An Intrinsic-hoc Framework for Heterogeneous Cellular Senescence Elucidation Using Deep Graph Representation Learning and Experimental Validation

Anjun Ma[1,2,†], Hao Cheng[1,†], Natalia Del Pilar Vanegas[3,†], Ahmed Ghobashi[1], Hu Chen[1], Jhonny Rodriguez[3], Lorena Rosas[3], Cankun Wang[1,2], Jianming Shao[4], Yi Jiang[1], Xiaoying Wang[1,2], Irfan Rahman[5], Jose Lugo-Martinez[6], Dongmei Li[7], Gloria S Pryhuber[8], Ziv Bar-Joseph[6,9], Oliver Eickelberg[10], Melanie Königshoff[10], Dongjun Chung[1,2], Rui Chen[11], Toren Finkel[12], Ana Mora[3,*], Mauricio Rojas[3,*], Qin Ma[1,2,*]

[1] Department of Biomedical Informatics, College of Medicine, Ohio State University, Columbus, OH 43210, USA.
[2] Pelotonia Institute for Immuno-Oncology, The James Comprehensive Cancer Center, The Ohio State University, Columbus, OH 43210, USA.
[3] Dorothy M. Davis Heart and Lung Research Institute, Division of Pulmonary, Critical Care, and Sleep Medicine, Department of Internal Medicine, Ohio State University, Columbus, OH, USA
[4] Department of Molecular and Human Genetics, Baylor College of Medicine, 1 Baylor Plaza, Houston, Texas, 77030 USA
[5] Environmental Medicine, University of Rochester, Rochester, NY, USA
[6] Ray and Stephanie Lane Computational Biology Department, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA
[7] Department of Clinical and Translational Research, University of Rochester Medical Center, Rochester, NY, USA
[8] Department of Pediatrics, University of Rochester Medical Center, Rochester, NY, USA
[9] Machine Learning Department, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA
[10] Center of Lung Aging and Regeneration, Division of Pulmonary, Allergy, Critical Care, and Sleep Medicine, University of Pittsburgh School of Medicine, Pittsburgh, PA, USA
[11] Department of Ophthalmology and Vision Science, University of California Irvine School of Medicine, Irvine, CA, USA
[12] Aging Institute, University of Pittsburgh School of Medicine, Pittsburgh, PA, USA
[†] These authors contributed equally
[*] Corresponding Authors

ABSTRACT

Characterizing senescent cells and identifying corresponding senescence-associated genes within complex tissues is critical for our understanding of aging and age-related diseases. We present DeepSAS, an intrinsic-hoc framework for elucidating the heterogeneity encoded in senescent cells and their associated genes from single-cell RNA-seq data using deep graph representation learning. Applied to both healthy eye cell atlas and in-house idiopathic pulmonary fibrosis datasets with Xenium spatial transcriptomics validation, DeepSAS reveals robust and biologically grounded senotypes and demonstrates superior benchmarking performance compared with existing methods.

MAIN

Cellular senescence is a hallmark of aging and age-related diseases, defined by stable cell-cycle arrest and a distinct secretory phenotype[1]. Senescent cells (**SnCs**) accumulate in aging tissues and contribute to the pathogenesis of multiple diseases, such as cancer, atherosclerosis, Alzheimer's disease, and idiopathic pulmonary fibrosis (IPF)[2]. Historically, senescence has been studied at the bulk tissue level, treating SnCs as a uniform "phenotype" and overlooking the diverse cellular origins from which they arise. The advent of single-cell RNA sequencing (scRNA-seq) enabled higher resolution, but most studies continued to model SnCs as a homogeneous cell type, defined largely by elevated expression of canonical markers such as CDKN1A (p21) and CDKN2A (p16). This oversimplification fails to capture the complexity of senescence in diverse cell types, tissues, and diseases. In reality, every SnC arises from a normal cell and retains transcriptional features of its origin[3]. While these cells may share hallmark features such as G1 arrest, lysosomal activation, and SASP expression, their senescent programs are highly context-specific and shaped by cell-intrinsic transcriptional states and microenvironmental cues. This layered heterogeneity gives rise to a diversity of senotypes (i.e., senescence trigger, characteristics of cellular states, molecular signatures, and functional properties associated with cellular senescence) that are embedded within individual cell types and not adequately resolved by existing approaches[4-6] (**Fig. 1a**). Existing senescence hallmark gene sets only demonstrate part of the complexity of SnCs at single-cell resolution, as they fail to capture the diverse, context-specific gene expression programs that define senescence within each cell type[7,8].

To achieve precise and comprehensive characterization of senescence at single-cell resolution, there is an urgent need for computational frameworks that uncover SnC states and their corresponding cell-type-specific senescence-associated genes (**SnGs**) in an unbiased and scalable manner. As far as we know, the identification of senescent cells (SnCs) and senescent gene signatures (SnGs) remains challenging, primarily due to their rarity and heterogeneity, which hinder confident detection at single-cell resolution. Furthermore, many SnCs have lost normal cellular function and exhibit globally reduced transcriptional activity[9], leading to two key complications: first, they are often misclassified as low-quality or damaged cells; second, the diminished expression levels make it difficult to reliably capture cell-type-specific SnGs patterns.

While current methods provide valuable insights into SnCs identification from scRNA-seq data[4-6], most of them rely on predefined senescence hallmarks curated from bulk data or predict new SnGs for a specific tissue/disease type that are not suitable for broader applications. Many apply global scoring schemes that assume a uniform senescence phenotype, disregarding cell-type-specific transcriptional contexts[9,10]. Others adopt sequential workflows that first identify senescent cells and subsequently extract differentially expressed genes (DEGs), making it difficult to resolve rare or lowly expressed SnGs and potentially introducing bias through thresholding or marker dependence.

To address these challenges, we developed DeepSAS (Deep graph representation learning framework for Senescent cells And Senescence-associated genes), a contrastive learning framework to identify rare SnCs and their corresponding SnGs in an intrinsic-hoc and context-aware manner. DeepSAS is conceptually innovative as we hypothesize that SnCs are distinguishable from non-SnCs within the same cell type based on intrinsic transcriptional programs ($d_1<d_3$ in **Fig. 1b**), and simultaneously, their cell-type specificity can be inferred based on SASP-driven microenvironmental cues ($d_1<d_2$ in **Fig. 1b**).

DeepSAS has four unique features (**Fig. 1c** and **Online Methods**). First, it constructs a heterogeneous graph that represents cells and genes, involving both intracellular signals and intercellular cell-cell communication (CCC). This design enables DeepSAS to infer SnCs and their associated SnGs simultaneously in a unified representation, offering an unbiased and biologically grounded alternative to hallmark- or differential expression-based approaches. Second, DeepSAS employs a graph transformer architecture within a contrastive learning framework to learn cell-gene relationships adaptively. Attention weights derived from the model capture both local (within-cell-type) and global (across-cell-type) dependencies, facilitating the detection of subtle, low-expression SnC and SnG signals. Third, we define a senescence score based on the transformer-derived attention weights and the graph's topological structure. This continuous score reflects the degree of senescence within the context of each cell's intrinsic state and neighborhood environment, enabling fine-resolution characterization of distinct senotypes. Lastly, DeepSAS uses a self-supervised learning scheme to distinguish SnCs from non-SnCs without requiring external labels, greatly enhancing its robustness and generalizability across tissues, diseases, and experimental conditions.

The evaluation and benchmarking of DeepSAS demonstrated robust and consistent performance across diverse conditions. Ablation studies (**Supplementary Fig. S1**) examined the impact of three major design choices: the number of genes used to construct the heterogeneous graph, the inclusion of CCC information, and the use of different hallmark gene lists for labeling initial SnG candidates. The results showed that incorporating CCC edges and integrating multiple hallmark gene sets improved model performance, while varying the number of genes had minimal impact. Robustness tests using different subsampling manners further demonstrated that DeepSAS maintains relatively consistent outputs against phenotype influences, confirming its resilience to sample imbalance and noise (**Supplementary Fig. S2** and **Table S1**). Due to the absence of an established gold standard for single-cell senescence prediction and the lack of existing tools capable of inferring cell-type-specific SnGs (see detailed justification in the **Online Methods**), we focused our comparative evaluation on SnC detection across tools applied to the same datasets (**Supplementary Fig. S3** and **Table S2**). DeepSAS produced biologically plausible predictions of SnC across cell types (average SnC proportion is 7.04%, S^2=3.36). In contrast, other methods often returned unrealistic results, either predicting an excessively large number of SnCs in multiple cell types (e.g., average SnC proportion is 56.26%, S^2=1071.72) or generating globally low SnC proportions with limited statistical power (e.g., average SnC proportion is 0.47%, S^2=0.69). These comparisons reinforce the reliability of DeepSAS in capturing rare but meaningful senescent subpopulations across diverse cellular contexts.

We first applied DeepSAS to a comprehensive human retinal pigment epithelium (RPE) and choroid single-cell atlas, which contains 716,410 cells collected from 53 healthy human donors, spanning from infants to 93 years old (**Fig. 1d** and **Supplementary Fig. S4**). We are interested in how senescence is interpreted and relevant to aging and cell types. To facilitate data analysis at this scale, we partitioned the dataset into balanced batches, ensuring that the cell type proportions were preserved in each batch. In total, we identified 35,185 SnCs, representing 4.91% of total cells (**Fig. 1e**). We found that fibroblasts, retinal pigment epithelium (RPE) cells, and endothelium cells are the top three cell types having the highest proportion of SnCs identified, and SnCs identified in each batch showed a relatively low divergence in terms of SnC proportions (**Fig. 1f** and **Supplementary Fig. S5**). It is noteworthy that we observed a significantly higher abundance of SnCs in adult donors compared to pediatric donors (**Fig. 1g** and

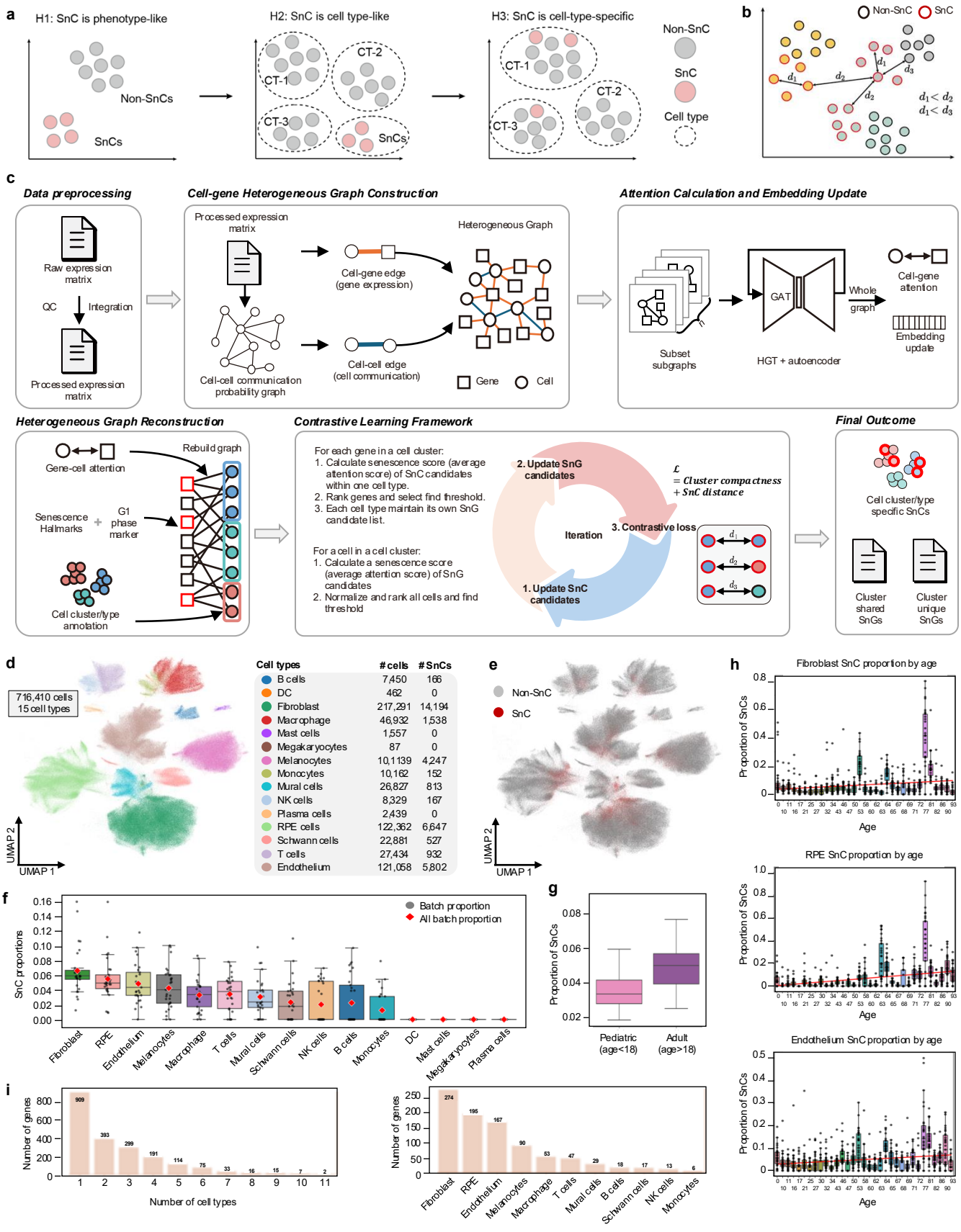
Supplementary Fig. S6). Furthermore, within fibroblast, RPE, and endothelial populations, we detected a positive correlation between SnC proportions and donor age, supporting the biological relevance of DeepSAS predictions (**Fig. 1h**). These findings align with the hypothesis that aging increases cellular stress, leading to progressive accumulation of SnCs, particularly within these three cell populations. In parallel, DeepSAS also identified cell-type-specific SnGs (**Fig. 1i** and **Supplementary Table S3-S4**), which may facilitate future mechanistic studies of senescence in aging-related retinal diseases, such as age-related macular degeneration, cataracts, and glaucoma.

We next applied DeepSAS to our newly generated in-house 10X Flex single-cell dataset from six IPF patients, encompassing (**Fig. 2a** and **Supplementary Table S5**). DeepSAS identified 1,694 senescent cells (7.02%) (**Fig. 2b**), revealing striking enrichment in IPF versus healthy samples, in lower versus upper lobes (UL vs. LL), and in old versus young samples (**Fig. 2c** and **Supplementary Table S6**). These findings align with our biological expectations that SnCs are more enriched in the lower lobes compared to the upper lobes, and are more prevalent in donors with advanced age and disease status[11,12]. Furthermore, we recognized 263 SnGs across 25 cell types (**Supplementary Table S7**), demonstrating the to-be-elucidated senescent heterogeneity in our IPF tissue. Among all the SnGs, 43 are uniquely identified in only one cell type (e.g., AT1, AT2, Adventitial fibroblast, Alveolar fibroblast, Fibrotic fibroblast, and Proliferating fibroblast), and the other SnGs are shared by at least two cell types (**Fig. 2d** and **Supplementary Table 8**). Functional enrichment of identified SnGs revealed strong enrichment in canonical senescence pathways (e.g., cellular senescence, p53 signaling, NF-κB signaling) (**Supplementary Fig. S7**). In addition, the enrichment of cell-type-specific programs, such as extracellular matrix remodeling in fibroblasts, vascular stress response in endothelial cells, and epithelial-associated cancer pathways, further supports that DeepSAS captures context-aware senotypes embedded within distinct cellular compartments. By overlapping predicted SnCs with known hallmark lists[13-20], we also labeled SnGs that have been previously recognized as hallmarks, and new SnGs have never been reported before (**Supplementary Table S9**). In addition, by comparing SnGs with phenotype-associated DEGs, we identified subsets of SnGs that are specifically associated with aging, uniquely linked to IPF, shared between aging and IPF, or exclusively present in disease-specific contexts independent of aging (**Supplementary Table S10**).

To validate the SnGs predicted by DeepSAS, we leveraged high-resolution spatial transcriptomics data generated using the Xenium platform from one normal parenchyma sample and paired UL and LL samples from two IPF patients (**Supplementary Fig. S8** and **Table S11**). We manually annotated two kinds of histologically distinct regions: fibrotic foci and the surrounding fibroblast-rich regions. Alveolar fibroblasts residing in the surrounding tissue are known to differentiate into *CTHRC1*$^{+}$ fibroblasts localized within fibrotic foci[21]. Therefore, we used the surrounding regions and fibrotic foci as spatial surrogates for alveolar fibroblasts and *CTHRC1*$^{+}$ fibroblasts, respectively. For Alveolar fibroblasts, we picked *CCL4*, which was identified as a shared SnG while also exhibiting strong expression signals (**Fig. 2e**). We compared *CCL4* transcript expressions across 19 random regions in normal parenchyma, 126 fibrotic foci (61 in UL and 65 in LL samples), 101 surrounding regions (50 in UL and 51 in LL samples). We found that the number of *CCL4*-expressing cells was substantially higher in the surrounding regions than in the fibrotic foci and notably elevated compared to the normal parenchyma (**Fig. 2f**). Furthermore, pathological *CTHRC1*+ fibroblasts were exclusively observed in both the UL and LL regions of IPF samples and were absent in healthy lung tissues. This finding aligns with the established role of *CTHRC1*$^{+}$ fibroblasts in promoting extracellular matrix

remodeling and fibrogenesis, further validating their relevance in IPF pathogenesis. To further confirm the presence of senescent cells, we examined the co-expression of *IL6*, a canonical senescence marker, with *CCL4*. This analysis also revealed elevated co-expression in the surrounding regions relative to fibrotic foci (**Supplementary Fig. S9**). In addition, we evaluated another SnG predicted by DeepSAS, *NFE2L2*, a transcription factor coding gene associated with senescent *CTHRC1*$^{+}$ fibroblasts. Co-expression analysis of *NFE2L2*, *CTHRC1*, and the canonical senescence marker *CDKN1A* showed a significant increase in triple-positive cells within fibrotic foci compared to surrounding regions in both UL and LL (**Fig. 2g-h**). These spatial findings support the predictive power of DeepSAS in identifying SnGs across fibroblast subtypes and highlight the spatial localization of distinct senescent states within fibrotic lung tissue.

In summary, DeepSAS is a generally applicable method for the identification of SnCs and SnGs from scRNA-seq data in diverse tissues and aging-related diseases. Nonetheless, several limitations remain. First, DeepSAS currently characterizes senescence in a binary fashion, SnCs or non-SnCs, but does not yet model intermediate or progressive senescence stages. The combination of dynamic graphs and diffusion models may capture distinct biological functions or temporal dynamics. Second, the model itself cannot deliver an end-to-end prediction of phenotypically associated SnGs. Appropriate hypergraph representations and causal inference methods can contribute to inferring the causal effect of individual SnGs in response to different phenotypic influences.



Cell types	# cells	# SnCs
B cells	7,450	166
DC	462	0
Fibroblast	217,291	14,194
Macrophage	46,932	1,538
Mast cells	1,557	0
Megakaryocytes	87	0
Melanocytes	10,1139	4,247
Monocytes	10,162	152
Mural cells	26,827	813
NK cells	8,329	167
Plasma cells	2,439	0
RPE cells	122,362	6,647
Schwann cells	22,881	527
T cells	27,434	932
Endothelium	121,058	5,802

Figure 1. **Overview workflow of DeepSAS and application in human choroid samples.**

(a) Conceptual models of senescent cell identity: Hypothesis 1 (H1) treats SnCs as a uniform phenotype; Hypothesis 2 (H2) models SnCs as a distinct cell type; Hypothesis 1 (H3; our hypothesis) considers SnCs to be cell-type-specific subpopulations requiring context-aware identification.

(b) DeepSAS assumes SnCs within the same cell type are more transcriptionally similar to each other than to SnCs from other lineages, enabling within-type separation of SnCs and non-SnCs.

(c) Workflow of DeepSAS: (1) Raw expression matrices are preprocessed. (2) A heterogeneous graph is built with cell–cell (intercellular signaling) and gene–cell (expression) edges. (3) A heterogeneous graph transformer is trained using a contrastive learning objective to jointly embed cells and genes. (4) Graph attention and topology are used to calculate senescence scores and update graph structure. (5) SnCs and SnGs are iteratively updated based on senescence scores and contrastive loss. (6) The final output includes cell-type-specific SnC and SnG assignments.

(d) UMAP visualization of the human RPE and choroid cell atlas including 15 predefined cell types.

(e) UMAP of predicted SnCs across all cell types in the human RPE and choroid snRNA-seq.

(f) SnC proportions across 15 cell types, colored by age group, showing expected enrichment in aged samples. Each dot represents a batch, and the red diamond indicates the proportion of total SnCs in each cell type, combining all batches.

(g) Proportion of SnCs identified in each batch between pediatric and adult donors, respectively.

(h) Boxplots comparing DeepSAS predicted SnC proportions across age groups in fibroblast, RPE, and endothelium cell populations.

(i) Bar plot showing the number of SnGs identified in all batches that shared in multiple cell types (left) and the number of SnGs identified only in one cell type (right).

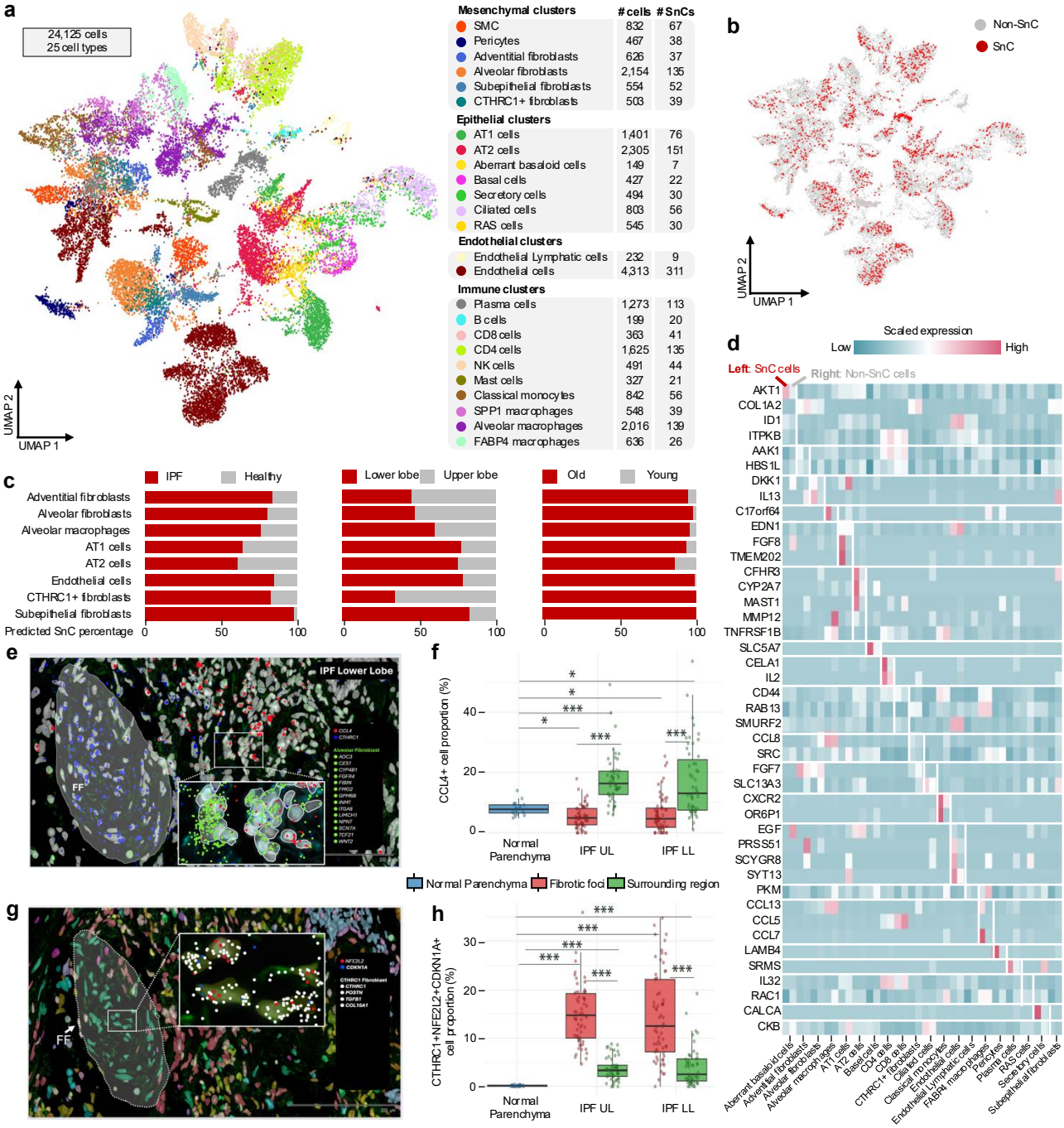

a
24,125 cells
25 cell types
UMAP 2
UMAP 1
Mesenchymal clusters # cells # SnCs
SMC 832 67
Pericytes 467 38
Adventitial fibroblasts 626 37
Alveolar fibroblasts 2,154 135
Subepithelial fibroblasts 554 52
CTHRC1+ fibroblasts 503 39
Epithelial clusters
AT1 cells 1,401 76
AT2 cells 2,305 151
Aberrant basaloid cells 149 7
Basal cells 427 22
Secretory cells 494 30
Ciliated cells 803 56
RAS cells 545 30
Endothelial clusters
Endothelial Lymphatic cells 232 9
Endothelial cells 4,313 311
Immune clusters
Plasma cells 1,273 113
B cells 199 20
CD8 cells 363 41
CD4 cells 1,625 135
NK cells 491 44
Mast cells 327 21
Classical monocytes 842 56
SPP1 macrophages 548 39
Alveolar macrophages 2,016 139
FABP4 macrophages 636 26
b
Non-SnC
SnC
UMAP 2
UMAP 1
c
IPF
Healthy
Lower lobe
Upper lobe
Old
Young
Adventitial fibroblasts
Alveolar fibroblasts
Alveolar macrophages
AT1 cells
AT2 cells
Endothelial cells
CTHRC1+ fibroblasts
Subepithelial fibroblasts
Predicted SnC percentage
0
50
100
d
Scaled expression
Low
High
Left: SnC cells
Right: Non-SnC cells
AKT1
COL1A2
ID1
ITPKB
AAK1
HBS1L
DKK1
IL13
C17orf64
EDN1
FGF8
TMEM202
CFHR3
CYP2A7
MAST1
MMP12
TNFRSF1B
SLC5A7
CELA1
IL2
CD44
RAB13
SMURF2
CCL8
SRC
FGF7
SLC13A3
CXCR2
OR6P1
EGF
PRSS51
SCYGR8
SYT13
PKM
CCL13
CCL5
CCL7
LAMB4
SRMS
IL32
RAC1
CALCA
CKB
Aberrant basaloid cells
Adventitial fibroblasts
Alveolar fibroblasts
Alveolar macrophages
AT1 cells
AT2 cells
Basal cells
CD4 cells
CD8 cells
CTHRC1+ fibroblasts
Ciliated cells
Classical monocytes
Endothelial cells
Endothelial Lymphatic cells
FABP4 macrophages
Pericytes
Plasma cells
RAS cells
Secretory cells
Subepithelial fibroblasts
e
IPF Lower Lobe
FF
f
CCL4+ cell proportion (%)
Normal Parenchyma
IPF UL
IPF LL
Normal Parenchyma
Fibrotic foci
Surrounding region
g
FF
h
CTHRC1+NFE2L2+CDKN1A+ cell proportion (%)
Normal Parenchyma
IPF UL
IPF LL

Figure 2. **DeepSAS identifies cell-type-specific SnCs and SnGs in IPF lungs, validated by spatial transcriptomics.**

(a) UMAP visualization of 24,125 cells from human IPF and healthy lungs, annotated into 25 cell types. The table summarizes the number and proportion of predicted SnCs per cell type identified by DeepSAS.

(b) UMAP highlighting all DeepSAS-predicted SnCs across the dataset, showing their distribution across multiple cell types.

(c) Bar plots showing enrichment of SnCs in IPF versus healthy lungs (left), in lower versus upper lobes (middle), and in older versus younger individuals (right), stratified by cell type.

(d) Heatmap of top senescence-associated genes (SnGs) identified by DeepSAS, showing scaled expression across SnCs (left) and non-SnCs (right) in each cell type.

(e) Spatial transcriptomics (Xenium platform) of the lower lobe of an IPF lung highlighting the expression of *CCL4* in alveolar and subepithelial fibroblasts within the fibroblast-rich surrounding regions.

(f) Boxplot quantification showing significantly higher proportions of CCL4-expressing cells in the surrounding regions compared to fibrotic foci, across anatomical lobes.

(g) Spatial co-expression of *NFE2L2*, *CTHRC1*, and *CDKN1A* in the fibrotic region of an IPF lung lower lobe, indicating the presence of senescent *CTHRC1*$^{+}$ fibroblasts.

(h) Boxplot showing significantly increased abundance of triple-positive (*NFE2L2*$^{+}$, *CTHRC1*$^{+}$, *CDKN1A*$^{+}$) cells in fibrotic foci relative to surrounding regions in both lobes. Statistical significance was assessed using the Kruskal–Wallis test. $p < 0.05$ (*), $p < 0.01$ (**), and $p < 0.001$ (***)

METHODS

1. The DeepSAS framework

1.1. Data preprocessing

A raw expression matrix was preprocessed through necessary quality control processes, including cell filtering, gene filtering, doublets detection and removal, batch removal, normalization, and scaling. The processed data was integrated for joint clustering and cell type annotation. High-quality clustering and accurate annotation results can enhance the SnC and SnG prediction in DeepSAS. In the absence of cell type labels, DeepSAS employed the standard Louvain clustering algorithm[22] to categorize cells into distinct clusters.

1.2. Heterogeneous graph construction

The preprocessed expression matrix, defined as $X^R \in \mathbb{R}^{N\times M}$, where N denotes the number of cells and M denotes the number of genes. We constructed an undirected cell-gene heterogeneous graph $G = (V, E)$ from X^R, incorporating the node set $V = V^C \cup V^G$ where $V^C = \{v_i^C \mid i = 1,2,\dots,N\}$ represent cells, and $V^G = \{v_j^G \mid j = 1,2,\dots,M\}$ represent genes. By default, DeepSAS utilizes all genes in the following steps to ensure that potentially lowly expressed SnGs can be captured. To improve computational efficiency, we also offer the option to select only top highly variable genes. The edge set E comprises cell-gene edge $\{(v_i^C, v_j^G) \mid i = 1,2,\dots,N, j = 1,2,\dots,M\}$ and cell-cell edge $\{(v_i^C, v_k^C) \mid i = 1,2,\dots,N, k = 1,2,\dots,N\}$. The cell-gene edge (v_i^C, v_j^G) represents the gene expression of v_j^G in v_i^C. For $X_{ij}^R > 0$, $w(v_i^G, v_j^C) = 1$, otherwise, $w(v_i^G, v_j^C) = 0$. The cell-cell edge (v_i^C, v_k^C) represents the CCC between v_i^C and v_k^C, in terms of SASP-related ligand-receptor interaction probability predicted by SoptSC[23]. Differing from other existing CCC prediction tools, SoptSC calculates the probability of CCC between two cells, rather than two cell types. Suppose we have a ligand-receptor pair p_k, with expression distribution given by two vectors $L^k \in \mathbb{R}^N$ and $R^k \in \mathbb{R}^N$ for ligand and receptor, respectively, the probability that a signal is sent from cell i to cell j by this ligand-receptor pair p_k is then given by:

$$P_{ij}^k = exp\left(-\frac{1}{L_i^k R_j^k}\right).$$

We then constructed the CCC probability matrix $P^k \in \mathbb{R}^{N\times N}$ regarding the ligand-receptor pair p_k. The overall probability of signaling summed over all ligand-receptor pairs, $P_1 \in R^{N\times N}$, is given by

$$P_1 = \frac{\sum_{k=1}^{K} P^k}{K},$$

where K is the number of all ligand-receptor pairs. P_1 serves as the adjacency matrix of the CCC graph. We symmetrized this matrix by considering bidirectional interactions:

$$P_1' = \frac{P_1 + P_1^T}{2}.$$

Thus, the edge weight between two cells in the heterogeneous graph is defined based on a threshold ϕ for signaling strength:

$$\begin{cases} w(v_i^C, v_j^C) = w(v_j^C, v_i^C) = 1, \text{if } P'_{1,ij} \geq \phi \quad , \\ w(v_i^C, v_j^C) = w(v_j^C, v_i^C) = 0, \text{if } P'_{1,ij} < \phi \quad . \end{cases}$$

Lastly, we computed initial embeddings for genes and cells via UMAP dimensionality reduction. The UMAP algorithm processes the cell-gene expression matrix X^R,

generating a low-dimensional embedding matrix $Z_c \in \mathbb{R}^{N\times D}$ for each cell, where D is the dimension of the embedding vector. The expression matrix is then transposed, and UMAP is applied to generate a low-dimensional embedding matrix for genes $Z_g \in \mathbb{R}^{M\times D}$. Thus, each cell and gene obtain the same initial embedding of D dimensions. The final embedding matrix $\hat{Z} \in \mathbb{R}^{(M+N)\times D}$ represents the embedding for all genes and cells.

1.3. Graph autoencoder

We adopted a graph autoencoder to learn graph embeddings for all the nodes and edges in G defined in the above section. The graph autoencoder aims to reconstruct the adjacency matrix A of the heterogeneous graph G. It consists of an encoder layer and a decoder layer, similar to a standard autoencoder. The encoder layer of the graph autoencoder utilizes a graph attention network (GAT), while the decoder layer is based on the inner product of the latent embeddings. The inputs to the encoder layer are the embedding matrices Z_c for cell nodes and Z_g for gene nodes, along with the adjacency matrix A. Due to the heterogeneity of nodes, different types of nodes reside in distinct feature spaces. Therefore, for each type of node (e.g., node with type ϕ_i), we designed a type-specific transformation matrix M_{ϕ_i} to project the features of different node types into a common feature space:

$$Z_c' = M_{\phi_1} Z_c,$$
$$Z_g' = M_{\phi_2} Z_g.$$

The encoder layer comprises L GAT layers. The embedding of node i in the l-th layer is $\mathbf{z}_i^l \in \mathbb{R}^D$, and the updated embedding of node i in the $(l+1)$-th layer is:

$$\mathbf{z}_i^{l+1} = \sum_{j\in\mathcal{N}(i)\cup i} \alpha_{i,j}^l \, W_t \mathbf{z}_j^l,$$

where the **attention coefficients** $\alpha_{i,j}^l$ are computed as

$$\alpha_{i,j}^l = \frac{\exp\left(\mathrm{LeakyReLU}\left(\mathrm{a}_s^\top \mathrm{W}_s \mathrm{z}_i^l + \mathrm{a}_t^\top \mathrm{W}_t \mathrm{z}_j^l\right)\right)}{\sum_{k\in\mathcal{N}(i)\cup i} \exp\left(\mathrm{LeakyReLU}\left(\mathrm{a}_s^\top \mathrm{W}_s \mathrm{z}_i^l + \mathrm{a}_t^\top \mathrm{W}_t \mathrm{z}_k^l\right)\right)},$$

where W_s and W_t are the weight matrices for source node and target node, respectively. When a cell is treated as a target node, its connected genes and neighboring cells serve as source nodes, similarly to how a gene acts when selected as the target. a is the weight vector of attention, $\mathcal{N}(i)$ is the neighbor nodes set of node i, $\alpha_{i,j}^l$ is the attention coefficient between node i and j. The outputs of the encoder layer are the embedding matrix of cells $\mathrm{Z}_c \in \mathbb{R}^{N\times D_1}$ and genes $\mathrm{Z}_g \in \mathbb{R}^{M\times D_1}$. Then the reconstructed adjacency matrix $\widehat{A'}$ is calculated by:

$$\widehat{A'} = \sigma\left(\mathbf{Z}_c' \mathbf{Z}_g'^\top\right).$$

σ is sigmoid function. The loss function of the graph autoencoder is defined as the binary cross entropy between the estimated and the ground truth edges:

$$\text{loss} = \sum_{ij} \mathrm{BCE}\left(\widehat{A'}_{ij}, A'_{ij}\right).$$

1.4. Reconstruct the heterogeneous graph

To support biologically guided contrastive learning, we reconstruct a weighted heterogeneous bipartite graph $G' = (V, E')$, in which edges represent connections derived from the graph transformer in Section 1.3. We define the edge set $E' \subseteq V^C \times V^G$, where each edge (v_i^C, v_j^G) is assigned a weight $w_{ij} = \alpha_{i,j}^l$, with $\alpha_{i,j}^l$ the attention score from the

final heterogeneous graph transformer layer l. Each gene node is assigned to a binary label as follows:

$$y_j^G = \begin{cases} 1, & \text{if } y_j \in \mathcal{H}, \\ 0, & \text{otherwise,} \end{cases}$$

where $\mathcal{H}$ is the union of senescence hallmark genes (e.g., SenMayo[17], Fridman[18], CellAge[19]) and cell-cycle markers. Each cell node is assigned a discrete label y_i^C based on the cell type or cell cluster labels defined in 1.1 Data preprocessing. Lastly, we retain the encoded latent embeddings of all nodes from the encoder output of the graph autoencoder.

1.5. Contrastive learning iteration

This framework iteratively identifies candidate cell-type-specific SnGs and SnCs. During the iteration process, we let the encoder f_θ learn how to distinguish between the SnCs and non-SnCs by maximizing the distance between them. Initially, the known senescent marker lists determine the senescent gene nodes. We defined a senescence score attention score α_i of cell i based on the average attention scores of SnG candidates:

$$\alpha_i = \frac{\sum_{j \in S} \alpha_{i,j}}{|S|},$$

where S is the set of SnG candidates, $|S|$ is the total number of SnG candidates. We select the top-D_1 cells with the highest senescence score as SnCs in the first iteration. D_1 is defined by the upper bound of the senescence score distribution, $Q_3 + 1.5 \times \text{IQR}$, where Q_3 is the third quartile, and IQR is the interquartile range. We retain cell types with more than 10 senescent cells.

We next calculated senescence scores for genes. Similar to the selection of SnC candidates, we computed, for each gene, the average attention scores across SnC candidates within the same cell type. The top-D_2 genes with the highest senescence scores were selected, where D_2 is defined as the upper bound of the senescence score distribution for SnG candidates. That is to say, starting from the second iteration, each cell type maintains its own SnG candidate list based on these senescence scores.

We defined three different distances d_1, d_2, d_3 with biological significance to calculate the distance between the cell nodes. d_1 is the Euclidean distance between a SnC and the center of the SnC hub in the same cell type. d_2 is the distance between two SnCs in different cell types. d_3 is the distance between a SnC and a non-SnC in the same cell type. We hypothesize that the relationship should satisfy the following conditions:

$$d_1 < d_2,\ d_1 < d_3.$$

To enforce these constraints, the loss function $\mathcal{L}$ is defined as follows:

$$\mathcal{L} = \sum_{i=1}^{n} -\log \frac{\exp(v_i \cdot c_i/\phi_s')}{\sum_{j=0}^{K} \exp\left(v_i \cdot c_j/\phi_j'\right)} + \frac{1}{nK} \sum_{i=0}^{K} \sum_{j=1}^{n} \sum_{s \in \mathcal{S}} g\left(\bar{d}(c_i, v_j), s\right) \cdot \left|\bar{d}(c_i, v_j) - s\right|,$$

where v_i is the embedding for cell node i, c_i is the embedding of the center of cluster where node i belongs to. K is the number of cell clusters. n is the total number of cells. The first term is the ProtoNCE loss function[24]. ProtoNCE could estimate the concentration for the feature distribution around each cell cluster. The distribution of embeddings for each cell cluster has different level of concentration. We use ϕ' to denote the concentration level of the feature distribution:

$$\phi' = \frac{\sum_{z=1}^{Z} \|v_z' - c\|_2}{Z \log(Z + \alpha)},$$

where $\{v_i\}_{i=1}^{Z}$ are momentum features and c is the hub center of SnCs in one cell type. α is a smooth parameter to ensure that small clusters do not have an overly large ϕ'. A smaller ϕ' indicates larger concentration. The desired ϕ' should be small (i.e., high concentration) if (*i*) the average distance between v_z' and c is small, and (*ii*) the cluster contains more feature points (i.e. Z is large). ϕ' is then normalized for all clusters. The second term is the loss function is the Multi-level Distance Regularization (MDR)[25]. The distance is defined as normalized Euclidean distance between two given embedding vectors:

$$\bar{d}(c_i, v_j) = \frac{d(c_i, v_j) - \mu}{\sigma}.$$

We define a set of levels $s \in \mathcal{S}$, and the levels are initialized with predefined values; each level s is interpreted as a multiplier of the standard deviation of the normalized distance. $g(d; s)$ is an assignment function that outputs whether the given distance $\bar{d}$ and the given level s are the closest or not, and is defined as:

$$g(d, s) = \begin{cases} 1, & \text{if type } (c_i, v_j) = \text{type}(s), \\ 0, & \text{otherwise.} \end{cases}$$

By adopting the assignment function, MDR selects valid regularization levels for each distance with the consideration of various degrees of similarities. After optimizing with the loss function $\mathcal{L}$, the cell embeddings are updated. The iteration converges when the overlap of candidate SnCs and SnGs between adjacent iterations reaches 1. The output of this step is a SnC label (0 represents for non-SnC and 1 represents for SnC) for each cell, an attention-based senescence score for each cell, and an unfiltered SnG list in each cell type.

1.6. Cell type shared and specific SnG determination

To determine SnGs, we applied a three-level selection strategy. The first level consists of the direct output of the contrastive learning framework. For predicted SnCs in each cell type, the adjacent genes with the highest senescence scores were selected. To enhance experimental interpretability, we next assessed expression differences within each cell type by grouping cells into SnCs and non-SnCs and performing differential expression analysis. SnG in each cell type identified in the last step were then intersected with the corresponding DEGs (log fold change ≥ 0.25) within the same cell type to ensure transcriptional distinction. Finally, to capture cell-type-specific signals, we retained only those genes from the second set that appeared exclusively in a single cell type. These were designated as cell-type-specific SnGs. Genes that appeared in more than one cell type after DEG filtering were referred to as cell-type-shared SnGs. This hierarchical strategy ensured that final SnGs are not only biologically relevant and discriminative between SnCs and non-SnCs, but also practical candidates for downstream validation within specific cellular contexts.

Using the same strategy, we also evaluated the relationship between SnGs and sample phenotypes. Specifically, within each cell type, we stratified cells by either age group (young vs. old) or disease status (healthy vs. IPF) and performed differential expression analysis. The resulting DEGs were then intersected with the cell-type-shared SnGs to identify age-related SnGs and IPF-related SnGs, respectively. This allowed us to further characterize the phenotypic relevance of senescence programs across distinct cellular contexts.

2. Evaluation and benchmarking

2.1. Robustness test

To assess the robustness of DeepSAS against variations in dataset composition, we conducted a series of subsampling experiments designed to simulate real-world scenarios with limited or imbalanced data. The full dataset (comprising IPF and healthy control samples) served as the reference for comparison. We generated multiple modified versions of the data by applying strategic subsampling or removal procedures and evaluated how consistently DeepSAS could recover SnCs and associated predictions. Specifically, we performed ten distinct subsampling schemes. For each, we randomly selected a fixed fraction (2/3, 1/2, and 1/3) of cells from the healthy samples and IPF samples, either jointly or separately. In total, six subsampling conditions were created:

(i) two-thirds of the cells from both healthy and IPF samples
(ii) one-half from both
(iii) one-third from both
(iv) two-thirds from only IPF samples
(v) one-half from only IPF samples
(vi) one-third from only IPF samples

In addition, to evaluate robustness to missing cellular contexts, we performed a cell-type dropout test: in each run, we randomly selected 10 out of the 25 annotated cell types and removed all cells from those types.

Each of the above subsampling scenarios was repeated five times to ensure consistency. DeepSAS was independently applied to each replicate. The set of SnCs identified in each subsampled run was compared to the original SnCs predicted from the full dataset. We recorded the overlap rate and number of overlapping SnCs between the subsampled and reference results to quantify stability.

2.2. Ablation test

To dissect the contribution of individual components in the DeepSAS framework, we conducted an ablation study by systematically modifying or removing specific modules. Each variant was benchmarked against the full DeepSAS model with default settings to evaluate its impact on SnC identification performance. First, we examined the role of CCC by disabling the CCC-based edge augmentation in the graph construction step. This allowed us to assess the importance of incorporating predicted intercellular signaling into the heterogeneous graph structure. Second, we tested the effect of node set size by varying the number of genes included per cell. Specifically, we reduced the default gene set to subsets of 8,000 and 3,000 genes, simulating conditions with limited gene coverage or lower sequencing depth. Third, we evaluated the sensitivity of DeepSAS to different senescence hallmark gene sets used to initialize senescence-associated gene (SnG) candidates. We tested four alternatives: (i) Fridman only, (ii) SenMayo only, (iii) SenMayo + CellAge, and (iv) SenMayo + Fridman. Each configuration was run independently, and outputs were compared to the default multi-source initialization. Performance metrics, including the number and distribution of identified SnCs and SnGs and overlap with reference results, were used to quantify changes in model behavior. This ablation study highlights the relative contributions of key architectural choices and confirms the robustness and interpretability of DeepSAS under varying input assumptions and configurations.

2.3. Comparison with other tools

To evaluate DeepSAS in the context of senescent cell identification, we performed a benchmarking analysis against several contemporary computational approaches selected for their focus on rare cell populations, transcriptomic heterogeneity, or senescence-specific modeling.

- ***SenCID***[4] trains a support vector machine using published bulk RNA-seq data from both normal and SnCs. One limitation of SenCID is that its training data were collected from in vitro studies, which cannot be transferred to implicate in vivo conditions. Also, it often predicts an unreasonably high proportion of cells (>70%) as senescent in every cell type, based on our tests, suggesting poor specificity and possible overfitting to general stress responses. Critically, rather than identifying cell-type-specific SnGs, SenCID identifies six senescence gene sets to cover senescence characterization in 30 cell types, yet it is not capable of generating new SnGs for new input scRNA-seq data.
- ***SenePy***[5] separates the derivation of senescence gene signatures from cell scoring, leading to biased or suboptimal identification of SnCs and SnGs. Its reliance on co-expression and binarized networks makes it sensitive to dropout, likely missing lowly expressed or transient markers. Moreover, our team tested the SenePy tool and found that it cannot generate new cell-type-specific SnGs from a new dataset, lacking the discovery power needed to derive cell-type-specific SnGs directly from input scRNA-seq data.
- ***DeepScence***[6] lacks cell-type specificity, relies on self-compiled senescence hallmarks for training, and does not infer SnGs. Its binary classification approach may also miss subtle or context-specific senescence phenotypes. These constraints position DeepScence as less flexible and less biologically contextualized. Since this work has not yet been published, we did not include it in the benchmark.
- **Seurat+ChatGPT**. We adapted a strategy inspired by the GPTCelltype framework[26], using Seurat for differential expression analysis alongside large language model (LLM)-assisted annotation. Specifically, for each cluster identified within the dataset, we extracted the top 20 DEGs using Seurat. These DEGs were embedded into a structured prompt—"This is a list that includes senescent cell marker:[*Genelist*] I have [*number*] clusters, some of them are senescent cells, provide the accurate cell names with whether they are senescent cell and the senescent biomarkers you use to do senescent dentification the gene markers are found here: (1) Differential expressed genes for group 0: [*Genelist*]; (2) Differential expressed genes for group 1: [*Genelist*]…"—which we input to ChatGPT-4. The resulting annotations leveraged both the statistical power of Seurat's clustering and the contextual reasoning provided by LLM to discern potential senescent subsets.
- **MarsGT[27]+ChatGPT**. Because senescent cells often represent a minority population, we additionally tested MarsGT, a method specifically tailored to detect rare cell types. MarsGT was first used to cluster the transcriptomic dataset, after which we applied Scanpy to identify the top 20 DEGs in each rare cell cluster. These gene lists were then incorporated into the same GPT prompt template described above and analyzed via ChatGPT-4 to further refine senescence-related cluster assignments.

3. Case study: human retinal pigment epithelium and choroid snRNA-seq data

snRNA-seq data from the human RPE and choroid were obtained from the posterior segments of postmortem eyes of 53 donors in Dr. Rui Chen's lab. Nuclei were isolated, and snRNA-seq libraries were generated using the Chromium Next GEM Single Cell 3′

Reagent Kit v3.1 (10x Genomics) and sequenced on the NovaSeq 6000 platform (Illumina). Data processing followed the methodology previously described in a preprint by Jin Li et al.[28]. Briefly, sequencing reads were aligned using Cell Ranger (v7.1.0, 10x Genomics), and quality control was performed using cellqc (https://github.com/lijinbio/cellqc) with the following criteria: correction for ambient RNA contamination; removal of low-quality nuclei with fewer than 300 detected features, fewer than 500 total transcript counts, or with >10% of reads mapping to mitochondrial genes; and doublet detection using DoubletFinder. Batch correction and data integration across multiple samples were conducted using scVI[29]. Major cell types were annotated based on canonical marker genes. The dataset comprising 716,410 single cells was randomly subsampled but preserved the same cell proportion into 29 batches. The first 28 batches involve 25,000 cells per batch, and the final (29th) batch included the remaining 16,410 cells. Each batch was independently analyzed using DeepSAS.

4. Case study: human IPF single-cell Flex and Xenium data

4.1 Flex scRNA-seq data generation and analysis

Sample collection and processing

Collection of human lung tissue was approved by the Institutional Review Board (IRB) of The Ohio State University (OSU) under study ID 2021H0180 and was performed by the Comprehensive Transplant Center (CTC) Human Tissue Biorepository at OSU with Total Transplant Care Protocol informed consent and research authorization from the donors. CTC operates in accordance with NCI and ISBER Best Practices for Repositories. The demographic summary of the study population is shown in **Supplementary Table S12.** Lung tissues were collected from the upper and lower lobes of lung explants of patients diagnosed with IPF and from the parenchymal anatomical region of age-matched and young non-IPF donors. The lung tissue was minced into pieces of 10 mm x 20 mm and fixed for 24 h in 4% paraformaldehyde and then was transferred to 70% ethanol and embedded in paraffin (wax) to create a Formalin-Fixed Paraffin-Embedded (FFPE) lung tissue block.

10X Flex scRNA-seq library preparation

Fixed RNA Profiling was performed on one biological sample and two biological samples from the parenchymal lung region of young non-IPF and age-matched normal donors, respectively, and on two biological samples from the parenchymal lung region of upper and lower lobes from two unique IPF donors. For each donor or explant, two FFPE blocks were obtained and preprocessed to evaluate lung architecture (H&E staining) and RNA quality (parameter >30% DV200) to select one FFPE block per sample. For each sample, two tissue sections at 50 µm were then cut from the FFPE blocks and placed onto 1.5 ml PCR clean tube. Then, from cells isolated from FFPE tissue using 1 mg/ml Liberase TH (Dissociation protocol, CG000632, Rev D, 10x genomics). The fixed RNA profiling and library preparation (Chromium Fixed RNA Profiling Reagent Kits for Multiplexed Samples, CG000527, Rev F, 10x Genomics) were performed based on human probes to identify whole transcriptome, with a target of 128,000 cells for capture, 8,000 cells per sample for a single reaction of 16 samples, each sample with a barcode. Next-generation sequencing was carried out in the Advanced Genomics Core at the University of Michigan and in the Novogene Corporation Inc. Gene expression libraries were sequenced on Illumina NovaSeq 6000 sequencer with sequencing depth of 15,000 read pairs per cell, Paired-end, dual indexing type, read lengths of 28 cycles Read 1, 10 cycles i7 index, 10 cycles i5 index, 90 cycles Read 2. scRNA-seq data was extracted from the raw sequencing data using Cell Ranger (version 7.1.0, 10x Genomics).

IPF raw sequence mapping and quantification
Cell Ranger[30] (version 2.1.0, 10x Genomics) was used for read alignment, transcript reconstruction/annotation, abundance estimation, read quality trimming, and sample quality analysis. Background expression due to ambient RNA was mitigated using CellBender[31], while doublets were identified using DoubletFinder[32]. The output gene expression file was analyzed using Seurat v4.0[33]. Low-quality cells were identified and removed based on high mitochondrial and/or ribosomal counts and low transcriptional counts. The preprocessed expression matrix was then used for dimensional reduction and cell clustering using Seurat. DEGs between different clusters were determined by the Wilcoxon Rank Sum test, where p-values were adjusted for multiple testing using the Bonferroni procedure. Clusters were annotated by comparing DEGs to canonical markers from public-domain (**Supplementary Table S5**).

4.2. Xenium data generation and validation

Sample preparation
Human lung tissue was obtained with IRB approval (2021H0180) from the OSU Comprehensive Transplant Center (CTC) Human Tissue Biorepository, adhering to NCI and ISBER Best Practices for Repositories. Lung tissues were collected from the parenchymal region of the upper and lower lobes of explanted lungs from patients diagnosed with IPF and from age-matched healthy donor lungs (**Supplementary Table S12**). The lung tissue was minced into pieces of 10 mm x 20 mm and fixed for 24 h in 4% paraformaldehyde and then was transferred to 70% ethanol and embedded in paraffin (wax) to create a Formalin-Fixed Paraffin-Embedded (FFPE) lung tissue block.

Gene panel design
Xenium in situ technology relies on a pre-defined gene panel. Each probe consists of two complementary sequences targeting the mRNA of interest and a unique gene-specific barcode. Upon hybridization, the probe circularizes and undergoes rolling circle amplification, amplifying the signal for target detection and decoding. A total of 480 genes were derived from a standalone human lung Xenium panel designed by Lung Aging Lab and analyzed using Chemistry version v1 on Xenium Analyzer instrument serial number XETG00239. The custom panel was curated based on human lung single-cell analysis data generated by Lung Aging Lab (**Supplementary Table S11**), focusing on genes relevant for cell type identification and potential involvement in aging, cellular senescence and IPF.

Cell segmentation and annotation
The Xenium In Situ Cell Segmentation Kit (Xenium Multi-Tissue Stain), PN-1000662 (10x genomics) was used to improve the determination of cell boundaries by using a stain and algorithmic technique developed and validated through custom-trained machine learning models. We manually annotated cell clusters and confirmed that Xenium spatial transcriptomics captured the heterogeneous cellular populations within lung tissue. Alveolar type 1 (AT1) cells were identified by the expression of *CLIC5*, *AGER*, and *RTKN2*, while alveolar type 2 (AT2) cells were characterized by *SFTPC*, *LAMP3*, and *HHIP*. Basal cells exhibited high expression of *KRT17*, *TP63*, and *KRT5*, whereas secretory cells showed increased levels of *MUC5B*, *MUC5AC*, and *SCGB3A1*. Ciliated cells, which are developmentally related to secretory cells, displayed partial expression of secretory markers alongside strong expression of *TMEM190*, *HYDIN*, and *TPPP3*. In addition to canonical epithelial populations, we were able to identify a pathological RAS cell population, characterized by low *SFTPC* and elevated *SCGB3A2* and *SFTPB*. Among

stromal populations, adventitial fibroblasts were defined by *PI16*, *MFAP5*, and *SCARA5* expression. CTHRC1+ fibroblasts expressed *CTHRC1*, *POSTN*, and *COL10A1*, while alveolar fibroblasts were marked by *CES1*, *WNT2*, and *TCF21*. Smooth muscle cells (SMCs) and pericytes were identified via *MYH11*, *MYL9*, *CSPG4*, and *RGS5*. We also distinguished lymphatic endothelial and blood endothelial cells through the expression of *CCL21*, *TFF3*, *MMRN1*, *LYVE1*, *DKK2*, *PLA1A*, and *PLVAP*. Immune populations were robustly represented, including mast cells (*KIT*, *MS4A2*, *SLC18A2*) and $CD4^+$ T cells (*CD4*, *CD40LG*, *FOXP3*, *IL7R*). $CD8^+$ T cells and natural killer (NK) cells were identified by the expression of *CD8A*, *GZMK*, *NKG7*, *KLRF1*, *KLRD1*, and *GNLY*. Alveolar macrophages and classical monocytes expressed *MERTK*, *CD163*, *S100A9*, and *S100A8*, while *SPP1*+ macrophages co-expressed *SPP1* with tissue-resident markers such as *MARCO*, *C1QA*, and *APOE*. B cells and plasma cells were distinguished by expression of *MS4A1*, *BANK1*, *CD19*, *IGLC1*, *TNFRSF18*, and *JCHAIN*. These results confirmed that the Xenium platform effectively captures the cellular and molecular complexity of fibrotic lung tissue, thereby providing a robust spatial validation framework for the senescence-associated genes predicted by DeepSAS.

Data Availability

The raw and processed IPF scRNA-seq and Xenium spatial transcriptomics data generated in this study are preserved and will be made available upon publication. Access to the data before publication can be provided upon reasonable request to the corresponding author, Dr. Qin Ma (qin.ma@osumc.edu). The human RPE and choroid snRNA-seq atlas is not yet publicly available, as they are currently under review for publication. Access to the data can be provided upon reasonable request to Dr. Rui Chen (rui.chen@uci.edu).

Code Availability

The DeepSAS source code is available at https://github.com/chthub/deepsas. All codes are publicly available and open source under the MIT License.

Acknowledgement

This work was supported by NIH grants U54AG075931 (TriState SenNet) and R01GM152585 (QM). Dr. Rui Chen and the generation of the human RPE and choroid snRNA-seq atlas were supported by the Chan Zuckerberg Initiative (CZI) under awards 2019-002425, 2021-239847, and 2021-237885. We thank Dr. Qing Nie (University of California, Irvine) for his invaluable support in helping the cell–cell communication graph design and implementation.

Author Contributions

Conceptual design: Qin Ma
Computational framework design: Anjun Ma and Qin Ma
IPF experimental design: Ana Mora and Mauricio Rojas
IPF sample processing and data analysis: Natalia Del Pilar Vanegas, Ahmed Ghobashi, Jhonny Rodriguez, and Lorena Rosas
Choroid sample support and analysis: Jianming Shao, Ahmed Ghobashi, and Rui Chen,
Code deployment and optimization: Hao Cheng and Hu Chen
Benchmarking and testing: Yi Jiang, Hu Chen, Xiaoying Wang, and Anjun Ma
Manuscript draft: Anjun Ma, Hao Cheng, Ana Mora, Mauricio Rojas, and Qin Ma

SenNet TriState knowledge support: Irfan Rahman, Jose Lugo-Martinez, Dongmei Li, Gloria S Pryhuber, Dongjun Chung, Ziv Bar-Joseph, Oliver Eickelberg, Melanie Königshoff, Toren Finkel
All authors revised the final manuscript.

Ethics declarations
Competing interests
The authors declare no competing interests.

REFERENCES

1 Di Micco, R., Krizhanovsky, V., Baker, D. & d'Adda di Fagagna, F. Cellular senescence in ageing: from mechanisms to therapeutic opportunities. *Nat Rev Mol Cell Biol* **22**, 75-95, doi:10.1038/s41580-020-00314-w (2021).PMC8344376

2 He, S. & Sharpless, N. E. Senescence in Health and Disease. *Cell* **169**, 1000-1011, doi:10.1016/j.cell.2017.05.015 (2017).PMC5643029

3 Kumari, R. & Jat, P. Mechanisms of Cellular Senescence: Cell Cycle Arrest and Senescence Associated Secretory Phenotype. *Front Cell Dev Biol* **9**, 645593, doi:10.3389/fcell.2021.645593 (2021).PMC8039141

4 Tao, W., Yu, Z. & Han, J.-D. J. Single-cell senescence identification reveals senescence heterogeneity, trajectory, and modulators. *Cell Metabolism* **36**, 1126-1143.e1125, doi:10.1016/j.cmet.2024.03.009 (2024)

5 Sanborn, M. A., Wang, X., Gao, S., Dai, Y. & Rehman, J. SenePy: Unveiling the Cell-Type Specific Landscape of Cellular Senescence through Single-Cell Analysis in Living Organisms. *bioRxiv*, 2023.2008.2030.555644, doi:10.1101/2023.08.30.555644 (2023)

6 Qu, Y. *et al.* Single-cell and spatial detection of senescent cells using DeepScence. *bioRxiv*, 2023.2011.2021.568150, doi:10.1101/2023.11.21.568150 (2025)

7 Suryadevara, V. *et al.* SenNet recommendations for detecting senescent cells in different tissues. *Nat Rev Mol Cell Biol* **25**, 1001-1023, doi:10.1038/s41580-024-00738-8 (2024).PMC11578798

8 NIH SenNet Consortium to map senescent cells throughout the human lifespan to understand physiological health. *Nat Aging* **2**, 1090-1100, doi:10.1038/s43587-022-00326-5 (2022).PMC10019484

9 Gil, J. The challenge of identifying senescent cells. *Nature Cell Biology* **25**, 1554-1556, doi:10.1038/s41556-023-01267-w (2023)

10 Lee, P. J. *et al.* NIH SenNet Consortium to map senescent cells throughout the human lifespan to understand physiological health. *Nature Aging*, doi:10.1038/s43587-022-00326-5 (2022)

11 Hernandez-Gonzalez, F. *et al.* Cellular Senescence in Lung Fibrosis. *International journal of molecular sciences* **22**, doi:10.3390/ijms22137012 (2021).PMC8267738

12 Schafer, M. J. *et al.* Cellular senescence mediates fibrotic pulmonary disease. *Nat Commun* **8**, 14532, doi:10.1038/ncomms14532 (2017).PMC5331226

13 Zhao, M., Chen, L. & Qu, H. CSGene: a literature-based database for cell senescence genes and its application to identify critical cell aging pathways and associated diseases. *Cell Death Dis* **7**, e2053, doi:10.1038/cddis.2015.414 (2016).PMC4816187

14 de Magalhães, J. P., Curado, J. & Church, G. M. Meta-analysis of age-related gene expression profiles identifies common signatures of aging. *Bioinformatics* **25**, 875-881, doi:10.1093/bioinformatics/btp073 (2009).PMC2732303

15 Freund, A., Orjalo, A. V., Desprez, P. Y. & Campisi, J. Inflammatory networks during cellular senescence: causes and consequences. *Trends Mol Med* **16**, 238-246, doi:10.1016/j.molmed.2010.03.003 (2010).PMC2879478

16 Casella, G. *et al.* Transcriptome signature of cellular senescence. *Nucleic Acids Res* **47**, 11476, doi:10.1093/nar/gkz879 (2019).PMC6868356

17 Saul, D. *et al.* A new gene set identifies senescent cells and predicts senescence-associated pathways across tissues. *Nat Commun* **13**, 4827, doi:10.1038/s41467-022-32552-1 (2022).PMC9381717

18 Fridman, A. L. & Tainsky, M. A. Critical pathways in cellular senescence and immortalization revealed by gene expression profiling. *Oncogene* **27**, 5975-5987, doi:10.1038/onc.2008.213 (2008).PMC3843241
19 Avelar, R. A. *et al.* A multidimensional systems biology analysis of cellular senescence in aging and disease. *Genome Biol* **21**, 91, doi:10.1186/s13059-020-01990-9 (2020).PMC7333371
20 Hasanaj, E. *et al.* SenSet, a novel human lung senescence cell gene signature, identifies cell-specific senescence mechanisms. *bioRxiv*, 2024.2012.2021.629928, doi:10.1101/2024.12.21.629928 (2024)
21 Tsukui, T., Wolters, P. J. & Sheppard, D. Alveolar fibroblast lineage orchestrates lung inflammation and fibrosis. *Nature* **631**, 627-634, doi:10.1038/s41586-024-07660-1 (2024)
22 Hao, Y. *et al.* Dictionary learning for integrative, multimodal and scalable single-cell analysis. *Nature biotechnology* **42**, 293-304 (2024)
23 Wang, S., Karikomi, M., MacLean, A. L. & Nie, Q. Cell lineage and communication network inference via optimization for single-cell transcriptomics. *Nucleic Acids Research* **47**, e66-e66, doi:10.1093/nar/gkz204 (2019)
24 Li, J., Zhou, P., Xiong, C. & Hoi, S. C. Prototypical contrastive learning of unsupervised representations. *arXiv preprint arXiv:2005.04966* (2020)
25 Kim, Y. & Park, W. in *Proceedings of the AAAI Conference on Artificial Intelligence.* 1827-1835.
26 Hou, W. & Ji, Z. Assessing GPT-4 for cell type annotation in single-cell RNA-seq analysis. *Nature Methods* **21**, 1462-1465, doi:10.1038/s41592-024-02235-4 (2024)
27 Wang, X. *et al.* MarsGT: Multi-omics analysis for rare population inference using single-cell graph transformer. *bioRxiv*, 2023.2008.2015.553454, doi:10.1101/2023.08.15.553454 (2023)
28 Li, J. *et al.* Integrated multi-omics single cell atlas of the human retina. *Res Sq*, doi:10.21203/rs.3.rs-3471275/v1 (2023).PMC10680922 Inc and Omniscope Ltd, and owns interests in Dermagnostix GmbH and Cellarity Inc. Other authors declare no competing interests.
29 Lopez, R., Regier, J., Cole, M. B., Jordan, M. I. & Yosef, N. Deep generative modeling for single-cell transcriptomics. *Nat Methods* **15**, 1053-1058, doi:10.1038/s41592-018-0229-2 (2018).PMC6289068
30 Zheng, G. X. *et al.* Massively parallel digital transcriptional profiling of single cells. *Nature communications* **8**, 14049, doi:10.1038/ncomms14049 (2017).PMC5241818 L.M., D.A.M., S.Y.N., M.S.L., P.W.W., C.M.H., R.B., A.W., K.D.N., T.S.M. and B.J.H. are employees of 10x Genomics.
31 Fleming, S. J. *et al.* Unsupervised removal of systematic background noise from droplet-based single-cell experiments using CellBender. *Nature methods*, doi:10.1038/s41592-023-01943-7 (2023)
32 McGinnis, C. S., Murrow, L. M. & Gartner, Z. J. DoubletFinder: Doublet Detection in Single-Cell RNA Sequencing Data Using Artificial Nearest Neighbors. *Cell Syst* **8**, 329-337 e324, doi:10.1016/j.cels.2019.03.003 (2019).PMC6853612
33 Hao, Y. *et al.* Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573-3587 e3529, doi:10.1016/j.cell.2021.04.048 (2021).PMC8238499