# Learning cell-specific networks from dynamics and geometry of single cells

Stephen Y Zhang *[1,2] and Michael P H Stumpf[1,2,3,4]

[1]School of Mathematics and Statistics, University of Melbourne, Australia
[2]Melbourne Integrative Genomics, University of Melbourne, Australia
[3]School of Biosciences, University of Melbourne, Australia
[4]ARC Centre of Excellence for the Mathematical Analysis of Cellular Systems (MACSYS)

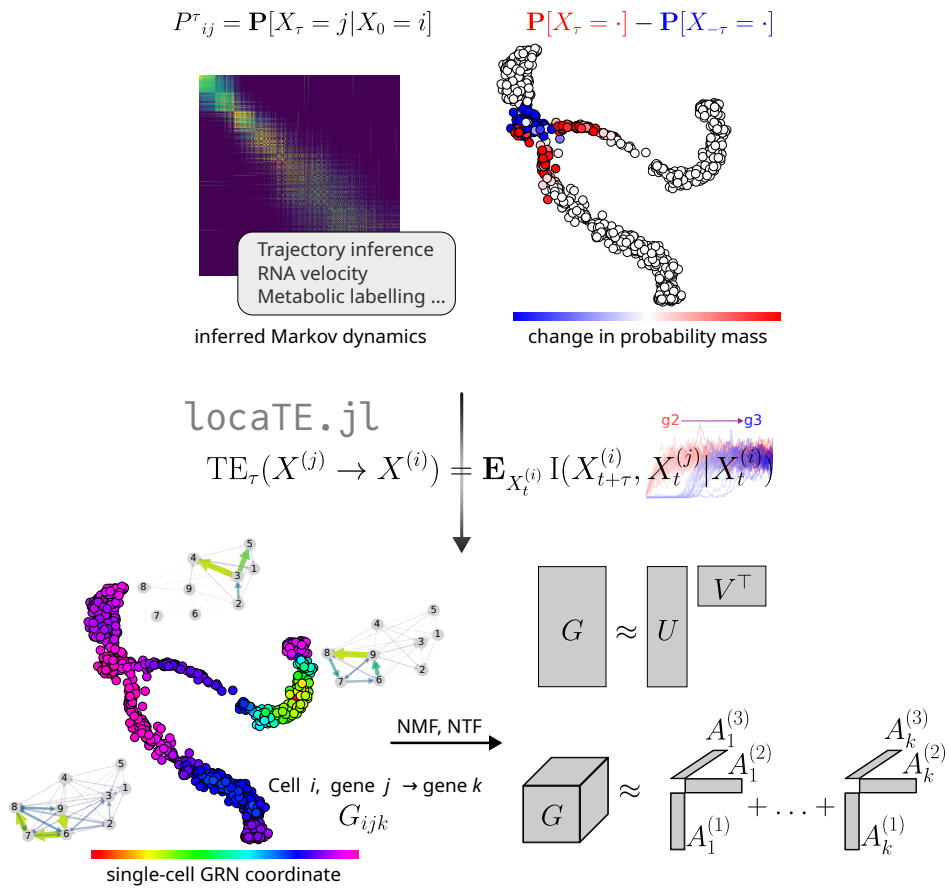
**Abstract**

Cell dynamics and biological function are governed by intricate networks of molecular interactions. Inferring these interactions from data is a notoriously difficult inverse problem. The majority of existing network inference methods work at the population level to construct population-averaged representations of gene interaction networks, and thus do not naturally allow us to infer differences in gene regulation activity across heterogeneous cell populations. We introduce locaTE, an information theoretic approach that leverages single cell dynamical information together with geometry of the cell state manifold to infer cell-specific, causal gene interaction networks in a manner that is agnostic to the topology of the underlying biological trajectory. We find that factor analysis can give detailed insights into the inferred cell-specific GRNs. Through a detailed simulation study and application to three experimental datasets spanning mouse primitive endoderm formation, pancreatic development, and haematopoiesis, we demonstrate superior performance and additional insights compared to standard static GRN inference methods. We find that locaTE provides a powerful, efficient and scalable network inference method that allows us to distil cell-specific networks from single cell data.

## 1 Introduction

Cell identity is determined by the dynamics underlying gene expression. Gene expression, in turn, is controlled by gene regulatory networks that direct the activity of genes over time, and which set the differences between cells. Identifying and characterising gene regulatory interactions is a central aim of cell and systems biology. Single cell assays have been enabling us to capture population heterogeneity in gene expression, allowing for the detection and study of rare and transient cellular phenotypes that would be unobservable from bulk studies [1]. This has demonstrated variability within groups of cells which would conventionally have been regarded as being of the same "type". The popularity of single cell data has spurred rapid advances in data analysis and modelling methods such as trajectory inference [2], dimensionality reduction [3], and network inference [4, 5].

*stephenz@student.unimelb.edu.au




**Graphical abstract: Cell-specific network inference from estimated dynamics and geometry.** LocaTE takes as input a transition matrix $P$ that encodes inferred cellular dynamics as a Markov chain on the cell state manifold. By considering the coupling $(X_\tau, X_{-\tau})$, locaTE produces an estimate of transfer entropy for each cell $i$ and each pair of genes $(j, k)$. Downstream factor analyses can extract coherent patterns of interactions in an unsupervised fashion.

Modern single cell datasets overcome the "large-$p$ small-$N$ problem" [6] that has plagued network inference from bulk data: the number of hypotheses $p$ is smaller than the number of data points samples [7].The number of hypotheses or predictions grows quadratically with the number of genes considered. In this vein, a collection of methods have been developed for network inference from single cell data, some of which build on methods originally intended for bulk expression data, and others which are specifically tailored to single cell data. The recent surveys in [8, 9, 5] provide an overview of the current methodology. The overwhelming majority of network inference methods aim to reconstruct a single *static* network [5, 8] that describes the set of possible interactions occurring within an observed population. Given the biological importance of cellular heterogeneity, a natural expectation is that variations in transcriptional state may correspond to variations in (cell state dependent) regulatory interactions which cannot be represented as static networks [7], and indeed has been observed empirically in previous studies [10, 11].

Single-cell analysis approaches that model cell phenotype as a continuous rather than discrete covariate have recently received attention in applications such as differential gene expression [12] and allelic variant analysis [13]; but in the context of network inference there has been comparatively little related work. The need for methods that can learn cell- or phenotype-specific networks [7, 14], is now leading to new classes of network inference methods [15, 16, 17, 18]. These methods seek to infer multiple networks (e.g. one per cell type, or even one per cell) from single-cell resolved data, thus allowing us to learn about variation in gene interactions over time and between conditions. Such methods generally rely on neighbourhood or cluster information to construct networks that are cell, time, or context-specific. Existing methods for cell-specific network inference have focused on inferring undirected networks [15, 16, 17]. If we hope to infer *causal* or

*directed* interactions, then information about underlying dynamics in a cell population is essential [19, 20]. Many methods for static network inference take advantage of pseudotemporal orderings to inform network inference [19, 21, 22, 23, 24]. However, the one-dimensional nature of pseudotime necessarily imposes a total ordering over cells, resulting in a loss of finer structure of the cell state manifold. Recent developments in trajectory inference from a Markov process viewpoint [25, 26, 27, 28] depart from this framework and are able to model more complex dynamics on the set of observed cell states directly, and are free from assumptions on the trajectory topology. Manifold learning approaches [29] avoid clustering of cell states and instead construct a cell-state graph by learning local neighbourhoods and are therefore flexible enough to model arbitrarily complex trajectory structures (see Box 1 for further discussion).

Here we adopt transfer entropy (TE) [30] as an information-theoretic measure of causality. Transfer entropy, like other information-theoretical measures [4], is a model-free framework for measuring dependencies between random variables which has been used widely in contexts such as neuroscience where abundant time-series data are available. In a Gaussian setting it is known to be equivalent to Granger causality [31]. Transfer entropy has been used to predict static networks using a pseudotemporal ordering [19, 21]. We demonstrate that the transfer entropy can be adapted to infer cell-specific networks directly from dynamics on the cell-state manifold without imposing a prescribed ordering of cells.

Inference of cell-specific networks helps us, for example, to tackle problems in developmental biology, where it is often the transient dynamics that are the most interesting [32] since these are likely to correspond to major epigenetic changes [7]. Our approach is tailored to such problems and we make no restricting simplifying assumptions: the cell state space is treated as a continuum, and no *a priori* partitioning or clustering is required to group or classify cells.

## Box 1: Beyond pseudotime: geometry of the cell-state manifold

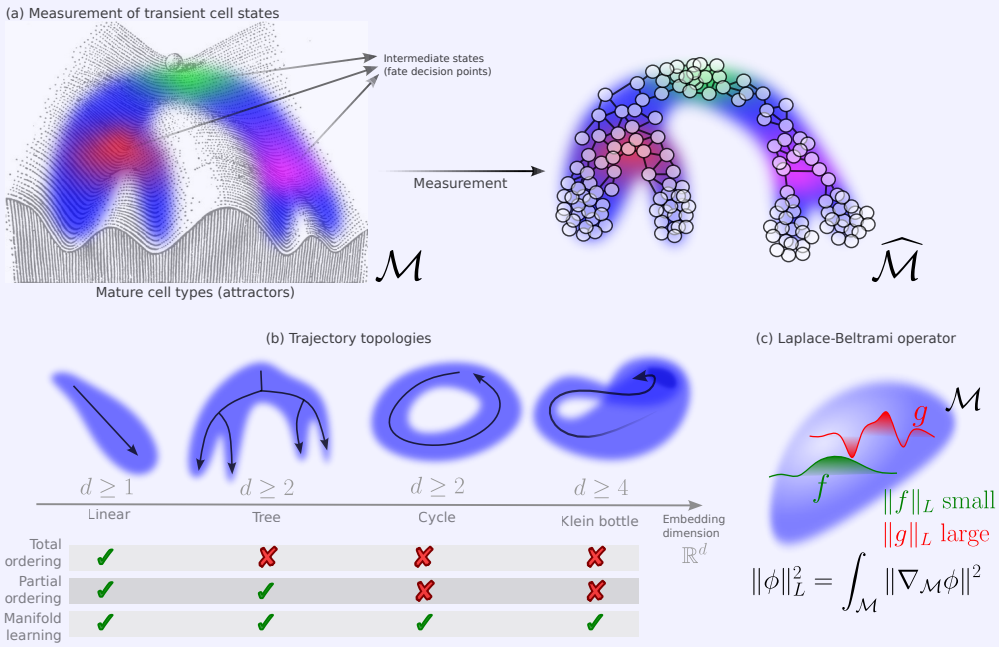



Figure 1: (a) Manifold $\mathcal{M}$ of cell states overlaid on the classical Waddington's landscape: intermediate cell states correspond to points of fate decision. Measurement by single-cell technologies produces a set of observed cell states, from which an approximation $\widehat{\mathcal{M}}$ of the underlying cell state manifold can be learned. (b) Total and partial orderings (and thus, pseudotemporal orderings and extensions thereof) are limited to linear and tree-like trajectory topologies, while manifold learning approaches, which rely on construction of *neighbourhoods* of points, are by nature topology-agnostic. (c) The Laplace-Beltrami operator is a central object describing a manifold $\mathcal{M}$ and in particular compares functions in a manner adapted to the underlying geometry.

Developmental systems are by nature dynamic: C. H. Waddington's classical landscape metaphor [33] conceptualised biological processes as populations of cells traversing down a landscape of epigenetic potential from an uphill progenitor state to mature celltypes, thought of as attractors at the bottom of the landscape. In this landscape, attractors corresponding to biological cell types are separated by saddle nodes which can be thought of as transitory states at which "fate decisions" are made by differentiating cells. Single cell profiling makes it possible to measure not only cells within attractors, but also transient cell states at these decision points [32, 34], enabling the examination of intermediate states that is not possible in bulk analyses [35].

Identification of transient cell states requires knowledge of the (temporal) structure of the underlying biological process. The problem of recovering this information from single cell measurements is called trajectory inference [2]. For example, pseudotime imposes an ordering of cells based on relative progression through the developmental process. Pseudotime is fundamentally limited to a small set of trajectory topologies that it can handle: a total ordering of cells can only accurately reflect a linear topology. Branched pseudotime methods produce a partial ordering of cells which can describe tree-like topologies. In two dimensions or higher, trajectories can exhibit cycles; in four or more dimensions, more complex topologies such as the Klein bottle are theoretically possible. Trajectories that depart from linear and tree-like topologies cannot be accurately represented by pseudotime methods. Approaches based on manifold learning, by contrast, aim to construct local neighbourhoods of cells, and thus can naturally handle arbitrary topologies to an extent limited only by the amount of available data. Since single cell datasets are naturally embedded in $\sim 10^4$ (the number of discernible transcripts) dimensions, there are a vast number of degrees of freedom and hence ways in which linear or tree-like topological assumptions may be violated.

## Box 2: Single cell dynamics and transfer entropy

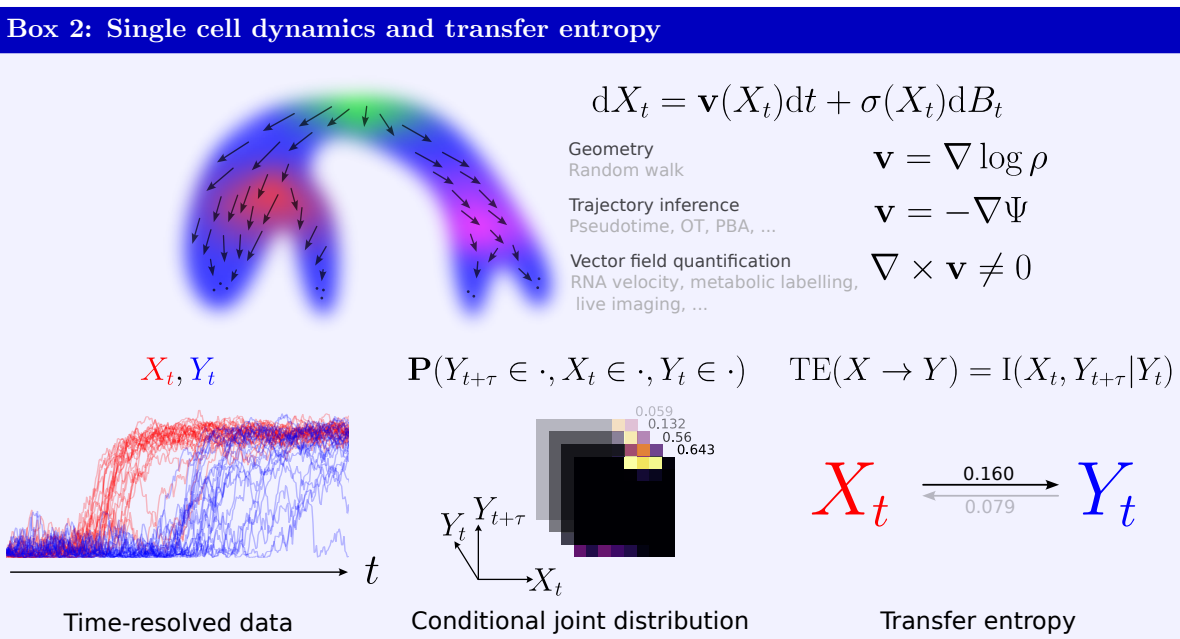



Figure 2: Top: Single cell dynamics reflect the underlying developmental program. Dynamics can be inferred or estimated using a range of computational and experimental techniques, which may incur various structural limitations on the inference. Bottom: Transfer entropy is an information theoretic approach for quantifying information flow through a stochastic dynamical system.

Trajectory inference methods aim to recover dynamics from single cell data, but have limitations on the dynamics they can model. From geometry of single cell measurements alone, construction of a random walk over observed cell states is consistent with an inferred drift of $\mathbf{v} = \nabla \log \rho$ where $\rho$ is the density of observed cells [36, 35] – if the observation process draws from a closed system at equilibrium then this recovers the potential $\Psi$. Methods such as pseudotime, optimal transport and population balance analysis implicitly assume potential-driven dynamics $\mathbf{v} = -\nabla\Psi$ and thus cannot model behaviours such as oscillations [25]. On the other hand, approaches such as RNA velocity, metabolic labelling and live imaging enable direct quantification of the vector field [37, 28]. These methods thus have the potential to identify dynamics that cannot be modelled by geometry or trajectory inference alone, one notable example being the periodic dynamics of the cell cycle. We provide further discussion of these points in S1.2.

The mutual information between two random variables is a non-parametric measure of statistical dependence that is free of distributional assumptions and can capture non-linear relationships between variables. For two random variables, $X, Y$, with joint distribution, $p_{XY}$, the mutual information is defined as

$$\mathrm{I}(X,Y) = \mathrm{KL}(p_{XY}|p_X \otimes p_Y) = \int \mathrm{d}p_{XY} \log\left(\frac{\mathrm{d}p_{XY}}{\mathrm{d}p_X \mathrm{d}p_Y}\right) = \mathrm{H}(p_{XY}) - \mathrm{H}(p_X) - \mathrm{H}(p_Y),$$

where $\mathrm{H}(p)$ is Shannon's (neg)entropy, $\mathrm{H}(p) = \int \mathrm{d}p \log\left(\frac{\mathrm{d}p}{\mathrm{d}x}\right)$. We have $I(X,Y) \geqslant 0$, and $I(X,Y) = 0$ if and only if $X$ and $Y$ are independent, i.e. $\mathbf{P}(X \in \cdot, Y \in \cdot) = \mathbf{P}(X \in \cdot)\mathbf{P}(Y \in \cdot)$. For a third random variable $Z$, the *conditional* mutual information can be defined in terms of a conditional expectation

$$\mathrm{I}(X,Y|Z) = \sum_z \mathrm{I}(X|Z=z, Y|Z=z)\mathbf{P}[Z=z],$$

where $X|Z=z, Y|Z=z$ denote random variables conditioned on the event $\{Z=z\}$.

The *transfer entropy* (TE) between two time-dependent random variables $(X_t, Y_t)$ measures information transfer from $X_t$ to $Y_t$ over time. For a chosen time lag parameter, $\tau > 0$, the transfer entropy $\mathrm{TE}[X_t \to Y_t]$ is defined as $\mathrm{I}(X_t, Y_{t+\tau}|Y_t)$, the mutual information between $X_t$ and $Y_{t+\tau}$ conditioned on $Y_t$ in order to remove the dependence of $Y$ on itself.

## 2 Method

**Local transfer entropy of expression dynamics** We represent the transcriptional state of a cell as a vector of expression levels $X = (X^{(1)}, \ldots, X^{(N)})$. Transfer entropy does not make assumptions on the specific form of a model, but for the sake of illustration let us consider a model of cellular dynamics where the expression level of each transcript $X^{(i)}$ is governed by a stochastic differential equation (SDE) of the form

$$\mathrm{d}X_t^{(i)} = f^{(i)}(X_t^{(1)}, \ldots, X_t^{(N)})\mathrm{d}t + \sigma^{(i)}(X_t^{(1)}, \ldots, X_t^{(N)})\mathrm{d}B_t. \tag{1}$$

Note that this framework includes the chemical Langevin equation (CLE) [38], a biophysically derived model of single cell expression dynamics. Under the Itô interpretation [39, 40], for a small time interval $\tau$, we may write approximately,

$$X_{t+\tau}^{(i)} = X_t^{(i)} + f^{(i)}(X_t^{(1)}, \ldots, X_t^{(N)})\tau + \sigma^{(i)}(X_t^{(1)}, \ldots, X_t^{(N)})\sqrt{\tau}Z^{(i)}, \tag{2}$$

where the $Z^{(k)}$ are independently and identically distributed Gaussian random variables. We can condition on $\{X_t^{(i)} = x\}$,

$$X_{t+\tau}^{(i)}|\{X_t^{(i)} = x\} = x + f^{(i)}(\{X_t^{(k)}\}_{k\neq i}, x)\tau + \sigma^{(i)}(\{X_t^{(k)}\}_{k\neq i}, x)\sqrt{\tau}Z^{(i)}. \tag{3}$$

For any other gene $X^{(j)}, j \neq i$, following [41, Section 4.2], the *transfer entropy* with time lag $\tau$ is defined as

$$\mathrm{TE}_\tau(X^{(j)} \to X^{(i)}) := \mathbf{E}_{X_t^{(i)}}\left[\mathrm{I}(X_{t+\tau}^{(i)}, X_t^{(j)}|X_t^{(i)})\right], \tag{4}$$

where

$$\mathrm{I}(X, Y|Z) := \mathbf{E}_{X,Y}\left[\log \rho_{XY|Z} - \log(\rho_{X|Z} \otimes \rho_{Y|Z})\right] \tag{5}$$

is the mutual information of two random variables, $X, Y$, conditional on a third variable, $Z$. Under the discrete approximation, the transfer entropy will be positive whenever $f^{(i)}$ or $\sigma^{(i)}$ depend on $X^{(j)}$, since we then have $\mathrm{I}(X_{t+\tau}^{(i)}, X_t^{(j)}|X_t^{(i)}) > 0$. In the biophysically relevant setting of the CLE, the functions $f$ and $\sigma$ share the same functional dependence on $\{X^{(k)}\}_k$ when $X^{(i)} = x$, and so for infinitesimal time-scales $\tau$ a non-zero transfer entropy implies regulation, i.e. $\partial_j f^{(i)} > 0, \partial_j \sigma^{(i)} > 0$. In the practically relevant setting where the time-scale, $\tau$, is short, the data processing inequality [42] guarantees that direct interactions will dominate indirect interactions, i.e. for a feed-forward loop $X^{(i)} \to X^{(j)} \to X^{(k)}$, one has that $\mathrm{I}_{ik} < \min\{\mathrm{I}_{ij}, \mathrm{I}_{jk}\}$ [43].

The concept of transfer entropy has previously been used for inference of static networks from pseudotemporally ordered expression data [19, 21]. However, in these methods the expectations in (5), (4) are taken with respect to the set of lagged tuples of cells along a total ordering. This introduces a strong dependence upon the provided pseudotemporal ordering; it also results in loss of resolution: any variation in regulatory activity over the expression space is integrated out to yield a static network. This can be problematic since biological trajectories may exhibit bifurcating, cyclic or converging trajectories, potentially requiring multiple runs of GRN inference on different subsets of a dataset.

We propose to take, for each observed cell $\hat{X}$, the expectation (5) with respect to the evolution described by the transition kernel $P^\tau$ of (1) starting from some neighbourhood $\mathcal{N}(\hat{X})$ of the cell. For short time scales, $\tau$, and an appropriately sized neighbourhood, $\mathcal{N}(\hat{X})$, (i.e. large enough to take expectations, but not so large as to lose local detail), the resulting dynamics are therefore local. This is best understood as a conditional transfer entropy,

$$\mathrm{TE}_\tau\left[X^{(j)} \to X^{(i)} \mid X_0 \in \mathcal{N}(\hat{X})\right] := \mathbf{E}_{X_0\in\mathcal{N}(\hat{X})}\,\mathrm{I}(X_\tau^{(i)}, X_0^{(j)}|X_0^{(i)}).$$

Inferred regulatory interactions are specific to the local cell context, which is specified by its neighbourhood. Some information-theoretic static inference methods [4, 19] condition on a third gene in order to remove indirect interactions. Instead of conditioning separately on each potential indirect interaction, our framework conditions on a local region of the expression space. This can be interpreted as conditioning on some (potentially non-linear) function of the overall gene expression profile that is determined by the construction of the cellular manifold (e.g. in the case of PCA, this would be an affine combination).

Finally, since we model dynamics as a general Markov process, no assumptions on the ordering of cells or topology of trajectories is made or necessary.

**Inference in practice** Typically we do not have information about the drift and diffusion terms [39] in (1); dynamical information must instead be *inferred* from observed expression data. This *trajectory inference* problem has been widely studied [26, 44, 2]. Since we are interested in cell-specific networks, we restrict our focus to a class of methods that model cellular dynamics as a Markov process involving drift and diffusion over a manifold of cell states [25, 27, 45, 28]. Let $\mathcal{X} = \{x_i\}_{i=1}^N$ be a sample from a population of cells, where $x_i \in \mathbb{R}^M$ is the vector of mRNA expression values for cell $i$; $\mathcal{M}$ denotes a graph constructed from $\mathcal{X}$ that can be thought of as approximating the underlying manifold of cell states [29]. We assume that we have access to a transition matrix $P \in \mathbb{R}^{N\times N}$ that is supported on $\mathcal{M}$ and that adequately describes the ("true") dynamics of (1). In practice, as we do in Section 3, we could use a range of recently reported methods to estimate $P$. Then $\mathcal{M}$ equipped with $P$ encodes a discrete Markov chain that approximates the true cellular dynamics. In what follows, we write $[\![n]\!] = \{1, \dots, n\}$ for $n \in \mathbb{N}$.

For any cell, $x \in \mathcal{M}$, we can construct (using e.g. a uniform distribution, truncated Gaussian, etc.) a probability distribution supported on its neighbourhood, $\pi_0^x$. We then consider a Markov process $X_t, t \geqslant 0$ that starts from $X_0 \sim \pi_0^x$ and evolves under $P$. For some fixed $\tau$ the resulting *coupling* is

$$(X_0, X_\tau) \sim \mathrm{diag}(\pi_0^x)P^\tau =: \gamma_\tau^x. \tag{6}$$

For a pair of genes, $(j, k)$, we quantify the local transfer entropy for the causal relationship, $j \to k$, from knowledge of the coupling, $\gamma_t^x$, and the gene expression states,

$$\mathrm{TE}\left[X_0^{(j)} \to X_\tau^{(k)}\right] = \mathbf{E}_{(X_\tau, X_0)\sim\gamma_\tau^x}\left[\mathrm{I}(X_0^{(j)}, X_\tau^{(k)} | X_0^{(k)})\right]. \tag{7}$$

Repeating this computation for all cells, $x_i \in \mathcal{M}$, and gene pairs, $j, k \in [\![M]\!]$, we arrive at a tensor, $\hat{G} \in \mathbb{R}^{N\times M\times M}$, such that, $\hat{G}_{ijk}$, is the transfer entropy score of $j \to k$ in the neighbourhood of cell, $x_i$. Since the TE score matrix, $\hat{G}_i$, for each cell has to be learned from a local neighbourhood, it is potentially noisy. Furthermore, it is impractical (due to limited data and computational resources) to condition on potential indirectly interacting genes [19]. Thus, TE scores may contain noise and spurious signals resulting from indirect interactions. To deal with this, we first filter interaction scores using context likelihood of relatedness (CLR) [46, 4] and then solve a manifold-regularised optimisation problem to both de-noise the interaction signals, and to attenuate noise and indirect interactions [19, 15].

**Context likelihood of relatedness.** Given raw TE scores, $\hat{G}$, we use the context likelihood of relatedness [47] to filter scored interactions. The CLR algorithm has been shown to filter out indirect interactions in the context of mutual information scores [47] and other information theoretical measures [4]. For a pair of genes, $(i, j)$, and given a interaction score matrix, $A$, we compute $z_i$ (resp. $z_j$) to be the $z$-score of $A_{ij}$ with respect to $A_{i\cdot}$ (resp. $A_{\cdot j}$). Then we define the *weighted* CLR score for $(i, j)$ to be

$$\widetilde{A}_{ij} = \frac{A_{ij}}{2}\sqrt{\max(0, z_i)^2 + \max(0, z_j)^2}. \tag{8}$$

Applying CLR filtering along the first axis of $\hat{G}$, we obtain the filtered tensor, $\widetilde{G}$. We modify the original approach of [47] and weigh the CLR score by the initial MI value $\hat{G}_{ij}$. This is important since CLR was originally designed to filter interactions in static networks. In cell-specific networks, very few edges may be "active" in any given cellular context; then entire rows or columns of $\hat{G}$ may consist only of noise. Computing $z$-scores along those rows or columns would put both noise and signal on the same scale.

**Smoothing and denoising** The tensor $\widetilde{G}$ contains a (noisy) matrix of filtered interaction scores, one for each cell. For notational convenience we write $\widetilde{G} \in \mathbb{R}^{N\times M^2}$, i.e. each row is a length-$M^2$ unfolded score matrix. We propose to solve the optimisation problem, $G = \arg\min_X \mathcal{L}(X; L, \widetilde{G})$, where

$$\mathcal{L}(X; L, \widetilde{G}) = \frac{1}{2}\sum_{i=1}^N w_i \|X_{i\cdot} - \hat{G}_{i\cdot}\|_2^2 + \frac{\lambda_1}{2}\,\mathrm{tr}(X^\top L X) + \lambda_2 \sum_{i=1}^N w_i \|X_{i\cdot}\|_1. \tag{9}$$

Here $\lambda_1, \lambda_2$ are hyper-parameters, $w$ is a vector of cell weights, and $L$ is the symmetric graph Laplacian on $\mathcal{M}$. The term associated with $\lambda_1$ is a manifold regularisation term corresponding to an expectation that

regulatory relationships should vary smoothly with changes in cell state, i.e. $\mathrm{tr}(X^\top LX)$ is large for rapidly fluctuating $X$. The term associated to $\lambda_2$ is a Lasso [6] term that encourages $X$ to be sparse, reflecting our knowledge of biological networks [20, 15]. Together, the objective (9) encourages both parsimony and sharing of information along the cell state manifold. The problem is a case of L1-L2-regularised least squares, and can be solved efficiently using an iterative scheme described in Section S1.4. Provided we have a transition matrix, the application of locaTE requires only the choice of parameters $(\tau, \lambda_1, \lambda_2)$. Section S2.1 provides discussion and guidance regarding the choice of these parameters.

**Using a backward transition matrix** Because we consider couplings of the process at times $(0, \tau)$, TE scores may reflect the underlying dynamics up to a shift in time. To remedy this, we consider couplings symmetric in time, i.e. between $(-\tau, \tau)$. While time-reversal of a diffusion-drift process away from equilibrium is generally not a well-posed problem, in practice, suitable backward operators have been constructed by transposing the transition matrix [27, 48] (see Appendix S1.3 for further discussions). Given an approximate backward operator, $Q$, a time-symmetric coupling, $\gamma^x_\tau = (Q^\top)^\tau \,\mathrm{diag}(\pi^x_0)P^\tau$, can be constructed.

**Undirected inference without dynamical information** Dynamical information is not always straightforward to obtain for a given dataset. Metabolic labelling data are experimentally non-trivial to obtain, quality of RNA velocity outputs can be strongly dependent on dataset and modelling assumptions, and trajectory inference methods can be sensitive to parameter and preprocessing choices. Similarly, some snapshot measurements are made from the equilibrium distribution of the underlying biological process, at which point causal information is lost. In these settings, inference of causal interactions between genes is not possible without additional (e.g. perturbation response) data or structural assumptions on a mechanistic model [49]. Instead, the goal of inferring undirected interactions in this setting has been addressed by a range of methods, of which GENIE3 [50] and PIDC [4] have been among the most successful [5]. The locaTE framework can be seamlessly applied to infer undirected cell-specific networks as a special case. In the absence of inferred dynamics, from an observed cell population, $\mathcal{X}$, a Markov chain can be constructed as an undirected random walk on the cell state manifold. There are a variety of ways in which this random walk can be constructed: for example, a random walker located at cell $x$ randomly chooses another cell $x'$ in its neighbourhood, $\mathcal{N}(x)$, with a probability proportional to $\exp(-\|x - x'\|^2/\sigma^2)$. Under an appropriate normalisation of the transition kernel [36, Section 3], the resulting random walk is consistent with the stochastic dynamics of a system at equilibrium, where $\sigma$ is constant and the vector field $f$ is

$$f(x) = \nabla \log p(x) = \frac{\nabla p(x)}{p(x)},$$

where $p(x)$ is the population density of cells.

**Normalisation of local transfer entropy scores** For information theoretic measures such as mutual information normalisation is not straightforward [47, 4], and therefore direct comparison of edge scores may not be an accurate reflection of relative edge confidence. We employ a cumulative distribution function (CDF) normalisation procedure to locaTE scores prior to downstream analysis which was found to work well in [4]. For a $M \times M$ reference matrix $A$, Gamma distributions are fit to each row and column to yield CDFs $F_{i\in[\![M]\!]}$, $F'_{j\in[\![M]\!]}$. A CDF-transformed matrix $\tilde{B} = \texttt{cdf_transform}(B; A)$ is then constructed as

$$\tilde{B}_{ij} = \frac{F_i(B_{ij}) + F'_j(B_{ij})}{2}.$$

LocaTE produces a $N \times M \times M$ tensor $G$ which describes the smoothed $M \times M$ cell-specific network in each of the $N$ input cells. We construct the reference matrix as $A_{jk} = \max_i G_{ijk}$, where the maximum is taken since interactions may only be non-zero in a small subset of cells. The CDF-transformed tensor $\tilde{G}$ is thus constructed cell-wise as

$$\tilde{G}_{i\cdot\cdot} = \texttt{cdf_transform}(G_{i\cdot\cdot}; A)$$

**Computational considerations** The problem of cell-specific network inference is at least $\mathcal{O}(NM^2)$ for $N$ cells and $M$ genes, since each cell-gene-gene triplet must be considered. For each triplet, $(i, j, k)$, locaTE calculates the joint distributions of $(X_\tau^{(k)}, X_0^{(j)}, X_0^{(k)})$ under the coupling $\gamma^{(i)}$. While in the worst-case (if $\gamma$ were dense) this would have complexity $\mathcal{O}(N^2)$, in all our applications $\gamma$ is sparse and transitions occur only between nearby cells; the complexity is therefore $\mathcal{O}(|\gamma^{(i)}|) \ll N^2$. The CLR filtering step is also $\mathcal{O}(NM^2)$, and the regression step primarily involves iterative matrix products of sparse $N \times N$ against $N \times M^2$. We implemented locaTE using the Julia programming language [51, 52] to take advantage of its speed for numerical computation.

With exception of the regression step, which is global, all computational steps are easily parallelised across cells. The calculation of TE scores can alternatively be parallelised across gene pairs. This allows locaTE to take advantage of large scale parallelisation when available, potentially drastically reducing the compute time required. We find empirically that a large portion of computational burden arises from the construction of the joint distribution, $(X_\tau^{(k)}, X_0^{(j)}, X_0^{(k)})$. Since this is an accumulation operation, this can be accelerated using GPU computation. Other steps in the locaTE pipeline (regression, factor analysis) amount to linear algebra operations and are also amenable to GPU acceleration. For datasets in excess of $\sim 10^4$ cells, the computational cost could be greatly reduced by first grouping sampled cells into metacells, using e.g. the approach in [53]. We present an example of this in Section S2.4.

# 3 Results

We first discuss two simulation examples where the true regulatory interactions are known by design. To illustrate the advantages of cell-specific networks, we consider examples where the network is rewired over time. Results for a wider range of example developmental trajectories are provided in Appendix S2.2.

## 3.1 Simulated data: branch-dependent logic

**Overview** We start by considering a scenario where a static network cannot capture the mechanisms underlying gene regulation. In the network in Figure 3(a) a bifurcation driven by a toggle-switch feeds into one of two "gene expression modules". Modules A and B involve the same genes $\{g6, g7, g8, g9\}$, but with very different interactions among the genes: the flow of information in each module is the mirror image of the other.

A single population-level network (in the graphical modelling sense) is insufficient to describe this system, since the directionality of some interactions depends on cellular context. Context dependence can alternatively be understood as arising from the presence of higher-order interactions. Consider gene 7, which may either be activated by gene 6 (in module A) or by gene 8 (in module B). This could be understood conceptually as a *hyperedge*, since the corresponding Boolean rule would be

$$g7 \leftarrow (g6 \wedge (g4 \wedge \neg g5)) \vee (g8 \wedge (\neg g4 \wedge g5)).$$

Inference of higher-order interactions is known to be an even more challenging statistical problem than inference for pairwise interactions [20, 4]. Without trying to address this problem in generality, cell-specific networks allow us to disentangle interactions for a certain subclass of higher-order interactions: those which can be understood to be locally first-order, conditional on some latent variable that is encoded in the cellular context. Since most existing cell-specific network inference methods [15, 16] infer undirected edges only, such methods are unable to distinguish the direction of information flow in this example and thus would predict identical networks for either branch.

**Simulation** We simulate the network in Figure 3(a) using the BoolODE software package [5]. This uses a chemical Langevin equation scheme for simulating biologically plausible expression dynamics: $\mathrm{d}X_t = f(X_t)\mathrm{d}t + g(X_t)\mathrm{d}B_t$ where $f, g$ have specific forms dictated by the reaction network. From 1,000 independent realisations of the system's trajectories, we generate a set of 1,000 sampled cells by sampling a cell state from each trajectory at a time chosen uniformly at random. As a measure of ground truth interactions for each cell $x_i$, we compute the corresponding Jacobians $J_{ijk} = \partial_j f_k(x_i)$. In Appendix S2.1 we document the Boolean rules used to implement this network.

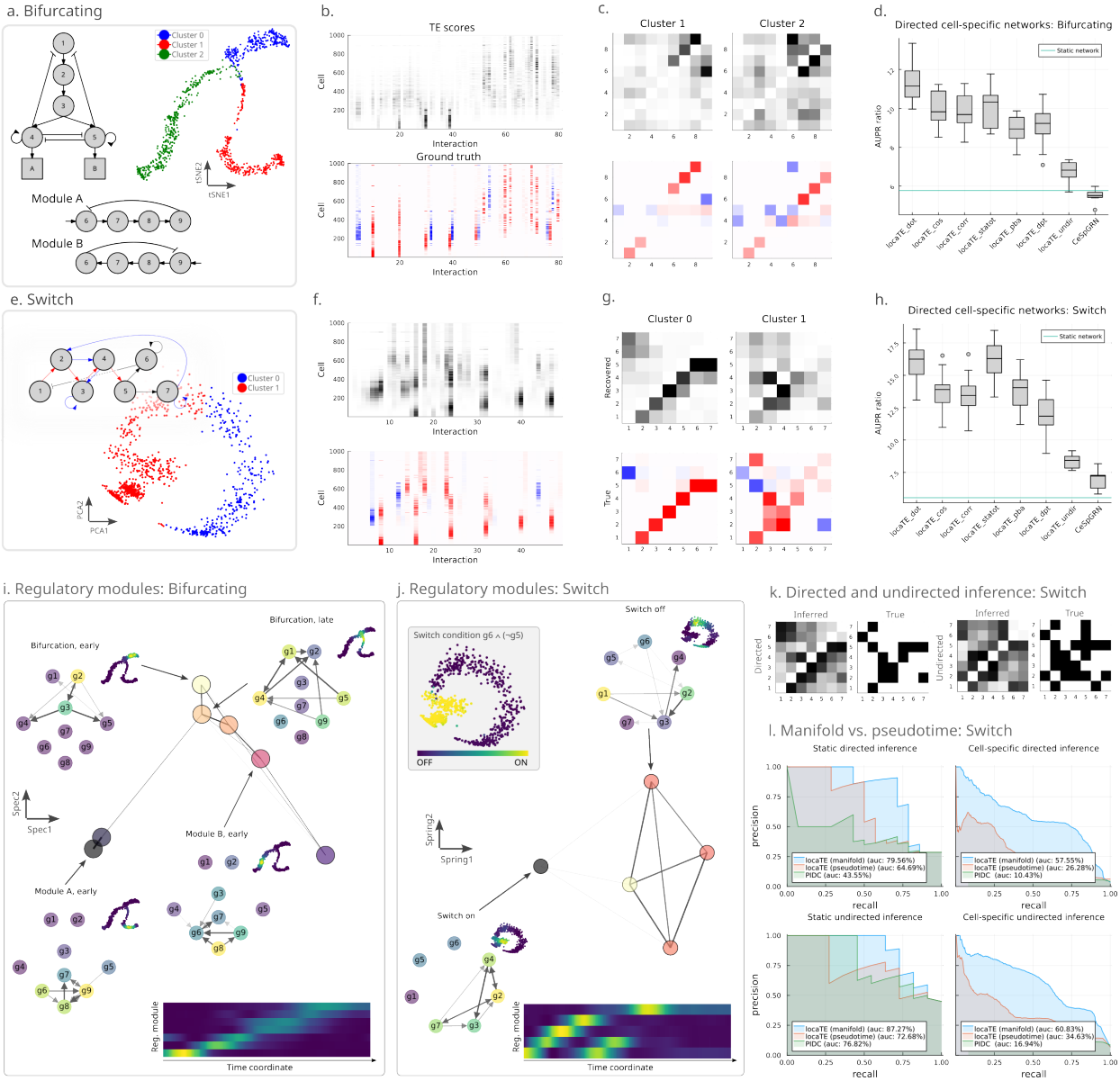



Figure 3: **Simulated networks.** **(a)** Bifurcating network: 5-gene bifurcating subnetwork feeding into two modules A, B involving the same 4 genes $g6, g7, g8, g9$. 4-gene modules A, B each involve different regulatory interactions involving the same species; tSNE embedding of simulated cell profiles, coloured by inferred diffusion pseudotime cluster. **(b)** LocaTE output as a $N \times M^2$ matrix in order of increasing pseudotime (top), shown alongside the ground truth Jacobians (bottom). **(c)** Interactions inferred by locaTE averaged by cluster (top) shown against ground truth interactions averaged by cluster. **(d)** AUPRC ratio for locaTE and other methods across 10 sampled simulation datasets, in the setting of cell-specific inference. We show as a horizontal line the AUPRC ratio achieved by perfect knowledge of the static network. **(e)** Switch network: Blue (resp. red) interactions are active when $(g6 \wedge (\neg g5))$ is FALSE (resp. TRUE). Notably, gene 2 interacts both directly and indirectly with gene 4, depending on the system state. **(f)** LocaTE output as a $N \times M^2$ matrix in order of increasing pseudotime (top), shown alongside the ground truth Jacobians (bottom). **(g)** Interactions inferred by locaTE averaged by cluster (top) shown against ground truth interactions averaged by cluster. **(h)** AUPRC ratio for locaTE and other methods across 10 sampled simulation datasets, in the setting of cell-specific inference. We show as a horizontal line the AUPRC ratio achieved by perfect knowledge of the static network. **(i)** Top: Regulatory modules for the simulated bifurcating network found by locaTE-NMF with $k = 8$, shown as a graph spectral layout. Selected modules are shown (top 10% of edges), along with the module activity on tSNE coordinates in inset. Bottom: smoothed module activities against simulated time. **(j)** Top: Regulatory modules for the simulated switch network found by locaTE-NMF with $k = 5$, shown as a graph spring layout. Selected modules are shown (top 20% of edges), along with the module activity on PCA coordinates in inset. Bottom: smoothed module activities against simulated time. **(k)** Comparison of population-averaged networks for switch system for inference of directed interactions utilising dynamical information (left) and inference of undirected interactions utilising only geometry (right). **(l)** Comparison of population-averaged and cell-specific inference, both in the directed and undirected settings, for switch system using graph construction from PCA coordinates (manifold) and pseudotime coordinates (pseudotime). We show for reference PIDC [4], an information-theoretic approach closely related to locaTE that does not utilise geometric information.

**Inferring dynamics** As input our method requires a cell-state transition matrix, $P$, which for instance can be obtained by using a variety of dynamical inference methods (see Section S1.2). For our simulated data, we construct $P$ from the velocity vectors of sampled cells in the expression space ("RNA velocity") as follows. For each observed cell, $x_i$, we calculate the corresponding velocity vector, $v_i = f(x_i)$. From this, we can construct transition probabilities to a neighbourhood of $x_i$ (in gene expression space) using a kernel function $K$:

$$P_{ij} = \frac{\exp\left(\sigma^{-1}K(x_j - x_i, v_i)\right)}{\sum_{k\in\mathcal{N}[i]}\exp\left(\sigma^{-1}K(x_k - x_i, v_i)\right)}\mathbf{1}\{j \in \mathcal{N}[i]\} \tag{10}$$

where $\mathbf{1}$ denotes the 0-1 indicator function. We set $\mathcal{N}$ to be the $k$-NN neighbourhood where we chose $k = 25$, and selected the bandwidth $\sigma$ following the median heuristic of [27]. Different choices of the function $K$ can be used: as studied in [48], $K(x, y)$ can be chosen as the inner product, $\langle x, y\rangle$, cosine similarity, $\cos(x, y) = \langle x, y\rangle/\|x\|\|y\|$, or Pearson correlation $\mathrm{corr}(x, y)$. While the inner product is sensitive to both the direction and magnitude of $v_i$, the cosine and correlation kernels are only sensitive to the direction.

In addition to approaches that directly quantify the vector field, we also consider inference methods such as a transition matrix construction based on diffusion pseudotime [54] corresponding to a random walk biased in the direction of increasing pseudotime, as well as the methods StationaryOT [45] and PBA [25], which were shown to have good performance for inferring stochastic dynamics without depending on pseudotime or velocity estimates. Finally, we also consider an undirected transition matrix which corresponds to inference of undirected interactions at equilibrium. We summarise the assumptions and limitations of dynamical inference approaches in Box 2 and provide a more detailed discussion in Section S1.2.

**Inferring interactions** We apply the method described in Section 2 with a backward matrix obtained via the transpose (see Section S1.3), for a range of parameter values, $(\tau, \lambda_1, \lambda_2)$. To construct the initial neighbourhood distribution, $\pi_0^x$, we use a bistochastic projection of the pairwise distance matrix [55], which produces a local neighbourhood density similar in effect to a truncated Gaussian. In Figure 3(b) we show the individual cell-specific networks obtained for $\tau = 3, \lambda_1 = 25, \lambda_2 = 10^{-3}$ represented as vectors against the pseudotemporal ordering, and in Figure 3(c) we show the averaged networks over the cell clusters corresponding to the two branches. We observe overall that the the de-noised result, $G$, resembles the ground truth. The averaged networks for module A and module B are directed and reflect the mirrored connectivities described in Figure 3(a). Averaging networks is convenient for summarising the inference outputs for cell-specific networks, but may obscure the cell-specific nature of the inference results. To illustrate this, in Figure S1(b) we show the interaction scores for the two edges $g6 \to g7, g7 \to g6$, which are specific to modules A and B, respectively. We observe that the flow of information between genes 6 and 7 is reversed on opposite branches, demonstrating that locaTE can resolve the cell-specific directionalities of edges.

**Comparison to other methods** CeSpGRN [15] is perhaps the method most similar to locaTE in terms of the problem it seeks to address: inferring a $N \times G \times G$ array of cell-specific gene-gene interaction scores. It is limited to undirected interactions due to use of the (Gaussian) Graphical Lasso. For a larger set of comparisons we also consider methods for static network inference. We consider three information-theoretic methods: TENET [21] and Scribe [19], both of which infer static directed networks from pseudotime-ordered single-cell data, as well as PIDC [4], which infers an undirected network using partial information decomposition. We also consider SINCERITIES, a method based on Granger causality for directed inference using pseudotime-ordered data [56] and GENIE3 [50], a random forest regression model for undirected inference, SCODE [22] and GRISLI [24] which are ODE-based models for directed inference. Among these methods, PIDC, GENIE3 and SINCERITIES were found to perform well in the benchmarking study [5].

**Cell-specific edge detection** For each triple $(i, j, k)$ with $i \in [\![N]\!]$ and $j, k \in [\![M]\!]$, we treat the problem of detecting an edge, $j \to k$, in cell $x_i$ as a binary classification problem with threshold $q$. To construct the set of true positives, we consider the matrix $\Pi J$, where $\Pi$ is a neighbourhood transition matrix such that $\Pi_{i\cdot} = \pi_0^{x_i}$. The motivation for considering this instead of simply using the raw Jacobians, $J$, is that cell-specific interactions are necessarily inferred using neighbourhood information. Strict cell-wise comparison to a ground truth would be overly stringent and sensitive to small perturbations in the expression space. By

computing $\Pi J$, the ground truth signal is smoothed over the same neighbourhood that was used for inference, allowing for more robust assessment of classifiers. We classify all potential edges for which $(\Pi J)_{ijk} > 0.5$ to be "true" edges.

In Figure S1(c) we show the Precision-Recall (PR) curves calculated for locaTE, as well as raw transfer entropy (TE), transfer entropy followed by CLR filtering (TE+CLR), and the method CeSpGRN [15], which infers cell-specific undirected networks. We find that the CLR filtering step improves upon raw TE scores in terms of prediction performance, and that locaTE improves substantially upon both TE and TE+CLR. This demonstrates that both the CLR filtering and the regularised regression step improve inference: this can be understood as a combined effect of (1) removal of indirect interactions using the CLR algorithm, and (2) sharing of information along the cell state manifold and attenuation of noise using the smooth and sparse regression. We also compute a ground-truth static network by aggregating the cell-specific ground truth networks and calculate a static network AUPRC, corresponding to the best-case performance for a static method, i.e. if all cells were estimated to have the same static network.

To investigate the performance of locaTE further we use the BoolODE package to generate 10 datasets using different kinetic parameters for each simulation [5, Methods]. Thus, variation across datasets results from both the random sampling of cells and differences in the simulated biophysics. For each dataset, we infer cell-specific networks using locaTE using several methods for trajectory inference. We consider three approaches using velocity information: dot-product, cosine similarity, and Pearson correlation (see Section S1.2 for details on the difference between each construction). We also consider StationaryOT and PBA, as well as a transition kernel based on diffusion pseudotime [54]. In Figure 3(d) we summarise performance in terms of AUPRC ratio, that is the AUPRC for inferred cell-specific networks relative to that of a random predictor.

These results show that the trajectory reconstructions used as input to locaTE can affect the inference result: the dot-product velocity kernel outperforms the cosine and correlation velocity kernels. This is not unexpected, since cosine and correlation kernels depend only on the orientation of the velocity vector and not on its magnitude – thus losing information. StationaryOT performs similarly to cosine and correlation kernels, and the PBA and the pseudotime kernels perform worst among the dynamical inference approaches. We find that all locaTE-based methods using dynamical information outperform locaTE with an undirected transition matrix (locaTE-undir), which in turn outperforms the undirected method CeSpGRN. This suggests that the availability of even potentially poor dynamical information can lead to improved inference of directed interactions. All methods perform at least as well as the static network baseline, illustrating the value of cell-specific network inference methods in general.

**Static networks as aggregates** As most methods in the literature deal with static inference problems, we investigate the performance of locaTE by averaging local networks across cells to form a static network (see Figure S1(e) for representative examples). We show AUPRC ratios in Figure S1(d) for networks inferred by locaTE, as well as other selected static inference methods. In this setting we find that locaTE performs similarly well across trajectory inference methods, and that CeSpGRN performs marginally better. This can be understood, at least in parts, because the aggregated networks for modules A and B together in Figure 3(a) would be indistinguishable from the corresponding undirected network (as we show in the next example, this is not always the case). Competing static methods all perform worse, with GENIE3 and PIDC performing best, and SCODE performing worst, broadly in keeping with the conclusions of [5]. In Figure S1(e) we provide examples of the inference output of the considered methods, together with the ground truth.

**Downstream factor analysis** Instead of using the de-noising regression described above, we can apply techniques from factor analysis to the tensor of TE values. Methods such as non-negative matrix factorisation (NMF) and non-negative tensor factorisation (NTF) seek a simplified, low-rank description of the data. In the context of cell-specific GRNs, we are faced with an $N \times G \times G$ array. By applying NMF (or NTF) to the flattened $N \times G^2$ matrix (resp. $N \times G \times G$ tensor), we effectively seek groups or *modules* of interactions that are active together in some region of the cell state space; here the regulatory state of a given cell is represented as a linear combination of some inferred archetypal regulation patterns. We derive a numerical scheme for finding matrix and tensor low-rank representations of the TE score matrix that incorporates the same smoothing and sparsity priors as (9). This can be used as a drop-in replacement for the (full-rank) problem (9). We describe these in detail in Appendix S1.5 (NMF) and Appendix S1.6 (NTF).

In Figure 3(i) we illustrate the utility of factor analysis for finding simplified representations of regulatory interactions inferred using locaTE for the bifurcating system. To choose the number of factors $k$ we employ the following heuristic: compute the singular values, $\sigma_1 \leqslant \sigma_2 \leqslant \ldots \leqslant \sigma_{\min\{N,G^2\}}$, of the full-rank matrix of TE values, and choose the largest $k$ such that $(\sum_{i=1}^{k} \sigma_i^2)/(\sum_i \sigma_i^2) \leqslant \alpha$. Applying locaTE-NMF with $\alpha = 0.8$, we find $k = 7$ regulatory modules whose activities are localised in the cell state manifold. Following a similar approach to [57], an undirected graph is constructed from the locaTE-NMF gene regulatory modules with edge affinities calculated from the Wasserstein distance between the normalised module activities. We present a spectral layout of this graph, which exhibits the bifurcating structure that reflects the underlying process. Examining the top edges of selected modules shows that key portions of the regulatory dynamics are recovered. Two distinct phases of the bifurcation are found: the first corresponding activation of genes 4 and 5 by gene 3 ("Bifurcation, early"), and the second corresponding to deactivation of gene 1 by genes 4 and 5 ("Bifurcation, late"). Along each of the branches, locaTE-NMF uncovers elements of the different regulatory logics that involve the same genes ("Module A", "Module B").

Tensor factorisation differs from matrix factorisation in that the learned atoms have the additional constraint of being separable. Because of this added structure, tensor decompositions are often found to have better properties (identifiability of factors, robustness to noise [58]). We apply locaTE-NTF to find 16 rank-one factors of the cell-specific network tensor (see Figure S2), and we apply $k$-means clustering to group the rank-one atoms into 8 clusters based on their activities. We find that the resulting clusters largely resemble the ones found by NMF. Unlike NMF however, the clusters can be decomposed into smaller rank-one components, and appear less noisy than the atoms found by NMF owing to the additional constraint. In Figure S2(c) we show how two example clusters from the NTF representation split into rank-one components, demonstrating that the rank constraint leads to finer grained representations of regulatory states.

## 3.2 Simulated data: switch-like behaviour

We next consider a scenario where differential regulation occurs along a linear trajectory instead of a bifurcation. This inference problem is more subtle than a bifurcating process given the lack of a clear branching or clustering structure upon which regulatory interactions depend. We consider a model network depicted in Figure 3(e): this network either contains a feed-forward loop $g2 \rightarrow g3 \rightarrow g4$, or $g2 \rightarrow g4 \rightarrow g3$, depending on the position of a cell along the trajectory. The interaction $g2 \rightarrow g4$ is first *indirect* and then becomes *direct*: the nature of this differential regulation is such that static GRN inference methods would observe a superposition of these interactions, i.e. $g2 \rightarrow g3 \rightarrow g4$ as well as $g2 \rightarrow g4$. As an example, this could potentially lead to the $g2 \rightarrow g4$ interaction being erroneously identified as an indirect interaction and thus being pruned, if filtering heuristics such as CLR or the data-processing inequality were used [47, 21]. Cell-specific networks may be able to disentangle these two effects.

We implement this network in BoolODE and simulate 1,000 cells. As before, we apply locaTE using a transition kernel derived from the simulated velocity information and the same parameters as in the bifurcating dataset. We find from Figure 3(f, g) that the inferred interactions accurately capture the true directed interactions. We demonstrate that locaTE significantly outperforms competing methods in terms of AUPRC for cell-specific inference in Figure 3(h) and for static inference in Figure S3(c). We find that locaTE substantially outperforms CeSpGRN in both settings for this example, in contrast to the previous example where CeSpGRN performed well for static inference.

We next apply locaTE-NMF with $\alpha = 0.8$ to find $k = 5$ regulatory modules following the same procedure as for the bifurcating example. In Figure 3(j) we show a force-directed layout of the regulatory module graph, upon which we label modules that correspond to the regulatory logic when the switch condition $g6 \wedge (\neg g5)$ is, respectively, on and off. In contrast to the visualisation in Figure 3(g), which made use of a separate clustering step, this demonstrates that locaTE-NMF can identify distinct regulatory states in an unsupervised manner independent of clustering.

In Figure 3(k) we demonstrate using the switch-like network the two settings in which locaTE can be applied. Where information on the dynamics is available, directed interactions are successfully inferred using a transition matrix that captures these dynamics. When no dynamical information is available, a transition matrix reflecting only neighbourhood information can be used to infer undirected interactions.

We also investigate the importance of accounting for the cell state manifold. The switch-like network exhibits a linear trajectory topology, allowing for a fair comparison against pseudotemporal ordering. We

applied locaTE to the same simulated data as previously, but constructed neighbourhood kernels and $k$-NN graphs based on the true simulation time of each sampled cell instead of PCA coordinates. We reasoned that using the ground-truth simulation time for the ordering provided a "best-case" scenario for pseudotime-inference approaches. In Figure 3(j) we compare the performance of locaTE using PCA-based manifold approximation (locaTE-manifold) against locaTE using pseudotime neighbourhoods (locaTE-pseudotime) as well as PIDC [4] as an additional baseline. We consider the directed setting, using velocity information with the dot-product kernel, as well as the undirected setting using an undirected transition matrix. Parameters for each case are the same as previous. LocaTE-manifold outperforms locaTE-pseudotime significantly in all cases: the performance gain is especially dramatic for cell-specific inference, with the AUPR roughly doubling. Of particular interest is that locaTE-manifold outperforms PIDC for undirected, static network inference: locaTE-manifold does not use velocity information in this setting, and so has exactly the same input data as PIDC. This provides further evidence that modelling appropriately the underlying geometry cell state is crucial to obtaining good results for network inference.

Together, our simulated network analyses demonstrate that locaTE is able to combine dynamical and geometric information to improve inference of both cell-specific and static networks, often with a substantial margin compared to existing approaches. Unsurprisingly, better estimates of dynamics lead to better results for network inference. Perhaps what is more remarkable is that, even after aggregating cell-specific networks produced by locaTE, the resulting static networks generally outperform competing static inference methods (see Figures S1(d, e), S3(d, e)).

## 3.3 mESC differentiation dataset using StationaryOT

We first consider the dataset of Hayashi et al. (2018) [59], consisting of 456 cells and 100 TFs. This dataset is a time-series of mouse embryonic stem cells differentiating through to primitive endoderm, sampled at 0, 12, 24, 48, 72h. Cells in this differentiation process follow a linear trajectory, as can be seen in Figure 4(a). To infer trajectories, we apply StationaryOT [45] using quadratically-regularised optimal transport to ensure a sparse transition kernel (see Section S2.5 for details). Using this transition matrix, we infer cell-specific networks with locaTE, resulting in a $456 \times 100 \times 100$ array of interaction scores.

Since no ground truth is available for cell-specific networks in this experimental dataset, we are only able to validate the output of locaTE against a static benchmark. In [23] the authors constructed a reference network from known regulatory interactions in the embryonic stem cell atlas from pluripotency evidence (ESCAPE) database [60], which consists of 19 regulators and 100 targets. Starting from the array of cell-specific interaction scores, we obtain a static $100 \times 100$ GRN by averaging interaction scores across cells and compute precision-recall curves relative to the interactions in our benchmark network (Figure 4(c)). For comparison, we run CeSpGRN, which infers cell-specific undirected networks, and a collection of static inference methods (TENET [21], SCODE [22], GRISLI [24], PIDC [4] and SINCERITIES [56]). We compare a $19 \times 100$ subnetwork of the inferred networks to the ESCAPE reference. In the directed setting, we find that locaTE outperforms other methods in terms of AUPRC, and performs very well in terms of EP (0.542, compared to a best EP of 0.553 for TENET and baseline of 0.343). For undirected inference locaTE is out-performed by SCODE (AUPRCs of 0.450 and 0.483 respectively). Together with the observation that SCODE performs poorly for directed inference, this suggests that, despite being a method for directed inference, SCODE may be prone to detecting interactions but with the wrong directionality; PIDC [4] is the second best performing method in terms of AUPRC (0.456).

Since locaTE infers cell-specific networks, we can ask questions at finer resolution than possible at the population level. To demonstrate the power of this type of analysis, we use locaTE-NMF with $\alpha = 0.75$ to find $k = 10$ regulatory modules that are active at different stages along the trajectory (Figure 4(d)). Hierarchical clustering of the inferred factor activities produces five groups, corresponding to "Early", "Mid(1,2)", and "Late(1,2)" trajectory stages. For each cluster we compute a representative network by taking weighted averages of NMF factors, and the top 0.25% of edges for each stage are shown in Figure 4(e). It is clear that the nature of the inferred regulatory interactions changes drastically between stages. Included among the top inferred interactions are the canonical ones previously reported to be associated with maintenance of pluripotency and embryonic development, including Nanog-Zfp42 [61], Foxd3-Nanog [62] and Sall4-Klf2 [63] which are active early in the trajectory. Top inferred interactions for the late trajectory include Epas1 (Hif2a)-Pou5f1 (Oct4), which is known to play a role in embryonic development [64]; and Epas1 (Hif2a)-Bhlhe40

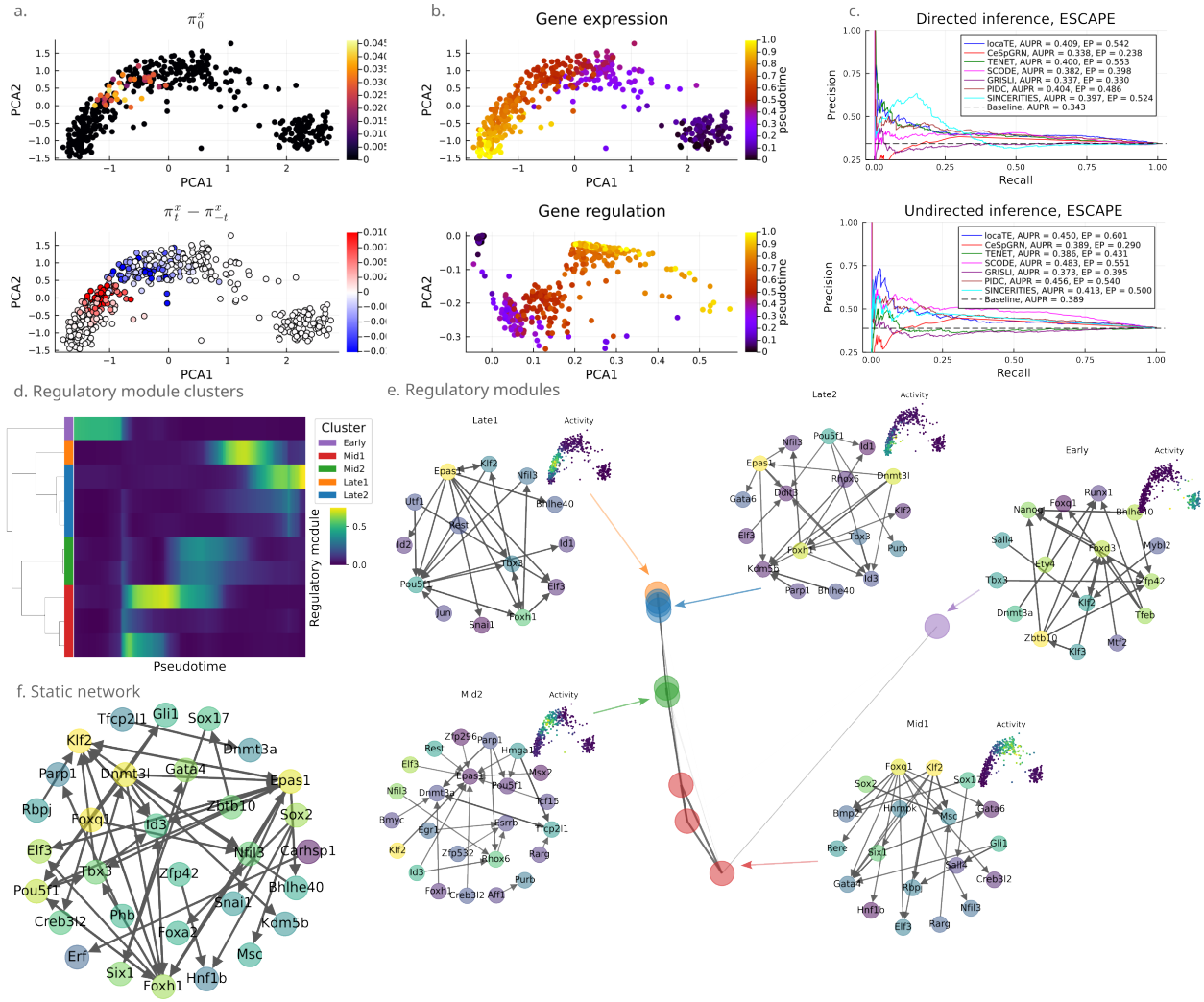



Figure 4: **ES-endoderm development dataset.** **(a)** PCA embedding of 456 cells along an ES-endoderm differentiation trajectory, coloured by (top) example neighbourhood density (bottom) change in density along evolution of Markov process. **(b)** PCA embedding of (top) expression profiles (bottom) cell-specific interaction networks, coloured by diffusion pseudotime. **(c)** Precision-recall curves against the ESCAPE reference for static inference, both directed and undirected. **(d)** Regulatory module activities found using locaTE-NMF, clustered by activity along the trajectory. **(e)** Graph layout of regulatory modules found by locaTE-NMF coloured by cluster together with inferred networks (top 0.25% of all edges by confidence) averaged over each cluster. Network nodes (TFs) are coloured by out-edge eigenvector centrality. Overall activity for each cluster of modules is shown (inset). **(f)** Static network produced by averaging locaTE cell-specific networks (top 0.5% of all edges by confidence).

(Dec1), which plays a role in suppression of Deptor [65], a stemness factor known to be downregulated during ES differentiation [66].

For each stage, we rank potential regulators by their out-edge eigenvector centrality and show the 25 top-ranked genes for each stage (see Figure S11(a)). We find that many of the top genes are known to be associated with stem cell development, including Nanog, Pou5f1 (Oct4), Sox2 [62], Gata6 [67], Zfp42 [61]. These gene lists are input to Metascape [68] to find corresponding Gene Ontology (GO) annotations (Figure S11). Among the gene set annotations found for the "early" and "mid1" trajectories, we find that enriched pathways with the highest confidence are those associated with pluripotency and embryonic morphogenesis. Annotations for the "mid2" and "late(1,2)" trajectories include epithelial cell differentiation and organ development. This reflects the developmental progression of embryonic stem cell differentiation to primitive endoderm and suggests that locaTE is able to identify key molecular drivers of the underlying process.

## 3.4 Pancreas development dataset with RNA velocity

Next we consider the murine pancreatic development dataset of Bastidas-Ponce et al. [69]; here cellular dynamics are captured in the form of RNA velocity estimates. The dataset consists of 2,531 cells, and we filter for 82 transcription factors following the preprocessing of [24]. We apply the scVelo [37] package to produce RNA velocity estimates, and CellRank [27] with a cosine kernel to produce a transition matrix encoding the inferred dynamics on the cell set.

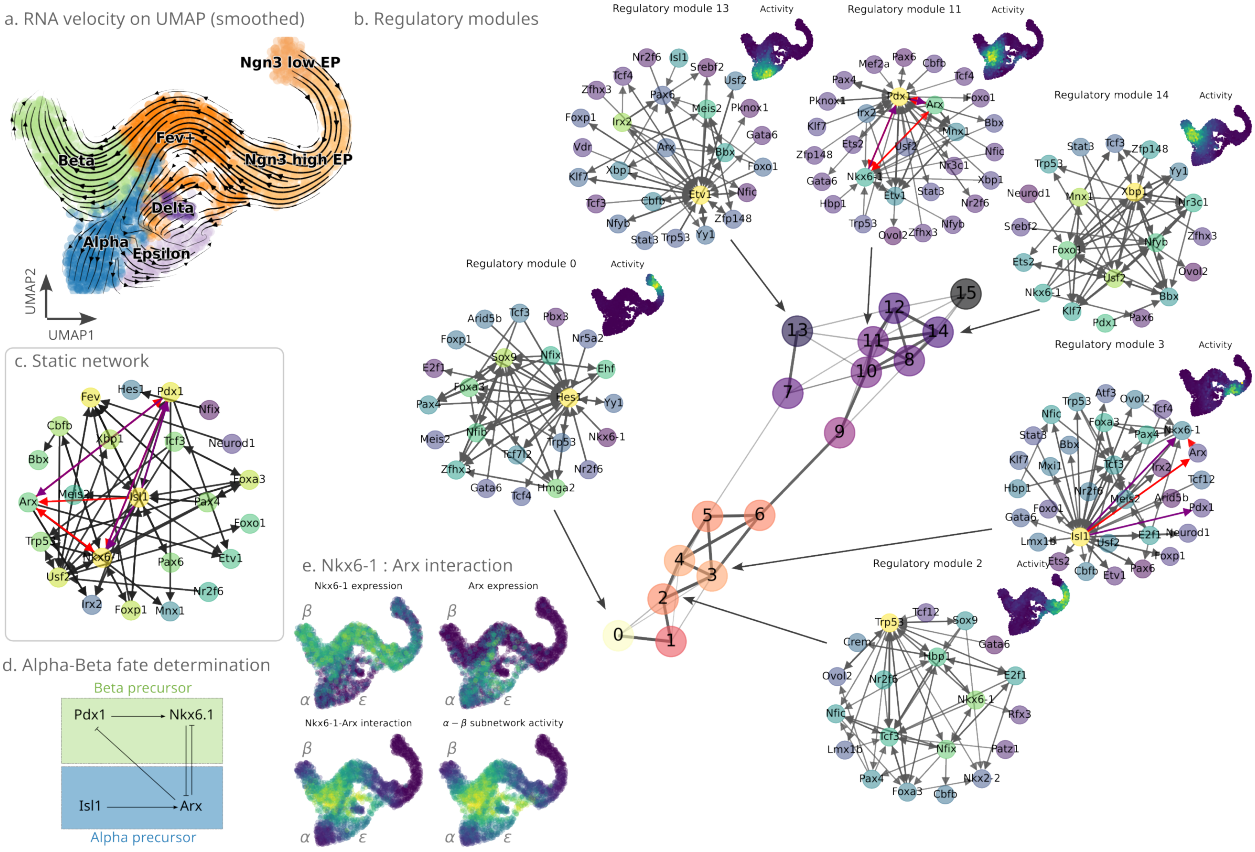



Figure 5: **Pancreatic development dataset.** **(a)** UMAP embedding of murine pancreatic development dataset [69], with smoothed RNA velocity vector field shown. Cells are coloured according to their cluster annotation. **(b)** Graph layout of regulatory modules found using locaTE-NMF, coloured by proximity to early trajectory. Selected regulatory modules are shown (top 1% of edges shown by confidence). **(c)** Static GRN constructed as population average of cell-specific networks (top 1% of edges by confidence shown, after filtering for stress response genes). **(d)** Nkx6.1-Arx expression and interaction scores in UMAP coordinates, and average activity of the 4-gene $\alpha/\beta$ subnetwork of [70]; **(e)** Alpha-beta fate determination network of [70].

We apply locaTE with the RNA-velocity derived transition matrix, and obtain a $2,531 \times 82 \times 82$ array of cell-specific interaction scores. In Figure 5(a) we show a UMAP embedding of the full dataset with streamlines visualising RNA-velocity dynamics. To illustrate the heterogeneity in the networks inferred by locaTE, we also colour cells by their corresponding cell-specific GRN using the first diffusion component of the inferred cell-specific networks (see Figure S13(b)). This reveals a clear gradient along the differentiation trajectory, suggesting that the networks inferred by locaTE capture a progression in regulatory interactions over developmental time. The highest-scoring interactions involve the genes Fos and Egr1, which are known to play a role in stress response [71]. Fos expression has been documented to be an dissociation-induced artifact in scRNA-seq [72], and we therefore conservatively filter out interactions involving Fos or Egr1 in subsequent analyses.

As a first summary of the inference output we produce an $80 \times 80$ static GRN by population averaging. Figure 5(c) shows the top 1% of edges from this network. We observe that the genes with highest out-edge eigenvector centrality include Nkx6-1, a known master regulator of pancreatic beta cell development [70, 73], as well as Isl1 and Pdx1, known determinants of alpha and beta cell fate [74]. In [70] the authors report evidence for a 4-gene network (henceforth referred to as the $\alpha/\beta$ network) involving Nkx6-1, Pdx1, Isl1 and Arx, that drives alpha and beta fate determination (Figure 5(d)). All 4 genes have high out-edge centrality in the inferred static network, and, furthermore, we find that the $\alpha/\beta$ network is recapitulated in the top-1% edges inferred network, as highlighted by the coloured edges in Figure 5(b, c). Comparisons to the networks inferred by PIDC and TENET (Figure S14), we find that both approaches also detect aspects of the same fate-determination network, albeit with some features missing (e.g. Isl1-Arx does not appear among the top 1% of edges).

Applying locaTE-NMF with $\alpha = 0.75$, we find 16 regulatory modules spanning the set of transcriptomic states. Computing pairwise similarities between regulatory modules as previously, we show in Figure 5(b) a graph layout of the learned regulatory modules in which each node is coloured by its activity over pseudotime. From these we illustrate six modules (top 1% edges by confidence), which correspond to key features of the developmental trajectory. Regulatory module 0 is active near the Ngn3− EP (endocrine progenitor) cell cluster, including a predicted interaction Sox9–Hes1. Sox9 is known to play a key role in maintaining the pancreatic progenitor cell state by regulating Hes1 [75], an inhibitor of Neurog3 (Ngn3) [74]. This interaction is no longer present in regulatory module 2, which is most active near the Ngn3+ cell cluster, further evidence for its role in maintaining the Ngn3− state. This interaction also does not appear in the static GRN in Figure 5(b), illustrating the need for cell-specific networks in capturing localised signals, which would be lost in static networks or population averaging. Out-edge eigenvector centrality of module 0 identifies additional potential regulators including Foxa3, known to be implicated in endodermal organ formation [76]; and Hmga2, a chromatin-associated protein expressed during development that does not directly regulate downstream genes, but appears to have widespread indirect effects via chromatin interactions [77].

Modules whose activities correspond to alpha and beta committed cells yield networks that are distinct from each other as well as from the progenitor clusters. The activity of module 13 is concentrated near the alpha cluster: examining out-edge eigenvector centrality, we find that Irx2 is featured prominently, and has been suspected to be involved in alpha fate specification [78]. In module 14, which is concentrated near the beta cluster, we find that Xbp1 is a top regulator recently found to be critical to maintenance of beta-cell identity [79].

Not all regulatory modules correspond to annotated cell type clusters, however: for instance, regulatory interactions that drive fate determination may instead be active at the boundary between cell states (thought of as a saddle point between two attractors in the Waddington's landscape, see Box 1). We highlight module 11, which is active at the boundary between the alpha and beta lineages. In this module we find interactions between Pdx1, Arx and Nkx6.1 to be among the top inferred edges. This module is consistent with a fate decision boundary between the two lineages. Additionally in module 3, which is located between the Ngn3+ EP and Fev+ clusters, Isl1 is identified as a top regulator along with Isl1-Arx. This agrees with the early behaviour of the fate determination network of [70] in which Isl1 activates Arx (consistent with module 3), which in turn competes with Nkx6.1 (consistent with module 11).

To investigate this at a finer scale we visualise expression levels and interaction scores for the Nkx6.1-Arx interaction in Figure 5(e). As before, we observe a fate decision boundary from expression levels that reflects the competition between Nkx6.1 and Arx, corresponding to beta and alpha fate. The interaction scores for both the Nkx6.1–Arx edge and the entire $\alpha/\beta$ subnetwork are also strongest near this boundary.

Modelling cellular dynamics as Markov processes naturally allows estimation of fate commitment in terms of absorption probabilities [25, 27, 45] and so examining correlation of gene interactions with fate commitment can be used to construct a putative "lineage driver" network that is correlated to a cell fate of interest. In Figure S13(d) we construct such networks for the alpha, beta, and epsilon cell lineages. As we expect, the $\alpha/\beta$ network is strongly correlated to beta cell fate. We find that Isl1 and Nkx-6.1 have higher out-edge eigenvector centrality towards the alpha and beta lineages respectively, reflecting their roles in the known network.

## 3.5 Haematopoiesis with extensions of RNA velocity

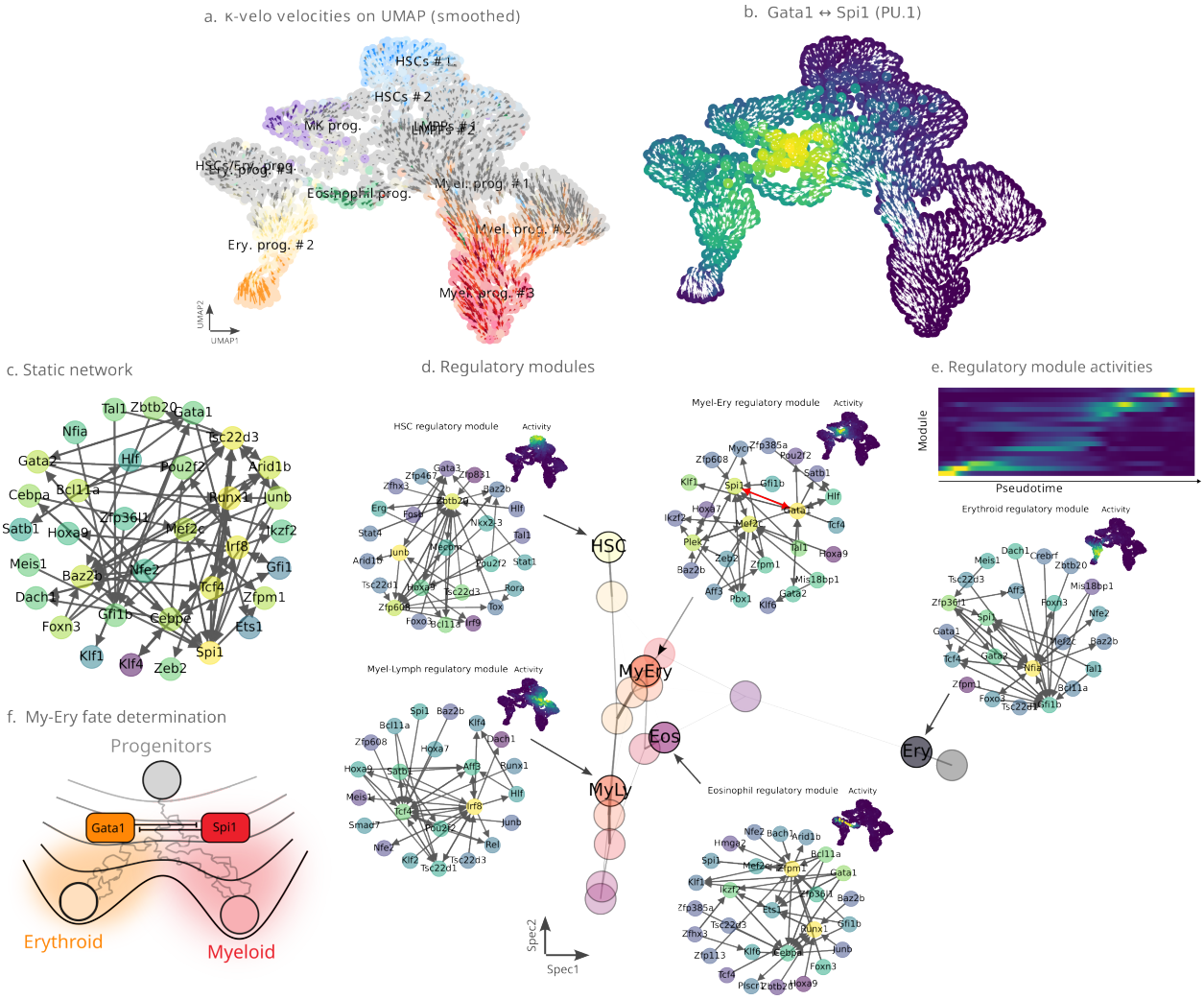



Figure 6: **Early-stage haematopoiesis.** **(a)** Smoothed velocities found by $\kappa$-velo shown on UMAP coordinates, coloured by cluster annotation. **(b)** Average locaTE score for Gata1-Spi1 interactions shown on UMAP coordinates along with smoothed velocities. **(c)** Static network (top 1% of all filtered edges) obtained as population average of cell-specific networks. **(d)** Graph layout of regulatory modules found by locaTE-NMF coloured by proximity (measured by mean first passage time) from the HSC module, shown alongside representative networks of selected modules (top 0.5% of all filtered edges) corresponding to HSC, Myeloid-Erythroid decision boundary, Erythroid, Eosinophil and Myeloid-Lymphoid. **(e)** Smoothed regulatory module activities shown against pseudotime. **(f)** Myeloid-Erythroid fate determination schematic.

One advantage of employing a Markov modelling framework is that it is agnostic to how the dynamics are

inferred. Currently, inference, interpretation and downstream use of RNA velocity estimates is a contentious topic [80, 81] and remains an active research area. Marot-Lassauzaie et al. [82] propose $\kappa$-velo, an alternative method for estimating RNA velocity and provide an example application to a haematopoiesis dataset, where the more established method scVelo [37] fails to recover expected trajectories. We apply locaTE to this dataset using inferred dynamics computed by $\kappa$-velo, producing a $2,430 \times 100 \times 100$ array of cell-specific interaction scores (see Section S2.7 for details). Initial observation of the inferred networks showed that the top inferred interactions were dominated by a handful of TFs: Hmgb2, Hmgb3, Myb, and Myc, all of which are known to be implicated in haematopoiesis [83, 84, 85]. For visualisation purposes we excluded interactions involving these TFs to prevent overcrowding of the displayed networks.

A bifurcation in the trajectories between myeloid and erythroid lineages is a prominent feature in this dataset (see Figure 6(a)). Erythroid-myeloid cell fate determination is governed by a genetic switch comprised of mutually antagonistic Gata1 and Spi1 (known also as PU.1) [86] (see illustration in Figure 6(f)). In Figure 6(b) we show the locaTE interaction score for the Gata1–Spi1 interaction alone along with the smoothed RNA velocity vector field. We find that the peak activity corresponds to a "dead zone" in the velocity field, while in the surrounding regions downstream, the velocity fields are coherent and indicative of lineage commitment. This is of practical interest [82], as one interpretation of this could be that it corresponds to cells which have high fate plasticity and are still in the "decision-making" process at a saddle node in the epigenetic landscape.

We next apply locaTE-NMF with $\alpha = 0.75$ to find $k = 18$ regulatory modules, and in Figure 6(d) we show a layout of the graph constructed from modules in the same manner as previously done in Figure 5 (see Figure S16(a) for all modules). From this we can see two clear branches corresponding to erythroid and myeloid fate. We find a module (see Figure 6(d), labelled Myel-Ery) that appears concentrated near the bifurcation point on the UMAP coordinates. This module is centered around activity of Spi1 and Gata1, as well as Mef2c, which is known to be involved in lymphoid-myeloid fate determination [87], and locaTE-NMF predicts the key Gata1-Spi1 (PU.1) interaction (edge highlighted in red) as one of the top interactions. In Figure 6(d, e) we display other regulatory modules that correspond to various degrees of fate commitment, including haematopoietic stem cells (HSCs). These networks uncover known hallmarks of haematopoietic development: the myeloid-lymphoid progenitor (labelled Myel-Lymph) module prominently features Irf8, understood to play a key role in the development of myeloid and lymphoid lineages [88, 89] (note that the lymphoid lineage is not captured in this dataset). The erythroid module is centered on a different set of genes such as Gfi1b [90]. The HSC module features Junb prominently, which has been implicated in homeostasis in the long-term HSC population [91]. The HSC module also contains interactions involving Baz2b, a master regulator which has been found to reprogram lineage-committed cells to a multipotent state [92].

In the static network produced by aggregating locaTE cell-specific networks (see Figure 6(c)) we find that interactions such as Gata1-Spi1 are obscured due to the population averaging. Comparing to the output of PIDC and TENET (Figure S17), we find that PIDC successfully uncovers this interaction within the top 1% confident interactions, while TENET misses it. On the other hand, Gata1–Spi1 features prominently in the interaction module uncovered by locaTE, with Gata1 and Spi1 being among the top four nodes by out-edge centrality in that module. Similarly, in Figure S16 we show cluster-specific networks produced by aggregating locaTE outputs across annotated cell types, and it is clear that Gata1-Spi1 does not appear among the top cluster-averaged interactions. This again illustrates the potential for important interactions to occur at the boundary between two defined cell types (and therefore missed by cluster-wise analyses), and the value of locaTE-NMF as an unsupervised approach for finding them. Overall, these results highlight the utility of cell-specific inference methods such as locaTE, which are able to learn interpretable representations of regulatory states without the need for clustering.

# 4 Discussion

Biology is dynamic. A static representation of a network can never do justice to what we observe phenotypically. Discerning molecular causes requires experimental and theoretical advances to progress in lock-step. The power of single cell data, and the insights we can gain from such data, rely on flexible statistical analyses that allow us to dissect the differences between gene regulatory networks (and programs) and how they differ between cells. Gaining deeper mechanistic insights into single-cell data requires a framework to reconstruct cell-specific, causal networks. Our method, locaTE, computes the transfer entropy (TE) [30, 41] for pairs

of genes to detect causal relationships. The approach uses approximations of the underlying dynamics but makes no extraneous assumptions on linearity of interactions, the distribution of gene expression counts, or the topology of the trajectory.

By using neighbourhoods in gene expression space we can infer cell-specific networks without going to the extreme $N = 1$ case. How to pool information across cells most efficiently will ultimately decide on the success of such methods and the insights they provide. Temporal ordering helps enormously, but we are still, by and large, not able to collect meaningful single cell transcriptomic time-course data. Pseudo-temporal ordering or some of the many versions of RNA velocity entering the literature can provide useful information, as can, related multi-omic data, such as scATAC-seq.

LocaTE is not limited any individual dynamical inference method: while RNA velocity data are currently among the most abundant in the literature, metabolic labelling datasets hold the promise of more accurate vector field quantification. Nor is our approach limited to inference for transcription factor networks: as an information theoretic network inference approach, locaTE aims to uncover information flow between observed random variables. Thus locaTE can in principle be applied to any gene expression data, but also surface marker and protein expression data. As an illustration, in Section S2.8 we applied locaTE to a metabolic labelling dataset of the cell cycle [93] to infer context-specific and causal coexpression networks between 200 genes found to be highly correlated with cell cycle. This highlights the flexibility of locaTE: using a learned vector field based on metabolic labelling data [28], non-linear trajectories can be handled with ease.

Directions for future work include investigating approaches using which local information can help dealing with technical zeros (dropout) in single cell data. We investigated the application of data-imputation methods for the pancreatic dataset; however, we found that imputation worsened performance for all inference methods considered. This effect has been reported previously to be due to spurious signals being amplified [94] between non-interacting genes.

The problem of inferring cellular dynamics, upon which application of methods such as locaTE depend, is far from solved. In particular, accurately modelling and quantifying RNA velocity remains challenging in practice [80, 48]. Data from additional modalities will help greatly: metabolic labelling assays can provide a more accurate source of dynamical information [28], and lineage-resolved data [95] can provide dynamical information beyond that which can be obtained from gene expression data alone. Since locaTE requires only that dynamics are described by a Markov process, we expect that advances in trajectory inference can be used to inform cell-specific network inference. Finally, a further route for improving locaTE and network inference in general is to use relevant prior information to guide network inference.

# Acknowledgements

SYZ would like to thank Geoffrey Schiebinger, Léo Diaz and Wilder Scott for insightful discussions. The first author acknowledges support from the Australian Government Research Training Program and an Elizabeth and Vernon Puzey Scholarship. MPHS gratefully acknowledges funding from the Australian Research Council through a Laureate Fellowship (FL220100005), and the University of Melbourne through DRM funding.

# References


[1] Florian Buettner, Kedar N Natarajan, F Paolo Casale, Valentina Proserpio, Antonio Scialdone, Fabian J Theis, Sarah A Teichmann, John C Marioni, and Oliver Stegle. Computational analysis of cell-to-cell heterogeneity in single-cell rna-sequencing data reveals hidden subpopulations of cells. *Nature biotechnology*, 33(2):155–160, 2015.

[2] Sophie Tritschler, Maren Büttner, David S Fischer, Marius Lange, Volker Bergen, Heiko Lickert, and Fabian J Theis. Concepts and limitations for learning developmental trajectories from single cell genomics. *Development*, 146(12):dev170506, 2019.

[3] Etienne Becht, Leland McInnes, John Healy, Charles-Antoine Dutertre, Immanuel WH Kwok, Lai Guan Ng, Florent Ginhoux, and Evan W Newell. Dimensionality reduction for visualizing single-cell data using umap. *Nature biotechnology*, 37(1):38–44, 2019.

[4] Thalia E Chan, Michael PH Stumpf, and Ann C Babtie. Gene regulatory network inference from single-cell data using multivariate information measures. *Cell systems*, 5(3):251–267, 2017.

[5] Aditya Pratapa, Amogh P Jalihal, Jeffrey N Law, Aditya Bharadwaj, and TM Murali. Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. *Nature methods*, 17(2):147–154, 2020.

[6] Bradley Efron and Trevor Hastie. *Computer Age Statistical Inference, Student Edition: Algorithms, Evidence, and Data Science*, volume 6. Cambridge University Press, 2021.

[7] Michael PH Stumpf. Inferring better gene regulation networks from single-cell data. *Current Opinion in Systems Biology*, 27:100342, 2021.

[8] Kyle Akers and TM Murali. Gene regulatory network inference in single-cell biology. *Current Opinion in Systems Biology*, 26:87–97, 2021.

[9] Shuonan Chen and Jessica C Mar. Evaluating methods of inferring gene regulatory networks highlights their lack of performance for single cell gene expression data. *BMC bioinformatics*, 19(1):232–221, 2018.

[10] Patrick S Stumpf, Rosanna CG Smith, Michael Lenz, Andreas Schuppert, Franz-Josef Müller, Ann Babtie, Thalia E Chan, Michael PH Stumpf, Colin P Please, Sam D Howison, et al. Stem cell differentiation as a non-markov stochastic process. *Cell Systems*, 5(3):268–282, 2017.

[11] Elias Ventre, Ulysse Herbach, Thibault Espinasse, Gérard Benoit, and Olivier Gandrillon. One model fits all: combining inference and simulation of gene regulatory networks. *bioRxiv*, 2022.

[12] Alsu Missarova, Leah Ursula Rosen, Emma Dann, Rahul Satija, and John Marioni. Sensitive cluster-free differential expression testing. *bioRxiv*, pages 2023–03, 2023.

[13] Anna SE Cuomo, Tobias Heinen, Danai Vagiaki, Danilo Horta, John C Marioni, and Oliver Stegle. Cellregmap: a statistical framework for mapping context-specific regulatory variants using scrna-seq. *Molecular Systems Biology*, 18(8):e10663, 2022.

[14] Robrecht Cannoodt, Wouter Saelens, Louise Deconinck, and Yvan Saeys. Spearheading future omics analyses using dyngen, a multi-modal simulator of single cells. *Nature Communications*, 12(1):1–9, 2021.

[15] Ziqi Zhang, Jongseok Han, Le Song, and Xiuwei Zhang. Inferring cell-specific gene regulatory networks from single cell gene expression data. *bioRxiv*, 2022.

[16] Xuran Wang, David Choi, and Kathryn Roeder. Constructing local cell-specific networks from single-cell data. *Proceedings of the National Academy of Sciences*, 118(51), 2021.

[17] Hao Dai, Lin Li, Tao Zeng, and Luonan Chen. Cell-specific network constructed by single-cell rna sequencing data. *Nucleic acids research*, 47(11):e62–e62, 2019.

[18] Federico Bocci, Peijie Zhou, and Qing Nie. splicejac: transition genes and state-specific gene regulation from single-cell transcriptome data. *Molecular Systems Biology*, 18(11):e11176, 2022.

[19] Xiaojie Qiu, Arman Rahimzamani, Li Wang, Bingcheng Ren, Qi Mao, Timothy Durham, José L McFaline-Figueroa, Lauren Saunders, Cole Trapnell, and Sreeram Kannan. Inferring causal gene regulatory networks from coupled single-cell expression dynamics using scribe. *Cell systems*, 10(3):265–274, 2020.

[20] Sophie Lèbre, Jennifer Becq, Frédéric Devaux, Michael P H Stumpf, and Gaëlle Lelandais. Statistical inference of the time-varying structure of gene-regulation networks. *BMC Systems Biology*, 4(1):130, 2010.

[21] Junil Kim, Simon T. Jakobsen, Kedar N Natarajan, and Kyoung-Jae Won. Tenet: gene network reconstruction using transfer entropy reveals key regulatory factors from single cell transcriptomic data. *Nucleic acids research*, 49(1):e1–e1, 2021.

[22] Hirotaka Matsumoto, Hisanori Kiryu, Chikara Furusawa, Minoru SH Ko, Shigeru BH Ko, Norio Gouda, Tetsutaro Hayashi, and Itoshi Nikaido. Scode: an efficient regulatory network inference algorithm from single-cell rna-seq during differentiation. *Bioinformatics*, 33(15):2314–2321, 2017.

[23] Atul Deshpande, Li-Fang Chu, Ron Stewart, and Anthony Gitter. Network inference with granger causality ensembles on single-cell transcriptomics. *Cell reports*, 38(6):110333, 2022.

[24] Pierre-Cyril Aubin-Frankowski and Jean-Philippe Vert. Gene regulation inference from single-cell rna-seq data with linear differential equations and velocity inference. *Bioinformatics*, 36(18):4774–4780, 2020.

[25] Caleb Weinreb, Samuel Wolock, Betsabeh K Tusi, Merav Socolovsky, and Allon M Klein. Fundamental limits on dynamic inference from single-cell snapshots. *Proceedings of the National Academy of Sciences*, 115(10):E2467–E2476, 2018.

[26] Geoffrey Schiebinger. Reconstructing developmental landscapes and trajectories from single-cell data. *Current Opinion in Systems Biology*, 27:100351, 2021.

[27] Marius Lange, Volker Bergen, Michal Klein, Manu Setty, Bernhard Reuter, Mostafa Bakhti, Heiko Lickert, Meshal Ansari, Janine Schniering, Herbert B Schiller, et al. Cellrank for directed single-cell fate mapping. *Nature methods*, pages 1–12, 2022.

[28] Xiaojie Qiu, Yan Zhang, Jorge D Martin-Rufino, Chen Weng, Shayan Hosseinzadeh, Dian Yang, Angela N Pogson, Marco Y Hein, Kyung Hoi Joseph Min, Li Wang, et al. Mapping transcriptomic vector fields of single cells. *Cell*, 185(4):690–711, 2022.

[29] Kevin R Moon, Jay S Stanley III, Daniel Burkhardt, David van Dijk, Guy Wolf, and Smita Krishnaswamy. Manifold learning-based methods for analyzing single-cell rna-sequencing data. *Current Opinion in Systems Biology*, 7:36–46, 2018.

[30] Thomas Schreiber. Measuring information transfer. *Physical review letters*, 85(2):461, 2000.

[31] Lionel Barnett, Adam B Barrett, and Anil K Seth. Granger causality and transfer entropy are equivalent for gaussian variables. *Physical review letters*, 103(23):238701, 2009.

[32] Naomi Moris, Cristina Pina, and Alfonso Martinez Arias. Transition states and cell fate decisions in epigenetic landscapes. *Nature Reviews Genetics*, 17(11):693–703, 2016.

[33] Conrad Hal Waddington. *The strategy of the genes.* 1957.

[34] Adam L MacLean, Tian Hong, and Qing Nie. Exploring intermediate cell states through the lens of single cells. *Current Opinion in Systems Biology*, 9:32–41, 2018.

[35] Peijie Zhou, Shuxiong Wang, Tiejun Li, and Qing Nie. Dissecting transition cells from single-cell transcriptome data through multiscale stochastic dynamics. *Nature communications*, 12(1):5609, 2021.

[36] Ronald R Coifman and Stéphane Lafon. Diffusion maps. *Applied and computational harmonic analysis*, 21(1):5–30, 2006.

[37] Volker Bergen, Marius Lange, Stefan Peidli, F Alexander Wolf, and Fabian J Theis. Generalizing rna velocity to transient cell states through dynamical modeling. *Nature biotechnology*, 38(12):1408–1414, 2020.

[38] Daniel T Gillespie. The chemical langevin equation. *The Journal of Chemical Physics*, 113(1):297–306, 2000.

[39] Crispin Gardiner. *Stochastic Methods: A Handbook For The Natural And Social Sciences*. Springer, 2009.

[40] MA Coomer, L Ham, and MPH Stumpf. Noise distorts the epigenetic landscape and shapes cell-fate decisions. *Cell Syst*, 13(1):83–102, 2022.

[41] Terry Bossomaier, Lionel Barnett, Michael Harré, and Joseph T Lizier. Transfer entropy. In *An introduction to transfer entropy*, pages 65–95. Springer, 2016.

[42] Thomas M. Cover and Joy A. Thomas. *Elements Of Information Theory*. J. Wiley, 2005.

[43] Adam A Margolin, Ilya Nemenman, Katia Basso, Chris Wiggins, Gustavo Stolovitzky, Riccardo Dalla Favera, and Andrea Califano. Aracne: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. In *BMC bioinformatics*, volume 7, pages 1–15. BioMed Central, 2006.

[44] Louise Deconinck, Robrecht Cannoodt, Wouter Saelens, Bart Deplancke, and Yvan Saeys. Recent advances in trajectory inference from single-cell omics data. *Current Opinion in Systems Biology*, 27:100344, 2021.

[45] Stephen Zhang, Anton Afanassiev, Laura Greenstreet, Tetsuya Matsumoto, and Geoffrey Schiebinger. Optimal transport analysis reveals trajectories in steady-state systems. *PLoS computational biology*, 17(12):e1009466, 2021.

[46] Aviv Madar, Alex Greenfield, Eric Vanden-Eijnden, and Richard Bonneau. Dream3: network inference using dynamic context likelihood of relatedness and the inferelator. *PloS one*, 5(3):e9803, 2010.

[47] Jeremiah J Faith, Boris Hayete, Joshua T Thaden, Ilaria Mogno, Jamey Wierzbowski, Guillaume Cottarel, Simon Kasif, James J Collins, and Timothy S Gardner. Large-scale mapping and validation of escherichia coli transcriptional regulation from a compendium of expression profiles. *PLoS biology*, 5(1):e8, 2007.

[48] Tiejun Li, Jifan Shi, Yichong Wu, and Peijie Zhou. On the mathematics of rna velocity i: theoretical analysis. *bioRxiv*, 2020.

[49] Ulysse Herbach, Arnaud Bonnaffoux, Thibault Espinasse, and Olivier Gandrillon. Inferring gene regulatory networks from single-cell data: a mechanistic approach. *BMC systems biology*, 11(1):1–15, 2017.

[50] Vân Anh Huynh-Thu, Alexandre Irrthum, Louis Wehenkel, and Pierre Geurts. Inferring regulatory networks from expression data using tree-based methods. *PloS one*, 5(9):e12776, 2010.

[51] Jeff Bezanson, Alan Edelman, Stefan Karpinski, and Viral B Shah. Julia: A fresh approach to numerical computing. *SIAM review*, 59(1):65–98, 2017.

[52] E Roesch, JG Greener, AL MacLean, H Nassar, C Rackauckas, TE Holy, and MPH Stumpf. Julia for biologists. *Nat Methods*, 20(5):655–664, 2023.

[53] Yael Baran, Akhiad Bercovich, Arnau Sebe-Pedros, Yaniv Lubling, Amir Giladi, Elad Chomsky, Zohar Meir, Michael Hoichman, Aviezer Lifshitz, and Amos Tanay. Metacell: analysis of single-cell rna-seq data using k-nn graph partitions. *Genome biology*, 20(1):1–19, 2019.

[54] Manu Setty, Vaidotas Kiseliovas, Jacob Levine, Adam Gayoso, Linas Mazutis, and Dana Pe'Er. Characterization of cell fate probabilities in single-cell data with palantir. *Nature biotechnology*, 37(4):451–460, 2019.

[55] T Matsumoto, S Zhang, and G Schiebinger. Beyond $k$nn: Adaptive, sparse neighborhood graphs via optimal transport. Submitted, 2022.

[56] Nan Papili Gao, SM Minhaz Ud-Dean, Olivier Gandrillon, and Rudiyanto Gunawan. Sincerities: inferring gene regulatory networks from time-stamped single cell transcriptional expression profiles. *Bioinformatics*, 34(2):258–266, 2018.

[57] Rihao Qu, Xiuyuan Cheng, Peggy Myung, Esen Sefik, Richard Flavell, and Yuval Kluger. Gene trajectory inference for single-cell data by optimal transport metrics. *bioRxiv*, pages 2022–07, 2022.

[58] Tamara G Kolda and Brett W Bader. Tensor decompositions and applications. *SIAM review*, 51(3):455–500, 2009.

[59] Tetsutaro Hayashi, Haruka Ozaki, Yohei Sasagawa, Mana Umeda, Hiroki Danno, and Itoshi Nikaido. Single-cell full-length total rna sequencing uncovers dynamics of recursive splicing and enhancer rnas. *Nature communications*, 9(1):1–16, 2018.

[60] Huilei Xu, Caroline Baroukh, Ruth Dannenfelser, Edward Y Chen, Christopher M Tan, Yan Kou, Yujin E Kim, Ihor R Lemischka, and Avi Ma'ayan. Escape: database for integrating high-content published data collected from human and mouse embryonic stem cells. *Database*, 2013, 2013.

[61] Wenjing Shi, Hui Wang, Guangjin Pan, Yijie Geng, Yunqian Guo, and Duanqing Pei. Regulation of the pluripotency marker rex-1 by nanog and sox2. *Journal of biological chemistry*, 281(33):23319–23325, 2006.

[62] Guangjin Pan and James A Thomson. Nanog and transcriptional networks in embryonic stem cell pluripotency. *Cell research*, 17(1):42–49, 2007.

[63] Chin Yan Lim, Wai-Leong Tam, Jinqiu Zhang, Haw Siang Ang, Hui Jia, Leonard Lipovich, Huck-Hui Ng, Chia-Lin Wei, Wing Kin Sung, Paul Robson, et al. Sall4 regulates distinct transcription circuitries in different blastocyst-derived stem cell lineages. *Cell stem cell*, 3(5):543–554, 2008.

[64] Kelly L Covello, James Kehler, Hongwei Yu, John D Gordan, Andrew M Arsham, Cheng-Jun Hu, Patricia A Labosky, M Celeste Simon, and Brian Keith. Hif-2$\alpha$ regulates oct-4: effects of hypoxia on stem cell function, embryonic development, and tumor growth. *Genes & development*, 20(5):557–570, 2006.

[65] Hong Doan, Alexander Parsons, Shruthi Devkumar, Jogitha Selvarajah, Francesc Miralles, and Veronica A Carroll. Hif-mediated suppression of deptor confers resistance to mtor kinase inhibition in renal cancer. *Iscience*, 21:509–520, 2019.

[66] Pooja Agrawal, Joseph Reynolds, Shereen Chew, Deepak A Lamba, and Robert E Hughes. Deptor is a stemness factor that regulates pluripotency of embryonic stem cells. *Journal of Biological Chemistry*, 289(46):31818–31826, 2014.

[67] Edward E Morrisey, Zhihua Tang, Kirsten Sigrist, Min Min Lu, Fang Jiang, Hon S Ip, and Michael S Parmacek. Gata6 regulates hnf4 and is required for differentiation of visceral endoderm in the mouse embryo. *Genes & development*, 12(22):3579–3590, 1998.

[68] Yingyao Zhou, Bin Zhou, Lars Pache, Max Chang, Alireza Hadj Khodabakhshi, Olga Tanaseichuk, Christopher Benner, and Sumit K Chanda. Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. *Nature communications*, 10(1):1–10, 2019.

[69] Aimée Bastidas-Ponce, Sophie Tritschler, Leander Dony, Katharina Scheibner, Marta Tarquis-Medina, Ciro Salinno, Silvia Schirge, Ingo Burtscher, Anika Böttcher, Fabian J Theis, et al. Comprehensive single cell mrna profiling reveals a detailed roadmap for pancreatic endocrinogenesis. *Development*, 146(12):dev173849, 2019.

[70] Ashleigh E Schaffer, Brandon L Taylor, Jacqueline R Benthuysen, Jingxuan Liu, Fabrizio Thorel, Weiping Yuan, Yang Jiao, Klaus H Kaestner, Pedro L Herrera, Mark A Magnuson, et al. Nkx6. 1 controls a gene regulatory network required for establishing and maintaining pancreatic beta cell identity. *PLoS genetics*, 9(1):e1003274, 2013.

[71] EC Baechler, FM Batliwalla, George Karypis, PM Gaffney, K Moser, WA Ortmann, KJ Espe, S Balasubramanian, KM Hughes, JP Chan, et al. Expression levels for many genes in human peripheral blood cells are highly sensitive to ex vivo incubation. *Genes & Immunity*, 5(5):347–353, 2004.

[72] Léo Machado, Frédéric Relaix, and Philippos Mourikis. Stress relief: emerging methods to mitigate dissociation-induced artefacts. *Trends in Cell Biology*, 31(11):888–897, 2021.

[73] Brandon L Taylor, Fen-Fen Liu, and Maike Sander. Nkx6. 1 is essential for maintaining the functional state of pancreatic beta cells. *Cell reports*, 4(6):1262–1275, 2013.

[74] H Efsun Arda, Cecil M Benitez, and Seung K Kim. Gene regulatory networks governing pancreas development. *Developmental cell*, 25(1):5–13, 2013.

[75] Philip A Seymour, Kristine K Freude, Man N Tran, Erin E Mayes, Jan Jensen, Ralf Kist, Gerd Scherer, and Maike Sander. Sox9 is required for maintenance of the pancreatic progenitor cell pool. *Proceedings of the National Academy of Sciences*, 104(6):1865–1870, 2007.

[76] Richard I Sherwood, Tzong-Yang Albert Chen, and Douglas A Melton. Transcriptional dynamics of endodermal organ formation. *Developmental dynamics: an official publication of the American Association of Anatomists*, 238(1):29–42, 2009.

[77] Shin-Heng Chiou, Madeleine Dorsch, Eva Kusch, Santiago Naranjo, Margaret M Kozak, Albert C Koong, Monte M Winslow, and Barbara M Grüner. Hmga2 is dispensable for pancreatic cancer development, metastasis, and therapy resistance. *Scientific reports*, 8(1):1–10, 2018.

[78] C Dorrell, J Schug, CF Lin, PS Canaday, AJ Fox, O Smirnova, R Bonnah, PR Streeter, CJ Stoeckert, KH Kaestner, et al. Transcriptomes of the major human pancreatic cell types. *Diabetologia*, 54:2832–2844, 2011.

[79] Kailun Lee, Jeng Yie Chan, Cassandra Liang, Chi Kin Ip, Yan-Chuan Shi, Herbert Herzog, William E Hughes, Mohammed Bensellam, Viviane Delghingaro-Augusto, Mark E Koina, et al. Xbp1 maintains beta cell identity, represses beta-to-alpha cell transdifferentiation and protects against diabetic beta cell failure during metabolic stress in mice. *Diabetologia*, 65(6):984–996, 2022.

[80] Gennady Gorin, Meichen Fang, Tara Chari, and Lior Pachter. Rna velocity unraveled. *bioRxiv*, 2022.

[81] Shijie C Zheng, Genevieve Stein-O'Brien, Leandros Boukas, Loyal A Goff, and Kasper D Hansen. Pumping the brakes on rna velocity–understanding and interpreting rna velocity estimates. *bioRxiv*, pages 2022–06, 2022.

[82] Valérie Marot-Lassauzaie, Brigitte Joanne Bouman, Fearghal Declan Donaghy, Yasmin Demerdash, Marieke Alida Gertruda Essers, and Laleh Haghverdi. Towards reliable quantification of cell state velocities. *PLoS Computational Biology*, 18(9):e1010031, 2022.

[83] Cuiping Zhang, Yvonne N Fondufe-Mittendorf, Chi Wang, Jin Chen, Qiang Cheng, Daohong Zhou, Yi Zheng, Hartmut Geiger, and Ying Liang. Latexin regulation by hmgb2 is required for hematopoietic stem cell maintenance. *haematologica*, 105(3):573, 2020.

[84] Michael J Nemeth, Martha R Kirby, and David M Bodine. Hmgb3 regulates the balance between hematopoietic stem cell self-renewal and differentiation. *Proceedings of the National Academy of Sciences*, 103(37):13783–13788, 2006.

[85] Xunde Wang, Nikolaos Angelis, and Swee Lay Thein. Myb–a regulatory factor in hematopoiesis. *Gene*, 665:6–17, 2018.

[86] Thomas Graf and Tariq Enver. Forcing cells to change lineages. *Nature*, 462(7273):587–594, 2009.

[87] Rachel M Gerstein. Deciding the decider: Mef2c in hematopoiesis. *Nature Immunology*, 10(3):235–236, 2009.

[88] Hongsheng Wang and Herbert C Morse. Irf8 regulates myeloid and b lymphoid lineage diversification. *Immunologic research*, 43:109–117, 2009.

[89] Tomohiko Tamura, Daisuke Kurotaki, and Shin-ichi Koizumi. Regulation of myelopoiesis by the transcription factor irf8. *International journal of hematology*, 101:342–351, 2015.

[90] Voahangy Randrianarison-Huetz, Benoit Laurent, Valérie Bardet, Gerard C Blobe, François Huetz, and Dominique Duménil. Gfi-1b controls human erythroid and megakaryocytic differentiation by regulating tgf-$\beta$ signaling at the bipotent erythro-megakaryocytic progenitor stage. *Blood, The Journal of the American Society of Hematology*, 115(14):2784–2795, 2010.

[91] Emmanuelle Passegué, Erwin F Wagner, and Irving L Weissman. Junb deficiency leads to a myeloproliferative disorder arising from hematopoietic stem cells. *Cell*, 119(3):431–443, 2004.

[92] Karthik Arumugam, William Shin, Valentina Schiavone, Lukas Vlahos, Xiaochuan Tu, Davide Carnevali, Jordan Kesner, Evan O Paull, Neus Romo, Prem Subramaniam, et al. The master regulator protein baz2b can reprogram human hematopoietic lineage-committed progenitors into a multipotent state. *Cell reports*, 33(10):108474, 2020.

[93] Nico Battich, Joep Beumer, Buys de Barbanson, Lenno Krenning, Chloé S Baron, Marvin E Tanenbaum, Hans Clevers, and Alexander van Oudenaarden. Sequencing metabolically labeled transcripts in single cells reveals mrna turnover strategies. *Science*, 367(6482):1151–1156, 2020.

[94] Lam-Ha Ly and Martin Vingron. Effect of imputation on gene network reconstruction from single-cell rna-seq data. *Patterns*, 3(2):100414, 2022.

[95] Shou-Wen Wang, Michael J Herriges, Kilian Hurley, Darrell N Kotton, and Allon M Klein. Cospar identifies early cell fate biases from single-cell transcriptomic and lineage information. *Nature Biotechnology*, 40(7):1066–1074, 2022.