

Reviewed Preprint

v2 • December 18, 2024

Revised by authors

Reviewed Preprint

v1 • August 21, 2024

MGPfact^{XMBD}: A Model-Based Factorization Method for scRNA Data Unveils Bifurcating Transcriptional Modules Underlying Cell Fate Determination

Jun Ren, Ying Zhou, Yudi Hu, Jing Yang, Hongkun Fang, Xuejing Lyu, Jintao Guo, Xiaodong Shi, Qiyuan Li

School of Informatics, Xiamen University, Xiamen, China • National Institute for Data Science in Health and Medicine, School of Medicine, Xiamen University, Xiamen, China • Department of Hematology, The First Affiliated Hospital of Xiamen University and Institute of Hematology, School of Medicine, Xiamen University, Xiamen, China

https://en.wikipedia.org/wiki/Open_access

Copyright information

eLife Assessment

MGPfact^{XMBD} is a novel computational method for investigating cell evolutionary trajectory for scRNA-seq samples. It is **important**, with several potential future applications. The authors benchmarked this method using synthetic and real-world samples and showed superior performance for some of the tasks in cell trajectory analysis compared to other methods with **compelling** evidence.

<https://doi.org/10.7554/eLife.97424.2.sa3>

Abstract

Manifold-learning is particularly useful to resolve the complex cellular state space from single-cell RNA sequences. While current manifold-learning methods provide insights into cell fate by inferring graph-based trajectory at cell level, challenges remain to retrieve interpretable biology underlying the diverse cellular states. Here, we described MGPfact^{XMBD}, a model-based manifold-learning framework and capable to factorize complex development trajectories into independent bifurcation processes of gene sets, and thus enables trajectory inference based on relevant features. MGPfact^{XMBD} offers more nuanced understanding of the biological processes underlying cellular trajectories with potential determinants. When bench-tested across 239 datasets, MGPfact^{XMBD} showed advantages in major quantity-control metrics, such as branch division accuracy and trajectory topology, outperforming most established methods. In real datasets, MGPfact^{XMBD} recovered the critical pathways and cell types in microglia development with experimentally valid regulons and markers. Furthermore, MGPfact^{XMBD} discovered evolutionary trajectories of tumor-associated CD8⁺ T cells and yielded new subtypes of CD8⁺ T cells with gene expression signatures significantly predictive of the responses to immune checkpoint inhibitor in independent cohorts. In summary, MGPfact^{XMBD} offers a manifold-learning framework in

scRNA-seq data which enables feature selection for specific biological processes and contributing to advance our understanding of biological determination of cell fate.

Introduction

Data-mining of single-cell RNA sequencing (scRNA-seq) is often transformed into learning of lower-dimensional embedding (Becht et al., 2019 [↗](#); Haghverdi et al., 2015 [↗](#); Van der Maaten and Hinton, 2008 [↗](#)) of the expression vectors, which represents the variation in the cellular space and helps explain the biological background. Previous single-cell studies used various embedding methods to characterize and visualize clustering of cells with unique biological functions (Saelens et al., 2019 [↗](#)). Among the existing methods, graph-based embedding can better capture nonlinear biological signals among cells hence yielded more insights of the diversity of cells. More recent studies also use graph-based embedding (Costa et al., 2018 [↗](#)) to reveal the dependency among cells and thereby reconstruct the evolutionary trajectory in the cellular space, which helps in understanding the determination of cell fate in development, differentiation and cancer.

To date, more and more manifold-learning methods are developed to infer lower-dimensional graphic embedding (manifolds) of scRNA-seq data, and yielded a number of trajectories corresponding to important cellular processes, such as TSCAN (Ji and Ji, 2016 [↗](#)), DPT (Haghverdi et al., 2016 [↗](#)), and scShaper (Smolander et al., 2022 [↗](#)) belong to linear topological classes and reveal major linear pathways based on embedding spaces or cell clustering. TSCAN employs the construction of minimum spanning trees to discover pathways, while DPT reconstructs cellular trajectories using random walks, and scShaper integrates multiple pseudo-temporal solutions to deduce the shortest trajectory within a linear context. Additionally, there have been many approaches capable of inferring complex tree topological structures, such as the widely used Monocle series of algorithms includes Monocle 2 and 3 (Cao et al., 2019 [↗](#); Qiu et al., 2017b [↗](#), 2017a [↗](#)). They leverage sophisticated graphing techniques to map intricate cell hierarchies; Monocle 2 creates DDRtree based on reversed graph embedding techniques, while Monocle 3 utilizes UMAP (Becht et al., 2019 [↗](#)) for embedding. TinGa (Todorov et al., 2020 [↗](#)) and scFates (Faure et al., 2023 [↗](#)) represent more recent innovations. TinGa utilizes the Growing Neural Gas (GNG) algorithm (Fritzke, 1994 [↗](#)) to construct an adaptive graph structure that effectively captures the density structure of the dataset. scFates, streamlines pseudotime analysis with flexible tree learning options, advanced feature extraction tasks, and specific functionalities for fork analysis.

Moreover, recent studies based on RNA velocity has provided insights into cell state transitions. These methods measure RNA synthesis and degradation rates based on the abundance of spliced and unspliced mRNA, such as CellRank (Lange et al., 2022 [↗](#)). Nevertheless, current RNA velocity analyses are still unable to resolve cell-fates with complex branching trajectory. Deep learning methods such as scTour (Li, 2023 [↗](#)) and TIGON (Sha, 2024 [↗](#)) circumvent some of these limitations, offering continuous state assumptions or requiring prior cell sampling information.

Despite these advances, trajectory prediction remains a major challenge in single cell analysis. Firstly, graph-based trajectories represent synergic effects multiple biological processes, making it difficult to disentangle the effects of specific process, hence limited model interpretability and the power to gain novel biological insights. Secondly, the inference of trajectory is highly dependent on the biases in the gene-selection, whereas conventional statistical feature-selection methods are less efficient for the learning of complex topologies, and adds to the difficulty of suggesting candidate genes for downstream functional study. Additionally, many approaches require additional prior information, which further limits the applicability.

Here, we describe MGPfact^{XMBD} (Factorization based on Mixtures of Gaussian Processes), a model-based, unsupervised manifold-learning method which factorize complex cellular trajectories into interpretable bifurcation Gaussian processes of transcription, and thereby enable discovery of specific biological determinants of cell fate. In the validation datasets, MGPfact recapitulated developmental trajectory of microglia and recovered key regulatory factors which have been proved experimentally. Moreover, MGPfact discovered highly specific subtypes of tumor-associated CD8⁺ T cells which associated with benefit to cancer immunotherapy.

Bring together, MGPfact is a knowledge discovery tool which conducts manifold-learning and factorization simultaneously. MGPfact offers two advantages in future scRNA-seq analyses: first, it provides highly interpretable, factorizable cellular trajectories with matched gene expression modules; then, it provides efficient feature-selection for graph-based embedding, thus enhancing our understanding of the determination of cell fate.

Results

Design of MGPfact

The analytical pipeline of MGPfact consists two major stages (**Fig. 1** [↗](#)). An algorithmic description is given in Algorithm 1.

First, we performed downsampling of the preprocessed scRNA-seq data \mathbf{Y} to yield a M -by- N expression matrix \mathbf{Y}' based on the “minimum unbiased representative points” (MURP) as described previously (Ren et al., 2022 [↗](#)), where M representative points were considered as landmarks of the cellular trajectory and N is the number of genes. Then, we computed L Principal Components (PCs) of the downsampled expression matrix to obtain the matrix $\mathbf{Y}^* = \{\mathbf{y}_1^*, \mathbf{y}_2^*, \mathbf{y}_3^*, \dots, \mathbf{y}_L^*\}$ (M -by- L),

$$\mathbf{y}_l^* = \mathbf{Y}' \cdot \mathbf{v}_l \quad (1)$$

where \mathbf{v}_l is projection vector, serve as the l -th initial state of embedding.

Next, we used typical Gaussian Process Regression of \mathbf{y}_l^* on pseudotime T :

$$\mathbf{y}_l^* = f(T) + \varepsilon \quad (2)$$

where $f(T)$ is a Gaussian Process (GP) with covariance matrix \mathbf{S} .

$$f(T) = \mathcal{GP}(0, \mathbf{S} + \sigma_s^2 \cdot \mathbf{I}) \quad (3)$$

And for all features:

$$p(\mathbf{Y}^*, f(T)) = p(\mathbf{Y}^* | f(T)) \cdot p(f(T)) \quad (4)$$

where $p(\mathbf{Y}^* | f(T))$ is defined as follow:

$$p(\mathbf{Y}^* | f(T)) = \prod_{l=1}^L p(\mathbf{y}_l^* | f(T)) = \prod_{l=1}^L \mathcal{N}(\mathbf{y}_l^* | 0, \mathbf{S} + \sigma_s^2 \cdot \mathbf{I}) \quad (5)$$

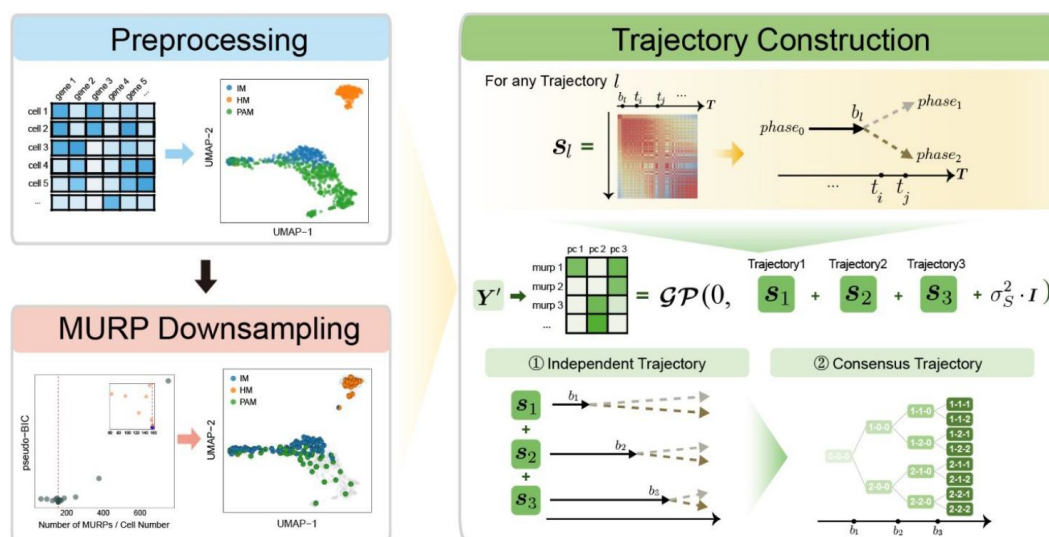


Fig 1.

Overview of MGPfact workflow.

The complete workflow comprises two major stages: MURP downsampling with preprocessed data and trajectory reconstruction. In the stage of trajectory reconstruction, the scRNA-seq data were first factorized into independent bifurcation processes based on mixtures of Gaussian processes, which were then merged into a consensus trajectory.

Specifically, we consider \mathbf{S} as a mixture of L independent bifurcation Gaussian processes (Schulz et al., 2018 [DOI](#)),

$$\mathbf{S} = \sum_{l=1}^L \mathbf{s}_l \quad (6)$$

To cope with the bifurcation processes in cell fate, each of the Gaussian processes is defined with a bifurcation point at b_l , branching labels \mathbf{c}_l , and the necessary hyperparameters. The branching labels $\mathbf{c}_l \in \{0, 1, 2\}$, correspond to different phases and states of cell fate, where $\mathbf{c}_l = \mathbf{0}$ corresponds the phase before branching, and $\mathbf{c}_l \in \{1, 2\}$, corresponds to the two cellular states of the bifurcation process, respectively. For any landmark (MURP) x ,

$$\begin{cases} c_{l,x} = 0, & \text{if } t_x < b_l \\ c_{l,x} \in \{1, 2\}, & \text{if } t_x \geq b_l \end{cases} \quad (7)$$

The covariance matrix \mathbf{s}_l for trajectory l can be expressed as follows,

$$[\mathbf{s}_l]_{x,y} = \mathcal{K}(t_x, t_y) \quad (8)$$

where \mathcal{K} is a kernel function. We employ radial basis function (rbf) and polynomial kernel function (pl). We chose these two kernel functions for the effectiveness in handling nonlinear and polynomial relationships, achieving a balance between model performance and computational efficiency.

$$k_{rbf}(t_x, t_y) = \lambda_{rbf} \cdot e^{(-\alpha_{rbf} \|t_x - t_y\|^2)} \quad (9)$$

$$k_{pl}(t_x, t_y) = (\lambda_{pl} t_x^T t_y + c_{pl})^{d_{pl}} \quad (10)$$

And the $\mathcal{K}(t_x, t_y)$ is calculated as follows:

$$\mathcal{K}(t_x, t_y) = \begin{cases} k_{rbf}(t_x, t_y) + k_{pl}(t_x, t_y) & t_x, t_y < b_l \\ k_{rbf}(t_x, t_y) + k_{pl}(t_x - b_l, t_y - b_l) & t_x, t_y > b_l, c_{l,x} = c_{l,y} \\ \frac{k_{rbf}(t_x, b_l) \cdot k_{rbf}(b_l, t_y)}{k_{rbf}(b_l, b_l)} & c_{l,x} \neq c_{l,y} \end{cases} \quad (11)$$

Therefore, $p(\mathbf{Y}^* | f(\mathbf{T}))$ is updated as follows:

$$p(\mathbf{Y}^* | f(\mathbf{T})) = \prod_{l=1}^L \mathcal{N}(\mathbf{y}_l^* | 0, \sum_{l=1}^L \mathbf{s}_l^* + \sigma_s^2) \quad (12)$$

We infer all parameters by maximizing the posterior likelihood using Markov Chain Monte Carlo (MCMC) methods available in Mamba (B J, 2014 [DOI](#)). The posterior distribution of pseudotime \mathbf{T} can be represented as:

$$p(\mathbf{T} | \mathbf{Y}^*) \propto p(\mathbf{Y}^* | f(\mathbf{T})) \cdot p(f(\mathbf{T})) \quad (13)$$

where $p(\mathbf{Y}^* | f(\mathbf{T}))$ is the likelihood function of the observed data \mathbf{Y}^* , and $p(f(\mathbf{T}))$ is the prior distribution of the Gaussian process. This posterior distribution integrates the observed data with model priors, enabling inference of pseudotime and trajectory simultaneously. Due to the high

autocorrelation of T in the posterior distribution, we use Adaptive Metropolis within Gibbs (AMWG) sampling (Roberts and Rosenthal, 2009 [\[10\]](#); Tierney, 1994 [\[11\]](#)). Other parameters are estimated using the more efficient SLICE sampling technique (Neal, 2003 [\[12\]](#)).

Algorithm 1 MGPfact: infer cell fate trajectory

Algorithm 1 MGPfact: infer cell fate trajectory

INPUT: expression matrix \mathbf{Y} , independent trajectories number L
OUTPUT: $\theta = \{T, B, C, \text{other hyperparameters}\}$; trajectory topology

- 1: initialize all parameters in θ
- 2: covariance matrix $S = 0$
- 3: object function $\mathcal{L} = 0$
- 4: $\mathbf{Y}' \leftarrow$ MURP downsampling $Y_{p,n}$;
- 5: $\mathbf{Y}^* \leftarrow$ PCA analysis
- 6: $\theta \leftarrow$ Optimized ObjectF
- 7: $Graph \leftarrow$ Create independent trajectory using θ
- 8: **function** OBJECTF($\mathbf{Y}^*, T, b_l, c_l$)
- 9: $Q \leftarrow \dim(\mathbf{Y}^*)[2]$
- 10: **for** $q = 1; q < Q; q++$ **do** ▷ object function
- 11: $v \leftarrow [\mathbf{Y}^*]_q$
- 12: **for** $l = 1; l < L; l++$ **do**
- 13: $s_l \leftarrow \text{Cov}(T, b_l, c_l)$
- 14: $S \leftarrow S + s_l$
- 15: **end for**
- 16: $\mathcal{L} \leftarrow \mathcal{L} + \text{MultivariateNormalPDF}(v, 0, S)$
- 17: **end for**
- 18: **return** \mathcal{L}
- 19: **end function**
- 20: **function** COV(T, b_l, c_l) ▷ construct covariance matrix
- 21: $P \leftarrow \text{length}(T)$
- 22: **for** $i = 1; i < P; i++$ **do**
- 23: **for** $j = 1; j < P; j++$ **do**
- 24: $s_l[i, j] \leftarrow \text{KERNEL}(t_i, t_j, b_l, c_l)$
- 25: **end for**
- 26: **end for**
- 27: **return** s_l
- 28: **end function**
- 29: **function** KERNEL($t_i, t_j, b_l, c_{l,i}, c_{l,j}$) ▷ kernel function
- 30: **if** $t_i, t_j < b_l$ **then**
- 31: **return** $k_{rbf}(t_i, t_j) + k_{pl}(t_i, t_j)$
- 32: **else if** $t_i, t_j > b_l \wedge c_{l,i} = c_{l,j}$ **then**
- 33: **return** $k_{rbf}(t_i, t_j) + k_{pl}(t_i - b_l, t_j - b_l)$
- 34: **else if** $c_{l,i} \neq c_{l,j}$ **then**
- 35: **return** $\frac{k_{rbf}(t_i, b_l) \cdot k_{rbf}(b_l, t_j)}{k_{rbf}(b_l, b_l)}$
- 36: **end if**
- 37: **end function**

Then, MGPfact can identify genes that have significant impacts on the branching events in the trajectories. we introduce a rotation matrix $R = \{r_1, r_2, \dots, r_L\}$ to obtain factor score w_l for each trajectory l by rotating \mathbf{Y}^* .

$$w_l = \mathbf{Y}^* \cdot \mathbf{r}_l + e_l^2 \quad (14)$$

For all trajectories,

$$p(\mathbf{W}|\mathbf{Y}^*) = \prod_{l=1}^L [\mathcal{N}(\mathbf{Y}^* \cdot \mathbf{r}_l + e_l | 0, \mathbf{s}_l) \cdot \mathcal{N}(e_l | 0, \sigma_{error}^2)] \quad (15)$$

where e_l is the error term for the l -th trajectory.

Specifically, the factor scores w_l for each gene onto the l -th trajectory can be expressed using equations (1) and (14) as follows,

$$\mathbf{w}_l = [\mathbf{Y}' \cdot \mathbf{v}_l] \cdot \mathbf{r}_l + e_l^2 = \mathbf{Y}' \cdot \mathbf{u}_l + e_l^2 \quad (16)$$

Here, \mathbf{u}_l is used to represent the contribution (gene weight) of each gene to the l -th trajectory, thus enabling gene-selection based on the inferred trajectories.

Additionally, we can combine independent bifurcation processes to form a consensus diffusion tree to represent the trajectory of cell fate (Supplementary Methods, Supplementary Table 1).

Performance evaluation of MGPfact

Robustness analysis of MGPfact

Before the performance evaluation, we performed a grid search on the number of independent trajectories in 100 training datasets and selected $L = 3$ for downstream testing (Supplementary Fig. 1, Supplementary Fig. 2, Methods).

To further validate the efficacy of MURP downsampling step in MGPfact, we employed an alternative downsampling using randomly selected cells for trajectory inference (Methods). This comparison revealed that the prediction accuracy substantially diminished without MURP, evidenced by a notable reduction in branch assignment ($F1_{branches}$, 20.5%) and cell ordering (cor_{dist} , 64.9%) (Supplementary Fig. 3). In contrast, trajectory predictions utilizing MURP-based downsampling realized an overall score increase of 5.31% to 185%, underscoring the indispensable role of MURP in the trajectory inference capabilities of MGPfact.

Furthermore, we performed a robustness analysis on the topological consistency of the predicted consensus trajectory by comparing prediction results from randomly sampled subsets of the original data. As a result, the consensus trajectory predicted from random subsets by MGPfact retained a high degree of congruence with those from the original datasets (Supplementary Fig. 4, $HIM_{mean} = 0.686$). This outcome attests to MGPfact's robustness and generalizability under varying data conditions, hence the capability of retrieving conserved bifurcative trajectories in the data.

MGPfact predicts cellular trajectories

Then, we assessed the performance of MGPfact for prediction of cellular trajectories in 239 test datasets, including 171 synthetic and 68 real datasets, alongside with another 7 existing algorithms (Supplementary Fig. 1). For overall performance score, MGPfact ($Overall_{mean} = 0.534$) ranked second only to TinGa ($Overall_{mean} = 0.563$) and outperformed the rest of 6 algorithms (Fig. 2a [↗](#)). Particularly, MGPfact demonstrated the highest accuracy in predicting cell fate in branching trajectory (Fig. 2b [↗](#), $F1_{branches_{mean}} = 0.482$). As for other three individual metrics, MGPfact ranked the 4th in HIM ($HIM_{mean} = 0.606$), the 6th in cor_{dist} ($cor_{dist_{mean}} = 0.507$), and the 4th in $wcor_{features}$ ($wcor_{features_{mean}} = 0.712$) (Fig. 2c-e).

Trajectory Type	DPT	Monocle DDRTree	Monocle3	scFates Tree	scShaper	TinGa	TSCAN
Acyclic Graph	0.169	0.735	0.341	0.167	0.006	0.485	0.002
Bifurcation	0.559	0.001	0.015	0.298	0.010	0.772	0.002
Convergence	0.275	0.059	0.000	0.027	0.120	0.871	0.008
Cycle	0.000	0.000	0.002	0.114	0.991	0.659	0.982
Disconnected Graph	0.020	0.001	0.108	0.007	0.001	0.475	0.000
Connected Graph	0.053	0.214	0.114	0.285	0.005	0.057	0.001
Linear	0.000	0.000	0.000	0.000	1.000	0.000	0.980
Multifurcation	0.033	0.001	0.041	0.717	0.552	0.758	0.051
Tree	0.021	0.809	0.918	0.086	0.000	0.993	0.000

Table 1.

MGPfact outperformed state-of-the-art methods in $F1_{branches}$.

P-values based one-sided paired t-tests suggest that the $F1_{branches}$ scores of MGPfact were significantly higher than those of the other methods for different trajectory types in the test set.

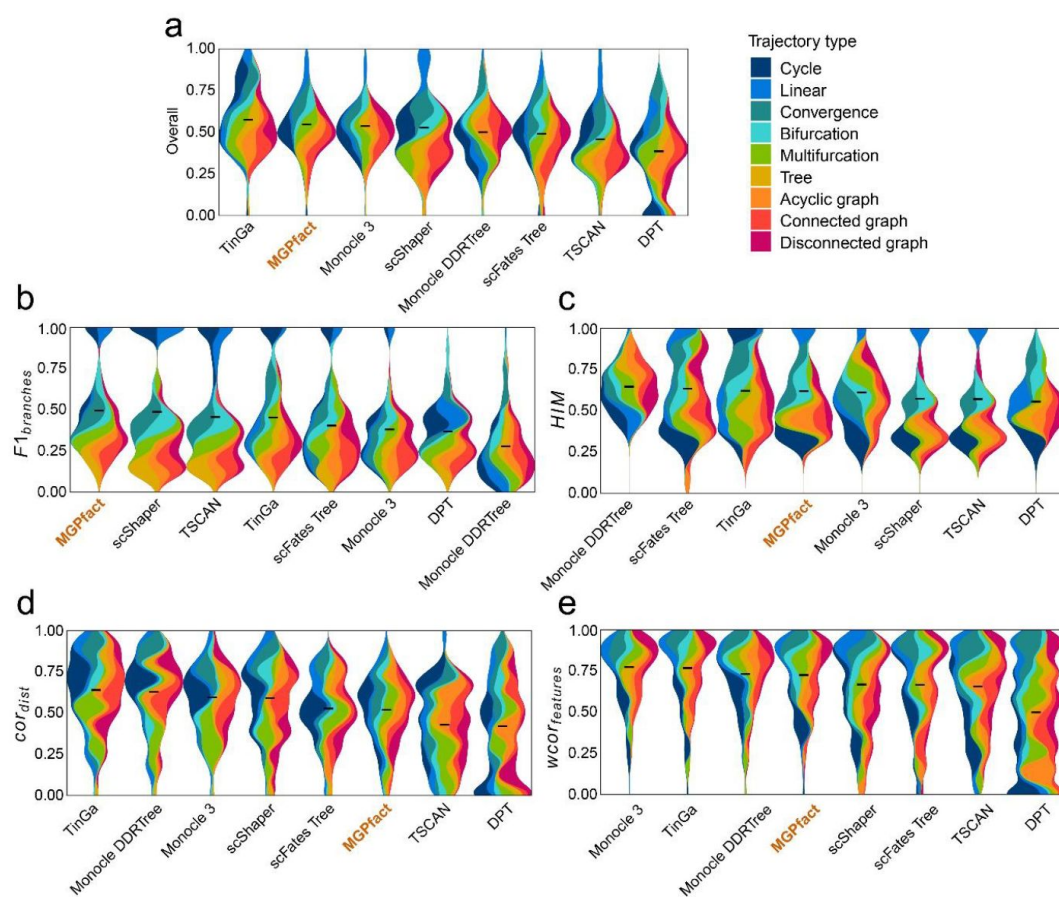


Fig 2.

Trajectory inference (TI) performance of state-of-the-art methods in 239 test datasets.

a. Overall scores; b. $F1_{branches}$; c. HIM ; d. cor_{dist} ; e. $wcor_{features}$. All results are color-coded based on the trajectory types, with the black line representing the mean value. The “Overall” assessment is calculated as the geometric mean of all four metrics.

Next, we compared the performance of MGPfact with the other algorithms in 9 different trajectory types, respectively, for predicting differentiated cell fate ($F1_{branches}$). As a result, MGPfact significantly outperformed more than half of the algorithms tested in the following trajectory types (T-test $P < 0.1$, **Table 1**): disconnected graph (N=5), linear (N=5), bifurcation (N=4), multifurcation (N=4) and tree (N=4). As for the other three metrics, MGPfact also showed advantages in HIM in linear (N=5), bifurcation (N=3), convergence (N=3); and in $wcor_{features}$ in multifurcation (N=6), bifurcation (N=5). Nevertheless, MGPfact showed limited predictive performance for cor_{dist} (Supplementary Table 2).

It is also worth noting that in 68 test datasets of real cell populations, MGPfact ranked the top of all 7 algorithms for “overall score” (**Fig. 3a**). As for the individual metric, MGPfact ranked to top for predicting trajectory topology ($HIM_{mean} = 0.721$, **Fig. 3c**); and the 2nd in trajectory branching ($F1_{branches_{mean}} = 0.600$, **Fig. 3b**) with no significant difference with that of the top predictor (scShaper, T-test, $P = 0.829$). These data show that MGPfact can effectively reconstruct the trajectory of cell fate and retrieve relevant biological processes. As for the other metrics in real test datasets, MGPfact was the 5th in the similarity of cell ($cor_{dist_{mean}} = 0.46$), and the 3rd in the similarity of gene significance ($wcor_{features_{mean}} = 0.735$). The differences between the metrics of MGPfact and those of the top performers were subtle (**Fig. 3d-e**, $\Delta cor_{dist_{mean}} = 0.07, \Delta wcor_{features_{mean}} = 0.05$).

We also analyzed three real-world datasets (Saelens et al., 2019), each representing a unique topology of trajectory: linear, single bifurcation, and multiple bifurcations. MGPfact excelled in capturing key developmental trajectory with branch points (Supplementary Fig. 5). In the linear trajectory, MGPfact accurately predicted the absence of bifurcations, aligning well with the ground truth ($overall = 0.871$). For the bifurcation trajectory, MGPfact successfully identified the main bifurcation ($overall = 0.636$). As for the multifurcation trajectory, MGPfact’s prediction is also close to the ground truth, as reflected by the overall score ($overall = 0.566$).

In summary, our data suggest that MGPfact is highly efficient in predicting cell fate in branching trajectory ($F1_{branches}$) and topological structure (HIM). These capabilities align with the primary objectives of the algorithm, namely, effective identification of the branching events in the development processes of cells. In addition, MGPfact performed better in real datasets, suggesting its robustness to noise from real experimental conditions. Nevertheless, trajectory inference by MGPfact is based on factorization of the covariance matrix, hence less performance in $wcor_{features}$ and cor_{dist} than methods based on full covariance matrix.

Comparative of Time Efficiency and Memory Consumption

We also compared the runtime and memory usage of different algorithms across 239 test datasets (Supplementary Table 3). MGPfact’s average maximum memory consumption ($memory_{mean}^{(max)} = 0.75GB$) is comparable to those of the other algorithms ($memory_{mean}^{(max)} \in [0.55, 0.91]GB$). As a trade-off for its advantages in feature-selection and factorization, MGPfact requires moderately longer execution time than the other algorithms ($time_{mean} = 3.42 min$).

MGPfact recovers the trajectory of early postnatal microglia development

The main advantage of MGPfact lies in the capability to factorize a complex cellular trajectory into bifurcation processes of selected co-expressed genes. To illustrate how MGPfact elucidates the biological process underlying cell fate determination, we applied MGPfact to a scRNA-seq data of microglia development and validated the results with experimental evidences (Li et al., 2019).

Table 2.

Comparison of the explanatory power for CD8⁺ T cell fate for MGPfact and three other different methods.

Adjusted R-squared values and P-values based on F-tests demonstrate the relative performance of MGPfact, Monocle 2, Monocle 3, and scFates Tree in fitting the experimentally characterized and annotated CD8⁺ T cell subtypes.

		MGPfact		Monocle 2		Monocle 3		scFates Tree	
		Adjust R ²	P	Adjust R ²	P	Adjust R ²	P	Adjust R ²	P
NSCLC (GSE99254)	CD8-LEF1	0.935	0.000	0.176	0.000	0.089	0.08	0.902	0.000
	CD8-CD28	0.195	0.002	0.170	0.000	0.108	0.06	0.006	0.145
	CD8-CX3CR1	0.634	0.000	0.259	0.000	0.629	0.000	0.882	0.000
	CD8-GZMK	0.259	0.000	0.189	0.000	0.855	0.000	0.547	0.000
	CD8-ZNF683	0.232	0.001	0.051	0.043	0.625	0.003	0.039	0.003
	CD8-LAYN	0.435	0.000	0.031	0.027	0.503	0.018	0.523	0.000
CRC (GSE108989)	CD8-LEF1	0.311	0.000	0.027	0.036	0.461	0.007	0.99	0.000
	CD8-GPR183	0.380	0.000	0.032	0.025	0.474	0.006	0.139	0.000
	CD8-CX3CR1	0.648	0.000	0.047	0.008	0.454	0.007	0.817	0.000
	CD8-GZMK	0.130	0.013	0.550	0.000	0.855	0.000	0.236	0.000
	CD8-CD6	0.277	0.000	0.109	0.007	0.45	0.008	0.054	0.000
	CD8-CD160	0.124	0.016	0.080	0.025	0.856	0.000	0.707	0.000
	CD8-LAYN	0.741	0.000	0.172	0.000	0.373	0.021	0.505	0.000

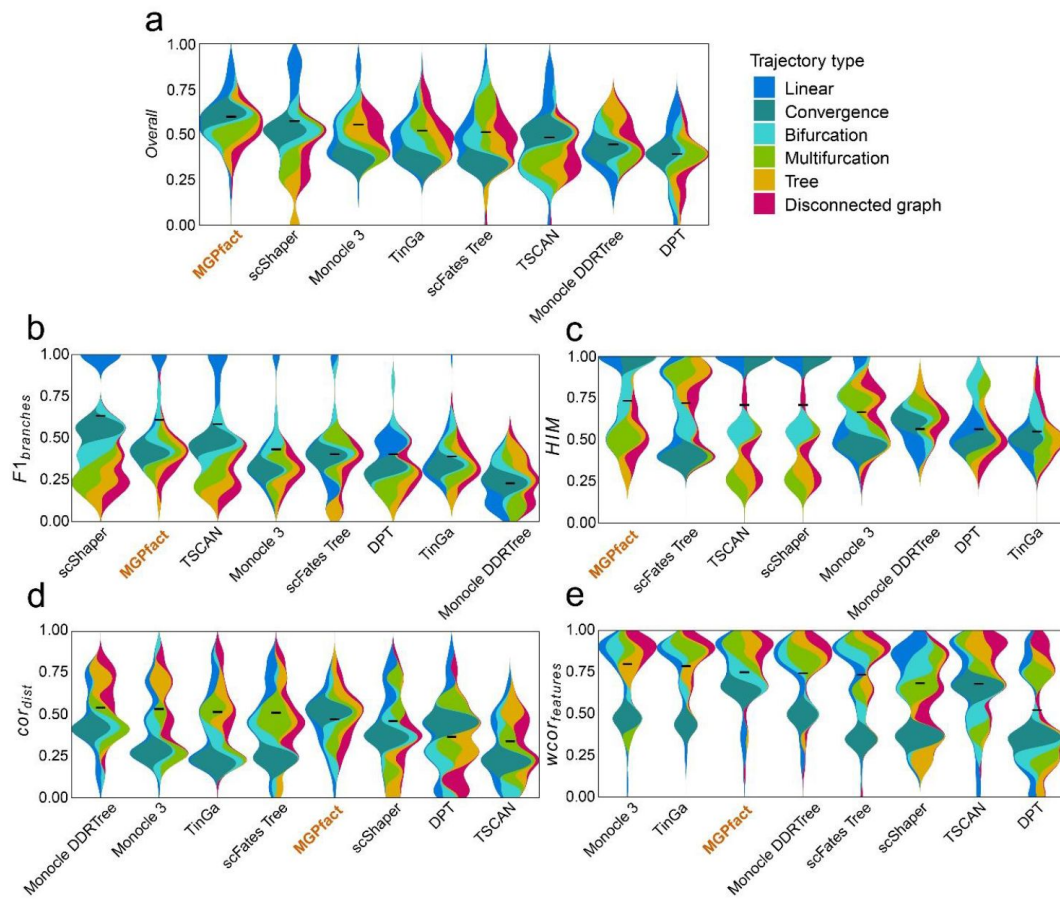


Fig 3.

Trajectory inference (TI) performance of state-of-the-art methods in 68 test datasets of real cell population.

a. Overall scores; b. $F1_{branches}$; c. HIM; d. cor_{dist} ; e. $wcor_{features}$. All results are color-coded based on the trajectory types, with the black line representing the mean value for ranking all methods. The "Overall" assessment is calculated as the geometric mean of all four metrics.

MGPfact recovers the determinants of the microglia development

Utilizing MGPfact, we reconstructed the developmental trajectories of microglia from immature microglia (IM at pseudotime 0) to homeostatic microglia (HM) and proliferative-region-associated microglia (PAM) (**Fig. 4a-c**, left panel). MGPfact identified three bifurcation processes (Supplementary Table 4), each corresponding to 74-90 highly weighted genes (HWG, absolute gene weight > 0.05) (**Fig. 4a-c**, right panel).

The first bifurcation determines the differentiated cell fates of PAM and HM, which involves a set of notable marker genes of both cell types, such as *Apoe*, *Selp1g* (HM), and *Gpnmb* (PAM). The second bifurcation determines the proliferative status, which is crucial for the development and function of PAM and HM (Guzmán, n.d.; Li et al., 2019). The genes affected by the second bifurcation are associated with cell cycle and proliferation, such as *Mki67*, *Tubb5*, *Top2a*. The third bifurcation influences the development and maturity of microglia, of which the highly weighted genes, such as *Tmem119*, *P2ry12*, and *Sepp1* are all previously annotated markers for establishment of the fates of microglia (Anderson et al., 2022; Li et al., 2019) (Supplementary Table 4).

Moreover, we retrieved highly active regulons from the HWG by MGPfact, of which the significance is quantified by the overall weights of the member genes. These data unveiled highly active transcription regulations in each bifurcation processes, which further traced back to the influential transcription factors as determinants of microglia development, such as *Hif1a*, *E2f5*, and *Nfkb1* (**Fig. 4d-e**, Methods). Specifically, *Hif1a* is crucial for microglial activation and directly linked to neurodegenerative disease progression (Wang et al., 2022). Our data showed an upregulation of *Hif1a* in the PAM-branch (phase 1) of the first bifurcation, reaffirming the role of *Hif1a* in PAM differentiation. The other two transcription factors, *E2f5* and *Nfkb1*, were active in phase 2 of the 2nd bifurcation and the 3rd bifurcation, respectively. Both are known for the roles in microglial development (Dresselhaus and Meffert, 2019; Nawal, n.d.).

Using consensus trajectories to delineate the development of microglial cells

We generated a consensus trajectory of microglial development from three independent bifurcation processes (Methods). Of note, the consensus trajectory revealed two distinct subtypes of proliferative-region-associated microglia (PAM), PAM-T1 (*Hif1a*+/*E2f5*-/*Nfkb1*+) and PAM-T2 (*Hif1a*+/*E2f5*-/*Nfkb1*-) (**Fig. 4f**). Particularly, the highly expressed genes in PAM-T2, including *Spp1*, *Gpnmb*, *Lgals1*, and *Cd63*, are previously identified in disease-associated microglia (DAM) (Li et al., 2019). Thus, our finding reaffirmed the connection between the two cell types, DAM and PAM, and suggested *Nfkb1* as a potential determinant of differentiation of PAM (**Fig. 4g**, Supplementary Table 5).

In conclusion, MGPfact reconstructed the cellular trajectory of microglial development, identified distinct cell types with marker genes and key regulators which are highly consistent to the experimental evidences (Dresselhaus and Meffert, 2019; Li et al., 2019; Nawal, n.d.; Wang et al., 2022).

MGPfact to decipher the evolution of tumor-associated CD8⁺ T cells

Next, we applied MGPfact to two populations of tumor-associated CD8⁺ T cells from non-small cell lung cancer (NSCLC) (Guo et al., 2018) and colorectal cancer (CRC) (Zhang et al., 2018), respectively. Using the same analytical pipeline as above, we identified a set of CD8⁺ T cell gene expression signatures (GES) from MGPfact-inferred trajectories, which are significantly predictive of clinical outcomes and immune treatment responses. Additionally, our data unveiled novel subtypes of tumor-associated CD8⁺ T cells with strong clinical implications.

MGPfact better explains the fate of tumor-associated CD8⁺ T cells

We assessed the goodness-of-fit (adjusted R-square) of the consensus trajectory derived by MGPfact and three methods (Monocle 2, Monocle 3 and scFates Tree) for the CD8⁺ T cell subtypes described in the original studies (Guo et al., 2018 [DOI](#); Zhang et al., 2018 [DOI](#)). The data showed that MGPfact significantly improved the explanatory power for most CD8⁺ T cell subtypes over Monocle 2, which was used in the original studies ($P < 0.05$, see **Table 2** [DOI](#) and Supplementary Table 6), except for the CD8-GZMK cells in the CRC dataset. Additionally, MGPfact demonstrated better explanatory power in specific cell types when compared to Monocle 3 and scFates Tree. For instance, in the NSCLC dataset, MGPfact exhibited higher explanatory power for CD8-LEF1 cells (**Table 2** [DOI](#), R-squared = 0.935), while Monocle 3 and scFates Tree perform better in other cell types.

MGPfact identifies T-cell gene expression signatures with clinical implications

MGPfact discerned the different cellular fates of tumor-associated CD8⁺ T cells by distinct bifurcation processes. To reveal the clinical implications of these bifurcation processes, we retrieved the mean expression vectors corresponding to either phase (branch) as Gene Expression Signatures (GES), and stratified cancer cohorts by quantifying the propensities to certain destiny of CD8⁺ T cells (Methods). Our data demonstrated pronounced disparities in the clinical outcomes associated with different fates of CD8⁺ T cells among patients (**Fig. 5a-b** [DOI](#), Supplementary Fig. 8e). In NSCLC, trajectory 1 is associated with cytotoxic T cells (96%, phase 1) and higher overall survival in TCGA-lung adenocarcinoma (LUAD) patients. Trajectory 2 is associated with exhausted T cells (91%, phase 2), and lower overall survival in the same cohort (**Fig. 5a** [DOI](#), Supplementary Fig. 6a-b). Similarly, in the CRC, trajectory 1 is associated with exhausted T cells (95%, phase 1) and poor overall survival in TCGA-COAD patients (**Fig. 5b** [DOI](#), Supplementary Fig. 8a-b).

In addition, for each trajectory identified in NSCLC and CRC, we selected a set of HWG (absolute gene weight > 0.05) to characterize the underlying biological processes and key transcription factors determining the bifurcation (Supplementary Table 7-8, Supplementary Fig. 6c-d, Supplementary Fig. 8c-d). In NSCLC, the HWG of trajectory 1 are primarily implicated in immune responses associated with antigen processing and presentation (Supplementary Table 9), while trajectory 2-HWG involve processes of immune cell migration (Supplementary Table 9). For CRC, trajectory 1 is enriched for genes of T cell activation and regulation (Supplementary Table 10).

Notably, our data showed that the weighted mean expression of the HWG of CD8⁺ T cell trajectories (Methods) are associated with responses to immune checkpoint inhibitors (ICIs) in multiple independent cohorts (**Fig. 5c-d** [DOI](#), Supplementary Fig. 8f). For instance, the weighted means of HWG of trajectories 1 and 2 in NSCLC which are associated with high activities of cytotoxic T cells, predicted better responses to anti-PD-130 and anti-CTLA-4 (Cho et al., 2020 [DOI](#); Liu et al., 2018 [DOI](#)), as well as their combination therapies (Auslander et al., 2018 [DOI](#)) ($AUC \in \{0.813, 0.964\}$, $P < 0.1$, Supplementary Fig. 7). Similarly, HWG pertaining to trajectory 1 in CRC is associated with high proportion of EMRA (87%, phase 1), hence better responses to immunotherapies in 4 cohorts (Auslander et al., 2018 [DOI](#); Cho et al., 2020 [DOI](#); Jung et al., 2019 [DOI](#); Liu et al., 2018 [DOI](#)) ($AUC \in \{0.796, 0.964\}$, $P < 0.01$, Supplementary Fig. 9).

Taken together, the factorization of scRNA-seq data by MGPfact provides highly relevant gene expression signatures of the fate of tumor associated CD8⁺ T cells, which advances the understanding of the evolution of tumor immune microenvironment (TIME) and predicts clinical outcomes.

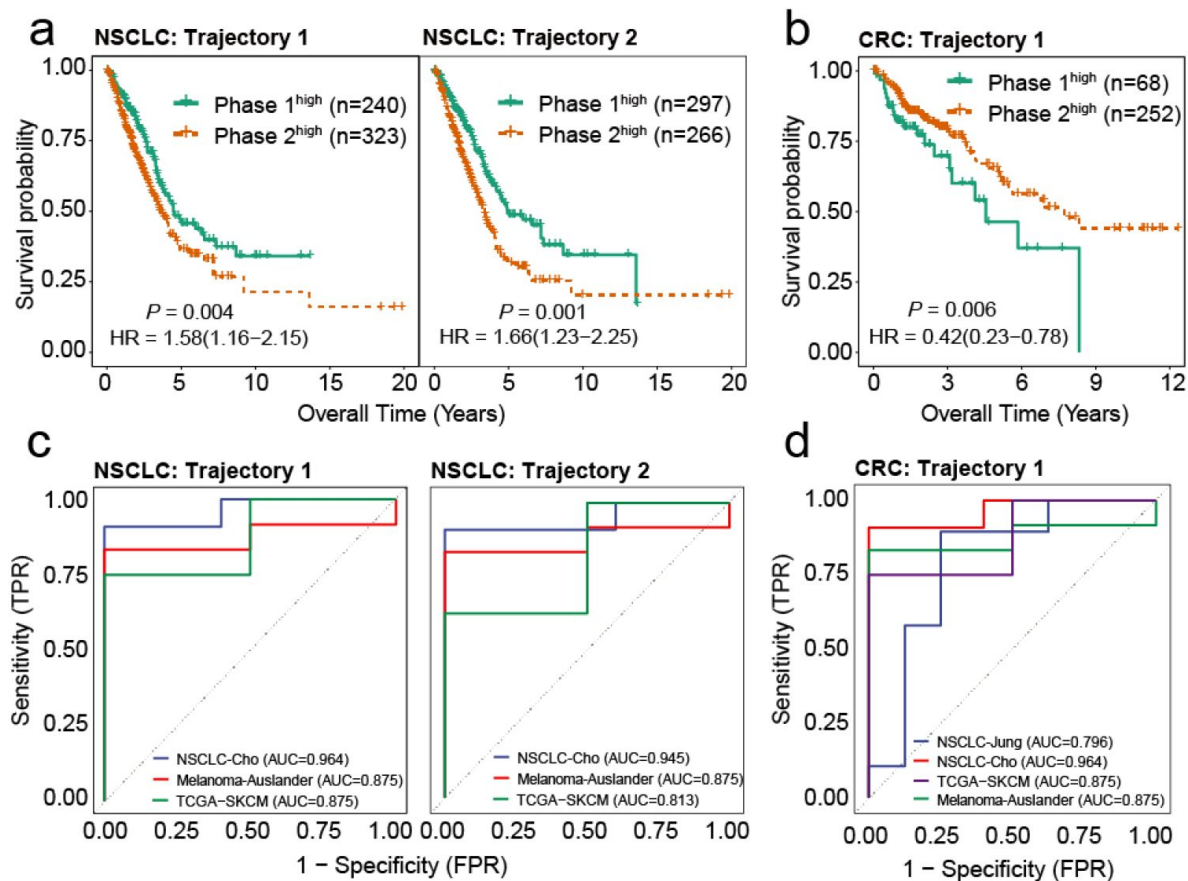


Fig 5.

Highly weighted genes (HWG) of the bifurcation processes of CD8⁺ T cells serve as reliable indicators for clinical outcome and ICI treatment response.

a. Gene expression signatures (GES) corresponding to HWG in CD8⁺ T cells trajectory 1 and 2 in NSCLC predict overall survival of the TCGA-LUAD cohort. **b.** Gene expression signatures (GES) corresponding to HWG in CD8⁺ T cells trajectory 1 in CRC predict overall survival of the TCGA-COAD cohort. **c.** ROC curve showing the weighted mean of HWG in Trajectories 1 and 2 in NSCLC significantly associated with ICI response across 3 independent studies. **d.** ROC curve showing the weighted mean of HWG in trajectories 1 and 2 in CRC significantly associated with ICI response across 4 independent studies.

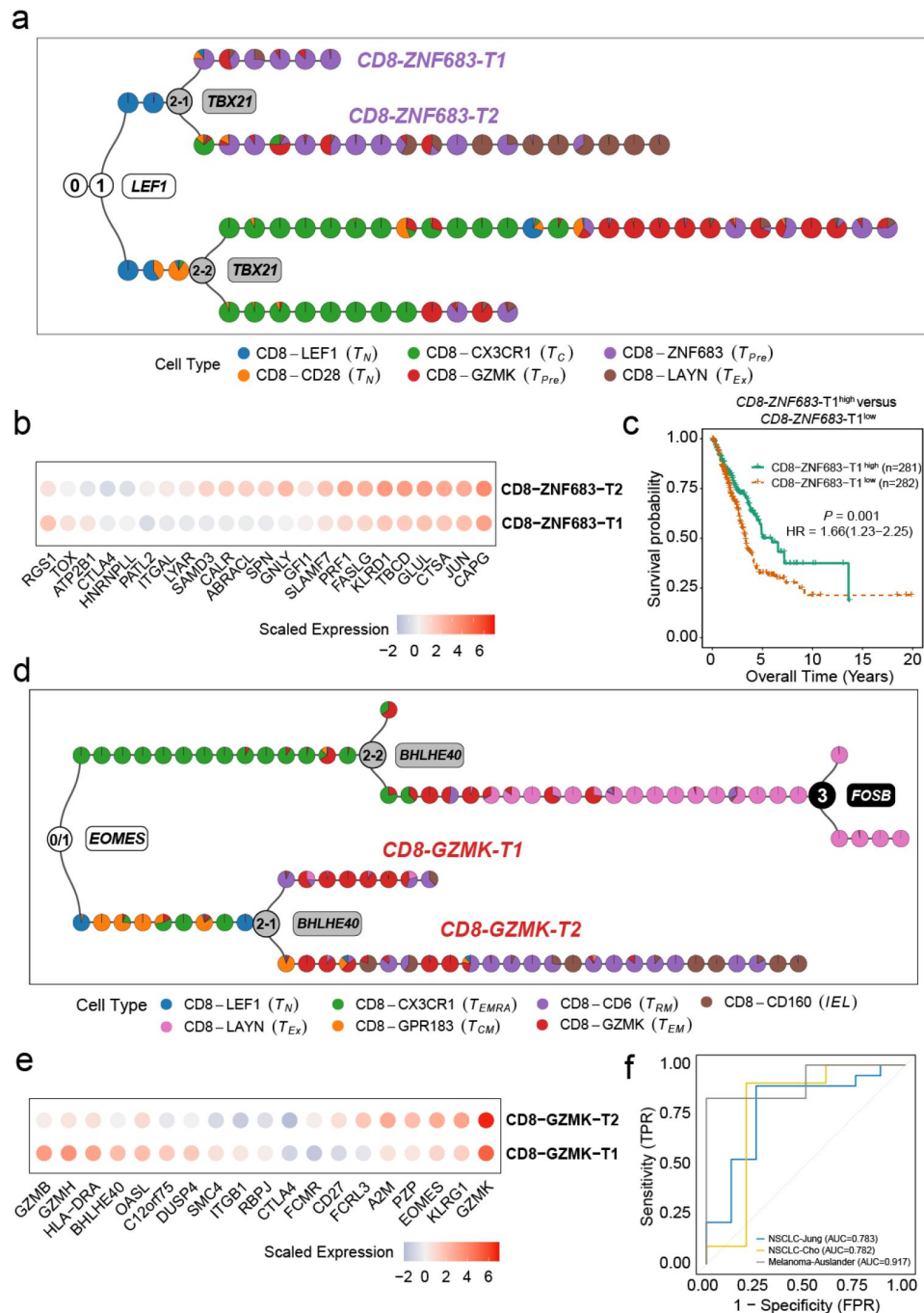


Fig 6.

MGPfact serves as an effective approach for characterization of new cellular subtypes.

a. The consensus trajectory of tumor-associated CD8⁺ T cells in NSCLC identified CD8-ZNF683-T1 and CD8-ZNF683-T2 as two subtypes of CD8-ZNF683, which are influenced by TBX21. **b.** Selected differentially expressed genes between CD8-ZNF683-T1 and CD8-ZNF683-T2 ($|\log_2FC| > 0.25$, adjusted P-value < 0.1). **c.** High expression of CD8-ZNF683-T1 signatures predicts good overall survival in the TCGA LUAD cohort (Methods). P-values were calculated through multivariate Cox regression analysis, and HR represents hazard ratio. **d.** The consensus trajectory of tumor-associated CD8⁺ T cells in CRC identified CD8-GZMK-T1 and CD8-GZMK-T2 as two subtypes of CD8-GZMK. **e.** Selected differentially expressed genes between CD8-GZMK-T1 and CD8-GZMK-T2 ($|\log_2FC| > 0.25$, adjusted P-value < 0.1). **f.** ROC curve showing high expression of CD8-GZMK-T1 signature associated with ICI treatment response in three independent studies. The consensus trajectory is formed by merging three bifurcation processes. Each colored circle represents a landmark (MURP), indicating the of cell type.

MGPfact identified new subtypes of CD8⁺ T cells with clinical implications

Furthermore, the consensus trajectories of tumor-associated CD8⁺ T cells inferred by MGPfact from NSCLC and CRC revealed new subtypes of lymphocytes. In NSCLC, we characterized CD8-ZNF683-T1 (LEF+/TBX21-) and CD8-ZNF683-T2 (LEF+/TBX21+) from CD8-ZNF683 (Fig. 6a, Supplementary Fig. 6c-d). The CD8-ZNF683-T2 cells highly expressed genes associated with “pre-exhausted” state, such as ITGAL, SAMD3, and SLAMF7 (Pritchard et al., 2023), many of which are target genes of TBX21. In contrast, CD8-ZNF683-T1 showed lower expression of these genes, hence repellency to the “pre-exhausted” state (Fig. 6b, Supplementary Table 11). In CRC, we identified two subtypes of effector memory T cells (CD8-GZMK), CD8-GZMK-T1 (EOMES-/BHLHE40+) and CD8-GZMK-T2 (EOMES-/BHLHE40-) (Fig. 6d, Supplementary Fig. 8c-d). CD8-GZMK-T2 strongly resembles CD8-GZMK, and potentially differentiating into CD8-CD6 (resident memory T cells, T_{RM}) and CD8-CD160 (intraepithelial lymphocytes, *IEL*); whereas CD8-GZMK-T1 cells demonstrated higher expression of GZMB, indicating an active cytotoxic cell-killing capability (Trapani, 2001). Simultaneously, these cells also marked by high expression levels of immune related genes, including OASL, RBPJ, and CTLA4, which are known targets of BHLHE40 (Lutter et al., 2022; Salmon et al., 2022), implying that BHLHE40 is a modulator of the higher effector activity in CD8-GZMK-T1 (Fig. 6e, Supplementary Table 12).

We further derived scores based on the differentially expressed genes of CD8-ZNF683-T1 and CD8-GZMK-T1 (Methods), as measures of the fraction of each subtype in cancer cohorts. In the LUAD cohort of TCGA, increased fraction of CD8-ZNF683-T1 in TIME was associated with favorable outcomes (Fig. 6c). And increased fractions of CD8-GZMK-T1 in TIME were associated with better responses to ICI therapy across 3 independent cohorts^{31–33} (Fig. 6f, Supplementary Fig. 10, $AUC \in \{0.782, 0.917\}$, $P < 0.1$), which were treated with anti-PD-1, anti-CTLA-4 and their combination therapies. In conclusion, our data showed the cellular trajectory inferred by MGPfact can be used to elucidate the complex evolutionary processes of tumor-associated CD8⁺ T cells, and further inform the characterization of new subtypes of T cells with significant clinical implications.

Discussion

Single-cell RNA sequencing (scRNA-seq) provides a direct, quantitative snapshot of a population of cells in certain biological conditions, thereby revealing the actual cell states and functions. Although existing clustering and embedding algorithms can effectively reveal discrete biological states of cells, these methods become less efficient when depicting continuous evolving of cells over the temporal domain. The introduction of manifold learning offers a new dimension for discovery of relevant biological knowledge in cell fate determination, allowing for a better representation of continuous changes in cells, especially in time-dependent processes such as development, differentiation, and clonal evolution. However, current manifold learning methods face major limitations, such as the need for prior information on pseudotime and cell clustering, and lack of explainability, which restricts their applicability. Additionally, many existing trajectory inference methods do not support gene selection, making it difficult to annotate the results to known biological entities, thereby hindering the interpretation of results and subsequent functional studies.

We developed MGPfact to overcome the limitations of the existing methods. Inspired by recent studies, MGPfact model the cell fate as mixture of Gaussian processes, which accommodate both continuous evolution pathway and biphasic destiny of cell fate. Thus, MGPfact is capable to distinguish discrete and continuous events in the same trajectory. In addition, by factorizing the mixture Gaussian processes, MGPfact offers the advantage to select genes corresponding to each

bifurcation process and thereby enable full biological annotation and interpretation of the trajectory. As a validation, we showed that gene-selection by MGPfact consistently recapitulated the development of microglia and tumor-associated CD8⁺ T cells; and recovered key regulators of distinct cell fate. So far, MGPfact is the only model-based manifold-learning framework which factorizes complex development trajectories into independent bifurcation processes of gene sets.

We conducted a comprehensive comparison of MGPfact with existing TI methods from various perspectives. This comparison included the correlation of cell sorting, accuracy of branch allocation, similarity of topological structures and differentially expressed features. It ought to be mentioned that number of principal components used should be determined by the intrinsic biological characteristics of the cell fate-determination. Our experiment based on a limited number of datasets may not represent more complex scenarios in other cell types. For the overall TI-performance, MGPfact demonstrated leading performance across 239 datasets, second only to TinGa. For the performance in branch allocation, which directly reflect the fitness to the outcomes of cell-fate, MGPfact outperformed its counterparts, especially in the topology groups of linear and bifurcation. As for $wcor_{features}$ and cor_{dist} , MGPfact performed less well mainly for two reasons. First, MGPfact is designed for bifurcation topology in the cellular trajectory, hence less efficient in inferring complex topologies. Then, MGPfact inference is based on selected gene sets instead of the whole transcriptome, the resulted trajectory correspond only to the bifurcation processes of interest hence does not necessarily reflect the whole topology of cellular trajectory. Furthermore, MGPfact performed significantly better in trajectory prediction in real cell population compared to the synthetic ones, suggesting the algorithm fit better to the true biological variation and noises.

To reconstruct the trajectory of cell fate, we merged all the bifurcation processes into a consensus trajectory. In the validation by microglia and tumor-associated CD8⁺ T cells, the consensus trajectory revealed highly consistent findings recovering known biology and the marker genes of specific cell states which further inform the transcription factor (TF) determining the fate of cell. In addition, the consensus trajectory revealed new subtypes of cells demonstrating highly relevant transcriptional characteristics. Particularly, we reported new subtypes of tumor associated CD8⁺ T cells characterized by different TBX21 and BHLHE40 activity, both of which are known regulators of CD8⁺ T cell functionality (Lutter et al., 2022 [DOI](#); Pritchard et al., 2023 [DOI](#); Salmon et al., 2022 [DOI](#); Trapani, 2001 [DOI](#)). These data suggest that MGPfact is capable to discover gene modules with strong and consistent transcriptional background hence better interpretability. Moreover, the results of MGPfact demonstrated strong clinical relevance. Using MGPfact we retrieved gene expression signatures (GES) which quantitatively measure the propensity and fraction of different fates of CD8⁺ T cell. These signatures correspond to important biological processes of T cell activity; and predict clinical outcome and ICI treatment responses from transcriptome data of bulk tumor biopsies, independent of any endogenous feature of the tumor cells.

Nevertheless, MGPfact also has some limitations, which shall be addressed in future study. Firstly, the complex definition of the bifurcation kernel introduces unfavorable singularity to the Gaussian kernel when considering highly complex trajectories. Additionally, the current trajectory inference by MGPfact is solely based on the temporal domain, neglecting bifurcation processes occurring in space. To overcome these limitations, future models should incorporate spatial dynamics of transcription and RNA velocity data to provide more comprehensive insights of cell fate. Moreover, the reconstruction of cellular trajectories by MGPfact implies independence of each bifurcation processes, which may not reflect real cellular behavior. Therefore, predictions from MGPfact should be interpreted with caution and validated experimentally.

Methods

Benchmarking MGPfact to state-of-the-art methods

We adopted a comprehensive evaluation framework from previous scRNA-seq study to assess the TI performance of MGPfact (Saelens et al., 2019 [DOI](#); Smolander et al., 2022 [DOI](#); Todorov et al., 2020 [DOI](#)). The validation dataset comprises 110 real data and 229 synthetic data, encompassing 9 different cellular trajectory topologies. The ground truth of cellular trajectories of each dataset were inferred and validated by the original study (Saelens et al., 2019 [DOI](#)). The synthetic datasets were generated using four simulators: dyngen (Saelens et al., 2019 [DOI](#)), dyntoy (Saelens et al., 2019 [DOI](#)), PROSSTT (Papadopoulos et al., 2019 [DOI](#)), and Splatter (Zappia et al., 2017 [DOI](#)), each modeling different trajectory topologies such as linear, branching, and cyclic. Splatter simulates branching events by setting expression states and transition probabilities, dyntoy generates random expression gradients to reflect dynamic changes, and dyngen focuses on complex branching structures within gene regulatory networks.

The evaluation of TI performance was based on five metrics.

- 1) The Hamming-Ipsen-Mikhailov (HIM) distance is a metric for assessing similarity between two topological structures. It integrates the normalized Hamming distance, which highlights differences in edge lengths, with the Ipsen-Mikhailov distance, which focuses on similarities in degree distributions. By linearly combining these two measures, the HIM distance offers a comprehensive evaluation of both local and global structural differences.
- 2) The $F1_{branches}$ is a metric used to evaluate a model's accuracy in branch allocation. It represents the harmonic mean of precision and recall, effectively capturing the performance of branch identification. In trajectory inference, $F1_{branches}$ are calculated by assessing the similarity between predicted and actual trajectory branches, emphasizing the Jaccard similarity of branch pairs.
- 3) This cor_{dist} metric measures similarity in intercellular distances between predicted and actual trajectories. It evaluates model accuracy in cell ordering by comparing relative positions of paired cells, highlighting changes in cell positions across states and reflecting cell differentiation dynamics.
- 4) The $wcor_{features}$ metric evaluates the similarity of key features, such as differentially expressed genes, between predicted and actual trajectories. Using weighted Pearson correlation, it highlights consistent features, reflecting the model's ability to capture biological variation. This helps identify crucial genes in trajectories and understand the molecular mechanisms of cell state transitions.
- 5) The *overall* score is calculated by taking the geometric mean of the four aforementioned metrics, providing an assessment of overall performance.

The dataset is divided into two groups: a training set and a testing set (Supplementary Fig. 1).

We use 100 training datasets to perform the following tasks:

- 1) Determine the optimal number of trajectories: With 3 set as the default for the number of factorized trajectories, we tested other values (1, 2, 4, and 5) and used paired T-tests to assess whether there are significant changes in MGPfact's prediction results under different parameter settings.
- 2) Verify the critical role of MURP: Randomly select 20, 40, 60, 80, and 100 cells for trajectory inference, map the inference results back to the original data using the KNN graph structure, and compare the prediction results with those obtained through MURP downsampling.

3) Robustness analysis of the consensus trajectory topology: Perform 60%, 70%, 80%, and 90% sampling on the original data, and then calculate the *HIM* similarity between the consensus trajectory predictions of MGPfact with and without sampling. A higher score indicates better robustness of the method.

Subsequently, in 239 test datasets, we compare MGPfact with 7 state-of-the-art TI methods using the aforementioned metrics, including Monocle DDRTree (Qiu et al., 2017b [DOI](#), 2017a [DOI](#)), TSCAN (Ji and Ji, 2016 [DOI](#)), and DPT (Haghverdi et al., 2016 [DOI](#)), as well as four new methods from recent studies: Monocle 3 (Cao et al., 2019 [DOI](#)), scShaper (Smolander et al., 2022 [DOI](#)), scFates Tree (Faure et al., 2023 [DOI](#)), and TinGa (Todorov et al., 2020 [DOI](#)).

The experimental comparisons were conducted on a CentOS system equipped with 48 CPU cores running at 2.2GHz and 250GB of memory. To ensure a uniform comparison, all experiments were performed using a single CPU core. For MGPfact, we tested each resulted trajectory and selected the one with the best “overall” score for comparison. For the other 7 methods, default settings are used unless otherwise specified.

Application of MGPfact in a Microglia Single-cell RNA-seq Dataset

We utilized the MGPfact reconstructed the developmental trajectory of microglia from a scRNA-seq dataset, including immature microglia (IM), proliferation-associated microglia (PAM), and homeostatic microglia (HM) (Li et al., 2019 [DOI](#)). In this analysis, we provided a detailed explanation of the analytical steps of MGPfact and the key results. Firstly, we identified three independent developmental pathways and pinpointed HWG associated with each bifurcation process. Then, we retrieved highly active regulons within each bifurcation process, tracing back to the potential influential determinants (transcription factors) in the development of microglia. Finally, we combined all the bifurcation processes into a consensus trajectory (Supplementary methods), which recovered the known biology of disease-related microglia (PAM), as represented by distinct cellular state and marker genes.

Predicting the evolutionary trajectory

of tumor-associated CD8⁺ T cells

We utilized MGPfact to conducted an exploratory analysis of the evolution of tumor-associated CD8⁺ T cells of non-small cell lung cancer (NSCLC) and colorectal cancer (CRC). We evaluated the goodness-of-fitness of the consensus trajectories from MGPfact to the CD8⁺ T cell subtypes identified in the original studies. For comparison, we used Monocle 2 (Qiu et al., 2017b [DOI](#), 2017a [DOI](#)) as a baseline model.

For the survival analysis, we extracted gene expression signatures (GES) from each independent bifurcation process to develop classifiers for evolutionary propensity of CD8⁺ T cells towards specific fates, based on which we stratified TCGA cancer cohorts and verified their association with clinical outcome.

To evaluated the association to ICI responses, we used HWG to retrieve key transcription factors related to each bifurcation process and characterized the underlying biological processes. We then assessed their connection to immunotherapy response using weighted means of the HWG.

Finally, we identified new subtypes based on distinct endpoints of the consensus trajectory and validated their association with clinical outcome and immunotherapy response using mean of the differently expressed gene (DEG, Supplementary Methods).

Single-cell sequencing data processing

We obtained the original mouse developmental microglia single-cell sequencing data from the GEO accession number GSE123025 (Li et al., 2019 [\[1\]](#)). Using Seurat (Butler et al., 2018 [\[2\]](#); Stuart et al., 2019 [\[3\]](#)), we replicated the processing steps described in the original study: 1) Normalization by dividing gene expression values by total RNA count, followed by log2 transformation; 2) Selection of highly variable genes (HVGs) using Seurat's mean.var.plot function, with controlled average expression [0.0125,3] and variance [0.5, ∞]; 3) Scaling and centering of the normalized matrix for HVGs, with regression of cell cycle effects. After preprocessing, we grouped cells into IM (P7-C0), PAM (P7-C1 and P7-GPNMB⁺CLEC7A⁺), and HM (P60), resulting in a 4,889-gene expression matrix across 1,009 cells.

For analyzing tumor-associated CD8⁺ T cells, we utilized scRNA-seq data from lung cancer (GSE99254) (Guo et al., 2018 [\[4\]](#)) and colorectal cancer (GSE108989) (Zhang et al., 2018 [\[5\]](#)) in the GEO database. We extracted preprocessed and centralized gene expression matrices of CD8⁺ T cells and analyzed them using the same genes and the same method (Monocle 2^{9,10}) as in the original papers or MGPfact for trajectory construction for a direct comparison. The NSCLC data yielded an 888-gene expression matrix across 3,700 cells, while the CRC data resulted in a 700-gene expression matrix across 3,177 cells.

Functional enrichment of highly weighted genes

For the highly weighted genes (HWG, absolute gene weight > 0.05) obtained from independent bifurcation processes, we utilized the R package clusterprofiler (Yu et al., 2012 [\[6\]](#)) to perform functional annotation using GO terms (Consortium, 2004 [\[7\]](#)), including biological process (BP), cellular component (CC), and molecular function (MF). The results with a Benjamini–Hochberg-adjusted P value less than 0.05 were retained.

Transcription factor program analysis

To comprehensively assess key regulatory factors within each independent trajectory, we performed SCENIC (Aibar et al., 2017 [\[8\]](#)) transcription factor regulatory program estimation for each analysis case. GENIE3 (Huynh-Thi et al., 2010 [\[9\]](#)) was used to identify co-expressed modules from the results of MURP (Ren et al., 2022 [\[10\]](#)) downsampling. RCisTarget (Aerts et al., 2010 [\[11\]](#); Aibar et al., 2017 [\[8\]](#)) was then used to identify regulons before AUCell (Aibar et al., 2017 [\[8\]](#)) was used to estimate the activity of each regulon. Each regulon comprises a specific transcription factor and its target genes. Finally, we utilize gene weights obtained from MGPfact analysis to evaluate the distinct impact of top regulons on each trajectory.

Generating the consensus trajectory

Following MGPfact decomposition, we obtained multiple independent bifurcative trajectories, each corresponds to a binary tree within the temporal domain. These trajectories were then merged to construct a coherent diffusion tree, representing the consensus trajectory of cells' fate. The merging process involves initially sorting all trajectories by their bifurcation time. The first (earliest) bifurcative trajectory is chosen as the initial framework, and subsequent trajectories are integrated to the initial framework iteratively by adding the corresponding branches at the bifurcation timepoints. As a result, the trajectories are ultimately merged into a comprehensive binary tree, serving as the consensus trajectory.

Assessing consistency of MGPfact-

derived CD8⁺ T Cell subtype trajectories

In the case study of CD8⁺ T cells, by combining independent trajectories, we derive a consensus trajectory representing the complex developmental pathway. To assess the goodness-of-fitness to the CD8⁺ T cell subtypes from the original study, we classified the trajectories into several states based on bifurcation points, each corresponding to a distinct stage of the evolutionary process. Then, we evaluated the interactive effects between the states and pseudotime on the fraction of the cell types using F-test (ANOVA). The resulted R-squared (R²), P-values, and F-statistics were used to evaluate the goodness-of-fitness of the models tested hence the explanatory power. For comparison, we used the Monocle 2 as the baseline model for trajectory inference. The differentiation trajectories of Monocle 2 were replicated following the workflow in the original study (Guo et al., 2018 [DOI](#); Zhang et al., 2018 [DOI](#)).

Survival analysis

We assessed the association of bifurcation processes and specific cell types with the clinical outcomes two cohorts of lung adenocarcinoma (LUAD, N=563) and colon adenocarcinoma (COAD, N=320) data from The Cancer Genome Atlas (TCGA). The gene expression and clinical data were downloaded from UCSC Xena platform (<http://xena.ucsc.edu/> [DOI](#)).

We assessed the survival impacts of NSCLC and CRC bifurcation processes in the TCGA LUAD and COAD cohorts, respectively. For each independent bifurcation process, we defined Gene Expression Signatures (GES) by the mean expression vectors of all trajectories in phase 1 and 2, respectively. Subsequently, we calculated the Pearson's correlation coefficients for each individual expression profile in TCGA LUAD or COAD cohorts to the phase 1 and 2 GES, where higher correlation correspond to stronger propensity to specific cell fate. This allowed us to classify patients into two groups: those exhibiting propensity to phase 1 and those exhibiting propensity to phase 2. To assess the survival impacts of specific cell states defined by the consensus trajectory, we developed a CD8-ZNF683-T1 score based on the signed average expression level of DEG associated with CD8-ZNF683-T1 (Supplementary methods). Subsequently, we classified the TCGA LUAD cohorts into two groups using the median of CD8-ZNF683-T1 scores, identifying those demonstrating a propensity towards CD8-ZNF683-T1 and those demonstrating a propensity towards CD8-ZNF683-T2.

To adjust for possible confounding effects, the relevant clinical features including age, sex and tumor stage were used as covariates. The Cox regression model was implemented using R-4.2 package "survival". And we generated Kaplan-Meier survival curves based on different classifiers to illustrate differences in survival time and report the statistical significance based on Log-rank test.

Immune-checkpoint inhibitor treatment response analysis

For the prediction of Immune-checkpoint inhibitor treatment response, we collected four datasets containing ICI treatment responses. These datasets consist of two non-small cell lung cancer related datasets (GSE135222 (Jung et al., 2019 [DOI](#)), n=27; GSE126044 (Cho et al., 2020 [DOI](#)), n=16) and two melanoma related datasets (GSE115821 (Auslander et al., 2018 [DOI](#)), n=14; TCGA-MENO (Liu et al., 2018 [DOI](#)), n=10). All data were processed with DESeq2 (Love et al., 2014 [DOI](#)) to fit gene dispersion to a negative binomial distribution, normalize raw counts, and stabilize variance, achieving standardization.

To validate the bifurcation processes identified by MGPfact predicting patients' response to immune-checkpoint inhibitor (ICI) treatments, we selected highly weighted genes (HWG) with absolute weights greater than 0.05 from each independent bifurcation process. Then, we

calculated the weighted mean expression of HWG in each ICI dataset to generate ROC curves for patient drug response.

To validate the specific cell states defined by the consensus trajectory predicting patients' response to ICI treatments, we used CD8-GZMK-T1 score based on the average expression level differences of upregulated and downregulated genes associated with CD8-GZMK-T1. Then, we calculated the CD8-GZMK-T1 score in each ICI dataset to generate ROC curves for patient drug response.

Additional information

Acknowledgments

We would like to thank Qingyun Li for providing the cell labels for the microglial dataset (Li et al., 2019). We would also like to express our gratitude to Nengming Xiao, and Guo Fu for their valuable input and constructive suggestions during the preparation of the manuscript.

Code availability

We have developed a comprehensive workflow for MGPfact. Firstly, a Docker container enables one-click program execution (details at: https://github.com/renjun0324/ti_mgpfact). Additionally, to fully harness MGPfact's capabilities, we have created the R package MGPfactR, accessible at: <https://github.com/renjun0324/MGPfactR>. Within this workflow, MCMC sampling for model parameter estimation is carried out using the Mamba library in the Julia environment. The related Julia package can be found here: <https://github.com/renjun0324/MGPfact.jl>. Additionally, other analysis scripts can be found on GitHub at https://github.com/renjun0324/mgpfact_paper. And the scFates Tree used in this paper is available as a performance comparison Docker container, constructed using the dendritic trajectory process in scFates (Faure et al., 2023), accessible at: https://github.com/renjun0324/ti_scFates_tree.

Data availability

The datasets used for performance comparison are archived on Zenodo by Saelens *et al.* (Saelens et al., 2019), with processed real and synthetic datasets available at <https://doi.org/10.5281/zenodo.1443566>. Data for specific cell instances of microglia and CD8⁺ T cells can be obtained from the GEO database with the following accession numbers: GSE123025 (Li et al., 2019), GSE99254 (Guo et al., 2018), and GSE108989 (Zhang et al., 2018). Expression matrices related to immune-checkpoint inhibitors (ICI) and clinical response information are downloadable from the CTR-DB (<https://ctrdb.cloudna.cn>).

Sources of Funding

This work was supported by the Fundamental Research Funds for the National Natural Science Foundation of China [82272944 to QL]; the National Natural Science Foundation of China [82203420 to JG].

Author Contributions

The project was conceived by Qiyuan Li. The model construction, data collection and analytical validation was done by Jun Ren, with the assistance of Yudi Hu and Jintao Guo. The manuscript was written by Jun Ren and Xuejing Lyu, with the assistance of Ying Zhou and Xiaodong Shi. All authors read and approved the final manuscript.

References

- Aerts S, Quan X-J, Claeys A, Naval Sanchez M, Tate P, Yan J, Hassan BA (2010) **Robust Target Gene Discovery through Transcriptome Perturbations and Genome-Wide Enhancer Predictions in *Drosophila* Uncovers a Regulatory Basis for Sensory Specification** *PLoS Biol* **8** <https://doi.org/10.1371/journal.pbio.1000435>
- Aibar S *et al.* (2017) **SCENIC: single-cell regulatory network inference and clustering** *Nat Methods* **14**:1083–1086 <https://doi.org/10.1038/nmeth.4463>
- Anderson SR, Roberts JM, Ghena N, Irvin EA, Schwakopf J, Cooperstein IB, Bosco A, Vetter ML (2022) **Neuronal apoptosis drives remodeling states of microglia and shifts in survival pathway dependence** *Elife* **11**
- Auslander N *et al.* (2018) **Robust prediction of response to immune checkpoint blockade therapy in metastatic melanoma** *Nat Med* **24**:1545–1549 <https://doi.org/10.1038/s41591-018-0157-9>
- Smith BJ (2014) **Mamba: Markov Chain Monte Carlo for Bayesian Analysis in julia**
- Becht E, McInnes L, Healy J, Dutertre C-A, Kwok IWH, Ng LG, Ginhoux F, Newell EW (2019) **Dimensionality reduction for visualizing single-cell data using UMAP** *Nat Biotechnol* **37**:38–44 <https://doi.org/10.1038/nbt.4314>
- Butler A, Hoffman P, Smibert P, Papalexi E, Satija R (2018) **Integrating single-cell transcriptomic data across different conditions, technologies, and species** *Nat Biotechnol* **36**:411–420 <https://doi.org/10.1038/nbt.4096>
- Cao J *et al.* (2019) **The single-cell transcriptional landscape of mammalian organogenesis** *Nature* **566**:496–502 <https://doi.org/10.1038/s41586-019-0969-x>
- Cho J-W, Hong MH, Ha S-J, Kim Y-J, Cho BC, Lee I, Kim HR (2020) **Genome-wide identification of differentially methylated promoters and enhancers associated with response to anti-PD-1 therapy in non-small cell lung cancer** *Exp Mol Med* **52**:1550–1563 <https://doi.org/10.1038/s12276-020-00493-8>
- Consortium GO (2004) **The Gene Ontology (GO) database and informatics resource** *Nucleic acids research* **32**:D258–D261
- Costa F, Grün D, Backofen R (2018) **GraphDDP: a graph-embedding approach to detect differentiation pathways in single-cell-data using prior class knowledge** *Nat Commun* **9** <https://doi.org/10.1038/s41467-018-05988-7>
- Dresselhaus EC, Meffert MK (2019) **Cellular Specificity of NF- κ B Function in the Nervous System** *Front Immunol* **10** <https://doi.org/10.3389/fimmu.2019.01043>
- Faure L, Soldatov R, Kharchenko PV, Adameyko I (2023) **scFates: a scalable python package for advanced pseudotime and bifurcation analysis from single-cell data** *Bioinformatics* **39** <https://doi.org/10.1093/bioinformatics/btac746>

- Fritzke B (1994) **A growing neural gas network learns topologies** *Advances in neural information processing systems* **7**
- Guo X *et al.* (2018) **Global characterization of T cells in non-small-cell lung cancer by single-cell sequencing** *Nat Med* **24**:978–985 <https://doi.org/10.1038/s41591-018-0045-3>
- Guzmán AU (2022) **Single-cell RNA sequencing of spinal cord microglia in a mouse model of neuropathic pain** McGill University
- Haghverdi L, Büttner F, Theis FJ (2015) **Diffusion maps for high-dimensional single-cell analysis of differentiation data** *Bioinformatics* **31**:2989–2998 <https://doi.org/10.1093/bioinformatics/btv325>
- Haghverdi L, Büttner M, Wolf FA, Büttner F, Theis FJ (2016) **Diffusion pseudotime robustly reconstructs lineage branching** *Nat Methods* **13**:845–848 <https://doi.org/10.1038/nmeth.3971>
- Hugo W *et al.* (2016) **Genomic and Transcriptomic Features of Response to Anti-PD-1 Therapy in Metastatic Melanoma** *Cell* **165**:35–44 <https://doi.org/10.1016/j.cell.2016.02.065>
- Huynh-Thu VA, Irrthum A, Wehenkel L, Geurts P (2010) **Inferring Regulatory Networks from Expression Data Using Tree-Based Methods** *PLoS ONE* **5** <https://doi.org/10.1371/journal.pone.0012776>
- Ji Z, Ji H (2016) **TSCAN: Pseudo-time reconstruction and evaluation in single-cell RNA-seq analysis** *Nucleic Acids Res* **44**:e117–e117 <https://doi.org/10.1093/nar/gkw430>
- Jung H, Kim HS, Kim JY, Sun J-M, Ahn JS, Ahn M-J, Park K, Esteller M, Lee S-H, Choi JK (2019) **DNA methylation loss promotes immune evasion of tumours with high mutation and copy number load** *Nat Commun* **10** <https://doi.org/10.1038/s41467-019-12159-9>
- Lange M *et al.* (2022) **CellRank for directed single-cell fate mapping** *Nat Methods* **19**:159–170 <https://doi.org/10.1038/s41592-021-01346-6>
- Li Q (2023) **scTour: a deep learning architecture for robust inference and accurate prediction of cellular dynamics** *Genome Biology*
- Li Q *et al.* (2019) **Developmental Heterogeneity of Microglia and Brain Myeloid Cells Revealed by Deep Single-Cell RNA Sequencing** *Neuron* **101**:207–223 <https://doi.org/10.1016/j.neuron.2018.12.006>
- Liu Jianfang *et al.* (2018) **An Integrated TCGA Pan-Cancer Clinical Data Resource to Drive High-Quality Survival Outcome Analytics** *Cell* **173**:400–416 <https://doi.org/10.1016/j.cell.2018.02.052>
- Love M, Anders S, Huber W (2014) **Differential analysis of count data—the DESeq2 package** *Genome Biol* **15**:10–1186
- Lutter L, Van Der Wal MM, Brand EC, Maschmeyer P, Vastert S, Mashreghi M, Van Loosdregt J, Van Wijk F (2022) **Human regulatory T cells locally differentiate and are functionally heterogeneous within the inflamed arthritic joint** *Clin & Trans Imm* **11** <https://doi.org/10.1002/cti2.1420>
- Nawal HS **A Systems Biology Perspective of Stem Cell Differentiation into Microglia**

Neal RM (2003) **Slice sampling** *The annals of statistics* **31**:705–767

Papadopoulos N, Gonzalo PR, Söding J (2019) **PROSSTT: probabilistic simulation of single-cell RNA-seq data for complex differentiation processes** *Bioinformatics* **35**:3517–3519 <https://doi.org/10.1093/bioinformatics/btz078>

Pritchard GH, Phan AT, Christian DA, Blain TJ, Fang Q, Johnson J, Roy NH, Shallberg L, Kedl RM, Hunter CA (2023) **Early T-bet promotes LFA1 upregulation required for CD8+ effector and memory T cell development** *Journal of Experimental Medicine* **220** <https://doi.org/10.1084/jem.20191287>

Qiu X, Hill A, Packer J, Lin D, Ma Y-A, Trapnell C (2017) **Single-cell mRNA quantification and differential analysis with Census** *Nat Methods* **14**:309–315 <https://doi.org/10.1038/nmeth.4150>

Qiu X, Mao Q, Tang Y, Wang L, Chawla R, Pliner HA, Trapnell C (2017) **Reversed graph embedding resolves complex single-cell trajectories** *Nat Methods* **14**:979–982 <https://doi.org/10.1038/nmeth.4402>

Ren J, Zhang Q, Zhou Y, Hu Y, Lyu X, Fang H, Yang J, Yu R, Shi X, Li Q (2022) **A downsampling method enables robust clustering and integration of single-cell transcriptome data** *Journal of Biomedical Informatics* **130** <https://doi.org/10.1016/j.jbi.2022.104093>

Roberts GO, Rosenthal JS (2009) **Examples of adaptive MCMC** *Journal of computational and graphical statistics* **18**:349–367

Saelens W, Cannoodt R, Todorov H, Saeys Y (2019) **A comparison of single-cell trajectory inference methods** *Nat Biotechnol* **37**:547–554 <https://doi.org/10.1038/s41587-019-0071-9>

Salmon AJ *et al.* (2022) **BHLHE40 Regulates the T-Cell Effector Function Required for Tumor Microenvironment Remodeling and Immune Checkpoint Therapy Efficacy** *Cancer Immunology Research* **10**:597–611 <https://doi.org/10.1158/2326-6066.CIR-21-0129>

Schulz E, Speekenbrink M, Krause A (2018) **A tutorial on Gaussian process regression: Modelling, exploring, and exploiting functions** *Journal of Mathematical Psychology* **85**:1–16 <https://doi.org/10.1016/j.jmp.2018.03.001>

Sha Y (2024) **Reconstructing growth and dynamic trajectories from single-cell transcriptomics data** *Nature Machine Intelligence* **6**:25–39

Smolander J, Junttila S, Venäläinen MS, Elo LL (2022) **scShaper: an ensemble method for fast and accurate linear trajectory inference from single-cell RNA-seq data** *Bioinformatics* **38**:1328–1335 <https://doi.org/10.1093/bioinformatics/btab831>

Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck WM, Hao Y, Stoeckius M, Smibert P, Satija R (2019) **Comprehensive Integration of Single-Cell Data** *Cell* **177**:1888–1902 <https://doi.org/10.1016/j.cell.2019.05.031>

Tierney L (1994) **Markov chains for exploring posterior distributions** *the Annals of Statistics* :1701–1728

Todorov H, Cannoodt R, Saelens W, Saeys Y (2020) **TinGa: fast and flexible trajectory inference with Growing Neural Gas** *Bioinformatics* **36**:i66–i74 <https://doi.org/10.1093/bioinformatics/btaa463>

Trapani JA (2001) **Granzymes: a family of lymphocyte granule serine proteases** *Genome Biol* **2** <https://doi.org/10.1186/gb-2001-2-12-reviews3014>

Van der Maaten L, Hinton G (2008) **Visualizing data using t-SNE** *Journal of machine learning research* **9**

Wang Q, Lu M, Zhu X, Gu X, Zhang T, Xia C, Yang L, Xu Y, Zhou M (2022) **The role of microglia immunometabolism in neurodegeneration: Focus on molecular determinants and metabolic intermediates of metabolic reprogramming** *Biomedicine & Pharmacotherapy* **153** <https://doi.org/10.1016/j.biopha.2022.113412>

Yu G, Wang L-G, Han Y, He Q-Y (2012) **clusterProfiler: an R package for comparing biological themes among gene clusters** *Omics: a journal of integrative biology* **16**:284–287

Zappia L, Phipson B, Oshlack A (2017) **Splatter: simulation of single-cell RNA sequencing data** *Genome Biol* **18** <https://doi.org/10.1186/s13059-017-1305-0>

Zhang L *et al.* (2018) **Lineage tracking reveals dynamic relationships of T cells in colorectal cancer** *Nature* **564**:268–272 <https://doi.org/10.1038/s41586-018-0694-x>

Author information

Jun Ren

School of Informatics, Xiamen University, Xiamen, China, National Institute for Data Science in Health and Medicine, School of Medicine, Xiamen University, Xiamen, China, Department of Hematology, The First Affiliated Hospital of Xiamen University and Institute of Hematology, School of Medicine, Xiamen University, Xiamen, China

Ying Zhou

National Institute for Data Science in Health and Medicine, School of Medicine, Xiamen University, Xiamen, China, Department of Hematology, The First Affiliated Hospital of Xiamen University and Institute of Hematology, School of Medicine, Xiamen University, Xiamen, China

Yudi Hu

National Institute for Data Science in Health and Medicine, School of Medicine, Xiamen University, Xiamen, China

Jing Yang

National Institute for Data Science in Health and Medicine, School of Medicine, Xiamen University, Xiamen, China

Hongkun Fang

National Institute for Data Science in Health and Medicine, School of Medicine, Xiamen University, Xiamen, China

Xuejing Lyu

National Institute for Data Science in Health and Medicine, School of Medicine, Xiamen University, Xiamen, China

Jintao Guo

National Institute for Data Science in Health and Medicine, School of Medicine, Xiamen University, Xiamen, China

Xiaodong Shi

School of Informatics, Xiamen University, Xiamen, China

Qiyuan Li

National Institute for Data Science in Health and Medicine, School of Medicine, Xiamen University, Xiamen, China, Department of Hematology, The First Affiliated Hospital of Xiamen University and Institute of Hematology, School of Medicine, Xiamen University, Xiamen, China

ORCID iD: [0000-0002-8934-8948](https://orcid.org/0000-0002-8934-8948)

For correspondence: qiyuan.li@xmu.edu.cn

Editors

Reviewing Editor

Mohammad Karimi

King's College London, London, United Kingdom

Senior Editor

Alan Moses

University of Toronto, Toronto, Canada

Reviewer #1 (Public review):

Summary:

Ren et al developed a novel computational method to investigate cell evolutionary trajectory for scRNA-seq samples. This method, MGPfact, estimates pseudotime and potential branches in the evolutionary path through explicitly modeling the bifurcations in a Gaussian process. They benchmarked this method using synthetic as well as real world samples and showed superior performance for some of the tasks in cell trajectory analysis. They further demonstrated the utilities of MGPfact using single cell RNA-seq samples derived from microglia or T cells and showed that it can accurately identify the differentiation timepoint and uncover biologically relevant gene signatures.

Strengths:

Overall I think this is a useful new tool that could deliver novel insights for the large body of scRNA-seq data generated in the public domain. The manuscript is written in a logical way and most parts of the method are well described.

Comments on revisions:

In this revision, the authors have sufficiently addressed all of my concerns. I don't have any follow-up comments.

<https://doi.org/10.7554/eLife.97424.2.sa2>

Reviewer #2 (Public review):

Summary of the manuscript:

Authors present MGPfactXMBD, a novel model-based manifold-learning framework designed to address the challenges of interpreting complex cellular state spaces from single-cell RNA sequences. To overcome current limitations, MGPfactXMBD factorizes complex development trajectories into independent bifurcation processes of gene sets, enabling trajectory inference based on relevant features. As a result, it is expected that the method provides a deeper understanding of the biological processes underlying cellular trajectories and their potential determinants.

MGPfactXMBD was tested across 239 datasets, and the method demonstrated similar to slightly superior performance in key quality-control metrics to state-of-the-art methods. When applied to case studies, MGPfactXMBD successfully identified critical pathways and cell types in microglia development, validating experimentally identified regulons and markers. Additionally, it uncovered evolutionary trajectories of tumor-associated CD8⁺ T cells, revealing new subtypes with gene expression signatures that predict responses to immune checkpoint inhibitors in independent cohorts.

Overall, MGPfactXMBD represents a relevant tool in manifold-learning for scRNA-seq data, enabling feature selection for specific biological processes and enhancing our understanding of the biological determinants of cell fate.

Summary of the outcome:

The novel method addresses core state-of-the-art questions in biology related to trajectory identification. The design and the case studies are of relevance.

Comments on revisions:

The authors have addressed all my previous comments to satisfaction.

<https://doi.org/10.7554/eLife.97424.2.sa1>

Author response:

The following is the authors' response to the original reviews.

Reviewer #1:

Comment#1: Ren et al developed a novel computational method to investigate cell evolutionary trajectory for scRNA-seq samples. This method, MGPfact, estimates pseudotime and potential branches in the evolutionary path by explicitly modeling the bifurcations in a Gaussian process. They benchmarked this method using synthetic as well as real-world samples and showed superior performance for some of the tasks in cell trajectory analysis. They further demonstrated the utilities of MGPfact using single-cell RNA-seq samples derived from microglia or T cells and showed that it can accurately identify the differentiation timepoint and uncover biologically relevant gene signatures. Overall I think this is a useful new tool that could deliver novel insights for the large body of scRNA-seq data generated in the public domain. The manuscript is written in a logical way and most parts of the method are well described.

Thank you for reviewing our manuscript and for your positive feedback on MGPfact. We are pleased that you find it useful for identifying differentiation timepoints and uncovering gene

signatures. We will continue to refine MGPfact and explore its applications across diverse datasets. Your insights are invaluable, and we appreciate your support.

Comment#2: Some parts of the methods are not clear. It should be outlined in detail how pseudo time T is updated in Methods. It is currently unclear either in the description or Algorithm 1.

Thanks to the reviewers' comments. We've added a description of how pseudotime T is obtained between lines 138 and 147 in the article. In brief, the pseudotime of MGPfact is inferred through Gaussian process regression on the downsampled single-cell transcriptomic data. Specifically, T is treated as a continuous variable representing the progression of cells through the differentiation process. We describe the relationship between pseudotime and expression data using the formula:

$$\mathbf{y}_i^* = f(\mathbf{T}) + \varepsilon \quad (1.)$$

Where $f(\mathbf{T})$ is a Gaussian Process (GP) with covariance matrix \mathbf{S} , and represents the error term. The Gaussian process is defined as:

$$f(\mathbf{T}) = \mathcal{GP}(\mathbf{0}, \mathbf{S} + \sigma_s^2 \cdot \mathbf{I}) \quad (2.)$$

Where σ_s^2 is the variance set to 1e-6.

During inference, we update the pseudotime by maximizing the posterior likelihood. Specifically, the posterior distribution of pseudotime T can be represented as:

$$p(\mathbf{T} | \mathbf{Y}^*) \propto p(\mathbf{Y}^* | f(\mathbf{T})) \cdot p(f(\mathbf{T})) \quad (3.)$$

Where $p(\mathbf{Y}^* | f(\mathbf{T}))$ is the likelihood function of the observed data \mathbf{Y}^* , and $p(f(\mathbf{T}))$ is the

prior distribution of the Gaussian process. This posterior distribution integrates the observed data with model priors, enabling inference of pseudotime and trajectory simultaneously. Due to the high autocorrelation of in the posterior distribution, we use Adaptive Metropolis within Gibbs (AMWG) sampling (Roberts and Rosenthal, 2009; Tierney, 1994). Other parameters are estimated using the more efficient SLICE sampling technique (Neal, 2003).

Comment#3: There should be a brief description in the main text of how synthetic data were generated, under what hypothesis, and specifically how bifurcation is embedded in the simulation.

Thank you for the reviewers' comments. We have added descriptions regarding the synthetic dataset in the methods section. The revised content is from line 487 to 493:

The synthetic datasets were generated using four simulators: dyngen (Saelens et al., 2019), dyntoy (Saelens et al., 2019), PROSSTT (Papadopoulos et al., 2019), and Splatter (Zappia et al., 2017), each modeling different trajectory topologies such as linear, branching, and cyclic. Splatter simulates branching events by setting expression states and transition probabilities,

dyntoy generates random expression gradients to reflect dynamic changes, and dyngen focuses on complex branching structures within gene regulatory networks.

Comment#4: Please explain what the abbreviations mean at their first occurrence.

We appreciate the reviewers' feedback. We have thoroughly reviewed the entire manuscript and made sure that all abbreviations have had their full forms provided upon their first occurrence.

Comment#5: In the benchmark analysis (Figures 2/3), it would be helpful to include a few trajectory plots of the real-world data to visualize the results and to evaluate the accuracy.

We appreciate the reviewer's feedback. To more clearly demonstrate the performance of MGPfact, we selected three representative cases from the dataset for visual comparison. These cases represent different types of trajectory structures: linear, bifurcation, and multifurcation. The revised content is between line 220 and 226.

As shown in Supplementary Fig. 5, it is evident that MGPfact excels in capturing main developmental paths and identifying key bifurcation points. In the linear trajectory structure, MGPfact accurately predicted the linear structure without bifurcation events, showing high consistency with the ground truth (*overall*=0.871). In the bifurcation trajectory structure, MGPfact accurately captured the main bifurcation event (). In the multifurcation trajectory structure, although MGPfact predicted only one bifurcation point, its overall structure remains close to the ground truth, as evidenced by its high overall score (*overall*=0.566). Overall, MGPfact demonstrates adaptability and accuracy in reconstructing various types of trajectory structures.

Comment#6: It is not clear how this method selects important genes/features at bifurcation. This should be elaborated on in the main text.

Thanks to the reviewers' comments. To enhance understanding, we've added detailed descriptions of gene selection in the main text and appendix, specifically from lines 150 to 161. In brief, MGPfact employs a Gaussian process mixture model to infer cell fate trajectories and identify independent branching events. We calculate load matrices using formulas 1 and 14 to assess each gene's contribution to the trajectories. Genes with an absolute weight greater than 0.05 are considered predominant in specific branching processes. Subsequently, SCENIC (Aibar et al., 2017; Bravo González-Blas et al., 2023) analysis was conducted to further infer the underlying regulons and annotate the biological processes of these genes.

Comment#7: It is not clear how survival analysis was performed in Figure 5. Specifically, were critical confounders, such as age, clinical stage, and tumor purity controlled?

To evaluate the predictive and prognostic impacts of the selected genes, we utilized the Cox multivariate regression model, where the effects of relevant covariates, including age, clinical stage, and tumor purity, were adjusted. We then conducted the Kaplan-Meier survival analysis again to ensure the reliability of the results. The revisions mainly include the following sections:

(1) We modified the description of adjusting for confounding factors in the survival analysis, from line 637 to 640:

To adjust for possible confounding effects, the relevant clinical features including age, sex and tumor stage were used as covariates. The Cox regression model was implemented using R-4.2 package “survival”. And we generated Kaplan-Meier survival curves based on different

classifiers to illustrate differences in survival time and report the statistical significance based on Log-rank test.

(2) We updated the images in the main text regarding the survival analysis, including Fig. 5a-b, Fig. 6c, and Supplementary Fig. 8e.

Comment#8: I recommend that the authors perform some sort of 'robustness' analysis for the consensus tree built from the bifurcation Gaussian process. For example, subsample 80% of the cells to see if the bifurcations are similar between each bootstrap.

We appreciate the reviewers' feedback. We performed a robustness analysis of the consensus tree using 100 training datasets. This involved sampling the original data at different proportions, and then calculating the topological similarity between the consensus trajectory predictions of MGPfact and those without sampling, using the Hamming-Ipsen-Mikhailov () metric. A higher score indicates greater robustness. The relevant figure is in Supplementary Fig. 4, and the description is in the main text from line 177 to 182.

The results indicate that the consensus trajectory predictions based on various sampling proportions of the original data maintain a high topological similarity with the unsampled results ($HIM_{mean}=0.686$). This demonstrates MGPfact's robustness and generalizability under different data conditions, hence the capability of capturing bifurcative processes in the cells' trajectory.

Reviewer #2:

Comment#1: The authors present MGPfact^{XMBD}, a novel model-based manifold-learning framework designed to address the challenges of interpreting complex cellular state spaces from single-cell RNA sequences. To overcome current limitations, MGPfact^{XMBD} factorizes complex development trajectories into independent bifurcation processes of gene sets, enabling trajectory inference based on relevant features. As a result, it is expected that the method provides a deeper understanding of the biological processes underlying cellular trajectories and their potential determinants. MGPfact^{XMBD} was tested across 239 datasets, and the method demonstrated similar to slightly superior performance in key quality-control metrics to state-of-the-art methods. When applied to case studies, MGPfact^{XMBD} successfully identified critical pathways and cell types in microglia development, validating experimentally identified regulons and markers. Additionally, it uncovered evolutionary trajectories of tumor-associated CD8+ T cells, revealing new subtypes with gene expression signatures that predict responses to immune checkpoint inhibitors in independent cohorts. Overall, MGPfact^{XMBD} represents a relevant tool in manifold learning for scRNA-seq data, enabling feature selection for specific biological processes and enhancing our understanding of the biological determinants of cell fate.

Thank you for your thoughtful review of our manuscript. We are thrilled to hear that you find MGPfact^{XMBD} beneficial for exploring cellular evolutionary paths in scRNA-seq data. Your insights are invaluable, and we look forward to incorporating them to further enrich our study. Thank you once again for your support and constructive feedback.

Comment#2: How the methods compare with existing Deep Learning based approaches such as TIGON is a question mark. If a comparison would be possible, it should be conducted; if not, it should be clarified why.

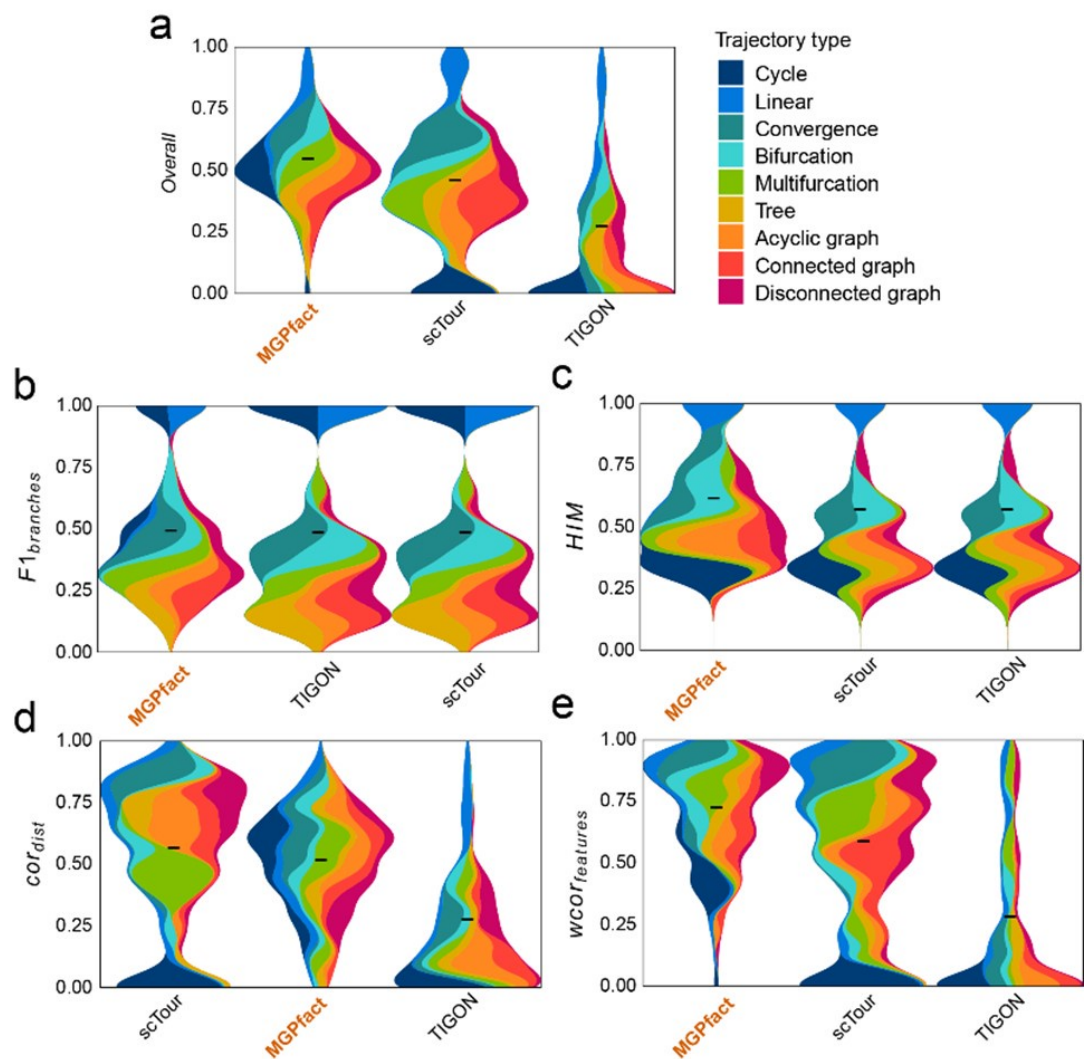
We appreciate the reviewer's comments. We have added a comparison with the sctour (Li, 2023) and TIGON methods (Sha, 2024).

It is important to note that the encapsulation and comparison of MGPfact are based on traditional differentiation trajectory construction. Saelens et al. established a systematic evaluation framework that categorizes differentiation trajectory structures into topological subtypes such as linear, bifurcation, multifurcation, graph, and tree, focusing on identifying branching structures in the cell differentiation process (Saelens et al., 2019). The scTour and TIGON methods mentioned by the reviewer are primarily used for estimating RNA velocity, focusing on continuous temporal evolution rather than explicit branching structures, and do not explicitly model branches. Therefore, we considered the predictions of these two methods as linear trajectories and compared them with MGPfact. While scTour explicitly estimates pseudotime, TIGON uses the concept of "growth," which is analogous to pseudotime, so we made the necessary adaptations.

Author response image 1 show that within this framework, compared to scTour ($overall_{mean}=0.448$) and TIGON ($overall_{mean}=0.263$), MGPfact still maintains a relatively high standard ($overall_{mean}=0.534$). This indicates that MGPfact has a significant advantage in accurately capturing branching structures in cell differentiation, especially in applications where explicit modeling of branches is required.

Author response image 1.

Comparison of MGPfact with scTour and TIGON in trajectory inference performance across 239 test datasets. a. Overall scores ; b. $F1_{branches}$; c. HIM ; e. $wcor_{features}$. All results are color-coded based on the trajectory types, with the black line representing the mean value. The "Overall" assessment is calculated as the geometric mean of all four metrics.



Comment#3: Missing Methods:

- The paper lacks a discussion of Deep Learning approaches for bifurcation analysis. e.g. scTour, Tigon.

- I am missing comments on methods such CellRank, and alternative approaches to delineate a trajectory.

We thank the reviewer for these comments.

(1) As mentioned in response to Comments#2, the scTour and TIGON methods are primarily used for estimating RNA velocity, focusing on continuous temporal evolution rather than explicit branching structures, and they do not explicitly model branches. We consider the predictions of these two methods as linear trajectories and compare them with MGPFact. The relevant description and discussion have been addressed in the response.

(2) We have added a description of RNA velocity estimation methods (scTour, TIGON, CellRank) in the introduction section. The revised content is from line 66 to 71:

Moreover, recent studies based on RNA velocity has provided insights into cell state transitions. These methods measure RNA synthesis and degradation rates based on the

abundance of spliced and unspliced mRNA, such as CellRank (Lange et al., 2022). Nevertheless, current RNA velocity analyses are still unable to resolve cell-fates with complex branching trajectory. Deep learning methods such as scTour (Li, 2023) and TIGON (Sha, 2024) circumvent some of these limitations, offering continuous state assumptions or requiring prior cell sampling information.

Comment#4: Impact of MURP:

The rationale for using MURP is well-founded, especially for trajectory definition. However, its impact on the final results needs evaluation.

How does the algorithm compare with a random subselection of cells or the entire cell set?

Thank you for the comments. We fully agree that MURP is crucial in trajectory prediction. As a downsampling method, MURP is specifically designed to address noise issues in single-cell data by dividing the data into several subsets, thereby maximizing noise reduction while preserving the main structure of biological variation (Ren et al., 2022). In MGPfact, MURP typically reduces the data to fewer than 100 downsampled points, preserving the core biological structure while lowering computational complexity. To assess MURP's impact, we conducted experiments by randomly selecting 20, 40, 60, 80, and 100 cells for trajectory inference. These results were mapped back to the original data using the KNN graph structure for final predictions, which were then compared with the MURP downsampling results. Supplementary results can be found in Supplementary Fig. 3, with additional descriptions in the main text from line 170 to 176.

The results indicate that trajectory inference using randomly sampled cells has significantly lower prediction accuracy compared to that using MURP. This is particularly evident in branch assignment ($F1_{branches}$) and correlation cor_{dist} , where the average levels decrease by 20.5%-64.9%. In contrast, trajectory predictions using MURP for downsampling show an overall score improvement of 5.31%-185%, further highlighting MURP's role in enhancing trajectory inference within MGPfact.

Comment#5: What is the impact of the number of components selected?

Thank you for the comments. In essence, MGPfact consists of two main steps: 1) trajectory inference; 2) calculation of factorized scores and identification of high-weight genes. After step 1, MGPfact estimates parameters such as pseudotime T and bifurcation points B . In step

2, we introduce a rotation matrix $R = \{r_1, r_2, \dots, r_L\}$ to obtain factor scores for each trajectory by rotating Y^* .

$$w_l = Y^* \cdot r_l + e_l^2 \quad (4.)$$

For all trajectories,

$$p(W|Y^*) = \prod_{l=1}^L [\mathcal{N}(Y^* \cdot r_l + e_l | 0, s_l) \cdot \mathcal{N}(e_l | 0, \sigma_{error}^2)] \quad (5.)$$

where is the error term for the i -th trajectory. The number of features in \mathbf{Y}^* must match the dimensions of the rotation matrix \mathbf{R} to ensure the factorized score matrix \mathbf{W} contains factor scores for trajectories, achieving effective feature representation and interpretation in the model.

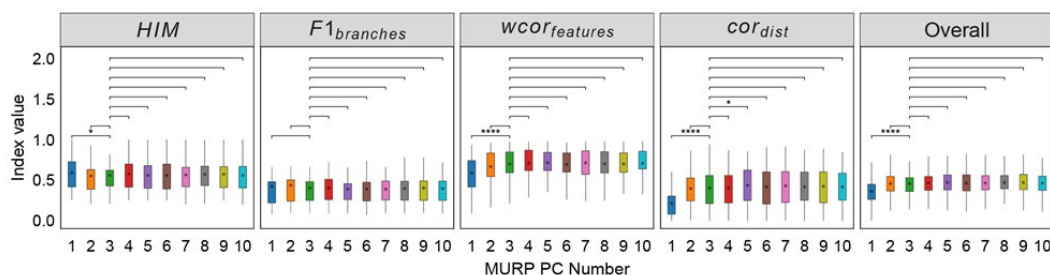
Additionally, to further illustrate the impact of the number of principal components (PCs) on model performance in step 1, we conducted additional experiments. We used 3 PCs as the default and adjusted the number to evaluate changes from this baseline. As shown in Author response image 2, setting the number of PCs to 1 significantly decreases the overall performance score ($overall_{mean}=0.363$), as well as the $wcor_{features}$

($wcor_{features_{mean}} = 0.586$) and cor_{dist} ($cor_{dist_{mean}} = 0.21$) metrics. In contrast, increasing

the number of PCs does not significantly affect the metrics. It ought to be mentioned that number of components used should be determined by the intrinsic biological characteristics of the cell fate-determination. Our experiment based on a limited number of datasets may not represent more complex scenarios in other cell types.

Author response image 2.

Robustness testing of the number of MURP PCA components on 100 training datasets. With the number of principal components (PCs) set to 3 by default; we tested the impact of different number of components (1-10) on the prediction results. In all box plots, the asterisk represents the mean value, while the whiskers extend to the farthest data points within 1.5 times the interquartile range. Significance is denoted as follows: not annotated indicates non-significant; * $P < 0.05$; ** $P < 0.01$; *** $P < 0.001$; two-sided paired Student's T-tests.



Comment#6: Please comment on the selection of the kernel functions (rbf and polynomial) and explain why other options were discarded.

Thank you for the comments. We have added a description regarding the selection of radial basis functions and polynomial kernels in lines 126-130. As the reviewers mentioned, the choice of kernel functions is crucial in the MGPfact analysis pipeline for constructing the covariance matrix of the Gaussian process. We selected the radial basis function (RBF) kernel and the polynomial kernel to balance capturing data complexity and computational efficiency. The RBF kernel is chosen for its ability to effectively model smooth functions and capture local variations in the data, making it well-suited to the continuous and smooth characteristics of biological processes; its hyperparameters offer modeling flexibility. The polynomial kernel is used to capture more complex nonlinear relationships between input features, with its hyperparameters also allowing further customization of the model. In contrast, other complex kernels, such as Matérn or spectral kernels, were omitted due to their interpretability challenges and the risk of overfitting with limited data. However, as suggested by the reviewers, we will consider and test the impact of other kernel functions on

the covariance matrix of the Gaussian process and their role in trajectory inference in our subsequent phases of algorithm design.

Comment#7: What is the impact of the Pseudotime method used initially? This section should be expanded with clear details on the techniques and parameters used in each analysis.

We are sorry for the confusion. We've added a description of how pseudotime is obtained between line 138 and 147 in the main text. And the specific hyperparameters involved in the model and their prior settings are detailed in the supplementary information.

In brief, the pseudotime and related topological parameters of the bifurcative trajectories in MGPfact are inferred by Gaussian process regression from downsampled single-cell transcriptomic data (MURP). Specifically, is treated as a continuous variable representing the progression of cells through the differentiation process. We describe the relationship between pseudotime and expression data as:

$$\mathbf{y}_i^* = f(\mathbf{T}) + \varepsilon \quad (6.)$$

where is a Gaussian Process (GP) with covariance matrix , and represents the error term. The Gaussian process is defined as:

$$f(\mathbf{T}) = \mathcal{GP}(\mathbf{0}, \mathbf{S} + \sigma_s^2 \cdot \mathbf{I}) \quad (7.)$$

where is the variance set to 1e-6. During inference, we update the pseudotime by maximizing the posterior likelihood. Specifically, the posterior distribution of pseudotime is obtained by combining the observed data \mathbf{Y}^* with the prior distribution of the Gaussian process model.

$$p(\mathbf{T}|\mathbf{Y}^*) \propto p(\mathbf{Y}^* | f(\mathbf{T})) \cdot p(f(\mathbf{T})) \quad (8.)$$

We use the Markov Chain Monte Carlo method for parameter estimation, particularly employing the adaptive Metropolis-within-Gibbs (AMWG) sampling to handle the high autocorrelation of pseudotime.

Comment#8: Enhancing Readability: For clarity, provide intuitive descriptions of each evaluation function used in simulated and real data. The novel methodology performs well for some metrics but less so for others. A clear understanding of these measurements is essential.

To address the concern of readability, we have added descriptions of 5 evaluation metrics in the methodology section (Benchmarking MGPfact to state-of-the-art methods) in line 494 to 515. Additionally, we have included a summary and discussion of these metrics in the conclusion section in line 214-240 to help the readers better understand the significance and impact of these measurements.

(1) In brief, the Hamming-Ipsen-Mikhailov () distance measures the similarity between topological structures, combining the normalized Hamming distance and the Ipsen-Mikhailov distance, which focus on edge length differences and degree distribution similarity, respectively. The is used to assess the accuracy of a model's branch assignment via Jaccard

similarity between branch pairs. In trajectory inference, cor_{dist} quantifies the similarity of inter-cell distances between predicted and true trajectories, evaluating the accuracy of cell ordering. The $wcor_{features}$ assesses the similarity of key features through weighted Pearson correlation, capturing biological variation. The score is calculated as the geometric mean of these metrics, providing an assessment of overall performance.

(2) For MGPfact and the other seven methods included in the comparison, each has its own focus. MGPfact specializes in factorizing complex cell trajectories using Gaussian process mixture models, making it particularly capable of identifying bifurcation events. Therefore, it excels in the accuracy of branch partitioning and similarity of trajectory topology. Among other methods, scShaper (Smolander et al., 2022) and TSCAN (Ji and Ji, 2016) are more suited for generating linear trajectories and excel in linear datasets, accurately predicting pseudotime. The Monocle series, as typical representatives of tree methods, effectively capture complex topologies and are suitable for analyzing cell data with diversified differentiation paths.

Comment#9: Microglia Analysis: In Figures 3A-C, the genes mentioned in the text for each bifurcation do not always match those shown in the panels. Please confirm this.

Thank you for pointing this out. We have carefully reviewed the article and corrected the error where the genes shown in the figures did not correspond to the descriptions in the article. The specific corrections have been made between line 257 and 264:

The first bifurcation determines the differentiated cell fates of PAM and HM, which involves a set of notable marker genes of both cell types, such as *Apoe*, *Selpg* (HM), and *Gpnmb* (PAM). The second bifurcation determines the proliferative status, which is crucial for the development and function of PAM and HM (Guzmán, n.d.; Li et al., 2019). The genes affected by the second bifurcation are associated with cell cycle and proliferation, such as *Mki67*, *Tubb5*, *Top2a*. The third bifurcation influences the development and maturity of microglia, of which the highly weighted genes, such as *Tmem119*, *P2ry12*, and *Sepp1* are all previously annotated markers for establishment of the fates of microglia (Anderson et al., 2022; Li et al., 2019) (Supplementary Table 4).

Comment#10: Regulons:

- The conclusions rely heavily on regulons. The Methods section describes using SCENIC, GENIE3, RCisTarget, and AUCell, but their relation to bifurcation analysis is unclear.

- Do you perform trajectory analysis on all MURP-derived cells or within each identified trajectory based on bifurcation? This point needs clarification to make the outcomes comprehensible. The legend of Figure 4 provides some ideas, but further clarity is required.

Thank you for the comments.

(1) To clarify, we used the tools like SCENIC to annotate the highly weighted genes (HWG) resulted from the bifurcation analysis for transcription factor regulation activity and possible impacts on biological processes. We have added descriptions to the analysis of our microglial data. The revised content is between line 265 and 266:

Moreover, we retrieved highly active regulons from the HWG by MGPfact, of which the significance is quantified by the overall weights of the member genes.

(2) We apologize for any confusion caused by our description. It is important to clarify that we performed an overall trajectory analysis on all MURP results, rather than analyzing within each identified trajectory. Specifically, we first used MURP to downsample all

preprocessed cells, where each MURP subset represents a group of cells. We then conducted trajectory inference on all MURP subsets and identified bifurcation points. This process generated multiple independent differentiation trajectories, encompassing all MURP subsets. To clearly convey this point, we have added descriptions in the legend of Figure 4. The revised content is between line 276 and 283:

“Fig. 4. MGPfact reconstructed the developmental trajectory of microglia, recovering known determinants of microglia fate. a-c. The inferred independent bifurcation processes with respect to the unique cell types (color-coded) of microglia development, where phase 0 corresponds to the state before bifurcation; and phases 1 and 2 correspond to the states post-bifurcation. Each colored dot represents a metacell of unique cell type defined by MURP. The most highly weighted regulons in each trajectory were labeled by the corresponding transcription factors (left panels). The HWG of each bifurcation process include a set of highly weighted genes (HWG), of which the expression levels differ significantly among phases 1, 2, and 3 (right panels).”

Comment#11: CD8+ T Cells: The comparison is made against Monocle2, the method used in the publication, but it would be beneficial to compare it with more recent methods. Otherwise, the added value of MGPfact is unclear.

Per your request, we have expanded our comparative analysis to include not only Monocle2 but also more recent methods such as Monocle3 (Cao et al., 2019) and scFates Tree (Faure et al., 2023). We used adjusted R-squared values to evaluate each method's ability to explain trajectory variation. The results have been added to Table 2 and Supplementary Table 6. The revised content is between line 318 and 326:

We assessed the goodness-of-fit (adjusted R-square) of the consensus trajectory derived by MGPfact and three methods (Monocle 2, Monocle 3 and scFates Tree) for the CD8+ T cell subtypes described in the original studies (Guo et al., 2018; Zhang et al., 2018). The data showed that MGPfact significantly improved the explanatory power for most CD8+ T cell subtypes over Monocle 2, which was used in the original studies ($P < 0.05$, see Table 2 and Supplementary Table 6), except for the CD8-GZMK cells in the CRC dataset. Additionally, MGPfact demonstrated better explanatory power in specific cell types when compared to Monocle 3 and scFates Tree. For instance, in the NSCLC dataset, MGPfact exhibited higher explanatory power for CD8-LEF1 cells (Table 2, R-squared = 0.935), while Monocle 3 and scFates Tree perform better in other cell types.

Comment#12: Consensus Trajectory: A panel explaining how the consensus trajectory is generated would be helpful. Include both visual and textual explanations tailored to the journal's audience.

Thank you for the comments. Regarding how the consensus trajectory is constructed, we have illustrated and described this in Figure 1 and the supplementary methods. Taking the reviewers' suggestions into account, we have added more details about the generation process of the consensus trajectory in the methods section to enhance the completeness of the manuscript. The revised content is from line 599 to 606:

Following MGPfact decomposition, we obtained multiple independent bifurcative trajectories, each corresponds to a binary tree within the temporal domain. These trajectories were then merged to construct a coherent diffusion tree, representing the consensus trajectory of cells' fate. The merging process involves initially sorting all trajectories by their bifurcation time. The first (earliest) bifurcative trajectory is chosen as the initial framework, and subsequent trajectories are integrated to the initial framework iteratively by adding the corresponding branches at the bifurcation timepoints. As a result, the trajectories are ultimately merged into a comprehensive binary tree, serving as the consensus trajectory.

Comment#13: Discussion:

- Check for typos, e.g., line 382 "pseudotime."
- Avoid considering HVG as the entire feature space.
- The first three paragraphs are too similar to the Introduction. Consider shortening them to succinctly state the scenario and the implications of your contribution.

Thank you for pointing out the typos.

(1) We conducted a comprehensive review of the document to ensure there are no typographical errors.

(2) We restructured the first three paragraphs of the discussion section to clarify the limitations in the use of current manifold-learning methods and removed any absolute language regarding treating HVGs as the entire feature space. The revised content is from line 419 to 430:

Single-cell RNA sequencing (scRNA-seq) provides a direct, quantitative snapshot of a population of cells in certain biological conditions, thereby revealing the actual cell states and functions. Although existing clustering and embedding algorithms can effectively reveal discrete biological states of cells, these methods become less efficient when depicting continuous evolving of cells over the temporal domain. The introduction of manifold learning offers a new dimension for discovery of relevant biological knowledge in cell fate determination, allowing for a better representation of continuous changes in cells, especially in time-dependent processes such as development, differentiation, and clonal evolution. However, current manifold learning methods face major limitations, such as the need for prior information on pseudotime and cell clustering, and lack of explainability, which restricts their applicability. Additionally, many existing trajectory inference methods do not support gene selection, making it difficult to annotate the results to known biological entities, thereby hindering the interpretation of results and subsequent functional studies.

Comment#14: Minor Comments:

- (1) Review the paragraph regarding the "current manifold-learning methods are faced with two major challenges." The message needs clarification.
- (2) Increase the quality of the figures.
- (3) Update the numbering of equations from #(.x) to (x).

We thank the reviewer for these detailed suggestions.

(1) We have thoroughly revised the discussion section, addressing overly absolute statements. The revised content is from line 426 to 428:

However, current manifold learning methods face major limitations, such as the need for prior information on pseudotime and cell clustering, and lack of explainability, which restricts their applicability.

(2) We conducted a comprehensive review of the figures in the article to more clearly present our results.

(3) We have meticulously reviewed the equations in the article to ensure there are no display issues with the indices.

Reference

- Aibar S, González-Blas CB, Moerman T, Huynh-Thu VA, Imrichova H, Hulselmans G, Rambow F, Marine J-C, Geurts P, Aerts J, van den Oord J, Atak ZK, Wouters J, Aerts S. 2017. SCENIC: single-cell regulatory network inference and clustering. *Nat Methods* 14:1083–1086. doi:10.1038/nmeth.4463
- Anderson SR, Roberts JM, Ghena N, Irvin EA, Schwakopf J, Cooperstein IB, Bosco A, Vetter ML. 2022. Neuronal apoptosis drives remodeling states of microglia and shifts in survival pathway dependence. *Elife* 11:e76564.
- Bravo González-Blas C, De Winter S, Hulselmans G, Hecker N, Matetovici I, Christiaens V, Poovathingal S, Wouters J, Aibar S, Aerts S. 2023. SCENIC+: single-cell multiomic inference of enhancers and gene regulatory networks. *Nat Methods*. doi:10.1038/s41592-023-01938-4
- Cao J, Spielmann M, Qiu X, Huang X, Ibrahim DM, Hill AJ, Zhang F, Mundlos S, Christiansen L, Steemers FJ, Trapnell C, Shendure J. 2019. The single-cell transcriptional landscape of mammalian organogenesis. *Nature* 566:496–502. doi:10.1038/s41586-019-0969-x
- Faure L, Soldatov R, Kharchenko PV, Adameyko I. 2023. scFates: a scalable python package for advanced pseudotime and bifurcation analysis from single-cell data. *Bioinformatics* 39:btac746. doi:10.1093/bioinformatics/btac746
- Guo X, Zhang Y, Zheng L, Zheng C, Song J, Zhang Q, Kang B, Liu Z, Jin L, Xing R, Gao R, Zhang L, Dong M, Hu X, Ren X, Kirchhoff D, Roider HG, Yan T, Zhang Z. 2018. Global characterization of T cells in non-small-cell lung cancer by single-cell sequencing. *Nat Med* 24:978–985. doi:10.1038/s41591-018-0045-3
- Guzmán AU. n.d. Single-cell RNA sequencing of spinal cord microglia in a mouse model of neuropathic pain.
- Ji Z, Ji H. 2016. TSCAN: Pseudo-time reconstruction and evaluation in single-cell RNA-seq analysis. *Nucleic Acids Res* 44:e117–e117. doi:10.1093/nar/gkw430
- Lange M, Bergen V, Klein M, Setty M, Reuter B, Bakhti M, Lickert H, Ansari M, Schniering J, Schiller HB, Pe'er D, Theis FJ. 2022. CellRank for directed single-cell fate mapping. *Nat Methods* 19:159–170. doi:10.1038/s41592-021-01346-6
- Li Q. 2023. scTour: a deep learning architecture for robust inference and accurate prediction of cellular dynamics. *Genome Biology*.
- Li Q, Cheng Z, Zhou L, Darmanis S, Neff NF, Okamoto J, Gulati G, Bennett ML, Sun LO, Clarke LE, Marschallinger J, Yu G, Quake SR, Wyss-Coray T, Barres BA. 2019. Developmental Heterogeneity of Microglia and Brain Myeloid Cells Revealed by Deep Single-Cell RNA Sequencing. *Neuron* 101:207–223.e10. doi:10.1016/j.neuron.2018.12.006
- Neal RM. 2003. Slice sampling. *The annals of statistics* 31:705–767.
- Papadopoulos N, Gonzalo PR, Söding J. 2019. PROSSTT: probabilistic simulation of single-cell RNA-seq data for complex differentiation processes. *Bioinformatics* 35:3517–3519. doi:10.1093/bioinformatics/btz078
- Ren J, Zhang Q, Zhou Y, Hu Y, Lyu X, Fang H, Yang J, Yu R, Shi X, Li Q. 2022. A downsampling method enables robust clustering and integration of single-cell transcriptome data. *Journal of Biomedical Informatics* 130:104093. doi:10.1016/j.jbi.2022.104093
- Roberts GO, Rosenthal JS. 2009. Examples of adaptive MCMC. *Journal of computational and graphical statistics* 18:349–367.

- Saelens W, Cannoodt R, Todorov H, Saeys Y. 2019. A comparison of single-cell trajectory inference methods. *Nat Biotechnol* 37:547–554. doi:10.1038/s41587-019-0071-9
- Sha Y. 2024. Reconstructing growth and dynamic trajectories from single-cell transcriptomics data 6.
- Smolander J, Junttila S, Venäläinen MS, Elo LL. 2022. scShaper: an ensemble method for fast and accurate linear trajectory inference from single-cell RNA-seq data. *Bioinformatics* 38:1328–1335. doi:10.1093/bioinformatics/btab831
- Tierney L. 1994. Markov chains for exploring posterior distributions. *the Annals of Statistics* 1701–1728.
- Zappia L, Phipson B, Oshlack A. 2017. Splatter: simulation of single-cell RNA sequencing data. *Genome Biol* 18:174. doi:10.1186/s13059-017-1305-0
- Zhang L, Yu X, Zheng L, Zhang Y, Li Y, Fang Q, Gao R, Kang B, Zhang Q, Huang JY, Konno H, Guo X, Ye Y, Gao S, Wang S, Hu X, Ren X, Shen Z, Ouyang W, Zhang Z. 2018. Lineage tracking reveals dynamic relationships of T cells in colorectal cancer. *Nature* 564:268–272. doi:10.1038/s41586-018-0694-x

<https://doi.org/10.7554/eLife.97424.2.sa0>