Exploring variational autoencoders potential to classify single-cell RNA-seq data

Begoña Bolos Sierra[1], Felix Pacheco Pastor[1], Paula Rodriguez[1], Laura Sans-Comerma[1] and Ole Winther[2,3]

1 DTU Bioinformatics, Technical University of Denmark; 2 The Bioinformatics Centre, Department of Biology, University of Copenhagen, 3 DTU Compute, Technical University of Denmark

Introduction

Gene expression levels of individual cells are measured by **single-cell RNA sequencing** (scRNA-seq). This procedure generates large amounts of count data, which is characterized for being highly dimensional and sparse. Therefore, current analysis approaches require intensive pre-processing steps.

Deep generative models, such as **variational auto-encoders** (VAE) can be used to analyze scRNA-seq data, generating latent representations which capture most variability of gene expression levels. The library **scVAE** applies VAE on raw scRNA-seq data, omitting all pre-processing steps and producing a latent representation useful for downstream analyses [1].

In this study, the potential of generative models such as scVAE will be evaluated on a cell type classification problem. To do so, a latent representation of the data will be trained by a Feed Forward Neural Network (FFNN) to classify the cell types. The classification will be compared to a FFNN used on sparse raw data and a sub composition by performing Principal Component Analysis (PCA). The data set used corresponds to 10x-PBMC-68k [2].

Hypothesis and validation steps

Hypothesis: scVAE latent representation improves cell type classification in comparison with the direct use of sparse count data and/or other dimensionality reduction strategies.

Steps to validate hypothesis:

- Compute 16 scVAE models with different hidden units and latent dimensions.
- Evaluation and selection of best models, based on ELBO and Rand Index.
- Perform PCA to raw count data.
- Build a FFNN to classify cell type.
- Compare classification accuracy of the FFNN trained on:
  1. Raw data.
  2. scVAE latent representation.
  3. PCA components.

Workflow

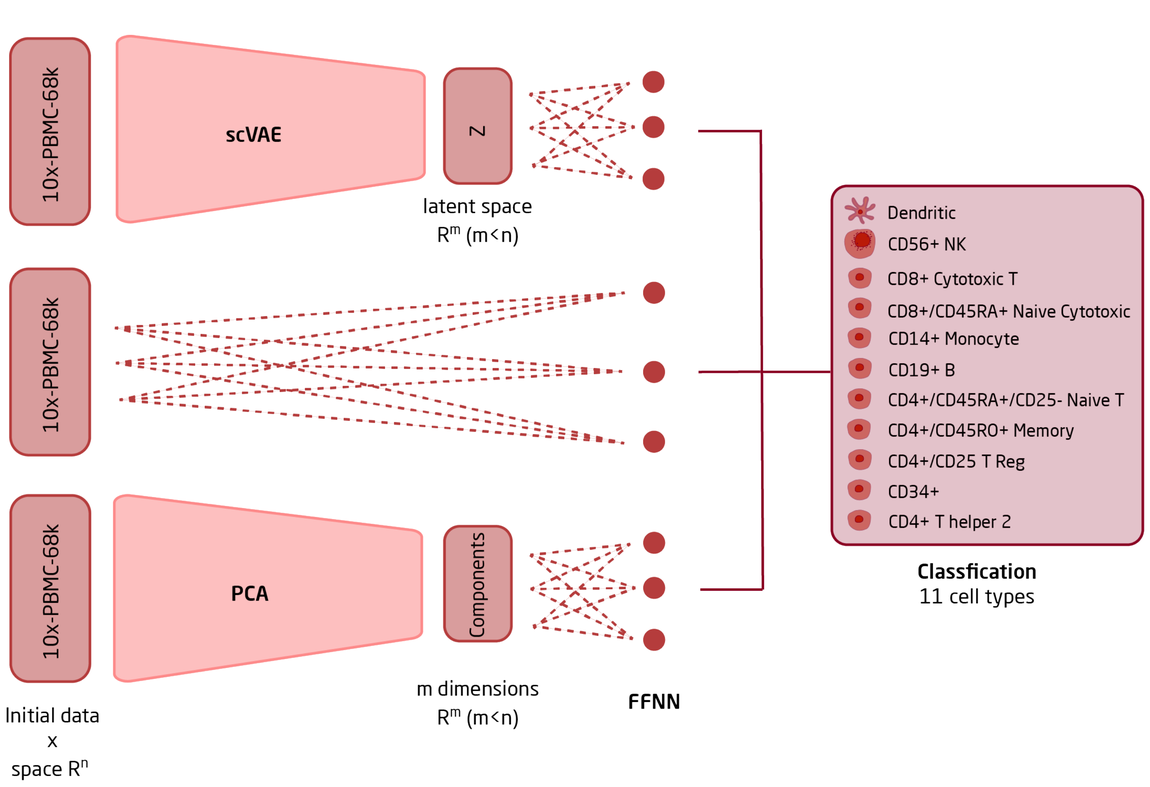

Figure 1: Overview of the workflow

scVAE model performance

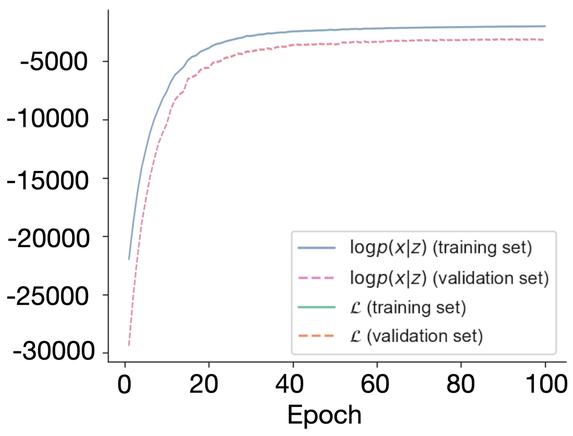

Figure 2: ELBO values across epochs.
Model hyperparameters: Z=100, H=500

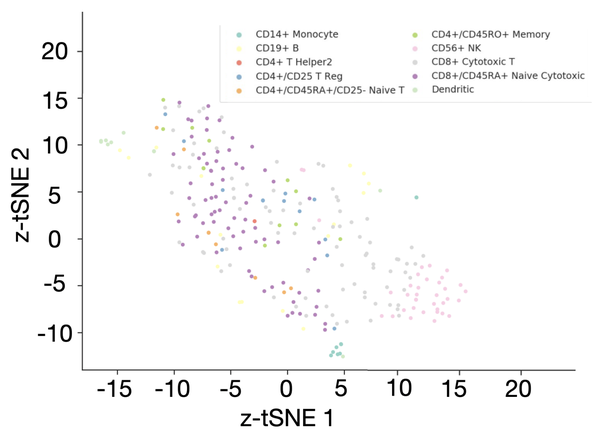

Figure 3: tSNE values of latent representation.
Model hyperparameters: Z=100, H=500

Table 1: Statistics of top-performance models.

Z	H	ELBO	Rand Index
10	250	-4420.9	0.118
25	500	-2704.2	0.110
50	250	-2868.5	0.111
100	**500**	**-2387.5**	**0.106**

Z: Latent space dimensions
H: Number of hidden units
Rand index: Clustering quality.
ELBO: Evidence Lower Bound.

Cell type classification with feed-forward neural network

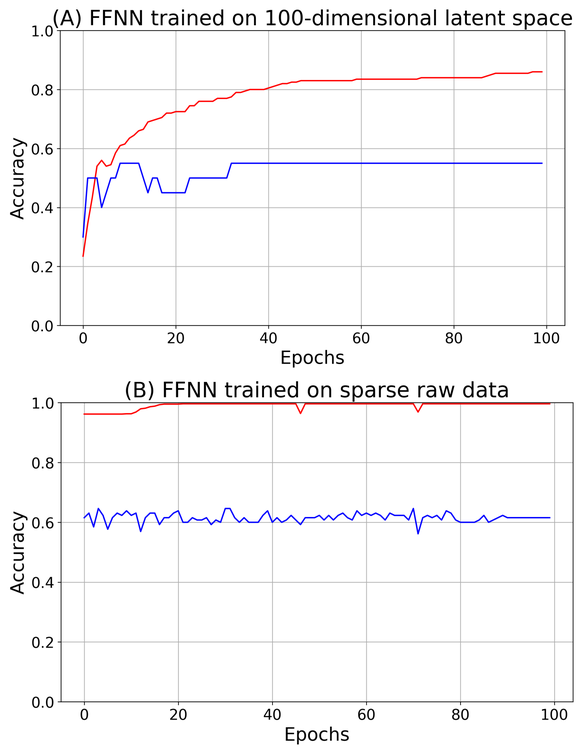

Figure 4: Learning curves of FFNN trained with (A) scVAE latent representation (Z=100, H=500) (B) raw data (C) PCA (k=100)

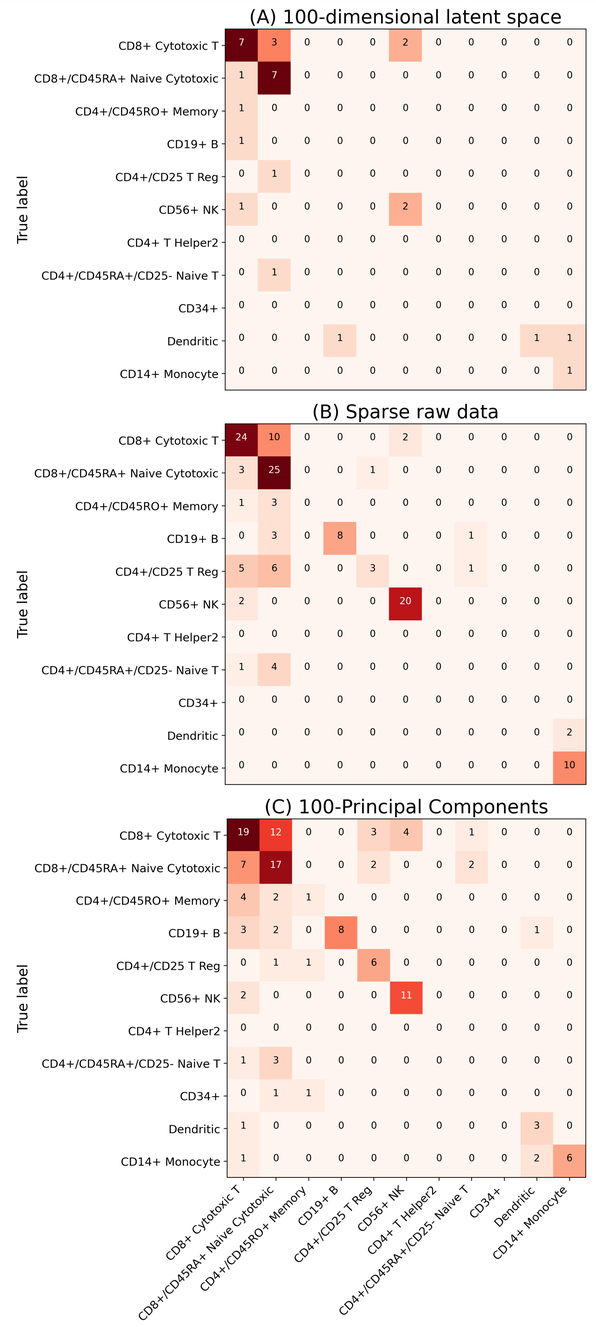

Figure 5: Confusion matrices computed with test set from (A) scVAE latent representation (Z=100, H=500) (B) raw data (C) PCA (k=100)

Effect of dimensionality

To evaluate the use of latent representations to classify cell types. An FFNN was computed using latent representations with different dimensions.

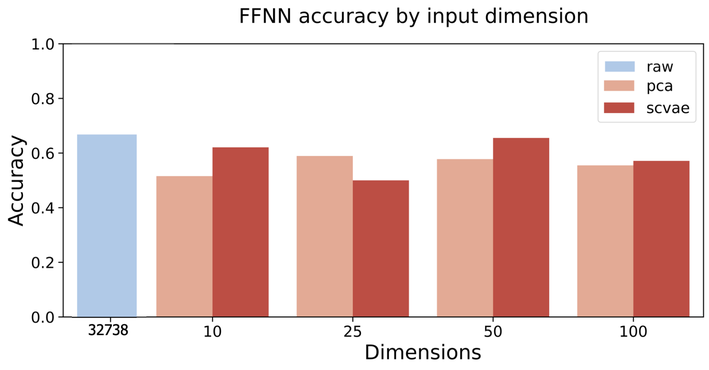

Figure 6: Classification accuracy using input raw data, m PCA components and m dimensional latent space.

Conclusions

- **scVAE** model performance was low in terms of Rand Index and ELBO.
- Overfitting and no differences across inputs were observed in the **FFNN**.
- These results could be explained by the **data set size**.

Future perspectives

- Increase training **data set size** for training scVAE models.
- **Validate** current scVAE model on other external data sets.
- Train scVAE on **different data sets**.

References

[1] C. H. GrÃ¸nbech, M. F. Vording, P. N. Timshel, C. K. SÃ¸nderby, T. H. Pers, and O. Winther. scVAE: variational auto-encoders for single-cell gene expression data. *Bioinformatics*, 36(16):4415–4422, 05 2020. ISSN 1367-4803. doi: 10.1093/bioinformatics/btaa293. URL https://doi.org/10.1093/bioinformatics/btaa293.

[2] G. X. Zheng, J. M. Terry, P. Belgrader, P. Ryvkin, Z. W. Bent, R. Wilson, S. B. Ziraldo, T. D. Wheeler, G. P. McDermott, J. Zhu, M. T. Gregory, J. Shuga, L. Montesclaros, J. G. Underwood, D. A. Masquelier, S. Y. Nishimura, M. Schnall-Levin, P. W. Wyatt, C. M. Hindson, R. Bharadwaj, A. Wong, K. D. Ness, L. W. Beppu, H. J. Deeg, C. McFarland, K. R. Loeb, W. J. Valente, N. G. Ericson, E. A. Stevens, J. P. Radich, T. S. Mikkelsen, B. J. Hindson, and J. H. Bielas. Massively parallel digital transcriptional profiling of single cells. *Nature Communications*, 8(1):14049, 2017. doi: 10.1038/ncomms14049.

s193036@student.dtu.dk, fepac@biosustain.dtu.dk, s192448@student.dtu.dk, s192437@student.dtu.dk