Deep generative single-cell embedding models

Jacob Krol

Course CSCI 5922, University of Colorado Boulder

1 Introduction

1.1 Motivation

Single-cell sequencing technologies measure the biological state of cells by counting the number of ribonucleic acid (RNA) transcripts (called RNA-seq), quantifying chromatin accessibility (called ATAC-seq), and more. Studying cell states is useful to characterize disease by differential normal/diseased-sample analysis, finding rare cell types, mapping unknown cell types to their most likely cluster (called "reference mapping"). Despite the research efforts to collect multi-omics single-cell data [1], data analysis faces significant challenges due to the high dimensionality of the RNA-seq and ATAC-seq modalities, plus data artifacts caused by experimental setup (called "batch effects"). To address this, recent research is investigating the deep generative models to learn cross-modality, low dimensional representations (embeddings) of cells.

1.2 Gap

Non-neural network methods were developed to learn cross-modal embeddings of RNA-seq and ATAC-seq, but these methods have common downsides: no batch effect correction, computational cost, and do not generalize to new modalities. As mentioned, neglecting batch effects can cause cell embeddings to cluster by experimental protocols rather than true biological signals. Furthermore, large single-cell databases house millions of cells produced by different experimental protocols. Therefore, cell embedding models must be scalable to millions of cells measured by diverse experimental setups. Often, not all modalities (RNA-seq, ATAC-seq, CITE-seq, etc.) are available for each cell which further motivates models capable of modality imputation and still projecting each cell into a shared embedding space.

1.3 Proposal

This review discusses, compares, and reproduces the results of state-of-the-art deep generative models for multi-modal single-cell representation learning.

The goal is to find which generative models are best at multi-modal single-cell representation learning based on reproducing experiments from recent publications. Experiments will quantify the quality of single-cell representations, and the model's ability to overcome batch effects. The results presented will give

insight on the pros/cons and shared/distinct attributes of state-of-the-art deep, generative cell embedding models. After establishing a foundation of experimental results, the current limitations and trends will be discussed to offer future research directions.

2 Related Work

2.1 Topic 1: generative neural networks for learning multi-modal single-cell embeddings

Related work uses variational autoencoder (VAE) models as an unsupervised approach to learn low dimensional representations of single-cell data. Namely, Cobolt [2] and MultiVI [3] produce embeddings of arbitrary single-cell multimodal data, but in different fashions.

Cobolt receives input cells with positive count data features (from arbitrary sequencing modalities) and outputs a latent representation. Cobolt's architecture is a VAE where the encoder neural network produces the mean and log-variance parameters for a softmax-transformed multivariate Gaussian. However, the decoder component includes no neural network and instead assumes a multinomial generative model, used during training as the generative distribution for the evidence lower bound objective (ELBO) function. Cobolt's encoder has a single hidden layer before outputting the variational distribution parameters.

Meanwhile, MultiVI's VAE uses distinct encoders and decoders for each modality, and the joint representation is acquired by taking the mean of each modality's encoder representations. The variational distribution is also a multivariate Gaussian, and the ELBO training objective is used. However, the generative distribution was chosen by hand for each modality in the paper: negative binomial for RNA-seq and Bernoulli (simplified to accessible or not 0,1) for ATAC-seq.

The primary model of interest for this review, MultiDGD [4], differs from Cobolt and MultiVI since it is a decoder-only model and assumes that latent single-cell data is from a mixture of multivariate Guassians.

2.2 Topic 2: VAEs versus decoders

[5] introduced VAEs and showed that stochastic gradient descent is an efficient alternative to computationally expensive maximum likelihood-based methods for estimating the parameters of both latent variable distributions and their generative counterparts. However, recent works suggest using the representations directly from the decoder and excluding the encoder can improve training data efficiency [6].

MultiDGD is a decoder-only model and was experimentally shown to be more data efficient than MultiVI, a VAE. And multiDGD produced embeddings more similar to input cells and latent representations were less clustered by batch.

3 Methodology

To find models for reviewing single-cell multiomics representation learning, traditional and AI-assisted search engines were queried with the key words of "generative", "neural network", "single-cell", "multiomics", "deep", and related terms.

For a focused evaluation on generative models for single-cell multiomic representation learning, papers and models were filtered by the following criteria:

- The model must have a neural network component.
- The model training objective must be generative, not discriminative.
- The model must support at least two single-cell modalities.
- The model must be capable of outputting an embedding, given multi-modal single-cell input.
- The model must have an associated peer-reviewed publication.

Publications from the search results (MultiDGD and MultiVI) were read and their relevant citations were explored to find other models (e.g., Cobolt), reviews, and relevant related work. After reading multiDGD, MultiVI, and Cobolt papers, it seemed that multiDGD and multiVI were the most performant, recent models for closer investigation and experimental reproduction.

4 Experiments

Experiments will compare the performance of multiDGD and MultiVI to demonstrate which model learns cell embeddings similar to input sequencing data while mitigating batch effects.

4.1 Data

Processed data from the multiDGD publication will be used to conduct experiments. The data includes paired single-cell RNA-seq and ATAC-seq from three datasets: 1) human bone marrow (num. cells=69249), 2) human brain (num. cells=3534), and 3) mouse gastrulation (num. cells=56861) tissues. The latter two datasets underwent feature selection preprocessing where if a feature was not present (zero-valued) in 1 percent or more of the cells, then it is dropped. The preprocessed data is available at

`https://figshare.com/articles/dataset/multiDGD_-_processed_data_and_models/23796198/1`

.

4.2 Experiment 1: reconstruction error

Experiment 1 will measure the accuracy of model representations by measuring the reconstruction loss of the decoder compared to an input cell feature vector.

Within each dataset, a subset of cells will be held out for evaluation while remainder are used to train the multiDGD and multiVI models. Then, reconstruction error of the test RNA-seq cell data will be quantified by the root mean square error (RMSE) between the decoder output and input data. Similarly for ATAC-seq, the decoder output will be evaluated by the area under the precision recall curve (AUPRC); note, in the multiDGD publication ATAC-seq peaks are binarized such that any value of 1 or greater is set to 1, otherwise 0. The expected result is that RMSE will be lower for multiDGD compared to MultiVI, and the AUPRC will be higher for multiDGD compared to multiVI.

4.3 Experiment 2: batch effect removal

Experiment 2 will measure each model's tendency to cluster representations by batch effects.

For each dataset with more than 1 batch (human bone marrow and mouse gastrulation), the tendency of cell representations to cluster by batch will be quantified by 1 - average silhouette width (ASW). By assigning cells to clusters by their batch label, the silhouette width will score high if cells of the same batch are closer in the embedding space compared to cells from other batches; the max ASW score is 1. Intuitively, a higher 1-ASW value indicates cell data does not cluster by batch labels. The expected result is that multiDGD will have a higher 1-ASW compared to multiVI.

References

1. Pan, L., Parini, P., Tremmel, R., et al.: Single cell atlas: a single-cell multi-omics human cell encyclopedia. Genome Biology **25** (2024) 104
2. Gong, B., Zhou, Y., Purdom, E.: Cobolt: integrative analysis of multimodal single-cell sequencing data. Genome Biology **22**(1) (2021) 351
3. Ashuach, T., Gabitto, M.I., Koodli, R.V., et al.: Multivi: deep generative model for the integration of multimodal data. Nature Methods **20** (2023) 1222–1231
4. Schuster, V., Dann, E., Krogh, A., et al.: multidgd: A versatile deep generative model for multi-omics data. Nature Communications **15**(1) (2024) 10031
5. Kingma, D.P., Welling, M.: Auto-encoding variational bayes (2022)
6. Schuster, V., Krogh, A.: A manifold learning perspective on representation learning: Learning decoder and representations without an encoder. Entropy **23**(11) (2021)