PROS: A Panoramic Representation Of multi-cellular Samples

Patrick Yu
University of Washington
pzy@cs.washington.edu

Wei Qiu
University of Washington
wqiu0528@cs.washington.edu

Ethan Weinberger
University of Washington
ewein@cs.washington.edu

Maria Brbić*
EPFL
mbrbic@epfl.ch

Su-In Lee*
University of Washington
suinlee@cs.washington.edu

Single-cell omics has emerged as a valuable tool for investigating cellular and molecular diversity. With improvements in sequencing techniques, large-scale single-cell studies now routinely involve collecting donor samples, each consisting of thousands of cells. Collectively, these cells serve as biological representations of the donor, and it may be of interest to learn a *donor embedding* from donor samples. Such a donor embedding can then be used to stratify donors by disease subtype, or identify cell subtype-specific changes across disease progression.

An ideal donor embedding should transform an arbitrary number of donor cells into a consistent set of features, each representing variations in a particular cell subtype of a donor. One way to obtain donor embeddings is by averaging (pseudobulking) across a donor's cell representations. Recent approaches utilize attention-based multiple instance learning (ABMIL) [1] to model donor embeddings as an attention-weighted average of a donor's cell representations [3, 4, 9]. However, by taking the unweighted or weighted average of all cell representations in a donor, pseudobulk and ABMIL-based methods are unable to capture disease-critical mechanisms that are *specific* to a particular cell subtype - such as cell subtype-specific changes in location, spread, and proportion across disease progression.

To this end, we propose PROS (A **P**anoramic **R**epresentation **O**f multi-cellular **S**amples), a donor embedding that provides a *panoramic* view of the location, spread, and proportion of each cell subtype in a given donor (Figure 1). We used PROS embeddings to train a MLP classifier for predicting disease and age phenotypes from held-out donors, and we found that classifiers trained on PROS embeddings outperformed alternative disease imputation methods when evaluated on four different scRNA-seq studies.

1 PROS Donor Embeddings

For a given study, let $z_i^s \in \mathbb{R}^d$ denote the scVI [5] representation for the i-th cell collected from the s-th donor. For all cells i from all donors s, PROS assumes that z_i^s is sampled from a global Gaussian Mixture Model (GMM) $p(z_i^s|\theta^{global})$ that models C distinct cell subtypes, with each global mixture component $\theta_c^{global} = \{\pi_c^{global}, \mu_c^{global}, \Sigma_c^{global}\}$ representing a distinct cell subtype:

$$p(z_i^s|\theta^{global}) = \sum_{c=1}^{C} \pi_c^{global} \mathcal{N}(z_i^s|\mu_c^{global}, \Sigma_c^{global}), s.t. \sum_{c=1}^{C} \pi_c^{global} = 1$$

The global cluster weight $\pi_c^{global} \in \mathbb{R}$ represents the abundance of cell subtype c for all donors in the study, while the global cluster mean $\mu_c^{global} \in \mathbb{R}^d$ and diagonal covariance $\Sigma_c^{global} \in \mathbb{R}^d$ describe the location and spread of cell subtype c for all donors in the study.

For each donor s, we can do a donor-specific M-step to estimate donor-specific cluster weights, mean, and covariances, by using cells from donor s and the cluster responsibilities from our global GMM. Let $p(c|z_i^s, \theta^{global})$ denote the responsibility of cluster c to cell z_i^s under our global GMM model, θ^{global}. Then, our donor-specific parameters for donor s are defined as follows:

$$\pi_c^s = \frac{\sum_{i=1}^{n_s} p(c|z_i^s, \theta^{global})}{n_s},\ \mu_c^s = \frac{\sum_{i=1}^{n_s} z_i^s \cdot p(c|z_i^s, \theta^{global})}{\sum_{i=1}^{n_s} p(c|z_i^s, \theta^{global})},\ \Sigma_c^s = \frac{\sum_{i=1}^{n_s} (z_i^s - \mu_c^s)^2 \cdot p(c|z_i^s, \theta^{global})}{\sum_{i=1}^{n_s} p(c|z_i^s, \theta^{global})} \tag{1}$$

where n_s denotes the number of cells in donor s.

Then, the PROS embedding of donor sample s is defined as the concatenation of $\pi_c^s, \mu_c^s, \Sigma_c^s$ across all C cell subtypes: $PROS(s) = concat(\pi_1^s, \mu_1^s, \Sigma_1^s, \pi_2^s, \mu_2^s, \Sigma_2^s, \ldots, \pi_C^s, \mu_C^s, \Sigma_C^s) \in \mathbb{R}^{C(1+2d)}$.

Preprint. Under review.

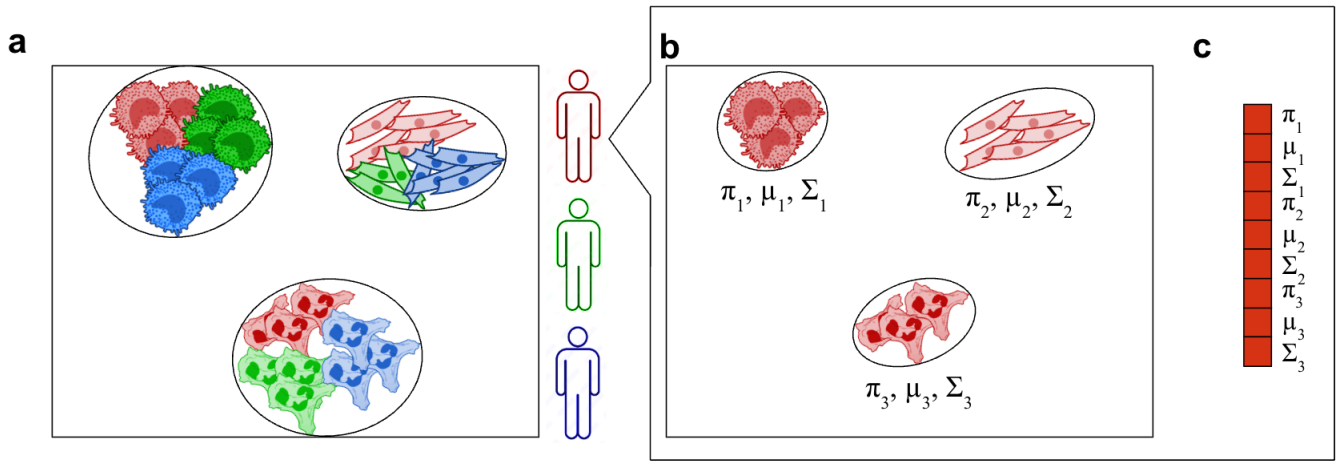


Figure 1: **a**) Each study consists of multiple donors and multiple cell types. Using a Gaussian Mixture Model, we can identify global cell subtypes that are shared across donors. **b**) The global cell subtypes are mapped onto each donor's cells to obtain donor-specific cluster parameters. **c**) The PROS embedding for a donor is obtained by concatenating the donor-specific cluster parameters across all clusters. In doing so, PROS embeddings describe the abundance, location, and spread of each cell subtype in a particular donor.

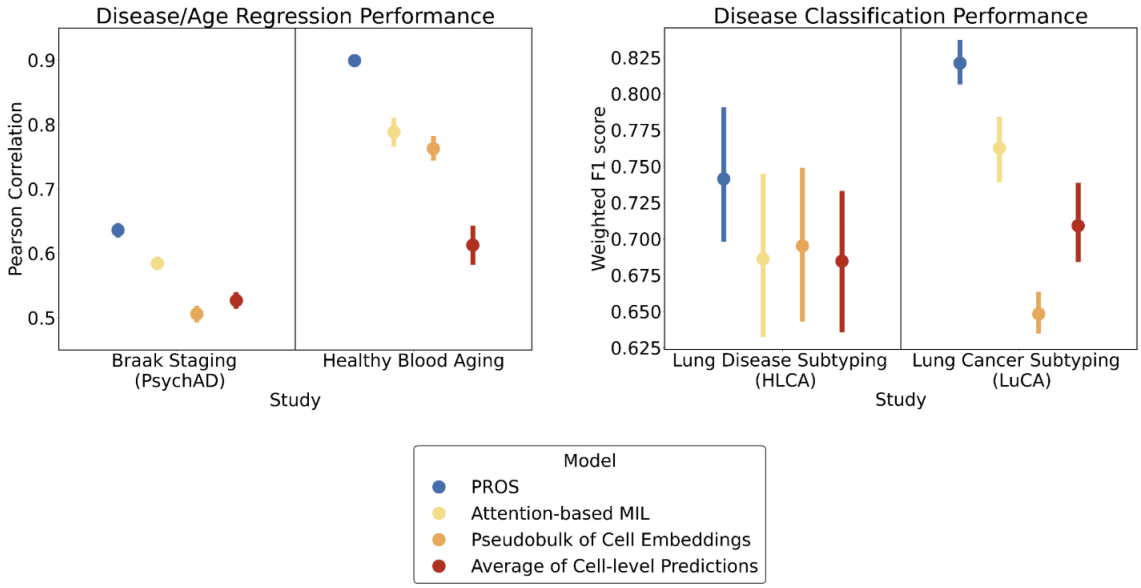


Figure 2: Classifiers trained on PROS embeddings outperforms baseline models' performance at predicting held-out donor phenotype scores. Plots depict mean $\pm$ standard error across 10 train-validation-test splits for each metric.

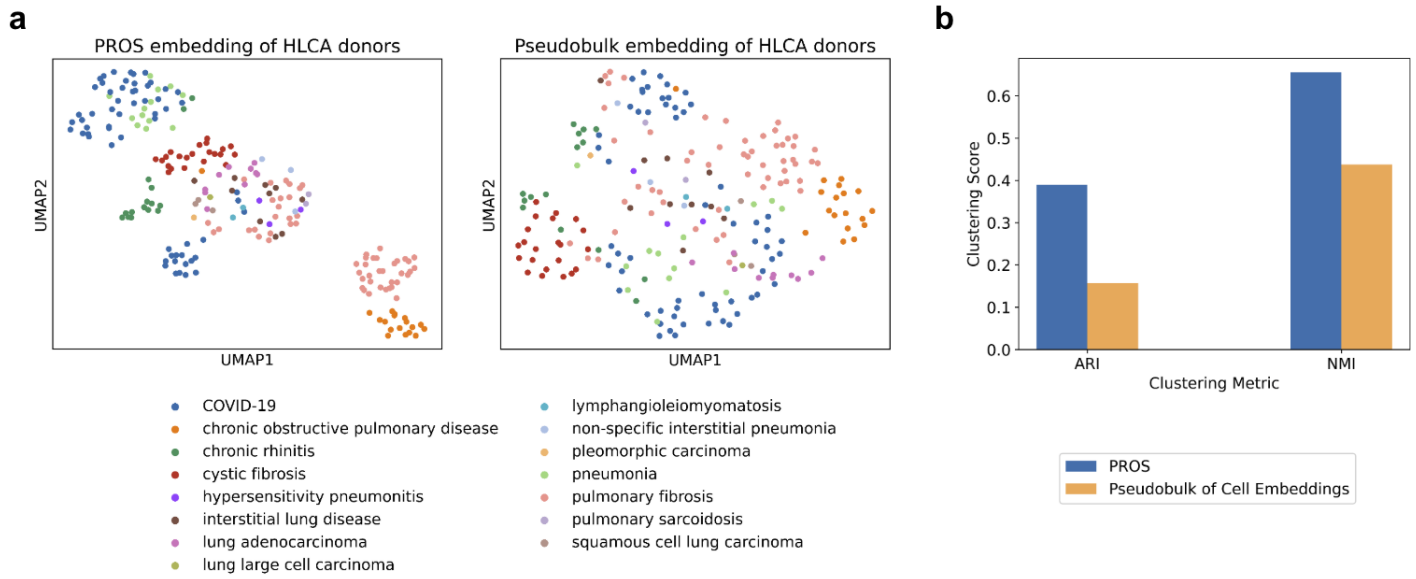
2 Results

To measure PROS' capabilities as a donor representation, we used PROS embeddings to train MLP classifiers in predicting held-out age and disease phenotype labels across four different scRNA-seq studies. We applied PROS embeddings on two regression tasks - predicting Braak neuropathology scores from 1,176 Alzheimer's donors from the PsychAD study [2], and chronological age using 318 healthy donor blood samples from [8]. We also applied PROS embeddings on two classification tasks - predicting lung disease subtypes in 486 donors from 49 substudies in the Human Lung Cell Atlas (HLCA) [7], and lung cancer subtypes in 309 donors from the Lung Cancer Atlas (LuCA) [6].

For all datasets except HLCA, we randomly held out phenotype labels from 20% of donors during training, and predicted their labels during evaluation. For the HLCA dataset, we tested PROS' ability to impute phenotypes in unseen study cohorts, where we held out phenotype labels for all donors from 10 randomly sampled substudies and predicted their labels during evaluation.

Results indicate that disease classifiers trained using PROS embeddings outperformed all baseline methods across all four datasets tested (Figure 2). We also investigated whether PROS embeddings could be directly used to cluster donors by disease subtype in an unsupervised manner, and found out that PROS embeddings performed better at stratifying HLCA donors by disease than pseudobulking (Supplementary Figure 3). Taken together, these results demonstrate PROS' ability to serve as a powerful donor representation, particularly in relating donor-level scRNA-seq data with disease phenotypes, in both a supervised and unsupervised manner. In future work, we will investigate PROS' capabilities to facilitate new biological insights (e.g. locating cell type-specific drivers of disease).

*Denotes co-senior authorship



Supplementary Figure 3: PROS embeddings can be directly used to stratify donors by disease subtype in an unsupervised manner. **a**) HLCA donors are more visibly separated by disease subtype under PROS embeddings than pseudobulk embeddings. **b**) When compared with ground-truth disease subtype labels, disease clusterings obtained from PROS embeddings achieve a higher ARI and NMI score than disease clusterings obtained from pseudobulk embeddings.

3 Supplementary Information

We used 50-dimensional scVI cell representations [5] for all experiments, that is, $z_i^s \in \mathbb{R}^{50}$ for all donors s and cells i.

For PROS, we used scikit-learn's GaussianMixture with diagonal covariance and k-means initialization with C = {80, 160} clusters. This GMM was fit on all cells, from all donors in a given study, until convergence. Then, for each donor, we did a donor-specific M-step as described in Equation 1, and concatenated the resulting donor-specific cluster parameters to obtain a PROS embedding for each donor. We used the PROS embedding as input to a two-layer multilayer perceptron with a 50-dimensional hidden layer, which was trained to predict disease phenotypes.

For ABMIL, we used the formulation from [1], where the donor embedding for donor s with n_s cells is defined as $ABMIL(s) = \sum_{k=1}^{n_s} a_k h_k$, where the attention mechanism is defined as $a_k = \frac{exp\{w^T tanh(Vh_k^T)\}}{\sum_{j=1}^{n_s} exp\{w^T tanh(Vh_j^T)\}}$ with $w \in \mathbb{R}^{50\times 1}$, $V \in \mathbb{R}^{50\times 50}$, and $h_k = f_\psi(z_k^s)$ where f_ψ is a linear layer with a 50-dimensional output. We then used another linear layer to obtain the phenotype prediction from the ABMIL embedding.

For pseudobulking, we defined the donor embedding for donor s as $PB(s) = \sum_{k=1}^{n_s} z_k^s$. We used the pseudobulk embedding as input to a two-layer multilayer perceptron with a 50-dimensional hidden layer, which was trained to predict disease phenotypes.

For average of cell-level predictions, we trained a two-layer multilayer perceptron with a 50-dimensional hidden layer to predict disease phenotype for each individual cell in the training dataset. For classification tasks, we predicted the disease phenotype for each cell in a test donor, and reported the most-frequently classified phenotype class as the test donor's predicted phenotype. For regression tasks, we reported the average of phenotype predictions for all cells in the test donor.

For each model tested, we used a batch size of 1 and trained for 200 epochs or until convergence, and sweeped over learning rates of $\{10^{-2}, 10^{-3}, 10^{-4}\}$. We used a held out validation set (20% of a study's donors) to select the optimal learning rate, epoch, and for PROS, the optimal number of clusters, and evaluated this optimal model on a held out test set (another 20% of a study's donors). We repeated this process for 10 randomly generated train-val-test splits, and reported the average test performance and standard error across the 10 splits.

In Supplementary Figure 3, we used K-means clustering with 15 clusters (the number of HLCA disease subtypes) on both PROS and pseudobulk embeddings to obtain predicted clusters. We then used ARI (Adjusted Rand Index) and NMI (Normalized Mutual Information) metrics to evaluate the quality of the predicted clusterings against the true disease clusters.

References

[1] Maximilian Ilse, Jakub M. Tomczak, and Max Welling. Attention-based deep multiple instance learning. *CoRR*, abs/1802.04712, 2018. URL `http://arxiv.org/abs/1802.04712`.

[2] Donghoon Lee, Mikaela Koutrouli, Nicolas Y. Masse, Gabriel E. Hoffman, Seon Kinrot, Xinyi Wang, N.M. Prashant, Milos Pjanic, Tereza Clarence, Fotios Tsetsos, Deepika Mathur, David Burstein, Karen Therrien, Aram Hong, Clara Casey, Zhiping Shao, Marcela

Alvia, Stathis Argyriou, Jennifer Monteiro Fortes, Pavel Katsel, Pavan K. Auluck, Lisa L. Barnes, Stefano Marenco, David A. Bennett, PsychAD Consortium, Lars Juhl Jensen, Kiran Girdhar, Georgios Voloudakis, Vahram Haroutunian, Jaroslav Bendl, John F. Fullard, and Panos Roussos. Single-cell atlas of transcriptomic vulnerability across multiple neurodegenerative and neuropsychiatric diseases. *medRxiv*, 2024. doi: 10.1101/2024.10.31.24316513. URL https://www.medrxiv.org/content/early/2024/11/04/2024.10.31.24316513.

[3] Anastasia Litinetskaya, Maiia Shulman, Soroor Hediyeh-zadeh, Amir Ali Moinfar, Fabiola Curion, Artur Szałata, Alireza Omidi, Mohammad Lotfollahi, and Fabian J. Theis. Multimodal weakly supervised learning to identify disease-specific changes in single-cell atlases. *bioRxiv*, 2024. doi: 10.1101/2024.07.29.605625. URL https://www.biorxiv.org/content/early/2024/07/29/2024.07.29.605625.

[4] Tianyu Liu, Edward De Brouwer, Tony Kuo, Nathaniel Diamant, Alsu Missarova, Hanchen Wang, Minsheng Hao, Tommaso Biancalani, Hector Corrada Bravo, Gabriele Scalia, Aviv Regev, and Graham Heimberg. Learning multi-cellular representations of single-cell transcriptomics data enables characterization of patient-level disease states. *bioRxiv*, 2025. doi: 10.1101/2024.11.18.624166. URL https://www.biorxiv.org/content/early/2025/04/02/2024.11.18.624166.

[5] Romain Lopez, Jeffrey Regier, Michael B. Cole, Michael I. Jordan, and Nir Yosef. Deep generative modeling for single-cell transcriptomics. *Nature Methods*, 15(12):1053–1058, Dec 2018. ISSN 1548-7105. doi: 10.1038/s41592-018-0229-2. URL https://doi.org/10.1038/s41592-018-0229-2.

[6] Stefan Salcher, Gregor Sturm, Lena Horvath, Gerold Untergasser, Georgios Fotakis, Elisa Panizzolo, Agnieszka Martowicz, Georg Pall, Gabriele Gamerith, Martina Sykora, Florian Augustin, Katja Schmitz, Francesca Finotello, Dietmar Rieder, Sieghart Sopper, Dominik Wolf, Andreas Pircher, and Zlatko Trajanoski. High-resolution single-cell atlas reveals diversity and plasticity of tissue-resident neutrophils in non-small cell lung cancer. *bioRxiv*, 2022. doi: 10.1101/2022.05.09.491204. URL https://www.biorxiv.org/content/early/2022/05/10/2022.05.09.491204.

[7] Lisa Sikkema, Ciro Ramírez-Suástegui, Daniel C. Strobl, Tessa E. Gillett, Luke Zappia, Elo Madissoon, Nikolay S. Markov, Laure-Emmanuelle Zaragosi, Yuge Ji, Meshal Ansari, Marie-Jeanne Arguel, Leonie Apperloo, Martin Banchero, Christophe Bécavin, Marijn Berg, Evgeny Chichelnitskiy, Mei-i Chung, Antoine Collin, Aurore C. A. Gay, Janine Gote-Schniering, Baharak Hooshiar Kashani, Kemal Inecik, Manu Jain, Theodore S. Kapellos, Tessa M. Kole, Sylvie Leroy, Christoph H. Mayr, Amanda J. Oliver, Michael von Papen, Lance Peter, Chase J. Taylor, Thomas Walzthoeni, Chuan Xu, Linh T. Bui, Carlo De Donno, Leander Dony, Alen Faiz, Minzhe Guo, Austin J. Gutierrez, Lukas Heumos, Ni Huang, Ignacio L. Ibarra, Nathan D. Jackson, Preetish Kadur Lakshminarasimha Murthy, Mohammad Lotfollahi, Tracy Tabib, Carlos Talavera-López, Kyle J. Travaglini, Anna Wilbrey-Clark, Kaylee B. Worlock, Masahiro Yoshida, Yuexin Chen, James S. Hagood, Ahmed Agami, Peter Horvath, Joakim Lundeberg, Charles-Hugo Marquette, Gloria Pryhuber, Chistos Samakovlis, Xin Sun, Lorraine B. Ware, Kun Zhang, Maarten van den Berge, Yohan Bossé, Tushar J. Desai, Oliver Eickelberg, Naftali Kaminski, Mark A. Krasnow, Robert Lafyatis, Marko Z. Nikolic, Joseph E. Powell, Jayaraj Rajagopal, Mauricio Rojas, Orit Rozenblatt-Rosen, Max A. Seibold, Dean Sheppard, Douglas P. Shepherd, Don D. Sin, Wim Timens, Alexander M. Tsankov, Jeffrey Whitsett, Yan Xu, Nicholas E. Banovich, Pascal Barbry, Thu Elizabeth Duong, Christine S. Falk, Kerstin B. Meyer, Jonathan A. Kropski, Dana Pe'er, Herbert B. Schiller, Purushothama Rao Tata, Joachim L. Schultze, Sara A. Teichmann, Alexander V. Misharin, Martijn C. Nawijn, Malte D. Luecken, Fabian J. Theis, and Lung Biological Network Consortium. An integrated cell atlas of the lung in health and disease. *Nature Medicine*, 29(6):1563–1577, Jun 2023. ISSN 1546-170X. doi: 10.1038/s41591-023-02327-2. URL https://doi.org/10.1038/s41591-023-02327-2.

[8] Marina Terekhova, Amanda Swain, Pavla Bohacova, Ekaterina Aladyeva, Laura Arthur, Anwesha Laha, Denis A. Mogilenko, Samantha Burdess, Vladimir Sukhov, Denis Kleverov, Barbora Echalar, Petr Tsurinov, Roman Chernyatchik, Kamila Husarcikova, and Maxim N. Artyomov. Single-cell atlas of healthy human blood unveils age-related loss of NKG2C+GZMB-CD8+ memory T cells and accumulation of type 2 memory T cells. *Immunity*, 56(12):2836–2854.e9, Dec 2023. ISSN 1074-7613. doi: 10.1016/j.immuni.2023.10.013. URL https://doi.org/10.1016/j.immuni.2023.10.013.

[9] Qinhua Wu, Junxiang Ding, Ruikun He, Lijian Hui, Junwei Liu, and Yixue Li. Exploring phenotype-related single-cells through attention-enhanced representation learning. *bioRxiv*, 2024. doi: 10.1101/2024.10.31.619327. URL https://www.biorxiv.org/content/early/2024/11/01/2024.10.31.619327.