MultiMAP: Dimensionality Reduction and Integration of Multimodal Data

Mika Sarkin Jain* [1 2] **Krzysztof Polanski** [2] **Cecilia Dominguez Conde** [2] **Xi Chen** [2 3] **Jongeun Park** [2 4] **Lira Mamanova** [2] **Andrew Knights** [2] **Rachel A. Botting** [5] **Emily Stephenson** [5] **Muzlifah Haniffa** [2 5] **Austen Lamacraft** [1] **Mirjana Efremova** [2 6] **Sarah A. Teichmann** [1 2]

Abstract

Multimodal data is rapidly growing in single-cell biology and other fields of science and engineering. We introduce MultiMAP, an approach for dimensionality reduction and integration of multiple datasets. MultiMAP is a nonlinear manifold learning technique that recovers a single manifold on which all datasets reside and then projects the data into a single low-dimensional space so as to preserve the manifold structure. MultiMAP has several advantages over existing integration strategies for single-cell data, including that it can integrate any number of datasets, leverages features that are not present in all datasets (i.e. datasets can be of different dimensionalities), is not restricted to a linear mapping, allows the user to specify the influence of each dataset on the embedding, and is extremely scalable to large datasets. We apply MultiMAP to the integration of a variety of single-cell transcriptomics, chromatin accessibility, methylation, and spatial data, and show that it outperforms current approaches in preservation of high-dimensional structure, alignment of datasets, visual separation of clusters, transfer learning, and runtime. The MultiMAP codebase is available at this https url.

1. Introduction

Multimodal data is rapidly growing in single-cell biology and many other fields of science and engineering. Emerging single-cell technologies are providing high-resolution measurements of different features of cellular identity, including single-cell assays for gene expression, protein abundance[2,3], chromatin accessibility[4], DNA methylation[5], and spatial resolution[6]. Large scale collaborations including the Human Cell Atlas international consortium[7],[8] are generating an exponentially increasing amount of data of many biological tissues, using a myriad of technologies. Each technology provides a unique view of cellular biology and has different strengths and weaknesses. Integrating these measurements in the study of a single biological system will open avenues for more comprehensive study of cellular identity, cell-cell interactions, developmental dynamics, and tissue structure[9].

The integration of multi-omic data poses several challenges[10]. Different omics technologies measure distinct unmatched features with different underlying distributions and properties and hence produce data of different dimensionality. This makes it difficult to place data from different omics in the same feature space. Additionally, omics technologies can also have different noise and batch characteristics which are challenging to identify and correct. Further, as multi-omic data grows along two axes, the number of cells per omic and the number of omics per study, integration strategies need to be extremely scalable.

Most data integration methods project multiple measurements of information into a common low-dimensional representation to assemble multiple modalities into an integrated embedding space. Recently published methods employ different algorithms to project multiple datasets into an embedding space, including canonical correlation analysis (CCA)[11], nonnegative matrix factorization (NMF)[]12] or neural network models[13]. While these methods can be tremendously powerful, they suffer from several shortcomings. Current methods require correspondence between the features profiled across omics technologies. A further drawback is that methods that use linear models, such as CCA and NMF, are not able to capture non-linear differences between datasets. Another limitation of these methods is they cannot scale to large datasets, failing on datasets of hundreds of thousands to millions of cells.

*Equal contribution [1]Theory of Condensed Matter, Dept Physics, Cavendish Laboratory, University of Cambridge [2]Wellcome Sanger Institute, Wellcome Genome Campus, Cambridge [3]Southern University of Science and Technology [4]KAIST, South Korea [5]Biosciences Institute, Newcastle University [6]Barts Cancer Institute, Queen Mary University of London. Correspondence to: Mika Sarkin Jain <mikasarkinjain@gmail.com>, Mirjana Efremova <me5@sanger.ac.uk>, Sarah A. Teichmann <st9@sanger.ac.uk>.

The 2021 ICML Workshop on Computational Biology.

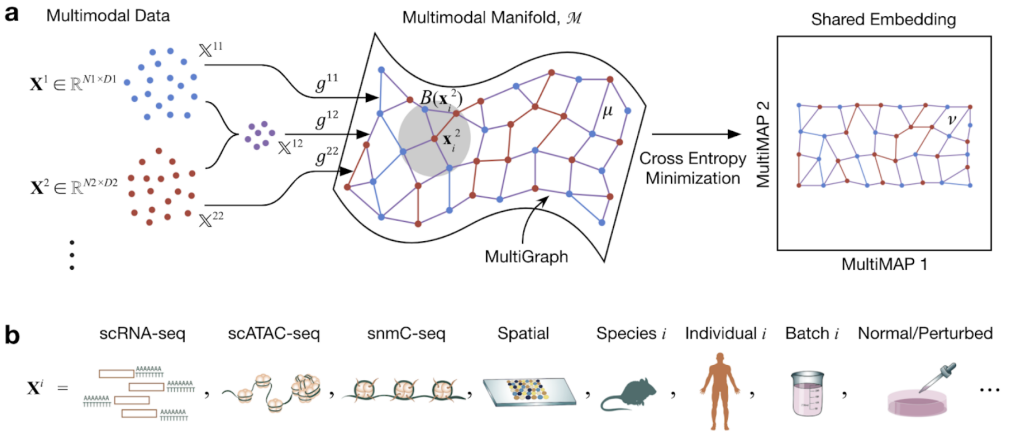


Figure 1. **Schematic of MultiMAP.** a. MultiMAP takes any number of datasets, including those of differing dimensions, recovers geodesic distances on a single latent manifold on which all data lie, constructs a neighborhood graph (MultiGraph) on the manifold, and then projects the data into a single low-dimensional embedding. Integrated analysis and visualisation can be performed on the embedding or graph. Variables are discussed in Methods. $\mathbf{X}^i$ is dataset i, $\mathbf{x}_{ji}$ is a point in $\mathbf{X}^i$, $\mathcal{M}$ is the shared manifold, $B(\mathbf{x}_i^2)$ is a ball on $\mathcal{M}$ centered at $\mathbf{x}_i^2$, $\mathbb{X}^{ij}$ is the ambient space of $\mathcal{M}$ in the coordinate space with data containing points from datasets i and j, g_{ij} is the metric of $\mathcal{M}$ in the space $\mathbb{X}_{ij}$, μ is the membership function of the fuzzy simplicial set on the manifold, ν is the membership function of the fuzzy simplicial set in the low-dimensional space. The mathematical formulation of MultiMAP is elaborated in Supplementary Methods. b. In the field of cell atlas technologies, encompassing single cell genomics and spatial technologies, MultiMAP can be applied to integrate across different omics modalities, species, individuals, batches, and normal/perturbed states.

Here we introduce a method that overcomes all these limitations: MultiMAP, an approach for the dimensionality reduction and integration of multiple datasets. MultiMAP integrates data by constructing a non-linear manifold on which diverse high-dimensional data reside and then projecting the manifold and data into a shared low-dimensional space. In contrast to other integration strategies for single-cell data, MultiMAP can integrate any number of datasets, is not restricted to a linear mapping, leverages features that are not present in all datasets (i.e. datasets can be of different dimensionalities), allows the user to specify the influence of each dataset on the embedding, and is effortlessly scalable to large datasets. The ability of MultiMAP to integrate datasets of different dimensionalities allows the strategy to leverage information that is not considered by methods that operate in a shared feature space. (e.g. MultiMAP can integrate the 20,000-feature gene space of scRNAseq data together with a 100,000-feature peak space of scATACseq data).

We apply MultiMAP to challenging synthetic multimodal data, and demonstrate its ability to integrate a wide range of single-cell omics datasets. We show that MultiMAP can co-embed datasets across different technologies and modalities, while at the same time preserving the structure of the data, even with extensive biological and technical differences. The resulting embedding and shared neighborhood graph (MultiGraph) can be used for simultaneous visualisation and integrative analysis of multiple datasets. With respect to single cell genomics data, this allows for standard analysis on the integrated data, such as cluster label transfer, joint clustering, and trajectory analysis.

2. Results

2.1. The MultiMAP Framework

We introduce MultiMAP, an approach for integration and dimensionality reduction of multimodal data based on a framework of Riemannian geometry and algebraic topology. MultiMAP takes as input any number of datasets of potentially differing dimensions. MultiMAP recovers geodesic distances on a single latent manifold on which all of the data is uniformly distributed. The distances are calculated between data points of the same dataset by normalizing distances with respect to a neighborhood distance specific to the dataset, and between data points of different datasets by normalizing distances between the data in a shared feature space with respect to a neighborhood parameter specific to the shared feature space. These distances are then used to construct a neighborhood graph (MultiGraph) on the manifold. Finally, the data and manifold space are projected into a low-dimensional embedding space by minimizing the cross entropy of the graph in the embedding space with respect to the graph in the manifold space. MultiMAP allows the user to modify the weight of each dataset in the cross entropy loss, allowing the user to modulate the contribution

of each dataset to the layout. Integrated analysis can be performed on the embedding or the graph, and the embedding also provides an integrated visualization. The mathematical formulation of MultiMAP is elaborated in Supplementary Methods.

In order to study MultiMAP in a controlled setting, we first applied it to two synthetic examples of multimodal data (Methods). The first synthetic data consists of points sampled randomly from the canonical 3D "Swiss Roll" surface and the 2D rectangle (Figure 2a). The dataset is considered multimodal data, because samples are drawn from different feature spaces but describe the same rectangular manifold. In addition, we are given the position along the manifold of 1% of the data. This synthetic setting illustrates that MultiMAP can integrate data in a nonlinear fashion and operate on datasets of different dimensionality, because data points along a similar position on the manifold are near each other in the embedding (Figure 2b). The MultiMAP embedding properly unrolls the Swiss Roll dataset, indicating that the projection is nonlinear. The embedding also appears to preserve aspects of both datasets; the data is curved and at the same time unrolled.

To determine if MultiMAP can effectively leverage features unique to certain datasets, we used the MNIST database[14], where handwritten images were split horizontally with thin overlap (Figure 2c; see Methods for details). The two datasets can be considered multimodal because they have different feature spaces but describe the same set of digit images. The thin overlapping region of the two halves is not enough information to create a good embedding of the data (Figure 2c). Many distinct digits are similar in this thin central sliver, and hence they cluster together in the feature space of this sliver. Indeed, in a UMAP projection of the data in the shared feature space of this overlap, the clusters of different digits are not as well separated as in the UMAP projections of each half (Figure 2c).

A multimodal integration strategy that effectively leverages all features would use the features unique to each half to separate different digits, and the shared space to bring the same digits from each dataset close together (Figure 2d). We show that with MultiMAP the different modalities are well mixed in the embedding space and the digits cluster separately, despite mostly different feature spaces and noise being added to only the second dataset. This indicates that MultiMAP is leveraging the features unique to each dataset and is also robust to datasets with different noise.

2.2. Benchmarking

We assessed and benchmarked the performance of MultiMAP against several popular approaches for integrating single-cell multi-omics, including Seurat 3[11], LIGER[12], Conos[22] and GLUER[23].

These integration approaches differ in key regards, summarized in Figure 3c. We used a diversity of performance metrics to comprehensively compare MultiMAP with other approaches, including transfer accuracy, silhouette score, alignment, preservation of the structure, and runtime. With these metrics, we quantified the separation of the joint clusters, how well mixed the datasets were after integration and how well they preserved the structure in the original datasets to investigate whether the methods integrate populations across datasets without blending distinct populations together.

We benchmarked MultiMAP using a variety of multi-omic data with both newly generated and published cell type annotations. This includes scATAC-seq and scRNA-seq data of the mouse spleen (n=1) [15], scATAC-seq and scRNA-seq data data of human bone marrow and peripheral blood mononuclear cells [16], and scRNA-seq and spatial STARmap (n=2) data of the mouse brain [18]. For all datasets, MultiMAP achieves top or near top performance on all metrics (Figure 3a). The embeddings produced by MultiMAP prove superior for transferring cell type annotations between datasets, separating clusters of different cell populations, integrating datasets in a well-mixed manner, and capturing the high-dimensional structure of each dataset.

Critically, MultiMAP is significantly faster and more scalable than all other benchmarked methods, and significantly faster than LIGER and Seurat 3 (Figure 3b). Seurat 3 and LIGER were not able to scale to the primary cortex data of 600k, producing out-of-memory errors despite access to 218 GB of RAM.

3. Discussion

Here we present a novel approach for dimensionality reduction and integration of multimodal data. MultiMAP estimates a non-linear manifold on which all data reside and then projects this manifold space into a low dimensional embedding. This enables both visualization and integrated downstream analyses of all datasets simultaneously. Crucially, our method takes into account the full data, even when they have different feature spaces, and thus takes advantage of the full power of multi-omics data. Ignoring the features unique to one dataset (as in existing methods), may omit important information, for instance distinguishing features of certain subpopulations of cells and yield an integrated embedding that does not distinctly cluster all subpopulations. Comparison with existing methods for integration shows that MultiMAP outperforms or has close to best performance in every performance metric studied. In particular, MultiMAP far more fast and scalable than current approaches.

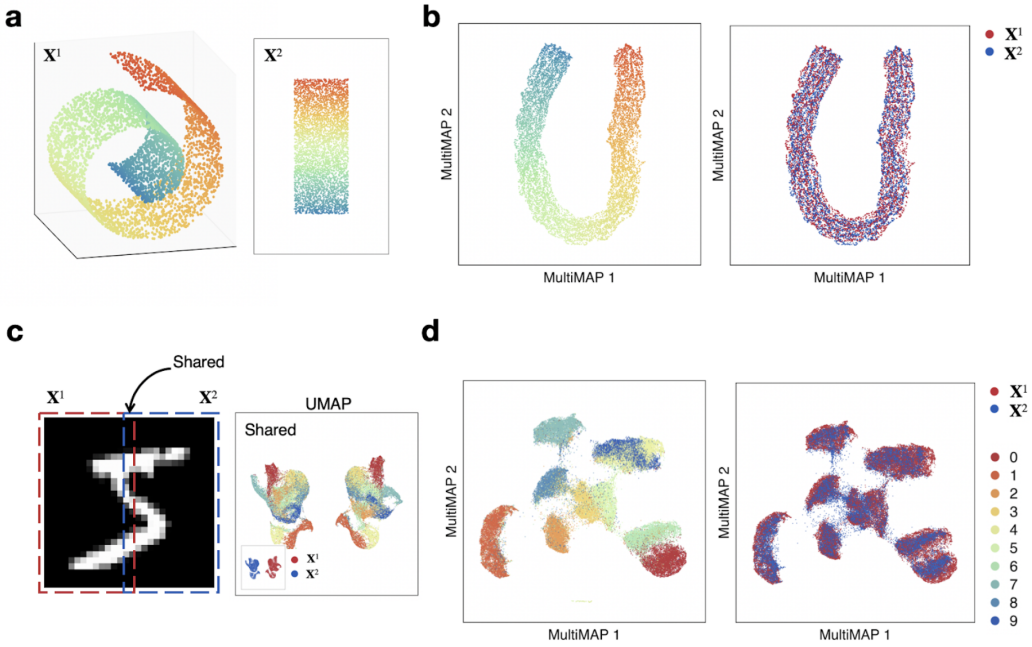


Figure 2. **MultiMAP applied to synthetic data.** a. Data sampled from the 3D Swiss Roll (X1) and a 2D rectangle (X2). b. Shared embedding of both datasets produced by MultiMAP. Color indicates position along the manifold (a,b). c. Left (X1) and right (X2) halves of MNIST handwritten digit images with a 2 pixel wide shared region. Gaussian noise is added to the left half. UMAP projections of each half and the shared region. d. Shared embedding of both MNIST halves (including Gaussian noise introduced for the left half) produced by MultiMAP. Each color is a different handwritten digit (0-9 as shown in the key). This illustrates that MultiMAP leverages both shared and unshared features to integrate multimodal datasets.

Using synthetic examples to illustrate the power of the method, we show that MultiMAP leverages the features unique to each dataset, is capable of discovering non-linear transformations, and is robust to data with noise. Throughout our applications of MultiMAP to diverse single-cell multi-omic data, we demonstrate that our method can facilitate integration across transcriptomic, epigenomic, and spatially resolved datasets, and derive biological insights jointly from multi-omic single-cell data. This demonstrates that MultiMAP can align datasets across different technologies and modalities even with extensive biological and technical differences. The ability of MultiMAP scale to massive datasets and integrate more than two omics technologies opens many opportunities for the comprehensive study of tissues. Crucially, we show that MultiMAP is flexible enough to integrate datasets with different clusters and cell populations, illustrating that MultiMAP is applicable even when datasets contain some different cell type populations.

Perhaps the greatest potential lies in applying MultiMAP to datasets beyond those considered here. Integrative analysis with MultiMAP can be used to compare healthy and diseased states, and identify pathologic features, or to uncover cell-type specific responses to perturbations. Other examples include the integration of data across species to enable studying the evolution of cell states and identifying conserved cell types and regulatory programs. Along similar lines, the integration of in vivo with in vitro models such as organoids will reveal the quality or faithfulness of cells in a dish relative to their native counterparts. Finally, given the rapid development of joint multimodal single cell genomics methods (e.g. CITEseq for protein and RNA, joint snRNA- and ATACseq), it is relevant to emphasize that MultiMAP can be applied to multi-omic data acquired both from different cells as well as from the same cells.

In summary, given the broad appeal of dimensionality reduction methods (e.g. PCA, tSNE, UMAP), and the growth of multimodal data in many areas of science and engineering, we anticipate that MultiMAP will find wide and diverse use.

4. References

1. Becht, E. et al. Dimensionality reduction for visualizing single-cell data using UMAP. Nat. Biotechnol. (2018) doi:10.1038/nbt.4314.

2. Stoeckius, M. et al. Simultaneous epitope and transcriptome measurement in single cells. Nature Methods vol. 14

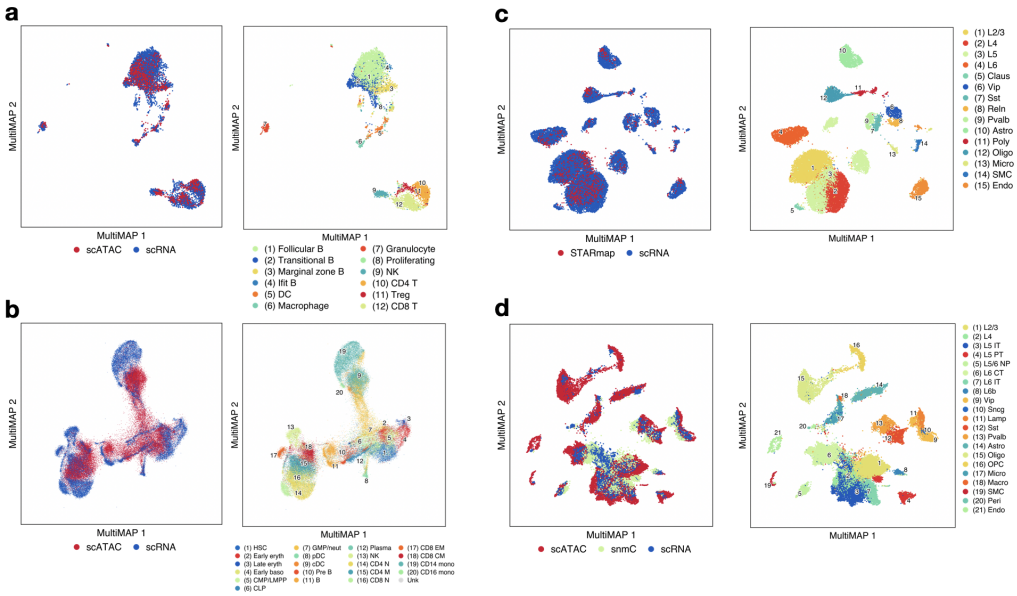


Figure 3. **MultiMAP applied to single-cell data.** a. MultiMAP embedding of the integration of published scATAC-seq15 and newly generated scRNA-seq data of the mouse spleen (n=1), colored by omic technology and independent cell type annotations of each omic technology. b. MultiMAP embedding of the integration of single-cell transcriptomics and chromatin accessibility of human bone marrow and peripheral blood mononuclear cells[16] colored by omic technology and by the published cell type annotation. c. MultiMAP embedding of scRNA-seq17 (n=2) and spatial STARmap[18] (n=2) data of the mouse brain, colored by omic technology and joint clusters identified with the MultiGraph. d. MultiMAP embedding of the integration of single-cell transcriptomics, chromatin accessibility, and DNA methylation of the mouse primary cortex, colored by omic technology and the published cell type annotation[20].

865–868 (2017).

3. Peterson, V. M. et al. Multiplexed quantification of proteins and transcripts in single cells. Nature Biotechnology vol. 35 936–939 (2017).

4. Klemm, S. L., Shipony, Z. Greenleaf, W. J. Chromatin accessibility and the regulatory epigenome. Nat. Rev. Genet. 20, 207–220 (2019).

5. Karemaker, I. D. Vermeulen, M. Single-Cell DNA Methylation Profiling: Technologies and Biological Applications. Trends Biotechnol. 36, 952–965 (2018).

6. Mayr, U., Serra, D. Liberali, P. Exploring single cells in space and time during tissue development, homeostasis and regeneration. Development 146, (2019).

7. Regev, A. et al. The Human Cell Atlas. Elife 6, (2017).

8. HuBMAP Consortium. The human body at cellular resolution: the NIH Human Biomolecular Atlas Program. Nature 574, 187–192 (2019).

9. Efremova, M. Teichmann, S. A. Computational methods for single-cell omics across modalities. Nat. Methods 17, 14–17 (2020).

10. Lähnemann, D. et al. Eleven grand challenges in single-cell data science. Genome Biol. 21, 31 (2020).

11. Stuart, T. et al. Comprehensive Integration of Single-Cell Data. Cell vol. 177 1888–1902.e21 (2019).

12. Welch, J. D. et al. Single-Cell Multi-omic Integration Compares and Contrasts Features of Brain Cell Identity. Cell vol. 177 1873–1887.e17 (2019).

13. Lopez, R. et al. A joint model of unpaired data from scRNA-seq and spatial transcriptomics for imputing missing gene expression measurements. arXiv [cs.LG] (2019).

14. GradientBased Learning Applied to Document Recognition. Intelligent Signal Processing (2009) doi:10.1109/9780470544976.ch9.

15. Chen, X., Miragaia, R. J., Natarajan, K. N. Teichmann, S. A. A rapid and robust method for single cell chromatin accessibility profiling. Nat. Commun. 9, 5345 (2018).

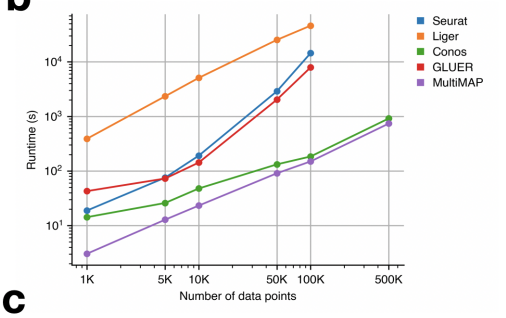
16. Granja, J. M. et al. Single-cell multiomic analysis identifies regulatory programs in mixed-phenotype acute leukemia. Nat. Biotechnol. 37, 1458–1465 (2019).

17. Saunders, A. et al. Molecular Diversity and Specializations among the Cells of the Adult Mouse Brain. Cell 174, 1015–1030.e16 (2018).

a

Method	Transfer	Silhouette	Alignment	Structure	Runtime
Spleen					
None	0.46	0.16	0.00	0.47	–
Seurat 3	0.17	0.15	**0.38**	0.21	133 s
LIGER	0.32	0.22	0.14	0.17	1311 s
Conos	0.46	0.17	0.20	0.35	22 s
GLUER	0.15	0.12	0.13	0.21	103 s
MultiMAP	**0.76**	**0.29**	0.21	**0.56**	**12s**
Visual Cortex					
None	0.26	0.22	0.00	0.33	–
Seurat 3	0.41	0.07	0.08	0.31	1674 s
LIGER	0.62	0.26	0.09	0.28	1524 s
Conos	0.58	0.19	0.04	0.05	280 s
GLUER	0.45	0.14	0.07	0.24	3903 s
MultiMAP	**0.65**	**0.35**	**0.11**	**0.39**	**13 s**
Primary Cortex					
None	0.32	0.15	0.00	0.42	–
Seurat 3	×	×	×	×	×
LIGER	×	×	×	×	×
Conos	0.12	0.15	0.21	0.03	3978 s
GLUER	×	×	×	×	×
MultiMAP	**0.63**	**0.39**	**0.16**	**0.47**	**3811 s**

b



c

Method	Mapping	Max no. datasets	Scalable to large data	Dataset-specific features	Dataset influence on integration
Seurat 3	Nonlinear	2	No	Ignores	Fixed
LIGER	Linear	No max	No	Ignores	Fixed
Conos	Linear	No max	Yes	Ignores	Fixed
GLUER	Nonlinear	No max	No	Ignores	Fixed
MultiMAP	Nonlinear	No max	Yes	Leverages	Can be modulated

Figure 4. **Benchmarking MultiMAP against existing approaches.** a. Comparison of each method in terms of transfer learning accuracy ("Transfer"), separation of cell type clusters as quantified by Silhouette coefficient ("Silhouette"), mixing of different datasets as measured by fraction of nearest neighbours that belong to a different dataset ("Alignment"), preservation of high-dimensional structure as measured by the Pearson correlation between distances in the high- and low-dimensional spaces ("Structure"), and runtime. b. Wall-clock time of multi-omic integration methods on different sized datasets. Seurat 3 and LIGER produced out-of-memory errors when run on 500,000 data points (218 GB RAM). To produce these datasets we subsampled the mouse primary cortex scRNA-seq and scATAC-seq data[20] using geometric sketching[33]. The datasets were subsampled so that there are equal number of cells in the scRNA-seq and scATAC-seq data until 100,000 cells. Since the scATAC-seq data had 81,196 cells in total, for the 500,000 cells comparison, we used an scRNA-seq of 418,804 cells. c. Comparison of capabilities and properties of each method. "Mapping" refers to the nature of the mapping employed by the method; "Max no. datasets" refers to the upper limit in terms of numbers of datasets accepted by the method; "Scalable to large data" refers to allowing a total of over 500,000 cells; "Data-set specific features" is whether the integration method allows information that is not shared across datasets; and "Dataset influence on integration" is whether the user can modulate the weighting of a given dataset relative to the others during the integration.

18. Wang, X. et al. Three-dimensional intact-tissue sequencing of single-cell transcriptional states. Science 361, (2018).

19. Brodmann, K. Brodmann's: Localisation in the Cerebral Cortex. (Springer Science Business Media, 2007).

20. Yao, Z. et al. An integrated transcriptomic and epigenomic atlas of mouse primary motor cortex cell types. 2020.02.29.970558 (2020) doi:10.1101/2020.02.29.970558.

21. Yamawaki, N., Borges, K., Suter, B. A., Harris, K. D. Shepherd, G. M. G. A genuine layer 4 in motor cortex with prototypical synaptic circuit connectivity. Elife 3, e05422 (2014).

22. Barkas, N. et al. Joint analysis of heterogeneous single-cell RNA-seq dataset collections. Nat. Methods 16, 695–698 (2019).

23. Peng, T., Chen, G. M. Tan, K. GLUER: integrative analysis of single-cell omics and imaging data by deep neural network. doi:10.1101/2021.01.25.427845.

24. Muraro, M. J. et al. A Single-Cell Transcriptome Atlas of the Human Pancreas. Cell Syst 3, 385–394.e3 (2016).

25. Segerstolpe, Å. et al. Single-Cell Transcriptome Profiling of Human Pancreatic Islets in Health and Type 2 Diabetes. Cell Metab. 24, 593–607 (2016).

26. Baron, M. et al. A Single-Cell Transcriptomic Map of the Human and Mouse Pancreas Reveals Inter- and Intra-cell Population Structure. Cell Syst 3, 346–360.e4 (2016).

27. Chazarra-Gil, R., van Dongen, S., Kiselev, V. Y. Hemberg, M. Flexible comparison of batch correction methods for single-cell RNA-seq using BatchBench. Nucleic Acids Res. (2021) doi:10.1093/nar/gkab004.

28. Roels, J. et al. Distinct and temporary-restricted epigenetic mechanisms regulate human and T cell development. Nat. Immunol. 21, 1280–1292 (2020).

29. Jia, G. et al. Single cell RNA-seq and ATAC-seq analysis of cardiac progenitor cell transition states and lineage settlement. Nat. Commun. 9, 4877 (2018).

30. Chen, H. et al. Single-cell trajectories reconstruction, exploration and mapping of omics data with STREAM. Nat. Commun. 10, 1903 (2019).

31. Park, J.-E. et al. A cell atlas of human thymic development defines T cell repertoire formation. Science 367, (2020).

32. Hosokawa, H. Rothenberg, E. V. How transcription factors drive choice of the T cell fate. Nature Reviews Immunology (2020) doi:10.1038/s41577-020-00426-6.

33. Hie, B., Cho, H., DeMeo, B., Bryson, B. Berger, B. Geometric Sketching Compactly Summarizes the Single-Cell Transcriptomic Landscape. Cell Syst 8, 483–493.e7 (2019).

34. Lecun, Y., Bottou, L., Bengio, Y. Haffner, P. Gradient-based learning applied to document recognition. Proceedings of the IEEE vol. 86 2278–2324 (1998).

35. Hern, W. M. Correlation of fetal age and measurements between 10 and 26 weeks of gestation. Obstet. Gynecol. 63, 26–32 (1984).

36. van den Brink, S. C. et al. Single-cell sequencing reveals dissociation-induced gene expression in tissue subpopulations. Nat. Methods 14, 935 (2017).

37. Fang, R. et al. Fast and Accurate Clustering of Single Cell Epigenomes Reveals Cis-Regulatory Elements in Rare Cell Types. doi:10.1101/615179.

38. Schep, A. N., Wu, B., Buenrostro, J. D. Greenleaf, W. J. chromVAR: inferring transcription-factor-associated accessibility from single-cell epigenomic data. Nat. Methods 14, 975–978 (2017).

39. Blondel, V. D., Guillaume, J.-L., Lambiotte, R. Lefebvre, E. Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment vol. 2008 P10008 (2008).

40. Haghverdi, L., Büttner, M., Wolf, F. A., Buettner, F. Theis, F. J. Diffusion pseudotime robustly reconstructs lineage branching. Nat. Methods 13, 845 (2016).

41. Wolf, F. A., Angerer, P. Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. Genome Biol. 19, 15 (2018).

42. Van den Berge, K. et al. Trajectory-based differential expression analysis for single-cell sequencing data. Nat. Commun. 11, 1201 (2020).

43. Wolock, S. L., Lopez, R. Klein, A. M. Scrublet: Computational Identification of Cell Doublets in Single-Cell Transcriptomic Data. Cell Syst 8, 281–291.e9 (2019).

44. Grytten, I. et al. Graph Peak Caller: Calling ChIP-seq peaks on graph-based reference genomes. PLoS Comput. Biol. 15, e1006731 (2019).

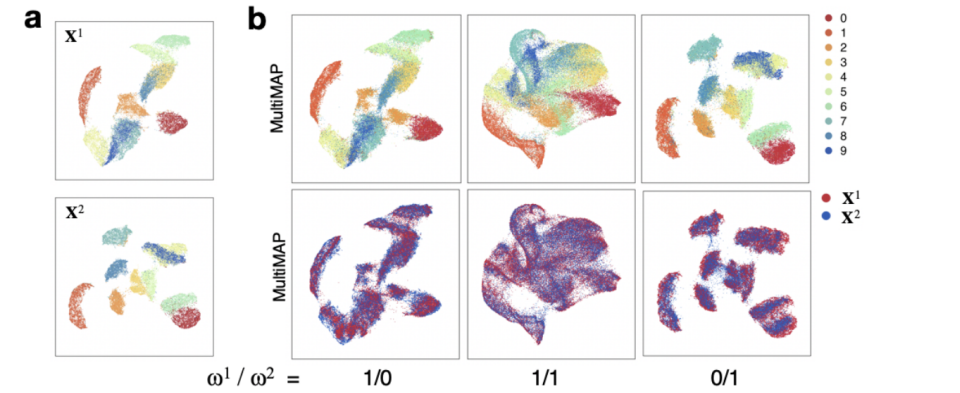
45. Zhu, X., Ghahramani, Z. Lafferty, J. D. Semi-supervised learning using gaussian fields and harmonic functions. in Proceedings of the 20th International conference on Machine learning (ICML-03) 912–919 (2003).

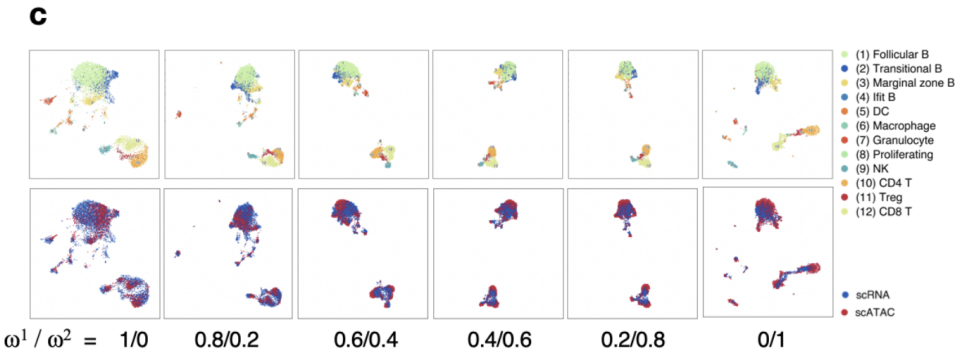
46. Pliner, H. A. et al. Cicero Predicts cis-Regulatory DNA Interactions from Single-Cell Chromatin Accessibility Data. Mol. Cell 71, 858–871.e8 (2018).

47. Kozareva, V. et al. A transcriptomic atlas of the mouse cerebellum reveals regional specializations and novel cell types. doi:10.1101/2020.03.04.976407.

48. Rousseeuw, P. J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics vol. 20 53–65 (1987).

5. Supplemental

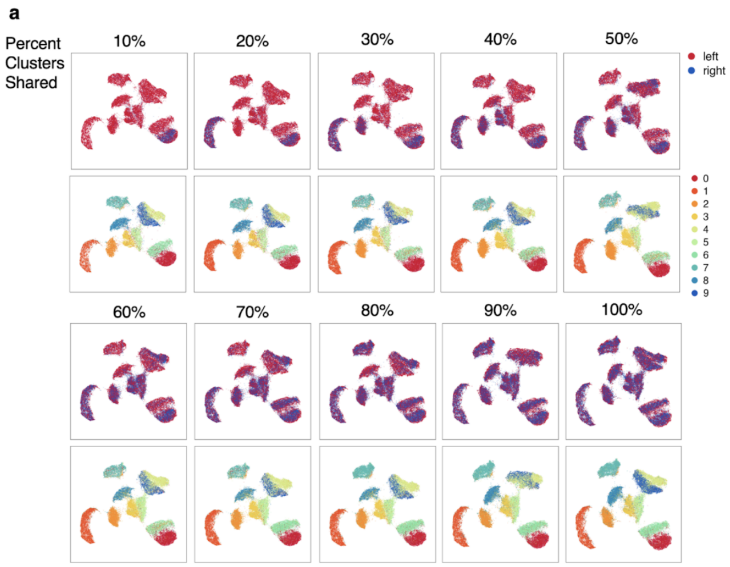




d

ω^1 / ω^2	1/0	0.8/0.2	0.6/0.4	0.4/0.6	0.2/0.8	0/1
Transfer	0.76	0.75	0.76	0.72	0.71	0.73
Silhouette	0.33	0.29	0.30	0.31	0.28	0.28
Alignment	0.23	0.25	0.24	0.26	0.25	0.24
Structure	0.57	0.58	0.58	0.60	0.59	0.55

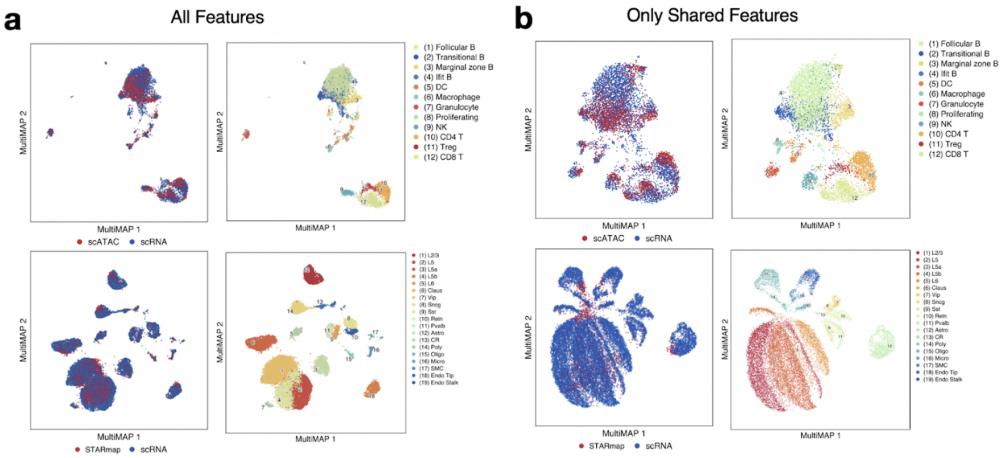
Figure S1. **MultiMAP's weight parameter.** a. UMAP projections of the two halves of the MNIST handwritten digit images. b. MultiMAP embeddings as the weight parameters are varied. Each color is a different handwritten digit (0-9). When ω^1 is larger than ω^2, the embedding more closely resembles the projection of only $\mathbf{X}^1$; when ω^2 is larger than ω^1, the embedding more closely resembles the projection of only $\mathbf{X}^2$. For different choices of ω^v, the datasets are well integrated in the embedding space. c. MultiMAP integration with varied weight parameters of published scATAC-seq and newly generated scRNA-seq data of the mouse spleen (n=1) [15]. d. Comparison of the MultiMAP integration of the spleen data as the weight parameter is varied – in terms of transfer learning accuracy ("Transfer"), separation of cell type clusters as quantified by Silhouette coefficient ("Silhouette"), and preservation of high-dimensional structure as measured by the Pearson correlation between distances in the high- and low-dimensional spaces ("Structure")



b

Percent Clusters Shared	Transfer	Silhouette	Structure
10%	0.70	0.28	0.21
20%	0.79	0.28	0.43
30%	0.78	0.27	0.34
40%	0.76	0.27	0.31
50%	0.74	0.24	0.27
60%	0.75	0.24	0.24
70%	0.77	0.25	0.23
80%	0.80	0.27	0.23
90%	0.77	0.24	0.24
100%	0.80	0.26	0.24

Figure S2. **MultiMAP integration with non-shared clusters.** a. MultiMAP integration of the left and right halves of MNIST handwritten digit images with a 2 pixel wide shared region. Gaussian noise is added to the left half. MultiMAP integration is performed with a varying number of digit clusters removed from the right dataset, so that the integration ranges from one shared cluster (10%) to all clusters shared (100%). b. Comparison of the MultiMAP integration of the modified MNIST dataset as the percent of clusters shared is varied – in terms of transfer learning accuracy ("Transfer"), separation of cell type clusters as quantified by Silhouette coefficient ("Silhouette"), and preservation of high-dimensional structure as measured by the Pearson correlation between distances in the high- and low-dimensional spaces ("Structure").



c

Spleen	All Features	Only Shared Features
Transfer	0.75	0.55
Silhouette	0.29	0.15
Alignment	0.25	0.10
Structure	0.58	0.10

Visual Cortex	All Features	Only Shared Features
Transfer	0.67	0.28
Silhouette	0.34	0.17
Alignment	0.12	0.14
Structure	0.38	0.24

Figure S3. **MultiMAP integration with all features vs. only shared features in the spleen scRNA-seq + scATACseq, and visual cortex STARmap + scRNAseq datasets.** a. MultiMAP embeddings using all genes present in each dataset (intended use of MultiMAP). b. MultiMAP embeddings using only genes shared by all datasets in each integration. c. Comparison of the MultiMAP integration with all features vs. only shared features – in terms of transfer learning accuracy ("Transfer"), separation of cell type clusters as quantified by Silhouette coefficient ("Silhouette"), mixing of different datasets as measured by fraction of nearest neighbours that belong to a different dataset ("Alignment"), and preservation of high-dimensional structure as measured by the Pearson correlation between distances in the high- and low-dimensional spaces ("Structure").