# Deep Generative Modeling for Single-cell Transcriptomics

Romain Lopez

University of California, Berkeley

Slide credits: Jeffrey, Nir, Adam, Chenling & Romain

# How do you define normalization?

Algorithmic query:

- Normalization aims at providing a value (or a distribution) **for each individual entry of the gene expression matrix** which satisfies some statistical properties of interest.

Motivation:

- Normalization as a **computational artifact** to make the data Gaussian, and amenable to standard machine learning algorithms (i.e., for PCA, CCA, some autoencoders etc.);
- Normalization as a way to **control for covariates**, which are either technical artifacts or unwanted biological signal (RUV and others);

# How do you normalize?

We posit a **generative model** (scVI & HCV) for the gene expression counts $x_{ng}$ of a cell $n$, with unwanted covariate $s_n$ and a gene $g$ is

$$
\begin{aligned}
z_n &\sim \mathrm{Normal}(0, I) && \text{Cell embedding} \\
\ell_n &\sim \mathrm{LogNormal}(\ell_\mu, \ell_\sigma^2) && \text{Library size} \\
\rho_n &= f_w(z_n, s_n) && \text{Normalized expression} \\
\pi_n &= f_h(z_n, s_n) && \text{Dropout rate} \\
x_{ng} &\sim \mathrm{ZINB}(\ell_n \rho_{ng}, \theta_g, \pi_{ng}) && \text{Raw data}
\end{aligned}
$$

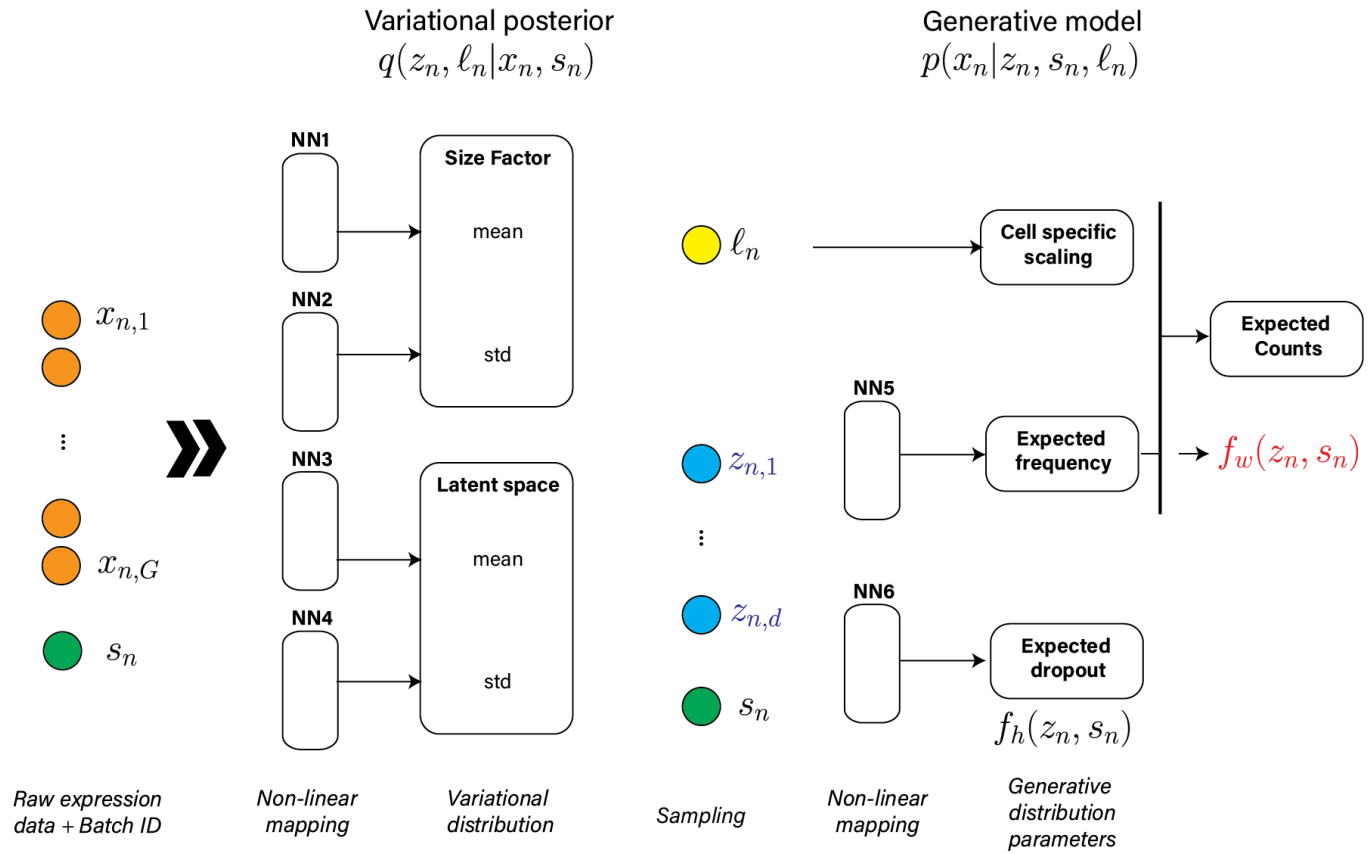
where $f_w$ and $f_h$ are two neural networks. $z_n$ is made **invariant** to $s_n$ as well as $\ell_n$.

Lopez et al., *Nature Methods*, 2018 & Lopez et al. *NeurIPS*, 2018

# How do you normalize?

Inference can be done with Auto-encoding Variational Bayes!

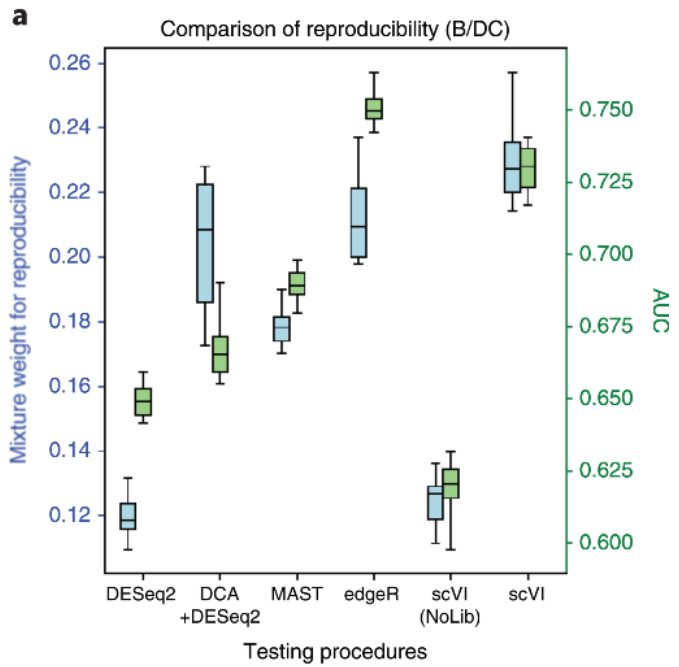
All CompBio tasks are well defined!




Lopez et al., *Nature Methods*, 2018

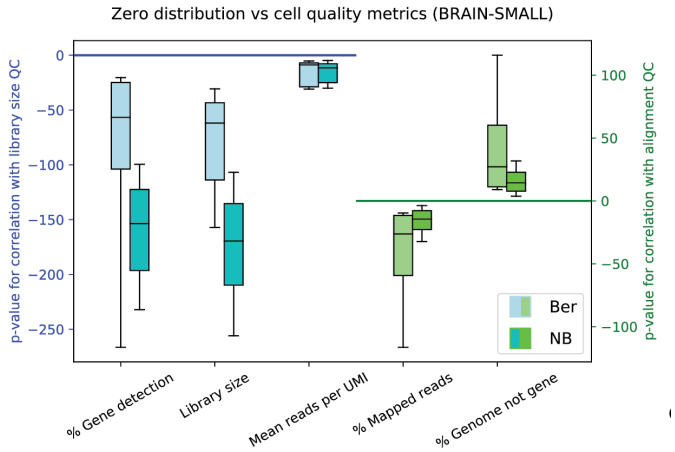
# How do you demonstrate success?

- The latent space is **more correlated with biological information** (i.e., cell types);
- The latent space is **less correlated with quality control metrics** (i.e., library size or sequencing errors);
- Differential expression picks up **more reproducible genes**



# Latent variable $\pi_n$ both captures sequencing errors and transcriptional bursting

## Correlation of $\pi_n$ with quality control metrics in scVI




## AutoZI: spike and slab prior

$$\delta_g \sim \text{Beta}(\alpha, \beta)$$

$$m_g \sim \text{Bernoulli}(\delta_g)$$

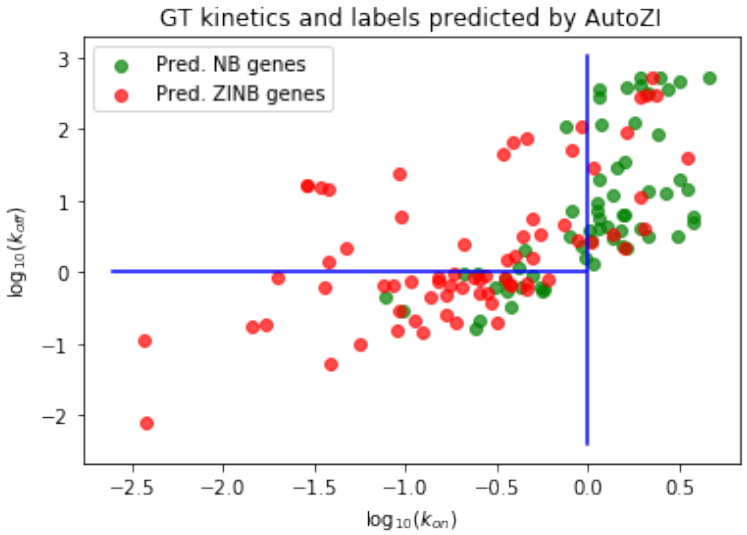
$$\pi_{ng} = (1 - m_g) f_h(z_n, s_n)$$

AutoZI: Clivio et al. 2019

## Empirical findings of AutoZI

- ERCCs are not ZI
- Biological genes are mainly ZI

## Bimodality of stochastic gene expression might recover ZI genes



# Where does your method break?

- Learning this model requires a certain number of cells (or gene filtering);
- Adding gene-specific variables for accounting for gene length bias for example is not straightforward;
- It can be **hard to perform disentanglement** (in case of general removal of unwanted variation) with Auto-encoding Variational Bayes;

# What is your suggestion for the second day of the workshop?

1. Extending model-based selection methods (AutoZI) for a data-driven choice of conditional distribution? How do you show the NB is better than lognormal?
2. How to further constrain the models to avoid overfitting or under-interpretability? Where is the line – if there is one?
3. Introduce new terms / communicate about current misusage? a) imputation b) denoising c) smoothing d) normalization

https://yoseflab.github.io/2019/06/25/ZeroInflation/

# Open-source scientific research

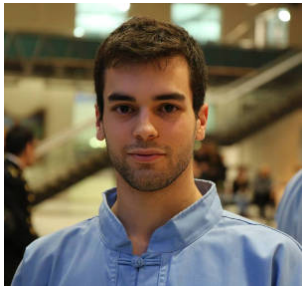
# scVI is a public repository

http://www.github.com/YosefLab/scVI

- Our codebase is maintained and contains the software scVI, scANVI, gimVI and totalVI (more to come !) as well as tutorials;
- The codebase is modular and research oriented. It is simple to quickly create **novel research outcomes** (several manuscripts from outside of our team);
- Feedback and utilisation from academia and industry;

Come contribute !

# The scVI collaboration



Romain Lopez

Chenling Xu

Adam Gayoso

Jeff Regier

Mike Jordan

Nir Yosef

& Maxime Langevin, Edouard Melhman, Jules Samaran, Achille Nazaret, Gabriel Misrachi, Oscar Clivio, Pierre Boyeau, Yining Liu