Geometric Quantification of Cell Phenotype Transition Manifolds with Information Geometry

Miao Huang[1,2], Junda Ying[3], Yuxuan Wang[4], Haijun Zhou[1,2], Lei Zhang[5], Weikang Wang[1,2]

[1]CAS Key Laboratory for Theoretical Physics, Institute of Theoretical Physics, Chinese Academy of Sciences, Beijing 100190, China

[2] School of Physical Sciences, University of Chinese Academy of Sciences, Beijing 100049, China

[3]School of Mathematical Sciences, Peking University, Beijing 100871, China

[4]Center for Data Science, Peking University, Beijing 100871, China

[5]Beijing International Center for Mathematical Research, Center for Quantitative Biology, Center for Machine Learning Research, Peking University, Beijing 100871, China.

Email: wangwk@itp.ac.cn, zhangl@math.pku.edu.cn

Key word:
Cell phenotype transition manifolds, Information geometry, Curvature, Information velocity, Critical point

Abstract

Cell phenotype transition (CPT) plays a pivotal role in various biological processes like development. Recent advancements in single-cell sequencing techniques have uncovered that cell transition dynamics during development are confined on low-dimensional manifolds. However, existing methods are inadequate for directly quantifying the manifolds from experimental data. Here we present SCIM (single cell information manifolds), a novel geometry-guided method to quantify the CPT manifolds using information geometry. In particular, we convert single cells' high-dimensional gene vectors into probability distributions via Gaussian embedding. The Fisher metric is naturally defined in this embedding space. With the transformed

Gaussian distributions, we calculate the coarse Ricci curvature of each single cell. Our analyses reveal that the cells with low curvature are associated with critical transitions. To further examine the invariant characteristics of the manifolds of CPT, we compute the information velocity of each single cell based on RNA velocity. Remarkably, the regions with high information velocity correspond with the low curvature regions, indicating that the geometry can guide the dynamics of single cells on the manifolds. The proposed method not only unveils the invariant characteristics of the CPT manifolds, but also establishes a generic approach for quantifying the intricate dynamics on the CPT manifolds.

Introduction

Biological systems, despite their inherent complexity, exhibit remarkable precision in various biological processes. A prime example is observed in embryo development, where the transition of stem cells into distinct phenotypes orchestrates the formation of well-organized organisms. Waddington's landscape analogy is a well-known metaphor for illustrating embryo development. Recent advances in single-cell RNA sequencing have enabled a quantitative description of the landscape [1-3]. Most current methods define scalar potentials and fall short in fully capturing the dynamics of development, as sequencing is destructive to cells. To tackle this issue, algorithms have been developed to infer RNA velocity using counts of spliced and unspliced RNA, facilitating the learning of a vector field in the gene space of individual cells [4-8].

It is widely recognized that dynamics of cell phenotype transition (CPT) is confined to low-dimensional manifolds in phase space, which are governed by the underlying gene regulatory networks [5, 9-12]. However, current manifold learning methods merely preserve topology information in the original space and are often utilized as dimension reduction techniques that project high-dimensional data to a low-dimensional embedding space [13-16]. For instance, the coordinates for projecting scalar potentials or RNA velocity vectors with t-SNE or UMAP for visualization lack biological meaning and the capacity for quantitative study. Another major category of methods for learning the underlying manifold is probabilistic embedding. This category includes techniques such as diffusion maps, variational autoencoders (VAE), and various methods derived from these approaches [17-22]. Among these methods, scVI has been developed to address tasks such as batch correction, visualization, and clustering in single-cell RNA sequencing datasets [20]. Additionally, siVAE can identify gene modules and hubs by embedding cells and features simultaneously [21]. Furthermore, an integrated pipeline for deep probabilistic analysis of single-cell datasets, known as *scvi-tools*, is now available online [22].

In mathematics, Gauss and Riemann introduced the concept of intrinsic geometry of manifold, where the Riemann metric determines the structure of manifold. Defining a Riemann metric on the low-dimensional manifold with real data is challenging. Fisher information is a quantitative measure of the information that observables carry about parameters of probability distributions [23]. Rao demonstrated that Fisher information can be utilized as a Riemann metric in

probability space [24]. Amari developed information geometry based on the Fisher metric [24, 25]. From the perspective of the master equation, the dynamics of CPT can be viewed as the evolution of a probability distribution in phase space. By converting single cell gene vectors into probability distributions, the Riemann metric can be established with Fisher information. One simple strategy for approximating the probability of single cells' gene vectors is Gaussian mixture model (GMM), where each Gaussian distribution component represents the probability density of a stable cell state [26, 27]. However, GMM is a coarse-graining approach that is incapable of yielding smooth manifolds. To overcome this limitation, we employ the Gaussian embedding method, in which each individual cell is modeled as a multivariate Gaussian distribution. This enables us to calculate the Fisher information of each cell and define the Riemann metric in a smooth and continuous manner [28, 29]. The dynamics on the manifold are also important. Information velocity, which measures the rate of change in the probability distribution, can be used to quantify the speed of cell state variation on this manifold.[30-32]. Information velocity can be computed based on the Fisher information metric and RNA velocity. Consequently, RNA velocity can be incorporated into the manifold formalism through the calculation of information velocity.

In this work, we introduce SCIM (single cell information manifolds), a novel approach to quantitatively analyze low-dimensional manifolds from single cell data. Initially, we transform each single cell's sequencing data into a multivariate Gaussian distribution, calculate the Fisher information of each cell and quantify the manifolds of CPT [28, 29]. By representing each cell as a Gaussian distribution, we can capture its uncertainty and calculate the information distance between individual cells explicitly [23, 33]. Moreover, we can assess the Fisher metric on pre-defined coordinates such as eigengenes, using the coordinate transformation with backward propagation in neural networks. By mapping the manifold onto coordinates with biological meaning, we can uncover the relationship between critical characteristics of cell dynamics and gene regulations. Unlike previous analyses, we concentrate on the invariant properties of the CPT manifolds, such as curvature and information velocity [30, 31, 34, 35]. Through incorporating information geometry with single cell geonomics analyses, SCIM provides a new framework for studying single cell transition dynamics, which can be further extended more rigorously from the perspective of geometry.

Results

Overview of SCIM

In this study, we harness neural networks to transform high-dimensional single cell data into a low-dimensional probability distribution space, represented with multivariate Gaussian distributions. To achieve this, we utilize a two-layer neural network in accordance with the Gaussian embedding method [28]. First, each single cell is represented as a vector X_i in the multidimensional expression space (Fig. 1a left), from which we compute the pairwise distance $|X_i - X_j|$ between cell i and cell j. Next, we use the k-nearest neighbor (kNN) cell graph (Fig. 1a right) to train a neural network (Fig. 1b top). The neural network ϕ, which takes individual cell gene vector as input, comprises two hidden layers and simultaneous outputs mean (μ) and standard deviation (σ) of Gaussian distribution ($\phi(X_i) = N(\mu_i, \sigma_i)$). Cell triplets randomly sampled from the kNN cell graph (Fig. 1b bottom), serve as additional input alongside the cells' gene vectors in training. We train the neural network with an energy-based learning method (Fig. 1b bottom) [36], where the energy is defined as the Fisher distance between the transformed Gaussian distributions and the loss function incorporates the graph structure information. The transformed Gaussian distribution captures the uncertainty of each cell node and reflects the diversity of its neighborhood (Fig. 1c) [28].

By defining Fisher-Rao metric on these transformed Gaussian distributions, the CPT manifolds can be quantitatively studied. The metric enables us to measure the distance on the manifold between different cells. Additionally, we can define other properties of the CPT manifolds like curvature and information velocity (Fig. 1d). Curvature, one of the most important geometric features on Riemann manifolds, is approximated with a method developed by Yann Ollivier, bypassing the need to define a connection (Methods) [35]. Information velocity, a property of dynamics, characterizes the variation rate of probability distribution and remains invariant under coordinate transformations (Methods).

In contrast to current manifold learning methods, which are tailored for visualization, our method, SCIM, is not constrained to this purpose. Although we can simply project high-dimensional cell vectors onto a 2D or 3D (μ, σ) space for visualization, it is crucial to recognize that such projections can misrepresent or distort the original manifold's geometry. This is due to the inherent limitations of lower-dimensional representations in capturing the true structure of high-dimensional manifolds. Different from previous studies that rely on visualization, SCIM

enables us to study the intrinsic geometry of the manifold quantitatively and identify invariant properties that are independent of coordinate settings.

Application of SCIM on toy datasets

To evaluate the effectiveness of SCIM, we apply it to several test datasets. The first dataset consists of data points sampled on a saddle surface (Fig. 2a top). The second dataset consists of data points sampled on the surface of a 3D ellipsoid (Fig. 2a middle). We also generate a third dataset by sampling on a four-well potential surface through simulating multiple trajectories (Fig. 2a bottom). For each dataset, we add seven-dimensional random features (Gaussian distribution with small amplitude) as noise to form a ten-dimensional data, and construct kNN graphs from the generated samples. Using SCIM, we project the embedded data of saddle surface, ellipsoid and four-well potential surface on the plane of (μ_1, μ_2), (μ_1, σ_1) and (σ_1, σ_2) respectively (Fig. 2b). For comparison, we also apply current manifold learning methods, such as t-SNE, UMAP and diffusion map, to the test datasets (Fig. S1a).

To evaluate the preservation of local metrics, we calculate ratios of geodesic lengths of multiple semi-circles sampled on a sphere surface before and after embedding, and compare the resulting ratio distributions between SCIM and other manifold learning methods (Methods). As manifold learning methods like t-SNE and UMAP primarily tend to preserve global topology for visualization purposes, their length distributions show larger coefficient of variation (CV) value comparing with that of SCIM. And diffusion maps and VAE embeddings also exhibit larger CV value in length distributions (Fig. S1b).

Curvature is a fundamental property in Riemann geometry, but its calculation requires a defined connection, which can be challenging due to the limitation of data and requirement on resolution of difference method. To circumvent this, we use coarse Ricci curvature (CRC) for estimating curvature on complex networks (Methods) [34, 35]. This is an approximation method that utilizing the relation of transportation distances between probability distributions and Ricci curvature. For the saddle surface data, data points near the saddle point have lower CRC values compared to other regions (Fig. 2c top). In the ellipsoid data, we observe that the data points at both ends of the major axis show higher CRC values (Fig. 2c middle). In the four-well potential

landscape, data points on the barriers between basins have lower CRC values, consistent with the findings in the saddle surface data (Fig. 2c bottom). Previous studies on complex networks has shown that CRC values of edges between nodes in different communities are generally negative, whereas edges inside each community usually have positive CRC values [37]. Our results, in conjunction with prior research, reveal that saddle points or the intermediate region between communities display lower CRC value. We further validate the robustness of our method by calculating CRC on saddle surface with varying parameter settings (Fig. S2).

Additionally, in the four-well potential data, points in the basins show high CRC values (Fig. 2c bottom). An intuitive explanation is that data points within a basin share similar neighbors, leading to close embedded distributions, short Wasserstein distances and high CRC values. These insights from toy datasets pave the way for identifying meta-stable cell states and critical points in CPTs of single-cell RNA sequencing data.

We applied the VAE to all the toy datasets (Fig. S3a). Except for the ellipsoid data (Fig. S3b middle), the distributions of CRC differ from those obtained with SCIM (Fig. S3b). A major problem with VAE is that the CRC results are not self-consistent. In the saddle surface data, not only do the points around the saddle points exhibit low curvature, but the entire ridge does as well (Fig. S3b top). Conversely, in the four-well potential, the barriers exhibit high curvature (Fig. S3b bottom). Additionally, the CRC value is significantly affected by the relative weight of the KL-divergence term in the loss function (Fig. S3c).

Fisher information analyses of single cell RNA sequencing data

We first apply SCIM to single-cell RNA sequencing data of dentate gyrus neurogenesis (Fig. 3a). Throughout the process of neurogenesis in dentate gyrus, radial glia-like cells differentiate through neuronal intermediate progenitor cells (nIPCs), neuroblast 1 and 2, immature granule cells, and eventually mature into granule cells [38].

For visualization, we calculate the pseudo-time of each single cell and plot the single-cell data in the probability distribution space (Fig. 3b and Fig. S4a). We also show the Fisher information

(g_{11}) of μ_1 in Fig. 3c (Other elements in Fisher information matrix are shown in supplement Fig. S4b).

The coordinates (μ, σ) lack direct biological interpretation. To elucidate the relationship between the geometry of CPT manifold and the biological characteristics of the system, it is necessary to calculate the Fisher metric in different coordinate systems, particularly those with explicit biological meaning. Here we perform a coordinate transformation to obtain the Fisher metric

$$g_{ij} = \int_X \frac{\partial logP}{\partial \theta} \frac{\partial \theta}{\partial \lambda_i} \frac{\partial logP}{\partial \theta} \frac{\partial \theta}{\partial \lambda_j} P dx$$

in the new coordinates $\{\lambda\}$ (partial derivative $\frac{\partial \theta}{\partial \lambda_i}$ calculations are detailed in Methods). We first convert the (μ, σ) coordinates to principal component (PC) coordinates. The Fisher information is a powerful tool for analyzing informative orthogonal parameters [39]. With SCIM, we can directly obtain the diagonal Fisher matrix. However, the Fisher matrices in PC coordinates may not necessarily be diagonal after coordinate transformation. If a non-diagonal element g_{ij} $(i \neq j)$ of Fisher matrix equals zero, parameters λ_i and λ_j are defined as information orthogonal [39]. By analyzing the Fisher matrix in coordinates with biological meaning, we can discern the orthogonality between these parameters. As shown in Fig. 3d, the Fisher information of PC1 and PC2 is relatively low (around 0) most of time except for a peak at pseudo-time around 0.6, coinciding with drastic changes in PC1 and PC2 values.

We also perform coordinate transformation from (μ, σ) to eigengene coordinates, where eigengene represent the PC1 of gene modules identified through gene correlation clustering (Fig. 3e) (Methods). During neurogenesis, the non-diagonal element g_{13} in the Fisher matrix of eigengenes remains nearly zero all the time. This information orthogonality between both eigengenes reflects their distinct roles at different neurogenesis stages. Similar with that of PC coordinates, this non-diagonal element g_{13} also peaks at pseudo-time around 0.6 (Fig. 3f). This consistency between two different coordinate systems reflects the underlying regulatory mechanism independent of coordinates. The Fisher information and information orthogonality are dynamical as the Fisher matrix evolves over time. Through Fisher information analysis, we gain diverse perspectives on the dynamics of CPT process. It is important to note that Fisher

information is affected by the choice of coordinates, for instance, the Fisher information of two gene modules' PC2 behave differently from that of eigengenes (Fig. S4c).

Fisher information reflects the sensitivity of probability density to control parameters. Stiff parameters are those with large values of Fisher information, while sloppy parameters have small values (Fig. 3g) [40]. By calculating the Fisher information for each gene (Methods), we can identify the stiff and sloppy genes in the context of CPT. Our analyses reveal that the stiffness and sloppiness of genes vary throughout the CPT process, suggesting that different genes may play different roles at different stages of CPT (Fig. 3h). Specifically, the Fisher information of transcriptional factors reveals their probable regulatory functions in CPT. For instance, the Fisher information of Neurod2 against pseudo-time exhibits a peak around a pseudo-time of 0.6 (Fig. 3i & 3j), which aligns with the trends of non-diagonal Fisher information in PC coordinates (Fig. 3d) and eigengenes (Fig. 3f). In contrast, the Fisher information of Bcl11b and Lhx2 behaves differently (Fig. 3i). Therefore, Fisher information analysis provides a potential method for identifying the roles of different genes in regulating CPT.

We compared the embeddings of SCIM and scVI on dentate gyrus neurogenesis (Fig. S5). As shown in Fig. S5b, the scVI embeddings form spherical clouds around the origin in the coordinate system due to the regularization in VAE, which minimizes its Kullback-Leibler divergence (KL-divergence) with the standard normal distribution $\mathcal{N}(0, I)$ [18]. And the latent embedding in VAE is designed to facilitate generative sampling for the decoder. In contrast, SCIM focuses on preserving local metrics and provides a framework for quantifying geometric properties of the underlying low-dimensional manifolds.

We also apply Fisher information analysis on two additional datasets. One is the development of pancreatic endocrine cells (Fig. 4a) [41]. During embryonic development Ngn3-low progenitors first transform into Ngn3-high precursors, then Fev-high (Fev+) cells. The latter further develop into endocrine cells including glucagon producing α-cells and β-cells. Here we concentrate on the α-cell branch. The variation of Fisher information of μ_1 is shown in Fig. 4b. And the Fisher information of the two eigengenes (non-diagonal element) exhibits one local minimum value and one local maximum value as CPT proceeds (Fig. 4c). And both local extrema are deviated from zero, indicating the complex interplay between eigengenes and their synergetic influences on the

probability distributions vary as CPT proceeds. The variation in Fisher information for multiple genes during pancreatic endocrinogenesis reveals the switching of roles as stiff or sloppy genes throughout this CPT process (Fig. 4d). Additionally, the changes in Fisher information for transcriptional factors may indicate the dynamics of gene regulation at different stages (Fig. 4e). For example, the Fisher information and expression levels of Neurog3 are plotted against pseudo-time (Fig. 4f).

The other is an induced in-vitro transition process, the epithelial-to-mesenchymal transition (EMT) of human A549 cells treated with TGF-β for different durations (0 day, 8 hours, 1 day, and 3 days) (Fig. 4g) [42]. At the initial stage of EMT (0 day), most A549 cells are in the epithelial state. And most cells transit into the mesenchymal state after three days of treatment. Comparing with the development datasets, cancer cells exhibit greater heterogeneity, reflected in the significant Fisher information variation of μ_1 in the 8h sample (Fig. 4h). The non-diagonal element of Fisher information matrix between two PCs is non-zero initially (mostly 0 day). And the two PCs become information-orthogonal after TGF-β treatment (Fig. 4i). Due to cell state heterogeneity, we calculate pseudo-time for a clearer dynamic representation. The changes in Fisher information for multiple genes (Fig. 4j) and transcriptional factors (Fig. 4k) reveal variations in control during the EMT process. We also presented the Fisher information and expression levels of FOXP1 across pseudo-time (Fig. 4l). It appears that there is no clear correlation between the magnitude of Fisher information and gene expression (Fig. 3j, 4f, and 4l).

Identification of critical points with curvature

Analyzing Fisher information in specific coordinates can shed light on the regulatory dynamics in CPT. However, the variation of Fisher matrix is dependent on the coordinate system. To learn more about the geometry of the information manifold, we further computed CRC on the single-cell RNA sequencing datasets (Fig. 5a) (Methods).

In the dentate gyrus data, cells with the lowest and highest pseudo-time values correspond to radial glia cells and granule cells respectively and these cells displayed high CRC values. As demonstrated in Fig. 2d, high-CRC value characterizes the basins on the energy landscape, indicating the two types of cells are in stable states. Following our finds on toy datasets, cells in this low-CRC region are posited to be around the saddle point, a critical point in dynamical

system theory that connects different attractors or basins. And we do observe a low-CRC region (Fig. 5b box ii) bridges the two high-CRC regions (Fig. 5b box i and iii), suggesting a transition from radial glia cells to granule cells. Therefore, this low-CRC region is associated with the intermediate or transition states which often display mixed properties of different cell phenotypes and are typically unstable and revisable. Characterization of transition states is crucial for understanding the control of cell fate decision [43-45]. In addition, the Ricci curvature should be invariant under coordinate transformation. Here we use SCIM to transform the single cell data of dentate gyrus to another Gaussian embedding coordinates (μ', σ') (Fig.5c). We find that the CRC distribution is similar with that of (μ, σ) in Fig. 5b and one can still observe a low-CRC region (Fig. 5c box ii) bridges the two high-CRC regions (Fig. 5c box i and iii).

In the pancreatic endocriogenesis data, two low-CRC regions are identified, bridging three high-CRC regions (Fig. 5d left). And there is a subset of Ngn3-high precursors in the meta-stable state (high CRC region in the middle) between the two transitory regions with low-CRC values, suggesting a step-like transition process in endocriogenesis. For the EMT dataset, curvature analysis reveals that cells treated with TGF- β for one day have the lowest CRC value, connecting two high-CRC regions that are the epithelial and mesenchymal states (Fig. 5d right). The intermediate state is commonly recognized as a partial EMT state which is related to the collective migration of tumors [46, 47].

For the dentate gyrus neurogenesis data, we analyzed genes that are highly correlated with CRC values in the marked regions (boxes in Fig. 5b&5c), i.e. the basins and critical point. These genes serve as markers of the corresponding cell regions. And they correspond to three eigengenes respectively (except a few genes in region ii belong to the first eigengene), highlighting the modular regulation in CPT (Fig. 5e). We further examined the biological functions of the eigengenes by performing gene ontology (GO) analysis on dentate gyrus neurogenesis data (Fig. 5f) [48]. The GO terms highly associated with the eigengenes are listed. For example, the third eigengene (EG3) is closely associated with neural functions.

It is important to note that CRC is an approximation of Ricci curvature and is not derived from Fisher metric. Nevertheless, our analyses reveal its association with the metric. The low-CRC

region coincides with the peaks of non-diagonal element of Fisher information matrix (Fig. 3d & 3f).

Geometry guides single cell dynamics on CPT manifolds

CRC elucidates the structure of the CPT manifold, with low-CRC regions closely related to saddle points between basins and its derivation does not require dynamics information. In the spirit of general relativity theory, where *spacetime tells matter how to move*, we hypothesize that the geometry of CPT manifolds can similarly guide the dynamics of single cells in the transitions.

To extract dynamics information from snapshot sequencing data, Gioele La Manno and his collaborators proposed the RNA velocity method [4]. The RNA velocity method assumes simple dynamic balance between spliced RNA and unspliced RNA and estimates the time derivative of spliced RNA counts. This method has been further developed by several labs [5-7]. By integrating RNA velocity with Fisher metric, we can compute the information velocity of each individual cell in different coordinate systems (Fig. 6a) (Methods) [30-32], which quantifies the cell dynamics on the information manifolds.

Analyses of various CPT processes revealed that cells in the low-CRC region display high information velocity values (Fig. 6b box), indicating a correlation between high information velocity and low CRC value (Fig. 6b). As the low-CRC region is around the saddle point, probability distributions on the opposite sides of the saddle point are distinct. Transitions between the two distributions are drastic, resulting in high information velocity values in this region. Conversely, low information velocity values in the basins indicates that the distribution shows little variation and cells are in stable states. These findings reflect that the geometry of underlying manifold guides the dynamics of single cells during CPT. Robustness checks using different parameters for CRC and information velocity calculations confirmed this conclusion (Fig. S6&S7). We also applied an alternative gene selection method named *DUBstepR* to preprocess dentate gyrus data [49], yielding similar results of CRC and information velocity (Fig. S8) with that in Fig. 6d.

Additionally, we performed coordinate transformations and calculate information velocity in different coordinate systems. Using non-negative matrix factorization (NMF) for dimension reduction, we derived the Fisher metric in this new coordinate system. The velocities of each NMF dimension can be obtained by the linear combination of different velocity genes' RNA velocities. We found that the distribution of information velocity is similar to that in the (μ, σ) coordinates (Fig. 6c and Fig. S9). From the definition of information velocity, it is evident that it is a value that is independent of coordinates. Our analyses confirm its invariance under coordinate transformation, although choice of coordinates can affect the absolute value of information velocity which is probably due to limitation of data and algorithm in practice. However, high information velocity value is not necessarily related to the high absolute value of RNA velocity. As shown by the distributions of norm of the RNA velocity in different datasets (Fig. S10), there is no direct correlation between them with that of the information velocity.

Eigengenes (EG) represent control modules in the gene regulation network, and switching between these modules is critical in cell state transition. Plotting eigengene values across the cells reveals a pattern that eigengenes form a relay in the dentate gyrus neurogenesis (Fig. 6e). Especially, the increase of EG3 is right after the decrease of EG1 and EG2. Furthermore, the low-CRC region or high information velocity region (Fig. 6d) coincides with the transitory region between eigengenes (Fig. 6e), suggesting the interaction between eigengenes or gene modules probably determine the geometry of CPT information manifolds. The dynamic pattens of CPT emerge from the complex interactions between genes. Specifically, the saddle region, characterized by low-CRC values and high information velocity, probably involves more gene interactions even global reprogramming of gene profiles [43, 50].

Characterizing branching points with Fisher information

In this work, we primarily analyzed the transition of cells from one phenotype to another. However, stem cells often differentiate into multiple cell types, leading to complex critical transitions and manifolds. Identifying these branch points is crucial for lineage reconstruction in embryo development. We also tested our method on datasets with different branching structures. First, we applied it to a toy dataset with two branches (Fig. 7a). The CRC value at the branch points was low (Fig. 7b). Moreover, we performed eigen decomposition on the Fisher

information matrix and found that the ratio between the second largest eigenvalue and the largest eigenvalue increases in cells around the branch points (Fig. 7c). Large eigenvalues of the Fisher information matrix correspond to stiff parameter combinations that affect the embedded distributions, i.e., the cell states. The increase in the relative ratio of the second largest eigenvalue to the largest eigenvalue indicates that there are two stiff parameters with similar effects on the cell states. In other words, there are two parameters that can significantly regulate the cell states. When cells transition from the trunk to different branches of states, the increase in stiff parameters likely indicates an increase of stiff direction in dynamics space (Fig. 3g) due to the activation of different functional gene modules.

During development, pancreatic endocrine cells differentiate into various branches, including glucagon-producing α-cells and β-cells (Fig. 7d). We observed that the distribution of CRC is essentially the same as in Fig. 5d when the dataset includes only the α-cell branch. The two low-CRC regions correspond to transitional states (Fig. 7e). However, CRC alone cannot differentiate between the two transitional states, as the latter represents a branch point.

To address this, we analyzed the ratio of the second largest eigenvalue to the largest eigenvalue and found that only the latter transitional region exhibited an increase in this ratio (Fig. 7f). This indicates that the second largest eigenvalue becomes more significant relative to the largest eigenvalue at the branch point, suggesting the presence of two stiff parameters that significantly influence cell states. While information velocity cannot directly identify branch points, we projected the information velocity onto the coordinates of different eigengenes corresponding to different branches (Fig. S11). This projection helps distinguish the branch points by highlighting changes in the underlying gene regulatory networks.

Discussion

The exploration of underlying manifolds of CPT has recently emerged as a novel research field. In this study, we have developed a method based on information geometry to investigate the intrinsic geometric properties of the CPT manifolds. Our approach, the Single-Cell Information Manifold (SCIM) method, diverges from traditional manifold learning methods that focus solely on low-dimensional visualization. Instead, SCIM examines the CPT manifold through its

intrinsic geometry, independent of visualization techniques. It transforms the single cell gene vectors into multivariate Gaussian distributions, allowing for the computation of Fisher metric and CRC of information manifold in probability space. We discover that the geometric properties of manifold, such as CRC, correlates with dynamical properties like information velocity in CPT. Moreover, the CRC and information velocity are invariant under coordinate transformation.

When calculating the Fisher metric, constructing a cell graph from experimental data is necessary. However, sampling biases can affect the accuracy of this step. For instance, the local cell density variations probably alter the neighbor diversity in kNN graph, potentially impacting the computation of the Fisher metric. Although we do not find direct correlation between CRC and local cell density, the sampling bias should be taken into consideration, particularly when dealing with time-course datasets like the EMT dataset of A549 cells. Moreover, as experimental techniques constrain sampling, extrapolation the analyses into the regions where there are no sampled cells is not feasible. Our focus is on leveraging local neighborhood information to investigate the properties of single cells.

The intrinsic geometry of the CPT manifold is related to the intrinsic dimensionality, the estimation of which is still an open question [51]. An accurate intrinsic dimension can increase the training efficiency and accuracy of neural networks, while an incomplete coordinate system may distort the calculation of the Fisher metric. We vary the number of latent dimensions to ensure that the transformation is robust (Fig. S6). But the methodology for estimation of intrinsic dimension and its reliability will be further explored.

Duluxan Sritharan and colleagues proposed a method to calculate Riemann curvature with extrinsic differential geometry in single cell RNA sequencing data [52]. Different from the extrinsic geometry method, SCIM is rooted in intrinsic geometry which is invariant under coordinate transformation. It allows us to study the system by transforming into different biological meaningful coordinates, such as eigengenes.

Another prevalent technique for embedding high-dimensional data with probability distribution is diffusion map [53], which employs kernel based method for constructing Markov transition matrix between neighboring data points. PHATE further develop this approach with information-

theoretic potential distance for improved visualization of high-dimensional biological data [17]. Oluwadamilola Fasina and colleagues developed neural FIM with Markovian diffusion operator based on PHATE for computing Fisher information metric [54]. While neural FIM represents data points with diffusion probability distributions and defines Fisher metric on the original feature space, SCIM captures the uncertainty of individual cells with the transformed Gaussian distributions and learns Fisher metric in various coordinate systems through transformation.

Nowadays, RNA velocity has become a potent tool for learning the dynamics of CPT. However, it relies on some strong assumptions and suffers from some flaws when applied to different datasets [55]. Considering that the geometry of the CPT manifold is related to the cells' dynamics, this method can be further refined to regularize the inferred dynamics with geometric properties. For instance, the fact that information length computed with information velocity and with local Fisher metric should be consistent can be used as constrain on inferring RNA velocity.

Study of critical points in complex systems is still an active research field [45, 56-58]. Information geometry, along with Fisher information, offers a powerful means to characterize phase transitions and can be readily adapted to non-equilibrium systems because it does not require defining Hamiltonian or ensemble function [59, 60]. The peaks observed in the non-diagonal elements of the Fisher information matrix are similar with the divergence of order parameter in phase transition, which indicates CPT probably share similar characteristics with phase transition (Fig. 3d and f) [59]. Therefore, SCIM not only offers a framework for analyzing single cell transitions, but may also reveal underlying principles of transitions in a broader spectrum of complex systems [60, 61].

The manifolds may evolve due to intricate interactions among cells or external signals in complex biological processes. In all the analyses in this work, the parameter systems do not explicitly contain time. However, some parameters implicitly contain information about the direction of time. For example, the EMT dataset is collected from different time points and clearly shows that μ_1 is correlated with time (Fig. 4h). And the parameters that control bifurcation like X_2 in Fig.7 can be also time dependent [43, 45]. Currently, we use all the data to learn the geometry only in spatial coordinates. The exploration of pseudo-Riemannian geometry, which incorporates both time and space, will be pursued in future studies. Quantitative

descriptions and analyses of the geometry of evolving manifolds can help us characterize different stages in complex developmental processes.

Figures

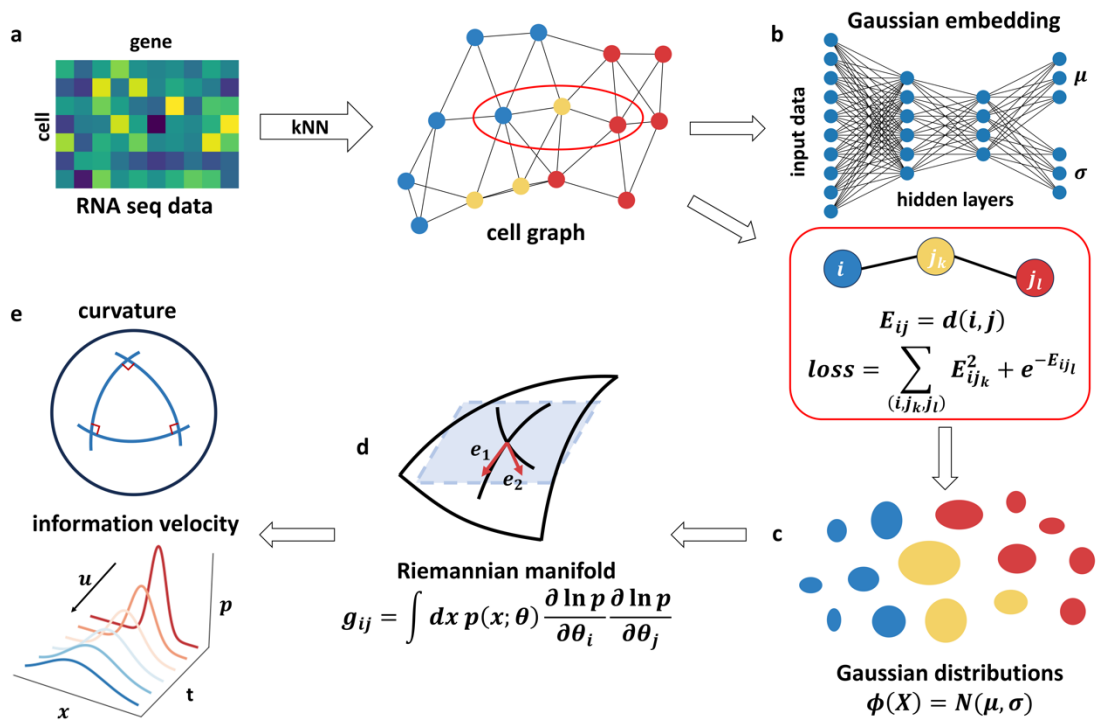

Figure 1 Illustrative overview of the SCIM framework.

(a) Construction of kNN cell graph with single cell RNA sequencing data. Each single cell is depicted as a multi-dimensional vector, representing the counts of different genes' RNA (left). The Euclidean distance between cells are calculated to build the graph. Colors represent different cell phenotypes.

(b) The neural network architecture of the Gaussian embedding and its training method. The neural network is a multi-layer perceptron with two hidden layers. And it outputs μ and σ of Gaussian distribution respectively. Training is conducted using energy-based loss functions on sampled triplets.

(c) Representation of single cells as multivariate Gaussian distributions. The transformed distributions are visualized as ellipses. And positions and sizes of the ellipses represent μ and σ respectively.

(d) Defining Fisher metric on transformed Gaussian distributions give rise to information manifolds of CPT.

(e) The geometric property of CPT manifolds such as curvature and dynamics on the manifolds like information velocity are computed with SCIM.

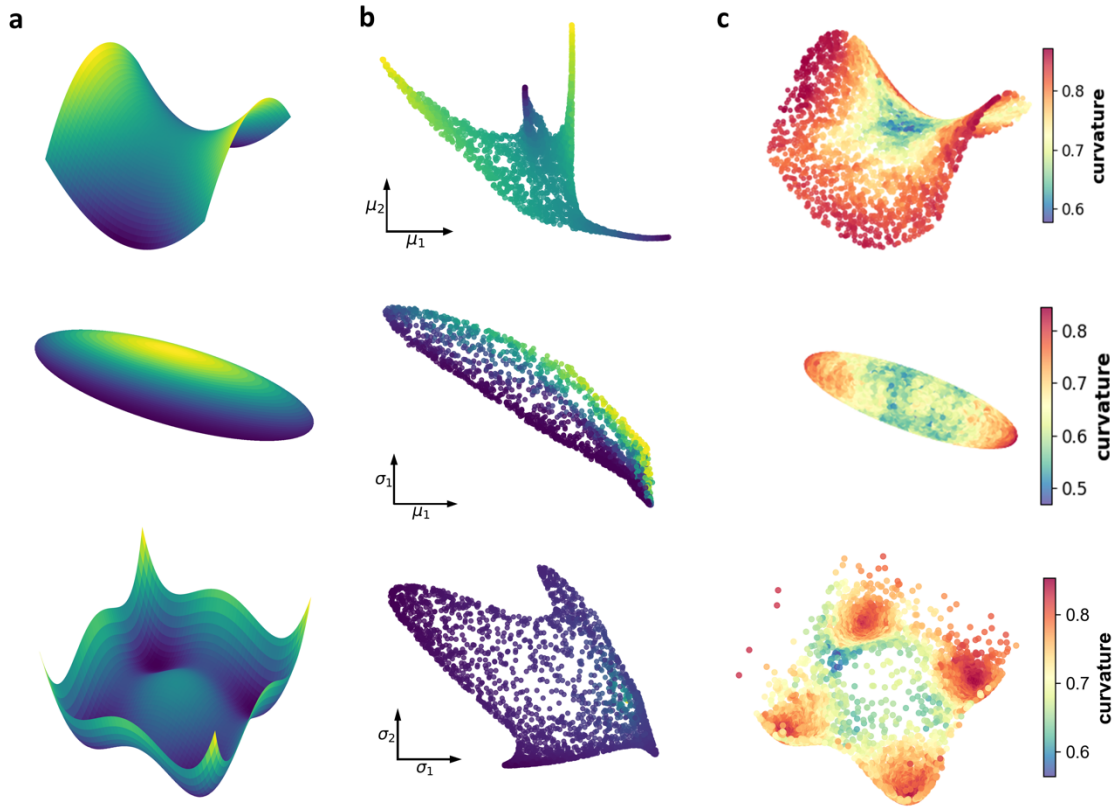

Figure 2 Evaluation of SCIM on toy models including saddle surface, ellipsoid and four-well potential.

(a) Sampling of data points on various surfaces: saddle surface (top), ellipsoid surface (middle), and simulated trajectories on a four-well potential landscape (bottom).

(b) Planar projections of saddle surface on the plane of (μ_1, μ_2) (top), ellipsoid on the plane of (μ_1, σ_1) (middle) and four-well potential on the plane of (σ_1, σ_2) (bottom) respectively.

(c) CRC distributions across the sampled data points in saddle surface (top), ellipsoid surface (middle) and four-well potential (bottom). Color represents the value of CRC.

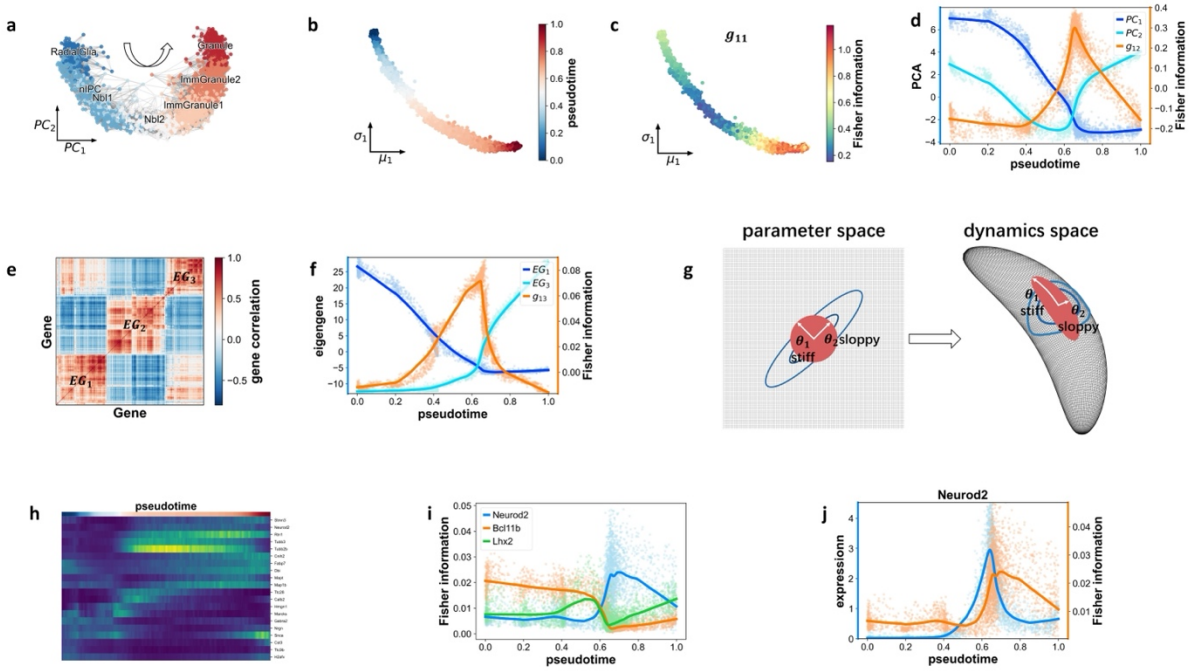

Figure 3 Fisher information analyses of dentate gyrus neurogenesis.

(a) KNN graph of dentate gyrus neurogenesis. Color represents cell type. The direction of transition denoted by the arrow is from radial glia cells to granule cells.

(b) Pseudo-time analysis of dentate gyrus neurogenesis (projection on the plane of (μ_1, σ_1)). Color represents pseudo-time.

(c) Variation of g_{11} (Fisher information of μ_1) in dentate gyrus neurogenesis. Color represents value of Fisher information.

(d) Variation of one non-diagonal element in Fisher information matrix (g_{12}) on PC coordinates (PC1-PC2) and values of PC1 and PC2 (with respect to pseudo-time). This non-diagonal element exhibits a peak during the transition process. The lines are fitted with LOWESS (Methods).

(e) Eigengenes (EG) derived with clustering on gene correlations. Color represents correlations among different genes. The value of eigengene is defined as PC1 of each cluster.

(f) Variation of one non-diagonal element in Fisher information matrix (g_{12}) on eigengene coordinates (EG1-EG2) and the corresponding values of eigengenes (with respect to

pseudo-time). This non-diagonal element also exhibits a peak during the transition process and position of the peak is close to that in PC coordinates in panel (e). The lines are fitted with LOWESS (Methods).

(g) Schematic illustration of stiff and sloppy parameters. The system's dynamics are determined by the parameters, shown as a mapping from the parameter space to the dynamics space. Variations in stiff parameters significantly impact the system's dynamics, while the influence of sloppy parameters is negligible (red discs). Given some samples from the dynamics space, stiff parameters can be inferred with small uncertainties, whereas sloppy parameters are poorly determined (blue circles).

(h) Heatmap of Fisher information for multiple genes with respect to pseudo-time. The variation in Fisher information reflects changes in the susceptibilities of different genes during the CPT process.

(i) Fisher information of several transcriptional factors, including Neurod2, Bcl11b, and Lhx2, in each single cell with respect to pseudo-time. The variation in Fisher information reflects the different regulatory functions of these transcriptional factors during the CPT process. Each dot represents a single cell, and the lines are fitted using LOWESS (Methods).

(j) Fisher information (orange) and expression value (blue) of Neurod2 in each single cell with respect to pseudo-time. Each dot represents a single cell, and the lines are fitted using LOWESS (Methods).

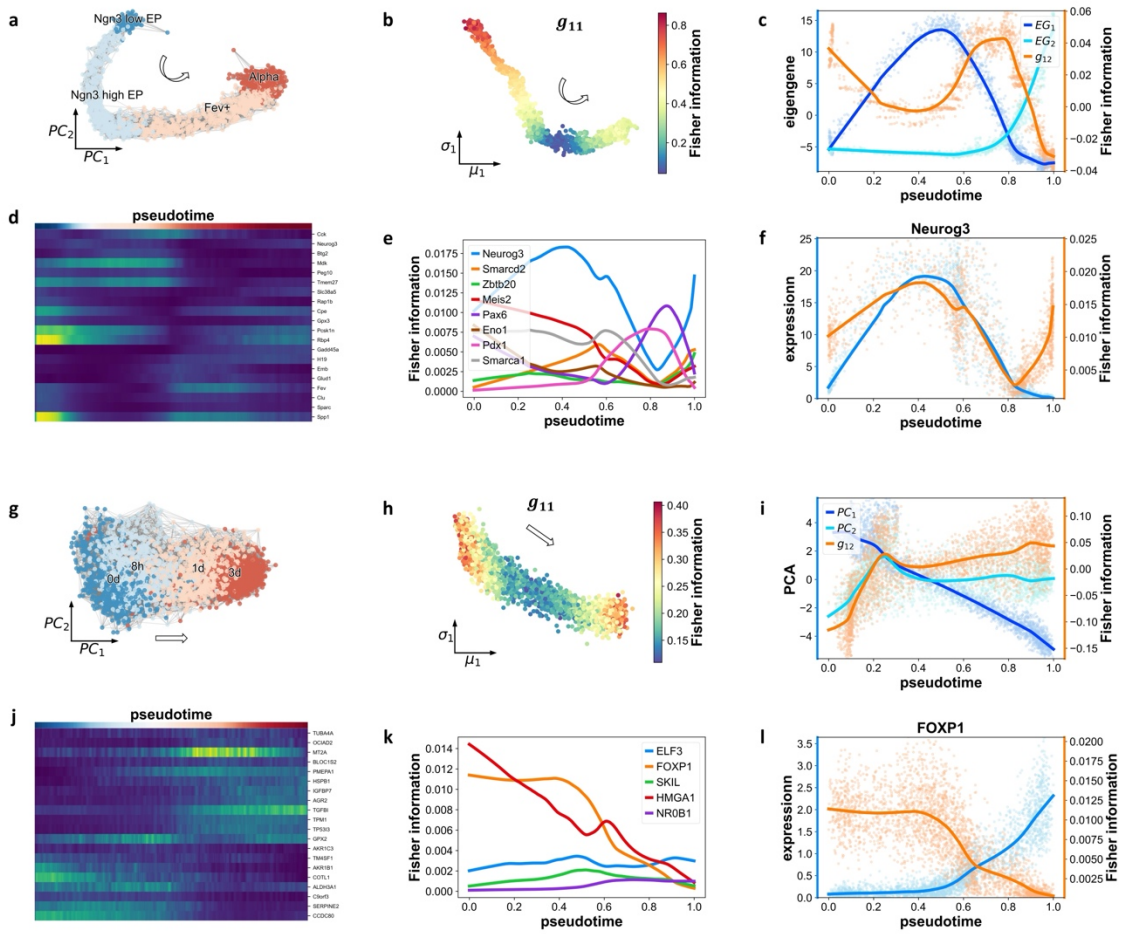

Figure 4 Fisher information analyses of pancreatic endocrinogenesis and EMT of A549 cells induced by TGF-β.

(a) kNN graph of pancreatic endocrinogenesis. Color represents cell type.

(b) Variation Fisher information of μ_1 in pancreatic endocrinogenesis. Color represents value of Fisher information.

(c) Variation of one non-diagonal element in Fisher information matrix (g_{12}) on eigengene coordinates and the corresponding values of eigengenes (with respect to pseudo-time) in pancreatic endocrinogenesis. The lines are fitted with LOWESS (Methods).

(d) Heatmap of Fisher information for multiple genes with respect to pseudo-time. The variation in Fisher information reflects changes in the susceptibilities of different genes during pancreatic endocrinogenesis.

(e) Fisher information of various transcriptional factors in each single cell with respect to pseudo-time. The lines are fitted to single-cell data using LOWESS (Methods).

(f) Fisher information (orange) and expression value (blue) of Neurog3 in each single cell with respect to pseudo-time. Each dot represents a single cell, and the lines are fitted using LOWESS (Methods).

(g) Same as in panel (a), except for EMT of A549 cells.

(h) Same as in panel (b), except for EMT of A549 cells.

(i) Same as in panel (c), except for Fisher information of PC1-PC2 in EMT of A549 cells.

(j) Same as in panel (d), except for EMT of A549 cells.

(k) Same as in panel (e), except for EMT of A549 cells.

(l) Same as in panel (f), except for FOXP1 in EMT of A549 cells.

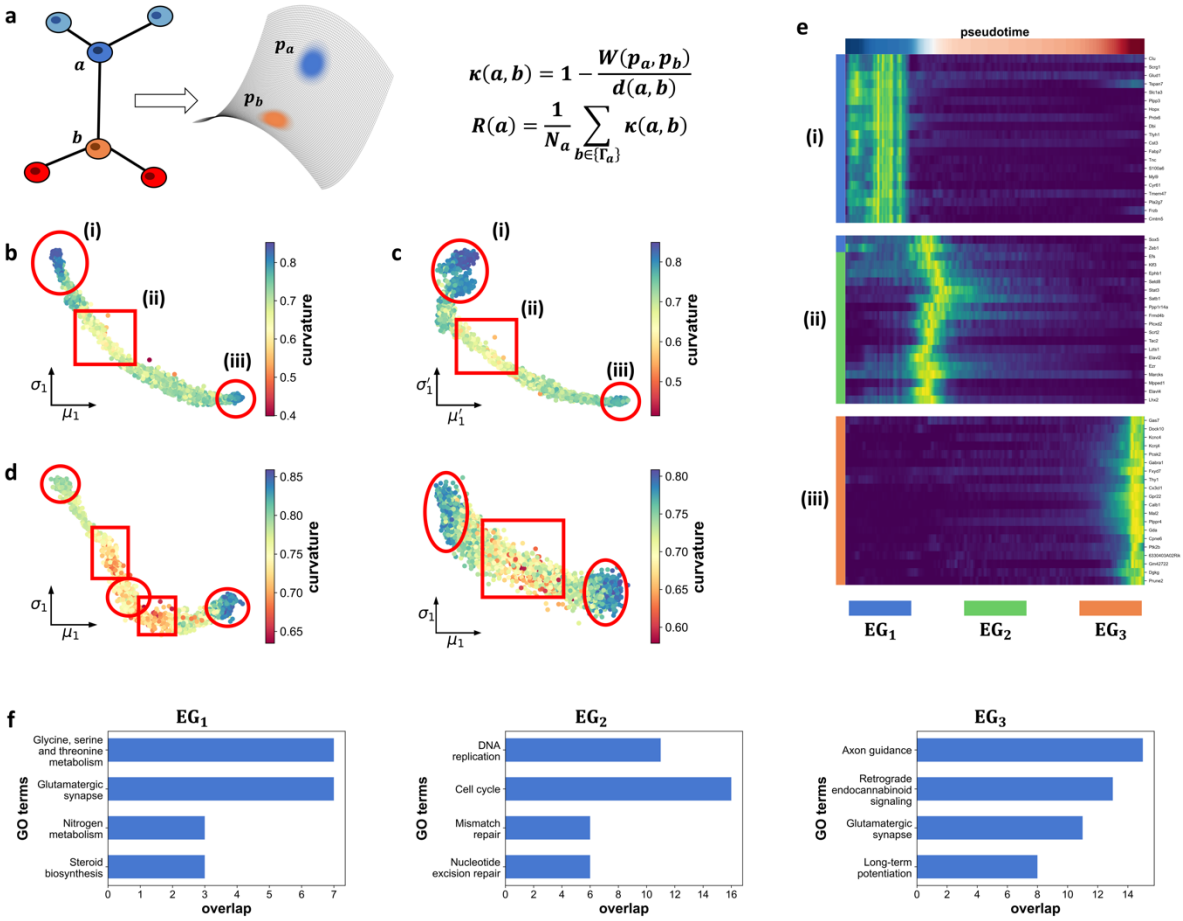

Figure 5 Identification of critical points and basins in different CPT processes with CRC.

(a) Schematic representation of CRC calculation from single cell RNA sequencing data. Single cell gene vectors are transformed into Gaussian distributions. The ratio between Euclidean distance of cells and Wasserstein distance of transformed distributions is calculated to derive CRC. In the equation, $W(P_a, P_b)$ represents the Wasserstein distance between the Gaussian embeddings of cell a and cell b and $d(a, b)$ denotes the Euclidean distance between cell a and cell b. $\kappa(a, b)$ is the CRC of the edge between cell a and cell b, and $R(a)$ is the CRC of cell a. $\{\Gamma_a\}$ represents all the neighboring cells of cell a and N_a is the number of neighbors of cell a.

(b) Identification of the critical points and basins with CRC in dentate gyrus neurogenesis. Regions where cells have local extreme (maximum and minimum) values of CRC are enclosed in red box. High CRC values are found in regions (box i and iii) corresponding to radial glia cells and granule cells, whereas the intermediate region exhibits low CRC values (box ii) that corresponds to saddle node. The color bar is reversed in this panel for better visualization.

(c) Invariance of CRC in different Gaussian embedding coordinates. The distribution of CRC calculated within a different Gaussian embedding coordinate (μ', σ') is similar to that within (μ, σ) in panel (b), which reflects the invariant properties of curvature under coordinate transformation. The color bar is reversed in this panel for better visualization.

(d) Identification of the critical point and basins with CRC in pancreatic endocrinogenesis (left) and EMT of A549 cells induced by TGF-β (right). In the pancreatic endocrinogenesis, besides the high-CRC regions at the initial and final stages in transition, an additional high-CRC region is detected in the middle. Two regions with low CRC values (rectangles) bridge the three high-CRC regions (circles). In the EMT dataset, the low CRC cells are mainly found in the intermediate region (rectangles) between the 8-hour and 1-day samples. The color bar is reversed in this panel for better visualization.

(e) Heatmap of genes correlated with CRC values. Genes correlated with CRC values are highlighted within the red boxes labeled i, ii, and iii. Adjacent color bars on the left side of the heatmap denote the eigengenes associated with each gene in the row.

(f) Gene ontology (GO) enrichment analysis of different eigengenes in dentate gyrus neurogenesis. The various GO terms reveal the association between geometric properties and biological function modules.

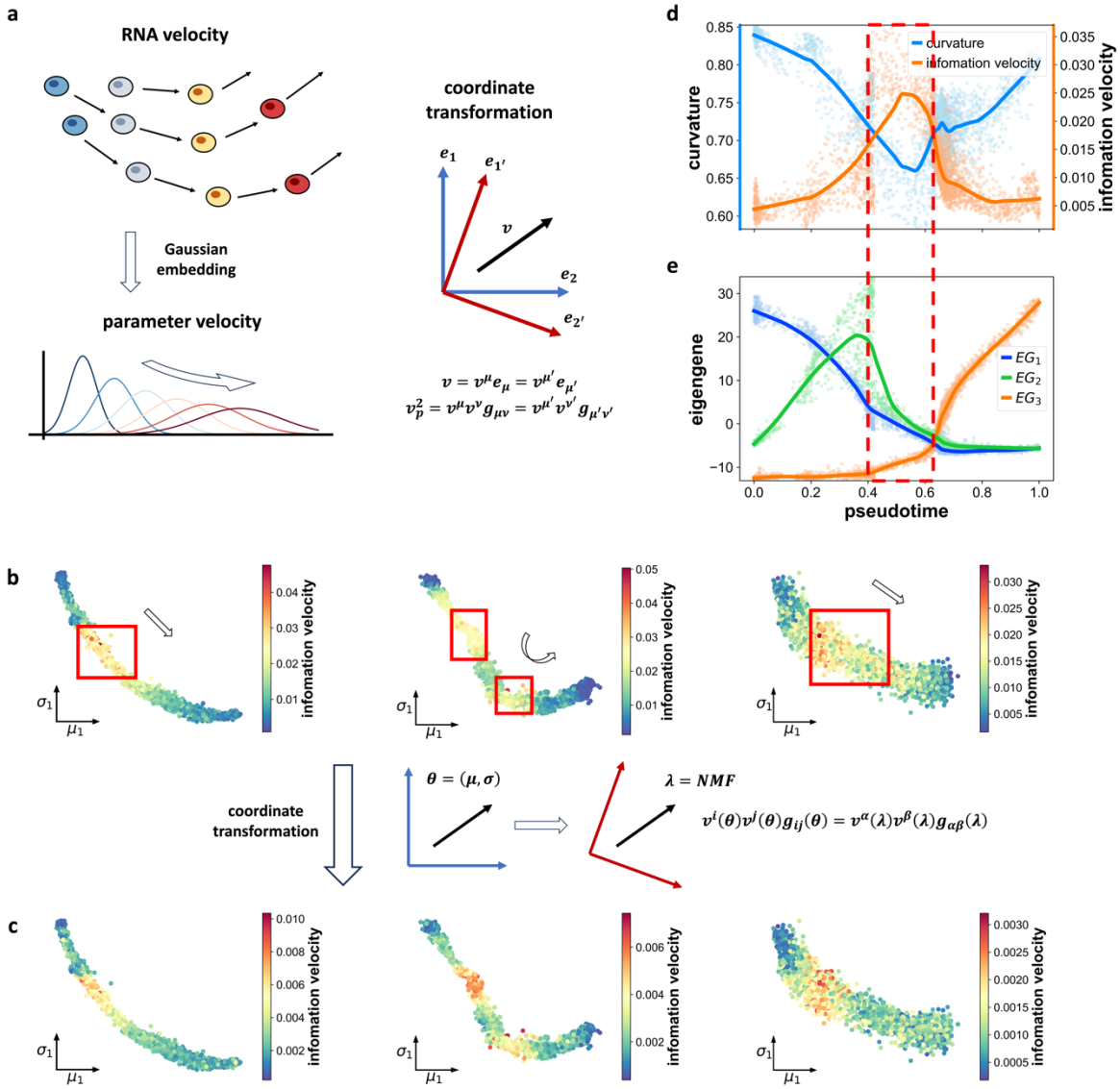

Figure 6 Information velocity analyses in different CPT processes with SCIM.

(a) Flowchart of information velocity calculation using RNA velocity method. The RNA velocity is inferred from the dynamic equilibrium between unspliced mRNA and spliced mRNA.

Information velocity is then calculated with the velocities and Fisher metrics. It can be calculated in different coordinates and is invariant under coordinate transformation. (e_1, e_2) and (e_1', e_2') denote different coordinates. $\boldsymbol{v}$ is the parameter velocity and $\boldsymbol{v_p}$ denotes the information velocity.

(b) Variation of information velocity calculated within (μ, σ) coordinates in dentate gyrus neurogenesis (left), pancreatic endocrinogenesis (middle), and EMT of A549 cells induced by TGF-β (right). Cells with low CRC values mainly coincide with those exhibiting high information velocity values (boxes).

(c) Variation of information velocity calculated within NMF coordinates in dentate gyrus neurogenesis (left), pancreatic endocrinogenesis (middle), and EMT of A549 cells induced by TGF-β (right). The similar distributions with that in panel (b) reflects the invariance of information velocity under coordinate transformation.

(d) Association of low-CRC value and high information velocity in dentate gyrus neurogenesis (with respect to pseudo-time). The lines are fitted with LOWESS (Methods).
(e) Association between manifold geometry and gene regulation networks (switch of eigengenes). Comparing with the minimum point of CRC and the peak of information velocity in panel (c), all three eigengenes show significant variations in slopes at this critical point (dashed box). The lines are fitted with LOWESS (Methods).

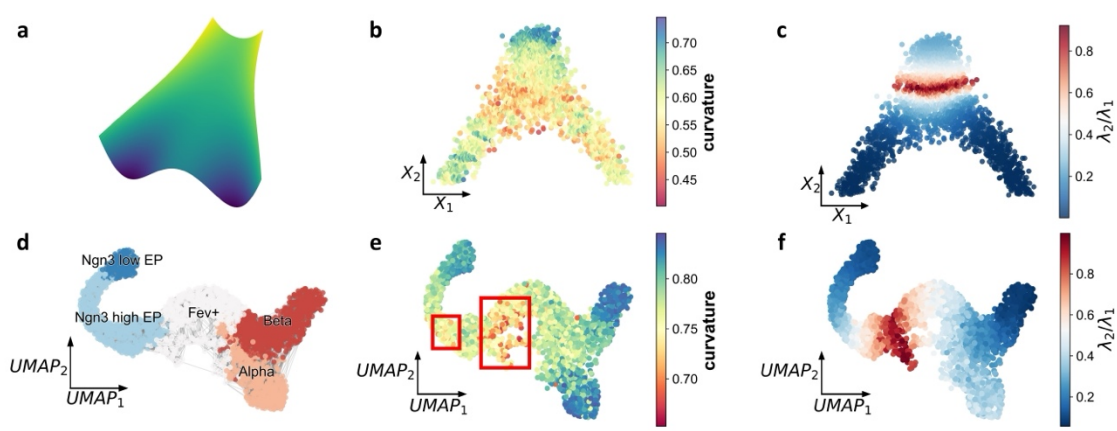

Figure 7 Application of SCIM on data with multiple branching structures.

(a) Synthetic data generated through simulations on landscapes with multiple branching structures.

(b) CRC distributions across synthetic data points.

(c) Ratio of the second largest eigenvalue to the largest eigenvalue of the Fisher information matrix in each cell. Cells near the branch points show high ratios, indicating increased number of stiff parameters at these points.

(d) Cell graph of endocrinogenesis data including both α and β cells.

(e) CRC distribution reveals two transition regions in this process. The CRC analysis identifies two distinct transition regions, consistent with previous analyses on single branches.

(f) Ratio of the second largest eigenvalue to the largest eigenvalue of the Fisher information matrix in each cell. Similar to the synthetic data, cells in the transition region around the branch point exhibit high ratios, while cells in the other transition region do not.

Methods

1 Gaussian embedding of single cell RNA sequencing

Single cell RNA sequencing is a powerful technique that measures abundance of mRNA in individual cells, enabling a detailed understanding of cell heterogeneity. Each cell is characterized as a point in a high-dimensional gene space. We construct a k-nearest neighbor (kNN) graph by computing pair-wise Euclidean distances between cells in this space (Fig. 1a). The number of neighbors used in kNN cell graph is denoted with k_{nei}.

To represent each cell as a probability distribution, we employ the Gaussian embedding method, a technique previously applied in natural language processing [29]. The SCIM framework utilizes neural networks to learn the Gaussian embedding of each cell based on its high-dimensional gene vectors and its proximity to neighboring cells in the gene space R^D (D is the original high-dimensional gene expression space) (Fig. 1b) [28]. Each cell is transformed into a multivariate Gaussian distribution in an L-dimensional space R^L (Fig. 1c), where L is significantly lower than D. For simplicity, the covariance matrices of transformed Gaussian distributions are diagonal.

In this study, the neural network is trained with an energy-based learning method. The energy-based model is a method of capturing dependencies between variables through mapping each configuration of the variables to a scalar energy. And the loss function is used to evaluate the qualities of different energy functions. During training, triplets (i, j_k, j_l) as illustrated in Fig. 1b are sampled uniformly from the cell graph. The individual cells' gene vectors are used as inputs in the neural network. The loss function is defined as

$$Loss = \sum_{i} \sum_{k<l} \sum_{j_k \in N_{ik}} \sum_{j_l \in N_{il}} (E_{ij_k}^2 + exp(-E_{ij_l}^2)),$$

where N_{ik} represents the set of nodes j_k that the length of shortest paths on the graph between node i and node j_k equal k. The maximal k considered is denoted by k_{hop}. The energy E_{ij} represents the distance between nodes i and j. By minimizing this loss, the E_{ij_k} in each term is minimized as much as possible while E_{ij_l} is maximized.

In the original work, the energy function is Kullback-Leibler divergence $E_{ij} = D_{KL}(P_i \| P_j)$ [28].

In our approach, we use the Fisher distance as the energy function,

$$E_{ij} = \sqrt{2\sum_{d=1}^{L}\left(\ln\frac{\left|\left(\frac{\mu_{id}}{\sqrt{2}},\sigma_{id}\right)-\left(\frac{\mu_{jd}}{\sqrt{2}},-\sigma_{jd}\right)\right|+\left|\left(\frac{\mu_{id}}{\sqrt{2}},\sigma_{id}\right)-\left(\frac{\mu_{jd}}{\sqrt{2}},\sigma_{jd}\right)\right|}{\left|\left(\frac{\mu_{id}}{\sqrt{2}},\sigma_{id}\right)-\left(\frac{\mu_{jd}}{\sqrt{2}},-\sigma_{jd}\right)\right|-\left|\left(\frac{\mu_{id}}{\sqrt{2}},\sigma_{id}\right)-\left(\frac{\mu_{jd}}{\sqrt{2}},\sigma_{jd}\right)\right|}\right)^2}$$

, where (μ_{id}, σ_{id})represents the coordinate of the d_{th} component of the i_{th} Gaussian distribution. In this way the geometric structure of data points represented by the cell graph is preserved into the embedded Gaussian distributions. Once the neural network is trained, it takes single cell's gene vector as input and transforms this vector into a Gaussian distribution in application. Typically, we take $k_{nei} = 10,\ L = 10,\ k_{hop} = 2$ for experimental data. We also test other parameters (Fig. S6&S7), and the result is robust to the change of the parameters.

In contrast, the loss function of VAE is $loss = -E_z[p(x|z)] + \beta KL\big(p_\theta(z|x)||q(z)\big)$, where x represents the input data, z is the variable of the embedded distribution and β is the weight of KL-divergence. As can be seen, the second term pushes the embedded distribution $p_\theta(z|x)$ to a given prior $q(z)$, so keeping geometric structures in latent embedding space is not the goal of VAE embedding.

2 Riemann manifold with Fisher metric

The single cell RNA sequencing data typically forms a manifold structure, which means the data points are concentrated to a hypersurface in the gene space, and the manifold typically is low-dimensional. Manifolds can be quantified by adding a metric tensor, which specifies the infinitesimal distances on it. A manifold equipped with a positive-definite metric tensor is called a Riemannian manifold. However, in many cases, there is no meaningful way to define the metric. By representing single cells as Gaussian distributions, there is a natural way to define the metric. Using theories from information geometry, we calculate the Fisher information of each cell, which acts as the Riemann metric in the probability distribution space. The Fisher information matrix is defined as

$$g_{ij} = \int dx \frac{\partial p(x,\theta)}{\partial \theta_i}\frac{\partial p(x,\theta)}{\partial \theta_j},$$

where θs are the parameters of the distribution. For 1-D Gaussian distribution, the Fisher information of μ and σ are $\frac{1}{\sigma^2}$ and $\frac{2}{\sigma^2}$ respectively [23, 33].And the Fisher information matrix G of a L-dimensional diagonal Gaussian distribution is

$$G = \begin{bmatrix} \frac{1}{\sigma_1^2} & & & & & & & \\ & \frac{2}{\sigma_1^2} & & & & & & \\ & & \ddots & & & & & \\ & & & \frac{1}{\sigma_l^2} & & & & \\ & & & & \frac{2}{\sigma_l^2} & & & \\ & & & & & \ddots & & \\ & & & & & & \frac{1}{\sigma_L^2} & \\ & & & & & & & \frac{2}{\sigma_L^2} \end{bmatrix}.$$

For clearance, each cell has a Fisher information matrix as each single cell X is transform into a Gaussian distribution $N(\boldsymbol{\mu}, \boldsymbol{\sigma})$.

Defining connections such as Levi-Civita connection which is fundamental to curvature in this space, remains a complex challenge when working with experimental data. Therefore, we employ an alternative method proposed by Yann Ollivier, which estimates the curvature on a graph by assigning a probability measure to each node, and then calculates the coarse Ricci curvature of each edge using the following equation [35]:

$$\kappa(a, b) = 1 - \frac{W(P_a, P_b)}{d(a, b)}$$

where $W(P_a, P_b)$ represents the Wasserstein distance between the probability measures of node a and node b and $d(a, b)$ is the length of shortest path between node a and node b [34]. The Wasserstein distance measures the minimum cost of transportation required when transforming one probability distribution to another. Ricci curvature, on the other hand, controls the overlap of two balls or the overlap between distributions, and is directly related to the Wasserstein distance [35].

By representing each single cell as a multivariate Gaussian distribution with mean $\boldsymbol{\mu}$ and diagonal covariance Σ, we can directly calculate the Wasserstein distance between neighboring cells:

$$W(P_a, P_b) = (|\mu_a - \mu_b|^2 + tr(\Sigma_a + \Sigma_b - 2(\Sigma_b^{1/2}\Sigma_a\Sigma_b^{1/2})^{1/2})^{1/2},$$

and $d(a, b)$ is the Euclidean distance between neighboring cells in gene space. Then we can calculate the coarse Ricci curvature (CRC) of each edge, and the CRC of each cell is then determined by the average of CRC across all its connecting edges $R(a) = \frac{1}{N_a}\sum_{b\in\{\Gamma_a\}}\kappa(a, b)$[62]. Here $\{\Gamma_a\}$ represent all the neighboring cells of cell a and N_a is the number of neighbors of cell a.

3 Test data generation

In our research, we use a series of toy models to generate data for validation of SCIM approach. Specifically, we employ four different models:

(1) Saddle surface:
Data points are randomly sampled from a saddle surface, a two-dimensional manifold embedding in a three-dimensional space, represented by the equation $z = k(x^2 - y^2)$. Here we take $k = 0.5$. To these points, we add seven additional dimensions of Gaussian distributed noise, which introduce minor variances in the data compared to the dominant first three dimensions. Despite the introduction of noise, the resulting ten-dimensional data points remain to be confined on the 2D saddle surface.

(2) Ellipsoid surface: We randomly sampled data points on an ellipsoid surface, where the original data are three-dimensional and defined by the equation

$$x = a sin\phi cos\psi$$
$$y = b sin\phi sin\psi$$
$$z = a cos\phi$$

(ϕ, ψ) are angles and (a, b) are the length of the semi-axes (a=1, b=4).

As with the saddle surface model, seven dimensions of noise were added to the original three-dimensional data.

(3) Four-well potential landscape: Data points were obtained from a four-well potential surface described by the function $U = a(k\,x^4 - x^2 + k\,y^4 - y^2)$ (where U is the potential), generated through trajectory simulations using the Langevin equation. Here we take $a = 0.5$, $k = 0.05$. Subsequent to the simulation, the trajectory data are added with seven dimensions of noise, in a manner consistent with the previous models.

(4) Simulated differentiation process with double branches

Data points were sampled with trajectory simulation on the double-branch potential surface described by the function $U = k_1x^4 + k_2yx^2 + k_3y$ (where U is the potential). The simulations were terminated when the range of y exceeds 35.

Here we use $k_1 = 0.05,\ k_2 = -0.06,\ k_3 = -1$. The trajectory data were also added with seven dimensions of noise.

By applying SCIM to these toy models, we can assess its feasibility and effectiveness in exploring the manifolds of CPT. These models allow us to evaluate the performance of SCIM in different scenarios and gain insights into its advantages and limitations.

4 Comparison of local metric preservation capability

To evaluate the preservation of metric in original space, we sampled points on the ellipsoid surface with equal major and minor axes (a=b=1) i.e. a sphere surface. We then sampled multiple semicircles on this surface, which represent geodesic lines on the sphere.

We first calculate the length of these geodesic lines (L_X) on the kNN graph using shortest path method in *Networkx* (Dijkstra algorithm). The weights assigned between neighboring nodes correspond to the pairwise Euclidean distances (d_X) and the lengths of these semi-circles approximate π, although the true values are subject to perturbation by seven-dimensional noise. Subsequently, we calculated the length of semi-circles (L_F) in the embedding spaces. For Gaussian embedding, the length is calculated with pairwise Fisher distances (d_F). For other manifold learning methods, since they lack defined metrics, we use the pairwise Euclidean

distances (d_E) within the embedding space to compute the length of semi-circles (L_E). By comparing the distribution of L_F/L_X with that of L_E/L_X, we assess the effectiveness of the Gaussian embedding in maintaining the integrity of the original local metric. And the coefficient of variation (CV) of the ratio distribution provides a quantitative measure of this comparison.

5 Eigengene analysis of regulatory mechanisms of CPT manifolds

To investigate the underlying regulatory mechanism of CPT manifolds and their characteristics, we compute the co-expression profile of genes based on their correlations. Utilizing the gene-gene correlation as a measure of distance, we conducted hierarchical clustering and separating the genes into distinct modules. The eigengene (EG) of each module is defined as its first principal component (PC1) value $EG = W_m X_m$ [63, 64], representing the collective behaviors of genes, where W_m is weight vector of gene module X_m. These eigengenes reveals the regulation of underlying gene networks and are treated as parameters that modulate the probability distributions in our analysis.

6 Identification of stiff and sloppy genes

In addition to being used for calculating invariant properties such as curvature and information velocity, Fisher information can also be employed to measure the stiffness and sloppiness of different parameters. The probability distribution is highly sensitive to changes in stiff parameters, whereas sloppy parameters have minimal impact on the distribution.

To calculate the Fisher information of different genes, we need to perform coordinate transformation from (μ, σ) space to gene space:

$$g_{\alpha\beta} = g_{ij} \frac{\partial \theta^i}{\partial x^\alpha} \frac{\partial \theta^j}{\partial x^\beta},$$

where $\theta = (\mu, \sigma)$ are the parameters, and x represent genes (here not the variables of Gaussian distribution). Instead of directly performing this transformation, we first transform to the PCA space and then to the gene space:

$$\frac{\partial \theta}{\partial x} = \frac{\partial \theta}{\partial x_{PCA}} \frac{\partial x_{PCA}}{\partial x}.$$

7 Information velocity and Information length

Information length is defined as the accumulated variation of information along a path in the manifold and is given by this equation,

$$L = \int \sqrt{g_{ij} d\theta^i d\theta^j}$$

(where $g_{ij} = \int_x \frac{\partial logP}{\partial \theta_i} \frac{\partial logP}{\partial \theta_j} P dx$ is the element of the fisher information matrix G, and $P = P(x, t)$ is a distribution function, here is one result of the Gaussian embedding. Einstein summation convention is applied). This concept is widely used in analyzing thermodynamics and attractor structure [30, 31]. To compute the information length between two cells, it is necessary to define the path that links them. Typically, this path is the geodesic, which is the shortest route connecting them. And the measure of the change of information along the geodesic in the probability space is referred to as the Fisher distance. To examine the variation of information length along the transition path at different stages of CPT, we employ pseudo-time to denote the progression of transitions. We first order the single cells according to their pseudo-time values, and divide them into 100 bins with equal numbers of cells. Then we calculate the mean cell in each bin and transform it into a Gaussian distribution. Next, we compute the Fisher distance between mean cells in neighboring bins. The information length along geodesic line on the information manifold, allow us to characterize the optimal transition paths between different phenotypes. We assess the correlation between mean expression levels of transcription factors and Fisher distances between neighboring bins, identifying transcription factors that are pivotal in modulating information variation in CPT (Fig. S12). This could be a complementary method of stiffness and sloppiness for identifying import regulatory genes from the perspective of information variation.

Information length can also be calculated with integration of information velocity along a path.

$$L = \int dt \sqrt{g_{ij} \frac{d\theta^i}{dt} \frac{d\theta^j}{dt}} = \int dt\, v_p.$$

By applying the chain rule twice, the information velocity v_p is defined and calculated with the following equation:

$$v_p := \sqrt{\int_x \left(\frac{dlogP}{dt}\right)^2 Pdx} = \sqrt{\frac{d\theta^i}{dt} g_{ij} \frac{d\theta^j}{dt}} = \sqrt{\frac{dx}{dt}\frac{\partial\theta^i}{\partial x} g_{ij} \frac{\partial\theta^j}{\partial x}\frac{dx}{dt}}$$

where $\frac{d\theta}{dt}$ represents the velocity of the parameter and θ is (μ, σ) in Gaussian distribution, $\frac{dx}{dt}$ represents RNA velocity. Information velocity measures the variation rate of the probability density function and reflects the speed of information variation along a path [30-32].

$$L = \int dt \sqrt{g_{ij} \frac{d\theta^i}{dt}\frac{d\theta^j}{dt}} = \int dt\, v_p.$$

We calculate the information velocity of each cell using RNA velocity. RNA velocity reflects the direction of single cell along the path of CPT in the gene expression space ($\boldsymbol{X}$), thus information velocity of single cell represents the speed of information variation along the transition path of CPT in probability distribution space. With RNA velocity $\frac{dx}{dt}$, $\frac{d\theta}{dt}$ can be expressed via chain rule $\frac{d\theta}{dt} = \frac{\partial\theta}{\partial x}\frac{dx}{dt}$. The $\frac{\partial\theta}{\partial x}$ is calculated using *Pytorch autograd* function numerically. For the eigen-gene or PCA coordinates (λ), the velocity is defined as $\frac{d\lambda}{dt} = \sum_m w_m \frac{dx_m}{dt}$, where w_m is the weight of velocity gene x_m in the corresponding coordinates.

8 Coordinate transformation and eigen-decomposition of Fisher matrix

We use a neural network to learn the mapping function between vector pairs. Specifically, we train a simple multi-layer perceptron neural network (two hidden layers) with *PyTorch*, where the input is the user-defined coordinates λ and the output is the (μ, σ) (represented with θ) of the Gaussian distribution. The derivative of this mapping function $\theta = f(\lambda)$ can be estimated with *autograd* in *PyTorch* numerically. This allows us to compute the transformation between (μ, σ) and other coordinates. With the following equation

$$\tilde{G}(\lambda) = \left(\frac{\partial\theta}{\partial\lambda}\right)^T G(\theta) \frac{\partial\theta}{\partial\lambda} = (\widetilde{g_{ij}}(\lambda))$$

Fisher metric $\tilde{G}$ in the coordinates λ with biological meaning such as eigengenes and principal components can be computed, thereby facilitating the quantitative analysis of CPT.

The Fisher information matrix in the (μ, σ) space is diagonal. However, this is not necessarily the case in other coordinate systems. Given that the Fisher information matrix is always symmetric and real, it is natural to consider its eigen decomposition in different coordinates. This can be easily accomplished after coordinate transformation.

9 Density estimation

To estimate local cell density, we employ a method using k-nearest neighbor and Gaussian kernel [65]. The local density is defined with $\rho_i = \Sigma_j exp(-d_{ij}^2/d_c^2)$, where d_{ij} is the distance between cell i to its neighbor cell j and d_c is a cutoff distance. We set d_c as the mean distance between all neighbor cell pairs. Given the experiment data is often noisy, we incorporate a smoothing step to refine our analyses. For a scaler function f like curvature and information velocity defined on single cell, its value is smoothed by the following equation $f_i^s = \Sigma_j f_j exp(-d_{ij}^2/d_c^2)/\rho_i$.

10 Locally Weighted Scatterplot Smoothing (LOWESS)

For better visualizing the variation of CRC or information velocity, we use LOWESS in *Scipy* to model the relation between the CRC or information velocity and pseudo-time. LOWESS is a non-parametric regression method that employs local linear estimation for smoothing scatter plot, providing a better view of the underlying trends in data.

11 pseudo-time analysis

The pseudo-time in all the single cell RNA sequencing datasets are determined with Palantir [66]. The Palantir method employs a probabilistic method for characterizing cell fate. With Palantir, we establish a pseudo-time ordering for cells within the single cell RNA sequencing dataset. The dataset for EMT of A549 cells includes samples that treated with different durations with TGF-β. Unlike actual experiment time, the pseudo-time uncovers the heterogeneity of single cell dynamics on a more granular temporal scale.

12 Dataset preprocessing

The single cell RNA sequencing datasets of dentate gyrus neurogenesis, pancreatic endocrinogenesis, and EMT of A549 treated by TGF-b were obtained from the GEO website with accession number GSE95753 [38], GSE132188 [41], GSE121861 [42] respectively.

To concentrate on the process of CPT within each dataset and minimize the influence of genes that are not associated with transition, we use two methods to select the genes exhibiting significant variations during CPT. The first one is selecting genes that display switch-like behaviors by using approach described in [50]. The second method is DUBStepR, which is based on gene correlation analysis and helps filter out those genes that are not cell-type-specific [49]. We run 30 regression steps to select seed genes which can best explain the gene-gene correlation, iteratively add the most correlated gene to this seed gene set, and take the first 600 genes as the feature genes.

Code availability

The scripts are available on GitHub at https://github.com/HuangMiao999/SCIM.

The authors declare no conflict of interest.

Acknowledgement

W.W. is supported by National Natural Science Foundation of China Grants No.12247104 and National Innovation Institute of Defense Technology Grant No. 22TQ0904ZT01025. L.Z. is supported by the National Key Research and Development Program of China 2021YFF1200500 and the National Natural Science Foundation of China (No.12225102, T2321001, and 12288101). We also thank Dr. Li Li (Institute of Theoretical Physics, Chinese Academy of Sciences) for insightful discussion on Riemann geometry.

References

1. Teschendorff, A.E. and T. Enver, *Single-cell entropy for accurate estimation of differentiation potency from a cell's transcriptome.* Nature Communications, 2017. 8(1): p. 15599.
2. Shi, J., et al., *Quantifying Waddington's epigenetic landscape: a comparison of single-cell potency measures.* Briefings in Bioinformatics, 2020. 21(1): p. 248-261.
3. Jin, S., et al., *scEpath: energy landscape-based inference of transition probabilities and cellular trajectories from single-cell transcriptomic data.* Bioinformatics, 2018. 34(12): p. 2077-2086.
4. La Manno, G., et al., *RNA velocity of single cells.* Nature, 2018. 560(7719): p. 494-498.
5. Qiu, X., et al., *Mapping transcriptomic vector fields of single cells.* Cell, 2022. 185(4): p. 690-711.e45.
6. Bergen, V., et al., *Generalizing RNA velocity to transient cell states through dynamical modeling.* Nature Biotechnology, 2020. 38(12): p. 1408-1414.
7. Chen, Z., et al., *DeepVelo: Single-cell transcriptomic deep velocity field learning with neural ordinary differential equations.* Science Advances. 8(48): p. eabq3745.
8. Gao, M., C. Qiao, and Y. Huang, *UniTVelo: temporally unified RNA velocity reinforces single-cell trajectory inference.* Nature Communications, 2022. 13(1): p. 6586.
9. François, P. and L. Jutras-Dubé, *Landscape, bifurcations, geometry for development.* Current Opinion in Systems Biology, 2018. 11: p. 129-136.
10. Corson, F. and E.D. Siggia, *Geometry, epistasis, and developmental patterning.* Proceedings of the National Academy of Sciences, 2012. 109(15): p. 5568-5575.
11. Lange, M., et al., *CellRank for directed single-cell fate mapping.* Nature Methods, 2022. 19(2): p. 159-170.
12. Puy, A., et al., *Models with higher effective dimensions tend to produce more uncertain estimates.* Science Advances, 2022. 8(42): p. eabn9450.
13. Luecken, M.D. and F.J. Theis, *Current best practices in single-cell RNA-seq analysis: a tutorial.* Molecular Systems Biology, 2019. 15(6): p. e8746.
14. Ye, Z. and C.A. Sarkar, *Towards a Quantitative Understanding of Cell Identity.* Trends Cell Biol, 2018. 28(12): p. 1030-1048.
15. Raimundo, F., et al., *Machine learning for single-cell genomics data analysis.* Current Opinion in Systems Biology, 2021. 26: p. 64-71.
16. Saelens, W., et al., *A comparison of single-cell trajectory inference methods.* Nature Biotechnology, 2019. 37(5): p. 547-554.
17. Moon, K.R., et al., *Visualizing structure and transitions in high-dimensional biological data.* Nature Biotechnology, 2019. 37(12): p. 1482-1492.
18. Kingma, D.P. and M. Welling, *Auto-Encoding Variational Bayes.* CoRR, 2013. abs/1312.6114.
19. Xu, J., et al., *Graph embedding and Gaussian mixture variational autoencoder network for end-to-end analysis of single-cell RNA sequencing data.* Cell Rep Methods, 2023. 3(1): p. 100382.
20. Lopez, R., et al., *Deep generative modeling for single-cell transcriptomics.* Nature Methods, 2018. 15(12): p. 1053-1058.

21. Choi, Y., R. Li, and G. Quon, *siVAE: interpretable deep generative models for single-cell transcriptomes.* Genome Biology, 2023. 24(1): p. 29.
22. Gayoso, A., et al., *A Python library for probabilistic analysis of single-cell omics data.* Nature Biotechnology, 2022. 40(2): p. 163-166.
23. Pinele, J., J.E. Strapasson, and S.I.R. Costa, *The Fisher–Rao Distance between Multivariate Normal Distributions: Special Cases, Bounds and Applications.* Entropy, 2020. 22(4): p. 404.
24. Amari, S.-i., *Information Geometry and Its Applications*. 2016: Springer Publishing Company, Incorporated.
25. Nielsen, F., *An Elementary Introduction to Information Geometry.* Entropy, 2020. 22(10): p. 1100.
26. Kang, X. and C. Li, *A Dimension Reduction Approach for Energy Landscape: Identifying Intermediate States in Metabolism-EMT Network.* Advanced Science, 2021. 8(10): p. 2003133.
27. Pearce, P., et al., *Learning dynamical information from static protein and sequencing data.* Nature Communications, 2019. 10(1): p. 1-8.
28. Bojchevski, A. and S. Günnemann, *Deep Gaussian Embedding of Graphs: Unsupervised Inductive Learning via Ranking.* arXiv: Machine Learning, 2017.
29. Vilnis, L. and A. McCallum, *Word Representations via Gaussian Embedding.* CoRR, 2014. abs/1412.6623.
30. Kim, E.-j., *Investigating Information Geometry in Classical and Quantum Systems through Information Length.* Entropy, 2018. 20(8): p. 574.
31. Kim, E.J., et al., *Geometric structure and geodesic in a solvable model of nonequilibrium process.* Phys Rev E, 2016. 93(6): p. 062127.
32. Crooks, G.E., *Measuring Thermodynamic Length.* Physical Review Letters, 2007. 99(10): p. 100602.
33. Costa, S.I.R., S.A. Santos, and J.E. Strapasson, *Fisher information distance: A geometrical reading.* Discrete Applied Mathematics, 2015. 197: p. 59-69.
34. Sia, J., E. Jonckheere, and P. Bogdan, *Ollivier-Ricci Curvature-Based Method to Community Detection in Complex Networks.* Scientific Reports, 2019. 9(1): p. 9800.
35. Ollivier, Y., *Ricci curvature of Markov chains on metric spaces.* Journal of Functional Analysis, 2009. 256(3): p. 810-864.
36. Lecun, Y., S. Chopra, and R. Hadsell, *A tutorial on energy-based learning*. 2006.
37. Ni, C.-C., et al., *Community Detection on Networks with Ricci Flow.* Scientific Reports, 2019. 9(1): p. 9984.
38. Hochgerner, H., et al., *Conserved properties of dentate gyrus neurogenesis across postnatal development revealed by single-cell RNA sequencing.* Nature Neuroscience, 2018. 21(2): p. 290-299.
39. Cox, D.R. and N. Reid, *Parameter Orthogonality and Approximate Conditional Inference.* Journal of the Royal Statistical Society: Series B (Methodological), 1987. 49(1): p. 1-18.
40. Transtrum, M.K., et al., *Perspective: Sloppiness and emergent theories in physics, biology, and beyond.* The Journal of Chemical Physics, 2015. 143(1).
41. Bastidas-Ponce, A., et al., *Comprehensive single cell mRNA profiling reveals a detailed roadmap for pancreatic endocrinogenesis.* Development, 2019. 146(12).

42. Cook, D.P. and B.C. Vanderhyden, *Context specificity of the EMT transcriptional response.* Nature Communications, 2020. 11(1): p. 2142.
43. Moris, N., C. Pina, and A.M. Arias, *Transition states and cell fate decisions in epigenetic landscapes.* Nature Reviews Genetics, 2016. 17(11): p. 693-703.
44. Brackston, R.D., E. Lakatos, and M.P.H. Stumpf, *Transition state characteristics during cell differentiation.* PLOS Computational Biology, 2018. 14(9): p. e1006405.
45. Mojtahedi, M., et al., *Cell Fate Decision as High-Dimensional Critical State Transition.* PLOS Biology, 2016. 14(12): p. e2000640.
46. Nieto, M.A., et al., *EMT: 2016.* Cell, 2016. 166(1): p. 21-45.
47. Pastushenko, I. and C. Blanpain, *EMT Transition States during Tumor Progression and Metastasis.* Trends Cell Biol, 2019. 29(3): p. 212-226.
48. Fang, Z., X. Liu, and G. Peltz, *GSEApy: a comprehensive package for performing gene set enrichment analysis in Python.* Bioinformatics, 2022. 39(1).
49. Ranjan, B., et al., *DUBStepR is a scalable correlation-based feature selection method for accurately clustering single-cell data.* Nature Communications, 2021. 12(1): p. 5849.
50. Wang, W., et al., *Transiently increased intercommunity regulation characterizes concerted cell phenotypic transition.* bioRxiv, 2021: p. 2021.09.21.461257.
51. Camastra, F. and A. Staiano, *Intrinsic dimension estimation: Advances and open problems.* Information Sciences, 2016. 328: p. 26-41.
52. Sritharan, D., S. Wang, and S. Hormoz, *Computing the Riemannian curvature of image patch and single-cell RNA sequencing data manifolds using extrinsic differential geometry.* Proc Natl Acad Sci U S A, 2021. 118(29).
53. Coifman, R.R., et al., *Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps.* Proceedings of the National Academy of Sciences, 2005. 102(21): p. 7426-7431.
54. Fasina, O.T., et al., *Neural FIM for learning Fisher Information Metrics from point cloud data.* ArXiv, 2023. abs/2306.06062.
55. Bergen, V., et al., *RNA velocity—current challenges and future perspectives.* Molecular Systems Biology, 2021. 17(8): p. e10282.
56. Scheffer, M., et al., *Anticipating Critical Transitions.* Science, 2012. 338(6105): p. 344-348.
57. Chen, L., et al., *Detecting early-warning signals for sudden deterioration of complex diseases by dynamical network biomarkers.* Scientific Reports, 2012. 2(1): p. 342.
58. Scheffer, M., et al., *Early-warning signals for critical transitions.* Nature, 2009. 461(7260): p. 53-59.
59. Prokopenko, M., et al., *Relating Fisher information to order parameters.* Physical Review E, 2011. 84(4): p. 041116.
60. Brody, D.C. and A. Ritz, *Information geometry of finite Ising models.* Journal of Geometry and Physics, 2003. 47(2): p. 207-220.
61. Ruppeiner, G., *Riemannian geometry in thermodynamic fluctuation theory.* Reviews of Modern Physics, 1995. 67(3): p. 605-659.
62. Wu, W., G. Hu, and F. Yu *Ricci Curvature-Based Semi-Supervised Learning on an Attributed Network*. Entropy, 2021. 23, DOI: 10.3390/e23030292.

63. Langfelder, P. and S. Horvath, *Eigengene networks for studying the relationships between co-expression modules.* BMC Systems Biology, 2007. 1(1): p. 54.
64. Langfelder, P. and S. Horvath, *WGCNA: an R package for weighted correlation network analysis.* BMC Bioinformatics, 2008. 9(1): p. 559.
65. Du, M., S. Ding, and H. Jia, *Study on density peaks clustering based on k-nearest neighbors and principal component analysis.* Knowledge-Based Systems, 2016. 99: p. 135-145.
66. Setty, M., et al., *Characterization of cell fate probabilities in single-cell data with Palantir.* Nature Biotechnology, 2019. 37(4): p. 451-460.