Briefings in Bioinformatics, 2025, **26(4)**, bbaf415
https://doi.org/10.1093/bib/bbaf415
Problem Solving Protocol

Boosting data interpretation with GIBOOST to enhance visualization of complex high-dimensional data

Komlan Atitey[1], Jiaqi Li[1], Brian Papas[1], Osafu A. Egbon[1], Jian-Liang Li[1], Musa Kana[2,3], Idowu Aimola[4], Benedict Anchang[1,*]

[1]Biostatistics and Computational Biology Branch, National Institute of Environmental Health Sciences, 111 T W Alexander Dr, Research Triangle Park, Durham, NC 27709, United States
[2]Department of Community Medicine, Faculty of Clinical Sciences, College of Medicine, Kaduna State University, Tafawa Balewa Way, Kaduna, Kaduna State, 800241, Nigeria
[3]Barau Dikko Teaching Hospital, No 1 Lafiya Road, Kaduna, Kaduna State, 800227, Nigeria
[4]Africa Centre of Excellence for Neglected Tropical Diseases and Forensic Biotechnology, Department of Biochemistry, Ahmadu Bello University, Sokoto Road, Zaria, 800001, Nigeria

*Corresponding author. Biostatistics and Computational Biology Branch, National Institute of Environmental Health Sciences, 111 T W Alexander Dr Rall Building, Research Triangle Park, 27709, NC, USA. E-mail: benedict.anchang@nih.gov

Abstract

High-dimensional single-cell data analysis is crucial for understanding complex biological interactions, yet conventional dimensionality reduction methods (DRMs) often fail to preserve both global and local structures. Existing DRMs, such as t-distributed Stochastic Neighbor Embedding (t-SNE), Uniform Manifold Approximation and Projection (UMAP), Principal Component Analysis (PCA), and Potential of Heat-diffusion for Affinity-based Transition Embedding (PHATE), optimize different visualization objectives, resulting in trade-offs between cluster separability, spatial organization, and temporal coherence. To overcome these limitations, we introduce GIBOOST, an AI-driven framework that integrates outputs from multiple DRMs using a Bayesian framework and an optimized autoencoder. GIBOOST systematically selects and integrates the two most informative DRMs by evaluating key visualization features, including separability, spatial continuity, uniformity, cellular dynamics, and cluster sensitivity. Rather than prioritizing a single DRM, it identifies the optimal combination that maximizes clustering sensitivity (GI) while preserving biologically relevant spatial and temporal structures. This integration is further refined through a GI-optimized autoencoder, which optimizes the joint distribution of GI, neuron count, and batch size effects to improve visualization quality. We demonstrate GIBOOST's efficacy across multiple dynamic biological processes, including epithelial–mesenchymal transition, CiPSC reprogramming, spermatogenesis, and placental development. Compared to nine individual DRMs, GIBOOST enhances clustering sensitivity and biological relevance by ~30%, enabling more accurate interpretation of differentiation trajectories and cell–cell interactions. When applied to a large single-cell RNA-seq dataset (~400 000 cells, 28 cell types, seven placental regions), GIBOOST uncovers novel immune-placenta interactions, providing deeper insights into cross-tissue communication during pregnancy. By improving both the visualization and interpretability of high-dimensional data, GIBOOST serves as a powerful tool for computational systems biology, enabling a more accurate exploration of complex cellular systems.

Keywords: dimensionality reduction; single-cell analysis; AI-driven data integration; immune-placental interactions; cell–cell communication; data visualization

Introduction

The advent of single-cell technologies has revolutionized our understanding of cellular heterogeneity [1], differentiation [2], and the molecular mechanisms underpinning complex diseases [3]. Despite these advances, the high dimensionality, scale, and complexity of single-cell data pose significant challenges to the effective visualization and interpretability (EVI) of cell–cell interaction, communication and transitions, which are critical for understanding disease progression and drug response. Popular dimensionality reduction methods (DRMs) such as t-distributed Stochastic Neighbor Embedding (t-SNE) [4], Uniform Manifold Approximation and Projection (UMAP) [5], Principal Component Analysis (PCA) [6], and Potential of Heat-diffusion for Affinity-based Transition Embedding (PHATE) [7] and their extensions have been widely adopted to address these challenges by reducing the complexity of the data while preserving essential features for effective visualization. They each have unique strengths and limitations [8]. For instance, t-SNE excels in preserving local geometric features by maintaining the relative distances between nearby points, which helps in identifying clusters of similar cells. However, it often struggles with global relationships, making it difficult to interpret the overall structure of the data [9]. UMAP is effective in separating clusters and preserving both local and some global geometric features, providing a more balanced view of the data. However, it also may distort distances between clusters, which can affect the interpretation of relationships between different clusters [9]. PCA identifies linear combinations of biological features that maximize variance, which may not always translate to biologically relevant patterns, especially when the data is highly nonlinear. PHATE focuses on preserving the progression of cell states, making it suitable for visualizing developmental trajectories and capturing both local and global geometric features related to

Received: February 18, 2025. **Revised:** June 13, 2025. **Accepted:** July 14, 2025
Published by Oxford University Press 2025.

cell state transitions [10]. However, it may not capture other critical features as effectively, such as the fine-grained structure within clusters. By leveraging on the strengths and limitations of these DRMs in terms of their ability to preserve local and global geometric features for EVI, we propose an AI-driven integrative visualization framework called GIBOOST which is designed to enhance the visualization and interpretation of high-dimensional single-cell data. GIBOOST optimizes the integration of the outputs of multiple DRMs, leveraging on each of their important features, to provide a more holistic representation of the data. The core idea behind GIBOOST is to combine the strengths of different DRMs while mitigating their individual limitations. By doing so, GIBOOST enables researchers to gain a more comprehensive understanding of complex biological processes in the data in order to generate more precise hypotheses.

Rather than prioritizing a single DRM, GIBOOST systematically identifies and integrates the most informative pair of methods to enhance both local and global interpretability. It uses the MIBCOVIS Bayesian framework [8] to assess key visualization features, such as cluster separability, spatial continuity, uniformity, temporal structure, and cluster sensitivity, through a set of quantitative metrics. Based on these evaluations, GIBOOST selects the optimal DRM pair and integrates them using a tailored autoencoder that maximizes clustering sensitivity while optimizing critical hyperparameters like neuron count and batch size. This integration produces a unified visualization that captures complementary biological insights, such as spatial ordering and cell-state transitions, enabling more accurate interpretation of dynamic processes across diverse tissues and disease states.

A few recently published methods share certain principles with GIBOOST, particularly in their efforts to enhance the visualization and interpretability of high-dimensional single-cell data through integration or hybrid approaches. Methods like scPhere [11] and Ensemble UMAP [12] combine multiple dimensionality reduction techniques to capture both local and global structures. scPhere [11] employs a spherical autoencoder framework that allows for geometry-aware embeddings well-suited for circular or branching processes. Ensemble UMAP builds consensus visualizations by averaging outputs from multiple runs or parameter settings of UMAP to improve robustness and generalizability. scVI [13] leverages a variational autoencoder for probabilistic modeling of gene expression in single cells, enabling dimensionality reduction while accounting for batch effects and dropout noise, making it particularly effective for integrating heterogeneous datasets. SIMLR (Single-cell Interpretation via Multi-kernel LeaRning) [14] combines multiple kernels to capture both local and global similarities across cells, producing embeddings optimized for visualization and clustering. SAUCIE (Sparse Autoencoder for Unsupervised Clustering, Imputation, and Embedding) [15] integrates dimensionality reduction with downstream tasks such as batch correction, clustering, and denoising, using a deep learning-based autoencoder with regularized loss functions. Additionally, scCoGAPS [16] and SCENIC+ [17] leverage hybrid and integrative approaches to improve pattern discovery and visualization across different data modalities, while MOFA+ [18], though primarily focused on multi-omics integration, also aligns with GIBOOST's goal of reducing dimensionality while preserving critical biological insights.

However, GIBOOST distinguishes itself by utilizing a Bayesian framework (MIBCOVIS) to optimize the selection of DRMs based on desired performance features such as cluster sensitivity, spatial organization, and temporal continuity. It then integrates these optimized outputs using a targeted-guided neural network-based autoencoder, further refining visualization and interpretability. This combination of Bayesian optimization and deep learning sets GIBOOST apart as a unique and powerful tool for enhancing the visualization of complex biological processes compared to other existing methods.

To validate the efficacy of GIBOOST, we applied it to four distinct dynamic biological processes: Epithelial–mesenchymal transition (EMT) [19], chemically induced pluripotent stem-cell reprogramming (CiPSC) [20], spermatogenesis [21], and placenta development [22]. By preserving both local and global dynamic structures, GIBOOST provides clearer visualization of cell–cell-tissue communication and cellular transitions compared to individual DRMs, revealing previously obscured communication patterns and differentiation trajectories between intermediate states from standard methods. For example, the placenta, a complex organ that facilitates nutrient and gas exchange between the mother and fetus, undergoes dynamic changes during gestation. Recent single-cell studies with data collected at various gestation periods have revealed the diverse cellular composition of the placenta and the critical role of immune cells in regulating its development. However, the high dimensionality and multi-tissue complexity of the data make it challenging to fully capture the complex maternal-fetal interface interactions using traditional DRMs. By integrating the outputs of UMAP and PHATE, GIBOOST provides a more comprehensive view of cell–cell communication during placentation, highlighting key interactions between trophoblasts, immune cells, and stromal cells that are crucial for placental function.

In summary, GIBOOST is a powerful tool for enhancing the visualization and interpretation of high-dimensional data, especially in the analysis of dynamic biological systems, involving multiple multiscale interactions.

Result

Overview of GIBOOST

High-dimensional biological data often require multiple DRMs to reveal both local cluster separability and global spatial structure. GIBOOST is designed to integrate complementary strengths of DRMs to improve visualization and interpretability. As an illustrative example (Fig. 1A), one DRM may generate well-separated clusters (e.g. Method A), while another maintains spatial continuity (e.g. Method B). Rather than selecting a single DRM, GIBOOST synthesizes their outputs to generate embeddings that reflect both cluster boundaries and biological context. To guide this integration, GIBOOST uses the MIBCOVIS framework [8] to evaluate each DRM according to five biologically relevant metrics: separability index (SI), which measures cluster distinction, occupation index (OI), which assesses evenness of embedding space usage, uniformity index (UI), which captures spatial uniformity, time order structure index (TI), which preserves temporal patterns, and gradient boosting classifier index (GI), which quantifies cluster sensitivity in the reduced space.

GIBOOST proceeds in three stages (Fig. 1B–F). First, it profiles candidate DRMs across multiple visualization objectives using biologically relevant metrics (Fig. 1B). Next, it selects the top DRM pair that maximizes additive clustering sensitivity through a Bayesian model conditioned on the GI and other metrics (Fig. 1C–E). Finally, it integrates the selected pair using an optimized autoencoder, producing a fused low-dimensional embedding that enhances clustering clarity based on the GI score (Fig. 1F). The next sections provide a brief mathematical overview

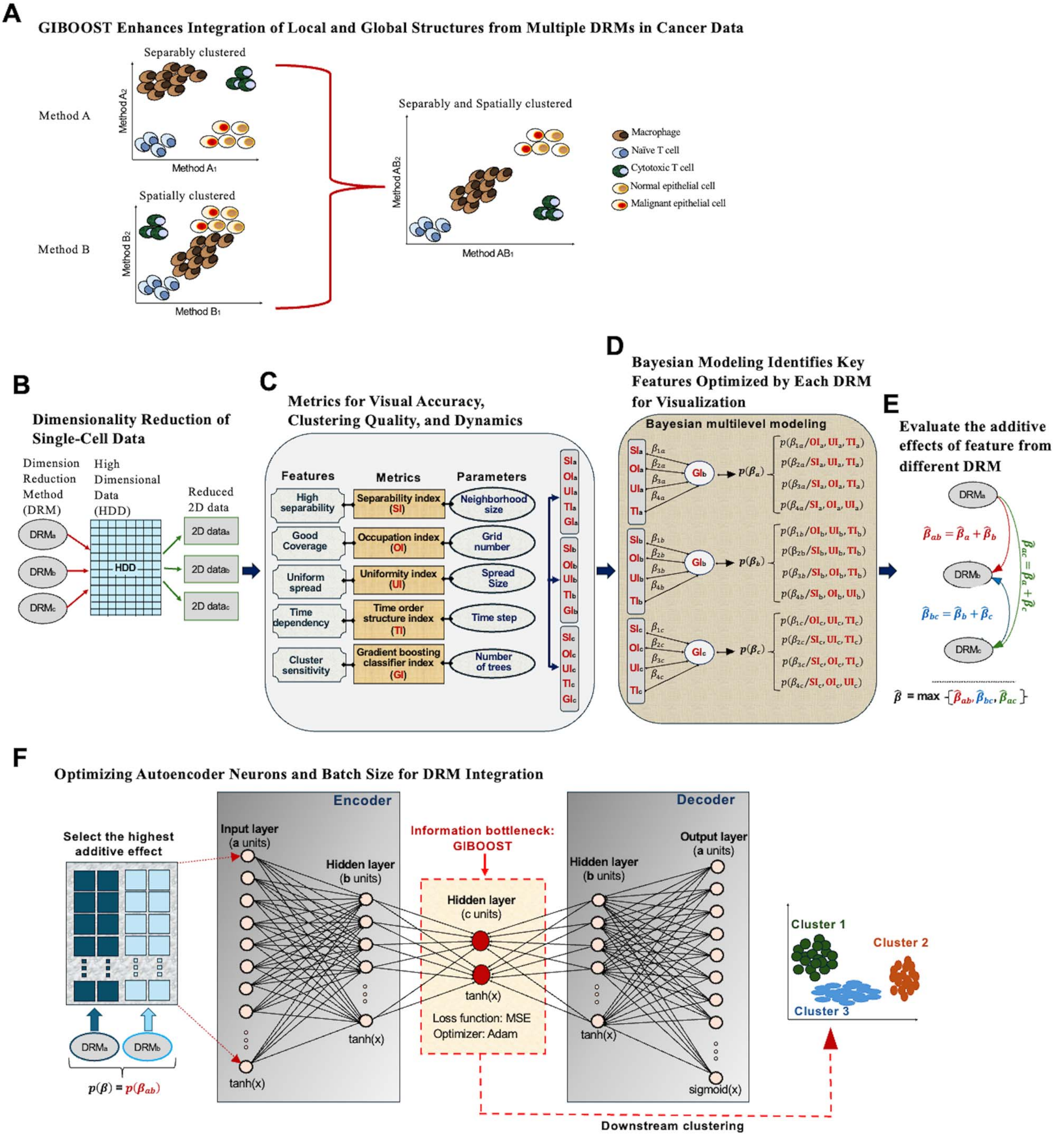

A
GIBOOST Enhances Integration of Local and Global Structures from Multiple DRMs in Cancer Data
Separably clustered
Method A
Method A2
Method A1
Spatially clustered
Method B
Method B2
Method B1
Separably and Spatially clustered
Method AB2
Method AB1
Macrophage
Naïve T cell
Cytotoxic T cell
Normal epithelial cell
Malignant epithelial cell
B
Dimensionality Reduction of Single-Cell Data
Dimension Reduction Method (DRM)
High Dimensional Data (HDD)
Reduced 2D data
DRMa
DRMb
DRMc
HDD
2D dataa
2D datab
2D datac
C
Metrics for Visual Accuracy, Clustering Quality, and Dynamics
Features
Metrics
Parameters
High separability
Separability index (SI)
Neighborhood size
Good Coverage
Occupation index (OI)
Grid number
Uniform spread
Uniformity index (UI)
Spread Size
Time dependency
Time order structure index (TI)
Time step
Cluster sensitivity
Gradient boosting classifier index (GI)
Number of trees
SIa OIa UIa TIa GIa
SIb OIb UIb TIb GIb
SIc OIc UIc TIc GIc
D
Bayesian Modeling Identifies Key Features Optimized by Each DRM for Visualization
Bayesian multilevel modeling
SIa OIa UIa TIa
β_{1a} β_{2a} β_{3a} β_{4a}
GIb
$p(\beta_a)$
$p(\beta_{1a}/OI_a, UI_a, TI_a)$
$p(\beta_{2a}/SI_a, UI_a, TI_a)$
$p(\beta_{3a}/SI_a, OI_a, TI_a)$
$p(\beta_{4a}/SI_a, OI_a, UI_a)$
SIb OIb UIb TIb
β_{1b} β_{2b} β_{3b} β_{4b}
GIb
$p(\beta_b)$
$p(\beta_{1b}/OI_b, UI_b, TI_b)$
$p(\beta_{2b}/SI_b, UI_b, TI_b)$
$p(\beta_{3b}/SI_b, OI_b, TI_b)$
$p(\beta_{4b}/SI_b, OI_b, UI_b)$
SIc OIc UIc TIc
β_{1c} β_{2c} β_{3c} β_{4c}
GIc
$p(\beta_c)$
$p(\beta_{1c}/OI_c, UI_c, TI_c)$
$p(\beta_{2c}/SI_c, UI_c, TI_c)$
$p(\beta_{3c}/SI_c, OI_c, TI_c)$
$p(\beta_{4c}/SI_c, OI_c, UI_c)$
E
Evaluate the additive effects of feature from different DRM
DRMa
DRMb
DRMc
$\hat{\beta}_{ab} = \hat{\beta}_a + \hat{\beta}_b$
$\hat{\beta}_{bc} = \hat{\beta}_b + \hat{\beta}_c$
$\hat{\beta}_{ac} = \hat{\beta}_a + \hat{\beta}_c$
$\hat{\beta} = \max \{\hat{\beta}_{ab}, \hat{\beta}_{bc}, \hat{\beta}_{ac}\}$
F
Optimizing Autoencoder Neurons and Batch Size for DRM Integration
Select the highest additive effect
DRMa
DRMb
$p(\beta) = p(\beta_{ab})$
Encoder
Input layer (a units)
tanh(x)
Hidden layer (b units)
tanh(x)
Information bottleneck: GIBOOST
Hidden layer (c units)
tanh(x)
Loss function: MSE
Optimizer: Adam
Decoder
Hidden layer (b units)
tanh(x)
Output layer (a units)
sigmoid(x)
Cluster 1
Cluster 2
Cluster 3
Downstream clustering

Figure 1. Overview of the GIBOOST framework for enhanced data visualization and interpretability. (A). Plots showing two independent methods (A and B) yielding distinct outputs after high-dimensional data reduction. Instead of selecting one method, the goal of GIBOOST is to integrate these outputs to achieve data that is both separably and spatially clustered. (B). GIBOOST performs dimension reduction using targeted input methods. 2D data$_a$, 2D data$_b$, 2D data$_c$ are the 2D outputs from three different DRMs (DRMa, DRMb, and DRMc). (C). GIBOOST evaluates the performance of each DRM. We used a set of metrics, including the separability index (SI), occupation index (OI), uniformity index (UI), time order structure index (TI), and gradient boosting classifier index (GI), to assess the visualization and interpretability of DRMa, DRMb, and DRMc. (D). GIBOOST uses Bayesian modeling to score metric performance. Conditional posterior effects of metric features ($\beta_1, \beta_2, \beta_3,$ and β_4) are modeled to minimize clustering variance and bias (GI), thereby preserving key information from the high-dimensional dataset (HDD). (E). GIBOOST uses additive effects of features to select optimal combination of methods. The means of the conditional effects ($\beta_1, \beta_2, \beta_3,$ and β_4) of metrics are combined across methods to identify the pair with the most complementary information. (F). GIBOOST uses an autoencoder to perform data integration: The selected DRM$_a$ and DRM$_b$ serve as inputs to an optimized autoencoder, with the hidden layer output representing the integrated 2D data.

of how GIBOOST evaluates the GI, selects the optimal DRM pairs using Bayesian modeling, and optimizes an autoencoder-based framework to integrate the top DRM pair. Detailed information can be found in Supplementary Text 1, 2 and 3.

GIBOOST uses gradient boosting classifier index to enhance visualization and interpretability

The GI score quantifies how well an embedding preserves label-based separability. Given a projected dataset $\varepsilon = \{(x_j, y_j)\}_{j=1}^{D_r}$, where each $x_j \in \mathbb{R}^2$ is a 2D embedding of a data point (with D_r, the number of reduced data points) and $y_j \in \{1, \cdots, K\}$ is its corresponding class label, a multinomial gradient boosting classifier [23–25] (via xgboost R package [26]) is trained. Classification accuracy in the embedded space is used as the GI score:

$$GI = \frac{1}{D_r} \sum_{j=1}^{D_r} \mathbb{I}\left[\hat{y}_j = y_j\right]$$

where $\hat{y}_j = \mathbf{argmax}_k P\left(y = k|x_j\right)$ is the predicted class label (Supplementary Text 2). A higher GI indicates better preservation of class structure, improving downstream interpretability (Supplementary Text 3). We generate a distribution of GI by varying the number of decision trees or weak learners used by xgboost. This distribution is used to select optimal data reduction methods within a Bayesian framework for data integration.

GIBOOST selects optimal dimensionality reduction method pairs via Bayesian modeling

To identify the best DRM pair, GIBOOST uses a Bayesian conditional effect regression [27], to estimate the influence of each visualization metric on GI. For each DRM v, we model its expected GI score z_v (Fig. 1D) as a function of a predictor vector $S_v = \left(x_{1,v}, x_{2,v}, x_{3,v}, x_{4,v}\right)$, representing SI, OI, UI, and TI:

$$\mathbb{E}\left(z_v|S_v\right) = \frac{e^{S_v \boldsymbol{\phi_v}}}{1 + e^{S_v \boldsymbol{\phi_v}}}$$

Here $\boldsymbol{\phi_v} = \left(\phi_{1,v}, \phi_{2,v}, \phi_{3,v}, \phi_{4,v}\right)$ denotes regression coefficients associated with the predictors index by u, and priors following a Student's t-distribution to ensure robustness to outliers. The conditional variance is modeled as:

$$Var\left(z_v|S_v\right) = \varphi_{u,v}.V_a\left[\mathbb{E}\left(z_v|S_v\right)\right]$$

where $\varphi_{u,v}$ is a dispersion parameter and V_a is a variance function that relates the conditional variance to the mean. Posterior inference [28] is performed via MCMC with M = 4000 draws (first 2000 as burn-in). The posterior mean of standardized effects $\hat{\beta}_{u,v}$ is used to compute the additive sensitivity score:

$$\hat{\beta}_n = \max_{u \in \{1,2,3,4\}} \hat{\beta}_{u,v}$$

Among all $\binom{N}{2}$ DRM pairs, the pair with the maximum sum of additive effects is selected, ensuring high clustering sensitivity with minimal redundancy (Fig. 1E). See Method section in Supplementary Text 1 for more details.

GIBOOST optimizes an autoencoder to integrate the outputs of the two most effective dimensionality reduction methods

To synthesize the top two DRMs into a unified 2D embedding, GIBOOST employs an autoencoder (AE) [8] optimized for clustering quality. The AE consists of a single hidden-layer encoder and symmetric decoder, with the encoder mapping $x \in \mathbb{R}^4$ to a latent representation $h = \mathbf{tanh}\left(Wx + b\right)$. The decoder reconstructs the input via:

$$\hat{x} = k\left(\hat{W}h + \hat{b}\right),$$

where k is the sigmoid function.

The AE is trained to minimize mean squared reconstruction error (MSE):

$$\mathcal{L}\left(x, \hat{x}\right) = \frac{1}{N} \sum_{i=1}^{N} \left(x_i - \hat{x}_i\right)^2$$

Optimization is carried out using the Adam optimizer [29] to enhance convergence speed and stability. The number of neurons controls the model's capacity to compress and represent information. Too few neurons may lead to underfitting and loss of critical signal, while too many may retain noise and reduce generalizability. Similarly, batch size affects training dynamics such that smaller batch sizes can introduce more stochasticity, potentially helping escape local minima, whereas larger batches stabilize training but may obscure subtle patterns. To ensure the output remains 2D, the bottleneck layer is constrained to 2 neurons. GIBOOST systematically evaluates AE configurations across a range of hidden units (8–120) and batch sizes (32–128) (Fig. 2A), selecting configuration that yields the highest GI score after clustering. This approach balances fine-grained feature capture with robust generalization, producing a fused low dimensional embedding that enhances interpretability while preserving biologically meaningful structure (Fig. 1F). See Method section in Supplementary Text 1 for further details.

GIBOOST effectively improves the clustering structure of the high-dimensional data with significant reduction in variance and bias

GIBOOST was next applied to three high-dimensional datasets representing different biological processes: EMT, CiPSCs, and spermatogenesis. We used four common DRMs: t-SNE, UMAP, PCA, and PHATE as input. The datasets correspond to EMT [19] (Data set 1), CiPSCs [20] (Data set 2), and spermatogenesis [21] (Data set 3). Data set 1 comprises of 96 000 single-cell observations spanning a 20-day period following in vitro stimulation of lung cancer cell lines with TGF-β for 10 days, succeeded by TGF-β withdrawal for an additional 10 days. This dataset delineates eight states of EMT-MET transition (Fig. 2), validated by six canonical EMT markers (Vimentin, Ecadherin, Cd44, Cd24, MUC1, Twist1). Data set 2 encompasses 50 000 cells and 102 signature gene expression (Supplementary Table S1) sampled across 12 time points over 21 days, exploring pluripotency via chemically induced cellular reprogramming. Lastly, Data set 3 contains ∼110 000 cells and 174 significant marker genes (Supplementary Table S1) sampled from 16 postnatal stages throughout spermatogenesis. These cells are categorized into 29 clusters spanning various cell types implicated in spermatogenesis and supporting the testicular microenvironment.

Figure 2B–F illustrates the reduced representations of the three high-dimensional datasets (EMT, CiPSC, and spermatogenesis) using five methods. From left to right, the columns display results for the EMT, CiPSC, and spermatogenesis datasets, while each row corresponds to t-SNE, UMAP, PCA, PHATE, and GIBOOST. GIBOOST optimization involved tuning the autoencoder's neuron count (nc) and batch size (bs) as binary variables optimized for GI. The optimal (nc, bs) values corresponding to the highest GI were (48, 82) for the EMT dataset, (58, 60) for the CiPSC dataset, and (91, 38) for the spermatogenesis dataset (Fig. 2A). Accordingly, t-SNE–PHATE, PCA–PHATE, and PCA–PHATE were selected as the best pairs for integrating complementary information in the EMT, CiPSC, and spermatogenesis datasets, respectively. The distinct clustering of points highlights underlying similar structures or relationships within the data, enabling more efficient exploration and interpretation.

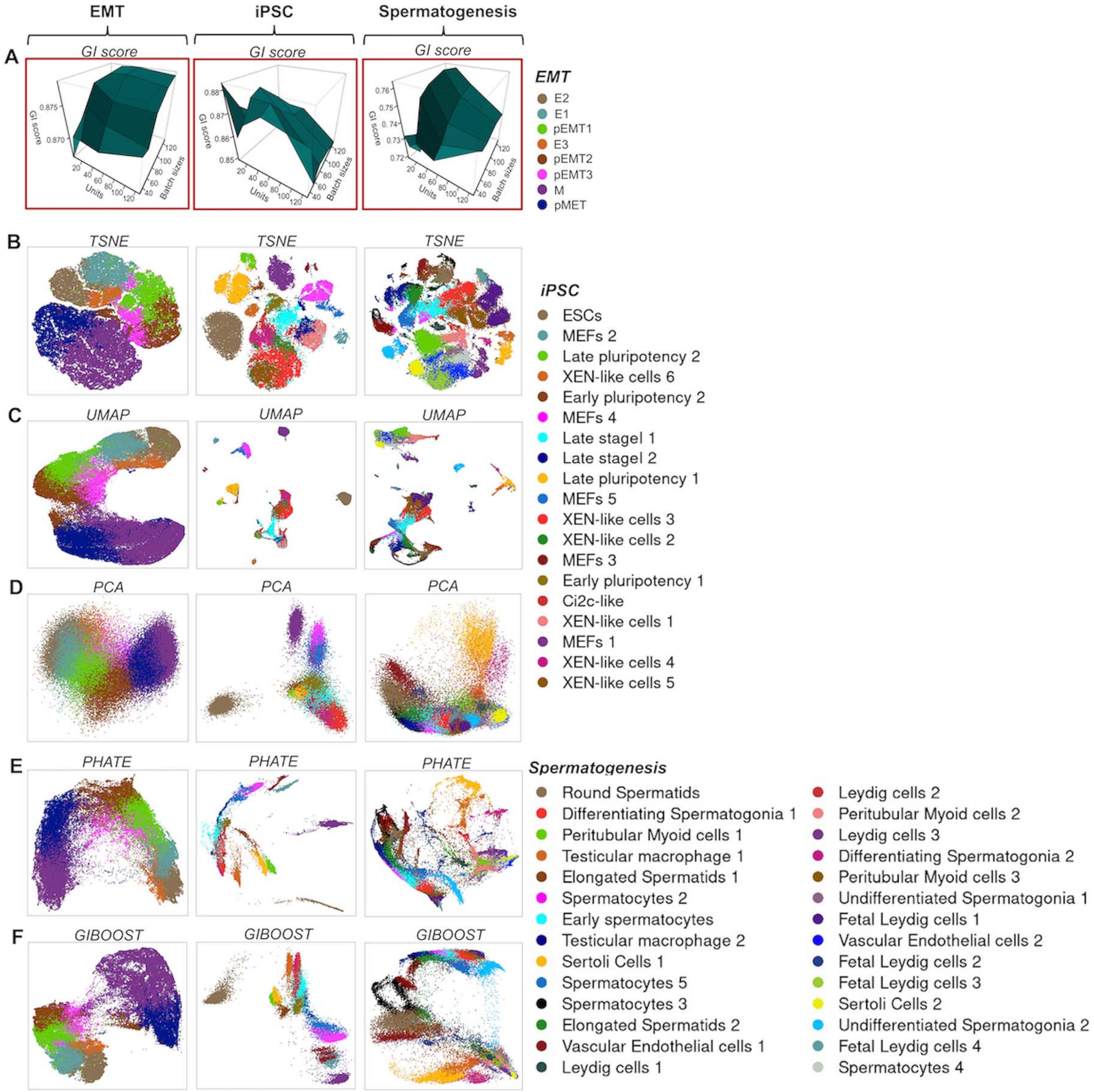


Figure 2. Comparison of reduced 2D data for three single-cell time course datasets. (A). Multivariate plot of GI score by number of neurons in the input layer and the batch size of the Autoencoder (AE) architecture across EMT(first column), CiPSC (second column) and spermatogenesis data (third column). (B). t-SNE row plots for EMT, CiPSC and spermatogenesis data. (C). UMAP row plot for EMT, CiPSC and spermatogenesis data. (D). PCA row plots showing reduced outputs for EMT, CiPSC and spermatogenesis data. (E). 2D plots of PHATE from EMT, CiPSC and spermatogenesis data analysis. (F). GIBOOST optimized projections after reducing EMT, iPSC, and spermatogenesis data.

We next compared the performance of t-SNE, UMAP, PCA, PHATE, and GIBOOST in terms of cluster sensitivity after data reduction. GIBOOST consistently outperformed the other methods across all three datasets (Fig. 3A). Specifically, compared to the top-performing single methods; t-SNE for EMT, PCA for CiPSC, and PCA for spermatogenesis, GIBOOST enhanced accuracy performance by 12.2%, 15.0%, and 17.3%, respectively. We also conducted a cophenetic correlation matrix [30] analysis to assess the relative cluster relationships between the original HDD (EMT, CiPSC, and spermatogenesis) and their reduced representations (Supplementary text 4). A high correlation coefficient indicates strong structural similarity between the original and reduced data [31], suggesting that the data reduction method effectively preserves both global and local features [32]. GIBOOST consistently showed the highest correlation coefficients compared to the other four methods (Fig. 3B, Supplementary Figs S1–S3).

Accurate visualization and interpretation of the communication and transitions between EMT and mesenchymal–epithelial transition (MET) during tumor progression is crucial for understanding cancer metastasis [33]. EMT enables cancer cells to detach from the primary tumor and invade surrounding tissues, while MET allows them to establish new colonies at distant sites [34–36]. Single-cell data reduction EMT–MET maps have been used to infer these transitions to better understand the mechanisms of cancer spread and identify potential therapeutic targets for preventing or treating metastasis [19]. However, given that the structure of the map may vary depending on the data reduction method, we show that an integrative approach like GIBOOST can indeed enhance interpretability by resolving some of the controversial transitions or relationships. Figure 3C–H highlights eight EMT–MET states after data reduction; three epithelial states (E1–E3), three partial states (pEMT1–3), one mesenchymal state (M) and one partial MET state (pMET) using t-SNE, UMAP, PCA,

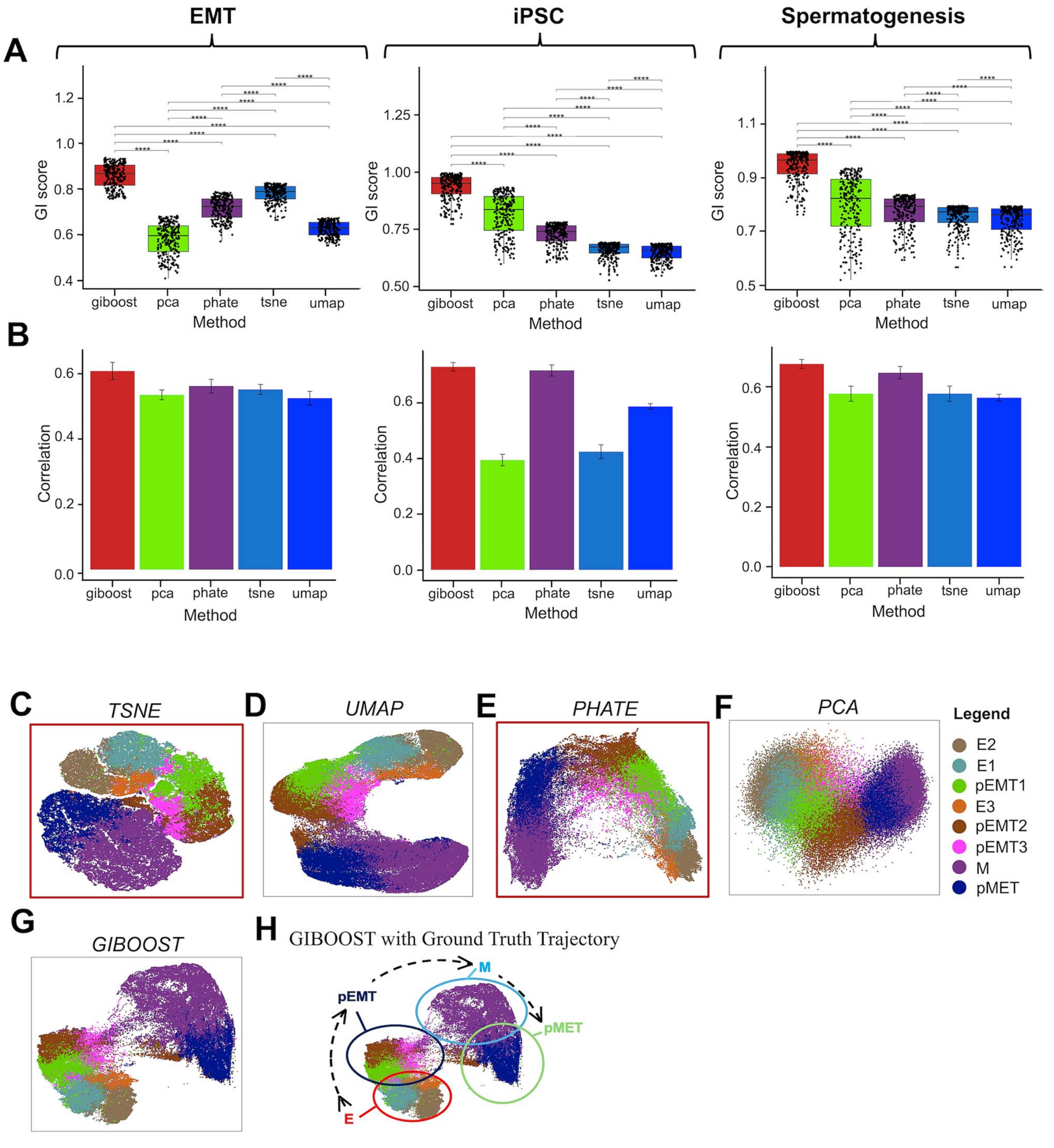


Figure 3. GIBOOST validations. The effectiveness of GIBOOST is assessed for cluster sensitivity using GI index and structural fidelity using cophenetic correlations. Additionally, we also qualitatively validate the sensitivity of the relationships using prior knowledge between the clusters during the evolution of the biological processes as further evidence of the efficiency of GIBOOST. Quantitative validation. (A). GI scores for various DRM on the EMT, CiPSC and spermatogenesis data from left to right respectively. In all 3 cases, the GI score of the GIBOOST is the highest. (B). Bar plots of correlation coefficients which measure the structural relationship between the high-dimensional EMT, CiPSC, and spermatogenesis data and their reduced data across five different DRM. From left to right, the plots represent EMT, CiPSC, and spermatogenesis data, with GIBOOST always exhibiting the highest correlation coefficient. Qualitative validation. Visual investigation using prior knowledge of how different DRMs; (C). t-SNE. (D). UMAP, (E). PHATE. (F). PCA. (G). GIBOOST align cells states during EMT-MET process. GIBOOST integrates the outputs of t-SNE(C) and PHATE(E) highlighted in red. (H). Unlike t-SNE, UMAP, PCA, and PHATE, GIBOOST accurately aligns cell states with the expected biology ordering: Epithelial (E), partial EMT (pEMT), Mesenchymal (M), and partial MET (pMET).

PHATE, and GIBOOST techniques based on an EMT CyTOF dataset. GIBOOST selected t-SNE and PHATE as the best combination for enhancing visualization and interpretability. While the transition between the differentiated epithelial and mesenchymal states can easily be identified from methods like t-SNE, UMAP, PCA and PHATE, the transition trajectories between the partial EMT and MET processes are not obvious. GIBOOST, is the only output that shows a clear path from E-pEMT–M-pMET compared to the other methods effectively distinguishing several transitions from mesenchymal to epithelial phenotypes during EMT–MET.

Benchmarking GIBOOST for sensitivity to trustworthiness, clustering number uncertainty, and trajectory accuracy across integrative dimensionality reduction methods

To assess the effectiveness of GIBOOST in improving trajectory inference, we benchmarked its performance using the Slingshot algorithm (see Supplementary Text 5), aligning inferred pseudotime values with a known biological trajectory. Specifically, we leveraged the EMT, a widely studied process that progresses through epithelial (E), partial EMT (pEMT), mesenchymal (M), and partial mesenchymal–epithelial transition (pMET) states. This biological sequence served as a ground truth to evaluate the fidelity of pseudo-time estimates generated by various DRMs.

We applied GIBOOST alongside established DRMs including Ensemble UMAP [37], SAUCIE [15], SIMLR [14], scVI [13], and scPHERE [11] to EMT CyTOF data (Fig. 4A). These methods summarized briefly in the introduction, are advanced DRMs designed to extract meaningful low-dimensional representations from high-dimensional single-cell data. Collectively, these tools are designed to uncover cellular heterogeneity and trajectories while addressing the unique challenges of single-cell omics data.

Each method's reduced representation was used as input to the Slingshot algorithm [38] for trajectory inference, using known biological cluster labels to anchor the trajectories. Pseudotime values were then computed for all cells, reflecting their inferred positions along the differentiation path. To quantify how accurately each method captured the biological progression, we computed the Spearman rank correlation [39] between the true cluster ordering and the inferred pseudo-time. This nonparametric metric provides a robust measure of concordance between biological and computational orderings.

GIBOOST outperformed all other methods, achieving the highest Spearman correlation ($\rho = 0.820$), indicating superior alignment with the known EMT trajectory. Ensemble UMAP and scPHERE followed with moderate correlations of $\rho = 0.807$ and $\rho = 0.739$, respectively, while SIMLR, SAUCIE and scVI performed less effectively, with correlations ranging from $\rho = 0.195$ to 0.627 (Fig. 5). These results highlight GIBOOST's advantage in preserving the global structure of dynamic cellular transitions, supporting more accurate and biologically meaningful trajectory inference. Overall, GIBOOST offers a robust framework for modeling differentiation, plasticity, and other temporally structured processes in single-cell omics data.

Additionally, to robustly evaluate GIBOOST's performance in capturing clustering structure and uncertainty, we employed the Silhouette score, which quantifies how well-separated and cohesive the identified cell populations are in the reduced space, in an unsupervised manner. Higher Silhouette scores indicate better-defined and more distinct clusters. GIBOOST consistently outperformed competing methods, including Ensemble UMAP, SAUCIE, SIMLR, scVI, and scPHERE, demonstrating superior cluster separation and structural integrity (Fig. 4B).

We also assessed local neighborhood preservation using the Trustworthiness score [40], which measures how well the reduced embedding maintains local relationships from the original high-dimensional space. Specifically, we calculated the proportion of nearest neighbors in the low-dimensional space that were not among the true neighbors in the high-dimensional space. By varying the neighborhood size, GIBOOST consistently outperformed all other methods in terms of Trustworthiness (Fig. 4C), further reinforcing its effectiveness in preserving local structure through data integration.

GIBOOST significantly enhances visualization and interpretation of cell–cell interaction effects during various differentiation processes

We next evaluated GIBOOST's performance in identifying clusters and their local and global relationships along staging trajectories in two distinct biological processes: CiPSC and spermatogenesis processes, respectively.

Chemically induced pluripotent stem-cell analysis

GIBOOST analysis of scRNA-seq data from the CiPSC process using mouse embryonic fibroblasts (MEFs) successfully identified five distinct MEF cell states at the beginning after induction (Fig. 6A), a heterogeneous XEN-like intermediate cell state and a continuum of early and late pluripotency cell states that ultimately differentiate into embryonic stem cells (ESCs) (Supplementary Fig. S4). MEFs, the starting somatic cell population, are characterized by the expression of key markers such as Prrx1, Twist2, Zeb2, Thy1, and Fbn1 (Supplementary Fig. S5). During reprogramming, cells transition through an intermediate XEN-like state, characterized by the expression of markers like SALL4, GATA4, and SOX17. This complex state space represents a critical phase of chemical reprogramming, with all subgroups of XEN-like cells expressing key XEN master genes, signifying a shared identity. However, differential gene expression (DEG) analysis (Supplementary Fig. S6) highlights heterogeneity within the XEN-like population revealing distinct gene expression profiles across subgroups and underscoring the plasticity of these intermediate cells. As reprogramming progresses, cells move toward pluripotency, passing through early and late pluripotent states. Early pluripotent cells co-express core pluripotency genes, such as Sall4 and Lin28a, alongside XEN-associated genes, reflecting their transitional nature (Supplementary Figs S4, S5 and S7). Late pluripotent cells, in contrast, silence XEN-related genes and upregulate additional pluripotency markers, including Nanog and Sox2, signifying the acquisition of a fully reprogrammed, pluripotent state. Together, these findings provide a detailed characterization of the dynamic cellular transitions during CiPSC reprogramming and reveal the intricate molecular interplay underlying somatic-to-pluripotent transformation.

Spermatogenesis analysis

GIBOOST's analysis of scRNA-seq data from mouse spermatogenesis and the testicular microenvironment (Fig. 6B) reveals a clear spatial separation of spermatogenesis from the surrounding supporting cells. Supplementary Fig. S8 highlights key markers expressed by the major cell types regulating spermatogenesis and supporting the testicular microenvironment. The spermatogenesis trajectory begins with undifferentiated sperm cells and progresses through early differentiated spermatogonia, early spermatocytes, spermatocytes, round spermatids, and finally elongated spermatids. DEG analysis (Supplementary Fig. S9) reveals distinct gene expression profiles marking the transitions from undifferentiated stem cells to early spermatogenesis. This developmental progression is supported by the testicular microenvironment, which comprises diverse cell types, including Leydig cells, myoid cells, vascular endothelial cells, testicular macrophages, and fetal Leydig cells, each contributing uniquely to the development and function of the testis.

Sertoli cells emerge as a critical niche component, which is intimately associated with spermatogenesis. These cells provide structural and paracrine support to germ cells at various stages of differentiation, supply nutrients, isolate developing cells from

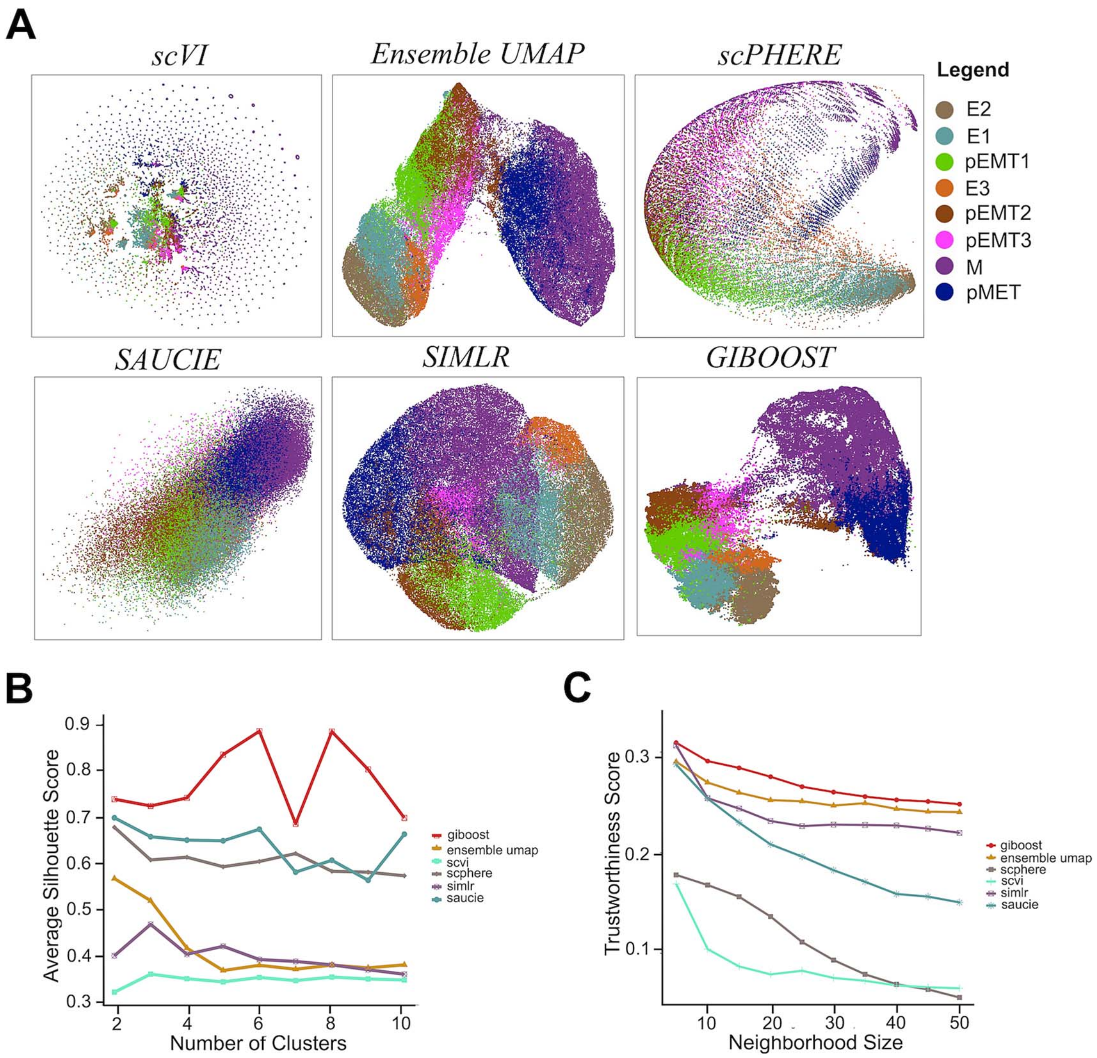


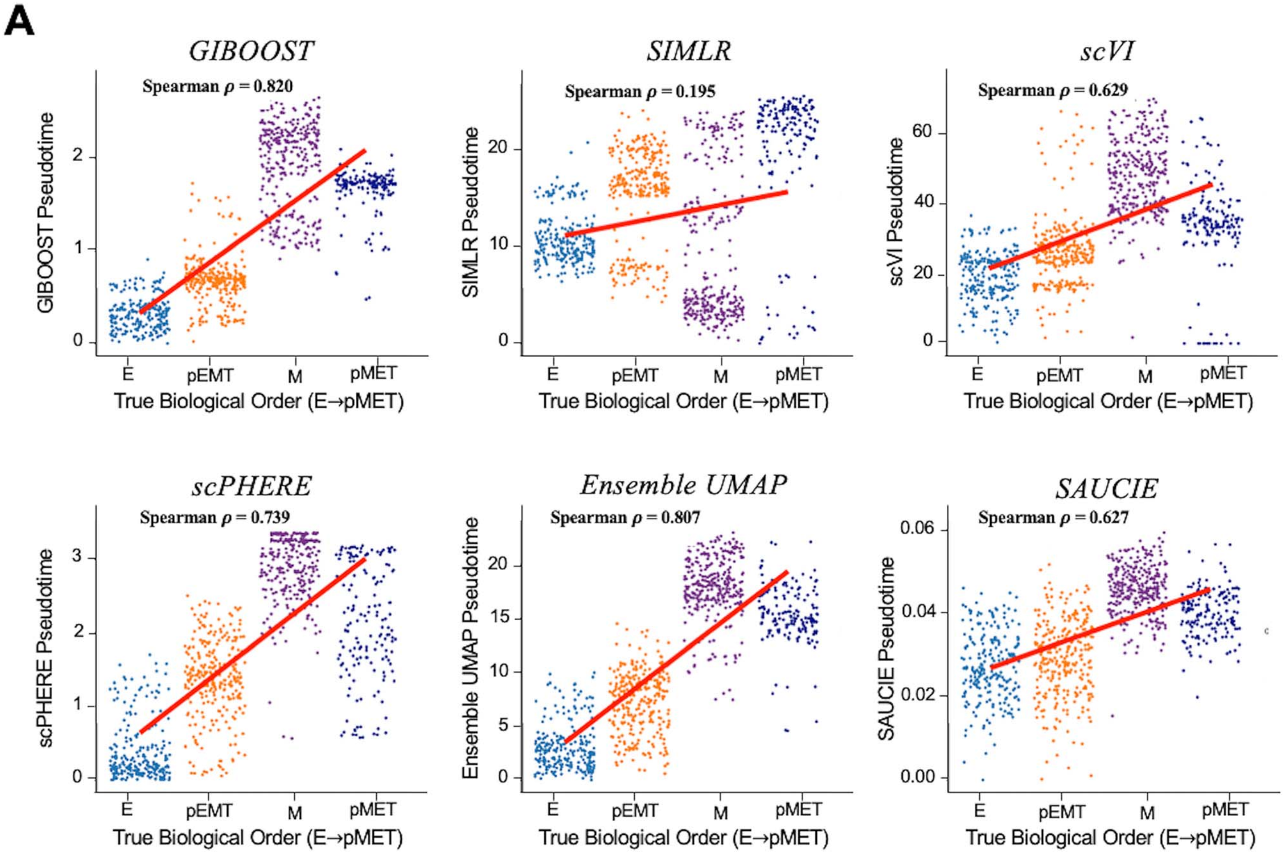
Figure 4. Global and local structure validation: (A). Visual assessment of DRMs based on prior biological knowledge and expected cellular structures during EMT-MET process. Unlike other methods, GIBOOST accurately aligns cell states with the expected biology ordering: Epithelial (E), partial EMT (pEMT), Mesenchymal (M), and partial MET (pMET). (B). Silhouette score comparison across DRMs. Quantitative assessment of clustering quality using silhouette scores for GIBOOST and competing DRMs (ensemble UMAP, SAUCIE, SIMLR, scVI, and scPHERE). GIBOOST consistently achieves the highest silhouette score, reflecting its superior ability to preserve biologically meaningful cluster structures in high-dimensional single-cell data. (C). Trustworthiness score comparison across DRMs. Evaluation of local structure preservation in reduced representations using trustworthiness scores for GIBOOST, ensemble UMAP, SAUCIE, SIMLR, scVI, and scPHERE. GIBOOST outperforms other methods, demonstrating its effectiveness in maintaining local data integrity during dimensionality reduction.

bloodborne molecules, and facilitate their movement toward the lumen of the seminiferous tubules. Sertoli cells also promote the self-renewal of spermatogonial stem cells.

Interestingly, the data reveal two distinct Sertoli cell states: Sertoli cells 1 and Sertoli cells 2 (Fig. 6B, Supplementary Fig. S10). Sertoli cells 2 are closely linked to undifferentiated stem cells and express high levels of marker genes such as Lgals7, Socs2, Ccnd2, and Wnt6. In contrast, Sertoli cells 1 are associated with early spermatogonia differentiation and are characterized by markers such as Ldhc, Meig1, Pcsk1n, and Fabp9. These findings underscore the dynamic interplay between germ cells and the testicular microenvironment, with Sertoli cells playing a pivotal regulatory and supportive role throughout spermatogenesis. The distinct states of Sertoli cells highlight their specialized contributions at different stages, ensuring proper germ cell development and the maintenance of male fertility.

Effectiveness of GIBOOST in capturing specific biological interactions involved in placentation

Pregnancy is a complex physiological state involving dynamic changes in the placenta and immune systems to protect both the mother and the developing fetus [41]. A pivotal component of these changes is cell–cell communication. Cell–cell communication is essential for placenta development and pregnancy, facilitating trophoblast differentiation, invasion, nutrient and gas exchange, immune regulation, and hormonal signaling for pregnancy maintenance. It also supports angiogenesis, ensuring proper blood supply and waste removal to maintain a healthy intrauterine environment. Disruptions in these pathways can lead to pregnancy complications such as preeclampsia, intrauterine growth restriction, and other disorders. Understanding the interactions between maternal immune cells (e.g. macrophages and natural killer cells) and fetal-derived immune cells, along



B

Methods	tSNE	UMAP	PCA	PHATE	GIBOOST
Spearman Correlation for Trajectory Sensitivity	0.598	0.768	0.701	0.612	0.820

Figure 5. Benchmarking GIBOOST for accurate trajectory inference using EMT progression. (A). Quantitative evaluation of pseudotime alignment accuracy using the Slingshot algorithm applied to EMT CyTOF data. The benchmark compared GIBOOST with ensemble UMAP, SAUCIE, SIMLR, scVI, and scPHERE by aligning inferred pseudo-time values with the known biological sequence of the EMT process (E → pEMT → M → pMET). Spearman rank correlation was used to assess the consistency between inferred pseudo-time and true biological ordering. GIBOOST achieved the highest correlation ($\rho = 0.820$), outperforming all other methods and demonstrating superior ability to preserve biologically meaningful transitions in trajectory inference. (B). Pseudotime alignment accuracy for EMT CyTOF data using Slingshot. GIBOOST outperforms input DRMs (t-SNE, UMAP, PCA, PHATE) by achieving the highest Spearman correlation ($\rho = 0.820$) with the known EMT trajectory (E → pEMT → M → pMET), highlighting its enhanced ability to preserve biologically meaningful transitions.

with the cellular and molecular changes in placental disorders like preeclampsia, gestational diabetes, and preterm birth, remains challenging [42]. The additional goal of this part of the study is to utilize GIBOOST to unveil the complex maternal-fetal cellular network communication between different types of trophoblast, mesenchymal and endothelial and immune cells during placental growth by integrating a large cohort of data spanning different placenta regions across various gestational periods during pregnancy (Supplementary Text 6).

Recent studies in single-cell placenta research present exciting opportunities for gaining insights into the development, function, and disorders of this vital organ. For instance, Vento-Tormo *et al.* (2018) [43] reconstructed the early maternal-fetal interface in humans at the single-cell level. Liu *et al.* (2018) [44] used single-cell RNA-seq to explore the diversity of trophoblast subtypes and differentiation patterns in the human placenta providing a valuable insight into placental cellular heterogeneity. Suryawanshi *et al.* (2018) [45] conducted single-cell RNA sequencing on human placental villous and decidual tissue during early pregnancy. More recently, Guo *et al.* (2021) [46] conducted a single-cell transcriptomic analysis of decidual immune cells from patients with recurrent pregnancy loss (RPL) and healthy controls. Pique-Regi *et al.* (2019) [47] investigated the cell types and transcriptional signatures of the human placenta in term and preterm parturition. Yang *et al.* (2021) [48] identified various cell types in the human placenta potentially associated with gestational diabetes mellitus through scRNA-seq analysis. Garcia-Flores *et al.* (2022) [49] explored immune responses during COVID-19 disease using single-cell data from maternal and cord blood. Together, the scRNA-seq datasets from normal placenta (placenta collected during normal delivery) tissue in these studies offer a comprehensive view of placental development across different gestational stages, highlighting the dynamic changes in cellular composition, structural development, and functional specialization required to further our understanding of placental biology. Supplementary Text 6 provides a summary of meta

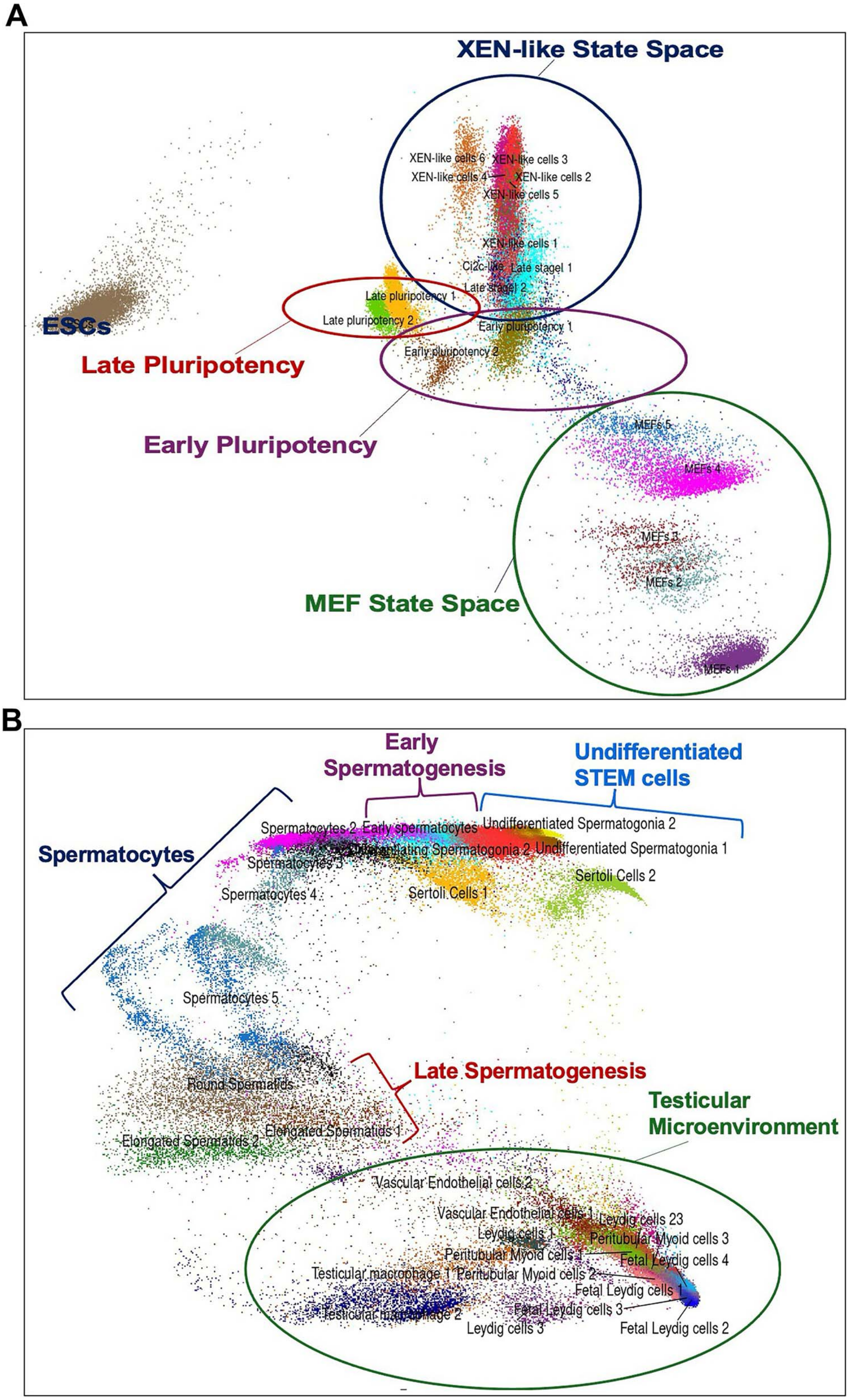


Figure 6. GIBOOST enhances visualization and interpretation of stem cell reprogramming and spermatogenesis. (A). GIBOOST effectively preserves both the local and global structures of the CiPSC data by globally ordering the different differentiation stages involved in the CiPSC process from MEFs to intermediate XEN-like, early and late pluripotency cells that ultimately differentiate into ESCs and locally illustrating clearly the heterogeneous cell states within each stage of the process. (B). GIBOOST provides a clear spatial separation of the cells involved in spermatogenesis and the surrounding supporting testicular environment. It effectively highlights the diversity of various heterogeneous trajectories in spermatogenesis, including undifferentiated stem cells, early spermatogenesis, spermatocytes, late spermatogenesis, and the testicular microenvironment.

scRNA-seq datasets from four different studies: E-MTAB-6701 [43], GSE89497 [50], phs001886.v4.p1 [47] and PRJNA492902 [22] used to build an integrated placenta map. In total, the pooled placenta data consists of expression values from 33 837 genes across 341 090 cells spanning 28 cell types and/or states after quality control, batch correction and normalization using reciprocal PCA integration tool from Seurat v5. The annotated cell types include: dendritic cells (DC1, DC2), decidual macrophages (dM1, dM2, dM3), decidual natural killer cells (dNK1, dNK2), NK cells (NK CD16-, NK CD16+) decidual perivascular cells (PV1, PV2), endothelial cells [Endo (m), Endo L], decidual stromal cells (dS1, dS2, dS3), epithelial glandular cells (Epi1, Epi2), fibroblast (F1, F2), Hofbauer cells (HB), innate lymphocyte cells (ILC3), monocytes (MO), plasma cells, T cell, extravillous trophoblast (EVT), syncytiotrophoblast (SCT), and villous cytotrophoblast (VCT).

Individual dimensionality reduction methods provide a sub optimal view of cell–cell communication

We investigated different trajectories of placentation by exploring the connections between cluster states resulting from the use of t-SNE, UMAP, PCA, and PHATE to reduce the high-dimensional placenta data. While t-SNE and UMAP are effective at highlighting clusters and local neighborhoods, they distort the overall distances and relationships between clusters. For example, despite the close connections between dendritic cells (DC1, DC2), trophoblasts (EVT, SCT, VCT), macrophages (dM1, dM2), and T cells in processes like pregnancy (placental immunity), these methods struggle to accurately represent the relationships between dendritic cells, trophoblasts, and macrophages with T cells in the resulting graphs (Fig. 7A–B), leading to misinterpretations of their biological connections. Specifically, t-SNE fails to accurately depict the continuity or relationships between macrophages, DCs, and T cells. While UMAP generally provides better visualization, it introduces excessive gaps between these clusters in complex placental data, inadequately representing cell–cell communication. For instance, there is an exaggerated gap between macrophages and immune cells compared to other cell types, which is not biologically relevant for interpretation.

While PHATE (Fig. 7D) provides valuable insights into the continuity and transitions within the placenta data, particularly the relationships between dendritic cells (DC1, DC2), trophoblasts (EVT, SCT, VCT), macrophages (dM1, dM2), and T cells, its high degree of cluster overlap obscures the discrete boundaries between cell types or states. This is critical for analyses such as defining distinct cell–cell communication pathways during processes like placentation. On the other hand, PCA, which is not suited for capturing the complex nonlinear relationships in the data, failed to provide meaningful insights into these continuities and transitions within the placenta data (Fig. 7C).

GIBOOST reveals cell–cell tissue specific interactions during placental development

Unlike traditional methods, GIBOOST presents cluster specificity and interactions in a biologically relevant form, effectively showing continuity and transitions within the placenta data (Fig. 7E). The left half of the projection is enriched with immune cells communicating with the right half of the map enriched with mostly epithelial, endotheial, mesenchymal celltypes. This allows for the effective inference of cross tissue heterogeneity in cell–cell communication during placenta development. GIBOOST selected UMAP and PHATE as the two methods that displayed the most complementary information for EVI. Subsequently, it utilized an optimized autoencoder to capture the combined effects of UMAP's separability and PHATE's continuity and space maximization (Fig. 7L). GIBOOST highlights various cell–cell communication involved in different biological processes during placenta development. Notable examples include interactions between dendritic cells and T cells (Fig. 7G), trophoblast and T cells (Fig. 7H), and fibroblast—endothelial cells (Fig. 7I).

For quantitative validation, we compared the performance of t-SNE, UMAP, PCA, PHATE, and GIBOOST in terms of their ability to preserve cluster sensitivity after data dimension reduction. Our findings indicate that GIBOOST outperforms the other methods across the placenta data (Fig. 7J), achieving an accuracy of 99%, which is 11.4% higher than the top accuracy scores of the other four methods. Additionally, our analysis reveals that GIBOOST consistently has the highest cophenetic correlation coefficient among the five DRMs (Fig. 7K). We further examined the feature proportion contribution scores of each of the five methods in terms of SI, OI, UI, and TI (Fig. 7L). The results indicate that GIBOOST performs best with a combined SI (37%) and OI (52%) additive effects of integrating PHATE and UMAP, which individually show the highest performance in OI and SI, respectively. GIBOOST also shows an 8% contribution in TI. In summary, when evaluating the performance of the DRMs by averaging the GI scores across the four datasets used in this study, GIBOOST increases accuracy by 26.7% (Fig. 7M), outperforming the top accuracy scores of the other methods.

Discussion

Accurate identification of cell–cell-tissue interactions using single-cell data is crucial for understanding signal communication during biological processes, developing effective treatments, and ensuring healthy outcomes, as exemplified by the critical role these interactions play in placenta development during pregnancy [51]. However, the high dimensionality and complexity of the data pose significant challenges for the EVI motivating the need for data reduction. Even after data reduction, ensuring that the lower-dimensional representation still accurately reflects the underlying biology, including cellular transitions and cell–cell communication, is difficult, especially when subtle interactions are involved as evidence from most traditional DRM when applied independently to single-cell data. To address these limitations, we developed GIBOOST, a computational tool designed to enhance the visualization and interpretability of complex high-dimensional data. Given a set of targeted data reduction methods, the GIBOOST first applies Bayesian regression modeling on EVI features to select the two most complementary DRMs. It then uses an optimized autoencoder to integrate the two outputs from the selected pair. In this study, we demonstrate extensively that GIBOOST achieves superior results in terms of preserving global and local cluster relationship sensitivities compared to using DRMs individually.

Popular DRMs such as t-SNE, UMAP, PCA, and PHATE [52] offer different strengths for visualizing high-dimensional data [8], while tools like CellphoneDB [53] and SingleCellSignalR [54] leverage on these DRMs to infer cell–cell communication. However, these DRMs often fall short when used independently, either distorting global structures or failing to capture fine-grained subpopulations. GIBOOST overcomes these limitations by leveraging complementary properties of multiple DRMs under supervised optimization, significantly enhancing the preservation of both local and global features crucial for downstream biological interpretation. To demonstrate GIBOOST's advantages, we applied it to diverse datasets representing dynamic biological processes

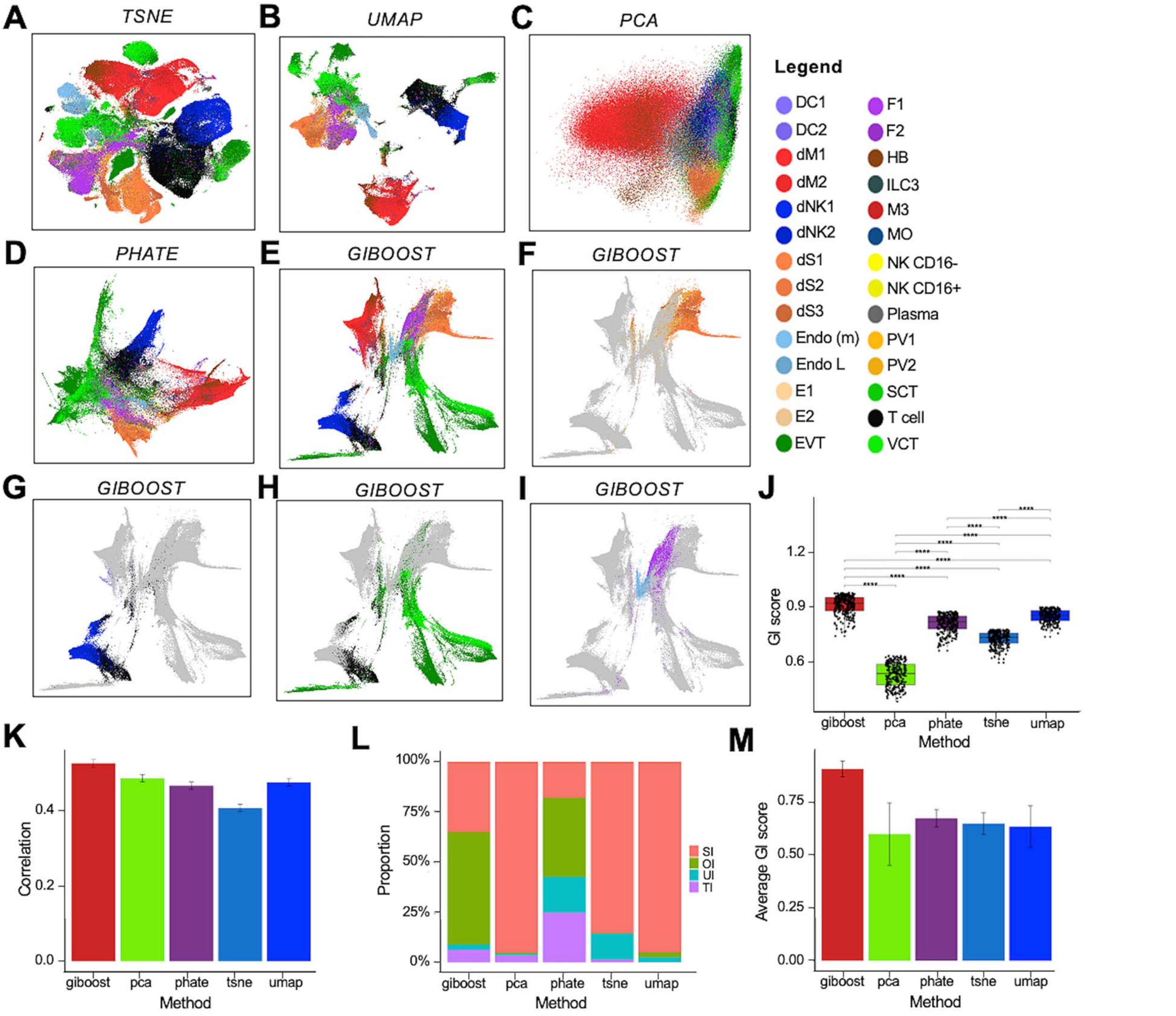


Figure 7. GIBOOST analysis of cell–cell communication during placenta development. Projection of single-cell placenta using (A). t-SNE map. (B). UMAP map. (C). PCA map. (D). PHATE map (E). GIBOOST map. GIBOOST plots highlighting major cell–cell interactions during placenta development. These include; (F). Interaction between decidual perivascular cells (PV)–decidual stromal cells (dS)–epithelial cells (E). (G). Interaction between T cells–NK cells. (H). Interaction in trophoblasts–T cells. (I). Interaction involving fibroblast–endothelial cells. (J). GI score distribution for various DRM after reducing the placenta data. (K). Correlation coefficients measuring the structural relationship between the high-dimensional placenta data and their reduced data using different DRM. (L). Proportion contributions of metric features for enhancing visualization and interpretability for GIBOOST, PCA, PHATE, t-SNE, and UMAP. (M). Average GI scores for DRMs when reducing the EMT, CiPSC, spermatogenesis, and placenta data.

including EMT, CiPSC reprogramming, and spermatogenesis. Across these datasets, GIBOOST consistently outperformed standalone DRMs, preserving structural transitions such as epithelial-to-mesenchymal progression (Fig 3G–H), intermediate pluripotent states in reprogramming (Fig. 6A), and developmental stages in spermatogenesis (Fig. 6B). This consistent performance demonstrates GIBOOST's capacity to maintain biologically meaningful trajectories and heterogeneity.

To further assess GIBOOST's ability to preserve meaningful biological structure, we evaluated how well local neighborhood relationships were maintained following dimensionality reduction using the trustworthiness score. By varying neighborhood sizes, we found that GIBOOST consistently outperformed leading DRMs like scPHERE, scVI, SIMLR, SAUCIE, and Ensemble UMAP (Fig. 4C). Beyond local structure, we also compared GIBOOST to these methods in the context of reconstructing biologically relevant trajectories. Specifically, we analyzed the EMT, a well-characterized process involving transitions through epithelial (E), partial EMT (pEMT), mesenchymal (M), and pMET states. As shown in Figs 3 and 5, GIBOOST inferred trajectories that closely aligned with the known EMT progression, whereas other methods produced distorted trajectories that misrepresented key transitions, particularly between M and pMET states. Overall, across a comprehensive suite of metrics capturing local neighborhood fidelity, global structure preservation, and trajectory inference accuracy (Fig. 4A–C), GIBOOST consistently outperformed current state-of-the-art approaches, demonstrating its robust utility for high-resolution visualization and interpretation of single-cell data.

Applying GIBOOST to placental development further demonstrated its utility. Compared to individual DRMs (Fig. 7A–E), GIBOOST showed the highest performance in preserving clustering structure, as measured by the GI and cophenetic correlation. Notably, the integration of PHATE and UMAP within GIBOOST optimized cluster separability and continuity (Fig. 7J), while cophenetic correlation analysis revealed GIBOOST preserved the highest structural integrity relative to the original data (Fig. 7K).

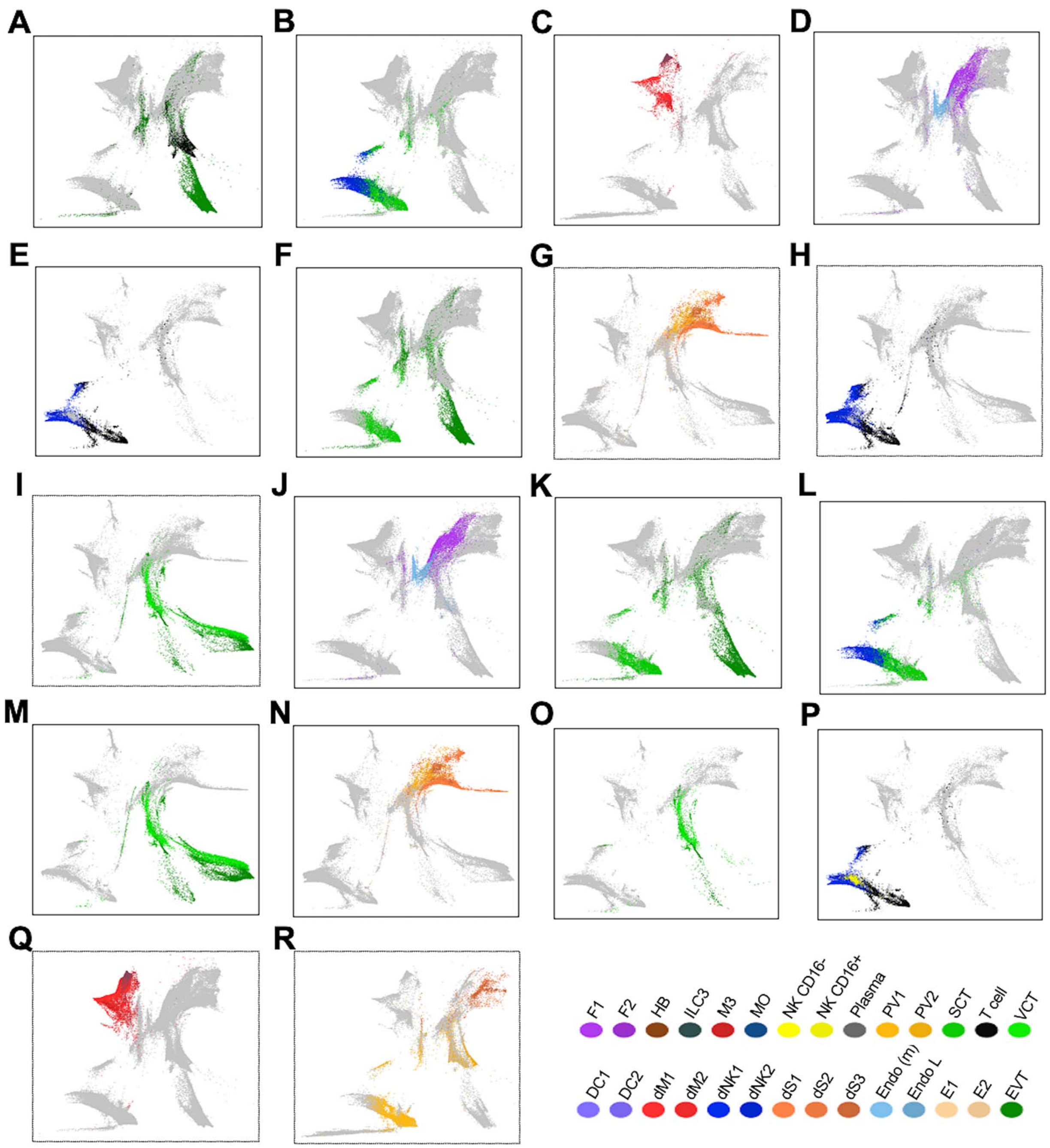


Figure 8. Visualizing cell–cell interactions in cross tissues/organs. Cell–cell interactions in Placental villi and basal plate (A-F) include: (A). T cells–EVT, (B). dNK–SCT. (C). dM–Hofbauer (HB). (D). Fibroblast–dendritic cell (DC). (E). T cells–NK. (F). EVT–SCT–VCT (villous cytotrophoblast). In the decidua (G–I) we observe: (G). Epithelial-decidual perivascular cell-decidual stromal cell interactions. (H). T cell–NK cells. (I). EVT–SCT–VCT. In the Chorioamniotic membranes (J–L) we have: (J). Fibroblast–endothelial. (K). NK–SCT. (L). EVT-SCT-VCT. In the trophoblast enriched placenta tissue (M-N) we observe: M. Epithelial-decidual perivascular cells-decidual stromal cells. (N). EVT–SCT–VCT. In blood (O-P) we find: (O). EVT–SCT–VCT interactions. (P). Macrophage–NK cells. (Q). In the basal plate; GIBOOST identifies macrophage–Hofbauer (HB) cells. (R). In a mixture of placental and basal plate cells, GIBOOST highlights interactions in epithelial-decidual perivascular cells-decidual stromal cells.

Additional metrics; separability (SI), occupation (OI), uniformity (UI), and time order (TI), confirmed that the PHATE–UMAP pair yielded complementary strengths (Fig. 7L), resulting in a 26.7% improvement over individual methods (Fig. 7M). This enhanced structural preservation enabled detailed insights into cell–cell communication during placental development. For example, GIBOOST highlighted: (i) the interconnected roles of decidual perivascular cells, decidual stromal cells, and epithelial cells in placental development (Fig. 7F); (ii) the interaction between dendritic cells and T cells, promoting the formation of regulatory T cells (Fig. 7G); (iii) the secretion of chemokines (CXCL12 and CCL2) by trophoblasts, which attract T cells to the placenta (Fig. 7H); and (iv) the stimulation of endothelial cells by fibroblasts (Fig. 7I). Focusing on specific tissues or organ samples, GIBOOST further identified a range of crucial cell–cell interactions, including: fibroblast–endothelial in chorioamniotic membranes, placental villi, and basal plate; NK–SCT in chorioamniotic membranes, placental villi, and basal plate; T cells–EVT in

Table 1. Cell–cell interactions involving immune and various placenta tissues or organ

Tissue/Organ	Interactions					
Chorioamniotic membranes	F–Endo	NK–S	E–S–V	E–S–V	dM–HB	T cell–NK
Placental villi, basal plate	F–Endo	NK–S	T cell–E			
Placental basal plate	Ep–dP–dS	T cell–NK	E–S–V			
Decidua	Ep–dP–dS	E–S–V				
Placental	Ep–dP–dS					
Basal Plate	dM–HB					
Blood	E–S–V	dM–NK				

F: Fibroblast; Endo: Endothelial cell; NK: Natural Killer cell; E: Extravillous trophoblast; S: Syncytiotrophoblast; V: Villous cytotrophoblast; Ep: Epithelial cell; dP: Decidual perivascular cells; dS: Decidual stromal cells; dM: decidual Macrophage; HB: Hofbauer cells.

placental villi and basal plate; epithelial–decidual perivascular cells–decidual stromal cells in decidua, placenta, and basal plate; Macrophage–HBs in basal plate and placental villi; T cells–NK in decidua, placental villi, and basal plate; EVT–SCT–VCT in chorioamniotic membranes, decidua, placenta, blood, placental villi, and basal plate; and T cells–NK in the blood (Fig. 8, Table 1). These findings, supported by existing experimental studies (see Supplementary Text 7), validate GIBOOST's ability to illuminate biologically relevant communication networks at scale.

To evaluate computational efficiency (Supplementary Tables S2–S5), we compared the run times of GIBOOST and several commonly used DRMs across EMT, CiPSC, and spermatogenesis datasets under standardized conditions, including identical preprocessing steps, software environments, hardware specifications, and default or comparable hyperparameter settings. GIBOOST demonstrated moderate computational demands across all datasets, with runtimes of 34, 38, and 43 min, respectively. These times were longer than lightweight methods such as PCA or UMAP, but competitive with other advanced techniques like scPHERE or SAUCIE. Notably, dataset size and modality influenced performance: the EMT dataset was considerably smaller and less complex than the CiPSC and spermatogenesis datasets, which were generated from single-cell RNA-seq and included more cell types (8 vs. 19 and 28). We acknowledge these inherent differences in data structure and complexity, as they may partially contribute to observed variation in computational runtime despite our controlled testing framework.

GIBOOST's success is rooted in its systematic evaluation of DRMs, including UMAP, t-SNE, PCA, and PHATE as candidate inputs. Rather than relying on default parameterizations or a single technique, it uses the GI metric to select the two most informative DRMs based on features such as cluster separability, spatial continuity, and biological substructure sensitivity. These are then fused via an optimized autoencoder whose hyperparameters (e.g. neuron count, batch size) are selected via grid search to maximize GI. This yields embeddings that retain both structure and interpretability, with clear traceability to the contributions of each DRM. Looking forward, we plan to replace GI with the Spearman Correlation Trajectory (SCT) score to directly optimize trajectory inference. SCT evaluates alignment between inferred and known biological trajectories, enabling GIBOOST to better support continuous processes such as differentiation and lineage progression. This shift will further strengthen GIBOOST's role in uncovering dynamic biological phenomena.

GIBOOST is designed to be modular and fully compatible with standard upstream preprocessing workflows that implement different dimensionality reduction tools. Users are encouraged to apply existing batch correction methods such as Harmony, Seurat's integration, or ComBat prior to applying GIBOOST to minimize batch-related variation. Future iterations of GIBOOST will incorporate native batch effect correction capabilities to enhance visualization and interpretability.

Additional improvements include treating DRM parameters as tunable hyperparameters in the Bayesian framework, allowing end-to-end optimization of the full EVI pipeline. GIBOOST could also be extended to integrate three or more DRMs, increasing flexibility. Finally, given the growing variety of interpretability metrics [8] (e.g. silhouette coefficient [41]), GIBOOST offers a generalizable framework for enhancing clustering sensitivity in both supervised and unsupervised settings. Furthermore, while deep learning models like autoencoders are powerful, they are stochastic and often lack interpretability [55]. Methods like GIBOOST could help enhance the interpretability nature of neural networks.

In summary, GIBOOST offers an innovative and efficient solution for improving the visualization and interpretation of high-dimensional single-cell data. By systematically selecting and integrating complementary DRMs, it maximizes the preservation of local and global biological structure, enabling deeper insights into cellular dynamics and tissue-level communication. Its modular design and superior performance across diverse tasks make it a powerful tool for advancing single-cell analysis.

Data availability

The spermatogenesis dataset used in this study includes ∼110 000 cells across 25 developmental stages, processed from raw scRNA-seq data available from the Gene Expression Omnibus (GEO) under accession codes GSE121904 [56], GSE124904 [21] and GSE117707 [57], as well as from the ArrayExpress database under accession code EMTAB-6946 [58]. The processed CiPSC dataset, containing about 50 000 cells across 12 time points, was obtained from raw scRNA-seq data of mouse MEFs and chemically induced CiPSCs from GEO under accession code GSE114952 [20]. Detailed preprocessing of these scRNA-seq datasets, including the generation of cluster labels and marker selection, is described in Anchang *et al.* [30]. Additionally, arcsinh-transformed CyTOF time-course data for EMT analysis was downloaded from Karacosta *et al.* [19].

For the placenta development GIBOOST analysis, we used decision-tree models described in Anchang *et al.* [30] to identify 41 (Supplementary Table S1) most relevant genes across 341 090 cells that capture the dynamic and heterogeneous nature of the data. In addition, cell type annotation was performed on the integrated Seurat scRNA-seq dataset (see Supplementary Text 6). We used a two-pronged approach: first, curating cell type-specific marker genes from Vento R. *et al.* paper [43], and second, utilizing scType [59], a computational tool for cell type annotation using scRNA-seq data. By integrating these marker gene sets with immune system markers within scType, each cell was assigned a putative cell type label based on its gene expression profile. Supplementary processed data for the EMT, CiPSC, spermatogenesis, and placenta development datasets used in this study are available as R Source Data files at https://github.com/NIEHS/GIBOOST.git.

Key Points

- GIBOOST is an AI-driven framework that integrates multiple dimensionality reduction methods to enhance visualization and interpretability of high-dimensional single-cell data.
- Unlike traditional DRMs that optimize specific visualization objectives, GIBOOST selects and integrates the most informative methods, preserving both global structure and local structure.
- GIBOOST improves clustering sensitivity and biological relevance by about 30%, enabling more accurate identification of differentiation trajectories, cell–cell communication, and tissue interactions.
- GIBOOST uncovered novel immune-placenta interactions, providing insights into cross-tissue communication during pregnancy and broader applications in cancer, stem cell biology, and disease modeling.

Author contributions

K.A. and B.A. contributed to the concept and algorithm of GIBOOST. K.A. and B.A. were involved in the algorithm implementation. B.A., M.K., I.A., B.P., J.L., and J-L. L. were involved in data generation, pre-processing and data integration. K.A. J.L., O.E., M.K,. I.A., J-L. L., and B.A. were involved in writing the initial manuscript. All authors were involved in interpreting the results.

Supplementary data

Supplementary data are available at *Briefings in Bioinformatics* online.

Conflict of interest: None declared.

Funding

This work was supported by funding from grant number 1ZIAES103350-05 from NIH/NIEHS/DIR and from Chan Zuckerberg Initiative DAF, an advised fund of Silicon Valley Community Foundation with grant numbers: 2021-240134 (5022).

References

1. Buenrostro JD, Wu B, Litzenburger UM. *et al.* Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* 2015;**523**:486–90. https://doi.org/10.1038/nature14590
2. Chen H, Albergante L, Hsu JY. *et al.* Single-cell trajectories reconstruction, exploration and mapping of omics data with STREAM. *Nat Commun* 2019;**10**:1–14. https://doi.org/10.1038/s41467-019-09670-4
3. Satpathy AT, Granja JM, Yost KE. *et al.* Massively parallel single-cell chromatin landscapes of human immune cell development and intratumoral T cell exhaustion. *Nat Biotechnol* 2019;**37**:925–36. https://doi.org/10.1038/s41587-019-0206-z
4. Van der Maaten L, Hinton G. Visualizing data using t-SNE. *J Mach Learn Res* 2008;**9**:2579–2605.
5. Becht E, McInnes L, Healy J. *et al.* Dimensionality reduction for visualizing single-cell data using UMAP. *Nat Biotechnol* 2019;**37**:38–44. https://doi.org/10.1038/nbt.4314
6. Van Der Maaten L, Postma E, Van den Herik J. Dimensionality reduction: A comparative. *J Mach Learn Res* 2009;**10**:1–41.
7. Moon KR, Van Dijk D, Wang Z. *et al.* Visualizing structure and transitions in high-dimensional biological data. *Nat Biotechnol* 2019;**37**:1482–92. https://doi.org/10.1038/s41587-019-0336-3
8. Atitey K, Motsinger-Reif AA, Anchang B. Model-based evaluation of spatiotemporal data reduction methods with unknown ground truth through optimal visualization and interpretability metrics. *Brief Bioinform* 2024;**25**:bbad455. https://doi.org/10.1093/bib/bbad455
9. Marx V. Seeing data as t-SNE and UMAP do. *Nat Methods* 2024;**21**:930–3. https://doi.org/10.1038/s41592-024-02301-x
10. VanHorn KC, Çobanoğlu MC. Haisu: Hierarchically supervised nonlinear dimensionality reduction. *PLoS Comput Biol* 2022;**18**:e1010351. https://doi.org/10.1371/journal.pcbi.1010351
11. Ding J, Regev A. Deep generative model embedding of single-cell RNA-Seq profiles on hyperspheres and hyperbolic spaces. *Nat Commun* 2021;**12**:2554. https://doi.org/10.1038/s41467-021-22851-4
12. Mcinnes L, Healy J, Melville J. Umap: Uniform manifold approximation and projection for dimension reduction, arXiv preprint arXiv:1802.03426 2018.
13. Lopez R, Regier J, Cole MB. *et al.* Deep generative modeling for single-cell transcriptomics. *Nat Methods* 2018;**15**:1053–8. https://doi.org/10.1038/s41592-018-0229-2
14. Wang B, Ramazzotti D, De Sano L. *et al.* SIMLR: A tool for large-scale single-cell analysis by multi-kernel learning. *Proteomics* 2018;**18**:2. https://doi.org/10.1002/pmic.201700232
15. Amodio M, Van Dijk D, Srinivasan K. *et al.* Exploring single-cell data with deep multitasking neural networks. *Nat Methods* 2019;**16**:1139–45. https://doi.org/10.1038/s41592-019-0576-7
16. Stein-O'Brien GL, Clark BS, Sherman T. *et al.* Decomposing cell identity for transfer learning across cellular measurements, platforms, tissues, and species. *Cell systems* 2019;**8**:e398.
17. Bravo González-Blas C, De Winter S, Hulselmans G. *et al.* SCENIC+: Single-cell multiomic inference of enhancers and gene regulatory networks. *Nat Methods* 2023;**20**:1355–67. https://doi.org/10.1038/s41592-023-01938-4
18. Argelaguet R, Arnol D, Bredikhin D. *et al.* MOFA+: A statistical framework for comprehensive integration of multi-modal single-cell data. *Genome Biol* 2020;**21**:1–17. https://doi.org/10.1186/s13059-020-02015-1
19. 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**:5587. https://doi.org/10.1038/s41467-019-13441-6
20. 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. https://doi.org/10.1016/j.stem.2018.05.025
21. Law NC, Oatley MJ, Oatley JM. Developmental kinetics and transcriptome dynamics of stem cell specification in the spermatogenic lineage. *Nat Commun* 2019;**10**:2787. https://doi.org/10.1038/s41467-019-10596-0
22. Suryawanshi H, Morozov P, Straus A. *et al.* A single-cell survey of the human first-trimester placenta and decidua. *Sci Adv* 2018;**4**:eaau4788. https://doi.org/10.1126/sciadv.aau4788

23. Sibuya M, Yoshimura I, Shimizu R. Negative multinomial distribution. *Ann Inst Stat Math* 1964;**16**:409–26. https://doi.org/10.1007/BF02868583
24. Ruder S. An overview of gradient descent optimization algorithms, arXiv preprint arXiv:1609.04747 2016.
25. Wang F, Ross CL. Machine learning travel mode choices: Comparing the performance of an extreme gradient boosting model with a multinomial logit model. *Transp Res Rec* 2018;**2672**:35–45. https://doi.org/10.1177/0361198118773556
26. Chen T, He T, Benesty M. *et al.* Package ‘xgboost’. *R version* 2019;**90**:40.
27. Rohmer J, Gehl P. Sensitivity analysis of Bayesian networks to parameters of the conditional probability model using a Beta regression approach. *Expert Syst Appl* 2020;**145**:113130. https://doi.org/10.1016/j.eswa.2019.113130
28. Atitey K. DEGBOE: Discrete time evolution modeling of gene mutation through Bayesian inference using qualitative observation of mutation events. *J Biomed Inform* 2022;**134**:104197. https://doi.org/10.1016/j.jbi.2022.104197
29. Zhang Z. Improved Adam optimizer for deep neural networks. In: Zhang Z (ed.), *Proceedings of the 2018 IEEE/ACM 26th International Symposium on Quality of Service (IWQoS); June 4–6, 2018; Banff, Alberta, Canada*. Piscataway, NJ: IEEE; 2018, 1–2.
30. 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. https://doi.org/10.1093/bib/bbac017
31. Schober P, Boer C, Schwarte LA. Correlation coefficients: Appropriate use and interpretation. *Anesth Analg* 2018;**126**:1763–8. https://doi.org/10.1213/ANE.0000000000002864
32. So S-S, Karplus M. Three-dimensional quantitative structure– activity relationships from molecular similarity matrices and genetic neural networks. 1. Method and validations. *J Med Chem* 1997;**40**:4347–59. https://doi.org/10.1021/jm970487v
33. Demirkan B. The roles of epithelial-to-mesenchymal transition (EMT) and mesenchymal-to-epithelial transition (MET) in breast cancer bone metastasis: Potential targets for prevention and treatment. *J Clin Med* 2013;**2**:264–82. https://doi.org/10.3390/jcm2040264
34. Celià-Terrassa T, Kang Y. How important is EMT for cancer metastasis? *PLoS Biol* 2024;**22**:e3002487. https://doi.org/10.1371/journal.pbio.3002487
35. Bakir B, Chiarella AM, Pitarresi JR. *et al.* EMT, MET, plasticity, and tumor metastasis. *Trends Cell Biol* 2020;**30**:764–76. https://doi.org/10.1016/j.tcb.2020.07.003
36. Li C-H, Hsu T-I, Chang Y-C. *et al.* Stationed or relocating: The seesawing emt/met determinants from embryonic development to cancer metastasis. *Biomedicine* 2021;**9**:1265. https://doi.org/10.3390/biomedicines9091265
37. Kawakami E, Kobayashi N, Ichihara Y. *et al.* Monitoring of blood biochemical markers for periprosthetic joint infection using ensemble machine learning and UMAP embedding. *Arch Orthop Trauma Surg* 2023;**143**:6057–67. https://doi.org/10.1007/s00402-023-04898-8
38. Street K, Risso D, Fletcher RB. *et al.* Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. *BMC Genomics* 2018;**19**:477. https://doi.org/10.1186/s12864-018-4772-0
39. Zar JH. Spearman rank correlation. In: Armitage P, Colton T (eds.), *Encyclopedia of biostatistics* 2005;**7**. https://doi.org/10.1002/0470011815.b2a15150
40. Venna J, Peltonen J, Nybo K. *et al.* Information retrieval perspective to nonlinear dimensionality reduction for data visualization. *J Mach Learn Res* 2010;**11**:451–90.
41. Carter AM. Animal models of human placentation–a review. *Placenta* 2007;**28**:S41–7. https://doi.org/10.1016/j.placenta.2006.11.002
42. Digitale E. Immune system changes during pregnancy are precisely timed, *Lastet ned* 2017;**24**:2020. Hentet fra https://med.stanford.edu/news/all-news/2017/09/immune-systemchanges-during-pregnancy-are-precisely-timed.html
43. Vento-Tormo R, Efremova M, Botting RA. *et al.* Single-cell reconstruction of the early maternal–fetal interface in humans. *Nature* 2018;**563**:347–53. https://doi.org/10.1038/s41586-018-0698-6
44. Liu Y, Fan X, Wang R. *et al.* Single-cell RNA-seq reveals the diversity of trophoblast subtypes and patterns of differentiation in the human placenta. *Cell Res* 2018;**28**:819–32. https://doi.org/10.1038/s41422-018-0066-y
45. Suryawanshi H, Max K, Bogardus KA. *et al.* Dynamic genome-wide gene expression and immune cell composition in the developing human placenta. *J Reprod Immunol* 2022;**151**:103624. https://doi.org/10.1016/j.jri.2022.103624
46. Guo C, Cai P, Jin L. *et al.* Single-cell profiling of the human decidual immune microenvironment in patients with recurrent pregnancy loss. *Cell discovery* 2021;**7**:1. https://doi.org/10.1038/s41421-020-00236-z
47. Pique-Regi R, Romero R, Tarca AL. *et al.* Single cell transcriptional signatures of the human placenta in term and preterm parturition. *elife* 2019;**8**:e52004. https://doi.org/10.7554/eLife.52004
48. Yang Y, Guo F, Peng Y. *et al.* Transcriptomic profiling of human placenta in gestational diabetes mellitus at the single-cell level. *Front Endocrinol* 2021;**12**:679582. https://doi.org/10.3389/fendo.2021.679582
49. Garcia-Flores V, Romero R, Xu Y. *et al.* Maternal-fetal immune responses in pregnant women infected with SARS-CoV-2. *Nat Commun* 2022;**13**:320. https://doi.org/10.1038/s41467-021-27745-z
50. Liu Y, Fan X, Wang R. *et al.* Single-cell RNA-seq reveals the diversity of trophoblast subtypes and patterns of differentiation in the human placenta. *Cell Res* 2018;**28**:819–32. https://doi.org/10.1038/s41422-018-0066-y
51. Barrozo ER, Aagaard KM. Human placental biology at single-cell resolution: A contemporaneous review, BJOG. *An International Journal of Obstetrics & Gynaecology* 2022;**129**:208–20. https://doi.org/10.1111/1471-0528.16970
52. Malepathirana T, Senanayake D, Vidanaarachchi R. *et al.* Dimensionality reduction for visualizing high-dimensional biological data. *Biosystems* 2022;**220**:104749. https://doi.org/10.1016/j.biosystems.2022.104749
53. Efremova M, Vento-Tormo M, Teichmann SA. *et al.* CellPhoneDB: Inferring cell–cell communication from combined expression of multi-subunit ligand–receptor complexes. *Nat Protoc* 2020;**15**:1484–506. https://doi.org/10.1038/s41596-020-0292-x
54. Cabello-Aguilar S, Alame M, Kon-Sun-Tack F. *et al.* SingleCellSignalR: Inference of intercellular networks from single-cell transcriptomics. *Nucleic Acids Res* 2020;**48**:e55–5. https://doi.org/10.1093/nar/gkaa183
55. Wysocka M, Wysocki O, Zufferey M. *et al.* A systematic review of biologically-informed deep learning models for cancer: Fundamental trends for encoding and interpreting oncology data. *BMC bioinformatics* 2023;**24**:198. https://doi.org/10.1186/s12859-023-05262-8

56. 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. https://doi.org/10.1371/journal.pgen.1007810

57. 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**:5196. https://doi.org/10.1038/s41467-019-13193-3

58. Ernst C, Eling N, Martinez-Jimenez CP. *et al.* Staged developmental mapping and X chromosome transcriptional dynamics during mouse spermatogenesis. *Nat Commun* 2019;**10**:1251. https://doi.org/10.1038/s41467-019-09182-1

59. Ianevski A, Giri AK, Aittokallio T. Fully-automated and ultra-fast cell-type identification using specific marker combinations from single-cell transcriptomic data. *Nat Commun* 2022;**13**:1246. https://doi.org/10.1038/s41467-022-28803-w

Briefings in Bioinformatics, 2025, **26(4)**, bbaf415
https://doi.org/10.1093/bib/bbaf415
Problem Solving Protocol