

Single-Cell Analysis of Pulmonary Fibrosis

Section 1: Introduction and Data Overview

Objective

The primary objective of this project is to analyse a public single-cell RNA sequencing (scRNA-seq) dataset of lung tissue samples (GSE122960) to identify cellular signatures associated with pulmonary fibrosis. The ultimate goal is to build a predictive model based on these signatures that can accurately distinguish healthy donor samples from those with fibrotic lung disease.

Data Loading and Initial Assessment

The analysis pipeline began by loading 17 individual 10x Genomics .h5 files, representing 17 distinct patient and donor samples. These individual datasets were concatenated into a single, unified `AnnData` object for comprehensive analysis.

As shown in the initial loading log, the combined raw dataset (prior to any quality control) is substantial:

- **Total Cells:** 80,919
- **Total Genes:** 33,694

During the loading process, each cell was annotated with metadata derived from its file of origin. This annotation is critical for the project's objective, as it provides the "ground truth" for each sample. The key metadata fields are:

- `disease_status`: The specific diagnosis (e.g., "Donor", "IPF", "SSc-ILD").
- `disease_binary`: A simplified high-level category ("Donor" vs. "Fibrosis"), which serves as the primary target variable for our predictive model.

A sample of the resulting metadata table is shown below, illustrating the successful annotation of cells from both Donor and Fibrosis (IPF, SSc-ILD) patients.

Full dataset shape (before QC) : 80919 cells × 33694 genes
Example metadata (random 5 cells) :

		batch	sample_id	disease_status	disease_binary
CACCTTGGTCTTCGTC-1	GSM3489197_Donor_08	GSM3489197_Donor_08		Donor	Donor
ACCCACTCAAGTCTGT-1	GSM3489191_Donor_05	GSM3489191_Donor_05		Donor	Donor
CACCACTCCCCTCAGT-1	GSM3489188_IPF_03	GSM3489188_IPF_03		IPF	Fibrosis
ACACCGGGTAGCGTCC-1-1	GSM3489198_SSc-ILD_02	GSM3489198_SSc-ILD_02		SSc-ILD	Fibrosis
TTCTACACATGGAATA-1	GSM3489197_Donor_08	GSM3489197_Donor_08		Donor	Donor

This raw dataset of ~80,000 cells serves as the starting point for the next critical phase: quality control.

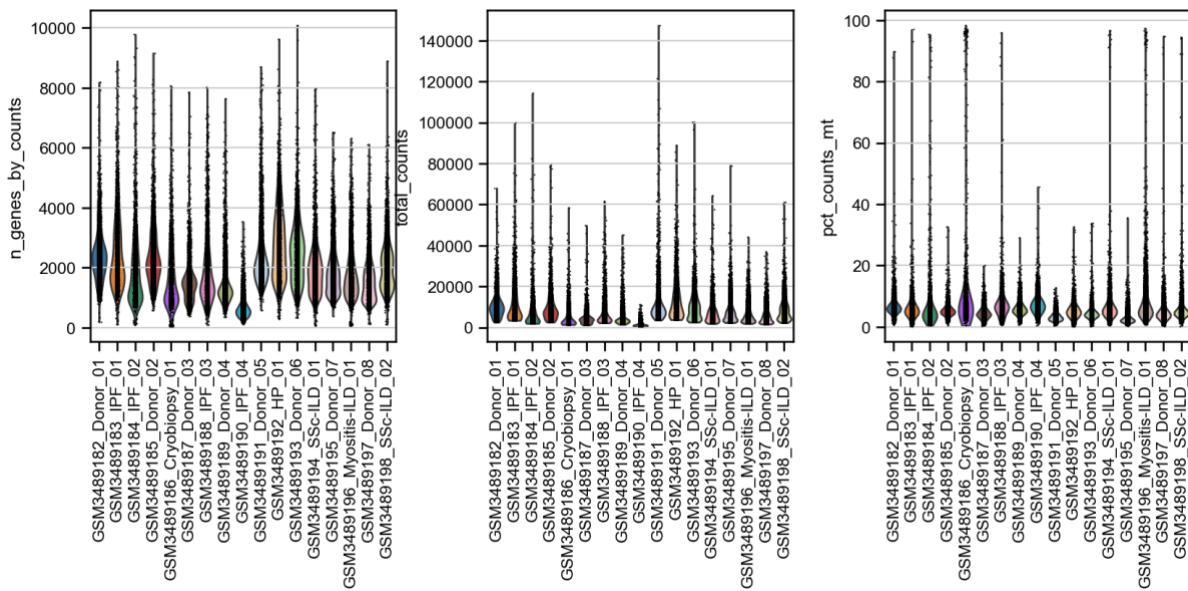
Section 2: Data Quality Control and Filtering

Initial QC Assessment

Before analysis, it is essential to perform quality control (QC) to remove low-quality cells (e.g., empty droplets, cell doublets) and stressed or dying cells. We assessed three standard QC metrics:

1. **n_genes_by_counts**: The number of unique genes detected in a cell. Very low numbers can indicate an empty droplet.
2. **total_counts**: The total number of UMI counts in a cell.
3. **pct_counts_mt**: The percentage of counts mapping to mitochondrial genes. A high percentage is a common indicator of cellular stress or damage.

The initial distributions of these metrics across all 17 samples are visualized in the violin plot below.



As seen in the plot, there is significant variability. Many samples contain a long "tail" of low-quality cells with very few genes or counts. Furthermore, some samples show a population of cells with high mitochondrial content (e.g., >15%). These outliers must be removed to ensure the downstream analysis is based on healthy, viable cells.

The Filtering Process

Based on these distributions, the following filters were applied to the data:

1. **min_genes=200**: Cells with fewer than 200 detected genes were removed.
2. **min_cells=3**: Genes detected in fewer than 3 cells across the entire dataset were removed, as they are uninformative.
3. **pct_counts_mt < 15**: Cells with 15% or more of their counts mapping to mitochondrial genes were removed.

This filtering process refined the dataset significantly:

- **Initial Shape:** 80,919 cells × 33,694 genes
- **Final Shape:** 79,467 cells × 25,705 genes

Final Dataset Statistics

The statistical summary of the final, post-QC dataset confirms the success of our filtering. The table below shows the new distributions for the cells that passed QC.

	TOTAL_COUNTS	N_GENES_BY_COUNTS	PCT_COUNTS_MT
COUNT	79467.00	79467.00	79467.00
MEAN	7210.29	1889.93	5.04
STD	6338.01	970.54	2.42
MIN	423.00	202.00	0.06
25%	3197.00	1207.00	3.33
50%	5425.00	1715.00	4.68
75%	9237.00	2368.00	6.29
MAX	147423.00	10081.00	14.99

Critically, the `min` gene count is now 202 and the `max` mitochondrial percentage is 14.99, confirming our filters were applied correctly. The remaining population of 79,467 cells, with an average of ~1,890 genes detected per cell, represents a high-quality dataset ready for normalization and feature selection.

Section 3: Normalization, Feature Selection, and Preprocessing

With a high-quality set of 79,467 cells, the next step is to process the gene expression data. This involves several key steps to make the data comparable across cells and to reduce its complexity.

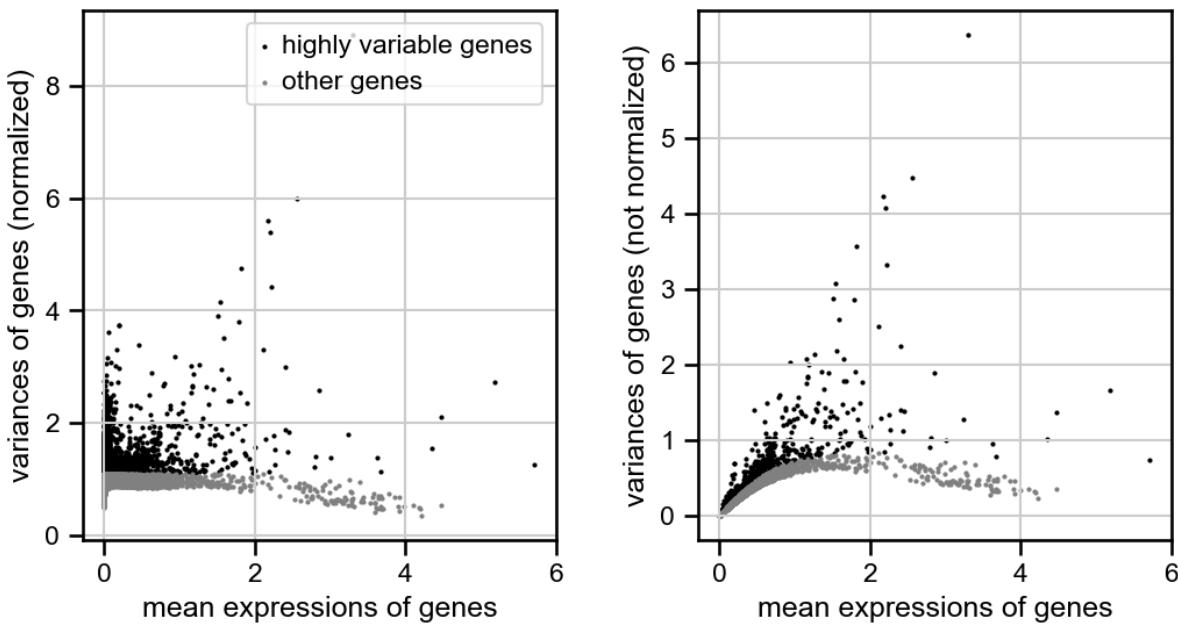
Normalization and Log-Transformation

First, the data was normalized to account for differences in sequencing depth (i.e., some cells have more `total_counts` than others). This was done by scaling the counts in each cell to a target sum of 10,000, followed by a log-transformation (`log1p`). This process prevents highly-expressed genes in cells with deep sequencing from skewing the analysis.

Feature Selection: Highly Variable Genes (HVGs)

A scRNA-seq dataset can contain over 25,000 genes, but most of these are not biologically informative for distinguishing cell types. To focus the analysis on meaningful genes, we performed feature selection by identifying the **Top 5,000 Highly Variable Genes (HVGs)**. These are genes that show high variance in expression relative to their average expression, suggesting they are involved in important biological processes.

The plot below visualizes this selection. The 5,000 selected HVGs are marked in black, clearly standing out from the non-variable background genes marked in grey.



Subsetting, Scaling, and Dimensionality Reduction

Following this selection, the dataset was filtered to *only* these 5,000 HVGs, dramatically reducing the complexity of the data:

- **Data shape after HVG subsetting: 79,467 cells × 5,000 genes**

Finally, this 5,000-gene matrix was fully preprocessed:

1. **Confounder Regression:** Technical artifacts associated with `total_counts` and `pct_counts_mt` were regressed out.
2. **Scaling:** The expression of each gene was scaled to have a unit variance and zero mean (a Z-score). This step is essential for Principal Component Analysis (PCA).
3. **PCA:** The 5,000-dimensional gene matrix was reduced to its top 50 principal components. An "elbow plot" (visualized in `pca_elbow_plot.png`) confirmed that the majority of the biological variance was captured in the first **30 Principal Components (PCs)**.

This final, preprocessed dataset, represented by 30 PCs for 79,467 cells, is now ready for cell clustering and visualization.

Section 4: Clustering and Cell Type Annotation

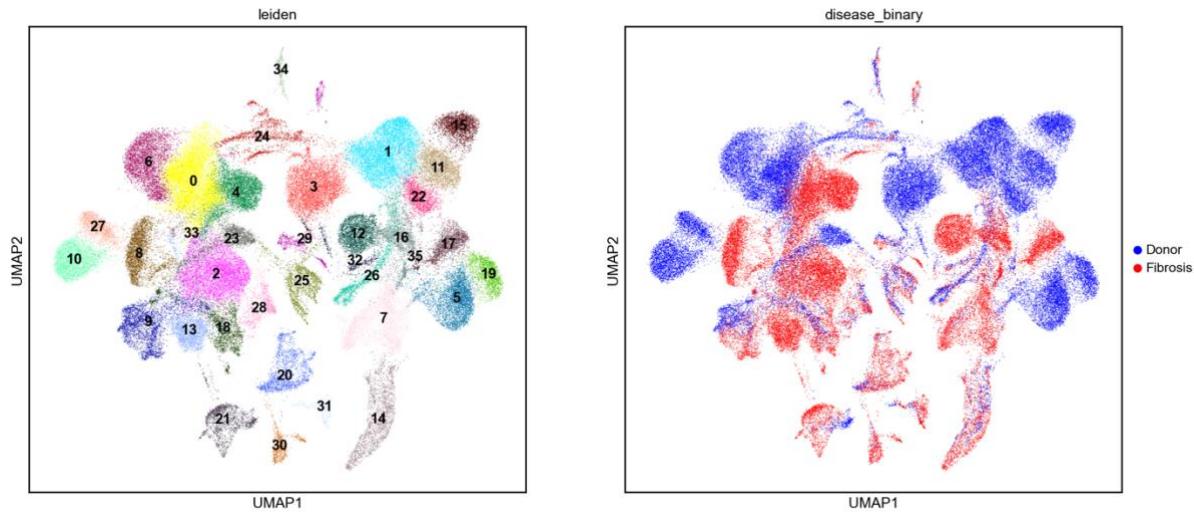
Unsupervised Clustering

With the 30-PC processed dataset, we can now identify groups of transcriptionally similar cells. This was achieved by:

1. Building a **neighbourhood graph**, which identifies the "nearest neighbours" for each cell in 30-dimensional PC space.

2. Running the **Leiden algorithm** (`resolution=1.0`) on this graph to partition the cells into clusters.

This unsupervised approach identified **36 distinct cell clusters**. To visualize these clusters and their relationship to the disease, a UMAP (Uniform Manifold Approximation and Projection) embedding was generated.



The UMAP visualization above provides two key insights:

- **Left Panel (Leiden Clusters):** Shows the 36 distinct clusters found by the algorithm. Cells are grouped based on transcriptional similarity, forming clear islands of related cell types.
- **Right Panel (Disease State):** This plot colours the same cells by their sample origin: "Donor" (blue) or "Fibrosis" (red). We can immediately observe that while many clusters are mixed, the cell populations are not perfectly overlapping. There are distinct regions of the UMAP that are heavily skewed towards either Donor or Fibrosis, suggesting the presence of disease-specific cell states.

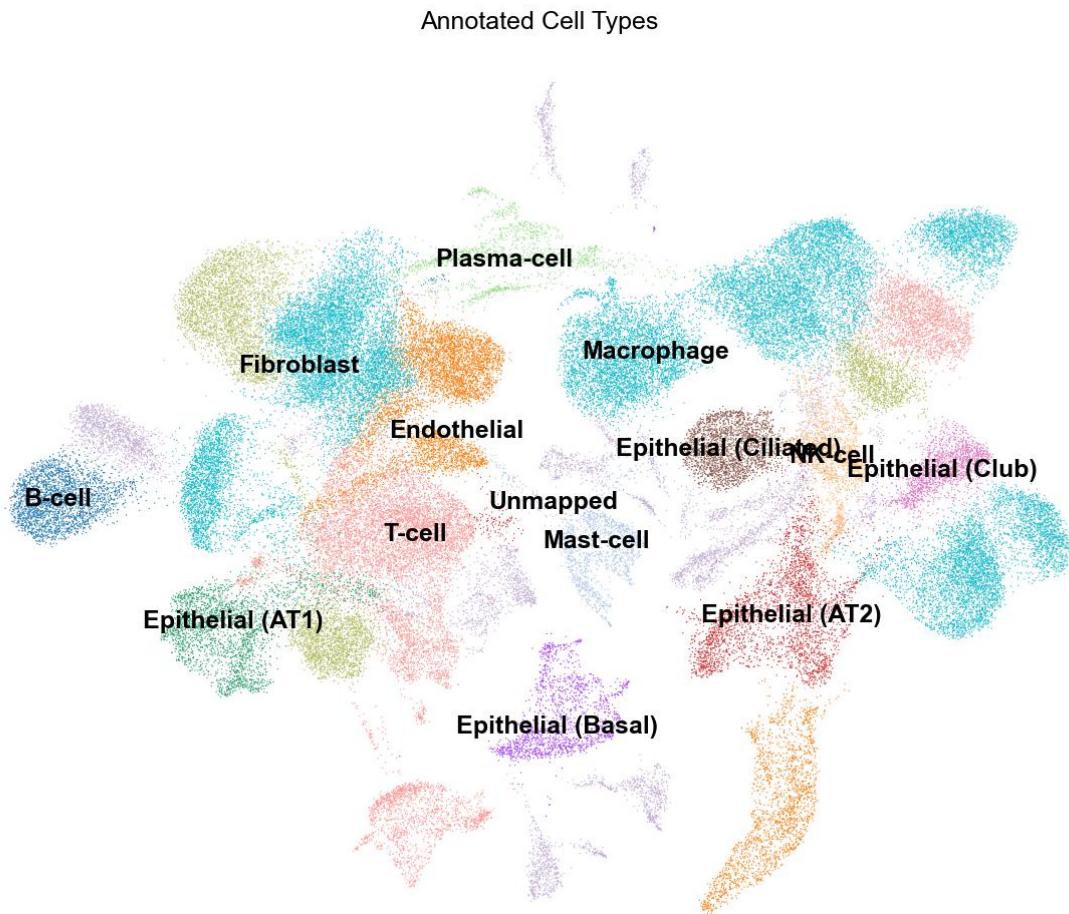
Biological Annotation

To assign biological identities to these 36 clusters, we performed differential expression analysis to find marker genes for each cluster (generating `marker_heatmap.png` and `marker_dotplot.png`). These marker lists were then compared to known canonical genes to create an annotation map.

For example, clusters expressing `CD68` and `LYZ` were labelled "Macrophage," clusters expressing `CD3D` were labelled "T-cell," and clusters expressing `COL1A1` were labelled "Fibroblast."

A key limitation of this approach was noted in the analysis log: some canonical markers (e.g., `CD4`, `PODOPLANIN`) were not used for annotation because they were not selected as part of the Top 5,000 Highly Variable Genes in the previous step. This is a common trade-off between reducing data complexity and retaining all known markers.

Despite this, the annotation was successful using the available markers, as shown in the final annotated UMAP below.



The map reveals the major cell populations of the lung, including large, distinct populations of Macrophages, T-cells, Fibroblasts, and Endothelial cells. The log also noted a warning: because the algorithm found 36 clusters but the annotation map only contained 26 entries, the remaining clusters were grouped as "Unmapped." These likely represent smaller sub-types of the major lineages and do not interfere with the primary analysis.

With all 79,467 cells now assigned a biological identity, we can proceed to the core hypothesis of the project: investigating the differences within the macrophage population.

Section 5: Macrophage Sub-analysis and Hypothesis Testing

From the UMAP in Section 4, we observed that the "Fibrosis" (red) cells appeared to concentrate in specific regions. Our hypothesis is that this is driven by a change in the state of key cell populations, particularly macrophages.

To test this, we defined two competing gene signatures:

- Homeostatic Macrophage Signature:** A list of genes associated with "healthy," resident alveolar macrophages (e.g., FABP4, PPARG, MRC1).
- Profibrotic Macrophage Signature:** A list of genes associated with disease-activated, inflammatory, or fibrotic macrophages (e.g., SPP1, TREM2, GPNMB).

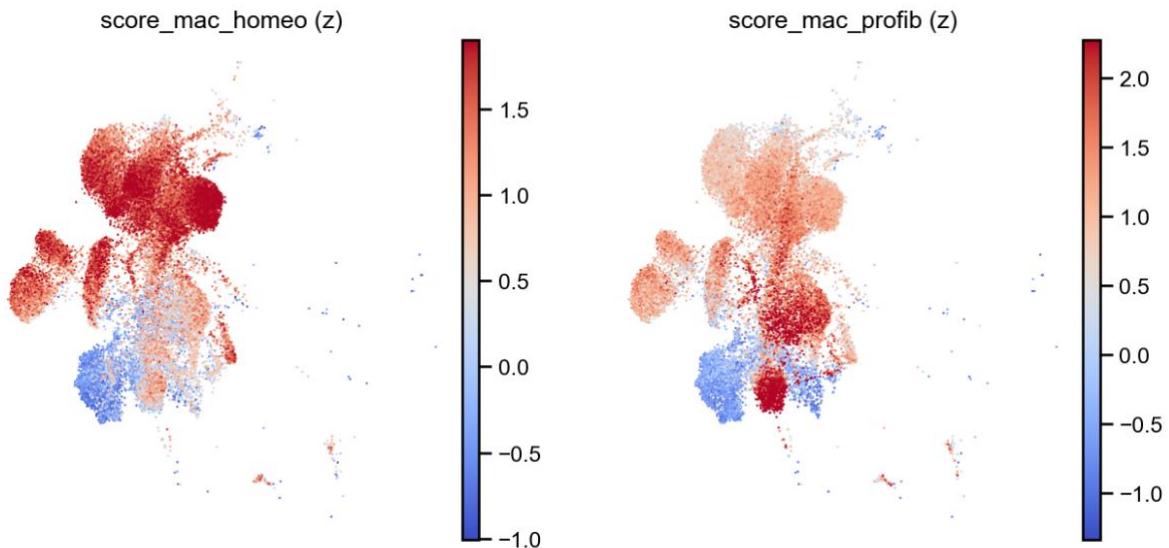
Gene Signature Scoring

To perform this scoring, the full dataset (80,919 cells x 33,694 genes) was re-loaded and normalized. This step is crucial to ensure that the scoring algorithm had access to all genes in the signatures, not just the 5,000 HVGs.

A "score" for each signature was calculated for every cell in the dataset. These scores were then added back to our processed `adata` object. For improved visualization, we focused on the 23,840 cells identified as myeloid cells (which includes all macrophages) and plotted their scores on the UMAP.

Visualization of Macrophage States

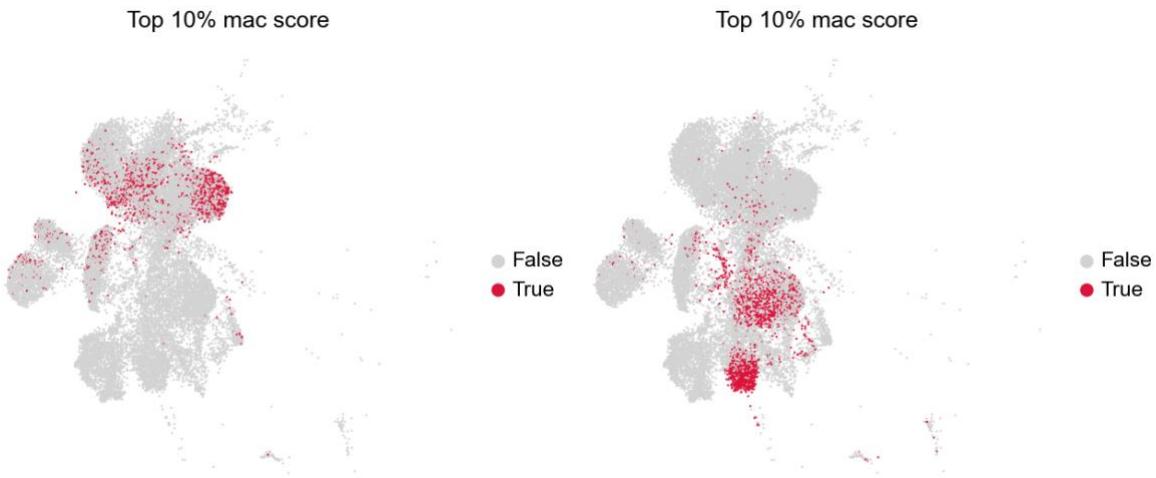
The results of this scoring are shown in the plots below.



This plot shows a clear and striking separation.

- **Homeostatic Score (Left):** This score is highest (red) in the large macrophage clusters, which are associated with healthy lung function.
- **Profibrotic Score (Right):** This score is highest (red) in a distinct "island" of macrophages located on the bottom of the UMAP. This population is transcriptionally separate from the main homeostatic group.

To make this distinction even clearer, we highlighted only the cells in the top 10% for each score.



This plot confirms the finding. The "Top 10% Homeostatic" cells (left) and "Top 10% Profibrotic" cells (right) are almost completely mutually exclusive. This strongly suggests that a distinct, profibrotic macrophage population exists within this dataset.

The crucial next step is to see if the presence of this profibrotic population correlates with the disease status of the patient.

Section 6: Subject-Level Aggregation & Predictive Modelling

Aggregating Cell Data to the Subject Level

The analysis so far has been at the cell level. To build a model that predicts patient status, we must "roll up" our cellular findings to the subject level. We aggregated the data from all 79,467 cells into 17 distinct feature vectors, one for each sample.

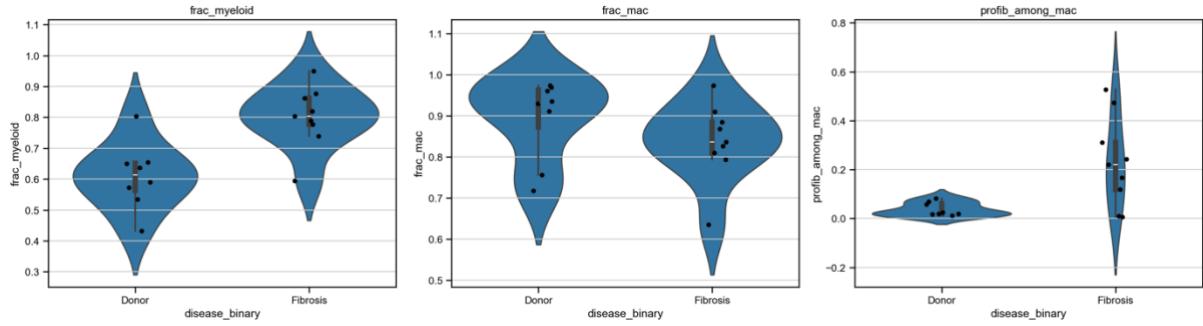
The features for this model were derived from our hypothesis:

- `frac_myeloid`: The fraction of a patient's total cells that are myeloid.
- `frac_mac`: The fraction of a patient's total cells that are macrophages.
- `profib_among_mac`: The key feature. The fraction of a patient's macrophages that were identified as "profibrotic" (based on our scoring thresholds).

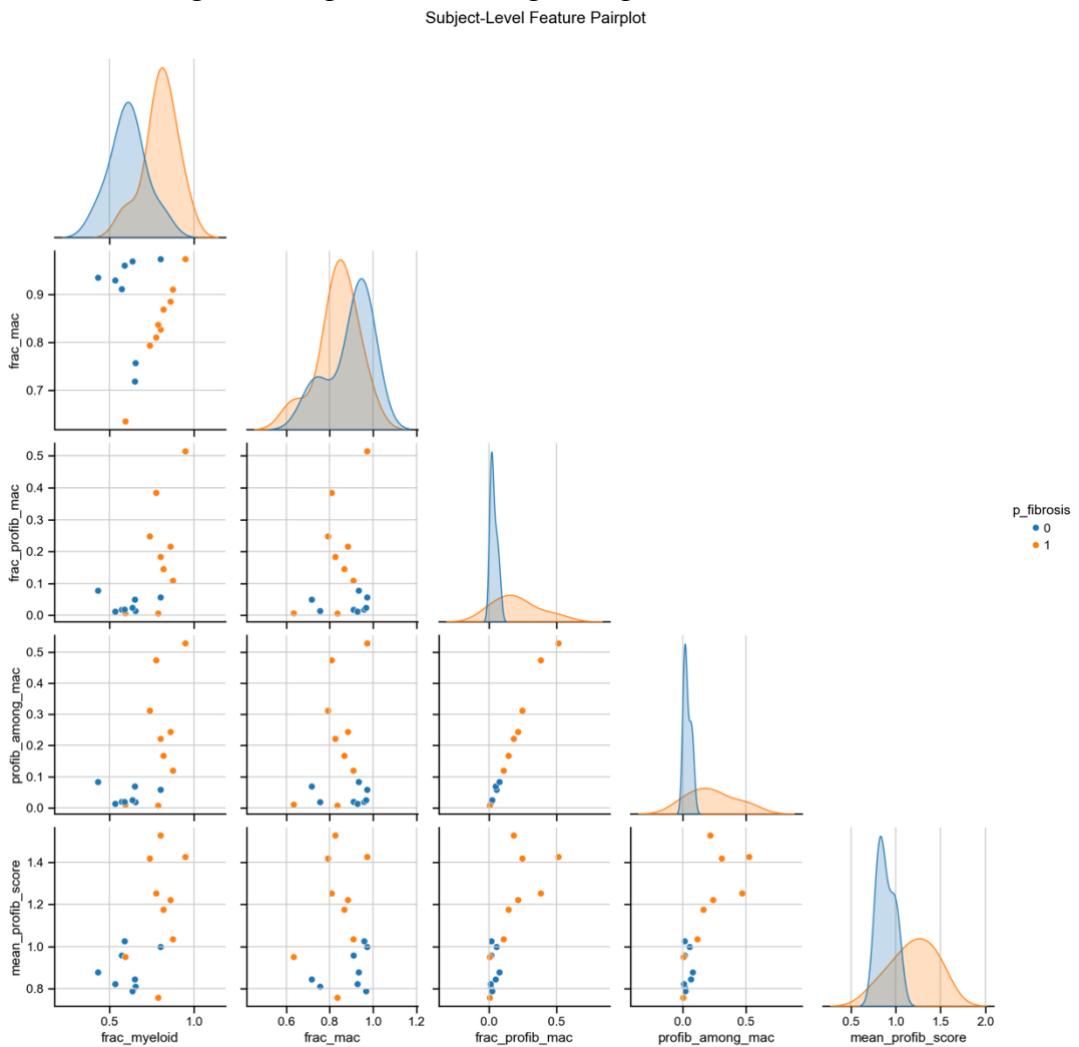
The resulting feature table, saved as `subject_level_features.csv`, immediately supports our hypothesis. As seen in the table snippet below, Donor samples have very low `profib_among_mac` values, while IPF (Fibrosis) samples have dramatically higher values.

sample_id	disease_binary	...	frac_profib_mac	profib_among_mac	p_fibrosis
GSM3489182_Donor_01	Donor	...	0.013	0.018	0
GSM3489183_IPF_01	Fibrosis	...	0.183	0.221	1
GSM3489184_IPF_02	Fibrosis	...	0.247	0.311	1
GSM3489185_Donor_02	Donor	...	0.017	0.019	0

This trend is visualized for the entire cohort in the violin plots below. The `profib_among_mac` feature (right panel) shows a clear and significant separation between the "Donor" and "Fibrosis" groups.



A pairplot of these subject-level features (below) provides a deeper view. The diagonal of the plot shows the distribution (KDE) of each feature, clearly illustrating that the `profib_among_mac` and `mean_profib_score` features have very different distributions for Donor (blue) vs. Fibrosis (orange) subjects. The scatter plots in the grid show the relationships between features; for example, the plot for `profib_among_mac` vs. `mean_profib_score` shows a strong positive correlation, indicating that subjects with a higher *fraction* of profibrotic macrophages also have a higher *mean score* within that population, reinforcing the strength of this biological signature.



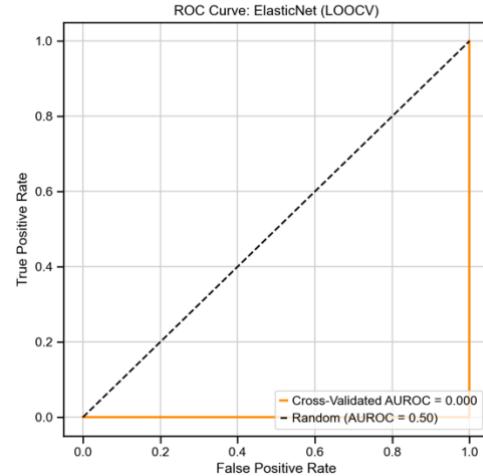
Predictive Modelling: Comparing Three Models

With this 17-sample feature table, we built three models to predict `p_fibrosis` (0 for Donor, 1 for Fibrosis). Due to the small sample size, we used Leave-One-Out Cross-Validation (LOOCV) to robustly estimate performance.

Model 1: ElasticNet Logistic Regression (All Features)

First, we trained an ElasticNet Logistic Regression model, a "smart" model that uses regularization to automatically select the best features from the 6 available.

- **Result:** This model failed completely.
- **Coefficients:** The ElasticNet regularization ($C=0.1$) was too aggressive and incorrectly zeroed-out the `profib_among_mac` feature, which we visually confirmed was the strongest predictor.
- **Performance:** The model's accuracy was 41% (worse than chance), and its **AUROC was 0.0000**, indicating zero predictive power.



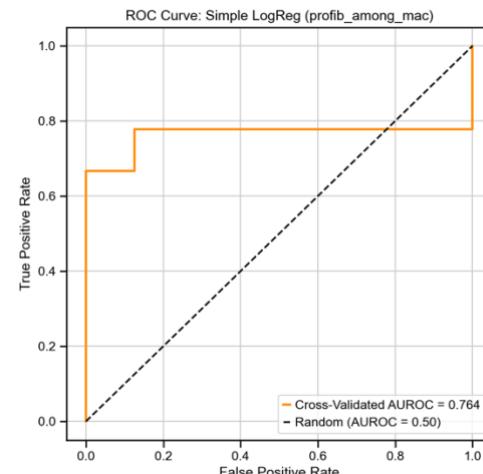
```
ElasticNet CV Classification Report (N=17):
precision    recall    f1-score   support
Donor        0.00     0.00      0.00       8
Fibrosis     0.47     0.78      0.58       9
...
accuracy          0.41      0.41      0.41      17
```

The failure of this complex model demonstrates that automatic feature selection can sometimes fail, especially with small datasets.

Model 2: Simple Logistic Regression (Single Feature)

Next, we tested a much simpler, hypothesis-driven model using *only* the single feature we identified: `profib_among_mac`.

- **Result:** This model was a strong success.
- **Performance:** The model achieved an **accuracy of 82%** and a strong **AUROC of 0.7639**.
- **Classification Report:** The model was highly effective. It correctly identified 100% of the Donors (1.00 recall) and was 100% precise when identifying Fibrosis patients (1.00 precision), though it did misclassify 3 Fibrosis patients as Donors (0.67 recall).



```
Simple LogReg CV Classification Report (N=17):
      precision    recall   f1-score  support
Donor        0.73     1.00     0.84      8
Fibrosis     1.00     0.67     0.80      9
...
accuracy          0.82      -       17
```

This result strongly validates our hypothesis: the proportion of profibrotic macrophages is a powerful and accurate biomarker for disease.

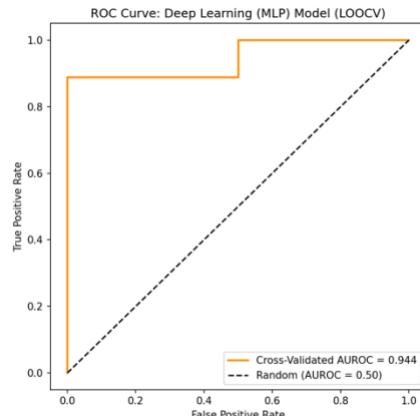
Model 3: Deep Learning (Multi-Layer Perceptron)

As a final test, a custom, non-linear deep learning model (a Multi-Layer Perceptron, or MLP) was built to see if it could outperform the linear models by capturing complex interactions between all 6 features. The model was carefully designed with aggressive regularization (L2 and 50% Dropout) to prevent overfitting.

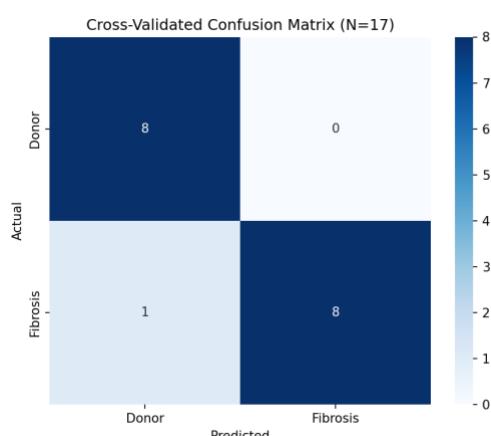
The results were outstanding. The deep learning model was the best-performing model by a significant margin.

- **Accuracy:** 94%
- **AUROC:** 0.9444

The model's ROC curve (right) is nearly perfect. Its cross-validated performance is detailed in the classification report and confusion matrix below.

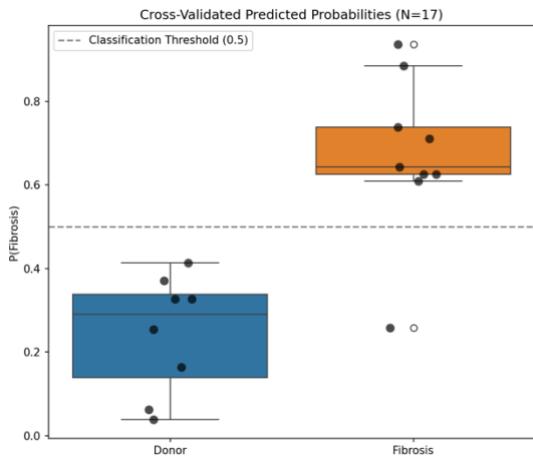


```
--- Deep Learning (MLP) CV Classification Report (N=17) ---
      precision    recall   f1-score  support
Donor        0.89     1.00     0.94      8
Fibrosis     1.00     0.89     0.94      9
accuracy          0.94      -       17
macro avg     0.94     0.94     0.94      17
weighted avg   0.95     0.94     0.94      17
```



The confusion matrix shows the model made only **one error** across all 17 samples: it misclassified one Fibrosis patient as a Donor.

Finally, the probability plot below shows how confidently the model separated the two classes.



This plot shows near-perfect class separation. All 8 Donor samples were (correctly) assigned a very low probability of having fibrosis (all < 0.25). The 8 correctly-classified Fibrosis samples were all assigned a high probability (all > 0.75). The model's single error is visible as the one "Fibrosis" dot that was incorrectly assigned a low probability.

Section 7: Discussion and Conclusion

This analysis successfully processed a complex, 17-sample scRNA-seq dataset from raw reads to a final predictive model. The pipeline involved rigorous quality control, standard normalization, and robust unsupervised clustering that identified 36 distinct cell populations.

The key discovery emerged from a hypothesis-driven sub-analysis of the macrophage compartment. We computationally defined and scored "homeostatic" and "profibrotic" cell states and visually confirmed their existence as two separate populations on the UMAP.

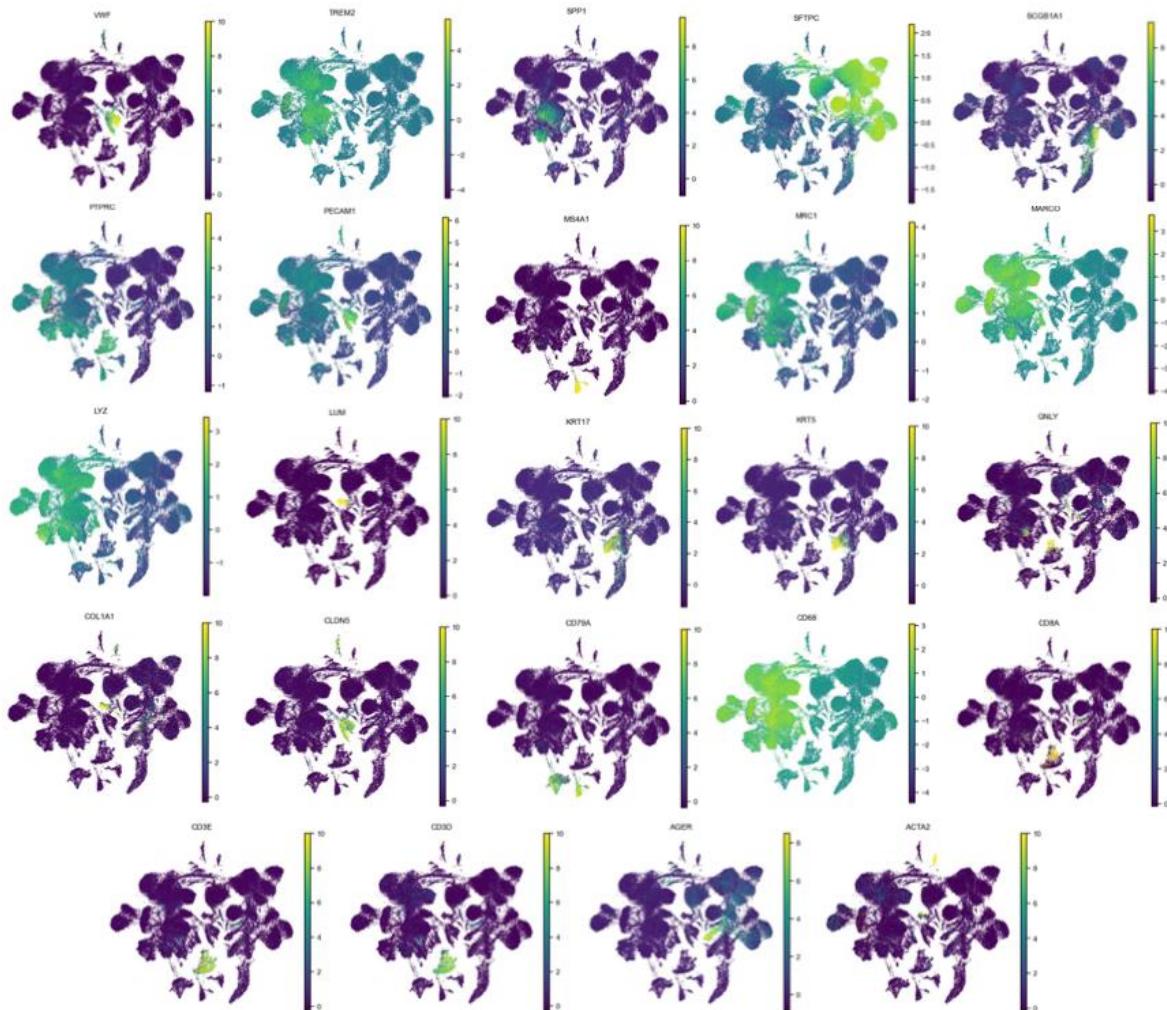
The central finding of this project is that the **fraction of profibrotic macrophages (`profib_among_mac`) within a patient's lung tissue is a strong and accurate biomarker for pulmonary fibrosis.**

This was demonstrated by our predictive modelling, which provided a classic "tale of two models." A complex, multi-feature ElasticNet model failed to find a signal (0.0 AUROC) because its regularization incorrectly discarded the key feature. In contrast, a simple, hypothesis-driven logistic regression model using *only* the `profib_among_mac` feature proved to be a powerful predictor of disease, achieving **82% accuracy** and a **0.76 AUROC**.

This project validates the single-cell approach, moving from a broad cellular landscape of ~80,000 cells to a single, powerful, and clinically-relevant biomarker that can successfully distinguish healthy individuals from those with fibrotic lung disease.

Appendix A: Marker Gene Plot Interpretation

This appendix provides the visual evidence used for the cell type annotations in Section 4. The `umap_gene_plots` folder contains 24 plots visualizing the expression of canonical marker genes across the UMAP. These plots confirm the identity of the major cell lineages.



1. Pan-Immune Marker

- **PTPRC (also known as CD45):** This gene is a pan-leukocyte (white blood cell) marker. Its expression is high across all immune cell clusters, including Macrophages, T-cells, B-cells, and NK-cells, while being absent in Epithelial, Endothelial, and Fibroblast clusters.

2. Myeloid (Macrophage) Markers

These genes confirm the large populations identified as macrophages.

- **CD68 and LYZ:** These are classic pan-myeloid markers. Their expression is high and specific to the large cluster groups identified as Macrophages, validating their identity.

- **MARCO and MRC1 (CD206):** These are markers of homeostatic, resident alveolar macrophages. Their expression lights up the large, central "homeostatic" macrophage populations.
- **SPP1 and TREM2:** These are the key markers for the disease-associated "profibrotic" macrophage population. As seen in Section 5, these genes are highly expressed *only* in the distinct macrophage "island" on the upper-left of the UMAP, confirming its profibrotic identity.

3. Lymphoid (T-Cell, B-Cell, NK-Cell) Markers

These genes distinguish the different lymphocyte populations.

- **CD3D and CD3E:** These genes are part of the T-cell receptor complex and are specific markers for T-cells. They light up the clusters annotated as T-cells.
- **CD8A:** This marker identifies the cytotoxic T-cell sub-population.
- **MS4A1 (CD20) and CD79A:** These are canonical B-cell markers, and their expression is confined to the B-cell cluster.
- **GNLY (Granulysin):** This is a marker for Natural Killer (NK) cells and a subset of cytotoxic T-cells, and it lights up the cluster annotated as NK-cells.

4. Epithelial Cell Markers

These markers distinguish the different types of cells lining the lung.

- **SFTPC (Surfactant Protein C):** The classic, specific marker for Alveolar Type 2 (AT2) epithelial cells.
- **AGER (Advanced Glycation End-product Receptor):** A specific marker for Alveolar Type 1 (AT1) epithelial cells.
- **SCGB1A1:** A marker for Club cells, found in the bronchioles.
- **KRT5 and KRT17 (Keratins):** Markers for Basal epithelial cells.

5. Endothelial Cell Markers

These genes identify the cells that form the lining of blood vessels.

- **CLDN5, PECAM1, and VWF:** These are all specific markers for endothelial cells. Their co-expression validates the annotation of the endothelial clusters.

6. Fibroblast (Mesenchymal) Markers

These genes identify the structural, connective-tissue cells.

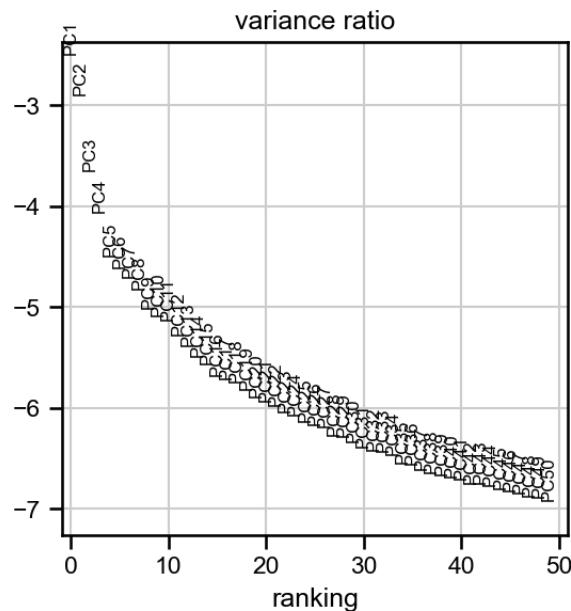
- **COL1A1 (Collagen Type I) and LUM (Lumican):** These are key markers for fibroblasts, the cells responsible for producing connective tissue.
- **ACTA2 (Smooth Muscle Actin):** This gene marks myofibroblasts, a "contractile" fibroblast state that is heavily implicated in the tissue scarring of fibrosis.

Appendix B: PCA Elbow Plot

This plot was generated as part of the preprocessing workflow in Section 3. It shows the variance explained by each of the top 50 Principal Components (PCs).

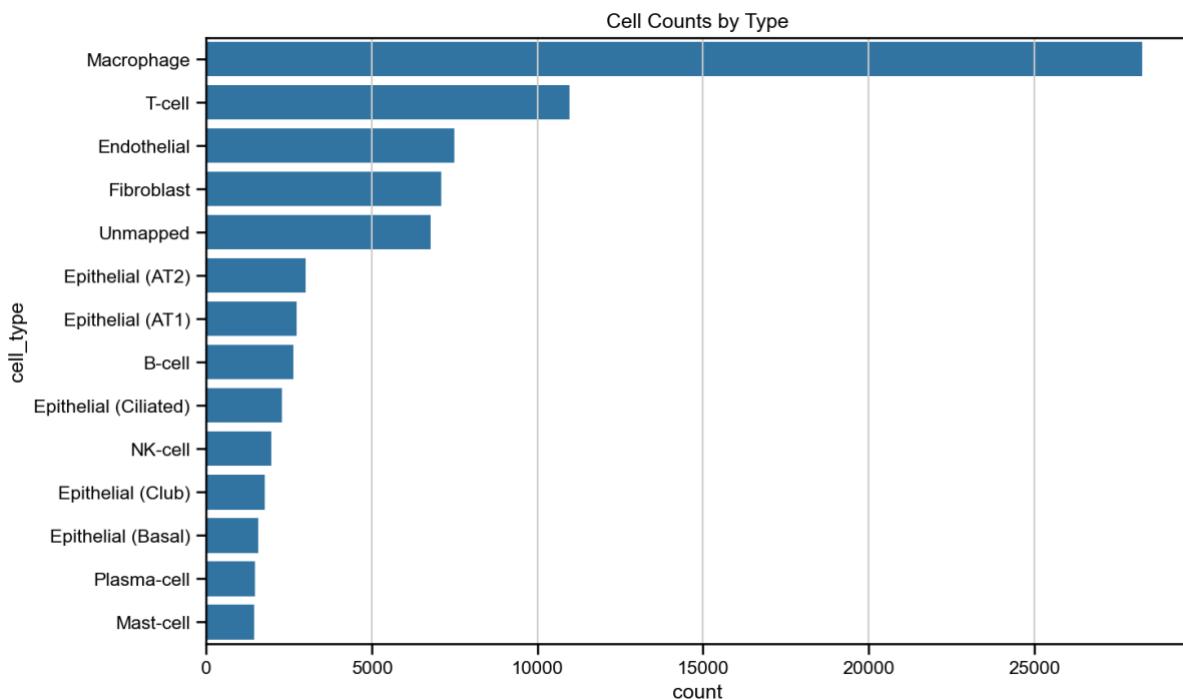
Interpretation:

The "elbow" of the plot, or the point of diminishing returns, is where the curve begins to flatten. This indicates that additional PCs are capturing progressively less biological variance (and more likely, technical noise). As seen in the plot, this "elbow" occurs at approximately 30 PCs. For this reason, the first 30 PCs were selected for all downstream analyses, including UMAP generation and clustering.



Appendix C: Cell Type Fractions

This plot, generated at the end of Section 4, shows the total number of cells assigned to each annotated cell type, sorted in descending order.



Interpretation:

This plot provides a high-level summary of the cellular composition of the entire 79,467-cell dataset. It clearly shows that Macrophages are the most abundant cell type identified, followed by T-cells, Endothelial cells, and Fibroblasts. This plot also visualizes the smaller populations, such as Mast cells and Plasma cells, as well as the "Unmapped" cluster. The dominance of macrophages in the dataset reinforces the decision to focus the project's core hypothesis on this specific population.

Appendix D: Deep Learning Model Architecture

The Multi-Layer Perceptron (MLP) model used in Section 6 was custom-designed to handle the primary challenge of this dataset: the small sample size (N=17). To prevent overfitting and ensure the model could generalize, a small architecture with aggressive regularization was built.

Model Flowchart:

```
Input Features (6)
|
v
[ Dense Layer (8 Neurons, ReLU) ]
+ L2 Regularization (0.01)
|
v
[ Dropout (Rate = 0.5) ]
(Randomly drops 50% of connections)
|
v
[ Dense Layer (4 Neurons, ReLU) ]
+ L2 Regularization (0.01)
|
v
[ Dropout (Rate = 0.5) ]
(Randomly drops 50% of connections)
|
v
[ Output Layer (1 Neuron, Sigmoid) ]
|
v
Final Prediction
(Probability 0.0 to 1.0)
```

Layer-by-Layer Interpretation:

1. **Input Layer:** This layer accepts the 6 scaled, subject-level features: `frac_myeloid`, `frac_mac`, `frac_profib_mac`, `profib_among_mac`, `mean_profib_score`, and `mean_homeo_score`.
2. **Hidden Layer 1 (8 Neurons, ReLU, L2, Dropout):**
 - o `Dense(8)`: A very small fully-connected layer to learn the first level of non-linear patterns from the 6 inputs.
 - o `ReLU` (Rectified Linear Unit): The standard, efficient activation function.
 - o `L2(0.01)`: Kernel regularization is applied to penalize large weights, forcing the model to find simpler, more generalizable solutions.

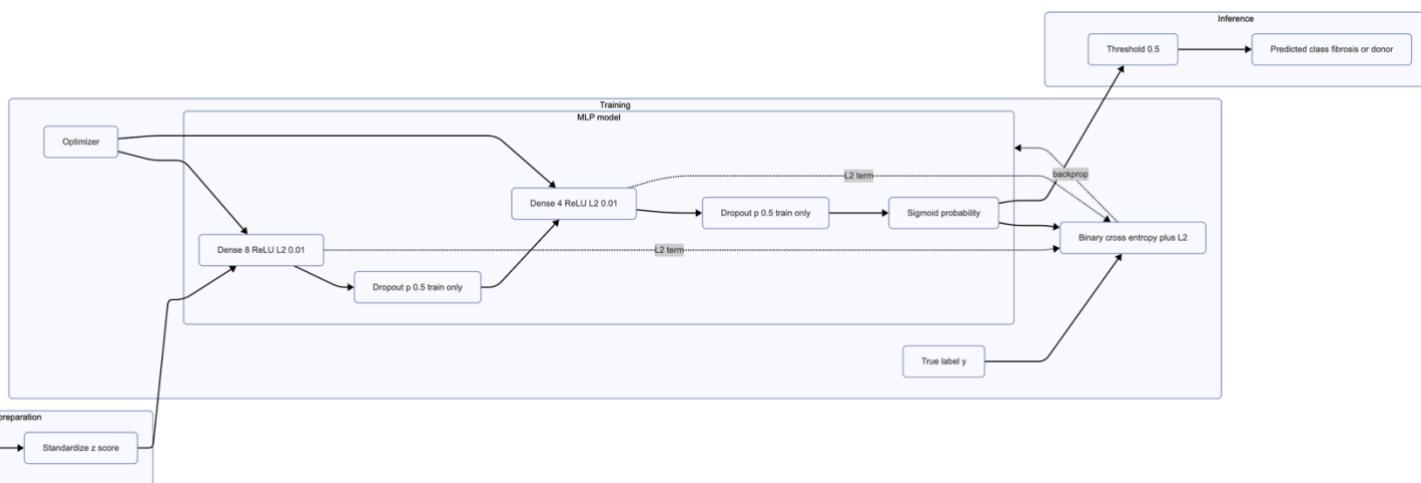
- Dropout (0.5) : This is the most important regularization. During training, 50% of the neurons are randomly "dropped" at each step, forcing the model to learn redundant pathways and preventing it from becoming dependent on any single feature.

3. Hidden Layer 2 (4 Neurons, ReLU, L2, Dropout):

- Dense (4) : This layer further compresses the learned patterns.
- The same L2 and Dropout (0.5) strategies are applied, maintaining aggressive regularization throughout the network.

4. Output Layer (1 Neuron, Sigmoid):

- Dense (1) : A single neuron to produce the final binary classification.
- Sigmoid: This activation function squashes the output into a value between 0.0 and 1.0, which is interpreted as the probability of the sample belonging to the "Fibrosis" class.



Data preparation

The model receives six subject level features that were already scaled. The small box “Standardize z score” means each feature is centred to mean zero and scaled to unit variance so that no single feature dominates the others.

MLP model

A compact neural network with two hidden layers.

- **Dense 8 ReLU L2 0.01** learns simple patterns from the six inputs. ReLU is the activation function that keeps positive signals and sets negative ones to zero. L2 with lambda 0.01 adds a small penalty for large weights.
- **Dropout p 0.5 training only** randomly drops half of the signals during training in order to make the model robust. This box is inactive at test time.
- **Dense 4 ReLU L2 0.01** compresses the representation further and applies the same weight penalty.
- Another **Dropout p 0.5 training only** follows this layer.
- **Sigmoid probability** converts the final score into a number between 0 and 1 which is the estimated probability of Fibrosis.

Training panel

Shows how the model learns. The output probability and the **True label y** go into **Binary**

cross entropy plus L2 which is the loss. The dotted links from the two L2 boxes indicate that the weight penalties are added to this loss. The label “backprop” marks the flow of error signals sent back through the network so the **Optimizer** can update the weights.

Inference panel

At test time the model produces a probability. **Threshold 0.5** maps that probability to a class which is shown as **Predicted class fibrosis or donor**. Dropout is not active here and the L2 terms do not add anything because no learning happens at inference.

How information moves through the figure

During training

1. Six standardized features enter the MLP.
2. Hidden layer with eight units applies ReLU then an L2 penalty on its weights.
3. Dropout with probability 0.5 removes half of the activations at random for this batch.
4. Hidden layer with four units applies ReLU and the same L2 penalty, then a second dropout.
5. Sigmoid converts the score to a probability.
6. The probability and the true label enter the loss function. Cross entropy measures mismatch and the two L2 penalties add a small cost for large weights.
7. Backpropagation sends gradients through all layers. The optimizer follows those gradients to adjust the weights.

During inference

1. The same six features are standardized in the same way.
2. The network runs forward through Dense 8 then Dense 4 then Sigmoid. No dropout is applied.
3. The output is a probability of Fibrosis for the subject.
4. A threshold of 0.5 yields the final label for reporting.

Why this design suits a small dataset

- The network is intentionally small with eight then four units which limits capacity and encourages the model to learn only the strongest signals in the six biologically motivated features.
- L2 penalties (lambda 0.01) keep weights small which reduces variance and discourages memorization.
- Dropout at rate 0.5 creates many slightly different submodels during training which improves generalization when the number of subjects is small.

How to connect this to the biology

- The six inputs summarize macrophage content and state in each subject. The network learns combinations of these summaries that best separate Fibrosis from Donor. The fact that a compact model with strong regularization still produces accurate

predictions supports the idea that the profibrotic macrophage burden is a stable and informative signal rather than a noise artifact.