nature methods

Registered Report

https://doi.org/10.1038/s41592-025-02856-3

Multitask benchmarking of single-cell multimodal omics integration methods

Received: 10 December 2023

Accepted: 8 September 2025

Published online: 13 October 2025

Check for updates

Chunlei Liu[1,2,3], Sichang Ding[1,3], Hani Jieun Kim[1,2,3,5], Siqu Long[1,3], Di Xiao[1,3], Shila Ghazanfar[2,3,4] & Pengyi Yang[1,2,3,4]

Single-cell multimodal omics technologies have empowered the profiling of complex biological systems at a resolution and scale that were previously unattainable. These biotechnologies have propelled the fast-paced innovation and development of data integration methods, leading to a critical need for their systematic categorization, evaluation and benchmarking. Navigating and selecting the most pertinent integration approach poses a considerable challenge, contingent upon the tasks relevant to the study goals and the combination of modalities and batches present in the data at hand. Understanding how well each method performs multiple tasks, including dimension reduction, batch correction, cell type classification and clustering, imputation, feature selection and spatial registration, and at which combinations will help guide this decision. Here we develop a much-needed guideline on choosing the most appropriate method for single-cell multimodal omics data analysis through a systematic categorization and comprehensive benchmarking of current methods. The stage 1 protocol for this Registered Report was accepted in principle on 30 July 2024. The protocol, as accepted by the journal, can be found at https://springernature.figshare.com/articles/journal_contribution/Multi-task_benchmarking_of_single-cell_multimodal_omics_integration_methods/26789902.

The recent development of single-cell multimodal omics technologies has revolutionized our ability to simultaneously profile multilayered molecular programs at a global scale in individual cells[1,2]. The availability of single-cell multimodal omics data provides new opportunities to investigate molecular programs that underlie cell identity, cell-fate decisions and disease mechanisms at a resolution previously inaccessible from studying bulk cell populations[1–3]. A considerable number of technologies and platforms have been developed, which have led to the generation of a variety of combinations of data modalities, each capturing a unique molecular feature in the cell. So far, prevalent modalities encompass gene expression (typically denoted as RNA), surface protein abundance (through antibody-derived tags (ADT)) and chromatin accessibility (through assay for transposase-accessible chromatin (ATAC)) and can be profiled jointly by popular single-cell multimodal omics technologies such as CITE-seq[4], SHARE-seq[5] and TEA-seq[6] to name a few (Fig. 1a).

Integrating modalities of data generated from single-cell multimodal omics technologies is essential and greatly impacts the utility of such data for downstream biological interpretation[7–9]. A fast-growing number of computational methods have been developed to integrate various combinations of data modalities captured by different single-cell multimodal omics technologies. Nevertheless, there is a lack of study that comprehensively outlines the similarities and differences between current methods including the data integration strategies, the tasks they intend to address and the data modalities they

[1]Computational Systems Biology Unit, Children's Medical Research Institute, Faculty of Medicine and Health, University of Sydney, Westmead, New South Wales, Australia. [2]School of Mathematics and Statistics, Faculty of Science, University of Sydney, Sydney, New South Wales, Australia. [3]Sydney Precision Data Science Centre, University of Sydney, Sydney, New South Wales, Australia. [4]Charles Perkins Centre, University of Sydney, Sydney, New South Wales, Australia. [5]Present address: The Kinghorn Cancer Centre and Cancer Research Theme, Garvan Institute of Medical Research, Darlinghurst, New South Wales, Australia. e-mail: pengyi.yang@sydney.edu.au

https://doi.org/10.1038/s41592-025-02856-3

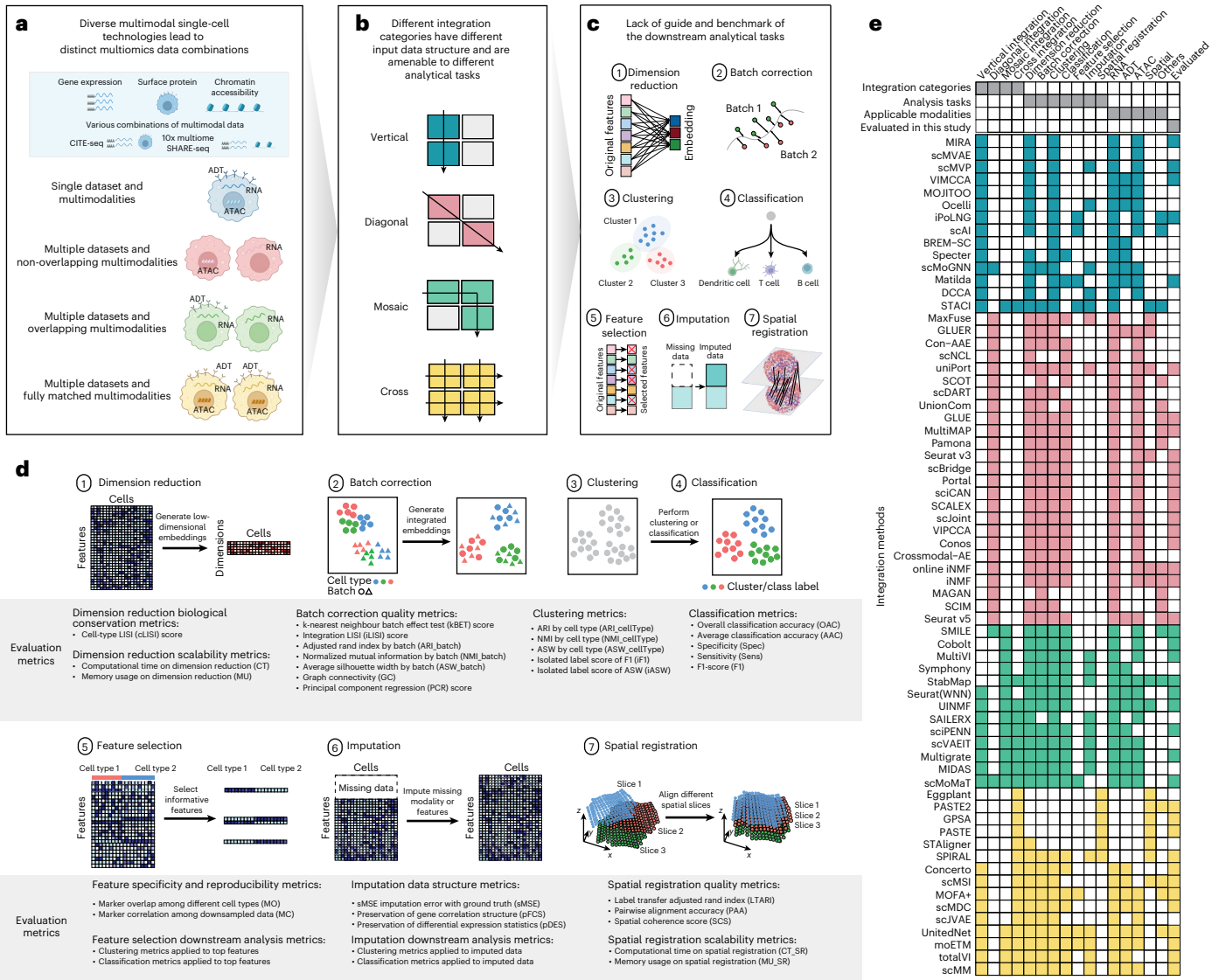


Fig. 1 | An overview of benchmarking for single-cell multimodal omics integration methodologies. a, Diverse technological combinations arising from different modalities. **b**, Four prototypical single-cell multimodal omics data integration categories. **c**, Schematics of common tasks for single-cell multimodal omics analysis. **d**, Schematics for each task and the panel of evaluation metrics tailored for its performance evaluation. **e**, Summary of integration methods and their alignment with specific integration categories, computational tasks, modality capacities and evaluation status.

can integrate. Furthermore, while a few works have been conducted to compare integration methods for cancer studies using both bulk and single-cell omics data[10,11], there is a lack of systematic evaluation and benchmark of the currently available methods specifically designed for single-cell multimodal omics data integration for their performance on the broad range of tasks and data modalities they are designed to handle. These, together, have created a considerable challenge for navigating and selecting the most pertinent integration method for single-cell multimodal omics data analysis.

To address the challenge of selecting the most appropriate methods for single-cell multimodal omics data analysis, here we set out to systematically categorize and comprehensively benchmark the current integration methods on common computational tasks and data modalities. Building on previous works[7,12] and based on input data structure and modality combination, we define four prototypical single-cell multimodal omics data integration categories: 'vertical', 'diagonal', 'mosaic' and 'cross' integration (Fig. 1b). Depending on the applications, we further introduce seven common tasks that many methods are designed to address (Fig. 1c), including (1) dimension reduction, (2) batch correction, (3) clustering, (4) classification, (5) feature selection, (6) imputation and (7) spatial registration. Using panels of evaluation metrics, each tailor-made for a specific task (Fig. 1d), we evaluated 40 integration methods (Fig. 1e) across the 4 data integration categories on 64 real datasets and 22 simulated datasets. In particular, we included 18 vertical integration methods, 14 diagonal integration methods, 12 mosaic integration methods and 15 cross integration methods. This Registered Report provides a comprehensive view of methods designed for single-cell multimodal omics data integration and serves as (1) a much-needed guide for single-cell integration across modalities to the research community and (2) a foundation to foster future research and application of computational methods for single-cell multimodal omics data analysis.

The study was conducted in accordance with the registered, peer-reviewed protocol at https://springernature.figshare.com/articles/journal_contribution/Multi-task_benchmarking_of_single-cell_multimodal_omics_integration_methods/26789902. Except for preregistered data, all analysis results reported in the Registered Report were collected after the date of the registered protocol publication.

https://doi.org/10.1038/s41592-025-02856-3

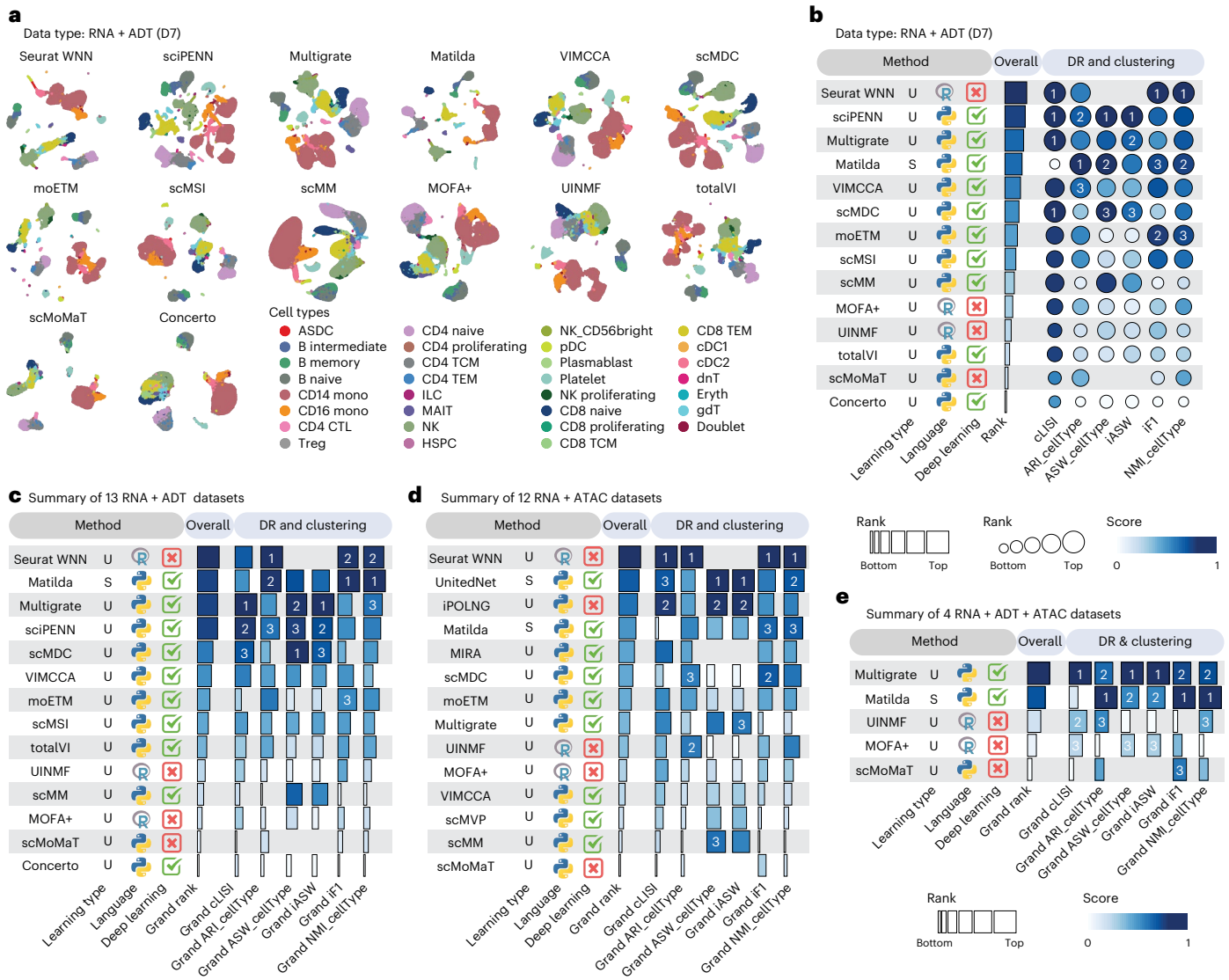


Fig. 2 | Benchmark results of vertical integration methods for dimension reduction and clustering. a, UMAP visualization of the vertical integration methods applied to a representative RNA + ADT dataset (D7). Cell type labels were obtained from the original publications. **b**, Method performance on the dataset D7. Dimension reduction (DR) and clustering metrics are used to evaluate method performance. Overall rank scores (horizontal bars) are computed as the min–max scaled mean ranks across evaluation metrics. Individual metric scores (bubbles) are computed as the min–max scaled raw values. **c**–**e**, Performance summary of vertical integration methods from all RNA + ADT datasets (**c**), all RNA + ATAC datasets (**d**) and all RNA + ADT + ATAC datasets (**e**). Overall grand rank scores are computed as the min–max scaled mean rank of the grand ranks across each metric. Grand ranks of individual metrics are computed as the min–max scaled mean ranks across all applicable datasets. For all summary panels, the learning types are categorized into three groups: supervised (S), unsupervised (U) and semi-supervised (Semi). Metrics that do not apply to a given method were not assigned a value.

Results

Vertical integration for dimension reduction and clustering

To evaluate the performance of vertical integration methods on dimension reduction and clustering tasks, we systematically benchmarked all applicable methods and datasets of varying modalities. This included evaluations of 14 methods[13–26] on 13 paired RNA and ADT (RNA + ADT) datasets, 14 methods[17–30] on 12 paired RNA and ATAC (RNA + ATAC) datasets, and 5 methods[17–21] on 4 datasets containing all three modalities (RNA + ADT + ATAC).

As an illustrative example, the results from different methods on a representative dataset D7, with paired RNA and ADT data, are visualized in Fig. 2a and quantitatively summarized in Fig. 2b. Among the evaluated methods, Seurat WNN[22], sciPENN[14] and Multigrate[19] demonstrated generally better performance on this dataset, effectively preserving the biological variation of cell types. While the evaluation metrics generally agreed in method assessment, notable differences in ranking were observed. For instance, moETM[25] was one of the top-ranked methods by iF1 and NMI_cellType, yet received comparatively low rankings based on ASW_cellType and iASW. Uniform Manifold Approximation and Projection (UMAP) visualizations and metric quantifications for additional representative datasets, D15 (RNA + ATAC) and D22 (RNA + ADT + ATAC), are shown in Supplementary Fig. 1a,b and Supplementary Fig. 1c,d, respectively. We note that Seurat WNN, MIRA and scMoMaT generate graph-based outputs rather than embeddings, making ASW_cellType and iASW metrics inapplicable for these methods.

To summarize the method performance across all datasets for the dimension reduction and clustering tasks, we first calculated the overall rank score for each dataset by summarizing method performance across evaluation metrics (Supplementary Fig. 1e–g). Boxplots of ranks for each method visualize the variation in method performance across individual datasets (Supplementary Fig. 1h–j).

Methods were then ranked on the basis of their overall grand rank scores for 13 bimodal datasets of RNA + ADT (Fig. 2c), 12 bimodal datasets of RNA + ATAC (Fig. 2d) and 4 trimodal datasets of RNA + ADT + ATAC (Fig. 2e). These analyses reveal that dataset complexity may affect the performance of integration methods. Simulated datasets, which may lack latent data structure observed in real data, can be easier to integrate, and some methods (for example, scMM[23]) that do not perform well on real datasets tend to perform better on simulated datasets (Supplementary Fig. 1e,f). While Seurat WNN, Multigrate, Matilda[17] and UnitedNet[29] in general performed well across diverse datasets for particular data modality combinations (Supplementary Fig. 1e,f), method performance is both dataset dependent and, more notably, modality dependent.

Benchmark vertical integration for feature selection

Feature selection is typically used to identify molecular markers associated with specific cell types[31]. Among the vertical integration methods, only Matilda[17], scMoMaT[21] and MOFA+[18] support feature selection of molecular markers from single-cell multimodal omics data. Notably, Matilda and scMoMaT are capable of identifying distinct markers for each cell type in a dataset, whereas MOFA+ selects a single cell-type-invariant set of markers for all cell types. Extended Data Fig. 1a visualizes the top cell-type-specific markers selected from the RNA and ADT modalities of dataset D8 (RNA + ADT) for three example cell types (that is, CD14 monocytes ('CD14 mono'), natural killer (NK) cells ('NK') and plasmablast cells ('Plasmablast')) by Matilda and scMoMaT. MOFA+ is excluded from this visualization as its marker selection does not associate with a particular cell type. For both Matilda and scMoMaT, the selected markers show higher gene expression (RNA) and protein abundance (ADT) in their respective cell types compared with other cell types. Moreover, for RNA of NK cells, ADT of CD14 monocytes and ADT of plasmablast cells, the two methods identified the same top markers. The quantification of method performance on feature selection for each data modality, using the union of top-5 markers selected from all cell types, in D8 is presented in Extended Data Fig. 1b. To ensure comparability, the number of markers selected by MOFA+ was matched with the other two methods. Results from D8 suggest that the assessments of selected markers can differ across clustering and classification metrics. Visualization and quantification of feature selection results from additional representative datasets with varying combinations of data modalities shown in Extended Data Fig. 2a–d, which demonstrates more consistent evaluation results across clustering and classification metrics.

To summarize the method performance across all datasets for the feature selection task, we first calculated the overall rank score of each method in each dataset for the three evaluation categories of clustering, classification and reproducibility for datasets with RNA + ADT modalities (Supplementary Fig. 2a), RNA + ATAC modalities (Supplementary Fig. 2b) and RNA + ADT + ATAC modalities (Supplementary Fig. 2c). Variations in method performance across individual datasets are visualized in Extended Data Fig. 2e,f and Supplementary Fig. 2d. Methods were then ranked on the basis of their overall grand rank scores for each combination of data modalities (Extended Data Fig. 1c–e). These results reveal that MOFA+, while unable to select cell-type-specific markers, generated more reproducible feature selection results across different data modalities. Note that marker correlation (MC) is not influenced by which top features are chosen. Instead, it measures the correlation across all features, leading to the same values even when the selected markers vary within the same method. However, features selected by scMoMaT and Matilda generally led to better clustering and classification of cell types than those by MOFA+. It is also interesting to note that markers selected from ATAC modality (that is, peaks) typically underperform in clustering and classification evaluations compared with those selected from RNA and ADT modalities (Supplementary Fig. 2b,c). Furthermore, increasing the number of top markers did not always improve downstream evaluation results. Importantly, method performance varies depending on the datasets and evaluation metrics, underscoring the necessity of using multiple evaluation metrics for robust method assessment.

Benchmark diagonal integration methods

We systematically benchmarked 14 diagonal integration methods[32–45] on dimension reduction, clustering, batch correction and classification tasks, including evaluations on 12 datasets with a single batch of RNA data and a single batch of ATAC data, and 5 methods[40–44] on 6 datasets comprising multiple batches of RNA data and multiple batches of ATAC data. As an example, the evaluation results of different methods on a representative dataset D27, which consists of one RNA batch and one ATAC batch, are visualized in Fig. 3a and quantified using different evaluation metrics in Fig. 3b. Among the methods, scBridge[32], GLUE[44], Seurat v5[38], uniPort[45] and scJoint[43] achieved the best overall performance on this dataset, effectively removing batch effects while preserving the biological variation of cell types. Of note, GLUE and Seurat v5 integrate RNA and ATAC data using peak features, whereas the other methods require transforming ATAC peaks into gene activity scores. While the evaluation metrics largely agreed in method assessment, some differences in ranking were observed on this dataset. For instance, scBridge was top-ranked in dimension reduction and clustering, whereas the other high-performing method GLUE was top-ranked in batch correction. Moreover, rankings of methods in the classification category further deviate from those in dimension reduction and clustering and batch correction and could be confounded by the use of the multilayer perceptron (MLP) classifier implemented in this work and, if available, the original classifiers implemented in the methods. Results from a representative dataset D37 with multiple batches of RNA and ATAC data are shown in Supplementary Fig. 3a,b. Fewer methods can handle such data, and among those applicable, GLUE and scJoint demonstrated the best performance on this dataset.

To summarize the method performance across all datasets for the dimension reduction, clustering, batch correction and classification tasks, we first calculated the overall rank score for each dataset (Supplementary Fig. 3c,d), visualized the performance variability across datasets (Supplementary Fig. 3e,f) and then ranked methods on the basis of their overall grand rank scores for the datasets with single RNA and ATAC batches (Fig. 3c) and for datasets with multiple RNA and ATAC batches (Fig. 3d). These results reveal that methods such as scBridge and scJoint consistently demonstrated top performance in dimension reduction and clustering tasks across datasets. However, their effectiveness in batch correction was moderate, suggesting limitations in their ability to fully harmonize batches. By contrast, methods such as GLUE, uniPort and Portal[33], which excel at batch correction, did not demonstrate similarly strong performance in dimension reduction and clustering. This observation indicates a potential trade-off between preserving biological signals, captured by dimension reduction and clustering results, and effective batch correction. In addition, except for Seurat v3[36], which performed better with its specifically designed classifier, most integration methods showed similar overall classification rank scores regardless of the classifier used, indicating that the performance depends more on the quality of the integration method itself rather than the choice of classifier. This highlights that, while classifier choice may influence results for some methods, the effectiveness of the integration method is the primary determinant of classification performance.

Taken together, in datasets with single RNA and ATAC batches, scBridge was the top performer across all metrics for dimension reduction and clustering, while GLUE excelled in batch correction metrics. By contrast, for datasets with multiple RNA and ATAC batches, scJoint ranked highest across all dimension reduction and clustering metrics, with GLUE ranking second and demonstrating strong performance on batch correction metrics, highlighting its effectiveness in addressing batch effects.

https://doi.org/10.1038/s41592-025-02856-3

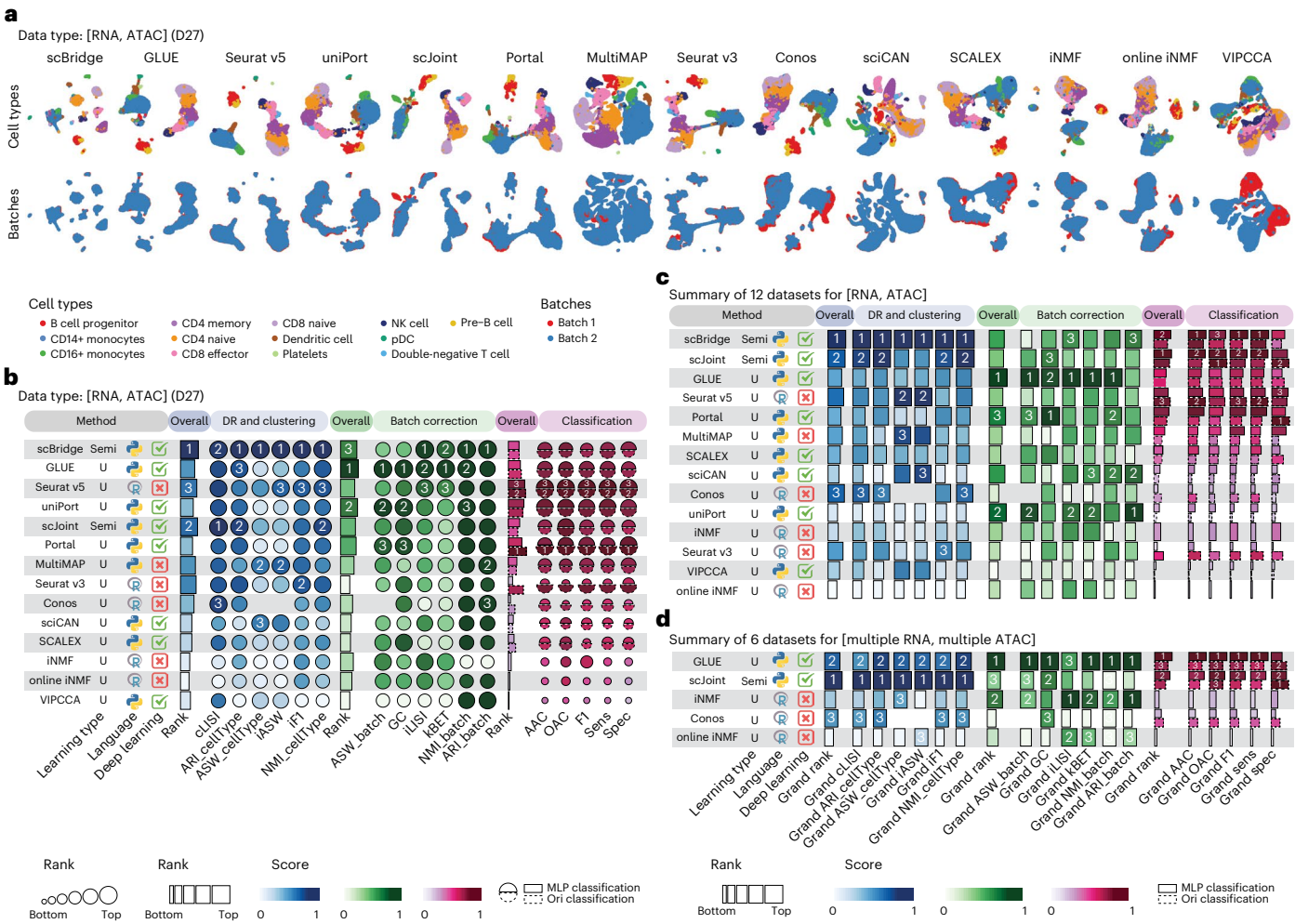


Fig. 3 | Benchmark results of diagonal integration methods. a, Visualization of diagonal integration methods applied to a representative dataset of D27 with a single batch of RNA data and a single batch of ATAC data. **b**, Method performance on the dataset D27. Metrics are categorized into dimension reduction (DR) and clustering (blue tab), batch correction (green tab) and classification (pink tab). For classification, the solid line represents the MLP classifier implemented in this work, while the dashed line represents the classifier proposed in the original papers. **c**,**d**, Performance summary of diagonal integration from all datasets with a single batch of RNA data and a single batch of ATAC data (**c**) and all datasets with multiple batches of RNA data and multiple batches of ATAC data (**d**).

Benchmark mosaic integration methods

We next benchmarked seven mosaic integration methods (that is, MultiVI[46], Cobolt[47], SMILE[48], scMoMaT[21], Multigrate[19], StabMap[49] and UINMF[20]) for dimension reduction, clustering, batch correction and classification tasks, on 17 datasets spanning four data types. As an illustrative example, Fig. 4a visualizes UMAP projections of results generated by different methods on dataset D45, which comprises three batches, including an RNA batch, an ATAC batch and an RNA + ATAC batch serving as a bridge between data modalities. Method performance on this dataset was quantified in Fig. 4b. Among the methods, MultiVI, Cobolt and SMILE are specifically designed for this type of data. By contrast, scMoMaT, Multigrate and StabMap are more flexible, as they require only the presence of bridge features between different batches for any data modality combinations. Results from additional datasets with other data modality combinations are presented in Supplementary Fig. 4a–f. In particular, only scMoMaT, Multigrate and StabMap were applicable to D40, a dataset with three batches (RNA, ADT, RNA + ADT), and D49 (RNA + ADT, RNA + ATAC, ADT). Beside these three methods, UINMF was also applicable to D46 (RNA + ADT, RNA + ATAC, RNA). UINMF is unique in that it requires common features across all batches and therefore is limited in its applicability.

To evaluate the overall performance of mosaic integration methods, we first ranked the methods on the basis of their overall rank scores in each dataset and within each task category. In Supplementary Fig. 4g, datasets with (RNA, RNA + ADT, ADT) and those with mixed combinations with or without a shared modality are combined into a single panel, while datasets with (RNA, RNA + ATAC, ATAC) are shown separately in Supplementary Fig. 4h. The overall grand rank scores are presented in Fig. 4c–f, where Fig. 4c summarizes results for datasets with (RNA, RNA + ADT, ADT), Fig. 4d presents results for datasets with (RNA, RNA + ATAC, ATAC), and Fig. 4e,f summarizes results for datasets with mixed combinations of modalities and with or without shared features. The method performance variability across individual datasets is visualized in Supplementary Fig. 4i,j. These results reveal that StabMap performed particularly well in dimensionality reduction, clustering and classification tasks across most datasets. scMoMaT and Multigrate show strong performance in batch correction across datasets and data types. When analyzing datasets with (RNA, RNA + ATAC, ATAC), MultiVI and Cobolt consistently achieve strong performance across tasks and datasets, probably because they are specifically designed for such data types.

Overall, these analyses provide guidance on method selection for specific tasks and input data types but also highlight that there are relatively few mosaic integration methods available compared with other integration categories, as many of them impose strict restrictions, limiting their flexibility. This lack of flexibility can make these

https://doi.org/10.1038/s41592-025-02856-3

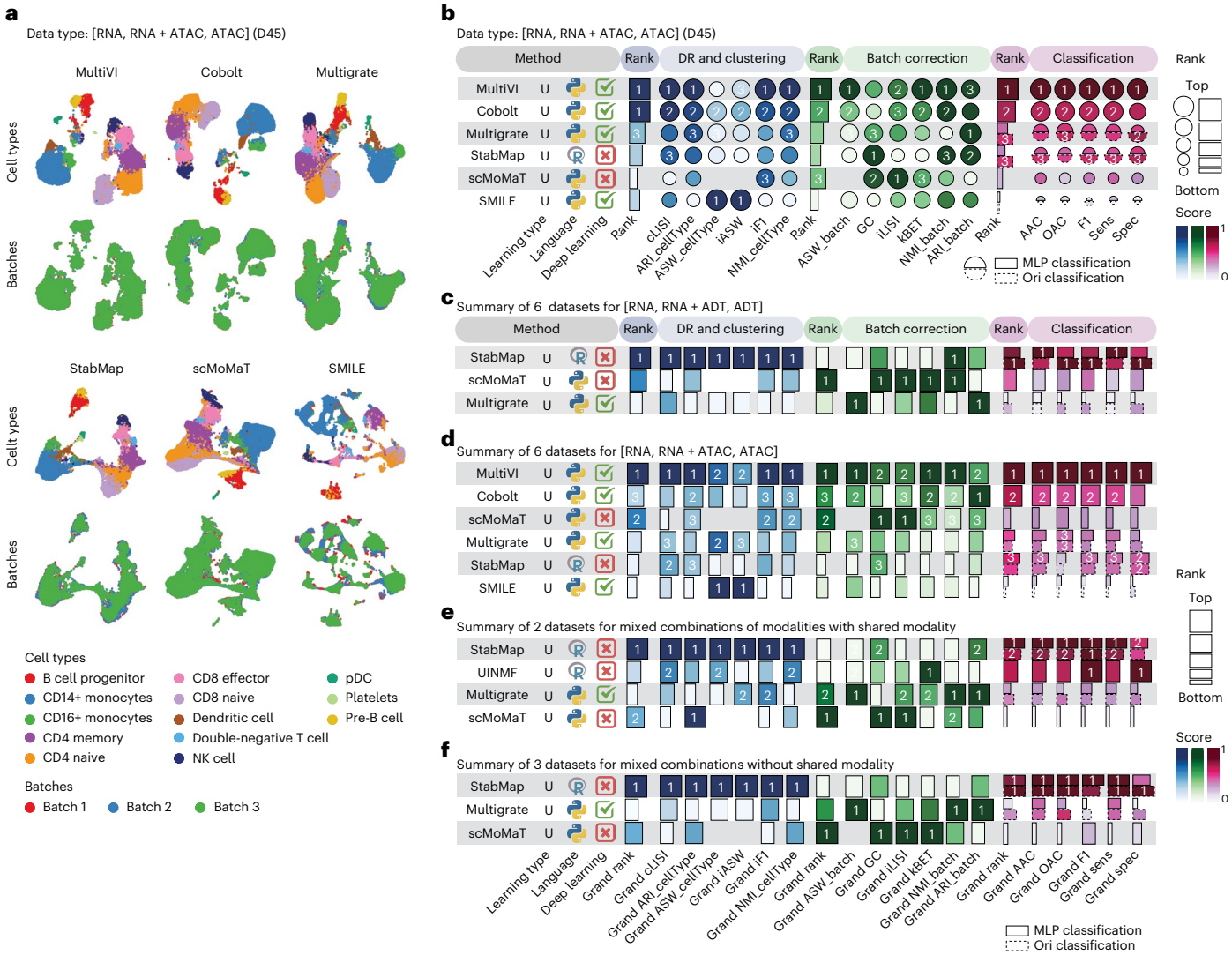


Fig. 4 | Benchmark results of mosaic integration methods. a, UMAP visualization of mosaic integration methods applied to a representative dataset D45 with data batches of RNA, paired RNA + ATAC and ATAC. **b**, Method performance on the dataset D45. Metrics are categorized into dimension reduction (DR) and clustering (blue tab), batch correction (green tab) and classification (pink tab). For classification, the solid line represents the MLP classifier implemented in this work, while the dashed line represents the classifier proposed in the original papers. **c**–**f**, Performance summary of mosaic integration from all datasets with data batches of RNA, paired RNA + ADT and ADT (**c**), all datasets with data batches of RNA, paired RNA + ATAC and ATAC (**d**), all datasets with mixed batch combinations with shared modality across batches (**e**) and all datasets with mixed batch combinations but without shared modality across batches (**f**).

methods less suitable for datasets with diverse modalities or limited shared features, highlighting the need for more flexible approaches in mosaic integration.

Benchmark methods on data imputation for mosaic data

While the mosaic integration methods benchmarked in the previous section can be directly applied to mosaic data, several additional methods can also be applied to mosaic data by first imputing missing data modalities and subsequently performing integrative analysis. These include sciPENN[14], moETM[25], scMM[23], totalVI[13] and UnitedNet[29]. Notably, MultiVI[46] and StabMap[49] also offer options for data imputation, although they are applicable to mosaic data integration without the imputation step. Among these methods, totalVI and sciPENN can only impute the ADT modality, whereas the other methods can impute both RNA and another modality (ADT or ATAC).

Extended Data Fig. 3a–c shows imputation results, where we imputed the missing ADT modality in a representative CITE-seq dataset D53. UnitedNet and MultiVI were excluded from this analysis as they do not support imputation in this scenario. On this dataset, we found that, while data imputed by scMM showed good structural similarity to the ground truth, this did not translate to better clustering and classification results. Conversely, data imputed by sciPENN showed good performance when assessed by clustering and classification metrics but have the lowest structural similarity to the ground truth compared with those imputed by other methods. Similar results were found from additional imputation analyses. These include imputing the RNA modality for the CITE-seq dataset D53 (Supplementary Fig. 5a–c) and imputing ATAC (Supplementary Fig. 5d–f) and RNA (Supplementary Fig. 5g–i) modalities for the 10x multiome dataset D56. These findings highlight that strong preservation of structural similarity during imputation does not necessarily lead to improved performance on downstream analytical tasks, emphasizing the importance of evaluating imputation methods across multiple complementary metrics.

To summarize the method performance across all datasets for the imputation task, we first calculated the overall rank score of each method in each dataset for the evaluation categories of clustering,

https://doi.org/10.1038/s41592-025-02856-3

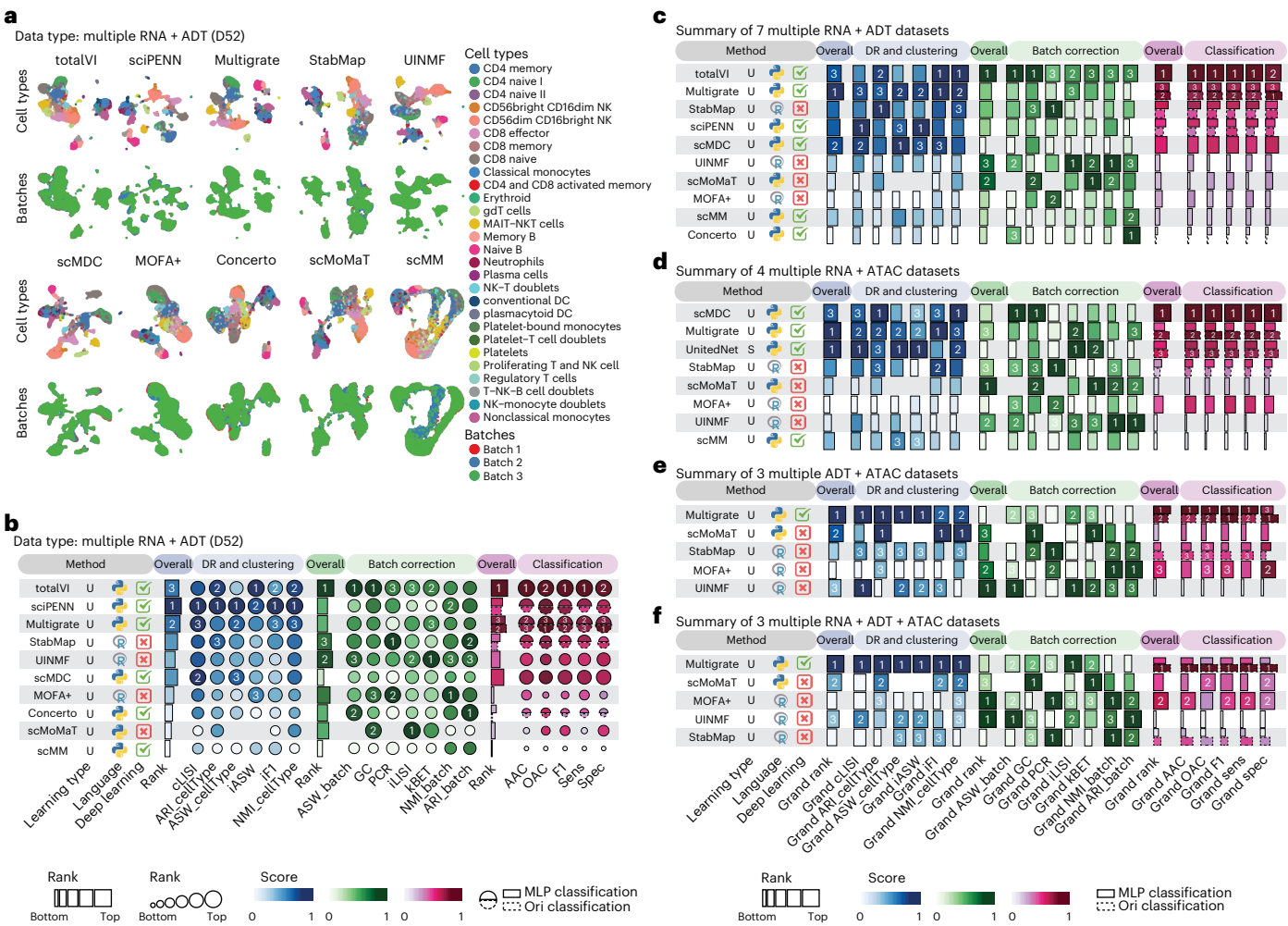


Fig. 5 | Benchmark results of cross integration methods. a, UMAP visualization of cross integration methods applied to a representative dataset of D52. **b**, Method performance on the dataset D52. Metrics are categorized into dimension reduction (DR) and clustering (blue tab), batch correction (green tab) and classification (pink tab). For classification, the solid line represents the MLP classifier implemented in this study, while the dashed line represents the classifier proposed in the original papers. **c**–**f**, Performance summary of cross integration methods applied to all datasets with paired RNA + ADT data across multiple batches (**c**), all datasets with paired RNA + ATAC data across multiple batches (**d**), all datasets with paired ADT + ATAC data across multiple batches (**e**) and all datasets with paired RNA + ADT + ATAC data across multiple batches (**f**).

classification and structural similarity (Supplementary Fig. 5j–m) and visualized their variability in Supplementary Fig. 5n–q. Methods were then ranked on the basis of their overall grand rank scores for each imputation scenario, including imputing ADT (Extended Data Fig. 3d) and RNA (Extended Data Fig. 3e) for CITE-seq datasets and imputing ATAC (Extended Data Fig. 3f) and RNA (Extended Data Fig. 3g) for multiome datasets. These results revealed that totalVI and StabMap imputed ADT data performed well in clustering and classification evaluations and demonstrated strong structural similarity to those in ground-truth data. Although data imputed by sciPENN achieved the best performance in clustering and classification, their structural similarity did not rank as high as the others. For imputing ATAC or RNA data from 10x multiome, MultiVI demonstrated good performance under clustering and classification metrics, while StabMap and scMM performed well under structural similarity metrics. These analyses highlight that imputation performance alone does not determine the overall effectiveness of a method in mosaic data analysis, and a comprehensive evaluation across multiple criteria is essential.

Benchmark cross integration methods

We next benchmarked methods for dimension reduction, clustering, batch correction and classification tasks on cross integration where all data modalities are present in each of all batches in a dataset. This included seven multibatch bimodal RNA + ADT datasets, four multibatch bimodal RNA + ATAC datasets, three multibatch bimodal ADT + ATAC datasets and three multibatch trimodal RNA + ADT + ATAC datasets. Results from different methods on a representative dataset D52 are visualized in Fig. 5a and quantified by evaluation metrics in Fig. 5b. These results highlight the agreements and differences in performance assessment depending on the evaluation metrics. In particular, performance rankings from dimension reduction, clustering and classification metrics tend to have higher concordance on this dataset, whereas those from batch correction metrics tend to lead to different rankings. Similar results were found in additional datasets including D56 (multibatch RNA + ATAC), D58 (multibatch ADT + ATAC) and D59 (multibatch RNA + ADT + ATAC) in Supplementary Fig. 6a–f.

We summarized the method performance in each dataset by their overall rank scores (Supplementary Fig. 6g–l) and then ranked methods by their grand overall rank scores across dataset categories, including those from RNA + ADT datasets (Fig. 5c), RNA + ATAC datasets (Fig. 5d), ADT + ATAC datasets (Fig. 5e) and RNA + ADT + ATAC datasets (Fig. 5f). These analyses reaffirmed the above-observed discordance between batch correction metrics and other evaluation metrics, highlighting the challenge of achieving balanced performance

across tasks. For example, scMDC performed well in dimension reduction, clustering and classification, but was less competitive in batch effect removal. Conversely, scMoMaT showed strong performance in batch effect removal but in many cases was less competitive in other tasks (Supplementary Fig. 6g–i). Overall, for bimodal RNA + ADT datasets, totalVI, a method specifically designed for such type of data, performed well across most evaluation metrics (Fig. 5c). For datasets with other combinations of data modalities, Multigrate appears to be the best option given its ability to handle more complex data modality combinations and consistently high performance across various evaluation metrics (Fig. 5d–f).

Benchmark spatial registration methods

Spatial registration aligns spatial transcriptomics data from different slices into a common spatial framework. Spatial coordinates can be viewed as an additional data modality, and its integration conceptually aligns with other integration tasks. To compare method performance on spatial registration for spatial transcriptomics data integration, we benchmarked five methods (that is, PASTE center alignment[50], PASTE pairwise alignment[50], SPIRAL[51], GPSA[52] and PASTE2[53]) on 12 datasets generated from various patients/donors and spatial techniques (that is, Visium, MERFISH, Stereo-seq and Xenium). An example data D61 generated using Visium technology is shown in Extended Data Fig. 4a, where each column visualizes the four slices of the spatial transcriptomics profiles of the human cortex aligned by a method. Extended Data Fig. 4b quantifies the alignment results from different methods using three evaluation metrics and then ranks the methods on the basis of the overall performance across these evaluation metrics. Among the evaluated methods, PASTE with center and pairwise alignments and PASTE2 rely solely on linear transformation and, therefore, preserve the relative spatial relationship within each slice. As a result, their spatial coherence scores (SCS), which quantifies the spatial coherence of tissue layers across different slices, are identical. In contrast, SPIRAL and GPSA apply nonlinear transformation to the original slices, and this may explain their relatively low performance when assessed by SCS. Alignment results and their quantification from another representative dataset D64 generated by the MERFISH technology are shown in Supplementary Fig. 7a,b.

The overall rankings for each individual dataset are summarized in Supplementary Fig. 7c, while the grand overall ranking across all datasets is presented in Extended Data Fig. 4c. Performance variability across individual datasets are summarized in Supplementary Fig. 7d. These results indicate that PASTE with center alignment performed consistently well across all datasets. PASTE2 also showed strong and stable performance, ranking competitively in most cases. However, its added flexibility in partial alignment across more diverse spatial transcriptomics datasets may offer limited benefit compared with PASTE in the alignment of adjacent and serial tissue sections. By comparison, the performance of PASTE with pairwise alignment, GPSA, and SPIRAL was more variable and appeared to depend on the specific dataset (Supplementary Fig. 7c). It is important to note that the SCS metric yields identical scores for methods that apply linear transformations of each slice, and SPIRAL, which uses a nonlinear transformation that regenerates coordinates for query slices, was not favorably evaluated under SCS (Extended Data Fig. 4c). However, SPIRAL demonstrated competitive performance under other metrics such as label transfer adjusted Rand index (LTARI) and pairwise alignment accuracy (PAA). Together, these findings highlight the importance of evaluating spatial registration methods using a broad range of datasets and diverse metrics, given that method performance can depend heavily on both the nature of the data and the chosen assessment criteria.

Computation time and peak memory usage

As the number of cells in a dataset increases, computational speed becomes a critical factor in integration. To evaluate computational efficiency, we monitored the time and peak memory usage (MU) of each method on a selection of large datasets with different data modality combinations for each integration category (Supplementary Fig. 8). Among the methods used for vertical integration, we found that UINMF was one of the most efficient methods in terms of both computation time (Supplementary Fig. 8a) and peak MU (Supplementary Fig. 8b) across different data modality combinations and scales well with increasing cell numbers. By contrast, some methods did not scale well with increasing numbers of cells (for example, scMSI and MIRA), and others consumed a large amount of memory (for example, scMM, MOFA+ and UnitedNet). For diagonal integration, while most methods performed very fast with minimal memory consumption (Supplementary Fig. 8c,d), Seurat appeared to consume substantially more time and memory. For mosaic integration, MultiVI showed a substantial increase in computation time usage and scMoMaT showed a large increase in MU (Supplementary Fig. 8e,f). Besides these two methods, Multigrate also required relatively longer computational time (CT) and used more memory in most applications. By contrast, StabMap performed most efficiently in terms of computation time and MU. For cross integration, Concerto and scMDC exhibited the highest runtime while maintaining moderate peak MU (Supplementary Fig. 8g,h). MOFA+ also consumed substantial computation time and memory, and Multigrate encountered an out of memory error when applied to integrate dataset D56. Again, StabMap was fast and efficient in MU in cross integration applications. In spatial registration, most methods encountered out-of-memory issues due to high memory requirements. SPIRAL was the only method that scaled to datasets with a large number of cells, although it required a long runtime and a substantial amount of memory (Supplementary Fig. 8i,j). Given sufficient memory, PASTE (with center and pairwise alignment) and PASTE2 demonstrated computational efficiency.

Impact of clustering algorithms on method evaluation

Given that different clustering algorithms can lead to varying clustering results, we assessed whether the choice of clustering algorithm would influence the ranks of methods in clustering evaluation. Specifically, we compared the overall rankings of methods obtained using three different clustering algorithms of *k*-means, Leiden and Louvain, across all integration categories and tasks that included clustering evaluation. The correlations of method rankings among different clustering algorithms are visualized in Supplementary Fig. 9. Our analysis revealed minimal impact on method assessment, as indicated by the high correlations of ranks across different clustering algorithms. For the vertical, diagonal, mosaic and cross integration categories (Supplementary Fig. 9a–d), most correlations exceeded 0.9. For the feature selection and imputation tasks (Supplementary Fig. 9e,f), the correlations were nearly 1, indicating that different clustering algorithms led to essentially identical assessment results on method performance. These findings demonstrate that our benchmark analyses are robust to the choice of clustering algorithm.

Method robustness and consistency

To assess the robustness and consistency of the methods, we first performed a stability analysis by randomly excluding approximately 20% of the cell types from the data (repeated ten times) and examining how each method's ranking changed relative to the others. This allows us to provide a comparative interpretation of method robustness and consistency that mitigates the lack of interpretation for individual evaluation metrics. Overall, most methods demonstrated robust performance, with relatively small error bars across iterations, indicating consistent rankings (Supplementary Fig. 10). However, a few methods showed notable variability. For instance, in vertical integration of RNA + ADT data modality, Matilda and UINMF exhibited larger error bars compared with other methods, suggesting greater sensitivity to data perturbations (Supplementary Fig. 10a). In feature selection

https://doi.org/10.1038/s41592-025-02856-3

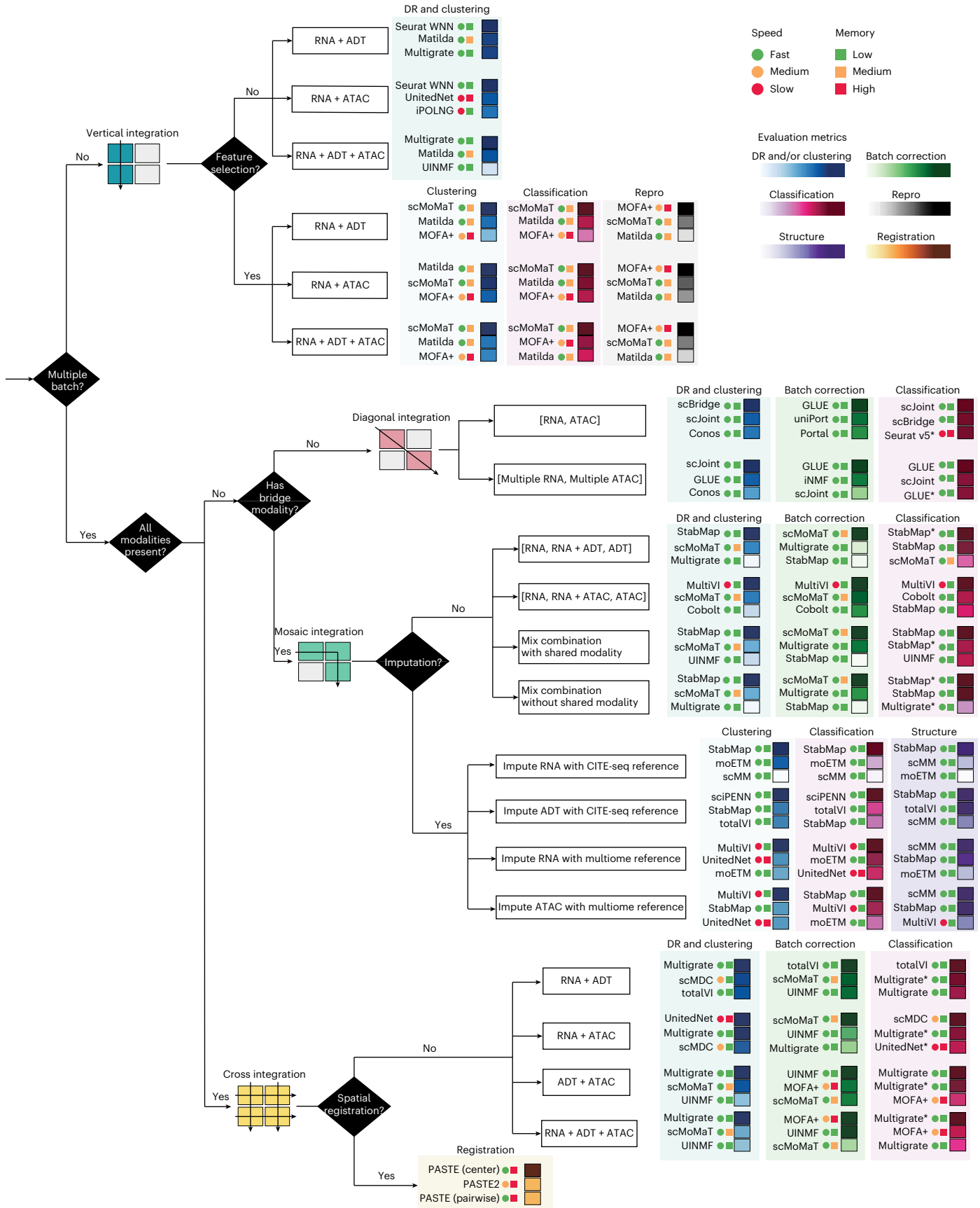


Fig. 6 | Guidelines for analyzing single-cell multimodal omics data. Recommendations of methods were based on data characteristics and the integration categories. The top-3 methods in each specific application and under each category of evaluation metrics were included in the recommendation. The speed and MU are color coded to provide a general guide on method scalability, while Repro represents reproducibility.

tasks, MOFA+ consistently outperformed other methods with the highest reproducibility, highlighting its robustness across perturbations (Supplementary Fig. 10b). In diagonal integration, scBridge showed stable performance across dimension reduction, clustering and classification metrics but displayed a larger error bar for batch correction, indicating potential variability in aligning batches when subsets of cell types were excluded (Supplementary Fig. 10c). We also assessed the methods that require random seeds. This analysis complements the evaluation using data perturbation, providing further insights into the sensitivity of methods to stochasticity of algorithms in data integration (Supplementary Fig. 11). We found that variabilities introduced by different random seeds generally are lower than those from data perturbation. Methods such as Matilda that are sensitive to data perturbation show more comparable stability in rankings compared with other methods. These results provide insights into the method robustness and strengthen the comparative analysis of method performance.

Discussion

We benchmarked 40 methods for vertical, diagonal, mosaic and cross integration of single-cell multimodal omics data across seven tasks: dimension reduction, clustering, batch correction, classification, feature selection, imputation and spatial registration. Using a battery of evaluation metrics, we summarize the overall rank score of each method for each integration category and task. This comprehensive ranking provides a detailed assessment of the strengths and weaknesses of methods across distinct evaluation criteria, offering valuable insights into their performance in addressing specific challenges of integration tasks. Overall, the performance of methods is found to be dependent on evaluation metrics, tasks and datasets. These results underscore the importance of evaluating methods using diverse metrics and datasets with different complexity and modality combinations for specific tasks.

In particular, it is important to be aware of the limitation of evaluation metrics. For example, ASW_cellType, iASW, ASW_batch and principal component regression (PCR) cannot be applied to integration methods that generate graph-based outputs. Many evaluation metrics, such as cLISI, rely heavily on accurate cell type annotations that may not always be available. Furthermore, SCS, a spatial registration metric, produces identical assessment scores for methods that rely solely on linear transformations of the data, making it unable to distinguish the relative performance of those methods. These findings highlight the need for careful consideration, selection and interpretation of evaluation metrics and results to ensure a meaningful method assessment for single-cell multimodal omics data integration.

The benchmark results also revealed a trade-off between preserving biological signals, as reflected by cell type separation, and batch correction. Methods that excel in one aspect may compromise its performance in the other. These findings emphasize the importance of considering specific integration objectives when selecting methods, as the trade-off between different aspects of performance may vary depending on the dataset and analysis priorities. In addition, we found that, for most methods, the performance on cell type classification is primarily influenced by the quality of data integration rather than by the choice of classification model. Indeed, a simple MLP classifier was able to achieve comparable performance to specialized classification models in most cases, suggesting that robust data integration plays a more critical role in downstream classification outcomes.

For the imputation task, we found that methods excelling in clustering and classification metrics do not necessarily perform well in preserving structural features. Likewise, for the feature selection task, approaches that identify cell-type-specific markers often compromise reproducibility when compared with methods that select a single set of markers for all cell types. Consequently, it is essential to consider the specific application, such as prioritizing the consistency of feature selection results or the importance of cell-type-specific information, when choosing appropriate methods for these tasks.

Although simulated datasets offer a valuable advantage by providing ground truth that is typically unavailable in real datasets, they may not fully capture the complexity and multidimensional variability inherent in real datasets. Therefore, the performance of integration methods evaluated on the simulated data may not always accurately reflect their performance on real datasets. Indeed, we observed that several methods that performed well on simulated datasets exhibited reduced performance when applied to real datasets. This discrepancy underscores the importance of cautiously interpreting benchmarking results derived from simulations and highlights the necessity of validating method robustness and generalizability using diverse, experimentally derived datasets.

In general, deep learning methods are popular, accounting for 26 out of the 40 integration methods evaluated. Among these, deep learning approaches consistently achieved top performance, securing the highest rank across all data types in both diagonal and cross integration tasks. These methods typically involve multilayer neural networks, which are computationally intensive and require graphics processing unit acceleration to achieve computational efficiency. In addition, the performance of deep learning methods is often affected by initial values and random seed selection, although the impact of these stochastic factors is generally small.

As a general guide, Fig. 6 summarizes the recommendation of top-3 integration methods based on the integration categories and tasks the user aims to accomplish. In particular, the selection of the methods depends on key data characteristics such as whether multiple batches are present, whether all data modalities are present and whether a bridge modality is present to connect datasets. For each category, we recommend three methods based on their overall rank score performance across all datasets for each evaluation criterion, and complemented by considerations of speed and MU. To further facilitate an in-depth exploration of method performance on specific tasks, we have also developed a Shiny app (http://shiny.maths.usyd.edu.au/scMultiBench/), which enables users to dynamically and interactively visualize benchmark results. Together, this benchmark study provides a comprehensive overview of computational methods for single-cell multimodal omics data integration, offering valuable guidance to researchers and method developers for applying and developing new methods tailored to specific aspects of data integration. In future work, we aim to address the lack of appropriate and broadly applicable metrics for evaluating integration methods across different tasks.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41592-025-02856-3.

References

1. Baysoy, A., Bai, Z., Satija, R. & Fan, R. The technological landscape and applications of single-cell multi-omics. *Nat. Rev. Mol. Cell Biol.* **24**, 695–713 (2023).
2. Vandereyken, K., Sifrim, A., Thienpont, B. & Voet, T. Methods and applications for single-cell and spatial multi-omics. *Nat. Rev. Genet.* **24**, 494–515 (2023).
3. Gaulton, K. J., Preissl, S. & Ren, B. Interpreting non-coding disease-associated human variants using single-cell epigenomics. *Nat. Rev. Genet.* **24**, 516–534 (2023).
4. Stoeckius, M. et al. Simultaneous epitope and transcriptome measurement in single cells. *Nat. Methods* **14**, 865–868 (2017).
5. Ma, S. et al. Chromatin potential identified by shared single-cell profiling of RNA and chromatin. *Cell* **183**, 1103–1116 (2020).
6. Swanson, E. et al. Simultaneous trimodal single-cell measurement of transcripts, epitopes, and chromatin accessibility using TEA-seq. *eLife* **10**, e63632 (2021).

7. Argelaguet, R., Cuomo, A. S. E., Stegle, O. & Marioni, J. C. Computational principles and challenges in single-cell data integration. *Nat. Biotechnol.* **39**, 1202–1215 (2021).
8. Heumos, L. et al. Best practices for single-cell analysis across modalities. *Nat. Rev. Genet.* **24**, 550–572 (2023).
9. Miao, Z., Humphreys, B. D., McMahon, A. P. & Kim, J. Multi-omics integration in the age of million single-cell data. *Nat. Rev. Nephrol.* **17**, 710–724 (2021).
10. Cantini, L. et al. Benchmarking joint multi-omics dimensionality reduction approaches for the study of cancer. *Nat. Commun.* **12**, 124 (2021).
11. Leng, D. et al. A benchmark study of deep learning-based multi-omics data fusion methods for cancer. *Genome Biol.* **23**, 171 (2022).
12. Luecken, M. D. et al. Benchmarking atlas-level data integration in single-cell genomics. *Nat. Methods* **19**, 41–50 (2022).
13. Gayoso, A. et al. Joint probabilistic modeling of single-cell multi-omic data with totalVI. *Nat. Methods* **18**, 272–282 (2021).
14. Lakkis, J. et al. A multi-use deep learning method for CITE-seq and single-cell RNA-seq data integration with cell surface protein prediction and imputation. *Nat. Mach. Intell.* **4**, 940–952 (2022).
15. Yang, M. et al. Contrastive learning enables rapid mapping to multimodal single-cell atlas of multimillion scale. *Nat. Mach. Intell.* **4**, 696–709 (2022).
16. Zhang, C. et al. Contrastively generative self-expression model for single-cell and spatial multimodal data. *Brief. Bioinform.* **24**, bbad265 (2023).
17. Liu, C., Huang, H. & Yang, P. Multi-task learning from multimodal single-cell omics with Matilda. *Nucleic Acids Res.* **51**, e45 (2023).
18. Argelaguet, R. et al. MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data. *Genome Biol.* **21**, 111 (2020).
19. Lotfollahi, M., Litinetskaya, A. & Theis, F. J. Integration and querying of multimodal single-cell data with PoE-VAE. Preprint at *bioRxiv* https://doi.org/10.1101/2022.03.16.484643 (2022).
20. Kriebel, A. R. & Welch, J. D. UINMF performs mosaic integration of single-cell multi-omic datasets using nonnegative matrix factorization. *Nat. Commun.* **13**, 780 (2022).
21. Zhang, Z. et al. scMoMaT jointly performs single cell mosaic integration and multi-modal bio-marker detection. *Nat. Commun.* **14**, 384 (2023).
22. Hao, Y. et al. Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573–3587.e29 (2021).
23. Minoura, K., Abe, K., Nam, H., Nishikawa, H. & Shimamura, T. A mixture-of-experts deep generative model for integrated analysis of single-cell multiomics data. *Cell Rep. Methods* **1**, 100071 (2021).
24. Lin, X., Tian, T., Wei, Z. & Hakonarson, H. Clustering of single-cell multi-omics data with a multimodal deep learning method. *Nat. Commun.* **13**, 7705 (2022).
25. Zhou, M. et al. Single-cell multi-omics topic embedding reveals cell-type-specific and COVID-19 severity-related immune signatures. *Cell Rep. Methods* **3**, 100563 (2023).
26. Wang, Y. et al. A multi-view latent variable model reveals cellular heterogeneity in complex tissues for paired multimodal single-cell data. *Bioinformatics* **39**, btad005 (2023).
27. Zhang, W. & Lin, Z. iPoLNG-an unsupervised model for the integrative analysis of single-cell multiomics data. *Front. Genet.* **14**, 998504 (2023).
28. Lynch, A. W. et al. MIRA: joint regulatory modeling of multimodal expression and chromatin accessibility in single cells. *Nat. Methods* **19**, 1097–1108 (2022).
29. Tang, X. et al. Explainable multi-task learning for multi-modality biological data analysis. *Nat. Commun.* **14**, 2546 (2023).
30. Li, G. et al. A deep generative model for multi-view profiling of single-cell RNA-seq and ATAC-seq data. *Genome Biol.* **23**, 20 (2022).
31. Yang, P., Huang, H. & Liu, C. Feature selection revisited in the single-cell era. *Genome Biol.* **22**, 321 (2021).
32. Li, Y. et al. scBridge embraces cell heterogeneity in single-cell RNA-seq and ATAC-seq data integration. *Nat. Commun.* **14**, 6045 (2023).
33. Zhao, J. et al. Adversarial domain translation networks for integrating large-scale atlas-level single-cell datasets. *Nat. Comput. Sci.* **2**, 317–330 (2022).
34. Xiong, L. et al. Online single-cell data integration through projecting heterogeneous datasets into a common cell-embedding space. *Nat. Commun.* **13**, 6118 (2022).
35. Hu, J., Chen, M. & Zhou, X. Effective and scalable single-cell data alignment with non-linear canonical correlation analysis. *Nucleic Acids Res.* **50**, e21 (2022).
36. Stuart, T. et al. Comprehensive integration of single-cell data. *Cell* **177**, 1888–1902 (2019).
37. Jain, M. S. et al. MultiMAP: dimensionality reduction and integration of multimodal data. *Genome Biol.* **22**, 346 (2021).
38. Hao, Y. et al. Dictionary learning for integrative, multimodal and scalable single-cell analysis. *Nat. Biotechnol.* **42**, 293–304 (2024).
39. Xu, Y., Begoli, E. & McCord, R. P. sciCAN: single-cell chromatin accessibility and gene expression data integration via cycle-consistent adversarial network. *npj Syst. Biol. Appl* **8**, 33 (2022).
40. Barkas, N. et al. Joint analysis of heterogeneous single-cell RNA-seq dataset collections. *Nat. Methods* **16**, 695–698 (2019).
41. Welch, J. D. et al. Single-cell multi-omic integration compares and contrasts features of brain cell identity. *Cell* **177**, 1873–1887 (2019).
42. Gao, C. et al. Iterative single-cell multi-omic integration using online learning. *Nat. Biotechnol.* **39**, 1000–1007 (2021).
43. Lin, Y. et al. scJoint integrates atlas-scale single-cell RNA-seq and ATAC-seq data with transfer learning. *Nat. Biotechnol.* **40**, 703–710 (2022).
44. Cao, Z.-J. & Gao, G. Multi-omics single-cell data integration and regulatory inference with graph-linked embedding. *Nat. Biotechnol.* **40**, 1458–1466 (2022).
45. Cao, K., Gong, Q., Hong, Y. & Wan, L. A unified computational framework for single-cell data integration with optimal transport. *Nat. Commun.* **13**, 7419 (2022).
46. Ashuach, T. et al. MultiVI: deep generative model for the integration of multimodal data. *Nat. Methods* **20**, 1222–1231 (2023).
47. Gong, B., Zhou, Y. & Purdom, E. Cobolt: integrative analysis of multimodal single-cell sequencing data. *Genome Biol.* **22**, 351 (2021).
48. Xu, Y., Das, P. & McCord, R. P. SMILE: mutual information learning for integration of single-cell omics data. *Bioinformatics* **38**, 476–486 (2022).
49. Ghazanfar, S., Guibentif, C. & Marioni, J. C. Stabilized mosaic single-cell data integration using unshared features. *Nat. Biotechnol.* **42**, 284–292 (2023).
50. Zeira, R., Land, M., Strzalkowski, A. & Raphael, B. J. Alignment and integration of spatial transcriptomics data. *Nat. Methods* **19**, 567–575 (2022).
51. Guo, T. et al. SPIRAL: integrating and aligning spatially resolved transcriptomics data across different experiments, conditions, and technologies. *Genome Biol.* **24**, 241 (2023).
52. Jones, A., Townes, F. W., Li, D. & Engelhardt, B. E. Alignment of spatial genomics data using deep Gaussian processes. *Nat. Methods* **20**, 1379–1387 (2023).
53. Liu, X., Zeira, R. & Raphael, B. J. Partial alignment of multislice spatially resolved transcriptomics data. *Genome Res.* **33**, 1124–1132 (2023).

Methods

Summary of integration methods in each category and task

The evaluation of integration methods from the four integration categories and seven tasks are summarized in Supplementary Fig. 12a–d. Please refer to Supplementary Notes for the criteria used for selecting integration methods. The details of the 64 real datasets and 22 simulated datasets that are used for the benchmarking are included in the 'Datasets and preprocessing' section, and the experimental settings for each of the 40 integration methods are described in the 'Single-cell multimodal omics data integration methods' section.

Datasets and preprocessing

Experimental datasets. We create a total of 64 datasets from 21 data sources (see 'Data availability' section in Supplementary Notes) generated using a variety of single-cell multimodal omics technologies. Details of each dataset grouped by integration categories are provided in Supplementary Notes.

All datasets are listed in Supplementary Table 1, and the datasets used for each method and task are summarized in Supplementary Fig. 12a–d. The same preprocessing procedure is applied to each dataset, including the removal of cell types with fewer than ten cells, the removal of genes expressed in less than 1% of cells in the RNA data modality, and the filtering of peaks quantified in less than 1% of cells in the ATAC data modality. All features are retained for the ADT data modalities. For multiple ATAC data batches with different genomic ranges, we followed https://stuartlab.org/signac/articles/merging using the 'reduce' function from the GenomicRanges package and the 'FeatureMatrix' function from the Signac package[54] to consolidate and quantify the intersecting peaks across all data batches to ensure a cohesive analysis of the data.

Data cleaning and reannotation. The reliance on predefined cell types can introduce bias toward methods that resemble those used to produce the original annotations. To minimize the potential bias in the original cell type annotation, we conduct data cleaning and reannotation using AdaSampling[55] and Annotatability[56] techniques, followed by manual inspection for additional verification. In AdaSampling, a soft-label learning step is used to train a classification model for inferring the mislabeling probability of each cell. Cells with high probabilities of mislabelling are excluded in further iterations of model training. This iterative training process enhances the accuracy of the classification model by systematically improving the quality of the cell type labels used for training. For each cell, the final trained model is used to either confirm its original cell type annotation or identify potential mislabeling and possible correction for the cell. Similarly, Annotatability refines cell type annotation with a deep neural network by monitoring the confidence (quantified as the average probability of annotations) and the variability (quantified as the standard deviation for the probability of annotations) over training epochs. Cells are then categorized as correctly annotated, erroneously annotated or ambiguous based on their confidence and variability. Possible corrections for cells in the erroneously annotated category are given by a deep neural network trained solely on cells from the correctly annotated category.

The data cleaning and reannotation procedure is summarized in Supplementary Fig. 12e. Specifically, we reannotate cells to their newly nominated cell types if both AdaSampling and Annotatability identified them as erroneously annotated and nominated them to the same new cell type. Cells that are identified as erroneously annotated by both methods, but the newly annotated cell type disagree for the two methods are removed. Finally, cells are considered correctly annotated if both methods classify them as their original cell type. The final dataset after the cleaning and reannotation procedure is manually validated by inspecting the expression of differentially expressed genes from each cell type and the annotated cells to verify if they are at a comparable level.

Simulation datasets. While the application of the above reannotation procedures may help to reduce the potential cell annotation bias from the predefined cell types of each data source, simulation datasets can also be used as an alternative for method evaluation. We use a generalized version of SymSim[57] to simulate single-cell multimodal omics for method evaluation. Specifically, we simulate multiome data with RNA and ATAC modalities by following the tutorial (https://github.com/PeterZZQ/Symsim2) provided by the scDART[58] using the functions 'SimulateTrueCounts', 'True2ObservedCounts' and 'DivideBatches' in the package. For the simulation of ADT data modality, we follow the instruction outlined in scMoMaT[21]. This involves choosing a reference ADT count dataset to establish a count distribution, from which we then simulate ADT counts that exhibit a positive correlation with gene expression data. For integration methods that require gene activity scores as input, we simulate the gene activity score using the same procedure as in simulating the ADT modality. After simulating all paired modalities, we create various modality combinations according to the integration categories. We create a total of 22 simulated datasets. Details of each dataset grouped by integration categories are provided in Supplementary Notes.

Single-cell multimodal omics data integration methods

In this section, we outline the single-cell multimodal omics data integration techniques included in this benchmark study. Count data are typically utilized as the input, unless otherwise indicated. For diagonal integration, we select methods specifically designed for aligning RNA and ATAC data modalities. Some methods require using original ATAC peaks directly as input, while others require peak data converted into gene activity scores. To facilitate this, we implement a cohesive pipeline, using the 'GeneActivity' function from the Signac package (v1.14.0), to transform the ATAC matrix from peak level quantifications to gene activity scores. Note that we use the 'CreateGeneActivityMatrix' function in the Seurat package to convert peak level quantifications into gene activity scores for the 'P0_BrainCortex' sample of SNARE-seq data as fragment files are not available. For diagonal integration, we use the gene activity scores from ATAC data as default unless indicated otherwise. For vertical, mosaic and cross integration, we use the peak data of ATAC as default unless indicated otherwise. The appropriate data combinations and formats for each method are included in Supplementary Table 2. Details for each method are provided in Supplementary Notes.

Evaluation metrics

Here, we summarize the metrics used for evaluating methods on single-cell multimodal omics data integration. The definition of each metric is provided in Supplementary Notes. For each metric, the advantages, disadvantages, requirements for cell type and batch labels, applicability to graph or embedding outputs, and other information are included in Supplementary Table 3.

Dimension reduction. Biological conservation was assessed via the graph cell-type local inverse Simpson's index (cLISI) score and dimension reduction scalability via CT and MU.

Batch correction. Batch correction quality was assessed via *k*-nearest neighbor batch effect test score; graph integration LISI score (iLISI); adjusted Rand index by batch (ARI_batch); normalized mutual information by batch (NMI_batch); average silhouette width by batch (ASW_batch); graph connectivity (GC); and PCR score.

Clustering. Clustering quality evaluation was assessed via adjusted Rand index by cell type (ARI_cellType); normalized mutual information by cell type (NMI_cellType); average silhouette width by cellType (ASW_cellType); isolated label score of F1 (iF1); and isolated label score ASW (iASW).

Classification. Classification performance was evaluated via the overall classification accuracy (OCA); average classification accuracy (ACA); specificity (Spec) and sensitivity (Sens); and F1-score (F1).

Feature selection. Feature specificity and reproducibility were assessed using marker overlap among different cell types (MO) and marker correlation in downsampled data (MC). Downstream analysis metrics, including clustering and classification, were applied to the top features.

Imputation. Imputation data structure metrics included standardized mean squared error (sMSE) compared with ground-truth data, preservation of feature correlation structure (pFCS) and preservation of differential expression statistics (pDES). Downstream analysis metrics for imputation involved clustering and classification applied to the imputed data.

Spatial registration. Spatial registration quality metrics included LTARI, PAA and SCS. Spatial registration scalability metrics included CT (CT_SR) and MU on spatial registration (MU_SR).

Evaluation pipelines

The following sections describe the seven task evaluation pipelines. The details of the datasets, integration methods, tasks and evaluation metrics involved are described in the corresponding sections above. The code for each pipeline is available via GitHub at https://github.com/PYangLab/scMultiBench/tree/main/evaluation_pipelines. A detailed description of the algorithms and pseudocode is included in Supplementary Notes.

Dimension reduction evaluation pipeline. The dimension reduction task evaluation pipeline is summarized in Supplementary Fig. 13a and by Algorithm 1 (Supplementary Notes). The input data can be either single-batch multimodal count data or multibatch multimodal count data, and the multimodal integration methods typically generate low-dimensional embeddings/graphs that capture the multimodal characteristics of individual cells. The quality of the embeddings/graphs generated by the integration methods can be evaluated using the cLISI score. We also assess the scalability of these integration methods by analyzing the CT and MU. Specifically, with our dimension reduction evaluation pipeline, we analyze 18 vertical integration methods on 23 real datasets and 6 simulated datasets, 14 diagonal integration methods on 14 real datasets and 4 simulated datasets, 7 mosaic integration methods on 13 real datasets and 4 simulated datasets and 11 cross integration methods on 9 real datasets and 8 simulated datasets (Supplementary Fig. 13a).

Batch correction evaluation pipeline. Batch correction is carried out using various multibatch integration methods, including 14 diagonal integration methods on 14 real datasets and 4 simulated datasets, 7 mosaic integration methods on 13 real datasets and 4 simulated datasets and 11 cross integration methods on 9 real datasets and 8 simulated datasets. These methods can generate batch-corrected low-dimensional embeddings/graphs, which is evaluated using the seven batch correction metrics. Note that PCR is calculated only for cross integration because it requires unintegrated data. Cross integration is the only category where there are unintegrated data for a specific modality across all batches. Supplementary Fig. 13b illustrates the evaluation pipeline for the batch correction task.

Clustering evaluation pipeline. The clustering evaluation pipeline is summarized in Supplementary Fig. 13c. In particular, clustering is performed on the low-dimensional integrated embeddings/graphs generated from the dimension reduction methods using Leiden clustering. All integration methods and datasets from the dimension reduction evaluation pipeline are assessed, including 18 vertical integration methods on 23 real datasets and 6 simulated datasets, 14 diagonal integration methods on 14 real datasets and 4 simulated datasets, 7 mosaic integration methods on 13 real datasets and 4 simulated datasets and 11 cross integration methods on 9 real datasets and 8 simulated datasets. The quality of the clustering results is measured using the five clustering evaluation metrics: ARI_cellType, NMI_cellType, ASW_cellType, iF1 and iASW.

Classification evaluation pipeline. Supplementary Fig. 13d summarizes the workflow of the classification evaluation pipeline. Similar to the batch correction pipeline above, classification is also performed and evaluated on the integrated low-dimensional embeddings generated from the dimension reduction methods, with the focus on the multibatch integration methods, including the 14 diagonal integration methods on 14 real datasets and 4 simulated datasets, the 7 mosaic integration methods on 13 real datasets and 4 simulated datasets and the 11 cross integration methods on 9 real datasets and 8 simulated datasets. All datasets included in this task contain multiple batches. For each dataset and method, each data batch is used as a reference while the remaining batches serve as queries to assess prediction results using a simple MLP classifier. If a method also provides its own classification model, we also include the classification results using its own model. Finally, we summarize the results from across all classification combinations for each dataset and method.

Feature selection evaluation pipeline. Summarized in the evaluation pipeline in Supplementary Fig. 13e, we focus on the three integration methods that enable feature selection: Matilda, scMoMaT and MOFA+. The quality of feature selection is evaluated on the 23 real datasets and 6 simulated datasets and using the aforementioned 5 classification metrics, 5 clustering metrics and the feature specificity and reproducibility metrics (that is, MO and MC). In particular, we train a simple MLP classifier and conduct fivefold cross-validation for evaluating the utility of selected features for classifying cell types and also apply the Leiden clustering pipeline as in the clustering task to evaluate their utility in cell clustering.

Imputation evaluation pipeline. Supplementary Fig. 13f illustrates the imputation evaluation pipeline. We focus on the seven mosaic integration methods applicable to imputation. This evaluation involves using reference and query data generated from datasets D51–57. In particular, we randomly select one batch as the reference data and then select another batch as the query, retaining only one data modality in the query dataset. This allows the missing data modality in the query data to be used as the ground truth for evaluating the quality of the imputed data. During the imputation, we utilize the reference data to impute the missing data modalities in the query data and then compare the imputed data with the ground-truth data. The quality of the imputed data is quantified using three imputation data structure metrics (that is sMSE, pFCS and pDES) as well as downstream analysis metric, including five classification metrics and five clustering metrics. Specifically, for the classification assessment, we first train a simple MLP classifier using the reference data subset by the features that are included in the imputed data (denoted by an asterisk in Supplementary Fig. 13f) and then classify the cells using the imputed data. This approach allows the assessment of imputed data using the five classification metrics.

Spatial registration evaluation pipeline. The spatial registration evaluation pipeline focuses on the four integration methods applicable to this task using the five spatial datasets with multiple slices, with each spatial slice consisting of the gene expression and spatial coordinates. Supplementary Fig. 13g illustrates the evaluation schematic which uses multiple spatial slices as the input. After conducting spatial registration

via the integration methods, these slices are aligned into a consistent coordinate system (that is, aligned coordinates). For evaluation, we use three established metrics from existing literature: LTARI, PAA and SCS. For LTARI, we perform Leiden clustering on all slices independently to obtain cluster labels for each slice. We next choose one slice as the reference, while the remaining as queries and transfer labels from the reference to the queries using 1-nearest neighbors based on the aligned coordinates. Finally, we apply the adjusted Rand index (ARI) metric to the transferred labels and ground truth to calculate LTARI. For PAA and SCS, these metrics are calculated based on the reference labels and the labels transferred to the query data. Besides, we also apply two spatial registration scalability metrics for the evaluation: CT_SR and MU_SR. As described in the 'Evaluation metrics' section in Supplementary Notes, we randomly sample from dataset 61 to create 11 subdatasets with different cell numbers ranging from 500 to 100,000. These subdatasets were used to evaluate runtime (CT_SR) and MU (MU_SR).

Other benchmark considerations

Robustness and consistency assessment. We randomly remove a subset of cell types (~20%) from datasets to assess the robustness of methods on data perturbation. The random removal of cell types is repeated ten times, and for each iteration, we evaluate the performance of the integration using metrics for their respective tasks. The change of performance on perturbed datasets is used to assess the robustness of each method on all integration categories and tasks. We also assess the impact of methods that rely on random initialization by randomizing initial values/seeds and repeating these methods (ten times) to evaluate their integration performance using metrics for their respective tasks. Finally, we assess the consistency of the methods by grouping and ranking their performance results by technologies and/or platforms. This allows us to quantify the change of their performance across different technologies and platforms.

Clustering algorithms and parameters. We use popular clustering algorithms of *k*-means, Leiden and Louvain for tasks that require clustering to evaluate method performance. Specifically, for *k*-means, we set *k*, the number of clusters, to the actual number of cell types in each dataset. For Leiden and Louvain clustering, we use the 'get_N_clusters' function from the Sinfonia package[59] to automatically select the resolution parameters that align the number of clusters to the number of cell types in each dataset. This allows us to quantify the performance of integration methods across different clustering algorithms.

R Shiny application for interactive exploration of benchmark results

We create an R Shiny application to host the benchmark results for interactive exploration, enabling the users to select specific integration categories, tasks, evaluation metrics and datasets of interest and dynamically display the relative performance of methods across the selections. The application will also enable the inclusion of new methods and datasets as they become available, ensuring up-to-date communication of benchmark results.

Guideline and recommendations

We provide a decision-tree-style guideline for each integration and task category as the final recommendation for method selection. Specifically, the decision-tree-style guideline includes scenario-specific recommendations for methods, taking into consideration data characteristics and computational efficiency. Each branch of the decision tree displays three recommended methods tailored to the specific integration category and task. These methods are accompanied by key performance metrics, benchmarked in this study and represented using color-coded boxes. The structured guidelines facilitate clear and comprehensive method recommendations for users.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

All real single-cell and spatial transcriptomics datasets used in this benchmark were obtained from publicly available repositories, as described in Supplementary Table 4. These include CITE-seq datasets from ArrayExpress (E-MTAB-10026) and GEO (GSE166489, GSE164378 and GSE194122), Zenodo (https://doi.org/10.5281/zenodo.6348128 (ref. 60)) and Software Heritage (https://archive.softwareheritage.org/browse/revision/1c7fcabb18a1971dc4d6e29bc3ed4f6f36b2361f/); 10x multiome datasets from GEO (GSE194122, GSE205117 and GSE204684) and Zenodo (https://doi.org/10.5281/zenodo.6348128 (ref. 60)); SHARE-seq and SNARE-seq datasets from GEO (GSE140203 and GSE126074); ASAP-seq, DOGMA-seq and TEA-seq datasets from GEO (GSE156478 and GSE158013); and spatial transcriptomics datasets including Visium (https://doi.org/10.5281/zenodo.6334774 (ref. 61) and https://github.com/raphael-group/paste_reproducibility/tree/main/data/DLPFC), Xenium (https://www.10xgenomics.com/products/xenium-in-situ/preview-dataset-human-breast), Stereo-seq (https://db.cngb.org/stomics/flysta3d/), MERFISH (https://cellxgene.cziscience.com/collections/31937775-0602-4e52-a799-b6acdd2bac2e) and Spatial ATAC-RNA (GEO: GSE205055). The processed input datasets used for benchmarking are publicly available via figshare at https://doi.org/10.6084/m9.figshare.29035586.v1 (ref. 62).

Code availability

The source code is available via GitHub at https://github.com/PYangLab/scMultiBench, including scripts for running the benchmarking methods, the evaluation pipeline and generating the figures associated with the paper. The code is also archived and available via Zenodo at https://doi.org/10.5281/zenodo.15385334 (ref. 63).

References

54. Stuart, T., Srivastava, A., Madad, S., Lareau, C. A. & Satija, R. Single-cell chromatin state analysis with Signac. *Nat. Methods* **18**, 1333–1341 (2021).
55. Yang, P. et al. AdaSampling for positive-unlabeled and label noise learning with bioinformatics applications. *IEEE Trans. Cyber.* **49**, 1932–1943 (2019).
56. Karin, J., Mintz, R., Raveh, B. & Nitzan, M. Interpreting single-cell and spatial omics data using deep neural network training dynamics. *Nat. Comput. Sci.* **4**, 941–954 (2024).
57. Zhang, X., Xu, C. & Yosef, N. Simulating multiple faceted variability in single cell RNA sequencing. *Nat. Commun.* **10**, 2611 (2019).
58. Zhang, Z., Yang, C. & Zhang, X. scDART: integrating unmatched scRNA-seq and scATAC-seq data and learning cross-modality relationship simultaneously. *Genome Biol.* **23**, 139 (2022).
59. Jiang, R., Li, Z., Jia, Y., Li, S. & Chen, S. SINFONIA: scalable identification of spatially variable genes for deciphering spatial domains. *Cells* **12**, 604 (2023).
60. Cheng, M., Li, Z. & Costa, I. G. MOJITOO: a fast and universal method for integration of multimodal single cell data [Data set]. *Zenodo* https://doi.org/10.5281/zenodo.6348128 (2022).
61. Zeira, R., Land, M., Strzalkowski, A. & Raphael, B. J. Alignment and integration of spatial transcriptomics data. *Zenodo* https://doi.org/10.5281/zenodo.6334774 (2021).
62. Liu, C. Datasets. *figshare* https://doi.org/10.6084/m9.figshare.29035586.v1 (2025).
63. Liu, C. Multi-task benchmarking of single-cell multimodal omics integration methods. *Zenodo* https://doi.org/10.5281/zenodo.15385334 (2025).

Acknowledgements
We thank all our colleagues, particularly at the School of Mathematics and Statistics, The University of Sydney and Sydney Precision Data Science Centre for their support and intellectual engagement. This work was supported by a National Health and Medical Research Council (NHMRC) Investigator Grant (1173469) to P.Y. S.G. was supported by an Australian Research Council DECRA Fellowship (DE220100964).

Author contributions
C.L., H.J.K. and P.Y. conceived the study. P.Y. supervised the study and C.L., S.D., S.L. and D.X. performed the initial feasibility analysis of the study. C.L. led the benchmark with major contributions from S.D. and P.Y., and additional contributions from H.J.K., S.L., D.X. and S.G. C.L., S.D. and P.Y. drafted the paper with input from all authors. All authors read, edited and approved this work.

Funding
Open access funding provided by the University of Sydney.

Competing interests
The authors declare no competing interests.

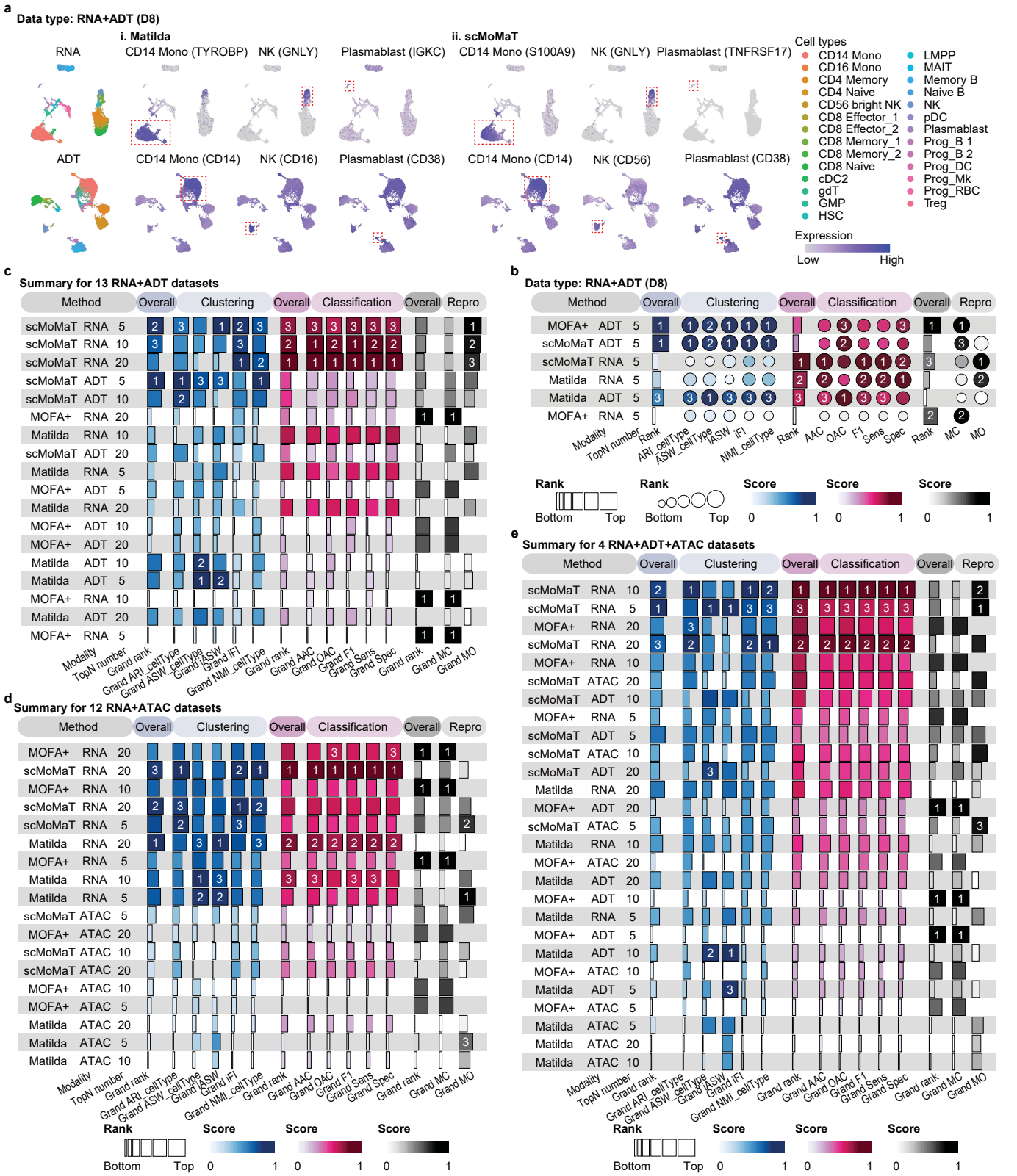
Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41592-025-02856-3.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41592-025-02856-3.

Correspondence and requests for materials should be addressed to Pengyi Yang.

Peer review information *Nature Methods* thanks Xiang Zhou and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available. Primary Handling Editor: Lin Tang, in collaboration with the *Nature Methods* team.

https://doi.org/10.1038/s41592-025-02856-3

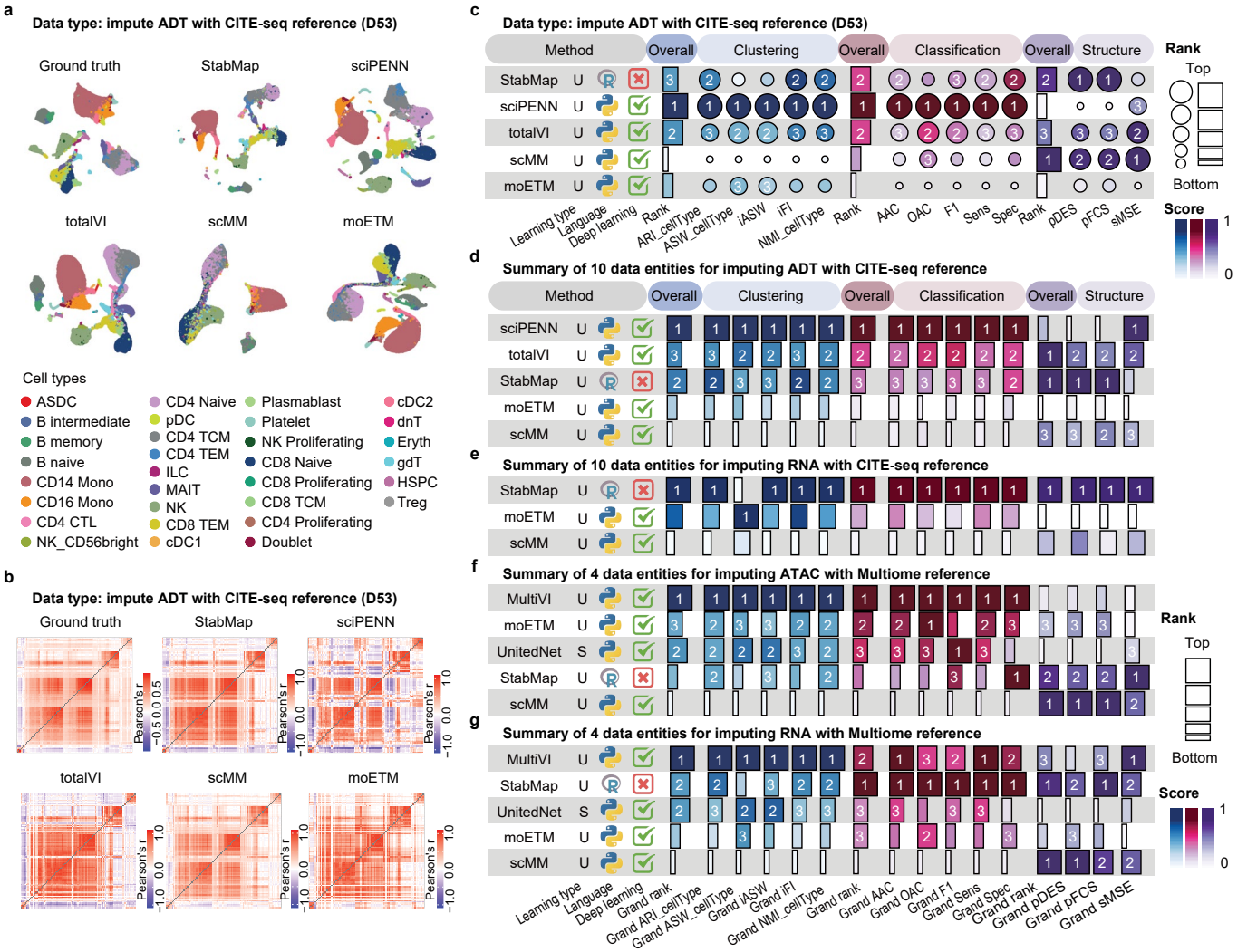


Extended Data Fig. 1 | Benchmark results of vertical integration methods for feature selection. a, UMAP visualisation of the top cell-type-specific markers selected from each data modality for three example cell types in a representative RNA + ADT dataset (D8). Note that MOFA+ is excluded from this visualisation since it selects a single set of markers for all cell types (that is cell-type-invariant). **b**, Method performance on the dataset D8 using top-5 markers. Metrics are categorised into clustering (blue tab), classification (pink tab), and reproducibility (grey tab). Performance summary of vertical integration for feature selection using top-5, 10, and 20 markers from **c**, All RNA + ADT datasets; **d**, All RNA + ATAC datasets; and **e**, All RNA + ADT + ATAC datasets. Repro represents reproducibility.

https://doi.org/10.1038/s41592-025-02856-3

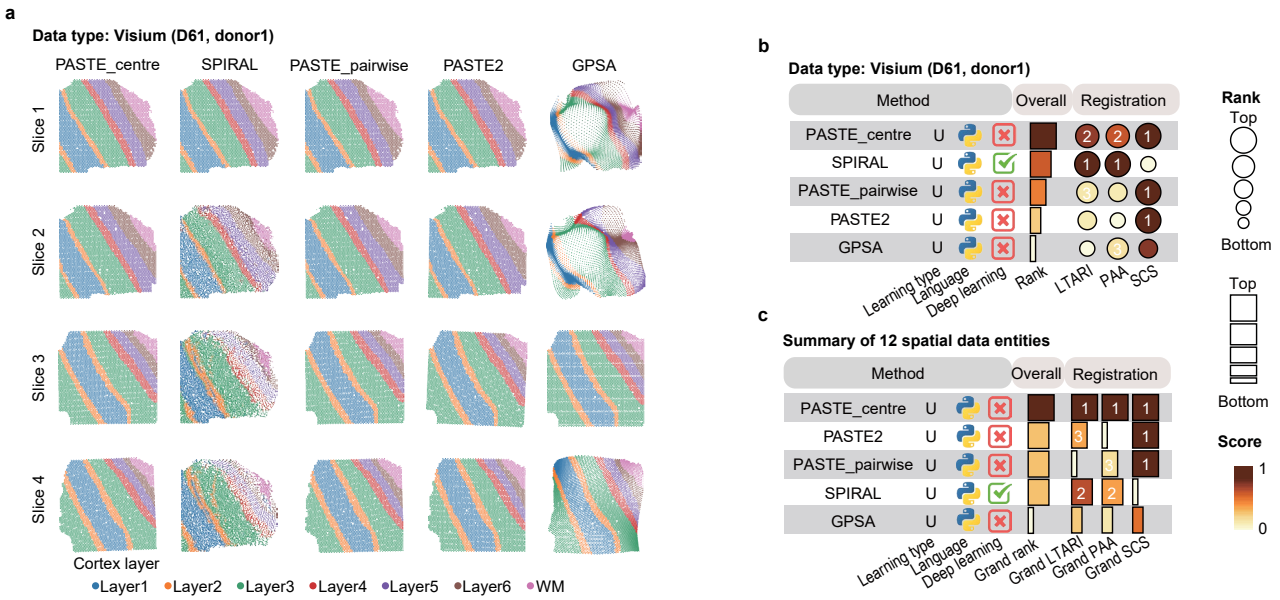
Extended Data Fig. 2 | Benchmark results of vertical integration methods for feature selection. a, UMAP visualisation of the top cell-type-specific markers selected from each data modality for three example cell types in a representative RNA + ATAC dataset (D15). **b**, UMAP visualisation of the top cell-type-specific markers selected from each data modality for three example cell types in a representative RNA + ADT + ATAC data (D22). **c**, Method performance on D15. **d**, Method performance on D22. Box plots showing the distribution of method ranks across **e**, All RNA + ADT datasets; and **f**, All RNA + ATAC datasets. In the box plots, the centre lines indicate the median, boxes indicate the interquartile range, and whiskers indicate 1.5× interquartile range. Each dot corresponds to the rank for an individual dataset.

https://doi.org/10.1038/s41592-025-02856-3



Extended Data Fig. 3 | Benchmark results of methods on data imputation for mosaic data analysis. a, UMAP visualisation of a representative dataset of D53 for ground truth and imputed ADT data modality for mosaic integration methods applicable to ADT data imputation. **b**, Pairwise correlation heatmap of top-100 highly variable ADTs in ground truth and imputed ADT data. **c**, Method performance on the dataset D53 for ADT imputation. Metrics are categorised into clustering (blue tab), classification (pink tab), and structure (eggplant tab). Performance summary of mosaic integration methods applied to **d**, All applicable data for imputing ADT data modality; **e**, All applicable data for imputing RNA data modality using an RNA + ADT reference; **f**, All applicable data for imputing ATAC data modality; and **g**, All applicable data for imputing RNA data modality using an RNA + ATAC reference.

https://doi.org/10.1038/s41592-025-02856-3



Extended Data Fig. 4 | Benchmark results of spatial registration methods. **a**, Visualisation of cortex spatial arrangements from dataset D61 (donor 1). Each row corresponds to the spatial transcriptomics profile of a tissue slice, and each column corresponds to the spatial registration results from a method across slices. **b**, Quantification of method performance on the dataset D61 (donor 1), comparing their alignment quality through multiple performance metrics. **c**, Performance summary of spatial registration methods across all examined datasets.

nature portfolio

Corresponding author(s): Pengyi Yang

Last updated by author(s): May 19, 2025

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	☐ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐ A description of all covariates tested
☒	☐ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No software was used during data collection.
Data analysis	Analysis of benchmark results was performed in R and Python using standard packages. We have uploaded the source code to a GitHub repository at https://github.com/PYangLab/scMultiBench, including scripts for running the benchmarking methods, the evaluation pipeline, and generating the figures associated with the paper. The code is also archived and available via Zenodo at https://doi.org/10.5281/zenodo.15385334. Key package versions: scvi-tools (version 1.1.2), sciPENN (version 1.0.0), Concerto (github version ab1fc7f), scMSI (github version dffcbb2), Matilda (github version 7d71480), MOFA+ (version 1.6.0), Multigrate (version 0.0.2), UNIMF (version 2.0.1), scMoMaT (version 0.2.2), scMM (github version c5c8579), scMDC (github version 43b0c3a), moETM (github version c2eaa97), VIMCCA (version 0.5.6), iPOLNG (version 0.0.2), MIRA (version 2.1.0), UnitedNet (github version 3689da8), scMVP (github version fc61e4d), scBridge (github version ff17561), Portal (version 1.0.2), SCALEX (version 1.0.2), VIPCCA (version 0.2.7), Seurat (version 5.0.2), MultiMAP (github version 681e608), SMILE (github version a2e2ca6), sciCAN (github version ad71bba), Conos (version 1.5.2), iNMF (version 2.0.1), online iNMF (version 2.0.1), scJoint (github version cbbfa5d), GLUE (version 0.3.2), uniPort (version 1.2.2), MultiVI (version 1.1.2), StabMap (version 0.1.8), Cobolt (version 1.0.1), PASTE (version 1.4.0), GPSA (version 0.8), SPIRAL (version 1.0), PASTE2 (github version b71ec88), scib (version 1.1.4), limma (version 3.44.3), SingleR (version 1.4.0), Signac (version 1.14.0), celldex (version 1.0.0).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

All real single-cell and spatial transcriptomics datasets used in this benchmark were obtained from publicly available repositories, as described in Supplementary Table 4. These include CITE-seq datasets from ArrayExpress (E-MTAB-10026) and GEO (GSE166489, GSE164378, GSE194122), Zenodo (https://zenodo.org/records/6348128), and Software Heritage (https://archive.softwareheritage.org/browse/revision/1c7fcabb18a1971dc4d6e29bc3ed4f6f36b2361f/); 10x multiome datasets from GEO (GSE194122, GSE205117, GSE204684) and Zenodo (https://zenodo.org/records/6348128); SHARE-seq and SNARE-seq datasets from GEO (GSE140203, GSE126074); ASAP-seq, DOGMA-seq, and TEA-seq datasets from GEO (GSE156478, GSE158013); and spatial transcriptomics datasets including Visium (https://zenodo.org/records/6334774, https://github.com/raphael-group/paste_reproducibility/tree/main/data/DLPFC), Xenium (https://www.10xgenomics.com/products/xenium-in-situ/preview-dataset-human-breast), Stereo-seq (https://db.cngb.org/stomics/flysta3d/), MERFISH (https://cellxgene.cziscience.com/collections/31937775-0602-4e52-a799-b6acdd2bac2e), and Spatial ATAC-RNA (GEO: GSE205055). The processed input datasets used for benchmarking are available in a publicly accessible Figshare repository (https://figshare.com/articles/dataset/datasets/29035586?file=54438737).

Human research participants

Policy information about studies involving human research participants and Sex and Gender in Research.

Reporting on sex and gender	n/a
Population characteristics	n/a
Recruitment	n/a
Ethics oversight	n/a

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No new data was generated as part of this study. The number of samples in each data source was provided by the original authors and we selected the samples whose data match the multi-modal data types covered in our study. The number of datasets in the study was selected to cover a representative sample of modality combinations and tissues.
Data exclusions	No data was excluded from the study.
Replication	Due to computational limitations, the main benchmark run was performed only once. However, to assess the influence of random seeds and the reproducibility of subsetting 80% of the cell types, we: (1) performed 10 rounds of random subsampling of the 80% cell type data and compared the rank summaries across methods for each task; and (2) conducted 10 rounds with different seeds to evaluate the impact of seed variation, where applicable.
Randomization	This is not relevant to our study because we do not include separate experimental groups.
Blinding	This is not relevant to our study because we do not include separate experimental groups.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging