# Benchmarking Spatiotemporal Data Reduction Models with Unknown Ground Truth using Optimal Visualization and Interpretability Metrics

Komlan Atitey

*Biostatistics and Computational Biology Branch National Institute of Environmental Health Sciences 111 T W Alexander Dr Rall Building, Research Triangle Park, NC 27709*

*E-mail: komlan.atitey@nih.gov*

Alison Anne Motsinger-Reif

*Biostatistics and Computational Biology Branch National Institute of Environmental Health Sciences 111 T W Alexander Dr Rall Building, Research Triangle Park, NC 27709*

*E-mail: alison.motsinger-reif@nih.gov*

Benedict Anchang*

*Biostatistics and Computational Biology Branch, National Institute of Environmental Health Sciences 111 T W Alexander Dr Rall Building, Research Triangle Park, NC 27709*

**E-mail: benedict.anchang@nih.gov*

**Abstract**

Data reduction is essential for understanding complex processes with high-dimensional variable interactions. However, different data reduction methods applied to the same temporal data yield different spatial maps, which complicates downstream analysis, interpretability, and data visualization. This highlights the necessity for robust visualization metrics and benchmarking models, particularly when ground truth data is absent, such as in single-cell analysis. We propose a new analytical framework called MIBCOVIS. It incorporates five robust visualization and interpretability metrics within a hierarchical Bayesian model. Our framework benchmarks the performance of linear, nonlinear, and artificial neural network dimensionality reduction methods applied to visualize three distinct dynamic biological processes. We discovered that no current method optimizes for joint visualization and interpretability, and provide optimal parameter regions, features, and methods, including an optimized variational autoencoder called oVAE for targeted visualization. We anticipate that MIBCOVIS would be useful for benchmarking single-cell atlases or spatio-temporal data reduction models in general.

## Introduction

The rise of big data in various fields presents challenges for extracting insights from complex datasets. Traditional visualization methods such as scatter plots and heatmaps, become less effective as the number of dimensions increases, leading to the use of more advanced techniques of dimensionality reduction like t-SNE[1] and UMAP[2]. However, different data reduction methods (DRM) may produce conflicting results, creating problems for feature anotations and interpretability. For example, in a field like single-cell analysis[2,3], several DRM when applied to the same data tend to produce different visual outputs (Fig. 1a, Supplementary Figs. 1-6) thereby, confounding the entire cell labeling process and interpretability of cellular relationships. 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 dataset. To address this issue, there is a need for advanced visualization techniques that are both informative and interpretable, as well as for careful benchmarking of statistical models and algorithms used to analyze high-dimensional data. Currently, there is no data reduction model or benchmarking performance metric that optimizes for data complexity, visualization, interpretability, and overfitting simultaneously. To tackle this problem, we propose a **m**ultivariate **i**nterpretable **b**enchmarking and **co**mputational framework for optimal **v**isualization and **i**nterpretability of high-dimensional **s**eparable data with and without ground truth, called **MIBCOVIS**. This framework can be applied to single-cell and non-single cell data from various fields.

Data reduction techniques are widely used in many scientific and statistical applications to analyze large and complex datasets. Principal component analysis (PCA), for example, is commonly used in genetics[3], neuroscience[4], and image processing[5] to identify patterns and structure in high-dimensional data. Feature selection techniques are used to identify relevant features for a given task. This is commonly used in machine learning[6] and data mining[7] to improve the performance of models and reduce the computational complexity of the analysis. Data compression techniques reduce storage and transmission costs of large datasets. This is particularly important in fields such as astronomy[8] and remote sensing[9]. Additionally, effective dimensionality reduction is crucial for single-cell analysis methods such as clustering and trajectory modeling[10]. The huge number of methodologies[11, 12] that are optimized for linear and non-linear visualizations (Supplementary Table 1), warrants the need for a unified benchmarking framework that accounts for these variations.

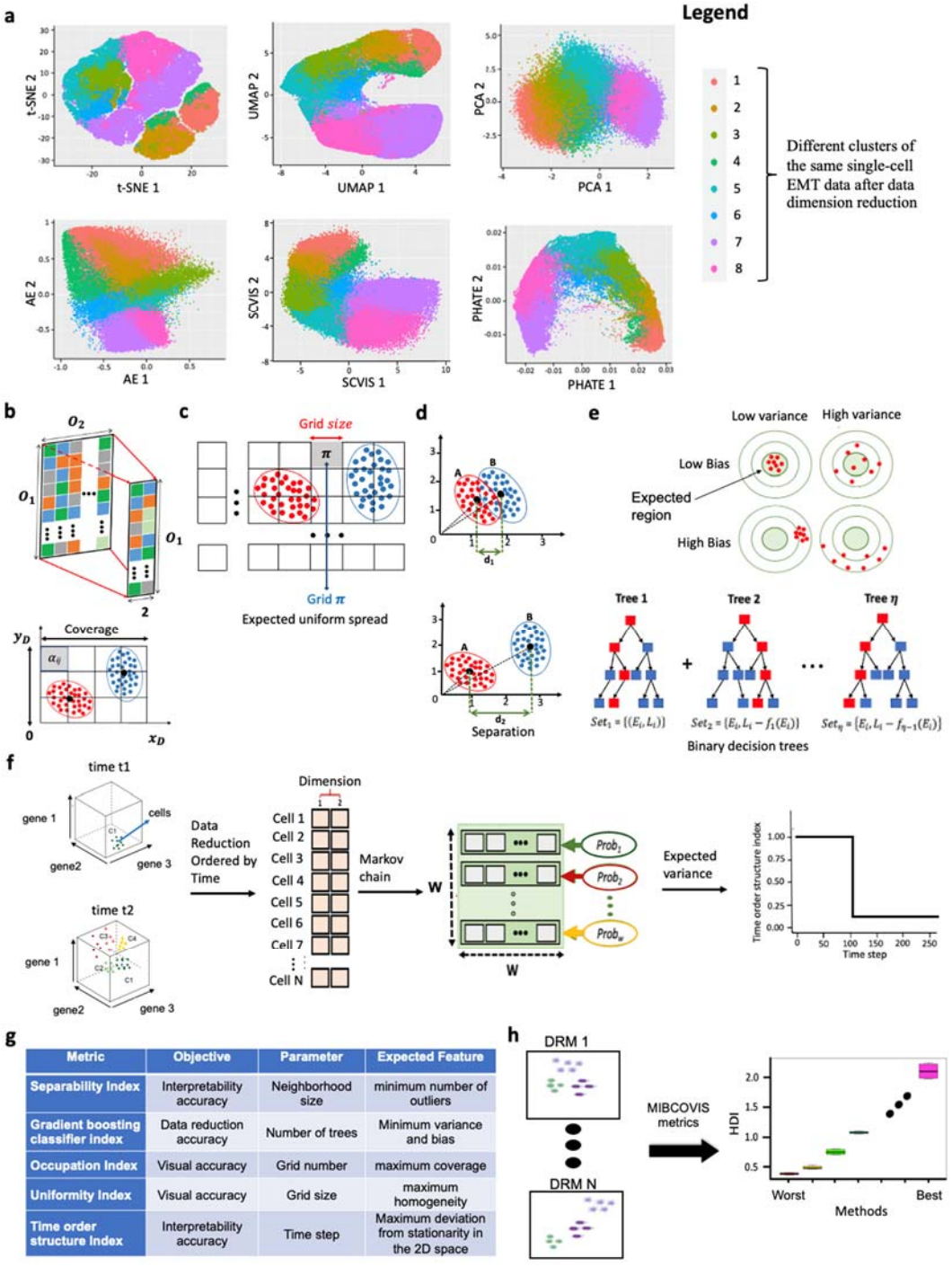
Benchmarking various data reduction models for **optimal visualization and interpretability** (OVI) involves evaluating their performance against a set of predefined criteria or metrics. Some common metrics for OVI optimize for; clustering accuracy e.g. F1 score[13], and the adjusted Rand index[14], dimensionality reduction accuracy e.g. Kullback–Leibler divergence[15], which measures 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, feature importance e.g. Gini importance score[16], 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 DRM in single-cell analysis, benchmarking studies typically use clustering accuracy metrics[17, 18, 19]. 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. Currently, there is no comprehensive and quantitative evaluation of spatio-temporal data reduction[10]. Moon et al. 2019[14] used non-linear correlations to benchmark the Potential of Heat-diffusion for Affinity-based Transition Embedding (PHATE) method based on diffusion metrics with other DRM, while Saelens et al. 2019[20] used a common trajectory representation and four independent metrics to compare various pseudotime trajectory methods from simulated scRNA-seq data. In a recent study[21], non-linear correlation and network similarity statistical models were used to benchmark pseudotime and temporal trajectory methods, respectively. However, these benchmarking metrics do not account for uncertainty in the ground truth or optimize for visualization features. Moreover, current benchmarking frameworks tend to compare these metrics independently and do

not help interpret which factors contribute to the observed differences in visualization or interpretability performance.

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 OVI. We reflect on recent investigations about OVI by paying specific attention to ways in which these abstract features can be addressed quantitatively particular for methods that generate low-dimensional metric maps also known as self-organizing maps (SOM) as outputs. We introduce two visual accuracy metrics, the occupation index (OI) and the uniformity index (UI), to assess the performance of a DRM for user clarity and persuasiveness[22]. Firstly, using a similar context in material science[23], 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 high dimensional space (Fig. 1b). Secondly, UI describes the uniform spread of data points in the low dimensional space, assessing both uniformity and orthogonality (Fig. 1c), two properties associated with many data reduction models[24]. Note our definition of UI is different from that used in other fields to assess the diversity or heterogeneity of a system. The OI and UI both range from 0 to 1 use boundary grids of points to grade how well the projected points occupy the whole space (Fig. 1b) and how close the projected points are uniformly distributed (Fig. 1c) respectively. To quantify data reduction accuracy (clustering accuracy), we introduce a third metric, the Gradient boosting Classifier Index (GI) which uses a Gradient Boosting Classifier[25] to assess the sensitivity of clusters in the projected space. To further evaluate the degree of separation or distinction between these clusters after data reduction, we introduce 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[26]. Various researchers have used different formulations of class separability as a benchmark for evaluating the performance of DRM in single-cell analysis[2, 27]. We classify SI as an interpretability accuracy metric and use it to estimate the average number of instances in a dataset that have a nearest neighbor with the same cluster label[28]. Furthermore, to assess the correlation and causation accuracy of a given DRM, we introduce a second interpretability accuracy metric, the Time Order Structure Index (TI) (Fig. 1f, Fig. 2) which uses Markov chains to assess the time dependency of ordered points (cells for single-cell data) in the low-dimensional projected space. In a nutshell, this study introduces 5 metrics to optimize 3 dependent performance objectives for OVI of data reduction models: (1) visual accuracy, (2) data reduction accuracy, and (3) interpretability accuracy. Fig. 1g summarizes the five metrics, objectives, key parameters and expectations to assess the performance of dimensionality reduction methods for OVI.

MIBCOVIS aims to integrate correlated features within a statistical framework for optimal benchmarking. Fig. 1h shows the input-output features of MIBCOVIS, which takes projected output data from linear, non-linear, and neural network DRM 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 (Fig. 3 and methods). To account for uncertainty in the ground truth, MIBCOVIS is applied to data under both supervised and semi-supervised learning frameworks. In the supervised learning framework, the source datasets used for benchmarking are discretely labeled, while under the semi-supervised approach, both labeled and unlabeled datasets are used to improve the classification accuracy and the benchmarking pipeline. MIBCOVIS is used to benchmark 6 major DRM spanning linear, non-linear, and neural network method space using data from three dynamic biological processes involved in normal development: Epithelial Mesenchymal Transition (EMT) plasticity, spermatogenesis, and stem cell reprogramming. We investigate the effect of increasing data complexity on visualization performance, defined in terms of feature dimension, unknown cell types, and dynamic differences in biological processes. Using MIBCOVIS, we demonstrate that no current method optimizes the proposed features for OVI jointly. We suggest oVAE, a joint variational and contractive autoencoder (Supplementary text 2), as an optimal benchmarking method when the user is unsure which OVI feature to target. This study offers a reliable

metric set and an unbiased benchmarking framework for OVI of relationships in spatio-temporal metric maps after data reduction.




| Metric | Objective | Parameter | Expected Feature |
|---|---|---|---|
| Separability Index | Interpretability accuracy | Neighborhood size | minimum number of outliers |
| Gradient boosting classifier index | Data reduction accuracy | Number of trees | Minimum variance and bias |
| Occupation Index | Visual accuracy | Grid number | maximum coverage |
| Uniformity Index | Visual accuracy | Grid size | maximum homogeneity |
| Time order structure Index | Interpretability accuracy | Time step | Maximum deviation from stationarity in the 2D space |

**Fig. 1: Motivation and overview of MIBCOVIS framework, including key parameters for data visualization and interpretability assessment. a** Dimension reduction of the CyTOF EMT dataset with six methods; tSNE, UMAP, PCA, Autoencoder (AE), SCVIS, PHATE presenting different projections with different scales from the lowest (PHATE) to the highest

(t-SNE) which can confound visualization. **b** The Occupation index (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 $\pi$ bins to evaluate the distribution of data points across the space. **d** The Separability Index (SI) measures the degree of separation between classes in a two-dimensional dataset 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 Gradient Boosting Classifier Index (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 bulls-eye target region (green color) represents a model with perfect sensitivity. As we move away from the bull-eye, the predictions get worse. **f** Time order structure index (TI) uses a weighted Markov process to convert an Nx2 ordered data reduced state space from time course data into an W x W stationary state space and estimates TI scores using expected variances between time steps derived from the MCMC chain. **g** Table outlining five metrics, objectives, key parameters and model expectations to assess the performance of dimensionality reduction methods for good visualization and interpretability **h** The MIBCOVIS framework takes the projected output data from DRM and selects key parameters directly associated with the performance of five visualization and interpretability metrics to rank the methods using high-density posterior interval estimates from Bayesian regression modeling.

## Results

We present mathematical definitions of the metrics used for OVI and highlight the DRM 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 three different biological datasets with discrete labels for each cell representing the ground truth, followed by a summary of the results from the semi-supervised analysis.

### MIBCOVIS uses Occupation index (OI) and Uniformity index(UI) to assess visual accuracy

A DRM that maximizes the coverage of projected space can reveal patterns, clusters, and outliers in high-dimensional data (Fig. 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[29]. 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 methods), $Cov_h$ represents coverage of the high dimensional space, determined by the total variance (See methods) and $\kappa$ is a weighting parameter proportional to the surface area per unit summary coverage defined by Kufer S. (2019)[30]. In this study we set $\kappa = 0.33$ . 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 UI, which is determined by the spread of the projected data. To estimate the uniformity index (UI), we use the goodness fit of the Pearson Chi-Squared ($\chi^2$) test of uniformity of the data distribution. The projected space is divided into $\pi$ grids with $\Pi$ the total number of grids (Fig. 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 methods) to define the uniformity index (UI).

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

where, $\Upsilon$ 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 Gradient boosting Classifier Index (GI) to assess accuracy of dimensionality reduction method**

MIBCOVIS uses 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 (Fig. 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[31] 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 is expected to improve the accuracy of the model.

**MIBCOVIS uses Separability index (SI) and Time oder structure index (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 dataset (Fig. 1d). The index ranges from 0 to 1, with higher values indicating better separability. A separability index 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 ($\rho_k$) in the projected dataset such that all cells have a nearest neighbor with the same cluster label $k$ (See methods).

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

where $|\Psi_k|$ 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 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 dependency of cells in a temporal reduced space (2D) ( Figs. 2a-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 (Fig. 2e-f). The framework involves six main steps, including partitioning the reduced data into consecutive submatrices (Fig 2c), estimating cell state probabilities (Fig. 2d), computing expected transitions (Fig. 2e), generating a W×W transition matrix (Fig. 2d), weighting the matrix (Fig. 2g), and computing the cumulative sum and column mean $\mu_j$ to estimate the expected variance ($V_j$). 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.

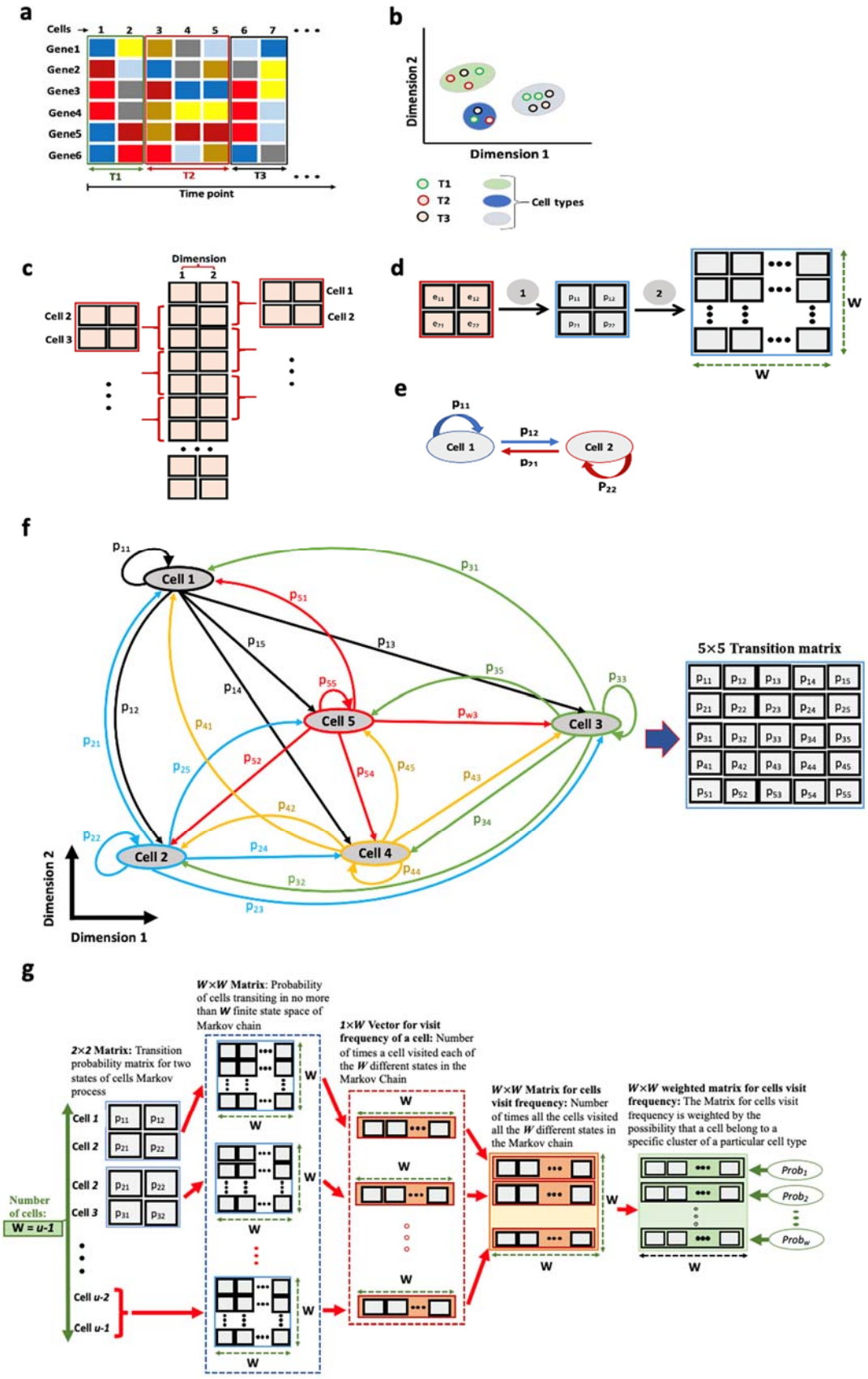

a
Cells
Gene1
Gene2
Gene3
Gene4
Gene5
Gene6
T1
T2
T3
Time point
b
Dimension 2
Dimension 1
T1
T2
T3
Cell types
c
Dimension
Cell 1
Cell 2
Cell 3
d
W
e
Cell 1
Cell 2
f
Cell 1
Cell 2
Cell 3
Cell 4
Cell 5
5×5 Transition matrix
Dimension 2
Dimension 1
g
W×W Matrix: Probability of cells transiting in no more than W finite state space of Markov chain
1×W Vector for visit frequency of a cell: Number of times a cell visited each of the W different states in the Markov Chain
2×2 Matrix: Transition probability matrix for two states of cells Markov process
W×W Matrix for cells visit frequency: Number of times all the cells visited all the W different states in the Markov chain
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
Number of cells:
W = u-1
Cell 1
Cell 2
Cell 2
Cell 3
Cell u-2
Cell u-1

**Fig. 2. Time order structure index (TI) for identifying changes in sub-populations over time**. **a** Single-cell RNA-seq gene expression data collected at multiple time points (T1, T2, 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 dataset into consecutive 2x2 square matrices to capture transition between time points **d** Large projected dataset is partitioned into $u-1$ consecutive 2x2 matrices to capture transition between time points resulting in a $W \times W$ square transition probability matrix. **e** Two cell state Markov chain model and its transition probabilities. **f** A Markov chain with 5 cells at 5 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 is generated by choosing an initial state from the 2 square transition matrices and computing the number of times a cell visits other states in the Markov chain. This choice results in a $W \times W$ transition probability matrix characterized by different initial states. The resulting $W \times W$ matrix is weighted using the proportion of cell states ( $Prob_w$ ) in the data to account for non-uniform distributions of cells during transitions.

## MIBCOVIS selects 6 dimensionality reduction methods to account for variations of all data reduction methods

We evaluated over 20 data reduction tools[11, 12] 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[32], general neural network AE[33] for denoising and imputation of missing values, variational autoencoder SCVIS[34] for scRNA-seq data, and diffusion theory PHATE. These methods account for variations in all DRM.

## MIBCOVIS uses single-cell time course labeled data sets from 3 biological processes for benchmarking

We evaluate MIBCOVIS using three datasets to investigate dynamic biological processes: epithelial-to-mesenchymal transition (EMT)[35] (dataset 1), chemically induced pluripotent stem cells (iPSCs) [36] (dataset 2), and spermatogenesis[37] (dataset 3). For dataset 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 dataset consists of 8 EMT-MET states: E1 (2), E2 (1), E3 (4), pEMT1(3), pEMT2 (5), pEMT3 (6), M (7) and pMET (8) validated using 6 canonical EMT markers. Dataset 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[36]. Dataset 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[35].

## 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 (Fig. 1a, Supplementary Figs. 1-6) for various dimensionality reduction methods challenge our ability to properly interprete dynamic interactions after data reduction. For example, in the study by Karacosta *et al*. they combined t-distributed stochastic neighbor embedding (t-SNE) and a artificial neural network to visualize the plasticity of 8 states of EMT[35] namely; three epithelial (E1, E2, and E3), three partial EMT (pEMT1, pEMT2, and pEMT3), one mesenchymal (M) and one partial MET states (pMET). Interpreting the dynamics visually between these 8 EMT states could be different as shown by the different Voronoi partitions[35] of t-SNE and Uniform Manifold Approximation and Projection (UMAP) maps (Supplementary Fig. 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 OVI performance under a supervised setting by using 2D dimensionality reduction metric maps and ground truth labeled data. It combines visual-based features in a hierarchical Bayesian regression framework to generate conditional posterior distributions of accuracy (Fig. 3a-c) and ranks the performance of methods using boxplots of high density intervals of conditional regression coefficients. Figs. 3a-b summarize six DRM and five metric set used for model building. Fig. 3b 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 Fig. 8a). The third panel (Fig. 3c) includes a two-step sequential analysis to benchmark dimensional reduction methods. In the first step, Spearman correlation coefficient is used to assess the correlation between features from panel b and method accuracy (Supplementary Fig. 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).

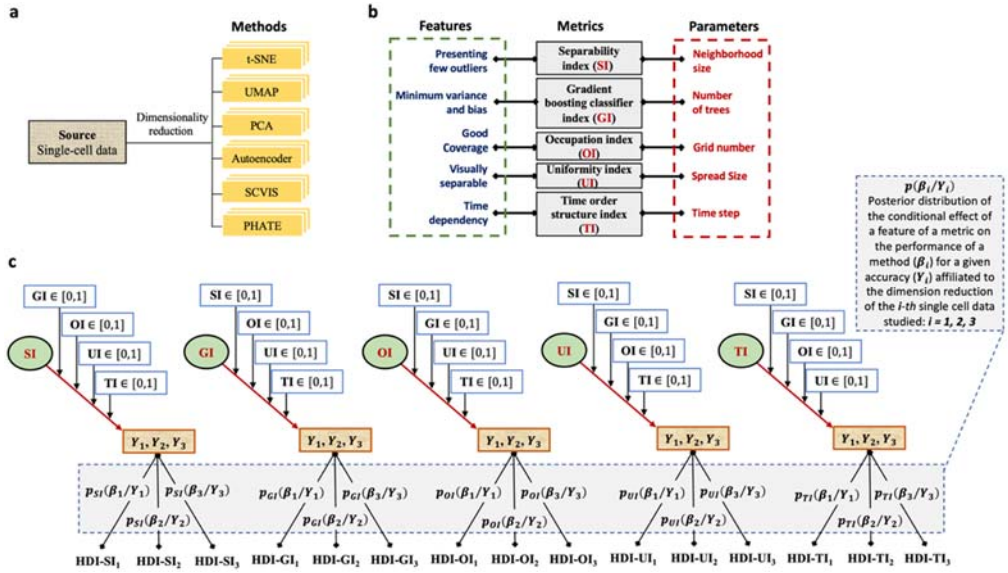



**Fig. 3: MIBCOVIS model framework optimizes visualization and interpretability**. **a** MIBCOVIS uses six dimensionality reduction methods spanning three different model classes (Linear, non-linear, and Neural Network). These include: t-SNE, PCA, UMAP, Autoencoder (AE), SCVIS, and PHATE. **b** MIBCOVIS uses 5 quantitative metrics - Separability index (SI), Gradient boosting classifier index (GI), Occupation index (OI), Uniformity index (UI), and Time order structure index (TI) to evaluate and compare the performance accuracy of the dimensionality reduction methods (DRM). Each of the five metrics depend on an independent parameter $R_p$ of length 250. **c** MIBCOVIS uses the GI to derive a direct overall measure of performance accuracy $Y$ of a given dimension reduction method. We generate uniform GI scores in the range $[S_{min}, S_{max}]$ of length 250 where $S_{min}$, and $S_{max}$ represents the minimum and maximum GI score respectively of all the methods such that:
$S_{min} = min\{min_{tsne}, min_{umap}, min_{pca}, min_{ae}, min_{scvis}, min_{phate}\}$,
$S_{max} = max\{max_{tsne}, max_{umap}, max_{pca}, max_{ae}, max_{scvis}, max_{phate}\}$. Fig. 3c shows accuracy scores ($Y_1$, $Y_2$, and $Y_3$) for reducing EMT, IPSC, and Spermatogenesis single cell data, respectively. We evaluate the correlation between metric features and a given accuracy $Y_1$, $Y_2$, or $Y_3$ using spearman correlation. To account for the dependency between metric features, we use Bayesian conditional regression modeling which describes the performance of a method for a given feature of a metric as a conditional model depending on the moderation effects of other metric features. MIBCOVIS models the relationship between multiple independent (metric features) and one dependent variable (accuracy of method defined by $Y_i$). We present the conceptual model for (SI), (GI), (OI), (UI), and (TI) in Fig. 1c. Using lognormal priors, we compute the posterior distribution of the conditional accuracy effect of a metric feature, assuming a beta distribution of accuracy scores [0, 1]. We analyze the posterior distribution using the Markov Chain Monte Carlo (MCMC) sample and summarize the distributions using a 95% high-density interval (HDI). We construct boxplots of HDI regression coefficients for all 3 data and sum up the averages to compare accuracy between methods.

## MIBCOVIS identifies diverse metric parameter regions for optimal data reduction visualization

We model the variability of five metrics across six dimensionality reduction methods (DRMs) on the

CyTOF EMT dataset (Fig. 4). t-SNE outperforms UMAP, SCVIS, Autoencoder, PHATE, and PCA in terms of SI (Fig. 4a), indicating its ability to better separate the EMT biological process phenotypes with few outliers. We obtain similar results when assessing DRMs on dataset 2 and dataset 3 (Supplementary Fig. 9a). Optimal separability index 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 datasets and find that t-SNE outperforms UMAP and SCVIS (Fig. 4b). Using a single tree displays the lowest performance of all DRMs, while the performance of all DRMs grows exponentially with 2 to 25 trees. Similar results are obtained for scRNA-seq datasets 2 and 3 (Supplementary Fig. 9b).

To evaluate optimal parameter values for good coverage of the projected space by DRMs, we assessed the performance of the six DRMs with varying grid number values (Fig. 4c). We found that all six DRMs perform best with a relatively small grid number, also when applied to dataset 2 and dataset 3 (Supplementary Fig. 9c). Additionally, for dataset 1 with 8 EMT states, Autoencoder, PCA, and t-SNE show similar high performance, but rankings change with an increase in the number of cell types. In dataset 2 and dataset 3, t-SNE, SCVIS, and PCA exhibit the highest performance for a larger number of cell types.

The UI measures a DRM's ability to uniformly spread the projected data in the low dimensional space. All the 6 DRMs perform well for small initial grid sizes (Fig. 3d and Supplementary Fig. 9d), but their performance decreases as the grid size increases. PHATE shows the biggest drop in performance among all 6 DRMs for dataset 1, with similar results for datasets 2 and 3.

Fig. 4e summarizes the relationship between the time order structure index and time steps for various DRMs. No method shows superior evolution dynamics for the various cell types, with all 6 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 performance among the 6 DRMs for reducing the CyTOF dataset 1 and for datasets 2 and 3 (Supplementary Fig. 9e).

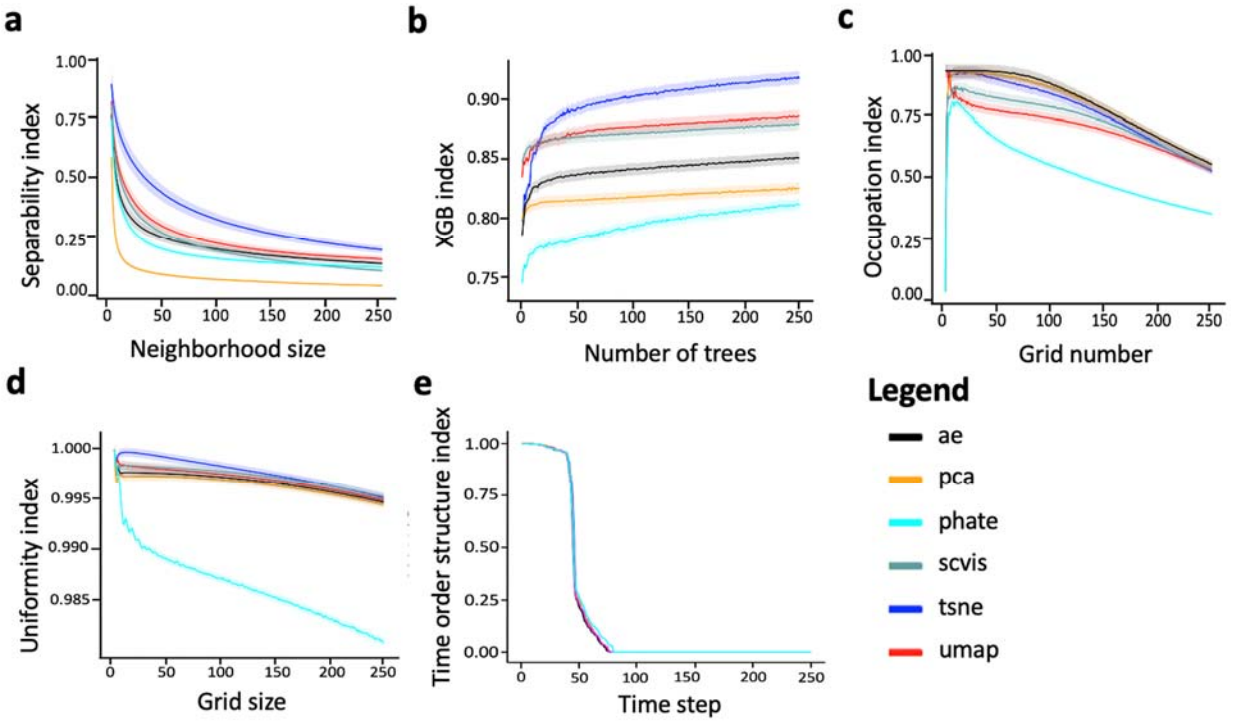



**Fig. 4 MIBCOVIS independent metric analysis for EMT data using 6 DRMs:** Five metric functions with estimated standard error corresponding to the 95% confidence interval are presented, showing the performance accuracy under varying parameters of neighborhood size, number of trees, grid number, grid size, and time steps. **a** For all the 6 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 Uniformity index 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

**MIBCOVIS exhibits strong correlation between metric features.**

To assess the accuracy of dimensionality reduction methods, we use the XGBoost algorithm to minimize the error between a fitted model and observed data. We calculate the gradient boosting classify index for t-SNE, UMAP, PCA, AE, SCVIS, and PHATE applied to three high-dimensional single cell data sets (1-3). To cover the entire accuracy distribution for all methods, we chose the lowest and the highest scores ($S_{min}$, and $S_{max}$) for all 3 single cell data sets. We obtained different ranges of accuracy scores $Y_1 = [0.73, 0.98]$, $Y_2 = [0.62, 0.93]$ and $Y_3 = [0.57, 0.75]$ respectively. By analyzing the accuracy scores of each method, we observe that an increase in the number of cell types decreases the accuracy of DRMs ($Y_1 > Y_2 > Y_3$). Next, we uniformly generate accuracy values of length 250 for each of the ranges $Y_1$, $Y_2$, and $Y_3$ to represent the observed accuracy of a given method. Using Spearman's rank correlation coefficient[38], we find strong correlation between variables that jointly describe the performance of a DRM (Supplementary Figs. 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 Fig. 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 DRMs across different metrics using a joint hierarchical regression Bayesian framework[53]. The model variables include five independent variables (SI,

GI, OI, UI, and TI) and one dependent variable (accuracy scores $Y_1$, $Y_2$, and $Y_3$) for evaluating DRMs for different datasets. The accuracy of a DRM for a given metric is defined as a random variable conditionally dependent on the moderation effects of other metric parameters. 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). The MIBCOVIS posterior distributions from moderation analysis on the log scale show that no single method optimizes all features driving optimal visualization and interpretability (Fig. 5).

We estimate the posterior distribution using the Metropolis algorithm of Markov Chain Monte Carlo (MCMC) [39]. We sample 2000 points from the posterior distribution using 4 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 4 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 4 chains and the model accuracy (Supplementary Figs. 13-17). We construct the 95% High-Density Interval (HDI) [40] of the posterior distribution (Supplementary Figs. 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 (Fig. 5, Supplementary Figs. 33-50). We use the square of absolute values of the standardized regression coefficients[41] to generate box plots quantifying the relative performance of each metric.

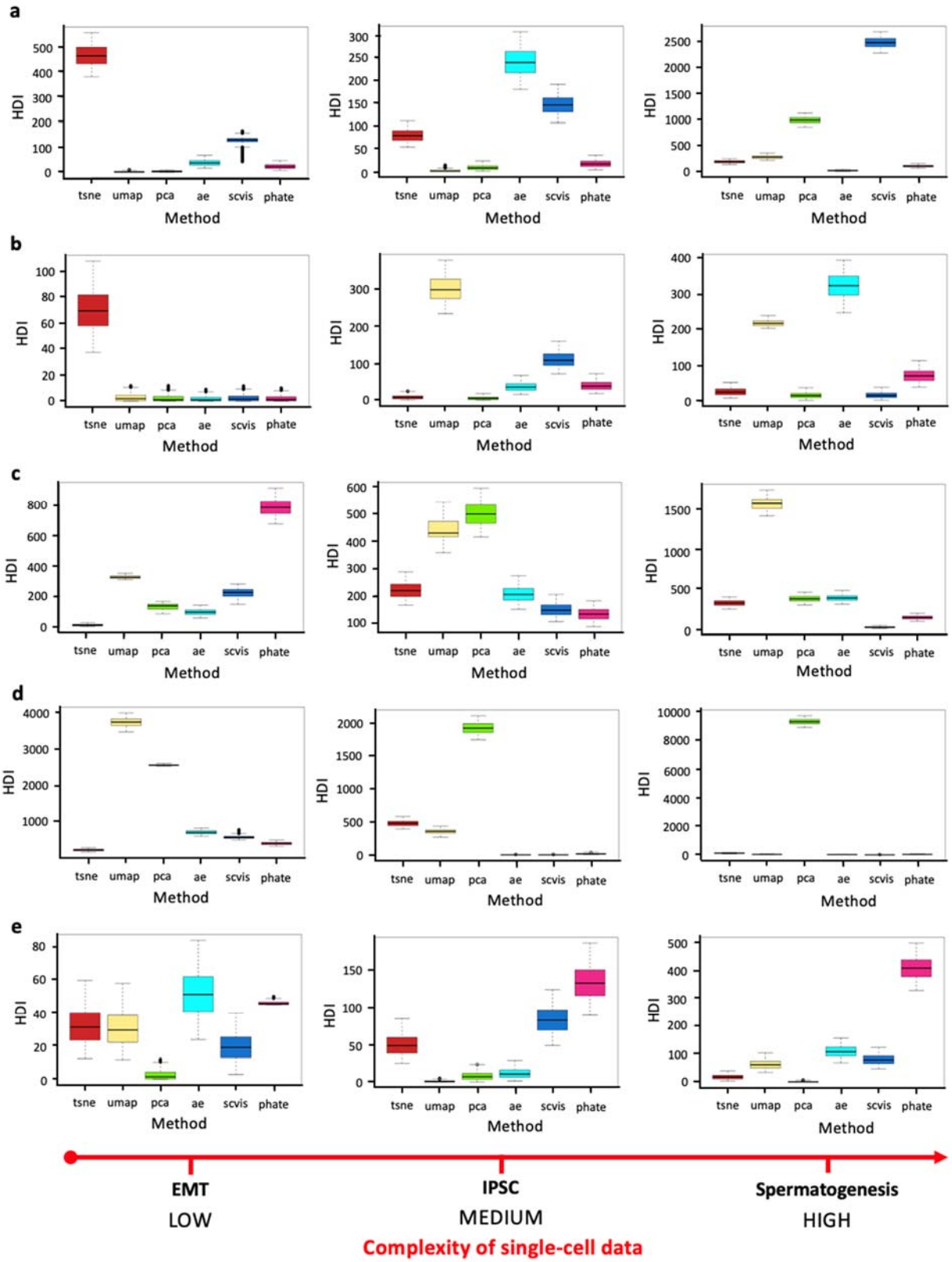



**Fig. 5 Boxplots comparing the performance accuracy distributions of six DRMs using High-Density Interval (HDI) of standardized regression coefficients for EMT, IPSC, and Spermatogenesis data, in order of increasing complexity**. **a** Occupation index (OI) performance varies between 0.1 to 600 for EMT dataset, with t-SNE having higher median values than

UMAP, PCA, PHATE, AE, and SCVIS. For IPSC dataset, 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 values than SCVIS, PHATE, AE, PCA, and t-SNE. HDI range is 2 to 395 for Spermatogenesis data with AE, and UMAP showing higher median values. **c** SI performance for EMT dataset varies from 0.4 to 900 with PHATE outperforming other methods. For IPSC dataset, SI performance varies from 95 to 600 with UMAP, and PCA having highest median value. The SI HDI for Spermatogenesis varies from 1 to 1900 with UMAP having higher medians. **d** MIBCOVIS UI performance for EMT dataset varies from 100 to 4000 with UMAP, and PCA outperforming other methods. For IPSC dataset, 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 dataset. For IPSC dataset, HDI varies from 0.035 to 200 with SCVIS, and PHATE having highest median. PHATE has superior TI performance for Spermatogenesis data with HDI ranges between 0.025 to 500 and highest median of 420.

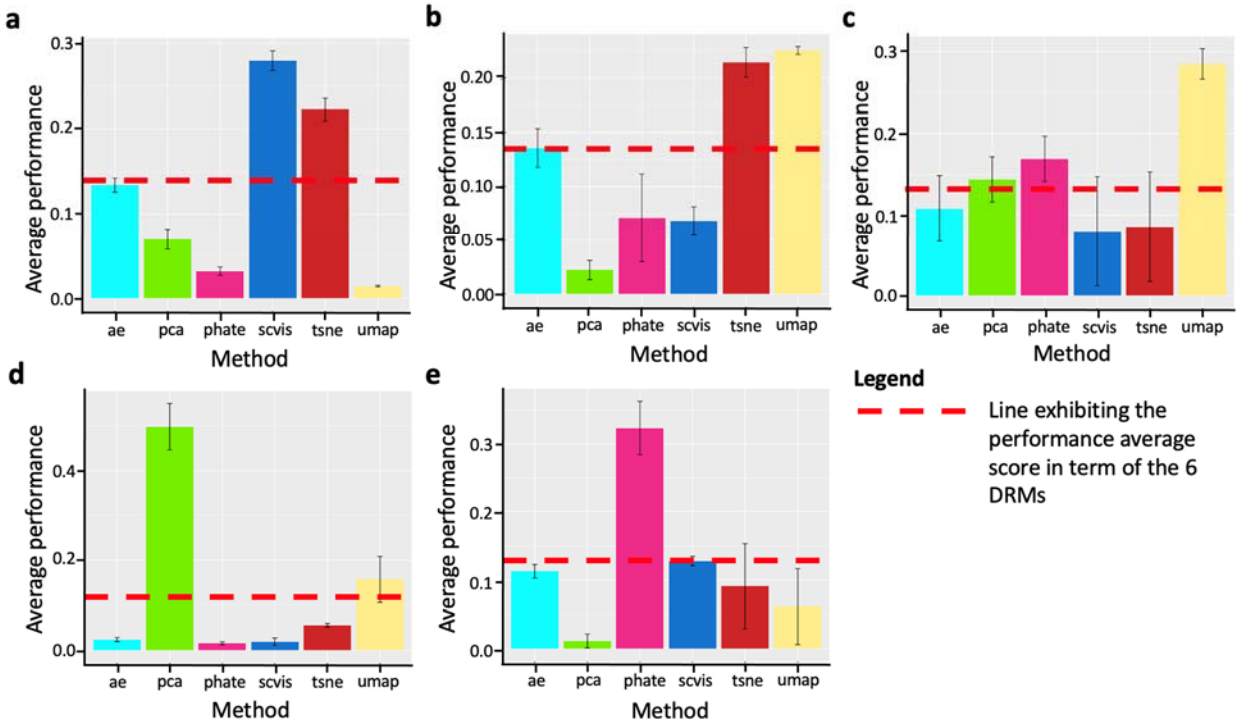



**Fig. 6 MIBCOVIS analysis of the average accuracy (mean of HDI) of DRMs related to the EMT, IPSC, and Spermatogenesis datasets for OVI.** Barplots with error bars summarizing the average performance of **a** the Occupation index, **b** the Gradient boosting classifier index effect, **c** Separability index **d** Uniformity index **e** Time order structure index when 6 DRMs methods are applied to the 3 different biological datasets. The overall average performance score line across all data sets and DRMs is 0.124 (a), 0.143 (b), 0.137 (c), 0.128 (d), and 0.121 (e) respectively.

We compute the overall average performance score per metric by normalizing the HDI scores and averaging the normalized values across DRM and datasets. The red dashed lines in Fig. 6a-e show the average performance scores for OI, GI, SI, UI, and TI, respectively. For instance, based on OI (Figs. 5a, 6a), t-SNE, AE and SCVIS exhibit the highest median performance across all datasets and data complexities, indicating good coverage of the 2D space. Moreover, SCVIS, and t-SNE outperform the average performance score (0.124) for OVI by providing maximum coverage of projected data in 2D space (Fig. 6a).

Regarding GI performance metric (Fig. 5b), t-SNE, UMAP, and AE demonstrate the highest performance on average for high-dimensional datasets, while t-SNE displays good performance for low-dimensional datasets. UMAP presents good performance for high-dimensional datasets. Model averaging (0.143) suggests that t-SNE, and UMAP are the best methods for OVI while minimizing classification bias and variance (Fig. 6b).

To summarize, for SI, while PHATE performs the best for low-dimensional datasets, UMAP operates the best for high-dimensional datasets (Fig. 5c). For SI, UMAP has the best performance in visualizing data with few outliers, and PHATE with PCA also performs well (Fig. 6c). For UI, PCA consistently performs the best across medium and high datasets resulting in a superior above average performance (Fig. 6d), while UMAP performs better with single-cell data of low dimensions. For TI, PHATE consistently outperforms most methods. These findings are shown in Figs. 5 and 6.

## Optimized VAE as a benchmarking tool for MIBCOVIS

Given that none of the six DRMs is superior simultaneously for all visualization features (Figs. 5-6), we propose using an optimized variational autoencoder (oVAE) as a benchmark model for OVI (Supplementary Text 2 and Supplementary Fig. 51). oVAE combines autoencoders, variational autoencoders[42], and contractive autoencoders[43] to improve data projection and and reduce overfitting, while providing a closed framework for 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 OVI (SGVI) using normalized average scores across datasets and methods including oVAE ( 0.142) or without oVAE (0.145) represented by dashed lines in Figs. 7a-b. Compared to non-VAE models (t-SNE, UMAP, PCA, AE, SCVIS, and PHATE), oVAE performs better on average (Fig. 7b and Supplementary Fig. 52-56).

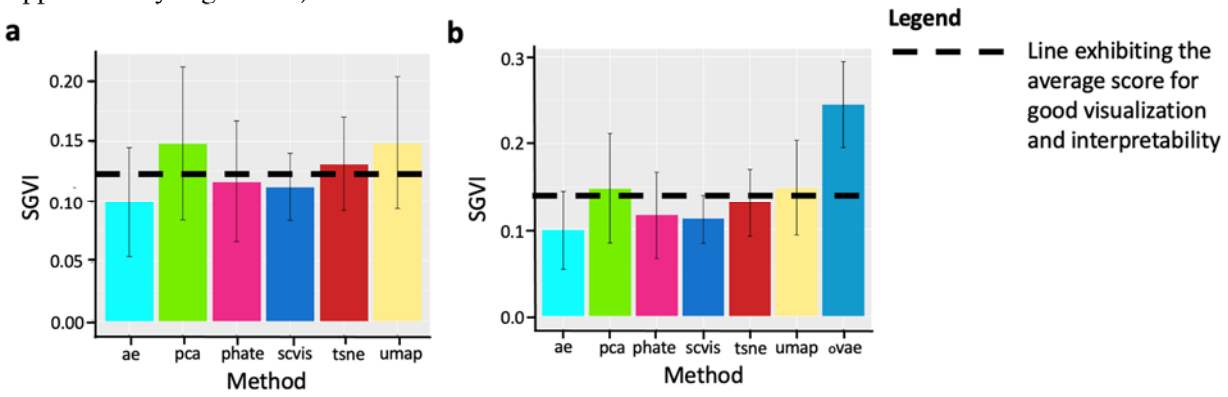



**Fig. 7 Analysis comparing the optimized variational autoencoder (oVAE) with other methods as a benchmarking tool for MIBCOVIS using barplots with error bars derived from an overall average score for OVI (SGVI). A** In methods without oVAE, t_SNE, UMAP, and PCA show above average performance (0.125), with PCA, and UMAP having the highest performances. **B** In methods including oVAE, AE, PHATE, and SCVIS have lower SGVI than the standard (0.142), while PCA, and UMAP display the second-highest performance (SGVI = 0.145). As oVAE has the highest SGVI (0.250), it is the most optimal benchmarking method on average.

## 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 DRMs using both labeled and unlabeled data in a semi-supervised analysis. Using the superior performance of support vector machines (SVM) (Supplementary Figs. 57), the results show significant improvement for many methods under semi-supervised learning (SSL) compared to supervised learning (SL), as seen in the ratio of average to total average performance (Fig. 8). For example, t-SNE improved sensitivity of classification by 21.8%, 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. MIBCOVIS exhibited similar top-ranking outcomes in the SSL analysis as in the SL analysis. (Supplementary Figs. 58-61)

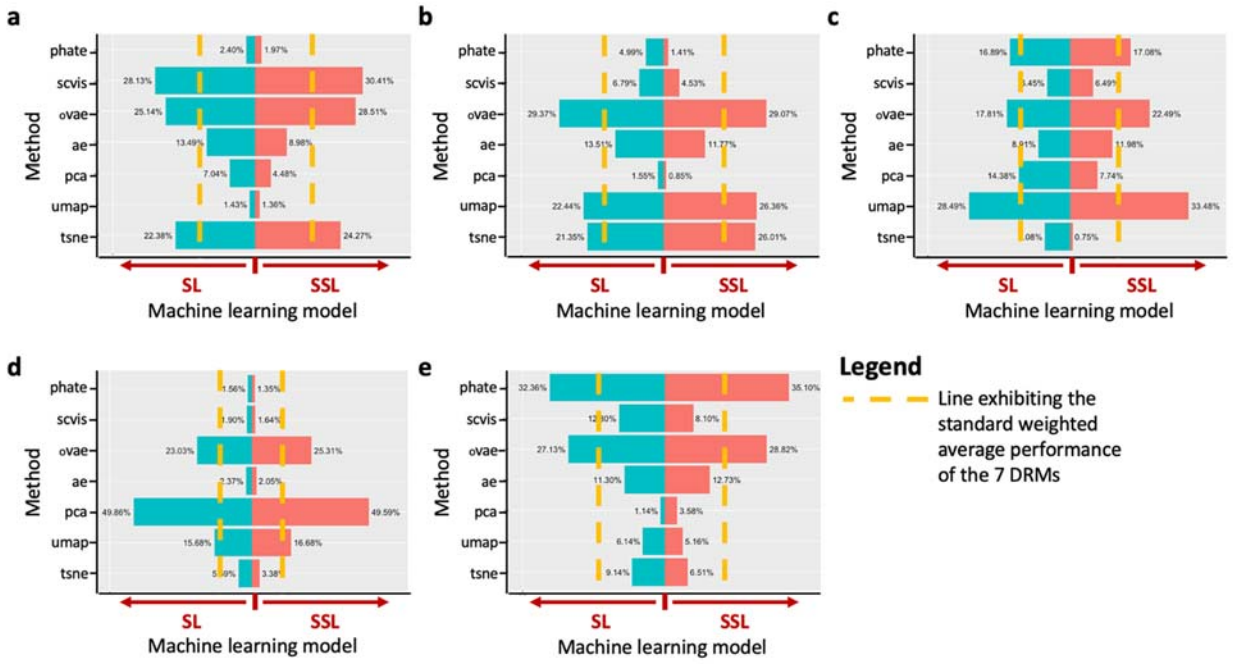



**Fig. 8**. **Weighted average performance of DRMs analysis for MIBCOVIS using Supervised learning (SL) and Semi-supervised learning (SSL).** The performance of each DRM is displayed as a percentage value beside the back-to-back green (SL) and red (SSL) barplots. The standard weighted average performance ($standard_{wp}$) is calculated as the mean of the weighted average performance, which is 14.29% for all 7 DRMs. The histograms display the weighted average performance of DRMs for **a** Occupation index, **b** the Gradient boosting classifier index **c** Separability index. **d** Uniformity index and **e** Time order structure index when applied to 3 biological data sets under both SL and SSL.

## 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 single-cell data is needed to understand complex biological interactions. To reduce data complexity, many dimensionality reduction approaches have been developed. However, methods that are reproducible, preserve data structure, maximize human interpretability, and are robustly benchmarked are needed to optimize visualization. This is especially important in single-cell analysis, where ground truth is often uncertain. Currently, UMAP and t-SNE models are widely used, but variations in projected outputs under different conditions and time points necessitate better approaches for benchmarking and characterizing underlying biological uncertainty.

In this study, we developed MIBCOVIS, a computational framework to assess and compare the performance of DRMs for OVI in terms of data reduction (clustering), visualization, and interpretability accuracy. MIBCOVIS ranks the performance of each method using five different metrics within a Bayesian framework. We optimize for visual accuracy using OI and UI, for data reduction accuracy using GI, and for interpretability accuracy using SI and TI. Although these metrics do not span the entire spectrum of features for OVI, they allow a fast, easy, and interpretable comparison of different methods. For example, we found that t-SNE tends to produce islands in the projected space (Fig. 1a), which may not represent maximum separability of homogeneous clusters (Supplementary text 6, 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 (**Fig. 5e** , Supplementary Figs. 7a-b), as captured by the TI metric and supported by PHENOSTAMP-TRACER[35] EMT visualization output.

TI models time-dependent patterns in data and can handle datasets with any data point ordering. It expands two-dimensional data into a larger matrix to capture how the data changes over time. TI analysis also reveals the relationship between data complexity and cellular dynamics (Fig. 9b), indicating the need for further investigation to identify key timepoints and cells driving high variable transitions. PHATE performs best using diffusion theory, which governs the movement of molecules and ions within and between cells[44]. However, the lack of variability between methods in the non-stationarity part of the TI Markov chain suggests a need for better visualization methods that account for spatio-temporal variations.

This study's benchmarking approach explains observed differences in data visualization by considering the correlation and moderation effects of model parameters defining an OVI's features. Performance is evaluated using five-feature metrics and applied to linear, non-linear, and neural network methods on three single-cell datasets. To account for uncertainty in the ground truth, a semi-supervised approach is developed for benchmarking. Semi-supervised MIBCOVIS allows optimal visualization of cell clusters based on gene expression patterns and supervised MIBCOVIS analysis can be used for selecting optimal methods for classifying samples into different subtypes based on known clinical outcomes. Simulations can also be used to compare different methods with synthetic datasets, but simple data structures like in Saelens et al. 2019[20] are insufficient for representing complex biological processes.

MIBCOVIS enables users to visualize and inteprete specific clustering features while accounting for data complexity. Fig. 9c 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 (Fig. 9a), 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. This study also shows the power of Neural Networks for learning latent features and its potential for data integration, as demonstrated in Karacosta et al[35].

MIBCOVIS has potential for non-visualization applications to leverage on the moderation effect framework and the weighted TI's Markov process. Benchmarking ($\geq$3D) spatio-temporal data reduction methods for optimal visualization and interpretability is a valuable extension of this work. In summary, MIBCOVIS provides a robust multivariate metric and unbiased benchmarking framework for OVI of spatio-temporal relationships.

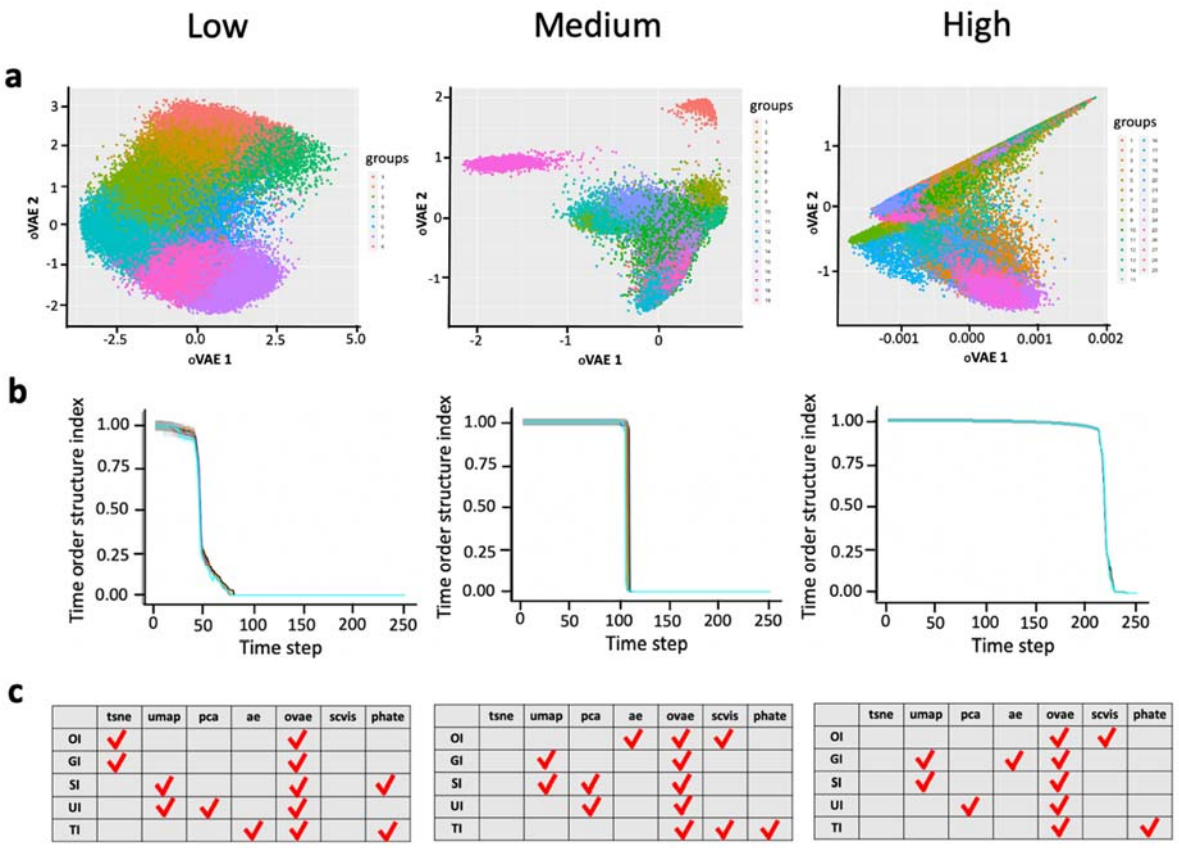



**Fig. 9. Summary of optimal visualization and intepretability of methods based on data complexity and feature accuracy.** **a** The data reduction maps illustrate the application of the optimal Variational Autoencoder (oVAE) to three distinct high-dimensional single-cell datasets: EMT data (Low), iPSC data (Medium), 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 by dimension size and number of cell types. Low, Medium, and High complexity is determined from analysing EMT, IPSC, and spermatogenesis datasets 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 DRM seem 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 datasets (EMT, IPSC, and spermatogenesis, respectively).

## Method

MIBCOVIS uses a set of metrics to evaluate the performance of DRM regarding OVI features. These metrics include the Separability index (SI), Gradient boosting classifier index (GI), Occupation index (OI), Uniformity index (UI), and Time order structure index (TI).

### Separability index

The separability index in this study quantifies the average number of instances in a dataset with a nearest neighbor that shares the same label[28]. Fig. 1d shows the separability parameters between two classes in a two-dimensional feature space. To determine the SI, we consider data points $x^k$ with coordinates $x^k = (x_1^k, x_2^k, \cdots, x_{D_r}^k)$ in the reduced data space of size $D_r$, which can be assigned to one of $k$ classes, $\Psi_k = \{\varphi_1, \varphi_2, \cdots \varphi_k\}$. For each data point $x_i$ belonging to class $\Psi_k$, we calculate the number of nearest neighbors based on Euclidean distance and determine the proportion of neighbors that belong to the same class $k$ .

The size of the nearest neighbor is a critical parameter of separability[45] as the SI is affected by the dominance of nearest neighbors belonging to the same labeled class in feature space. The resulting indicator function ($\xi$) given below for each point $x_i$ within a given neighborhood takes values 0 or 1.

$$\xi(x_i, x_j^k) = \begin{cases} 1, & if\ x_i, x_j^k\ \in\ \Psi_k \\ 0, & otherwise \end{cases}$$

Hence, from a total number of $Q_l$ nearest neighbor of cells for each cell $l$ we compute the average proportion ($\rho_k$) of neighbors that come from the same class $k$ as

$$\rho_k = \frac{\sum_{i=1}^{D_r} \sum_{l=1}^{Q_l} \frac{\xi(x_l, x_i^k)}{Q_l}}{D_r}$$

To compute the Separability index of the projected 2D data, a weighted average of the separability of different cells across all classes $\Psi_k$ is denoted as follows using a given set of 250 numbers of nearest neighbors ($Q_l = [1, 250]$);

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

where $|\Psi_k|$ refers to the total amount of classes. The *SI* provides output values defined between 0 and 1.

**Occupation index**

To measure the coverage of the projected space, a grid-based approach is used, where the window of the projected space is divided into tiles, and the number of tiles with at least one data point is counted. Let $x_0$, $x_{1,} \cdots\ x_{D_r}$ and $y_0, y_{1,} \cdots\ y_{D_r}$, be the split points of tiles on each axis (Fig. 2b) with $\alpha_{i,j}$ the number of points in the grid tile $[x_i, x_{i+1}] \times [y_i, y_{i+1}]$, and $D_r$ the total number of points in the projected space. The OI is expressed as the number of tiles with at least one point divided by the total number of tiles in the projected space given as;

$$Cov_p = \frac{\sum_{i=0}^{x_{D_r}-1} \sum_{j=0}^{y_{D_r}-1} \alpha_{i,j}}{Dr^2}$$

Since $\alpha < D_r$, the $Cov_p$ provides output values in the range 0 and 1.

We estimate high dimensional data coverage by approximating it as the volume of a hyper-rectangle, which can be factorized using the singular value decomposition of a matrix (*D*). This decomposition expresses *D* as the product of two matrices[46]:

$$D = U \sum V$$

where *U,* and *V* are orthogonal (respectively of size $o_1 \times o_2$ and $o_2 \times o_2$) and the variance $\sum = Diag(\sigma_1, \sigma_{2,} \ldots)$ with $\sigma_1 \geq \sigma_{2,} \geq \cdots \geq 0$[47] . The singular values $\sigma_1, \cdots, \sigma_{o_2}$ of the matrix $D$ of the high dimensional space is the square roots of the eigenvalues such that:

$$\sigma_i = \sqrt{\lambda_i}$$

The coverage of the high dimensional space ($Cov_h$) is determined by the total variance as:

$$Cov_h = \prod_{i=1}^{o_2} \sqrt{\lambda_i}$$

The absolute value of any eigenvalue of a stochastic matrix is less than or equal to 1[30]. In consequence, the total variance $Cov_h$ is less or equal to 1. We introduce κ, a weighting parameter proportional to the surface area per summary coverage. It represents the average proportion of surface area occupied by data defined by Kufer S. (2019) as 'surface area per summary'[30] describing in our case the average ratio of potential data region to grid area given all data points and number of grids. Given a projected data, the smaller the

'surface area per summary', the more the incorporated empty space is minimized as well as the probability of data features overlapping each other. Relying on recommendations by Kufer S. (2019)[30], we choose $\kappa = 0.33$ as the smallest surface area per summary to maximize coverage of the high dimensional data space on the projected 2D window.

The occupation index (*OI*) is described mathematically by the weighted product of the coverage of the projected and the high dimensional spaces[48] given as:

$$OI = 0.33 * Cov_p . Cov_h$$

Since the coverage in the projected space ($Cov_p$) is also less or equal to 1, the *OI* provides output values defined between 0 and 1.

**Gradient boosting classifier index**

GI uses a Gradient Boosting Classifier for a multinomial model which is a generalization of the binary case, and it involves fitting a separate binary classification tree for each class. Given our projected set of input features $E = \left[\left(x_{i,j}\right)_{i=1}^{2}, y_j\right]_{j=1}^{D_r}$ where each point $y_j$ corresponds to a class *k*=1:K, the likelihood of a cell belonging to a class *k* is given by:

$$P(y = k|x) = \frac{e^{f_k(x)}}{\sum_{k=1}^{K} e^{f_k(x)}}$$

where *K* is the number of classes, and $f_k(x)$ is the output of the $k^{th}$ class in the gradient boosting algorithm, which is trained to minimize the negative multinomial log-likelihood loss function:

$$L(y, f) = -\sum_{j=1}^{D_r}\sum_{k=1}^{K} y_{jk} \log\left(P\left(y_{jk} = k|x_j\right)\right)$$

where $y_{jk}$ is a binary indicator variable that is 1 if the $j^{th}$ observation belongs to class *k*, and 0 otherwise. The algorithm uses gradient descent to iteratively update the weights of each tree in the ensemble, in order to minimize the loss function. At each iteration, a new tree is added to the ensemble, which is trained to predict the negative gradient of the loss function with respect to the current predictions. The learning rate controls the step size of the gradient descent updates, and the maximum depth and number of trees are hyperparameters that can be tuned to optimize performance. We fit a Gradient Boosting classifier on the low-dimensional projection of the data ($E$) to predict the label of each point, and count the fraction of correct predictions as a value of GI. We use the extreme gradient boosting (xgboost) R package[31] to implement the gradient boosting classifier index.

**Uniformity index**

We adapt the goodness fit of Pearson Chi-Squared ($\chi^2$) test of uniformity of the data distribution to estimate the uniformity index (UI). In this test, the space is divided into $\pi$ grids with Π the total number of grids (Fig. 1c), determined by the grid size, defined as the length of each side of the square grid. The number of points is counted in each grid. Under the null hypothesis that the points are uniformly distributed, the statistic test is defined as:

$$\Upsilon = \sum_{\pi=1}^{\Pi} \frac{(O_\pi - E_\pi)^2}{E_\pi}$$

Where $O_\pi$ is the actual number of points in each grid, with $D_r$ the total number of points in the projected space, and Π number of grids. The expected number of points in each grid is defined as $E_\pi = D_r/\Pi$. It is expected that if the projected data is distributed uniformly, the p-values should be extremely large. Recall that the p-value is the probability of obtaining results at least as extreme as the number of observed points, assuming that the null hypothesis is correct. Hence, the level of Υ gives us an indication on how large there is deviation from uniformity which is indirectly related to the grid size through $E_\pi$. Since a decrease in grid size results in an increase in number of grids Π and varying the grid size from 1 to 250, we observe that $\Pi > \Upsilon$ , without loss of generalities. This allows us to define the uniformity index (UI) as:

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

which produces values defined in the range 0 and 1.

**Time order structure index**

Single-cell experiments capture snapshots of heterogeneous dynamic cell populations over time (Fig. 2a), sampled by sequencing single cells at multiple stages. High dimensionality reduction algorithms should account for temporal variations in both high and projected space to model time dependency independent of underlying data structure[49]. Given an ordered projected reduced space (2D) metric map, TI uses markov processes to quantify the time dependency of cells condition on the projected ordered with community structure[50] based on temporal networks.

***Background on Markov process***

A Markov chain is a process in which random choices are made among a finite (or enumerable) set of states at each time-step. The transition probability can be represented by a matrix $P = (p_{ij})$, where $p_{ij}$ is the probability of moving from state $s_i$ to state $s_j$, denoted as *i* and *j*, respectively, at time $t_r$. For homogeneous chains, these probabilities are stationary[50] and do not depend on time *t*. Markov chain estimation evaluates the dynamic evolution of cells based on a mixture of cells in different states with varying state compositions over time. It assumes that the sequence of variables has "no-past-effect," meaning that the distribution of future states is independent of past states when the present state is known. The evolution of a random time-dependent process $x(t)$, assuming a probability of transforming from state *i* to state *j* at time $t$+1 of $p_{ij}$ and a probability of being in state i at time t of $\gamma_i(t)$, is captured by:

$$\gamma_i(t+1) = \sum_{i=1}^{w} \gamma_i(t) p_{ij} \quad (i = 1, 2, \cdots, w)$$

Then, the initial distribution $\gamma_1$, together with the transition matrix $P$, determines the probability distribution for any state at all future times. $P$ defined as:

$$P = \begin{bmatrix} p_{11} & \cdots & p_{1w} \\ \vdots & \ddots & \vdots \\ p_{w1} & \cdots & p_{ww} \end{bmatrix}$$

is nonnegative square matrix with unit row sums that is, $0 \leq p_{ij} \leq 1$, $\sum_j p_{ij} = 1$ (stochastic matrix) for every $j \in G$ (discrete state space $\{1, \cdots, w\}$)[51].

***Dynamics of cellular state***

The typical output data from the dimensionality reduction methods (projected data) is a non-square $u \times v$ data matrix $E$ composed of $u$ rows and $v$ columns:

$$E = \begin{bmatrix} e_{11} & \cdots & e_{1v} \\ \vdots & \ddots & \vdots \\ e_{u1} & \cdots & e_{uv} \end{bmatrix}$$

In our case, $u \gg v$ with $v = 2$. Since $E$ is large and ordered, we assume $E$ is a combination of smaller submatrices ordered linearly by time points. We partition $E$ to create several consecutive square $2 \times 2$ submatrices for Markov chain manipulations. Specifically, we partition adjacent rows of $E$ to create $u - 1$ square submatrices while maintaining consecutive links between the squared submatrices in terms of their row ordering (Fig. 2c).

Let's consider the first 2 rows of the projected data as 2 different initial states of cellular motion within a given temporal trajectory of cells (Fig. 2d). The model shows the two possible states (state 1 modeled by cell 1, and state 2 modeled by cell 2). $S = \{s_1, s_2\}$ for $s_1$, and $s_2$ respectively equivalent to 0, and 1, Fig. 2e, their emissions, and probabilities for transition between them (state $i$ to state $j$ ) such that $\sum_{j=1}^{2} p_{ij} = 1$ as:

$$P_1 = \begin{pmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{pmatrix}$$

Where, $p_{11} = \frac{e_{11}}{e_{11}+e_{12}}$, $p_{12} = \frac{e_{12}}{e_{11}+e_{12}}$, $p_{21} = \frac{e_{21}}{e_{21}+e_{22}}$, and $p_{22} = \frac{e_{22}}{e_{21}+e_{22}}$

In general, we let expression $e_{ij}$ denote the coordinate of cells which were in state $i$ in period $t - 1$ and are in state $j$ in period $t$. We next estimate the probability of cells being in state $j$ in period $t$ given that they were in state $i$ in period $t - 1$, denoted by $p_{ij}$, using the following formula[52]:

$$p_{ij} = \frac{e_{ij}}{\sum_j e_{ij}}$$

The cell states are varying discretely with time so, for a given data, there is a finite number $W = u - 1$ of possible states in the state space $G$ such that: $G = \{1, \cdots, W\}$. Hence, based on the first original $2 \times 2$ square matrix $P_1$, and by choosing the initial state as cell 1, we uniformly generate a $W \times W$ square matrix $P_{W1}$ (Fig. 2d) such that:

$$P_{W1} = \begin{pmatrix} p_{11} & \cdots & p_{1W} \\ \vdots & \ddots & \vdots \\ p_{W1} & \cdots & p_{WW} \end{pmatrix}$$

with the conditions that $P_{W1} = (P_{i,j})$, $P_{i,j} \geq 0 \ \ \forall i, j$ and $\sum_{j=1}^{W} P_{ij} = 1$. It has the advantage to include all the transition states of the remaining square submatrices $P_2, P_3, \cdots P_{u-1}$. Example of a Markov chain with 5 states (cells) is illustrated in Fig 2f.

The integer number of the finite states $W$ defines the number of sequence state variables in the Markov chain between the starting state to the final state in the state space $G$. $W$ summarizes the different possible state transitions in the projected space.

By individually defining cell 2 to cell $u - 1$ as the initial state, the same procedure as above is applied to the consecutive 2 x 2 submatrices $P_2, P_3, \cdots P_{u-1}$ to generate $u - 1$ number of different $W \times W$ square matrices $P_{W2}, P_{W3}, \cdots, P_{Wu-1}$. In order to identify the temporal structure or time dependencies in the $W \times W$ square matrices $P_{W1}, P_{W2}, \cdots, P_{Wu-1}$, we proceed with the following three steps.

**Step 1**: given $m$ as the initial state of cell $m$, we compute the expected number of times a cell $m$ has visited other states in the Markov chain (Fig. 2g). Let $C_{ij}$ be the expected number of times (over all time steps) that the Markov chain visit state $j \in G$ for $G = \{1, \cdots, W\}$ given the initial state is $i \in G$.

$$C_{i,j} = E\left\{\sum_{n=0}^{\infty} I\{W_n = j / W_0 = i\}\right\} = \sum_{n=0}^{\infty} P_{ij}^n$$

$C = (C_{ij})$ is a $W \times W$ matrix. Let $I$ denote the $W \times W$ identity matrix. Then $C = I + P_W C$ yielding the solution $C = (I - P_W)^{-1}$.

Relying on the $W \times W$ matrix $P_{W1}$, we compute the expected number of times cell *1* at state 1 visited each of the $W$ different states in the Markov chain. These results in a $1 \times W$ vector $C_1$ of integers such that $C_1 = [c_{11}, c_{12}, \cdots, c_{1W}]$. We extend this to generate $C_{u-1}$ vectors corresponding to all the other $P_{W2}, \cdots, P_{Wu-1}$ matrices. For example, by choosing $P_{W2}$, we compute the expected number of times cell *2* at state 2 visited the other $W$ different states in the Markov chain to engender $C_2 = [c_{21}, c_{22}, \cdots, c_{2W}]$. We then collect all the row vectors $C_1, C_2, \cdots C_{u-1}$ computed from the $W \times W$ matrices respectively $P_{W1}, P_{W2}, \cdots, P_{Wu-1}$ ensemble together to give the $W \times W$ matrix $C$, such as:

$$C = \begin{pmatrix} c_{11} & \cdots & c_{1W} \\ \vdots & \ddots & \vdots \\ c_{u-1w} & \cdots & c_{u-1W} \end{pmatrix}$$

The matrix $C$ will serve as input in the next step.

**Step 2**: we weight each row of the matrix $C$ by the proportion of associated label of cell types in the reduced data (Fig. 2g). By way of illustration, let $k$ be a distinctive cell type in the labeled vector $C_{type} = (k_j)_{j=1}^{u-1}$ such that $k \in C_{type}$ , and $K$ stands for the finite number of distinct cell types. We then, outline $Num_k = |k_j|$, the total number of a particular cell type $k$ in the label vector $C_{type}$. Note that $Num_k < |C_{type}|$, where $|C_{type}|$ refers to the total amount of cells in $C_{type}$. The possibility $(s_{ij}|k_i \in C_{type})$ that a cell moves from its state $s_i$ at time *t-1* to the state $s_j$ at time *t* given that it belongs to label $k$ is described as:

$$P_{ij}(s_{ij}|k_i \in C_{type}) = P_{ij} \times Prob_k$$

with

$$Prob_k = \frac{Num_k}{|C_{type}|}$$

where $Prob_k$ represents the possibility that a cell belongs to a specific cluster of cell type $k$ during the transition.

Accounting for different cell types or states during transition, the weighted matrix $C$ in terms of the expected number of times a cell visits all W states in the reduced data is given as:

$$C = Prob_k \times \left[(C_{k_i,j})_{i=1}^{u-1}\right]_{j=1}^{W}$$

With $u - 1$ equal to the number of rows of the matrix $C$ computed in Step 1.

**Step 3**: We compute the cumulative sum of the weighted matrix $C$. It is used to display the cumulative distribution function of a given variable over time for example. Also, we calculate the column means of $C$ and return them in a row vector of length $W$ with entry $\mu_j$. We use the mean to estimate the expected variance ($V_j$) that describes the spread (amount of variability) of data around the expectation (mean $\mu_j$) such that for each column $j$ of the weighted matrix $C$:

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

Convergence and stationarity are theoretically linked for time-series forecasts, as both imply stability or equilibrium[53]. The Time Order Structure Index (TI) evaluates the stationary distribution of a Markov chain of W steps, where the variance converges to equilibrium. The sooner the variance ($V_j$) converge to equilibrium, the less time-dependent the outcomes of Dimensionality Reduction Methods (DRMs) are. To make TI comparable to other metrics, we divide the length of the Markov chain corresponding to the vector of the variance ($l_V$) by a regular interval $\tau$ such that $l_V/\tau = 250$. For each interval $\tau$ index by $\theta$ as $\tau_\theta$ ($\theta \in$

$[1, 250]$), we count the number of time ($n_\theta$) the variance does not converge to equilibrium. In a nutshell, we illustrate the *TI* as:

$$TI = \frac{n_\theta}{N_{step}}$$

where $n_\theta$ represents the number of steps in each interval $\tau$ which display a variance different to the equilibrium with $N_{step} = INT[(u-1)/250\ ]$, the integer total number of steps in each interval.

**MIBCOVIS benchmarking framework**

We next describe the various steps used by **MIBCOVIS** for evaluating the performance accuracy of a DRM in terms of a optimal visualization and interpretation (OVI).

**Step 1: Metric set identification and parameter evaluation**

**MIBCOVIS** starts with constructing a set X of measurable performance metrics associated with five different features characterizing features of OVI (Fig. 3b). We define the metric set as $X = \{X_i\}_{i=1}^{5}$ where variables $X_1, X_2, X_3, X_4,$ and $X_5$ encodes the SI, the GI, the OI, the UI, and the TI feature respectively. The detailed process is described in Supplementary text 4 and Supplementary Fig. 62. We next show how the optimal parameter values for each metric used by **MIBCOVIS** is determined.

**Step 2: Correlation between metric features and method accuracy**

This step assesses the relationship between the parameters of the metrics and the accuracy of DRMs in performing single cell data dimension reduction as a function of the labeled data (Fig. 3c). We use the Spearman correlation coefficient, which represents the strength of non-linear association between the features of the metrics and the performance of a DRM.

**Step 3: Bayesian conditional effect regression model**

We use a Bayesian conditional effect regression model for a beta family to estimate the size of the conditional effect of a feature of a metric on the continuous accuracy variable (*Y*) associated with OVI . This is formulated as follows:

Assuming that the response variable (accuracy) $Y$ follows a beta distribution with mean and dispersion shape parameters. Given the parameters $\varphi$ and $\beta$, corresponds to the dispersion and regression coefficients, the conditional expectation of $Y$ given the predictor (moderator) variables $X = (X_1, X_2, X_3, X_4, X_5)$ is modeled using a linear regression model:

$$E(Y/X) = g(X\beta)$$

where $g$ is a link function that maps the linear predictor $X\beta$ to the interval [0,1]. We use logistic function given below as

$$g(X\beta) = \frac{e^{X\beta}}{1 + e^{X\beta}}$$

The conditional variance of $Y$ given $X$ is modeled using a generalized linear model with a log-link function:

$$Var(Y/X) = \varphi V\left(g(X\beta)\right)$$

where $\varphi$ is a dispersion parameter and V is a variance function that relates the conditional variance to the mean.

We assune the regression coefficients $\beta$ follow a student t distribution with 3 degrees of freedom, the prior for the dispersion parameter $\varphi$ follows an inverse-gamma distribution and the residual error variance follows a half Cauchy prior. The posterior distribution of the parameters can be obtained using Bayes' theorem:

$$p(\beta, \varphi | Y, X) \propto p(X, \beta, \varphi) p(\beta) p(\varphi)$$

where $p(Y|X, \beta, \varphi)$ is the likelihood function that describes the probability of observing the data $Y$ given the parameters $\beta$ and $\varphi$, and $p(\beta)$and $p(\varphi)$are the prior distributions of the parameters.

We use Markov Chain Monte Carlo (MCMC) methods to draw samples from the posterior distribution. The posterior samples are used to estimate the high density intervals (HDI) for performance accuracy of a DRM. In the practice, we use the main function brm of the brms R package which implements the logit conditional beta regression model[54]. We provide performance accuracy scores in log scale.

**Step 4: MIBCOVIS uses standardized regression coefficients for benchmarking**

We estimate the 95% highest density interval (HDI)[55] of the posterior distribution $p(\beta/Y)$ for each regression coefficient. Further details can be found in Supplementary text 5 and Fig. 63. Condition on this interval, the performance accuracy of each DRM, while accounting for variability in clustering accuracy, visual accuracy and interpretability accuracy is defined as $Y_{HDI} = \left(\frac{\widehat{\beta}}{sd(\widehat{\beta})}\right)^2$, the square of the conditional standardized regression coefficients[56]. MIBCOVIS uses boxplots of $Y_{HDI}$ to rank the performance of various DRMs.

**Variability of cluster centroids**

Visual interpretability in single-cell data relies on deviations of points from cluster centers. Cluster centroids are used to measure cluster location and interpret variability. Centroids represent the average observation within a cell cluster and the maximum distance from observations to the centroid measures variability within the cluster. To verify if dense regions generated by data reduction methods truly indicate cluster separability, we compute the distance between cluster centroids. A loop is created over clusters $c_k (k = 1, \cdots, K)$ and the centroid of a cluster $k$ $(c_k)$ is calculated as the mean of the data points, $x_1, x_2, \cdots$ assigned to that cluster given as ;

$$c_l = \frac{1}{Num_k} \sum\nolimits_{j \in c_k} x_j$$

where $Num_k = |k_j|$ denotes the number of vectors assigned to the cluster $c_k$. Hence, for cluster 1, and 2 for example, the distance between their respective centroids $c_1$, and $c_2$ is quantified as $\|c_2 - c_1\|^2$ [57].

**Data availability**

The spermatogenesis dataset with about 110,000 cells across 25 developmental stages used in this study was processed from published raw scRNA-seq datasets for mouse spermatogenesis from Gene Expression Omnibus (GEO) under accession codes GSE121904[58], GSE124904[37], and GSE117707[59] and from the ArrayExpress database under accession code EMTAB-6946[60]. The processed IPSC dataset with about 50000 cells across 12 time points was obtained from raw scRNA-seq datasets for mouse MEFs and chemically induced pluripotent stem cells from GEO under accession code GSE114952[36]. 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. 2022[59]. We also downloaded arcsinh-transformed CyTOF time-course data for the EMT analysis from[35]. The source single cell data including EMT data, IPSC data and Spermatogenesis use in this study are provided as a Source Data file as well as 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 OVI using Bayesian multilevel modeling. While traditional benchmarking methods define the performance of a method as independent variable for OVI, the MIBCOVIS computational framework models the performance of methods as a multivariate function depending on conditional effect of multiple factors which illustrate OVI. MIBCOVIS including all scoring functions for various metrics is implemented using R available on Github (https://github.com/NIEHS/MUBCOVID.git).

**Reference**


1. Linderman GC, Rachh M, Hoskins JG, Steinerberger S, Kluger Y. Fast interpolation-based t-SNE for improved visualization of single-cell RNA-seq data. *Nature methods* **16**, 243-245 (2019).

2. Becht E*, et al*. Dimensionality reduction for visualizing single-cell data using UMAP. *Nature biotechnology* **37**, 38-44 (2019).

3. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. *Nature genetics* **38**, 904-909 (2006).

4. Haxby JV, Gobbini MI, Furey ML, Ishai A, Schouten JL, Pietrini P. Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science* **293**, 2425-2430 (2001).

5. Olshausen BA, Field DJ. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature* **381**, 607-609 (1996).

6. Guyon I, Elisseeff A. An introduction to variable and feature selection. *Journal of machine learning research* **3**, 1157-1182 (2003).

7. Kohavi R, John GH. Wrappers for feature subset selection. *Artificial intelligence* **97**, 273-324 (1997).

8. Ball NM, Brunner RJ. Data mining and machine learning in astronomy. *International Journal of Modern Physics D* **19**, 1049-1106 (2010).

9. Haut JM, Paoletti ME, Plaza J, Plaza A. Fast dimensionality reduction and classification of hyperspectral images with extreme learning machines. *Journal of Real-Time Image Processing* **15**, 439-462 (2018).

10. Heiser CN, Lau KS. A quantitative framework for evaluating single-cell data structure preservation by dimensionality reduction techniques. *Cell reports* **31**, 107576 (2020).

11. Sun S, Zhu J, Ma Y, Zhou X. Accuracy, robustness and scalability of dimensionality reduction methods for single-cell RNA-seq analysis. *Genome biology* **20**, 1-21 (2019).

12. Butler A, Hoffman P, Smibert P, Papalexi E, Satija R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature biotechnology* **36**, 411-420 (2018).

13. Powers DM. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. *arXiv preprint arXiv:201016061*, (2020).

14. Moon KR*, et al.* Visualizing structure and transitions in high-dimensional biological data. *Nature biotechnology* **37**, 1482-1492 (2019).

15. Kullback S, Leibler RA. On information and sufficiency. *The annals of mathematical statistics* **22**, 79-86 (1951).

16. Breiman L. Random forests. *Machine learning* **45**, 5-32 (2001).

17. Dries R*, et al.* Giotto, a pipeline for integrative analysis and visualization of single-cell spatial transcriptomic data. *BioRxiv* **701680**, (2019).

18. Habib N*, et al.* Massively parallel single-nucleus RNA-seq with DroNc-seq. *Nature methods* **14**, 955-958 (2017).

19. Duò A, Robinson MD, Soneson C. A systematic performance evaluation of clustering methods for single-cell RNA-seq data. *F1000Research* **7**, 1141 (2020).

20. Saelens W, Cannoodt R, Todorov H, Saeys Y. A comparison of single-cell trajectory inference methods: towards more accurate and robust tools. *BioRxiv*, 276907 (2018).

21. Anchang B*, et al.* Visualization, benchmarking and characterization of nested single-cell heterogeneity as dynamic forest mixtures. *Briefings in Bioinformatics* **23**, bbac017 (2022).

22. Kosslyn SM. *Graph design for the eye and mind*. OUP USA (2006).

23. De Graef M, McHenry ME. *Structure of materials: an introduction to crystallography, diffraction and symmetry*. Cambridge University Press (2012).

24. Sun F, Wang Y, Xu H. Uniform projection designs. *The Annals of Statistics* **47**, 641-661 (2019).

25. 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*) (2016).

26. Fang L*, et al.* Feature selection method based on mutual information and class separability for dimension reduction in multidimensional time series for clinical data. *Biomedical Signal Processing and Control* **21**, 82-89 (2015).

27. Samusik N, Good Z, Spitzer MH, Davis KL, Nolan GP. Automated mapping of phenotype space with single-cell data. *Nature methods* **13**, 493-496 (2016).

28. Mthembu L, Marwala T. A note on the separability index. *arXiv preprint arXiv:08121107*, (2008).

29. Gentle JE. Matrix algebra. *Springer texts in statistics, Springer, New York, NY, doi* **10**, 978-970 (2007).

30. Kufer S. *Effective and Efficient Summarization of Two-Dimensional Point Data: Approaches for Resource Description and Selection in Spatial Application Scenarios*. University of Bamberg Press (2019).

31. Chen T*, et al.* Xgboost: extreme gradient boosting. *R package version 04-2* **1**, 1-4 (2015).

32. Ovchinnikova S, Anders S. Exploring dimension-reduced embeddings with Sleepwalk. *Genome research* **30**, 749-756 (2020).

33. Kinalis S, Nielsen FC, Winther O, Bagger FO. Deconvolution of autoencoders to learn biological regulatory modules from single cell mRNA sequencing data. *BMC bioinformatics* **20**, 1-9 (2019).

34. Rashid S, Shah S, Bar-Joseph Z, Pandya R. Dhaka: variational autoencoder for unmasking tumor heterogeneity from single cell genomic data. *Bioinformatics* **37**, 1535-1543 (2021).

35. Karacosta LG*, et al.* Mapping lung cancer epithelial-mesenchymal transition states and trajectories with single-cell resolution. *Nature communications* **10**, 1-15 (2019).

36. Zhao T*, et al.* Single-cell RNA-seq reveals dynamic early embryonic-like programs during chemical reprogramming. *Cell stem cell* **23**, 31-45. e37 (2018).

37. Law NC, Oatley MJ, Oatley JM. Developmental kinetics and transcriptome dynamics of stem cell specification in the spermatogenic lineage. *Nature communications* **10**, 1-14 (2019).

38. Liesecke F*, et al.* Ranking genome-wide correlation measurements improves microarray and RNA-seq based global and targeted co-expression networks. *Scientific reports* **8**, 1-16 (2018).

39. Carlin BP, Chib S. Bayesian model choice via Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Methodological)* **57**, 473-484 (1995).

40. Bailer-Jones C, Rybizki J, Fouesneau M, Mantelet G, Andrae R. Estimating distance from parallaxes. IV. Distances to 1.33 billion stars in Gaia data release 2. *The Astronomical Journal* **156**, 58 (2018).

41. Kruschke J. Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan. (2014).

42. Dony L, König M, Fischer D, Theis FJ. 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*) (2020).

43. Rifai S, Vincent P, Muller X, Glorot X, Bengio Y. Contractive auto-encoders: Explicit invariance during feature extraction. In: *Icml*) (2011).

44. Phillips R, Kondev J, Theriot J. *Physical biology of the cell*. Garland Science (2009).

45. He J, Kumar S, Chang S-F. On the difficulty of nearest neighbor search. *arXiv preprint arXiv:12066411*, (2012).

46. Austin D. We recommend a singular value decomposition. *Feature Column*, (2009).

47. Ottaviani G, Paoletti R. A geometric perspective on the singular value decomposition. *arXiv preprint arXiv:150307054*, (2015).

48. Treves F. *Topological Vector Spaces, Distributions and Kernels: Pure and Applied Mathematics, Vol. 25*. Elsevier (2016).

49. Rashid S, Kotton DN, Bar-Joseph Z. TASIC: determining branching models from time series single cell data. *Bioinformatics* **33**, 2504-2512 (2017).

50. Peixoto TP, Rosvall M. Modelling sequences and temporal networks with dynamic community structures. *Nature communications* **8**, 1-12 (2017).

51. Ross SM*, et al. Stochastic processes*. Wiley New York (1996).

52. Jones MT. Estimating Markov transition matrices using proportions data: an application to credit risk. (2005).

53. Covarrubias G, Liu X. Relationship between Stationarity and Dynamic Convergence of Time Series. *Engineering Proceedings* **18**, 12 (2022).

54. Bürkner P-C. brms: An R package for Bayesian multilevel models using Stan. *Journal of statistical software* **80**, 1-28 (2017).

55. Kruschke J. Chapter 12-bayesian approaches to testing a point ("null") hypothesis in 478 Doing Bayesian Data Analysis , ed. Kruschke JK. *Academic Press, Boston* **479**, 335-358 (2015).

56. Bring J. How to standardize regression coefficients. *The American Statistician* **48**, 209-213 (1994).

57. Djordjevic IB. *Quantum Information Processing, Quantum Computing, and Quantum Error Correction: An Engineering Approach*. Academic Press (2021).

58. Grive KJ*, et al.* Dynamic transcriptome profiles within spermatogonial and spermatocyte populations during postnatal testis maturation revealed by single-cell sequencing. *PLoS genetics* **15**, e1007810 (2019).

59. Wang Z, Xu X, Li J-L, Palmer C, Maric D, Dean J. Sertoli cell-only phenotype and scRNA-seq define PRAMEF12 as a factor essential for spermatogenesis in mice. *Nature communications* **10**, 1-18 (2019).

60. Ernst C, Eling N, Martinez-Jimenez C, Marioni J, Odom D. Staged developmental mapping and X chromosome transcriptional dynamics during mouse spermatogenesis. Nat. Commun. 10: 1251.) (2019).

**Supplementary Information is available for this paper**

**Acknowledgments**

We would like to acknowledge Dr. Jian-Liang Li and Dr. Shyamal Peddada for providing some of the datatsets in the study and their valuable feedback in shaping this manuscript. This work is supported by funding from grant numbers 1ZIAES103350-03 and ES103388-01 from NIEHS. This research was supported by the Intramural Research Program of the NIH, National Institute of Environmental Health Sciences.

**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.

**Author declarations**

The authors declare that there are no competing interests.