Name: Shaista Madad
Title: Implementation of a Gaussian Process Latent Variable Model (GPLVM) for Non-linear Dimensionality Reduction of Single Cell RNA-seq Datasets
Supervisors: Dr Sarah Teichmann and Emma Dann

Summary
Single cell RNA sequencing (scRNA-seq) data is noisy and high dimensional which makes it difficult to find biologically meaningful patterns in the data. Dimensionality reduction is an important step before downstream analysis to find a representation of the data in a lower dimensional manifold with an enriched signal. Gaussian Process Latent Variable Model (GPLVM) learns a set of latent variables to explain the high dimensional data while preserving smooth biological patterns in the data, such as the cell cycle effect, which is not possible with linear dimensionality reduction methods such as Principal Component Analysis (PCA). We tested a state of the art implementation of a GPLVM on a range of scRNA-seq datasets to test its performance in comparison with standard RNA-seq workflows.
As a first step, we assessed whether the GPLVM recapitulates biological information captured by PCA. Next, we explored technical and biological effects not captured by PCA, focusing on the cell cycle effect. Our results did not find an improved performance of the GPLVM model to model the cell cycle effect in the datasets.

Introduction
Advances in single cell RNA sequencing (scRNA-seq) technologies provide exciting opportunities to derive useful insights about important biological processes in health and disease. However, scRNA-seq data is often noisy due to low capture efficiency and sequencing depth, and high dropout events lead to sparsity in the data (Sun *et al.*, 2019). Dimensionality reduction is an important data processing step to transform the original high dimensional matrix into a lower dimensional space with enriched signal, thus allowing downstream analysis such as clustering, trajectory analysis, 2D embeddings for visualisation (Luecken and Theis, 2019). Furthermore, dimensionality reduction also enables modelling algorithms to process single cell data in a computationally efficient manner (Xiang *et al.*, 2021).

Linear dimensionality reduction methods, such as principal component analysis (PCA) , have traditionally been preferred due to ease of computation (Lawrence and Hyvärinen, 2005). However, linear methods are limited in their ability to capture complex biological phenomena, e.g.,developmental processes, cell cycle (Verma and Engelhardt, 2020) as they assume a linear relationship between the high to low dimensional mapping. To address this challenge,

deep learning based non-linear dimensionality reduction methods such as ScVI (Lopez *et al.*, 2018)) and models based on variational autoencoders (Wang and Gu, 2018; Eraslan *et al.*, 2019; Grønbech *et al.*, 2020; Svensson *et al.*, 2020) are getting popular. However, one challenge to these neural network based models is their lack of interpretability. Gaussian Process Latent Variable Model (GPLVM) provides an advantage as easily interpretable models which learn a set of latent variables to model the high dimensional data by modelling the genes as a set of functions drawn from Gaussian distribution. GPLVMs have recently been applied as interpretable non-linear models for single-cell genomics (Lönnberg *et al.*, 2017; Verma and Engelhardt, 2020; Kumasaka *et al.*, 2021). These models allow us to incorporate prior knowledge about gene expression patterns (e.g. cycling patterns for proliferation genes, information about time points). However, one challenge to the GPLVMs is the scalability to large datasets.

Here, building upon the work by (Kumasaka *et al.*, 2021), we applied a state-of-the-art scalable GPLVM implementation (Lachland, Ravuri and Lawrence, unpublished) using Stochastic Variational Inference (Hoffman *et al.*, 2013) for dimensionality reduction of large scale single cell datasets. We assess the capabilities of this model compared to standard workflows for single-cell RNA-seq analysis in a range of scRNA-seq datasets.

As a first step, we assessed whether the scalable GPLVM recapitulates biological information captured by PCA. Next, we explored technical and biological effects not captured by PCA.

Results

First we investigated whether the scalable implementation of GPLVM, as a minimum, is able to capture the biological information captured by PCA **(Figure 1) i**n six publicly available datasets **(Table 1 in Methods).** For the GPLVM model, we compared two training regimes: with initialisation of the latent variables (LVs) by 7 PCA components (gplvm-pca) versus randomly generated values (gplvm-random). PCA initialization is commonly used as an optimization step for GPLVM training on large datasets (Kumasaka *et al.*, 2021) . As expected, the PCA-initialised LVs showed higher correlation with PCA components than randomly initialised LVs **(Figure 1A)**. Notably, in the majority of the tested datasets, the median Pearson correlation between PCA-initialised LVs and principal components was over 0.9, indicating that during training the values for LVs did not shift significantly from the initialization values. We used four metrics to compare the performance of PCA-initialised versus randomly initialised LVs (Methods). Visualisation of UMAP embeddings from all the

dimensionality reductions (standard PCA, gplvm-pca, gplvm-random) showed comparable levels of separation by cell types (**Figure 1B and Supplementary Figure S1**). Next, we used two common metrics of clustering agreement (adjusted Rand index (ARI) and the normalised mutual information (NMI) (Wu and Wu, 2020)) to evaluate the agreement between clustering based on embeddings derived from the three dimensionality reductions and clustering based on ground truth cell type labels. NMI is better suited for unbalanced and small clusters. Although PCA embeddings showed highest scores for both metrics across the six datasets **(Figure 1C)**, the ARI and NMI scores for GPLVM clusters were comparable. Since the ground truth cell type labels for all datasets were derived based on clustering in the PCA space, we expected higher scores with PCA embeddings. Notably, random GPLVM initialization did not lead to a severe drop in clustering agreement. Furthermore, we computed k-nearest neighbour purity (KNN) scores (see Methods) to assess the performance of our scalable GPLVM model in clustering cells belonging to a similar biological origin together. We found that the mean KNN purity scores of standard PCA were not significantly different from the GPLVM derived embeddings **(Figure 1D)**. Our findings indicate that the scalable implementation of the GPLVM model manages to capture biological information to a similar degree as PCA. In addition, we observe limited gain in the ability to capture cell type diversity using a PCA initialised GPLVM model, suggesting this step can be omitted to further reduce computation time during preprocessing.

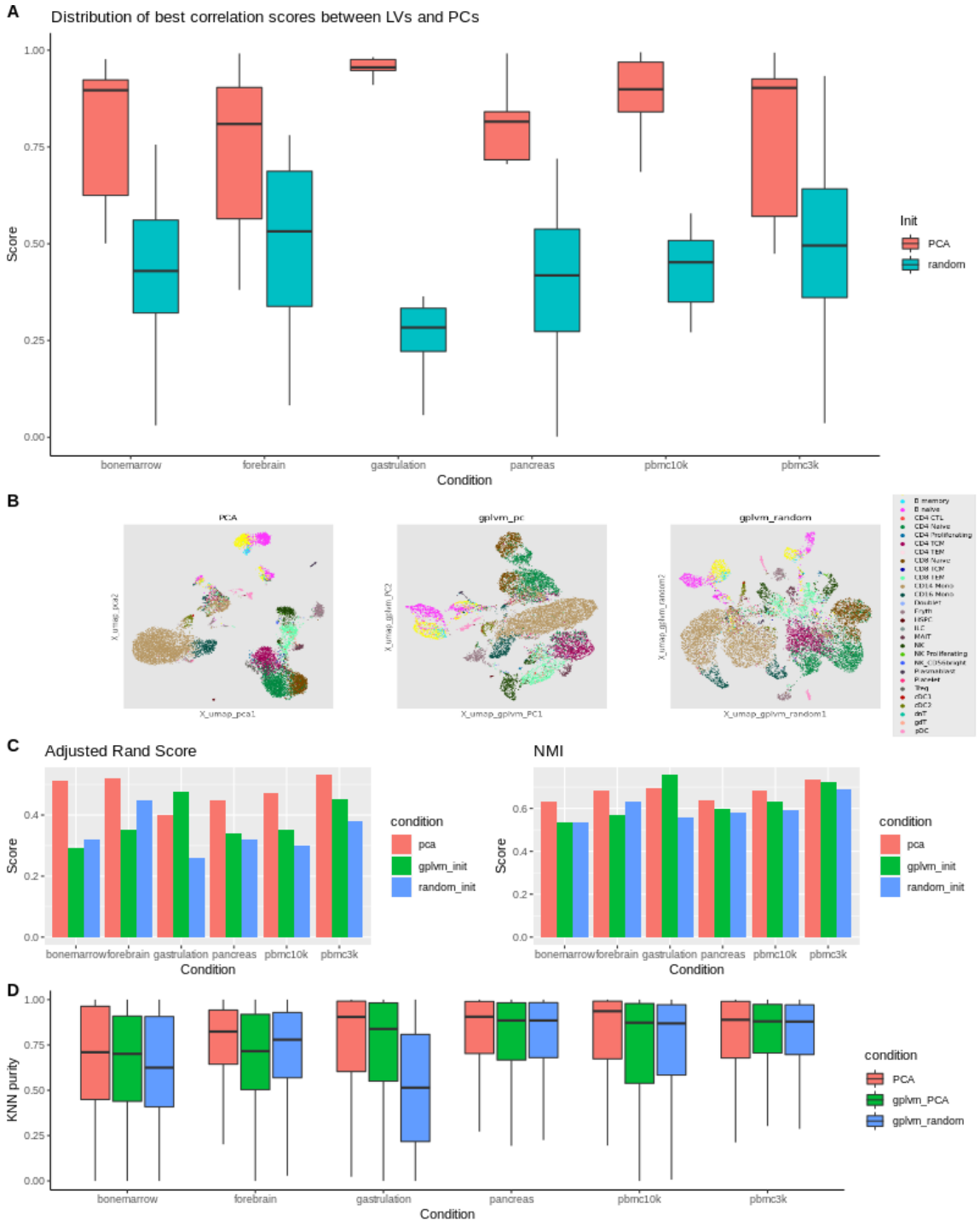

Figure 1

A) Comparison of the distribution of highest Pearson correlation score between each of the 7 latent variables (LVs) from GPLVM model and top 10 PCA components. Each box is a distribution of seven scores, one for each of the LVs. **B)** UMAP embeddings for a PBMC (Peripheral Blood Mononuclear Cells) dataset based on similarity in PCA (left) and similarity in the GPLVM reduced dimensionality (PC-initialised GPLVM: middle and randomly initialised GPLVM: right) **C)** ARI and NMI scores between ground truth cell type labels and cluster assignments from the three dimensionality reduction methods **D)** KNN purity scores for clustering based on the three dimensionality reduction methods

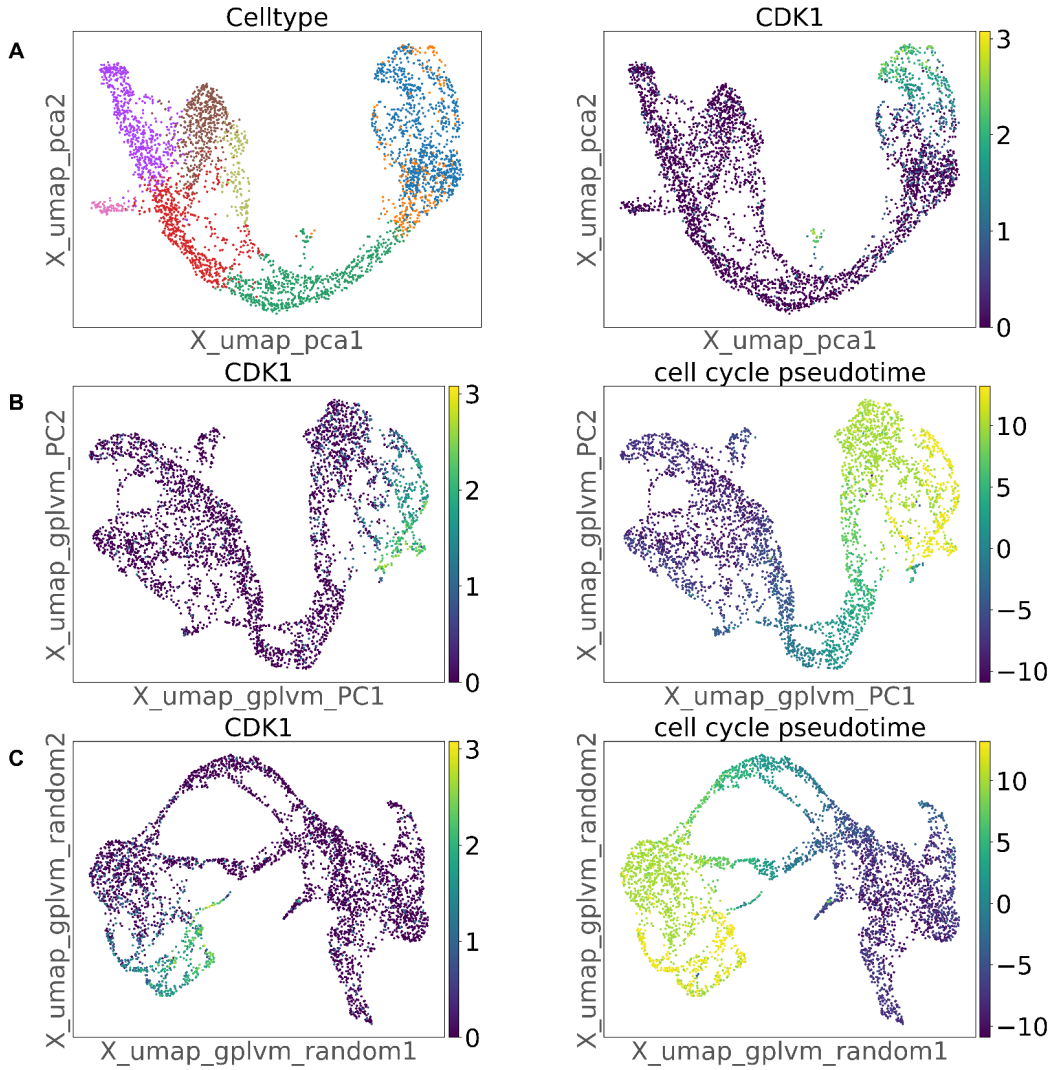

Figure 2
A) UMAP embeddings based on standard PCA coloured by celltype and CDK1 expression.UMAP embeddings based on GPLVM with PCA initialised LVs **(B)** and GPLVM with randomly initialised LVs **(C)** coloured by cell cycle pseudotime (right) and CDK1(left) for pancreas dataset. There is still clear separation along cell cycle stages.

Having shown that our GPLVM can capture cell type variation similarly to PCA on scRNA-seq datasets, we next evaluated the advantages of this model. One of the main advantages is that we can assign informative priors on some of the latent variables to better disentangle nuisance effects in gene expression data. For example, we can disentangle the cell cycle by incorporating a latent variable with a periodic kernel, thus forcing that variable to capture the main cyclic effect in the dataset (hereafter termed “cell cycle pseudotime”) which is not captured by PCA **(Figure 2A).**

In order to test whether the periodic kernel really captures the cell cycle, we tested whether excluding the cell cycle when deriving UMAP embeddings from neighbourhood graphs based on LVs loadings shows reduced separation between cells in different stages of cell cycle. For this analysis we started by using the pancreas dataset (Bastidas-Ponce *et al.*, 2019) in the scvelo package (Bergen *et al.*, 2020), where we observed that the clustering and embedding based on PCA still retains separation between cells driven by proliferation **(Figure 2A)** and this trend was observed across all six datasets. We observed separation between cell types driven by proliferation in the GPLVM embeddings **(Figure 2B and 2C).**

We defined another metric we have termed as cell cycle proliferation purity (CCP) to quantify the mixing of cells in different stages of cell cycle. As we did not observe a clear periodic pattern of gene expression against cell cycle pseudotime for a number of cell cycle associated genes **(Figure 3),** we chose a few cell cycle genes as a proxy for modelling the periodicity of cell cycle **(Figure 3).** The gene expression of these genes was categorised as low or high based on whether cells expressed zero or non-zero values respectively. For each cell, we then calculated the number of cells in its neighbourhood (100 cells) on the KNN graph expressing the same gene expression category as the query cell. We reasoned that the lower the CCP score, the better the performance of our GPLVM model in regressing cell cycle effects (i.e. better mixing of cells from different cell cycle stages). Taking into account the sparsity of scRNA-seq expression data, we subsetted to cells with non-zero gene expression when comparing the distribution of CCP scores and still observed very high scores for CCP for certain datasets (including the pancreas dataset) and little difference across the three dimensionality reduction methods **(Figure 4).**

We also compared the distribution of cell type KNN purity and CCP scores **(Figure 5).** The best performing dimensionality reduction method will have a higher proportion of cells with high KNN purity and low CCP (quadrant 2 in **Figure 5**). The standard PCA performed best for many cell cycle genes including CDK1 **(Figure 5A)** and CENPF **(Figure 5B)**.

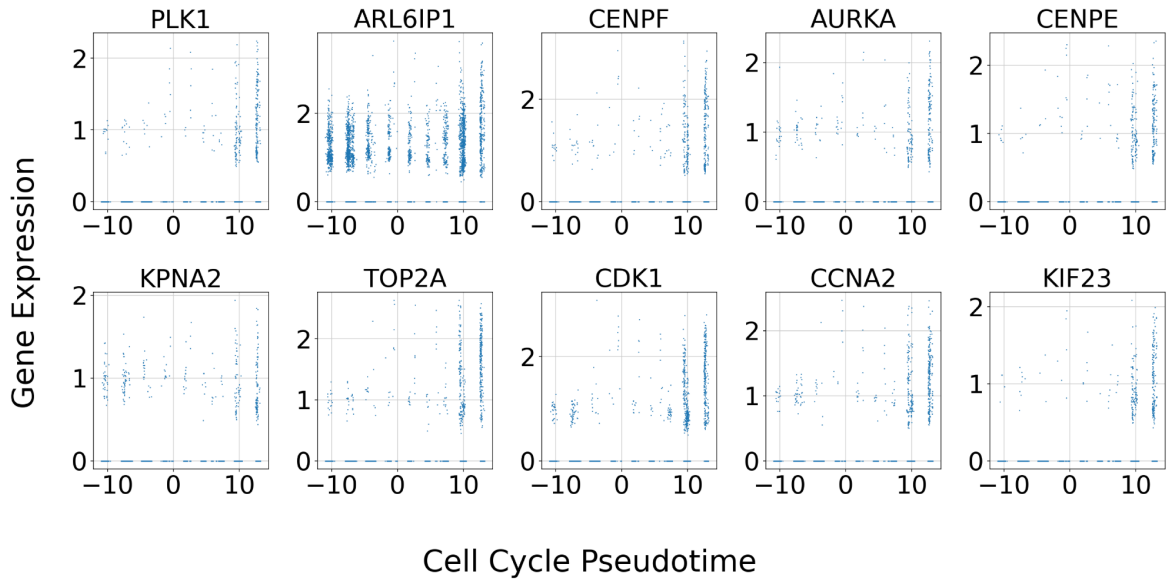

Figure 3
Scatterplots with cell cycle pseudotime on x-axis and gene expression data for M-Phase (top) and G-Phase (bottom) genes for pancreas dataset.

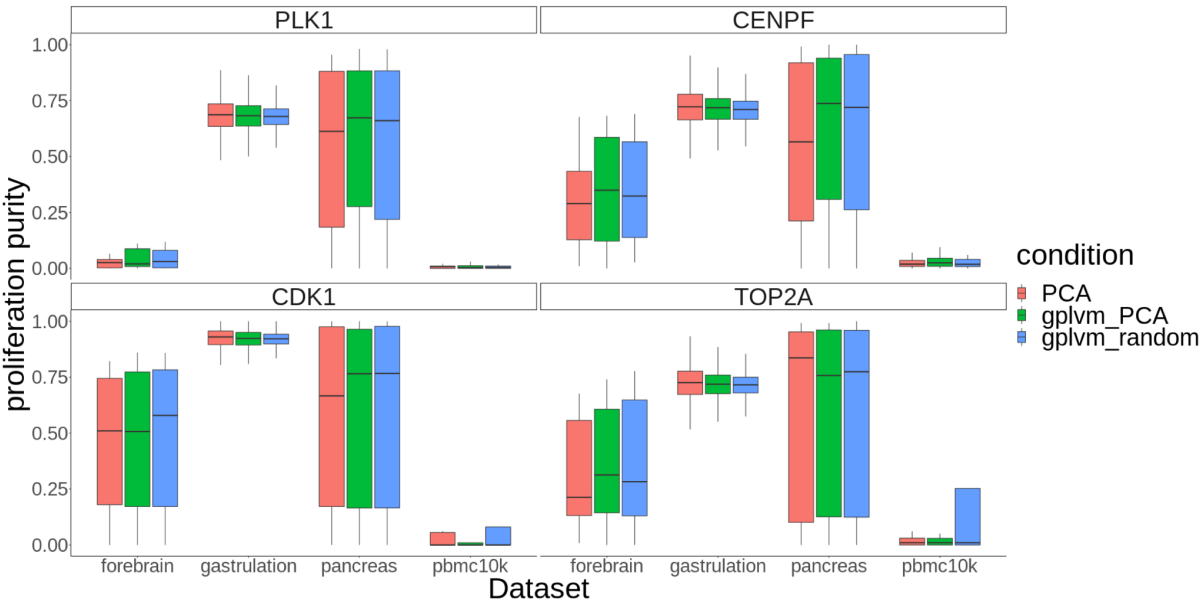

Figure 4
Distribution of cell cycle proliferation purity (CCP) scores for M-Phase (top) and G-Phase (bottom) genes. For all datasets, the scores are subsetted for cells with non-zero gene expression for the given gene

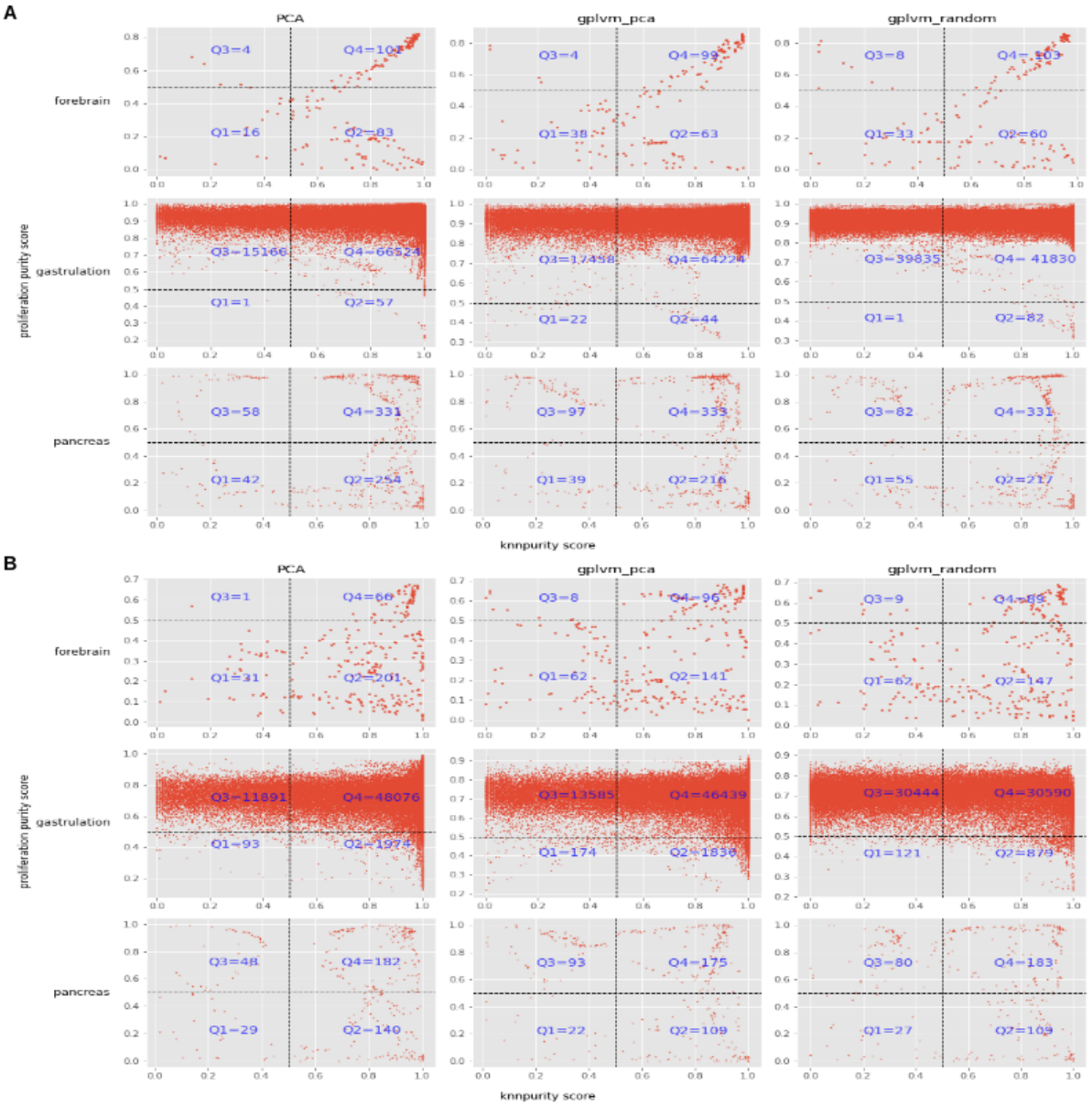

Figure 5
Scatterplots of KNN purity scores on x-axis and cell cycle proliferation purity (CCP) scores on y-axis for genes CDK1 (A) and CENPF (B) for three datasets: forebrain, gastrulation and pancreas for PCA (left column), GPLVM with PCA initialisation (middle) and GPLVM with random initialisation (right column). The best performing method should have a higher proportion of cells in quadrant 2: high KNN purity and low CCP scores.

Taken together these results indicate that the GPLVM model with the periodic kernel is no better than PCA at disentangling the cell cycle effect.

Discussion

In this analysis, we tested whether our scalable implementation of a GPLVM model based on Stochastic Variational Inference is able to capture smooth and continuous non-linear trends in data which the standard PCA is not able to do. We find the performance of our model does not supersede PCA in capturing the cell cycle effect in a range of datasets. Our original model (Kumasaka *et al.*, 2021) explicitly takes into account the donor variation as well as other known confounding effects (such as technical variation due to batch effect) as additional random effect terms in the model, whereas the current scalable implementation of the GPLVM only used the sparse GP to model the genes with a periodic kernel. Hence, it is possible that the addition of the random effect terms, and optimisation of the model by initialising the latent variables with customized and more informative PCs will lead to improved performance. It is possible that the periodic kernel is capturing another biological phenomena distinct from the cell cycle. Furthermore, the cell cycle may be confounded with other biological processes, e.g., differentiation in the pancreas dataset (Bastidas-Ponce *et al.*, 2019). It is also important to note that the original GPLVM model (Kumasaka *et al.*, 2021) was used on a very homogeneous dataset. Therefore, our analysis highlights the limitations of the scalable GPLVM model in disentangling the cell cycle effect in datasets with multiple cell types.

Code and Data Availability

The six datasets **(Table 1**) used in this analysis are available on public repositories and data portals https://scvelo.readthedocs.io/api/. The RNA counts matrices for the PBMC10K and PBMC3K RNA-seq datasets were downloaded from the 10X website. The bone marrow, gastrulation, forebrain and pancreas datasets were available as anndata objects within the scvelo package (Bergen *et al.*, 2020) in Python.
Code available at: https://github.com/shaistamadad/GPLVM_Shaista

Table 1: Summary of the Datasets. HV= Highly Variable

Dataset	No of Cells	No of HV Genes	Time to Train Model (min)
Bone Marrow (human) *(scvelo.datasets.bonemarrow — scVelo 0.2.5.dev6+g1805ab4.d20210829 documentation*, no date)	5780	2030	12
Gastrulation (mouse) (Pijuan-Sala *et al.*, 2019)	89267	1488	9
Forebrain (human) (La Manno *et al.*, 2018)	1720	2454	13
Pancreas (mouse) (Bastidas-Ponce *et al.*, 2019)	3696	1939	11
PBMC10K (human) *(pbmc_unsorted_10k -Datasets -Single Cell Multiome ATAC + Gene Exp. -Official 10x Genomics Support*, no date)	12016	4650	25
PBMC3K (human) *(pbmc_unsorted_3k -Datasets -Single Cell Multiome ATAC + Gene Exp. -Official 10x Genomics Support*, no date)	3012	4134	22

Methods

Overview of GPLVM Model

Let the gene expression data be represented by Y, where $Y \in R^{N \times D}$ where N represents the number of cells and D the total number of genes. Our GPLVM fast version model tries to capture the most significant factors of variation in a dataset by learning a Q dimensional latent encoding for each cell in the dataset where Q<D, thus providing dimensionality reduction.

We assumed the gene expression vector $Y_j = (Y_{ij}\ ;\ _i = 1, ..., N)^T$ for gene j across N cells is independently drawn from:

$$Y_j \sim N(\alpha_j, \sigma_j^2 \Omega)$$
$$\alpha_j \sim N(0, \sigma_j^2 K_\theta K_X)$$

Where α_j is a baseline gaussian process governed by 2 different kernel matrices, periodic kernel K_θ to capture the cell cycle state, and K_X for the latent biological and technical effects, where K_X is a smooth RBF kernel.

Pre-processing of gene expression data

Gene expression matrices were pre-processed following the workflow implemented in the python package Scanpy (Wolf, Angerer and Theis, 2018). Briefly, raw counts for each cell were normalised by total counts over all genes so that every cell has the same total count after normalisation, thus correcting for variable sequencing depth. The matrix was then log normalised and the highly variable genes were selected for training the GPLVM model. Finally, PCA was run on the processed gene expression data.

Seurat 4.0.1 was used for all analyses of scRNA-seq assays in the PBMC10K and PBMC3K datasets and is available on CRAN (https://cloud.r-project.org/package=Seurat). The cell type labels were obtained for the PBMC10K and PBMC3K datasets by finding anchors between the processed scRNA-seq data and a high quality CITE-seq dataset (Hao *et al.*, 2021) using the *FindTransferAnchors* function in Seurat. The cell type labels for the remaining datasets were already available in the metadata of the anndata object downloaded from the scvelo package with the function: *scvelo.datasets.[nameofdataset].()*

Correlation Scores

The Pearson correlation between the latent variables and top 10 PCA loadings were calculated using the corrcoef function in numpy (Harris *et al.*, 2020). For each latent variable, the highest correlation score was chosen.

Qualitative comparison with 2D embedding

A neighbourhood graph was computed based on euclidean distances between cells in the PCA or GPLVM space using the neighbours function in scanpy. A UMAP embedding based on the neighbourhood graph was created using the *sc.tl.umap()* function in scanpy.

Adjusted Rand Index and Normalised Mutual Information

We used the sklearn implementation of ARI. Cells were clustered using the leiden algorithm (*scanpy.tl.leiden*) based on the neighbourhood graphs constructed in PCA or GPLVM space.

KNN Purity Scores

For each cell, the purity metric was calculated as the average fraction of 100 nearest neighbours belonging to the same cell type as the given cell. The neighbourhood graphs were derived from the PCA or GPLVM space.

Cell Cycle Analysis

We conducted a literature search to find cell cycle genes(*Cyclebase 3.0*, no date, *Cell-Cycle Scoring and Regression*, no date; Scialdone *et al.*, 2015) . A total of 24 cell cycle genes were identified. Out of these genes, 8 were identified as growth phase (G-Phase) and 10 as mitosis phase (M-Phase) genes.

References

Bastidas-Ponce, A. *et al.* (2019) 'Comprehensive single cell mRNA profiling reveals a detailed roadmap for pancreatic endocrinogenesis', *Development* , 146(12). doi:10.1242/dev.173849.

Bergen, V. *et al.* (2020) 'Generalizing RNA velocity to transient cell states through dynamical modeling', *Nature biotechnology*, 38(12), pp. 1408–1414.

Cell-Cycle Scoring and Regression (no date). Available at: https://satijalab.org/seurat/archive/v3.1/cell_cycle_vignette.html (Accessed: 13 January 2022).

Cyclebase 3.0 (no date). Available at: https://cyclebase.org/Advanced%20Search (Accessed: 13 January 2022).

Eraslan, G. *et al.* (2019) 'Single-cell RNA-seq denoising using a deep count autoencoder', *Nature communications*, 10(1), p. 390.

Grønbech, C.H. *et al.* (2020) 'scVAE: variational auto-encoders for single-cell gene expression data', *Bioinformatics* , 36(16), pp. 4415–4422.

Hao, Y. *et al.* (2021) 'Integrated analysis of multimodal single-cell data', *Cell*, 184(13), pp. 3573–3587.e29.

Harris, C.R. *et al.* (2020) 'Array programming with NumPy', *Nature*, 585(7825), pp. 357–362.

Hoffman, M.D. *et al.* (2013) 'Stochastic variational inference', *Journal of machine learning research: JMLR* [Preprint]. Available at: https://www.jmlr.org/papers/volume14/hoffman13a/hoffman13a.pdf.

Kumasaka, N. *et al.* (2021) 'Mapping interindividual dynamics of innate immune response at single-cell resolution', *bioRxiv*. doi:10.1101/2021.09.01.457774.

La Manno, G. *et al.* (2018) 'RNA velocity of single cells', *Nature*, 560(7719), pp. 494–498.

Lawrence, N. and Hyvärinen, A. (2005) 'Probabilistic non-linear principal component analysis with Gaussian process latent variable models', *Journal of machine learning research: JMLR* [Preprint]. Available at: https://www.jmlr.org/papers/volume6/lawrence05a/lawrence05a.pdf.

Lönnberg, T. *et al.* (2017) 'Single-cell RNA-seq and computational analysis using temporal mixture modelling resolves Th1/Tfh fate bifurcation in malaria', *Science immunology*, 2(9). doi:10.1126/sciimmunol.aal2192.

Lopez, R. *et al.* (2018) 'Deep generative modeling for single-cell transcriptomics', *Nature methods*, 15(12), pp. 1053–1058.

Luecken, M.D. and Theis, F.J. (2019) 'Current best practices in single-cell RNA-seq analysis: a tutorial', *Molecular systems biology*, 15(6), p. e8746.

pbmc_unsorted_3k -Datasets -Single Cell Multiome ATAC + Gene Exp. -Official 10x Genomics Support (no date). Available at:

https://support.10xgenomics.com/single-cell-multiome-atac-gex/datasets/2.0.0/pbmc_unsorted_3k (Accessed: 6 January 2022).

pbmc_unsorted_10k -Datasets -Single Cell Multiome ATAC + Gene Exp. -Official 10x Genomics Support (no date). Available at: https://support.10xgenomics.com/single-cell-multiome-atac-gex/datasets/2.0.0/pbmc_unsorted_10k? (Accessed: 6 January 2022).

Pijuan-Sala, B. *et al.* (2019) 'A single-cell molecular map of mouse gastrulation and early organogenesis', *Nature*, 566(7745), pp. 490–495.

Scialdone, A. *et al.* (2015) 'Computational assignment of cell-cycle stage from single-cell transcriptome data', *Methods* , 85, pp. 54–61.

scvelo.datasets.bonemarrow — scVelo 0.2.5.dev6+g1805ab4.d20210829 documentation (no date). Available at: https://scvelo.readthedocs.io/scvelo.datasets.bonemarrow/ (Accessed: 6 January 2022).

Sun, S. *et al.* (2019) 'Accuracy, robustness and scalability of dimensionality reduction methods for single-cell RNA-seq analysis', *Genome biology*, 20(1), p. 269.

Svensson, V. *et al.* (2020) 'Interpretable factor models of single-cell RNA-seq via variational autoencoders', *Bioinformatics* , 36(11), pp. 3418–3421.

Verma, A. and Engelhardt, B.E. (2020) 'A robust nonlinear low-dimensional manifold for single cell RNA-seq data', *BMC bioinformatics*, 21(1), p. 324.

Wang, D. and Gu, J. (2018) 'VASC: Dimension Reduction and Visualization of Single-cell RNA-seq Data by Deep Variational Autoencoder', *Genomics, proteomics & bioinformatics*, 16(5), pp. 320–331.

Wolf, F.A., Angerer, P. and Theis, F.J. (2018) 'SCANPY: large-scale single-cell gene expression data analysis', *Genome biology*, 19(1), p. 15.

Wu, Z. and Wu, H. (2020) 'Accounting for cell type hierarchy in evaluating single cell RNA-seq clustering', *Genome biology*, 21(1), p. 123.

Xiang, R. *et al.* (2021) 'A Comparison for Dimensionality Reduction Methods of Single-Cell RNA-seq Data', *Frontiers in genetics*, 12, p. 646936.

Supplementary Figures

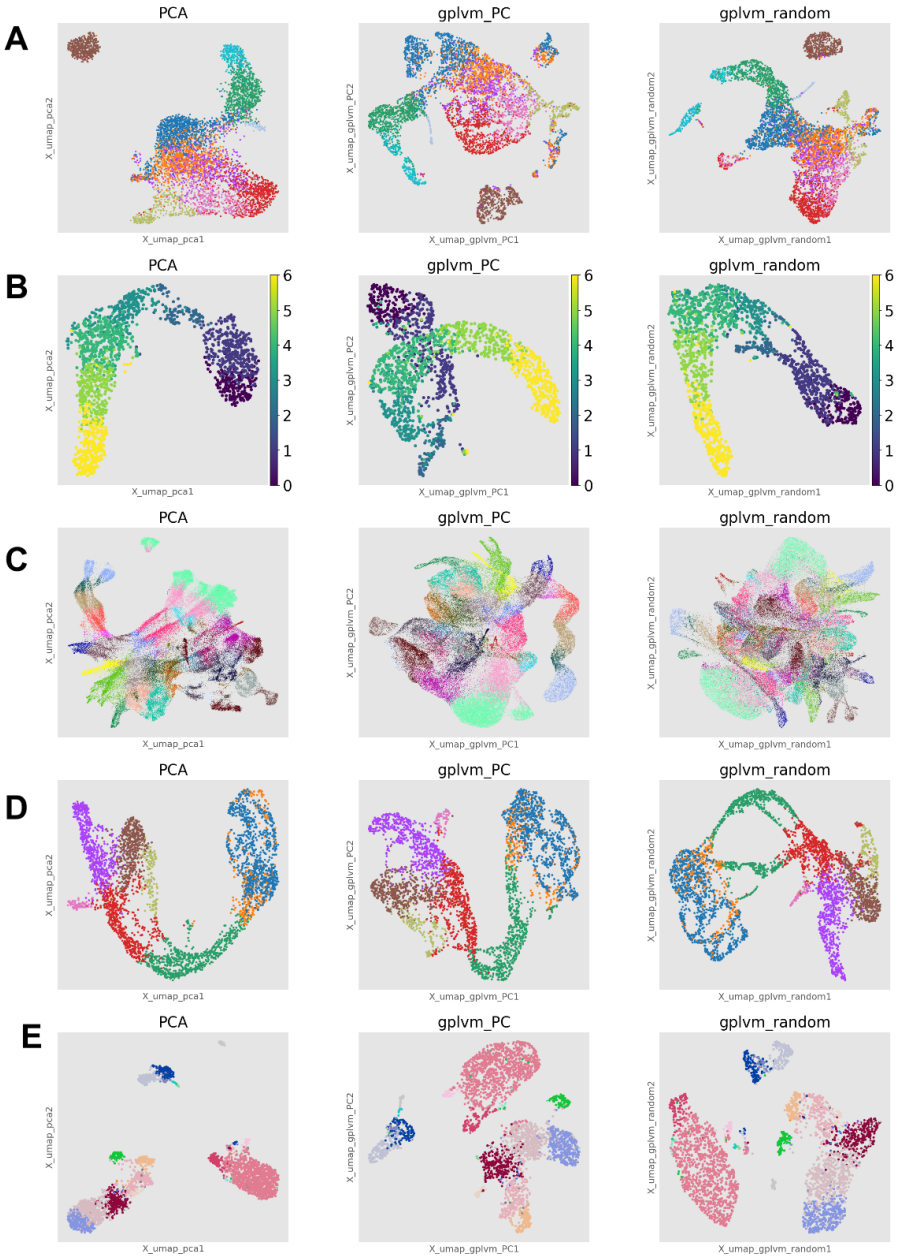

Figure S1: UMAP embeddings based on neighbourhood graphs in PCA (left), GPLVM with PCA initialisation (middle) and GPLVM with random initialisation (right) spaces respectively for bone marrow (**A),** forebrain **(B)**, gastrulation **(C)** , pancreas **(D)**, PBMC3K **(E)**. We observed comparable levels of separation by cell type for the three methods.

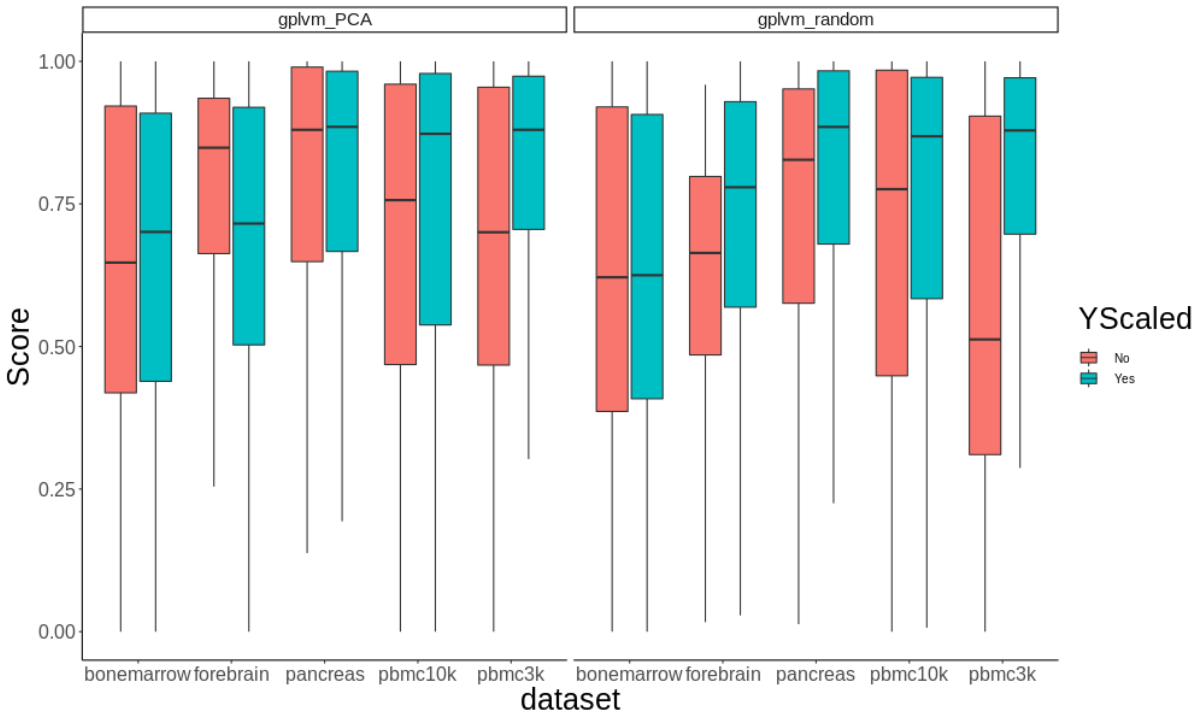

Figure S2
We assessed the effect of scaling or not scaling the gene expression data before training the GPLVM model with PC **(left)** and random **(right)** latent variable initialisation on the distribution of KNN purity scores. We observe a minor improvement in mean KNN purity scores with scaled data.