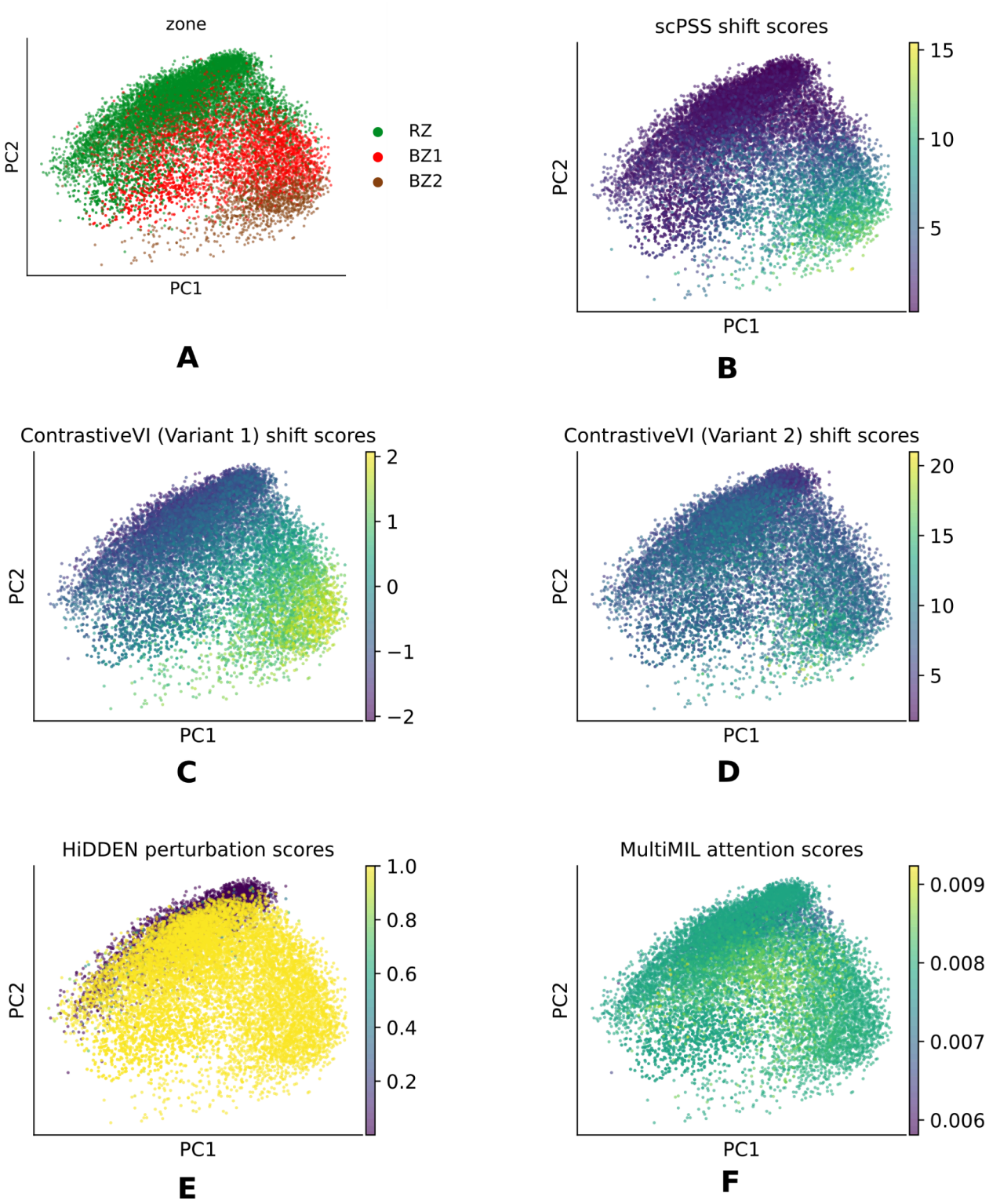
Supplemental Materials

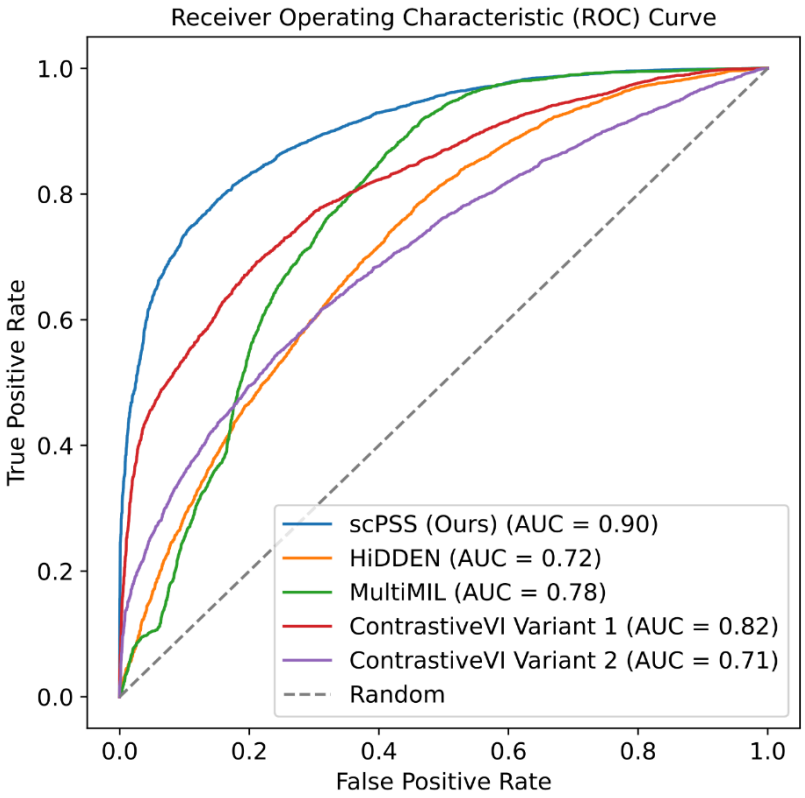
Quantifying Pathological Progression from Single-Cell Transcriptomic Data

Samin Rahman Khan[1], M Saifur Rahman[2], M. Sohel Rahman[2,*], Md. Abul Hassan Samee[3,4,*].

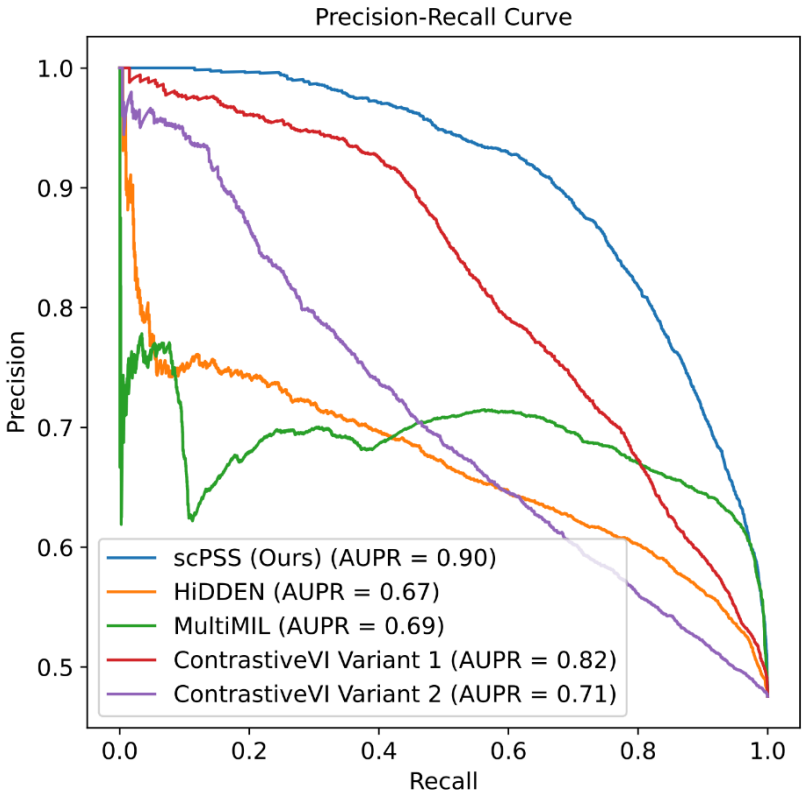
[1] Institute of Information and Communication Technology, Bangladesh University of Engineering and Technology, Dhaka, Bangladesh
[2] Department of Computer Science and Engineering, Bangladesh University of Engineering and Technology, Dhaka, Bangladesh
[3] Department of Integrative Physiology, Baylor College of Medicine, Houston, TX, USA
[4] Lead Contact
* Joint Corresponding Authors (msrahman@cse.buet.ac.bd, and samee@bcm.edu)



Supplemental Fig S1: (A-F) The PC embeddings of all the cells from both healthy (before infarction) and query (after infarction), colored according to regions of nearness to infarcted regions (A), the shift score assigned by different models: scPSS (Ours) (B), ContrastiveVI (Variant 1) (C), ContrastiveVI (Variant 2) (D), HiDDEN (E), and MultiMIL (F) for dataset (Calcagno et al., 2022).

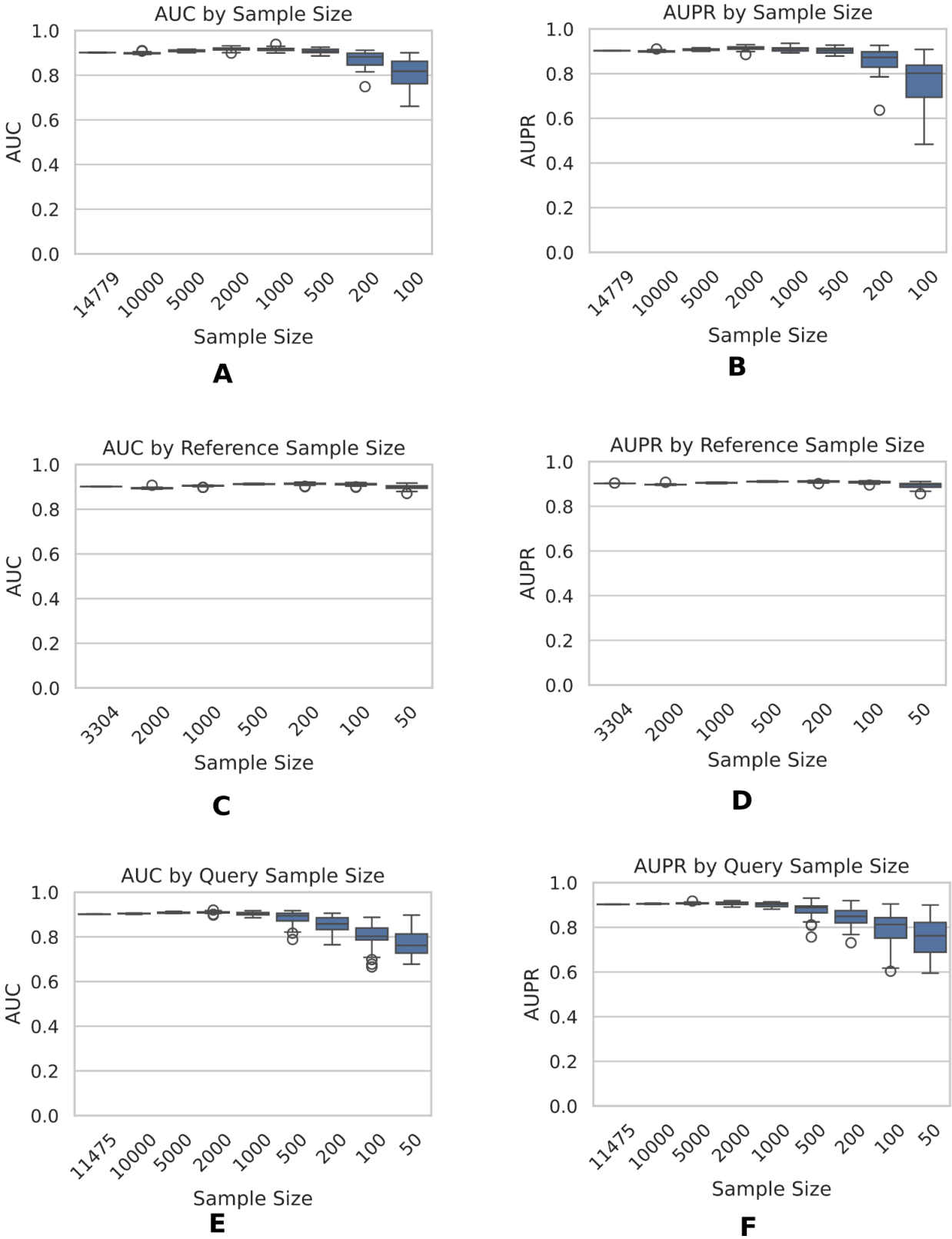


A

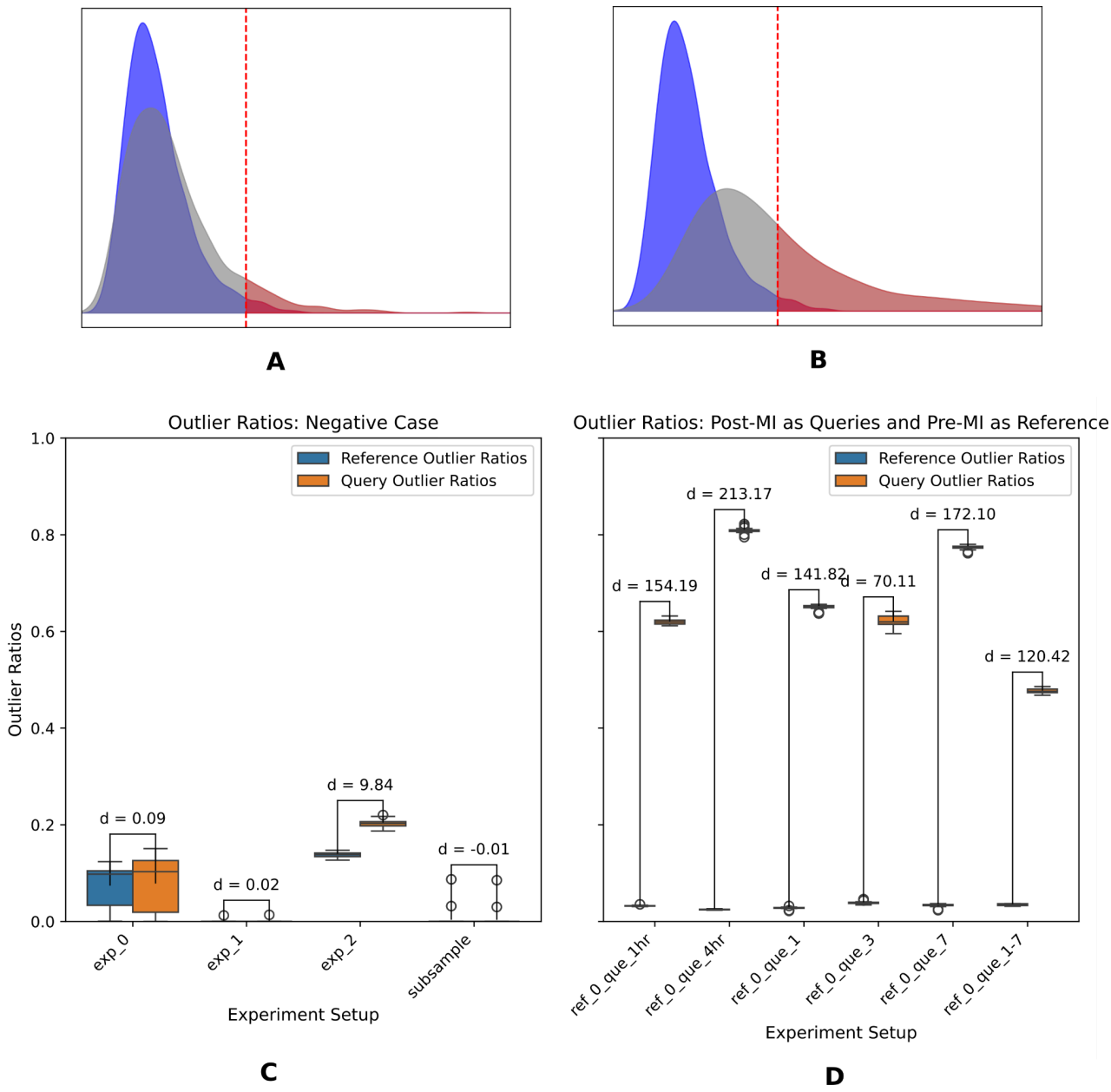


B

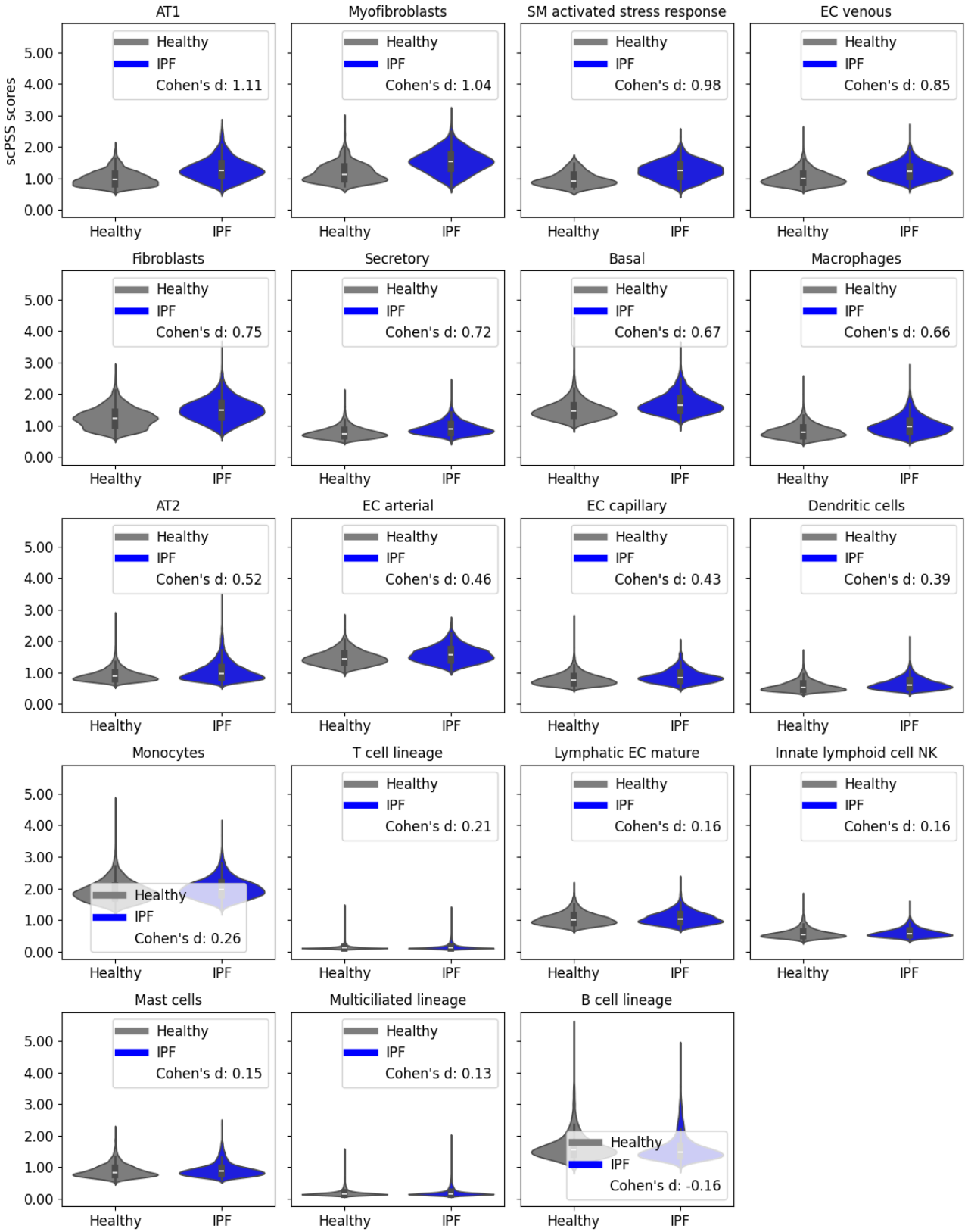
Supplemental Fig S2: (A-B) The Receiver Operating Characteristic (ROC) (A) and Precision-Recall (B) curves for the pathological scores provided by different models on the query dataset of (Calcagno et al., 2022).



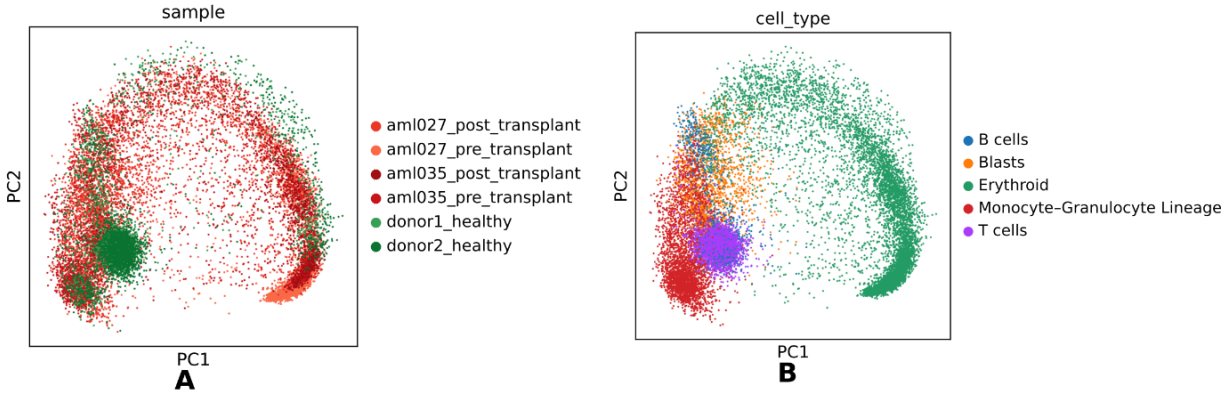
Supplemental Fig S3. Performance of scPSS at different sample sizes. (A-B) AUC and AUPR metrics of scPSS on (Calcagno et al., 2022) when the concatenated reference and query datasets were subsampled to different sizes. **(C-D)** AUC and AUPR metrics when only the reference dataset was subsampled and the entire query dataset was used, **(E-F)** AUC and AUPR metrics when only the query dataset was subsampled and the entire reference dataset was used.

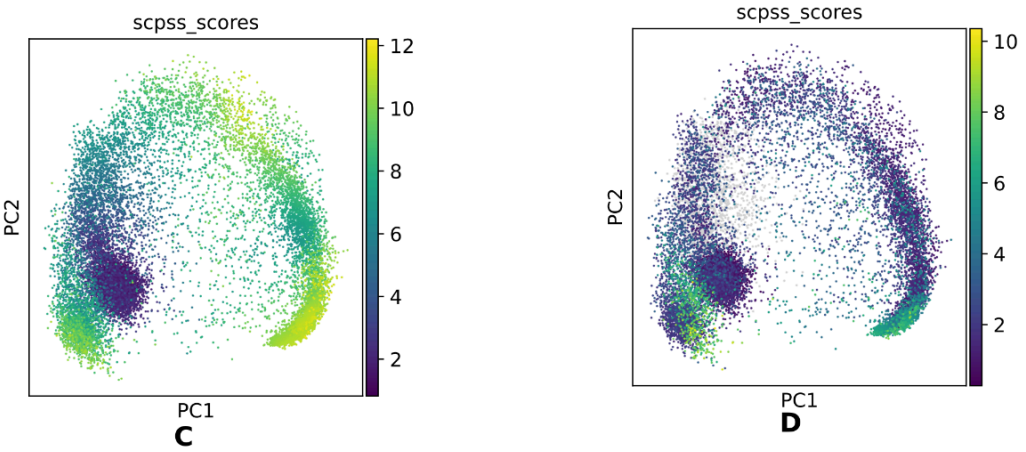


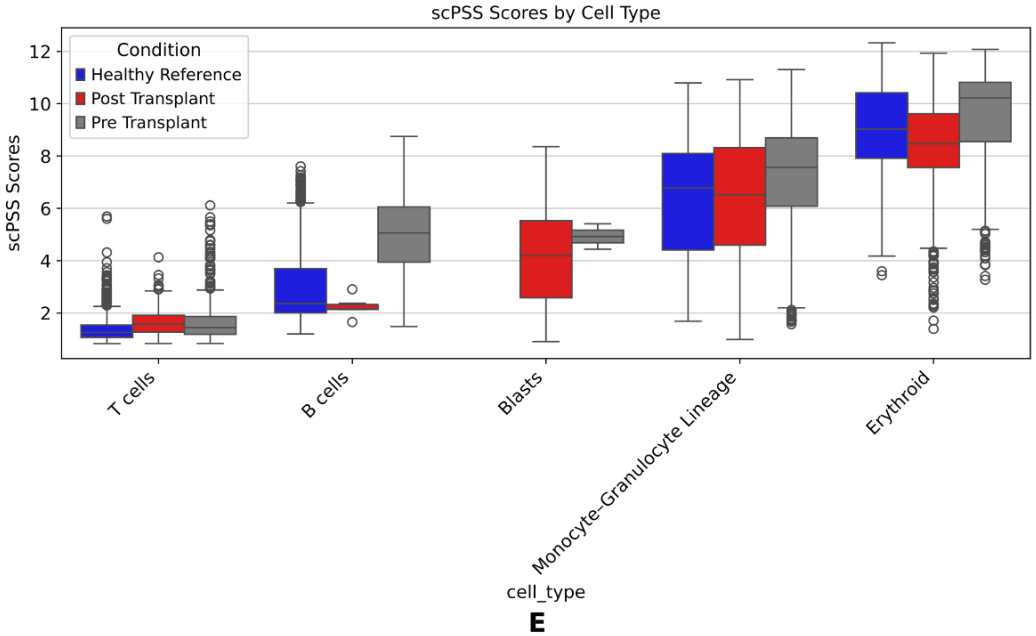
Supplemental Fig S4. Comparing reference and query outlier ratios in disease-positive and negative control cases. (A) In scPSS, the outlier ratios are expected to be nearly zero in the reference and query when there are few or no true diseased cells in the query (negative case), **(B)** whereas when diseased cells are present (positive case), the outlier ratio in the query should be much greater than the reference. **(C)** While testing on the negative control cases in the dataset (Calcagno et al., 2022), we found the outlier ratios in the reference and query are nearly zero, **(D)** whereas in the positive test cases, the outlier ratios are much higher in the query than in the reference.

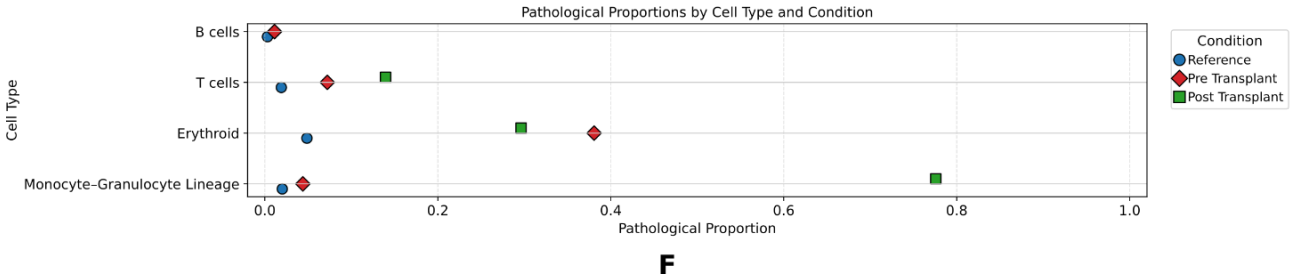


Supplemental Fig S5. The shift scores of reference and query cells when scPSS is applied separately on each cell type. The reference contains cells from healthy individuals, and the query contains cells from pulmonary fibrosis patients (Sikkema et al., 2023). The figures also show the Cohen's d measure between the reference and query scores.

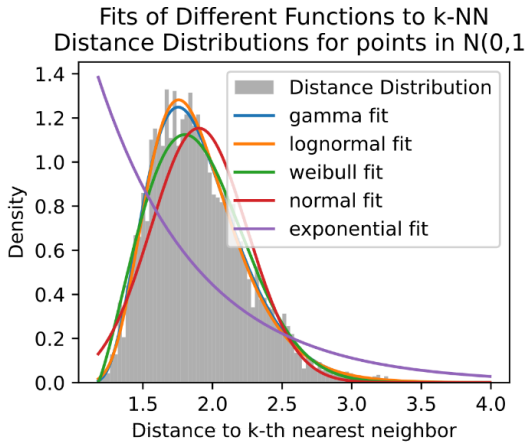

sample
PC2
PC1
aml027_post_transplant
aml027_pre_transplant
aml035_post_transplant
aml035_pre_transplant
donor1_healthy
donor2_healthy
A
cell_type
PC2
PC1
B cells
Blasts
Erythroid
Monocyte–Granulocyte Lineage
T cells
B


scpss_scores
PC2
PC1
12
10
8
6
4
2
C
scpss_scores
PC2
PC1
10
8
6
4
2
D

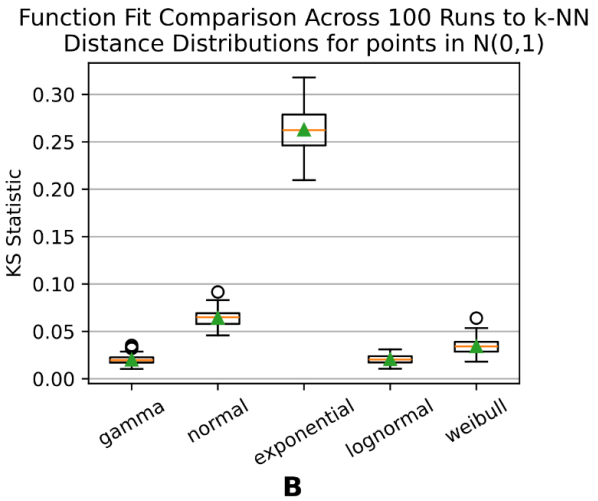

scPSS Scores by Cell Type
Condition
Healthy Reference
Post Transplant
Pre Transplant
scPSS Scores
12
10
8
6
4
2
T cells
B cells
Blasts
Monocyte-Granulocyte Lineage
Erythroid
cell_type
E


Pathological Proportions by Cell Type and Condition
Cell Type
B cells
T cells
Erythroid
Monocyte–Granulocyte Lineage
0.0
0.2
0.4
0.6
0.8
1.0
Pathological Proportion
Condition
Reference
Pre Transplant
Post Transplant
F

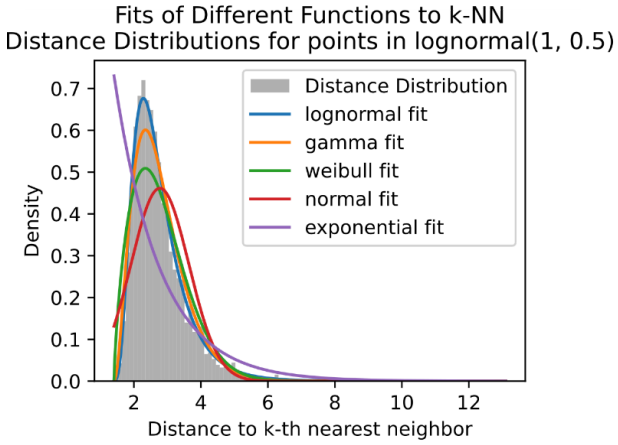
Supplemental Fig S6. (A-D) The PC embeddings of all the cells from both healthy and query (leukemia patients), colored according to source individual (A), cell type (B), and pathological scores assigned by scPSS when applied with larger k values (C) and default k-values (D). **(E)** The pathological scores of cells from each cell type in healthy, pre-transplant, and post-transplant individuals when scPSS is applied with large k values for detecting shifts across cell types. **(F)** The pathological proportion of each cell type in healthy, pre-transplant, and post-transplant individuals when scPSS is applied with default k-values for each cell type individually.



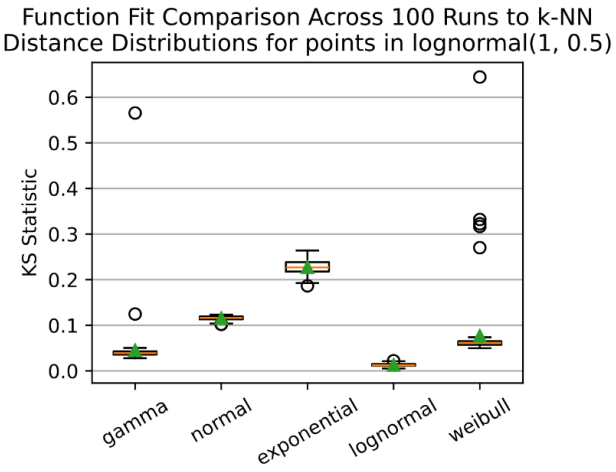
A



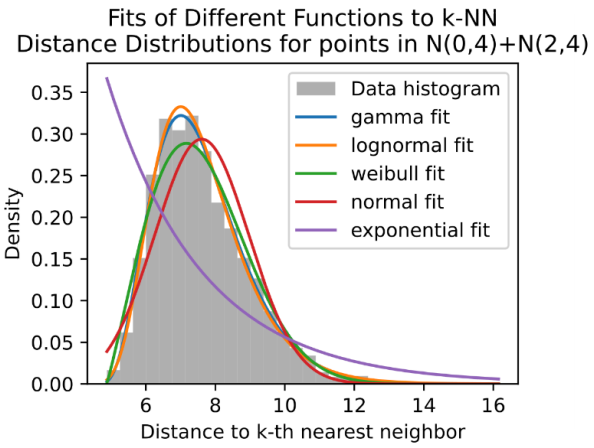
B



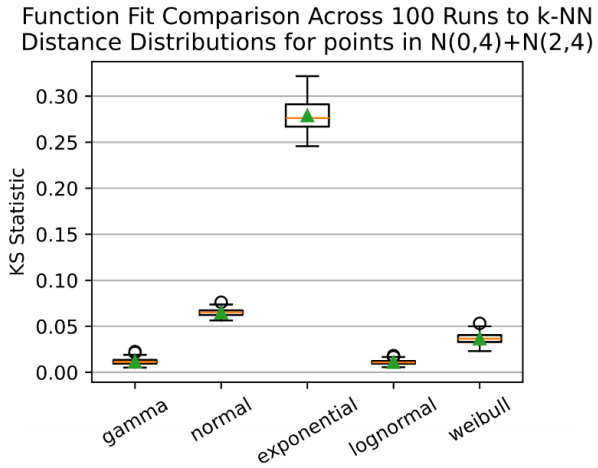
C



D

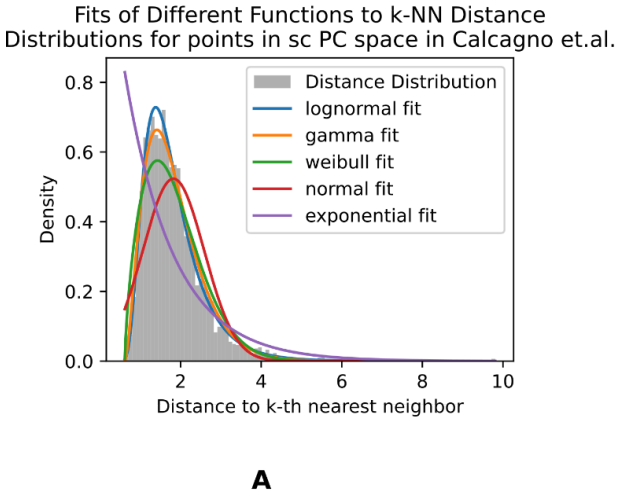


E

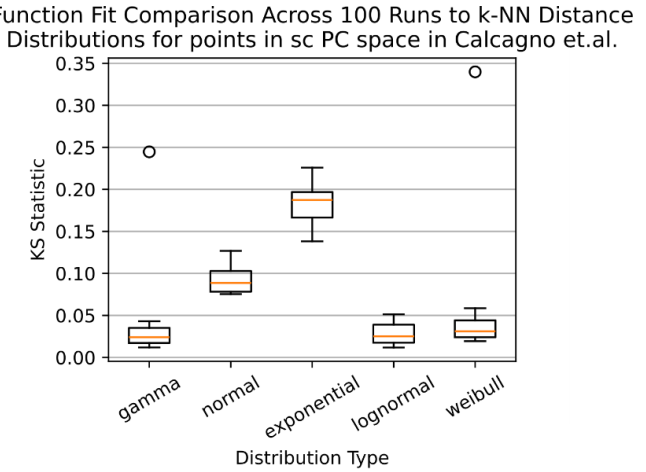


F

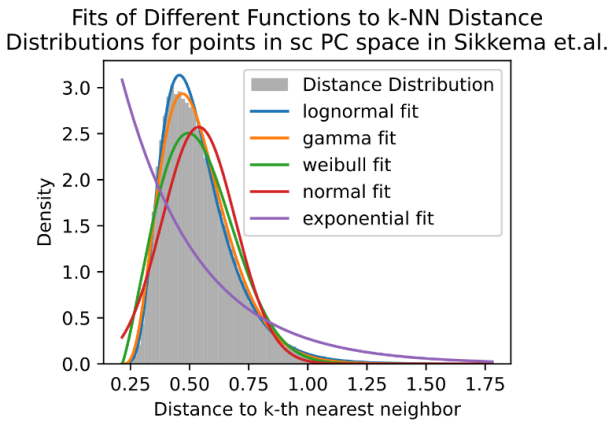
Supplemental Fig S7: Goodness-of-fit analysis of various probability distribution functions to a distribution of distances to the 5th nearest neighbor of 5,000 random points in 10-dimensional space **(A)** Distance distribution and fitted distribution functions when points are drawn from $N(0, 1)$. **(B)** Boxplot of KS statistic when this is repeated 100 times. (Median values are gamma: 0.020, lognormal: 0.020, Weibull: 0.034, normal: 0.065, and exponential: 0.262.) **(C)** Distance distribution and fitted distribution functions when points are drawn from $\mathrm{lognormal}(1,\ 0.5)$. **(D)** Boxplot of KS statistic when this is repeated 100 times. (Median values are lognormal: 0.012, gamma: 0.039, Weibull: 0.062, normal: 0.116, and exponential: 0.227.) **(E)** Distance distribution and fitted distribution functions when points are drawn from $N(0, 4) + N(2, 4)$. **(F)** Boxplot of KS statistic when this is repeated 100 times. (Median values are gamma: 0.011, lognormal: 0.011, Weibull: 0.037, normal: 0.065, and exponential: 0.276.)



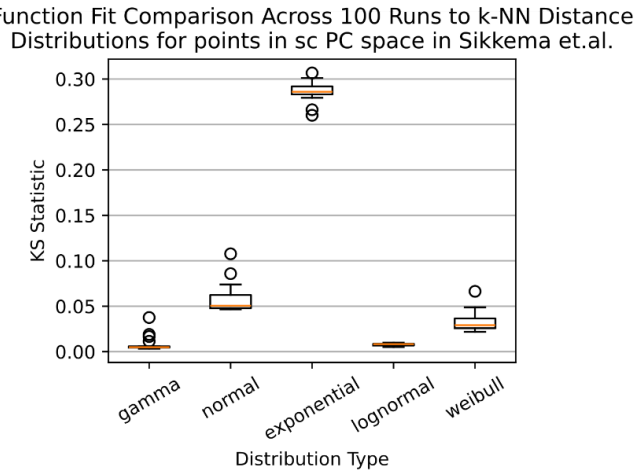
A



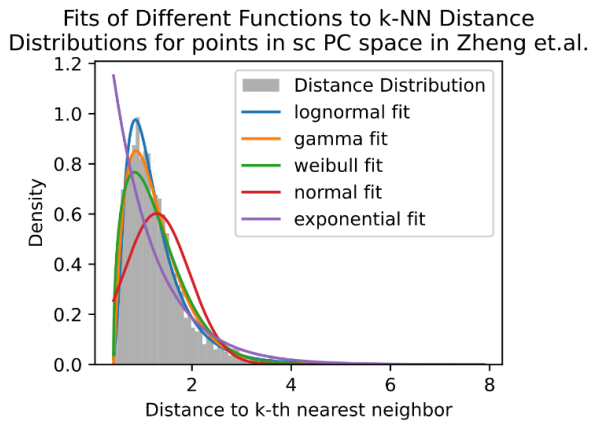
B



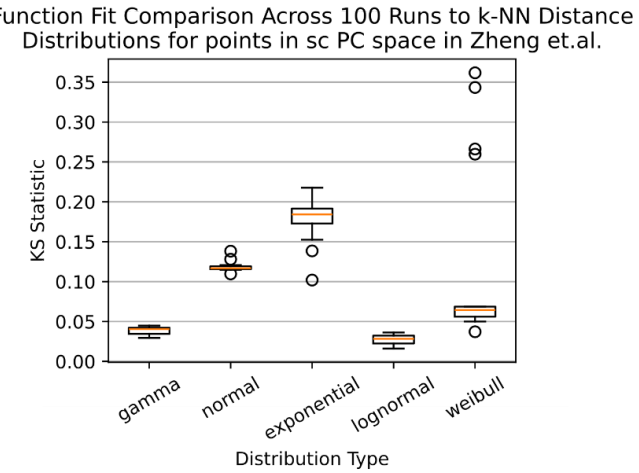
C



D



E



F

Supplemental Fig S8. Goodness-of-fit analysis of various probability distribution functions to a distribution of distances to the 5th nearest neighbor of principal component (PC) embeddings of single-cell transcriptomic datasets **(a)** Distance distribution and fitted distribution functions when top 6 PC embeddings of remote zone Cardiomyocyte cells from Calcagno et al (Calcagno et al., 2022) are used. **(b)** Boxplot of KS statistic when this is repeated with top 4 to 20 top PC embeddings (Median values are gamma: 0.038, lognormal: 0.028, Weibull: 0.051, normal: 0.092, and exponential: 0.182.) **(c)** Distance distribution and fitted distribution functions when top 6 PC embeddings of healthy Macrophage cells from Sikkema et al. (Sikkema et al., 2023) are used. **(d)** Boxplot of KS statistic when this is repeated with top 4 to 20 top PC embeddings (Median

values are gamma: 0.009, lognormal: 0.008, Weibull: 0.033, normal: 0.058, and exponential: 0.287.) **(e)** Distance distribution and fitted distribution functions when top 6 PC embeddings of healthy cells from Zhang et al. (Zheng et al., 2017) are used. **(f)** Boxplot of KS statistic when this is repeated with top 4 to 20 top PC embeddings (Median values are gamma: 0.039, lognormal: 0.027, Weibull: 0.117, normal: 0.119, and exponential: 0.177.)

Query Dataset	Harmony	Scanorama	scVI	No Integration
1hr	**0.9172 ± 0.0014**	0.7166 ± 0.0000	0.8629 ± 0.0142	0.9033 ± 0.0000
4hr	0.9403 ± 0.0006	0.9298 ± 0.0000	0.8714 ± 0.0083	**0.9485 ± 0.0000**
Day 1	**0.9509 ± 0.0002**	0.9242 ± 0.0000	0.8721 ± 0.0147	0.9443 ± 0.0000
Day 3	**0.9323 ± 0.0006**	0.6658 ± 0.0000	0.8104 ± 0.0234	0.9155 ± 0.0000
Day 7	**0.8515 ± 0.0027**	0.7381 ± 0.0000	0.7836 ± 0.0100	0.7926 ± 0.0000
all post-MI	0.9016 ± 0.0004	0.8253 ± 0.0000	0.8483 ± 0.0112	**0.9109 ± 0.0000**

Supplemental Table S1. Comparison of AUC measures for predicting healthy and damaged cells, with pre-MI (Myocardial Infarction) dataset as reference and cells collected at different time points after MI as query dataset, using different integration methods in the scPSS framework. Results show mean ± 95% confidence interval from 25 independent runs

Query Dataset	Harmony	Scanorama	scVI	No Integration
1hr	0.8417 ± 0.0033	0.6639 ± 0.0000	0.7820 ± 0.0195	**0.8495 ± 0.0000**
4hr	0.9590 ± 0.0004	0.9559 ± 0.0000	0.8905 ± 0.0082	**0.9646 ± 0.0000**
Day 1	**0.9591 ± 0.0002**	0.9389 ± 0.0000	0.8923 ± 0.0148	0.9536 ± 0.0000
Day 3	**0.8946 ± 0.0012**	0.5727 ± 0.0000	0.7444 ± 0.0261	0.8900 ± 0.0000
Day 7	**0.8243 ± 0.0040**	0.6501 ± 0.0000	0.7375 ± 0.0127	0.7706 ± 0.0000
all post-MI	0.9029 ± 0.0004	0.8174 ± 0.0000	0.8382 ± 0.0138	**0.9053 ± 0.0000**

Supplemental Table S2. Comparison of AUPR measures for predicting healthy and damaged cells, with pre-MI (Myocardial Infarction) dataset as reference and cells collected at different time points after MI as query dataset, using different integration methods in the scPSS framework. Results show mean ± 95% confidence interval from 25 independent runs.

Query Dataset	Harmony	Scanorama	scVI	No Integration
1hr	**0.7372 ± 0.0015**	0.5845 ± 0.0000	0.6536 ± 0.0383	0.5667 ± 0.0000
4hr	**0.8593 ± 0.0011**	0.8340 ± 0.0000	0.7726 ± 0.0166	0.7659 ± 0.0000
Day 1	**0.8685 ± 0.0010**	0.8589 ± 0.0000	0.6142 ± 0.0383	0.7154 ± 0.0000
Day 3	**0.7704 ± 0.0025**	0.4217 ± 0.0000	0.4772 ± 0.0351	0.6294 ± 0.0000
Day 7	0.7065 ± 0.0007	**0.7109 ± 0.0000**	0.6086 ± 0.0157	0.6745 ± 0.0000
all post-MI	**0.8074 ± 0.0007**	0.7397 ± 0.0000	0.5603 ± 0.0613	0.6958 ± 0.0000

Supplemental Table S3. Comparison of F1-score measures for predicting healthy and damaged cells, with pre-MI (Myocardial Infarction) dataset as reference and cells collected at different time points after MI as query dataset, using different integration methods in the scPSS framework. Results show mean ± 95% confidence interval from 25 independent runs.

Distance	1hr	4hr	Day 1	Day 3	Day 7	All Post-MI	Mean Rank
chebyshev	0.921	0.946	0.951	0.933	0.843	0.902	2.5
euclidean	0.917	0.939	0.951	0.932	0.85	0.901	4.167
mse	0.917	0.939	0.951	0.932	0.85	0.901	4.167
minkowski	0.917	0.939	0.951	0.932	0.85	0.901	4.167
sqeuclidean	0.917	0.939	0.951	0.932	0.85	0.901	4.167
mae	0.913	0.926	0.95	0.932	0.853	0.901	6.333
cityblock	0.913	0.926	0.95	0.932	0.853	0.901	6.333
mahalanobis	0.915	0.946	0.948	0.928	0.851	0.881	6.667
seuclidean	0.915	0.947	0.948	0.931	0.836	0.878	7.167
braycurtis	0.883	0.908	0.917	0.904	0.747	0.889	10
cosine	0.852	0.89	0.912	0.814	0.772	0.895	10.667
canberra	0.867	0.861	0.905	0.823	0.727	0.849	11.667

Supplemental Table S4. Comparison of AUC measures for predicting healthy and damaged cells, with pre-MI (Myocardial Infarction) dataset as reference and cells collected at different time points after MI as query dataset, using different distance metrics in the scPSS framework. Results show mean ± 95% confidence interval from 25 independent runs.

Distance	1hr	4hr	Day 1	Day 3	Day 7	All Post-MI	Mean Rank
chebyshev	0.859	0.963	0.959	0.897	0.811	0.903	2.167
euclidean	0.841	0.958	0.959	0.895	0.822	0.902	4
mse	0.841	0.958	0.959	0.895	0.822	0.902	4
minkowski	0.841	0.958	0.959	0.895	0.822	0.902	4
sqeuclidean	0.841	0.958	0.959	0.895	0.822	0.902	4
mae	0.821	0.945	0.958	0.894	0.824	0.902	6.333
cityblock	0.821	0.945	0.958	0.894	0.824	0.902	6.333
mahalanobis	0.837	0.959	0.957	0.888	0.815	0.888	7
seuclidean	0.839	0.961	0.957	0.893	0.791	0.887	7.167
braycurtis	0.754	0.928	0.907	0.832	0.626	0.871	10.5
cosine	0.705	0.909	0.9	0.654	0.679	0.878	11
canberra	0.762	0.848	0.869	0.742	0.615	0.823	11.5

Supplemental Table S5. Comparison of AUPR measures for predicting healthy and damaged cells, with pre-MI (Myocardial Infarction) dataset as reference and cells collected at different time points after MI as query dataset, using different distance metrics in the scPSS framework. Results show mean ± 95% confidence interval from 25 independent runs.

References

Calcagno, D. M., Taghdiri, N., Ninh, V. K., Mesfin, J. M., Toomu, A., Sehgal, R., Lee, J., Liang, Y., Duran, J. M., Adler, E., Christman, K. L., Zhang, K., Sheikh, F., Fu, Z., & King, K. R. (2022). Single-cell and spatial transcriptomics of the infarcted heart define the dynamic onset of the border zone in response to mechanical destabilization. *Nature Cardiovascular Research*, *1*(11), 1039–1055.

Sikkema, L., Ramírez-Suástegui, C., Strobl, D. C., Gillett, T. E., Zappia, L., Madissoon, E., Markov, N. S., Zaragosi, L.-E., Ji, Y., Ansari, M., Arguel, M.-J., Apperloo, L., Banchero, M., Bécavin, C., Berg, M., Chichelnitskiy, E., Chung, M.-I., Collin, A., Gay, A. C. A., … Theis, F. J. (2023). An integrated cell atlas of the lung in health and disease. *Nature Medicine*, *29*(6), 1563–1577.

Zheng, G. X. Y., Terry, J. M., Belgrader, P., Ryvkin, P., Bent, Z. W., Wilson, R., Ziraldo, S. B., Wheeler, T. D., McDermott, G. P., Zhu, J., Gregory, M. T., Shuga, J., Montesclaros, L., Underwood, J. G., Masquelier, D. A., Nishimura, S. Y., Schnall-Levin, M., Wyatt, P. W., Hindson, C. M., … Bielas, J. H. (2017). Massively parallel digital transcriptional profiling of single cells. *Nature Communications*, *8*, 14049.