Machine Learning Strategies for Biomarker Discovery: A Senescence Case Study

Euxhen Hasanaj

December 2024
CMU-ML-24-114

Machine Learning Department
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA

Thesis Committee
Ziv Bar-Joseph, Chair
Barnabás Póczos, Chair
Roni Rosenfeld
Oliver Eickelberg (University of Pittsburgh School of Medicine)

Submitted in partial fulfillment of the requirements
for the Degree of Doctor of Philosophy

This research was sponsored by: National Institutes of Health awards OT2OD026682 and U24CA268108, and National Science Foundation award 2134998

Keywords: biomarkers, senescence, aging, single-cell omics, steady state cells, dynamic systems, weak supervision, trajectory inference, supervised learning, unsupervised learning, graph structure learning, temporal graphs, multicommodity flow, transcriptomics, spatial omics, clinical data

Abstract

Genetic biomarkers play a crucial role in genome-to-phenotype mapping. They are essential for annotating cell identities, assessing treatment effects, monitoring cell progression and development, and exploring cell-cell interactions. The wealth and complexity of noisy genomic data presents a formidable challenge to the discovery of biomarkers, necessitating the development of computational approaches that can overcome experimental biases and sweep across vast genetic landscapes. Machine learning algorithms, which address both the scale of data and its inherent stochasticity, have emerged as a natural solution to these issues.

In this thesis, we explore machine learning strategies for biomarker discovery across different contexts. Our investigation is categorized into two primary themes; steady state and dynamic. In the steady state context our main objective is to identify biomarkers that differentiate conditions, cell types, and cell states within a given sample. We design computational methods that specifically target the static context across three learning settings, contingent upon the nature and availability of label information. These include supervised, weakly supervised, and unsupervised approaches. For the dynamic context, we are primarily concerned with the discovery of biomarkers that vary with time and explain the dynamics of the biological system being investigated. Here we focus on the problems of trajectory inference, as well as deep learning approaches for temporal graph learning. By applying these methods to senescence (a form of aging) and other real world datasets and diseases, we substantiate their practical value in unraveling the intricate relationship between genes and phenotypes.

To my parents

Acknowledgments

This thesis is the culmination of years of research, learning, and support from many remarkable people. I am deeply grateful to each of them for their contributions, encouragement, and inspiration throughout my academic journey.

First and foremost, I would like to express my profound gratitude to my advisors, Ziv Bar-Joseph and Barnabás Póczos. Their mentorship encouraged me to pursue my own intellectual path while maintaining the highest standards of rigor and integrity. Ziv has been with me from the very beginning of this journey. When I started, I was new to the field of genomics, yet under his insightful mentorship, I gained not only a strong foundation in the field but also a profound appreciation for approaching research with purpose. Barnabás's expertise and critical approach served as a guiding resource and a rigorous check on my theories, ensuring that my work was grounded in precision and scientific rigor. My appreciation extends to both their current and former lab members. It has been a privilege to work alongside such remarkable and dedicated individuals.

I am deeply grateful to Roni Rosenfeld and Oliver Eickelberg for serving on my thesis committee. Their invaluable advice and guidance have greatly shaped this work, with many of their suggestions incorporated into this text.

I am deeply appreciative of my collaborators, whose shared knowledge and enthusiasm have contributed immensely to the development of this thesis. My sincere thanks go to Jun Ding, who mentored me throughout my master's years. I also extend my special thanks to Oliver Eickelberg and Melanie Köenigshoff for giving me the opportunity to be involved in lab work at the University of Pittsburgh School of Medicine (UPMC). To everyone in the Blue and Pink labs, thank you for creating a productive, supportive, and enjoyable environment.

I am also fortunate to have found a wonderful community of friends and colleagues at CMU who provided a sense of balance, camaraderie, and encouragement throughout this journey. Whether it was through late nights in the office, lively discussions on AI, or gym sessions, their support made the achievements more meaningful. Special thanks to Youngseog Chung and Shantanu Gupta. Shantanu and I started this journey together, from the master's program to the PhD. Many late nights on the 8th floor were spent discussing ML over pizza. Youngseog and I understood each-other well and developed a friendship that was deepened by our shared interest on aging. Many other good friends in the department include Tanya Marwah, Zhili Feng, Chris Ki, Conor Igoe, Aakash Lahoti, Dylan Sam, Emily Byun, Sachin Goyal, Jacob Tyo, Brandon Trabucco, Daniel Jeong, Tom Yan, Rattana Pukdee, Yuda Song, Pratyush Maini, Chirag Gupta, Ian Char, Keegan Harris, Jacob Yeung, Ojash Neopane, Terrance Liu, Tomás González Lara, Kevin Li, Justin Whitehouse, YJ Choe, and many others. Other good friends outside the department include John Protano, Nicholas Roberts, Vishwak Srinivasan, Francis Anim, Tejus Surendran, Salil Singh, and more. Thank you to all for making this PhD journey fun.

Finally, I am deeply indebted to my family for their love and support; thank you for being my foundation and my strength. To my wife, Ketevan: thank you for believing in me at every step of the way.

Contents

List of Figures 5

List of Tables 12

1 Introduction 15

1.1 All of Molecular Biology . . . 16
1.2 All of Machine Learning for Molecular Biology . . . 17
1.3 Cellular Senescence . . . 19
1.4 Experimental Technologies for Characterizing Senescence . . . 19
1.5 Biomarker Discovery . . . 21
1.5.1 Biomarker Discovery in Steady State Cells . . . 21
1.5.2 Biomarker Discovery in Dynamically Changing Cells . . . 23

2 Biomarker Discovery in Steady State 24

2.1 Supervised Setting: Minimal Gene Sets for Accurate Cell Type Identification . . . 24
2.1.1 Results . . . 26
2.1.2 Materials and Methods . . . 32
2.1.3 Discussion . . . 35
2.2 Weakly Supervised Setting: Discovering Senescence Markers from Aging Atlases . . . 37
2.2.1 PU Learning Under Covariate Shift for Identifying Senescent Cells . . . 37
2.2.2 Results . . . 38
2.2.3 Materials and Methods . . . 52
2.3 Unsupervised Setting: Tools for Biomarker Discovery and Cell Type Annotation . . . 58
2.3.1 Typical Workflows in Cellar . . . 60
2.3.2 Materials and Methods . . . 62
2.3.3 Discussion . . . 63

3 Biomarker Discovery in a Dynamic Context 65

3.1 Endotype-Informed Biomarkers from Time Series Clinical Transcriptomics Data . . . 65
3.1.1 Materials and Methods . . . 66
3.1.2 Results . . . 70
3.1.3 Discussion . . . 75
3.1.4 Acknowledgments . . . 76
3.2 Recovering Time-Varying Networks From Single-Cell Data . . . 77
3.2.1 Materials and Methods . . . 78

3.2.2 Results 81
3.2.3 Discussion 87
3.2.4 Limitations 88
3.2.5 Acknowledgements 88

4 Discussion **89**
4.1 Future Work 91

Bibliography **93**

Appendices **110**
A Supervised Setting: Minimal Gene Sets for Accurate Cell Type Identification 110
B Endotype-Informed Biomarkers from Time Series Clinical Transcriptomics Data . . . 117
C Recovering Time-Varying Networks From Single-Cell Data 120

List of Figures

1.1 **Illustration of some sequencing technologies used in this thesis.** (Clockwise, starting at top left). 1) Bulk RNA-seq quantifies the average counts of mRNA molecules for the entire sample. 2) sn/scRNA-seq quantifies gene expression at the cellular level. 3) Spatial transcriptomics/proteomics offers spatial locations of cells in addition to their expression profiles. 4) ATAC-seq measures regions in the DNA with chromatin accessibility. 20

1.2 **Summary of the learning settings addressed in this thesis.** In the static context, the main objective is to identify biomarkers that differentiate steady states, such as cell types within a given sample. The dynamic setting is concerned with the discovery of markers that explain the dynamics of a system such as disease progression or the evolution of gene interactions over time. Each setting is detailed in the rest of this thesis. 21

2.1 **Illustration of Phenotype Cover and the multiset multicover approach.** Given a binary score matrix (left), each feature induces a bipartite graph between classes (center left). Edges in this graph form a set $\mathcal{E}_s$. Multiset multicover is then performed on the collection of $\mathcal{E}_s$ to select a small number of features which “distinguish” all phenotypic pairs (at least K times). The idea can be naturally extended to non-binary score matrices by assigning a multiplicity to each element $e_{ij} \in \mathcal{E}_s$ (See Methods). 26

2.2 **Comparison of feature selection methods for the IPF dataset.** Performance scores for (a) and (b) were averaged across five different random train and test splits. Standard deviation is shown as a shaded region. (**a**) Performance of a logistic regression model trained on the selected features. G-PC achieves the highest F1 score across all coverage factors, followed by CEM-PC and decision trees. (**b**) Jensen-Shannon divergence (lower is better) between CIBERSORT-predicted mixture proportions and the ground truth. (**c**) Stability scores for all eight methods over 5 runs. Sequential methods like G-PC, decision trees and CEM-PC suffer slightly in stability compared to other, more global methods. Nonetheless, G-PC shares about 70% of the features across runs. (**d**) Biomarker membership matrix. Gene s (columns) is assigned to cell type i (rows) if there exists another cell type j such that $\mathbf{M}_{i,s} - \mathbf{M}_{j,s} \geq 1$ (see Section 2.4). Rows and columns were ordered based on hierarchical clustering. Colors and shapes only serve visibility. (**e**) For every phenotypic pair, we compute the "coverage" (i.e., the score difference between the two phenotypes) provided by the selected gene set. A histogram of these coverage factors corresponding to a coverage of 10 is shown for each method. As can be seen, for G-PC and CEM-PC which optimize for coverage, each element is covered at least 10 times. Other methods provide high coverage for some elements but miss out on others. 27

2.3 **GSEA q-values for HCA** We select markers that provide coverage for each cell type for both G-PC and decision trees and perform gene set enrichment analysis (GSEA) using the HuBMAP ASCT+B gene set[79]. We first record q-values for the top entry which contains the correct cell type or the correct tissue independently. When comparing the ability of each method to assign the correct cell type, G-PC obtains a lower q-value (i.e., higher $-\log(q\text{-value})$) in 42% (3/7) of the cases (**a**). A similar analysis shows that G-PC obtains lower q-values in 54% (6/11) of the tissues (**b**). We did not find markers for four tissues in the gene set (common bile duct, muscle, rectum, stomach). (**c**) When tested for the ability to identify both the correct tissue *and* the correct cell type, G-PC obtained lower q-values in 71% (30/42) of the cases. The remaining tissue/cell type pairs (33) either belonged to a tissue which was not present in the marker set, or were not identified by either method. (**d**) Connected by an edge are known markers for CD4 and myeloid cells that were assigned to the correct tissue/cell type pair by G-PC. Some markers are assigned to multiple cell types (multiple outgoing edges), while others are pair specific. 30

2.4 **PCA plots of classes in HCA** A total of 121 markers were selected via G-PC (coverage= 5). For every tissue (**a**) and cell type (**b**), top two principal components of the markers providing coverage (≥ 1) for that phenotype (number in parenthesis) are plotted. There is visible separation between classes. 31

2.5 **Schematic illustration of PUc learning for identifying SnCs.** Given a large single-cell lung cohort from young and old individuals (top), we designate cells from young individuals as non-senescent cells (positive). Cells in older individuals are unlabeled initially. We then use positive-unlabeled learning under covariate shift (PUc) to identify SnCs in older individuals (middle). Using these cells we develop an expression profile for senescence markers in several different cell types (bottom). . 39

2.6 **Summary of the The Human Lung Cell Atlas**. (A) Distribution of donors by age and smoking history. (B) Number of donors by age group and sex. (C) Number of donors broken down by tissue and sampling method (non-smokers). (D-F) UMAP plots describing level 2 cell types, age groups, and sex ((NS = non-smoker)). (G) Cell type proportion among non-smokers. (H) Age group representation across cell types (non-smokers). (I) Average cell total counts for each donor. (J) Normalized gene expression values for *CDKN1A* and *CDKN2A*. 40

2.7 **PUc identified a novel SenSet senescence signature**. (A, left to right) The number of cells (≥ 50) identified as senescent in $\mathcal{A}$ and the corresponding percentage; the most frequent genes enriched for most cell types; the total number of SenSet marker genes assigned to a cell type. (B) Overlap sizes of SenSet with the prior lists. (C) Selected marker genes for some of the cell types and distribution among healthy and SnCs. (D) UMAP plot of senescent cells in $\mathcal{A}$. (E) UMAP plot of cell types assigned at least one marker in $\mathcal{A}$. (F) Top GO, Jensen, and MSigDB terms enriched for SenSet. (G) SenSet marker genes enriched in basal$^{(-)}$ cells, fibroblasts$^{(-)}$ and AT2$^{(-)}$ cells. 42

2.8 **Senescence induction in human PCLS by DNA damage.** (A) PCLS were generated from healthy donors lower left lobe lung with an age range of 20 to 78 years-old. Senescence was induced by treatment with bleomycin (Bleo) at 15 mg/mL, or doxorubicin (Doxo) at 0.1 µM for 6 days, and PCLS and supernatants were collected. (B) Hematoxylin eosin (H&E) staining on 4 µm sliced formalin-fixed paraffin embedded human PCLS at day 6. (C) β-galactosidase staining on whole PCLS at day 6. (D) p21 immunohistofluorescence (IHF) on 4 µm sliced formalin-fixed paraffin embedded human PCLS at day 6. (E) p21 positive cells quantification based on p21 IHF staining presented in (D) after bleomycin ($n = 7$) or doxorubicin ($n = 6$) treatment. (F) Quantification of p21 protein level by Western blot (WB) after bleomycin ($n = 8$) or doxorubicin ($n = 6$) treatment and representative blot. (G) SASP factor, *GDF-15*, measured by Luminex$^{\text{TM}}$ assay on human PCLS supernatants after bleomycin or doxorubicin treatment ($n = 5$). Paired t-test: $^{**}p < 0.005$, $^{*}p < 0.05$. . 46

2.9 **Validation of SenSet**. (A) Average total counts per cell across samples and conditions. (B) For each subject, we show the fraction of the genes in each list which were significantly up or downregulated with treatment. (C) SenSet genes up (down)regulated in most samples. (D) UMAP plot of scVI integrated data. (E) Clusters identified using Leiden clustering on scVI embeddings. (F) Normalized expression of *CDKN1A* and *CDKN2A* across conditions. 47

2.10 **Cell Type-specific signatures**. (A) For each cell type and gene set, we ran DE tests between the two conditions and show combined p-values (Pearson's method) for each marker gene. (B-C) Rank sum test statistics for every marker gene and cell type in PCLS. 49

2.11 **Comparisons of senescence marker genes between smokers and non-smokers in the HLCA.** (A) Wasserstein distance between the gene expression profiles of smokers and non-smokers across different age groups. Cell types inside the red box exhibited a smaller distance for the pair (young smokers, old non-smokers) when compared to (young smokers, young non-smokers). (B) Fraction of genes enriched in smokers compared to non-smokers among young ($\mathcal{Y}$) and old ($\mathcal{A}$) patients for selected cell types. (C-D) STEM significant profiles and the corresponding gene curves, categorized by smoking status and sex. Genes from significant profiles were combined into one plot. 53

2.12 **Cellar's workflow.** (a-c) Preprocessing (optional). (d, e) Dimensionality reduction and visualization. Several methods for dimensionality reduction are implemented as part of Cellar. The reduced data is then visualized by running another (possibly the same) dimensionality reduction method. (f-i) Clustering. Cellar supports several unsupervised and semi-supervised clustering methods. It also implements supervised label transfer methods. (j-l) Cell type assignment. Cellar enables the use of several functional annotation databases for the assignment of cell types. 59

2.13 **CODEX data analysis in Cellar.** (ID: 19-003 lymph node R2) **a** UMAP visual representation of a lymph node CODEX dataset with 46,840 cells, clustered via Leiden. **b** Projection of the assignments on the spatial CODEX image that can be visualized side-by-side in Cellar. Cluster assignments were copied from a. Not all clusters could be assigned to unique cell types given that only a few ten protein levels are measured, though several have been assigned based on differential gene analysis in Cellar. The B-Cell clusters are surrounded by T-cells and other cells types in the lymph. The B-Cell clusters also contain a subset of proliferating cells. 61

2.14 **SNARE-seq data analysis in Cellar.** (IDs: kidney SNARE ATAC/RNA 20201005) **a** UMAP plot of the chromatin modality for the kidney SNARE-seq dataset with 31,758 cells. First, we obtain a cell-by-cistopic matrix by running cisTopic which is then used to define clusters via Leiden clustering. **b** Corresponding UMAP plot of the expression matrix with cluster assignments copied from a. Cellar's dual mode allows a cell ID based label transfer from one modality to the other. 62

3.1 **Schematic illustration of Truffle.** For each patient, our flow algorithm returns a trajectory that passes through intermediate nodes for a smoother response. These trajectories are then aligned with the clustering results to obtain a state diagram. Finally, by estimating state initial and final probabilities from the data, we can compute and study the top directed trajectories. 68

3.2 **Clustering analysis of the psoriasis dataset.** (a-b) Distribution of visits across patients. (c) Uniform Manifold Approximation and Projection (UMAP) plot of cluster assignments. (d) Boxplots of PASI scores for each cluster. (e) Relative frequency of visits by cluster. (f) Top Gene Ontology (GO) terms for each cluster against healthy samples. We used a Kolmogorov-Smirnov (KS) test to rank the genes. A $(*)$ symbol means the category was statistically significant $((**) \equiv q \approx 0$ and $(*) \equiv q \leq 0.05)$. 70

3.3 **Truffle state diagram and top trajectories for the psoriasis dataset**. (a) Original connectivity graph obtained using fuzzy simplicial sets and (b) the graph corresponding to all the low-cost trajectories selected by Truffle (right). We used an edge capacity of 1 and a node capacity of 3 for this dataset. (c) The pruned state diagram describing the main state transitions in the Truffle network. Repeated states were collapsed into one, hence, no self-loops are shown. (d) The top paths identified by Truffle. 71

3.4 **Trajectories uncovered by Tempora and psupertime for the psoriasis dataset.** (a) Transition graph identified by Tempora. Five trajectories of length 1 were identified. (b) Separation of time points by psupertime. The y axis is the density of each time point and the x axis is the temporal ordering. (c) The top 5 genes identified as relevant by psupertime. These correspond to the genes with the largest absolute coefficients. (d) The top GO terms for all the relevant genes (294). Subfigures (b) and (c) were generated using psupertime. 72

3.5 **Selected STEM profiles for the top three Truffle trajectories in the psoriasis dataset.** (a-c) Two selected profiles for each of the three trajectories. In (c, right) “IL-27 Mediated Signaling Pathway” obtained a very high combined score (1e6), hence, was removed from the plot for clarity. The full list of profiles can be found in the appendix. 73

3.6 **Clustering and trajectory analysis for the COVID-19 dataset.** (a-b) Distribution of visits and distribution of visit counts per patient. (c) Clustering 650 samples from 304 patients. (d) Relative frequency of visits and (e) acuity scores per cluster. (f) A pruned diagram of top state transitions identified by Truffle. Pruning was performed by taking the fewest top edges that amount to $\geq 50\%$ of a node’s outgoing weight. (g) The tree learned by Tempora. Final states are 3, 1, and 5. (h) The top genes that vary with time according to psupertime (plot obtained from psupertime). (i) Selected STEM profiles for Truffle trajectories P_1 (green), P_3 (red), P_4 (blue). 74

3.7 **Overview of Marlene**. Marlene takes as input gene expression data in the form of a cell-by-gene matrix. It then performs gene featurization via the pooling by multihead attention (PMA) mechanism which returns a gene feature matrix. This matrix is then inputted into a self-attention module to obtain a gene network in the form of an adjacency matrix. The weights of the self-attention module evolve from one time point to the next via a gated recurrent unit (GRU). The expression of transcription factors and the recovered graph are used to reconstruct the full gene expression vector. Finally, the reconstructed matrix is used to predict the cell type for the batch. The network is trained in a model-agnostic meta-learning fashion where each cell type is treated as a “task” to be learned, thus enabling the model to quickly adapt to cell types with low representation. 79

3.8 **Overlap analysis of the SARS-CoV-2 vaccination dataset**. Showing $-\log_{10}(\text{FDR})$ values from a Fisher’s exact test measuring the overlap between predicted TF-gene interactions in reconstructed networks and two TF-gene interaction databases, TRRUST (top) and RegNetwork (bottom). Cell types are shown as columns. Best performing method is starred. 83

3.9 **Temporal analysis of the predicted gene regulatory networks for the SARS-CoV-2 vaccine dataset**. (a) Intersection-over-union (IoU) scores between consecutive graphs. (b) For each method, top 3 MSigDB terms enriched for genes that were regulated at day 2 but not day 0. 84

3.10 **Results on the HLCA dataset**. (a) FDR corrected p-values of Fisher exact tests reflecting the number of links that overlap with TRRUST and RegNetwork databases. (b) Top 3 Jensen Diseases terms enriched for genes added between the first and second age group. 86

3.11 **Enrichment for senescence using the SenMayo set**. For 4 cell types, there was statistically significant enrichment for the oldest age group. We only used the top 200 regulated genes. 87

A.1 **Comparison of feature selection methods for MC.** Performance scores for (a) and (b) were averaged across five different random train and test splits. (**a**) Performance of a logistic regression model trained on the selected features. (**b**) Jensen-Shannon divergence (lower is better) between CIBERSORT-predicted mixture proportions and the ground truth. (**c**) Stability scores for all eight methods when ≈ 139 features were selected (coverage= 40) over 5 runs. (**d**) Histogram of coverage factors per element (phenotypic pair) for the test set. Number of genes selected (66) corresponds to a coverage of 20. 113

A.2 **Comparison of feature selection methods for HCA.** Performance scores for (a) and (b) were averaged across five different random train and test splits. Standard deviation is shown as a shaded region in (a) but removed from (b) for better visibility. (**a**) Performance of a logistic regression model trained on the selected features. (**b**) Jensen-Shannon divergence (lower is better) between CIBERSORT-predicted mixture proportions and the ground truth. (**c**) Stability scores for all eight methods when ≈ 121 features were selected (coverage= 5) over 5 runs. (**d**) Histogram of coverage factors per element (phenotypic pair) for the test set. Number of genes selected (53) corresponds to a coverage of 2. 114

A.3 **Performance of a KNN classifier** trained on features selected by each method (a-c) and deconvolution performance based on linear least squares (d-f) for all three datasets. 115

A.4 **Robustness to batch effects and preprocessing.** (**a**) Batch effects. To test the impact of batch effects, we ran feature selection on a pancreas dataset from `https://pubmed.ncbi.nlm.nih.gov/27667667/` and used the selected features to train a Logistic Regression model on a pancreas dataset from a different study `https://pubmed.ncbi.nlm.nih.gov/27693023/` (after a train/test split). We only used overlapping genes and cell types (7) between the two datasets, reducing the number of cells for the former to 1,864 and for the latter to 1,941. The chart shows that G-PC and TopDE are the most robust to batch effects. (**b**) Comparison against different scGeneFit variants for the MC dataset. All hyperparameters were taken from `https://github.com/solevillar/scGeneFit-python/blob/master/examples/scGeneFit_example.ipynb`. (**c**) IPF - Distribution of gene counts for the top 10 markers selected via G-PC ($k = 10$). We observe that our method selects a combination of genes with different expression levels at baseline, showing that it is not affected by the basal expression of genes. (**d**) MC - Performance of each method on log-transformed and scaled MC data versus non log-transformed and unscaled data. 116

B.5 **Toy example demonstrating the need for node capacities.** Assume a situation where we have a single outlier (in green). Without capacities, we would incorrectly identify the trajectory 0-2-1 since all patients would travel through the green node. With node capacities, however, only one patient is allowed to travel through 2, and the rest travel through the more likely disease path 0-1. The possibility of such outliers impacting the results depends on the size of the dataset and the sampling rates. 117

B.6 **Truffle on simulated data.** (a) We randomly generated 15-dimensional samples from 6 different states and constructed random patient trajectories from these (by randomly deleting some nodes to simulate sparse visits). We designated 0-5-1-4, 5-1-3-4, and 4-3-1-2-5 as “true trajectories”. From these, Truffle accurately identified 4-3-1-2-5 as the top trajectory of length 4, and also 5-1-3-4 as the top trajectory of length 3. The other path 0-5-1-4 was one of the top three trajectories of length 3. . . 117

B.7 **Supplementary plots for the psoriasis dataset.** (a) PASI scores for each patient. (b) All significant STEM profiles for all three trajectories identified by Truffle. (c) Profile 0 for the trajectory $0 - 5 - 1$ shows a high score for “IL-27-Mediated Signaling Pathway.” (d) Complete heatmap with all significant GO processes (FDR < 0.05). 118

B.8 **Supplementary plots for the COVID-19 dataset.** (a) Selected STEM profiles and their GO processes. (b) STEM profiles for Tempora’s trajectory $8-7-9-2-3$ along with GO processes for profiles 37, 7, and 10. 118

B.9 **Supplementary plots for the Crohn’s disease dataset.** (a) Information on the number of visits per patient and cluster. (b) Clustering results. (c) Top GO processes for each cluster when compared to healthy samples. (d) STEM profiles for the top two paths $3-4-1$ and $2-5-0$ along with GO processes for the top profile for each. 119

C.10 (a) FDR corrected p-values of Fisher exact tests reflecting the number of links that overlap with the two TF-gene databases. (b) IoU scores across time reflecting the overlap between consecutive graphs. 121

C.11 (a) FDR corrected p-values of Fisher exact tests reflecting the number of links that overlap with the two mouse databases. (b) IoU scores across time reflecting the overlap between consecutive graphs. 122

List of Tables

2.1 **Three datasets were used in the PhenotypeCover study**. For HCA, we consider a combination of tissues and cell types (85). For IPF only healthy samples were kept. Endothelial-mural and astrocyte-ependymal pairs of cells were grouped for MC. 28

2.2 **Runtimes for all methods across all three datasets** (seconds). Names were abbreviated. 178 features were selected for IPF, 66 for MC, and 121 for HCA. Our C++ implementation of G-PC takes less than 2 seconds for all three datasets, making it the fastest along with F value computation. Performance tests are conducted on a machine with a 2.3 GHz 8-Core Intel Core i9 CPU and 32GB of memory. 32

2.3 **All 106 SenSet genes with full names**. Marked are genes upregulated in fibroblasts$^{(-)}$ (▲), downregulated in fibroblasts$^{(-)}$, (▼), upregulated in basal$^{(-)}$ cells (△), and downregulated in basal$^{(-)}$ cells (▽) in the HLCA. 44

2.4 **Donor information**. ^{1}IHF: immunohistofluorescence; ^{2}WB: western blot; ^{3}Peritumor tissue; ^{4}COPD Gold II; B: Bleomycin; D: Doxorubicin; I: Irradiation. If parentheses are missing, it is assumed to be (B, D). 48

3.1 Clinical data used to benchmark Truffle. 67

3.2 Time series scRNA-seq datasets used to benchmark Marlene 81

Acronyms

ARI adjusted Rand index.

ASCT+B Anatomical Structures, Cell Types, plus Biomarkers.

ATAC-seq Assay for transposase-accessible chromatin with sequencing.

BSOD Blue Screen of Death.

CCIs Cell-Cell Interactions.

CNNs convolutional neural networks.

CNVs copy number variations.

CODEX co-detection by indexing.

CPM counts per million.

DE differentially expressed.

EM expectation-maximization.

GC guanine-cytosine.

GO Gene Ontology.

GRNs gene regulatory networks.

GSEA gene set enrichment analysis.

HuBMAP Human BioMolecular Atlas Program.

KS Kolmogorov-Smirnov.

mRNA messenger RNA.

NGS next-generation sequencing.

NIH National Institutes of Health.

NK natural killer.

NP nondeterministic polynomial-time.

PBMCs Peripheral Blood Mononuclear Cells.

PCA principal component analysis.

PU Positive-Unlabeled.

RMA Robust Multichip Analysis.

RNA-seq Bulk RNA sequencing.

SASP senescence-associated secretory phenotype.

scRNA-seq Single-cell RNA sequencing.

SenNet The Cellular Senescence Network.

SnCs senescent cells.

SNPs single nucleotide polymorphisms.

snRNA-seq Single-nucleus RNA sequencing.

STEM Short Time-series Expression Miner.

TF transcription factor.

TFs transcription factors.

TMM trimmed mean of M-values.

UI user interface.

UMAP Uniform Manifold Approximation and Projection.

Chapter 1

Introduction

In recent decades, the field of genomics has undergone a seismic transformation. Technological developments enabled the study of the human genome at an unprecedented resolution. In particular, next-generation sequencing (NGS) allows for the rapid and comprehensive sequencing of genetic material from millions of cells at increasingly affordable rates. This has allowed researchers and labs to analyze gene expression patterns in individual cells (also known as single-cell data), identify mutations, and understand cellular heterogeneity with greater precision.

However, the wealth and complexity of genomic data presents new challenges, necessitating the development of computational approaches that process, analyze and model large high-throughput genomic datasets. In addition to size, biological data is inherently noisy making it even more challenging to study and model. Machine learning algorithms, which address both scale and stochasticity, have emerged as a natural solution for these issues. Using these methods researchers can now identify inter- and intra-cellular patterns that enhance our understanding of the genotype-phenotype axis.

Central to the relationship between genotypes and phenotypes is the identification of biomarkers. A **biomarker** or **biological marker** is an indicator of a phenotype or a biological state, such as a cell type, disease, or biological response[1]. Within the context of modern genomics, where data from high-throughput sequencing plays a crucial role, molecular biomarkers such as genes and proteins take center stage. Their significance lies in their ability to help experts annotate cell identities, determine the effect of treatments, monitor cell progression and development, study cell-cell interactions, and many more biological questions.

Several international efforts have focused on characterizing and identifying genetic markers in different tissues, organs, and diseases. Among these, Human BioMolecular Atlas Program (HuBMAP) is a National Institutes of Health (NIH) effort to map the human body at single-cell resolution[2]. An important outcome of this consortium was the curation of the Anatomical Structures, Cell Types, plus Biomarkers (ASCT+B) database which houses information regarding cell type and tissue-specific biomarkers[3]. Another NIH initiative, The Cellular Senescence Network (SenNet), was established to understand mechanisms involved in cellular senescence—a driver of aging characterized by cell cycle arrest. The SenNet Biomarker Working Group is responsible for discovering genetic markers that can be used to identify senescent cells (SnCs)[4].

The existence of such consortia suggests that biomarker discovery is not a trivial task. While certain genes can be activated in specific cell types or as a response to diseases, this does not qualify them as biomarkers. Their activation might stem from factors unrelated to the disease state, emphasizing the need for rigorous validation and context-aware interpretation.

In this thesis, we look at biomarker discovery from several lenses and propose context-aware solutions. We break down this work into two main chapters by differentiating between a static context and a dynamic one. In the static context, we assume a fixed cellular or tissue environment, where genetic markers are identified on the basis of stable characteristics that do not change over time. This approach focuses on identifying markers that reliably distinguish cell types, disease states, or treatment effects under consistent conditions. Conversely, in the dynamic context, we explore how genetic markers evolve in response to temporal changes, such as disease progression, aging, or environmental factors. Here, our goal is to capture markers that provide insights into transitional states and cellular responses, offering a more comprehensive understanding of biological processes in fluctuating conditions.

1.1 All of Molecular Biology[1]

Fitting all of biology into one page is, of course, impossible, but this overview aims to provide a quick primer on key biological processes relevant to this work, with an eye toward making these concepts accessible to readers from engineering backgrounds.

In eukaryotic cells, most genetic information is stored in the nucleus, organized in long, tightly coiled strands of **DNA** (deoxyribonucleic acid). DNA consists of nucleotide building blocks arranged in specific sequences and is divided into discrete units called **chromosomes**. Each chromosome contains numerous sections, or **genes**, which are sequences of nucleotides with defined start and end points. Genes serve as templates for **proteins**, which are crucial molecules that carry out nearly all structural and functional roles in the cell and, collectively, in an organism.

The process of converting genetic information from DNA into functional proteins begins with **transcription**. During transcription, the nucleotide sequence of a gene is copied into messenger RNA (mRNA). This **mRNA** molecule then exits the nucleus and is translated by ribosomes, cellular structures that read the mRNA sequence and assemble amino acids in the correct order to form a polypeptide chain. Once folded into a specific three-dimensional shape, the polypeptide becomes a functional protein.

Not all genes in a cell are “active” or **expressed** at any given time. **Gene expression** refers to the extent to which mRNA molecules are produced from a gene. A high level of mRNA for a specific gene generally correlates with higher levels of its corresponding protein(s), although this relationship is moderated by additional regulatory layers such as noncoding regions and post-translational modifications. The pattern of gene expression is what defines the cell’s identity, type, and function, for instance, whether a cell acts as a muscle cell, a neuron, or an immune cell.

To measure gene expression, scientists often count the number of mRNA molecules transcribed from each gene. Modern sequencing technologies, some of which are described in section 1.4, provide powerful tools for capturing these mRNA levels across thousands of genes simultaneously. This data enables us to compare gene expression patterns across different cell types, conditions, and even over time, giving insight into cellular responses and functions.

Regulation of gene expression is highly complex, involving various molecular players. Among the most influential regulators are **transcription factors** (TFs)—proteins that bind to specific DNA regions to either promote or inhibit the transcription of nearby genes. Transcription factors themselves are encoded by genes, creating a network of feedback and interaction. Understanding these connections, such as which transcription factors regulate which genes, reveals insights into

[1] Inspired by Larry Wasserman’s book titled “All of Statistics”. For a book that studies “all” of molecular biology, see the wonderful “Molecular Biology of the Cell” by Alberts, et al., 7th edition (2022).

how and why certain genes are activated under specific conditions. This network of gene and transcription factor interactions forms the basis of **gene regulatory networks**.

A few other concepts that are relevant are described below.

Epigenetics. Gene expression is also regulated by epigenetic modifications such as DNA methylation and histone modifications. These do not alter the DNA sequence, but can influence which genes are expressed. DNA is packaged in a layer of a substance called **chromatin**. Closed chromatin regions restrict access to DNA, and open regions allow TFs to bind and regulate gene expression. Epigenetic processes control the accessibility of chromatin and therefore play a key role in gene regulation. Technologies such as ATAC-seq (section 1.4) enable the sequencing of open chromatin regions, which can be insightful when studied alongside gene expression. However, this information is typically absent in standard, unimodal gene expression datasets.

Genetic Variation. Mutations, single nucleotide polymorphisms (SNPs), and copy number variations (CNVs) introduce variations in DNA sequences. These variations can influence gene expression and contribute to different phenotypes, further complicating comparison of gene expression across different cells, individuals, or organisms.

Cell-Cell Interactions (CCIs). Cells do not function in isolation; they constantly communicate with one another through molecular signals that shape cellular functions and responses. For instance, upon detecting an infected cell, helper T cells release signals to mobilize other immune cells to eliminate the infected cell. These CCIs are often overlooked by computational methods that analyze cells individually. Recent technologies such as **spatial transcriptomics** and **proteomics** allow researchers to capture both the spatial location of the cells and their expression profiles, facilitating the study of cell signaling and interactions within tissue contexts.

1.2 All of Machine Learning for Molecular Biology

With recent advances in machine learning and deep learning, the computational biology community has embraced these methods to address complex challenges in systems biology and beyond[5,6]. Problems in the field span a wide range of applications, from classification and regression tasks based on gene expression data to the analysis of spatial data and histopathological images, integration of multimodal data across various omics, graph learning of gene or cell networks, and trajectory inference for cellular differentiation.

Sequencing technologies produce an expression vector for each cell but do not inherently identify cell types reliably or cost-effectively. Cell type identity is a function of the genes expressed in the cell, which often leads to the use of **unsupervised approaches** for initial analysis. Most publicly available annotated datasets use cell type inference algorithms or human annotation, both of which are prone to biases. This reliance has made establishing true “golden standards” in the field nearly impossible. Furthermore, due to the continuous nature of gene expression, well-defined cell type classes may not even exist in some cases, with cells often existing on a spectrum between types, further complicating such categorical approaches.

A typical pipeline for analyzing gene expression data involves several key steps. Initially, **quality control** is performed, filtering out cells and genes with low counts, as these are unlikely to contribute meaningfully to the study. **Normalization** techniques, often including log-transformation, convert integer count data into continuous values more suitable for machine learning algorithms. Given the high dimensionality of gene expression data, **dimensionality reduction** approaches are frequently used. The most common strategy uses principal component analysis (PCA), but other methods have also been studied, including deep approaches such as **autoencoders**[7]. The reduced data is then typically **clustered**[8] to group cells by type. Cell types are often determined for each cluster by incorporating prior knowledge, such as cell type-specific marker genes curated from biological literature. Therefore, the reliance on established databases and pathways is essential.

Another approach to analyzing gene expression data is to infer cellular trajectories, assuming that observed variations in gene expression reflect a developmental or differentiation path. **Trajectory inference** algorithms aim to arrange cells in an order that reflects this progression, often assuming smooth changes in gene expression. Basic approaches include linear trajectory methods, while more complex approaches may incorporate tree-like structures, diffusion processes, or differential equations[9]. More recently, **RNA velocity**, a mathematical framework that infers cell trajectories by leveraging the underlying kinetics of gene expression, has emerged as an alternative approach to dynamic modeling[10,11].

In many cases, gene expression data is augmented with other modalities such as chromatin accessibility, DNA methylation, or protein expression. Integrating these datasets allows for a more holistic understanding of cellular processes. Approaches include translating between datasets or identifying shared latent spaces, which are critical for capturing multiple facets of cellular activity. **Multimodal machine learning**, which focuses on integrating diverse data types, has become a vast area of research[12], with many methods specifically tailored to biological data[13].

The complex interactions between cells and genes within an organism are often modeled as graphs. For instance, a directed graph of TFs and genes can model regulatory relationships, with the edge direction indicating the regulatory influence[14]. Similarly, graphs that connect cells can be used to model CCI networks[15]. Therefore, **graph learning** has become a significant area within computational biology.

Machine learning models that specialize in **image analysis**, such as convolutional neural networks (CNNs), are widely used to analyze histopathological images and spatial transcriptomics or proteomics data. These methods have shown success in identifying malignant regions within tissue[16,17] and in understanding extracellular signaling[18].

Other machine learning paradigms have also found applications in biology. For instance, **distribution shifts** across biological datasets are massive and necessitate shift-aware approaches or batch correction algorithms[19]; predicting gene expression responses to drugs or treatments has led to the development of **time-series forecasting** methods[20]; **few-shot learning** addresses challenges with rare cell populations; privacy concerns around sensitive medical data encourage the use of **differential privacy** techniques[21,22]; and **generative models** are leveraged to generate synthetic cell data[23].

Recently, the field has seen the emergence of **foundation models** designed to learn generalizable representations of genes, proteins, and cells, which can be fine-tuned for specific downstream tasks[24,25]. Given the intricate and interconnected nature of biological systems, it is increasingly challenging to manually capture all relevant features in a statistical model. Foundation models offer an appealing alternative, aiming to cover “all” of biology end-to-end. However, **interpretability** remains a crucial consideration, as the stakes in biological and medical contexts are higher than in other domains such as text or video. When a model predicts a drug or treatment outcome, understanding the reasoning behind that prediction is essential.

As biological technologies advance and sequencing methods improve, machine learning and artificial intelligence are increasingly positioned to unlock insights from the genome, promising new breakthroughs in biology and medicine.

1.3 Cellular Senescence

Several chapters in this thesis are dedicated to the analysis of senescent cells (SnCs). Cellular senescence refers to a permanent arrest of cell division triggered by the accumulation of DNA damage and exposure to other stressors such as genotoxic agents, oxidative stress, or nutrient deprivation[26]. The absence of cell division can detrimentally impact tissue regeneration and repair, thereby contributing to various age-related diseases. A growing body of research has explored the impact of clearing SnCs either genetically or by using senolytic drugs which target SnCs and kill these cells. Several experiments conducted on mice have indicated potential benefits, including amelioration of age-related pathologies such as cancer or chronic diseases, and a reduction in mortality[27,28].

Despite this success, a precise characterization of SnC identity remains elusive. Currently, there exists no agreed-upon definition of senescence, and there are no known universally expressed biomarkers for this process. SnCs maintain viability and resist apoptosis (cell death) through mechanisms that are poorly understood[26]. It is hypothesised that this characteristic might partly account for the accumulation of SnCs with advancing age. An alternative hypothesis implicates the senescence-associated secretory phenotype (SASP)—an amalgam of secreted proteins, proinflammatory cytokines, chemokines, and other factors—as a contributing mechanism[29]. There is evidence that SASP can induce senescence in normal cells[30], further exacerbating tissue dysfunction in aged individuals.

Discovery of genetic markers for senescence is a first and crucial step in understanding the mechanisms involved. Not only will such markers facilitate the identification of SnCs, but they will also allow us to track the trajectory of SnC development, their spatial location and the cell-cell interactions that drive senescence. As such, senescence biomarkers could help identify potential targets for therapeutic interventions.

1.4 Experimental Technologies for Characterizing Senescence

The comprehensive study of the diverse array of molecules within the human body requires the adoption of a variety of sequencing methods, subsequently leading to different types of omics data. Here we describe some of the main types of omics data used in this thesis (Fig. 1.1).

1. **Bulk RNA sequencing (RNA-seq)**. RNA-seq technology is an approach for quantifying the average count of mRNA molecules within a sample[31]. While this method yields a low resolution in terms of cell diversity, its cost-effectiveness renders it a favorable option in some cases. For instance, in clinical trials where many samples and individuals need to be profiled, this method becomes particularly advantageous. A single sample is represented as a vector of gene counts.

2. **Single-cell RNA sequencing (scRNA-seq)**. scRNA-seq is similar to RNA-seq, however, instead of obtaining the transcriptome of the entire sample, it quantifies the mRNA for individual cells[32]. During this process, mRNA molecules are selectively captured and uniquely associated with specific cells using barcodes. Following a series of other preparation steps, the data analyst is presented with count data organized in the form of a sparse cell-by-gene matrix. As of present, scRNA-seq technology supports a few million cells. The number of protein-coding genes in the human genome varies from 25k to 30k, with the precise number still being an open question[33].

3. **Spatial transcriptomics / proteomics**. Similar to scRNA-seq, spatial transcriptomics technologies quantify the amount of mRNA, while also providing a two-dimensional spatial

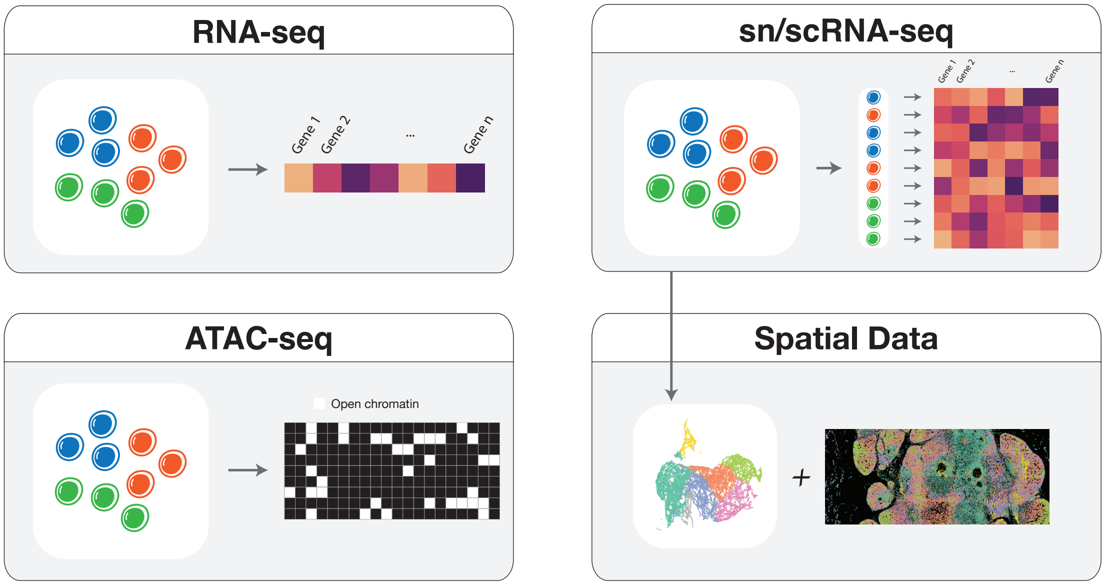


Fig. 1.1: **Illustration of some sequencing technologies used in this thesis.** (Clockwise, starting at top left). 1) Bulk RNA-seq quantifies the average counts of mRNA molecules for the entire sample. 2) sn/scRNA-seq quantifies gene expression at the cellular level. 3) Spatial transcriptomics/proteomics offers spatial locations of cells in addition to their expression profiles. 4) ATAC-seq measures regions in the DNA with chromatin accessibility.

mapping of the cells' location. In addition to understanding cell function, this type of data enables the study of cell-cell interactions and how cells are distributed in various conditions[34]. The data format is similar to scRNA-seq with the addition of several tiles encoding the spatial information. An example is Visium by 10x Genomics™.

Spatially resolved proteomics is a similar approach for mapping the cellular proteome (i.e., protein expression) rather than the transcriptome. An example of such a technology is co-detection by indexing (CODEX)[35].

4. **Assay for transposase-accessible chromatin with sequencing (ATAC-seq)**. Within the nucleus, DNA is surrounded by a layer of chromatin. Regions where chromatin is missing, known as accessible chromatin regions, enable physical access for regulatory elements to bind and regulate gene transcription. This is important in understanding what genes can be activated at a given moment. ATAC-seq is a method for assessing chromatin accessibility across the genome. The data takes the form of a cell-by-peak matrix, where peaks encode these open regions. Typically, the number of peaks ranges in the hundreds of thousands. This data is very challenging to work with as it is extremely sparse.

5. **Single-nucleus RNA sequencing (snRNA-seq)**. This method resembles scRNA-seq, with the difference that it analyzes transcripts inside the nucleus rather than the cell[36]. It is typically used for cells which are hard to isolate, such as cells from frozen tissue.

Time series data can be obtained using either sequencing technique from samples obtained at different time points. All the technologies mentioned thus far destroy the cell upon sequencing, hence, making it impossible to monitor the same cell over time. Newer technologies such as Live-seq allow for the extraction of RNA while preserving cell viability, however, their scalability remains limited at present, and can only sequence a few cells at a time[37].

1.5 Biomarker Discovery

This thesis focuses on the study of biomarkers across various biological contexts. While the scope of biology, as outlines in section 1.2 is broad, much of it ultimately aims to identify disease state, cell types, or biological pathways to advance human health—whether by deepening our understanding of biological systems or by directly contributing to the development of new treatments. Identifying these states or classes requires insights into the genes or markers specifically enriched for each context, making biomarker discovery a critical step. A context-aware approach helps ensure that markers are accurately associated with specific biological states. We divide this thesis into two main settings: steady state and dynamic.

1.5.1 Biomarker Discovery in Steady State Cells

In the static context our main objective is to identify biomarkers that differentiate steady states, such as cell types within a given sample. We design computational methods that specifically target the static context across three learning settings, contingent upon the nature and availability of label information.

1. **Supervised Setting**. This setting constitutes learning from biological data where cell type identities are known. The goal is to recover biomarkers that exhibit specificity for a particular cell type or cell type-tissue pair. We formulate this problem as a linear program that searches for a minimal set of genes that maximally separates classes of interest. This work is presented in section 2.1.

2. **Weakly Supervised Setting**. Unlike the supervised case, where cell identities are explicitly known, in weakly supervised learning, the learner has access to noisy or incomplete label information[38]. Despite these limitations, it can effectively leverage this information to classify previously unseen samples. We argue that this setting is relevant to the study of senescence and aging more broadly, where a consensus on what constitutes a senescent cell is lacking. By constructing weak clusters using genes that have been previously linked to senescence, we construct a Positive-Unlabeled (PU) learning framework that can distinguish SnCs from healthy cells. Finally, by comparing these two cell populations, we identify biomarkers that may potentially play a role in senescence. Details can be found in section 2.2.

3. **Unsupervised Setting**. The field of computational biology suffers from a dearth of ground truth data, primarily stemming from incomplete information regarding cell identities, and the inherent challenges posed by analyzing high-dimensional sparse omics data. Consequently, the assignment of cell types often necessitates unsupervised approaches. The availability of high-quality biomarker databases plays a pivotal role in ensuring the accurate annotation of cell types. In this context, we introduce Cellar, a user interface (UI) designed to streamline the entire process—from data quality control, to clustering and cell annotation. Cellar has been used by several research laboratories associated with HuBMAP and beyond, simplifying data analysis pipelines. Cellar is presented in section 2.3.

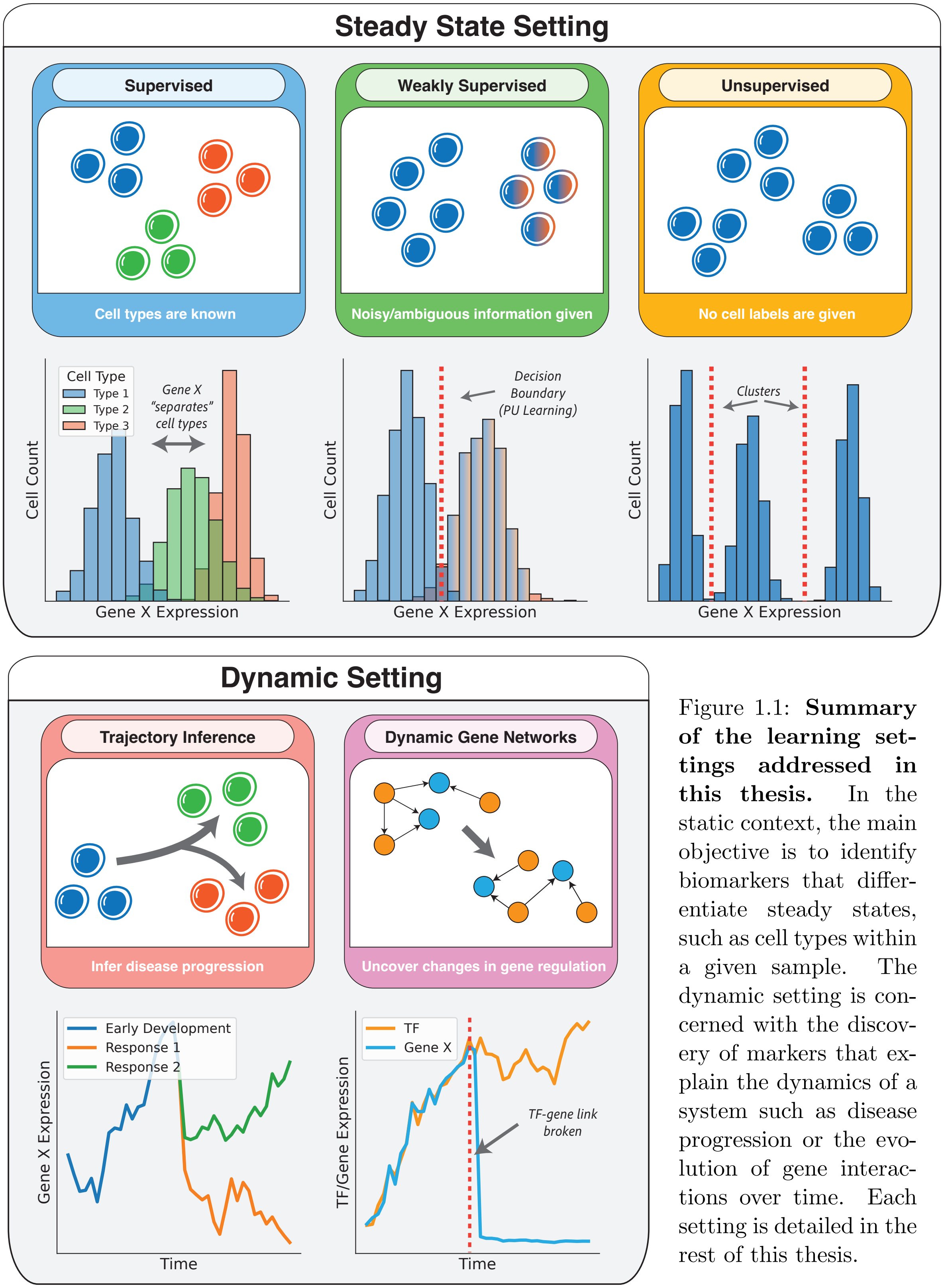


Figure 1.1: **Summary of the learning settings addressed in this thesis.** In the static context, the main objective is to identify biomarkers that differentiate steady states, such as cell types within a given sample. The dynamic setting is concerned with the discovery of markers that explain the dynamics of a system such as disease progression or the evolution of gene interactions over time. Each setting is detailed in the rest of this thesis.

1.5.2 Biomarker Discovery in Dynamically Changing Cells

All biological processes and systems are dynamic. A dynamic cellular context is often captured using measurements over time. The temporal dimension impacts experimental design in two ways. First, it introduces complexity into the algorithms, which now need to accommodate or satisfy timing restrictions. Second, it enables the exploration of questions pertinent to dynamic processes such as disease progression and individual responses to a drug or vaccine. Thereby, the discovery of biomarkers that vary with time and explain the dynamics of the system becomes an important question for personalized medicine and genetic diversity. We tackle dynamics on two fronts.

1. **Endotype-Informed Biomarkers**. Endotypes are subtypes of a disease defined by different pathogenic mechanisms. For instance, at least six asthma endotypes have been identified based on clinical characteristics, biomarkers, and physiological factors[39]. Utilizing clinical time series transcriptomics data, we design an algorithm based on multicommodity-flow[40] that infers distinct disease trajectories from data. Once these potential endotypes are learned, we identify endotype-specific biomarkers which have implications for personalized medicine. We specifically target psoriasis, COVID-19, and Crohn's disease. This is detailed in section 3.1.

2. **Temporal Gene Regulatory Networks**. Biology is complex. Gene regulation is orchestrated by various factors, including regulatory proteins and epigenetic modifications. Therefore, studying genes in isolation may fail to capture the intricate interplay of these regulatory mechanisms. To this end, we propose a meta-learning approach to learn evolving gene regulatory networks from time series data. Our proposed approach uses an evolving self-attention mechanism and is described in section 3.2.

Chapter 2

Biomarker Discovery in Steady State

In the context of steady state systems, our primary focus lies on the static traits of a dataset. Consider, for instance, a scRNA-seq dataset. Here, each cell can be mapped to a specific type, even though over time these cells may undergo differentiation processes. Given that each cell type is characterized by a unique combination of molecular markers, examining gene expression levels at a single time point should, in principle, enable the identification of these cell identities[41].

The converse is also true. Given distinct cell populations, a comparative analysis of their gene expression profiles should enable the discovery of cell type-specific markers.

This reciprocal relationship between marker sets and cell types forms an intriguing feedback loop—one akin to the classic "chicken and egg" paradox. In this context, precision begets precision, and the exploration of unsupervised or weakly supervised methodologies become central.

We explore this duality by beginning with the straightforward case—when all cell identities are known.

2.1 Supervised Setting: Minimal Gene Sets for Accurate Cell Type Identification

Based on: *Hasanaj, E.* et al. *Multiset multicover methods for discriminative marker selection. en.* Cell Rep. Methods *2, 100332 (Nov. 2022)*

Most biomarker discovery methods focus on differentially expressed (DE) genes. In these types of analyses, statistical tests—such as the t-Test or the Wilcoxon rank-sum test[43]—are used to compare gene expression levels between two samples of interest. If the test yields statistically significant results, the gene is categorized as either upregulated or downregulated in the experimental group.

When dealing with datasets containing multiple cell types, a common practice is to compare the cell group of interest against *all* other cells using these two-sided tests. This approach, while widely used, presents a challenge due to the hierarchical nature of cell types. Specifically, two cell types may exhibit greater similarity to each-other than to a third type. For instance, fibroblasts are expected to share more DE genes with myofibroblasts than with B cells, despite their distinct categorization. Consequently, including myofibroblasts within the broader "all" group may not return optimal DE genes.

Furthermore, consider large multi-organ scRNA-seq datasets. In such datasets we may be interested in markers that are specific for both a cell type *and* a tissue (i.e., markers that are uniquely found only in cell types from this tissue). Such markers may be less significant than overall DE genes since they may only distinguish between two similar types, but are still of major importance. An example is given by the Tabula Muris dataset[44], a collection of scRNA-seq profiles of over 100,000 cells from over 20 different organs and tissues in Mus musculus. When analyzing this data, the authors used traditional clustering and DE analysis without considering the issue of cell type/tissue combination. Another example are T cells, which mature in the thymus[45,46]. While T cells later migrate and reside in tissues throughout the body, the identification of T cells that have recently left the thymus (recent thymic emigrants, or RTEs) plays a role in treatment decisions[47]. Similarly, the role of resident and infiltrating immune cell types is still an active area of research for neurodegenerative diseases. A key challenge is the current inability to distinguish the resident central nervous system (CNS) immune cells and the bone marrow-derived immune cells[48]. Better signatures of CNS-specific immune cells and signatures of infiltrating immune cells are needed to understand the immune responses to therapies.

Beyond their specificity for particular cell types, markers are central to other questions in functional genomics. For instance, deconvolution of cell types from bulk data is highly dependent on the ability to select, not just good markers for each individual cell type, but also a set of discriminatory markers between all types[49,50]. In addition, a number of technologies—such as CODEX[35], Cell DIVE™, and Luminex© xMAP©—require the pre-selection of a limited number of markers to profile. A careful selection of a small number of markers that would explain the heterogeneity of the sample is a key criterion for such a selection.

Broadly, marker selection represents a feature selection problem. Feature selection methods can be largely divided into three categories: filters, wrappers, and embedded[51–53]. Wrapper and embedded methods interact with a specific classifier. Wrapper methods select (often in a greedy manner) a subset of the features that lead to a classifier with the highest accuracy. Examples include sequential forward and backward selection methods[54,55]. Embedded methods use the output of the classifier itself, which comes in the form of an explicit ranking of the features or implicitly via a scoring system (e.g., information gain in decision trees[56]). Since these methods are geared towards classification, they may not be applicable to other problems, including deconvolution.

Filter methods, on the other hand, are not tied to a specific classifier. For example, scGeneFit[57] selects those genes which maintain a separation of the different cell types similar to that of the original space. This method supports both a flat partition or a hierarchy of labels (e.g., major cell types and subtypes). RankCorr[58] works in a one-vs-all fashion and selects markers for a fixed cell type by performing a rank transformation. Another algorithm, Relief[59], and its extension ReliefF[60], penalize features that cannot distinguish a given instance from its negative (having a different label) neighbors, while assigning high scores to features that take similar values among instances from the same class. Minimum-redundancy-maximum-relevance (mRMR) selects features that are relevant to the target class but are not similar to each-other[61]. CIBERSORT[49] and a number of prior methods[50,62,63] analyzed a signature matrix of DE genes to identify submatrices with low condition number for use in deconvolution of bulk mixtures. Thus, while these methods can successfully select discriminative features when the overlap between sets is small, the ability of such methods to select markers that discriminate all pairs of phenotypes has not been extensively studied.

In this paper, we explore the problem of determining a *global* set of biomarkers. These represent features that collectively distinguish between higher context phenotypes. We begin with a *phenotype* × *feature*, binary or real score matrix $\mathbf{M}$, whose (i, s) entry represents the relevance of feature s (e.g., average gene expression) for phenotype i. We formulate the task as a combinatorial optimization problem where the goal is to identify the smallest set of features such that for every phenotypic pair (i, j) there exists a set of features that can be used to "distinguish" between i and j. We term this problem **Phenotype Cover** (PC). We show that PC is equivalent to multiset multicover which is NP-complete[64], and propose two algorithms that can approximate it in polynomial time.

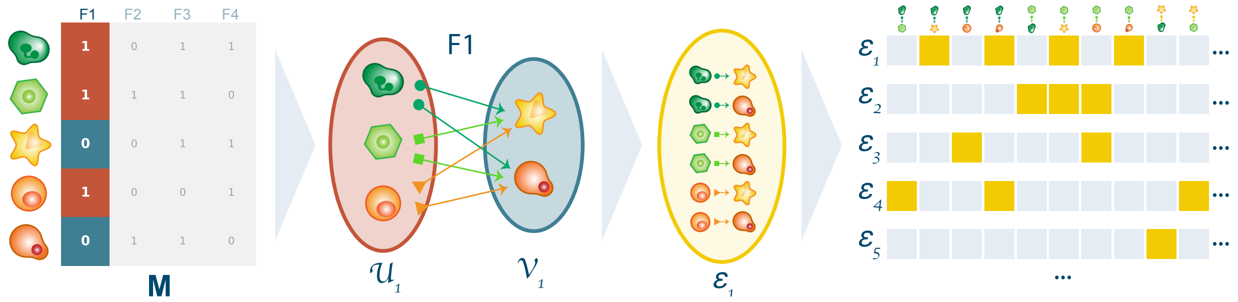


Fig. 2.1: **Illustration of Phenotype Cover and the multiset multicover approach.** Given a binary score matrix (left), each feature induces a bipartite graph between classes (center left). Edges in this graph form a set $\mathcal{E}_s$. Multiset multicover is then performed on the collection of $\mathcal{E}_s$ to select a small number of features which "distinguish" all phenotypic pairs (at least K times). The idea can be naturally extended to non-binary score matrices by assigning a multiplicity to each element $e_{ij} \in \mathcal{E}_s$ (See Methods).

The first is based on the extended greedy algorithm to set cover (G-PC)[65], and the second is based on the cross-entropy method (CEM-PC)[66,67]. By analyzing several marker selection problems, we show that the greedy algorithm outperforms competitors across a variety of tasks. We also analyze some of the specific markers selected by the method and discuss their ability to distinguish between similar cell types.

2.1.1 Results

We developed methods to select discriminative features from a large set of (potentially overlapping) signatures. The goal of the features we select is to enable the separation of the different components in the set. This can either be for a supervised learning (for example, classification) or for other learning approaches such as deconvolution or dimensionality reduction. Our method takes as input a signature or score matrix $\mathbf{M}$ which is used to estimate the importance of a feature for a phenotype of interest. Features are then selected by reformulating the problem as a multiset multicover instance where the goal is to select features such that every phenotypic pair is covered at least K times, for some positive K (Fig. 2.1). We developed two solutions to the multiset multicover problem: the first is based on a greedy approach (G-PC), and the second based on the cross-entropy method (CEM-PC). See Methods for details.

We tested G-PC and CEM-PC, and compared them to eight prior methods: scGeneFit[57], decision trees[68], top differentially expressed genes (TopDE), RankCorr[58], ReliefF[60], mRMR[61], ANOVA F-values, and mutual information[69,70]. We used three scRNA-seq datasets (Table 2.1). We vary the coverage factor K from 1 to 20 for the IPF dataset, from 1 to 40 for MC, and from 1 to 9 for HCA. For all baselines but TopDE and RankCorr, we select a number of features that matches the solution size returned by G-PC. For TopDE, we take the union of the top k differentially expressed genes for each phenotype (k varying from 1 to < 10). For RankCorr, we tuned the hyperparameters until a similar number of features was returned. Finally, for CEM-PC, all the features with a probability score greater than 0.98 after convergence were chosen (Supplementary Algorithm 3). We compare all methods in terms of phenotype classification performance, deconvolution of bulk mixtures, and feature stability. We also validate the features selected by G-PC by performing gene set enrichment analysis (GSEA) and comparing with known markers in the literature.

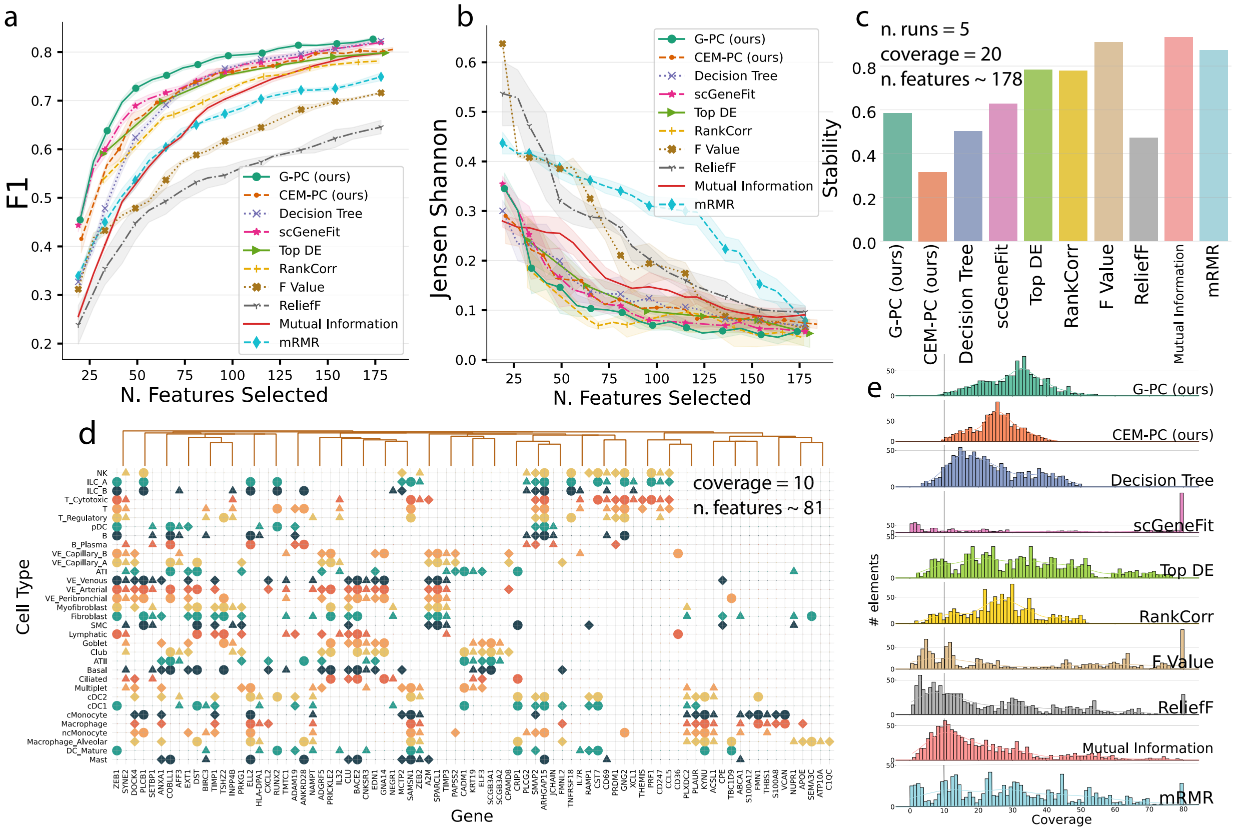


Fig. 2.2: **Comparison of feature selection methods for the IPF dataset.** Performance scores for (a) and (b) were averaged across five different random train and test splits. Standard deviation is shown as a shaded region. (**a**) Performance of a logistic regression model trained on the selected features. G-PC achieves the highest F1 score across all coverage factors, followed by CEM-PC and decision trees. (**b**) Jensen-Shannon divergence (lower is better) between CIBERSORT-predicted mixture proportions and the ground truth. (**c**) Stability scores for all eight methods over 5 runs. Sequential methods like G-PC, decision trees and CEM-PC suffer slightly in stability compared to other, more global methods. Nonetheless, G-PC shares about 70% of the features across runs. (**d**) Biomarker membership matrix. Gene s (columns) is assigned to cell type i (rows) if there exists another cell type j such that $\mathbf{M}_{i,s} - \mathbf{M}_{j,s} \geq 1$ (see Section 2.4). Rows and columns were ordered based on hierarchical clustering. Colors and shapes only serve visibility. (**e**) For every phenotypic pair, we compute the "coverage" (i.e., the score difference between the two phenotypes) provided by the selected gene set. A histogram of these coverage factors corresponding to a coverage of 10 is shown for each method. As can be seen, for G-PC and CEM-PC which optimize for coverage, each element is covered at least 10 times. Other methods provide high coverage for some elements but miss out on others.

Datasets	Genes	High Var.	Cells	Tissues	Cell Types	Ref
Idiopathic Pulmonary Fibrosis (IPF)	4,443	Yes	96,301	1 (lung)	33	[71]
Mouse Cortex (MC)	20,006	No	3,005	1 (cortex)	7	[72]
Human Cell Atlas (HCA)	2,968	Yes	84,363	15	7	[73]

Table 2.1: **Three datasets were used in the PhenotypeCover study**. For HCA, we consider a combination of tissues and cell types (85). For IPF only healthy samples were kept. Endothelial-mural and astrocyte-ependymal pairs of cells were grouped for MC.

Classification Performance

We first test the ability of a classifier to predict the correct phenotype given only a subset of the features. For each method we select a feature set $\mathcal{S}$ using a subset of the data, train a Logistic Regression model on the same subset, and evaluate performance on left out data. G-PC exhibits strong performance on the IPF and MC datasets across a wide range of coverage factors. For example, when 42 genes are selected on the IPF data, G-PC obtains an F1 score of 0.70, followed by scGeneFit (0.65) and CEM-PC (0.61) (Fig. 2.2a). On the MC data (Fig. A.1a), G-PC again performs best when 30-140 genes are selected (F1 $\approx$ 0.94-0.95). mRMR also performs well on this data except when the number of genes selected is small (< 30). Decision trees, on the other hand, do not improve in performance when more than 30 genes are selected (F1 $\approx$ 0.92).

These two datasets are obtained from a single tissue. We thus next tested the ability of PC to differentiate between the same cell types across multiple tissues. For this, we used all tissue/cell type combinations present in the HCA dataset. Decision trees outperform other methods on this classification task (Fig. A.2a). G-PC is the second best method when more than 100 genes are selected, while scGeneFit is the second best when less than 100 genes are selected. scGeneFit, however, does not improve in performance when more than 100 genes are selected. At 235 genes, decision trees converge at 0.70, while G-PC and mutual information reach an F1 of 0.68.

We note that scGeneFit can take the hierarchy of labels into account and the authors describe improved performance when cell subtypes are considered in the MC dataset. For a fair comparison, we ran three different variants of scGeneFit that take advantage of this hierarchical structure, and evaluated performance by using a nearest centroid classifier fit on the entire data. All the hyperparameters we used were identical to those provided by the authors. While G-PC does not use cell subtype information, it still outperforms all three variants across a different number of markers (Supplement).

We also tested an additional classifier (k nearest neighbors) and observed very similar results to those obtained with Logistic Regression (Fig. A.3). Finally, we tested the impact of batch effects by using two pancreas datasets[74,75], and observed that our method, G-PC along with TopDE are the most robust to batch effects (Fig. A.4).

Deconvolution

Inferring cell type proportions from bulk transcriptomics data is an important task in understanding composition of tumors and other tissues. Many methods have been developed to perform deconvolution of bulk mixtures[50,62,63,76]. Deconvolution typically requires solving a linear equation of the form $m = Sp$, where m is a given mixture vector, S is a signature matrix containing cell type-specific expression signatures (known), and p is the unknown class proportion vector. One widely used method for deconvolution is CIBERSORT[49] which uses ν-Support Vector Regression

(ν-SVR). CIBERSORT constructs the signature matrix S by considering the top k DE genes for every cell type subset (which leads to the exact same selection as the TopDE baseline we consider in this study). Next, CIBERSORT selects the k that leads to a signature matrix S with the lowest condition number. Finally, ν-SVR is fit on the data and the regression coefficients in the solution are used to estimate p.

To test the usefulness of the features selected by our method for deconvolution, we constructed pseudo-bulk mixtures using the IPF, MC, and HCA datasets by averaging expression levels across all single cells in the test sets. The signature matrix S was constructed with features selected from the training set and deconvolution via ν-SVR was then applied to the pseudo-bulk mixtures. As recommended by the authors, we initialize three linear ν-SVR instances with $\nu \in \{0.25, 0.5, 0.75\}$ and save the model that achieves the lowest root-mean-square error between the deconvolution result Sp and m. We compute the Jensen-Shannon (JS) divergence[77] between the predicted mixture p and the ground truth. G-PC performs well on the IPF data with RankCorr doing better only when 50-80 genes are selected (Fig. 2.2b). For example, when 163 genes are selected, G-PC achieves an average JS$= 0.045$, followed by RankCorr (0.056) and scGeneFit (0.062). For the MC dataset, G-PC is also the top ranking method, though TopDE and RankCorr also accurately resolve mixture proportions (Fig. A.1b). All three methods obtain a JS score of less than ≈ 0.025 across all K. CEM-PC performs well on some instances for both datasets, however, the results are unstable and vary between runs. None of the methods clearly outperforms all others on the HCA dataset (Fig. A.2b). These results demonstrate the challenges of trying to distinguish cell types across tissues.

Finally, we also tested another version of deconvolution which uses linear least squares (LLS) as the target. We observed that for this method G-PC performs no worse than other methods on IPF and MC (Supplement).

Stability Analysis

The focus of the comparison so far has been on accuracy. However, other considerations are also important, especially when selecting features that will be used across different platforms and potentially modalities. One such important issue is feature stability[78]. The stability index measures the average size of the overlap divided by the size of the union for all pairs of feature sets (Methods). To test the stability of different methods we randomly sample half the data and compute the stability index for the features selected by each method over 5 runs. Stability scores are shown in Fig. 2.2c, and the supplementary material. G-PC is more stable than decision trees for IPF and MC. However, due to their greedy sequential nature, both G-PC and decision trees are less stable then more global methods such as ReliefF and F values. Nonetheless, G-PC uses from $60\% - 70\%$ of the same genes across all runs. Perhaps not surprisingly, due to its random sampling nature, CEM-PC is the least stable method.

Biomarker Validation

To validate the set of biomarkers $\mathcal{S}$ selected by G-PC and decision trees, we performed enrichment analysis for the HCA dataset. We fix a coverage of 8, and for every phenotype i, we select from the solution $\mathcal{S}$ all those genes s for which there exists some phenotype j satisfying $\mathbf{M}_{i,s} \geq \mathbf{M}_{j,s} + 1$. We consider each of these sets as a biomarker set for the given phenotype for both G-PC (Fig. 2.2d, Supplement) and decision trees.

We next performed gene set enrichment analysis (GSEA)[80] using the HuBMAP ASCT+B marker set[79] to determine if the selected markers sets for a specific cell type are enriched for pathways associated with these cell types. We test the ability of G-PC and decision trees to identify the

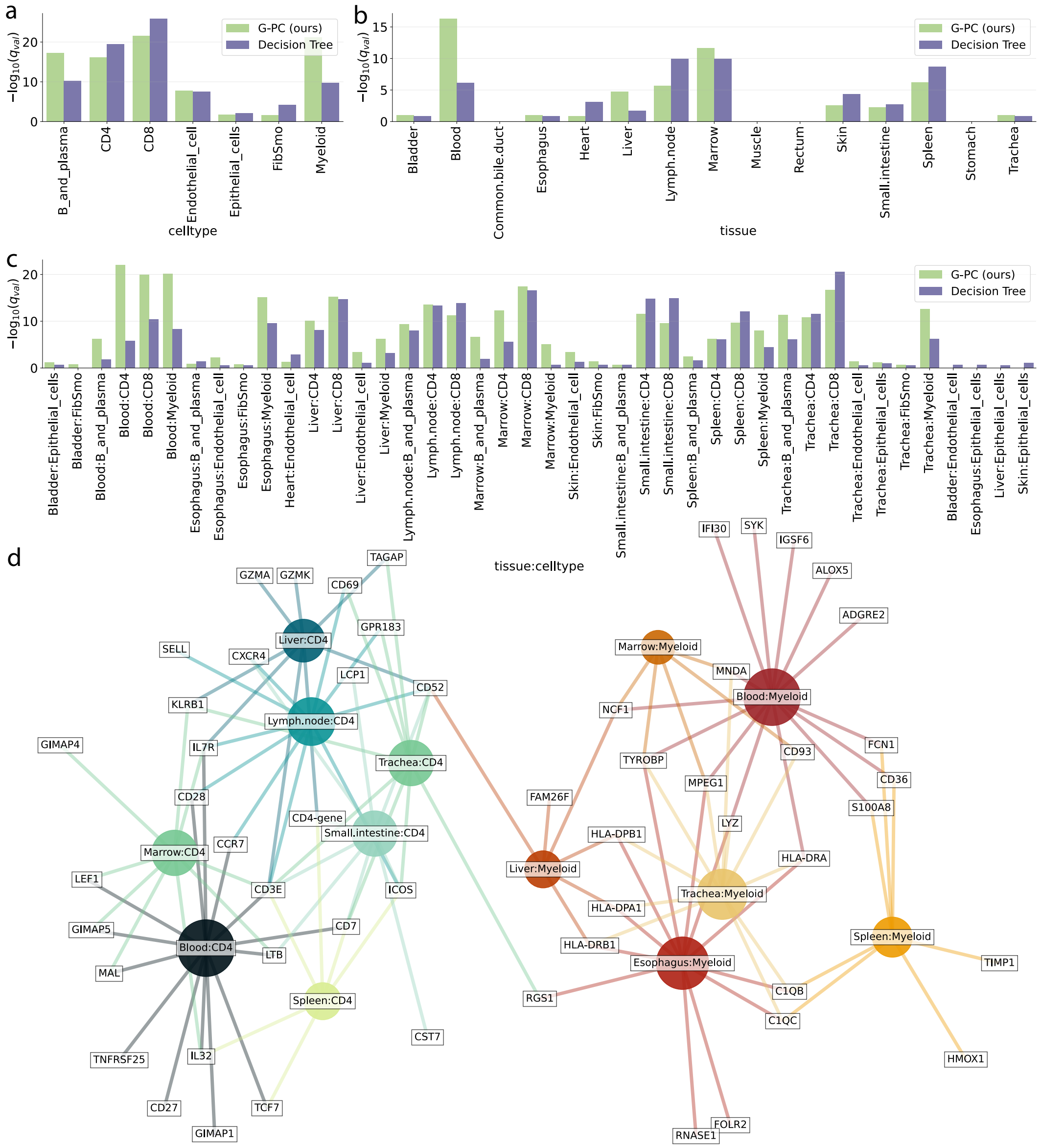


Fig. 2.3: **GSEA q-values for HCA** We select markers that provide coverage for each cell type for both G-PC and decision trees and perform gene set enrichment analysis (GSEA) using the HuBMAP ASCT+B gene set[79]. We first record q-values for the top entry which contains the correct cell type or the correct tissue independently. When comparing the ability of each method to assign the correct cell type, G-PC obtains a lower q-value (i.e., higher $-\log(q\text{-value})$) in 42% (3/7) of the cases (**a**). A similar analysis shows that G-PC obtains lower q-values in 54% (6/11) of the tissues (**b**). We did not find markers for four tissues in the gene set (common bile duct, muscle, rectum, stomach). (**c**) When tested for the ability to identify both the correct tissue *and* the correct cell type, G-PC obtained lower q-values in 71% (30/42) of the cases. The remaining tissue/cell type pairs (33) either belonged to a tissue which was not present in the marker set, or were not identified by either method. (**d**) Connected by an edge are known markers for CD4 and myeloid cells that were assigned to the correct tissue/cell type pair by G-PC. Some markers are assigned to multiple cell types (multiple outgoing edges), while others are pair specific.

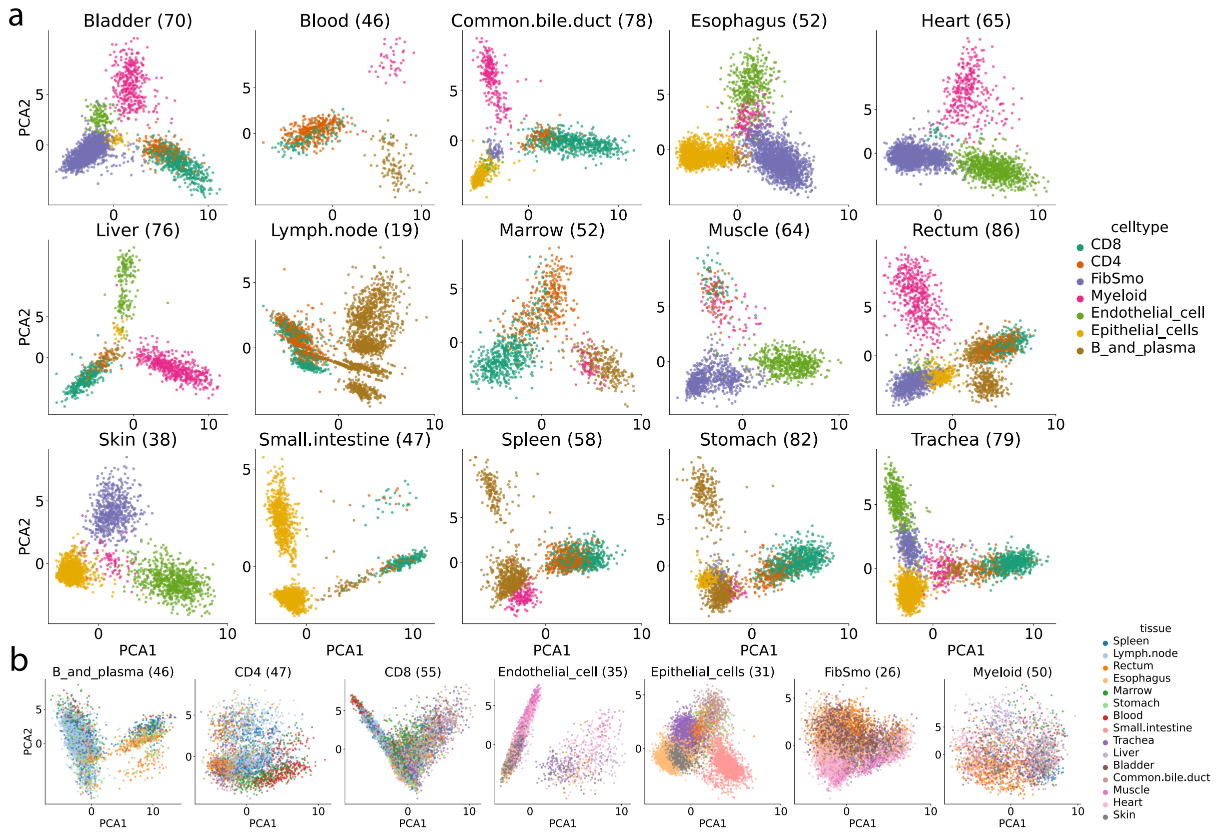


Fig. 2.4: **PCA plots of classes in HCA** A total of 121 markers were selected via G-PC (coverage= 5). For every tissue (**a**) and cell type (**b**), top two principal components of the markers providing coverage (≥ 1) for that phenotype (number in parenthesis) are plotted. There is visible separation between classes.

correct 1) tissue, 2) cell type, and 3) tissue/cell type combination. G-PC obtains lower q-values for 42% (3/7) of the cell types and 54% (6/11) of the tissues (Fig. 2.3a, b). No markers were found for four tissues. When tested against the correct tissue *and* cell type pair, G-PC obtained lower q-values for 71% (30/42) of the pairs (Fig. 2.3c). The remaining combinations (33) either belonged to a tissue which was not present in the marker set, or were not identified by either method. Some known markers assigned correctly by G-PC are shown in Fig. 2.3d. Esophagus and trachea tissues were mapped to *respiratory system* in the ASCT+B set. The top two principal components of the markers that provide coverage for a given tissue or cell type show visible separation between different classes (Fig. 2.4).

Due to the limitations of the marker set we are using, only 20 cell types could be identified for IPF. Among these, G-PC obtains lower q-values for 12 (60%) (Fig. 2.3). We observe good agreement between genes selected using our greedy procedures and genes known to be involved in specific cell types. For example, G-PC correctly assigns KRT19, ADGRF5 to Type I and Type II epithelial cells (ATI and ATII,[81–83]). CD69 is assigned to both B and T cells[84,85], COBLL1 is assigned to B cells[86], JCHAIN to B and B plasma cells[87], CXCL2 to macrophages[88], CCL5, PRF1, CD247 to natural killer cells[89–91]. See Fig. 2.2d for a larger list of identified markers.

In addition to selecting known cell type markers, G-PC is also able to select markers that distinguish between similar cell types. For example, it assigns CXCL2 to ATII and not to ATI[92],

Data	G-PC	CEM-PC	DT	scGF	DE	RC	FVal	ReliefF	MI	mRMR
IPF	**1.3**	55	304	102	41	90	1.5	388	1039	980
MC	0.17	227	5	39	2.7	3	**0.15**	11	116	139
HCA	1.25	88	50	52	30	95	**0.8**	340	790	310

Table 2.2: **Runtimes for all methods across all three datasets** (seconds). Names were abbreviated. 178 features were selected for IPF, 66 for MC, and 121 for HCA. Our C++ implementation of G-PC takes less than 2 seconds for all three datasets, making it the fastest along with F value computation. Performance tests are conducted on a machine with a 2.3 GHz 8-Core Intel Core i9 CPU and 32GB of memory.

CD69 and AFF3 to B cells and not plasma cells[84,85,93]. Another example is A2M and CST7 which are assigned to cytotoxic T cells[94] whereas NAMPT, TNFRSF18 are assigned to regulatory T cells[95].

2.1.2 Materials and Methods

Notation

Let $\mathbf{M} \in \mathbb{R}_{\geq 0}^{P \times F}$ represent a score matrix. We denote by P the number of phenotypes (e.g., cell types) and by F the number of features (e.g., genes). We assume that $\mathbf{M}$ is non-negative, with higher values representing stronger relationships between phenotypes and features. In this paper, we use scRNA-seq read count data denoted by $\mathbf{X} \in \mathbb{R}_{\geq 0}^{N \times G}$. Here, N denotes the number of cells and G denotes the number of genes. Given a known vector $\mathbf{y}$ of length N representing class labels, we derive a matrix $\mathbf{M}$ for the scRNA-Seq data $\mathbf{X}$ by averaging expression values of cells having the same class label. Thus, for such data $P = \{\text{number of distinct classes}\}$ and $F = G$. We denote by $[n]$ the set $\{1, 2, \ldots, n\}$ and for any given subset $\mathcal{S} \subset [G]$, let $\mathbf{X}_{:,\mathcal{S}}$ denote the submatrix formed by selecting the columns in $\mathcal{S}$ from $\mathbf{X}$. Finally, let $(x)^+ = \max\{x, 0\}$.

Problem Formulation and Complexity

Phenotype Cover (PC) Given a score (signature) matrix $\mathbf{M} \in \mathbb{R}_{\geq 0}^{P \times F}$, find a subset $\mathcal{S} \subset [F]$ of minimal cardinality, such that for every $i, j \in [P]$ with $i \neq j$, and some fixed positive K, the following holds

$$\sum_{s \in \mathcal{S}} (\mathbf{M}_{i,s} - \mathbf{M}_{j,s})^+ \geq K. \tag{2.1}$$

PC is asking for a small subset of features such that for *any* given ordered pair of phenotypes (i, j), one can find enough features which collectively distinguish i from j by a factor of at least K. This problem allows the selection of a gene which could cover several phenotypic pairs, e.g., multiple cell subtypes vs another major cell type, but also demands sufficient coverage between subtypes themselves. The straightforward solution of iterating over all possible feature subsets satisfying (2.1) and selecting the one with the smallest cardinality suffers from an exponential complexity in the number of subsets considered. In fact, PC is equivalent to multiset multicover which is NP-complete[65].

To establish this equivalence, it may help to first consider a simplified version of the problem where we restrict $\mathbf{M}$ to be binary and $K = 1$; call this problem PC-B. In this case, we require a small subset of features $\mathcal{S}$, such that for any two phenotypes $i \neq j$ there exists some index $s \in \mathcal{S}$

where $\mathbf{M}_{i,s} - \mathbf{M}_{j,s} = 1$. Note that in this simplified version, every feature s induces a bipartite graph $\mathbb{G}_s = (\mathcal{U}_s, \mathcal{V}_s, \mathcal{E}_s)$, where

$$\mathcal{U}_s := \{i \mid \mathbf{M}_{i,s} = 1\}, \qquad \mathcal{V}_s := \{j \mid \mathbf{M}_{j,s} = 0\} = [P] \setminus \mathcal{U}_s.$$

Every edge $e \in \mathcal{E}_s$ corresponds to an ordered pair of phenotypes (Fig. 2.1).

Now, given the collection of sets $\mathcal{E} := \{\mathcal{E}_s \mid s \in [F]\}$, set cover asks to find the smallest subset $\mathcal{E}_{sol} \subset \mathcal{E}$ such that for every element $e \in \bigcup \mathcal{E}_s$, there exists a set in $\mathcal{E}_{sol}$ which contains e. It is easy to see that the features corresponding to $\mathcal{E}_{sol}$ are the solution to PC-B.

So far we only considered a binary score matrix. However, a solution to the binary problem can be naturally extended to solve non binary scoring matrices by assigning multiplicities to the elements of $\mathcal{E}_s$. To every $e = (i, j)$ we assign the multiplicity $(\mathbf{M}_{i,s} - \mathbf{M}_{j,s})^+$ and view $\mathcal{E}_s$ as a *multiset*. Note that since we are working with real numbers, we need to round the multiplicities to integers. Higher precision can be easily obtained by first scaling both $\mathbf{M}$ and K by some scalar c and performing the rounding after. Finally, the requirement for $K = 1$ can also be relaxed by solving for a *multicover*, where we require each element to be contained at least K times in $\bigcup \mathcal{E}_{sol}$ (counting multiplicities).

Approximating a Solution to Phenotype Cover

Given the NP-Completeness of PC, we present two greedy solutions that run in polynomial time.

Greedy Phenotype Cover (G-PC) First, we consider the well-known greedy approach to solving set cover that iteratively picks the set which covers the greatest number of elements not covered yet[96,97]. The algorithm can be trivially extended to solve multiset multicover[65]. The full algorithm is presented in the Supplement (Algorithms 1, 2). Every time we select a set, we need to correct the multiplicities of all the remaining $O(F)$ sets, each of which may contain up to $O(P^2)$ elements (all phenotypic pairs). Therefore, if we denote the solution size by k, the run-time complexity of G-PC is $O(kP^2F)$. In practice, P is small and $k \ll F$, therefore, the method is almost linear in the number of features considered. The approximation accuracy for this solution was previously analyzed and it was shown that the greedy algorithm for multiset multicover is upper bounded by a factor of H_m increase in the solution size, where $H_m = 1 + \frac{1}{2} + \cdots + \frac{1}{m} \leq \log m + 1$ and m is the cardinality of the largest multiset[65].

Cross-Entropy Method Phenotype Cover (CEM-PC) In addition to the greedy multiset multicover approach, we developed a new method based on cross-entropy (CEM)[66]. CEM was originally used to estimate probabilities of rare events and it was later extended to solve combinatorial problems[98]. Roughly, CEM consists of two steps: 1) generate a random sample based on a specific distribution, and 2) update distribution parameters such that "high-scoring" samples are more likely to be produced in the next iteration. This two-step procedure is repeated until convergence, or until a maximal number of iterations is reached. The final parameters determine the solution to the combinatorial problem (in our case, selecting features whose probability is greater than some threshold). For a more detailed analysis of CEM, the reader may refer to the excellent tutorial of De Boer *et al.*[67].

We present a variant of CEM for solving set cover by introducing a scoring function that encourages high coverage, but penalizes a large number of features (Supplementary Algorithm 3). The run-time complexity of CEM-PC depends on the maximum number of iterations I, the number of random samples per iteration R_s, and the complexity of the scoring function (in this case, the smallest coverage attained per random sample). This leads to a total run-time complexity of $O(IR_sP^2F)$. In this paper, we use $I = 500$ and $R_s = 1000$. In practice, convergence is attained in fewer iterations.

Baselines

As mentioned above, several prior methods have been developed for marker and feature selection. We thus compared our method against several baselines on traditional supervised learning tasks, ability to construct signature matrices for deconvolution of bulk mixtures, and feature stability. Specifically, we compare our method to scGeneFit and RankCorr which were used for discriminative marker selection. We use the implementations provided by the authors of each method. For scGeneFit, we used a redundancy of 0.1 and kept the remaining parameters at defaults. We compare against an embedded method that uses decision trees with the Gini Index criterion to rank features. Note that here we use decision trees as a feature selection method only and not as a classifier. The performance of decision trees as a classifier was worse than that of Logistic Regression using the same features, hence, we excluded these results from the manuscript. We also compare against several other filter methods. We consider the union of the top differentially expressed genes per phenotype as determined by Welch's t-test[99] (TopDE). We compare against ReliefF which uses nearest neighbors' information to update feature weights. Since computing exact neighbors is slow for the single cell data we are using, we developed a variant of ReliefF that uses approximate neighbors based on the faiss package[100]. We compute 30 neighbors per sample. ANOVA F-values and mutual information between gene expression and phenotype are also computed using the popular package scikit-learn. Finally, we compare against minimum-redundancy-maximum-relevance (mRMR). For mRMR, we use the open-source Python package mrmr (`https://github.com/smazzanti/mrmr`) which measures relevance via the F-value and measures redundancy via Pearson's correlation. For all the baselines but TopDE and RankCorr, we take the top k scoring features, where k equals the size of the solution returned by G-PC.

Metrics

To compare the performance of Logistic Regression classifiers, we use the macro-average F1 score. This score equally weighs the F1 score of each class, which is desirable as we are interested in finding markers for all phenotypes, regardless of any class imbalance in the data. For a single class p, the F1 score is the harmonic mean between precision and recall

$$\mathrm{F1}_p = \frac{2}{\frac{1}{\mathrm{Precision}_p} + \frac{1}{\mathrm{Recall}_p}} = 2\frac{\mathrm{Precision}_p \cdot \mathrm{Recall}_p}{\mathrm{Precision}_p + \mathrm{Recall}_p}.$$

The macro-average F1 score is simply the unweighted mean of per-class F1 scores

$$\mathrm{F1}_{\mathrm{macro}} = \frac{1}{P}\sum_{p=1}^{P} \mathrm{F1}_p.$$

To evaluate the deconvolution performance, we use the Jensen-Shannon divergence[77] which is a symmetric measure between two probability distributions. Given two discrete probability distributions P and Q, the Kullback-Leibler divergence[101] is given by

$$\mathrm{KL}(P \parallel Q) = \sum_{x \in \mathcal{X}} P(x) \log\left(\frac{P(x)}{Q(x)}\right)$$

where $\mathcal{X}$ is a probability space. Letting $M = \frac{1}{2}(P+Q)$, the Jensen-Shannon divergence is

$$\mathrm{JS}(P \parallel Q) = \frac{1}{2}\mathrm{KL}(P \parallel M) + \frac{1}{2}\mathrm{KL}(Q \parallel M).$$

Feature stability computes the average size of the overlap divided by the size of the union for all pairs of feature sets. More precisely, given a collection of feature sets $\mathcal{C} = \{S_1, \ldots, S_k\}$, stability is given by

$$s = \frac{2}{k(k-1)} \sum_{i=1}^{k} \sum_{j>i}^{k} \frac{|S_i \cap S_j|}{|S_i \cup S_j|}.$$

Finally, we performed gene set enrichment analysis (GSEA) using the Python package GSEAPY (`https://gseapy.readthedocs.io`) and the Enrichr API[102]. We used the `HuBMAP_ASCTplusB_augmented_2022` gene set[79].

Datasets and Preprocessing

We use three public scRNA-seq datasets to validate our method (Table 2.1). For all three datasets we remove classes with less than 50 cells. This leads to 75 tissue/cell type pairs for HCA. We also filter for genes expressed in at least 10 cells, and for runtime efficiency purposes, we only consider highly variable genes for IPF and HCA for all methods. Also, scGeneFit was slow for MC, so we considered only highly variable genes for MC when running this method. Each dataset is normalized using Scanpy[103] so that the total counts for all cells are equal. The data is then $\log(x+1)$ transformed and each feature scaled to unit variance and zero mean. scGeneFit performed very poorly when the data was scaled, hence, for a fair comparison we skipped the scaling step when running scGeneFit. Log-transforming and scaling the data had a positive effect on the F1 score for all the other methods. We show these results for the MC dataset in Fig. A.1. On the other hand, deconvolution via CIBERSORT works best if the data is in linear space as recommended by the authors, hence, we did not log the data during deconvolution. Feature selection, however, is applied to logged data. We split all datasets into a train and test set of equal size in a stratified fashion. To obtain a signature matrix $\mathbf{M}$ for G-PC and CEM-PC, we average expression values for every phenotype. While it is true that this operation summarizes the data and leads to information loss, we note that our goal is not reconstruction or dimensionality reduction but rather marker selection. We argue that for such a task the individual cell based expression is less important since we are looking for markers that are generally observed across most or all cells. Furthermore, commonly used DE tests such as t-test also rely on a small set of sufficient statistics. Regarding the choice of K, in this paper we test the performance of our methods across multiple values of K. In practice, a single value for K could be obtained in a cross-validation fashion.

Code Availability

We implemented a general purpose package for running the greedy multiset multicover algorithm in C++ and expose it to Python. The code can be found on our GitHub repository at `https://github.com/euxhenh/multiset-multicover`. The G-PC and CEM-PC algorithms for feature selection can be found at `https://github.com/euxhenh/phenotype-cover`. Installation instructions are available in each repository. The code for running experiments in this paper is available from `https://github.com/euxhenh/phenotype-cover-experiments`.

2.1.3 Discussion

Selection and use of markers is a common step in many analysis pipelines. Most recently this topic received increased attention due to the large number of new cell types that have been identified and characterized using scRNA-seq data[104–107]. To date, such selection was mainly based on methods that focused on each cell type separately and did not consider the relationship between markers selected for different types. Such methods can select overlapping marker sets for different cell types making it hard to discriminate between similar cell types. This is especially important for large datasets where multiple cell types in multiple tissues are being profiled[44,73].

To address this issue and to improve the ability to select a discriminating set of markers, we defined a new optimization function for biomarker selection that takes the overlap into account. Specifically, we defined the phenotype cover problem that aims to optimize the accuracy of identifying different sets when using the selected markers. We presented two heuristic filter methods since these lead to solutions that can be used in several different analyses pipelines including classification, deconvolution, experimental design and more. The first is based on a greedy approximation algorithm (G-PC) and the second is based on the cross-entropy method (CEM-PC).

We evaluated these methods and compared them to prior methods developed for marker selection using several high throughput scRNA-seq datasets. Our analysis indicates that G-PC assigns equal importance to all different phenotypes in the data and is not affected by class imbalance as shown by the F1 score. Other methods tend to select features that discriminate only dominating classes. Furthermore, G-PC can be used with signature matrices rather than direct expression measurements. In such cases there is only a single score for all phenotype/gene pairs which makes using other methods difficult. This allows G-PC to construct signature matrices for deconvolution which leads to an accurate estimation of cell type proportions from bulk mixtures. While G-PC is slightly less stable than some other methods, it nonetheless retains the majority of the features ($\sim 70\%$) across runs. Decision trees outperform G-PC with regard to the F1 score in one of the datasets we analyzed (HCA). However, even for HCA, G-PC seems to obtain a more accurate list of cell type markers based on enrichment analysis. We note that that our method is best suited for datasets that require detailed annotations which usually mean that several cell types partially overlap in their markers. In contrast, for large datasets where the focus is on more coarser cell types we see less advantage compared to standard marker selection methods. Finally, we provide a C++ implementation of G-PC with Python bindings which makes it the fastest method we tested (Table 2.2). Speed is an important consideration when working with large scRNA-seq datasets.

We observed that CEM-PC sometimes selects a smaller set of genes that achieves the same coverage as G-PC. However, due to its random sampling nature, CEM-PC is very unstable and can lead to a completely different set of features across runs.

While G-PC worked well for the data analyzed in this paper, it definitely does not provide an optimal solution. It is interesting to see if other approximation algorithms that optimize for coverage will lead to better results when tested on biological data.

2.2 Weakly Supervised Setting: Discovering Senescence Markers from Aging Atlases

Based on: *Work done in collaboration with Blue and Pink labs at University of Pittsburgh School of Medicine (UPMC)*

Senescence was first discovered by Hayflick & Moorhead[108] in 1961. They demonstrated that normal human cells cultured in the lab had a limited proliferative capacity—typically spanning 40 to 60 generations. Beyond this threshold, these cells transition into a senescence phase. Since then, research investigating the factors linked to senescence has been on the rise. These studies employ diverse techniques such as gene knockout models, functional assays, and rigorous validation processes to determine biomarkers and understand how senescence affects the body. Several works have built on top of these studies by compiling marker gene sets enriched for senescence[109–112].

A principal limitation of these studies is the absence of a precise and universally accepted definition of senescence. This is, in part, due to the lack of a known gene that is universally activated across all SnCs, making formalization and standardization difficult. Indeed, existing senescence marker sets have very little agreement between them. Furthermore, studies on senescence reveal that both SASP and senescence markers are highly dependent on cell function, hence, it is unclear whether such universal signatures even exist[4].

This problem once again underscores the importance of cell-specific markers. While the method discussed in section 2.1 is appropriate for such task, it assumes the existence of a label for each cell. In the context of senescence, such labels are missing from virtually all public datasets because there is no adopted standard for identifying SnCs. Therefore, supervised approaches such as Phenotype Cover cannot be applied.

Nonetheless, the genes associated with senescence, while not universally consistent, can still serve as a valuable starting point. Learning from such incomplete and ambiguous information is a well-studied problem in the machine learning community that falls under the paradigm of *weakly supervised learning*. Weak supervision applies to problems where label information is noisy or incomplete[38]. Senescence fits this description. A class of learners known as Positive-Unlabeled (PU) learning algorithms are particularly relevant for addressing the challenge of identifying SnCs. We provide an introduction to PU learning below.

2.2.1 PU Learning Under Covariate Shift for Identifying Senescent Cells

Positive-Unlabeled (PU) learning is a variant of binary classification where the goal is to distinguish between positive and negative samples, with the restriction that only positive samples are seen during the training phase[113]. More concretely, the real data distribution is a mixture given by

$$p(x) = \alpha p_+(x) + (1 - \alpha)p_-(x) \tag{2.2}$$

where α is the mixture proportion, and p_+, p_- are the probability density functions of the positive and negative samples, respectively. In traditional binary classification, we are given training data $\mathcal{D}_{\text{tr}} \sim p(x)$ and test data $\mathcal{D}_{\text{te}} \sim p(x)$. However, in PU learning, $\mathcal{D}_{\text{tr}} \sim p_+(x)$ which makes it more challenging. Many methods have been proposed which typically assume some form of smoothness or separability of the classes, or treat the negative samples as noise[114–116].

PU learning fits our setup as follows. Given the current uncertainty in identifying senescent cells (SnCs) using gene expression data, we treat the population of SnCs as negative samples within an unlabeled set. The objective is to recover a PU classifier capable of distinguishing SnCs from healthy cells. Subsequently, differential expression (DE) analysis between these two groups can be used to identify marker genes.

To obtain (labeled) training and (unlabeled) test sets for PU learning, we rely on different age groups present in the data, where $\mathcal{D}_{\text{tr}}$ models young individuals, and $\mathcal{D}_{\text{te}}$ models more senior ones. To accomodate the PU learning setup, we require the following assumptions to hold: a) the young group contains few to no SnCs, b) the negative (non-healthy) class in the senior group consists mostly of SnCs, and c) the healthy cells in the senior group come from the same distribution as the healthy cells in the young. The last assumption is problematic as literature has shown that aging patients suffer from other hallmarks of aging not related to senescence such as inflammation, epigenetic alterations, or mitochondrial dysfunction[117–119]. Stated differently:

$$p_{+}^{\text{young}}(x) \neq p_{+}^{\text{senior}}(x). \tag{2.3}$$

Therefore, we face a covariate shift[120]. In covariate shift, the dependence of the response variable y (i.e., senescence status) on gene expression x is the same for both the training and test sets, however, the input distributions may not be:

$$p_{\text{tr}}(y \mid x) = p_{\text{te}}(y \mid x) \tag{2.4}$$

$$p_{\text{tr}}(x) \neq p_{\text{te}}(x). \tag{2.5}$$

To address this, the PU learner needs to accommodate a covariate shift. Here, we rely on the formulation of Sakai & Shimizu which propose an unbiased risk estimator for covariate shift adaptation on PU learning, termed PUc[121].

PUc assumes we are given three sets of samples: labeled training data $\{x_i^{\text{Ptr}}\}_i \sim p_{\text{tr}}(x \mid y = 1)$, unlabeled training data $\{x_j^{\text{Utr}}\}_j \sim p_{\text{tr}}(x)$, and unlabeled test data $\{x_k^{\text{Ute}}\}_k \sim p_{\text{te}}(x)$. When covariate shift occurs, the PU risk on the test distribution p_{te} differs from the PU risk on the train distribution p_{tr}. PUc addresses this by importance-weighting, where the ratio between test and train densities $w(x) := p_{\text{te}}(x)/p_{\text{tr}}(x)$ is used to weight each sample during the computation of the risk[122,123]. In this case, the PUc risk becomes

$$R_{\text{PUc}}(g) := \alpha \mathbb{E}_{x \sim p_{\text{tr}}(x|y=1)}[\tilde{\ell}(g(x))w(x)] + \mathbb{E}_{x \sim p_{\text{tr}}(x)}[\ell(-g(x))w(x)], \tag{2.6}$$

where g is a classifier and ℓ is a loss function with $\tilde{\ell}(x) := \ell(x) - \ell(-x)$. The PUc risk on training data can be shown to be an unbiased estimator of the PU risk on test data. The mixture proportion α is estimated from prior knowledge. Similar to the original work, we employ a linear-in-parameter classifier with a Gaussian kernel basis function. For more details on PUc, please refer to Sakai & Shimizu[121]. An illustration of the proposed approach is given in Fig. 2.5.

2.2.2 Results

Demographics and characterization of the HLCA

We developed a computational method that uses the largest publicly available single-cell lung dataset—the Human Lung Cell Atlas (HLCA)[124]—to identify senescent cell populations in the lung across different ages. Our approach is based on positive-unlabeled learning under covariate shift (PUc)[121] that enables the derivation of a list of senescence markers via direct differential expression (DE) analysis of healthy (i.e., non-senescent) and (Fig. 2.5). To achieve this, we trained and tested this PUc learning approach by treating different age groups in the HLCA as (un)labeled data (Methods).

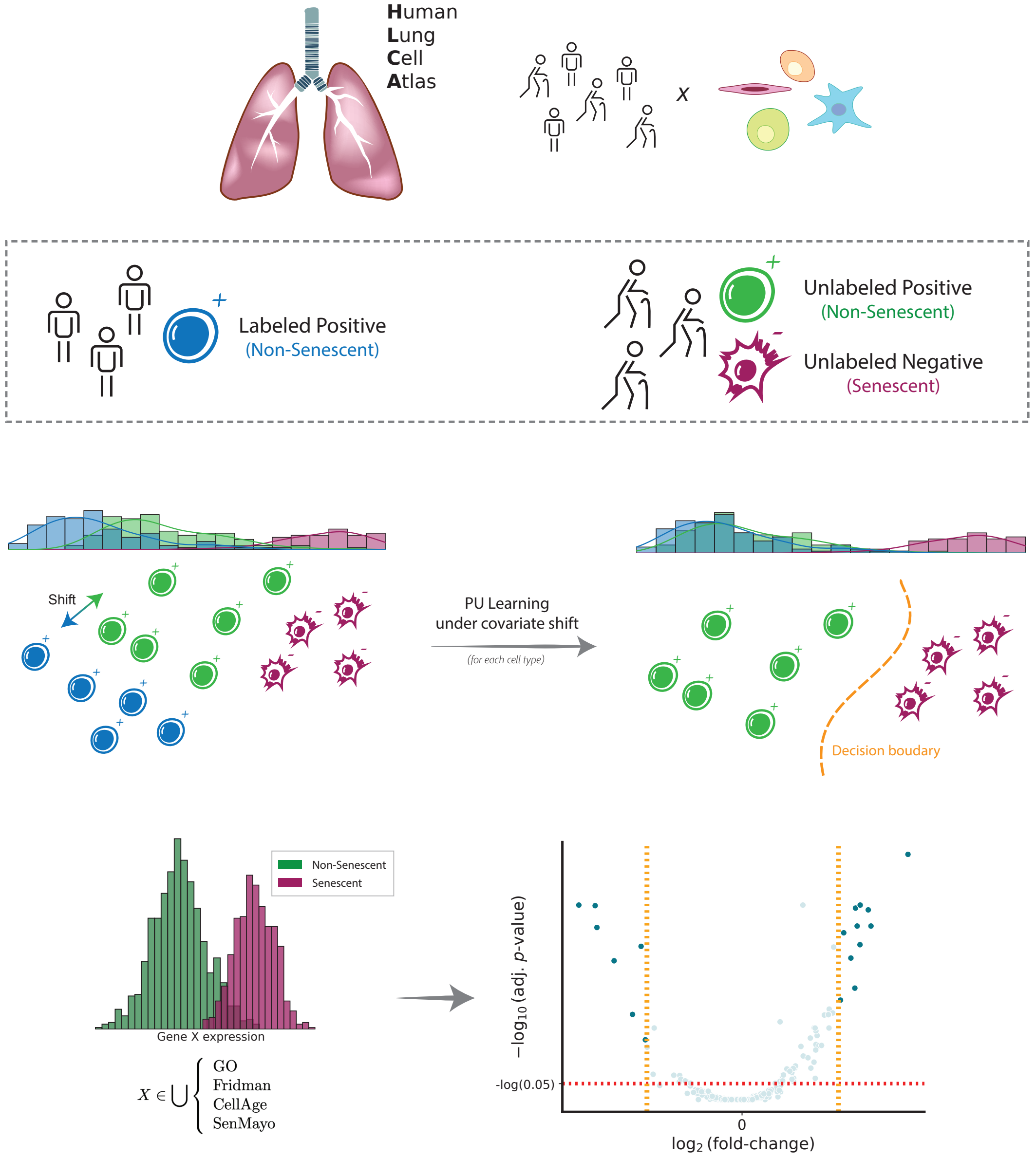


Fig. 2.5: **Schematic illustration of PUc learning for identifying SnCs.** Given a large single-cell lung cohort from young and old individuals (top), we designate cells from young individuals as non-senescent cells (positive). Cells in older individuals are unlabeled initially. We then use positive-unlabeled learning under covariate shift (PUc) to identify SnCs in older individuals (middle). Using these cells we develop an expression profile for senescence markers in several different cell types (bottom).

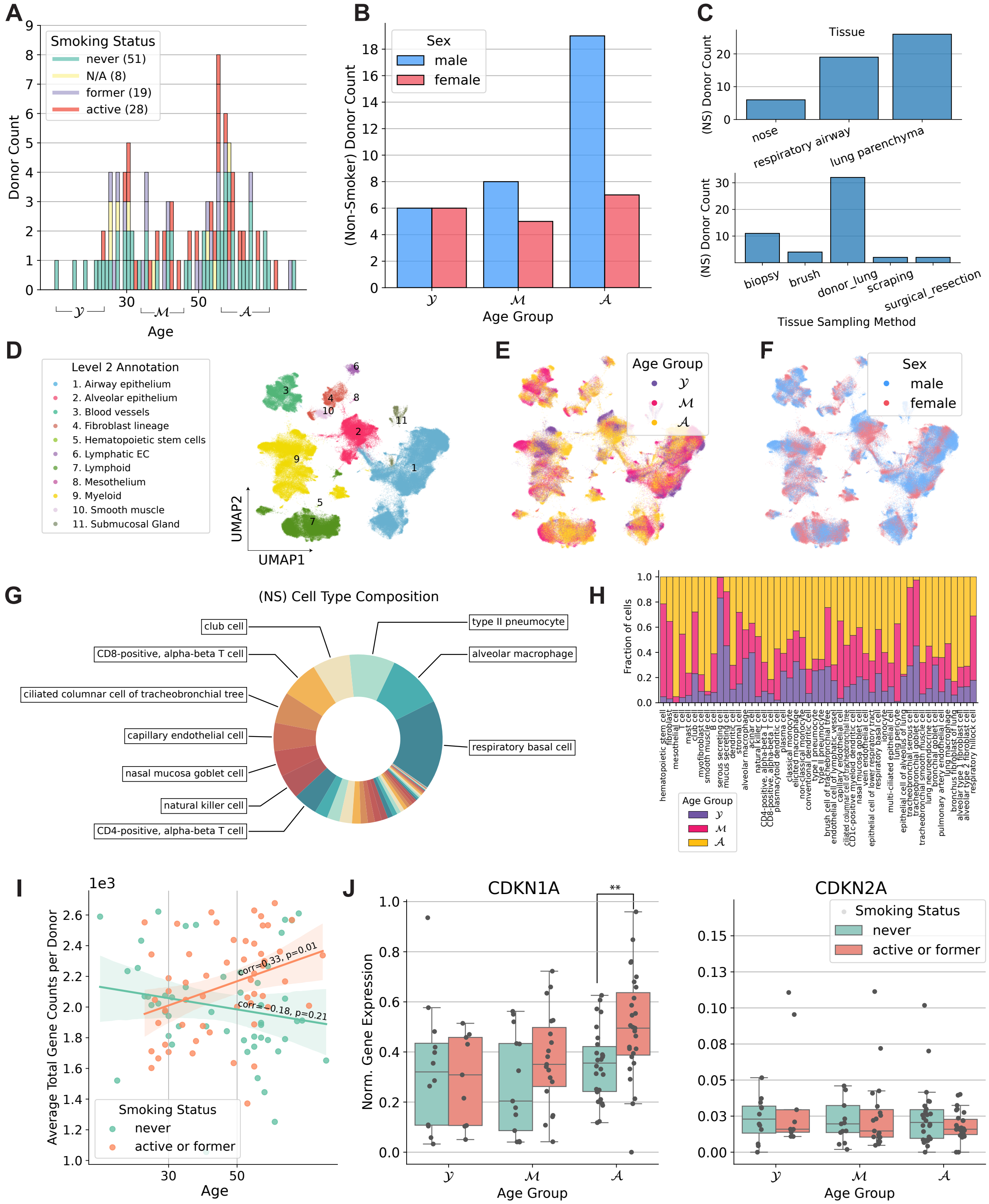


Fig. 2.6: **Summary of the The Human Lung Cell Atlas**. (A) Distribution of donors by age and smoking history. (B) Number of donors by age group and sex. (C) Number of donors broken down by tissue and sampling method (non-smokers). (D-F) UMAP plots describing level 2 cell types, age groups, and sex ((NS = non-smoker)). (G) Cell type proportion among non-smokers. (H) Age group representation across cell types (non-smokers). (I) Average cell total counts for each donor. (J) Normalized gene expression values for *CDKN1A* and *CDKN2A*.

The 107 tissue samples in the HLCA were derived from individuals aged 10 to 76 years, including 51 never-smokers, 19 former smokers and 28 active smokers (Fig. 2.6A, E). Overall, within the HCLA, tissue samples were derived from 69 males and 38 females. Among the 51 never-smokers, 33 were male with most of them belonging to the older age group (Fig. 2.6B, F). Thus, this dataset enables analysis of sex differences as well as analysis based on cigarette smoke exposure. Most of the samples originated from lung parenchyma from donor lungs that have not been deemed suitable for transplantation (Fig. 2.6C). A total of 50 cell types were present in the atlas at the finest annotation level (Fig. 2.6H), with respiratory basal cells and alveolar macrophages being the two most prevalent cell types among never-smokers (Fig. 2.6G). Among active smokers, type II pneumocytes and basal cells were the most common. The total number of cells was 584,944 with 301,791 cells from never-smokers. Total gene counts increased with age among smokers with a correlation of 0.33 ($p = 0.01$), while a slight decrease was observed for never-smokers, although non-significant (Fig. 2.6I). We first analyzed the expression of *CDKN1A* and *CDKN2A*, which encode the senescence markers p16 and p21, respectively. Notably, *CDKN1A* was upregulated in smokers for the older two age groups (Fig. 2.6J). The most significant upregulation was observed for the oldest smoker group compared to non-smokers (two-tailed t-Test, $p = 0.004$). No significant differences across the ages, nor between smokers and never smokers were found for *CDKN2A* (Fig. 2.6J).

Generation of SenSet from the HLCA

Several lung senescence gene sets have been published to date (GO:0090398, Fridman, SenMayo, and CellAge[110–112,125]). We first examined the extent of overlap between these gene sets. We observed that the pairwise overlap between them is relatively small compared to the size of each individual set, with the highest overlap of 34 genes shared between the GO and CellAge sets (Fig. 2.7B). The union **U** of all sets contains 501 unique marker genes, of which 434 were detected in the HLCA.

We sought to identify a subset of **U** that demonstrates greater specificity for senescence. The PUc estimator[121] constructs a model of healthy cells based on data from young and middle-aged individuals (groups $\mathcal{Y}$, $\mathcal{M}$, respectively) and then applies this model to identify cells that are senescent in the aged group (group $\mathcal{A}$, Fig. 2.5, 2.6A). PUc accounts for potential covariate shifts in $\mathcal{A}$, which may arise due to other aging processes and hallmarks, including inflammation or epigenetic alterations[117,118]. The advantage of this approach is that it allows for a direct comparison of SnCs against non-senescent cells within the same group—the oldest age group $\mathcal{A}$. This addresses the challenge of age-related confounding factors that may arise when comparing older and younger individuals. While PUc identifies cells that generally deviate from the healthy (young) profile, we hypothesize that a significant proportion of these non-healthy cells identified by PUc are indeed senescent. We denote these cells with a $(-)$ superscript to signify that they belong to the negative (senescent) class. Similarly a $(+)$ superscript will denote non-senescent cells for that group.

We applied PUc to 31 cell types in the HLCA with a sufficient number of cells per age group (at least 50), using data exclusively from non-smokers to study senescence genes without the impact of cigarette smoke exposure. PUc identified at least 10 cells in the negative class—claimed here to be senescent—within 22 of these cell types (Fig. 2.7A,D,E). The proportion of cells assigned to this class ranged from 0 to 76 %. Major cell types with high enrichment for senescence included alveolar type 1 fibroblasts, respiratory basal cells, and tracheobronchial smooth muscle cells (respectively, 44 %, 39 %, and 53 %). Of note, for some of the cell types with a high percentage of SnCs, such as tracheobronchial serous cells (76 %), few differentially expressed (DE) genes in these cell types were consistent with other types, suggesting mislabeling or insufficient data.

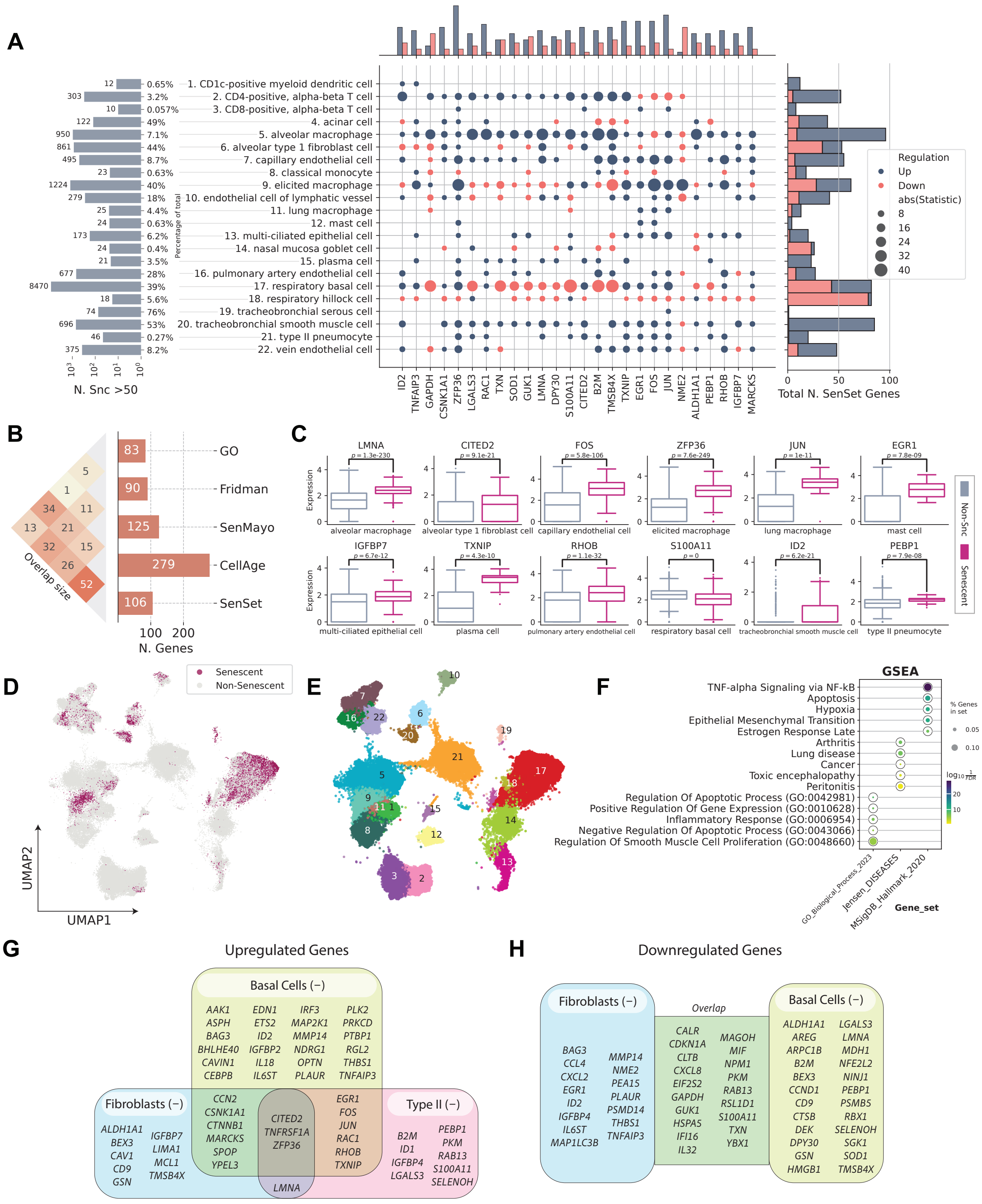


Fig. 2.7: **PUc identified a novel SenSet senescence signature**. (A, left to right) The number of cells (≥ 50) identified as senescent in $\mathcal{A}$ and the corresponding percentage; the most frequent genes enriched for most cell types; the total number of SenSet marker genes assigned to a cell type. (B) Overlap sizes of SenSet with the prior lists. (C) Selected marker genes for some of the cell types and distribution among healthy and SnCs. (D) UMAP plot of senescent cells in $\mathcal{A}$. (E) UMAP plot of cell types assigned at least one marker in $\mathcal{A}$. (F) Top GO, Jensen, and MSigDB terms enriched for SenSet. (G) SenSet marker genes enriched in basal$^{(-)}$ cells, fibroblasts$^{(-)}$ and AT2$^{(-)}$ cells.

Differential expression analysis of SenSet genes

Next, we performed a DE test between the senescent $\mathcal{A}^{(-)}$ and non-senescent $\mathcal{A}^{(+)}$ cells for each cell type within the old age group, and identified the genes in the overlap senescence signature that were significantly enriched in at least six cell types (FDR = 0.05). This number was chosen to get a set of approximately 100 genes, which we termed SenSet (Table 2.3, Fig. 2.7G). SenSet showed the highest overlap with CellAge (52 out of 106 genes), followed by Fridman (32) and SenMayo (26) (Fig. 2.7B).

The senescence hallmark gene *CDKN1A* was enriched in 7 cell types, thus is included in SenSet, while *CDKN2A*, which was enriched in only alveolar macrophages, is not. SenSet also contains SASP protein members, such as C-X-C motif chemokine ligand 8 (*CXCL8*), interleukin 18 (*IL18*), and insulin growth factor binding protein 7 (*IGFBP7*)[126–129]. Genes upregulated in most cell types include ZFP36 ring finger protein (*ZFP36*, 16 cell types), Jun proto-oncogene (*JUN*, 13), and thioredoxin interacting protein (*TXNIP*), early growth response 1 (*EGR1*), Fos proto-oncogene (*FOS*) (11 each), all of which encode for proteins known to be involved in signaling in transcriptional response to hypoxia and cellular stress[130,131]. In contrast, nucleoside diphosphate kinase 2 (*NME2*), a suppressor of apoptosis, was down-regulated in 9 cell types, followed by nucleophosmin 1 (*NPM1*) and glyceraldehyde-3-phosphate dehydrogenase (*GAPDH*) (7 each), involved in DNA replication and cell cycle[132–134].

Gene set enrichment analysis (GSEA) using the MSigDB gene set[135] revealed that SenSet is significantly enriched for genes involved in TNF-alpha signaling via NF-κB (27 genes, $q = 1\text{e}{-29}$), apoptosis (15 genes, $q = 1\text{e}{-13}$), and hypoxia (15 genes, $q = 1\text{e}{-12}$) (Fig. 2.7F). Additionally, SenSet is enriched for genes associated with arthritis (12 genes, $q = 1\text{e}{-8}$) and lung disease (10 genes, $q = 1\text{e}{-8}$) based on Jensen's disease set[136]. Notably, gene ontology (GO) analysis highlighted enrichment for the process "regulation of smooth muscle cell proliferation" ($q = 1\text{e}{-6}$). This finding aligns with the observation that TSM$^{(-)}$ cells showed an upregulation of most SenSet genes.

Cell type specific enrichment of SenSet genes

Not surprisingly, cell types displayed considerable heterogeneity in the expression of SenSet markers. From all SenSet genes, 87 were upregulated in alveolar macrophages$^{(-)}$ and 84 in tracheobronchial smooth muscle cells (TSM)$^{(-)}$, representing the highest numbers among the 22 cell types considered. Conversely, basal$^{(-)}$ and type 1 fibroblasts$^{(-)}$ showed a downregulation of 43 and 34 genes, respectively (Fig. 2.7A,C, Table 2.3).

Type II pneumocytes and fibroblasts are crucial structural cell types in the lung that have been implicated in senescence[108,137,138]. In fibroblasts$^{(-)}$, 19 SenSet genes were upregulated. Type II pneumocytes$^{(-)}$ also show an upregulation of 19 SenSet genes (different set), and 1 downregulated gene, *CTNNB1*. We found substantial overlap in upregulated genes between basal$^{(-)}$ cells and type II$^{(-)}$ pneumocytes, fibroblasts$^{(-)}$, respectively, with *TNFRSF1A*, *CITED2*, and *ZFP36* in common across all three. For instance, 9 genes were upregulated in both fibroblasts$^{(-)}$ and basal$^{(-)}$ cells, and 9 genes were also upregulated in both type II pneumocytes$^{(-)}$ and basal$^{(-)}$ cells (Fig. 2.7G).

Basal cells represent bona fide stem cells of the lung, and stem cell exhaustion—recognized as a hallmark of aging—has been associated with senescence[139]. Among 34 downregulated SenSet genes in fibroblasts$^{(-)}$, 19 of these were also downregulated in basal$^{(-)}$ cells (Fig. 2.7H).

Several genes were found to be upregulated in one cell type and downregulated in the other. For instance, *LMNA* was downregulated in basal$^{(-)}$ cells, but upregulated in both fibroblasts$^{(-)}$ and type II pneumocytes$^{(-)}$. Another example, *IGFBP4*, was downregulated in fibroblasts and upregulated in type II pneumocytes. In basal$^{(-)}$ cells, 39 SenSet genes were upregulated.

Symbol	Full Name
AAK1$^{\triangle}$	AP2 Associated Kinase 1
AKR1B1	Aldo-Keto Reductase Family 1 Member B
ALDH1A1$^{\blacktriangle,\triangledown}$	Aldehyde Dehydrogenase 1 Family Member A1
AREG$^{\triangledown}$	Amphiregulin
ARPC1B$^{\triangledown}$	Actin Related Protein 2/3 Complex Subunit 1B
ASPH$^{\triangle}$	Aspartate Beta-Hydroxylase
B2M$^{\triangledown}$	Beta-2-Microglobulin
BAG3$^{\blacktriangledown,\triangle}$	BAG Cochaperone 3
BEX3$^{\blacktriangle,\triangledown}$	Brain Expressed X-Linked 3
BHLHE40$^{\triangle}$	Basic Helix-Loop-Helix Family Member E40
CALR$^{\blacktriangledown,\triangledown}$	Calreticulin
CAV1$^{\blacktriangle}$	Caveolin 1
CAVIN1$^{\triangle}$	Caveolae Associated Protein 1
CCL3	C-C Motif Chemokine Ligand 3
CCL3L1	C-C Motif Chemokine Ligand 3 Like 1
CCL4$^{\blacktriangledown}$	C-C Motif Chemokine Ligand 4
CCN2$^{\blacktriangle,\triangle}$	Cellular Communication Network Factor 2
CCND1$^{\triangledown}$	Cyclin D1
CD44	CD44 Molecule (IN Blood Group)
CD9$^{\blacktriangle,\triangledown}$	CD9 Molecule
CDKN1A$^{\blacktriangledown,\triangledown}$	Cyclin Dependent Kinase Inhibitor 1A
CEBPB$^{\triangle}$	CCAAT Enhancer Binding Protein Beta
CITED2$^{\blacktriangle,\triangle}$	Cbp/P300 Interacting Transactivator With Glu/Asp Rich Carboxy-Terminal Domain 2
CLTB$^{\blacktriangledown,\triangledown}$	Clathrin Light Chain B
CSNK1A1$^{\blacktriangle,\triangle}$	Casein Kinase 1 Alpha 1
CTNNB1$^{\blacktriangle,\triangle}$	Catenin Beta 1
CTSB$^{\triangledown}$	Cathepsin B
CXCL2$^{\blacktriangledown}$	C-X-C Motif Chemokine Ligand 2
CXCL8$^{\blacktriangledown,\triangledown}$	C-X-C Motif Chemokine Ligand 8
DEK$^{\triangledown}$	DEK Proto-Oncogene
DPY30$^{\triangledown}$	Dpy-30 Histone Methyltransferase Complex Regulatory Subunit
EDN1$^{\triangle}$	Endothelin 1
EGR1$^{\blacktriangledown,\triangle}$	Early Growth Response 1
EIF2S2$^{\blacktriangledown,\triangledown}$	Eukaryotic Translation Initiation Factor 2 Subunit Beta
ETS2$^{\triangle}$	ETS Proto-Oncogene 2, Transcription Factor
EWSR1	EWS RNA Binding Protein 1
FOS$^{\triangle}$	Fos Proto-Oncogene, AP-1 Transcription Factor Subunit
GAPDH$^{\blacktriangledown,\triangledown}$	Glyceraldehyde-3-Phosphate Dehydrogenase
GMFG	Glia Maturation Factor Gamma
GSN$^{\blacktriangle,\triangledown}$	Gelsolin
GUK1$^{\blacktriangledown,\triangledown}$	Guanylate Kinase 1
HDAC1	Histone Deacetylase 1
HMGB1$^{\triangledown}$	High Mobility Group Box 1
HSPA5$^{\blacktriangledown,\triangledown}$	Heat Shock Protein Family A (Hsp70) Member 5
ID1	Inhibitor Of DNA Binding 1
ID2$^{\blacktriangledown,\triangle}$	Inhibitor Of DNA Binding 2
IFI16$^{\blacktriangledown,\triangledown}$	Interferon Gamma Inducible Protein 16
IGFBP2$^{\triangle}$	Insulin Like Growth Factor Binding Protein 2
IGFBP4$^{\blacktriangledown}$	Insulin Like Growth Factor Binding Protein 4
IGFBP7$^{\blacktriangle}$	Insulin Like Growth Factor Binding Protein 7
IL18$^{\triangle}$	Interleukin 18
IL32$^{\blacktriangledown,\triangledown}$	Interleukin 32
IL6ST$^{\blacktriangledown,\triangle}$	Interleukin 6 Cytokine Family Signal Transducer
IRF3$^{\triangle}$	Interferon Regulatory Factor 3
ISG15	ISG15 Ubiquitin Like Modifier
JUN$^{\triangle}$	Jun Proto-Oncogene, AP-1 Transcription Factor Subunit
LGALS3$^{\triangledown}$	Galectin 3
LIMA1$^{\blacktriangle}$	LIM Domain And Actin Binding 1
LMNA$^{\blacktriangle,\triangledown}$	Lamin A/C
MAGOH$^{\blacktriangledown,\triangledown}$	Mago Homolog, Exon Junction Complex Subunit
MAP1LC3B$^{\blacktriangledown}$	Microtubule Associated Protein 1 Light Chain 3 Beta
MAP2K1$^{\triangle}$	Mitogen-Activated Protein Kinase Kinase 1
MAP2K3	Mitogen-Activated Protein Kinase Kinase 3
MARCKS$^{\blacktriangle,\triangle}$	Myristoylated Alanine Rich Protein Kinase C Substrate
MCL1$^{\blacktriangle}$	MCL1 Apoptosis Regulator, BCL2 Family Member
MDH1$^{\triangledown}$	Malate Dehydrogenase 1
MIF$^{\blacktriangledown,\triangledown}$	Macrophage Migration Inhibitory Factor
MMP14$^{\blacktriangledown,\triangle}$	Matrix Metallopeptidase 14
NDRG1$^{\triangle}$	N-Myc Downstream Regulated 1
NFE2L2$^{\triangledown}$	NFE2 Like BZIP Transcription Factor 2
NINJ1$^{\triangledown}$	Ninjurin 1
NME2$^{\blacktriangledown}$	NME/NM23 Nucleoside Diphosphate Kinase 2
NPM1$^{\blacktriangledown,\triangledown}$	Nucleophosmin 1
OPTN$^{\triangle}$	Optineurin
PEA15$^{\blacktriangledown}$	Proliferation And Apoptosis Adaptor Protein 15
PEBP1$^{\triangledown}$	Phosphatidylethanolamine Binding Protein 1
PKM$^{\blacktriangledown,\triangledown}$	Pyruvate Kinase M1/2
PLAUR$^{\blacktriangledown,\triangle}$	Plasminogen Activator, Urokinase Receptor
PLK2$^{\triangle}$	Polo Like Kinase 2
PRKCD$^{\triangle}$	Protein Kinase C Delta
PSMB5$^{\triangledown}$	Proteasome 20S Subunit Beta 5
PSMD14$^{\blacktriangledown}$	Proteasome 26S Subunit, Non-ATPase 14
PTBP1$^{\triangle}$	Polypyrimidine Tract Binding Protein 1
RAB13$^{\blacktriangledown,\triangledown}$	RAB13, Member RAS Oncogene Family
RAC1$^{\triangle}$	Rac Family Small GTPase 1
RBX1$^{\triangledown}$	Ring-Box 1
RGL2$^{\triangle}$	Ral Guanine Nucleotide Dissociation Stimulator Like 2
RHOB$^{\triangle}$	Ras Homolog Family Member B
RSL1D1$^{\blacktriangledown,\triangledown}$	Ribosomal L1 Domain Containing 1
S100A11$^{\blacktriangledown,\triangledown}$	S100 Calcium Binding Protein A11
SELENOH$^{\triangledown}$	Selenoprotein H
SGK1$^{\triangledown}$	Serum/Glucocorticoid Regulated Kinase 1
SMARCB1	SWI/SNF Related, Matrix Associated, Actin Dependent Regulator Of Chromatin, Subfamily B, Member 1
SOD1$^{\triangledown}$	Superoxide Dismutase 1
SPOP$^{\blacktriangle,\triangle}$	Speckle Type BTB/POZ Protein
THBS1$^{\blacktriangledown,\triangle}$	Thrombospondin 1
TMSB4X$^{\blacktriangle,\triangledown}$	Thymosin Beta 4 X-Linked
TNFAIP3$^{\blacktriangledown,\triangle}$	TNF Alpha Induced Protein 3
TNFRSF1A$^{\blacktriangle,\triangle}$	TNF Receptor Superfamily Member 1A
TPR	Translocated Promoter Region, Nuclear Basket Protein
TXN$^{\blacktriangledown,\triangledown}$	Thioredoxin
TXNIP$^{\triangle}$	Thioredoxin Interacting Protein
VIM	Vimentin
YBX1$^{\blacktriangledown,\triangledown}$	Y-Box Binding Protein 1
YPEL3$^{\blacktriangle,\triangle}$	Yippee Like 3
ZFP36$^{\blacktriangle,\triangle}$	ZFP36 Ring Finger Protein

Table 2.3: **All 106 SenSet genes with full names**. Marked are genes upregulated in fibroblasts$^{(-)}$ ($\blacktriangle$), downregulated in fibroblasts$^{(-)}$, ($\blacktriangledown$), upregulated in basal$^{(-)}$ cells ($\triangle$), and downregulated in basal$^{(-)}$ cells ($\triangledown$) in the HLCA.

Senescence validation in human tissue model

To further refine and validate SenSet, we utilized a highly complex ex vivo human 3D tissue culture model based on precision cut-lung slices (PCLS). PCLS recapitulate the complexity of the lungs environment in situ, enabling the study of various lung cell types and lineages present in the parenchymal region, along with the extracellular matrix (ECM) within the lungs' native 3D architecture at high temporal and spatial resolution. We have previously demonstrated that this model can be applied to mimic the onset and progression of lung injury and diseases and enables testing potential therapeutics in living, diseased human tissue[140–142].

PCLS were generated from the lower left lung lobe (Fig. 2.8A) of healthy donors aged 20 to 78 (Table 2.4). To induce cellular senescence, PCLS were treated with either bleomycin (15 µg/mL) or doxorubicin (0.1 µM) for up to six days. Lung structure remained intact after six days in culture, as shown by hematoxylin and eosin (H&E) staining (Figure 2.8B).

Senescence induction was validated by several state-of-the-art readouts, with a significant increase in the number of β-galactosidase-positive cells (Fig. 2.8C) and p21-positive cells in PCLS (Fig. 2.8D, E). The percentage of p21 positive cells was significantly increased after bleomycin (2.9–fold ±0.8) and doxorubicin treatment (2.0–fold ±0.9, Fig. 2.8E). This was further confirmed by an increase in p21 protein after bleomycin (4.1–fold ±4.17) and doxorubicin treatment (1.9–fold ±1.49), respectively (Fig. 2.8F). Moreover, *GDF-15*, a known SASP protein, was significantly increased after bleomycin (4.6–fold ±2.7) and doxorubicin treatment (4.1–fold ±3.1) (Fig. 2.8G).

To validate that our SenSet list—which was derived from the HCLA using lung tissue across the ages—is indeed a senescence signature, we subjected our human senescence induction model to single-nucleus RNA sequencing (snRNA-seq) and further analyzed cell type specific gene expression. Senescence was induced in PCLS by bleomycin or doxorubicin as described above and we further included PCLS subjected to irradiation, as previously reported[143]. The samples used for irradiation originated from peritumor tissue.

Importantly, we identified all major cell lineages in our ex vivo human tissue model and identified four major epithelial cell types (Fig. 2.9E) based on lung canonical markers[144], with no discernible batch condition or cell cycle effects on data (Fig. 2.9D).

For each gene in SenSet as well as each of the prior gene sets, we performed a rank sum DE test to determine if the gene is up or downregulated in our PCLS senescence models. Notably, SenSet achieves the highest proportion of significantly regulated genes in all samples when compared to all prior lung senescence lists (FDR = 0.05, Fig. 2.9B). Several SenSet genes that were upregulated in $\mathcal{A}^{(-)}$ (HCLA) were also upregulated after senescence induction in human PCLS ex vivo, including *JUN*, and *IGFBP7*. The transcription factor *TXNIP*, which is known to be suppressed by p21, was downregulated in response to all three senescence inducers as well as the cell cycle regulators *NME2* and *NPM1*, which were also the top downregulated genes in $\mathcal{A}^{(-)}$ (Fig. 2.9C). *CDKN1A* was increased after all three treatments, with the highest induction after bleomycin treatment. No such increase was observed for *CDKN2A* (Figure 2.9C, F).

So far, we have compared the expression of senescence markers across conditions using the entire sample. We next turned our attention to cell type-specific DE analysis of the marker genes, using the manually annotated cell types. For each cell type, we performed a similar rank sum DE test as before between treatment and control samples, and combined the p-values of these tests using Pearson's method (Fig. 2.10A). This analysis revealed that SenSet markers showed significant enrichment across 10 cell types ($p \leq 0.05$). In comparison, markers from Fridman and SenMayo lists showed significant enrichment in only four cell types. CellAge, on the other hand, was not significantly enriched in any cell type, likely due to the high number of non-DE markers in that list.

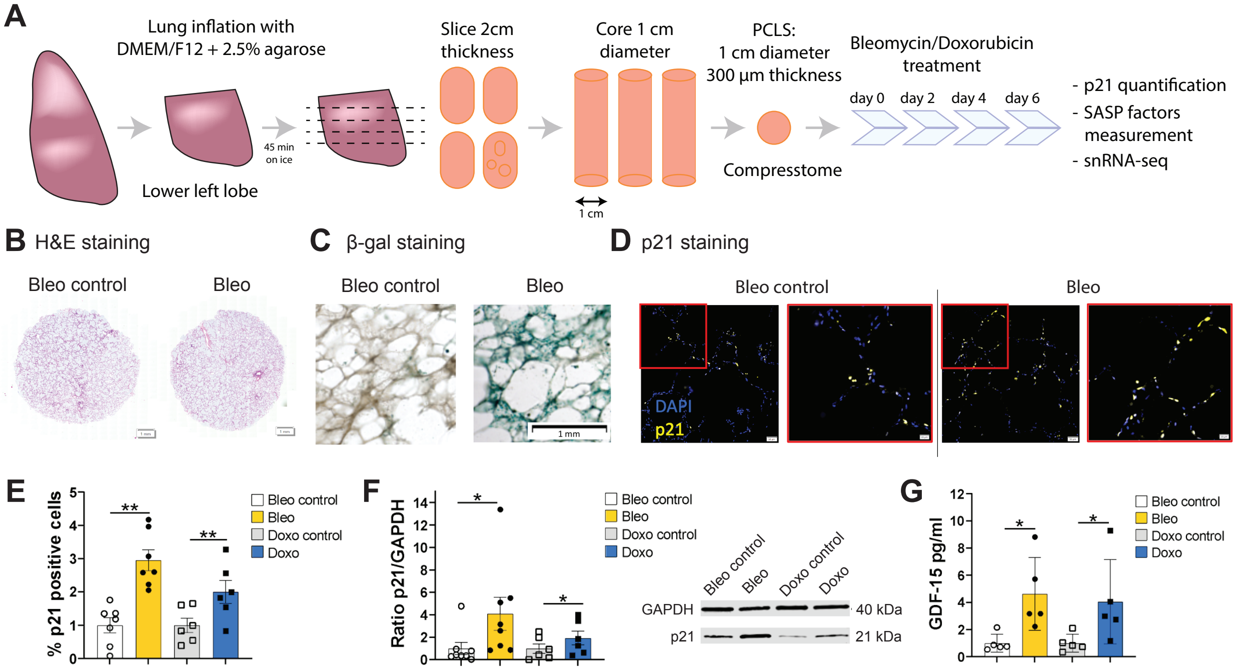


Fig. 2.8: **Senescence induction in human PCLS by DNA damage.** (A) PCLS were generated from healthy donors lower left lobe lung with an age range of 20 to 78 years-old. Senescence was induced by treatment with bleomycin (Bleo) at 15 mg/mL, or doxorubicin (Doxo) at 0.1 µM for 6 days, and PCLS and supernatants were collected. (B) Hematoxylin eosin (H&E) staining on 4 µm sliced formalin-fixed paraffin embedded human PCLS at day 6. (C) *β*-galactosidase staining on whole PCLS at day 6. (D) p21 immunohistofluorescence (IHF) on 4 µm sliced formalin-fixed paraffin embedded human PCLS at day 6. (E) p21 positive cells quantification based on p21 IHF staining presented in (D) after bleomycin ($n = 7$) or doxorubicin ($n = 6$) treatment. (F) Quantification of p21 protein level by Western blot (WB) after bleomycin ($n = 8$) or doxorubicin ($n = 6$) treatment and representative blot. (G) SASP factor, *GDF-15*, measured by Luminex$^{\text{TM}}$ assay on human PCLS supernatants after bleomycin or doxorubicin treatment ($n = 5$). Paired t-test: $^{**}p < 0.005$, $^{*}p < 0.05$.

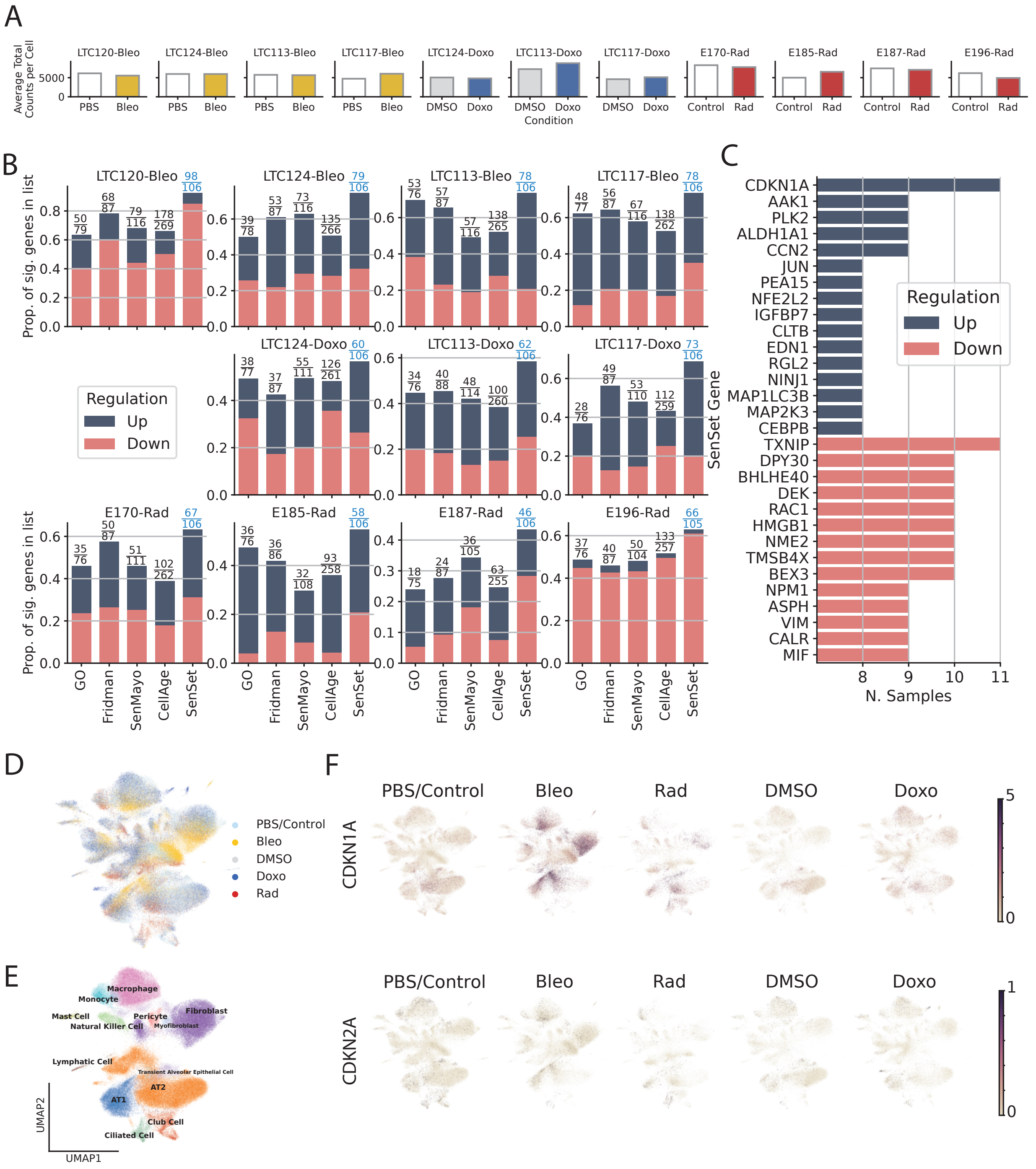


Fig. 2.9: **Validation of SenSet**. (A) Average total counts per cell across samples and conditions. (B) For each subject, we show the fraction of the genes in each list which were significantly up or downregulated with treatment. (C) SenSet genes up (down)regulated in most samples. (D) UMAP plot of scVI integrated data. (E) Clusters identified using Leiden clustering on scVI embeddings. (F) Normalized expression of *CDKN1A* and *CDKN2A* across conditions.

Donor	Sex	Age	Smoking status	IHF[1] p21	WB[2] p21	Multiplex assay	snRNA-seq
LTC 113	F	56	Never	✓			✓(B, D)
LTC 117	M	73	Never	✓		✓	✓(B, D)
LTC 118	F	23	Never	✓		✓(D)	
LTC 119	F	38	Never	✓(D)		✓(D)	
LTC 120	M	21	Never	✓(B)		✓(B)	✓(B)
LTC 121	M	62	Former	✓(B)	✓(B)	✓(B)	
LTC 124	M	36	Never	✓	✓	✓	✓(B, D)
LTC 127	M	41	Never	✓	✓	✓	
LTC 137	F	78	Former	✓	✓		
LTC 152	F	27	N/A		✓(B)		
LTC 164	M	57	Former		✓		
LTC 176	M	25	Never		✓		
LTC 200	M	20	Never		✓		
E170[3]	M	75	Former				✓(I)
E185[3,4]	F	81	Former				✓(I)
E187[3]	M	64	Former				✓(I)
E196[3]	F	70	Never				✓(I)

Table 2.4: **Donor information**. [1]IHF: immunohistofluorescence; [2]WB: western blot; [3]Peritumor tissue; [4]COPD Gold II; B: Bleomycin; D: Doxorubicin; I: Irradiation. If parentheses are missing, it is assumed to be (B, D).

A close inspection of the markers across cell types confirmed that many genes in the prior lists many genes were not DE in any cell type (Fig. 2.10C). In contrast, nearly all SenSet genes (all but seven) were DE in at least one cell type (Fig. 2.10B). SenSet markers were predominantly upregulated in AT1$^{(-)}$ and AT2$^{(-)}$ cells, while mostly downregulated in fibroblasts$^{(-)}$ and macrophages$^{(-)}$. Notably, 15 genes downregulated in fibroblasts$^{(-)}$ in the HLCA dataset were also downregulated in fibroblasts$^{(-)}$ in the treated PCLS data: *CALR*, *GAPDH*, *GUK1*, *IL32*, *MIF*, *NME2*, *NPM1*, *PKM*, *PLAUR*, *RAB13*, *S100A11*, *THBS1*, *TNFAIP3*, *TXN*, *YBX1*. A similar correspondence was observed for (elicited) macrophages and AT2 cells. For AT2$^{(-)}$ cells, 13 SenSet markers were upregulated in both the HLCA (19 total) and the PCLS (73 total).

Analysis of smokers in the HLCA

Air pollutants and cigarette smoke exposure are a major risk factor for the development and exacerbation of age-related lung diseases, including chronic obstructive pulmonary disease (COPD). COPD incidence and prevalence increase with age[145–148]. The disease is characterized by impaired lung repair and progressive distal lung tissue destruction (emphysema) and airways remodeling and inflammation (chronic bronchitis)[149]. First, we computed the Wasserstein distance between pairs of smokers and non-smokers from different age groups *across all genes*. The Wasserstein distance is a measure of the difference between two probability distributions and provides a notion of how "close" the gene expression of two populations for a given cell type is. We found that for 12 out of 18 cell types, the expression space of young smokers (< 30) was closer in distribution to that of old non-smokers (≥ 50) than that of young non-smokers (Fig. 2.11A).

Given that our SenSet signature was based on non-smokers, we next aimed to investigate whether SenSet is altered upon smoking. We performed DE testing on a few cell types of in-

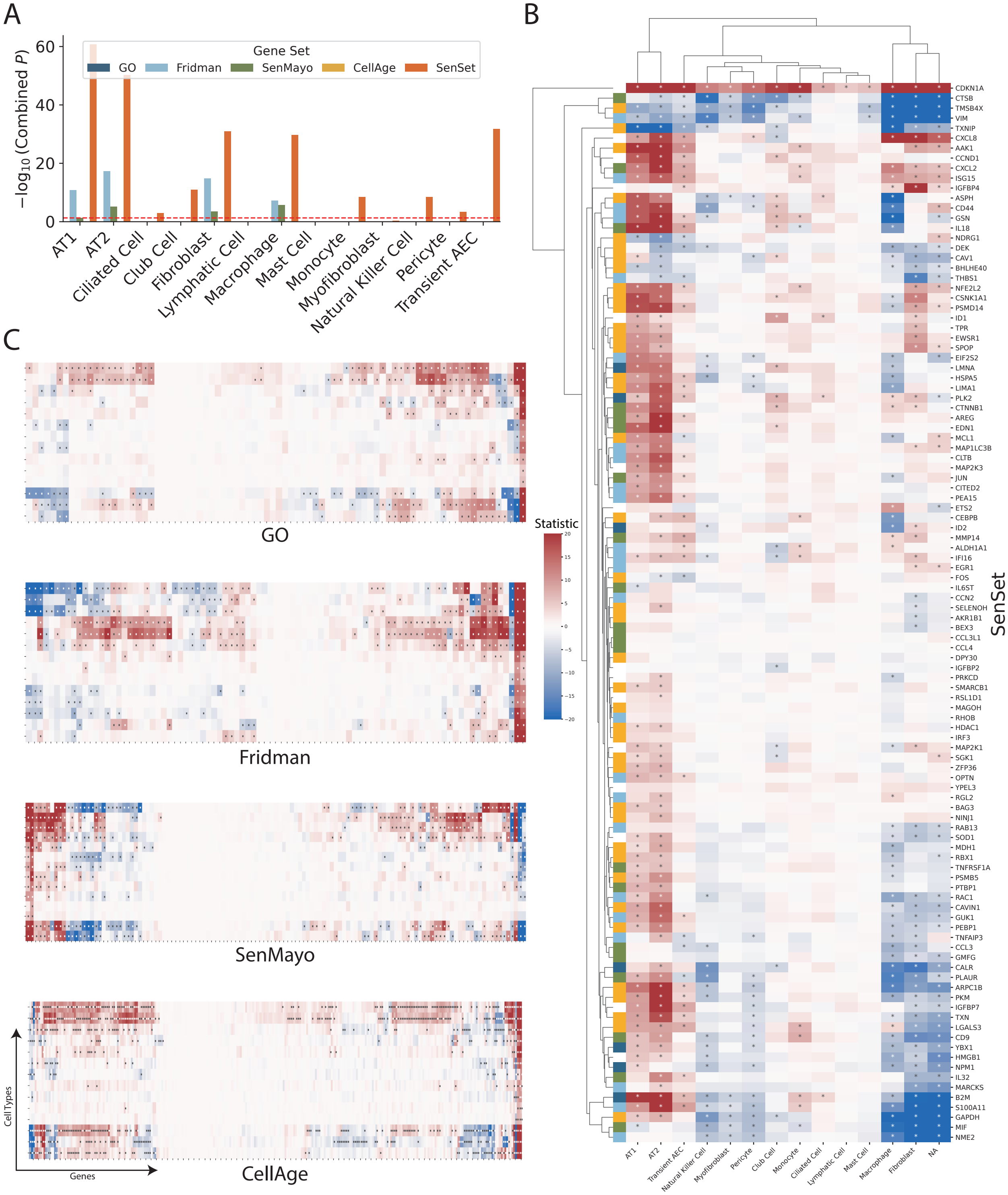


Fig. 2.10: **Cell Type-specific signatures**. (A) For each cell type and gene set, we ran DE tests between the two conditions and show combined p-values (Pearson's method) for each marker gene. (B-C) Rank sum test statistics for every marker gene and cell type in PCLS.

terest to see if senescence markers were enriched in smokers for the youngest $\mathcal{Y}$ and oldest $\mathcal{A}$ age groups. The analysis showed that there are differences in senescence enrichment between smokers of varying ages (Fig. 2.11B). Specifically, we found that most markers were upregulated in smoker basal cells (around 80% of the genes across all lists except for SenMayo at 50%). For other cell types, including CD4/CD8-positive, alpha-beta T cells and type II pneumocytes, we found that aged smokers showed an upregulation of more senescence markers than young smokers, across all markers lists, with SenSet containing most such genes. Lastly, type II pneumocytes showed a change in the regulation of marker genes with age, with most genes being downregulated in the young, but upregulated in the older age group.

Discussion

This study presents a machine learning-based framework to identify a novel gene list specific to senescent cells (SnCs) in the Human Lung Cell Atlas (HLCA). By analyzing single-cell transcriptomic data, our Positive-Unlabeled (PU) learning approach helped us characterize SnCs across age groups. Differential expression tests between SnCs and non-SnCs led to the generation of SenSet, which we validated in an ex vivo human lung model induced to undergo senescence.

Aging remains the strongest risk factor for chronic lung diseases, with SnCs accumulating in tissues and contributing to pathology through mechanisms such as extracellular matrix remodeling and pro-inflammatory signaling. Cellular senescence is induced by diverse stimuli, including oncogene activation associated with tumor suppressor inactivation[150], oxidizing agents inducing DNA damage[151], or chemotherapeutic agents, such as bleomycin and doxorubicin[152,153], with pathways varying by cell type, inducer, and time course. Hallmarks such as the DNA damage response (γH2AX activation), cyclin-dependent kinase inhibitors (p16, p21), SASP (mTOR, cGAS–STING, NF–κB), and apoptosis resistance (BCL-2)[154] define SnCs, yet these features are not unique to senescence, often overlapping with other cellular states.

Our findings revealed that fibroblasts and basal cells were assigned a high proportion of SnCs, accounting for 44 % and 39 % of all cells of that type in the oldest age group $\mathcal{A}$. Fibroblasts play a pivotal role in lung repair and extracellular matrix production but are also central to the pathogenesis of age-related diseases such as idiopathic pulmonary fibrosis (IPF)[155]. Their SASP is known to be pro-fibrotic, exacerbating tissue damage and scarring. Previous studies have shown that senolytic treatments targeting SnCs can alleviate fibrosis in mouse models[156]. Similarly, basal cells, which serve as progenitors for cells in the proximal airways, such as secretory and ciliated cells[157], showed significant SnC-associated gene regulation. These cells are critical for maintaining epithelial integrity and repair, particularly after injury. The number of basal cells gradually decreases in the proximal-distal axis in airway epithelium.

Alveolar type 2 (AT2) cells, essential for surfactant production and alveolar repair, exhibited notable senescence-associated changes. Although only 0.27 % of AT2 cells in healthy HLCA samples were identified as senescent (Fig. 2.7), their transcriptional changes after senescence induction were among the most pronounced, with most SenSet genes upregulated after senescence induction in human PCLS. In contrast, fibroblasts were the main cell type with most downregulated SenSet genes after senescence induction in PCLS (Fig. 2.10).

Several genes were consistently upregulated across fibroblasts$^{(-)}$, basal$^{(-)}$ cells, and AT2$^{(-)}$ cells, emphasizing their shared senescence-associated pathways. Among the 19 upregulated genes in fibroblasts$^{(-)}$, 8 genes are also upregulated in basal$^{(-)}$ cells and/or in AT2$^{(-)}$ cells such as *TNFRSF1A*, *YPEL3*, *SPOP*, *ZFP36*, *CITED2*, and *MARCKS* (Fig. 2.7G). Of these, *TNFRSF1A* encodes a receptor mediating inflammatory cytokine production, while *YPEL3*, a downstream target of p53, plays a critical role in inducing senescence[158]. *SPOP*, a tumor suppressor frequently mutated in cancers, was also upregulated and has been implicated in senescence induction and myofibroblast activation[159,160]. *ZFP36*, a gene regulating inflammatory cytokine production and metabolic pathways[161,162], showed consistent expression patterns across SnCs, reflecting its broad involvement in cellular stress responses. *ZFP36* is known to be induced in cellular senescence in fibroblasts and in different human tissues[163]. Similarly, *CITED2*, which modulates TGF-β signaling[164], was associated with cellular proliferation and senescence, with its reduced expression linked to aging and senescence in tendon-derived stem cells[165]. *MARCKS*, an actin-binding protein involved in cell motility and secretion, was also highly expressed in these cells.

A total of 19 genes were consistently downregulated in both fibroblasts$^{(-)}$ and basal$^{(-)}$ cells. For instance, *IL32*, a pro-inflammatory cytokine, was reduced in both, aligning with its role in immune regulation and oxidative metabolism[166–168]. *YBX1*, which regulates SASP translation and senescence markers[169], displayed similar downregulation.

In contrast, some genes were downregulated in fibroblasts$^{(-)}$ while being upregulated in other cell types, highlighting cell type-specific senescence responses. For instance, *PLAUR*, involved in fibroblast-to-myofibroblast differentiation[170], was downregulated in fibroblasts$^{(-)}$ but upregulated in basal$^{(-)}$ cells. Interestingly, *PLAUR* was identified as an upregulated gene in datasets from murine and human SnCs. *PLAUR* encodes for urokinase plasminogen activator receptor (uPAR), and treatment with uPAR-directed CAR T cells was a good senolytic strategy to decrease SnCs in vivo and in vitro[171]. Other genes included *ID2*, which is known to antagonize the growth-suppressive activities of p16 and p21[172], and *TNFAIP3* which encodes for TNF-α induced protein 3 (A20). In mice, fibroblast A20 deletion recapitulate major pathological features of systemic sclerosis[173].

TXNIP, associated with oxidative stress responses, was upregulated in both basal$^{(-)}$ cells and AT2$^{(-)}$ cells. *TXNIP* was shown to have a role in cellular senescence; its expression increases with age in β-cells and serum samples from humans, and it aggravates age-related and obesity-induced structural failure associated with an induction of cell cycle arrest and oxidative stress[174]. Other genes included *FOS* and *JUN*.

Validation of the SenSet gene list in an ex vivo model confirmed the induction of hallmark senescence markers, including p21. We also performed a multiplex assay of 47 secreted proteins, of which 28 were detectable in the PCLS supernatants. Among secreted proteins, GDF-15 exhibited the largest increase (Fig. 2.8G), aligning with its established role as an age-associated marker and stress-responsive factor[175–177]. Other validated genes included *ALDH1A1*, linked to SASP regulation and senescence in cancer stem cells[178], and *PLK2*, a kinase implicated in senescence pathways with reduced expression in glioblastoma[179]. *CNN2*, known to promote fibroblast senescence[180], further supported the relevance of these markers in defining SnCs. Interestingly, *TXNIP*, which was upregulated in 11 cell types$^{(-)}$ in HLCA, was downregulated in the ex vivo model, indicating its inducer-specific role in oxidative stress-mediated senescence. Similarly, genes such as *NME2* and *NPM1*, which play a role in cell cycle and tumor supression[132,133], were consistently downregulated in both datasets. This finding aligns with a prior study that *NPM1* upregulation inhibits p53-mediated senescence[181].

Our findings underscore how environmental factors interact with aging processes to drive senescence in specific lung cell types. Despite these insights, the study has limitations. The assumption that cells from individuals younger than 30 are universally healthy may not hold true in all cases, though the PUc classifier was robust against potential contamination by SnCs. Additionally, the use of arbitrary age thresholds (30 and 50) could influence the PUc classifier, although the framework remains valid as long as healthy cells in patients younger than 50 are similar. Recent studies have shown that significant developmental dysregulation occurs at the ages of 44 and 60[182], leading us to believe that most healthy cells in this age range (30-50) are similar to those in the youngest group. The upper threshold was set to 50 in order to expand the set of individuals for this age group. Furthermore, the identification of marker genes is performed using one group only (ages 50+), and not by comparing age groups against each-other, so this is not a major limitation of the subsequent analysis.

Some cell types such as tracheobronchial serous cells were assigned a large fraction of SnCs (76 %) by the PUc learner (Fig. 2.7A). Only 1 marker was assigned to these cells which may reflect mislabeling or insufficient data. Finally, only 18 of the 31 cell types analyzed using the PUc framework contained no SnCs in $\mathcal{Y}$, suggesting that the method is robust to a small number of SnCs in the young.

In cell type-specific analysis, certain cell types, such as lymphatic cells and mast cells, showed minimal or no enrichment for SenSet. This could be attributed to the small size of their clusters. Further investigation will be necessary to establish a true lack of enrichment for senescence in these cell types.

Furthermore, to establish a mechanistic link between SenSet genes and the senescent state, additional experiments may need to be performed. This is challenging as we wish to validate these putative senescence markers, but we lack a definitive "gold standard" for senescence itself. However, we can use several widely accepted senescence assays as a reference standard. For example, following functional manipulation of the candidate genes—such as CRISPR/Cas9-mediated gene knockout, or overexpression experiments—we can measure senescence-associated-β-Gal activity which SnCs are typically enriched for. Morphological changes are another potential reference, with SnCs often becoming enlarged and flattened. These are crude characteristics but can be useful phenotypic cues. Finally, measuring SASP factors, such as *IL6* or *IL8* could provide another important functional dimension to senescence identification.

In conclusion, this study advances our understanding of SnCs in the lung by identifying cell type-specific senescence markers using a robust machine learning framework. The validation of these findings in ex vivo models strengthens their relevance, offering a foundation for future research into the role of SnCs in aging and chronic lung diseases. By linking cellular senescence to environmental stressors like smoking, this work highlights potential targets for therapeutic interventions aimed at mitigating the detrimental effects of senescence on lung health.

2.2.3 Materials and Methods

The Human Lung Cell Atlas

The (core) Human Lung Cell Atlas (HLCA)[124] was downloaded from the humancellatlas portal. Counts were already normalized. The HLCA harmonizes scRNA-seq data from 14 datasets, encompassing 106 individuals aged between 10 and 76 years. We removed one individual for whom the age was not available. While five levels of annotation are available in the data, we used the finest level assigning one of 50 cell types to over 500,000 cells for the analysis in this study. The dataset also contains individuals with a smoking history, including 19 former and 28 active smokers. Smoking status is not available for 8 individuals and these were not included in the analysis (Fig. 2.6).

Deriving SenSet from the Human Lung Cell Atlas

During the gene set generation step, we kept only individuals without a smoking history from the HLCA (approximately 300,000 cells), to minimize potential confounding effects on the results. As described in the previous section, the PUc estimator requires three sets of samples: positive training samples, and unlabeled training and test samples. We estimated positive samples from individuals

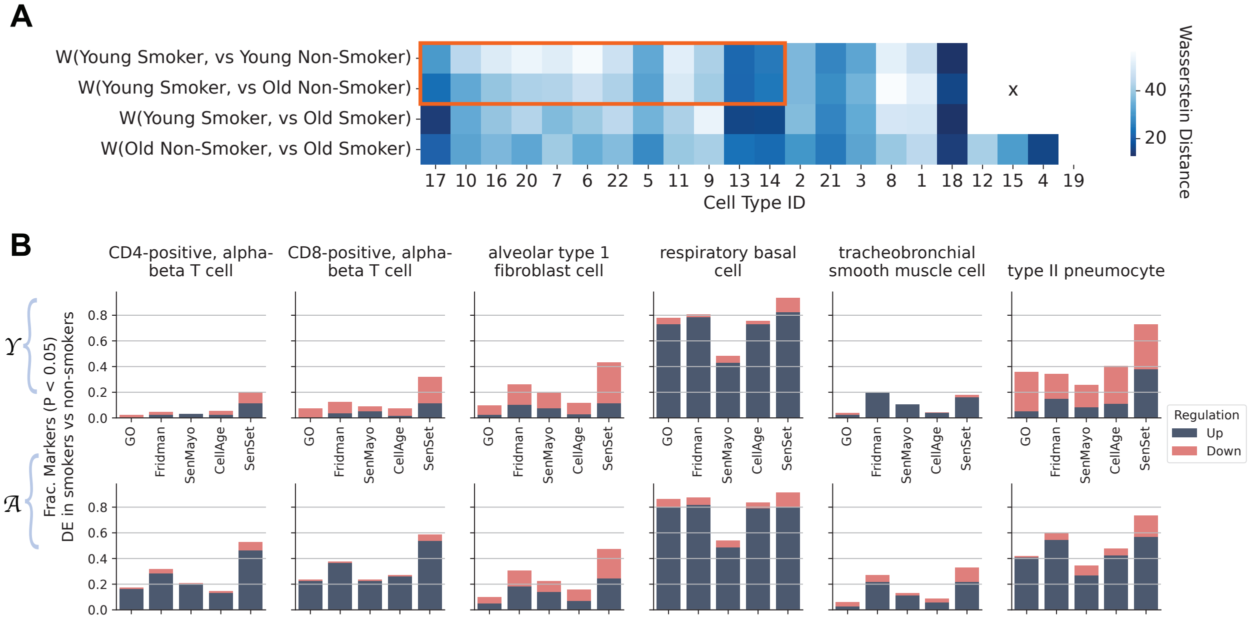


Fig. 2.11: **Comparisons of senescence marker genes between smokers and non-smokers in the HLCA.** (A) Wasserstein distance between the gene expression profiles of smokers and non-smokers across different age groups. Cell types inside the red box exhibited a smaller distance for the pair (young smokers, old non-smokers) when compared to (young smokers, young non-smokers). (B) Fraction of genes enriched in smokers compared to non-smokers among young ($\mathcal{Y}$) and old ($\mathcal{A}$) patients for selected cell types. (C-D) STEM significant profiles and the corresponding gene curves, categorized by smoking status and sex. Genes from significant profiles were combined into one plot.

under the age of 30, assuming that the prevalence of SnCs in this group is minimal. The unlabeled training samples were estimated from individuals aged 30 to 50. Recent studies have shown that significant developmental dysregulation occurs at the ages of 44 and 60[182], leading us to believe that most healthy cells in this age range (30-50) are similar to those in the youngest group. The upper threshold was set to 50 in order to expand the set of individuals for this age group. Finally, the test samples were obtained from older individuals aged 50 and above, which is when covariate shift occurs. We call these three age groups $\mathcal{Y}, \mathcal{M}$, and $\mathcal{A}$, respectively (Fig. 2.6A, E). Cell types with fewer than 50 cells in any age group were excluded from the analysis.

To prepare the data for PUc learning, we first applied principal component analysis (PCA) independently for each cell type. We restricted the gene set used for PCA to the union $\mathbf{U}$ of all four existing senescence gene sets. By incorporating all known senescence-associated genes, we aim to achieve a "weak" separation of healthy cells and SnCs, which can be leveraged by the PUc learner. The top 10 components were used as training data for the PUc classifier. For all experiments, we set the mixture proportion α to 0.9, based on the prior assumption that approximately 10% of the cells are senescent. However, the estimator was robust to this value and returned percentages in the range $0 - 40\%$.

Differential expression analysis is performed exclusively on the oldest age group, comparing healthy cells with SnCs. This approach is in contrast with methods that compare old and young individuals, where other aging signatures could introduce confounding factors. By directly comparing these two cell populations within the oldest age group, the analysis is specifically focused on senescence. A two-sided Wilcoxon rank-sum test was used to determine differentially expressed (DE) genes ($\text{FDR} = 0.05$). We tested only genes that belong to $\mathbf{U}$. FDR-adjusted p-values were obtained using the Benjamini & Hochberg procedure[183]. Cell types with fewer than 20 SnCs were not considered due to the limited sample size. We selected DE genes that were enriched in at least six cell types (either up or downregulated), resulting in a set of 106 genes that constitute SenSet. A detailed table of all cell-specific SenSet genes and their test statistics is provided in the supplementary material.

Gene set enrichment analysis (GSEA)[184] was performed using the GSEApy package[185]. To compute the Wasserstein distances in Fig. 2.11A, we sampled at random 5000 cells for cell types with too many cells to speed up computation.

Integrating PCLS data and validating SenSet

For the overall sample comparison presented in Fig. 2.9B-C, we performed basic cell and gene filtering for all 11 samples. Cells with fewer than 500 total counts and fewer than 400 expressed genes were excluded. Genes with fewer than 50 total counts were also removed. Next, we normalized the total counts of cells. Normalized gene counts between treatment and control samples were compared using a two-sided Wilcoxon rank-sum test ($\text{FDR} = 0.05$) for each gene set.

For the cell-specific analysis, we first integrated the data using scVI[186] focusing on the top 2000 variable genes. We used 2 hidden layers with 1000 nodes each. The dimensionality of the latent space was set to 30. Since we used raw counts for scVI, the gene likelihood was modeled as a zero-inflated negative binomial distribution. Nearest neighbors were computed in the scVI latent space, and clusters were identified via Leiden clustering[187]. Clusters were manually annotated based on canonical markers of lung cell types[144]. Clusters where no marker was significantly expressed were excluded from the cell-specific analysis. Wilcoxon rank-sum tests were used to assess whether SenSet genes were enriched in treatment samples compared to controls. For this analysis, bleomycin, doxorubicin, and irradiation cells were combined into one group.

For each gene set, we obtain a list of p-values for each gene based on the DE test. In order to perform a meta-analysis, we combine these p-values for each set using Pearson's method, which emphasizes larger p-values. I.e., given a set of p-values $\{p_i\}_{i=1}^n$, Pearson's method computes the statistic

$$P := -2\sum_{i=1}^{n}\log(1-p_i). \tag{2.7}$$

Under the null hypothesis $H_0 : p_i \sim U[0,1], i = [n]$, the test statistic P follows a χ^2 distribution with $2n$ degrees of freedom[188]. The combined p-value provides an overall assessment of the enrichment of a gene set within a given cell type.

PCLS culture and senescence induction (Bleomycin and Doxorubicin)

Lungs from healthy donors (ages 20-78) were collected, and the lower left lobes were inflated with 2.5 % agarose in DMEM/F-12 with HEPES (Thermo Fisher Scientific Cat#12400024). After 45 minutes on ice, the lobes were sliced, and 1 cm diameter cores were extracted. Precision-cut lung slices (PCLS) were prepared using a Compresstome® to obtain 300 µm thick tissue slices with a diameter of 1 cm. The PCLS were then placed in a 24-well plate containing 1 mL of DMEM/F-12 medium supplemented with 1 % FBS, 1 % Penicillin-Streptomycin (Merck MilliporeSigma, Sigma-Aldrich Cat#P00781), and 0.3 µg/mL Amphotericin B solution (Merck MilliporeSigma, Sigma-Aldrich Cat#A2942) and incubated for 24 hours at 37 °C with 5 % CO2 (day -1).

At 24 hours (day 0), 72 hours (day 2) and 120 hours (day 4), the medium was replaced with fresh medium containing treatments diluted in DMEM/F-12 with 0.1 % FBS, 1 % Penicillin-Streptomycin, and 0.3 µg/mL Amphotericin B. The PCLS were treated under the following conditions: in PBS as the control, with 15 µg/mL bleomycin (Fresenius Kabi Cat#10361), with DMSO diluted at 1:100,000 in medium (Merck Millipore Sigme, Sigma-Aldrich, Cat#D2438), or with 0.1 µM doxorubicin hydrochloride (Merck Millipore Sigma, Sigma-Aldrich, Cat#D1515-10MG), originally dissolved in DMSO at 10 mM. After 168 hours (day 6), the supernatants were collected and frozen at −80 °C for future multiplex immunoassay analysis by Luminex™. The PCLS were snap-frozen in liquid nitrogen for future protein extraction and snRNA-seq using the 10x Genomics™platform. Additionally, PCLS were fixed for 30 minutes in 4 % formaldehyde diluted in PBS (Life Technologies Cat#28908), washed in PBS, and embedded in paraffin. Separate PCLS were fixed for 30 minutes in the fixative solution from the β-galactosidase staining kit (Cell Signaling Technology Cat#9860) and then washed in PBS.

PCLS culture and senescence induction (Irradiation)

Peritumor control tissue from non-chronic lung diseases (N-CLD) patients were obtained from the CPC-M bioArchive at the Comprehensive Pneumology Center (CPC Munich, Germany). Patients were male. Human lung tissue was filled with 3 % of low gelling temperature agarose in DMEM/F-12 (Thermo Scientific, USA) with phenol red supplemented with 0.1 % FCS, 1 % P/S and 1 % amphotericin B and kept at 4 °C for at least 1 hour. 500 µm PCLS were generated using either a vibratome HyraxV50 (Zeiss, Germany) or 7000smz-2 Vibratome (Campden Instruments, England). The day after slicing, fresh medium was added and PCLS were exposed to ionizing radiation using the RS225 X-ray cabinet (Xstrahl, Camberley, UK). Dose was calculated according to exposure time (30 Gray (Gy) = 12 min 24 sec) at 195kV and 15mA. Then, PCLS were kept in culture for up to 5 days at 37 °C, 5 % CO2 and medium was changed every 2-3 days.

Ethic Statement

The study was approved by the local ethics committee of the Ludwig-Maximilians University of Munich, Germany (Ethic vote 19-630). Written informed consent was obtained for all study participants.

Histology and immunohistostaining

Paraffin-embedded PCLS were sliced at a thickness of 4 µm. One slide was subjected to H&E staining, and another slide was used for p21 immunostaining. The slides were rehydrated through a series of baths in xylene, followed by 100 %, 95 %, 85 % ethanol, and finally water. The slides were then incubated at 105 °C for 20 minutes in 1X DAKO high pH buffer (Agilent Technologies, Dako Target Retrieval Solution pH 9 10X, Cat#S236784-2). Then the slides were washed in buffer A from Duolink® In Situ Wash Buffers (MilliporeSigma, Sigma-Aldrich Cat#DUO82049-20L), followed by incubation in 300 mM glycine for 30 minutes, and then in PBS containing 0.1 % Tween and 0.5 % Triton for 15 minutes.

The slides were then incubated with a blocking solution consisting of 2 % BSA, 0.1 % Triton, and 0.1 % Tween at 37 °C for 45 minutes, followed by an overnight incubation at 4 °C with the primary antibody, anti-p21 antibody [EPR362] (Abcam, Cat#ab109520), diluted 1:500 in the blocking solution. After washing in buffer A, the slides were incubated for 1 hour at room temperature with a 1:1000 dilution of the secondary antibody, anti-rabbit IgG 647 (Biotium, Cat#20047). Following further washes in buffer B from Duolink® In Situ Wash Buffers, the nuclei were stained with DAPI. The slides were then washed in buffer B and mounted with Fluoroshield Mounting Medium (Abcam, Cat#ab104135). Images were captured using an IX83 Olympus microscope, acquiring the entire PCLS area at 20x magnification.

Quantification of p21-positive cells was performed using Fiji macros, where the total cell number was determined by DAPI staining, and the number of p21-positive cells was identified by the nuclear p21 signal overlapping with DAPI staining. The percentage of p21-positive cells was calculated as the ratio of p21-positive cells to the total number of cells in each image.

Protein extraction and western blotting

Four PCLS were sonicated three times at 35 % amplitude for 10 seconds in 200 µL of buffer (TPER buffer, Thermo Scientific, Cat#78510, supplemented with Halt^{TM} Protease and Phosphatase Inhibitor Cocktail, EDTA-free (100X), Thermo Scientific, Cat#1861281, to a final concentration of 1X) and kept on ice. The samples were then homogenized for 15 seconds using a Thermo Fisher Scientific homogenizer and centrifuged at 300×g for 5 minutes at 4 °C. The supernatants were transferred to new 1.5 mL tubes and centrifuged at 10,000 rpm for 10 minutes at 4 °C. The supernatants were gently collected, and protein concentrations were quantified in triplicate using the Pierce detergent Compatible Bradford Assay kit (Thermo Scientific, Cat#23246), with 150 µL of reagent and 5 µL of sample per well. A standard curve was generated using the Prediluted Protein Assay Standards BSA Set (Thermo Scientific, Cat#23208). Absorbance was measured at 595 nm using a spectrophotometer.

Protein extracts were denatured in 1X Protein Loading Buffer (Li-COR, Cat#928-40004) containing 0.1 M dithiothreitol (DTT) (Sigma-Aldrich, Cat#43816) for 10 minutes at 95 °C. A total of 10 µg of denatured protein was loaded onto Criterion TGX long shelf-life Precast Gels (4-15 %, Bio-Rad, Cat#5671083) using 1X Tris/Glycine/SDS buffer (Bio-rad, Cat#1610772). The proteins were then transferred onto an Odyssey Nitrocellulose Membrane (Li-COR, Cat#926-31092).

The membrane was then blocked for 1 hour at room temperature in a 1:1 mixture of 1X TBS and Intercept Blocking Buffer (Li-COR, Cat#927-70001). The blocked membrane was incubated overnight at 4 °C with p21 antibody (Abcam, Cat#ab109520) diluted 1:1000 or GAPDH antibody (Abcam, Cat#ab9485) diluted 1:2500 in a 1:1 mixture of 1X TBST and Intercept Blocking Buffer.

The membrane was then washed three times for 10 minutes each with TBST and incubated with secondary antibodies: for p21, anti-rabbit red (Cat#926-68073) and for GAPDH, anti-rabbit green diluted 1:20000 in a 1:1 mixture of 1X TBST and Intercept Blocking Buffer for 1 hour at room temperature.

Images were captured and analyzed using an Odyssey imaging system with Image Studio software.

SASP assessment by multiplex immunoassay

PCLS supernatants were harvested after six days in culture and stored at −80 °C. The multiplex immunoassay was performed using the Luminex^{TM} platform, following the manufacturer's instructions (Bio-Techne).

Single-nucleus RNA sequencing by 10x Genomics™

PCLS were snap-frozen in cryotubes using liquid nitrogen after six days of culture and stored in liquid nitrogen. Four PCLS per experimental condition were then used for snRNA-seq, following the manufacturer's instructions (10x Genomics™).

2.3 Unsupervised Setting: Tools for Biomarker Discovery and Cell Type Annotation

Based on: *Hasanaj, E.* et al. *Interactive single-cell data analysis using Cellar. en.* Nat. Commun. ***13,*** *1998 (Apr. 2022)*

The previous two sections discuss learning in the presence of either strong or weak labels. However, scRNA-seq datasets lack inherent cell identity information. Researchers must manually annotate cells based on domain expertise and knowledge of cell-specific marker genes. Due to the intricate nature of biological systems, understanding the motive behind every mRNA expressed by a cell is impossible. Consequently, it is unlikely to determine a single cell's identity in isolation solely based on its transcriptome.

Nonetheless, cell types can still be inferred from cell clusters by examining the collective behavior of these cells. Unsupervised approaches are a useful strategy for such analysis. The standard single-cell pipeline begins by reducing the high-dimensional scRNA-seq data to a lower-dimensional space using techniques like PCA. Subsequently, cells are clustered to identify distinct cell populations. Finally, one-vs-all DE tests, as outlined in section 2.1, are employed to discover cluster-specific markers and facilitate cell type annotation. This approach has spurred the development of numerous computational packages that play an indispensable role in navigating this pipeline. These packages offer a diverse array of tools, encompassing essential functionalities such as data quality control, clustering, visualization, and marker discovery.

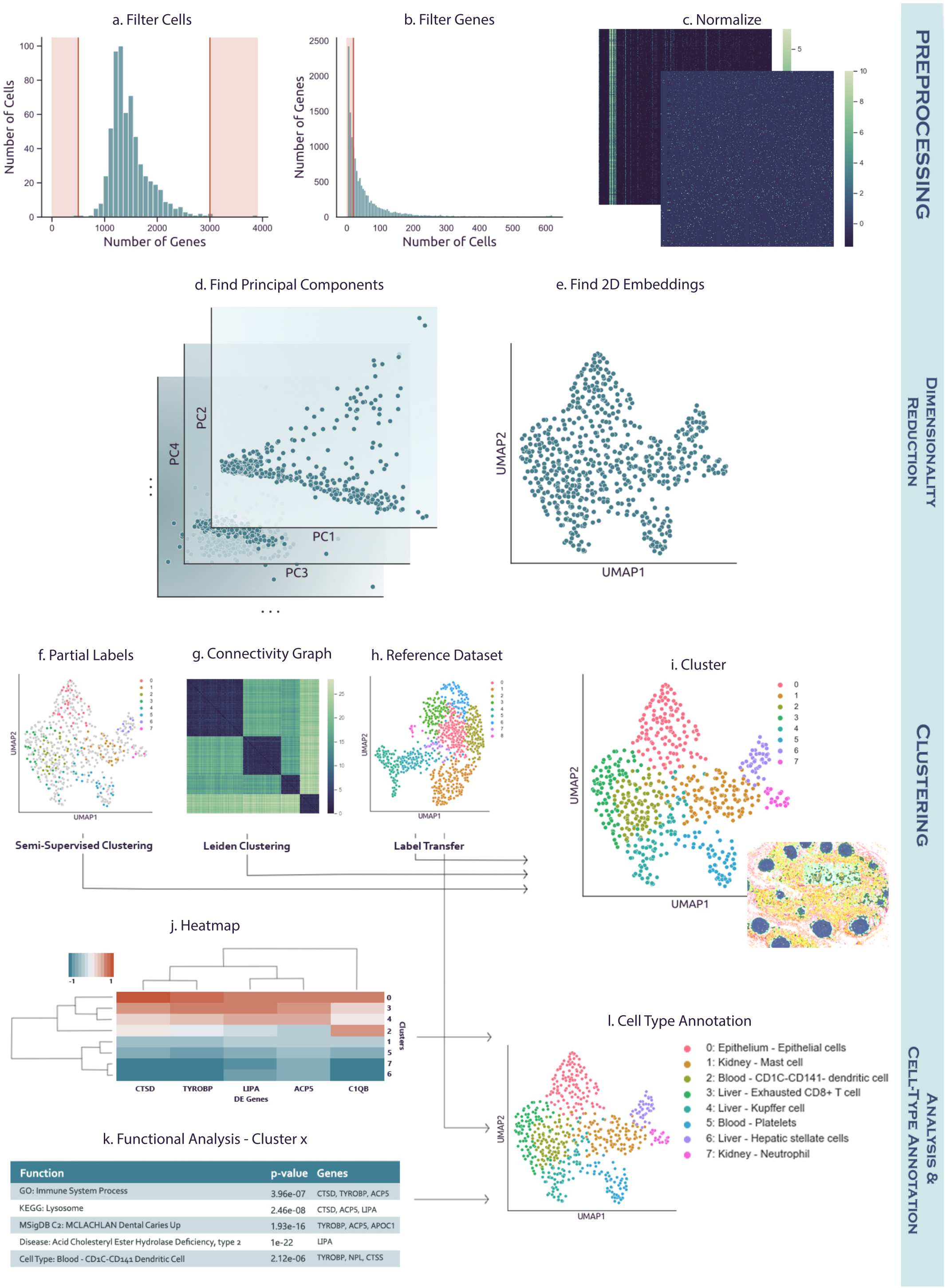
The proliferation of single-cell analysis tools has ushered in an era of abundance and, paradoxically, uncertainty. Within the scientific community, different research groups—whether part of the same consortia or distinct entities—often adopt different sets of tools when handling multiple types of single-cell data. As a consequence, integrating and comparing data across groups becomes challenging as researchers use different assignment techniques, markers, and even cell-type naming conventions. Moreover, each tool has its own unique programming interface which poses an additional barrier for biologists and researchers with limited programming experience.

To facilitate large-scale collaborations, seamless integration, and comparisons across diverse single-cell omics platforms and modalities, we developed Cellar: an interactive and graphical cell type assignment web server. Cellar implements a comprehensive suite of methods, both existing and new, to address every facet of the cell type assignment process. These include methods for dimensionality reduction and representation, clustering, reference-based alignment, identification of DE genes, intersection with functional and marker sets, as well as a dual mode for analyzing and comparing two datasets simultaneously (Fig. 2.12). Cellar includes a manually-curated marker set for many different cell types, as well as several other functional gene sets.

As cell type assignment often requires user input in the form of domain knowledge, Cellar adopts a semi-automatic solution that acknowledges the value of user expertise. This methodology allows users to intervene and tailor each processing step as necessary, ensuring flexibility and precision. To enable such interactive analysis, Cellar incorporates methods for semi-supervised clustering and projection of expression clusters in spatial single-cell images.

Cellar has been tested by members of HuBMAP[2] and used to annotate several single-cell datasets from different organs, platforms, and modalities.

Cellar is open-source and includes several public datasets. We provide below some examples of typical workflows in Cellar.



Function	p-value	Genes
GO: Immune System Process	3.96e-07	CTSD, TYROBP, ACP5
KEGG: Lysosome	2.46e-08	CTSD, ACP5, LIPA
MSigDB C2: MCLACHLAN Dental Caries Up	1.93e-16	TYROBP, ACP5, APOC1
Disease: Acid Cholesteryl Ester Hydrolase Deficiency, type 2	1e-22	LIPA
Cell Type: Blood - CD1C-CD141 Dendritic Cell	2.12e-06	TYROBP, NPL, CTSS

Fig. 2.12: **Cellar's workflow.** (a-c) Preprocessing (optional). (d, e) Dimensionality reduction and visualization. Several methods for dimensionality reduction are implemented as part of Cellar. The reduced data is then visualized by running another (possibly the same) dimensionality reduction method. (f-i) Clustering. Cellar supports several unsupervised and semi-supervised clustering methods. It also implements supervised label transfer methods. (j-l) Cell type assignment. Cellar enables the use of several functional annotation databases for the assignment of cell types.

2.3.1 Typical Workflows in Cellar

Analysis of scRNA-seq Data

We used Cellar to analyze 11 HuBMAP scRNA-seq datasets (10x Genomics™), with an average of 7,500 cells per dataset from five different tissues: kidney, heart, spleen, thymus, and lymph node[2]. Each of these datasets is available within Cellar. The Cellar pipeline begins with quality control, filtering out unreliable cells and low-count genes. Additional normalization and scaling are then applied based on user-defined criteria. Following this, Cellar clusters a lower-dimensional representation of the data, and further reduces dimensions for visualization.

To demonstrate this pipeline, we analyzed a spleen dataset containing 5,273 cells (Cellar ID: HBMP3-spleen-CC2). We applied PCA, followed by UMAP[190] for dimensionality reduction, and the Leiden algorithm for clustering[187], resulting in 16 distinct clusters. For each cluster, Cellar identified the top differentially expressed (DE) genes. Using the top 500 DE genes, functional enrichment analysis was conducted using GO, KEGG, and MSigDB gene sets[125,135,191]. This analysis identified cluster 0 as B-cells, with "B-Cell Activation" ($q = 0$) and "B-Cell Receptor Signaling Pathway" ($q = 0$) emerging as the top categories in GO and KEGG, respectively. This classification was further supported by visualizing the concurrent expression of known B-cell markers, including *CD79A* and *TNFRSF13C*[192,193].

In addition to unsupervised clustering, Cellar offers methods for supervised cell-type assignment based on reference datasets, enabling direct utilization of Cellar's dual mode and other classification methods. For example, this functionality can be combined with Cellar's semi-supervised clustering to reduce noise during the label transfer process. To illustrate this, we used Scanpy's Ingest function[194], available in Cellar, to integrate two expert-annotated spleen datasets (Cellar IDs: HBMP2-spleen-2 and HBMP3-spleen-CC3). Using HBMP3-CC3 as a the reference, we transferred labels from it to HBMP2-2 and compared the results with the ground truth annotations for HBMP2-2. The label transfer achieved an adjusted Rand index (ARI) of 0.39, whereas clustering HBMP2-2 with Leiden alone resulted in a lower ARI score of 0.27. We then refined the label transfer results using a semi-supervised adaptation of the Leiden algorithm, in which low-noise clusters were fixed as constraints and preserved during future iterations of the algorithm. This approach achieved a much-improved ARI score of 0.66, demonstrating the advantage of label transfer combined with semi-supervised clustering. Full details of these results are provided in the supplementary file linked in the paper above.

Analysis of ATAC-seq Data

While scRNA-seq remains the most widely used modality for single-cell data, other molecular data types are also being profiled at the single-cell level. To illustrate Cellar's flexibility, we used it to annotate an ATAC-seq dataset[195]. Cellar supports ATAC-seq data in two formats: cell-by-gene and cell-by-cistopic. The cell-by-gene format is based on open chromatin accessibility linked to regions near each gene, whereas the cell-by-cistopic format uses cisTopic[196], which applies Latent Dirichlet Allocation (LDA)[197] to identify cis-regulatory topics.

The resulting cell-by-gene or cell-by-cistopic matrices can then be used for downstream analyses such as visualization and clustering. We used Cellar to annotate an ATAC-seq dataset profiling Peripheral Blood Mononuclear Cells (PBMCs)[198] (Cellar ID: PBMC 10k Cell-By-Gene) using the cell-by-gene representation. DE analysis of clusters 0 and 4 identified *KLRD1* as a marker for natural killer (NK) cells[199].

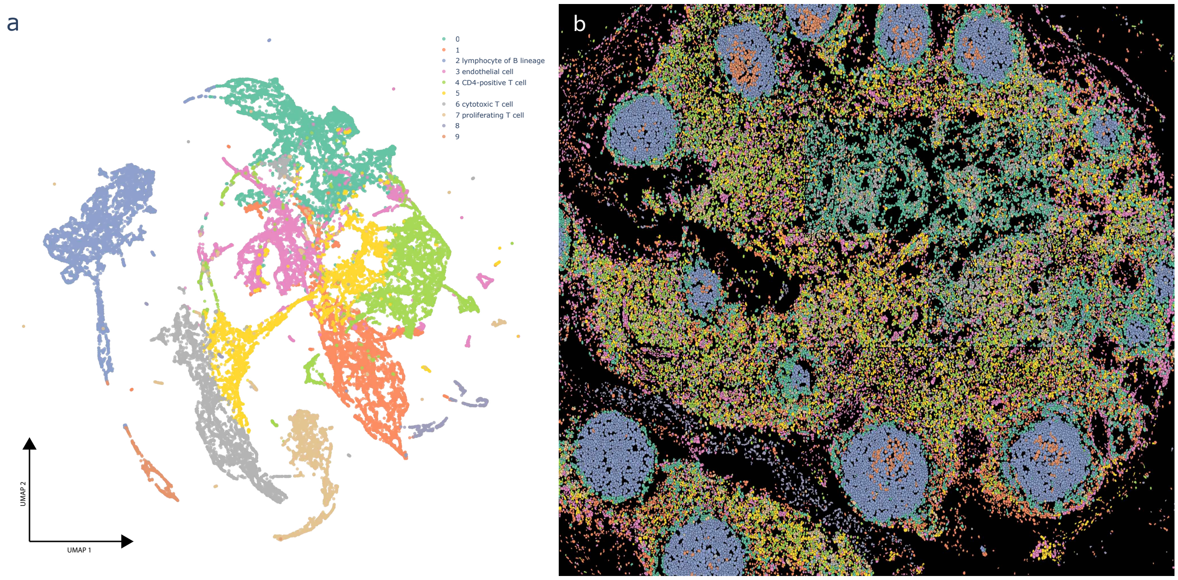


Fig. 2.13: **CODEX data analysis in Cellar.** (ID: 19-003 lymph node R2) **a** UMAP visual representation of a lymph node CODEX dataset with 46,840 cells, clustered via Leiden. **b** Projection of the assignments on the spatial CODEX image that can be visualized side-by-side in Cellar. Cluster assignments were copied from a. Not all clusters could be assigned to unique cell types given that only a few ten protein levels are measured, though several have been assigned based on differential gene analysis in Cellar. The B-Cell clusters are surrounded by T-cells and other cells types in the lymph. The B-Cell clusters also contain a subset of proliferating cells.

Analysis of Spatial Proteomics Data (CODEX)

In addition to sequencing assays, recent imaging assays now enable the measurement of gene or protein expression at the single-cell level. Cellar can be used to analyze such imaging data, providing a side-by-side view of expression clusters and their spatial organization. To demonstrate this, we analyzed spatial proteomics data generated using co-detection by indexing (CODEX)[35]. Specifically, we used a lymph node dataset containing 46,840 cells (Cellar ID: 19-003 lymph node R2). Clustering results are shown in Fig. 2.13 alongside a spatial tile for these cells, with projected cluster annotations.

Given the limited number of proteins profiled in this dataset (19), not all clusters could be assigned unique cell types, however, several clusters were identified based on DE analysis in Cellar. Cellar maintains consistent cell color mapping across clustering and spatial images, facilitating the identification of spatial organization patterns and their relationship to specific cell types. In the spatial tile, B cells are shown to cluster tightly together, surrounded by T cells and other cell types within the lymph node. Additionally, the B-cell clusters include a subset of proliferating cells.

Joint Analysis of Multiple Modalities

Finally, we used Cellar to jointly analyze data from two different modalities. For this, we used a SNARE-seq dataset[200] from kidney, which profiled both the transcriptome and chromatin accessibility of 31,758 cells (Cellar IDs: kidney SNARE ATAC/RNA 20201005). In this analysis, we first ran cisTopic on the chromatin accessibility data to identify cis-regulatory topics and then applied

Leiden clustering to these inferred topics to assign clusters (Fig. 2.14a). These cluster labels were then used to visualize the corresponding expression data in Fig. 2.14b. Cellar's dual mode enables this seamless transfer of cell IDs across modalities.

Cellar identified DE genes within each cluster, which were used to map cell types. For example, cluster 1 was identified as Proximal Tubule Cells based on known markers, such as *SLC5A12*, ($p=0$), and GO term analysis, with "Apical Plasma Membrane" ($p=1\text{e}{-4}$) further supporting this assignment[201],[202].

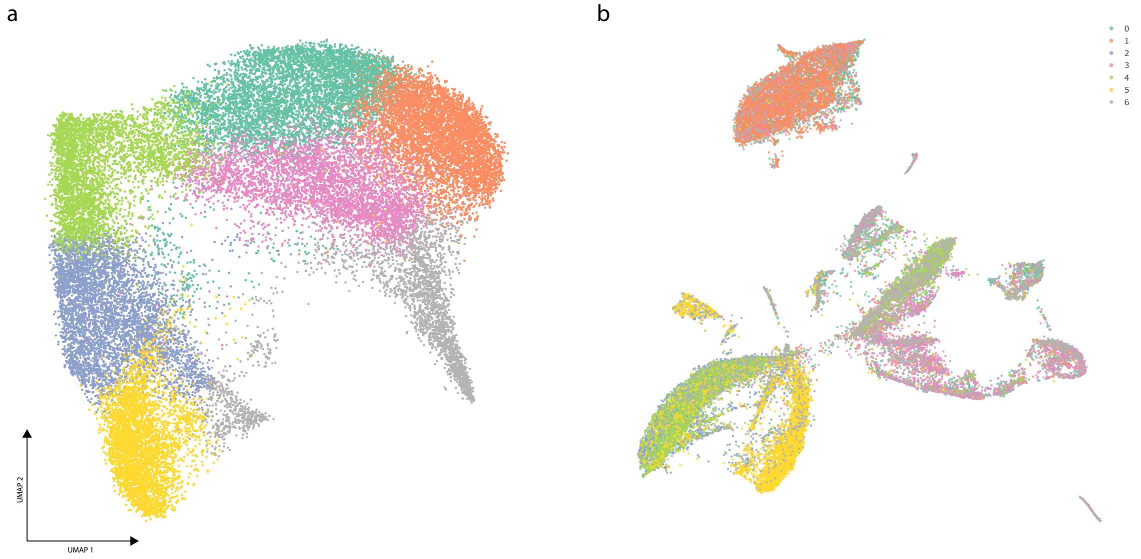


Fig. 2.14: **SNARE-seq data analysis in Cellar.** (IDs: kidney SNARE ATAC/RNA 20201005) **a** UMAP plot of the chromatin modality for the kidney SNARE-seq dataset with 31,758 cells. First, we obtain a cell-by-cistopic matrix by running cisTopic which is then used to define clusters via Leiden clustering. **b** Corresponding UMAP plot of the expression matrix with cluster assignments copied from a. Cellar's dual mode allows a cell ID based label transfer from one modality to the other.

2.3.2 Materials and Methods

Preprocessing

Data preprocessing was conducted using Scanpy[194]. For all scRNA-seq datasets, we filtered out cells with fewer than 50 or more than 3000 expressed genes. Genes expressed in fewer than 50 or more than 3000 cells were also excluded. The data matrix was then counts per million (CPM) normalized, with a total count of 1e5 , followed by a log1p transformation. Finally, we scaled the data to unit variance and zero mean.

For the PBMC ATAC-seq dataset, we generated a gene activity score matrix by summing peaks that intersect the genomic regions around each gene as defined in GENCODE v35[203]. Gene ranges were extended by 5000 base pairs downstream and 1000 base pairs upstream. The resulting cell-by-gene matrix was then normalized and log1p-transformed using the same procedure as for the scRNA-seq data.

No normalization was applied to the CODEX data.

Clustering, Visualization, and Functional Analysis

The scRNA-seq and gene activity matrices were reduced to a 40-dimensional space using PCA. We applied the PCA implementation from the scikit-learn package with a randomized SVD solver[204]. For the lymph node CODEX data, we performed dimensionality reduction with UMAP[190] to 10 dimensions, using the umap-learn Python package.

The resulting embeddings were used to construct an approximate neighbors graph with 15 neighbors, using the faiss library[205]. Clustering was then performed using the Leiden community detection algorithm[187] with a default resolution of 1. For the lymph node CODEX data only, we used a lower resolution of 0.1 to achieve a reasonable number of clusters. All datasets were further reduced to 2 dimensions with UMAP for visualization purposes.

DE analysis was conducted with diffxpy (`https://github.com/theislab/diffxpy`) using a Welch's t-test. The 500 DE genes with the highest fold-change values were selected for enrichment analysis using the GSEApy package[185], which implements the GSEA method[184]. For the CODEX data, where fewer than 20 channels were available, we used all DE proteins for enrichment analysis.

Label Transfer and Semi-supervised Clustering

Label transfer between HBMP2-spleen-2 and HBMP3-spleen-CC3 was performed using Scanpy's Ingest function (`https://scanpy.readthedocs.io/en/stable/generated/scanpy.tl.ingest.html`). Ingest projects the query dataset into a latent space fitted on the reference data, using PCA with 40 components and considering only overlapping genes between the two datasets.

Following label transfer, we applied a semi-supervised version of the Leiden algorithm (resolution=1) to refine cluster assignments, "freezing" clusters 0, 4, 9, 10. The ARI score was calculated based on ground truth annotations provided by a human expert. For the unconstrained Leiden clustering used in the experiment, we also set a default resolution of 1.

Joint Analysis and cisTopic

The SNARE-seq dataset was created by combining four separate kidney SNARE-seq datasets. Cells lacking annotations were excluded. The chromatin accessibility data was processed with cisTopic[196] to identify 40 topics, selected based on cisTopic's log-likelihood model selection method. These topics served as a reduced representation of the data and were used for clustering and visualization, following the same approach described previously for scRNA-seq data.

Data Availability

For a list of all datasets and download links used in the study, visit the Cellar website or see `https://www.nature.com/articles/s41467-022-29744-0#data-availability`.

2.3.3 Discussion

In conclusion, Cellar is a user-friendly, interactive, and comprehensive tool designed for cell type assignment in single-cell studies. Developed in Python with the Dash framework, Cellar incorporates efficient operations and data structures optimized for large datasets. These include the use of Annotated Data object in memory-mapping mode, which enables analysis of large datasets with minimal system memory usage, and approximate nearest neighbors via faiss, which accelerates neighbor graph construction for Leiden clustering. Cellar also offers various interactive components for maximum flexibility. For instance, selecting the appropriate number of neighbors to build the connectivity graph, as well as determining a suitable resolution parameter for Leiden clustering, can both significantly influence the ability to select small, rare cell type clusters. Drawing on their specialized understanding of the data, the domain experts can fine-tune these parameters to achieve more accurate results.

Cellar supports multiple types of molecular sequencing and imaging data, and implements a range of popular methods for visualization, clustering, and analysis. It has been used to annotate single-cell data across diverse platforms and tissues, with many of these annotated datasets (primarily from HuBMAP) available as references for label transfer to other datasets. For tissues not covered by the existing HuBMAP references, Cellar integrates several external functional enrichment datasets. These resources, along with user insights on specific markers, assist in accurate cell-type assignment.

Clustering algorithms included in Cellar, such as Leiden[187], do not suffer from clusters of same size, such as KMeans, and can identify rare cell type populations.

We anticipate that Cellar will enhance both the accuracy and ease of cell-type assignment in single-cell studies. A web server running Cellar is accessible at https://cellar.cmu.hubmapconsortium.org/app/cellar.

Chapter 3

Biomarker Discovery in a Dynamic Context

With few exceptions, all cells in a person's body share the same DNA and genes. However, our physiological systems are far from static. Gene expression is constantly being regulated by proteins, epigenetic modifications, and even environmental factors such as diet and temperature[206]. Therefore, studying cells and genes in the static context does not provide a full picture of the dynamic processes at play within our bodies.

Consider, for instance, the COVID-19 pandemic. Remarkably varied responses to the disease were evident even among individuals who shared many characteristics such as diet, living environment, and age[207]. Due to baseline differences between people, understanding the genetic underpinnings of these diverse responses cannot be accomplished through static analysis alone. Instead, a longitudinal approach—tracking disease progression over time and capturing the body's evolving response to the virus—is essential. By integrating the temporal dimension into genomic data, we can identify biomarkers that effectively capture these dynamic processes.

To discover such important biomarkers, we first need to develop tools that model the dynamics of these systems. In this thesis, we explore the dynamics of endotypes, as well as temporal gene regulatory networks (GRNs).

3.1 Endotype-Informed Biomarkers from Time Series Clinical Transcriptomics Data

Based on: *Hasanaj, E.* et al. *Integrating patients in time series clinical transcriptomics data. en.* Bioinformatics ***40,*** *i151–i159 (June 2024)*

Transcriptomics data has been collected and profiled in clinical and drug response studies for over a decade[209]. In most cases, researchers profile bulk expression, though more recently single-cell data was also profiled in such studies[210]. The main goal of these studies is to reconstruct networks and systems that are activated in response to the disease, drug, or vaccine, over time[211,212].

A major challenge in the analysis of data from clinical trials is the fact that different individuals may display different response *dynamics*[20,213]. Even if the same biological process is activated, based on baseline differences (related to age, gender, prior disease history, etc), these individuals may respond faster or slower to the same treatment. Another challenge is the heterogeneous responses from different individuals. While a single response trajectory is possible, often we observe a (small) number of endotypes. **Endotypes** are subtypes of a disease characterized by different pathogenic mechanisms[39,214,215] which can have an impact on the specific optimal treatment. Each of the endotype groups may respond differently to the same treatment and so the overall set of patients cannot be directly integrated when studying treatment or vaccine response.

Several methods have been developed to address the first challenge (aligning patients)[216,217]. These often use expectation-maximization (EM) like methods. In these approaches, genes are represented as continuous curves and individuals are assigned to different time points along these[218]. Such methods have been widely applied[219,220] but they still suffer from several drawbacks. First, the continuous expression assumption may be problematic when sampling rates are sparse (genes can change a lot between two consecutive measurements) and second, they cannot reconstruct trajectories for multiple subsets of patients but rather assume a homogeneous response among all.

Another direction that was explored, especially in the single-cell space, is that of trajectory inference. Unlike the EM methods, these approaches assume the presence of multiple states in the data and allow for multiple subsets or branching. These methods range from linear or tree-based, to more recent adaptations of RNA velocity[9,221]. However, most of these methods assume no relationships between cells or samples. Only a few methods have focused on the case when samples come from different time points as is often the case with clinical trials data[222,223]. However, these single-cell methods assume a very large number of samples (in the thousands or tens of thousands) which is not available for most clinical studies including the ones analyzed in this paper. In addition, they usually do not explicitly map the different subgroups within the data, leaving it for subsequent, post-processing, analysis.

Here, we present **Tr**ajectory Inference via M**u**lti-commodity **Fl**ow with Nod**e** Constraints (Truffle), a method that performs pseudotime ordering of samples in short time series data (Fig. 3.7). Truffle is based on the multi-commodity flow algorithm[40] which generalizes minimum cost flow problems to include multiple source and sink nodes. Each sample in our data can be seen as either a source or a sink node and we are interested in recovering directed paths between these that minimize a cost function (typically some distance in gene space). The advantage of Truffle is that these trajectories can be constrained to satisfy timing restrictions and to pass through other nodes which correspond to intermediate disease states not present in the patient specific time series. Endotypes are then determined by constructing a state diagram for different subsets of patients. Truffle allows for the possibility of recovering contrasting endotypes since trajectories are inferred for each patient rather than for the entire dataset.

Truffle helps us perform biomarker discovery through the discovery of these developmental trajectories. For example, if two trajectories pass through distinct disease states (clusters), performing a DE analysis between these states can reveal genes that are differentially activated in one group of patients compared to the other. These genes are potential biomarkers for the endotype associated with each trajectory.

We tested Truffle on several microarray and bulk RNA-seq datasets. As we show, Truffle can accurately identify relevant disease trajectories and pathways, improving upon prior methods for clinical time series data and methods for single-cell data. A number of novel trajectories identified by Truffle suggest new subsets of patients that can benefit from precision medicine. We also compare some of these endotypes and show processes where they differ, thus enabling the identification of markers that are endotype-specific.

3.1.1 Materials and Methods

Data and preprocessing

We used three public time series datasets with the following GEO accession numbers GSE171012 (psoriasis), GSE212041 (COVID-19), and GSE112366 (Crohn's disease)[224–226] (Table 3.1).

Table 3.1: Clinical data used to benchmark Truffle.

	Number of				Metadata		
Disease	Samples	Genes	Patients$^{+(-)}$	Visits	Time Points	Tissue	Treatment
Crohn's	231	11,133	108 (26)	3	WK0, WK8, WK44	ileum	ustekinumab
COVID-19	650	33,142	304 (8)	3	D0, D3, D7	blood	N/A
Psoriasis	55	16,369	15 (11)	4	Pre[1], WK2, WK4, WK12	lesion	secukinumab

Notes: All three datasets contain missing values. We show both the number of patients who tested positive (+) and the number of healthy control patients (−).

[1] Pretreatment week.

Raw gene counts were downloaded from NCBI GEO for the two RNA-seq datasets (psoriasis and COVID-19). Only protein-coding genes that had more than 0.25 counts per million (CPM) in at least 1% of the samples were kept. In the case of duplicated gene identifiers, the gene with the highest mean expression was considered. Datasets were then normalized for their guanine-cytosine (GC) content and trimmed mean of M-values (TMM) was performed[227]. If batch information was present, ComBat was used to extract batch-corrected expression values[228]. Only samples with disease/treatment were used for pseudo-ordering. For microarray data, in the case of multiple probesets belonging to a protein-coding gene, only the one with the highest expression was kept. The Crohn's dataset was pre-normalized by Robust Multichip Analysis (RMA).

We removed symptomatic COVID-19^{-} from the COVID-19 data and kept only the patients who tested positive for the disease.

Assignment of disease states through clustering

To obtain disease states, we clustered the samples. We followed a standard practice that is also adopted by other computational tools such as Seurat[229]. We first ran principal component analysis (PCA) to obtain low dimensional embedding vectors which were then used to construct a fuzzy simplicial set as done by Uniform Manifold Approximation and Projection (UMAP)[190]. We adjusted the number of neighbors based on the total number of samples—using 15 for Crohn's, 20 for COVID-19, and 5 for psoriasis. Larger numbers resulted in highly connected graphs. This connectivity graph is the input for both Leiden clustering[187], and multi-commodity flow (below).

To assign states to biological processes, we performed gene set enrichment analysis (GSEA)[184] using the prerank function of GSEApy[185]. Genes were ranked based on the following score:

$$\text{gene score}_i = -\log_{10}(\text{adj. } p\text{-value}) \cdot \log_2(\text{FC})$$

where in the first term, adjusted p values were obtained from a two-sided Kolmogorov-Smirnov (KS) test[230] comparing the diseased and healthy sets of patients, and the second term is the log fold-change in gene expression between the two sets. We rely on the Gene Ontology (GO)[125] biological processes marker set for the enrichment analysis in this work[125].

Multi-commodity flow with node capacity constraints

The multi-commodity flow problem with node capacity constraints is defined as follows. Consider a directed graph $\mathcal{G} = (V, E)$, where an edge $(u, v) \in E$ has an associated cost $c_{u,v}$. We are given a set of K commodities $\mathcal{K} \coloneqq [K]$. The i^{th} commodity is defined by a source and sink node (s_i, t_i).

Multi-commodity flow can be used to model patient trajectories. Assume for simplicity patients with only two visits each. In this setup, each patient corresponds to one commodity, and the two

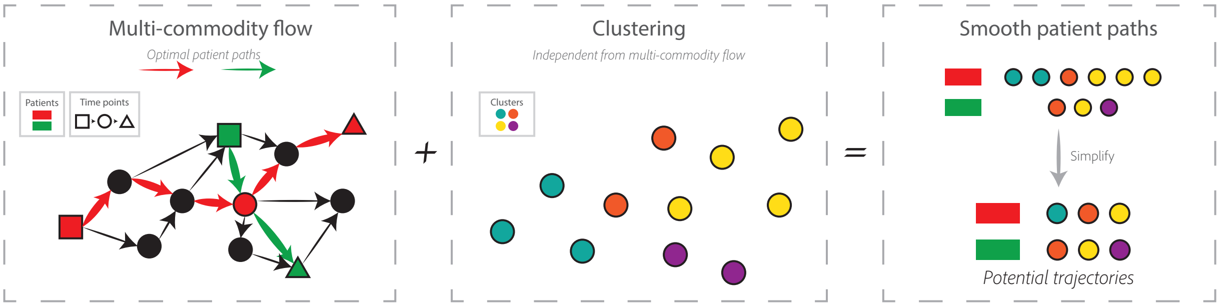


Fig. 3.1: **Schematic illustration of Truffle.** For each patient, our flow algorithm returns a trajectory that passes through intermediate nodes for a smoother response. These trajectories are then aligned with the clustering results to obtain a state diagram. Finally, by estimating state initial and final probabilities from the data, we can compute and study the top directed trajectories.

visits represent its source s and sink t. The objective is to recover a smooth disease trajectory between these two endpoints. If the data contains patients with diverse disease states, we can assume that some of the samples will lie "in between" s and t. The shortest path between these two nodes in the neighbors graph captures this smooth transition. By setting edge and node capacities we force the algorithm to look for robust paths (defined here as paths with similar state transitions even though they share no edges). Finally, if a patient has more than two time points, we consider each transition separately. E.g., a time series $a \to b \to c$ is split into two commodities $a \to b$ and $b \to c$.

Specifically to use multi-commodity for trajectory inference, we use the following constrains. For every commodity i, we wish to learn separate functions $f_i : E \to \{0, 1\}$ that satisfy the following constraints:

1. **Max edge capacity**: the total amount of commodity that passes over an edge does not exceed its capacity
$$\forall (u, v) \in E : \sum_{i \in \mathcal{K}} f_i(u, v) \leq C.$$

2. **Flow conservation**: flow must fully exit source nodes and enter sink nodes. For all $i \in \mathcal{K}$:
$$\forall n \in V : \sum_{w \in V} f_i(n, w) - f_i(w, n) = \begin{cases} 1 & \text{if } n \text{ is the } i^{\text{th}} \text{ source} \\ -1 & \text{if } n \text{ is the } i^{\text{th}} \text{ sink} \\ 0 & \text{otherwise} \end{cases}$$

Given a node capacity $N > 0$, we also consider the following constraint:

3. **Max node capacity**: the total amount of commodity that passes through a node does not exceed its capacity
$$\forall w \in V : \sum_{i \in \mathcal{K}} \sum_{u \in V, u \neq w} f_i(u, w) \leq N.$$

Along with flow conservation, constraint 3 guarantees limits on both incoming and outgoing flow. This variant of multi-commodity flow with node capacity constraints has also been explored before[231]. The integer problem has been shown to be nondeterministic polynomial-time (NP)-complete[232], however, its fractional form (setting the codomain of f to be $[0, 1]$) can be solved in polynomial time through linear programming. We use the open source Python optimization library pyomo[233] and the glpk solver[234]. It is worth noting that faster commercial solvers exist[235].

In the general formulation of the problem, each commodity can have a demand D, and each edge can have a capacity C[40]. Since a priori we do not have any preference for individuals, we set $D = 1$ for all commodities. We set $C = 1$ for psoriasis and Crohn's datasets. For the COVID-19 data, the problem was infeasible for $C = 1$, so we used $C = 2$. Enforcing edge and node capacities prevents outliers and errors in the data from having a large impact. An example has been provided in Fig. B.5.

Obtaining flow satisfying solutions

We learn f by optimizing the following target function

$$U = \sum_{(u,v)\in E} \left(c_{u,v} \sum_{i \in \mathcal{K}} f_i(u, v) \right)$$

Recall that $c_{u,v}$ is a cost function. As we are concerned with smooth trajectories, this is initialized as the Euclidean distance between the PCA embeddings for nodes u and v.

Note that for any given commodity defined by source s_i and target t_i, most of the edges "far away" from s_i and t_i will not be picked by the solver. We can incorporate this observation into our problem by considering only edges that belong to any path $s_i \to t_i$ of length $\leq \ell$ for some ℓ. This reduces the runtime for large datasets without compromising the optimality of the solution. For the smaller datasets, we found that the solution to this modified problem was similar to the original one. For the COVID-19 data, we set $\ell = 4$. Unreachable commodities were removed (17%).

Trajectory inference from optimal flow paths

After obtaining a path for each patient, we aggregate this information in the form of a state-transition matrix. In this work, we estimate initial and final state probabilities from the data, although domain expertise or priors determined from larger knowledge bases can be also used. Finally, we can then compute the most likely trajectories by performing random walks of a desired length. This is preferred over simply counting the occurrence of each path since in that case we could miss trajectories which are not identical, but show the same trend. For example, the paths $0-5-2-7$ and $0-5-3-2-7$ are different, but likely correspond to a similar disease trajectory. Our setup would assign a high probability to transitions $0 - 5$ and $2 - 7$.

STEM analysis of learned trajectories

To determine groups of genes that follow similar transcriptional programs, we perform Short Time-series Expression Miner (STEM) analysis[236]. We performed STEM normalization on gene expression values and used the default number of profiles (50), except for paths of length 2 where the maximum possible number is 16. Larger values for the number of profiles resulted in many redundant profiles that were nearly identical. For psupertime only, we reduced the "Minimum Absolute Expression Change" to 0, since psupertime normalized expression values were in a much smaller range than for the other two methods.

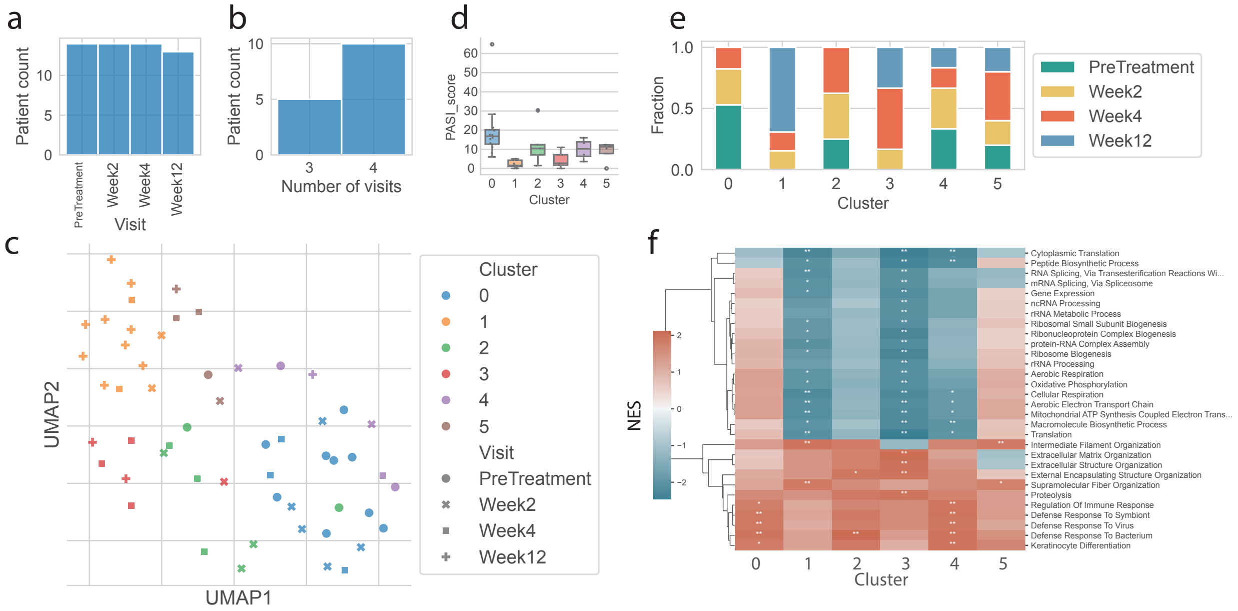


Fig. 3.2: **Clustering analysis of the psoriasis dataset.** (a-b) Distribution of visits across patients. (c) UMAP plot of cluster assignments. (d) Boxplots of PASI scores for each cluster. (e) Relative frequency of visits by cluster. (f) Top GO terms for each cluster against healthy samples. We used a KS test to rank the genes. A $(*)$ symbol means the category was statistically significant $((**) \equiv q \approx 0$ and $(*) \equiv q \leq 0.05)$.

3.1.2 Results

We developed a method to perform pseudotime ordering of multiple short times series clinical data based on optimal flow algorithms. Our method takes as input gene expression data from multiple subjects along with their specific time point, and tries to reconstruct trajectories that describe distinct disease endotypes. As a proof of concept, we first performed a simulation study with randomly generated data. Truffle accurately recovered the simulated trajectories in this study (Fig. B.6). To further validate our method, we used clinical data for psoriasis, COVID-19, and Crohn's disease (Table 3.1). We compare our method against prior work developed for similar tasks including Tempora, psupertime, as well as a baseline that assigns endotypes based solely on clustering analysis. The set GO Biological Processes was used for Tempora.

Truffle recovers trajectories that indicate regeneration and reduction of inflammation in patients with psoriasis

We tested Truffle on bulk RNA data from psoriasis patients treated with secukinumab. The data spans 12 weeks and most patients have data for all four time points (Fig. 3.2a-b). Leiden clustering identified six states (Fig. 3.2c). Cluster 0 predominantly consists of pre-treatment samples (50%) and contains no samples from week 12. Judging by the PASI scores (Fig. 3.2d), this cluster represents severe chronic plaque psoriasis. GO analysis shows significant upregulation of genes involved in the regulation of immune response (FDR $\leq$ 0.001) and defense response to virus & bacterium (FDR $\approx$ 0, Fig. 3.2f) when compared to healthy samples. We also see significant upregulation for keratinocyte differentiation (FDR $\leq$ 0.001) which is a hallmark of moderate-severe

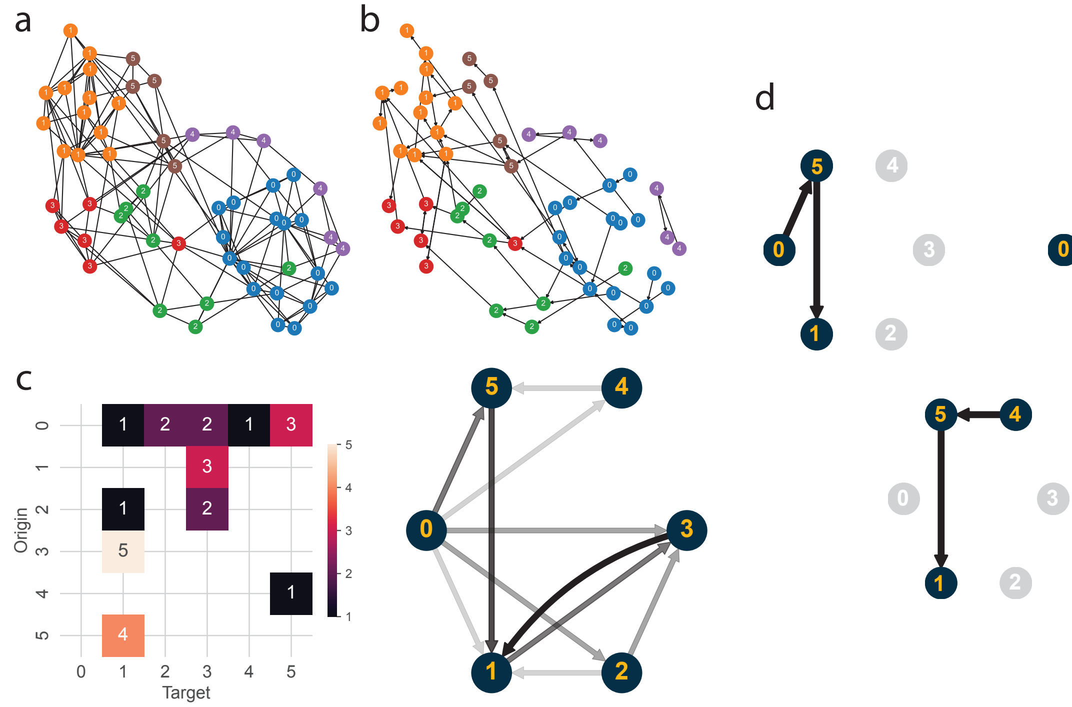


Fig. 3.3: **Truffle state diagram and top trajectories for the psoriasis dataset**. (a) Original connectivity graph obtained using fuzzy simplicial sets and (b) the graph corresponding to all the low-cost trajectories selected by Truffle (right). We used an edge capacity of 1 and a node capacity of 3 for this dataset. (c) The pruned state diagram describing the main state transitions in the Truffle network. Repeated states were collapsed into one, hence, no self-loops are shown. (d) The top paths identified by Truffle.

disease states[237]. Other immune-related processes such as Neutrophil Chemotaxis, Antimicrobial Humoral Response and Regulation Of Interferon-Beta Production were also up-regulated in this cluster (Fig. B.7d). In contrast, for cluster 1 approximately 70% of the samples are from week 12 and there are no samples assigned to this cluster from the pre-treatment week. The PASI scores for cluster 1 were also the lowest among all clusters (an average of 2.3). This cluster is enriched for intermediate filament and supramolecular structure organization, and keratinocyte differentiation is no longer significant. Downregulation of processes related to regulation of gene expression is also seen as a result of drug action, along with a reduced immune response.

We first looked at the most common cluster transitions using patients' samples timeline without cost constraints. We found that three patients transitioned from state 0 → 1, and two remained at state 4. All the remaining transitions were exclusive to only 1 patient. Next, we ran Truffle to uncover smoother response trajectories. Fig. 3.3 shows the state diagram identified by Truffle as well as the top 3 paths. The transition 0 → 1 was supplemented with two intermediate states, 5 and 3. GO analysis (Fig. 3.2f) shows that state 5 is characterized by a downregulation of defense response mechanisms when compared to state 0, while serving as an intermediary for a number of downregulated terms in state 1. On the other hand, state 3 is characterized by an upregulation of extracellular matrix organization which plays a role in tissue regeneration. Among the baselines, Tempora was able to recover paths of length 1 only (Fig. 3.4a). However, it correctly

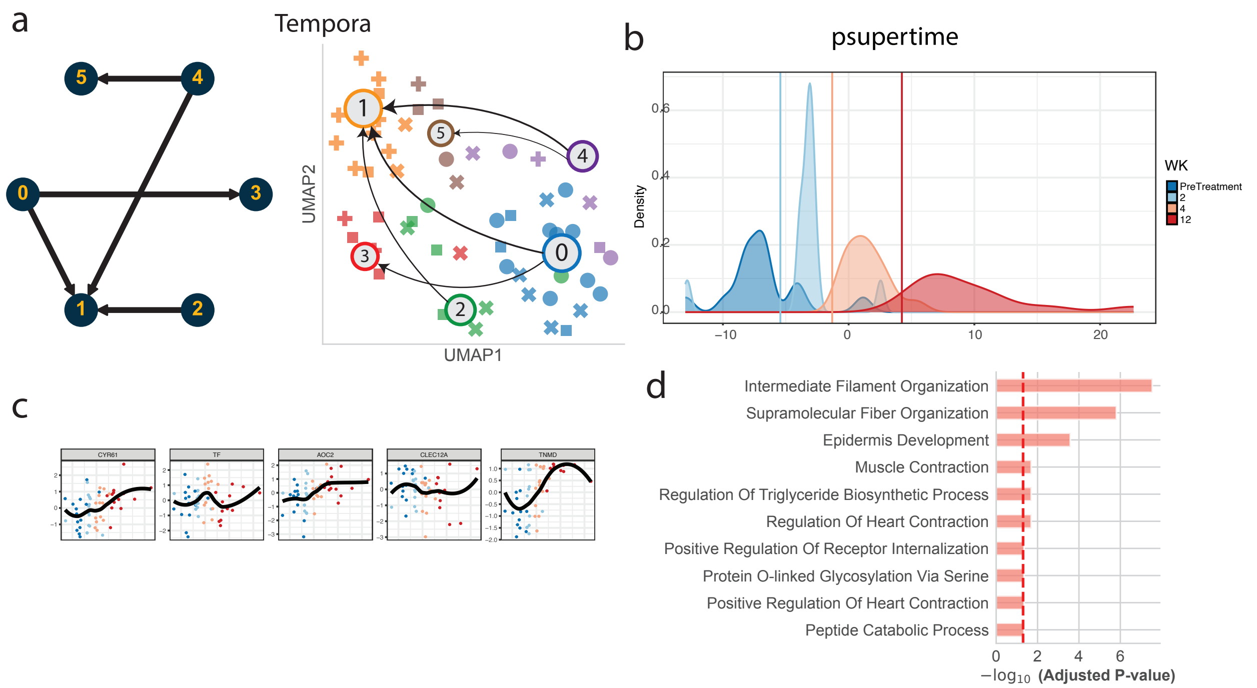


Fig. 3.4: **Trajectories uncovered by Tempora and psupertime for the psoriasis dataset.** (a) Transition graph identified by Tempora. Five trajectories of length 1 were identified. (b) Separation of time points by psupertime. The y axis is the density of each time point and the x axis is the temporal ordering. (c) The top 5 genes identified as relevant by psupertime. These correspond to the genes with the largest absolute coefficients. (d) The top GO terms for all the relevant genes (294). Subfigures (b) and (c) were generated using psupertime.

identified state 1 as a terminal state, but also 3 and 5. Psupertime identified 294 genes which vary coherently with time. GO analysis shows that these genes are enriched for intermediate filament and supramolecular fiber organization, as well as epidermis development. However, no significant terms involving defense response were found for the psupertime results.

Finally, we performed STEM analysis on the top three trajectories identified by Truffle. Profiles involving upregulation of epidermis development and downregulation of defense response overlapped across all three trajectories. Trajectories $0-5-1$ and $4-5-1$ contained decreasing profiles which were significantly enriched for genes involved in “IL-27-Mediated Signaling Pathway” (Combined Score $\geq 1e6$, Fig. 3.5c (right) and Fig. B.7c). These two trajectories differ in their initial state only. While states 0 and 4 are both enriched for defense response, state 4 shows a downregulation of terms such as cytoplasmic translation and other biosynthetic processes.

Truffle identifies different immune responses to COVID-19

We repeated the analysis with samples from a larger dataset of COVID-19 patients collected at days 0, 3, and 7. Clustering analysis identified 10 states (Fig. 3.6c). State 8 consisted of day 0 samples, and showed the highest acuity scores (Fig. 3.6d-e). State 0 showed significant upregulation of inflammatory response and other defense mechanisms when compared to healthy samples (FDR ≈ 0). State 1 was similarly enriched for “Defense Response to Virus”, but not for inflammation. About 20% of all patients ended in state 2, which differed from healthy samples only in it being

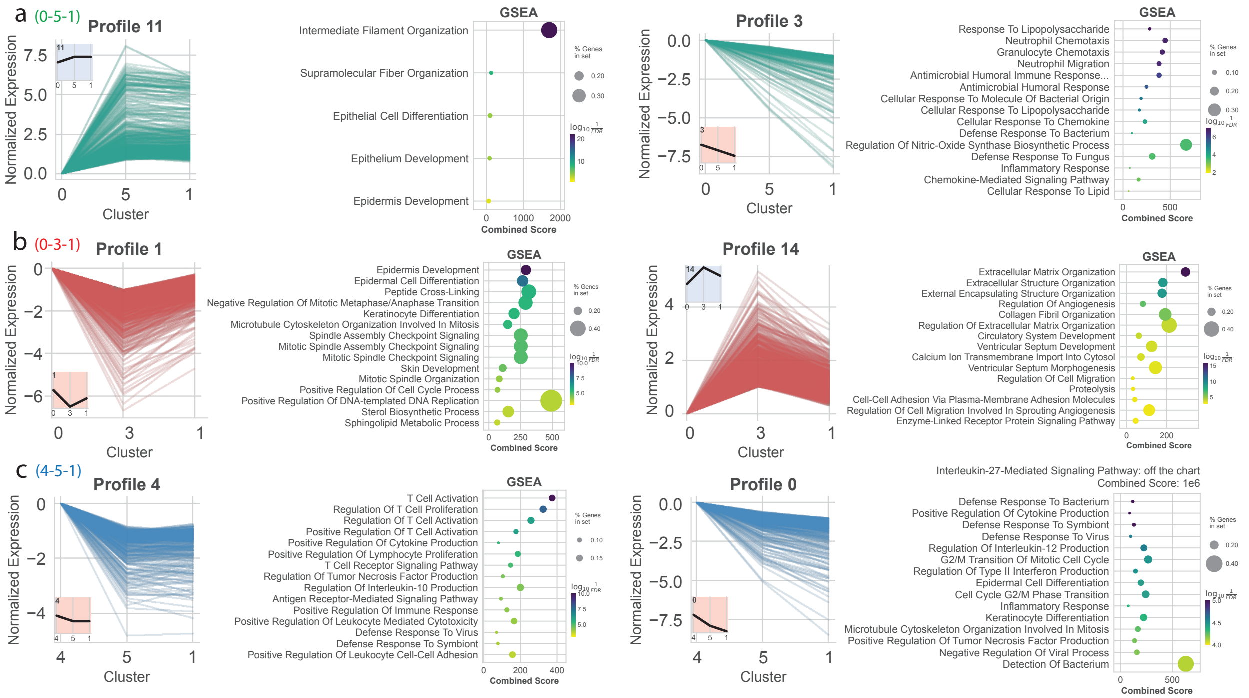


Fig. 3.5: **Selected STEM profiles for the top three Truffle trajectories in the psoriasis dataset.** (a-c) Two selected profiles for each of the three trajectories. In (c, right) “IL-27 Mediated Signaling Pathway” obtained a very high combined score (1e6), hence, was removed from the plot for clarity. The full list of profiles can be found in the appendix.

significantly enriched for Antimicrobial Humoral Response and Defense Response To Bacterium (FDR $\approx$ 0). This suggests that this is a milder state than the previous two, also confirmed by acuity scores where cluster 2 is the only one containing no samples with acuity 4 or 5 (Fig. 3.6e). Across all 3 time points, most patients (10) moved from state 0 to state 2. This was also the top trajectory captured by Truffle (factoring in initial and terminal probabilities for each state, Fig. 3.6f). In contrast, this trajectory was not recovered by Tempora (Fig. 3.6g).

Next, we studied the top trajectories identified by Truffle at varying levels of resolution. The top trajectories of length 3 and 4 were $T_1 \coloneqq 0-1-5-2$, $T_2 \coloneqq 0-1-5-4$, and $T_3 \coloneqq 0-1-2-5-4$, $T_4 \coloneqq 0-2-5-4-3$, respectively. For brevity, since T_2 is a subsequence of T_3, we only look at T_3, although T_2 could be an endotype in its own right describing a “faster” response.

STEM analysis of T_1 assigned more than 4,000 genes to profile 49 (Fig. B.8a). GO analysis showed that ~ 50 genes in profile 49 were involved in sensory perception of smell (FDR = 0.02), a common symptom of COVID-19[238]. We see an upregulation of these genes from 0 → 5, but a downregulation from 5 → 2.

On the other end, for T_3, STEM assigned more than 9,000 genes to a strictly increasing profile (profile 41, Fig. B.8a). This profile was also enriched for processes related to sensory perception of smell, but this time we see an upregulation of related genes across all 4 temporal steps. Profile 2 (T_3) and profile 9 (T_4) indicate downregulation of immune response. Profile 9 is gradual. Looking at GO enrichment of the final state of T_4 (cluster 3), we observe a return to baseline (healthy) for various defense response processes and downregulation of gene regulation activities.

Tempora, on the other hand, identified only two paths of length ≥ 2. These were $Q_1 \coloneqq 6-2-3$

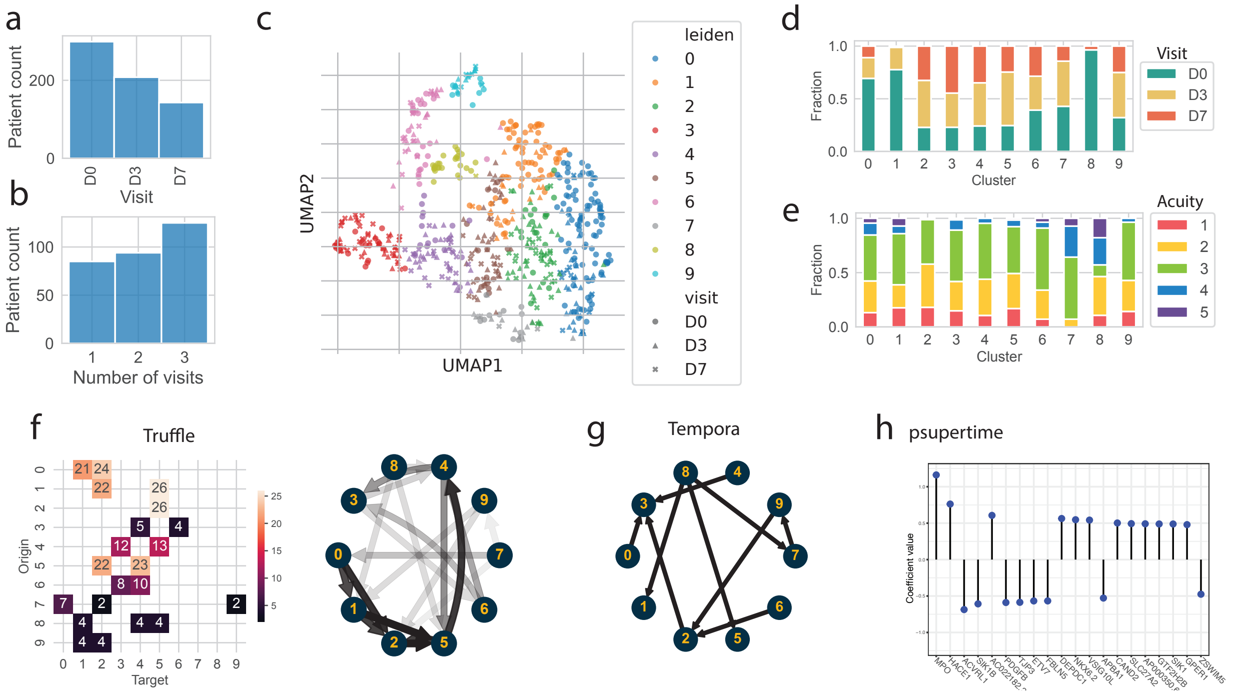


Fig. 3.6: **Clustering and trajectory analysis for the COVID-19 dataset.** (a-b) Distribution of visits and distribution of visit counts per patient. (c) Clustering 650 samples from 304 patients. (d) Relative frequency of visits and (e) acuity scores per cluster. (f) A pruned diagram of top state transitions identified by Truffle. Pruning was performed by taking the fewest top edges that amount to $\geq 50\%$ of a node's outgoing weight. (g) The tree learned by Tempora. Final states are 3, 1, and 5. (h) The top genes that vary with time according to psupertime (plot obtained from psupertime). (i) Selected STEM profiles for Truffle trajectories P_1 (green), P_3 (red), P_4 (blue).

and $Q_2 := 8-7-9-2-3$. Three significant STEM profiles were determined for Q_1, none of which was significantly enriched for any GO process (FDR $= 0.05$). For Q_2, STEM returned 11 significant profiles. Among these, only three were enriched for GO processes (Fig. B.8b). Profile 10 was enriched for sensory perception of smell, and profile 7 was enriched for the only term "Positive Regulation of NF-kappaB Transcription Factor Activity". Meanwhile, profile 37 showed an initial increment, followed by a monotone decrement of processes related to signaling. Finally, psupertime identified 462 relevant genes. GO analysis using these genes returned only one process: "Hydrogen Peroxide Catabolic Process" (FDR $= 0.007$).

Truffle identifies two contrasting response mechanisms to ustekinumab in patients with Crohn's disease

Finally, we tested Truffle on microarray data from patients with Crohn's disease treated with ustekinumab[226]. The data was collected at weeks 0, 8, and 44. Clustering analysis revealed 8 distinct states. States 1 and 4 were not statistically different from healthy samples. States 0, 3, and 6 expressed genes enriched for inflammatory response, while cluster 2 showed a downregulation of the process (Fig. B.9b-c).

The top Truffle trajectories of length 2 were $C_1 := 3-4-1$ and $C_2 := 2-5-0$. C_1 transitions from a state with inflammation into two healthy states, suggesting that patients along this path saw improvement from the drug. In contrast, for C_2 we see an activation of immune response in its final state (cluster 0). Indeed, about 14 patients were clustered under state 0 at week 44, suggesting that they showed partial response to the drug. STEM analysis of C_1 returned several decreasing profiles which were enriched for inflammatory response. In contrast, C_2 was assigned increasing profiles enriched for immune response and activation of T cells (Fig. B.9d). Thus, Truffle was able to recover two contrasting endotypes for patients in this study.

3.1.3 Discussion

Several trajectory inference methods have been developed to date and these differ in representation power and assumptions made[9]. Most of the work has focused on single-cell with much less focus on data collected in clinical studies. Here we focus on studies that profile a small number of time-points in multiple patients. To analyze such data, we developed Truffle which respects the time ordering of samples for a given patient, and obtains patient journeys through the disease/treatment process. Truffle is based on multi-commodity flow by splitting short time series into source and target nodes. These are then connected through a path that travels through other intermediate nodes in order to generate a smooth path. We tested Truffle on several time series datasets and compared it to two other methods developed for similar tasks.

For the psoriasis dataset, all patients display a significant health improvement after treatment with secukinumab as indicated by their PASI scores and GO analysis of the terminal state. Since patients respond differently to the treatment, we sought to understand different endotypes within the patient population. Clustering analysis does not lead to accurate grouping of disease subtypes. Some of the other methods were able to capture the improvement either by identifying a healthy final state (Truffle, Tempora) or by showing enrichment for healing biological processes (psupertime). However, Tempora identified only paths of length 1, thus providing lower resolution into the drug response progression, while psupertime does not provide details into different response mechanisms or endotypes due to its linearity assumption. Only Truffle was able to capture temporal dynamics of the treatment process among different patients and obtain different endotypes. For example, Truffle recovered two paths which end in a healthy state but travel through different states. Both show the downregulation of *IL-27* and its pathway genes. Reduction of type I & II interferons (IFNs) and/or tumor necrosis family (TNF) receptors, which are regulators of *IL-27*, has been previously observed as part of the recovery[239]. Furthermore, *IL-27* was previously reported to promote the onset of psoriasis[240]. However, they also differ in other pathways. One of these trajectories was characterized by an upregulation of extracellular matrix organization (ECM) and downregulation of intermediate filament organization (IFO), while for the other trajectory we observed the opposite. Prior work has shown that activation of ECM is related to the severity of psoriasis[241]. We hypothesize that the upregulation of ECM may be an intermediary stage of slow responders. Results show that a subset of patients quickly attained normalization of keratinocyte differentiation (Fig. 2-3 clusters 1, 3, 5). Such patients can be deemed as super/fast responders to therapy. These patients can be further investigated to better tailor personalized therapy.

For the COVID-19 dataset, prior methods failed to recover smooth trajectories with any significant GO terms. Tempora recovered trajectories that oscillate between time points, which makes them hard to interpret, and psupertime returned only one significant GO process, likely because this linear method was forced to combine heterogeneous subtypes in its trajectories. Truffle identified several trajectories, including ones which showed a downregulation of defense response over time and others where this response was reinstated at day 7. This was confirmed by a reduction of sensory perception of smell during this time step.

While the applications we presented are mainly focused on immunology, we believe that Truffle can also be applied to oncology time series data and that it can also be integrated with time series data from other sources including electronic health records (EHRs) or claims databases. Additionally, clinical outcomes can be integrated into the analytical framework to identify trajectories that better align with observed patient results. For instance, semi-supervised clustering methods can be used to group terminal samples from patients who share similar clinical endpoints, thus uncovering outcome-aligned trajectories.

While successful, Truffle has a few limitations. The datasets we used in this study contained at most 650 samples. The open-source linear solver we used to optimize a graph of this size may not scale to graphs with several thousands of samples. In this case, several simplifications to the problem may need to be introduced, such as limiting the set of edges a commodity can be transported over. For the specific datasets we evaluated, Truffle took 0.12s to run for the small psoriasis dataset and 22s for the larger COVID-19 dataset ($\ell = 4$)[1]. In addition, faster commercial solvers can also be used.

To conclude, Truffle is a method for integrating patient data in time series transcriptomics studies. It is able to both, identify patient trajectories and subgroups within a population, thus enabling marker discovery through the use of DE analysis methods on the discovered trajectories. Truffle is available as an open source software from the link in the abstract.

3.1.4 Acknowledgments

The authors thank the anonymous reviewers for their valuable suggestions. The authors thank Hamid Mattoo from Precision Medicine Computational Biology at Sanofi, Cambridge MA, for suggesting the Crohn's disease dataset.

[1] Tests performed on a MacBook Pro with an M3 Pro Max chip.

3.2 Recovering Time-Varying Networks From Single-Cell Data

Based on: *Hasanaj, E.* et al. *Recovering time-varying networks from single-cell data.* arXiv [q-bio.QM] *(Oct. 2024)*

The previous chapter focused on general pathways activated over time in disease and development. While such analysis is important, it does not fully explain the root causes of the observed responses. Such information is crucial when attempting to intervene in a biological process, for example for drug development or diagnostics.

To construct such accurate models of biological activity during development, disease progression, treatment response, and other biological processes, it is essential to track their evolution over time[213]. Studying the *regulation* of these dynamic processes is key for understanding the underlying mechanisms that drive the response and for identifying potential interventions that can serve as cures for diseases[243].

Much of the research in this area is focused on the reconstruction of regulatory networks[14,244]. These networks comprise a subset of proteins known as transcription factors (TFs), which regulate the activity of all other genes and proteins within the cell. However, these gene regulatory networks (GRNs) are not static. Instead, both the active nodes (proteins) and the edges (genes) change over time[245,246]. To reconstruct such networks, researchers often integrate static data—such as the type of nodes in the network—with dynamic data, such as time series measurements of node activity (gene expression profiles). Early work in this area employed microarrays and ChIP-chip data[247–250] followed by time series next-generation RNA-seq data[251], and most recently, scRNA-seq data[14,252,253].

Several computational methods have been proposed over the last two decades to reconstruct such dynamic GRNs[254–257]. Some of these methods utilized time-varying graphical models including Hidden Markov models, Markov random fields, and Dynamic Bayesian Networks[258–261]. Other approaches attempted to use regression or to learn temporal precision matrices using extensions of the graphical lasso algorithm[262,263].

While such models successfully reconstructed some processes[258,264], they are less suitable for more recent types of data, most notably scRNA-seq time series. First, the larger size of the data presents a challenge for traditional graphical models. Also, prior methods do not directly account for the fact that multiple cells are profiled for each time point. Finally, prior methods do not leverage larger models, such as neural networks, which have demonstrated significant performance improvements across various learning tasks[6,265].

Very recently, a few methods have been proposed for using deep learning to recover static GRNs[266,267]. However, they cannot be directly used to capture dynamic GRNs (i.e., enforcing learning between time points). Two recent approaches, Dictys and CellOracle[268,269] can infer dynamic GRNs, however, these methods depend on data types like ATAC-seq, which provides direct information about transcription factor (TF) binding sites but is less prevalent and harder to obtain.

Beyond the realm of biology, the inference of dynamic graphs using neural networks has garnered significant attention. This problem has found applications in diverse domains, including information retrieval, molecular graphs, and traffic forecasting[270,271]. While there are similarities between these problems and the dynamic GRN problem, there are also significant differences that make it hard to extend these methods for time series scRNA-seq. The problem of inferring temporal graphs is usually defined by recovering a series of graph adjacency matrices $\mathbf{A}_t \in \mathbb{R}^{n\times n}$ where n is the number of nodes and each node is a k-dimensional feature vector. However, when dealing with scRNA-seq data, the problem becomes: given a gene expression matrix $\mathbf{X}_t \in \mathbb{R}^{c\times g}$, where c is the number of cells (samples) and g is the number of genes (features), we are interested in recovering gene networks $\mathbf{A}_t \in \mathbb{R}^{g \times g}$, i.e., *graphs of features* rather than nodes (cells).

In this paper, we present a novel deep learning framework that effectively addresses the challenges discussed above for reconstructing dynamic GRNs. Our contribution is three-fold. First, we demonstrate that existing deep learning methods for temporal graph structure learning can be adapted for scRNA-seq data analysis. To achieve this, we perform a *gene featurization* step by leveraging set-like architectures such as DeepSets or Set Transformers[272,273]. Second, we construct dynamic graphs by applying a self-attention mechanism[274] to these gene feature vectors. To model dynamics, we draw inspiration from EvolveGCN where a gated recurrent unit (GRU) evolves the weights of a graph neural network[275]. However, unlike EvolveGCN, our approach uses a GRU to evolve the weights of key and value projection matrices in the self-attention module. This allows for the construction of dynamic graphs that capture regulatory interactions over time. Lastly, GRNs are highly dependent on cell functions, hence, separate GRNs need to be learned for each cell type. A single scRNA-seq dataset may combine cells of multiple types, some of which are rare cell populations. To this end, we employ a model-agnostic meta-learning (MAML)[276] training procedure by treating each cell type as a "task" to be learned. With this approach, the model quickly adapts to tasks with few samples, enabling the reconstruction of dynamic graphs even for rare cell types.

We apply our **m**et**a** lea**r**ning approach for inferring tempora**l** g**ene** regulatory networks (Marlene) to three publicly available scRNA-seq datasets. The first is a time series SARS-CoV-2 mRNA vaccination dataset of human Peripheral Blood Mononuclear Cells (PBMCs)[277]. The second dataset is a human lung aging atlas from the Human Cell Atlas Project[124,278]. The third dataset is from a study of lung fibrosis using a mouse lung injury model[279]. All three datasets incorporate several time points, thus enabling a longitudinal analysis of the relevant biological responses through the inference of dynamic, cell type-specific GRNs. As we show, our method is able to reconstruct accurate networks for these datasets, significantly improving upon prior methods proposed for this task. Finally, by studying the transcription factors (TFs) or genes that are added between time points, we can identify markers for the specific response and pinpoint the time they were activated during the time course.

3.2.1 Materials and Methods

Problem Setup

Consider a gene expression matrix $\mathbf{X} \in \mathbb{R}^{c \times g}$ where c is the number of cells and g is the number of genes. In the human genome, g varies from 25,000 to 30,000, while the number of cells could be between a couple thousand to a few million. In the setting of dynamic graphs, we assume the existence of a time point for each row (cell), leading to a time series $\widetilde{\mathbf{X}} := \{\mathbf{X}_1, \ldots, \mathbf{X}_T\}$ with $\mathbf{X}_t \in \mathbb{R}^{c_t \times g}$. Here, the number of cells c_t may vary with t. We are interested in recovering a series of directed graphs $\widetilde{\mathcal{G}} := \{\mathcal{G}_1, \ldots, \mathcal{G}_T\}$ where each $\mathcal{G}_t = \{\mathcal{N}, \mathcal{E}_t\}$. The set of nodes is the set of genes, i.e., $\mathcal{N} = [g]$, and we assume this set is static over time. The dynamic edge sets $\mathcal{E}_t = \{(u, v, w)\}_{u,v \in \mathcal{N}, w \in \mathbb{R}}$ denote directed weighted links between genes, where the source u is the gene that regulates the expression of its target v, and w is the strength of this relationship. The source nodes are called transcription factor genes (TFs).

We can alternatively characterize each graph $\mathcal{G}_t$ by the corresponding adjacency matrix $\mathbf{A}_t \in \mathbb{R}^{g \times g}$. Denote $\widetilde{\mathbf{A}} := \{\mathbf{A}_1, \ldots, \mathbf{A}_T\}$. Since TFs control the expression of their target genes, the underlying GRNs should, in principle, allow the recovery of the full expression profile for a cell. In other words, $\widetilde{\mathbf{X}} = f(\widetilde{\mathbf{X}}^{\mathrm{TF}}, \widetilde{\mathbf{A}})$ where $\widetilde{\mathbf{X}}^{\mathrm{TF}}$ denotes the expression of all TFs. The function f is unknown as it involves intricate interactions among genes, including combinatorial effects. For

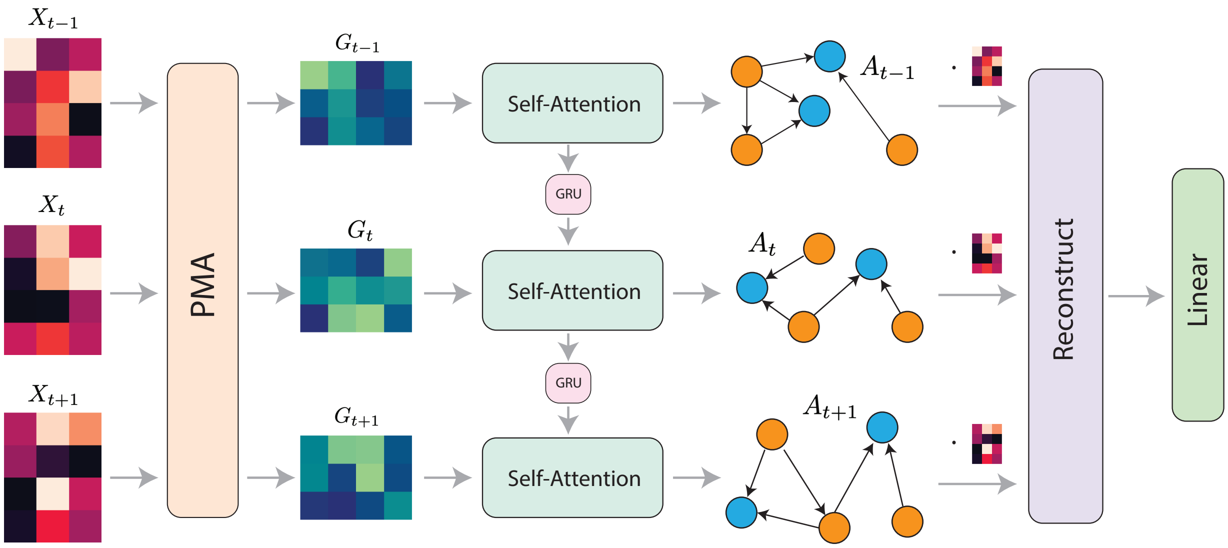


Fig. 3.7: **Overview of Marlene**. Marlene takes as input gene expression data in the form of a cell-by-gene matrix. It then performs gene featurization via the pooling by multihead attention (PMA) mechanism which returns a gene feature matrix. This matrix is then inputted into a self-attention module to obtain a gene network in the form of an adjacency matrix. The weights of the self-attention module evolve from one time point to the next via a gated recurrent unit (GRU). The expression of transcription factors and the recovered graph are used to reconstruct the full gene expression vector. Finally, the reconstructed matrix is used to predict the cell type for the batch. The network is trained in a model-agnostic meta-learning fashion where each cell type is treated as a "task" to be learned, thus enabling the model to quickly adapt to cell types with low representation.

instance, certain scenarios exist where TFs cooperate with each-other to activate a gene, while in other instances, the activation requires some TFs to be active and others to be inactive[280].

In existing deep learning literature, f is sometimes modeled using autoencoders[267,281,282]. However, reconstructing the full gene expression vector is challenging as the data is extremely sparse and conventional reconstruction losses, such as mean squared error, tend to emphasize overall averages. Since GRNs are dependent on cell function, we hypothesize that simplifying the problem by predicting cell types may improve the accurate recovery of GRNs. In other words, given a temporal batch of cells of the same type $\widetilde{x}^{\mathrm{TF}} \in \mathbb{R}^{T\times \text{batch size}\times|\text{TFs}|}$, we consider the classification problem $y = f(\widetilde{x}^{\mathrm{TF}}, \widetilde{\mathbf{A}})$ where y is the known cell type label for the batch. Finally, the task of learning $\widetilde{\mathbf{A}}$ given a batch $\widetilde{x}$ becomes

$$\arg\min_{\widetilde{\mathbf{A}}} \text{CrossEntropyLoss}(y, f(\widetilde{x}^{\mathrm{TF}}, h(\widetilde{x}))), \qquad \widetilde{\mathbf{A}} := h(\widetilde{x}) \tag{3.1}$$

for choices of functions f and h where h uses the expression data to obtain the adjacency matrices.

Architecture of Marlene

In this work, we propose a neural network architecture called Marlene that effectively learns dynamic GRNs (Fig. 3.7). Marlene consists of three main steps. The first two steps address the choice of h in (3.1), while the last addresses the choice of f.

In the first step, we apply a gene featurization step by treating a batch of cells as a set of elements. The DeepSet architecture introduces a pooling operator that allows the neural network to be invariant to the order of input samples, effectively treating the input as a set[272]. Similarly, the Set Transformer architecture is designed to process sets of data via attention-based operators that are permutation invariant[273]. Specifically, the pooling by multihead attention (PMA) aggregation scheme introduced in Set Transformers outputs a matrix of k vectors $\mathbf{H} \in \mathbb{R}^{k\times g}$ for an arbitrary input set $\mathbf{X} \in \mathbb{R}^{c\times g}$. Each of the k output vectors has a specific meaning such as statistics of the input data. By applying the PMA operator to a temporal batch of cells, we obtain a gene-by-feature matrix $\mathbf{G} = \mathbf{H}^\top \in \mathbb{R}^{g\times k}$ that encodes information about the cells from each time point. PMA consists of a multihead attention block (MAB) where the input $\mathbf{X}$ consists of key vectors, and the query is a learnable set of k vectors $\mathbf{S} \in \mathbb{R}^{g\times k}$. In this work, we use a shared PMA layer for all time points assuming that the specific key statistical properties are invariant (though their value obviously changes for different time points). Given $\widetilde{x} = [x_1, \ldots, x_T]$, we have

$$\widetilde{\mathbf{G}} = \mathrm{PMA}(\widetilde{x})^\top := \mathrm{MAB}(\mathbf{S}, \widetilde{x})^\top := (\widetilde{\mathbf{M}} + \mathrm{rFF}(\widetilde{\mathbf{M}}))^\top \tag{3.2}$$

where $\mathbf{M} = \mathbf{S} + \mathrm{Multihead}(\mathbf{S}, x, x) \in \mathbb{R}^{g\times k}$ and rFF is a row-wise feedforward layer. For completeness, these operations are defined in the appendix.

Note that if cells of different types are mixed in the same batch, the statistics derived by the PMA step may not capture cell type-specific information. Consequently, in a single batch, we only include cells of one type.

In the second step, Marlene learns temporal adjacency matrices using a self-attention mechanism. To model dynamics, we draw inspiration from EvolveGCN which performs model adaptation using a GRU[275]. Unlike EvolveGCN, which uses the GRU to update the weights of a graph convolution layer, we use a GRU to evolve the key and query projection weights of the self-attention module. Since most time series we deal with contain very few time points, a GRU should suffice and not suffer from vanishing gradient problems. Similar to EvolveGCN, we apply a summarization step via top k pooling to reduce the gene feature matrix to a square matrix for the GRU (Appendix).

More precisely, we initialize self-attention weights $\mathbf{W}_0^Q, \mathbf{W}_0^K \in \mathbb{R}^{k\times k}$ and two recurrent units $\mathrm{GRU}_Q, \mathrm{GRU}_K$. Given the time sequence of gene feature matrices $\widetilde{\mathbf{G}}$ obtained from the previous step, temporal adjacency matrices are constructed in the following recurrent fashion for all $t \in [T]$:

$$\mathbf{Z}_t = \mathrm{TopK}(\mathbf{G}_t) \in \mathbb{R}^{k\times k} \tag{3.3}$$

$$\mathbf{W}_t^Q = \mathrm{GRU}_Q(\mathbf{Z}_t, \mathbf{W}_{t-1}^Q), \quad \mathbf{W}_t^K = \mathrm{GRU}_K(\mathbf{Z}_t, \mathbf{W}_{t-1}^K) \tag{3.4}$$

$$\mathbf{Q}_t = \mathbf{G}_t\mathbf{W}_t^Q, \quad \mathbf{K}_t = \mathbf{G}_t\mathbf{W}_t^K \tag{3.5}$$

$$\mathbf{A}_t = \mathrm{softmax}\left(\frac{\mathbf{Q}_t\mathbf{K}_t^\top}{\sqrt{k}}\right). \tag{3.6}$$

Here, $\mathbf{W}_t^Q$ and $\mathbf{W}_t^K$ serve as hidden states for the respective GRUs. The GRUs dynamically adapt self-attention weights, influencing which TFs specific genes should attend to in subsequent time steps. Consequently, the evolution of these weights is constrained. We also restrict the columns of $\mathbf{A}_t$ (i.e., sources) to p known TFs in the TRRUST database[283] which greatly reduces the number of parameters to be learned. Therefore, in our implementation $\mathbf{A}_t \in \mathbb{R}^{g\times p}$.

Next, we perform a gene expression reconstruction step based on the expression of TFs and the inferred adjacency matrices. This is followed by any number of fully connected layers with nonlinear activation functions σ. Finally, we sum across output vectors to obtain a logit vector

Table 3.2: Time series scRNA-seq datasets used to benchmark Marlene

	Number of				Metadata	
Dataset	Cells	Genes[1]	TFs	Cell Types	Time Points	Sample
SARS-CoV-2	113,271	1899	556	7	d0, d2, d10, d28	PBMCs (human)
HLCA	27,953[2]	2433	674	11	Ages $< 35, 35 - 50, \geq 50$	lung (human)
Fibrosis	22,758	1217	433	6	PBS, d3, d7, d10, d14, d21, d28	lung (mouse)

[1] Only showing the number of genes overlapping with the TRRUST database.

[2] We randomly sampled cells from 11 cell types.

with the same dimension as the number of cell types in the data:

$$\tilde{y} = \text{Pool}(\text{Linear}(\ldots\sigma(\text{Linear}(\widetilde{x}^{\text{TF}}\widetilde{\mathbf{A}}^{\top})))). \tag{3.7}$$

Network depth can be introduced at all three levels by stacking MAB layers during gene featurization, stacking GRUs, or stacking linear layers at the end.

Meta learning for rare cell types

ScRNA-seq datasets often originate from biological samples that exhibit cellular heterogeneity, potentially containing multiple distinct cell types. Some of these cell subpopulations are rare and are represented by a small number of cells in the sample[284]. Since we are concerned with the discovery of cell type-specific temporal GRNs, learning such large graphs for these rare cell types may not be feasible and lead to overfitting. Since many interactions are shared across cell types[285], we employ the model-agnostic meta-learning framework (MAML)[276]. MAML is specifically designed to enable neural networks to adapt to novel tasks with limited training samples (i.e., few shot learning). By treating each cell type as a "task", the MAML training paradigm facilitates the recovery of dynamic graphs for rare cell types. We begin by adapting model parameters through multiple optimization steps using a batch of support examples (cells). These adapted parameters are then evaluated on a separate set of query cells, followed by a meta-update.

During the adaptation step, we perform gradient descent, while for the meta-update, we employ the Adam optimizer[286]. During training, gradient clipping proves crucial to prevent overfitting of the MAML adaptation step to the cell type under consideration.

3.2.2 Results

To validate our approach, we use three public scRNA-seq datasets (Table 3.2): a human SARS-CoV-2 mRNA vaccination dataset, a lung aging atlas (The Human Lung Cell Atlas—HLCA), and a mouse lung fibrosis dataset[124,277,279]. To assess the quality of the inferred networks, we draw upon two databases of TF-gene regulatory interactions, which have been curated from the scientific literature—TRRUST and RegNetwork[283,287]. For the human genome, TRRUST contains 8427 unique validated regulatory edges, while RegNetwork contains 150,405. Note that certain edges lack a corresponding TF or gene in the expression data, so the numbers used for the analysis are smaller. We used only the genes that were present in TRRUST for all three datasets. For the mouse lung dataset we used the corresponding mouse networks for both databases. To match the number of links in these databases, we selected the top 2% of edges for all methods. For Marlene, this was done by sparsifying the self-attention matrix to retain only the top scoring edges. The significance of the overlap was carried out via Fisher's exact test[288]. All p-values were corrected for multiple testing using the Benjamini-Hochberg procedure[183].

We compare Marlene against several popular static gene regulatory network inference methods included in the BEELINE benchmark[289] and beyond, such as PIDC, GENIE3, GRNBoost2, SCODE, and DeepSEM[267,290–293], which are applied independently to each time point. DeepSEM is a deep generative model based on structural equation modeling. We also compare against time-varying graphical lasso (TVGL)[2], a method that models temporal precision matrices[263], and to a deep neural network that utilizes the S4 module (GraphS4mer)[294,295].

During inference, we obtain multiple A_t for different batches and average them. We train Marlene with a batch size of 16 cells and also use 16 seeds in the PMA layer. For MAML, we use 5 inner steps. The model with the lowest loss is selected for GRN inference. For the meta-update, we use a decaying learning rate starting with $1{-}4$, while for the inner step we use $1{-}3$ for both datasets. Experiments were performed using an NVIDIA RTX 3060 and took only a few minutes per run.

Case study 1: SARS-CoV-2 vaccination

The SARS-CoV-2 vaccination dataset consists of peripheral blood mononuclear cells (PBMCs) from six healthy donors at four time points (days 0, 2, 10, and 28)[277]. Day 0 samples were obtained before vaccination. We removed the "Other" cell type group and kept the remaining seven. These include B cells, dendritic cells (DC), monocytes (Mono), natural killer cells (NK), and various types of T cells.

Marlene recovers accurate gene regulatory networks Analysis results using Marlene and prior methods is presented in Fig. 3.8. As can be seen, Marlene outperformed competing methods in the dynamic GRN inference task for 5 of the cell types, yielding statistically significant results across time points. Specifically, for B cells, Marlene successfully identified more than 800 regulatory links within the RegNetwork framework at each time point (FDR $\leq 1{-}67$), surpassing the performance of the second-best method, SCODE, which detected 579 links (day 2, FDR $\leq 1{-}15$). Analogous findings were observed for natural killer cells, where Marlene identified over 600 RegNetwork links at each time point (FDR $\leq 1{-}18$). In comparison, the second-ranking method, SCODE, showed a significant overlap for only one time point (day 28). Upon examining the TRRUST database, we observed less pronounced differences in the results. Nonetheless, Marlene obtained higher overlap for 5 out of 7 cell types followed by GENIE3, which performed well for monocytes and the "Other T" category.

Marlene recovers realistic dynamic transitions The analysis so far has primarily focused on individual time points. Next, we turned our attention to assessing the quality of graph transitions between consecutive time points. Specifically, we examined whether the learned graphs demonstrated smooth transitions over time. To evaluate this, we computed the intersection-over-union (IoU) score for edges between time points t and $t+1$ (Fig. 3.9a). Notably, our findings revealed that for most cell types, Marlene exhibited the lowest IoU score during the initial period (days $0 \rightarrow 2$), followed by higher scores during days $2 \rightarrow 10$, and $10 \rightarrow 28$. This pattern aligns with our expectations, as variations in gene expression are likely to be most pronounced during the early post-vaccination period (days $0 \rightarrow 2$). From the methods we compare against, GENIE3, GRNBoost2, SCODE, PIDC, and DeepSEM exhibited significantly lower IoU scores across all temporal

[2] We used the implementation of https://github.com/fdtomasi/regain

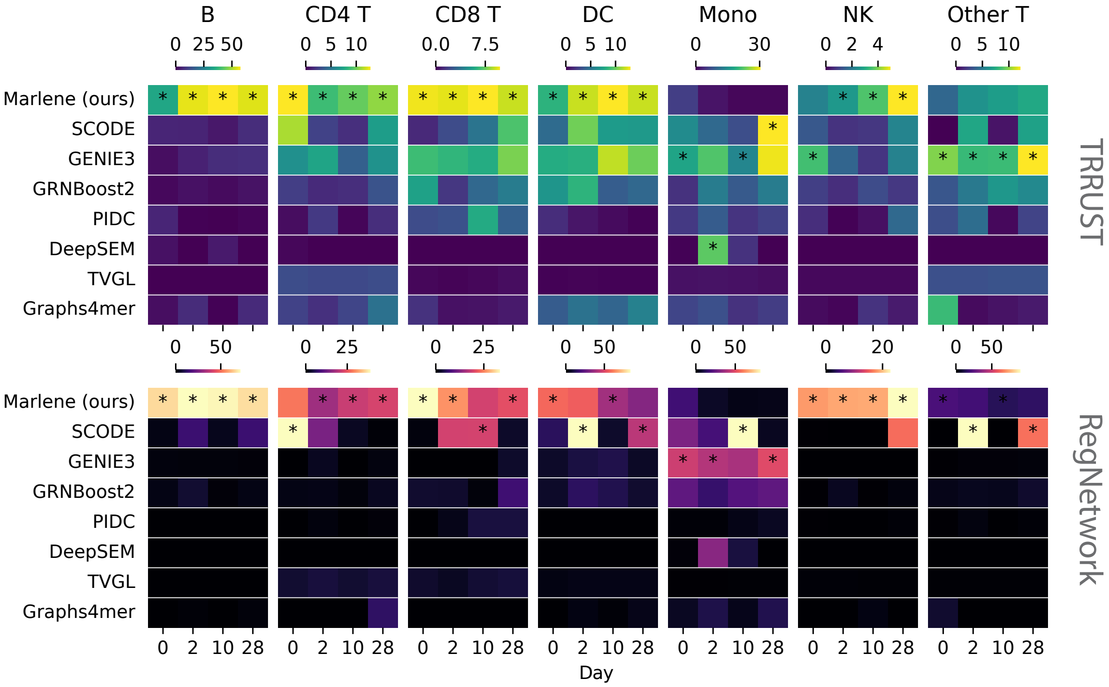


Fig. 3.8: **Overlap analysis of the SARS-CoV-2 vaccination dataset**. Showing $-\log_{10}(\text{FDR})$ values from a Fisher's exact test measuring the overlap between predicted TF-gene interactions in reconstructed networks and two TF-gene interaction databases, TRRUST (top) and RegNetwork (bottom). Cell types are shown as columns. Best performing method is starred.

transitions, likely due to their lack of dynamic modeling and the fact that they were ran independently per time point. TVGL, on the other hand, showed high IoU scores which remained close to constant over time. Finally, Graphs4mer displayed a reduction in IoU scores over time, which is unlikely given the expected immediate immune response.

Next, we sought to assess the quality of the TF-gene regulatory links added between time points. For brevity, we focused specifically on the initial temporal transition (days 0 → 2), as this period is likely to witness a more significant biological response. For each cell type, we took note of all the genes that were regulated by some TF at day 2 but not at day 0. Using this set of genes z, we performed gene set enrichment analysis (GSEA)[184,185] using the molecular signatures database (MSigDB)[135]. Through permutation tests, GSEA assigns an enrichment score (ES) to z reflecting its overrepresentation within the MSigDB gene set collection. We found that for many cell types, genes added by Marlene at day 2 greatly overlapped with COVID-19 and SARS-CoV-2 related gene sets. For instance "Interferon Gamma Response", which was identified as a SARS-CoV-2 antiviral response[296], was significantly enriched in dendritic cells (15 genes, FDR = 1−6). Similarly, "TNF-alpha Signaling via NF-kB"—a pathway involved in the immune response and inflammation[297]—was enriched in several cell types, as well as processes such as "Apoptosis" (cell death) and "p53 Pathway" (inhibits replication of infected cells)[298,299]. Other methods, while being enriched for relevant terms, showed a smaller gene overlap for these types (Fig. 3.9b) or were not consistent across cell types (e.g., DeepSEM, SCODE).

Overall, these results suggest that Marlene is able to capture both known TF-gene links, but also genes that are relevant to the response being studied.

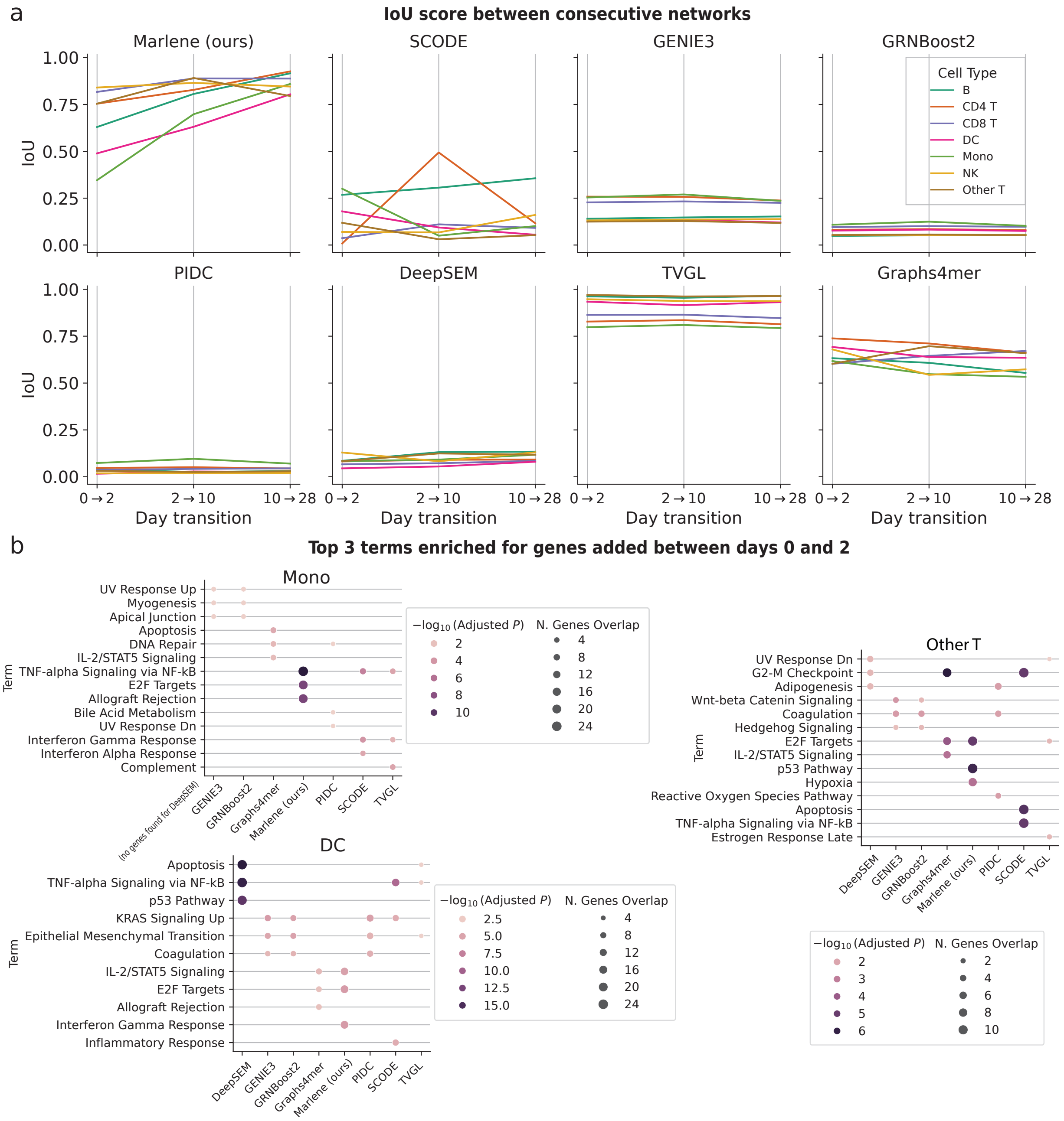


Fig. 3.9: **Temporal analysis of the predicted gene regulatory networks for the SARS-CoV-2 vaccine dataset**. (a) Intersection-over-union (IoU) scores between consecutive graphs. (b) For each method, top 3 MSigDB terms enriched for genes that were regulated at day 2 but not day 0.

Case study 2: Aging and senescence in the lung

The Human Lung Cell Atlas (HLCA) is a large data integration effort by the Human Cell Atlas Project[124,278]. This data combines scRNA-seq samples from 107 individuals spanning an age range of 10 to 76 years, making it particularly attractive for studying aging and senescence (a form of aging characterized by the absence of cell division)[4,300].

We split the atlas into three age groups at 35 and 50 years old, thus forming a pseudotime series of length 3. We removed smokers from the dataset as these will likely confound the results. To accommodate the data in the GPU, we randomly selected cells from 11 cell types, including type II pneumocytes, endothelial cells, and monocytes.

Similar to the vaccination dataset, we begin the analysis by evaluating the set of regulatory links using the TRRUST and RegNetwork databases. For this dataset we find that Marlene and SCODE are the top two performing methods (Fig. 3.10a). For some of the cell types, Marlene achieves significant results, recovering more than 1000 RegNetwork links (classical monocytes, FDR $= 1{-}76$). Even for cell types with fewer cells, such as non-classical monocytes (with only 138 cells for the second age group), Marlene still recovered more than 800 known TF-gene links for each transition (FDR $\leq 1{-}27$). SCODE performed well for some cell types such as CD1c-positive myeloid dendritic cells and CD4-positive, alpha-beta T cells. For all other methods, the overlap was smaller (Fig. C.10a). Note that while SCODE is comparable for the static network (single time point) inference task, it does not utilize dynamic information.

We next examined the ability of different methods to capture the dynamics of the biological processes. For this, we looked at graph transitions. IoU scores show that only the temporal methods (Marlene, TVGL, Graphs4mer) capture the smooth temporal transition between time points, while other methods, including SCODE, achieve low IoU scores (Fig. C.10b). We performed GSEA using Jensen Diseases gene set to see if genes added by Marlene in these transitions were enriched for any age-related diseases[136,301]. We found that Marlene added genes are enriched for several diseases such as arthritis, lung disease, and coronary artery disease. Other dynamic baselines were also enriched for relevant terms, but contained fewer marker genes (Fig. 3.10b).

Finally, we also investigated whether the genes regulated at different age groups were enriched for senescence. Cellular senescence refers to a permanent arrest of cell division triggered by the accumulation of DNA damage[302]. The absence of cell division can detrimentally impact tissue regeneration and repair, thereby contributing to various age-related diseases. Here, we use the SenMayo gene set which contains 125 genes reported to be enriched for senescence[303]. Only 81 of these genes overlapped with our data. We found that for 4 cell types, there was an increase in SenMayo gene regulation at the oldest age group (age > 50), suggesting that senescent cells accumulate with age as hypothesized (Fig. 3.11).

Case study 3: Fibrosis in a mouse lung injury model

Next, we evaluated whether Marlene could perform effectively across different species by analyzing a dataset from a mouse model of lung injury induced by the chemotherapeutic agent bleomycin[279]. The dataset included seven time points: one pre-treatment and six post-treatment intervals. After filtering out cell types with low representation and genes with low counts, we retained six cell types, including B cells, T cells, and macrophages.

In this analysis, Marlene outperformed competing methods in four of the six cell types when benchmarked against the RegNetwork database, specifically in alveolar epithelial cells, dendritic cells, endothelial cells, and macrophages. For T cells, TVGL showed slightly better performance. When evaluated against the TRRUST database, SCODE performed well in four cell types, while

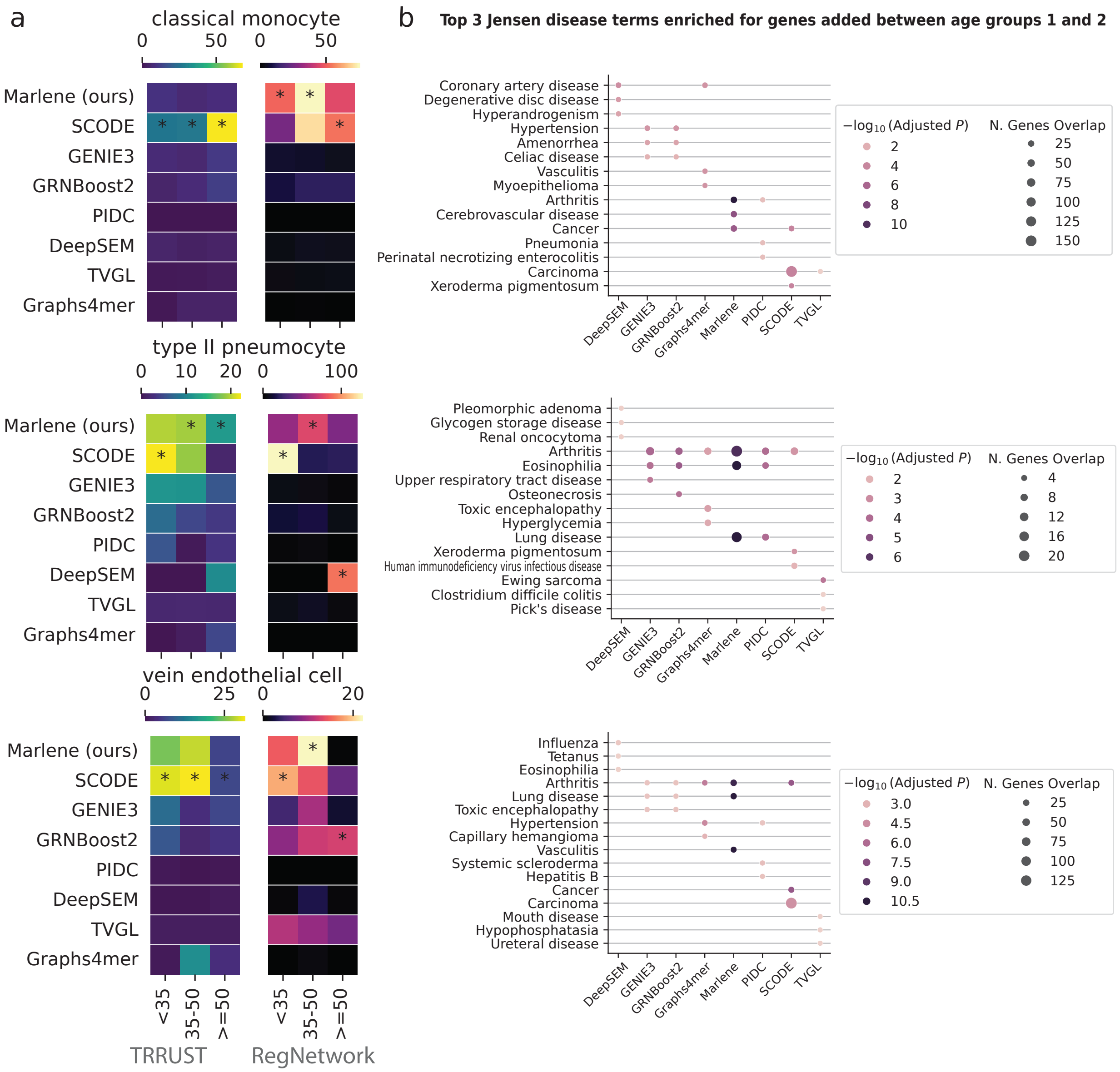


Fig. 3.10: **Results on the HLCA dataset**. (a) FDR corrected p-values of Fisher exact tests reflecting the number of links that overlap with TRRUST and RegNetwork databases. (b) Top 3 Jensen Diseases terms enriched for genes added between the first and second age group.

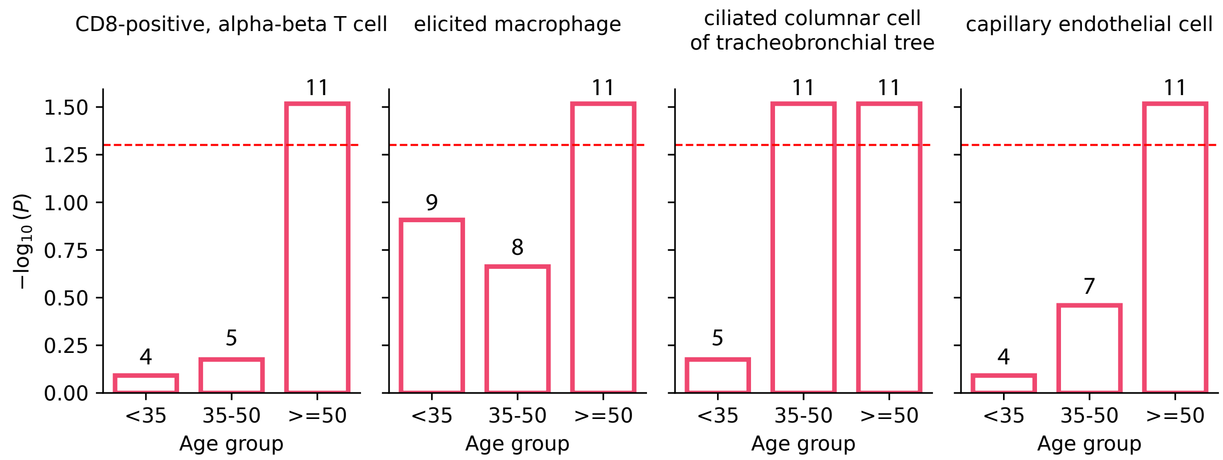


Fig. 3.11: **Enrichment for senescence using the SenMayo set**. For 4 cell types, there was statistically significant enrichment for the oldest age group. We only used the top 200 regulated genes.

Marlene surpassed it in the remaining two. The differing results between two databases may reflect their incomplete coverage, highlighting the need for further refinement.

Finally, all static baselines, including SCODE, showed low IoU scores across time points, indicating their inability to capture temporal evolution. In contrast, Marlene, showed increasing IoU scores over time, suggesting ongoing lung regeneration following treatment which slowly stabilizes. Figures illustrating these findings are provided in the appendix.

3.2.3 Discussion

Gene regulation is a dynamic process that underlies all biological systems. Understanding which TFs regulate which genes, and when this regulation occurs, provides insights into these dynamic processes which can lead to better treatment options. For instance, understanding what TF-gene links are disrupted could help researchers discover drugs targets for specific TF-gene connections.

To improve on current methods for reconstructing time varying regulatory networks, we use the expressive capabilities of deep neural networks to model the dynamic regulation of genes. Specifically, we focused on inferring dynamic networks from scRNA-seq data.

Our proposed method, Marlene, constructs dynamic graphs from time series data. Marlene begins with a set pooling operator based on PMA to extract gene features. These gene features are then used to construct dynamic graphs via a self-attention mechanism. The weights of the self-attention block are updated through the use of GRUs. Additionally, by employing MAML, we help Marlene uncover graphs even for rare cell types However, Marlene optimizes the prediction of cell type label rather than gene expression. As such, Marlene is not currently equipped to determine the impact of perturbations including gene knockouts or overexpression experiments. Exploring the integration of causal inference capabilities into Marlene represents a promising direction for future research.

We demonstrated the effectiveness of Marlene in recovering dynamic GRNs using three datasets: a SARS-CoV-2 vaccination dataset, a lung aging atlas, and a mouse dataset of fibrosis. In all three datasets, Marlene successfully identified many validated TF-gene links from the TRRUST and

RegNetwork databases across various cell types. It also accurately modeled the temporal dynamics of these connections. Some prior methods ignored the temporal aspect, leading to little similarity between consecutive networks. Other methods integrated all time points together, leading to very similar networks for each time point. In contrast, Marlene accurately recovered the variation dynamics, which is often characterized by strong rewiring following treatment that later stabilizes. In addition, Marlene identified many relevant edges. For instance, in the lung aging data, several dynamic edges were enriched for age-related diseases, such as arthritis. Meanwhile, in the SARS-CoV-2 data, these dynamic links were enriched for immune response processes. Prior methods captured some known edges, however, the overall results were less significant. By providing better models to explain disease and vaccine response, researchers can zoom in on the specific mechanisms targeted which in turn can lead to better treatments. Code will be made publicly available on publication.

3.2.4 Limitations

While successful, Marlene has a few limitations. The datasets we used in this study, while typical for scRNA-seq time series, consisted of only a few time points. For longer sequences, the GRU operation may suffer from vanishing gradient problems[304]. In such scenarios, the S4 module may be preferred as it has been shown to model long sequences better than traditional GRUs[295]. In addition, using a large number of genes for training, results in quadratic growth in memory consumption due to the need to store adjacency matrices. This led us to restrict the set of genes for each of the two studies. A more efficient implementation or alternative approaches such as FlashAttention[305] can lead to better ability to utilize all genes profiled.

3.2.5 Acknowledgements

The authors would like to thank Tanya Marwah for suggesting the Graphs4mer baseline.

Chapter 4

Discussion

Biomarker discovery is fundamental to understanding tissue composition, disease mechanisms, and developing treatment strategies. The context of the sample from which biomarkers are derived is one of the most important considerations. Key questions to address during marker discovery include:

- Are the classes in the sample well-defined?
- How significant is the potential for labeling errors, and to what extent might they affect the analysis?
- For unsupervised learning, does the inductive bias of the chosen clustering method align with the biological nature of the sample (e.g., different strategies may be necessary for clustering scRNA-seq versus ATAC-seq data)?
- Can prior knowledge be leveraged to simplify marker identification, such as through weakly supervised approaches with incomplete label information?
- If temporal dynamics are present in the data, how should the system's evolution over time be incorporated into the analysis?

Addressing these questions will guide the selection of the appropriate marker discovery strategy. This becomes especially important when studying complex processes like senescence. A useful analogy for understanding senescence is the infamous Blue Screen of Death (BSOD) in Windows systems. While the outcomes appeared identical, the underlying causes—such as hardware failures, overheating, corrupt drivers, or malware—were highly diverse. Similarly, senescence has evolved as a protective mechanism against DNA damage. Its triggers range from sun-induced irradiation and oxidative stress to genotoxic agents, each activating unique pathways but converging on the same endpoint: senescence.

Ignoring the specific context of a cell's function, type, tissue, and location risks identifying markers that fail to correspond with the precise biological response, underscoring the importance of context-sensitive approaches to marker discovery. This context sensitivity is not only critical for biomarker discovery but also has far-reaching implications for drug discovery and beyond.

The complexity of biomarker analysis is further amplified by the fact that many genes have multiple roles and are involved in diverse biological processes. While this multifaceted nature may not directly impact marker discovery—especially if only one process dominates the sample of interest—it significantly complicates the interpretation of the underlying biological mechanisms. For example, in GSEA analysis, a marker might appear enriched in a sample, but determining the specific biological process responsible for this enrichment can be challenging due to the gene's involvement in multiple pathways. This insight challenges the notion that biomarkers directly map to real biological states and highlights the need for domain expertise. For example, in a typical scRNA-seq analysis pipeline, the final interpretation often relies on the expert annotator's understanding of the sample, tissue, and the collective enrichment patterns of multiple genes. Even if biomarkers are more of an interpretive tool than a definitive representation of biology, their utility in bridging complex transcriptomic data to phenotypic outcomes remains undeniable.

Additionally, there are cases where a biological response is driven not by the activation but by the inactivation of a specific marker gene. Detecting such genes is particularly challenging in static scRNA-seq samples, as low-count or underexpressed genes are often excluded from analysis. Even when included, distinguishing between a gene being actively suppressed versus simply lowly expressed can be difficult. In such cases, time series data can provide valuable insights by enabling the detection of decreasing expression patterns across time points. This highlights the necessity of time series data for marker discovery in many biological processes. Consequently, this approach requires the development of tools with an inherent inductive bias to effectively capture and analyze the dynamic nature of these systems.

Lastly, due to the intricate interactions within a cell, the same gene can serve as a marker for a biological process in one specific endotype and not in another. Recognizing these endotypic differences is crucial in precision medicine, where a treatment may be effective for some patients but not for others. Trajectory inference algorithms have shown potential in identifying these branching trajectories, providing insights into the divergent biological pathways that underpin patient variability.

This thesis proposes several strategies to mitigate some of these issues. Our focus is on scRNA-seq data, where the classes (cell types) may or may not be well-defined.

For cases where cell types are known, we propose an approach based on set cover algorithms, **Greedy-PC**, to select genes that maximally separate these types while prioritizing the most informative markers. When prior information about classes is incomplete—such as in the case of senescence—we propose a framework that employs **PUc** learning to first separate the classes of interest. This is followed by DE analysis to identify marker genes between the now well-defined classes. In the absence of label information, unsupervised methods are required. To support this, we developed **Cellar**, a web server that facilitates end-to-end analysis of scRNA-seq data.

Additionally, we propose two algorithms tailored for analyzing time series data. The first, **Truffle**, identifies patient endotypes from clinical transcriptomics data, enabling the discovery of marker genes that differentiate between endotypes and shed light on patient variability. The second, **Marlene**, is an attention-based neural network designed to construct temporal GRNs. Marlene allows the study of how TF-gene regulatory links evolve over time, highlighting relevant TFs and genes that could serve as therapeutic targets.

4.1 Future Work

Several challenges and directions for future research remain. We discuss these below.

1. **Role of spatial context in marker discovery**. Spatial transcriptomics and proteomics are rapidly emerging as powerful tools for studying gene and protein expression within their tissue context. These approaches are particularly attractive for investigating GRNs, where cell-cell communication plays a role in determining which marker genes are activated in the receiving cell. By integrating gene expression data with the spatial location of cells—especially in tissues with complex architectures—it becomes possible to reconstruct more accurate GRNs that account for both intracellular regulation and intercellular interactions.

   Such methods inherently adopt a dual-network framework, simultaneously modeling the regulatory network of genes within individual cells and the interaction network between neighboring cells. Attention-based networks, like those employed in Marlene, provide a promising foundation for this dual-network approach. By extending these models to incorporate both cell-level and gene-level features, they could effectively capture the dynamic interplay of spatial and regulatory factors, paving the way for deeper insights into tissue-specific gene regulation.

2. **Data shifts in marker discovery**. Distribution shifts in biological data are massive, which complicates the validation of computational methods. It is often the case that methods perform well on certain datasets but fail on others due to these shifts. This underscores the need for developing methods that are robust to data shifts and that can effectively separate biological signal from technical or patient variability.

   Our proposed PUc approach for identifying senescence markers was one attempt to address this issue. We assumed that a shift occurs in aged individuals due to aging hallmarks such as inflammation or epigenetic alterations. However, in this work, we grouped patients into discrete age categories, treating these groups as well-defined classes to model shifts.

   A more natural and biologically accurate approach would be to model age as a continuous variable, reflecting the gradual and progressive nature of aging hallmarks rather than a stepwise change. Methods that account for progressive distribution shifts[306] are better suited to capture the smooth transitions inherent to aging processes. Investigating these approaches and comparing them against the discretized variant is a promising direction for future marker discovery efforts in senescence and beyond.

3. **Challenging the assumption that cell types are discrete**. Much of the computational biology literature relies on the core assumption that cell types represent distinct, well-defined classes. However, this assumption often oversimplifies the biological reality, where gene expression exists on a continuum, particularly in processes like differentiation. Most current approaches, such as DE analysis, are not inherently designed to handle such continuous transitions; instead, they test the hypothesis that two samples originate from two different distributions.

   Developing clustering, DE, and cell-type annotation methods that account for these "fuzzy" boundaries between cell types could significantly alter analytical outcomes of many studies. The ongoing discovery of new cell (sub)types—often representing subtle variations within the existing hierarchy of cell types—suggests that the true diversity of cell types might be vast. This raises the question of where to draw the boundaries of this continuum. Adopting methods that explicitly model these gradual transitions would help overcome these limitations.

4. **Foundation models for biomarker discovery**. Inspired by the success of large language foundation models such as ChatGPT[307,308], a number of scientific labs and industry institutions are exploring the potential of biological foundation models[24,25,309]. These models are trained on large biological datasets, such as scRNA-seq data, and leverage techniques pioneered in natural language processing. For instance, scGPT and Geneformer[24,25] employ transformer-based architectures, and train on gene expression data using variants of masked modeling. CellPLM[309] goes further by integrating spatial information with gene expression data, acknowledging that cell-cell interactions are as critical as gene expression.

   These models vary in several aspects:

   (a) **Input Representation**: How genes and/or cells are encoded for the model.

   (b) **Data Modalities**: The types of biological data used as input.

   (c) **Model Architecture**: The specific deep learning framework employed.

   Of these, input representation is arguably one of the most important. Most methods draw inspiration from language models, structuring genes and cells to mimic sentences of words. For example, Geneformer uses as input an ordered vector of the top 2000 regulated genes for each cell, masking some genes during training to predict their values based on the surrounding context. However, this approach ignores the nuanced expression levels of individual genes, and, furthermore, it is not clear how biologically relevant the exact order of these top regulated genes is. Similarly, scGPT[25] applies a "binning" step to group gene expression values into bins, in order to assign the same semantic value to genes falling in the same bin across cells. These methods highlight an ongoing challenge: finding biologically meaningful representation with a high inductive bias tailored to biological data.

   Foundation models offer several promising directions for advancing biomarker discovery.

   Robust Gene Embeddings. One promise of such models is the fact that they can generate robust gene embeddings that are less sensitive to technical variability and noise across datasets. This robustness can help identify markers that reliably activate under specific conditions. However, this same robustness might overlook critical individual-specific information, such as variations between endotypes. Careful validation that accounts for these nuances, including varying time points and conditions, is essential.

   Multi-Modal Insights. Models trained on multi-modal data such as scGPT, can identify markers in the context of the tissue and cellular microenvironments. Additionally, transfer learning enables these models to generalize insights from one domain (e.g., cancer biology) to another (e.g., senescence), thus helping identify shared or unique markers across biological contexts.

   Rare Cell Populations. Marker discovery for rare cell populations often suffers from limited statistical power in traditional DE analyses. Foundation models, due to their scalability and few-shot transfer learning capabilities, can overcome these limitations and capture fine-grained patterns.

   Biological foundation models represent a paradigm shift in computational biology, with the potential to make marker discovery more robust, scalable, and context-sensitive. Their applications in the field are truly exciting, with the potential to significantly enhance and accelerate progress in biomedical sciences.

Bibliography

1. Strimbu, K. & Tavel, J. A. What are biomarkers? en. *Curr. Opin. HIV AIDS* **5,** 463–466 (Nov. 2010).
2. HuBMAP Consortium. The human body at cellular resolution: the NIH Human Biomolecular Atlas Program. en. *Nature* **574,** 187–192 (Oct. 2019).
3. Börner, K. *et al.* Anatomical structures, cell types and biomarkers of the Human Reference Atlas. en. *Nat. Cell Biol.* **23,** 1117–1128 (Nov. 2021).
4. SenNet Consortium. NIH SenNet Consortium to map senescent cells throughout the human lifespan to understand physiological health. en. *Nat Aging* **2,** 1090–1100 (Dec. 2022).
5. Sapoval, N. *et al.* Current progress and open challenges for applying deep learning across the biosciences. en. *Nat. Commun.* **13,** 1728 (Apr. 2022).
6. Angermueller, C., Pärnamaa, T., Parts, L. & Stegle, O. Deep learning for computational biology. en. *Mol. Syst. Biol.* **12,** 878 (July 2016).
7. Sun, S., Zhu, J., Ma, Y. & Zhou, X. Accuracy, robustness and scalability of dimensionality reduction methods for single-cell RNA-seq analysis. en. *Genome Biol.* **20,** 269 (Dec. 2019).
8. Yu, L., Cao, Y., Yang, J. Y. H. & Yang, P. Benchmarking clustering algorithms on estimating the number of cell types from single-cell RNA-sequencing data. en. *Genome Biol.* **23,** 49 (Feb. 2022).
9. Saelens, W., Cannoodt, R., Todorov, H. & Saeys, Y. A comparison of single-cell trajectory inference methods. en. *Nat. Biotechnol.* **37,** 547–554 (May 2019).
10. La Manno, G. *et al.* RNA velocity of single cells. en. *Nature* **560,** 494–498 (Aug. 2018).
11. Bergen, V., Soldatov, R. A., Kharchenko, P. V. & Theis, F. J. RNA velocity-current challenges and future perspectives. en. *Mol. Syst. Biol.* **17,** e10282 (Aug. 2021).
12. Baltrusaitis, T., Ahuja, C. & Morency, L.-P. Multimodal machine learning: A survey and taxonomy. en. *IEEE Trans. Pattern Anal. Mach. Intell.* **41,** 423–443 (Feb. 2019).
13. Baysoy, A., Bai, Z., Satija, R. & Fan, R. The technological landscape and applications of single-cell multi-omics. en. *Nat. Rev. Mol. Cell Biol.* **24,** 695–713 (Oct. 2023).
14. Badia-I-Mompel, P. *et al.* Gene regulatory network inference in the era of single-cell multi-omics. en. *Nat. Rev. Genet.* **24,** 739–754 (Nov. 2023).
15. Liu, Z., Sun, D. & Wang, C. Evaluation of cell-cell interaction methods by integrating single-cell RNA sequencing data with spatial information. en. *Genome Biol.* **23,** 218 (Oct. 2022).

16. Munir, K., Elahi, H., Ayub, A., Frezza, F. & Rizzi, A. Cancer diagnosis using deep learning: A bibliographic review. en. *Cancers (Basel)* **11,** 1235 (Aug. 2019).

17. Abdelrahman, L., Al Ghamdi, M., Collado-Mesa, F. & Abdel-Mottaleb, M. Convolutional neural networks for breast cancer detection in mammography: A survey. en. *Comput. Biol. Med.* **131,** 104248 (Apr. 2021).

18. Yuan, Y. & Bar-Joseph, Z. GCNG: graph convolutional networks for inferring gene interaction from spatial transcriptomics data. en. *Genome Biol.* **21,** 300 (Dec. 2020).

19. Tran, H. T. N. *et al.* A benchmark of batch-effect correction methods for single-cell RNA sequencing data. en. *Genome Biol.* **21,** 12 (Jan. 2020).

20. Ding, J., Sharon, N. & Bar-Joseph, Z. Temporal modelling using single-cell transcriptomics. en. *Nat. Rev. Genet.* **23,** 355–368 (June 2022).

21. Ficek, J., Wang, W., Chen, H., Dagne, G. & Daley, E. Differential privacy in health research: A scoping review. en. *J. Am. Med. Inform. Assoc.* **28,** 2269–2276 (Sept. 2021).

22. Adnan, M., Kalra, S., Cresswell, J. C., Taylor, G. W. & Tizhoosh, H. R. Federated learning and differential privacy for medical image analysis. en. *Sci. Rep.* **12,** 1953 (Feb. 2022).

23. Guo, Z. *et al.* Diffusion models in bioinformatics: A new wave of deep learning revolution in action. *arXiv [cs.LG]* (Feb. 2023).

24. Theodoris, C. V. *et al.* Transfer learning enables predictions in network biology. en. *Nature* **618,** 616–624 (June 2023).

25. Cui, H. *et al.* scGPT: toward building a foundation model for single-cell multi-omics using generative AI. en. *Nat. Methods* **21,** 1470–1480 (Aug. 2024).

26. Campisi, J. & d'Adda di Fagagna, F. Cellular senescence: when bad things happen to good cells. en. *Nat. Rev. Mol. Cell Biol.* **8,** 729–740 (Sept. 2007).

27. Baker, D. J. *et al.* Clearance of p16Ink4a-positive senescent cells delays ageing-associated disorders. en. *Nature* **479,** 232–236 (Nov. 2011).

28. Baker, D. J. *et al.* Naturally Occurring p16Ink4a-positive Cells Shorten Healthy Lifespan. *Nature* **530,** 184–189 (2016).

29. Coppé, J.-P., Desprez, P.-Y., Krtolica, A. & Campisi, J. The senescence-associated secretory phenotype: the dark side of tumor suppression. en. *Annu. Rev. Pathol.* **5,** 99–118 (2010).

30. Acosta, J. C. *et al.* A complex secretory program orchestrated by the inflammasome controls paracrine senescence. en. *Nat. Cell Biol.* **15,** 978–990 (Aug. 2013).

31. Ozsolak, F. & Milos, P. M. RNA sequencing: advances, challenges and opportunities. en. *Nat. Rev. Genet.* **12,** 87–98 (Feb. 2011).

32. Haque, A., Engel, J., Teichmann, S. A. & Lönnberg, T. A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. en. *Genome Med.* **9,** 75 (Aug. 2017).

33. Salzberg, S. L. Open questions: How many genes do we have? en. *BMC Biol.* **16,** 94 (Aug. 2018).

34. Moses, L. & Pachter, L. Museum of spatial transcriptomics. en. *Nat. Methods* **19,** 534–546 (May 2022).

35. Goltsev, Y. *et al.* Deep profiling of mouse splenic architecture with CODEX multiplexed imaging. en. *Cell* **174,** 968–981.e15 (Aug. 2018).

36. Lacar, B. *et al.* Nuclear RNA-seq of single neurons reveals molecular signatures of activation. en. *Nat. Commun.* **7,** 11022 (Apr. 2016).

37. Chen, W. *et al.* Live-seq enables temporal transcriptomic recording of single cells. en. *Nature* **608,** 733–740 (Aug. 2022).

38. Zhou, Z.-H. A brief introduction to weakly supervised learning. en. *Natl. Sci. Rev.* **5,** 44–53 (Jan. 2018).

39. Lötvall, J. *et al.* Asthma endotypes: a new approach to classification of disease entities within the asthma syndrome. en. *J. Allergy Clin. Immunol.* **127,** 355–360 (Feb. 2011).

40. Leighton, T. *et al.* Fast approximation algorithms for multicommodity flow problems. en. *J. Comput. Syst. Sci.* **50,** 228–243 (Apr. 1995).

41. Zeng, H. What is a cell type and how to define it? en. *Cell* **185,** 2739–2755 (July 2022).

42. Hasanaj, E., Alavi, A., Gupta, A., Póczos, B. & Bar-Joseph, Z. Multiset multicover methods for discriminative marker selection. en. *Cell Rep. Methods* **2,** 100332 (Nov. 2022).

43. Wilcoxon, F. Individual comparisons by ranking methods. *Biom. Bull.* **1,** 80 (Dec. 1945).

44. Consortium, T. M. *et al.* Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. *Nature* **562,** 367–372 (2018).

45. Janeway Jr, C. A., Travers, P., Walport, M. & Shlomchik, M. J. in *Immunobiology: The Immune System in Health and Disease. 5th edition* (Garland Science, 2001).

46. Heath, W. R. in *Encyclopedia of Immunology (Second Edition)* (ed Delves, P. J.) Second Edition, 2341–2343 (Elsevier, Oxford, 1998). ISBN: 978-0-12-226765-9.

47. Ravkov, E., Slev, P. & Heikal, N. Thymic output: Assessment of CD4+ recent thymic emigrants and T-Cell receptor excision circles in infants. *Cytometry Part B: Clinical Cytometry* **92,** 249–257 (2017).

48. Ronning, K. E., Karlen, S. J., Miller, E. B. & Burns, M. E. Molecular profiling of resident and infiltrating mononuclear phagocytes during rapid adult retinal degeneration using single-cell RNA sequencing. *Scientific reports* **9,** 1–12 (2019).

49. Newman, A. M. *et al.* Robust Enumeration of Cell Subsets from Tissue Expression Profiles. *Nature Methods 2015 12:5* **12,** 453–457 (Mar. 2015).

50. Gong, T. *et al.* Optimal Deconvolution of Transcriptional Profiling Data Using Quadratic Programming with Application to Complex Clinical Blood Samples. *PLOS ONE* **6,** e27156 (Nov. 2011).

51. Bolón-Canedo, V., Sánchez-Maroño, N., Alonso-Betanzos, A., Benítez, J. & Herrera, F. A Review of Microarray Datasets and Applied Feature Selection Methods. *Information Sciences* **282,** 111–135 (Oct. 2014).

52. Tadist, K., Najah, S., Nikolov, N. S., Mrabti, F. & Zahi, A. Feature Selection Methods and Genomic Big Data: A Systematic Review. *Journal of Big Data* **6,** 1–24 (Dec. 2019).

53. Saeys, Y., Inza, I. & Larrañaga, P. A Review of Feature Selection Techniques in Bioinformatics. *Bioinformatics* **23,** 2507–2517 (Oct. 2007).

54. Whitney, A. A Direct Method of Nonparametric Measurement Selection. *IEEE Transactions on Computers* **C-20,** 1100–1103 (Sept. 1971).

55. Marill, T. & Green, D. On the Effectiveness of Receptors in Recognition Systems. *IEEE Transactions on Information Theory* **9,** 11–17 (Jan. 1963).

56. Breiman, L., Friedman, J. H., Olshen, R. A. & Stone, C. J. *Classification And Regression Trees* (Routledge, New York, 1984).

57. Dumitrascu, B., Villar, S., Mixon, D. G. & Engelhardt, B. E. Optimal Marker Gene Selection for Cell Type Discrimination in Single Cell Analyses. *Nature Communications* **12,** 1186 (Feb. 2021).

58. Vargo, A. H. S. & Gilbert, A. C. A Rank-Based Marker Selection Method for High Throughput scRNA-seq Data. *BMC Bioinformatics* **21,** 477 (Oct. 2020).

59. Kira, K. & Rendell, L. A. *A Practical Approach to Feature Selection* in *Proceedings of the Ninth International Workshop on Machine Learning* (Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, July 1992), 249–256.

60. Kononenko, I. in *Machine Learning: ECML-94* (eds Carbonell, J. G. *et al.*) 171–182 (Springer Berlin Heidelberg, Berlin, Heidelberg, 1994).

61. Peng, H., Long, F. & Ding, C. Feature Selection Based on Mutual Information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **27,** 1226–1238 (Aug. 2005).

62. Abbas, A. R., Wolslegel, K., Seshasayee, D., Modrusan, Z. & Clark, H. F. Deconvolution of Blood Microarray Data Identifies Cellular Activation Patterns in Systemic Lupus Erythematosus. *PLOS ONE* **4,** e6098 (July 2009).

63. Gong, T. & Szustakowski, J. D. DeconRNASeq: A Statistical Framework for Deconvolution of Heterogeneous Tissue Samples Based on mRNA-Seq Data. *Bioinformatics* **29,** 1083–1085 (Apr. 2013).

64. Vazirani, V. V. *Approximation Algorithms* (Springer, Berlin, Heidelberg, 2003).

65. Rajagopalan, S. & Vazirani, V. V. Primal-Dual RNC Approximation Algorithms for (Multi)-Set (Multi)-Cover and Covering Integer Programs. *Annual Symposium on Foundatons of Computer Science (Proceedings),* 322–331 (1993).

66. Rubinstein, R. Y. Optimization of Computer Simulation Models with Rare Events. *European Journal of Operational Research* **99,** 89–112 (May 1997).

67. De Boer, P. T., Kroese, D. P., Mannor, S. & Rubinstein, R. Y. A Tutorial on the Cross-Entropy Method. *Annals of Operations Research 2005 134:1* **134,** 19–67 (Jan. 2005).

68. Quinlan, J. R. Induction of Decision Trees. *Machine Learning* **1,** 81–106 (Mar. 1986).

69. Kozachenko, L. F. & Leonenko, N. N. Sample Estimate of the Entropy of a Random Vector. *Problemy Peredachi Informatsii,* 9–16 (1987).

70. Kraskov, A., Stögbauer, H. & Grassberger, P. Estimating Mutual Information. *Physical Review E* **69,** 066138 (June 2004).

71. Adams, T. S. *et al.* Single-Cell RNA-seq Reveals Ectopic and Aberrant Lung-Resident Cell Populations in Idiopathic Pulmonary Fibrosis. *Science Advances* **6,** eaba1983 (2020).

72. Zeisel, A. *et al.* Cell Types in the Mouse Cortex and Hippocampus Revealed by Single-Cell RNA-seq. *Science* **347,** 1138–1142 (Mar. 2015).

73. He, S. *et al.* Single-Cell Transcriptome Profiling of an Adult Human Cell Atlas of 15 Major Organs. *Genome Biology* **21,** 294 (Dec. 2020).

74. Segerstolpe, Å. *et al.* Single-Cell Transcriptome Profiling of Human Pancreatic Islets in Health and Type 2 Diabetes. *Cell Metabolism* **24,** 593–607 (Oct. 2016).

75. Muraro, M. J. *et al.* A Single-Cell Transcriptome Atlas of the Human Pancreas. *Cell Systems* **3,** 385–394.e3 (Oct. 2016).

76. Tsoucas, D. *et al.* Accurate Estimation of Cell-Type Composition from Gene Expression Data. *Nature Communications 2019 10:1* **10,** 1–9 (July 2019).

77. Lin, J. Divergence Measures Based on the Shannon Entropy. *IEEE Transactions on Information Theory* **37,** 145–151 (Jan. 1991).

78. Nogueira, S., Sechidis, K. & Brown, G. On the Stability of Feature Selection Algorithms. *Journal of Machine Learning Research* **18,** 1–54 (2018).

79. Börner, K. *et al.* Anatomical Structures, Cell Types and Biomarkers of the Human Reference Atlas. *Nature Cell Biology* **23,** 1117–1128 (Nov. 2021).

80. Subramanian, A. *et al.* Gene Set Enrichment Analysis: A Knowledge-Based Approach for Interpreting Genome-Wide Expression Profiles. *Proceedings of the National Academy of Sciences* **102,** 15545–15550 (Oct. 2005).

81. Coulombe, P. A. & Wong, P. Cytoplasmic intermediate filaments revealed as dynamic and multipurpose scaffolds. *Nature Cell Biology 2004 6:8* **6,** 699–706. ISSN: 1476-4679 (8 Aug. 2004).

82. Saha, S. K., Kim, K., Yang, G. M., Choi, H. Y. & Cho, S. G. Cytokeratin 19 (KRT19) has a Role in the Reprogramming of Cancer Stem Cell-Like Cells to Less Aggressive and More Drug-Sensitive Cells. *International Journal of Molecular Sciences* **19,** 19. ISSN: 14220067 (5 May 2018).

83. Kubo, F. *et al.* Loss of the adhesion G-protein coupled receptor ADGRF5 in mice induces airway inflammation and the expression of CCL2 in lung endothelial cells 11 Medical and Health Sciences 1102 Cardiorespiratory Medicine and Haematology. *Respiratory Research* **20,** 1–21. ISSN: 1465993X (1 Jan. 2019).

84. Vazquez, B. N., Laguna, T., Carabana, J., Krangel, M. S. & Lauzurica, P. CD69 gene is differentially regulated in T and B cells by evolutionarily conserved promoter-distal elements. *Journal of immunology (Baltimore, Md. : 1950)* **183,** 6513. ISSN: 0022-1767 (10 Nov. 2009).

85. Ziegler, S. F., Ramsdell, F. & Alderson, M. R. The activation antigen CD69. *Stem cells (Dayton, Ohio)* **12,** 456–465. ISSN: 1066-5099 (5 1994).

86. Plešingerová, H. *et al.* Expression of COBLL1 encoding novel ROR1 binding partner is robust predictor of survival in chronic lymphocytic leukemia. *Haematologica* **103,** 313–324. ISSN: 1592-8721 (2 Jan. 2018).

87. Castro, C. D. & Flajnik, M. F. Putting J-chain back on the map: how might its expression define plasma cell development? *Journal of immunology (Baltimore, Md. : 1950)* **193,** 3248. ISSN: 0022-1767 (7 Oct. 2014).

88. Plaen, I. G. D. *et al.* Lipopolysaccharide induces CXCL2/macrophage inflammatory protein-2 gene expression in enterocytes via NF-kappaB activation: independence from endogenous TNF-alpha and platelet-activating factor. *Immunology* **118,** 153–163. ISSN: 0019-2805 (2 June 2006).

89. Robertson, M. J. Role of chemokines in the biology of natural killer cells. *Journal of Leukocyte Biology* **71,** 173–183 (2002).

90. Lee, S. M. *et al.* Characterisation of diverse PRF1 mutations leading to decreased natural killer cell activity in North American families with haemophagocytic lymphohistiocytosis. *Journal of Medical Genetics* **41,** 137–144. ISSN: 0022-2593 (2 Feb. 2004).

91. Valés-Gómez, M. *et al.* Natural killer cell hyporesponsiveness and impaired development in a CD247-deficient patient. *Journal of Allergy and Clinical Immunology* **137,** 942–945.e4. ISSN: 0091-6749 (3 Mar. 2016).

92. Vanderbilt, J. N. *et al.* CXC chemokines and their receptors are expressed in type II cells and upregulated following lung injury. *American journal of respiratory cell and molecular biology* **29,** 661–668. ISSN: 1044-1549 (6 Dec. 2003).

93. Shi, Y. *et al.* AFF3 upregulation mediates tamoxifen resistance in breast cancers. *Journal of Experimental & Clinical Cancer Research : CR* **37.** ISSN: 17569966 (1 Oct. 2018).

94. Maher, K., Konjar, S., Watts, C., Turk, B. & Kopitar-Jerala, N. Cystatin F regulates proteinase activity in IL-2-activated natural killer cells. *Protein and peptide letters* **21,** 957–965. ISSN: 1875-5305 (9 July 2014).

95. Ronchetti, S. *et al.* Glucocorticoid-Induced Tumour Necrosis Factor Receptor-Related Protein: A Key Marker of Functional Regulatory T Cells. *Journal of Immunology Research* **2015.** ISSN: 23147156 (2015).

96. Johnson, D. S. Approximation Algorithms for Combinatorial Problems. *Journal of Computer and System Sciences* **9,** 256–278 (Dec. 1974).

97. Chvatal, V. A Greedy Heuristic for the Set-Covering Problem. *Mathematics of Operations Research* **4,** 233–235 (1979).

98. Rubinstein, R. The Cross-Entropy Method for Combinatorial and Continuous Optimization. *Methodology And Computing In Applied Probability* **1,** 127–190 (Sept. 1999).

99. Welch, B. L. The Generalisation of Student's Problems When Several Different Population Variances Are Involved. *Biometrika* **34,** 28–35 (1947).

100. Johnson, J., Douze, M. & Jégou, H. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data* **7,** 535–547 (2019).

101. Kullback, S. & Leibler, R. A. On Information and Sufficiency. *The Annals of Mathematical Statistics* **22,** 79–86 (Mar. 1951).

102. Chen, E. Y. *et al.* Enrichr: Interactive and Collaborative HTML5 Gene List Enrichment Analysis Tool. *BMC bioinformatics* **14,** 128 (Apr. 2013).

103. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: Large-scale Single-Cell Gene Expression Data Analysis. *Genome Biology* **19,** 15 (Feb. 2018).

104. Fu, Y., Huang, X., Zhang, P., van de Leemput, J. & Han, Z. Single-Cell RNA Sequencing Identifies Novel Cell Types in Drosophila Blood. *Journal of genetics and genomics = Yi chuan xue bao* **47,** 175–186 (Apr. 2020).

105. Shekhar, K. & Menon, V. in *Computational Methods for Single-Cell Data Analysis* (ed Yuan, G.-C.) 45–77 (Springer, New York, NY, 2019).

106. Wilkerson, B. A. *et al.* Novel Cell Types and Developmental Lineages Revealed by Single-Cell RNA-seq Analysis of the Mouse Crista Ampullaris. *eLife* **10** (eds Whitfield, T. T. & Bronner, M. E.) e60108 (May 2021).

107. Wu, H., Kirita, Y., Donnelly, E. L. & Humphreys, B. D. Advantages of Single-Nucleus over Single-Cell RNA Sequencing of Adult Kidney: Rare Cell Types and Novel Cell States Revealed in Fibrosis. *Journal of the American Society of Nephrology* **30,** 23–32 (Jan. 2019).

108. Hayflick, L. & Moorhead, P. S. The serial cultivation of human diploid cell strains. en. *Exp. Cell Res.* **25,** 585–621 (Dec. 1961).

109. Carbon, S. *et al.* AmiGO: online access to ontology and annotation data. en. *Bioinformatics* **25,** 288–289 (Jan. 2009).

110. Fridman, A. L. & Tainsky, M. A. Critical pathways in cellular senescence and immortalization revealed by gene expression profiling. en. *Oncogene* **27,** 5975–5987 (Oct. 2008).

111. Saul, D. *et al.* A new gene set identifies senescent cells and predicts senescence-associated pathways across tissues. en. *Nat. Commun.* **13,** 4827 (Aug. 2022).

112. Avelar, R. A. *et al.* A multidimensional systems biology analysis of cellular senescence in aging and disease. en. *Genome Biol.* **21,** 91 (Apr. 2020).

113. Bekker, J. & Davis, J. Learning from positive and unlabeled data: a survey. en. *Mach. Learn.* **109,** 719–760 (Apr. 2020).

114. Scott, C. *A rate of convergence for mixture proportion estimation, with application to learning from noisy labels* en. in *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics* (eds Lebanon, G. & Vishwanathan, S. V. N.) **38** (PMLR, San Diego, California, USA, Feb. 2015), 838–846.

115. Garg, S., Wu, Y., Smola, A., Balakrishnan, S. & Lipton, Z. C. Mixture Proportion Estimation and PU learning: A modern approach. *arXiv [cs.LG]* (Nov. 2021).

116. Kiryo, R., Niu, G., du Plessis, M. C. & Sugiyama, M. Positive-Unlabeled Learning with Non-Negative Risk Estimator. *Advances in Neural Information Processing Systems* **30** (2017).

117. Li, X. *et al.* Inflammation and aging: signaling pathways and intervention therapies. en. *Signal Transduct. Target. Ther.* **8,** 239 (June 2023).

118. Wang, K. *et al.* Epigenetic regulation of aging: implications for interventions of aging and diseases. en. *Signal Transduct. Target. Ther.* **7,** 374 (Nov. 2022).

119. Amorim, J. A. *et al.* Mitochondrial and metabolic dysfunction in ageing and age-related diseases. en. *Nat. Rev. Endocrinol.* **18,** 243–258 (Apr. 2022).

120. Dharani., G., Nair, N. G., Satpathy, P. & Christopher, J. *Covariate shift: A review and analysis on classifiers* in *2019 Global Conference for Advancement in Technology (GCAT)* (IEEE, Oct. 2019), 1–6.

121. Sakai, T. & Shimizu, N. Covariate shift adaptation on learning from positive and unlabeled data. en. *Proc. Conf. AAAI Artif. Intell.* **33,** 4838–4845 (July 2019).

122. Sugiyama, M., Krauledat, M. & Müller, K.-R. Covariate shift adaptation by importance weighted cross validation. *J. Mach. Learn. Res.* **8,** 985–1005 (2007).

123. Sugiyama, M., Suzuki, T. & Kanamori, T. *Density Ratio Estimation in Machine Learning* (Cambridge University Press, Feb. 2012).

124. Sikkema, L. *et al.* An integrated cell atlas of the lung in health and disease. en. *Nat. Med.* **29,** 1563–1577 (June 2023).

125. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. en. *Nat. Genet.* **25,** 25–29 (May 2000).

126. Ortiz-Montero, P., Londoño-Vallejo, A. & Vernot, J.-P. Senescence-associated IL-6 and IL-8 cytokines induce a self- and cross-reinforced senescence/inflammatory milieu strengthening tumorigenic capabilities in the MCF-7 breast cancer cell line. en. *Cell Commun. Signal.* **15,** 17 (May 2017).

127. Zhang, L.-M. *et al.* Interleukin-18 promotes fibroblast senescence in pulmonary fibrosis through down-regulating Klotho expression. en. *Biomed. Pharmacother.* **113,** 108756 (May 2019).

128. Dinarello, C. A. Interleukin 1 and interleukin 18 as mediators of inflammation and the aging process. en. *Am. J. Clin. Nutr.* **83,** 447S–455S (Feb. 2006).

129. Siraj, Y. *et al.* IGFBP7 is a key component of the senescence-associated secretory phenotype (SASP) that induces senescence in healthy cells by modulating the insulin, IGF, and activin A pathways. en. *Cell Commun. Signal.* **22,** 540 (Nov. 2024).

130. Hettiarachchi, G. K. *et al.* Translational and transcriptional responses in human primary hepatocytes under hypoxia. en. *Am. J. Physiol. Gastrointest. Liver Physiol.* **316,** G720–G734 (June 2019).

131. Chinenov, Y. & Kerppola, T. K. Close encounters of many kinds: Fos-Jun interactions that mediate transcription regulatory specificity. en. *Oncogene* **20,** 2438–2452 (Apr. 2001).

132. Liu, Y.-F. *et al.* NME2 reduces proliferation, migration and invasion of gastric cancer cells to limit metastasis. en. *PLoS One* **10,** e0115968 (Feb. 2015).

133. Box, J. K. *et al.* Nucleophosmin: from structure and function to disease development. en. *BMC Mol. Biol.* **17,** 19 (Aug. 2016).

134. Mansur, N. R., Meyer-Siegler, K., Wurzer, J. C. & Sirover, M. A. Cell cycle regulation of the glyceraldehyde-3-phosphate dehydrogenase/uracil DNA glycosylase gene in normal human cells. en. *Nucleic Acids Res.* **21,** 993–998 (Feb. 1993).

135. Liberzon, A. *et al.* The Molecular Signatures Database (MSigDB) hallmark gene set collection. en. *Cell Syst.* **1,** 417–425 (Dec. 2015).

136. Grissa, D., Junge, A., Oprea, T. I. & Jensen, L. J. Diseases 2.0: a weekly updated database of disease-gene associations from text mining and data integration. en. *Database (Oxford)* **2022,** baac019 (Mar. 2022).

137. Yao, C. *et al.* Senescence of alveolar type 2 cells drives progressive pulmonary fibrosis. en. *Am. J. Respir. Crit. Care Med.* **203,** 707–717 (Mar. 2021).

138. Lin, Y. & Xu, Z. Fibroblast senescence in idiopathic pulmonary fibrosis. en. *Front. Cell Dev. Biol.* **8,** 593283 (Nov. 2020).

139. Ruzankina, Y. & Brown, E. J. Relationships between stem cell exhaustion, tumour suppression and ageing. en. *Br. J. Cancer* **97,** 1189–1193 (Nov. 2007).

140. Uhl, F. E. *et al.* Preclinical validation and imaging of Wnt-induced repair in human 3D lung tissue cultures. en. *Eur. Respir. J.* **46,** 1150–1166 (Oct. 2015).

141. Alsafadi, H. N. *et al.* An ex vivo model to induce early fibrosis-like changes in human precision-cut lung slices. en. *Am. J. Physiol. Lung Cell. Mol. Physiol.* **312,** L896–L902 (June 2017).

142. Alsafadi, H. N. *et al.* Applications and approaches for three-dimensional precision-cut lung slices. Disease modeling and drug discovery. en. *Am. J. Respir. Cell Mol. Biol.* **62,** 681–691 (June 2020).

143. Melo Narvaez, M. C. *et al.* *An ex vivo model of cellular senescence and inflammaging in precision-cut lung slices* en. in *Mechanisms of lung injury and repair* **8** (European Respiratory Society, Mar. 2022).

144. Travaglini, K. J. *et al.* A molecular cell atlas of the human lung from single-cell RNA sequencing. en. *Nature* **587,** 619–625 (Nov. 2020).

145. Tsuji, T., Aoshiba, K. & Nagai, A. Alveolar cell senescence exacerbates pulmonary inflammation in patients with chronic obstructive pulmonary disease. en. *Respiration* **80,** 59–70 (2010).

146. Tsuji, T., Aoshiba, K. & Nagai, A. Alveolar cell senescence in patients with pulmonary emphysema. en. *Am. J. Respir. Crit. Care Med.* **174,** 886–893 (Oct. 2006).

147. Meiners, S., Eickelberg, O. & Königshoff, M. Hallmarks of the ageing lung. en. *Eur. Respir. J.* **45,** 807–827 (Mar. 2015).

148. Easter, M., Bollenbecker, S., Barnes, J. W. & Krick, S. Targeting aging pathways in chronic obstructive pulmonary disease. en. *Int. J. Mol. Sci.* **21,** 6924 (Sept. 2020).

149. Christenson, S. A., Smith, B. M., Bafadhel, M. & Putcha, N. Chronic obstructive pulmonary disease. en. *Lancet* **399,** 2227–2242 (June 2022).

150. Serrano, M., Lin, A. W., McCurrach, M. E., Beach, D. & Lowe, S. W. Oncogenic ras provokes premature cell senescence associated with accumulation of p53 and p16INK4a. en. *Cell* **88,** 593–602 (Mar. 1997).

151. Duan, J., Duan, J., Zhang, Z. & Tong, T. Irreversible cellular senescence induced by prolonged exposure to H2O2 involves DNA-damage-and-repair genes and telomere shortening. en. *Int. J. Biochem. Cell Biol.* **37,** 1407–1420 (July 2005).

152. Aoshiba, K., Tsuji, T. & Nagai, A. Bleomycin induces cellular senescence in alveolar epithelial cells. en. *Eur. Respir. J.* **22,** 436–443 (Sept. 2003).

153. Fitsiou, E., Soto-Gamez, A. & Demaria, M. Biological functions of therapy-induced senescence in cancer. en. *Semin. Cancer Biol.* **81,** 5–13 (June 2022).

154. Hernandez-Segura, A., Nehme, J. & Demaria, M. Hallmarks of Cellular Senescence. en. *Trends Cell Biol.* **28,** 436–453 (June 2018).

155. Álvarez, D. *et al.* IPF lung fibroblasts have a senescent phenotype. en. *Am. J. Physiol. Lung Cell. Mol. Physiol.* **313,** L1164–L1173 (Dec. 2017).

156. Schafer, M. J. *et al.* Cellular senescence mediates fibrotic pulmonary disease. en. *Nat. Commun.* **8,** 14532 (Feb. 2017).

157. Parekh, K. R. *et al.* Stem cells and lung regeneration. en. *Am. J. Physiol. Cell Physiol.* **319,** C675–C693 (Oct. 2020).

158. Kelley, K. D. *et al.* YPEL3, a p53-regulated gene that induces cellular senescence. en. *Cancer Res.* **70,** 3566–3575 (May 2010).

159. Zhu, H. *et al.* SPOP E3 ubiquitin ligase adaptor promotes cellular senescence by degrading the SENP7 deSUMOylase. en. *Cell Rep.* **13,** 1183–1193 (Nov. 2015).

160. Yang, W. *et al.* Substrate-dependent interaction of SPOP and RACK1 aggravates cardiac fibrosis following myocardial infarction. en. *Cell Chem. Biol.* **30,** 1248–1260.e4 (Oct. 2023).

161. Angiolilli, C. *et al.* ZFP36 family members regulate the proinflammatory features of psoriatic dermal fibroblasts. en. *J. Invest. Dermatol.* **142,** 402–413 (Feb. 2022).

162. Cicchetto, A. C. *et al.* ZFP36-mediated mRNA decay regulates metabolism. en. *Cell Rep.* **42,** 112411 (May 2023).

163. Xu, P. *et al.* The landscape of human tissue and cell type specific expression and co-regulation of senescence genes. en. *Mol. Neurodegener.* **17,** 5 (Jan. 2022).

164. Chou, Y.-T. *et al.* CITED2 functions as a molecular switch of cytokine-induced proliferation and quiescence. en. *Cell Death Differ.* **19,** 2015–2028 (Dec. 2012).

165. Hu, C. *et al.* Downregulation of CITED2 contributes to TGFβ-mediated senescence of tendon-derived stem cells. en. *Cell Tissue Res.* **368,** 93–104 (Apr. 2017).

166. Alsaleh, G. *et al.* Innate immunity triggers IL-32 expression by fibroblast-like synoviocytes in rheumatoid arthritis. en. *Arthritis Res. Ther.* **12,** R135 (July 2010).

167. Wen, S. *et al.* Cancer-associated fibroblast (CAF)-derived IL32 promotes breast cancer cell invasion and metastasis via integrin β3-p38 MAPK signalling. en. *Cancer Lett.* **442,** 320–332 (Feb. 2019).

168. Aass, K. R. *et al.* Intracellular IL-32 regulates mitochondrial metabolism, proliferation, and differentiation of malignant plasma cells. en. *iScience* **25,** 103605 (Jan. 2022).

169. Kwon, E. *et al.* The RNA-binding protein YBX1 regulates epidermal progenitors at a post-transcriptional level. en. *Nat. Commun.* **9,** 1734 (Apr. 2018).

170. Bernstein, A. M., Twining, S. S., Warejcka, D. J., Tall, E. & Masur, S. K. Urokinase receptor cleavage: a crucial step in fibroblast-to-myofibroblast differentiation. en. *Mol. Biol. Cell* **18,** 2716–2727 (July 2007).

171. Amor, C. *et al.* Senolytic CAR T cells reverse senescence-associated pathologies. en. *Nature* **583,** 127–132 (July 2020).

172. Lasorella, A., Iavarone, A. & Israel, M. A. Id2 specifically alters regulation of the cell cycle by tumor suppressor proteins. en. *Mol. Cell. Biol.* **16,** 2570–2578 (June 1996).

173. Wang, W. *et al.* Fibroblast A20 governs fibrosis susceptibility and its repression by DREAM promotes fibrosis in multiple organs. en. *Nat. Commun.* **13,** 6358 (Oct. 2022).

174. Li, Y. *et al.* TXNIP exacerbates the senescence and aging-related dysfunction of β cells by inducing cell cycle arrest through p38-p16/p21-CDK-Rb pathway. en. *Antioxid. Redox Signal.* **38,** 480–495 (Mar. 2023).

175. Conte, M. *et al.* GDF15, an emerging key player in human aging. en. *Ageing Res. Rev.* **75,** 101569 (Mar. 2022).

176. Basisty, N. *et al.* A proteomic atlas of senescence-associated secretomes for aging biomarker development. en. *PLoS Biol.* **18,** e3000599 (Jan. 2020).

177. Schafer, M. J. *et al.* The senescence-associated secretome as an indicator of age and medical risk. en. *JCI Insight* **5** (June 2020).

178. Muralikrishnan, V. *et al.* A novel ALDH1A1 inhibitor blocks platinum-induced senescence and stemness in ovarian cancer. en. *Cancers (Basel)* **14,** 3437 (July 2022).

179. Ding, Y., Liu, H., Zhang, C., Bao, Z. & Yu, S. Polo-like kinases as potential targets and PLK2 as a novel biomarker for the prognosis of human glioblastoma. en. *Aging (Albany NY)* **14,** 2320–2334 (Mar. 2022).

180. Jun, J.-I. & Lau, L. F. CCN2 induces cellular senescence in fibroblasts. en. *J. Cell Commun. Signal.* **11,** 15–23 (Mar. 2017).

181. Wong, J. C. T. *et al.* Nucleophosmin 1, upregulated in adenomas and cancers of the colon, inhibits p53-mediated cellular senescence: NPM1 inhibits p53-mediated cellular senescence. en. *Int. J. Cancer* **133,** 1567–1577 (Oct. 2013).

182. Shen, X. *et al.* Nonlinear dynamics of multi-omics profiles during human aging. en. *Nat. Aging,* 1–16 (Aug. 2024).

183. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society series b-methodological* **57,** 289–300 (1995).

184. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. en. *Proc. Natl. Acad. Sci. U. S. A.* **102,** 15545–15550 (Oct. 2005).

185. Fang, Z., Liu, X. & Peltz, G. GSEApy: a comprehensive package for performing gene set enrichment analysis in Python. en. *Bioinformatics* **39,** btac757 (Jan. 2023).

186. Lopez, R., Regier, J., Cole, M. B., Jordan, M. I. & Yosef, N. Deep generative modeling for single-cell transcriptomics. en. *Nat. Methods* **15,** 1053–1058 (Dec. 2018).

187. Traag, V. A., Waltman, L. & van Eck, N. J. From Louvain to Leiden: guaranteeing well-connected communities. en. *Sci. Rep.* **9,** 5233 (Mar. 2019).

188. Heard, N. A. & Rubin-Delanchy, P. Choosing between methods of combining *p*-values. en. *Biometrika* **105,** 239–246 (Mar. 2018).

189. Hasanaj, E., Wang, J., Sarathi, A., Ding, J. & Bar-Joseph, Z. Interactive single-cell data analysis using Cellar. en. *Nat. Commun.* **13,** 1998 (Apr. 2022).

190. McInnes, L., Healy, J., Saul, N. & Großberger, L. UMAP: Uniform Manifold Approximation and Projection. *J. Open Source Softw.* **3,** 861 (Sept. 2018).

191. Ogata, H. *et al.* KEGG: Kyoto Encyclopedia of genes and genomes. en. *Nucleic Acids Res.* **27,** 29–34 (Jan. 1999).

192. Mason, D. Y. *et al.* CD79a: a novel marker for B-cell neoplasms in routinely processed tissue samples. en. *Blood* **86,** 1453–1459 (Aug. 1995).

193. Smulski, C. R. & Eibel, H. BAFF and BAFF-receptor in B cell selection and survival. en. *Front. Immunol.* **9,** 2285 (Oct. 2018).

194. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. en. *Genome Biol.* **19,** 15 (Feb. 2018).

195. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. en. *Nat. Methods* **10,** 1213–1218 (Dec. 2013).

196. Bravo González-Blas, C. *et al.* cisTopic: cis-regulatory topic modeling on single-cell ATAC-seq data. en. *Nat. Methods* **16,** 397–400 (May 2019).

197. Blei, D. M., Ng, A. Y. & Jordan, M. I. Latent dirichlet allocation. en. *J. Mach. Learn. Res.* **3,** 993–1022 (Mar. 2003).

198. Genomics, 1. *Peripheral Blood Mononuclear Cells (PBMCs) from a healthy donor (v1)* https://www.10xgenomics.com/resources/datasets/10-k-peripheral-blood-mononuclear-cells-pbm-cs-from-a-healthy-donor-1-standard-1-1-0. 2020. (2020).

199. Bongen, E., Vallania, F., Utz, P. J. & Khatri, P. KLRD1-expressing natural killer cells predict influenza susceptibility. en. *Genome Med.* **10,** 45 (June 2018).

200. Chen, S., Lake, B. B. & Zhang, K. High-throughput sequencing of the transcriptome and chromatin accessibility in the same cell. en. *Nat. Biotechnol.* **37,** 1452–1457 (Dec. 2019).

201. Gopal, E. *et al.* Cloning and functional characterization of human SMCT2 (SLC5A12) and expression pattern of the transporter in kidney. en. *Biochim. Biophys. Acta* **1768,** 2690–2697 (Nov. 2007).

202. Molitoris, B. A. & Wagner, M. C. Surface membrane polarity of proximal tubular cells: alterations as a basis for malfunction. en. *Kidney Int.* **49,** 1592–1597 (June 1996).

203. Frankish, A. *et al.* GENCODE reference annotation for the human and mouse genomes. en. *Nucleic Acids Res.* **47,** D766–D773 (Jan. 2019).

204. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *arXiv [cs.LG]* (Jan. 2012).

205. Douze, M. *et al.* The Faiss library. *arXiv [cs.LG]* (Jan. 2024).

206. Hunter, D. J. Gene-environment interactions in human diseases. en. *Nat. Rev. Genet.* **6,** 287–298 (Apr. 2005).

207. Deng, H., Yan, X. & Yuan, L. Human genetic basis of coronavirus disease 2019. en. *Signal Transduct. Target. Ther.* **6,** 344 (Sept. 2021).

208. Hasanaj, E., Mathur, S. & Bar-Joseph, Z. Integrating patients in time series clinical transcriptomics data. en. *Bioinformatics* **40,** i151–i159 (June 2024).

209. Meyer, U. A., Zanger, U. M. & Schwab, M. Omics and Drug Response. en (Jan. 2013).

210. Wang, Y. *et al.* Changing technologies of RNA sequencing and their applications in clinical oncology. en. *Front. Oncol.* **10,** 447 (Apr. 2020).

211. Huang, T. *et al.* Prediction of pharmacological and xenobiotic responses to drugs based on time course gene expression profiles. en. *PLoS One* **4,** e8126 (Dec. 2009).

212. Almon, R. R., DuBois, D. C., Pearson, K. E., Stephan, D. A. & Jusko, W. J. Gene arrays and temporal patterns of drug response: corticosteroid effects on rat liver. en. *Funct. Integr. Genomics* **3,** 171–179 (Dec. 2003).

213. Bar-Joseph, Z., Gitter, A. & Simon, I. Studying and modelling dynamic biological processes using time-series gene expression data. en. *Nat. Rev. Genet.* **13,** 552–564 (July 2012).

214. Battaglia et al, M. Introducing the endotype concept to address the challenge of disease heterogeneity in type 1 diabetes. en. *Diabetes Care* **43,** 5–12 (Jan. 2020).

215. Czarnowicki, T., He, H., Krueger, J. G. & Guttman-Yassky, E. Atopic dermatitis endotypes and implications for targeted therapeutics. en. *J. Allergy Clin. Immunol.* **143,** 1–11 (Jan. 2019).

216. Listgarten, J., Neal, R., Roweis, S. & Emili, A. Multiple Alignment of Continuous Time Series. *Advances in Neural Information Processing Systems* **17** (2004).

217. Lin, T.-H., Kaminski, N. & Bar-Joseph, Z. Alignment and classification of time series gene expression in clinical studies. en. *Bioinformatics* **24,** i147–55 (July 2008).

218. Bar-Joseph, Z., Gerber, G. K., Gifford, D. K., Jaakkola, T. S. & Simon, I. Continuous representations of time-series gene expression data. en. *J. Comput. Biol.* **10,** 341–356 (2003).

219. Behnke, M. S. *et al.* Coordinated progression through two subtranscriptomes underlies the tachyzoite cycle of Toxoplasma gondii. en. *PLoS One* **5,** e12354 (Aug. 2010).

220. Czarnewski, P. *et al.* Conserved transcriptomic profile between mouse and human colitis allows unsupervised patient stratification. en. *Nat. Commun.* **10,** 2892 (June 2019).

221. Lange, M. *et al.* CellRank for directed single-cell fate mapping. en. *Nat. Methods* **19,** 159–170 (Feb. 2022).

222. Tran, T. N. & Bader, G. D. Tempora: Cell trajectory inference using time-series single-cell RNA sequencing data. en. *PLoS Comput. Biol.* **16,** e1008205 (Sept. 2020).

223. Macnair, W., Gupta, R. & Claassen, M. psupertime: supervised pseudotime analysis for time-series single-cell RNA-seq data. en. *Bioinformatics* **38,** i290–i298 (June 2022).

224. Liu, J. *et al.* Transcriptomic profiling of plaque psoriasis and cutaneous T-cell subsets during treatment with secukinumab. en. *JID Innov.* **2,** 100094 (May 2022).

225. LaSalle et al, T. J. Longitudinal characterization of circulating neutrophils uncovers phenotypes associated with severity in hospitalized COVID-19 patients. en. *Cell Reports Medicine* **3** (Oct. 2022).

226. VanDussen, K. L. *et al.* Abnormal small intestinal epithelial microvilli in patients with crohn's disease. en. *Gastroenterology* **155,** 815–828 (Sept. 2018).

227. Robinson, M. D. & Oshlack, A. A scaling normalization method for differential expression analysis of RNA-seq data. en. *Genome Biol.* **11,** R25 (Mar. 2010).

228. Johnson, W. E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. en. *Biostatistics* **8,** 118–127 (Apr. 2006).

229. Hao, Y. *et al.* Dictionary learning for integrative, multimodal and scalable single-cell analysis. en. *Nat. Biotechnol.,* 1–12 (May 2023).

230. Massey, F. J. The kolmogorov-smirnov test for goodness of fit. *J. Am. Stat. Assoc.* **46,** 68 (Mar. 1951).

231. Charikar, M., Naamad, Y., Rexford, J. & Zou, X. K. en. in *Algorithmic Aspects of Cloud Computing* 73–101 (Springer International Publishing, Cham, 2019).

232. Even, S., Itai, A. & Shamir, A. *On the complexity of time table and multi-commodity flow problems* in *16th Annual Symposium on Foundations of Computer Science (sfcs 1975)* (IEEE, Oct. 1975), 184–193.

233. Bynum, M. L. *et al. Pyomo — optimization modeling in python* 3rd ed. en (Springer Nature, Cham, Switzerland, Mar. 2021).

234. Oki, E. in *Linear Programming and Algorithms for Communication Networks* 19–38 (CRC Press, Aug. 2012).

235. Meindl, B. & Templ, M. Analysis of Commercial and Free and Open Source Solvers for the Cell Suppression Problem. *Trans. Data Priv.* **6,** 147–159 (2013).

236. Ernst, J. & Bar-Joseph, Z. STEM: a tool for the analysis of short time series gene expression data. en. *BMC Bioinformatics* **7,** 191 (Apr. 2006).

237. Ma, F. *et al.* Single cell and spatial sequencing define processes by which keratinocytes and fibroblasts amplify inflammatory responses in psoriasis. en. *Nat. Commun.* **14,** 3455 (June 2023).

238. Parma, V. *et al.* More than smell-COVID-19 is associated with severe impairment of smell, taste, and chemesthesis. en. *Chem. Senses* **45,** 609–622 (Oct. 2020).

239. Povroznik, J. M. & Robinson, C. M. IL-27 regulation of innate immunity and control of microbial growth. en. *Future Sci. OA* **6,** FSO588 (June 2020).

240. Shibata, S. *et al.* Possible roles of IL-27 in the pathogenesis of psoriasis. en. *J. Invest. Dermatol.* **130,** 1034–1039 (Apr. 2010).

241. Wagner, M. F. M. G., Theodoro, T. R., Filho, C. D. A. S. M., Oyafuso, L. K. M. & Pinhal, M. A. S. Extracellular matrix alterations in the skin of patients affected by psoriasis. en. *BMC Mol. Cell Biol.* **22,** 55 (Oct. 2021).

242. Hasanaj, E., Póczos, B. & Bar-Joseph, Z. Recovering time-varying networks from single-cell data. *arXiv [q-bio.QM]* (Oct. 2024).

243. Silverman, E. K. *et al.* Molecular networks in Network Medicine: Development and applications. en. *Wiley Interdiscip. Rev. Syst. Biol. Med.* **12,** e1489 (Nov. 2020).

244. Karlebach, G. & Shamir, R. Modelling and analysis of gene regulatory networks. en. *Nat. Rev. Mol. Cell Biol.* **9,** 770–780 (Oct. 2008).

245. Gasch, A. P. *et al.* Genomic expression programs in the response of yeast cells to environmental changes. en. *Mol. Biol. Cell* **11,** 4241–4257 (Dec. 2000).

246. Luscombe, N. M. *et al.* Genomic analysis of regulatory network dynamics reveals large topological changes. en. *Nature* **431,** 308–312 (Sept. 2004).

247. Lee, T. I. *et al.* Transcriptional Regulatory Networks in Saccharomyces Cerevisiae. *Science* **298,** 799–804 (Oct. 2002).

248. Wang, Y., Joshi, T., Zhang, X.-S., Xu, D. & Chen, L. Inferring Gene Regulatory Networks from Multiple Microarray Datasets. *Bioinformatics* **22,** 2413–2420 (Oct. 2006).

249. Blais, A. & Dynlacht, B. D. Constructing Transcriptional Regulatory Networks. *Genes & Development* **19,** 1499–1511 (July 2005).

250. Gilchrist, D. A., Fargo, D. C. & Adelman, K. Using ChIP-chip and ChIP-seq to Study the Regulation of Gene Expression: Genome-wide Localization Studies Reveal Widespread Regulation of Transcription Elongation. *Methods. Analysis of RNA Polymerase II Elongation* **48,** 398–408 (Aug. 2009).

251. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: A Revolutionary Tool for Transcriptomics. *Nature Reviews Genetics* **10,** 57–63 (Jan. 2009).

252. Ding, J., Sharon, N. & Bar-Joseph, Z. Temporal Modelling Using Single-Cell Transcriptomics. *Nature Reviews Genetics* **23,** 355–368 (June 2022).

253. Nguyen, H., Tran, D., Tran, B., Pehlivan, B. & Nguyen, T. A Comprehensive Survey of Regulatory Network Inference Methods Using Single Cell RNA Sequencing Data. *Briefings in Bioinformatics* **22,** bbaa190 (May 2021).

254. Bar-Joseph, Z. *et al.* Computational Discovery of Gene Modules and Regulatory Networks. *Nature Biotechnology* **21,** 1337–1342 (Nov. 2003).

255. Yosef, N. *et al.* Dynamic Regulatory Network Controlling TH17 Cell Differentiation. *Nature* **496,** 461–468 (Apr. 2013).

256. Schulz, M. H. *et al.* DREM 2.0: Improved Reconstruction of Dynamic Regulatory Networks from Time-Series Expression Data. *BMC Systems Biology* **6,** 104 (Aug. 2012).

257. Yan, M. *et al.* Dynamic Regulatory Networks of T Cell Trajectory Dissect Transcriptional Control of T Cell State Transition. *Molecular Therapy - Nucleic Acids* **26,** 1115–1129 (Dec. 2021).

258. Ahmed, A. & Xing, E. P. Recovering Time-Varying Networks of Dependencies in Social and Biological Studies. *Proceedings of the National Academy of Sciences* **106,** 11878–11883 (July 2009).

259. Song, L., Kolar, M. & Xing, E. *Time-Varying Dynamic Bayesian Networks* in *Advances in Neural Information Processing Systems* **22** (Curran Associates, Inc., 2009).

260. Dondelinger, F., Lèbre, S. & Husmeier, D. Non-Homogeneous Dynamic Bayesian Networks with Bayesian Regularization for Inferring Gene Regulatory Networks with Gradually Time-Varying Structure. *Machine Learning* **90,** 191–230 (Feb. 2013).

261. Zhu, S. & Wang, Y. Hidden Markov Induced Dynamic Bayesian Network for Recovering Time Evolving Gene Regulatory Networks. *Scientific Reports* **5,** 17841 (Dec. 2015).

262. Wang, H. *et al.* Time-Varying Gene Network Analysis of Human Prefrontal Cortex Development. *Frontiers in Genetics* **11** (2020).

263. Hallac, D., Park, Y., Boyd, S. & Leskovec, J. Network inference via the time-varying graphical lasso. en. *KDD* **2017,** 205–213 (Aug. 2017).

264. Kim, M.-S., Kim, J.-R., Kim, D., Lander, A. D. & Cho, K.-H. Spatiotemporal Network Motif Reveals the Biological Traits of Developmental Gene Regulatory Networks in Drosophila Melanogaster. *BMC Systems Biology* **6,** 31 (May 2012).

265. Ching, T. *et al.* Opportunities and obstacles for deep learning in biology and medicine. en. *J. R. Soc. Interface* **15** (Apr. 2018).

266. Shrivastava, H., Zhang, X., Song, L. & Aluru, S. GRNUlar: A deep learning framework for recovering single-cell gene regulatory networks. en. *J. Comput. Biol.* **29,** 27–44 (Jan. 2022).

267. Shu, H. *et al.* Modeling gene regulatory networks using neural network architectures. en. *Nat. Comput. Sci.* **1,** 491–501 (July 2021).

268. Wang, L. *et al.* Dictys: dynamic gene regulatory network dissects developmental continuum with single-cell multiomics. en. *Nat. Methods* **20,** 1368–1378 (Sept. 2023).

269. Kamimoto, K. *et al.* Dissecting cell identity via network inference and in silico gene perturbation. en. *Nature* **614,** 742–751 (Feb. 2023).

270. Zhu, Y. *et al.* A survey on Graph Structure Learning: Progress and opportunities. *arXiv [cs.LG]* (Mar. 2021).

271. Zhang, Q., Chang, J., Meng, G., Xiang, S. & Pan, C. Spatio-temporal graph structure learning for traffic forecasting. en. *Proc. Conf. AAAI Artif. Intell.* **34,** 1177–1185 (Apr. 2020).

272. Zaheer, M. *et al.* Deep Sets. *Advances in Neural Information Processing Systems* **30** (2017).

273. Lee, J. *et al. Set Transformer: A Framework for Attention-based Permutation-Invariant Neural Networks* in *Proceedings of the 36th International Conference on Machine Learning* (eds Chaudhuri, K. & Salakhutdinov, R.) **97** (PMLR, 2019), 3744–3753.

274. Bahdanau, D., Cho, K. & Bengio, Y. Neural machine translation by jointly learning to align and translate. *arXiv [cs.CL]* (Sept. 2014).

275. Pareja, A. *et al.* EvolveGCN: Evolving graph convolutional networks for dynamic graphs. en. *Proc. Conf. AAAI Artif. Intell.* **34,** 5363–5370 (Apr. 2020).

276. Finn, C., Abbeel, P. & Levine, S. Model-agnostic meta-learning for fast adaptation of deep networks. *arXiv [cs.LG]. Proceedings of Machine Learning Research* (eds Precup, D. & Teh, Y. W.) 1126–1135 (Mar. 2017).

277. Zhang, B. *et al.* Multimodal single-cell datasets characterize antigen-specific CD8+ T cells across SARS-CoV-2 vaccination and infection. en. *Nat. Immunol.* **24,** 1725–1734 (Oct. 2023).

278. Regev, A. *et al.* The Human Cell Atlas. en. *Elife* **6** (Dec. 2017).

279. Strunz, M. *et al.* Alveolar regeneration through a Krt8+ transitional stem cell state that persists in human lung fibrosis. en. *Nat. Commun.* **11,** 3559 (July 2020).

280. Wise, A. & Bar-Joseph, Z. cDREM: inferring dynamic combinatorial gene regulation. en. *J. Comput. Biol.* **22,** 324–333 (Apr. 2015).

281. Seninge, L., Anastopoulos, I., Ding, H. & Stuart, J. VEGA is an interpretable generative model for inferring biological network activity in single-cell transcriptomics. en. *Nat. Commun.* **12,** 5684 (Sept. 2021).

282. Wang, J., Chen, Y. & Zou, Q. Inferring gene regulatory network from single-cell transcriptomes with graph autoencoder model. en. *PLoS Genet.* **19,** e1010942 (Sept. 2023).

283. Han, H. *et al.* TRRUST v2: an expanded reference database of human and mouse transcriptional regulatory interactions. en. *Nucleic Acids Res.* **46,** D380–D386 (Jan. 2018).

284. Jindal, A., Gupta, P., Jayadeva & Sengupta, D. Discovery of rare cells from voluminous single cell expression data. en. *Nat. Commun.* **9,** 4719 (Nov. 2018).

285. Chasman, D. & Roy, S. Inference of cell type specific regulatory networks on mammalian lineages. en. *Curr. Opin. Syst. Biol.* **2,** 130–139 (Apr. 2017).

286. Kingma, D. P. & Ba, J. Adam: A method for stochastic optimization. *arXiv [cs.LG]* (Dec. 2014).

287. Liu, Z.-P., Wu, C., Miao, H. & Wu, H. RegNetwork: an integrated database of transcriptional and post-transcriptional regulatory networks in human and mouse. en. *Database (Oxford)* **2015,** bav095 (Sept. 2015).

288. Fisher, R. A. On the interpretation of χ^2 from contingency tables, and the calculation of P. *J. R. Stat. Soc.* **85,** 87 (Jan. 1922).

289. Pratapa, A., Jalihal, A. P., Law, J. N., Bharadwaj, A. & Murali, T. M. Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. en. *Nat. Methods* **17,** 147–154 (Feb. 2020).

290. Chan, T. E., Stumpf, M. P. H. & Babtie, A. C. Gene regulatory network inference from single-cell data using multivariate information measures. en. *Cell Syst.* **5,** 251–267.e3 (Sept. 2017).

291. Huynh-Thu, V. A., Irrthum, A., Wehenkel, L. & Geurts, P. Inferring regulatory networks from expression data using tree-based methods. en. *PLoS One* **5,** e12776 (Sept. 2010).

292. Moerman, T. *et al.* GRNBoost2 and Arboreto: efficient and scalable inference of gene regulatory networks. en. *Bioinformatics* **35,** 2159–2161 (June 2019).

293. Matsumoto, H. *et al.* SCODE: an efficient regulatory network inference algorithm from single-cell RNA-Seq during differentiation. en. *Bioinformatics* **33,** 2314–2321 (Aug. 2017).

294. Tang, S. *et al.* Modeling multivariate biosignals with graph neural networks and Structured State Space models. *arXiv [cs.LG]* (Nov. 2022).

295. Gu, A., Goel, K. & Ré, C. Efficiently modeling long sequences with structured state spaces. *arXiv [cs.LG]* (Oct. 2021).

296. Hilligan, K. L. *et al.* Bacterial-induced or passively administered interferon gamma conditions the lung for early control of SARS-CoV-2. en. *Nat. Commun.* **14,** 8229 (Dec. 2023).

297. Hayden, M. S. & Ghosh, S. NF-κB in immunobiology. en. *Cell Res.* **21,** 223–244 (Feb. 2011).

298. Elmore, S. Apoptosis: a review of programmed cell death. en. *Toxicol. Pathol.* **35,** 495–516 (June 2007).

299. Harris, S. L. & Levine, A. J. The p53 pathway: positive and negative feedback loops. en. *Oncogene* **24,** 2899–2908 (Apr. 2005).

300. Van Deursen, J. M. The role of senescent cells in ageing. en. *Nature* **509,** 439–446 (May 2014).

301. Pletscher-Frankild, S., Pallejà, A., Tsafou, K., Binder, J. X. & Jensen, L. J. DISEASES: text mining and data integration of disease-gene associations. en. *Methods* **74,** 83–89 (Mar. 2015).

302. Suryadevara, V. *et al.* SenNet recommendations for detecting senescent cells in different tissues. en. *Nat. Rev. Mol. Cell Biol.,* 1–23 (June 2024).

303. Saul, D. *et al.* A new gene set identifies senescent cells and predicts senescence-associated pathways across tissues. en. *Nat. Commun.* **13,** 4827 (Aug. 2022).

304. Pascanu, R., Mikolov, T. & Bengio, Y. On the difficulty of training Recurrent Neural Networks. *arXiv [cs.LG]* (Nov. 2012).

305. Dao, T., Fu, D. Y., Ermon, S., Rudra, A. & Ré, C. FlashAttention: Fast and memory-efficient exact attention with IO-awareness. *arXiv [cs.LG]* (May 2022).

306. Zhang, Y.-J., Zhang, Z.-Y., Zhao, P. & Sugiyama, M. Adapting to continuous covariate shift via online density ratio estimation. *arXiv [cs.LG]* (Feb. 2023).

307. OpenAI. *ChatGPT (Nov 22 version) [Large language model]* https://chat.openai.com/chat. 2024.

308. OpenAI. GPT-4 Technical Report. *arXiv [cs.CL]* (Mar. 2023).

309. Wen, H. *et al. CellPLM: Pre-training of Cell Language Model Beyond Single Cells* in *The Twelfth International Conference on Learning Representations* (Oct. 2023).

310. Vaswani, A. *et al.* Attention is All you Need. *Advances in Neural Information Processing Systems* **30** (2017).

311. Cangea, C., Veličković, P., Jovanović, N., Kipf, T. & Liò, P. Towards sparse hierarchical graph classifiers. *arXiv [stat.ML]* (Nov. 2018).

312. Paszke, A. *et al.* PyTorch: An imperative style, high-performance deep learning library. *arXiv [cs.LG]* (Dec. 2019).

A Supervised Setting: Minimal Gene Sets for Accurate Cell Type Identification

Supplementary Algorithms

Algorithm 1: Greedy Phenotype Cover

```
Data: Signature matrix $\mathbf{M} \in \mathbb{R}^{P \times F}$, coverage factor $K > 0$
Result: List of selected features indices $\mathcal{S} \subset [F]$
$\mathcal{E} \leftarrow$ empty list;
$\mathcal{D} \leftarrow$ dictionary mapping every pair $(i, j)$ to some unique integer;
for $s \leftarrow 1$ to $F$ do
    $\mathcal{E}_s \leftarrow$ empty list;
    for every ordered pair $(i, j)$ in $[P]$ do
        if $i = j$ then continue;
        coverage $\leftarrow \mathbf{M}[i][s] - \mathbf{M}[j][s]$;
        if coverage $> 0$ then $\mathcal{E}_s$.append(Pair($\mathcal{D}[(i, j)]$, coverage)) ;
        else if coverage $< 0$ then $\mathcal{E}_s$.append(Pair($\mathcal{D}[(j, i)]$, $-$coverage)) ;
    end
    $\mathcal{E}$.append ($\mathcal{E}_s$);
end
$\mathcal{S} \leftarrow$ MultisetMulticover(Union($\mathcal{E}$), $\mathcal{E}$, $K$);
```

Algorithm 2: Multiset Multicover

```
input  : Universe $\mathcal{U} = \{e_1, \ldots, e_n\}$, collection of multisets $\mathcal{E} = \{\mathcal{E}_1, \ldots, \mathcal{E}_F\}$, coverage
         $K > 0$
output: Indices of selected sets $\mathcal{S} \subset [F]$
Function MultisetMulticover($\mathcal{U}, \mathcal{E}, K$):
    leftToCover $\leftarrow$ $K$-initialized list of length $n$;
    setValues $\leftarrow$ zero-initialized list of length $F$;
    $\mathcal{S} \leftarrow$ empty list;
    while max(leftToCover)$> 0$ do
        for $s \leftarrow 1$ to $F$ do
            sv $\leftarrow 0$;
            for $i \leftarrow 1$ to length($\mathcal{E}[s]$) do
                (element, multiplicity) $\leftarrow$ getElementAndMult($\mathcal{E}[s][i]$);
                sv $+=$ min(multiplicity, leftToCover[element]);
            end
            setValues$[s] \leftarrow$ sv;
        end
        bestSet $\leftarrow$ argmax(setValues);
        $\mathcal{S}$.append(bestSet);
        for $i \leftarrow 1$ to length($\mathcal{E}$[bestSet]) do
            (element, multiplicity) $\leftarrow$ getElementAndMult($\mathcal{E}[s][i]$);
            leftToCover$[i] \leftarrow$ max(leftToCover$[i]$ $-$ multiplicity, $0$);
        end
        clearAllElements($\mathcal{E}$[bestSet]);
    end
    return $\mathcal{S}$;
```

Algorithm 3: Cross-Entropy Method

```
Data: Signature matrix $\mathbf{M} \in \mathbb{R}^{P \times F}$, coverage factor $K$, maximum number of iterations $I$,
      number of random samples per iteration $R_s$, smoothing parameter $\theta$, quantile $\rho$,
      mixing parameter $\alpha$
Result: List of selected features indices $\mathcal{S} \subset [F]$
$\widehat{\mathbf{M}} \leftarrow$ ComputePairMatrix($\mathbf{M}$);
$\widehat{\mathbf{p}} \leftarrow$ list of length $F$ with every element initialized to 0.5;
for $i \leftarrow 1$ to $I$ do
    Sample $\mathbf{X}_1, \dots, \mathbf{X}_{R_s} \sim$ Bernoulli($\widehat{\mathbf{p}}$)   # feature $s$ is a Bernoulli($\widehat{\mathbf{p}}[s]$)
    scores $\leftarrow$ [Score($\mathbf{X}_1$, $\widehat{\mathbf{M}}$),...,Score($\mathbf{X}_{R_s}$, $\widehat{\mathbf{M}}$)];
    $\widehat{\gamma} \leftarrow$ Quantile(scores, $1 - \rho$);
    successes $\leftarrow$ number of scores that are $\geq \widehat{\gamma}$;
    $\widehat{\mathbf{p}}_{new} \leftarrow$ zero-initialized list of length $F$;
    for $j \leftarrow 1$ to $R_s$ do
        if scores[$j$] $\geq \widehat{\gamma}$ then
            $\widehat{\mathbf{p}}_{new}$ += ($\mathbf{X}_j$/successes);
        end
    end
    $\widehat{\mathbf{p}} \leftarrow \theta \cdot \widehat{\mathbf{p}}_{new} + (1 - \theta) \cdot \widehat{\mathbf{p}}$;
    if Converged($\widehat{\mathbf{p}}$, $\widehat{\mathbf{p}}_{history}$) then break;
end
$\mathcal{S} \leftarrow$ [indices $j$ for which $\widehat{\mathbf{p}}[j] > 0.98$];

Function ComputePairMatrix($\mathbf{M}$):
    $\widehat{\mathbf{M}} \leftarrow$ zero-initialized matrix of shape $(P(P-1), F)$;
    $\mathcal{D} \leftarrow$ dictionary mapping every pair $(i, j)$ to some unique integer in $[1, P(P-1)]$;
    for every ordered pair $(i, j)$ in $[P]$ do
        if $i = j$ then continue;
        coverage $\leftarrow \mathbf{M}[i, :] - \mathbf{M}[j, :]$;
        $\widehat{\mathbf{M}}$ $[\mathcal{D}[(i, j)], :]$ = elementwiseMax(coverage, 0);
    end
    return $\widehat{\mathbf{M}}$;

Function Score($\mathbf{X}$, $\widehat{\mathbf{M}}$):
    coverage $\leftarrow$ zero-initialized list of length $P(P-1)$;
    featuresSelected $\leftarrow 0$;
    for $s \leftarrow 1$ to $F$ do
        if $\mathbf{X}[s] = 0$ then continue;
        coverage += $\widehat{\mathbf{M}}[:, s]$;
        featuresSelected += 1;
    end
    minCoverage $\leftarrow$ min(coverage);
    return min(minCoverage, $K$) $- \alpha \cdot$ featuresSelected;
```

Supplementary Figures

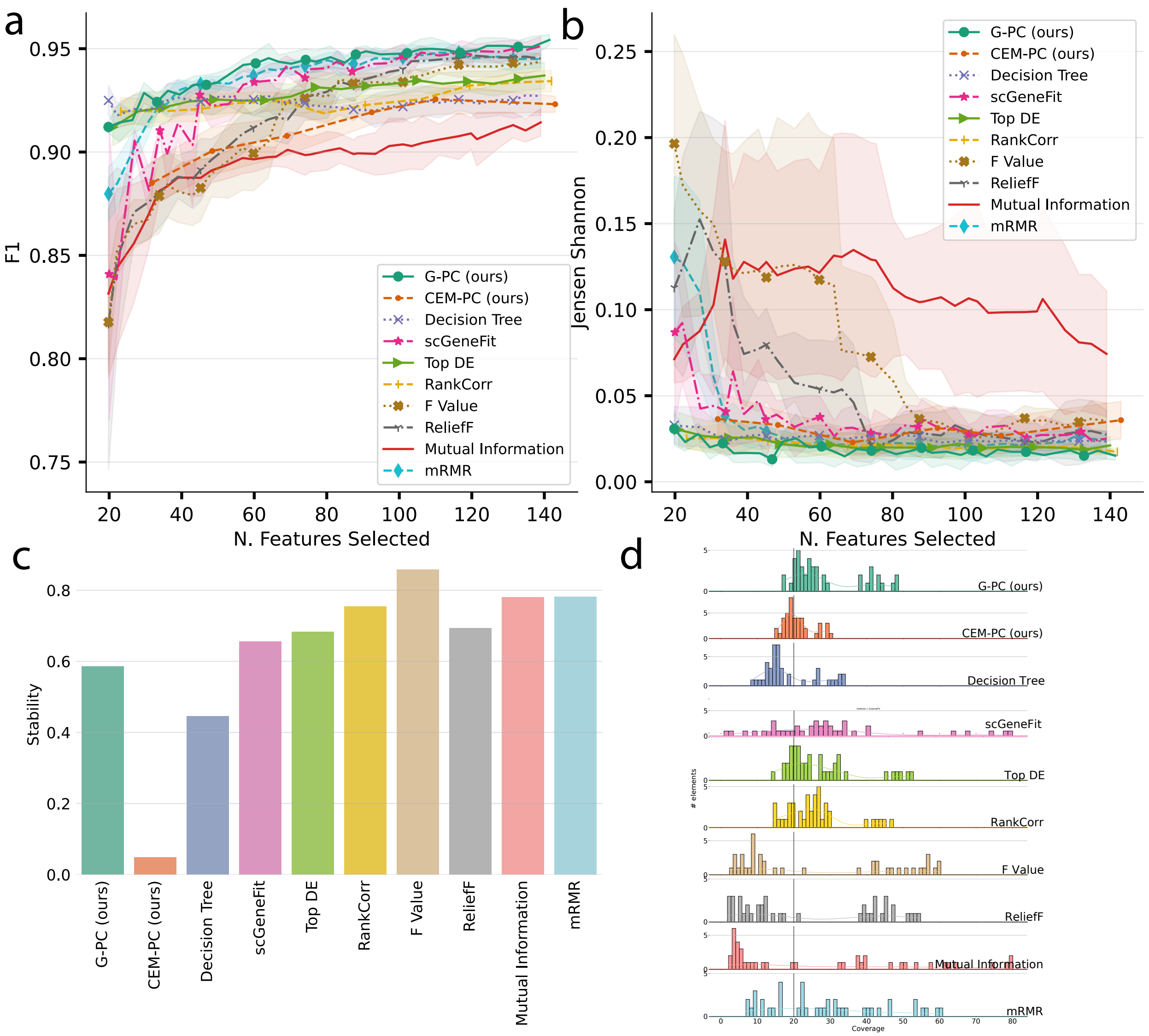


Fig. A.1: **Comparison of feature selection methods for MC.** Performance scores for (a) and (b) were averaged across five different random train and test splits. (**a**) Performance of a logistic regression model trained on the selected features. (**b**) Jensen-Shannon divergence (lower is better) between CIBERSORT-predicted mixture proportions and the ground truth. (**c**) Stability scores for all eight methods when ≈ 139 features were selected (coverage= 40) over 5 runs. (**d**) Histogram of coverage factors per element (phenotypic pair) for the test set. Number of genes selected (66) corresponds to a coverage of 20.

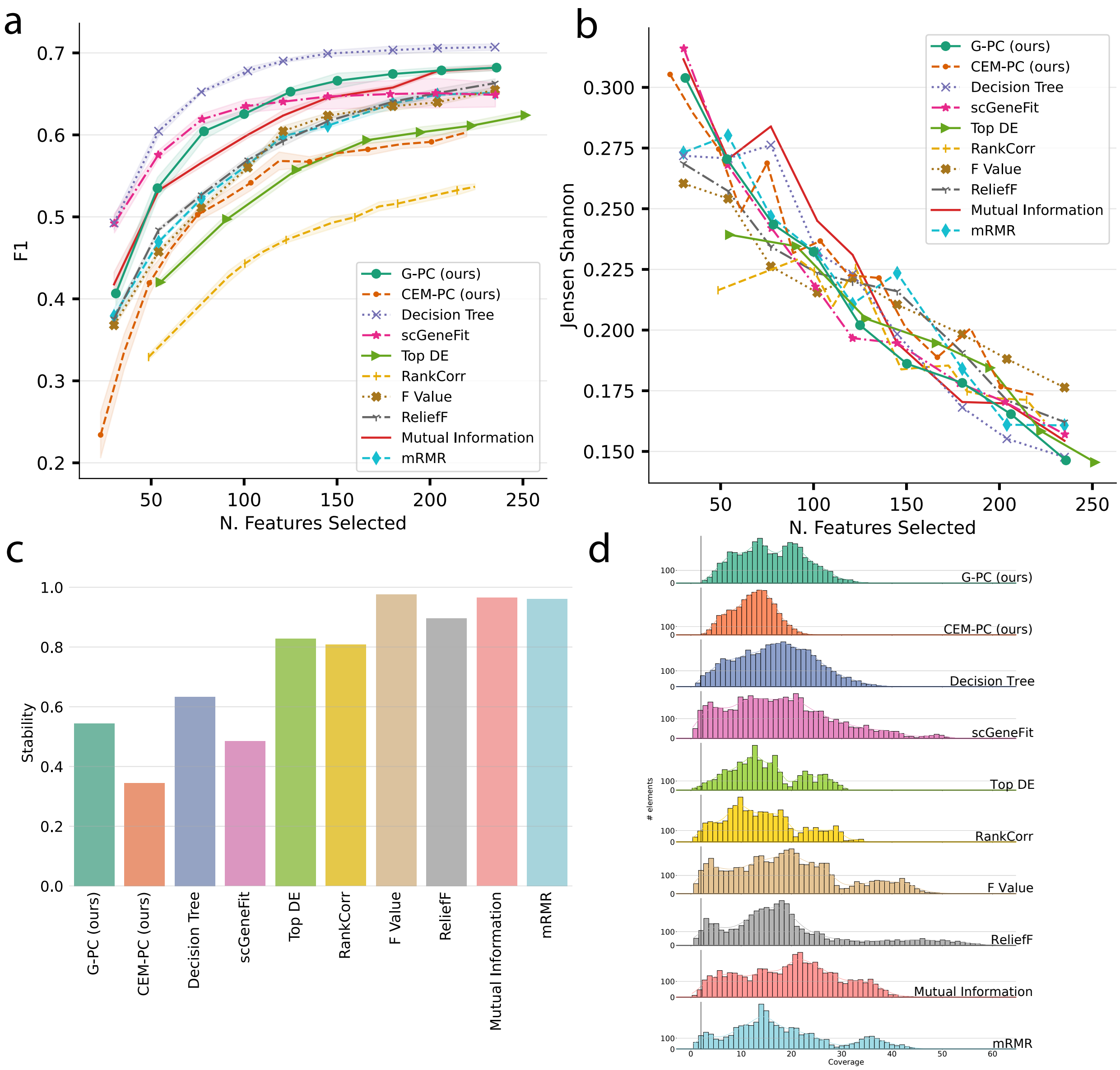


Fig. A.2: **Comparison of feature selection methods for HCA.** Performance scores for (a) and (b) were averaged across five different random train and test splits. Standard deviation is shown as a shaded region in (a) but removed from (b) for better visibility. (**a**) Performance of a logistic regression model trained on the selected features. (**b**) Jensen-Shannon divergence (lower is better) between CIBERSORT-predicted mixture proportions and the ground truth. (**c**) Stability scores for all eight methods when ≈ 121 features were selected (coverage= 5) over 5 runs. (**d**) Histogram of coverage factors per element (phenotypic pair) for the test set. Number of genes selected (53) corresponds to a coverage of 2.

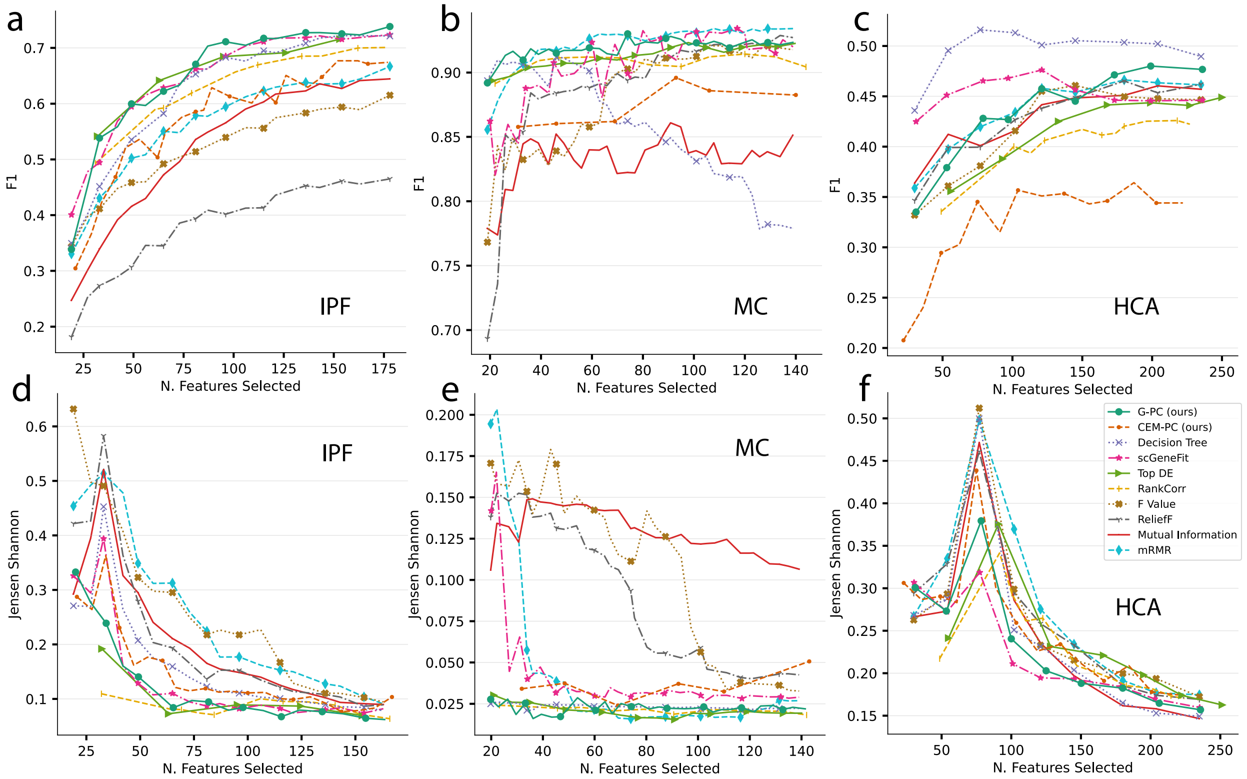


Fig. A.3: **Performance of a KNN classifier** trained on features selected by each method (a-c) and deconvolution performance based on linear least squares (d-f) for all three datasets.

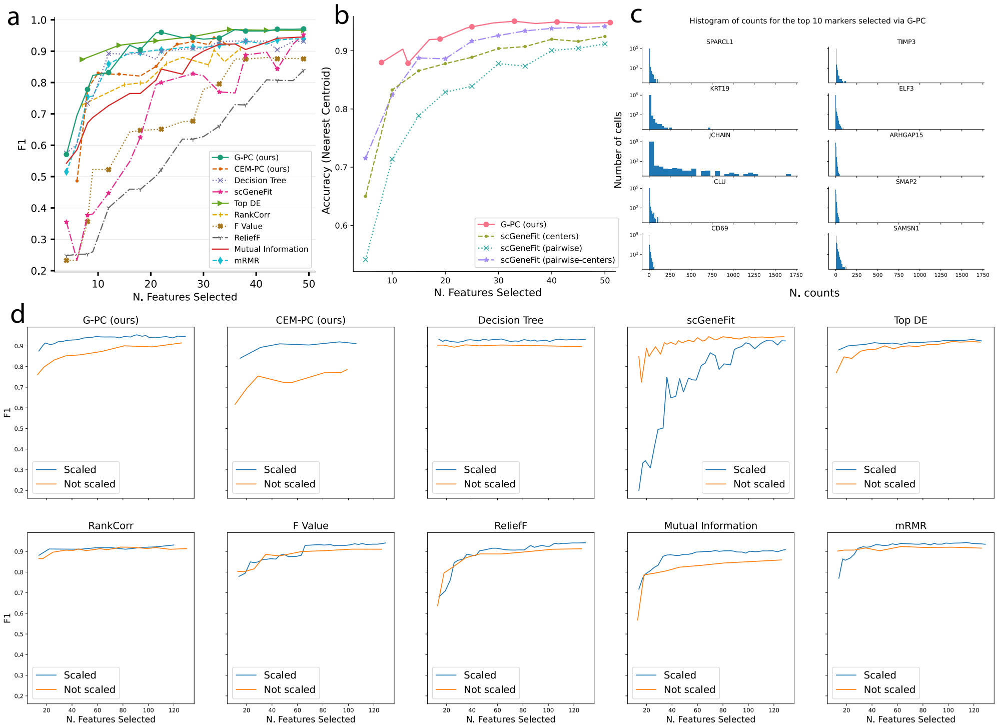


Fig. A.4: **Robustness to batch effects and preprocessing.** (**a**) Batch effects. To test the impact of batch effects, we ran feature selection on a pancreas dataset from `https://pubmed.ncbi.nlm.nih.gov/27667667/` and used the selected features to train a Logistic Regression model on a pancreas dataset from a different study `https://pubmed.ncbi.nlm.nih.gov/27693023/` (after a train/test split). We only used overlapping genes and cell types (7) between the two datasets, reducing the number of cells for the former to 1,864 and for the latter to 1,941. The chart shows that G-PC and TopDE are the most robust to batch effects. (**b**) Comparison against different scGeneFit variants for the MC dataset. All hyperparameters were taken from `https://github.com/solevillar/scGeneFit-python/blob/master/examples/scGeneFit_example.ipynb`. (**c**) IPF - Distribution of gene counts for the top 10 markers selected via G-PC ($k = 10$). We observe that our method selects a combination of genes with different expression levels at baseline, showing that it is not affected by the basal expression of genes. (**d**) MC - Performance of each method on log-transformed and scaled MC data versus non log-transformed and unscaled data.

B Endotype-Informed Biomarkers from Time Series Clinical Transcriptomics Data

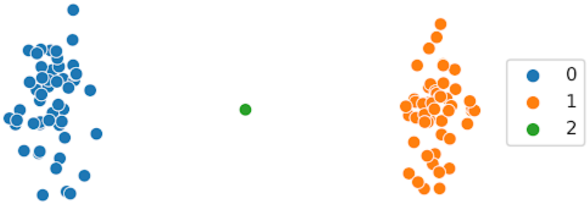


Fig. B.5: **Toy example demonstrating the need for node capacities.** Assume a situation where we have a single outlier (in green). Without capacities, we would incorrectly identify the trajectory 0-2-1 since all patients would travel through the green node. With node capacities, however, only one patient is allowed to travel through 2, and the rest travel through the more likely disease path 0-1. The possibility of such outliers impacting the results depends on the size of the dataset and the sampling rates.

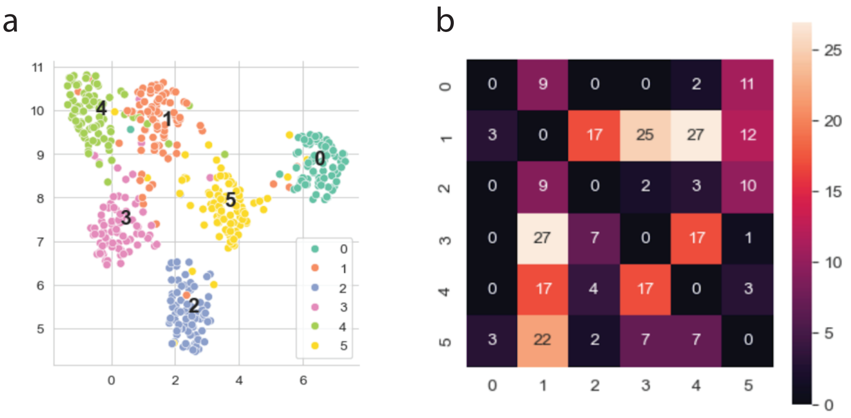


Fig. B.6: **Truffle on simulated data.** (a) We randomly generated 15-dimensional samples from 6 different states and constructed random patient trajectories from these (by randomly deleting some nodes to simulate sparse visits). We designated 0-5-1-4, 5-1-3-4, and 4-3-1-2-5 as "true trajectories". From these, Truffle accurately identified 4-3-1-2-5 as the top trajectory of length 4, and also 5-1-3-4 as the top trajectory of length 3. The other path 0-5-1-4 was one of the top three trajectories of length 3.

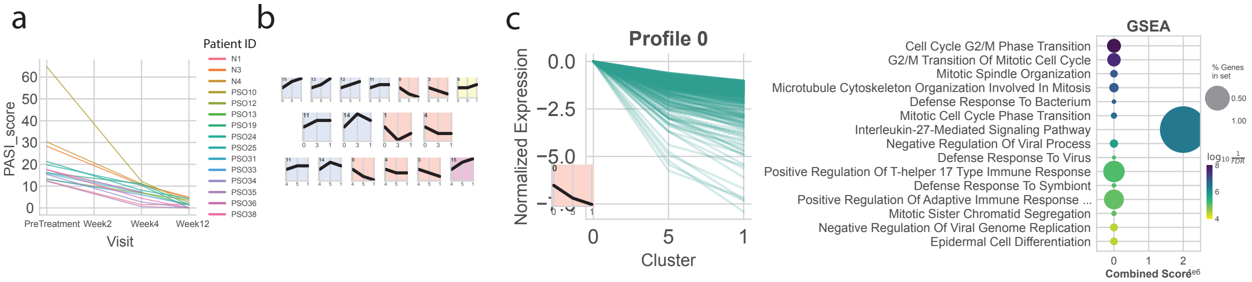


Fig. B.7: **Supplementary plots for the psoriasis dataset.** (a) PASI scores for each patient. (b) All significant STEM profiles for all three trajectories identified by Truffle. (c) Profile 0 for the trajectory $0-5-1$ shows a high score for "IL-27-Mediated Signaling Pathway." (d) Complete heatmap with all significant GO processes (FDR < 0.05).

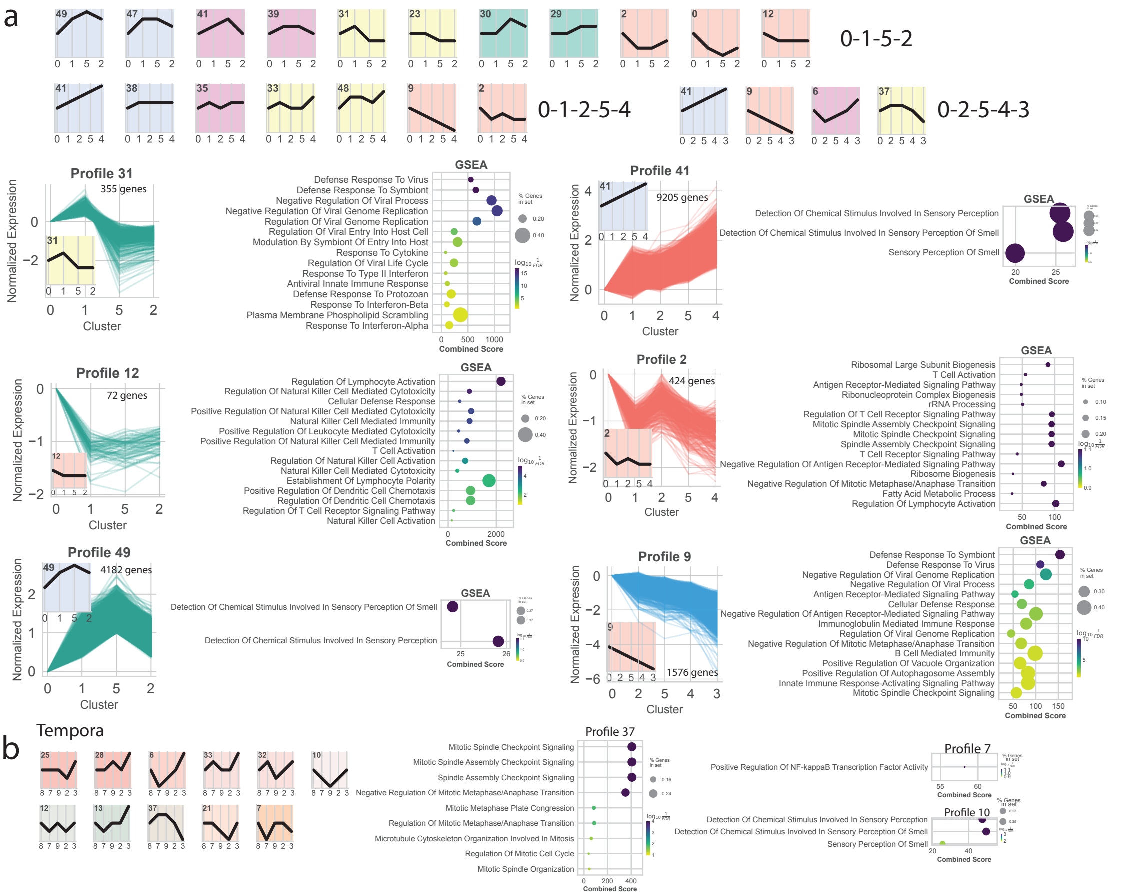


Fig. B.8: **Supplementary plots for the COVID-19 dataset.** (a) Selected STEM profiles and their GO processes. (b) STEM profiles for Tempora's trajectory $8-7-9-2-3$ along with GO processes for profiles 37, 7, and 10.

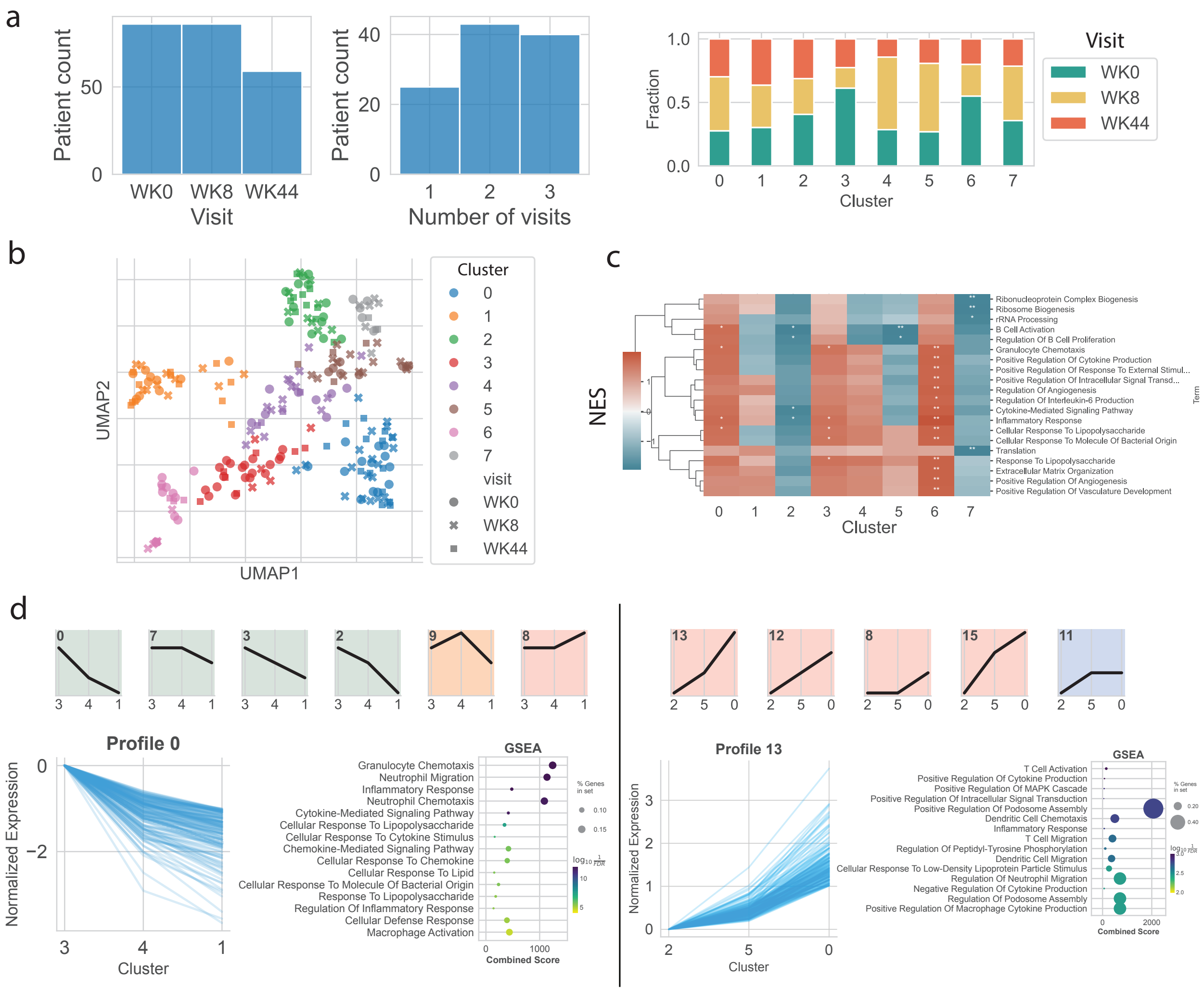


Fig. B.9: **Supplementary plots for the Crohn's disease dataset.** (a) Information on the number of visits per patient and cluster. (b) Clustering results. (c) Top GO processes for each cluster when compared to healthy samples. (d) STEM profiles for the top two paths $3-4-1$ and $2-5-0$ along with GO processes for the top profile for each.

C Recovering Time-Varying Networks From Single-Cell Data

Set Transformer Operations

We redefine the Multihead and rFF operations from Set Transformers [273] to the ones used for Marlene here.

First, we define the **Attention** operation. Let $Q \in \mathbb{R}^{k \times g}$ be the query matrix of k elements and g dimensions. The Attention operation used for MAB is

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{g}}\right) V \tag{1}$$

where the key and value matrices are $K, V \in \mathbb{R}^{c \times g}$. Next, the Multihead attention operation with h heads [310] is given by

$$\text{Multihead}(Q, K, V) = \text{concat}(O_1, \ldots, O_h) W^O \tag{2}$$

where $O_j = \text{Attention}(QW_j^Q, KW_j^K, VW_j^V)$ for weight matrices $W_j^Q, W_j^K, W_j^V \in \mathbb{R}^{g \times g/h}$ and $W^O \in \mathbb{R}^{g \times g}$ (these matrices are not to be confused with self-attention weights used consequently for Marlene). In our implementation, k is the number of seeds or output vectors used for the PMA layer. This is a hyperparameter that corresponds to the number of "statistic" vectors we expect to learn from data. Finally, rFF is a feedfoward layer such as a linear layer.

EvolveGCN Operations

Here, we introduce the GRU and topK pooling operations used in the second step of Marlene.

The topK pooling operation is needed to summarize nodes into k representative ones [275, 311]. Here k is the same as the number of seeds used for PMA. Given an input $\mathbf{G} \in \mathbb{R}^{g \times k}$ and a learnable vector q, the TopK operation performs the following steps:

$$\begin{aligned}
\rho &= \frac{\mathbf{G}q}{\|q\|} \\
i &= \text{Top-k-indices}(\rho) \\
\mathbf{Z} &= [\mathbf{G} \odot \tanh(\rho)]_i .
\end{aligned}$$

At time step t, given a pooled matrix $\mathbf{Z}_t$ and hidden state $\mathbf{W}_{t-1}$ (i.e., self-attention weights $\mathbf{W}_{t-1}^Q$ or $\mathbf{W}_{t-1}^K$), the standard GRU operation is:

$$\begin{aligned}
r_t &= \sigma(M_{ir}\mathbf{Z}_t + b_{ir} + M_{hr}\mathbf{W}_{t-1} + b_{hr}) \\
z_t &= \sigma(M_{iz}\mathbf{Z}_t + b_{iz} + M_{hz}\mathbf{W}_{t-1} + b_{hz}) \\
n_t &= \tanh(M_{in}\mathbf{Z}_t + b_{in} + r_t \odot (M_{hn}\mathbf{W}_{t-1} + b_{hn})) \\
\mathbf{W}_t &= (1 - z_t) \odot n_t + z_t \odot \mathbf{W}_{t-1}
\end{aligned}$$

where σ is the sigmoid function and $\odot$ is the Hadamard product. See also Paszke *et al.* [312].

Supplementary Figure for the HLCA dataset

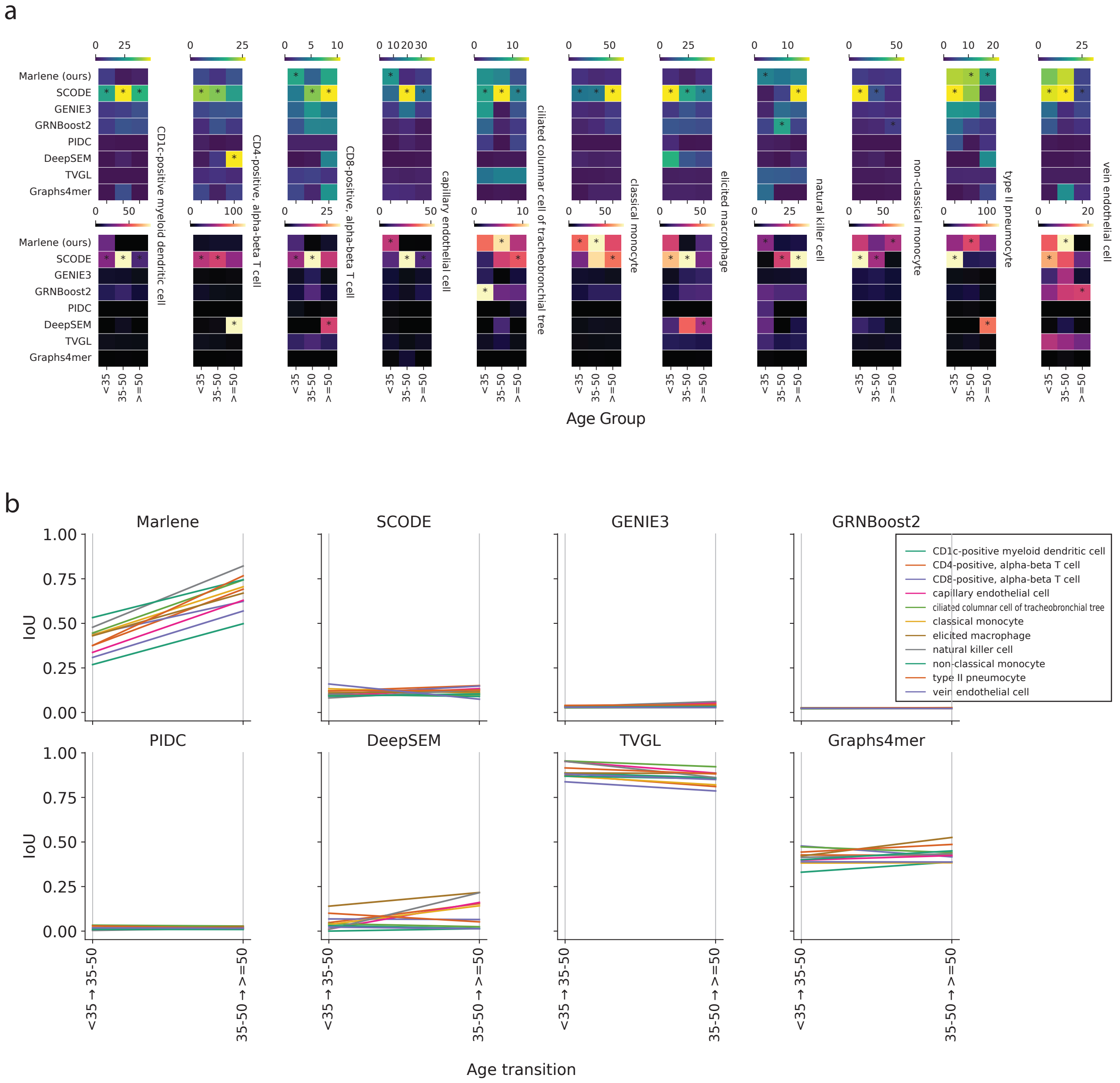


Fig. C.10: (a) FDR corrected p-values of Fisher exact tests reflecting the number of links that overlap with the two TF-gene databases. (b) IoU scores across time reflecting the overlap between consecutive graphs.

Supplementary Figure for the Mouse Fibrosis Dataset

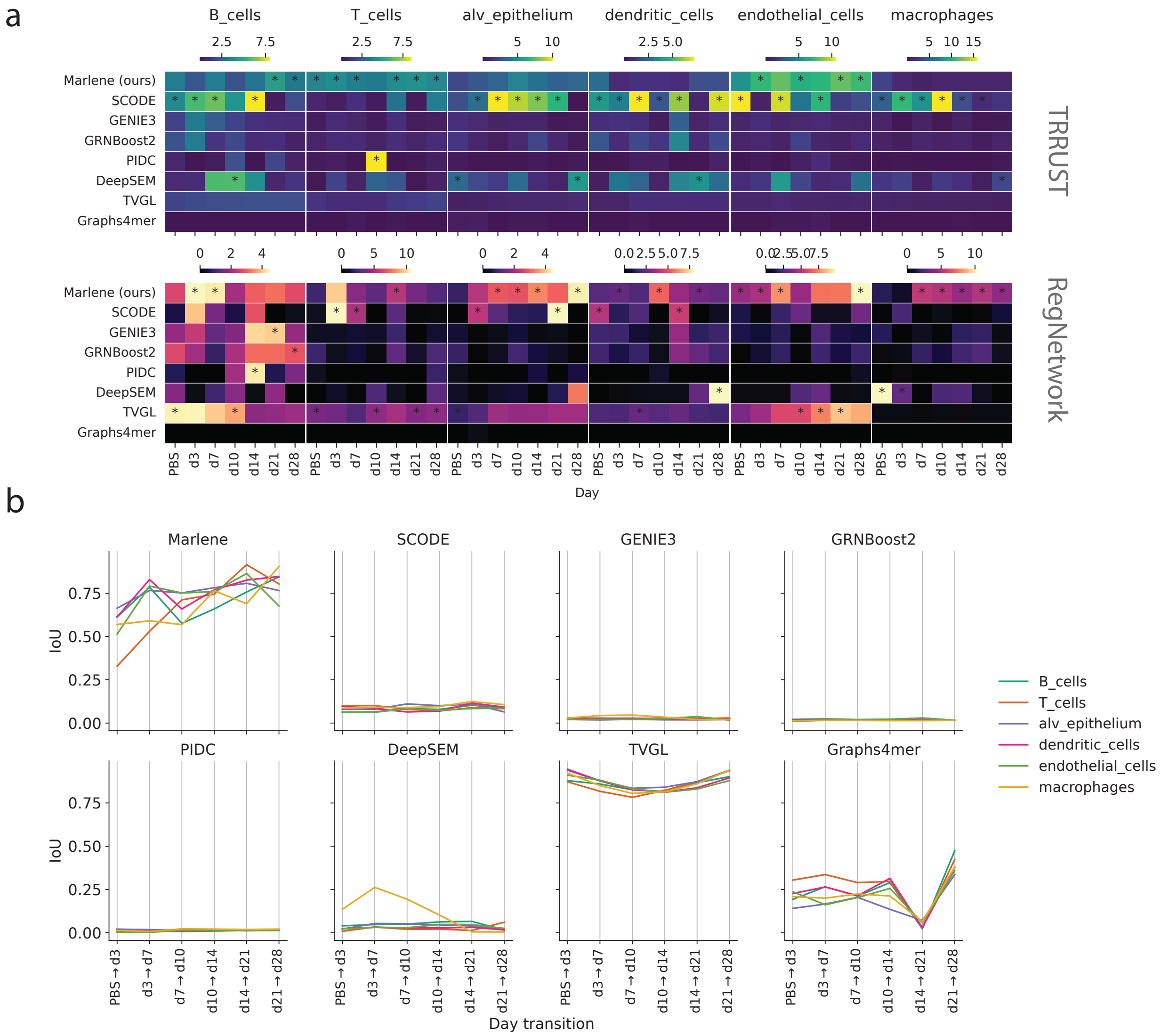


Fig. C.11: (a) FDR corrected p-values of Fisher exact tests reflecting the number of links that overlap with the two mouse databases. (b) IoU scores across time reflecting the overlap between consecutive graphs.