Semi-supervised Nonnegative Matrix Factorization for Type-1 Diabetes Classification from scRNA-seq Data

Elle Marsyla
Harvey Mudd College

Aditi Bonthu
Harvey Mudd College

Tian Xie
Harvey Mudd College

Edward Gao
Pomona College

***Abstract*—Single cell RNA sequencing data takes the read counts for all RNA transcripts throughout an entire single cell. In this way, it represents gene expression at the cell level, since genes that are more highly expressed will have higher numbers of transcripts. Our dataset contains scRNAseq data for PBMCs, or peripheral blood mononuclear cells. It includes cells from individuals with Type-1 Diabetes and healthy donors. We preprocessed this data through Scanpy to identify the subsets of cells present in the data. Then, we selected four cell subtypes to train a SSNMF model on the labeled disease data. We validated whether or not these models could predict disease using K-fold validation. We then analyzed significant genes that contribute to either disease predictions or healthy predictions. We found that monocytes, erythroblasts, and B cells produced successful predictive models. Only T cells were unsuccessful at prediction. We verified these results using the biological basis for diabetes.**

I. Introduction

Our project is looking at single cell RNA sequencing data. Sc-RNA-seq sequences every RNA transcript in an entire single cell. Since RNA is transcribed from genes, this data essentially represents the gene expression across an entire cell. Our dataset collected blood samples from 24 individuals, 12 who are healthy and 12 who have type-1 diabetes. The cell sample type that was extracted are PBMCs. PBMCs are important immune cells that circulate in the bloodstream. The inflammation associated with diabetes activates these immune cells (Gu et. al 2024). We hypothesized that this biological response could be detected in an SSNMF model, and be used for disease prediction and classification. To accurately train this model, we first needed to isolate cell subtypes from our full dataset. To do this, we used Scanpy, a bioinformatics tool that uses PCA to cluster cell types. This workflow first normalized the data, and then found 50 PCs to reduce the dimension of the data. This produced a neighborhood graph of cells using the PCA representation of the data matrix. Then, the Leiden graph-clustering method (community detection based on optimizing modularity) was run to produce a UMAP plot to display the 21 predicted cell clusters (Fig 1).

A. Cell Type Labeling

We then annotated the cell types using the Decoupler package in python. This takes a database of cells and assigns cell types to cell clusters based on marker genes that are highly expressed. This produced a UMAP plot containing the predicted cell types (Fig 2).

Thank you to Prof Haddock for an amazing semester!

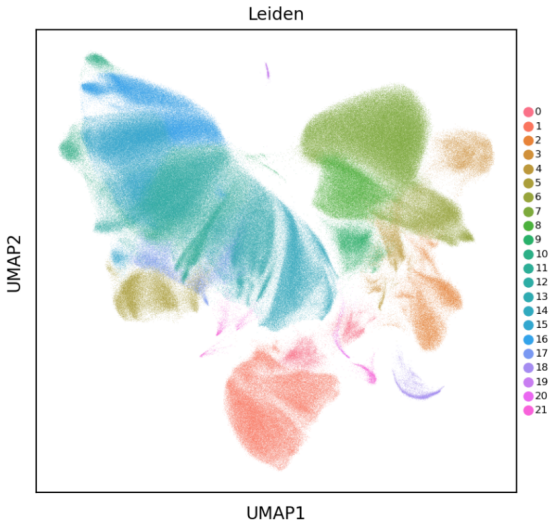

Fig. 1. Original UMAP plot produced by the Leiden graph clustering method shows 21 predicted cell clusters

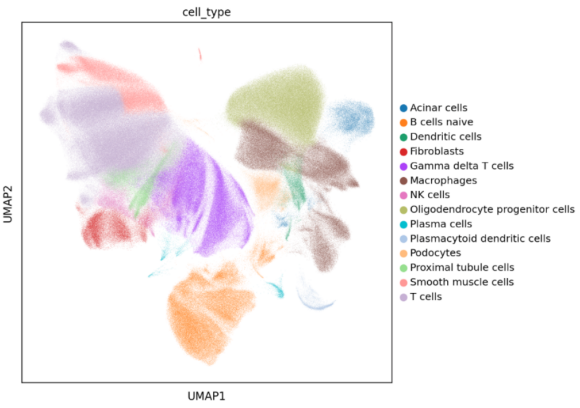

Fig. 2. Cell type labeling from Decoupler package

We then selected the cell subtypes B cells naive, T cells, Gamma delta T cells, and macrophages for developing the SSNMF models as these were the largest clusters with the

most even distribution of healthy and diseased cells. For these four cell types that we isolated, we verified their cellular annotations using comparative marker gene expression analysis. We verified that "T cells" are likely T cells and that "B cells naive" are likely B cells (Fig 3A, B). "Gamma delta T cells" were inconclusive, or are possibly Erythroblasts (Fig 3C). "Macrophages" were verified to be monocytes (Fig 3D).

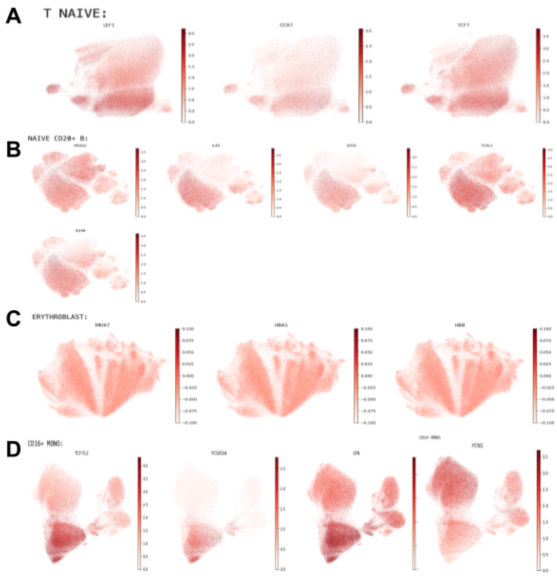

Fig. 3. Verification of cell subtypes. Brighter red coloring means that marker genes are more highly expressed, and a cell is likely of the corresponding type. (A) "T cells" verified as T naive cells. (B) "B cells naive" verified as naive B cells. (C) "Gamma delta t cells" verified as erythroblasts. (D) "Macrophages" verified as monocytes.

B. Biology of Cell Subtypes

1) T-Cells: T-cells are a type of white blood cell called a lymphocyte. They are important immune cells because they attack harmful pathogens and also send signals to the immune system. They come in many different types such as cytotoxic T-cells, helper T-cells, and even suppressor T-cells (Cleveland Clinic, T Cells).

2) B-cells: B-cells are also included in lymphocytes and are important immune cells. They produce antibodies, which are proteins that learn to detect pathogens (Cleveland Clinic, B Cells).

3) Monocytes: Monocytes are another type of white blood cell that is a key part of the immune system. They can differentiate into macrophages, which fight bacteria, fungi, and protozoa.

4) Erythroblasts: Erythroblasts, also known as normoblasts, are precursors to red blood cells (Butina, Michelle 2020).

C. Nonnegative Matrix Factorization

Non-negative matrix factorization is the basis for the SSNMF model that we used in this paper. X is a data matrix $\mathbf{X} \in \mathbb{R}_{\geq 0}^{n_1 \times n_2}$ that is constrained to be nonnegative. A model rank k, dimension $k \in \mathbb{N}$ is chosen to factorize $\mathbf{X}$ into a product of two low-dimensional matrices. These two low-dimensional matrices are constrained to be non-negative as well, hence why this process is called non-negative matrix factorization. This low dimensional representation is in the form of $\mathbf{A}$ and $\mathbf{S}$ so that $\mathbf{X} \approx \mathbf{AS}$. $\mathbf{A} \in \mathbb{R}_{\geq 0}^{n_1 \times r}$ Can be interpreted as the topics matrix. The model rank k can be the number of topics, and each topic has the same number of features as the number of rows in X. $\mathbf{S} \in \mathbb{R}_{\geq 0}^{r \times n_2}$ can be interpreted as the weights matrix. There are the same number of weight columns as there are data points (columns) in the X matrix. There are k weights, each corresponding to one of the k topics. This decomposition is visualized in figure 4. This

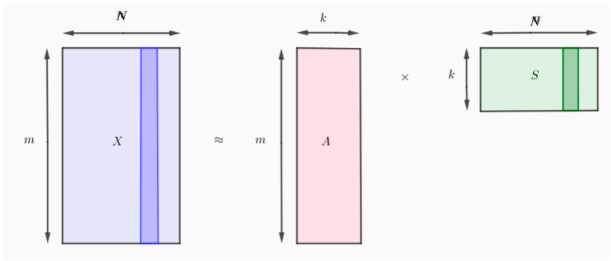

Fig. 4. NMF decomposition

decomposition attempts to satisfy the following objective:

$$\underset{\mathbf{A}\geq 0,\mathbf{S}\geq 0}{\operatorname{argmin}} \|\mathbf{X} - \mathbf{AS}\|_F^2 \tag{1}$$

D. Semi-supervised NMF

SSNMF is the model we use in this paper to perform our classification predictions. It is a form of NMF that also considers a class matrix. According to Prof Haddock, given a data matrix $\mathbf{X} \in \mathbb{R}_{\geq 0}^{n_1 \times n_2}$ and a class label matrix $\mathbf{Y} \in \mathbb{R}_{\geq 0}^{k \times n_2}$, $(\|\cdot\|_F, \|\cdot\|_F)$-SSNMF is defined by

$$\underset{\mathbf{A},\mathbf{S},\mathbf{B}\geq 0}{\operatorname{argmin}} \underbrace{\|\mathbf{W} \odot (\mathbf{X} - \mathbf{AS})\|_F^2}_{\text{Reconstruction Error}} + \lambda \underbrace{\|\mathbf{L} \odot (\mathbf{Y} - \mathbf{BS})\|_F^2}_{\text{Classification Error}}, \tag{2}$$

where $\mathbf{A} \in \mathbb{R}_{\geq 0}^{n_1 \times r}$, $\mathbf{B} \in \mathbb{R}_{\geq 0}^{k \times r}$, $\mathbf{S} \in \mathbb{R}_{\geq 0}^{r \times n_2}$, and the regularization parameter $\lambda > 0$ is the weight of the importance of the supervision term (Haddock et. al).

E. Project Outline

For each cell type, we annotated a one-hot encoded matrix to represent whether the cell donor had diabetes (disease) or was healthy. The first row of the one-hot matrix corresponds to health, while the second row corresponds to diabetes. For each of the four cell subtypes, we ran an SSNMF model fitting with model rank k=2 and 100 iterations. We validated the model through K-fold cross validation with 80% train, 20% test splits and developed confusion plots to measure the rates of true positives, false positives, true negatives, and false negatives. We chose a verification threshold of greater than 70% to validate the models. Once the models were verified, we found the genes that most highly predict disease or healthiness in this model. We ultimately hope to draw biological conclusions

from this data, and inspire future research into the important genes we find.

II. Experimental Data and Results

K-Fold Cross Validation of the SSNMF Model on B Cells

For the SSNMF model run of B-cells, We found a mean accuracy of 74.78% +- 0.32%. We used a 5-fold cross-validation by randomly sampling the dataset 5 times (Figure 5). By doing so, we can make sure that our model works on the whole dataset and is consistent throughout. The standard deviation of accuracy is 0.0032. This is a very low value so consistency is verified. By looking at the confusion matrix, we can see that the dataset has a good true positive and true negative ratio with the false positive and false negative. This could also be shown through a decent mean accuracy of 0.7478. In addition, since the ratio of true negative and false negative (about 7500:2000) is higher than that of the true positive and false negative), we can conclude that for this particular cell (B cell), the model is better at predicting the healthy category than the disease category.

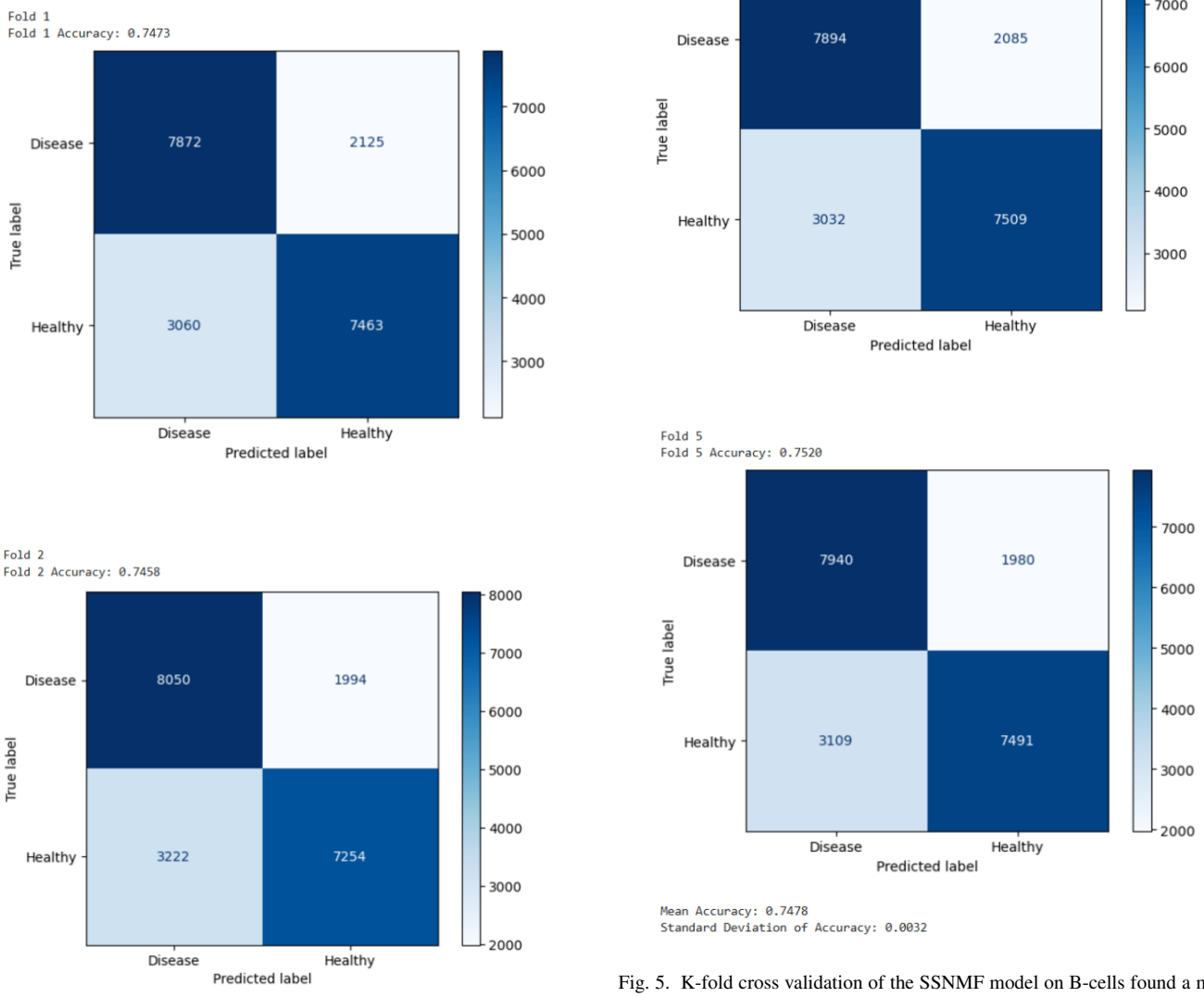

Fig. 5. K-fold cross validation of the SSNMF model on B-cells found a mean accuracy of 74.78%

A. Interpretation of the factored matrices on B-cells

The Y matrix is our learning matrix (Fig 6). It has two rows. The first row represents healthiness, and the second row represents disease. It has the same number of columns as matrix X. Hence, each column represents a cell in matrix X. However, we want to give the algorithm some guidance. This is why we use one-hot encoding of each cell on whether it is considered healthy or malicious. This completes our semi-supervised non-negative matrix factorization. By running a semi-supervised non-negative matrix factorization onto this problem, we are able to find the matrix B, a 2 x 2 matrix that encodes the information in column A. Since we know that the first row of Y represents healthiness and the second column of Y represents diseases, we can interpret matrix B. Given the fact that 7.7310-1 is much bigger than 1.48 10-12, we can conclude that the first column of Y (and thus essentially the first column of A) represents healthiness. Similarly, since 9.10*10-1 is much bigger than 1.78*10-3, the second column represents diseases (Fig 6).

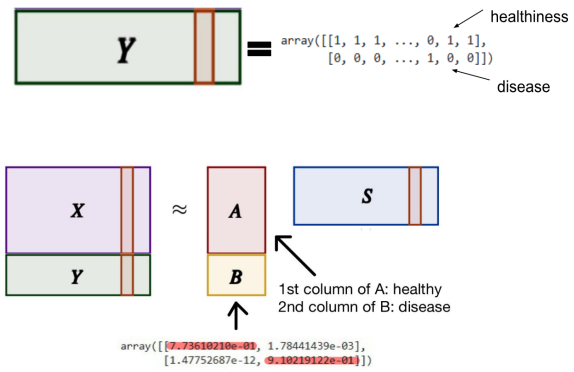

Fig. 6. Interpretation of the SSNMF matrices involved

B. Extraction of important genes (extremely large values)

As shown in figure 6, the mean of matrix A is just about 0.10. In interpreting matrix factorization, a common way is to first find the extremely large values. Typically, a greater numerical value implies a greater importance in the column in which it represents. Here, we found all the genes with a value greater than 3. From this, we can conclude genes such as PRDM11, WDR76, CD83, etc, play an important role in cellular function for both healthy and diseased individuals, since the values are large in both columns (Fig 7).

C. Extraction of important genes (Large and small ratios)

In the first part, we look at extremely large values as a guide. However, a numerical value doesn't always tell the full story. In particular, we hope to understand which gene contributes more to healthiness vs diseases. For example, in figure 7, WDR76 has similar values in both columns, making it hard to determine the individual contribution to healthiness and disease. Thus, we need to look at the overall ratio rather than simple numerical values. For example, TOB2 from figure

```
Value: 4.471208, Row: 96, Gene: PRDM11, Column: 0
Value: 3.286481, Row: 96, Gene: PRDM11, Column: 1
Value: 3.056630, Row: 523, Gene: WDR76, Column: 0
Value: 3.219838, Row: 523, Gene: WDR76, Column: 1
Value: 7.706994, Row: 1073, Gene: CD83, Column: 0
Value: 6.733536, Row: 1073, Gene: CD83, Column: 1
Value: 4.342328, Row: 1182, Gene: MORN1, Column: 0
Value: 3.878683, Row: 1182, Gene: MORN1, Column: 1
Value: 3.463984, Row: 1859, Gene: TLR2, Column: 0
Value: 3.093114, Row: 1859, Gene: TLR2, Column: 1
Value: 3.715929, Row: 1904, Gene: SEMA7A, Column: 0
Value: 10.185510, Row: 2121, Gene: ZC3H8, Column: 0
Value: 8.923961, Row: 2121, Gene: ZC3H8, Column: 1
Value: 7.801981, Row: 2385, Gene: CARHSP1, Column: 0
Value: 6.788574, Row: 2385, Gene: CARHSP1, Column: 1
Value: 3.114657, Row: 2478, Gene: FAM161B, Column: 0
Value: 3.616470, Row: 2794, Gene: CMYA5, Column: 0
Value: 3.817788, Row: 4726, Gene: C8orf37-AS1, Column: 0
Value: 4.615613, Row: 4726, Gene: C8orf37-AS1, Column: 1
mean of A is 0.10546612190929278
figure 6
```

Fig. 7. Interpretation of the SSNMF A matix: Finding large values

7 has a value of 0.042 in the healthiness column and 0.0013 in the disease column. The ratio of healthiness/disease for this particular gene is thus 30.7 (figure 8). While the ratio might not be exact, we can conclude that higher expression of the TOB2 gene likely contributes to a healthy cell response. This is verified by a study that found that TOB2 inhibits inflammatory responses in autoimmune diseases, which type-1 diabetes is (Jiang, Guosheng et al). The fact that our model was able to correctly identify this gene as significant for healthy cells further validates our model. Our model can thus be used to interpret genes as being important for healthy cellular responses, and diseased cellular responses. This could contribute to novel research into these genes to determine their biological pathways.

```
Ratio: 30.731564, Row: 3631, Gene: TOB2
Ratio: 5.524879, Row: 4063, Gene: TEX45
Ratio: 14.715007, Row: 4181, Gene: ZDHHC11B
Ratio: 5.085403, Row: 4471, Gene: LINC00665
Ratio: 17.592066, Row: 4639, Gene: LINC02086
Ratio: 38.463034, Row: 4867, Gene: A1BG-AS1
Ratio: 0.264866, Row: 1866, Gene: FXYD2
Ratio: 0.337574, Row: 1962, Gene: SORD
Ratio: 0.300629, Row: 2579, Gene: CBR3
Ratio: 0.375863, Row: 4018, Gene: SNHG12
Ratio: 0.377521, Row: 4418, Gene: ZRANB2-AS2
Ratio: 0.305917, Row: 4447, Gene: KLF3-AS1
Ratio: 0.237816, Row: 4460, Gene: TRAF3IP2-AS1
```

Fig. 8. Interpretation of the SSNMF A matrix: Finding large ratios. Ratio greater than 5 means a gene is important for healthy classification. Ratio less than 0.4 means a gene is important for disease classification.

D. K-Fold Cross Validation of the SSNMF Model on T cells

Across the 5 K-fold validations of B-cells, we found an average accuracy of 67% +- 0.19% (Fig. 9). While this is a low value, it could be due to the imbalanced representations

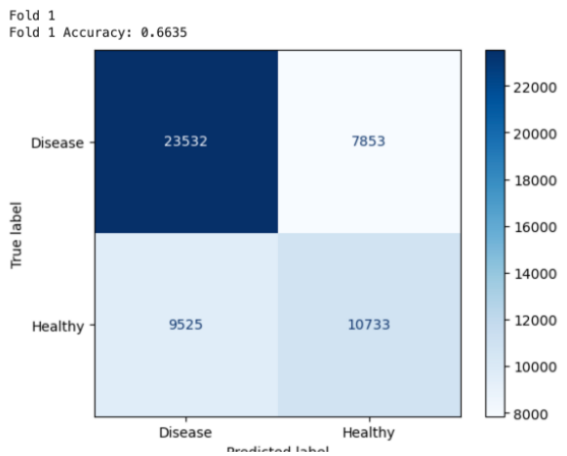

Fig. 9. One representative confusion matrix for the 5 fold test on T-cells. This shows a very low prediction accuracy with a greater number of diseased individuals represented

of healthy and diseased data points. We observe that there are approximately 1.5 times as many diseased patients as healthy patients in this dataset.

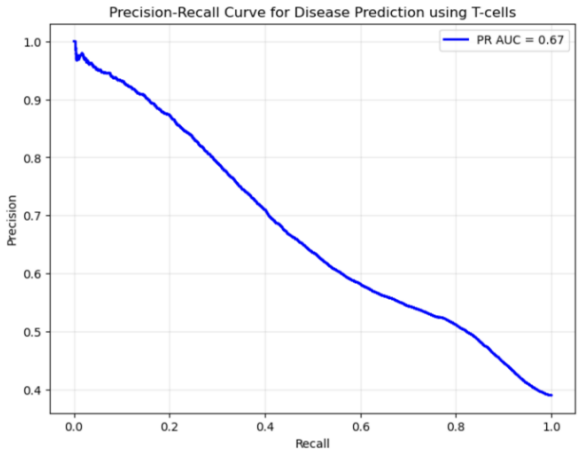

Fig. 10. Precision recall curve for the k-fold validation on T-cells gives an area of 0.67

In the case where a cell subtype has an unequal distribution of healthy data points and diseased data points, this may artificially increase or decrease the predictive ability of the model. To address these imbalances, we constructed precision recall curves. Precision is the percentage of correctly identified positives, or $\frac{T_p}{T_p+F_p}$ and recall is the percentage of true positives over the total number of positives in the dataset, or $\frac{T_p}{T_p+F_n}$. An ideal precision recall curve will have a 90 degree angle in its shape, and have an area of 1 underneath its curve. This method can be used to verify these models on top of the k-fold validation. Because of the imbalance in the T-cell dataset, we used a precision recall curve to determine the quality of prediction (Figure 10). However, we found that the precision recall curve had an area of 0.67, meaning this prediction is almost no better than random. Therefore, we were not able to verify the model developed on T-cells (Figure 10).

We also found significant genes with high ratios in this model (Fig 11). Even though this model did not have very high accuracy, finding these significant genes could be useful. This shows a flaw in the method of using ratios to detect important genes. In This scenario, gene PUS7 has an extremely large ratio. However, this is only because its denominator is 3.35E-12. This is most likely because that gene has near zero representation in T cells dataset, making it impossible for the model to pick up the signal. This emphasizes the importance of proceeding with caution when using ratios to determine important genes, as these genes may not actually be the most important. They could artificially show higher or lower ratios.

```
Ratio: 49.933062, Row: 507, Gene: PUS7
Ratio: 6.360083, Row: 1639, Gene: ENOSF1
Ratio: 5.762199, Row: 1793, Gene: GPR55
Ratio: 5.420217, Row: 2379, Gene: PRDM8
Ratio: 5.723048, Row: 2495, Gene: CD109
Ratio: 5.386202, Row: 2573, Gene: MPZ
Ratio: 32.417222, Row: 2663, Gene: JOSD2
Ratio: 25.913596, Row: 2746, Gene: TRAT1
Ratio: 8.220736, Row: 3231, Gene: HPSE
Ratio: 10.292866, Row: 3407, Gene: CRACR2B
Ratio: 14.531173, Row: 4181, Gene: ZDHHC11B
Ratio: 6.134046, Row: 4317, Gene: HLA-DPB1
Ratio: 8.188385, Row: 4338, Gene: SLC9A3-AS1
Ratio: 5.516919, Row: 4362, Gene: LINC02848
Ratio: 15325.195886, Row: 4639, Gene: LINC02086
Ratio: 125850186014.793091, Row: 4867, Gene: A1BG-AS1
Ratio: 5.434632, Row: 4928, Gene: LINC00540
Ratio: 5.410132, Row: 4962, Gene: C13orf46
Ratio: 0.215315, Row: 320, Gene: XRCC1
Ratio: 0.323472, Row: 1256, Gene: ARMC2
Ratio: 0.365491, Row: 3500, Gene: CXCR2
Ratio: 0.207961, Row: 3694, Gene: CCDC137
Ratio: 0.398054, Row: 4413, Gene: ITGB1-DT
Ratio: 0.194027, Row: 4447, Gene: KLF3-AS1
Ratio: 0.254506, Row: 4627, Gene: HLA-L
Ratio: 0.304135, Row: 4690, Gene: C5orf17
```

Fig. 11. Significant genes in T-cell prediction. Ratio greater than 5 means a gene is important for healthy classification. Ratio less than 0.4 means a gene is important for disease classification.

E. K-Fold Cross Validation of the SSNMF Model on Erythroblasts

In the case of erythroblasts, we found a mean accuracy of 77.15% with a standard deviation of 0.27% (Figure 12). There was little variation, validating the model's consistency. The mean accuracy is also relatively high, so we concluded that this was a successful prediction model. Upon examining the confusion matrix, it was evident that the dataset exhibited a favorable ratio of true positives and true negatives in relation to false positives and false negatives. Notably, for the erythroblasts cell, the model demonstrated greater proficiency in predicting the disease category as compared to the health category.

Since there is a slight imbalance in the proportion of healthy and diseased individuals, we included a precision recall curve that contains an area of 0.85 (Figure 13). Since an ideal area is

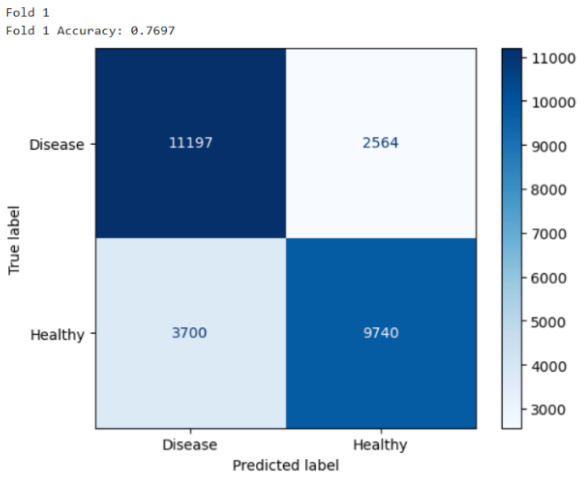

Fig. 12. One representative confusion matrix for the 5 fold test on Erythroblasts. This shows a prediction accuracy of 77% and a higher representation of diseased individuals

1, this shows that this model is somewhat successful at disease prediction.

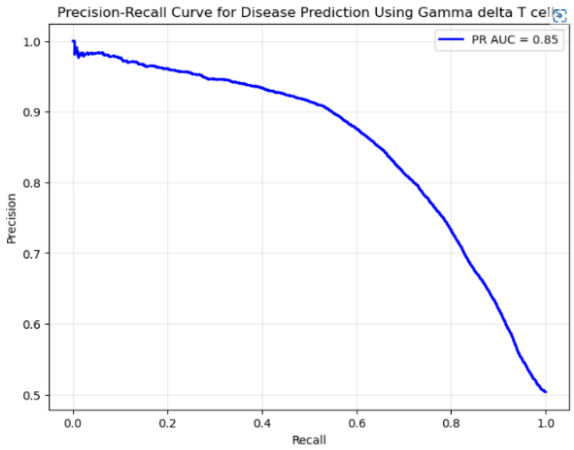

Fig. 13. Precision recall curve for the k-fold validation on Erythroblasts gives an area of 0.85

We also found significant genes using the ratio test (Figure 14).

F. K-Fold Cross Validation of the SSNMF Model on Monocytes

In this case, we ran 5-fold cross validation on Macrophages, randomly sampling each time. The mean accuracy was 83.74% with a standard deviation of 0.27%, which means we have consistent results (Figure 15).

Our random sampling values demonstrate that the Macrophage dataset had almost three times as many healthy patients as diseased. To address this imbalance, we plotted the precision-recall curve. This found that monocytes are almost a perfect model for disease prediction, with an area of 0.97

```
Ratio: 12.931141, Row: 132, Gene: TIMP2, Num
Ratio: 5.517480, Row: 1312, Gene: HSPH1, Num
Ratio: 7.606133, Row: 4362, Gene: LINC02848,
Ratio: 16.993566, Row: 4867, Gene: A1BG-AS1,
Ratio: 0.214395, Row: 320, Gene: XRCC1
Ratio: 0.330532, Row: 392, Gene: EPB41L2
Ratio: 0.284093, Row: 575, Gene: BCL2L13
Ratio: 0.394539, Row: 856, Gene: ARMC6
Ratio: 0.274501, Row: 1395, Gene: ATP8A1
Ratio: 0.389961, Row: 2465, Gene: KDM8
Ratio: 0.192270, Row: 2579, Gene: CBR3
Ratio: 0.207475, Row: 3694, Gene: CCDC137
Ratio: 0.385930, Row: 4063, Gene: TEX45
Ratio: 0.243136, Row: 4191, Gene: IGLC2
Ratio: 0.187010, Row: 4193, Gene: TRGC1
Ratio: 0.196214, Row: 4195, Gene: TRBC1
Ratio: 0.209472, Row: 4447, Gene: KLF3-AS1
Ratio: 0.213650, Row: 4690, Gene: C5orf17
Ratio: 0.374601, Row: 4710, Gene: GMDS-DT
Ratio: 0.377488, Row: 4711, Gene: LINC02506
```

Fig. 14. Significant genes in Erythroblast prediction. Ratio greater than 5 means a gene is important for healthy classification. Ratio less than 0.4 means a gene is important for disease classification.

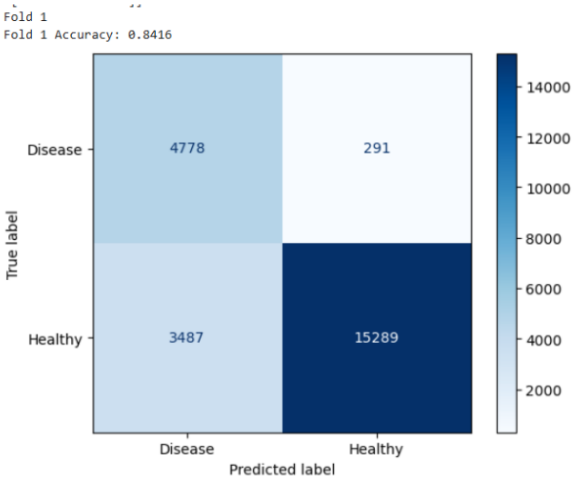

Fig. 15. One representative confusion matrix for the 5 fold test on Monocytes. This shows a prediction accuracy of 84.16% and a higher representation of healthy individuals

underneath the curve (Figure 16). This makes monocytes the strongest model that we developed in this study. We also found

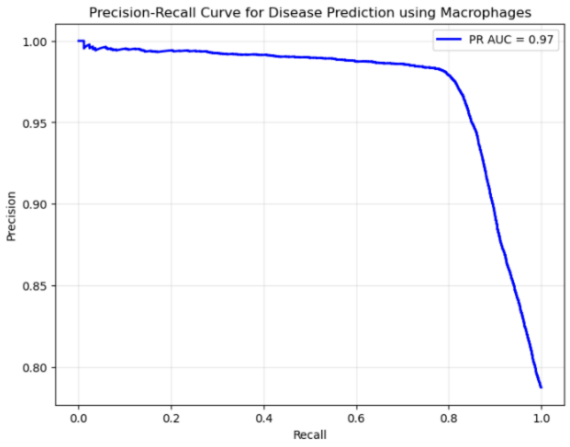

Fig. 16. Precision recall curve for the k-fold validation on Monocytes gives an area of 0.97

significant genes using the ratio test (Figure 17).

```
Ratio: 175281135082913.000000, Row: 4018, Gene: SNHG12
Ratio: 5.477574, Row: 4674, Gene: LINC02150
Ratio: 12.741598, Row: 4704, Gene: LINC02762
Ratio: 0.105044, Row: 5, Gene: BAD
Ratio: 0.154054, Row: 6, Gene: CD99
Ratio: 0.167851, Row: 7, Gene: KLHL13
Ratio: 0.382102, Row: 11, Gene: PLXND1
Ratio: 0.225448, Row: 18, Gene: MEOX1
Ratio: 0.385927, Row: 23, Gene: CROT
Ratio: 0.323809, Row: 28, Gene: ITGA2B
Ratio: 0.071989, Row: 29, Gene: OSBPL7
Ratio: 0.114431, Row: 33, Gene: IFRD1
Ratio: 0.312613, Row: 53, Gene: CELSR3
Ratio: 0.116761, Row: 55, Gene: MPND
Ratio: 0.104611, Row: 58, Gene: NFIX
```

Fig. 17. Significant genes in Monocyte prediction. Ratio greater than 5 means a gene is important for healthy classification. Ratio less than 0.4 means a gene is important for disease classification.

III. Conclusion

In conclusion, we determined data clusters using the Leiden graph clustering method after reducing the dimension of the data using PCA. We then used Decoupler to label the cell subtypes using a database of marker genes, However, this cell type labeling may have initially been slightly inaccurate, so the cell type labeling was confirmed using manual marker gene comparison. We selected the cell types T cells, B cells, Erythroblasts, and Macrophages to train our models. We trained the model using semi-supervised nonnegative matrix factorization with a model rank of two. Our class label matrix was a one-hot encoded matrix with the first row corresponding to healthiness and the second row corresponding to disease. We ran k-fold validation on each of the four models, and if there was unequal representation in the healthy and diseased groups, we included a precision recall curve to further verify the model. Ultimately, we found that Monocytes were the best model, with erythroblasts and B cells being slightly worse, but still verified as successful. T cells were not a successful basis for Type-1 Diabetes disease prediction. We can conclude that at least in this dataset, monocytes, erythroblasts, and B cells can be used to predict disease outcome solely based on the single cell RNA sequencing for that cell type.

We attempted to find a biological basis to the ability of monocytes and B cells to predict type-1 diabetes, and why T cells were less successful. We found that diabetes is associated with increased monocyte inflammatory status (Kanter et al, 2020). This could explain why monocytes were a successful predictor of disease in this case. We also found that Erythroblasts are early stages of red blood cells, which are impacted by diabetes (Obeagu, Emmanuel 2024). Even though erythroblasts are an early form of red blood cells, gene expression at this stage could be determining whether it differentiates into a healthy or diseased form. This could explain why early stages of this cell are successful predictors of disease. However, we also found that T cells play a role in Type 1 diabetes. In this case, there is a T-cell mediated immune response, where T-cells destroy pancreatic cells, leading to disease progression (Roep, Bart 2003). Due to this information, it seems unclear as to why T-cells were not a successful model in this experiment. This could be due to the limitations of our knowledge of cell subtype labeling. The T-cells involved in this immune response are Autoreactive T-cells. The T-cells we looked at may have been a different type of T-cells.

Limitations of our work include the fact that our cell labeling step may need to be advised by a biologist to ensure accuracy. Additionally, while this was a relatively large dataset with many different data points, it also comes from a very small donor set. There were only 12 healthy individuals and 12 individuals with Type-1 Diabetes in this study. This could have introduced bias in the model due to biological variation on the individual level rather than the disease response level.

In the future, it would be important to consult a biologist with experience in scRNAseq data to assist in the cell labeling step. This could help narrow down the cells of interest, and provide clearer biological relevance to Type-1 Diabetes. Additionally, I would like to perform SSNMF with greater ranks K to try to improve the model accuracy. Since our project was focused on the interpretability of these matrices, adding greater K ranks may hurt interpretability since there would no longer be a streotypically “healthy” or “diseased” cell. However, a diagnostic model would need to be more accurate to be useful. Another future experiment could combine many different scRNAseq datasets, selecting one representative cell from each individual. This would ensure less bias due to a smaller donor set such as the one in this study. We also hope that our characterization of important genes could inspire research into their pathways.

Our work in this project hopefully serves as a proof of concept study that it is possible to apply machine learning techniques to scRNAseq data and create predictive models. While these models were not perfect, the interpretability of

the outcome of our model can show which genes may be important for healthy and diseased cellular function. This study has shown that future experiments using this type of data could result in stronger models that could be used for disease diagnostics.

ACKNOWLEDGEMENTS

Thank you to Prof Haddock for an amazing semester! Thank you to the Math177 class for creating a positive learning environment.

REFERENCES

[1] Gu, Doeon, et al. "Single-Cell Analysis of Human PBMCs in Healthy and Type 2 Diabetes Populations: Dysregulated Immune Networks in Type 2 Diabetes Unveiled through Single-Cell Profiling." Frontiers in Endocrinology, vol. 15, 12 July 2024, https://doi.org/10.3389/fendo.2024.1397661.

[2] Cleveland Clinic. "T Cells: Types and Function." Cleveland Clinic, Cleveland Clinic, 17 Jan. 2023, my.clevelandclinic.org/health/body/24630-t-cells.

[3] Cleveland Clinic. "B Cells." Cleveland Clinic, Cleveland Clinic, 1 Feb. 2023, my.clevelandclinic.org/health/body/24669-b-cells.

[4] "Normoblast - an Overview — ScienceDirect Topics." Www.sciencedirect.com, www.sciencedirect.com/topics/medicine-and-dentistry/normoblast.

[5] Jiang, Guosheng, et al. "Tob2 Inhibits TLR-Induced Inflammatory Responses by Association with TRAF6 and MyD88." The Journal of Immunology, vol. 205, no. 4, 15 Aug. 2020, pp. 981–986, www.jimmunol.org/content/205/4/981.abstract, https://doi.org/10.4049/jimmunol.2000057. Accessed 21 Oct. 2022.

[6] Kanter, Jenny E., et al. "Monocytes and Macrophages as Protagonists in Vascular Complications of Diabetes." Frontiers in Cardiovascular Medicine, vol. 7, 14 Feb. 2020, p. 10, www.ncbi.nlm.nih.gov/pmc/articles/PMC7033616/, https://doi.org/10.3389/fcvm.2020.00010. Accessed 7 Dec. 2022.

[7] Emmanuel Ifeanyi Obeagu. "Red Blood Cells as Biomarkers and Mediators in Complications of Diabetes Mellitus: A Review." Medicine, vol. 103, no. 8, 23 Feb. 2024, pp. e37265–e37265, journals.lww.com/md-journal/fulltext/2024/02230/red blood cells as biomarkers and mediators in.44.aspx, https://doi.org/10.1097/md.0000000000037265.

[8] Roep, Bart O. "The Role of T-Cells in the Pathogenesis of Type 1 Diabetes: From Cause to Cure." Diabetologia, vol. 46, no. 3, Mar. 2003, pp. 305–321, link.springer.com/article/10.1007/s00125-003-1089-5, https://doi.org/10.1007/s00125-003-1089-5.

[9] Haddock, Jamie et al. "Semi-supervised Nonnegative Matrix Factorization for Document Classification" 2021 55th Asilomar Conference on Signals, Systems, and Computers, Oct. 2021, DOI:10.1109/IEEECONF53345.2021.9723109