**Methods**

*Imputation via scImpute*

To test whether the imputation improves the performance of GMM-VGAE clustering, we imputed the three real labeled datasets (*i.e.*, Baron 3, Baron 4, and Darmanis) with a state-of-the-art imputation algorithm scImpute. Based on the previous comparative study of single-cell RNA-sequencing imputation methods (Hou *et al.*, 2020), most methods at the current stage do not improve the performance of the downstream traditional clustering analyses. Among those imputation methods showing improvements, we further selected the methods that can impute the technical zeros while preserving the biological zeros, which are scImpute and SAVER (Li and Li, 2018; Huang *et al.*, 2018). However, in our preliminary study, the results from SAVER caused singularity problems during the following clustering. Thus, we picked scImpute for imputation in our research.

During the imputation, scImpute first learns each gene’s dropout probability in each cell by fitting a Gamma-Normal mixture model. Specifically, the dropout probability () of gene in cell is estimated as

where  is gene ’s dropout rate in cell subpopulation ,  are the shape and rate parameters of Gamma distribution, and  are the mean and standard deviation of Normal distribution.

Next, scImpute imputes the expression of gene with high dropout probability ( as the default) in the cell by borrowing information of the same gene in other similar cells. The imputed expression value () is estimated as

where is the expression vector of the same gene in other similar cells. Since the datasets we used in our study contain true labels, we used other cells within cell ’s cluster as the “similar cells”. are the imputation coefficients, which are estimated from the non-negative least squares (NNLS) regression as

where and are the expression values of other genes in cell and those similar cells.

The preprocessing steps (*i.e.*, normalization and log transformation) were as same as the unimputed datasets, which were done to the data before imputation as the algorithm requires. After imputation, the top 1200 genes were selected via variance stabilizing transformation.

**References**

Hou, W., Ji, Z., Ji, H., & Hicks, S. C. (2020). A systematic evaluation of single-cell RNA-sequencing imputation methods. *Genome biology*, *21*(1), 218.

Li, W. V., & Li, J. J. (2018). An accurate and robust imputation method scImpute for single-cell RNA-seq data. *Nature communications*, *9*(1), 997.

Huang, M., Wang, J., Torre, E., Dueck, H., Shaffer, S., Bonasio, R., Murray, J. I., Raj, A., Li, M., & Zhang, N. R. (2018). SAVER: gene expression recovery for single-cell RNA sequencing. *Nature methods*, *15*(7), 539–542.