ECE 449 Project Milestone

Group Name: Feipan 449

Project Number: #7

Group Number: Yao Wentao, Xu Ke, Kong Zitai, Liu Chang

1. Paper Research

   Here, we mainly research the paper “Learning interpretable cellular and gene signature embeddings from single-cell transcriptomic data”. This paper presents an unsupervised model “scETM” based on “Topic embedding” model. To be specific, this model learns an encoder network to infer cell type mixture and a set of highly interpretable gene embeddings, topic embeddings and batch-effect linear intercepts from RNA-seq datasets. In this model, it uses the “topic embedding”, which is a method in NLP field. So, we search its original paper “Topic Modeling in Embedding Spaces” and have some deeper insights into it. We will explain this in brief below:

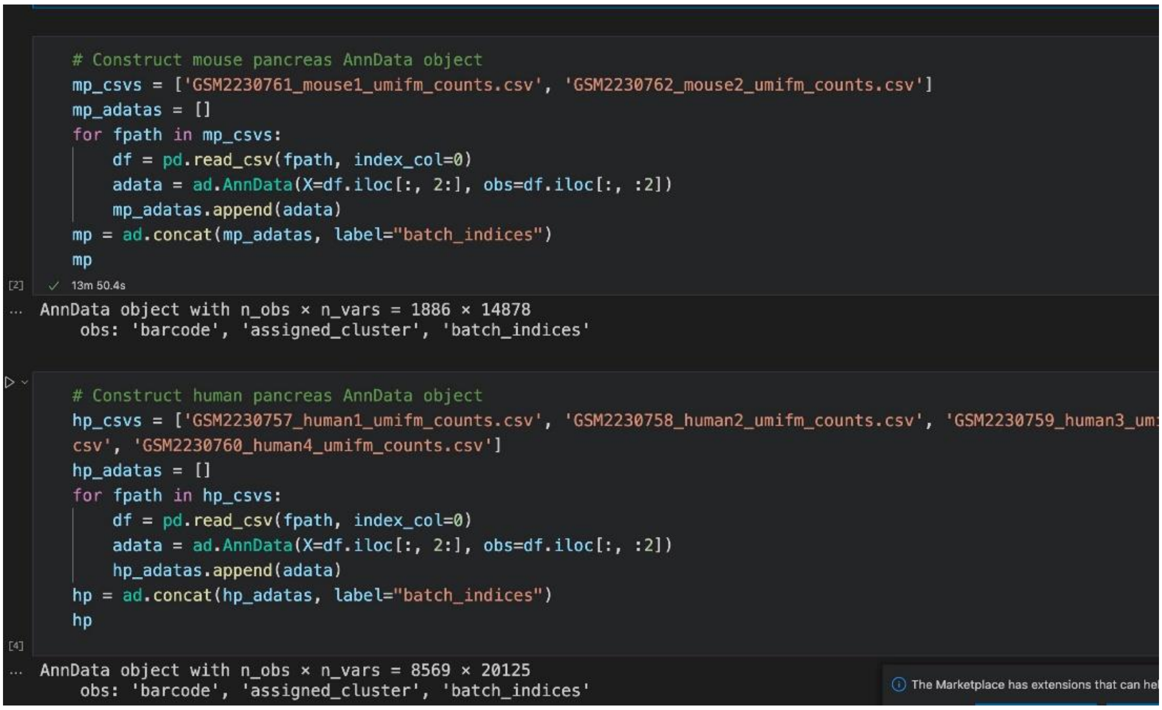
   The Embedded Topic model (ETM) uses the embedding representation of both words and topics. Simply speaking, for an arbitrary article, it will have several topics. And for each word from this article, it will have a distribution over all these topics. The Embedded topic model is a method to embed these features. To be specific, the vocabulary is embedded in a L-dimensional space, and the latent topics are embedded into K-dimensional space. In that case, the k^{th} topic is a vector $\alpha_k \in R^L$ in the embedding space, which is topic embedding. With this form, the ETM will assign high probability to a word v in topic k by measuring the agreement between the word’s embedding and the topic embedding.

   So, how does ETM connect to the RNA-seq task? In the sc-ETM model, each cell is considered as a document, and each scRNA-seq read is considered as a token in the document. Then gene give rise to each RNA-seq read is a word. And the latent types of cells are the latent topics. In that case, the ETM, an NLP method is well fit for the RNA-seq task because of their similarities. It is a good method that can help to enables the incorporation of know gene sets into the gene embeddings, so that it can also help to learn the association between pathways and topics through the topic embeddings.

2. Data preparation

   The original data is stored in csv files, via using AnnData, we process it in a way that will facilitate the data manipulation.

   AnnData is capable of reading from and writing to csv and h5ad files. In this case, both the data of mouse pancreas and human pancreas are in the form of csv files. By processing the data, we extract the matrix with raw data and their oberservations (in this case, the labels).



```
# Construct mouse pancreas AnnData object
mp_csvs = ['GSM2230761_mouse1_umifm_counts.csv', 'GSM2230762_mouse2_umifm_counts.csv']
mp_adatas = []
for fpath in mp_csvs:
    df = pd.read_csv(fpath, index_col=0)
    adata = ad.AnnData(X=df.iloc[:, 2:], obs=df.iloc[:, :2])
    mp_adatas.append(adata)
mp = ad.concat(mp_adatas, label="batch_indices")
mp
```

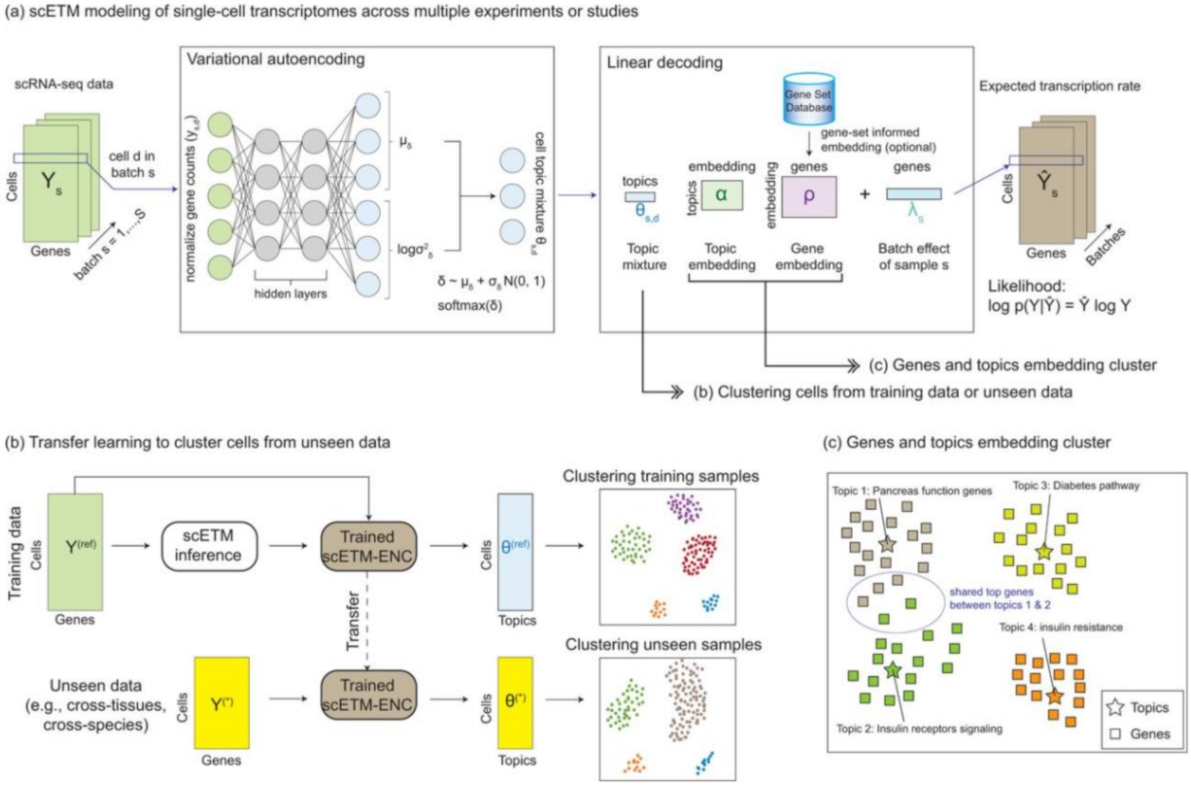
```
[2]  ✓ 13m 50.4s
...  AnnData object with n_obs × n_vars = 1886 × 14878
         obs: 'barcode', 'assigned_cluster', 'batch_indices'
```

```
# Construct human pancreas AnnData object
hp_csvs = ['GSM2230757_human1_umifm_counts.csv', 'GSM2230758_human2_umifm_counts.csv', 'GSM2230759_human3_um
csv', 'GSM2230760_human4_umifm_counts.csv']
hp_adatas = []
for fpath in hp_csvs:
    df = pd.read_csv(fpath, index_col=0)
    adata = ad.AnnData(X=df.iloc[:, 2:], obs=df.iloc[:, :2])
    hp_adatas.append(adata)
hp = ad.concat(hp_adatas, label="batch_indices")
hp
```

```
[4]
...  AnnData object with n_obs × n_vars = 8569 × 20125
         obs: 'barcode', 'assigned_cluster', 'batch_indices'
```

3. Experimental Method

We establish preliminary experiments by reproducing the method in the paper "Learning interpretable cellular and gene signature embeddings from single-cell transcriptomic data", i.e., scETM. The structure of scETM is shown below.



(a) scETM modeling of single-cell transcriptomes across multiple experiments or studies
scRNA-seq data
Cells
Y_s
Genes
cell d in batch s
batch s = 1,...,S
Variational autoencoding
normalize gene counts (y_s,d)
hidden layers
μ_δ
logσ²_δ
δ ~ μ_δ + σ_δ N(0, 1)
softmax(δ)
cell topic mixture θ_s,d
Linear decoding
Gene Set Database
gene-set informed embedding (optional)
topics
θ_s,d
Topic mixture
embedding
topics
α
Topic embedding
genes
embedding
ρ
Gene embedding
+
genes
λ_s
Batch effect of sample s
Expected transcription rate
Cells
Ŷ_s
Genes
Batches
Likelihood:
log p(Y|Ŷ) = Ŷ log Y
(c) Genes and topics embedding cluster
(b) Clustering cells from training data or unseen data
(b) Transfer learning to cluster cells from unseen data
Training data
Cells
Y^(ref)
Genes
scETM inference
Trained scETM-ENC
Transfer
Cells
θ^(ref)
Topics
Clustering training samples
Unseen data (e.g., cross-tissues, cross-species)
Cells
Y^(*)
Genes
Trained scETM-ENC
Cells
θ^(*)
Topics
Clustering unseen samples
(c) Genes and topics embedding cluster
Topic 1: Pancreas function genes
Topic 3: Diabetes pathway
shared top genes between topics 1 & 2
Topic 4: insulin resistance
Topic 2: Insulin receptors signaling
☆ Topics
□ Genes

The input scRNA-seq data is preprocessed into a couple of matrixes, a matrix will a cell in a row and the genes expressed in a column, and each matrix represent a batch of the biological experiments. Then we feed the matrixes into the model. The model has a variational autoencoder and a linear decoder. The autoencoder for inferring the cell topic mixture is a 2-layer neural network plus a softmax layer, with hidden sizes of 128, ReLU activations, 1D batch normalization, and 0.1 dropout rate between layers. The gene embedding dimension is set to be 400, and the number of topics is set to be 50. For each cell, the variational autoencoder can learn the mean and logarithm of variance, each pair will be put into a Normal distribution after softmax, which represents the topic of each cell. All these cell-topic relation will be put together to be a cell topic mixture matrix.

$$\delta_d \sim N(0, I), \qquad \theta_d = softmax(\delta_d) = \frac{e^{\delta_{d,k}}}{\sum_{k=1}^{K} e^{\delta_{d,k}}}$$

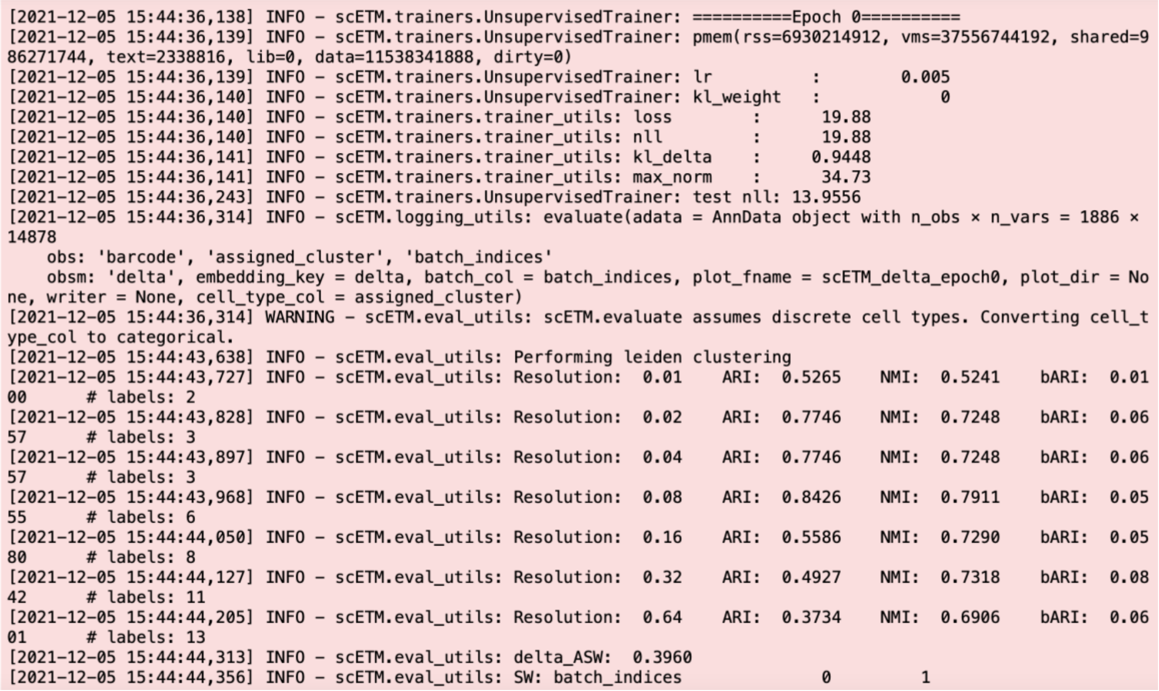
In the linear decoder, we will have three matrixes. One is the cell topic mixture (cells-by-topics) matrix θ comes from encoder, the second is a mixture of topic and gene information from gene set database, i.e., topic embedding matrix α and the last one is the gene embedding matrix ρ coming from gene set database. Then we add a batch effect correction matrix λ to remove the batch effect. Finally, we can get the expected amount of gene expression (of cell d in batch s) by

$$softmax(\theta_{s,d} \alpha \rho + \lambda_s)$$

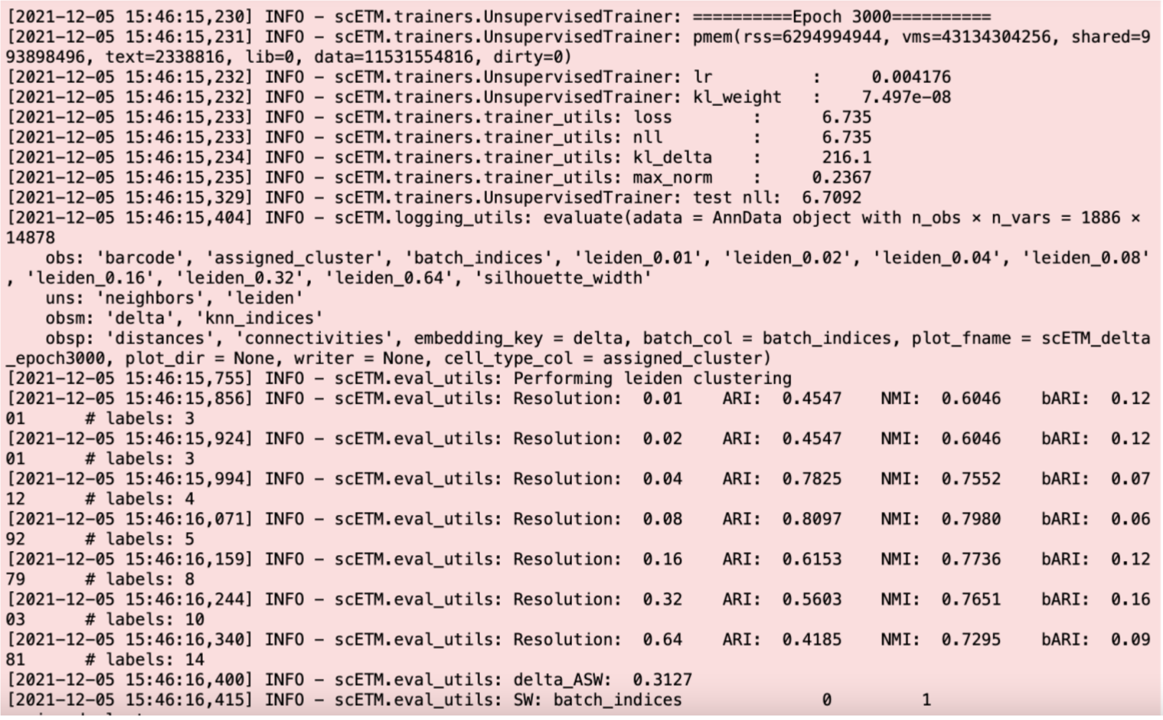
For training, we use evidence lower bound (ELBO) of the marginal categorical likelihood of the scRNA-seq counts as the loss function and optimize our model with Adam Optimizer and a 0.005 learning rate.

4. Preliminary Result

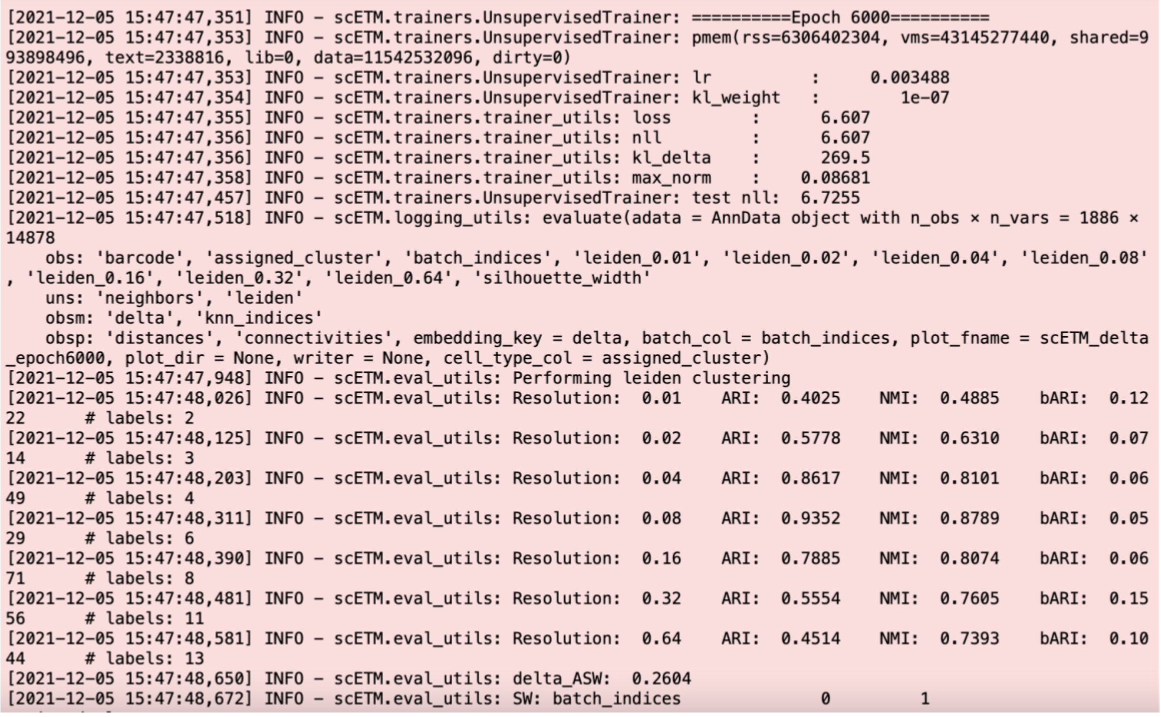
We use an UnsupervisedTrainer to train the scETM model. Since the scETM requires about 6k steps to converge (observe the test NLL to confirm that), so for the MP dataset whose size is smaller than the training minibatch size, we will train it for at least 6k epochs, some details are shown below:



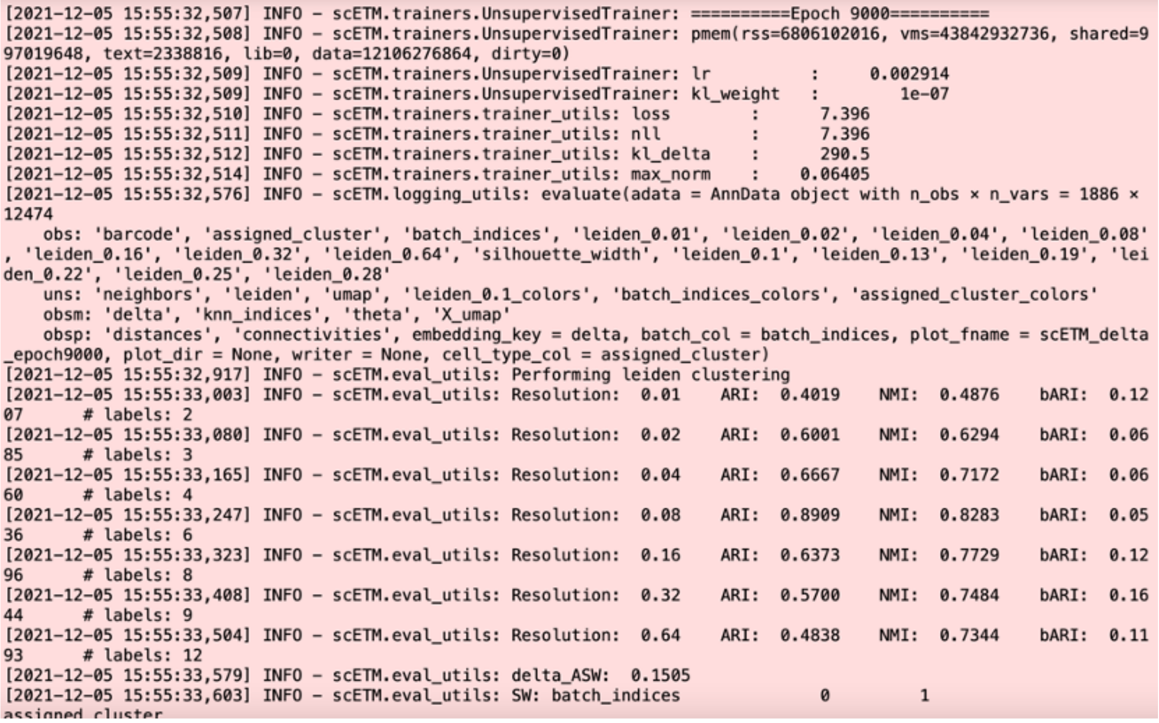
```
[2021-12-05 15:44:36,138] INFO - scETM.trainers.UnsupervisedTrainer: ==========Epoch 0==========
[2021-12-05 15:44:36,139] INFO - scETM.trainers.UnsupervisedTrainer: pmem(rss=6930214912, vms=37556744192, shared=9
86271744, text=2338816, lib=0, data=11538341888, dirty=0)
[2021-12-05 15:44:36,139] INFO - scETM.trainers.UnsupervisedTrainer: lr          :        0.005
[2021-12-05 15:44:36,140] INFO - scETM.trainers.UnsupervisedTrainer: kl_weight   :            0
[2021-12-05 15:44:36,140] INFO - scETM.trainers.trainer_utils: loss        :      19.88
[2021-12-05 15:44:36,140] INFO - scETM.trainers.trainer_utils: nll         :      19.88
[2021-12-05 15:44:36,141] INFO - scETM.trainers.trainer_utils: kl_delta    :     0.9448
[2021-12-05 15:44:36,141] INFO - scETM.trainers.trainer_utils: max_norm    :      34.73
[2021-12-05 15:44:36,243] INFO - scETM.trainers.UnsupervisedTrainer: test nll: 13.9556
[2021-12-05 15:44:36,314] INFO - scETM.logging_utils: evaluate(adata = AnnData object with n_obs × n_vars = 1886 ×
14878
    obs: 'barcode', 'assigned_cluster', 'batch_indices'
    obsm: 'delta', embedding_key = delta, batch_col = batch_indices, plot_fname = scETM_delta_epoch0, plot_dir = No
ne, writer = None, cell_type_col = assigned_cluster)
[2021-12-05 15:44:36,314] WARNING - scETM.eval_utils: scETM.evaluate assumes discrete cell types. Converting cell_t
ype_col to categorical.
[2021-12-05 15:44:43,638] INFO - scETM.eval_utils: Performing leiden clustering
[2021-12-05 15:44:43,727] INFO - scETM.eval_utils: Resolution:  0.01    ARI:  0.5265    NMI:  0.5241    bARI:  0.01
00      # labels: 2
[2021-12-05 15:44:43,828] INFO - scETM.eval_utils: Resolution:  0.02    ARI:  0.7746    NMI:  0.7248    bARI:  0.06
57      # labels: 3
[2021-12-05 15:44:43,897] INFO - scETM.eval_utils: Resolution:  0.04    ARI:  0.7746    NMI:  0.7248    bARI:  0.06
57      # labels: 3
[2021-12-05 15:44:43,968] INFO - scETM.eval_utils: Resolution:  0.08    ARI:  0.8426    NMI:  0.7911    bARI:  0.05
55      # labels: 6
[2021-12-05 15:44:44,050] INFO - scETM.eval_utils: Resolution:  0.16    ARI:  0.5586    NMI:  0.7290    bARI:  0.05
80      # labels: 8
[2021-12-05 15:44:44,127] INFO - scETM.eval_utils: Resolution:  0.32    ARI:  0.4927    NMI:  0.7318    bARI:  0.08
42      # labels: 11
[2021-12-05 15:44:44,205] INFO - scETM.eval_utils: Resolution:  0.64    ARI:  0.3734    NMI:  0.6906    bARI:  0.06
01      # labels: 13
[2021-12-05 15:44:44,313] INFO - scETM.eval_utils: delta_ASW:  0.3960
[2021-12-05 15:44:44,356] INFO - scETM.eval_utils: SW: batch_indices             0         1
```



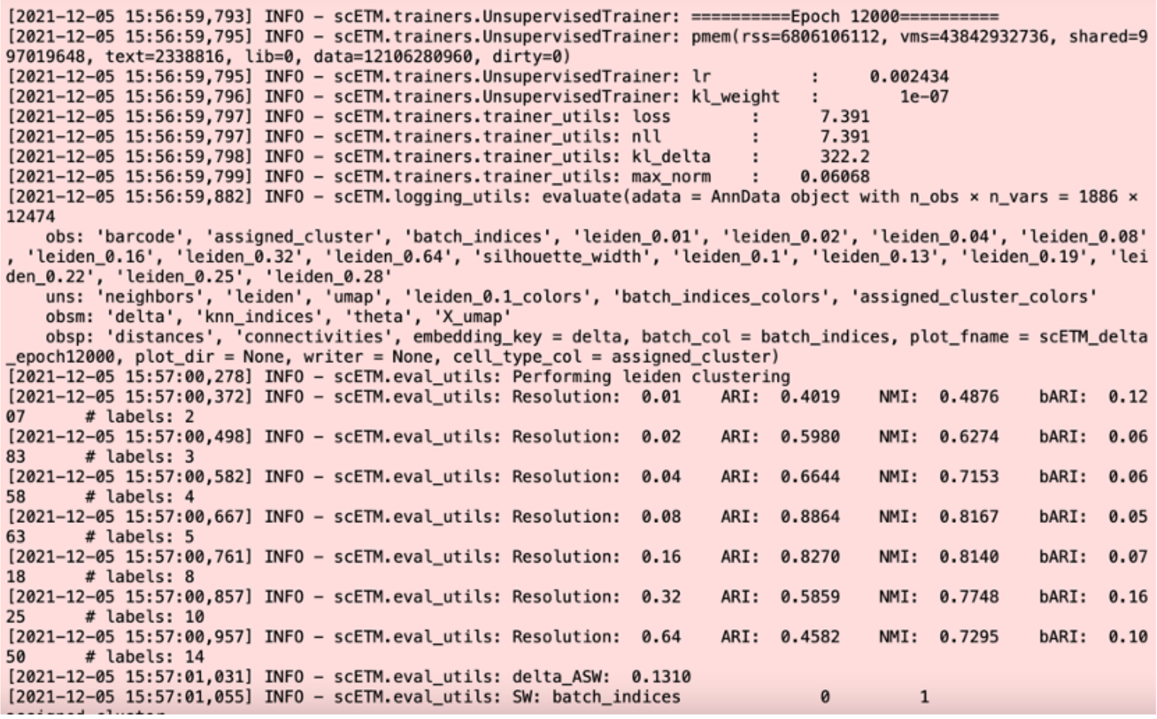
```
[2021-12-05 15:46:15,230] INFO - scETM.trainers.UnsupervisedTrainer: ==========Epoch 3000==========
[2021-12-05 15:46:15,231] INFO - scETM.trainers.UnsupervisedTrainer: pmem(rss=6294994944, vms=43134304256, shared=9
93898496, text=2338816, lib=0, data=11531554816, dirty=0)
[2021-12-05 15:46:15,232] INFO - scETM.trainers.UnsupervisedTrainer: lr           :     0.004176
[2021-12-05 15:46:15,232] INFO - scETM.trainers.UnsupervisedTrainer: kl_weight    :    7.497e-08
[2021-12-05 15:46:15,233] INFO - scETM.trainers.trainer_utils: loss        :      6.735
[2021-12-05 15:46:15,233] INFO - scETM.trainers.trainer_utils: nll         :      6.735
[2021-12-05 15:46:15,234] INFO - scETM.trainers.trainer_utils: kl_delta    :      216.1
[2021-12-05 15:46:15,235] INFO - scETM.trainers.trainer_utils: max_norm    :     0.2367
[2021-12-05 15:46:15,329] INFO - scETM.trainers.UnsupervisedTrainer: test nll:  6.7092
[2021-12-05 15:46:15,404] INFO - scETM.logging_utils: evaluate(adata = AnnData object with n_obs × n_vars = 1886 ×
14878
    obs: 'barcode', 'assigned_cluster', 'batch_indices', 'leiden_0.01', 'leiden_0.02', 'leiden_0.04', 'leiden_0.08'
, 'leiden_0.16', 'leiden_0.32', 'leiden_0.64', 'silhouette_width'
    uns: 'neighbors', 'leiden'
    obsm: 'delta', 'knn_indices'
    obsp: 'distances', 'connectivities', embedding_key = delta, batch_col = batch_indices, plot_fname = scETM_delta
_epoch3000, plot_dir = None, writer = None, cell_type_col = assigned_cluster)
[2021-12-05 15:46:15,755] INFO - scETM.eval_utils: Performing leiden clustering
[2021-12-05 15:46:15,856] INFO - scETM.eval_utils: Resolution:  0.01    ARI:  0.4547    NMI:  0.6046    bARI:  0.12
01      # labels: 3
[2021-12-05 15:46:15,924] INFO - scETM.eval_utils: Resolution:  0.02    ARI:  0.4547    NMI:  0.6046    bARI:  0.12
01      # labels: 3
[2021-12-05 15:46:15,994] INFO - scETM.eval_utils: Resolution:  0.04    ARI:  0.7825    NMI:  0.7552    bARI:  0.07
12      # labels: 4
[2021-12-05 15:46:16,071] INFO - scETM.eval_utils: Resolution:  0.08    ARI:  0.8097    NMI:  0.7980    bARI:  0.06
92      # labels: 5
[2021-12-05 15:46:16,159] INFO - scETM.eval_utils: Resolution:  0.16    ARI:  0.6153    NMI:  0.7736    bARI:  0.12
79      # labels: 8
[2021-12-05 15:46:16,244] INFO - scETM.eval_utils: Resolution:  0.32    ARI:  0.5603    NMI:  0.7651    bARI:  0.16
03      # labels: 10
[2021-12-05 15:46:16,340] INFO - scETM.eval_utils: Resolution:  0.64    ARI:  0.4185    NMI:  0.7295    bARI:  0.09
81      # labels: 14
[2021-12-05 15:46:16,400] INFO - scETM.eval_utils: delta_ASW:  0.3127
[2021-12-05 15:46:16,415] INFO - scETM.eval_utils: SW: batch_indices             0         1
```



```
[2021-12-05 15:47:47,351] INFO - scETM.trainers.UnsupervisedTrainer: ==========Epoch 6000==========
[2021-12-05 15:47:47,353] INFO - scETM.trainers.UnsupervisedTrainer: pmem(rss=6306402304, vms=43145277440, shared=9
93898496, text=2338816, lib=0, data=11542532096, dirty=0)
[2021-12-05 15:47:47,353] INFO - scETM.trainers.UnsupervisedTrainer: lr           :     0.003488
[2021-12-05 15:47:47,354] INFO - scETM.trainers.UnsupervisedTrainer: kl_weight    :        1e-07
[2021-12-05 15:47:47,355] INFO - scETM.trainers.trainer_utils: loss        :      6.607
[2021-12-05 15:47:47,356] INFO - scETM.trainers.trainer_utils: nll         :      6.607
[2021-12-05 15:47:47,356] INFO - scETM.trainers.trainer_utils: kl_delta    :      269.5
[2021-12-05 15:47:47,358] INFO - scETM.trainers.trainer_utils: max_norm    :    0.08681
[2021-12-05 15:47:47,457] INFO - scETM.trainers.UnsupervisedTrainer: test nll:  6.7255
[2021-12-05 15:47:47,518] INFO - scETM.logging_utils: evaluate(adata = AnnData object with n_obs × n_vars = 1886 ×
14878
    obs: 'barcode', 'assigned_cluster', 'batch_indices', 'leiden_0.01', 'leiden_0.02', 'leiden_0.04', 'leiden_0.08'
, 'leiden_0.16', 'leiden_0.32', 'leiden_0.64', 'silhouette_width'
    uns: 'neighbors', 'leiden'
    obsm: 'delta', 'knn_indices'
    obsp: 'distances', 'connectivities', embedding_key = delta, batch_col = batch_indices, plot_fname = scETM_delta
_epoch6000, plot_dir = None, writer = None, cell_type_col = assigned_cluster)
[2021-12-05 15:47:47,948] INFO - scETM.eval_utils: Performing leiden clustering
[2021-12-05 15:47:48,026] INFO - scETM.eval_utils: Resolution:  0.01    ARI:  0.4025    NMI:  0.4885    bARI:  0.12
22      # labels: 2
[2021-12-05 15:47:48,125] INFO - scETM.eval_utils: Resolution:  0.02    ARI:  0.5778    NMI:  0.6310    bARI:  0.07
14      # labels: 3
[2021-12-05 15:47:48,203] INFO - scETM.eval_utils: Resolution:  0.04    ARI:  0.8617    NMI:  0.8101    bARI:  0.06
49      # labels: 4
[2021-12-05 15:47:48,311] INFO - scETM.eval_utils: Resolution:  0.08    ARI:  0.9352    NMI:  0.8789    bARI:  0.05
29      # labels: 6
[2021-12-05 15:47:48,390] INFO - scETM.eval_utils: Resolution:  0.16    ARI:  0.7885    NMI:  0.8074    bARI:  0.06
71      # labels: 8
[2021-12-05 15:47:48,481] INFO - scETM.eval_utils: Resolution:  0.32    ARI:  0.5554    NMI:  0.7605    bARI:  0.15
56      # labels: 11
[2021-12-05 15:47:48,581] INFO - scETM.eval_utils: Resolution:  0.64    ARI:  0.4514    NMI:  0.7393    bARI:  0.10
44      # labels: 13
[2021-12-05 15:47:48,650] INFO - scETM.eval_utils: delta_ASW:  0.2604
[2021-12-05 15:47:48,672] INFO - scETM.eval_utils: SW: batch_indices             0         1
```

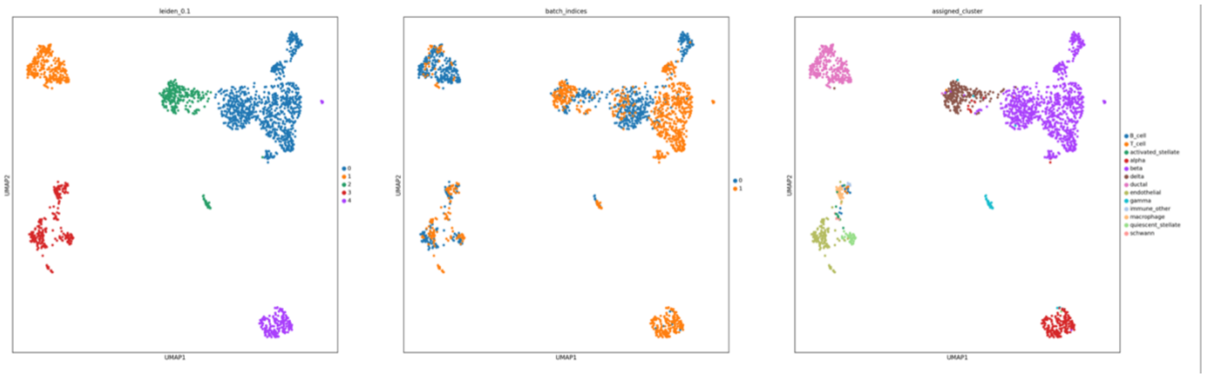


```
[2021-12-05 15:55:32,507] INFO - scETM.trainers.UnsupervisedTrainer: ==========Epoch 9000==========
[2021-12-05 15:55:32,508] INFO - scETM.trainers.UnsupervisedTrainer: pmem(rss=6806102016, vms=43842932736, shared=9
97019648, text=2338816, lib=0, data=12106276864, dirty=0)
[2021-12-05 15:55:32,509] INFO - scETM.trainers.UnsupervisedTrainer: lr          :     0.002914
[2021-12-05 15:55:32,509] INFO - scETM.trainers.UnsupervisedTrainer: kl_weight   :        1e-07
[2021-12-05 15:55:32,510] INFO - scETM.trainers.trainer_utils: loss        :      7.396
[2021-12-05 15:55:32,511] INFO - scETM.trainers.trainer_utils: nll         :      7.396
[2021-12-05 15:55:32,512] INFO - scETM.trainers.trainer_utils: kl_delta    :      290.5
[2021-12-05 15:55:32,514] INFO - scETM.trainers.trainer_utils: max_norm    :    0.06405
[2021-12-05 15:55:32,576] INFO - scETM.logging_utils: evaluate(adata = AnnData object with n_obs × n_vars = 1886 ×
12474
    obs: 'barcode', 'assigned_cluster', 'batch_indices', 'leiden_0.01', 'leiden_0.02', 'leiden_0.04', 'leiden_0.08'
, 'leiden_0.16', 'leiden_0.32', 'leiden_0.64', 'silhouette_width', 'leiden_0.1', 'leiden_0.13', 'leiden_0.19', 'lei
den_0.22', 'leiden_0.25', 'leiden_0.28'
    uns: 'neighbors', 'leiden', 'umap', 'leiden_0.1_colors', 'batch_indices_colors', 'assigned_cluster_colors'
    obsm: 'delta', 'knn_indices', 'theta', 'X_umap'
    obsp: 'distances', 'connectivities', embedding_key = delta, batch_col = batch_indices, plot_fname = scETM_delta
_epoch9000, plot_dir = None, writer = None, cell_type_col = assigned_cluster)
[2021-12-05 15:55:32,917] INFO - scETM.eval_utils: Performing leiden clustering
[2021-12-05 15:55:33,003] INFO - scETM.eval_utils: Resolution:  0.01    ARI:  0.4019    NMI:  0.4876    bARI:  0.12
07      # labels: 2
[2021-12-05 15:55:33,080] INFO - scETM.eval_utils: Resolution:  0.02    ARI:  0.6001    NMI:  0.6294    bARI:  0.06
85      # labels: 3
[2021-12-05 15:55:33,165] INFO - scETM.eval_utils: Resolution:  0.04    ARI:  0.6667    NMI:  0.7172    bARI:  0.06
60      # labels: 4
[2021-12-05 15:55:33,247] INFO - scETM.eval_utils: Resolution:  0.08    ARI:  0.8909    NMI:  0.8283    bARI:  0.05
36      # labels: 6
[2021-12-05 15:55:33,323] INFO - scETM.eval_utils: Resolution:  0.16    ARI:  0.6373    NMI:  0.7729    bARI:  0.12
96      # labels: 8
[2021-12-05 15:55:33,408] INFO - scETM.eval_utils: Resolution:  0.32    ARI:  0.5700    NMI:  0.7484    bARI:  0.16
44      # labels: 9
[2021-12-05 15:55:33,504] INFO - scETM.eval_utils: Resolution:  0.64    ARI:  0.4838    NMI:  0.7344    bARI:  0.11
93      # labels: 12
[2021-12-05 15:55:33,579] INFO - scETM.eval_utils: delta_ASW:  0.1505
[2021-12-05 15:55:33,603] INFO - scETM.eval_utils: SW: batch_indices             0         1
assigned_cluster
```

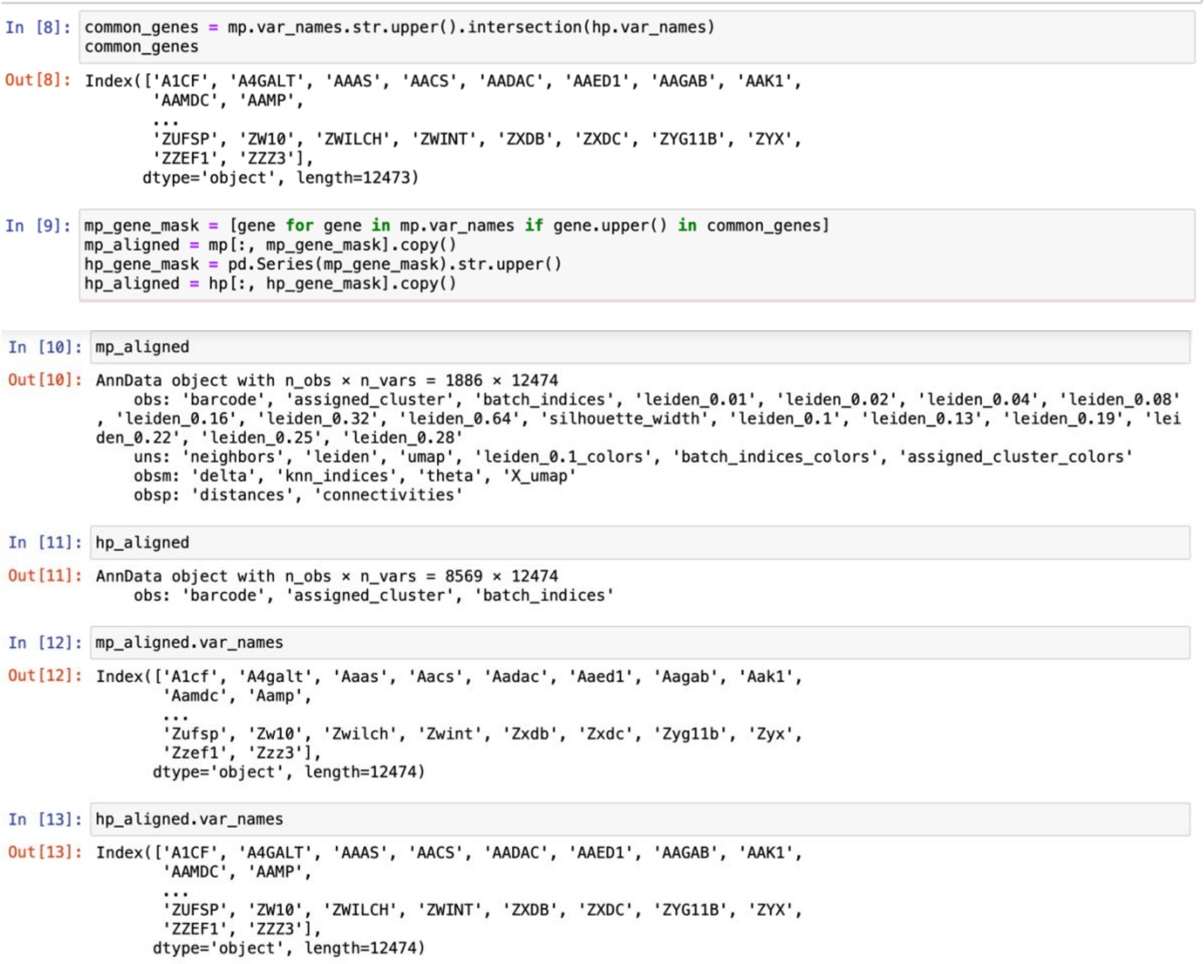


```
[2021-12-05 15:56:59,793] INFO - scETM.trainers.UnsupervisedTrainer: ==========Epoch 12000==========
[2021-12-05 15:56:59,795] INFO - scETM.trainers.UnsupervisedTrainer: pmem(rss=6806106112, vms=43842932736, shared=9
97019648, text=2338816, lib=0, data=12106280960, dirty=0)
[2021-12-05 15:56:59,795] INFO - scETM.trainers.UnsupervisedTrainer: lr          :     0.002434
[2021-12-05 15:56:59,796] INFO - scETM.trainers.UnsupervisedTrainer: kl_weight   :        1e-07
[2021-12-05 15:56:59,797] INFO - scETM.trainers.trainer_utils: loss        :      7.391
[2021-12-05 15:56:59,797] INFO - scETM.trainers.trainer_utils: nll         :      7.391
[2021-12-05 15:56:59,798] INFO - scETM.trainers.trainer_utils: kl_delta    :      322.2
[2021-12-05 15:56:59,799] INFO - scETM.trainers.trainer_utils: max_norm    :    0.06068
[2021-12-05 15:56:59,882] INFO - scETM.logging_utils: evaluate(adata = AnnData object with n_obs × n_vars = 1886 ×
12474
    obs: 'barcode', 'assigned_cluster', 'batch_indices', 'leiden_0.01', 'leiden_0.02', 'leiden_0.04', 'leiden_0.08'
, 'leiden_0.16', 'leiden_0.32', 'leiden_0.64', 'silhouette_width', 'leiden_0.1', 'leiden_0.13', 'leiden_0.19', 'lei
den_0.22', 'leiden_0.25', 'leiden_0.28'
    uns: 'neighbors', 'leiden', 'umap', 'leiden_0.1_colors', 'batch_indices_colors', 'assigned_cluster_colors'
    obsm: 'delta', 'knn_indices', 'theta', 'X_umap'
    obsp: 'distances', 'connectivities', embedding_key = delta, batch_col = batch_indices, plot_fname = scETM_delta
_epoch12000, plot_dir = None, writer = None, cell_type_col = assigned_cluster)
[2021-12-05 15:57:00,278] INFO - scETM.eval_utils: Performing leiden clustering
[2021-12-05 15:57:00,372] INFO - scETM.eval_utils: Resolution:  0.01    ARI:  0.4019    NMI:  0.4876    bARI:  0.12
07      # labels: 2
[2021-12-05 15:57:00,498] INFO - scETM.eval_utils: Resolution:  0.02    ARI:  0.5980    NMI:  0.6274    bARI:  0.06
83      # labels: 3
[2021-12-05 15:57:00,582] INFO - scETM.eval_utils: Resolution:  0.04    ARI:  0.6644    NMI:  0.7153    bARI:  0.06
58      # labels: 4
[2021-12-05 15:57:00,667] INFO - scETM.eval_utils: Resolution:  0.08    ARI:  0.8864    NMI:  0.8167    bARI:  0.05
63      # labels: 5
[2021-12-05 15:57:00,761] INFO - scETM.eval_utils: Resolution:  0.16    ARI:  0.8270    NMI:  0.8140    bARI:  0.07
18      # labels: 8
[2021-12-05 15:57:00,857] INFO - scETM.eval_utils: Resolution:  0.32    ARI:  0.5859    NMI:  0.7748    bARI:  0.16
25      # labels: 10
[2021-12-05 15:57:00,957] INFO - scETM.eval_utils: Resolution:  0.64    ARI:  0.4582    NMI:  0.7295    bARI:  0.10
50      # labels: 14
[2021-12-05 15:57:01,031] INFO - scETM.eval_utils: delta_ASW:  0.1310
[2021-12-05 15:57:01,055] INFO - scETM.eval_utils: SW: batch_indices             0         1
```

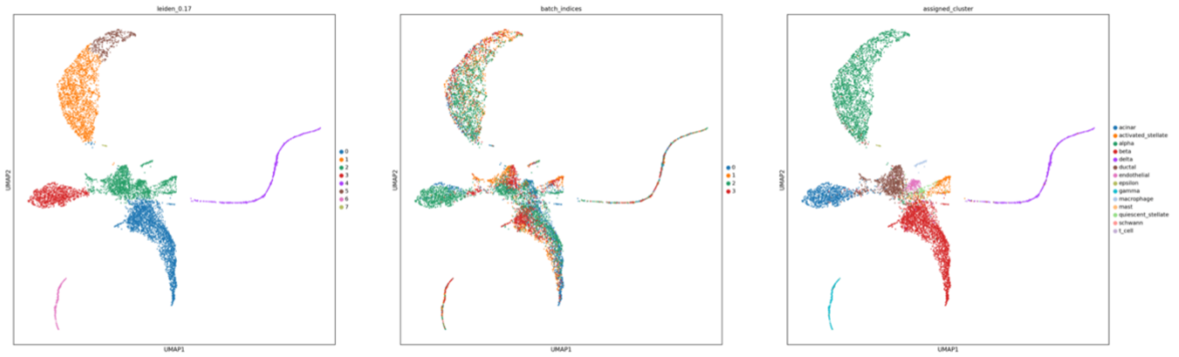
We use the evaluate function provided by scETM to explicitly evaluate the learned embedding. The evaluate function looks for the embedding_key (which defaults to "delta") in adata.obsm, evaluates its ARI with cell type and batch, NMI with cell type, batch mixing entropy and kBET, then plots the embedding shown as following:



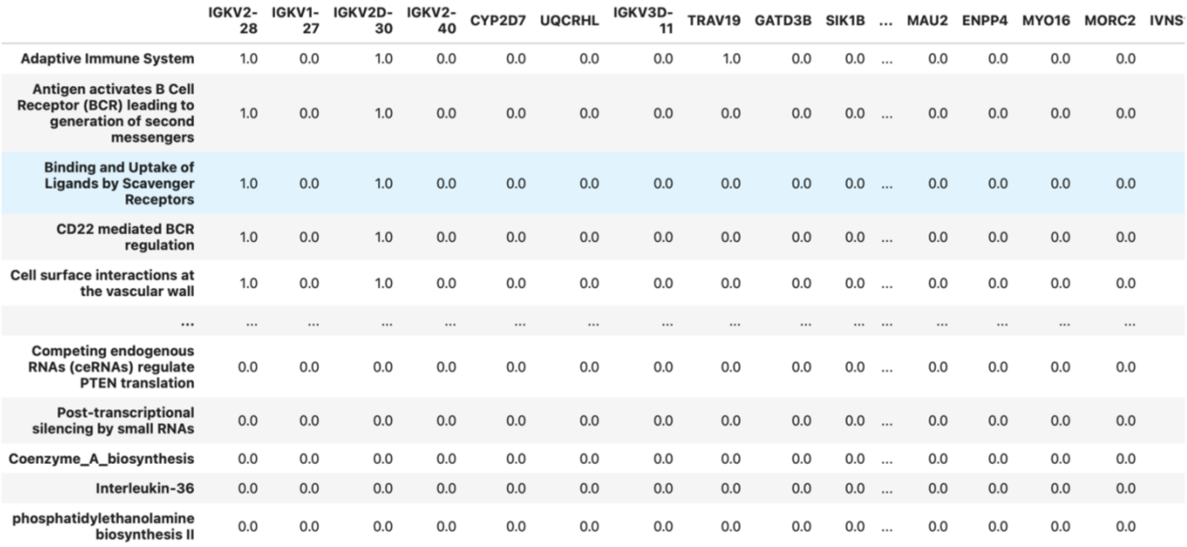
Transfer learning with scETM is extremely simple, we just train scETM on the reference dataset and apply it to the query dataset. We will demonstrate the aligning procedure in the code below.



```
In [8]:  common_genes = mp.var_names.str.upper().intersection(hp.var_names)
         common_genes

Out[8]:  Index(['A1CF', 'A4GALT', 'AAAS', 'AACS', 'AADAC', 'AAED1', 'AAGAB', 'AAK1',
                'AAMDC', 'AAMP',
                ...
                'ZUFSP', 'ZW10', 'ZWILCH', 'ZWINT', 'ZXDB', 'ZXDC', 'ZYG11B', 'ZYX',
                'ZZEF1', 'ZZZ3'],
               dtype='object', length=12473)

In [9]:  mp_gene_mask = [gene for gene in mp.var_names if gene.upper() in common_genes]
         mp_aligned = mp[:, mp_gene_mask].copy()
         hp_gene_mask = pd.Series(mp_gene_mask).str.upper()
         hp_aligned = hp[:, hp_gene_mask].copy()

In [10]: mp_aligned

Out[10]: AnnData object with n_obs × n_vars = 1886 × 12474
             obs: 'barcode', 'assigned_cluster', 'batch_indices', 'leiden_0.01', 'leiden_0.02', 'leiden_0.04', 'leiden_0.08'
         , 'leiden_0.16', 'leiden_0.32', 'leiden_0.64', 'silhouette_width', 'leiden_0.1', 'leiden_0.13', 'leiden_0.19', 'lei
         den_0.22', 'leiden_0.25', 'leiden_0.28'
             uns: 'neighbors', 'leiden', 'umap', 'leiden_0.1_colors', 'batch_indices_colors', 'assigned_cluster_colors'
             obsm: 'delta', 'knn_indices', 'theta', 'X_umap'
             obsp: 'distances', 'connectivities'

In [11]: hp_aligned

Out[11]: AnnData object with n_obs × n_vars = 8569 × 12474
             obs: 'barcode', 'assigned_cluster', 'batch_indices'

In [12]: mp_aligned.var_names

Out[12]: Index(['A1cf', 'A4galt', 'Aaas', 'Aacs', 'Aadac', 'Aaed1', 'Aagab', 'Aak1',
                'Aamdc', 'Aamp',
                ...
                'Zufsp', 'Zw10', 'Zwilch', 'Zwint', 'Zxdb', 'Zxdc', 'Zyg11b', 'Zyx',
                'Zzef1', 'Zzz3'],
               dtype='object', length=12474)

In [13]: hp_aligned.var_names

Out[13]: Index(['A1CF', 'A4GALT', 'AAAS', 'AACS', 'AADAC', 'AAED1', 'AAGAB', 'AAK1',
                'AAMDC', 'AAMP',
                ...
                'ZUFSP', 'ZW10', 'ZWILCH', 'ZWINT', 'ZXDB', 'ZXDC', 'ZYG11B', 'ZYX',
                'ZZEF1', 'ZZZ3'],
               dtype='object', length=12474)
```



Pathway-informed scETM (p-scETM) uses a pathway-gene matrix from external database as part/all of scETM gene embedding rho. And we will use the pathDIP data which we download from http://ophid.utoronto.ca/pathDIP/Download.jsp and shown as following:



	IGKV2-28	IGKV1-27	IGKV2D-30	IGKV2-40	CYP2D7	UQCRHL	IGKV3D-11	TRAV19	GATD3B	SIK1B	...	MAU2	ENPP4	MYO16	MORC2	IVNS
Adaptive Immune System	1.0	0.0	1.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	...	0.0	0.0	0.0	0.0	
Antigen activates B Cell Receptor (BCR) leading to generation of second messengers	1.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	
Binding and Uptake of Ligands by Scavenger Receptors	1.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	
CD22 mediated BCR regulation	1.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	
Cell surface interactions at the vascular wall	1.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
Competing endogenous RNAs (ceRNAs) regulate PTEN translation	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	
Post-transcriptional silencing by small RNAs	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	
Coenzyme_A_biosynthesis	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	
Interleukin-36	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	
phosphatidylethanolamine biosynthesis II	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	

In the next week, we will extend the pathway-gene matrix to all genes in the human pancreas dataset, filling missing values with 0.0 and instantiate a p-scETM model, passing the pathway-gene matrix to rho_fixed_emband train the p-scETM model.