Benchmarking MOFA+ against StabMap in mosaic data integration

Protocol for the “Working in Biosciences Module”

Leoni Zimmermann

Ruprecht-Karls-Universität Heidelberg

Velten Group

04.10.2024

Table of Contents

1. Introduction 3

2. Methods 4

2.1 MOFA+ 4

2.2 StabMap 5

2.3 PBMC data 6

2.4 Evaluation metrics 6

2.5 Optimal parameter selection 8

2.6 Mosaic data integration with varying bridge sizes 8

2.7 Mosaic data integration with an incomplete reference 9

2.8 Mouse gastrulation scRNA-seq data and multi-hop mosaic data integration simulation 10

3. Results 11

3.1 Determination of optimal model parameters 11

3.2 Reducing the number of shared features 14

3.3 Simulating incomplete reference data 16

3.4 Mosaic data integration across multiple datasets 18

4. Discussion 20

Supplementary 22

Acknowledgment 24

Code availability 24

References 25

1. Introduction

The advent of new technologies has enabled the measurement of multiple modalities, or "omics," from a single sample (Angermueller et al., 2016; Clark et al., 2018; Ma et al., 2020; Stoeckius et al., 2017). These multimodal datasets offer important insights into cellular processes but also present complex challenges, such as their integration to enable comparable analysis.

Data integration can be categorized broadly into the following types: horizontal, vertical, and mosaic integration (Argelaguet et al., 2021). Horizontal integration, also known as batch correction, involves combining datasets of the same modality sampled from independent groups of cells. On the other hand, vertical integration deals with integrating different modalities measured from the same sample. This can be done locally, focusing on specific features in a defined search space, or a globally. Global vertical integration methods aim to analyze broader cell states driven by gene regulatory networks using all measured features. These methods typically employ unsupervised dimensionality reduction.

Well-known tools are, among others, canonical correlation analysis (Stuart et al., 2019) and Weighted Nearest Neighbor Analysis (Hao et al., 2021), both integrated into Seurat and MOFA+ (Argelaguet et al., 2020). MOFA+ (Multi-Omics Factor Analysis) is a well-established and widely used method, which uses variational inference to infer latent factors that capture variance both between and within different data modalities. Additionally, MOFA+ is capable of imputing missing values in the data by utilizing the latent factors.

However, the scalability of multimodal assays remains a significant limitation (Argelaguet *et al.*, 2021). As a result, experimental designs often involve multiple sets of modalities measured across different samples, leading to what is known as “mosaic data”- a patchwork of data matrices with some features or whole modalities missing. Integrating mosaic data presents a unique challenge, as there is no single anchor connecting all datasets like in horizontal or vertical integration.

Several tools have been developed to address mosaic data integration, including UINMF (Kriebel and Welch, 2022), MultiMAP (Jain et al., 2021), Cobolt (Gong et al., 2021), MultiVI (Ashuach et al., 2023), and StabMap (Ghazanfar et al., 2024). Among these, StabMap offers a unique approach: Unlike the other mentioned mosaic data integration tools, StabMap does not rely on a subset of features common to all datasets or existing single-cell-modality datasets. It projects all cells onto reference coordinates derived from supervised or unsupervised learning, leveraging all available features.

The increasing prevalence of mosaic data in single-cell and trough extensions of currently available atlases with additional molecular layers introduces the need to systematically compare the performance of tools able to integrate mosaic data. This study aims to compare MOFA+, which can handle mosaic data integration through its ability to deal with missing data, to StabMap in their performance in integration of mosaic data and imputation of the missing values. In this work, I evaluate their performance across different simulated scenarios with varying degrees of feature overlap between datasets and a multi-hop integration task, where data must be integrated through several intermediate datasets. I show that both tools yield useful results and illustrate the circumstances under which MOFA+ is more effective than StabMap, and vice versa.

2. Methods

2.1 MOFA+

MOFA+ (Argelaguet *et al.*, 2020) is a statistical framework designed to learn a low-dimensional representation of multimodal single-cell data. The input data may consist of several datasets from different sets of cells (groups) with multiple measured modalities (views). MOFA assumes that the input data Y_{gm}, comprising g groups and m modalities, can be expressed as a linear combination of latent factors. These factors capture the variation across the modalities. The relationship can be expressed as follows:

$$Y_{gm} = Z_g W_m^T + \epsilon_{gm}$$

In this equation, Z represents the factor matrix for each group g, while W corresponds to the weight matrix for each modality m. The weights reflect the contribution of each feature to the variations in the data. The term ϵ_{gm} accounts for residual noise not covered by the factors. The model employs Bayesian variational inference (VI) to infer these equation components. VI approximates the true posterior distribution by introducing a variational distribution. The goal is to minimize the Kullback-Leibler divergence (KL) between the true posterior and the variational approximation. This equation can be rewritten to the evidence lower bound (ELBO) formula, transforming it into a maximization computation.

MOFA+ introduces a spike-and-slab prior that sets individual weights and factor values to zero, thereby enforcing sparse solutions and identifying the most relevant features for each factor. Another noteworthy feature of MOFA is its ability to address missing data in Y. Following the model training, the missing values can be imputed by multiplying the factor and weight matrices, $Z_g \times W_m^T$.

Version 1.12.1 of the MOFA2 R package was used for benchmarking. Input data did not include groups, and unless otherwise specified, it consisted of one view with missing features for a subset of cells and another with a complete set.

2.2 StabMap

StabMap (Ghazanfar *et al.*, 2024) is a bioinformatic tool designed to perform mosaic data integration leveraging the shared features between datasets. Input datasets must be connected via shared features, directly or through intermediate datasets, to form a Mosaic Data Topology (MDT). The MDT is an undirected network with each dataset as a node and the edges representing the shared features, weighted by the amount of overlap. The first step in generating the low dimensional embedding is to designate a reference dataset onto which query datasets will be projected. Dimension reduction of the reference data D_r is performed by Principal Component Analysis (PCA) if no cell type labels are available, estimating a score matrix S_r and a loadings matrix A_r so that

$$S_r = D_r^T \times A_r.$$

For every non-reference dataset, the shortest path to the reference in the MDT is calculated. In the case of a direct connection where not all features are common, the PC scores are estimated by a multivariable linear model trained on the shared features. For each row j containing the PC scores in S_r, the model is fit as follows:

$$S_r[j] = X_{<i,r>}[j]\beta_{<r,i>}[j] + \varepsilon$$

$X_{<i,r>}$ denotes the submatrix of the reference D_r features shared with the query dataset D_i. $\beta_{<r,i>}$ represents the fitted coefficients stored in the matrix $B_{<r,i>}$, while ε accounts for normally distributed noise. The estimated score matrix for dataset i can then be calculated by

$$S_i^r = X_{<r,i>}B_{<r,i>}$$

with $X_{<r,i>}$ being the submatrix of D_i features that are also common to the reference D_r. If there is no direct feature overlap between the reference and query datasets, this process will be repeated iteratively over intermediated datasets along the shortest weighted path in the MDT. Alternatively, the StabMap parameter “ProjectAll” can be activated to ensure that not only the query data but both the query and reference are projected together. After integration, missing features can be estimated by using the mean value of the nearest neighbors in the low-dimensional embedding. For this study, version 0.1.8 of the StabMap R package was used. The reference dataset was the complete dataset unless otherwise specified.

2.3 PBMC data

The dataset utilized for this project was obtained through the R package “SingleCellMultiModal” (Eckenrode et al., 2023). The original dataset consists of 10032 peripheral blood mononuclear cells (PBMCs) from a healthy human donor. Annotated with 14 cell types, the data includes 108344 accessible genomic regions and 36549 expressed genes.

The data underwent log10-normalization and variance filtering. The biological component of variation for each feature was estimated by fitting a trend to the variance against the mean for all features. Gene expression features with a mean count of less than 0.01 and accessible gene regions with a mean count of less than 0.25 were excluded. Hypothesis testing was performed under the assumption that the biological variance was ≤ 0. Features with a p-value below 0.05 were retained. After filtering, the final matrix comprised 1740 features (952 RNA and 788 ATAC) across 10032 cells.

The data was artificially separated for the benchmark into two halves with 5016 cells each.

2.4 Evaluation metrics

Three criteria were selected for evaluation: mean silhouette width and cell type accuracy for assessing the integration, and root-mean-squared error (RMSE) for measuring the quality of the imputation. The integration assessment was conducted based on UMAP coordinates derived from the model output, namely the MOFA factors and StabMap PCs.

The silhouette width measures the degree of similarity between a data point i (in this case, a cell) and other data points within the same cluster (cell type). Two components are calculated for every data point i of cluster C_I: Cohesion $a(i)$, the average distance to other points in cluster C_I, and separation $b(i)$, the mean distance to the data points of the nearest neighboring cluster C_J. These are defined by the following equations:

$$a(i) = \frac{1}{|C_i| - 1} \sum_{j \in C_I, i \neq j} d(i,j)$$

$$b(i) = min_{J \neq I} \frac{1}{|C_J|} \sum_{j \in C_J} d(i,j)$$

$d(i,j)$ denotes the distance between data points i and j. The silhouette width is then calculated as:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}.$$

The silhouette width can range from $-1 < s(i) < 1$, with larger values representing better clustering since the data point is closer to its cluster than others. For tool comparison, the mean silhouette width was computed for each cluster and then averaged to calculate the silhouette score representing the overall clustering quality.

Cell type accuracy was assessed by removing a subset of cell type labels from the UMAP embedding and predicting them using k-nearest neighbors ($k = 5$). The cell type accuracy A represents the proportion of correctly classified cell type labels and is calculated as:

$$A = \frac{\sum_i I\{C_i^{pred} = C_i^{true}\}}{\sum_i 1}$$

where $I\{C_i^{pred} = C_i^{true}\}$ equals 1, when the cell type label is predicted correctly, and 0, when otherwise.

To account for varying numbers of cells in each cluster, A was first calculated per cluster and then averaged across all clusters. Higher values of A indicate better classification accuracy and, therefore, better clustering of the cells.

The RMSE was calculated on the imputed values of the data's artificially removed features, measuring the prediction error:

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (\hat{x}_i - x_i)^2}$$

with $\hat{x}_i$ representing the predicted value, x_i the actual value, and n the number of data points. A lower RMSE indicates a smaller difference between predicted and true values, demonstrating a better imputation.

2.5 Optimal parameter selection

To determine the optimal parameters for the benchmark, the following model parameters' influence on the results were analyzed:

Parameter	**MOFA+**	**StabMap**
Model options	Number of factors	Number of reference and query PCs
	Activate spike-and-slab prior	
		Activate Project all
Data options	Scale views to unit variance	Scale data to a mean of zero Scale data to a standard deviation of one

For the number of MOFA+ factors and StabMap PCs, the following values were tested: 10, 15, 20, 30, 40, 50, 60, and 70. All other parameters were either enabled or disabled. For both models, all possible parameter combinations were tested on the PBMC data with all ATAC features removed for the first dataset, resulting in 32 combinations for MOFA+ and 512 for StabMap.

The optimal parameters were selected by analyzing various plots depicting the trend or parameter influence on the evaluation metrics. Data option and model option parameters except for the number of factors and PCs were set. The models were subsequently re-evaluated using the same range of values for the number of factors and PCs as previously tested.The best number of MOFA+ factors and StabMap PCs were chosen based on graphical trend representation for all three evaluation criteria.

2.6 Mosaic data integration with varying bridge sizes

The first object of the benchmark was to quantify the effect of the number of shared features, referred to as “bridge size”, on the model's performance. Features were randomly removed from the PBMC's first dataset as shown in the table below.

Bridge size	**Shared features [%]**
10	0.5%
52	3%
104	6%

156	9%
218	12.5%
438	25%
653	37.5%
870	50%
953	54.5%
1088	62.5%
1305	75%
1523	87.5%
1730	99.5%

Both MOFA+ and StabMap were trained using their respective optimal parameters. Each model was trained five times per bridge size, with a different random set of features removed in each iteration. After training, the evaluation metrics were calculated. Trends across the various bridge sizes were plotted and compared between MOFA+ and StabMap.

2.7 Mosaic data integration with an incomplete reference

For this simulation, the PBMC data with the removed ATAC features in the first dataset was used. At the same time, RNA features were randomly removed from the second dataset to simulate an incomplete reference dataset. Thus, the maximum bridge size was limited to 952. The specific bridge sizes applied are shown below:

Bridge size	**Common features [%]**
10	1%
56	6%
112	12%
167	17.5%
238	25%
476	50%
714	75%
942	99%

Both models were trained using the optimal parameters identified earlier, with MOFA+ modeling ATAC features in one view and RNA features in another, while StabMap set the incomplete reference as the reference dataset. Each model ran three times per bridge size, with a unique set of features removed for each run. The trends of the evaluation metrics were compared for all runs and both tools.

2.8 Mouse gastrulation scRNA-seq data and multi-hop mosaic data integration simulation

The count data of embryonic day 8.5 was accessed through the MouseGastrulationData Bioconductor package (Griffiths and Lun, 2024). The data contained 29452 RNA features for 20978 cells. Preprocessing followed the same steps as for the PBMC dataset, with features filtered for a mean count above 0.05 and a variance greater than 0. The resulting dataset contained 9572 highly variable genes.

The data was organized into eight datasets, each containing an equal number of randomly selected cells and features. Approximately 50% of the features are shared between neighboring datasets, meaning Dataset 2 shared half of its features with Dataset 1 and the other half with Dataset 3, but Dataset 1 and Dataset 3 do not have any features in common. These datasets varied in size, containing 500, 1000, or 2000 cells, and 100, 200, 500, or 1000 features. Each model completed five rounds of training, exploring all combinations of cell and feature numbers. Parameters were the optimal parameters determined but the data was scaled to a mean of 0 and standard deviation of 1, and the number of MOFA factors and StabMap PCs was set to 50. For StabMap, the first dataset was set as the reference. MOFA+ was set up to have eight views, each containing the features of one of the artificially created datasets. A UMAP embedding was calculated based on the MOFA+ factors. The cell type accuracy metric was calculated and compared across different settings to assess the performance of both tools in multi-hop mosaic data integration.

3. Results

In this study, I benchmarked MOFA+, designed for vertical data integration, with StabMap, created specifically for mosaic data integration, in their performance in both integration and imputation of mosaic data while keeping the conditions for both tools as comparable as possible. For performance evaluation, I selected three metrics: silhouette score and cell type accuracy for integration, and RMSE for imputation.

3.1 Determination of optimal model parameters

The first step of the benchmark involved finding the parameters yielding the most optimal results for both tools, thus enabling a comparative analysis.

I formed all possible combinations of model parameters for both tools. StabMap was trained with 512 combinations, while MOFA+ was trained with 32 different parameter combinations on the PBMC data missing ATAC features in one dataset (Fig. 1a). I plotted the trends of the three evaluation criteria for the spike-and-slab prior (MOFA+), project all (StabMap), and data scaling parameters. I calculated two averages, one for the results from training with the parameter activated, and another when it was turned off, allowing for a direct comparison between both conditions (Fig. S1).

For StabMap, the number of reference PCs changed the results the most (Fig. 1b). Conversely, the amount of query PCs did not affect the results. Scaling the data to a mean of 0 and the parameter “project all” improved integration but worsened the imputation results. Scaling the standard deviation to 1 resulted in the opposite effect. Based on these findings, I determined that enabling the "project all" parameter was optimal, as it made the integration process more similar to MOFA+. However, due to the conflicting effects of data scaling, I decided to disable all data scaling options for StabMap.

For MOFA+, the only parameter with a significant influence on results was the number of factors (Fig. 1c). MOFA+’s ability to introduce sparsity through a spike-and-slab prior showed no significant effect on integration and imputation. Given that StabMap has no comparable function, I decided to disable this feature. As with StabMap, I disabled the data scaling option due to the lack of significant impact.

Based on these initial results, I revised the parameter optimization process. I set all parameters except the number of MOFA+ factors and StabMap PCs to their newly determined optimal values. This approach allowed for a more focused analysis of the effects of these key parameters on the model

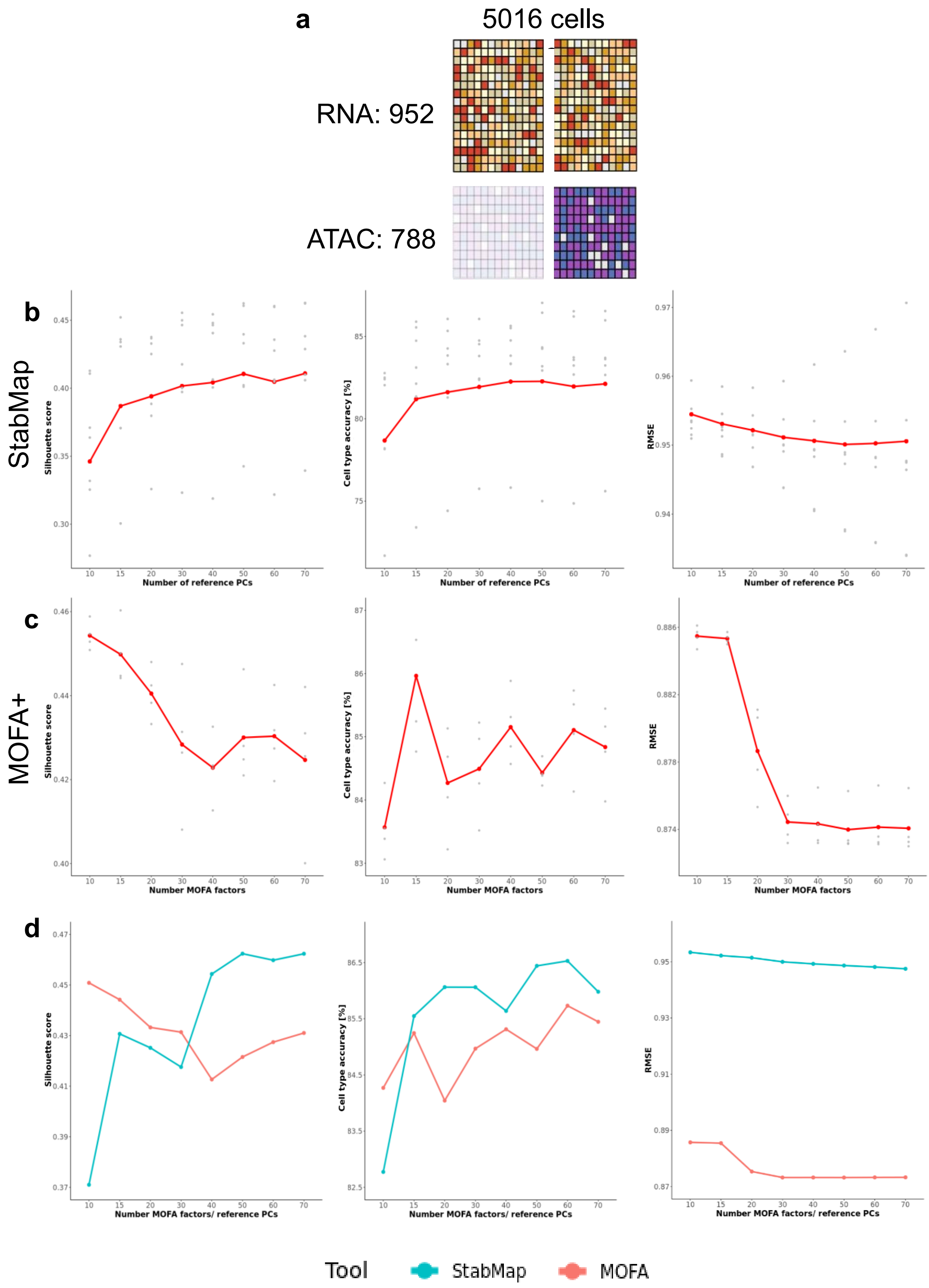

a
5016 cells
RNA: 952
ATAC: 788
b
StabMap
Silhouette score
0.30
0.35
0.40
0.45
Number of reference PCs
10
15
20
30
40
50
60
70
Cell type accuracy [%]
75
80
85
Number of reference PCs
RMSE
0.94
0.95
0.96
0.97
Number of reference PCs
c
MOFA+
Silhouette score
0.40
0.42
0.44
0.46
Number MOFA factors
Cell type accuracy [%]
83
84
85
86
87
Number MOFA factors
RMSE
0.874
0.878
0.882
0.886
Number MOFA factors
d
Silhouette score
0.37
0.39
0.41
0.43
0.45
0.47
Number MOFA factors/ reference PCs
Cell type accuracy [%]
82.5
83.5
84.5
85.5
86.5
Number MOFA factors/ reference PCs
RMSE
0.87
0.89
0.91
0.93
0.95
Number MOFA factors/ reference PCs
Tool
StabMap
MOFA

FIGURE 1

The evaluation metrics indicate that integrating with MOFA+ factors is optimal with fewer factors, while imputation accuracy and StabMap performance improve with more factors or PCs.

a The layout of the PBMCs data with all ATAC features missing for the query dataset (left), shown as grayed out. **b** Mean values of the three evaluation criteria (y-axis) from all tested parameter combinations against the number of reference PCs for StabMap **c** and the number of factors for MOFA+. The values for the individual model runs are represented by the grey dots. **d** Trends in the evaluation metrics for StabMap (blue) and MOFA+ (red) as the number of factors or PCs varied, while all other parameters were set to their identified optimal values.

performance, allowing me to isolate their effects and determine their optimal values without interference from less important factors.

I again visualized the trend of the evaluation metrics but combined the plots for StabMap and MOFA+ for easy comparison between the results for both tools (Fig. 1d). StabMap showed a clear trend of superior outcomes with a higher number of PCs. Thus, I determined that 70 was the optimal number of PCs. MOFA+'s imputation achieved a better RMSE with more factors. To additionally keep it comparable with StabMap, I set the number of factors for imputation to 70. The integration showed the opposite trend when observing the silhouette score. Although it starts to increase with more than 40 factors, the best results were obtained with 10 factors, followed by 15 factors. The trend of the cell type accuracy metric shows no clear pattern, but the third-best value is scored with 15 factors. Hence, I chose 15 as the number of factors for integration with MOFA+. Four of these 15 factors are associated with different cell types (Fig. S2). The final optimal model settings can be found in Table 1.

Number of MOFA+ factors	Integration: 15; Imputation: 70
Number of StabMap PCs	Integration and Imputation: 70, equally for reference and query
Spike-and-slab prior (MOFA+)	Disabled
Project all (StabMap)	Enabled
Data scaling options	Disabled

TABLE 1

Results for the optimal model parameter.

The parameter optimization revealed that StabMap outperforms MOFA+ in the integration. StabMap's silhouette score was 0.46, and the cell type accuracy was 85.9%, compared to 0.44 and 85.2%, respectively, for MOFA+. On the contrary, MOFA+'s imputation achieved more accurate results, with an RMSE value of 0.87, in contrast to the RMSE value of 0.95 observed for StabMap. These results indicate that, in this specific test case, while StabMap excels in integration, MOFA+ demonstrates superior performance in imputation.

3.2 Reducing the number of shared features

With the optimal model parameters established, I conducted a series of simulations to investigate the impact of reducing the number of shared features between the datasets in varying degrees of severity when integrating mosaic data.

I removed varying numbers of randomly selected features from one dataset of the PBMC data (Fig. 2a) and trained both models with optimal settings. I calculated the evaluation criteria and visualized their changes along the tested number of shared features, called bridge sizes.

As shown in Figures 2b and 2c, MOFA+ consistently outperformed StabMap across all bridge sizes for both silhouette score and cell type accuracy. Both tools showed a significant decline in performance as the number of shared features decreased, with silhouette scores turning negative and cell type accuracy dropping sharply at smaller bridge sizes. However, as more features were shared, the results recovered and stabilized. StabMap's performance plateaued beyond 438 shared features, while MOFA+ leveled off at different points depending on the evaluation criteria: around 870 features for the silhouette score and 218 for cell type accuracy. The best results for both models were achieved at the largest bridge sizes tested.

In the imputation task, MOFA+ demonstrated superior performance. Both tools showed similar trends of the RMSE decreasing as the number of shared features increased, with both models reaching their lowest error rates at larger bridge sizes. The most significant improvement in RMSE occurred when increasing the number of shared features from 10 to 52. However, the standard deviation of the RMSE also increased when more features were shared, particularly at the biggest bridge size.

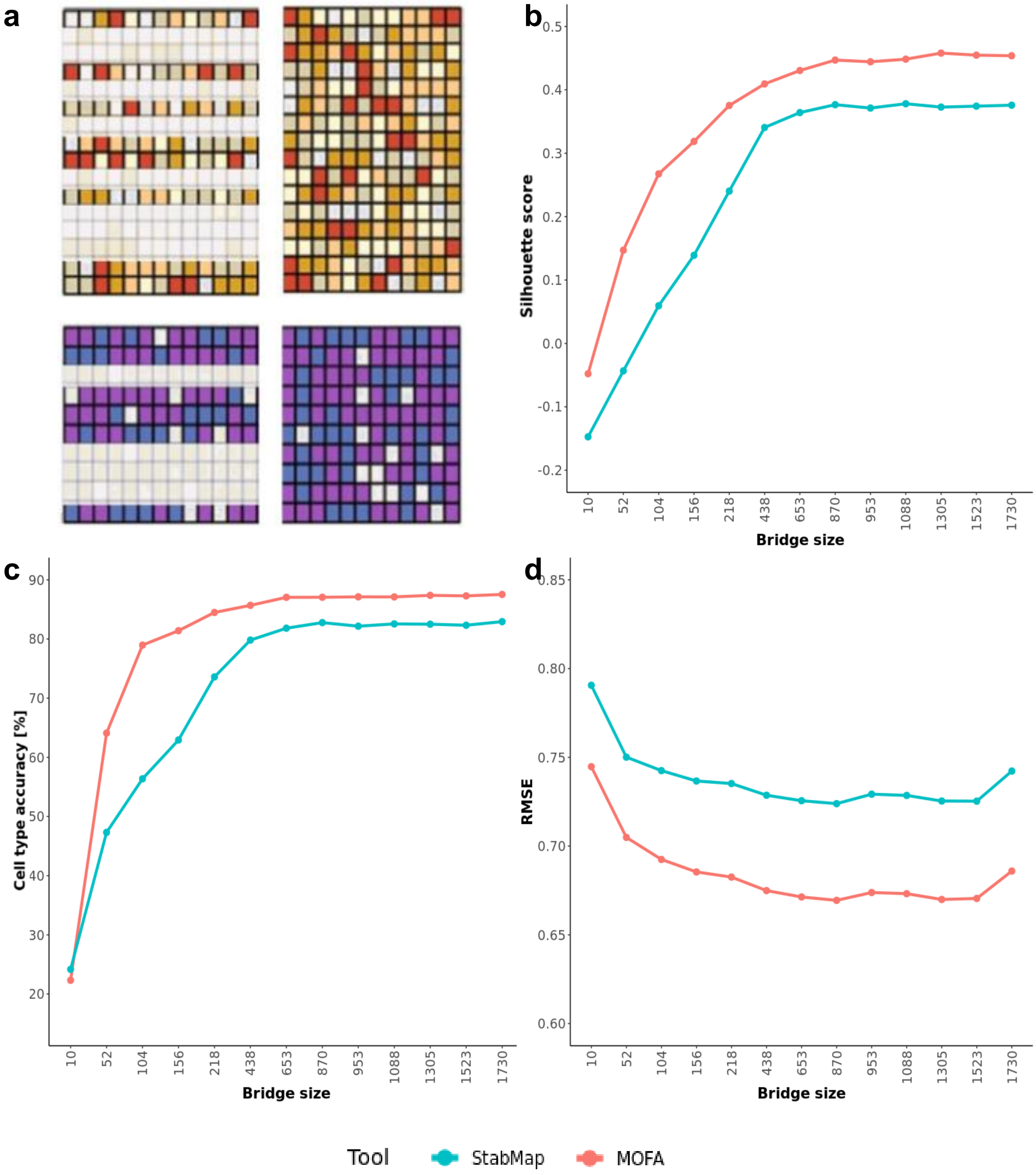


FIGURE 2

Reduction in the number of shared features significantly impairs integration and imputation, with MOFA+ achieving overall better results.

a An example of the PBMC data layout: RNA and ATAC features were randomly removed from the query dataset (left). Different numbers of shared features, called bridge sizes, were tested. **b** The results of the model training with varying bridge sizes (x-axis) from StabMap (blue) and MOFA+ (red) for the silhouette score, **c** cell type accuracy, **d** and RMSE.

Overall, MOFA+ consistently outperformed StabMap in both integration and imputation. While both tools struggled at lower bridge sizes, their performance stabilized with more shared features.

3.3 Simulating incomplete reference data

In the previous setup, one dataset contained all RNA and ATAC features, while the other had random missing features. To create a more realistic mosaic data integration scenario, I removed varying numbers of random RNA features from the second, previously complete, dataset, simulating an incomplete reference. (Fig. 3a).

Both models were trained using the optimal parameters, and evaluation metrics were calculated based on the results. As in the previous experiment, small bridge sizes significantly decreased the performance of both models. MOFA+ achieved better results overall, although StabMap generally came closer to the results of MOFA+, and in a few cases, surpassed them.

For the silhouette score (Fig 3b), both models performed poorly at the smallest bridge size, with MOFA+ outperforming StabMap. MOFA+'s silhouette score increased at a faster rate than that of StabMap, reaching a plateau at a bridge size of 112. StabMap's performance improved more gradually, eventually surpassing MOFA+ at a bridge size of 942.

A similar trend as in the silhouette score’s improvement was observed for cell type accuracy (Fig. 3c). While StabMap had a higher accuracy at the smallest bridge size, MOFA+ cell type accuracy improved more rapidly and maintained a higher accuracy at larger bridge sizes. However, StabMap's results were only slightly worse at these higher feature-sharing levels.

The RMSE was calculated separately for the ATAC and RNA features (Fig 3d, 3e). MOFA+ showed a higher RMSE value at the smallest bridge size but also a faster decline in RMSE than StabMap for ATAC features. Both tools reach a stable performance of approximately 112 shared features, with MOFA+ outperforming StabMap across all bridge sizes other than ten. For RNA features, MOFA+ shows less accurate imputed values at small bridge sizes but improves faster with an increasing number of shared features compared to StabMap. In comparison to the RMSE of the ATAC features, the RMSE for the RNA features does not reach a plateau but continues to decrease steadily until the largest bridge size is reached.

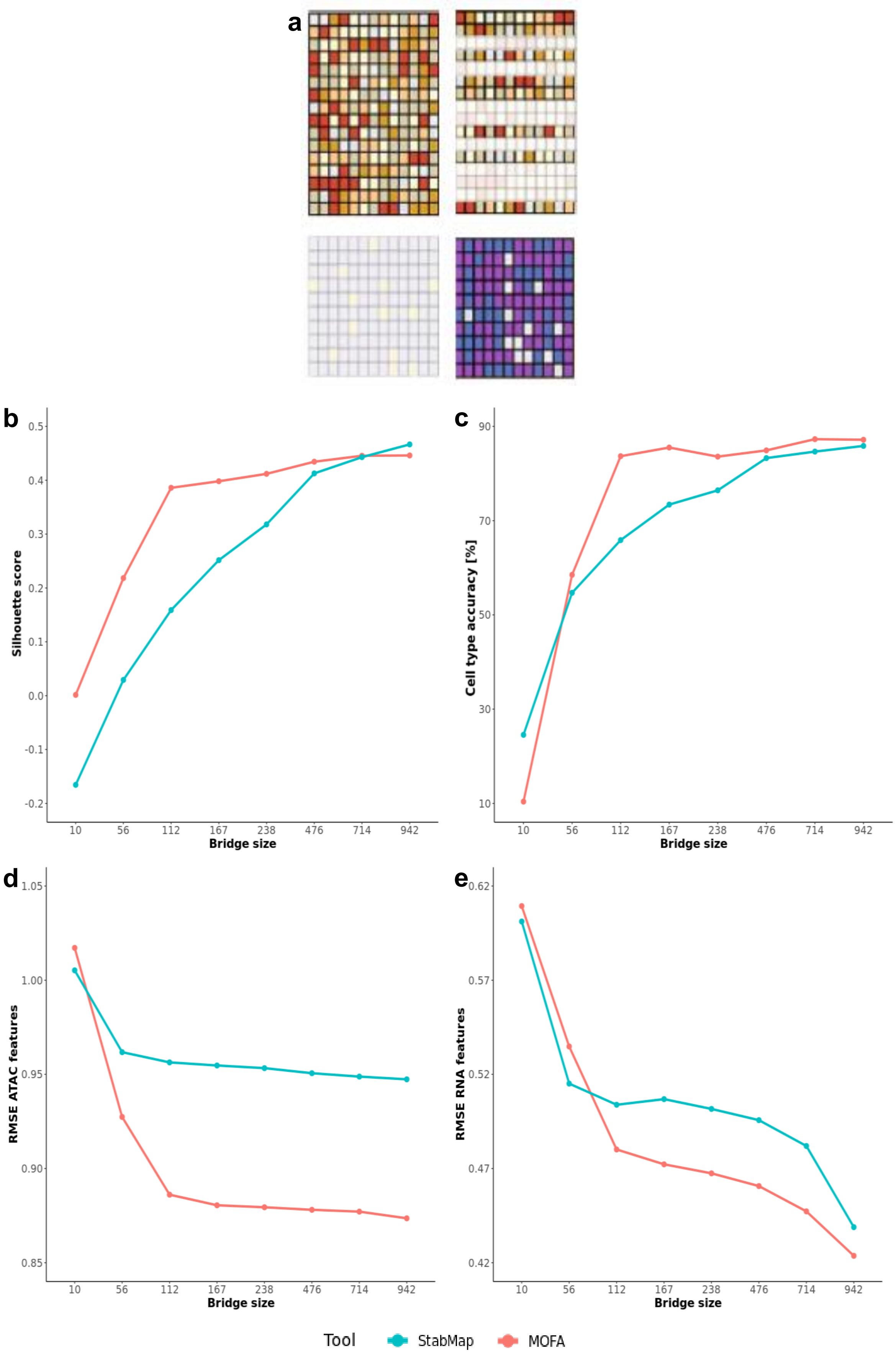

a
b
Silhouette score
0.5
0.4
0.3
0.2
0.1
0.0
-0.1
-0.2
10
56
112
167
238
476
714
942
Bridge size
c
Cell type accuracy [%]
90
70
50
30
10
d
RMSE ATAC features
1.05
1.00
0.95
0.90
0.85
e
RMSE RNA features
0.62
0.57
0.52
0.47
0.42
Tool
StabMap
MOFA

FIGURE 3

The removal of features from the reference results in a notable decline in performance, particularly with small bridge sizes. In a few extreme cases, StabMap outperforms MOFA+. However, in most cases, MOFA+ demonstrates superior performance over StabMap.

a Example of the layout of the PBMC data: All ATAC features were removed for the query dataset (left). Varying numbers of randomly selected RNA features were removed in the reference data set (right), resulting in different numbers of shared features (bridge sizes). **b** The results of the evaluation criteria silhouette score, **c** cell type accuracy, **d** and the RMSE calculated separately for the ATAC **e** and RNA features.

In conclusion, MOFA+ demonstrated superior performance, although StabMap performed better in some of the more extreme cases. Both models showed improved performance as the number of shared features increased, with their results stabilizing at larger bridge sizes.

3.4 Mosaic data integration across multiple datasets

Finally, I tested an extreme case of mosaic data integration using one experimental setup from the StabMap paper (Ghazanfar *et al.*, 2024). In this setup, scRNA-seq data from a mouse gastrulation atlas was artificially separated into eight datasets (Fig. 4a). The parameters for model training were adjusted to match those used in the original study. The datasets varied in the number of cells and features per dataset, and, after training, cell type accuracy was calculated to evaluate the integration quality.

For StabMap, a clear improvement in cell type label prediction was observed when there were fewer intermediate datasets between Dataset 1 and any other dataset, across all tested numbers of cells and features. Increasing the number of cells per dataset generally led to better results, particularly in runs with 500 and 1000 features per dataset. However, the number of features had a stronger influence on cell type accuracy than the number of cells.

On the other hand, MOFA+ produced a consistently low cell type accuracy between 5% and 10%, regardless of the number of intermediate datasets. The number of cells and features per dataset had little effect on the accuracy.

To summarize, StabMap demonstrated better adaptability in extreme mosaic data integration, with performance improving when fewer intermediate datasets were involved and with more features available. MOFA+, however, showed consistently low cell type accuracy, unaffected by changes in dataset composition.

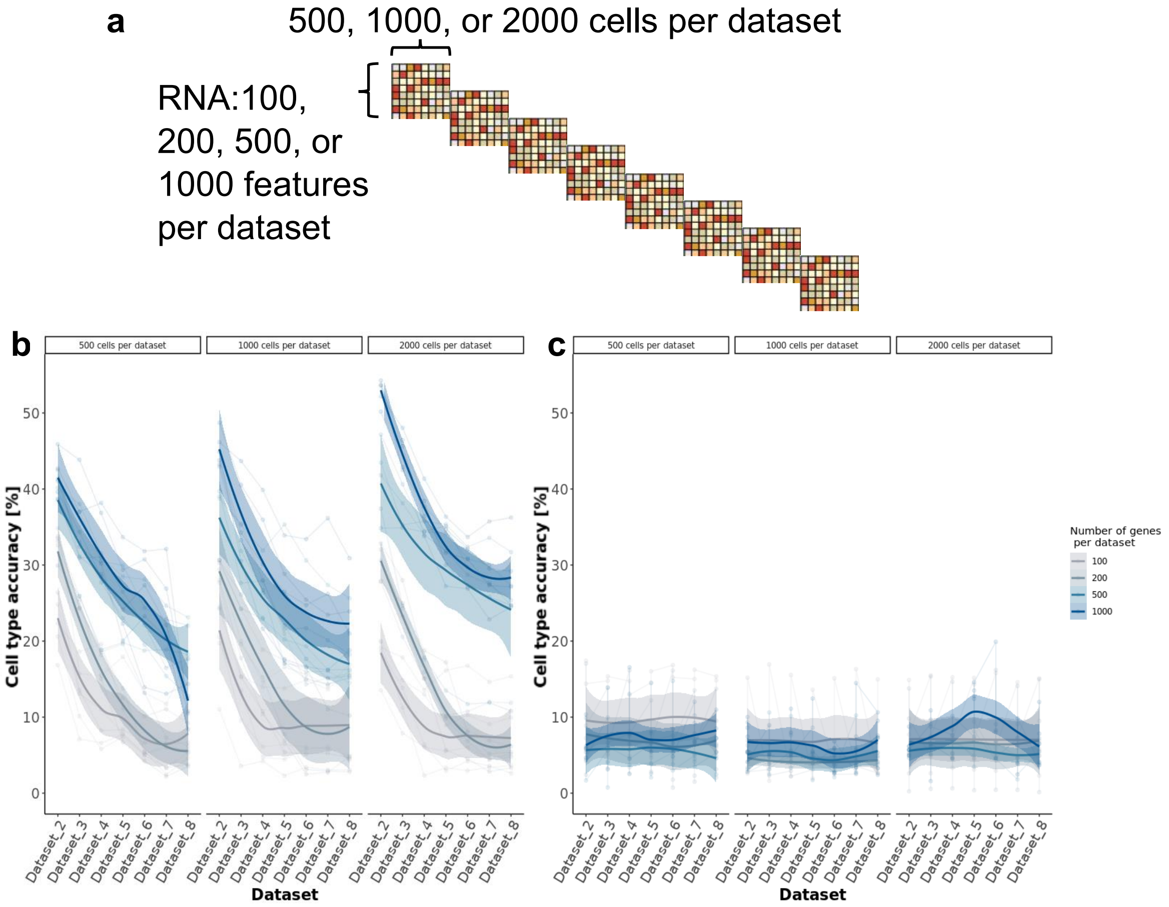


Figure 4

In extreme multi-hop mosaic integration scenarios, StabMap significantly outperforms MOFA+, showing better performance with fewer intermediate datasets, as well as with more features and cells per dataset. In contrast, MOFA+ maintains consistent but lower cell type accuracy across all datasets, regardless of the number of features or cells per dataset.

a Design of the mouse gastrulation RNA-seq dataset: Artificially separated into eight datasets, each exhibiting a 50% overlap in features with the preceding and subsequent datasets, respectively. The training was performed on combinations of varying numbers of features and cells per dataset. **b** Cell type accuracy for each dataset after training for StabMap **c** and MOFA+. The three columns separate the different numbers of cells used per dataset, while the line colors denote the number of genes per dataset. Each model run is represented by a dot, and the ribbons show the 95% confidence interval.

4. Discussion

In this benchmark study, I compared the performance of MOFA+ and StabMap in mosaic data integration, evaluating their quality in both integration and imputation across different simulated conditions. While both tools demonstrated strong capabilities, each demonstrated its strengths in different scenarios.

The parameter optimization process revealed important insights into the behavior of both tools. For MOFA+, the number of factors emerged as the most influential parameter. Interestingly, 15 factors were identified as optimal for integration, with four of these factors showing correlation to sets of different cell types. This finding suggests that the optimal number of factors may be related to the underlying biological structure of the data, potentially varying across different datasets. For StabMap the number of reference PCs had the most significant impact on results, while the number of query PCs showed little effect. This discrepancy may be attributed to the fact that the reference PCs are utilized for training the multivariable linear model, whereas the query PCs are merely projected onto it. The decision to disable the data scaling options for both tools was based on their minimal or conflicting effects on the results. This choice aimed to create a more comparable baseline between both tools, although in other situations, data scaling might still be beneficial, depending on the specific characteristics of the dataset.

Both tools exhibited a performance decline as the number of shared features decreased, observable across the simulations in sections 3.2 and 3.3. This can be attributed to an information bottleneck, with fewer shared features increasing the chance of losing relevant biological information crucial for accurate integration. Furthermore, the increase in data sparsity increases the susceptibility to noise. These factors make identifying patterns within the data more challenging and thus impede the ability to integrate the data effectively. The RMSE trend of both tools is very similar, possibly due to a strong correlation between the imputed values. However, MOFA+ showed better performance. The MOFA+ method has demonstrated superior imputation accuracy in the past compared to the k-nearest neighbor method that is used by StabMap (Argelaguet et al., 2018). Overall, MOFA+ outperformed StabMap consistently in these two simulations.

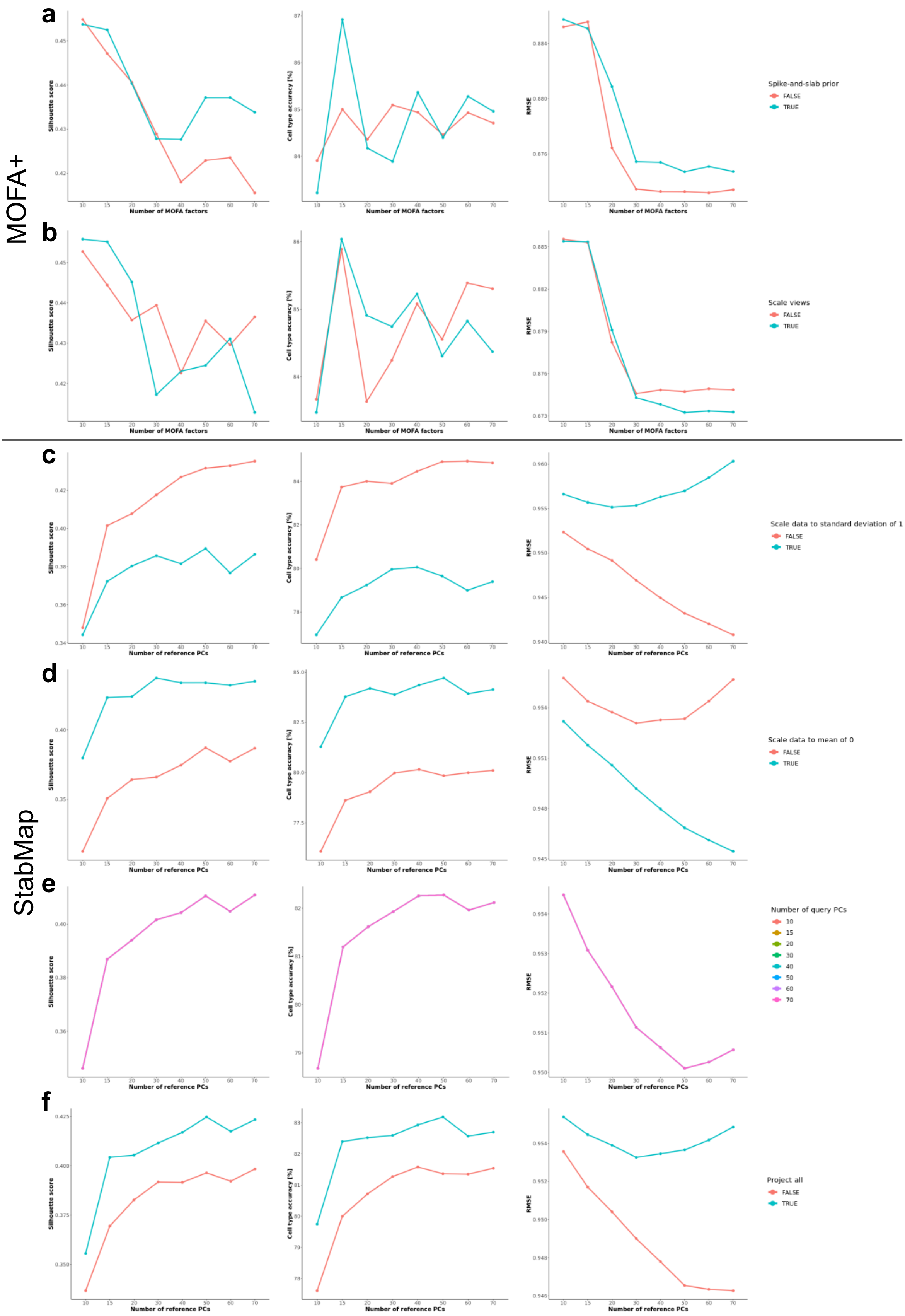
The multi-hop mosaic data integration experiment, mimicking a more complex scenario, revealed significant differences in tool performance. StabMap significantly outperformed MOFA+ in cell type accuracy, especially when fewer intermediate datasets and more features were involved. MOFA+ struggled to achieve accurate cell type label predictions but kept consistent results across all permutations. The reason for this might be how the tools weigh the features when integrating the data: StabMap puts greater weight on the shared features which are used to train the multivariable linear model needed for inferring the PCs of the unshared features. Furthermore, StabMap integrates along the shortest path in the MDT, thus Dataset 3 gets integrated with Dataset 2 before being integrated with the first dataset. This approach could prove advantageous when integrating data in this specific layout. In contrast, MOFA+ assigns equal weight to all existing features and lacks a step-by-step technique. This may contribute to the observed consistency in the results, irrespective of the number of intermediate datasets, cells, and features. It must be acknowledged that the simulated data layout is not representative of a realistic scenario. Consequently, it would be advisable to test the actual advantage of the StabMap method over MOFA+ in a more natural simulation.

It is important to note certain limitations of my study. The parameter optimization process also highlighted the complexity of choosing the right settings for these tools. The optimal parameters identified in this study may not be universally applicable, as they were determined based on a specific dataset and set of conditions. Furthermore, the benchmarking was primarily based on a single PBMC dataset, and while I introduced various artificial permutations to this data, the generalizability to other tissues or experimental contexts still has to be established. Comparing the final integration results from the parameter optimization with the results from section 3.2 for the bridge size of 953-matching the number of removed features used in the parameter search-reveals a discrepancy in the outcomes. This suggests that my approach is sensitive to the specific features that have been removed, namely whether all the features of a view or an omic are absent for a given dataset or not.

Future studies should therefore investigate the performance of MOFA+ and StabMap on other datatypes and real mosaic data that have not been separated artificially. Another direction to explore would be mosaic data integration with spatial data, mapping gene expressions from a comprehensive atlas on spatial coordinates.

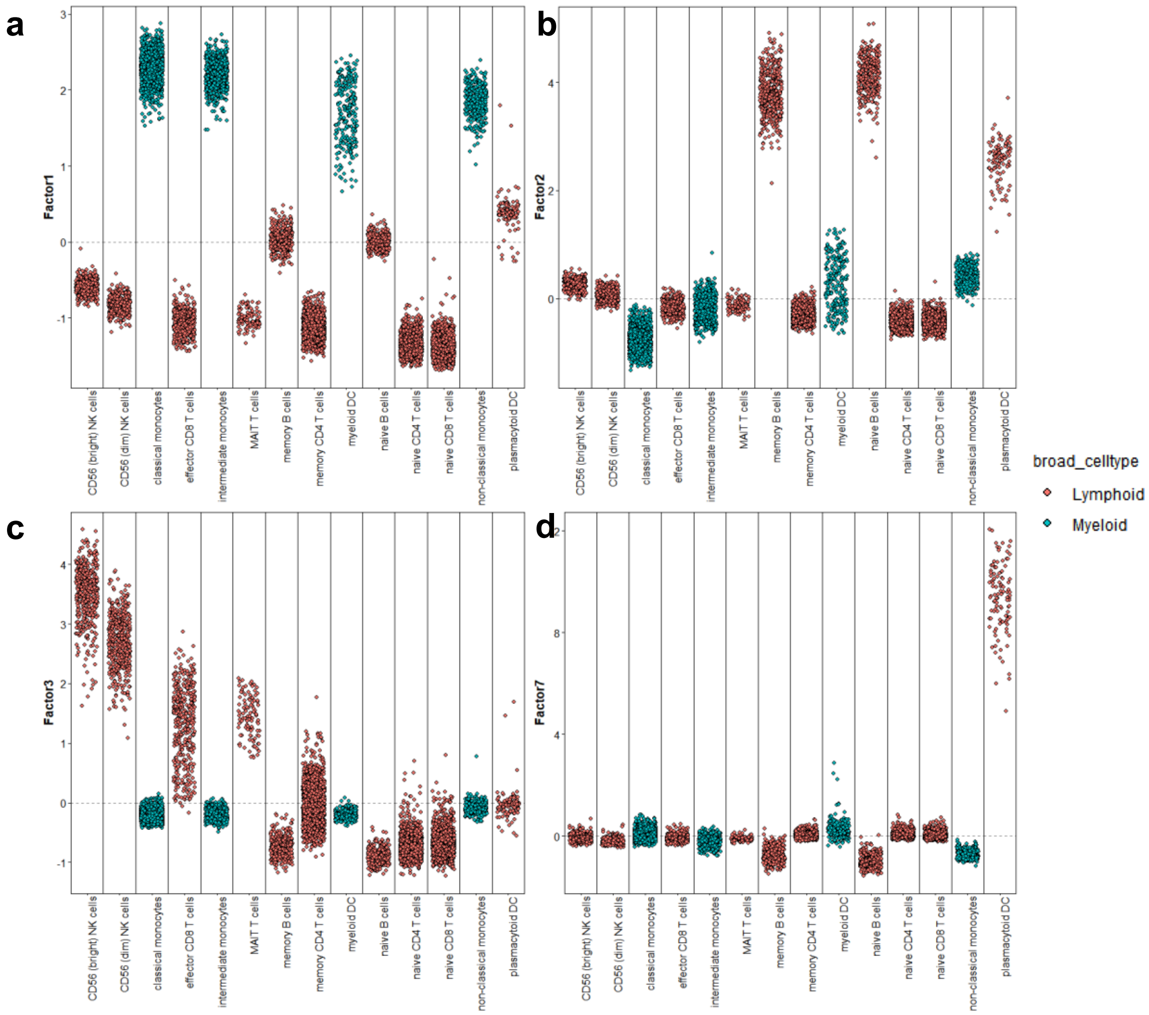
In conclusion, both MOFA+ and StabMap demonstrate effective integration of mosaic data integration. MOFA+ excelled in simpler dataset layouts where at least some features for every view were available, while StabMap showed superior integration capabilities in a complex multi-hop scenario. The selection between those tools should therefore be based on the characteristics of the dataset to be integrated.

Supplementary



SUPPLEMENTARY FIGURE 1

Effects of enabling or disabling model parameters. The lines represent the averages calculated from runs with the parameter activated compared to those with it turned off. Tested MOFA+ parameters were the **a** spike-and-slab prior and **b** view scaling. The analyzed options from StabMap included **c** scaling the data mean to 0 and **d** the standard deviation to 1, **e** number of query PCs, **f** and project all.



SUPPLEMENTARY FIGURE 2

If the MOFA+ model is trained with 15 Factors, four of them can be traced back to cell types.

a Factor 1 seperates lymphoid and myeloid cells, **b** factor 2 is correlated to B cells, **c** factor 3 to NK cells, and **d** factor 7 captures variation associated with the plasmacytoid DC.

Acknowledgment

I would like to express my gratitude to Britta Velten for giving me the opportunity to join her group for my lab rotation and for her assistance with my project. I also want to thank the entire group for being so welcoming and for the fruitful discussion on my work.

Furthermore, I acknowledge support by the state of Baden-Württemberg through bwHPC and the German Research Foundation (DFG) through grant INST 35/1597-1 FUGG.

Code availability

The analysis was conducted under R version 4.3.3 (2024-02-29). The code is available at https://github.com/velten-group/Benchmarking-MOFA-against-StabMap-in-mosaic-data-integration.

References

Angermueller, C., Clark, S.J., Lee, H.J., Macaulay, I.C., Teng, M.J., Hu, T.X., Krueger, F., Smallwood, S., Ponting, C.P., Voet, T., et al. (2016). Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity. Nat Methods *13*, 229-232. 10.1038/nmeth.3728.

Argelaguet, R., Arnol, D., Bredikhin, D., Deloro, Y., Velten, B., Marioni, J.C., and Stegle, O. (2020). MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data. Genome Biol *21*, 111. 10.1186/s13059-020-02015-1.

Argelaguet, R., Cuomo, A.S.E., Stegle, O., and Marioni, J.C. (2021). Computational principles and challenges in single-cell data integration. Nat Biotechnol *39*, 1202-1215. 10.1038/s41587-021-00895-7.

Argelaguet, R., Velten, B., Arnol, D., Dietrich, S., Zenz, T., Marioni, J.C., Buettner, F., Huber, W., and Stegle, O. (2018). Multi-Omics Factor Analysis-a framework for unsupervised integration of multi-omics data sets. Mol Syst Biol *14*, e8124. 10.15252/msb.20178124.

Ashuach, T., Gabitto, M.I., Koodli, R.V., Saldi, G.A., Jordan, M.I., and Yosef, N. (2023). MultiVI: deep generative model for the integration of multimodal data. Nat Methods *20*, 1222-1231. 10.1038/s41592-023-01909-9.

Clark, S.J., Argelaguet, R., Kapourani, C.A., Stubbs, T.M., Lee, H.J., Alda-Catalinas, C., Krueger, F., Sanguinetti, G., Kelsey, G., Marioni, J.C., et al. (2018). scNMT-seq enables joint profiling of chromatin accessibility DNA methylation and transcription in single cells. Nat Commun *9*, 781. 10.1038/s41467-018-03149-4.

Eckenrode, K.B., Righelli, D., Ramos, M., Argelaguet, R., Vanderaa, C., Geistlinger, L., Culhane, A.C., Gatto, L., Carey, V., Morgan, M., et al. (2023). Curated single cell multimodal landmark datasets for R/Bioconductor. PLoS Comput Biol *19*, e1011324. 10.1371/journal.pcbi.1011324.

Ghazanfar, S., Guibentif, C., and Marioni, J.C. (2024). Stabilized mosaic single-cell data integration using unshared features. Nat Biotechnol *42*, 284-292. 10.1038/s41587-023-01766-z.

Gong, B., Zhou, Y., and Purdom, E. (2021). Cobolt: integrative analysis of multimodal single-cell sequencing data. Genome Biol *22*, 351. 10.1186/s13059-021-02556-z.

Griffiths, J., and Lun, A. (2024). MouseGastrulationData: Single-Cell -omics Data across Mouse Gastrulation and Early Organogenesis.

Hao, Y., Hao, S., Andersen-Nissen, E., Mauck, W.M., 3rd, Zheng, S., Butler, A., Lee, M.J., Wilk, A.J., Darby, C., Zager, M., et al. (2021). Integrated analysis of multimodal single-cell data. Cell *184*, 3573-3587 e3529. 10.1016/j.cell.2021.04.048.

Jain, M.S., Polanski, K., Conde, C.D., Chen, X., Park, J., Mamanova, L., Knights, A., Botting, R.A., Stephenson, E., Haniffa, M., et al. (2021). MultiMAP: dimensionality reduction and integration of multimodal data. Genome Biol *22*, 346. 10.1186/s13059-021-02565-y.

Kriebel, A.R., and Welch, J.D. (2022). UINMF performs mosaic integration of single-cell multi-omic datasets using nonnegative matrix factorization. Nat Commun *13*, 780. 10.1038/s41467-022-28431-4.

Ma, S., Zhang, B., LaFave, L.M., Earl, A.S., Chiang, Z., Hu, Y., Ding, J., Brack, A., Kartha, V.K., Tay, T., et al. (2020). Chromatin Potential Identified by Shared Single-Cell Profiling of RNA and Chromatin. Cell *183*, 1103-1116 e1120. 10.1016/j.cell.2020.09.056.

Stoeckius, M., Hafemeister, C., Stephenson, W., Houck-Loomis, B., Chattopadhyay, P.K., Swerdlow, H., Satija, R., and Smibert, P. (2017). Simultaneous epitope and transcriptome measurement in single cells. Nat Methods *14*, 865-868. 10.1038/nmeth.4380.

Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W.M., 3rd, Hao, Y., Stoeckius, M., Smibert, P., and Satija, R. (2019). Comprehensive Integration of Single-Cell Data. Cell *177*, 1888-1902 e1821. 10.1016/j.cell.2019.05.031.