Diabetes in mice

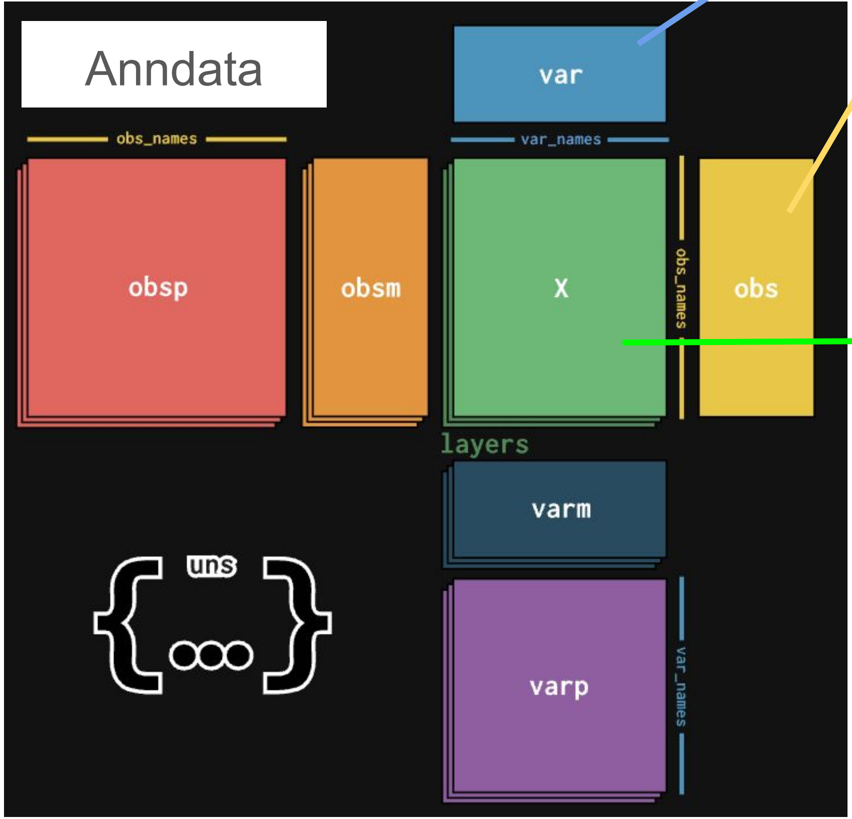
Detecting diabetes from gene expression with ML

By: Quinn Saul, Maja Lindholm, Jacob Flynn, Jakob B. Hansen and Ling Jun Zhou

Overview

	Supervised	Supervised & Unsupervised	Unsupervised	
The Data	**Classification**	**AutoEncoder**	**Clustering**	**Conclusion**
Mice with diabetes Cell x Gene	Can we detect sick cells?	Can we detect sick cells from knowing just how healthy cells look? Anomaly detection Preprocessing	Can we detect structural differences between healthy and sick cells? PCA Latent space of autoencoder	

The data



Information about genes (we don’t really care, but biologist do)

Information about cells, like disease labels

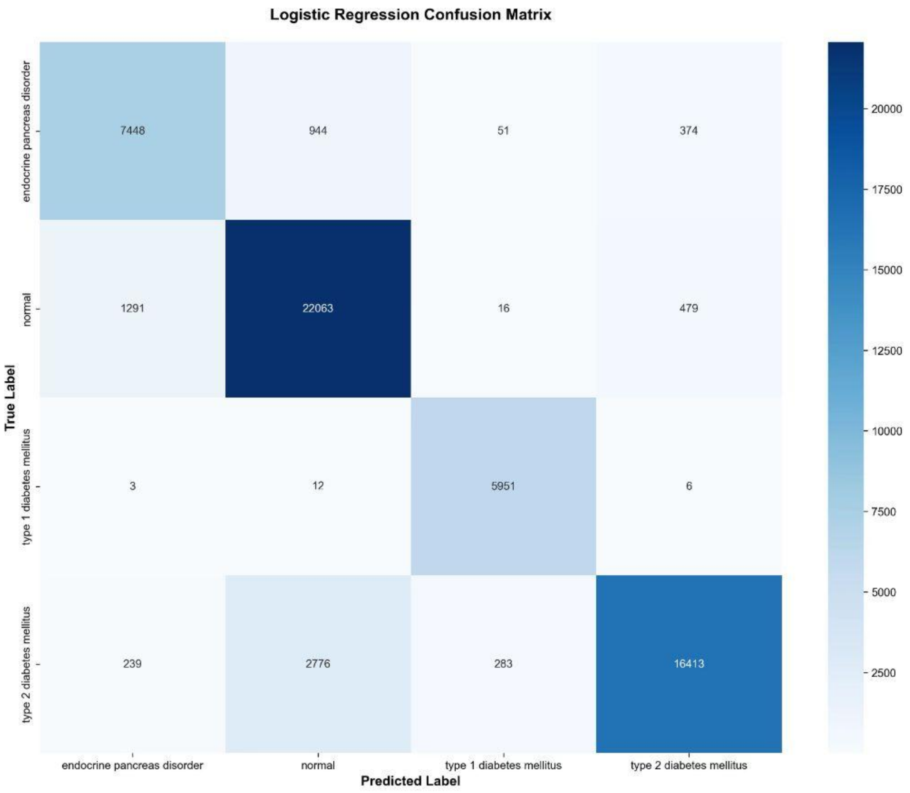
Genes

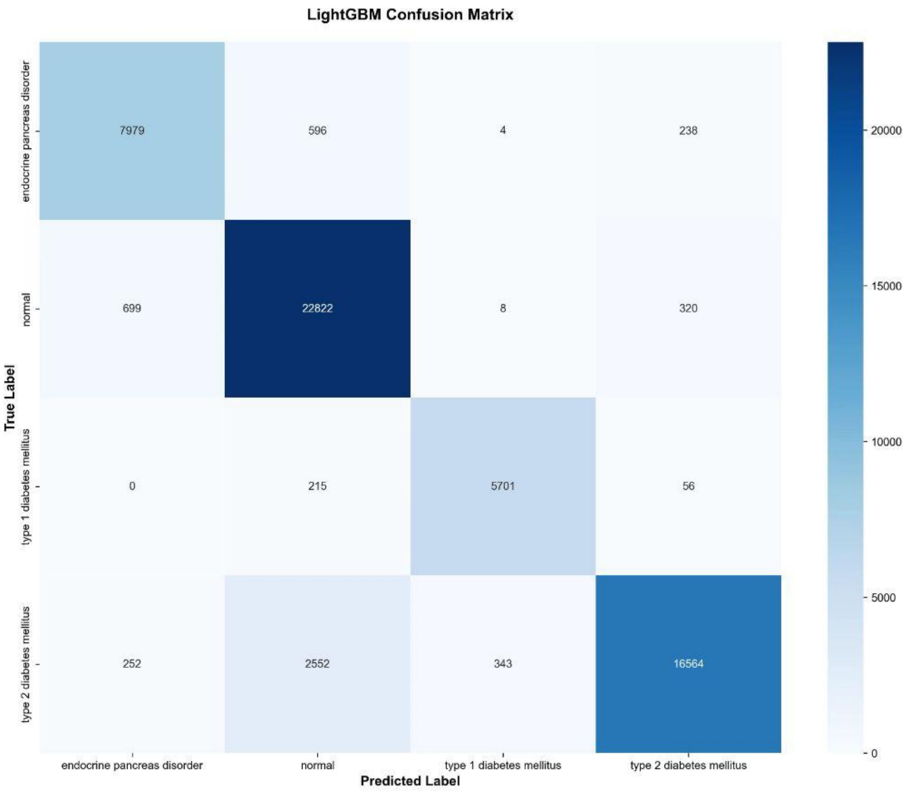
Cells

data.X	0	1	..	31.202
0	0.4	6	…	10
1	1.8	0	…	6
…	…	…	…	…
301.796	0.9	4	…	0.7

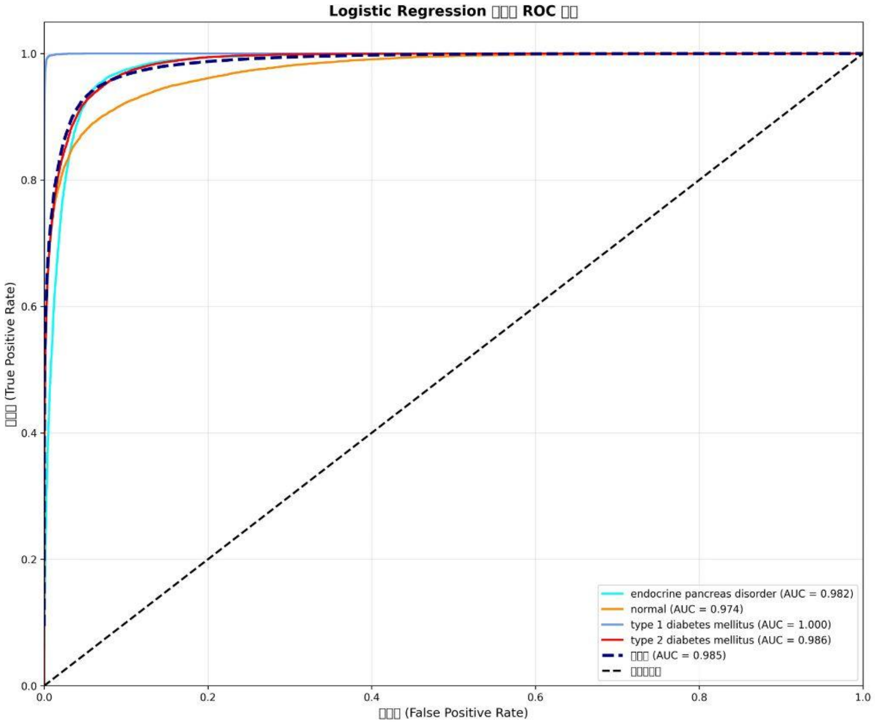
How much does cell 1 express of gene 0

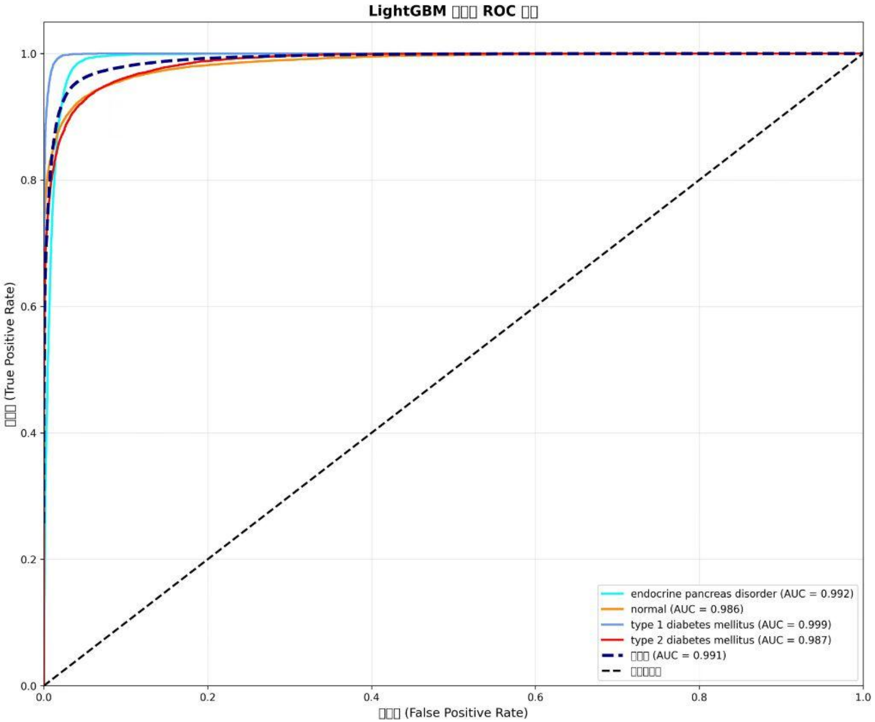
Confusion matrix



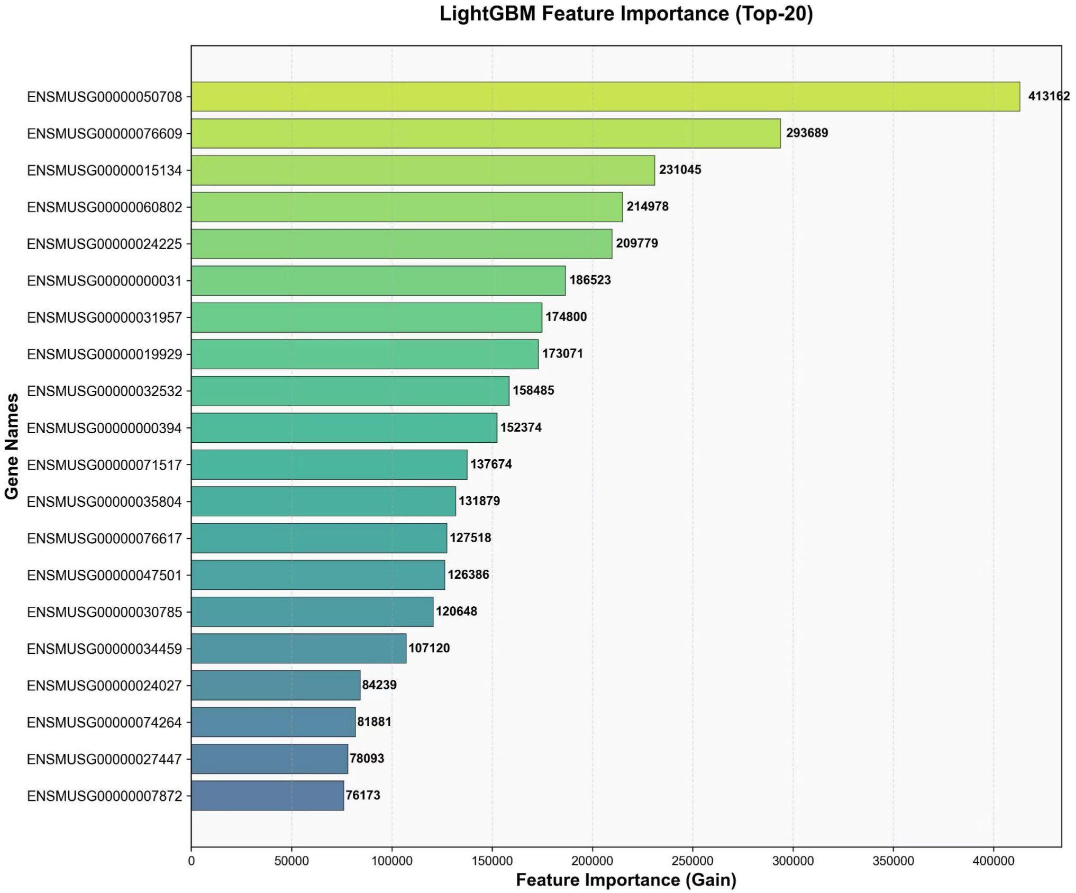


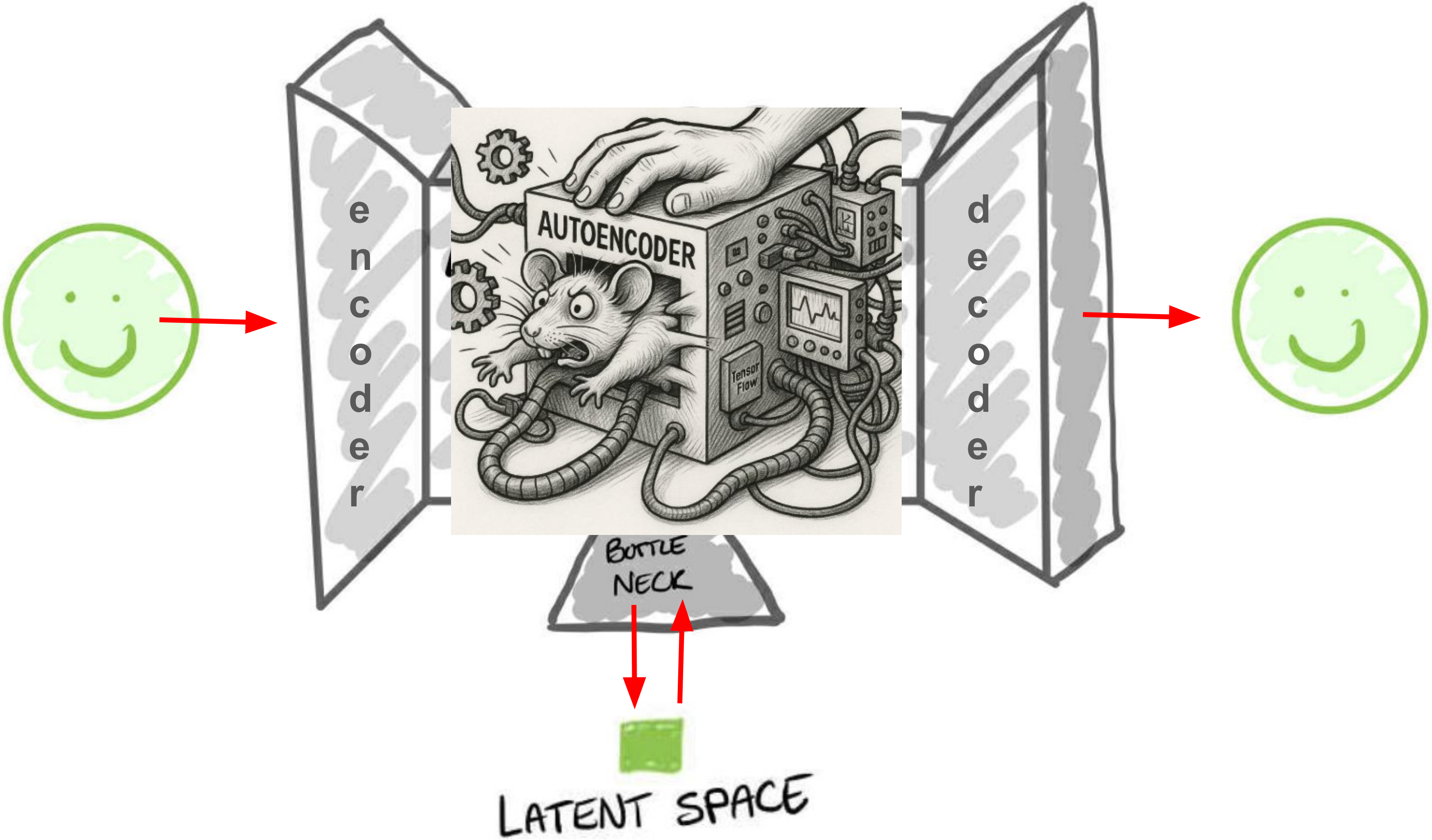
One-vs-Rest ROC Curves





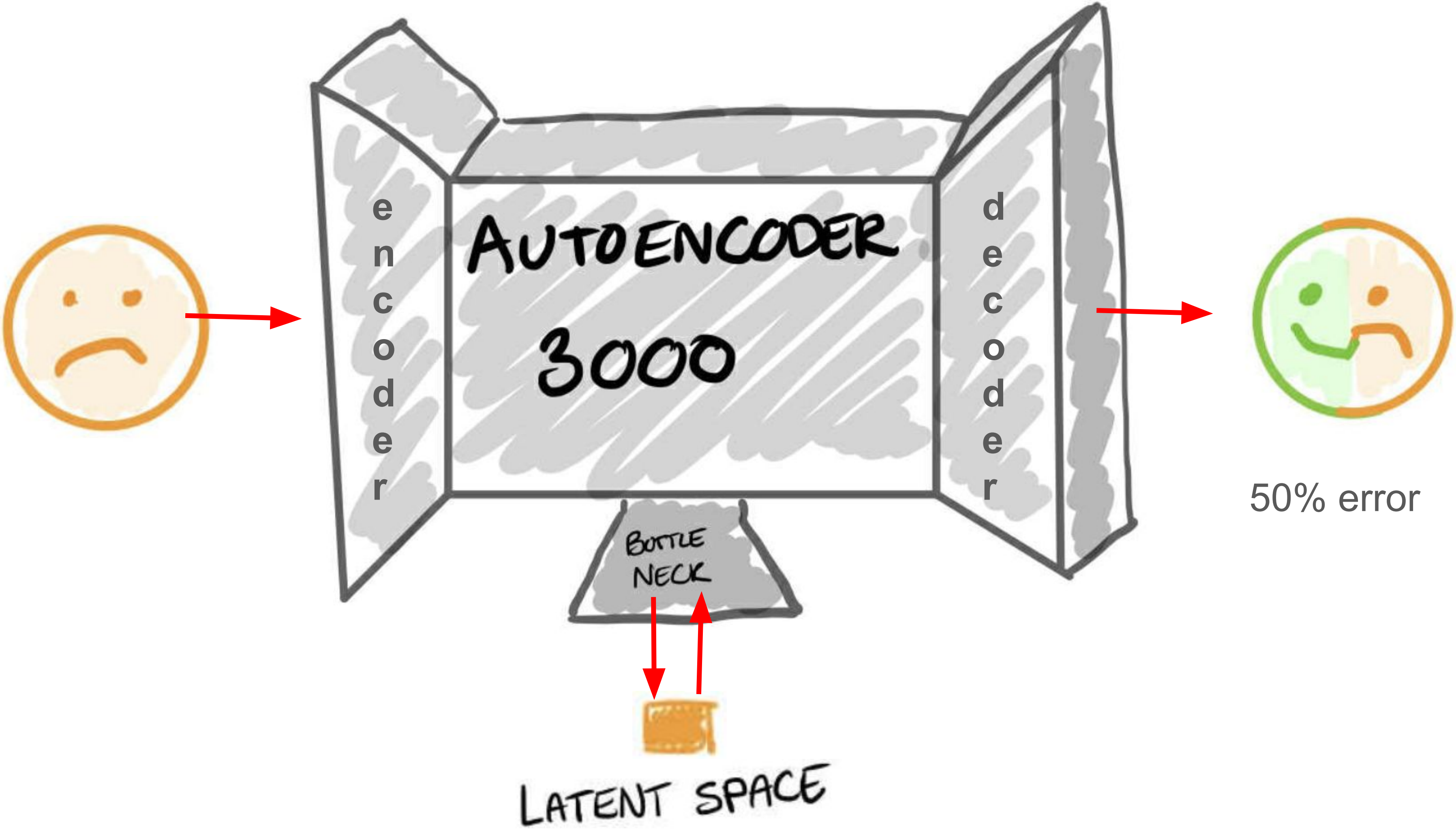
LightGBM Feature Importance





encoder

decoder



encoder

decoder

50% error

Initial processing of data

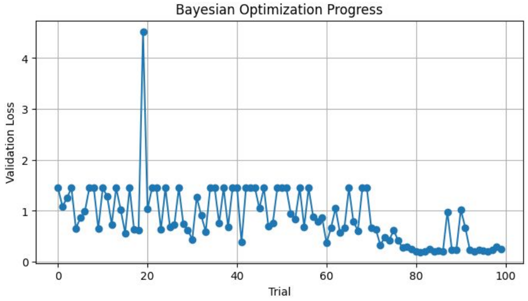
Different ways to do so:

- Raw data
- initial PCA
- most important genes

Should initial processing include sick cells?

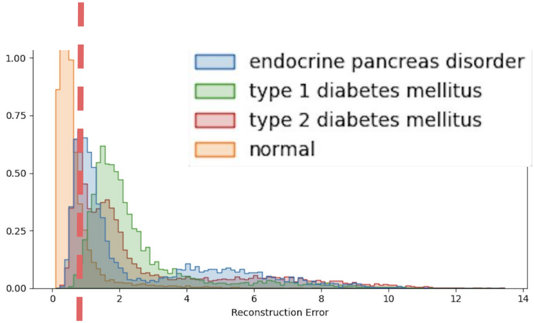
Training on **healthy** cells

Optimizing HP with Bayesian Optimization



Determine threshold

How many false positive or false negative do we accept?



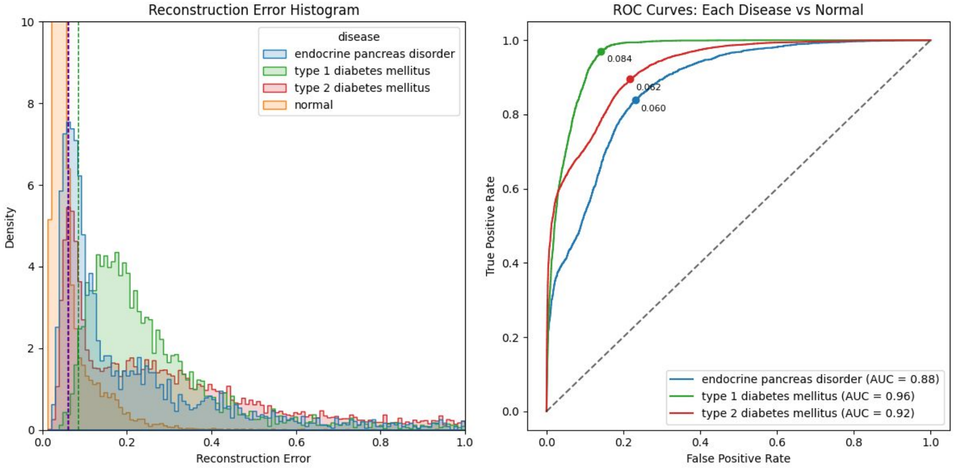
sick

Initial processing

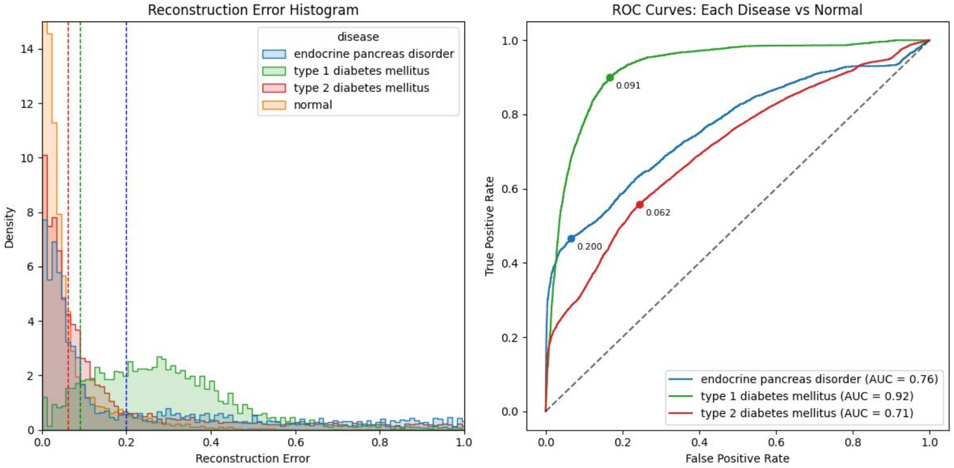
	Raw data	**PCA**	**Highly variable**
Pros	No information loss	Reduces dimensionality, keeps main variation	Very fast, easy to interpret
Cons	Slow optimization, (10+ min pr. trial)	Linear method - misses nonlinear structure	Removes many genes without modeling relationships

Results with preprocessing

Does quite well, but the preprocessing includes sick cells



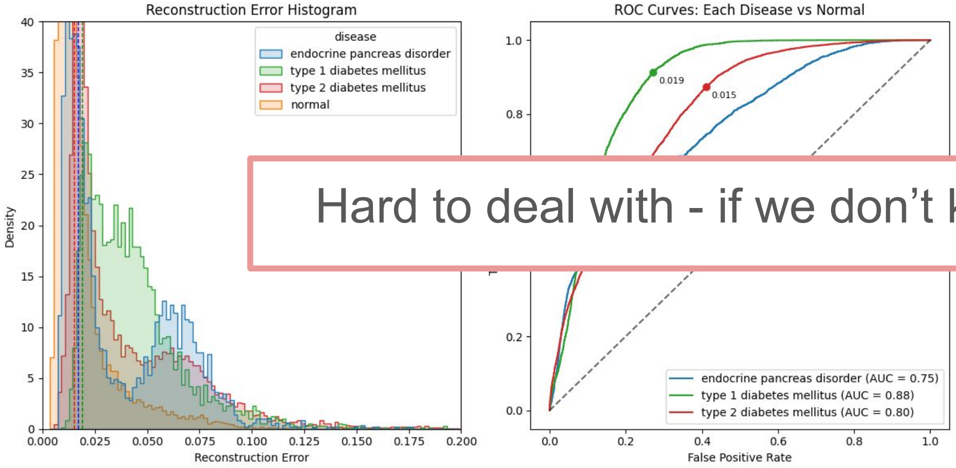
PCA 80 components



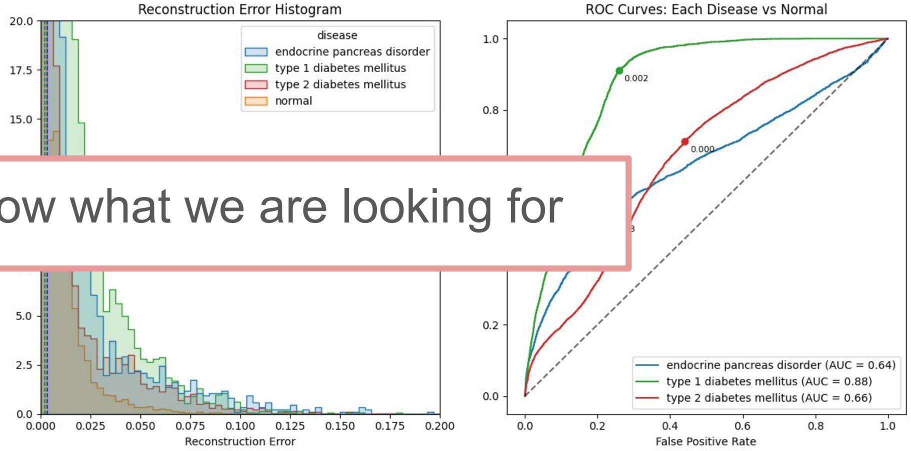
Highly variable top 80 genes

Can it be made more general?

Only seen healthy data, also in preprocessing data



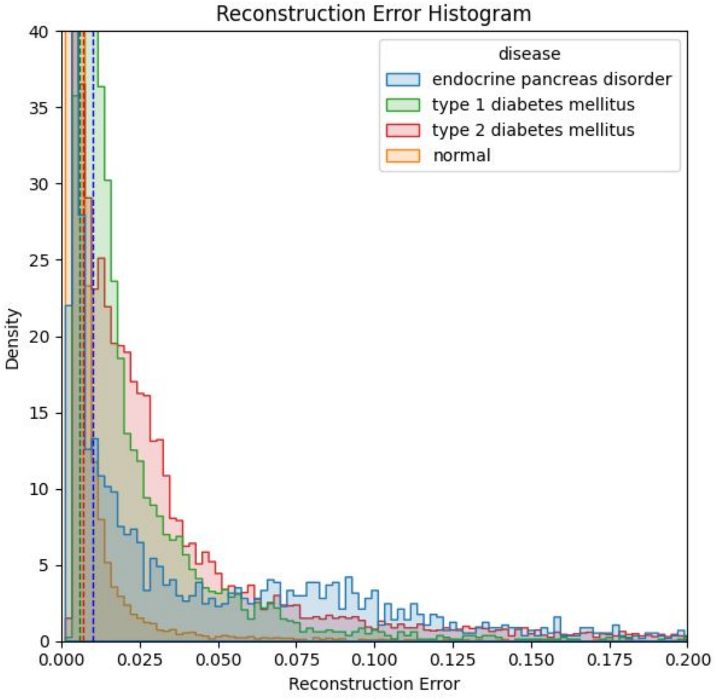
PCA 80 components

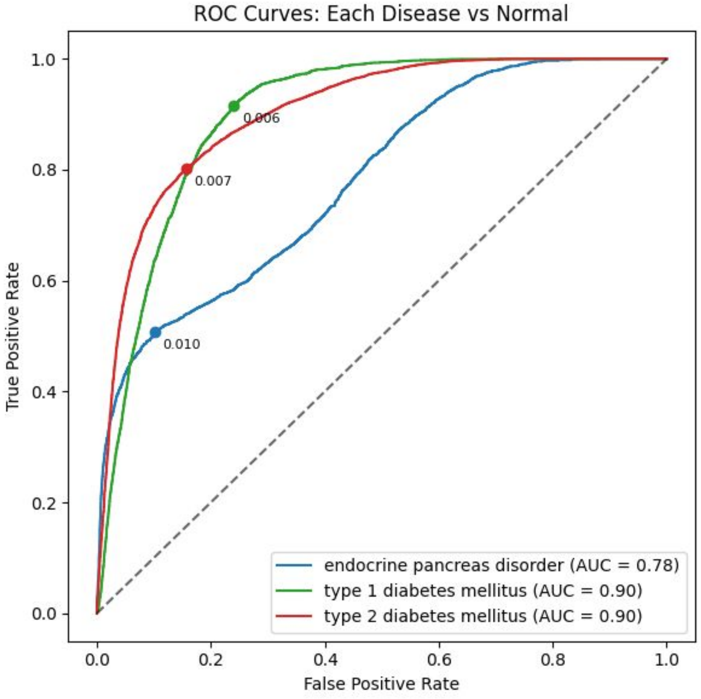


Highly variable top 80 genes

Hard to deal with - if we don’t know what we are looking for

Combined supervised and unsupervised

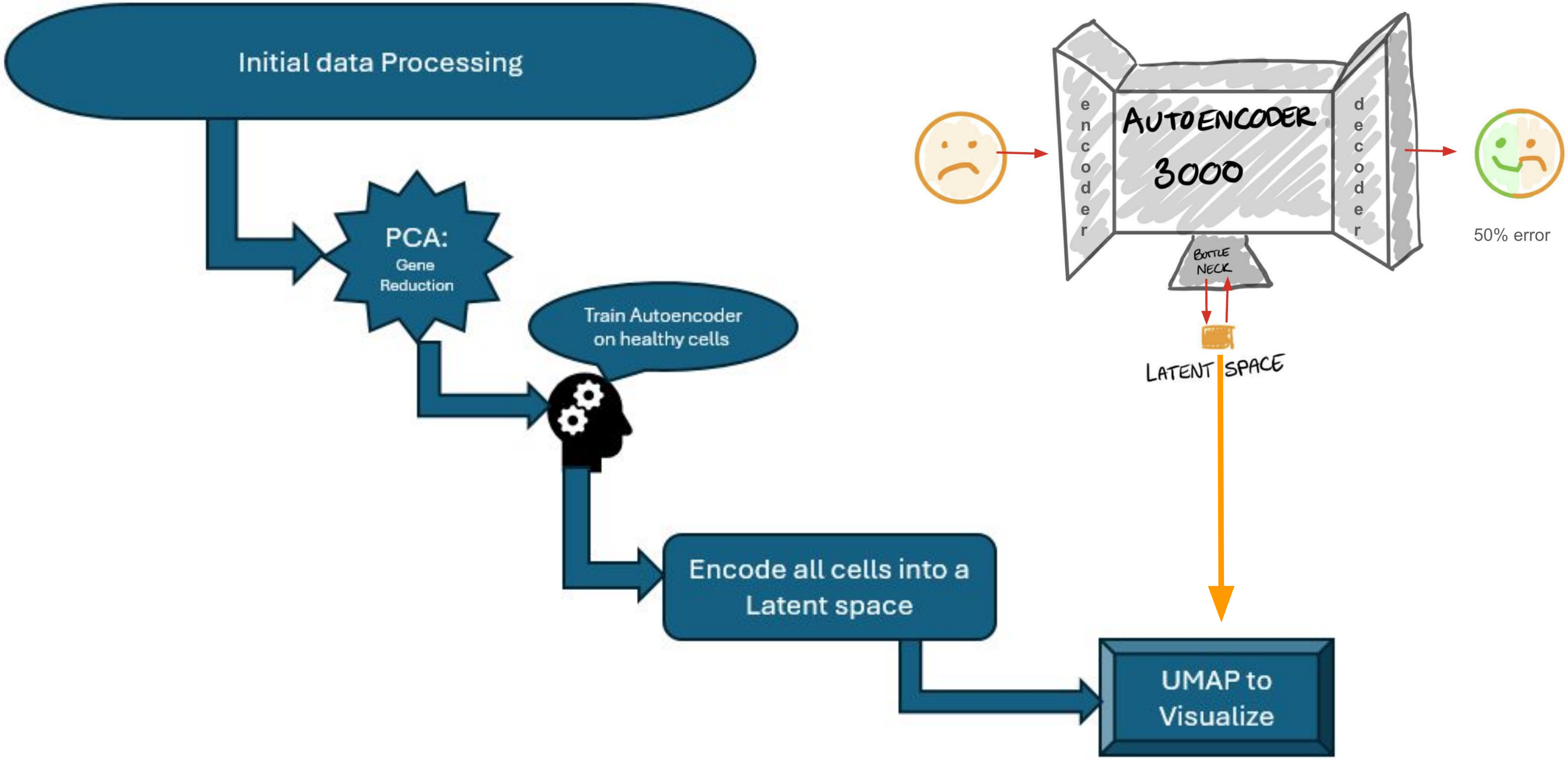




Better than top 80 variable, only slightly worse than PCA

- actually, it wasn’t top 80 shap, but random from top 1000 shap

First use of Auto encoder → workflow



Initial data Processing

PCA:
Gene Reduction

Train Autoencoder on healthy cells

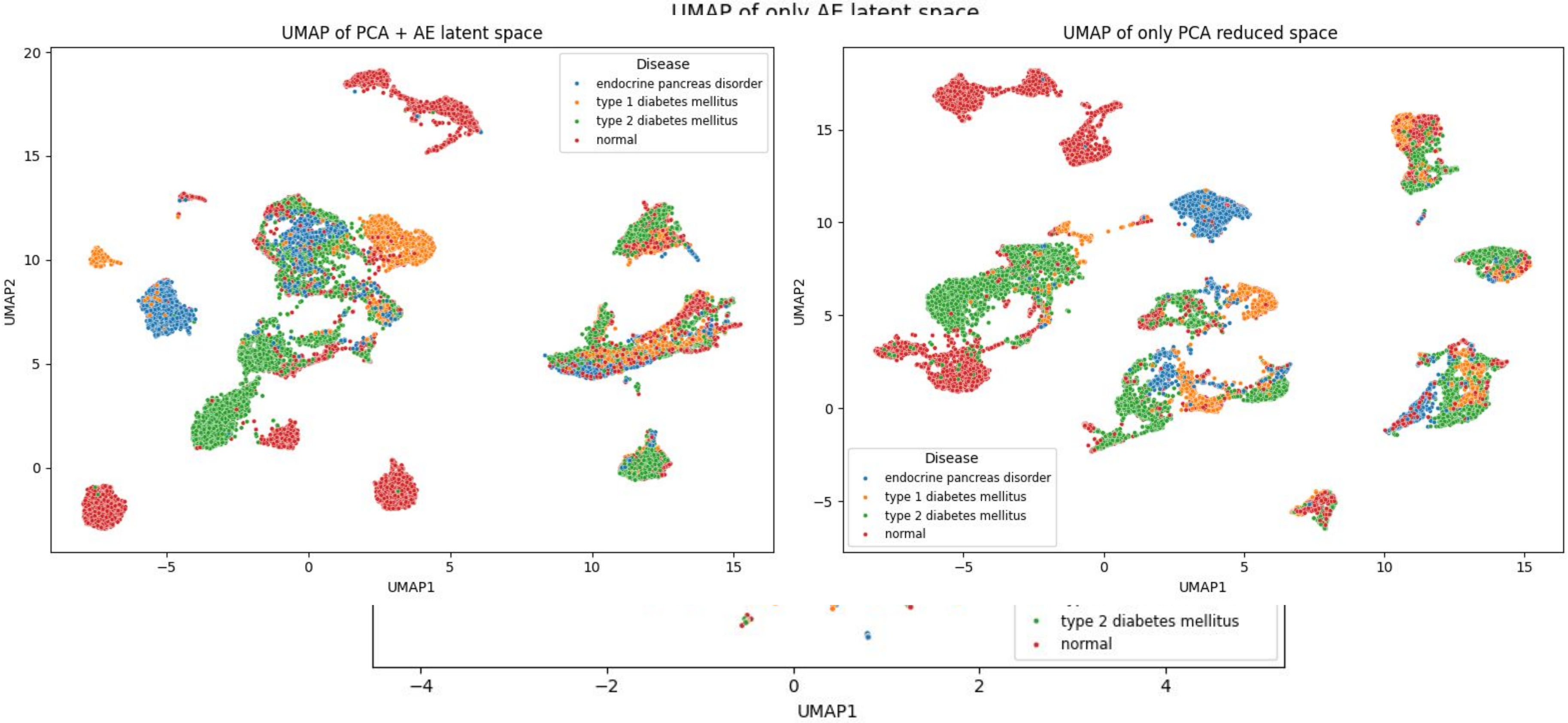
Encode all cells into a Latent space

UMAP to Visualize

50% error

UMAP Visualization of combined data

Only training on Normal Cells



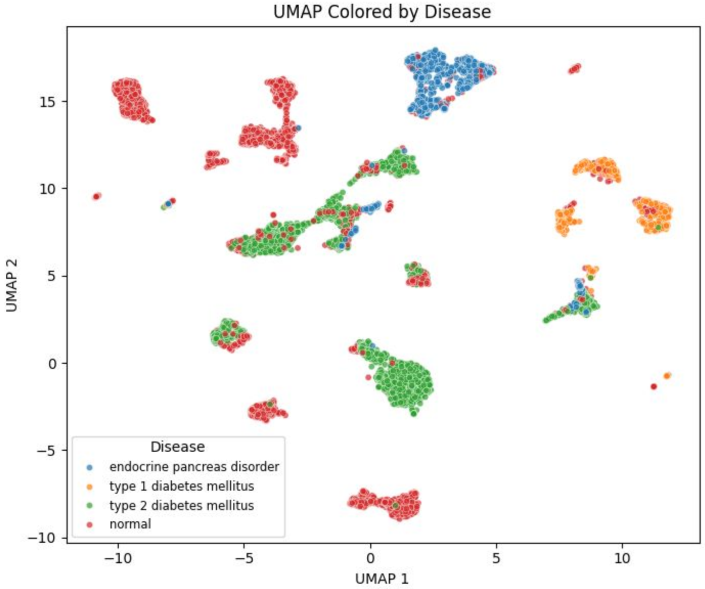
UMAP of only AE latent space

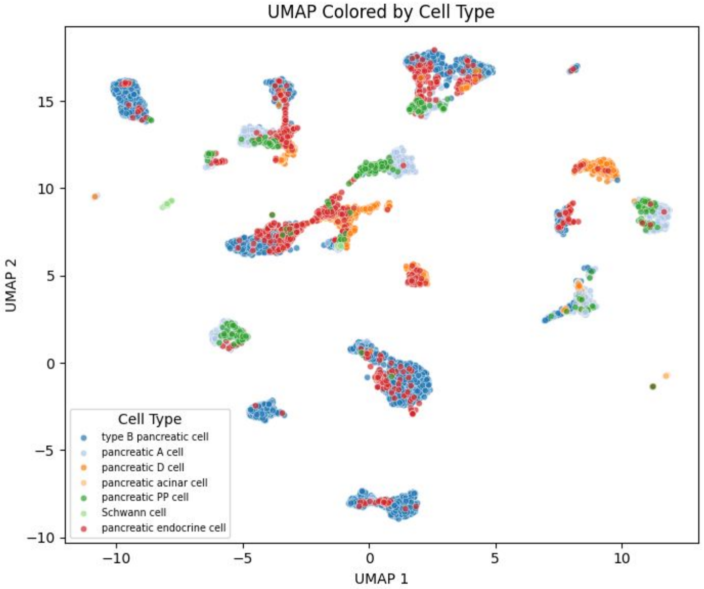
UMAP of PCA + AE latent space

UMAP of only PCA reduced space

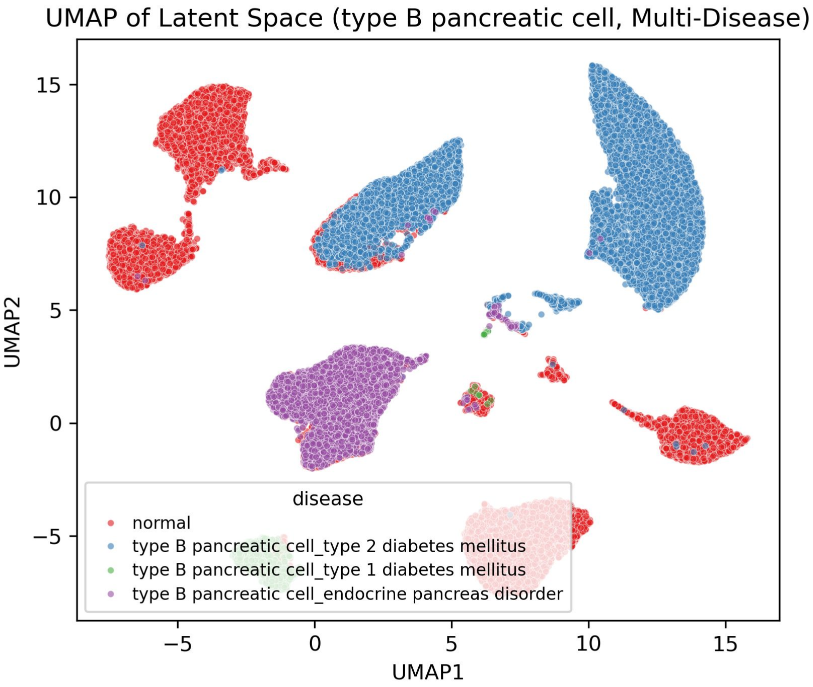
Can we find better clustering?

Using only PCA

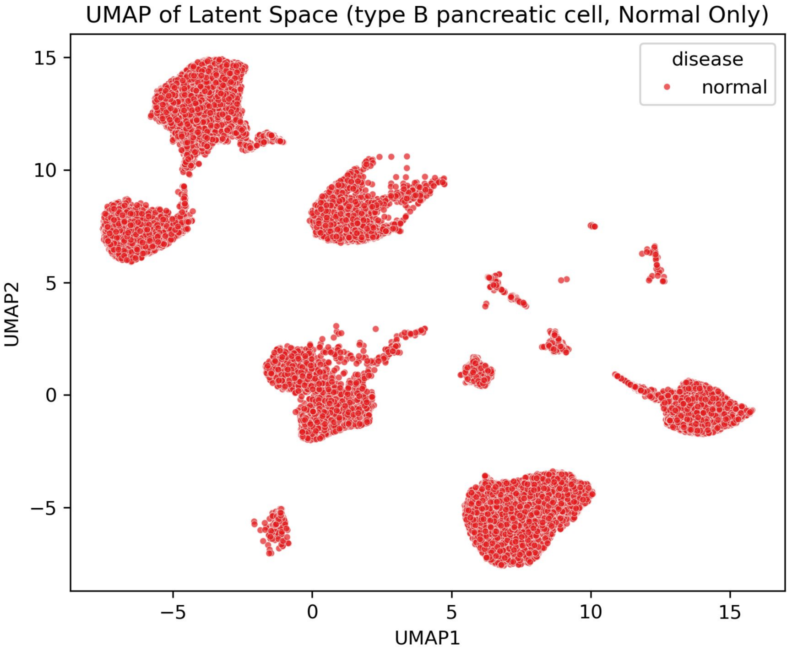




Visualization: Beta Cells



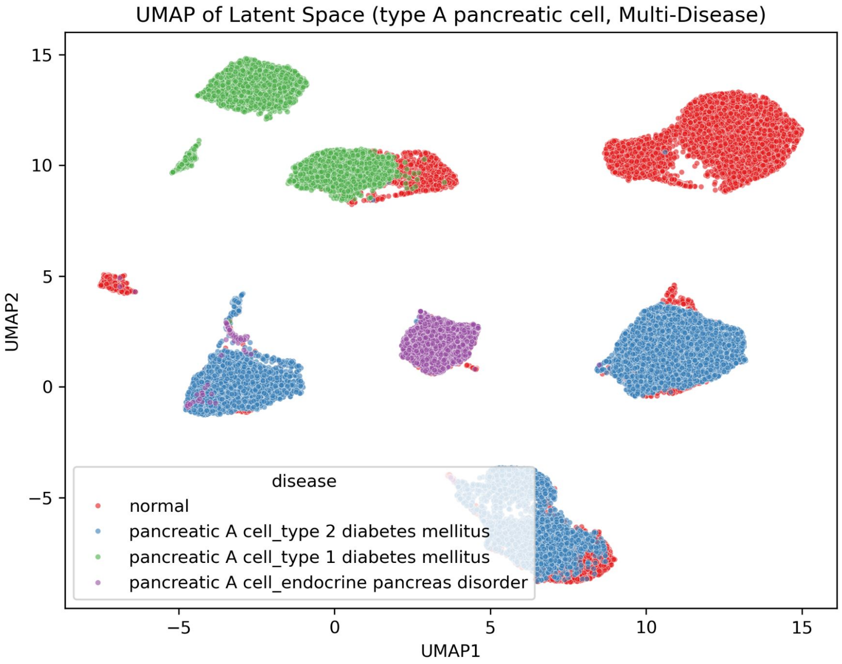
UMAP of Latent Space (type B pancreatic cell, Multi-Disease)



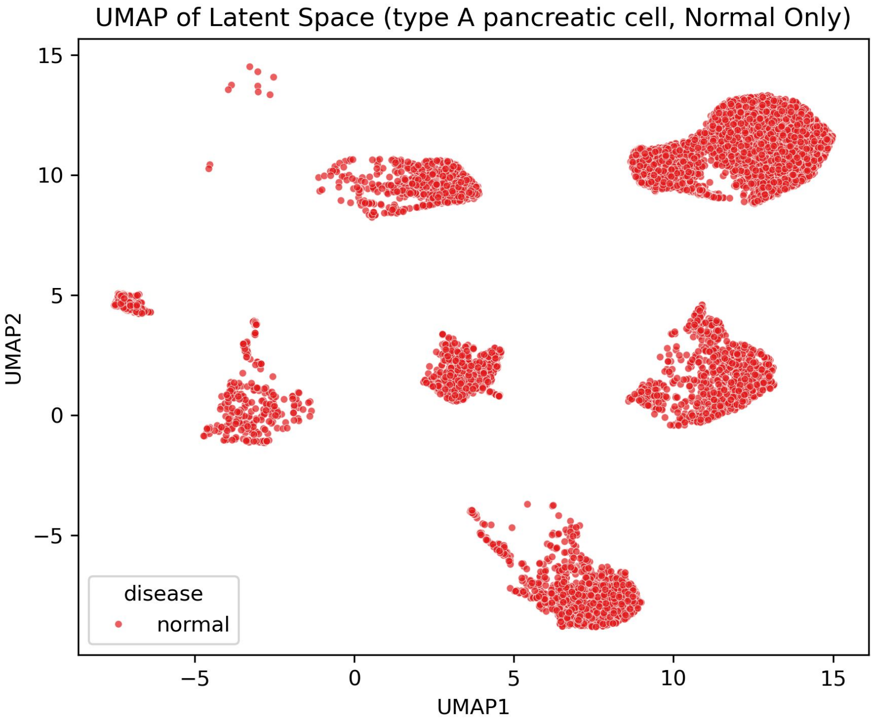
UMAP of Latent Space (type B pancreatic cell, Normal Only)

Indicating that beta cells with T2D behave differently from healthy ones, based on patterns learned by the model on healthy ones.

Visualization: Alpha Cells



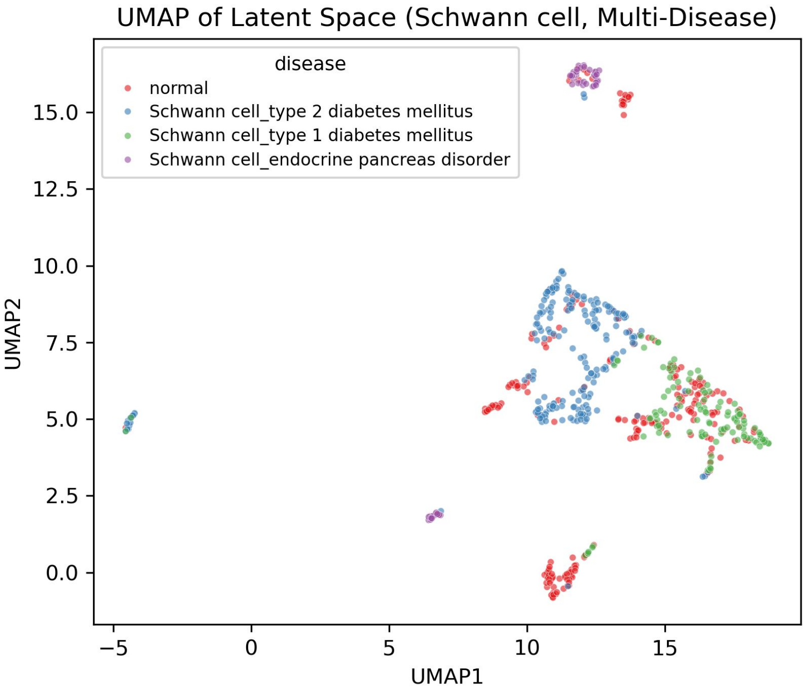
UMAP of Latent Space (type A pancreatic cell, Multi-Disease)

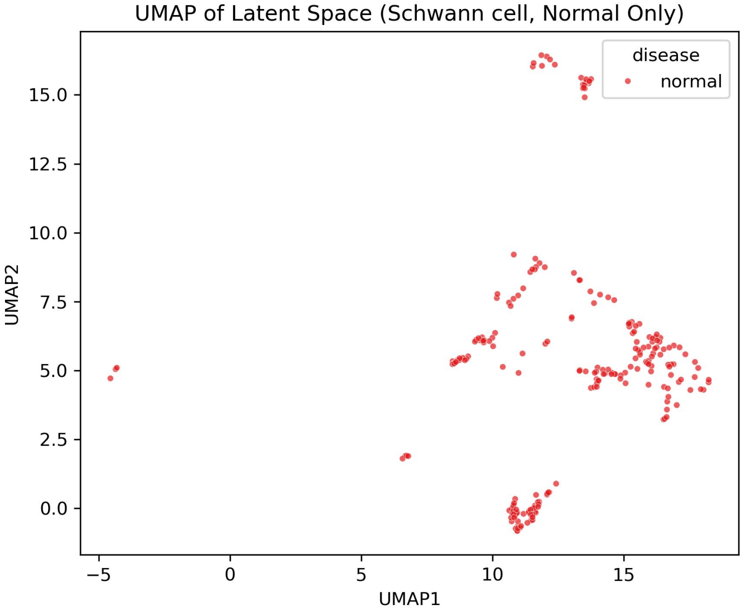


UMAP of Latent Space (type A pancreatic cell, Normal Only)

Providing insight on how Alpha cells with T1D behave differently from healthy ones, based on patterns learned by the model on healthy ones.

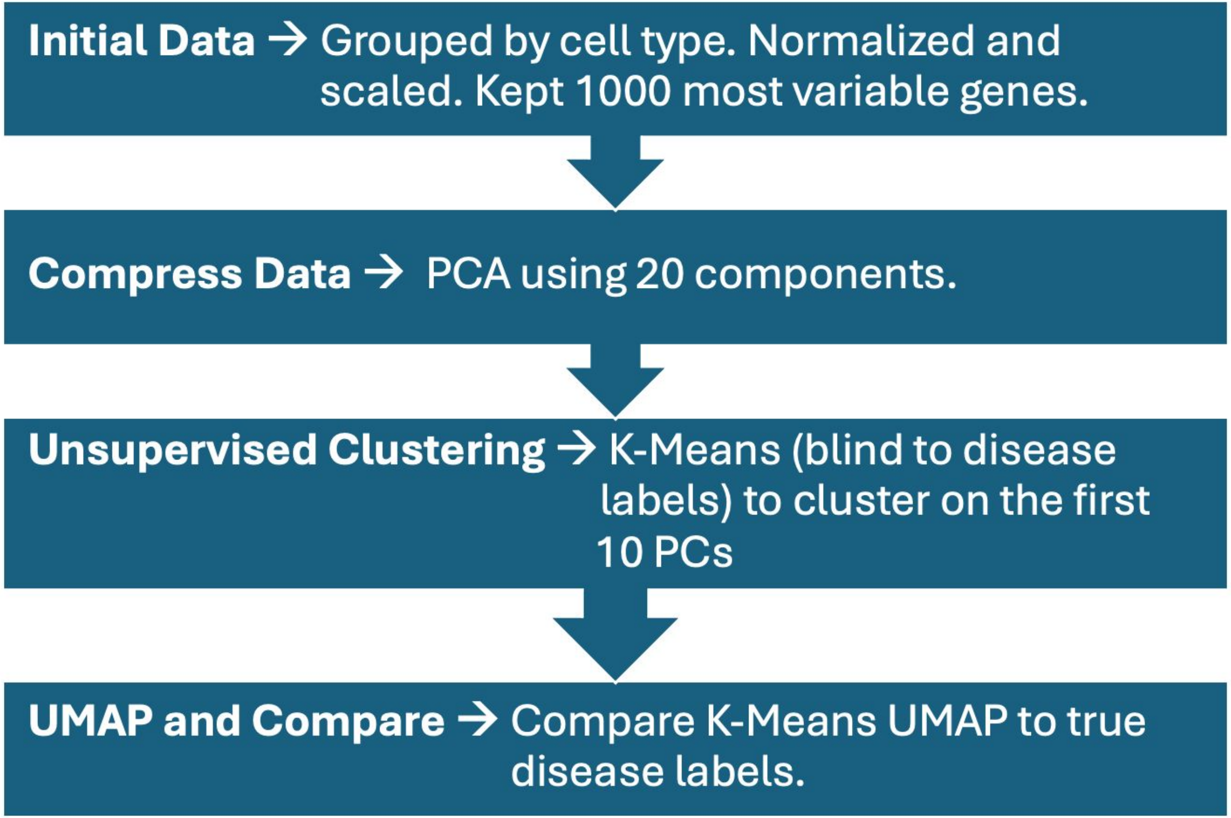
Visualization: Schwann cells





Model has limited ability to distinguish or characterize disease behaviour in Schwann cells. BUT WHY?

Clustering → Workflow

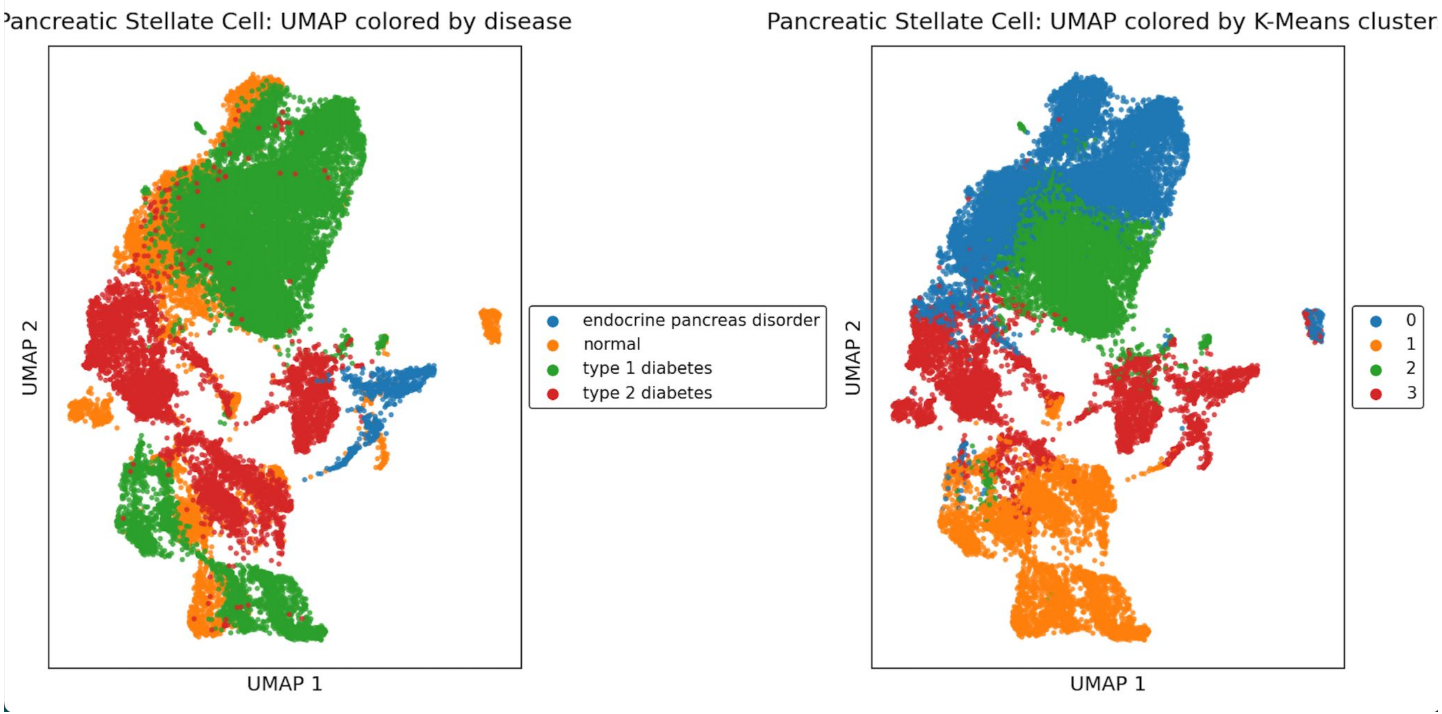


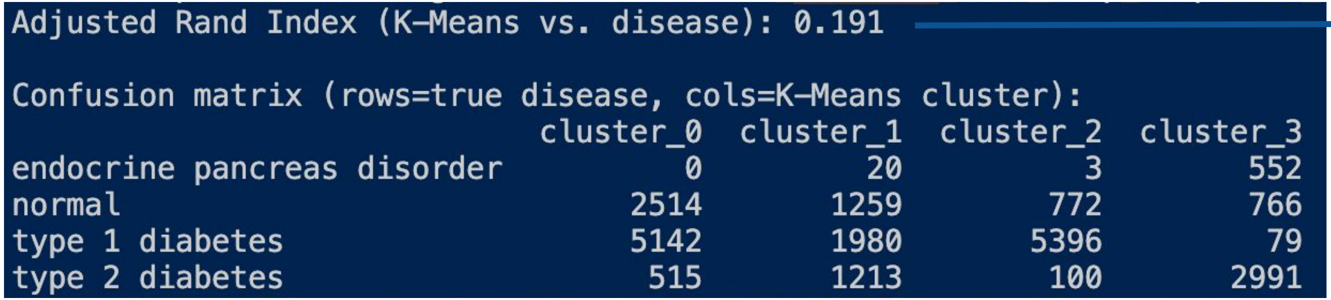
Initial Data → Grouped by cell type. Normalized and scaled. Kept 1000 most variable genes.

Compress Data → PCA using 20 components.

Unsupervised Clustering → K-Means (blind to disease labels) to cluster on the first 10 PCs

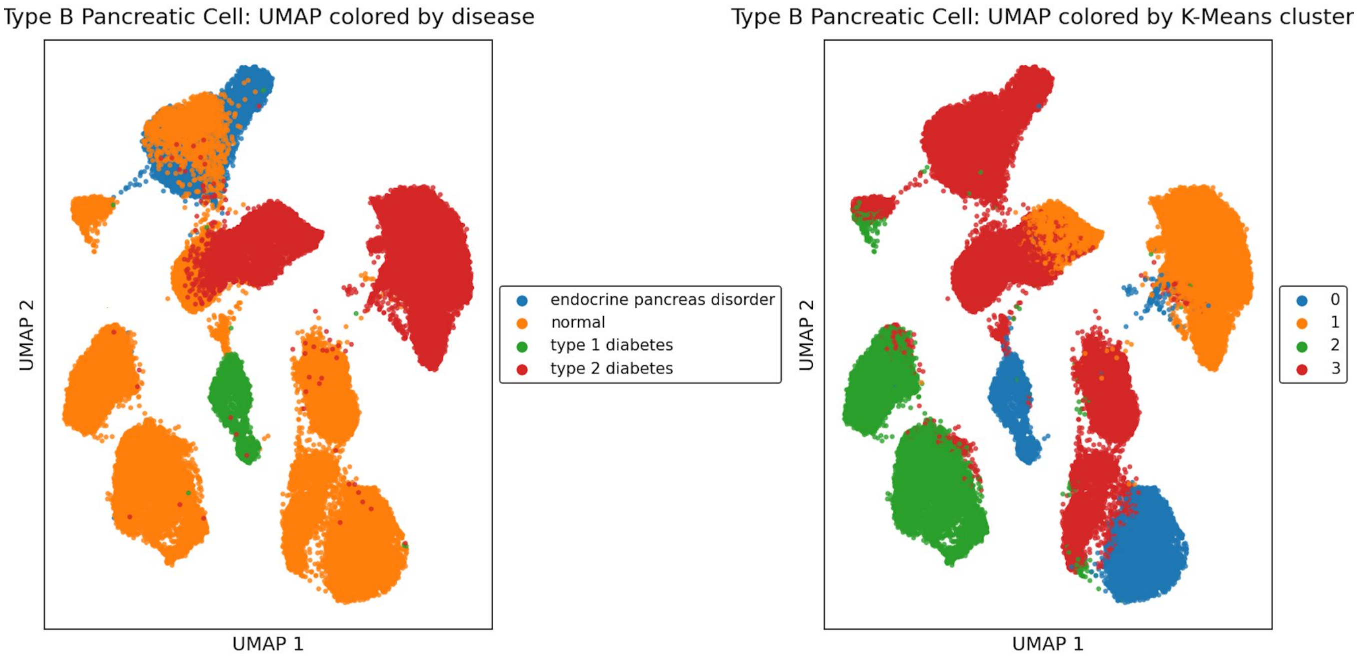
UMAP and Compare → Compare K-Means UMAP to true disease labels.

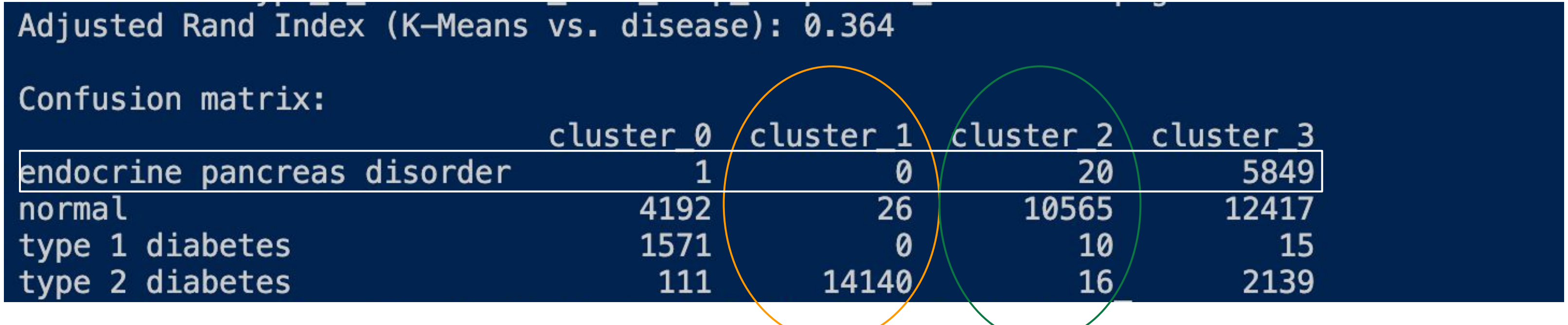




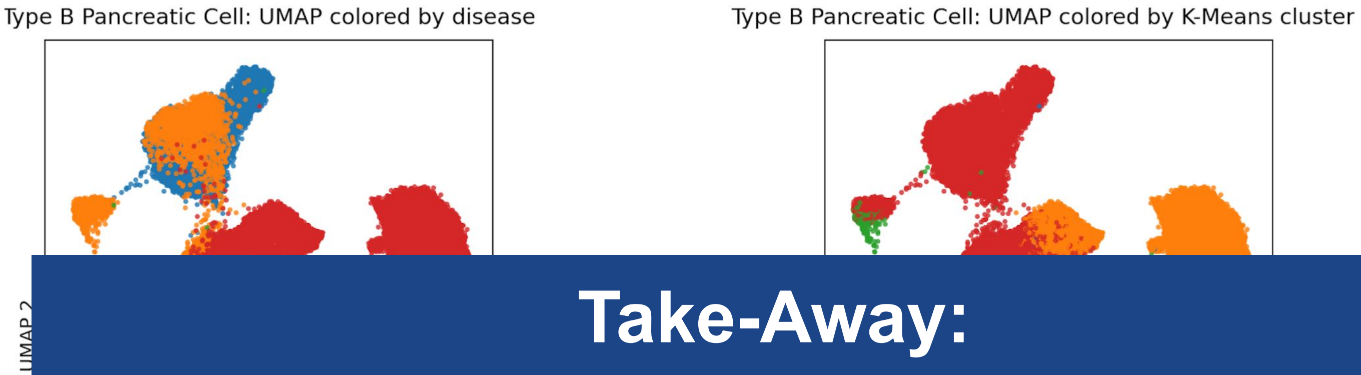
```
Adjusted Rand Index (K-Means vs. disease): 0.191

Confusion matrix (rows=true disease, cols=K-Means cluster):
                            cluster_0  cluster_1  cluster_2  cluster_3
endocrine pancreas disorder         0         20          3        552
normal                           2514       1259        772        766
type 1 diabetes                  5142       1980       5396         79
type 2 diabetes                   515       1213        100       2991
```

ARI:
1 = complete agreement
0 = random clusters





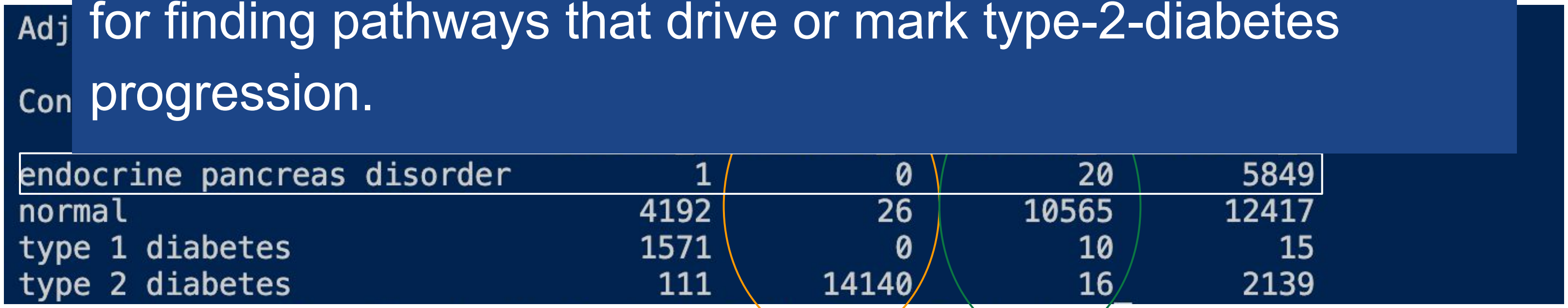
```
Adjusted Rand Index (K-Means vs. disease): 0.364

Confusion matrix:
                             cluster_0  cluster_1  cluster_2  cluster_3
endocrine pancreas disorder          1          0         20       5849
normal                            4192         26      10565      12417
type 1 diabetes                   1571          0         10         15
type 2 diabetes                    111      14140         16       2139
```

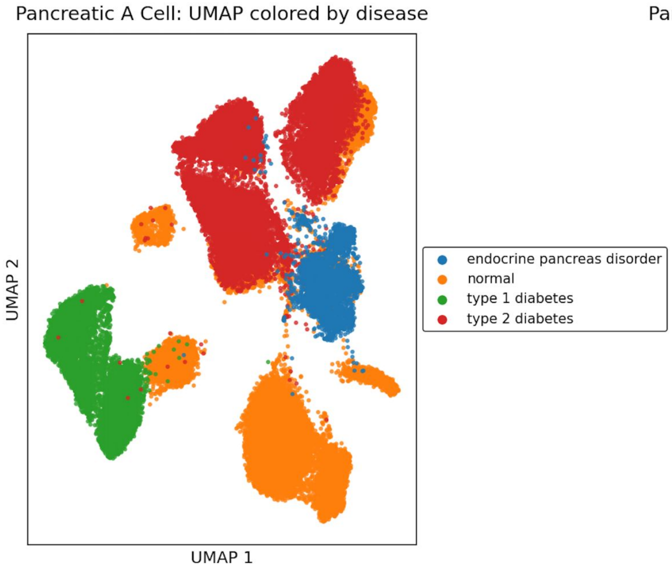

Type B Pancreatic Cell: UMAP colored by disease

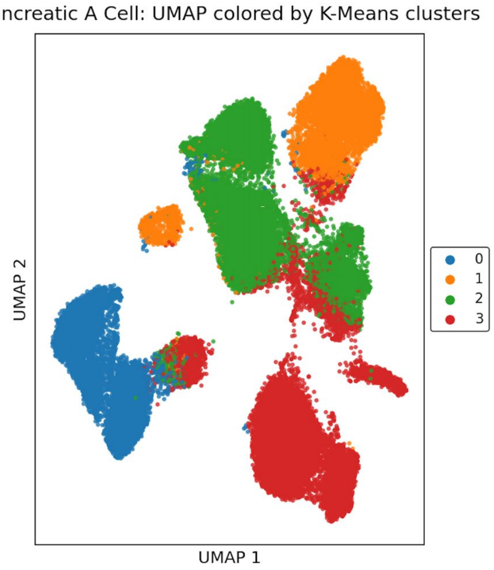
Type B Pancreatic Cell: UMAP colored by K-Means cluster

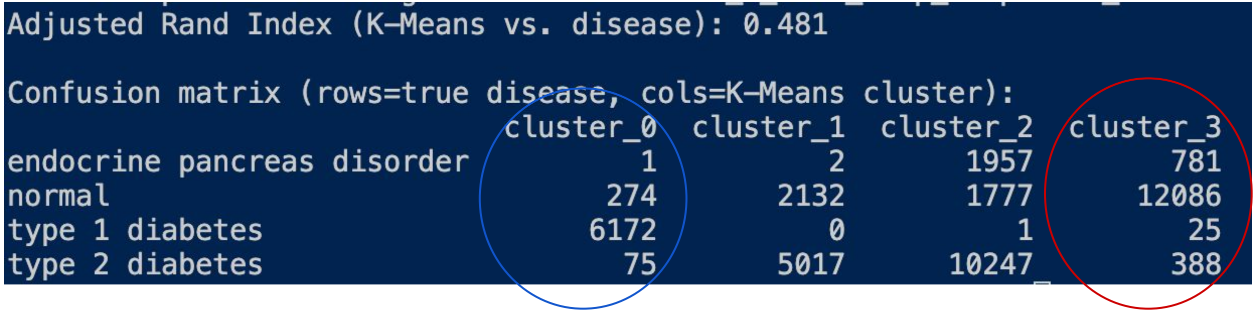
Take-Away:

A cluster made up solely of T2D pancreatic B cells means these cells share a *distinct gene-expression signature* that the algorithm can spot without knowing their diagnosis (unsupervised - remember!). That signature could be a target for finding pathways that drive or mark type-2-diabetes progression.


```
Adj

Con

endocrine pancreas disorder        1      0     20    5849
normal                          4192     26  10565   12417
type 1 diabetes                 1571      0     10      15
type 2 diabetes                  111  14140     16    2139
```







```
Adjusted Rand Index (K-Means vs. disease): 0.481

Confusion matrix (rows=true disease, cols=K-Means cluster):
                            cluster_0  cluster_1  cluster_2  cluster_3
endocrine pancreas disorder         1          2       1957        781
normal                            274       2132       1777      12086
type 1 diabetes                  6172          0          1         25
type 2 diabetes                    75       5017      10247        388
```

Cluster 0 = 95% T1D
Accounts for 99% of total T1D

Cluster 3 = 91% Normal
Accounts for 74% of total Normal

Disease State	**Cell types where the signature is very clear –** > 80% of the cells in at least one cluster are the same disease	**Cell types where the signature is visible, but mixed –** 50 – 80 % purity	**Cell types where the signature is indistinguishable-** < 50 % purity
T1D	A cells D cells	B cells Ductal cells PP cells	Stellate cells Endocrine cells
T2D	B cells Ductal cells	A cells PP cells D cells Endocrine cells	Stellate cells
Pancreatic Endocrine Disorder	Endocrine cells	PP cells Ductal cells	A cells B cells D cells Stellate cells
Normal	D cells A cells B cells Ductal Cells PP Cells	Endocrine cells	Stellate cells

Concluding remarks

Biological data is COMPLICATED

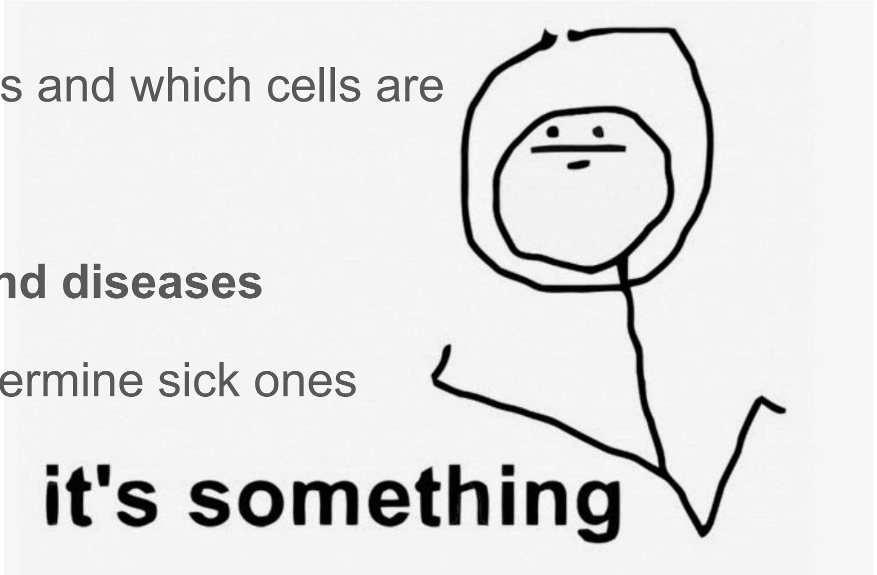
Can be good and optimized if the question is very specific

Know what disease we are looking for

different cross checks to see which genes and which cells are important for that specific disease

Hard to generalize across different cells and diseases

Not able to just see healthy cells and determine sick ones



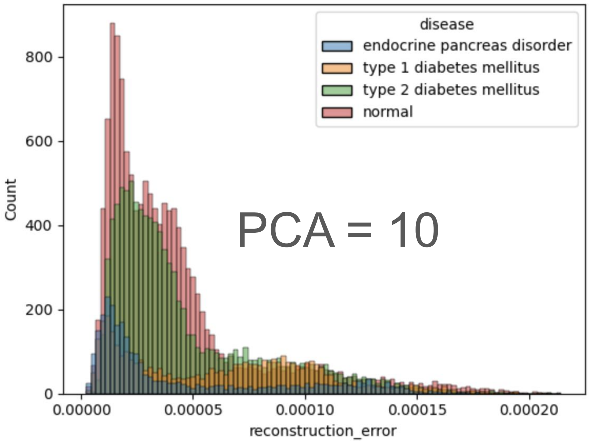
Appendix Slides

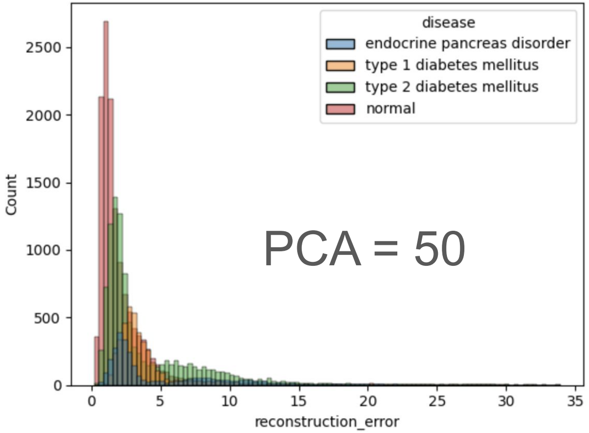
Top 5 important genes

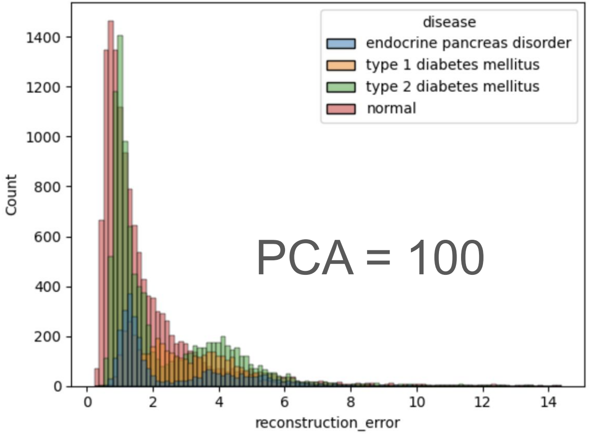
ENSMUSG00000050708	ferritin light polypeptide 1 Study shows strong connection between ferritin levels and diabetes
ENSMUSG00000076609	immunoglobulin kappa constant
ENSMUSG00000015134	aldehyde dehydrogenase family 1, subfamily A3
ENSMUSG00000060802	beta-2 microglobulin
ENSMUSG00000024225	colipase, pancreatic

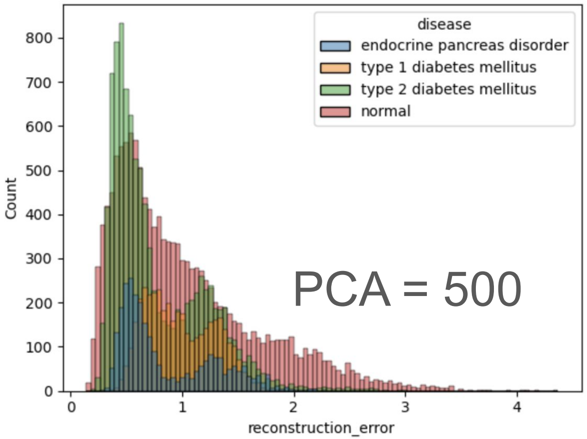
input(50) - encode(64) - encode(16) - decode(64) - output(50)

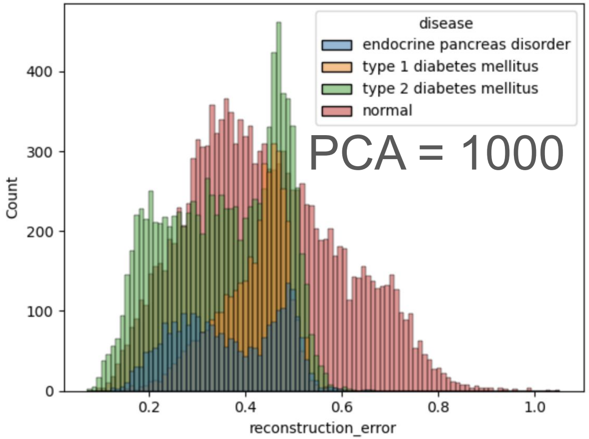
Maja’s (bonus) slide

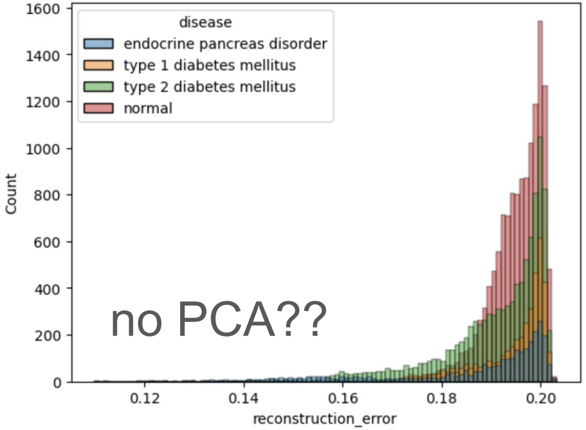

PCA = 10


PCA = 50

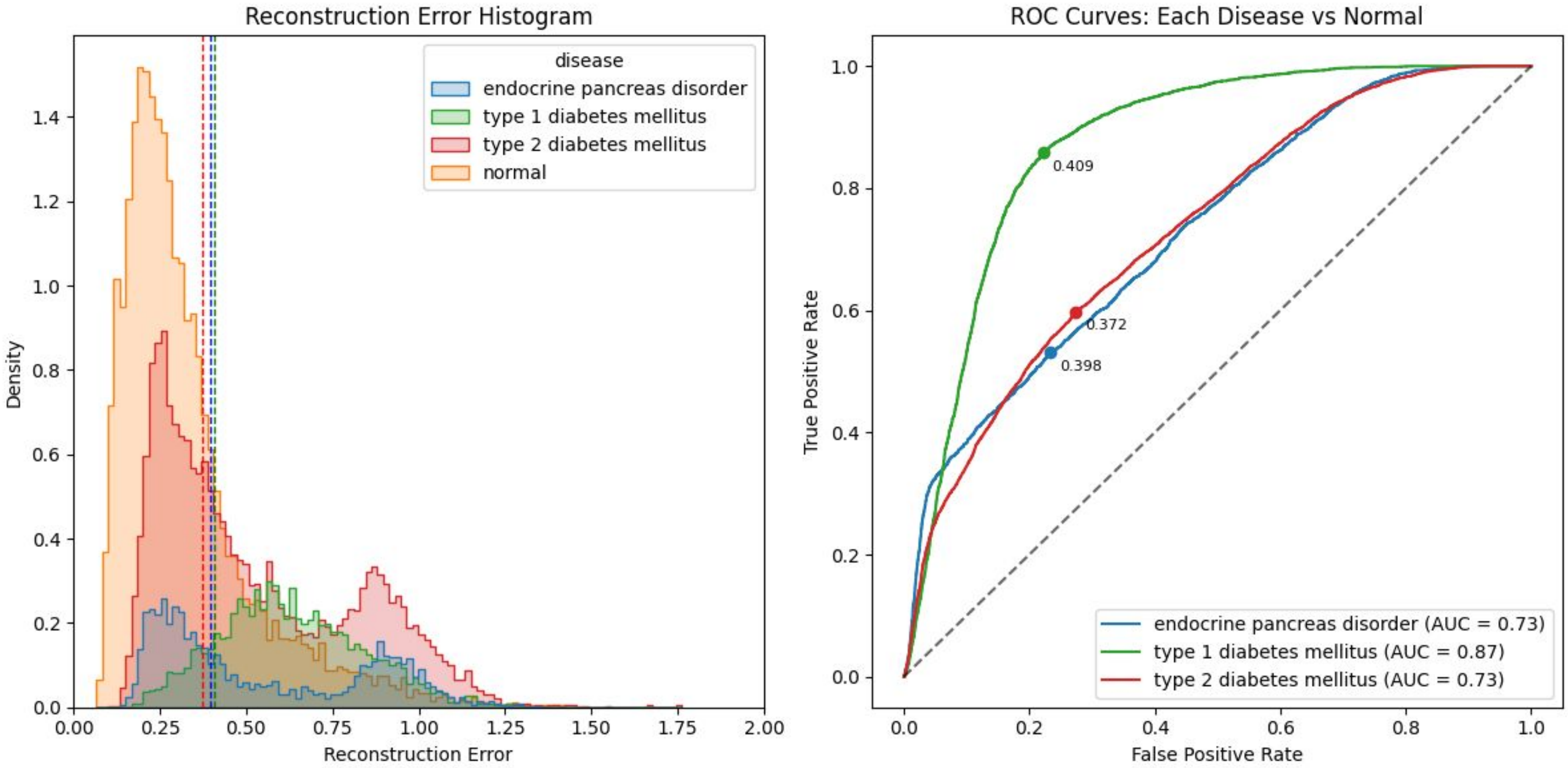

PCA = 100


PCA = 500


PCA = 1000

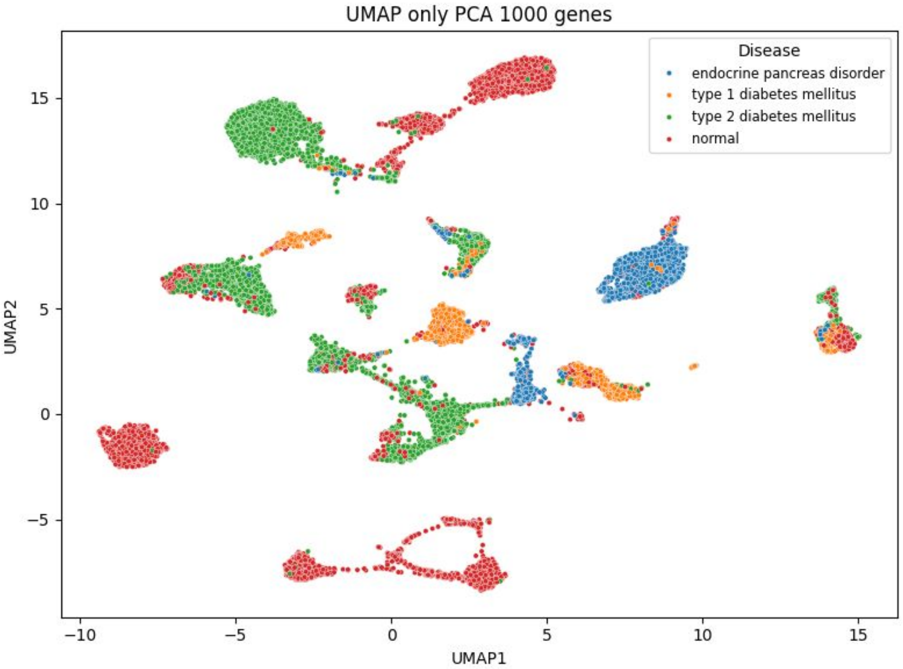

no PCA??

Maja’s (bonus) slide

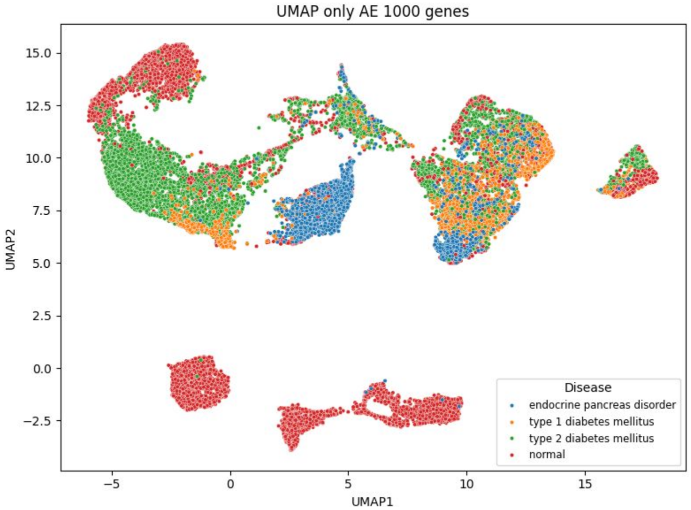


Top 1000 shap values on smaller AutoEncoder - too much noise, or needs larger AutoEncoder model

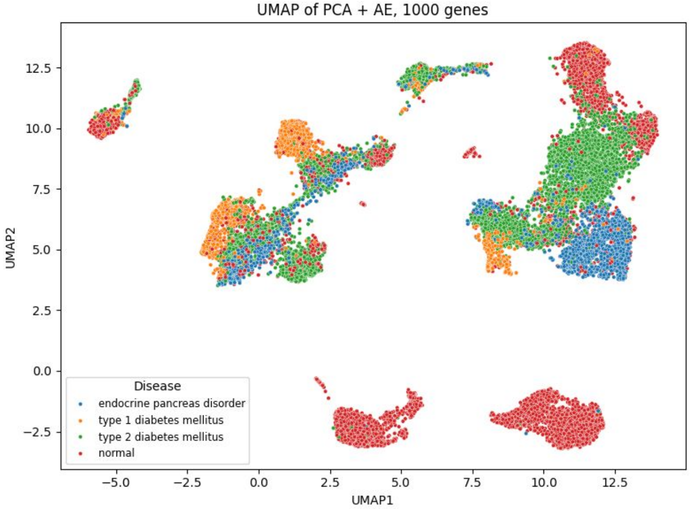
Reducing dimensions from 1000 genes



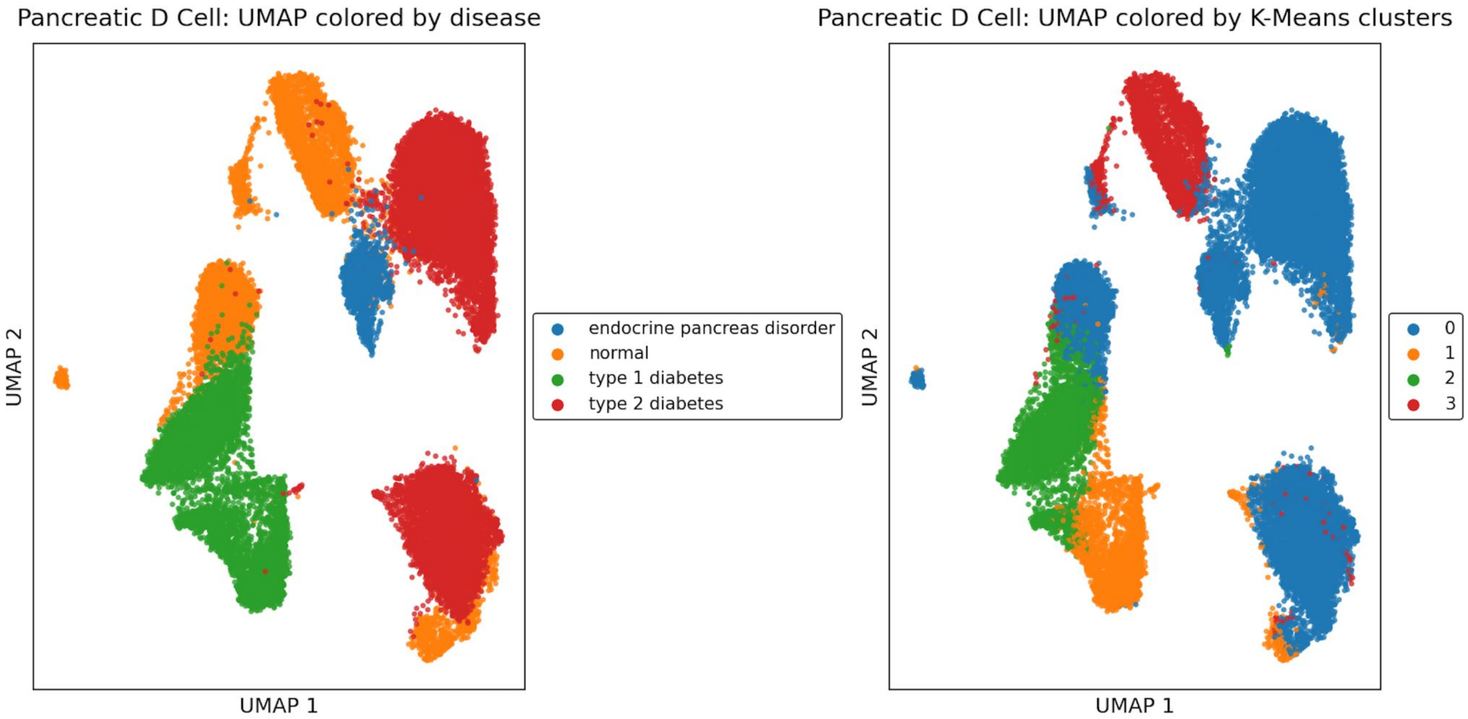
UMAP only PCA 1000 genes



UMAP only AE 1000 genes

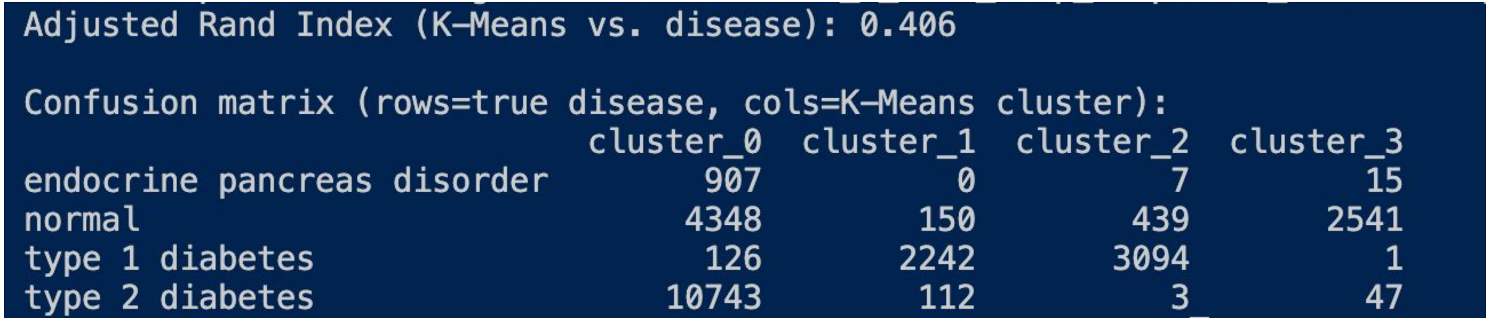


UMAP of PCA + AE, 1000 genes

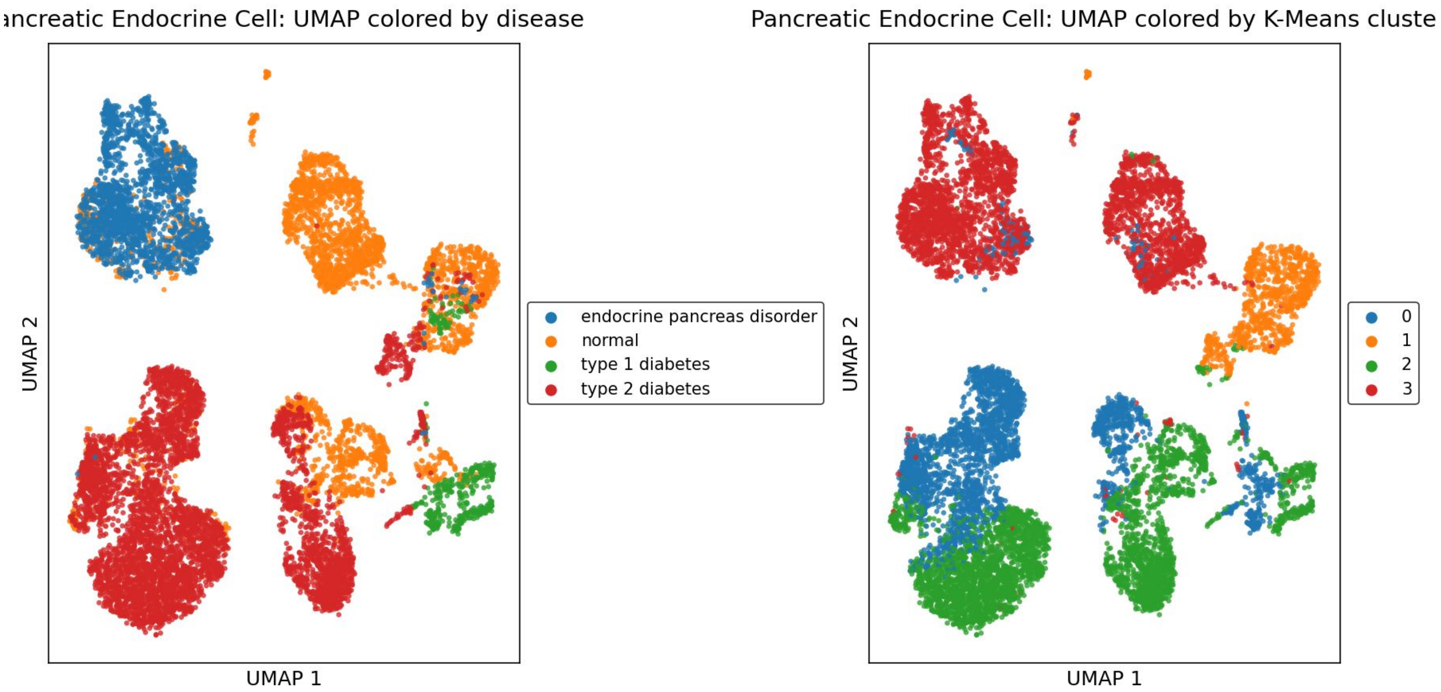


Pancreatic D Cell: UMAP colored by disease

Pancreatic D Cell: UMAP colored by K-Means clusters

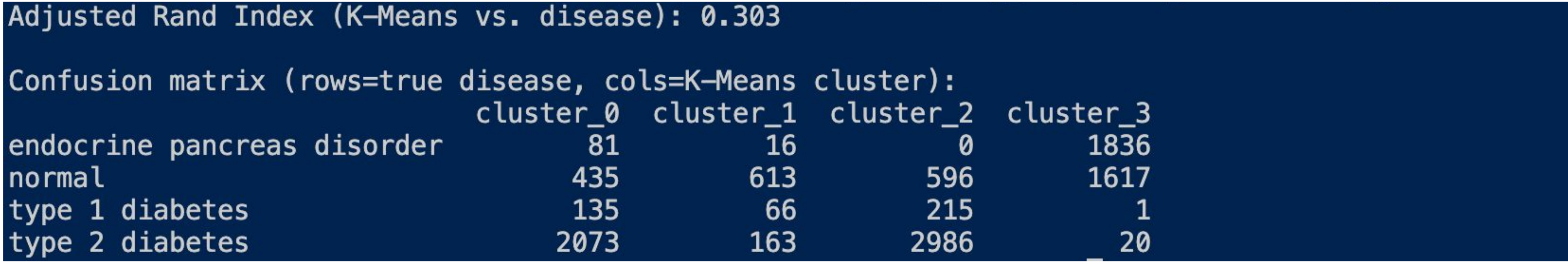


```
Adjusted Rand Index (K-Means vs. disease): 0.406

Confusion matrix (rows=true disease, cols=K-Means cluster):
                             cluster_0  cluster_1  cluster_2  cluster_3
endocrine pancreas disorder        907          0          7         15
normal                            4348        150        439       2541
type 1 diabetes                    126       2242       3094          1
type 2 diabetes                  10743        112          3         47
```

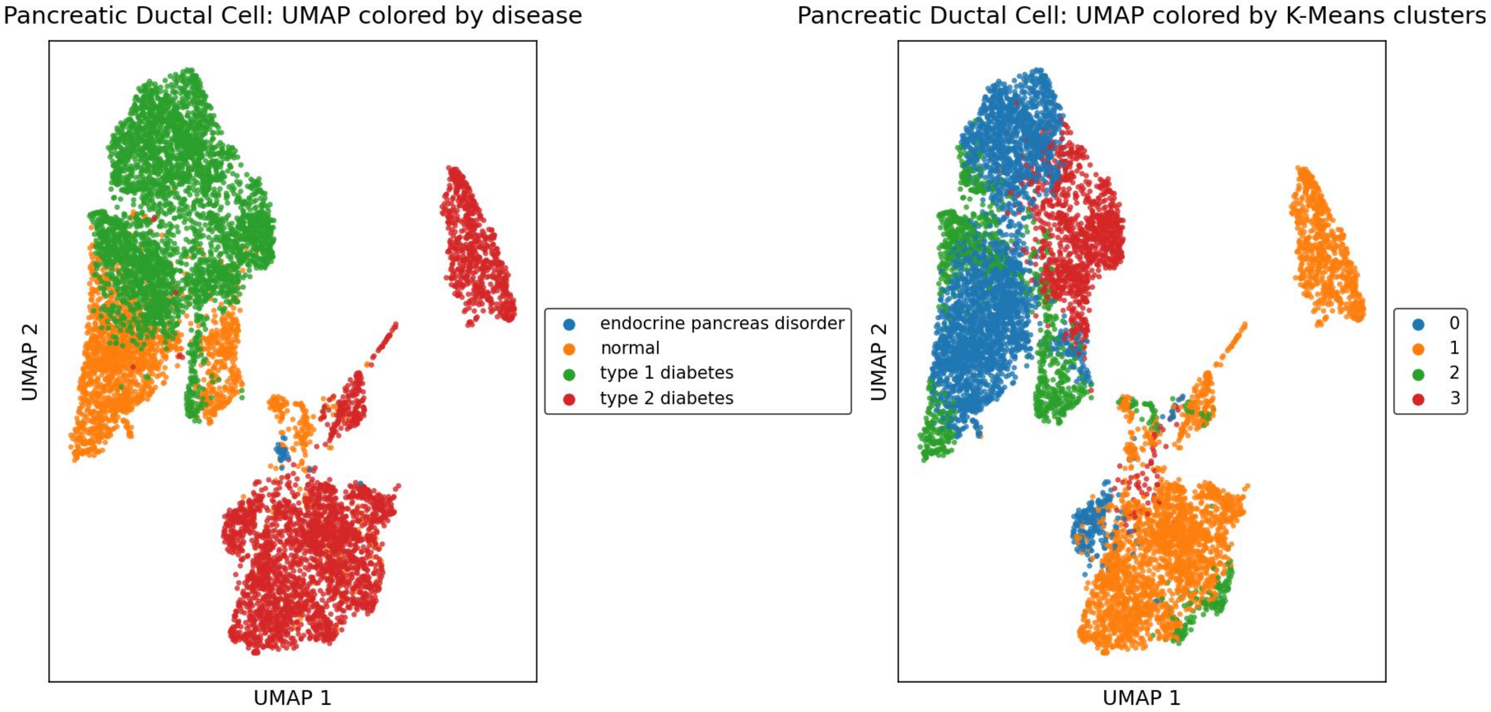


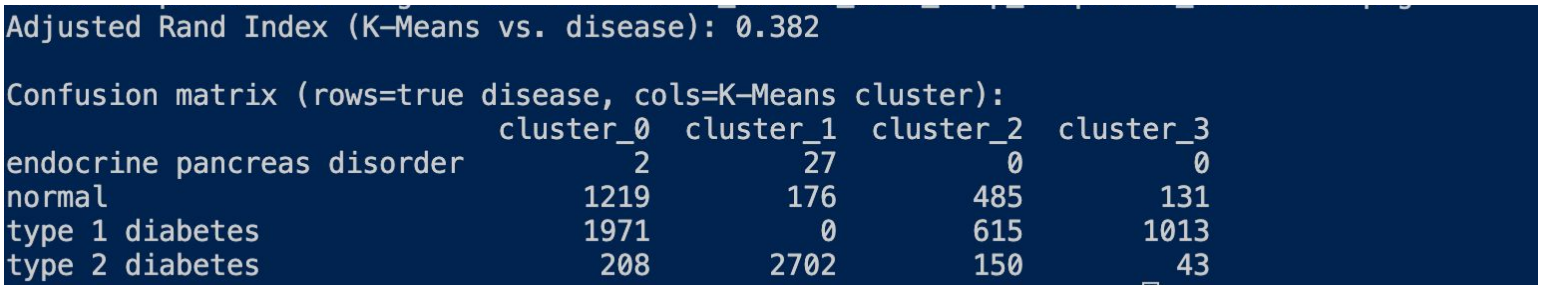
ancreatic Endocrine Cell: UMAP colored by disease

Pancreatic Endocrine Cell: UMAP colored by K-Means cluste

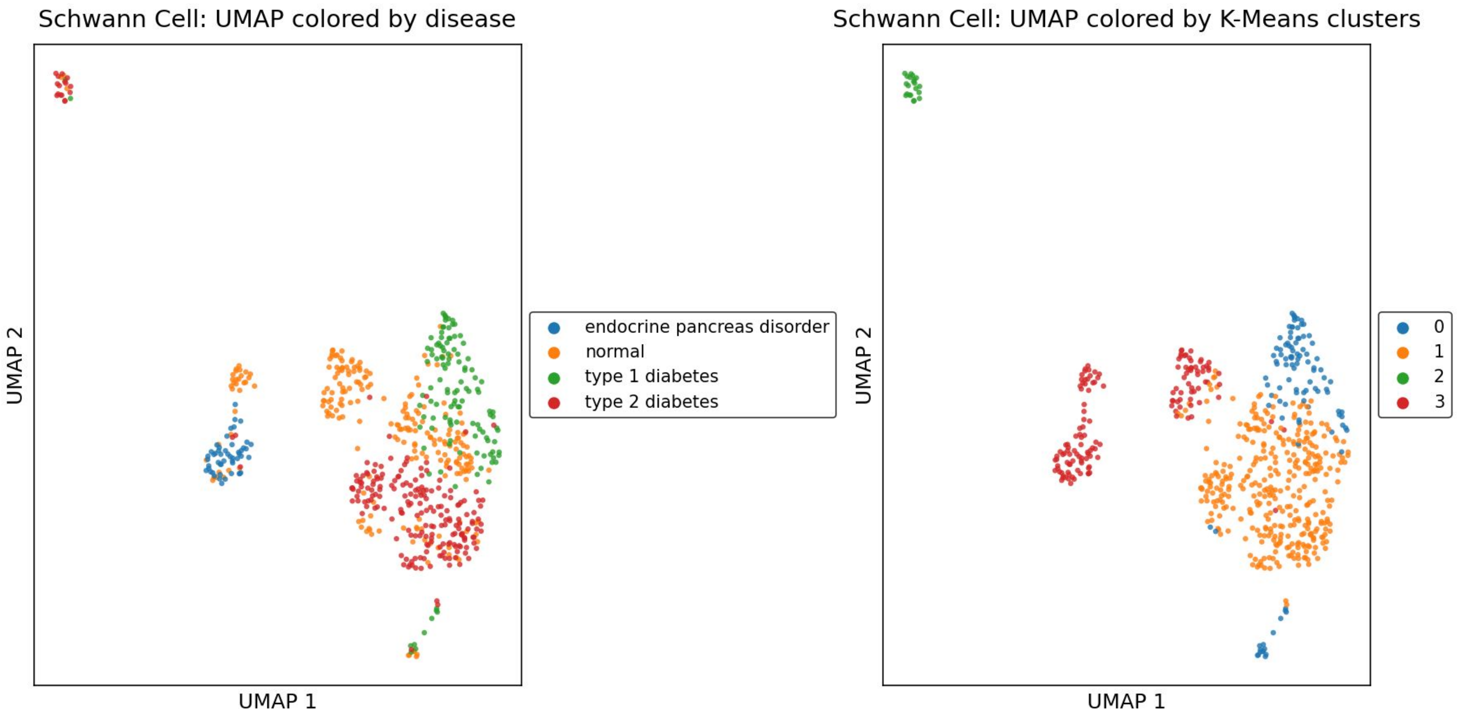


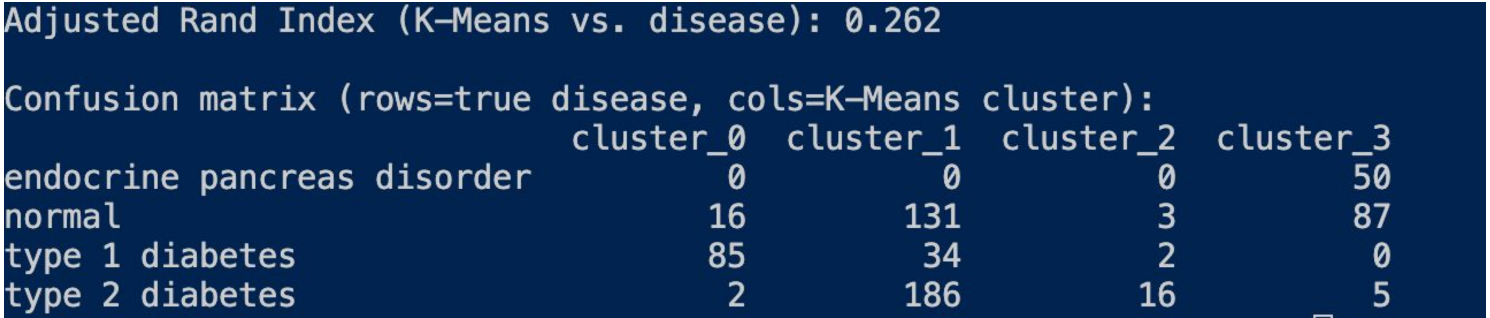
```
Adjusted Rand Index (K-Means vs. disease): 0.303

Confusion matrix (rows=true disease, cols=K-Means cluster):
                             cluster_0  cluster_1  cluster_2  cluster_3
endocrine pancreas disorder         81         16          0       1836
normal                             435        613        596       1617
type 1 diabetes                    135         66        215          1
type 2 diabetes                   2073        163       2986         20
```



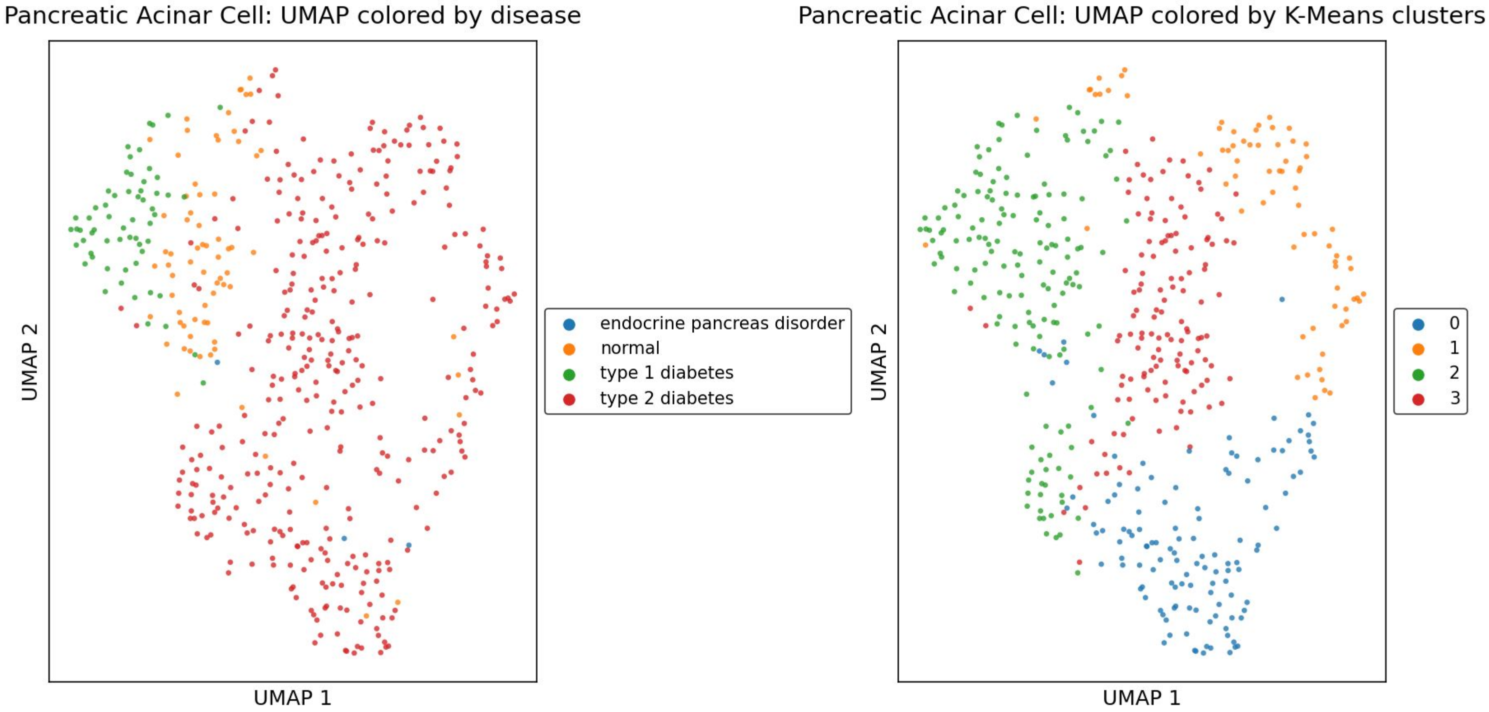


```
Adjusted Rand Index (K-Means vs. disease): 0.382

Confusion matrix (rows=true disease, cols=K-Means cluster):
                             cluster_0  cluster_1  cluster_2  cluster_3
endocrine pancreas disorder          2         27          0          0
normal                            1219        176        485        131
type 1 diabetes                   1971          0        615       1013
type 2 diabetes                    208       2702        150         43
```



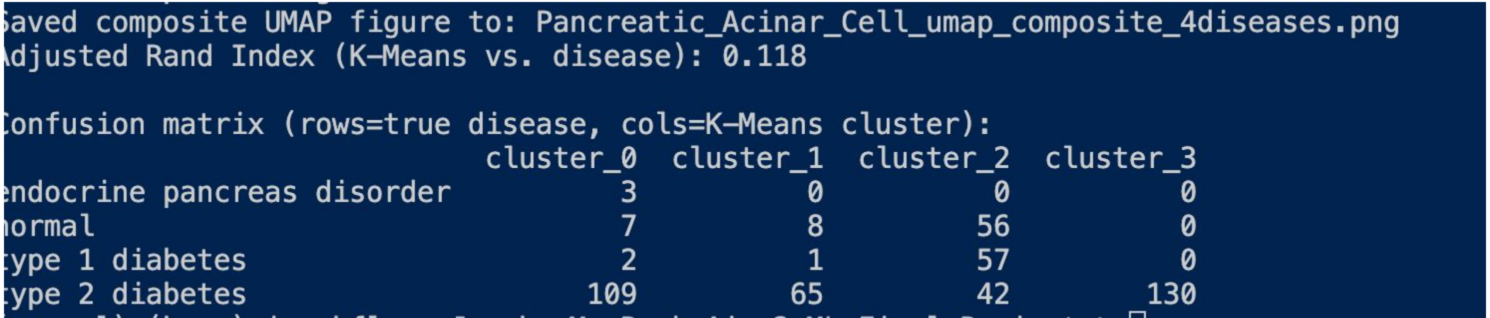


```
Adjusted Rand Index (K-Means vs. disease): 0.262

Confusion matrix (rows=true disease, cols=K-Means cluster):
                             cluster_0  cluster_1  cluster_2  cluster_3
endocrine pancreas disorder          0          0          0         50
normal                              16        131          3         87
type 1 diabetes                     85         34          2          0
type 2 diabetes                      2        186         16          5
```



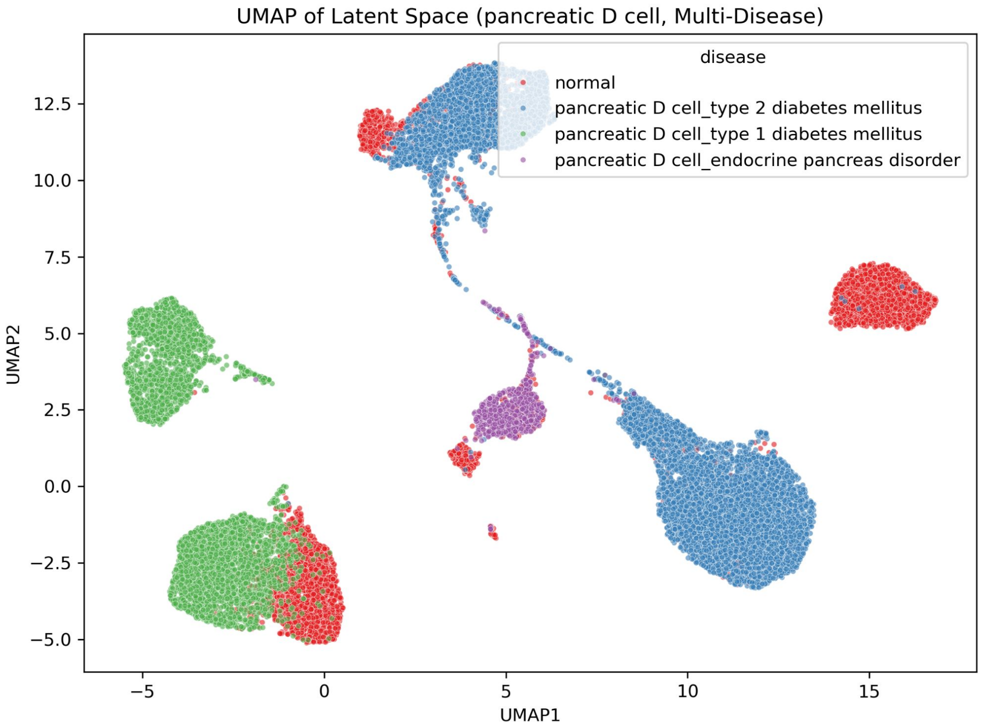
Pancreatic Acinar Cell: UMAP colored by disease

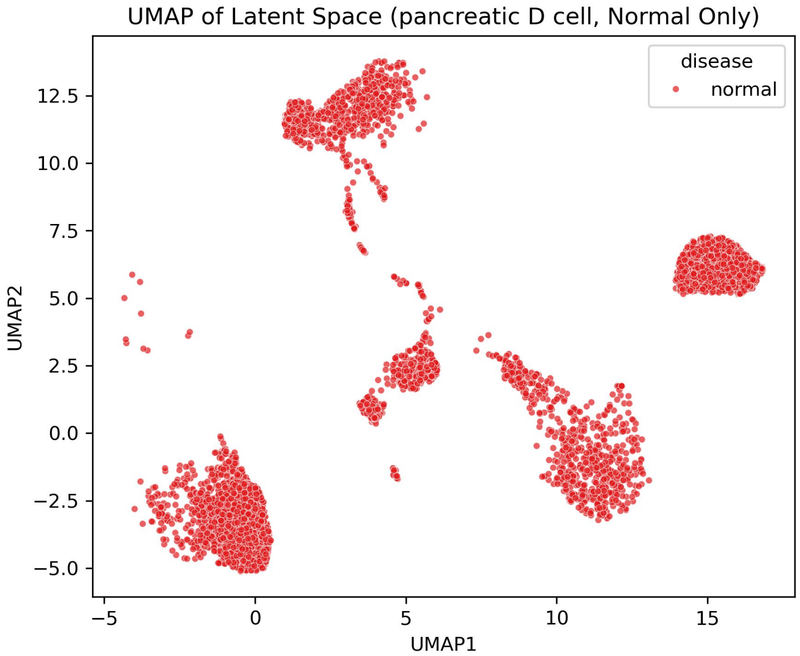
Pancreatic Acinar Cell: UMAP colored by K-Means clusters



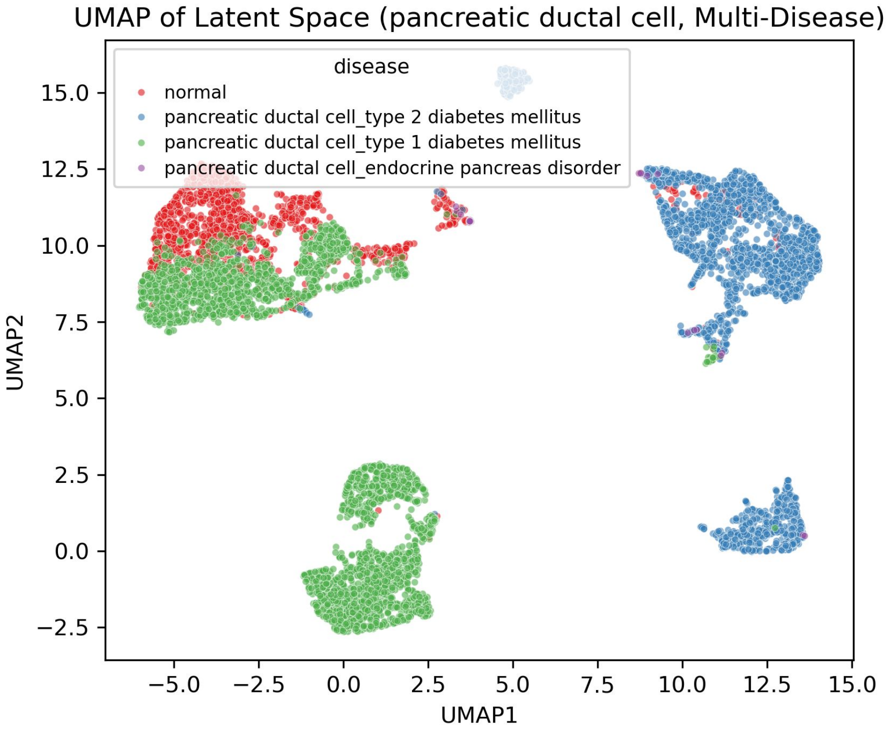
```
Saved composite UMAP figure to: Pancreatic_Acinar_Cell_umap_composite_4diseases.png
Adjusted Rand Index (K-Means vs. disease): 0.118

Confusion matrix (rows=true disease, cols=K-Means cluster):
                            cluster_0  cluster_1  cluster_2  cluster_3
endocrine pancreas disorder          3          0          0          0
normal                               7          8         56          0
type 1 diabetes                      2          1         57          0
type 2 diabetes                    109         65         42        130
```

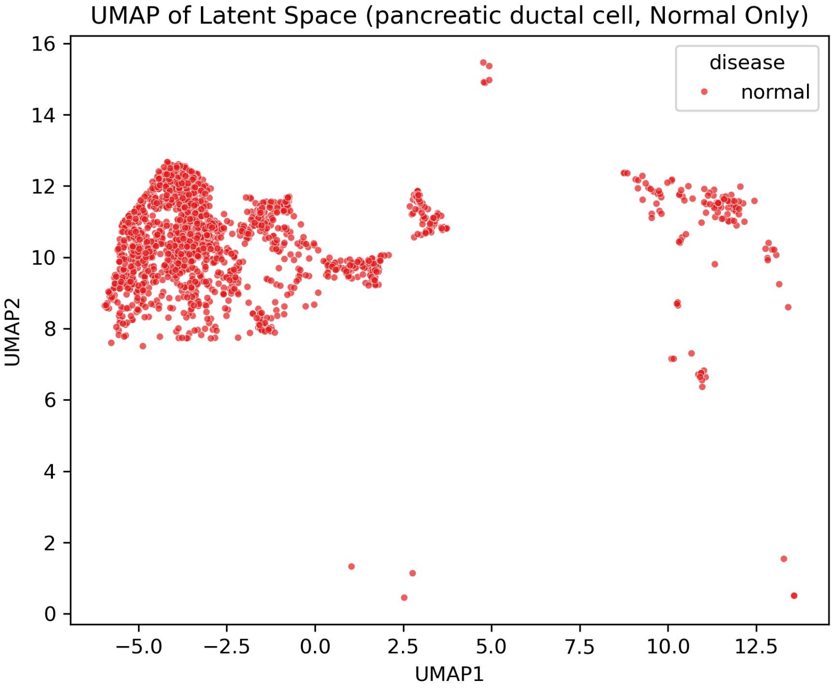
Visualization: Pancreatic D Cell → Some information gained



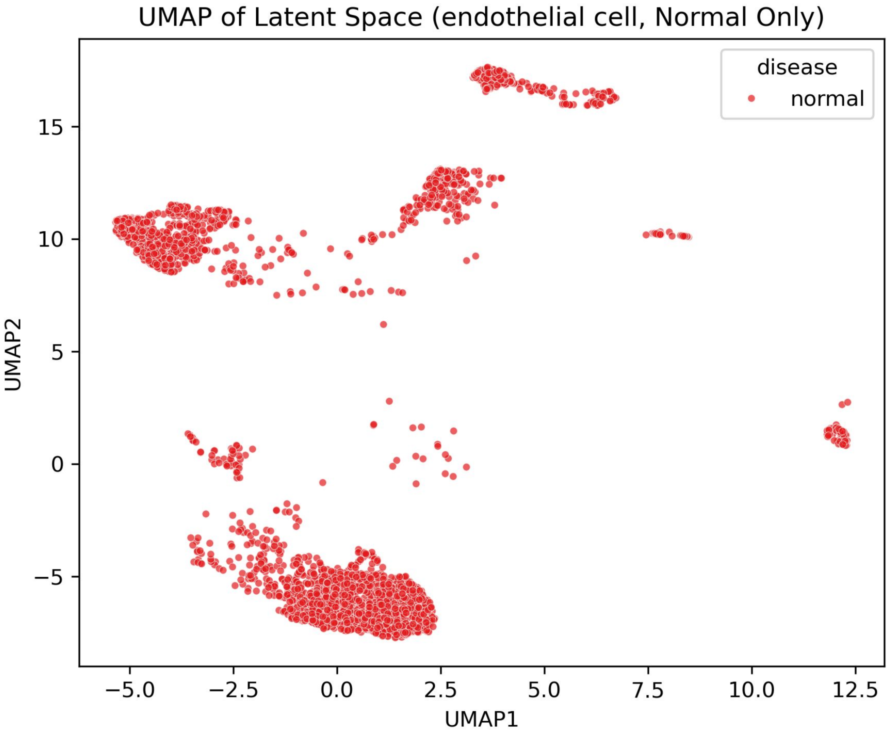


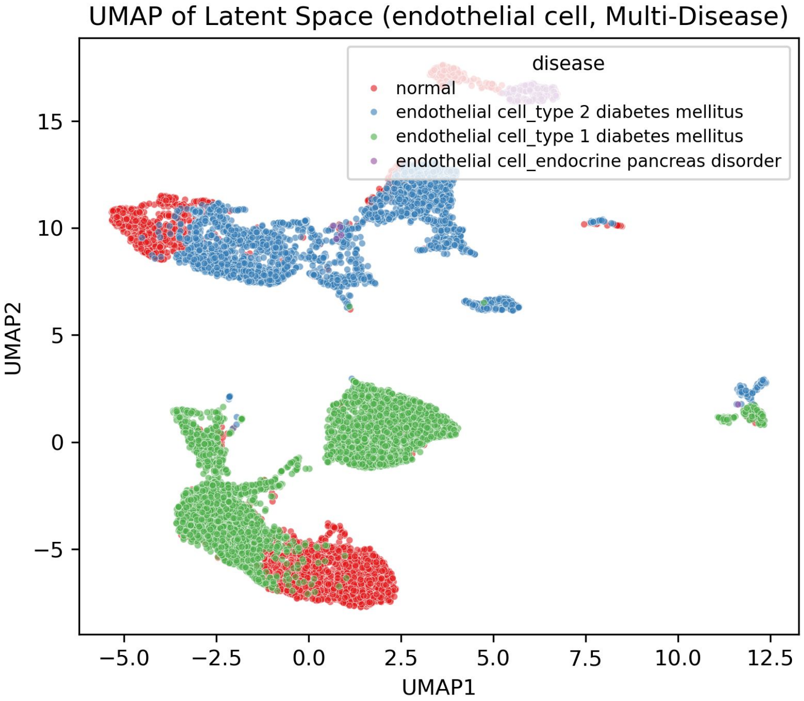
Visualization: Pancreatic Ductal Cell





Visualization: Endothelial cell





No clear distinction