Department of Mathematics
TUM School of Computation, Information and Technology
Technical University of Munich

A multi-modal model of RNA velocity describing nuclear export and splicing kinetics

Robin Mittas

Thesis for the attainment of the academic degree

Master of Science

at the TUM School of Computation, Information and Technology of the Technical University of Munich

Supervisor:
Prof. Dr. Fabian J. Theis

Advisors:
Dr. Fabiola Curion
Philipp Weiler

Submitted:
Munich, 27.10.2023

I hereby declare that this thesis is entirely the result of my own work except where otherwise indicated. I have only used the resources given in the list of references.

Munich, 27.10.2023 Robin Mittas

Abstract

Recent advances in single-cell technologies allow us to measure transcriptomic and epitomic information simultaneously. For instance, it is possible to profile regions of open chromatin or surface proteins together with transcriptomic protocols. These technologies include the extraction of nuclei, therefore resulting in single-nucleus RNA transcripts. To guide the interpretation of transcriptional dynamics, the concept of RNA velocity has been introduced and rapidly adopted. However, splicing dynamics underlying RNA velocity are defined for single-cell sequencing transcripts which measure cellular RNA abundances spanning the nucleus and cytoplasm. Here, we investigate the inference of RNA velocity of single-nucleus transcripts and propose a new model to integrate single-cell and single-nucleus measurements into a unified system. The proposed model, referred to as *Nucleus-cytosol model*, describes splicing and nuclear-exporting dynamics by introducing an additional kinetic parameter that characterizes the nuclear export rate per gene. This enhancement allows us to describe nucleic and cytosolic RNA abundances, resulting in a more precise description of the underlying biological processes governing the life cycle of RNA. As a first step, we show the model's ability to learn kinetic parameters with a high correlation to true rate parameters and latent variables of simulated data. However, to apply the model to real-world sequencing datasets, we need to estimate the missing abundances: For single-cell protocols, we lack information regarding the proportions of nucleic and cytosolic RNA, whereas single-nucleus protocols only capture nucleic RNA. Therefore, we introduced a method to estimate the respective missing abundances for both measurements. Finally, we apply the model on subpopulation kinetics in pancreatic endocrinogenesis and benchmark the *Nucleus-cytosol model* against the single-modal models solely trained on either single-cell or nucleus transcripts. We show that the single-nucleus model exhibits lower permutation scores compared to the single-cell and *Nucleus-cytosol* models, yet it performs equally in terms of cell cycle phase transition probabilities. Despite the observation that many cycling genes in the single-nucleus dataset do not conform to the underlying assumptions of RNA velocity, the velocity estimates still provide valuable information about a cell's future state. Here, we explored uncertainty measures of the velocity vector field that can be used to gain insights into a cell's fate and future state. We further investigated and compared the inferred velocity vector fields of the single-modal models and the *Nucleus-cytosol model*. Specifically, we evaluated velocity confidences as well as velocity correlations between models. We observed that the proposed *Nucleus-cytosol model* has significantly higher velocity correlations to the single-cell model compared to the single-nucleus model. This observation further highlights the necessity of a unified system to integrate single-cell and nucleus measurements.

Acknowledgements

First of all, I want to thank Prof. Dr. Fabian Theis for giving me the opportunity to write my master thesis within his research group at Helmholtz Zentrum Munich. I truly believe that the Theis lab with all members is shaping and defining the future of single-cell research and it was an honor to be part of such a great research team and contribute to its amazing achievements. I am really thankful for all the interesting insights.

I owe particular thanks to Dr. Fabiola Curion and Philipp Weiler who have defined the scope and topic of my thesis. I am genuinely grateful for the chance to work on such an exciting research topic as RNA velocity.

I further want to express my deep gratitude to my main supervisor, Philipp Weiler, who guided me through this thesis. His excellent expertise, insightful ideas, and very constructive feedback not only enhanced my critical thinking and independence but also elevated the scientific rigor of this work.

Lastly, I want to thank my family and my partner Karolina for supporting me throughout my entire master's studies. I am really grateful for all the love and support you give me.

Contents

1 Introduction

Understanding the mechanisms underlying cellular differentiation, disease progression and identifying genes driving such dynamics, requires us to study the process of cellular development. Traditionally, bulk RNA sequencing has been used to study gene expression of entire biological samples, resulting in cell-averaged expression profiles [1]. However, as bulk measurements do not separate different cell types within a sample, they might hide some of the complexity such as cell expression profile heterogeneity. For instance, it is possible that only particular cell types or interactions between cell types are affected by drugs or perturbations. Consider oncology as an example, here we might encounter tumor cells that are resistant to drugs, leading to a relapse. These cells are challenging if not impossible to detect using bulk transcripts, even when performed on cultured cells [1].

Consequently, cellular differentiation, disease progression, and cellular heterogeneity need to be considered at the level of single cells. Nowadays, with the introduction of single-cell RNA-sequencing (scRNA-seq), it is possible to profile gene expression of thousands of individual cells in parallel. The technology allows studying cellular heterogeneity at an unprecedented resolution [2], discovering previously unknown cellular populations [3][4], and reconstructing trajectories of cellular state dynamics and investigating cell fate decisions [5]. However, as common single-cell sequencing protocols include cell lysis, *i.e.*, breaking down the cell membrane, they are destructive by nature. Therefore, cells can only be measured once, and their gene expression can not be tracked over time [1].

Recently live-sequencing has been introduced, which enables temporal transcriptomic recording of single cells [6]. However, live-sequencing is experimentally challenging and expensive to conduct [1]. Consequently, given the measured snapshot data, the underlying dynamical process needs to be estimated instead.

The mechanisms driving cellular differentiation are asynchronous, and this characteristic allows cells to be ordered along a trajectory. Within the task of trajectory inference, we can estimate a differentiation tree of cells by leveraging the asynchronous nature of dynamical biological processes [1]. Furthermore, we can assign a so-called pseudotime to a cell, indicating its relative stage and position in the developmental process compared to other cells. Trajectory inference is a well-studied field, providing different methods to construct cell-specific pseudotimes [7]. However, common approaches are solely based on gene expression counts, they usually require specifying an initial root cell, *i.e.*, where the overall process starts, and they cannot infer the directions or relative rates of cellular transitions [7][8].

To overcome the limitations of traditional methods for inferring pseudotime, and to model the dynamics in biological systems over time, the concept of RNA velocity has been proposed [9]. It uses the fact that unspliced and spliced mRNA reads can be distinguished in common scRNA-seq protocols. The latter is detectable by the presence of introns, *i.e.*, non-coding regions of unspliced mRNA [9]. RNA velocity describes the change of spliced mRNA over time and, thus, gives an estimate of the future state of a single cell, which can then be used as a starting point for the application of further trajectory inference methods. The concept of RNA velocity comes with the advantage that root cell specification is not necessary, while simultaneously incorporating additional directional information determined by splicing dynamics [7].

By considering the respective measurements of unspliced and spliced mRNA, splicing dynamics can directly be modeled. To actively drive change in expression, DNA is first transcribed to produce unspliced, precursor messenger-RNA (pre-mRNA). Following, newly transcribed pre-mRNA is transformed into mature, spliced mRNA by a process called splicing. Within this process, intronic regions are removed and adjacent exons, *i.e.*, expressed regions of unspliced mRNA, are spliced together to produce mature, spliced mRNA. Finally, mature mRNA is degraded. This dynamical behavior can mathematically be described by a two-dimensional set of differential equations [10]. RNA velocity is then defined as the derivative of spliced counts with respect to time, which is determined by the production of spliced mRNA from un-

spliced mRNA, and the mRNA degradation [9]. Therefore, the velocity estimates can be used to infer the direction and magnitude of gene expression changes at the single-cell level in order to predict the future state of a cell on a timescale of hours [9].
The first model to infer RNA velocity, which we will refer to as the *steady-state model*, made two fundamental simplifying assumptions: (1) All genes share a common splicing rate and (2) steady states are observed. However, these assumptions are restricting its potential usefulness. On the one hand, splicing is a gene-dependent property with potentially largely varying rates between genes. On the other hand, if transcription ends prematurely before mRNA level saturation, steady states are oftentimes not observed [11][10].
To overcome the assumptions of the *steady-state model*, the *EM-model* (Expectation-Maximization model) has been introduced [10]. The *EM-model* solves the full gene-wise transcriptional dynamics and infers gene-specific transcription, splicing, and degradation rates, as well as a gene-shared latent time. Time needs to be inferred as a hidden variable as it is unobserved due to the destructive nature of sequencing protocols [10]. Further, assuming that cell-gene pairs can be associated with a state (induction, repression, steady state), we receive a second hidden, unobserved variable. Following, in order to solve the dynamical model for the rate parameters and to infer the latent variables, an Expectation-Maximization (EM) algorithm is required. Although the *EM-model* enables unobserved steady states to be faithfully captured while inferring gene-specific rate parameters [10], the model still assumes full-gene-wise independence as well as constant kinetic rate parameters, again leading to incorrectly inferred cellular state changes in hematopoietic systems, for example, [12][13]. The lack of extensibility and flexibility of the outlined methods motivated new approaches to infer the parameters as well as the cell's hidden states.
To adapt to more complicated scenarios, recent advances in generative modeling allow us to reformulate the inference problem of RNA velocity within an easily extensible framework. To this end, *veloVI*, a deep-learning-based model, has recently been proposed [14]. VeloVI first encodes the unspliced and spliced abundances of each cell into a low-dimensional latent variable, called the cell representation. This embedding captures the notion that the observed state of a cell is a composition of multiple concomitant processes that together span the phenotypic manifold [15]. The cell representation is then used to assign a transcriptional state to each gene in the cell. Following, the cell-gene-specific latent time is inferred as a function of state assignment and cell representation. The justification for this modeling choice stems from the observation that when employing the *EM-model*, which is fitted independently for each gene, the inferred latent time matrix (with dimensions cells by genes) demonstrates a low-rank structure. It is noteworthy, however, that this low-rank structure is not of rank one [14] justifying that *veloVI* assigns a latent time to each gene within a cell. By coupling latent time and state via the cell representation, the model relaxes the gene-wise independence assumption. Finally, the cell-gene-specific latent time and transcriptional state are used to solve the kinetic equations, fit kinetic parameters, and compute RNA velocity. Due to the flexible nature of the model, it can be adapted to more complex real-world scenarios [14].
So far, the outlined methods start to model the dynamical process with the transcription of DNA to produce unspliced pre-mRNA and end the modeling with the degradation of spliced mRNA. However, splicing dynamics in general neither begins with a sudden transcription nor does it end with the degradation of mature mRNA. Cells express proteins called transcription factors, which bind to DNA-encoding sequence elements called transcription factor binding sites, which play a major role in determining the position of transcription initiation and the frequency of transcription. These binding sites are regions of open chromatin. Afterwards, DNA-bound transcription factors recruit the RNA polymerase which initiates transcription of a nearby gene. Once mRNA is transcribed and fully processed in the nucleus, mature mRNA is exported to the cytoplasm, where mRNA is translated into functional proteins before degradation [16]. Therefore, to capture a more accurate picture of cellular processes, it is in our interest to measure regions of open chromatin or protein levels in single-cells [1]. The quantification of these aspects is a critical step towards understanding cell differentiation and fate, disease progression, and clinical diagnostics [1].
New sequencing technologies allow measuring regions of open chromatin (ATAC-seq [17]) as well as surface proteins (CITE-seq [18]). These modalities can simultaneously be profiled with transcriptomic protocols, resulting in multi-modal sequencing information of single cells. To measure chromatin accessibility

and gene expression in parallel, nuclei are first extracted before DNA fragments of open regions are tagged [17]. Nuclei are then loaded onto the sequencing machines, thus resulting in gene expression profiles of the nucleus (snRNA-seq), instead of the entire cell. So far, little work has been done to incorporate the additional modalities into the estimation of RNA velocity. To incorporate chromatin accessibility, *MultiVelo* [8] has been proposed. This model has been showcased on snRNA-seq data, where the assumptions of RNA velocity, such as constant degradation and nuclear export have not been conclusively verified [19]. Therefore, it is still unclear how RNA velocity performs on snRNA-seq data compared to scRNA-seq. Our objective is therefore twofold: (1) Study the applicability of the existing method to estimate RNA velocity based on snRNA-seq measurements; (2) propose a new model integrating snRNA-seq and scRNA-seq measurements in one system.

Here, we introduce a model which integrates single-cell and nucleus data into a unified system. We thereby describe unspliced and spliced abundances in the nucleus, as well as spliced abundance in the cytoplasm. Therefore, we extended the differential equations describing splicing dynamics by introducing a rate parameter describing the nuclear export of spliced nucleic RNA. This enhancement allows us to depict the kinetics of splicing and nuclear exporting through a three-dimensional system of ordinary differential equations (ODEs), relating the provided quantities of unspliced and spliced mRNA in the nucleus and cytoplasm over time. In this work, we have analytically solved the three-dimensional system of ODEs and extended the existing veloVI framework. Future perspectives based on this work include the integration of further modalities, such as chromatin accessibility, to form a biological more sound model. The transcriptional rate can then be inferred based on the ATAC-seq measurements, resulting in a time-dependent transcriptional rate, similar to the *MultiVelo*-model or the suggested extension presented in [14].

2 Model specification and mathematical foundations

2.1 Model definition and analytical solution

We model the splicing dynamics by the following three-dimensional set of ordinary differential equations (ODEs). Within this work, we will refer to this model as the *Nucleus-cytosol model*. Explanations for the parameters can be found in Table 2.1.

$$\begin{aligned}
\frac{du_n^{(g)}(t)}{dt} &= \alpha_{gk} - \beta_g u_n^{(g)}(t) \\
\frac{ds_n^{(g)}(t)}{dt} &= \beta_g u_n^{(g)}(t) - \nu_g s_n^{(g)}(t) \\
\frac{ds_c^{(g)}(t)}{dt} &= \nu_g s_n^{(g)}(t) - \gamma_g s_c^{(g)}(t)
\end{aligned} \tag{2.1}$$

Notation	Description
Subscript (g)	Gene g
t	Time (unobserved)
$u_n^{(g)}$	Unspliced abundance of pre-mRNA in the nucleus of gene g
$s_n^{(g)}$	Spliced abundance of mRNA in the nucleus of gene g
$s_c^{(g)}$	Spliced abundance of mRNA in the cytoplasm of gene g
$k \in \{1, 2, 3, 4\}$	Different States: Induction ($k = 1$), Induction Steady State (SS) ($k = 2$), Repression ($k = 3$), Repression SS ($k = 4$)
α_{gk}	Gene-state specific transcription rate
β_g	Splicing rate in nucleus of gene g
ν_g	Nuclear export rate of spliced mRNA for gene g
γ_g	Degradation rate of spliced mRNA in cytoplasm
Index i	Cell
Index pair ig	Cell-gene pair
Subscript/ index pair gk	Gene-state
t_g^s	Switching time of gene g: System switches from induction to repression state
$G \in \mathbb{N}$	Number of genes
$N \in \mathbb{N}$	Number of cells

Table 2.1 Notations used within this work.

The differential equation (2.1) can be rewritten as $\dot{x}(t) = Ax(t) + f(t)$ which is equivalent to

$$\begin{pmatrix} \dot{u}_n^{(g)}(t) \\ \dot{s}_n^{(g)}(t) \\ \dot{s}_c^{(g)}(t) \end{pmatrix} = \begin{pmatrix} -\beta_g & 0 & 0 \\ \beta_g & -\nu_g & 0 \\ 0 & \nu_g & -\gamma_g \end{pmatrix} \begin{pmatrix} u_n^{(g)}(t) \\ s_n^{(g)}(t) \\ s_c^{(g)}(t) \end{pmatrix} + \begin{pmatrix} \alpha_{gk} \\ 0 \\ 0 \end{pmatrix}.$$

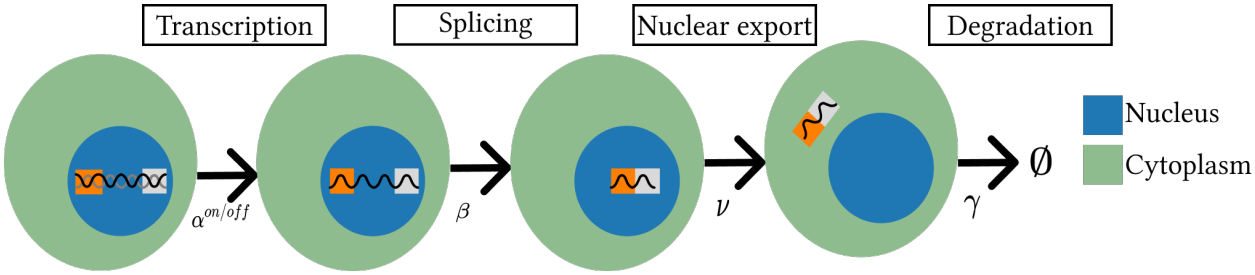

Figure 2.1 Splicing and exporting dynamics are modeled as follows: First, DNA is transcribed into unspliced precursor mRNA, which is then transformed into mature spliced mRNA. This entire process occurs within the cell nucleus. Afterward, spliced nucleic mRNA is transported from the nucleus into the cytoplasm before its final degradation.

Following, the matrix exponential of $A \in \mathbb{R}^{3\times 3}$, given its eigen-decomposition QDQ^{-1}, can be calculated with Theorem A.3.1 as

$$e^{tA} = Q \begin{pmatrix} e^{-t\beta_g} & 0 & 0 \\ 0 & e^{-t\nu_g} & 0 \\ 0 & 0 & e^{-t\gamma_g} \end{pmatrix} Q^{-1}$$

Let the initial conditions for gene g be given by $(u_n^{(g)}(t_0), s_n^{(g)}(t_0), u_c^{(g)}(t_0)) = (u_{n0}^{(g)}, s_{n0}^{(g)}, u_{c0}^{(g)})$ at time point t_0. Finally, applying Theorem A.3.2 yields the following analytical solution to the differential equations

$$u_n^{(g)}(t) = u_{n0}^{(g)} e^{-\beta_g(t-t_0)} + \frac{\alpha_{gk}}{\beta_g}(1 - e^{-\beta_g(t-t_0)})$$

$$s_n^{(g)}(t) = s_{n0}^{(g)} e^{-\nu_g(t-t_0)} + \frac{\alpha_{gk}}{\nu_g}(1 - e^{-\nu_g(t-t_0)}) + \frac{a_{gk} - \beta_g u_{n0}^{(g)}}{\nu_g - \beta_g}(e^{-\nu_g(t-t_0)} - e^{-\beta_g(t-t_0)})$$

$$\begin{aligned} s_c^{(g)}(t) = \nu_g\beta_g \Bigg(& \frac{\frac{\alpha_{gk}}{\beta_g}(1 - e^{-\beta_g(t-t_0)}) + u_{n0}^{(g)} e^{-\beta_g(t-t_0)}}{(\nu_g - \beta_g)(\gamma_g - \beta_g)} - \frac{\frac{\alpha_{gk}}{\nu_g}(1 - e^{-\nu_g(t-t_0)}) + u_{n0}^{(g)} e^{-\nu_g(t-t_0)}}{(\gamma_g - \nu_g)(\nu_g - \beta_g)} \\ & + \frac{\frac{\alpha_{gk}}{\gamma_g}(1 - e^{-\gamma_g(t-t_0)}) + u_{n0}^{(g)} e^{-\gamma_g(t-t_0)}}{(\gamma_g - \nu_g)(\gamma_g - \beta_g)} \Bigg) \\ & + \frac{\nu_g}{\gamma_g - \nu_g}(e^{-\nu_g(t-t_0)} - e^{-\gamma_g(t-t_0)}) s_{n0}^{(g)} + e^{-\gamma_g(t-t_0)} s_{c0}^{(g)}. \end{aligned} \tag{2.2}$$

2.2 State specifications

Transcriptional regulation involves the cellular control of DNA to RNA conversion, thereby effectively coordinating gene activity. This suggests that changes in the state-dependent transcriptional rate α_{gk} are responsible for the regulation of both gene upregulation and downregulation [10]. Consequently, the initial conditions as well as the initial time t_0^{gk} of the system in state k are state-dependent, resulting in $(u_n^{gk}(t_0^{gk}), s_n^{gk}(t_0^{gk}), u_c^{gk}(t_0^{gk})) = (u_{n0}^{gk}, s_{n0}^{gk}, u_{c0}^{gk})$.
Given the solution to the ODEs in equations (2.2), the unspliced abundance of RNA in the nucleus of the cell i and gene g at time t_{ig} is thus given by

$$\overline{u}_n^{(g)}(t_{ig}, k) := u_{n0}^{gk} e^{-\beta_g(t_{ig} - t_0^{gk})} + \frac{\alpha_{gk}}{\beta_g}(1 - e^{-\beta_g(t_{ig} - t_0^{gk})}). \tag{2.3}$$

Similarly, the spliced transcript abundance in the nucleus and the cytoplasm is defined as

$$\bar{s}_n^{(g)}(t_{ig}, k) := s_{n0}^{gk} e^{-\nu_g (t_{ig} - t_0^{gk})} + \frac{\alpha_{gk}}{\nu_g}(1 - e^{-\nu_g (t_{ig} - t_0^{gk})}) + \frac{a_{gk} - \beta_g u_{n0}^{gk}}{\nu_g - \beta_g}(e^{-\nu_g (t_{ig} - t_0^{gk})} - e^{-\beta_g (t_{ig} - t_0^{gk})}) \tag{2.4}$$

$$\begin{aligned}\bar{s}_c^{(g)}(t_{ig}, k) := \nu_g \beta_g \Bigg(& \frac{\frac{\alpha_{gk}}{\beta_g}(1 - e^{-\beta_g (t_{ig} - t_0^{gk})}) + u_{n0}^{gk} e^{-\beta_g (t_{ig} - t_0^{gk})}}{(\nu_g - \beta_g)(\gamma_g - \beta_g)} - \frac{\frac{\alpha_{gk}}{\nu_g}(1 - e^{-\nu_g (t_{ig} - t_0^{gk})}) + u_{n0}^{gk} e^{-\nu_g (t_{ig} - t_0 gk)}}{(\gamma_g - \nu_g)(\nu_g - \beta_g)} \\ & + \frac{\frac{\alpha_{gk}}{\gamma_g}(1 - e^{-\gamma_g (t_{ig} - t_0^{gk})}) + u_{n0}^{gk} e^{-\gamma_g (t_{ig} - t_0^{gk})}}{(\gamma_g - \nu_g)(\gamma_g - \beta_g)} \Bigg) \\ & + \frac{\nu_g}{\gamma_g - \nu_g}(e^{-\nu_g (t_{ig} - t_0^{gk})} - e^{-\gamma_g (t_{ig} - t_0^{gk})}) s_{n0}^{gk} + e^{-\gamma_g (t_{ig} - t_0^{gk})} s_{c0}^{gk}.\end{aligned} \tag{2.5}$$

This model assumes for a gene g, that cells are at the initial time of the system in induction phase $k = 1$. Once transcription starts with rate α_{g1}, the transcript abundances (u_n, s_n, s_c) increase. Spliced mRNA is produced from unspliced mRNA and therefore the abundances of s_n increase with a a temporal delay with respect to u_n. Similarly, the abundance of s_c increases with a temporal delay with respect to s_n, as spliced mRNA in the cytoplasm s_c is regulated by the nuclear export rate of spliced nucleic RNA s_n. Eventually, cells reach the induction steady state. Next, at the switching time point t_g^s the system switches from induction to repression state, where the unspliced and spliced expression starts decreasing. Finally, cells reach the repression steady state in which there is no expression. This yields the following transcript abundances for the respective states.

Induction state

For the induction state, $k = 1$ we have $u_{n0}^{g1} = s_{n0}^{g1} = s_{c0}^{g1} = 0$, $\alpha_{g1} > 0$ and $t_0^{g1} = 0$. The transcript abundance then simplify to

$$\bar{u}_n^{(g)}(t_{ig}, k = 1) := \frac{\alpha_{g1}}{\beta_g}(1 - e^{-\beta_g t_{ig}})$$

$$\bar{s}_n^{(g)}(t_{ig}, k = 1) := \frac{\alpha_{g1}}{\nu_g}(1 - e^{-\nu_g t_{ig}}) + \frac{a_{g1}}{\nu_g - \beta_g}(e^{-\nu_g t_{ig}} - e^{-\beta_g t_{ig}})$$

$$\bar{s}_c^{(g)}(t_{ig}, k = 1) := \nu_g \beta_g \left(\frac{\frac{\alpha_{g1}}{\beta_g}(1 - e^{-\beta_g t_{ig}})}{(\nu_g - \beta_g)(\gamma_g - \beta_g)} - \frac{\frac{\alpha_{g1}}{\nu_g}(1 - e^{-\nu_g t_{ig}})}{(\gamma_g - \nu_g)(\nu_g - \beta_g)} + \frac{\frac{\alpha_{g1}}{\gamma_g}(1 - e^{-\gamma_g t_{ig}})}{(\gamma_g - \nu_g)(\gamma_g - \beta_g)} \right)$$

Induction steady state

For the Induction steady state $k = 2$ the abundances are defined as the time limits of the system in state $k = 1$. Thus, we receive

$$\bar{u}_n^{(g)}(t_{ig}, k = 2) := \lim_{t_{ig} \to \infty} \bar{u}_n^{(g)}(t_{ig}, k = 1) = \frac{\alpha_{g1}}{\beta_g} \tag{2.6}$$

$$\bar{s}_n^{(g)}(t_{ig}, k = 2) := \lim_{t_{ig} \to \infty} \bar{s}_n^{(g)}(t_{ig}, k = 1) = \frac{\alpha_{g1}}{\nu_g} \tag{2.7}$$

$$\bar{s}_c^{(g)}(t_{ig}, k = 2) := \lim_{t_{ig} \to \infty} \bar{s}_c^{(g)}(t_{ig}, k = 1)$$

$$= \nu_g \beta_g \left(\frac{\frac{\alpha_{g1}}{\beta_g}}{(\nu_g - \beta_g)(\gamma_g - \beta_g)} - \frac{\frac{\alpha_{g1}}{\nu_g}}{(\gamma_g - \nu_g)(\nu_g - \beta_g)} + \frac{\frac{\alpha_{g1}}{\gamma_g}}{(\gamma_g - \nu_g)(\gamma_g - \beta_g)} \right) \tag{2.8}$$

Repression state

For the repression state $k = 3$, we have $\alpha_{g3} = 0$ and $t_0^{g3} = t_g^s > 0$. The initial conditions for the repression state are defined by the abundances of the induction model at the switching time point t_g^s, such that

$$\begin{aligned} u_{n0}^{g3} &:= \overline{u}_n^{(g)}(t_g^s, k = 1) \\ s_{n0}^{g3} &:= \overline{s}_n^{(g)}(t_g^s, k = 1) \\ s_{c0}^{g3} &:= \overline{s}_c^{(g)}(t_g^s, k = 1). \end{aligned}$$

The abundance can then be simplified to

$$\begin{aligned} \overline{u}_n^{(g)}(t_{ig}, k = 3) &:= u_{n0}^{g3} e^{-\beta_g (t_{ig} - t_g^s)} \\ \overline{s}_n^{(g)}(t_{ig}, k = 3) &:= s_{n0}^{g3} e^{-\nu_g (t_{ig} - t_g^s)} - \frac{\beta_g u_{n0}^{g3}}{\nu_g - \beta_g} (e^{-\nu_g (t_{ig} - t_g^s)} - e^{-\beta_g (t_{ig} - t_g^s)}) \end{aligned}$$

$$\begin{aligned} \overline{s}_c^{(g)}(t_{ig}, k = 3) := \nu_g \beta_g &\left(\frac{e^{-\beta_g (t_{ig} - t_g^s)}}{(\nu_g - \beta_g)(\gamma_g - \beta_g)} - \frac{e^{-\nu_g (t_{ig} - t_g^s)}}{(\gamma_g - \nu_g)(\nu_g - \beta_g)} + \frac{e^{-\gamma_g (t_{ig} - t_g^s)}}{(\gamma_g - \nu_g)(\gamma_g - \beta_g)} \right) u_{n0}^{g3} \\ &+ \frac{\nu_g}{\gamma_g - \nu_g} (e^{-\nu_g (t_{ig} - t_g^s)} - e^{-\gamma_g (t_{ig} - t_g^s)}) s_{n0}^{g3} + e^{-\gamma_g (t_{ig} - t_g^s)} s_{c0}^{g3}. \end{aligned}$$

Repression steady state

For the repression steady state $k = 4$, the time limit for t_{ig} of the repression state is equal to 0, thus there is no expression

$$\overline{u}_n^{(g)}(t_{ig}, k = 4) := \lim_{t_{ig} \to \infty} \overline{u}_n^{(g)}(t_{ig}, k = 3) = 0$$

$$\overline{s}_n^{(g)}(t_{ig}, k = 4) := \lim_{t_{ig} \to \infty} \overline{s}_n^{(g)}(t_{ig}, k = 3) = 0$$

$$\overline{s}_c^{(g)}(t_{ig}, k = 4) := \lim_{t_{ig} \to \infty} \overline{s}_c^{(g)}(t_{ig}, k = 3) = 0$$

2.3 Generative process

Here, we assume that each gene g is on the same timescale for all cells. Therefore, we fix the latent time to the interval $t_{ig} \in [0, 20]$, resulting in $t_{\max} := 20$. We further use a latent dimension of $d = 10$ for the cell representation. Within the generative process for each cell we first sample the so-called cell-representation as

$$z_i \sim \mathcal{N}(0, I_d),$$

which summarizes the latent state of the cell i. Next, for each gene g in cell i we draw a distribution over the state assignments and assign a state to each gene as

$$\begin{aligned} \pi_{ig} &\sim \text{Dirichlet}(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}) \\ k_{ig} &\sim \text{Categorical}(\pi_{ig}). \end{aligned}$$

If $k_{ig} = 1$, then the cell-gene specific time is a function of the cell representation z_i, defined as

$$\rho_{ig}^{(1)} = [h_{\text{ind}}(z_i)]_g$$
$$t_{ig}^{(1)} = \rho_{ig}^{(1)} t_g^s,$$

where $h_{\text{ind}} : \mathbb{R}^d \to (0,1)^G$ is modelled as a fully connected neural network, with G denoting the number of genes. This function maps the cell representation to a gene-specific scaling factor, resulting in a latent time that is strictly smaller than the switching time t_g^s.
For the repression phase $k_{ig} = 3$, we similarly define a function $h_{\text{rep}} : \mathbb{R}^d \to (0,1)^G$, modeled as a fully connected network, resulting in the following scaling factor and latent time

$$\rho_{ig}^{(3)} = [h_{\text{ind}}(z_i)]_g$$
$$t_{ig}^{(3)} = (t_{\max} - t_g^s)\rho_{ig}^{(3)} + t_g^s.$$

If a gene g for a cell i is sampled within a steady state (e.g. $k_{ig} \in \{2, 4\}$), we consider the respective time limits. This is equivalent to

$$t_{ig}^{(2)} = \lim_{\rho_{ig} \to 1} t_{ig}^{(1)} = t_g^s$$
$$t_{ig}^{(4)} = \lim_{\rho_{ig} \to 1} t_{ig}^{(3)} = t_{\max}.$$

Finally, the observed data is sampled from normal distributions as

$$u_n^{(ig)} = \mathcal{N}(\overline{u}_n^{(g)}(t_{ig}^{k_{ig}}, k_{ig}), (c_k \sigma_g^{u_n})^2)$$
$$s_n^{(ig)} = \mathcal{N}(\overline{s}_n^{(g)}(t_{ig}^{k_{ig}}, k_{ig}), (c_k \sigma_g^{s_n})^2)$$
$$s_c^{(ig)} = \mathcal{N}(\overline{s}_c^{(g)}(t_{ig}^{k_{ig}}, k_{ig}), (c_k \sigma_g^{s_c})^2),$$

where $c_k = 1$ for $k \in \{1, 2, 3\}$ and $c_4 = 0.1$ is accounting for a state-dependent scaling factor on the variance. We consider the observed data to be smoothed expression data that has been preprocessed, such that for each gene the smoothed abundances are independently min-max scaled to values in $[0, 1]$.
In the following, we will refer to θ as the set of parameters of the generative process $(\alpha, \beta, \nu, \gamma, t^s)$ as well as the parameters of the neural network.

2.4 Inference

Variational posterior

The objective is to find (1) point estimates for θ, e.g. the gene-specific rate parameters (transcription, splicing, nuclear export, and degradation rate), the switching time point t^s and the parameters of the neural network, and (2) a posterior distribution over the hidden variables z and π. As the marginal likelihood defined as

$$p_\theta(u_n, s_n, s_c) = \sum_{k=1}^{4} \int p_\theta(z, \pi_k) p(u_n, s_n, s_c | z, \pi_k) dz$$

is intractable, we use variational inference. Following, we define the factorized approximate posterior distribution as

$$q_\phi(z, \pi | u_n, s_n, s_c) := \prod_{i=1}^{N} q_\phi(z_i | u_n^i, s_n^i, s_c^i) \prod_{g=1}^{G} q_\phi(\pi_{ig} | z_i),$$

where ϕ denotes the parameter set of the neural network.
The likelihoods for the unspliced and spliced abundances are a mixture of normal distributions of the different transcriptional states

$$p_\theta(u_n^{(ig)}|z_i,\pi_i) = \sum_{k_{ig}\in\{1,2,3,4\}} \pi_{igk_{ig}} \mathcal{N}(\overline{u}_n^{(g)}(t_{ig}^{k_{ig}},k_{ig}),(c_k\sigma_g^{u_n})^2) \tag{2.9}$$

$$p_\theta(s_n^{(ig)}|z_i,\pi_i) = \sum_{k_{ig}\in\{1,2,3,4\}} \pi_{igk_{ig}} \mathcal{N}(\overline{s}_n^{(g)}(t_{ig}^{k_{ig}},k_{ig}),(c_k\sigma_g^{s_n})^2) \tag{2.10}$$

$$p_\theta(s_c^{(ig)}|z_i,\pi_i) = \sum_{k_{ig}\in\{1,2,3,4\}} \pi_{igk_{ig}} \mathcal{N}(\overline{s}_c^{(g)}(t_{ig}^{k_{ig}},k_{ig}),(c_k\sigma_g^{s_c})^2) \tag{2.11}$$

Objective loss function

The objective that is minimized during inference is composed of two terms

$$\mathcal{L}_{\text{velo}}(\theta,\phi;u_n,s_n,s_c) = \mathcal{L}_{\text{ELBO}}(\theta,\phi;u_n,s_n,s_c) + \lambda\mathcal{L}_{\text{switch}}(\theta;u_n,s_n,s_c),$$

where $\mathcal{L}_{\text{ELBO}}$ is the evidence-lower-bound (ELBO) [20] on the log-likelihood $p_\theta(u_n,s_n,s_c)$ and $\mathcal{L}_{\text{switch}}$ is an additional penalty term regulazing the location of the transcriptional switch in the phase portrait. In more detail, let KL denote the Kullback-Leibler divergence, then the ELBO is defined as

$$\begin{aligned}\mathcal{L}_{\text{ELBO}}(\theta,\phi;u_n,s_n,s_c) = \sum_i -\mathbb{E}_{q_\phi(z_i,\pi_i|u_n^i,s_n^i,s_c^i)}[\log(p_\theta(u_n^i,s_n^i,s_c^i|z_i,\pi_i)] + \text{KL}[q_\phi(z_i|u_n^i,s_n^i,s_c^i) \parallel p(z)] \\ + \mathbb{E}_{q_\phi(z_i|u_n^i,s_n^i,s_c^i)}\left[\sum_g \text{KL}[q_\phi(\pi_{ig}|z_i) \parallel p(\pi_{ig}]\right],\end{aligned}$$

where the first term is often referred to as the reconstruction term and the remaining two terms as the regularization terms. The ELBO can be optimized using mini-batches of data, in particular, we will here use mini-batches of 256 cells for inference. For the penalty term $\mathcal{L}_{\text{switch}}$, we start by only considering cells that are above the 99-th percentile of unspliced abundance for each gene, thus these cells are approximately in induction steady state. Considering these cells we compute the median unspliced and spliced abundance in the respective part of the cell for each gene separately. Let $(u_n^*,s_n^*,s_c^*) \in \mathbb{R}^G$ be the outcome of this procedure, then

$$\mathcal{L}_{\text{switch}}(\theta;u_n,s_n,s_c) = \sum_g (u_{n0}^{g3}-u_n^{g*})^2 + (s_{n0}^{g3}-s_n^{g*})^2 + (s_{c0}^{g3}-s_c^{g*})^2,$$

where $(u_{n0}^{g3},s_{n0}^{g3},s_{c0}^{g3})$ are the initial conditions of the repression phase at the switching time t_g^s.

Rate parameter initialization

To initialize kinetics parameters, we will consider the calculated medians (u_n^*,s_n^*,s_c^*) of cells above the 99-th percentile of unspliced abundance for each gene. By definition, these cells are approximately in induction steady state. Thus, with equation (2.6) and (2.7) we obtain the abundances in the induction steady state for the particular cells. Similar to the *steady-state model*, we will assume a constant splicing rate $\beta = 1$ to initialize the transcription and nuclear export rate as

$$\begin{aligned}u_n^{*g} = \frac{\alpha_{g1}}{\beta_g} \Leftrightarrow \alpha_{g1} := u_n^{*g} \\ s_n^{*g} = \frac{\alpha_{g1}}{\nu_g} \Leftrightarrow s_n^{*g} = \frac{u_n^{*g}}{\nu_g} \Leftrightarrow \nu_g := \frac{u_n^{*g}}{s_n^{*g}}.\end{aligned}$$

Finally, using the definition of RNA velocity, e.g. in our case we will use the derivative of spliced mRNA in the cytoplasm $v = ds_c/dt$, and the fact that cells in the steady state have zero velocity, we initialize the degradation rate as

$$\frac{ds_c^{*g}}{dt} = \nu_g s_n^{*g} - \gamma_g s_c^{*g} = 0 \Leftrightarrow \gamma_g := \frac{u_n^{*g}}{s_c^{*g}}.$$

To guarantee numerical stability for the given initialization we further apply the inverse softplus function, defined as

$$f(x) = \log(e^x - 1),$$

to the given initialized rate parameters.

Optimization

For optimization of neural network parameters and kinetic rate parameters, we will use stochastic gradient descent along with the Adam optimizer with learning rate $\eta := 5e-3$ and weight decay as implemented in PyTorch. For the objective loss function, we will use $\lambda := 0.2$ as the scaling factor for the regularization term. The model's memory usage is constant throughout the training due to using a fixed batch size. The complete architecture of the extended veloVI model manifests as a variational autoencoder (VAE), consisting of decoder and encoder modules. All sub-modules of the VAE are fully connected feedforward networks that use standard activation functions like ReLU for hidden layers and softplus or exponential for parameterizing non-negative distributional parameters.

2.5 Downstream tasks

Fitted abundance values

The fitted values for unspliced and spliced abundance in the respective part of a cell are the posterior predictive means. Thus, for a cell $i \in \{1, ..., N\}$ the abundances can be calculated as

$$\begin{aligned}
\hat{u}_n^{(i)} &= \mathbb{E}_{p(u_n^{(i*)}|u_n^{(i)}, s_n^{(i)}, s_c^{(i)})}[u_n^{(i*)}] \\
\hat{s}_n^{(i)} &= \mathbb{E}_{p(s_n^{(i*)}|u_n^{(i)}, s_n^{(i)}, s_c^{(i)})}[s_n^{(i*)}] \\
\hat{s}_c^{(i)} &= \mathbb{E}_{p(s_c^{(i*)}|u_n^{(i)}, s_n^{(i)}, s_c^{(i)})}[s_c^{(i*)}],
\end{aligned}$$

with posterior predictive in the case of unspliced RNA defined as

$$p(u_n^{(i*)}|u_n^{(i)}, s_n^{(i)}, s_c^{(i)}) = \mathbb{E}_{q_\phi(z_i, \pi_i|u_n^{(i)}, s_n^{(i)}, s_c^{(i)})}[p_\theta(u_n^{(i*)}|z_i, \pi_i)],$$

which uses the variational posterior distribution as a plug-in estimator for the true (unknown) posterior distribution.
For later comparison, we oftentimes consider the mean-squared-error (MSE) between the model's fit and the true underlying abundances. Let $N \in \mathbb{N}$ be the number of cells, then for each gene g the MSE of the abundances can be calculated as

$$\mathrm{MSE}_{u_n}(\overline{u}_n^{(g)}, u_n^{(g)}) = \frac{1}{N}\sum_{i=1}^{N}(\overline{u}_n^{(g)} - u_n^{(g)})^2 \tag{2.12}$$

$$\mathrm{MSE}_{s_n}(\overline{s}_n^{(g)}, s_n^{(g)}) = \frac{1}{N}\sum_{i=1}^{N}(\overline{s}_n^{(g)} - s_n^{(g)})^2 \tag{2.13}$$

$$\mathrm{MSE}_{s_c}(\overline{s}_c^{(g)}, s_c^{(g)}) = \frac{1}{N}\sum_{i=1}^{N}(\overline{s}_c^{(g)} - s_c^{(g)})^2, \tag{2.14}$$

where $(\overline{u}_n^{(g)}, \overline{s}_n^{(g)}, \overline{s}_c^{(g)})$ denote the estimated abundances, *i.e.* the posterior predictive means, and $(u_n^{(g)}, s_n^{(g)}, s_c^{(g)})$ denote the actual unspliced and spliced abundances.

State assignment

The state assignment for each gene and cell is the approximate posterior mean

$$\mathbb{E}_{q_\phi(z_i|u_n^{(i)},s_n^{(i)},s_c^{(i)})}[\mathbb{E}_{q_\phi(\pi_{ig}|z_i)}[\pi_{ig}]]$$

Gene-wise latent time assignment

The latent time is computed for each gene and cell as

$$\mathbb{E}_{q_\phi(z_i|u_n^{(i)},s_n^{(i)},s_c^{(i)})}[\mathbb{E}_{q_\phi(\pi_{ig}|z_i)}[t_{ig}^{k_{ig}}]],$$

where the outer expectation with respect to $q_\phi(z_i|u_n^{(i)}, s_n^{(i)}, s_c^{(i)})$ is estimated with Monte Carlo samples, while the inner expectation is computed analytically over the transcriptional states k_{ig}.

Computation of RNA velocity

RNA velocity of a gene g in a particular cell i is similarly a function of the variational posterior. Recall that the velocity is computed as the time derivative of spliced RNA abundance, here we will define it as the time derivative of spliced RNA in the cytosol as

$$v_{s_c}^{(g)}(t^{(k)},k) := \left.\frac{ds_c^{(g)}(t,k)}{dt}\right|_{t^{(k)}} = \nu_g s_n^{(g)}(t^{(k)},k) - \gamma_g s_c^{(g)}(t^{(k)},k). \tag{2.15}$$

Thus, we can compute samples of a posterior predictive velocity distribution via the following process

1. Sample z_i from $q_\phi(z_i|u_n^{(i)}, s_n^{(i)}, s_c^{(i)})$
2. Compute $\mathbb{E}_{q_\phi(\pi_{ig}|z_i)}[v^{(g)}(t_{ig}^{(k_{ig})}, k_{ig})]$ for each gene.

The outcome of this procedure will provide samples from a distribution over the velocity for every gene, and cell pair, which we then use in downstream tasks.
Similarly for a gene g in a particular cell i, we can define unspliced and spliced velocity in the nucleus as its time derivatives

$$v_{s_n}^{(g)}(t^{(k)},k) := \left.\frac{ds_n^{(g)}(t,k)}{dt}\right|_{t^{(k)}} = \beta_g u_n^{(g)}(t^{(k)},k) - \nu_g s_n^{(g)}(t^{(k)},k) \tag{2.16}$$

$$v_{u_n}^{(g)}(t^{(k)},k) := \left.\frac{du_n^{(g)}(t,k)}{dt}\right|_{t^{(k)}} = \alpha_{gk} - \beta_g u_n^{(g)}(t^{(k)},k). \tag{2.17}$$

Lastly, we define the velocity $v_s^{(g)}$ as the sum of both spliced time derivatives. For a gene g the velocity can be calculated as

$$\begin{aligned} v_s^{(g)}(t^{(k)},k) := v_{s_n}^{(g)}(t^{(k)},k) + v_{s_c}^{(g)}(t^{(k)},k) &= \left.\frac{d(s_n^{(g)} + s_c^{(g)})(t,k)}{dt}\right|_{t^{(k)}} \\ &= \beta_g u_n^{(g)}(t^{(k)},k) - \nu_g s_n^{(g)}(t^{(k)},k) + \nu_g s_n^{(g)}(t^{(k)},k) - \gamma_g s_c^{(g)}(t^{(k)},k) \\ &= \beta_g u_n^{(g)}(t^{(k)},k) - \gamma_g s_c^{(g)}(t^{(k)},k). \end{aligned} \tag{2.18}$$

All different velocities can be calculated by following the same procedure as for spliced-cytoplasmic RNA velocity. We refer to these three different velocity computations based on different time derivatives as *velocity modes*.

Intrinsic uncertainty

Let $\bar{v}_i$ be the posterior predictive velocity mean from the procedure above. Let $c(\cdot,\cdot)$ denote the cosine similarity, then the intrinsic uncertainty is defined as

$$\mathbb{V}ar_{q_\phi(v_i|u_n^{(i)},s_n^{(i)},s_c^{(i)})}[c(\bar{v}_i, v_i)].$$

Let $\{v_i^{(l)}\}_{l=1}^L$ be the set of $L \in \mathbb{N}$ velocity vector samples of cell i in from the variational posterior. Then the intrinsic uncertainty can be calculated as

$$\hat{\sigma}_i^2 = \frac{1}{L-1}\sum_{l=1}^{L}\left(\frac{v_i^{(l)}\bar{v}_i}{\|v_i^{(l)}\|\|\bar{v}_i\|} - \frac{1}{L}\sum_{j=1}^{L}\frac{v_i^{(j)}\bar{v}_i}{\|v_i^{(j)}\|\|\bar{v}_i\|}\right).$$

Within this work, we used $L = 50$ due to computational limitations.

Extrinsic uncertainty

Let $T(v_{s_c}^{1:N}, s_c^{1:N})$ be a function mapping the velocity vectors and the spliced abundances in the cytoplasm of the full dataset, e.g. of all N cells, to a cell-cell transition matrix computed as described in ref. [10]. The function compares the similarity of the displacement δ_{ij} of nearest neighbors $s_c^{(i)}$ and $s_c^{(j)}$ to the velocity v_i of cell i via the cosine similarity

$$\cos(\delta_{ij}, v_i) = \frac{\delta_{ij}^T v_i}{\|\delta_{ij}^T\|\|v_i\|}$$

as the basis for computing transition probabilities between pairs of cells. The resulting transition matrix for one sample of velocity is used to compute the matrix multiplication $T(v^{1:N}, s_c^{1:N})S_c$, where $S_c \in \mathbb{R}^{N\times G}$ denotes the matrix of spliced cytosolic RNA abundance of all cell-gene pairs. The predicted future cell state vectors are then used to compute the variance of the cosine similarity, similar to the intrinsic uncertainty. Notably, to compute extrinsic uncertainty based on the velocity v_{s_n}, we let $T(v_{s_n}^{1:N}, s_n^{1:N})$ be a function of nucleic spliced abundance. Finally, we multiply the output with $S_n \in \mathbb{R}^{N\times G}$ denoting the matrix of spliced nucleic RNA abundance of all cell-gene pairs. Similarly, for v_s we let $T(v_{s_n}^{1:N}, s_n^{1:N} + s_c^{1:N})$ be a function of the sum of spliced abundance which is then multiplied by $S_n + S_c \in \mathbb{R}^{N\times G}$ denoting the matrix containing the sum of nucleic and cytosolic spliced abundance.

Velocity confidence

Velocity confidence is a measure to validate the coherence of the velocity vector field. It compares the velocity of a reference cell i with the velocity of its k-nearest neighbors within a neighborhood graph. The velocity confidence for a cell i can be formulated as

$$c_i = \frac{1}{k}\sum_{j=1}^{k}\mathrm{corr}(v_i, v_j),$$

where the index j refers to neighboring cells. Within this work, we have used the function `velocity_confidence` implemented in the *scVelo* package.

Permutation score

To assess the robustness of the inferred dynamics with respect to random permutations in the input data, we introduce a gene- and cell-type specific permutation score [14]. To this end, for each gene and cell type, the unspliced and spliced abundances are independently permuted. Specifically, we permute the

indices of unspliced and spliced abundances of the observations within each cell type using permutations $(\pi_{u_n}, \pi_{s_n}, \pi_{s_c})$. After applying this procedure we receive permuted vectors per gene

$$
\begin{aligned}
u_n^{(p)} &= \left((u_n)_{(\pi_{u_n}^{-1}(1))}, (u_n)_{(\pi_{u_n}^{-1}(2))}, ..., (u_n)_{(\pi_{u_n}^{-1}(N))} \right) \\
s_n^{(p)} &= \left((s_n)_{(\pi_{s_n}^{-1}(1))}, (s_n)_{(\pi_{s_n}^{-1}(2))}, ..., (s_n)_{(\pi_{s_n}^{-1}(N))} \right) \\
s_c^{(p)} &= \left((s_c)_{(\pi_{s_c}^{-1}(1))}, (s_c)_{(\pi_{s_c}^{-1}(2))}, ..., (s_c)_{(\pi_{s_c}^{-1}(N))} \right).
\end{aligned}
$$

Following, the absolute errors between the permuted data and the model's fit of the permuted data, *i.e.* the posterior predictive means, denoted as $(\hat{u}_n^{(p)}, \hat{s}_n^{(p)}, \hat{s}_c^{(p)})$, is calculated as

$$
\mathcal{L}_{\text{AE}}^{(p)} = |u_n^{(p)} - \hat{u}_n^{(p)}| + |s_n^{(p)} - \hat{s}_n^{(p)}| + |s_c^{(p)} - \hat{s}_c^{(p)}|.
$$

We use the same calculation for none-permuted data. The mean absolute error between permuted observations and inferred dynamics is denoted as $\mu^{(p)}$, while the mean absolute error for unpermuted data is denoted as μ^0. To test whether the mean absolute errors of the permuted and unpermuted samples are equal, we define the permutation score as the independent, cell-type specific t-test statistic

$$
T = \frac{\mu^{(p)} - \mu^0}{\sqrt{2\frac{S^2}{n}}},
$$

where n denotes the cell-type specific number of cells and S^2 denotes the pooled variance of the absolute errors. To reduce the effect of dataset size we use a maximum sample size of $n = 200$. Note that permutation scores are aggregated on a gene level and if the statistics are equal, the permutation score will be close to 0. We assume that this scenario occurs for complete steady-state populations.

3 Methods and data pre-processing

3.1 Data simulation

Simulating data based on the specified *Nucleus-cytosol model* requires the following parameters:

- Number of genes, $G \in \mathbb{N}$. In the following, we used $G = 1000$.
- Number of cells, $N \in \mathbb{N}$. In the following, we used $N = 25000$.
- Minimum and maximum number for cell-gene times defined as $t_{\text{min}} = 0$ and $t_{\text{max}} = 20$. Thus, the starting time of the induction phase is $t_0^{g1} = 0$ for all genes.
- Cell times $t \in \mathbb{R}^N$ either sampled from a Poisson distribution or a Uniform distribution. Cell times are then scaled s.t. $t_i \in [t_{\text{min}}, t_{\text{max}}] = [0, 20]$. Here, we used uniformly distributed cell times to guarantee "complete" phase portraits containing the characteristic almond shape.
- Switching times t_g^s uniformly distributed in the interval $[0.1 * t_{\text{max}}, 0.5 * t_{\text{max}}]$.
- Rate parameters $(\alpha_g, \beta_g, \nu_g, \gamma_g) \in \mathbb{R}^{4\times G}$ sampled from a multivariate log-normal distribution. The mean $\mu \in \mathbb{R}^4$ and covariance matrix $\Sigma \in \mathbb{R}^{4\times 4}$ for the normal distribution were exemplary defined as

$$\mu = (2, 1, 0.5, 0.1)$$
$$\Sigma = \begin{pmatrix} 0.16 & 0.128 & 0.08 & 0.032 \\ 0.128 & 0.16 & 0.08 & 0.032 \\ 0.08 & 0.08 & 0.16 & 0.08 \\ 0.032 & 0.032 & 0.08 & 0.16 \end{pmatrix}.$$

  The transcription rate α_{gk} is set to 0 for cell-gene pairs in repression phase or repression steady state ($k \in \{3, 4\}$), *i.e.*, where $t \geq t_g^s$. Eliminating large outliers among the sampled rate parameters involves the exclusion of genes with values exceeding the 99-th percentile of each kinetic parameter.
- Initial values for induction and repression phase. For the induction phase, the initial values are set to 0. The initial values for the repression phase are defined as the unspliced/ spliced abundances evaluated at the switching time t_g^s, similar as defined in section 2.2

$$\begin{aligned} u_{n0}^{g1} &= s_{n0}^{g1} = s_{c0}^{g1} = 0 \\ u_{n0}^{g3} &= \overline{u}_n^{(g)}(t_g^s, k = 1) \\ s_{n0}^{g3} &= \overline{s}_n^{(g)}(t_g^s, k = 1) \\ s_{c0}^{g3} &= \overline{s}_c^{(g)}(t_g^s, k = 1). \end{aligned}$$

- Noise level σ to sample noise from a normal distribution with $\mu = 0$. Without modeling noise, the simulated data would otherwise perfectly follow the underlying model. Assume that for some gene g we have already sampled unspliced and spliced abundances $(u_n^g, s_n^g, s_c^g) \in \mathbb{R}^{N\times 3}$ following the parameters and procedure defined beforehand. We then define the 99-th percentile $p_{99} \in \mathbb{R}^3$ of (u_n^g, s_n^g, s_c^g). Finally, noise $\epsilon \in \mathbb{R}^{N\times 3}$ is sampled from $\mathcal{N}(0, \frac{\sigma}{10} p_{99})$ and added to the unspliced and spliced abundances (u_n^g, s_n^g, s_c^g). Unless specified otherwise, we used $\sigma = 0.8$.

3.2 Integration of snRNA-seq and scRNA-seq data

Current RNA velocity models consider splicing kinetics at the level of the entire cell. However, these dynamics have a sub-cellular component: unspliced pre-mRNA is spliced in the nucleus and RNA molecules are exported from the nucleus. Consequently, for a more accurate description of the biological process, we introduced the *Nucleus-cytosol model*, as seen in Chapter 2, differentiating between nucleic and cytosolic RNA.
However, single-cell protocols measure RNA abundance in the whole cell, spanning the nucleus and cytoplasm, while single-nucleus protocols only measure nucleic RNA. Hence, for single-cell protocols, we lack information regarding the proportions of nucleic and cytosolic RNA, whereas single-nucleus protocols do not capture cytosolic RNA. However, inferring the rates of the *Nucleus-cytosol model* requires RNA measurements in both parts of the cell. Therefore, the missing abundance, *i.e.*, nucleic RNA in the case of single-cell, and cellular RNA for single-nucleus observations need to be estimated. The estimate of cellular abundance for single-nucleus cells automatically yields an estimate of cytosolic abundance. Similarly, cytosolic abundance for single-cell observations is automatically given by their nucleic RNA estimates.
Estimation is achieved by integrating single-cell and single-nucleus observations into one model. Figure 3.1 illustrates the Uniform Manifold Approximation and Projection [21] (UMAP) projection of unintegrated, *i.e.*, none batch corrected single-cell and nucleus observations of the pancreas E14.5 datasets [22][23]. This plot clearly highlights the necessity of a batch-corrected shared latent space. Therefore, as a first step a joint latent space is constructed, such that observations from one protocol have a sufficient number of neighboring cells from the other protocol.
Once we embed cells from both measurements into this batch-corrected latent space, we compute a k-nearest neighbor graph. This graph then indicates the degree of connectivity between cells, which we can use to calculate the estimates of nucleic RNA for single-cell observations and to calculate the estimates of cellular RNA for single-nucleus observations.

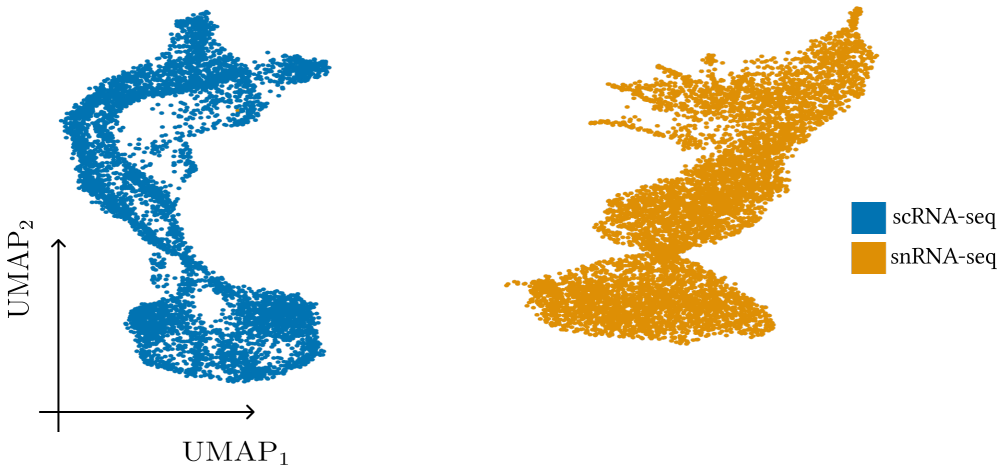

Figure 3.1 Unintegrated UMAP embedding of common PCA space for single-cell and nucleus measurements of pancreas E14.5 datasets [22][23].

3.2.1 Construction of joint latent space

To construct a joint latent space we used models provided within the Python modules *scglue* ("single-cell graph linked unified embedding") [24] and *scVI* ("single-cell variational inference") [25]. Both models have shown to work well for the task of multi-omics data integration, which are shortly described in Appendix A.2.3 and A.2.4. For *scglue*, we trained one model using gene expression counts of both modalities and another model using chromatin accessibility profiles in combination with gene expression counts of the scRNA-seq data. Contrastingly, *scVI* is solely trained on gene expression data of both modalities. For both modules, we relied on the Tutorial sections by following the same pre-processing and training steps as

suggested [26][27].
To evaluate the performance of the outlined multi-omics data integration methods, we benchmarked method performance by comparing the *scIB* metrics for data integration [28] briefly summarized in Appendix A.2.5. Additionally, we relied on the integration consistency score, a metric proposed by Cao et al. [24]. This metric measures the consistency between the integrated multi-omics space and the prior knowledge represented in the guidance graph, which should remain above the threshold of 0.05 to result in a reliable integration (Appendix A.2.4). The consistency scores for the pancreas E14.5 and E15.5 datasets [22][23] are reported in Figure A.16. From here we can conclude that the integration is reliable.
Based on the scIB metrics as shown in Figure 3.3, we can conclude that none of the models performed significantly better. However, for most metrics, the *scglue* model trained on gene expression counts of both modalities consistently reports the highest values. Following, the aggregated batch correction and bio conservation scores are the highest for this model's latent embedding.
Notably, the *scVI* model was fitted faster compared to *scglue* models. Within subsequent downstream tasks, we will further compare *scglue* and *scVI* based models as they result in different latent embeddings and neighbor graphs on which we rely for many pre-processing steps. However, as the *scglue* model trained on gene expression counts resulted in higher benchmarking scores, we decided to exclude the model trained on chromatin accessibility profiles.

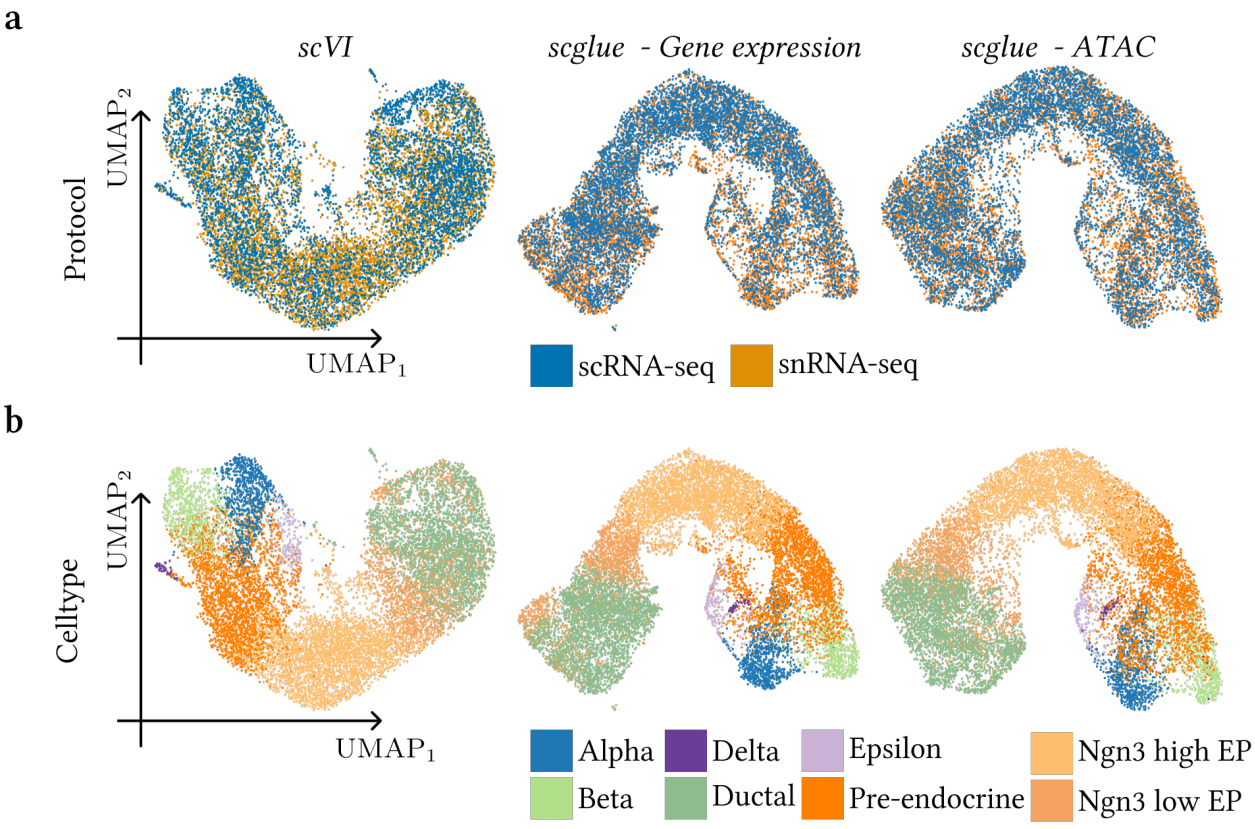

Figure 3.2 UMAP embeddings of pancreas E14.5 datasets [22][23]. **a.** Integrated UMAP embedding computed on batch-corrected latent spaces for the three different models colored by protocol. **b.** Same as **a**, but colored by cell type.

3.2.2 Estimation of nuclear and cytosolic mRNA abundance

Given the embedded cell representations of the batch corrected latent space, we can calculate a k-nearest neighbor graph $G \in \mathbb{R}^{N \times N}$ via the pre-processing function `neighbors` as implemented in the *scVelo* framework. Here, we used the default value $k = 30$. Let the cell-cell neighbor graph be given, where

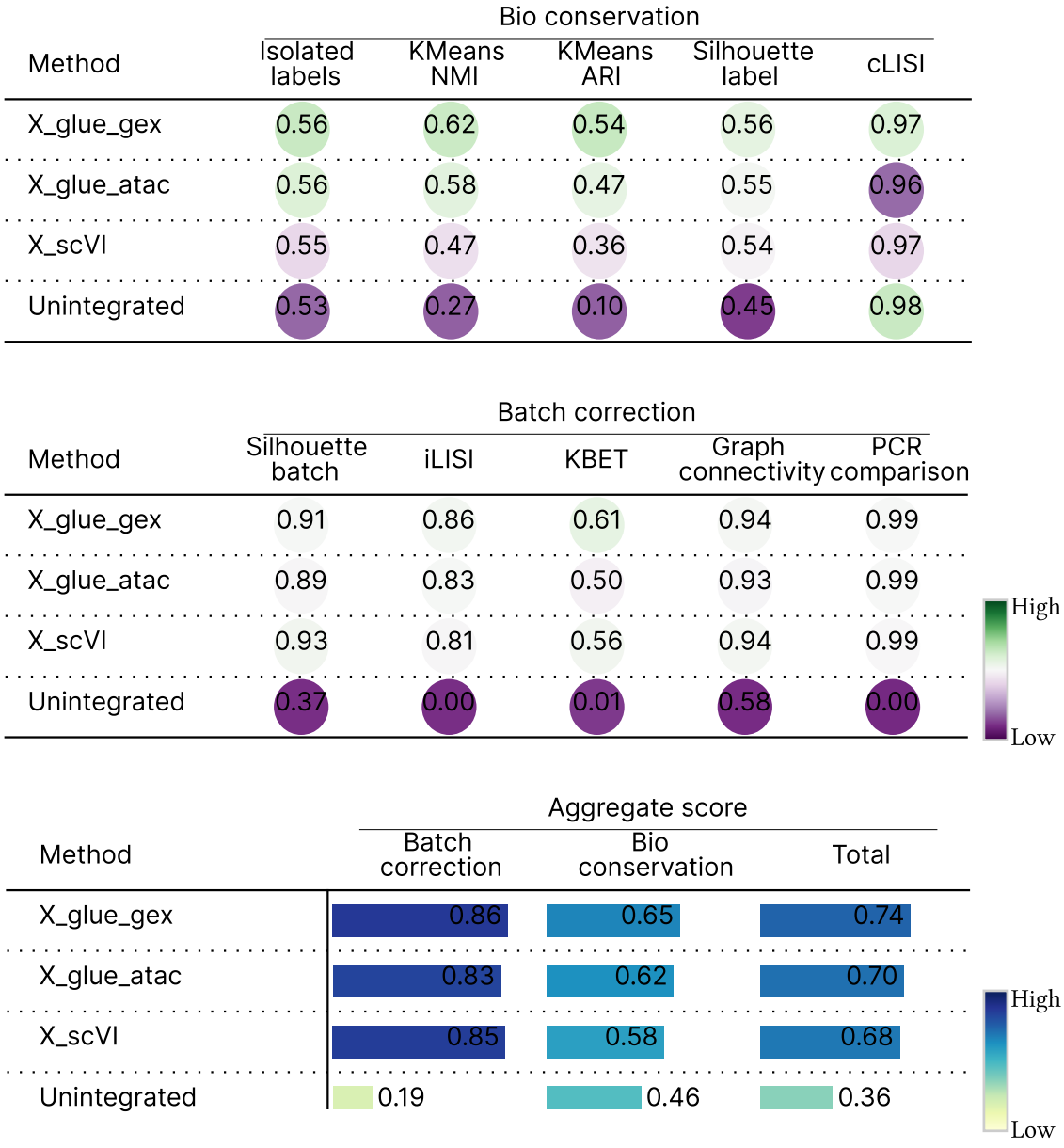

Figure 3.3 scIB metrics for integration of the pancreas E14.5 datasets [22][23]: *X_glue_gex* refers to the latent embeddings generated by the *scglue* model trained on gene expression counts of both modalities; *X_glue_atac* refers to the latent embeddings generated by the *scglue* model trained on single-cell gene expression transcripts and chromatin accessibility profiles; *X_scVI* refers to the latent embeddings generated by the *scVI* model; *Unintegrated* refers to the none batch corrected PCA embeddings.

$g_{ij} \in [0, 1]$ corresponds to the strength of the connectivity between cell i and cell j. For simplicity, we denote the set of observations from the single-cell protocol as A and cells from the single-nucleus protocol as B. Further, let $(u^{(ig)}, s^{(ig)})$ and $(u_n^{(ig)}, s_n^{(ig)})$ be the measured unspliced and spliced abundances for a gene g and cell i for scRNA-seq and snRNA-seq protocols, respectively. Estimated variables are characterized by a circumflex.
Let us first consider a cell $i \in A$ stemming from a scRNA-seq protocol. As an initial step, we define its neighborhood as

$$\mathcal{N}(i) := \{j \in A \cup B | g_{ij} > 0\}.$$

Then, to estimate nucleic RNA, we restrict the neighborhood of cell i to neighbor cells from snRNA-seq measurements as

$$\mathcal{N}_{\text{snRNA}}(i) := \{j \in B | g_{ij} > 0\}.$$

Finally, an estimation of unspliced and spliced abundance in the nucleus for a gene g of cell i can be calculated as a weighted average of RNA abundance of its neighbor cells

$$\hat{u}_n^{(ig)} = \frac{1}{\sum_{j \in \mathcal{N}_{\text{snRNA}}(i)} g_{ij}} \sum_{j \in \mathcal{N}_{\text{snRNA}}(i)} g_{ij} u_n^{(jg)} \tag{3.1}$$

$$\hat{s}_n^{(ig)} = \frac{1}{\sum_{j \in \mathcal{N}_{\text{snRNA}}(i)} g_{ij}} \sum_{j \in \mathcal{N}_{\text{snRNA}}(i)} g_{ij} s_n^{(jg)}. \tag{3.2}$$

Eventually, spliced abundance in the cytosol can be calculated as the difference between measured cellular counts and estimated nucleic counts

$$\hat{s}_c^{(ig)} = s^{(ig)} - \hat{s}_n^{(ig)}. \tag{3.3}$$

If $s^{(ig)} < \hat{s}_n^{(ig)}$, we then clip the cytosolic abundance as $\hat{s}_c^{(ig)} = 0$.
Similarly, for a cell $i \in B$ from snRNA measurements we can define its neighborhood restricted on cells measured from scRNA protocols as

$$\mathcal{N}_{\text{scRNA}}(i) := \{j \in A | g_{ij} > 0\}.$$

Here, we are interested in estimating spliced abundance in the cytosol. To do so, we first calculate an estimate for the abundance of the whole cell spanning nucleus and cytoplasm, which is then given by

$$\hat{u}^{(ig)} = \frac{1}{\sum_{j \in \mathcal{N}_{\text{scRNA}}(i)} g_{ij}} \sum_{j \in \mathcal{N}_{\text{scRNA}}(i)} g_{ij} u^{(jg)} \tag{3.4}$$

$$\hat{s}^{(ig)} = \frac{1}{\sum_{j \in \mathcal{N}_{\text{scRNA}}(i)} g_{ij}} \sum_{j \in \mathcal{N}_{\text{scRNA}}(i)} g_{ij} s^{(jg)}. \tag{3.5}$$

Finally, spliced abundance in the cytosol can be estimated as the difference between the estimate for the abundance of the whole cell and the known nucleic abundance

$$\hat{s}_c^{(ig)} = \hat{s}^{(ig)} - s_n^{(ig)}. \tag{3.6}$$

Similar as before, if $\hat{s}^{(ig)} < s_n^{(ig)}$, we set $\hat{s}_c^{(ig)} = 0$.

3.2.3 Lambda correction

A central challenge in analyzing sequencing protocols is presented by batch effects. They arise when examining cells derived from different batches, *i.e.*, when handling cells in distinct groups. They can also occur when the variation in sample groups is caused by technical arrangement, *e.g.* when labs use different

techniques to dissociate samples or when using different sequencing depths, which in turn can lead to false conclusions [1]. When analyzing the pancreas E15.5 datasets [22][23], we observed that snRNA-seq cells in general have higher gene expression counts compared to scRNA-seq cells, even though scRNA-protocols measure RNA abundance of the whole cell, while snRNA-protocols only measures nucleic RNA. Figure 3.4 displays unspliced and spliced count densities and distributions for both experimental protocols for the pancreas datasets E14.5 and E15.5 [22][23]. For E15.5 the distributions indicate that single-nucleus observations in general exhibit elevated unspliced counts, whereas spliced counts approximately follow the same distribution for both protocols, despite the missing cytosolic spliced abundance for the single-nucleus observations. Hence, it appears that the snRNA-seq pipeline demonstrates greater sensitivity in capturing spliced and unspliced mRNA molecules, resulting in higher overall counts. We refer to this observation as a batch effect. For the pancreas E14.5 datasets [22][23] we observed expected distributions, *i.e.*, spliced abundances are elevated for single-cell observations, while unspliced abundances approximately follow the same distribution as shown in Figure 3.4. However, it is noteworthy that the two-sided Welch's test yielded $P < 0.001$ for all comparisons between the two modalities.

To account for and mitigate this batch effect, we integrated a so-called *lambda correction*. Here, we want to ensure that unspliced abundances of single-cell and single-nucleus observations follow the same distribution as unspliced molecules are primarily found within the nucleus. Therefore, single-cell and nucleus observations should approximately follow the same unspliced count distribution. Omitting the lambda correction led to the estimation of substantial quantities of spliced cytosolic abundance as negative values, which can be attributed to variations in count distributions.

As a first step, unspliced abundance is estimated for both measurements as shown in equations (3.1) and (3.4), resulting in cell-gene matrices $U, U_n \in \mathbb{R}^{N\times G}$, where $U_n = (u_n, \hat{u}_n)^T$ denotes measured and estimated unspliced abundance based on snRNA-seq data and $U = (\hat{u}, u)^T$ the estimated and measured unspliced abundance based on scRNA-seq data, with $u_n, \hat{u} \in \mathbb{R}^{|B|\times G}$; $u, \hat{u}_n \in \mathbb{R}^{|A|\times G}$ and $N = |A \cup B|$. Thus, the matrix U consists of unspliced abundances solely based on scRNA-seq abundances, while U_n only takes snRNA-seq measurements into account. We then define the matrix $\Lambda \in \mathbb{R}^{N\times G}$ as

$$\Lambda_{ij} = \begin{cases} \frac{U_{ij}}{(U_n)_{ij}} & \text{if } U_{ij}, (U_n)_{ij} \neq 0 \\ 0 & \text{else.} \end{cases}$$

As a final step, we similarly define the cell-gene matrix of nucleic spliced counts based on snRNA-seq data as $S_n = (s_n, \hat{s}_n)^T$, where $s_n \in \mathbb{R}^{|B|\times G}$; $\hat{s}_n \in \mathbb{R}^{|A|\times G}$. Finally, nucleic abundances are scaled as

$$(S_n)_{ij} = \Lambda_{ij}(S_n)_{ij}$$
$$(U_n)_{ij} = \Lambda_{ij}(U_n)_{ij}$$

before estimating spliced abundance in the cytosol as outlined in equation (3.3) and (3.6). As a last data quality check we consider the summed entries of Λ over all cells for each gene g

$$\lambda_g = \sum_{i=1}^{N} \Lambda_{ig}.$$

Finally, we drop genes, where $\lambda_g = 0$, as otherwise unspliced and spliced abundances in the nucleus are set to zero for all cells. Summarizing, the *lambda correction* calculates a ratio between the unspliced counts of single-cell and single-nucleus measurements to then scale the nucleic abundance by this ratio.

3.3 Data pre-processing

In the following, the single-modal pancreas datasets [22][23] were pre-processed by following the same steps. First, we needed to integrate single-cell and nucleus data as outlined in Section 3.2.1 to receive batch-corrected latent embeddings for each observation. Then, we concatenated single-cell and nucleus

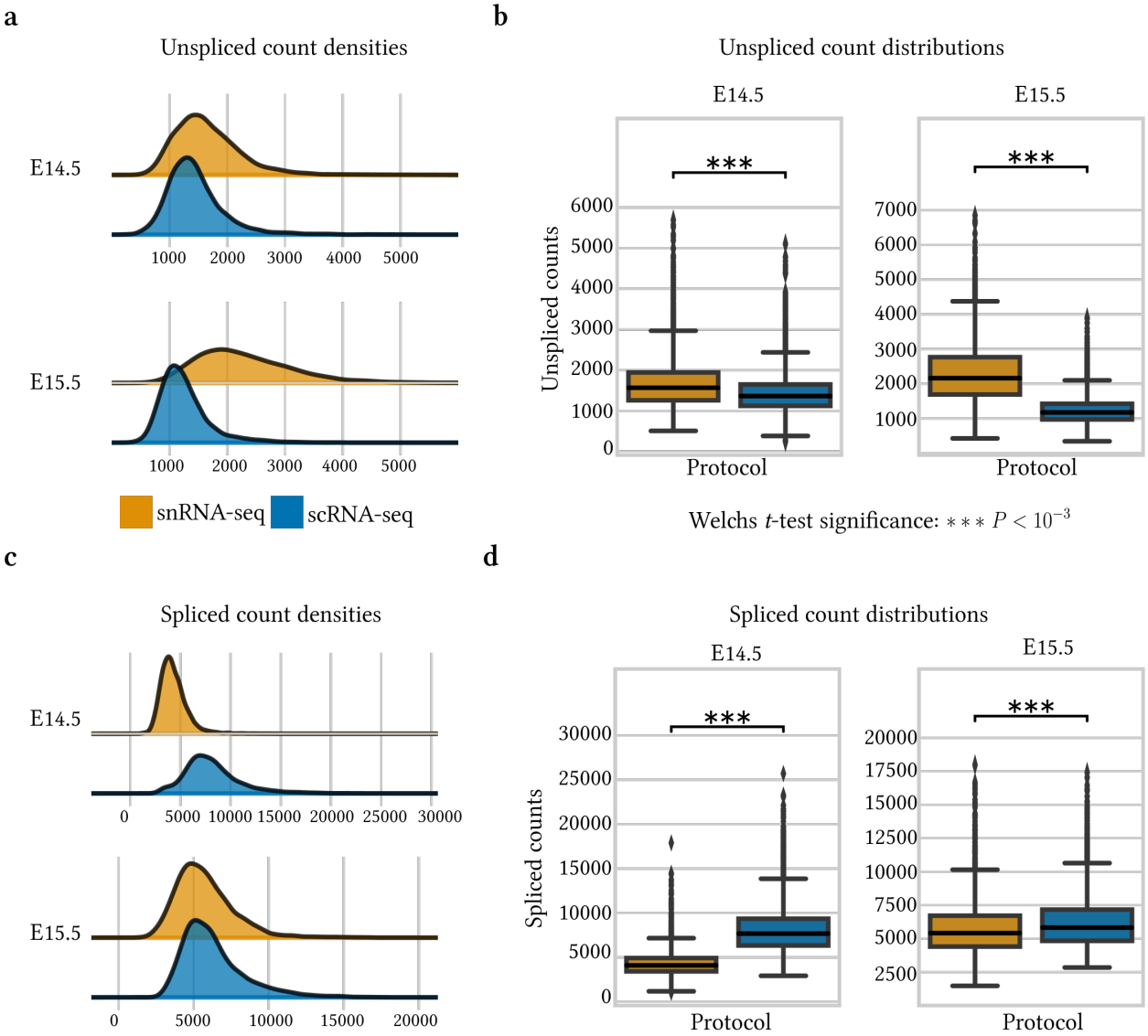

Figure 3.4 Unspliced and spliced counts visualizations of pancreas datasets [22][23]. Here, gene expression counts of all genes present for both modalities are summed up for each cell. **a.** Unspliced count densities of cells per protocol for E14.5 and E15.5 [22][23]. **b.** Unspliced count distribution of cells per protocol for E14.5 and E15.5 [22][23]. Box plots indicate the median (center line), interquartile range (hinges), and whiskers at 1.5× interquartile range. **c.** Spliced count densities of cells per protocol for E14.5 and E15.5 [22][23]. **d.** Spliced count distribution of cells per protocol for E14.5 and E15.5 [22][23]. Box plots indicate the median (center line), interquartile range (hinges), and whiskers at 1.5× interquartile range.

data containing all cells and genes. Given the batch corrected latent representations we calculated a k-nearest neighbor graph with the default value of $k = 30$ as implemented in *scanpy*'s [29] pre-processing function `neighbors`. Next, we estimated the missing abundances as described in Section 3.2.2 which adds the layers *unspliced_nucleus*, *spliced_nucleus*, *spliced_cytoplasm* to the concatenated `adata` object. Afterward, genes with fewer than 20 spliced counts are removed before extracting the 2000 most highly variable genes based on dispersion. This is achieved by first executing *scvelo*'s `filter_genes` function with `min_counts=20`. Subsequently, *scanpy*'s [29] pre-processing function `highly_variable_genes` is executed with `flavor=''seurat_v3''` (count data), `n_top_genes=2000` and `batch_key` specified as the variable containing the cell's batch information. Then, transcriptomic counts of each cell are normalized by their median by running *scvelo*'s pre-processing `normalize_per_cell` function with `layers` defined as the layers containing unspliced and spliced abundances in the respective part of a cell. Using the neighbor graph which was constructed based on the common latent space representation, counts are smoothed by the mean expression across neighbor cells to receive the final RNA abundances. These calculations are performed by following the same steps as within *scvelo*'s `moments` pre-processing function. However, the function needed to be adapted according to our new three-layer structure. Finally, the pre-processed unspliced and spliced moments in the respective part of a cell are independently scaled to the unit interval per gene. Scaled moments are then used to train our *Nucleus-cytosol model*.

4 Results

4.1 Evaluation on simulated data

4.1.1 Inference of kinetic parameters and latent times

As a first step, we evaluated whether the *Nucleus-cytosol model* is capable of learning the parameters underlying the modeled splicing and nuclear-exporting dynamics on simulated data. The data has been generated as described in Section 3.1. Since the model is none-identifiable as shown in Lemma A.3.5, in the following, we will consider rate parameter ratios, such that the scaling factor $\lambda_g \in \mathbb{R}$ is canceled out for each gene g. To assess the quality of the model's fit we therefore considered correlations between inferred and true kinetic parameter ratios and correlations between inferred and true sampled latent times as well as switching times.
The kinetic parameter ratio plots are shown in Figure 4.1a. Here, we calculated the Pearson correlations between the true $\frac{\alpha}{\beta}, \frac{\alpha}{\nu}, \frac{\alpha}{\gamma}, \frac{\nu}{\beta}, \frac{\gamma}{\beta}, \frac{\gamma}{\nu} \in \mathbb{R}^G$ and the respective inferred ratio pairs $\frac{\hat{\alpha}}{\hat{\beta}}, \frac{\hat{\alpha}}{\hat{\nu}}, \frac{\hat{\alpha}}{\hat{\gamma}}, \frac{\hat{\nu}}{\hat{\beta}}, \frac{\hat{\gamma}}{\hat{\beta}}, \frac{\hat{\gamma}}{\hat{\nu}} \in \mathbb{R}^G$. The correlations are reported in Figure 4.1b, which clearly highlights that the model is able to infer a scaled version of the true underlying rate parameters for most genes.
To evaluate time assignments, let $T \in \mathbb{R}^{N \times G}$ be the inferred latent times and $t \in \mathbb{R}^N$ be the observed underlying cell times. First, we scaled the inferred cell-gene times for each gene g as

$$T[:, g] = T[:, g] \frac{t_{\max}}{\max_g\{T[:, g]\}},$$

where $t_{\max} = 20$. The scaling is done to ensure that $T \in [t_{\min}, t_{\max}] = [0, 20]$. Again we need to cancel the scaling factor $\lambda_g \in \mathbb{R}$ for each gene g. Therefore, for each gene g we calculated the Spearman correlations between the scaled inferred cell-gene times and the scaled observed cell times as $\mathrm{corr}(\hat{\alpha}_g T[:, g], \alpha t)$. The resulting correlations for all genes are reported in Figure 4.1c. From here we conclude that the inference of the *Nucleus-cytosol model* results in reliable time assignments with generally high correlations to the true underlying cell times.
Lastly, we compared inferred against true switching times as plotted in Figure 4.1d. As for the kinetic parameters and latent times, we need to cancel the scaling factor and therefore report the switching times scaled by the transcription rate. Notably, the estimation of switching times resulted in lower correlations in comparison to the correlations achieved for kinetic parameter ratios and latent times.
Overall, the extended model accurately fits the underlying simulated data by correctly inferring rate parameters, cell-gene latent times, and switching times.

4.1.2 Investigation of outlier genes

The ratio plots as shown in Figure 4.1a suggest that the model very accurately fits the underlying dynamics. However, there are genes where the inference is not working as well as for others. To identify these genes, let $\hat{\alpha}, \hat{\beta}, \hat{\nu}, \hat{\gamma} \in \mathbb{R}^G$ be the estimated rate parameters and $\alpha, \beta, \nu, \gamma \in \mathbb{R}^G$ the true underlying rate parameters. Then, by considering the element-wise absolute errors between the ratios, *i.e.*, the absolute errors per gene, we define the ratio deviation vectors $r_i \in \mathbb{R}^G$ for $i \in \{1, ..., 6\}$ as

$$\begin{aligned} r_1^{(g)} &= \left|\frac{\hat{\alpha}^{(g)}}{\hat{\beta}^{(g)}} - \frac{\alpha^{(g)}}{\beta^{(g)}}\right|, \quad r_2^{(g)} = \left|\frac{\hat{\alpha}^{(g)}}{\hat{\nu}^{(g)}} - \frac{\alpha^{(g)}}{\nu^{(g)}}\right|, \quad r_3^{(g)} = \left|\frac{\hat{\alpha}^{(g)}}{\hat{\gamma}^{(g)}} - \frac{\alpha^{(g)}}{\gamma^{(g)}}\right| \\ r_4^{(g)} &= \left|\frac{\hat{\nu}^{(g)}}{\hat{\beta}^{(g)}} - \frac{\nu^{(g)}}{\beta^{(g)}}\right|, \quad r_5^{(g)} = \left|\frac{\hat{\gamma}^{(g)}}{\hat{\beta}^{(g)}} - \frac{\gamma^{(g)}}{\beta^{(g)}}\right|, \quad r_6^{(g)} = \left|\frac{\hat{\gamma}^{(g)}}{\hat{\nu}^{(g)}} - \frac{\gamma^{(g)}}{\nu^{(g)}}\right|. \end{aligned} \tag{4.1}$$

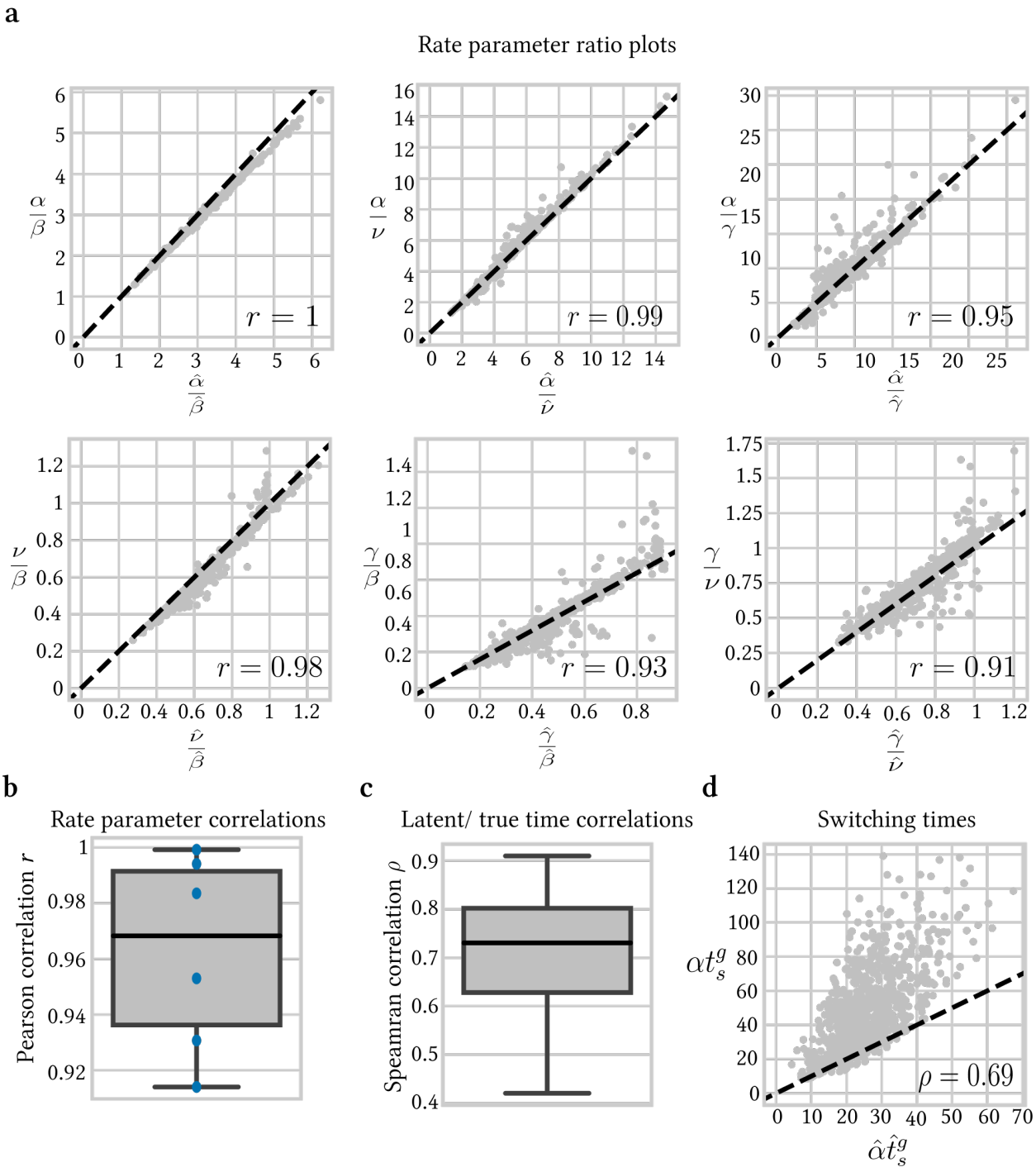

Figure 4.1 a. Rate parameter ratio plots of estimated against true underlying rate parameters. Estimated ratios are characterized by a circumflex. Black dashed lines represent the identity line. Pearson correlation coefficient r rounded to two decimal places is reported. **b.** Pearson correlation coefficient for all 6 rate parameter ratios. One blue point represents the Pearson correlation coefficient for one kinetic ratio pair. Box plots indicate the median (center line), interquartile range (hinges), and whiskers at 1.5× interquartile range. **c.** Spearman correlations of true scaled cell times and inferred scaled cell-gene latent times. Box plots indicate the median (center line), interquartile range (hinges), and whiskers at 1.5× interquartile range. **d.** Scaled inferred switching time against true scaled switching time. Spearman correlation coefficient ρ rounded to two decimal places is reported.

Subsequently, we defined the set of outlier genes as genes for which at least one of the gene-wise ratio deviations is above the respective 99-th percentile of the ratio vector r_i. The set of none-outlier genes is defined as the set of remaining genes.
Then, we compared mean-squared-errors (MSE) between the model's fit and true unspliced and spliced abundances by following the steps outlined in equations (2.12)-(2.14). In Figure 4.2b we reported $\log_{10}$ MSE scores. From here we can conclude that outlier genes report higher MSE scores in comparison to none-outlier genes. The one-sided Welch's test yielded $P < 10^{-3}$ for the MSE comparison of outlier genes and none-outlier genes for the nucleic abundances. For cytosolic spliced abundance we received $P < 10^{-1}$..
Subsequently, we investigated these outlier genes by plotting phase portraits including the inferred dynamics. The phase portraits depicted in Figure 4.2d show that the learned dynamics do not perfectly follow the underlying data. However, even for genes that exhibit the most significant deviations from the standard diagonal, oftentimes the predicted dynamics still align well with the simulated data as seen in Figure 4.2c.
Finally, it is worth noting that there are no evident indications of specific genes where the model fits the data more accurately. For example, the phase portrait of spliced cytosolic against unspliced nucleic abundances in Figure 4.2d demonstrates that spliced cytosolic abundance continues to increase after reaching the induction steady state. We identified this particular gene as an outlier gene based on the calculated ratio deviations in equation (4.1). However, the phase portrait in Figure 4.2e yields a similar characteristic, which was not identified as an outlier gene. For this none-outlier gene, the inference works better compared to the gene in Figure 4.2b.

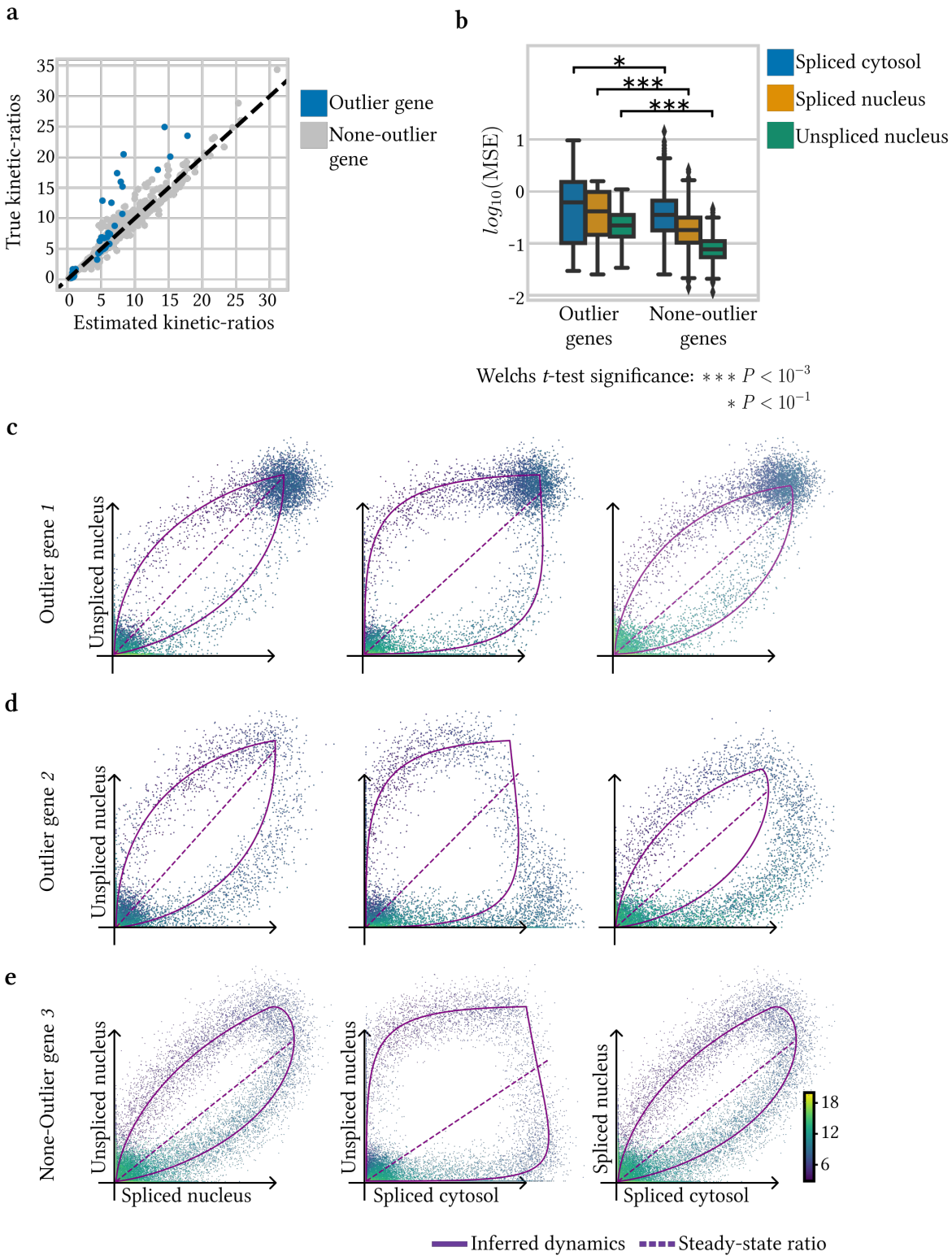

Figure 4.2 a. All rate parameter ratios are stacked together into one plot colored by outlier and none-outlier genes. Black dashed lines represent the identity line. **b.** Mean-squared-error comparisons between outlier and none-outlier genes for each feature. Box plots indicate the median (center line), interquartile range (hinges), and whiskers at 1.5× interquartile range. **c-d.** Phase portraits for two exemplary outlier genes colored by inferred latent time. The purple curve represents the inferred dynamics and the dashed line represents the respective steady state ratios. **e.** Phase portrait for none outlier gene colored by inferred latent time. The purple curve and dashed line are analogous to **c-d**.

4.1.3 Effect of introduced noise

Lastly, we explored the impact of introduced noise on simulated data. Employing identical parameters for data simulation, the only alteration involved adjusting the noise level parameter as described in Section 3.1. Specifically, we utilized $\sigma = [0.1, 0.3, 0.5, 0.7, 1, 1.25]$, resulting in six distinct datasets. For each of these simulated datasets, a separate model was trained by using the same hyperparameters for training. We compared the models based on metrics including MSE loss, Pearson correlations between rate parameter ratios, and the assignment of latent times.
Figure 4.3a reports the Pearson correlations between kinetic parameter ratios. Here, we further conducted one-sided Welch's tests between the model with the lowest noise level, *i.e.*, $\sigma = 0.1$, and all other noise level models. The lowest noise level model does not significantly better fit the rate parameters (P-value not significant).
Subsequently, we calculated Spearman correlations between the underlying and inferred cell times by following the same steps as described in Section 4.1.1. Here, we again conducted Welch's test between the lowest noise level model and all remaining noise level models. Figure 4.3b shows that for $\sigma \geq 0.7$, the Spearman correlations decrease, suggesting that the lower noise level models have higher Spearman correlations (for $\sigma \leq 0.5$ it holds $P < 0.001$ and $\sigma \geq 0.7$ results in none-significant P).
Afterward, we compared MSE losses between different noise levels. Figure 4.4 depicts $\log_{10}$-MSE scores for all different models per feature. From here we can conclude, that the lowest noise level model performs better than the highest noise level model, as the one-sided Welch's test yielded $P < 0.001$ for nucleic MSE scores and $P < 0.1$ for cytosolic spliced MSE scores. However, for the remaining noise level models, the MSE scores of the lowest noise level model are not significantly lower. Here, only the $\sigma = 1$ model yielded $P < 0.1$ for spliced nucleic MSE scores. In general, the MSE score remains consistent for $\sigma \leq 1$.
Finally, we computed the Pearson correlations of true and inferred switching times. The correlation coefficients varied between 0.68 and 0.75, with the highest correlation for the noise levels $\sigma \in \{0.7, 1.25\}$ and the lowest Pearson coefficient for the noise level $\sigma = 0.1$ as reported in Table.
In general, the *Nucleus-cytosol model* demonstrates robustness to varying levels of introduced noise, which

Noise level	Spearman correlation coefficient
$\sigma = 0.1$	$\rho = 0.68$
$\sigma = 0.3$	$\rho = 0.74$
$\sigma = 0.5$	$\rho = 0.71$
$\sigma = 0.7$	$\rho = 0.75$
$\sigma = 1$	$\rho = 0.73$
$\sigma = 1.25$	$\rho = 0.75$

Table 4.1 Spearman correlations of true and inferred scaled switching times calculated as $\text{corr}(\alpha t^s, \hat{\alpha}\hat{t}^s)$, where $\alpha, \hat{\alpha}, t^s, \hat{t}^s \in \mathbb{R}^G$.

is an encouraging indicator for its application to real-world sequencing protocols. As these protocols are not generated by exactly following the laws of the data simulation process, the ability of the model to handle noise in the data is a valuable attribute.

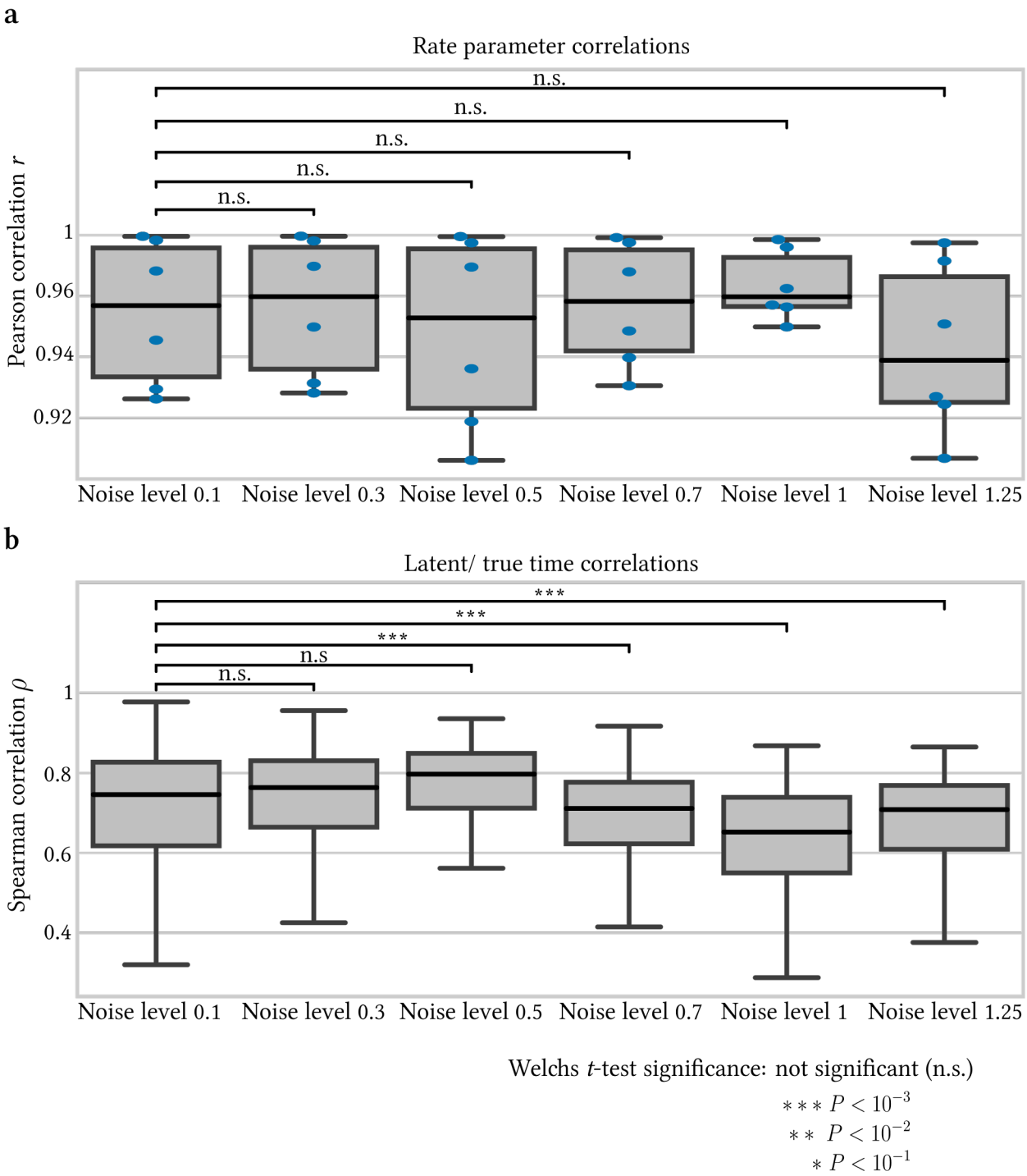

Figure 4.3 a. Pearson correlation coefficient distribution for all 6 rate parameter ratios per noise level. One blue point represents the Pearson correlation coefficient for one kinetic ratio pair. Box plots indicate the median (center line), interquartile range (hinges), and whiskers at 1.5× interquartile range. **b.** Spearman correlation between inferred and true latent times per noise level. Box plots indicate the median (center line), interquartile range (hinges), and whiskers at 1.5× interquartile range.

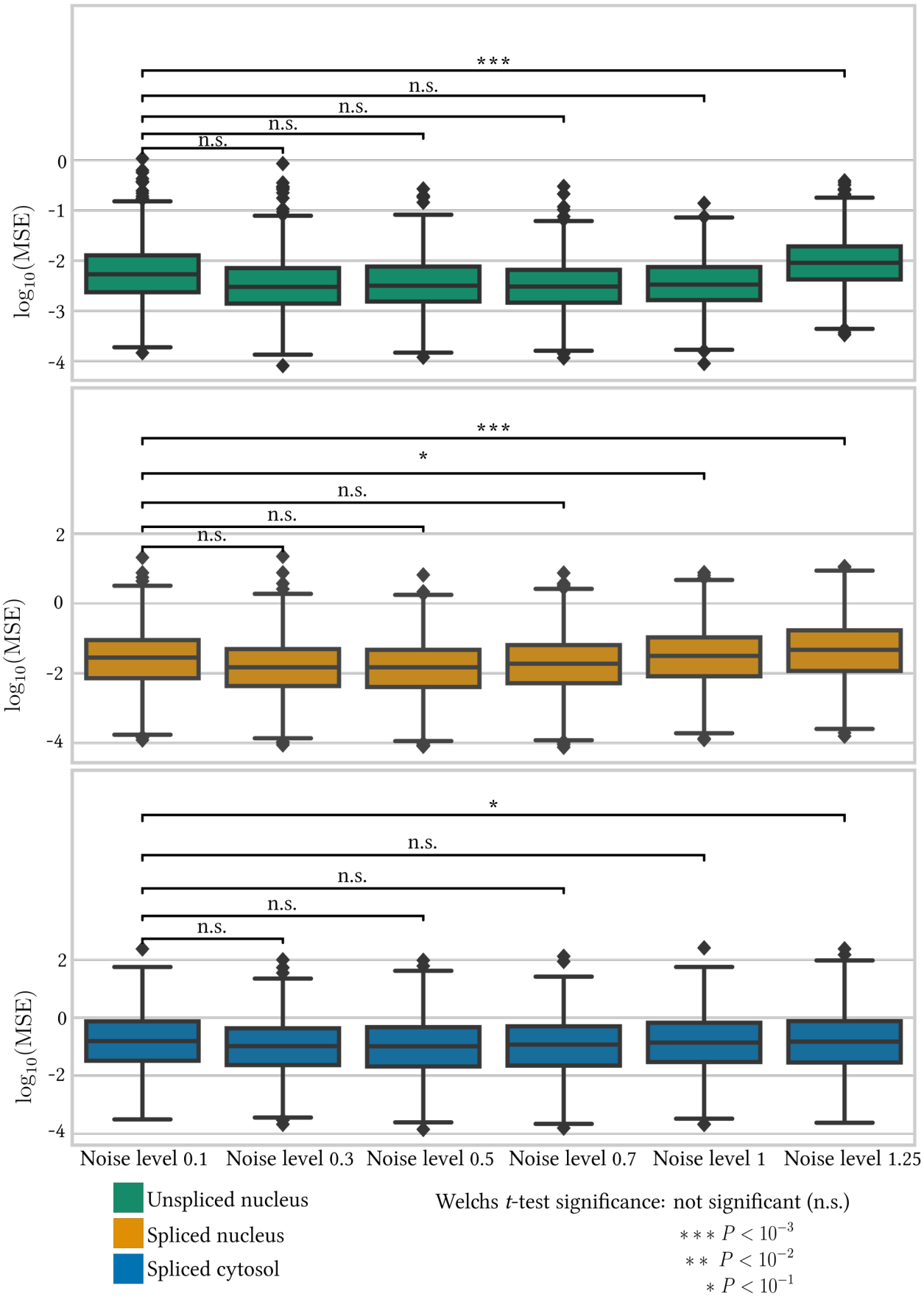

Figure 4.4 $\log_{10}$-MSE comparison between model's fit and true abundances for different noise levels. Box plots indicate the median (center line), interquartile range (hinges), and whiskers at 1.5× interquartile range.

4.2 Evaluation on pancreatic endocrinogenesis

In the following, we will evaluate the *Nucleus-cytosol model* on a dataset of pancreatic endocrinogenesis [22][23]. To be precise, the pancreas dataset contains single-nucleus transcripts and single-cell transcripts at embryonic day 14.5 [22][23]. To train the model on this dataset, we first constructed a batch-corrected latent space for the single-cell and nucleus observations as described in Section 3.2.1. Then, we computed a k-nearest neighbor graph on the joint latent space. Finally, we estimated nucleic and cytosolic RNA abundances by following the steps outlined in Sections 3.2.2 and 3.2.3.
During the assessment of the model on the pancreas datasets [22][23], we conducted a comparison involving three distinct models: (1) The *Nucleus-cytosol model* (2) a model trained exclusively on single-cell RNA sequencing data using the original *veloVI* framework and (3) a model exclusively trained on single-nucleus RNA sequencing data using the original *veloVI* framework. Here, we note that the direct application of RNA velocity models on single-nucleus data only describes the dynamics of the sub-cellular component within the nucleus

$$\frac{du_n^{(g)}(t)}{dt} = \alpha_{gk} - \beta_g u_n^{(g)}(t)$$
$$\frac{ds_n^{(g)}(t)}{dt} = \beta_g u_n^{(g)}(t) - \nu_g s_n^{(g)}(t).$$

Instead of modeling the degradation of RNA, for single-nucleus data, we describe the kinetics of nuclear export of spliced RNA.
To ensure consistency with model comparisons, we employed identical hyperparameters and an equal number of epochs to train the models.
Notably, the evaluation of the pancreas E15.5 datasets [22][23] yielded very similar results and can be found in Appendix A.1.

4.2.1 Permutation score analysis

The permutation score serves as a metric to evaluate if genes exhibit some structure in the phase portrait. It can further be used to identify partially observed dynamics or systems in steady states [14]. Hence, this metric can help assess the appropriateness of RNA velocity analysis for a specific dataset and, as a consequence, can be utilized to filter genes for further analyses [14]. In the following, we use permutation scores to compare the different models trained on their respective modalities.
As a first step, we trained all three models independently. Subsequently, we computed permutation scores following the procedure outlined in Section 2.5. For single-modal models trained using the *veloVI* framework, the permutation score was computed in a similar manner, with the exception of the third term describing spliced abundance in the cytosol.
The comparison between models involved the calculation of permutation score log ratios between two models for each gene and cell type cluster. To ensure consistency, only genes that were present across all datasets after pre-processing were considered for this analysis. If models perform equally well, we expect log ratios between permutation scores to be approximately 0. The box plots in Figure 4.5a show the distribution of the permutation log-ratios across common genes per cell type. Here, we see that all ratios are close to 0 with no noticeable differences between different model comparisons.
To summarize the permutation for a gene, we used the maximum permutation score across cell types [14]. The resulting permutation density is depicted in Figure 4.5b and the permutation distribution in Figure 4.5c. We performed one-sided Welch's tests between the permutation scores of the models. From here, we can conclude that applying data permutations yields similar permutation scores for the *Nucleus-cytosol model* and the single-cell model. However, the permutation scores of the single-nucleus model are significantly lower compared to the other two models (one-sided Welch's test $P < 0.1$ and $P < 0.01$). For data visualizations of specific genes, we selected the genes *Top2a*, a cell cycle marker in the pancreas datasets [22][23], and *Sulf2*, a marker of endocrine progenitor cells. The cell type-specific permutation scores for

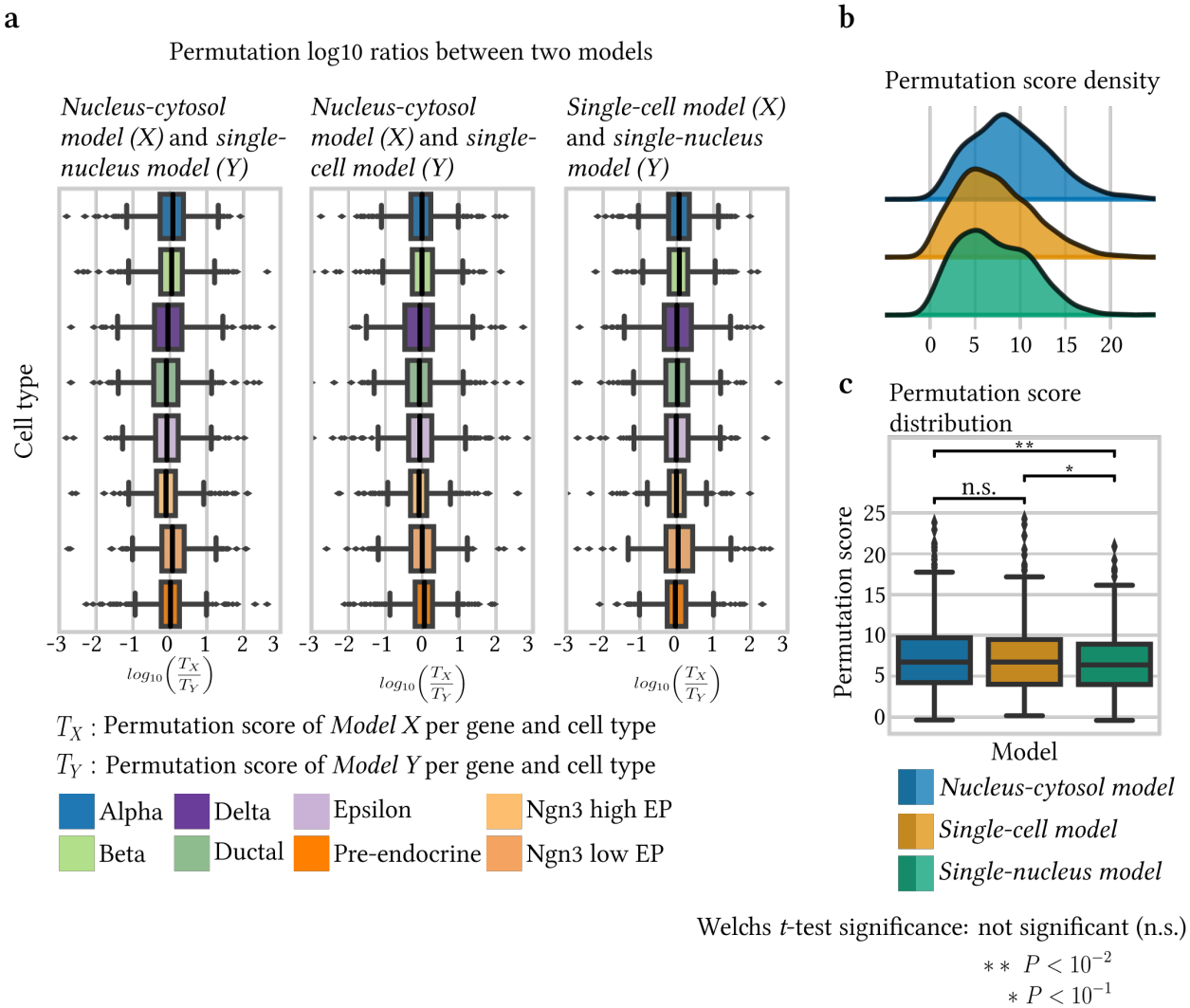

Figure 4.5 a. Permutation $\log_{10}$ ratios between two respective models colored by cell type for genes present in both respective datasets. Box plots indicate the median (center line), interquartile range (hinges), and whiskers at 1.5× interquartile range. **b.** Permutation score density per model. **c.** Permutation score distribution per model. Box plots indicate the median (center line), interquartile range (hinges), and whiskers at 1.5× interquartile range.

these genes are presented in Figures 4.6 and 4.7.
First, from Figure 4.6b we note that the model trained on the single-cell dataset exhibits the highest permutation scores for *Top2a*. In contrast, the model trained exclusively on single-nucleus data shows the lowest permutation scores. This preliminary insight possibly implies that applying *veloVI* directly on single-nucleus transcripts might not identify the cell cycle as *Top2a* is well suited to explain the vector field in the cycling progenitors [10]. The lower permutation scores can be attributed to the missing cytosolic spliced abundances in the single-nucleus dataset. The unpermuted phase portraits in Figure 4.6a show, that cytosolic spliced abundances have largely varying values within Ductal and Ngn3 high EP cells, leading to greater effects when permuting the data. However, when examining single-nucleus phase portraits, it appears that many Ductal cells are close to repression steady state, specifically, positioned in the lower-left corner of the phase portrait as shown in Figure 4.6b, therefore resulting in lower permutation scores.
However, for the gene *Sulf2* all models approximately exhibit similar permutation scores as shown in Figure 4.7d. Analyzing phase portraits of single-cell and nucleus data in Figure 4.7b,c suggest that both follow the assumptions of RNA velocity, resulting in the characteristic almond-shaped phase portraits.

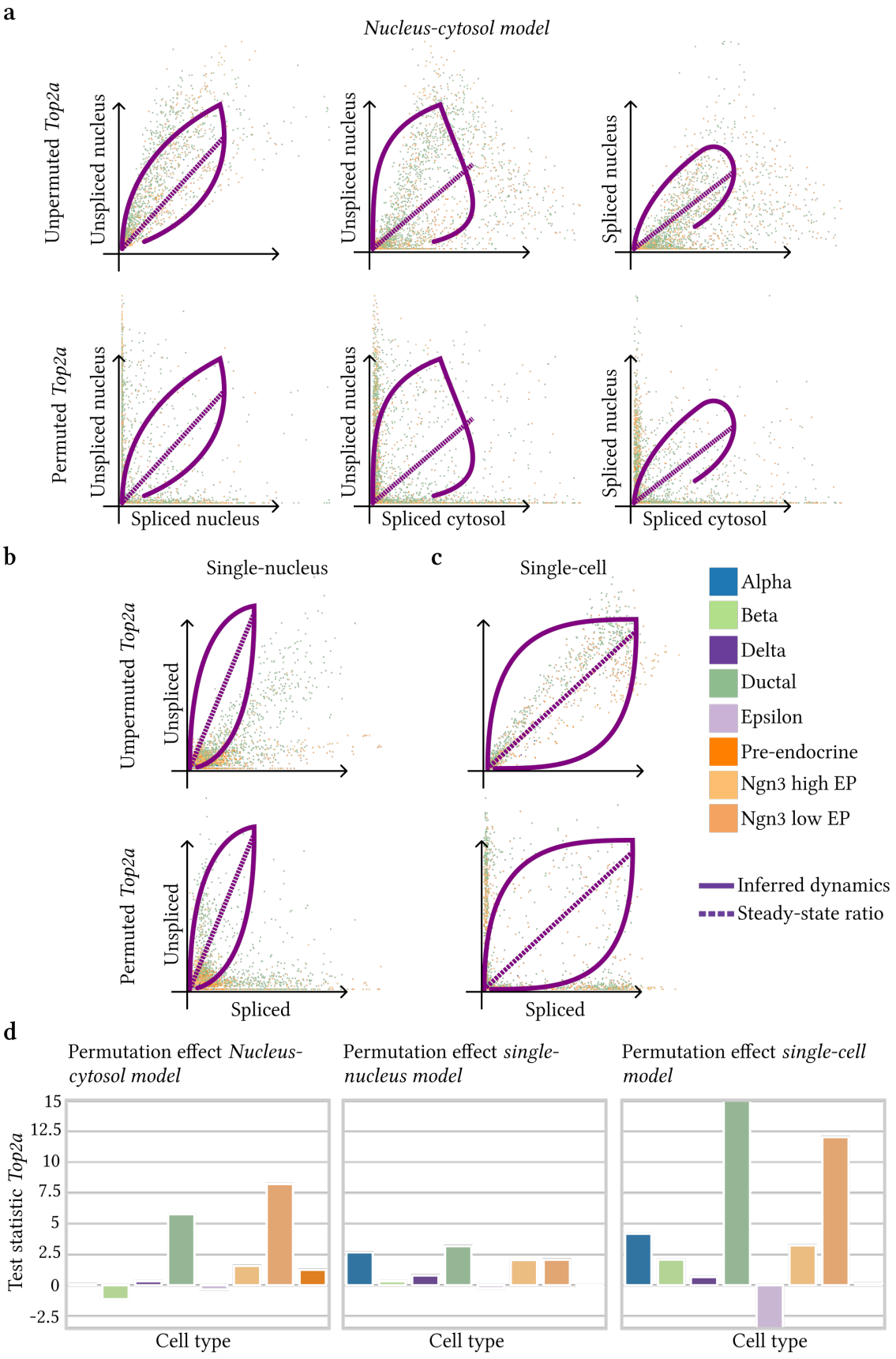

Figure 4.6 a-c. Unpermuted and permuted phase portraits of the gene *Top2a* for the *Nucleus-cytosol model* (**a**), the single-nucleus model (**b**), and the single-cell model (**c**). The purple curve represents the inferred dynamics and the dashed line represents the respective steady state ratios. **d.** Permutation scores for the gene *Top2a* per cell type for each model.

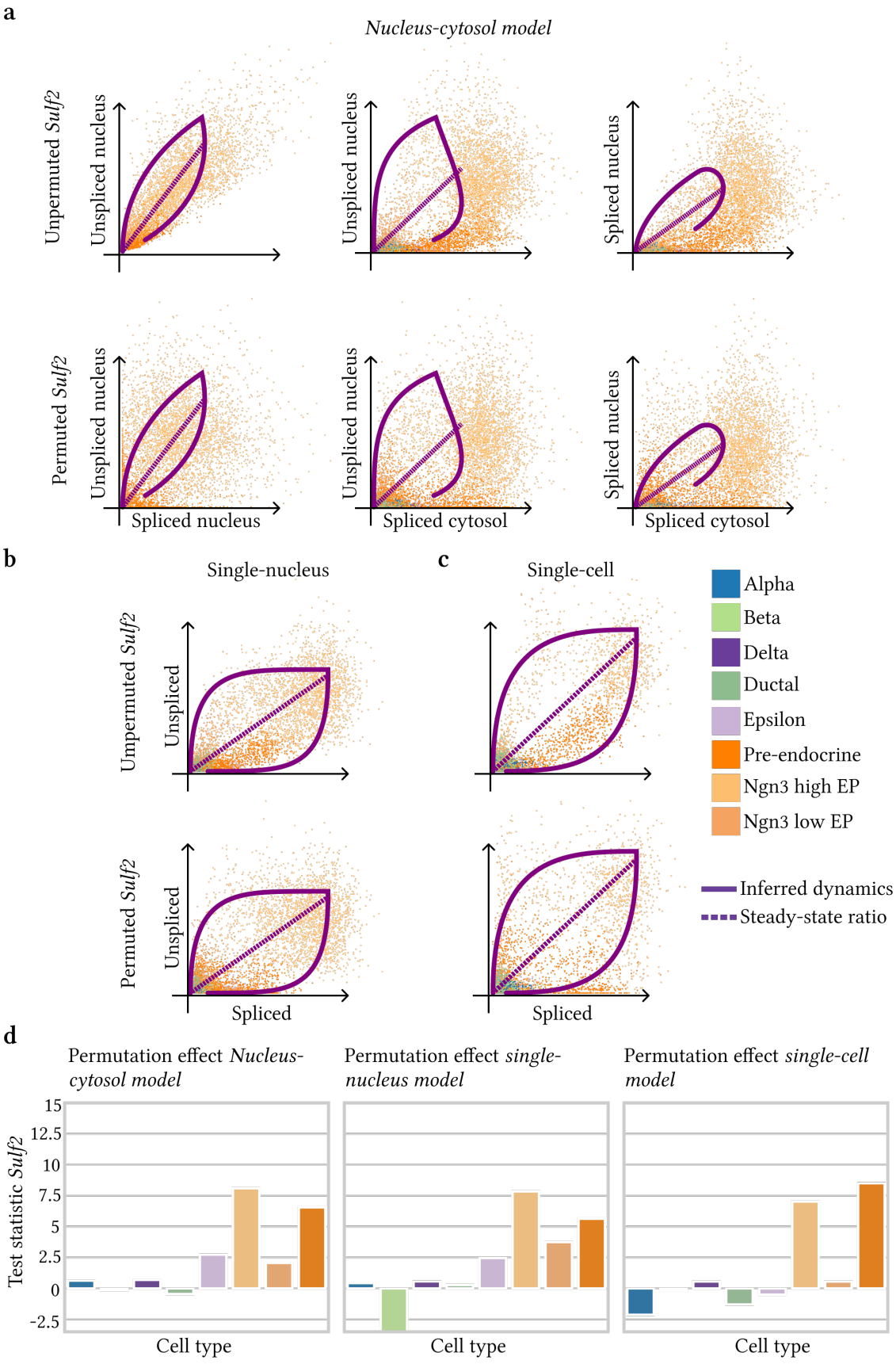

Figure 4.7 a-c. Unpermuted and permuted phase portraits of the gene *Sulf2* for the *Nucleus-cytosol model* (**a**), the single-nucleus model (**b**), and the single-cell model (**c**). The purple curve represents the inferred dynamics and the dashed line represents the respective steady state ratios. **d.** Permutation scores for the gene *Sulf2* per cell type for each model.

4.2.2 Velocity vector field comparison

In the following, we evaluated the velocity vector field generated by *scVI* and *scglue* based models. The evaluation includes the comparison of three different velocity modes as defined in equations (2.15), (2.17) and (2.18): (1) The time derivative of nucleic spliced mRNA denoted as v_{s_n}, (2) the time derivative of cytosolic spliced mRNA denoted as v_{s_c} and (3) the sum of both spliced time derivatives denoted as v_s. Unless specified otherwise, we refer to the single-modal velocities as v_s.
As a first step, we plotted phase portraits including the inferred velocity vectors for the genes *Hells* and *Top2a* which are well-suited to explain the vector field in the cycling progenitors, as well as for *Sulf2*, a marker of endocrine progenitor cells. The phase portraits are depicted in 4.8. However, from here we can not conclude if one velocity mode best captures the underlying dynamics.

Therefore to further evaluate and to decide which velocity mode best captures the dynamics, we considered Pearson correlations between the inferred velocities of the *Nucleus-cytosol model* and the inferred velocities of models trained on single-cell and single-nucleus protocols. To compare velocities of single-nucleus cells, we subset the dataset used for training the *Nucleus-cytosol model* to just consider single-nucleus observations. Similarly, to calculate Pearson correlations of velocities between single-cell observations, we subset the data used for training the *Nucleus-cytosol model* to just consider single-cell observations. Then, for each gene g present in both datasets after pre-processing, we calculated the Pearson correlation between the velocity vectors from two different models. Given that *veloVI* effectively captures the dynamics within the single-cell pancreas dataset [22][23], a strong correlation between the velocities of the *Nucleus-cytosol model* and the single-cell model serves as evidence that the *Nucleus-cytosol model* is similarly capable of capturing the underlying dynamics.
In Figure 4.9a we reported the velocity correlations between the *Nucleus-cytosol model* and the single-modal models for all three velocity modes. Here, we compared *scVI* [25] and *scglue* [24] based models. Figure 4.9a suggests that both models have equal correlations to the single-modal models, as the two-sided Welch's test yielded a non-significant P-value. However, we further compared the correlations of the *Nucleus-cytosol model* and the single-modal models. From here we can conclude that for all velocity modes and models, *i.e.*, *scVI* [25] and *scglue* [24] based models, the correlations to the single-cell model are elevated (one-sided Welch's test $P < 0.001$). This finding further underscores the need for a model integrating both modalities into a unified system. Moreover, from Figure 4.9b we note that the correlations for the summed spliced velocities v_s are highest for both, single-cell and nucleus models. Here, the correlations of the velocities v_s of the *Nucleus-cytosol model* and the single-modal velocities are significantly higher compared to the correlations of v_{s_n} (one-sided Welch's test $P < 0.1$). Subsequently, v_{s_n} yields significant higher correlations compared to v_{s_c} (one-sided Welch's test $P < 0.001$). Since the velocity v_s yields the highest correlations to the single-modal models, we decided to use this velocity mode for further visualization purposes.
Figure 4.9c further shows the correlations between *scglue* and *scVI* based *Nucleus-cytosol models* for the different velocity modes. Here, we note that the correlations are not much elevated compared to the velocity correlations reported in 4.9a,b. We can conclude, that *scglue* and *scVI* based models resulted in different velocity estimates as the correlations are generally not very high, and even models trained using exactly the same model definition, can result in different velocity estimates. We attribute this outcome to the different neighbor graphs used for the estimation process as described in Sections 3.2.1 and 3.2.2. As there is no clear preference in terms of high-velocity correlations, in the following we decided to use the *scglue* based model. This decision is further supported by the overall scIB metrics, which suggest the highest integration metrics for the *scglue* model as shown in Figure 3.3.
We further compared the velocity estimates at the level of individual genes. Here, we selected the markers *Sulf2* and *Top2a* and the velocity estimate v_s. But it is noteworthy that the velocity v_{s_n} resulted in very similar plots and correlations. Figure 4.9d illustrates the velocity estimates for *Sulf2* where the *Nucleus-cytosol model* has high correlations to both, single-cell and nucleus, models, therefore resulting in consistent velocity estimates for all models. However, for the cycling marker *Top2a*, the correlation to the single-cell velocities is elevated compared to single-nucleus velocities as depicted in Figure 4.9e. As par-

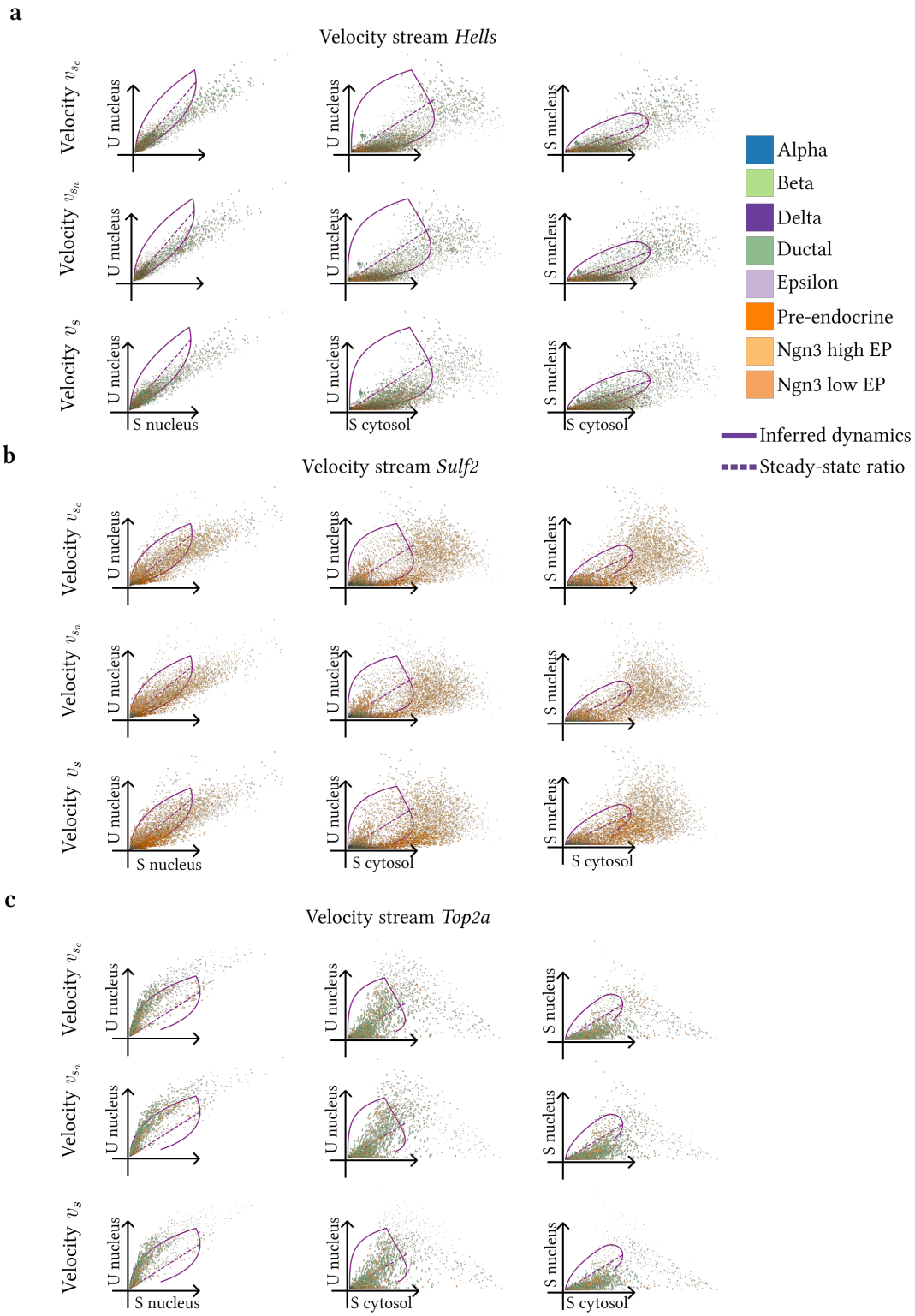

Figure 4.8 Phase portraits of *Nucleus-cytosol model* colored by cell type. Left column: Spliced nucleus against unspliced nucleus. Middle column: Spliced cytosol against unspliced nucleus. Right column: Spliced cytosol against spliced nucleus. We used the abbreviation "U" for "Unspliced" and "S" for "Spliced". The purple curve represents the inferred dynamics and the dashed line represents the respective steady state ratios. **a-c.** Phase portraits for the genes *Hells* (**a**), *Sulf2* (**b**) and *Top2a* (**c**) with velocity mode v_{s_c} on the top row, v_{s_n} in the middle row and v_s at the bottom row.

ticularly *Top2a* is well-suited to explain the vector field in the cycling progenitors, the low correlations to the single-nucleus model provide an indication that solely considering single-nucleus transcripts might not be sufficient to capture cell cycles.

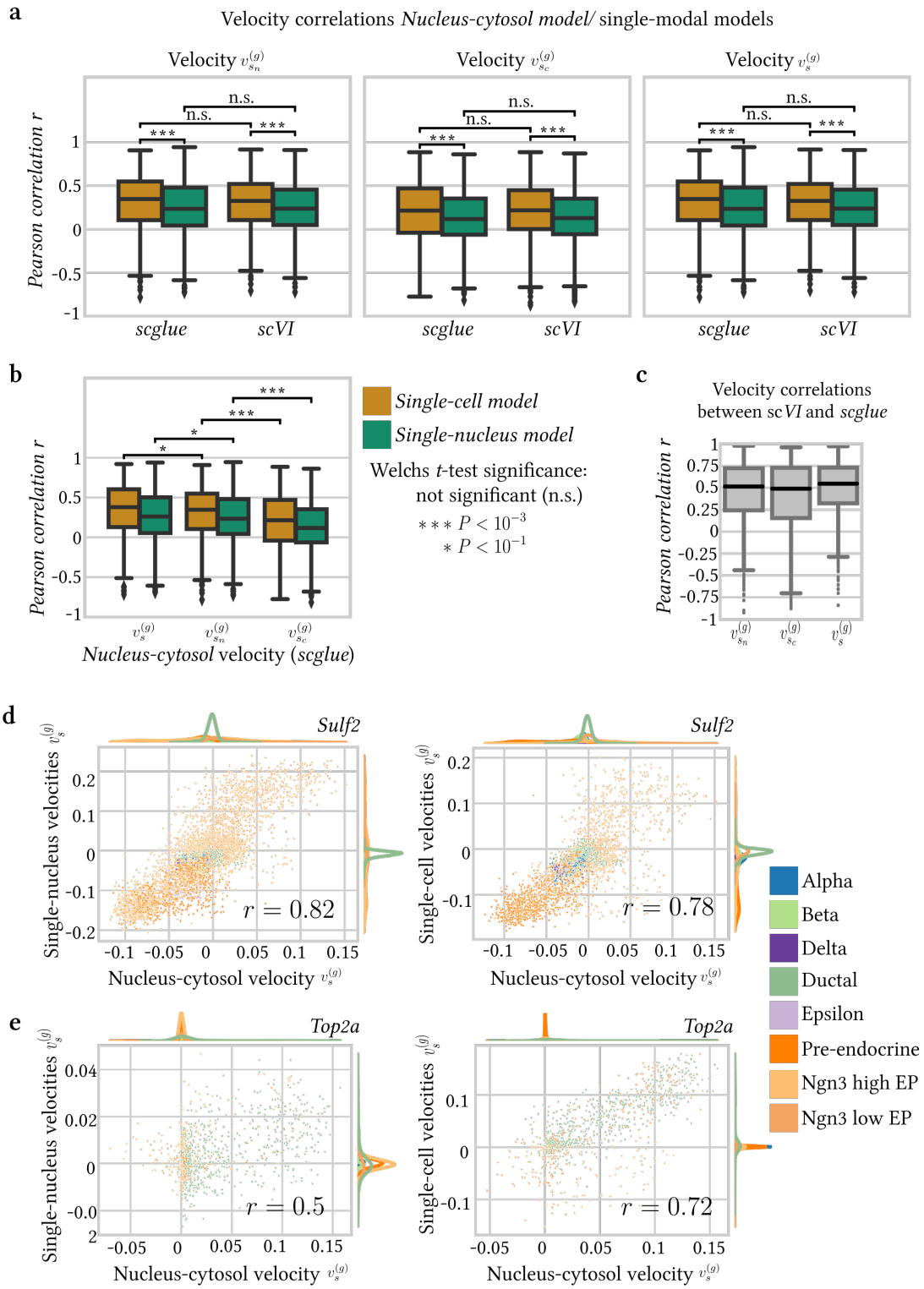

Figure 4.9 a. Velocity correlations between *Nucleus-cytosol model* and the single-nucleus and single-cell models for all three different velocity modes. We report correlations for the *scglue* (gene-expression) and *scVI* based models. Here, we compared *scglue* and *scVI* based models as well as their correlations to the single-modal models. Box plots indicate the median (center line), interquartile range (hinges), and whiskers at 1.5× interquartile range. **b.** Velocity correlations between *Nucleus-cytosol model* and the single-nucleus and single-cell models for all three different velocity modes. Here, we compared different velocity modes. Box plots indicate the median (center line), interquartile range (hinges), and whiskers at 1.5× interquartile range. **c.** Velocity correlations between *scglue* and *scVI* based models for different velocity modes. Box plots indicate the median (center line), interquartile range (hinges), and whiskers at 1.5× interquartile range. **c-d.** Velocity correlation for the genes *Sulf2* (**c**) and *Top2a* (**d**) between *Nucleus-cytosol model* (*scglue* based model) and single-nucleus on the left and single-cell on the right colored by cell type.

4.2.3 Velocity confidence

The metric velocity confidence [10] has been introduced to verify if a cell's velocity aligns well with the inferred velocities of neighboring cells. Here, we want to verify if the velocity of a cell stemming from one modality aligns well with the velocities of neighboring cells from the respective other modality. To do so, we trained a model on single-cell and nucleus transcripts respectively. Then, we reused the latent representations and neighbor graph computed on the batch-corrected latent space to calculate velocity confidences. We adjust the calculation as presented in Section 2.5 by only considering neighbor cells from its modality counterpart, for a single-cell observation i we define the "restricted" velocity confidence as

$$c_i = \frac{1}{k} \sum_{j \in \mathcal{N}_{\text{snRNA}}(i)}^{k} \text{corr}(v_i, v_j), \tag{4.2}$$

where we reused the notations for the restricted neighborhood of cell i as presented in Section 3.2.2. Similarly, for a cell i stemming from the single-nucleus protocol, we define the "restricted" velocity confidence as

$$c_i = \frac{1}{k} \sum_{j \in \mathcal{N}_{\text{scRNA}}(i)}^{k} \text{corr}(v_i, v_j). \tag{4.3}$$

The *Nucleus-cytosol model* in general has higher velocity confidences compared to the single-modal models as shown in Figure 4.10a. Here, we performed one-sided Welch's tests suggesting that v_{s_n} yield the highest confidences ($P < 0.001$ compared to v_{s_c} confidences). Subsequently, the velocity mode v_{s_c} yields similar confidences as v_s, as the P-value is not significant. Finally, the velocity confidences of the *Nucleus-cytosol model* are significantly higher compared to the single-modal models, as $P < 0.001$. We attribute the higher confidences of the *Nucleus-cytosol model* to the estimation process, as the missing abundances are calculated as the weighted average of its neighboring cells abundances. Therefore, the abundances of a cell are similar to its neighboring cells and subsequently, the velocities will likely also be very similar, thus resulting in high-velocity confidences. Lastly, we observe that the single-cell confidences are significantly higher compared to single-nucleus confidences ($P < 0.001$).
Figure 4.10b-e shows that the velocity confidences for all models are in general very high. However, when calculating the confidences as described in equations (4.2) and (4.3), we clearly see a decrease of confidences as shown in Figure 4.10a,e. We can conclude that the velocity estimates of the model trained on the single-nucleus dataset do not perfectly align with the velocity estimates of its neighboring cells from the single-cell dataset. This observation indicates that the application of *veloVI* on single-nucleus data does not result in similarly reliable velocity estimates as for single-cell transcripts. The highest confidences reported within Ngn3-low endocrine are progenitor cells, with a maximal value of approximately 0.6.
Additionally, we observe from Figure 4.10b-e that the single-cell model exhibits greater confidence levels within the cycling Ductal cell population compared to other cell types. Similarly, the *Nucleus-cytosol model* has greater confidence in Ductal cells compared to other cell types. Contrastingly, the single-nucleus model displays lower confidence for Ductal cells when compared to other cell types of the respective model. In this context, we employed the identical UMAP embeddings as depicted in 4.15, which allows us to determine the regions for all cell types in the UMAP embeddings.

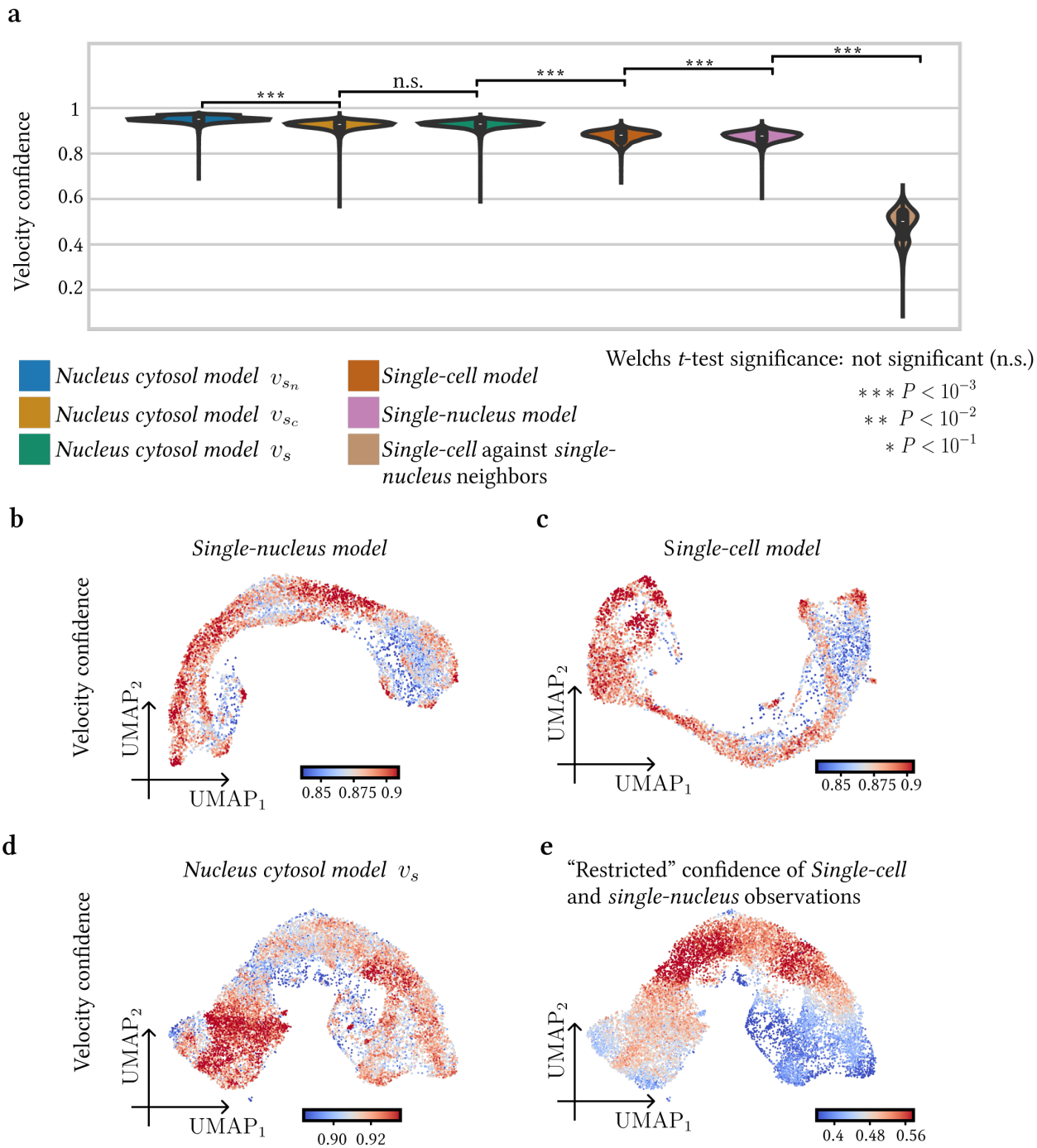

Figure 4.10 a. Velocity confidences for all models. **b.** Velocity confidence for single-nucleus model. **c.** Velocity confidence for the single-cell model. **d.** Velocity confidence for single-nucleus model for velocity mode v_s. **e.** Velocity confidence as described in equations (4.2) and (4.3).

4.2.4 Cell cycle analysis

To evaluate if the *Nucleus-cytosol model* is capable of faithfully identifying the cell cycle within the pancreas datasets [22][23], we again started by training all three models on their corresponding data. Then, for single-cell as well as for single-nucleus observations we calculated S-scores as well as G2M-scores using the list of cell cycle genes [30], which are present in the respective datasets. The scores are defined as the average expression of the two sets of genes associated with the S- and G2M phase respectively, subtracted with the average expression of a reference set of genes [31]. Here, the reference set was sampled randomly. Based on the calculated scores we assigned a cell cycle phase, including gap 1 (G1), synthesis (S), and gap 2/ mitosis (G2M) phases, to each cell. All steps are implemented within *scvelo*'s tooling function `score_genes_cell_cycle`.
After assigning a cell cycle phase to each cell, we permuted unspliced and spliced abundances within each cell type. For this analysis, we only consider the permutation effect for cycling genes on Ductal cells. For comparisons between trained models on different modalities, we will just consider the list of cell cycle genes present within all datasets after pre-processing.
The permutation scores, as depicted in Figure 4.11a,b, reveal that the single-nucleus model shows significantly lower permutation scores for cycling genes compared to the single-cell model and the *Nucleus-cytosol model* (one-sided Welch's test $P < 0.001$). The single-cell model and the *Nucleus-cytosol model* have no significantly different permutation scores for cycling genes. The observations indicate that solely considering the single-nucleus pancreas dataset [23] does not capture cell cycles as Ductal cells do not have largely varying abundances and are thus resistant to data permutations. However, when calculating permutation scores it is noteworthy that for the *Nucleus-cytosol model*, we calculate the sum of mean-absolute-errors (MAE) between three layers, *i.e.*, nucleic unspliced and spliced abundances as well as cytosolic spliced abundances. Subsequently, the differences between the MAEs of unpermuted and permuted input features can be elevated by nature compared to the single-cell or nucleus permutations. Therefore, to compare the models trained on different modalities, we further need to investigate cell cycle transition probabilities. Then, we can validate whether one of the models performs preferably in terms of cell cycle identification.
In the following, we will calculate cell cycle phase transition probabilities. To achieve that, we first compute a cell-cell transition matrix $T \in \mathbb{R}^{N \times N}$. Here, we relied on the functionalities implemented in *Cellrank* [32] to calculate cell-cell transition probabilities based on the inferred velocities of the cycling genes. Given the k-nearest neighbor graph $G \in \mathbb{R}^{N \times N}$, for each cell i, the transition probability T_{ij} to a neighboring cell j is calculated as

$$T_{ij} = \frac{e^{\text{corr}(v_i, \delta_{ij})}}{\sum_{k \in \mathcal{N}(i)} e^{\text{corr}(v_i, \delta_{ik})}},$$

where $v_i \in \mathbb{R}^G$ denotes the velocity vector of cell i, δ_{ij} the difference in gene expression between cells j and i and "corr" denotes the Pearson correlation between both vectors [32]. All steps are implemented within *Cellrank*'s `VelocityKernel` class with `gene_subset` defined as the list of cycling genes present in the datasets.
Consequently, for every cell i, we can compute the transition probabilities from its current cell cycle phase to other phases, *i.e.*, by summing over the transition probabilities of all cells belonging to the target cycle phase. Let C_{G1}, C_S and C_{G2M} denote the sets of all cells assigned to the respective phases. Then, for any cell $i \in [C_{G1}, C_S, C_{G2M}]$ it holds that

$$\sum_{j \in C_{G1}} P(i \rightarrow j) + \sum_{j \in C_S} P(i \rightarrow j) + \sum_{j \in C_{G2M}} P(i \rightarrow j) = 1,$$

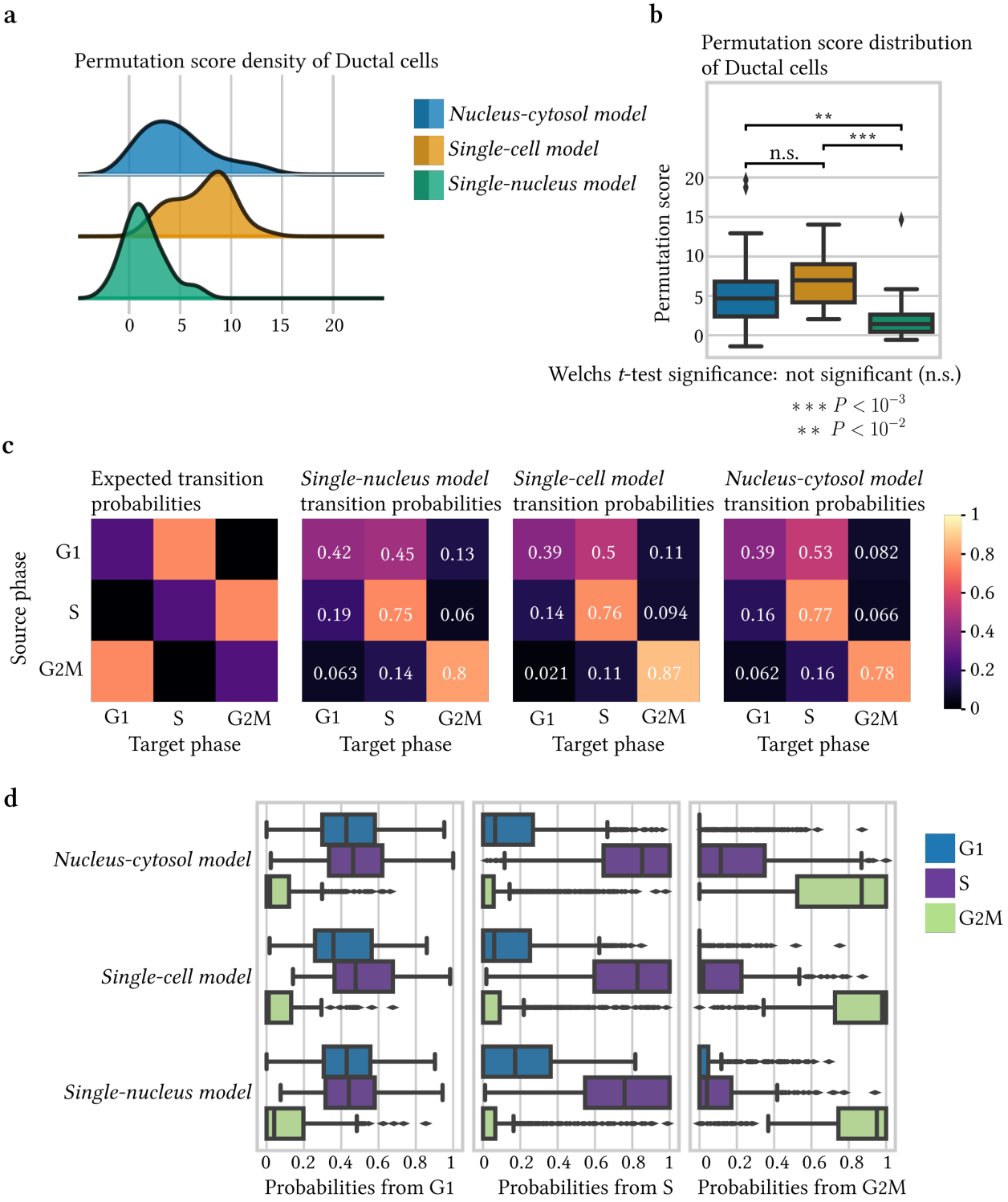

Figure 4.11 a. Permutation score densities on Ductal cells solely for cycling genes present within all three datasets after pre-processing. **b.** Permutation score distribution of Ductal cells per model. Box plots indicate the median (center line), interquartile range (hinges), and whiskers at 1.5× interquartile range. **c.** Expected cell cycle transition matrix and transition matrices for all models colored by probability. **d.** Transition probability distributions for cell cycle phases and trained models. Box plots indicate the median (center line), interquartile range (hinges), and whiskers at 1.5× interquartile range.

where $P(i \rightarrow j) = T_{ij}$. In other words, the sum over columns of the matrix T yields the unit vector. Subsequently, we define the overall cell cycle transition probabilities as the average probabilities over cells from a source phase $X \in [\text{G1}, \text{S}, \text{G2M}]$ to a target phase $Y \in [\text{G1}, \text{S}, \text{G2M}]$ as

$$P(X \rightarrow Y) := \frac{1}{|C_X|} \sum_{i \in C_X} \sum_{j \in C_Y} P(i \rightarrow j),$$

where C_X, C_Y denote the sets of cells assigned to the source and target phase respectively.
Figure 4.11c illustrates an expected transition matrix. We reason that a cell $i \in C_X$ assigned to a source phase X will either point to cells within the same cycle phase, *i.e.*, if cell i is at the beginning of phase X, or towards the next cell cycle phase, *i.e.* if cell i is at the end of the current cell cycle phase. The underlying biological cell cycle transitions are given by $G1 \rightarrow S \rightarrow G2M \rightarrow G1$. In more detail, for a cell $i \in C_{\text{G1}}$ we expect its transition probabilities to be highest for S and G1. We reason that, cells in the early stages of the G1-phase should rather point towards other G1 cells, while cells approaching the end of G1-phase should point towards the S-phase. Similarly, for a cell $i \in C_{\text{S}}$ we expect the transition probabilities to be highest for G2M and S and for a cell $i \in C_{\text{G2M}}$ the transition probabilities for G1 and G2M are expected to be highest.
The transition matrices for all three models are reported solely for the cycling population of Ductal cells in Figure 4.11c. Interestingly, none of the models exhibit a clear preference for solving the underlying biological cell cycle transitions $G1 \rightarrow S \rightarrow G2M \rightarrow G1$. While the *Nucleus-cytosol model* and single-cell model reveal the highest probabilities for $P(G1 \rightarrow S)$, the single-nucleus model and the *Nucleus-cytosol model* get assigned the highest probability for $P(G2M \rightarrow G1)$. This observation indicates that the *Nucleus-cytosol model* incorporates the respective higher transition probabilities from both single-modal models, highlighting that the model is able to faithfully resolve the underlying cell cycle transitions. However, for the transition from S to $G2M$, the single-cell model has the highest probability.
Furthermore, Figure 4.11d reports transition probability distributions. Here, one value refers to the transition probability $P(i \rightarrow j)$ of a cell i assigned to a source phase to a cell j from the target phase. Figure 4.11d suggests that the transition probabilities of the models approximately follow the same distributions. However, it is noteworthy that as seen in Figure 4.11b the average transition probabilities slightly differ for the models.
To visualize the results we further plotted phase portraits of cycling genes. If all models equally likely capture cell cycles, we expect to observe clear patterns within the phase portraits, *i.e.*, cells pointing from one phase to the next cycle phase $G1 \rightarrow S \rightarrow G2M \rightarrow G1$. Considering phase portraits of cycling genes for single-nucleus observations as depicted in Figure 4.12 and 4.13 revealed that many genes do not display the underlying cell cycle dynamics in contrast to the single-cell phase portraits. For many genes, *i.e. Birc5*, *Cenpe*, *Cdc20*, *Ccnb2* or *Tpx2*, we observe that unspliced abundance is highest for cells with low spliced abundance. Therefore, many cells are located in the left corner of the phase portrait as highlighted in Figure 4.12b for the gene *Cenpe*. This observation might be explained by the missing cytosolic abundance in the case of single-nucleus observations. With the additional spliced cytosolic abundance, some of these cells will have higher spliced abundances and therefore shift to the right in the phase portrait, resulting in phase portraits that follow the assumptions of RNA velocity more accurately. Moreover, we note that the observation aligns with the significantly lower permutation scores of the single-nucleus model, as the phase portraits do not follow the assumptions of RNA velocity and therefore exhibit less structure in the phase portraits.
However, for the *Nucleus-cytosol model*, we depend on single-nucleus observations. The abundances are processed as measured during the estimation steps, with the exception of applying the lambda correction to mitigate batch effects as discussed in Sections 3.2.2 and 3.2.3. Within the phase portraits as depicted in Figure 4.12 and 4.13 we see that nucleic phase portraits approximately follow the dynamics. However, examining the phase portraits including spliced cytosolic abundances, we see that many cells with $s_c^{(g)} \geq 0$ have near zero nucleic abundances, resulting in "noisy" phase portraits that do not align very closely with the modeled dynamics. This observation was consistent for most cycling genes.
In summary, the single-nucleus model reported the lowest permutation scores, and examining phase por-

traits of cycling genes revealed that these genes do not closely align with the underlying model assumptions. However, by considering transition probabilities, all models exhibited relatively similar performance in terms of resolving the underlying cell cycle transitions. This observation demonstrates that all models learned useful information regardless of permutation scores or the underlying phase portraits. Therefore, we can conclude that *Cellrank* provides reliable transition matrices for all models.

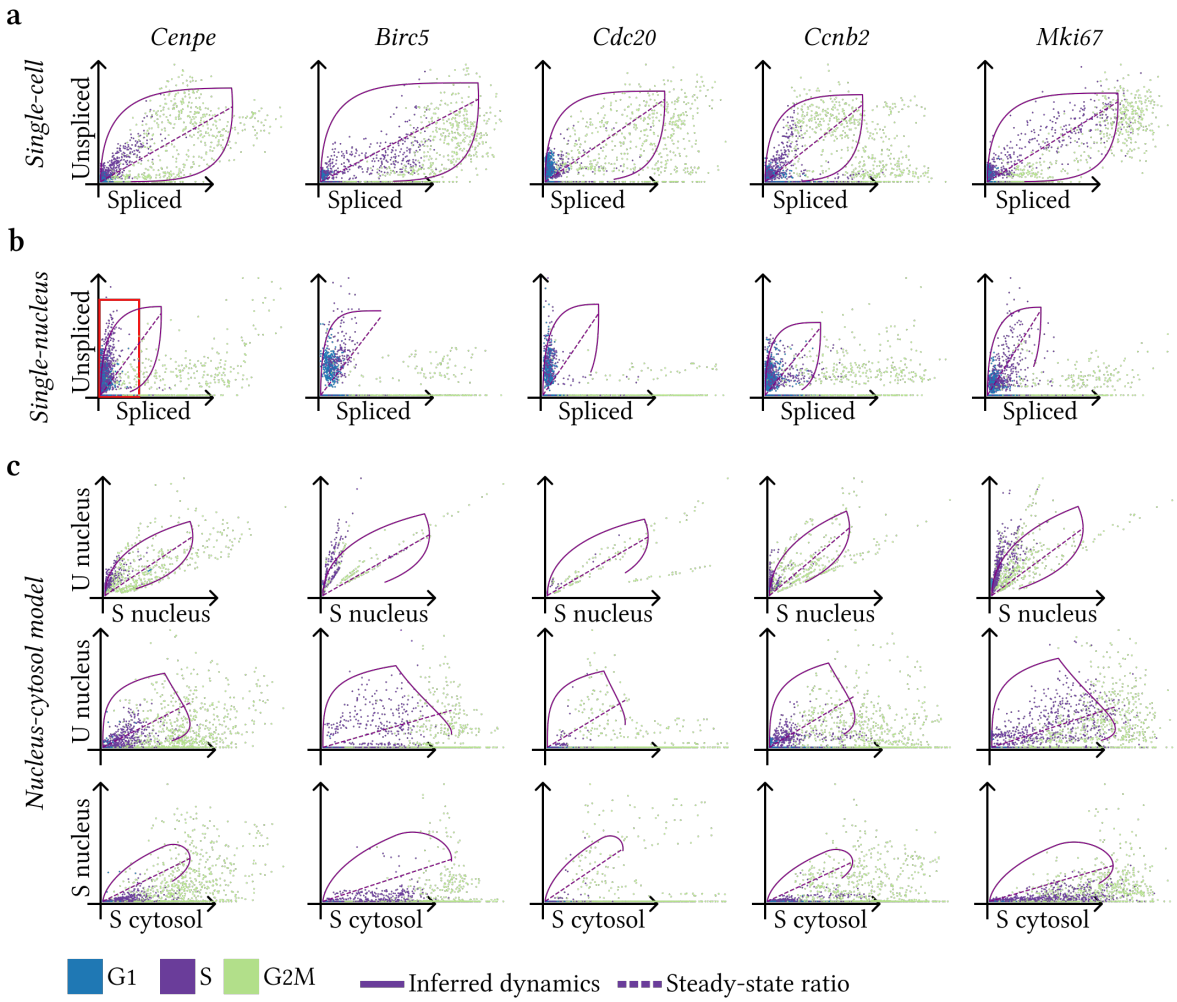

Figure 4.12 a-c. Single-cell (**a**), single-nucleus (**b**), and imputed phase portraits (**c**) for multiple cell cycle genes colored by cell cycle phase. We used the abbreviation "U" for "Unspliced" and "S" for "Spliced". The purple curve represents the inferred dynamics and the dashed line represents the respective steady state ratios. The red rectangle in **b** indicates cells with very small spliced abundance but with largely varying unspliced abundance. This characteristic can be found in multiple other single-nucleus phase portraits of cycling genes.

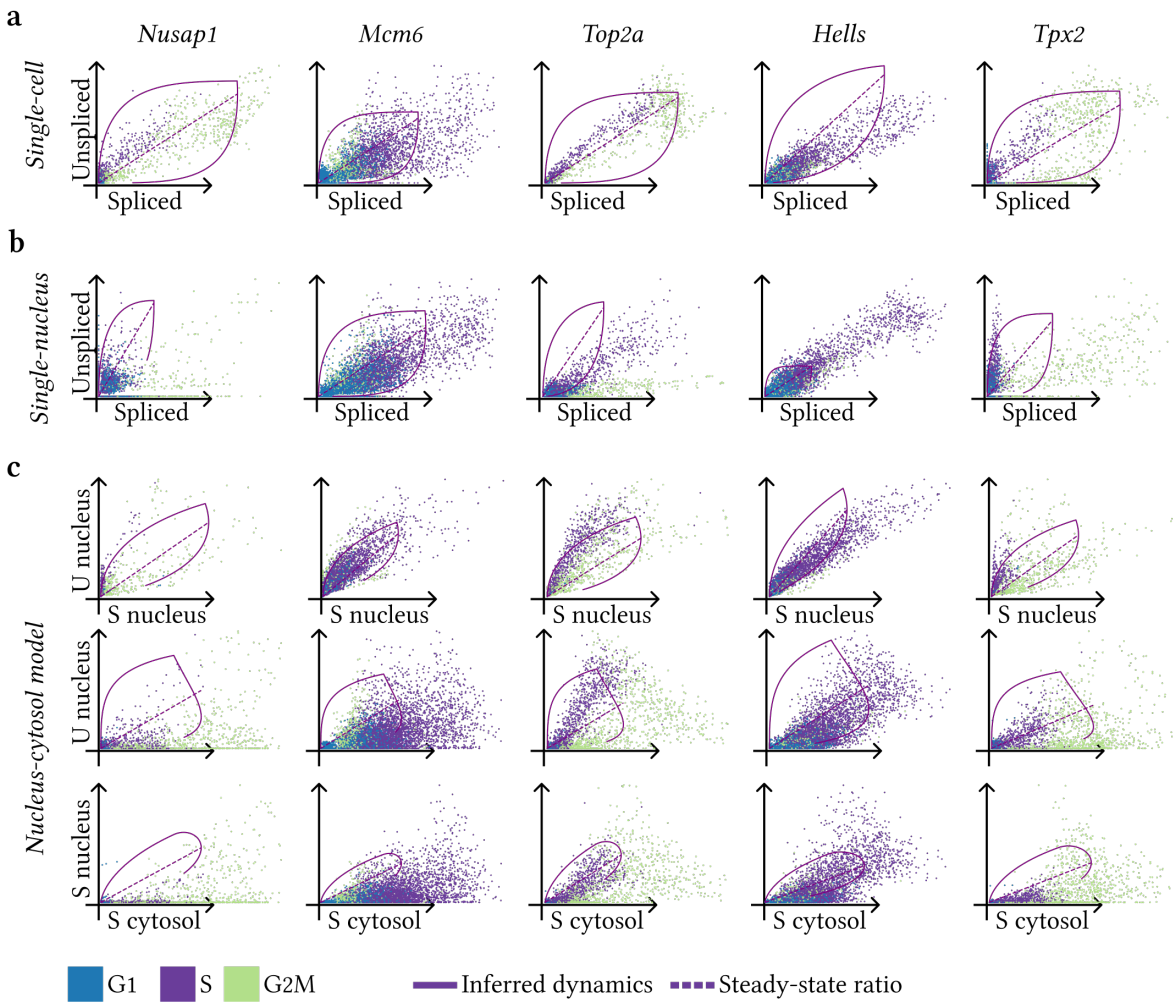

Figure 4.13 a-c. Single-cell (**a**), single-nucleus (**b**), and imputed phase portraits (**c**) for multiple cell cycle genes colored by cell cycle phase. We used the abbreviation "U" for "Unspliced" and "S" for "Spliced". The purple curve represents the inferred dynamics and the dashed line represents the respective steady state ratios.

4.2.5 Uncertainty analysis

As demonstrated in [14], Ductal cells, in which cycling cells are observable within the pancreas datasets [22][23], exhibit an elevated level of intrinsic and extrinsic uncertainty as defined in Sections 2.5 and 2.5. Therefore, employing uncertainty measures could serve as another valuable metric to gain insight into the model's ability to capture cell cycles.
Intrinsic uncertainty can be used to measure the variability in the phenotypic directionality suggested by the velocity vector in each cell [14]. Extrinsic uncertainty on the other hand can quantify the variability of predicted future cell states under the velocity-induced cell-cell transition matrix. Notably, the extrinsic uncertainty measure encompasses both, the variability among the cell's neighborhood and the intrinsic uncertainty due to processing posterior velocity samples through the calculation [14].
First, we compared the scales of uncertainty measures between models in Figure 4.14. While the intrinsic uncertainty of the *Nucleus-cytosol model* is significantly lower compared to the uncertainties of the single-modal models (one-sided Welch's test $P < 0.001$), the extrinsic uncertainty is significantly higher (one-sided Welch's test $P < 0.001$). In the following, we will only compare uncertainty values within one model, as each model trained on its respective modality produces uncertainty values on different scales. Hence, the illustrated uncertainty scales refer to the scale within the uncertainty value range for one model. For instance, comparisons of regions with high and low uncertainties always refer to the uncertainty scale of its respective model if not stated otherwise.
From Figure 4.15a,b we observe that the single-cell as well as the single-nucleus models approximately highlight the same regions and cell types for the intrinsic and extrinsic uncertainties as shown in Figure 4.15a,b. Here, intrinsic uncertainty was notably higher in Ductal and Ngn3-low endocrine progenitor populations. Additionally, extrinsic uncertainty highlighted the same populations along with terminal Alpha and Beta cells. As already shown in ref. [14], this finding demonstrates that lower intrinsic uncertainty does not necessarily indicate lower extrinsic uncertainty. In the case of Alpha and Beta cells, which are terminal populations in the pancreas datasets [22], the high extrinsic uncertainty can be explained as there are no observed successor states. Conversely, transient cell populations like Ngn3-high endocrine progenitors and pre-endocrine cells exhibit lower uncertainties. The low intrinsic uncertainty in these cells likely results from their dynamics aligning well with the underlying model assumptions [14]. Moreover, the low extrinsic uncertainty suggests that these cell types have distinct successor populations in the respective dataset [14].
For the *Nucleus-cytosol model* we included the uncertainty embeddings for the velocity modes v_{s_n}, v_{s_c}, v_s in Figure 4.15c. While the intrinsic uncertainty is elevated for Ngn3 high EP cells in the case of spliced nucleic velocity, the velocities v_{s_c}, v_s highlighted Ductal, Ngn3 high and low EP cells. This observation clearly aligns with the dynamics of the population. The elevated uncertainty for the initial states can be interpreted as the cell's fate whether to transition towards terminal Alpha or Beta cells or not. Once the decision about the future state is made, the uncertainty decreases.
The extrinsic uncertainty however is elevated for most cell types, but especially in Ductal, Ngn3-high endocrine progenitors, Pre-endocrines, and terminal Alpha and Beta cells for all velocity modes. However, while v_s has a high uncertainty in most regions, the velocities v_{s_n}, v_{s_c} have lower uncertainty regions for Ductal, Ngn3 high EP, Pre-endocrine, and Alpha cells. Overall, the uncertainties indicate high variability in the phenotypic directionality suggested by the velocity vector in each cell, which possibly can be attributed to the estimation process.
The abundance in the respective part of the cell depends on its nearest neighbors from its modality-counterpart. Namely, nucleic abundances for single-cell observations rely on its single-nucleus neighbors of the batch corrected k-nearest neighbor graph and vice versa. Therefore, for a cell i with two neighboring cells l, m pointing in the opposite direction with similar connectivity strength, the extrinsic uncertainty can be elevated. Here, this scenario could be applicable to many cells within the pancreas datasets [22].
As a next step, for the *Nucleus-cytosol model*, we assigned a pseudotime to each cell which is computed by *scvelo*'s [10] `velocity_pseudotime` function with `vkey` defined as the layer containing the velocities v_s. Subsequently, we fitted *Cellrank*'s [32] `GPCCA` estimator based on the transition matrix calculated by the `VelocityKernel` to obtain fate probabilities for each cell. Finally, Figure 4.16a shows the UMAP

embedding of the *Nucleus-cytosol model* colored by the inferred pseudotimes. From here we can conclude, that pseudotimes are correctly assigned from initial states within Ductal cells and terminal Alpha and Beta cells. Subsequently, we plotted pseudotimes against fate probabilities as seen in Figure 4.16b-d. From here we can see that the extrinsic uncertainty is elevated especially within Ductal and Ngn3 high EP cells. The elevated uncertainty in initial Ductal cells can be attributed to the cycling nature of this population. Here, a cell might continue to cycle or it might exit the cell cycle, therefore yielding higher uncertainties. For Ngn3 high EP cells the elevated uncertainties can be interpreted as the cell's fate point into which future state it will transition. This region serves as a major decision point on whether to transition into the direction of terminal Alpha or Beta cells or not.
Similarly, Figure 4.16d suggests that the uncertainty within velocity samples decreases once the fate decision is conducted. As pseudotime and fate probability increase, the intrinsic uncertainty decreases.
The same plots are reported for the single-cell and nucleus pancreas datasets [22][23] in Figures 4.17 and 4.18. For both modalities, we can infer similar interpretations. The single-cell model exhibits elevated extrinsic uncertainties within Ductal cells. Moving in time, the uncertainty first decreases before it increases for the Pre-endocrine cell type cluster. We can conclude, that in the single-cell model, in contrast to the *Nucleus-cytosol model*, the cell's fate is decided within Pre-endocrines, rather than within Ngn3-high EP cells. The intrinsic uncertainty is strongly elevated for Ductal cells and decreases over time. Towards terminal states, the intrinsic uncertainty slightly increases.
The single-nucleus model exhibits elevated extrinsic uncertainties for Ductal, Ngn3-high EP cells, and Pre-endocrine cells. Figure 4.18c therefore suggests that a cell's fate decision is made towards Ngn3-high EP and Pre-endocrine cells, which is supported by the elevated intrinsic uncertainties in the same regions.

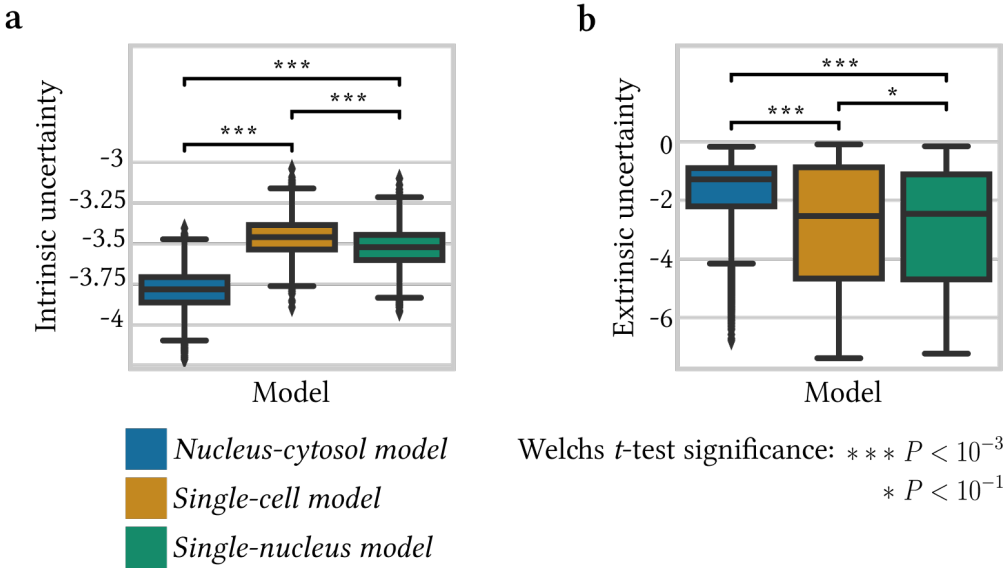

Figure 4.14 a. Comparison of intrinsic uncertainty measures between models. A high value corresponds to high uncertainty and a low value to low uncertainty. **b.** Comparison of extrinsic uncertainty measures between models. A high value corresponds to high uncertainty and a low value to low uncertainty.

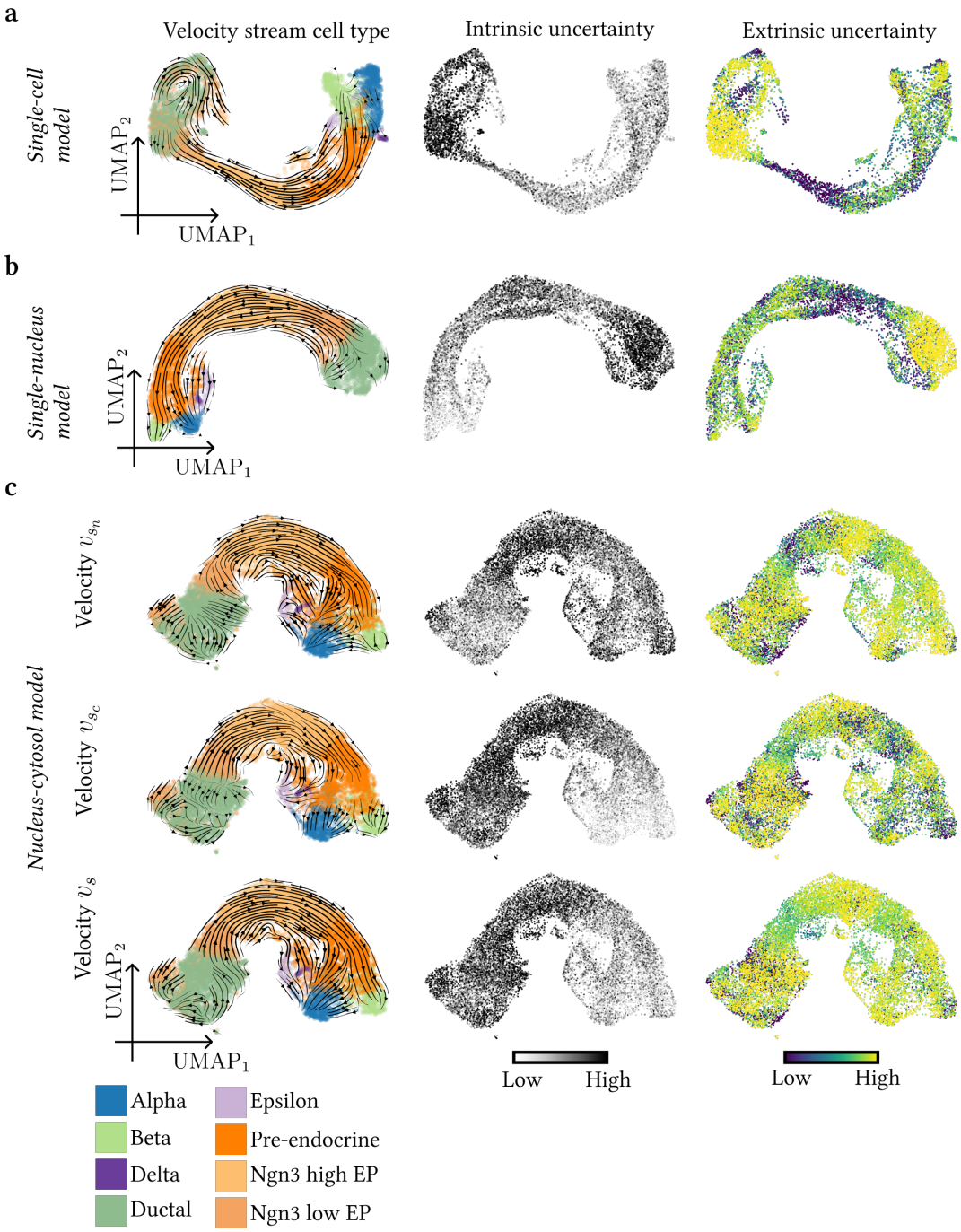

Figure 4.15 The left column shows the velocity stream colored by cell types. The middle column shows UMAP embeddings colored by intrinsic uncertainties. The right column shows UMAP embedding colored by extrinsic uncertainty. **a.** UMAP embeddings for the single-cell model. **b.** UMAP embeddings for single-nucleus model. Same color codes as in a. **c.** UMAP embeddings for *Nucleus-cytosol model.*

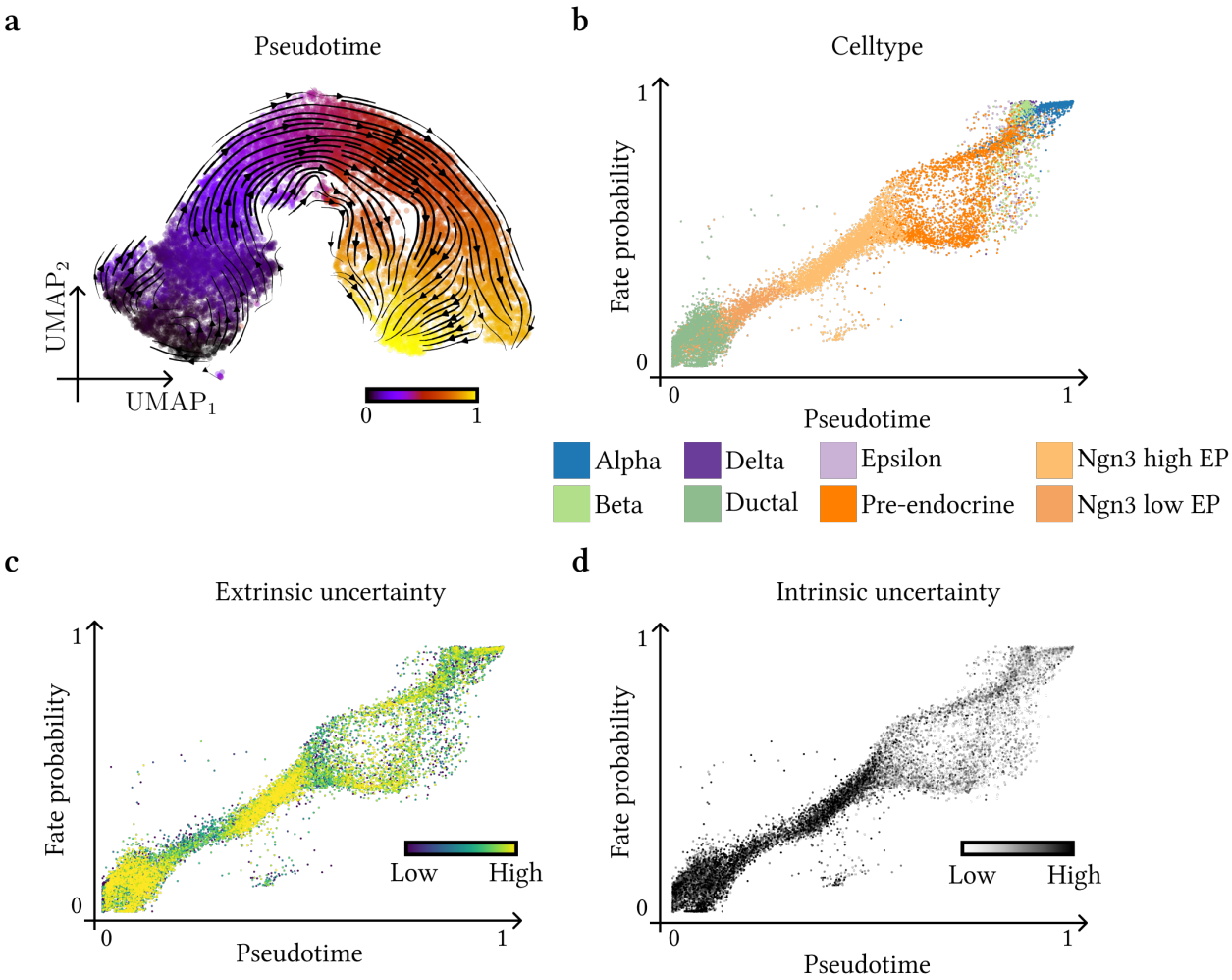

Figure 4.16 a. UMAP embedding of integrated pancreas E14.5 datasets [22][23] for the *Nucleus-cytosol model* colored by inferred pseudotime. **b.** Pseudotime against cell's fate probability colored by cell type. **c.** Same as **b.** but colored by intrinsic uncertainty. **d.** Same as **b.** but colored by extrinsic uncertainty.

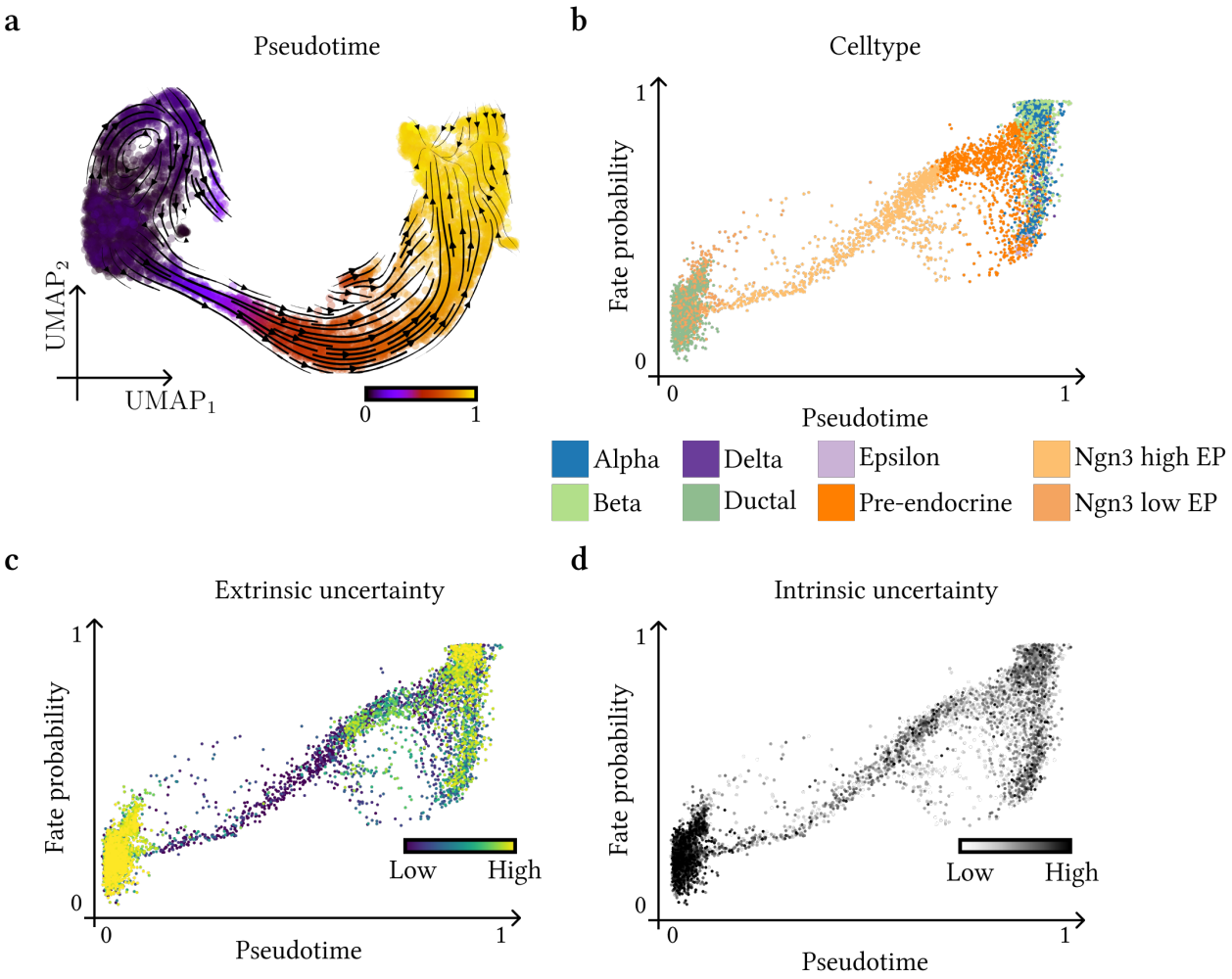

Figure 4.17 a. UMAP embedding of single-cell pancreas E14.5 dataset [22] colored by inferred pseudotime. **b.** Pseudotime against cell's fate probability colored by cell type. **c.** Same as **b.** but colored by intrinsic uncertainty. **d.** Same as **b.** but colored by extrinsic uncertainty.

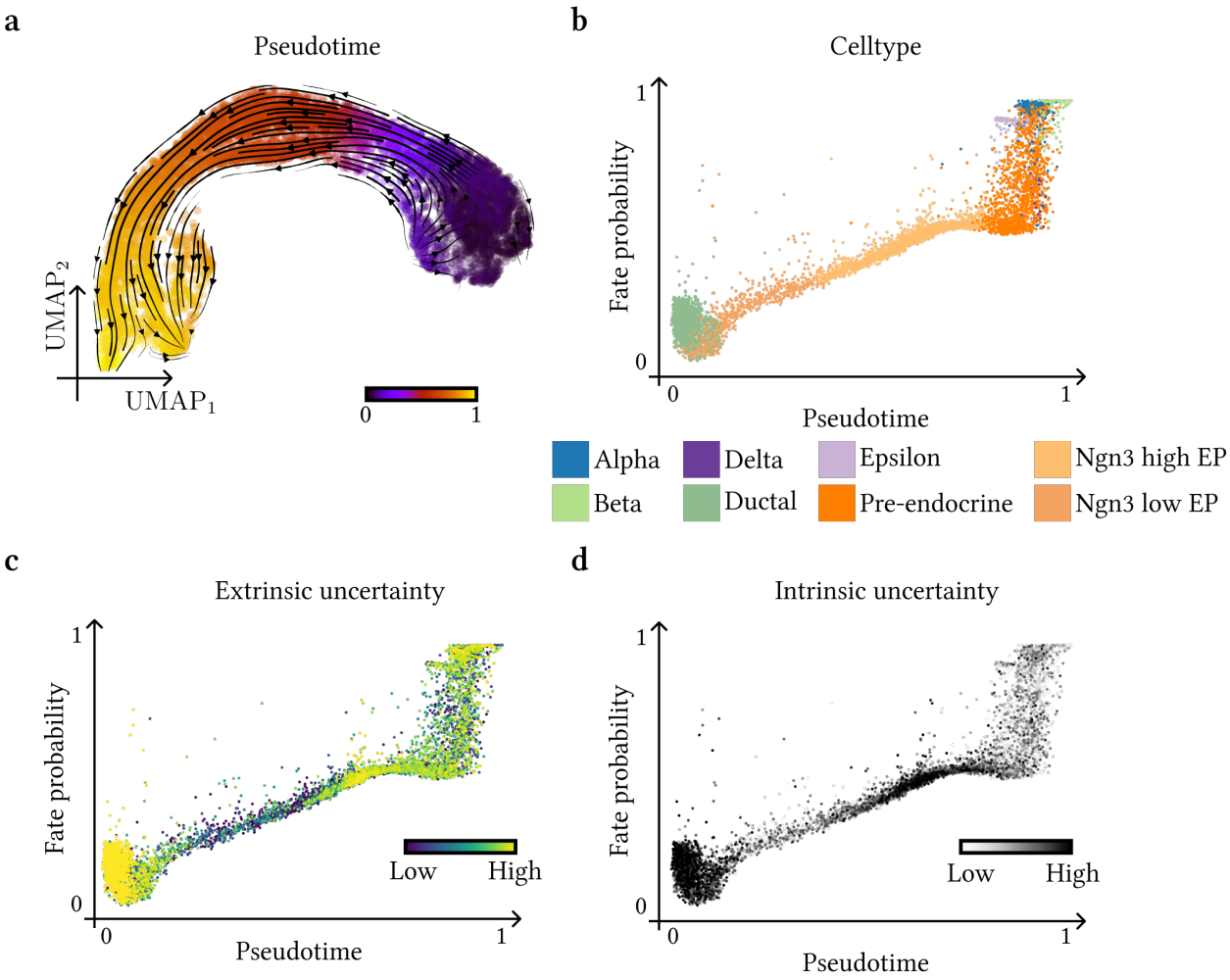

Figure 4.18 a. UMAP embedding of single-nucleus pancreas E14.5 dataset [23] colored by inferred pseudotime. **b.** Pseudotime against cell's fate probability colored by cell type. **c.** Same as **b.** but colored by intrinsic uncertainty. **d.** Same as **b.** but colored by extrinsic uncertainty.

5 Discussion

For eukaryotic cells, the transport of mRNA from the nucleus to the cytoplasm is one of the essential steps in the process of gene expression [33] and the underlying exporting rates are therefore of special interest. It is well known that most mRNA export is directly coupled to splicing [34]. However, nuclear export rates have rarely been measured [35][36][37].
Batich et al. have determined transcription, nuclear export, and, degradation rates for thousands of genes in Drosophila cells by using combined metabolic labeling of mRNA with cell fractionation and mathematical modeling [37]. Their findings indicate a connection between nuclear export and degradation rates, as well as between transcription and degradation rates, suggesting global coordination between mRNA degradation and both nuclear export and transcription [37].
To describe the biological process of mRNA metabolism, we here adapted the rate equations underlying the concept of RNA velocity by incorporating nuclear-exporting mechanisms. We therefore introduced a kinetic parameter describing the nuclear export of spliced RNA. By incorporating the nuclear export rate, we characterize the dynamics using a three-dimensional system of differential equations. Subsequently, the rate equations describe the change of unspliced and spliced nucleic RNA abundance as well as cytosolic spliced RNA abundance over time. Notably, a recent study has shown that the nuclear degradation of faulty or improperly spliced mRNAs is negligible [38], which supports our modeling approach.
As a first step, we solved the system of differential equations analytically and adapted the *veloVI* [14] framework accordingly. For the model extension, the *veloVI* framework has proven highly adaptable.
Subsequently, we evaluated if the *Nucleus-cytosol model* is able to infer kinetic parameters and latent variables, *i.e.* transcriptional states, cell-gene times, and switching times, of simulated data, *i.e.*, data following the modeled dynamics. We showed that the model infers rate parameter ratios that exhibit a strong correlation with the actual underlying parameters. Similarly, the time assignments yield high correlations with true underlying cell times. We further investigated genes where the inference did not work as well as for others. Here, we concluded that there are no clear indications for which genes the model fits the dynamics well or not. Moreover, for many genes that we identified as "outlier genes", the inference still aligned well with the underlying dynamics. Lastly, we investigated the impact of introduced noise. Here, the model has shown consistent results for most noise levels. The investigation involved the comparison of mean-squared errors and the comparison of inferred latent variables. We assessed the correlations between actual and estimated kinetic parameter ratios, between actual and estimated latent times, and between actual and estimated switching times. The investigation yielded consistent results across noise levels, especially for kinetic parameters and switching times. Since real-world sequencing transcripts do not follow the steps of the underlying data simulation process, the model's ability to handle simulated noise becomes a valuable attribute.
However, to apply the *Nucleus-cytosol model* on real-world sequencing data, we rely on having knowledge of the abundances in the respective part of a cell. Since single-cell experiments measure cellular abundances, spanning the nucleus and cytoplasm, we are missing information on the proportions of nucleic and cytosolic spliced RNA abundance. In contrast, single-nucleus experiments only measure nucleic RNA, therefore missing cytosolic spliced abundance. Subsequently, our model depends on the estimation of the abundances within the respective part of a cell as well as on the actual measured quantities. Therefore, we proposed a method to estimate the respective missing abundance on a batch-corrected neighbor graph.
Within the estimation process of cytosolic spliced mRNA abundance, as defined in equations (3.3) and (3.6), we assume that the single-nucleus protocols primarily capture nucleic RNA molecules and that we can disregard cytoplasmic RNA attached to the nuclei after their extraction. However, it is noteworthy that the physical isolation of nuclei introduces a complex distortion to the ratio between spliced and unspliced mRNA compared to single-cell sequencing pipelines [19]. This distortion results from a variety of

factors, including nuclear transport rates and the likelihood of these mRNA molecules being present in the residual cytoplasmic structures that remain attached to the outer surface of purified nuclei after their extraction, *i.e.*, to the rough endoplasmic reticulum. As a result, the relative abundance of spliced and unspliced mRNAs may undergo shifts and variations [19].
Therefore, we proposed the *lambda correction* in Section 3.4, which ensures that unspliced abundances of both measurements, *i.e.*, single-cell and nucleus, follow approximately the same count distribution and simultaneously scales measured and estimated spliced nucleic abundances. Within the pancreas E15.5 [22][23] datasets, single-nucleus observations yield higher unspliced counts per cell compared to single-cell observations and a similar level of spliced counts. Notably, single-nucleus spliced counts have shown to be significantly lower. However, omitting the lambda correction resulted in estimating substantial quantities of spliced cytosolic abundance as negative values. This observation further highlighted the necessity of the lambda correction as single-nucleus observations are expected to capture less spliced RNA molecules.
After final pre-processing and estimation, we validated the *Nucleus-cytosol model* on the pancreas datasets [22][23]. We started by assessing permutation scores and compared the resulting test statistics with the single-modal models. Here, we considered permutation score log ratios between two models for each gene and cell type. The overall permutation score distribution for common genes suggested that the *Nucleus-cytosol model* and the single-cell model have significantly higher permutation scores compared to the single-nucleus model. This observation indicated that the single-nucleus transcripts have less varying abundances within a cell type cluster and therefore less structure in the phase portraits. However, we noted that the permutation scores for the *Nucleus-cytosol model* can be elevated by nature due to its three-layer structure describing unspliced and spliced nucleic abundance as well as cytosolic spliced abundance.
Subsequently, we compared permutation effects on single genes. Here, we considered the genes *Top2a*, a cycling marker in the pancreas datasets [22][23], and the gene *Sulf2*, a marker of endocrine progenitor cells. While the permutation effect is similar for all models for the gene *Sulf2*, the models showed different permutation effects for *Top2a*. Here, the single-cell model, where induction and repression states for Ductal and Ngn3 low EP cells are fully observed as depicted in Figure 4.6c, exhibits the highest permutation score. In comparison, the phase portraits of the single-nucleus as well as of the *Nucleus-cytosol model* suggested that they do not align very closely with the inferred dynamics.
As a next step, we evaluated the inferred velocity vector field of the *Nucleus-cytosol model*. For the estimation process, we relied on constructing a batch-corrected latent space and therefore we here included a comparison between *scglue* [24] and *scVI* [25] based models. Both models yielded very similar velocity correlations to the single-modal models, and therefore for further analyses we only concluded the *scglue* based *Nucleus-cytosol model*. This choice is further supported by the higher overall scIB metrics as shown in Figure 3.3.
We then demonstrated that, in general, there are stronger velocity correlations between *Nucleus-cytosol model* and the single-cellin comparison to the correlations to the single-nucleus. In order to conduct a more detailed comparison of velocities between single-nucleus and single-cell data, we introduced and computed restricted velocity confidences. In this context, we assessed the alignment of velocities between a single-nucleus observation and its neighboring single-cell observations, and vice versa. The low confidence levels, compared to the confidence levels when considering the respective modality neighbors, further highlighted the need for a unified system providing reliable and closely aligned velocity estimates for both single-cell and nucleus observations.
Moreover, we investigated cell cycles within the cycling Ductal cell population in the pancreas datasets [22][23]. First, we assigned a cell cycle phase to each cell by computing G2M- and S-scores using the list of cell cycle genes [30] present in the pancreas datasets [22][23] after pre-processing. Then, for each cell cycle gene, we evaluated permutation scores. Here, the single-nucleus model revealed the lowest permutation effect, again indicating to exhibit less structure within the phase portraits. Therefore, we investigated phase portraits which confirmed that the assumptions of RNA velocity are violated by many cycling genes within the single-nucleus pancreas dataset [23]. We observed that many cells with very low spliced abundance have largely varying unspliced abundance, resulting in numerous cells being located on the very left side of the phase portraits. This finding can be attributed to the missing cytosolic RNA abundance, as

otherwise many cells would be shifted towards the right, *i.e.* the induction state could be more observed. However, as the *Nucleus-cytosol model* strongly depends on the single-nucleus transcripts, the phase portraits for the model oftentimes are very noisy, not following the inferred dynamics. Another contributing factor to the presence of noisy phase portraits is our approach to estimating the missing abundances by calculating a weighted average from neighboring cells' abundances. This approach has the potential to introduce additional noise.

To further investigate if the model resolves the underlying biological cell cycle transitions, we calculated a cell-cell transition matrix by using the functionalities of *Cellrank* [32]. Afterward, we defined and computed cell cycle phase transition probabilities. Here, we noticed that the single-modal models and the *Nucleus-cytosol model* equally well resolve the underlying cell cycle transitions, even though the different cell cycle phases are more clearly evident in the phase portraits of the single-cell pancreas dataset [22]. This finding underscored that *Cellrank* [32] provides reliable transition matrices.

Future cell cycle analysis could include the investigation of cell cycle phase assignments. Here, we relied on the standardized scores of mean expression levels of phase marker genes for the cell cycle assignments. One could, for example, use different cell cycle assignment methods as for instance proposed in [39][39][40] or [41] to name only a few. Furthermore, if there existed a cell-cycle dataset of fluorescent ubiquitination-based cell-cycle indicator (FUCCI) cells [42][43], but with the differentiation of nucleic and cytosolic RNA abundances, we could undertake a more in-depth examination of the model's ability to identify cell cycles.

Finally, we investigated intrinsic and extrinsic velocity uncertainties. We showed that the uncertainty measures can be used to identify regions where cells likely make decisions about their future state. For instance, the *Nucleus-cytosol model* exhibited elevated extrinsic uncertainties within Ductal and Ngn3 high endocrine progenitor cells compared to other cell types. The elevated uncertainty within initial Ductal cells can be attributed to the cycling population. Here, a cell might continue to cycle or it might exit the cell cycle, therefore yielding higher uncertainties. In contrast, the elevated uncertainty of Ngn3 high EP cells can be interpreted as a decision point of a cell whether it will transition towards terminal Alpha or Beta cells or not.

As we arranged cells according to their assigned pseudotimes, we observed that the intrinsic uncertainty was elevated for cells in the initial states. The elevated uncertainty persisted as we progressed along the pseudotime until we reached the Ngn3 high Ep cells. Subsequently, after the determination of future cell states, the uncertainty starts to decrease. It is noteworthy that the single-modal models revealed similar patterns within their respective uncertainty scales.

In summary, we proposed a model that integrates single-nucleus and single-cell transcripts into a unified system, describing the underlying biological process of mRNA splicing and exporting kinetics. Simultaneously, we have studied the applicability of RNA velocity on single-nucleus transcripts. Although many genes do not conform to the assumptions of RNA velocity, the single-nucleus model is still able to provide insightful information on the kinetics in pancreatic endocrinogenesis. Similarly, the *Nucleus-cytosol model* after estimating the respective missing abundances revealed noisy phase portraits. However, the inferred dynamics and velocity estimates still align well with the true underlying dynamics within the pancreas datasets [22][23]. Notably, the proposed estimation process automatically yields cellular abundances of single-nucleus observations. The cellular abundances could also be used to infer RNA velocity. The resulting velocity estimates will likely be well-aligned with the velocity estimates of single-cell observations.

Future perspectives based on this work include the verification and evaluation of the model on further real-world sequencing protocols. Here, the effect of using single-nucleus and single-cell transcripts from the same tissue can be investigated. Using cells from the same sample will likely increase the accuracy and also remove noise within the estimation process.

Furthermore, the model can be extended by integrating additional modalities such as chromatin accessibility. As open chromatin is a requirement for RNA to be transcribed, open chromatin profiles can be very useful to infer transcription rates. This extension results in a time-dependent transcriptional rate, similar to the *MultiVelo model* [8].

Moreover, it is possible to infer cell type-specific transcription rates. This modeling approach is justified by

real-world scenarios in which transcriptional bursts cause a rapid and sudden change in the transcription rates among different cell types.
Future perspectives could further include the exploration of alternative batch correction methods for estimating nucleic and cytosolic abundances, which can lead to phase portraits with reduced noise.
Finally, we note that the application of RNA velocity models on snRNA-seq data has shown promising results, as indicated by prior studies [44][45]. Therefore, the benchmarking of single-nucleus against single-cell transcripts can be extended by exploring additional metrics.

A Appendix

A.1 Evaluation pancreas E15.5 datasets

Here, we will report the same figures as presented in Section 4.2.

Construction of joint latent space and scIB metrics

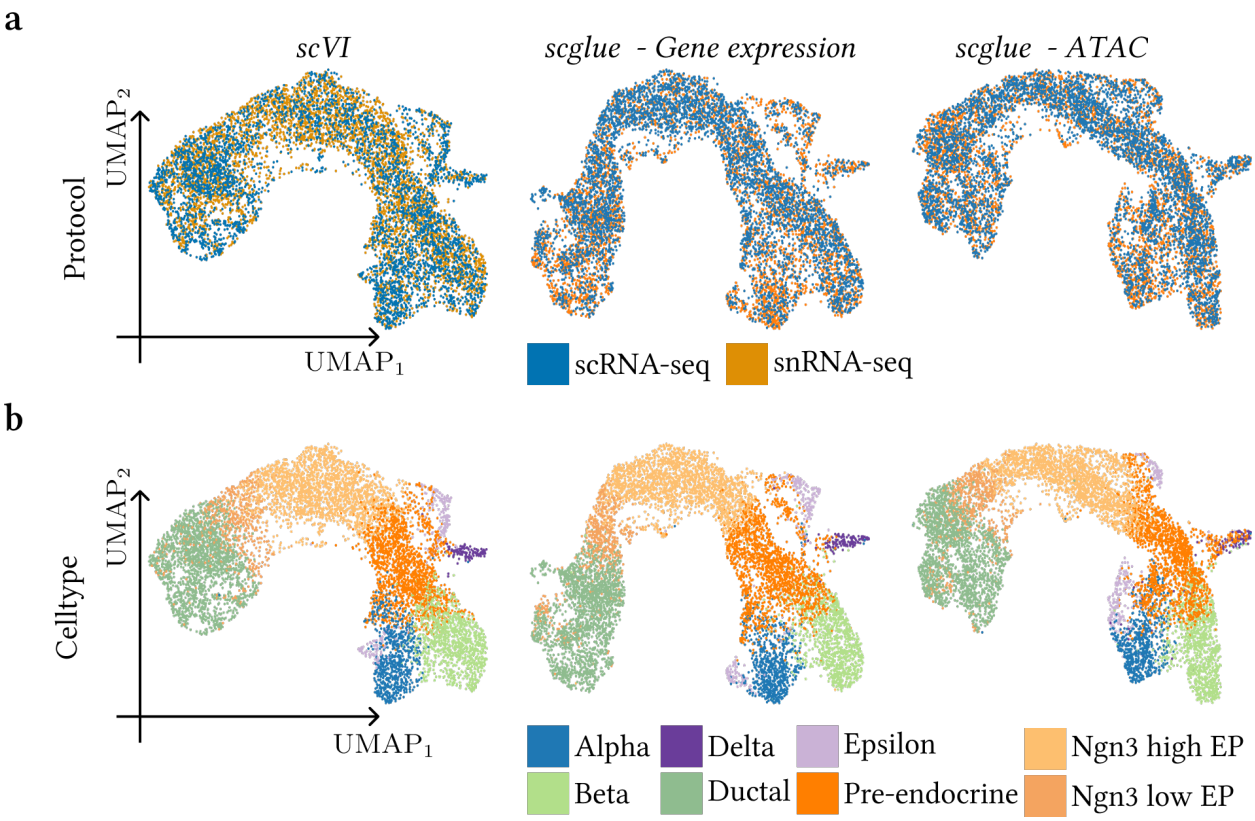

Figure A.1 Analogous to Figure 3.2. UMAP embeddings of pancreas E15.5 datasets [22][23]. **a.** Integrated UMAP embedding computed on batch-corrected latent spaces for the three different models colored by protocol. **b.** Same as **a**, but colored by cell type.

Figure A.2 Analogous to Figure 3.3. scIB metrics for integration of the pancreas E15.5 datasets: *X_glue_gex* refers to the latent embeddings generated by the *scglue* model trained on gene expression counts of both modalities; *X_glue_atac* refers to the latent embeddings generated by the *scglue* model trained on single-cell gene expression transcripts and chromatin accessibility profiles; *X_scVI* refers to the latent embeddings generated by the *scVI* model; *Unintegrated* refers to the none batch corrected PCA embeddings.

Permutation score analysis

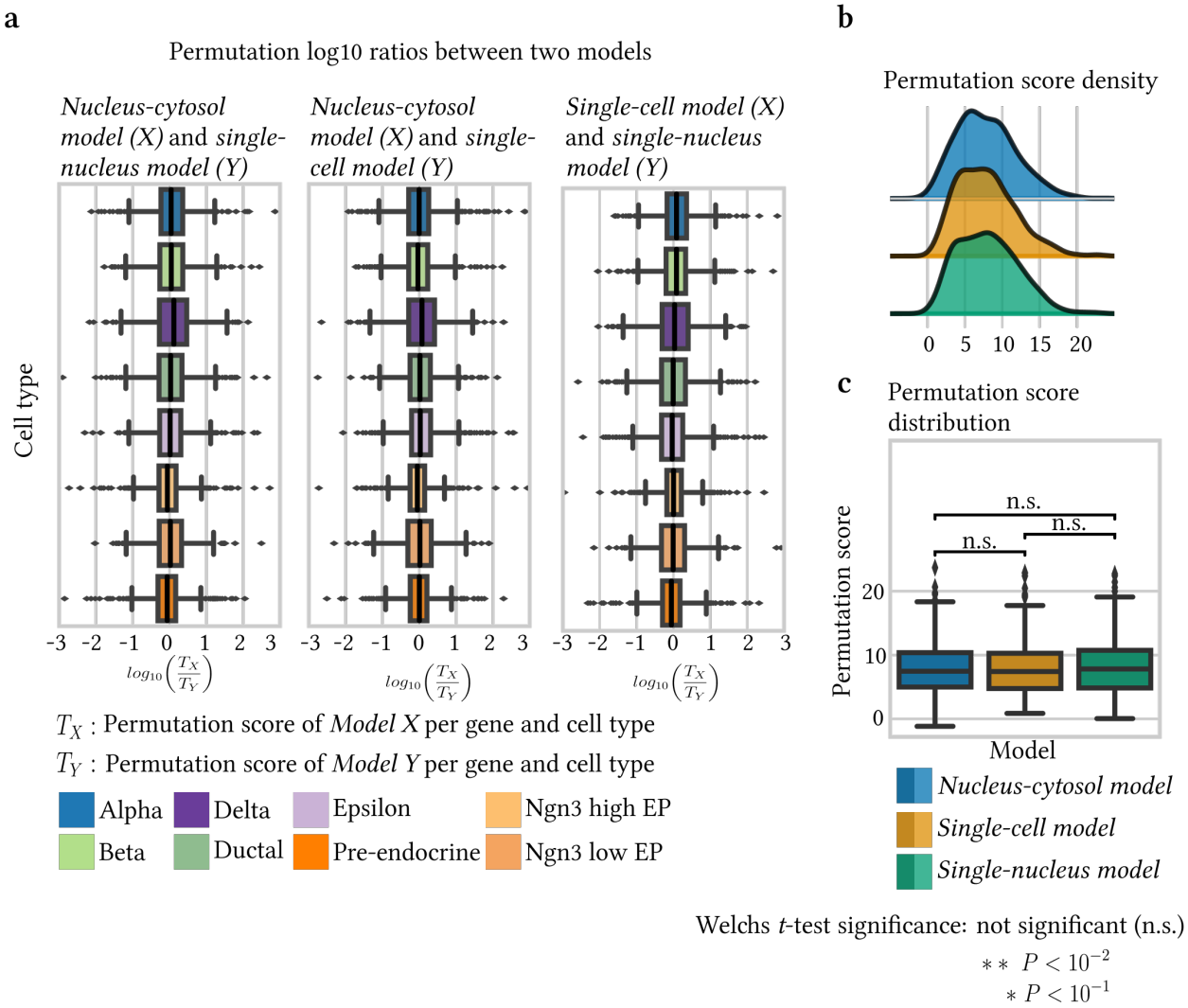

Figure A.3 Analogous to Figure 4.5 **a.** Permutation $\log_{10}$ ratios between two respective models colored by cell type for genes present in both respective datasets. Box plots indicate the median (center line), interquartile range (hinges), and whiskers at 1.5× interquartile range. **b.** Permutation score density per model. **c.** Permutation score distribution per model. Box plots indicate the median (center line), interquartile range (hinges), and whiskers at 1.5× interquartile range.

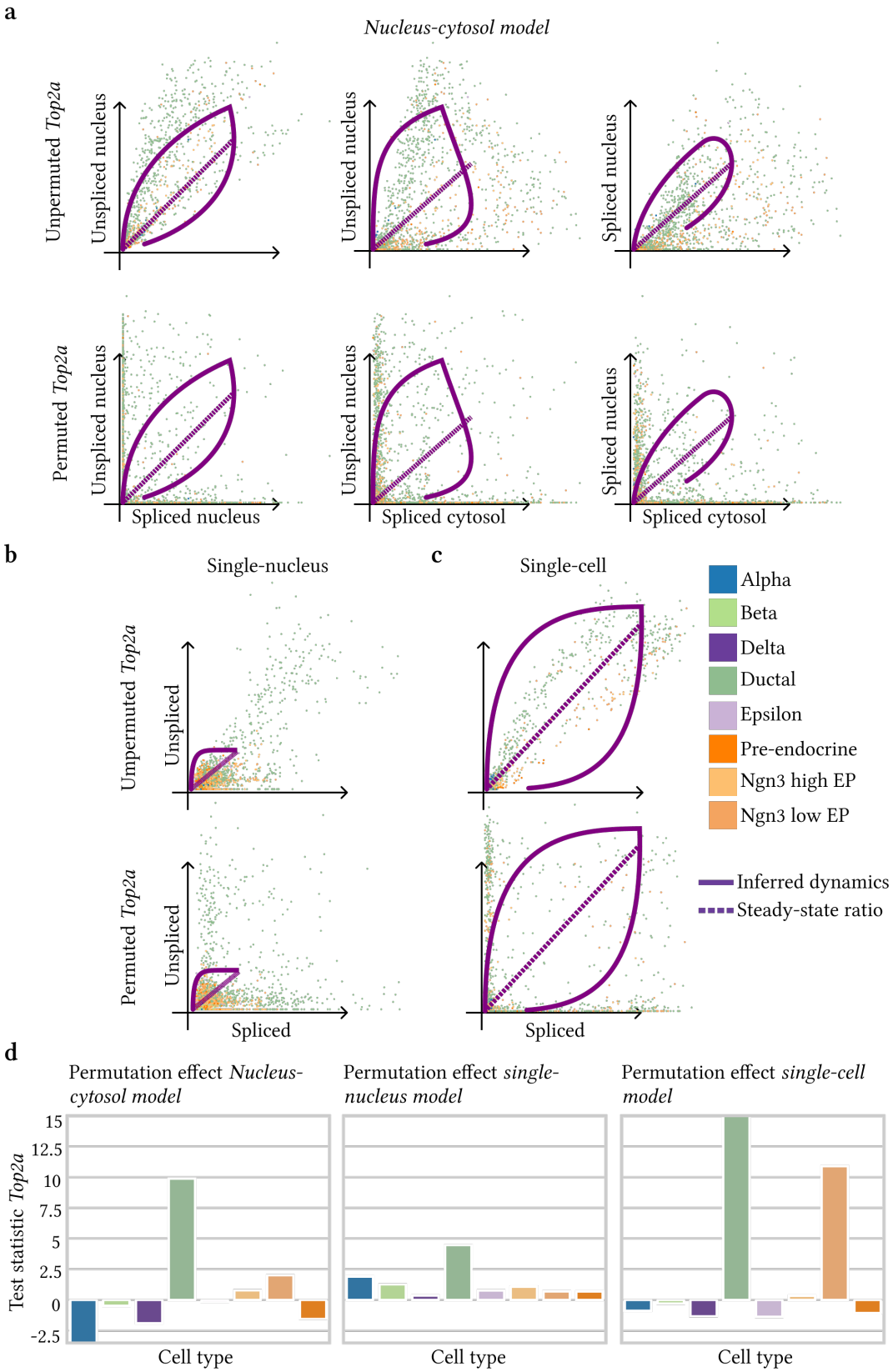

Figure A.4 Analogous to Figure 4.6. **a-c.** Unpermuted and permuted phase portraits of the gene *Top2a* for the *Nucleus-cytosol model* (**a**), the single-nucleus model (**b**), and the single-cell model (**c**). The purple curve represents the inferred dynamics and the dashed line represents the respective steady state ratios. **d.** Permutation scores for the gene *Top2a* per cell type for each model.

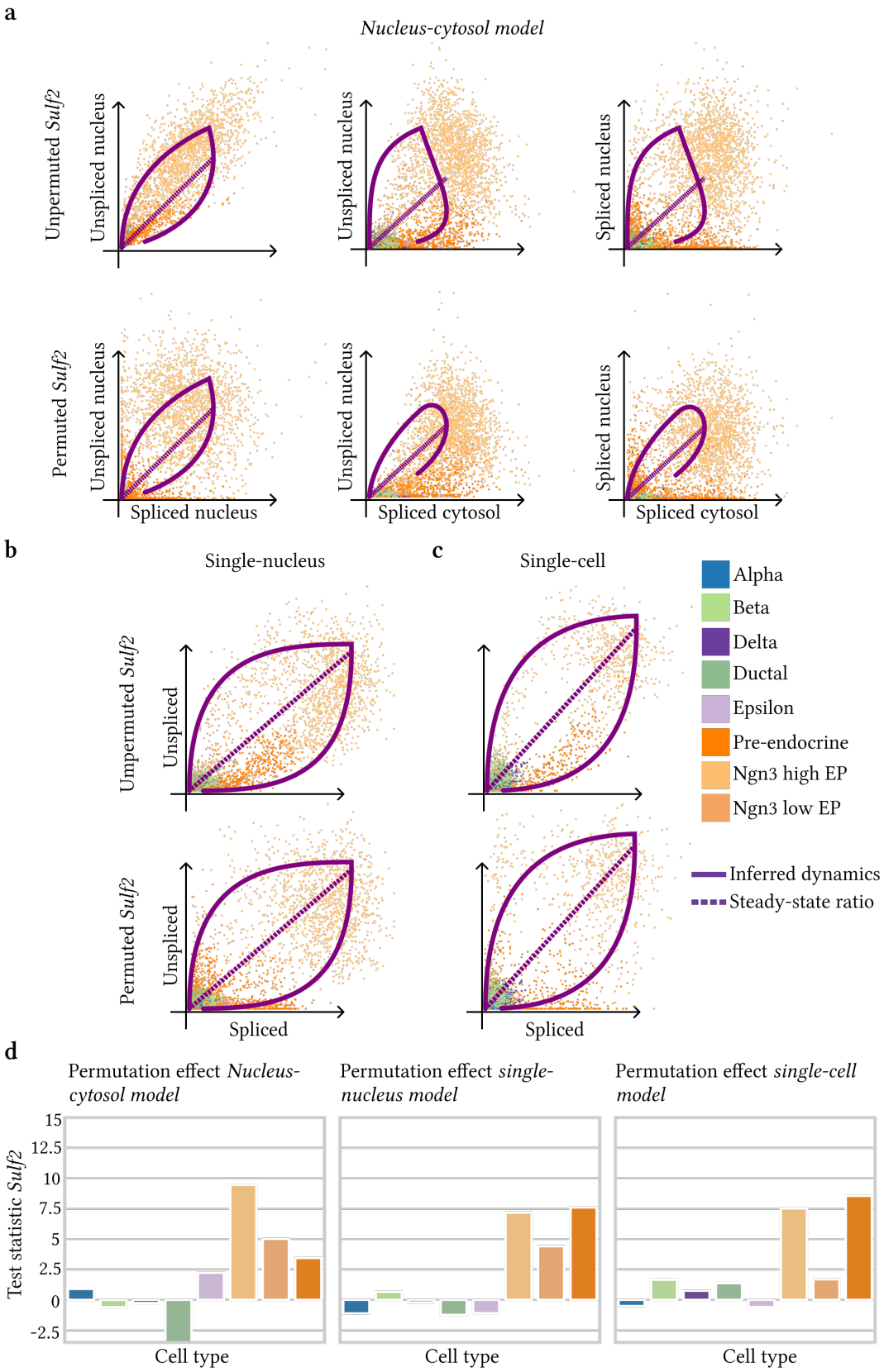

Figure A.5 Analogous to Figure 4.7. **a-c.** Unpermuted and permuted phase portraits of the gene *Sulf2* for the *Nucleus-cytosol model* (**a**), the single-nucleus model (**b**), and the single-cell model (**c**). The purple curve represents the inferred dynamics and the dashed line represents the respective steady state ratios. **d.** Permutation scores for the gene *Sulf2* per cell type for each model.

Velocity vector field comparison

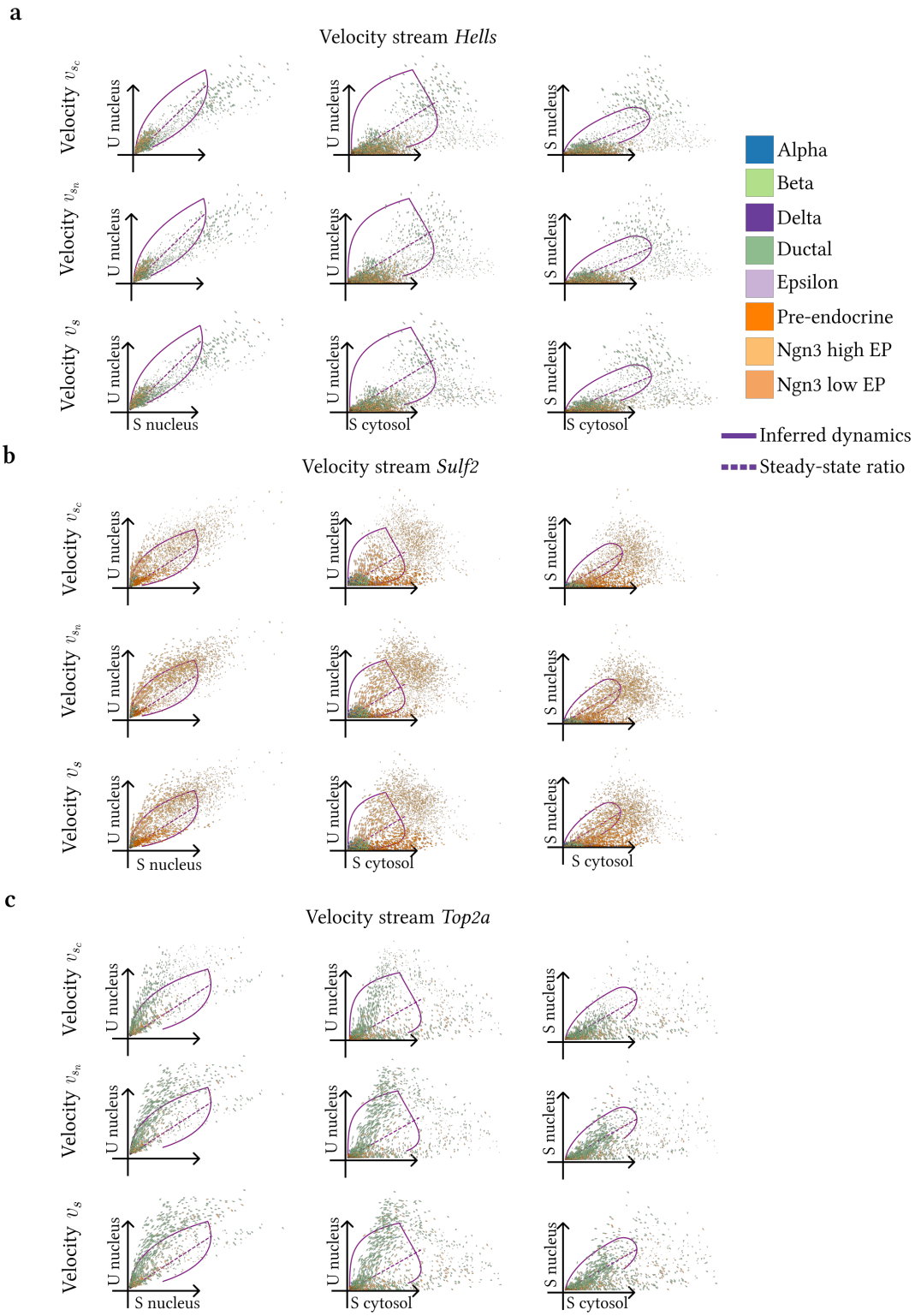

Figure A.6 Analogous to Figure 4.8. Phase portraits of *Nucleus-cytosol model* colored by cell type. Left column: Spliced nucleus against unspliced nucleus. Middle column: Spliced cytosol against unspliced nucleus. Right column: Spliced cytosol against spliced nucleus. **a.** Phase portraits for the gene *Hells* with velocity mode v_{s_c} on the top and v_{s_n} at the bottom. **b.** Same as **a.** but for the gene *Sufl2*. **c.** Same as **a.** but for the gene *Top2a*.

Velocity confidence

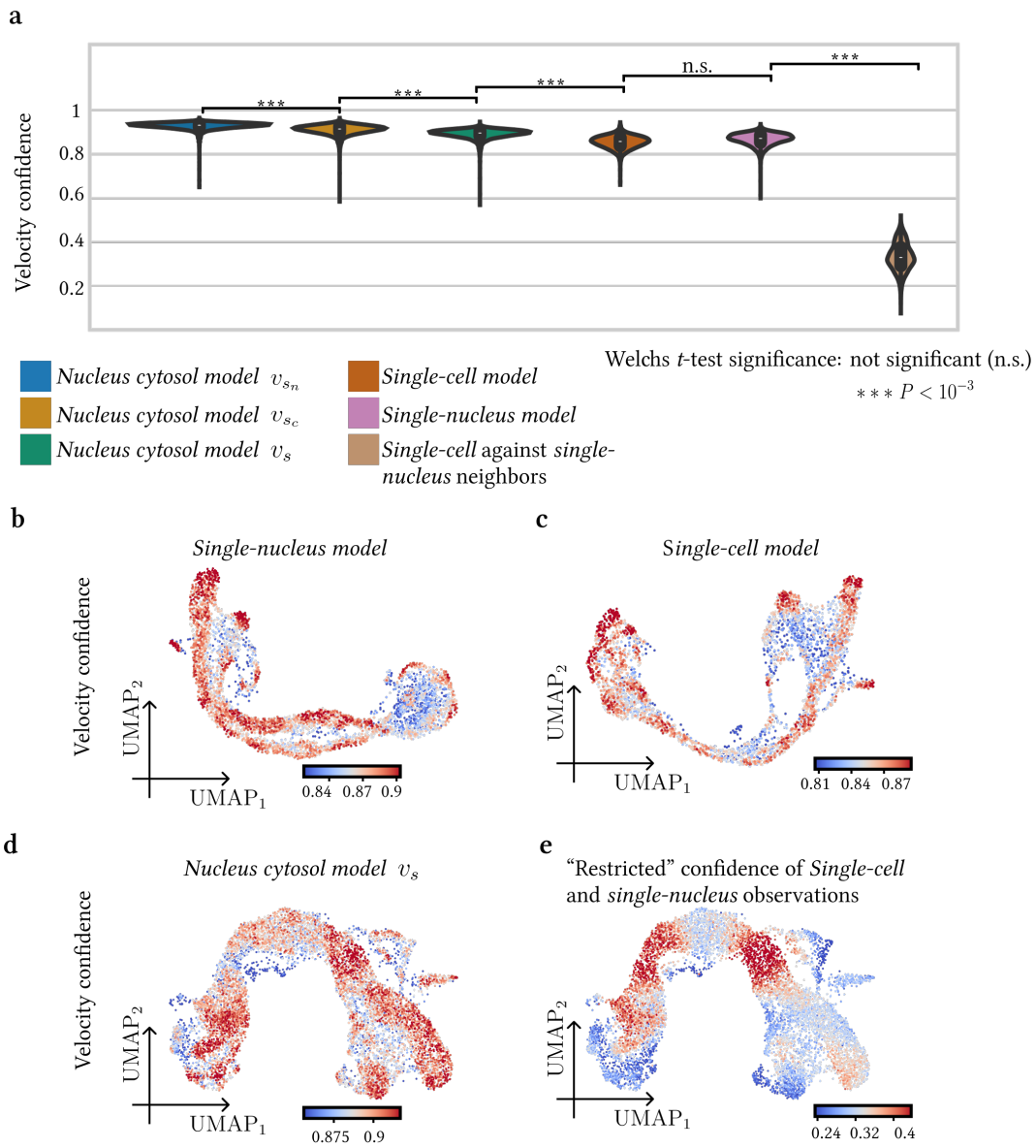

Figure A.7 Analogous to Figure 4.10. **a.** Velocity confidences for all models. **b.** Velocity confidence for single-nucleus model. **c.** Velocity confidence for the single-cell model. **d.** Velocity confidence for single-nucleus model for velocity mode v_s. **e.** Velocity confidence as described in equations (4.2) and (4.3).

Cell cycle analysis

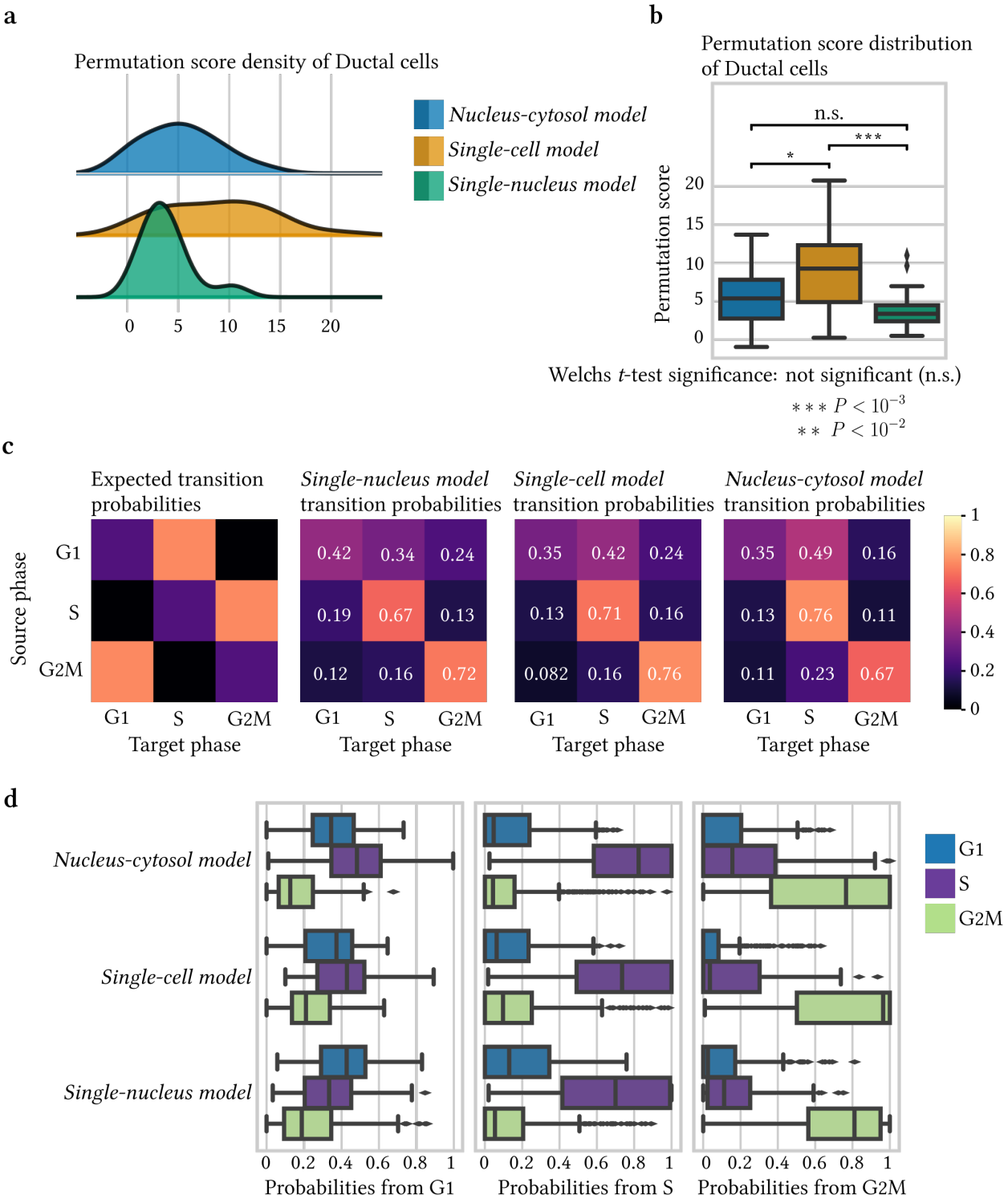

Figure A.8 Analogous to Figure 4.11. **a.** Permutation score densities on Ductal cells solely for cycling genes present within all three datasets after pre-processing. **b.** Permutation score distribution of Ductal cells per model. Box plots indicate the median (center line), interquartile range (hinges), and whiskers at 1.5× interquartile range. **c.** Expected cell cycle transition matrix and transition matrices for all models colored by probability. **d.** Transition probability distributions for cell cycle phases and trained models. Box plots indicate the median (center line), interquartile range (hinges), and whiskers at 1.5× interquartile range.

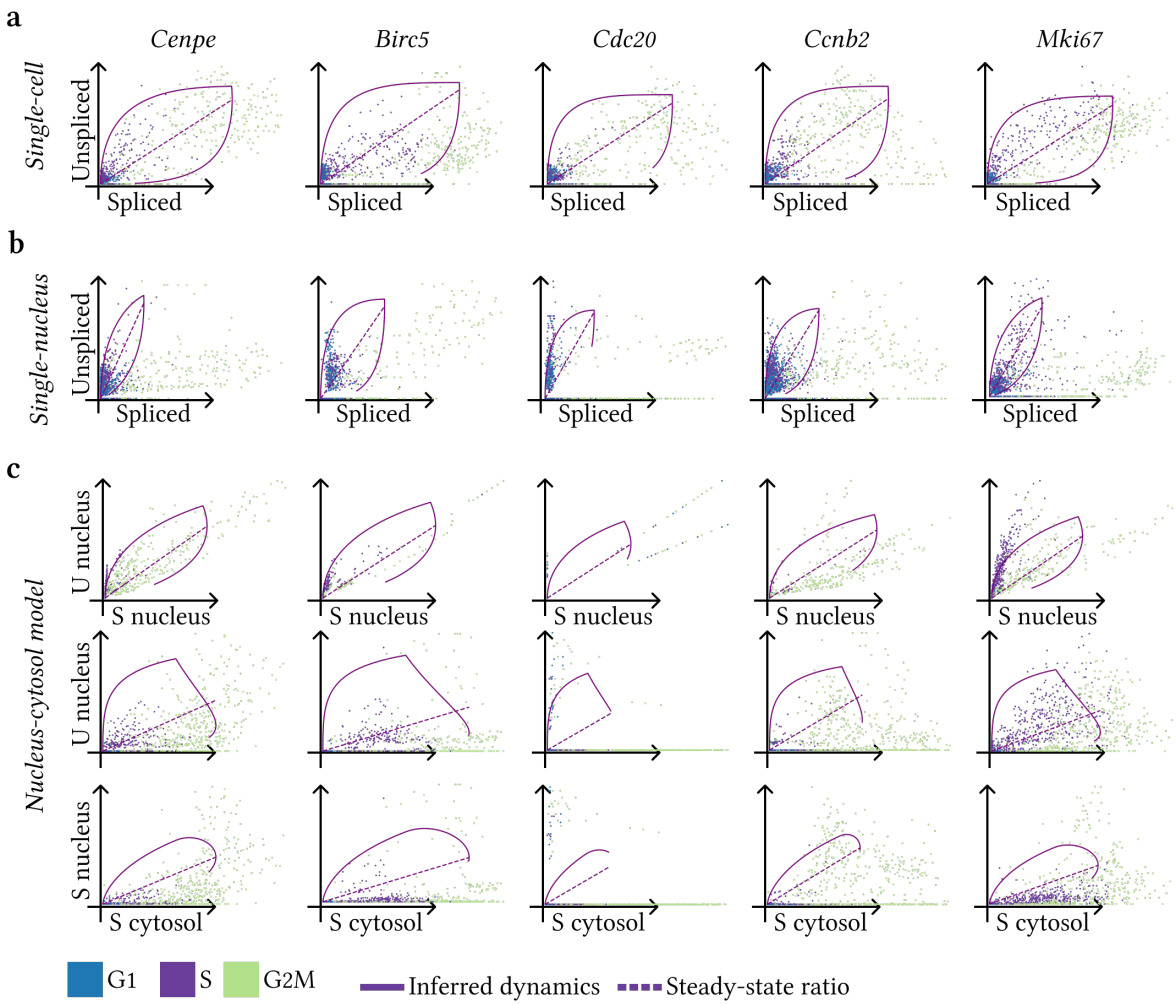

Figure A.9 Analogous to Figure 4.12. **a-c.** Single-cell (**a**), single-nucleus (**b**), and imputed phase portraits (**c**) for multiple cell cycle genes colored by cell cycle phase. We used the abbreviation "U" for "Unspliced" and "S" for "Spliced". The purple curve represents the inferred dynamics and the dashed line represents the respective steady state ratios.

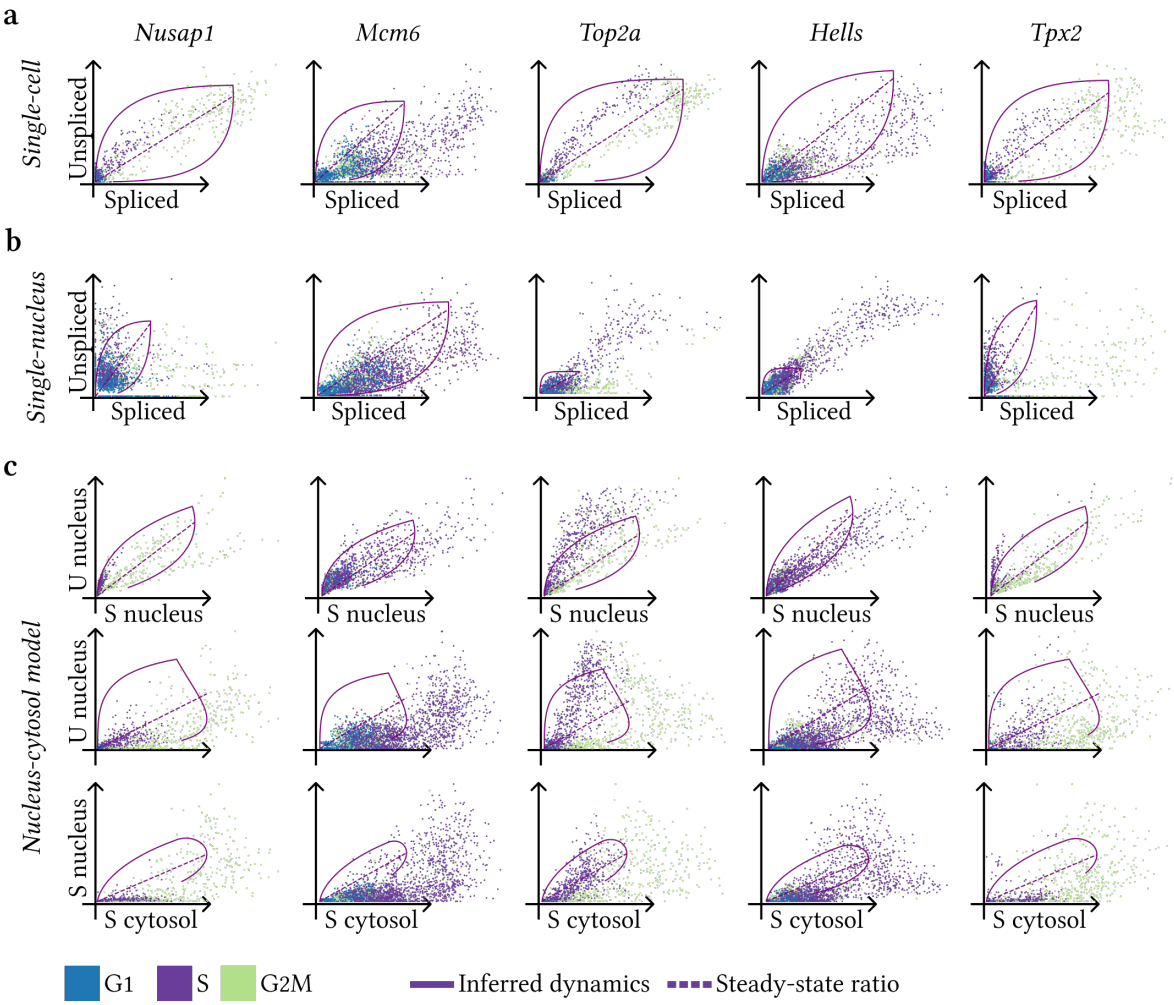

Figure A.10 Analogous to Figure 4.13. **a-c.** Single-cell (**a**), single-nucleus (**b**), and imputed phase portraits (**c**) for multiple cell cycle genes colored by cell cycle phase. We used the abbreviation "U" for "Unspliced" and "S" for "Spliced". The purple curve represents the inferred dynamics and the dashed line represents the respective steady state ratios.

Uncertainty analysis

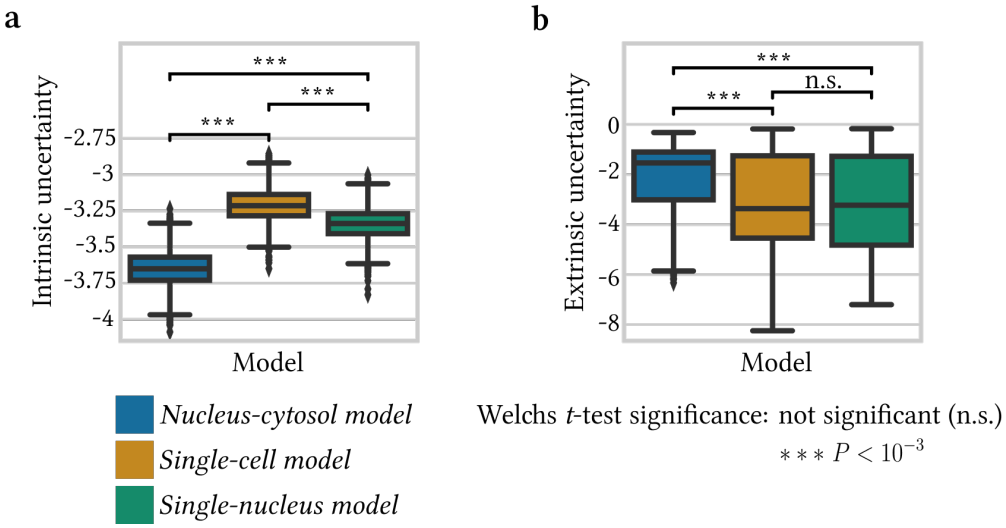

Figure A.11 Analogous to Figure 4.14. **a.** Comparison of intrinsic uncertainty measures between models. A high value corresponds to high uncertainty and a low value to low uncertainty. **b.** Comparison of extrinsic uncertainty measures between models. A high value corresponds to high uncertainty and a low value to low uncertainty.

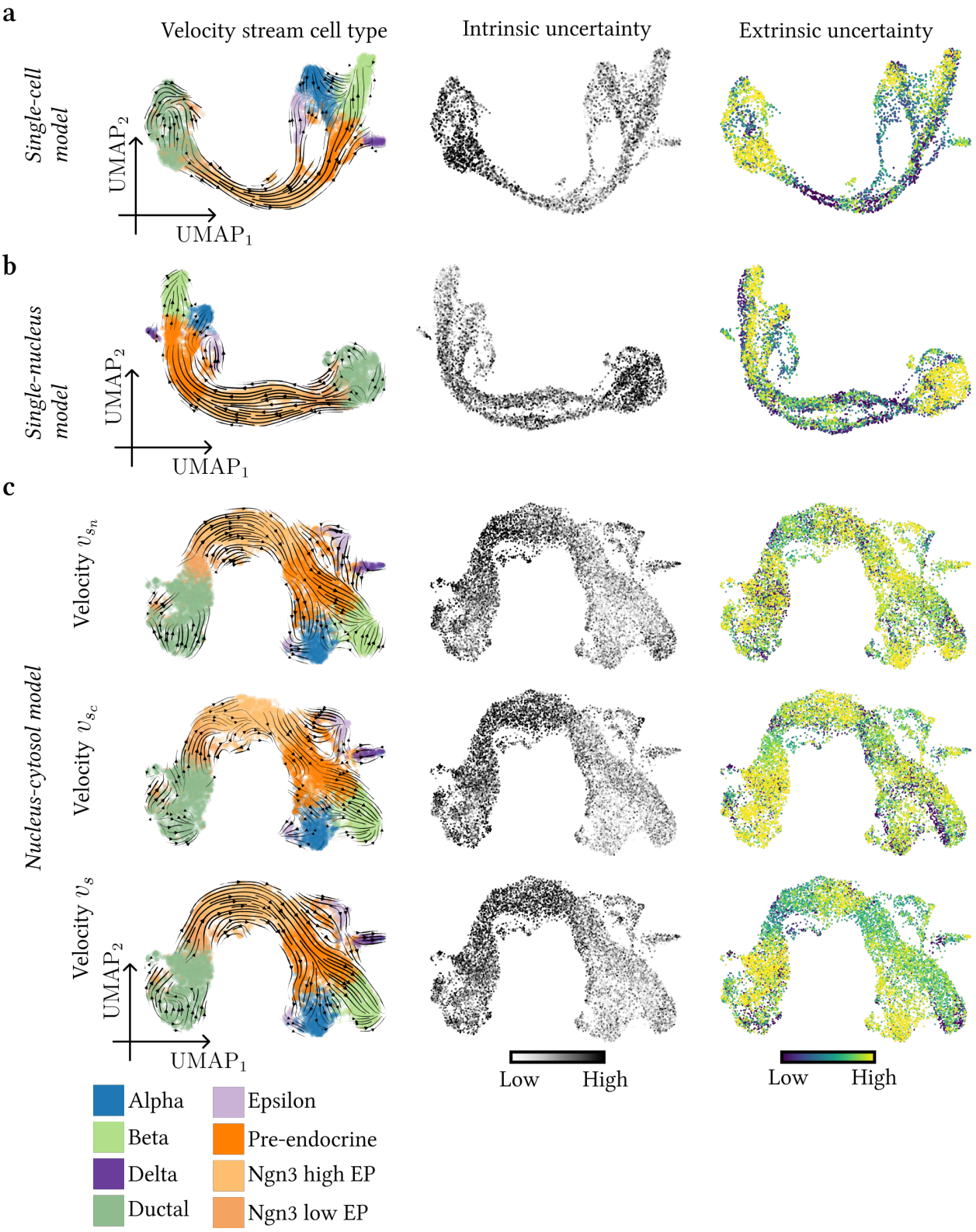

Figure A.12 Analogous to Figure 4.15. The left column shows the velocity stream colored by cell types. The middle column shows UMAP embeddings colored by intrinsic uncertainties. The right column shows UMAP embedding colored by extrinsic uncertainty. **a.** UMAP embeddings for the single-cell model. **b.** UMAP embeddings for single-nucleus model. Same color codes as in a. **c.** UMAP embeddings for *Nucleus-cytosol model.*

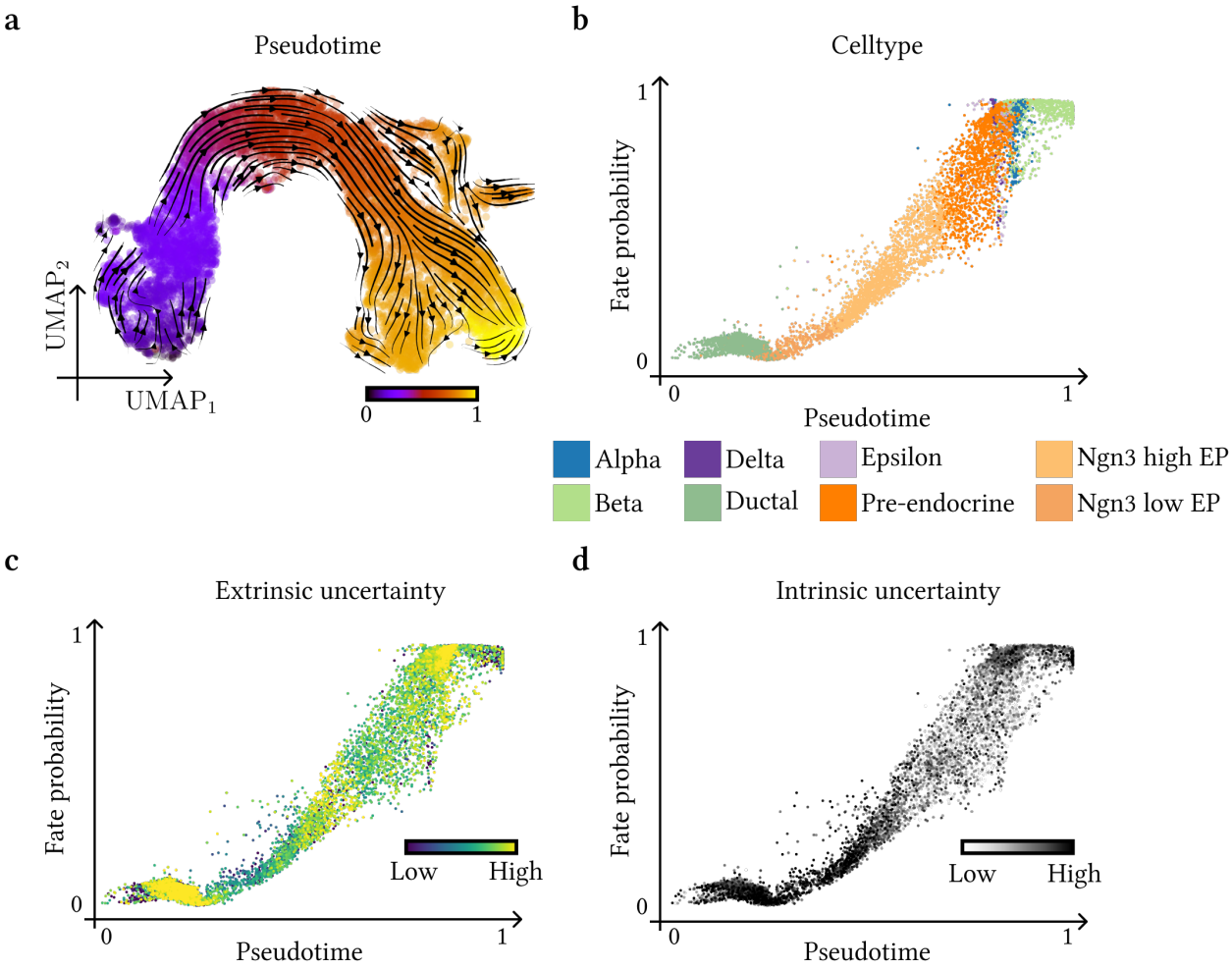

Figure A.13 Analogous to Figure 4.16. **a.** UMAP embedding of integrated pancreas E15.5 datasets [22][23] for the *Nucleus-cytosol model* colored by inferred pseudotime. **b.** Pseudotime against cell's fate probability colored by cell type. **c.** Same as **b.** but colored by intrinsic uncertainty. **d.** Same as **b.** but colored by extrinsic uncertainty.

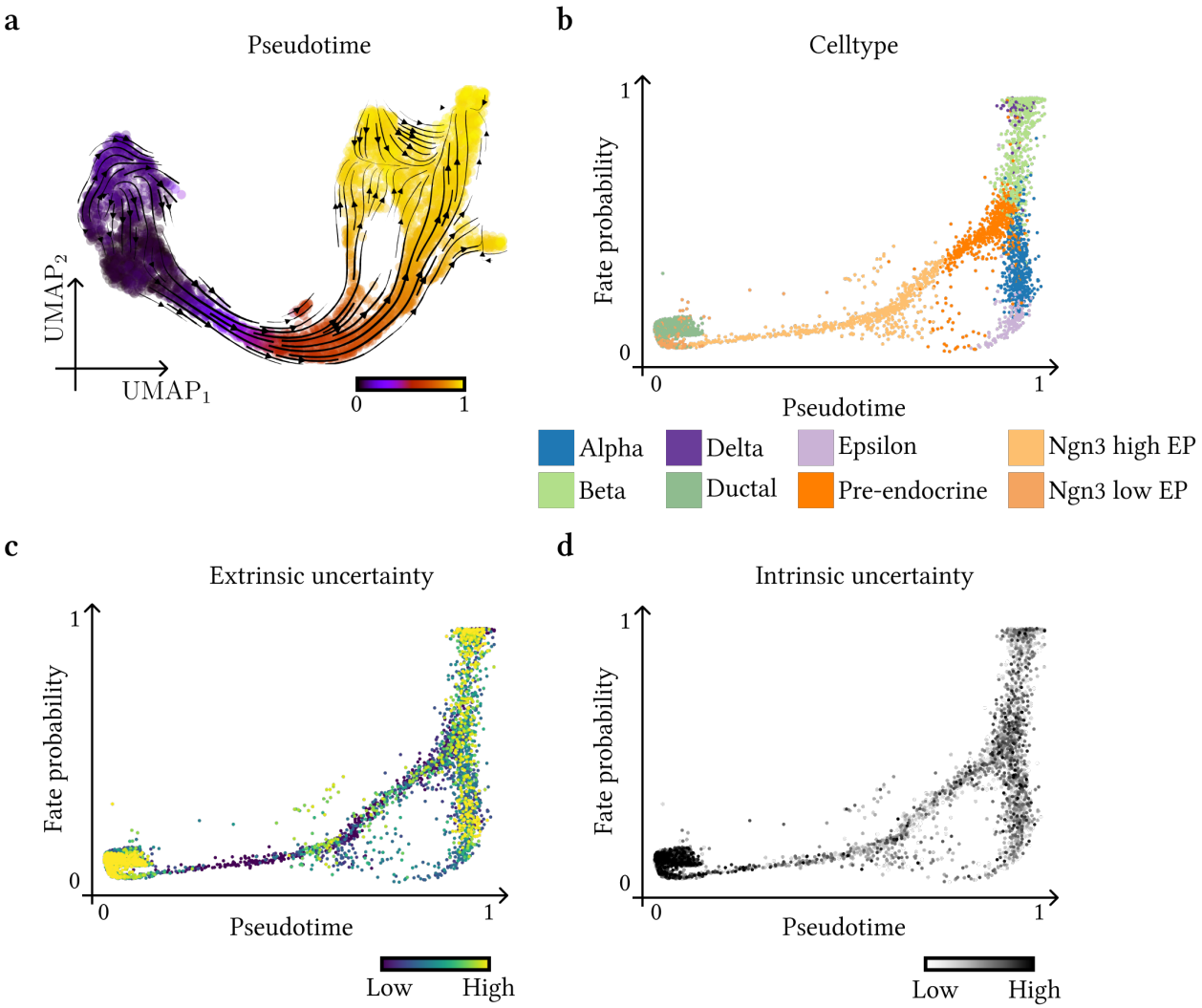

Figure A.14 Analogous to Figure 4.17. **a.** UMAP embedding of single-cell pancreas E15.5 dataset [22] colored by inferred pseudotime. **b.** Pseudotime against cell's fate probability colored by cell type. **c.** Same as **b.** but colored by intrinsic uncertainty. **d.** Same as **b.** but colored by extrinsic uncertainty.

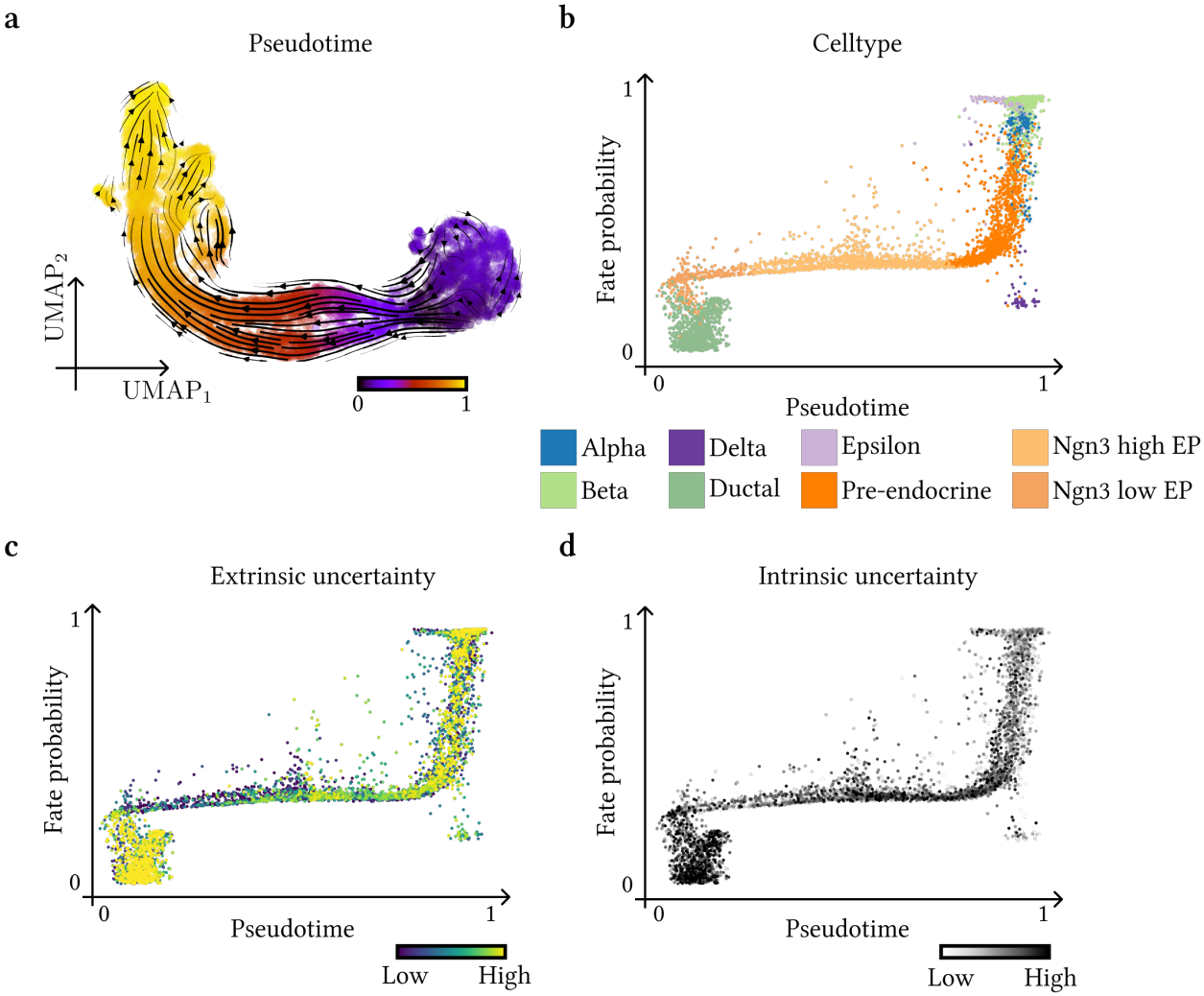

Figure A.15 Analogous to Figure 4.18. **a.** UMAP embedding of single-nucleus pancreas E15.5 dataset [23] colored by inferred pseudotime. **b.** Pseudotime against cell's fate probability colored by cell type. **c.** Same as **b.** but colored by intrinsic uncertainty. **d.** Same as **b.** but colored by extrinsic uncertainty.

A.2 Background

Eukaryotic transcripts are constructed from exons which are divided by longer non-coding sequences, referred to as introns. As a consequence, the coding sequence of the gene often only makes up a fraction of the total gene length. While both, introns and exons are transcribed into unspliced pre-mRNA, the introns and also some exons are later cut out through a process called splicing. The transcripts of many eukaryotic genes are spliced in multiple ways, providing different sets of mRNAs, depending on which exons were left in the final mRNA. This makes it possible to encode multiple proteins from just one gene [46]. Now being able to identify unspliced and spliced reads from scRNA-seq data allows the formulation of a dynamical model describing splicing kinetics [47]. Current models of RNA velocity assume the gene-specific model defined as

$$\begin{aligned} \frac{du_g(t)}{dt} &= \alpha_g - \beta_g u_g \\ \frac{ds_g(t)}{dt} &= \beta_g u_g - \gamma_g s_g, \end{aligned} \tag{A.1}$$

where u_g, s_g denote the unspliced and spliced RNA of the gene g respectively, a_g denotes the transcription rate, β_g the splicing rate and γ_g the degradation rate. The rate parameters can not be measured during sequencing and need to be inferred instead [11].

During a dynamic process, an increase in the transcription rate α_g first results in a rapid increase in unspliced mRNA, which is followed by a subsequent increase in spliced mRNA until a new steady state is reached [9]. RNA velocity is then defined as the derivative of spliced counts with respect to time. A gene is thereby up-regulated if it has a positive RNA velocity, which occurs for cells that have a higher abundance of unspliced mRNA than expected in a steady state for that gene. We refer to this state as the induction phase. The contrary holds for negative velocities, indicating that a gene is down-regulated, which we will refer to as the repression phase. Given the velocities of the genes in a cell, the future state of an individual cell can be estimated [10].

A.2.1 Steady-State-Model

The first model of RNA velocity, which we will refer to as *Steady-State-Model*, introduced by Manno et. al. [9] made simplifying assumptions of constant rates of transcription α_g and degradation γ_g of mRNA, a single, global splicing rate, chosen as $\beta = 1$, that the cellular dynamics reached an equilibrium in the induction phase, and gene-wise independence [14]. Using the assumption that steady states have been reached it holds for cells in steady states, that

$$\frac{ds_g(t)}{dt} = 0 \Leftrightarrow \beta_g u_g - \gamma_g s_g = 0 \Leftrightarrow \frac{u_g}{s_g} = \frac{\gamma_g}{\beta_g} =: \gamma_g' \Leftrightarrow u_g = \gamma_g' s_g,$$

resulting in a steady-state ratio γ_g' of unspliced to spliced mRNA, indicating where mRNA synthesis and degradation are in balance. Since steady states are expected at the lower and upper quantiles in phase space, we can estimate the steady-state ratio through an extreme quantile linear regression fit in closed form using the least squares loss as [10] [11]

$$\gamma_g' = \frac{u_g^T s_g}{\|s_g\|}.$$

The vectors $u_g, s_g \in \mathbb{R}^n$ represent size-normalized unspliced and spliced counts for a particular gene g for cells which lie in the lower or upper extreme quantile, thus $n \in \mathbb{N}$ is only a fraction of the total number of cells [10]. The estimated RNA velocity for a cell c and a given gene g is then defined as the residual of the unspliced mRNA to the linear regression fit

$$v_{g,c} = u_{g,c} - \gamma_g' s_{g,c}.$$

This model requires the steady-states to be observed in the data, which might not be the case if the induction phase terminates before the mRNA-level saturates [10]. The assumption of a common constant splicing rate across genes might not reflect the true nature of the splicing dynamics. Further, only using a subset of the available data is an additional drawback.

A.2.2 EM Model

The *EM model* does not assume that steady states are observed and it introduces a gene-dependent splicing rate β_g, overcoming the two limiting assumptions of the steady state model. It further introduces a state-dependent transcription rate $\alpha^{(k)}$ for states $k \in \{1, 2, 3, 4\}$ (induction, induction steady state, repression, repression steady-state). The splicing dynamics are then solved in a likelihood-based EM (expectation-maximization) framework, by iteratively estimating the parameters of reaction rates and the latent cell-specific variables. Within the expectation step the hidden variables k and t (the transcriptional state and cell-internal latent time) are updated, and the maximization step consists of updating the rate parameters ($\theta = \{\alpha^{(k)}, \beta, \gamma\}$). The model considers the analytical solution to the system given in A.1, which can again be found with Theorem A.3.1 and A.3.2. As rate parameters are time-independent (constant across time), a positive transcription rate $\alpha > 0$ is assumed for the induction phase, while during the repression phase it is set to 0. Further, assigning an optimal time to each cell-gene pair is computationally expensive. Therefore the optimal solution can be approximated, leading to a 30-fold speedup in computational time [10].

The steady-state, as well as the EM model, lack a notion of uncertainty, thus making it difficult to evaluate the robustness of the RNA velocity estimate. In the case of dentate gyrus neurogenesis, the visualization of RNA velocity suggests that Granule mature cells develop into their immature counterparts, even though the EM model returns genes with a high likelihood suggesting the RNA velocity pointing into the reverse and correct direction [14].

A.2.3 Single-cell variational inference - scVI

scVI models the observed gene expression for each gene g and cell i as a zero-inflated negative binomial (ZINB) distribution $p(x_{ig}|z_i, s_i, l_i)$ conditioned on the observed batch annotation s_i and two unobserved latent variables [25]. The first hidden variable l_i is a one-dimensional Gaussian used to account for nuisance variation arising from differences in capture efficiency and sequencing depth. It functions as a cell-specific scaling factor. The second variable, z_i, is a low-dimensional Gaussian, that captures the residual variation, primarily associated with biological distinctions among cells. We have used a latent dimension of $n = 10$. Within the *scVI* model, a neural network is employed to link the latent variables to the parameters of the ZINB distribution. This connection involves intermediary values denoted as p_i^g, which furnish a batch-corrected and normalized estimation of the proportion of transcripts from each gene g within each cell i [25]. To approximate the posterior distribution of the latent variables $q(z_i, \log l_i|x_i, s_i)$, a separate neural network is trained using variational inference along with an efficient stochastic optimization procedure, as proposed by Kingma et. al [20]. The architecture of the *scVI* model manifests as a variational autoencoder.

A.2.4 Single-cell graph linked unified embedding - scglue

scglue is able to integrate multiple omics layers into one system. It thereby models cell states as low-dimensional cell embeddings learned via variational autoencoders [20]. Due to the inherent differences in biological characteristics of distinct omics layers, each layer is equipped with a dedicated autoencoder that uses a probabilistic generative model tailored to its specific feature space [24].
Building on existing biological knowledge, the approach involves the utilization of a knowledge-based graph, referred to as the "guidance graph". This graph explicitly captures cross-layer regulatory interactions to connect layer-specific feature spaces. In this graph, vertices represent features from various omics layers, while edges signify signed regulatory relationships. For instance, when integrating scRNA-seq and scATAC-seq data, the vertices represent genes and accessible chromatin regions, *i.e.*, ATAC peaks, and

positive edges may link an accessible region to its potential downstream gene. It is further used to link the cell embeddings learned for different omics layers as they could have inconsistent semantic meanings as each embedding is generated by distinct autoencoders.
Subsequently, an adversarial multi-modal alignment of the cells is executed as an iterative optimization process. This alignment is guided by feature embeddings derived from the graph [24]. To verify if the model provides a reliable integration of the modalities, Cao et al. presented a metric called *consistency score*. This metric measures the consistency between the integrated multi-omics space and the prior knowledge represented in the guidance graph [24].
The calculation involves several steps: (1) Initially, cells from all omics layers are clustered together in the aligned cell embedding space using the k-means clustering algorithm; (2) within each omics layer, the cells within each cluster are grouped into a metacell. These metacells are then paired up, forming sample pairs; (3) feature correlation is computed using these paired metacells as samples. Specifically, Spearman's correlation is calculated for each edge in the guidance graph; (4) finally, the integration consistency score is determined by averaging the correlation values across all edges in the graph. Each correlation value is negated according to the sign of the respective edge and weighted by the edge weight [24].
When integrating data from dissimilar tissues, significantly lower consistency scores were reported, often approaching 0, in contrast to the integration of data within the same tissue. This finding confirms the metric's credibility as an indicator of integration quality, as outlined in [24]. It is noted that, backed by empirical evidence, it is reasonable to consider the integration as reliable when the curve consistently stays above the 0.05 threshold [24]. However, since edges in the guidance graph represent various features, such as genes for gene expression data and ATAC peaks for scATAC data, this functionality is primarily intended for omics layers that comprise different features. This is particularly relevant because single-cell and single-nucleus data have numerous genes in common, resulting in shared edges within the guidance graph. The functionality to calculate consistency scores for omics layers with shared features is not yet implemented within the package. Thus we have reported the consistency scores for varying numbers of metacells for the integrated ATAC and single-cell measurements in Figure A.16. Given that the curve consistently remains above the threshold of 0.05, we can confidently assert that the model provides a reliable integration of both modalities.

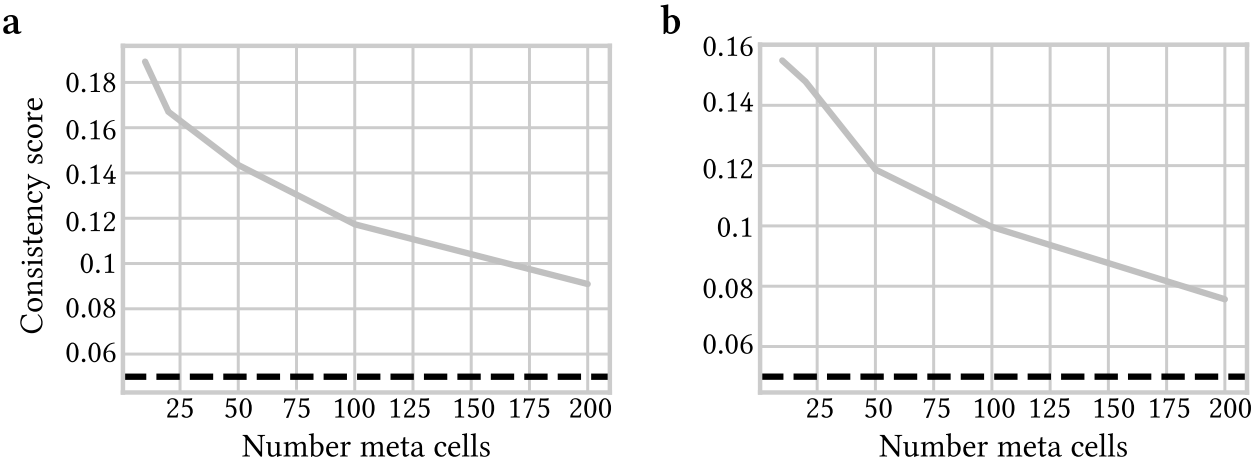

Figure A.16 a-b. Consistency score for the integration of the pancreas E14.5 (**a**) and E15.5 (**b**) [22][23] datasets. The black dashed line represents the consistency threshold of 0.05.

A.2.5 scIB metrics

Luecken et. al have categorized the metrics into two main groups: (1) metrics related to batch effect removal and (2) metrics related to the preservation of biological variation [28]. In the following, these metrics are shortly described.
NMI - Normalized mutual information

Normalized Mutual Information (NMI) is used to assess the alignment between cell type labels and Louvain clusters computed on the integrated dataset. The NMI values fall within the range of 0 to 1, where 0 signifies no correlation between the clusters, and 1 represents a perfect match. For optimization, Louvain clustering at a range of resolutions from 0.1 to 2 in steps of 0.1 is performed and the clustering output with the highest NMI is selected when compared to the label set [28].

ARI - Adjusted Rand Index

The adjusted Rand index (ARI) assesses the agreement between two clusterings by considering both correct overlaps and correct disagreements. Like NMI, the ARI is used to compare cell-type labels with the NMI-optimized Louvain clustering generated from the integrated dataset. The Adjusted Rand Index (ARI) corrects for random labeling, and an ARI score of 0 signifies random labeling, while a score of 1 denotes a perfect match [28].

ASW- Average silhouette width

The average silhouette width (ASW) quantifies the relationship between a cell's within-cluster distances and its between-cluster distances to the nearest cluster. Calculating the average silhouette widths for a group of cells, yields the ASW score, which falls within the range of -1 and 1. ASW is frequently employed to assess cluster separation: a value of 1 indicates tightly packed and well-separated clusters, while 0 or -1 suggests overlapping clusters (resulting from equal within- and between-cluster variability) or significant misclassification (stemming from greater within-cluster than between-cluster variability), respectively [28].

Isolated labels (ASW)

The Isolated label score is used to assess the performance of data integration methods when handling cell identity labels shared by only a small number of batches. More precisely, isolated cell labels were identified as the those found in the fewest number of batches during the integration process. The score quantifies the effectiveness of separating these isolated labels from other cell identities using the using the average-width silhouette score [28].

Graph LISI - Local Inverse Simpson's Index

The Local Inverse Simpson's Index (LISI) is a diversity metric designed to evaluate both batch mixing (referred to as iLISI) and the separation of different cell types (known as cLISI). These scores are calculated based on neighborhood lists for each node within integrated k-nearest neighbor graphs. To be more specific, the inverse Simpson's index is employed to determine how many cells can be selected from a neighbor list before encountering a cell from the same batch again. Consequently, LISI scores fall within the range of 1 to N, with N representing the total number of batches present in the dataset.

kBET - k-nearest neighbour Batch Effect Test

The kBET algorithm assesses whether the composition of cell labels within the k nearest neighbors of a cell aligns with the expected global label composition. This test is conducted iteratively on a random subset of cells, and the outcomes are consolidated into a rejection rate, considering all examined neighborhoods. kBET operates within the context of a k-nearest neighbor graph [28].

Graph connectivity

The graph connectivity metric evaluates whether the k-nearest neighbor graph representation, denoted as G, for the integrated data establishes direct connections between all cells sharing the same cell identity label, *e.g.* the same cell-type [28].

PCR - Principal component regression

Principal component regression, which is derived from Principal Component Analysis (PCA), has been employed to quantify batch effect removal [48]. In a nutshell, this method involves calculating R^2 values through linear regression of the covariate of interest, such as the batch variable B, against each principal component (PC). The contribution of batch effect variance for each principal component is then determined by multiplying the variance explained by the respective principal component PC_i with the corresponding R-squared value $R^2(PC_i|B)$ [28].

Aggregated scores

The aggregated bio conservation and batch correction metrics are the simple average over all calculated metrics belonging to the respective group. Finally, the overall aggregated score is calculated by taking

the sum over both aggregated scores and weighting the aggregated batch score by 0.6 and weighting the aggregated bio conservation score by 0.4 [28].

A.3 Mathemtical background

Theorem A.3.1. *Let $A = QDQ^{-1}$ be the eigenvalue-decomposition of a matrix $A \in \mathbb{R}^{n\times n}$. Then it holds, that*

$$e^A = e^{QDQ^{-1}} = Qe^DQ^{-1} = Qdiag(e^{d_{11}}, ..., e^{d_{nn}})Q^{-1},$$

where $diag(e^{d_{11}}, ..., e^{d_{nn}})$ refers to the diagonal matrix with $e^{d_{ii}}$ on the diagonal.

Theorem A.3.2. *The general solution for $A \in \mathbb{R}^{n\times n}$, $f(t) = (f_1(t), ..., f_n(t))^T$ a given vector function and $x(t) = (x_1(t), ..., x_n(t))^T$ to the none-homogenous system with the initial condition $x(t_0) = x_0$*

$$\frac{dx(t)}{dt} = Ax(t) + f(t)$$

is given by

$$x(t) = e^{(t-t_0)A}x_0 + \int_{t_0}^{t} e^{(t-s)A}f(s)ds$$

Lemma A.3.3. *The solution to the system given in (2.2) is none-negative for given none-negative initial conditions $(u_{n0}^{(g)}, s_{n0}^{(g)}, s_{c0}^{(g)}) \geq 0$ for all genes g. We assume that rate parameters for all genes $(\alpha_{gk}, \beta_g, \nu_g, \gamma_g)$ are none-negative.*

Proof. We will start by considering the differential equation describing the temporal change of unspliced abundance of RNA in the nucleus

$$\frac{du_n^{(g)}(t)}{dt} = \alpha_{gk} - \beta_g u_n^{(g)}(t).$$

Assuming the system starts at time t_0 and $u_n^{(g)}(t_0) \geq 0$, it directly follows that if there exists a time point $t^* > t_0$ such that $u_n^{(g)}(t^*) < 0$, there need to exist a time point $t' \in [t_0, t^*)$ such that $u_n^{(g)}(t') = 0$. W.l.o.g. we can assume that t' is the smallest root of u_n. By definition of the dynamical system, the function is increasing in t' and by the continuity of the solution to the system, it follows that

$$\frac{du_n^{(g)}(t')}{dt} = \alpha_{gk} \geq 0$$
$$u_n^{(g)}([t', t' + \epsilon]) \geq 0.$$

Thus it holds that after the first potential root of the function, the function increases to be positive again

$$u_n^{(g)}(t) \geq 0 \ \ \forall t \in [t_0, t')$$
$$u_n^{(g)}(t') = 0$$
$$u_n^{(g)}(t) \geq 0 \ \ \forall t \in [t', t' + \epsilon]$$

Inductively we can conclude that $u_n^{(g)}(t) \geq 0$ for all $t \in [t_0, \infty)$, contradicting the assumption that there exists a time point $t^* > t_0$ such that $u_n^{(g)}(t^*) < 0$.
Let us now consider

$$\frac{ds_n^{(g)}(t)}{dt} = \beta_g u_n^{(g)}(t) - \nu_g s_n^{(g)}(t).$$

Similarly as in the first case, and using that $u_n^{(g)}(t) \geq 0$ for all $t \in [t_0, \infty)$, we can follow exactly the same steps to show that $s_n^{(g)}(t) \geq 0$ for all $t \in [t_0, \infty)$. Finally, using the none-negativity of the first two solutions, and applying the same steps, we can conclude that $s_c^{(g)}(t) \geq 0$ for all $t \in [t_0, \infty)$. □

Definition A.3.4. *A model is identifiable if it is theoretically possible to learn the true values of the model's underlying parameters after obtaining an infinite number of observations from it.*

Lemma A.3.5. *The rate parameters $(\alpha_{gk}, \beta_g, \nu_g, \gamma_g)$, as well as the time parameters (t, t_0) are none-identifiable for the system given in (2.2).*

Proof. Let $\theta_g^k = (\alpha_{gk}, \beta_g, \nu_g, \gamma_g, t, t_0)$ be the set of ground truth parameters underlying the system describing the splicing dynamics in (2.1) for a gene in state k and let $\phi_g^k = (\lambda_g\alpha_{gk}, \lambda_g\beta_g, \lambda_g\nu_g, \lambda_g\gamma_g, \frac{1}{\lambda_g}t, \frac{1}{\lambda_g}t_0)$ be a rescaled version of the ground truth parameters for $\lambda_g \in \mathbb{R}^+$. We aim to show that both sets of parameters will result in the same transcript abundances. Let us first consider the unspliced abundance for the parameter set ϕ_g^k

$$\begin{aligned} u_n^{(g)}(\phi_g^k) &= u_{n0}^{(g)} e^{-\lambda_g\beta_g(t-t_0)\frac{1}{\lambda_g}} + \frac{\lambda_g\alpha_{gk}}{\lambda_g\beta_g}(1 - e^{-\lambda_g\beta_g(t-t_0)\frac{1}{\lambda_g}}) \\ &= u_n^{(g)}(\theta_g^k), \end{aligned}$$

where the scaling factor λ_g is canceled out, thus resulting in the same solution. The same holds for the spliced abundances in the nucleus

$$\begin{aligned} s_n^{(g)}(\phi_g^k) &= s_{n0}^{(g)} e^{-\lambda_g\nu_g(t-t_0)\frac{1}{\lambda_g}} + \frac{\lambda_g\alpha_{gk}}{\lambda_g\nu_g}(1 - e^{-\lambda_g\nu_g(t-t_0)\frac{1}{\lambda_g}}) \\ &\quad + \frac{\lambda_g(a_{gk} - \beta_g u_{n0}^{(g)})}{\lambda_g(\nu_g - \beta_g)}(e^{-\lambda_g\nu_g(t-t_0)\frac{1}{\lambda_g}} - e^{-\lambda_g\beta_g(t-t_0)\frac{1}{\lambda_g}}) \\ &= s_n^{(g)}(\theta_g^k) \end{aligned}$$

and the cytoplasm

$$\begin{aligned} s_c^{(g)}(\phi_g^k) &= \lambda_g^2\nu_g\beta_g\Bigg(\frac{\frac{\lambda_g\alpha_{gk}}{\lambda_g\beta_g}(1 - e^{-\lambda_g\beta_g(t-t_0)\frac{1}{\lambda_g}}) + u_{n0}^{(g)} e^{-\lambda_g\beta_g(t-t_0)\frac{1}{\lambda_g}}}{\lambda_g^2(\nu_g - \beta_g)(\gamma_g - \beta_g)} \\ &\quad - \frac{\frac{\lambda_g\alpha_{gk}}{\lambda_g\nu_g}(1 - e^{-\lambda_g\nu_g(t-t_0)\frac{1}{\lambda_g}}) + u_{n0}^{(g)} e^{-\lambda_g\nu_g(t-t_0)\frac{1}{\lambda_g}}}{\lambda_g^2(\gamma_g - \nu_g)(\nu_g - \beta_g)} \\ &\quad + \frac{\frac{\lambda_g\alpha_{gk}}{\lambda_g\gamma_g}(1 - e^{-\lambda_g\gamma_g(t-t_0)\frac{1}{\lambda_g}}) + u_{n0}^{(g)} e^{-\lambda_g\gamma_g(t-t_0)\frac{1}{\lambda_g}}}{\lambda_g^2(\gamma_g - \nu_g)(\gamma_g - \beta_g)}\Bigg) \\ &\quad + \frac{\lambda_g\nu_g}{\lambda_g(\gamma_g - \nu_g)}(e^{-\lambda_g\nu_g(t-t_0)\frac{1}{\lambda_g}} - e^{-\lambda_g\gamma_g(t-t_0)\frac{1}{\lambda_g}})s_{n0}^{(g)} + e^{-\lambda_g\gamma_g(t-t_0)\frac{1}{\lambda_g}} s_{c0}^{(g)} \\ &= s_n^{(g)}(\theta_g^k). \end{aligned}$$

Thus, the model is none-identifiable and we will consider rate parameter ratios to compare the inferred rates with the true underlying rates of the simulated data, such that the scaling factor λ_g is canceled out. The none identifiability can be attributed to the inferred time scale, which is a rescaled version of the true time scale. As a consequence, the rate parameters are rescaled versions of the true parameters. □

List of Figures

Bibliography

[1] L. Heumos et al. “Best practices for single-cell analysis across modalities”. In: *Nature Reviews Genetics* (Mar. 2023).

[2] J. A. Briggs et al. “The dynamics of gene expression in vertebrate embryogenesis at single-cell resolution”. In: *Science* 360.6392 (June 2018).

[3] D. T. Montoro et al. “A revised airway epithelial hierarchy includes CFTR-expressing ionocytes”. In: *Nature* 560.7718 (Aug. 2018), pp. 319–324.

[4] K. R. Moon et al. “Visualizing structure and transitions in high-dimensional biological data”. In: *Nature Biotechnology* 37.12 (Dec. 2019), pp. 1482–1492.

[5] C. Trapnell et al. “The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells”. In: *Nature Biotechnology* 32.4 (Mar. 2014), pp. 381–386.

[6] A. Zbiciak and T. Markiewicz. “A new extraordinary means of appeal in the Polish criminal procedure: the basic principles of a fair trial and a complaint against a cassatory judgment”. en. In: *Access to Justice in Eastern Europe* 6.2 (Mar. 2023), pp. 1–18.

[7] L. Deconinck et al. “Recent advances in trajectory inference from single-cell omics data”. In: *Current Opinion in Systems Biology* 27 (2021), p. 100344. ISSN: 2452-3100.

[8] C. Li et al. “Single-cell multi-omic velocity infers dynamic and decoupled gene regulation”. In: (Dec. 2021).

[9] G. L. Manno et al. “RNA velocity in single cells”. In: *bioRxiv* (2017). eprint: `https://www.biorxiv.org/content/early/2017/10/19/206052.full.pdf`.

[10] V. Bergen et al. “Generalizing RNA velocity to transient cell states through dynamical modeling”. In: *bioRxiv* (2019). eprint: `https://www.biorxiv.org/content/early/2019/10/29/820936.full.pdf`.

[11] P. Weiler et al. “A Guide to Trajectory Inference and RNA Velocity”. In: *Single Cell Transcriptomics: Methods and Protocols*. Ed. by R. A. Calogero and V. Benes. New York, NY: Springer US, 2023, pp. 269–292. ISBN: 978-1-0716-2756-3.

[12] P. Weiler et al. “Unified fate mapping in multiview single-cell data”. In: (July 2023).

[13] B. Pijuan-Sala et al. “A single-cell molecular map of mouse gastrulation and early organogenesis”. In: *Nature* 566.7745 (Feb. 2019), pp. 490–495.

[14] A. Gayoso et al. “Deep generative modeling of transcriptional dynamics for RNA velocity analysis in single cells”. In: *bioRxiv* (2022). eprint: `https://www.biorxiv.org/content/early/2022/08/15/2022.08.12.503709.full.pdf`.

[15] A. Wagner, A. Regev, and N. Yosef. “Revealing the vectors of cellular identity with single-cell genomics”. In: *Nature Biotechnology* 34.11 (Nov. 2016), pp. 1145–1160. ISSN: 1546-1696.

[16] B. Alberts et al. *Molecular Biology of the Cell*. 4th. Garland, 2002.

[17] J. Buenrostro et al. “ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide”. In: (2015).

[18] M. Stoeckius et al. “Simultaneous epitope and transcriptome measurement in single cells”. In: *Nature Methods* 14.9 (July 2017), pp. 865–868.

[19] V. Bergen et al. “RNA velocity-current challenges and future perspectives”. In: *Molecular systems biology* 17 (Aug. 2021), e10282.

[20] D. P. Kingma and M. Welling. *Auto-Encoding Variational Bayes*. 2022. arXiv: 1312.6114 [stat.ML].

[21] L. McInnes, J. Healy, and J. Melville. *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction*. 2018.

[22] A. Bastidas-Ponce et al. “Massive single-cell mRNA profiling reveals a detailed roadmap for pancreatic endocrinogenesis”. In: *Development* (Jan. 2019).

[23] D. Klein et al. “Mapping cells through time and space with moscot”. In: (May 2023).

[24] Z.-J. Cao and G. Gao. “Multi-omics single-cell data integration and regulatory inference with graph-linked embedding”. In: *Nature Biotechnology* 40.10 (May 2022), pp. 1458–1466.

[25] R. Lopez et al. “Deep generative modeling for single-cell transcriptomics”. In: *Nature Methods* 15.12 (Nov. 2018), pp. 1053–1058.

[26] R. Lopez et al. *Tutorials*. https://docs.scvi-tools.org/en/stable/tutorials/index.html. [Online; accessed 20-September-2023]. 2018.

[27] Z.-J. Cao and G. Gao. *Tutorials*. https://scglue.readthedocs.io/en/latest/tutorials.html. [Online; accessed 20-September-2023]. 2022.

[28] M. D. Luecken et al. “Benchmarking atlas-level data integration in single-cell genomics”. In: *Nature Methods* 19.1 (Dec. 2021), pp. 41–50.

[29] F. A. Wolf, P. Angerer, and F. J. Theis. “SCANPY: large-scale single-cell gene expression data analysis”. In: *Genome Biology* 19.1 (Feb. 2018).

[30] I. Tirosh et al. “Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq”. In: *Science* 352.6282 (Apr. 2016), pp. 189–196.

[31] R. Satija et al. “Spatial reconstruction of single-cell gene expression data”. In: *Nature Biotechnology* 33.5 (Apr. 2015), pp. 495–502.

[32] M. Lange et al. “CellRank for directed single-cell fate mapping”. In: *Nature Methods* 19.2 (Jan. 2022), pp. 159–170.

[33] K. Tokunaga et al. “Nucleocytoplasmic transport of fluorescent mRNA in living mammalian cells: nuclear mRNA export is coupled to ongoing gene transcription”. In: *Genes to Cells* 11.3 (Mar. 2006), pp. 305–317.

[34] M.-j. Luo and R. Reed. “Splicing is required for rapid and efficient mRNA export in metazoans”. In: *Proceedings of the National Academy of Sciences* 96.26 (Dec. 1999), pp. 14937–14942.

[35] D. Lefaudeux et al. “Kinetics of mRNA nuclear export regulate innate immune response gene expression”. In: *Nature Communications* 13.1 (Nov. 2022).

[36] S. Gueroussov et al. “Analysis of mRNA Nuclear Export Kinetics in Mammalian Cells by Microinjection”. In: *Journal of Visualized Experiments* 46 (Dec. 2010).

[37] T. Chen and B. van Steensel. “Comprehensive analysis of nucleocytoplasmic dynamics of mRNA in Drosophila cells”. In: *PLOS Genetics* 13.8 (Aug. 2017). Ed. by M. Choder, e1006929.

[38] J. M. Müller et al. “Nuclear export is a limiting factor in eukaryotic mRNA metabolism”. In: (May 2023).

[39] A. Scialdone et al. “Computational assignment of cell-cycle stage from single-cell transcriptome data”. In: *Methods* 85 (Sept. 2015), pp. 54–61.

[40] C. J. Hsiao et al. “Characterizing and inferring quantitative cell cycle phase in single-cell RNA-seq data analysis”. In: *Genome Research* 30.4 (Apr. 2020), pp. 611–621.

[41] D. Schwabe et al. “The transcriptome dynamics of single cells during the cell cycle”. In: *Molecular Systems Biology* 16.11 (Nov. 2020).

[42] D. Mahdessian et al. “Spatiotemporal dissection of the cell cycle with single-cell proteogenomics”. In: *Nature* 608.7924 (Aug. 2022), E32–E32.

[43] N. Battich et al. “Sequencing metabolically labeled transcripts in single cells reveals mRNA turnover strategies”. In: *Science* 367.6482 (Mar. 2020), pp. 1151–1156.

[44] B. Marsh and R. Blelloch. “Single nuclei RNA-seq of mouse placental labyrinth development”. In: *eLife* 9 (Nov. 2020).

[45] Q. Adewale et al. “Single-nucleus RNA velocity reveals synaptic and cell-cycle dysregulations missed by gene expression in neuropathologic Alzheimer disease”. In: (Nov. 2022).

[46] Wikipedia. *Alternatives Spleißen — Wikipedia, die freie Enzyklopädie*. [Online; Stand 3. Mai 2023]. 2021.

[47] A. Zeisel et al. “Coupled pre-mRNA and mRNA dynamics unveil operational strategies underlying transcriptional responses to stimuli”. In: *Molecular Systems Biology* 7.1 (2011), p. 529. eprint: `https://www.embopress.org/doi/pdf/10.1038/msb.2011.62`.

[48] M. Büttner et al. “A test metric for assessing single-cell RNA-seq batch correction”. In: *Nature Methods* 16.1 (Dec. 2018), pp. 43–49.