure of the data and the evaluation criteria used. In practice, it remains unclear when complex generative models provide tangible advantages over simpler baselines, particularly when the analytical objective is to recover known immune cell identities or disease-associated structure.

In this project, we systematically benchmark representation-learning methods for scRNA-seq data in MG across two biological contexts: peripheral blood mononuclear cells (PBMCs) and thymic tissue. The input to our analysis consists of raw scRNA-seq count matrices and associated cell- and sample-level metadata from publicly available MG and healthy control datasets. The output of each method is a low-dimensional latent representation of cells, which is evaluated based on its ability to preserve immune cell-type structure, maintain biologically meaningful neighborhood relationships, and reflect disease-associated variation.

PBMCs are analyzed as a controlled, single-tissue setting in which immune populations are relatively well-defined and close to linearly separable. This setting allows us to assess whether deep generative modeling provides benefits beyond classical linear dimensionality reduction. In contrast, thymic tissue presents a substantially more complex modeling problem due to its architectural heterogeneity, developmental gradients, and disease-specific remodeling. By comparing multiple representation-learning approaches across these two settings, our goal is not to identify a universally “best” method, but rather to clarify how modeling assumptions, data complexity, and evaluation metrics interact in practice. This work contributes a principled, comparative perspective on representation learning for autoimmune scRNA-seq data and highlights the importance of aligning method choice with specific analytical goals.

II. RELATED WORK

Recent advances in single-cell and spatial transcriptomics have substantially improved our understanding of immune dysregulation in myasthenia gravis (MG). Early single-cell RNA sequencing studies of peripheral blood demonstrated pronounced transcriptional heterogeneity among immune populations in MG patients, revealing disease-associated B-cell and T-cell subtypes that are obscured in bulk analyses. Jin et al. (2021) showed that PBMC scRNA-seq can identify immune subgroups correlated with disease activity, highlighting the importance of cell-resolved approaches for studying MG-related immune remodeling.

Subsequent work shifted attention toward the thymus as a central site of MG pathogenesis. Yasumizu et al. (2022) characterized thymoma-associated immune and stromal populations using scRNA-seq, uncovering aberrant gene expression in medullary thymic epithelial cells and altered immune–stromal interactions. More recently, spatial transcriptomics has enabled direct localization of disease-relevant immune niches within thymic tissue. Yasumizu et al. (2024) integrated Visium spatial transcriptomics with scRNA-seq to map medullary and germinal center–like structures in MG-associated thymomas, providing compelling evidence that spatially organized immune environments contribute to autoantibody production. Additional studies have reported clonal B-cell expansion and dysregulated cytokine signaling within MG thymus, further reinforcing the need to analyze peripheral and thymic immune compartments in parallel.

From a computational perspective, integrating and representing heterogeneous single-cell datasets remains an active area of research. Classical approaches such as principal component analysis (PCA), are widely used due to their simplicity, interpretability, and computational efficiency. However, these linear methods may struggle to capture complex, nonlinear biological variation and require careful tuning to balance batch mixing against signal preservation. Deep generative models, particularly variational autoencoder–based frameworks such as scVI and scANVI, have emerged as state-of-the-art tools for single-cell analysis, offering probabilistic modeling of technical noise and improved robustness to dataset heterogeneity. Bayesian factor models such as MOFA+ provide an alternative probabilistic framework for uncovering shared and dataset-specific sources of variation across conditions and modalities.

While these methods have been applied successfully in diverse biological contexts, their relative performance in capturing immune structure in MG single-cell datasets has not been systematically evaluated across tissues and model classes. Most prior MG studies emphasize biological discovery using a single computational approach rather than comparing multiple representation-learning methods under a unified evaluation framework. In this work, we address this gap by benchmarking classical, probabilistic, and deep generative models on PBMC and thymic scRNA-seq datasets from MG patients and healthy controls, with the goal of clarifying how modeling assumptions, data complexity, and evaluation criteria jointly influence representation quality.

III. DATA

Dataset 1: Peripheral Blood Mononuclear Cell (PBMC) scRNA-seq (Okuzono et al., 2024)

The first dataset consists of single-cell RNA sequencing (scRNA-seq) profiles of peripheral blood mononuclear cells (PBMCs) collected from patients with myasthenia gravis (MG) and healthy controls. Generated using the DNBSEQ-G400 platform, the dataset consists of 40 samples with a total of tens of thousands of cells spanning multiple MG subtypes, including seronegative MG and AChR-positive MG, alongside healthy individuals. The single-cell resolution enables detailed characterization of circulating immune cell types, including CD4^{+} and cytotoxic T cells, B cells, monocytes, dendritic cells, and natural killer cells. Both raw count matrices and extensive sample- and cell-level metadata are publicly available through the Gene Expression Omnibus (GEO).

In this project, the PBMC dataset is analyzed independently to characterize how myasthenia gravis affects circulating immune cell states and compositions in peripheral blood. This dataset is directly relevant to the project as it serves as a controlled benchmark for comparing a deep generative model (scVI) and classical baseline in their ability to capture disease-associated transcriptional structure within a single tissue modality. By restricting the analysis to PBMCs, this dataset helps us evaluate model performance in recovering known MG-related immune patterns without confounding effects from cross-tissue or cross-modality integration.

Preprocessing: .

Raw count matrices were downloaded from GEO and loaded on a per-sample basis, with cells subsampled per donor to control memory usage. Samples were concatenated into a single AnnData object while preserving donor identity (`sample_id`) and condition labels (Healthy, AChR-positive MG, Seronegative MG). Standard quality control was applied, including filtering cells with high mitochondrial content (>15%), low or extreme gene counts (<200 or >6000), and removing genes expressed in fewer than 10 cells. The resulting dataset was saved as a QC object that preserves counts.

This QC-filtered, raw-count version (`pbmc_qc.h5ad`) was used for our deep generative model (scVI), which requires integer counts to run. In parallel, the data were normalized and log-transformed to create an intermediate representation (`pbmc_normalized_log1p.h5ad`), which served as a parent object for downstream analyses.

Before highly variable gene (HVG) selection, a copy of the log-normalized data was saved (`pbmc_DE_ready.h5ad`) to retain the full gene set for differential expression analyses, which were used as sanity checks to ensure datasets were not corrupted (see code for figures). Finally, highly variable genes were selected using a batch-aware procedure (`seurat_v3`, accounting for `sample_id`), and the dataset was restricted to the top 2,000 HVGs (`pbmc_HVG_2000.h5ad`) for classical baseline analyses, including PCA, clustering, and UMAP visualization.

Datasets 2 and 3: Thymus scRNA-seq (MG and Healthy Reference)

To study disease-associated immune dysregulation in the thymus, we use two single-cell RNA sequencing datasets that together represent pathological and healthy thymic states. These datasets are analyzed jointly to enable direct comparison between MG-associated thymic alterations and normal thymic organization.

Dataset 2: MG Thymus scRNA-seq (Single Cell Portal accession SCP1532)

Dataset 2 consists of single-cell RNA sequencing (scRNA-seq) profiles derived from thymic tissue of patients with myasthenia gravis (MG). The original study includes both thymus samples and matched peripheral blood mononuclear cells (PBMCs); however, only thymus-derived cells are retained in this project. The dataset captures a diverse range of thymic immune and stromal populations, including developing T cells, B cells, myeloid cells, thymic epithelial cells, and other structural cell types implicated in MG pathophysiology.

Due to its large size, the dataset was processed in chunks during file reading and then concatenated into a single AnnData object. Author-provided cell-type annotations and ontology-based labels were preserved where available, enabling biologically grounded interpretation of thymic cell states.

Dataset 3: Healthy Human Thymus scRNA-seq Atlas (Yayon et al., 2024)

Dataset 3 is a large-scale single-cell RNA sequencing atlas of the healthy human thymus, comprising approximately 482,651 cells profiled across four complementary single-cell assays. Generated as part of a comprehensive effort to map thymic cellular composition and developmental trajectories, the dataset includes a wide range of thymic cell populations, such as thymocytes at multiple maturation stages, thymic epithelial cell subtypes, stromal cells, fibroblasts, and endothelial cells. In our project this dataset represents baseline, non-diseased thymic biology and serves as a high-resolution reference for normal thymic cell states and transcriptional programs.

Preprocessing (Datasets 2 and 3: Thymus scRNA-seq).
Both datasets were loaded into Scanpy, with the MG thymus data processed in chunks due to file size. Only thymus-derived cells were retained from the MG dataset, and original cell-type annotations and ontology labels were kept where available. Quality control was applied separately to each dataset, removing cells with high mitochondrial content (>15%), low or extreme gene counts (<200 or >6000), extreme library sizes, and genes expressed in fewer than 10 cells. The QC-filtered MG and healthy thymus datasets were saved as intermediate objects. To balance the comparison, the healthy thymus dataset was randomly downsampled to match the post-QC size of the MG thymus dataset. Gene names were standardized, and both datasets were restricted to their shared gene set to ensure identical feature spaces. The resulting subsets were concatenated into a single AnnData object with dataset and condition labels preserved. A raw counts layer was explicitly stored to support deep generative modeling, and this combined thymus dataset served as the primary input for downstream integration and benchmarking.

IV. METHODS

Overview of Analytical Strategy

We adopt a structured benchmarking framework to evaluate how different representation-learning approaches capture disease-associated immune structure in myasthenia gravis (MG). Specifically, we compare classical linear methods, batch-correction approaches, probabilistic factor models, and deep generative models across two biological contexts: peripheral blood mononuclear cells (PBMCs) and thymic tissue.

PBMCs are analyzed independently as a controlled, single-tissue setting in which disease effects can be examined without cross-tissue heterogeneity. In contrast, thymic datasets from MG patients and healthy donors are analyzed jointly, reflecting the complex tissue architecture and disease-specific immune remodeling central to MG pathophysiology. Across both contexts, models are evaluated using a combination of qualitative visualization and quantitative representation metrics to enable principled comparison.

PBMC Analysis Pipeline

Classical Baseline: Principal Component Analysis (PCA)

As a classical baseline, PBMC gene expression data were embedded using principal component analysis (PCA). Given a centered and scaled gene expression matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$, PCA identifies orthogonal directions of maximal variance by solving $\max_{\mathbf{W}^\top \mathbf{W}=I} \mathrm{Var}(\mathbf{XW})$, where the columns of W are the eigenvectors of the sample covariance matrix.

Cells were projected into the resulting low-dimensional space, from which neighborhood graphs were constructed. These graphs were used for UMAP visualization and Leiden clustering to identify putative immune populations. Cell-type identities were assigned using marker-gene scoring, and cluster-level labels were determined by majority vote. This PCA-based representation serves as a transparent and widely used baseline against which more complex models are compared.

Deep Generative Model: scVI

To learn a probabilistic latent representation of PBMC transcriptional structure, we applied scVI, a variational autoencoder (VAE) specifically designed for single-cell RNA-sequencing data. scVI models observed gene counts x_{ig} for cell i and gene g using a negative binomial likelihood:

$$x_{ig} \sim \mathrm{NB}(\mu_{ig}, \theta_g), \quad \mu_{ig} = s_i \cdot f_\theta(\mathbf{z}_i)_g$$

where $\mathbf{z}_i \in \mathbb{R}^d$ is a low-dimensional latent variable, s_i is a cell-specific size factor, and $f_\theta(\cdot)$ is a neural decoder network.

Latent variables are inferred via amortized variational inference by maximizing the evidence lower bound (ELBO):

$$\mathcal{L}_{\mathrm{ELBO}} = \mathbb{E}_{q(\mathbf{z}|\mathbf{x})}\left[\log p(\mathbf{x}|\mathbf{z})\right] - \mathrm{KL}\left(q(\mathbf{z}|\mathbf{x}) \,\|\, p(\mathbf{z})\right)$$

The model was trained with a 20-dimensional latent space. The learned latent representation was used to construct neighborhood graphs, UMAP embeddings, and Leiden clusters directly in scVI space. Marker-based annotation was repeated to enable direct comparison with PCA-based results.

Thymus Analysis Pipeline (MG and Healthy)

Thymic tissue presents a substantially more challenging modeling problem due to strong biological heterogeneity, disease-associated architectural changes, and inter-dataset variation. To address this, we benchmarked four complementary approaches: PCA, Harmony, MOFA+, and scVI.

Linear Baseline: PCA

As in the PBMC analysis, PCA was used as a linear baseline for thymic data. PCA provides a low-dimensional summary of transcriptional variance but does not explicitly account for batch effects or disease-specific structure. PCA embeddings were used to assess the extent to which disease-associated thymic variation can be recovered without correction or probabilistic modeling.

Probabilistic Factor Model: MOFA+

To model shared and condition-specific sources of variation, we applied MOFA+, a Bayesian latent factor model. MOFA+ represents gene expression as

$$\mathbf{X}^{(v)} = \mathbf{W}^{(v)}\mathbf{Z} + \boldsymbol{\varepsilon}^{(v)},$$

where $X^{(v)}$ denotes data from view v, Z is a matrix of latent factors, $W^{(v)}$ are view-specific loadings, and $\varepsilon^{(v)}$ is Gaussian noise.

In our setting, MOFA+ was applied to the combined MG–healthy thymus dataset to identify latent factors capturing disease-associated variation. Factors were examined in terms of variance explained and biological interpretability and compared to scVI representations.

Deep Generative Model: scVI

Finally, we applied scVI to the combined thymus dataset to learn a nonlinear, probabilistic latent representation. By operating directly on raw count data and explicitly modeling technical noise, scVI provides a flexible framework for capturing subtle disease-associated transcriptional programs and continuous thymic cell-state trajectories.

V. EXPERIMENTS, RESULTS AND DISCUSSION

Experimental Setup

We conducted a benchmarking study to evaluate representation-learning methods for single-cell RNA-sequencing (scRNA-seq) data in myasthenia gravis (MG) across two biological contexts: peripheral blood mononuclear cells (PBMCs) and thymic tissue. The two datasets were analyzed independently, reflecting their distinct biological structure and complexity, while using a consistent evaluation framework.

For the PBMC dataset, we compared a linear baseline (Principal Component Analysis, PCA) with a deep generative model (scVI). PBMCs are relatively homogeneous and serve as a controlled setting for evaluating whether nonlinear generative modeling provides benefits beyond linear dimensionality reduction.

For the thymus dataset, which exhibits substantially higher cellular and structural heterogeneity, we evaluated a broader set of models: PCA, MOFA+ (Bayesian factor model), and scVI. This allowed us to assess how increasingly expressive integration methods perform in a complex tissue environment.

Model Training and Hyperparameters: PBMC dataset

For PCA, we retained the top 50 principal components, a common choice in scRNA-seq analysis that balances variance preservation with noise reduction. Empirically, the leading 30–50 components typically capture the dominant biological structure in immune datasets, while additional components largely reflect technical variation. Retaining 50 components therefore ensures that relevant signal is preserved for downstream clustering and evaluation without overfitting to noise.

The scVI model was trained using the default optimization settings provided by the scvi-tools framework, including the default learning rate and batch size. These defaults are designed to provide stable optimization across a wide range of scRNA-seq datasets. Training was performed for a maximum of 150 epochs, a value chosen based on the size of the quality-controlled PBMC dataset (approximately 33,000 cells) to allow sufficient convergence without unnecessary overtraining. Validation loss was monitored every 10 epochs during training. No batch covariate was specified because sample-level identifiers were perfectly confounded with disease status in the PBMC dataset, such that each sample contained cells from only a single condition. Under this design, treating sample or condition as a batch would necessarily remove biologically meaningful variation, making batch correction statistically inappropriate. Optimization was performed using the Adam optimizer, which is standard for deep generative models.

Regularization was implicitly enforced through the KL-divergence term of the variational objective, which constrains the latent space and discourages overly complex representations. Where applicable, default random initialization settings were used, and analyses were conducted within a consistent computational environment to support reproducibility.

Model Training and Hyperparameters: Thymus dataset

For the thymus dataset, PCA was used as a linear baseline by retaining the top 50 principal components. In addition to PCA, one more classical integration baseline was evaluated. MOFA+ was used as a Bayesian factor model to learn shared latent factors across thymic samples, with default configuration parameters. These methods do not rely on stochastic gradient optimization and are therefore less sensitive to hyperparameter tuning.

scVI was applied to the thymus dataset using the same model architecture as in the PBMC analysis, with a latent dimensionality of 10 and default optimization settings provided by scvi-tools. Because the thymus dataset contained approximately 128,000 quality-controlled cells and exhibited increased biological and technical heterogeneity, training was performed for up to 500 epochs to allow sufficient convergence. Regularization was enforced through the KL-divergence term of the variational objective, and validation loss was monitored during training to ensure stable convergence.

Metrics:

Model performance on the PBMC dataset was evaluated using cell-type–based latent-space metrics, selected to assess whether learned representations preserve biologically meaningful immune structure. The silhouette score by cell type was used to quantify within–cell-type cohesion and between–cell-type separation, directly reflecting how well immune populations are resolved in the latent space. The silhouette score by condition (Healthy vs MG) was included as a diagnostic to assess whether disease status disproportionately structures the representation, which would indicate potential confounding. To measure agreement between latent-space clustering and known cell-type annotations, we computed the Adjusted Rand Index (ARI) and Normalized Mutual Information (NMI), which provide complementary assessments of clustering consistency and information overlap. All metrics were computed on the learned latent representations to enable direct comparison between PCA and scVI.

Model performance on the thymus dataset was evaluated using similar metrics. To assess preservation of thymic cell identity, we computed the silhouette score by cell type, which measures within–cell-type cohesion and between–cell-type separation in the latent space, as well as Adjusted Rand Index (ARI) and Normalized Mutual Information (NMI), which quantify agreement between latent-space clustering and reference cell-type annotations. To evaluate the integrity of the integrated neighborhood structure, we report graph connectivity, which measures whether cells of the same type remain connected in the kNN graph after integration.

Results: PBMC

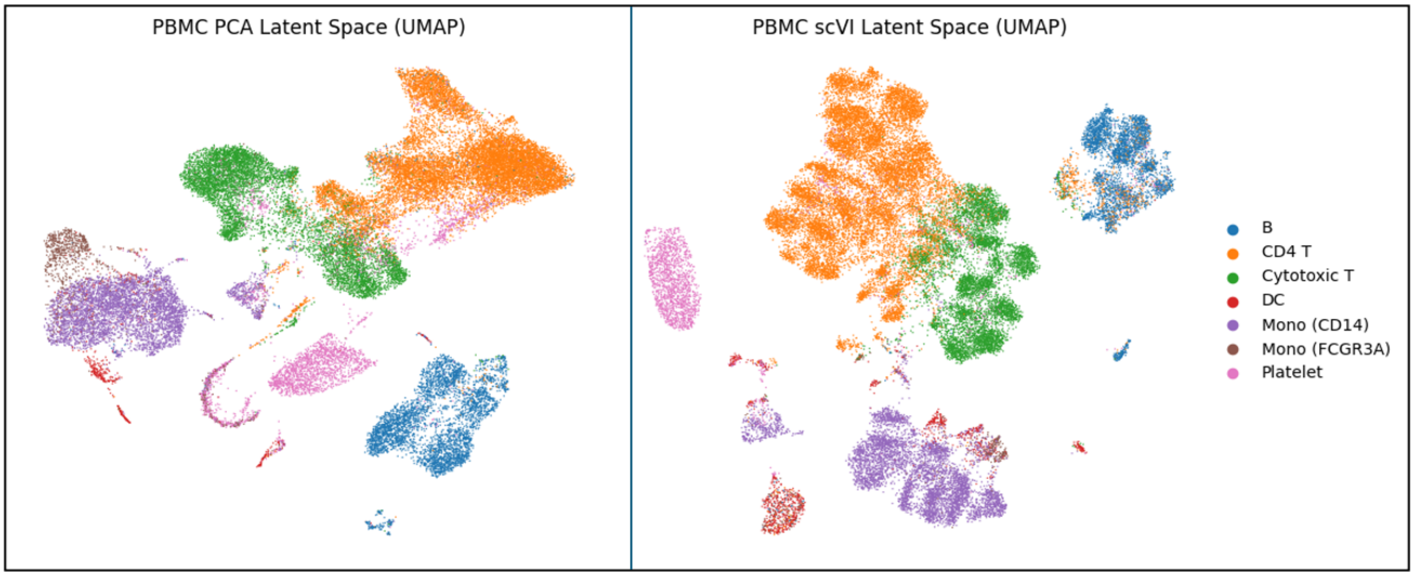

Figure 1. Comparison of PCA and scVI latent representations for PBMCs.
UMAP visualizations of the PBMC dataset computed from (left) the PCA latent space and (right) the scVI latent space, with cells colored by reference cell-type annotations.

Metric	**PCA**	**scVI**
Silhouette score (cell type)	0.266	0.134
Silhouette score (condition)	−0.015	0.024
Adjusted Rand Index (ARI)	0.773	0.547
Normalized Mutual Information (NMI)	0.723	0.629

Table 1. PBMC representation learning performance for PCA and scVI.
Quantitative evaluation of latent-space representations on the PBMC dataset using cell-type–based metrics. Silhouette scores by cell type measure within–cell-type cohesion and between–cell-type separation, while silhouette scores by condition (Healthy vs MG) are reported as a diagnostic of disease-driven structure. Adjusted Rand Index (ARI) and Normalized Mutual Information (NMI) quantify agreement between latent-space clustering and reference cell-type annotations. Neighborhood overlap reflects the degree of local geometric similarity between PCA and scVI representations.

In the PBMC dataset, PCA shows stronger agreement with existing cell-type annotations than scVI, as reflected by higher ARI, NMI, and cell-type silhouette scores. This is not surprising: PBMC cell types are largely defined by strong marker genes and are often close to linearly separable, making linear methods particularly effective at reproducing curated annotation schemes. As a result, PCA performs better when evaluation focuses on alignment with predefined PBMC labels.

By contrast, scVI learns a lower-dimensional, probabilistic latent space that explicitly models technical noise and applies regularization. This produces more compact and visually coherent clusters but also smooths fine-grained subtype boundaries, especially among transcriptionally similar populations such as CD14 and FCGR3A monocytes. Consequently, scVI shows reduced agreement with fine-grained PBMC annotations, despite preserving all major immune populations and maintaining a biologically plausible overall structure.

Silhouette scores by condition remain near zero for both methods, indicating that disease status does not dominate the learned representations. The low neighborhood overlap between PCA and scVI further confirms that scVI substantially reshapes the latent geometry rather than acting as a minor refinement of PCA. Overall, these results suggest that PCA is sufficient for PBMC cell-type separation, while scVI emphasizes smoother latent structure at the cost of fine-grained annotation fidelity.

Results: Thymus

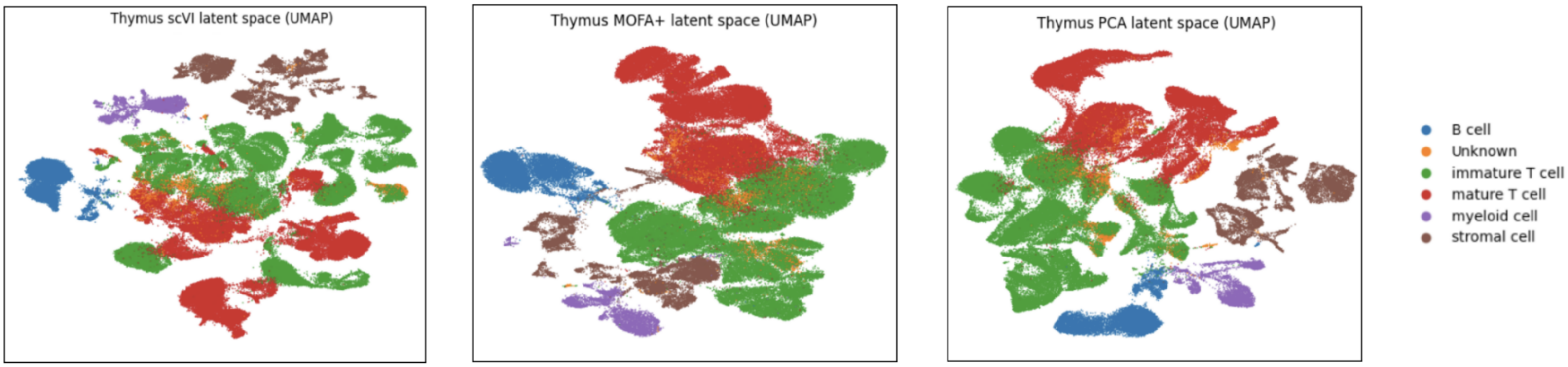

Figure 2. Comparison of latent representations learned from thymus scRNA-seq data.
UMAP visualizations of latent spaces obtained using scVI (left), MOFA+ (center), and PCA (right), with cells colored by annotated cell type. PCA and MOFA+ produce clearer global separation between major thymic cell populations, including immature and mature T cells, stromal cells, myeloid cells, and B cells. In contrast, scVI yields more overlapping clusters, reflecting reduced global cell-type separation despite preservation of local neighborhood structure. These qualitative differences are consistent with quantitative clustering and connectivity metrics reported in Table X.

Metric	**PCA**	**MOFA+**	**scVI**
Graph connectivity	0.895	0.940	0.000
Silhouette score (cell type)	0.577	0.537	0.534
Adjusted Rand Index (cell type)	0.143	0.133	0.008
Normalized Mutual Information (cell type)	0.495	0.471	0.191

Table 2. Quantitative evaluation of latent representations on thymus scRNA-seq data.
This table reports graph connectivity, silhouette score, Adjusted Rand Index (ARI), and Normalized Mutual Information (NMI) for PCA, MOFA+, and scVI. PCA and MOFA+ achieve higher clustering agreement with annotated cell types, while scVI exhibits substantially reduced global cell-type separation despite comparable silhouette scores, indicating differences in how each method captures thymic cellular structure.

Quantitative evaluation of latent representations learned from thymus scRNA-seq data reveals clear differences in how PCA, MOFA+, and scVI capture thymic cellular structure. PCA and MOFA+ achieve substantially higher graph connectivity scores (0.895 and 0.940, respectively), indicating strong preservation of local neighborhood relationships in the low-dimensional space. In contrast, scVI exhibits near-zero graph connectivity, suggesting limited preservation of global neighborhood structure under this metric. This behavior reflects differences in modeling objectives rather than a failure of scVI. Graph connectivity favors methods that preserve discrete, cluster-based structure, which aligns well with PCA and MOFA+, whereas scVI learns a smooth, probabilistic latent manifold optimized for denoising and continuous transcriptional variation. As a result, cells from the same annotated thymic population may be distributed across multiple regions of the scVI latent space, reducing graph connectivity despite preserving biologically meaningful local structure.

PCA attains the highest clustering agreement with annotated cell types, as reflected by the largest silhouette score (0.577), Adjusted Rand Index (0.143), and Normalized Mutual Information (0.495). MOFA+ performs comparably but slightly worse across these metrics, indicating moderate preservation of thymic cell-type organization. By comparison, scVI yields markedly lower ARI (0.008) and NMI (0.191), despite achieving silhouette scores similar to the other methods, suggesting that while local separation between some cell populations is maintained, global cell-type structure is less coherently captured.

Together, these results indicate that linear and factor-based methods better preserve discrete thymic cell-type organization in this dataset, whereas scVI prioritizes alternative aspects of transcriptional variation that do not align as closely with annotated cell-type boundaries. These quantitative findings are consistent with the qualitative differences observed in UMAP visualizations of the respective latent spaces.

Across both PBMC and thymus datasets, scVI consistently underperforms PCA and MOFA+ on metrics that quantify agreement with discrete, curated cell-type annotations, despite its greater model complexity. This pattern reflects a mismatch between the evaluation criteria and the modeling objective of scVI rather than an intrinsic weakness of the generative framework. In PBMCs, where immune populations are strongly marker-defined and close to linearly separable, linear methods such as PCA are particularly well suited to reproducing existing annotation schemes, leading to higher ARI, NMI, and cell-type silhouette scores. Similarly, in thymic tissue, PCA and MOFA+ better preserve discrete cell-type organization as measured by graph connectivity and clustering agreement.

By contrast, scVI is explicitly designed to learn a smooth, denoised latent manifold that captures continuous transcriptional variation while accounting for technical noise. In both datasets, this leads to compression of fine-grained subtype boundaries and redistribution of cells across the latent space, which reduces alignment with label-based metrics that assume well-separated clusters. While this behavior results in lower scores under clustering-focused evaluations, it may be advantageous for modeling gradual cell-state transitions or subtle biological gradients not well captured by discrete labels. Thus, the observed performance differences highlight the importance of aligning model choice with analysis goals: in settings where faithful reproduction of curated cell types is the primary objective, simpler linear or factor-based methods may be sufficient, whereas generative models such as scVI may be better suited for tasks emphasizing denoising, trajectory inference, or latent structure discovery beyond predefined annotations.

VI. CONCLUSION AND FUTURE WORK

In this project, we benchmarked representation-learning methods for single-cell RNA sequencing data in myasthenia gravis (MG) across two biologically distinct contexts: peripheral blood mononuclear cells (PBMCs) and thymic tissue. Motivated by the growing use of increasingly expressive models in scRNA-seq analysis, we compared a classical linear baseline (PCA), a probabilistic factor model (MOFA+), and a deep generative model (scVI) to assess how modeling assumptions interact with data complexity and evaluation criteria. Across both datasets, PCA and MOFA+ consistently achieved stronger agreement with curated cell-type annotations, as measured by silhouette scores, ARI, NMI, and graph connectivity. These results reflect the strongly marker-driven and often discrete structure of immune cell identities in PBMCs, as well as the relatively well-defined organization of many thymic populations. In contrast, scVI learned smoother, probabilistic latent representations that preserved biologically plausible local structure but showed reduced alignment with label-based clustering metrics. Together, these findings underscore that model complexity alone does not guarantee superior performance and that representation quality must be interpreted relative to the biological structure of the data and the analytical objectives.

A key limitation of this study is the absence of spatial transcriptomics analysis, which was part of the original project design. We initially aimed to integrate MG thymus Visium spatial transcriptomics data and apply DestVI to jointly model single-cell and spatial gene expression. Visium enables spatial localization of transcriptional programs within tissue architecture, while DestVI allows inference of cell-type proportions and continuous cell states at each spatial location, offering direct insight into disease-associated immune niches. However, access to raw spatial count data for MG patients is restricted due to privacy considerations, precluding the use of DestVI. If such data were available, future work would focus on spatial integration to map MG-associated thymic remodeling in situ and to evaluate whether generative models better capture spatially continuous biological processes than discrete clustering metrics suggest. More broadly, extending evaluation beyond clustering-based metrics to include trajectories, lineage relationships, longitudinal data, or immune receptor sequencing would provide a more comprehensive assessment of when generative models offer advantages over simpler baselines in autoimmune disease settings..

VII. REFERENCES AND CONTRIBUTIONS

MG research

Jin, X., Wang, S., Zhang, Y., Li, Y., Zhao, Y., Liu, Y., … Zhang, X. (2021). Single-cell RNA sequencing reveals transcriptional heterogeneity and immune subtypes associated with disease activity in human myasthenia gravis. *Cell Discovery, 7*(1), Article 83. https://doi.org/10.1038/s41421-021-00314-w

Yasumizu, Y., Kondo, T., Ishikawa, Y., Kato, S., Kobayashi, T., Takahashi, K., … Takahashi, Y. (2022). Myasthenia gravis–specific aberrant neuromuscular gene expression by medullary thymic epithelial cells in thymoma. *Nature Communications, 13*(1), Article 4119. https://doi.org/10.1038/s41467-022-31951-8

Yasumizu, Y., Kondo, T., Ishikawa, Y., Kato, S., Kobayashi, T., Takahashi, K., … Takahashi, Y. (2024). Spatial transcriptomics elucidates a medullary niche supporting germinal center responses in myasthenia gravis–associated thymoma. *Cell Reports, 43*(2), 114677. https://doi.org/10.1016/j.celrep.2024.114677

Datasets

Okuzono, K., Suzuki, R., Matsumoto, T., *et al.* (2024). *Single-cell transcriptomic profiling of peripheral blood immune cells in myasthenia gravis* (GSE227835) [Dataset]. Gene Expression Omnibus. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE227835

Broad Institute Single Cell Portal. (2025). *Thymoma and PBMC of myasthenia gravis patients* (Dataset SCP1532). Retrieved from https://singlecell.broadinstitute.org/single_cell/study/SCP1532/thymoma-and-pbmc-of-myasthenia-gravis-patients

CZ CELLxGENE Discover. (2024). *Annotated fetal and paediatric human thymus single-cell RNA-seq atlas* [Data set]. CZ CELLxGENE Discover. https://cellxgene.cziscience.com/collections/fc19ae6c-d7c1-4dce-b703-62c5d52061b4 Cellxgene Data Portal

Authors

First Author – María Dolores (Lola) Navarro Maturana, mn3312
Second Author – Daris (Dasha) Lykova, dl3415
Third Author – Yitian (Evan) Tan, yt2916