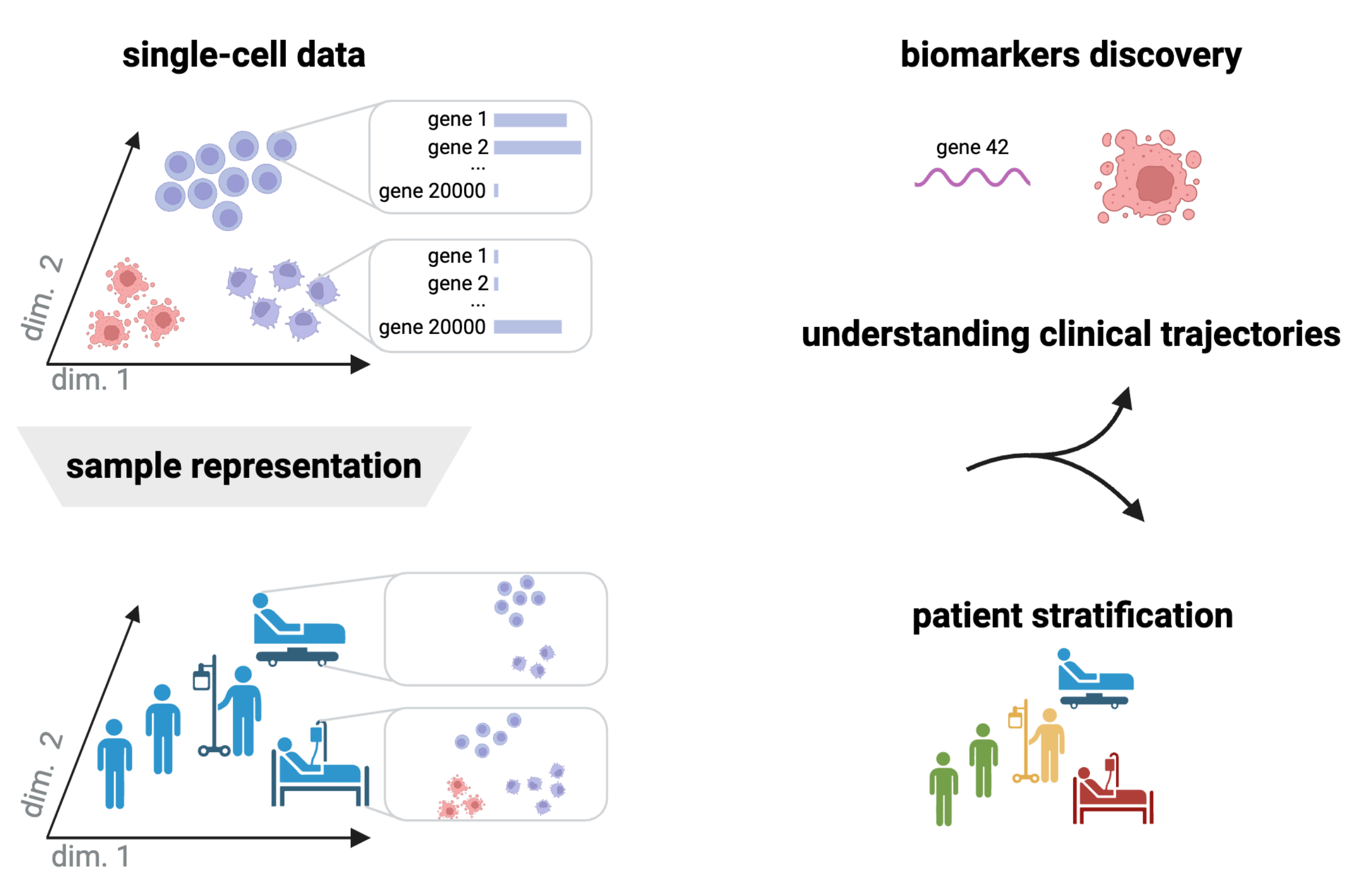
# Benchmarking Sample Representations from Single-Cell Data: Metrics for Biologically Meaningful Embeddings

Vladimir A. Shitov[1,2], Mohammad Dehkordi[3], Malte D. Luecken[1,2]

[1] Department of Computational Health, Institute of Computational Biology, Helmholtz Munich, Munich, Germany
[2] Comprehensive Pneumology Center (CPC) with the CPC-M bioArchive and Institute of Lung Health and Immunity (LHI), Helmholtz Munich; Member of the German Center for Lung Research (DZL), Munich, Germany
[3] TUM School of Computation, Information and Technology

## From cells to humans with sample representation

**Single-cell data** enables understanding cell biology with unprecedented resolution

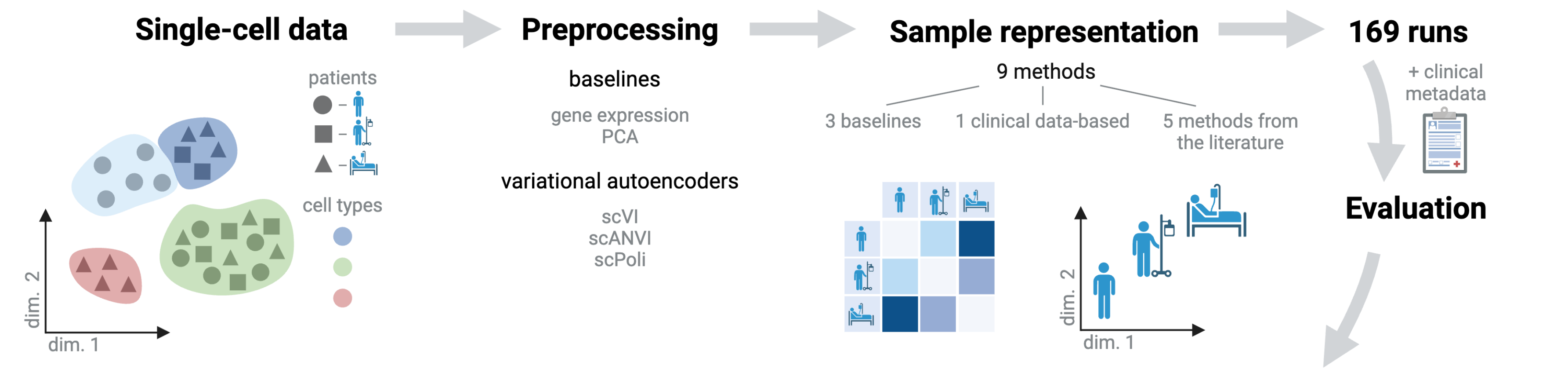


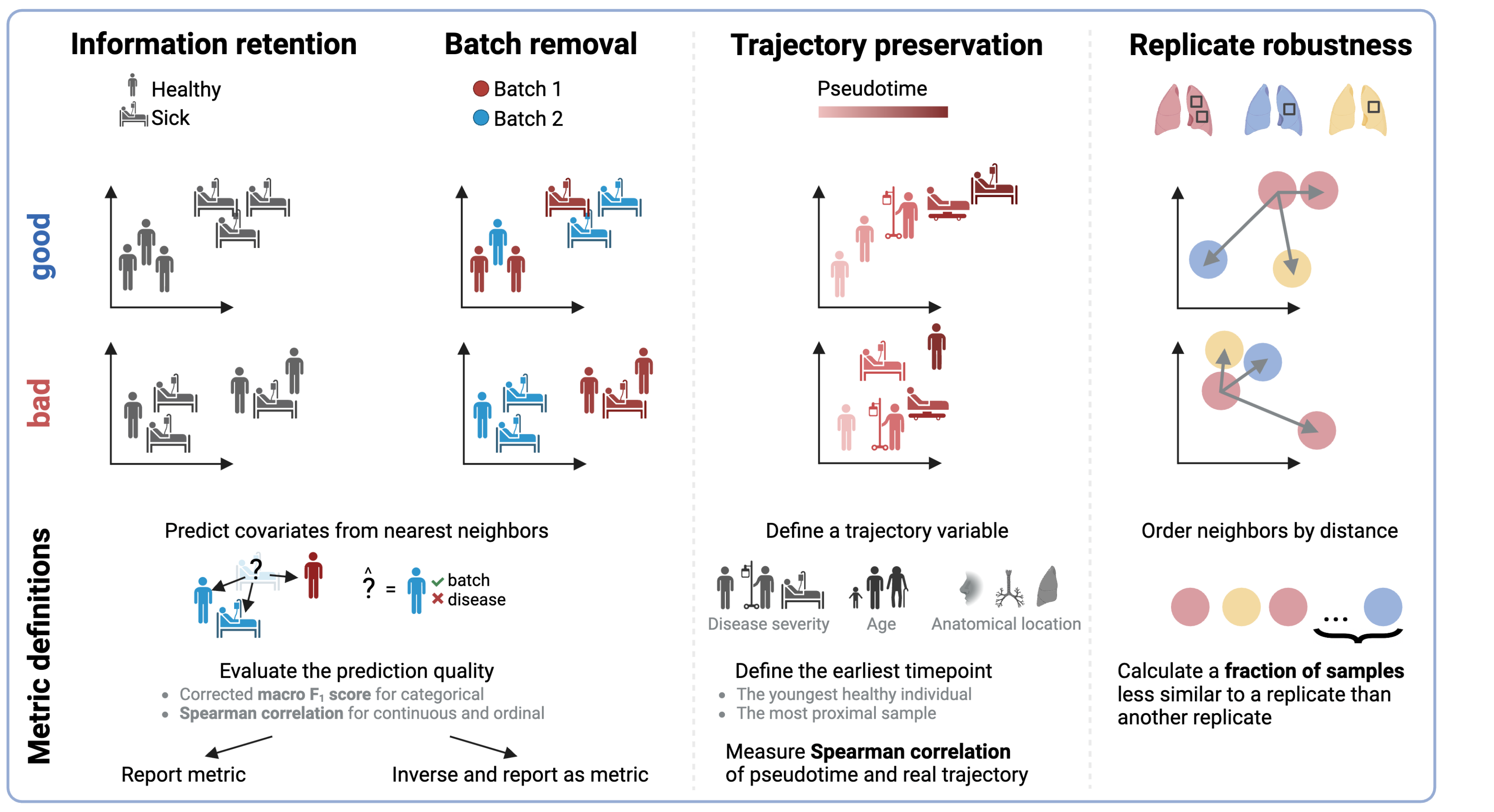
With datasets growing in size[1], it is now possible to analyze **sample-level variation**.

Several methods for sample representation are published[2-8], but a **consistent comparison** and **biologically relevant metrics** are still lacking.

We present a Single-cell-based Patient Representation Evaluation (SPARE) benchmark. We defined **4 application-inspired metrics**, and used these to compare **9 sample representation methods** on **5 datasets** from **2 tissues** comprising **4.5 million cells** and **1668 samples**.

## Establishing best practices for sample representation



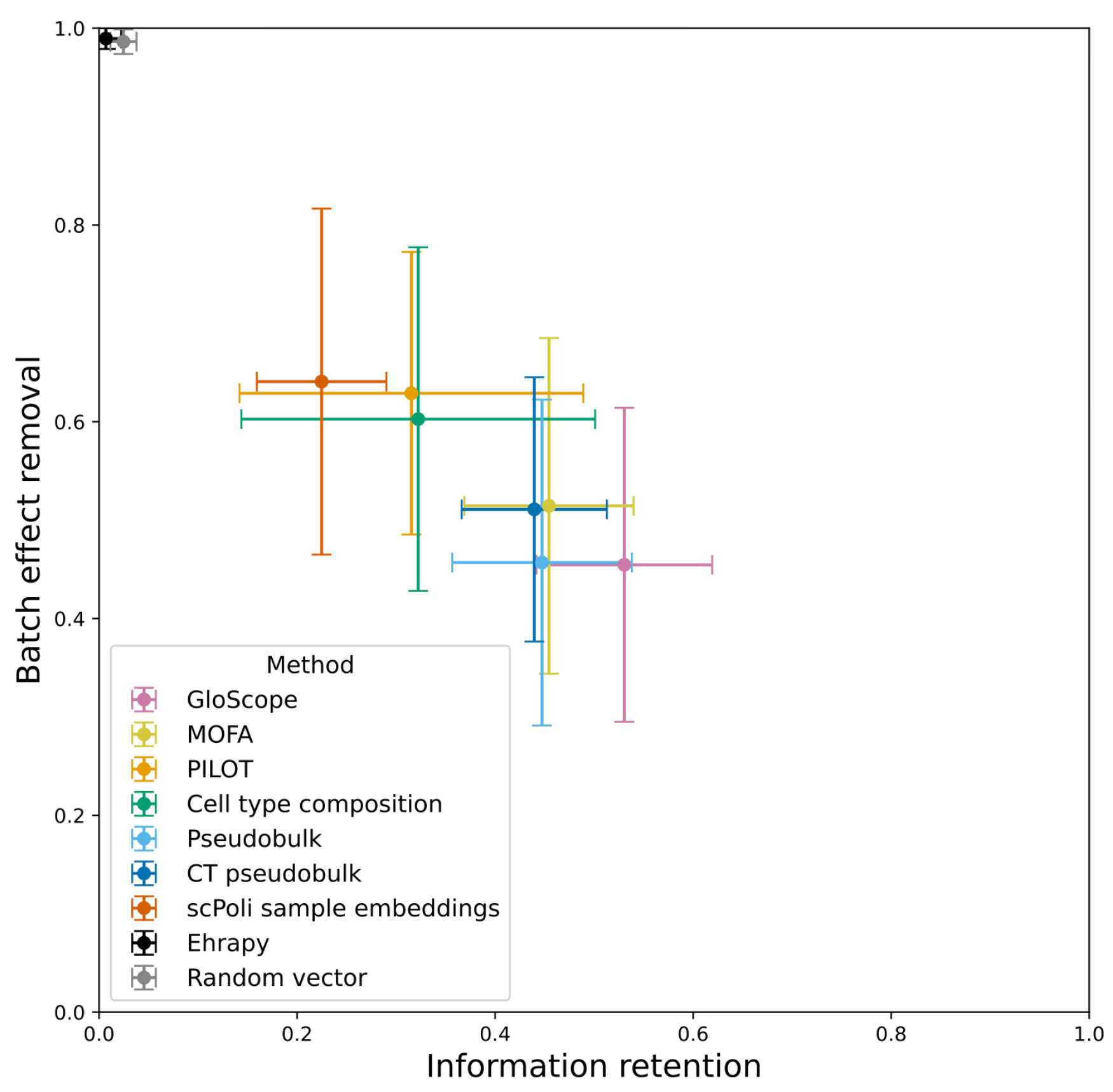

Information retention
Healthy
Sick
Batch removal
Batch 1
Batch 2
Trajectory preservation
Pseudotime
Replicate robustness
good
bad
Metric definitions
Predict covariates from nearest neighbors
?
$\hat{?}$ =
✓ batch
✗ disease
Evaluate the prediction quality
Corrected macro $F_1$ score for categorical
Spearman correlation for continuous and ordinal
Report metric
Inverse and report as metric
Define a trajectory variable
Disease severity
Age
Anatomical location
Define the earliest timepoint
The youngest healthy individual
The most proximal sample
Measure Spearman correlation of pseudotime and real trajectory
Order neighbors by distance
...
Calculate a fraction of samples less similar to a replicate than another replicate

## Density-based method, GloScope, often performs the best

| Dataset | Representation | Information retention | Batch removal | Replicate robustness | Trajectory preservation | Total |
|---|---|---|---|---|---|---|
| COMBAT | scPoli – GloScope | **0.37** | 0.57 | – | **0.79** | **0.58** |
| | scANVI$_s$ – GloScope | 0.31 | 0.47 | – | 0.71 | 0.50 |
| | scANVI$_b$ – CT pseudobulk | 0.24 | **0.70** | – | 0.62 | 0.48 |
| | counts – MOFA | 0.23 | 0.53 | – | 0.03 | 0.21 |
| Stephenson | scVI$_s$ – MOFA | **0.48** | 0.47 | – | **0.45** | **0.47** |
| | scVI$_b$ – MOFA | 0.47 | 0.44 | – | 0.45 | 0.45 |
| | scANVI$_b$ – MOFA | 0.41 | 0.42 | – | 0.45 | 0.43 |
| | scPoli – Pseudobulk | 0.06 | **0.58** | – | 0.07 | 0.17 |
| Onek1k | scPoli – GloScope | 0.60 | 0.55 | – | 0.41 | **0.52** |
| | scANVI$_s$ – GloScope | **0.63** | 0.47 | – | **0.42** | 0.51 |
| | Cell type composition | 0.54 | 0.65 | – | 0.40 | 0.50 |
| | scPoli – MOFA | 0.00 | **0.97** | – | 0.00 | 0.20 |
| HLCA | scVI$_b$ – Pseudobulk | **0.54** | 0.36 | – | **0.81** | **0.61** |
| | scANVI$_b$ – Pseudobulk | 0.48 | 0.46 | – | 0.81 | 0.61 |
| | scVI$_s$ – Pseudobulk | 0.54 | 0.37 | – | 0.80 | 0.61 |
| | Ehrapy | 0.00 | **0.98** | – | 0.07 | 0.23 |
| COPD | PCA – GloScope | **0.54** | 0.69 | **0.98** | 0.26 | **0.61** |
| | scANVI$_b$ – CT pseudobulk | 0.48 | 0.61 | 0.96 | **0.33** | 0.59 |
| | scANVI$_b$ – GloScope | 0.47 | 0.61 | 0.99 | 0.32 | 0.59 |
| | Random vector$_{10}$ | 0.01 | **0.97** | 0.37 | 0.02 | 0.25 |

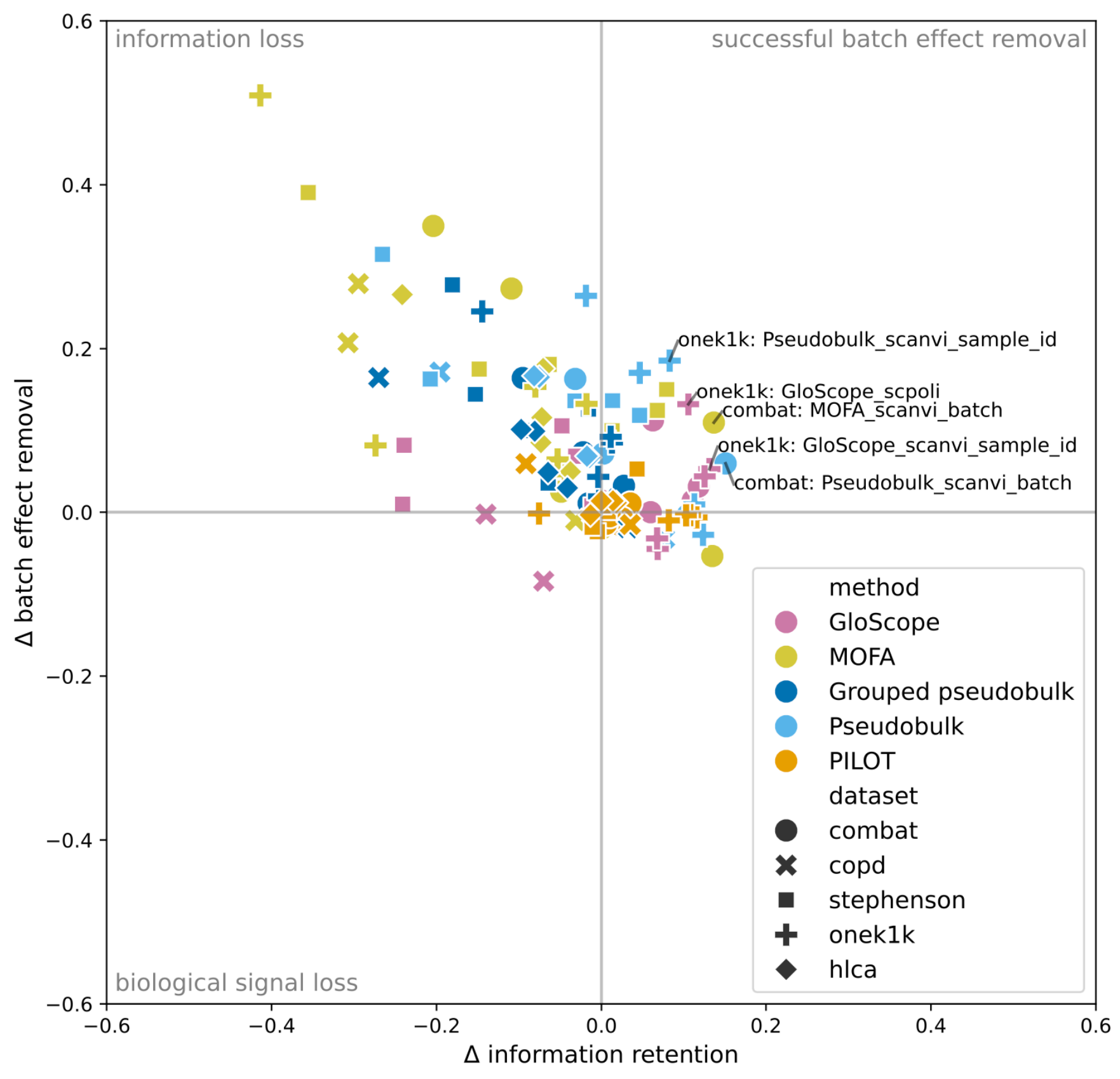
Top **3 best** and **1 worst representation** per dataset according to the total score. Representation names consist of input space (where applicable) and sample representation method. sc[AN]VI$_{b/s}$ refers to a sc[AN]VI model trained with the batch covariate “batch” or "sample" to integrate the data. Total score is weighted average with the Batch removal score weighted half as much as others.

## There is a trade-off between Information retention and Batch removal



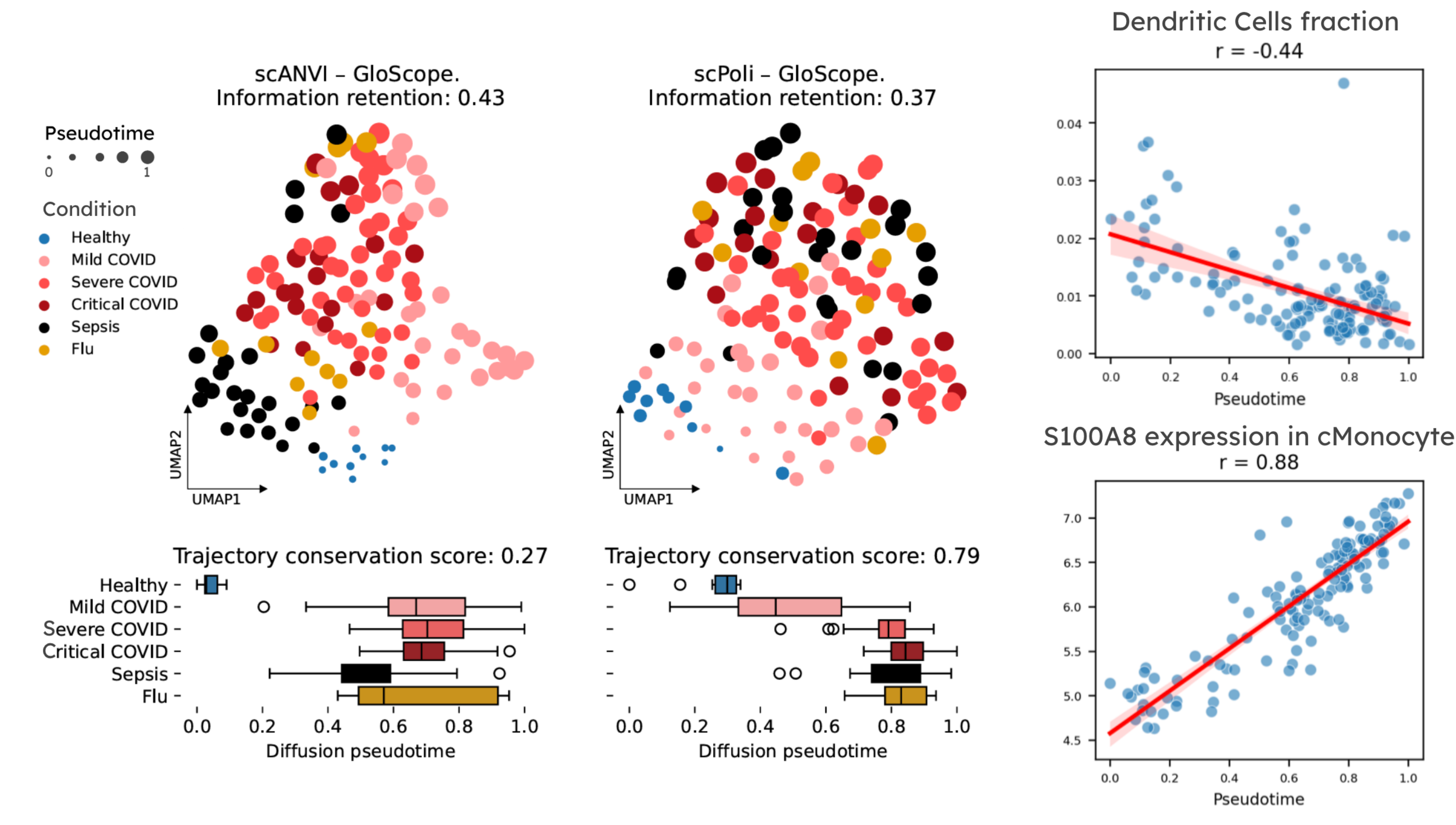
Mean and standard deviation across datasets are shown for the best sample representation from each method

## Batch correction improves sample representation



Differences in information retention and batch removal scores for each method in comparison to PCA-based representation with the same method.

## Best COMBAT representation reveals markers of COVID-19 severity



Left: best representation according to the Information retention score, middle: best representation according to the Trajectory conservation score (and overall). Right: top pseudotime-correlated cell type fraction and gene expression in a cell type.

## Datasets

| Dataset | COMBAT | Stephenson | Onek1k | HLCA | COPD |
|---|---|---|---|---|---|
| **#donors** | 140 | 130 | **982** | 344 | 61 (72 samples) |
| **#cells** | 784k | 639k | 1.25M | **1.68M** | 176k |
| **Tissue** | PBMC | PBMC | PBMC | Lung and airways | Lung parenchyma |
| **Relevant information** | Condition, Severity, Death in 28 days, Duration | Condition, Severity, Outcome, Duration | Age | Tissue anatomical location, Condition, Smoking status | Severity, Lung function tests, Progression |
| **Technical covariates** | Institute, Pool ID | Site | Sex | Suspension type, Fresh or frozen, Sequencing platform, Assay | Batch, Lung lobe, Cancer |

## Code and paper

## References

1. Hrovatin, K., Sikkema, L., Shitov, V. A., Heimberg, G., Shulman, M., Oliver, A. J., Mueller, M. F., Ibarra, I. L., Wang, H., Ramírez-Suástegui, C., He, P., Schaar, A. C., Teichmann, S. A., Theis, F. J., & Luecken, M. D. (2025). Considerations for building and using integrated single-cell atlases. Nature methods, 22(1), 41–57. https://doi.org/10.1038/s41592-024-02532-y
2. De Donno, C. et al. Population-level integration of single-cell datasets enables multi-scale analysis across samples. Nat Methods 20, 1683–1692 (2023).
3. Boyeau, P. et al. Deep generative modeling for quantifying sample-level heterogeneity in single-cell omics. 2022.10.04.510898 Preprint at https://doi.org/10.1101/2022.10.04.510898 (2022).
4. Argelaguet, R. et al. MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data. Genome Biology 21, 111 (2020).
5. Chen, W. S. et al. Uncovering axes of variation among single-cell cancer specimens. Nat Methods 17, 302–310 (2020).
6. Joodaki, M. et al. Detection of PatIent-Level distances from single cell genomics and pathomics data with Optimal Transport (PILOT). Molecular Systems Biology 20, 57–74 (2024).
7. Tong, A. et al. Diffusion Earth Mover’s Distance and Distribution Embeddings. Preprint at http://arxiv.org/abs/2102.12833 (2021).
8. Wang, H., Torous, W., Gong, B. et al. Visualizing scRNA-Seq data at population scale with GloScope. Genome Biol 25, 259 (2024). https://doi.org/10.1186/s13059-024-03398-1
9. Heumos, L., Ehmele, P., Treis, T. et al. An open-source framework for end-to-end analysis of electronic health record data. Nat Med 30, 3369–3380 (2024). https://doi.org/10.1038/s41591-024-03214-0
10. COvid-19 Multi-omics Blood ATlas (COMBAT) Consortium. Electronic address: julian.knight@well.ox.ac.uk, & COvid-19 Multi-omics Blood ATlas (COMBAT) Consortium (2022). A blood atlas of COVID-19 defines hallmarks of disease severity and specificity. Cell, 185(5), 916–938.e58. https://doi.org/10.1016/j.cell.2022.01.012
11. Stephenson, E., Reynolds, G., Botting, R.A. et al. Single-cell multi-omics analysis of the immune response in COVID-19. Nat Med 27, 904–916 (2021). https://doi.org/10.1038/s41591-021-01329-2
12. Yazar, S., Alquicira-Hernandez, J., Wing, K., Senabouth, A., Gordon, M. G., Andersen, S., Lu, Q., Rowson, A., Taylor, T. R. P., Clarke, L., Maccora, K., Chen, C., Cook, A. L., Ye, C. J., Fairfax, K. A., Hewitt, A. W., & Powell, J. E. (2022). Single-cell eQTL mapping identifies cell type-specific genetic control of autoimmune disease. Science (New York, N.Y.), 376(6589), eabf3041. https://doi.org/10.1126/science.abf3041
13. Sikkema, L., Ramírez-Suástegui, C., Strobl, D.C. et al. An integrated cell atlas of the lung in health and disease. Nat Med 29, 1563–1577 (2023). https://doi.org/10.1038/s41591-023-02327-2

The authors would like to thank Francesca Polverino from Baylor College of Medicine for providing the COPD dataset, members of Luecken and Theis labs, scverse community, and all the people in the world involved in the coffee production.

Created with BioRender Poster Builder