Imputation for single-cell multi-omics data

Zhao Yimin

Johns Hopkins University

November 7, 2021

Contents

Introduction

- Multiple types of measurement can be made simultaneously on single cells in the same experiment
- Method of the Year 2019 by Nature Methods

Existing protocols

Time	Protocol	Data type
2016	scMT-seq	Gene expression; Methylation
2018	sci-CAR	Chromatin accessibility; Gene expression
2018	scNMT-seq	Gene expression; Chromatin accessibility; Methylation
2018	SNARE-seq	Chromatin accessibility; Gene expression
2020	SHARE-seq	Chromatin accessibility; Gene expression
2020	10X Multiome	Chromatin accessibility; Gene expression
2020	CITE-seq	Gene expression; Protein levels
2021	DOGMA-seq	Chromatin accessibility; Gene expression; Protein levels
2021	Paired-Tag	Gene expression; Histone Modification

Motivation

- Data quality is not as good as protocols which produce data of one modality(i.e. Data may be more sparse)
- Expect to get better imputation result since we incorporte information from gene expression and chromatin accessibility

Contents

Concatenation Model

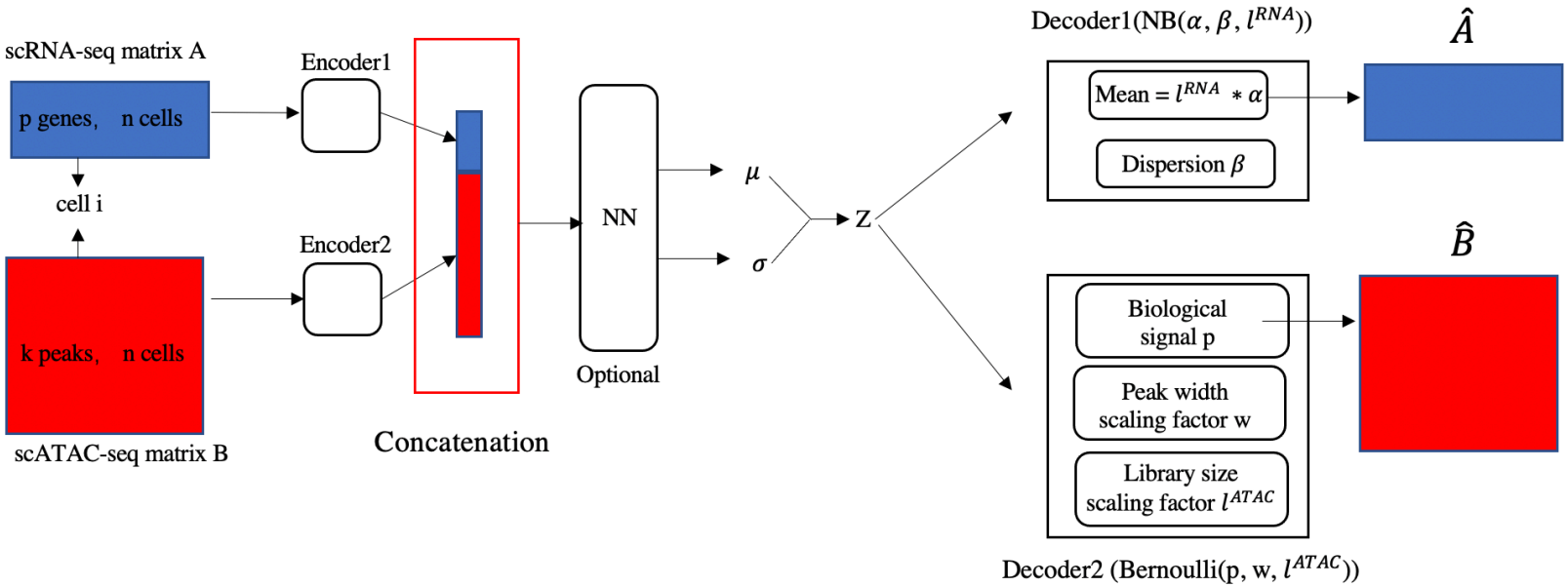


Figure: Model framework

Contents

Product of Expert(PoE) Model

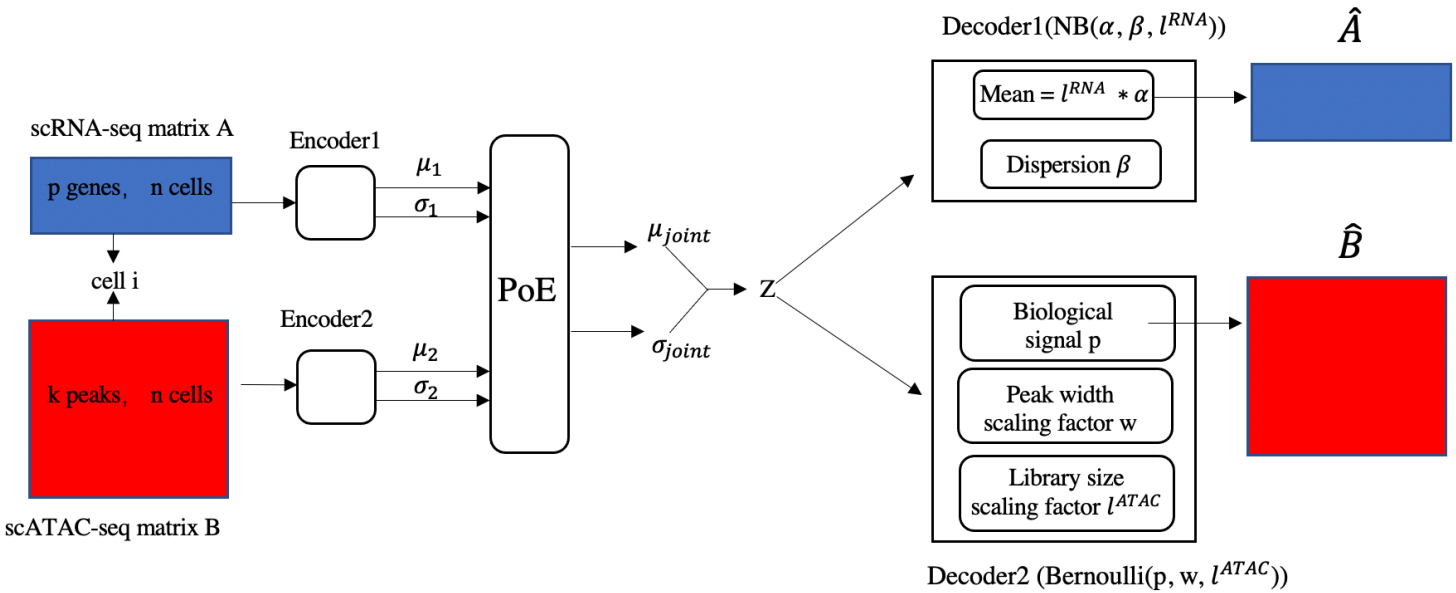


Figure: Model framework

PoE Model

PoE

Product of Experts (PoE) framework models the joint posterior as the product of the conditional marginal posteriors.

In our framework, it can be represented by

$$q_\phi\left(Z^{joint}|A,B\right) = q_{\phi_1}\left(Z|A\right) q_{\phi_2}\left(Z|B\right)$$

setting $q_{\phi_1}\left(Z|A\right)$ or $q_{\phi_2}\left(Z|B\right)$ to 1 if any one of them is missing.

PoE Model

$$q_\phi(Z^{joint}|X) \propto e^{-\frac{1}{2}(z-\mu_1)^T\sigma_1^{-1}(z-\mu_1)-\frac{1}{2}(z-\mu_2)^T\sigma_2^{-1}(z-\mu_2)}$$

$$\propto e^{-\frac{1}{2}z^T\sigma_1^{-1}z+\frac{1}{2}\mu_1^T\sigma_1^{-1}z+\frac{1}{2}z^T\sigma_1^{-1}\mu_1-\frac{1}{2}z^T\sigma_2^{-1}z+\frac{1}{2}\mu_2^T\sigma_2^{-1}z+\frac{1}{2}z^T\sigma_2^{-1}\mu_2}$$

$$\propto e^{-\frac{1}{2}z^T(\sigma_1^{-1}+\sigma_2^{-1})z+\frac{1}{2}(\mu_1^T\sigma_1^{-1}+\mu_2^T\sigma_2^{-1})z+\frac{1}{2}z^T(\sigma_1^{-1}\mu_1+\sigma_2^{-1}\mu_2)}$$

$$\propto e^{-\frac{1}{2}(z-(\sigma_1^{-1}\mu_1+\sigma_2^{-1}\mu_2)(\sigma_1^{-1}+\sigma_2^{-1})^{-1})^T(\sigma_1^{-1}+\sigma_2^{-1})(z-(\sigma_1^{-1}\mu_1+\sigma_2^{-1}\mu_2)(\sigma_1^{-1}+\sigma_2^{-1})^{-1})}$$

The paramters of joint distribution can be given by

$$\mu_{\text{joint}} = \left(\mu_1\sigma_1^{-1} + \mu_2\sigma_2^{-1}\right)\left(\sigma_1^{-1} + \sigma_2^{-1}\right)^{-1}$$

$$\sigma_{\text{joint}} = \left(\sigma_1^{-1} + \sigma_2^{-1}\right)^{-1}$$

ELBO

The ELBO(Evidence Lower Bound) can be given by

$$\mathcal{L} = -\beta KL(q_\phi(Z|A,B)||p_\theta(Z)) + E_{q_\phi(Z|A,B)}[\log p_1(A|Z)] + E_{q_\phi(Z|A,B)}[\log p_2(B|Z)]$$

Contents

Structure-RNA

Experiment/Parameter	Divide mean of NB into library size * proportion	Fixed library size	Library size learnt by NN	dispersion parameter of NB is constant per gene across cells	Normalization for input data	Add scaling factor for decoder
1	X	✓		X	X	X
2	X				✓	X
3	X				✓	✓
4	X		✓	✓		
5	X		✓	✓	✓	
6	✓	✓		✓		
7	✓		✓	✓		

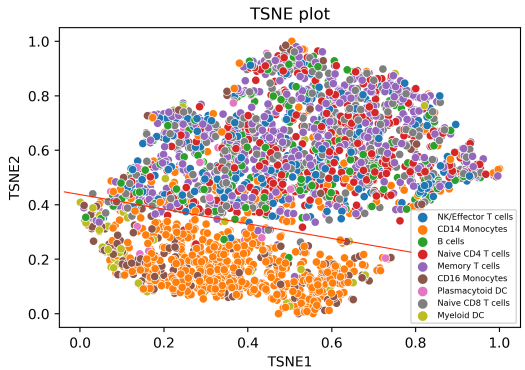
Table: Experiments

Notes:

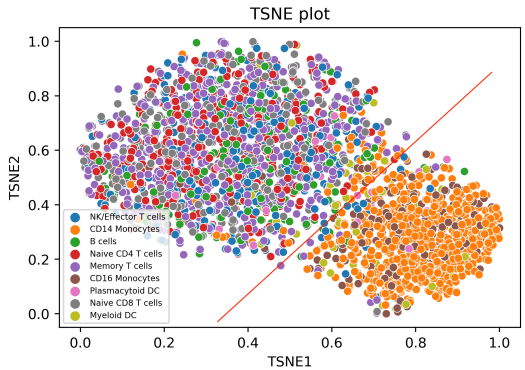
- Main difference between 6, 7 and others is whether dividing mean paramter of NB into library size * proportion
- Difference between 6 and 7 is whether using fixed library size

Result

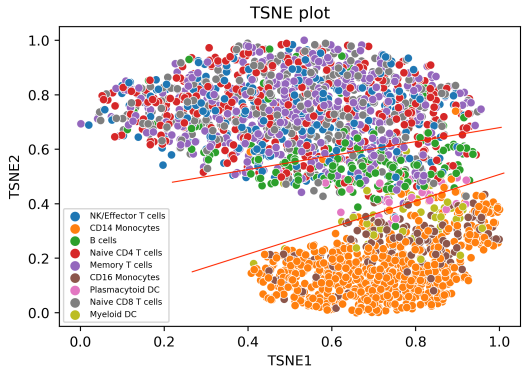
I just use result of PBMC 3k data for illustration. The plots are got by latent space of the model. (Each plot corresponds to structure in last slide)



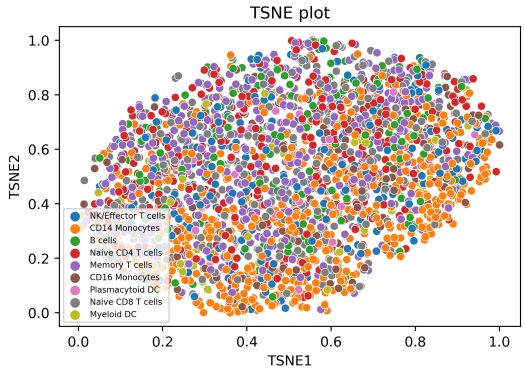
(a) Exp1



(b) Exp2

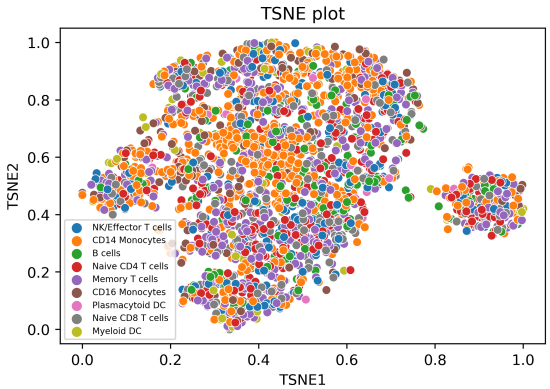


(c) Exp3

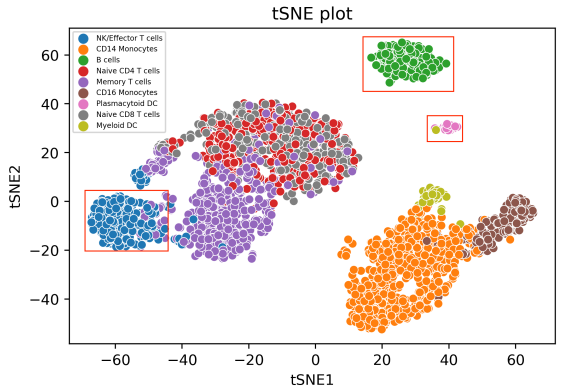


(d) Exp4

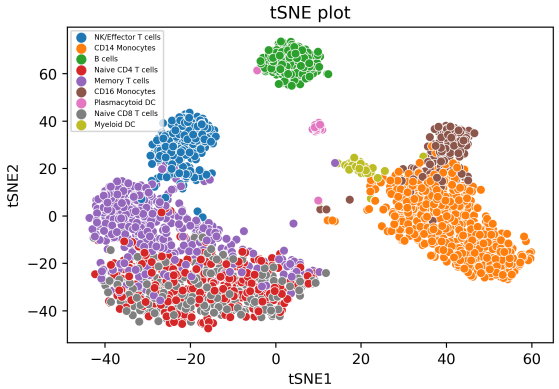
Result



(a) Exp5



(b) Exp6

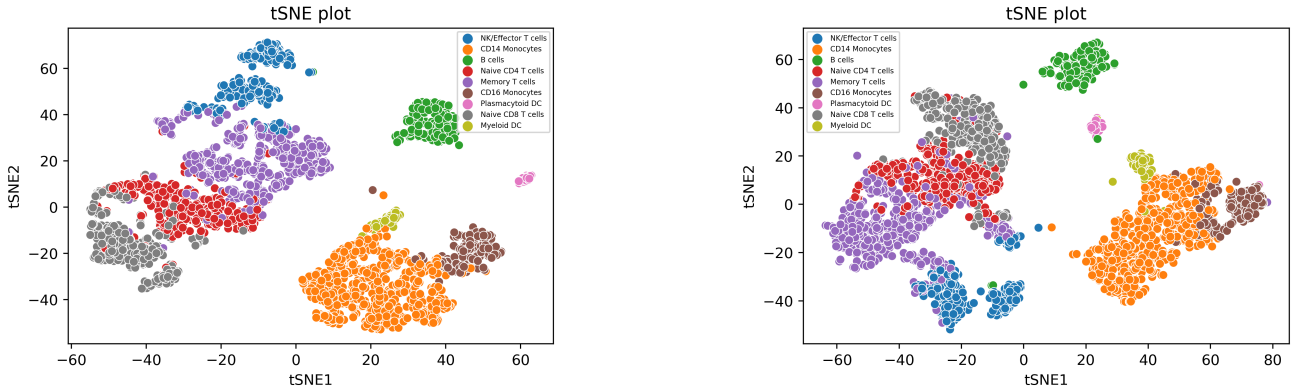


(c) Exp7

Figure: Exp5 - 7

Variable in all out

Data of some genes may contain more noises than signals.



(a) Variable in all out

(b) All in all out

Figure: TSNE plot for latent space

- The procedure of selecting variable genes is totally the same with Seurat(Top 5, 000 genes are selected)
- Same or better performance
- Less parameters, higher efficiency

Final Structure

Final Structure for RNA part

- Use NB as default decoding distribution(UMI or not UMI, that is the question for scRNA-seq zero-inflation, Nature Biotechnology2021)
- Use fixed library size(sum of counts) as default setting and provide option of library size learnt by NN
- Use proportion * library size to be mean of NB
- Dispersion paramter is fixed per gene across cells
- Use variable genes as input and output all genes for imputation

Structure-ATAC

probability of whether a peak region$_i$ is open in a cell$_j$ $= p_{ij} * w_i * l_j^{ATAC}$

- p_{ij}: true probability of whether a peak region is open in one cell
- w_i: scaling factor to capture the variance of peak region width(peak region specific)
- l_j^{ATAC}: scaling factor to capture the variance of library size(cell specific)

Structure-ATAC

We proposed 3 scaling factors for library size.

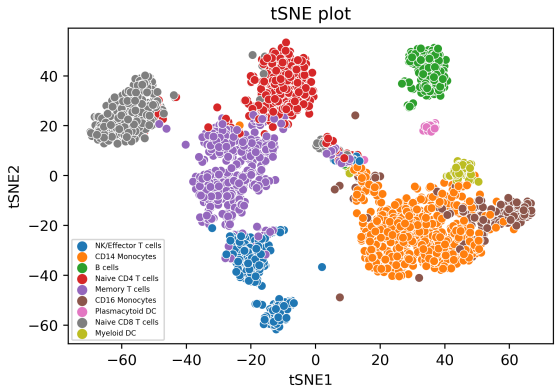
- Linear scaling factor: library size$_i$/max library size
- Nonlinear scaling factor: z score transformation + sigmoid

$$\text{normalized library size}_i = \frac{\text{library size}_i - \text{mean(library size)}}{\text{sd(library size)}}$$

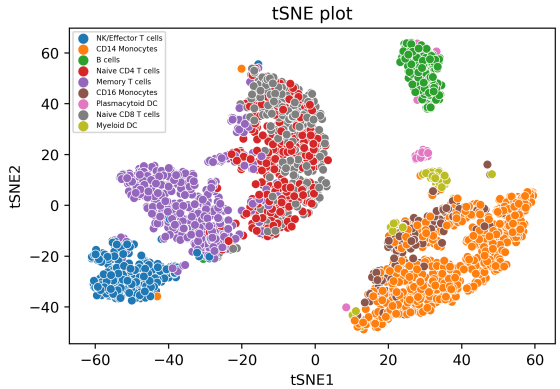
$$\text{nonlinear scaling factor}_i = \frac{1}{1 + e^{-\text{normalized library size}_i}}$$

- Scaling factor learnt by NN(sigmoid)

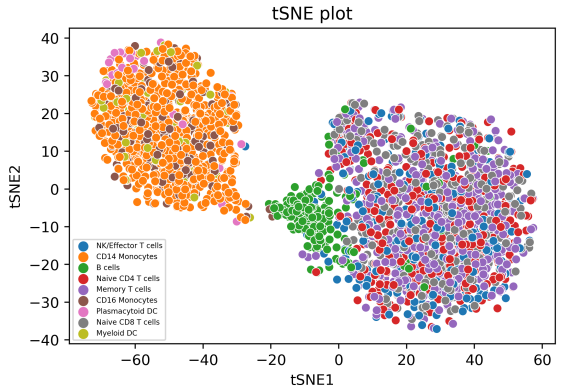
Result



(a) Nonlinear scaling factor



(b) NN



(c) Linear scaling factor

Figure: TSNE plot for latent space

Tricks

Common tricks for VAE model:

- Log transformation(Make the gradient more stable)
- KL divergence warm up(How to Train Deep Variational Autoencoders and Probabilistic Ladder Networks. ICML2016)

$$-\beta KL\left(q_\phi(z \mid x) \| p_\theta(\mathbf{z})\right) + E_{q_\phi(z|x)}\left[\log p_\theta(\mathbf{x} \mid \mathbf{z})\right]$$

  instead of optimizing ELBO directly
- Early stopping

Tricks for PoE model

There are at least 3 training roots in total. RNA in RNA, ATAC out; ATAC in RNA, ATAC out; RNA, ATAC in RNA, ATAC out;

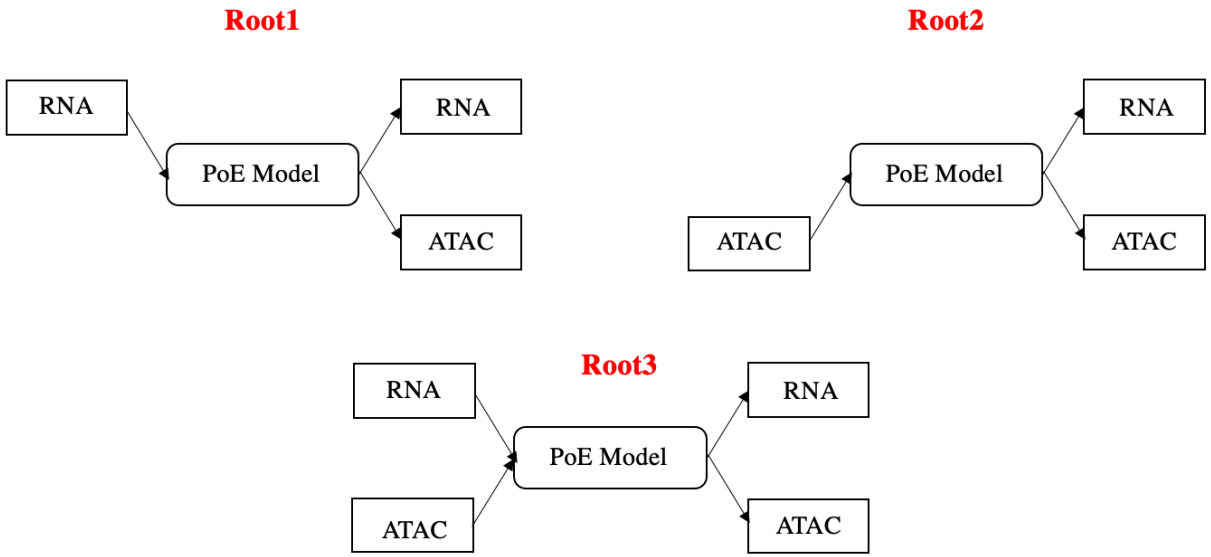


Figure: 3 roots

Tricks for PoE model

Ob1

$$-\beta KL(q_\phi(Z|A)||p_\theta(Z)) + E_{q_\phi(Z|A)}[\log p_1(A|Z)] + E_{q_\phi(Z|A)}[\log p_2(B|Z)]$$

Ob2

$$-\beta KL(q_\phi(Z|B)||p_\theta(Z)) + E_{q_\phi(Z|B)}[\log p_1(A|Z)] + E_{q_\phi(Z|B)}[\log p_2(B|Z)]$$

Ob3

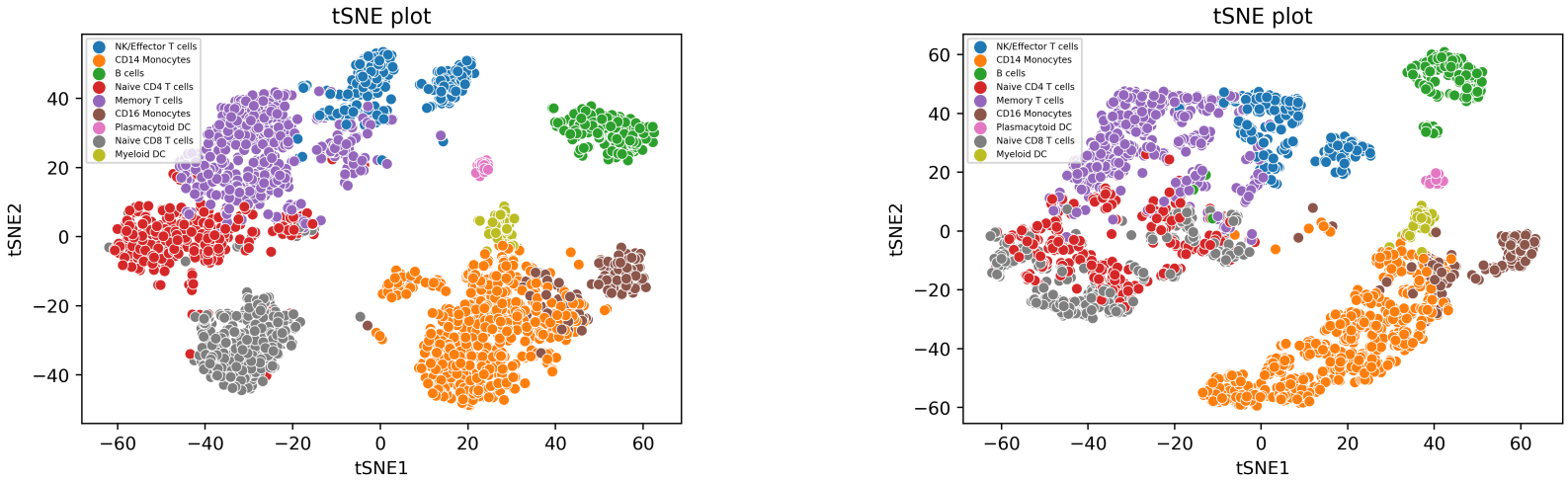
$$-\beta KL(q_\phi(Z|A,B)||p_\theta(Z)) + E_{q_\phi(Z|A,B)}[\log p_1(A|Z)] + E_{q_\phi(Z|A,B)}[\log p_2(B|Z)]$$

For each epoch, the model will randomly select one objective to optimize. Ob1 and Ob2 can let the model learn how to predict data of another modality when the input is only one modality.

Tricks for PoE model

- Design a quantitive criterion to evaluate model fitting(We use plot of latent space to evaluate now)
- Use cross validation to determine the probability of choosing each Ob(Now the probabilities for each Ob are equal)

Initial result



(a) PoE (b) Concatenation

Figure: TSNE plot for latent space

Heatmap for latent space

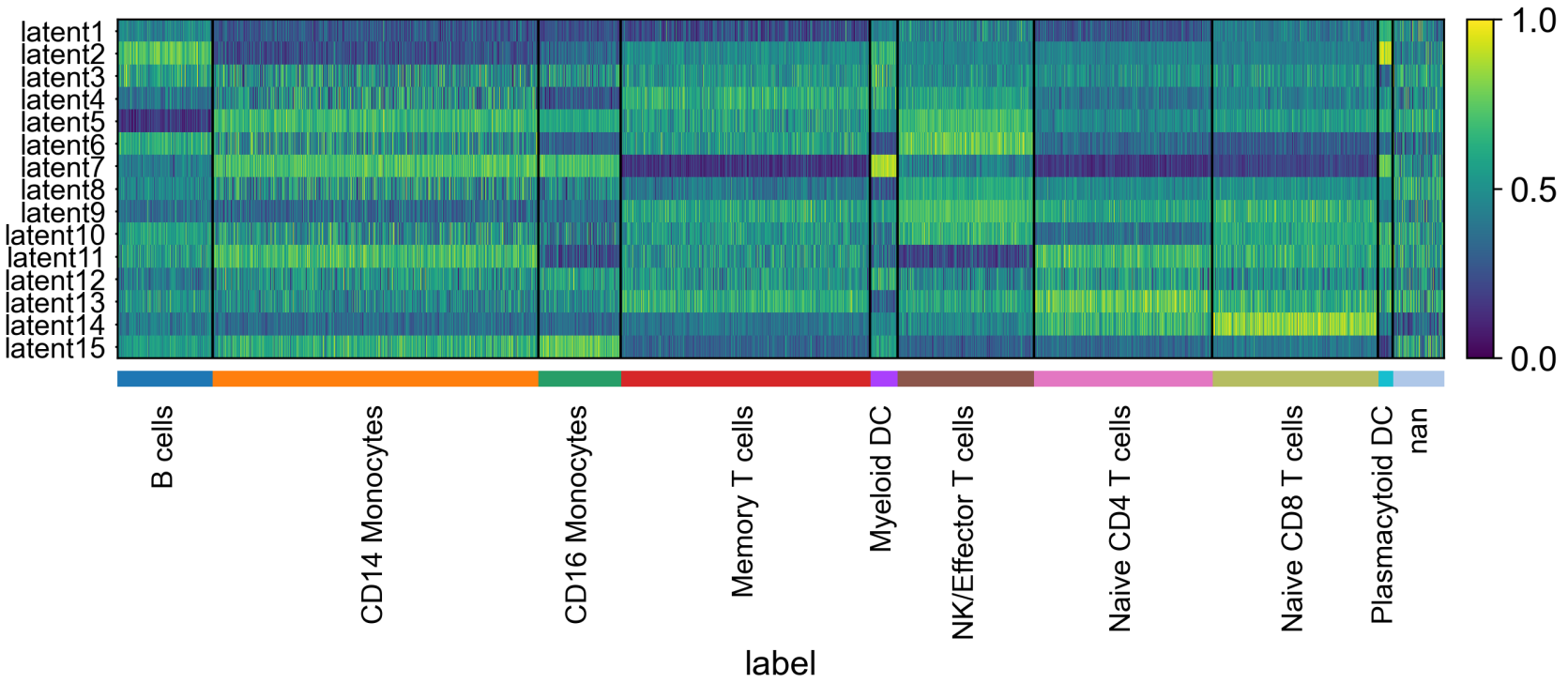
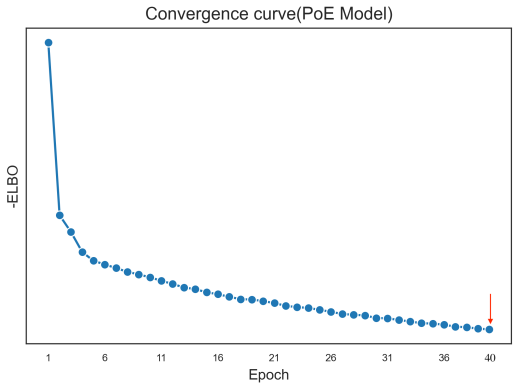


Figure: Heatmap for latent space

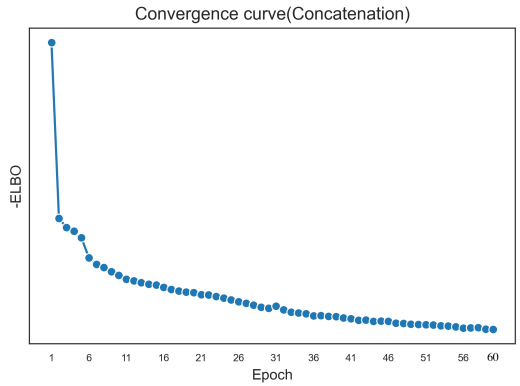
There are different patterns for different cell types.

Contents

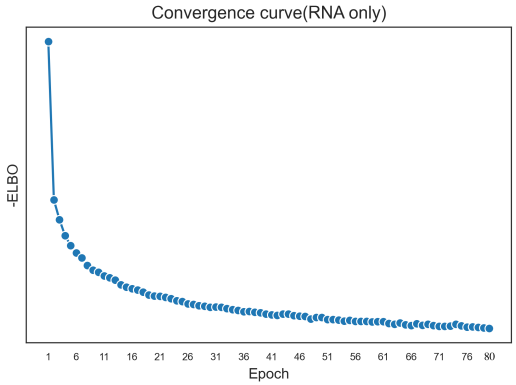
Convergence



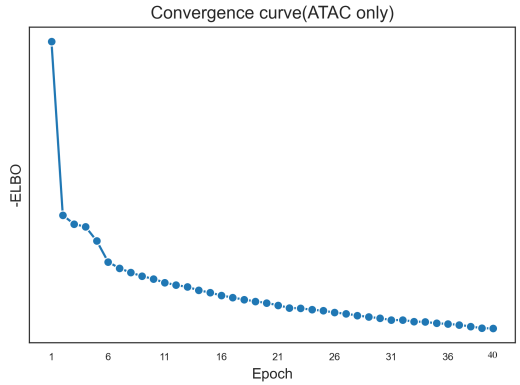
(a) PoE Model



(b) Concatenation Model



(c) Only RNA



(d) Only ATAC

Figure: Convergence Curve

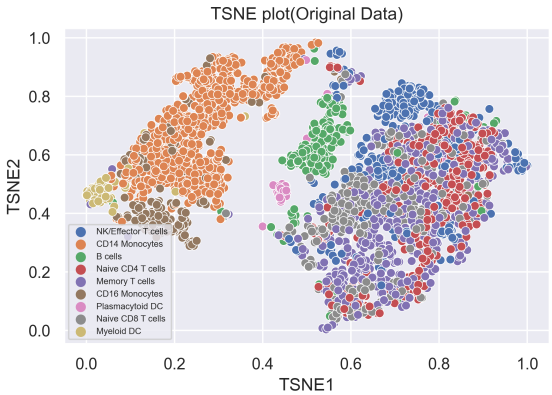
Contents

RNA-seq

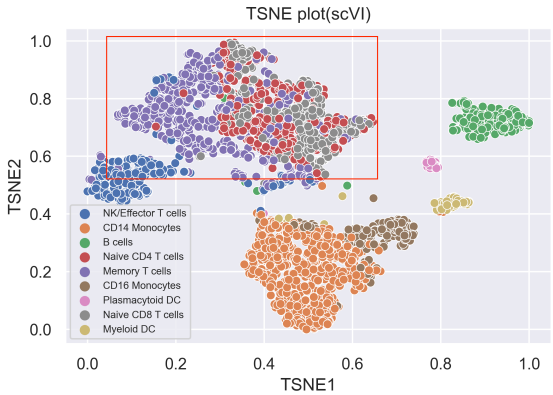
Benchmark Models: SAVER, scVI, MAGIC
Metrics: Low dimension embedding, Unsupervised clustering, Correlation with bulk sample

Low dimensional embedding-tSNE

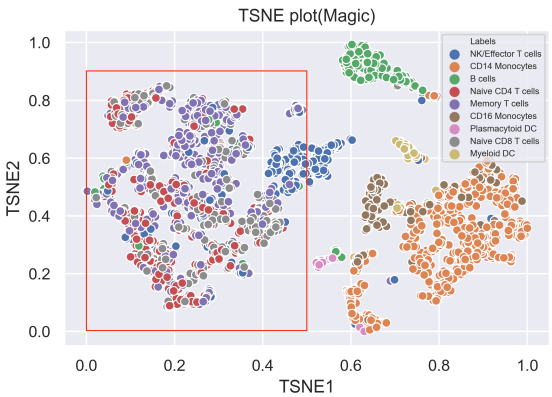
I chose **tSNE** and **UMAP** to get low dimensional embedding for imputed data.



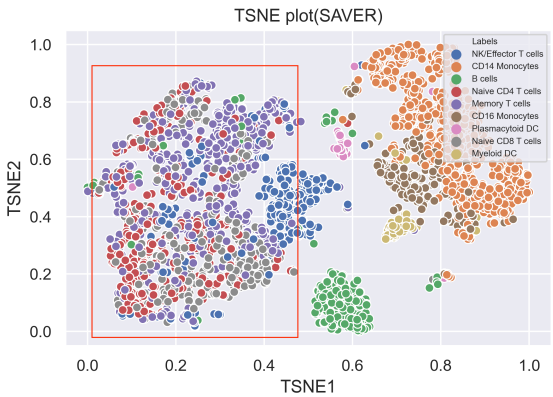
(a) Original data



(b) scVI

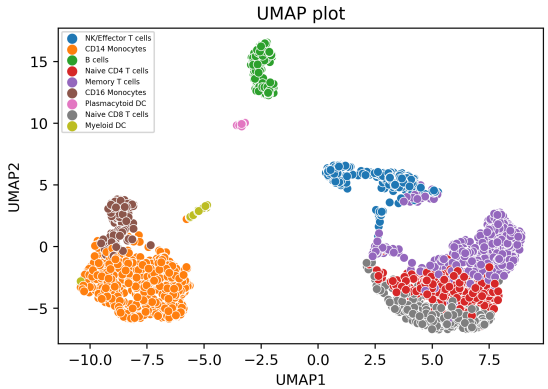


(c) Magic

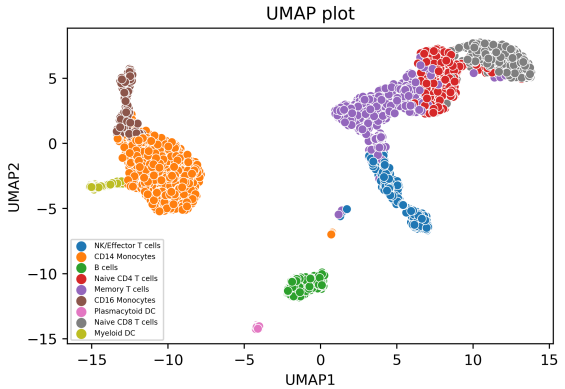


(d) SAVER

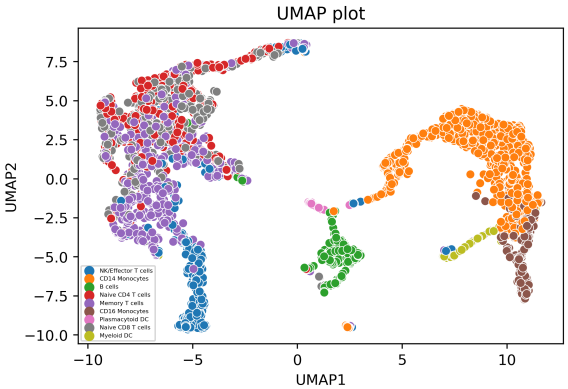
Low dimensional embedding-UMAP



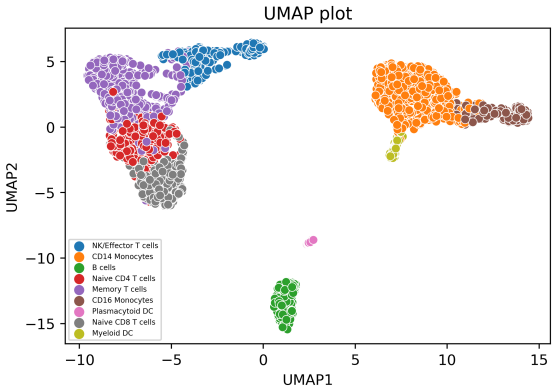
(a) Original data



(b) SAVER



(c) MAGIC



(d) scVI

Figure: UMAP plot

Correlation with bulk samples

I downloaded bulk RNA-seq data from **ENCODE**. The details are listed in below table. (All samples are sequenced by Paired-end protocols.)

Cell type	**File name**	**GEO Number**
B cell	ENCFF195UAA.tsv	GSM220576
T cell	ENCFF038KPM.tsv	GSM220574
CD14 Monocytes	ENCFF248RZF.tsv	GSM1220575

Table: Bulk data information

Correlation between bulk and pesudo-bulk

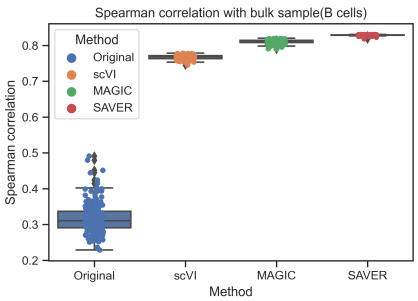
The SCC is calculated comparing the bulk sample and pesudo-bulk imputed scRNA-seq profiles.

	Original	scVI	SAVER	MAGIC
B cell	0.806	0.793	0.831	0.824
CD14	0.834	0.817	0.846	0.840
T cells	0.836	0.823	0.848	0.848

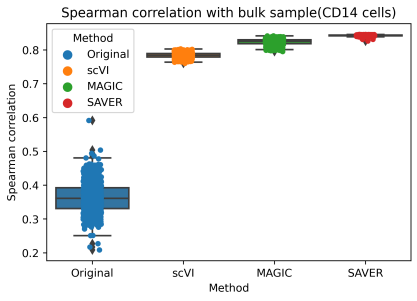
Table: Scc statistics

Result

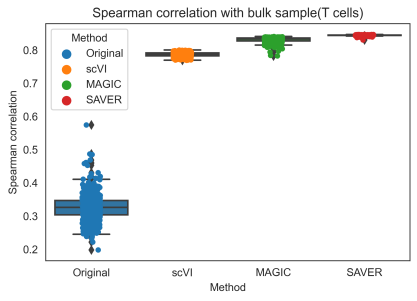
I used **Spearman Correlation** to evaluate the correlation between bulk data and imputed single cell data.



(a) B cells



(b) CD14



(c) T cells

- SAVER behaves best among all benchmark methods
- How to evaluate within cluster variance? Too high correlation may not be good. (Loss of variance)

Correlation with difference

The SCC is calculated comparing the difference (log fold change or LFC) in two bulk cell type profiles compared to two pesudo-bulk imputed scRNA-seq cell type profiles.

	Original	scVI	MAGIC	SAVER
T vs B	0.083	0.069	0.091	0.061
B vs CD14	0.390	0.337	0.385	0.390
T vs CD14	0.292	0.289	0.305	0.266

Table: Spearman Correlation Coefficient

- Imputation methods don't behave well in learning cell type difference

Unsupervised clustering

I chose **K-means**, **Louvain** as the methods of unsupervised clustering and used **ARI** as index for evaluation.

- For SAVER and MAGIC, I do clustering in top 20 pcs(explained variance > 0.99)
- For scVI, I do clustering on both latent space and top 20 pcs

	scVI(latent)	scVI(Imputed)	SAVER	Magic
K-means	0.684	0.38	0.238	0.289
Louvain	0.627	0.41	0.32	0.349

Table: Evaluation for unsupervised clustering

Contents

ATAC-seq

Benchmark Models: SCALE, RA3, PeakVI
Metrics: Low dimensional embedding, Correlation with bulk ATAC-seq or DNase-seq

Correlation with bulk ATAC-seq or DNase-seq

Again, I downloaded bulk ATAC-seq or DNase-seq data from **ENCODE**.

Cell type	**GEO Number**
T cell	SAMN18514556
B cell	SAMN18514446
CD14	GSE169981

Table: Bulk data information

Contents

Only RNA Expert

If we only have scRNA-seq data, the posterior for ATAC expert is viewed as 1.

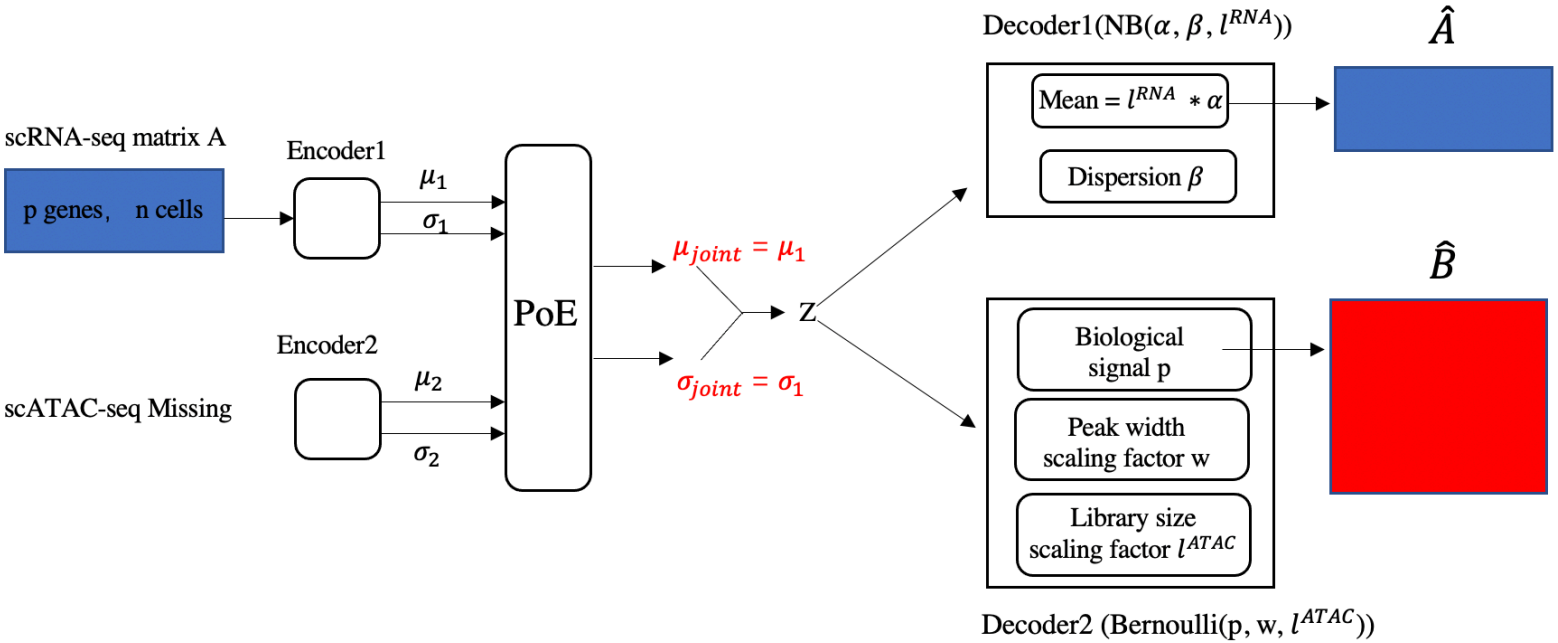


Figure: RNA Expert Only

Only ATAC Expert

If we only have scATAC-seq data, the posterior for RNA expert is viewed as 1.

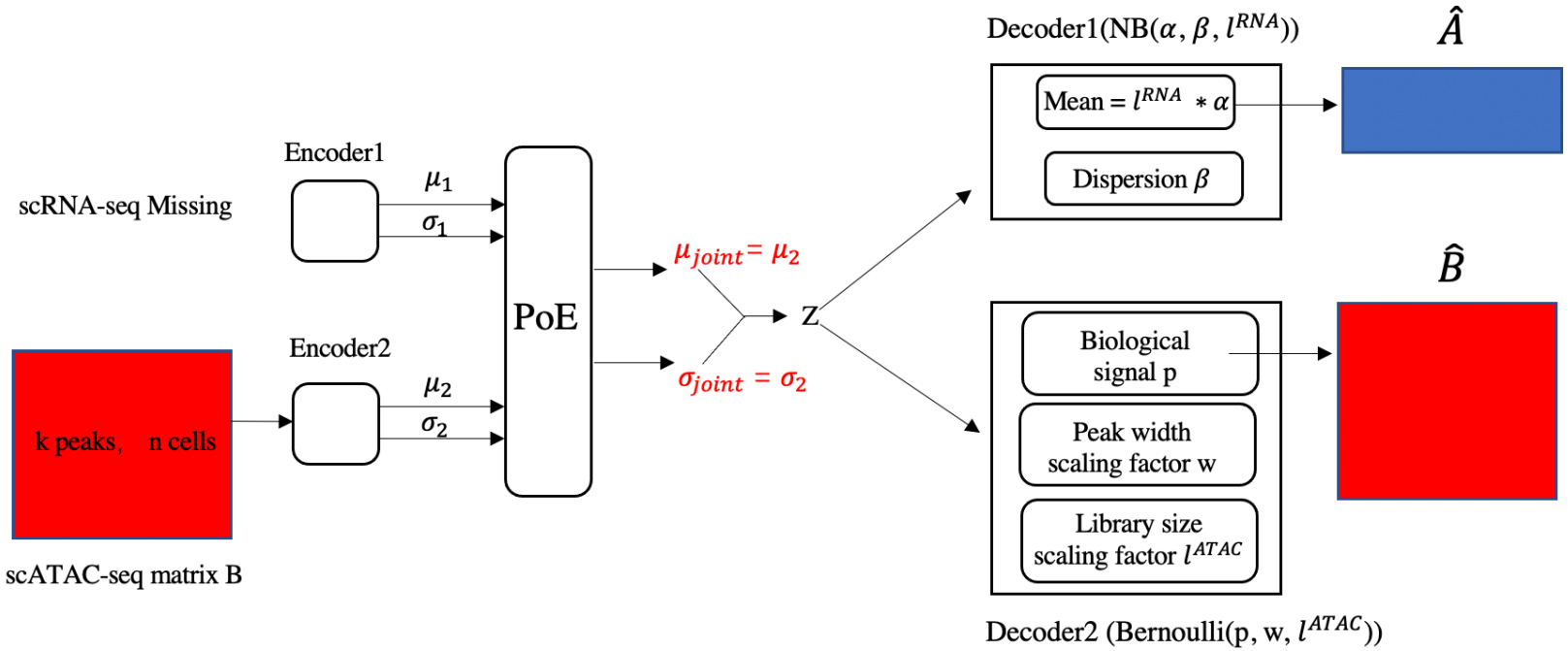


Figure: ATAC Expert Only

Future work

- Test model in more datastes(Especially multi-omics data from other protocols)
- Utilize relation between gene and peak region
- Generalize the model to RNA + X(i.e. RNA + Methylation)