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

Vladimir A. Shitov[1,2], Mohammad Dehkordi[3], and 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

March 18, 2025

Abstract

As single-cell datasets are growing, it is becoming possible to analyse differences between groups of samples on a cellular and molecular level. The promise of patient stratification, disease classification, and early-stage diagnosis has led to the development of several so-called sample representation methods. However, consistent standards for the evaluation of sample representation methods are lacking. We developed SPARE – a modular and extendable sample representation benchmark, defining 3 application-inspired metrics, and used these to compare 8 sample representation methods on 5 datasets, testing different preprocessing regimes. We find that the density-based method GloScope outperforms other methods on most datasets and identify general best-practice preprocessing strategies for sample representation methods. We envision that this study will set standards for the development of sample representation methods and facilitate users in selecting an optimal tool, leading to improved outcomes for single-cell applications in precision medicine.

1 Introduction

Single-cell transcriptomics profiles cells at unprecedented resolution in health and disease [Regev et al., 2018]. With an ever-growing number of donors in single-cell tran-

[*]Corresponding author: malte.luecken@helmholtz-munich.de

scriptomics datasets [Hrovatin et al., 2025], it has become possible to study variation on a donor level, which has led to the development of sample representation methods. These methods enable researchers to stratify patient populations [Boyeau et al., 2024], infer disease trajectories and connect them to changes in cell type proportions and gene expression [Joodaki et al., 2024].

While a variety of sample representation methods[1] have been suggested, systematic comparisons are lacking. Existing publications [Appendix A.3] use different datasets to compare methods, define baselines in an inconsistent way, and apply different metrics, typically without assessing their biological relevance. A common way to evaluate the methods in the sample representation literature is the silhouette score [Joodaki et al., 2024, Boyeau et al., 2024, Wang et al., 2024] measuring how well patients with different health status are separated. However, this metric is not extendable to continuous sample-level covariates, such as age, and is not reliable in nested batch-effect scenarios [Rautenstrauch and Ohler, 2025], which are common in single-cell datasets.

Here, we present a Single-cell-based Patient Representation Evaluation (SPARE) benchmark. We developed 4 evaluation metrics measuring clinically relevant information retention, batch effect removal, biological trajectory preservation and robustness of sample representation methods. We used these to systematically compare 8 sample representation methods on single-cell transcriptomics data from 5 large-scale datasets on COVID-19 [COvid-19 Multi-omics Blood ATlas (COMBAT) Consortium, 2022, Stephenson et al., 2021], aging [Yazar et al., 2022], Chronic Obstructive Pulmonary Disease (COPD)[2], and Human Lung Cell Atlas (HLCA) [Sikkema et al., 2023]. As data preprocessing has been shown to be crucial for cell-level analysis, we additionally explore the effect of 5 preprocessing approaches for sample representations. We demonstrate the impact of using top-performing methods by showcasing the recovery of COVID-19 biomarkers. Our work provides a framework for sample representation evaluation and an easily extendable Nextflow pipeline [Di Tommaso et al., 2017] to facilitate future method development. We envision that the SPARE benchmark will standardize the problem definition of generating meaningful sample representations from single-cell data and guide both users and developers of these methods to better derive relevant patient-level insights from single-cell data, paving the way for single-cell-informed personalised medicine.

2 Benchmarking setup

Sample representation methods have been developed using a variety of approaches, from simple averaging to optimal transport Joodaki et al. [2024], Chen et al. [2020] or tensor decomposition [Mitchel et al., 2024]. As a result, method input and output formats vary. To consistently benchmark sample representation methods, we developed a benchmarking pipeline (Figure 1) that defines common data input and output formats to which all methods were adapted. We collected 5 population-scale datasets with comprehensive sample-level metadata from 2 tissues comprising 4.5 million cells

[1] Sometimes also called patient representation methods. To disambiguate from cases when all the donors in a dataset are healthy or when donors have multiple samples taken, we prefer using the term "sample".

[2] Unpublished, provided by collaborators

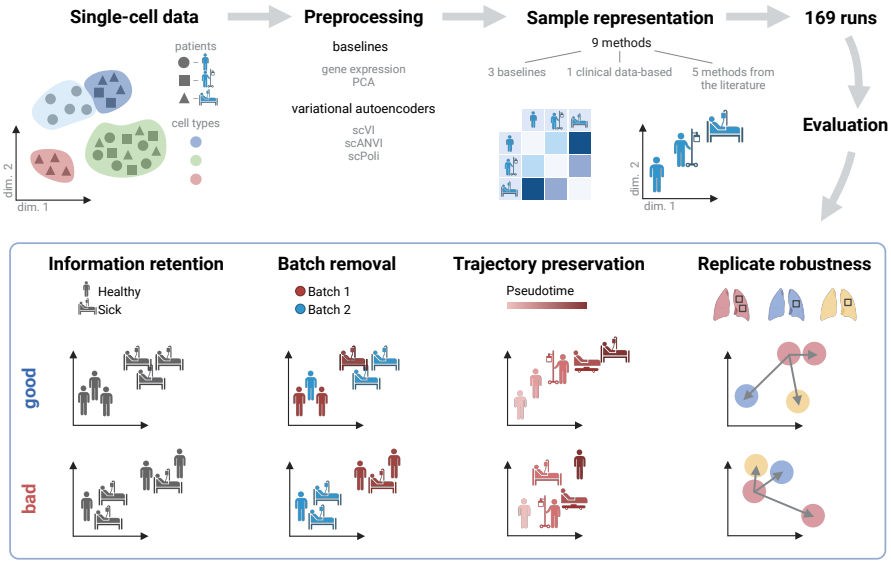

Figure 1: Benchmark overview. Top: benchmarking pipeline, bottom: metrics and examples of a good and a bad representation for each metric.

and 1668 samples (Table 1). We classified the metadata into technical (e.g., batch, number of cells) and relevant (e.g., disease severity, symptoms duration). Here, relevant metadata are signals relevant to the main focus of the corresponding study and are regarded as most important to accurately represent and technical metadata signals are regarded as a nuisance signal. Single-cell datasets are preprocessed using 5 strategies that comprise 3 batch correct methods (scVI [Gayoso et al., 2022], scANVI [Xu et al., 2021], and scPoli [De Donno et al., 2023]) and 2 non-batch corrected cellular representations (PCA [Pearson, 1901] and gene expression), and the preprocessed single-cell embeddings are input into 8 methods comprising baseline methods and sample representation methods [Argelaguet et al., 2020, Wang et al., 2024, Joodaki et al., 2024, De Donno et al., 2023, Heumos et al., 2024]. Excluding the methods that did not finish the computation in 24 hours, this gives a total of 169 runs in the benchmark.

Overall, this setup is sufficiently flexible to evaluate different types of sample representation methods while ensuring that methods are evaluated in comparison with simple baseline methods that have performed well in the past on a range of datasets that cover diverse application scenarios.

3 Defining biologically meaningful metrics

In this section, we suggest metrics inspired by biological applications. We define information retention and batch effect removal scores to rank sample representation methods by their diagnostic applicability, a trajectory preservation score to evaluate the

Table 1: Datasets overview.

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 covariates	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

representation of continuous processes, and a replicate robustness score to check the consistency of results across replicate samples. Rigorous definitions of these metrics can be found in the appendix A.4.

3.1 Information retention and batch effect removal

The structure of a meaningful sample representation should be determined by relevant biological and clinical parameters of interest and not by technical artefacts. Mathematically, this can be quantified by assessing the proximity of biologically similar samples in an embedding space. A good sample representation can be used to annotate samples with unknown labels based on similar samples in this embedding. This can be helpful for personalised medicine applications, such as disease diagnostics, and for atlassing, where researchers often struggle due to missing labels in the data [Huang et al., 2023].

To overcome the limitations of a typically used silhouette score, we propose evaluating embeddings using a KNN-based prediction of sample-level features. In this setup, the prediction performance reflects the preservation of the corresponding effect in the sample embedding. We measure the F_1 score corrected for random prediction (see Appendix A.5) for categorical covariates and the Spearman correlation score for ordinal and continuous variables. For relevant features (Table 1), we call this metric information retention. For batch effect removal, we use the same KNN prediction approach but invert the metric so that score 0 means an embedding, where technical features are grouped perfectly, and score 1 means a complete removal of technical effect from a sample representation. We report the average metric for all relevant or technical covariates, thus focusing not only on one covariate of interest, but on all accessible metadata.

3.2 Biological trajectory preservation

Many biological processes that are likely of interest in a sample representation (such as infection, development, or aging) are continuous in nature. It is therefore important to not only group samples in a meaningful way but also to order them correctly. To measure how well such effects are preserved in the sample embeddings, we assess whether sample-level trajectories can be identified that order ordinal or continuous relevant metadata covariates correctly. For this, we calculate diffusion pseudotime starting from the putatively earliest point in the trajectory (see Appendix) and compute its correlation with various trajectories from the metadata.

3.3 Replicate robustness

Assuming that a tissue sample represents the health of the underlying organ, replicate samples taken from the same individual at the same time point should capture technical variability, potentially some biological variability from differences in sampling location, but limited or no clinical variability. Thus, these samples should have a very similar sample representation. We test this assumption with different samples from the same patients in the COPD dataset. We use the fraction of samples less similar to a given sample than its replicate as a metric. A value 0 means that all samples are more similar to a sample than its replicate, and a value 1 means that replicates are the most similar samples in the representation. We report the average value for 6 replicate pairs as a final metric.

3.4 Metric aggregation

To aggregate scores into one metric for ranking, we used a weighted average with weights equal to 1 for every metric except for batch removal, where it is set to 1/2. This is done to prevent prioritizing poor representations because the batch removal score is equal to 1 for a random embedding. It is not easy to reach a high score in other metrics so the batch removal score without down-weighting biases results towards representations closer to random. The aggregated total score is then scaled to the [0; 1] interval.

4 Benchmarking results

4.1 GloScope and baselines often outperform other methods

We find that for all tasks, most or all top-performing sample representations are built with cell embeddings from variational autoencoders (Table 2). Even baseline sample representation methods, such as pseudobulk or cell-type pseudobulk, often show good performance and outperform most other tools when batch-corrected cell embeddings are used. The only methods with a higher total score than baselines were GloScope [Wang et al., 2024], which showed consistent performance across datasets, and MOFA [Argelaguet et al., 2020], which provided the best sample representations for the Stephenson dataset.

4.2 GloScope is the best method for information retention

We find that across all sample representation tasks, the density estimation-based method GloScope performed best in information retention (score 0.448 ± 0.123 across datasets). Furthermore, GloScope also showed the lowest standard deviation among top-performing methods, suggesting its robust performance across tasks: individual embeddings of GloScope were top performers for 4 out of 5 datasets in our benchmark. PCA-GloScope representation scored best for HLCA and COPD datasets, while embeddings built on scANVI features performed best for Onek1k and COMBAT datasets. For the Stephenson dataset, the highest information retention score was obtained with MOFA trained

Table 2: 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$_b$ refers to a sc[AN]VI model trained with the batch covariate "batch" to integrate the data, while sc[AN]VI$_s$ uses sample ID as a batch covariate.

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

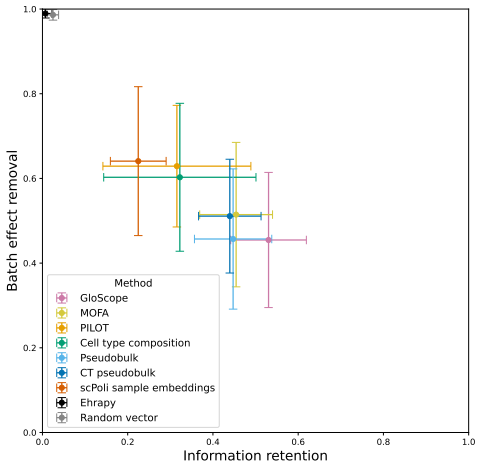

Figure 2: Information retention and batch effect removal trade-off. Mean and standard deviation across datasets are shown for the best sample representation from each method.

on scVI or scANVI cell embeddings, with several GloScope representations closely following.

Comparing sample representation methods with varying preprocessing strategies, scPoli [De Donno et al., 2023] stands out as a preprocessing strategy because it provides sample embeddings in addition to cell representations. Notably, scPoli is one of the worst performing methods for sample embedding with a mean information retention score of 0.225 ± 0.065 and a batch removal score of 0.641 ± 0.176, outperforming only negative baseline with random vectors and ehrapy representation in capturing relevant information. However, scPoli cell embeddings were a valuable input for GloScope, providing top-performing representations for COMBAT and Onek1k.

Notably, all methods in our benchmark outperformed the Ehrapy [Heumos et al., 2024] representation built on accessible clinical metadata. This highlights the significant information gain from single-cell transcriptomics data and its potential application in a clinical scenario.

4.3 GloScope recovers biomarkers of COVID-19 severity

After evaluating all methods on all datasets using trajectory preservation score, we again found that GloScope embeddings are in the top 3 of all datasets. Gloscope produces the best representation for COVID-19 severity in the COMBAT (score = 0.787) and aging in the Onek1k (score = 0.420) datasets. MOFA represented COVID-19 severity in the Stephenson dataset with a score of 0.453. Cell-type pseudobulk captured COPD severity best (score = 0.332), and pseudobulk obtained the best embedding for continuous anatomical location in HLCA (score = 0.814). All the best trajectory representations were based on cellular features from a variational autoencoder (scANVI for all except COMBAT, where scPoli-based representation scored the highest).

To demonstrate the biological utility of this metric, we further investigated the top-performing embedding in a COVID-19 severity case study. While the scANVI-based GloScope representation scored the highest for information retention (score = 0.429), pseudotime built on this embedding only had a correlation score of 0.266 with COVID-19 severity. In contrast, the scPoli-GloScope embedding had an information retention score of 0.373 but a trajectory preservation score of 0.787. UMAP visualisations of the sample embeddings (Figure 3) provide some context for this discrepancy. The scANVI-based representation places patients with sepsis separately, thus achieving a better KNN-based score, while the scPoli-based embedding mixes these patients with COVID-19 cases, thereby potentially representing inflammation patterns that are common in different diseases.

We confirm this by computing the correlation of pseudotime with cell type proportions. The pseudotime trajectory for the scPoli-GloScope representation correlates with the proportion of classical monocytes (Spearman correlation 0.489, adj. p-value 1.95e-08), platelets (Spearman correlation 0.462, adj. p-value 1.02e-07) and other hallmarks of COVID-19 severity [COvid-19 Multi-omics Blood ATlas (COMBAT) Consortium, 2022], thus acting as a severity score. In contrast, the scANVI-based representation only correlates with B-cell proportions (Spearman correlation 0.297, adj. p-value 0.007). This result suggests that our trajectory-evaluation metric is a useful

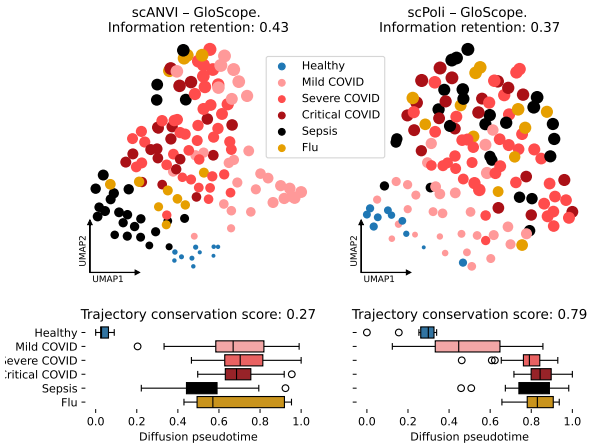

Figure 3: Comparison of the best representations of the COMBAT dataset according to information retention (left) and trajectory conservation (right) score. Top row: UMAP visualisations of scANVI-GloScope and scPoli-GloScope sample representations, point size represents diffusion pseudotime values and colors represent health status. Bottom row: distribution of diffusion pseudotime values among health status groups.

orthogonal measure of sample representations, and it can be used as an analysis tool to obtain biological insights on likely disease biomarkers.

4.4 Many sample representation methods fail to robustly represent replicates

The only methods that correctly represent all the replicates as the closest samples were scANVI-GloScope and PCA-MOFA (Table 4). Aggregated results with the best representation per input space according to the robustness metric suggest that GloScope is the most robust method (score of 0.986), followed by cell type composition (score of 0.974, 1 data point only), and PILOT (score of 0.937).

4.5 Scalability limits the application of certain methods

The number of cells in datasets has grown exponentially over the years, and recently, datasets with close to a thousand samples have become available. With expectations of even larger datasets, sample representation methods must be not only efficient but also scalable.

Our experiments were run on a computational cluster with 300 GB RAM and 32 CPU cores. Preprocessing was performed on a computational cluster with a GPU. We find that most methods compared in this study take minutes to run (at most 11 minutes 29 seconds for one PILOT run). However, GloScope, despite great performance in all

metrics, took 6 hours and 53 minutes to finish on average, with the longest run taking 16 hours. For HLCA, GloScope did not finish computations in 24 hours and was not evaluated. Moreover, some published methods were not described in this benchmarking study [Boyeau et al., 2024, Tong et al., 2021] as they caused out-of-memory error, even on a 50% subset of the data. We encourage the method developers community to use our benchmarking setup to suggest new methods or improve the efficiency of the existing tools.

5 Conclusion

In this study, we defined biologically meaningful metrics to evaluate sample representation methods from single-cell data and showed how they can select methods that provide valuable insights into disease. SPARE can distinguish good from poor performing sample representation methods and therefore sets standards for the development of these methods and prioritizes tools for differnet use cases. We find that cell embeddings from variational autoencoder-based models provide valuable input for sample representation, and a density estimation method, GloScope, outperforms other approaches in all metrics despite worse scalability. We see great potential for method development in this direction facilitated by our sample representation benchmark setup and easily extendable Nextflow pipeline.

6 Code availability

SPARE benchmark pipeline is available on GitHub: `https://github.com/lueckenlab/SPARE`. For running sample representation methods, evaluating the methods and biomarker analysis, we used the open-source patpy package: `https://github.com/lueckenlab/patpy`.

Acknowledgments

This work was supported by the Chan Zuckerberg Initiative Foundation (CZIF; grant CZIF2022-007488 (Human Cell Atlas Data Ecosystem)). This project has received funding from the European Union's Horizon 2023 HORIZON MISS CANCER-01-01 programme under Grant Agreement No. 101136552. V.A.S. is supported by the Helmholtz Association under the joint research school "Munich School for Data Science - MUDS". biorender.com was used for making the figures.

Meaningfulness Statement

We consider a meaningful representation of life to be a biological sample representation that reflects the biological and clinical features of donors. Our work builds a benchmark for sample representation from single-cell data and evaluates existing methods with biologically meaningful metrics. This paves the way for single-cell genomics applications for human health and personalised medicine.

References

R. Argelaguet, D. Arnol, D. Bredikhin, et al. Mofa+: a statistical framework for comprehensive integration of multi-modal single-cell data. *Genome Biology*, 21:111, 2020. doi: 10.1186/s13059-020-02015-1.

P. Boyeau, J. Hong, A. Gayoso, M. Kim, J. L. McFaline-Figueroa, M. I. Jordan, E. Azizi, C. Ergen, and N. Yosef. Deep generative modeling of sample-level heterogeneity in single-cell genomics. *bioRxiv*, 2024. doi: 10.1101/2022.10.04.510898. URL https://www.biorxiv.org/content/early/2024/05/10/2022.10.04.510898.

W. S. Chen, N. Zivanovic, D. van Dijk, G. Wolf, B. Bodenmiller, and S. Krishnaswamy. Uncovering axes of variation among single-cell cancer specimens. *Nature Methods*, 17(3):302–310, 2020. doi: 10.1038/s41592-019-0689-z. URL https://doi.org/10.1038/s41592-019-0689-z.

COvid-19 Multi-omics Blood ATlas (COMBAT) Consortium. A blood atlas of covid-19 defines hallmarks of disease severity and specificity. *Cell*, 185(5):916–938.e58, 2022. doi: 10.1016/j.cell.2022.01.012.

C. De Donno, S. Hediyeh-Zadeh, A. A. Moinfar, et al. Population-level integration of single-cell datasets enables multi-scale analysis across samples. *Nature Methods*, 20:1683–1692, 2023. doi: 10.1038/s41592-023-02035-2.

P. Di Tommaso, M. Chatzou, E. Floden, P. Prieto Barja, E. Palumbo, and C. Notredame. Nextflow enables reproducible computational workflows. *Nature Biotechnology*, 35:316–319, 2017. doi: 10.1038/nbt.3820. URL https://nextflow.io.

A. Gayoso, R. Lopez, G. Xing, P. Boyeau, V. Valiollah Pour Amiri, J. Hong, K. Wu, M. Jayasuriya, E. Mehlman, M. Langevin, Y. Liu, J. Samaran, G. Misrachi, A. Nazaret, O. Clivio, C. Xu, T. Ashuach, M. Gabitto, M. Lotfollahi, V. Svensson, E. da Veiga Beltrame, V. Kleshchevnikov, C. Talavera-López, L. Pachter, F. J. Theis, A. Streets, M. I. Jordan, J. Regier, and N. Yosef. A python library for probabilistic analysis of single-cell omics data. *Nature Biotechnology*, Feb 2022. ISSN 1546-1696. doi: 10.1038/s41587-021-01206-w. URL https://doi.org/10.1038/s41587-021-01206-w.

L. Haghverdi, M. Büttner, F. A. Wolf, F. Buettner, and F. J. Theis. Diffusion pseudotime robustly reconstructs lineage branching. *Nature Methods*, 13(10):845–848, 2016. ISSN 1548-7105. doi: 10.1038/nmeth.3971. URL https://doi.org/10.1038/nmeth.3971.

L. Heumos, P. Ehmele, T. Treis, J. Upmeier zu Belzen, E. Roellin, L. May, A. Namsaraeva, N. Horlava, V. A. Shitov, X. Zhang, L. Zappia, R. Knoll, N. J. Lang, L. Hetzel, I. Virshup, L. Sikkema, F. Curion, R. Eils, H. B. Schiller, A. Hilgendorff, and F. J. Theis. An open-source framework for end-to-end analysis of electronic health record data. *Nature Medicine*, 30:3369–3380, 2024. doi: 10.1038/s41591-024-03214-0.

K. Hrovatin, L. Sikkema, V. A. Shitov, et al. Considerations for building and using integrated single-cell atlases. *Nature Methods*, 22:41–57, 2025. doi: 10.1038/s41592-024-02532-y.

Y.-N. Huang, P. V. Jaiswal, A. Rajesh, A. Yadav, D. Yu, F. Liu, G. Scheg, G. Boldirev, I. Nakashidze, A. Sarkar, J. H. Mehta, K. Wang, K. K. Patel, M. A. B. Mirza, K. C. Hapani, Q. Peng, R. Ayyala, R. Guo, S. Kapur, T. Ramesh, M. S. Abedalthagafi, and S. Mangul. The systematic assessment of completeness of public metadata accompanying omics studies. *bioRxiv*, 2023. doi: 10.1101/2021.11.22.469640. URL https://www.biorxiv.org/content/early/2023/12/27/2021.11.22.469640.

M. Joodaki, M. Shaigan, V. Parra, R. D. Bülow, C. Kuppe, D. L. Hölscher, M. Cheng, J. S. Nagai, M. Goedertier, N. Bouteldja, et al. Detection of patient-level distances from single cell genomics and pathomics data with optimal transport (pilot). *Molecular systems biology*, 20(2):57–74, 2024.

M. D. Luecken, M. Büttner, K. Chaichoompu, et al. Benchmarking atlas-level data integration in single-cell genomics. *Nature Methods*, 19:41–50, 2022. doi: 10.1038/s41592-021-01336-8.

J. Mitchel, M. G. Gordon, R. K. Perez, E. Biederstedt, R. Bueno, C. J. Ye, and P. V. Kharchenko. Coordinated, multicellular patterns of transcriptional variation that stratify patient cohorts are revealed by tensor decomposition. *Nature Biotechnology*, 2024. doi: 10.1038/s41587-024-02411-z. URL https://doi.org/10.1038/s41587-024-02411-z.

K. Pearson. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2(11):559–572, 1901.

P. Rautenstrauch and U. Ohler. Metrics matter: Why we need to stop using silhouette in single-cell benchmarking. *bioRxiv*, 2025. doi: 10.1101/2025.01.21.634098. URL https://www.biorxiv.org/content/early/2025/01/24/2025.01.21.634098.

A. Regev, S. Teichmann, O. Rozenblatt-Rosen, M. Stubbington, K. Ardlie, I. Amit, P. Arlotta, G. Bader, C. Benoist, M. Biton, B. Bodenmiller, B. Bruneau, P. Campbell, M. Carmichael, P. Carninci, L. Castelo-Soccio, M. Clatworthy, H. Clevers, C. Conrad, R. Eils, J. Freeman, L. Fugger, B. Goettgens, D. Graham, A. Greka, N. Hacohen, M. Haniffa, I. Helbig, R. Heuckeroth, S. Kathiresan, S. Kim, A. Klein, B. Knoppers, A. Kriegstein, E. Lander, J. Lee, E. Lein, S. Linnarsson, E. Macosko, S. MacParland, R. Majovski, P. Majumder, J. Marioni, I. McGilvray, M. Merad, M. Mhlanga, S. Naik, M. Nawijn, G. Nolan, B. Paten, D. Pe'er, A. Philippakis, C. Ponting, S. Quake, J. Rajagopal, N. Rajewsky, W. Reik, J. Rood, K. Saeb-Parsy, H. Schiller, S. Scott, A. Shalek, E. Shapiro, J. Shin, K. Skeldon, M. Stratton, J. Streicher, H. Stunnenberg, K. Tan, D. Taylor, A. Thorogood, L. Vallier, A. van Oudenaarden, F. Watt, W. Weicher, J. Weissman, A. Wells, B. Wold, R. Xavier, X. Zhuang, and H. C. A. O. Committee. The human cell atlas white paper, 2018. URL https://arxiv.org/abs/1810.05192.

L. Sikkema, C. Ramírez-Suástegui, D. C. Strobl, et al. An integrated cell atlas of the lung in health and disease. *Nature Medicine*, 29:1563–1577, 2023. doi: 10.1038/s41591-023-02327-2.

C. Spearman. The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72–101, 1904.

E. Stephenson, G. Reynolds, R. A. Botting, et al. Single-cell multi-omics analysis of the immune response in covid-19. *Nature Medicine*, 27:904–916, 2021. doi: 10.1038/s41591-021-01329-2.

T. Stuart, A. Butler, P. Hoffman, C. Hafemeister, E. Papalexi, W. M. Mauck III, Y. Hao, M. Stoeckius, P. Smibert, and R. Satija. Comprehensive integration of single-cell data. *Cell*, 177(7):1888–1902.e21, 2019. ISSN 0092-8674. doi: 10.1016/j.cell.2019.05.031.

A. Tong, G. Huguet, A. Natik, K. MacDonald, M. Kuchroo, R. Coifman, G. Wolf, and S. Krishnaswamy. Diffusion earth mover's distance and distribution embeddings, 2021. URL `https://arxiv.org/abs/2102.12833`.

H. Wang, W. Torous, B. Gong, et al. Visualizing scrna-seq data at population scale with gloscope. *Genome Biology*, 25:259, 2024. doi: 10.1186/s13059-024-03398-1.

F. A. Wolf, P. Angerer, and F. J. Theis. Scanpy: large-scale single-cell gene expression data analysis. *Genome Biology*, 19(1):15, 2018. doi: 10.1186/s13059-017-1382-0.

C. Xu, R. Lopez, E. Mehlman, J. Regier, M. I. Jordan, and N. Yosef. Probabilistic harmonization and annotation of single-cell transcriptomics data with deep generative models. *Molecular Systems Biology*, 17(1):e9620, 2021. doi: 10.15252/msb.20209620.

S. Yazar, J. Alquicira-Hernandez, K. Wing, A. Senabouth, M. G. Gordon, S. Andersen, Q. Lu, A. Rowson, T. R. P. Taylor, L. Clarke, K. Maccora, C. Chen, A. L. Cook, C. J. Ye, K. A. Fairfax, A. W. Hewitt, and J. E. Powell. Single-cell eqtl mapping identifies cell type-specific genetic control of autoimmune disease. *Science*, 376 (6589):eabf3041, 2022. doi: 10.1126/science.abf3041.

A Appendix

A.1 Sample representation task

Let $\mathbf{x}_i^s = \{g_{i1}^s, g_{i2}^s, \ldots, g_{im}^s, l_i^s\}$ be a cell of sample s with $g_{\cdot j} \in \mathbb{Z}_{\geq 0}$ representing gene expression and l representing a cell type label. A single-cell sample is a collection of n_s cells $\mathbf{X}_s = \{\mathbf{x}_1^s, \mathbf{x}_2^s, \ldots, \mathbf{x}_{n_s}^s\}$. We define a sample representation task as calculating a divergence or distance between single-cell datasets:

$$d(f(\mathbf{X}_j), f(\mathbf{X}_k)) \tag{1}$$

Where f is a preprocessing function. Sample representation is, therefore, a matrix of pairwise distances (or divergencies) between samples. This is a convenient format because every output of every sample representation method can be converted to a distance between samples and because all of the metrics in our study require only the distances between samples, not sample embeddings. Distance matrix can be used for other downstream tasks, such as clustering to find groups of similar samples, and for visualisation with TSNE, UMAP or Multidimensional scaling.

A.2 Preprocessing methods

A.2.1 Normalized expression

We use scanpy [Wolf et al., 2018] to preprocess single-cell data. Raw gene expression is normalized to 10,000 and log1p-transformed. 3000 highly-variable genes are selected in a batch-aware approach with `seurat_v3` method [Stuart et al., 2019].

A.2.2 PCA

We use Principal Components Analysis [Pearson, 1901] (PCA)-transformed normalized expression as a low-dimensional representation of cells to have a non-batch corrected input space. 50 principal components are used.

A.2.3 scVI

scVI [Gayoso et al., 2022] is a variational autoencoder trained to embed cells in a latent space with the normal distribution of the features and reconstruct counts from latent features. We train the scVI model with parameters `n_layers=2`, `n_latent=30`, `gene_likelihood="nb"`, following the recommendations in the scvi-tools documentation. For all datasets except COPD, we train 2 scVI models with a sample or batch as a batch covariate to eliminate technical effects. Latent cell representations of the trained models are then used as input to sample representation methods.

A.2.4 scANVI

scANVI [Xu et al., 2021] is a semi-supervised extension of scVI that incorporates cell type information in the model. It was shown to be the best method for atlassing level integration of single-cell data [Luecken et al., 2022] and is widely used in the atlassing [Sikkema et al., 2023]. We initialize scANVI model with a trained scVI model and train it for additional 20 epochs.

A.2.5 scPoli

scPoli is a variational autoencoder model for population-level integration of single-cell datasets. Its distinct features are learning sample embeddings instead of one-hot encoding sample IDs in scVI and scANVI models and using prototype loss to learn cell-type prototypes. We train scPoli with parameters `latent_dim=3`, `n_epochs=50`, `pretraining_epochs=40`, `eta=5`.

A.3 Sample representation methods

A.3.1 Random vector

As a negative baseline to validate the correct behaviour of our metrics, we generate random representations for samples in each dataset. This method is completely data-independent. I terms of the equation 1,

$$f(\mathbf{X}) \sim \mathcal{N}_K(\mathbf{0}, \mathbf{I}_K)$$

Where K is a dimensionality of a random vector. We use $K = 10, 30$ and Euclidean distance as d.

A.3.2 Pseudobulk

This simple baseline method aggregates all the cells of a sample in a single vector representing an average cell of this sample. Hence the name: it simulates a sample from a bulk RNA-sequencing study.

$$f(\mathbf{X}_s) = \frac{1}{n_s} \sum_{i=1}^{n_s} \mathbf{x}_i$$

Where $\mathbf{x}_i$ is ith cell representation of a given sample. Features of $\mathbf{x_i}$ must not be raw expression counts at this point and can be any real number. As d, we use Euclidean and cosine distance but discard the latter due to its identical performance with the Euclidean distance.

A.3.3 Cell type pseudobulk

This baseline method aggregates cells per cell type. Each sample is therefore represented as a vector with length $K * C$, where K is the dimensionality of input space and C is the number of cell types in a dataset. Let n_{c_i} be the number of cells with cell type label c_i. Then:

$$f(\mathbf{X}) = \{\mathbf{p}_{c_1}, \mathbf{p}_{c_2}, \ldots, \mathbf{p}_{c_C}\}$$

Where p_{c_i} is a pseudobulk of all cells with label c_i. Euclidean distance between aggregated representations of overlapping cell types between samples is then calculated.

A.3.4 Cell type composition

This baseline method does not use expression information at all and only compares samples based on the differences in cell type proportions.

$$f(\mathbf{X_s}) = \{\sum_{i=1}^{n_s} \frac{\mathbf{1}_{l_i=c_1}}{n_s}, \sum_{i=1}^{n_s} \frac{\mathbf{1}_{l_i=c_2}}{n_s}, \ldots, \sum_{i=1}^{n_s} \frac{\mathbf{1}_{l_i=c_C}}{n_s}\}$$

Where, $\mathbf{1}_{l_i=c_i}$ is the indicator function that equals 1 if the cell i has label l_i equal to cell type c_i, and 0 otherwise. We use Euclidean distance as d.

A.3.5 Ehrapy

This baseline stands out of the others because it is the only one using sample metadata. We select easily accessible covariates that can be routinely measured in a clinical practice or provided by a patient. We then one-hot encode categorical variables, build PCA and calculate Euclidean distances between PCA features of the samples.

Table 3: Metadata features used to build Ehrapy representation.

Dataset	**Accessible metadata features**
COMBAT	Age, Sex, BMI, Hospitalstay, PreExistingHeartDisease, PreExistingLungDisease, PreExistingKidneyDisease, PreExistingDiabetes, PreExistingHypertension, PreExistingImmunocompromised, Smoking, Requiredvasoactive, Respiratorysupport, SARSCoV2PCR, TimeSinceOnset
Stephenson	Swab_result, Smoker, Days_from_onset, sex, development_stage
Onek1k	age, sex
HLCA	BMI, age_or_mean_of_age_range, age_range, anatomical_region_ccf_score, smoking_status, sex, tissue, self_reported_ethnicity, development_stage
COPD	Smoking_Status, Age, Sex, FEV1, FEV1_FVC, Bronchodilator_Use, Leukotriene_Use, Steroid_Use, Pack_Year, Smoking_History, Quit_Data

A.3.6 scPoli

We use scPoli sample embeddings to calculate distances between samples. f, in this case, is a lookup table, which returns a sample embedding from a trained variational autoencoder. d is an Euclidean distance.

A.3.7 PILOT

PILOT [Joodaki et al., 2024] is an optimal transport method that calculates the Wasserstein distance between 2 samples represented as cell type proportions, taking into account cell type similarity. f is similar to cell type composition, but instead of cell type proportions, maximum a posteriori estimates of parameters of the multinomial distribution are taken. d is the Earth Moving Distance between estimated cell type proportions. To compute optimal transport cost, it uses cosine distance between cell types defined as medoids of cell type (similar to pseudobulk, but per-feature median is used instead of average). For details, see the original publication [Joodaki et al., 2024].

A.3.8 GloScope

GloScope [Wang et al., 2024] estimates distance between distributions of cells of different samples. To do so, it uses Kullback-Leibler (KL) divergence. Let $r_j(\mathbf{x}_{i,u})$ be the

distance from the uth cell in sample i to its kth nearest neighbour in sample j. Then, the KL divergence is estimated as:

$$\hat{KL}(\mathbf{X}_i \parallel \mathbf{X}_j) = \frac{K}{n_i} \sum_{u=1}^{n_i} \log \frac{r_j(x_{i,u})}{r_i(x_{i,u})} + \log \frac{n_j}{n_i - 1}$$

Where K is the cell representation dimensionality. The resulting divergence between samples is obtained by symmetrizing KL divergence:

$$d(\mathbf{X}_i, \mathbf{X}_j) = KL(\mathbf{X}_i \parallel \mathbf{X}_j) + KL(\mathbf{X}_j \parallel \mathbf{X}_i)$$

We run GloScope with parameters `dist_mat="KL"`, `dens="KNN"`, `k=25`.

A.3.9 MOFA

MOFA [Argelaguet et al., 2020] uses factor decomposition to decompose cell type pseudobulks P_i:

$$\mathbf{P}_i = \mathbf{ZW} + \epsilon$$

Where $\mathbf{Z} \in \mathbb{R}^{S \times F}$ is a sample by factor matrix with S samples and F factors, and $W \in \mathbb{R}^{F \times K}$ is the factor loading matrix for cell features. We then calculate the Euclidean distance between factor values for each sample.

A.4 Metrics definition

A.4.1 Information retention and batch removal scores

Let $NN(\mathbf{X}_i, k)$ be the indices of the k nearest neighbors of ith sample in a sample representation. For a categorical metadata covariate G, kNN-based prediction is obtained as:

$$\hat{G}_i = \underset{l \in unique(G)}{\operatorname{argmax}} \sum_{j \in NN(\mathbf{X}_i, k)} \mathbf{1}_{G_j = l}$$

Where $unique(G)$ is the set of unique values of G. For a numerical or ordinal covariate R, a prediction is obtained as:

$$\hat{R}_i = \frac{1}{k} \sum_{j \in NN(\mathbf{X}_i, k)} R_j$$

An information retention score for categorical covariates is the macro F_1 score corrected for random prediction (see Appendix A.5 for the motivation and proof):

$$I(G) = \tilde{F_1}^{macro} = \frac{L}{L-1}({F_1}^{macro}(G, \hat{G}) - \frac{1}{L})$$

For categorical and ordinal covariates, an information retention score is the absolute value of the Spearman correlation [Spearman, 1904] between true and predicted values:

$$I(R) = |\rho(R, \hat{R})|$$

Batch removal score is an inverted information retention score for technical covariates:

$$B(G) = 1 - I(G)$$
$$B(R) = 1 - I(R)$$

We use $k = 3$ for all datasets and covariates in the study.

A.4.2 Trajectory preservation

To evaluate trajectory preservation, we use disease severity in COMBAT, Stephenson and COPD datasets, age in Onek1k dataset, and continuous anatomical location in the HLCA. We select the earliest time point in each trajectory, i.e. the youngest non-smoking healthy donor or the most proximal sample, as the start of the trajectory and apply diffusion pseudotime [Haghverdi et al., 2016] to order the other samples in a representation. Let T be a vector of diffusion pseudotime values and R a ground truth trajectory covariate. The trajectory preservation score TS is defined as the absolute value of Spearman correlation between T and R:

$$TS = |\rho(T, R)|$$

A.4.3 Replicate robustness

Let $U(\mathbf{X}_i, \mathbf{X}_j)$ be the number of samples with distance to $\mathbf{X}_i$ smaller than distance between $\mathbf{X}_i$ and its replicate $\mathbf{X}_j$:

$$U(\mathbf{X}_i, \mathbf{X}_j) = \sum_{k=1, k\neq i,j}^{N} \mathbf{1}_{d(\mathbf{X}_k, \mathbf{X}_j) < d(\mathbf{X}_i, \mathbf{X}_j)}$$

The replicate robustness score RS is the fraction of samples that are **less** similar to $\mathbf{X}_i$ than its replicate:

$$RS(i, j) = 1 - \frac{U(\mathbf{X}_i, \mathbf{X}_j)}{N - 1}$$

The final replicate robustness score is the average across all replicates in a dataset:

$$RS = \frac{1}{|replicates|} \sum_{i,j \in replicates} RS(i, j)$$

We use different samples from the same patients in the COPD dataset to calculate the replicate robustness score.

A.5 F_1 score corrected for random prediction

To measure how well a sample representation preserves information about categorical covariates, we use the macro F1 score for the values predicted from the nearest neighbours for every sample. However, for the default implementation of this score, its value depends on the number of levels L in a covariate. Below, we prove that for a random prediction, the expected value of F_1^{macro} score is $\frac{1}{L}$. For example, if a covariate is binary, the value would be $\frac{1}{2}$, and for a covariate with 6 levels, the score would be $\frac{1}{6}$. Such a behaviour makes it unclear which covariate is represented better and which has a higher value merely due to fewer classes. To be able to rank covariates by the quality of prediction, and make an averaged score more interpretable, we use the corrected version of the metric:

$$\tilde{F_1}^{macro} = \frac{L}{L-1}(F_1{}^{macro} - \frac{1}{L}) \tag{2}$$

This score takes values from $[-\frac{1}{L-1}; 1]$, where a score of 1 means a perfect prediction for every sample, 0 means a random prediction and negative values mean a prediction worse than random. We further clip negative values to 0 to not distinguish between the latter 2 cases.

Proof. In the `sklearn` implementation, macro F_1 score is defined as an unweighted average of F_1 scores for each of L classes:

$$F_1^{macro} = \frac{1}{L}\sum_{i=1}^{L} F_1^i \tag{3}$$

Where the score for each class F_1^i is a harmonic mean of precision and recall considering instances of class i as a "true" label:

$$F_1^i = 2\frac{precision_i * recall_i}{precision_i + recall_i} \tag{4}$$

Precision and recall are defined through true positive (TP), false positive (FP), and false negative (FN) predictions:

$$precision_i = \frac{TP_i}{TP_i + FP_i} \tag{5}$$

$$recall_i = \frac{TP_i}{TP_i + FN_i} \tag{6}$$

Let p_i be a fraction of class i to the total number of observations. For a random prediction, expectations of both $precision$ and $recall$ are equal to p_i as well as their harmonic mean F_1^i. Summing p_i over all classes removes the sum in 3 and finishes the proof. Note that the expected value of F_1^{macro} does not depend on the proportion of the classes but only on their number.

A.6 Batch effect correction is a necessary step of optimal sample embedding

Batch effect correction is a crucial analysis step to draw biologically meaningful conclusions that are not affected by technical artefacts from genomics data. For single-cell transcriptomics, variational autoencoder-based methods [Gayoso et al., 2022, Xu et al., 2021, De Donno et al., 2023] have shown good performance for large-scale data integration [Luecken et al., 2022]. However, the effect of batch effect correction on sample representation was not previously explored. Our analysis shows that technical and relevant features are often entangled, and better retention of relevant information usually means worse batch effect removal (Figure 2). In an attempt to break this relation, we tested if sample representations based on batch-corrected cellular features outperform representations based on PCA-transformed count data.

We compared the information retention and batch removal scores for all methods that could take different cell representations as input with those of the same method based on PCA (Figure 4). Notably, using a batch-corrected space often did not improve the relevant feature representation. Instead, the information retention score was often reduced while batch correction was improved, showing information loss in the embedding. Most frequently, this was the case for MOFA and pseudobulk experiments. For some methods, however, batch correction prior to sample representation improved the information retention score while reducing batch effects. GloScope, PILOT and pseudobulk benefitted from batch-corrected cell representations the most. Notably, many of the inproved representations used cell embeddings from scANVI or scPoli, which both leverage cell type labels on training. This result suggests that using cell-type aware integration methods benefits sample representation. This can be explained by the fact that they use more prior information or that they converge better with default parameters.

A.7 Replicate robustness results

Table 4: Average replicate robustness metric for different methods

Sample representation method	Average replicate robustness metric
GloScope	0.986 ± 0.001
Cell type composition	0.974
PILOT	0.937 ± 0.025
Pseudobulk	0.912 ± 0.103
CT pseudobulk	0.911 ± 107
MOFA	0.751 ± 0.235
Random vector	0.370

A.8 Not aggregated results for the KNN score

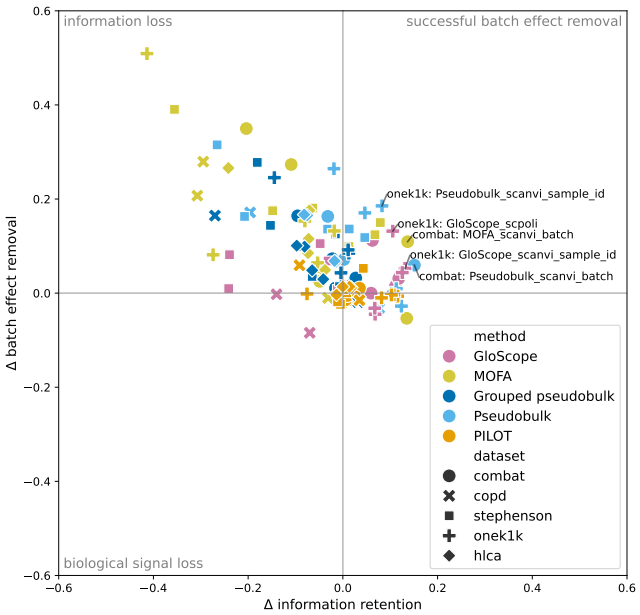

Figure 4: Effect of batch correction on sample representation. Axes represent differences in information retention (horizontal) and batch removal (vertical) scores for each method in comparison to PCA-based representation with the same method.

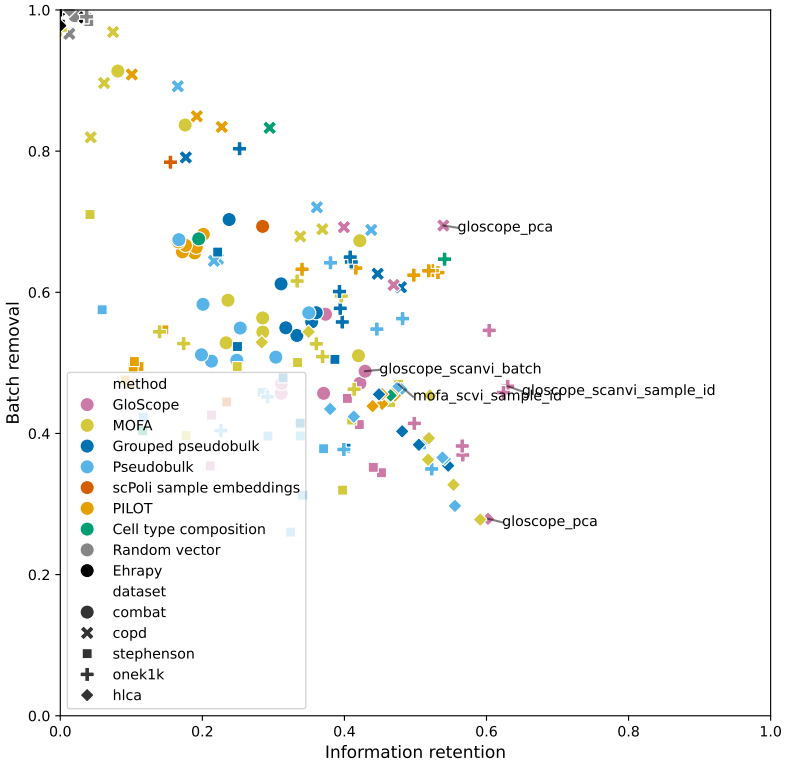

Figure 5: Not aggregated results for Figure 2