

Mosaic Cell Relation Factorization for Single-Cell Multi-omics Integration Across Simulated Populations and Dataset Sizes

Sandeep chowdary Vejandla

vsc4u@uab.edu

University of Alabama at Birmingham

Hemant Tiwari

University of Alabama at Birmingham

Article

Keywords: single-cell multi-omics, mosaic integration, cell relation factorization, scMultiSim, simulation benchmarking, clustering agreement

Posted Date: October 14th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11144436/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Mosaic Cell Relation Factorization for Single-Cell Multi-omics Integration Across Simulated Populations and Dataset Sizes

Sandeep C. Vejandla^{1*}, Hemant K. Tiwari¹

¹ Department of Biostatistics, School of Public Health, University of Alabama at Birmingham, Birmingham, AL 35294, USA

*Corresponding Author: Sandeep C. Vejandla

Email: vsc4u@uab.edu

Keywords single-cell multi-omics; mosaic integration; cell relation factorization; scMultiSim; simulation benchmarking; clustering agreement.

Abstract

Single-cell multi-omics integration must combine cells with paired and unpaired modalities. Here we introduce Mosaic Cell Relation Factorization (MCRF), which integrates RNA and chromatin accessibility by factorizing observed cell similarities. We compare MCRF with UINMF, COBOLT, and MIDAS using two collections of scMultiSim datasets: 144 smaller datasets from the original scMultiSim study and 102 larger datasets generated for this study. Both collections span three population structures and two levels of cell identity factor variability, while differing in cell and feature counts. Clustering agreement is evaluated using the adjusted Rand index (ARI) and normalized mutual information (NMI), with reference labels used to select clustering resolutions. MCRF shows favorable NMI performance in many configurations in the smaller collection and often lower variability across simulation replicates. In the larger collection, MCRF's absolute scores are generally higher than in the smaller collection, but UINMF and MIDAS achieve higher median ARI in all configurations with four replicates. These findings support cell relation factorization under several simulated conditions while identifying limitations on larger datasets. Evaluation on experimental data is needed to establish its value in biological applications.

Introduction

Multi-omics examines biological systems through several molecular layers, including the genome, transcriptome, epigenome, proteome, and metabolome. Each layer describes a different aspect of the same system. Studying these layers together can help connect molecular differences associated with disease and identify relationships that would be difficult to interpret from one data type alone. The resulting associations still require appropriate experimental evidence before they can be interpreted as causal relationships (1).

Integration methods differ in how they combine this information. Correlation approaches seek combinations of features that vary together across molecular layers. Matrix factorization represents measurements through a smaller set of factors, which can describe patterns shared across modalities and patterns specific to each modality.

Probabilistic models additionally describe how the observed measurements arise from underlying variables, while deep generative models use neural networks to learn more flexible relationships between these variables and the data. Another approach represents samples by their similarities to other samples, allowing different feature spaces to be compared through relations among the same observations (2). For example, similarity network fusion constructs a separate sample similarity network for each data type and combines these networks for joint analysis (3).

Single-cell multi-omics brings these questions to the level of individual cells. A molecular difference measured across an entire tissue may reflect altered activity within cells, a change in the proportions of different cell populations, or both. Measurements from individual cells help distinguish these possibilities. Joint analysis of transcriptomes, surface proteins, and lymphocyte receptors in COVID-19 illustrates this approach (4). For gene regulation, RNA sequencing and the assay for transposase-accessible chromatin (ATAC) sequencing provide complementary information about gene expression and chromatin accessibility. Technologies such as SHARE-seq measure both in the same cell and have revealed relationships between accessibility changes and subsequent gene expression during differentiation (5). Integration should therefore preserve the molecular differences that distinguish cell types and states, while allowing each modality to contribute additional information about their biology.

The available measurements are often incomplete across experiments. Some cells have both RNA-seq and ATAC-seq measurements, whereas others have only one modality. Combining these datasets creates a mosaic integration problem: the observed modalities overlap across batches, but no complete feature matrix is available for every cell. Paired cells can connect the modalities, while cells measured in only one modality contribute additional observations. The integration method must use these connections to place cells in a common representation. RNA-seq and ATAC-seq measure different aspects of a cell's biology, so differences between their measurements do not necessarily mean that the cells belong to different types or states. The method must account for these differences while preserving the biological information provided by each modality. Separately, samples processed in different experiments or batches may differ because of technical factors, such as sample preparation or sequencing depth. The method should reduce these technical effects so that biologically similar cells can be represented together, while retaining genuine differences between cell types and states.

UINMF (6), COBOLT (7), and MIDAS (8) address this problem using different representations. Unshared integrative nonnegative matrix factorization (UINMF) allows both shared and unshared features to contribute to cell factors (6). COBOLT uses a multimodal variational autoencoder to represent paired and unpaired RNA-seq and ATAC-seq measurements in a common latent space (7). MIDAS uses a probabilistic framework that separates biological and technical latent variables and aligns representations inferred from different observed modalities (8). Their different objectives make a common evaluation useful, but clustering performance also

depends on preprocessing and the procedures used to construct clusters. A benchmark must therefore define both the integration problem and the evaluation conditions.

Simulated data provide known population labels and allow selected conditions to be varied in a controlled way. scMultiSim (9) is particularly relevant because it generates RNA-seq and chromatin accessibility through related biological processes, incorporating cell identity, gene regulation, and technical variation. Its Cell Identity Factors (CIFs) represent underlying differences among cells, and the parameter σ_{CIF} controls variability in their construction. In a linear population structure, cell states vary gradually along a single continuous path without branches. In a branching structure, this path divides into different trajectories. A discrete structure represents distinct groups of cells without a continuous transition between them. These structures allow clustering to be evaluated across different forms of cellular organization. The original scMultiSim study used 144 datasets to benchmark mosaic integration, including UINMF and COBOLT. Its simulations thus provide an established starting point for examining how integration results change with population structure and simulation conditions.

Subsequent work has extended integration benchmarking in several directions. Fu *et al.* (10) included UINMF, COBOLT, and MIDAS within a broader comparison of single-cell multimodal integration algorithms. scMultiSim has also been used to evaluate integration on paired datasets containing 10,000 cells, including a study of binarized gene expression and the SCMBench evaluation of methods across different batch effect sizes (11, 12). Building on this work, we introduce Mosaic Cell Relation Factorization (MCRF), a new integration method, and compare it with UINMF, COBOLT, and MIDAS across population structures, cell and feature counts, CIF variability levels, and simulated replicates in a defined mosaic RNA–ATAC sequencing design.

This study makes two related contributions. First, we evaluate the four methods on the original 144 scMultiSim datasets and an additional collection of 102 larger datasets generated for this study, extending the evaluated cell counts from 500 and 800 to 2,000, 5,000, and 10,000 while also increasing feature counts. We assess agreement with the simulated population labels using the adjusted Rand index (ARI) and normalized mutual information (NMI) and distinguish performance level from variation across simulated replicates. Second, MCRF provides an approach to integration that represents cells through their observed similarities and factorizes those relations to obtain an integrated representation. Shared factors for paired cells connect RNA-seq and ATAC-seq, while separate modality matrices describe how those factors contribute to similarity in each modality. MCRF shows favorable NMI performance under many original simulation conditions, whereas UINMF and MIDAS achieve higher median ARI throughout the replicated larger configurations. These findings identify both strengths and limitations of cell relation factorization for mosaic integration.

Results

We compared MCRF, UINMF, COBOLT, and MIDAS across 144 original and 102 larger simulated RNA–ATAC datasets. Dataset dimensions and simulation settings are summarized in Table 1 and Fig. 1a; the mosaic

design combines paired RNA–ATAC cells with ATAC-only and RNA-only cells (Fig. 1b). Model and preprocessing settings are summarized in Table 2, with clustering and resolution selection described in Methods. Figures 2 and 3 show ARI for the original and larger collections, respectively; Figs. 4 and 5 show NMI. Each panel pools dataset sizes and simulation seeds. Configuration-specific medians and interquartile ranges (IQRs) instead summarize four simulation replicates with the remaining settings held fixed; the six datasets with 10,000 cells have one replicate each. Dataset-level scores are provided in Supplementary Tables 1–4, and configuration summaries in Supplementary Tables 5 and 6.

MCRF showed favorable NMI in the original collection

MCRF compared more favorably with UINMF and MIDAS in NMI than in ARI in the original collection (Figures 2 and 4). Comparing median scores across the four simulation replicates for each fixed configuration, MCRF had higher NMI than UINMF in 22 of 36 configurations and than MIDAS in 28 of 36. The corresponding counts for higher ARI were seven and 15. Its NMI performance was particularly favorable in continuous/branching populations, where its median exceeded each method in 10 of 12 configurations (Supplementary Tables 5 and 6, panel A).

Some configurations combined higher NMI with less variation across simulation seeds. For discrete populations with 500 cells, 200 genes, and $\sigma_{\text{CIF}} = 0.5$, MCRF had a median NMI of 0.501 and an IQR of 0.015. The corresponding medians and IQRs were 0.389 and 0.074 for UINMF, 0.368 and 0.042 for MIDAS, and 0.159 and 0.069 for COBOLT. Across the original collection, MCRF combined higher median NMI with a smaller IQR in 11 of 36 configurations relative to UINMF and 17 relative to MIDAS (Supplementary Table 6, panel A). The favorable NMI pattern also had some limits. MCRF had lower median ARI than UINMF in all 12 linear configurations. For discrete populations with 500 genes, its median ARI was lower than both UINMF and MIDAS in all four combinations of cell count and CIF variability. Its median NMI was also below UINMF in all four of these configurations (Supplementary Tables 5 and 6, panel A).

MCRF often showed less variation across simulated replicates

MCRF generally had a smaller ARI IQR than UINMF and MIDAS when the simulation configuration was held fixed. In the original collection, its ARI IQR was smaller in 27 of 36 configurations relative to UINMF and 28 relative to MIDAS. In the larger collection, the corresponding counts were 23 of 24 and 18 of 24 (Supplementary Table 5).

For NMI, MCRF had a smaller IQR than MIDAS in 25 of 36 original configurations and 18 of 24 larger configurations. Compared with UINMF, its NMI IQR was smaller in only half of the original configurations, but in 22 of 24 larger configurations (Supplementary Table 6). Lower dispersion was therefore a recurring property, although its frequency depended on the metric and comparator. These IQRs describe variation across four

separately simulated datasets per configuration; they do not measure variability from repeated initialization on the same dataset.

MCRF ranked lower in the larger collection

MCRF's relative performance was weaker in the larger collection (Figures 3 and 5). Its median ARI was below both UINMF and MIDAS in all 24 configurations with four simulation replicates. Its median NMI was higher than each method in only four configurations; one advantage over MIDAS was less than 0.001 (Supplementary Tables 5 and 6, panel B). MCRF also had lower ARI than both methods on all six datasets with 10,000 cells, for which only one simulation replicate was available.

The distinction between performance and dispersion was clear in discrete populations with 2,000 cells, 2,000 genes, and $\sigma_{\text{CIF}} = 0.1$. MCRF's median ARI was 0.444 with an IQR of 0.001, compared with 0.683 and 0.079 for UINMF and 0.667 and 0.058 for MIDAS. MCRF therefore produced closely grouped scores across replicates, but at a lower performance level. Its median NMI in the same configuration was 0.744, also below UINMF at 0.800 and MIDAS at 0.831 (Supplementary Tables 5 and 6, panel B).

Comparisons with COBOLT remained more favorable. Across the 60 configurations with four replicates in the two collections, MCRF had higher median NMI in every configuration and higher median ARI in 56. The four ARI exceptions occurred in continuous/branching populations (Supplementary Tables 5 and 6).

Performance across population structures and simulation conditions

MCRF generally achieved higher ARI and NMI in discrete populations than in linear or branching populations (Figures 2–5). At $\sigma_{\text{CIF}} = 0.1$, its median scores across the datasets shown in each panel were approximately 0.39 for ARI and 0.66 for NMI in the original discrete populations, compared with 0.18 and 0.35 in linear populations and 0.23 and 0.47 in branching populations. UINMF showed a similar pattern, with its highest scores generally occurring in discrete populations.

MCRF's absolute scores were generally higher in the larger collection, despite its weaker performance relative to UINMF and MIDAS. In discrete populations at $\sigma_{\text{CIF}} = 0.1$, its panel medians reached approximately 0.44 for ARI and 0.74 for NMI. Linear populations also showed higher scores, whereas branching populations showed a modest increase in NMI with little change in ARI. UINMF showed clearer increases across collections, particularly in discrete populations, where its panel medians at $\sigma_{\text{CIF}} = 0.1$ rose from approximately 0.51 to 0.69 for ARI and from 0.64 to 0.80 for NMI (Figures 2–5). These differences describe performance across the two collections; they do not isolate the effect of increasing cell counts because feature counts and other simulation conditions also differ.

Higher CIF variability generally reduced MCRF's scores, with a clearer pattern for NMI than for ARI (Figs. 4 and 5). Holding cell count, gene count, and population structure fixed, increasing σ_{CIF} from 0.1 to 0.5 lowered median NMI in 17 of 18 original comparisons and all 12 larger comparisons (Supplementary Table 6). UINMF

also showed lower NMI in linear and discrete populations at higher CIF variability, but this pattern was less consistent in branching populations (Figs. 4 and 5). MCRF's favorable comparisons under some higher variability conditions therefore coexisted with a decline in its own absolute scores.

Discussion

The benchmark supports MCRF as a useful approach to mosaic integration under some of the evaluated conditions, with its clearest relative strength appearing in NMI in the original collection (Fig. 4). Its median NMI exceeded UINMF in 22 of 36 configurations and MIDAS in 28 of 36, with favorable comparisons occurring particularly often in continuous/branching populations (Supplementary Table 6, panel A). These findings indicate that the clusters obtained from MCRF retained information about the reference populations. The less favorable ARI comparisons show that this information did not consistently translate into better recovery of which cell pairs belonged together. The evidence therefore supports a specific strength in the correspondence measured by NMI, alongside limitations in the agreement measured by ARI.

ARI and NMI summarize different properties of a partition, so their rankings need not agree. ARI assesses agreement between cell pairs after adjusting for chance, whereas NMI measures shared information between cluster and population labels without that adjustment (13, 14). MCRF's relatively favorable NMI indicates that its cluster assignments were informative about reference population membership, even where competing methods achieved better agreement in how individual cell pairs were grouped. These findings distinguish the information captured by a partition from how closely it reproduces the reference grouping. MCRF's NMI performance is therefore a meaningful strength, while its weaker ARI shows why both measures are needed to assess clustering performance.

The repeated simulations also showed that MCRF often produced more consistent scores across datasets generated under the same simulation settings. Its IQR was frequently smaller than those of UINMF and MIDAS, particularly for ARI and in the larger collection (Supplementary Tables 5 and 6). In the original collection, this consistency was sometimes accompanied by better performance: MCRF achieved both higher median NMI and a smaller IQR in 11 of 36 configurations compared with UINMF and 17 compared with MIDAS. In these settings, MCRF captured more information about population membership while showing less variation across simulation replicates. In the larger collection, however, its frequently narrower ARI IQR accompanied a lower median ARI than both methods in all 24 configurations with four replicates. MCRF therefore showed consistency in many tested conditions, although this did not overcome its lower clustering performance in the larger collection.

MCRF's weaker relative ranking in the larger collection was accompanied by generally higher absolute scores, particularly for discrete and linear populations. For example, its approximate NMI panel median for discrete populations at $\sigma_{\text{CIF}} = 0.1$ increased from 0.66 to 0.74, while UINMF's increased from 0.64 to 0.80 (Figs. 4 and 5). MCRF therefore achieved better agreement with population labels in these larger datasets, but competing

methods improved more. These comparisons summarize performance across collections with different cell counts, feature counts, and simulation settings. A similar distinction applies to CIF variability: increasing σ_{CIF} generally reduced MCRF's NMI, even where its relative performance became more favorable (Supplementary Table 6). In those settings, its advantage reflected a smaller decline than the comparator rather than unchanged performance.

Population structure also affects how clustering agreements should be interpreted. MCRF's generally higher scores in discrete populations indicate closer recovery of the reference groups. For linear populations, the reference labels divide a continuous trajectory into thirds, so cells near a boundary can be biologically similar despite receiving different labels. Lower agreement can therefore reflect difficulty reproducing these boundaries, which is a different task from preserving the underlying trajectory. The original scMultiSim study also reported lower clustering agreement for continuous populations than for discrete populations (9). MCRF's favorable relative NMI in branching populations suggests that its clusters retained information about the reference populations, while preservation of branch structure and developmental ordering remains a separate question.

MCRF's methodological contribution is a model of observed cell relations that connects the mosaic batches through shared cell factors, as detailed in Methods. Each modality supplies similarities among cells measured in that modality. The paired cells use the same factors in the RNA and ATAC reconstructions, allowing the two sets of relations to constrain a common representation. This makes it possible to integrate the specified mosaic design without converting ATAC regions into gene activity features or reconstructing the molecular measurements of an unobserved modality. The broader use of sample similarities in integration has established precedents, including similarity network fusion (3). MCRF's contribution is its particular factorization of the observed relation blocks, together with modality matrices and a resulting cell embedding, rather than the general idea of using similarities.

This representation gives a possible interpretation of why useful population information can be retained across modalities with different feature spaces. Cells are connected through patterns of similarity within each modality, and the model seeks factors that explain those patterns jointly. The block intercepts account for constant residual offsets, while normalization of each reconstruction loss compares its error with a mean baseline. Together, these components define how similarities contribute to the fit. It is plausible that combining many cell relations and balancing the block losses contributes to the recurring low dispersion, but this remains a hypothesis. The benchmark did not isolate the effects of relation construction, loss normalization, factor constraints, or restart selection. Their individual contributions cannot be inferred from the observed rankings. Relation representation also has limits. Cosine similarities depend on preprocessing and on which features are observed, so replacing feature measurements with relations does not ensure that technical variation is removed. In addition, the fitted cell factors are not validated regulatory programs, and the model does not directly provide feature loadings that explain individual genes or regions.

COBOLT's less favorable clustering results help define the scope of the comparison. MCRF had higher median NMI than COBOLT in all 60 replicated configurations, but this is a statement about population-label agreement under the present benchmark. In the original scMultiSim evaluation, COBOLT also had relatively weak ARI and NMI, while showing better batch mixing than another evaluated approach (9). In a separate benchmark using experimental mosaic RNA and ATAC datasets, COBOLT performed well across several tasks (15). These observations show why its present scores should not be interpreted as a general failure of the method. Our results do not measure batch mixing separately, and the available implementation records do not establish whether COBOLT's published correction for representations inferred from different modality combinations was used. Neither an implementation failure nor a particular explanation involving optimization, preprocessing, or model assumptions has been demonstrated.

More generally, the benchmark compares complete analysis procedures rather than isolated integration objectives. The methods differ in feature processing, latent dimensions, neighborhood construction, and clustering algorithms (Table 2; Methods). MCRF, UINMF, and MIDAS also use the reference labels to select the resolution with the highest NMI, while COBOLT uses its recorded internal optimal-resolution procedure. MCRF's embedding is learned without population labels, but its primary reported clustering is selected with those labels. The scores therefore describe agreement after this selection and should not be read as an estimate of clustering performance when reference labels are unavailable. The selection criterion is particularly relevant to NMI, although its use by several methods does not by itself explain their different rankings.

Finally, the known ground truth supplied by scMultiSim is also a boundary on the conclusions. Both collections arise from one simulator, two modalities, and a three-batch mosaic arrangement. They do not establish performance across experimental tissues, disease states, different modality combinations, or other patterns of missing measurements. ARI and NMI assess agreement with the simulated population labels, but they do not independently assess batch removal, recovery of regulatory relationships, or preservation of all biologically meaningful variation. Broader integration benchmarks distinguish preservation of biological variation from removal of batch effects (15, 16). Experimental validation and evaluation beyond clustering agreement would therefore be needed to determine how the observed strengths of MCRF translate to biological applications.

This study introduces MCRF, a method that integrates cells by factorizing their observed relations. We evaluated MCRF alongside three established methods across 144 original and 102 larger scMultiSim datasets. MCRF compared favorably in NMI under many conditions in the original collection and often produced more consistent scores across simulated replicates. In the larger collection, its absolute scores generally improved, but UINMF and MIDAS achieved higher median ARI in all configurations with four replicates. Together, these findings support cell relation factorization as a promising approach to mosaic integration and identify opportunities to improve its clustering performance on larger datasets. Evaluation on experimental data is the next step toward establishing its value in biological research.

Methods

Simulation framework and biological ground truth

We evaluated the integration of single-cell RNA sequencing and single-cell ATAC sequencing data generated using scMultiSim (9). These two modalities measure gene expression and chromatin accessibility, respectively. To connect them, scMultiSim begins with a specified population structure and an underlying biological state for each cell. It then uses this state, together with information about gene regulation and chromatin accessibility, to generate RNA and ATAC measurements. Technical variation, including observation noise and batch effects, is added afterward. Because the population structure is specified during simulation, the resulting population labels provide a reference for evaluating how well clustering recovers that structure.

The underlying state of each cell is represented by Cell Identity Factors (CIFs). These factors contain four components that describe different sources of biological variation. Nondifferential CIFs represent variation associated with factors shared across populations. Differential CIFs describe the specified population structure. For continuous populations, these factors change gradually along a differentiation tree through a Gaussian random walk, allowing nearby cells to have similar states. For discrete populations, cells vary around the states assigned to the terminal populations. The remaining components represent regulatory influences. Transcription factor CIFs describe the effects of the gene regulatory network (GRN), while ligand CIFs describe how signals from neighboring cells influence a receiving cell. Both components act together with weights that determine their effects on individual genes. The parameter σ_{CIF} controls variability in the construction of CIFs and therefore influences how much cells differ within a population or a nearby part of a trajectory. Its effect on integration performance must be assessed from the results rather than assumed in advance (9).

CIFs describe cells, but genes can respond differently to the same cellular state. Gene Identity Vectors (GIVs) represent these differences by specifying how strongly each CIF contributes to each gene. Combining the CIF and GIV matrices produces a value for each cell and gene. These values are then transformed into three kinetic parameters: k_{on} , the rate at which a gene becomes active; k_{off} , the rate at which it becomes inactive; and s , the rate of RNA synthesis while the gene is active. The effects of the GRN and signals from neighboring cells contribute to the RNA synthesis component through their corresponding weights (9).

Chromatin accessibility is generated through a similar process. Region Identity Vectors (RIVs) specify how the CIFs contribute to each chromatin region. Combining CIFs with RIVs produces accessibility values, which are then rescaled to a reference distribution. These accessibility values are generated before RNA because chromatin states help determine whether a gene can become active. Specifically, accessibility influences k_{on} through the relationships linking chromatin regions to genes. The supplied GRN and the simulator's links between transcription factors, chromatin regions, and genes therefore connect the two modalities through a shared regulatory system (9).

The kinetic parameters are then used to generate RNA counts. In the steady state model, the RNA degradation rate is fixed to define the time scale. For each cell and gene, the simulator first samples an activation fraction, $v \sim \text{Beta}(k_{\text{on}}, k_{\text{off}})$, representing the fraction of time the gene is active. It then samples the RNA count as $x | v \sim \text{Poisson}(vs)$. Thus, the expected count depends on both how often the gene is active and how quickly it produces RNA during those active periods. The full kinetic model also supports RNA velocity, although this option was disabled for the larger dataset collection.

Finally, the simulated biological measurements are subjected to technical effects. These include RNA capture, amplification and sequencing effects, ATAC observation noise, and batch effects. The observed data therefore contains variation beyond the differences between the specified populations. The population labels provide biological ground truth for clustering evaluation, while the added biological and technical variation makes recovering those labels a meaningful integration task (9).

Dataset collections and mosaic observations

We used the 144 simulated multi-omics datasets provided by the original scMultiSim study (9) as the first benchmark collection. Its full design crossed three population structures, two cell counts (500 and 800), three gene counts (110, 200, and 500), two CIF variability levels ($\sigma_{\text{CIF}} = 0.1$ and 0.5), and four simulation replicates (seeds 1–4). The structures were linear, continuous/branching, and discrete, corresponding to the predefined trees Phyla1, Phyla3, and Phyla5. Each cell/gene combination contributed 24 datasets. ATAC contained three times as many regions as RNA genes, giving 330, 600, or 1,500 regions (Table 1; Fig. 1a). Dataset counts refer to joint simulated RNA–ATAC datasets, with the two modalities counted together.

To extend the benchmark to larger datasets, we generated an additional 102 simulated RNA–ATAC datasets using scMultiSim, following a biological simulation design similar to that of the original study. We retained the same population structures and CIF variability levels while increasing the numbers of cells and features. We used four combinations of cell and gene counts: 2,000 cells with 1,000 genes, 5,000 cells with 1,000 genes, 2,000 cells with 2,000 genes, and 5,000 cells with 2,000 genes. Each combination included three population structures, two CIF variability levels, and four simulation replicates (seeds 1–4), yielding 96 datasets. We generated six additional datasets with 10,000 cells and 2,000 genes, using one replicate for each population structure and CIF variability combination. Each dataset contained three times as many ATAC regions as RNA genes. The complete design is summarized in Table 1 and Fig. 1a.

For the linear datasets, cells were divided into three groups based on their position along the trajectory: the first, middle, and final thirds. These groups represent stages along a continuous trajectory, rather than three distinct cell types. For the branching and discrete datasets, we used the population labels provided by the simulator.

Both collections used three mosaic batches: batch 1 contained paired RNA and ATAC, batch 2 contained ATAC only, and batch 3 contained RNA only (Fig. 1b). In the larger simulations, cells were assigned uniformly to three

technical batches with batch-effect size 3, producing approximately equal batch sizes. Full observed matrices retained values for all cells, so modality availability had to be enforced by batch membership. Hidden RNA from batch 2 and hidden ATAC from batch 3 were outside the integration inputs. Population labels were used as the clustering reference rather than as measured molecular features.

Integration methods

Mosaic Cell Relation Factorization

Representing cells through their relations

Matrix factorization represents molecular measurements using cell factors and feature loadings. Mosaic Cell Relation Factorization (MCRF) begins with a different representation of the same cells. Instead of directly factorizing their molecular measurements, it first constructs a matrix $K = XX^T$ whose entries describe how similar each cell is to other cells. The factorization then aims to explain these similarities.

The connection between conventional matrix factorization and MCRF can be seen directly through their mathematical forms. Let $\tilde{X}$ be a preprocessed cell by feature matrix whose rows have unit length. A conventional factorization represents this matrix as $\tilde{X} \approx UV^T$,

where U contains cell factors and V contains feature loadings. The corresponding matrix of cell similarities is $K = \tilde{X}\tilde{X}^T$. Substituting the factorized representation gives

$$K = \tilde{X}\tilde{X}^T \approx (UV^T)(UV^T)^T = U(V^TV)U^T.$$

Thus, factorizing molecular measurements also implies a factorization of cell similarities. Writing $Q = V^TV$ gives simpler expression

$$K \approx UQU^T.$$

MCRF takes this relation model as its starting point. Rather than first fitting the feature matrix, it constructs K from the observed data and directly estimates cell factors Z and a matrix Q that reconstruct the similarities. It also includes an intercept β to account for a constant level of similarity:

$$K \approx ZQZ^T + \beta\mathbf{1}\mathbf{1}^T.$$

Equivalently, the fitted matrix consists of a structured component H and a constant component:

$$\hat{K} = H + \beta\mathbf{1}\mathbf{1}^T, \quad H = ZQZ^T.$$

For two cells i and j , the fitted similarity has the simple form

$$\hat{K}_{ij} = z_i^T Q z_j + \beta,$$

where z_i and z_j are the factor representations of the two cells. Their representations, together with Q , determine how much their fitted similarity differs from the shared intercept.

For mosaic RNA and ATAC data, the same idea is applied to each observed relation block:

$$K_{bb'}^{(m)} \approx Z_b Q_m Z_{b'}^T + \beta_{bb'}^{(m)} \mathbf{1}_{n_b} \mathbf{1}_{n_{b'}}^T.$$

Here, Z_b and $Z_{b'}$ represent the cells in the two batches, while Q_m describes how their factors contribute to similarity in modality m . Cells with both modalities have one shared factor representation, allowing them to connect RNA and ATAC within the same model. Separate matrices Q_R and Q_A allow those shared factors to contribute differently to RNA and ATAC similarities. Integration therefore takes place by jointly modeling observed cell relations, without reconstructing molecular measurements for an unobserved modality.

Let n_b denote the number of cells in batch $b \in 1,2,3$, with $N = n_1 + n_2 + n_3$. Batch 1 contains both RNA and ATAC measurements, batch 2 contains ATAC only, and batch 3 contains RNA only. The observed RNA matrices are $X_1^{(R)} \in \mathbb{R}_+^{n_1 \times G}$ and $X_3^{(R)} \in \mathbb{R}_+^{n_3 \times G}$, where G is the number of genes. The observed ATAC matrices are $X_1^{(A)} \in \mathbb{R}_+^{n_1 \times P}$ and $X_2^{(A)} \in \mathbb{R}_+^{n_2 \times P}$, where P is the number of chromatin regions.

Cell identifiers align the measurements with the metadata, while feature identifiers ensure that batches sharing a modality use the same feature order. MCRF uses all aligned genes and regions, without selecting variable features or requiring a map between regions and genes. Observed counts must be finite and nonnegative, and each observed cell must have a positive total count. Preprocessing and model fitting are performed separately for each dataset.

Preprocessing and relation matrices

Each RNA count is divided by the cell's total count and multiplied by 10,000, then transformed using the natural logarithm after adding one. This standard normalization (17) expresses counts relative to a common total and compresses large values while preserving zeros.

$$Y_{ig}^{(R)} = \log \left(1 + 10,000 \frac{X_{ig}^{(R)}}{\sum_{h=1}^G X_{ih}^{(R)}} \right).$$

ATAC counts are transformed as $Y_{ip}^{(A)} = \log(1 + X_{ip}^{(A)})$. This compresses large accessibility counts while preserving zeros and differences between positive counts that would be lost through binarization. The subsequent normalization of each transformed cell vector to unit length allows cosine similarities to compare accessibility patterns rather than vector magnitudes. However, this does not guarantee removal of sequencing depth effects, because depth can also affect which regions are detected.

For both modalities, each transformed cell vector is then divided by its Euclidean norm:

$$\tilde{X}_{b,i}^{(m)} = \frac{Y_{b,i}^{(m)}}{\|Y_{b,i}^{(m)}\|_2}, \quad m \in \{R, A\}.$$

This gives each row unit length, so the inner product of two rows measures their cosine similarity. A row with zero norm is invalid because this normalization would be undefined. Transformations and relation calculations use double precision.

For batches b and b' that share modality m , the observed relation matrix is

$$K_{bb'}^{(m)} = \tilde{X}_b^{(m)} (\tilde{X}_{b'}^{(m)})^T \in \mathbb{R}^{n_b \times n_{b'}}.$$

Each entry compares one cell in batch b with one cell in batch b' . When $b = b'$, the matrix describes relations within a batch. Otherwise, it describes relations between two batches. Because the transformed measurements are nonnegative, these cosine similarities lie in $[0,1]$. RNA is available in batches 1 and 3, giving the block set $\mathcal{E}_R = (1,1), (1,3), (3,3)$. ATAC is available in batches 1 and 2, giving $\mathcal{E}_A = (1,1), (1,2), (2,2)$. These six blocks contain the observed relations used by MCRF.

A binary mask $M_{bb'}^{(m)}$ determines which entries of the relation matrix contribute to the reconstruction error during fitting. It has the same dimensions as $K_{bb'}^{(m)}$ and contains only zeros and ones. A value of one includes the corresponding cell pair in the loss, while a value of zero excludes it. The mask is applied through entrywise multiplication with the reconstruction residual, $M_{bb'}^{(m)} \odot (K_{bb'}^{(m)} - \hat{K}_{bb'}^{(m)})$, so excluded entries contribute no error to the objective.

All entries between different batches are included. Within a batch, diagonal similarities equal one by construction and contain no information about relations between distinct cells, so the mask excludes them:

$$M_{bb'}^{(m)} = \begin{cases} \mathbf{1}_{n_b} \mathbf{1}_{n_{b'}}^T, & b \neq b', \\ \mathbf{1}_{n_b} \mathbf{1}_{n_b}^T - I_{n_b}, & b = b'. \end{cases}$$

Here, $\mathbf{1}_{n_b} \mathbf{1}_{n_{b'}}^T$ is a matrix of ones, and I_{n_b} is the identity matrix. Subtracting the identity matrix changes the diagonal entries to zero while leaving all other entries equal to one. Note that within a batch, the similarity between cells i and j appears in both positions (i,j) and (j,i) . Both entries remain included, even though they contain the same value.

Cell factors, modality metrics, and objective

MCRF represents each cell using $d = 30$ factors. For batch b , unconstrained parameters $\Theta_b \in \mathbb{R}^{n_b \times d}$ are converted into cell factors through a softmax transformation:

$$Z_{b,ik} = \frac{\exp(\Theta_{b,ik})}{\sum_{h=1}^d \exp(\Theta_{b,ih})}, \quad Z_b \geq 0, \quad Z_b \mathbf{1}_d = \mathbf{1}_{n_b}.$$

Each row of Z_b therefore contains nonnegative weights that sum to one. These weights describe a cell's representation across the 30 factors; they are not assigned to population labels. The paired cells use the same matrix Z_1 in both RNA and ATAC reconstruction, which connects the two modalities during fitting.

The modality matrices determine how the cell factors produce similarities. For each modality, an unconstrained matrix $\Phi_m \in \mathbb{R}^{d \times d}$ is transformed elementwise using $L_m = \text{softplus}(\Phi_m)$, where $\text{softplus}(x) = \log(1 + e^x)$. The modality matrix is then constructed as $Q_m = L_m L_m^T$.

This construction makes Q_m symmetric, positive semidefinite, and entrywise nonnegative. The same Q_R is used for every RNA-seq block, and the same Q_A is used for every ATAC-seq block. Thus, each modality has a common rule for translating cell factors into similarities across its observed batches.

Returning to the basic model introduced above, the structured component for a particular block is

$$H_{bb'}^{(m)} = Z_b Q_m Z_{b'}^T.$$

Each block also has its own intercept. Given the current factors, this intercept is calculated as the mean difference between the observed similarities and the structured component over the included entries:

$$\beta_{bb'}^{(m)} = \frac{\langle M_{bb'}^{(m)}, K_{bb'}^{(m)} - H_{bb'}^{(m)} \rangle_F}{\langle M_{bb'}^{(m)}, M_{bb'}^{(m)} \rangle_F}.$$

The complete fitted block is therefore

$$\hat{K}_{bb'}^{(m)} = H_{bb'}^{(m)} + \beta_{bb'}^{(m)} \mathbf{1}_{n_b} \mathbf{1}_{n_{b'}}^T.$$

Here, $\langle \cdot, \cdot \rangle_F$ denotes the sum of entrywise products. Because the mask is binary, the denominator in the intercept expression counts the included entries. The intercept shifts every fitted entry in the block by the same amount, accounting for the residual mean. It is recalculated from the current factors during each differentiable forward calculation rather than fitted as a separate optimization parameter.

The reconstruction loss measures how closely the fitted similarities match the observed similarities. To make this error easier to interpret, we compare it with the error from a simple baseline that assigns the same mean similarity to every included entry in a block. This follows the idea behind the coefficient of determination, R^2 , which compares a model's residual sum of squares (SSE) with the total sum of squares around the observed mean (SST):

$$R^2 = 1 - \frac{\text{SSE}}{\text{SST}}.$$

MCRF uses a loss corresponding approximately to $1 - R^2$, rather than R^2 itself. A smaller loss therefore indicates better reconstruction relative to the mean baseline. This normalization allows us to combine errors from different blocks on a relative scale, instead of comparing their raw sums of squared errors.

For each block, we first calculate the mean of the observed similarities included by the mask:

$$\bar{K}_{bb'}^{(m)} = \frac{\left\langle M_{bb'}^{(m)}, K_{bb'}^{(m)} \right\rangle_F}{\left\langle M_{bb'}^{(m)}, M_{bb'}^{(m)} \right\rangle_F}.$$

Because the mask contains only zeros and ones, the numerator sums the included similarities and the denominator counts them. This mean is calculated from the observed data and remains fixed during fitting.

Next, we measure how much the observed similarities vary around this mean:

$$S_{bb'}^{(m)} = \left\| M_{bb'}^{(m)} \odot \left(K_{bb'}^{(m)} - \bar{K}_{bb'}^{(m)} \mathbf{1}_{n_b} \mathbf{1}_{n_{b'}}^T \right) \right\|_F^2.$$

This quantity plays the role of SST. It is the reconstruction error we would obtain if we ignored differences between cell pairs and predicted the block mean everywhere. Here, $\odot$ denotes entrywise multiplication, and the squared Frobenius norm, $\|\cdot\|_F^2$, sums the squared entries. The mask ensures that only the selected comparisons contribute.

The model's reconstruction loss is then

$$\ell_{bb'}^{(m)} = \frac{\left\| M_{bb'}^{(m)} \odot \left(K_{bb'}^{(m)} - \bar{K}_{bb'}^{(m)} \right) \right\|_F^2}{S_{bb'}^{(m)} + 10^{-12}}.$$

The numerator plays the role of SSE: it measures the remaining error after fitting MCRF. Dividing by the baseline error expresses how much variation remains unexplained. Ignoring the small numerical stabilizer, a loss of zero indicates perfect reconstruction, a loss of one indicates the same error as predicting the block mean, and a loss above one indicates worse reconstruction than that baseline. Each block must satisfy $S_{bb'}^{(m)} > 10^{-12}$ to avoid normalization by an effectively constant block. The additional 10^{-12} in the denominator provides numerical stabilization.

The full objective adds the six normalized block losses and a penalty on the modality matrices:

$$\mathcal{L}_{\text{MCRF}} = \sum_{(b,b') \in \mathcal{E}_R} \ell_{bb'}^{(R)} + \sum_{(b,b') \in \mathcal{E}_A} \ell_{bb'}^{(A)} + 10^{-3} (\|Q_R\|_F^2 + \|Q_A\|_F^2).$$

Each normalized block loss receives weight one. The final term discourages large values in Q_R and Q_A .

Minimizing the objective therefore balances accurate reconstruction of the observed cell similarities with control over the magnitude of the learned modality matrices.

Optimization and integrated embedding

MCRF fits the model using Adam, with all six relation blocks included in every optimization step. The quantities updated during training are the cell parameters Θ_b and the modality parameters Φ_m . These determine the cell factors Z_b and modality matrices Q_m through the transformations described above. The learning rate is 0.01, weight decay is zero, and calculations use double precision. The global gradient norm is clipped at 10 to limit unusually large gradients.

The unconstrained cell parameters are initialized independently as $\Theta_{b,ik} \sim \mathcal{N}(0, 0.01^2)$. Initializing Θ_b near zero gives initial softmax weights close to $1/30$, with small random differences across cells and factors.

The metric factors L_m are initialized near the identity, with intended diagonal entries of 1 and other entries of 10^{-3} . The corresponding starting values of Φ_m are obtained by applying inverse softplus, followed by Gaussian noise with standard deviation 10^{-6} .

MCRF is fitted from three independent random initializations using seeds 123, 456, and 789, with the dataset, objective, and training settings unchanged. These fitting seeds are separate from simulation seeds. Fixed seeds and deterministic operations support reproducibility under the same computational setup.

During each restart, MCRF saves the parameter values with the smallest finite objective encountered so far. Training continues while the objective is making meaningful progress. Specifically, an improvement at step t must satisfy

$$\frac{L_{\text{best}} - L_t}{\max(|L_{\text{best}}|, 10^{-12})} > 10^{-6},$$

where L_{best} is the smallest loss observed before that step. Training runs for at least 1,000 steps and stops after 200 consecutive steps without meaningful improvement, subject to a maximum of 4,000 steps.

The best saved solution from each restart is compared, and the one with the smallest objective is retained. If objectives are tied within 10^{-12} , the smaller seed is chosen. Thus, selection depends only on the unsupervised fitting objective. Population labels, clustering scores, and the visual appearance of the embedding are not used.

The clustering representation must incorporate the learned modality matrices because structured similarity depends on $Z_b Q_m Z_b^T$, rather than Z_b alone. Euclidean distances on Z_b would omit this learned weighting.

To obtain one representation incorporating both modalities, MCRF first divides each modality matrix by its trace and then averages the two matrices:

$$\tilde{Q}_m = \frac{Q_m}{\text{tr}(Q_m) + 10^{-12}}, \quad Q_* = \frac{1}{2}(\tilde{Q}_R + \tilde{Q}_A).$$

The trace is the sum of the diagonal entries. This normalization puts the matrices on a common trace scale, so one does not receive greater weight in the average simply because its overall trace is larger. Each trace must exceed 10^{-12} . The resulting Q_* provides a combined rule for how the cell factors contribute to similarity.

The goal is now to build cell coordinates whose ordinary inner products reproduce this combined similarity. This is achieved by multiplying the cell factors by the matrix square root of Q_* :

$$E_b = Z_b Q_*^{1/2} \in \mathbb{R}^{n_b \times 30}.$$

The square root is used because an inner product combines two embedding vectors. Each contributes one factor of $Q_*^{1/2}$, and the two factors multiply to give Q_* :

$$E_b E_{b'}^T = Z_b Q_*^{1/2} (Q_*^{1/2})^T Z_{b'}^T = Z_b Q_* Z_{b'}^T.$$

In this way, the information in Q_* becomes part of the coordinates themselves. The same connection applies to distances. If z_i and z_j are two rows of the cell factor matrix, and e_i and e_j are their transformed coordinates, then

$$\|e_i - e_j\|_2^2 = (z_i - z_j) Q_* (z_i - z_j)^T.$$

Thus, ordinary distances between rows of E account for the learned importance and combinations of the factors. This is why downstream clustering uses E rather than Z alone.

To calculate the matrix square root, Q_* is first numerically symmetrized and decomposed as $Q_* = U \Lambda U^T$. Its symmetric positive semidefinite square root is then

$$Q_*^{1/2} = U \Lambda_+^{1/2} U^T.$$

Small negative eigenvalues in $[-10^{-10}, 0)$ may be set to zero to accommodate numerical error. Eigenvalues below -10^{-8} fail validation.

The embedding contains the structured cell representation. The block intercepts are not added to the coordinates because they describe constant offsets in the relation blocks rather than individual cell factors.

Finally, the batch embeddings are concatenated and restored to the cell order specified in the metadata. The resulting $N \times 30$ matrix E is used for neighborhood construction and downstream clustering, without additional standardization, row normalization, principal component transformation, or batch correction.

Comparison methods

UINMF

Unshared integrative nonnegative matrix factorization (UINMF) integrates datasets using both features shared across datasets and features available in only one dataset (6). For input group q , let $A_q \in \mathbb{R}_+^{n_q \times g}$ contain the

shared features and $B_q \in \mathbb{R}_+^{n_q \times u_q}$ contain the unshared features. Both matrices are represented using the same cell factors $F_q \in \mathbb{R}_+^{n_q \times k}$:

$$A_q \approx F_q(W + V_q), \quad B_q \approx F_q U_q.$$

Here, W describes feature patterns shared across groups, while V_q describes patterns specific to group q within the shared features. The matrix U_q describes the unshared features. Because both reconstructions use F_q , information from shared and unshared features contributes to the same cell representation.

The published objective is

$$\min_{F_q, W, V_q, U_q \geq 0} \sum_q \left[\|A_q - F_q(W + V_q)\|_F^2 + \|B_q - F_q U_q\|_F^2 + \lambda_q (\|F_q V_q\|_F^2 + \|F_q U_q\|_F^2) \right].$$

The first two terms measure reconstruction error. The remaining terms penalize the contributions specific to each group, balancing their influence against the shared component. The model is fitted by alternating updates to the factor matrices while keeping their entries nonnegative. UINMF therefore provides a comparison that factorizes feature measurements, whereas MCRF factorizes cell similarities.

The benchmark used rliqr 1.0.1 with $k = 15$, regularization parameter 5, a maximum of 30 alternating least squares iterations, and convergence threshold 10^{-10} . The resulting cell factors underwent quantile normalization before downstream analysis. RNA preprocessing included normalization by total counts, variable gene selection using package defaults, and scaling without centering. For cells with ATAC only, the supplied map between regions and genes was used to construct gene activity features that could be matched to RNA genes. Peaks without gene links were retained as unshared features (Table 2).

COBOLT

COBOLT uses a multimodal variational autoencoder (VAE) to represent cells in a shared latent space using their observed RNA and ATAC counts (7). Its inference networks map the available measurements to a latent representation, while its generative model describes how that representation produces counts in each modality.

The generative model converts a Gaussian latent variable into nonnegative category proportions that sum to one. These categories describe patterns across features, and each modality has its own feature probabilities. Together, the category proportions and feature probabilities define multinomial distributions for the observed counts. Separate inference networks allow cells to contribute RNA, ATAC, or both modalities.

The published framework also accounts for technical batches and describes a correction learned from paired cells to align representations inferred from different combinations of modalities. The available implementation records do not establish whether this correction was used in the present benchmark.

We used COBOLT commit cd8015b, a latent dimension of 10, and 20 training epochs, with learning rates of 0.005 for the original collection and 0.001 for the larger collection. Features were filtered separately within batches using lower and upper expression quantile thresholds of 0.50 and 0.99. The saved latent representation was used for clustering (Table 2).

MIDAS

MIDAS is a probabilistic model for mosaic integration that represents biological differences between cells separately from technical variation (8). Its modular VAE uses an encoder for each available modality and combines Gaussian estimates from these encoders and batch information through a product of experts. This allows the model to construct a representation from whichever modalities are observed.

The model uses biological and technical latent variables to reconstruct the measurements, with Poisson distributions for RNA and Bernoulli distributions for ATAC. A separate decoder predicts batch identity from the technical variables. Additional training objectives encourage the biological representation to remain consistent when inferred from different modality subsets and encourage biological and technical information to remain separate. Although MIDAS also supports imputation and knowledge transfer, this benchmark evaluates only its biological representation.

The benchmark used scmidas 0.3.0 with 32 biological latent dimensions and two technical latent dimensions. Training used a maximum of 1,000 epochs, batch size 128, and learning rate 10^{-4} . ATAC counts were binarized within the pipeline, and only the biological embedding was used for clustering evaluation (Table 2).

Clustering evaluation and presentation of results

We evaluated clustering using the adjusted Rand index (ARI) and normalized mutual information (NMI). Both statistics compare the clusters obtained from an integration method with the known population labels generated by the simulator. They assess whether cells are grouped in a way that agrees with the biological reference provided by the simulation tool scMultiSim. All cells across the three batches were included in each evaluation. Each method supplied clusters derived from its own latent representation.

Clustering and resolution selection

For both dataset collections, MCRF constructed a graph connecting each cell to its 30 nearest neighbors using Euclidean distances in the integrated embedding. Leiden clustering was applied at resolutions from 0.1 to 2.0 in increments of 0.1 to generate candidate partitions, with higher resolutions generally producing more clusters (18). After generating these partitions, the population labels were used to select the one with the highest NMI. Ties within 10^{-12} were resolved by higher ARI, proximity of the resolution to 0.9, and finally the lower resolution. A separate partition at resolution 0.9 was retained but was not used for the primary score. Thus, MCRF learned its

embedding without population labels, while the primary clustering result used those labels to select the resolution.

The comparison methods used different graph and clustering settings. UINMF applied Louvain clustering to a graph with 15 neighbors, searched the same resolution grid, and selected the partition with the highest NMI (19). MIDAS applied Leiden clustering to a graph with 15 neighbors, searched the same resolution grid, and used the same selection and tie rules as MCRF. COBOLT used a Euclidean graph with the Scanpy default of 15 neighbors, followed by the internal optimal resolution clustering and scoring procedure in scIB 1.1.7 (16).

Adjusted Rand index

ARI measures agreement by checking whether pairs of cells are grouped together or kept apart in both the inferred clusters and reference populations. A pair agrees when both partitions group the cells together or both separate them, and disagrees otherwise. Because some agreement can occur by chance, ARI adjusts for the agreement expected between random partitions with the same cluster and population sizes (13).

Let n_{ij} be the number of cells assigned to the inferred cluster i and reference population j . Let $a_i = \sum_j n_{ij}$ be the size of the inferred cluster i , $b_j = \sum_i n_{ij}$ the size of the reference population j , and N the total number of cells.

Define

$$S = \sum_{ij} \binom{n_{ij}}{2}, \quad A = \sum_i \binom{a_i}{2}, \quad B = \sum_j \binom{b_j}{2}, \quad T = \binom{N}{2}.$$

Here, $\binom{n}{2} = n(n-1)/2$ counts the distinct pairs among n cells. Thus, S counts pairs placed together in both partitions, A counts pairs placed together by clustering, B counts pairs belonging to the same reference population, and T counts all cell pairs.

Under the random partition model, the expected value of S is AB/T . ARI compares the observed value with this expectation and normalizes the difference:

$$\text{ARI} = \frac{S - AB/T}{\frac{1}{2}(A + B) - AB/T}.$$

An ARI of 1 indicates identical partitions, allowing for different cluster names. An ARI of 0 indicates agreement equal to the expectation under the specified random partition model. Negative values indicate agreement below that expectation. ARI therefore measures recovery of the reference grouping after accounting for chance. MCRF used the adjusted Rand implementation in scikit-learn, while UINMF used the corresponding implementation in mclust.

Normalized mutual information

NMI measures how much knowing a cell's inferred cluster reduces uncertainty about its reference population. Clusters concentrated within particular populations provide more information than clusters containing the same population mixture as the full dataset. Higher NMI indicates closer correspondence between the partitions.

This idea is expressed through mutual information and entropy (14). Using the same counts as above, define the observed proportions

$$p_{ij} = \frac{n_{ij}}{N}, \quad p_i = \frac{a_i}{N}, \quad p_j = \frac{b_j}{N}.$$

The joint proportion p_{ij} describes how frequently cluster i and population j occur together. If cluster and population membership were independent, this proportion would equal $p_i p_j$. Mutual information measures their departure from independence:

$$I = \sum_{i,j:p_{ij}>0} p_{ij} \log \left(\frac{p_{ij}}{p_i p_j} \right).$$

Entropy describes uncertainty in membership before information from the other partition is supplied. The entropies of the inferred clusters and reference populations are

$$H_C = - \sum_i p_i \log p_i, \quad H_T = - \sum_j p_j \log p_j.$$

For example, uncertainty is greater when cells are spread evenly across several populations than when almost all cells belong to one population. Because the amount of available information depends on this distribution, mutual information is normalized by the entropies of the two partitions.

Under arithmetic normalization, NMI is

$$\text{NMI}_{\text{arith}} = \frac{I}{(H_C + H_T)/2} = \frac{2I}{H_C + H_T}.$$

This divides the shared information by the average uncertainty in the two partitions. A value of 1 indicates identical partitions, allowing for different labels. A value of 0 indicates independence in the observed proportions, so cluster membership provides no information about population membership. Intermediate values indicate partial correspondence.

Unlike ARI, NMI does not subtract the agreement expected by chance. Random partitions can therefore have positive NMI in finite datasets. ARI and NMI provide complementary descriptions of clustering agreement: ARI evaluates cell pairs with a chance correction, while NMI evaluates shared information relative to the uncertainty in the partitions.

Presentation and interpretation of results

The figures present ARI and NMI separately for the original and larger dataset collections summarized in Table 1 and Fig. 1a.

Each performance figure (Figures 2–5) contains six panels, one for each combination of population structure and σ_{CIF} level. Within each panel, all methods are compared on the same datasets, pooling the different combinations of cell and gene counts and available simulation seeds. The spread of scores therefore reflects performance across different simulated datasets. Even when all other simulation settings remain unchanged, a different simulation seed generates a new dataset; it does not represent another run of the method on the same dataset with a different initialization seed.

Supplementary Tables 1–4 report one ARI and NMI pair for each method and dataset, rounded to two decimal places. For descriptive comparisons across simulation replicates, datasets were grouped by configuration, holding all simulation settings fixed except the simulation seed. For each method, we calculated the median and interquartile range of ARI and NMI across four replicates, where $\text{IQR} = Q_3 - Q_1$ and Q_1 and Q_3 are the 25th and 75th percentiles. Quartiles used linear interpolation, and comparisons used unrounded scores, treating differences within 10^{-12} as ties. Supplementary Tables 5 and 6 report ARI and NMI as median (Q_1, Q_3) to three decimal places, with 36 original configurations in Panel A and 30 larger configurations in Panel B. The six configurations with one replicate show a single score marked with an asterisk and were excluded from variability comparisons. Medians and IQRs were compared between methods within each configuration. These summaries describe variation across separately simulated datasets, not repeated method initialization, and do not establish statistical significance.

Software environments: The original benchmark ran on the UAB Cheaha computing cluster using CPU execution and separate method environments. MCRF used Python 3.10.4, PyTorch 2.5.1+cpu, scikit-learn 1.0.2, Scanpy 1.9.6, and leidenalg 0.9.1. UINMF used R 4.4.1 and rligier 1.0.1. COBOLT environment used Python 3.12.3, COBOLT 0.0.1 at the specified source commit, and scIB 1.1.7. MIDAS used Python 3.12.3 and scmidas 0.3.0. The larger datasets were generated with R 4.4.1, Bioconductor 3.20, and scMultiSim 1.2.0.

Data availability: The original 144 simulated datasets are publicly available through the scMultiSim authors' Dropbox share at <https://www.dropbox.com/sh/sfkn5hweaejbrir/AAB9liDyL8QuXd7LAGUsihGfa?dl=0>. Code and instructions for generating the 102 larger simulated datasets are available in the MCRF-manuscript <https://github.com/sandeepcvej-yogi/MCRF-manuscript>, together with the MCRF implementation and benchmarking workflows. Dataset-level ARI and NMI results are provided in Supplementary Tables 1–4, and summaries across simulation replicates are provided in Supplementary Tables 5 and 6.

Code availability: The MCRF source code, benchmarking workflows for COBOLT, UINMF, MIDAS, and MCRF, and scripts for generating the 102 larger scMultiSim datasets are publicly available at

<https://github.com/sandeepcvej-yogi/MCRF-manuscript>. The repository includes software requirements and instructions for generating datasets, running the integration methods, and calculating ARI and NMI.

References

1. Hasin Y, Seldin M, Lusis A. Multi-omics approaches to disease. *Genome Biol.* 2017;18(1):83.
2. Baiao AR, Cai Z, Poulos RC, Robinson PJ, Reddel RR, Zhong Q, et al. A technical review of multi-omics data integration methods: from classical statistical to deep generative approaches. *Brief Bioinform.* 2025;26(4).
3. Wang B, Mezlini AM, Demir F, Fiume M, Tu Z, Brudno M, et al. Similarity network fusion for aggregating data types on a genomic scale. *Nature methods.* 2014;11(3):333–7.
4. Stephenson E, Reynolds G, Botting RA, Calero-Nieto FJ, Morgan MD, Tuong ZK, et al. Single-cell multi-omics analysis of the immune response in COVID-19. *Nature medicine.* 2021;27(5):904–16.
5. Ma S, Zhang B, LaFave LM, Earl AS, Chiang Z, Hu Y, et al. Chromatin potential identified by shared single-cell profiling of RNA and chromatin. *Cell.* 2020;183(4):1103–16. e20.
6. Kriebel AR, Welch JD. UINMF performs mosaic integration of single-cell multi-omic datasets using nonnegative matrix factorization. *Nature communications.* 2022;13(1):780.
7. Gong B, Zhou Y, Purdom E. Cobolt: integrative analysis of multimodal single-cell sequencing data. *Genome biology.* 2021;22(1):351.
8. He Z, Hu S, Chen Y, An S, Zhou J, Liu R, et al. Mosaic integration and knowledge transfer of single-cell multimodal data with MIDAS. *Nature biotechnology.* 2024;42(10):1594–605.
9. Li H, Zhang Z, Squires M, Chen X, Zhang X. scMultiSim: simulation of single-cell multi-omics and spatial data guided by gene regulatory networks and cell–cell interactions. *Nature Methods.* 2025;22(5):982–93.
10. Fu S, Wang S, Si D, Li G, Gao Y, Liu Q. Benchmarking single-cell multi-modal data integrations. *Nature Methods.* 2025;22(11):2437–48.
11. Misra R, Ferrena A, Zheng D. Facilitate integrated analysis of single cell multiomic data by binarizing gene expression values. *Nature Communications.* 2025;16(1):5763.
12. Wang Y, Fan Y, Wang X, Yu T, Zong Y, Liu X, et al. SCMBench: benchmarking domain-specific and foundation models for single-cell multi-omics data integration. *Nature Communications.* 2026.
13. Hubert L, Arabie P. Comparing partitions. *Journal of classification.* 1985;2(1):193–218.
14. Vinh NX, Epps J, Bailey J, editors. Information theoretic measures for clusterings comparison: is a correction for chance necessary? *Proceedings of the 26th annual international conference on machine learning*; 2009.
15. Liu C, Ding S, Kim HJ, Long S, Xiao D, Ghazanfar S, et al. Multitask benchmarking of single-cell multimodal omics integration methods. *Nature Methods.* 2025;22(11):2449–60.
16. Luecken MD, Büttner M, Chaichoompu K, Danese A, Interlandi M, Müller MF, et al. Benchmarking atlas-level data integration in single-cell genomics. *Nature methods.* 2022;19(1):41–50.
17. Satija L. Seurat - Guided Clustering Tutorial. Seurat documentation. 2026.
18. Traag VA, Waltman L, Van Eck NJ. From Louvain to Leiden: guaranteeing well-connected communities. *Scientific reports.* 2019;9(1):5233.
19. Blondel VD, Guillaume J-L, Lambiotte R, Lefebvre E. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment.* 2008;2008(10):P10008.

Acknowledgements: The analyses were performed using the Cheaha high-performance computing system at the University of Alabama at Birmingham.

Funding: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Author contributions: SCV and HKT conceived the study. SCV developed and implemented MCRF. SCV generated the larger simulated datasets, performed the benchmarking analyses, and prepared the figures and tables. HKT supervised the research. SCV drafted the manuscript, and HKT reviewed and revised it.

Competing interests: The authors declare no competing interests.

Table 1. Summary of the simulated scMultiSim datasets

Panel A. Dataset dimensions				
Collection	Number of Cells	Number of Genes	Number of ATAC-seq regions	Unique datasets
Original	500	110	330	24
	500	200	600	24
	500	500	1,500	24
	800	110	330	24
	800	200	600	24
	800	500	1,500	24
Large	2,000	1,000	3,000	24
	5,000	1,000	3,000	24
	2,000	2,000	6,000	24
	5,000	2,000	6,000	24
	10,000	2,000	6,000	6

Note. Original datasets: 144; large datasets: 102; combined total: 246. RNA and ATAC are counted together per dataset. ATAC region counts are three times gene counts.

Panel B. Common simulation design	
Design element	Specification
Population structures	Linear; continuous/branching; discrete
CIF variability σ_{CIF}	0.1 and 0.5
Replicate seeds	Original: 1, 2, 3, 4; larger 2,000/5,000-cell settings: 1, 2, 3, 4; 10,000 cells: 1 only
Mosaic batches	Batch 1: RNA and ATAC; Batch 2: ATAC only; Batch 3: RNA only

Table 2. Implementation details for the four integration methods

Method	Latent dimension	Main model and training settings	Preprocessing affecting the embedding
COBOLT	10	• Training: 20 epochs; learning rates of 0.005 for the original collection and 0.001 for the larger collection. • Implementation: epurdom/cobolt@cd8015b (GitHub commit).	• Feature filtering separately within each batch. • Expression quantile thresholds: lower = 0.50; upper = 0.99.
UINMF (rliger 1.0.1)	15 factors	• Optimization: optimizeALS; lambda = 5. • Maximum 30 iterations; convergence threshold = 1e-10. • Factor alignment: quantile_norm.	• RNA: Normalize, selectGenes, and scaleNotCenter. • ATAC-only batch: gene activity scores from region_to_gene. • Peaks without gene links are retained as unshared features.
MIDAS (scmidas 0.3.0)	32 biological dimensions	• Training: maximum 1,000 epochs; batch size = 128. • Learning rate = 1e-4; num_workers = 0. • Scored representation: 32 biological latent dimensions. • Two additional technical latent dimensions excluded.	• ATAC counts binarized within MIDAS before learning the embedding.
MCRF	30	• Optimization: Adam; learning rate = 0.01. • 1,000–4,000 steps; early-stopping patience = 200. • Relative loss tolerance = 1e-6. • Numerics: gradient clipping = 10; float64. • Restarts: seeds 123, 456, and 789.	• RNA: library-size normalization to 10,000 counts per cell, then log1p. • ATAC: log1p transformation. • Both modalities: row-wise L2 normalization of transformed counts before calculating cosine similarities.

Note 1. Fitting seed for UINMF, COBOLT, and MIDAS: 123. MCRF used three fitting seeds: 123, 456, and 789. These differ from the simulation seeds.

Note 2. ARI and NMI are calculated against simulated population labels across all three batches using clusters derived from each method's own latent representation.

Note 3. Method-specific graph construction and clustering resolution selection are described in Methods.

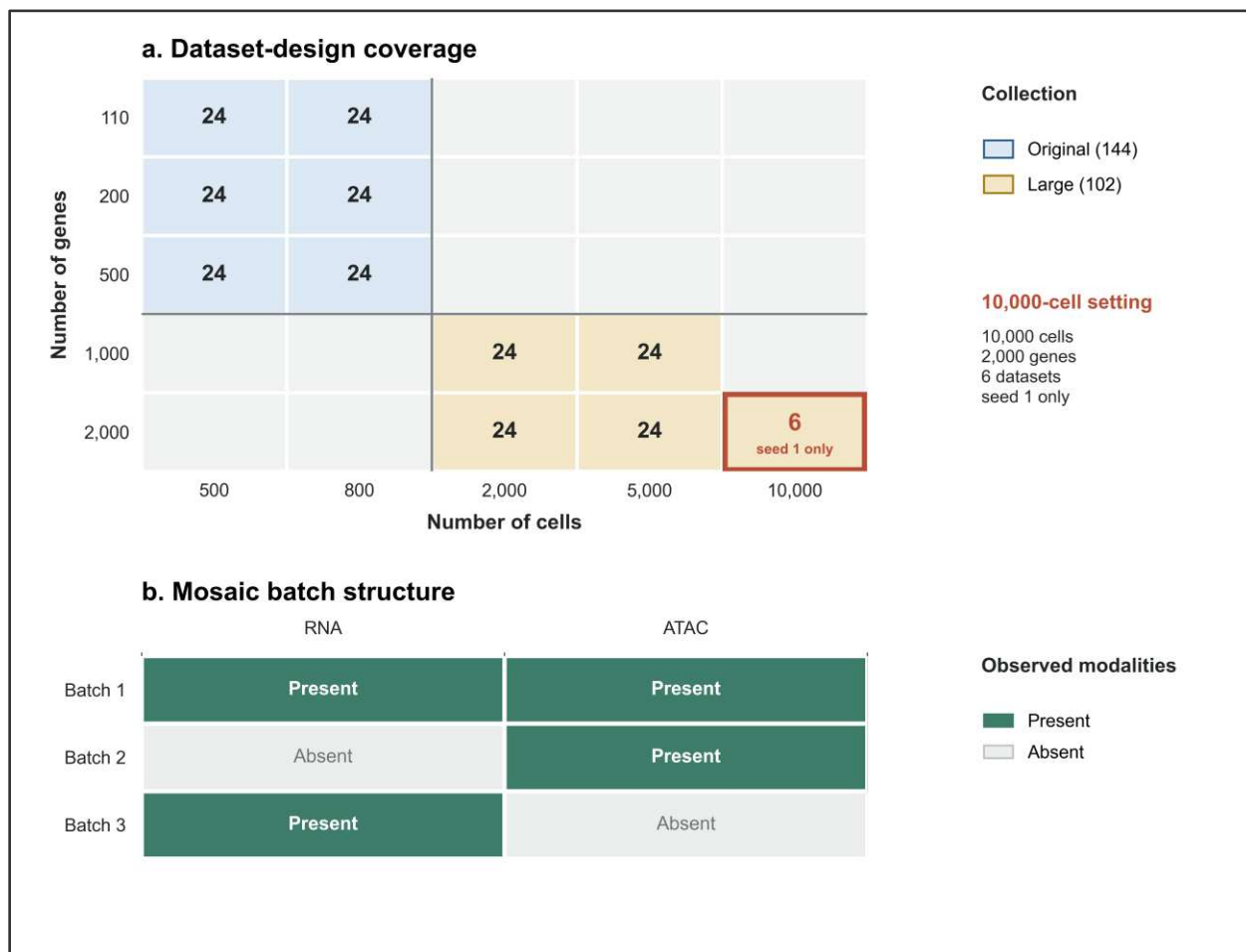


Figure 1. Dataset design and mosaic batch structure

a. Coverage of cell/gene combinations. Blue and gold cells identify the original and larger collections; grey cells indicate combinations not included. Counts refer to jointly simulated RNA-seq and ATAC-seq datasets, with the modalities counted together. Each count of 24 combines three population structures, two CIF variability levels, and four simulation seeds. The 10,000 cells, 2,000 genes setting contains six datasets using seed 1 only.

b. Modality availability across mosaic batches: batch 1 contains RNA-seq and ATAC-seq, batch 2 ATAC-seq only, and batch 3 RNA-seq only. Green indicates an observed modality and grey an unavailable modality for integration.

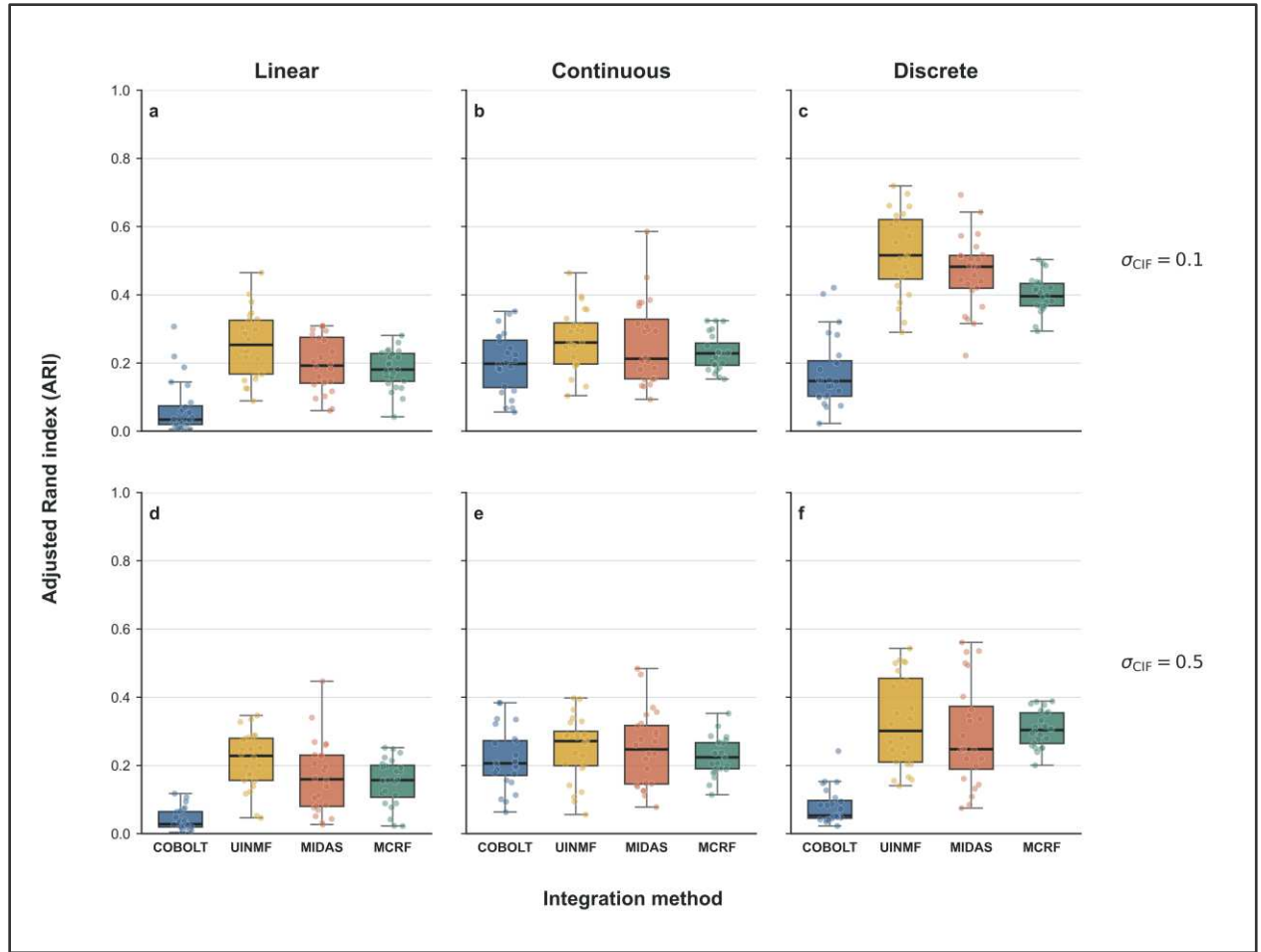


Figure 2. ARI performance on the original 144 datasets

Panels a, b, and c show linear, continuous/branching, and discrete populations at $\sigma_{CIF} = 0.1$; panels d, e, and f show the same structures at $\sigma_{CIF} = 0.5$. Continuous denotes continuous/branching populations. Each point is one simulated dataset evaluated by that method. Boxes span the 25th to 75th percentiles, with the median marked inside. Whiskers extend to the most extreme observations within 1.5 times the interquartile range of the box edges. All observations, including those beyond the whiskers, are overlaid as points; separate outlier symbols are suppressed. Each panel contains the same 24 datasets per method: six cell/gene combinations and four simulation seeds. Boxes pool sizes and seeds, not repeated method initializations on one dataset.

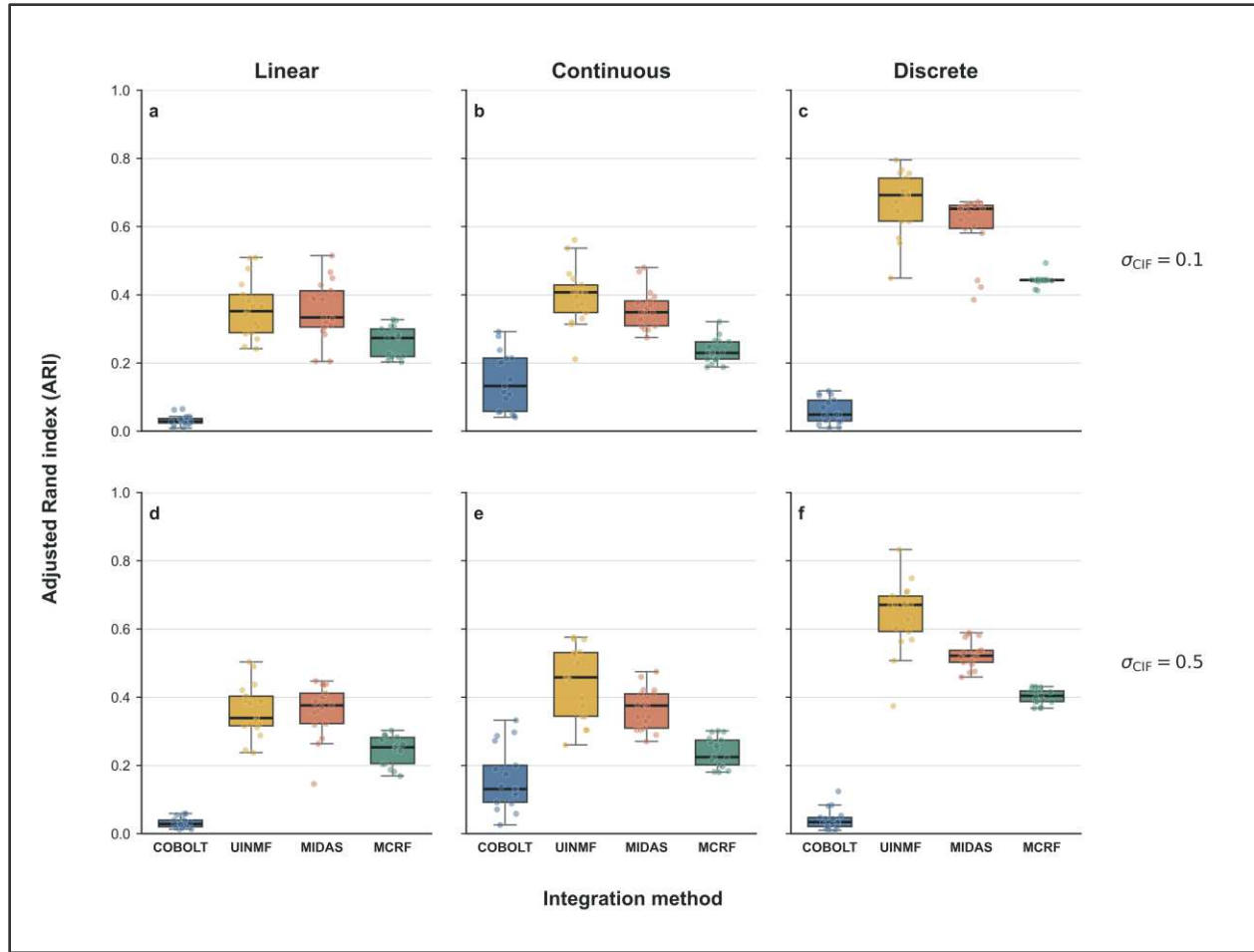


Figure 3. ARI performance on the 102 large datasets

Panels a, b, and c show linear, continuous/branching, and discrete populations at $\sigma_{CIF} = 0.1$; panels d, e, and f show the same structures at $\sigma_{CIF} = 0.5$. Continuous denotes continuous/branching populations. Each point is one simulated dataset evaluated by that method. Boxes span the 25th to 75th percentiles, with the median marked inside. Whiskers extend to the most extreme observations within 1.5 times the interquartile range of the box edges. All observations, including those beyond the whiskers, are overlaid as points; separate outlier symbols are suppressed. Each panel contains the same 17 datasets per method: four cell/gene combinations with four simulation seeds each, plus one 10,000-cell dataset. Boxes pool sizes and seeds, not repeated method initializations on one dataset.

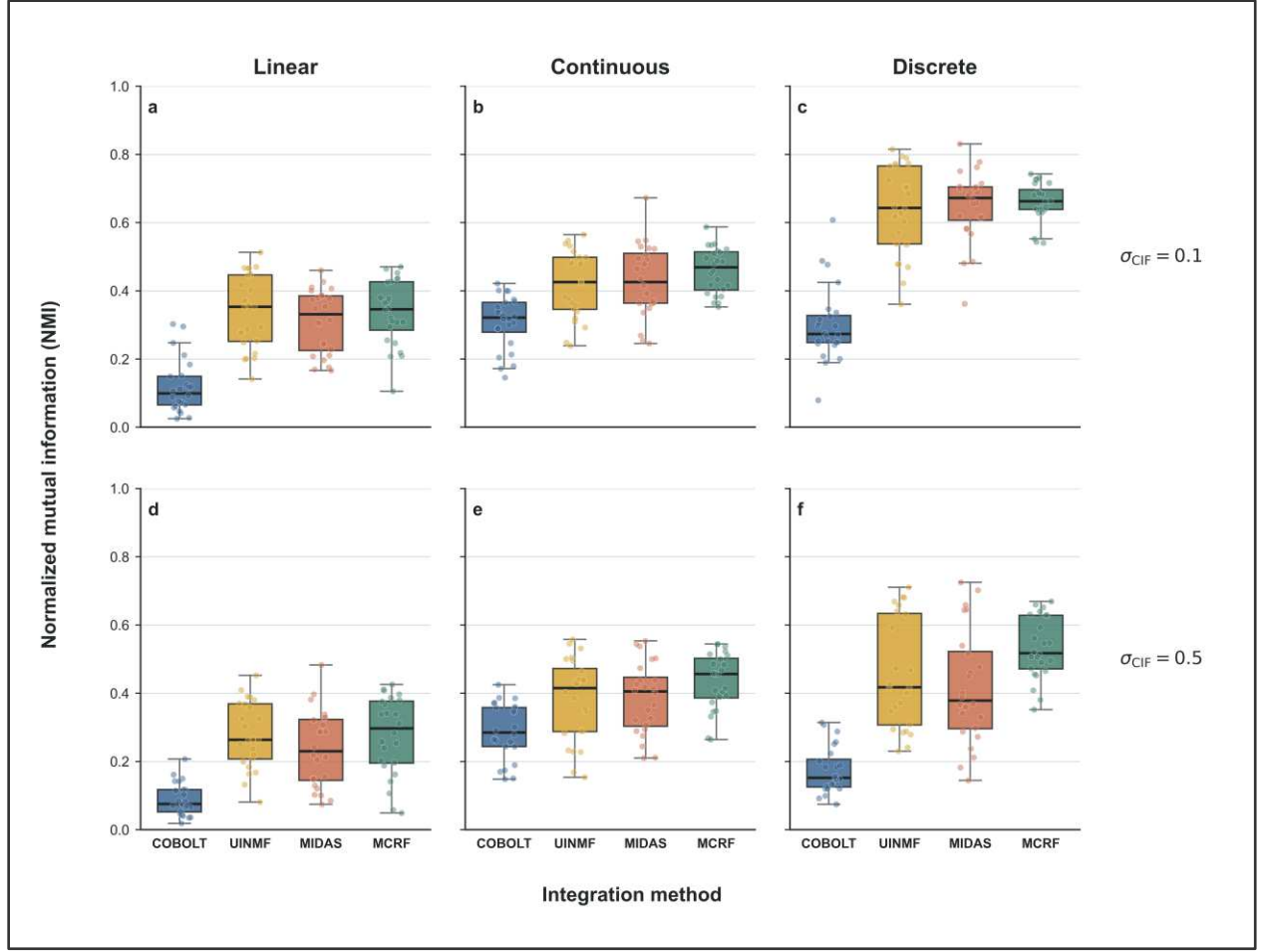


Figure 4. NMI performance on the original 144 datasets

Panels a, b, and c show linear, continuous/branching, and discrete populations at $\sigma_{CIF} = 0.1$; panels d, e, and f show the same structures at $\sigma_{CIF} = 0.5$. Continuous denotes continuous/branching populations. Each point is one simulated dataset evaluated by that method. Boxes span the 25th to 75th percentiles, with the median marked inside. Whiskers extend to the most extreme observations within 1.5 times the interquartile range of the box edges. All observations, including those beyond the whiskers, are overlaid as points; separate outlier symbols are suppressed. Each panel contains the same 24 datasets per method: six cell/gene combinations and four simulation seeds. Boxes pool sizes and seeds, not repeated method initializations on one dataset.

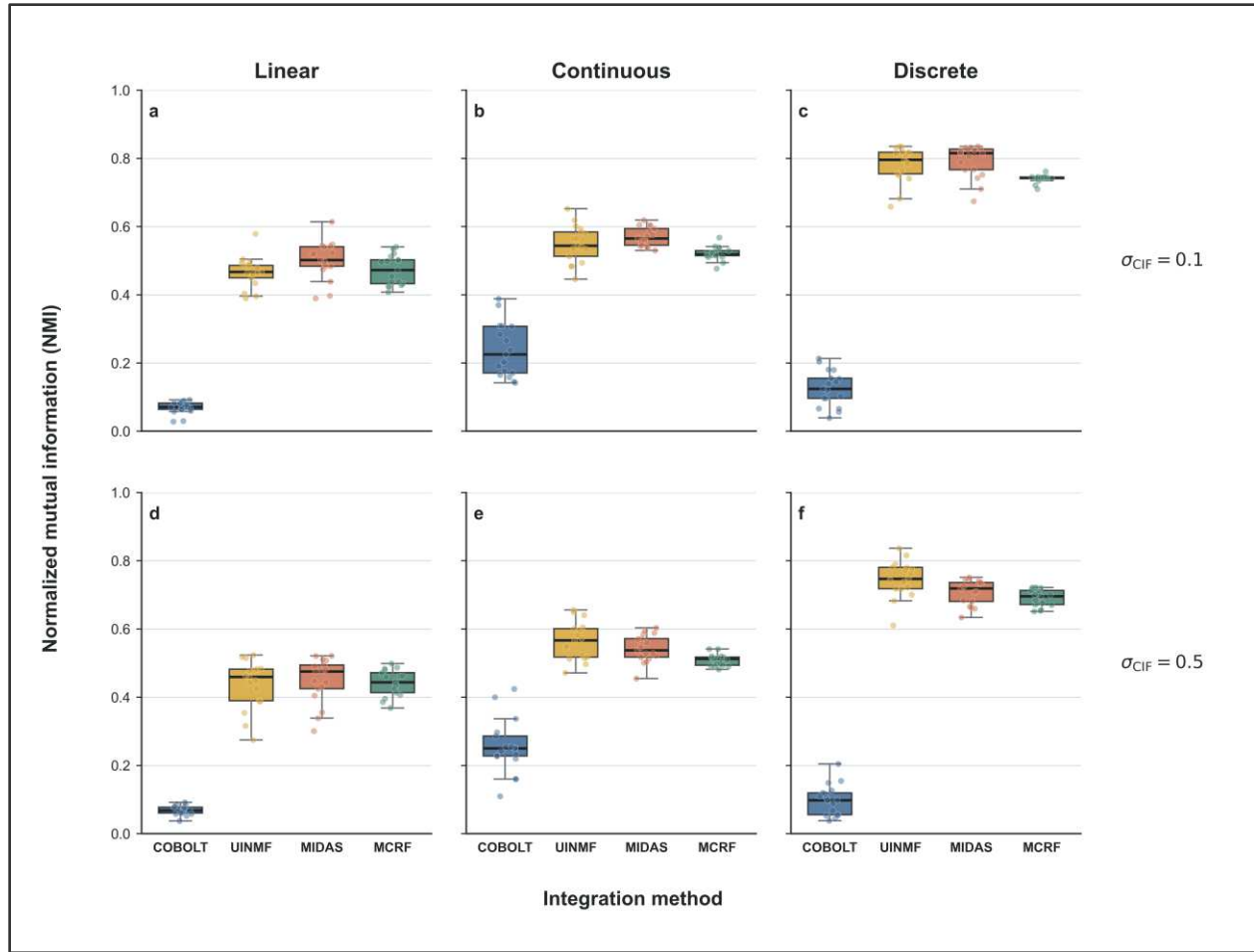


Figure 5. NMI performance on the 102 large datasets

Panels a, b, and c show linear, continuous/branching, and discrete populations at $\sigma_{CIF} = 0.1$; panels d, e, and f show the same structures at $\sigma_{CIF} = 0.5$. Continuous denotes continuous/branching populations. Each point is one simulated dataset evaluated by that method. Boxes span the 25th to 75th percentiles, with the median marked inside. Whiskers extend to the most extreme observations within 1.5 times the interquartile range of the box edges. All observations, including those beyond the whiskers, are overlaid as points; separate outlier symbols are suppressed. Each panel contains the same 17 datasets per method: four cell/gene combinations with four simulation seeds each, plus one 10,000-cell dataset. Boxes pool sizes and seeds, not repeated method initializations on one dataset.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [SupplementaryTables.xlsx](#)