Identification of High-Risk Cells in Single-Cell Spatially Resolved Transcriptomics Data Using Deep Transfer Learning

Debolina Chatterjee[1] Justin L. Couetil[1] Ziyu Liu[2] Tianhan Dong[3]
Kun Huang[1] Chao Chen[4] Jie Zhang[1] Michael A. Kalwat[5]
Travis S.Johnson[1]

[1]Indiana University School Of Medicine, Indianapolis, IN
[2]Purdue University, West Lafayette, IN
[3]Takeda Oncology, Cambridge, MA
[4]Stony Brook University, Stony Brook, NY
[5] Indiana Biosciences Research Institute, Indianapolis, IN

Abstract

Background: The examination of high-risk cells and regions in tissue samples from spatially resolved transcriptomics platforms, offers meaningful insights into specific disease processes. For existing methods, while cell types or clusters can be identified and associated with disease attributes, individual cells are unable to be associated in the same manner which may result in failing to identify subsets of cells associated with disease attributes especially if the disease-associated cells cluster together with non-disease-associated cells.
Method: DEGAS (Diagnostic Evidence Gauge of Single-cells) [5], solves the above problem with a sophisticated deep transfer learning algorithm designed to identify high-risk components in single-cell RNA sequencing data from tumor samples. DEGAS employs latent representations of gene expression data and domain adaptation to transfer disease attributes from patients to individual cells. In this research, we present DEGAS's versatility in adapting to data arising from single cell spatially resolved transcriptomics platforms such as the 10X Genomics Xenium platforms, and Nanostring's CosMx platform. By integrating spatial location information from the above platforms, DEGAS can not only identify high-risk components in tissue samples but also pinpoint locations within the slides associated with disease status.
Results: We evaluated DEGAS across multiple platforms, including 10X Genomics Xenium and Nanostring CosMx. DEGAS successfully identified high-risk cells and regions, which were validated through known markers. Additionally, DEGAS was applied to our newly generated T2D Xenium dataset and a publicly available melanoma Xenium dataset. We tested DEGAS on publicly available Nanostring CosMx FFPE samples of normal and Hepatocellular Carcinoma tissues, revealing high-risk cells and topologies associated with key pathways. Notably, high-risk regions were predominantly enriched in tumor tissue, with DEGAS uncovering heterogeneity that correlates with aggressive disease markers and cell type diversity.
Keywords: Single-cell RNA Sequencing, Spatially Resolved Transcriptomics, Domain adaptation, Deep transfer learning

1 Introduction

Single cell spatially resolved transcriptomics (scSRT) combines sequencing and imaging technologies to analyze tissue samples at high resolution, providing valuable insights into disease processes. The scSRT data that have evolved within the last few years provide both spatial and gene expression

information, along with associated histological images of the tissue samples. Therefore, like scRNA-seq, deep transfer learning can also link scSRT data and bulk tissue transcriptomics, as scSRT data and sc-RNA seq data also share the same molecular feature space. Using transfer learning methods like domain adaptation and multitask learning, models can connect patient outcomes to scSRT cells and spatial locations within the tissue sections. DEGAS (Diagnostic Evidence Gauge of Single-cells) [5] successfully integrated different levels of omic data and identified latent molecular relationships between populations of cells and disease attributes and thereby accurately predicting patient characteristics at the single-cell level. Furthermore, DEGAS is a highly general framework which was applied across multiple diseases and data types. Using the rationale that single cell RNA-seq data and patient-level transcriptomic data (e.g., RNA-seq with clinical information) share a common set of genes, and hence same feature space, therefore, there exists a natural connection between the two data types and efficient transfer of such information between the data types can be extremely informative in identifying the associations between patients and cells. In this article, we show how DEGAS was adapted to incorporate spatial information from scSRT datasets. This adaptation not only allows for the identification of high-risk cell types but also locates regions within the tissue where these cells are concentrated, providing deeper insights into tissue niches that may be at high risk for developing a disease—ultimately aiding in the development of targeted therapies. This application note contains the following information i) a review of the DEGAS framework, ii) the smoothing strategy adapted for scSRT data, iii) results from various scSRT datasets from different platforms, and iv) a discussion of key findings summarizing the results.

2 Methodology

The DEGAS framework (Fig. 1) is designed using Python's TensorFlow library and provides an interface in R. A detailed explanation of the framework can be found in the original publication [5]. DEGAS is available for download on GitHub at https://github.com/tsteelejohnson91/DEGAS. Briefly described below are the key steps in the original DEGAS framework focused on integrating single-cell and bulk expression data for disease analysis. *Input:* Initially, a single-cell RNA sequencing (scRNA-seq) dataset and a bulk RNA sequencing (RNA-seq) dataset for patient-level expression data. Labels for patient samples (e.g., clinical outcomes) are matched to bulk data samples. Labels include categorical attributes (such as tumor, normal, etc.) or survival data with time and event indicators. scRNA-seq and bulk RNA-seq data are aligned to have rows as samples or cells and columns as genes. Patient sample labels are one-hot encoded into a matrix. For survival data, the first column represents the time to event, and the second column is the event indicator. This same approach can also be applied to spatial transcriptomics data [3] and scSRT data. *Preprocessing:* The main rationale is to find a set of common features among the clinical and the scRNA seq data. This is achieved by selecting genese associated with clinical outcomes that intersect between scRNA-seq and bulk RNA-seq data. It is worth noting that DEGAS is not a feature selection algorithm and if genes are selected with no association with the clinical outcomes of interest, then DEGAS will not perform optimally. Specific feature selection methods, such as LASSO (least absolute shrinkage and selection operator) selection, or statistical tests (e.g., t-tests), should be applied to define the the feature set. Gene expression values are then normalized by converting them to sample-wise z-scores for comparability. These z-scores are further scaled to the range $[0, 1]$ to improve compatibility with the deep learning models. *Model training:* A multitask deep learning model (e.g., a DenseNet with n layers) is employed. Separate output layers are used for cellular subtype predictions and clinical attribute predictions. To address differences between single cell and bulk expression data, domain adaptation techniques like Maximum Mean Discrep-

ancy (MMD) are employed. MMD minimizes the distributional discrepancies between the two data types in the latent representation. *Loss functions* for different tasks are defined: *Cox Proportional Hazards Loss*: For survival prediction using clinical data. *Classification Loss*: For both patient-level attributes and cell subtype predictions. *MMD Loss*: For aligning the distributions of single cell and bulk expression data.

If the single cell expression data includes cell type (i.e., fibroblast, epithelial, tumor clone, etc.) or disease-attribute labels (such as tumor vs. normal or diabetic vs. non-diabetic, etc.), and the pbulk expression data also contains class labels, a *ClassClass* model is used. If the single cell expression data lacks patient attributes but the patient-level data includes disease labels, the *BlankClass* model is applied. When the single cell expression data contains patient attributes, but the bulk expression data lacks disease labels, the *ClassBlank* model is implemented. If the single cell expression data has no attributes and the bulk expression data contains overall survival information, the *BlankCox* model is used. Finally, when both the single cell expression data contains different class attributes and the bulk expression data include survival information, the *ClassCox* model is used.

Training Steps: Input data is split into sets for bagging and bootstrap aggregation (BAg) is performed m times to improve the robustness of predictions. Key hyperparameters include: Number of training steps, Batch sizes for single-cell and bulk datasets, Dropout retention rates, and Regularization weights for different loss functions can be adjuseted to improve performance. Based on the original publication, default hyperparameter values are preset that have performed well across experiments [5]. *Disease Attribute Mapping*: Predictions of disease attributes are transferred from patient-level data to single-cell data and vice versa. The output is converted into interpretable scores (i.e., disease, survival, or progression associations) for downstream analysis. *Impressions*: These association scores are called *impressions* and represent correlative associations with disease or cellular attributes of interest. *Smoothing to accommodate scSRT data:* The DEGAS pipeline is largely the same between scRNA-seq and scSRT data with key differences in the smoothing steps. To adapt to spatial platforms, the following types of smoothing may be performed.

K-nearest neighbor (kNN) smoothing is applied to smooth the predictions for single cells at their spatial locations, enabling the visualization of disease attribute scores within the tissue context. Using spatial coordinates, we compute the K-nearest neighbors for each spatial point and perform smoothing based on their distances. Additionally, low-dimensional techniques like Uniform Manifold Approximation and Projection (UMAP) provide insights into cell clustering based on gene expression. By integrating both UMAP and spatial information into the smoothing process, we obtain more precise and accurate smoothing results. The disease attribute probabilities are transformed to a correlation coefficient like score within $[-1, 1]$ for disease risk. Let us denote the scaled scores by S_{scaled}. Following are the smoothing equations.

UMAP-Based kNN Smoothing

$$\mathrm{S}_{\text{umap}}[i] = \frac{1}{k} \sum_{j \in \mathcal{N}_k(i)} \mathrm{S}_{\text{scaled}}[j] \tag{1}$$

where $\mathcal{N}_k(i)$ represents the set of k nearest neighbors of i in the UMAP embedding space (UMAP1, UMAP2).

Spatial kNN Smoothing

$$\mathrm{S}_{\text{scaled_spatial}}[i] = \frac{1}{k} \sum_{j \in \mathcal{N}_k(i)} \mathrm{S}_{\text{scaled}}[j] \tag{2}$$

where $\mathcal{N}_k(i)$ is defined in the 2D spatial coordinates (x, y).

Combined UMAP and Spatial kNN Smoothing

$$\mathrm{S}_{\text{scaled_umap_spatial}}[i] = \frac{1}{k} \sum_{j \in \mathcal{N}_k(i)} \mathrm{S}_{\text{scaled_umap}}[j] \tag{3}$$

where neighbors $\mathcal{N}_k(i)$ are derived from the spatial coordinates (x, y).

Spatial-to-UMAP kNN Smoothing

$$\mathrm{S}_{\text{scaled_spatial_umap}}[i] = \frac{1}{k} \sum_{j \in \mathcal{N}_k(i)} \mathrm{S}_{\text{scaled_spatial}}[j] \tag{4}$$

where neighbors $\mathcal{N}_k(i)$ are determined based on UMAP coordinates $(\mathrm{UMAP.1}, \mathrm{UMAP.2})$. Where k denotes the Number of nearest neighbors; $\mathcal{N}_k(i)$ is a set of k nearest neighbors of sample i. $(\mathrm{UMAP1}, \mathrm{UMAP2})$ are Coordinates in the UMAP embedding space, and (x, y) are the Spatial coordinates. These equations iteratively smooth the data by averaging over nearest neighbors in different embedding spaces, enhancing the interpretability of spatial and transcriptomic patterns.

Gene-based smoothing: During model training, we start with a set of highly variable genes and gradually expand the number of included genes. To predict DEGAS hazard scores for each cell in a tissue sample, we aggregate model predictions from all iterations. The final score for each cell is obtained by averaging these predictions, which helps smooth the results when overlaying risk scores onto tissue images. Incorporating more genes refines the visualization, improving the clarity of risk score distributions and highlighting high-risk regions within areas of interest. This method, though based solely on gene information, can be combined with other smoothing approaches for enhanced results.

3 Single cell spatially resolved transcriptomic analyses

3.1 DEGAS identifies high-risk cell types in CosMx Hepatocellular carcinoma samples

The NanoString CosMx Spatial Molecular Imager (SMI) platform uses advanced RNA hybridization and imaging techniques to achieve spatially resolved gene expression data at single-molecule resolution. The gene expression data from the CosMx platform are presented as counts of RNA molecules, mapped to specific spatial coordinates within tissue sections, along with the information on the type of each individual cell. The key distinction of scSRT data from this platform compared to others lies in its resolution and approach: 10X Visium provides spatially resolved transcriptomics at the level of spots (50-55 μm) that encompass multiple cells, while 10X Xenium also achieves single-cell resolution but primarily utilizes multiplexed in situ hybridization.

The implementation of DEGAS on CosMx data offers the potential to uncover disease risks across various cell types within tissue samples. CosMx data is characterized by multiple Fields of View (FOVs) that collectively cover the entire tissue area, with each FOV containing specific details about its area and alignment position. Within each FOV, individual cells are mapped with precise spatial coordinates, and the locations of individual RNA molecules within these cells are also captured, enabling detailed spatial resolution of gene expression. To demonstrate the versatility of DEGAS in analyzing such datasets, we utilized publicly available hepatocellular carcinoma (LIHC) data from NanoString's website https://nanostring.com/products/cosmx-spatial-molecular-imager/ffpe-dataset/human-liver-rna-ffpe-dataset/. The dataset comprises two Formalin-Fixed, Paraffin-Embedded (FFPE) tissue sections analyzed with the Human Universal

Cell Characterization Panel (1000-plex). Sample 1 represents normal liver tissue (Male, Caucasian, Age 35), while Sample 2 corresponds to hepatocellular carcinoma tissue (Female, Caucasian, Age 65, Grade G3, Stage II). The gene expression metadata contains individual cell IDs; the global as well as within FOV X and Y-coordinates of the centroid of each cell; and the cell type of each individual cell are specified. There are 22 unique cell types. The data contains 340,517 cells in the Normal sample and 464,126 cells in the LIHC tumor sample.

The clinical metadata required for implementing the DEGAS workflow was sourced from Firebrowse, which provided bulk RNA-seq data from 377 human individuals along with their clinical information, including sex, age, and overall survival. The gene expression count data for each of the two samples was individually preprocessed according to DEGAS's specified preprocessing steps. However, there were no labels distinguishing single cells as tumor or normal, nor was there a disease/no disease classification for the patient-level data. As a result, the DEGAS BlankCox model was applied following the identification of a common gene set between the scSRT and the patient data. Subsequently, the 5-times aggregated, 3-layer Densenet model of DEGAS was utilized. The gene-bassed smoothing was performed for each single cell, and the average hazard prediction across the iterative selection of genes was retained as the final smoothed score. The output from the BlankCox model was then scaled using min-max scaling, resulting in a hazard score ranging from $[-1, 1]$. For patient-level predictions, we used binary overall survival labels (0 or 1) as the ground truth and compared them with the predicted hazard scores from DEGAS. The model exhibited strong discriminatory power, achieving a high ROC-AUC of 0.7664 (Fig. 2). This result further highlights the effectiveness of the trained model. For the single-cell or tissue level prediction, the heatmap of mean hazard scores (Fig. 3) illustrates high-risk regions (red) and low-risk regions (blue) for all cells within a field of view in both the normal and LIHC samples. As expected, there is a highly significant difference between the predicted risk scores in the normal and LIHC (liver hepatocellular carcinoma) samples (p-value $\leq 2e^{-16}$). Specifically, in the normal sample, 296,865 low-risk and 36,012 high-risk cells were identified, while in the LIHC sample, 336,896 low-risk and 123,545 high-risk cells were found. This reveals a significantly higher number of high-risk cells in the cancer sample (($\chi^2 = 30833$,p-value$\leq 2.2e^{-16}$), based on a Chi-squared test assessing the difference in the number of high-risk areas between the normal and cancer samples. To further investigate the risk associated with individual cell types, we plotted the overall hazard scores for each cell type within the liver tissues (Fig. 4).

In the normal liver sample, we observed that stellate cells, macrophages, and erythrocytes—representing stromal, immune, and blood cells—exhibited elevated risk. In contrast, in the LIHC sample, tumor cells, hepatocytes, and cholangiocytes showed higher average risk compared to endothelial or gamma delta T cells. Notably, regions in the normal tissue exhibiting high hazard scores for LIHC were enriched with stellate cells, non-inflammatory macrophages, mature B cells, and erythroid cells. Macrophages, in particular, are known to contribute to tumor growth by suppressing immune surveillance through the inhibition of cytotoxic T cells [8]. A similar pattern was observed in a DEGAS-based study of prostate cancer [3], where specific normal tissues were linked to cancer progression. In that study, high-risk normal regions were found to be infiltrated by macrophages, detectable in both transcriptomic data and histopathological images. In our study, we observed a comparable trend.

3.2 DEGAS identifies high-risk pancreatic cells associated with Type II diabetes using analysis of Xenium data

Pancreatic β-cell dysfunction has long been associated with the development of type II diabetes (T2D), and recent studies have further elucidated its role in disease pathogenesis [6, 2, 1, 4]. In

particular, β-cell dysfunction is now recognized as a key factor in the progression of T2D, with its impact on insulin secretion and glucose regulation becoming a central focus of research. To better understand the underlying mechanisms of T2D, our study aims to explore the spatial relationships between various regions in pancreatic tissue that have an elevated potential for developing T2D, even in nondiabetic (ND) individuals. By integrating this spatial information with the locations of β-cells within the tissue, we aim to provide a more comprehensive understanding of T2D pathogenesis and identify other cell types that may exhibit high-risk characteristics for the disease.

De-identified FFPE human pancreas tissue was obtained through the National Disease Research Interchange (NDRI) (Table 1). Tissue blocks were processed into 5 μm sections, and sections from two donors (two Non-diabetic and two Type 2 Diabetic) were mounted directly within the bounding box of the Xenium slide by Indiana University's Histology Lab Service Core. Slides were processed according to the Xenium v1 workflow for in-situ Gene Expression (CG000582) using the Human Multi-tissue and Cancer panel (377 genes) by the Indiana University Center for Medical Genomics according to the instructions from 10X Genomics Xenium protocol https://www.10xgenomics.com/. Data were processed and visualized in the Xenium Explorer software. The Xenium data was preprocessed using the DEGAS preprocessing steps, which involve filtering for high-quality cells and normalizing gene expression to account for technical biases, ensuring robust downstream analysis. For the patient-level bulk RNA-sequencing data, clinical metadata obtained from the GEO dataset GSE159984 [7], which includes data on 58 non-diabetic and 27 T2D patients, was used. This dataset provides important information on patient characteristics such as body mass index (BMI), sex, and disease status, which are critical for interpreting the results within a clinical context.

Since the scSRT data did not have any known labels for ND versus T2D or cell type labels for individual cells, and the patient-level clinical data included Normal vs. T2D classification information, we employed the 5-times aggregated 3-layer BlankClass DEGAS model. This model predicted the likelihood of each individual cell being categorized as associated with T2D or ND. Gene-based smoothing was implemented over the dEGAS risk predictions. The spatial plot of the DEGAS T2D risk association scores highlighted individual cells as low risk (blue) or high risk (red)(Fig. 5). As expected, the normal tissues exhibit a larger number of cells associated with low risk as compared to the diabetic samples.

3.3 Melanoma results

The DEGAS framework for Xenium can be utilized in situations where the patient-level bulk data contains survival information, where the patients either died or were censored. The bulk human subcutaneous melanoma (SKCM) data downloaded from Firebrowse. This reference data contained 293 patients and their overall survival in months. The Xenium scSRT data was sourced from publicly available 10X Xenium FFPE Human Skin dataset https://www.10xgenomics.com/. The dataset consists of a single tissue sample. Using overall survival data from patients as a reference (as outlined in subsection 3.2), we applied the BlankCox DEGAS model to analyze the melanoma data, with overall survival as the patient-level outcome. The gene-based smoothing was implemented to find the smoothed DEGAS hazard scores. The output from the BlankCox model was then scaled using min-max scaling, resulting in a hazard score ranging from $[-1, 1]$. The disease association scores were calculated The smoothed disease association score was overlaid with the tissue's spatial coordinates, highlighting areas of disease association within the tissue (Fig.6). Cells were categorized as low-risk if their disease association score was below the median, and as high-risk if the score exceeded the median. after conducting differential gene expression analysis, and calculating the log base 2 of the fold changes ($\log_2$FCs) between low- and high-risk genes, we identified several differentially expressed genes, such as CLDN1 ($\log_2$FC = 0.323), KIT

($\log_2$FC = 0.08), and DMKN ($\log_2$FC = 0.51), which were significantly upregulated in high-risk cells. Conversely, genes like ADAM12 ($\log_2$FC = -0.03), CD68 ($\log_2$FC = -0.41), COL6A1 ($\log_2$FC = -0.38), CTSZ ($\log_2$FC = -0.33), and MITF ($\log_2$FC = -0.026) were significantly downregulated in high-risk cells when compared to low-risk cells. Some of these genes, including KIT and MITF [9, 10], have been associated with melanoma progression. In the DEGAS risk overlay, we observed that the tissue's outer margins and larger central regions were predominantly highlighted as high-risk, while the inner areas were largely classified as low-risk. The segregation of high- and low-risk cells led to the identification of differentially expressed genes, some of which could serve as potential therapeutic targets. Our results demonstrate the effectiveness of DEGAS in identifying high-risk regions within Xenium tissue samples.

4 Summary

In summary, this article highlights the versatility and effectiveness of the DEGAS method in integrating spatial information from scSRT technologies patient-level bulk expression datasets to identify high-risk regions and cell populations within tissue samples. We demonstrated DEGAS's seamless application across multiple scSRT platforms and diseases, including 10X Genomics Xenium and Nanostring's CosMx and LIHC, T2D, and Melanoma, respectively. Additionally, we implemented smoothing techniques that account for spatial localization and created a novel pipeline for scSRT datasets. The application of DEGAS led to meaningful findings, such as identifying high-risk cell types in liver cancer (LIHC) and revealing potential normal tissue markers linked to cancer progression. Furthermore, DEGAS provided a novel approach to understanding the pathogenesis of Type 2 Diabetes (T2D) by uncovering key interactions between pancreatic β-cells and other cell types. These discoveries position DEGAS as a powerful and efficient tool for advancing translational research involving scSRT, offering significant potential for early diagnosis, prevention, and therapeutic interventions.

5 Software and data availability statement

The DEGAS software is publicly available at https://github.com/tsteelejohnson91/DEGAS. The sources of the above datasets are mentioned in the respective sections.

Acknowledgement

This work was funded by the following grants. T.S.J and J.Z. received funding from NIH-NCI (R21CA264339) and T.S.J., J.Z., and C.C. received funding from NIH-NIGMS (1R01GM148970). General funding support from the IU Precision Health Initiative and AnalytixIN are also acknowledged.

References

[1] Marianne Böni-Schnetzler and Daniel T Meier. Islet inflammation in type 2 diabetes. In *Seminars in Immunopathology*, volume 41, pages 501–513. Springer, 2019.

[2] Marlon E Cerf. Beta cell dysfunction and insulin resistance. *Frontiers in Endocrinology*, 4:37, 2013.

[3] Justin L Couetil, Ziyu Liu, Ahmed K Alomari, Jie Zhang, Kun Huang, and Travis S Johnson. Diagnostic evidence gauge of spatial transcriptomics (degas): Using transfer learning to map clinical data to spatial transcriptomics in prostate cancer. *bioRxiv*, pages 2023–04, 2023.

[4] Phiwayinkosi V Dludla, Sihle E Mabhida, Khanyisani Ziqubu, Bongani B Nkambule, Sithandiwe E Mazibuko-Mbeje, Sidney Hanser, Albert Kotze Basson, Carmen Pheiffer, and Andre Pascal Kengne. Pancreatic β-cell dysfunction in type 2 diabetes: Implications of inflammation and oxidative stress. *World journal of diabetes*, 14(3):130, 2023.

[5] Travis S Johnson, Christina Y Yu, Zhi Huang, Siwen Xu, Tongxin Wang, Chuanpeng Dong, Wei Shao, Mohammad Abu Zaid, Xiaoqing Huang, Yijie Wang, et al. Diagnostic evidence gauge of single cells (degas): a flexible deep transfer learning framework for prioritizing cells in relation to disease. *Genome medicine*, 14(1):11, 2022.

[6] Masato Kasuga et al. Insulin resistance and pancreatic β cell failure. *The Journal of clinical investigation*, 116(7):1756–1760, 2006.

[7] Lorella Marselli, Anthony Piron, Mara Suleiman, Maikel L Colli, Xiaoyan Yi, Amna Khamis, Gaelle R Carrat, Guy A Rutter, Marco Bugliani, Laura Giusti, et al. Persistent or transient human β cell dysfunction induced by metabolic stress: specific signatures and shared gene expression with type 2 diabetes. *Cell Reports*, 33(9), 2020.

[8] Stefano Ugel, Francesco De Sanctis, Susanna Mandruzzato, Vincenzo Bronte, et al. Tumor-induced myeloid deviation: when myeloid-derived suppressor cells meet tumor-associated macrophages. *The Journal of clinical investigation*, 125(9):3365–3376, 2015.

[9] Kevin Yang, Allen SW Oak, Radomir M Slominski, Anna A Brożyna, and Andrzej T Slominski. Current molecular markers of melanoma and treatment targets. *International journal of molecular sciences*, 21(10):3535, 2020.

[10] Lamberto Zocchi, Alberto Lontano, Martina Merli, Emi Dika, Eduardo Nagore, Pietro Quaglino, Susana Puig, and Simone Ribero. Familial melanoma and susceptibility genes: a review of the most common clinical and dermoscopic phenotypic aspect, associated malignancies and practical tips for management. *Journal of clinical medicine*, 10(16):3760, 2021.

Tables

Table 1: Details of the Xenium tissue samples used in Section 3.2. ICH stands for intracerebral hemorrhage

Sample ID	**Gender**	**Race**	**Age**	**BMI**	**Diagnosis**	**Cause of Death**	**T2D Duration**
ND_1	Male	White	82	28.4	Normal	Stroke, ICH	NA
$T2D_2$	Male	White	82	37.8	T2D	Respiratory arrest	30 yrs
$T2D_3$	Male	White	77	26.3	T2D	Stroke/ICH	15 yrs
ND_4	Female	White	78	23.7	Normal	Stroke, ICH	NA

Figure Captions

Figure 1. DEGAS workflow: **Procesing data:** scSRT data and bulk expression data are first preprocessed into expression matrices. **Training:** Bootstrap-aggregated DenseNet DEGAS models are trained using both single-cell and patient disease attributes through a multitask learning neural network. This network learns latent representations that minimize differences between patient and single-cell data at the final hidden layer, using maximum mean discrepancy (MMD). **Predicting with DEGAS:** The model's output layer allows for the simultaneous inference of disease attributes in single cells and cellular composition in patients.

Figure 2. ROC curve of the DEGAS model for classifying overall survival status.

Figure 3. Spatial visualization of scaled hazard scores from the DEGAS output for the CosMX normal tissue (left) and cancer tissue (right). The LIHC cancer tissue exhibits a higher proportion of high-risk cells, which are spatially clustered, in contrast to the normal tissue.

Figure 4. Mean hazard scores for different cell types in Normal (left) and LIHC cancer (right) tissues. Tumor cells in the LIHC sample show higher mean hazard scores, while the normal tissue sample highlights stellate cells, macrophages, and erythroid cells as having stronger associations with LIHC hazard scores.

Figure 5. Spatial visualization of Type 2 Diabetes scaled hazard scores generated by the DEGAS model, with the legend displaying scaled hazard scores on all four Xenium tissue samples.

Figure 6. Spatial visualization of SKCM scaled hazard scores generated by the DEGAS model for the Xenium SKCM sample, with the legend displaying scaled hazard scores.

Figures

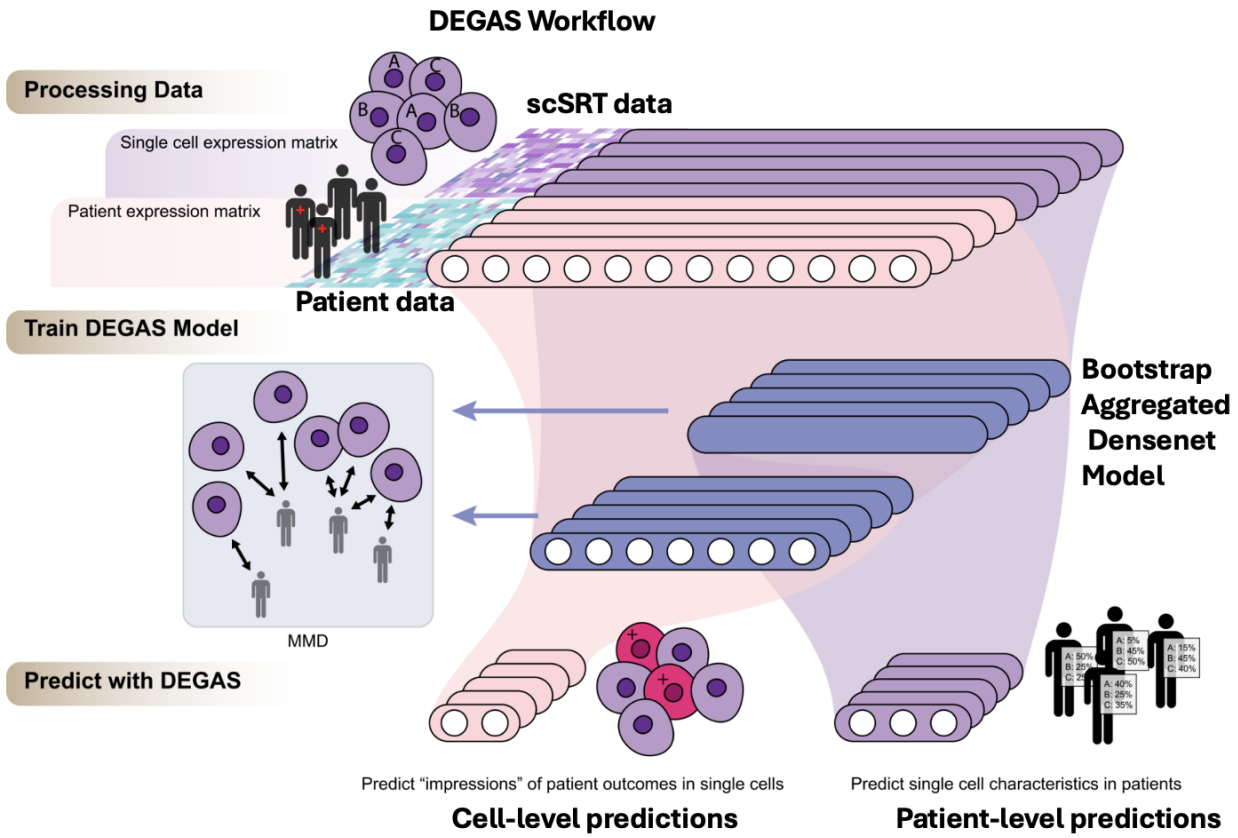

Figure 1: DEGAS workflow: **Procesing data:** scSRT data and bulk expression data are first preprocessed into expression matrices. **Training:** Bootstrap-aggregated DenseNet DEGAS models are trained using both single-cell and patient disease attributes through a multitask learning neural network. This network learns latent representations that minimize differences between patient and single-cell data at the final hidden layer, using maximum mean discrepancy (MMD). **Predicting with DEGAS:** The model's output layer allows for the simultaneous inference of disease attributes in single cells and cellular composition in patients.

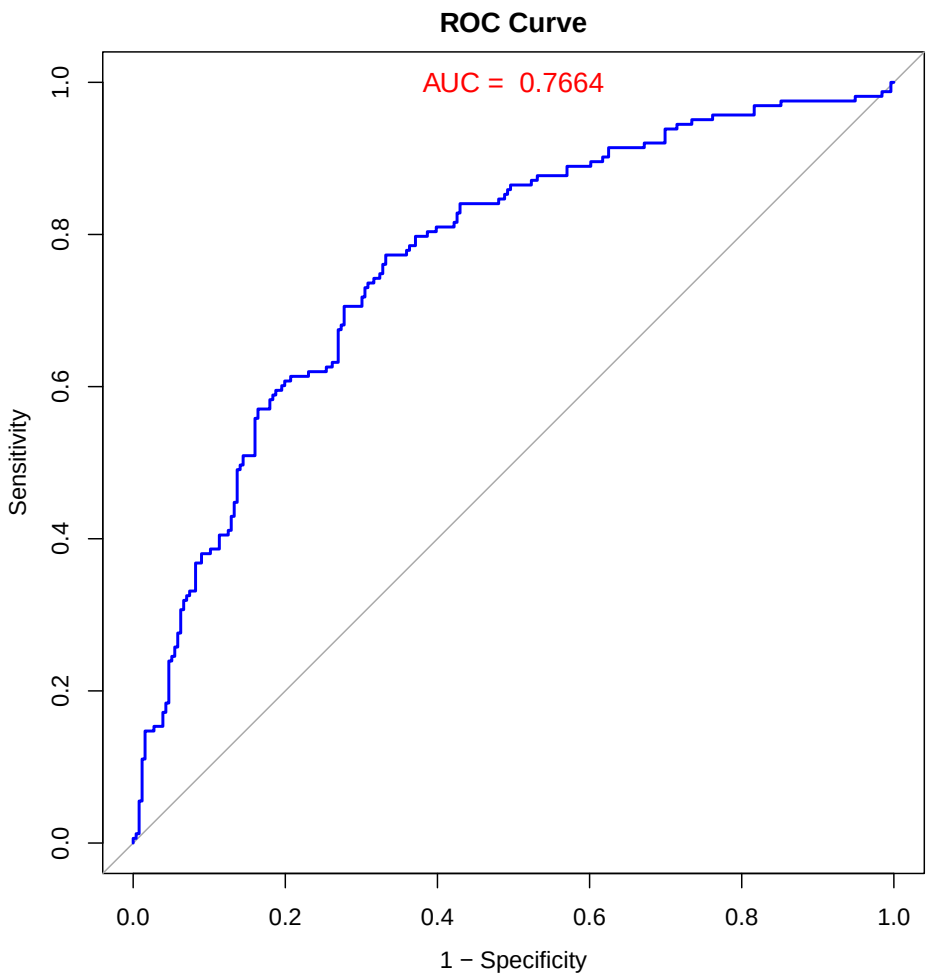

Figure 2: ROC curve of the DEGAS model for classifying overall survival status.

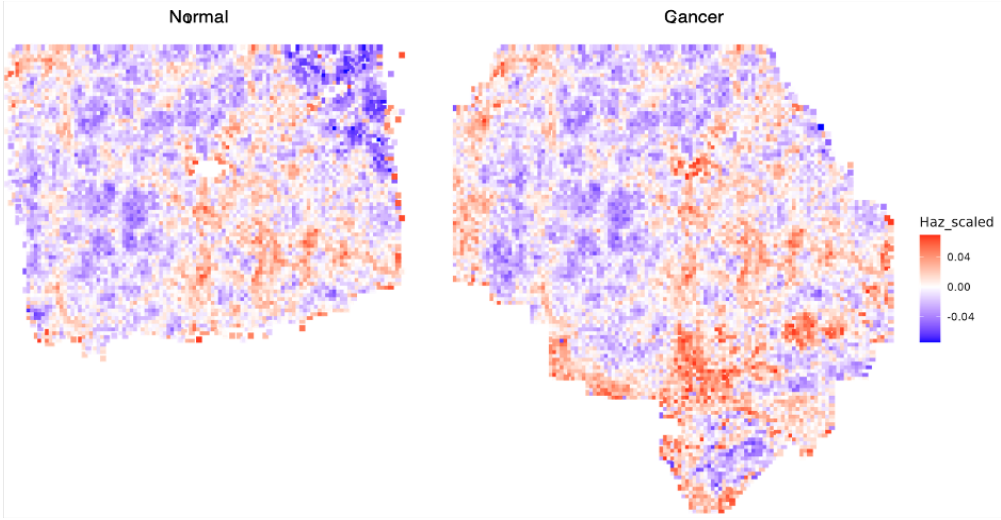

Figure 3: Spatial visualization of scaled hazard scores from the DEGAS output for the CosMX normal tissue (left) and cancer tissue (right). The LIHC cancer tissue exhibits a higher proportion of high-risk cells, which are spatially clustered, in contrast to the normal tissue.

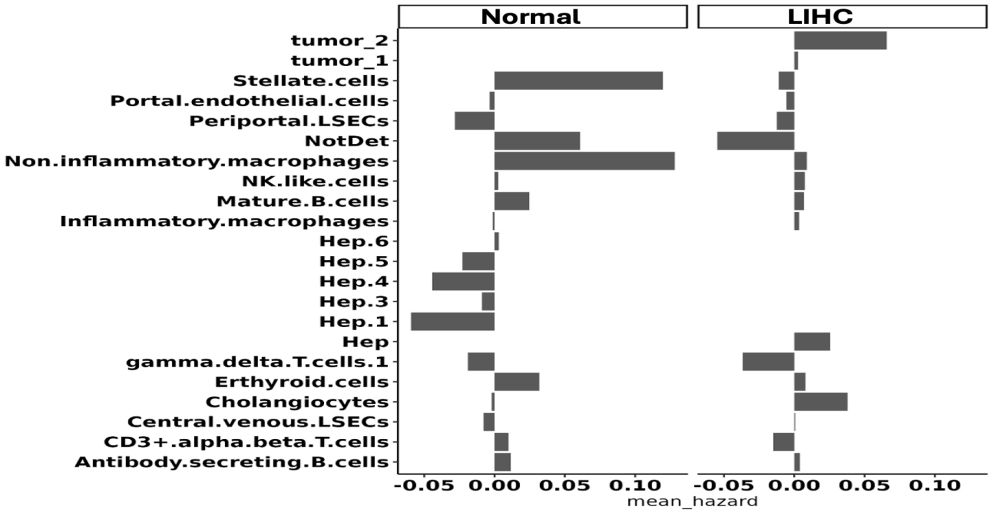

Figure 4: Mean hazard scores for different cell types in Normal (left) and LIHC cancer (right) tissues. Tumor cells in the LIHC sample show higher mean hazard scores, while the normal tissue sample highlights stellate cells, macrophages, and erythroid cells as having stronger associations with LIHC hazard scores.

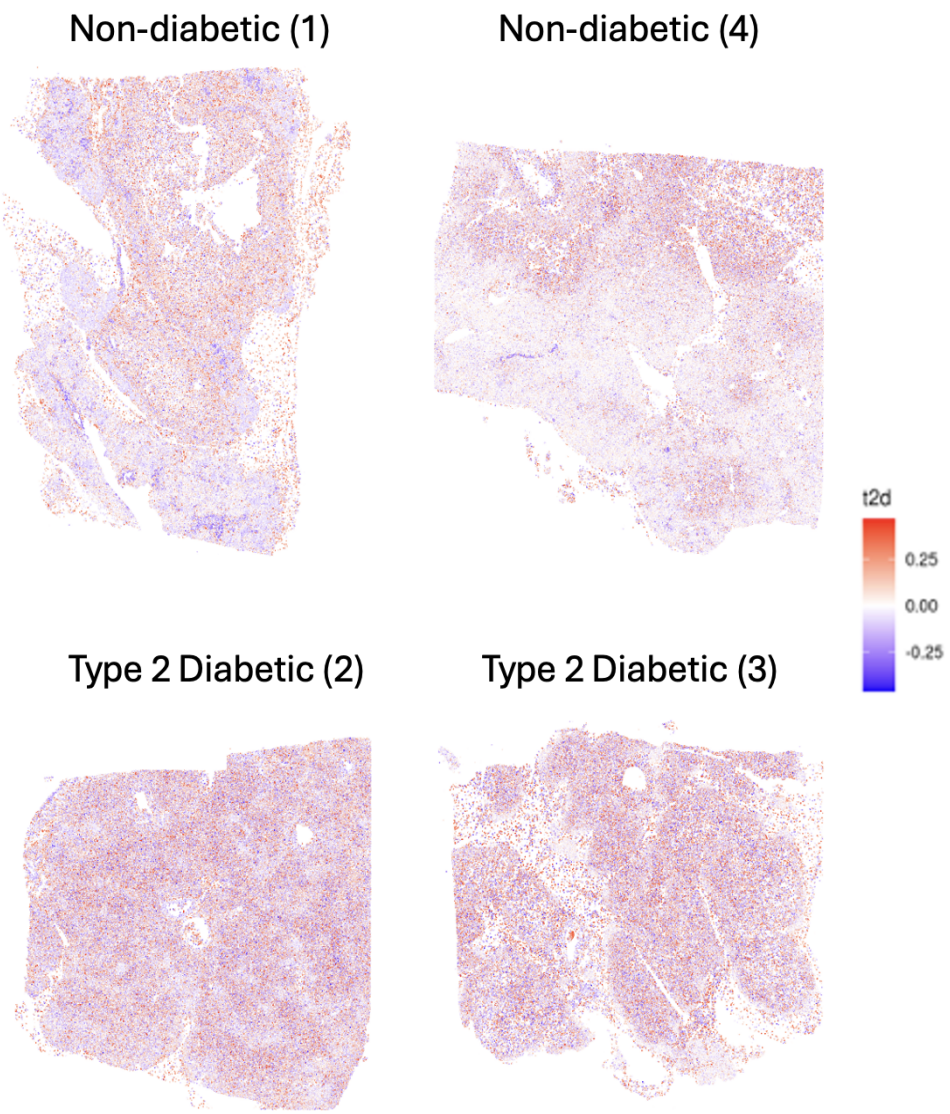

Figure 5: Spatial visualization of Type 2 Diabetes scaled hazard scores generated by the DEGAS model, with the legend displaying scaled hazard scores on all four Xenium tissue samples.

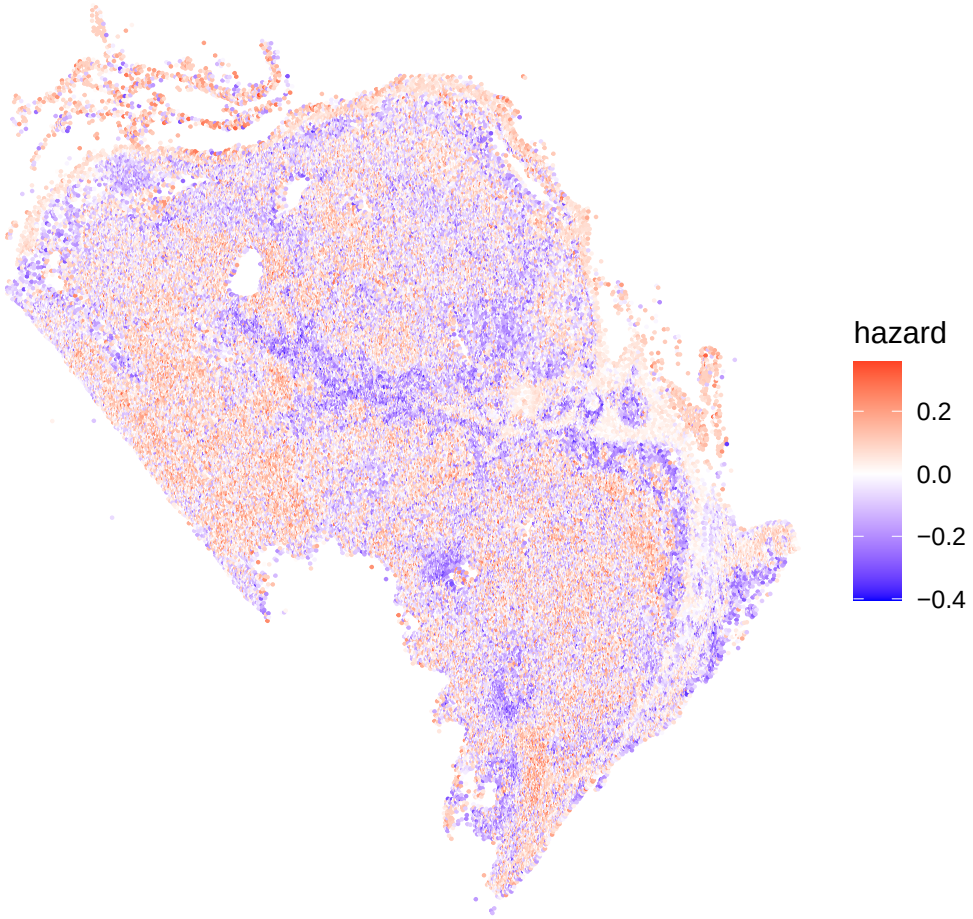

Figure 6: Spatial visualization of SKCM scaled hazard scores generated by the DEGAS model for the Xenium SKCM sample, with the legend displaying scaled hazard scores.