# FRONT MATTER

## Title

### Long Title

Diagnostic Evidence GAuge of Single cells (DEGAS): A flexible deep-transfer learning framework for prioritizing cells in relation to disease

### Short Title

Diagnostic Evidence GAuge of Single cells (DEGAS)

## Authors

Travis S. Johnson[1,2,3], Christina Y. Yu[1,2], Zhi Huang[4], Siwen Xu[5], Tongxin Wang[6], Chuanpeng Dong[5], Wei Shao[1], Mohammad Abu Zaid[1], Xiaoqing Huang[3], Yijie Wang[6], Christopher Bartlett[7], Yan Zhang[2,8], Brian A. Walker[9], Yunlong Liu[5,10], Jie Zhang[10*], Kun Huang[1,3,10,11*]

## Affiliations

[1]Department of Medicine, Indiana University School of Medicine
[2]Department of Biomedical Informatics, College of Medicine, The Ohio State University
[3]Department of Biostatistics and Health Data Science, Indiana University School of Medicine
[4]School of Electrical and Computer Engineering, Purdue University
[5]Center for Computational Biology and Bioinformatics, Indiana University School of Medicine
[6]Department of Computer Science, Indiana University
[7]Nationwide Children's Hospital
[8]The Ohio State University Comprehensive Cancer Center (OSUCCC - James)
[9]Division of Hematology Oncology, Melvin and Bren Simon Comprehensive Cancer Center, Indiana University
[10]Department of Medical and Molecular Genetics, Indiana University School of Medicine
[11]Regenstrief Institute
* To whom correspondence should be addressed (jizhan@iu.edu or kunhuang@iu.edu)

## Abstract

The advancement of single cell sequencing technology and corresponding analytical methods allow for granular studies of cellular heterogeneity in tissues. However, there lacks a direct approach to infer clinical information in single-cell data to prioritize cells with respect to disease status or patient outcomes. We propose *DEGAS* (Diagnostic Evidence GAuge of Single cells), a novel deep learning and transfer learning framework, to transfer prognostic information from patients to cells. We call such transferrable information "impressions," which allow individual cells to be associated with disease outcomes. Using simulated data as well as nine diverse single cell and patient bulk tissue transcriptomic datasets from Glioblastoma Multiforme (GBM), Alzheimer's Disease (AD), and Multiple Myeloma (MM), we demonstrate the feasibility, flexibility, and broad applications of the *DEGAS* framework. The analysis on MM led to the identification of a *PHF19*$^{high}$ myeloma cell subtype associated with myeloma progression. *DEGAS* is freely available as an R package.

## Teaser

By leveraging preexisting patient cohorts, *DEGAS* provides novel tools to identify high-risk cells in complex diseases at single cell resolution.

## MAIN TEXT

## Introduction

The emergence of single cell RNA sequencing (scRNA-seq) in 2009 has revolutionized the medical research community with single cell level resolution, providing a much deeper understanding of transcriptomic heterogeneity in tissues and diseases. Now that scRNA-seq is a standard part of the biomedical research toolbox, increasing numbers of scRNA-seq studies have been published [1, 2], and databases have quickly accumulated with scRNA-seq data, such as Hemberg lab [3], scRNASeqDB [4], SCPortalen [5], Allen Institute Cell Types Database, and the NCBI Gene Expression Omnibus (GEO) [5]. Many methods have been developed to analyze scRNA-seq data, the most notable being *Seurat*, which includes ways to cluster and normalize cell expression as well as perform integrative analysis with other data types (e.g., CITE-seq and ATAC-seq) [6]. For *Seurat* and similar methods, while cell types/clusters can be identified and associated with disease outcomes [7-10], individual cells are unable to be associated in the same manner. This may result in failing to identify subsets of cells associated with disease status or outcomes, especially if the disease-associated cells cluster together with non-disease-associated cells.

Currently, disease associated cell types are identified by transferring molecular heterogeneity information from cells to patients using expression deconvolution [11-13]. However, this approach is limited as it focuses on the changes in relative abundance of subtypes of cells instead of transcription changes of these cells. And the resolution of the cell subtyping is constrained by the clustering experiment. Therefore, novel machine learning methods that can transfer information from patients to cells and identify latent links between them are sorely needed to leverage the relative strengths of single cell and patient level data. For example, in cancer studies, bulk transcriptomic data is ideal for studying inter-tumor heterogeneity and scRNA-seq is ideal for studying intra-tumor heterogeneity. However, such integration faces numerous challenges since different data modalities and different data sources can have different characteristics in terms of quantity, quality, distribution and resolution [1]. For instance, it is common to find studies with a large number of patient samples for bulk tissue RNA sequencing (RNA-seq), whereas studies with scRNA-seq data usually contain a small number of patient samples. Most scRNA-seq experiments generate a large number of cells per sample, making the scaling of such data to multiple tissue samples computationally difficult [1]. On the other hand, a large patient sample size is often required for statistical studies such as outcome prediction and survival analysis [14]. If traditional methods were used, the resulting scRNA-seq data could end up with cell numbers on the scale of millions making such studies more difficult.

To address these challenges, previous studies have directly established associations of diseases with cell types derived from scRNA-seq without using deconvolution. These methods mainly utilize unsupervised methods and focused primarily on the number of differentially expressed genes (DEGs) in a given cell type corresponding to DEGs related to some clinical outcome [15, 16]. For example, Gawel *et al.* used enrichment of the cell cluster specific DEGs and multicellular disease models (MCDMs) to visualize the cell types for prioritization [7]. Alternatively, the cell type prioritization tool *Augur* did not primarily rely on DEGs, but still focused the biological resolution to the cell type level [17]. They trained classifiers on each cell type with respect to the disease state of the tissue from which those cells were sampled. The accuracy of the classifier in each cell type was used to prioritize its relation to the disease state of interest [17]. Both of these methods rely on either prior knowledge to calculate enrichment of DEGs or require scRNA-seq data from both disease and normal samples. Furthermore, all of these existing methods are reliant on accurately defining the cell types within a scRNA-seq experiment. In summary, these methods assign disease associations to the previously defined cell types and not to the individual cells.

To address such challenges as prioritizing individual cells in relation to disease with considerations on sample size and computational cost, we established the combined deep learning and transfer learning framework called *DEGAS* (Diagnostic Evidence GAuge of Single cells) to integrate scRNA-seq and bulk tissue transcriptomic data with the goal to transfer clinical information from patients to cells. The ability of *DEGAS* to assign patient-level disease attributes to single cells, among other functions, provides a flexible and useful tool to prioritize cells, cell types, patients, and patient subtypes in relation to disease outcomes. In this paper, we focus on the most relevant use case of assigning disease status and associating outcomes from patients to individual cells since there is no current state-of-the art technique to perform this task.

Taking transcriptomic data as an example, the rationale behind the *DEGAS* framework is that scRNA-seq data and patient-level transcriptomic data (e.g., RNA-seq with clinical information) share the same feature space (i.e., common set of genes). In addition, a natural connection exists between the two data types that can be leveraged to further identify the associations between patients and cells. Viewing this association as a graph (**Fig. 1**), we can connect the outcomes in patients to individual cells, via a latent representation of the common feature space (selected genes). This latent representation fitting two datasets can be learned using a transfer learning technique called domain adaptation [18-21]. Domain adaptation applies linear or non-linear transformations on the features for both datasets so that their distributions are similar after the transformations. Our biological intuition is thus: the expression patterns of genes in cells and tissues should carry a portion of the same biological patterns such as molecular pathways, signaling cascades, and/or metabolic processes, making the information learned from this portion of gene expression patterns transferable between patients and cells. Our hypothesis is that the latent representation learned from these shared gene expression patterns will be simultaneously predictive of patient outcomes and cellular subtypes. Similar hypotheses are already adopted to transfer information between different single cell experiments [6, 22-26] and to transfer information from bulk transcriptomic cell type atlases to single cell experiments [27].

In our *DEGAS* workflow, we incorporate patient-level outcome information with cell type information from disparate datasets to perform cell prioritization on scRNA-seq data. These disease associations in cells can be attributed to disease-related biological perturbations identified in the patients. This novel deep transfer learning approach simultaneously trains a model on single cell data and patient data along with their labels and learns a representation in which the cells and patients occupy the same latent space. Multitask learning, also known as parallel transfer learning, is precisely designed to achieve these two goals. Used extensively in computer vision, multitask learning learns a low dimensional representation of the input data to optimally address multiple tasks. Examples of such application in medical science include predicting benign versus malignant tumor samples and subclassification in breast cancer histology images [28, 29]. In this paper, we further extend this line of research to include datasets with patient outcomes that can be trained simultaneously so that the outcomes can be transferred (or cross-mapped) between single cells and patients. Specifically, our framework enables knowledge learned from patients using deep learning models to be transferred to single cells and vice versa. The major advantages of our transfer learning framework are that the single-cell gene expression data and clinical bulk gene expression data can come from different patient cohorts of the same disease without matched data while the disease associations can still be directly assigned to individual cells. This flexibility not only presents an ingenious way to integrate molecular omics data analysis in different levels, but also virtually merges them into the same cohort, which makes studying a broad variety of heterogeneous diseases possible.

The *DEGAS* framework in its simplest form can be broken into three tasks during model training: 1) correctly labeling cells with a cellular subtype using multitask learning; 2) correctly assigning clinical labels to patients using multitask learning; and 3) generating a latent space in which patients and cells are comparable using domain adaptation (**Fig. 1A**). To perform *DEGAS* analysis, first we select representative gene features that are predictive of cell type, predictive of patient outcome, and present at measurable levels in both scRNA-seq and bulk transcriptomic data. Secondly, we apply deep learning models to learn the latent representation of the single-cell and patient-level transcriptomic data, with the goal to simultaneously minimize cell type classification error, patient outcome prediction error, and the differences between cells and patients in their latent representation. Finally, the patient-level clinical information such as survival and clinical subtypes is predicted in the single cells using the patient label output layer and cell types are predicted in patients using the cell type output layer (**Fig. 1B**). We call these transferrable label probabilities "impressions" since information from gene expression of disparate data types and studies can be extracted and the characteristics from one data type can be mapped to another. We emphasize the ability to make predictions of patient outcomes in individual cells since there are currently not methods available to perform this task. *DEGAS* is developed as a generalizable model generating deep transfer learning framework that can be applied to any disease data as long as the data contain clinical information for a cohort of patients or a separate clustering analysis result on sets of cells from single cell level omic experiments of the same disease. There is not an inherent limitation to the input transcriptomic data. Furthermore, *DEGAS* can be potentially expanded to accommodate other modalities of data with proper normalization steps.

To demonstrate the feasibility and effectiveness of the *DEGAS* framework, we first tested it on simulated data and glioblastoma (GBM) transcriptomic data, which contain ground-truth labels of cell types on single cell gene expression data and clinical labels for patient bulk tissue gene expression data. Then we applied *DEGAS* to multiple Alzheimer's disease (AD) gene expression datasets from MSBB, Allen Institute, and Grubmann et al. [30] in which certain cell type changes (microglia and neuron) are largely known [31-37]. Finally, as an exploratory tool, we applied *DEGAS* to study multiple myeloma (MM) transcriptomic data, where the disease associated subtypes of cells are largely unknown.

MM stems from the proliferation of aberrant clonal plasma cells in the bone marrow that secrete monoclonal immunoglobulins and is the second most common blood cancer in the United States [38]. Patient level transcriptomic data for MM has been widely available for some time and has been used to identify subtypes MM with different prognoses [39]. However, only recently has scRNA-seq become available for MM [9, 40, 41] and few studies have identified the most high risk subtypes of cells [9]. Here we combined our newly generated MM scRNA-seq data from four local samples and bulk tissue data from the Multiple Myeloma Research Foundation CoMMpass study, and then applied *DEGAS* to infer clinical impressions for myeloma cell subtypes and successfully identified a *PHF19*$^{high}$ myeloma cell subgroup associated with a high risk of relapse.

## Results

### *DEGAS clinical impression framework*

In this study, we applied *DEGAS* to integrate and analyze scRNA-seq, bulk gene expression, and clinical data from simulated data as well as three different diseases: GBM, AD, and MM. The simulated, GBM, and AD datasets primarily served as validation to demonstrate the feasibility and universality of the *DEGAS* transfer learning approach since the ground truth of the simulated data is known, the correct GBM subtypes are known, and neuron loss with microglia gain in AD brains are also known. We then further expand our study to MM data, which serves as the discovery

dataset, since the myeloma cell subtypes and high-risk factors related to MM are not as well understood at the single-cell level. In the MM study, we applied *DEGAS* on patient data from the Multiple Myeloma Research Foundation CoMMpass study (MMRF) and scRNA-seq data from patients at Indiana University School of Medicine (IUSM). Our aim was to identify the cell subtypes using the impressions of relapse risk on the single cells. We then applied the results to two separate MM validation datasets, one of which contained plasma cells from normal bone marrow (NHIP), two MM precursor conditions - monoclonal gammopathy of undetermined significance (MGUS) and smoldering multiple myeloma (SMM), and MM. We tested if *DEGAS* assignment of relapse risk to cell subtypes were higher for more malignant conditions. An additional external validation dataset of patient level expression data with overall survival was used to evaluate whether the patient stratification learned by *DEGAS* was robust enough to be generalized to an external survival dataset.

***DEGAS correctly identifies high risk cell types and subtypes in simulated data***

To evaluate *DEGAS* in a controlled context, 5,000 single cells were generated with *Splatter [42]* (**Fig. 2A**) where 2,000 of the cells were held-out to generate simulated patients. Using this group of held-out cells, 600 simulated patients were generated by aggregating sets of 400 simulated cells (**Fig. 2B-D**). We conducted three simulation experiments, denoted Simulation 1, Simulation 2, and Simulation 3, where the single cells were aggregated in known proportions for each patient so that we could generate a "disease" patient group with different cellular composition than the "normal" patient group (**see Methods**). To highlight the utility of *DEGAS*, the experiments were: Simulation 1: cell type 1 is enriched in disease patients (**Fig. 2B**); Simulation 2: one subtype of cell type 4, *i.e.,* cell type 4 disease, is enriched in disease patients (**Fig. 2C**); and Simulation 3: both subtypes of cell type 4 are enriched in disease patients (**Fig. 2D**).

Please note that the optimal number of clusters for the simulated single cells would be determined to be four based on a standard scRNA-seq workflow (*i.e.,* tSNE followed by K-Medoids where optimal cluster number is selected based on average silhouette width) (**Fig. 2E**). This would cluster the cells into the four cell types while ignoring the two subtypes in the cell type 4 (**Fig. 2F**). As a result, deconvolution algorithms will not be able to detect the subtype level risk associations. Fortunately, cell type prioritization algorithms like *Augur* can detect these changes within cell types due to disease. However, for situations that do not have new cell type or missing cell type in the disease (Simulation 1), *Augur* cannot detect the association between cell type 1 and disease since there is no disease associated cell type change (**Fig. 2G**). *Augur* can detect the disease-associated cell type 4 in Simulation 2 (**Fig. 2H**). In simulation 3 where there is a mix of disease and normal subtypes for the cell type 4 in the disease group, *Augur* again has difficulty in identifying the cell type 4 disease association (**Fig. 2I**). In contrast to *Augur*, deconvolution can easily identify the correct cell type for Simulation 1 (**Fig. 2J**) and Simulation 3 (**Fig. 2L**) but not for simulation 2 (**Fig. 2K**). In comparison, *DEGAS* not only identified the correct cell type and subtypes in each experiment, it also correctly detected all of the simulated disease associations (**Fig. 2G-L**). Additionally, *DEGAS* had high precision-recall area under the curve (PR-AUC) predicting disease status of simulated patients (0.96-0.98) (**Table S1**) and almost perfectly predicted the cell type of simulated cells (~1.0) during cross-validation (**Table S2**). Since *DEGAS* directly assigns disease risk to cells, many of the problems with cell type level analyses can be avoided and the correct groups of cells can be identified by overlaying impressions of disease risk.

***DEGAS correctly mapped single cells to corresponding GBM subtypes***

We first demonstrate *DEGAS* in a straightforward case to show the performance of our framework using real data from GBM. We use single-cell data from Patel *et al*. [43], in which researchers

assigned four major GBM tumor subtypes (Proneural, Mesenchymal, Classical, and Neural) to the scRNA-seq data obtained from five GBM tumors. Of the five tumor samples, four had been labeled in the original publication with a single subtype based on the major proportion of cells assigned to each GBM subtype. For GBM bulk tumor tissue expression data, we obtained microarray data for 111 GBM patients from The Cancer Genome Atlas (TCGA), for which the same labels of GBM subtypes were also provided. As the simplest form of validation, we used these two datasets as input for the *DEGAS* model to test if it could re-identify the same GBM subtypes for both single cells and the TCGA GBM cohort simultaneously. The resulting *DEGAS* models also proved to be accurate with high PR-AUCs (0.79-0.97) when predicting each of the GBM subtypes in the TCGA patients during 10-fold cross-validation (**Table S3**). *DEGAS* correctly re-identified the same labels for all four tumors by overlaying GBM subtypes associations on each single cell, as indicated by the groups of cell subtypes with the highest association score determined by the median value (indicated with a red box) (**Fig. 3A-D**). For the fifth tumor sample, MGH31, it was labeled as a combination of multiple GBM subtypes in the original study, so we did not use it in our evaluation although *DEGAS* identified mesenchymal as its most associated GBM subtype (**Fig. 3E**). Additionally, these relationships can be visualized by plotting the single cells and overlaying the GBM subtype association. It is clear that MGH28 and MGH29, for instance, have a high association with the mesenchymal GBM subtype (**Fig. 3F, Fig. S1**).

***DEGAS identifies AD patients with increased microglia and reduced neuron populations***

Aside from GBM, AD also has well documented characteristics that can be used as a test bed for *DEGAS*. Specifically, there is a well-documented reduction in neurons [34-36] and increase in microglia [31-33, 37] in AD patients. Brain scRNA-seq data were obtained from the Allen Institute for Brain Science and bulk AD RNA-seq were retrieved from the Mount Sinai Brain Bank (MSBB) [44]. The *DEGAS* models were trained using neuron, oligodendrocyte, astrocyte, OPC, and microglia cell types. The brain samples were split into groups based on AD diagnosis status ($AD^+$ or $AD^-$). During 10-fold cross-validation, *DEGAS* models achieved high AD diagnosis PR-AUC (0.82) in MSBB patients **(Table S4**) and high cell type prediction PR-AUCs (>0.99) for single cells (**Table S5**).

From *DEGAS* results, we confirmed that in the single cell level, the AD association was negative in neurons as previously described [45], which is shown by the dark shade of neurons compared to other cell types (**Fig. 3G**, **Table 1**). In contrast, we observed positive AD association in microglia cells (**Fig. 3G**, **Table 1**). A strength of the *DEGAS* framework is that it can detect intra-cell type differences in disease risk. Within the astrocyte cell type, we identified an astrocyte subtype that had a positive association with AD (**Fig. 3G**) that corresponded to the *Astro L1 FGFR3 FOS* subtype from the Allen Institute brain cell atlas (i.e. *FOS* is a marker for disease associated astrocytes, or DAAs) [46] (**Fig. 3H**), and was enriched for *GFAP*, another marker for DAAs (**Fig. 3I**) [47]. When we performed *DEGAS* analysis on a separate dataset [30] with single cells from both AD and normal brains, we found that the major cell types from AD brains were significantly more associated with AD than their counterparts in normal brains as judged by median value (**Fig. 3J**). Additionally, the astrocytes in AD brains were highly positively associated with AD (AD association = 0.22, p-value = $7.89\times10^{-7}$) whereas the astrocytes in normal brains were negatively associated with AD (AD association = -0.06, p-value = $1.06\times10^{-2}$, **Fig. 3J**). Astrocytes from AD brains also expressed *GFAP* at greater levels than astrocytes from normal brains (t-test p-value < $2.20\times10^{-16}$). This supports our finding of an AD-associated *GFAP*+ subtype of astrocytes. In contrast, we saw that all microglia were associated with AD and did not see any subcluster of disease associated microglia (DAM) that was more associated with AD (**Fig. 3G**). This could be a function of the general large positive association of all microglia with AD that makes identifying DAM more difficult when all cell types are plotted together.

### *DEGAS identifies plasma cell subtypes in IUSM CD138+ scRNA-seq of MM*

In the MM study, unlike the previous two datasets, there were no predefined cell type labels, but *DEGAS* was still capable of analyzing such data and give clinical perspective to the clusters of cells in the MM scRNA-seq data. In order to cluster cells into groups, we first used Seurat [26], a commonly used scRNA-seq data analysis tool, to merge and cluster all the CD138+ bone marrow cells from four patients (two SMM and two MM) whose samples were collected at the IUSM. Using Seurat, five major clusters of cells were identified (**Fig. 4A**). These five clusters were used as the subtype labels in the *DEGAS* framework. We verified these cell clusters by clustering cells from each patient individually with Seurat and another scRNA-seq normalization tool *BERMUDA* (Batch Effect ReMoval Using Deep Autoencoders) [23] for all four patients. We found that the individual clustering results closely mirrored the Seurat-CCA clusters (**Fig. S2A-D**, **Table S6**) and that the subtype 2 was consistent across all MM patients using *BERMUDA* (**Fig. S2E**). For bulk tissue data from MMRF, the clinical outcomes of relapse free survival for 647 patients were used as the patient-level input to *DEGAS* and overlaid onto the CD138+ single cells from the four IUSM patients (**Fig. 4B**).

### *DEGAS patient stratification and cell type classification on MM*

A *DEGAS* model was trained on IUSM patient scRNA-seq data with subtype labels defined above and MMRF patients with bulk tissue data and relapse free survival information. The performance metrics were calculated via 10-fold cross-validation. It is worth noting that for PR-AUC, random no skill classifiers will achieve a performance equal to the percentage of the class of interest and in the case of uncommon classes like subtype 4, the random classifier performance will be close to zero (0.02). When predicting cellular subtype label in single cells, *DEGAS* was able to achieve a PR-AUC between 0.44-0.98 for all of the five CD138+ cellular subtypes identified in the above scRNA-seq data while the PR-AUC for subtype 2 reached 0.91 (**Table S7**). The receiver operating curve AUCs (ROC-AUCs) were between 0.90-0.98 for these five subtypes (**Table S7**). Due to class imbalance some of the subtypes did not perform as well as others based on PR-AUC but all of the PR-AUCs were substantially greater than a purely random model. Aside from correctly classifying the single cells, *DEGAS* was able to stratify the MMRF patients into high and low risk groups based on median relapse risk (p-value = $4.72\times10^{-10}$, **Fig. 4C**). We then applied the trained model on an external patient transcriptomic dataset from Zhan *et al*. [39] for validation. We demonstrated that the Cox proportional hazards portion for patient overall survival time of the *DEGAS* model was robust across datasets, and the impressions extracted from the *DEGAS* framework were capable of stratifying patients into low- and high-risk groups in the validation dataset (p-value = $1.12\times10^{-3}$, **Fig. 4D**).

### *DEGAS identifies CD138+ cellular subtypes that are associated with patient relapse*

The *DEGAS* model for the MM study transfers clinical impressions to single cells (i.e., single cells were directly assigned a relapse association score), as well as transfers cellular/molecular impressions to patients (i.e., patients are assigned subtype enrichment score). We found that the subtype 2 cells were the most important for prognosis (**Fig. 4B**). Specifically, the subtype 2 cells were associated with a shorter time to relapse (**Fig. 4E**, p-value < $2.2\times10^{-16}$). On an external validation scRNA-seq dataset from Ledergor *et al*. [40], the relapse association increased from NHIP (no disease) to SMM (**Fig. 5F** p-value = $1.50\times10^{-2}$) and MM (**Fig. 5F**, p-value = $1.70\times10^{-2}$), which is consistent with the order of precursor conditions for MM (NHIP → MGUS → SMM → MM). In addition, the enrichment of the subtype 2 cells increased from NHIP to near-MM stage SMM (**Fig. 4G**, p-value = $3.10\times10^{-2}$) and MM (**Fig. 4G**, Kruskal-Wallis p-value = $3.40\times10^{-2}$).

***MM prognostic subtypes have distinct gene signatures***

Differential gene expression analysis was performed between subtype 2 and all other subtypes (**Supplementary File 2**), and we found that subtype 2 had significantly up-regulated *PHF19* expression in all four of the patients (**Fig. 4H**). *PHF19* is a known marker for malignant disease in MM [48]. Besides *PHF19*, its associated markers such as *HELLS*, *EZH2*, *TYMS*, *ZWINT*, and *MKI67* were also significantly up-regulated in subtype 2. These results suggested the existence of an especially malignant CD138+/*PHF19*$^{high}$ subpopulation of plasma cells represented by the subtype 2 cluster. It is important to notice that the gene feature set that was used as input into *DEGAS* only contained the *HELLS* gene, which further highlights the ability of *DEGAS* to identify high risk cellular subtypes.

***DEGAS is robust to hyper-parameter choice***

To assess the robustness of *DEGAS*, we also analyzed how the hyper-parameter choices influence its results using a set of 100 randomly generated hyper-parameters with 10-fold cross-validation on each set of those 100 sets of hyper-parameters on the GBM datasets. The hyper-parameters that we evaluated include: the number of training steps, batch size for single cells, batch size for patients, number of hidden layer nodes, drop-out retention rate (the percentage of nodes randomly retained at the hidden layer), patient loss weight, MMD loss weight, and $L_2$-regularization weight. The detailed information about the range of hyper-parameters that were randomly sampled can be found in subsection titled *Evaluation of DEGAS robustness to hyper-parameters* in the **Methods** section while the default parameters used for all previous experiments can be found in the subsection titled *Transfer learning using DEGAS*. We discovered that among the eight hyper-parameters, the majority of them did not significantly affect the ROC-AUC for predicting GBM subtypes in TCGA GBM patients with the exception of three hyperparameters – namely the drop-out retention rate, number of hidden layer nodes, and $L_2$-regularization weight with p-value < 0.1 (**Fig. S3, Table 8**). Similarly, the majority of hyper-parameters did not significantly affect the correct assignment of subtype to GBM scRNA-seq tumor, except for a few exceptions in training steps, patient loss weight, and MMD loss weight with p-value < 0.1 (**Fig. S4**, **Table S9**). We therefore suggest users to keep default settings for at least patient loss weight, MMD loss weight, and $L_2$-regularization weight. The percentage of GBM subtype labels ranking in the top two predicted labels improves from 74% to 82% if the default parameters or greater values are used for patient loss weight, MMD loss weight, and $L_2$-regularization weight (**Fig. S5**).

## Discussion

In this work, we developed the transfer learning framework *DEGAS* to integrate scRNA-seq and patient-level transcriptomic data in order to infer the transferrable "impressions" between patient characteristics in single cells and cellular characteristics in patients. Using transfer learning, we trained a model with both scRNA-seq and patient bulk tissue gene expression data, then reduced the differences between the distributions of the representations for the two data types in the final hidden layer of our model via domain adaptation. This process allows information about patient outcome as well as cell types to be transferred between the two data types. We tested and validated the *DEGAS* framework on datasets from two diseases: GBM, which contained ground truth tumor subtype labels, and AD, which contained ground truth cell type-disease associations. For the GBM single cell patient cohort, each GBM tumor, from which scRNA-seq data was generated, had a GBM subtype label [43]. The *DEGAS* results showed that the majority of cells in each tumor were labeled with the same GBM subtype as previously defined in Patel *et al.* [43]. Specifically, *DEGAS* correctly mapped Proneural, Mesenchymal, Classical, and Neural GBM subtypes to single cells in four GBM tumor samples. This experiment also shows the broad applicability of the model since the single cells had no labels and the patient samples had multiclass labels. *DEGAS* is highly flexible and allows for different categories of output labels to be combined, which may include but are not

limited to classification labels, Cox proportional hazard, and even no labels. This allows for a wide variety of applications to adopt the *DEGAS* framework so that impressions are not limited to only one type of outcome.

The *DEGAS* analysis on AD data further validated our model by correctly identifying the decreased neuron and increased microglia proportions in AD patients. Aside from these known characteristics of AD pathology, we also identified a *GFAP*+ astrocyte subtype taken from normal human brain tissue that is associated with AD and is supported from AD mouse models [47]. We further validated this by finding that *GFAP* expression in Astrocytes was significantly increased in Astrocytes taken from AD patients and concluded that there may be an expansion of this Astrocyte subtype in AD. This is also a convincing example of the utility of *DEGAS* as it assigned disease association at the single cell level, allowing us to identify intra-cell type differences in disease risk that constitute disease-associated cells.

To explore disease with less understood cellular subtypes, we applied *DEGAS* to multiple MM datasets. The models were able to assign relapse free survival metrics to subtype populations of CD138+ cells identified by cell type clustering methods Seurat [26] and *BERMUDA* [23]. Among the identified subtypes of cells, subtype 2 was the most consistent between IUSM patients visualized by *BERMUDA* (**Fig. S2E**). Furthermore, we found that the subtype 2 cell population appeared to have a gradient of cells moving away from the main subtype 1 group, possibly associated with a certain degree of differentiation (**Fig. S2A-D**). We did experience a lower PR-AUC for subtype 4 than the other subtypes used during model training. However, this subtype was extremely uncommon in the samples and as a result the random PR-AUC would be close to zero making the PR-AUC of 0.44 well above random. Considering that subtype 4 was not found to be highly associated with relapse, the lower PR-AUC did not greatly affect our interpretation of the data, which mainly focused on subtype 1 and subtype 2. We believe that *DEGAS* could be improved for highly imbalanced data.

Upon further examination, we found evidence that the subtype 2 cells may represent a population of malignant plasma cells expressing high levels of *PHF19*. *PHF19* is known to play a role in hematopoietic stem cell state and differentiation [49-51] and is a marker for aggressive disease in MM [48]. Furthermore, knock down of *PHF19* has been shown to shift myeloma cells into a less proliferative state [48]. The subtype 2 cells express *SDC1* (also known as *CD138)* and showed significantly increased *PHF19* expression in comparison to the other subtypes. Since all of the IUSM MM cells in our study had already been FACS sorted for CD138+, it is possible we have identified a subpopulation of CD138+/$PHF19^{high}$ cells in MM tumors. This could prove a useful finding since currently the association between *PHF19* and tumor aggressiveness is at the patient level whereas our results imply that only a fraction of malignant plasma cells in a MM tumor actually overexpress *PHF19.*

This subtype could be targeted using precision immunotherapies that are not restricted to a single patient since the CD138+/$PHF19^{high}$ cells (i.e., subtype 2) were found to be present in multiple (3/4) patients. Of the three patients with detectable levels of subtype 2 in the CD138+ fraction, two patients (patient 2 and patient 4) had relapsed MM at time of biopsy and the other patient (patient 5) was SMM at biopsy and later progressed to MM. The other patient (patient 3) had little to no detectable subtype 2 cells in the CD138+ fraction and was SMM at time of biopsy and has not progressed to MM. These signs again seem to indicate a common cellular phenotype associated with relapse in MM.

Based on the validated results in a variety of disease data analyses, we find that DEGAS has broad applications in virtually all diseases with available patient-level and single cell level omic data. The tensorflow [52] machine learning code is integrated with a simple R package interface (https://github.com/tsteelejohnson91/DEGAS) which will facilitate researchers to manipulate scRNA-seq and bulk expression data on their own.

*DEGAS* is a powerful transfer learning tool for integrating different levels of omic data and identifying the latent molecular relationships between populations of cells and clinical outcomes, which we refer to as impressions. We validated the *DEGAS* framework on simulated data, GBM and AD by showing *DEGAS* models were capable of accurately predicting patient characteristics at single-cell level. We then leveraged this transfer learning approach on MM data and identified a CD138+/*PHF19*$^{high}$ subtype population in MM that was significantly associated with disease relapse. This subtype contains unique RNA profiles and gene correlations that could be both leveraged as a prognostic biomarker and possibly targeted directly to reduce the risk of relapse. We believe that *DEGAS* can be a powerful solution to overcome the challenge of integrating patient single-cell data with bulk tissue data so that researchers can identify populations of cells associated with an outcome of interest. Furthermore, *DEGAS* can accommodate flexible data types. This makes it a highly general framework that can be applied in multiple diseases and data types to identify cellular populations that are associated with prognosis or treatment response, or to identify specific patient groups with certain cell subtypes for personalized treatment.

## Materials and Methods

### *Experimental Design and Datasets*

In this study we analyzed simulated data and data from three different diseases, GBM, AD, and MM, to test the *DEGAS* framework and apply it for novel discoveries. The simulation, GBM, and AD experiments were primarily used as validation datasets since the ground truth is known. The simulated data were generated so that cell types are directly related to disease status in patients. For GBM data, we used scRNA-seq data for five tumors from Patel *et al*. [43] and microarray data for the GBM TCGA cohort [53]. For AD data, we used human scRNA-seq from Allen Institute Cell Types Database (https://celltypes.brain-map.org/) and AD patient RNA-seq from the Mount Sinai/JJ Peters VA Medical Center Brain Bank (MSBB) study [44].

We further expanded our inquiry into MM, which served as a discovery study. Since the plasma cell subtypes are less understood in relation to clinical outcomes, we aimed to identify subtypes of plasma cells associated with worse prognosis. We first utilized 647 CD138$^{+}$-enriched bone marrow patient samples from the Multiple Myeloma Research Foundation CoMMpass study (MMRF). These data were generated as part of the Multiple Myeloma Research Foundation Personalized Medicine Initiatives (https://research.themmrf.org). The dataset consisted of tumor tissue RNA-seq data and corresponding clinical information including relapse-free survival time and survival status. Relapse-free survival was defined as the time taken for a patient to relapse after treatment of the initial tumor or the time of death if relapse was not reached. The demographic information of the MMRF patients are shown in **Table 2**. The first scRNA-Seq data used in this study were generated at Indiana University School of Medicine (IUSM) using samples consisting of CD138$^{+}$ plasma cells purified from bone marrow from four MM patients. The low number of patients was a good test case considering most scRNA-seq experiments have few patients. The single cells were sequenced using 10x Genomics and Illumina NovaSeq6000 sequencer. *CellRanger 2.1.0* (http://support.10xgenomics.com/) was utilized to process the raw sequence data. Briefly, *CellRanger* used *bcl2fastq* (https://support.illumina.com/) to demultiplex raw base sequence calls generated from the sequencer into sample-specific FASTQ files. The FASTQ files were then aligned to the human reference genome GRCh38 with RNA-seq aligner

*STAR*. The aligned reads were traced back to individual cells and the gene expression level of individual genes were quantified based on the number of UMIs (unique molecular indices) detected in each cell. The filtered gene-cell barcode matrices generated by *CellRanger* were used for further analysis. A second publicly available scRNA-seq dataset was used for validation, which consisted of NHIP (normal control), MGUS, SMM, and MM patients [40]. A second bulk tissue dataset was used for validating the proportional hazards modeling. This dataset consisted of bulk expression profiling by microarray of CD138+ plasma cells with overall survival information for 559 MM patients [39]. The detailed information of the four datasets is shown in **Table 3**.

***Transfer learning using DEGAS***

Several types of labels including Cox proportional hazards, patient classification, and cell type classification, along with maximum mean discrepancy (MMD), a technique used to match distributions across different sets of data [20], were combined to create the multitask transfer learning framework *DEGAS*.

The first step was to find a set of gene expression features that were both informative of cell type and of patient clinical outcome (e.g., recurrence). The intersection of high variance genes found in the scRNA-seq and bulk expression data of patient samples are used for further analysis. The definition of this gene set is up to the user but *Seurat-CCA*, LASSO selection, and even statistical tests such as Student t-test and F-test can be used to define the gene set. Since these features are the same between patients and single cells, the patients and cells share the same input layer. This makes it possible to predict proportional hazard and cell type regardless of the input sample type (patient or single cell data).

All experiments in this manuscript use a bootstrap aggregated 3-layer DenseNet-based implementation of *DEGAS*, but the simplest form of the *DEGAS* framework is a single layer network. In our description of the overall architecture below (shown in Fig. 1), we used a single layer network for the purpose of simplicity. The following **Eq. 1** can nevertheless be extrapolated to multiple layers and architectures, some of which we have already included in our open-source software package. First, a hidden layer was used to transform the genes into a lower dimension using a sigmoid activation function (**Eq. 1**). Where $X$ represents an input expression matrix, $\theta_{Hidden}$ represents the hidden layer weights, and $b_{Hidden}$ represents the hidden layer bias.

$$f_{Hidden}(X) = sigmoid(X^{\mathrm{T}}\theta_{Hidden} + b_{Hidden}) \qquad \textbf{Eq. 1}$$

Next, output layers were added for both the patient output and for the single cell output. For the single cells, there could be classification output or no output. No output means there is no known labels for the single cells to match. Similarly, patients could have Cox proportional hazard output, classification output, or no output (implying no known labels for patients).

The Cox proportional hazards estimates consisted of a linear transformation to a single output followed by a sigmoid activation function (**Eq. 2**):

$$f_{Cox}(X) = sigmoid(f_{Hidden}(X)^{\mathrm{T}}\theta_{Cox} + b_{Cox}), \qquad \textbf{Eq. 2}$$

where the variable $X$ represents an input expression matrix, $\theta_{Cox}$ represents the Cox proportional hazard layer weights [54], and $b_{Cox}$ represents the Cox proportional hazard layer bias. The classification output consisted of a transformation to the same number of outputs as the number of labels, *i.e.*, patient subtypes, cellular subtypes, using a softmax activation function (**Eq. 3**).

$$f_{Class}(X) = softmax(f_{Hidden}(X)^{\mathrm{T}}\theta_{Class} + b_{Class}), \qquad \textbf{Eq. 3}$$

where $\theta_{Class}$ represents the classification layer weights, and $b_{Class}$ represents the classification layer bias.

To train the *DEGAS* model, we need to compute three types of loss functions for the Cox proportional hazards output, classification output, and MMD [20] respectively. The Cox proportional hazards loss [54] was calculated only for the patient expression data ($X_{Pat}$) using the followup period ($C$), and event status ($t$) (**Eq. 4**). Similarly, the patient classification loss was only calculated for the patient data ($X_{Pat}$) using the patient labels ($Y_{Pat}$). Alternatively, the cellular classification loss was only calculated for the single cell expression data ($X_{Cell}$) and true subtype label ($Y_{Cell}$) (**Eq. 5**). The MMD loss was calculated between the patient expression data ($X_{Pat}$) and the single cell expression data ($X_{Cell}$) (**Eq. 6**), which is the key for mapping the distributions of the data representations between the single-cell and patient bulk tissue data.

$$Loss_{Cox} = \sum_{C(i)=1}\left(f_{Cox}(X_{Pat})_i - \sum_{t_j \geq t_i}(\exp\left(f_{Cox}(X_{Pat})_j\right)\right) \qquad \textbf{Eq. 4}$$

$$Loss_{Class} = \frac{1}{n}\sum_{i=1}^{n}\left(\sum(Y_{type,i} - f_{Class}\left(X_{type}\right)_i)\right) \; where\ type \in \{Pat, Cell\} \qquad \textbf{Eq. 5}$$

$$Loss_{MMD} = MMD(X_{Cell}, X_{Pat}) \qquad \textbf{Eq. 6}$$

Besides the three losses, we also add a $L_2$-regularization loss term to constrain for the complexity of the model. The overall loss function was the weighted sum of the four types of loss using the hyper-parameters $\lambda_0$ (single cell loss function), $\lambda_1$ (patient loss function), $\lambda_2$ (MMD loss), and $\lambda_3$ (regularization loss), so that the importance of each loss term and regularization term could be adjusted (**Eq. 7**):

$$Loss_{ClassCox} = \lambda_0 Loss_{Class} + \lambda_1 Loss_{Cox} + \lambda_2 Loss_{MMD} + \lambda_3 \|\theta\|_2^2. \qquad \textbf{Eq. 7}$$

To address more diverse scenarios, we can also adapt **Eq. 7** for two classification outputs (**Eq. 8**), a single classification output without patient outcome (**Eq. 9**), a single classification output without cell type label (**Eq. 10**), or a single Cox output without cell type label (**Eq. 11**):

$$Loss_{ClassClass} = \lambda_0 Loss_{Class} + \lambda_1 Loss_{Class} + \lambda_2 Loss_{MMD} + \lambda_3 \|\theta\|_2^2, \qquad \textbf{Eq. 8}$$

$$Loss_{ClassBlank} = \lambda_0 Loss_{Class} + \lambda_2 Loss_{MMD} + \lambda_3 \|\theta\|_2^2, \qquad \textbf{Eq. 9}$$

$$Loss_{BlankClass} = \lambda_1 Loss_{Class} + \lambda_2 Loss_{MMD} + \lambda_3 \|\theta\|_2^2, \qquad \textbf{Eq. 10}$$

$$Loss_{BlankCox} = \lambda_1 Loss_{Cox} + \lambda_2 Loss_{MMD} + \lambda_3 \|\theta\|_2^2. \qquad \textbf{Eq. 11}$$

In summary, a common hidden layer was used to merge the single cells and patients data. Next, an output layer was added to predict the proportional hazards or classes of the patient samples [54]. The loss function for the proportional hazards prediction or patient classification was back-propagated across both layers for each patient. The single cells also had an output layer consisting of a softmax output to predict the cellular subtype of each cell. Error was back-propagated across both layers from the label output for each cell. Finally, a model was learned that can model both the single cells and the patients. To perform this task, we utilized the MMD method [20] to reduce the differences between patients and cells in a low dimensional representation. Both single cell and patient bulk tissue data were combined into a single group such that the MMD loss was minimized between patient bulk tissue data and single cells data from multiple patients. Because there are many different combinations of these outputs, *i.e.*, single cell output followed by patient output, we implemented ClassCox, ClassClass, ClassBlank, BlankClass, and BlankCox models based on equations (7)-(11) in the current version but intend to provide more options in the future.

To keep the analyses consistent, we used the same network architecture and hyperparameters throughout all of the experiments. Specifically, we used a 3-layer DenseNet architecture bootstrap aggregated five times such that **Eq. 1** would consist of a DenseNet instead of a single layer feedforward network and five such models were trained. The same set of hyper-parameters were used in all of the experiments in this study, except for the robustness to hyper-parameters experiment, where they were intentionally altered to test the influences on the output results. These are considered the default hyper-parameters in the *DEGAS* package but can be changed. They are:

training steps 2000, single cell batch size 200, patient batch size 50, hidden layer nodes 50, drop-out retention rate 50%, single cell loss weight ($\lambda_0$) 2, patient loss weight ($\lambda_1$) 3, MMD loss weight ($\lambda_2$) 3, and $L_2$-regularization weight ($\lambda_3$) 3.

***Validating DEGAS using Simulated single cell data***

First we generated 5,000 single cells in four cell types where the cell type 4 had two subtypes (cell type 4 disease and cell type 4 normal). Each of these five groups described above contains 1,000 cells. We split randomly these cells into 2 parts with 2,000 cells used for patient bulk tissue data generation and 3,000 cells to use directly as single cell data. The 2,000 single cells used to generate 600 patients across three different experiments (designated as Simulation 1, 2, and 3) where in Simulation 1 the cell type 1 is associated with disease, in simulation 2 only the cell type 4 disease is associated with disease, and in simulation 3 the entire cell type 4 is associated with disease. Each patient bulk tissue data was generated by randomly combining 400 single cells using the proportions in **Table 4**.

We then performed 10-fold cross validation by training the *DEGAS* ClassClass models using cell type and the disease status as outcomes. We evaluated the model capacity for mapping patient labels on patients and cell type labels on single cells using PR-AUC and ROC-AUC. We then recapitulated the known cell type associations in each simulation by overlaying disease association onto the simulated cells. As a comparison, we also deconvoluted the patients using the 4 cell types using least squares. Deconvolution should be able to correctly identify the cells of interest in simulation 1 and simulation 3. In contrast, cell type prioritization using *Augur* [17] should be able to correctly identify the disease associated cell types in simulation 2. In the simulation 1 Augur experiment, cell type 1, cell type 2, cell type 3, and cell type 4 normal were randomly assigned to the disease or normal groups. In the simulation 2 *Augur* experiment, cell type 1, cell type 2, and cell type 3 were randomly assigned to disease or normal groups. The cell type 4 disease cells were all assigned to the disease group and the cell type 4 normal cells were all assigned to the normal group. In the simulation 3 Augur experiment, cell type 1, cell type 2, and cell type 3 were randomly assigned to disease and normal groups. The cell type 4 cells assigned to the disease group consisted of 60% cell type 4 normal and 40% cell type 4 disease and cell type 4 cells assigned to the normal group consisted of 100% cell type 4 normal. These proportions match those in the simulation 3 patients used by *DEGAS*.

***Validating DEGAS using GBM data***

The scRNA-seq data from the Patel *et al*. study [43] were downloaded from NCBI Gene Expression Omnibus (GSE57872). The single cell expression values were previously normalized to TPM containing 5,948 genes with *mean*($\log_2$(TMP))>4.5 retained in the data table. The top 20% variance genes were retained for training. These values were converted to z-scores then standardized to a range of [0,1] for each sample. The TCGA GBM microarray expression data was downloaded from Firebrowse (http://firebrowse.org/). Microarray data were used since it contains more patient samples for training with GBM subtype information than RNA-seq data. Likewise, the top 20% variance genes were retained for training and these expression values were converted to z-scores then standardized to a range of [0,1] for each sample. The GBM subtype labels for the TCGA patients were downloaded from Verhaak *et al*. [55]. The intersection of genes between single cells and patients were used for the final model training. Since subtype labels were only available for the GBM patient samples, we trained a BlankClass *DEGAS* model (**Eq. 10**). This model minimizes the MMD loss between single cells and patients while minimizing the classification loss only in GBM patients. We split the dataset into 10 groups and performed 10-fold cross-validation by leaving out a single patient group during training. After cross-validation, we converted the [0,1] *DEGAS* output to an association [-1,1] using the *DEGAS toCorrCoeff* function. These association scores were

overlaid on the GBM single cells and now referred to as GBM subtype association scores because GBM subtype from patients is overlaid on single cells. We plotted these association scores stratified by GBM subtype for each tumor individually. We then compared the proportions of these cell types to the previously defined GBM types from the original publication were marked with red boxes. We also visualized the GBM subtypes association in single cells by calculating a low dimensional representation using *tSNE* and overlaying the *kNN* smoothed GBM subtype associations. To make the scatter plots of cells and patients more informative, *kNN* smoothing was used by averaging each point's GBM subtype association value with its five nearest neighbors in *tSNE*. The model performance was shown with the PR-AUC and ROC-AUC for each of the GBM subtype labels in the TCGA patients from cross-validation.

***Validating DEGAS and exploration using AD data***

For AD datasets, we were primarily interested in identifying known relationships between cell types and AD diagnosis. For these reasons, we downloaded all of the adult Human scRNA-seq data from the Allen Brain Institute. Only inhibitory neurons, excitatory neurons, oligodendrocytes, astrocytes, microglia, and oligodendrocyte progenitor cells (OPCs) were retained in the analysis due to the extremely low sample sizes in the remaining cell types. The inhibitory and excitatory neuron groups were merged into a single neuron group. These data were then $\log_2$ transformed, converted to sample-wise z-scores, and then standardized to [0,1] by each sample. Only the top 50 up-regulated DEGs for each cell type (calculated by *Seurat*) were retained in the single cell data. The labels for the single cells consisted of the major cell types listed above. The AD brain data was downloaded from Mount Sinai/JJ Peters VA Medical Center Brain Bank (https://www.synapse.org/#!Synapse:syn3157743). Each of the RNA-seq samples were either from an AD patient or a normal brain sample. The binary outcomes of AD case or normal were used as the label for the model. As in the other experiments, the RNA-seq values were $\log_2$ transformed, converted to sample-wise z-scores, and standardized to [0,1] for each sample. The top 50% variance genes were retained for training to keep the feature set larger. The intersection of the patient genes and single cell genes were using to train the final model. Using the cell type classification for each single cell and the AD/normal classification for each MSBB patient we were able to train a *DEGAS* ClassClass model (**Eq. 8**). The performance was evaluated using 10-fold cross-validation by leaving out each group during training once. As in the GBM experiments, we converted the DEGAS output to an association using the *DEGAS toCorrCoeff* function for each single cell so that each single cell now had an AD association. Correlation analysis was performed on AD association scores for different cells with each cell type by taking the median score and calculating the p-value by treating it as a correlation. In addition, patients were plotted overlaid with *kNN* smoothed cell type enrichment and single cells were plotted overlaid with kNN smoothed AD association (**Fig. 3**). Furthermore, to evaluate *DEGAS* performance, PR-AUC and ROC-AUC were computed for the single cells during cross-validation for each cell type in the single cell data. Similarly, AD diagnosis PR-AUC and ROC-AUC were computed from the MSBB patient RNA-seq (**Fig. S6**).

***Preprocessing of IUSM MM single cell data***

The scRNA-seq data generated at IUSM were first combined into a dataset using *Seurat-CCA* [26]. This initial dataset integration allowed conserved subtypes of cells to be identified across datasets. All four patient dataset counts were loaded into a *Seurat* object. They were normalized, scaled, biased cells removed, and high variance genes identified following the Seurat online vignette. Using the union of high variance genes, multi-canonical correlation analysis was run across all four datasets, the subspaces were aligned across patients, the aligned single cells were plotted with *tSNE* [56], and clusters of cells were identified. The raw expression values for the high variance genes identified by *Seurat* were $\log_2$ transformed, converted to z-scores, and then scaled to [0,1].

Furthermore, each IUSM scRNA-seq patient was individually clustered using *Seurat* to check the replicability of the clusters and were plotted with *UMAP* [57]. We used Rand, Fowlkes and Mallows's index (FM), and Jaccard index (JI) to measure the cluster consistency between single patient clustering experiments and the merged all-patient clustering results. The four single patient clustering results, one for each IUSM scRNA-seq patient, were used as input into *BERMUDA* [23] to visualize and evaluate the original Seurat clustering.

***Preprocessing of MMRF patient data***

MMRF patients with bulk tissue RNA-seq and clinical data were used in MM analysis. We used relapse-free survival (RFS), i.e., the time to first relapse or death. TPM values for the MMRF patient gene expression data and the RFS survival data were used as the input for *DEGAS*, these values were $\log_2$ transformed, converted to z-scores, and scaled to [0,1]. The union of the features identified by Seurat in the single cell data and the features selected in the MMRF patient data were used as the final feature set. The features retained in the MMRF data were identified by fitting an elastic-net Cox model [58] to the TPM values based on the RFS.

***Evaluate DEGAS performance on MM datasets***

PR-AUC and AUC were calculated for each of the output labels for the single cells and for patient labels if a classification output was used for the patient data. Cox proportional hazard output was used on patients, a log-rank test was calculated for each patient so that the hazard ratio and p-value could be evaluated based on patient stratification by median proportional hazard. Additionally, the same models were used to predict risk in the GSE2658 dataset which had information on OS. The output for each GSE2658 sample averaged across all 10 *DEGAS* models and stratified by median risk to show the robustness of the cox output across datasets.

***Identification of CD138+ cell types associated with MM prognosis***

The single cells from MM patients can be assigned proportional hazards based on the MMRF Cox output of the model. Each single cell in the validation set was assigned relapse risk by feeding those samples through the Cox output layer. In this way, we can infer the association with relapse risk of specific cell types as well as the cell type enrichment contained in each MMRF sample. Since the Cox output is a proportional hazard, we centered the outputs to zero for each step of cross validation to produce a RFS association using the *DEGAS centerFunc*. We plotted these relationships and conducted Student's t-tests on the subtype vs. RFS association in IUSM single cells, RFS association vs. MM malignancy from Ledergor *et al*., and subtype 2 enrichment vs. MM malignancy from Ledergor *et al* [40].

***Analysis of differential gene expression in prognostic cell types***

Student's t-tests were calculated cell subtype 1 vs all cell subtypes and cell subtype 2 vs. all cell subtypes using the batch corrected gene expression values from *Seurat*. These values were stored in (**Supplementary File 1** and **Supplementary File 2**) respectively. For the marker set of *PHF19*, *HELLS*, *EZH2*, *TYMS*, *ZWINT*, and *MKI67* we performed Student's t-tests for each patient individually.

***Evaluation of DEGAS robustness to hyper-parameters in GBM***

Using the GBM dataset, we evaluated the robustness of DEGAS model outputs to hyper-parameters by repeating 10-fold cross-validation 100 times with randomly generated hyper-parameters following a uniform distribution. The range of hyper-parameters used in training consisted of training steps 1,000-3,000, single cell batch size 100-300, patient batch size 20-100, hidden features 10-100, drop-out retention rate 0.1-0.9, Cell loss weight ($\lambda_0$) held at constant 2, Patient loss weight

($\lambda_1$) 0.2-5, MMD loss weight ($\lambda_2$) 0.2-5, $L_2$ regularization weight ($\lambda_3$) 0.2-5. The output for each of the 100 hyper-parameters was quantile normalized.

Using these outputs, we performed two tests. One was to evaluate the loss in performance based on changing the hyper-parameters where performance was measured with AUC among the TCGA GBM patients labeled by patient GBM subtype (Mesenchymal, Classical, Proneural, Neural). In this test, we calculated the spearman correlation and plotted the scatter plot between the AUC of each of the four GBM subtype labels and the hyper-parameters used.

Next, we evaluated whether or not the correct GBM subtype labels (Mesenchymal, Classical, Proneural, Neural) could be recapitulated in the GBM scRNA-seq tumors that had known GBM subtypes (MGH26: Proneural, MGH28: Mesenchymal, MGH29: Mesenchymal, MGH30: Classical). To do this for each tumor (MGH26, MGH28, MGH29, MGH30), the rank of the correct label was calculated by calculating the mean of each GBM subtype association across all of the cells in that tumor. This resulted in each of the 100 random hyper-parameters having a rank for each GBM subtype for each of the GBM scRNA-seq tumors (4 highest ranked, 1 lowest ranked). Ideally all GBM scRNA-seq tumors would have a rank of 4 indicating the correct GBM subtype was ranked the highest regardless of hyper-parameters. Similarly, we also calculated the Spearman correlation and plotted the scatter plot between correct label rank and the hyper-parameters used.

## References


1. Lahnemann, D., et al., *Eleven grand challenges in single-cell data science.* Genome Biol, 2020. **21**(1): p. 31.
2. Ma, A., et al., *Integrative Methods and Practical Challenges for Single-Cell Multi-omics.* Trends Biotechnol, 2020. **38**(9): p. 1007-1022.
3. Kiselev, V.Y., A. Yiu, and M. Hemberg, *scmap: projection of single-cell RNA-seq data across data sets.* Nat Methods, 2018. **15**(5): p. 359-362.
4. Cao, Y., et al., *scRNASeqDB: A Database for RNA-Seq Based Gene Expression Profiles in Human Single Cells.* Genes (Basel), 2017. **8**(12).
5. Abugessaisa, I., et al., *SCPortalen: human and mouse single-cell centric database.* Nucleic Acids Res, 2018. **46**(D1): p. D781-D787.
6. Stuart, T., et al., *Comprehensive integration of single-cell data.* Cell, 2019. **177**(7): p. 1888-1902. e21.
7. Gawel, D.R., et al., *A validated single-cell-based strategy to identify diagnostic and therapeutic targets in complex diseases.* Genome Med, 2019. **11**(1): p. 47.
8. Chen, S., et al., *Single-cell analysis reveals transcriptomic remodellings in distinct cell types that contribute to human prostate cancer progression.* Nature Cell Biology, 2021. **23**(1): p. 87-98.
9. Jang, J.S., et al., *Molecular signatures of multiple myeloma progression through single cell RNA-Seq.* Blood cancer journal, 2019. **9**(1): p. 1-10.
10. Maynard, A., et al., *Therapy-induced evolution of human lung cancer revealed by single-cell RNA sequencing.* Cell, 2020. **182**(5): p. 1232-1251. e22.
11. Cobos, F.A., et al., *Benchmarking of cell type deconvolution pipelines for transcriptomics data.* Nature communications, 2020. **11**(1): p. 1-14.
12. Johnson, T.S., et al., *Combinatorial analyses reveal cellular composition changes have different impacts on transcriptomic changes of cell type specific genes in Alzheimer's Disease.* Sci Rep, 2021. **11**(1): p. 353.
13. Johnson, T.S., et al., *Spatial cell type composition in normal and Alzheimers human brains is revealed using integrated mouse and human single cell RNA sequencing.* Sci Rep, 2020. **10**(1): p. 18014.
14. Jung, S.-H. and S.-C. Chow, *On sample size calculation for comparing survival curves under general hypothesis testing.* Journal of biopharmaceutical statistics, 2012. **22**(3): p. 485-495.
15. Mathys, H., et al., *Single-cell transcriptomic analysis of Alzheimer's disease.* Nature, 2019. **570**(7761): p. 332-337.
16. Rossi, M.A., et al., *Obesity remodels activity and transcriptional state of a lateral hypothalamic brake on feeding.* Science, 2019. **364**(6447): p. 1271-1274.
17. Skinnider, M.A., et al., *Cell type prioritization in single-cell data.* bioRxiv, 2019: p. 2019.12.20.884916.
18. Kouw, W.M. and M. Loog, *A Review of Domain Adaptation without Target Labels.* IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021. **43**(3): p. 766-785.
19. Hardoon, D.R., S. Szedmak, and J. Shawe-Taylor, *Canonical correlation analysis: An overview with application to learning methods.* Neural computation, 2004. **16**(12): p. 2639-2664.
20. Gretton, A., et al., *A kernel two-sample test.* Journal of Machine Learning Research, 2012. **13**(Mar): p. 723-773.
21. Andrew, G., et al. *Deep canonical correlation analysis*. in *International conference on machine learning*. 2013. PMLR.
22. Zhang, F., Y. Wu, and W. Tian, *A novel approach to remove the batch effect of single-cell data.* Cell discovery, 2019. **5**(1): p. 1-4.

23. Wang, T., et al., *BERMUDA: a novel deep transfer learning method for single-cell RNA sequencing batch correction reveals hidden high-resolution cellular subtypes.* Genome Biol, 2019. **20**(1): p. 165.
24. Tran, H.T.N., et al., *A benchmark of batch-effect correction methods for single-cell RNA sequencing data.* Genome biology, 2020. **21**(1): p. 1-32.
25. Johnson, T.S., et al., *LAmbDA: label ambiguous domain adaptation dataset integration reduces batch effects and improves subtype detection.* Bioinformatics, 2019. **35**(22): p. 4696-4706.
26. Butler, A., et al., *Integrating single-cell transcriptomic data across different conditions, technologies, and species.* Nat Biotechnol, 2018. **36**(5): p. 411-420.
27. Aran, D., et al., *Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage.* Nature immunology, 2019. **20**(2): p. 163-172.
28. Araujo, T., et al., *Classification of breast cancer histology images using Convolutional Neural Networks.* PLoS One, 2017. **12**(6): p. e0177544.
29. Bardou, D., K. Zhang, and S.M. Ahmad, *Classification of Breast Cancer Based on Histology Images Using Convolutional Neural Networks.* IEEE Access, 2018. **6**: p. 24680-24693.
30. Grubman, A., et al., *A single-cell atlas of entorhinal cortex from individuals with Alzheimer's disease reveals cell-type-specific gene expression regulation.* Nature neuroscience, 2019. **22**(12): p. 2087-2097.
31. Holtman, I.R., et al., *Induction of a common microglia gene expression signature by aging and neurodegenerative conditions: a co-expression meta-analysis.* Acta Neuropathol Commun, 2015. **3**: p. 31.
32. Hemonnot, A.L., et al., *Microglia in Alzheimer Disease: Well-Known Targets and New Opportunities.* Front Aging Neurosci, 2019. **11**: p. 233.
33. Glass, C.K., et al., *Mechanisms underlying inflammation in neurodegeneration.* Cell, 2010. **140**(6): p. 918-34.
34. Donev, R., et al., *Neuronal death in Alzheimer's disease and therapeutic opportunities.* J Cell Mol Med, 2009. **13**(11-12): p. 4329-48.
35. DeKosky, S.T. and S.W. Scheff, *Synapse loss in frontal cortex biopsies in Alzheimer's disease: correlation with cognitive severity.* Ann Neurol, 1990. **27**(5): p. 457-64.
36. de Wilde, M.C., et al., *Meta-analysis of synaptic pathology in Alzheimer's disease reveals selective molecular vesicular machinery vulnerability.* Alzheimers Dement, 2016. **12**(6): p. 633-44.
37. Akiyama, H., *Inflammatory response in Alzheimer's disease.* Tohoku J Exp Med, 1994. **174**(3): p. 295-303.
38. Institute, N.C., *Cancer Statistics*, N.C. Institute, Editor. 2019: Cancer.gov.
39. Zhan, F., et al., *The molecular classification of multiple myeloma.* Blood, 2006. **108**(6): p. 2020-8.
40. Ledergor, G., et al., *Single cell dissection of plasma cell heterogeneity in symptomatic and asymptomatic myeloma.* Nat Med, 2018. **24**(12): p. 1867-1876.
41. Cohen, Y.C., et al., *Identification of resistance pathways and therapeutic targets in relapsed multiple myeloma patients through single-cell sequencing.* Nature medicine, 2021: p. 1-13.
42. Zappia, L., B. Phipson, and A. Oshlack, *Splatter: simulation of single-cell RNA sequencing data.* Genome Biol, 2017. **18**(1): p. 174.
43. Patel, A.P., et al., *Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma.* Science, 2014. **344**(6190): p. 1396-401.
44. Wang, M., et al., *The Mount Sinai cohort of large-scale genomic, transcriptomic and proteomic data in Alzheimer's disease.* Sci Data, 2018. **5**: p. 180185.

45. Fu, H., et al., *Tau Pathology Induces Excitatory Neuron Loss, Grid Cell Dysfunction, and Spatial Memory Deficits Reminiscent of Early Alzheimer's Disease.* Neuron, 2017. **93**(3): p. 533-541 e5.
46. Xu, J., et al., *Multimodal single-cell/nucleus RNA sequencing data analysis uncovers molecular networks between disease-associated microglia and astrocytes with implications for drug repurposing in Alzheimer's disease.* Genome research, 2021: p. gr. 272484.120.
47. Habib, N., et al., *Disease-associated astrocytes in Alzheimer's disease and aging.* Nature Neuroscience, 2020. **23**(6): p. 701-706.
48. Mason, M.J., et al., *Multiple Myeloma DREAM Challenge reveals epigenetic regulator PHF19 as marker of aggressive disease.* Leukemia, 2020. **34**(7): p. 1866-1874.
49. Bagger, F.O., S. Kinalis, and N. Rapin, *BloodSpot: a database of healthy and malignant haematopoiesis updated with purified and single cell mRNA sequencing profiles.* Nucleic acids research, 2019. **47**(D1): p. D881-D885.
50. Lara-Astiaso, D., et al., *Chromatin state dynamics during blood formation.* science, 2014. **345**(6199): p. 943-949.
51. Vizán, P., et al., *The Polycomb-associated factor PHF19 controls hematopoietic stem cell state and differentiation.* Science advances, 2020. **6**(32): p. eabb2745.
52. Abadi, M., et al. *Tensorflow: A system for large-scale machine learning*. in *12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16)*. 2016.
53. Cancer Genome Atlas Research, N., *Comprehensive genomic characterization defines human glioblastoma genes and core pathways.* Nature, 2008. **455**(7216): p. 1061-8.
54. Ching, T., X. Zhu, and L.X. Garmire, *Cox-nnet: An artificial neural network method for prognosis prediction of high-throughput omics data.* PLoS Comput Biol, 2018. **14**(4): p. e1006076.
55. Verhaak, R.G., et al., *Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1.* Cancer Cell, 2010. **17**(1): p. 98-110.
56. Maaten, L.v.d. and G. Hinton, *Visualizing data using t-SNE.* Journal of machine learning research, 2008. **9**(Nov): p. 2579-2605.
57. Becht, E., et al., *Dimensionality reduction for visualizing single-cell data using UMAP.* Nat Biotechnol, 2018.
58. Friedman, J., T. Hastie, and R. Tibshirani, *Regularization Paths for Generalized Linear Models via Coordinate Descent.* J Stat Softw, 2010. **33**(1): p. 1-22.

## Acknowledgments

**Acknowledgements:** We thank the Center for Computational Biology and Bioinformatics for the computational resources and work space to complete the research. We also thank the MMRF for the data generated as part of the Multiple Myeloma Research Foundation Personalized Medicine Initiatives (https://research.themmrf.org and www.themmrf.org), the Allen Institute for Brain Science for the data generated as part of their cell types database, and Mount Sinai/JJ Peters VA Medical Center for the data generated as a part of their brain bank.

**Funding:** National Institutes of Health NLM-NRSA Fellowship F31LM013056 to TSJ and The Ohio State University (Columbus, OH) and departmental start-up funding from the Indiana University School of Medicine (Indianapolis, IN) to KH and TSJ.

**Author contributions:** TSJ, CYY, JZ, and KH conceived and designed the project. TSJ, CYY, SX performed the analyses. TSJ and ZH designed the software package. TSJ, CYY, XH, SX, CD, MA, YW, CB, YZ, YL, JZ, BW, and KH interpreted the results. ZH, TW, WS, YW, and CB provided technical guidance. TSJ, CYY, JZ, and KH wrote the manuscript. JZ and KH supervised the project.

**Competing interests:** The authors declare that this research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

**Data and materials availability:** The DEGAS R package is freely available on GitHub (https://github.com/tsteelejohnson91/DEGAS). The MM scRNA-seq data generated at IUSM will be deposited into GEO upon publication and will be made available to reviewers upon request. All other data are publicly available.

## Figures and Tables

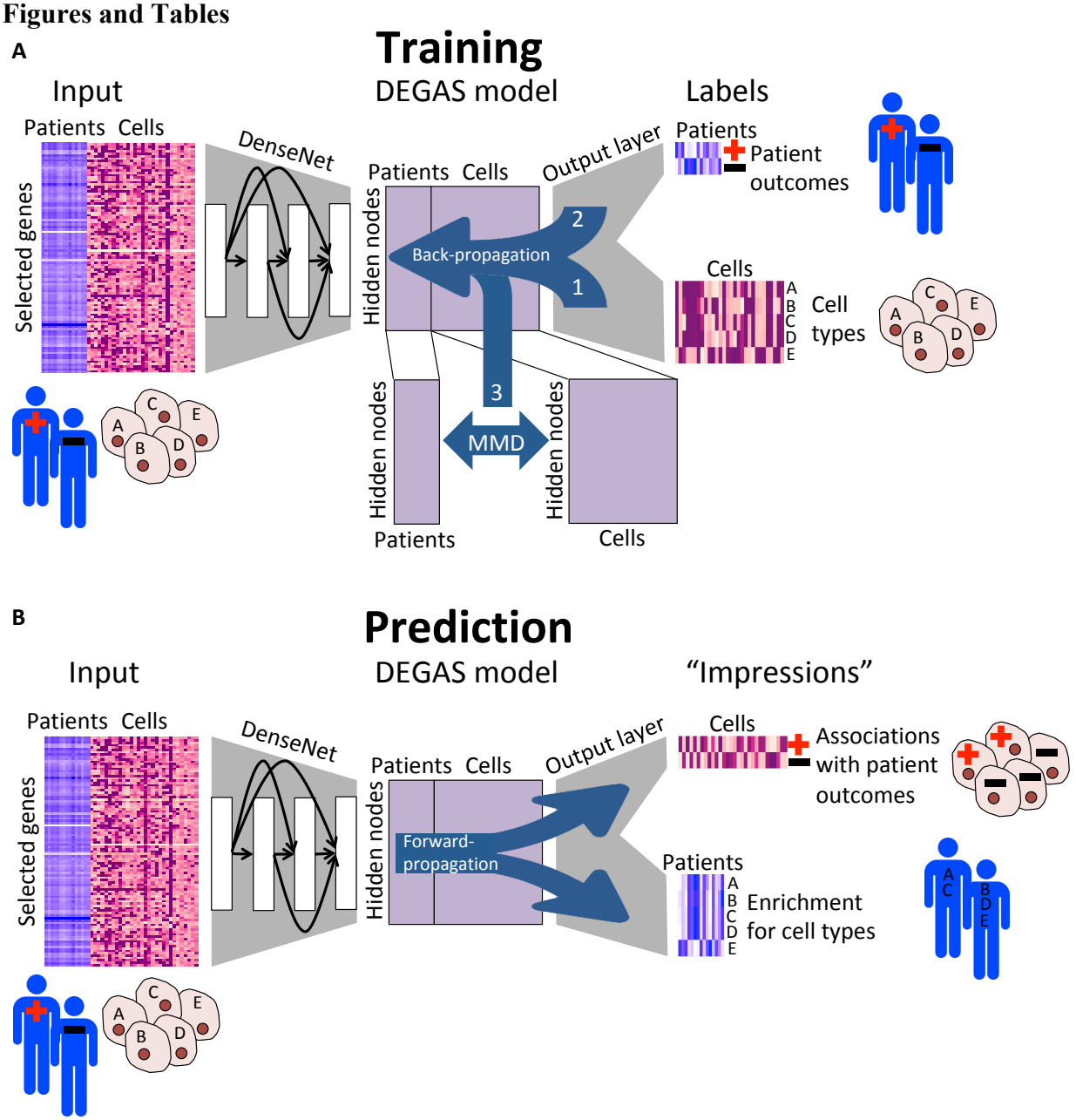


**Fig. 1 A workflow diagram of the *DEGAS* framework**. **A**) The scRNA-seq and patient expression data are preprocessed into expression matrices. Next, Bootstrap aggregated DenseNet *DEGAS* models are trained using both single cell and patient outcomes using a multitask learning neural network that learns latent representation reducing the differences between patients and single cells at the final hidden layer using maximum mean discrepancy (MMD). **B**) The output layer of this model can be used to simultaneously infer clinical outcome impressions in single cells and cellular composition impressions in patients.

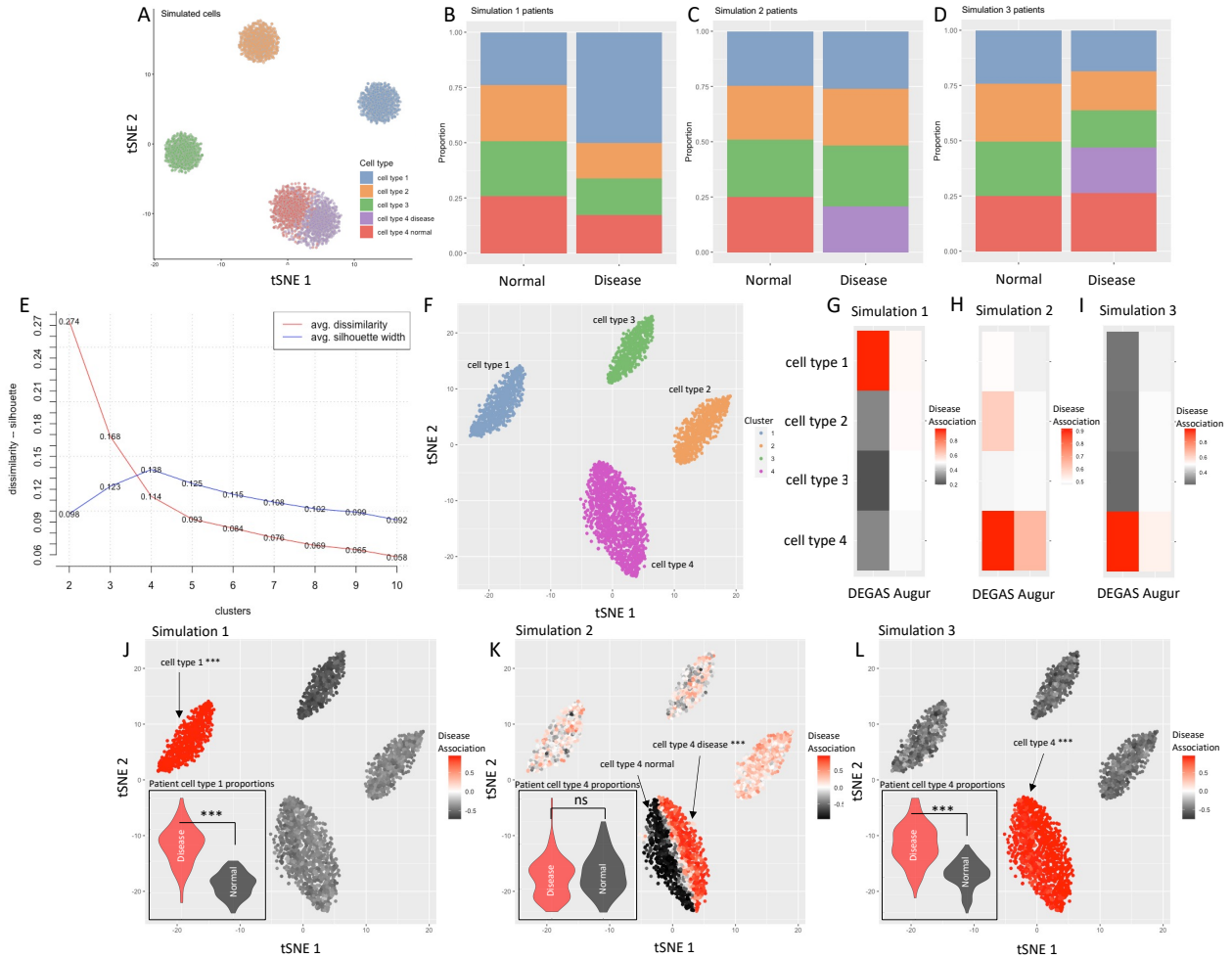


**Fig. 2 Simulation study and baseline comparisons of *DEGAS* framework. A)** 5,000 simulated cells from splatter with 4 cell types where one of the cell types has two subtypes. Cell type 4 is composed of two subtypes that are specific to either disease or normal patients. 2,000 of these cells were used to generate the 600 simulated patients in **B-D** and 3,000 were used as the cell input to our models. **E)** Optimal cluster number (4 clusters) based on average silhouette width for 3,000 cells not used to generate patients. **F)** The same 3,000 cells used as the cellular input colored by their cluster. **G)** *DEGAS* comparison to Augur in simulation 1. **H)** *DEGAS* comparison with *Augur* in simulation 2. **I)** *DEGAS* comparison with *Augur* in simulation 3. **J-L)** *DEGAS*-calculated disease association from each simulation overlaid onto 3,000 cells. The violin plot in the bottom left corner is deconvolution cell type proportion for cell type 1 in simulation 1 patients (**J**), cell type 4 proportion in simulation 2 patients (**K**), and cell type 4 proportion in simulation 3 patients (**L**).

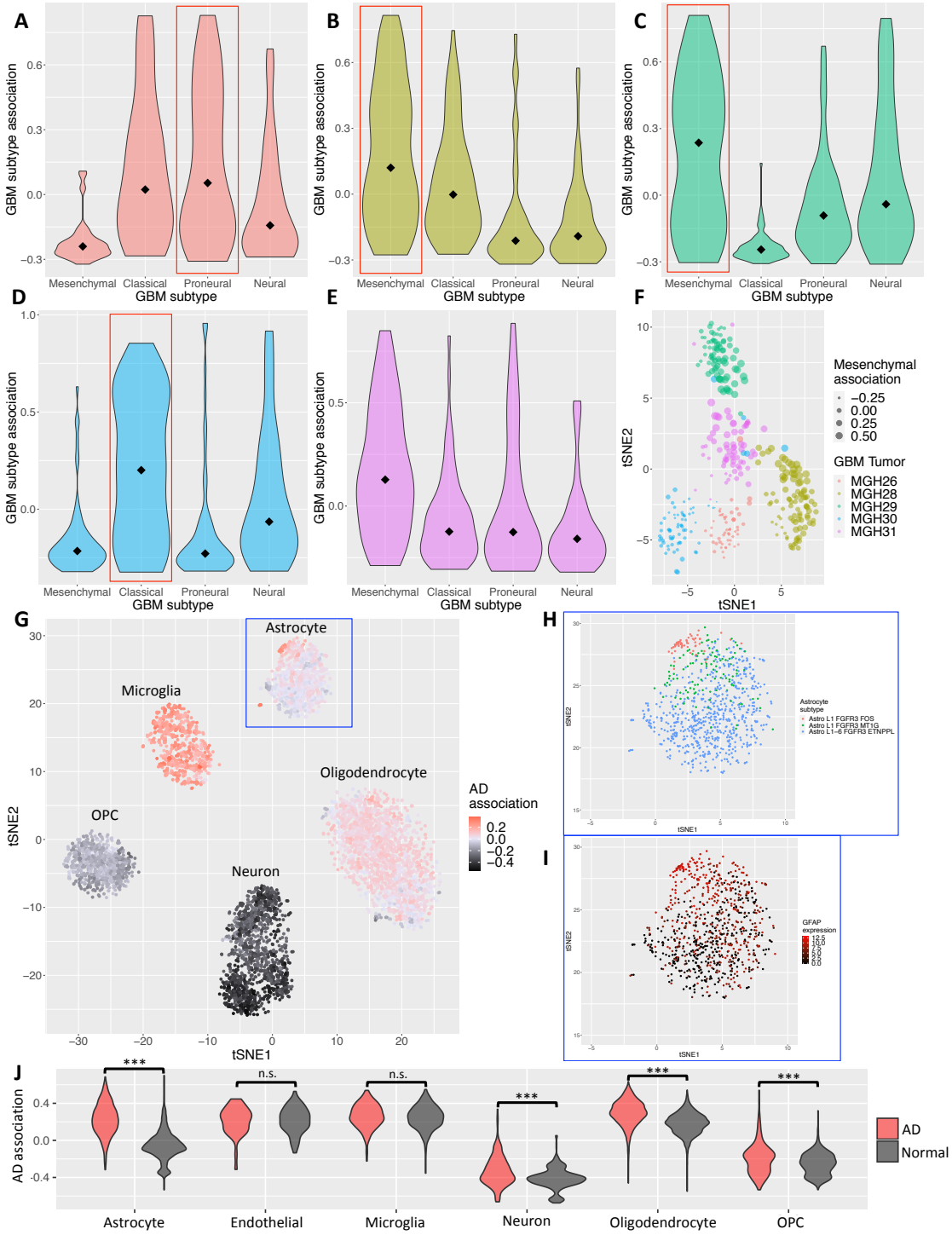


**Fig. 3 *DEGAS* validation in GBM and AD.** *DEGAS* output of the distribution of GBM subtypes in single cells from five GBM tumors. Four of the five tumors had known GBM subtype information from Patel *et al*. (MGH26: Proneural, MGH28: Mesenchymal, MGH29: Mesenchymal, and MGH30: Classical, indicated by red boxes) which were recapitulated by *DEGAS*. The subtype information for the tumors, MGH26, MGH28, MGH29, and MGH30 were derived from Patel *et al.* where MGH31 did not have a clearly defined subtype in Patel *et al.* The association of cells assigned to each subtype were plotted for each tumor; **A)** MGH26, **B)** MGH28, **C)** MGH29, **D)** MGH30 and **E)** MGH31. Median values are marked by a diamond in each of the violin plots. **F)** The association with the mesenchymal subtype is overlaid on all of the single cells from the five tumors (indicated by the size of the dots). It clearly shows that the single cell from MGH28 and MGH29 have the highest association with the mesenchymal subtype. Note: the tumors in **(F)** are color coded the same

as the boxplots in **(A-E)**. **G)** *DEGAS* output of AD association score for each single cell. The AD association score is indicated by the color and is overlaid onto Allen Institute single cells. This plot shows the negative AD association in neuron cells and positive AD association in Microglia. **H-I)** There also appeared to be a subpopulation of astrocytes with positive AD association. The astrocytes were plotted separately and colored by Allen Institute Astrocyte subtypes **(H)** and *GFAP* expression, a disease-associated astrocyte marker **(I)**. **J)** Comparison of *DEGAS*-derived AD association scores for single cells from AD and Normal control samples from Grubman *et al.* Significance values: n.s. (not significant), • (0.1), * (0.05), ** (0.01), *** (0.001).

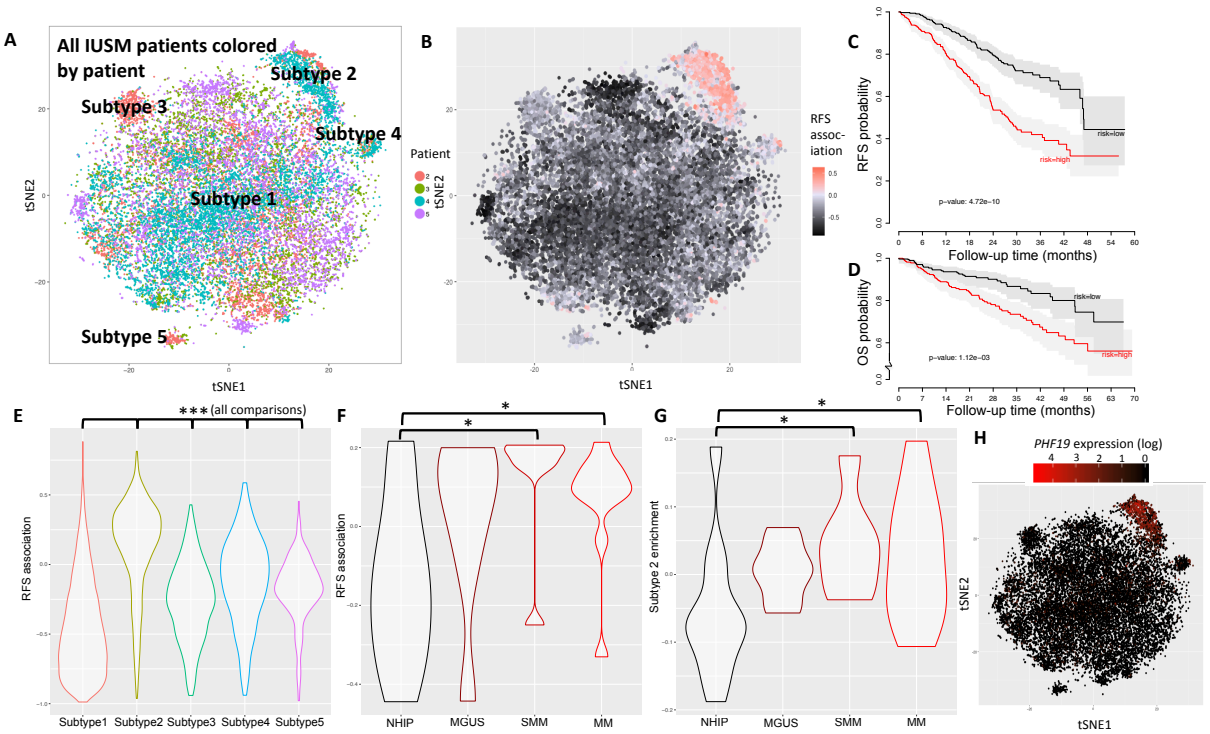


**Fig. 4 Association between subtypes and relapse risk in MM**. IUSM CD138+ scRNA-seq subtype clusters generated from *Seurat* colored by **A)** cluster, i.e., subtype and **B)** relapse association. **C)** Kaplan-Meier curves from cross-validation for the MMRF patients stratified by median proportional hazard. **D)** Kaplan-Meier curves of overall survival of Zhan *et al.* external dataset stratified by median proportional hazard. **E)** Relapse association for IUSM CD138+ subtypes **F)** Relapse association prediction for NHIP, MGUS, SMM, and MM in the external dataset Ledergor *et al*. **G)** Subtype 2 enrichment for NHIP, MGUS, SMM, and MM in the external dataset Ledergor *et al*. NHIP: normal hip bone marrow, MGUS: monoclonal gammopathy of undetermined significance, SMM: smoldering multiple myeloma, MM: multiple myeloma. Significance values: • (0.1), * (0.05), ** (0.01), *** (0.001). All median values in violin plots are marked with a diamond. All plots were generated using the default parameters for the *DEGAS* package described in the section of **Methods**: *Transfer learning using DEGAS*.

**Table 1 Comparison of AD association scores in single cells between cell types as visualized in Fig. 3G.** The *DEGAS* models were trained using neuron, oligodendrocyte, astrocyte, OPC, and microglia cell types. The single cells were split into groups based on their cell type and the mean AD associations of each cell type was evaluated as a correlation. The Neuron and microglia groups are bolded to highlight their much higher mean AD association

| Cell type | Cell-type mean association | Number of cells | p-value |
|---|---|---|---|

| **Neuron** | **-0.35** | **1329** | **$<2.2\times10^{-16}$** |
|---|---|---|---|
| Oligodendrocyte | 0.05 | 1795 | $3.42\times10^{-2}$ |
| Astrocyte | 0.03 | 809 | $3.94\times10^{-1}$ |
| OPC | -0.12 | 738 | $1.09\times10^{-3}$ |
| **Microglia** | **0.22** | **741** | **$1.42\times10^{-9}$** |

**Table 2. Summary of the clinical features in each patient cohorts used in training.** * Final age category is >90 years.

| **Glioblastoma Multiforme TCGA** | |
|---|---|
| **Feature** | **Details** |
| Sex | 74 Male, 37 Female |
| Age (years) | Range: 14-83, Mean: 56, Median: 58 |
| Clinical GBM subtype | 34 Classical, 33 Mesenchymal, 9 Neural, 35 Proneural |
| **Alzheimer's Disease MSBB** | |
| **Feature** | **Details** |
| Sex | 90 Male, 131 Female |
| Age (years) | Range: 61-90+, Mean* > 82, Median = 84 |
| AD diagnosis | 135 AD, 86 Control |
| **Multiple Myeloma MMRF** | |
| **Feature** | **Details** |
| Sex | 387 Male, 260 Female |
| Age (years) | Range: 27-93, Mean: 64, Median: 64 |
| Relapse-free survival time (days) | Range: 13-1753, Mean: 665.4, Median: 629<br>200 patients relapsed |

**Table 3. Overview of all datasets used in the analysis.** *The simulated patients were generated from the splatter simulated cells by combining known proportions of cell types and subtypes.

| Study | Dataset | Sample size | Data type | Outcome |
|---|---|---|---|---|
| **Simulation** | Simulated cells* | 5000 cells | scRNA-seq | Cell type |
| | Simulated patients* | 600 patients | RNA-seq | Disease status |

| | | | | |
|---|---|---|---|---|
| **Glioblastoma** | Patel *et al*., 2014 | 532 cells (5 patients) | scRNA-seq (SMART-seq) | None |
| | TCGA GBM | 111 patients | Microarray | GBM subtype |
| **Alzheimer's disease** | Allen Institute | 47,396 cells (11 patients) | scRNA-seq (SMART-seq) | Brain cell types |
| | Grubman *et al.*, 2019 | 13,214 cells (12 patients) | snRNA-seq (10x Genomics) | AD and normal brain cell types |
| | MSBB | 682 samples (221 patients) | RNA-seq | AD diagnosis |
| **Multiple myeloma** | MMRF | 647 patients | RNA-seq | Relapse-Free Survival |
| | IUSM | 22,968 cells (4 patients) | scRNA-seq (10x Genomics) | Subtype cluster (Subtype 1-5) |
| | Ledergor *et al*., 2019 | 13,440 cells (35 patients) | scRNA-seq (MARS-seq) | Malignancy (NHIP, MGUS, SMM, MM) |
| | Zhan *et al*., 2006 | 559 patients | Microarray | Overall Survival |

0
1

**Supplementary Materials**

Supplementary Materials for

# Diagnostic Evidence GAuge of Single cells (DEGAS): A flexible deep-transfer learning framework for prioritizing cells in relation to disease

Travis S. Johnson, Christina Y. Yu, Zhi Huang, Siwen Xu, Tongxin Wang, Chuanpeng Dong, Wei Shao, Mohammad Abu Zaid, Xiaoqing Huang, Yijie Wang, Christopher Bartlett, Yan Zhang, Brian A. Walker, Yunlong Liu, Jie Zhang*, Kun Huang*

*Corresponding author. Email: jizhan@iu.edu, kunhuang@iu.edu

**This PDF file includes:**

Figs. S1 to S5
Tables S1 to S9

**Other Supplementary Materials for this manuscript include the following:**

Supplementary File 1 to 2

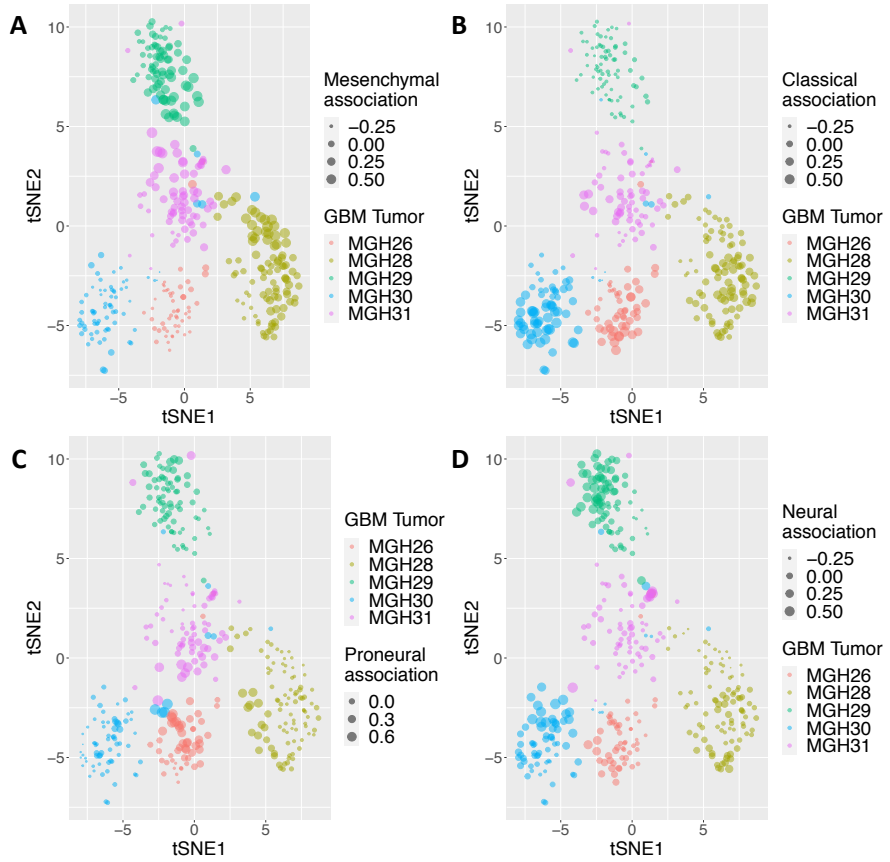


**Fig. S1** Scatterplots for the single cells in each of the Patel et al. GBM tumors overlaid with GBM subtype associations: **A)** Mesenchymal association, **B)** Classical association, **C)** Proneural association, and **D)** Neural association.

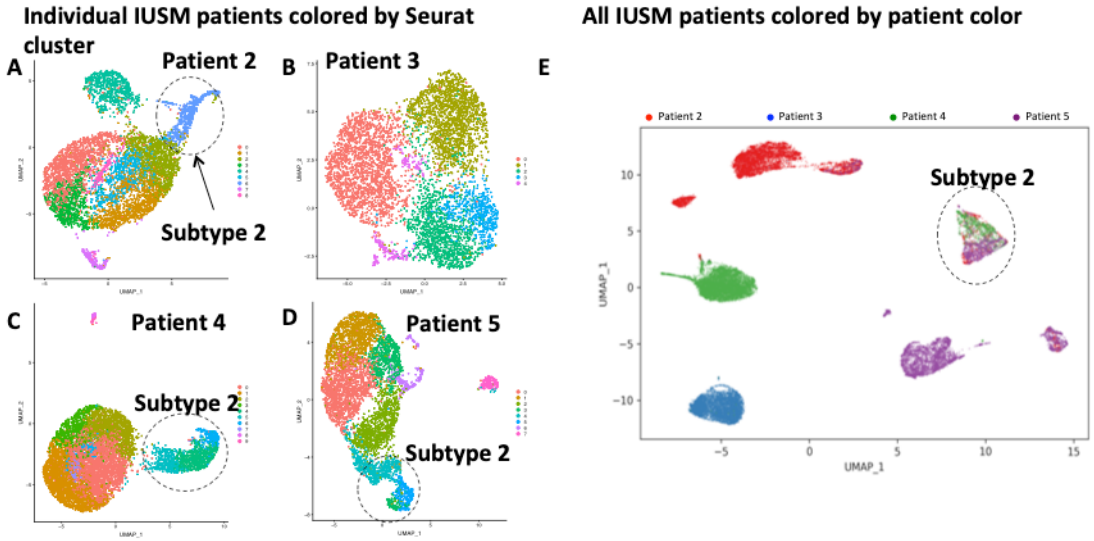


**Fig. S2** Clustering of IUSM myeloma patient single cells. **A-D)** The Seurat individual patient clustering results colored by cluster for **A)** IUSM patient 2, **B)** IUSM patient 3, **C)** IUSM patient 4, and **D)** IUSM patient 5. **E)** The clustering results from *BERMUDA* for all four patients colored by patient.

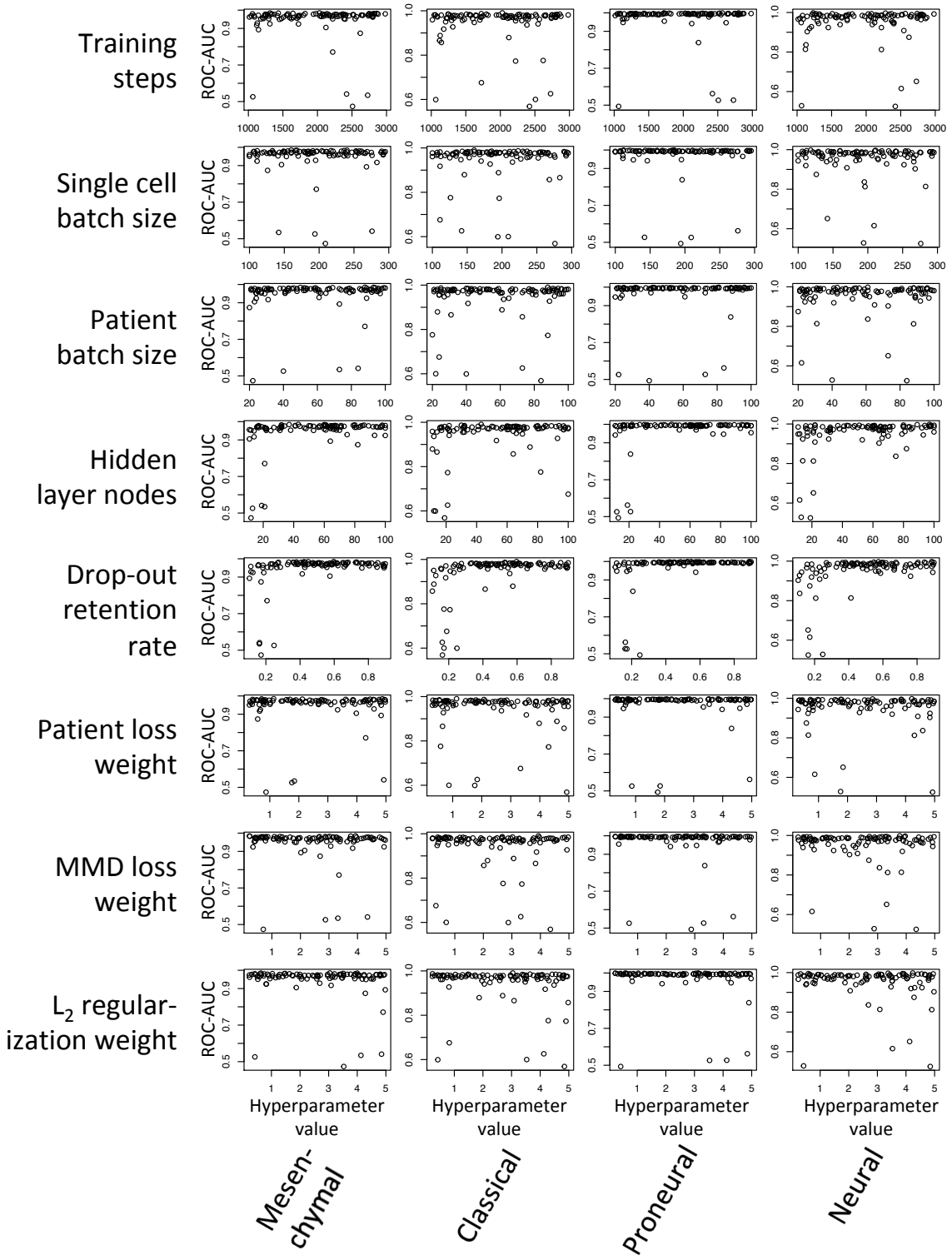


**Fig. S3** All scatterplots between the choices of hyperparameter values and the classification performance. The y-axis is the performance as measured by ROC-AUC for a given GBM subtype label, as shown in the bottom of the figure. Each specific parameter type is shown on the left side of the figure. The corresponding spearman correlation coefficients are computed and shown in **Table S8**.

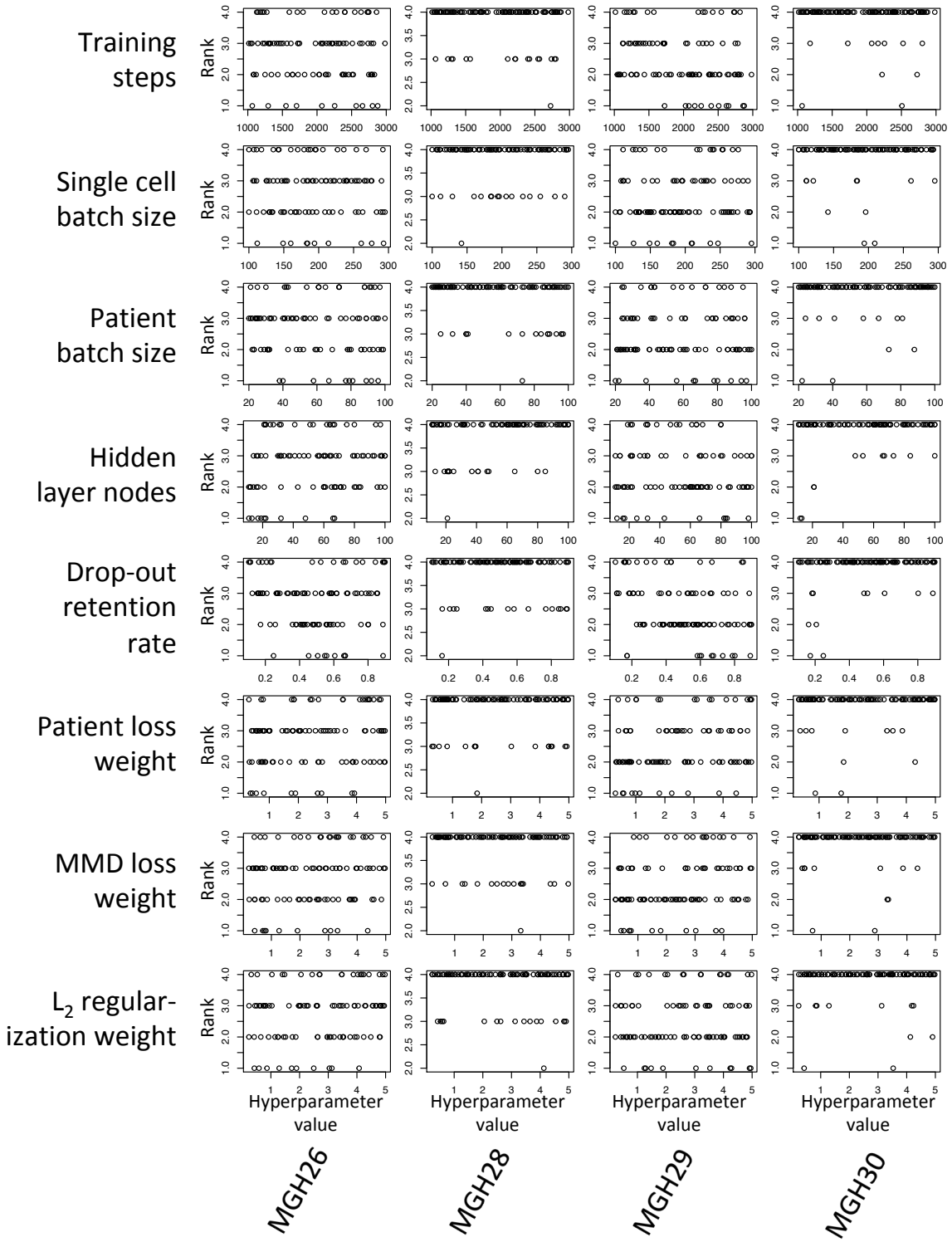


**Fig. S4** All scatterplots between hyperparameter values and correct label rank such that 4 is the desired rank. The x-axis values in each plot is the hyperparameter value of a specific hyperparameter label, as shown on the left side of the figure. The y-axis is the rank for the correct label in a given GBM tumor. The corresponding spearman correlation coefficients are computed and shown in **Table S9**.

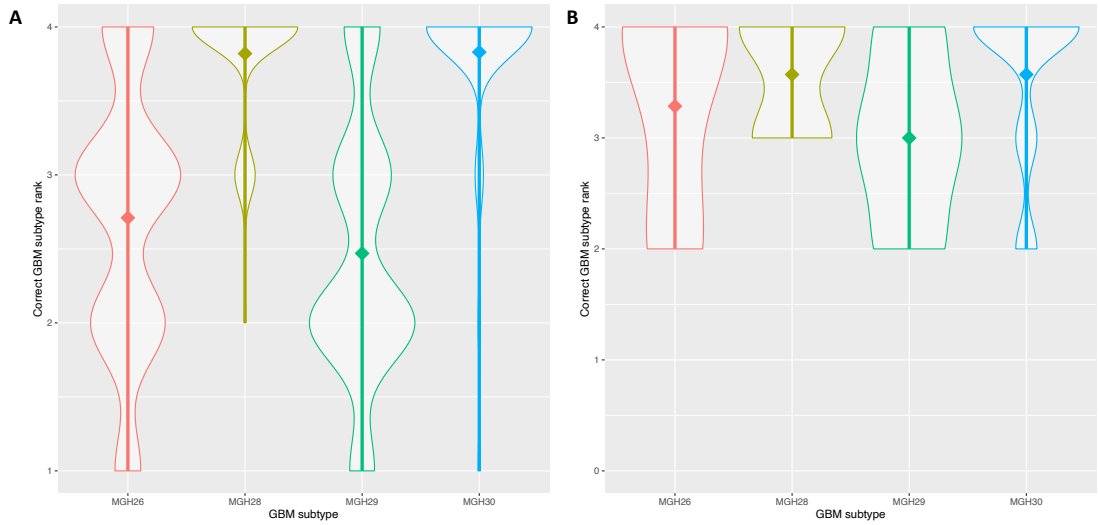


**Fig. S5** Rank of the correct cell label for each of the Patel et al. GBM tumors. **A)** These are the results including all 100 randomly selected sets of hyperparameters. **B)** These are the results including only the hyperparameters where $\lambda_1$, $\lambda_2$, and $\lambda_3$ are greater than 3 that represent the default hyperparameters. The rank of the correct label was calculated by calculating the mean of each GBM subtype association across all of the cells in that tumor. This resulted in each of the 100 random hyperparameters having a rank for each GBM subtype for each of the GBM scRNA-seq tumors (4 highest ranked, 1 lowest ranked). Ideally the rank for correct labels would be 4 since there are 4 total labels (Mesenchymal, Classical, Proneural, Neural). The mean rank value is marked by a diamond in both **A** and **B**.

**Table S1.** Prediction performance metrics for simulated patients including receiver operating curve (ROC) area under the curve (AUC) and precision recall (PR) curve AUC.

| | **Simulation 1** | | **Simulation 2** | | **Simulation 3** | |
|---|---|---|---|---|---|---|
| | **ROC-AUC** | **PR-AUC** | **ROC-AUC** | **PR-AUC** | **ROC-AUC** | **PR-AUC** |
| **Disease status** | 0.94 | 0.96 | 0.99 | 0.99 | 0.96 | 0.98 |

**Table S2.** Prediction performance metrics for simulated cells including ROC-AUC and PR-AUC.

| | **Simulation 1** | | **Simulation 2** | | **Simulation 3** | |
|---|---|---|---|---|---|---|
| | **ROC-AUC** | **PR-AUC** | **ROC-AUC** | **PR-AUC** | **ROC-AUC** | **PR-AUC** |
| **Cell type 1** | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| **Cell type 2** | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| **Cell type 3** | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| **Cell type 4** | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |

**Table S3.** Prediction performance metrics for TCGA GBM patients into molecular subtypes including ROC-AUC and PR-AUC.

| | **ROC-AUC** | **PR-AUC** |
|---|---|---|
| **Mesenchymal** | 0.93 | 0.89 |
| **Classical** | 0.94 | 0.86 |
| **Proneural** | 0.99 | 0.97 |
| **Neural** | 0.97 | 0.79 |

**Table S4.** Prediction performance metrics for MSBB patients including ROC-AUC and PR-AUC.

| | **ROC-AUC** | **PR-AUC** |
|---|---|---|
| **AD status** | 0.78 | 0.82 |

**Table S5.** Prediction performance metrics for Allen Brain Institute cells including ROC-AUC and PR-AUC.

| | ROC-AUC | PR-AUC |
|---|---|---|
| **Neuron** | 0.99 | 0.99 |
| **Oligodendrocyte** | 0.99 | 0.99 |
| **Astrocyte** | 0.99 | 0.99 |
| **OPC** | 0.99 | 0.99 |
| **Microglia** | 0.99 | 0.99 |

**Table S6.** Consistency of subtype clusters between multi-patient clustering (**Fig. 4A**) and single patient (**Fig. S2A-D**) clustering experiments. Rand, Fowlkes and Mallows's index (FM), and Jaccard index (JI) were used to measure the subtype cluster consistency.

| | Rand | FM | JI |
|---|---|---|---|
| **Patient 2** | 0.7059083 | 0.7446245 | 0.5872991 |
| **Patient 3** | 0.6362535 | 0.7976550 | 0.6362535 |
| **Patient 4** | 0.6064240 | 0.6417339 | 0.4611969 |
| **Patient 5** | 0.6683671 | 0.8059836 | 0.6637983 |

**Table S7.** Prediction performance metrics for IUSM myeloma single cells including ROC-AUC and PR-AUC.

| | ROC-AUC | PR-AUC |
|---|---|---|
| **Subtype 1** | 0.90 | 0.98 |
| **Subtype 2 (*PHF19*$^{high}$)** | 0.98 | 0.91 |
| **Subtype 3** | 0.90 | 0.62 |
| **Subtype 4** | 0.91 | 0.44 |
| **Subtype 5** | 0.98 | 0.84 |

**Table S8.** Spearman correlation coefficients between AUC for each of the patient output labels in patients and 100 randomly selected sets of hyperparameters. Significance values: • (0.1), * (0.05), ** (0.01), *** (0.001).

| | Mesenchymal | Classical | Proneural | Neural |
|---|---|---|---|---|
| **Training steps** | 0.20* | 0.11 | 0.19 | 0.17• |
| **Single cell batch size** | 0.03 | 0.07 | -0.02 | 0.01 |
| **Patient batch size** | 0.17• | 0.09 | 0.17• | 0.08 |
| **Hidden layer nodes** | 0.31** | 0.27** | 0.19• | 0.33*** |
| **Drop-out retention rate** | 0.28** | 0.38*** | 0.42*** | 0.40*** |
| **Patient loss weight ($\lambda_1$)** | 0.03 | -0.08 | 0.03 | -0.04 |
| **MMD loss weight ($\lambda_2$)** | -0.18• | -0.12 | -0.10 | -0.05 |
| **L2 regularization weight ($\lambda_3$)** | -0.09 | -0.12 | -0.19• | -0.06 |

**Table S9.** Spearman correlation coefficients between correct label rank (MGH26: Proneural, MGH28: Mesenchymal, MGH29: Mesenchymal, MGH30: Classical) in single cells and 100 randomly selected sets of hyperparameters. Since there are 4 labels the best rank is 4, which indicates the correct label was assigned across the cells in that tumor. Significance values: • (0.1), * (0.05), ** (0.01), *** (0.001).

| | MGH26 | MGH28 | MGH29 | MGH30 |
|---|---|---|---|---|
| **Training steps** | -0.07 | -0.10 | -0.25* | -0.07 |

| | | | | |
|---|---|---|---|---|
| **Single cell batch size** | -0.06 | 0.01 | 0.06 | 0.07 |
| **Patient batch size** | -0.06 | -0.22* | 0.01 | 0.06 |
| **Hidden layer nodes** | 0.10 | 0.27** | -0.08 | 0.06 |
| **Drop-out retention rate** | -0.07 | -0.06 | -0.34*** | 0.15 |
| **Patient loss weight ($\lambda_1$)** | 0.15 | -0.02 | 0.24* | 0.10 |
| **MMD loss weight ($\lambda_2$)** | 0.11 | -0.10 | 0.28* | 0.07 |
| **L2 regularization weight ($\lambda_3$)** | 0.18• | -0.09 | 0.01 | 0.03 |