VAE application to single-cell RNAseq data

Agnieszka Michalak

1. Data Exploration

The scRNA-seq data was split into training and testing dataset containing 72 208 and 18 052 barcodes (observations) respectively with 5000 genes (variables). Barcodes are the molecular tag of each cell, which correspond to the number of single cells, however oftentimes due to the technology artifacts this assumption is violated (ex. duplicates).

adata.obs table contains metadata of the experiments. The metadata includes information on the number of donors (n=9) and their information such as: 'DonorNumber', 'DonorID', 'DonorAge', 'DonorBMI', 'DonorBloodType', 'DonorRace', 'Ethnicity', 'DonorGender', 'QCMeds' 'DonorSmoker'.
Cell information includes: 'GEX_n_genes_by_counts', 'GEX_pct_counts_mt', 'GEX_size_factors' 'GEX_phase', 'ADT_n_antibodies_by_counts' 'ADT_total_counts', 'ADT_iso_count' 'cell_type', 'batch', 'ADT_pseudotime_order',
'GEX_pseudotime_order'.
The number of cell types is 45 (Fig. 1.).
The data was prepared in 4 laboratories. The batches were prepared by the laboratory id and the donor id, resulting in 12 batches.
On the Figure 1 is the distribution of cell type by donor. Donor with the id 15078 has significantly more observations, which corresponds to the higher number of raw counts (Fig. 2).

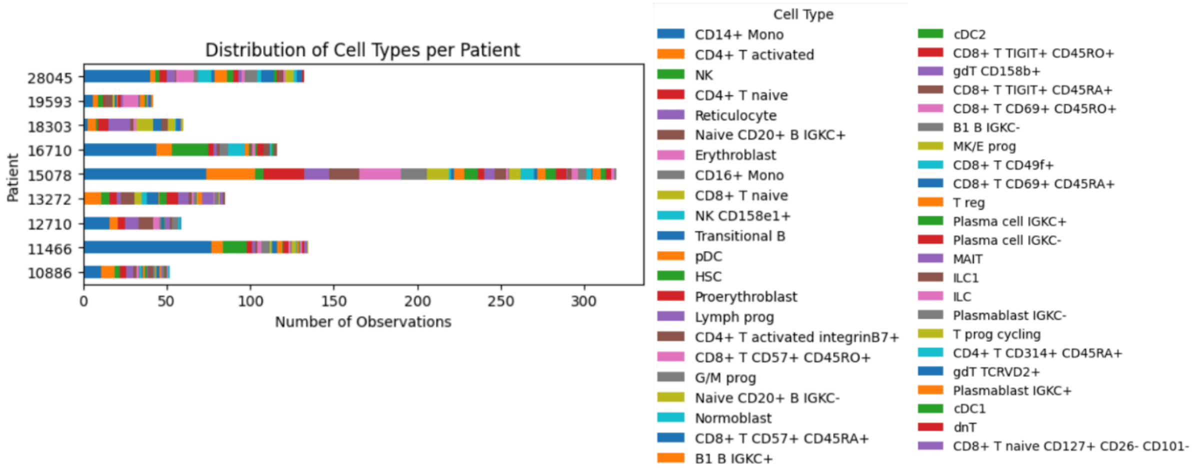


Figure 1. Distribution of cell types per Donor (Patient).

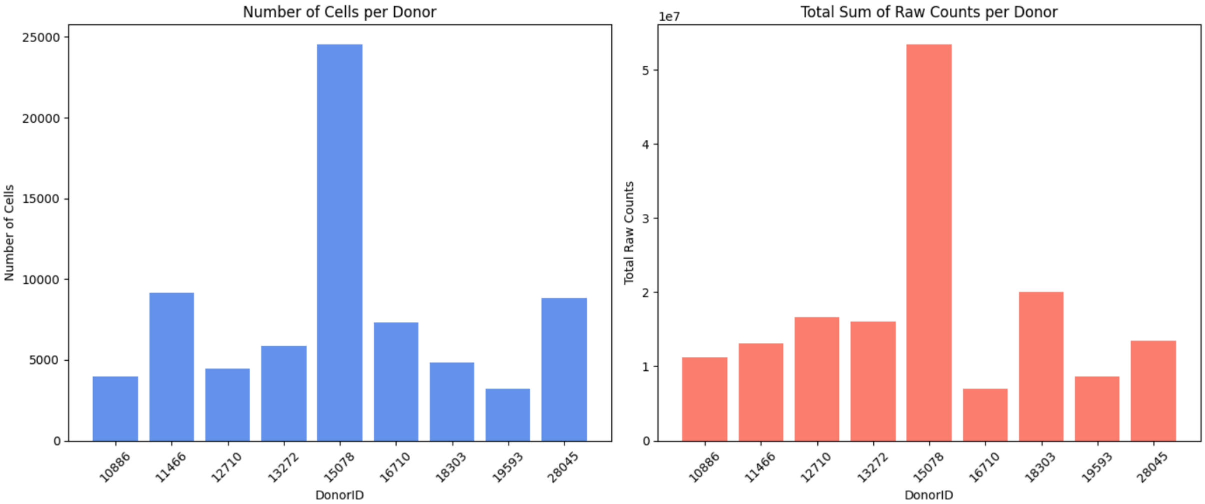


Figure 2. Number of cells and raw counts by each Donor.

Because the number of cells and the number of raw counts correlate to each other, we can conclude that the reason behind this is the bigger sample size of cells that were sequenced.

Preprocessed and raw counts were plotted (Fig. 3.) Their distribution corresponds to the negative binomial distribution which is often used in analyzing tools, for example, to model top variable features.

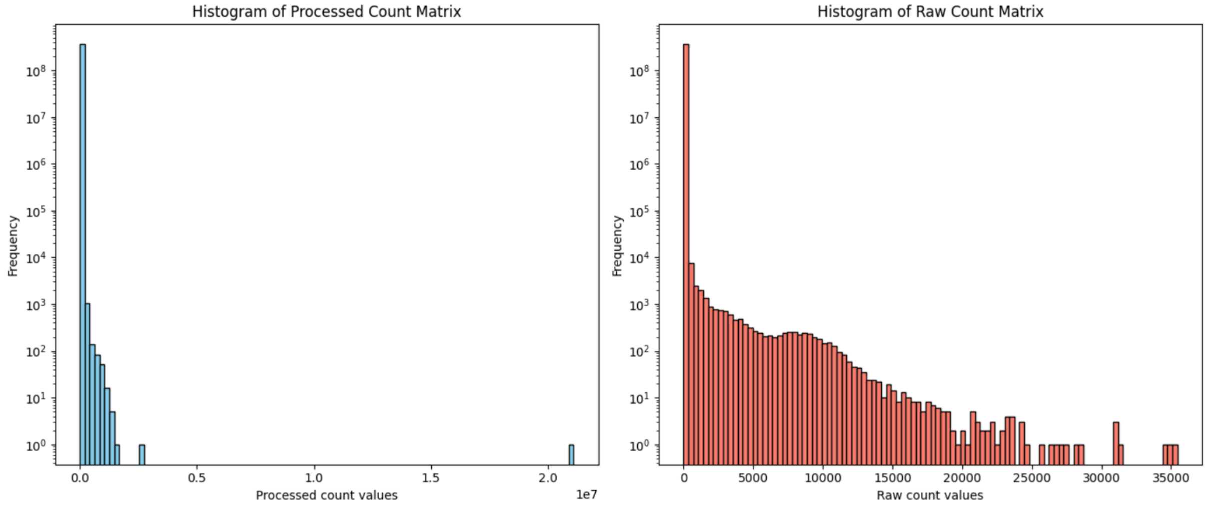


Figure 3. Distributions of preprocessed and raw counts.

The 0 count values are abundant in the distribution (around 10^8-10^9). The 0 count values are divided into “biological zeroes” and “technical zeroes”. Biological zeroes correspond to the true biological non-expressed genes and technical zeroes (dropouts) where mRNA is expressed but not detected.

Due to the dropout effect and cell heterogeneity, the fraction of zeroes is even at 90%.

After filtering the highest counts, we can notice how much the 0 frequency is more abundant compared to the expressive frequency (Fig. 4).

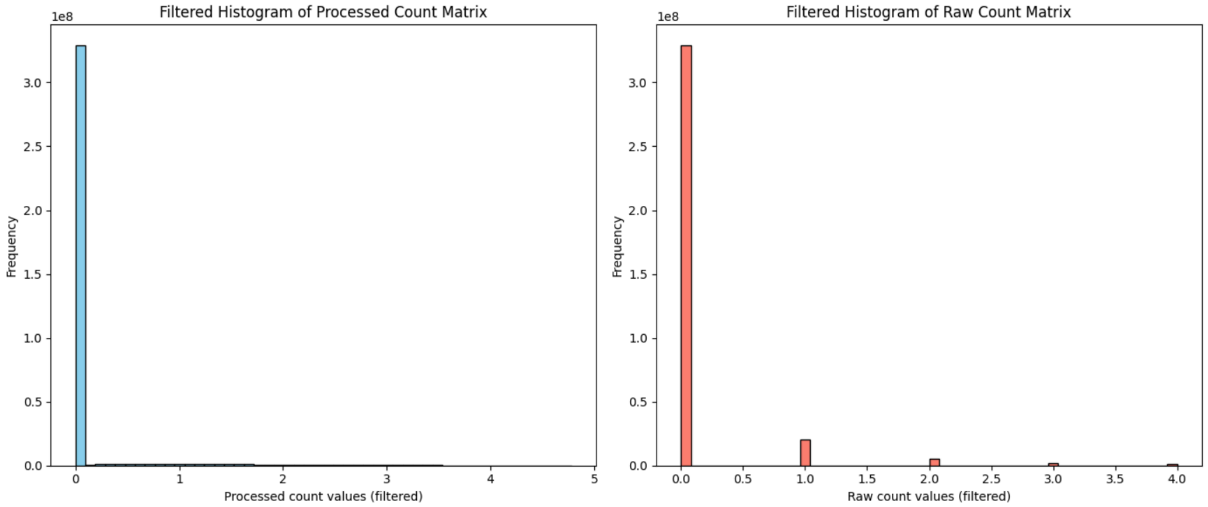


Figure 4. Distribution of preprocessed and raw counts after filtering out the highest counts.

The goal of data normalization is to remove technical variation to correctly draw further conclusions based on biological variation. Normalization consists of two steps: 1) Scaling - calculating size factors and scaling counts by them. 2) Transformation - addressing the compositional bias.
The aim of transformation is to reduce the domination by highly expressed genes. It allows for data comparison across samples. The most common method is log-transformation. However, this method favors genes with low expression and dampens highly expressed genes. The proposed solution to this problem is using Pearson residuals.

From the preprocessed and raw data parameters (Table 1) we can draw the conclusion that the raw data is extremely sparse - most entries are 0, and only a few entries have high counts. In contrast, the processed data shows a much higher maximum value and a higher mean. The processed data were likely normalized using size factor to adjust for differences in sequencing depth across cells. However, the very large maximum indicates that the data still has a long tail of very high values - so it remains highly skewed.

Table 1. Preprocessed and raw counts parameters.

	Raw counts	Preprocessed counts
Min	0	0
Max	35451	21078940
Mean	0.4	3.4

To the preprocessed data the log-transformation should be applied to reduce the dominance of high values.

2. Variational Autoencoder

The encoder network maps an input x into a distribution over the latent space z. The encoder outputs the parameters of a Gaussian distribution mean and variance.

$$q_\phi(z|x) = N(z; \mu_\phi(x), \mathrm{diag}(\sigma_\phi(x)^2))$$

The decoder network maps a latent variable z back into the data space. It defines a likelihood over x given z.

$$p_\theta(x|z) = N(x; z, \sigma^2)$$

When training VAE, we need to backpropagate through a sampling operation from $q_\phi(z|x)$, but stochastic sampling is not differentiable. The trick is to rewrite the sampling step as:

$$z = \mu_\phi(x) + \sigma_\phi(x) \odot \varepsilon,\ \varepsilon \sim N(0,I)$$

This transformation “moves” the randomness into ε so that gradients can flow through $\mu\phi(x)$ and $\sigma\phi(x)$ during backpropagation.

The Kullback-Leibler (KL) divergence is a measure of how one probability distribution ($q_\phi(z|x)$) diverges from a second, reference probability distribution p(z).

$$\mathrm{KL}(q(z) \,\|\, p(z)) = \int q(z) \log \frac{q(z)}{p(z)}\, dz$$

In VAE, the KL divergence is used to regularize the encoder’s output $q_\phi(z|x)$ by penalizing deviations from the chosen prior p(z).
The reconstruction loss measures how well the decoder p(z) can reconstruct the original input x from the latent representation z. It is typically the negative log-likelihood of x under the decoder’s distribution.

$$L_{recon} = -E_{q\phi(z|x)} [\log p_\theta(x|z)]$$

The Evidence Lower Bound (ELBO) is a lower bound on the log-likelihood of the observed data and serves as the objective function for training VAE. It balances the trade-off between reconstruction accuracy and regularization of the latent space.

$$\mathrm{ELBO} = -E_{q\phi(z|x)} [\log p_\theta(x|z)] + \mathrm{KL}(q_\phi(z|x) \,||\, p(z)$$

The closer the ELBO is to log p(x), the better $q\phi(z|x)$ approximates p(z|x).

β-VAE introduces a hyperparameter β to control the trade-off between the reconstruction term and the KL divergence term in the ELBO. The modified loss function becomes:

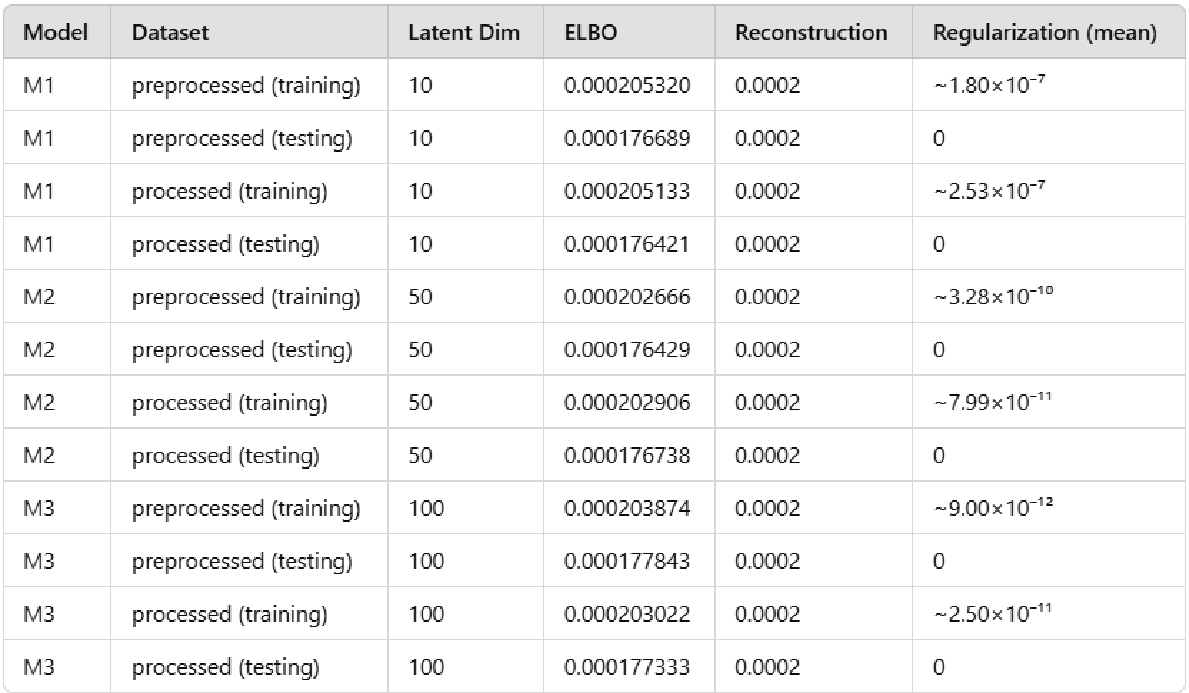
$$L = -E_{q\phi(z|x)} [\log p\theta(x|z)] + \beta KL(q\phi(z|x) \,||\, p(z))$$

β-VAE is often beneficial when learning disentangled representations is important.
In tasks where high-fidelity reconstructions are important, such as medical imaging or tasks over-regularization may lead to poor reconstruction performance.

3. Results

Raw, preprocessed and log-transform processed dataset was prepared for the VAE model. However, when using raw data in the merged dataset, the gradient was unstable, therefore only preprocessed and processed data was used.
Model variants were trained and tested with various latent space size (M1: 10, M2: 50 and M3: 100) and hidden layer size (M2: 256, M4: 512, and M5: 1024). Out of the tested latent space size, the 50 size performed the best (Table 2).

Table 2. Comparing models (M) with various latent sizes.

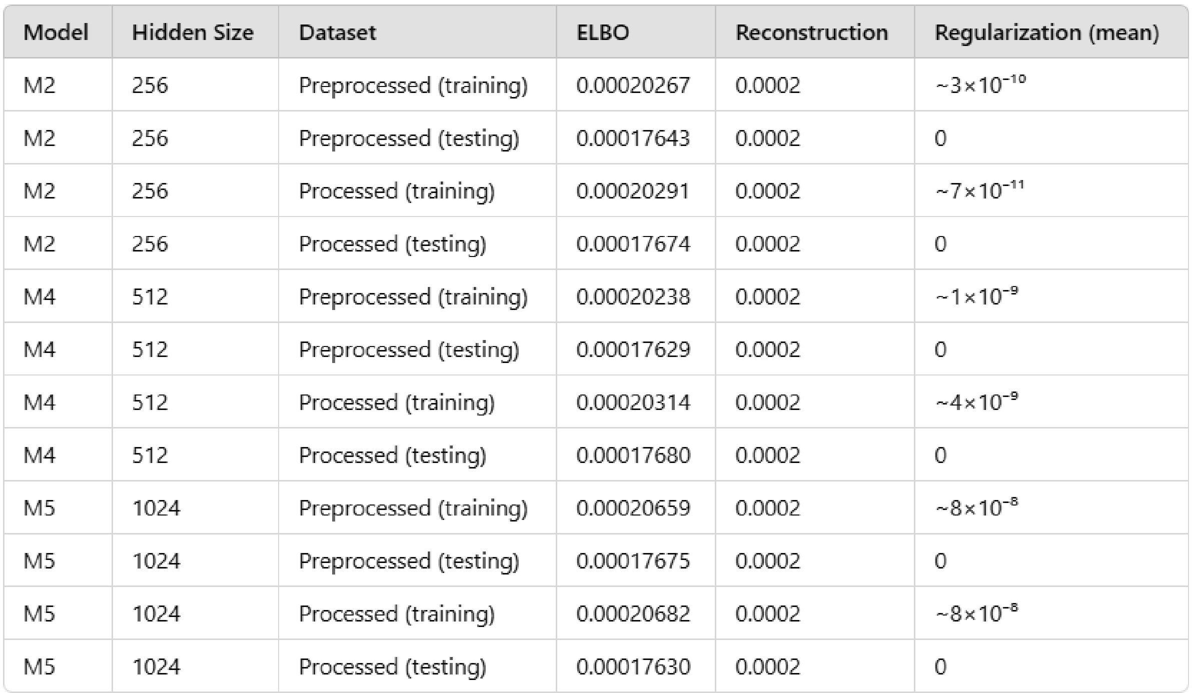


Model	Dataset	Latent Dim	ELBO	Reconstruction	Regularization (mean)
M1	preprocessed (training)	10	0.000205320	0.0002	$\sim 1.80\times10^{-7}$
M1	preprocessed (testing)	10	0.000176689	0.0002	0
M1	processed (training)	10	0.000205133	0.0002	$\sim 2.53\times10^{-7}$
M1	processed (testing)	10	0.000176421	0.0002	0
M2	preprocessed (training)	50	0.000202666	0.0002	$\sim 3.28\times10^{-10}$
M2	preprocessed (testing)	50	0.000176429	0.0002	0
M2	processed (training)	50	0.000202906	0.0002	$\sim 7.99\times10^{-11}$
M2	processed (testing)	50	0.000176738	0.0002	0
M3	preprocessed (training)	100	0.000203874	0.0002	$\sim 9.00\times10^{-12}$
M3	preprocessed (testing)	100	0.000177843	0.0002	0
M3	processed (training)	100	0.000203022	0.0002	$\sim 2.50\times10^{-11}$
M3	processed (testing)	100	0.000177333	0.0002	0

Every result reports a reconstruction error of 0.0002. This indicates that all models reconstruct the data equally well. The overall ELBO in these experiments is the sum of the reconstruction term and the regularization (KL divergence) term. Since the reconstruction term is identical across the board, differences in ELBO come solely from the regularization term.
These results are very similar, however, 50-dimensional latent space has slightly lower ELBO values, so might treat it as the best optimization in terms of latent space.

The model variants were further tested on hidden layer size (Table 2).

Table 2. Comparing models (M) with various hidden layer sizes.



Model	Hidden Size	Dataset	ELBO	Reconstruction	Regularization (mean)
M2	256	Preprocessed (training)	0.00020267	0.0002	$\sim 3\times10^{-10}$
M2	256	Preprocessed (testing)	0.00017643	0.0002	0
M2	256	Processed (training)	0.00020291	0.0002	$\sim 7\times10^{-11}$
M2	256	Processed (testing)	0.00017674	0.0002	0
M4	512	Preprocessed (training)	0.00020238	0.0002	$\sim 1\times10^{-9}$
M4	512	Preprocessed (testing)	0.00017629	0.0002	0
M4	512	Processed (training)	0.00020314	0.0002	$\sim 4\times10^{-9}$
M4	512	Processed (testing)	0.00017680	0.0002	0
M5	1024	Preprocessed (training)	0.00020659	0.0002	$\sim 8\times10^{-8}$
M5	1024	Preprocessed (testing)	0.00017675	0.0002	0
M5	1024	Processed (training)	0.00020682	0.0002	$\sim 8\times10^{-8}$
M5	1024	Processed (testing)	0.00017630	0.0002	0

Similarly to the first comparison, the reconstruction error is fixed at 0.0002 across all variants. Model 4 has better trade-off among the tested hidden layer sizes, having slightly better ELBO than Model 4, but not as computationally expensive as Model 5.

On Figure 5 is depicted latent space of model 3 and space is colored by cell type annotation. We can notice that the cell types form clusters. The closer related cells are to each other, the closer their clusters are on the latent space. The latent space on the Figure 5 is dimensionally redacted using Uniform Manifold Approximation and Projection (UMAP). UMAP's belief is that although data may exist in a high-dimensional space, it actually lies on a manifold of much lower dimensionality. This means that locally, the data behaves as if it were embedded in a simpler, lower-dimensional space. By focusing on these local neighborhoods, UMAP can capture the true structure of the data.

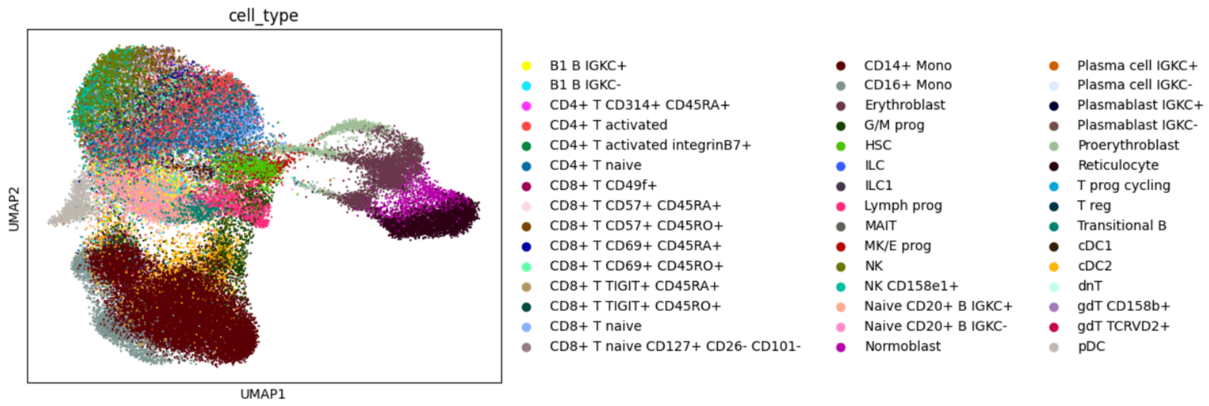


Figure 5. UMAP dimension reduction of model 3 latent space.

Batch effects refer to the technical errors resulting from the sample preparation in batches, for instance in different laboratories, or different personnel. In the Figure 6. we check for the batch effects.

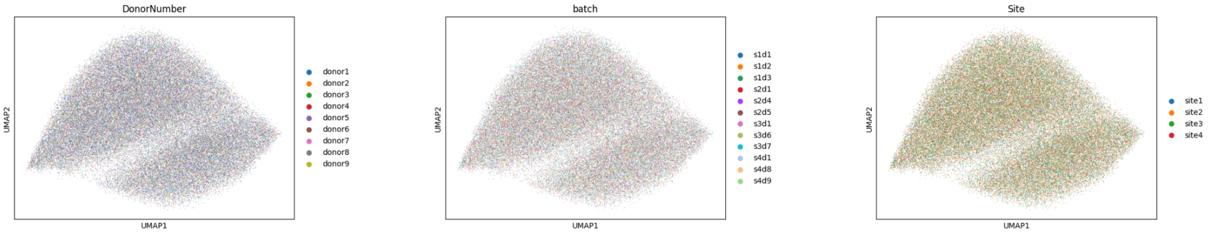


Figure 6. Batch effect was checked on model 3 latent space.

Batch effects were checked on donors, batches (sites x donors) and sites. There are no batch effects in the data because data points are mixed.

Figure 7 visualizes batch effects on the latent space of the final model on the test dataset.

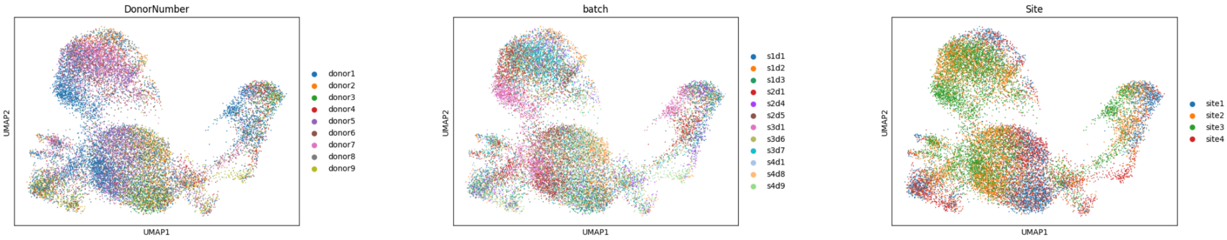


Figure 7. UMAP dimension reduction of model 4 latent space.

This UMAP visualization (Fig. 7) is compared to the Figure 8 where the batch effects on the test dataset are visualized with the Principal Component Analysis (PCA).

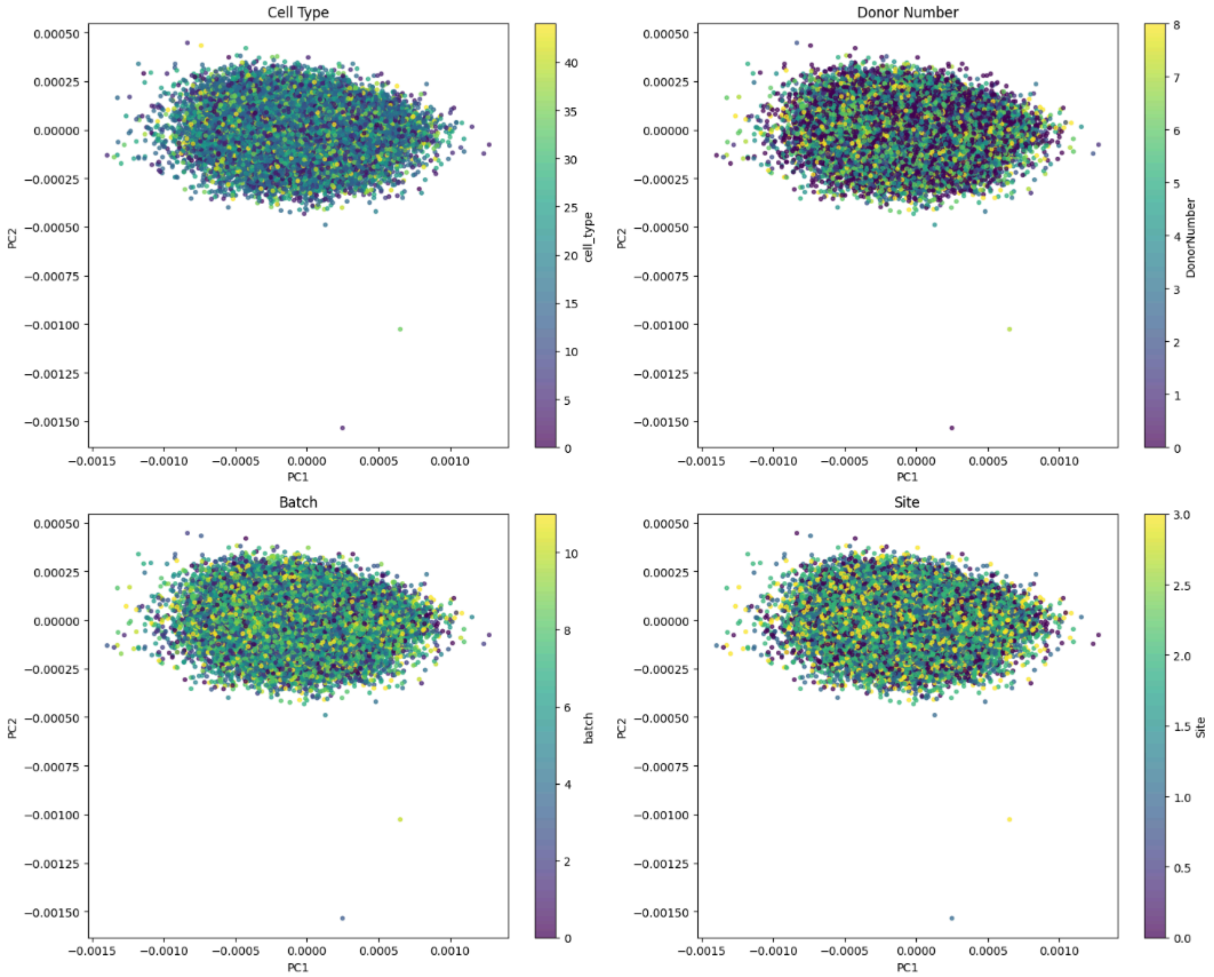


Figure 8. PCA dimension reduction of the model 4 latent space.

Both UMAP and PCA capture the lack of batch effects in the test dataset, characterized by the mixed data points across batches.

4. Conclusions

Raw dataset is not compatible with using VAE model, because of the non-normalized values it causes unstable gradients.

All tested model variants achieved very similar losses, suggesting that these parameters might not be that important to optimize and it would be interesting to test for further improvements.

Datasets were batch-effect free and the latent space visualizations using UMAP and PCA are equally informative.