LUXEMBOURG INSTITUTE OF HEALTH

Journal Club: Learning interpretable cellular and gene signature embeddings from single-cell transcriptomic data

P. Nazarov

2022-05-04

Bioinformatics Seminars

LUXEMBOURG INSTITUTE OF HEALTH

ARTICLE

Check for updates

https://doi.org/10.1038/s41467-021-25534-2 OPEN

Learning interpretable cellular and gene signature embeddings from single-cell transcriptomic data

Yifan Zhao[1,5,6], Huiyu Cai[2,6], Zuobai Zhang[3], Jian Tang[4]✉ & Yue Li[1]✉

Single cell data → Single-cell Embedded Topic Model (scETM) = VAE + tri-factorization

Features:

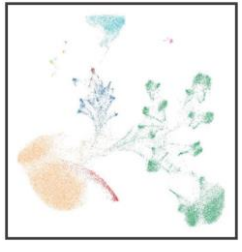
- clustering (transferable!)
- batch effects taken into account
- explainable
- scalable (> 10^6 cells)

The idea originated from NLP:

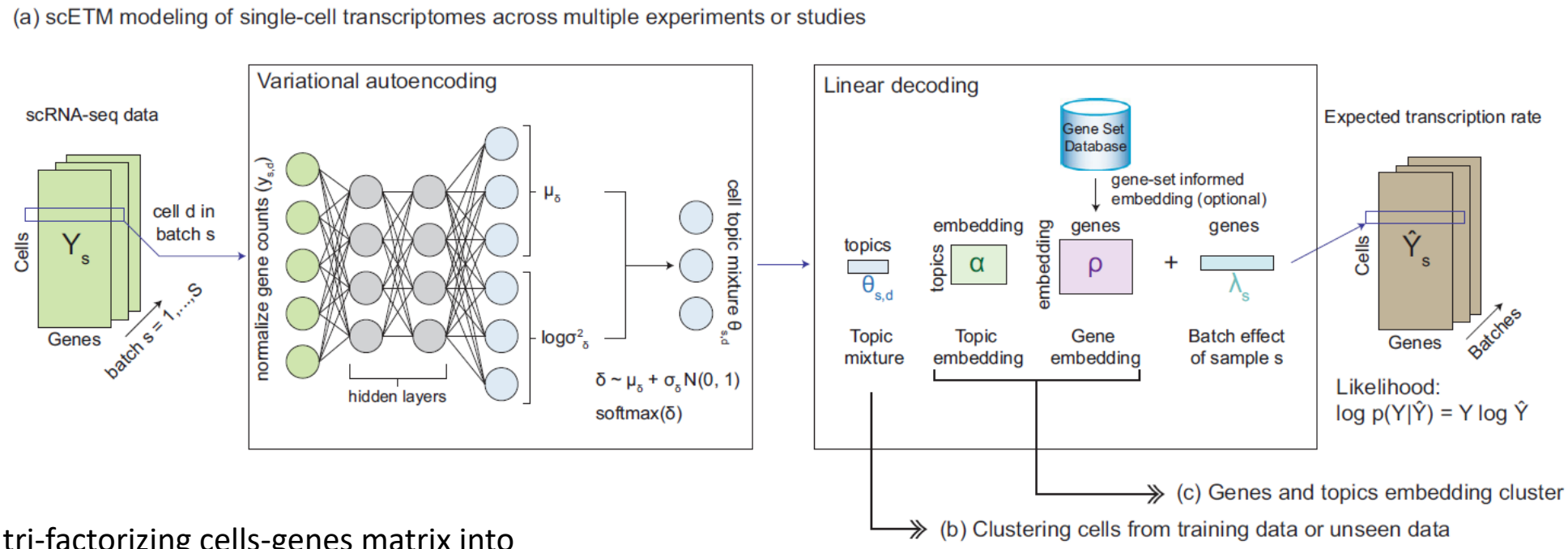
each cell → document

gene that gives rise to the read → a word

Assumption: a cell transcriptome can be represented as a mixture of "latent cell types"



General Approach: VAE

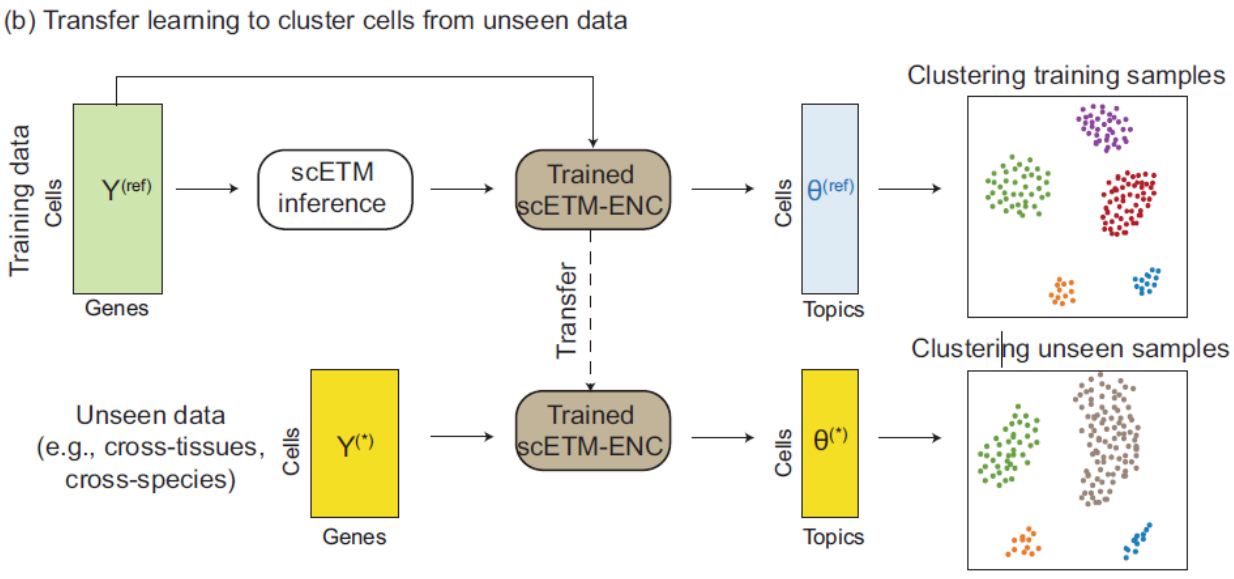


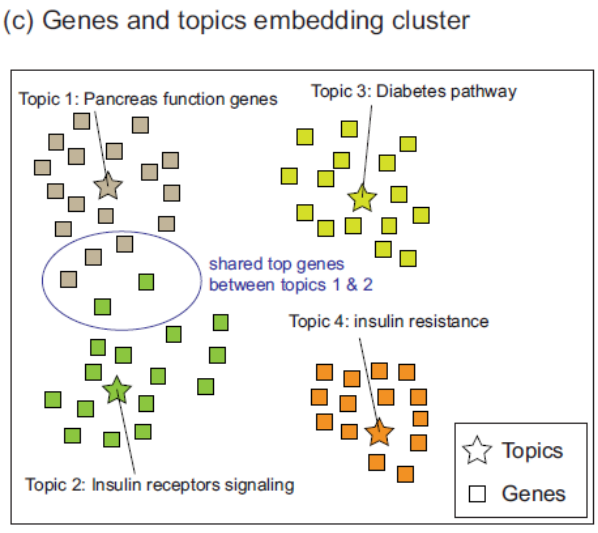
By tri-factorizing cells-genes matrix into

- cells-by-topics θ
- topics-by-embeddings α
- embeddings-by-genes ρ

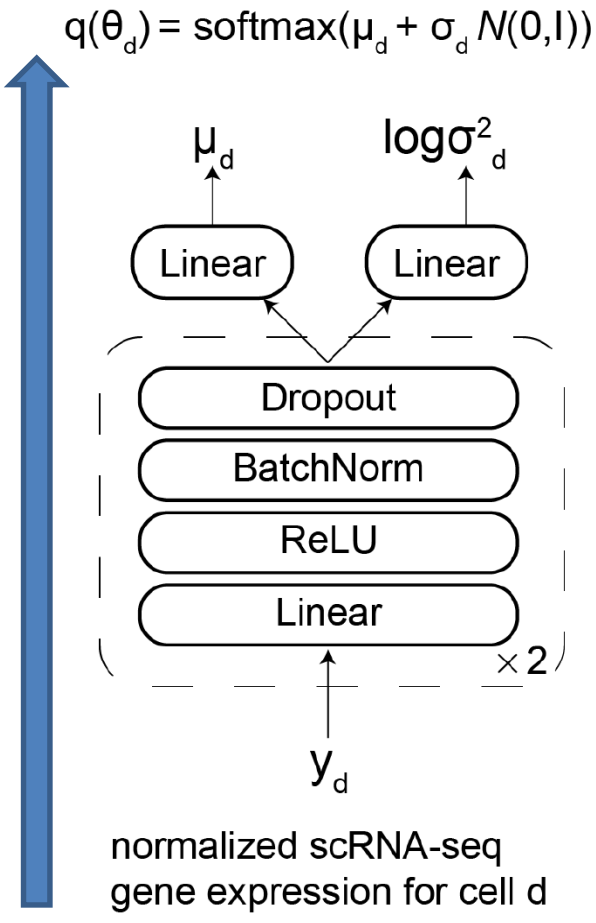
they are able to incorporate existing pathway information into gene embeddings ρ during the model training to further improve interpretability.

"Transfer learning" for clustering





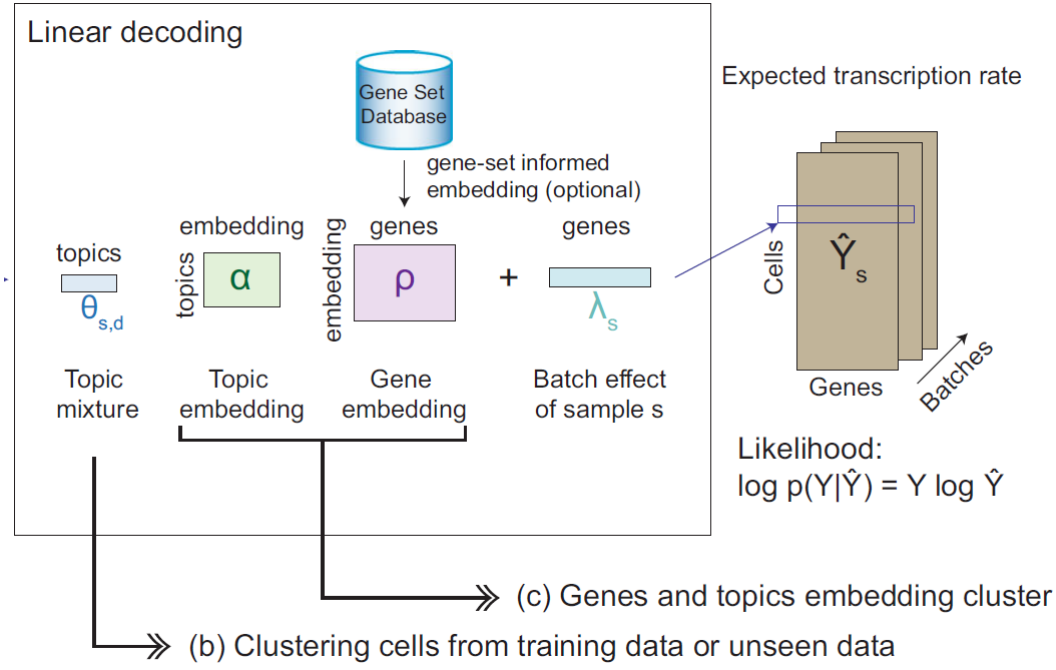
Encoder



- 2-layer neural-network
- hidden sizes of 128, ReLU activations
- 1D batch normalization
- 0.1 drop-out rate between layers.

The gene embedding dimension to 400, and the number of topics to 50. Adam Optimizer and a 0.005 learning rate.

Linear Decoder



Genes embeddings:

- similar to gene2vec – present the genes in the space where genes from similar gene sets (e.g. GOs) are located together

Gene2vec embeddings

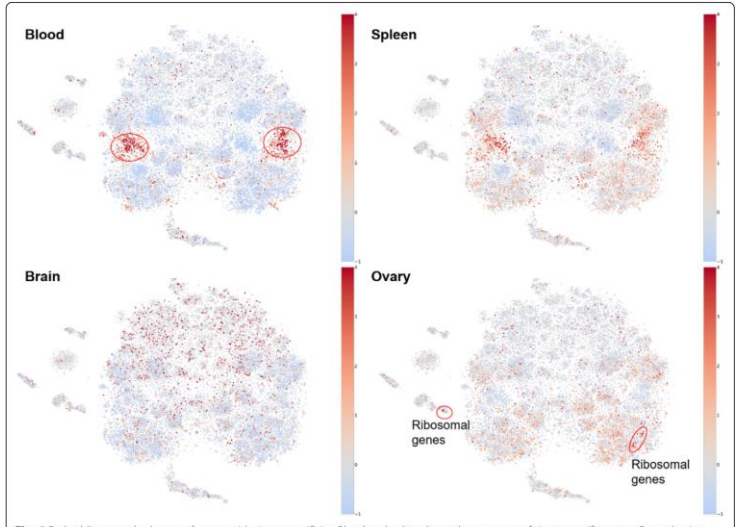
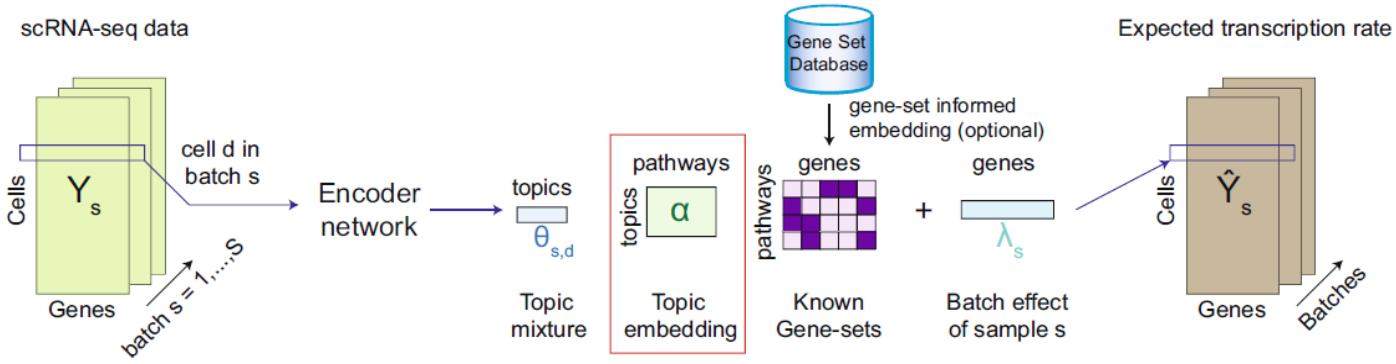


Fig. 4 Embedding reveals clusters of genes with tissue-specificity. Blood and spleen have clear patterns of tissue-specific genes. Reproductive system (e.g., ovary) also showed distinguished genes. Genes not available in GTEx data were colored grey

Linear Decoder: Topic Embeddings



Batch correction: Mouse Retina

Figure 2 Integration and batch correction on the Mouse Retina dataset. Each panel shows the Mouse Retina cell clusters using UMAP based on the cell embeddings obtained by each of the 9 methods. The cells are colored by cell types in the first two rows and by batches, which are the two source studies, in the last two rows.

kBET: k-nearest-neighbor Batch-Effect Test (low is good)

ARI: Adjusted Rand's index (high is good)

Batch correction: Mouse Pancreas

Figure 3: Integration and batch correction on the Mouse Pancreas (MP) dataset. Each panel shows the MP cell clusters using UMAP based on the cell embeddings obtained by each of the 9 methods. The cells are colored by cell types in the first two rows and by batches, which are the two mouse strains, in the last two rows.