SAD2022Z_Report_Ada_Hryniewicka

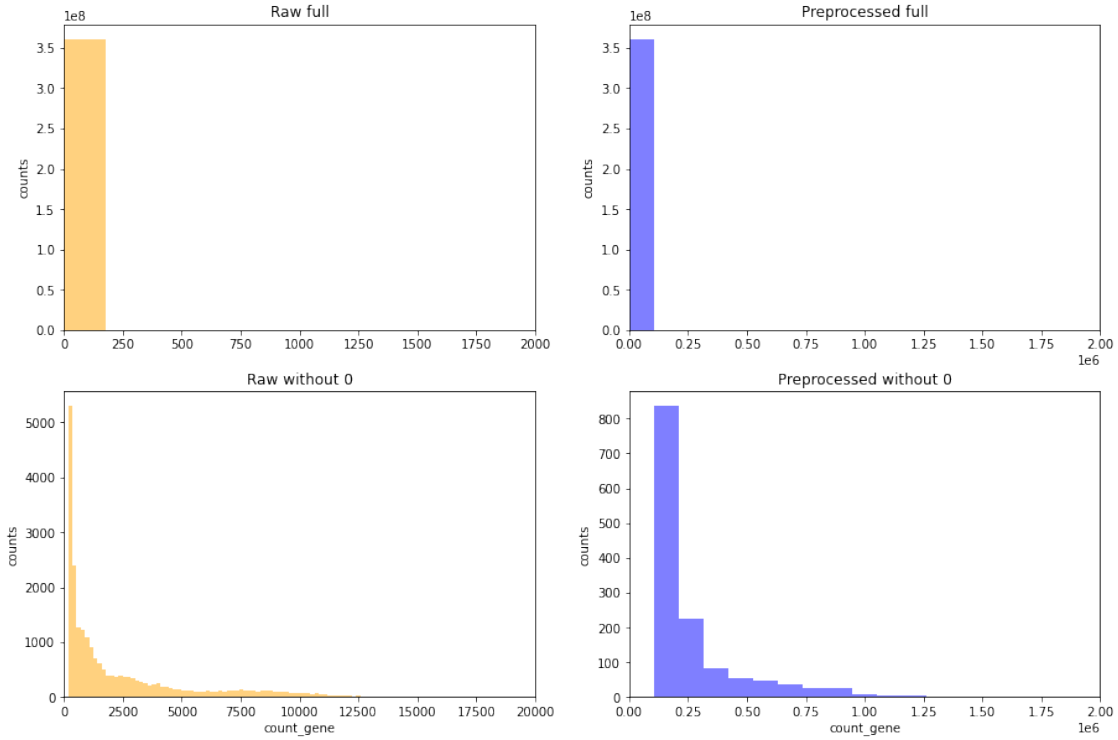
December 14, 2022

The goal was to apply different Variational Autoencoders (VAEs) to biological data-single cell RNA sequencing. The first step was examination of the data. Later it was trained on Vanilla VAE and the last step was changing the decoder in VAE for better results. Output was visualized by plotting the first two PCAs.

1 Exploration

```
Observations and variables numbers for train dataset: 72208 5000
Observations and variables numbers for test dataset: 18052 5000
```

Histograms showing the data distribution.



The abundance of zeros overwhelms the plots. The second histograms are showing data excluding zeros.

Researchers view vast zeros in single-cell RNA-seq data differently: some regard zeros as biological signals representing no or low gene expression, while others regard zeros as missing data to be corrected.

Firstly, Poisson distribution comes to mind because it is for discrete data but here the decision was to choose exponential distribution which is a kind of generalization of Poisson for continuous data. It can be helpful when doing operations like normalization and work with no-integers also. What is more, later when there is modification of the decoder according to distribution there is no sampling from it for loss function (using probability) so choosing continuous distribution instead of discrete should not have negative effect.

```
Means
Raw: 1805200.0
Cleaned: 1805200.0

Median
Raw: 3.0
Cleaned: 0.0

Max
Raw: 360994644
Cleaned: 361037234

Min
Raw: 0
Cleaned: 0
```

Cleaned dataset (adata.X) was prepared by dividing raw dataset counts(adata.layers[”counts”]) by GEX_size_factor for each cell- normalization. The authors of dataset say that log1p transformation data is stored in adata.layers[”log_norm”] but this column is absent in our dataset.

```
For index: 20
preprocessed: 0.4978048
raw: 1.0
raw after dividing by GEX_size_facore: 0.4978047829330648
```

The adata_train.obs contains information about detailed information about cells. They are shown below.

```
Index(['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', 'Samplename', 'Site', 'DonorNumber', 'Modality',
       'VendorLot', 'DonorID', 'DonorAge', 'DonorBMI', 'DonorBloodType',
       'DonorRace', 'Ethnicity', 'DonorGender', 'QCMeds', 'DonorSmoker',
       'is_train'],
      dtype='object')

Number of patients: 9
Number of cell types: 45
```

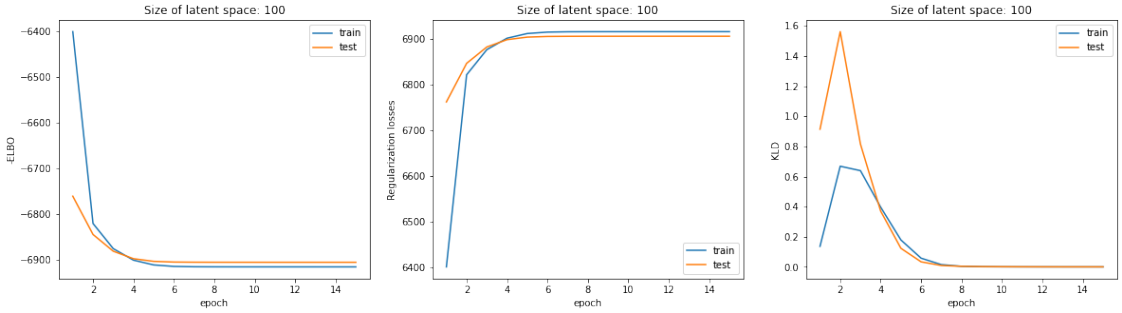
```
Number of labs: 4
```

2 Vanilla VAE training

The learning curves for Vanila VAE and latent space- 100.

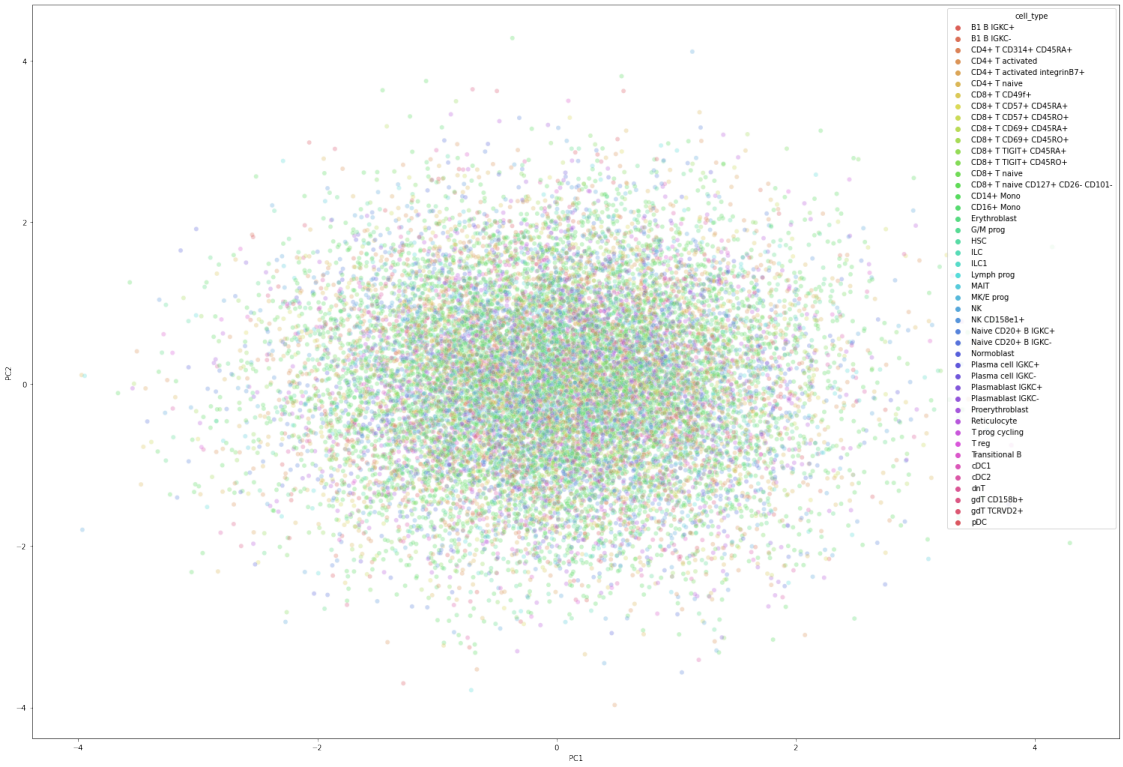
Comparing the -ELBO losses for different latent spaces.

```
        Latent space 100  Latent space 50  Latent space 300
-ELBO       -6906.421685     -6906.379411       -6906.420672
```



The lowest -ELBO is for latent space 100 and this size was chosen for further computations.

```
The number of principal components that explains more than 95% of the variance:
95
```



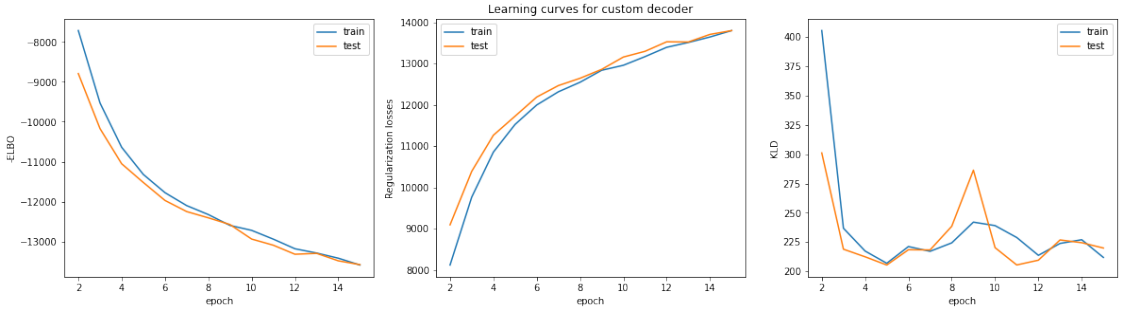
The Vanilla VAE does not provide reasonable results, so it needs to be improved by changing the distribution in Custom Decoder. The exponential distribution was chosen. The same results were on different labels coloring. This decoder has not done a proper job.

3 Custom Decoder

Custom decoder modification was about changing the distribution in probability loss function to exponential. The first step was deciding which latent space to choose. Comparison of -ELBO are shown below.

	Latent space 100	Latent space 50	Latent space 300
-ELBO	-13433.67627	-13582.166731	-13435.678794

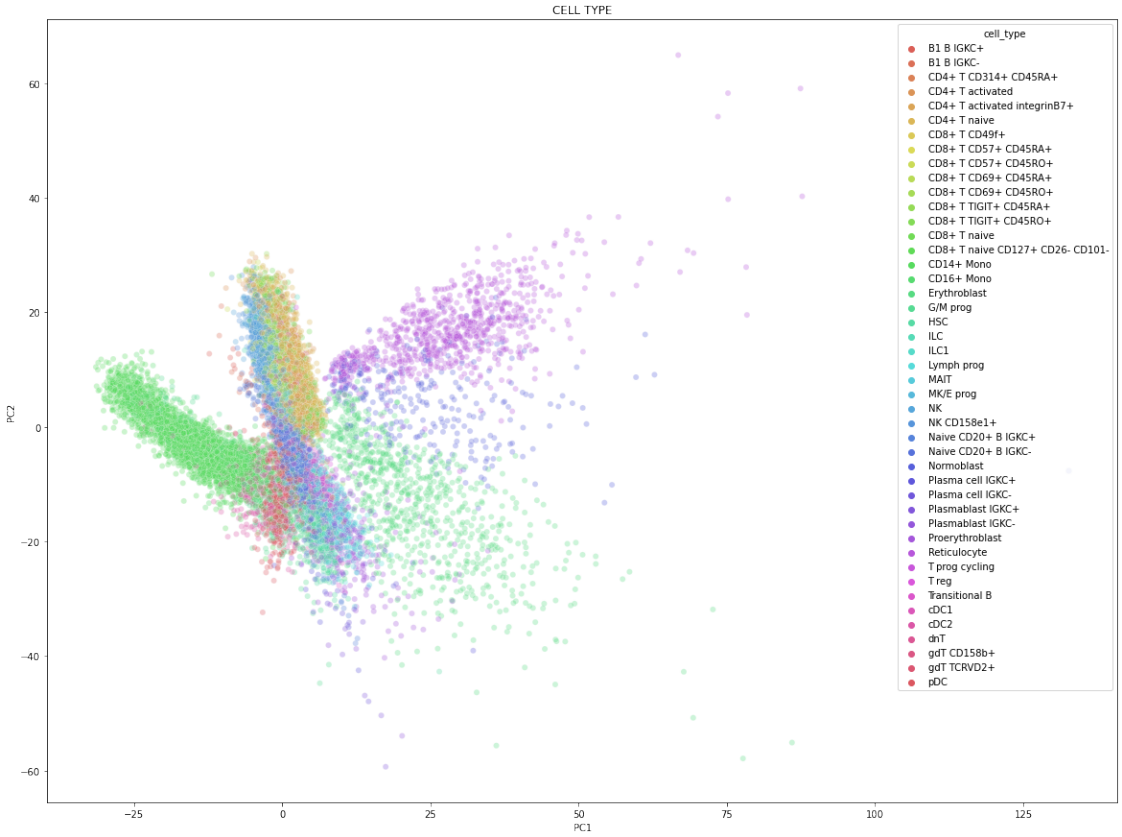
The lowest -ELBO is for latent space 50. This model was chosen for PCA analysis. The fraction of explained variance was the same for latent space 100 and latent space 50.

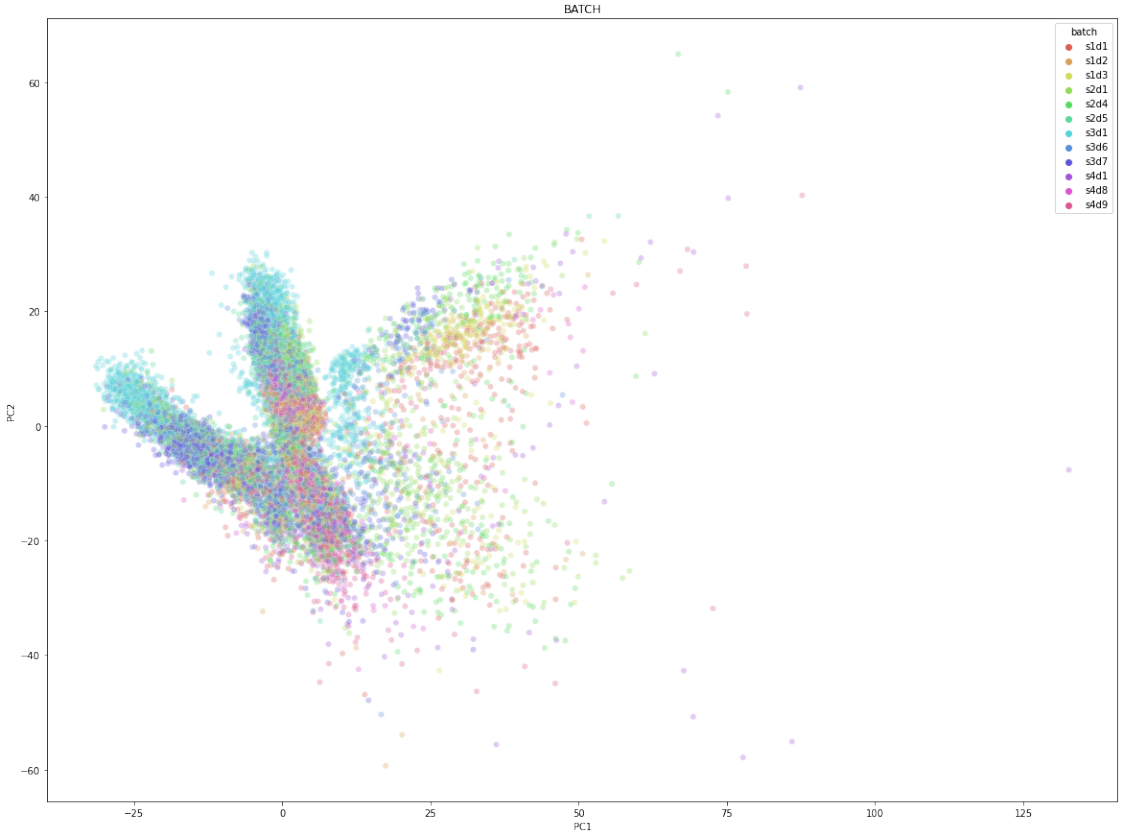


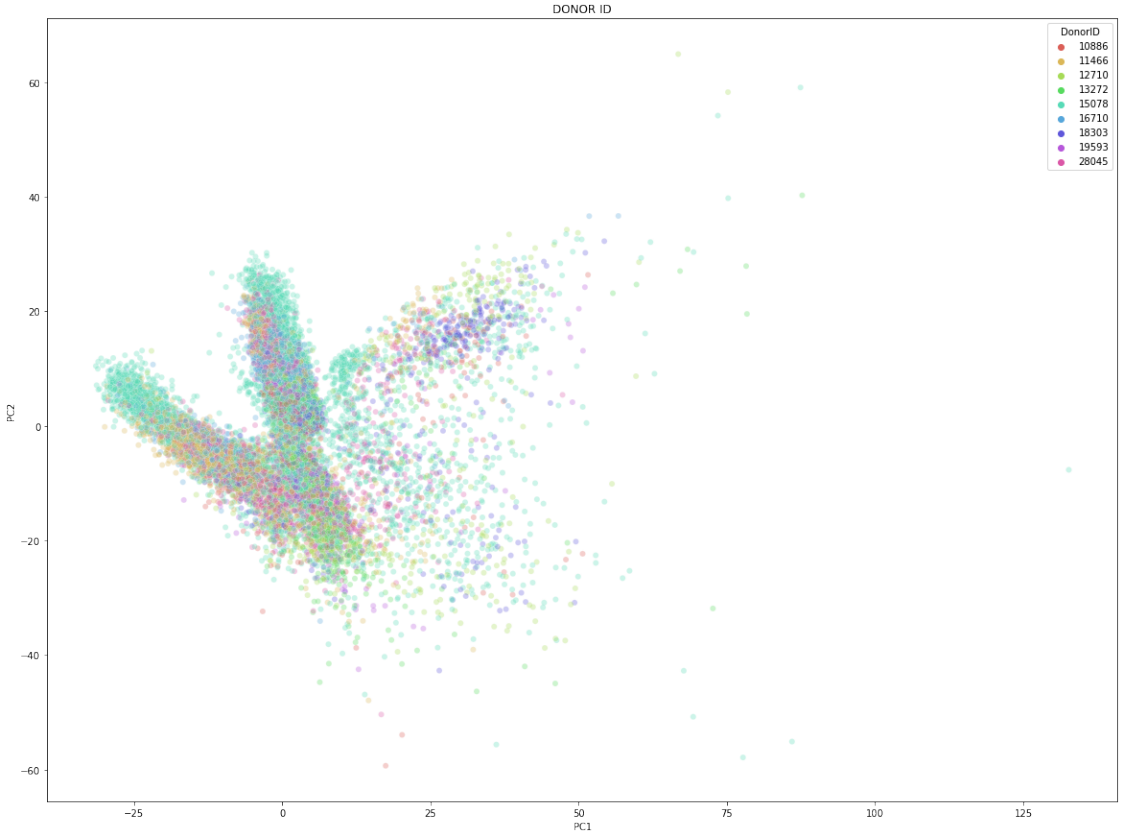
`The number of principal components that explains more than 95% of the variance for new decoder: 29`

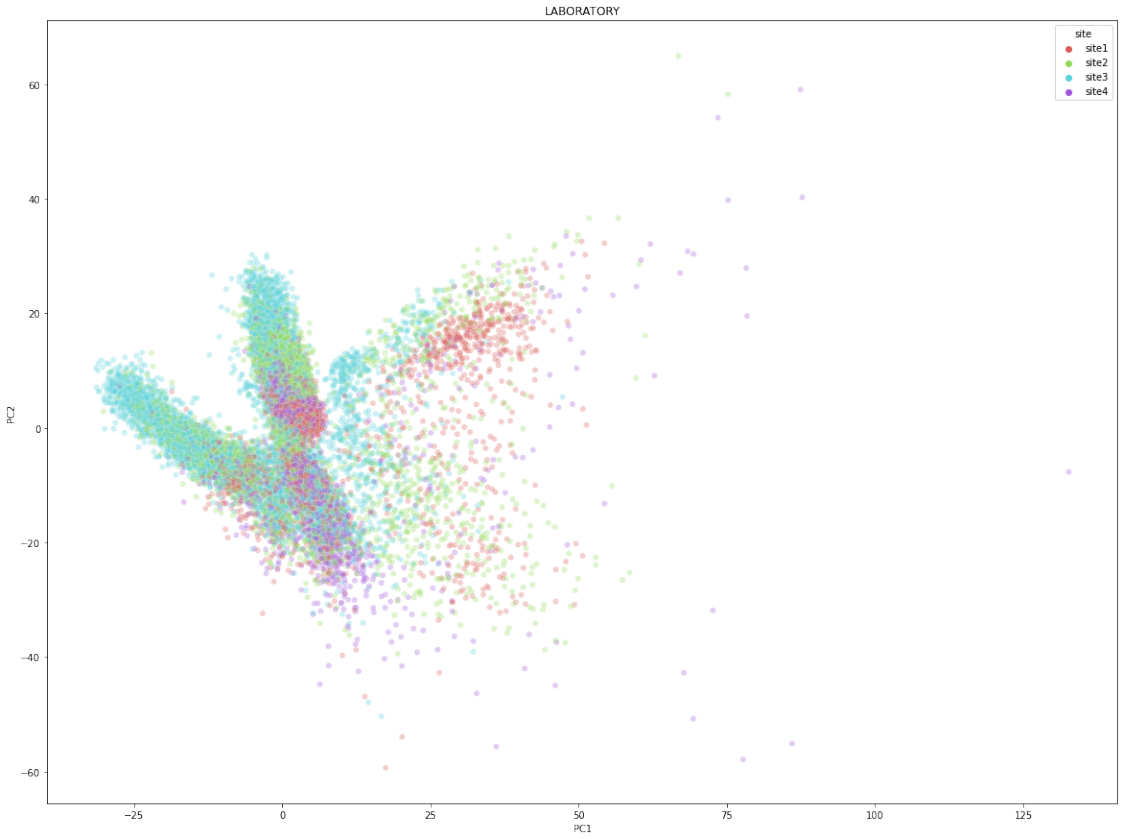
Which shows that new decoder imporved the result from 95/100 in Vanilla to 29/50 for custom decoder.

The next step is to present first two principal components coloured by specific label.




BATCH
batch
s1d1
s1d2
s1d3
s2d1
s2d4
s2d5
s3d1
s3d6
s3d7
s4d1
s4d8
s4d9
PC2
PC1
60
40
20
0
−20
−40
−60
−25
0
25
50
75
100
125


DONOR ID
DonorID
10886
11466
12710
13272
15078
16710
18303
19593
28045
PC2
PC1
60
40
20
0
−20
−40
−60
−25
0
25
50
75
100
125



Plots of PCA show that there are not significant proofs of batch effect. The best clustering is present in cell type. Other factors such as site, donor ID or batch do not cluster.

4 Summary

Customizing the decoder provides much better results, which are shown on the plots. Moreover, the fraction of explained 95% of variance improved from 95/100 to 29/50. Concluding, the normal distribution in decoder is not a good approach, when dealing with this type of data. Exponential distribution is more accurate.