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

Jun Ren[1,2,3], Ying Zhou[2,3], Yudi Hu[2], Jing Yang[2], Hongkun Fang[2], Xuejing Lyu[2], Jintao Guo[2], Xiaodong Shi[1], Qiyuan Li[2,3*]

1 School of Informatics, Xiamen University, Xiamen, 361105, China.

2 National Institute for Data Science in Health and Medicine, School of Medicine, Xiamen University, Xiamen, 361102, China.

3 Department of Hematology, The First Affiliated Hospital of Xiamen University and Institute of Hematology, School of Medicine, Xiamen University, Xiamen, 361102, China.

*Corresponding authors

Qiyuan Li, National Institute for Data Science in Health and Medicine, School of Medicine, Xiamen University, Xiamen, 361102, China

Fax: 0592-2185175; Tel: 0592-2185175; Email: qiyuan.li@xmu.edu.cn

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.

Key Words: scRNA-seq, factorization, manifold-learning, bifurcation process, mixtures of Gaussian processes

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.

Despite the recent advancements in reconstructing single-cell trajectory, the current manifold-learning methods are faced with two major challenges. As the inferred graph-based embedding is the result of multiple complex biological processes, the effect of specific process cannot be directly distinguished from each other, which hinders the interpretability of model and discovery of novel, relevant knowledge in biology. Then, the inferred graph embedding is highly dependent on the biases in the feature-selection, whereas conventional linear feature-selection methods are less efficient for the learning of complex topologies, hence adds to the difficulty of suggesting candidate genes for downstream functional study.

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

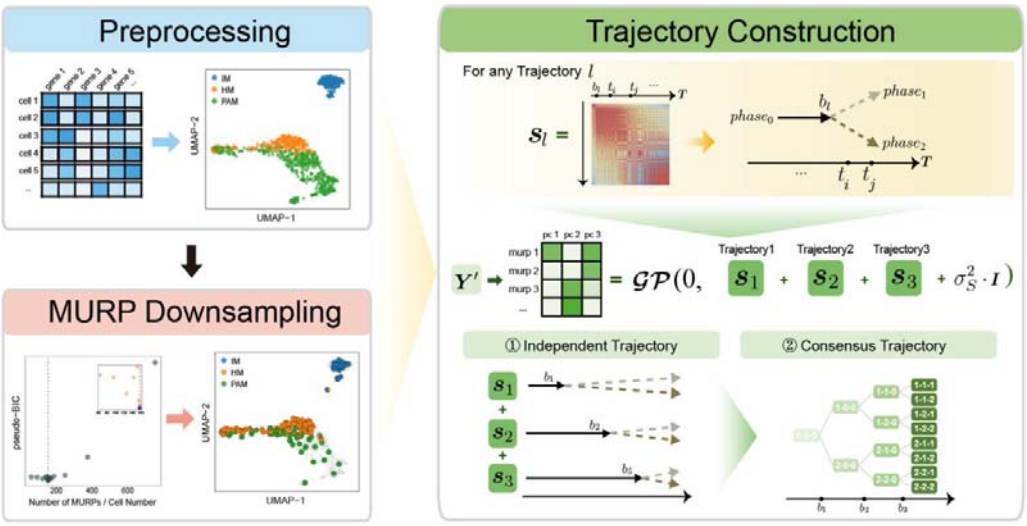

Fig 1. Overview of MGPfact workflow. The complete workflow comprises two major stages: MURP downsampling with preprocessed data and trajectory reconstruction. In the stage of trajectory reconstruction, the scRNA-seq data were first factorized into independent bifurcation processes based on mixtures of Gaussian processes, which were then merged into a consensus trajectory.

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 $\boldsymbol{Y}$ to yield a M-by-N expression matrix $\boldsymbol{Y}'$ 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 $\boldsymbol{Y}^* = \{\boldsymbol{y}_1^*, \boldsymbol{y}_2^*, \boldsymbol{y}_3^*, \dots \boldsymbol{y}_L^*\}$ (M-by-L),

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

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

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

$$\boldsymbol{y}_l^* = f(\boldsymbol{T}) + \varepsilon \#(2.)$$

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

$$f(\boldsymbol{T}) = \mathcal{GP}(0, \boldsymbol{S} + \sigma_S^2 \cdot \boldsymbol{I}) \#(3.)$$

And for all features:

$$p(\boldsymbol{Y}^*, f(\boldsymbol{T})) = p(\boldsymbol{Y}^* | f(\boldsymbol{T})) \cdot p(\boldsymbol{f}(\boldsymbol{T})) \#(4.)$$

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

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

Specifically, we consider $\boldsymbol{S}$ as a mixture of L independent bifurcation Gaussian processes(Schulz et al., 2018),

$$\boldsymbol{S} = \sum_{l=1}^{L} \boldsymbol{s}_l \#(6.)$$

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 $\boldsymbol{c_l}$, and the necessary hyperparameters. The branching labels $\boldsymbol{c_l} \in \{0,1,2\}$, correspond to different phases and states of cell fate, where $\boldsymbol{c_l} = \mathbf{0}$ corresponds the phase before branching, and $\boldsymbol{c_l} \in \{1,2\}$. corresponds to the two cellular states of the bifurcation process, respectively. For any landmark (MURP) x,

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

The covariance matrix $\boldsymbol{s_l}$ for trajectory l can be expressed as follows,

$$[s_l]_{x,y} = \mathcal{K}(t_x, t_y) \#(8.)$$

where $\mathcal{K}$ is a kernel function. We employ radial basis function (rbf) and polynomial kernel function (pl).

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

$$k_{pl}(t_x, t_y) = \left(\lambda_{pl} {t_x}^T t_y + c_{pl}\right)^{d_{pl}} \#(10.)$$

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

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

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

$$p\left(\boldsymbol{Y}^{*}\middle|f(\boldsymbol{T})\right)=\prod_{l=1}^{L}\mathcal{N}\left(\boldsymbol{y}_{l}^{*}\middle|0,\sum_{l=1}^{L}\boldsymbol{s}_{l}^{*}+\sigma_{S}^{2}\right)\#(12.)$$

We infer parameters by maximizing the posterior likelihood using Markov Chain Monte Carlo (MCMC) methods available in Mamba(B J, 2014). $\boldsymbol{T}$ is estimated using the Adaptive Metropolis within Gibbs (AMWG) sampling(Roberts and Rosenthal, 2009; Tierney, 1994), as the $\boldsymbol{T}$ exhibits high autocorrelation in the posterior distribution. Other parameters are estimated using the more efficient SLICE sampling technique(Neal, 2003).

Algorithm 1 MGPfact: infer cell fate trajectory

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

```
initialize all parameters in θ
covariance matrix S = 0
object function L = 0
Y' ← MURP downsampling Y_{p,n};
Y* ← PCA analysis
θ ← Opmized ObjectF
Graph ← Greate independent trajectory using θ
function ObjectF(Y*, T, b_l, c_l)
    Q ← dim(Y*)[2]
    for q = 1; q < Q; q++ do                          ▷ object function
        v ← [Y*]_q
        for l = 1; l < L; l++ do
            s_l ← Cov(T, b_l, c_l))
            S ← S + s_l
        end for
        L ← L + MultivariateNormalPDF(v, 0, S)
    end for
    return L
end function
function Cov(T, b_l, c_l)                             ▷ construct covariance matrix
    P ← length(T)
    for i = 1; i < P; i++ do
        for j = 1; j < P; j++ do
            s_l[i, j] ← Kernel(t_i, t_j, b_l, c_l)
        end for
    end for
    return s_l
end function
function Kernel(t_i, t_j, b_l, c_{l,i}, c_{l,j})      ▷ kernel function
    if t_i, t_j < b_l then
        return k_rbf(t_i, t_j) + k_pl(t_i, t_j)
    else if t_i, t_j > b_l ∧ c_{l,i} = c_{l,j} then
        return k_rbf(t_i, t_j) + k_pl(t_i − b_l, t_j − b_l)
    else if c_{l,i} ≠ c_{l,j} then
        return k_rbf(t_i,b_l)·k_rbf(b_l,t_j) / k_rbf(b_l,b_l)
    end if
end function
```

Then, we introduce a rotation matrix $R=\{r_1,r_2,\ldots,r_L\}$ to obtain factor score $\boldsymbol{w_l}$ for each trajectory l by rotating $\mathbf{Y}^{*}$.

$$\mathbf{w}_{l}=\mathbf{Y}^{*}\cdot\mathbf{r}_{l}+e_{l}^{2}\#(13.)$$

For all trajectories,

$$p(\boldsymbol{W}|\boldsymbol{Y}^*) = \prod_{l=1}^{L} [\mathcal{N}(\boldsymbol{Y}^* \cdot \boldsymbol{r}_l + e_l|0, \boldsymbol{s}_l) \cdot \mathcal{N}(e_l|0, \sigma_{error}^2)] \#(14.)$$

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

Specifically, the loading matrix for each gene onto the l-th trajectory can be expressed using equations (1) and (13) as follows,

$$\mathbf{w}_l = [\boldsymbol{Y}' \cdot \mathbf{v}_l] \cdot \mathbf{r}_l + e_l^2 = \mathbf{Y}' \cdot \mathbf{u}_l + e_l^2 \#(15.)$$

which enables factorization and 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

MGPfact predicts cellular trajectories

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). 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).

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).

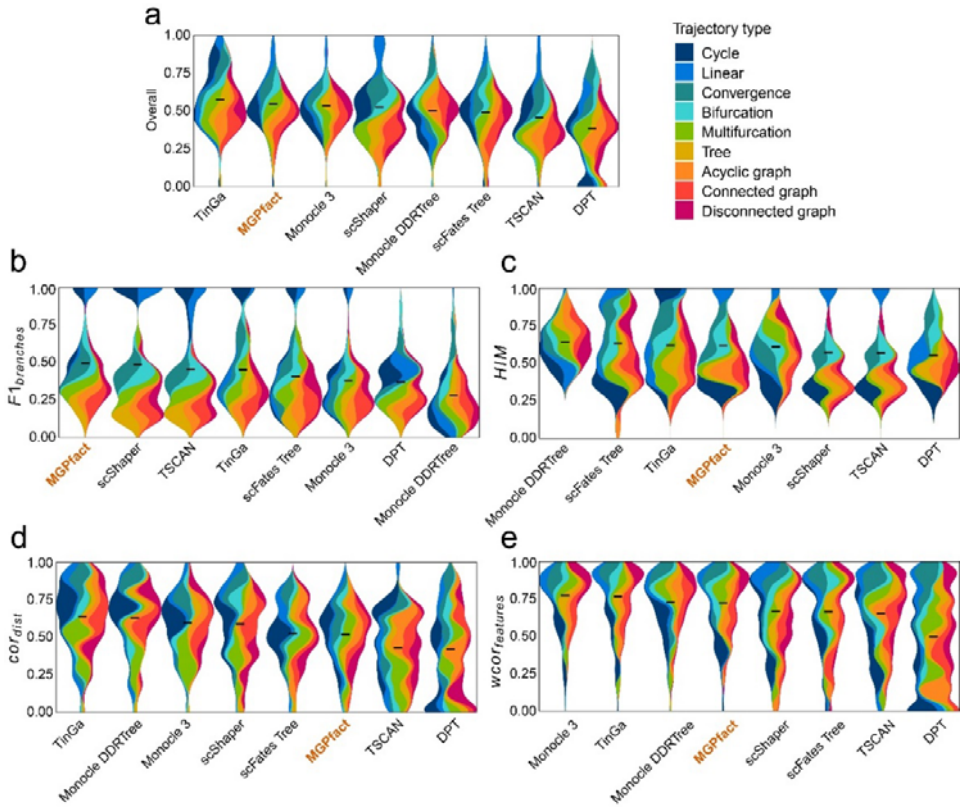

Fig. 2. Trajectory inference (TI) performance of state-of-the-art methods in 239 test datasets. a. Overall scores; b. $F1_{branches}$; c. HIM; d. cor_{dist}; e. $wcor_{features}$. All results are color-coded based on the trajectory types, with the black line representing the mean value. The "Overall" assessment is calculated as the geometric mean of all four metrics.

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.

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**

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). As for the other metrics in real test datasets, MGPfact was the 5th in the similarity of cell locations ($cor_{dist_{mean}} = 0.46$), 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, $\Delta\, cor_{dist_{mean}} = 0.07, \Delta\, wcor_{features_{mean}} = 0.05$).

In summary, our data suggest that MGPfact is highly efficient in predicting cell fate in branching trajectory ($F1_{branches}$) and topological structure (HIM), which meets the primary goals of the algorithm. 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

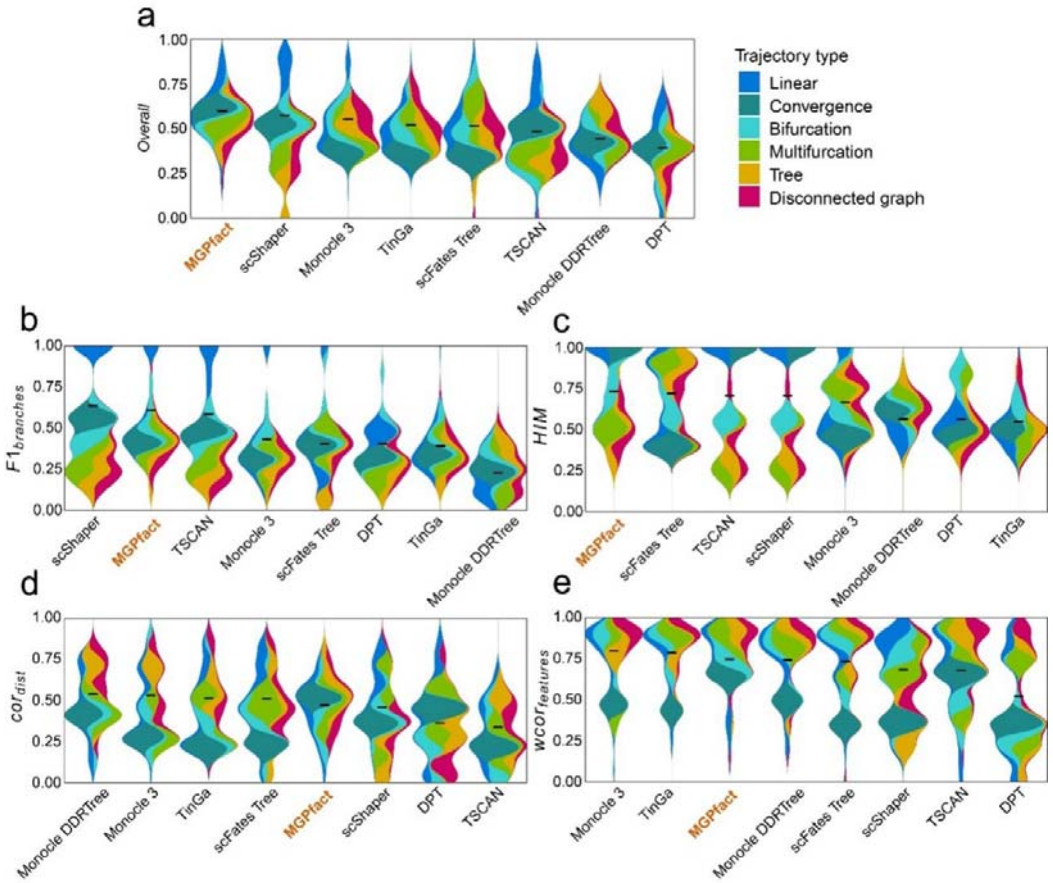

Fig 3. Trajectory inference (TI) performance of state-of-the-art methods in 68 test datasets of real cell population. a. Overall scores; b. $F1_{branches}$; c. HIM; d. cor_{dist}; e. $wcor_{features}$. All results are color-coded based on the trajectory types, with the black line representing the mean value for ranking all methods. The "Overall" assessment is calculated as the geometric mean of all four metrics.

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

196 the biological process underlying cell fate determination, we applied MGPfact to a scRNA-seq data
197 of microglia development and validated the results with experimental evidences (Li et al., 2019).

198 **MGPfact recovers the determinants of the microglia development**

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

204 The first bifurcation determines the differentiated cell fates of PAM and HM, which involves a
205 set of notable marker genes of both cell types, such as Spp1, Apoe, Gpnmb (PAM), and Tmem119
206 (HM). The second bifurcation determines the proliferative status, which is crucial for the
207 development and function of PAM and HM(Guzmán, n.d.; Li et al., 2019). The genes affected by the
208 second bifurcation are associated with cell cycle and proliferation, such as Mki67, Tubb5, Arl6ip1,
209 Top2a. The third bifurcation influences the development and maturity of microglia, of which the
210 highly weighted genes, such as Selplg, P2ry12, Siglech, Cst3, 4632428N05Rik are all previously
211 annotated markers for establishment of the fates of microglia(Anderson et al., 2022; Li et al., 2019)
212 (Supplementary Table 4).

213 Moreover, MGPfact unveiled highly active regulons in each bifurcation processes, which further
214 traced back to the influential transcription factors as determinants of microglia development, such
215 as Hif1a, E2f5, and Nfkb1 (Fig. 4d-e, Methods). Specifically, Hif1a is crucial for microglial activation
216 and directly linked to neurodegenerative disease progression(Wang et al., 2022). Our data showed
217 an upregulation of Hif1a in the PAM-branch (phase 1) of the first bifurcation, reaffirming the role of
218 Hif1a in PAM differentiation. The other two transcription factors, E2f5 and Nfkb1, were active in
219 phase 2 of the 2nd bifurcation and the 3rd bifurcation, respectively. Both are known for the roles in
220 microglial development(Dresselhaus and Meffert, 2019; Nawal, n.d.).

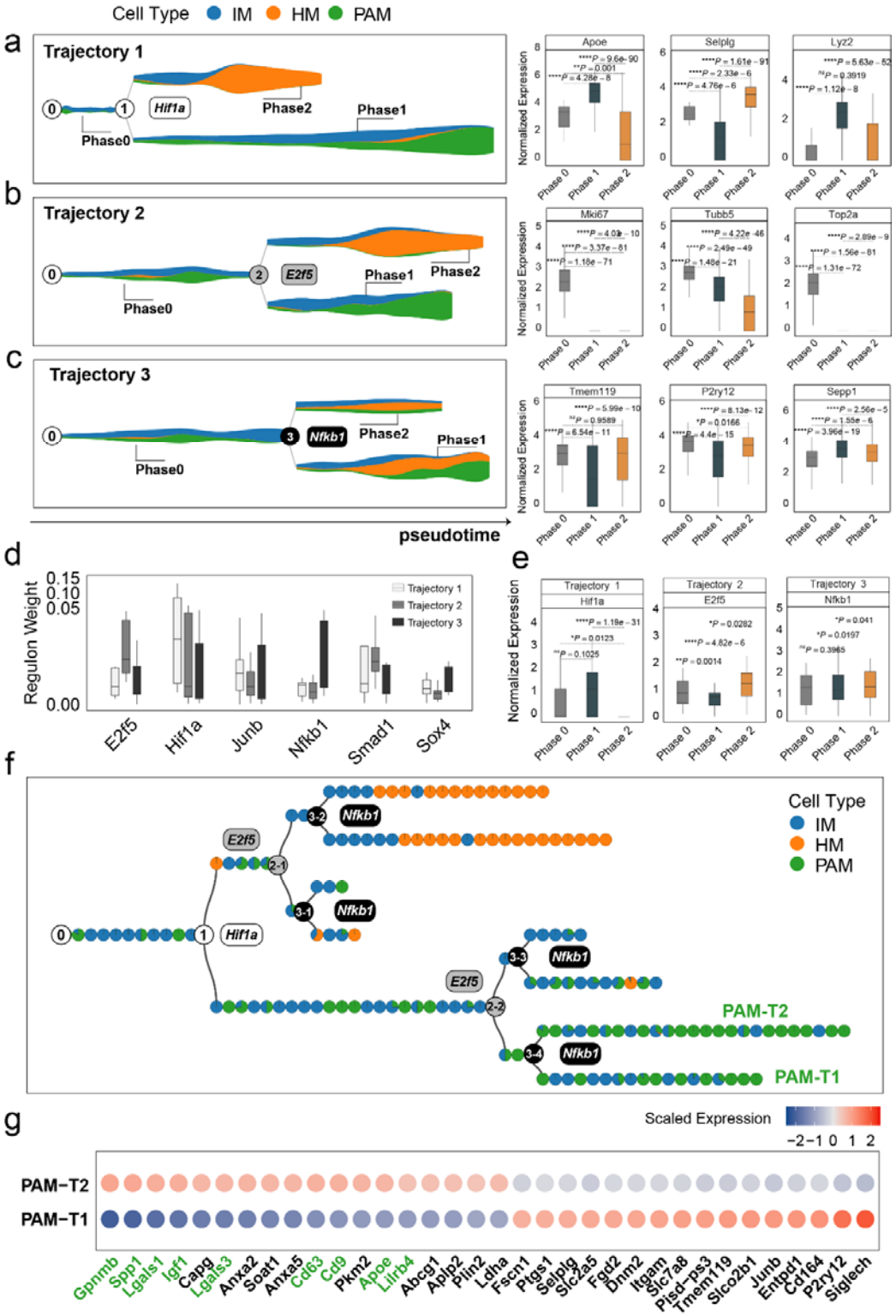

Fig. 4. **MGPfact reconstructed the developmental trajectory of microglia, recovering known determinants of microglia fate. a-c.** The inferred independent bifurcation processes with respect to the unique cell types (color-coded) of microglia development, where phase 0 corresponds to the state before bifurcation; and phases 1 and 2 correspond to the states post-bifurcation. The most highly weighted regulons in each trajectory were labeled by the corresponding transcription factors (left panels). The top three HWG of each bifurcation are significantly differently expressed in phase 1, 2 and 3 (right panel). **d.** The most highly weighted regulons influencing the three developmental

trajectories of microglia. **e.** The expression levels of the transcription factors of highly weighted regulons in each trajectory significantly differ among different phases. **f.** The consensus developmental trajectory by merging the three bifurcation processes. Point 0 denotes the initial of differentiation, whereas the notion of "*n*-*m*" denotes the *m*-th branch from the branching point *n*. Each colored circle represents a landmark (MURP) of the trajectory, showing the fraction of cell types. The transcription factors of highly weighted regulons in each bifurcation process were used to label each branching point. Particularly, PAM-T1 and PAM-T2 are the two newly defined subtypes of PAM. **g.** Selected differently expressed genes between PAM-T1 and PAM-T2 (|logfc| > 0.25, adjusted P-value < 0.1) are shown by colored-dots corresponding to the mean expression levels in either cell type. The IDs validated marker genes for PAM are labeled in green. In all box plots, the horizontal line represents the median value, and the whisker extends to the furthest data point within 1.5 times the interquartile range. Significance is denoted as: ns, not significant, * P < 0.05, ** P < 0.01, *** P < 0.001, **** P < 0.0001, Wilcoxon rank-sum test.

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) .

Using 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

261 subtypes of tumor-associated CD8$^+$ T cells with strong clinical implications.

262 **MGPfact better explains the fate of tumor-associated CD8$^+$ T cells**

263 We first evaluated the goodness-of-fit of MGPfact-derived consensus trajectory with the CD8$^+$
264 T cell subtypes characterized by the original studies. Following the original study, we used Monocle
265 2 as a reference baseline model for trajectory inference (Methods). As a result, MGPfact showed
266 significantly improved explanatory power for all CD8$^+$ T cell subtypes only with the exception for
267 CD8-GZMK in CRC ($P < 0.05$, Table 2), suggesting that MGPfact's prediction power of cellular
268 trajectory offers better consistence to the validated cell fate of CD8$^+$ T cells hence improved
269 interpretability.

270

271 **Table 2 The consensus trajectory by MGPfact offered better explanatory power of cell fate. R-square and**
272 ***P*-values based on F-tests suggest that the consensus trajectory of MGPfact showed significantly better**
273 **fitness to the subtypes of CD8$^+$ T cells characterized and annotated by experimental evidence than the**
274 **baseline model (Monocle 2).**

		MGPfact			Monocle 2		
		Adjust R^2	**P**	**F**	**Adjust R^2**	**P**	**F**
NSCLC (GSE99254)	CD8-LEF1	**0.935**	0.000	80.657	0.176	0.000	4.007
	CD8-CD28	**0.195**	0.002	3.906	0.170	0.000	3.887
	CD8-CX3CR1	**0.634**	0.000	10.581	0.259	0.000	5.924
	CD8-GZMK	**0.259**	0.000	4.597	0.189	0.000	8.386
	CD8-ZNF683	**0.232**	0.001	4.105	0.051	0.043	2.375
	CD8-LAYN	**0.435**	0.000	5.262	0.031	0.027	5.008
CRC (GSE108989)	CD8-LEF1	**0.311**	0.000	4.021	0.027	0.036	4.479
	CD8-GPR183	**0.380**	0.000	5.105	0.032	0.025	5.124
	CD8-CX3CR1	**0.648**	0.000	13.326	0.047	0.008	7.161
	CD8-GZMK	0.130	0.013	2.628	0.550	0.000	18.106
	CD8-CD6	**0.277**	0.000	5.158	0.109	0.007	2.716
	CD8-CD160	**0.124**	0.016	2.544	0.080	0.025	2.224
	CD8-LAYN	**0.741**	0.000	20.148	0.172	0.000	3.916

275

276 **MGPfact identifies T-cell gene expression signatures with clinical implications**

277 MGPfact discerned the different cellular fates of tumor-associated CD8$^+$ T cells by distinct
278 bifurcation processes. To reveal the clinical implications of these bifurcation processes, we

279 retrieved the mean expression vectors corresponding to either phase (branch) as Gene Expression
280 Signatures (GES), and stratified cancer cohorts by quantifying the propensities to certain destiny of
281 $CD8^+$ T cells (Methods). Our data demonstrated pronounced disparities in the clinical outcomes
282 associated with different fates of $CD8^+$ T cells among patients (Fig. 5a-b, Supplementary Fig. 5e).
283 In NSCLC, trajectory 1 is associated with cytotoxic T cells (96%, phase 1) and higher overall
284 survival in TCGA-lung adenocarcinoma (LUAD) patients. Trajectory 2 is associated with exhausted
285 T cells (91%, phase 2), and lower overall survival in the same cohort (Fig. 5a, Supplementary Fig.
286 3a-b). Similarly, in the CRC, trajectory 1 is associated with exhausted T cells (95%, phase 1) and
287 poor overall survival in TCGA-COAD patients (Fig. 5b, Supplementary Fig. 5a-b).

288 In addition, for each trajectory identified in NSCLC and CRC, we selected a set of HWG
289 (absolute gene weight > 0.05) to characterize the underlying biological processes and key
290 transcription factors determining the bifurcation (Supplementary Table 6-7, Supplementary Fig.
291 3c-d, Supplementary Fig. 5c-d). In NSCLC, the HWG of trajectory 1 are primarily implicated in
292 immune responses associated with antigen processing and presentation (Supplementary Table 8),
293 while trajectory 2-HWG involve processes of immune cell migration (Supplementary Table 8). For
294 CRC, trajectory 1 is enriched for genes of T cell activation and regulation (Supplementary Table 9).

295 Notably, our data showed that the weighted mean expression of the HWG of $CD8^+$ T cell
296 trajectories (Methods) are associated with responses to immune checkpoint inhibitors (ICIs) in
297 multiple independent cohorts (Fig. 5c-d, Supplementary Fig. 5f). For instance, the weighted means
298 of HWG of trajectories 1 and 2 in NSCLC which are associated with high activities of cytotoxic T
299 cells, predicted better responses to anti-PD-130 and anti-CTLA-4 (Cho et al., 2020; Liu et al., 2018),
300 as well as their combination therapies(Auslander et al., 2018) ($AUC \in \{0.813, 0.964\}$, $P < 0.1$,
301 Supplementary Fig. 4). Similarly, HWG pertaining to trajectory 1 in CRC is associated with high
302 proportion of $EMRA$ (87%, phase 1), hence better responses to immunotherapies in 4
303 cohorts(Auslander et al., 2018; Cho et al., 2020; Jung et al., 2019; Liu et al., 2018) ($AUC \in$
304 $\{0.796, 0.964\}$, $P < 0.01$, Supplementary Fig. 6).

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

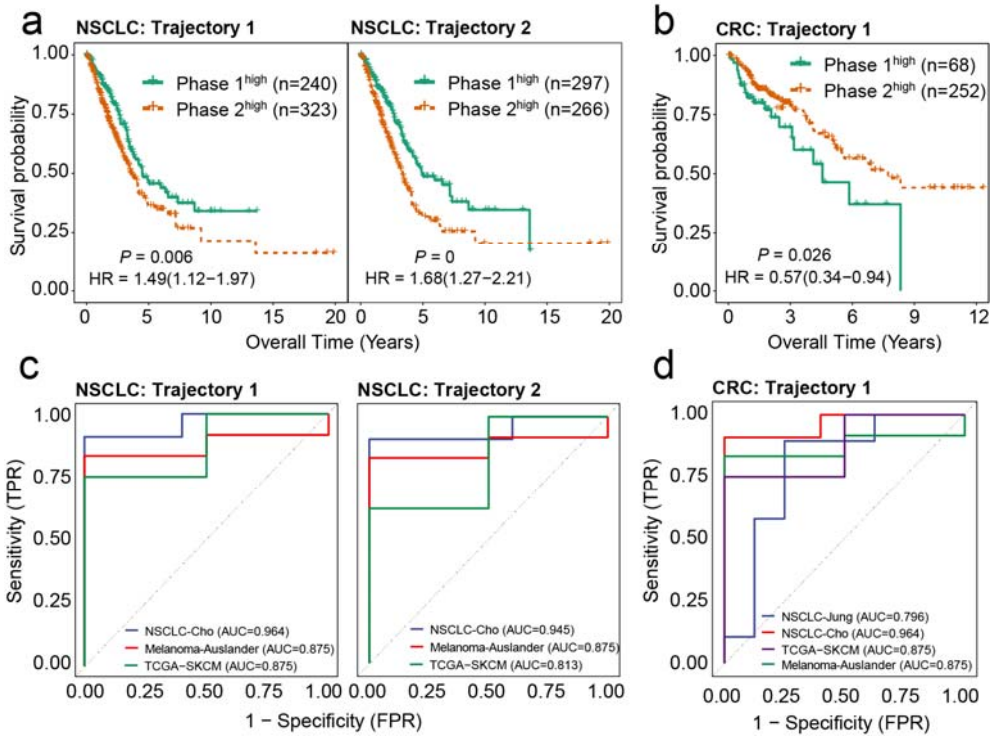

Fig. 5. Highly weighted genes (HWG) of the bifurcation processes of CD8$^+$ T cells serve as reliable indicators for clinical outcome and ICI treatment response. a. Gene expression signatures (GES) corresponding to HWG in CD8$^+$ T cells trajectory 1 and 2 in NSCLC predict overall survival of the TCGA-LUAD 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.

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. 3c-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 10). In CRC, we identified two subtypes of effector memory T cells (CD8-GZMK), CD8-GZMK-T1

329 (EOMES-/BHLHE40+) and CD8-GZMK-T2 (EOMES-/BHLHE40-) (Fig. 6d, Supplementary Fig.
330 5c-d). CD8-GZMK-T2 strongly resembles CD8-GZMK, and potentially differentiating into CD8-CD6
331 (resident memory T cells, T_{RM}) and CD8-CD160 (intraepithelial lymphocytes, IEL); whereas
332 CD8-GZMK-T1 cells demonstrated higher expression of GZMB, indicating an active cytotoxic
333 cell-killing capability (Trapani, 2001). Simultaneously, these cells also marked by high expression
334 levels of immune related genes, including OASL, RBPJ, and CTLA4, which are known targets of
335 BHLHE40 (Lutter et al., 2022; Salmon et al., 2022), implying that BHLHE40 is a modulator of the
336 higher effector activity in CD8-GZMK-T1 (Fig. 6e, Supplementary Table 11).

337 We further derived scores based on the differentially expressed genes of CD8-ZNF683-T1 and
338 CD8-GZMK-T1 (Methods), as measures of the fraction of each subtype in cancer cohorts. In the
339 LUAD cohort of TCGA, increased fraction of CD8-ZNF683-T1 in TIME was associated with
340 favorable outcomes (Fig. 6c). And increased fractions of CD8-GZMK-T1 in TIME were associated
341 with better responses to ICI therapy across 3 independent cohorts[31–33] (Fig. 6f, Supplementary Fig.
342 7, $AUC \in \{0.782, 0.917\}$, $P < 0.1$), which were treated with anti-PD-1, anti-CTLA-4 and their
343 combination therapies.

344 In conclusion, our data showed the cellular trajectory inferred by MGPfact can be used to
345 elucidate the complex evolutionary processes of tumor-associated $CD8^+$ T cells, and further inform
346 the characterization of new subtypes of T cells with significant clinical implications.

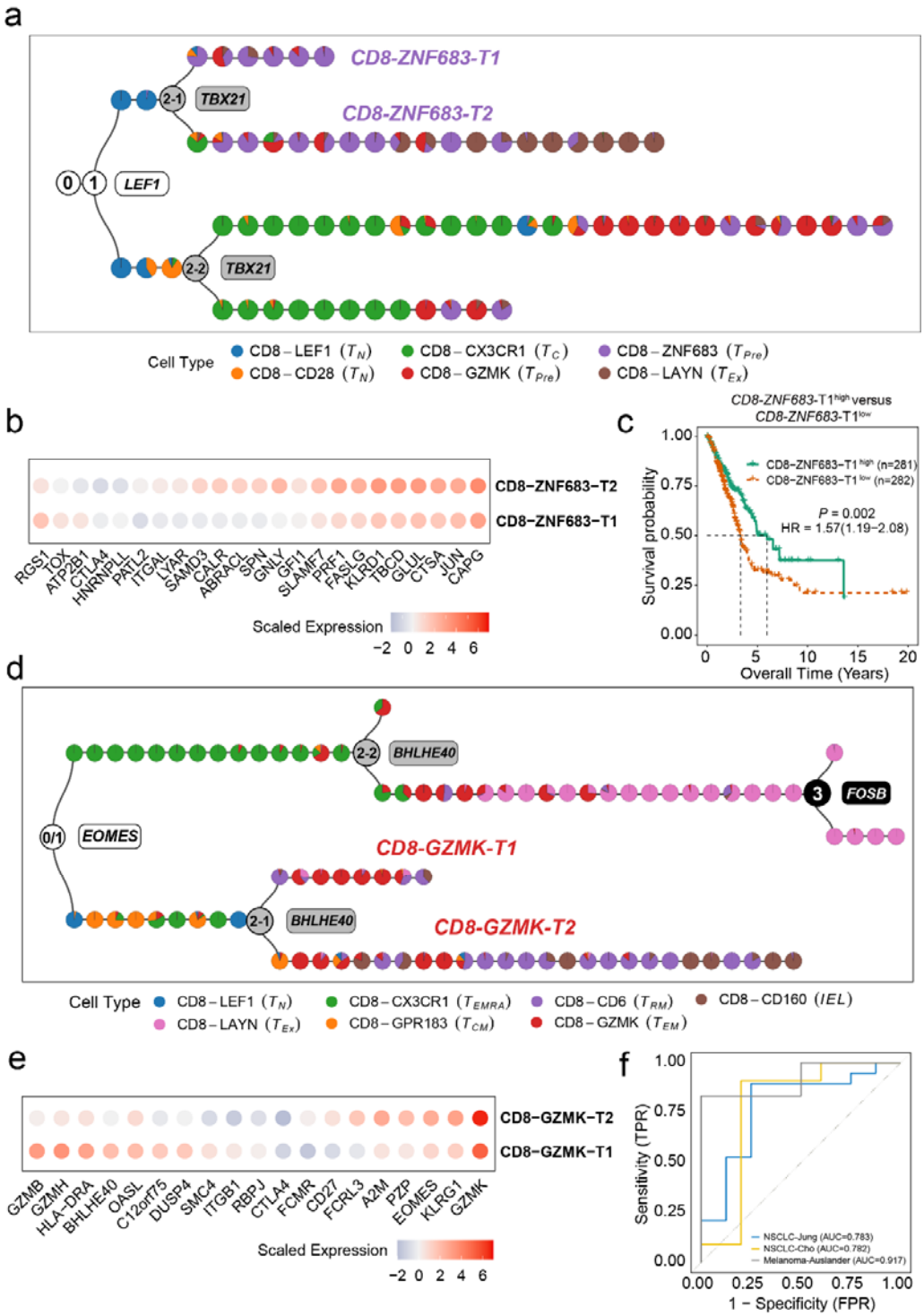

Fig. 6. MGPfact serves as an effective approach for characterization of new cellular subtypes. a. The consensus trajectory of tumor-associated CD8$^+$ T cells in NSCLC identified CD8-ZNF683-T1 and CD8-ZNF683-T2 as two subtypes of CD8-ZNF683, which are influenced by TBX21. **b.** Selected differently expressed genes between CD8-ZNF683-T1 and CD8-ZNF683-T2 (|logfc| > 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 differently expressed genes between CD8-GZMK-T1 and CD8-GZMK-T2 (|logfc| > 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.

Discussion

Single-cell RNA sequences (scRNA-seq) offer direct, quantitative snapshot of a cell population involved in actual biological activities hence tremendous insight of the variability of cells' states and functions. So far, most studies use scRNA-seq to characterize new subtypes of cells by various clustering and embedding algorithms. These algorithms are usually optimized for efficiently revealing discrete biological states of cells corresponding to distinct functions. Nevertheless, as the resolution increases, clustering algorithms becomes more sensitive to noises and less efficient in delineating continuous biological process(Ren et al., 2022). The challenges are reflected in the discrepancy of definition of new cell types, over- and under-clustering due to imbalanced sampling.

The introduction of manifold-learning in scRNA-seq analysis provides a new dimension for biological knowledge discovery. The trajectories inferred by manifold-learning methods are more suitable representation of continuous variation in cells, which are usually implicated in many time-dependent biological processes, such as development, differentiation and clonal evolution. Particularly, manifold-learning is capable to reconstruct pseudotime trajectory from cross-sectional scRNA-seq data, which offers the unique advantage to delineating complex evolutionary processes of cells sampled from real clinical biopsies. Thus far, trajectory-inference is used by many recent studies to elucidate the origin and developmental pathways of different cell states in the temporal space, also known as cell fate.

Despite of the latest booming of new methods, existing manifold-learning methods are faced with major limitations. First, many trajectory inference (TI) methods still requires prior information on pseudtime and cell clustering, hence restrictions in the applicability. Thus, the resulted cellular trajectory based on cell clustering represents the admixture of both continuous temporal variation and discrete subclones of cells. Then, most of the current manifold-learning methods are based on either single feature (gene), or the whole feature space, such as all highly variable genes (HVG), hence the results are highly dependent and even biased on the input feature (genes). Besides, the lack of necessary gene-selection in the inference makes it challenging to annotate the resulting

388 trajectory to known ontological terms or pathways, which hinders the interpretability of the results
389 and the follow-up functional study. In some cases, discrepancy can arise between trajectory
390 inferred from all genes and cell type characterized by specific gene sets, which further restrict the
391 use of manifold-learning in scRNA-seq analysis.

392 We developed MGPfact to overcome the limitations of the existing methods. Inspired by recent
393 studies, MGPfact model the cell fate as mixture of Gaussian processes, which accommodate both
394 continuous evolution pathway and biphasic destiny of cell fate. Thus, MGPfact is capable to
395 distinguish discrete and continuous events in the same trajectory. In addition, by factorizing the
396 mixture Gaussian processes, MGPfact offers the advantage to select genes corresponding to each
397 bifurcation process and thereby enable full biological annotation and interpretation of the trajectory.
398 As a validation, we showed that gene-selection by MGPfact consistently recapitulated the
399 development of microglia and tumor-associated $CD8^+$ T cells; and recovered key regulators of
400 distinct cell fate. So far, MGPfact is the only model-based manifold-learning framework which
401 factorizes complex development trajectories into independent bifurcation processes of gene sets.

402 We conducted a comprehensive comparison of MGPfact with existing TI methods from various
403 perspectives. This comparison included the correlation of cell sorting, accuracy of branch allocation,
404 similarity of topological structures and differentially expressed features. For the overall
405 TI-performance, MGPfact demonstrated leading performance across 239 datasets, second only to
406 TinGa. For the performance in branch allocation, which directly reflect the fitness to the outcoms of
407 cell-fate, MGPfact outperformed its counterparts, especially in the topology groups of linear and
408 bifurcation. As for $wcor_{features}$ and cor_{dist}, MGPfact performed less well mainly for two reasons.
409 First MGPfact is designed for bifurcation topology in the cellular trajectory, hence less efficient in
410 inferring complex topologies. Then, MGPfact inference is based on selected gene sets instead of
411 the whole transcriptome, the resulted trajectory correspond only to the bifurcation processes of
412 interest hence does not necessarily reflect the whole topology of cellular trajectory. Furthermore,
413 MGPfact performed significantly better in trajectory prediction in real cell population compared to
414 the synthetic ones, suggesting the algorithm fit better to the true biological variation and noises.

415 To reconstruct the trajectory of cell fate, we merged all the bifurcation processes into a
416 consensus trajectory. In the validation by microglia and tumor-associated $CD8^+$ T cells, the
417 consensus trajectory revealed highly consistent findings recovering known biology and the marker
418 genes of specific cell states which further inform the transcription factor (TF) determining the fate of
419 cell. In addition, the consensus trajectory revealed new subtypes of cells demonstrating highly
420 relevant transcriptional characteristics. Particularly, we reported new subtypes of tumor associated

421 $CD8^+$ T cells characterized by different TBX21 and BHLHE40 activity, both of which are known
422 regulators of $CD8^+$ T cell functionality(Lutter et al., 2022; Pritchard et al., 2023; Salmon et al., 2022;
423 Trapani, 2001). These data suggest that MGPfact is capable to discover gene modules with strong
424 and consistent transcriptional background hence better interpretability. Moreover, the results of
425 MGPfact demonstrated strong clinical relevance. Using MGPfact we retrieved gene expression
426 signatures (GES) which quantitatively measure the propensity and fraction of different fates of
427 $CD8^+$ T cell. These signatures correspond to important biological processes of T cell activity; and
428 predict clinical outcome and ICI treatment responses from transcriptome data of bulk tumor
429 biopsies, independent of any endogenous feature of the tumor cells.

430 Nevertheless, the current method also has some limitations, which shall be addressed in future
431 study. The complex definition of the bifurcation kernel introduces unfavorable singularity to the
432 Gaussian kernel when considering numerous trajectories. Additionally, the current
433 MGPfact-inferred trajectory is based solely on pseudotime, neglecting potential bifurcation
434 processes occurring in space. Future models should incorporate spatial dynamics of transcription
435 and analyze cellular velocity data to provide more comprehensive insights. Furthermore, the
436 reconstruction of cellular trajectories by MGPfact implies independence of each bifurcation
437 processes, which may not reflect real cellular behavior. Therefore, predictions from MGPfact should
438 be interpreted with caution and validated experimentally.

439 Methods

440 Benchmarking MGPfact to state-of-the-art methods

441 We adopted a comprehensive evaluation framework from previous scRNA-seq study to assess
442 the TI performance of MGPfact(Saelens et al., 2019; Smolander et al., 2022; Todorov et al., 2020).
443 The validation dataset comprises 110 real data and 229 synthetic data, encompassing 9 different
444 cellular trajectory topologies. The ground truth of cellular trajectories of each dataset were inferred
445 and validated by the original study(Saelens et al., 2019). The evaluation of TI performance was
446 based on five metrics: HIM for assessing trajectory topology similarity, $F1_{branches}$ for measuring
447 branch allocation similarity, cor_{dist} for quantifying inter-cell distance similarity, $wcor_{features}$ for
448 evaluating the consistency of significant features between predicted and actual trajectories, and the
449 $overall$ performance score, which is the geometric mean of the above four metrics.

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

We use 100 training datasets to determine the optimal parameter settings. 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, 2017a), TSCAN(Ji and Ji, 2016), and DPT(Haghverdi et al., 2016), as well as four new methods from recent studies: Monocle 3(Cao et al., 2019), scShaper(Smolander et al., 2022), scFates(Faure et al., 2023), and TinGa(Todorov et al., 2020). 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). 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 highly weighted genes (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, 2017a) 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). Using Seurat (Butler et al., 2018; Stuart et al., 2019), 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 $[0.5,\infty]$; 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) and colorectal cancer (GSE108989) (Zhang et al., 2018) 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 (HWG)

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) to perform functional annotation using GO terms(Consortium, 2004), 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) transcription factor regulatory program estimation for each analysis case. GENIE3(Huynh-Thu et al., 2010) was used to identify co-expressed modules from the results of MURP(Ren et al., 2022) downsampling. RCisTarget(Aerts et al., 2010; Aibar et al., 2017) was then used to identify regulons before AUCell(Aibar et al., 2017) 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.

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; Zhang et al., 2018).

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/).

We assessed the survival impacts of NSCLC and CRC bifurcation processes in the TCGA

539 LUAD and COAD cohorts, respectively. For each independent bifurcation process, we defined
540 Gene Expression Signatures (GES) by the mean expression vectors of all trajectories in phase 1
541 and 2, respectively. Subsequently, we calculated the Peason's correlation coefficients for each
542 individual expression profile in TCGA LUAD or COAD cohorts to the phase 1 and 2 GES, where
543 higher correlation correspond to stronger propensity to specific cell fate. This allowed us to classify
544 patients into two groups: those exhibiting propensity to phase 1 and those exhibiting propensity to
545 phase 2.

546 To assess the survival impacts of specific cell states defined by the consensus trajectory, we
547 developed a CD8-ZNF683-T1 score based on the signed average expression level of DEG
548 associated with CD8-ZNF683-T1 (Supplementary methods). Subsequently, we classified the TCGA
549 LUAD cohorts into two groups using the median of CD8-ZNF683-T1 scores, identifying those
550 demonstrating a propensity towards CD8-ZNF683-T1 and those demonstrating a propensity
551 towards CD8-ZNF683-T2.

552 The Cox regression model was implemented using R-4.2 package "survival". And we
553 generated Kaplan-Meier survival curves based on different classifiers to illustrate differences in
554 survival time and report the statistical significance based on Log-rank test.

555 Immune-checkpoint inhibitor treatment response analysis

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

563 To validate the bifurcation processes identified by MGPfact predicting patients' response to
564 immune-checkpoint inhibitor (ICI) treatments, we selected highly weighted genes (HWG) with
565 absolute weights greater than 0.05 from each independent bifurcation process. Then, we
566 calculated the weighted mean expression of HWG in each ICI dataset to generate ROC curves for
567 patient drug response.

568 To validate the specific cell states defined by the consensus trajectory predicting patients'
569 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.

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. Furthermore, 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).

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.

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.

Reference

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**:e1000435. doi:10.1371/journal.pbio.1000435

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.

Auslander N, Zhang G, Lee JS, Frederick DT, Miao B, Moll T, Tian T, Wei Z, Madan S, Sullivan RJ, Boland G, Flaherty K, Herlyn M, Ruppin E. 2018. Robust prediction of response to immune checkpoint blockade therapy in metastatic melanoma. *Nat Med* **24**:1545–1549. doi:10.1038/s41591-018-0157-9

B J S. 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. doi: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. doi:10.1038/nbt.4096

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

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. doi: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**:3685. doi:10.1038/s41467-018-05988-7

Dresselhaus EC, Meffert MK. 2019. Cellular Specificity of NF-κB Function in the Nervous System. *Front Immunol* **10**:1043. doi: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**:btac746. doi:10.1093/bioinformatics/btac746

Fritzke B. 1994. A growing neural gas network learns topologies. *Advances in neural information processing systems* **7**.

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.

Haghverdi L, Buettner F, Theis FJ. 2015. Diffusion maps for high-dimensional single-cell analysis of differentiation data. *Bioinformatics* **31**:2989–2998. doi:10.1093/bioinformatics/btv325

Haghverdi L, Büttner M, Wolf FA, Buettner F, Theis FJ. 2016. Diffusion pseudotime robustly reconstructs lineage branching. *Nat Methods* **13**:845–848. doi:10.1038/nmeth.3971

Hugo W, Zaretsky JM, Sun L, Song C, Moreno BH, Hu-Lieskovan S, Berent-Maoz B, Pang J, Chmielowski B, Cherry G, Seja E, Lomeli S, Kong X, Kelley MC, Sosman JA, Johnson DB, Ribas A, Lo RS. 2016. Genomic and Transcriptomic Features of Response to Anti-PD-1 Therapy in Metastatic Melanoma. *Cell* **165**:35–44. doi: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**:e12776. doi: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. doi: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**:4278. doi:10.1038/s41467-019-12159-9

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

Liu Jianfang, Lichtenberg T, Hoadley KA, Poisson LM, Lazar AJ, Cherniack AD, Kovatich AJ, Benz CC, Levine DA, Lee AV, Omberg L, Wolf DM, Shriver CD, Thorsson V, Hu H, Caesar-Johnson SJ, Demchok JA, Felau I, Kasapi M, Ferguson ML, Hutter CM, Sofia HJ, Tarnuzzer R, Wang Z, Yang L, Zenklusen JC, Zhang J (Julia), Chudamani S, Liu Jia, Lolla L, Naresh R, Pihl T, Sun Q, Wan Y, Wu Y, Cho J, DeFreitas T, Frazer S, Gehlenborg N, Getz G, Heiman DI, Kim J, Lawrence MS, Lin P, Meier S, Noble MS, Saksena G, Voet D, Zhang Hailei, Bernard B, Chambwe N, Dhankani V, Knijnenburg T, Kramer R, Leinonen K, Liu Y, Miller M, Reynolds S, Shmulevich I, Thorsson V, Zhang W, Akbani R, Broom BM, Hegde AM, Ju Z, Kanchi RS, Korkut A, Li J, Liang H, Ling S, Liu W, Lu Y, Mills GB, Ng K-S, Rao A, Ryan M, Wang Jing, Weinstein JN, Zhang J, Abeshouse A, Armenia J, Chakravarty D, Chatila WK, De Bruijn I, Gao J, Gross BE, Heins ZJ, Kundra R, La K, Ladanyi M, Luna A, Nissan MG, Ochoa A, Phillips SM, Reznik E, Sanchez-Vega F, Sander C, Schultz N, Sheridan R, Sumer SO, Sun Y, Taylor BS, Wang Jioajiao, Zhang Hongxin, Anur P, Peto M, Spellman P, Benz C, Stuart JM, Wong CK, Yau C, Hayes DN, Parker JS, Wilkerson MD, Ally A, Balasundaram M, Bowlby R, Brooks D, Carlsen R, Chuah E, Dhalla N, Holt R, Jones SJM, Kasaian K, Lee D, Ma Y, Marra MA, Mayo M, Moore RA, Mungall AJ, Mungall K, Robertson AG, Sadeghi S, Schein JE, Sipahimalani P, Tam A, Thiessen N, Tse K, Wong T, Berger AC, Beroukhim R, Cherniack AD, Cibulskis C, Gabriel SB, Gao GF, Ha G, Meyerson M, Schumacher SE, Shih J, Kucherlapati MH, Kucherlapati RS, Baylin S, Cope L, Danilova L, Bootwalla MS, Lai PH, Maglinte DT, Van Den Berg DJ, Weisenberger DJ, Auman JT, Balu S, Bodenheimer T, Fan C, Hoadley KA, Hoyle AP, Jefferys SR, Jones

681 CD, Meng S, Mieczkowski PA, Mose LE, Perou AH, Perou CM, Roach J, Shi Y, Simons JV, Skelly T,
682 Soloway MG, Tan D, Veluvolu U, Fan H, Hinoue T, Laird PW, Shen H, Zhou W, Bellair M, Chang K,
683 Covington K, Creighton CJ, Dinh H, Doddapaneni H, Donehower LA, Drummond J, Gibbs RA, Glenn R,
684 Hale W, Han Y, Hu J, Korchina V, Lee S, Lewis L, Li W, Liu X, Morgan M, Morton D, Muzny D, Santibanez
685 J, Sheth M, Shinbro E, Wang L, Wang M, Wheeler DA, Xi L, Zhao F, Hess J, Appelbaum EL, Bailey M,
686 Cordes MG, Ding L, Fronick CC, Fulton LA, Fulton RS, Kandoth C, Mardis ER, McLellan MD, Miller CA,
687 Schmidt HK, Wilson RK, Crain D, Curley E, Gardner J, Lau K, Mallery D, Morris S, Paulauskis J, Penny R,
688 Shelton C, Shelton T, Sherman M, Thompson E, Yena P, Bowen J, Gastier-Foster JM, Gerken M, Leraas
689 KM, Lichtenberg TM, Ramirez NC, Wise L, Zmuda E, Corcoran N, Costello T, Hovens C, Carvalho AL, De
690 Carvalho AC, Fregnani JH, Longatto-Filho A, Reis RM, Scapulatempo-Neto C, Silveira HCS, Vidal DO,
691 Burnette A, Eschbacher J, Hermes B, Noss A, Singh R, Anderson ML, Castro PD, Ittmann M, Huntsman
692 D, Kohl B, Le X, Thorp R, Andry C, Duffy ER, Lyadov V, Paklina O, Setdikova G, Shabunin A, Tavobilov M,
693 McPherson C, Warnick R, Berkowitz R, Cramer D, Feltmate C, Horowitz N, Kibel A, Muto M, Raut CP,
694 Malykh A, Barnholtz-Sloan JS, Barrett W, Devine K, Fulop J, Ostrom QT, Shimmel K, Wolinsky Y, Sloan
695 AE, De Rose A, Giuliante F, Goodman M, Karlan BY, Hagedorn CH, Eckman J, Harr J, Myers J, Tucker K,
696 Zach LA, Deyarmin B, Hu H, Kvecher L, Larson C, Mural RJ, Somiari S, Vicha A, Zelinka T, Bennett J,
697 Iacocca M, Rabeno B, Swanson P, Latour M, Lacombe L, Têtu B, Bergeron A, McGraw M, Staugaitis SM,
698 Chabot J, Hibshoosh H, Sepulveda A, Su T, Wang T, Potapova O, Voronina O, Desjardins L, Mariani O,
699 Roman-Roman S, Sastre X, Stern M-H, Cheng F, Signoretti S, Berchuck A, Bigner D, Lipp E, Marks J,
700 McCall S, McLendon R, Secord A, Sharp A, Behera M, Brat DJ, Chen A, Delman K, Force S, Khuri F,
701 Magliocca K, Maithel S, Olson JJ, Owonikoko T, Pickens A, Ramalingam S, Shin DM, Sica G, Van Meir
702 EG, Zhang Hongzheng, Eijckenboom W, Gillis A, Korpershoek E, Looijenga L, Oosterhuis W, Stoop H,
703 Van Kessel KE, Zwarthoff EC, Calatozzolo C, Cuppini L, Cuzzubbo S, DiMeco F, Finocchiaro G, Mattei L,
704 Perin A, Pollo B, Chen C, Houck J, Lohavanichbutr P, Hartmann A, Stoehr C, Stoehr R, Taubert H, Wach
705 S, Wullich B, Kycler W, Murawa D, Wiznerowicz M, Chung K, Edenfield WJ, Martin J, Baudin E, Bubley G,
706 Bueno R, De Rienzo A, Richards WG, Kalkanis S, Mikkelsen T, Noushmehr H, Scarpace L, Girard N,
707 Aymerich M, Campo E, Giné E, Guillermo AL, Van Bang N, Hanh PT, Phu BD, Tang Y, Colman H, Evason
708 K, Dottino PR, Martignetti JA, Gabra H, Juhl H, Akeredolu T, Stepa S, Hoon D, Ahn K, Kang KJ,
709 Beuschlein F, Breggia A, Birrer M, Bell D, Borad M, Bryce AH, Castle E, Chandan V, Cheville J, Copland
710 JA, Farnell M, Flotte T, Giama N, Ho T, Kendrick M, Kocher J-P, Kopp K, Moser C, Nagorney D, O'Brien D,
711 O'Neill BP, Patel T, Petersen G, Que F, Rivera M, Roberts L, Smallridge R, Smyrk T, Stanton M,
712 Thompson RH, Torbenson M, Yang JD, Zhang L, Brimo F, Ajani JA, Angulo Gonzalez AM, Behrens C,
713 Bondaruk J, Broaddus R, Czerniak B, Esmaeli B, Fujimoto J, Gershenwald J, Guo C, Lazar AJ,
714 Logothetis C, Meric-Bernstam F, Moran C, Ramondetta L, Rice D, Sood A, Tamboli P, Thompson T,
715 Troncoso P, Tsao A, Wistuba I, Carter C, Haydu L, Hersey P, Jakrot V, Kakavand H, Kefford R, Lee K,
716 Long G, Mann G, Quinn M, Saw R, Scolyer R, Shannon K, Spillane A, Stretch J, Synott M, Thompson J,
717 Wilmott J, Al-Ahmadie H, Chan TA, Ghossein R, Gopalan A, Levine DA, Reuter V, Singer S, Singh B, Tien
718 NV, Broudy T, Mirsaidi C, Nair P, Drwiega P, Miller J, Smith J, Zaren H, Park J-W, Hung NP, Kebebew E,
719 Linehan WM, Metwalli AR, Pacak K, Pinto PA, Schiffman M, Schmidt LS, Vocke CD, Wentzensen N,
720 Worrell R, Yang H, Moncrieff M, Goparaju C, Melamed J, Pass H, Botnariuc N, Caraman I, Cernat M,
721 Chemencedji I, Clipca A, Doruc S, Gorincioi G, Mura S, Pirtac M, Stancul I, Tcaciuc D, Albert M,
722 Alexopoulou I, Arnaout A, Bartlett J, Engel J, Gilbert S, Parfitt J, Sekhon H, Thomas G, Rassl DM, Rintoul
723 RC, Bifulco C, Tamakawa R, Urba W, Hayward N, Timmers H, Antenucci A, Facciolo F, Grazi G, Marino M,
724 Merola R, De Krijger R, Gimenez-Roqueplo A-P, Piché A, Chevalier S, McKercher G, Birsoy K, Barnett G,
725 Brewer C, Farver C, Naska T, Pennell NA, Raymond D, Schilero C, Smolenski K, Williams F, Morrison C,
726 Borgia JA, Liptay MJ, Pool M, Seder CW, Junker K, Omberg L, Dinkin M, Manikhas G, Alvaro D, Bragazzi
727 MC, Cardinale V, Carpino G, Gaudio E, Chesla D, Cottingham S, Dubina M, Moiseenko F, Dhanasekaran
728 R, Becker K-F, Janssen K-P, Slotta-Huspenina J, Abdel-Rahman MH, Aziz D, Bell S, Cebulla CM, Davis A,
729 Duell R, Elder JB, Hilty J, Kumar B, Lang J, Lehman NL, Mandt R, Nguyen P, Pilarski R, Rai K,

Schoenfield L, Senecal K, Wakely P, Hansen P, Lechan R, Powers J, Tischler A, Grizzle WE, Sexton KC, Kastl A, Henderson J, Porten S, Waldmann J, Fassnacht M, Asa SL, Schadendorf D, Couce M, Graefen M, Huland H, Sauter G, Schlomm T, Simon R, Tennstedt P, Olabode O, Nelson M, Bathe O, Carroll PR, Chan JM, Disaia P, Glenn P, Kelley RK, Landen CN, Phillips J, Prados M, Simko J, Smith-McCune K, VandenBerg S, Roggin K, Fehrenbach A, Kendler A, Sifri S, Steele R, Jimeno A, Carey F, Forgie I, Mannelli M, Carney M, Hernandez B, Campos B, Herold-Mende C, Jungk C, Unterberg A, Von Deimling A, Bossler A, Galbraith J, Jacobus L, Knudson M, Knutson T, Ma D, Milhem M, Sigmund R, Godwin AK, Madan R, Rosenthal HG, Adebamowo C, Adebamowo SN, Boussioutas A, Beer D, Giordano T, Mes-Masson A-M, Saad F, Bocklage T, Landrum L, Mannel R, Moore K, Moxley K, Postier R, Walker J, Zuna R, Feldman M, Valdivieso F, Dhir R, Luketich J, Mora Pinero EM, Quintero-Aguilo M, Carlotti, Jr CG, Dos Santos JS, Kemp R, Sankarankuty A, Tirapelli D, Catto J, Agnew K, Swisher E, Creaney J, Robinson B, Shelley CS, Godwin EM, Kendall S, Shipman C, Bradford C, Carey T, Haddad A, Moyer J, Peterson L, Prince M, Rozek L, Wolf G, Bowman R, Fong KM, Yang I, Korst R, Rathmell WK, Fantacone-Campbell JL, Hooke JA, Kovatich AJ, Shriver CD, DiPersio J, Drake B, Govindan R, Heath S, Ley T, Van Tine B, Westervelt P, Rubin MA, Lee JI, Aredes ND, Mariamidze A. 2018. An Integrated TCGA Pan-Cancer Clinical Data Resource to Drive High-Quality Survival Outcome Analytics. *Cell* **173**:400-416.e11. doi: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**:e1420. doi:10.1002/cti2.1420

Nawal HS. n.d. A Systems Biology Perspective of Stem Cell Differentiation into Microglia.

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

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**:e20191287. doi:10.1084/jem.20191287

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

Qiu X, Mao Q, Tang Y, Wang L, Chawla R, Pliner HA, Trapnell C. 2017b. Reversed graph embedding resolves complex single-cell trajectories. *Nat Methods* **14**:979–982. doi: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**: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

Salmon AJ, Shavkunov AS, Miao Q, Jarjour NN, Keshari S, Esaulova E, Williams CD, Ward JP, Highsmith AM, Pineda JE, Taneja R, Chen K, Edelson BT, Gubin MM. 2022. BHLHE40 Regulates the T-Cell Effector Function Required for Tumor Microenvironment Remodeling and Immune Checkpoint Therapy Efficacy. *Cancer Immunology Research* **10**:597–611. doi: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. doi:10.1016/j.jmp.2018.03.001

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

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.e21.

doi: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. doi:10.1093/bioinformatics/btaa463

Trapani JA. 2001. Granzymes: a family of lymphocyte granule serine proteases. *Genome Biol* **2**:reviews3014.1. doi: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**:113412. doi: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.

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