Briefings in Bioinformatics, 2024, **25(1)**, 1–16
https://doi.org/10.1093/bib/bbad455
Problem Solving Protocol

Model-based evaluation of spatiotemporal data reduction methods with unknown ground truth through optimal visualization and interpretability metrics

Komlan Atitey, Alison A. Motsinger-Reif and Benedict Anchang

Corresponding author: Benedict Anchang. E-mail: benedict.anchang@nih.gov

Abstract

Optimizing and benchmarking data reduction methods for dynamic or spatial visualization and interpretation (DSVI) face challenges due to many factors, including data complexity, lack of ground truth, time-dependent metrics, dimensionality bias and different visual mappings of the same data. Current studies often focus on independent static visualization or interpretability metrics that require ground truth. To overcome this limitation, we propose the MIBCOVIS framework, a comprehensive and interpretable benchmarking and computational approach. MIBCOVIS enhances the visualization and interpretability of high-dimensional data without relying on ground truth by integrating five robust metrics, including a novel time-ordered Markov-based structural metric, into a semi-supervised hierarchical Bayesian model. The framework assesses method accuracy and considers interaction effects among metric features. We apply MIBCOVIS using linear and nonlinear dimensionality reduction methods to evaluate optimal DSVI for four distinct dynamic and spatial biological processes captured by three single-cell data modalities: CyTOF, scRNA-seq and CODEX. These data vary in complexity based on feature dimensionality, unknown cell types and dynamic or spatial differences. Unlike traditional single-summary score approaches, MIBCOVIS compares accuracy distributions across methods. Our findings underscore the joint evaluation of visualization and interpretability, rather than relying on separate metrics. We reveal that prioritizing average performance can obscure method feature performance. Additionally, we explore the impact of data complexity on visualization and interpretability. Specifically, we provide optimal parameters and features and recommend methods, like the optimized variational contractive autoencoder, for targeted DSVI for various data complexities. MIBCOVIS shows promise for evaluating dynamic single-cell atlases and spatiotemporal data reduction models.

Keywords: dimensionality data reduction; metric; benchmarking; dynamic and spatial visualization and interpretability; CODEX multiplex imaging; single-cell protein and gene expression analysis

INTRODUCTION

High-dimensional dynamic or spatial data often have intricate patterns and variations that change over time or space presenting challenges for data reduction, visualization and interpretability. Traditional visualization methods, such as scatter plots and heatmaps, become less effective as the data complexity increases, leading to the use of more advanced techniques of dimensionality reduction like t-distributed stochastic neighbor embedding (t-SNE) [1], UMAP [2] and their extensions including other methods [3, 4]. While they are powerful tools, when applied to time-series data, they might produce visualizations that mix time points or fail to capture the temporal order, making it difficult to interpret the dynamics. For example, in a field like single-cell analysis [3, 4], several data reduction methods (DRMs) when applied to the same time course data tend to produce different visual outputs (Figure 1A, Supplementary Figures 1–6), thereby confounding the entire cell labeling process and making it difficult to ensure consistency in interpretation. More so, many end users tend to pick and choose their favorite method in a bias and subjective manner based on their past experiences or familiarity with the method, even if it may not be the most suitable method for the data set, e.g. using a DRM that is optimal for dynamic instead of a static interpretation. Furthermore, in many dynamic scenarios, the true underlying patterns are not always known or easily measurable. For example, in single-cell analysis, the ground truth is often unknown due to the exploration of novel biological systems, rare cell populations, complex cellular heterogeneity, technical limitations, biological variability, dynamic nature of biological processes and ethical and practical constraints. In addition, recently advanced data reduction techniques that integrate both feature data (e.g. protein expression) and spatial information [5] and the huge number of data reduction methodologies [3, 4] that are optimized for linear and nonlinear visualizations (Supplementary Table 1) warrant the need for a unified benchmarking framework that accounts for these variations. Thus, there is a need for a careful evaluation and benchmarking of data reduction models

Komlan Atitey is a Research fellow at the Biostatics and Computational Biology Branch in the National Institute of Environmental Health Sciences. He is developing computational methods to study spatiotemporal biological processes.
Alison Motsinger-Reif is a Senior Investigator at the National Institute of Environmental Health Sciences. Her work focuses on the development and application of methods in statistical genetics and genomics.
Benedict Anchang is a Stadtman principal investigator at National Institute of Environmental Health Sciences and National Cancer Institute leading a computational and systems biology group on multi-scale modeling, visualization and integration of biological processes using single-cell analysis.

Received: August 24, 2023. **Revised:** November 6, 2023. **Accepted:** November 20, 2023
Published by Oxford University Press 2023.

and algorithms used to analyze high-dimensional dynamic and spatial data. Currently, to the best of our knowledge, optimizing or benchmarking the performance of data reduction models for spatiotemporal data complexity, visualization, interpretability and overfitting simultaneously is still very challenging. To address the above concerns, we propose a multivariate interpretable benchmarking and computational framework for optimal visualization and interpretability of high-dimensional stochastic data without ground truth, called MIBCOVIS. This framework can be applied to single-cell and non-single-cell data from various fields.

Benchmarking various data reduction models for optimal dynamic or spatial visualization and interpretability (ODSVI) involves evaluating their performance against a set of predefined criteria or metrics. Some common metrics for ODSVI optimize for clustering accuracy in the presence of ground truth, e.g. normalized mutual information (NMI) [6], the adjusted mutual information (AMI) [7], the Fowlkes–Mallows index (FMI) [8] and the adjusted Rand index (ARI) [9] while others are useful in the absence of ground truth, namely, the silhouette coefficient (SC) [10] and the Davies–Bouldin (DB) [10]. Some metrics optimize for dimensionality reduction accuracy, e.g. Kullback–Leibler divergence [11] and manifold preservation (MP) [12], measure the ability of a visualization technique to accurately represent high-dimensional data in a lower-dimensional space while preserving the underlying structure and patterns of the data. Other key ODSVI metrics target feature importance, e.g. Gini importance score [13] and correlation and causation accuracy, which measure the strength and direction of relationships between different features or variables in the data. To evaluate the performance of DRMs in single-cell analysis, benchmarking studies typically use clustering accuracy metrics [14–16]. However, these metrics are not fully effective at accounting for the observed variability driving the differences in visual geometry and interpretability after data reduction of temporal processes. Moon *et al.* [17] used nonlinear correlations to benchmark the Potential of Heat-diffusion for Affinity-based Transition Embedding (PHATE) method based on diffusion metrics with other DRMs, while Saelens *et al.* [18] used a common trajectory representation and four independent metrics to compare various pseudotime trajectory methods from simulated scRNA-seq data. In a recent study [19], nonlinear correlation and network similarity statistical models were used to benchmark pseudotime and temporal trajectory methods, respectively. However, these benchmarking metrics tend to focus mainly on clustering performance and do not account for uncertainty in the ground truth or the dynamic or spatial stochastic variations at the single-cell level driving cell state transitions. Moreover, current benchmarking frameworks tend to compare these metrics independently producing performance rankings that may be different when the correlations between the metrics are accounted for.

In this study, we aim to formulate a unified computational benchmarking framework that can capture variability at multiscale and multivariate levels, including identifying moderators contributing to ODSVI. We reflect on recent investigations about ODSVI by paying specific attention to ways in which these abstract features can be addressed quantitatively particular for methods that generate time or spatial dependent low-dimensional metric maps including self-organizing maps (SOMs) as outputs. We introduce five metrics to optimize three dependent performance objectives for ODSVI of data reduction models: (1) visual accuracy, (2) data reduction or clustering accuracy and (3) dynamic interpretability accuracy. Figure 1B–F summarizes the five metrics. We define two visual accuracy metrics, the occupation index (OI; Figure 1B) and the uniformity index (UI; Figure 1C), to assess the performance of a DRM for user clarity and persuasiveness [20]. The OI quantifies the shape of the projected data in the low-dimensional space by assessing coverage of the projection space in relation to the coverage in the high-dimensional space (Figure 1B). The UI describes the uniform spread of data points in the low-dimensional space, assessing both uniformity and orthogonality (Figure 1C), two properties associated with many data reduction models [21]. To quantify data reduction accuracy (clustering accuracy for single-cell analysis), we introduce a third metric, the gradient boosting classifier index (GI), which uses a gradient boosting classifier [22] to assess the sensitivity of clusters in the projected space and compare its performance to some of the existing metrics listed above. To further evaluate the degree of separation or distinction between these clusters after data reduction, we introduce the fourth metric, the separability index (SI), commonly used in the fields of pattern recognition, machine learning and data mining to assess the quality of the feature representation or feature selection for classification tasks [23]. Various researchers have used different formulations of class separability as a benchmark for evaluating the performance of DRM in single-cell analysis [2, 24]. We classify the SI as an interpretability accuracy metric and use it to estimate the average number of instances in a data set that have a nearest neighbor with the same cluster label [25]. Furthermore, to assess the dynamic or causation relationships produced by a given DRM, we introduce the fifth interpretability accuracy metric, the time order structure index (TI; Figures 1F and 2), which uses Markov chains with community structure to assess the time dependency of ordered points (cells for single-cell data) in the low-dimensional projected space. Figure 1G summarizes the five metrics, objectives and key parameters to assess the performance of dimensionality reduction methods (DRMs) for ODSVI.

MIBCOVIS aims to integrate correlated features within a statistical framework for optimal evaluation and benchmarking. Figure 1G and H show the input–output features of MIBCOVIS, which takes projected output data from linear, nonlinear and neural network DRMs as input. Key parameters directly associated with the performance of each of the five metrics described above are selected, and then the visualization and interpretability accuracy of each method is ranked using high-density posterior interval estimates from Bayesian conditional regression modeling (Figure 4 and see Method section in Supplementary Text). To account for uncertainty in the ground truth, MIBCOVIS is applied to a semi-supervised learning (SSL) framework, which uses both labeled and unlabeled data sets to improve the classification accuracy and the benchmarking pipeline. MIBCOVIS is used to evaluate the performance of six major DRMs as a representative sample of all linear, nonlinear and neural network method space applied on three distinct data modalities; protein (CyTOF), gene expression (scRNA-seq) and imaging (CODEX) to study four dynamics (spatial) biological processes involved in normal development: epithelial-to-mesenchymal transition (EMT) plasticity, spermatogenesis, stem cell reprogramming and small to large intestinal cell–cell interactions. We investigate the effect of increasing data complexity on visualization performance, defined in terms of feature dimension, unknown cell types and dynamic or spatial interaction differences in biological processes. Using MIBCOVIS, we demonstrate that no current method optimizes the proposed features for ODSVI jointly and show that ranking of DRM under independence performance is different from ranking of methods under joint performance. We suggest oVAE, a joint variational and contractive autoencoder (AE; Supplementary Text 2),

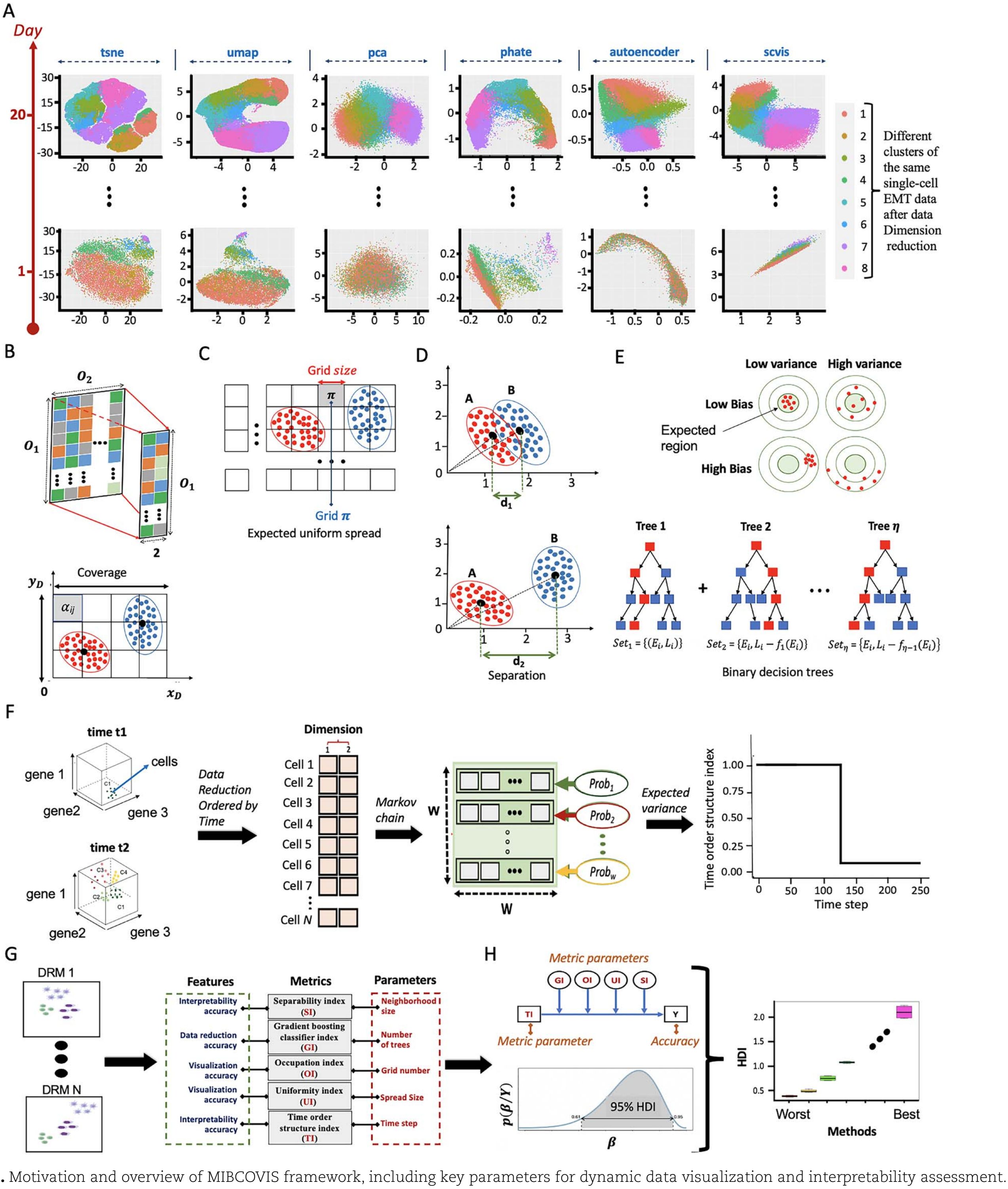
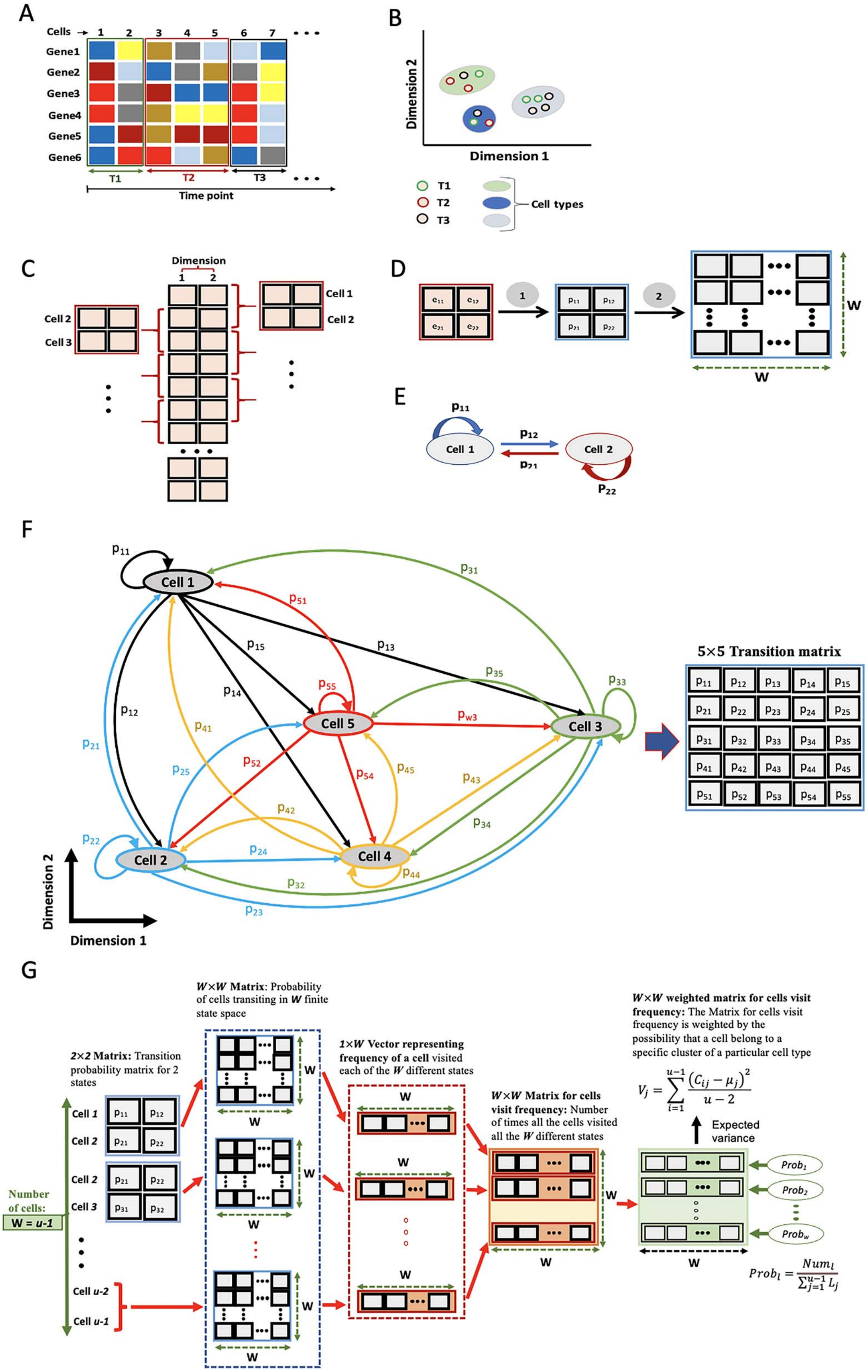


Figure 1. Motivation and overview of MIBCOVIS framework, including key parameters for dynamic data visualization and interpretability assessment. (**A**) Dimensional reduction techniques employed to assess cellular evolution dynamics in a CyTOF EMT single-cell data set spanning 20 days. The data are visualized using nonlinear methods (t-SNE, UMAP, PHATE), a linear approach (PCA) and neural network models (AE, SCVIS). The resulting projections exhibit varying shapes and scales, ranging from the most compact (PHATE) to the most dispersed (t-SNE), potentially introducing challenges in visualization and interpretation. (**B**) The OI quantifies the coverage of the projected 2D space and how well it is occupied by cells. (**C**) The uniformity (UI) of the projected 2D data space is estimated by dividing it into π bins to evaluate the distribution of data points across the space. (**D**) The SI measures the degree of separation between classes in a 2D data set based on the distance between their centroids. Comparing d_1(top) and d_2(bottom), shows that (A) and (B) are more separated in the second case than the first case due to $d_2 > d_1$. (**E**) The GI uses gradient boosting classification trees to minimize bias and variance in order to predict the label of each point and count the fraction of correct predictions. The center of the bull's-eye target region represents a model with perfect sensitivity. As we move away from the bull's-eye, the predictions get worse. (**F**) TI uses a weighted Markov process to convert an $N \times 2$ ordered data reduced state space from time course data into a $W \times W$ stationary state space and estimates TI scores using expected variances between time steps derived from the MCMC chain. (**G**) MIBCOVIS takes the projected output data from DRMs and selects key parameters directly associated with the performance of five data reduction, visualization and interpretability metrics. (**H**) A multivariate conditional effect model used for the evaluation performance of DRM condition on the key TI metric moderated by the effects of other metric features. Given the accuracy data Y, MIBCOVIS ranks the performance of DRMs by quantifying the HDI of the posterior probability density function $p(\beta/Y)$ for model parameters β.



p11	p12	p13	p14	p15
p21	p22	p23	p24	p25
p31	p32	p33	p34	p35
p41	p42	p43	p44	p45
p51	p52	p53	p54	p55

2×2 Matrix: Transition probability matrix for 2 states

W*×*W* Matrix**: Probability of cells transiting in ***W finite state space

1*×*W* Vector representing frequency of a cell** visited each of the ***W different states

W*×*W* Matrix for cells visit frequency:** Number of times all the cells visited all the ***W different states

***W*×*W* weighted matrix for cells visit frequency:** The Matrix for cells visit frequency is weighted by the possibility that a cell belong to a specific cluster of a particular cell type

$$V_j = \sum_{i=1}^{u-1} \frac{\left(C_{ij} - \mu_j\right)^2}{u-2}$$

$$Prob_l = \frac{Num_l}{\sum_{j=1}^{u-1} L_j}$$

Figure 2. TI for identifying changes in sub-populations over time. (**A**) Single-cell RNA-seq gene expression data collected at multiple time points (T1–T3), is clustered and projected to a low dimensional space. (**B**) Expected dynamic relationship of cells in the projected space. (**C**) A partitioning of the large non-square projected data set into consecutive 2 × 2 square matrices to capture transition between time points. (**D**) Large projected data set is partitioned into $u - 1$ consecutive 2 × 2 matrices to capture transition between time points resulting in a W × Wsquare transition probability matrix. (**E**) Two cell–state Markov chain model and its transition probabilities. (**F**) A Markov chain with five cells at five different states with selected states transition. The state transition probability matrix of the Markov chain gives the probabilities of transitioning from one state to another in a single time unit. (**G**) Weighted transition probability matrix for the Markov chain based on the distribution of cell states ($Prob_w$) used to estimate the expected variance (V_j).

as an optimal benchmarking method when the user is unsure which ODSVI feature to target. This study offers a robust metric set and an unbiased benchmarking framework for ODSVI of relationships in spatio-temporal metric maps after data reduction.

RESULTS

We present mathematical definitions of the metrics used for ODSVI and highlight the DRMs selected by MIBCOVIS to account for variations due to dimensionality reduction. The variability resulting from different visualization tools applied to a dynamic biological process like EMT motivates the need for a spatiotemporal benchmarking framework like MIBCOVIS to determine the best method for visual biological interpretability. We then present a supervised analysis of MIBCOVIS applied to four different biological data sets with discrete labels for each cell representing the ground truth, followed by a summary of the results from the semi-supervised analysis.

MIBCOVIS uses OI and UI to assess visual accuracy

A DRM that maximizes the coverage of projected space can reveal patterns, clusters and outliers in high-dimensional data (Figure 1B). To evaluate the coverage of the data distribution in the projected space, we used the OI, which is a proportion of the total surface area indexed by the underlying cells to the projected space area [26]. The OI is described mathematically by the weighted product of the coverage of the projected and the high-dimensional spaces given as

$$OI = \kappa * Cov_p . Cov_h$$

where Cov_p corresponds to the coverage in the projected space, expressed in terms of the number of grid tiles (see Method section in Suplementary text), Cov_h represents coverage of the high-dimensional space, determined by the total variance (see Method section in Supplementary text) and κ is a weighting parameter proportional to the surface area per unit summary coverage defined by Kufer [27]. The choice of surface area per unit summary coverage is a challenging task for different applications, as it depends on the specific goals and context of visualization. By minimizing projected unused space and overlapping summary coverages, a lower surface area per unit summary coverage indicates that the visualization or summary captures a lot of detail in a small area, potentially resulting in improvement of readability and usability of a visualization map. A smaller grid size generally corresponds to higher detail in a smaller area, which aligns with a lower surface area per unit summary coverage because more detail is captured in a given unit of summary coverage. Given that the surface area per unit summary coverage varies with the number of grids, which is constant for this study as well as the coverage of the high-dimensional space (Cov_h) for all the DRMs, without loss of generality, we set $\kappa = 0.33$ as estimated in Kufer [27] from several applications. We vary the number of grid from 1 to 250 to generate a univariate distribution for OI.

MIBCOVIS assesses the uniformity of projected data using the UI, which is determined by the spread of the projected data. To estimate the UI, we use the goodness fit of the Pearson chi-square (χ^2) test of uniformity of the data distribution. The projected space is divided into π grids with Π the total number of grids (Figure 1C), determined by the grid size, defined as the length of each side of the square grid. We count the number of points in each grid and use the relation between grid size and total number of grids (see Method section) to define the UI.

$$UI = 1 - \frac{\Upsilon}{\Pi}$$

where Υ represents the test statistic under the null hypothesis that the points are uniformly distributed.

The UI distribution with values in the range of 0 to 1 is derived from varying the grid size from 1 to 250.

MIBCOVIS uses GI to assess accuracy of DRM

MIBCOVIS uses the GI, which depends on gradient boosting classifier algorithm, a machine learning algorithm for classification. It iteratively trains weak learners on the residuals of the previous weak learners to minimize bias and variance (Figure 1E, Methods). Observations that are most difficult to classify correctly are assigned higher weights in each iteration. In practice, we fit a gradient boosting classifier on the low-dimensional embedding of the data, predict the label of each point and count the fraction of correct predictions. The xgboost R package [28] is used to generate 250 GI values by varying the number of trees, a parameter that controls the number of weak learners. A larger number of trees are expected to improve the accuracy of the model.

MIBCOVIS uses SI and TI to assess interpretability accuracy

MIBCOVIS uses the SI to evaluate the ability of a data reduction or feature selection method to discriminate between different groups or classes in a data set (Figure 1D). The index ranges from 0 to 1, with higher values indicating better separability. An SI of 1 means perfect separation, while a value of 0 indicates complete overlap or confusion between groups. The SI is calculated as the average number of instances (ρ_k) in the projected data set such that all cells have a nearest neighbor with the same cluster label k (see Method section).

$$SI = \frac{1}{\mid \Psi_k \mid} \sum_{k=1}^{|\Psi_k|} \rho_k$$

where $\mid \Psi_k \mid$ refers to the number of distinct classes. We vary the numbers of nearest neighbors for each point from 1 to 250 to estimate the distribution for SI.

MIBCOVIS uses the TI to assess the accuracy of interpreting the correlation and causation of a visualization technique. It applies multi-state Markov processes to evaluate the time or spatial dependency of cells in a temporal reduced space (2D; Figure 2A and B). A Markov chain model with community structure is used to describe the possible states of individual cells during consecutive time steps in the projected space (Figure 2E and F). The framework involves six main steps, including partitioning the reduced data into consecutive submatrices (Figure 2C), estimating cell state probabilities (Figure 2D), computing expected transitions (Figure 2E), generating a W × W transition matrix (Figure 2D), weighting the matrix (Figure 2G) and computing the cumulative sum and column mean μ_j to estimate the expected variance (V_j) (Figure 2G).

Markov processes assume stationarity, which simplifies analysis but may not hold for all biological processes. The TI leverages on this non-stationarity to assess how a trajectory method accounts for time ordering variations and how the variance converges to the equilibrium. A faster convergence indicates less time dependency in the metric map.

MIBCOVIS selects six DRMs to account for variations of all DRMs

We evaluated over 20 data reduction tools [3, 4] for analyzing and visualizing large biological data in single-cell analysis, categorizing them into linear, nonlinear and neural network methods in Supplementary Table 1 and summarizing their advantages and computational limitations in Supplementary Text 1. MIBCOVIS uses widely accepted linear PCA, popular nonlinear t-SNE and UMAP [29], general neural network AE [30] for denoising and imputation of missing values, variational AE SCVIS [31] for scRNA-seq data and diffusion theory PHATE. These methods account for variations in all DRM.

MIBCOVIS uses three single-cell time course data sets and one single-cell spatially ordered imaging data characterizing low, medium and high data complexities for benchmarking

We evaluate MIBCOVIS using three data sets to investigate dynamic biological processes: EMT [32] (Data set 1), chemically induced pluripotent stem cells (iPSCs) [33] (Data set 2) and spermatogenesis [34] (Data set 3). For Data set 1, 96 000 single-cell data were collected for 20 days after *in vitro* stimulation of lung cancer cell lines with TGF-β for 10 days, followed by TGF-β withdrawal for another subsequent 10 days. The data set consists of eight EMT-MET states: E1 (2), E2 (1), E3 (4), pEMT1(3), pEMT2 (5), pEMT3 (6), M (7) and pMET (8) validated using six canonical EMT markers. We define this data set as 'low complexity'. Data set 2 comprises 50 000 cells collected from 12 time points over 21 days, investigating pluripotency using chemically induced cellular reprogramming. Each cell is associated with one of 19 labeled clusters derived from 102 significant genes [33]. We categorize Data set 2 as 'medium complexity'. Data set 3 includes approximately 110 000 cells from 16 postnatal stages during spermatogenesis. These cells are associated with 29 clusters derived from 174 significant genes driving spermatogenesis [32]. We classify this data set as 'highly complexed'.

We also evaluate MIBCOVIS on normal patient CODEX multiplex imaging (CMI) protein data of small bowel and large intestine spanning the duodenum to the sigmoid colon comprising of eight individual tissue regions [5] covering about 300 000 cells, characterizing 25 cell types, 20 multicellular neighborhoods and 10 communities of neighborhoods. Instead of dynamic cellular relationship, we use the spatial ordering of regions beginning from duodenum, proximal jejunum, mid-jejunum, ileum, ascending, transverse, descending to descending-sigmoid characterized in Hickey *et al.* [5] to evaluate 'functional' cellular relationship from the small intestine to the large intestine. Given the joint high-dimensional and very high number of neighborhoods by cell-type interactions, we consider this data 'highly complexed'.

GI is a robust ODSVI clustering accuracy metric compared to traditional metrics

Single-cell analysis results in noisy cell labeling of the data, which can impact the sensitivity of downstream analysis. Thus, we introduced the GI as a metric for assessing classification accuracy. It predicts the label of each data point in the low-dimensional space and quantifies the fraction of accurate predictions, providing a summary score of model performance. However, since several metrics for ODSVI have been published to evaluate clustering accuracy in the presence of ground truth (including NMI [6], AMI [7], FMI [8] and ARI [9]) as well as in its absence (SC [10], DB [10] and MP [12]), we conducted a comparison of some of these metrics. We compared a supervised GI (sGI), which relies on cluster labels, with a semi-supervised GI (uGI) that uses both labeled and unlabeled data in addition to the unsupervised metrics, including SC, DB and MP, to evaluate the clustering accuracy of DRMs. Without loss of generality, we select the t-SNE dimension reduction method due to the fact that it yielded the highest supervised GI score when compared to other DRM. For the unsupervised clustering accuracy metrics, we assume a fixed number of expected clusters from prior knowledge. In a nutshell, we evaluate the performance of t-SNE by varying the following specific parameters: (1) the number of trees for the GI, (2) the maximum number of iterations for the SC, (3) the power of dispersion measure of a cluster for DB and (4) the number of nearest neighbors for MP.

We vary the number of iterations in SC resulting in 250 different SC scores when applied to each of the four data sets (Figure 3). In the same way, to generate 250 different scores for DB, we vary the parameter of power of the dispersion measure of a cluster in the range of 1–250 as well as the parameter of number of nearest neighbors in MP in the range 2–251, generating 250 values for each data set.

We next use a two-sample paired *t*-test to evaluate the statistical difference between the five metrics across all data complexity. The results of the investigation revealed a significant difference among the five metrics, with small *P*-values typically < 0.05 (Figure 3). Furthermore, comparing each metric's median value relative to the overall median (dotted line, Figure 3) across data complexity, demonstrates the uGI metric performance exceeds the overall median by 13% making it suitable clustering accuracy metric for ODSVI.

t-SNE or UMAP produces different visualization and interpretable dynamics during EMT and MET

The high-dimensional nature of single-cell data including the observed different visual outputs (Figure 1A, Supplementary Figures 1–6) for various DRMs challenges our ability to properly interpret dynamic interactions after data reduction. For example, in the study by Karacosta *et al.*, they combined t-SNE and an artificial neural network to visualize the plasticity of eight states of EMT [32], namely, three epithelial (E1–E3), three partial EMT (pEMT1–3), one mesenchymal (M) and one partial MET states (pMET). Interpreting the dynamics visually between these eight EMT states could be different as shown by the different Voronoi partitions [32] of t-SNE and Uniform Manifold Approximation and Projection (UMAP) maps (Supplementary Figure 7). The different relative positions of M and pMET states by both methods confounds the visual interpretability of EMT-MET trajectories. One of the goals of this study is to use a metric like TI and other features to evaluate which of the methods provides a better visualization of the time-dependent EMT transitions.

MIBCOVIS-supervised framework

MIBCOVIS evaluates ODSVI performance under a supervised setting by using 2D dimensionality reduction metric maps and ground truth-labeled data. Since the GI method has been shown to yield lower prediction errors and higher accuracy results compared to other metrics from Figure 3, we next transform it into a global accuracy measure for evaluation of DRMs by the MIBCOVIS Bayesian framework. This new variable enables us to assess the accuracy of a DRM condition on all metric features. By relying on the 250 scores of each of the six DRMs (t-SNE, UMAP, PCA, AE, SCVIS and PHATE), we generate two sets, one for minimum scores (S_{min}= {$\mathbf{min}_{tsne}$, $\mathbf{min}_{umap}$, $\mathbf{min}_{pca}$, $\mathbf{min}_{ae}$, $\mathbf{min}_{scvis}$, $\mathbf{min}_{phate}$}) and the other for maximum scores (S_{max}= {$\mathbf{max}_{tsne}$,

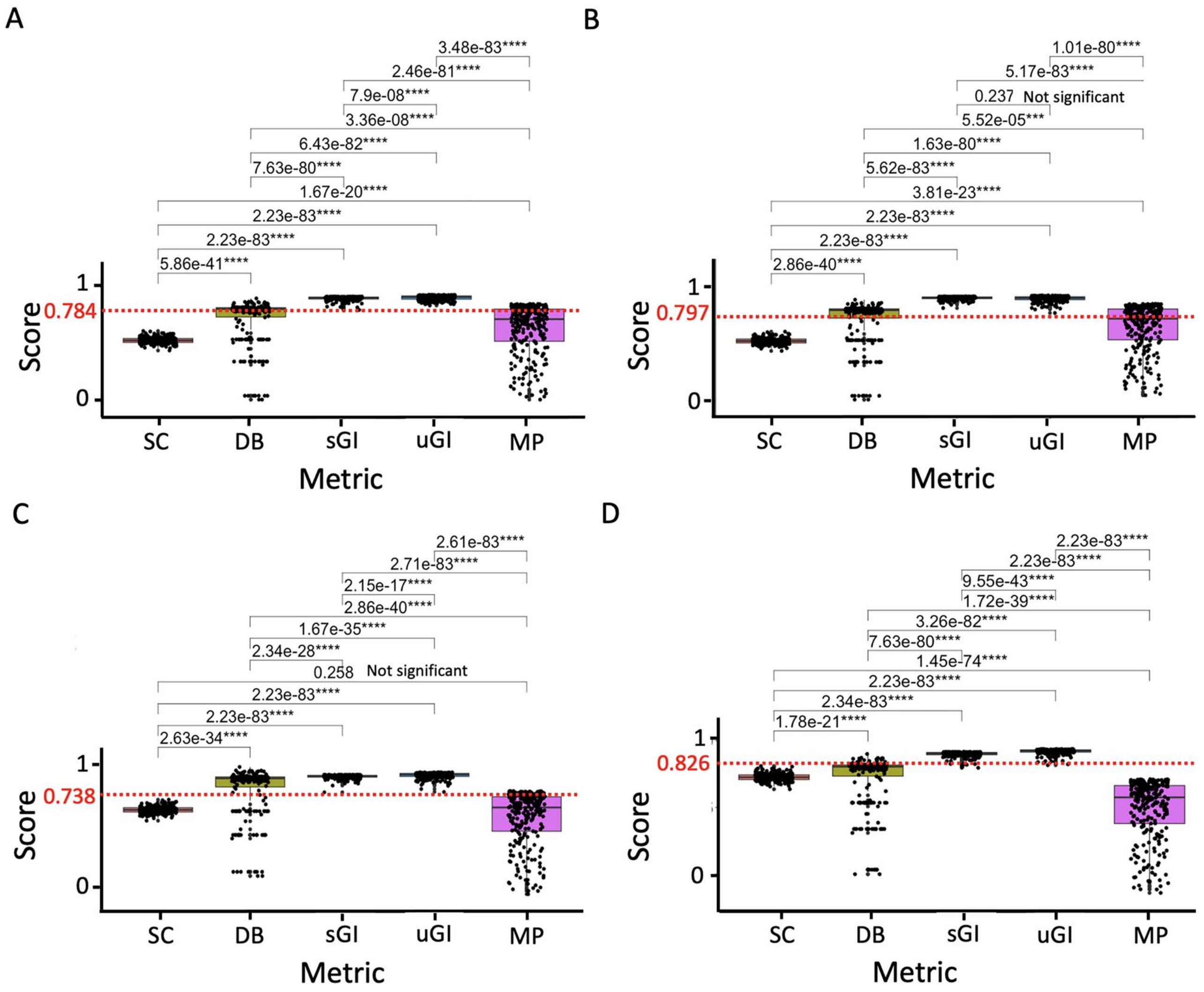


Figure 3. MIBCOVIS uses the GI as a robust clustering accuracy metric for ODSVI compared to traditional metrics. (**A**) Box plots with associated significant differential *P*-values comparing the t-SNE performance scores between supervised GI (sGI), semi-supervised GI (uGI) and the unsupervised metrics; SC, DB and MP for the EMT single-cell data. The median normalized score of uGI and sGI surpass not only the overall median but also the median of SI, DB and MP. The same inference can be made for (**B**) IPSC data reduction analysis using t-SNE, (**C**) spermatogenesis single-cell data analysis using t-SNE and (**D**) CODEX multiplex imaging single-cell data reduction analysis using t-SNE.

$\mathbf{max}_{umap}$, $\mathbf{max}_{pca}$, $\mathbf{max}_{ae}$, $\mathbf{max}_{scvis}$, $\mathbf{max}_{phate}\}$). We set $\mathbf{Min}_{score} = \mathbf{min}\,(S_{min})$ and $\mathbf{Max}_{score} = \mathbf{max}\,(S_{min})$, ensuring that the score range of the six DRM falls within the range $[\mathbf{Min}_{score}, \mathbf{Max}_{score}]$. We then generate 250 uniformly distributed accuracy scores, denoted as Y, within the range $[\mathbf{Min}_{score}, \mathbf{Max}_{score}]$. It leads to obtaining Y_1, Y_2, Y_3 and Y_4 (Figure 4C), defining the accuracy variables after reducing the EMT, IPSC, spermatogenesis and CMI single-cell data.

Considering the accuracy variable (Y_i), as a response variable, MIBCOVIS combines metric-based features in a multivariate hierarchical Bayesian regression framework to generate conditional posterior distributions of accuracy (Figure 4A–C) and ranks the performance of methods using boxplots of high-density intervals (HDIs) of conditional regression coefficients. Figure 4A and B summarizes six DRMs and five metric set used for model building. Figure 4B highlights key parameters; neighborhood size, number of trees, grid number, grid size and time steps associated with each metric, used to investigate the conditional effects of the metric accuracy (Supplementary Figure 8A). The third panel (Figure 4C) includes a two-step sequential analysis to benchmark dimensional reduction methods. In the first step, the Spearman correlation coefficient is used to assess the correlation between features from Panel B and method accuracy (Supplementary Figure 8B). In the second step, Bayesian multilevel modeling is used to evaluate the conditional effect of metric features and method accuracy outcome for dimensionality reduction (see Methods).

MIBCOVIS identifies diverse metric parameter regions for optimal data reduction visualization

We model the variability of five metrics across six DRMs on the CyTOF EMT data set (Figure 5). t-SNE outperforms UMAP, SCVIS, Autoencoder, PHATE and PCA in terms of SI (Figure 4A), indicating its ability to better separate the EMT biological process phenotypes with few outliers. We obtain similar results when assessing DRM on Data sets 2–4 (Supplementary Figure 9A). Optimal SI values can be achieved with low neighborhood size values in the range [1, 6] for all six DRMs.

To evaluate classification bias and variance, we use the GI across all three data sets and find that t-SNE outperforms UMAP and SCVIS (Figure 5B). Using a single tree displays the lowest performance of all DRM, while the performance of all DRM grows

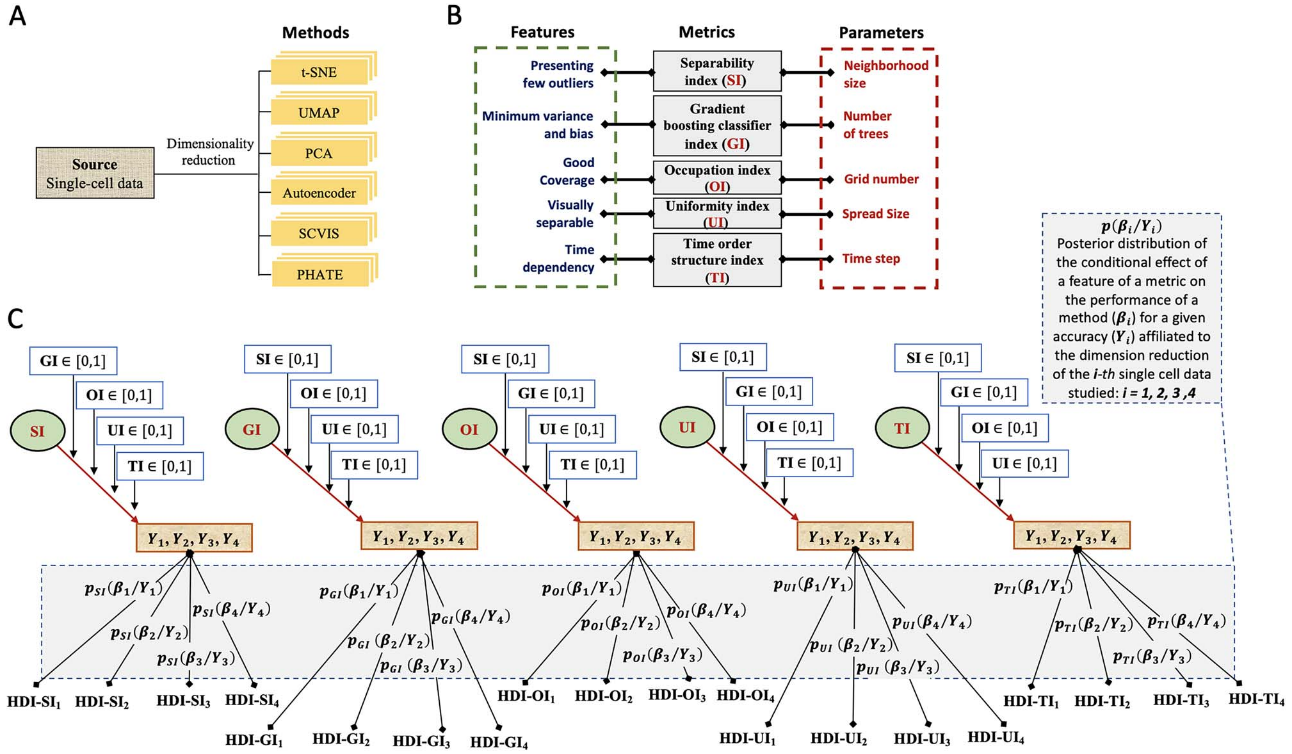

A
Methods
t-SNE
UMAP
PCA
Autoencoder
SCVIS
PHATE
Source Single-cell data
Dimensionality reduction
B
Features
Metrics
Parameters
Presenting few outliers
Minimum variance and bias
Good Coverage
Visually separable
Time dependency
Separability index (SI)
Gradient boosting classifier index (GI)
Occupation index (OI)
Uniformity index (UI)
Time order structure index (TI)
Neighborhood size
Number of trees
Grid number
Spread Size
Time step
$p(\beta_i/Y_i)$ Posterior distribution of the conditional effect of a feature of a metric on the performance of a method (β_i) for a given accuracy (Y_i) affiliated to the dimension reduction of the *i-th* single cell data studied: *i = 1, 2, 3 ,4*
C
SI, GI, OI, UI, TI
GI ∈ [0,1], OI ∈ [0,1], UI ∈ [0,1], TI ∈ [0,1], SI ∈ [0,1]
Y_1, Y_2, Y_3, Y_4
$p_{SI}(\beta_1/Y_1)$ $p_{SI}(\beta_2/Y_2)$ $p_{SI}(\beta_3/Y_3)$ $p_{SI}(\beta_4/Y_4)$
$p_{GI}(\beta_1/Y_1)$ $p_{GI}(\beta_2/Y_2)$ $p_{GI}(\beta_3/Y_3)$ $p_{GI}(\beta_4/Y_4)$
$p_{OI}(\beta_1/Y_1)$ $p_{OI}(\beta_2/Y_2)$ $p_{OI}(\beta_3/Y_3)$ $p_{OI}(\beta_4/Y_4)$
$p_{UI}(\beta_1/Y_1)$ $p_{UI}(\beta_2/Y_2)$ $p_{UI}(\beta_3/Y_3)$ $p_{UI}(\beta_4/Y_4)$
$p_{TI}(\beta_1/Y_1)$ $p_{TI}(\beta_2/Y_2)$ $p_{TI}(\beta_3/Y_3)$ $p_{TI}(\beta_4/Y_4)$
HDI-SI$_1$ HDI-SI$_2$ HDI-SI$_3$ HDI-SI$_4$
HDI-GI$_1$ HDI-GI$_2$ HDI-GI$_3$ HDI-GI$_4$
HDI-OI$_1$ HDI-OI$_2$ HDI-OI$_3$ HDI-OI$_4$
HDI-UI$_1$ HDI-UI$_2$ HDI-UI$_3$ HDI-UI$_4$
HDI-TI$_1$ HDI-TI$_2$ HDI-TI$_3$ HDI-TI$_4$

Figure 4. MIBCOVIS model framework optimizes visualization and interpretability. (**A**) MIBCOVIS uses six DRMs spanning three different model classes (linear, nonlinear and neural network). These include t-SNE, PCA, UMAP, AE, SCVIS and PHATE. (**B**) MIBCOVIS uses five quantitative metrics—SI, GI, OI, UI and TI to evaluate and compare the performance accuracy of the DRMs. Each of the five metrics depend on an independent parameter of length 250. (**C**) MIBCOVIS uses the GI to derive a direct overall measure of performance accuracy Y of a given DRM. MIBCOVIS models the relationship between multiple independent (metric features) and the dependent variable Y_i. Using lognormal priors, MIBCOVIS computes the posterior distribution of the conditional accuracy effect of a metric feature, assuming a beta distribution of accuracy scores. It next analyzes the posterior distribution using the MCMC sample and summarize the distributions using a 95% HDI.

exponentially with 2–25 trees. Similar results are obtained for scRNA-seq Data sets 2 and 3 and CODEX Data set 4 (Supplementary Figure 9B).

To evaluate optimal parameter values for good coverage of the projected space by DRM, we assessed the performance of the six DRMs with varying grid number values (Figure 5C). We found that all six DRMs perform best with a relatively small grid number, also when applied to Data sets 2–4 (Supplementary Figure 9C). Additionally, for Data set 1 with eight EMT states and CODEX Data set 4, AE, PCA and t-SNE show similar high performance, but rankings change for Data sets 2 and 3, where t-SNE, SCVIS and PCA exhibit the highest performance.

The UI measures a DRM's ability to uniformly spread the projected data in the low-dimensional space. All the six DRMs perform well for small initial grid sizes (Figure 4D and Supplementary Figure 9D), but their performance decreases as the grid size increases. PHATE shows the biggest drop in performance among all six DRMs for Data set 1, with similar results for Data sets 2–4.

Figure 5E summarizes the relationship between the TI and time steps for various DRMs. No method shows superior evolution dynamics for the various cell types, with all six methods taking approximately 48 time steps to achieve stationarity of state transitions. However, within a small number of time steps (48–75 units), DRMs exhibit variability in performance relative to the time ordering of single-cell data in the low-dimensional space. For example, PHATE exhibits the highest TI performance among the six DRMs, especially after reducing the CODEX data (Supplementary Figure 9E). It also shows a slightly better performance on the CyTOF Data set 1 (Figure 5E) and for Data sets 2 and 3 (Supplementary Figure 9E).

MIBCOVIS exhibits strong correlation between metric features

Using the GI distributions, the following [S_{min} and S_{max}] ranges of accuracy scores $Y_1 = [0.73, 0.98]$, $Y_2 = [0.62, 0.93]$, $Y_3 = [0.57, 0.75]$ and $Y_4 = [0.54, 0.82]$ for various DRMs were estimated. We observe that an increase in data complexity is associated with a decrease in the accuracy of DRM ($Y_1 > Y_2 > Y_3$) or ($Y_1 > Y_2 > Y_4$). Using Spearman's rank correlation coefficient [35], we also observe a strong correlation between these variables (Supplementary Figures 10–12). We use Bayesian modeling to study how the effect of one metric on visualization and interpretability accuracy is moderated by the effect of the others (Supplementary Figure 8C).

The conditional effect analysis using MIBCOVIS posterior distributions suggests that no single method optimizes all features necessary for optimal visualization and interpretability

We investigate the conditional performance accuracy of DRM across different metrics using a joint hierarchical regression Bayesian framework [36]. The model variables include five independent variables (SI, GI, OI, UI and TI) and one dependent variable (accuracy scores Y_1, Y_2, Y_3 and Y_4) for evaluating DRMs for different data sets. We use the 'brms' R package to compute the posterior distribution of the conditional effect of a metric feature on the accuracy performance of a DRM by combining a beta model with a logistic link (Supplementary Text 4). We

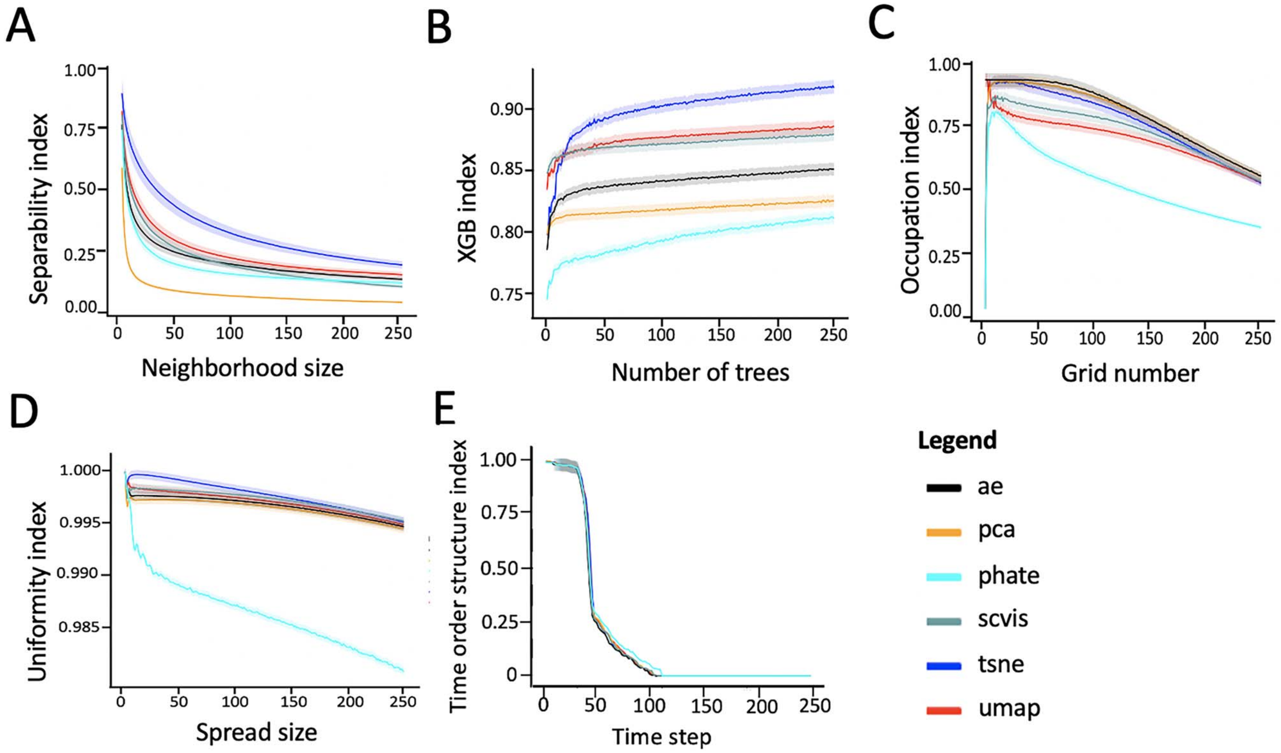


Figure 5. MIBCOVIS-independent metric analysis for EMT data using six DRMs: five metric functions with estimated standard error corresponding to the 95% confidence interval are presented, showing the performance accuracy under varying parameters of neighborhood size, number of trees, grid number, grid size and time steps. (**A**) For all the six methods, the optimal range of neighborhood size (SI) with fewest outliers in the low-dimensional embedding is [1, 6]. (**B**) The optimal number of trees for minimizing classification bias and variance after data reduction (GI) is in the range [50, 100]. (**C**) The coverage (OI) of the low dimensional space is maximized when the grid number is in the range [5, 20]. (**D**) The UI decreases slowly with the increase of grid size, and all the methods maximize uniformity of projected space when the grid size is in the range [2, 5]. (**E**) All DRMs except PHATE perform similarly in terms of TI, which shows a slight increase in performance between the time steps 48 and 75.

estimate the posterior distribution using the metropolis algorithm of Markov chain Monte Carlo (MCMC) [37]. We sample 2000 points from the posterior distribution using four chains with different initial states and examine the dependencies of the MCMC chains and accuracy in terms of (1) the overlap of the density of the four chains and (2) the clumpiness of the chains measured by the autocorrelation of chain values. The analysis results show a good mixture of the density of the four chains and the model accuracy (Supplementary Figures 13–17). We construct the 95% high-density interval (HDI) [38] of the posterior distribution (Supplementary Figures 18–32, Supplementary Text 5) to illustrate the distribution of the conditional effect of a targeted feature of a metric, taking into account the moderation effects of other features on the performance accuracy of a DRM (Figure 6, Supplementary Figures 33–50). We use the square of absolute values of the standardized regression coefficients [39] to generate box plots quantifying the relative performance of each metric. In a nutshell, comparing the MIBCOVIS posterior distributions on the log scale shows that no single method optimizes all features driving optimal visualization and interpretability (Figure 6).

We compute the overall average performance score per metric by normalizing the HDI scores and averaging the normalized values across DRM and data sets. The red dashed lines in Figure 7A–E show the average performance scores for OI, GI, SI, UI and TI, respectively. For instance, based on OI (Figures 6A and 7A), t-SNE, AE and SCVIS exhibit the highest median performance across all data sets and data complexities, indicating good coverage of the 2D space. Moreover, SCVIS and t-SNE outperform the average performance score (0.165) for OVI by providing maximum coverage of projected data in 2D space (Figure 7A).

Regarding the GI performance metric (Figure 6B), t-SNE, UMAP and AE demonstrate the highest performance on average for high-dimensional data sets, while t-SNE displays good performance for low-dimensional data sets. UMAP presents good performance for high-dimensional data sets. Model averaging (0.149) suggests that AE, t-SNE and UMAP are the best methods for ODSVI while minimizing classification bias and variance (Figure 7B).

To summarize for SI, while PHATE performs the best for low-complexity data sets, UMAP operates the best for highly complexed data sets (Figure 6C). Across all data modalities, for SI, UMAP, PHATE and PCA has the best performance in visualizing data with few outliers (Figure 7C). For UI, PCA consistently performs the best across medium and high data sets resulting in a superior above-average performance (Figure 7D), while UMAP performs better with low data complexity. For TI, PHATE consistently outperforms most methods.

Optimized VAE as a benchmarking tool for MIBCOVIS

Given that none of the six DRMs is superior simultaneously for all visualization features (Figures 6 and 7), we propose using an oVAE as a benchmark model for ODSVI (Supplementary Text 2 and Supplementary Figure 51). oVAE combines AEs, variational AEs [40] and contractive AEs [41] to improve data projection and reduce overfitting while providing a closed framework for

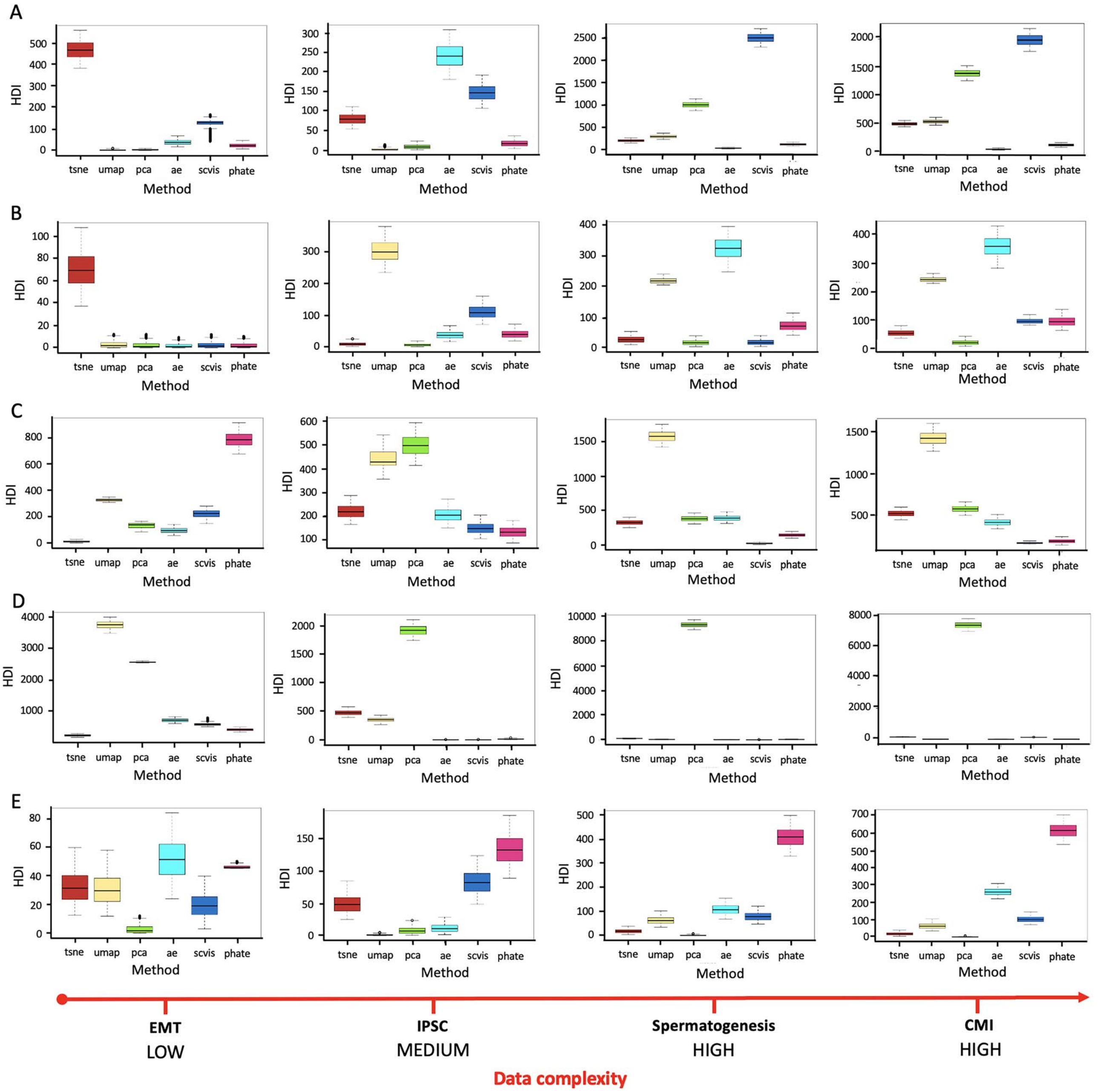


Figure 6. Boxplots comparing the performance accuracy distributions of six DRMs using HDI of conditional metric effects on EMT, IPSC, spermatogenesis and CODEX intestine data, in order of increasing complexity. (**A**) OI performance varies between 0.1 and 600 for EMT data set, with t-SNE having higher median values than UMAP, PCA, PHATE, AE and SCVIS. For the IPSC data set, HDI varies from 0.2 to 350 with AE and SCVIS outperforming other methods. The HDI performance for spermatogenesis varies from 0.1 to 2800 with SCVIS showing highest median values. (**B**) GI HDI varies from 1 to 890 with t-SNE having higher median values than AE, PCA, PHATE and SCVIS for EMT data. For IPSC data, UMAP has a higher median value than SCVIS, PHATE, AE, PCA and t-SNE. HDI range is 2–395 for spermatogenesis data with AE and UMAP showing higher median values. (**C**) SI performance for EMT data set varies from 0.4 to 900 with PHATE outperforming other methods. For IPSC data set, SI performance varies from 95 to 600 with UMAP and PCA having the highest median value. The SI HDI for spermatogenesis varies from 1 to 1900 with UMAP having higher medians. (**D**) MIBCOVIS UI performance for EMT data set varies from 100 to 4000 with UMAP and PCA outperforming other methods. For IPSC data set, HDI values range from 1 to 2300 with PCA having best performance. HDI for spermatogenesis ranges from 1 to 9900 with PCA showing higher median values. (**E**) HDI for TI ranges from 0.20 to 82 with AE and PHATE surpassing other methods for EMT data set. For IPSC data set, HDI varies from 0.035 to 200 with SCVIS and PHATE having the highest median. PHATE has superior TI performance for both spermatogenesis and CMI data with HDI highest median values > 420.

model predictions. oVAE uses a weighted likelihood for projection (see Supplementary Text 2). We assess oVAE's performance using MIBCOVIS and compute an average performance score for ODSVI (SGVI) using normalized average scores across data sets and methods including oVAE (0.172) or without oVAE (0.160) represented by black dashed lines in Figure 8A and B. Compared to non-VAE models (t-SNE, UMAP, PCA, AE, SCVIS and PHATE), oVAE performs better on average (Figure 8B and Supplementary Figures 52–56).

MIBCOVIS semi-supervised analysis

We extended the MIBCOVIS framework (see Supplementary Text 3) to account for uncertainty in data labeling and compared the performance of DRM using both labeled and unlabeled data

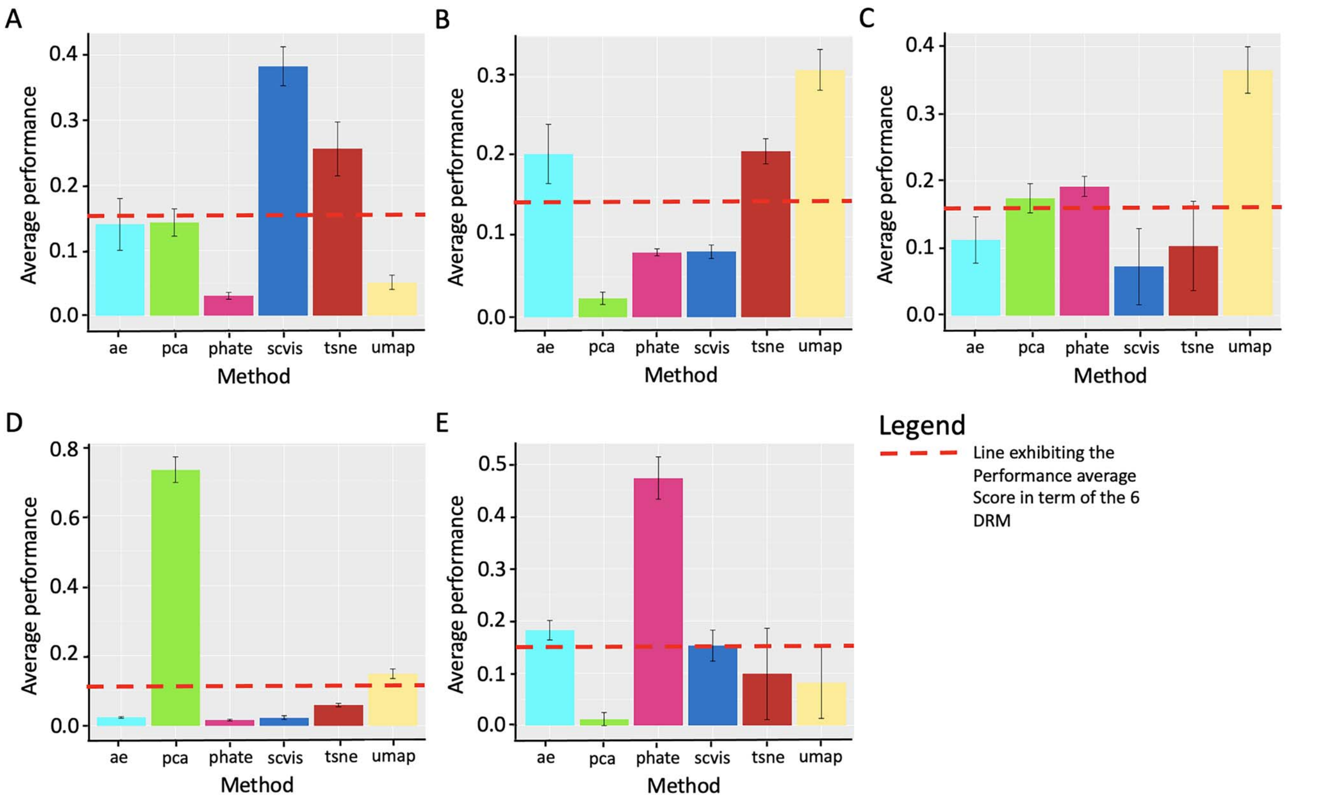


Figure 7. MIBCOVIS analysis of the average accuracy (mean of HDI) of DRM related to the EMT, IPSC, spermatogenesis and CMI data sets for ODSVI. Barplots with error bars summarizing the average performance of (**A**) the OI, (**B**) the GI effect, (**C**) SI, (**D**) UI and (**E**) TI when six DRM methods are applied to the four different biological data sets. The overall average performance score line across all data sets and DRM is 0.165 (A), 0.149 (B), 0.162 (C), 0.166 (D) and 0.166 (E), respectively.

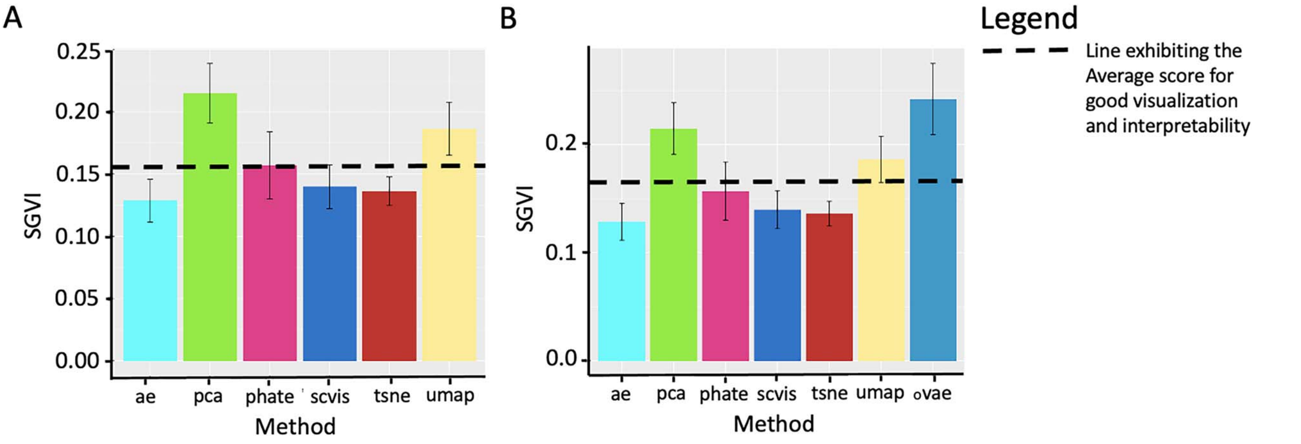


Figure 8. Analysis comparing the oVAE with other methods as a benchmarking tool for MIBCOVIS using barplots with error bars derived from an overall average score for ODSVI (SGVI). (**A**) In methods without oVAE, PHATE, UMAP and PCA show above-average performance (0.160), with PCA and UMAP having the highest performances. (**B**) In methods including oVAE, AE, PHATE, t-SNE and SCVIS have a lower SGVI than the standard (0.172), while PCA and UMAP display the second-highest performance (SGVI = 0.220, SGVI = 0.181, respectively). As oVAE has the highest SGVI (0.240), it is the most optimal benchmarking method on average.

in a semi-supervised analysis. Using the superior performance of support vector machines (SVMs) (Supplementary Figure 57), the results show significant improvement for many methods under SSL compared to supervised learning (SL), as seen in the ratio of average to total average performance (Figure 9). For example, t-SNE improved sensitivity of classification by 21.8% and oVAE improved separability (26%), coverage (13.4%) and uniformity (10%), while PHATE improved time dependency (8.5%). Overall, oVAE, t-SNE and SCVIS performed best for OI, while t-SNE, UMAP and oVAE performed best for GI. However, PHATE, SCVIS and AE did not perform well for UI. In a nutshell, MIBCOVIS exhibited similar top-ranking outcomes in the SSL analysis as in the SL analysis (Supplementary Figures 58–61).

DISCUSSION

Visualizing and interpreting high-dimensional, dynamic and spatial data sets is a major challenge in various scientific fields, such as developmental biology, where reduction of heterogeneous

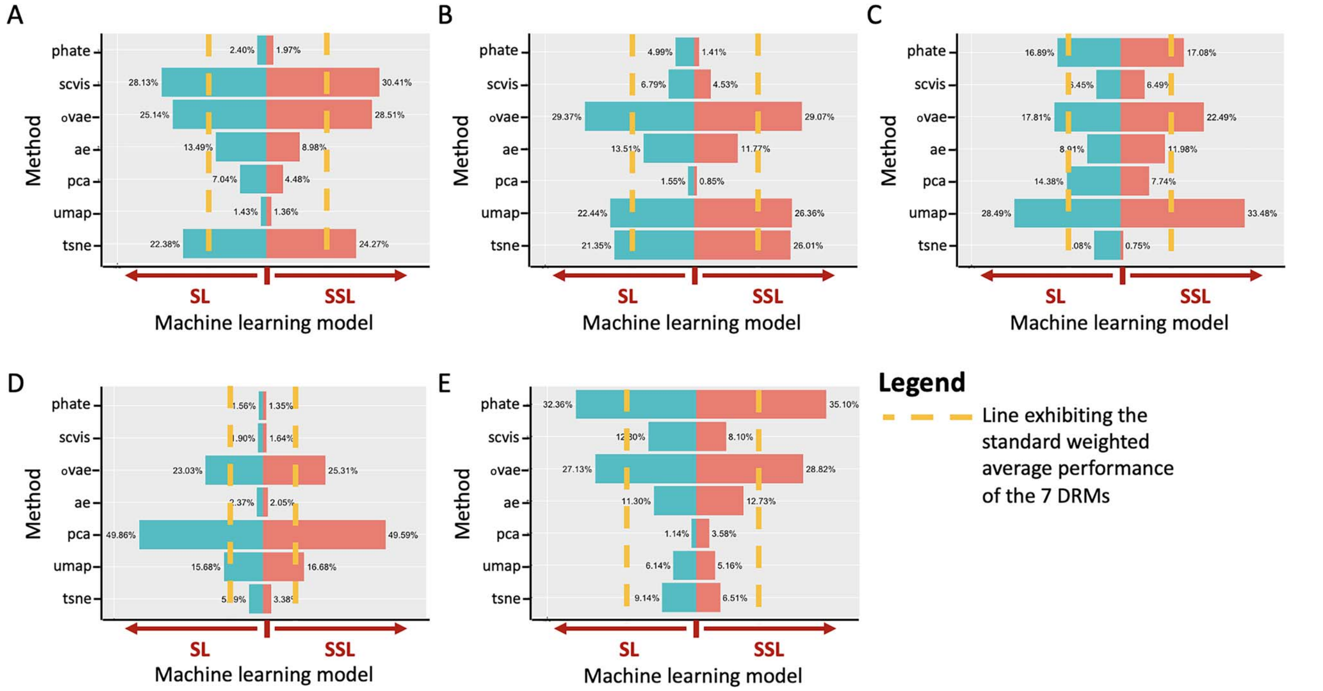


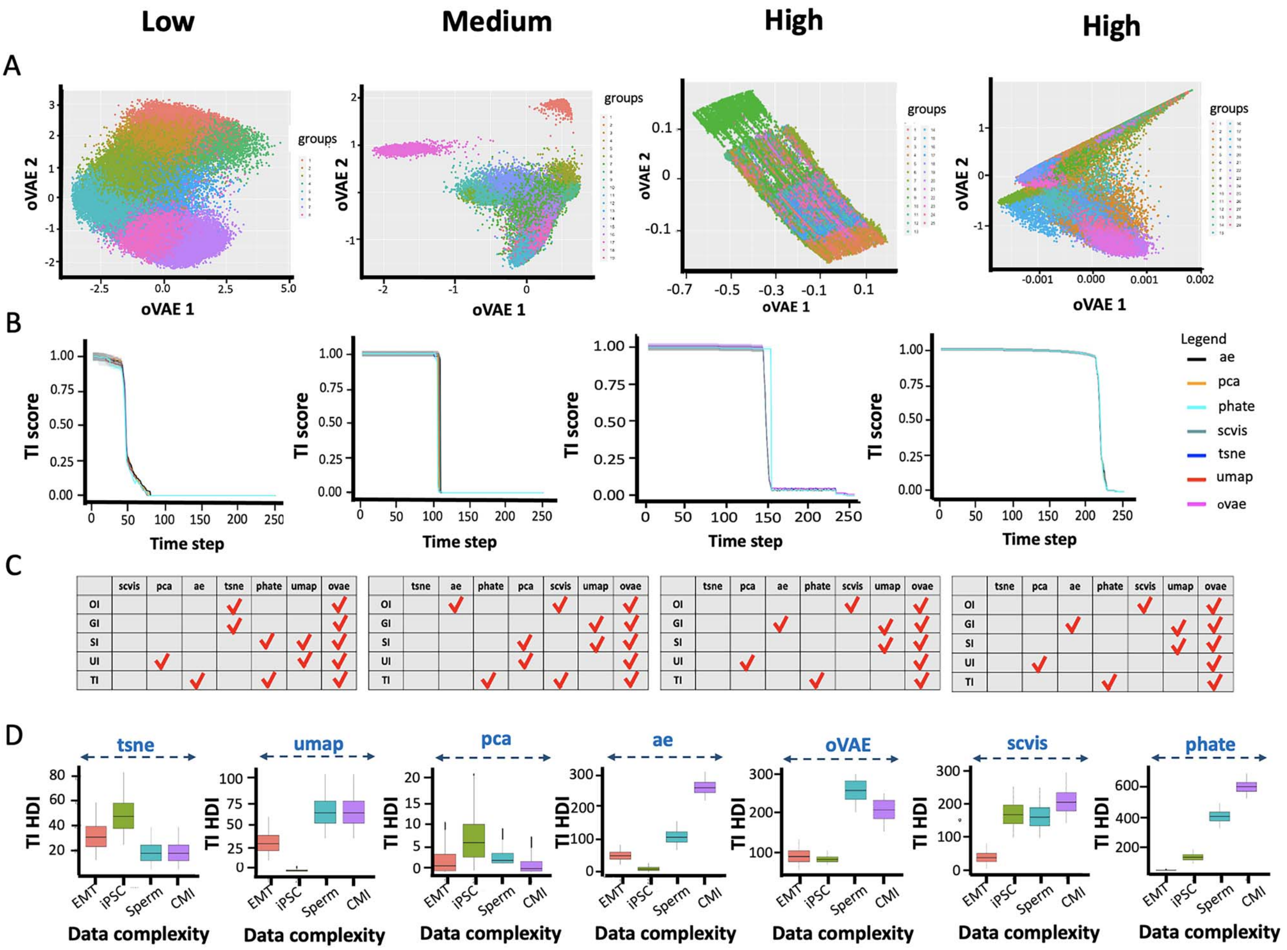
Figure 9. Weighted average performance of DRM analysis for MIBCOVIS using supervised learning (SL) and semi-supervised learning (SSL). The performance of each DRM is displayed as a percentage value beside the back-to-back) SL and SSL barplots. The dashed line corresponding to standard weighted average performance $standard_{wp}$ is calculated as the mean of the weighted average performance, which is 14.29% for all seven DRMs. The histograms display the weighted average performance of DRM for the (**A**) OI, (**B**) GI, (**C**) SI, (**D**) UI and (**E**) TI when applied to three biological data sets under both SL and SSL.

single-cell data is needed to understand complex biological interactions. To reduce data complexity, many dimensionality reduction approaches have been developed. However, evaluation of these methods [42] primarily focus on independently static visualization or interpretability metrics, whose performance depends on knowing the ground truth. In many applications, e.g. trajectory single-cell analysis, there is no known ground truth for the reduced data, making it challenging to directly evaluate accuracy. Effective benchmarking methods that can account for this uncertainty and dynamic stochasticity are needed to provide alternative ways to assess the visualization and interpretability quality of dimensionality reduction techniques.

In this study, we developed MIBCOVIS, a computational framework to assess and compare the performance of DRMs for ODSVI in terms of data reduction (clustering), visualization and dynamic (spatial) interpretability accuracy across various data complexities. MIBCOVIS ranks the performance of each method using five different metrics within a hierarchical Bayesian framework that accounts for the joint correlation effect of all metric parameters. We optimize for visual accuracy using OI and UI, for data reduction accuracy using GI and for dynamic interpretability accuracy using SI and TI. Although these metrics do not span the entire spectrum of features for ODSVI, they allow for a robust, easy, reproducible and interpretable comparison of different methods. For example, we show in Figure 3 that GI is a robust metric for evaluating clustering accuracy in single-cell analysis compared to traditional metrics like the SC and the DB measures. We also demonstrated that no current method optimizes the proposed features for ODSVI jointly (Figures 6 and 7) and that ranking of DRMs under independence performance (Figure 5) is different from ranking of methods under joint performance (Figure 7). Also, we found that t-SNE tends to produce islands in the projected space (Figure 1A), which may not represent maximum separability of homogeneous clusters (Supplementary Texts 6 and 7, Supplementary Table 2). PHATE separates EMT clusters better due to their higher variances in comparison with the variances of other methods from centroids. We also observed that t-SNE displays time-dependent EMT transitions better than UMAP (Figure 6E, Supplementary Figures 7A and B), as captured by the TI metric and supported by PHENOSTAMP-TRACER [32] EMT visualization output while PHATE visualizes various cell states, transitions, spatial interactions and trajectories the best.

TI models time-dependent patterns in data while considering cell–cell similarity. The TI analysis of the CODEX imaging data set (Supplementary Figure 65) demonstrates its effectiveness in scoring data sets with spatial information. The flexibility of TI makes it a suitable metric for evaluating joint temporal and spatial relationships in biological processes. Furthermore, TI analysis uncovers the relationship between cellular dynamics (Figure 10B) and data complexity (Figure 10C and D). These results highlight the need for further investigation to identify key timepoints or spatial neighborhoods of cells driving high-variability transitions.

It is worth noting that the TI performance of methods like PHATE and Autoencoder (Figure 10D) appears to increase with data complexity. PHATE employs the diffusion theory, governing the movement of molecules and ions within and between cells [43]. The relatively high TI performance of neural network projections (Figure 10D) of highly complex data sets underscores their capability to learn latent features and model temporal dependencies and variations over time, as demonstrated in Karacosta *et al.* [32]. However, the absence of variability between methods in the non-stationarity aspect of the TI Markov chain (Figure 10B) indicates the necessity for improved visualization techniques that consider non-homogeneous spatiotemporal structural variations, especially in the context of single-cell analysis.



	scvis	pca	ae	tsne	phate	umap	ovae
OI				✓			✓
GI				✓			✓
SI					✓	✓	✓
UI		✓				✓	✓
TI			✓		✓		✓

	tsne	ae	phate	pca	scvis	umap	ovae
OI		✓			✓		✓
GI						✓	✓
SI				✓		✓	✓
UI				✓			✓
TI			✓		✓		✓

	tsne	pca	ae	phate	scvis	umap	ovae
OI					✓		✓
GI			✓			✓	✓
SI						✓	✓
UI		✓					✓
TI				✓			✓

	tsne	pca	ae	phate	scvis	umap	ovae
OI					✓		✓
GI			✓			✓	✓
SI						✓	✓
UI		✓					✓
TI				✓			✓

Figure 10. Summary of optimal visualization and interpretability of methods based on data complexity and feature accuracy. (**A**) The data reduction maps illustrate the application of the oVAE to four distinct high-dimensional single-cell data sets: EMT data (low), iPSC data (medium), CMI (high) and spermatogenesis data (high). The oVAE showcases robust local clustering, successfully grouping similar categories even at low scale values. (**B**) Time order structure index evaluates the DRM performance based on data complexity measured by dimension size and number of cell types. Low, medium and high complexity is determined from analyzing EMT, IPSC, CMI and spermatogenesis data sets, respectively. Changes in data complexity and cellular dynamics are associated with changes in DRM performance, with an increase in delay to stationarity observed with increased complexity. The performance of all DRMs seems to be associated with changes in the complexity as well as cellular dynamics. An increase in delay to stationarity can be observed with increase in data complexity. (**C**) Tables highlight above-average method-metric performance for low-, medium- and high-complexity data sets (EMT, IPSC, CMI and spermatogenesis, respectively). (**D**) Box plots showing the variation of TI performance scores by data complexity for t-SNE, UMAP, PCA, AE, oVAE, SCVIS and PHATE, respectively.

This study's benchmarking approach explains observed differences in data visualization by considering the correlation and moderation effects of model parameters defining an ODSVI's features. Performance is evaluated using five-feature metrics and applied to linear, nonlinear and neural network methods on four single-cell data sets. To account for uncertainty in the ground truth, a semi-supervised approach is developed for benchmarking. Semi-supervised MIBCOVIS showed a strong improvement in performance accuracy for the best methods compared to the supervised analysis. Such an approach allows for optimal visualization of cell state transitions based on gene expression patterns. Simulations can also be used to compare different methods with synthetic data sets, but simple deterministic data structures like in Saelens *et al.* [18] are insufficient for representing stochasticity in complex dynamic biological processes.

MIBCOVIS enables users to visualize and interpret specific data reduction, visualization and dynamic features while accounting for data complexity. Supplementary Figures 52–56 and Figure 10D show the variation of SI (Supplementary Figure 52), GI (Supplementary Figure 53), OI (Supplementary Figure 54), UI (Supplementary Figure 55) and TI (Figure 10D) performance scores by data complexity for t-SNE, UMAP, PCA, AE, oVAE, SCVIS and PHATE, respectively. Figure 10C summarizes which combination of methods and metric features yield above-average visualization. For example, oVAE and UMAP are optimal if one requires a DRM that maximizes clustering accuracy, visual separability of classes and coverage of the projected space with increasing data complexity. oVAE offers a robust baseline choice (Figure 10A), combining most of the advantages of PCA with a higher cluster variability and optimal visual output, making it useful for users who are uncertain about which feature to optimize for visualization. It also has a shorter training time than SCVIS and a closed-form formula for prediction or sensitivity analysis. Recent implementations of variational AEs (scVAE [44], bmVAE [45] and siVAE [46]) have been used for optimal static visualization of single-cell data sets.

Given the increasing interest in computational tools that integrate feature data (protein expression, gene expression) and spatial information (cellular neighborhoods, spatial locations) [5], MIBCOVIS provides a powerful framework to evaluate the visualization of spatial or temporal relationships of these data reduction integration tools. For example, Supplementary Figure 65 shows that, by integrating a spatially constrained clustering of a CODEX image intestinal single-cell data and spatially constrained data reduction model, MIBCOVIS shows a superior overall visualization performance compared to the spatially unconstrained model. Furthermore, the posterior distributions from the MIBCOVIS framework can be used to integrate targeted DRMs and features to improve dynamic or spatial data visualization and interpretability.

MIBCOVIS has the potential for applications beyond visualization. It can be utilized to leverage the moderation effect framework and the flexible TI Markov process for structure learning and feature selection. Additionally, further evaluation of data reduction methods for the integration of various single-cell data modalities and benchmarking spatio-temporal data reduction methods in dimensions greater than or equal to 3D would be a valuable extension of this work. In summary, MIBCOVIS offers a robust set of metrics and an unbiased benchmarking framework for optimizing the visualization and interpretability of spatio-temporal relationships, thereby enabling deeper insights into complex biological systems.

Key Points

- MIBCOVIS offers a robust set of metrics and an unbiased model-based benchmarking framework to achieve optimal visualization and interpretability of spatio-temporal relationships, thereby facilitating deeper insights into complex biological systems.
- MIBCOVIS uses a time-ordered Markov-based community structure metric to assess single-cell dynamic or spatial variations and cell–cell interactions in distinct dynamic and spatial biological processes.
- In contrast to conventional methods relying on single-summary statistics, MIBCOVIS conducts comprehensive comparisons across data reduction methods by considering entire parameter and accuracy distributions.
- The study highlights that existing data reduction methods do not optimize for dynamic visualization and interpretability simultaneously. Notably, the ranking of methods differs when assessed for independence performance compared to dependency performance.
- The study provides valuable insights by offering optimal parameter ranges, visualization features and methods, including the optimized Variational Contractive Autoencoder (oVAE), as a benchmarking tool for visualization and interpretability across diverse data modalities and complexity scenarios.

ACKNOWLEDGEMENTS

We would like to acknowledge Dr John Hickey (Stanford University), Dr Jian-Liang Li (NIEHS) and Dr Shyamal Peddada (NIEHS) for providing some of the data sets in the study or valuable feedback in shaping this manuscript.

FUNDING

This work is supported by funding from Chan Zuckerberg Initiative DAF, an advised fund of Silicon Valley Community Foundation and National Institute of Environmental Health Sciences (NIEHS) with grant numbers: 2021-240134 (5022) and 1ZIAES103350-03 and ES103388-01, respectively. This research was supported by the Intramural Research Program of the NIH, NIEHS.

AUTHOR CONTRIBUTIONS

K.A. and B.A. contributed to the concept and algorithm of MIBCOVIS. K.A. and B.A. were involved in the metric implementation. K.A. and B.A developed and implemented the time order structure metric and performed the benchmarking analysis. B.A. was involved in data generation and pre-processing. K.A. A.M.R. and B.A. were involved in writing the initial manuscript. All authors were involved in interpreting the results.

DATA AVAILABILITY

The spermatogenesis data set with about 110 000 cells across 25 developmental stages used in this study was processed from published raw scRNA-seq data sets for mouse spermatogenesis from Gene Expression Omnibus (GEO) under accession codes GSE121904 [47], GSE124904 [34] and GSE117707 [48] and from the ArrayExpress database under accession code EMTAB-6946 [49]. The processed IPSC data set with about 50 000 cells across 12 time points was obtained from raw scRNA-seq data sets for mouse MEFs and chemically iPSCs from GEO under accession code GSE114952 [33]. The detailed preprocessing of scRNA-seq data sets and generation of cluster labels and selection of markers using the above data sets is described in Anchang *et al.* [19]. We also downloaded arcsinh-transformed CyTOF time-course data for the EMT analysis from Karacosta *et al.* [32]. The preprocessed CODEX multiplexed imaging single-cell protein data of the human intestine from donor B005 comprising of about 282 000 cells from tile stitching, drift compensation, cycle concatenation, background subtraction, deconvolution and determination of best focal plane, followed by single-cell segmentation, and column marker z-normalization by tissue can be downloaded from https://doi.org/10.5061/dryad.pk0p2ngrf. MIBCOVIS Supplementary data for EMT, IPSC, spermatogenesis and intestine used in this study also provided as R Source Data files can be found in https://github.com/NIEHS/MUBCOVID.git.

CODE AVAILABILITY

In summary, MIBCOVIS is a computational framework to evaluate dimension reduction methods for ODSVI using Bayesian multilevel modeling. While traditional benchmarking methods define the performance of a method as independent variable for ODSVI, the MIBCOVIS computational framework models the performance of methods as a multivariate function with focus on the conditional effects of multiple factors which illustrate ODSVI. MIBCOVIS including all scoring functions for various metrics is implemented using R available on GitHub (https://github.com/NIEHS/MUBCOVID.git).

REFERENCES

1. Linderman GC, Rachh M, Hoskins JG, *et al.* Fast interpolation-based t-SNE for improved visualization of single-cell RNA-seq data. *Nat Methods* 2019;**16**:243–5.
2. Becht E, McInnes L, Healy J, *et al.* Dimensionality reduction for visualizing single-cell data using UMAP. *Nat Biotechnol* 2019;**37**: 38–44.

3. Sun S, Zhu J, Ma Y, Zhou X. Accuracy, robustness and scalability of dimensionality reduction methods for single-cell RNA-seq analysis. *Genome Biol* 2019;**20**:1–21.
4. Butler A, Hoffman P, Smibert P, *et al.* Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat Biotechnol* 2018;**36**:411–20.
5. Hickey JW, Becker WR, Nevins SA, *et al.* Organization of the human intestine at single-cell resolution. *Nature* 2023;**619**: 572–84.
6. Amelio A, Pizzuti C. Is normalized mutual information a fair measure for comparing community detection methods? In: *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015*. Paris, France, 2015, pp. 1584–5.
7. Walkowiak T, Gniewkowski M. Evaluation of vector embedding models in clustering of text documents. In: *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*. Varna, Bulgaria. INCOMA Ltd., 2019, pp. 1304–11.
8. Campello RJ. A fuzzy extension of the Rand index and other related indexes for clustering and classification assessment. *Pattern Recogn Lett* 2007;**28**:833–41.
9. Steinley D. Properties of the Hubert-Arable adjusted Rand index. *Psychol Methods* 2004;**9**:386–96.
10. Wang G, Wang Z, Chen W, Zhuang J. Classification of surface EMG signals using optimal wavelet packet method based on Davies-Bouldin criterion. *Med Biol Eng Comput* 2006;**44**: 865–72.
11. Kullback S, Leibler RA. On information and sufficiency. *Ann Math Stat* 1951;**22**:79–86.
12. Ying S, Wen Z, Shi J, *et al.* Manifold preserving: an intrinsic approach for semisupervised distance metric learning. *IEEE Trans Neural Netw Learn Syst* 2017;**29**:1–12.
13. Breiman L. Random forests. *Machine learning* 2001;**45**:5–32.
14. Dries R, Zhu Q, Dong R. *et al.* Giotto: a toolbox for integrative analysis and visualization of spatial expression data. *Genome Biol* 2021;**22**:78.
15. Habib N, Avraham-Davidi I, Basu A, *et al.* Massively parallel single-nucleus RNA-seq with DroNc-seq. *Nat Methods* 2017;**14**: 955–8.
16. Duò A, Robinson MD, Soneson C. A systematic performance evaluation of clustering methods for single-cell RNA-seq data. *F1000Research* 2020;**7**:1141.
17. Moon KR, van Dijk D, Wang Z, *et al.* Visualizing structure and transitions in high-dimensional biological data. *Nat Biotechnol* 2019;**37**:1482–92.
18. Saelens W, Cannoodt R, Todorov H, Saeys Y. A comparison of single-cell trajectory inference methods. *Nature biotechnology*. 2019;**37**:547–54.
19. Anchang B, Mendez-Giraldez R, Xu X, *et al.* Visualization, benchmarking and characterization of nested single-cell heterogeneity as dynamic forest mixtures. *Brief Bioinform* 2022;**23**: bbac017.
20. Kosslyn SM. *Graph Design for the Eye and Mind*. (2006; online edn, Oxford Academic, 2012). https://doi.org/10.1093/acprof:oso/9780195311846.001.0001 (6 December 2023, date last accessed). Oxford University Press.
21. Sun F, Wang Y, Xu H. Uniform projection designs. *Ann Stat* 2019;**47**:641–61.
22. Chen T, Guestrin C. Xgboost: A scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. (pp. 785–94). New York, NY, U SA: ACM. https://doi.org/10.1145/2939672.2939785.
23. Fang L, Zhao H, Wang P, *et al.* Feature selection method based on mutual information and class separability for dimension reduction in multidimensional time series for clinical data. *Biomed Signal Process Control* 2015;**21**:82–9.
24. Samusik N, Good Z, Spitzer MH, *et al.* Automated mapping of phenotype space with single-cell data. *Nat Methods* 2016;**13**: 493–6.
25. Greene J. Feature subset selection using thornton's separability index and its applicability to a number of sparse proximity-based classifiers. In *Proceedings of annual symposium of the pattern recognition association of South Africa*, 2001.
26. Gentle JE. *Matrix Algebra*. Springer Texts in Statistics. New York, NY: Springer, 2007.
27. Kufer S. *Effective and Efficient Summarization of Two-Dimensional Point Data: Approaches for Resource Description and Selection in Spatial Application Scenarios*. Germany: University of Bamberg Press, 2019.
28. Chen T, He T, Benesty M, *et al.* Xgboost: extreme gradient boosting. In: CRAN *R Package Version 0.4–2*, 2015;**1**:1–4.
29. Ovchinnikova S, Anders S. Exploring dimension-reduced embeddings with sleepwalk. *Genome Res* 2020;**30**:749–56.
30. Kinalis S, Nielsen FC, Winther O, Bagger FO. Deconvolution of autoencoders to learn biological regulatory modules from single cell mRNA sequencing data. *BMC Bioinformatics* 2019;**20**:1–9.
31. Rashid S, Shah S, Bar-Joseph Z, Pandya R. Dhaka: variational autoencoder for unmasking tumor heterogeneity from single cell genomic data. *Bioinformatics* 2021;**37**:1535–43.
32. Karacosta LG, Anchang B, Ignatiadis N, *et al.* Mapping lung cancer epithelial-mesenchymal transition states and trajectories with single-cell resolution. *Nat Commun* 2019;**10**:1–15.
33. Zhao T, Fu Y, Zhu J, *et al.* Single-cell RNA-seq reveals dynamic early embryonic-like programs during chemical reprogramming. *Cell Stem Cell* 2018;**23**:31–45.e7.
34. Law NC, Oatley MJ, Oatley JM. Developmental kinetics and transcriptome dynamics of stem cell specification in the spermatogenic lineage. *Nat Commun* 2019;**10**:1–14.
35. Liesecke F, Daudu D, Dugé de Bernonville R, *et al.* Ranking genome-wide correlation measurements improves microarray and RNA-seq based global and targeted co-expression networks. *Sci Rep* 2018;**8**:1–16.
36. Bowman FD, Caffo B, Bassett SS, Kilts C. A Bayesian hierarchical framework for spatial modeling of fMRI data. *Neuroimage* 2008;**39**:146–56.
37. Carlin BP, Chib S. Bayesian model choice via Markov chain Monte Carlo methods. *J R Stat Soc B Methodol* 1995;**57**:473–84.
38. Bailer-Jones C, Rybizki J, Fouesneau M, *et al.* Estimating distance from parallaxes. IV. Distances to 1.33 billion stars in Gaia data release 2. *Astron J* 2018;**156**:58.
39. Kruschke J. *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan*. 2nd edn (Academic, 2015).
40. Dony L, König M, Fischer D *et al.* Variational autoencoders with flexible priors enable robust distribution learning on single-cell RNA sequencing data. In: *ICML 2020 Workshop on Computational Biology (WCB) Proceedings Paper* Vol 37, Vienna, Austria, 2020.
41. Rifai S, Vincent P, Muller X, *et al.* Contractive auto-encoders: explicit invariance during feature extraction. In *Proceedings of the 28th International Conference on Machine Learning, Bellevue*, WA, USA, Vol. 32, 2011 (pp. 833–40).
42. Huang H, Wang Y, Rudin C, Browne EP. Towards a comprehensive evaluation of dimension reduction methods for transcriptomic data visualization. *Commun Biol* 2022;**5**:719.
43. Phillips R, Kondev J, Theriot J. *Physical Biology of the Cell*. Boca Raton: Garland Science, 2012.

44. Grønbech CH, Vording MF, Timshel PN, *et al.* scVAE: variational auto-encoders for single-cell gene expression data. *Bioinformatics* 2020;**36**:4415–22.
45. Yan J, Ma M, Yu Z. bmVAE: a variational autoencoder method for clustering single-cell mutation data. *Bioinformatics* 2023;**39**:btac790.
46. Choi Y, Li R, Quon G. siVAE: interpretable deep generative models for single-cell transcriptomes. *Genome Biol* 2023;**24**:29.
47. Grive KJ, Hu Y, Shu E, *et al.* Dynamic transcriptome profiles within spermatogonial and spermatocyte populations during postnatal testis maturation revealed by single-cell sequencing. *PLoS Genet* 2019;**15**:e1007810.
48. Wang Z, Xu X, Li J-L, *et al.* Sertoli cell-only phenotype and scRNA-seq define PRAMEF12 as a factor essential for spermatogenesis in mice. *Nat Commun* 2019;**10**: 1–18.
49. Ernst C, Eling N, Martinez-Jimenez C, *et al.* Staged developmental mapping and X chromosome transcriptional dynamics during mouse spermatogenesis. *Nat Commun* 2019;**10**: 1251.