Learning cell fate landscapes from spatial transcriptomics using Fused Gromov-Wasserstein

Geert-Jan Huizing, Gabriel Peyré, Laura Cantini

Geert-Jan Huizing, Gabriel Peyré, Laura Cantini. Learning cell fate landscapes from spatial transcriptomics using Fused Gromov-Wasserstein. 2024. hal-04730091

HAL Id: hal-04730091
https://hal.science/hal-04730091v1

Preprint submitted on 10 Oct 2024

Learning cell fate landscapes from spatial transcriptomics using Fused Gromov-Wasserstein

Geert-Jan Huizing[1], Gabriel Peyré[2], Laura Cantini[1]

[1]Institut Pasteur, Université Paris Cité, CNRS UMR 3738, Machine Learning for Integrative Genomics Group, F-75015, Paris, France
[2]CNRS and DMA de l'Ecole Normale Supérieure, CNRS, Ecole Normale Supérieure, Université PSL, 75005, Paris, France

Abstract

In dynamic biological processes such as development, spatial transcriptomics is revolutionizing the study of the mechanisms underlying spatial organization within tissues. Inferring cell fate trajectories from spatial transcriptomics profiled at several time points has thus emerged as a critical goal, requiring novel computational methods. Wasserstein gradient flow learning is a promising framework for analyzing sequencing data across time, built around a neural network representing the differentiation potential. However, existing gradient flow learning methods cannot analyze spatially resolved transcriptomic data.

Here, we propose STORIES, a method that employs an extension of Optimal Transport to learn a spatially informed potential. We benchmark our approach using three large Stereo-seq spatiotemporal atlases and demonstrate superior spatial coherence compared to existing approaches. Finally, we provide an in-depth analysis of axolotl neural regeneration and mouse gliogenesis, recovering gene trends for known markers as *Nptx1* in neuron regeneration and *Aldh1l1* in gliogenesis and additional putative drivers.

Introduction

Spatial transcriptomics technologies are revolutionizing the study of how cells organize within tissues[1]. Techniques based on high-throughput sequencing have enabled the unbiased discovery of gene expression patterns within their spatial context. For instance, recent studies have revealed previously unknown spatial organization at the tumor-microenvironment interface in melanoma and Alzheimer's disease amyloid plaque microenvironment[2,3]. The most widely used spatially-resolved sequencing techniques (e.g. 10X Visium) measure spots larger than the typical cell size. However, recent technological developments based on barcoded arrays like Stereo-seq and HDST have reached single-cell resolution, effectively bridging functional and structural characterizations of the cell[4,5]. Recent works have leveraged Stereo-seq to produce large spatiotemporal atlases of various biological processes by profiling a system with spatial transcriptomics at several points in time[4,6,7]. These datasets are ideal for studying cellular dynamics within the tissue during processes such as development and the onset of complex diseases, where cells undergo coordinated transcriptomic changes and spatial reorganization.

Inferring the dynamics of biological processes from single-cell sequencing data requires tailored computational approaches known as trajectory inference methods[8]. Monocle initiated the field of trajectory inference by ordering cells along a pseudotime axis based on their transcriptomic similarities and analyzing gene expression trends along pseudotime[9]. While pseudotime represents the progression along a differentiation process, pseudotime-based methods do not provide a model for the underlying transcriptomic changes, and thus cannot predict a cell's future transcriptomic state[10]. RNA velocity has thus been proposed to predict changes in gene expression based on splicing dynamics[11]. However, velocity-based methods rely on simple kinetic models that can misinterpret cell dynamics, for instance in the case of transient boosts in transcription[12].

Multiple methods based on Optimal Transport (OT) have been developed for cases when several time points are available along differentiation. Waddington OT infers trajectories by computing probabilistic cell-cell transitions between adjacent time points[13]. However, it delivers neither a notion of pseudotime nor a notion of velocity. Another class of OT-based methods proposes a continuous model of population dynamics by training neural networks representing a generalized notion of velocity[14,15]. However, these methods do not order cells along a pseudotime axis. A promising OT-based framework for trajectory inference consists of learning a potential function governing a causal model of differentiation[16–18]. Framing cellular differentiation as the minimization of a potential function is rooted in systems biology and formalizes Waddington's idea of epigenetic landscape[19,20]. Furthermore, the potential function is a natural alternative to pseudotime, and its gradient yields a rigorous notion of velocity.

OT-based approaches for trajectory inference using spatial transcriptomics through time have recently been developed[21–23]. For instance, Moscot computes cell-cell transitions between adjacent time points using an extension of OT called Fused Gromov-Wasserstein (FGW), and stVCR learns a spatial velocity along with a gene expression velocity[22–24]. However, these methods do not provide an intuitive ordering of cells along differentiation, which limits their ability to discover the mechanisms driving dynamic biological processes. In addition, Moscot cannot predict the evolution of cells at future time points.

Here, we propose STORIES, a novel trajectory inference method capable of learning a causal model of cellular differentiation from spatial transcriptomics through time using FGW. Unlike Moscot and stVCR, STORIES learns a potential function that defines each cell's stage of differentiation. In addition, unlike Moscot, STORIES allows one to predict the evolution of cells at future time points. Indeed, STORIES uses FGW as a machine learning loss to learn a continuous model of differentiation, while Moscot uses FGW to connect adjacent time points.

We benchmarked our approach on three large-scale spatiotemporal Stereo-seq atlases, covering mouse development, zebrafish development, and axolotl regeneration[6,6,7]. Furthermore, we used STORIES for the in-depth analysis of cellular trajectories in axolotl neural regeneration and mouse gliogenesis. We recover gene trends for known markers, such as *Nptx1* in Nptx+ excitatory neuron regeneration and *Aldh1l1* in gliogenesis. In addition, STORIES uncovers other possible driver genes and transcriptional regulators of cellular differentiation in these contexts, which may be of interest for further biological investigation.

Finally, we provide STORIES as an open-source and user-friendly Python package (github.com/cantinilab/stories). It is based on the Scverse ecosystem, making it easy to interface STORIES with existing tools for single-cell analysis such as Scanpy and CellRank[25–27]. In addition, STORIES benefits from the JAX ecosystem for deep learning and OT computation, enabling the fast handling of large datasets[28,29].

Results

STORIES: a new single-cell trajectory inference method for spatial transcriptomics profiled through time

We developed SpatioTemporal Omics eneRgIES (STORIES), a novel tool for single-cell trajectory inference using omics data profiled through spatial and temporal dimensions (github.com/cantinilab/stories). STORIES allows studying dynamic biological processes in their spatial context by identifying cell fates, gene trends, and candidate transcriptional regulators (see Fig. 1).

STORIES is based on the Optimal Transport (OT), a mathematical framework that enables the geometrically meaningful comparison of distributions, using various flavors of the Wasserstein distance[30,31]. OT also provides a valuable model for population dynamics: the so-called Wasserstein gradient flows were popularized by Jordan, Kinderlehrer, and Otto for their connection with the Fokker-Planck equation and were recently used for trajectory inference in single-cell transcriptomics[16–18,32]. However, existing methods for trajectory inference based on Wasserstein gradient flows are not equipped to deal with spatially resolved omics data. STORIES introduces key methodological innovations that allow one to address the specific challenges of including spatial information.

As an input, STORIES takes slices of spatial transcriptomics profiled at several time points. For instance, Fig. 1A displays sections of axolotl brains profiled at different stages during regeneration. STORIES then learns the parameters θ of a neural network J_θ, which assigns a differentiation potential to each cell according to its gene expression profile x (see Fig. 1B). The function J_θ formalizes the Waddington epigenetic landscape, where undifferentiated cells have a high potential and, as they differentiate, move towards low-potential transcriptomic states, which correspond to mature cell types[20]. The transition to these low-potential attractor states defines a causal model of cellular dynamics capable of predicting future gene expression patterns and suggesting potential driver genes and transcriptional regulators (see Fig. 1C).

STORIES's potential-based approach provides two interpretable and biologically meaningful outputs: (i) the potential $J_\theta(\mathrm{x})$ naturally orders cells x along a differentiation process (ii) the vector $-\nabla_{\mathrm{x}} J_\theta(\mathrm{x})$ gives the direction of the evolution of gene expression. On the contrary, pseudotime-based methods[9,33,34] focus on the first aspect, and velocity-based[11,35] methods focus on the second. Crucially, STORIES also innovates compared to state-of-the-art potential-based methods[16–18] by enabling the use of spatial coordinates.

Briefly, STORIES trains the neural network J_θ by predicting a distribution $\rho_{t_k}(\theta)$ of gene expression profiles for each time point t_k where $k \in [1, \dots, K]$. These distributions are then compared to the ground-truth distributions μ_{t_k}, and the parameters θ are updated to improve the predictions. Unlike existing potential-based methods, STORIES allows one to take into account the spatial coordinates of cells when comparing the distributions of gene expression.

Formally, let $\mu_t = \sum_i a_i \delta_{(\mathbf{x}_i, \mathbf{r}_i)}$ denote the empirical distribution of cells at time t, characterized by their gene expression profile $\mathbf{x}_i \in \mathbb{R}^d$, spatial coordinates $\mathbf{r}_i \in \mathbb{R}^2$ and weight $a_i \in \mathbb{R}_+$ where $\sum_i a_i = 1$. Similarly, let us denote $\rho_t(\theta) = \sum_j b_j \delta_{(\mathbf{y}_j, \mathbf{s}_j)}$ the predictions of STORIES at time t. Unlike the gene expression profiles $\mathbf{x}_i, \mathbf{y}_j$, the spatial coordinates $\mathbf{r}_i, \mathbf{s}_j$ are not directly comparable because the slices are not necessarily aligned between time points. In other words, the spatial coordinates $\mathbf{r}_i, \mathbf{s}_j$ are defined up to an isometry (e.g., a rotation or translation).

Existing potential-based methods train the neural network using a linear OT objective, which is sensitive to isometries. Our approach instead uses a recently developed quadratic extension of OT called Fused Gromov-Wasserstein (FGW), which renders the model invariant to spatial isometries[24]. The FGW distance, defined below and explained more thoroughly in the Methods section, allows one to compare the distributions μ_t and ρ_t directly on gene expression profiles, and up to an isometry on spatial coordinates.

$$\mathcal{FGW}_\alpha^\varepsilon(\mu_t, \rho_t) = \min_{\mathbf{P} \in \mathcal{U}(\mathbf{a}, \mathbf{b})} (1-\alpha)\mathrm{L}(\mathbf{P}) + \alpha \mathrm{Q}(\mathbf{P}) - \varepsilon \mathrm{E}(\mathbf{P}) \tag{1}$$

FGW seeks a matrix $\mathbf{P}$ mapping cells from μ_t to ρ_t such that $\mathbf{P}$ minimizes the sum of three terms: (i) the linear term L compares the gene expression coordinates $\mathbf{x}_i, \mathbf{y}_j$ (ii) the quadratic term Q compares pairwise distances $d(\mathbf{r}_i, \mathbf{r}_{i'})$ and $d(\mathbf{s}_j, \mathbf{s}_{j'})$, which are not affected by translating or rotating the tissue (iii) an entropic regularization term E. The parameter $\alpha \in [0, 1[$ denotes the relative weight of spatial information.

Our proposed objective function evaluates the predictions across all time points using a debiased version of the FGW distance denoted $\overline{\mathcal{FGW}_\alpha^\varepsilon}$ (see Methods):

$$\mathcal{L}(\theta) = \sum_{k=1}^{K-1} (t_{k+1} - t_k) \overline{\mathcal{FGW}_\alpha^\varepsilon}(\mu_{t_{k+1}}, \rho_{t_{k+1}}(\theta)) \tag{2}$$

For $\alpha = 0$, Equation 2 corresponds to a model relying purely on linear OT and which does not leverage spatial information, as proposed in the state-of-the-art[16–18]. In the following, we refer to this as the linear method. Existing methods[16–18] propose different strategies to make the predictions $\rho_{t_{k+1}}(\theta)$. They vary in terms of teacher-forcing, number of steps between t_k and t_{k+1}, and whether steps are implicit or explicit (see "Discretization" in Methods). The linear method presented here ($\alpha = 0$) incorporates the best-performing choices for our experiments (see "Discretization" in Methods and Supplementary Figure 1). The linear method thus offers an unbiased way to assess the performance of STORIES compared to the state-of-the-art, as explored in the next section.

STORIES is implemented as an open-source Python package seamlessly integrated into the classical Python single-cell analysis pipeline (github.com/cantinilab/stories). Users can thus take advantage of scverse tools like Scanpy, Squidpy, and CellRank for preprocessing and downstream analysis[26,27,36]. In addition, STORIES provides a user-friendly visualization of driver genes and enriched transcription factors, thus helping biological interpretability.

In the following sections, we extensively benchmark STORIES against the state-of-the-art[16–18] using large-scale spatiotemporal atlases. To this end, we compare the linear method to our proposed model across different values of the parameter α.

STORIES's predictions are more biologically relevant than the state-of-the-art

We assessed the effectiveness of STORIES in predicting cell states over time across three Stereo-seq spatiotemporal atlases: a mouse development atlas, a zebrafish development atlas, and an axolotl brain regeneration atlas[4,6,7]. Details on data processing are provided in the Methods section.

From each atlas, we created three sets: a training set, an early test set, and a late test set (see Fig. 2A). The test sets are composed of two time points, and the goal is to use the first time point to predict the second time point's gene expression. The late test set is particularly challenging because its second slice comes from an entirely new time point, which may contain cell states not seen during training. For example, in the zebrafish atlas, fast muscle cells only appear at 24 hours post-fertilization (i.e. hpf), whereas the training set includes slices only up to 18 hpf.

Two scores, defined in the Methods section, have been used for the benchmark. First, the gene expression prediction score evaluates how closely the prediction approaches the real gene expression at the next time point. Second, the spatial coherence score evaluates whether the predictions are consistent with space. For instance, predicting lung cells from liver cells would yield a bad spatial coherence score.

We compared STORIES to the linear method described in the previous section, which incorporates the best-performing aspects of state-of-the-art methods[16–18] for our experiments (see "Discretization" in Methods and Supplementary Fig. 1). Since the linear method is formally equivalent to STORIES with $\alpha = 0$, it offers an unbiased way to assess the usefulness of spatial information compared to the state-of-the-art.

To better evaluate the impact of spatial information, we reported results with a varying weight α, representing the importance of the spatial term compared to the gene expression term (see Figure 2B). As expected, smaller values of α perform better in terms of gene expression prediction score, while larger values perform better in terms of spatial coherence score. A good compromise between gene expression and space is obtained for $\alpha = 5 \times 10^{-3}$, which we set as a default value.

For this choice of α, STORIES outperforms the linear method in terms of spatial coherence across all datasets for both early and late test sets (see Fig. 2B). In addition, STORIES outperforms the linear method in terms of gene expression prediction in the mouse atlas's early test set and the zebrafish atlas's early test set. This demonstrates that STORIES's predictions are biologically relevant on two levels: they are accurate in gene expression and coherent in space.

The biological relevance of STORIES’s predictions is confirmed in Fig. 2C, which compares the training losses of STORIES and the linear method. Since both methods involve matching predictions with a reference population of cells, we compared their matchings for specific cell types (see Methods).

First, in the axolotl atlas, STORIES correctly matches predictions from immature neurons (IMN) with Nptx+ excitatory neurons in the lateral pallium (NptxEX), and predictions from regeneration intermediate progenitor cells (rIPC2) with excitatory neurons in the dorsal pallium (dpEX)[7]. The linear method, on the contrary, incorrectly matches rIPC2 predictions with microglial cells (MCG) from a different anatomical region (see Fig. 2C).

Second, in the zebrafish atlas, STORIES accurately matches predictions from the optic vesicle with cells located around the eye, and predictions from the polster with cells located within the head. In contrast, the linear method incorrectly matches optic vesicle predictions with a broad group of cells across different anatomical regions, and polster predictions with cells from the tail area (see Fig. 2C).

Third, in the mouse atlas, STORIES correctly matches predictions from liver and lung cells with their respective organs. The linear method, instead, incorrectly matches lung cell predictions with a broad group of cells across organs (see Fig. 2C), suggesting poor biological coherence of the linear method.

STORIES's superior performance in achieving biologically coherent and accurate gene expression predictions demonstrates the significant benefit of considering spatial information when learning a gradient flow model on spatial transcriptomics data.

STORIES identifies trajectories and potential drivers of neuron regeneration in axolotls

To further assess the potential of STORIES for trajectory inference in spatial transcriptomics through time, we first focused on axolotl brain regeneration.

We trained STORIES as described in Methods on the subset of cells described in the original publication as involved in neuron regeneration: Wnt+ and reactive ependymoglial cells (wntEGC and reaEGC), regeneration intermediate progenitor cells (rIPC1 and rIPC2), immature neurons (IMN), Nptx+ lateral pallium excitatory neurons (nptxEX), dorsal pallium excitatory neurons (dpEX), and medial pallium excitatory neurons (mpEX)[7]. As shown in Fig. 3A, STORIES learns an energy landscape consistent with the original publication. Indeed, the potential J_θ assigns a high potential to progenitor states (wntEGC and reaEGC), a medium potential to intermediary states (rIPC1, rIPC2, and IMN), and a low potential to mature states (nptxEX, dpEX, and mpEX).

We computed cell-cell transitions by applying CellRank on the gradient of the trained potential (see Methods), as visualized in Fig. 3B. These transitions highlight that STORIES not only detects the correct stage of differentiation, but also recovers the three trajectories described in the original publication: wntEGC-mpEX, reaEGC-rIPC2-dpEX, and reaEGC-rIPC1-IMN-nptxEX[7]. Importantly, the authors identified these three trajectories by applying Moncole separately on three spatial regions and specifying EGCs as the starting point in each case. In contrast, STORIES achieves the same results without the need to isolate specific spatial regions and specify the starting point of the trajectory. Indeed, STORIES leverages spatial information to process all regions simultaneously and leverages temporal information to infer progenitor states from the data.

We then narrowed further into the reaEGC-rIPC1-IMN-nptxEX trajectory, which the original publication studies in most detail[7]. First, we sought to confirm expected gene trends along this trajectory. The original study suggests *Vim*, which encodes a critical cytoskeletal protein, as a marker of reaEGC cells and *Nptx1*, which is involved in synaptic plasticity, as a marker of NptxEX. Accordingly, STORIES recovered a clear decreasing trend for *Vim* expression along differentiation and a clear increasing trend for *Nptx1* expression (see Fig. 3C).

Next, we performed unsupervised discovery of gene trends by fitting a spline regression model along the previously mentioned trajectory (see Methods). Fig. 3D reports the best candidate driver genes across differentiation stages. Interestingly, the early stages of differentiation coincide with high expression of *Hes5*, which is known to maintain stemness in the context of neural differentiation[37], and *Cdc25b*, a cell-cycle regulator key to neuron production[38,39]. Conversely, late stages of differentiation coincide with high expression of the microtubule-associated protein gene *Map1a*, crucial to neural development and regeneration[40], and *L1cam*, shown to promote axonal regeneration[41]. STORIES also outputs additional genes that represent possible drivers of neuron regeneration and would require further biological investigation. For instance, STORIES uncovered a trend for late expression of the scarcely studied *Nsg2*, which is thought to be involved in synaptic function and, like *Nptx1*, interacts with AMPA receptors[42].

Finally, our analysis revealed possible transcriptional regulators of the differentiation process (see Fig. 3F) by testing transcription factor (TF) enrichment using the curated literature-based TRRUST database (see Methods). The most significantly enriched TF, CTNNB1, encodes β-catenin, which has been described as an essential regulator in neuron regeneration in mouse models and in limb regeneration in axolotl[43–45]. Other top TFs include SP1 and MYC, described in the context of neuron regeneration and computationally retrieved in axolotl limb regeneration[46–49]. Additionally, we identify SOX6, MYCN, and REST, which are not widely studied in the context of regeneration but are known regulators in development[50–52]. Interestingly, a recent study predicted REST as a regulator of neuron regeneration and validated this role in a mouse model[53].

STORIES thus learned a Waddington landscape that captures cell fate trajectories, continuous changes of gene expression, and the regulatory landscape underlying neuron regeneration in axolotls. Our model allowed for the unbiased discovery of potential drivers and mechanisms, possibly relevant for further biological investigations.

STORIES identifies trajectories and potential drivers of gliogenesis in the developing mouse dorsal midbrain

We then sought to highlight STORIES's potential in trajectory inference by studying mouse dorsal midbrain development.

We trained STORIES as described in Methods on the subset of cells described in the original article as exhibiting a branching trajectory: radial glial cells (RGC) differentiating into either neuroblasts (NeuB) or glioblasts (GlioB)[4]. As visualized in Fig. 4A, STORIES learns an energy landscape consistent with the original publication. Indeed, the potential J_θ assigns a high potential to RGC and a low potential to the more differentiated NeuB and GlioB.

We computed cell-cell transitions by applying CellRank on the gradient of the trained potential (see Methods), as visualized in Fig. 4B. These transitions highlight that STORIES not only detects the correct stage of differentiation but also recovers the expected branching from RGC to glial and neural cell fates[4]. Importantly, the original publication identified this branching using Monocle 3, which required manually setting RGC as the trajectories' starting point. On the contrary, STORIES achieves the same results without manual input by leveraging temporal information to infer the starting point from the data.

Glial cells outnumber neurons in the brain, but their development has been studied less extensively[54]. Moreover, understanding gliogenesis is of critical therapeutic importance because of its parallels with glioma, the most common and deadliest form of brain cancer[55]. Thus, we focused further on the RGC-GlioB trajectory. We first sought to confirm expected gene trends along this trajectory. The original study identifies *Mki67*, a proliferation marker, as highly expressed in RGC, and *Aldh1l1*, an astrocyte marker, as highly expressed in GlioB[4]. Accordingly, STORIES recovered a decreasing trend for *Mki67* expression along differentiation and an increasing trend for *Aldh1l1* expression (see Fig. 4C).

Next, we performed unsupervised discovery of gene trends by fitting a spline regression model along the previously mentioned trajectory (see Methods). Fig. 4D reports the best candidate driver genes across differentiation stages. The early stages of differentiation coincide with a high expression of cell cycle genes *Gmnn*, *Rrm2*, and *Hmgb2*[56]. Additionally, we observed a high expression of the alpha-tubulin gene *Tuba1b* in the early stages of differentiation, as previously described in the developing brain[57]. Conversely, the late stages of differentiation coincide with the high expression of the glutamine synthetase gene *Glul*, a key astrocyte marker[58,59]. STORIES also outputs additional genes that represent possible drivers of gliogenesis and would require further biological investigation. For instance, *Glis3* displays an increasing trend along gliogenesis (see Fig. 4E) but is little studied in this context. However, *Glis3* was recently suggested as a therapeutic target to suppress proliferation in glioma[60].

Finally, our analysis revealed candidate transcriptional regulators of the differentiation process (see Fig. 4F) by testing transcription factor (TF) enrichment using the curated literature-based TRRUST database (see Methods). Among the most enriched TFs, SOX4 and NOTCH2 have been studied in gliogenesis[61,62]. Additionally, STORIES recovers MYC, MYCN, and MAX, which have been studied in the context of glioma[63].

This second experiment in the context of gliogenesis confirms STORIES's ability to learn a Waddington landscape that captures cell fate trajectories, continuous gene expression changes, and regulatory processes. In doing so, we highlight known and potential mechanisms that motivate further biological investigations.

Discussion

Recent technological advances in spatial transcriptomics have enabled the tracking of gene expression at single-cell resolution in the spatial context of the tissue. Large datasets of spatial transcriptomics profiled through time give a unique opportunity to understand dynamic biological processes such as development and disease onset. However, their analysis requires trajectory inference tools tailored to the specific challenges of spatial data.

In this article, we proposed SpatioTemporal Omics eneRgIES (STORIES), a novel computational framework for trajectory inference from spatial transcriptomics profiled at several time points. STORIES enables a rich and spatially informed analysis of differentiation trajectories. To evaluate STORIES's performance, we benchmarked it against the state-of-the-art in three large Stereo-seq datasets and highlighted the advantage of considering spatial information in trajectory inference from time-course single-cell data. We further showcased STORIES's abilities in two concrete settings: axolotl neuron regeneration and mouse gliogenesis.

STORIES offers a model of population dynamics tailored for single-cell resolution spatial transcriptomics technologies like Stereo-seq or Visium HD. Given the fast-paced developments in spatial transcriptomics[1], the number of spatiotemporal atlases at single-cell resolution can be expected to increase steadily. At the same time, STORIES could be applied to low-resolution data (e.g. 10x Visium), which have a spot size larger than the typical cell, using deconvolution techniques[64,65]. In addition, STORIES could be adapted to imaging-based technologies like MERFISH which offer high resolution but can only detect a limited panel of genes[66].

STORIES provides an interpretable model of differentiation relying on a potential energy. Previous work shows that such potential landscapes arise naturally from simple gene regulatory networks (GRNs)[19]. However, potential energies cannot model complex GRNs, cell-cell communication, or oscillations within a cell state[67]. Extensions to more complex energy functionals could thus lead to further insights into biological processes such as development, immune response, and the onset of complex disease. For instance, recent work in gradient flow learning explored interaction energies[68], which could represent cell-cell communication in biology.

The major novelty of STORIES is its ability to learn a spatially-informed potential. This methodological development is critical because dynamic processes such as development involve coordinated expression changes and tissue reorganization[69]. However, the learned potential operates only on gene expression, so it does not allow the prediction of future positions of cells. Including a spatial component in the energy function may provide a more comprehensive view of biological processes by predicting cell migration. Relating this to existing models for morphogenesis, such as Alan Turing's reaction-diffusion model[70–72], is an exciting avenue for further research.

Methods

Data collection

Zebrafish

For the training and validation sets, we used the five first time points showcased in Figure 2A of the original publication[6]: 3.3hpf slice 1, 5.25hpf slice 10, 10hpf slice 11, 12hpf slice 8, 18hpf slice 8.

For the test set, we used:

- 10hpf slice 17 and 12hpf slice 5 to evaluate prediction within the time range seen during training. These two slices were studied together in Figure S4A of the original publication[6].
- 18hpf slice 11 and 24hpf slice 4 to evaluate prediction outside the time range seen during training. These two slices were studied together in Figure S4E of the original publication[6].

Altogether, this represented 17,920 cells after preprocessing.

Mouse

For the training and validation sets, we used the 7 first time points showcased in Figure 3A of the original publication[4]: E9.5 E1S1, E10.5 E1S1, E11.5 E1S1, E12.5 E1S1, E13.5 E1S1, E14.5 E1S1, E15.5 E1S1.

For the test set, we used:

- E13.5 E1S2 and E14.5 E1S2 to evaluate prediction within the time range seen during training. These two slices were studied in Figure S2C of the original publication[4].
- E15.5 E1S2 and E16.5 E1S1 to evaluate prediction outside of the time range seen during training. The first slice was studied in Figure S2C and the second in Figure 3A of the original publication[4].

Altogether, this represented 794,063 cells after preprocessing.

Dorsal midbrain

We retained one slice per time point. We used the 3 slices showcased in Figure S7A of the original publication[4]: E12.5 (E1S3), E14.5 (E1S3), E16.5 (E1S3). As in the original publication, we subset the analysis to the RGC, NeuB, and GlioB cell types.
Altogether, this represented 4,581 cells after preprocessing. Note that the "Mouse" dataset and the "Dorsal midbrain" dataset originate from the same experiments but have different resolutions (bin 50 vs image-based segmentation), so the same neural network weights cannot be used in both cases.

Axolotl

Benchmark. For the training and validation sets, we used the five first time points showcased in Figure 3B of the original publication[7]: 2DPI (rep1), 5DPI (rep1), 10DPI (rep1), 15DPI (rep4), 20DPI (rep2). We manually removed spatial outliers in 10DPI (rep1), 15DPI (rep4), and 20DPI (rep2). For the test set, we used:

- 10DPI (rep2) and 15DPI (rep1) to evaluate prediction within the time range seen during training. These two slices were showcased in Figure S9 of the original publication[7].
- 20DPI (rep3) and 30DPI (rep2) to evaluate prediction outside of the time range seen during training. These two slices were showcased in Figure S9 of the original publication[7].

As in the original publication, we restricted the analysis to the dorsal part of the injured hemisphere. Altogether, this represented 22,083 cells.

In-depth analysis. For the analysis in Section 3 of Results, we used slices 2DPI (rep1), 5DPI (rep1), 10DPI (rep1), 15DPI (rep4), 20DPI (rep2), and 30DPI (rep2). As before, we manually removed spatial outliers and restricted the analysis to the dorsal part of the injured hemisphere. As in the original publication, we subset the data to the following cell types: nptxEX, reaEGC, wntEGC, dpEX, mpEX, IMN, rIPC1, rIPC2. This represented 5,904 cells.

Preprocessing

For all datasets, we performed the following preprocessing steps.

Cell and gene quality control. Using Scanpy's sc.pp.filter_cells, we removed cells with less than 200 expressed genes. Then, we removed the top 0.1% of cells with the most expressed genes. Finally, we removed genes expressed in less than 3 cells using Scanpy's sc.pp.filter_genes.

Normalization and highly variable gene selection. We applied Scanpy's "Pearson residuals normalization" and selected 10,000 highly variable genes. Scanpy computed highly variable genes for each batch and merged them to avoid selecting batch-specific genes.

Dimensionality reduction. Using Scanpy's sc.tl.pca, we applied Principal Component Analysis (PCA) to reduce the data to 50 dimensions. Section 3 and Section 4 of Results used a subset of relevant cell types. This was done after PCA but before batch correction.

Batch correction. We applied Harmony on the PCA components to correct the batch effects, using Scanpy's sc.external.pp.harmony_integrate, a wrapper around `harmonypy`[73].

Visualization. Using Scanpy's sc.tl.umap, we applied UMAP to project the batch-corrected data into two dimensions. In Section 3 and Section 4 of Results we applied Isomap instead of UMAP. Indeed, we found that visually, Isomap respected cell type transitions better than UMAP. We used scikit-learn's sklearn.manifold.Isomap.

Wasserstein gradient flow learning with a quadratic objective

Notations

Let us consider a time point $t \in \mathbb{R}$ and $\mu_t \in \mathcal{P}(\mathbb{R}^d)$ a discrete distribution of n cells.
We denote $\mu_t = \sum_{i=1}^n a_i \delta_{\mathbf{x}_i}$ where the vector $\mathbf{x}_i \in \mathbb{R}^d$ represents the gene expression of the i-th cell, and the weights $a_i \in \mathbb{R}_+$ are such that $\sum a_i = 1$. In the following, for a function $f : \mathbb{R}^d \to \mathbb{R}^d$, we define the pushforward measure as $f_{\#}\mu_t = \sum_i a_i \delta_{f(\mathbf{x}_i)}$.

Wasserstein gradient flow

Similarly to previous works[16–18], we model the evolution of μ_t as a Wasserstein gradient flow for a potential energy. For one single cell x, the Euclidean gradient flow is a $\mathbf{x}_t$ that verifies $\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t} = -\nabla J(\mathbf{x})$, the continuous counterpart of gradient descent. Wasserstein gradient flows extend this to the space of measures[74,75].
We say that μ_t is a Wasserstein gradient flow for the energy $\mathcal{F} : \mu \mapsto \sum_{i,j} J(\mathbf{x})\mathrm{d}\mu(\mathbf{x}) \in \mathbb{R}$ if its density verifies the continuity equation

$$\frac{\partial \mu(\mathbf{x}, t)}{\partial t} = \mathrm{div}\big(\mu(\mathbf{x}, t)\nabla J(\mathbf{x})\big).$$

As in previous works[16–18], we do not know the potential $J : \mathbb{R}^d \to \mathbb{R}$ a priori and aim instead to learn a neural network J_θ from snapshots $\mu_{t_k} \in \mathcal{P}(\mathbb{R}^d)$ for $t_1 < \cdots < t_k < \cdots < t_K$. See "neural network architecture" for details about J_θ.

Discretization

Our approach boils down to learning J_θ such that for given parameters θ and an initial population $\rho_{t_1} = \mu_{t_1}$, the predicted populations ρ_{t_k} are close to the observed snapshots μ_{t_k}. To make these predictions, existing potential-based methods[16–18] differ in three main aspects: number of steps, teacher-forcing, and discretization scheme. As detailed below, we select the best-performing choices for these three aspects, as measured by validation loss in the Zebrafish atlas. The linear method ($\alpha = 0$) thus differs from existing works by combining their best-performing aspects.

Number of steps. Hashimoto et al. and Yeo et al.[16,17] make intermediary predictions between ρ_{t_k} and $\rho_{t_{k+1}}$, i.e. $\tau \ll (t_{k+1} - t_k)$. Bunne et al. perform a single step instead, i.e. $\tau = (t_{k+1} - t_k)$[18]. In our experiments, multiple steps did not improve results (see Supplementary Figure 1A) so we chose the computationally less expensive single-step method.

Teacher-forcing. Hashimoto et al. and Yeo et al.[16,17] predict $\rho_{t_{k+1}}$ from ρ_{t_k}. Bunne et al. introduce teacher-forcing, i.e. predicting $\rho_{t_{k+1}}$ from μ_{t_k}[18]. In our experiments, teacher-forcing improved results (see Supplementary Figure 1B), so we used it throughout this work.

Discretization scheme. To predict $\rho_{t+\tau}$ from an earlier population ρ_t, we used the forward Euler discretization scheme $\rho_{t+\tau} = (\mathrm{Id} - \tau\nabla J_\theta)_{\#}\rho_t$ as Hashimoto et al. and Yeo et al.[16,17]. In our discrete setting, this corresponds to $\mathbf{x}_{t+\tau} = \mathbf{x}_t - \tau\nabla J_\theta(\mathbf{x}_t)$ for each cell x. Bunne et al. propose using a backward Euler scheme to improve stability for large τ[18]. However we did not find this to improve results in our experiments (see Supplementary Figure 1C), so we used the computationally less expensive forward method.

Pairwise information

Our setting differs from previous works[16–18] in that for each cell $\mathbf{x} \in \mathbb{R}^d$ we have access to spatial coordinates $\mathbf{r} \in \mathbb{R}^2$. Spatial coordinates are defined up to an isometry. For instance, a slice of an embryo may be rotated without changing the problem. Consequently, one cannot simply concatenate $\mathbf{x}$ and $\mathbf{r}$ to leverage the spatial coordinates. The next paragraph details how we used the coordinates $\mathbf{r}$ to inform the problem, while still defining $J_\theta : \mathbf{x} \in \mathbb{R}^d \mapsto J_\theta(\mathbf{x}) \in \mathbb{R}$ only on gene expression.

Learning the potential

For each $k \in [1, \dots, K]$, we compared the prediction ρ_{t_k} to the reference snapshot μ_{t_k}.

Linear model. Let us consider two discrete probability distributions $\mu = \sum_i a_i \delta_{\mathbf{x}_i}$ and $\rho = \sum_j b_j \delta_{\mathbf{y}_j} \in \mathcal{P}(\mathbb{R}^d)$. To compare μ and ρ, previous works[16–18] use the Sinkhorn divergence[76], defined as

$$\overline{\mathcal{W}}_\varepsilon(\mu,\rho) = \mathcal{W}_\varepsilon(\mu,\rho) - \tfrac{1}{2}\big(\mathcal{W}_\varepsilon(\mu,\mu) + \mathcal{W}_\varepsilon(\rho,\rho)\big),$$

where $\mathcal{W}_\varepsilon$ is the entropy-regularized Optimal Transport[77] (entropy-regularized OT), defined as

$$\mathcal{W}_\varepsilon(\mu,\rho) = \min_{\mathbf{P}\in\mathcal{U}(\mathbf{a},\mathbf{b})} \sum_{i,j} \|\mathbf{x}_i - \mathbf{y}_j\|_2^2 \, \mathbf{P}_{i,j} - \varepsilon \mathrm{E}(\mathbf{P}).$$

Here, $\mathcal{U}(\mathbf{a},\mathbf{b}) \overset{\text{def.}}{=} \{\mathbf{P} \in \mathbb{R}_+^{n\times m} \text{ such that } \sum_j P_{ij} = a_i \text{ and } \sum_i P_{ij} = b_j\}$ and $\mathrm{E}(\mathbf{P}) = -\sum_{i,j} P_{i,j}\,(\log P_{i,j} - 1)$ is the Shannon entropy. $\overline{\mathcal{W}}_\varepsilon$ is a debiased version of $\mathcal{W}_\varepsilon$, such that $\forall \rho, \overline{\mathcal{W}}_\varepsilon(\rho,\rho) = 0$.

Quadratic model. The Sinkhorn divergence between ρ and μ only compares distributions of gene expression. Instead, we propose a debiased Fused Gromov-Wasserstein (debiased FGW) loss to enforce the spatial coherence of the predictions. Let us consider $\mu = \sum_i a_i \delta_{(\mathbf{x}_i,\mathbf{r}_i)}$ and $\rho = \sum_j b_j \delta_{(\mathbf{y}_j,\mathbf{s}_j)} \in \mathcal{P}(\mathbb{R}^d \times \mathbb{R}^2)$. Gromov-Wasserstein (GW) is a quadratic extension of OT well suited to compare measures defined up to an isometry[78]. A debiased version of GW has been used to learn a Generative Adversarial Network (GAN)[79]. Fused Gromov-Wassertein (FGW) combines a linear and a quadratic OT term[80]. In our setting, it is natural to use the linear term for gene expression and the quadratic term for spatial coordinates:

$$\mathcal{FGW}_\alpha^\varepsilon(\mu,\rho) = \min_{\mathbf{P}\in\mathcal{U}(\mathbf{a},\mathbf{b})} (1-\alpha)\mathrm{L}(\mathbf{P}) + \alpha \mathrm{Q}(\mathbf{P}) - \varepsilon \mathrm{E}(\mathbf{P}),$$

$$\text{where } \mathrm{L}(\mathbf{P}) = \sum_{i,j} \|\mathbf{x}_i - \mathbf{y}_j\|_2^2 \, \mathbf{P}_{i,j}$$

$$\mathrm{Q}(\mathbf{P}) = \sum_{i,j,i',j'} \left| \|\mathbf{r}_i - \mathbf{r}_{i'}\|_2^2 - \|\mathbf{s}_j - \mathbf{s}_{j'}\|_2^2 \right|^2 \mathbf{P}_{i,j}\mathbf{P}_{i',j'}.$$

Analogously to Bunne et al., we introduce a debiased FGW to ensure the loss vanishes for an exact match[79].

$$\overline{\mathcal{FGW}_\alpha^\varepsilon}(\mu,\rho) = \mathcal{FGW}_\alpha^\varepsilon(\mu,\rho) - \tfrac{1}{2}\big(\mathcal{FGW}_\alpha^\varepsilon(\mu,\mu) + \mathcal{FGW}_\alpha^\varepsilon(\rho,\rho)\big)$$

A weight $\alpha = 1$ corresponds to the debiased entropy-regularized GW[81]. As $\alpha = 0$, we recover the Sinkhorn divergence $\overline{\mathcal{W}_\varepsilon}$.

Final loss function. For $\alpha \in [0, 1[$, the full objective reads

$$\mathcal{L}(\theta) = \sum_{k=2}^{K} (t_k - t_{k-1}) \overline{\mathcal{FGW}_\alpha^\varepsilon}(\mu_{t_k}, \rho_{t_k})$$

where $\rho_{t_k} = (f_\theta)_\# \mu_{t_{k-1}}$ with $f_\theta : (\mathbf{x}, \mathbf{r}) \mapsto \left(\mathbf{x} - \tau \nabla J_\theta(\mathbf{x}), \mathbf{r}\right)$. In other words, gene expression is predicted with teacher-forcing and a single forward step (see "Discretization"), and the spatial coordinates remain unchanged. This loss is optimized using minibatches, as done by Yeo et al. and Bunne et al.[17,18] A theoretical study of minibatch Optimal Transport for debiased linear and quadratic OT is provided by Fatras et al.[82]

Choice of the quadratic weight

To investigate the effect of the relative weight of the linear term in FGW, we reported results for $\alpha \in \{1 \times 10^{-5}, 1 \times 10^{-4}, 1 \times 10^{-3}, 5 \times 10^{-3}, 1 \times 10^{-2}, 1 \times 10^{-1}\}$. Choosing $\alpha = 1$ would not allow learning the potential J_θ because it would ignore gene expression. As expected, with a low weight ($\alpha = 10^{-5}$), STORIES behaves as the linear method (see Section 2 of Results).

To compare our approach with the linear model proposed in previous works[16–18], we also trained the model with a Sinkhorn divergence, i.e. $\alpha = 0$.

The benchmark in Section 2 of Results suggests good performances for values of α of the order of 10^{-3}, with $\alpha = 5 \times 10^{-3}$ performing best. The value of α can be adjusted by the user depending on the dataset. In Section 3 and Section 4 of Results we set a value of $\alpha = 10^{-3}$.

Computational Optimal Transport

OTT solvers

We use the OTT package to solve OT problems in a fast, GPU-enabled, and differentiable manner[29]. In particular, we rely on the Sinkhorn and GromovWasserstein solvers. We set the entropic regularization $\varepsilon = 0.01$.

Linear term: gene expression

The linear OT terms are defined on gene expression space, for which we chose the d first components of the Harmony-aligned Principal Component Analysis. We chose $d = 20$ components as Bunne et al.[18] Before training the neural network, we normalized the points as $\mathbf{x}_i \leftarrow \frac{\mathbf{x}_i}{\max_j \|\mathbf{x}_j\|_\infty}$ to make the linear and quadratic terms comparable.

Quadratic term: Spatial coordinates

Before training the neural network and for each slice separately, we centered the spatial coordinates and scaled them to unit variance to make linear and quadratic terms comparable.

Neural network architecture

We implemented a Multi-Layer Perceptron (MLP) with two hidden layers of dimension 128 and GeLU activations using Flax[83]. This is a similar architecture to previous works[16–18]. The linear output layer has no bias since it would not influence the values of ∇J_θ. Likewise, a soft activation like GeLU is preferable to the classical ReLU because we manipulate ∇J_θ. Indeed, the derivative of ReLU is simply a unit step function, which is discontinuous and not very expressive.

Neural network training

Data loading

For each time point described in “Data Collection", 75% of the cells were used as training samples and 25% as validation. At each training or validation iteration, a batch containing 1,000 cells per time point was sampled uniformly without replacement. In development, the early time points contain fewer cells than the later time points. If less than 1,000 cells were available for a time point, we used all available cells. To reflect the train/validation split, one in four iterations performs a validation step.

Optimizer

We used Optax's implementation of the AdamW optimizer, with parameters `b1=0.9`, `b2=0.999`, `eps=1e-8`, and `weight_decay=1e-4`[28,84]. We set the learning rate using Optax's cosine scheduler, with an initial value of `1e-2` and `10000` decay steps. To ensure convergence, when performing 10 steps, we set the learning rate to `1e-3`. Similarly, when performing an implicit step, we set the learning rate to `1e-4`.

Early stopping

We set the maximum number of iterations to 15,000 but stopped the training when the validation loss had not improved in 150 iterations. We kept the weights associated with the lowest validation loss.

Seeds

We ran every experiment with 10 random seeds: `17158`, `20181`, `12409`, `5360`, `21712`, `21781`, `24802`, `13630`, `9668`, and `651`. The random seed reproducibly determines the train/validation split and weight initialization. For the plots in Figure 2C, and the analysis in Section 3 and Section 4 of Results, the experiments correspond to the randomly chosen seed `20181`.

Computational growth rate

Weight of marginals

When comparing a prediction ρ_{t_k} to the reference snapshot μ_{t_k}, Yeo et al. proposes setting the weights b_j of the prediction $\rho_{t_k} = \sum_j b_j \delta_{\mathbf{y}_j}$ proportionally to a computationally derived growth rate[17]. The motivation is that cells with a larger growth rate should be matched to more descendants. This idea was introduced in Waddington OT and recently reimplemented in MOSCOT[13,22]. In the next paragraphs, we follow MOSCOT's implementation.

Proliferation and apoptosis

Computing the growth rate relies on cell-wise proliferation and apoptosis scores. We used Scanpy's `sc.tl.score_genes` with lists of genes collected from the literature. The gene lists are described in more detail at the end of the section. We calculated gene scores on raw counts, after quality filtering of cells and genes.

Calculating the growth rate

For a given cell $\mathbf{x}_i$, let us call $\mathrm{prol}_i \in \mathbb{R}$ the proliferation score and $\mathrm{apo}_i \in \mathbb{R}$ the apoptosis score. We then define the birth rate $\beta_i \in \mathbb{R}$ and the death rate $\gamma_i \in \mathbb{R}$ as

$$\beta_i = \beta_{\min} + \frac{\beta_{\max}}{1 + \exp\left(-4.0 \times \frac{\mathrm{prol}_i - \beta_{\text{center}}}{\beta_{\text{width}}}\right)}, \quad \gamma_i = \gamma_{\min} + \frac{\gamma_{\max}}{1 + \exp\left(-4.0 \times \frac{\mathrm{apo}_i - \gamma_{\text{center}}}{\gamma_{\text{width}}}\right)}$$

We used MOSCOT's default parameters $\beta_{\min} = \gamma_{\min} = 0.3$, $\beta_{\max} = \gamma_{\max} = 1.7$, $\beta_{\text{center}} = 0.25$, $\beta_{\text{width}} = 0.5$, $\gamma_{\text{center}} = 0.1$, $\gamma_{\text{width}} = 0.2$. Finally, we defined the cell's growth rate as

$$g_i = \exp\left(\Delta t \times (\beta_i - \gamma_i)\right)$$

where Δt is the time difference between populations. We obtained a_i by normalizing the growth rate

$$a_i = \frac{g_i}{\sum_j g_j}, \quad \text{i.e.} \quad \mathbf{a} = \mathrm{softmax}\left(\Delta t \times (\beta - \gamma)\right).$$

In this equation, Δt plays the role of the softmax's inverse temperature. The histogram $\mathbf{a}$ will thus be sharper for large values of Δt. Most slices in our experiments were evenly sampled, so we set a fixed $\Delta t = 1$, which yielded sharp enough weight differences between cell types.

In the benchmark

Previous works tested gene sets to compute the growth rate in the case of mice and humans, but not of zebrafish and axolotl[13,17,22]. We thus kept uniform marginals for the benchmark in Section 2 of Results. In Section 3 and Section 4 of Results, we analyzed a restricted number of cell types, where we could validate the biological coherence of the computed growth rate and of the learned potential (see details in paragraphs below).

Dorsal midbrain

For Section 3 of Results, we used murine proliferation and apoptosis gene sets from MOSCOT. Proliferation genes come from https://doi.org/10.1038/nature20123 and apoptosis genes from https://www.gsea-msigdb.org/gsea/msigdb/cards/HALLMARK_P53_PATHWAY. See Supplementary Text 1 and Supplementary Text 2 for the gene names.

Axolotl neuron regeneration

For Section 4 of Results, we used an NSC axolotl gene set described in the original publication https://www.science.org/doi/10.1126/science.abp9444 to represent proliferation and a human apoptosis gene set from MOSCOT, originally from gsea-msigdb's HALLMARK_APOPTOSIS. See Supplementary Text 3 and Supplementary Text 4 for the gene names.

Evaluation

We solve the FGW problem with $\alpha = 10^{-3}$ and $\varepsilon = 10^{-3}$ and report the terms $\mathrm{L}(\mathbf{P})$ and $\mathrm{Q}(\mathbf{P})$ separately. Since $\mathrm{L}(\mathbf{P})$ quantifies the error in terms of gene expression, we call this quantity the *gene expression prediction score*. Similarly, since $\mathrm{Q}(\mathbf{P})$ quantifies the error in terms of spatial coordinates, we call this quantity the *spatial coherence score*.

Figure 3C compares the transport plans involved in the linear method and STORIES with $\alpha = 5 \times 10^{-3}$. Formally, let us consider the indicator vector $\mathbf{a} \in \mathbb{R}^n$ where $a_i = 1$ if the i-th cell corresponds to a given cell type, and $a_i = 0$ otherwise. The transport plan $\mathbf{P}$ between the prediction ρ_t and the ground truth μ_t is applied to the indicator, yielding a vector $\mathbf{b} = \mathbf{P}^\top \mathbf{a}$ representing the mass transported from $\mathbf{a}$ towards each cell in the second time point.

Trajectory inference

Gene imputation

In our analysis, the gene expression trends (Figure 3 and Figure 4) would be negatively affected by the sparsity of gene expression. Cellrank demonstrated good performances in identifying gene expression trends with MAGIC[27,85]. We thus applied MAGIC gene imputation after all other preprocessing steps. We computed the exponentiated Markov transition matrix on the Harmony-aligned PCA space instead of the original PCA. We did not use the imputed signal for tasks other than gene expression trends.

Potential visualization

The neural potential J_θ is a functional defined on the 20-dimensional space of Harmony-aligned principal components. To visualize the potential as a Waddington-like landscape defined on 2 dimensions, we proceed similarly to Qin et al.[86].

- First, we compute the potential $J_\theta(\mathbf{x})$ associated with each cell $\mathbf{x}$.
- Then, we use Scipy's RBF interpolation and the 2-dimensional Isomap coordinates of the cells to define a potential on a 2-dimensional grid[87].
- The cells are projected on the surface using the interpolator.
- Finally, the maximum value is thresholded.

We rendered the resulting surface and point cloud using Blender's Python API.

Cell-cell transition matrix

We used CellRank's VelocityKernel with a velocity $v(\mathbf{x}) = -\nabla J_\theta(\mathbf{x})$ for a trained J_θ. Based on this kernel, we computed a cell-cell transition matrix, and the trajectory plots Figure 3B, Figure 4B.

Gene expression trends

We fit the MAGIC-imputed gene expression as a function of the learned potential, using Scipy's Spline Regression[87]. To plot the gene expression cascades in Figure 3D, Figure 4D, we order genes by the value of potential for which the maximum regressed expression value is achieved. Genes are then split into equally sized groups illustrating different stages of differentiation (10 groups for Figure 3D and 2 groups for Figure 4D). Finally, regressed values for the genes with the best regression scores in each group are displayed (3 per group in Figure 3D and 15 per group in Figure 4D).

Transcription factor enrichment

We perform TF-target enrichment based on the TRRUST dataset, which contains “Activation", "Repression", and "Unknown" links based on a curated literature search[88]. For each TF in the database, we perform a Wilcoxon rank-sum test comparing the list of regression scores of its target genes, and the list of regression scores of the other genes. Figure 3B and Figure 4B display the TFs ranked by p-value.

Data availability

- We retrieved the mouse Stereo-seq atlas from Chen et al.[4], available at https://db.cngb.org/stomics/mosta/.
- We retrieved the zebrafish Stereo-seq atlas from Liu et al.[6], available at https://db.cngb.org/stomics/zesta/.
- We retrieved the axolotl Stereo-seq atlas from Wei et al.[7], available at https://db.cngb.org/stomics/artista/.
- We retrieved mouse proliferation genes from Tirosh et al.[89] (see Supplementary Text 1) and apoptosis genes from gsea-msigdb's HALLMARK_P53_PATHWAY (see Supplementary Text 2).
- We retrieved an NSC axolotl gene set from Wei et al.[7] (see Supplementary Text 3) and a human apoptosis gene set from gsea-msigdb's HALLMARK_APOPTOSIS (see Supplementary Text 4).

Code availability

The Python package for STORIES is hosted at https://github.com/cantinilab/stories. It can be installed easily by running “pip install stories-jax”. Code to reproduce the experiments and figures is available at https://github.com/cantinilab/stories_reproducibility/.

Figure captions

Figure 1. Overview of STORIES. (A) STORIES takes as an input spatial transcriptomics through time. **(B)** STORIES learns the parameters θ of a neural network J_θ representing the differentiation potential of a cell based on its transcriptomic profile. The objective function is based on Fused Gromov-Wasserstein, which leverages both the transcriptomic profile and the spatial coordinates. **(C)** The gradient of the function J_θ delivers a velocity that can be used to perform trajectory inference. The potential itself is a natural alternative to pseudotime and allows the study of gene trends along differentiation. Finally, STORIES can highlight possible transcription factors regulating differentiation.

Figure 2. Benchmark of STORIES on three large datasets. (A) Visual representation of the three datasets in our benchmark. From left to right, an axolotl brain regeneration dataset, zebrafish development, and mouse development. Slices in each dataset are split into a train set (blue), an early test set (orange), and a late test set (green). **(B)** Gene expression prediction score (top) and spatial coherence score (bottom) in early test sets (orange) and late test sets (green) across the three datasets. Scores are reported for ten seeds, across seven values of quadratic weight parameter α, including the linear method ($\alpha = 0$, light orange/green). **(C)** Visual representation of the optimal transport matching involved in the loss of the linear method (top) and STORIES for $\alpha = 5 \times 10^{-3}$ (bottom), for the three datasets. In each dataset, the left slice displays two cell types and the right slice displays the cells they are matched with at the following time point.

Figure 3. Trajectory inference with STORIES in axolotl neuron regeneration. (A) 3-D representation of the potential landscape learned with STORIES. The *x* and *y* axes are Isomap coordinates, and the *z*-axis is an interpolation of the potential. Colors represent cell types involved in the regeneration process. **(B)** Visual representation of cell-cell transitions computed using CellRank from STORIES's velocity vectors. **(C)** Smoothed gene expression for *Vim* and *Nptx1* along the potential computed by STORIES**.** The blue line is a spline regression of expression from potential. **(D)** Normalized gene expression regressed using a spline model along the potential computed by STORIES**.** Genes are ordered by the potential for which they achieve maximum expression **(E)** Smoothed gene expression for *Hes5* and *Nsg2* along the potential computed by STORIES**.** The blue line is a spline regression of expression from potential. **(F)** Enrichment scores of Transcription Factors targeting candidate driver genes.

Figure 4. Trajectory inference with STORIES in mouse gliogenesis. (A) 3-D representation of the potential landscape learned with STORIES. The *x* and *y* axes are Isomap coordinates, and the *z-axis* is an interpolation of the potential. Colors represent radial glial cells which differentiate into either neuroblasts or glioblasts. **(B)** Visual representation of cell-cell transitions computed using CellRank from STORIES's velocity vectors. **(C)** Smoothed gene expression for *Mki67* and *Aldh1l1* along the potential computed by STORIES**.** The blue line is a spline regression of expression from potential. **(D)** Normalized gene expression regressed using a spline model along the potential computed by STORIES**.** Genes are ordered by the potential for which they achieve maximum expression **(E)** Smoothed gene expression for *Tuba1b* and *Glis3* along the potential computed by STORIES**.** The blue line is a spline

regression of expression from potential. **(F)** Enrichment scores of transcription factors targeting candidate driver genes.

References

1. Palla, G., Fischer, D. S., Regev, A. & Theis, F. J. Spatial components of molecular tissue biology. *Nat. Biotechnol.* **40**, 308–318 (2022).
2. Hunter, M. V., Moncada, R., Weiss, J. M., Yanai, I. & White, R. M. Spatially resolved transcriptomics reveals the architecture of the tumor-microenvironment interface. *Nat. Commun.* **12**, 6278 (2021).
3. Chen, W.-T. *et al.* Spatial transcriptomics and in situ sequencing to study Alzheimer's disease. *Cell* **182**, 976–991 (2020).
4. Chen, A. *et al.* Spatiotemporal transcriptomic atlas of mouse organogenesis using DNA nanoball-patterned arrays. *Cell* **185**, 1777–1792 (2022).
5. Vickovic, S. *et al.* High-definition spatial transcriptomics for in situ tissue profiling. *Nat. Methods* **16**, 987–990 (2019).
6. Liu, C. *et al.* Spatiotemporal mapping of gene expression landscapes and developmental trajectories during zebrafish embryogenesis. *Dev. Cell* **57**, 1284–1298 (2022).
7. Wei, X. *et al.* Single-cell Stereo-seq reveals induced progenitor cells involved in axolotl brain regeneration. *Science* **377**, eabp9444 (2022).
8. Saelens, W., Cannoodt, R., Todorov, H. & Saeys, Y. A comparison of single-cell trajectory inference methods. *Nat. Biotechnol.* **37**, 547–554 (2019).
9. Trapnell, C. *et al.* The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.* **32**, 381–386 (2014).
10. Tritschler, S. *et al.* Concepts and limitations for learning developmental trajectories from single cell genomics. *Development* **146**, (2019).
11. La Manno, G. *et al.* RNA velocity of single cells. *Nature* **560**, 494–498 (2018).
12. Bergen, V., Soldatov, R. A., Kharchenko, P. V. & Theis, F. J. RNA velocity—current challenges and future perspectives. *Mol. Syst. Biol.* **17**, e10282 (2021).

13. Schiebinger, G. *et al.* Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming. *Cell* **176**, 928–943 (2019).

14. Tong, A., Huang, J., Wolf, G., Van Dijk, D. & Krishnaswamy, S. Trajectorynet: A dynamic optimal transport network for modeling cellular dynamics. in *International conference on machine learning* 9526–9536 (PMLR, 2020).

15. Huguet, G. *et al.* Manifold interpolating optimal-transport flows for trajectory inference. *Adv. Neural Inf. Process. Syst.* **35**, 29705–29718 (2022).

16. Hashimoto, T., Gifford, D. & Jaakkola, T. Learning population-level diffusions with generative RNNs. in *International Conference on Machine Learning* 2417–2426 (PMLR, 2016).

17. Yeo, G. H. T., Saksena, S. D. & Gifford, D. K. Generative modeling of single-cell time series with PRESCIENT enables prediction of cell trajectories with interventions. *Nat. Commun.* **12**, 3222 (2021).

18. Bunne, C., Papaxanthos, L., Krause, A. & Cuturi, M. Proximal optimal transport modeling of population dynamics. in *International Conference on Artificial Intelligence and Statistics* 6511–6528 (PMLR, 2022).

19. Huang, S., Guo, Y.-P., May, G. & Enver, T. Bifurcation dynamics in lineage-commitment in bipotent progenitor cells. *Dev. Biol.* **305**, 695–713 (2007).

20. Allen, M. Compelled by the diagram: thinking through CH Waddington’s epigenetic landscape. *Contemporaneity* **4**, 119 (2015).

21. Shen, X. *et al.* Inferring cell trajectories of spatial transcriptomics via optimal transport analysis. *bioRxiv* 2023–09 (2023).

22. Klein, D. *et al.* Mapping cells through time and space with moscot. *bioRxiv* 2023–05 (2023).

23. Peng, Q., Zhou, P. & Li, T. stVCR: Reconstructing spatio-temporal dynamics of cell development using optimal transport. *bioRxiv* 2024–06 (2024).

24. Vayer, T., Chapel, L., Flamary, R., Tavenard, R. & Courty, N. Optimal transport for structured data with application on graphs. *ArXiv Prepr. ArXiv180509114* (2018).

25. Virshup, I. *et al.* The scverse project provides a computational ecosystem for single-cell omics data analysis. *Nat. Biotechnol.* **41**, 604–606 (2023).

26. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 15 (2018).

27. Weiler, P., Lange, M., Klein, M., Pe'er, D. & Theis, F. Unified fate mapping in multiview single-cell data. (2023) doi:10.1101/2023.07.19.549685.

28. DeepMind *et al.* The DeepMind JAX Ecosystem. (2020).

29. Cuturi, M. *et al.* Optimal Transport Tools (OTT): A JAX Toolbox for all things Wasserstein. *ArXiv Prepr. ArXiv220112324* (2022).

30. Peyré, G., Cuturi, M., & others. Computational optimal transport: With applications to data science. *Found. Trends® Mach. Learn.* **11**, 355–607 (2019).

31. Villani, C. & others. *Optimal Transport: Old and New*. vol. 338 (Springer, 2009).

32. Jordan, R., Kinderlehrer, D. & Otto, F. The variational formulation of the Fokker–Planck equation. *SIAM J. Math. Anal.* **29**, 1–17 (1998).

33. Street, K. *et al.* Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. *BMC Genomics* **19**, 1–16 (2018).

34. Chen, H. *et al.* Single-cell trajectories reconstruction, exploration and mapping of omics data with STREAM. *Nat. Commun.* **10**, 1903 (2019).

35. Bergen, V., Lange, M., Peidli, S., Wolf, F. A. & Theis, F. J. Generalizing RNA velocity to transient cell states through dynamical modeling. *Nat. Biotechnol.* **38**, 1408–1414 (2020).

36. Palla, G. *et al.* Squidpy: a scalable framework for spatial single cell analysis. *BioRxiv* 2021–02 (2021).

37. Ohtsuka, T., Sakamoto, M., Guillemot, F. & Kageyama, R. Roles of the basic helix-loop-helix genes Hes1 and Hes5 in expansion of neural stem cells of the developing brain. *J. Biol. Chem.* **276**, 30467–30474 (2001).

38. Peco, E. *et al.* The CDC25B phosphatase shortens the G2 phase of neural progenitors and promotes efficient neuron production. *Development* **139**, 1095–1104

(2012).

39. Bonnet, F. *et al.* Neurogenic decisions require a cell cycle independent function of the CDC25B phosphatase. *Elife* **7**, e32937 (2018).

40. Nunez, J. & Fischer, I. Microtubule-associated proteins (MAPs) in the peripheral nervous system during development and regeneration. *J. Mol. Neurosci.* **8**, 207–222 (1997).

41. Schäfer, M. K. & Frotscher, M. Role of L1CAM for axon sprouting and branching. *Cell Tissue Res.* **349**, 39–48 (2012).

42. Chander, P., Kennedy, M. J., Winckler, B. & Weick, J. P. Neuron-specific gene 2 (NSG2) encodes an AMPA receptor interacting protein that modulates excitatory neurotransmission. *eneuro* **6**, (2019).

43. Marchetti, B. Wnt/β-catenin signaling pathway governs a full program for dopaminergic neuron survival, neurorescue and regeneration in the MPTP mouse model of Parkinson's disease. *Int. J. Mol. Sci.* **19**, 3743 (2018).

44. Duraikannu, A., Martinez, J. A., Chandrasekhar, A. & Zochodne, D. W. Expression and manipulation of the APC-β-catenin pathway during peripheral neuron regeneration. *Sci. Rep.* **8**, 13197 (2018).

45. Kawakami, Y. *et al.* Wnt/β-catenin signaling regulates vertebrate limb regeneration. *Genes Dev.* **20**, 3232–3237 (2006).

46. Miras-Portugal, M. T. *et al.* Nucleotides in neuroregeneration and neuroprotection. *Neuropharmacology* **104**, 243–254 (2016).

47. Zhang, Y., Zhao, Q., Chen, Q., Xu, L. & Yi, S. Transcriptional control of peripheral nerve regeneration. *Mol. Neurobiol.* **60**, 329–341 (2023).

48. Yang, S.-G., Wang, X.-W., Qian, C. & Zhou, F.-Q. Reprogramming neurons for regeneration: The fountain of youth. *Prog. Neurobiol.* **214**, 102284 (2022).

49. Jhamb, D. *et al.* Network based transcription factor analysis of regenerating axolotl limbs. *BMC Bioinformatics* **12**, 1–12 (2011).

50. Tolchin, D. *et al.* De novo SOX6 variants cause a neurodevelopmental syndrome

associated with ADHD, craniosynostosis, and osteochondromas. *Am. J. Hum. Genet.* **106**, 830–845 (2020).

51. Chen, J. & Guan, Z. Function of oncogene Mycn in adult neurogenesis and oligodendrogenesis. *Mol. Neurobiol.* **59**, 77–92 (2022).

52. Ballas, N. & Mandel, G. The many faces of REST oversee epigenetic programming of neuronal genes. *Curr. Opin. Neurobiol.* **15**, 500–506 (2005).

53. Cheng, Y. *et al.* Transcription factor network analysis identifies REST/NRSF as an intrinsic regulator of CNS regeneration in mice. *Nat. Commun.* **13**, 4418 (2022).

54. He, F. & Sun, Y. E. Glial cells more than support cells? *Int. J. Biochem. Cell Biol.* **39**, 661–665 (2007).

55. Laug, D., Glasgow, S. M. & Deneen, B. A glial blueprint for gliomagenesis. *Nat. Rev. Neurosci.* **19**, 393–403 (2018).

56. Tirosh, I. *et al.* Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science* **352**, 189–196 (2016).

57. Hausrat, T. J., Radwitz, J., Lombino, F. L., Breiden, P. & Kneussel, M. Alpha-and beta-tubulin isotypes are differentially expressed during brain development. *Dev. Neurobiol.* **81**, 333–350 (2021).

58. Anlauf, E. & Derouiche, A. Glutamine synthetase as an astrocytic marker: its cell type and vesicle localization. *Front. Endocrinol.* **4**, 144 (2013).

59. Jurga, A. M., Paleczna, M., Kadluczka, J. & Kuter, K. Z. Beyond the GFAP-astrocyte protein markers in the brain. *Biomolecules* **11**, 1361 (2021).

60. Liu, Z., Liu, L., Qi, Y., Li, H. & Pan, S. GLIS family zinc finger 3 promoting cell malignant behaviors and NF-κB signaling in glioma. *Brain Res.* **1770**, 147623 (2021).

61. Braccioli, L., Vervoort, S. J., Puma, G., Nijboer, C. H. & Coffer, P. J. SOX4 inhibits oligodendrocyte differentiation of embryonic neural stem cells in vitro by inducing Hes5 expression. *Stem Cell Res.* **33**, 110–119 (2018).

62. Taylor, M. K., Yeager, K. & Morrison, S. J. Physiological Notch signaling promotes gliogenesis in the developing peripheral and central nervous systems. *Development* **134**,

2435–2447 (2007).

63. Hirvonen, H. *et al.* Differential expression of myc, max and RB1 genes in human gliomas and glioma cell lines. *Br. J. Cancer* **69**, 16–25 (1994).

64. Andersson, A. *et al.* Single-cell and spatial transcriptomics enables probabilistic inference of cell type topography. *Commun. Biol.* **3**, 565 (2020).

65. Kleshchevnikov, V. *et al.* Cell2location maps fine-grained cell types in spatial transcriptomics. *Nat. Biotechnol.* **40**, 661–671 (2022).

66. Chen, K. H., Boettiger, A. N., Moffitt, J. R., Wang, S. & Zhuang, X. Spatially resolved, highly multiplexed RNA profiling in single cells. *Science* **348**, aaa6090 (2015).

67. Weinreb, C., Wolock, S., Tusi, B. K., Socolovsky, M. & Klein, A. M. Fundamental limits on dynamic inference from single-cell snapshots. *Proc. Natl. Acad. Sci.* **115**, E2467–E2476 (2018).

68. Terpin, A., Lanzetti, N. & Dörfler, F. Learning Diffusion at Lightspeed. *ArXiv Prepr. ArXiv240612616* (2024).

69. Zhou, R., Yang, G., Zhang, Y. & Wang, Y. Spatial transcriptomics in development and disease. *Mol. Biomed.* **4**, (2023).

70. Turing, A. M. The chemical basis of morphogenesis. *Bull. Math. Biol.* **52**, 153–197 (1990).

71. Zhu, J., Zhang, Y.-T., Alber, M. S. & Newman, S. A. Bare bones pattern formation: a core regulatory network in varying geometries reproduces major features of vertebrate limb development and evolution. *PLoS One* **5**, e10892 (2010).

72. Economou, A. D. *et al.* Periodic stripe formation by a Turing mechanism operating at growth zones in the mammalian palate. *Nat. Genet.* **44**, 348–351 (2012).

73. Korsunsky, I. *et al.* Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat. Methods* **16**, 1289–1296 (2019).

74. Ambrosio, L., Gigli, N. & Savaré, G. *Gradient Flows: In Metric Spaces and in the Space of Probability Measures.* (Springer Science & Business Media, 2005).

75. Santambrogio, F. {Euclidean, metric, and Wasserstein} gradient flows: an overview.

Bull. Math. Sci. **7**, 87–154 (2017).

76. Genevay, A., Peyré, G. & Cuturi, M. Learning generative models with Sinkhorn divergences. in *International Conference on Artificial Intelligence and Statistics* 1608–1617 (2018).

77. Cuturi, M. Sinkhorn Distances: Lightspeed Computation of Optimal Transport. in *Advances in Neural Information Processing Systems* (eds. Burges, C. J., Bottou, L., Welling, M., Ghahramani, Z. & Weinberger, K. Q.) vol. 26 (Curran Associates, Inc., 2013).

78. Mémoli, F. Gromov–Wasserstein distances and the metric approach to object matching. *Found. Comput. Math.* **11**, 417–487 (2011).

79. Bunne, C., Alvarez-Melis, D., Krause, A. & Jegelka, S. Learning generative models across incomparable spaces. *ArXiv Prepr. ArXiv190505461* (2019).

80. Vayer, T., Chapel, L., Flamary, R., Tavenard, R. & Courty, N. Fused Gromov-Wasserstein distance for structured objects: theoretical foundations and mathematical properties. *ArXiv Prepr. ArXiv181102834* (2018).

81. Peyré, G., Cuturi, M. & Solomon, J. Gromov-wasserstein averaging of kernel and distance matrices. in *International Conference on Machine Learning* 2664–2672 (2016).

82. Fatras, K. *et al.* Minibatch optimal transport distances; analysis and applications. *ArXiv Prepr. ArXiv210101792* (2021).

83. Heek, J. *et al.* Flax: A neural network library and ecosystem for JAX. (2023).

84. Loshchilov, I. & Hutter, F. Decoupled weight decay regularization. *ArXiv Prepr. ArXiv171105101* (2017).

85. Van Dijk, D. *et al.* Recovering gene interactions from single-cell data using data diffusion. *Cell* **174**, 716–729 (2018).

86. Qin, X. *et al.* A single-cell perturbation landscape of colonic stem cell polarisation. *bioRxiv* 2023–02 (2023).

87. Virtanen, P. *et al.* SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nat. Methods* **17**, 261–272 (2020).

88. Han, H. *et al.* TRRUST v2: an expanded reference database of human and mouse

transcriptional regulatory interactions. *Nucleic Acids Res.* **46**, D380–D386 (2018).

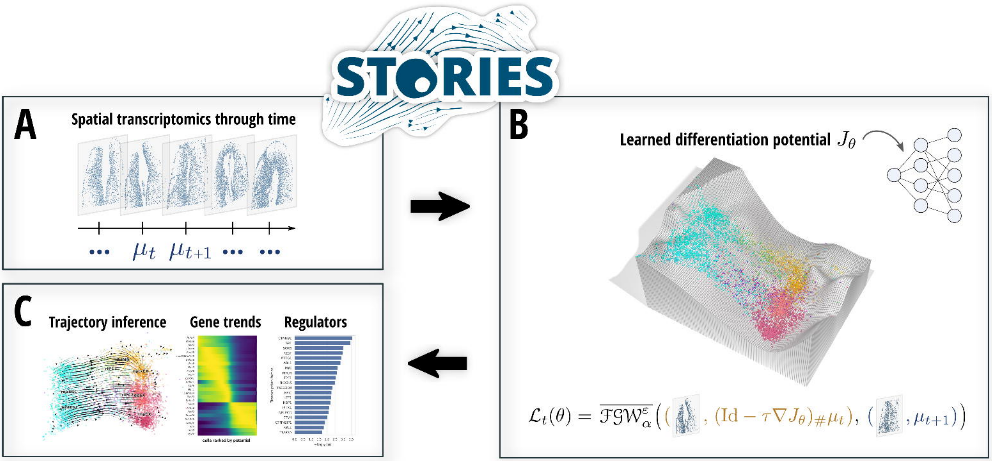
89. Tirosh, I. *et al.* Single-cell RNA-seq supports a developmental hierarchy in human oligodendroglioma. *Nature* **539**, 309–313 (2016).

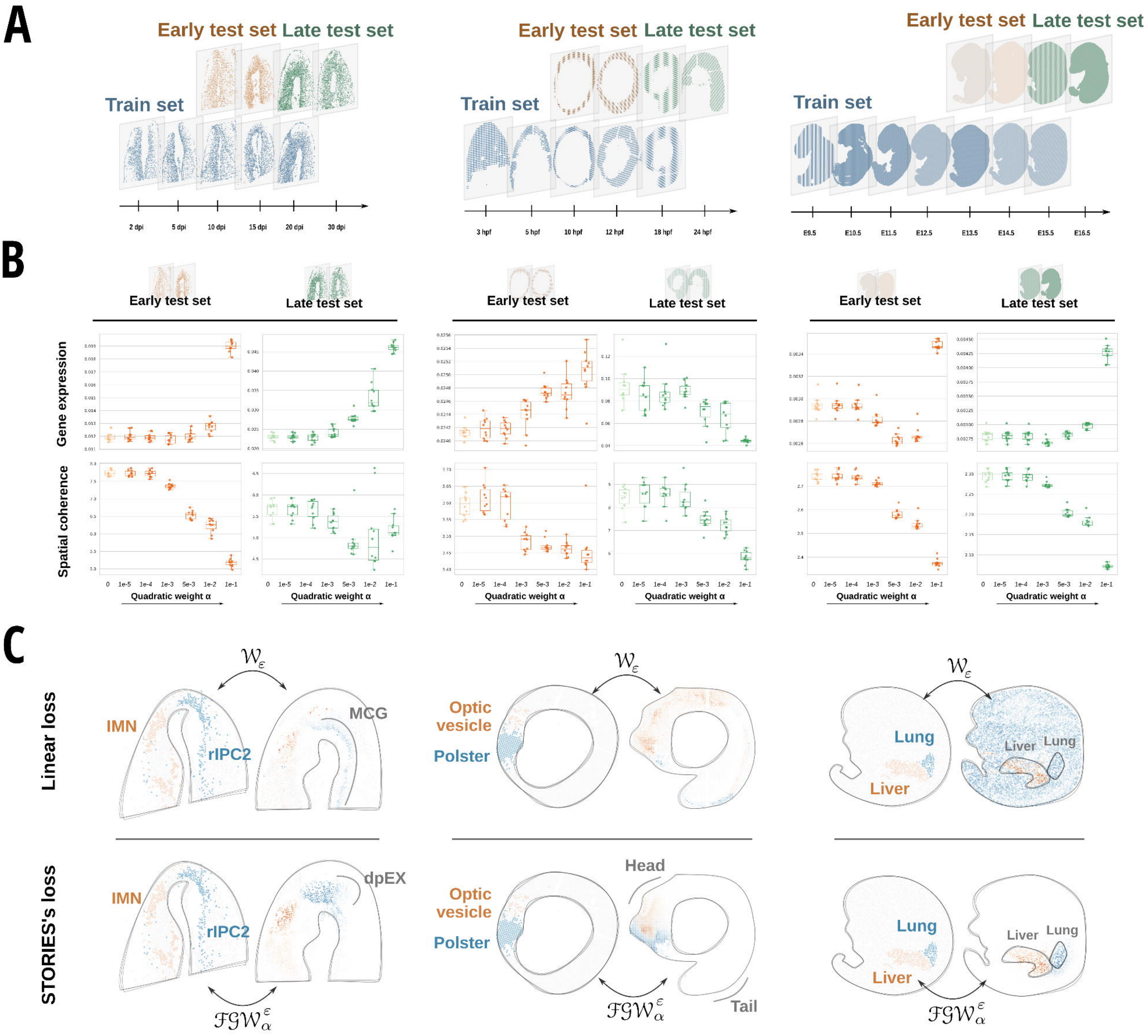
Acknowledgements

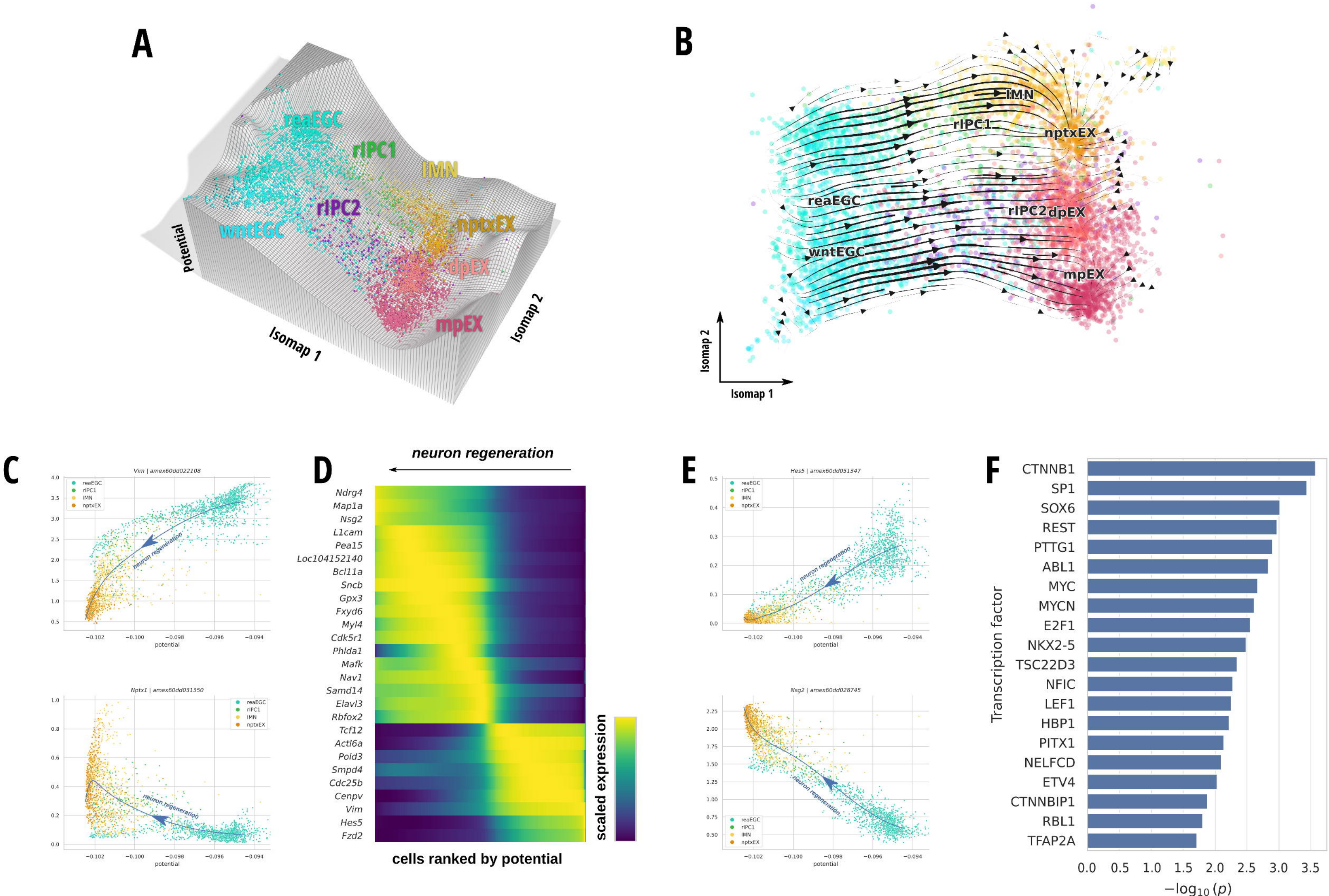
The project leading to this manuscript has received funding from the European Union (ERC StG, MULTIview-CELL, 101115618). In addition, this work has been funded by the French government under management of Agence Nationale de la Recherche as part of the "Investissements d'avenir" program, reference ANR-19-P3IA-0001 (PRAIRIE 3IA Institute) and by the Inception program (Investissement d’Avenir grant ANR-16-CONV-0005). The work of G. Peyré was supported by the European Research Council (ERC project NORIA) and the French government under management of Agence Nationale de la Recherche as part of the ‘Investissements d’avenir’ program, reference ANR19-P3IA-0001 (PRAIRIE 3IA Institute). We acknowledge the help of the HPC Core Facility of the Institut Pasteur and Déborah Philipps for the administrative support.

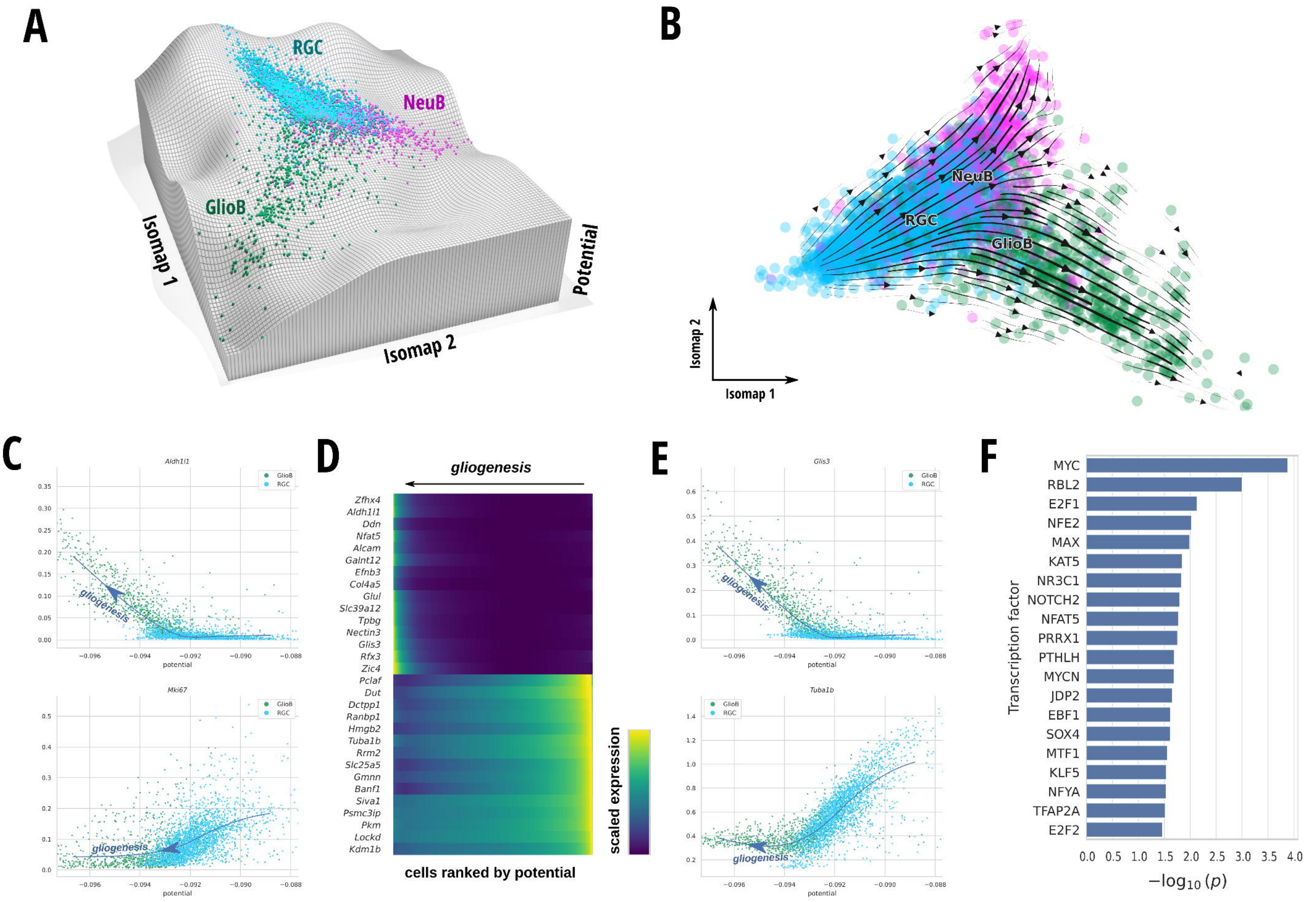
Competing interests

The authors declare no competing interests.


STORIES
A
Spatial transcriptomics through time
$\cdots$ μ_t μ_{t+1} $\cdots$ $\cdots$
B
Learned differentiation potential J_θ
$\mathcal{L}_t(\theta) = \overline{\mathcal{FGW}}^{\varepsilon}_{\alpha}\big((\ , (\mathrm{Id} - \tau\nabla J_\theta)_{\#}\mu_t), (\ , \mu_{t+1})\big)$
C
Trajectory inference
Gene trends
Regulators
cells ranked by potential


A
Early test set
Late test set
Train set
2 dpi
5 dpi
10 dpi
15 dpi
20 dpi
30 dpi
3 hpf
5 hpf
10 hpf
12 hpf
18 hpf
24 hpf
E9.5
E10.5
E11.5
E12.5
E13.5
E14.5
E15.5
E16.5
B
Early test set
Late test set
Gene expression
Spatial coherence
0
1e-5
1e-4
1e-3
5e-3
1e-2
1e-1
Quadratic weight α
C
Linear loss
STORIES's loss
$\mathcal{W}_\varepsilon$
$\mathcal{FGW}_\alpha^\varepsilon$
IMN
rIPC2
MCG
dpEX
Optic vesicle
Polster
Head
Tail
Lung
Liver


A
reaEGC
rIPC1
IMN
rIPC2
wntEGC
nptxEX
dpEX
mpEX
potential
Isomap 1
Isomap 2
B
IMN
rIPC1
nptxEX
reaEGC
rIPC2
dpEX
wntEGC
mpEX
Isomap 2
Isomap 1
C
Vim | amex60dd022108
reaEGC
rIPC1
IMN
nptxEX
neuron regeneration
potential
Nptx1 | amex60dd031350
neuron regeneration
potential
D
neuron regeneration
Ndrg4
Map1a
Nsg2
L1cam
Pea15
Loc104152140
Bcl11a
Sncb
Gpx3
Fxyd6
Myl4
Cdk5r1
Phlda1
Mafk
Nav1
Samd14
Elavl3
Rbfox2
Tcf12
Actl6a
Pold3
Smpd4
Cdc25b
Cenpv
Vim
Hes5
Fzd2
scaled expression
cells ranked by potential
E
Hes5 | amex60dd051347
neuron regeneration
potential
Nsg2 | amex60dd028745
neuron regeneration
potential
F
CTNNB1
SP1
SOX6
REST
PTTG1
ABL1
MYC
MYCN
E2F1
NKX2-5
TSC22D3
NFIC
LEF1
HBP1
PITX1
NELFCD
ETV4
CTNNBIP1
RBL1
TFAP2A
Transcription factor
0.0 0.5 1.0 1.5 2.0 2.5 3.0 3.5
$-\log_{10}(p)$


A
RGC
NeuB
GlioB
Isomap 1
Isomap 2
Potential
B
NeuB
RGC
GlioB
Isomap 2
Isomap 1
C
Aldh1l1
GlioB
RGC
gliogenesis
potential
Mki67
GlioB
RGC
gliogenesis
potential
D
gliogenesis
Zfhx4
Aldh1l1
Ddn
Nfat5
Alcam
Galnt12
Efnb3
Col4a5
Glul
Slc39a12
Tpbg
Nectin3
Glis3
Rfx3
Zic4
Pclaf
Dut
Dctpp1
Ranbp1
Hmgb2
Tuba1b
Rrm2
Slc25a5
Gmnn
Banf1
Siva1
Psmc3ip
Pkm
Lockd
Kdm1b
scaled expression
cells ranked by potential
E
Glis3
GlioB
RGC
gliogenesis
potential
Tuba1b
GlioB
RGC
gliogenesis
potential
F
MYC
RBL2
E2F1
NFE2
MAX
KAT5
NR3C1
NOTCH2
NFAT5
PRRX1
PTHLH
MYCN
JDP2
EBF1
SOX4
MTF1
KLF5
NFYA
TFAP2A
E2F2
Transcription factor
0.0 0.5 1.0 1.5 2.0 2.5 3.0 3.5 4.0
$-\log_{10}(p)$